Review of "Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data"

Summary

This manuscript compares six approaches for building predictive models/algorithms using a single spatial binary data set. In addition, the authors illustrate some important considerations when selecting tuning parameters and evaluating predictive performance, which includes block cross-validation and a similar technique to partition the data for tuning parameter selection.

General comments

I congratulate the authors on their work. Much of the literature that focuses on the analysis of "spatial data" doesn't consider machine learning methods like random forests or boosted regression trees and tends to almost exclusively focus on generalized linear mixed models (e.g., Gotway and Stroup 1997; Diggle et al. 1998) or generalized additive models (Wood and Augustin 2002; Kammann and Wand 2003). I think the author's work is needed and is a step forward in providing much need practical guidance. Overall, however, I feel the study is poorly executed, limited in scope, and sometimes missing relevant literature. I have several specific comments listed below. Also, I have a few comments that I consider major in the list below.

1. By using only a single data set, it is not clear how generalizable the results are. It seems that more than one data example should be used. I suggest framing the comparison using a common task framework (*sensu* Wikle et al. 2017). Heaton et al. (2017) provides an excellent example applying the common task framework to evaluate methods for spatial prediction.

2. I think there needs to be more attention fouused on how predictive performance is measured. For example, the authors use area under the receiver operating characteristics curve (AUROC). While AUROC is a common way to measure the predictive performance in a binary setting, it is well known that this measure (i.e., scoring rule) has some problems (most notability AUROC is not a proper scoring rule; Byrne et al. 2016). I would suggest the authors take a look at Gneiting and Raftery (2007) and give more consideration on how predictive performance is measured.

3. I suspect that the data used in this paper (e.g., Figure 1) are spatio-temporal (i.e., not all collected at the same time). Although spatial prediction is widely used in ecology, it seems the trend is towards making spatio-temporal predictions using statistical models and machine learning algorithms (e.g., Hefley et al. 2017b).

4. Many of the ideas presented in this paper (e.g., block cross-validation) do not reference the primary literature on the topic (e.g., citation of Brenning et al. 2005). Given how inclined the authors are to cite there own work, it seems that they should make a better effort to cite the original sources.

5. Overall the writing and structure could be improved. Some of the text seems out of place and multiple terms appear to be used to describe the same thing. I have tried to point these out in the specific comments below. It might help to have a table with a glossary of terms.

Specific comments

1. pg. 1 line 30: Although your data example is a classification problems. Machine learning methods can be applied to other types of ecological data (e.g., count data, proportions).

2. pg. 1 line 36: It is not clear to me what "unbiased performance estimation" actually is. I think what you're talking about relates to the properties of "local and proper" scoring rules (e.g., Gneiting and Raftery 2007). If this is the case the AUROC won't have these properties. If you mean something else, please explain.

3. pg. 2 line 14: "The effect of hyperparameter tuning saturates at around 50 iterations for this data." I really have no clue what this means.

4. pg. 3 line 26: "Consequently, the early detection and subsequent containment of fungal diseases is of great importance. Statistical and machine-learning models play an important role in this process." The data example you give is about mapping and not early detection. There are statistical methods for early detection of disease, but what you present is mapping.

5. pg. 3 line 47: I think it is fair to say that no one would recommend GLMs for spatial data. It is more common to use generalized linear mixed models (Gotway and Stroup 1997; Diggle et al. 1998). See Hefley et al. (2017a) for a recent review of modern approaches.

6. pg. 3 line 49: "These have gained popularity due to their ability to handle high-dimensional and highly correlated data and the lack of explicit model assumptions." The statement that machine learning methods lack assumptions is not true. It would be more accurate to say that the assumptions don't matter if you don't care about any formal statistical properties (e.g., unbiased predictions). Many machine learning algorithms can be re-cast in a statistical framework to understand the implied assumptions (e.g., Friedman et al. 2000). Well-known texts on machine learning cover this basic concept (e.g., Friedman et al. 2008).

7. pg. 6 line 33: Can you give more details on how this binary covariate was predicted. It seems that you could only predict the probability of hail damage. To get a '1' or a '0' you would have to pick an arbitrary cut off unless you used Bayesian imputation via the posterior predictive distribution. Also, I am interested in how you propagated uncertainty in the predicted covariates? See for example Stoklosa et al. (2015).

8. pg. 8 line 21-25: Can you cite the main literature on the topic.

9. pg. 8 line 43-45: "A random search with a varying number of iterations (0, 10, 50, 100, 200) was applied to each fold of the tuning level." I have no clue what this means. It seems that this would be algorithm specific. You start to explain this later on (e.g., bottom of pg. 9), but by that point I am confused and lost trying to figure out what is being done.

10. pg. 9 line 34: "The AUROC was selected as a goodness of fit measure..." AUROC is used measure predictive accuracy (i.e., its a scoring rule) and not to determine how well the model (algorithm) fits the data.

11. pg. 10 line 10: "bias-reduced assessment of a model's predictive power." Can you please explain what this is. Also, it is not clear how you would measure bias (in any quantity) without making formal assumptions about a statistical model.

12. pg. 10 line 13: "While (semi-)parametric algorithms cannot be tuned in the same way as machine-learning algorithms." I disagree with this. For example, in the mgcv package you can fit a gam with a Gaussian process spatial random effect where the range parameter can be treated as a tuning parameters. Wood (2017) give this example in his book.

13. pg. 11 line 14-19: Can you please explain this in more detail. I have no clue how to implement this based on the description.

14. pg. 14 all: It seems that this should come sooner. This material is very good, but would have been better presented earlier in the paper (e.g., to make sense of the tuning parameter optimization steps).

15. pg. 17 line 52: Can you please mention what type of basis function you used with the gam? I am skeptical of your results. On one hand, what the RF will make create a very "rough" predictive surface (i.e., heat map), while gams (depending on the basis function used) will typically create a very smooth predictive surface. It seems odd to me that such contrasting approaches have similar predictive performance. Hefley et al. (2017b) has some discussion about this.

16. pg. 24 line 20: "environmental as well as medical contexts to reduce Bias" bias of what? Please explain.

References

Byrne, S. et al. (2016). A note on the use of empirical auc for evaluating probabilistic forecasts. *Electronic Journal of Statistics*, 10(1):380–393.

Diggle, P. J., Tawn, J., and Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). *The elements of statistical learning*. Springer series in statistics New York.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gotway, C. A. and Stroup, W. W. (1997). A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 157–178.

Heaton, M. J., Datta, A., Finley, A., Furrer, R., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., et al. (2017). Methods for analyzing large spatial data: A review and comparison. *arXiv preprint arXiv:1710.05013*.

Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., Tipton, J. R., Williams, P. J., and Hooten, M. B. (2017a). The basis function approach for modeling autocorrelation in ecological data. *Ecology*, 98(3):632–646.

Hefley, T. J., Hooten, M. B., Russell, R. E., Walsh, D. P., and Powell, J. A. (2017b). When mechanism matters: Bayesian forecasting using models of ecological diffusion. *Ecology Letters*, 20(5):640–650.

Kammann, E. and Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1):1–18.

Stoklosa, J., Daly, C., Foster, S. D., Ashcroft, M. B., and Warton, D. I. (2015). A climate of uncertainty: accounting for error in climate variables for species distribution models. *Methods in Ecology and Evolution*, 6(4):412–423.

Wikle, C. K., Cressie, N., Zammit-Mangion, A., and Shumack, C. (2017). A common task framework (ctf) for objective comparison of spatial prediction methodologies. *Statistics Views*.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Wood, S. N. and Augustin, N. H. (2002). Gams with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological modelling*, 157(2-3):157–177.