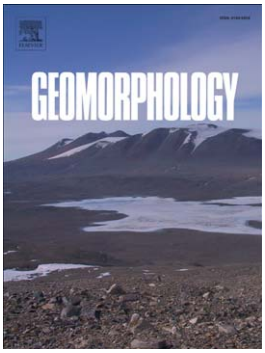

Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods

Jung-Hyun Lee, Maher Ibrahim Sameen, Biswajeet Pradhan, Hyuck-Jin Park

PII: S0169-555X(16)31106-0
DOI: doi:10.1016/j.geomorph.2017.12.007
Reference: GEOMOR 6248

To appear in: *Geomorphology*

Received date: 19 November 2016
Revised date: 7 August 2017
Accepted date: 3 December 2017

Please cite this article as: Lee, Jung-Hyun, Sameen, Maher Ibrahim, Pradhan, Biswajeet, Park, Hyuck-Jin, Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods, *Geomorphology* (2017), doi:10.1016/j.geomorph.2017.12.007

Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods

Jung-Hyun Lee[1], Maher Ibrahim Sameen[2], Biswajeet Pradhan[2,3,4*], Hyuck-Jin Park[1*]

[1]Department of Geoinformation Engineering, Sejong University 209 Neungdong-ro Gwangjin-gu, Seoul 05006 Republic of Korea

[2]Department of Civil Engineering, Faculty of Engineering, Universiti Putra Malaysia, Malaysia, 43400, UPM, Serdang, Tel. +603-8946-6383

[3]School of Systems, Management and Leadership, Faculty of Engineering and Information Technology, University of Technology Sydney, Building 11, Level 06, 81 Broadway, Ultimo NSW 2007 (PO Box 123), Australia

[3]Department of Energy and Mineral Resources Engineering, Sejong University 209 Neungdong-ro Gwangjin-gu, Seoul 05006 Republic of Korea

*Email. Biswajeet24@gmail.com, hjpark@sejong.ac.kr (Corresponding author)

Abstract

This study evaluated the generalizability of five models to select a suitable approach for landslide susceptibility modeling in data-scarce environments. In total, 418 landslide inventories and 18 landslide conditioning factors were analyzed. Multicollinearity and factor optimization were investigated before data modeling, and two experiments were then conducted. In each experiment, five susceptibility maps were produced based on support vector machine (SVM), random forest (RF), weight-of-evidence (WoE), ridge regression (Rid_R), and robust regression (RR) models. The highest accuracy (*AUC* = 0.85) was achieved with the SVM model when either the full or limited landslide inventories were used. Furthermore, the RF and WoE models were severely affected when less landslide samples were used for training. The other models were affected slightly when the training samples were limited.