Applied geostatistics

Lecture 6 – Assessing the quality of spatial predictions

D G Rossiter
University of Twente.
Faculty of Geo-information Science & Earth Observation (ITC)

June 27, 2014

ITC

Topics for this lecture

1. Assessment of model quality: overview
2. Model evaluation with an independent data set
3. Cross-validation
4. Kriging prediction variance
5. Spatial simulation

Topic 1: 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**.

- cf. model **calibration**, when we are building (fitting) the model.

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.

However, we still use the term **cross-validation**, for historical reasons and because the `gstat` function is so named.

Internal vs. external quality assessment

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

1. Completely separate **evaluation dataset**
2. **Cross-validation** using the **calibration dataset**, leaving parts out or resampling

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

- **Linear models**: **coefficient of determination**
  - 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)
- **Generalized linear models**: **residual deviance**
- **Kriging**: the variability of each prediction, i.e. the **kriging prediction variance**

To check your understanding . . .

Q1 : *What is a major* **advantage** *of an external quality assessment?* *Jump to A1* •

Q2 : *What is a major* **disadvantage** *of an external quality assessment?* *Jump to A2* •

Prediction error

The prediction (fitted value) will in general *not* be the same as the observed value.

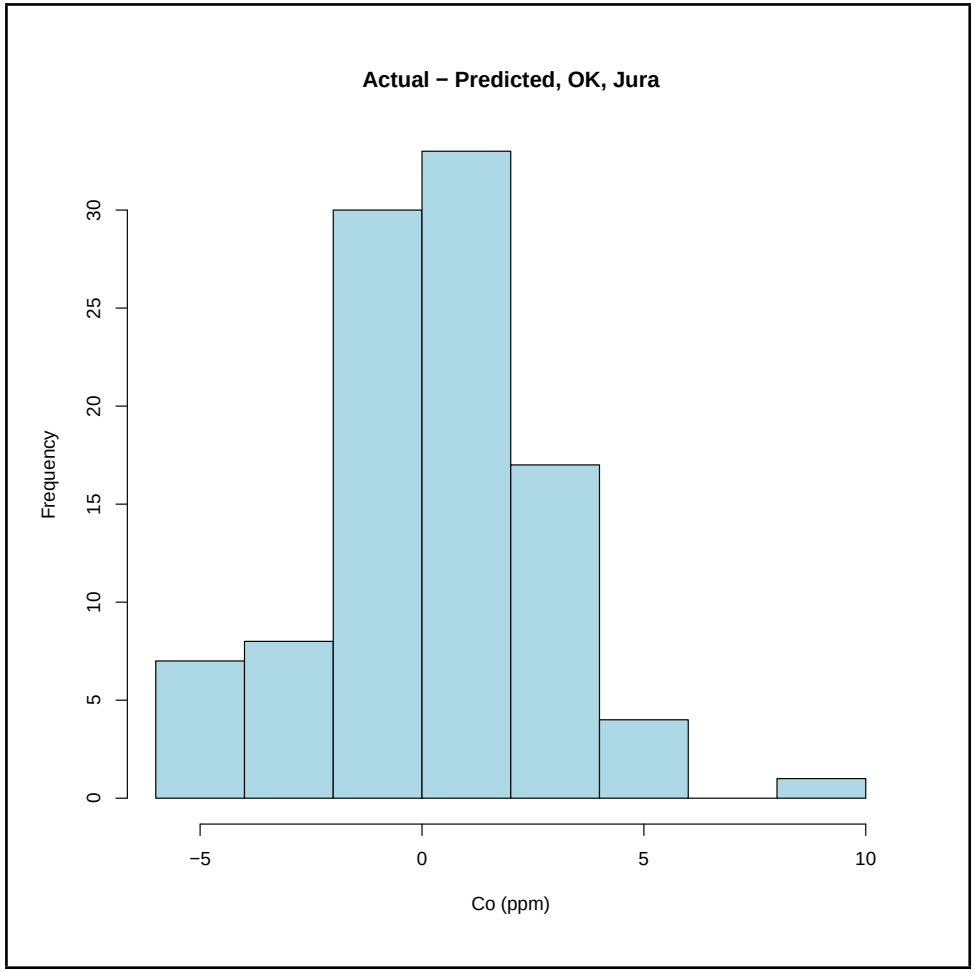
In linear modelling it is usual to define the **residual** e_i for one observation as:

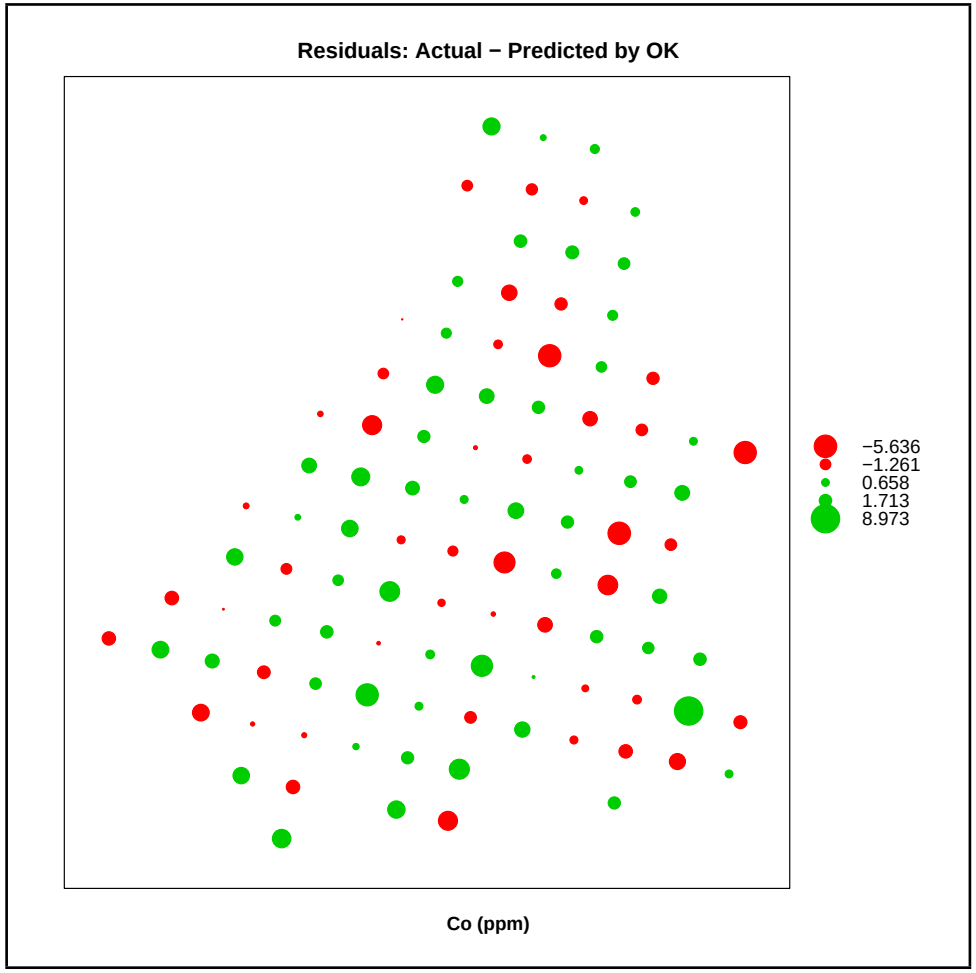
- the **actual** value as measured y_i; less ...
- ...the **estimate** from the model $\hat{y}_i$ (Note the use of the "hat" notation)
- $e_i \equiv y_i - \hat{y}_i$

For **model evaluation** the sign is often inverted, because we want to express this as **prediction error**: how wrong was the prediction? So, **prediction** less **actual**:

- $\hat{e}_i \equiv \hat{y}_i - y_i$

Residuals from evaluation and their location; Jura cobalt





To check your understanding . . .

Q3 : *What are the largest over- and under-estimates? Does the distribution of residuals appear to be normal, as expected by theory?* *Jump to A3* •

Topic 2 : Model evaluation with an independent dataset

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

- **Advantages**: objective measure of quality
- **Disadvantages**: requires more samples; not all samples can be used for modelling (→ poorer calibration?)

Selecting the validation 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

Measures of validity

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

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

  - 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

- **Bias** or mean prediction error (MPE) of estimated vs. actual mean of the **validation** dataset; should be zero (0)

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

Relative measures of validaity

The MPE and RMSE are expressed in the original units of the target variable, as absolute differences.

The magnitude of these can be judged by absolute criteria, but is also relevant to compare them to the dataset itself:

- 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?

To check your understanding . . .

Q4 : *Why, in the RMSE, are the differences between predicted and actual values squared?* *Jump to A4* •

Q5 : *Why then is the square root of the sum taken?* *Jump to A5* •

Model efficiency

Another measure is the **Nash-Sutcliffe model efficiency coefficient**

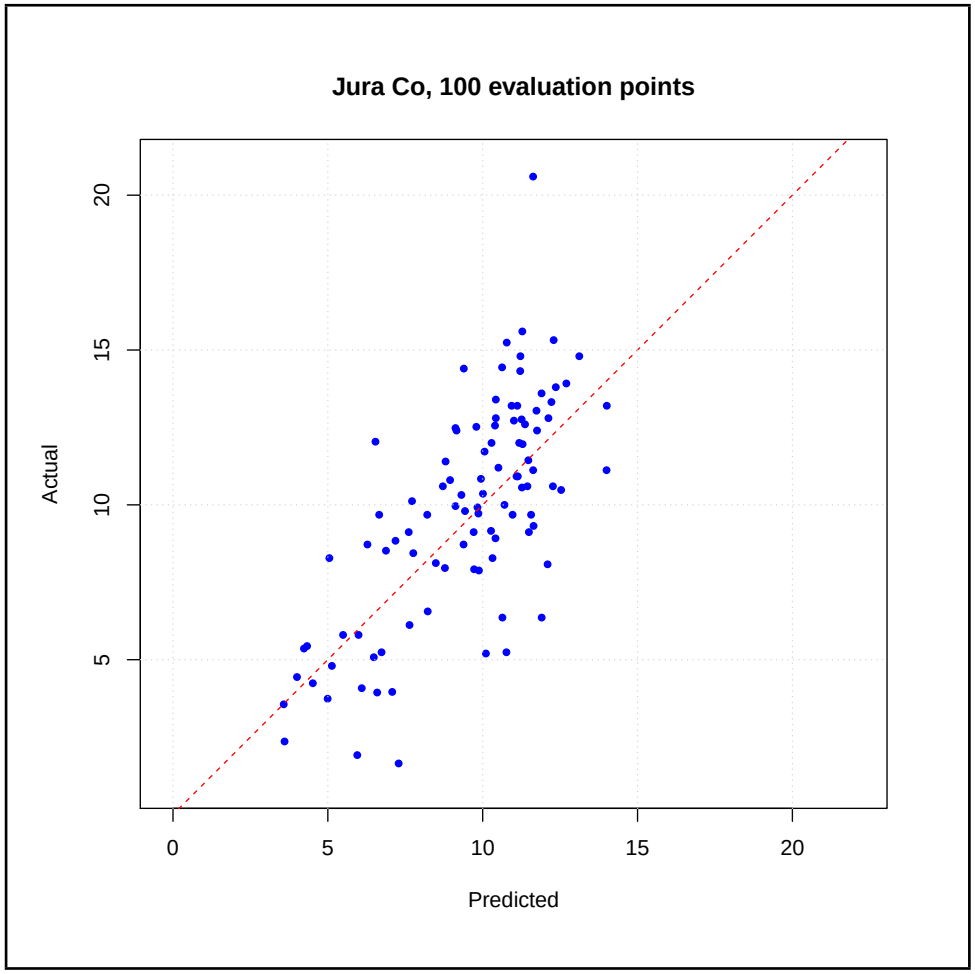
- Proposed (1970) to validate hydrologic models against real-world output
- Can apply to any **actual-vs-predicted**
- **Standardizes** prediction error vs. spread of data
- Equivalent to **coefficient of determination** ("R^2")

$$N = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{SS_{\text{error}}}{SS_{\text{total}}}$$

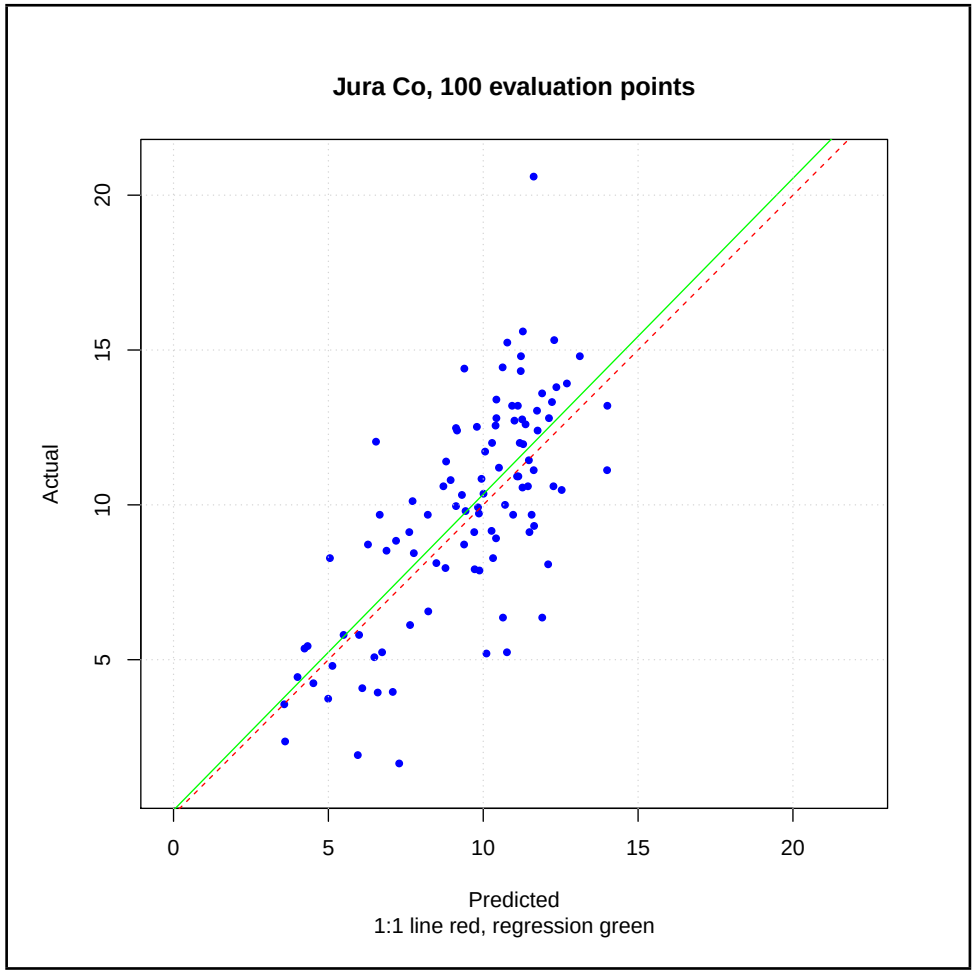
- ranges from $-\omega \ldots 1$
- $N = 0$: model = mean; could have just used the mean $\bar{y}$ of the observations
- $N = 1$: model explains all the variation in the observations (no residuals)

Visualizing actual vs. predicted

Scatterplot against 1:1 line



Regression



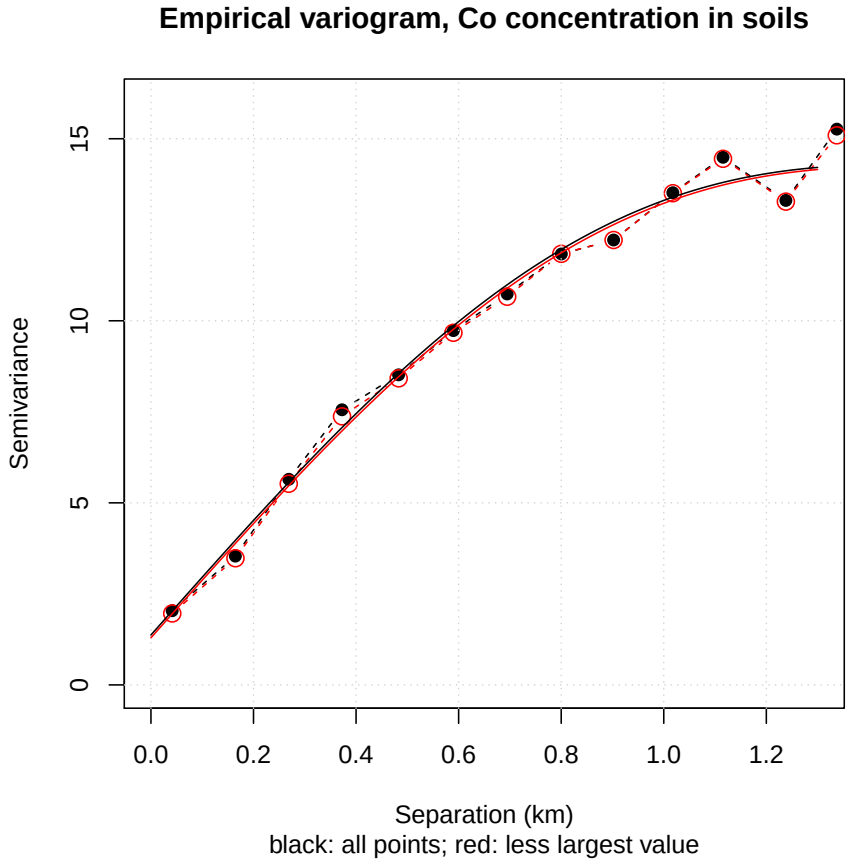
Topic 3: Cross-validation

If we don't have an independent data set to evaluate a model, we can ause the **same sample points** that were used to estimate the model to validate that same model.

This seems a bit dubious, but with enough points, the effect of the removed point on the model (which was estimated using that point) is minor.

Note: This is not legitimate for non-geostatistical models, because there is no theory of spatial correlation.

Effect of removing an observation on the variogram model



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

so it is legitimate to use the variogram fitted from all points in LOOCV

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

   (a) **Remove the point** from the sample set;
   (b) predict **at that point** using the **other points** and the modelled variogram;

3. Summarize the deviations of the model from the actual point.

This is called **leave-one-out cross-validation** (LOOCV).

Then models can be compared by their summary statistics, also by looking at individual predictions of interest.

To check your understanding . . .

Q6 : *What would be the kriging prediction at a sample point, if it were included in the prediction dataset?*

Jump to A6 •

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
  - this is almost guaranteed because kriging is unbiased

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.

Note: the **nugget** has a large effect on the MSDR, since the nugget sets a lower limit on the kriging variance at any point.

Summary statistics for cross–validation (3)

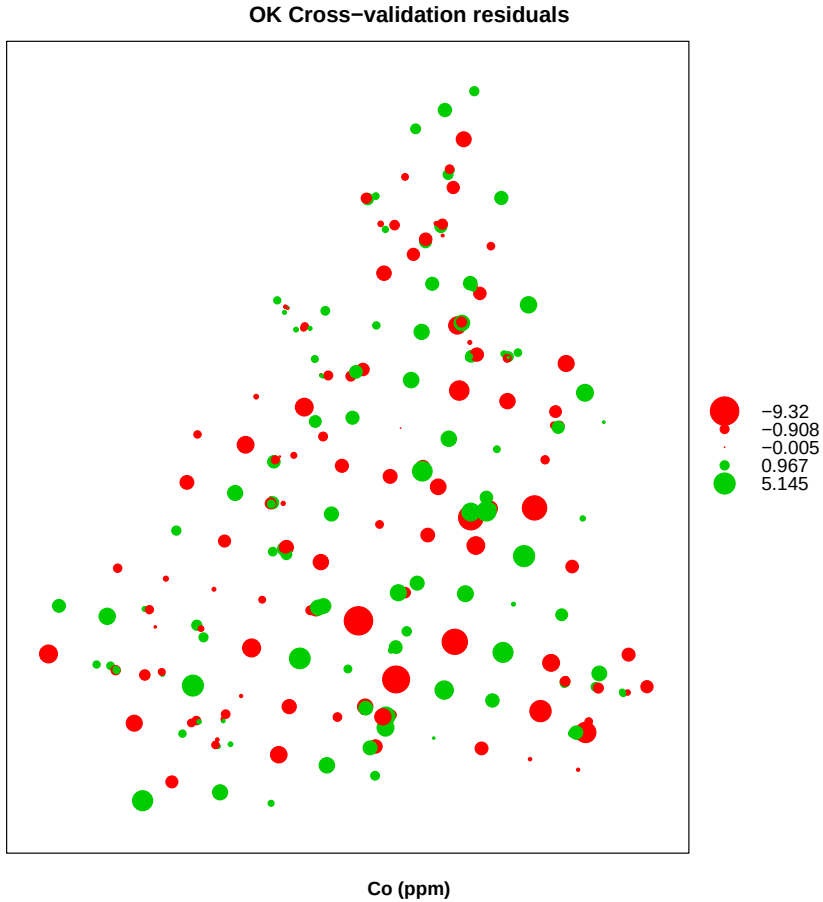
Another way to summarize the variability is the **median** of the Squared Deviation Ratio:

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

If a correct model is used for kriging, $\text{MeSDR} = 0.455$, which is the median of the standard χ^2 distribution (used here for the ratio of two variances) with one degree of freedom.

> "If the sample median is significantly less than 0.455 then this suggests that kriging overestimates the variance (possibly because of the effects of outliers on the variogram estimator) ...
> ... significantly greater ... suggests that kriging underestimates the variance."
>
> – 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.

Residuals from cross-validation and their location; Jura cobalt



actual – predicted; green are underpredictions

Topic 4: Kriging prediction variance

Recall from Lecture 4 that kriging is "optimal" with respect to a given model of spatial dependence, because the **kriging equations minimize the prediction variance** at each point to be predicted.

This is an **internal** measure of quality, because there is no independent dataset.

- Advantage: gives a measure of quality at **all** points
- Disadvantage: depends on the correctness of the **variogram model**

This **variance** presumably is from the **normally-distributed errors**, so we can use it accordingly to compute confidence intervals or threshold probabilities. This is quite useful in risk assessment.

Important: this makes the strong assumption that the random field is **Gaussian** – that is, the distribution of (spatially-correlated) deviations from the constant spatial mean is Gaussian!

Confidence intervals

Recall from Lecture 4:

The **two-sided interval** which has **probability** $(1 - \alpha)$ of containing the **true value** $z(\mathbf{x_0})$ is:

$$(\hat{z}(\mathbf{x_0}) - \zeta_{\alpha/2} \cdot \sigma) \leq \hat{z}(\mathbf{x_0}) \leq (\hat{z}(\mathbf{x_0}) + \zeta_{\alpha/2} \cdot \sigma)$$

where:

- $\hat{z}$ is the **estimated value** from OK;
- $\zeta_{\alpha/2}$ is the value of the standard normal distribution at **confidence level** $\alpha/2$;
- σ is the **square root of the prediction variance** from OK;

Topic 5: Spatial simulation

Simulation is the process or result of representing what reality *might* look like, given a model.

In geostatistics, this reality is usually a spatial distribution (map).

What is stochastic simulation?

- **"Simulation"** is a general term for studying a system without physically implementing it.
- **"Stochastic"** simulation means that there is a random component to the simulation model: quantified uncertainty is included so that each simulation is different.
- Non-spatial example: planning the number and timing of clerks in a new branch bank; customer behaviour (arrival times, transaction length) is stochastic and represented by probability distributions.
- Reference for spatial simulation:
  Goovaerts, P., 1997. *Geostatistics for natural resources evaluation*. Applied Geostatistics Series. Oxford University Press, New York; Chapter 8.

Why spatial simulation?

- Recall: the **theory of regionalized variables** assumes that the values we observe come from some **random process**; in the simplest case, with one **expected value** (first-order stationarity) with a **spatially-correlated error** that is the same over the whole area (second-order stationarity).

- So we'd like to see **"alternative realities"**; that is, spatial patterns that, by this theory, *could have* occurred in some "parallel universe" (i.e. another **realization** of the **spatial process**).

- In addition, **kriging maps are unrealistically smooth**, especially in areas with low sampling density.

  * Even if there is a high nugget effect in the variogram, this variability is *not* reflected in adjacent prediction points, since they are computed from the same observations, with almost the same weights.

When must simulation be used?

Goovaerts: "Smooth interpolated maps should not be used for applications sensitive to the presence of **extreme values and their patterns of continuity**." (p. 370)

Example: ground water travel time depends on sequences of large or small values ("critical paths"), not just on individual values.

Local uncertainty vs. spatial uncertainty

- Recall: kriging prediction also provides a **prediction error**; this is the BLUP and its error **for each prediction location separately**.
- So, at each prediction location we obtain a probability distribution of the prediction, a measure of its **uncertainty**. This is fine for evaluating each prediction individually.
- But, it is *not* valid to evaluate the set of predictions! Errors are *by definition* spatially-correlated (as shown by the fitted variogram model), so we can't simulate the error in a field by simulating the error in each point separately.
- **Spatial uncertainty** is a representation of the error over the **entire field of prediction locations** at the same time.

Practical applications of spatial simulation

- If the distribution of the target variable(s) over the study area is to be used as input to a **model**, then the uncertainty is represented by a number of simulations.

- Procedure:

  1. Simulate a "large" number of realizations of the spatial field
  2. Run the model on each simulation
  3. Summarize the output of the different model runs

- The statistics of the output give a direct measure of the **uncertainty** of the model in the light of the sample and the model of spatial variability.

Conditional simulation

This simulates the field, while **respecting the sample**.

The simulated maps resemble the best (kriging) prediction, but usually much more spatially-variable (depending on the magnitude of the nugget).

These can be used as inputs into **spatially-explicit models**, e.g. hydrology.

What is preserved in conditional simulation?

1. **Mean** over field

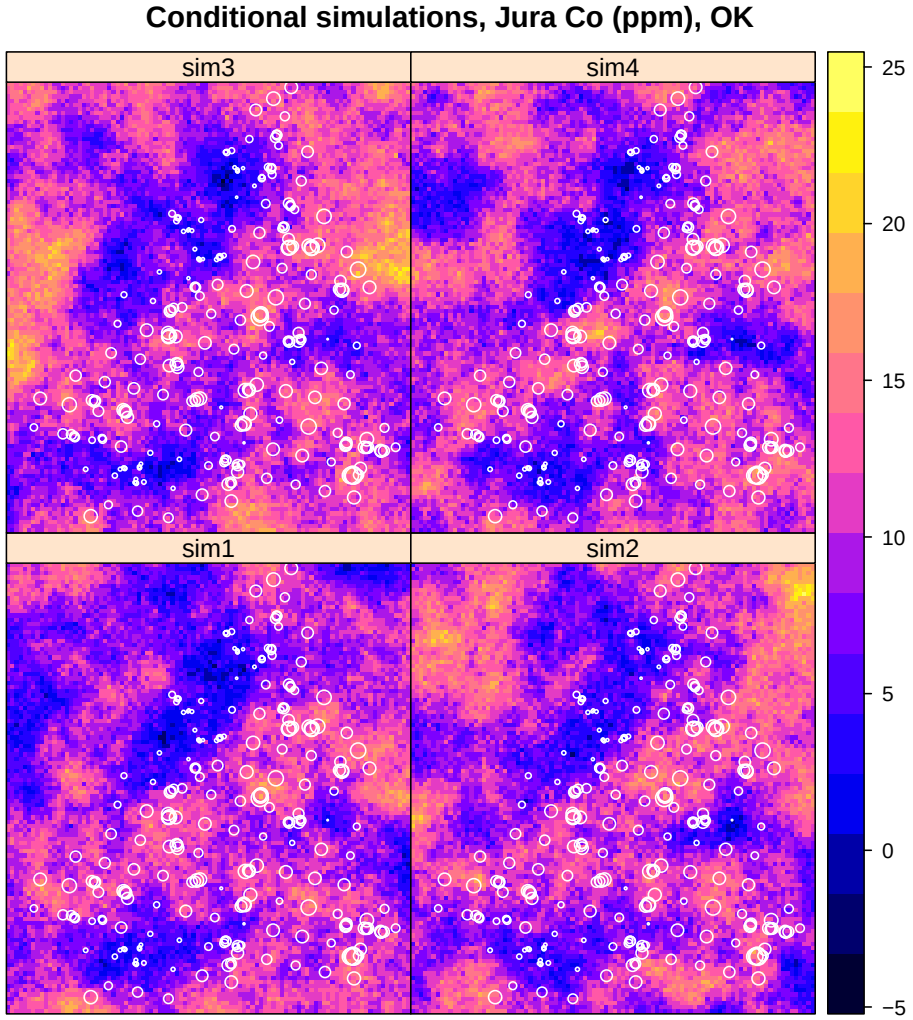
2. **Covariance structure**

3. **Observations** (sample points are predicted exactly)

See figures on the next page.

The OK prediction is then reproduced for comparison.

Conditional simulations: same field, different realizations



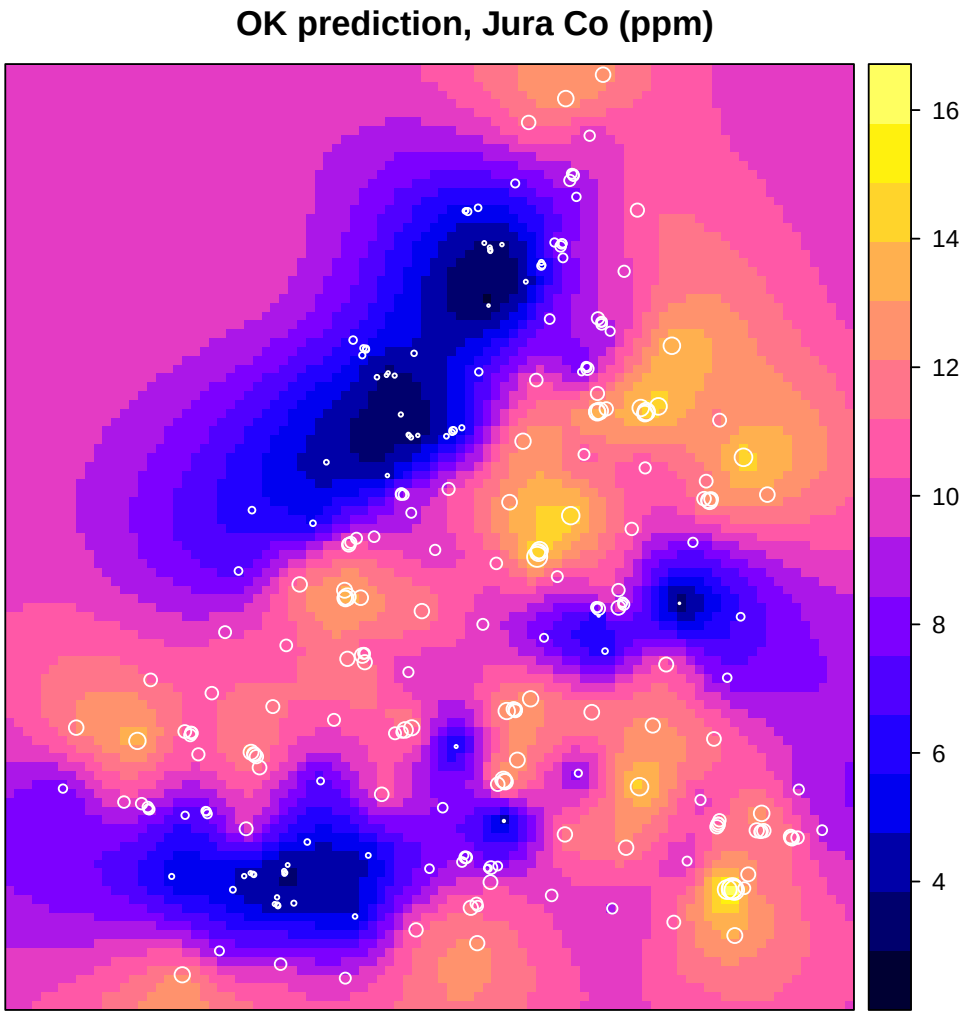
Jura Co concentration; known points over-printed as post-plot

To check your understanding . . .

Q7 : *In what respect do the conditional simulations resemble each other? In what respect do they not? In both cases, why?* *Jump to A7* •

OK prediction

Compare the conditional simulations with the single "best" prediction made by OK:



To check your understanding . . .

Q8 : *What is the principal difference between the conditional simulations and the OK prediction?* *Jump to A8* •

Unconditional simulation

In **unconditional** simulation, we simulate the field with *no reference to the actual sample*, i.e. the data we have. (It's only one realisation, no more valid than any other.)

This is used to **visualise a random field** as modelled by a variogram, *not* for prediction.

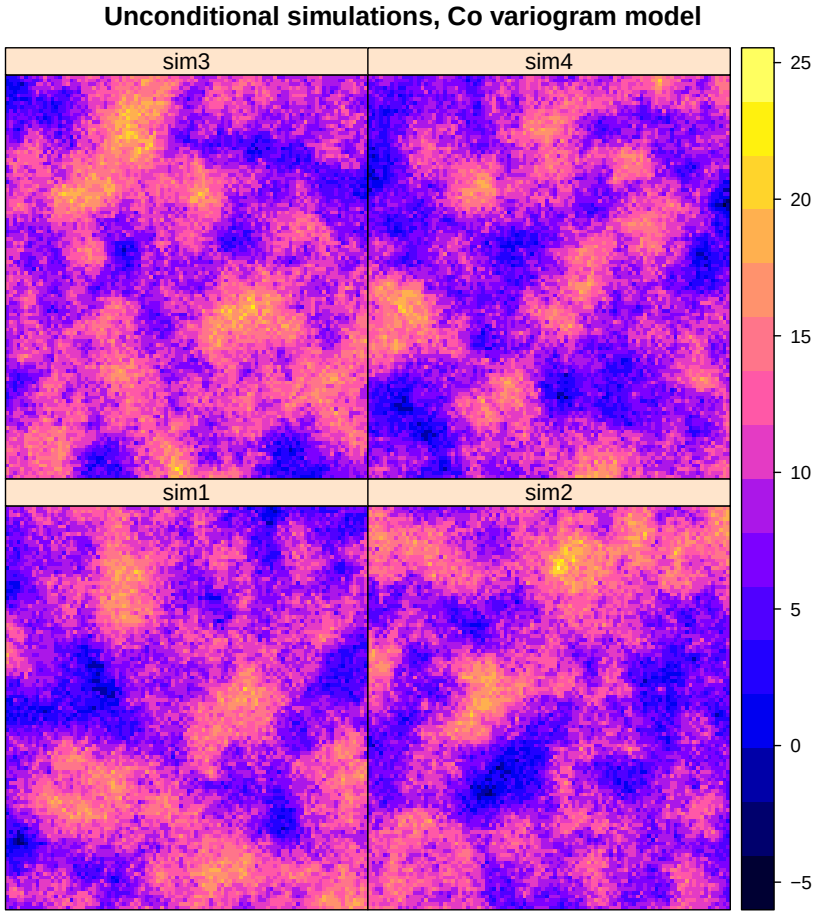
What is preserved in unconditional simulation?

1. **Mean** over field

2. **Covariance structure**

See figure on the next page. Note the similar degree of spatial continuity, but with no regard to the values in the sample.

Unconditional simulations: same field, different realizations

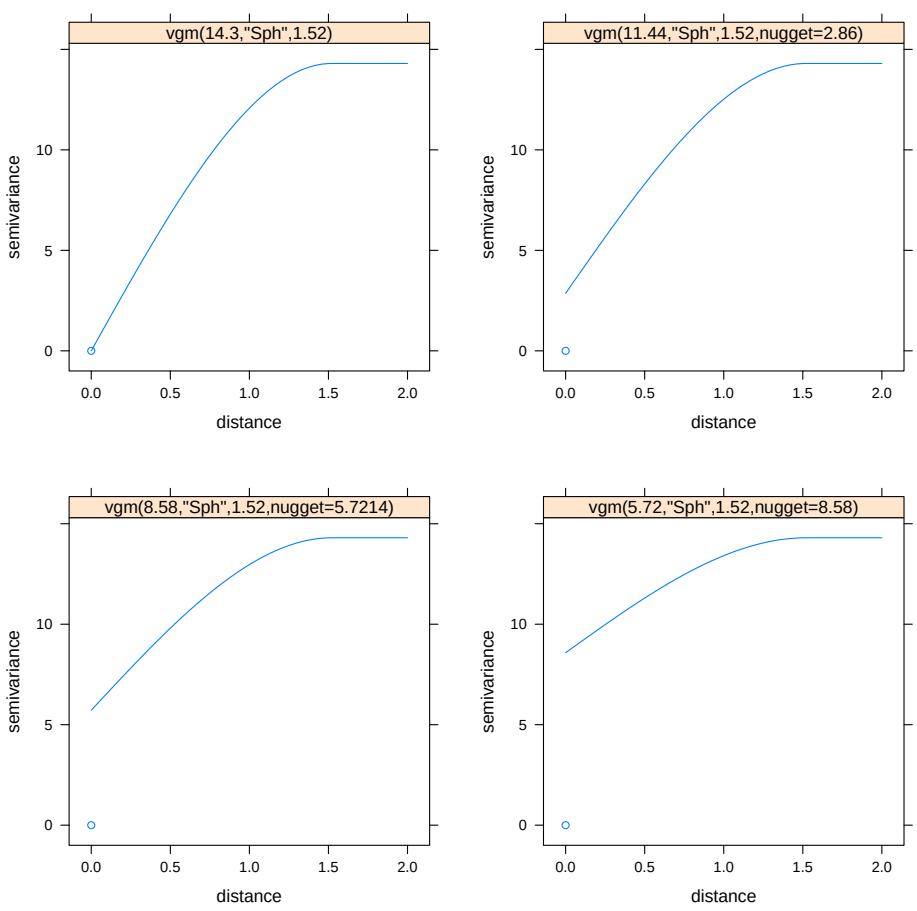


Model based on variogram analysis of Jura Co concentration

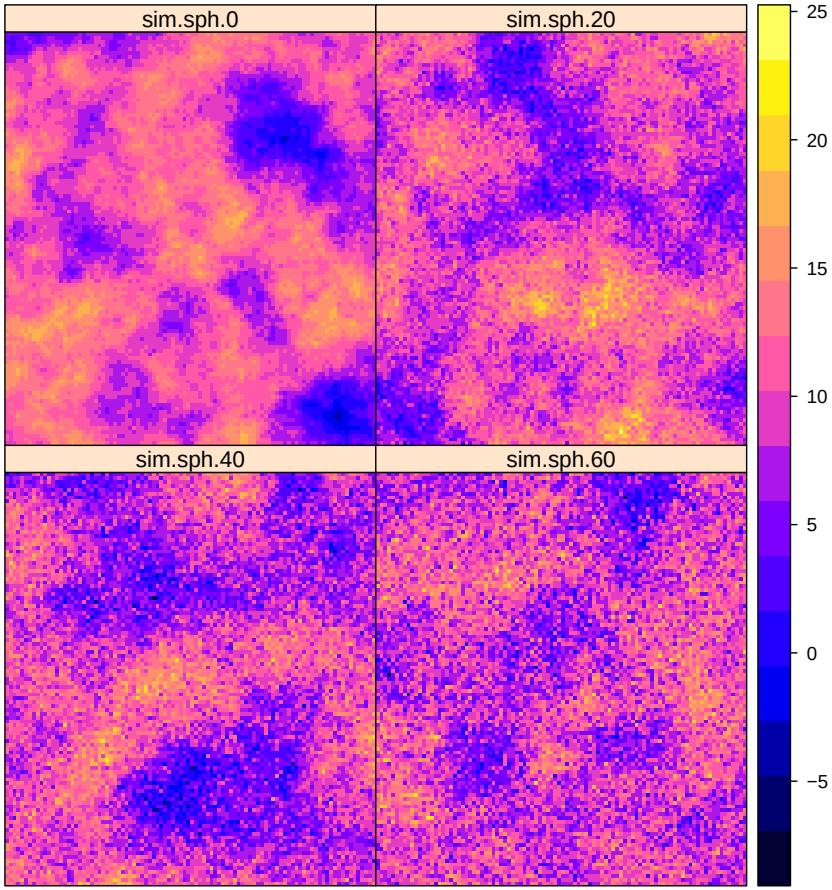
To check your understanding . . .

Q9 : *In what respect do the unconditional simulations resemble each other? In what respect do they not? In both cases, why?* *Jump to A9* •

Unconditional simulation: increasing nugget

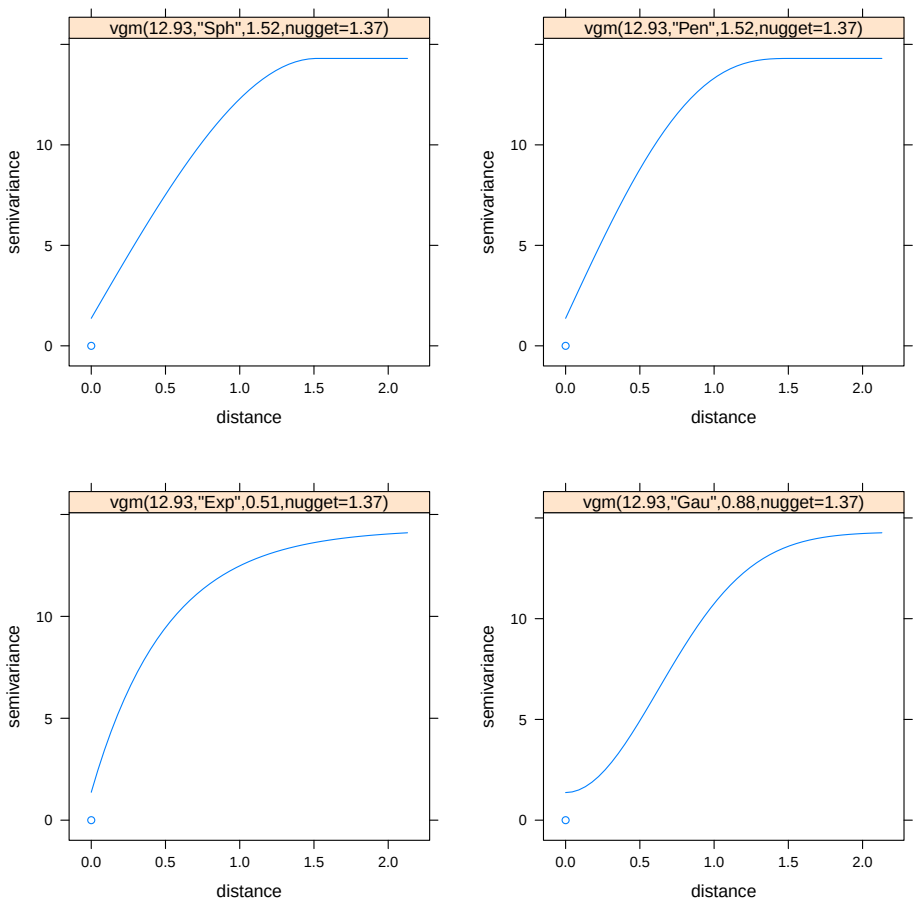


Variogram models

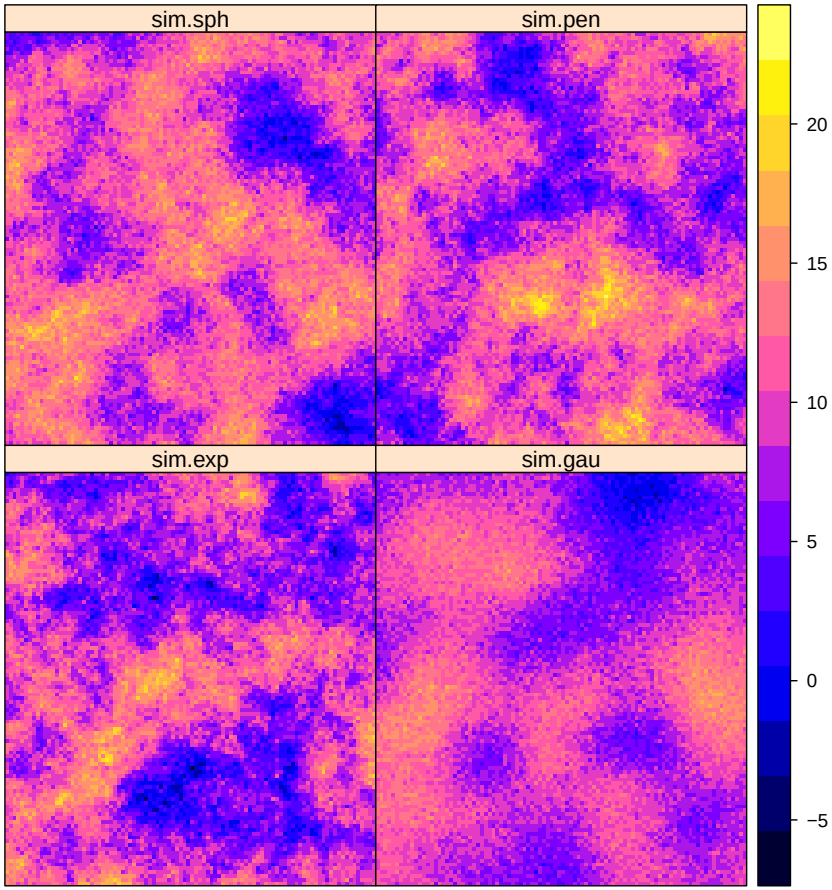


Simulated fields

Unconditional simulation: different models



Variogram models



Simulated fields

Simulation algorithm

There are several ways to simulate; see Emery, X. (2008). *Statistical tests for validating geostatistical simulation algorithms*. **Computers & Geosciences**, 34(11), 1610-1620. doi:10.1016/j.cageo.2007.12.012.

One algorithm is **sequential simulation** as used in the **gstat** package; in simplified form:

1. If conditional, place the data on the prediction grid
2. Pick a random unknown point; make a kriging prediction, along with its prediction variance
3. Assuming a normally-distributed prediction variance, simulate one value from this; add to the kriging prediction and place this at the previously-unknown point
4. This point is now considered "known"; repeat steps (2)-(3) until no more points are left to predict

Pebesma, E. J., & Wesseling, C. G. (1998). *Gstat: a program for geostatistical modelling, prediction and simulation*. **Computers & Geosciences**, 24(1), 17-31.

Answers

Q1 : *What is a major* **advantage** *of an external quality assessment?* •

A1 : *The data on which the quality assessment is based is* **independent** *of the data used to construct the model.* *Return to Q1* •

Q2 : *What is a major* **disadvantage** *of an external quality assessment?* •

A2 : *The expense of collecting another dataset; or, the loss of precision in calibration by holding out a portion for evaluation.* *Return to Q2* •

Answers

Q3 : *What are the largest over- and under-estimates? Does the distribution of residuals appear to be normal, as expected by theory?* •

A3 : *Largest over-estimate (negative residual):* $5.646\ mg\ kg^{-1}$ *more Co predicted than found; largest under-estimate (positive resdidual):* $8.973\ mg\ kg^{-1}$ *more found than predicted.*

Except for the one very large positive residual the distribution is symmetric; with a small sample size it's difficult to judge normality. *Return to Q3* •

Answers

Q4 : *Why, in the RMSE, are the differences between predicted and actual values squared?* •

A4 : *Because both positive and negative deviations are equally wrong.* *Return to Q4* •

Q5 : *Why then is the square root of the sum taken?* •

A5 : *To express the results in the original, not squared, units.* *Return to Q5* •

Answers

Q6 : *What would be the kriging prediction at a sample point, if it were included in the prediction dataset?* •

A6 : *Kriging is an exact predictor at known points, so would predict the value itself, if that point were included. So the "cross-validation" would appear perfect.* *Return to Q6* •

Answers

Q7 : *In what respect do the conditional simulations resemble each other? In what respect do they not? In both cases, why?* •

A7 : *The simulations all have similar degree of spatial continuity, as in unconditional simulation.*

In addition, all have a similar pattern with respect to the sample (e.g. hot spots are the same in all realizations). This is the difference with unconditional simulation, where there are no observations to respect. *Return to Q7* •

Q8 : *What is the principal difference between the conditional simulations and the OK prediction?* •

A8 : *The simulations are (realistically) "grainy"; the OK prediction is (unrealistically) smooth.* *Return to Q8*
•

Answers

Q9 : *In what respect do the unconditional simulations resemble each other? In what respect do they not? In both cases, why?* •

A9 : *The "patchiness" and "graininess" of all realizations is similar. This is because they all use the same model of spatial dependence.*

The overall pattern of high and low patches is the same. This is because they use the same observation points for conditioning.

The detailed local pattern is quite different, especially away from clusters of sample points. This is because the simulation has freedom to choose values as long as the covariance structure and sample values are respected. *Return to Q9* •