Geomorphology 262 (2016) 8–23

Contents lists available at ScienceDirect

Geomorphology

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

Exploring discrepancies between quantitative validation results and the geomorphic plausibility of statistical landslide susceptibility maps

Stefan Steger [a,*], Alexander Brenning [b], Rainer Bell [a], Helene Petschko [b], Thomas Glade [a]

[a] Department of Geography and Regional Research, University of Vienna, Universitätsstraße 7, A-1010 Vienna, Austria
[b] Department of Geography, Friedrich Schiller University Jena, Loebdergraben 32, 07743 Jena, Germany

ARTICLE INFO

Article history:
Received 22 August 2015
Received in revised form 11 March 2016
Accepted 15 March 2016
Available online 17 March 2016

Keywords:
Landslide susceptibility
Geomorphic plausibility
Spatial cross-validation
Statistical methods
Machine learning techniques

ABSTRACT

Empirical models are frequently applied to produce landslide susceptibility maps for large areas. Subsequent quantitative validation results are routinely used as the primary criteria to infer the validity and applicability of the final maps or to select one of several models. This study hypothesizes that such direct deductions can be misleading. The main objective was to explore discrepancies between the predictive performance of a landslide susceptibility model and the geomorphic plausibility of subsequent landslide susceptibility maps while a particular emphasis was placed on the influence of incomplete landslide inventories on modelling and validation results.

The study was conducted within the Flysch Zone of Lower Austria (1,354 km^2) which is known to be highly susceptible to landslides of the slide-type movement. Sixteen susceptibility models were generated by applying two statistical classifiers (logistic regression and generalized additive model) and two machine learning techniques (random forest and support vector machine) separately for two landslide inventories of differing completeness and two predictor sets. The results were validated quantitatively by estimating the area under the receiver operating characteristic curve (*AUROC*) with single holdout and spatial cross-validation technique. The heuristic evaluation of the geomorphic plausibility of the final results was supported by findings of an exploratory data analysis, an estimation of odds ratios and an evaluation of the spatial structure of the final maps.

The results showed that maps generated by different inventories, classifiers and predictors appeared differently while holdout validation revealed similar high predictive performances. Spatial cross-validation proved useful to expose spatially varying inconsistencies of the modelling results while additionally providing evidence for slightly overfitted machine learning-based models. However, the highest predictive performances were obtained for maps that explicitly expressed geomorphically implausible relationships indicating that the predictive performance of a model might be misleading in the case a predictor systematically relates to a spatially consistent bias of the inventory. Furthermore, we observed that random forest-based maps displayed spatial artifacts. The most plausible susceptibility map of the study area showed smooth prediction surfaces while the underlying model revealed a high predictive capability and was generated with an accurate landslide inventory and predictors that did not directly describe a bias. However, none of the presented models was found to be completely unbiased.

This study showed that high predictive performances cannot be equated with a high plausibility and applicability of subsequent landslide susceptibility maps. We suggest that greater emphasis should be placed on identifying confounding factors and biases in landslide inventories. A joint discussion between modelers and decision makers of the spatial pattern of the final susceptibility maps in the field might increase their acceptance and applicability.

1. Introduction

Landslides are potentially damaging phenomena caused by interacting natural and anthropogenic factors. However, in many cases their underlying processes are yet to be fully understood (Crozier, 1989; Carrara et al., 1999; Glade et al., 2005). Within this study, the term "landslide" is used for the slide-type movement of earth masses according to Cruden and Varnes (1996) and Dikau et al. (1996).

The likelihood or spatial probability of a landslide event occurring at a specific site is referred to as landslide susceptibility (Brabb, 1984; Guzzetti et al., 1999; Glade et al., 2005). Landslide susceptibility maps display a relative estimate of where landslides are more likely to occur in the future due to a set of environmental conditions. Thus, they do

* Corresponding author.
E-mail addresses: stefan.steger@univie.ac.at (S. Steger), alexander.brenning@uni-jena.de (A. Brenning), rainer.bell@univie.ac.at (R. Bell), helene.petschko@uni-jena.de (H. Petschko), thomas.glade@univie.ac.at (T. Glade).

http://dx.doi.org/10.1016/j.geomorph.2016.03.015

not provide information on the magnitude or temporal occurrence of upcoming events (Guzzetti et al., 2005).

At a regional scale, statistical classification approaches are most commonly applied to construct landslide susceptibility maps (Cascini, 2008; Van Westen et al., 2008). They are based on the assumption that future landslides occur under similar conditions as past landslides (Carrara et al., 1995; Orme, 2002). Subsequently, an empirical and generalized relationship between a landslide inventory (response variable) and multiple predisposing factors (predictors) is built. Finally, this relation is then transferred to every unit (e.g. raster cell, terrain unit) of a study area (Van Westen et al., 2006). Besides "classical" multiple variable statistical approaches (e.g. logistic regression and discriminant analysis), more flexible supervised machine learning techniques (e.g. support vector machines and random forest) have recently become popular to model landslide susceptibility (Brenning, 2005; Ballabio and Sterlacchini, 2012; Vorpahl et al., 2012; Catani et al., 2013).

The number of comparative studies in this field demonstrates that modelers have to make several choices that will influence the modelling results (Brenning, 2012a). For instance, the selection of the classification method (Devkota et al., 2012; Schicker and Moon, 2012; Kavzoglu et al., 2014; Goetz et al., 2015) and predictors (Iovine et al., 2014) as well as the type and quality of the landslide inventory (Zêzere, 2002; Steger et al., 2015) highly influence the modelling outcomes. Furthermore, the final results are also dependent on the spatial resolution of input data (Catani et al., 2013), the number of terrain units (Guzzetti et al., 2006; Catani et al., 2013), the applied sampling strategy (Nefeslioglu et al., 2008; Regmi et al., 2014) and the sample sizes (Heckmann et al., 2014).

The large number of possible modifications makes an in-depth validation of the final modelling results even more important. Early quantitative landslide susceptibility models were validated heuristically by comparing mapped landslides of an area with the produced susceptibility maps (Carrara, 1983; Brabb, 1984). Meanwhile, quantitative validation approaches replaced qualitative procedures to objectively evaluate the performance of statistical landslide susceptibility models (Frattini et al., 2010). In this context, Chung and Fabbri (1999) were among the first to highlight the necessity of a thorough quantitative evaluation of such predictions. Nowadays, threshold-independent performance measures like the area under the receiver operating characteristic curve (*AUROC*) (Brenning, 2005; Beguería, 2006; Frattini et al., 2010) or the area under the prediction rate curve (Chung and Fabbri, 2003; Remondo et al., 2003) are regularly assessed for an independent test sample or using cross-validation (Brenning, 2005) to estimate the predictive performance of a landslide susceptibility model. A review of the literature revealed that many authors consider these quantitative estimates as the only decision basis to favor a certain model over another and/or to deduce the usability of landslide susceptibility maps. Nevertheless, these conclusions are put in doubt when susceptibility maps with substantially different appearances achieve similar predictive model performances (Bell, 2007; Sterlacchini et al., 2011; Steger et al., 2015), putting modelers as well as decision-makers in a difficult position.

Several studies successfully expanded the quality evaluation of statistical landslide susceptibility models by quantifying uncertainties in terms of standard errors and confidence intervals of predicted probabilities (Rossi et al., 2010; Petschko et al., 2014a; Reichenbach et al., 2014). However, due to the lack of process-related knowledge and the assumptions that have to be taken as well as limitations in the availability of crucial spatial information, many uncertainties inherent in landslide prediction models are known to be not ascertainable by means of quantitative procedures (Ardizzone et al., 2002; Guzzetti et al., 2006; Van Westen et al., 2008). Only a few studies additionally discuss the appearance of landslide susceptibility maps (e.g. Bell, 2007; Demoulin and Chung, 2007; Sterlacchini et al., 2011; Goetz et al., 2015).

The main objective of this study is to examine possible discrepancies between the geomorphic plausibility of landslide susceptibility maps and statistical validation results. In particular, this study explores the effect of different classification methods and input data on the appearance of the final landslide susceptibility maps and associated model performance measures while a focus is set on the influence of possible mapping or reporting biases regularly inherent in landslide inventories (e.g. Brardinoni et al., 2003; Malamud et al., 2004; Guzzetti et al., 2012).

Within this study, geomorphic plausibility is used in analogy to the concept of biological plausibility regularly addressed in the fields of medicine and epidemiology to evaluate whether an observed association "*makes biological sense*" (Hoffer, 2003, p.180) or is in apparent conflict with scientific knowledge (Hill, 1965; Holland, 1986). In this context, we propose that a geomorphically plausible statistical landslide susceptibility map should demonstrate neither biases related to input data nor algorithm based artifacts while a high *AUROC* value should verify the success of the underlying prediction. We consider findings of an exploratory data analysis, odds ratios of modelling results and the spatial structure of the final maps to support our intrinsically subjective evaluations.

2. Study area

The study area, the landslide-prone Rheno-Danubian Flysch Zone of the province of Lower Austria, covers an area of 1,354 km^2 and its elongated shape extents over 149 km from west to east (Fig. 1). The prevalent alternating sandstone-marl and sandstone-siltstone layers are a result of large turbidity currents, which episodically transported large amounts of uncompact sediments from the continental shelf into the oceanic basin during the Cretaceous to Early Tertiary. In the periods between this deposition, hemipelagic marls and silts accumulated resulting in rhythmically interbedded strata. During the Alpidic orogenesis, these layers were compressed, thrust towards the north and uplifted. Thus, the typical undulating Flysch landscape is a result of a turbulent past (Wessely et al., 2006).

The widespread sequences of sedimentary rocks are highly erodible, deeply incised and well known to be susceptible to landsliding (Schwenk, 1992; Wessely et al., 2006; Damm and Terhorst, 2009; Petschko et al., 2014a). The prevalent clayey to silty layers, as well as specific deeply weathered sandstones (e.g. Mürbsandsteine), are likely to promote landslides triggered by precipitation and/or snow melt (Wessely et al., 2006). Typically, these downslope movements are related to an increase in water saturation inducing consistency changes of the fine material (e.g. swelling clays) while also water-related surcharge loading increases shear stresses. Several cases are known where artificial interventions destabilized slopes of the study area such as water supply, surcharge, and slope undercutting (Schwenk, 1992; Wessely et al., 2006). According to reports archived in the Building Ground Registry (BGR) since 1953, the Flysch- and Klippen Zones are the most landslide-prone areas within Lower Austria. BGR records show that 42% of all reported landslides were located within the Flysch Zone (7% of the total province area).

The gentle hilly landscape of the northeast and east is characterized by mean annual precipitation of <600 mm. Towards the southwest, the undulating landscape transitions into a low mountain range with steeper slopes and altitudes higher than 850 m a.s.l. The subalpine climatic conditions of the southwestern portion of the study area include mean annual precipitation above 1,300 mm (Skoda and Lorenz, 2007).

Intensively cultivated pastures cover 33% of the area and arable land 11%. Four percent of the study area is classified as settlements, whereas particularly towards the west a relatively dense road network links the prevalent dispersed farms. Large areas, especially in the eastern parts, are covered by forests. In total 52% of the study area is forested (deciduous forest = 40% and coniferous forest = 12%; Eder et al., 2011).