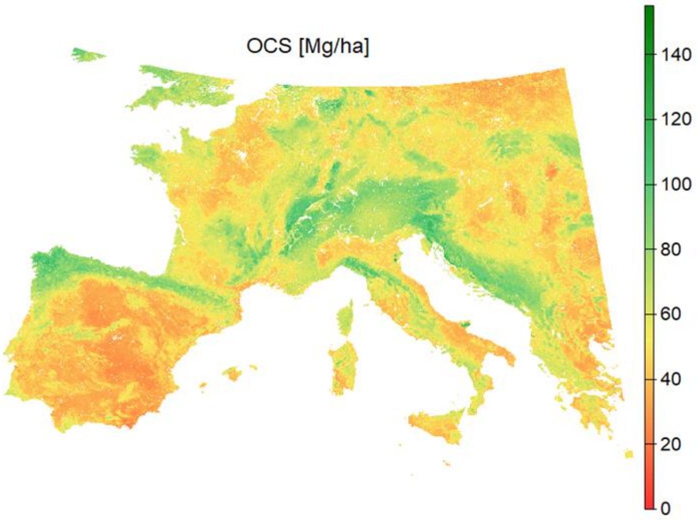
STATISTICAL VALIDATION AND CROSS-VALIDATION



In contrast to the previous lectures and labs focussing on input uncertainty propagation and uncertainty concerning model structure and model parameters, today we deal with validation of model results against reference data. In environmental sciences, the model results are often maps and today afternoon we validate soil organic carbon stock (OCS) maps created from observational data and environmental covariates, using a random forests (RF) model as the predictor. Note that in this course we only briefly address the mapping itself; we focus on the validation of the output of the mapping procedure. Nevertheless, the prediction is included in the script, so you can see how it works. Apart from that, in a cross validation setting we need to repeatedly predict at subsets of the data points, which are not used for training the model.

The map shown at the top right of this page was extracted from Soilgrids (https://soilgrids.org/; Poggio et al., 2021) and it serves as the reference (sometimes referred to as ground truth) OCS in our practical. We extract training data from that map and employ that data for training a RF model using a stack of environmental covariates (https://zenodo.org/records/6513429) as the predictors. Next, we predict OCS by the trained RF model. Using a second sample (while avoiding training data leakage) drawn from the reference OCS map we assess the accuracy of the RF predictions by the mean error (ME), the root mean squared error (RMSE) and the model efficiency coefficient (MEC) explained in this morning’s lecture. Furthermore, we compute the prediction interval coverage probability (PICP), which assesses the prediction model’s uncertainty predictions.

First, we rely on classical map accuracy assessment, which is solidly rooted in sampling theory wherein an unbiased estimate of map accuracy (e.g. mean squared map error) is obtained by design-based inference from a probability sample that is independent of the data used for model fitting (e.g., Stehman, 2009). We also address cross-validation, which repeatedly splits the full data set into subsets used for model fitting and for testing its predictions.

Input data

The input data are provided on MS Teams and they originate from the paper by de Bruin et al. (2022). The study area is western Europe, constrained in the north at 52° latitude and at -10° and 24° longitude The projection is `IGNF:ETRS89LAEA` (Lambert azimuthal equal area projection).

`OCSstack.tif`	Reference OCS and covariates used for predicting OCS at 500m resolution. The covariates are described at https://zenodo.org/records/6513429.
`OCSdata050.Rdata`	Strongly clustered sample of size 5000, holding reference OCS and covariates. It is one of 100 strongly clustered samples used in the above paper.

Download (and extract) the data to a folder with a convenient name on your local drive. Notice that the `.tif` file is fairly large as it covers large part of Europe and has 20 data layers at 500m resolution.

Script

Also download the R script for this practical that is provided in MS Teams (`ScriptThursday2024.R`). Make sure the required libraries are installed and let line 13 point to the correct data folder.

Predict OCS with RF model

The first 17 lines of the script load the libraries required in this afternoon's practical, as well as the OCS reference data along with 19 covariates. In lines 19-20, the reference OCS map is plotted using a customized colour palette. Next, the random forest model is fitted on a replicable random sample (n = 2500) in lines 26-35. Lines 40-43 define a wrapper function around `predict.ranger()` to facilitate prediction of the study area using the `terra` package. The actual prediction with the random forest models is called in lines 44-45.

Q1 *Run lines 1-48 of the script. Owing to the size of the study area and the resolution of the data, the prediction takes several minutes. Visually compare the reference map and the predicted OCS map. What main difference do you observe between the maps and how can it be explained?*

Validation with independent sample

Validation using an exhaustive sample

The next lines of the script employ samples of independent reference data. As explained in the lecture, this allows model-free, unbiased estimation of several validation metrics. Lines 55-65 implement an exhaustive validation. Since we have reference data for all map units (pixels) we can afford comparing all map predictions against OCS reference values. In real life this is of course never the case, as there would be no point in making map predictions if the true values are known everywhere. However, here it allows us to get reference validation metrics for comparison purposes.

Q2 *Run the code of lines 55-65. Take note of the values of the validation metrics. How would you judge the accuracy of the produced OCS map?*

Q3 *In line 55, many reference cells' values are set to NA (no data). What would be the reason for doing so?*

Validation using simple random sampling

The simplest form of probability sampling is simple random sampling (SRS). As implied by the name, it randomly samples from the population giving equal inclusion probability to each population unit, which means that each unit has an equal probability of being included in the sample. In lines 71 – 84, a random sample of size 350 is drawn and estimators are used for computing validation metrics.

Q4 *What are the inclusion probabilities of the map pixels used in the exhaustive validation and what of those in the simple random sampling case?*

Q5 *Run the code of lines 71 – 84. Why are the values of the validation metrics different from those of the exhaustive sample?*

Q6 *In the script also the standard error of the ME and upper and lower 90% confidence limits for the RMSE are computed. What do those measures of precision tell us and why were they not computed for the exhaustive validation?*

Q7 *If you many times take a sample of size 350 from the map(s) and average the validation metrics (ME and RMSE) computed over all samples, what values would you expect?*

Lines 86 – 101 implement the repeated sampling referred to in Q7.

Q8 *Run the code in lines 86 - 101. How does the mean of the ME computed over 100 samples compare to that obtained from the exhaustive sample? What would happen in case you increase the number of replicates?*

Q9 *Lines 98 – 101 of the script compute how often the reference ME falls within the confidence intervals computed from the 100 samples. Is the result according to your expectation and what do you expect to see upon increasing the number of replicates?*

Stratified random sampling

In stratified random sampling, the population is first divided into several strata, which are next independently and randomly sampled. In the subsequent statistical inference, the parameters for the full population are estimated from the strata. Compared to SRS, stratification can increase the efficiency of sampling, particularly if the allocation of sample units over the different strata is carefully considered. Increasing the efficiency means that population parameters (e.g., the mean of residuals) are more precisely estimated using the same sample size as used in SRS, or that a smaller sample size suffices to reach a target precision.

Here we demonstrate stratification using the land cover categories provided in the ninth layer of the `OCSstack.tif` file. That is because these strata are readily available in the dataset and they may be meaningful. Note that the estimators of the validation metrics for stratified random sampling are different from those of simple random sampling. You can find the estimators implemented in the code, where it is readily observed that the stratum size is taken into account. Several textbooks on sampling theory provide estimators for different sampling designs, including those for stratified random sampling.

Q10 *Run lines 112 – 118 of the script. What sample size is allocated to each of the strata? Do you expect this to be a good choice? Why?*

Q11 *What do the strata weights computed in lines 117 – 118 represent?*

Q12 *Run lines 120 – 138 of the script. Note how the strata weights are used in the statistical inference. How do the validation metrics compare to those computed from the SRS? What could explain this result?*

In lines 139 – 163, the stratified random sampling and inference from the sample are repeated but now using proportional sample allocation to the strata. This implies each stratum is sampled using a sample size that is proportional to the area fraction it takes over the entire study area. For estimating the stratum variance, at least two sample units per stratum are needed.

Q13 *Run lines 139 – 163 of the script. How do the validation metrics compare to those computed from the SRS and those of equal allocation (Q12)? What allocation scheme would you prefer?*

Prediction interval coverage probability (PICP)

As explained in the lecture, the PICP assesses the fraction of sampled reference values that are within the prediction intervals (PIs) obtained from our analysis. Quantile regression forests (QRF) predict quantiles of the target variable. By predicting a sequence of those quantiles, PIs of different widths are obtained. For each of these widths we then assess the fraction of sampled reference values within the PI. This is implemented in lines 170 – 193 of the script.

Q14 *Run lines 170 – 193 of the script. What do you conclude about the uncertainty quantification of the QRF? Is it too pessimistic, too optimistic or on the spot?*

Cross-validation

Above we have used an independent reference data set for validating map predictions. In practice, such so-called post-mapping probability samples that are exclusively used for map evaluation are often not available and therefore alternative methods have been proposed. The widely used *k*-fold cross-validation method splits the full dataset into *k* approximately equally-sized disjoint subsets or folds, where repeatedly (i.e. *k* times) the model is calibrated on *k*-1 folds, whilst the remaining fold is used for assessing prediction accuracy. The overall cross-validation accuracy is estimated by aggregating the (squared) residuals over the *k* folds. In conventional *k*-fold cross-validation, the folds are chosen randomly.

There has been ample discussion about validity of such approach (see Wadoux et al., 2021; de Bruin et al., 2022) and several alternatives have been proposed. These are beyond the scope of today's practical, but we may discuss some of them if you are interested. Here we only consider weighing the (squared) residuals at sampled points by the inverse of their sampling intensity to account for (strong) clustering of the existing sample. Otherwise, we give most weight to the areas that are also most intensely sampled and probably most accurately mapped. Lines 196 – 267 implement a 10-fold cross-validation using a strongly clustered sample from a recent paper we wrote (de Bruin et al., 2022).

Q15 *Run lines 202 – 227 of the script. Are the validation statistics too optimistic or too pessimistic? How come?*

In the remaining code the (squared) residuals obtained by random 10-fold cross-validation are weighted by the inverse of the sampling intensity that is estimated using a kernel density procedure from the `spatstat` library.

Q16 *Run lines 233– 267 of the script. Are the validation statistics too optimistic or too pessimistic and how do they compare to the results obtained in Q15? Can you explain?*

References

de Bruin, S., Brus, D.J., Heuvelink, G.B.M., van Ebbenhorst Tengbergen, T., Wadoux, A.M.C., 2022. Dealing with clustered samples for assessing map accuracy by cross-validation. Ecol. Inform., 69, 101665. https://doi.org/10.1016/j.ecoinf.2022.101665

Poggio, L., de Sousa, L.M. Batjes, N.H., Heuvelink, G.B.M., Kempen, B, Ribeiro E, Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7(1): 217-240. https://doi.org/10.5194/soil-7-217-2021

Stehman, S.V., 2009. Sampling designs for accuracy assessment of land cover. International Journal of Remote Sensing 30(20), 5243-5272. https://doi.org/10.1080/01431160903131000

Wadoux, A.M.J.C., Heuvelink, G.B.M., de Bruin, S., Brus, D.J., 2021. Spatial cross-validation is not the right way to evaluate map accuracy. Ecological Modelling 457, 109692. https://doi.org/10.1016/j.ecolmodel.2021.109692