Nat. Hazards Earth Syst. Sci. Discuss., 1, C486–C493, 2013
www.nat-hazards-earth-syst-sci-discuss.net/1/C486/2013/

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Exploring model sensitivity issues across different scales in landslide susceptibility" *by* F. Catani et al.

Anonymous Referee #3

Received and published: 10 July 2013

The manuscript by Catani et al. explores the sensitivity of random forest landslide susceptibility models to several choices in model setup, data preparation and variable selection. This topic is of substantial interest to the landslide modeling community since machine-learning methods are becoming increasingly popular in this field.

The manuscript does not describe the methodology in sufficient detail. Some methods, even the random forest technique itself, are not described well enough. The use of different accuracy measures (misclassification error in the context of variable importance, and area under the ROC curve for general model performance) and estimation methods (out-of-bag estimation for the misclassification error, and an unspecified method for the AUC, probably estimation on the training set) is confusing and seems to contradict the authors' claim that the ROC curve is the best performance measure available.

Better consideration should be given to findings reported in the recent literature related to computational statistics / machine learning and the application of machine-learning methods in landslide modeling and in the broader field of geomorphometry. In particular, random forests and related tree-based ensemble techniques are known to overfit to the training data, and different authors have argued that spatial cross-validation techniques are required in order to obtain bias-reduced estimates of accuracy measures when using highly flexible modeling techniques such as random forests. This limitation clearly affects the results and conclusions.

The terminology used is at least unconventional and in some cases incorrect.

Thus, while the topic would be of interest to readers of NHESS, from my perspective the manuscript in its present form does not adhere to the journal's high scientific standards. Key parts of the manuscript need to be rewritten and additional calculations are required to resolve the methodological issues. I am aware that this assessment is quite contrary to the assessment provided by another reviewer.

A number of detailed comments are listed below; however, due to the large number of issues and the implications of methodological concerns for the validity of results and conclusions, this is not a full and detailed review of the Results and Discussion sections. I very much hope that they will use the comments to improve the study's methodology and write-up in the future.

Detailed comments:

Title: Add "modeling using random forests" at the end

P584L3 – what is "LSM"?

P584L8 – "model tuning choice" – the first two issues mentioned in the previous paragraphs, which are related to the scale/resolution of LCVs and MURs, don't seem to be "model tuning choices" but choices related to the scale/resolution of the data.

P584L11 "family" – unusual word choice

P584L13 – “bayesian predictors” – capitalize Bayesian – I haven’t seen the concept of “Bayesian predictors” in the context of random forests before, and I don’t think this characterization of random forests is adequate, but I’d be happy to accept this terminology if I’m mistaken. Since “predictor” is most commonly used in the sense of “predictor variable” (as opposed to prediction model, for example), “tree predictors” seems to be an unconventional word choice for “(classification) trees”.

P584L13-14 – “permits to relate ... with” – change to “relates ... to”; “contributing factors” is unconventional wording, I suggest to stick to common terminology: “predictor variables” or “predictors”; L15 “data layers” should also read “predictors” – use consistent terminology

P584L16 – “no need to select unimodal training data” – this is not a strong argument, which model (other than maybe linear discriminant analysis) requires training data to be unimodal?; omit “classical and”

P584L19-20 standard deviation and “variety” over what? E.g. moving window (of what size)? “parameter set” refers to “set of predictors” (also known as feature set)?

P584L22 “input parameters” – change to “predictors”, use consistent terminology; “regression model” – this is a classification model

P584L22-24 (and throughout the manuscript): “optimal configuration”; “progressively smaller subsamples”; “best set of parameters” – The stepwise procedure used in this paper (described in more detail later in the manuscript) does not guarantee the optimality of the selected subset of predictors. It appears that the present paper, by adding or removing variables based on their importance ranking, effectively only compared up to p different random forest models, if p predictors are available (one model with one variable, one with two variables, etc.). However, simple combinatorics shows that 2ˆp models can be constructed based on p different predictors. While stepwise selection procedures may still be “reasonable”, pragmatic, computationally feasible approaches, the result should not be referred to as an “optimal configuration” or “best set” of pre-

dictors. Again, use more conventional terminology, e.g. “subsamples” should read “subsets”

P585L20 – “his judgement” – or her?

P585L24 – “spatial positioning inconsistencies” suggests that some of the maps were poorly georeferenced – is this the intended meaning, or should it read “thematic inconsistencies” or “thematic disagreement”

P586L12 – this seems to be in disagreement with Brenning (2005 in NHESS) and Vorpahl et al. (2012 in Ecological Modelling), who found great differences in the performance of different modeling techniques. But even when performance results are similar, my experience shows that the actual prediction maps produced by, e.g., logistic regression and random forests are often quite distinct and hardly “quite equivalent” (L10). This is of course not surprising since random forests have difficulties representing linear relationships, while logistic regression will only model interactions between predictor variables if the analyst decides to include such interaction terms in the model.

P586L22-24 – Saisana et al. (2004 in Environmental Science and Technology) and Brenning (2005 in NHESS) emphasize that false positive and false negative predictions may be associated with different “costs” in practice, which is a serious limitation for the use of the area under the ROC curve. Goetz et al. (2011) assessed model performance in terms of the sensitivity at a fixed (high) specificity to account for the higher “cost” that may be associated with false negative predictions. Brenning (2012, in Eberhardt et al., Proc. ISL/NASL, CRC Press) further elaborates on available performance measures.

P587L20-23 – see however Brenning (2005 in NHESS) who used the closely related bagging method, and Vorpahl et al. (2012 in Ecological Modelling) who reproduced these results in the same area using random forests in their comparison. Both studies show that this kind of technique strongly overfits to the training data, that this overfitting remains undetected when using non-spatial error estimation techniques, and that spatial resampling approaches reveal an ability to generalize from the training data that is

not superior to simpler statistical models such as the generalized linear model.

P588L3 – indicate software version(s)

P588L7 – Strobl et al. (2009) is in a psychology journal – is this a sociological study as stated in the manuscript? “usually” also doesn’t seem to be a good word choice, I would argue that random forests are “rarely” adopted in sociological and psychological studies.

P588L10-11 – “subset of the parameter space through bootstrapping” – two resampling approaches are involved in random forests, (1) bootstrap sampling of cases (observations), and (2) random subsampling of predictors; it seems that the authors are mixing these two distinct resampling steps.

P588L12 “random component” – sampling variability?

P588L18 “mixed use of categorical and numerical variables” – more traditional classification methods such as logistic regression also allow the use of numerical and categorical predictors, the latter have to be coded using indicator variables.

P588L19 – “interrelationships and non-linearities between variables” – rephrase: “non-linearities between variables” doesn’t seem to have a clear meaning; “interrelationships” should probably read “interactions”

P588L21-23 – The problem of errors in variables doesn’t seem to relate to the advantages of random forests mentioned in the previous sentence. I doubt that random forests are per se better at handling errors in variables; outliers and missing data are a different issue, that’s where random forests are more robust than classical statistical methods. – What is a “non-point location”?

P588L24-25 – “statistical weight” – aren’t parametric statistical models such as logistic regression much better at this? Model coefficients in logistic regression have a direct, relatively straightforward interpretation. See also Brenning (2012, in Eberhardt et al.) regarding variable importance and effect size.

P588L26-27 – this bootstrap mechanism hasn’t really been explained yet even though it is a key aspect of random forests. The bootstrap and the variable subsampling mechanism are also mixed up again.

P588L5-9 – the authors previously stated that the ROC curve is the “best quantitative tool to measure LSM quality” – why is the misclassification error now used to assess model performance and variable importance?

P589L24-25 – This requires a reference. However, I would like to argue that this statement is incorrect, as long as a “reasonably large” number of trees is used, i.e. at least hundreds of trees, see comment below. A larger number of trees does not influence mean performance, but it does influence robustness with respect to influential cases and stability of the classifier, i.e. it reduces the prediction variance. The acronym “T#” seems to be redundant, especially “T# value” is not much shorter than “number of trees”. Similarly, LCV# is much harder to read than “number of predictors”.

P589L25-P590L1 – reference required to support this statement

P589L4 explain the concept of out-of-bag estimation to the reader

P589L5-6 – I would like to disagree with this, but I can be convinced of the contrary if a suitable reference is provided that supports this statement.

P589L6-7 – I would argue that a large T# is recommendable regardless of the sample size.

P589L13 – Need to specify what constitutes an “acceptable” OOBE. The lower the better of course, so why stop at an “acceptable” one if there are better ones down the road for higher T#, according to what was stated earlier in this section? What T# values were tried out, and why? Table 1 suggests that T# = 1, 10 and 100 were used; while 100 may be marginally acceptable, T#=1 and 10 are unacceptable choices. Breiman (2001) uses asymptotic arguments to justify the development of the random forest method, and in

his manual (http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf) he suggests that “it does not hurt to put in a large number of trees”, mentioning studies where he used 1000 or even 5000 trees.

P592L1-3 “It is indeed known...” – provide reference.

P592L3 define AUC

P592L6 “calibrated ... samples” – incorrect use of the word “calibrated”?

P592L9 and throughout the paper: AUC estimated on the training sample? On the out-of-bag sample? This information is of critical importance because different estimators have different biases and “external” or spatial cross-validation approaches have been established as being suitable for detecting overfitting in the presence of spatial auto-correlation (see comment below with references). Provide details on the calculation of the AUC. Figure 4 and other figures seem to be using parametric approximations of ROC plots. It is common practice (especially in large-sample situations like this one) to use ROC plots that are non-parametrically estimated from the predicted values and observed class membership, and to calculate the area under these curves.

P595L24-25 – The second derivative of elevation in twodimensional geographical space is not a single numerical value but a 2 x 2 (Hessian) matrix. How is the overall curvature obtained from it? Provide a reference?

P599L10-12 – More details on interpolation required, e.g. leave-one-out cross-validation error. The sample size (N=111) would have been large enough to use a better interpolation method, in particular kriging (or splines), and to characterize the degree of spatial autocorrelation using a semivariogram.

P600L3-4 – since individual trees are all drawn from the same tree population regardless of whether T#=1, 10 or 100, their predictions obviously follow the same distribution and in particular share the same mean value. Any significant test result in the t-test is therefore necessarily a false-positive test result, and the application of t-tests is redun-

dant.

P600L15-16 – reducing the number of predictors is not what is commonly referred to as pruning, this is usually referred to as feature selection or variable selection. Pruning refers to reducing the depth of classification trees (i.e. cutting off branches, as in gardening).

In general, out-of-bag estimates used throughout this study should be expected to be biased (overoptimistic) because nearby or adjacent grid cells in the bootstrap and out-of-bag sample may be spatially autocorrelated. See e.g. approaches used by Brenning (2005 in NHESS) and Vorpahl et al. (2012 in Ecological Modelling) and implemented in open-source software (Brenning, 2012, Proc. IGARSS / R package ‘sperrorest’).

P600L27-28 – This is not necessarily an indicator of “good performance”.

P605L26, P606L1 “dimension” used incorrectly in the sense of sample size

Numerous additional typos and minor grammatical and idiomatic mistakes should be corrected to improve readability. E.g. P584L18 “from of”; P584L19 “constrains”; P584L15 “numeric” vs. L20 “numerical”, to name but a few such issues that occurred within a few lines of text.

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., 1, 583, 2013.