Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2021-218-RC2, 2021

Comment on gmd-2021-218

Anonymous Referee #2

Referee comment on "Nested leave-two-out cross-validation for the optimal crop yield model selection" by Thi Lan Anh Dinh and Filipe Aires, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-218-RC2, 2021

General comments

This paper addresses the very important question of model complexity. How can one best choose the level of model complexity when the objective is to minimize error of prediction for out of sample cases. The paper presents two practical cases (prediction of coffee yield in Viet Nam, prediction of maize yield in France) and two prediction methods (linear regression, artificial neural network). The main objectives of the paper are to explain the leave two out cross validation approach (LTO) as a method of evaluating and choosing between models of different levels of complexity, and to compare it with the leave one out cross validation approach (LOO). The presentation of LTO is useful and interesting, but I have some issues with the way the linear regression examples are formulated. (I do not comment on the use of neural networks, with which I am not very familiar).

Specific comments

The authors consider models with a fixed number of explanatory variables and take as the measure of model complexity, the number of potential explanatory variables from which the explanatory variables in the model are chosen. This is not usually the way the problem is formulated, at least in linear regression. In general, the list of potential explanatory variables is fixed, and the question is how many and which to include in the final model. Then the larger the number of explanatory variables chosen, the more complex the model. Comparing LTO and LOO as a function of the number of potential explanatory variables

may then not be very relevant to the problem of determining model complexity for a linear model. A more relevant question would be: How do LTO and LOO compare, when the number of potential explanatory variables is fixed, and the objective is to choose the best ones to include in the regression model?

In addition, though LOO can be used for model selection (choosing best explanatory variables) in linear regression, there are other approaches which are probably more common, such as forward regression, stepwise regression, the Akaike Information Criterion etc. How does LTO compare with other methods of model selection?

Technical corrections

L138 Problem with English

L145 Should be "to choose"

L163-164 I don't understand the sentence

Section 3, introduction. It seems to me that much of this is said elsewhere. Maybe combine with section 2.3?

Fig 3. I find the portrayal of distribution functions confusing. First of all, in the figure the distribution functions are continuous, while in practice they are necessarily discrete. Secondly, as far as I can tell, the distribution functions are totally irrelevant. Only the average error is of interest.

Section 3.2.2. Talking of a "testing dataset" is a bit confusing (is this the full set of testing data, which is identical to the available data, or is this a single value for each fold). Perhaps refer to the "testing value" or "testing datum" when talking about a single fold.

L266 "implemented" is probably the wrong word.

L362 "request" is probably the wrong word

L366 I don’t understand this sentence.