Validation of (digital) soil maps

Bas Kempen

ISRIC
World Soil Information

Map validation

- No map is perfect. All maps, including soil maps, are representations of reality that are often based on an underlying model.
- This means that there will always be a deviation between the phenomenon depicted on the map and the phenomenon observed in the real world, i.e. each map will contain errors.
- Validation is defined an activity in which the soil map predictions are compared with observed values. From this comparison, the **map quality** can be **quantified** and **summarized** using map quality measures.

Why validate?

- Validation is an important step in the soil mapping workflow.

- Why do we want to validate: soil maps are not perfect!
  - One should want to check the quality of ones work before this is made public
  - Compare the performance of methods
  - End users must know the quality of maps to judge their usability for specific purposes

- Validation provides summary (global) measures of accuracy: how accurate the map is on average for the mapping area (i.e. what is the expected error at a randomly selected location in the mapping area)

Validation data

- Internal versus external accuracy.

- Validation should be done with independent data, i.e. data not used for the production of the soil map.

Quality measures quantitative soil maps

- Prediction error:

$$e(\mathbf{s}) = \hat{z}(\mathbf{s}) - z(\mathbf{s})$$

- Mean error (bias: systematic over- or underestimation):

$$\mathrm{ME} = \bar{e} = \frac{1}{N}\sum_{i=1}^{N} e(\mathbf{s}_{\boldsymbol{i}})$$

- Mean absolute error and (root) mean squared error:

$$\mathrm{MAE} = |\bar{e}| = \frac{1}{N}\sum_{i=1}^{N} |e|(\mathbf{s}_{\boldsymbol{i}})$$

$$\mathrm{MSE} = \overline{e^2} = \frac{1}{N}\sum_{i=1}^{N} e^2(\mathbf{s}_{\boldsymbol{i}})$$

Quality measures quantitative soil maps

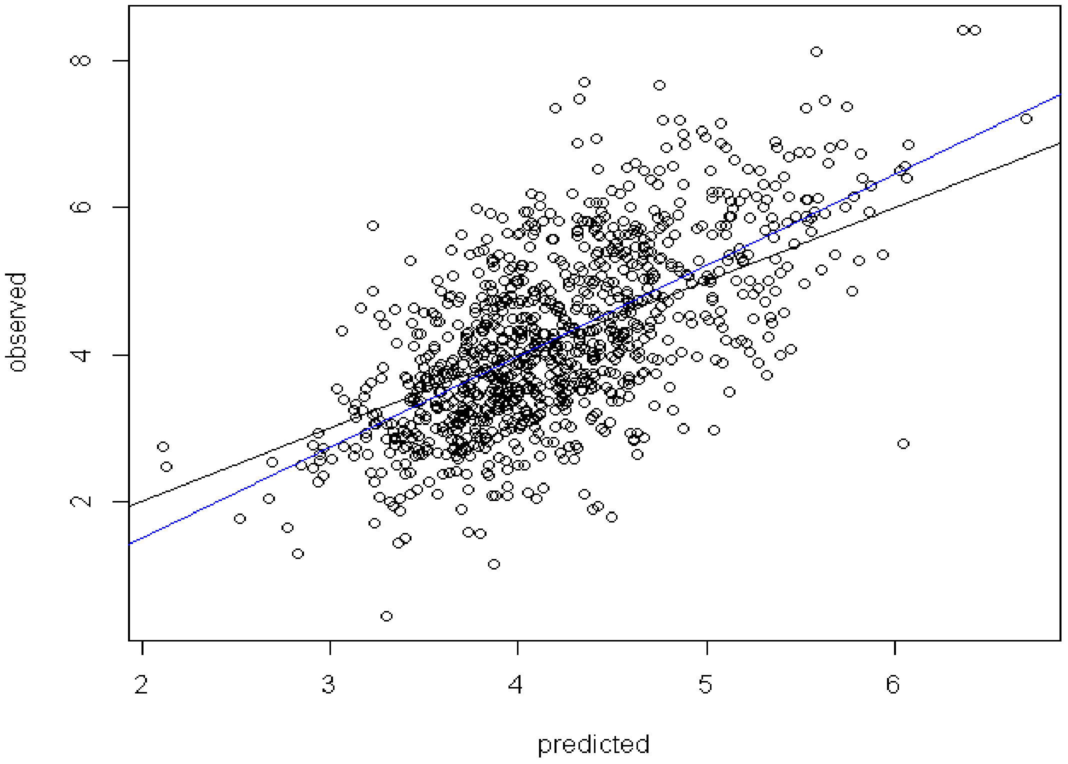
- Amount of variance explained:

$$AVE = 1 - \frac{\sum_{i=1}^{N} \left(\hat{z}(\mathbf{s}_{\boldsymbol{i}}) - z(\mathbf{s}_{\boldsymbol{i}}) \right)^2}{\sum_{i=1}^{N} (z(\mathbf{s}_{\boldsymbol{i}}) - \bar{z})^2}$$

- It is not good practice in validation to do a regression between the observed and predicted value and use the R^2 as measure for the amount of variance explained.

Quality measures quantitative soil maps

- Black: 1:1 line; Blue: regression line
- AVE = 0.40; $R^2 = 0.42$



Quality measures quantitative soil maps

- Mean square deviation ratio: measures how well the prediction model estimates the prediction uncertainty:

$$MSDR = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(\hat{z}(\mathbf{s}_{\boldsymbol{i}}) - z(\mathbf{s}_{\boldsymbol{i}})\right)^2}{\sigma^2(\mathbf{s}_{\boldsymbol{i}})}$$

- MSDR should be 1; median SDR 0.455
- MSDR < 1: model overestimates the uncertainty
- MSDR > 1: model underestimates the uncertainty

Estimating map quality measures

- We **estimate** the population means of the map quality measures from a sample taken from a limited number of locations in the mapping area.

- Because we estimate, we are **uncertain** about the estimations. [When the validation data are collected properly (with a probability sampling design) we can quantify this uncertainty.]

- We use the same equations as before but now *N* is replaced by *n:*

$$\text{ME} = \hat{\hat{e}} = \frac{1}{n}\sum_{i=1}^{n} e(\mathbf{s}_{\boldsymbol{i}})$$

Sampling types

- Purposive sampling: locations are selected purposively (e.g. representative, good spatial coverage)

- Haphazard sampling: locations are selected arbitrarily, ‘more-or-less random’. Be ware: this is **NOT** a random sample.

- Probability sampling: locations are selected randomly (based on a probability mechanism). Inclusions probabilities are known.

Validation methods

- Methods:
  - Additional probability sampling
  - Data-splitting
  - Cross-validation (*n*-fold, leave-one-out)

- Data-splitting and cross-validation if there is only one dataset available for calibration and validation.

Literature:

- Brus, Kempen, Heuvelink, 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62, 394-407.

Additional probability sampling

- For validation it is preferred to use data collected from randomly selected locations, because:
  - no model is needed for estimating map quality estimates. We can apply *design-based estimation*, meaning that model-free unbiased and valid estimates of the map quality measures can be obtained;
  - discussions on the validity of the estimated map quality are avoided;
  - model-free, valid estimates of the variance of the map quality measures can be obtained that allow for hypothesis testing, e.g. for comparison of model performance.

Additional probability sampling

- All sampling units have probability >0 of being selected, but the probabilities need not be equal
- Inclusion probabilities must be known for all sampling units in the population
- Inclusion probabilities are known by design and are used to estimate the quality measures: design-based estimation
- Various designs: sample, stratified, clustered, two-stage , systematic random sampling

Data-splitting

- Sample data set is split in two subsets.
- One subset is used to calibrate the prediction model. The other subset is used for validation.
- A frequently used splitting criterion is 70-30, where 70% of the sample data are used for calibration and 30% for validation.
- For sparse data sets, data-splitting can be inefficient since the information in the data set is not fully exploited for both calibration and validation.

Data-splitting

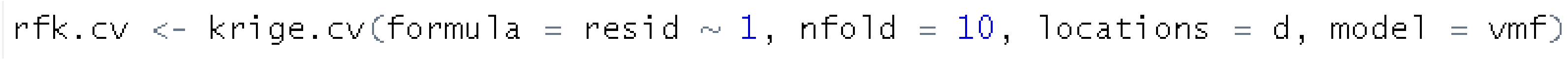
- A random subsample of a *non-probability sample* is not a probability sample of the study area -> design-based estimation of quality measures impossible.
- If the validation subset is a *non-probability sample* of the study area -> one must account for possible spatial autocorrelation of prediction (classification) error, i.e. model-based estimation.
- Often, spatial autocorrelation is not accounted for. Map quality measures cannot be considered unbiased and valid estimates of the population means.
- In this case, the map quality measures are only valid at the validation locations.

Cross-validation

- *n*-fold cross-validation
- Procedure:
  - Data is split in *n* subsets of equal size.
  - Model is calibrated using data from *n*-1 subsets.
  - Model is used to predict at the subset left out.
  - Repeated *n* times: each time setting aside a different subset.
- Special case: leave-one-out (*n*=number of samples)
- More efficient than data-splitting
- Problem of spatially correlated errors remains

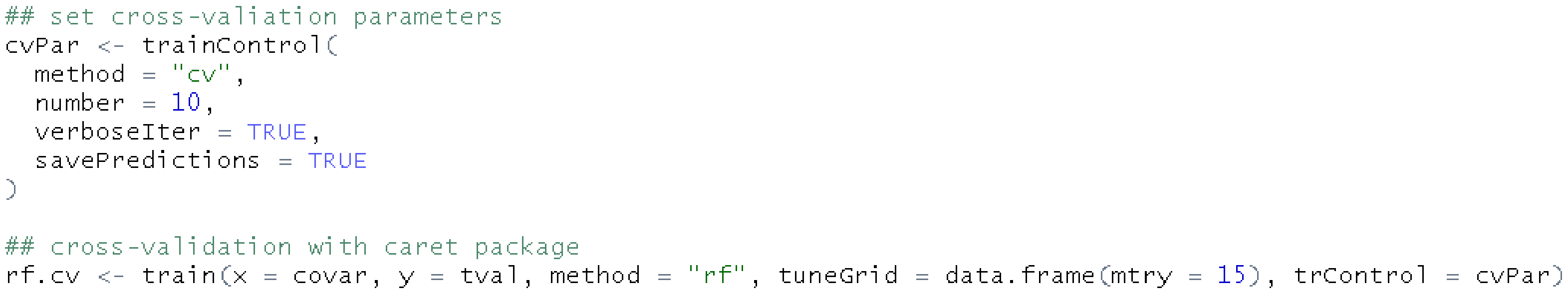
Tools in R

- Kriging *n*-fold cross-validation: **krige.cv** function of the **gstat** package.



```
rfk.cv <- krige.cv(formula = resid ~ 1, nfold = 10, locations = d, model = vmf)
```

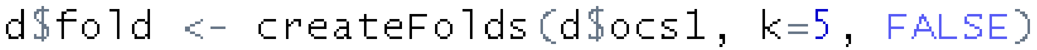
- *n*-fold cross-validation: **train** function of the **caret** package.



```
## set cross-valiation parameters
cvPar <- trainControl(
  method = "cv",
  number = 10,
  verboseIter = TRUE,
  savePredictions = TRUE
)

## cross-validation with caret package
rf.cv <- train(x = covar, y = tval, method = "rf", tuneGrid = data.frame(mtry = 15), trControl = cvPar)
```

- Subsetting data: **createFolds** function of the **caret** package.



```
d$fold <- createFolds(d$ocs1, k=5, FALSE)
```

- Note: using categorical covariates might give problems for cross-validation

Thanks!!