Nat. Hazards Earth Syst. Sci. Discuss., 1, C319–C326, 2013
www.nat-hazards-earth-syst-sci-discuss.net/1/C319/2013/

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Assessing the quality of landslide susceptibility maps – case study Lower Austria" *by* H. Petschko et al.

Anonymous Referee #1

Received and published: 12 June 2013

1 General Comments

This paper investigates several questions concerning statistical landslide susceptibility models, specifically i) model set-up in larger, heterogeneous study areas, ii) spatial vs. random sampling and model validation (cross-validation), and iii) model quality (e.g. variability in model selection, transferability and consistency, magnitude and spatial distribution of model uncertainty) in relation to, among others, sampling strategy and sample size. The method used (GAM) is relatively new to landslide susceptibility modelling, and has proven to be better suitable to model the frequently non-linear relationship of influencing factors and landsliding than existing approaches such as GLM. The research questions apply not only to GAM; they are very important for the appli-

cation of statistical methods to landsdlide susceptibility models - in previous studies, issues such as sample size, model transferability, consistency, and strategies of parameter estimation vs. model validations have often not been dealt with at all, and if so, in a less comprehensive and/or less rigorous manner.

The largest part of the manuscript is very well written, and the paper is clearly structured. The conclusions drawn from the results are comprehensible and well supported by the findings. I consider the paper well suited for publication in NHESS if a number of minor to moderate revisions are dealt with. For me, the most important issues arise in the methods section; some of the points may be due to incomplete understanding on my side, but should probably be explained more in detail in some cases in order i) to make the respective approach better reproducible and ii) to better justify the suggested indices and/or approaches.

2 Specific Comments

p1004 l4ff: In this paragraph the authors motivate their strategy of k-fold cross-validation. I feel that, in the introduction chapter, the generic strategy of 1-fold cross-validation (or "single hold-out") should be explained more generally, for the less experienced reader. In the methods section, the approach can (and must) be explained more in detail. Specifically, I suggest to briefly explain the term "single hold-out" as model estimation-validation-strategy in one short sentence instead of simply mentioning "single-holdout model performance measures".

p1004 l9: The abbreviation AUROC is used here without prior explanation what AUROC means and, even more importantly, what AUROC IS (the area under a receiver operating curve, and as such one possible concept to measure of model quality). As model validation is explained in the methods section, I suggest that the authors speak, less specifically, of "a range of possible validation outcomes instead of only one sin-

gle, "random" outcome...". Alternatively, the validation concept of ROC and the quality measure of AUROC would have to be explained prior to using the abbreviation.

p1010 l12f: a comment: The question is also to what extent missing data (from blurred, reworked, removed landslides) influences the result of a susceptibility model. If the latter is good, the locations of previously existing but now invisible landslide should be "predicted" as susceptible. This cannot be checked, but: cross-validation (which reserves part of the inventory for evaluation purposes) compared to model goodness-of-fit (i.e. the model estimated from and applied to the same area/sample) should give an idea of how an incomplete inventory affects model quality.

p1010 l25ff: Please add some explanation on the possible physical meaning of the DTM-derived variables. For some, it is more obvious (slope) than for others (catchment height - climatic proxy for precipitation ? slope aspect - orientation relative to bedding ?)

p1012 l25ff: Yes, tectonic faults are known to influence landslide (in a general sense) activity. In the present study, however, there is a focus on earth (and debris) slides (p1008 l17), which means a granulometry of >80 (earth) or <80 (debris) percent sand and finer. I wonder to what degree tectonics influence this type of process compared to, for example, lithology, degree of weathering, existence of cover beds, climate etc

p1014 l13ff: "landslide points" are mentioned here and in some other paragraphs, while throughout most of the papers, landslide cells are adressed. Please homogenise terminology. Generally speaking, "points" and "cells" are spatial units to which the sampling and the modelling are applied, and in "reality", you use raster cells, not point objects.

p1014 l24: Pls define sampling rate (as the number of sampled cells per unit area) as opposed to sample size. Could you also comment on the justification of the 1:1 ratio of

landslide to non-landslide cells ? I suppose it has to do with the binary target variable and the cut-off (of 0.5) to distinguish predicted events from non-events.

p1014 l26ff: Why is it necessary to adjust the predictions ? The model predicts, in each domain, the probability [0,1] of landslide occurrence, no matter what the sample size or sample rates are. Could you explain this further ? Furthermore, I could not comprehend the definitions of tau0 and tau1. tau0 is introduced as a "sampling rate for non-landslide points" and defined as the ratio of (land)slide to non-(land)slide points - but this ratio was described earlier as being unity. I understand "sampling rate" as the ratio of sampled cells and the total number of cells in the domain... Your "sampling rate tau1" is defined "for landslide points" and set to 1. This is confusing, because this seems to address the ratio of slide to non-slide cells that was explained earlier; your equation (4) simply becomes tau0*odds(x) because tau1 equals 1... As I understood it, the sampling rates of landslide and non-landslide "points" should be the same in each domain, because of the 1:1 ratio of sampled landlide and non-landslide points. Perhaps tau0 should be the sampling rate (ratio of slide and non-slide pixels to all pixels) in the study area, and tau1 this ratio in the domain ? Or is tau0 the ratio of slide to non-slide pixels in the study area, and tau1 is the ratio of slide to non-slide points in the sample (so tau1 = 1:1 = 1) ? To me, an "adjustment" only makes sense if a property of a domain (e.g. the ratio of sample size and total size, the ratio of landslide pixels to the total area, or the number of landslide pixels) is normalised with the same property of the complete study area. Please explain and clarify.

p1016 l16: It is easily understood that the random (non-spatial) partition gives different results for every replication, and the strategy of taking the median and IQR of n partitions is feasible. Two questions arise for me: (1) did you check if the median and IQR of the performance measure are already reasonably stable with 20 replications, or has the number of 20 been chosen arbitrarily ? (2) could you add a sentence explaining why/how the k-nearest-neighbour clustering of the "point" coordinates in the spatial partition approach leads to different results for every replication ? Are the k group cen-

troids chosen randomly ? Are the resulting spatial units similar in size or is it possible that models are estimated and validated in two partitions of very different size ?

p1017 l1ff: Why is the IQR an estimated one ? It is an empirical measure derived from 20 replications of a cross-validation procedure, i.e. of 100 empirical AUROC values (see p1016 l18f). Furthermore: Why does that measure have to be adjusted in order to be a measure of transferability ? Perhaps you need to explain this more thoroughly, in a more step-by-step fashion. Moreover, "Eq (1)" refers to the corresponding equation in the paper of Hanley and McNeil, not to Eq(1) in your study... I feel that this equation should be given here, or that "Eq(1)" should be removed. Concerning equation (5) for the calculation of the T index: Not knowing "Eq 1 of Hanley and McNeil", I suspect that the SE of the AUROC is estimated from the standard deviation of AUROCs, and will decrease with larger n. I do not see why T should be a better indicator of transferability than the empirical IQR, for example. A larger IQR means that a model could be very good, but also really bad, while a smaller IQR indicates that the models predict similarly well (or poorly). I feel that T should be better justified and explained. I understand from the paragraph that you "correct" a non-parametric empirical measure of AUROC variability (IQR) by subtracting a parametric, estimated measure of variability (that is multiplied by 1.35 to supposedly have the same value as IQR under the assumption of normality...). But why ? It may be correct, but it needs more explanation.

p1017 l17ff: assessing the thematic consistency with an index is a good idea.

p1019 l24: The classified susceptibility map is mentioned here, but the classification rules are introduced in the results chapter (p1020 l7ff). This should be done before/at the first instance when the classified map is introduced (here, or somewhere else in the methods section).

p1021 l10f: Considering that AUROC for the spatial (more meaningful) partition is 0.53 (very close to useless), the contrast to AUROC=0.79 (acceptable) for random partition

is very good evidence for the consequences of using over-optimistic validation strategies !

p1021 l13: You mention a (1st quantile) AUROC value of 0.35 - but the AUROC only takes values between 0.5 and 1 (see also p1016 l23) !

p1022 l12ff: "thus the transferability" - does "thus" also apply for n between 200 and 400 as in line 10f ? Moreover, can you recommend from your findings a minimum sample size ? If so, is it related to i) absolute sample size, or ii) to the corresponding sampling rate ? This question can possibly be answered with 16 large domains with different landslide densities...

p1022 l20f: "one specific random sample and variable selection repetition" - you should delete "repetition" (because the results are based on one sample and subsequent variable selection).

p1022 l27: the possible physical meaning of the variable "catchment height" is not explained, neither here nor in the section introducing the variables. See also my comment on p1010 l25ff)

p1023 l8: why "on average" ? You have x model runs, and p percent of them included the variable.

p1024 l20ff: the propagation of uncertainty to susceptibility classes is a very good idea in order not to over-interpret uncertainty while at the same time giving end-users the chance of having a closer look where uncertainty crosses the boundary of one or even two classes.

p1028 l23ff: what does "adverse effects" mean ? Does that mean that the performance measure is biased ? or wrong ? or that the performance could be better than estimated ?

p1028 l28: I slightly doubt that serious over- or underrepresentation is possible with large samples in the order of hundreds to tens of thousands of pixels. Perhaps in very inhomogeneous study areas - but that is being dealt with in your approach by establishing domains (at least with respect to lithology).

p1029 l27f: Does an underestimation not occur, for example, in the medium class as well ?

p1030 l3ff, especially l9ff: I feel that the comparison of your susceptibility map with the hazard zonation plans is not fully feasible. Risks are induced by mass movements not exclusively where they initiate, but also where they stop (in case of mass movements with a considerable runout). For some types of movement, the hazard zonation map needs to assess the runout zone as well. This might or might not be the case for earth and debris slides that represent the main focus of this paper.]

p1047 Fig. 4: Why does the AUROC for domain 230 scale on a 0-to-1 axis, while the AUROC has a range of [0.5,1]?? This is one of two inconsistent uses of AUROC range (see comment p1021 l13). Moreover, the legend for each boxplot should be changed: either use "spCV and nspCV" or (shorter and probably better) "sp and nsp".

p1048 Fig. 5b: Something is wrong with the y axis labels. Either it should be the numbers from 0.00 to 0.10, or from 0.00 to 1.00.

3 Technical Corrections

- p1013 l13ff: I suggest to split this sentence: "Among the currently available methods for landslide susceptibility modelling a GAM shows a compromise between the flexibility of machine learning algorithms and the smooth representation which results from GLMs such as logistic regression; meanwhile, it still gives the oppor-

tunity of a transparent and easy interpretable model (Brenning, 2008; Goetz et al., 2011). Alternatively: ...such as logistic regression while still giving the opportunity..."

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., 1, 1001, 2013.