Geophysical Research Abstracts
Vol. 15, EGU2013-13055, 2013
EGU General Assembly 2013

Sample size matters: Investigating the optimal sample size for a logistic regression debris flow susceptibility model

Tobias Heckmann, Katharina Gegg, and Michael Becht
Physical Geography, Cath. University of Eichstaett-Ingolstadt, Eichstaett, Germany (tobias.heckmann@ku.de)

Statistical approaches to landslide susceptibility modelling on the catchment and regional scale are used very frequently compared to heuristic and physically based approaches. In the present study, we deal with the problem of the optimal sample size for a logistic regression model. More specifically, a stepwise approach has been chosen in order to select those independent variables (from a number of derivatives of a digital elevation model and landcover data) that explain best the spatial distribution of debris flow initiation zones in two neighbouring central alpine catchments in Austria (used mutually for model calculation and validation).

In order to minimise problems arising from spatial autocorrelation, we sample a single raster cell from each debris flow initiation zone within an inventory. In addition, as suggested by previous work using the "rare events logistic regression" approach, we take a sample of the remaining "non-event" raster cells. The recommendations given in the literature on the size of this sample appear to be motivated by practical considerations, e.g. the time and cost of acquiring data for non-event cases, which do not apply to the case of spatial data. In our study, we aim at finding empirically an "optimal" sample size in order to avoid two problems:

First, a sample too large will violate the independent sample assumption as the independent variables are spatially autocorrelated; hence, a variogram analysis leads to a sample size threshold above which the average distance between sampled cells falls below the autocorrelation range of the independent variables.

Second, if the sample is too small, repeated sampling will lead to very different results, i.e. the independent variables and hence the result of a single model calculation will be extremely dependent on the choice of non-event cells. Using a Monte-Carlo analysis with stepwise logistic regression, 1000 models are calculated for a wide range of sample sizes. For each sample size, the "diversity" of the 1000 models is assessed using the Shannon Entropy index which is frequently applied in ecology as a diversity measure. In our study, a high model diversity is inferred where there exists a high number of different sets of independent variables that remain in the model after the stepwise procedure, and if these different combinations are more or less evenly distributed among the 1000 model runs. Conversely, model diversity is low when the number of different independent variable combinations is low, and/or only few combinations make up the lion's share of the 1000 models. Results show that model diversity decreases with increasing sample size to a local minimum before it increases again; this local minimum is reached with sample sizes that are closely below the critical threshold with respect to the autocorrelation range.

Finally, the magnitude and spatial distribution of model uncertainties is explored using an ensemble of 100 models calculated from independent samples (of the "optimal" size).