EGU2018-14922

A data driven approach to landslide susceptibility mapping in Great Britain

Chris Williams, Emma Bee, Claire Dashwood and Ben Marchant

British Geological Survey, Environmental Science Centre, Keyworth, Nottingham, NG12 5GG, United Kingdom

1. Introduction

Landslides are a geo-hazard which can have significant societal impacts including loss of human life, physical damage to infrastructure and financial loss. The ability to assess where landslides will occur is therefore of great interest for the public good and can be approached both theoretically and empirically.

With the ever increasing availability of spatial data, information on landslide events is now much more readily available ranging from initiation point coordinates to high (sub-metre) resolution topographic information and associated derivatives on affected (and unaffected) areas. Coupled with information on the geology of a region, it is possible to build up a detailed location specific profile of past events, all of which may prove useful for informing where future events may occur.

We present ***preliminary results from an assessment of various data to reassess current British landslide susceptibility datasets***. These could be used in future to provide additional information to support landslide forecasting. We define susceptibility as:

The potential for the occurrence of a hazard within a specified area. This is currently provided for by the BGS GeoSure Landslides product [1] which classifies landslide prone areas on an A-E (low-high) basis, based on heuristics as well as consideration of lithology, discontinuities and slope angle.

Data-driven analyses may provide further insights into where and why landslides occur. Using this knowledge, we hope to improve our current landslide susceptibility model. Consequently, this will enable us to be more confident in the identification of areas where landslides may occur in the future.

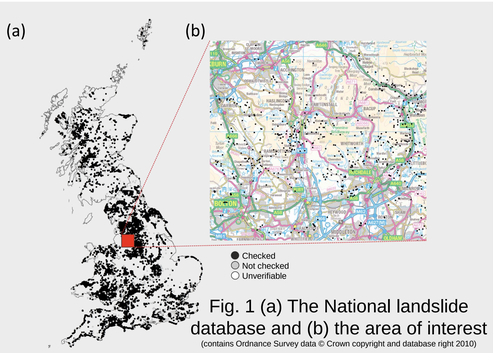

Fig. 1 (a) The National landslide database and (b) the area of interest

2. Data and study site overview

Our study area focuses on northern Greater Manchester and south east Lancashire in Northern England, on the western flank of the Pennines (Fig. 1). This is an area characterised by moorland in the open areas and the outskirts of a number of towns including Rochdale, Bolton and Blackburn. During the last ice age this region was covered by (~1 km thick) ice. Many of the landslides in this area occurred following the end of the last ice age (~13,000 yrs B.P.). Many of the resulting landslides have since been stable however they are at constant risk of reactivation [2]. Assessment therefore remains valid in such areas, especially with regard to future land use planning.

We consider a variety of landslide specific information and geological variables at locations where landslides have (478 records) and have not been recorded (1000 records). Recorded landslide locations were extracted from the National Landslide Database of Great Britain, managed by the BGS and freely available online [3] (Fig. 1a). At each location we consider the variables presented in Fig. 2, extracted from BGS datasets and using the NEXTMap® digital terrain model (5 m resolution).

3. Methodology

To ascertain the prediction importance of the variables extracted at each point as described in section 2, four models were used to test the classification of locations as being either *landslide* or *non-landslide*:

1. Generalized linear model (GLM) with slope as the only predictor;
2. Stepwise GLM automatically selecting predictors;
3. A single classification tree (with pruning based on cross-validation);
4. A random forest.

Model validation was achieved though spatial and random point removal through repeat selection over 100 model runs (Fig. 3). 200 points were omitted on each run. Spatial validation removed the closest 200 points to a randomly selected point identified from the 1478 sample points (section 2). Average model performance is presented in Table 1.

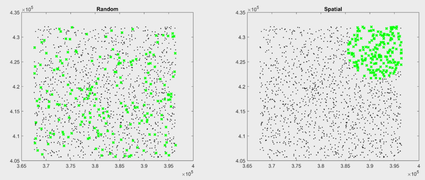

Fig. 3 Random and spatial point sampling.

To identify any internal relationships, the 478 verified landslides were subjected to a k-means clustering analysis (with a 5 cluster limit). Due to the mixed data types within the dataset (continuous, categorical, numerical and string), the distance matrix was constructed through the calculation of Gower distances, facilitated using the *daisy* tool in the R package *cluster* [4].

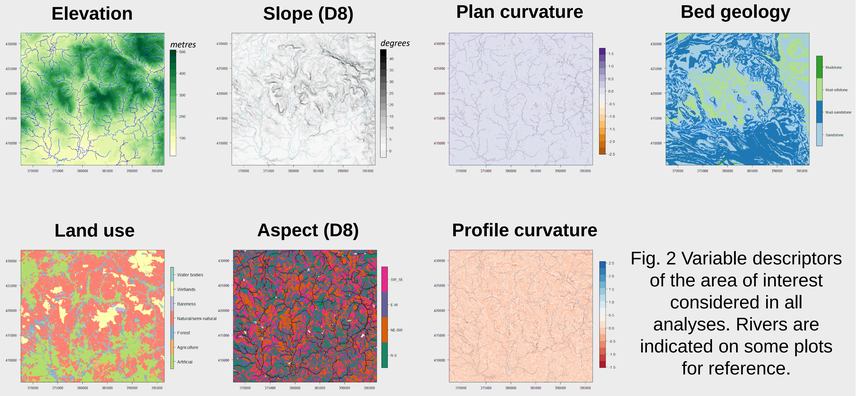

Fig. 2 Variable descriptors of the area of interest considered in all analyses. Rivers are indicated on some plots for reference.

4. Results

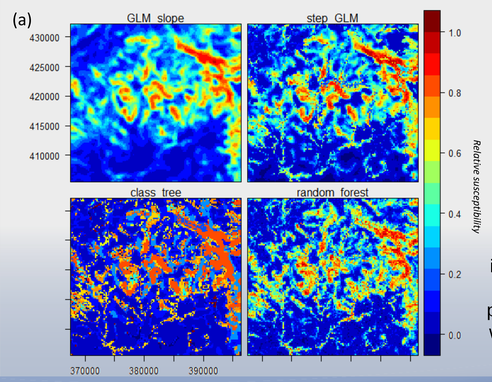

Fig. 4 (a) Relative landslide susceptibility per model implementation following averaging of random and spatial sampling permutations. These can be compared with the (b) current GeoSure landslide product.

Observed landslide clustering

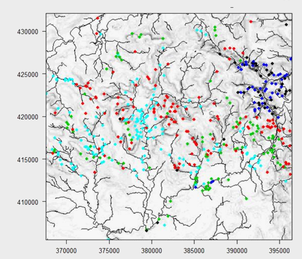

Fig. 6 Results of the cluster analysis using k-means on the verified records. Location was not incorporated into the clustering algorithm but is used here to display the cluster pattern.

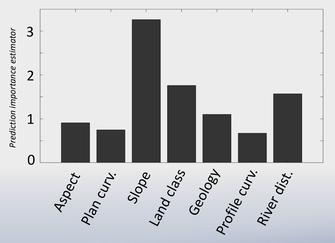

Fig. 5 variable importance within the random forest model implementation.

5. Description and discussion

Within the area of interest, greatest slopes and river channel density occur to the centre and eastern regions. These are areas where most landslide events have been recorded. In terms of land use, most past events are located within largely natural and semi-natural environments. Fig. 1b and 6 highlight the concentration of verified landslide events in these regions. The performance of all susceptibility models corroborate these patterns (Fig. 4) – the relative strength of predictors for the random forest model is presented in Fig. 5, with slope having been found to have the greatest importance.

Of the models (Fig. 4a), random forest provided the greatest degree of confidence but also the largest discrepancy between calibration and validation (see Table 1). This likely relates to model overfitting. To assess this, further analysis is required, especially where this approach is to be up-scaled (section 6).

Model	Calibration - random	Validation - random	Calibration – spatial	Validation – spatial
GLM (slope)	0.67	0.67	0.67	0.65
Stepwise GLM	0.72	0.70	0.72	0.69
Regression tree	0.76	0.70	0.76	0.67
Random forest	0.90	0.71	0.90	0.68

Table 1: Model calibration vs. validation

Comparing the susceptibility models (Fig. 4a) to GeoSure (Fig. 4b) – the current heuristic susceptibility product – subjectively, a similar pattern is exhibited when comparing most likely data-driven susceptibility and greatest heuristic hazard class value susceptibility. Modelled susceptibility exhibits greater variability in lower GeoSure hazard class areas. In addition, these model outputs provide information with regard to confidence unlike the heuristic approach displayed by the GeoSure product.

A number of considerations and limitations must be acknowledged with regard to the modelled values:

- Spatial autocorrelation is not accounted for;
- The random forest approach assumes data are independent;
- The area of interest is relatively stationary – results would differ should this approach be scaled up on a UK national scale.

The cluster analysis, with particular focus on the blue and black clusters in Fig. 6, group together in the areas of greatest modelled and heuristically defined susceptibility. This is again likely a function of the dominance of the slope variable within this area of interest.

6. Summary and outlook

Here we have presented an overview of current analyses which have been undertaken to identify the potential of data-driven approaches to assessing landslide susceptibility based on currently available BGS datasets, including the UK national landslide database. The approach presented here has been tested for only a small area of interest which exhibits a degree of stationarity, especially with regard to geology type.

With the intention of applying this methodology to a greater regional and ultimately national scale, a number of considerations must be made with regard to all of the presented analytical methods including:

- Effects of spatial stationarity on a regional and national level;
- Relative predictability potential of specific (and additional) variables;
- Uncertainty with regard to geomorphometric derivatives relating to their method of calculation and consequential impacts on model reliability.

7. References

[1] BGS GeoSure: http://www.bgs.ac.uk/products/geosure/home.html
[3] Crofts, D., Lancashire landslides: Integrated mapping of potential geological hazards. Earthwise, 20, British Geological Survey, 2004.
[2] BGS National Landslides Database: http://www.bgs.ac.uk/research/engineeringGeology/shallowGeohazardsAndRisks/landslides/nld.html
[4] Kaufman, L. and Rousseeuw, P.J., Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, 1990.

Contact: chrwil@bgs.ac.uk