Author's Accepted Manuscript

Analysis of Training Sample Selection Strategies for Regression-Based Quantitative Landslide Susceptibility Mapping Methods

Arzu Erener, A. Abdullah Sivas, A. Sevtap Selcuk-Kestel, H. Sebnem Düzgün

Cite this article as: Arzu Erener, A. Abdullah Sivas, A. Sevtap Selcuk-Kestel and H. Sebnem Düzgün, Analysis of Training Sample Selection Strategies for Regression-Based Quantitative Landslide Susceptibility Mapping Methods, *Computers and Geosciences,* http://dx.doi.org/10.1016/j.cageo.2017.03.022

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Analysis of Training Sample Selection Strategies for Regression-Based Quantitative Landslide Susceptibility Mapping Methods

Arzu Erener*, A. Abdullah Sivas**, A. Sevtap Selcuk-Kestel** and H. Sebnem Düzgün***

*Engineering Department, Geodetic and Geoinformation Engineering, Kocaeli University, 41380, Kocaeli,Turkey, arzu.erener@kocaeli.edu.tr
**Institute of Applied Mathematics, Middle East Technical University, 06800, Ankara, Turkey, sivas@metu.edu.tr, skestel@metu.edu.tr
***Mining Engineering Department, Geodetic and Geographic Information Technologies, Middle East Technical University, 06800, Ankara, Turkey, duzgun@metu.edu.tr

Abstract: All of the quantitative landslide susceptibility mapping (QLSM) methods requires two basic data types, namely, landslide inventory and factors that influence landslide occurrence (landslide influencing factors, LIF). Depending on type of landslides, nature of triggers and LIF, accuracy of the QLSM methods differs. Moreover, how to balance the number of 0 (nonoccurrence) and 1 (occurrence) in the training set obtained from the landslide inventory and how to select which one of the 1's and 0's to be included in QLSM models play critical role in the accuracy of the QLSM. Although performance of various QLSM methods is largely investigated in the literature, the challenge of training set construction is not adequately investigated for the QLSM methods. In order to tackle this challenge, in this study three different training set selection strategies along with the original data set is used for testing the performance of three different regression methods namely Logistic Regression (LR), Bayesian Logistic Regression (BLR) and Fuzzy Logistic Regression (FLR). The first sampling strategy is proportional random sampling (PRS), which takes into account a weighted selection of landslide occurrences in the sample set. The second method, namely non-selective nearby sampling (NNS), includes randomly selected sites and their surrounding neighboring points at certain preselected distances to include the impact of clustering. Selective nearby sampling (SNS) is the third method, which concentrates on the group of 1's and their surrounding neighborhood. A randomly selected group of landslide sites and their neighborhood are considered in the analyses similar to NNS parameters. It is found that LR-PRS, FLR-PRS and BLR-Whole Data set-ups, with order, yield the best fits among the other alternatives. The results indicate that in QLSM based on regression models, avoidance of spatial correlation in the data set is critical for the model's performance.

Keywords: Logistic Regression (LR); Bayesian Logistic Regression (BLR); Fuzzy Logistic Regression (FLR); Training Sample Strategies; Landslide

1. Introduction

The quantitative landslide susceptibility mapping (QLSM) methods require two basic data types, namely, landslide inventory and factors that influence landslide occurrence (landslide influencing factors, LIF). The LIF's are also called landslide predisposing factors, landslide conditional factors, or slope instability factors (Hussin, et al. 2016). Selection of LIF's and data preparation for the selected LIF play critical role in the performance of the QLSM method. The first step in applying any QLSM method is to define the mapping units, where the LIF's and inventory data are required to be prepared accordingly. There are basically five different types of mapping units used in susceptibility mapping which are grid-cells (pixels), terrain