Natural Hazards and Earth System Sciences
Discussions

Nat. Hazards Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/nhess-2021-168-RC1, 2021

Comment on nhess-2021-168

Anonymous Referee #1

Referee comment on "Automated determination of landslide locations after large trigger events: advantages and disadvantages compared to manual mapping" by David G. Milledge et al., Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2021-168-RC1, 2021

The authors propose a method for automatic mapping of earthquake-induced landslides and an objective method to compare the results with pre-existing ‘benchmarks’.

The method is based on the analysis of NDVI time series and a statistical hypothesis test to verify the significance of the changes in the series. The method is tuned and applied to some recent events for which enough long-time series of Landsat images exist.

The paper is quite long and dense of contents with concepts of geomorphology, remote sensing, statistic/theory of probabilities, and other disciplines. I’m to admit that my background does not cover all of these disciplines, still, I hope that my comments will be of some help

The writing is per se quite clear but I’m to say that I had to read the paper several times to pull all the things together. The description of some steps arrives too fragmented, making the reader somehow supposing what will be next or giving the reader the possibility of taking wrong directions/expectations (e.g. I had to wait almost till the end to understand what the ‘trade-off’ of the time series was and how the length was selected, and the connection with natural processes). 3.1 is maybe too generic in the introduction of the algorithm, in how to prepare the time series, what the distributions are, and that a fitting method is used to find the parameters (they sound to me all parts of the method). I suggest trying & anticipate some concepts.

I also think that in the first 4 paragraphs some of the preparatory steps, comments, results, and interpretations (see detailed comments) are 'one way, not necessarily wrong but too biased to demonstrate the outperforming of the automatic method. In the discussion, this is a bit relaxed.

In fact, the basic assumption that manual mapping is accepted as the most accurate method to map landslides is taken in a too broad sense and it is not critically reviewed neither contextualised.

The preference is in most of the paper given to the automatic mapping considering only some performance indices, but it does not take into consideration elements like the purposes of producing landslide inventories (in particular just after an event), the time needed to have long and adequate temporal series of satellite images to stabilize the signal in ALDI (at least one year if the sampling is consistent). The inability of the method to trap correctly small landslides is shown as a very secondary aspect, and the fact that, despite their presumed low qualities, manual inventories were used to tune the model (also the general one) is not remarked (without them ALDI could not be tuned). This is, as correctly stated by the authors, without a real benchmark

Nevertheless, I see some potentialities in the method (when better contextualised, and without unbalanced comparisons) to say that, given some inventories, it is possible to run it to update, or extend, or give homogeneity to the preexisting inventories (after many years of data acquisition), indicating the way of correctly using this type of product (for sure not in an emergency since it takes years to have the post-event time series).

In the detailed comments, I raise some issues related to some methodological steps that should be better explained or clarified.

Last, some of the elements in fig 4 (distributions), 8, and 9 are very difficult to catch.

I recommend for major review, and I strongly suggest the authors for a more adequate and multi-perspective contextualisation (maybe starting from the title, the outperformance is not absolute, but eventually relative to some choices).

°°°

Introduction

115 – 125 address the problem that we don’t currently have an objective comparison of manual mapping with automated classification: what is the meaning of objective here? Unbiased?

2 Case study sites

3 Methods

3.1 ALDI classifier: theory

170 – 175 reduces .. increases..: I understand the sense of the sentence but I don’t think it is formally correct, I suggest something like: reflection in vegetated areas is lower/higher than in bare soil because…

180 – 185 negative: since there are many ways to measure changes in NDVI, I suggest making clear that here the difference is used.

180 – 185 regrow slowly (years..): not completely true for areas in south-est Asia. Maybe slowly enough with respect to the availability of the first images after the event... I think this is quite well explained in the next sentence, I suggest to try and connect the two.

194 systematically: is it the right term?

195 – 205 Pt: a quite obscure part until at least the end of par 3.2, where it is possible to understand how to go from temporal series to distributions. I suggest finding a way to better introduce this concept.

Furthermore, Pt can be high also for agricultural practices, or fires, also the state of the vegetation can be completely different.

210 -215 eq 2d: what is ‘temperature’? What is the spatial res. Of this index?

225 - 226 landslide probability: the probability of the presence of landslides in the images?

ALDI does not seem to me so directly connected to a landslide probability unless we assume that Pt, the probability that two distributions are different can be called landslide probability (but then the H0 should be related to). Maybe in another way: is it enough to say that this is ‘landslide probability’?

231 – 232 due to a large variance between pre-event and post-event NDVI distributions: see my previous comment at 195 – 205. I guess here the authors refer to the boxplot of fig. 1-c difference, if yes, it sounds to me very far from being normal distributed, so I would not use variance unless the distribution is modeled, and the distribution has a variance. If not, what are the distributions?

3.2 ALDI classifier implementation and data pre-processing

250 – 255 trade off: actually it seems to me that the final/correct length of the time series is found according to a best fitting process and not related to a study of the physical processes that characterise the possible evolution of the time series. The only choice a priori is related to the length of the time series used in the fitting. This concept remains a

bit misleading until the calibration, at least for me, because I was waiting for a characterisation of the processes to monitor in the different areas.

A curiosity about the sample size: any limit imposed by Google?

259 for each pixel: in the satellite image?

260 - 265 short enough .. long enough …: (similar, and connected to some previous comments)

Generally: does it mean that some remain empty? If yes, how many? How do you cope with this?

Short enough to capture seasonal changes: please add references that this is true for the test areas. Furthermore, is the irregular initial sampling not influencing the result (see Nyquist)?

What does more robust to outliers mean here? How were outliers identified (what test and on what type of distribution)?

265 Vpost: I'm confused, Vpost is not obtained from the distribution of the differences. Is it?

265 – 266 t-test: assuming that the mean of the differences between the medians should be 0 (H0), it is also assumed that the distribution of the differences is approximately normal, something that I guess should be tested in the different situations, and in a robust way, including years with different rain seasons. I don't know how representative fig 1 is of all cases, but there is a 'constant' shift between the blue and the red curve before the event (mean different from 0). Is it a matter of different rainfall in the different years? Need to atmospheric correct the images?

3.3 Performance testing

Another general comment: I could not find the way the authors cope with the different resolutions/scales of the products in comparison and competition. Furthermore, how good is the co-registration among the data?

275 relative measure of the confidence: do the authors mean statistical confidence here?

290 – 295: I suggest mentioning in which situation this choice of working with TPR and FPR is preferred. Here (https://doi.org/10.3390/rs12030346) for example, the authors make a different choice for rapid mapping.

298 – 299 ALDI must first be thresholded to generate a binary classifier with the same FPR

as the competitor inventory with respect to the check inventory: sorry not sure I understand this important step: Is ALDI calibrated using the competitor inventory to obtain the same FPR, correct?

If yes, it seems to me that the outperforming of the automatic mapping is (or can be) strictly related to this choice: once the FP is under control (usually the biggest problem in automatic landslide detection), then TP outperforms. How about tuning to make TPR(aldi) = TPR(comp)? Would the results be similar? I think the choice is quite ‘smart’ because it tends to mitigate one of the issues in automatic mapping, but it must be highlighted that the choice strongly depends on this a priori choice. Is this what the authors mean when they talk about weakness? If yes, I think this point should be better discussed and highlighted.

3.4 Parameter calibration and uncertainty estimation

320 – 325 working from the most to least sensitive parameter

for each earthquake event and then checking for interaction between parameters: sorry, what does it mean? How was the sensitivity ranking established and how did the authors check for interactions?

Comment just for sake of discussion: since some of the parameters are exponents, a random sampling from a uniform distribution might not be able to trap the ranges in which the parameters landscape has a higher slope.

325 – 330: see my previous comment on the a-priori choice of the time series length. I suggest mentioning these criteria (not the numbers) before, in the methods paragraph. Still, I suggest listing the main ‘other landscape changes that might occur in the different areas and references in which it is said that the temporal window and the frequency sampling are adequate. Maybe fires? Drought?

330 – 340 : sorry but the entire subparagraph is not fully clear to me, maybe just because I'm not an expert in it.

measured in terms of AUC: to obtain the ALDI threshold which gives the best AUC, or the threshold which gives FPR(aldi) = FPR(comp)?

combine: what does combine mean here? Or: how did the authors combine the sets?

Parameter interaction: how do the parameters interact?

Equal weight: is it really necessary? In the end, the purpose is to get what gives the best final results.

3.5 landslide size

General comment: I understand the purpose but I have the feeling that this is not a real comparison because resampling the manual inventories introduces 2 issues: 1) you are not more using the original product, 2) you are losing one of the main pros of the manual inventories related to the capacity of the operator to distinguish between small landslides,

345 – 350 First...: the concept was somehow already introduced earlier.

345 - -350 current practice in landslide mapping: I repeat something already commented earlier. As far as I know, other authors prefer different strategies, according to the use that the map is devoted to. I suggest adding a reference to say that this is current practice.

4 Results

4.1 Spatial agreement: Gorkha case study

General comment of Fig. 2: It is not very easy to see the differences (actually I can't), maybe because the area is too large also in the windows c and d. or the quality of the figure I have in the pdf is too low. It would be nice to see some examples of mapped landslides in some VHR optical images (like 2e but without ALDI and more zoomed).

Another point, I see some parallel lines in the ALDI results (not only in fig.2), a sort of high-frequency noise, what is it? Striping problems? Or aliasing?

370 – 375 a number of false positives in the south and west of the study area: 2 comments: I think they become false positive once the threshold is selected, 2) actually with AUC the threshold is selected to make FPR(aldi) = FPR(comp), so, most of them if not all should disappear. Sorry, I don't understand the sentence.

4.2 ALDI calibration: Gorkha case study

405 – 410 This may be because longer stacks are more likely to include other landscape changes after the earthquake that disrupt the signal, such as post-seismic landslides or re-vegetation of co-seismic landslides: I guess this can be verified by looking at the series of year 2, 3, … and repeat the test of the differences year by year.

415 – 445: I'm sorry, I have some (maybe naive) doubts/questions related to this sub par, the topic is a bit out of my expertise, and it was also difficult to formulate my questions.

First, I think it would be useful to see also beta and lambda distributions to evaluate their contribution to the ratio with alpha but also their variability.

In general, to be honest, I'm not sure I fully understand the meaning of the ratio, or the meaning of 'controls' or 'less weight' (see 429): according to the shape of eq 4, the exponents should actually work all together to make numerically all the terms 'right' (to match the measure of goodness of fit with the manual inventory). So in the end what I don't understand is why if optimum performance always involves alpha / lambda < 1,

suggests that: 1) NDVI difference should be given less weight than the more complete t-test derived probability; and 2) the additional information on pixel variability provided in the t-test does adds considerable value to ALDI for this site, and so on. Maybe my mathematical lack but according to the adopted fitting process and the type of equation (no 'interaction betw. variables e.g. dV/V(post), no physical process), I don't see the connection.

If I understand well, the a-priori distribution is a uniform distribution for the parameters of eq 4. Apparently alpha does not change very much with the different measures of goodness of fit, which might be a symptom that more runs are needed or that another choice would be more appropriate (see also the histogram).

Why in col b in the first two rows there are empty spaces (I think beta distribution may help here too)?

Last: when dV rather than Pt is used: Sorry I don’t understand.

4.3 ALDI calibration: global comparison

445 – 450 sensitive: are the non-sensitive parameters useless?

450 – 455: L(post): see my previous comments. Probably there is a hidden connection between the different lengths of the time series and the processes that can cause changes in the time series in the different areas, but these connections are here not studied. I recommend reviewing this point (everywhere in the paper). I am also persuaded that these results can be influenced by the frequency sampling, and by the local characteristics, so I’m not so sure that they can be generalised (or they need further verification).

460 – 465 faster and simpler: faster because the inventory can be produced after 2 years instead of 5 (but potentially limiting the quality), correct?

485 out-performs it: but the competitor is now a part of the procedure to obtain the best ALDI, without a competitor, no ALDI. Does the confrontation make still sense? It sounds to me more like ALDI is a tool to improve the quality of existing inventories, or reduce the differences between inventories.

4.4 spatial agreement: global comparison to manual mapping

General comment, see the previous comment

4.5 Size distributions

550 – 555: the similarity: after resampling. So it is no more the original product.

5 Discussion

5.1 The problem of testing against check data of only comparable quality

560 – 565: Sorry, I don’t understand the meaning. Furthermore, what about the other 5?

5.2 Performance differences in manual mapping reflect inventory errors, not solely mapping errors.

580 severe warping: actually warping occurs when non-linear transformations and non-representative GCPs are used to orthorectify images, maybe better to say ‘distortions or deformations.

585 1 km: ok, but there is a clear mistake somewhere, co-registration / orthorectification processes are not so bad.

5.3 Both agreement between manual inventories and ALDI performance differ depending on the property of interest…

595 – 600: a more appropriate product: sorry but I keep thinking that this is not appropriate to say because, unless a fixed set of parameters is used (still, the manual inventories are somehow part of the process) in future earthquakes, ALDI needs the manual inventory in the flow chart. So, the competitor is used to get closer to the reference. Furthermore, hazard mapping needs an accurate definition of the size, so I don’t think ALDI can be considered as the best solution in this case.

5.4 Limitations to ALDI performance

General comments/curiosities:

Does still make sense to keep Haiti in the general model?

Back to some of my previous comments, it seems to me that the frequency of acquisition can be relevant.

Stripe problems: please, do see one of my previous comments (I have the feeling that I see parallel lines in many ALDI products.

5.5 Application of ALDI to future earthquakes

6 Conclusions

to be further verified