Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/nhess-2021-215-RC2, 2021

Comment on nhess-2021-215

R. S. Purves (Referee)

Referee comment on "Methodological and conceptual challenges in rare and severe event forecast verification" by Philip A. Ebert and Peter Milne, Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2021-215-RC2, 2021

This is a well written and interesting manuscript which discusses the challenges of verifying rare event forecasts. It does so from a particular perspective, that of a dichotomous categorical forecast. The paper argues convincingly as to why the Peirce Skill Score (known under a variety of names including the (Hanssen and)Kuipers skill score or the true skill score) is the best skill score for rare events since it considers three aspects: being better than chance, the direction of fit and the weighting of errors. It also advances an argument for using accuracy as a tie-breaker, which seems to me to make sense. The paper has been written in the context of avalanche forecasting, and uses a paper of which I was a co-author as an example (which I assume is why I was asked for a review). However, it is published in a journal on Natural Hazards and has more general relevance. I believe this is important, because many different problems wrestle with the challenge of calculating meaningful skill scores, and there is a great deal of work which could benefit from thinking about these issues.

The paper is written in an unusual style for the natural sciences, and is in many ways a commentary. Although this makes it very long, it is readable and engaging. In places the authors could perhaps try to think more about their tone - I often had the feeling that statements such as "is called the Kuipers Skill Score by..." implied one particular truth. In practice, many of the measures used have multiple names and we did our best to use those we found in the literature. It would be helpful to add these to Table D1.

Our work (Heirli et al. 2004) actually aimed to do two things: firstly to (re)introduce some different measures to the evaluation of nearest-neighbours avalanche forecasting and, secondly, to consider different ways of evaluating forecast quality (e.g. categorically, probabilistically and descriptively or qualitively). Based on these ideas we published a second paper at ISSW (Purves & Heirli, 2006) where we looked in more detail at the last aspect. The point of this, and the relevance to my review, was that we believed that simple measures were not terribly effective ways of evaluating forecast value a la Murphy. It would have been nice to see the authors recognise this a little more in using our work as an example, since what we were trying to do was rather similar - draw attention to the

importance of thinking about the use of different measures and their appropriateness.

By casting their net very wide the authors run some risks, in particular that they restate things known in other areas. The real strength of this work is in its (exhaustive) analysis, which allows the authors to make concrete and useful recommendations, which I think are very valuable. However, I also feel there is a need for spending a little more time looking at previous work more broadly, for example Manzato(2007) also discusses in detail properties of the Peirce skill score. Furthermore, the more general problem of imbalanced classification (of which Rare Event Classification is an example) is very well known in modern machine learning, to the point of being the subject of text books (Brownlee, 2020). I'm not very sure (sorry) what the best solution to this challenge is. But I think it would strengthen the paper to make links to more literature, and not make the mistake of treating a generic problem (coming up with meaningful measures for classification where data are imbalanced) as such a specific one that relevant previous work is completely missed. One (possibly useful) way of making the paper more relevant would be to include a table summarising measures used in recent papers in NHESS which explore this problem. I think this might also help to make the recommendations of the paper more widely read, and thus increase the contribution and relevance of the paper.

Purves, Ross S., and Joachim Heierli. "Evaluating nearest neighbours in avalanche forecasting-a qualitative approach to assessing information content?." ISSW, 2006.

Manzato, A. (2007). A note on the maximum Peirce skill score. Weather and Forecasting, 22(5), 1148-1154.

Brownlee, J. (2020). Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning. Machine Learning Mastery.