Clim. Past Discuss., 11, C3172–C3175, 2016
www.clim-past-discuss.net/11/C3172/2016/

Climate of the Past Discussions

Interactive comment on "Technical Note: Estimating unbiased transfer-function performances in spatially structured environments" *by* M. Trachsel and R. J. Telford

Anonymous Referee #2

Received and published: 15 February 2016

In this paper, the authors discuss the implementation of a type of cross-validation (h-block CV) used in palaeo-environmental reconstructions. H-block CV was first proposed as a CV method robust to spatial autocorrelation by Telford and Birks in 2009 (QSR 28:1309-1316). This was followed by a rather heated debate about the extent to which spatial autocorrelation compromises the evaluation of transfer functions (and by extension the need for h-block CV), e.g. in Guiot and Vernal (QSR 30:1965-1972) and in a reply by Telford and Birks (QSR 30:3210-3213). However, for the purposes of assessing this paper, I think it is fair to set this debate aside, and consider this paper for what it is, i.e., a technical note on the best implementation of the h-block CV.

Already in their original proposal of h-block CV, Telford and Birks noted a critical methodological issue, namely the selection of a correct value for h. If H is too small, effects of spatial autocorrelation are not fully removed and the CV performance is likely to be overly optimistic, while if H is too large, CV performance will be hurt by loss of data. Here the authors assess three possible approaches to estimate correct h. One of these, based on using an independent test set with similar distributions of environmental variables as in the calibration set used, is unlikely to be often used, as such comparable independent test sets are rarely available. However, using a particular case where such a test set is available (Atlantic foraminifera), the authors were able to use this method of estimating h to validate the h estimates produced by the other two methods, and this is very useful for the purposes of this paper.

Overall, I consider this a highly useful contribution, and I recommend publication with some minor revisions.

I especially appreciate the methodological approach which combines simulated biological datasets with real-world datasets. Simulated datasets can be very useful in analysing the mechanics and behaviour of reconstruction methods, as shown e.g. in other recent work by the authors (Juggins et al. (2015) Holocene 25:130-136). Here the results obtained with simulated datasets are then validated against real-world pollen and foraminifera datasets, which lends credence to the results from the simulated data.

The manuscript has a decidedly narrow scope (implementation of the h-block CV), however to me this specific topic is an extremely important one. Palaeo-reconstructions prepared from microfossil data are widely applied, from environmental monitoring to palaeoclimate science, and thus the robustness of these reconstructions is critical. Yet, the validation of these reconstructions has remained a persistent problem for decades. The prepared palaeo-curves themselves are hard to validate as the real past variability is inherently poorly understood. Thus, the possibility of running cross-validations which can provide a (relatively) unbiased estimate of predictive ability is extremely appealing, and it is crucial that this is explored, as is nicely done in this work.

From the technical side of things this is an extremely well-written and carefully prepared manuscript. The structure is good and the language is clear, concise and effective.

My one, overarching concern about this paper is how well it will reach its intended audience, despite its considerable merits. While the authors well illustrate the problems of spatial autocorrelation, and suggest concrete steps to take, the paper is also severely technical in its presentation. This is perhaps problematic considering that the core training of the target audience tends to lie somewhere around micro-palaeontology. One may observe that while the problems of spatial autocorrelation were highlighted by Telford and Birks already in 2005, and h-block CV suggested as a solution in 2009, actual applications of h-block CV have been few in later literature. I suspect this is in part due to technical obstacles in implementing the h-block CV, which e.g. requires some R coding which in my experience remains a major challenge for many palaeoecologists.

In Acknowledgements, the authors note that the code for implementing these methods will become available in a separate R vignette. This to me is an absolutely critical complement to this paper. In view of my previous concerns, I encourage the authors to make that R vignette as clear and easy-to-apply as possible. I suspect this would to some extent help make these methods more widely used in the future.

Finally, I have some minor comments relating to specific parts of the manuscript:

Page 4730, Line 3: "spatially-structured" –> "spatially structured"

Page 4732, bullet item iii. Would it possible to add one or two sentences to clarify the underlying reasoning of the variance-explained test? From the authors' description, I can readily follow what is done in this test, but I struggle to understand why this is a useful way to estimate the correct h.

Page 4735, Line 6: Regarding loess, the authors say that "shorter spans are expected to remove more local variance." Is this backwards? I'd expect a short span to remove

less of the local variance.

Figure 1: Question for authors. It seems to me based on this figure that for many datasets the variogram-distance method often considerably overestimates the optimal h compared to the other two methods. The authors suggest using both the variogram-distance and variance-explained methods, and choosing the smaller h. But in addition to that, would it be possible to roughly estimate by how much the variance-explained tends to overestimate optimal h? I ask because of the two methods suggested, the variogram-distance method is by far the easier one to run. The variance-explained method, in particular, seems like it might be very calculation-intensive. It can be run for MAT, probably the least calculation-intensive reconstruction method of all, but what if the test needs to be done for another reconstruction approach? Thus it might be helpful, in some cases, to be able to run the variogram-distance method only, and have some rule of thumb about how much the h is likely to be overestimated. Or does the relationship between the h suggested by the two methods vary too much between individual datasets to give any such guideline?

Figures 2-3: There appear to be two sets of results for the sum of variogram ranges of 30. Is one set of results perhaps for another x value?

Figure 5: I don't see this figure referenced anywhere in the text.

Interactive comment on Clim. Past Discuss., 11, 4729, 2015.