Applied geostatistics

Lecture 5 – Spatial prediction from point samples (Part 2)

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

October 26, 2010

Copyright © 2007–2010 University of Twente, Faculty ITC.
All rights reserved. Reproduction and dissemination of the work as a whole (not parts) freely permitted if this original copyright notice is included. Sale or placement on a web site where payment must be made to access this document is strictly prohibited. To adapt or translate please contact the author (`http://www.itc.nl/personal/rossiter`).

ITC

Topics for this lecture

1. Derivation of the Kriging equations: regression approach
2. *Derivation of the Kriging equations: minimization*
3. *Simple Kriging*
4. Block Kriging
5. Universal Kriging
6. *Kriging transformed variables*
7. *Kriging with External Drift*
8. *Regression Kriging*

The topics written *in italic script* are supplementary; there is not enough time to cover them in one lecture. They belong logically here, so are included for your reference.

Deriving the kriging system

In the previous lecture we saw how to apply a **model of local spatial dependence** (i.e. a variogram model) to **prediction by kriging**.

To avoid information overload, we deferred discussing the **kriging equations**, and in particular in what sense kriging is an **optimal** local predictor.

> **Note**: It is not necessary to understand this topic completely in order to correctly apply kriging. The derivation is necessarily mathematical and in places requires knowledge of matrix algebra or differential calculus. Still, everyone who uses kriging should be exposed to this at least once.

Two approaches

There are two approaches to this derivation:

Regression As a special case of **weighted least-squares** prediction in the **generalized linear model** with orthogonal projections in linear algebra

Minimization Minimizing the **kriging prediction** variance with calculus

Approach (1) is mathematically more elegant and provides a clear link to well-established linear modelling theory, so we present it as the main derivation.

Approach (2) is an application of standard minimization methods from differential calculus; but is not so transparent, because of the use of LaGrange multipliers; it is presented as an *optional* approach.

Topic 1: Regression derivation of the Ordinary Kriging equations

In this topic we show how to derive and solve the kriging equations. This provides a uniform framework for linear modelling ("regression") and kriging.

This approach has four steps:

1. Model the **spatial structure**, e.g. the covariance function or semivariogram function;
2. Estimate the **spatial mean**;
3. Set up a **kriging system** to **minimize the prediction variance**;
4. Compute the **kriging weights**

The spatial mean centres the predictions; note that in **Simple Kriging** (SK, see below), the mean is known somehow, and this step is not necessary.

Step (1) has been discussed in a previous topic.

Commentary

This topic requires a knowledge of **matrix algebra** and some familiarity with **general linear models** (e.g. weighted least-squares).

Step 2: estimate the spatial mean

The **spatial mean** of the variable z to be predicted, over the study area A, with area $|A|$ is:

$$\hat{\mu}(z) = \frac{1}{|A|} \int_A z$$

In practice this is **discretized** by summing over some fine **grid**:

$$\hat{\mu}(z) = \frac{1}{x_{\max}\, y_{\max}} \sum_{i=1 \ldots x_{\max}, j=1 \ldots y_{\max}} z_{i,j}$$

where $z_{i,j}$ is the value of the variable at grid location (i, j).

But in general we do not have measurements at all locations! That's what we want to find out. So this equation can't be applied. We need to estimate the spatial mean from **sparse observations** (i.e. our sample points)

The spatial mean is not the average!

Problem: because of **spatial autocorrelation**, it is not correct simply to **average** the observations to obtain the mean.

If there is any spatial dependence, in general:

Spatial mean $\neq$ **Average** of the observations

To check your understanding . . .

Q1 : *Consider a regular grid over some area; measure the data values and compute the ordinary average.*

If there is spatial dependence, what would happen to this average if we now make many observations near one of the grid points (i.e. a cluster), and not the others? *Jump to A1* •

The spatial mean

This is solved by weighted least squares, taking into account spatial correlation:

$$\hat{\mu} = (\mathbf{1}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{1})^{-1} \cdot (\mathbf{1}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{z})$$

- $\mathbf{1}$: a $n \times 1$ column vector of n 1's; so $\mathbf{1}^T$: a $1 \times n$ row vector of n 1's
  - In Universal Kriging (see below) this will be generalized to a design matrix $\mathbf{Q}$; here just a vector of 1's to estimate the mean
- $\mathbf{C}$: the covariance matrix $(n \times n)$ among known points;
  - Note that $C(\mathbf{0}) \equiv 1$ so all diagonals are 1.
- $\mathbf{z}$: the $n \times 1$ column vector of the n known data values

This is a special case of Generalized Least Squares (GLS).

To check your understanding . . .

Q2 : *How is the covariance matrix* $\mathbf{C}$ *computed?* *Jump to A2* •

Breaking this down ...

Two **quadratic forms**; both are $1 \times n \cdot n \times n \cdot n \times 1$ and so end up as scalars; multiply two scalars to get the spatial mean:

$$(\mathbf{1}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{1})^{-1}$$
$$(\mathbf{1}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{z})$$

Note: Let quadratic form $(\mathbf{1}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{1})^{-1} = \mathbf{V}$ in subsequent formulas.

If there is no spatial correlation, $\mathbf{C}$ reduces to $\mathbf{I}$: all diagonals are 1, all off-diagonals are 0, the inverse $\mathbf{I}^{-1} = \mathbf{I}$, and this reduces to Ordinay Least Squares (OLS) estimation.

In this case (design matrix $\mathbf{Q} = \mathbf{1}$), the OLS estimation of the spatial mean becomes the **arithmetic average**: the first quadratic term is $1/(\sum 1) = 1/n$ and the second $\sum z$; their product is $1/n \cdot \sum z$

Spatial mean computed with semivariances

Recall: the **semivariance** is the deviation of the covariance at some separation $\mathbf{h}$ from the total variance:

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$$

i.e.

$$C(\mathbf{h}) = C(\mathbf{0}) - \gamma(\mathbf{h})$$

But $C(\mathbf{0})$ is constant (1) in the covariance functions; further, both quadratic forms include the $\mathbf{C}$ matrix, so using its negative (plus a constant term), e.g. Γ, does not change the solution.

In fact, we can replace C by its difference from an arbitrary scalar α (element-wise):

$$\mathbf{C} \rightarrow \alpha \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}} - \mathbf{C}$$

So **semivariances** may be used rather than **covariances** in the formula for the spatial mean.

Step 3: Kriging prediction

Once we know the spatial mean, a kriging system can be set up without any additional constraints.

The special property of this system is that it is **BLUP**: "**B**est **L**inear **U**nbiased **P**redictor", given the modelled covariance structure.

Formula for kriging prediction variance

By definition, this is:

$$\mathrm{Var}(\hat{Z}_0 - Z_0)$$

where $\hat{Z}_0$ is the predicted value and Z_0 is the (unknown) true value.

For *any* **weighted average**, with weights λ_i, this is:

$$\mathrm{Var}(\sum \lambda_i \cdot Z_i - Z_0)$$

Kriging selects the weights λ to minimize this expression.

Continued . . .

To make the division of weights clear, define two vectors of length $n + 1$:

$$v = \begin{pmatrix} \lambda \\ -1 \end{pmatrix} \quad Z = \begin{pmatrix} Z \\ Z_0 \end{pmatrix}$$

where the first n elements relate to the observation points and the last element to the prediction point. Then

$$\begin{aligned} \mathrm{Var}(\sum \lambda_i \cdot Z_i - Z_0) &= v^T \mathrm{var}(Z) v \\ &= v^T \begin{pmatrix} C & c_0 \\ c_0^T & c_{00} \end{pmatrix} v \\ &= \lambda^T C \lambda - 2\lambda^T c_0 + c_{00} \end{aligned}$$

The key point here is that the variance-covariance matrix $\mathrm{var}(Z)$ is broken down into submatrices: C for the covariances between sample points, c_0 for the covariance of each sample point with the prediction point, and c_{00} for the variance at a point (the nugget).

Formula for kriging prediction

$$\hat{Z}_{OK} = \hat{\mu} + \mathbf{c}_0^T \cdot \mathbf{C}^{-1} \cdot (\mathbf{z} - \hat{\mu}\mathbf{1})$$

- $\hat{Z}_{OK}$ is the **prediction** at the prediction point
- $\hat{\mu}$ is the **spatial mean** computed in the previous step
- $\mathbf{c}_0$ is a $n \times 1$ column vector of the **covariance** between **each sample point** and the point to be **predicted**
- $\mathbf{C}$: the **covariance matrix** $(n \times n)$ among known points
- $\mathbf{z}$: the $n \times 1$ column vector of the n **known data values**
- $\mathbf{1}$: a $n \times 1$ column vector of n 1's, so that $\hat{\mu}\mathbf{1}$ is a column vector of the means, and $(\mathbf{z} - \hat{\mu}\mathbf{1})$ is a column vector of the **residuals** from the spatial mean

To check your understanding . . .

Q3 : *How is the covariance vector* $\mathbf{c_0}$ *computed?* *Jump to A3* •

Breaking this down . . .

- Note that the spatial mean $\hat{\mu}$ is subtracted from each data value; so the right summand

$$\mathbf{c}_0^T \cdot \mathbf{C}^{-1} \cdot (\mathbf{z} - \hat{\mu}\mathbf{1})$$

  is the **deviation** from the spatial mean at the prediction point.

- The spatial mean is added back in the left summand.

- If there is no spatial dependence, $\mathbf{c}_0 = \mathbf{0}$ and the prediction is just the spatial mean (which would then just be the average).

Step 4: Kriging weights

The OK prediction equation just derived does not explicitly give the weight to each observation; the kriging system can be used directly without computing these.

But we would often like to know the weights, to see the relative importance of each observation; also it may be more efficient to compute these and then the prediction as the weighted average.

Recall: the vector of kriging **weights**, is what to multiply each observation by in the weighted sum (prediction).

We get this by collecting all the terms that multiply the observed values $\mathbf{z}$ in the OK prediction equation:

$$\lambda^T = \mathbf{c}_0^T \cdot \mathbf{C}^{-1} - \mathbf{1}^T \cdot V\mathbf{1}^T \cdot \mathbf{C}^{-1}$$

Kriging variance without weights

Above the prediction variance was computed for *any* set of weights; now we've selected weights to minimize the variance. Then the variance can be expressed without explicitly showing the weights.

We do this by substituting the expression for weights (previous step) into the expression for prediction variance (Step 3), to obtain:

$$\mathrm{Var}(\hat{Z}_0 - Z_0) \quad = \quad c_{00} - \mathbf{c}_0^T \cdot \mathbf{C}^{-1} \cdot \mathbf{c}_0 + x_a^T \cdot \mathbf{V} \cdot x_a$$

where:

- c_{00} is the nugget covariance (at separation 0).
- $x_a = 1 - (\mathbf{c}_0^T \cdot \mathbf{C}^{-1} \cdot \mathbf{1})$

Note: For **block kriging**, replace c_{00} with c_{BB}, the average within-block variance (which will in general be smaller than the at-point variance); replace $\mathbf{c}_0$ and $\mathbf{C}$ with block-to-block covariances (see topic "Block Kriging").

Kriging variance in terms of semivariances

Because of the relation

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$$

the kriging variance can also be expressed as:

$$\mathrm{Var}(\hat{Z}_0 - Z_0) \quad = \quad c_{00} - \ \boldsymbol{\gamma}_0^T \cdot \boldsymbol{\Gamma}^{-1} \cdot \boldsymbol{\gamma}_0 - \ x_a/\mathbf{V}$$

where:

- $\boldsymbol{\Gamma}$ is the matrix of semivariances between sample points
- $\boldsymbol{\gamma}_0$ is the vector of semivariances between sample points and the point to be predicted

Note the change of sign, and that c_{00} is now implicit.

Topic 2: Minimization derivation of the Ordinary Kriging equations

This *optional* topic presents another apporach to deriving the kriging equations.

Commentary

In the detailed derivation that follows, keep your eye on the **big picture**:

- We want an **optimal** linear predictor, i.e. we want to compute **the best possible** weights for each sample point, to make each prediction;
- "Optimal" depends on some **objective function** which can be **minimized** with the best weights;
- We choose the **variance of the prediction** as the **objective function**; i.e. we want to **minimize** the **uncertainty** of the prediction.
- Given all this, we can derive a set of linear equations to derive the weights; this is known as the **kriging system**.
- The kriging system takes into account the **relative spatial positions** of each sample point, and their positions with respect to the prediction point.
- It does this with a **model of spatial covariance**.

You can appreciate that kriging is thus much more than "pressing a button" in some convenient software!

Prediction Variance

For **any** predictor (not just the kriging predictor):

- The **prediction** $\hat{z}(\mathbf{x}_0)$ at a given location $\mathbf{x}_0$ may be compared to the **true** value $z(\mathbf{x}_0)$; note the "hat" symbol to indicate an **estimated** value rather than a **measured** one.
- Even though we don't know the true (measured) value, we can write the expression for the **kriging variance**
- This is defined as the **expected value** of the **squared difference** between the **estimate** and the (unknown) **true value**:

$$\sigma^2(z(\mathbf{x}_0)) \equiv E\left[\{\hat{z}(\mathbf{x}_0) - z(\mathbf{x}_0)\}^2\right]$$

- If we can express this in some **computable form** (i.e. without the unknown true value) we can use it as an **optimality criterion**

Derivation of the kriging variance for OK

(Derivation based on P K Kitanidis, *Introduction to geostatistics: applications to hydrogeology*, Cambridge University Press, 1997; §3.9)

1. In OK, the estimated value is a **linear combination** of data values $\mathbf{x}_i$, with weights λ_i derived from the **kriging system**:

$$\hat{z}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i z(\mathbf{x}_i)$$

2. We don't yet know the weights λ but we do know how to compute the **estimated value** given the weights.

3. So, we can re-write the kriging variance with this weighted sum:

$$\sigma^2(z(\mathbf{x}_0)) = E\left[\{\sum_{i=1}^{N} \lambda_i z(\mathbf{x}_i) - z(\mathbf{x}_0)\}^2\right]$$

Expanding into parts

4. Add and subtract the (unknown but stationary) mean μ:

$$\sigma^2(z(\mathbf{x}_0)) = E[\{\sum_{i=1}^{N} \lambda_i(z(\mathbf{x}_i) - \mu) - (z(\mathbf{x}_0) - \mu)\}^2]$$

5. Expand the square:

$$\sigma^2(z(\mathbf{x}_0)) = E[(\sum_{i=1}^{N} \lambda_i z(\mathbf{x}_i) - \mu)^2 - 2\sum_{i=1}^{N} \lambda_i(z(\mathbf{x}_i) - \mu)(z(\mathbf{x}_0) - \mu) + (z(\mathbf{x}_0) - \mu)^2]$$

6. Replace the squared single summation (first term) by a **double summation**, i.e. with separate indices for the two parts of the square:

$$(\sum_{i=1}^{N} \lambda_i z(\mathbf{x}_i) - \mu)^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i\lambda_j(z(\mathbf{x}_i) - \mu)(z(\mathbf{x}_j) - \mu)$$

Bring expectations into each term

7. Bring the expectation into each term (n.b. expectation of a sum is the sum of expectations):

$$\begin{aligned}\sigma^2(z(\mathbf{x}_0)) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j E[(z(\mathbf{x}_i)-\mu)(z(\mathbf{x}_j)-\mu)] \\ &\quad -2\sum_{i=1}^{N}\lambda_i E[(z(\mathbf{x}_i)-\mu)(z(\mathbf{x}_0)-\mu)] + E[(z(\mathbf{x}_0)-\mu)^2]\end{aligned}$$

From expectations to covariances

8. Now, the three **expectations** in the previous expression are the definitions of **covariance** or **variance**:

   (a) $E[(z(\mathbf{x}_i) - \mu)(z(\mathbf{x}_j) - \mu)]$: **covariance** between **two sample points**
   (b) $E[(z(\mathbf{x}_i) - \mu)(z(\mathbf{x}_0) - \mu)]$: **covariance** between **one sample point** and the **prediction point**
   (c) $E[(z(\mathbf{x}_0) - \mu)^2]$: **variance** at the **prediction point**

9. So, replace the **expectations** with **covariances** and **variances**:

$$\begin{aligned}\sigma^2(z(\mathbf{x}_0)) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\mathrm{Cov}(z(\mathbf{x}_i), z(\mathbf{x}_j)) \\ &\quad -2\sum_{i=1}^{N}\lambda_i\mathrm{Cov}(z(\mathbf{x}_i), z(\mathbf{x}_0)) + \mathrm{Var}(z(\mathbf{x}_0))\end{aligned}$$

How can we evaluate this expression?

- **Problem 1**: how do we know the **covariances** between any two points?
  - Answer: by applying a **covariance function** which **only depends on spatial separation** between them.
    - Authorized functions e.g. spherical, exponential, ...
- **Problem 2**: how do we find the correct covariance function?
  - Answer: this is the subject of **variogram analysis**
- **Problem 3**: how do we know the **variance** at any point?
  - Answer: The actual value doesn't matter (it will be eliminated in the following algebra) but it **must be the same at all points**: this is the assumption of **second-order stationarity**.

Stationarity

- This is a term for **restrictions on the nature of spatial variation** that are required for OK to be correct

- **First-order** stationarity: the **expected values** (mean) at all locations in the field are the **same**:

$$E[Z(\mathbf{x}_i)] = \mu, \forall \mathbf{x}_i \in R$$

- **Second-order** stationarity:

1. The **variance** at any point is **finite** and the **same at all locations** in the field
2. The **covariance structure** depends only on **separation** between point pairs

- Note: it is easy to overcome problems with first-order stationarity (changing expected value across the field) but not second-order (changing variance and covariance)

Commentary

The concept of stationarity is often confusing, because stationarity refers to **expected** values, variances, or co-variances, rather than **observed** values.

In particular, of course the actual values change over the field! That is exactly what we want to use to predict at unsampled points. First-order stationarity just says that **before we sampled**, the **expected** value at all locations was the same.

That is, we **assume** the values result from a **spatially-correlated process** with a **constant mean** – **not** constant values!

Once we have some sample values, these influence the probability of finding values at other points, because of **spatial covariance**.

Unbiasedness

An **unbiased** estimate is one where the **expectation** of the **estimate** equals the expectation of the **true** (unknown) value: $E[\hat{z}(\mathbf{x}_0] \equiv E[z(\mathbf{x}_0)]$

For OK, we have assumed $E[z(\mathbf{x}_0)]$ is some unknown but **constant** μ. We have also decided to estimate $E[\hat{z}(\mathbf{x}_0]$ as a weighted sum, i.e. linear combination; see Step 1 previous slide. So we must have:

$$E[\hat{z}(\mathbf{x}_0] = \sum_{i=1}^{N} \lambda_i E[z(\mathbf{x}_i)] = \sum_{i=1}^{N} \lambda_i \mu = \mu \sum_{i=1}^{N} \lambda_i$$

because all the expected values are the same μ by **first-order stationarity**.

Since $E[\hat{z}(\mathbf{x}_0)] = \mu$ (unbiasedness), we must have $\sum_{i=1}^{N} \lambda_i = 1$.

This will be an additional restriction in the kriging system.

From point-pairs to separation vectors

As written above, the expressions are huge! The covariances between all point-pairs must be determined separately.

However, because of **second-order stationarity**, we assume that the covariances between any two points depend **only on their separation** and a single **covariance function**.

So rather than try to compute all the covariances, we just need to know this function, then we can apply it to any point-pair, just by knowing their separation.

We continue the derivation ...

Replace point-pairs with separation vectors

10. Substitute the **covariance function** of **separation** $\mathbf{h}$ into the expression; note that at separation zero (a point) this is the variance:

$$\sigma^2(z(\mathbf{x}_0)) = \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\text{Cov}(\mathbf{h}(i,j)) - 2\sum_{i=1}^{N}\lambda_i\text{Cov}(\mathbf{h}(i,0)) + \text{Cov}(\mathbf{0})$$

- $\mathbf{h}(i,0)$ is the separation between sample point $\mathbf{x}_I$ and the point to be predicted $\mathbf{x}_0$.
- $\mathbf{h}(i,j)$ is the separation between two sample points $\mathbf{x}_i$ and $\mathbf{x}_j$.

From covariances to semi-variances

11. Replace covariances by **semivariances**, using the relation $\mathrm{Cov}(\mathbf{h}) = \mathrm{Cov}(\mathbf{0}) - \gamma(\mathbf{h})$:

$$\sigma^2(z(\mathbf{x}_0)) = -\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\gamma(\mathbf{h}(i,j)) + 2\sum_{i=1}^{N}\lambda_i\gamma(\mathbf{h}(i,0))$$

Note that replacing covariances by semivariances changes the sign.

Done! This is now a **computable** expression for the **kriging variance** at any point $\mathbf{x}_0$, given the **locations** of the sample points $\mathbf{x}_i$, once the **weights** λ_i are known.

Next we will see how to use this expression to select optimum λ_i.

To check your understanding . . .

Q4 : *Does the kriging variance depend on the* **data values** *at the sample points, or on the predicted data value? How can you see this from the equation?* *Jump to A4* •

Summary: OK Kriging Variance

- Depends on the variogram function $\gamma(\mathbf{h})$ and the point configuration **around each point** to be predicted; we have derived this as:

$$\sigma^2(z(\mathbf{x}_0)) \equiv E[\{\hat{z}(\mathbf{x}_0) - z(\mathbf{x}_0)\}^2] = 2\sum_{i=1}^{N}\lambda_i\gamma(\mathbf{x}_i,\mathbf{x}_0) - \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\gamma(\mathbf{x}_i,\mathbf{x}_j)$$

  - First term: **lower semi-variances between a point and the sample points leads to a lower kriging variance**; different for each point to be predicted
  - Second term: **respect the co-variance structure of the sample points**; depends on configuration of sample points only

- We do not yet know what are the optimal weights λ, but once we do, we can calculate this kriging variance; so **we can selected the λ to minimize it**.

'Model globally, predict locally'

- The **kriging equations** are solved **separately for each point** $\mathbf{x}_0$, using the semivariances around that point, in a local neighbourhood; this gives a **different** set of weights λ for each point to be predicted.

- However, the **variogram model** $\gamma()$ used in these equations is estimated **only once**, using information about the spatial structure over the whole study area.

- Q: How is this possible?
  - A1: Assume **second-order stationarity**
  - A2: Assume at least **local first-order stationarity** (local weights will be high enough to mask long-distance non-stationarity)

Computing the weights

- There is one important piece of the puzzle missing: **How do we compute the weights** λ to predict at a given point?
  - (Recall the *ad-hoc* method: some power of inverse distance)
- We want these to be the "best", based on a computable **objective function**.
- There will be an optimum combination of weights at the point to be predicted, given the **point configuration** and the **modelled variogram**
- We compute these weights for each point to be predicted, by an **optimization criterion**, which in OK is **minimizing the kriging variance**.
- The previous slides have shown how to derive a **computable expression** for the kriging variance.

Commentary

The next several slides require a knowledge of **differential calculus**; however the central idea is simple: to **minimizing** an **objective function** (which we have just derived, i.e. the kriging variance).

The techniques of differential calculus are used to do the minimization.

Objective function (1): Unconstrained

In a minimization problem, we must define an **objective function** to be minimized. In this case, it is the kriging variance in terms of the N weights λ_i:

$$f(\lambda) = 2\sum_{i=1}^{N} \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) - \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j)$$

But this is **unbounded** and can be trivially solved by setting all weights to 0. We must add **another constraint** to **bound** it.

Objective function (2): Constrained

To bound the objective function, we need another constraint; here it is naturally **unbiasedness**; that is, the **weights must sum to 1**.

This is added to the system with a **LaGrange multipler** ψ:

$$f(\lambda, \psi) = 2 \sum_{i=1}^{N} \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) - \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) - 2\psi \left\{ \sum_{i=1}^{N} \lambda_i - 1 \right\}$$

Note that the last term $\equiv 0$, i.e. the prediction is unbiased. The LaGrange multiplier may be changed (we will see how) but that final term always drops out of the prediction.

Note: It is possible to derive the kriging system in other ways, without the LaGrange multipliers and in terms of covariances. See Topic "Regression approach to kriging" below.

Minimization

This is now a system of $N+1$ equations in $N+1$ unknowns.

Minimize by setting all $N+1$ **partial derivatives** to zero:

$$\begin{aligned} \frac{\partial f(\lambda_i, \psi)}{\partial \lambda_i} &= 0, \forall i \\ \frac{\partial f(\lambda_i, \psi)}{\partial \psi} &= 0 \end{aligned}$$

In the differential equation with respect to ψ, all the λ are constants, so the first two terms differentiate to 0; in the last term the ψ differentiates to 1 and we are left with the unbiasedness condition:

$$\sum_{i=1}^{N} \lambda_i = 1$$

The Kriging system

In addition to unbiasedness, the partial derivatives with respect to the λ_i give N equations (one for each λ_i) in $N+1$ unknowns (the λ_i plus the LaGrange multiplier ψ):

$$\sum_{j=1}^{N} \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) + \psi = \gamma(\mathbf{x}_i, \mathbf{x}_0), \quad \forall i$$

This is now a **system of $N+1$ equations in $N+1$ unknowns** and can be solved by linear algebra.

The **semivariances between sample points** $\gamma(\mathbf{x}_i, \mathbf{x}_j)$ are computed **only once** for any point configuration; however the **semivariances at a sample point** $\gamma(\mathbf{x}_i, \mathbf{x}_0)$ must be **computed separately for each point to be predicted**.

Solving the Kriging system

At each point to be predicted:

1. **Compute the semivariances** γ from the separation between the point and the samples, according to the **modelled variogram**

2. **Solve simultaneously** for the weights and multiplier

3. **Compute the predicted value** as the **weighted average** of the samples

4. **Compute the variable term** of the kriging variance

5. **Add the constant term** of the kriging variance to get the total variance.

Importance of the variogram model

- The kriging system is solved using the modelled semi-variances
- **Different models will give different kriging weights** to the sample points . . .
- . . . and these will give different predictions
- Conclusion: **bad model → bad predictions**

Matrix form of the Ordinary Kriging system

$$\mathbf{A}\lambda = \mathbf{b}$$

$$\mathbf{A} = \begin{bmatrix} \gamma(\mathbf{x}_1,\mathbf{x}_1) & \gamma(\mathbf{x}_1,\mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_1,\mathbf{x}_N) & 1 \\ \gamma(\mathbf{x}_2,\mathbf{x}_1) & \gamma(\mathbf{x}_2,\mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_2,\mathbf{x}_N) & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \gamma(\mathbf{x}_N,\mathbf{x}_1) & \gamma(\mathbf{x}_N,\mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_N,\mathbf{x}_N) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}$$

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \psi \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} \gamma(\mathbf{x}_1,\mathbf{x}_0) \\ \gamma(\mathbf{x}_2,\mathbf{x}_0) \\ \vdots \\ \gamma(\mathbf{x}_N,\mathbf{x}_0) \\ 1 \end{bmatrix}$$

Inside the Matrix

The **block matrix** notation shows the semivariances and LaGrange multiplier explicitly:

$$\mathbf{A} = \begin{pmatrix} \Gamma & \mathbf{1} \\ \mathbf{1}^T & 0 \end{pmatrix}$$

$$\lambda = \begin{bmatrix} \Lambda \\ \psi \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} \Gamma_0 \\ 1 \end{bmatrix}$$

Solution

This is a system of $N + 1$ equations in $N + 1$ unknowns, so can be solved uniquely, as long as $\mathbf{A}$ is positive definite; this is guaranteed by using authorized models. This has the solution (in matrix notation):

$$\lambda = \mathbf{A}^{-1}\mathbf{b}$$

Now we can **predict** at the point, using the weights:

$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i z(\mathbf{x}_i)$$

The **kriging variance** at a point is given by the scalar product of the weights (and multiplier) vector λ with the right-hand side of the kriging system: Note that λ includes as its last element ψ, which depends on covariance structure of the sample points:

$$\hat{\sigma}^2(\mathbf{x}_0) = \mathbf{b}^T\lambda$$

Exercise

At this point you should do the **first section** of **Exercise 5: Predicting from point samples (Part 2)** which is provided on the module CD:

- §1 **Kriging weights**

This should take less than an hour.

As in all exercises there are **Tasks**, followed by R code on how to complete the task, then some **Questions** to test your understanding, and at the end of each section the **Answers**. Make sure you understand all of these.

Topic 2: Simple Kriging

In OK:

- We must estimate the **regional** (stationary) mean along with the predicted values, in the OK system.
- However, there may be situations where **the regional mean is known**. Then we can use so-called **Simple Kriging** (SK)

Similarly for UK or KED:

- We must estimate both the intercept (β_0) and all other trend coeficients (β_i), along with the predicted values, in the UK system.
- Similarly, if **the trend is known**, we can use "Simple" variants of UK and KED.

Simple Kriging (SK)

- The (stationary) **regional mean** must be known *a priori*
  - in **Regression Kriging**, by definition the residuals have mean 0.
  - in **Indicator Kriging** we expect each quantile to have the corresponding proportion of 1's (if the sample was unbiased)
  - Note: the mean of a spatial sample is generally *not* the spatial mean, precisely because the observations are spatially-correlated.

- The mean may also be known from **previous studies** that give a better estimate than the (small) sample size we are working with; this could be a study of a larger area enclosing our study area.

- A **non-stationary** (moving) mean may be known from external evidence, e.g. a modelling exercise or regression analysis.

SK Predictions

- We reformulate the OK estimate **without the constraint** that weights sum to 1
- So, any bias from the weights must be compensated with respect to the known mean μ when predicting at a point:

$$\hat{Z}_{SK}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i z(\mathbf{x}_i) + \{1 - \sum_{i=1}^{N} \lambda_i\}\mu$$

- I.e. **Estimate = Weighted Sum + Bias Correction**
- The correction term is not present in OK prediction; its analog is the constraint that the weights sum to 1.

The SK system

There is **no need for a LaGrange multipler** in the SK system, since there is no unbiasedness condition on the weights at any point. The system thus has one less row and column than the OK system.

$$\mathbf{A}\lambda = \mathbf{b} \qquad \mathbf{A} = \begin{bmatrix} \gamma(\mathbf{x}_1,\mathbf{x}_1) & \gamma(\mathbf{x}_1,\mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_1,\mathbf{x}_N) \\ \gamma(\mathbf{x}_2,\mathbf{x}_1) & \gamma(\mathbf{x}_2,\mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_2,\mathbf{x}_N) \\ \gamma(\mathbf{x}_N,\mathbf{x}_1) & \gamma(\mathbf{x}_N,\mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_N,\mathbf{x}_N) \end{bmatrix}$$

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} \gamma(\mathbf{x}_1,\mathbf{x}_0) \\ \gamma(\mathbf{x}_2,\mathbf{x}_0) \\ \vdots \\ \gamma(\mathbf{x}_N,\mathbf{x}_0) \end{bmatrix}$$

Solution of the SK system

This is a system of N equations in N unknowns, so can be solved uniquely, as long as $\mathbf{A}$ is positive definite (as in OK); this is guaranteed by using authorized models.

However, since the weights λ_i are not constrained to sum to 1, in addition **the variogram must be bounded** (e.g. an exponential or spherical model; but not a power model). Then:

$$\lambda \quad = \quad \mathbf{A}^{-1}\mathbf{b}$$

Estimate at the point, using the weights and the known mean:

$$\hat{Z}_{SK}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i z(\mathbf{x}_i) + \{1 - \sum_{i=1}^{N} \lambda_i\}\mu$$

Note that the unbiasedness at each prediction point is restored here; if the λ did not sum to 1, it is corrected.

SK variance

The kriging variance at a point is given by:

$$\begin{aligned} \hat{\sigma}^2_{SK}(\mathbf{x}_0) &= \mathbf{b}^T \lambda \\ &= \sum_{i=1}^{N} \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) \end{aligned}$$

This formula has a simple interpretation:

- the kriging variance is the **weighted sum** of the **semivariances** between each sample point and the prediction point;
- the weights are the **kriging weights** found by solving the kriging system.

Topic 4: Block Kriging

Often we want to predict **average values** of some target variable in **blocks** of some defined size, not at points.

> **Example**: average woody biomass in a forest block of 40ha, if this is a mimimum management unit, e.g. we will decide to harvest or not the whole block. We don't care about any finer-scale information, we wouldn't use it if we had it.

Block kriging (BK) is quite similar in form to OK, but the **kriging variances are lower**, because the **within-block variability** is removed.

Commentary

There is only one new idea in this section: by predicting an **average over some block** larger than the support, we **reduce the kriging variance**.

Most of this section shows how this is expressed mathematically.

The practical implication is shown in the comparative figures at the end of the section.

Block Ordinary Kriging (BK)

- Estimate at blocks of a defined size, with unknown mean (which must also be estimated) and no trend

- Each **block** B is estimated as the weighted average of the values at all sample **points** $\mathbf{x}_i$:

$$\hat{Z}(B) = \sum_{i=1}^{N} \lambda_i z(\mathbf{x}_i)$$

- As with OK, the weights λ_i sum to 1, so that the estimator is **unbiased**, as for OK

The Block Kriging system

The same derivation as for the OK system produces these equations:

$$\sum_{j=1}^{N} \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) + \psi(B) = \overline{\gamma}(\mathbf{x}_i, B), \quad \forall i$$

- The semivariances are now between **sample points** and the **block** to be predicted
- The semivariance with a block is written as $\overline{\gamma}$, the overline indicating an **average**
- The left-hand side (semivariances between sample points) is the same as in OK

Matrix form of the Block Kriging system

$$\mathbf{A}\lambda = \mathbf{b}$$

$$\mathbf{A} = \begin{bmatrix} \gamma(\mathbf{x}_1,\mathbf{x}_1) & \gamma(\mathbf{x}_1,\mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_1,\mathbf{x}_N) & 1 \\ \gamma(\mathbf{x}_2,\mathbf{x}_1) & \gamma(\mathbf{x}_2,\mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_2,\mathbf{x}_N) & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \gamma(\mathbf{x}_N,\mathbf{x}_1) & \gamma(\mathbf{x}_N,\mathbf{x}_2) & \cdots & \gamma(\mathbf{x}_N,\mathbf{x}_N) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}$$

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \psi(B) \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} \overline{\gamma}(\mathbf{x}_1,B) \\ \overline{\gamma}(\mathbf{x}_2,B) \\ \vdots \\ \overline{\gamma}(\mathbf{x}_N,B) \\ 1 \end{bmatrix}$$

Solution

$$\lambda = \mathbf{A}^{-1}\mathbf{b}$$

Now we can estimate at the block, using the weights:

$$\hat{Z}(B) = \sum_{i=1}^{N} \lambda_i z(\mathbf{x}_i)$$

The kriging variance for the block is given by:

$$\hat{\sigma}^2(B) = \mathbf{b}^T\lambda - \overline{\gamma}(B,B)$$

Note that **the variance is reduced by the within-block variance**.

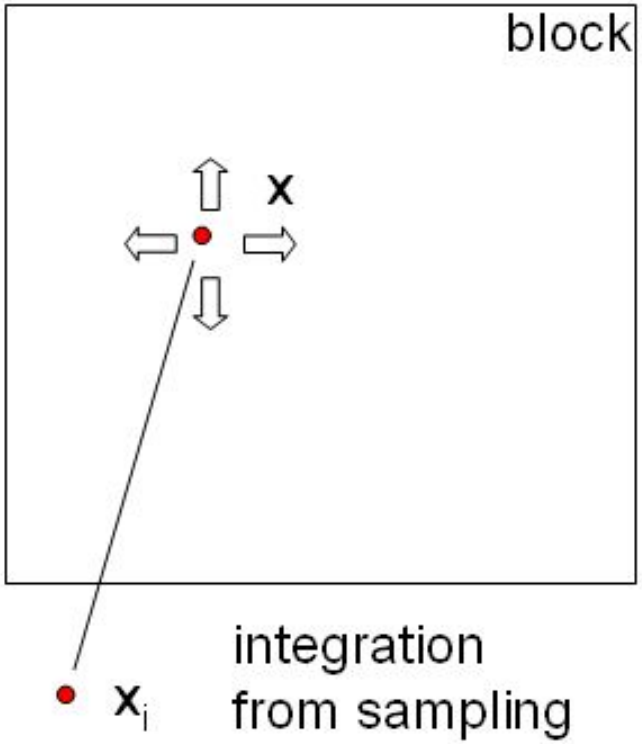
Commentary

The semivariances γ in the above formulation are not just a function of separation, because they are *not* between points. Instead, they are between sample **points** and prediction **blocks**. This is illustrated in the following figure (left side).

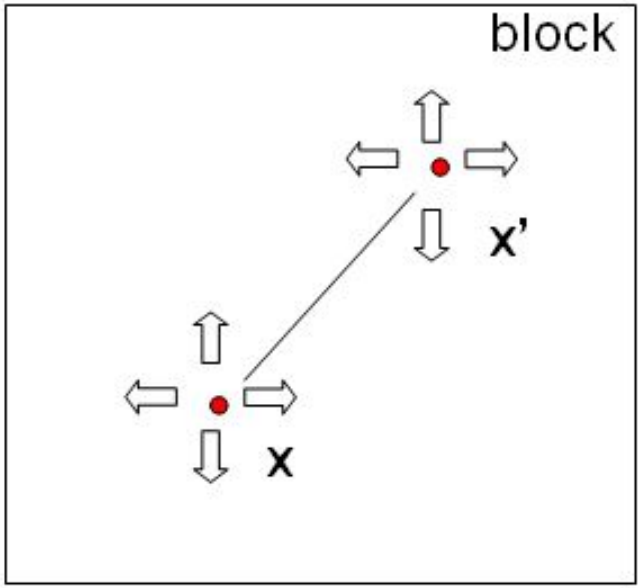
In addition, the variance at the prediction location is now not at a point, but rather at a **block**. So some of the kriging variance must be accounted for within that block, i.e. the variance that is due to short-range variability at distances **shorter than the block size**. This is illustrated in the following figure (right side).

We will then discuss in detail how to compute these two.

Integration of semivariances for block kriging



integration from sampling point to block



integration within block

Computing the semivariance between a point and a block

The complication here, compared to OK, is that the semivariances in the $\mathbf{A}$ matrix are between sample points and the entire **block** to be predicted: $\overline{\gamma}(\mathbf{x}_i, B)$

So there is not a single distance than can be substituted into the variogram model. We have to integrate over the block:

$$\overline{\gamma}(\mathbf{x}_i, B) = \frac{1}{|B|} \int_B \gamma(\mathbf{x}_i, \mathbf{x}) \, d\mathbf{x}$$

where $|B|$ is the area of the block, and $\mathbf{x}$ is a point within the block.

As written **all** points in the block (conceptually, an infinite number!) would be used; n practice, this is done by **discretization** of the block into a set of points.

Computing the within-block variances

This is the factor by which the estimation variance is reduced:

$$\overline{\gamma}(B,B) = \frac{1}{|B|^2}\int_B \int_B \gamma(\mathbf{x},\mathbf{x}')\, d\mathbf{x}d\mathbf{x}'$$

As the block size $|B|$ approaches zero, the double integral also approaches zero; in fact this is the limit. This shows that **OK is a special case of BK**.

In practice, this is also calculated by **discretizing** the block into n points:

$$\overline{\gamma}(B,B) \approx \frac{1}{|B|^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma(\mathbf{x}_i,\mathbf{x}_j)$$

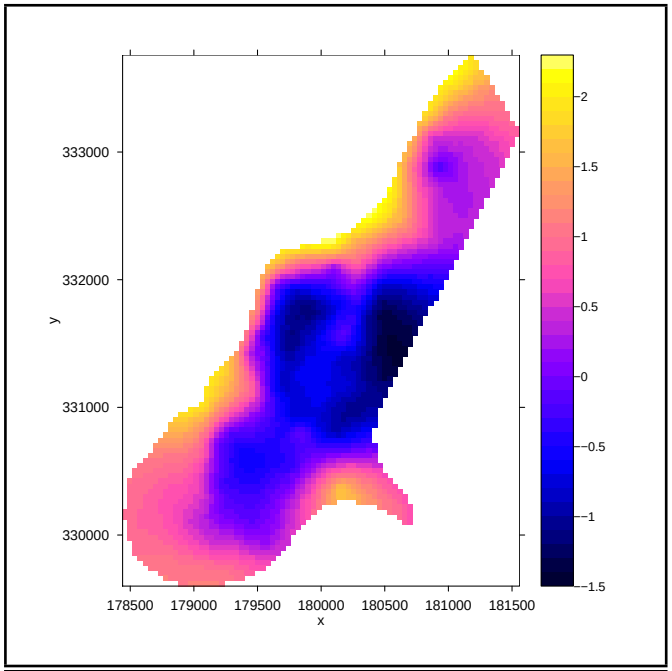
where of course $\sum_i w_i = 1$; the weights are set by their position within the unit block.

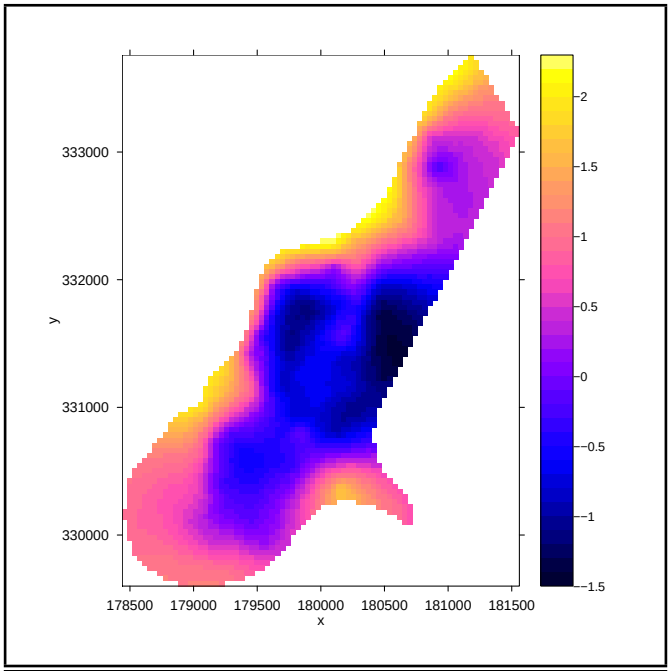
Visualizing the effect of block size

The following graphs show the changing **predictions** and their **variances** as the block size is increased. This is from the Meuse soil pollution study; target variable is log10(Cd).

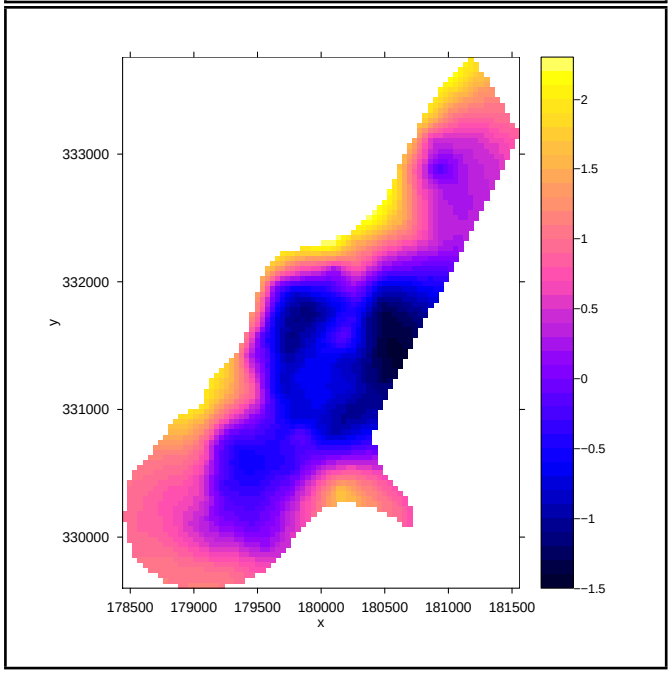
Each graph uses one scale, to allow direct comparaison.

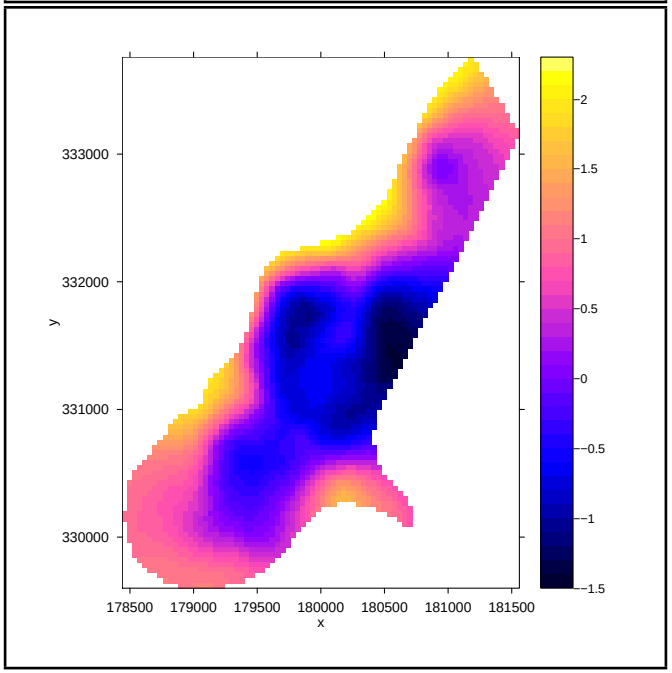
Predictions: OK, BK10





BK40, BK160



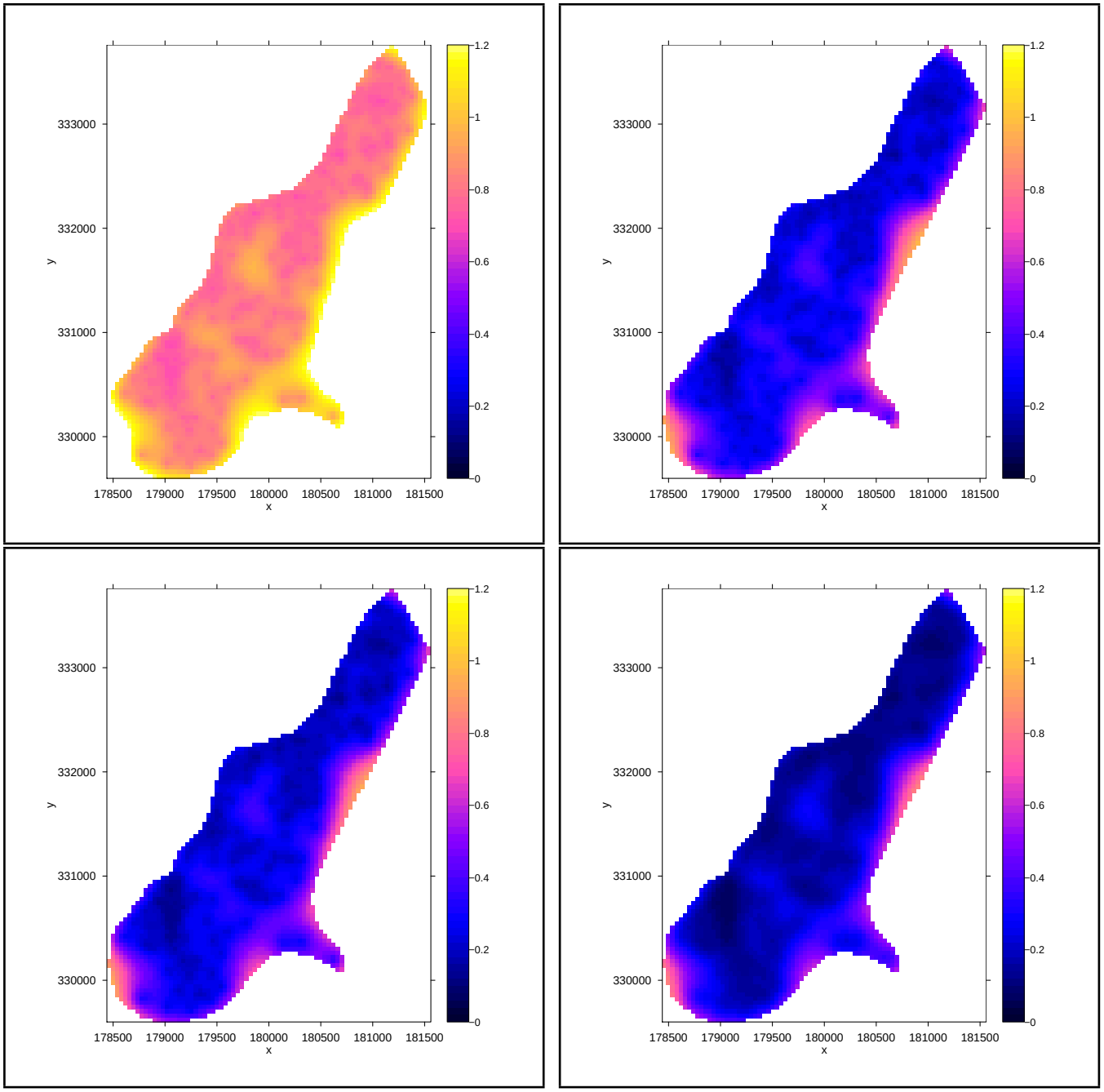


To check your understanding . . .

Q5 : *What are the differences between the* **predictions** *with different block sizes?* *Jump to A5* •

Variances: OK, BK10

BK40, BK160



To check your understanding . . .

Q6 : *What are the differences between the* **kriging variances** *with different block sizes?* *Jump to A6* •

Topic 5: Universal Kriging (UK)

This is a **mixed predictor** which includes a **global trend** as a function of the **geographic coördinates** in the kriging system, as well as **local spatial dependence**.

Example: The depth to the top of a given sedimentary layer may have a regional trend, expressed by geologists as the **dip** (angle) and **strike** (azimuth). However, the layer may also be locally thicker or thinner, or deformed, with spatial covariance in this local structure.

UK is recommended when there is evidence of **first-order non-stationarity**, i.e. the **expected value** varies across the map, but there is still **second-order stationarity** of the **residuals** from this trend.

Note: The "global" trend can also be fitted **locally**, within some user-defined radius, so that this interpolator can range from local (immediate neighbourhood) to global (whole area), according to the analyst's evidence on spatial structure.

Terminology: UK vs. KED

- There is confusion over the term "Universal Kriging" (UK)
- In these notes, we restrict it to the use of a **geographic trend** as co-variate
- The term "UK" is used by some authors also to include a trend in one or more **feature-space** predictors, i.e. co-variables.
- In this course we call this use of feature-space predictors **Kriging with External Drift** (KED)
- The mathematics are the same; it is the co-variables that are different:
  - UK : only geographic coördinates
  - KED : also (or only) feature-space co-variables.

Prediction with UK

In UK, we model the value of variable z at location $\mathbf{x}_i$ as the sum of:

- a regional **non-stationary trend** $m(\mathbf{x}_i)$ and a
- a local **spatially-correlated random component** $e(\mathbf{x}_i)$; the **residuals** from the regional trend

$$Z(\mathbf{x}_i) = m(\mathbf{x}_i) + e(\mathbf{x}_i)$$

Here $m(\mathbf{x})$ is **not a constant** as in OK, but instead is a **function of position**, i.e. the global **trend**.

Base functions

The trend is modelled as a **linear combination** of p known **base functions** $f_j(s)$ and p unknown constants β_j (these are the **parameters** of the base functions):

$$Z(\mathbf{x}_i) = \sum_{j=1}^{p} \beta_j f_j(\mathbf{x}_i) + e(\mathbf{x}_i)$$

Examples of base functions

- For **linear** drift:

$$f_0(\mathbf{x}) = 1,\ f_1(\mathbf{x}) = x_1,\ f_2(\mathbf{x}) = x_2$$

  where x_1 is one coördinate (say, E) and x_2 the other (say, N)

- Note that $f_0(\mathbf{x}) = 1$ estimates the global mean (as in OK).

- For **quadratric** drift: also include second-order terms:

$$f_3(\mathbf{x}) = x_1^2,\ f_4(\mathbf{x}) = x_1 x_2,\ f_5(\mathbf{x}) = x_2^2$$

Predictions at points

A point is **predicted** as in OK:

$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i z(\mathbf{x_i})$$

But, **the weights** λ_i for each sample point take into account both the **global trend** and **local effects**.

We need to set the UK system up to include both of these.

Unbiasedness of predictions

The **unbiasedness** condition is expressed with respect to the **trend** as well as the overall mean (as in OK):

$$\sum_{i=1}^{N} \lambda_i f_k(\mathbf{x_i}) = f_k(\mathbf{x_0}), \quad \forall k$$

The expected value at each point of all the **functions** must be that predicted by that function. The first of these is the overall mean (as in OK).

Example: If $f_1(\mathbf{x}_0) = x_1$ (linear trend towards the E), then at each point $\mathbf{x}_0$ the expected value must be x_1, i.e. the point's E coördinate:

$$\sum_{i=1}^{N} \lambda_i x_i = x_1$$

This is a **further restriction** on the weights λ.

Setting up the UK system

As for OK, the **objective function** to be **minimized** is the **kriging variance**; in addition to the unbiasedness constraint on the mean, there are unbiasedness constraints on the values of the base functions:

$$\sum_{j=1}^{N} \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) + \psi_0 + \sum_{k=1}^{K} \psi_k f_k(\mathbf{x}_i) = \gamma(\mathbf{x}_0, \mathbf{x}_i), \forall i \quad \text{(sample points)} \quad (1)$$

$$\sum_{i=1}^{N} \lambda_i = 1 \quad \text{(unbiasedness of mean)} \quad (2)$$

$$\sum_{i=j}^{N} \lambda_j f_k(\mathbf{x}_j) = f_k(\mathbf{x}_0), \forall k \quad \text{(base functions)} \quad (3)$$

Where is the trend surface?

Note that the coefficients β of the global trend surface are **not** mentioned in the UK system.

However, they are **implicit** in the solution and will affect the λ_i at each prediction point; that is, the local weights around each point will take into account the global trend.

(This has the advantage that we can restrict the search neighbourhood and fit the "global" trend only in a user-defined neighbourhood, i.e. in **patches**.)

Regression approach to Universal kriging

The general linear model applied to OK (above) can be extended to UK (or KED, see below) is natural. Instead of a vector of 1's we have the **design matrix** $\mathbf{Q}$ and the values $\mathbf{x}_0$ of the **basis functions** at the prediction point.

The spatial mean $\hat{\mu}$ is replaced by all the parameters $\hat{\beta}$ of the trend:

$$\hat{\beta} = (\mathbf{Q}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{Q})^{-1} \cdot (\mathbf{Q}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{z})$$

Note that the **GLS trend surface coefficients** are indeed produced by this method; analogous to the spatial mean in the regression formulation of OK.

This is an advantage of this formulation compared to the minimization approach.

(continued . . .)

UK prediction (regression formulation)

Then the kriging prediction becomes:

$$\hat{Z}_{OK} = \mathbf{x}_0^T \hat{\beta} + \mathbf{c}_0^T \cdot \mathbf{C}^{-1} \cdot (\mathbf{z} - \mathbf{Q}\hat{\beta})$$

- $\mathbf{x}_0^T \hat{\beta}$ here replaces the single estimate of the spatial mean $\hat{\mu}$ in OK; this term is the prediction from the GLS trend at the point.
- $(\mathbf{z} - \mathbf{Q}\hat{\beta})$ here replaces $(\mathbf{z} - \hat{\mu}\mathbf{1})$ in OK; this is again the residual from the GLS model.

UK weights (regression formulation)

From this we derive the vector of kriging **weights**, i.e. what to multiply each observation by in the weighted sum (prediction):

$$\lambda^T = \mathbf{c}_0^T \cdot \mathbf{C}^{-1} - \mathbf{x}_0^T \cdot (\mathbf{Q}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{Q})^{-1} \mathbf{Q}^T \cdot \mathbf{C}^{-1}$$

Note: derive this by collecting all the terms that multiply the observed values $\mathbf{z}$ in the OK prediction.

UK prediction variance (regression formulation)

And similarly for the kriging variance:

$$\mathrm{Var}(\hat{Z}_0 - Z_0) = c_{00} - \mathbf{c}_0^T \cdot \mathbf{C}^{-1} \cdot \mathbf{c}_0 + \frac{\mathbf{x}_0 - (\mathbf{c}_0^T \cdot \mathbf{C}^{-1} \cdot \mathbf{Q})}{\mathbf{Q}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{Q}}$$

Note that this variance includes that due to the global trend; we can see this from the presence of the $\mathbf{Q}$ (design) matrix, which includes the coordinates of each point.

For **block kriging**, replace c_{00} with c_{BB}, the average within-block variance (which will in general be smaller than the at-point variance).

Computing the experimental semivariogram for UK

- The semivariances γ are based on the **residuals**, not the original data, because the **random field** part of the spatial structure applies only **after** any trend has been removed.

- How to obtain?

  1. Calculate the **best-fit surface**, with the same base functions to be used in UK;
  2. **Subtract** the trend surface at the data points from the data value to get residuals;
  3. Compute the **variogram** of the **residuals**.

Note: Some programs, e.g. `gstat`, will do these all in one step.

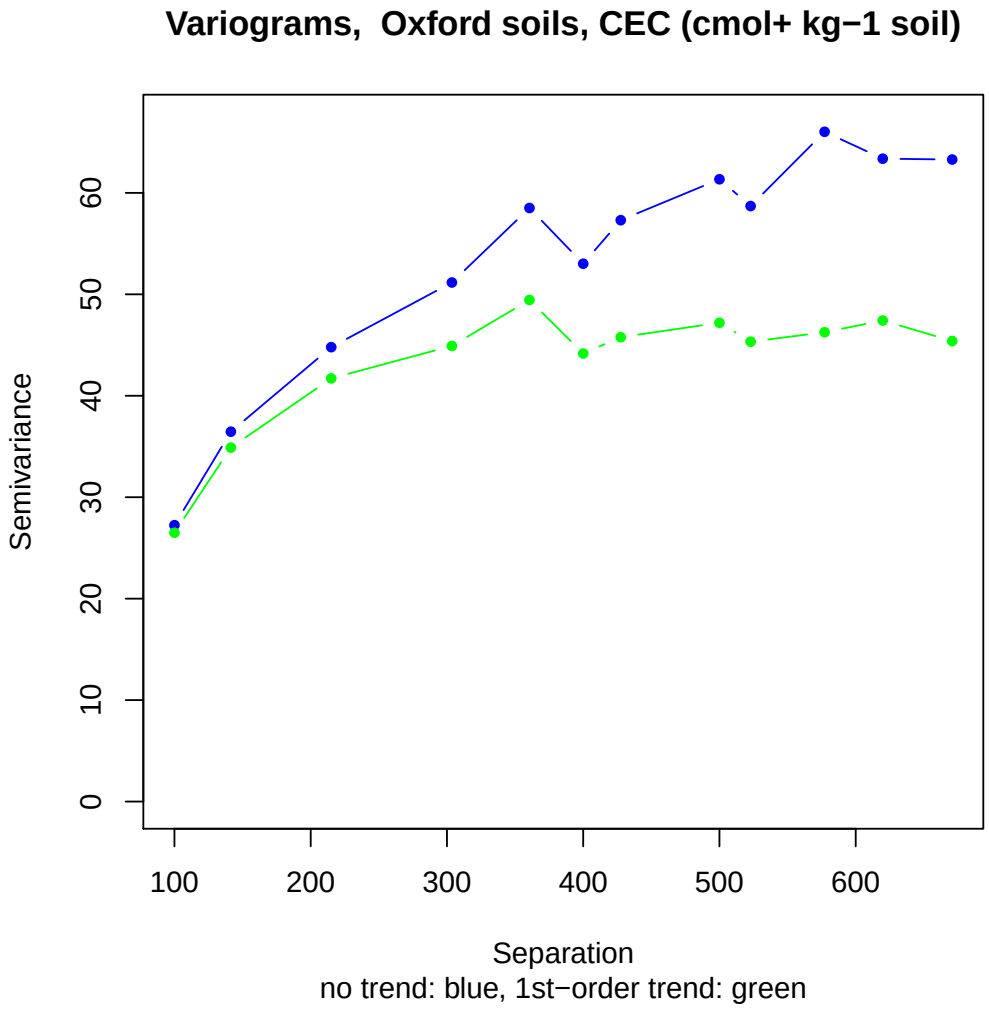
Characteristics of the residual variogram

If there is a strong trend, the variogram model **parameters** for the residuals will be very different from the original variogram model, since **the global trend has taken out some of the variation**, i.e. that due to the long-range structure.

The ususal case is:

- lower **sill** (less total variability)
- shorter **range** (long-range structure removed)

Example original vs. residual variogram



To check your understanding . . .

Q7 : *What are the approximate sill and range of the original (blue) and residual (green) variograms?*

The first-order trend surface in this example explained about 40% of the overall variance in the target variable. How is this reflected in the variogram of the residuals from this surface (as shown)? *Jump to A7* •

Q8 : *What is the relation between the nugget variance of the original (blue) and residual (green) variograms? What should the relation be, in theory?* *Jump to A8* •

The Universal Kriging system (1)

We now present the mathematics of the UK system; it is not necessary to fully understand this in order to apply UK correctly.

The discussion uses similar notation to that for the OK system in the previous lecture.

Solve: $\mathbf{A_U}\lambda_\mathbf{U} = \mathbf{b_U}$, where

$$\mathbf{A_U} = \begin{bmatrix} \gamma(\mathbf{x}_1,\mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_1,\mathbf{x}_N) & 1 & f_1(\mathbf{x}_1) & \cdots & f_k(\mathbf{x}_1) \\ \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \gamma(\mathbf{x}_N,\mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_N,\mathbf{x}_N) & 1 & f_1(\mathbf{x}_N) & \cdots & f_k(\mathbf{x}_N) \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ f_1(\mathbf{x}_1) & \cdots & f_1(\mathbf{x}_N) & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_k(\mathbf{x}_1) & \cdots & f_k(\mathbf{x}_N) & 0 & 0 & \cdots & 0 \end{bmatrix}$$

The upper-left block $N \times N$ block is the spatial correlation structure (as in OK); the lower-left $k \times n$ block (and its transpose in the upper-right) are the trend predictor values at sample points; the rest of the matrix fits with $\lambda_\mathbf{U}$ and $\mathbf{b_U}$ to set up the solution.

The Universal Kriging system (2)

$$\boldsymbol{\lambda}_\mathbf{U} = \begin{bmatrix} \lambda_1 \\ \cdots \\ \lambda_N \\ \psi_0 \\ \psi_1 \\ \cdots \\ \psi_k \end{bmatrix} \qquad \mathbf{b}_\mathbf{U} = \begin{bmatrix} \gamma(\mathbf{x}_1,\mathbf{x}_0) \\ \vdots \\ \gamma(\mathbf{x}_N,\mathbf{x}_0) \\ 1 \\ f_1(\mathbf{x_0}) \\ \vdots \\ f_k(\mathbf{x_0}) \end{bmatrix}$$

The $\boldsymbol{\lambda}_\mathbf{U}$ vector contains the N weights for the sample points and the $k+1$ LaGrange multipliers (1 for the overall mean and k for the trend model), and $\mathbf{b}_\mathbf{U}$ is structured like an additional column of $\mathbf{A}_\mathbf{u}$, but referring to the point to be predicted.

UK system as an extension of OK: (1) Semivariance matrix

$$\mathbf{A_U} = \begin{pmatrix} \Gamma & \mathbf{1} & \mathbf{F} \\ \mathbf{1}^T & 0 & \mathbf{0}^T \\ \mathbf{F}^T & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Only the upper-left is used in OK:

$$\mathbf{A_O} = \begin{pmatrix} \Gamma & \mathbf{1} \\ \mathbf{1}^T & 0 \end{pmatrix}$$

Dimensions:

$$(n+1+k)^2 = \begin{pmatrix} n \times n & n \times 1 & n \times k \\ 1 \times n & 1 \times 1 & 1 \times k \\ k \times n & k \times 1 & k \times k \end{pmatrix}$$

UK system as an extension of OK: (2) Weights and RHS

$$\boldsymbol{\lambda}_\mathbf{U} = \begin{bmatrix} \boldsymbol{\Lambda} \\ \boldsymbol{\psi}_0 \\ \boldsymbol{\Psi} \end{bmatrix} \quad \boldsymbol{\lambda}_\mathbf{O} = \begin{bmatrix} \boldsymbol{\Lambda} \\ \boldsymbol{\psi}_0 \end{bmatrix}$$

$$\mathbf{b}_\mathbf{U} = \begin{bmatrix} \boldsymbol{\Gamma}_0 \\ 1 \\ \mathbf{F}_0 \end{bmatrix} \quad \mathbf{b}_\mathbf{O} = \begin{bmatrix} \boldsymbol{\Gamma}_0 \\ 1 \end{bmatrix}$$

Universal Kriging: Solution & kriging variance

Exactly as for OK:

$$\begin{aligned} \lambda_{\mathbf{U}} &= {\mathbf{A}_{\mathbf{U}}}^{-1}\mathbf{b}_{\mathbf{U}} \\ \sigma^2_U &= {\mathbf{b}_{\mathbf{U}}}^T\lambda \end{aligned}$$

This gives the Lagrange multiplers ψ for the overall mean and the trend, but not the trend surface coefficients β as they would be computed by the linear model. The prediction variance *does* include uncertainty in the trend.

Universal Kriging: Local vs. Global trends

As with OK, UK can be used two ways:

- **Globally**: using **all** sample points when predicting each point
- **Locally**, or in **patches**: restricting the sample points used for prediction to some **search radius** (or sometimes **number of neighbours**) around the point to be predicted

We now discuss the properties of each:

UK over the region

Using **all** sample points when predicting each point:

- Appropriate if there is a **regional trend** across the entire study area
- Agrees with the global computation of the residual variogram
- This gives the same results as Regression Kriging on the coördinates

UK in a neighbourhood

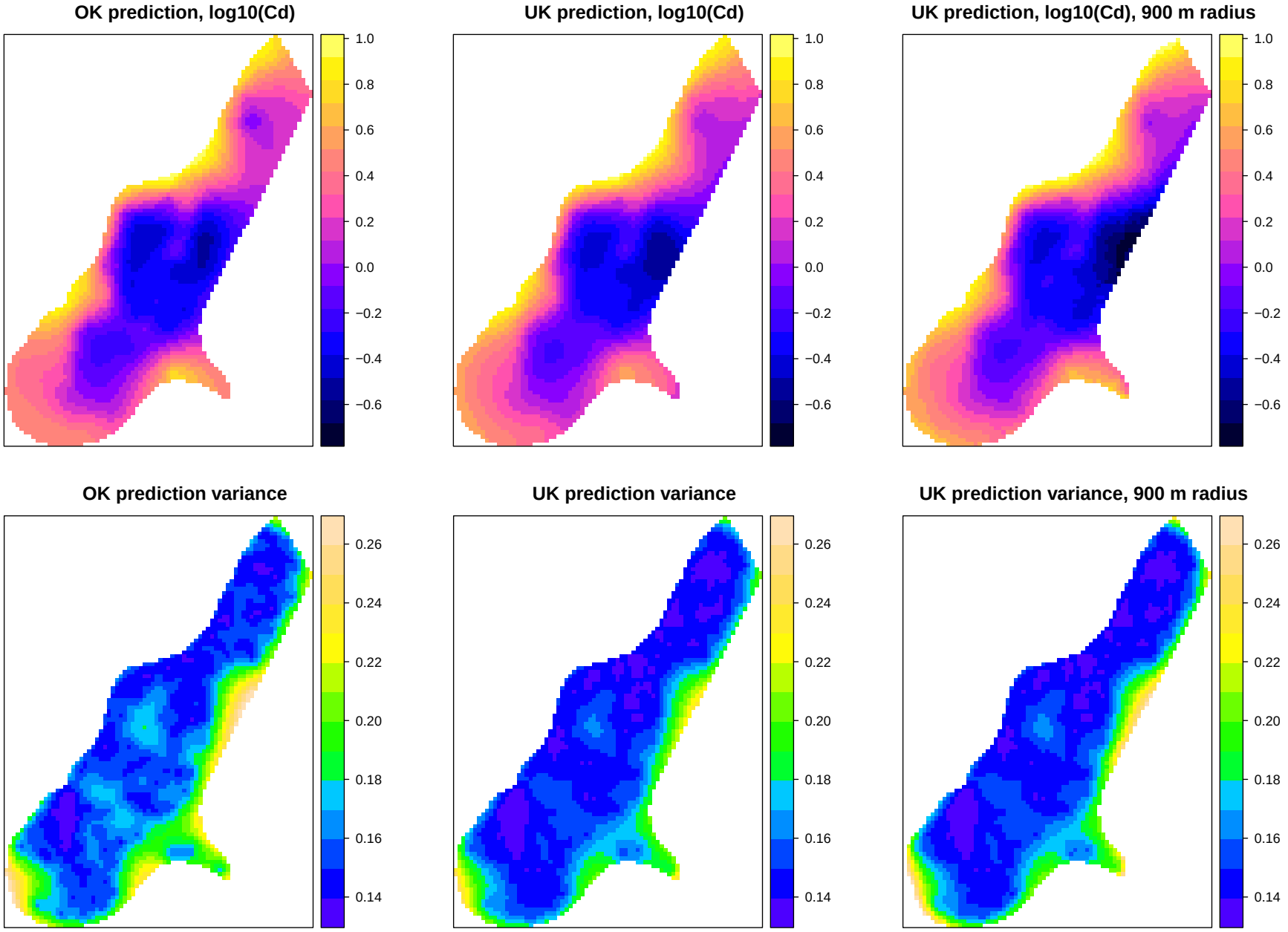
Using just the points in some **neighbourhood**:

- This allows the **trend surface to vary** over the study area, since it is **re-computed at each prediction point**
- Appropriate to smooth away some **local variation in a trend**, difficult to justify theoretically
- Note that the **residual variogram** was not computed in patches, but assuming a global trend
- Leads to some patchiness in the map
- There should be some **evidence of patch size**, perhaps from the original (*not* residual) variogram; this can be used as the search radius.

Compare UK to OK on the next page; the NW-SE trend somewhat modifies the predictions. Note the lower kriging variances with UK, due to the trend surface.

UK with a local trend is intermediate between global UK and OK.

All predictions are shown with the same scale, similarly for variances.


OK prediction, log10(Cd)
UK prediction, log10(Cd)
UK prediction, log10(Cd), 900 m radius
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
OK prediction variance
UK prediction variance
UK prediction variance, 900 m radius
0.26
0.24
0.22
0.20
0.18
0.16
0.14

To check your understanding . . .

Q9 : *What are the major differences in the above figure between OK, global UK, and neighbourhood UK predictions?* *Jump to A9* •

Commentary

UK in a neighbourhood is more difficult to justify than OK in a neighbourhood.

In the case of OK, we have a variogram showing the effective **range** of the local spatial dependence, so that points further than this from the prediction point only contribute to the expected value.

In the case of UK, the residual variogram is computed from the **global**, not local, trend. The range is generally much shorter than the "effective trend" near a prediction point.

Exercise

At this point you should do the **last sections** of **Exercise 5: Predicting from point samples (Part 2)** which is provided on the module CD:

- §2 **Block kriging**
- §3 **Universal kriging**

This should take about an hour.

As in all exercises there are **Tasks**, followed by R code on how to complete the task, then some **Questions** to test your understanding, and at the end of each section the **Answers**. Make sure you understand all of these.

Then do the **self-test** at the end of Exercise 5.

Topic 6: Kriging transformed variables

It may be desireable to **transform** a variable prior to variogram analysis and kriging.

Why transform?

Nothing in the formulation of kriging has made any assumptions about the **feature-space distribution** of the variable to be kriged! The only assumptions are about **spatial stationarity**.

However, the kriging estimator is a **weighted average** of nearby values. A distribution that is not **symmetric** may present problems:

- In a highly **positively-skewed** distribution (often found in earth sciences) the few **high** values will overwhelm all the others.

  Theoretically this may be correct but in practice, with small sample sizes, this can lead to over-prediciton.

- In a **multi-modal** distribution, the weighted average will predict intermediate values that fit none of the modes.

Thus, irregular distributions are often transformed.

Transformations

The problem with transformations for **kriging** in **geographic** space (as opposed to **linear modelling** in **feature** space) is that we must be able to back-transform not only the **prediction**, but also the **prediction variance**.

The most commonly-applied transformations are:

- **Logarithmic** for highly-skewed unimodal distributions
- **Normal-score** or **rank-order** for irregular histograms

We will only deal with the logarithmic transformation here.

Lognormal kriging

1. The target variable is transformed as the natural logarithm: $\mathcal{Y}(\mathbf{x})_i = \log z(\mathbf{x}_i)$; this should be then approximately symmetrically distributed.

2. **Model** and **interpolate** with the transformed variable (OK, block kriging, UK, KED, ...); at each point we get a predicted values $\hat{Y}(\mathbf{x}_0)$ instead of $\hat{Z}(\mathbf{x}_0)$

3. **Back-transform** to original units.

Note: the last step is optional if the transformed estimates can be used directly. For example, a regulatory threshold may be expressed directly in transformed units.

Back transformation of lognormal kriging

For the kriged estimate $\hat{Y}(\mathbf{x}_0)$, we want the back-transfomed estimate $\hat{Z}(\mathbf{x}_0)$. We can not just exponentiate the estimate, because the estimate is a weighted **sum** (not product!) of logarithms.

The back transformation is considered for two cases:

1. Simple Kriging (i.e. mean was **known a priori**)

2. Ordinary Kriging (i.e. mean was **estimated** with the prediction)

The derivations of these equations are found in:

Journel, A.G., 1980. *The lognormal approach to predicting local distributions of selective mining unit grades*. Mathematical Geology, 12(4): 285-303.

Back-transformation of log-SK

For SK, where the mean μ is known:

$$\hat{Z}(\mathbf{x}_0) = \exp\left(\hat{Y}(\mathbf{x}_0) + \frac{\sigma^2_{\mathrm{SK}}(\mathbf{x}_0)}{2}\right)$$

Note that the estimate is **increased** always, because the prediction variance σ^2_{SK} must be positive.

$$\mathrm{var}[\hat{Y}(\mathbf{x}_0)] = \mu^2 e^{\sigma^2_{\mathrm{SK}}}\left[1 - \exp\left(-\sigma^2_{\mathrm{SK}}(\mathbf{x}_0)/2\right)\right]$$

Note that the prediction variance now depends on the mean.

Back-transformation of log-OK

For OK, where the true spatial mean μ is not known:

$$\hat{Z}(\mathbf{x}_0) = \exp\left(\hat{Y}(\mathbf{x}_0) + \frac{\sigma^2_{\mathrm{OK}}(\mathbf{x}_0)}{2} - \psi\right)$$

Note that the estimate is increased by the prediction variance σ^2_{OK} must be positive; but also notice that the **LaGrange multiplier** decreases the prediction.

For OK, it is **not possible to back-transform the variance** in the original units, because μ is unknown (see formula for SK variance back-transformation, above).

Topic 7: Kriging with External Drift (KED)

This is a mixed interpolator that includes feature-space predictors that are not geographic coordinates.

The mathematics are exactly as for UK, but the *base functions* are different.

In UK, the base functions refer to the **grid coordinates**; these are by definition known at any prediction point.

In KED, the base functions refer to some **feature-space covariate** ...

- ... measured at the sample points (so we can use it to set up the predictive equations) and
- **also known at all prediction points** (so we can use it in the prediction itself).

Base functions for KED

There are two kinds of feature-space covariates:

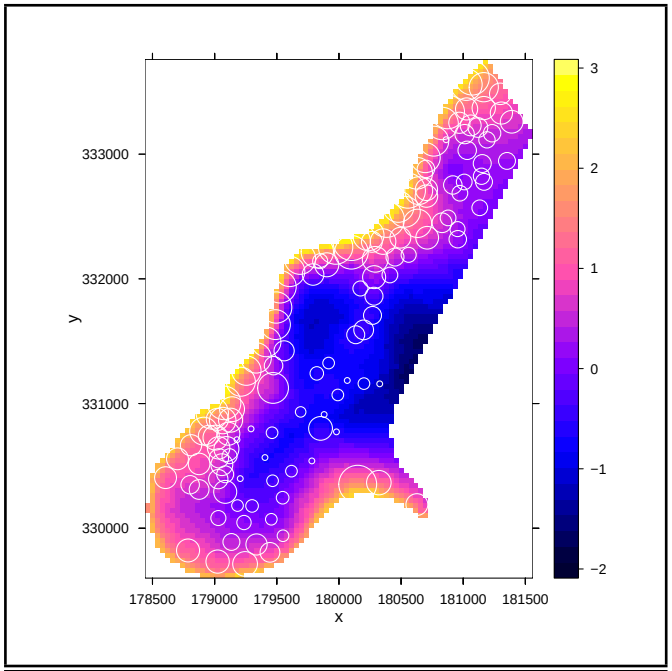
1. **strata**, i.e. factors. Examples: soil type, flood frequency class
   - Base function: $f_k(\mathbf{x}) = 1$ iff sample or prediction point $\mathbf{x}$ is in class k, otherwise 0 (class indicator variable)

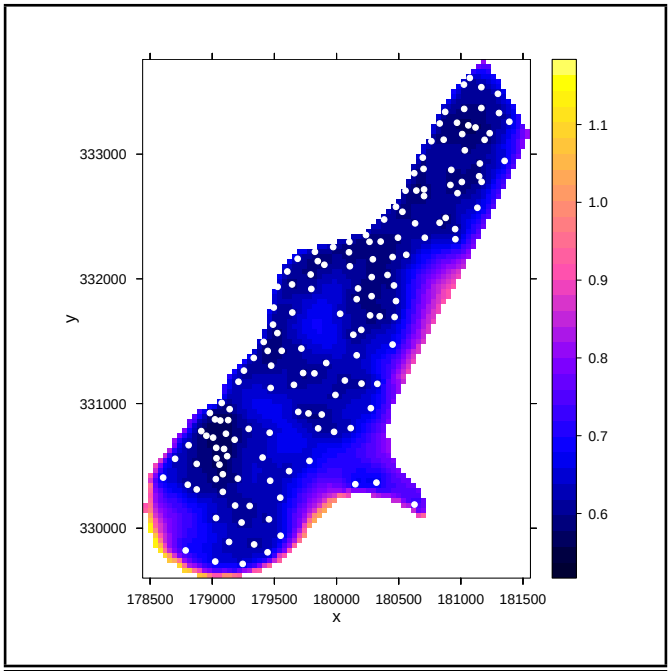
2. **continuous covariates**. Examples: elevation, NDVI
   - Base function: $f_k(\mathbf{x}) = v(\mathbf{x})$, i.e. the value of the predictor at the point.

Note that $f_0(\mathbf{x}) = 1$ for all models; this estimates the global mean (as in OK).

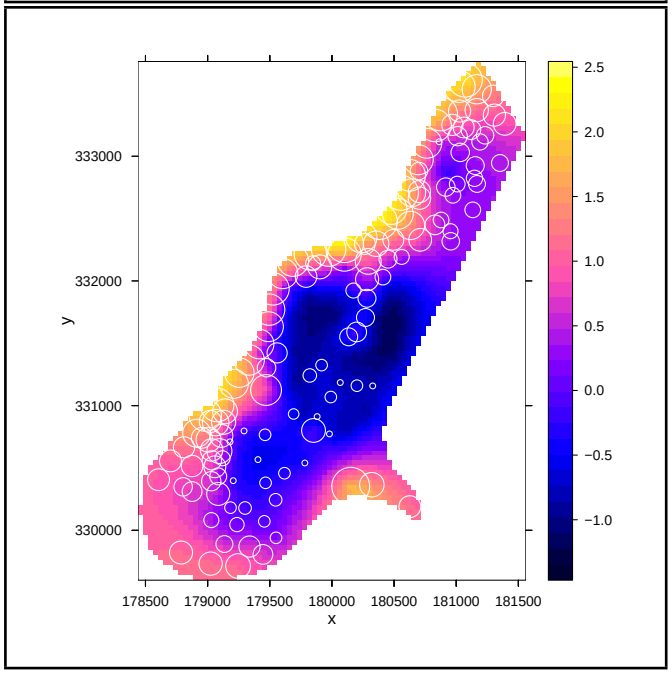
Compare KED to OK on the next page; the distance from the river somewhat modifies the predictions. Note the lower kriging variances with KED; this because the residual variogram has lower sill.

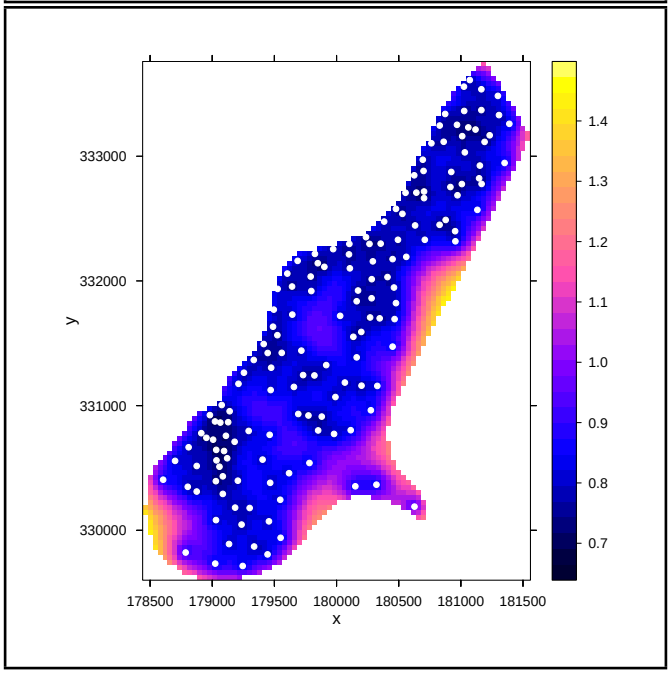
KED (distance to river)





OK





Topic 8: Regression Kriging (RK)

Regression Kriging (RK), also called "kriging after de-trending" models the *trend* (geographic or feature space) and its *residuals* separately.

UK/KED vs. RK (1)

Two methods for incorporating a trend in geographic or feature space:

1. **UK or KED**: Compute trend along with residuals in one UK/KED system
   - UK/KED gives a combined kriging variance
   - Recall: kriging systems are the same, the **base functions** are different

2. **RK**
   (a) Calculate **trend** (or **strata**) → prediction variance of linear model
   (b) subtract trend to get **residuals**
   (c) **model & krige residuals** using Simple Kriging (SK) with known mean of residuals $= 0$; → kriging variance of residuals
   (d) **add trend back** to get estimate
   (e) **sum the two prediction variances** at each point to get the overall error

UK/KED vs. RK (2)

- The coefficients β of the global trend surface are *not* explicit in the UK/KED system; they are *implicit* in the solution and will affect the λ_i at each prediction point
- By contrast, in RK we first find the best trend surface, by fitting coefficients β required by the linear model; then we model the residuals with OK
- The variable being kriged is different with the two approaches: original data values (OK) vs. residuals after fit (RK)
- The λ are different with the two approaches
- There are more ψ (LaGrange multipliers) in UK, and ψ_0 is different in any case
- The RK system does *not* account for the variance of the feature-space model; in the UK/KED system this is built-in.

Naïve RK : Using Ordinary Least Squares (OLS)

- The trend is computed by Ordinary Least Squares (OLS).
- This assumes **no correlation** between residuals ("errors" in statistical terminology):
- Computation:

$$z_{\mathrm{RK}}(\mathbf{x}_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(\mathbf{x}_0) + \sum_{i=0}^{n} \lambda_i z(\mathbf{x}_i)$$

- The two terms are the trend and local interpolation:
  - p is the degree of the trend surface
  - the $\hat{\beta}_k$ are estimated by OLS and multiply the covariates q_k evaluated at the prediction point
  - n is the number of sample points
  - the λ_i are estimated from the SK system of the residuals and multiply the sampled value of the target variable z at each sample point.

Problems with the naïve RK approach (1)

- Once the trend is removed, the residuals are supposed to be **independent samples** from the **same error distribution**.

- However, we have **evidence** from their variogram that **the regression residuals are in fact correlated** – this is the reason for RK instead of just using the trend.
  - Goodness-of-fit measure (R^2) of the OLS trend is too optimistic
  - The trend surface coefficients may be wrong
  - → the residuals may be wrong
  - → the variogram of the residuals may be poorly-modelled
  - → the SK of the residuals may be wrong.
  - → the prediction may be wrong.

- Four wrongs don't make a right . . . what do we do?

Problems with the naïve RK approach (2)

- In practice, if the sample points are **well-distributed** spatially (not clustered), the correlated residuals will be about the same everywhere, so that the OLS fit is satisfactory and naïve RK will give good results.

- *However*, if the sample points are **clustered** in some parts of the map (or feature space), we may well have *mis-estimated the regression coefficients* because we didn't weight the sample points.
  - N.b. This can happen in non-spatial regression also, where observations in feature space are clustered at certain predictor values.

- In particular, a large number of close-by points with similar values (as is expected by spatial dependence) will "pull" a trend surface or regression towards them

- This will tend also to lower the R^2 (supposed goodness-of-fit) → over-optimistic prediction variance
  - Thought experiment: add many close-by observations to one of the sample points; if their data values are quite similar (as expected if there is spatial correlation) these will all be well-fit, thereby decreasing the R^2 arbitrarily.

A more sophisticated RK approach: GLS/RK

- Use Generalised Least Squares (GLS), which allows a **covariance structure between residuals** to be **included directly in the least-squares solution** of the regression equation.

- This is a special case of **Weighted Least Squares** (WLS), in this case where samples are weighted according to the spatial structure using the theory of random fields.

- The GLS estimate of the regression coefficients is:

$$\hat{\beta}_{gls} = (q^T \cdot C^{-1} \cdot q)^{-1} \cdot q^T \cdot C^{-1} \cdot z$$

  where:

  * z is the data vector (observations)
  * q is the *design matrix*;
  * C is the *covariance matrix* of the (spatially-correlated) *residuals*, i.e. deviations from the trend.

GLS/RK (2)

Example: for a first-order trend, q consists of three column vectors:

$$q = [1 x y]$$

where x is one geographic coördinate and y the other.

Reference: Cressie, N., 1993. *Statistics for spatial data*. John Wiley & Sons, New York.

- The covariance structure must be known (or modelled).
- If there is no spatial dependence among the errors, C reduces to $I\sigma^2$ and the estimate to OLS.

Problem implementing GLS/RK

- The covariance structure refers to the *residuals*, but we can't compute these until we fit the trend . . . but we need the covariance structure to fit the trend . . . "which came first, the chicken or the egg?"

- In particular, **the covariance structure must be modelled from the variogram of residuals**; but the residuals can only be obtained **after the trend has been computed**.

- These are solved together in UK or KED, but not in RK; here an **iterative** approach is used.

Procedure with GLS/RK

1. Compute the **OLS trend** either on geographic coordinates (as in UK) or on some feature-space covariates (as in KED).

2. Subtract OLS trend from sample points to obtain **OLS residuals**.

3. Model the **covariance function** C of OLS residuals.

4. Compute the **GLS trend** on the geographic- or feature-space covariates using covariance function C to weight the observations.

   - So the trend is modelled **twice**, first with OLS just to get the residuals to model the spatial covariance, and second with GLS using this modelled covariance.

5. Subtract GLS trend from sample points to obtain **GLS residuals**.

6. Model the **semi-variance function** $\gamma(\mathbf{h})$ of GLS residuals ...

7. **Predict GLS residuals** at prediction points by SK ($\mu \equiv 0$)

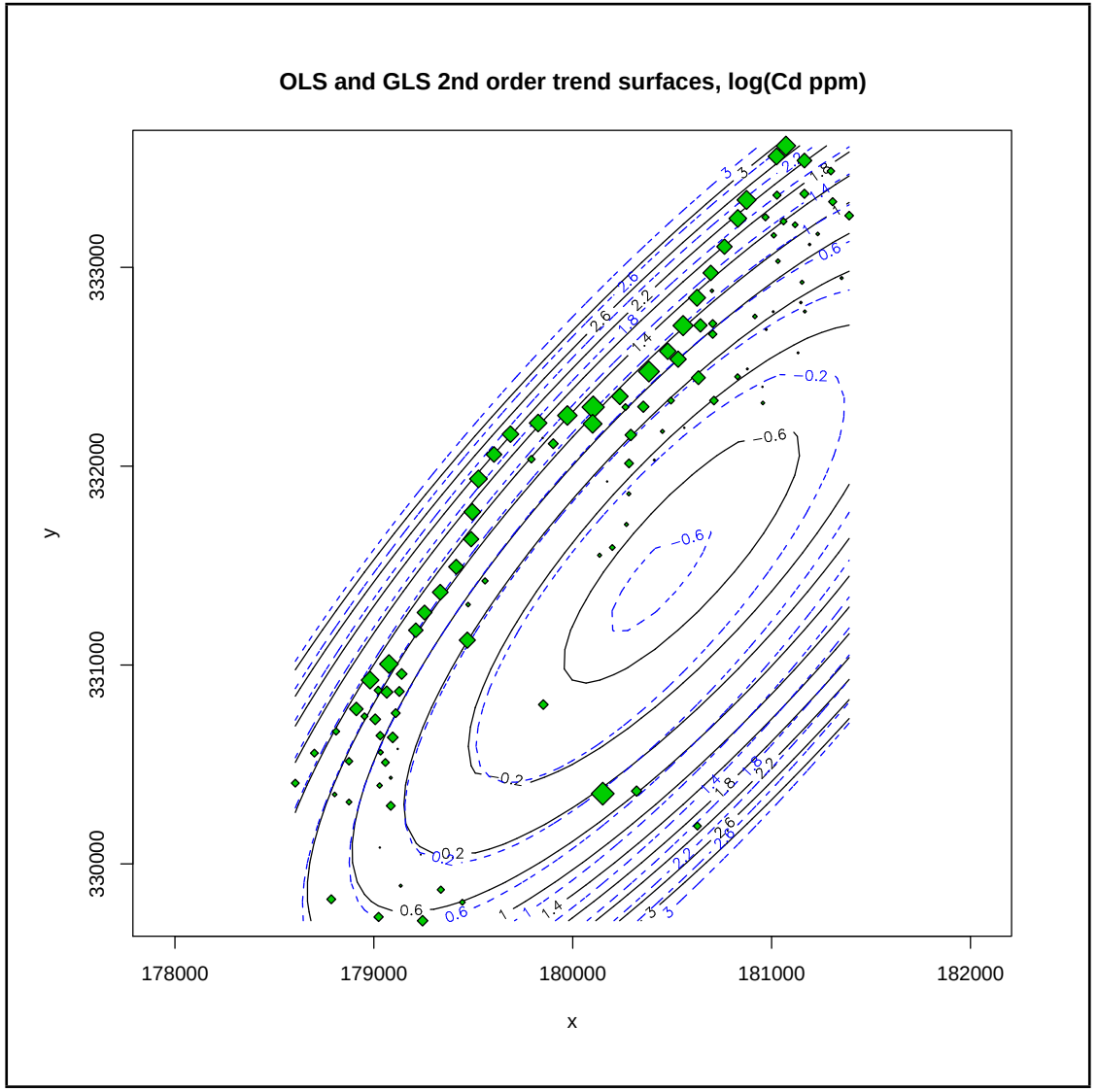
8. **Predict GLS trend** at prediction points from the GLS regression equation

9. **Add** predicted GLS trend and predicted GLS residuals at prediction points
   - → **final prediction**

10. **Add** GLS prediction variance and SK kriging variance at prediction points
    - → **final kriging variance**

Visualizing GLS/RK (1)

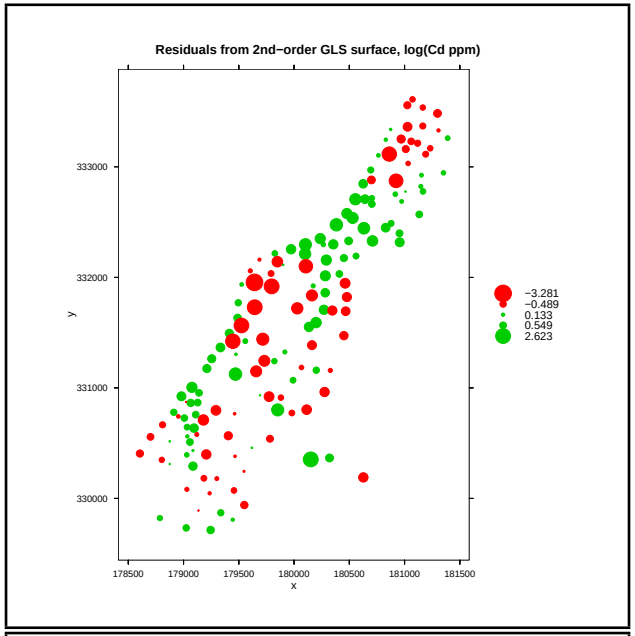
OLS and GLS trend surfaces compared

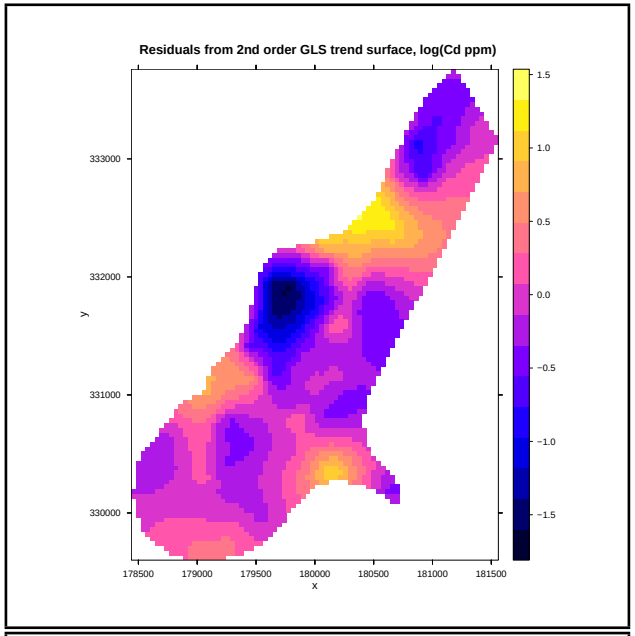


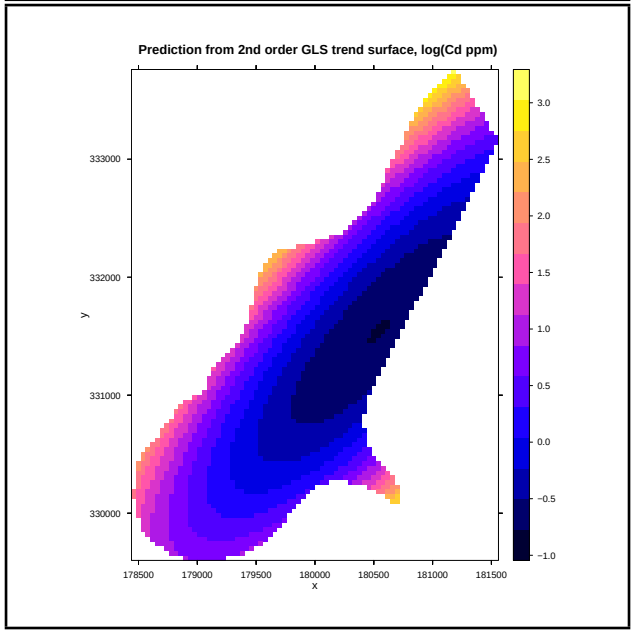
Visualizing GLS/RK (2)

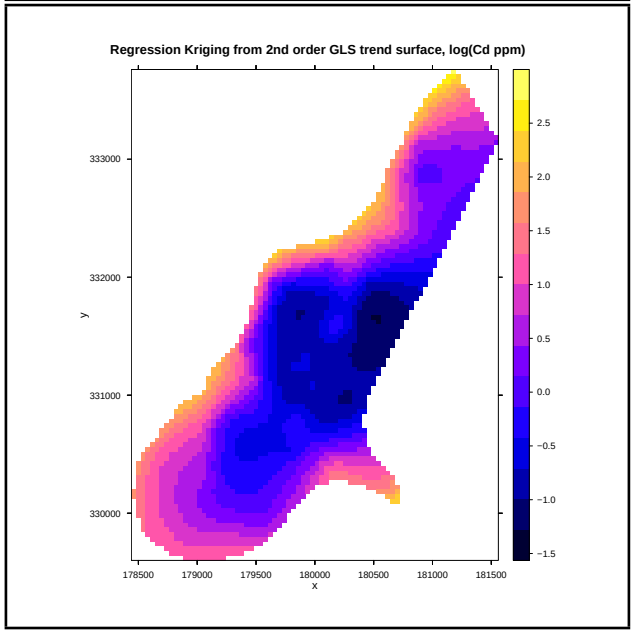
Residuals, SK

TS, RK



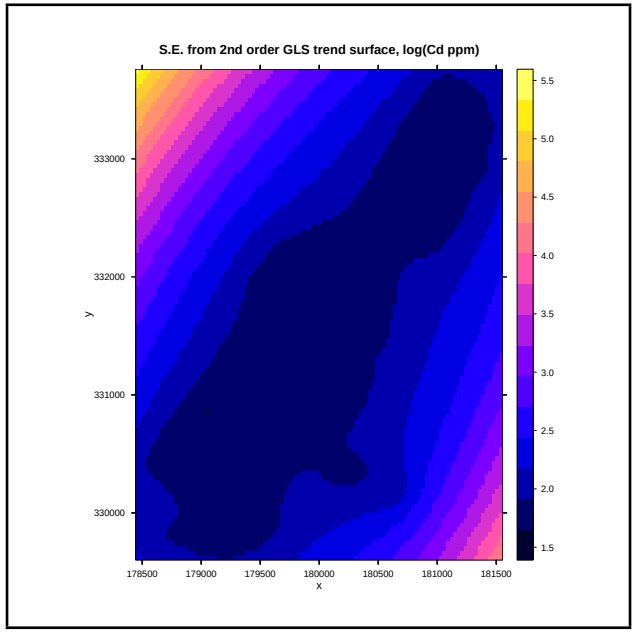


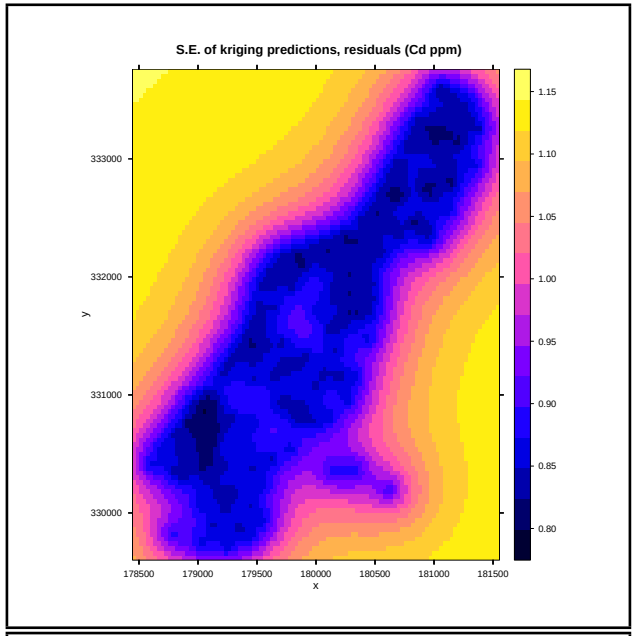




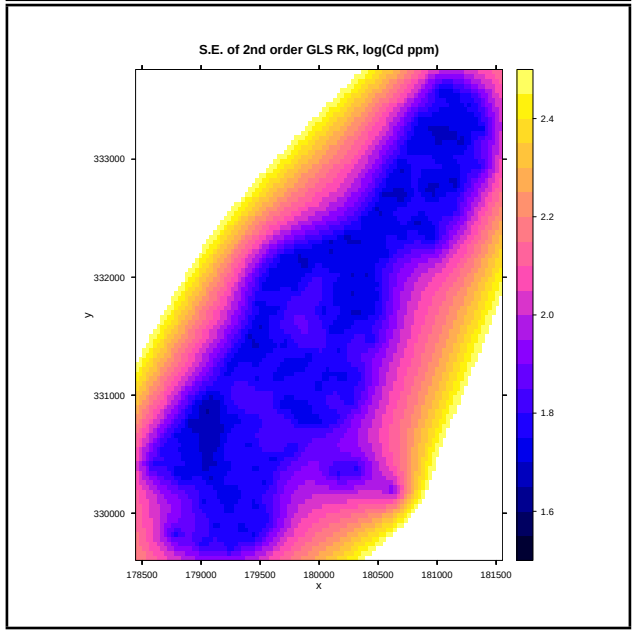
Visualizing GLS/RK (3)

Trend surface, SK S.E.





Regression Kriging S.E.



Answers

Q1 : *Consider a regular grid over some area; measure the data values and compute the ordinary average.*

If there is spatial dependence, what would happen to this average if we now make many observations near one of the grid points (i.e. a cluster), and not the others? •

A1 : *The average will get closer and closer to the average of the points in the cluster, which, because of spatial autocorrelation, will not necessarily be the average over the whole area.* *Return to Q1* •

Answers

Q2 : *How is the covariance matrix* $\mathbf{C}$ *computed?* •

A2 : *This is derived from the modelled* **covariance function**: *each matrix entry* c_{ij} *is computed from the* **separation** h_{ij} *between the two observations* z_i *and* z_j *(in the simple case, the distance between them), which is then used as an argument to the covariance function.*

For example, for an exponential model: $c_{ij} = ce^{(-h_{ij}/a)}$ *Return to Q2* •

Q3 : *How is the covariance vector* $\mathbf{c_0}$ *computed?* •

A3 : *As for the covariance matrix, but the separations are between each known point* z_j *and the point to be predicted,* z_0. *Return to Q3* •

Answers

Q4 : *Does the kriging variance depend on the* **data values** *at the sample points, or on the predicted data value? How can you see this from the equation?* •

A4 : *No and no. There is* **no** *reference to any* **data value** *on the right-hand side of the equation.*

Thus the kriging variance depends **only** *on the (1)* **sample point configuration** *and the (2)* **spatial covariance model**. *Return to Q4* •

Answers

Q5 : *What are the differences between the* **predictions** *with different block sizes?* •

A5 : *There not much difference, but as the block size increases the extremes are a bit softened; i.e. the hot and cold spots are a bit less pronounced.* *Return to Q5* •

Answers

Q6 : *What are the differences between the* **kriging variances** *with different block sizes?* •

A6 : *There is a very large reduction in variance with BK; increasing the block size reduces this still more but not as dramatically.* *Return to Q6* •

Answers

Q7 : *What are the approximate sill and range of the original (blue) and residual (green) variograms?*

The first-order trend surface in this example explained about 40% of the overall variance in the target variable. How is this reflected in the variogram of the residuals from this surface (as shown)? •

A7 : *Sill: 65 CEC-units squared (original) vs. 45 65 CEC-units squared (residual)*

Range: 550 m (original) vs. 350 m (residual).

The trend surface removed much of the total variability, thereby lowering the total sill by about 20/65 = 30%. The global trend model acts across the entire area, so that long-range variability due to the trend is removed, hence the shorter range of residuals. *Return to Q7* •

Q8 : *What is the relation between the nugget variance of the original (blue) and residual (green) variograms? What should the relation be, in theory?* •

A8 : *The nugget variance is almost the same, approx. 28 CEC-units squared. This agrees with theory: a trend surface can not take out the variability at a single point.* *Return to Q8* •

Answers

Q9 : *What are the major differences in the above figure between OK, global UK, and neighbourhood UK predictions?* •

A9 : *The NW-SE trend somewhat modifies the OK predictions. In particular, notice the lower UK predictions in the area near (180 600, 331 800). This is because the trend surface, which dips towards the SE, is not corrected by nearby observations. Also notice the smaller "halo" of high values in UK near the two high observations on the river bend near (180 200, 330 200). This is because the SE-dipping trend surface predicts low values in this area; the two high points compensate for this in their neighbourhood. In OK they have more influence because there is no contrary trend.*

UK has lower kriging variances than OK, especially at prediction locations with few observations nearby, due to the trend surface.

UK with a local trend is intermediate between global UK and OK in both predictions and variances. In some cases the local trend is even stronger than for UK, for example at the previously-mentioned "cold spot" centred on (180 600, 331 800). *Return to Q9* •