Model Evaluation

D G Rossiter

Cornell University, New York State College of Agriculture & Life Sciences
School of Integrative Plant Sciences, Section of Soil & Crop Sciences
南京师范大学地理学学院

February 23, 2023

1. Assessment of model quality
2. Internal evaluation
   - Empirical statistical models
   - Machine-learning models
   - Kriging models
3. External evaluation
   - Evaluation measures
   - Model efficiency coefficient
   - Regression of actual on predicted
   - Linn's Concordance
   - Confusion matrices
4. Resampling
5. Cross-validation
6. Spatial patterns

Assessment of model quality

- With any **predictive** method, we would like to know how good will be its predictions.
  - i.e., how well is the model expected to perform on **new** information?
- This is model **evaluation**, often called model **validation**.
- Contrast this with model **calibration**, when we are building (fitting) the model.
  - also called model **training**
  - i.e., how well does the model fit the **training** information?

Why the term "evaluation"?

- We 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 and how **useful** it is.
- "Evaluation" has many aspects, not just statistical.
  - Oreskes, N. (1998). Evaluation (not validation) of quantitative models. *Environmental Health Perspectives*, 106(Suppl 6), 1453–1460. DOI: `10.1289/ehp.98106s61453`[1]
  - Oreskes, N., *et al.* (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263, 641–646. DOI: `10.1126/science.263.5147.641`[2]

[1] `https://doi.org/10.1289/ehp.98106s61453`
[2] `https://doi.org/10.1126/science.263.5147.641`

Non-statistical aspects of model quality

- appropriate model form
- reasonable link between model form / assumptions and what is known about the process
- fitness for use
- interpretability, communication with model users
- ...

Internal vs. external evaluation

Internal Only using the calibration/training information and model diagnostics

External Using external information, not used in model calibration/training

Cross-validation Simulating external assessment with the same dataset used for model building

1 Assessment of model quality

2 Internal evaluation

Empirical statistical models

Machine-learning models

Kriging models

3 External evaluation

Evaluation measures

Model efficiency coefficient

Regression of actual on predicted

Linn's Concordance

Confusion matrices

4 Resampling

5 Cross-validation

6 Spatial patterns

Internal evaluation: fit to calibration set

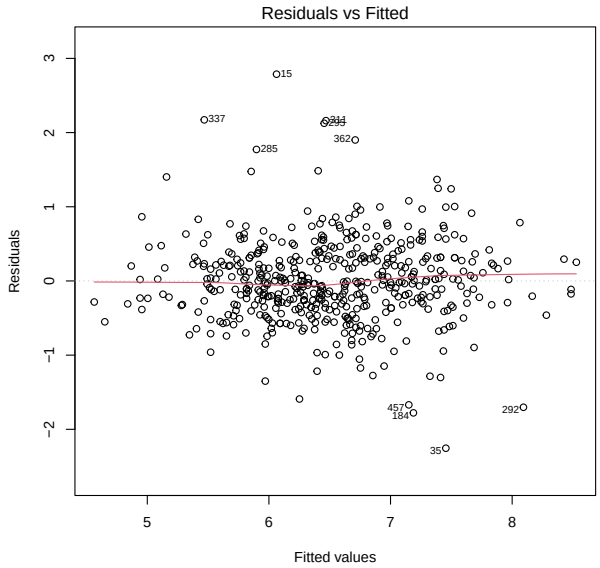
- How well does the **calibrated model** reproduce the **calibration set**?
- This is called the **goodness-of-fit** to the calibration data set.
- e.g., for Linear models: **coefficient of determination** R^2
  - Adding parameters to a model increases its fit; are we fitting **noise** rather than **signal**? Use adjusted measures, e.g. adjusted R^2 or Akaike Information Criterion (AIC)
- 1:1 scatterplot of actual vs. fit, compute
- no protection against **over-fitting**
- no evaluation of the model **form**

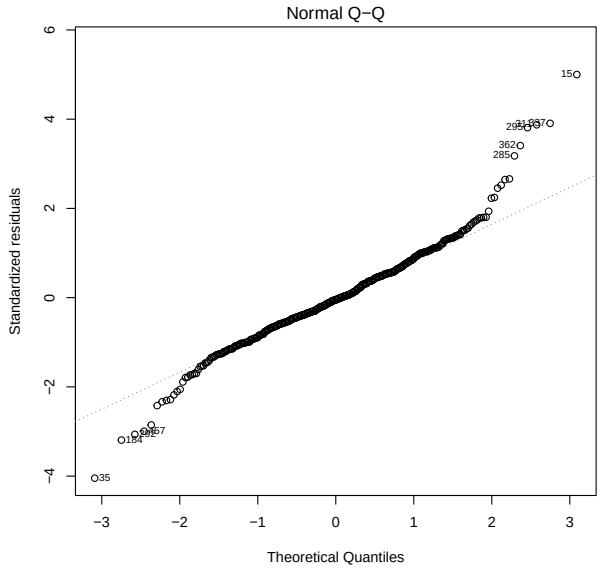
Internal evaluation of empirical-statistical models

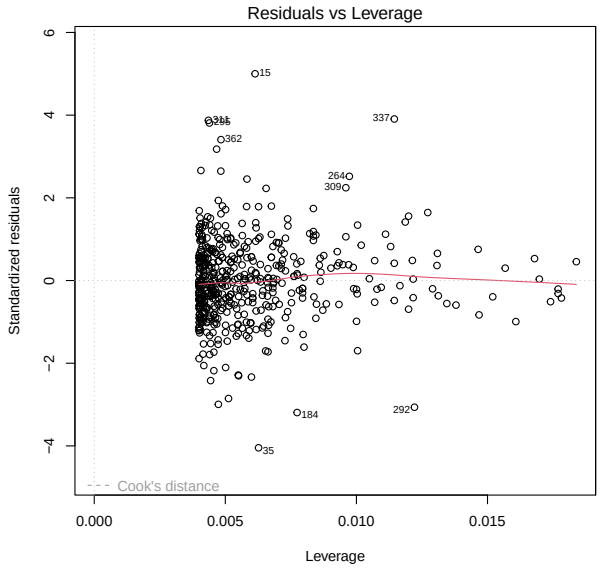
- e.g., (multiple) linear regression
- examine modelling assumptions (**model diagnostics**)[3]
  - homoscedascity, no relation between residuals and fitted values, normally-distributed residuals ...
  - spatial independence of residuals (otherwise, use GLS)
  - variance inflation factors of multiple predictors
- examine **residuals** – does the model fit equally well throughout the range?

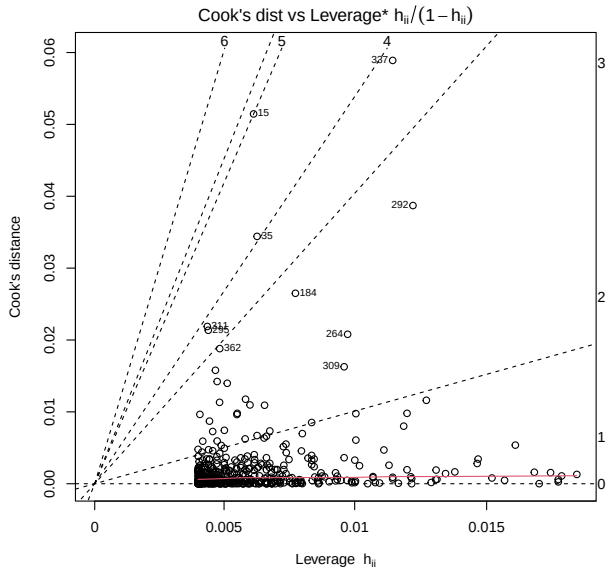
[3] Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression. Chapman and Hall. ISBN 978-0-412-24280-9

Non-spatial linear model diagnostics

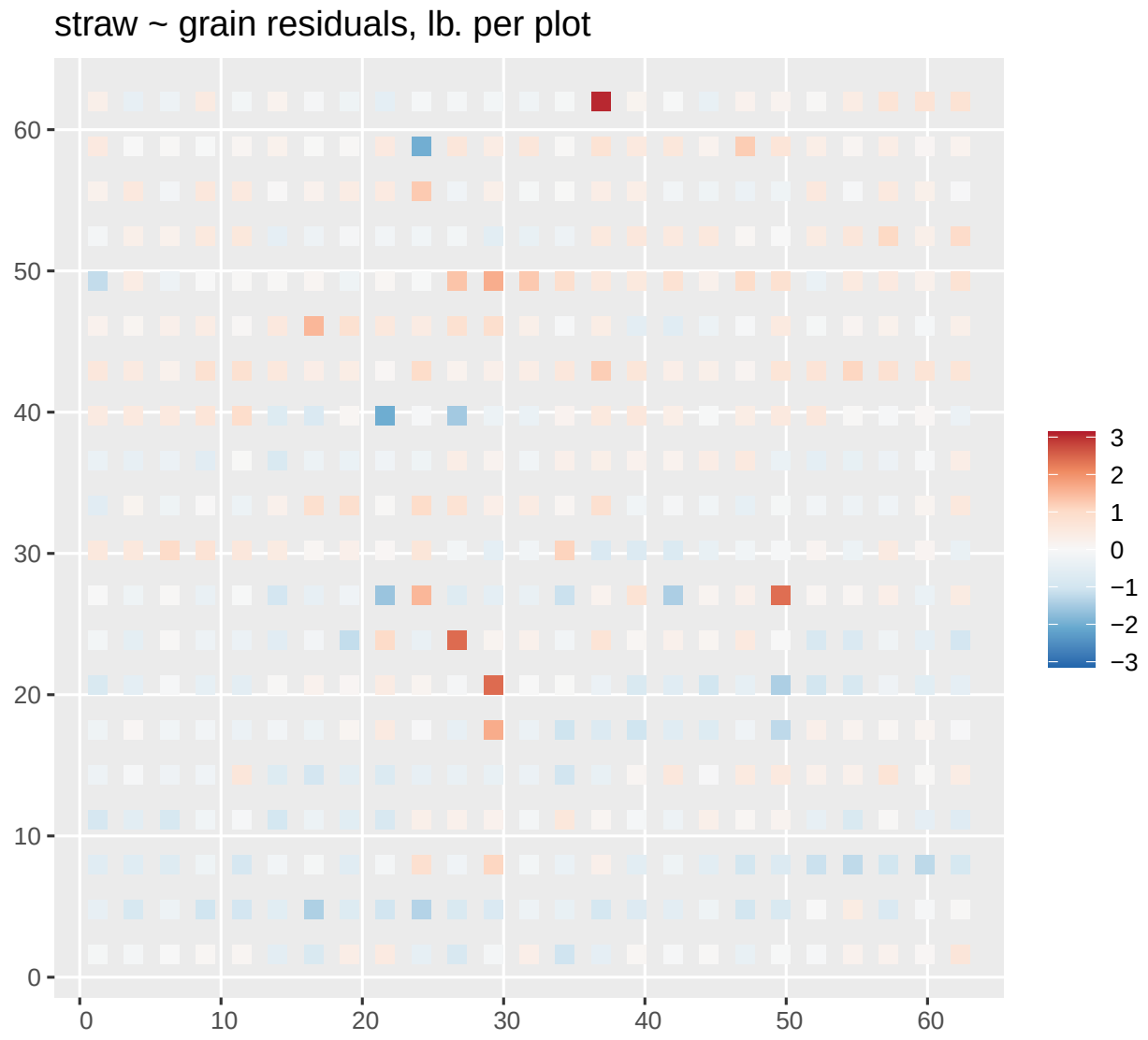




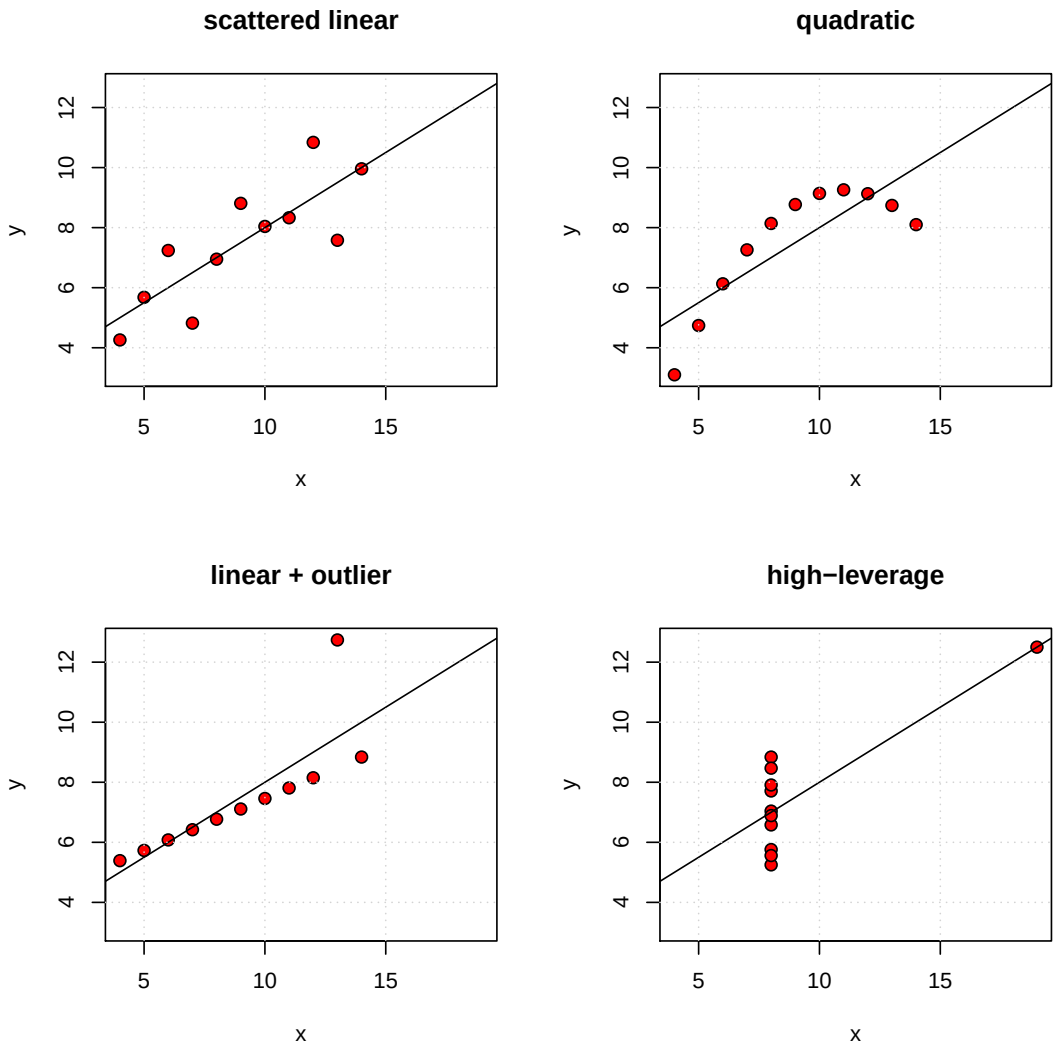




Spatial dependence of linear model residuals

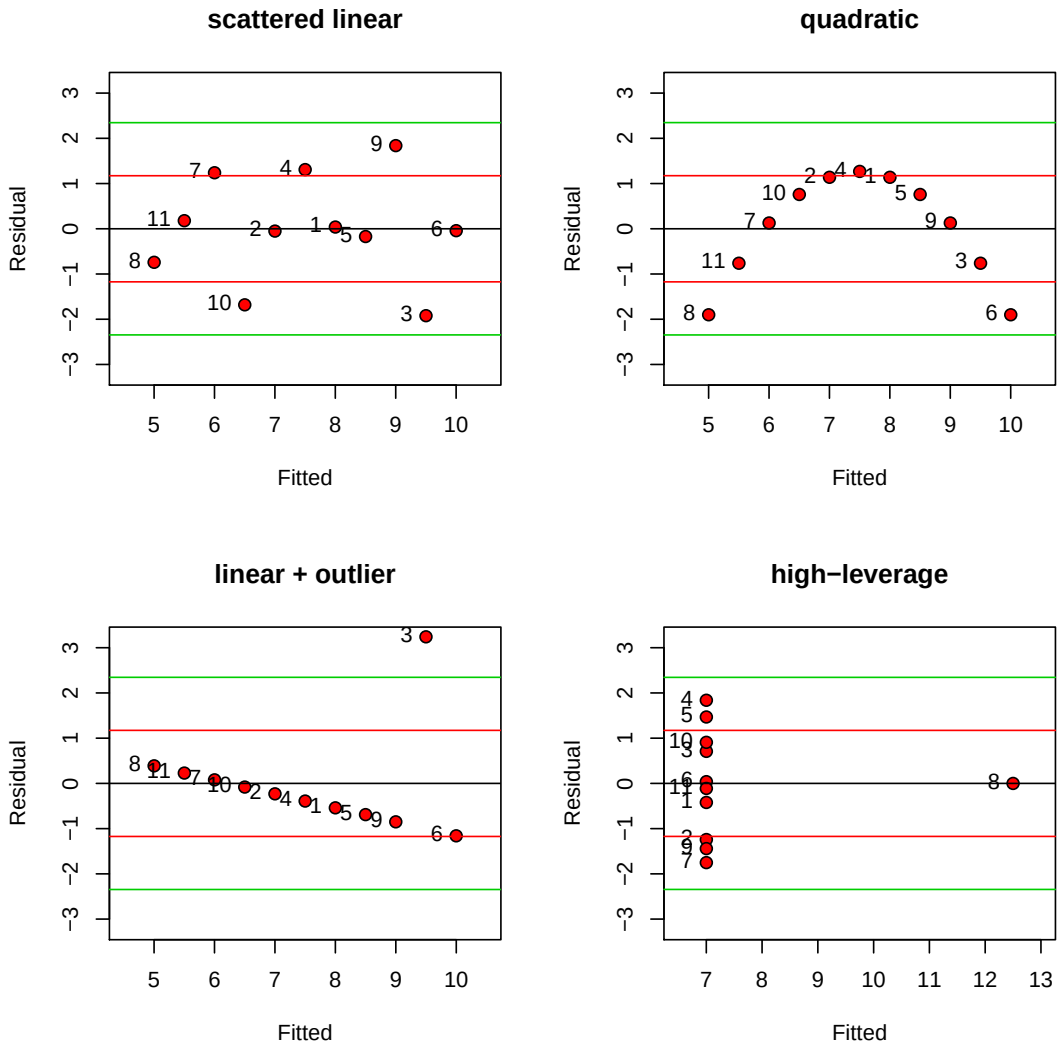


Is the model form justified? (1)



Anscombe, F. J. (1973). Graphs in Statistical Analysis. American Statistician, 27(1), 17–21. https://doi.org/10.2307/2682899

Is the model form justified? (2)



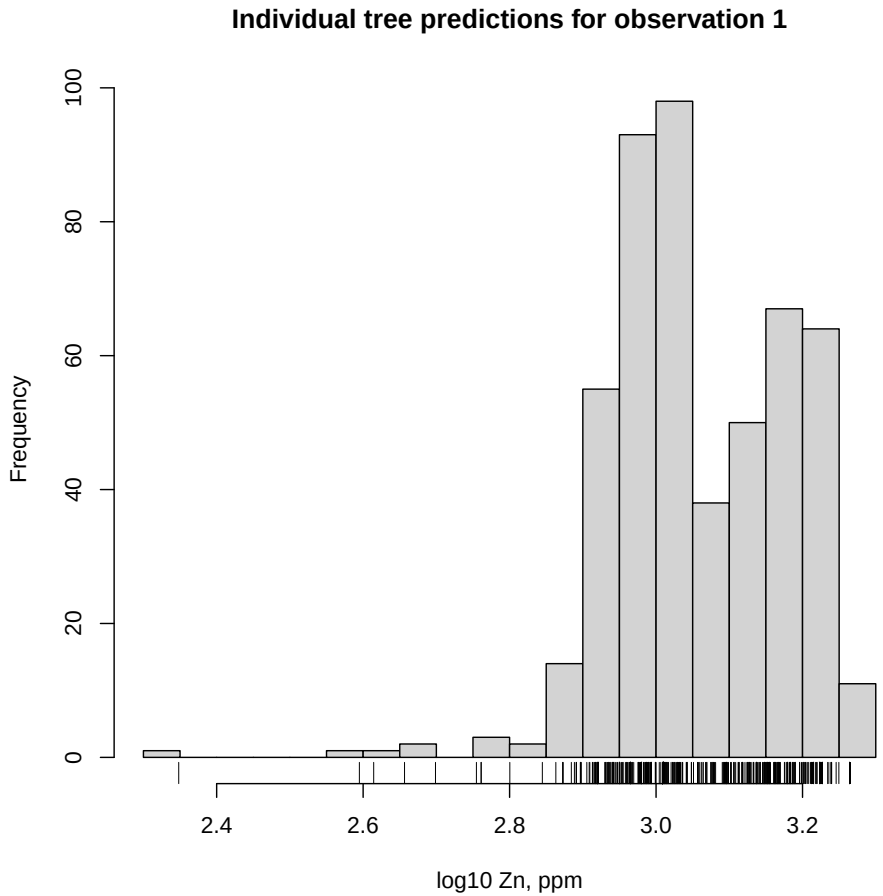
Diagnostic: residual vs. fit

Internal evaluation of machine-learning models

Each tree makes a separate prediction, so examine and summarize the **distribution** of the predictions of individual trees

- **Continuous** predictions (e.g., quantile random forests)
  - summarize by the standard deviation, (non-)normality ...
- **Categorical** (class) prediction (e.g., probability classification trees)
  - summarize by maximum probability, confusion index, Shannon entropy (see below)

Random forest: all trees



Summary statistics:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.348	2.976	3.026	3.057	3.155	3.265

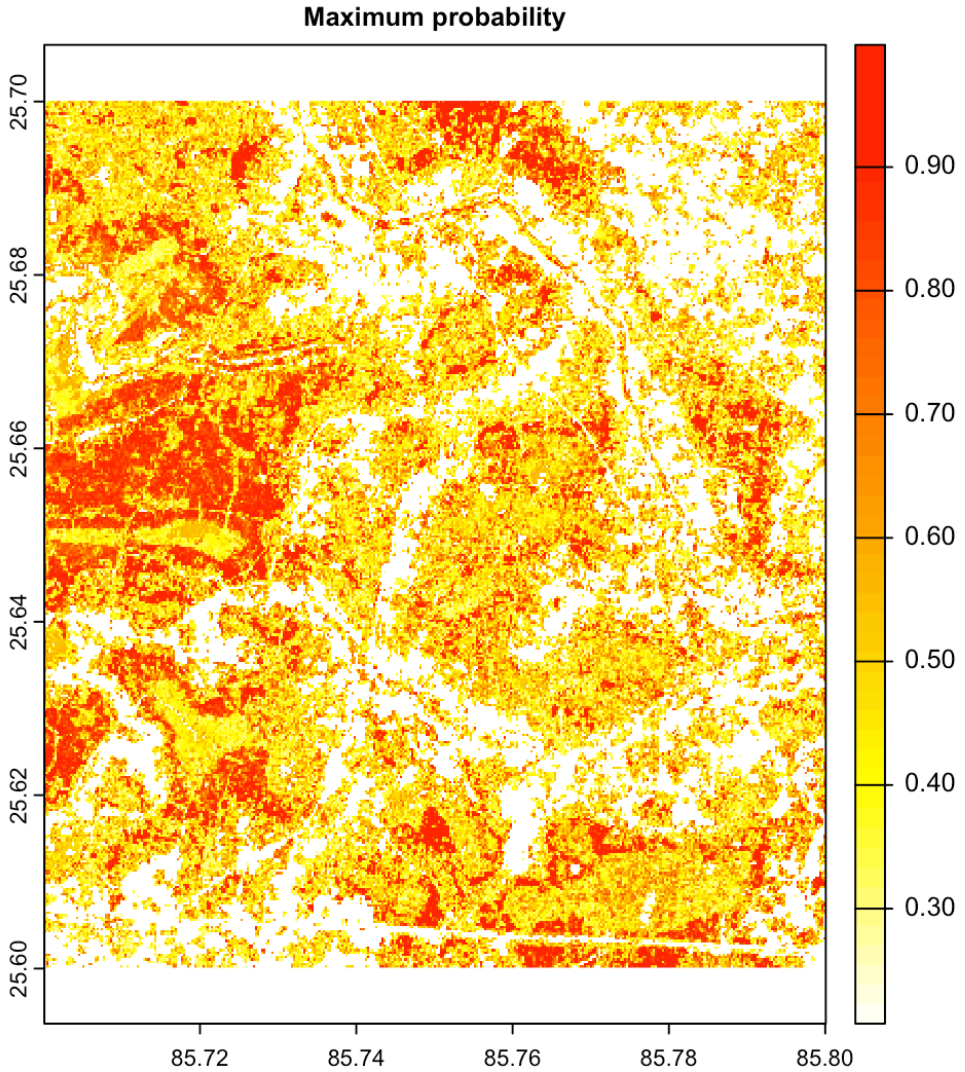
So RF prediction is 3.057.

Internal evaluation of classification models (1)

Maximum probability

- This shows how probable is the majority choice.
- Closer to 1 is better
- If too low, we can **refuse to predict** (the model is too uncertain)

Crop classification from remote sensing using random classification forest



Internal evaluation of classification models (2)

Confusion index

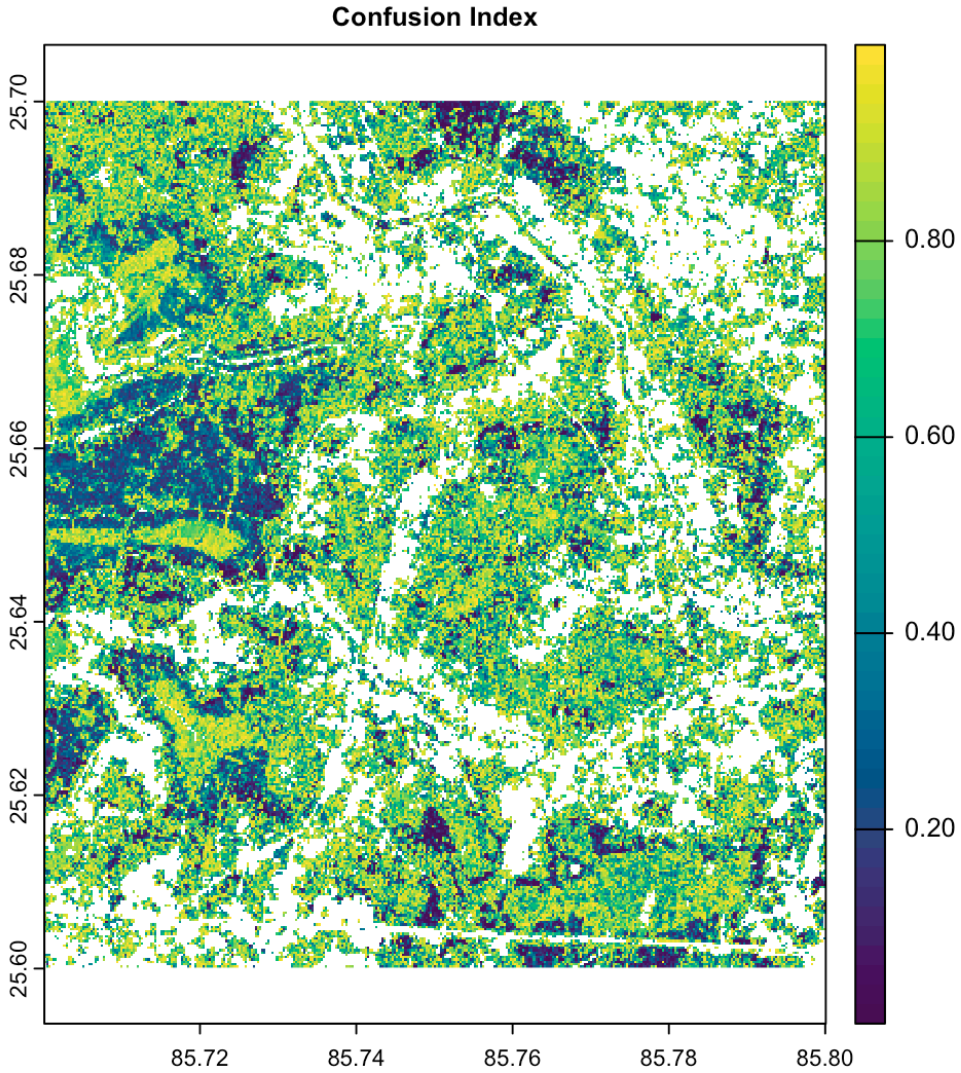
This shows how well the majority choice is separated from the next best choice:[4]

$$\mathrm{CI} = (1 - \{\mu_{\max} - \mu_{(\max-1}\})$$

- $\mu_{\max}$ is the probabilty of the most probable class
- $\mu_{\max-1}$ is the probabilty of the second most probable class.

[4] Burrough, P. A., van Gaans, P. F. M., & Hootsmans, R. (1997). Continuous classification in soil survey: Spatial correlation, confusion and boundaries. Geoderma, 77(2–4), 115–135 https://doi.org/10.1016/S0016-7061(97)00018-9)

Crop classification from remote sensing using random classification forest



Internal evaluation of classification models (3)

Shannon entropy

A measure of overall uncertainty[5]

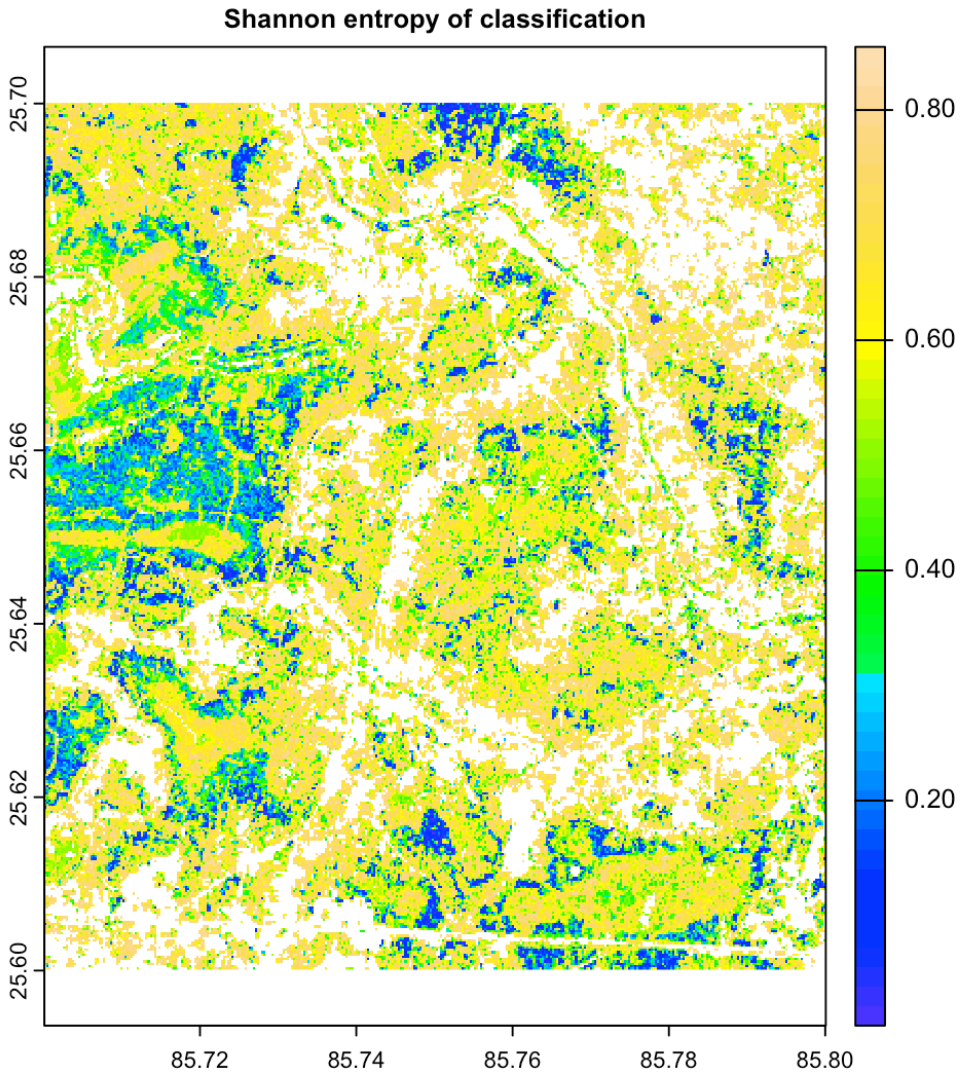
For a variable z with n classes, each of which has estimated proportion $\hat{\pi}(z_i)$:

$$H_z = -\sum_{i=1}^{n} \hat{\pi}(z_i) \cdot log_n \hat{\pi}(z_i)$$

The reason to use base-n logarithms is that 0 represents no uncertainty, and 1 maximum.

[5] Kempen, B., Brus, D. J., Heuvelink, G. B. M., & Stoorvogel, J. J. (2009). Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma, 151(3–4), 311–326. https://doi.org/10.1016/j.geoderma.2009.04.023

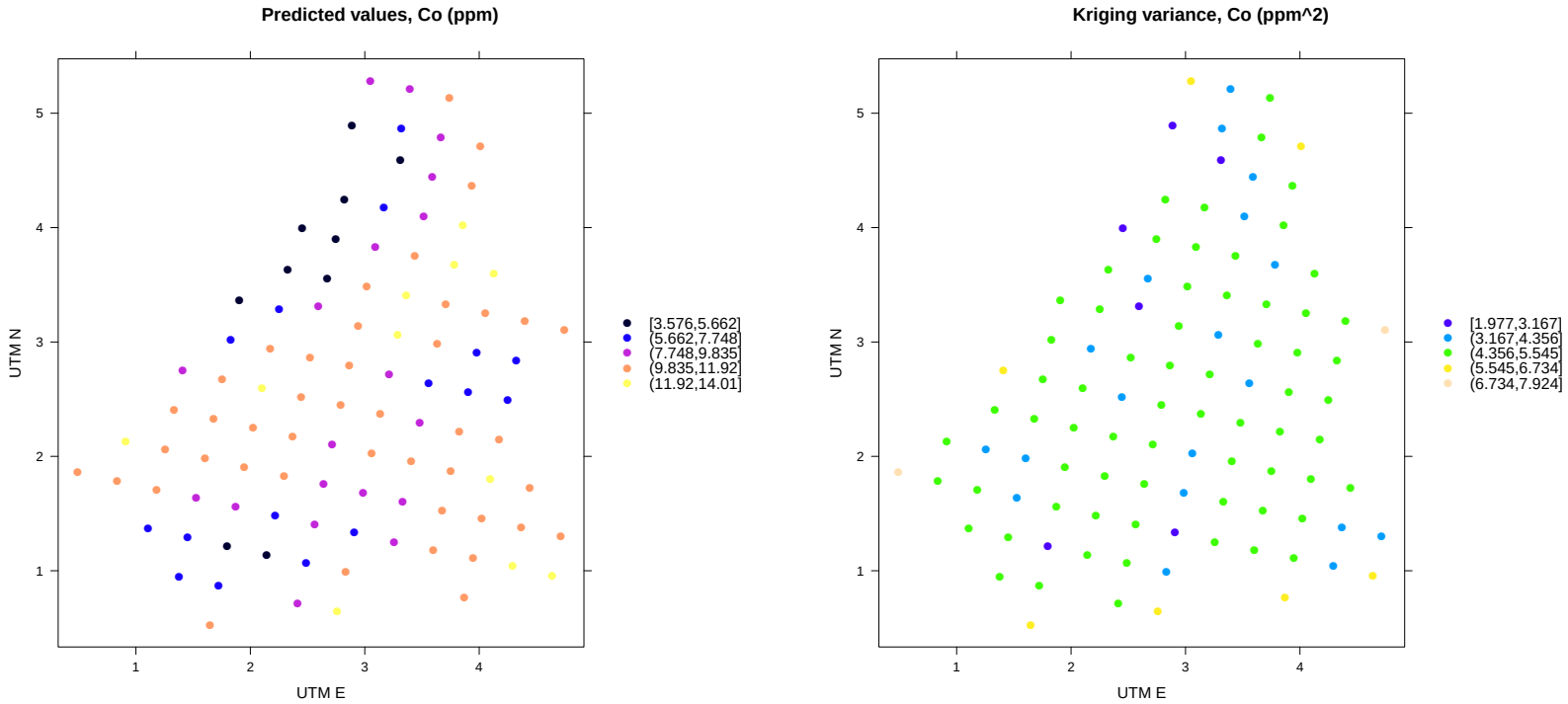
Crop classification from remote sensing using random classification forest



Internal evaluation of Kriging predictions

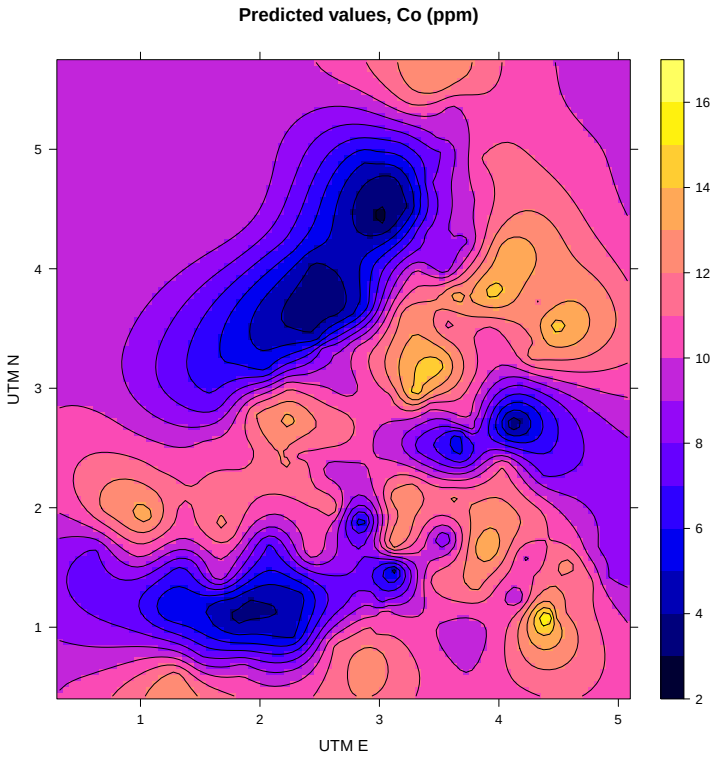
- Because of its model structure, Kriging automatically computes a **kriging prediction variance** to go with each prediction.
- This is because that variance is **minimized** in kriging, *assuming the model of spatial dependence is correct*!
  - Variogram form, variogram parameters
  - OK: Assumptions of 1^{st} and 2^{nd} order stationarity (mean, covariance among point-pairs)
  - KED/UK: Assumptions of 2^{nd} order stationarity (covariance among point-pairs model *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!

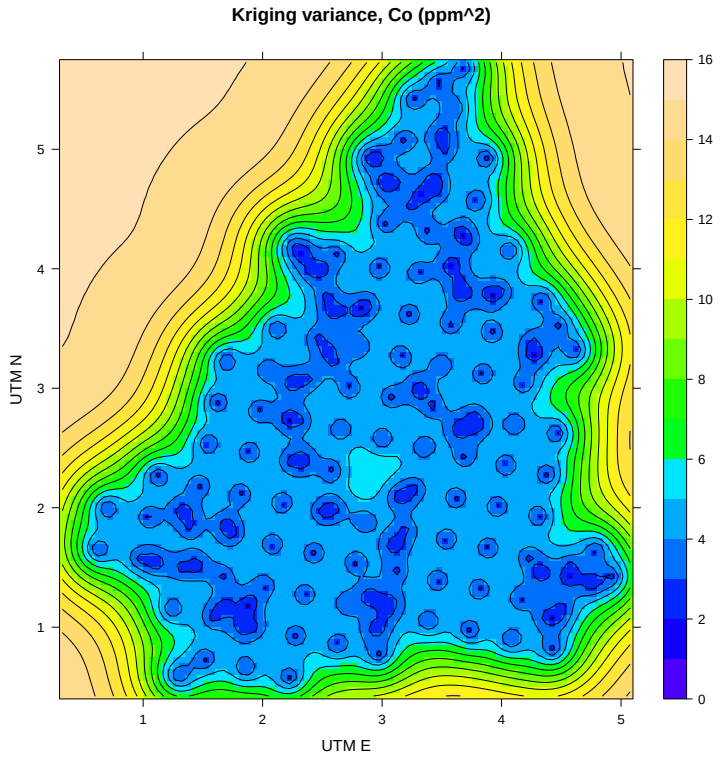
Kriging predictions and variance at points



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

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

1 Assessment of model quality

2 Internal evaluation

Empirical statistical models

Machine-learning models

Kriging models

3 External evaluation

Evaluation measures

Model efficiency coefficient

Regression of actual on predicted

Linn's Concordance

Confusion matrices

4 Resampling

5 Cross-validation

6 Spatial patterns

External evaluation

- This requires an **independent data set** that represents the **target population**
  - thus we can compare **model predictions** with **reality**.
- Two types:
  1. Completely separate **evaluation dataset** from a **target population** to be evaluated
     - specific to this population
  2. **Cross-validation** using the **calibration dataset**, leaving parts out or resampling
     - the calibration dataset must represent the target population

Model evaluation with an independent dataset

Compare model **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
- **Advantages**:
  - objective measure of quality
  - can be applied to a separate population to determine extrapolation power of the model
- **Disadvantages**:
  - Higher cost
  - Poorer model? Not all observations can be used for modelling.

Selecting the evaluation data set

- It must be a **representative** and **unbiased** sample of the **population** for which we want these statistics.
- Two methods:
  1. **Completely independent**
     - 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
     - i.e., *biased* calibration sample → *biased* evaluation statistics!
     - **Cross-validation** (see below) is often preferable

Probability vs. non-probability sample sets

- For evaluation statistics to represent the target population, the independent dataset must be from a **probability sample**[6]
- Otherwise, the statistics refer to a non-probability sample and **no inferences about the population** can be derived

[6] e.g., Brus, D. J., Kempen, B., & Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. European Journal of Soil Science, 62, 394–407. https://doi.org/10.1111/j.1365-2389.2011.01364.x

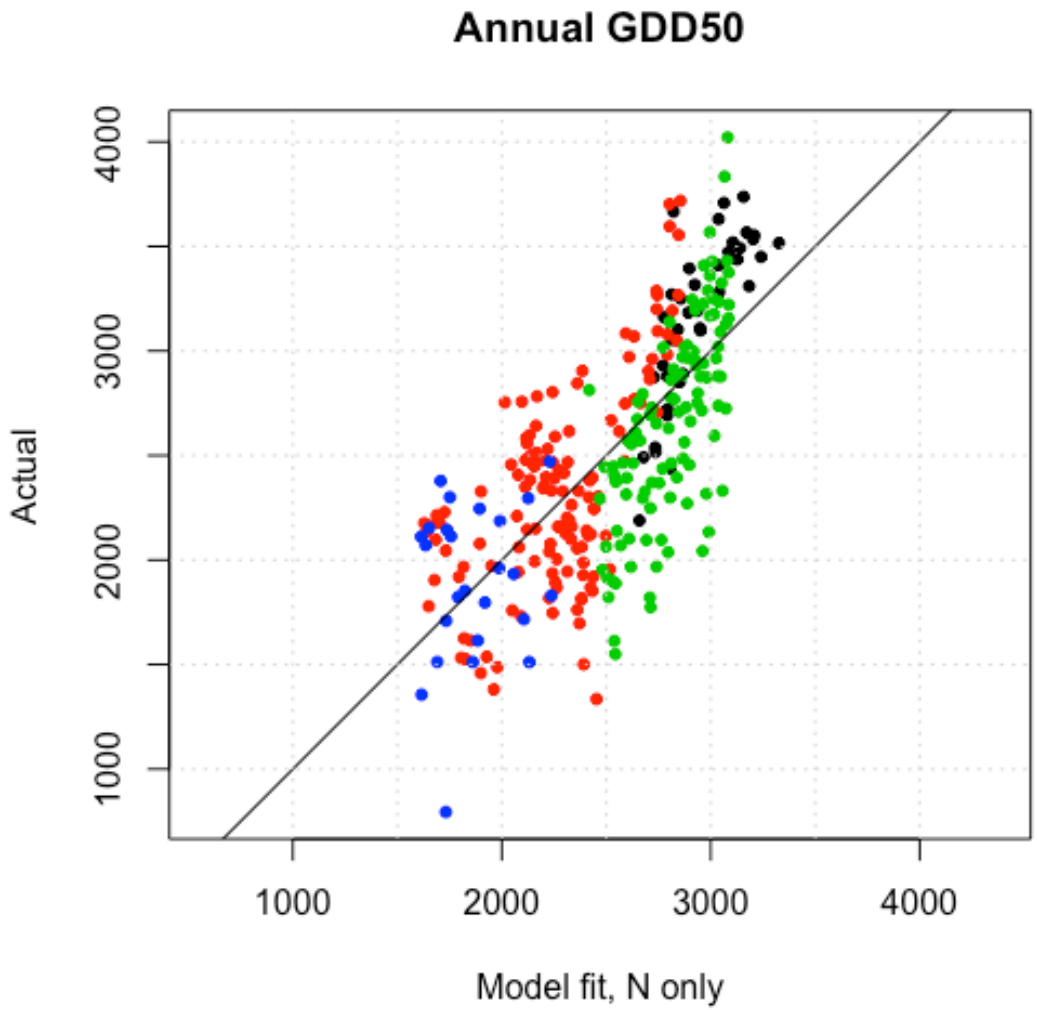
Evaluation measures: residuals

- The fundamental measure: (actual - predicted = **residual**)
- For each evaluation observation *i*:

$$r_i \quad = \quad (\hat{y}_i - y_i)$$

  where: $\hat{y}_i$ is a prediction; y_i is an actual (measured) value
- The entire **distribution** of the residuals can also be examined (max, min, median, quantiles) to make a statement about the model quality
- **unusual** individual residuals can be identified and examined – in which cases does the model fail badly?

Actual vs. predicted



Note 1:1 line: residual is the vertical distance to this line.

Evaluation measures: ME

- Summarize the individual residuals into a composite measure: what is the **average** deviation of predictions from reality?
- This is the **mean absolute error** (ME) of estimated vs. actual mean of the **evaluation** dataset

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

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}(r_i)$$

- closer to zero (0) is better
- Positive and negative residuals cancel each other
- This is the **prediction bias**

Evaluation measures: RMSE

- Summarize the individual residuals into a composite measure: how close **on average** are the predictions to reality?

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

$$\text{RMSE} = \left[\frac{1}{n} \sum_{i=1}^{n} r_i^2 \right]^{1/2}$$

- **lower is better**
- Positive and negative residuals are equally incorrect
- This is an estimate of the **average prediction error**

Relative evaluation measures

- The ME 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 *evaluation dataset values*:
  - ME 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[7], 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.

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

Model efficiency coefficient (MEC)

- This shows how much of the variability in the target population is **explained** by the model.

$$1 - (RMSE/SD)$$

  - RMSE: square root of the mean-square of the residuals
  - SD: standard deviation of the dataset
- $\mathrm{MEC} = 1$: RMSE is 0
- $\mathrm{MEC} = 0$: model explains nothing, because RMSE $=$ SD.
- For a target dataset with low vs. high variability, the same RMSE represents a less successful model.
- Sometimes called the Nash-Sutcliffe MEC, which refers to the **residual sum of squares** of model residuals vs. **total sum of squares** of the dataset:

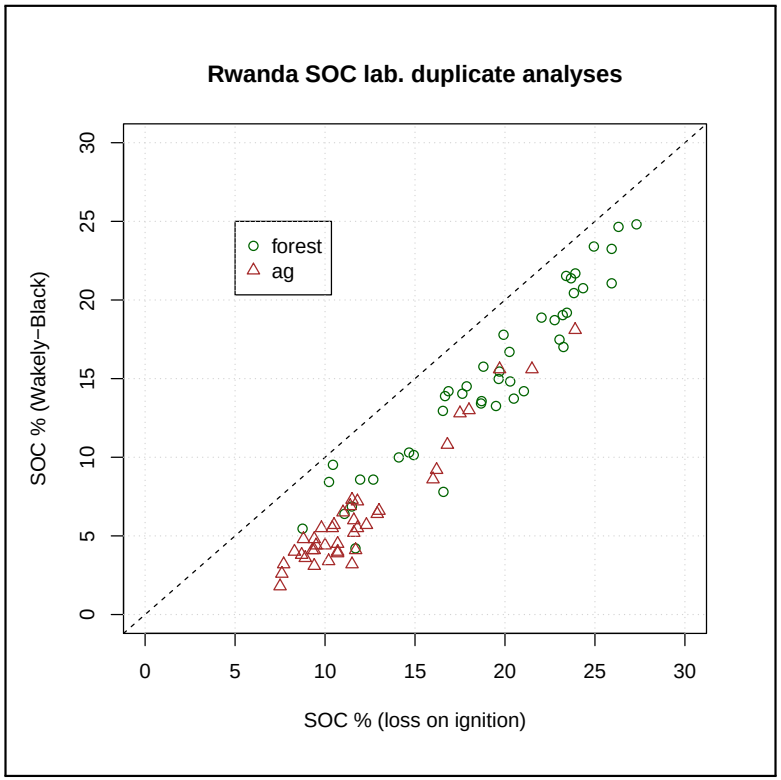
$$1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Regression of actual on predicted

- We can also compute a **linear regression of actual on predicted** values
  - $y = \beta_0 + \beta_1 \hat{y}$
- 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.

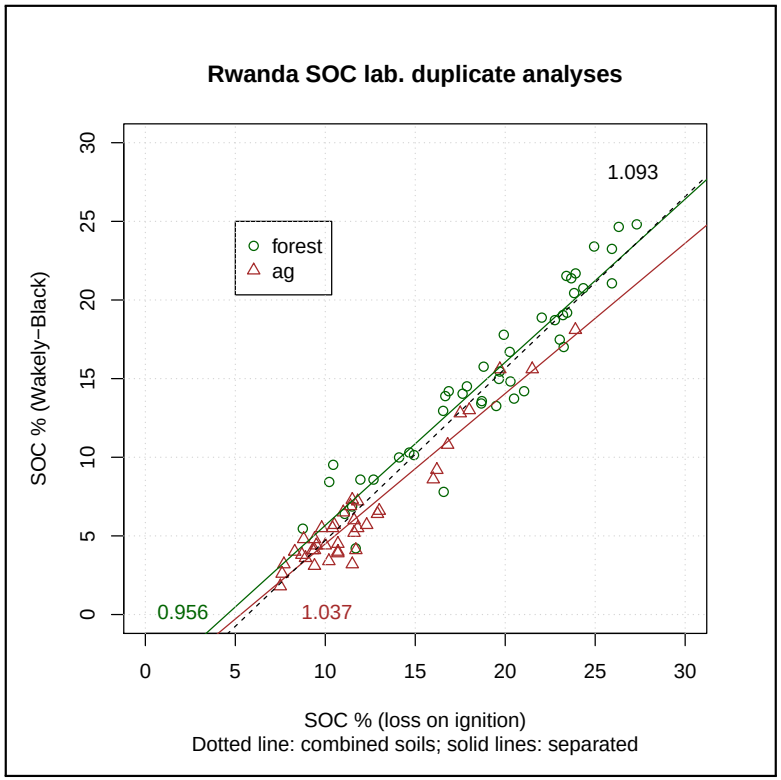
Visualizing actual vs. predicted

Scatterplot against 1:1 line



ME, RMSE

Regression



gain, bias

Lin's Concordance (CCC) (1)

- 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[8]
  - also valid to compare **actual vs. predicted** by any model, these are supposed to be the same

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

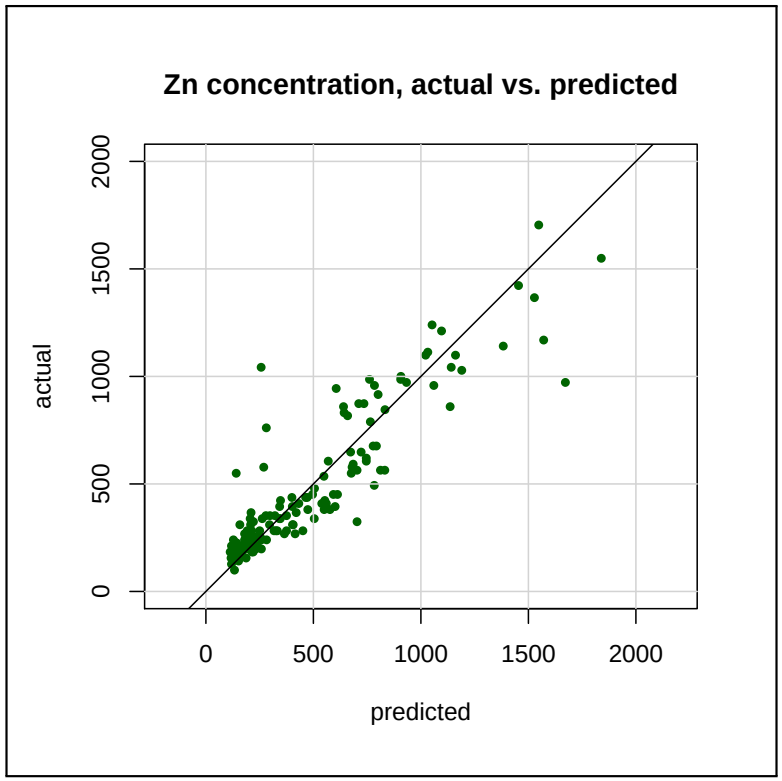
- σ_1 is the standard deviation of predictions (or "model 1")
- σ_2 is the standard deviation of actual values (or "model 2")
- $rho_{1,2}$ is the Pearson's (linear) correlation coefficient between the predictions and actual (or the two model predictions)
- μ_1, μ_2 are the two means

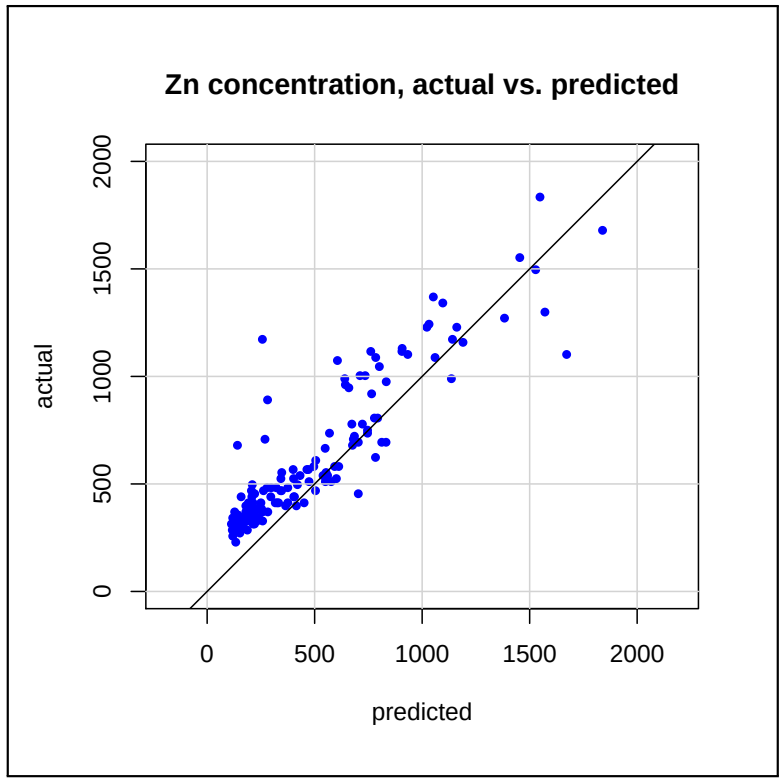
[8] Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45(1), 255–268. https://www.jstor.org/stable/2532051

Lin's Concordance (CCC) (2)

- The CCC includes all sources of deviation from a perfect model:
  - 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 evaluation points are from a *probability sample*, can use the *sample* estimates $r_{1,2}, S_1, S_2, \overline{Y_1}, \overline{Y_2}$ in place of the *population* statistics
- if not, CCC only refers to the evaluation set and not the population

Lin's Concordance – examples





Concordance: 0.900 (no bias)

Note same spread of values

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

External evaluation of classification models

- Compare **predicted** and **actual** classes in an evaluation dataset
  - Should be a **probability sample** from an independent dataset
- Organize as a **Confusion** (cross-classification) **matrix**
  - Rows: **predicted** class
  - Columns: **actual** class
- Shows **overall** accuracy and **per-class** accuracy/reliability
- Information for map **user** (how useful is the map?) and **producer** (how good was the mapping procedure?).

Example confusion matrix

	Alfisols	Andisols	Aridisols	Entisols	Gelisols	Histosol	Inceptisols	Mollisols	Oxisol	Spodosols	Ultisols	Vertisols
Alfisols	**8193**	29	23	1060	7	103	1068	1822	31	81	976	273
Andisols	41	**416**	1	24	1	3	92	30	28	21	24	0
Aridisols	37	15	**1206**	414	0	1	35	145	0	1	1	33
Entisols	124	15	144	**1958**	71	36	430	239	56	71	102	96
Gelisols	2	13	1	92	**1317**	43	180	7	0	95	0	2
Histosol	3	0	1	20	2	**313**	15	1	0	12	7	1
Inceptisols	158	106	20	886	222	140	**2802**	457	11	358	142	52
Mollisols	1223	102	335	1208	70	93	1089	**9265**	0	66	34	347
Oxisol	169	23	4	638	0	23	312	141	**4056**	20	793	61
Spodosols	73	49	0	154	45	100	295	69	2	**1409**	2	1
Ultisols	857	42	3	618	1	91	464	231	439	65	**3982**	89
Vertisols	19	0	5	39	1	0	20	43	1	0	1	**328**

Source: SoilGrids 250 m V1[9] internal evaluation from WoSIS soil profile dataset[10], ISRIC
Classes are Orders of USDA Soil Taxonomy

[9] Hengl, T., *et al.* SoilGrids250m: Global gridded soil information based on machine learning. PLOS ONE, 12(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

[10] https://www.isric.org/explore/wosis

Notation I

Symbol	Meaning	Computation
$\mathbf{X}$	confusion matrix → rows $[1 \dots r]$ are *classified* ("mapped") data → columns $[1 \dots r]$ are *reference* ("true") data	
r	number of rows and columns of $\mathbf{X}$	
x_{ij}	number of observations in row i, column j, i.e. in reference class j but mapped as class i	as observed
x_{i+}	marginal sum of row (mapped class) i	$\sum_{j=1}^{r} x_{ij}$
x_{+j}	marginal sum of column (reference class) j	$\sum_{i=1}^{r} x_{ij}$
n	total number of observations	$\sum_{i=1}^{r} \sum_{j=1}^{r} x_{ij}$
	...or	$\sum_{i=1}^{r} x_{i+}$
	...or	$\sum_{j=1}^{r} x_{+j}$

Notation II

Symbol	Statistic	Computation
C_i	User's 'accuracy', mapped class i	x_{ii}/x_{i+}
	...or	p_{ii}/p_{i+}
$\overline{C_i}$	Errors of commission, mapped class i	$1 - C_i$
O_j	Producer's 'reliability', reference class j	x_{jj}/x_{+j}
	...or	p_{jj}/p_{+j}
$\overline{O_j}$	Errors of omission, reference class j	$1 - O_j$
A_o	Overall accuracy	$\sum_{i=1}^{r} x_{ii}/n$
	...or	$\sum_{i=1}^{r} p_{ii}$
$\overline{A_o}$	Overall error	$1 - A_o$

Brus[11] uses the term "map unit purity" for C_i and "class representation" for O_j.

[11] e.g., Brus, D. J., Kempen, B., & Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. European Journal of Soil Science, 62, 394–407. https://doi.org/10.1111/j.1365-2389.2011.01364.x

Example confusion matrix – with evaluation statistics

								Actual								
			Alfisols	Andisols	Aridisols	Entisols	Gelisols	Histosol	Inceptisols	Mollisols	Oxisol	Spodosols	Ultisols	Vertisols			
		Alfisols	**8193**	29	23	1060	7	103	1068	1822	31	81	976	273	60.0%		
		Andisols	41	**416**	1	24	1	3	92	30	28	21	24	0	61.1%		
		Aridisols	37	15	**1206**	414	0	1	35	145	0	1	1	33	63.9%		
		Entisols	124	15	144	**1958**	71	36	430	239	56	71	102	96	58.6%		
Predicted		Gelisols	2	13	1	92	**1317**	43	180	7	0	95	0	2	75.2%		**User's**
		Histosol	3	0	1	20	2	**313**	15	1	0	12	7	1	83.5%		**accuracy**
		Inceptisols	158	106	20	886	222	140	**2802**	457	11	358	142	52	52.3%		
		Mollisols	1223	102	335	1208	70	93	1089	**9265**	0	66	34	347	67.0%		
		Oxisol	169	23	4	638	0	23	312	141	**4056**	20	793	61	65.0%		
		Spodosols	73	49	0	154	45	100	295	69	2	**1409**	2	1	64.1%		
		Ultisols	857	42	3	618	1	91	464	231	439	65	**3982**	89	57.9%		
		Vertisols	19	0	5	39	1	0	20	43	1	0	1	**328**	71.8%		
			75.2%	51.4%	69.2%	27.5%	75.8%	33.1%	41.2%	74.4%	87.7%	64.1%	65.7%	25.6%	**62.2%**		
							Producer's reliability										

Q: Which Order has the poorest map unit purity (user's perspective)? Which other Orders are most incorrectly mapped as this Order?

Q: Which Order has the poorest class representation (mapper's perspective)? Which other Orders are most incorrectly predicted to be in this Order?

Resampling

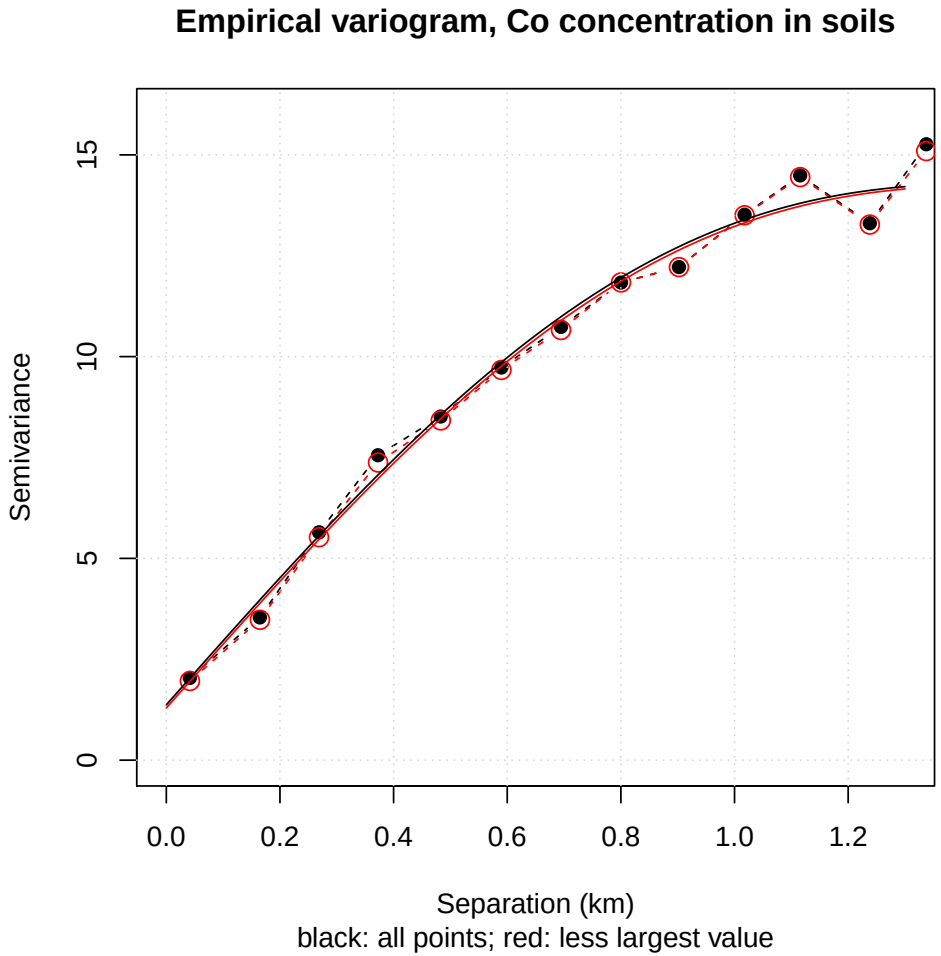
- If we don't have an independent data set to evaluate a model, we can use the **same sample points** that were used to estimate the model to evaluate that same model.
- For **geostatistical** models, see next section "Cross-validation"
- Non-geostatisical: Do many times:
  - Randomly split the dataset 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

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

  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.

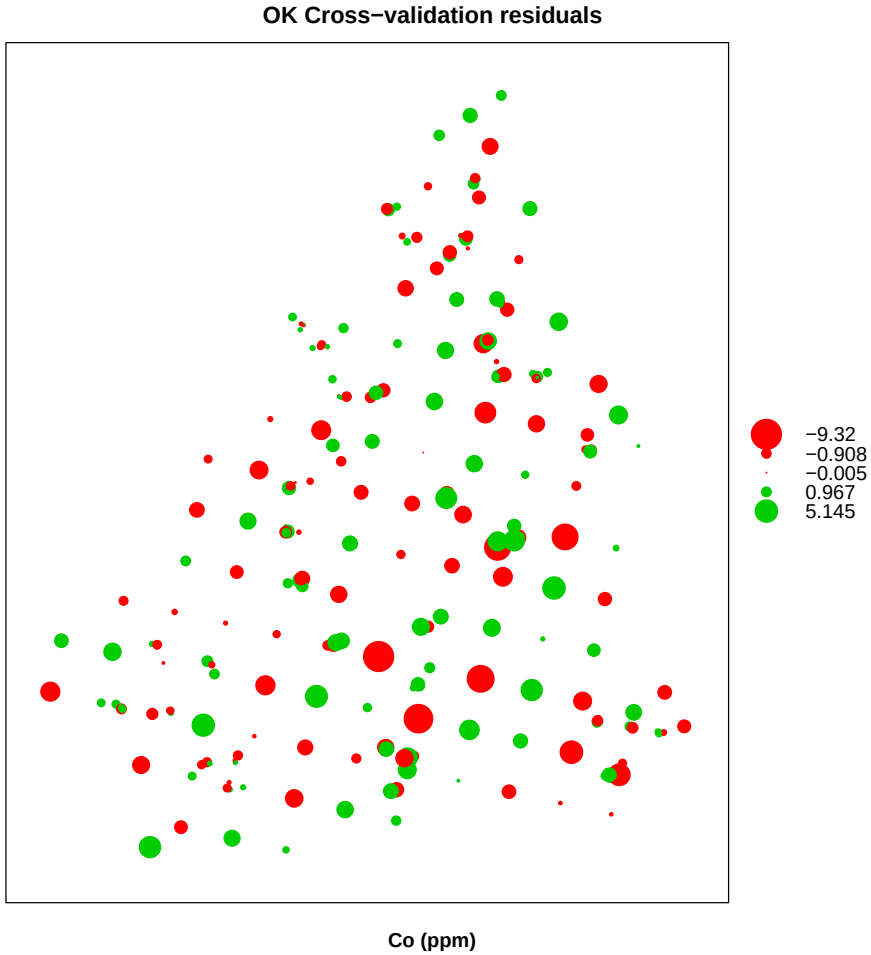
Summary statistics for cross–validation (3)

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

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

- If a correct model is used for kriging, $\mathrm{MeSDR} = 0.455$, which is the median of the χ^2 distribution (used for the ratio of two variances) with one degree of freedom.
- MeSDR $< 0.455 \rightarrow$ kriging **overestimates** the variance (possibly because of the effects of outliers on the variogram estimator)
- MeSDR $> 0.455 \rightarrow$ 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

1 Assessment of model quality

2 Internal evaluation
- Empirical statistical models
- Machine-learning models
- Kriging models

3 External evaluation
- Evaluation measures
- Model efficiency coefficient
- Regression of actual on predicted
- Linn's Concordance
- Confusion matrices

4 Resampling

5 Cross-validation

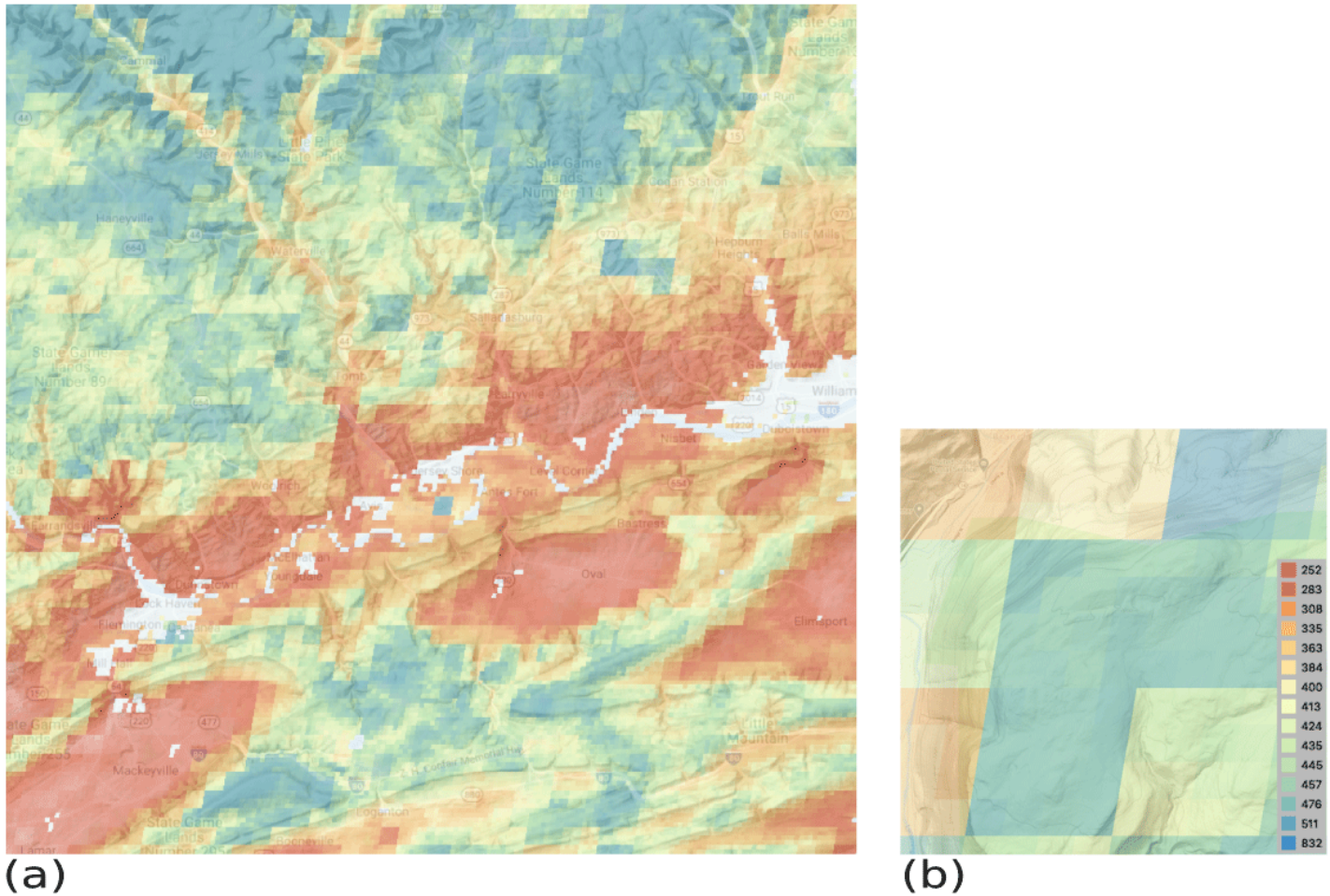
6 Spatial patterns

Spatial patterns

For models that produce a **map**, we also need a **spatial** evaluation

- If the mapping method provides a measure of uncertainty, where are the most uncertain areas?
  - (see above: Maximum Probability, Confusion Index, Shannon Entropy maps)
- what is the **spatial pattern** of the predictions?
  - Do the **patterns** agree with geographical knowledge from other sources?
  - Is the **scale** of the spatial variability realistic?

Spatial prediction pattern

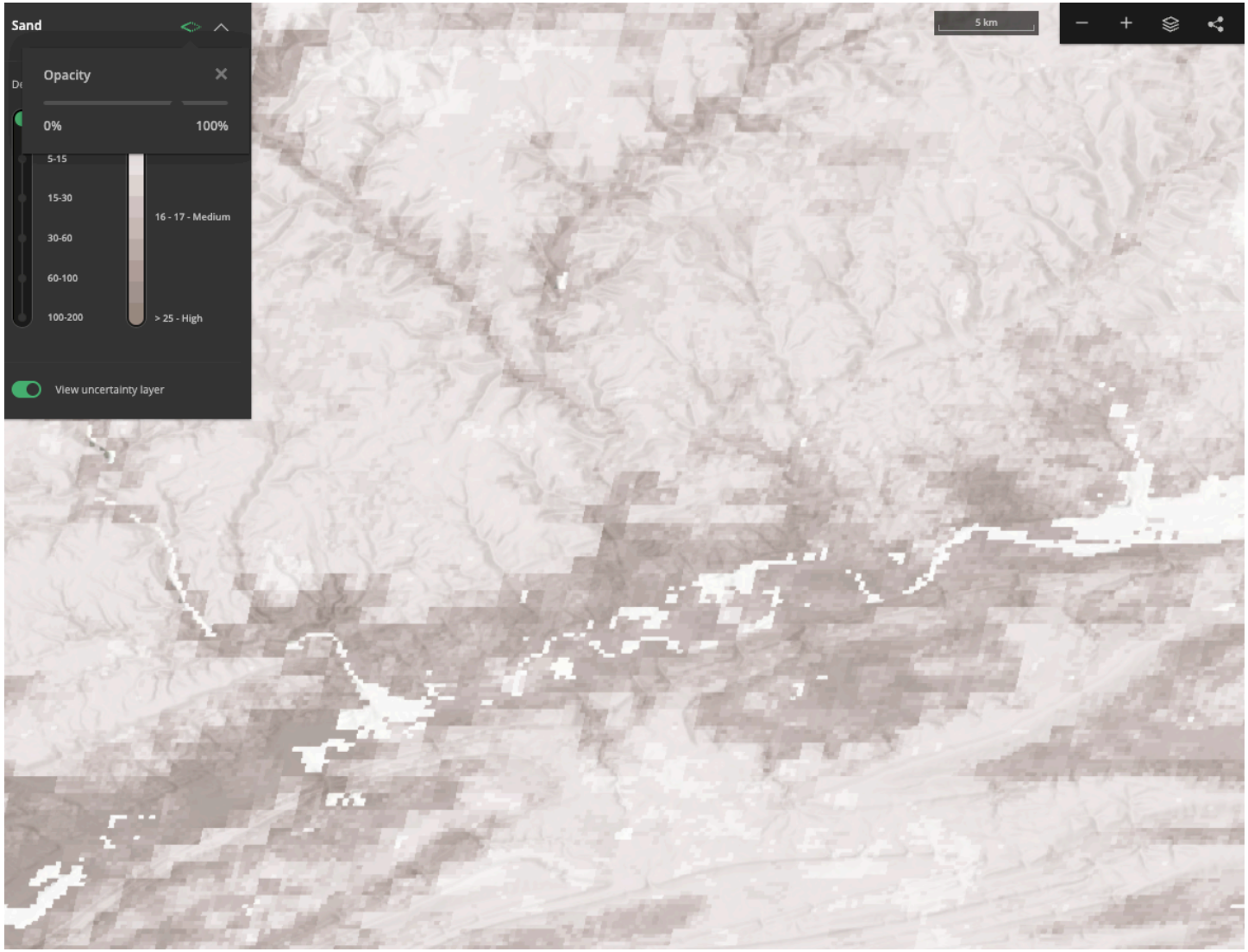


(a) (b)

Predicted sand concentration, %, 0–5 cm, ground overlay in ©Google Earth. (a) Overview; centre ≈ −77°14′ E, 41°14′ N, near Jersey Shore, PA. (b) Detail; centre ≈−76°56′ E, 41°33′ N. Grid cells ≈ 250×250 m

source: `https://dx.doi.org/10.5194/soil-7-217-2021`

Spatial uncertainty pattern



Relative uncertainty, Predicted sand concentration, %, 0–5 cm
source: `soilgrids.org`

End