# Conceptual basis of geostatistics

D G Rossiter

Cornell University, Section of Soil & Crop Sciences

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

June 23, 2022

# Table of Contents

# Concept

Attributes are distributed over space due to a combination of **processes**:

- Process 1: due to **other spatially-distributed attributes**
  - e.g., elevation → temperature; land cover class → vegetation density
- Process 2: due to a **spatial trend** (a function of the coördinates)
  - e.g., distance from source → rock stratum thickness
- Process 3: due to **local effects**
  - e.g., diffusion from a point source → disease/pest incidence in a crop field

# Autocorrelation

- "Auto" = self, i.e., an attribute correlated with itself
- Compare the attribute of one instance with that of another instance of the *same* attribute
- Define how to compare:
    - time: **temporal** autocorrelation (e.g., of time series)
    - space: **spatial** autocorrelation

# Equation

$$Z(\mathbf{s}) = Z^*(\mathbf{s}) + \varepsilon(\mathbf{s}) + \varepsilon'(\mathbf{s})$$

- $(\mathbf{s})$ a location in space, designated by a **vector** of coördinates
- $Z(\mathbf{s})$ **true** (unknown) value of some property at the location
- $Z^*(\mathbf{s})$ **deterministic** component, due to some known or modelled **non-stochastic** process
- $\varepsilon(\mathbf{s})$ **spatially-autocorrelated stochastic** component
- $\varepsilon'(\mathbf{s})$ pure (“white”) **noise**, no structure

# Two types of the deterministic components $Z^*(\mathbf{s})$

- as function of **spatially-distributed covariates** (Process 1)
- as a **trend**: a function of the coördinates (Process 2)
- these can have the same mathematical structure and be determined by the same algorithms
  - covariates: multiple regression, random forests ...
  - trend: low-degree polynomials, generalized additive models, thin-plate splines

# How do we fit the universal model?

$Z^*(\mathbf{s})$

- by a **process** model (simulation)
- by an **expert** or heuristic model, e.g., **stratification**, e.g., into map units (polygons)
- by an **empirical-statistical** ("regression") model in **feature** ("attribute") space
- by an **empirical-statistical** model in **geographic** space ("trend surface")

$\varepsilon(\mathbf{s})$

- as a realization of a spatially-correlated **random field** using geostatistics

$\varepsilon'(\mathbf{s})$

- can not not be modelled, but can be quantified → prediction **uncertainty**

# Table of Contents

# A 2D geographic example

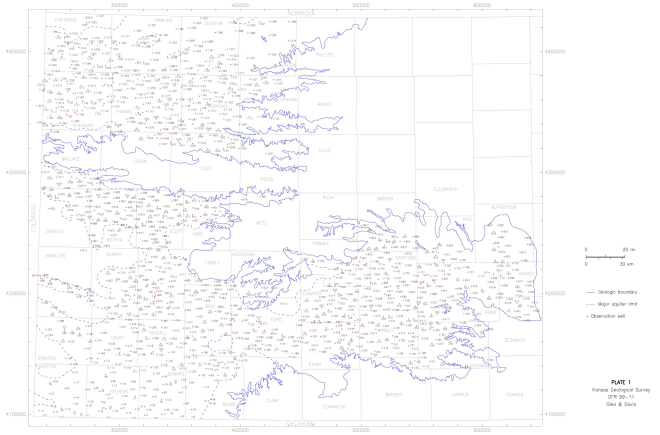
- Source: Olea, R. A., & Davis, J. C. (1999). *Sampling analysis and mapping of water levels in the High Plains aquifer of Kansas* (KGS Open File Report No. 1999-11). Lawrence, Kansas: Kansas Geological Survey.[1]
- **attribute**: elevation (US feet) above sea level of the top of an aquifer in Kansas (USA)[2]; NAVD 88 vertical datum
- **georeference**: observed at a large number of wells, position UTM Zone 14N, NAD83 meters
- Q: **What determines** the spatial variation?
- Q: How can we **model** this from the observations?
- Use the fitted model to **predict** at unsampled locations, either individual locations (proposed new wells) or over a fine-resolution grid

[1] Retrieved from `http://www.kgs.ku.edu/Hydro/Levels/OFR99_11/`

[2] `http://www.kgs.ku.edu/HighPlains/HPA_Atlas/`

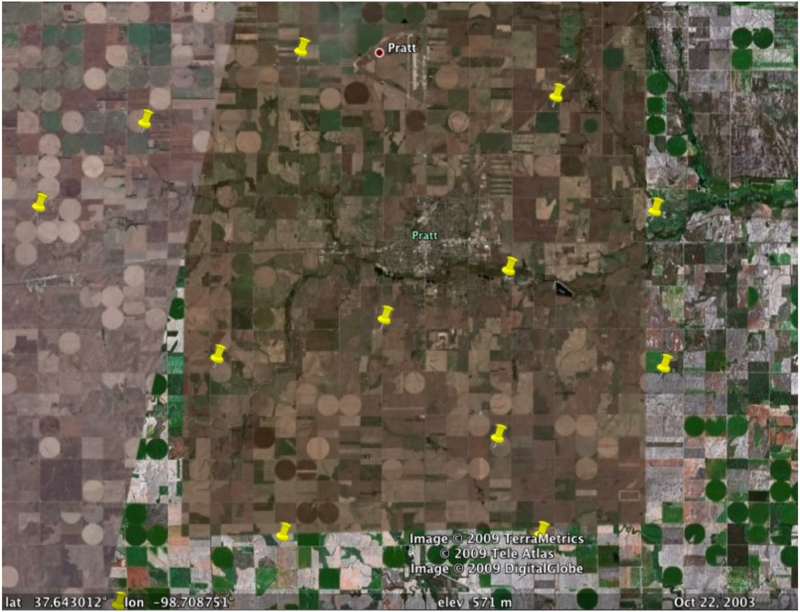
# Observations

MEASURED OBSERVATION WELLS IN THE HIGH PLAINS AQUIFER, JANUARY 1999

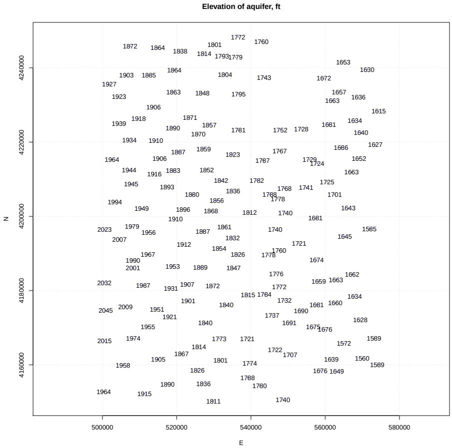



Reality: $Z(\mathbf{s}) = Z^*(\mathbf{s}) + \varepsilon(\mathbf{s}) + \varepsilon'(\mathbf{s})$

# Some well sites on imagery background

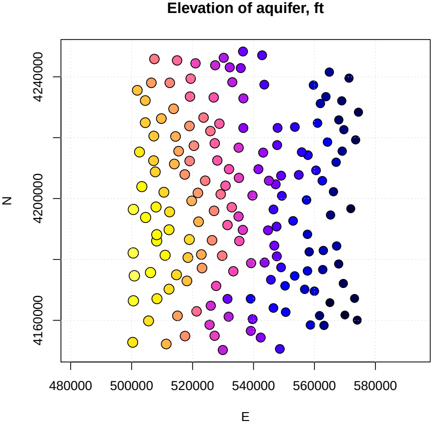


# Observations – text display



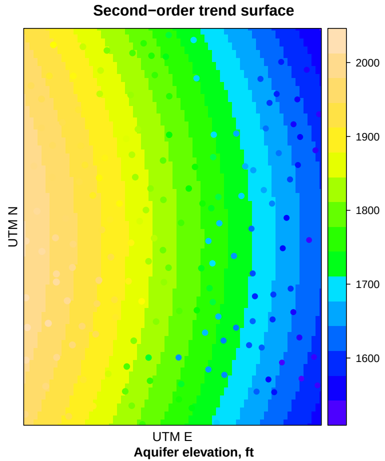

2D georeference, one attribute

# Observations – postplot



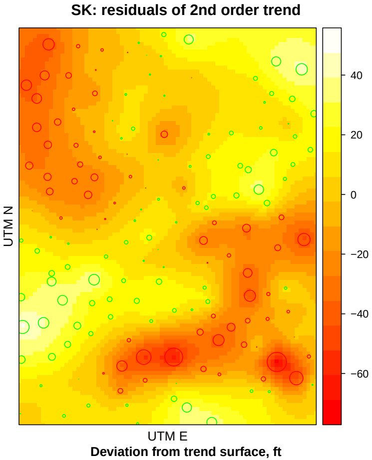

Q: How to divide these observations of $Z(\mathbf{s})$ into $Z^*(\mathbf{s})$, $\varepsilon(\mathbf{s})$, and $\varepsilon'(\mathbf{s})$?

# (1) A deterministic trend surface $Z^*(\mathbf{s})$




**process**: dipping and slightly deformed sandstone rock
**modelled** with a 2nd-order polynomial (empirical-statistical model) **trend surface**

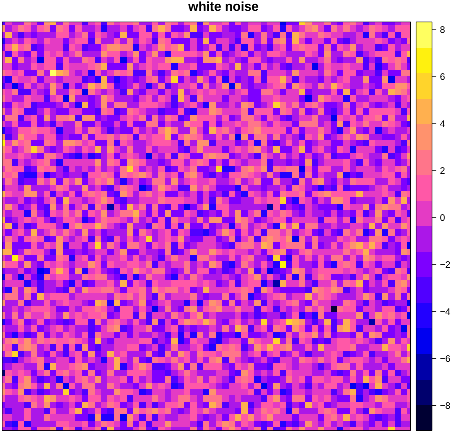
# (2) A spatially-correlated random field $\varepsilon(\mathbf{s})$




**process**: local variations from trend
**modelled** by **variogram modelling** of the random field and simple **kriging**

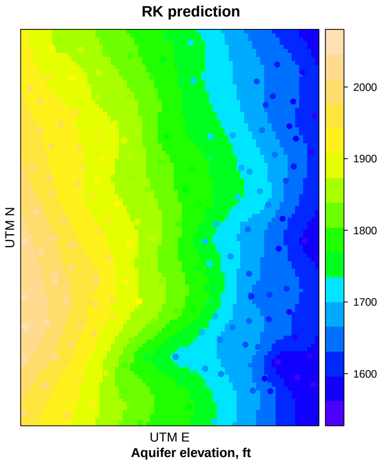
# (3) White noise $\varepsilon'(\mathbf{s})$

We do not know! but **assume** and **hope** it looks like this:

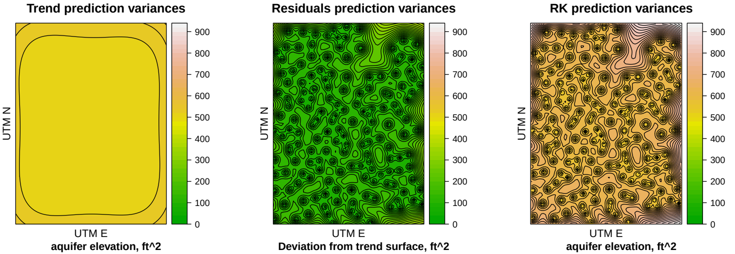



Quantified as **uncertainty** of the other fits

# Model with both trend and local variations $Z^*(\mathbf{s}) + \varepsilon(\mathbf{s})$

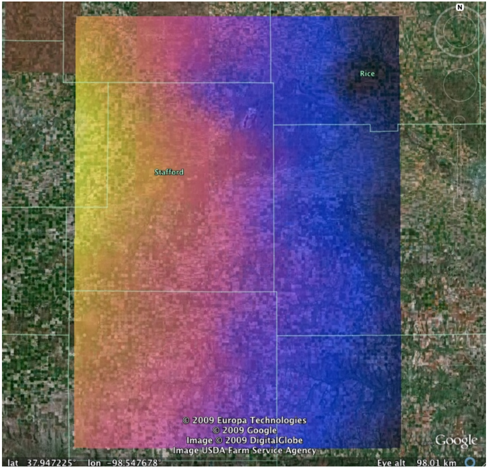


# Unexplained variation $\varepsilon'(\mathbf{s})$




Computation depends on model form (here: Generalized Least Squares trend + Simple Kriging of GLS residuals)

# Model predictions shown on the landscape




Google Earth, PNG ground overlay, KML control file

# Table of Contents

# Example: soil spatial variation (1)

$(\mathbf{s})$ area of interest; **discretized** and considered as **blocks** with some finite **support**

$Z(\mathbf{s})$ true block mean and within-block variation

$Z^*(\mathbf{s})$ effect of **soil-forming factors** that can be modelled

- *same* factors → *same* soil properties: Jenny (1941) '*clorpt*' model.
- includes strata (thematic maps units), "continuous" fields
- includes regional geographic trends (e.g., climate gradient)

...

# Example: soil spatial variation (2)

...

$\varepsilon(\mathbf{s})$ **spatially-correlated stochastic** component, modelled in **geographic** space

- local deviations from average effect of soil-forming factors
- some part of this is often **spatially-correlated**; this we can model

...

# Example: soil spatial variation (3)

...

$\varepsilon'(\mathbf{s})$ pure ("white") **noise**

- non-deterministic and not spatially-correlated
- includes variation at finer scale than support
- includes sampling and measurement imprecision ("error")

measurement imprecision (all included in "noise"):

- georeferencing / field location
- sampling protocol, sampling procedures
- lab. methods, lab. procedures, lab. quality control

# Table of Contents

# Conceptual issues with the universal model

1. **"Deterministic"** implies that some process always operates the same way with the same inputs.
    - Any deviations are considered noise and included in $\varepsilon(\mathbf{s})$ or $\varepsilon'(\mathbf{s})$.
    - "Deterministic" is **operationally** defined as "we can model it as if it were deterministic"
    - We are *not* really asserting that nature is deterministic.
2. The **spatially-autocorrelated stochastic** component is assumed to be one realization of a spatially-correlated random process
    - This is usually a convenient fiction to allow modelling.
    - It may include a spatially-correlated deterministic component that we don't know how to model.
    - It is a stochastic process, so there is **uncertainty** which is considered pure noise
3. The **"pure noise"** component may also have a structure but at a finer scale than we can measure.
    - It also contains our ignorance about the deterministic process and spatially-correlated random process

# Table of Contents

# Geostatistics

1. Definition
2. Detecting spatial autocorrelation
3. Modelling spatial autocorrelation
4. Predicting from a model of spatial autocorrelation and a set of observations

# Table of Contents

# What is “geostatistics”? (1)

> *“. . . a collection of methods used in the analysis of a* ***particular kind of spatial data*** *in which* ***measured values*** $Y_i$ *at* ***spatial locations*** $u_i$ *can be regarded as* ***noisy observations*** *from an* ***underlying process*** *in* ***continuous space****.”* [3]

Key phrase: “underlying process”. Note **continuous** space but **discrete** (noisy) observations.

---

[3] Ribeiro, P.J. Jr. and Diggle, P.J. (1999) *Bayesian inference in Gaussian model-based geostatistics*. Tech. Report ST-99-08, Dept Maths and Stats, Lancaster University.
`http://www.leg.ufpr.br/geoR/geoRdoc/bayeskrige.pdf`

# What is “geostatitics”? (2)

> *“ In particular, in a geostatistical analysis* ***spatial interpolation*** *or smoothing of the observed values is often carried out by a procedure known as* ***kriging****. In its basic form, kriging involves the construction of a* ***linear predictor for an unobserved value of the process****, and the form of this linear predictor is chosen with reference to the* ***covariance structure*** *of the data as estimated by a data-analytic tool known as the* ***variogram****.”*

Key phrase: “covariance structure” of the process.

# "Geostatistics" – definition

- Inferential statistics about a population with **spatial reference**, i.e. **coördinates**:
  - Any number of dimensions (1D, 2D, 3D . . . );
  - Any geometry;
  - Any coördinate reference system (CRS), including locally-defined coördinates;
  - There must be defined **distance** and **area** metrics.
- Key point: **observations** and **predictions** of the target variable (and possibly co-variables) are made at **known locations** in geographic space.

# Simple case: no deterministic component

- Suppose **no** geographic trend or spatially-distributed covariates
- Then $Z^*(\mathbf{s}) \equiv \mu$, where $\mu$ is the **stationary spatial mean.**
- The universal model:

$$Z(\mathbf{s}) = \mathbf{Z}^*(\mathbf{s}) + \varepsilon(\mathbf{s}) + \varepsilon'(\mathbf{s}) \quad (1)$$

- becomes:

$$Z(\mathbf{s}) = \mu + \varepsilon(\mathbf{s}) + \varepsilon'(\mathbf{s}) \quad (2)$$

This is a model **assumption!**

- The technical term here is **first-order stationarity**; later we relax this assumption.
- We want to model the **structure** of $\varepsilon(\mathbf{s})$ and ignore the **pure noise** $\varepsilon'(\mathbf{s})$
- the noise sets a lower bound on the precision of predictions made with the fitted model.

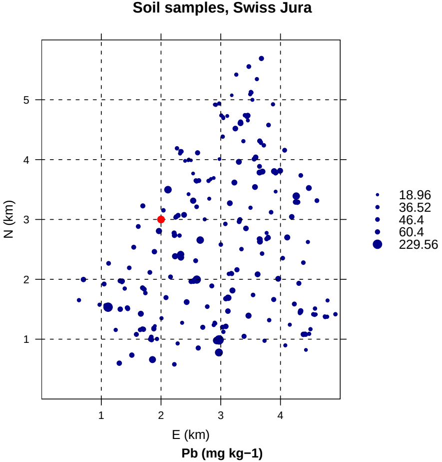
# Table of Contents

# A "point" observation dataset

**Post plot**:

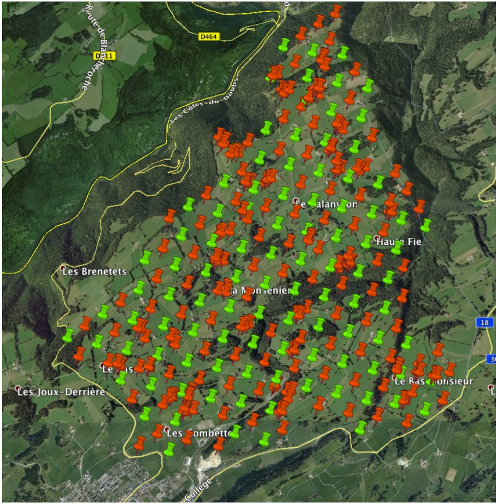
**Symbol size** proportional to **attribute** value

**Axes** are **geographic** coördinate




Q: Is there spatial autocorrelation of the Pb concentrations?

# Observation locations on the landscape



# Detecting local spatial autocorrelation

If there is local spatial autocorrelation, we need to **detect** it (empirically) and then **model** it (mathematically).

- **detection**: h-scatterplot, correlogram or **variogram**
- **modelling**: “authorized” model of spatial covariance

# Point-pairs

- Any two observations in geographic space are a **point-pair.**
  - We know (1) their coördinate $\mathbf{s}$; (2) their **attribute values** (what was measured about them) $z(\mathbf{s})$.
- For an $n$-observation dataset, there are $(n * (n-1)/2)$ **unique point-pairs.**
  - E.g., 150-point dataset has $150 \cdot 149/2 = 11\,175$ pairs!
- Each pair is separated in **geographic space** by a **distance** and (if $>$1D) **direction.**
- Each pair is separated in **feature** (attribute) **space** by the **difference** between their attribute values.

# Key point

- We now have a relation between the “distance” in **feature** (attribute) and the distance in **geographic** spaces.
- Question: Is there any **structure** to this relation? If so, how to **model** (quantify)?
- If so, there is **local spatial dependence** in the attribute.
- Note the relation betwen feature and geographic space was considered **globally** in **trend surfaces**, here we look at **local** and **distance-dependent** relations.

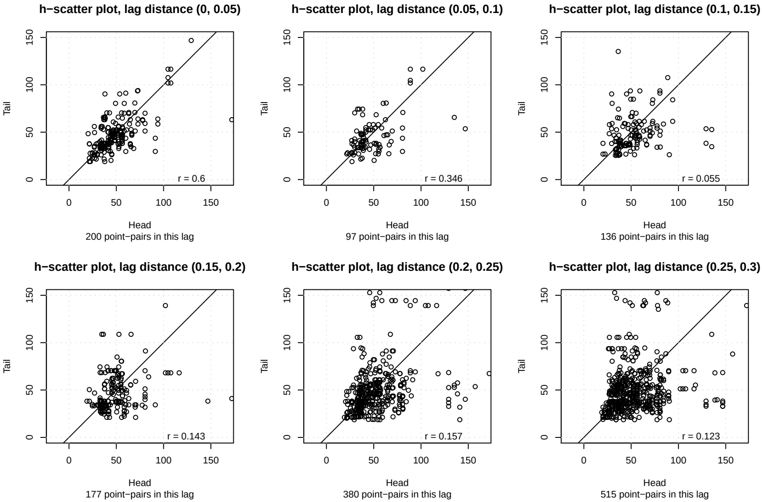
# Semivariance of a point-pair

- *Define* the **semivariance** $\gamma$ of one point-pair as:

$$\gamma(\mathbf{s}_i, \mathbf{s}_j) \equiv \frac{1}{2}[z(\mathbf{s}_i) - z(\mathbf{s}_j)]^2$$

- This quantifies the **difference** between the **attributes** values at the two points.
- Squared because the order of point-pairs is irrelevant
- 1/2 for technical reasons in the kriging equations (see later)

# *h*-scatterplot: correlation between point-pairs




Increasing lag distance $h$ → decreasing linear correlation $r$.

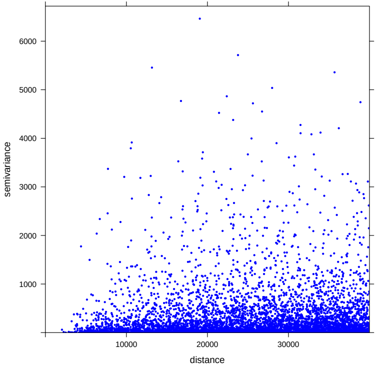
# Evidence of spatial autocorrelation from the *h*-scatterplot

- Point-pairs compared against the 1:1 line (equal values would be on the line)
- More scatter from the 1:1 line → less linear correlation
- If the **sequence of lags** from close to far also shows **increasing scatter** (i.e., decreasing correlation), this is evidence of local spatial autocorrelation.

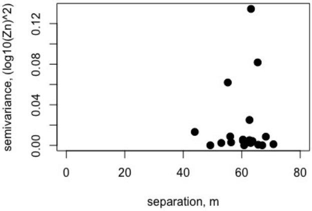
# Variogram cloud

- A scatterplot showing, for *all* point-pairs:

  (**x-axis**) the **separation distance** between the two observations
  (**y-axis**) their **semivariance** slide



# Variogram cloud – detail




```
> (vc <- variogram(logZn ~ 1, meuse,
                   cutoff=72, cloud=TRUE))
       dist        gamma  left right
1  70.83784 1.144082e-03     2     1
2  67.00746 9.815006e-05    11    10
3  62.64982 2.504076e-02    22    21
4  53.00000 2.375806e-03    23    22
5  49.24429 8.749351e-05    26    25
6  62.62587 5.128294e-03    33    32
7  65.60488 6.655118e-04    39    38
8  63.07139 2.403081e-03    72    71
9  63.63961 4.318603e-03    76    75
10 60.44005 4.486439e-03    84     9
11 43.93177 1.326441e-02    87    72
12 65.43699 8.178006e-02    87    80
13 56.04463 8.764773e-03    88    73
14 55.22681 6.198261e-02    88    79
15 60.41523 5.680995e-03   123    58
16 60.82763 5.583388e-05   124    52
17 63.15853 1.344946e-01   138    76
18 56.36488 2.996326e-03   139    77
19 68.24222 8.550172e-03   140    91
```

- Note the anomalous point-pair.
- This is difficult to interpret and model, so we summarize this with an **empirical variogram** (see next).

# Empirical semivariogram - equation

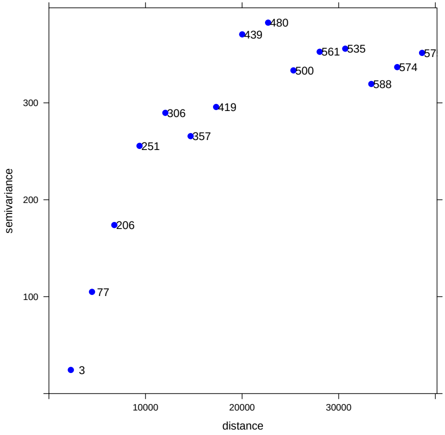
Summarize the cloud as **average semivariance** $\overline{\gamma}(\mathbf{h})$ in some **separation range h**

$$\overline{\gamma}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} [z(\mathbf{s}_i) - z(\mathbf{s}_i + \mathbf{h})]^2$$

- $m(\mathbf{h})$ is the number of **point pairs** separated by vector **h**, in practice some range of separations ("bin")
- these are indexed by $i$
- the notation $z(\mathbf{s}_i + \mathbf{h})$ means the "tail" of point-pair $i$, i.e., separated from the "head" $\mathbf{s}_i$ by the separation vector **h.**

# Empirical semivariogram - graph





# Empirical semivariogram - numerical

```
> (v <- variogram(logZn ~ 1, meuse, cutoff=1300, width=90))
    np       dist      gamma
1   41   72.24836 0.02649954
2  212  142.88031 0.03242411
3  320  227.32202 0.04818895
4  371  315.85549 0.06543093
5  423  406.44801 0.08025949
6  458  496.09401 0.09509850
7  455  586.78634 0.10656591
8  466  677.39566 0.10333481
9  503  764.55712 0.11461332
10 480  856.69422 0.12924402
11 468  944.02864 0.12290106
12 460 1033.62277 0.12820318
13 422 1125.63214 0.13206510
14 408 1212.62350 0.11591294
15 173 1280.65364 0.11719960
```

- np = number of point-pairs in bin
- `dist` = average separation between the point-pairs in bin (here, meters)
- gamma = average semivariance $\overline{\gamma}(\mathbf{h})$ between the point-pairs in bin (here, $\log_{10}Zn^2$)
- Obvious trend: wider separation → larger semivariance

# Separation types

For $>$1D geometries:

- **distance** only: the **omni-directional** variogram
- **distance and angle**: a **directional** variogram
  - includes a **tolerance angle** and/or maximum width

# Is there local spatial autocorrelation?

- **Spatial Dependence**: a relation between semivariance and separation.
- Closer in **geographic** space means closer in **feature** space.
  - i.e., knowing the attribute value at one observation gives some clue about the value at a “nearby” point
  - The closer to known points, the stronger the clue
- Visualize/infer by **plotting the empirical semivariogram**(s).
- If there appears to be evidence, then **model**

# Terminology of spatial autocorrelation

**sill** (also **total** Maximum semivariance at any separation

**range** separation at which the sill is reached or approximated

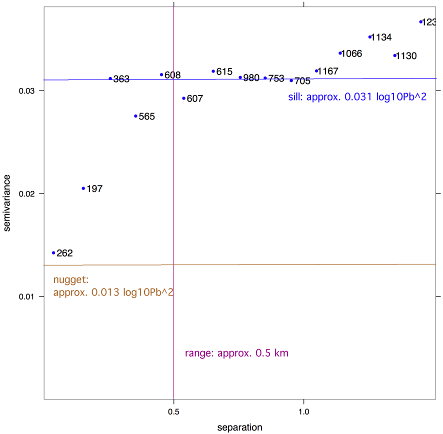
**nugget** semivariance at zero separation (at a point)

**structural sill** (also **partial** sill) the total sill less the nugget
- i.e., the portion due to spatial autocorrelation

**nugget/sill ratio** proportion of total sill due to the nugget, i.e., **unexplainable**

# Annotated empirical variogram

$\log_{10}$Pb, Jura soil samples



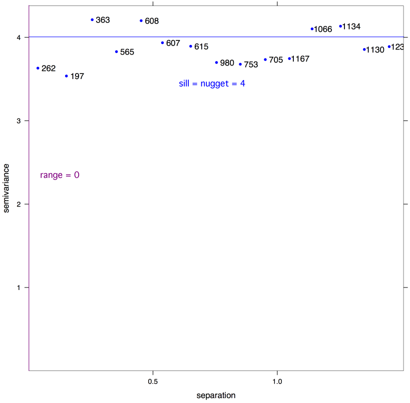

Nugget/sill ratio $\approx 0.42 \rightarrow$ variability not explained $\epsilon'(s)$

# Evidence of spatial autocorrelation from the variogram

The empirical variogram provides **evidence** that there is **local spatial autocorrelation.**

- The **variation** between point-pairs is **lower** if they are **closer** to each other; i.e. the separation is small.
- There is some distance, the **range** where this effect is noted; beyond the range there is no autocorrelation.
- The **relative magnitude** of the **total sill** and **nugget** give the **strength** of the local spatial autocorrelation; the **nugget** represents completely **unexplained** variation.
- If there is no spatial autocorrelation, we have a **pure nugget** variogram.

# Annotated empirical variogram – no spatial autocorrelation




Random fluctuations around sill, due to sampling variation and binning

# How reliable is the empirical variogram?

- Recall: it is based on some **sample** which represents the **population.**
- A different sample of the same size would give a different variogram. **Would they be consistent?**
  - i.e., when modelled (see below) would they result in more-or-less the same model?
- Simulation studies: e.g., Webster, R., & Oliver, M. A. (1992). *Sample adequately to estimate variograms of soil properties*. Journal of Soil Science, 43(1), 177–192.
- Conclusion: **150 to 200** observations allow reliable reconstruction of a known variogram model in the **isotropic** case.

# Table of Contents

# Modelling spatial autocorrelation

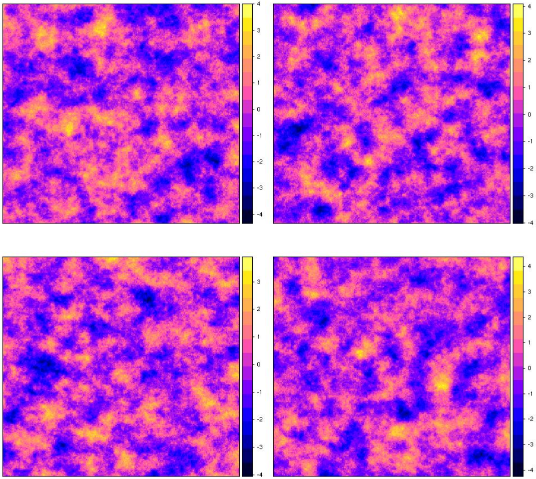
- Aim: To fit a **mathematical model** to an empirical variogram
- This model must be based on some **theory** – this is a modelling **assumption.**
- Theory: **random fields**[4]

[4]Source: Webster, R., & Oliver, M. A. (2001). *Geostatistics for environmental scientists*, John Wiley& Sons, Ltd.

# Spatially-autocorrelated random processes

- Assumption: The observed attribute values are only **one of many possible realisations** of a **random** ("stochastic") **process**
- This process is **spatially autocorrelated**, i.e., observations are **not independent**
- The result is called a **random field**
- Different stochastic processes are represented by different **models of spatial covariance**
- There is **only one reality** (which is sampled)
- From our one reality, we need to **infer the process** that produced it
- This dictates the proper **authorized variogram** (or, covariance) **function.**

# Four realizations of the same random field




256 x 256 grid; Spherical model; range 25; no nugget

# Modelling paradigm

1. **Assume** reality is one realization of a **regionalized variable** (structure *to be determined*)
2. **Assume** any spatial autocorrelation has the **same structure** everywhere
   - This is **2nd-order stationarity**
   - A major assumption and difficult to verify
3. Make observations; summarize as an **empirical variogram**
4. Select a **model of spatial autocorrelation**
5. **Parameterize** (fit) the selected model to the empirical variogram

# Selecting a model of spatial covariance

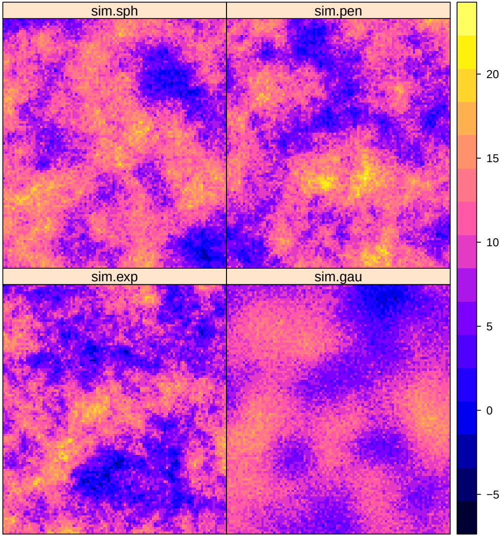
Various methods, more-or-less in order of preference:

1. What is known about the **spatial process** that produced the field
2. **Previous studies** of the same variable in similar circumstances
3. Visual assessment of the variogram form
4. Try to fit many, *maybe* automatic selection by "best" fit
5. Problem with "best" fit: depends on:
    1. variogram cutoff, bin width
    2. criterion for "best", e.g., more weight to more point-pairs and closer separations
    3. other forms may fit almost as well

# Four regionalized covariance models

same model parameters

Different processes



# Table of Contents

# Variogram model equations (1)

- Only some forms are **authorized**, i.e., will lead to positive-definite kriging matrices (see below). We review a few common models.
- All can be raised by the **nugget** variance $c_0$.
- The simplest is the **Exponential** model: sill $c$, **effective range** $3a$

$$\gamma(h) = c\left(1 - e^{\left(-\frac{h}{a}\right)}\right)$$

- Autocorrelation decreases exponentially with separation – the minimum spatial dependence.
- This is an *asymptotic* model: variance **approaches** a sill at some **effective range**, by convention, where$\gamma = 0.95c$.

# Variogram model equations (2)

**Gaussian** model: sill $c$, effective range $\sqrt{3}a$:

$$\gamma(h) = c\left(1 - e^{-\left(\frac{h}{a}\right)^2}\right)$$

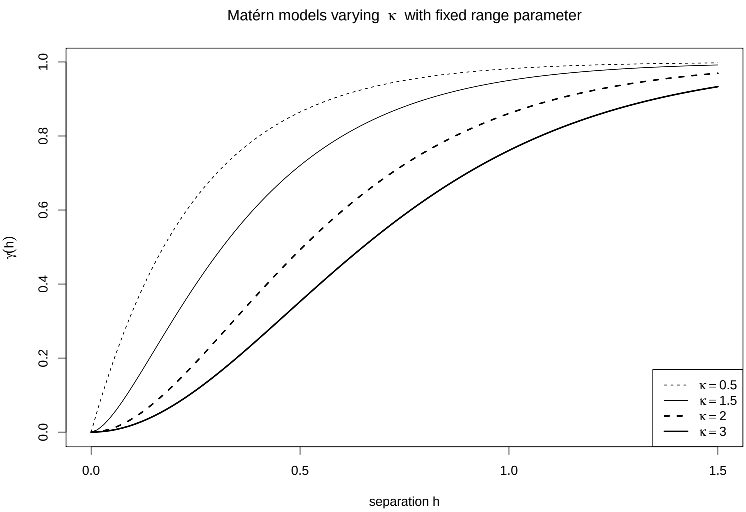
This has strong spatial continuity near the origin (0-separation), e.g., water table elevation, smoothly-varying terrain properties

# Variogram model equations (3)

**Matérn** model family: *generalizes* the Exponential, Power, Logarithmic and Gaussian models

$$\gamma(h) = c\left(1 - \frac{1}{2^{\kappa-1}\Gamma(\kappa)}\left(\frac{h}{a}\right)^{\kappa} K_{\kappa}\left(\frac{h}{a}\right)\right)$$

- smoothness parameter is $\kappa$; this adjust the variogram model to the process.
- small $\kappa$ implies that the spatial process is rough, large $\kappa$ smooth.
- $K_{\kappa}$ is a modified Bessel function of the second kind
- $\Gamma$ is the Gamma function (generalization of the factorial function)
- if $\kappa = 0.5$ this reduces to the exponential model
- if $\kappa = \infty$ this reduces to the Gaussian model
- most common values are $\kappa = 0.5, 1, 1.5, 2$


Matérn models varying κ with fixed range parameter
γ(h)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
separation h
κ = 0.5
κ = 1.5
κ = 2
κ = 3

# Variogram model equations (4)

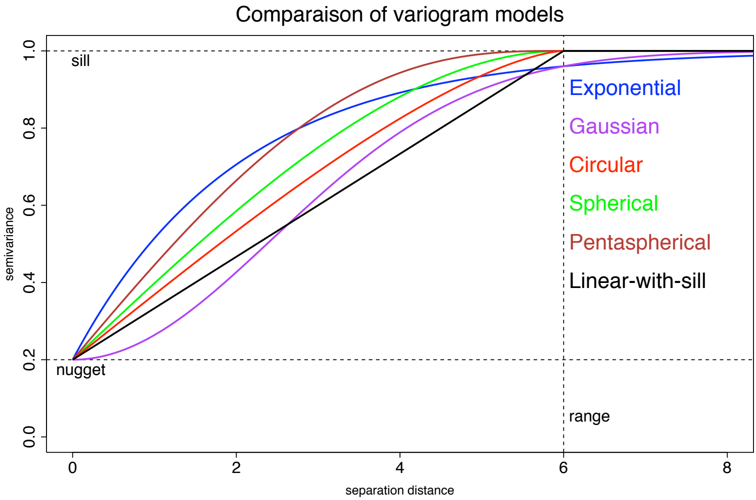
**Spherical** model: sill $c$, range $a$

$$\gamma(h) = \begin{cases} c\left(\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right) & : \quad h < a \\ c & : \quad h \geq a \end{cases}$$

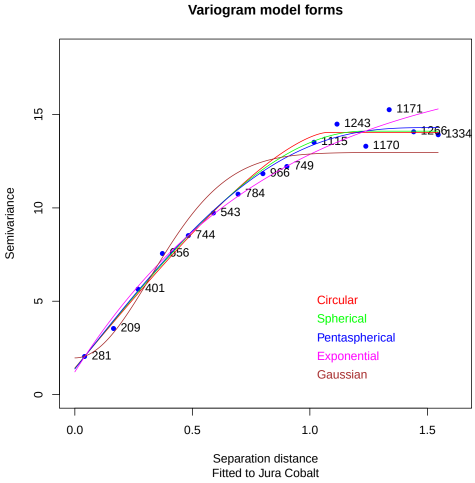
This is **almost linear** near the origin, reaches the sill $c$ at the range $a$ and is then **constant**, with a **“shoulder”** transition between.

It is often applied when the variable occurs in somewhat homogeneous **patches** with gradual boundaries, e.g., vegetation density, soil properties.

# Comparing variogram models – same parameters



# Fitted variogram models to same empirical variogram



# Combining models

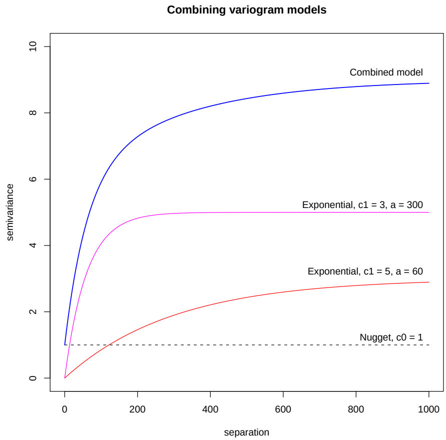
- Any **linear combination** of authorized models is also authorized
- Models with $> 1$ spatial structure at **different ranges**
- Common example: **nugget + structural**
- e.g. nugget + exponential

$$\gamma(h) = \mathbf{c_0} + c_1 \left(1 - e^{\left(-\frac{h}{a}\right)}\right)$$

- Structure at two ranges: e.g., **nugget** + exponential + exponential

$$\gamma(h) = \mathbf{c_0} + c_1 \left(1 - e^{\left(-\frac{h}{a_1}\right)}\right) + c_2 \left(1 - e^{\left(-\frac{h}{a_2}\right)}\right)$$

# Combining variogram models




Models nugget, short range ($3a = 180$) and long range ($3a = 900$) structures

# Table of Contents

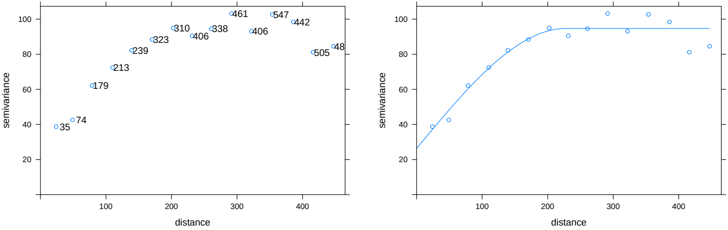
# Parameterizing an empirical variogram

- Problem: we do not know the random field, we only have an **empirical variogram**
  - i.e., the semivariances and separation distances for all point-pairs
- We select a **model form** (see above) . . .
  - this is the form of the assumed random field
- . . . now we need to **parameterize** it:
  - to interpret spatial structure
  - to be used in kriging

# Method 1: Method of moments

1. **Bin** the semivariances into a set of $m$ distance intervals
2. **Average** the semivariances in each bin; now we have a relation $\gamma \sim h$ for $m$ bins
3. Select a **variogram model**
4. Estimate (by eye, or semi-automatic) **initial parameters** (range $a$, partial sill $c$, nugget $c_0$)
5. **Fit** by **weighted non-linear least squares**: e.g., `gstat::fit.variogram`
    - *ad-hoc* weighting, generally $\propto n_h$, $\propto 1/h^2$
    - gives more weight to closely-separated pairs, bins with more point-pairs

# Empirical variogram and MoM fit




A

**spherical** model was selected by eye.

# Problems with the Method of Moments

- No objective way to select bins, cutoff
- Changing these will change the weighted least-squares fit (especially with small datasets)
- *ad-hoc* fitting method

# Method 2: Maximum likelihood

- Does not use binned point-pair semivariances; uses all point-pairs directly
- Assumption: the $n$ sample data come from a **multivariate normal distribution** (maybe after transformation)
- We model the **joint** probability density of all point-pairs and then solve for the covariance parameters
- This also requires the selection of a **model** to parameterize

# Computation

The **joint** probability density of all the points is computed as:

$$f(\mathbf{z}, \mu, \theta) = (2\pi)^{-\frac{n}{2}} |\mathbf{C}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\mu)^{\mathrm{T}}\,\mathbf{C}^{-1}\,(\mathbf{z}-\mu)\right\} \quad (3)$$

- $\mathbf{z}$ the vector with the $n$ sample data
- $\mu$ the vector with (unknown) means
- $\theta$ the vector with (unknown) **parameters** of the covariance function
- $\mathbf{C}$ the $n \times n$ **variance-covariance matrix** of the sample data.

**C** contains covariances (computed from the separation based $\theta$) on its off-diagonals. There is no binning.
This requires a **covariance model** which is applied to the separation between **all** pairs of points.

# Solution by ML

- The probability density of can be regarded as a function of $\mu$ and $\theta$ with the data $\mathbf{z}$ fixed
- This defines the likelihood $L(\mu, \theta|\mathbf{z})$, which can be **maximized** over the parameter space
- e.g., `geoR::likfit`
- Similar to the REML fit to the covariance function in **gls.**

# Table of Contents

# Predicting from a model of spatial autocorrelation and a set of observations

- Once we have a variogram model, it can be used to **predict** at unobserved locations.
- Model without trend: $Z(\mathbf{s}) = \mu + \varepsilon(\mathbf{s}) + \varepsilon'(\mathbf{s})$
- The realization of the random field at point **s** is:
  - some **mean value** $\mu$; plus ...
  - ... a **spatially-autocorrelated** random component $\varepsilon(\mathbf{s})$, with a defined covariance structure (e.g., a variogram model); plus ...
  - ... pure **noise** $\varepsilon'(\mathbf{s})$: nugget and lack of spatial correlation with increasing separation
  - Both the expected value (1$^{st}$-order) and covariance structure (2$^{nd}$-order) are **stationary**: the same everywhere in the field

# Non-geostatistical prediction methods

All of these have **no theory of spatial autocorrelation**, they have *ad hoc* implicit models of spatial structure:

- nearest neighbour (Thiessen polygons, Voronoi tessellation of space)
- average of nearest $k$-neighbours
- average of nearest $k$-neighbours weighted by inverse distance to some power
- average of all neighbours within some radius
- average of all neighbours within some radius weighted by inverse distance to some power
- ...with de-clustering of compact groups of known points

Choice of $k$, radius by cross-validation.

# Table of Contents

# A geostatistical prediction method: Ordinary Kriging (OK)

- The estimated value $\hat{z}$ at a point $\mathbf{x}_0$ is predicted as the **weighted average** of the values at *all* sample points $\mathbf{x}_i$:

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

- The weights $\lambda_i$ assigned to the sample points **sum to 1**: $\sum_{i=1}^{N} \lambda_i = 1$, therefore, the prediction is **unbiased.**
- Many other interpolators (e.g., inverse distance) are also linear unbiased, but OK is the **“best”** of all possible weightings

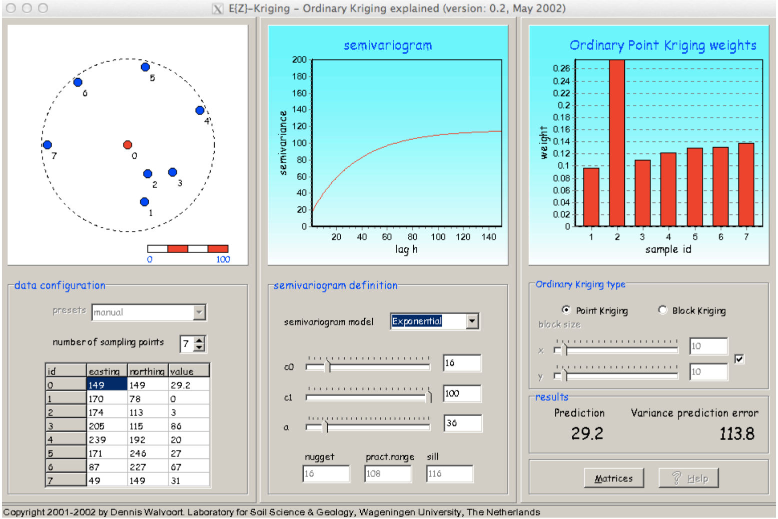
# In what sense is OK the "best" predictor?

- OK is called the **"Best Linear Unbiased Predictor"** (BLUP)
- "best" ≡ lowest **prediction variance** of all possible weightings
  - i.e., each prediction has the smallest possible confidence interval
- This criterion is used to derive the **OK system of equations**, which is solved to determine the **weights** for each sample point
- Weights depend on the **spatial covariance structure** as modelled by the **variogram model.**
- Spatial structure **between observations**, as well as **between observations and a prediction point**, is accounted for.

# Implications

- The **prediction** and its variance are only as good as the **model of spatial structure.**
- Points **closer** to the point to be predicted have **larger weights**, according to the modelled **spatial dependence**
- **Clusters** of points **“reduce to”** single equivalent points
  - i.e., over-sampling in a small area can’t bias result
  - automatically de-clusters
- Closer sample points **“mask”** further ones in the same direction
- Error estimate is based only on the **spatial configuration of the sample**, not the data values

# Experimenting with OK: **E{Z}-Kriging**




https://wiki.52north.org/AI_GEOSTATS/SWEZKriging

# Derivation of the OK system of equations

- Aim: **minimize the prediction variance**, subject to the **unbiasedness** and **spatial covariance** constraints.
- Two ways to derive the OK system:
  - 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 is an extension of linear modelling theory.
- Approach (2) is an application of standard minimization methods from differential calculus; but is not so transparent, because of the use of LaGrange multipliers.

# Matrix form of the Ordinary Kriging system

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

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

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

# Notation

- **kriging weights** $\lambda_i$ to be assigned to each observation point
- **semivariances** $\gamma$ between
  1. point to be predicted $\mathbf{x}_0$ and observation points $\mathbf{x}_i$;
  2. pairs of observation points $(\mathbf{x}_i, \mathbf{x}_j)$
- **LaGrange multiplier** $\psi$ which enters in the prediction variance

# Solution

- This is a system of $N+1$ equations in $N+1$ unknowns, so can be solved uniquely for the weights vector $\lambda$.

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

- But to compute the matrix inverse $\mathbf{A}^{-1}$ the $\mathbf{A}$ matrix (spatial structure) must be **positive definite**
- This is guaranteed for **authorized models** of spatial covariance

# OK prediction

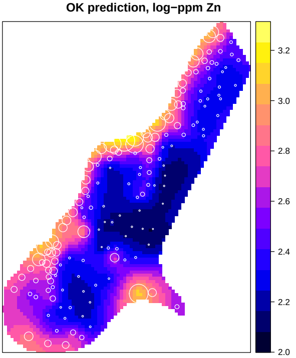
- Now we can **predict** at the point, as a weighted sum:

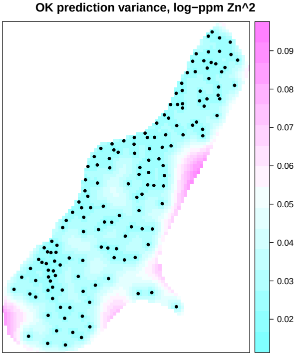
$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i z \mathbf{x}_i)$$

- The **kriging variance** at a point is computed as:

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

# Ordinary kriging (OK) predictions and variances




Predictions, log(Cd)
With postplot

Variance, Meuse log(Cd)$^2$
With sample points

# Characteristics of OK prediction

1. **smooth**: moving across the map, the kriging weights change smoothly, because the distance changes smoothly
2. Prediction is "best" (given the model and data) at each point separately
3. But the map is not realistic as a whole (smoother than reality)
4. Pure noise at each point represented by the prediction variance
5. Variance depends on the **configuration of the sample points**, *not* the **data values**!

# Table of Contents

# Geostatistics with the universal model

- Recall: the universal model is: $Z(\mathbf{s}) = Z^*(\mathbf{s}) + \varepsilon(\mathbf{s}) + \varepsilon'(\mathbf{s})$
- In the previous section we replaced $Z^*(\mathbf{s})$ with a constant $\mu \rightarrow$ **1st order stationarity.**
- Now we return to the full model: both the **deterministic** and **spatially-autocorrelated** must be modelled
- Question: How to separate the effects? or **how to model them in one step**?

# Adding non-stationarity to the model

- Depends on the **source** of the non-stationarity
  1. Regional trend → **Universal Kriging** (UK)
  2. Spatially-distributed covariate → **Kriging with External Drift** (KED)
- These are **mathematically equivalent**, the difference is in the **base functions** that define the non-stationarity.

# 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.**
- UK is recommended when there is evidence of **1st-order non-stationarity**, i.e. the **expected value** varies across the map according to the coördinates, but there is still **2nd-order stationarity** of the **residuals** from this trend.

# UK – application example

The elevation of 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 autocorrelation in this local structure – the **residuals** of the trend surface.

# Base functions for UK

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

$$Z^*(\mathbf{s}) = \sum_{j=1}^{p} \beta_j f_j(\mathbf{s})$$

- Base functions for **linear** drift:

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

  where $s_1$ is one coördinate (say, E) and $s_2$ the other (say, N)
- Note that $f_0(\mathbf{s}) = 1$ estimates the global mean (as in OK).
- Base functions for **quadratic** drift: also include second-order terms:

$$f_3(\mathbf{s}) = s_1^2,\ f_4(\mathbf{s}) = s_1 s_2,\ f_5(\mathbf{s}) = s_2^2$$

# Unbiasedness of UK predictions

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

$$\sum_{i=1}^{N} \lambda_i f_k(\mathbf{s_i}) = f_k(\mathbf{s_0}), \quad \forall\, k$$

- 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** for a linear trend: If $f_1(\mathbf{s}_0) = s_1$, then at each point $\mathbf{s}_0$ the expected value must be $s_1$, i.e. the point's E coördinate:

$$\sum_{i=1}^{N} \lambda_i s_i = s_1$$

This is a **further restriction** on the weights $\lambda$.

# The UK system (1)

$$\mathbf{A_U}\lambda_\mathbf{U} = \mathbf{b_U}$$

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

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 UK system (2)

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

The $\lambda_{\mathbf{U}}$ 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)

$\mathbf{b}_{\mathbf{U}}$ is structured like an additional column of $\mathbf{A_u}$, but referring to the point to be predicted.

# Predicting by UK

- Same as OK: a weighted linear combination of values at known points:

$$\hat{Z}(\mathbf{x_0}) = \sum_{i=1}^{N} \lambda_i z(\mathbf{x_i})$$

- But, **the weights** $\lambda_i$ for each sample point take into account both the **global trend** and **local spatial autocorrelation of the trend residuals.**
- The UK system must include both of these.

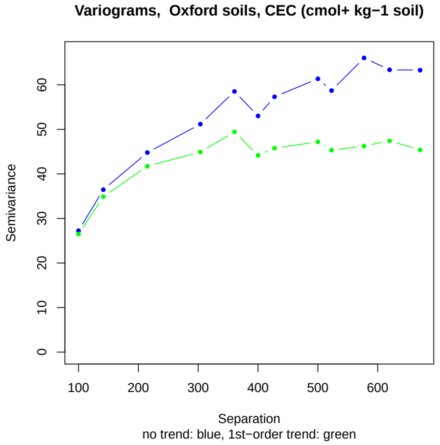
# Computing the empirical semivariogram for UK

- The semivariances $\gamma$ 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**.
  4. Note that `gstat::variogram` can do this in one step.
- **Problem**: the trend should have taken the spatial correlation into account!

# 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)
- In theory, the **nugget** should be unchanged (residual variance at a point is not removed by a trend)

# Example original vs. residual variogram




Note lower (partial, total) sill, shorter range, same nugget

# UK: 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

# Why use UK in a 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.

# Table of Contents

# Kriging with External Drift (KED)

- This is a **mixed predictor** that includes **feature-space predictors**, rather than geographic coördinates (as in UK).
- The **mathematics are exactly as for UK**, but the *base functions* are different.
- UK vs. KED:
  - In UK, the base functions refer to the **grid coördinates**; these are by definition known at any prediction point.
  - In KED, the base functions refer to some **feature-space covariates** . . .
    - . . . 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).

# KED – application example

Soil thickness partly depends topographic position (typically, thicker on hilltops and in valleys than on side slopes). This can be represented by geomorphometric parameters (relative elevation, slope gradient, profile curvature . . . ), which are available over the whole area from a digital elevation model (DEM). But there are local variations within this that are

spatially-correlated due to local factors not captured in the DEM (e.g., tree throw, small mass movements). There may be spatial autocorrelation in this local structure – the **residuals** of the deterministic model.

# Base functions for KED

There are two kinds of feature-space covariates:

1. **strata**, i.e., factors, categorical variables. Examples: soil type, flood frequency class
   - Base function: $f_k(\mathbf{s}) = 1$ iff sample or prediction point $\mathbf{s}$ is in class $k$, otherwise 0 (class indicator variable)
2. **continuous covariates.** Examples: elevation, NDVI
   - Base function: $f_k(\mathbf{s}) = v(\mathbf{s})$, i.e. the value of the predictor at the point.

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

# Conclusion

- Identify the sources of spatial variation in the universal model:
    - geographic trend
    - dependence on a spatially-distributed covariate
    - local spatial correlation
    - pure error
- Model accordingly.