Model Evaluation

D G Rossiter

Nanjing Normal University, Geographic Sciences Department
南京师范大学地理学学院

Cornell University, Section of Soil & Crop Sciences

November 26, 2018

1. Assessment of model quality
2. Internal evaluation
   - Kriging prediction variance
3. Independent evaluation
   - Evaluation measures
   - Linn's Concordance
4. Resampling
5. Cross-validation

Overview

- Assessment of model quality: overview
- Model evaluation with an independent data set
- Cross-validation

1 Assessment of model quality

2 Internal evaluation

Kriging prediction variance

3 Independent evaluation

Evaluation measures

Linn's Concordance

4 Resampling

5 Cross-validation

Assessment of model quality

- With any predictive method, we would like to know how good it is. This is model **evaluation**, often called model **validation**.
  - contrast with model **calibration**, when we are building (fitting) the model
- Prefer the term **evaluation** because "validation" implies that the model is correct ("valid"); that of course is never the case. We want to **evaluate** how close it comes to reality.
  - Oreskes, N. (1998). Evaluation (not validation) of quantitative models. *Environmental Health Perspectives*, 106(Suppl 6), 1453–1460.
  - Oreskes, N., *et al.* (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263, 641–646.[1]
- However, we still use the term **cross-validation**, for historical reasons and because the `gstat` function is so named.

[1] https://doi.org/10.1126/science.263.5147.641

Internal vs. external quality assessment (1)

External If we have an **independent data set** that represents the target population, we can **compare model predictions with reality**. Two types:

1. Completely separate **evaluation dataset** from a **target population** to be evaluated
2. **Cross-validation** using the **calibration dataset**, leaving parts out or resampling

Internal vs. external quality assessment (2)

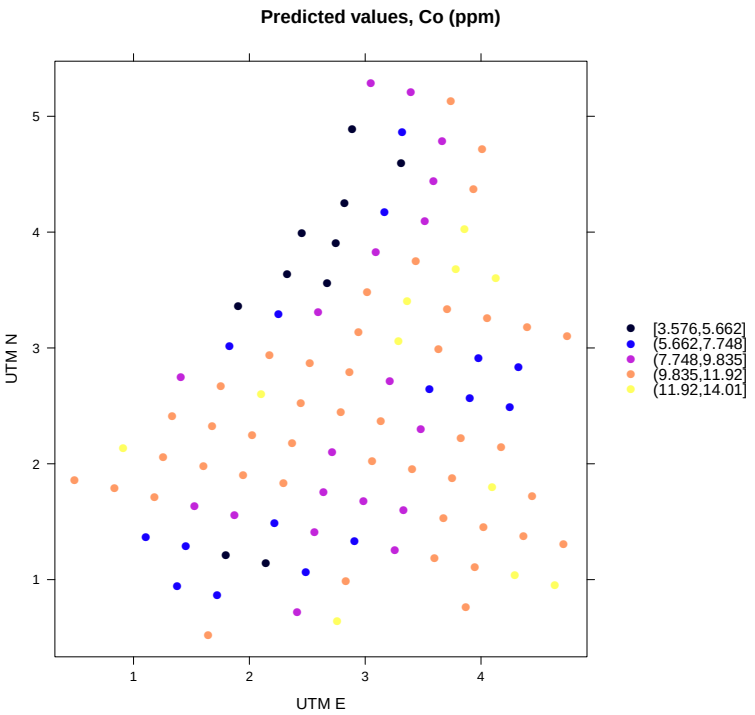
Internal Most prediction methods give some measure of **goodness-of-fit** to the **calibration data set**:

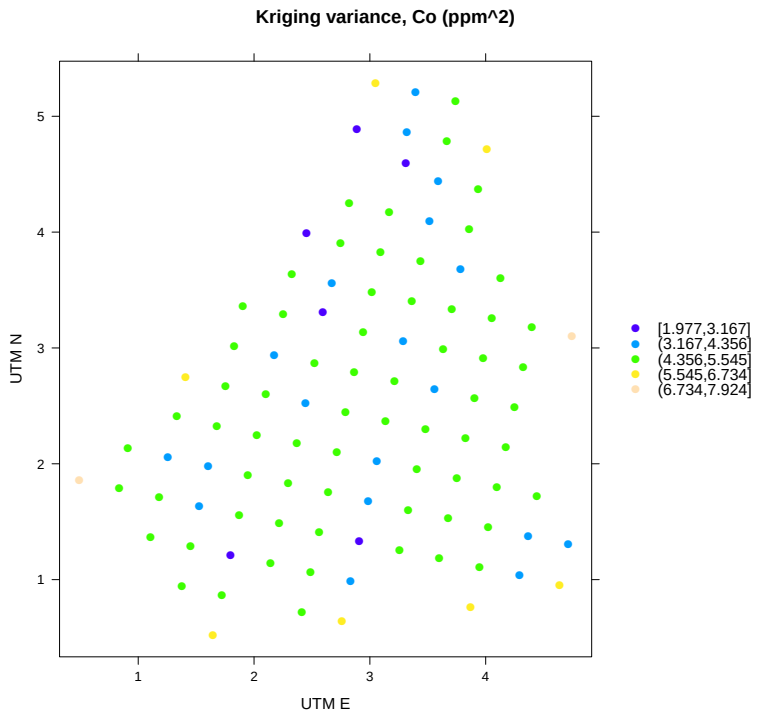
- **Linear models**: **coefficient of determination** R^2
  - Warning! Adding parameters to a model increases its fit; are we fitting **noise** rather than **signal**? Use adjusted measures, e.g. adjusted R^2 or Akaike Information Criterion (AIC)
- **Kriging**: the uncertainty of each prediction, i.e., the **kriging prediction variance**

Internal evaluation of Kriging predictions

- Because of its model structure, Kriging automatically computes a **kriging prediction variance** to go with each prediction.
- This is because that variance is **minimized** in kriging, *assuming the model of spatial dependence is correct*!
  - Variogram form, variogram parameters
  - OK: Assumptions of 1st and 2nd order stationarity (mean, covariance among point-pairs)
  - KED/UK: Assumptions of 2nd order stationarity (covariance among point-pairs model *residuals*)
  - KED/UK: Linear model assumptions to give 1st order stationarity of residuals
- This kriging prediction variance depends *only* on the **point configuration** of the known points, and the **model of spatial correlation**,*not* on the data values!
- In theory this gives the uncertainty of each prediction → internal evaluation

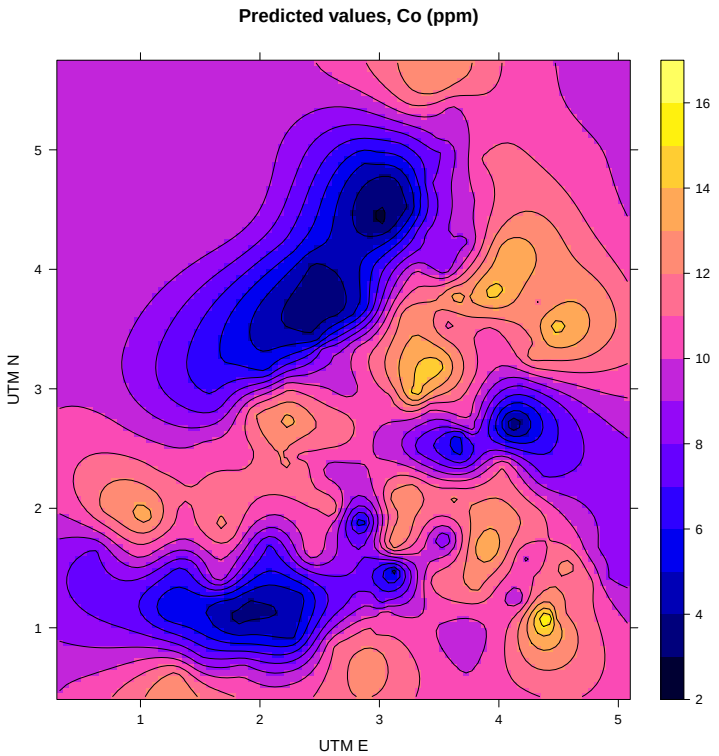
Kriging predictions and variance at points

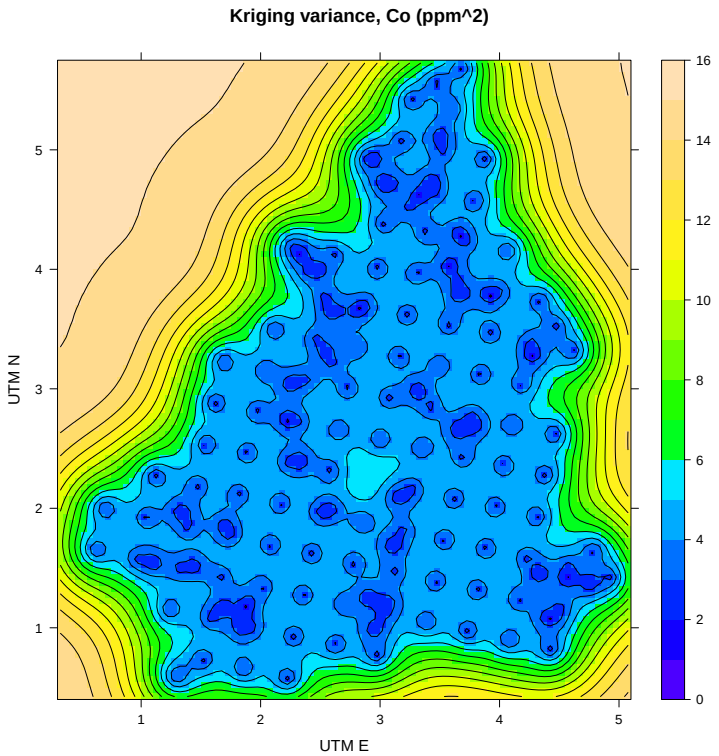




Jura (CH) topsoil heavy metals – Ordinary Kriging

Kriging predictions and variance over a grid





Jura (CH) topsoil heavy metals – Ordinary Kriging

Prediction outside the range of spatial dependence is the *spatial mean* and *covariance*

Numerical summaries of kriging variance

- Mean, maximum kriging prediction variance
  - mean: on average, how precise is the prediction?
  - maximum: what is the worst precision?
- These can be used as *optimization criteria* for comparing sampling plans, for samples to be used for Kriging

Model evaluation with an independent dataset

An excellent check on the quality of any model is to compare its **predictions** with **measured values** from an **independent data set**.

- This set can *not* be used in the calibration procedure!
- This set *must* be from the **target population** for the evaluation statistics
  - **same** sampling campaign, observations randomly removed from the calibration procedure
  - a **different** sampling campaign, either the same or another target population
- **Advantages**:
  - objective measure of quality
  - can be applied to a separate population to determine extrapolation power of the model
- **Disadvantages**:
  - Higher cost
  - Less precision? Not all observarions can be used for modelling (→ poorer calibration?)

Selecting the evaluation data set

- The validation statistics presented next apply to the **evaluation** ("validation") set.
- It must be a **representative** and **unbiased** sample of the **population** for which we want these statistics.
- Two methods:
  1. **Completely independent**, according to a sampling plan;
     - This can be from a different population than the calibration sample: we are testing the applicability of the fitted model for a **different target population**.
  2. A **representative** subset of the original sample.
     - A **random** splitting of the original sample
     - This evaluates the population from which the sample was drawn, only if the original sample was unbiased
     - If the original sample was taken to emphasize certain areas of interest, the statistics do not summarize the validity in the whole study area

Evaluation measures (1)

- **Root mean squared error** (RMSE) of the **residuals** (actual – predicted) in the **validation** dataset of n points; how close **on average** are the predictions to reality?
- **lower is better**
- computed as:

$$\mathrm{RMSE} = \left[\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right]^{1/2}$$

  - where: $\hat{y}$ is a prediction; y is an actual (measured) value
- This is an estimate of the **prediction error**
- An overall measure, can be compared to desired precision
- The entire **distribution of these errors** can also be examined (max, min, median, quantiles) to make a statement about the model quality

Evaluation measures (2)

- **Bias** or mean prediction error (MPE) of estimated vs. actual mean of the **validation** dataset
- closer to zero is better (0)

$$\mathrm{MPE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

Relative evaluation measures

- The MPE and RMSE are expressed in the original units of the target variable, as *absolute* differences.
- These can be compared to criteria external to the model, i.e., "fitness for use".
- These can also be compared to the *dataset values*:
  - MPE compared to the **mean** or **median**
    - Scales the MPE: how signficant is the bias when compared to the overall "level" of the variable to be predicted?
  - RMSE compared to the **range**, **inter-quartile range**, or **standard deviation**
    - Scales the RMSE: how significant is the prediction variance when compared to the overall variability of the dataset?

Putting RMSE in context

- The RMSE tells us how closely the model **on average** predicts to the true values
- But, *is this significant in the real world*?
  - relative to the *values* of the target variable;
  - relative to *precision* needed for an application of the model.
- Relative to target variable: RMSE as a **proportion** of the mean
- Relative to application: RMSE as uncertainty, e.g., deciding whether a value is above or below a *critical value*

Example: Relative to population

- Meuse heavy metals dataset: Cross-validation RMSE from OK of log10(Zn) is 0.173.
- How does this compare to the population?
- Estimate from the sample:

```
> summary(log10(meuse$zinc))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.053   2.297   2.513   2.556   2.829   3.265
> rmse <- 0.173
> rmse/mean(log10(meuse$zinc))
[1] 0.06767965
```

- This is about 7% of the mean value of *this* sample of *this* population.

Example: Regulatory threshold

- According to the Berlin Digital Environmental Atlas[2], the critical level for Zn is 150 mg kg^{-1}; crops to be eaten by humans or animals should not be grown in these condition.
- $log_{10}(150) = 2.177$; suppose we have a RMSE of 0.173.
- So to be sure we are *not* in a polluted spot with 95% confidence we should measure no more than 77 mg kg^{-1}.

```
> (lower.limit <- log10(150)-(qnorm(.95)*0.173))
[1] 1.891532
> 10^(lower.limit)
[1] 77.89895
```

- So we may be forcing farmers out of business for no reason.

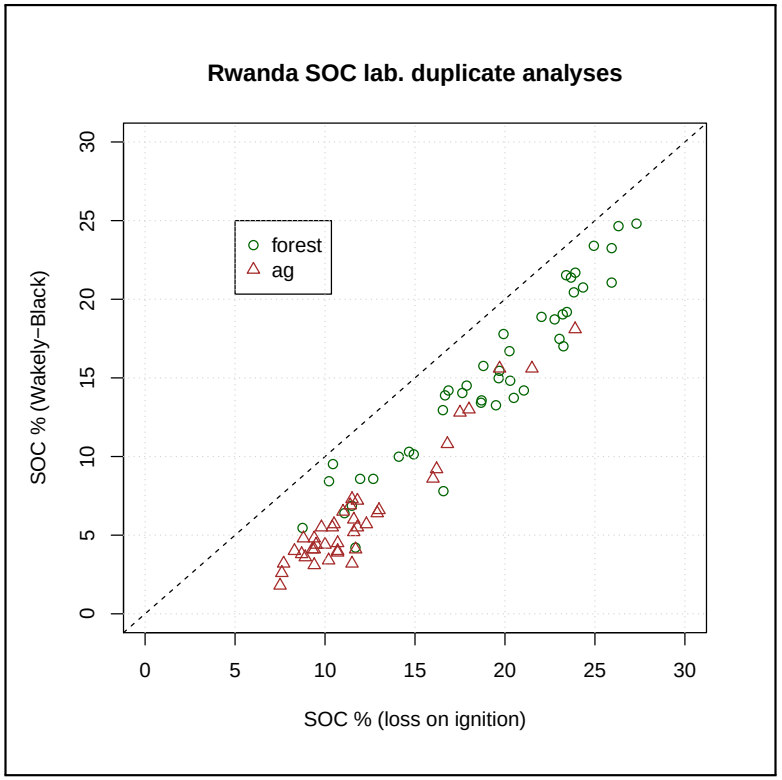
[2] http://www.stadtentwicklung.berlin.de/umwelt/umweltatlas/ed103103.htm

Regression of actual on predicted

- We can also compute a **linear regression**: $y = \beta_0 + \beta_1 \hat{y}$
- This shows how predictions made by the model from the calibration set could be adjusted to fit the evaluation set.
- β_0 is the **bias** of the fitted model; this should be 0.
- β_1 is the **gain** of the fitted model vs. the evaluation set; this should be 1.
- The R^2 of this equation is *not* an evaluation measure of the model!
  - It *does* tell us how well the adjustment equation is able to match the two sets.

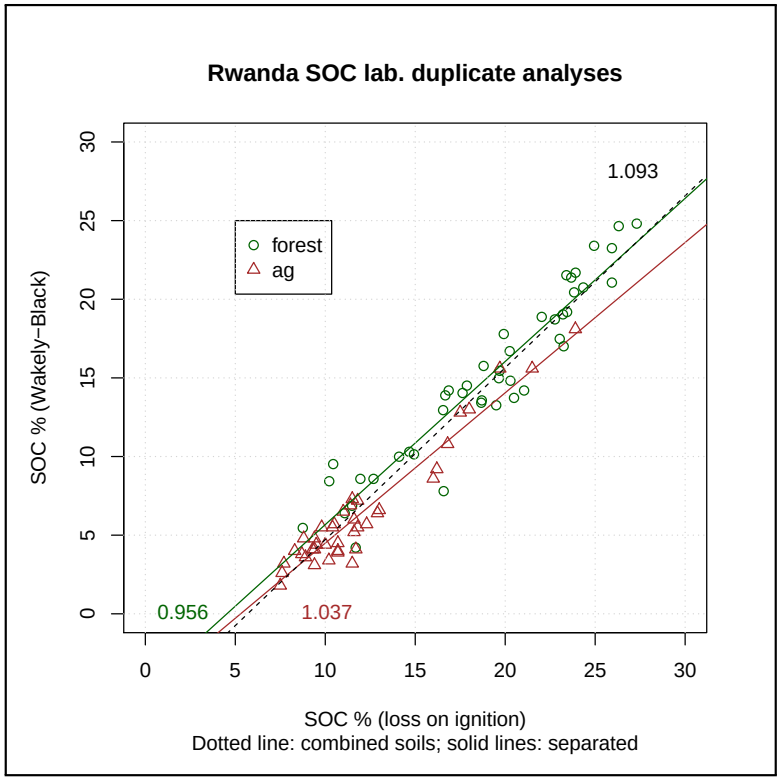
Visualizing actual vs. predicted

Scatterplot against 1:1 line



ME, RMSE

Regression



gain, bias

Lin's Concordance

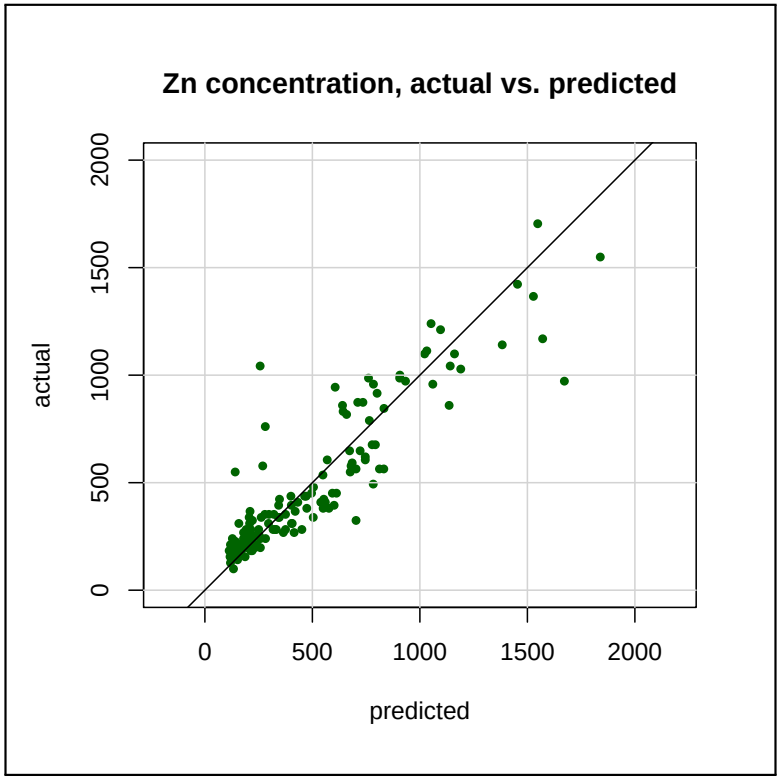
- A measure of the deviation from the 1:1 line
  - first developed to evaluate **reproducibility** of test procedures that are supposed to give the same result
  - also valid to compare **actual vs. predicted** by any model, these are supposed to be the same
-

$$\rho_c = \frac{2\rho_{1,2}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}$$

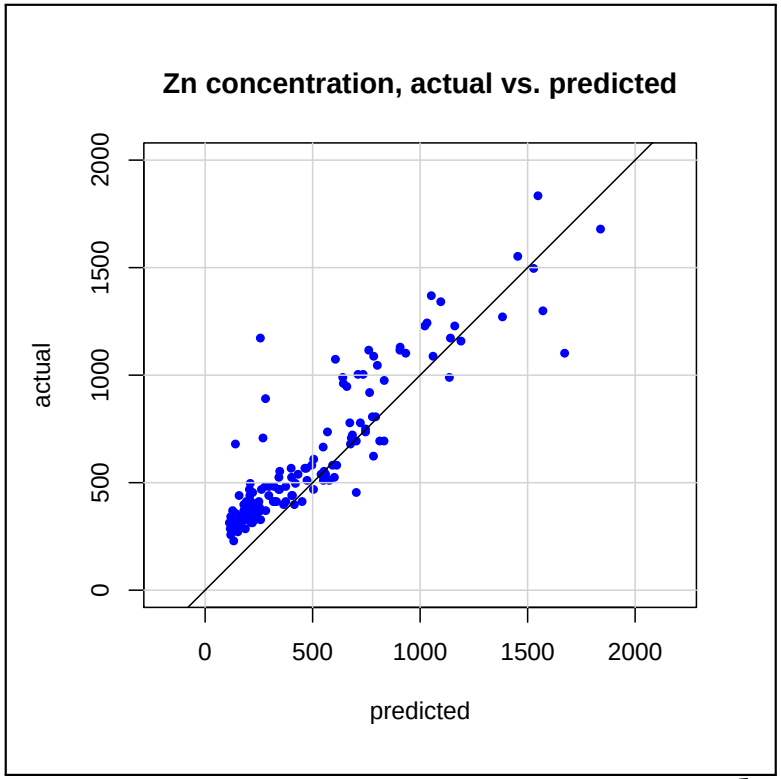
- Includes all sources of deviation:
  - location shift (bias) $(\mu_1 - \mu_2)/\sqrt{\sigma_1\sigma_2}$
  - scale shift (slope not 1) σ_1/σ_2
  - lack of correlation (spread) $1 - \rho_{1,2}$
- if points are *independent* use the *sample* estimates $r_{1,2}, S_1, S_2, \overline{Y_1}, \overline{Y_2}$

Reference: Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45(1), 255–268.

Lin's Concordance – examples



Concordance: 0.900 (no bias)



0.846 (bias +100 mg kg^{-1})

Resampling

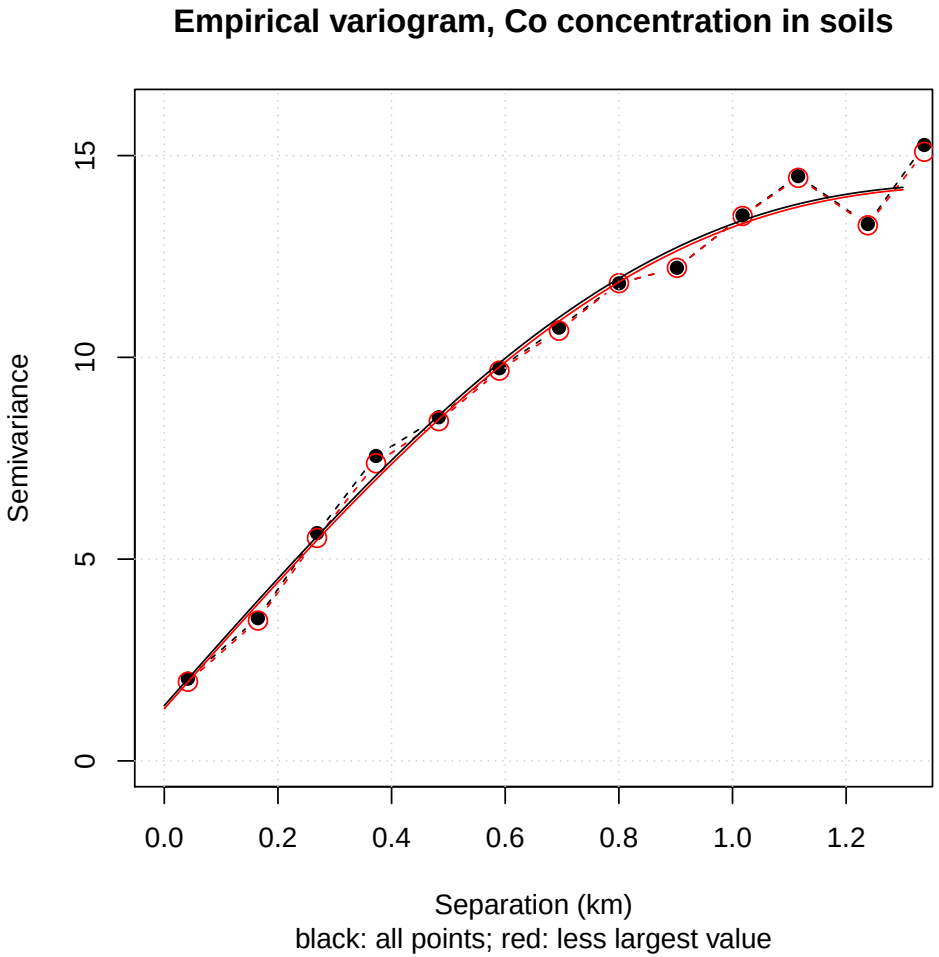
- If we don't have an independent data set to evaluate a model, we can use the **same sample points** that were used to estimate the model to validate that same model.
- For **geostatistical** models, see next section "Cross-validation"
- Non-geostatisical: Do many times:
  - Randomly split the dataset into calibration and evaluation parts.
  - Build the model using only the calibration part
  - Evaluate it against the evaluation part as in "independent evaluation", above

  Then, summarize the evaluation statistics.
- Build a final model using all the observations; but report the evaluation statistics from resampling.

Cross-validation

- For **geostatistical** models, if we don't have an independent data set to evaluate a model, we can use the **same sample points** that were used to estimate the model to validate that same model.
- With enough points, the effect of the removed point on the **model** (which was estimated using that point) is minor.

Effect of removing an observation on the variogram model



hardly any effect – both empirical variogram and fitted models are nearly identical

Cross-validation procedure

1. Compute experimental variogram with all sample points in the normal way; model it to get a parameterized variogram model
2. For each sample point
   1. **Remove the point** from the sample set;
   2. predict **at that point** using the **other points** and the modelled variogram;
3. This is called **leave-one-out cross-validation** (LOOCV).
4. Summarize the deviations of the model from the actual point.

Summary statistics for cross-validation (1)

Two are the same as for independent evaluation and are computed in the same way:

- **Root Mean Square Error** (RMSE): lower is better
- **Bias** or mean error (MPE): should be 0

Summary statistics for cross-validation (2)

Since we have variability of the cross-validation, and variability of each prediction (i.e. kriging variance), we can compare these:

- **Mean Squared Deviation Ratio** (MSDR) of residuals with kriging variance

$$\text{MSDR} = \frac{1}{n}\sum_{i=1}^{n}\frac{\{z(\mathbf{x}_i) - \hat{z}(\mathbf{x}_i)\}^2}{\hat{\sigma}^2(\mathbf{x}_i)}$$

  where $\hat{\sigma}^2(\mathbf{x}_i)$ is the kriging variance at cross-validation point $\mathbf{x}_i$.
- The MSDR is a measure of the **variability of the cross-validation vs. the variability of the sample set**. This ratio should be 1. If it's higher, the kriging prediction was too optimistic about the variability.
- The **nugget** has a large effect on the MSDR, since it sets a **lower limit** on the kriging variance at all points.

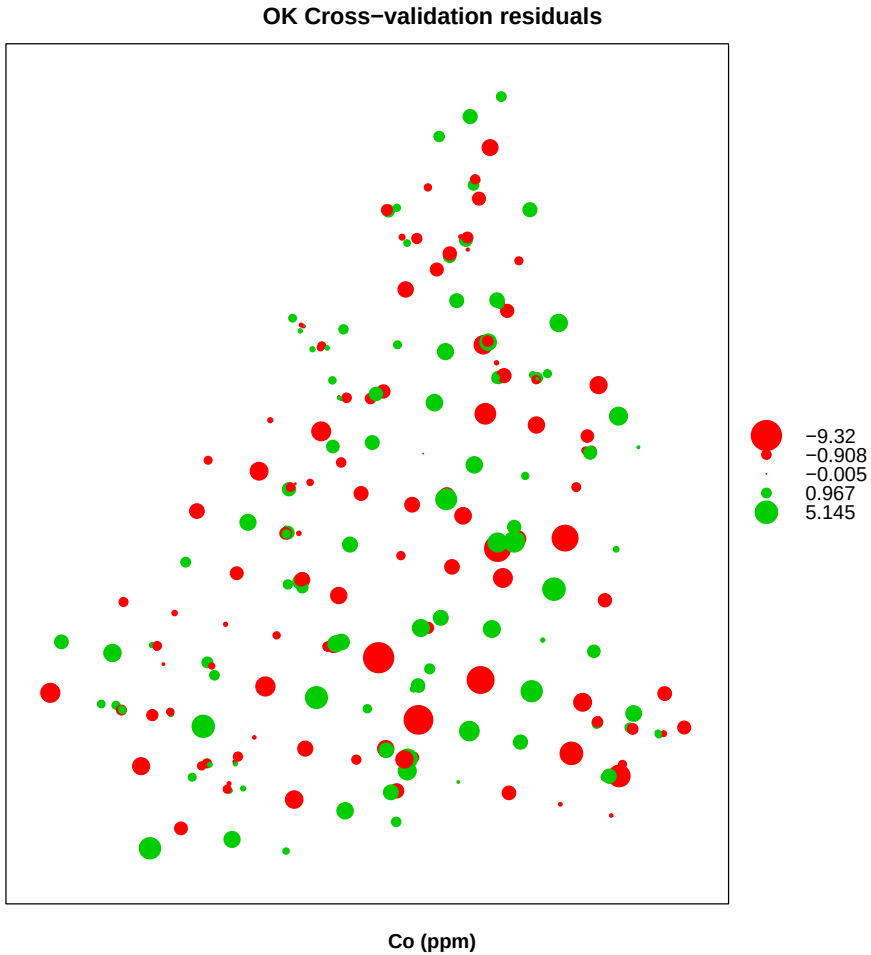
Summary statistics for cross-validation (3)

- Another way to summarize the variability is the **median** of the Squared Deviation Ratio:

$$\text{MeSDR} = \text{median}\left[\frac{\{z(\mathbf{x}_i) - \hat{z}(\mathbf{x}_i)\}^2}{\hat{\sigma}^2(\mathbf{x}_i)}\right]$$

- If a correct model is used for kriging, $\text{MeSDR} = 0.455$, which is the median of the χ^2 distribution (used for the ratio of two variances) with one degree of freedom.
- MeSDR < 0.455 → kriging **overestimates** the variance (possibly because of the effects of outliers on the variogram estimator)
- MeSDR > 0.455 → kriging **underestimates** the variance
- *Reference*: Lark, R.M. 2000. A comparison of some robust estimators of the variogram for use in soil survey. *European Journal of Soil Science* 51(1): 137–157.

Spatial distribution of cross-validation residuals



actual – predicted; green are underpredictions

End