Nat. Hazards Earth Syst. Sci. Discuss., 1, C1325–C1331, 2013
www.nat-hazards-earth-syst-sci-discuss.net/1/C1325/2013/

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.

Anonymous Referee #2

Received and published: 2 October 2013

The paper NHESS_2013-95 by T. Heckmann and co-authors gives a contribution to an important and often neglected problem in landslide susceptibility studies, that is, the influence of sample size on the model tuning and on the quality of final results.

The manuscript is well thought and organized and has a potential interest for the community of landslide hazard analysis scholars. It still has, however, some quite important weaknesses concerning the conceptual approach and basic hypothesis. For this reason, I think the manuscript could be published on NHESS after some revisions which would require a new review round.

The main issues I have concern the following points:

1. The study aims at providing a framework for the assessment of the impact of sample size on landslide susceptibility maps (LSM) using quantitative methods and some new approaches. However, the hypotheses and methods are tested against a very limited target, represented by 2 very small study areas where a very high homogeneity in physical settings can be expected (a proof of such homogeneity could be the fact that using only 81 samples the authors attain an overall ROC AUC = 0.87, see e.g. Page 31 – rows 23-25). How can conclusions drawn from this very specific domain be translated to larger geomorphological contexts? Areas smaller than 10-20 km2 are not the usual target for LSM due to the fact that direct field survey and detail scale analysis (including field tests and deterministic modeling) are still possible with reasonable costs (see e.g. Berti and Simoni, 2010 JGR, 115). Most of the basic assumptions of the paper (such e.g. the sampling distance, the factor autocorrelation and so on) are influenced by this choice. A much larger area of study (e.g. > 100 km2 at least) would be needed to discuss the actual effects of sample size on LSM, an area where the relative importance of susceptibility factors could vary due to the presence of different types of movements, settings and so on. The fact that this study only concerns debris flows does not exempt the authors to consider the practical fact that, actually, LSM application must be performed in most cases without any a-priori knowledge of active processes or at least without any possibility of discriminating different ensembles of environmental settings to perform limited-extent mapping. In this context, sampling for model calibration should also encompass the problem of statistical significance in different non-homogeneous areas with implications that are not negligible for the aim of the paper. 2. All the analysis and results are influenced by the impact of proportion between event and non-event pixels selected for the calibration process. That is well discussed in the manuscript and this is also one of a very few papers that directly raise the issue and propose a solution. However, there is no accounting for "false negatives" in the study. We all know that landslide inventories are highly affected by missed landslides, which have not been recognized due to the absence of visible effects at the

scale of mapping. This loss of information, due to the normal processes of geomorphic and environmental slope evolution after failure, is especially important for shallow landslides (such as the debris flows considered in the study) whose scars have a very low persistence in time. How the presence of such missed positives influences the results of analysis? Usually, one way of reducing the noise introduced by false positives would be to increase sampling dimension but, instead, this study seem to reveal that no gain is obtained by using larger samples. How is this conclusion influenced by the hidden presence of missed old debris scars within the non-event samples? How can this be translated (see also note 1) to larger areas where this problem could even be worse? 3. There is, throughout the paper, a diffuse contradiction: in the methodology part (and somewhere else as well) you report that step-wise model selection (or, I would say, factor selection since you always use one GLM model) is not appropriate for testing theory and to draw conclusions on the physical significance of single factors with respect to debris flows. In several other parts (mainly results and discussion) you nevertheless use this step-wise selection results to make hypotheses on the role of susceptibility factors. Furthermore, basically, I wonder why would you choose to perform all these preliminary analyses on sampling dimension and factor selection to reach an overall performance (AUC=0.8-0.9) similar to those found by many other authors without recurring to it if you are also unable to use the results of factor selection to draw some interesting conclusions on the physical processes actually producing debris flows? Please clarify this point. 4. Another, somewhat minor, conceptual issue: in the introduction it is not clear what are the novel aspects proposed by the paper with respect to recent similar works such as e.g. Brenning (2005) who already attempted sample size sensitivity analysis. Please add some more explanation here. 5. The authors test model stability against random sampling after selecting the optimal sample size. This comes a little bit too late in the paper, in my opinion. Why not try to understand which is the impact of local mapping heterogeneities at different sample sizes? I would think that, the smaller the sample, the larger the impact of outliers on the model calibration should be. Your results seem to contradict this quite common belief. But you only perform random tests

for 1 sample size. For example, how errors in the estimation of local DEM elevation influence the results of your analysis? The origin of the data (LiDAR acquisitions) usually implies high precision but high frequency of local errors, especially on mountain topography. 6. Autocorrelation. The authors rise here a very important issue that concerns almost all LSMs published so far. Their approach of using average variogram range as the limiting inter-distance for sampling (as similarly proposed by Brenning, 2005) is very understandable and should become common practice in environmental analysis. However, I believe that the whole geostatistical issue is in general treated with too much superficiality in the manuscript. In particular, I have two doubts regarding the specific case which I hope the authors will be able to clarify: a. Many of the commonly used variables (factors) influencing landslides are not continuous in space and cannot be treated as random space functions (or fields). Very clear examples are geology and land cover, whose spatial representation (also for mapping constrains) is that of areal objects with properties which are constant within polygons and suddenly vary when a polygon boundary is crossed. Due to the very nature of DEMs, this also applies (even though only partially) to elevation data and their derivatives (slope and so on) who are always measured (or computed) over finite spatial domains (pixels of pixel clusters). This implies the existence of very specific and complex spatial autocorrelation patterns that cannot be captured by simple pixel-based variogram analysis. A corollary of this is that strong spatial asymmetries are present in geological and geomorphological data, which in most of the cases would render a simple representation of average range completely useless. I believe the solution of taking the lowest range for staying on the safe side is not the right solution here. b. The authors discuss autocorrelation and its physical meaning in the landscape as a possible negative influence on factor selection due to lack of sample independence. The problem is, that the authors modify the (already artifact) autocorrelation properties of the factors by arbitrarily averaging some of them to different scales. It is clear, and I agree on that, that slope curvature calculated at the 1 m pixel scale has no relationship with debris flow initiation. However, why use 5 m aggregation and not e.g. 10 m or 2 m? This choice clearly influences the variogram

analysis. I would be grateful if the authors could clarify this point and how they think this influences their analysis and conclusions. 7. In all the manuscript the authors use the diversity and the diversity indexes as measures of model performances. While I agree on the fact that a model should be parsimonious and that over-parametrisation is to be avoided, it is still not clear to me why they automatically assume that low model diversity is perforce better. In some conditions low diversity could also be a warning of possible overfitting problems, especially at large sample sizes. 8. A very big concern: at page 29-30 row 28 to 37 you say that GLM models (and in particular the logistic regression used in the study) generates some errors in the susceptibility maps because "linear modeling approach is not capable of modeling complex non-linear relationships such as the one of slope and debris flow...". Now, slope angle, if I am not mistaken, is one of the most important factors resulting from your study. Why then did you choose to adopt logistic regression in the first place? A methodology more flexible towards non-linearities such as machine learning non-parametric models could have been a better choice. 9. The entire analysis set could have gained in clarity had the authors introduced, among the geofactors, also a dummy (completely random) parameter to act as a benchmark. The latter should have resulted as discarded in all model species so as to ensure that the metrics used to measure diversity and performances were correctly working in the model tests. 10. There are several sentences in the manuscript that are not clear and difficult to understand (and also many small grammar errors and typos). The readability of the manuscript would gain a lot from a careful English spelling and grammar revision.

Other minor issues:

1. (all manuscript) It would be useful for improving readability and comprehension, to present sample size not only in frequency terms (n) but also in relative areal extent. For example, how large is the proportion of the area sampled (over the total) for n=81? And for n=350? 2. Page 3 – row 11-14: the sentence is not clear, please rephrase 3. Page 3 – row 22: please add some more recent literature here 4. Page 5 – row

2: as far as I know, the first important applications of ANN to LSM studies are those by Lee et al (2004, Env.Geol.) and Erminni et al. (2005, Geomorphology) 5. Page 8 – row 11-14: the sentence is not clear, please rephrase 6. Page 11 – row 23-26: a recent example that you may refer to and cite here is probably Catani et al., (2013, NHESS Discussions Online) which has recently been out. The authors perform some tests which are quite similar to those presented in your study. 7. Page 15: you did not discuss the problem of noise introduced in the DEM by the presence of low standing vegetation. You say that the areas are not much vegetated but on the other hand (Tab. 1) you say that almost 20-25% has some patchy, shrub and woodland cover that could potentially introduce errors of an order of magnitude compatible to the DEM local differences. Such errors are known to occur even after LAS data filtering through last-impulse selection. 8. Page 15 – row 26: DHM5 should probably read DEM5 9. Page 16 – rows 10-23: the parameter SCA is quite important in your study, but you say it has been computed by using the multiple-direction algorithm proposed by Freeman (1991) which is notoriously known to work only for convex shaped areas. The same algorithm, in concave areas (such as channels and hollows for debris flow initiation) always underestimate flow accumulation because it does not account for convergence prevalence. A different algorithm should thus be used for concave areas. 10. Page 18 – row 28 and Page 19 - rows 1-3: I personally do not agree with this sentence which is in my opinion very strong and contradicts most of the published literature on LSM. I would advise the authors to use a different argument or to sustain the present one with strong proofs. 11. Page 22 – rows 8-9: this is not necessarily so. Points separated by distances shorter than the range can still have very different values and be not autocorrelated 12. Page 24 – rows 7-8: after devoting a lot of space to the problems of factor collinearity it is not contradictory that you propose the combined SCA*Slope factor? 13. Page 25 – row 2 and also Page 24 – rows 10-23: the combined parameter SCA*Slope is better known as distributed (or discrete) stream power index and has been early used and proposed by Bagnold (1966) for rivers and then by Moore at al (1991) and Seidl and Dietrich, (1992) for slopes. Please add some refs 14. Page 32 –

row 21: not clear, spelling error? 15. Figure 3 : could this figure be compressed into one single plot to improve readability in comparing different geofactors? 16. There are many other small issues connected to the main problems raised so far that do not need to be corrected at this stage

In summary, the paper is very promising and very much needed. I strongly encourage the authors in improving the manuscript along the suggested lines to achieve publication in NHESS.

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