Journal of African Earth Sciences 118 (2016) 53–64

Contents lists available at ScienceDirect

Journal of African Earth Sciences

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

Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression

Ismail Colkesen, Emrehan Kutlug Sahin, Taskin Kavzoglu*

Gebze Technical University, Department of Geodetic and Photogrammetric Engineering, Gebze-Kocaeli, Turkey

ARTICLE INFO

Article history:
Received 13 October 2015
Received in revised form
19 February 2016
Accepted 23 February 2016
Available online 26 February 2016

Keywords:
Gaussian process
Support vector machine
Logistic regression
Landslide
Susceptibility mapping
Kernel learning

ABSTRACT

Identification of landslide prone areas and production of accurate landslide susceptibility zonation maps have been crucial topics for hazard management studies. Since the prediction of susceptibility is one of the main processing steps in landslide susceptibility analysis, selection of a suitable prediction method plays an important role in the success of the susceptibility zonation process. Although simple statistical algorithms (e.g. logistic regression) have been widely used in the literature, the use of advanced non-parametric algorithms in landslide susceptibility zonation has recently become an active research topic. The main purpose of this study is to investigate the possible application of kernel-based Gaussian process regression (GPR) and support vector regression (SVR) for producing landslide susceptibility map of Tonya district of Trabzon, Turkey. Results of these two regression methods were compared with logistic regression (LR) method that is regarded as a benchmark method. Results showed that while kernel-based GPR and SVR methods generally produced similar results (90.46% and 90.37%, respectively), they outperformed the conventional LR method by about 18%. While confirming the superiority of the GPR method, statistical tests based on ROC statistics, success rate and prediction rate curves revealed the significant improvement in susceptibility map accuracy by applying kernel-based GPR and SVR methods.

1. Introduction

Throughout the history, natural disasters have played a major role in the economic development and survival of humanity. During the past decades, natural hazards such as droughts, floods, earthquakes and landslides have caused major loss of human lives and livelihoods, the destruction of economic and social infrastructure, as well as environmental damages (Sivakumar 2005). Landslides, one of the most destructive natural disasters affecting a large number of people and properties, are geological phenomena that include a wide range of ground movement, such as rock falls, deep failure of slopes, and shallow debris flow (Nemčok et al., 1972; Varnes, 1978; Hutchinson, 1988; Sivakumar 2005; Guzzetti, 2006). Obtaining reliable and accurate information about landslide-prone areas in terms of susceptibility level and spatial distribution is essential to ensure the safety of human life and property. For this reason, determining landslide-susceptible zones and producing up-to-date landslide susceptibility maps providing valuable information for government agencies, planners and decision makers are required for hazard management and prevention studies.

A hazard zonation map that aims at predicting where slope failures (or mass movements) are most likely to occur is more accurately defined as a landslide susceptibility map (Brabb, 1984). In the production of these maps terrain features together with geological and climatic conditions of a given area must be considered in relation to past landslides. The effectiveness of susceptibility assessment is dependent on the method applied to produce the landslide susceptibility maps, as well as the quality of the data used. Without a proper evaluation and/or validation, resulting thematic maps would have a poor scientific value for the decision makers who cannot perform an adequate economic cost-benefit analysis to make the proper land-use planning (Chung and Fabbri, 2003; Trigila et al., 2015).

In recent years, various methods have been applied for landslide susceptibility mapping for study areas all over the world. These methods can be categorized as qualitative and quantitative approaches, also as heuristic, probabilistic, statistical and deterministic (Aleotti and Chowdhury, 1999; Pardeshi et al., 2013). A qualitative approach is based on the subjective judgment of an

* Corresponding author.
E-mail address: kavzoglu@gtu.edu.tr (T. Kavzoglu).

http://dx.doi.org/10.1016/j.jafrearsci.2016.02.019
1464-343X/

expert or a group of experts whereas the quantitative approach is based on mathematically rigorous objective methodologies (Neaupane and Piantanakulchai, 2006). To be more specific, the qualitative approach is based on expert opinion and determining the weights of criteria (Ayalew et al., 2004; Yalcin, 2008; Kavzoglu et al., 2014). Quantitative methods are based on numerical expressions of the relationship between controlling factors and landslides (Ayalew and Yamagishi, 2005; Akgun and Turk, 2010; Kavzoglu et al., 2015a). Within the quantitative methods, probabilistic and statistical approaches have been widely used to determine landslide susceptibility levels. The probabilistic models like frequency ratio, bivariate analysis, and multivariate analysis are more frequently used to determine the landslide susceptibility zones (Ahmed, 2014). The logistic regression has been widely used for landslide susceptibility assessment at local and regional scale (Atkinson and Massari, 1998; Ayalew and Yamagishi, 2005; Bui et al., 2011; Althuwaynee et al., 2014; Dou et al., 2015a).

Lately, non-parametric methods, such as random forests, neural networks and decision trees have been also employed for landslide susceptibility mapping (Gómez and Kavzoglu, 2005; Catani et al., 2013; Kavzoglu et al., 2015a; Were et al., 2015). Particularly, support vector machine, a popular kernel-based machine learning algorithm, has drawn considerable interest in the past few years for landslide susceptibility mapping due to its robustness and effectiveness in solving complex-structured data sets (Yilmaz, 2010; Kavzoglu et al., 2014; Goetz et al., 2015). Another kernel-based machine learning algorithm called Gaussian process regression has been recently applied to many problems in machine learning (Pasolli et al., 2010; Hultquista et al., 2014; Stulp and Sigaud, 2015), but it has very limited use in landslide susceptibility zonation. It is formulated and interpreted as a Bayesian version of support vector machine (Rasmussen and Williams, 2006). This study attempts to evaluate the kernel-based machine learning algorithms, namely Gaussian process and support vector machine for landslide susceptibility assessment of a mountainous region facing active landslides. Their performances were compared with the result of logistic regression to evaluate their effectiveness. Well-known statistical accuracy measures namely, overall accuracy, area under the curve (AUC), success rate and prediction rate curves were estimated for statistical comparison of the method performances.

2. Study area and data

The area selected for this study covers the Tonya district of Trabzon in Turkey (Fig. 1). The overall area of the district is about 200 km^2, located between 39° 13′ and 39° 21′ west–east longitudes and 40° 57′ and 40° 43′ north–south latitudes. The study area is characterized by steep slopes associated with shallow soils in mountainous regions. Annual rainfall of 2200 mm and high slope gradients reaching to 64° are the key triggering factors for landslides. It can be stated that climatic conditions, topography and soil attributes have major contributions for landslide occurrences in the study area.

The geological map of the study area, including 8 lithological units, was created from 1:100,000-scale geological map published by the General Directorate of Mineral Research and Exploration of Turkey in 1998. Geological units of the study area are mainly formed by Lias–Dogger (Jlh) Upper Cretaceous–Paleocene (Cru1, Cru2, Cru3, Cru4b and Cru5b) and Eocene (Gama2, Ev) epochs (Fig. 2). The geological formations of the study area are primarily Eocene and Upper Cretaceous–Paleocene system. Eight geologic formations exist in the study area. Within these formations, Ev is the main geological formation with 38% coverage which is mainly composed of pyroclastic, greenish-grey basalt and andesite, sandy limestone and tuff units.

In order to evaluate the kernel-based method performances, thematic maps of a number of causative factors were created and utilized in this study. The causative factors considered here were slope angle, slope aspect, elevation, lithology, land cover/land use (LULC), NDVI, profile curvature and TWI. It should be mentioned that recent studies conducted in the Black Sea region also considered similar conditioning factors (Akgun et al., 2008; Nefeslioglu et al., 2008a; Yalcin et al., 2011; Kavzoglu et al., 2015b). A grid cell model was applied to estimate the susceptibility since it is the most popular approach for spatial representation of the datasets. The factor maps were available at different scales or intervals, so each factor map was standardized to the same scale (i.e. 30 × 30 m) for further analysis. As stated by Fressard et al. (2014), the choice of the raster images grid cell size is guided by both reference to the imposed cartographic scale and the original scale/resolution of the available data sets. Regarding the original cell size and contour line density on the available thematic maps and satellite imagery (Landsat TM) from which LULC map was produced through supervised image classification. In addition, the maps in a continuous data format were reclassified into discrete subclasses as an essential step to standardize the factor maps, which is a prerequisite for analysis in GIS. Therefore, the factor maps with continuous data (slope, NDVI, elevation, profile curvature, and TWI) were classified using natural break approach so that continuous data were converted to classes in specific intervals (Table 1). Other factor maps (lithology, aspect and LULC) were reclassified into classes in order to make the same output scaling. The Digital Elevation Model (DEM) produced from 1:25,000 scale topography maps was used to create the topographic parameters (elevation, slope angle, slope aspect, profile curvature and TWI) related to the landslide activity. While elevations in the study area were ranging from 201 m to 2412 m, slope angles varied between 0° and 64°.

Landslide inventory maps documenting the extent of landslide phenomena in a specific region and representing information about their spatial distribution, types, vulnerability, recurrence and statistics of slope failures are important sources for mapping landslide susceptibilities (Guzetti et al., 2012; Dou et al., 2015b). Landslide inventory maps were produced to introduce the locations of the past landslides and potential non-landslide sites. Several sampling strategies have been applied in the literature to construct landslide susceptibility maps. For example Nefeslioglu et al. (2008b) investigated conceptual difference of susceptibility models by applying two different sampling methods: from all landslide area with depletion and accumulation zones and from a zone which almost represents pre-failure conditions. On the other hand, some researchers preferred to use points to represent the spatial location of landslides (Neuhäuser et al. 2012; Bui et al., 2012). Dai and Lee (2002) considered only the data source areas defined by a surface of rupture which comprises the main scarp and the scarp floor and used a spatially uniform sampling scheme excluding a 40-m buffer zone. Furthermore, the landslide area with depletion and accumulation zones named as seed-cells to represent pre-failure conditions was proposed by Suzen and Doyuran (2004). ‘Sampling Circle’ approach was proposed to define shallow landslide initiation in the mapping units in susceptibility evaluations by Nefeslioglu et al. (2011). Several previous researches preferred to use rapture zones of landslides as a sampling unit in landslide susceptibility mapping (Remondo et al., 2003; Santacana et al., 2003).

Landslide inventory map used in this study was prepared after “Turkish Landslide Inventory Mapping Project”, started in 1997, by General Directorate of Mineral Research and Exploration (MTA). In the project, landslides were delimited on 1:25,000 scale topographic base maps by interpretation of aerial photographs (1:35,000–1:10,000 scale) and field investigations. The landslides (i.e. active and inactive) were then digitized and stored in GIS