Nat. Hazards Earth Syst. Sci. Discuss., 1, C1790–C1798, 2013
www.nat-hazards-earth-syst-sci-discuss.net/1/C1790/2013/

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Sample size matters: investigating the effect of sample size on a logistic regression debris flow susceptibility model" *by* T. Heckmann et al.

T. Heckmann et al.

tobias.heckmann@ku.de

Received and published: 18 November 2013

Reply to review by M. Bertrand
We thank Melanie Bertrand for her review. Below, we address the issues raised in the review, and outline changes that we made in our manuscript. On the following pages, the comments are written in boldface, followed by our replies in standard text style.

The multicollinearity between the geofactors has been measured thanks to the VIF indicator, and has been showed to be a little bit high for some of them. However, the authors judiciously chose a backward stepwise procedure to

select explanatory factors, so multicollinearity is a minor issue.

We agree with the reviewer; we were not aware that the stepwise selection procedure mitigates the multicollinearity issue. More on the latter is discussed in our reply to the second review. No changes have been made to the text.

Questions about the diversity of model species: The definition of model species is not so easy to understand at the first reading. I am still not sure to have properly understood.

We acknowledge that the short definition of model species may be hard to understand. In the revised manuscript, we have made an effort for a better description and explanation in section 3.3.1:
"We analyse model diversity by repeating the stepwise model selection with 1000 independent samples of a given sample size (between 50 and 5000)" (...)
"The set of selected geofactors for one sample defines a 'model species' (if, for example, the geofactors A, B and D are selected from the candidate geofactors A,B,...E, the species of the resulting model is ABD). The term 'model species' was used in order to highlight the similarity of the proposed method for model diversity assessment with investigations of biodiversity in ecology. Theoretically, $k_{max} = 2^g - 1$ different model species can exist if g candidate geofactors are available for model selection, and if the resulting model has to contain at least one geofactor. The diversity of the 1000 replicate models calculated for each sample size is evaluated using three measures:" (...)
"H and D combine the number of different model species ("species richness") and their relative frequency (relative "abundance") in one single number: A large diversity associated with a high species richness (k different terms have to be summed up for H and D, respectively) and/or an even distribution of model species across the 1000 samples. Conversely, diversity is low when there is only a small number of different

species, and/or one or few species strongly dominate. Shannon's Entropy has been interpreted in terms of the "average surprise a probability distribution will evoke" (see e.g. Thomas, 1981). The result of a stepwise selection with a sample size for which low diversity (low H) has been measured is not expected to be surprising, because one or few species have a very high probability of occurrence. We hypothesize that the diversity of model species, and the degree of surprise with which we see one particular outcome of the selection given the results of 1000 models, will reflect the sample-dependence of the stepwise selection. Therefore, we propose the "model diversity" as a measure of model quality in terms of reproducibility; similarly, Petschko (2013, NHESSD) recently proposed a "thematic consistency" index that uses variable-selection frequencies in model replications and is based on the Gini impurity index."

Could the authors confirm that only models considered as relevant (according to the AIC) are taken into account for the model species diversity calculation ?

Yes, the backward selection keeps variables based on the change in model AIC (variables are removed or re-included when AIC decreases upon removal/re-inclusion). The selection process is performed on each of 1000 random samples of a given sample size (one pixel for each debris flow initiation zone, plus n pixels from the non-event population), and the diversity of model species is assessed. This procedure is applied for different (non-event) sample sizes between 50 and 5000.
No changes have been made in the revised manuscript; however, the 1000fold sampling and model selection procedure was described in more detail, see previous paragraph.

What about the richness calculation ? The interpretation of the diversity indicators is not very clear. The author should explain a little bit more why the lowest diversity value corresponds to a minimal dependence of model selection on the

sample and its size.

In our approach, we assume that a high sample dependence is reflected in a large diversity of model species. If we get >20 different model species in 1000 runs, we conclude that what is the best model depends highly on the specific sample. Conversely, if the vast majority of 1000 stepwise selections results in the same model species (not necessarily with similar coefficients), we conclude that the model selection is not so much dependent on the specific sample (as many samples result in the same model composition). See our changes to the manuscript reported above (section 3.3.1).

Questions about the reliability / predictive power of the models: How can the authors explain the differences in AUC observed between the model calculated in the LT area and applied to the ZBT and the model calculated in the ZBT and applied to the LT area ? Is it due to the sampling process or to the spatial variability of geofactors between the two areas ? Can the spatial autocorrelation be considered as identical? Is it due to the difference in debris-flow event density between the two areas?

These are important questions that we cannot answer completely. We assume(d) that the two study areas have very similar characteristics, so probably the event density might play a role. In order to find out what controls the different performance (and hence, the model transferability), we would have to apply the modelling and validation procedure to a larger number of different study areas (see e.g. Heckmann and Becht, 2009, Erdkunde). We have added some text to the discussion:

"The different performance of the ZBT model in the LT area and vice versa is an interesting fact. This could be caused by different characteristics of the study areas

(different range, spatial and statistical distribution of geofactor values); the two neighbouring areas, however, are regarded as very similar and homogeneous. Heckmann and Becht (2009) investigated the transferability of a debris flow susceptibility model among different study areas and reported that the predictive power of models are largely independent of the degree of similarity of training and test area; their model approach (certainty factor), however, strongly differs from logistic regression. Besides computational and conceptual differences, continuous geofactors such as slope are classified using the same scheme in all study areas. Conversely, in our study, a different range of geofactors in training and test areas could lead to different coefficients and different model performance due to extrapolation. Another reason for the different performance could be the different debris flow density. In order to determine the controls of model performance, future research will have to use a larger number of different study areas with different debris flow densities. The methodological framework for the assessment of model variability and performance proposed here is considered useful for such investigations."

Similarly, we'd like to point the attention to the paragraph of the results and discussion section that deals with the observed sample dependence:
"(...) The LT is smaller than the ZBT, has a smaller number of debris flows, but a higher debris flow density (events per square kilometer), hence there does not appear any conspicuous relationship of the existence and location of plateaus or local minima, absolute or relative sample size, and the aforementioned study area properties. The investigation of these problems is left open to future research."

Did the authors try to measure the effects of other predictive modeling approaches (not the median) on the susceptibility map built from model ensembles?

We assume that what is meant here is not other modelling approaches but other possibilities for calculating a model ensemble (besides the median). The answer is no, we chose the median as it is common measure of location which has very good statistical properties like robustness and ease of computation. Consequently, we used the IQR as a non-parametric measure of dispersion. We know of other techniques of summarising a model ensembles, for example a weighted average that is recommended by Marmion et al. (2009, Comp. Geosci.)
. If the comment refers to different model approaches, the answer is no. We focused on logistic regression as i) it has performed well in several comparative studies and ii) it has been used very often. Moreover, the way the parameters are estimated (maximum likelihood) strongly relies on stochastic independence of the pixels under study (which is not given in a spatially autocorrelated sample). To overcome this point, we use the sampling procedure described in the paper.

Questions about the geofactors used: About the spatial units, i.e. pixels, on which the geofactors are calculated, did the authors try to measure the effects of radius size on the geofactors extraction. Of course 1m DEMs show micro-topography, but why has the particular 5m resolution been chosen for resampling purpose? Why were roughness and curvature geofactors calculated based on various moving window sizes (5 and 10) ?

We reported the moving window radii in order to make our approach reproducible, and both reviews are right in demanding justification for the specific choices. However, in our study, we have not rigorously checked several possible radii of moving window filters (which would admittedly be an interesting study on its own). The reason why we used a radius of 10 m (not 5) for DEM smoothing (contrary to the the manuscript, we had smoothed the DEM5, not the DEM1, now corrected) before curvature derivation is the (coarse) estimation of the “typical” scale of channels within the rock faces and

talus cones that are both prone to and indicative of debris flow activity. With a larger window, topography is smoothed too much as to indicate locations of concentrated flow; moreover, radii of profile curvature become too large, for example, to represent the boundary of rockface and talus). While we wanted to “smooth out” detail for the curvature calculation, we kept the window smaller (5m) for the roughness calculation, as we were more interested in material properties within smaller surroundings (for example on the sidewall of a gully rather than capturing the medium or large-scale roughness of the gully itself).

Added to/changed in the manuscript:
"Roughness was calculated as the "vector ruggedness measure" (Sappington, 2007) on the DEM1 within a moving window of radius 5 m, and the result was resampled to the same resolution and extent as the DEM5 using the nearest neighbour approach. The comparatively small radius was chosen to capture the roughness of surfaces rather than the roughness induced by landforms, e.g. by gullies."
"**Plan and profile curvatures** were derived with the same algorithm as slope, but from a DEM5 smoothed with a moving window mean filter with a radius of 10m. This was deemed necessary because of the extremely noisy character of fine-scale curvature. Medium-scale curvature based on a DEM that retains details on the typical spatial scale of channels within the rock faces and talus cones (that are both prone to and indicative of debris flow activity) is expected to be a better proxy variable for convergent flow of water (plan curvature) and changes in flow velocity (profile curvature)."

Another geofactors which could have been used to measure the convergence flow of water instead of plan curvature, is the convergence index. This has been used in some hydrological studies and could perhaps be less collinear with other geofactors. Indeed the plan curvature coefficients have been proved very variable depending on the sample and its size. Moreover this geofactor has been

selected in only 20% of the models species, it shows a stronger autocorrelation compared to other geofactors...
That is not correct, the autocorrelation range is the smallest in our set of geofactors (together with profile curvature). **and is also collinear with some of them. Did the authors try to measure the spatial autocorrelation of geofactors with other indicators?**
No, we essentially did the same as Brenning (2005, NHESS) who used the correlogram (of the residuals of the prediction model, however) to assess the autocorrelation range. We know of no other measure of autocorrelation range (which we need in order to compare it to the average distance between neighbouring pixels in the sample)
. We further concede that PLC is not a very useful geofactor because the coefficient is very variable (positive or negative values). This has been stated and discussed already in the original manuscript: "The coefficient for plan curvature has the largest range, and it takes positive and negative values, which makes the interpretation very difficult; this is probably caused by the fact that the random sampling of event cells from the upper erosional zones in the debris flow inventory will select locations in the center of channelised debris flow paths (with highly concave plan curvature), but also at the boundary of these areas, which are highly (plan) convex." This problem would supposedly persist if we chose e.g. a convergence measure instead of PLC.

On the figure 4, is the first geofactor falling within its autocorrelation range, encountered for the smallest sample size, often the same one?

We use the ranges of the variograms in fig. 2 for the shaded areas in fig. 4. There is probably a misunderstanding of the meaning of fig. 4 that estimates the mean distance between neighbouring pixels in a sample as a function of sample size, and compares this mean distance to the (fixed, because estimated for the whole study area) autocorrelation ranges of the (candidate) geofactors. It has no relation to the geofactors

contained in the respective model, and is only used to assess the risk of autocorrelation hazard associated with a given sample size.
The factor with the two largest autocorrelation ranges is slope (slope is autocorrelated on multiple scales), with ranges of ca. 800 m (due to the symmetry of the trough-shaped valley, where large areas of +/- the same slope exist on either side) and ca. 200 m. It is contained in every model (see fig. 3).

p2745, L26 is "DHM5" used instead of "DEM5" ? p2749, L10 "of" instead of "or" p2759, L5 "previous chapter" = "previous section" ?

The corrections are much appreciated.

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., 1, 2731, 2013.