Testing sensitivity of spatial clustering and extrapolation in predictive modeling using Machine Learning hackathon

28.08.2022 - 03.09.2022 (Sunday to Sunday)

Friendly City Hotel Oktopus, Siegburg

Prepared by: T. Hengl & A. Bouasria

Backgrounds

The capacity of existing scientific methods (especially Machine Learning models) to produce accurate global maps of past vegetation, climate or human footprint is often approached or exceeded. In a recent Comment, Meyer and Pebesma (2022) discuss the challenges of machine learning approaches in generating global maps of ecological variables. This calls for data producers to extensively test, cross-validate and report mapping accuracy, and do a thorough job of highlighting the dangers that can result from extrapolating predictions beyond

the bounds of available training data. De Bruin et al. (2022) have tested effects of spatial clustering and extrapolation, and concluded that for strongly clustered design (points) there is basically no way to avoid bias in estimation except to collect additional samples. How dangerous is it to extrapolate using ML and are there remedies for extrapolation? How to quantify uncertainty in the extrapolation space? How to correctly deal with spatial clustering in training data? We show in this blog that Ensemble ML could potentially help with extrapolation problems, but are these conclusions universally applicable?

The focus of this hackathon is, thus, to test some of these concepts using real datasets, run repeated resampling modeling and Cross-Validation and evaluate effects of different experimental settings on accuracy. We are specifically interested in **sensitivity analysis** of standard machine learning techniques (e.g. Random Forest and similar) under (1) spatial clustering of training points, (2) extrapolation effects in feature space. These two problems of modeling also overlap, but can be treated as two separate problems.

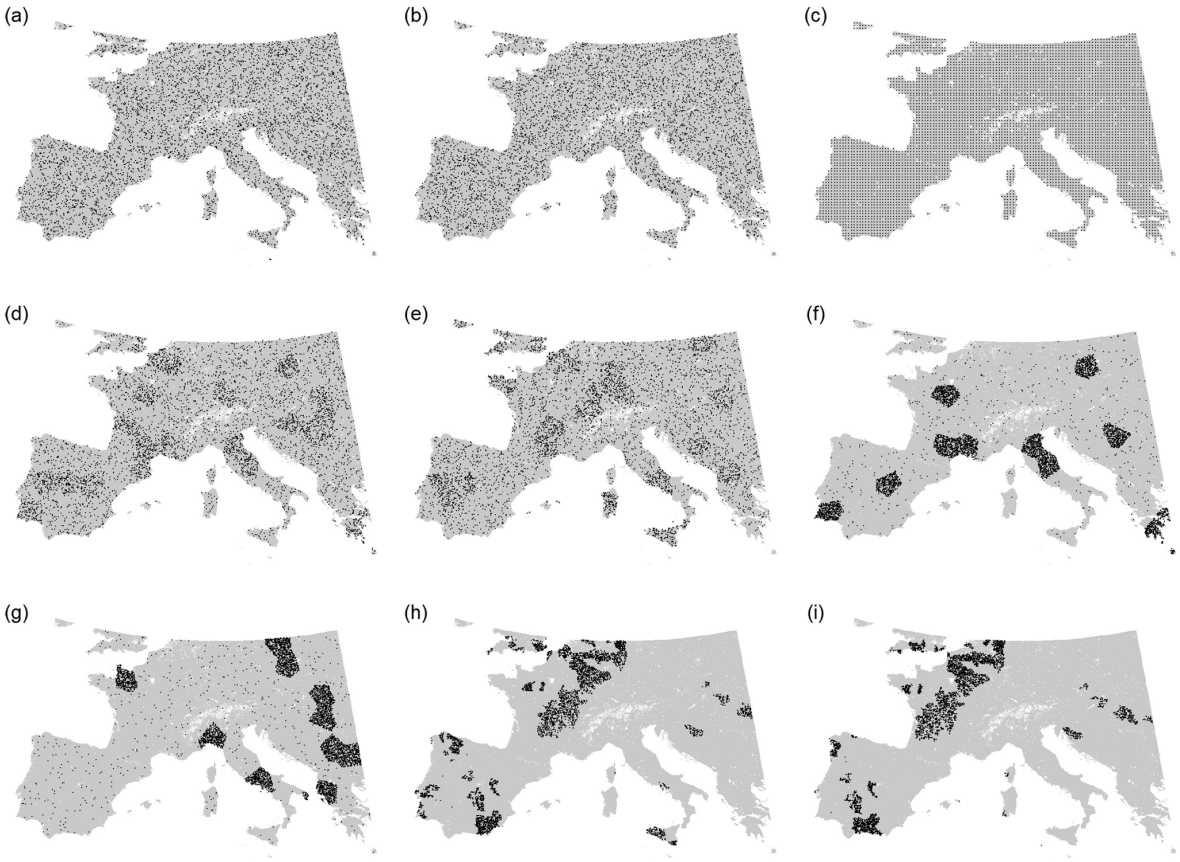


Fig: Examples of resampled spatial samples from De Bruin et al. (2022) indicating different levels of clustering (experimental design).

Task definition

Your task is to develop a R / Python computational notebook (R markdown or similar) that extends the **USGS NGS soil geochemicals mapping tutorial**. The computational notebook

should explain step-by-step all analysis you have conducted and should provide evidence for any claims you make. This is an **example** of how to test predictive performance of a ML method using repeated resampling.

Specific tasks include:

- ☐ Create a function to resample and subset existing probability samples to create pseudo-extrapolation samples e.g. with >90% of validation points in extrapolation (feature) space.
- ☐ Create a function to resample existing points to produce pseudo-clustered samples (geographical clustering) as in e.g. De Bruin et (2022).
- ☐ Create a function that tests mapping accuracy and returns summary accuracy metrics for different combinations of parameters (e.g. 2–3 prediction algorithms / learners, different clustering levels, different resampling strategies etc).
- ☐ Create a function that plots results of accuracy assessment in relation to: (a) N — size of training samples used to build models, (b) percent of validation points in extrapolation space and/or level of spatial clustering, (c) different geochemical variables of interest.

We are further looking at answering the following questions:

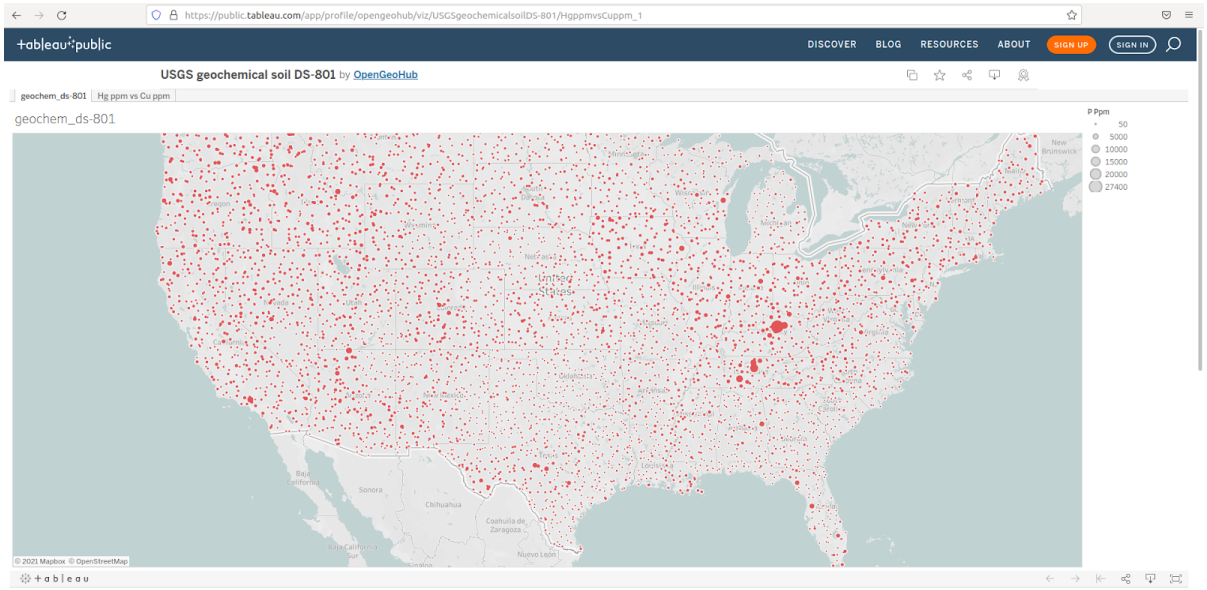
I. If more than >50% of validation samples are in extrapolation space, can ML generate unbiased estimates of the values of the target variable?
II. If more than >90% of validation samples are in extrapolation space, can ML generate unbiased estimates of the values of the target variable?
III. If more than >90% of training points are highly clustered (spatially) can ML produce unbiased estimates of the target variable (e.g. same distribution as without clustering)?
IV. What is the most accurate approach to estimate prediction errors regardless of the amount of extrapolation?
V. What is the most accurate approach to estimate the mapping accuracy (RMSE in log-space)?

We recommend focusing on RMSE at validation points and differences in distributions in predicted values at the validation points as the KPI's for testing. For validation we recommend using 10% of data randomly subset, then repeating the process 10×, then taking the global average (i.e. 10-fold Cross-Validation with refitting).

Data set

For testing of the extrapolation we use the Geochemical and minerological data set for USA48 (Smith, Woodruff, Solano, Ellefsen, & Karl, 2014). This is a public data set produced and maintained by the USA Geological Survey and contains laboratory measurements of chemical elements and minerals in soil at **4,857 sites** (for three depths 0 to 5 cm, A horizon and C horizon. The individual records can be browsed directly via https://mrdata.usgs.gov/ngdb/soil/. The points were produced using the generalized random tessellation stratified (GRTS) design, hence it is a point data set with perfect spread around the country and almost no gaps.

In addition to the point data we also provide a number of covariate layers (all at 1km spatial resolution) which were already tested for spatial prediction purposes.



You should try modeling and predicting at least three different target variables:

1. Pb concentration (**pb_ppm**);
2. Zinc concentration (**zn_ppm**);
3. Total soil carbon in pct (**c_org_pct**);

For all three variables above please use **log-transformed** values to report accuracy / compare distributions and to derive RMSE at validation points.

How will be the winner decided

The winner will be decided by the Hackathon committee that will evaluate your approach and code you produced, then assign points per different segment of your work. We will especially look at:

1. Are you using the correct methodological steps?
2. Have all tasks been completed?
3. Are conclusions clear and based on evidence?
4. Have you repeated resampling and refitting enough times so the results are not by chance?
5. Are all steps reproducible?

The committee will rate all answers and then sum-up points received and declare a winner. The winner then has to present all steps and provide a link to the github repository or similar.

References

1. De Bruin, S., Brus, D. J., Heuvelink, G. B., van Ebbenhorst Tengbergen, T., & Wadoux, A. M. C. (2022). Dealing with clustered samples for assessing map accuracy by cross-validation. Ecological Informatics, 69, 101665. https://doi.org/10.1016/j.ecoinf.2022.101665
2. Lovelace, R., Nowosad, J., & Muenchow, J. (2022). Geocomputation with R. CRC Press. 2nd edition; https://geocompr.robinlovelace.net/
3. Lu, B., & Hardin, J. (2021). A unified framework for random forest prediction error estimation. Journal of Machine Learning Research, 22(8), 1–41. Retrieved from http://jmlr.org/papers/v22/18-558.html
4. Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21(15), 3301-3307. https://doi.org/10.1093/bioinformatics/bti499
5. Meyer, H. & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12, 1620–1633. https://doi.org/10.1111/2041-210X.13650
6. Meyer, H., Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. Nat Commun 13, 2208 (2022). https://doi.org/10.1038/s41467-022-29838-9
7. Smith, D. B. C., Woodruff, W. F., Solano, L. G., Ellefsen, F., & Karl, J. (2014). Geochemical and mineralogical maps for soils of the conterminous united states. Denver, CO: USGS Geology, Energy,; Minerals Science Center. Retrieved from https://pubs.usgs.gov/ds/801/