Applied geostatistics

Lecture 2 – Exploring and visualizing spatial data

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

January 7, 2014

Topics for this lecture

1. Visualizing spatial structure: point distribution, postplots, quantile plots
2. Visualizing regional trends
3. Visualizing local spatial dependence: h-scatterplots, variogram cloud, experimental variogram
4. Visualizing anisotropy: variogram surfaces, directional variograms

Commentary

Spatial data by its nature can be **visualised** on a map, since each observation has coördinates in geographic space.

So before beginning any analysis, we can **explore** the nature of this spatial data with the **visualisation** tools provided by computing environments.

Topic 1: Visualizing spatial structure

1. Distribution in space

2. Postplots

3. Quantile plots

Commentary

We begin by examining sets of **sample points** in space.

The first question is how these points are **distributed** over the space. Are they **clustered**, **dispersed**, in a **regular** or **irregular** pattern, **evenly** or **un-evenly** distributed over subregions?

There are statistical techniques to answer these questions objectively; here we are concerned only with **visualization**

Distribution in space

This is examined with a **scatterplot** on the coördinate axes, showing only the position of each sample.

This can be in:

1. 1D : along a **line** or curve

2. 2D: in the **plane** or on a surface

3. 3D: in a **volume** of space

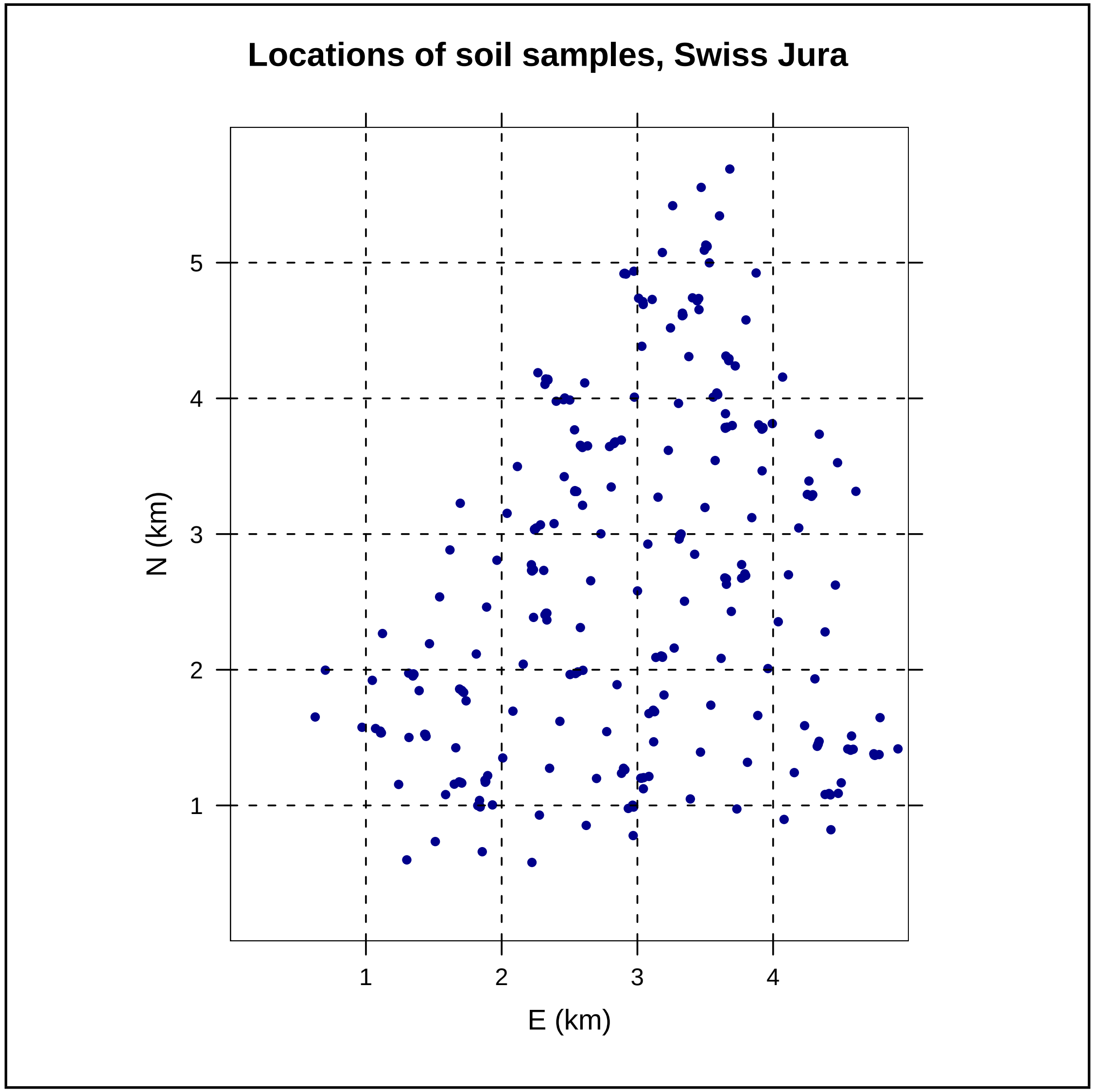
Note: We can not easily visualize higher dimensions, and they are not necessary for strictly geographic data. However, if **time** is added as a dimension, we can only visualize 2D points over time.

Commentary

Let's look at some distributions of sample points in 2D space.

The first plot is of sample points in a study area of the Jura mountain region of Switzerland.

Distribution plot of soil samples, Swiss Jura



To check your understanding . . .

Q1 : *Do the points cover the entire 5x5 km square? What area do they cover?* *Jump to A1* •

Q2 : *Within the* **convex hull** *of the points (i.e. the area bounded by the outermost points), is the distribution:*

1. *Regular or irregular?*
2. *Clustered or dispersed?*

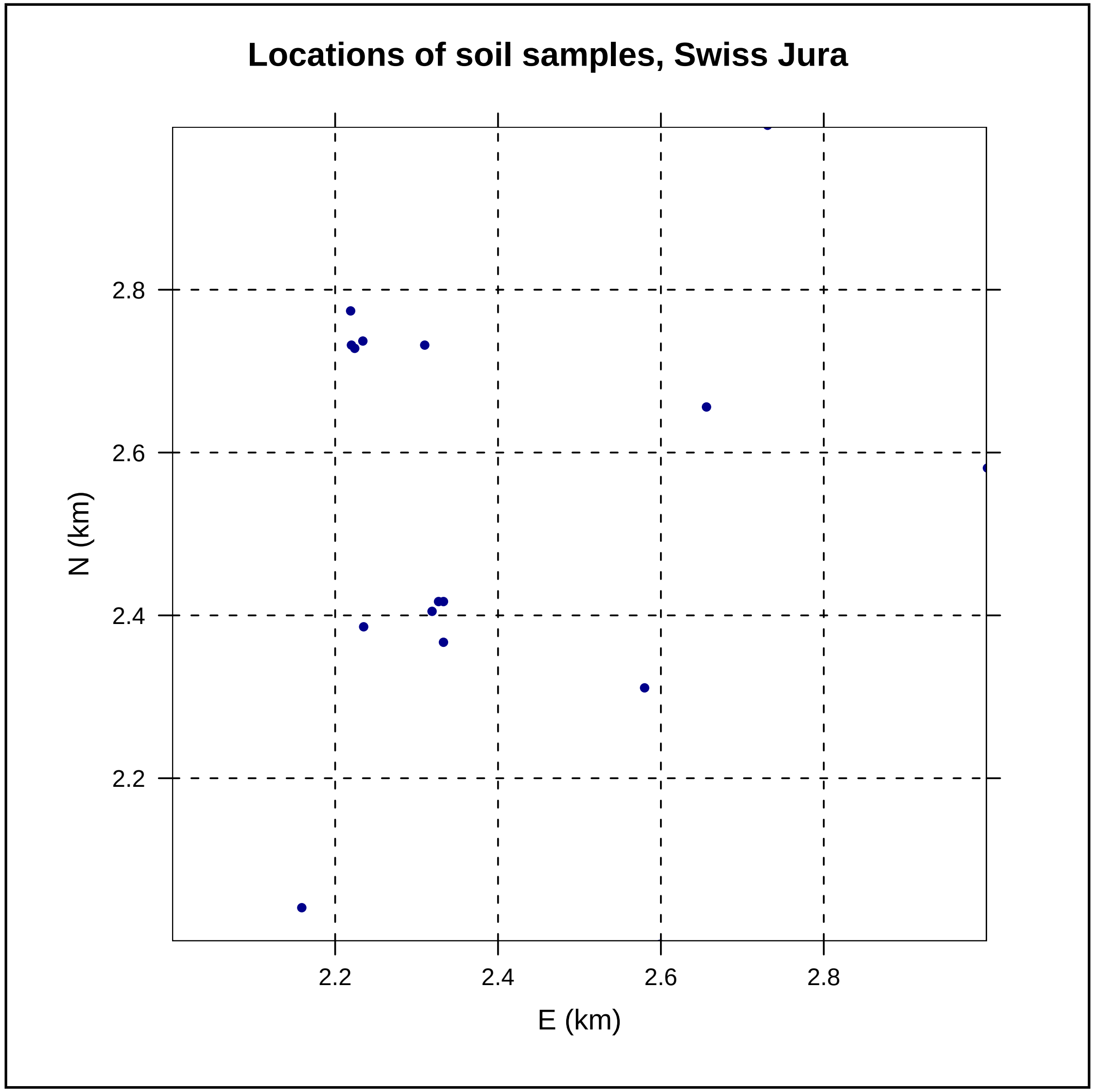
Jump to A2 •

Commentary

Now we look at a selected 1x1 km square within the study area.

We can evaluate the point distribution at this higher resolution.

Distribution plot of soil samples, Swiss Jura within a 1 x 1 km area



To check your understanding . . .

Q3 : *Do the points cover the entire 1x1 km square?* *Jump to A3* •

Q4 : *Within the* **convex hull** *of the points (i.e. the area bounded by the outermost points), is the distribution:*

1. *Regular or irregular?*
2. *Clustered or dispersed?*

Jump to A4 •

Postplots

These are the same as distribution plots, except they show the **relative value** of each point in its **feature space** by some graphic element:

1. relative **size**, or

2. **colour**, or

3. both

Showing feature-space values with symbol size

One way to represent the value in feature space in by the symbol **size**, in some way proportional to the value. But, which size?

- Radius proportional to **value**

- Radius proportional to **transformed value**

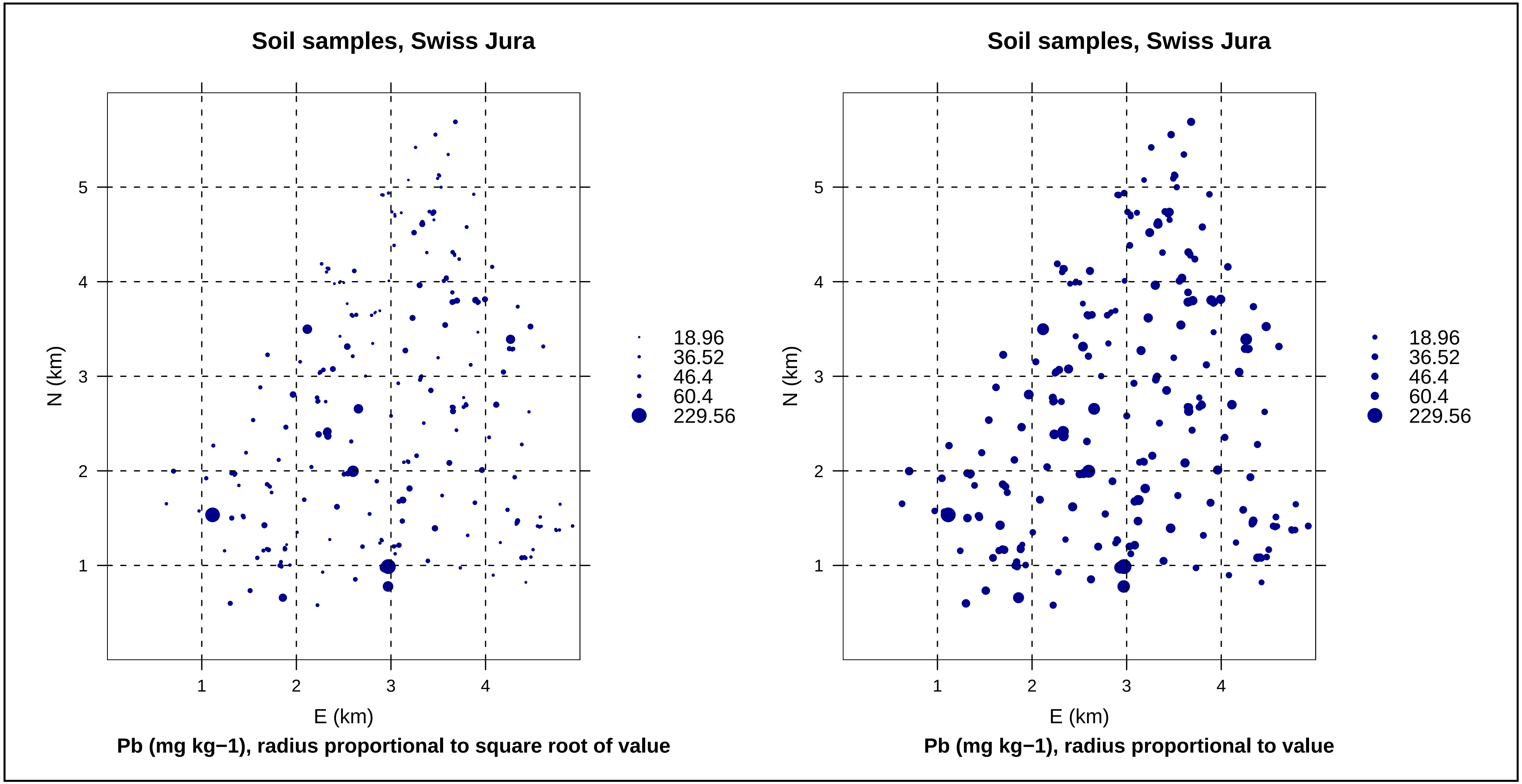
  1. Square root (because radius is 1D)
  2. Logarithm to some base

Commentary

We now look at the same points from the Swiss Jura, but print them with a size proportional to their lead (Pb) concentration.

Two versions are compared: radius proportional to value, and to the square root of the value.

Post-plot of Pb values, Swiss Jura; size



To check your understanding . . .

Q5 : *Which of this two graphs best shows the highest values ("hot spots")?* *Jump to A5* •

Q6 : *Which of this two graphs best shows the distribution of the values in feature space?* *Jump to A6* •

Showing feature-space values with a colour ramp

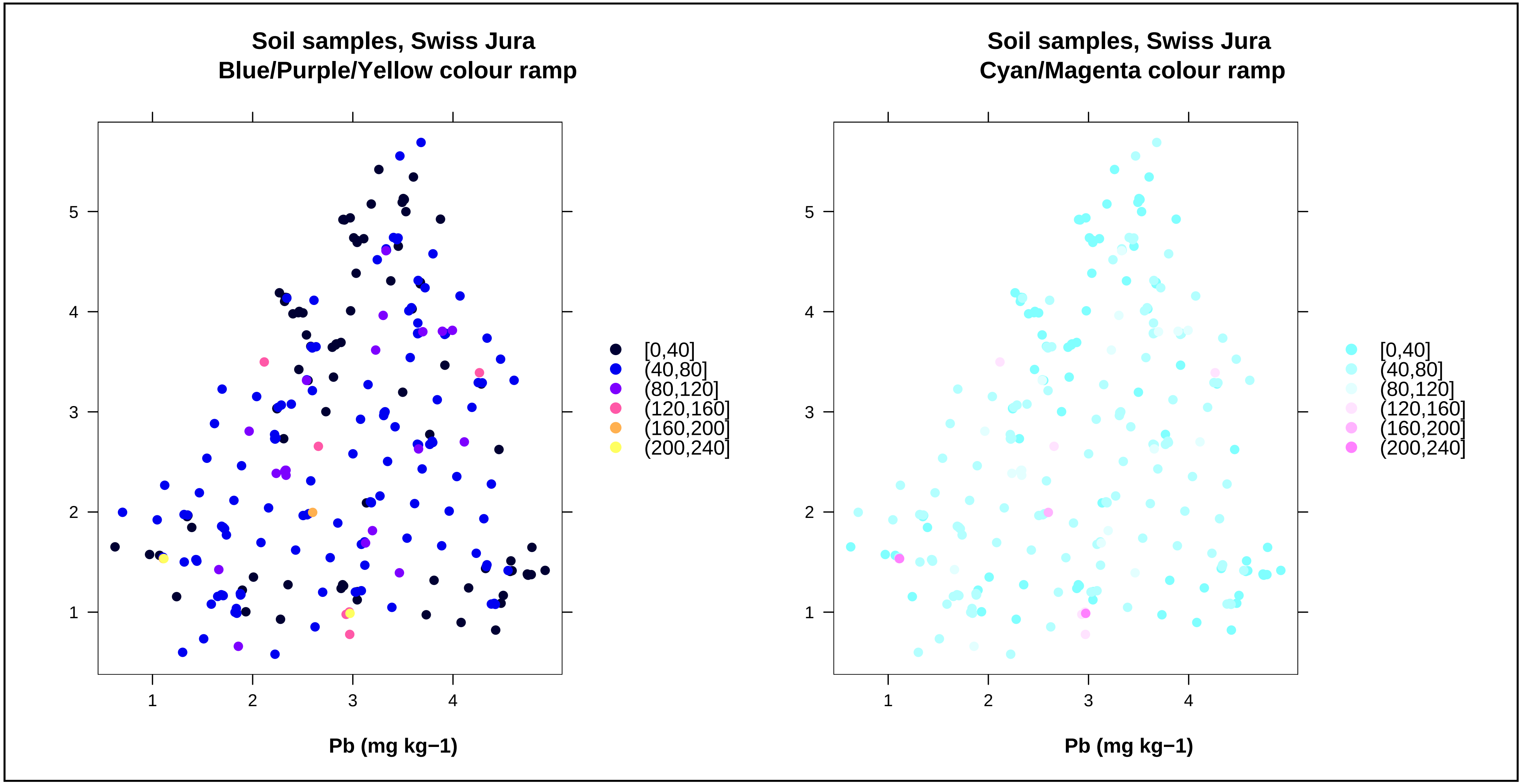
Another way to show the difference between feature-space values is with a **colour ramp**, i.e. a sequence of colours from low to high values.

Different colour ramps give very different impressions of the same data, as we now illustrate.

Warning: When points are widely-spaced, it is very difficult to pick out patterns shown by the colour ramp, because the eye is confused by the background (no matter what colour) where there are no points. The postplot using symbol size is usually a better choice.

Still, we show the colour ramp here, so you can form your own opinion.

Post-plot of Pb values, Swiss Jura; colours



To check your understanding . . .

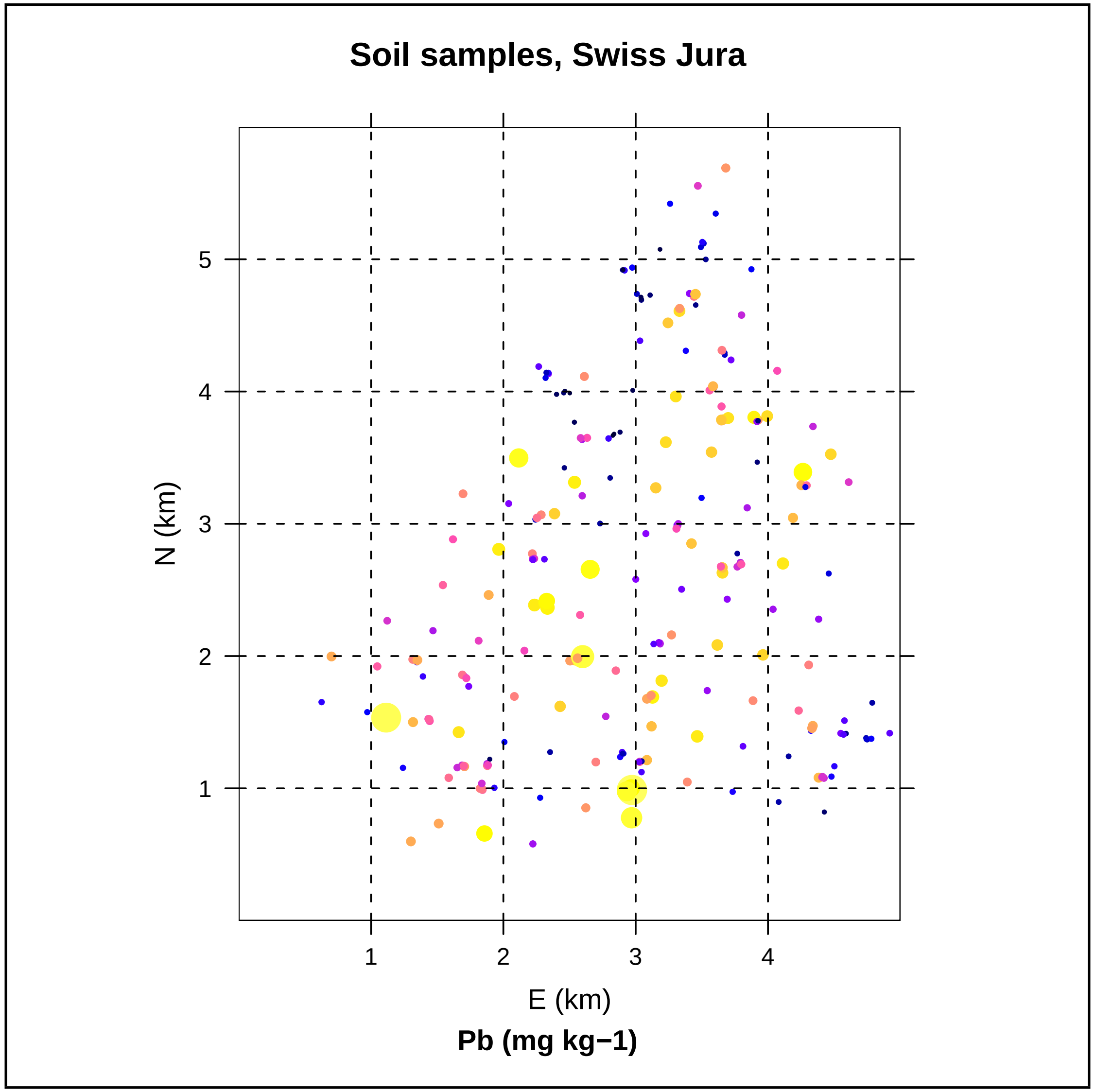
Q7 : *Which of the two ramps (blue/purple/yellow and cyan/magenta) do you think better highlights the hot spots?* *Jump to A7* •

Q8 : *Do you prefer the size or colour to show the feature-space value?* *Jump to A8* •

Showing feature-space values with sizes and a colour ramp

And of course we can combine both visualization techniques in one graph.

Post-plot of Pb values, Swiss Jura; colours and symbol size



Quantile plots

A **quantile plot** is a postplot where one quantile is represented in a **contrasting size or colour**. This shows how that quantile is distributed.

A **quantile** is a defined range of the **cumulative** empircal distribution of the variable.

The quantiles can be:

- **quartiles** (0-25%, 25-50% . . .);
- **deciles** (0-10%, 10-20% . . .);
- any cutoff point intersting to the analyst, e.g 95% (i.e. highest 5%)

Quantiles

Examples of quantiles of the Pb contents of 259 soil samples from the Swiss Jura:

Here are all the values, sorted ascending:

```
R> sort(jura.cal$Pb)
  [1]  18.96  20.20  21.48  21.60  22.36  22.56  23.68  24.64  25.40  26.00  26.76
 [12]  26.84  26.96  27.00  27.04  27.04  27.20  27.68  28.56  28.60  28.80  29.36
...
[243]  91.20  93.92 101.92 104.68 107.60 116.48 118.00 129.20 135.20 138.56 141.00
[254] 146.80 157.28 172.12 195.60 226.40 229.56
```

Here are some quantiles:

```
R> quantile(jura.cal$Pb)
    0%     25%     50%     75%    100%
 18.96   36.52   46.40   60.40 229.56
R> quantile(jura.cal$Pb, seq(0.1, 1, by=0.1))
    10%      20%      30%      40%      50%      60%      70%      80%      90%     100%
 30.792   34.952   37.680   41.656   46.400   51.200   56.472   65.360   80.480  229.560
R> quantile(jura.cal$Pb, 0.95)
   95%
104.97
```

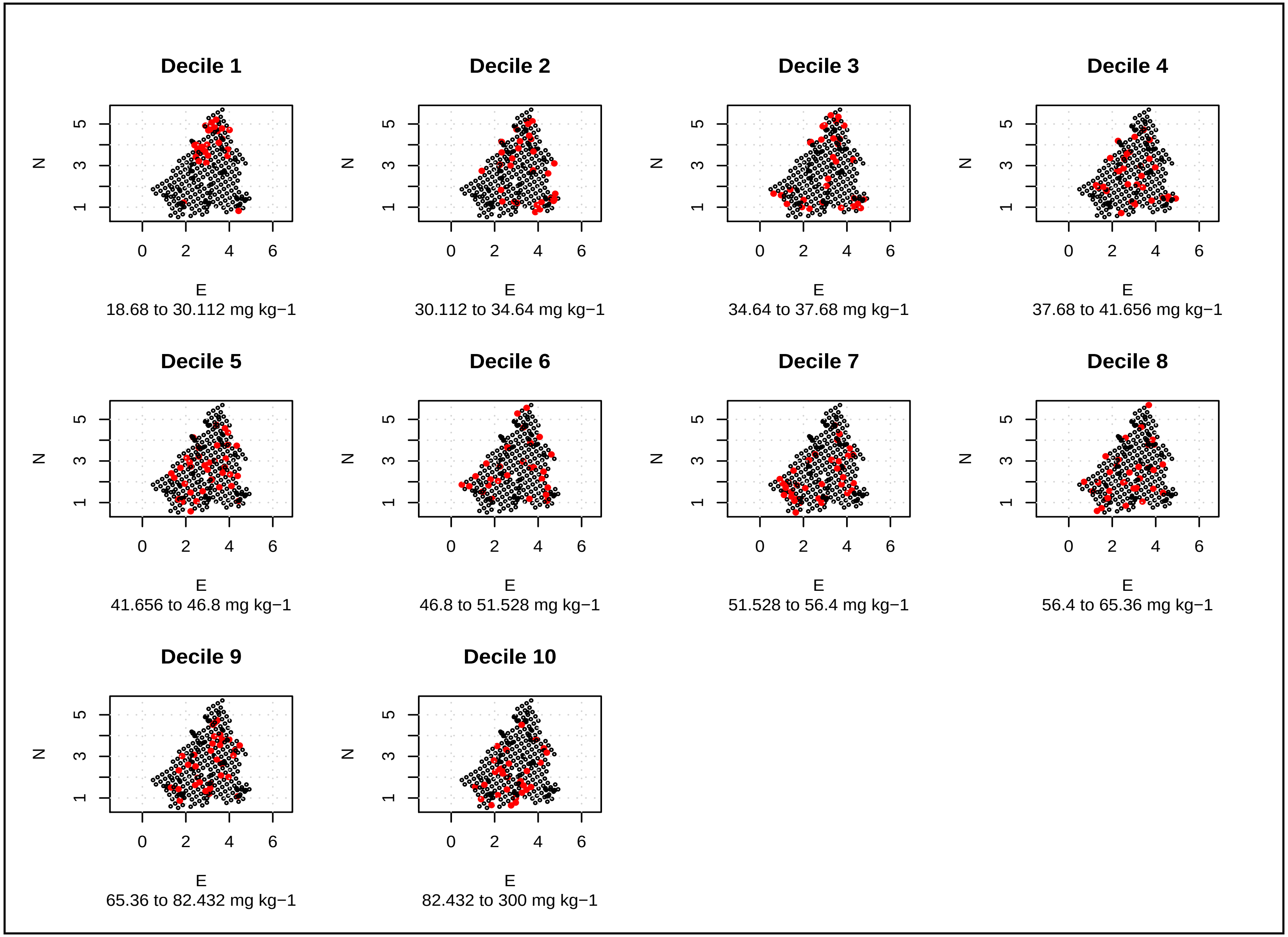
Commentary

The figure on the next page has 10 graphs, one for each **decile**. Points whose Pb content is in the decile are shown with a large symbol and others with a small symbol.

This sequence of graphs shows:

- whether certain deciles are concentrated in parts of the area; this would suggest a **regional trend**
- whether the deciles are clustered; if all are about equally clustered this suggests **local spatial dependence**

Quantile plot of Pb values (deciles), Swiss Jura



To check your understanding . . .

Q9 : *Are any of the deciles concentrated in parts of the area?* *Jump to A9* •

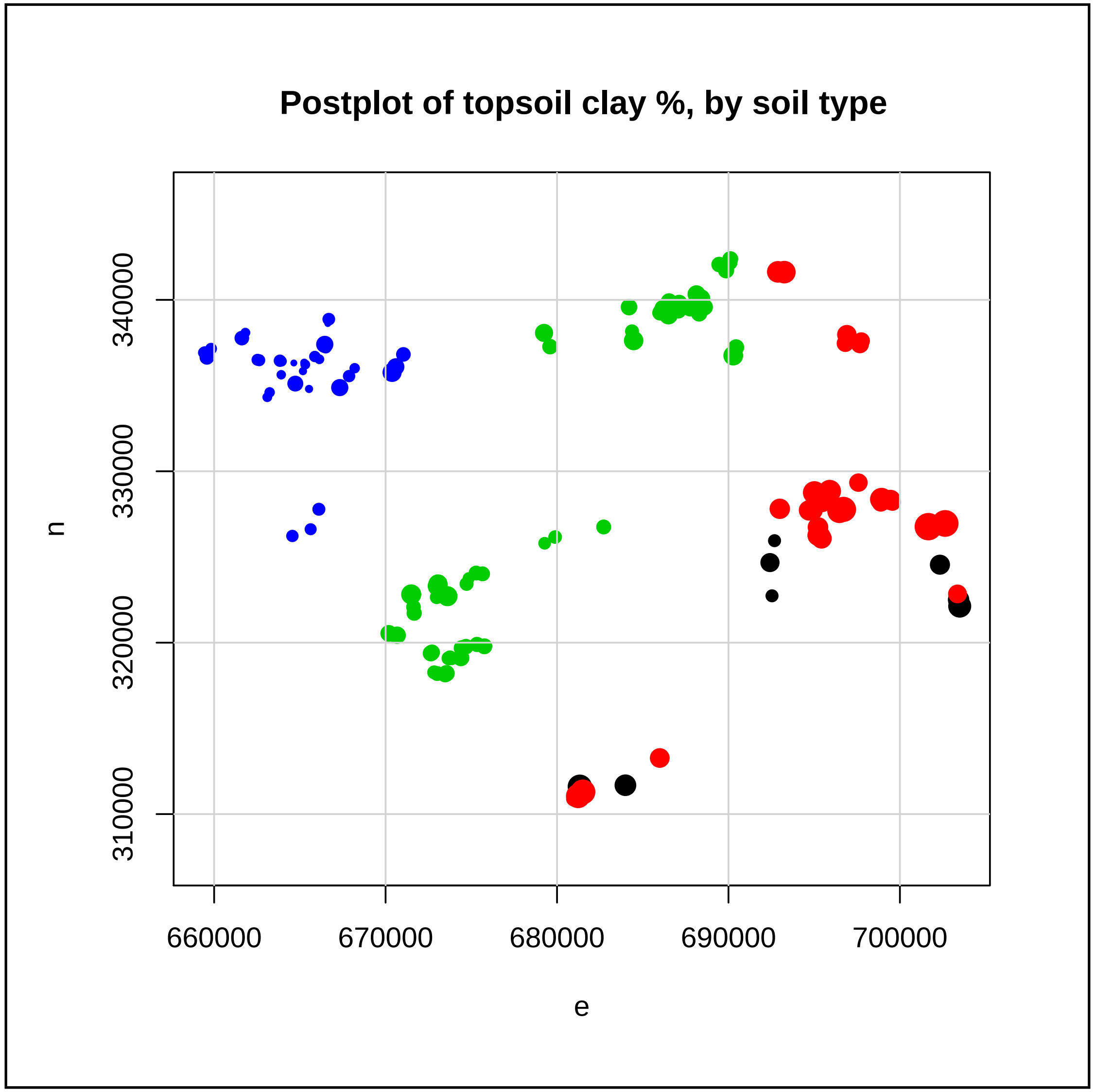
Q10 : *Are points in any or all of the deciles clustered?* *Jump to A10* •

Classfied postplots

If the observations come from **classes**, it is possible to visualize these:

- **without** the values of some continuous attribute: just visualize where the different classes are located;
- **with** the values of some continuous attribute: also see if the continuous attribute depends on:
  - ∗ the **class**;
  - ∗ a **regional** trend;
  - ∗ a **local** spatial dependence;
  - ∗ or some **combination**.

Classified post-plot, with a continuous attribute



To check your understanding . . .

Q11 : *What is the* **classifying** *attribute in this postplot?* *Jump to A11* •

Q12 : *What is the* **continuous** *attribute in this postplot?* *Jump to A12* •

Q13 : *Does the clay content appear to depend, at least to some extent, on the soil type?* *Jump to A13* •

Postplots with another variable as colour

Instead of using colour to repeat the information already shown with the symbol size, it can be used to show another variable.

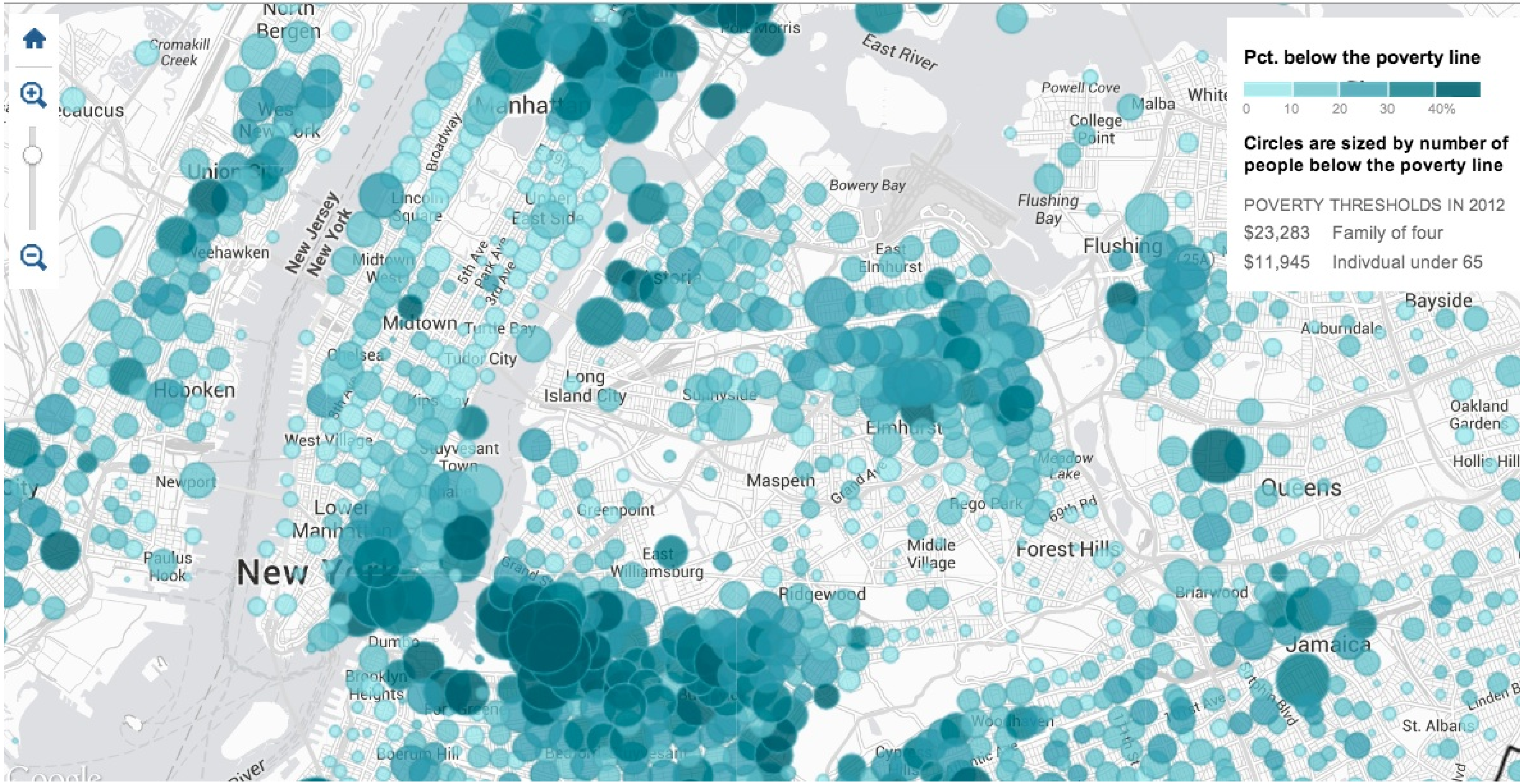
The next page shows a quite sophisticated postplot of this type, from the *New York Times* of 05-January-2014.

Points are the centres of census districts in a portion of New York City (USA) and adjacent New Jersey.

Circle size is proportional to the *number* of people below a defined income level ("poverty line") in the census district.

Circle colour is a colour ramp based on the *percentage* of the population below this income level.

Thus the circle *size* gives an *absolute* value (number of people), whereas the circle *colour* gives a *relative* value (proportion in poverty).



Pct. below the poverty line

0 10 20 30 40%

Circles are sized by number of people below the poverty line

POVERTY THRESHOLDS IN 2012

$23,283	Family of four
$11,945	Indivdual under 65

source: `http://www.nytimes.com/newsgraphics/2014/01/05/poverty-map/`

Topic 2: Visualizing a regional trend

1. Origin of regional trends
2. Looking for a trend in the **post-plot**;
3. Computing a **trend surface**

Regional trends

One kind of spatial structure is a **regional trend**, where the **feature space** value of the variable changes **systematically** with the **geographic space** coördinates.

A common example in many parts of the world is annual precipitation across a region, which decreases away from a source.

In a later lecture we will see how to combine a regional trend with local structure; here we just want to visualize any trend and assess it qualitatively.

Orders of trends

A systematic trend is an approximation to some **mathematical surface**; in a later lecture we will see how to find the mathematical representation.

For now, we are concerned with the **form** of the surface:

- **First-order**, where the surface is a **plane** (also called **linear**): the attribute value changes by the **same** amount for a given change in distance away from an origin;
- **Second-order**, where the surface is a **paraboloid** (2D version of a parabola), i.e. a **bowl** (lowest in the middle) or **dome** (highest in the middle)
- **Higher-order**, where the surface has **saddle points** or **folds**

To check your understanding . . .

Q14 : *What* **order of trend surface** *would you* **expect** *from these situations:*

1. *Mean annual precipitation in a coastal region where there is a low coastal plain, bounded inland by a mountain range; both of these are more or less linear in a given direction;*
2. *The depth to a gas deposit trapped in a salt dome;*
3. *Clay content in a soil developed from a sedimentary series of shales (claystones), siltstones and sandstones, outcropping as parallel strata?*

Jump to A14 •

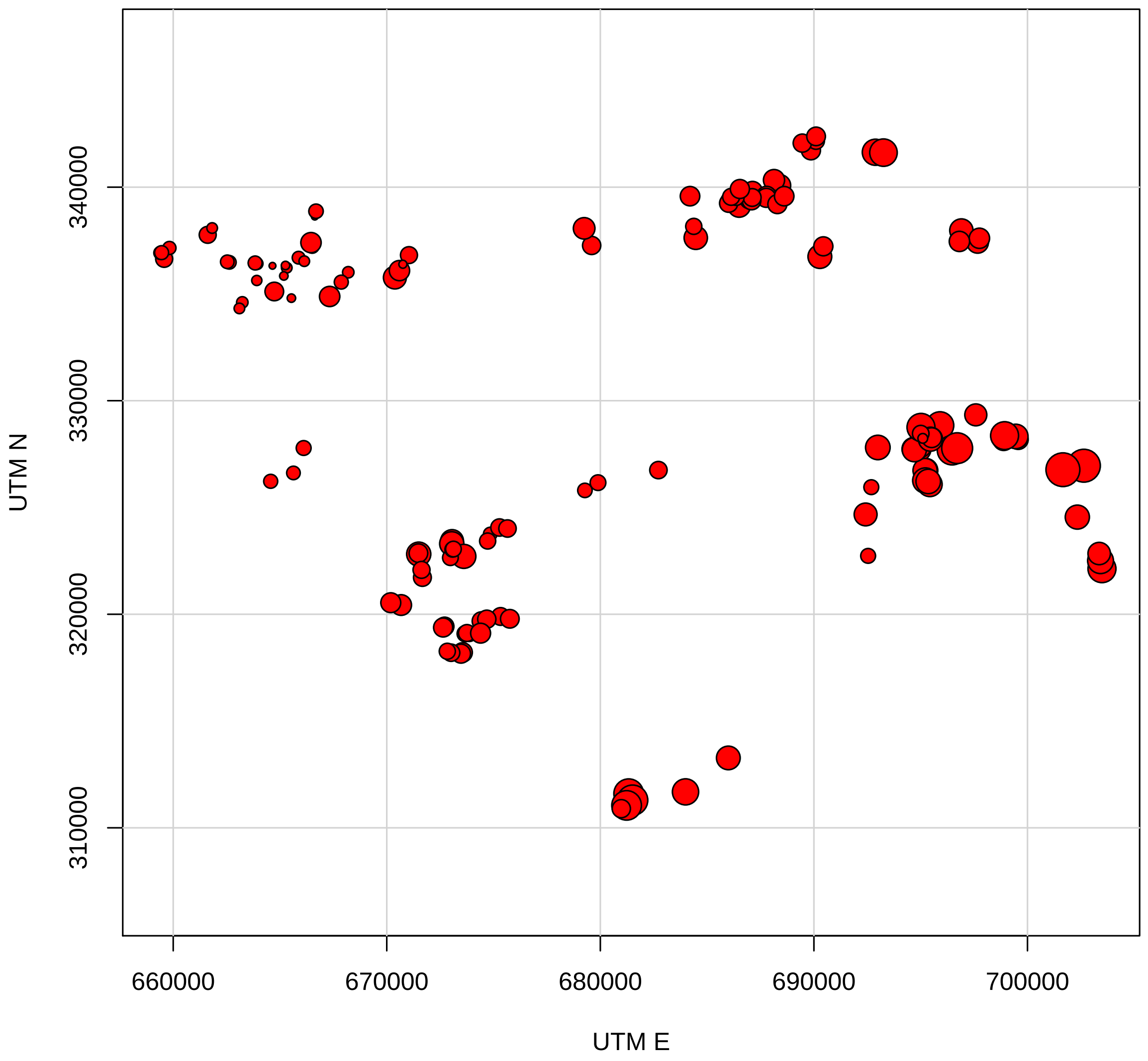
Looking for a trend in the post-plot

The post-plot shows **local** spatial dependence by the **clustering** of similar attribute values.

It can also show a **regional trend** if the size or colour of the symbols (related to the attribute values) **systematically change** over the whole plot.

Post-plot showing a regional trend

Tropenbos Cameroon Project, Clay content in 0−10~cm layer, %



To check your understanding . . .

Q15 : *Describe the regional pattern of the clay content shown in the previous graph.* *Jump to A15* •

Q16 : *How much of the variability in clay content explained by the trend? (No computation, just your impression of the plot.)* *Jump to A16* •

Visualizing a trend with a trend surface

From the set of points we can **fit an empirical model** which expresses the attribute as a function of the coördinates:

$$z = f(\mathbf{x})$$

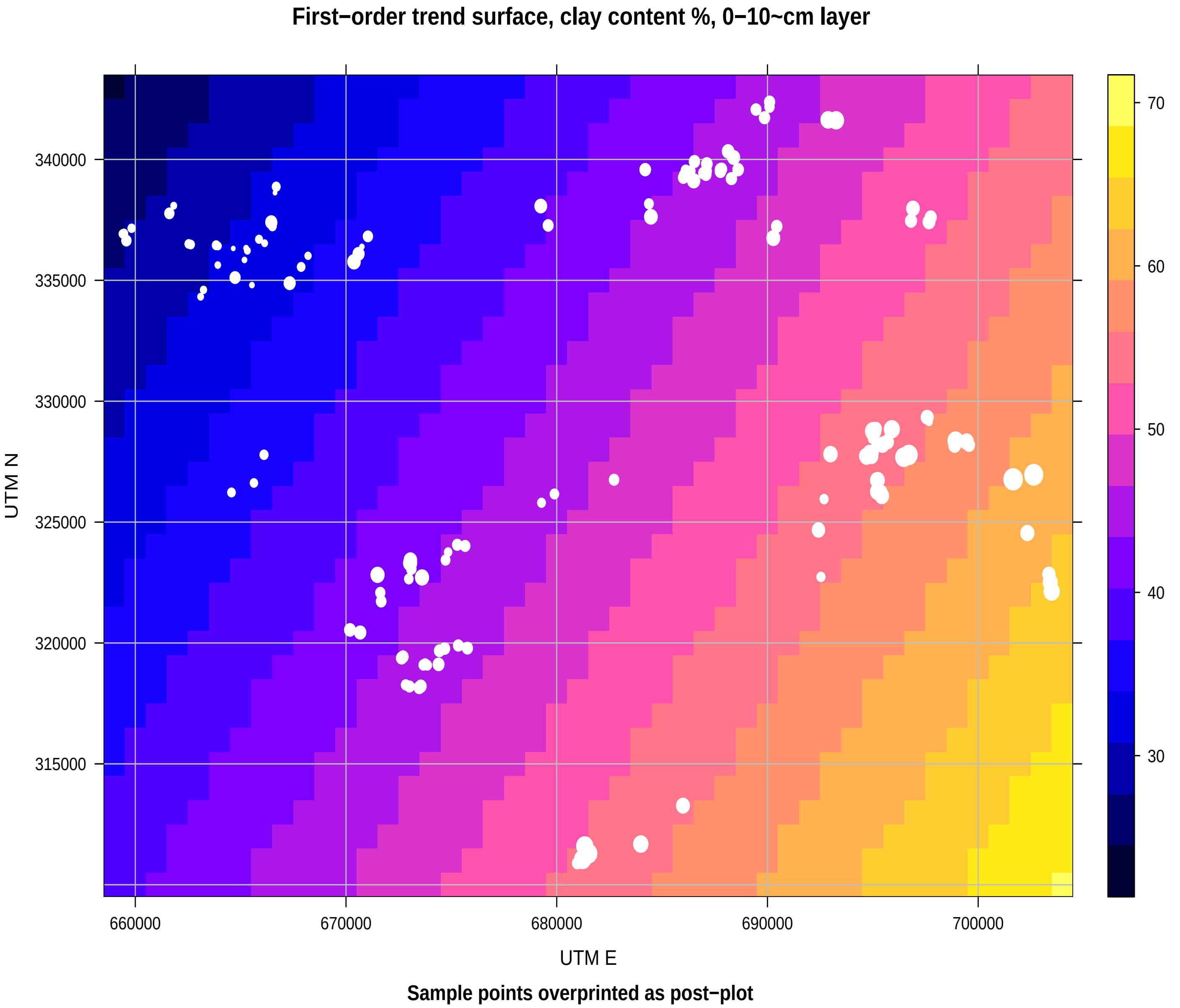
where $\mathbf{x}$ is a coördinate vector, e.g. in 2D it might be:

$$(x, y) = (\mathrm{UTM\_E}, \mathrm{UTM\_N})$$

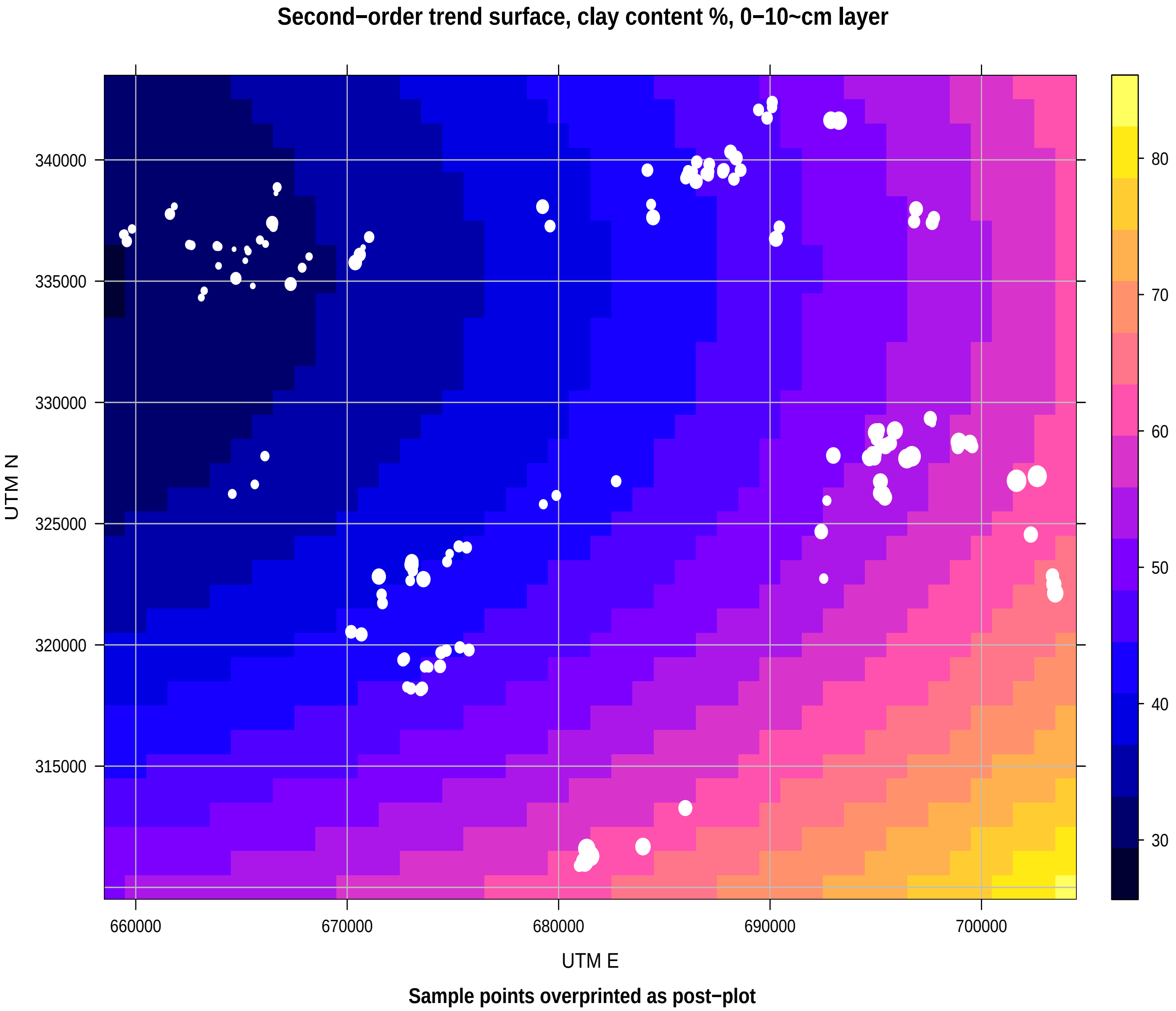
Then we apply the fitted function over a **regular grid** of points, and display this as a **map**. Normally the points are represented as pixels.

We will see how to fit the surface in a later lesson; for now we visualize the results.

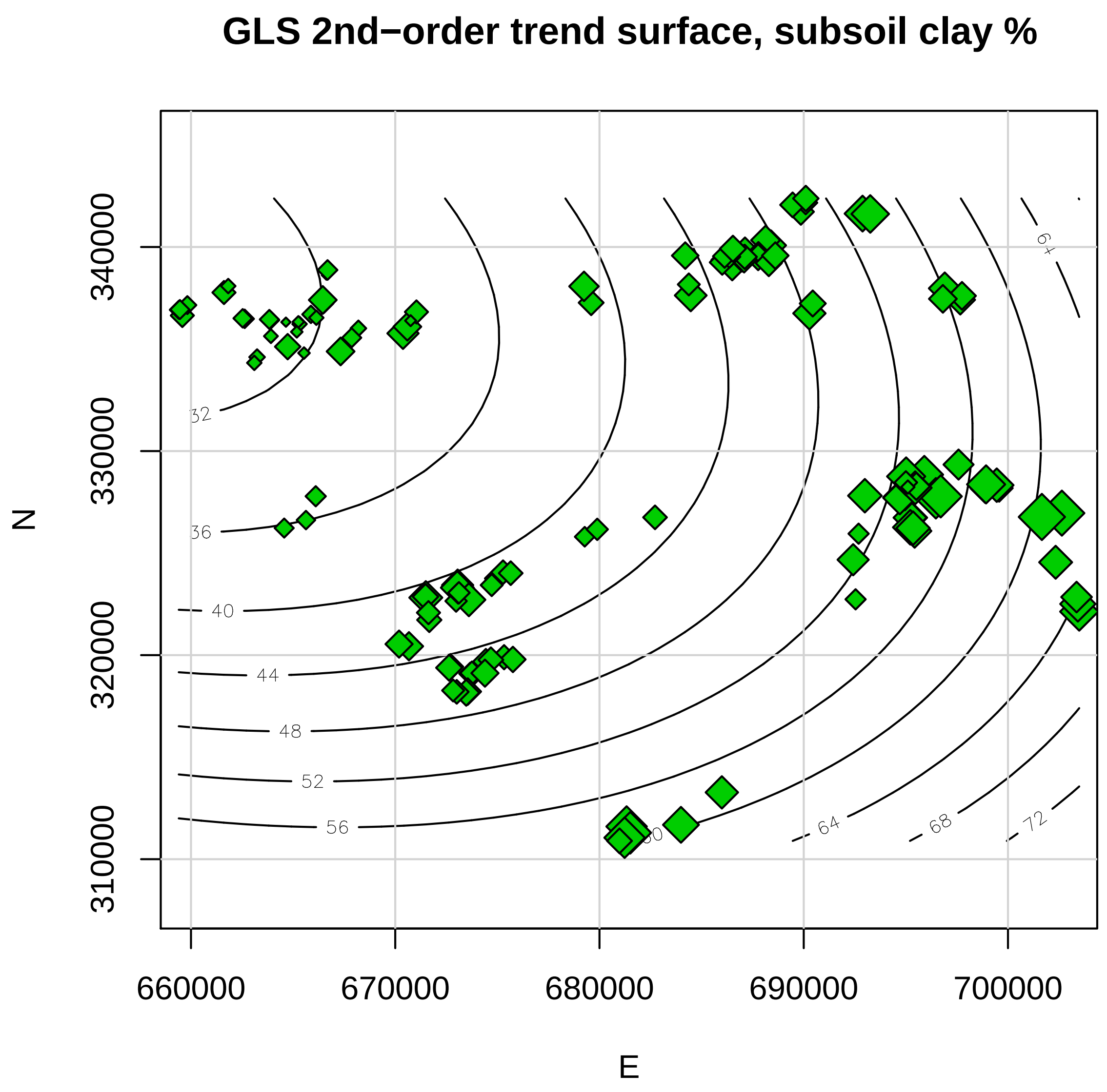
First-order trend surface


First−order trend surface, clay content %, 0−10~cm layer
UTM N
340000
335000
330000
325000
320000
315000
UTM E
660000
670000
680000
690000
700000
70
60
50
40
30
Sample points overprinted as post−plot

Second-order trend surface


Second−order trend surface, clay content %, 0−10~cm layer
UTM N
340000
335000
330000
325000
320000
315000
UTM E
660000
670000
680000
690000
700000
30
40
50
60
70
80
Sample points overprinted as post−plot

Second-order trend surface as a contour map

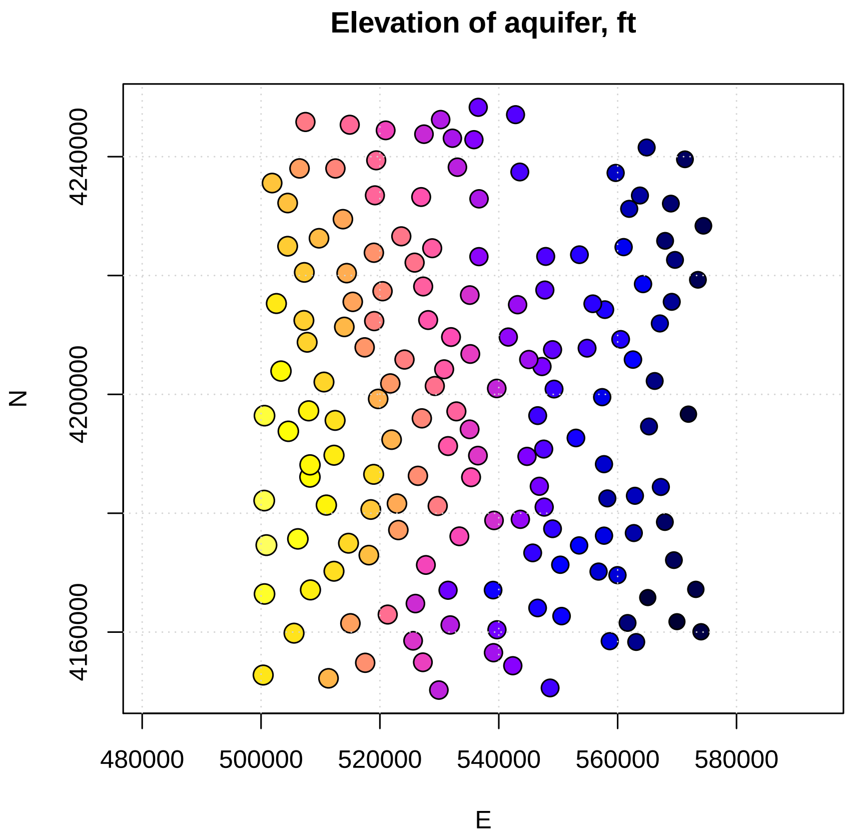


To check your understanding . . .

Q17 : *Which of these two trend surfaces best fits the sample points? (Compare the overprinted post-plot with the surface).* *Jump to A17* •

Q18 : *Is the second-order surface a* **bowl** *or* **dome***?* *Jump to A18* •

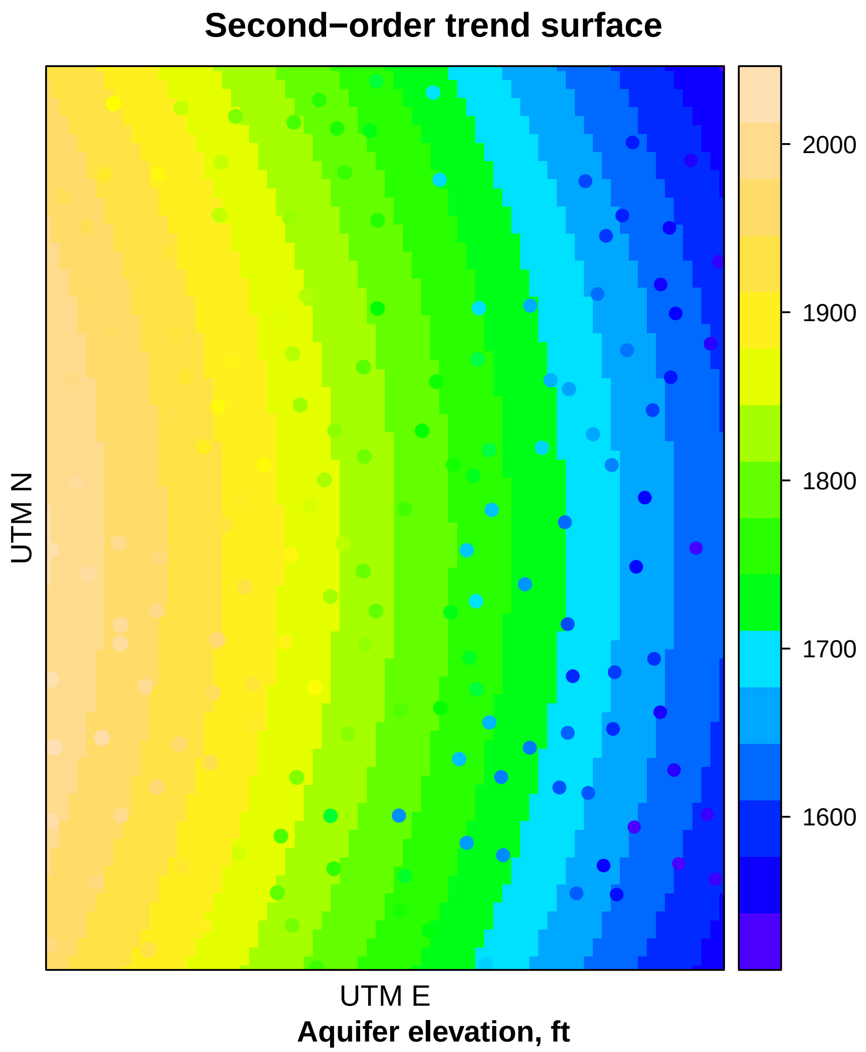
Another example of a trend surface: aquifer elevation



Elevation above mean sea level shown by circle size and colour

Source: R. A. Olea and J. C. Davis. *Sampling analysis and mapping of water levels in the High Plains aquifer of Kansas*. KGS Open File Report 1999-11, Kansas Geological Survey, May 1999. `http://www.kgs.ku.edu/Hydro/Levels/OFR99_11/`

Second-order trend surface fitting these elevations



Well elevations shown in the same colours as the trend surface; this shows the discrepency (residuals) from the best fit.

Topic 3: Visualizing local spatial dependence

1. Point-pairs

2. h-scatterplots

3. Variogram clouds

4. Empirical variograms

Commentary

We have seen in the postplots that there seems to be local spatial dependence. How can we visualize this?

The most common tool to evaluate local spatial dependence is the **empirical variogram**, which will be explained later. This is difficult for many people to understand when they first see this kind of graph.

So, we will begin with the concept of **point-pairs** and then see how we can relate point pairs at **lag distances** (to be defined).

Then the variogram should be easier to understand.

Point-pairs

Any two points are a **point-pair**.

If there are n points in a dataset, there are $(n * (n - 1)/2)$ **unique point-pairs**; that is, any of the n points can be compared to the other $(n - 1)$ points.

To check your understanding . . .

Q19 : *The Meuse data set has 155 sample points. How many* **point-pairs** *can be formed from these?*

Jump to A19 •

Q20 : *Are you surprised by the size of this number?* *Jump to A20* •

Comparing points in a point-pair

These can be compared two ways:

1. Their **locations**, i.e. we can find the **distance** and **direction** between them in **geographic** space;

2. Their **attribute values** in **feature space**.

The combination of **distance** and **direction** is called the **separation vector**.

Practical considerations for the separation vector

Except with gridded points, there are rarely many point-pairs with exactly the same separation vector.

Therefore, in practice we set up a **bin**, also called (for historical reasons) a **lag**, which is a **range** of distances and directions.

h-scatterplots

This is a **scatterplot** (i.e. two variables plotted against each other), with these two axes:

X-axis The attribute value at a point

Y-axis The attribute value at a second point, at some **defined distance** (and possibly direction, to be discussed as anisotropy, below), from the first point.

All **pairs of points** (for short usually called **point-pairs**) separated by the defined distance are shown on the scatterplot.

Interpreting the h-scatterplot

If there is **no relation** beween the values at the separation, the h-scatterplot will be a **diffuse cloud** with a **low correlation coefficient**.

If there is a **strong relation** beween the values at the separation, the h-scatterplot will be a **close to the 1:1 line** with a **high correlation coefficient**.

Commentary

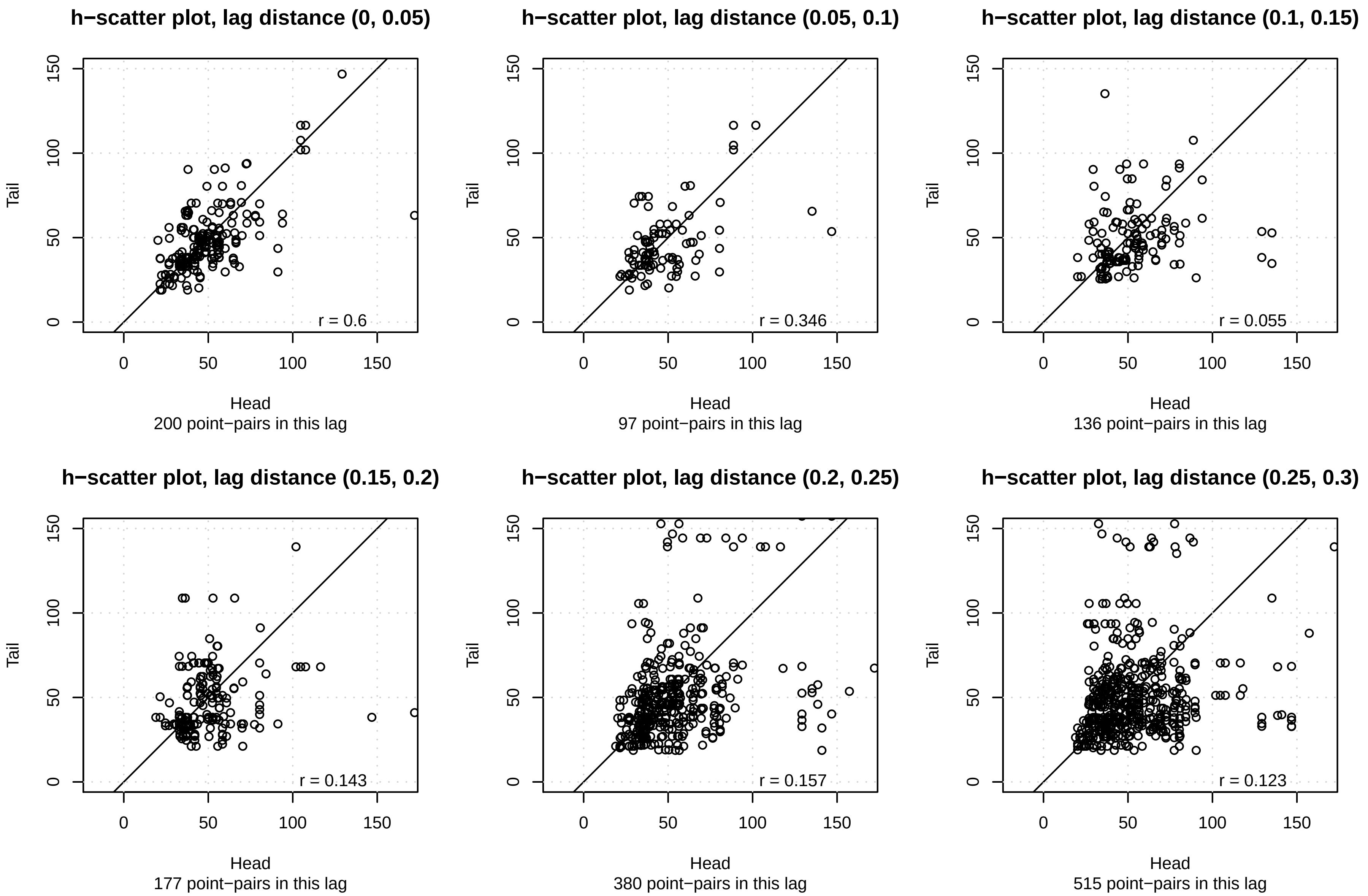
The next two slides show h-scatterplots for the Pb concentration in Jura soil samples at two resolutions:

- Bins of 50 m (0.05 km) width up to 300 m (0.3 km) separation
- Bins of 100 m (0.1 km) width up to 600 m (0.6 km) separation

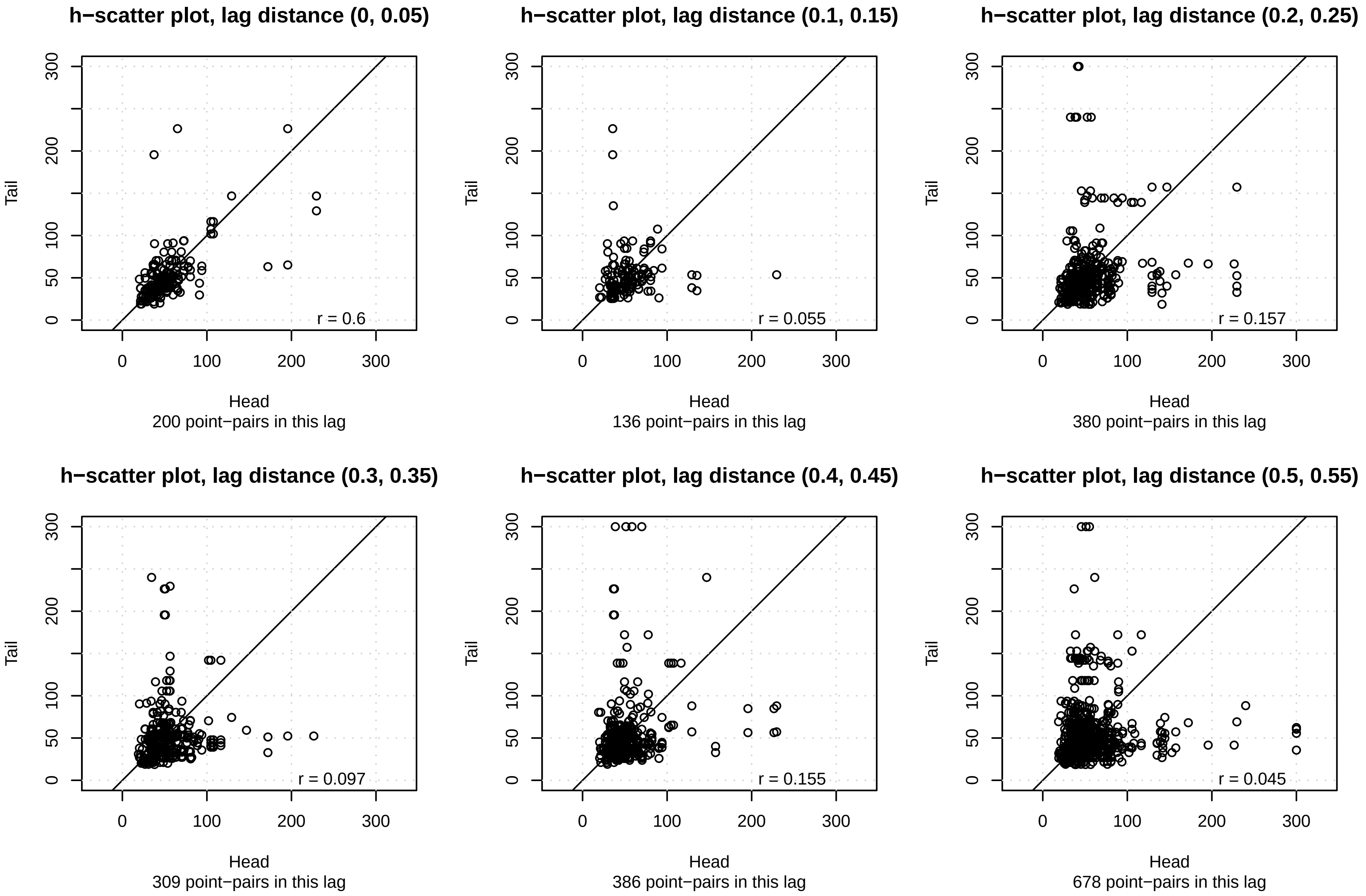
Note that the x- and y-axes are in units of the **attribute**, in this case ppm Pb.

The 50 m bins only show the lower part of the Pb attribute value range (to 150 ppm Pb); the 100 m bins show the full attribute range (to about 300 ppm Pb).

h-scatterplots for the Jura soil samples, Pb; 50 m bins to 0.3 km separation



h-scatterplots for the Jura soil samples, Pb; 100 m bins to 0.6 km separation


h−scatter plot, lag distance (0, 0.05)

h−scatter plot, lag distance (0.1, 0.15)

h−scatter plot, lag distance (0.2, 0.25)

h−scatter plot, lag distance (0.3, 0.35)

h−scatter plot, lag distance (0.4, 0.45)

h−scatter plot, lag distance (0.5, 0.55)

To check your understanding . . .

Q21 : *Describe the "evolution" of the cloud of point pairs as the separation distance increases.* *Jump to A21* •

Commentary

The most common way to visualize local spatial dependence is the **variogram**, also called (for historical reasons) the **semivariogram**.

To understand this, we have to first define the **semivariance** as a mathematical measure of the difference between the two points in a point-pair.

Semivariance

This is a mathematical measure of the **difference** between the two points in a point-pair. It is expressed as **squared difference** so that the order of the points doesn't matter (i.e. subtraction in either direction gives the same results).

Each **pair** of observation points has a **semivariance**, usually represented as the Greek letter γ ('gamma'), and defined as:

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

where $\mathbf{x}$ is a geographic point and $z(\mathbf{x})$ is its **attribute value**.

(Note: The 'semi' refers to the factor $1/2$, because there are two ways to compute for the same point pair.)

So, the **semivariance** between two points is half the **squared difference** between their values. If the values are **similar**, the semivariance will be **small**.

To check your understanding . . .

Q22 : *What are the* **units of measure** *for the semivariance?*

Jump to A22 •

Q23 : *Here are the first two points of Jura soil sample dataset:*

	coordinates	Rock	Land	Cd	Cu	Pb	Co	Cr	Ni	Zn
1	(2.386, 3.077)	Sequanian	Meadow	1.740	25.72	77.36	9.32	38.32	21.32	92.56
2	(2.544, 1.972)	Kimmeridgian	Pasture	1.335	24.76	77.88	10.00	40.20	29.72	73.56

For this point-pair, compute:

1. *The Euclidean* **distance** *between the points;*
2. *The* **difference** *between the Pb values;*
3. *The* **semivariance** *between the Pb values;*

Jump to A23 •

Commentary

Now we know two things about a point-pair:

1. The **distance** between them in **geographic** space;
2. The **semivariance** between them in **attribute** space.

So . . . it seems natural to see if points that are **'close by' in geographical space** are also **'close by' in attribute space**.

This would be **evidence** of **local spatial dependence**.

The variogram cloud

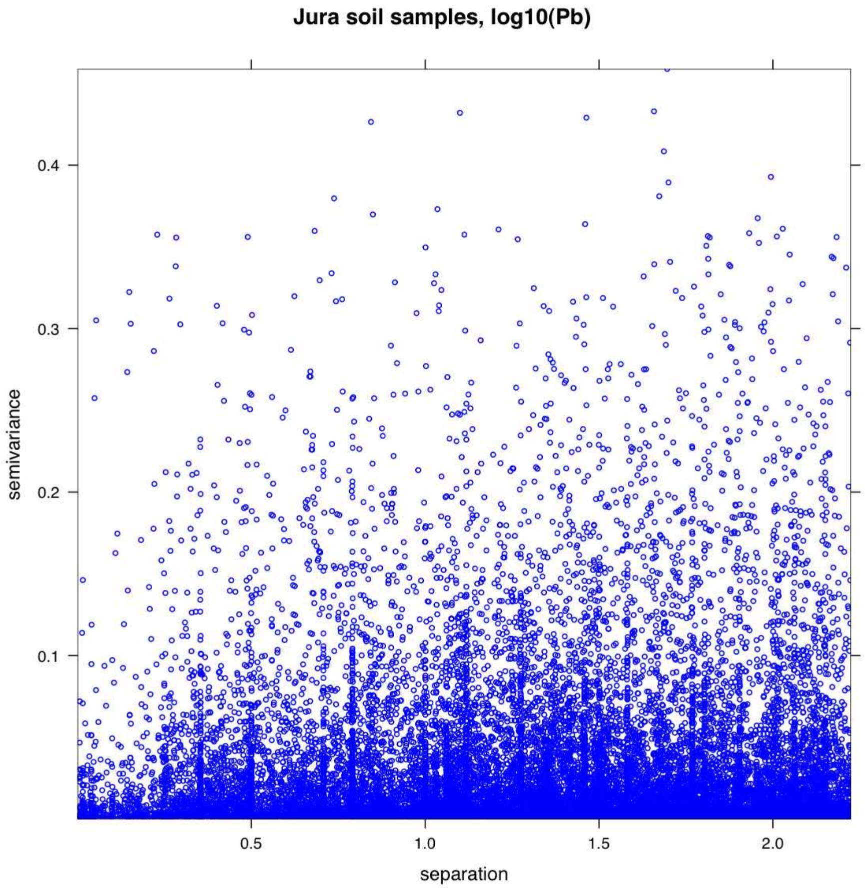
This is a graph showing **semivariances** between all point-pairs:

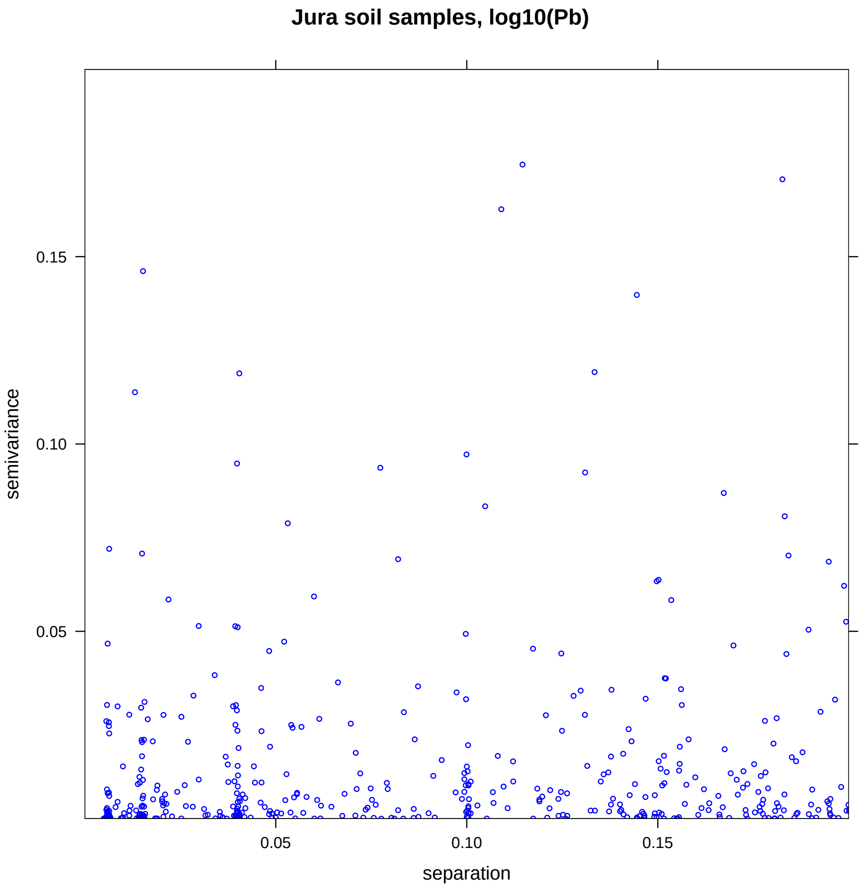
X-axis The **separation distance** within the point-pair

Y-axis The **semivariance**

- **Advantage**: Shows the comparaison between all point-pairs as a function of their separation;
- **Advantage**: Shows which point-pairs do not fit the general pattern
- **Disadvantage**: too many graph points, hard to interpret

Examples of a variogram cloud
Left: separations to 2 km; Right: separations to 200 m





To check your understanding . . .

Q24 : *Can you see a trend in the semi-variances as the separation distance increases?* *Jump to A24* •

Q25 : *What is the difficulty with interpreting this graph?* *Jump to A25* •

Commentary

Clearly, the variogram cloud gives too much information. If there is a relation between separation and semi-variance, it is hard to see. The usual way to visualize this is by **grouping** the point-pairs into **lags** or **bins** according to some separation range, and computing some **representative semi-variance** for the entire lag.

Often this is the arithmetic average, but not always.

The empirical variogram

- To **summarize** the variogram cloud, group the separations into **lags** (separation bins, like a histogram)

- Then, compute the **average** semivariance of all the point-pairs in the bin

- This is the **empirical variogram**, as the so-called **Matheron estimator**:

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

  ∗ $m(\mathbf{h})$ is the number of **point pairs** separated by vector $\mathbf{h}$, in practice some range (bin)

  ∗ These are indexed by i; the notation $z(\mathbf{x}_i + \mathbf{h})$ means the "tail" of point-pair i, i.e. separated from the "head" $\mathbf{x}_i$ by the separation vector $\mathbf{h}$.

Another mathematical formulation

Isaaks & Srivastava[1] present another way to write this formula, which may be more intuitive:

$$\overline{y}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{(i,j)|\mathbf{h}_{ij}=\mathbf{h}} (z_i - z_j)^2$$

Here we consider **points** numbered $1, 2, \ldots, i, \ldots, j, \ldots n$, i.e. the list of points, out of which we make **point-pairs**.

- $\mathbf{h}_{ij}$ is the distance between points i and j
- the notation $(i, j)|$ reads "pairs of points indexed as i and j, **such that** ...
- $= \mathbf{h}$ means that this point-pair has the separation vector. In practice $\mathbf{h}$ is some small range, the **bin** (see next).

[1] (1990) *An introduction to applied geostatistics*, Oxford University Press

Commentary

The variogram estimator given above is subject to high-leverage observations: one very unusual value has great influence when included in the squared differences.

This is the same issue with, e.g., linear regression by least-squares.

So, other ways to estimate the empirical variogram have been developed, so-called **robust** estimators, by analogy with other "robust" statistical methods; these are not unduly influenced by a few unusual values.

- For general robust statistics see e.g., Maronna, R.A., R.D. Martin, and V.J. Yohai. 2006. *Robust statistics: theory and methods*. John Wiley & Sons, Ltd, Chichester, West Sussex, England.
- For robust variogram estimators, see the review of Lark, R.M. 2000. *A comparison of some robust estimators of the variogram for use in soil survey*. European Journal of Soil Science 51(1): 137â€“157.

We continue with the standard Matheron estimator.

Defining the bins

There are some practical considerations, just like defining bins for a histogram:

- Each bin should have enough points to give a robust estimate of the representative semi-variance; otherwise the variogram is **erratic**;
- If a bin is too wide, the **theoretical variogram model** will be hard to estimate and fit; note we haven't seen this yet, it is in the next lecture;
- The **largest separation** should not exceed **half** the **longest separation** in the dataset;
- In general the largest separation should be somewhat shorter, since it is the **local** spatial dependence which is most interesting.

All computer programs that compute variograms use some **defaults** for the largest separation and number of bins; `gstat` uses 1/3 of the longest separation, and divides this into 15 equal-width bins.

Numerical example of an empirical variogram

Here is an empirical variogram of $\log_{10}$Pb from the Jura soil samples; for simplicity the maximum separation was set to 1.5 km:

```
     np     dist    gamma
1   262 0.037054 0.014245
2   197 0.151837 0.020510
3   363 0.255571 0.031182
4   565 0.353183 0.027535
5   608 0.452477 0.031559
6   607 0.538099 0.029263
7   615 0.651635 0.031891
8   980 0.755513 0.031293
9   753 0.851272 0.031229
10  705 0.951857 0.030979
11 1167 1.048865 0.031923
12 1066 1.140061 0.033649
13 1134 1.254365 0.035221
14 1130 1.350202 0.033418
15 1235 1.450247 0.036693
```

`np` are the **number of point-pairs** in the bin; `dist` is the **average separation** of these pairs; `gamma` is the **average semivariance** in the bin.

To check your understanding . . .

Q26 :

1. *What is the* **minimum and maximum separation** *for bin 2?*
2. **How many point-pairs** *are in this bin 2?*
3. *What is the* **average separation** *of all the point-pairs in bin 2?*
4. *What is the* **average semivariance** *of all the point-pairs in bin 2?*

Jump to A26 •

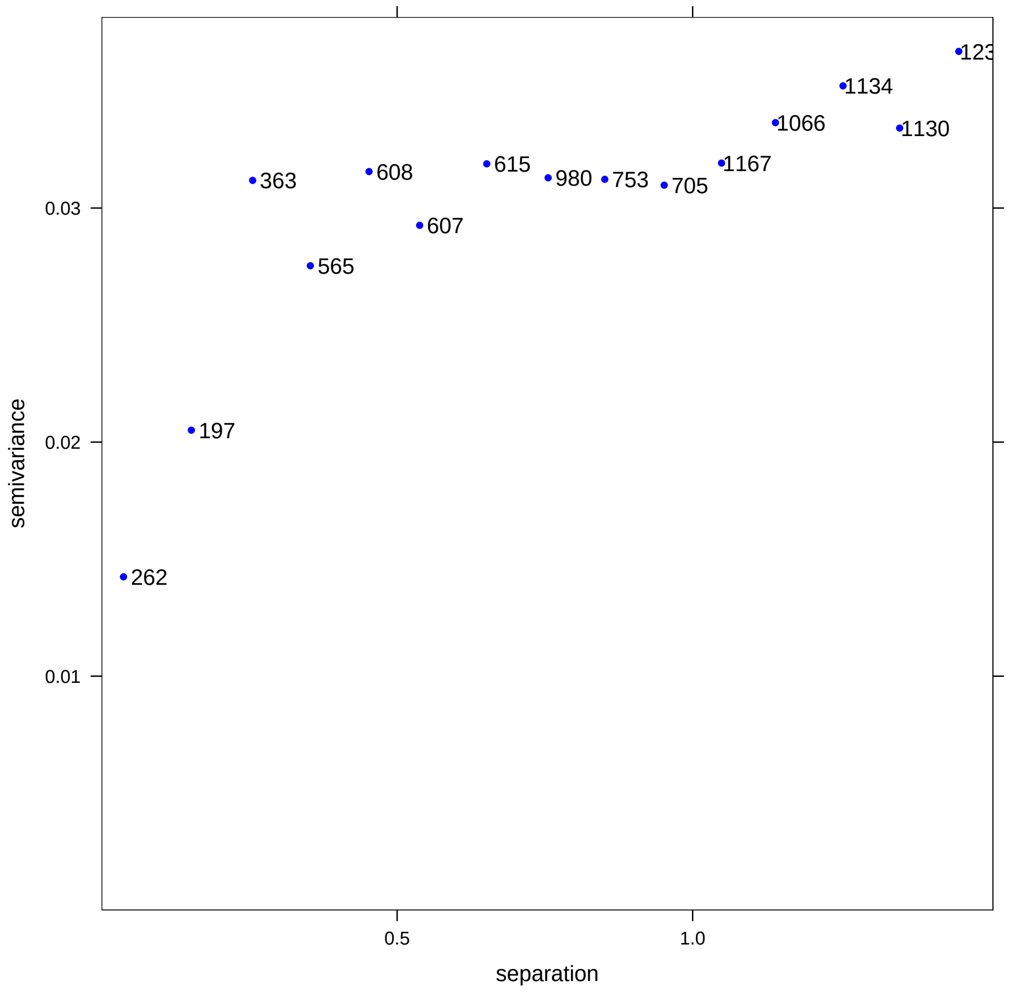
Q27 : *What is the trend in the* **average** *semi-variances as the* **average** *separation distance increases?*

Jump to A27 •

Plotting the empirical variogram

This can be plotted as semivariance `gamma` against average separation `dist`, along with the number of points that contributed to each estimate `np`

Empirical variogram of $\log_{10}$Pb, Jura soil samples



To check your understanding . . .

Now looking at the variogram plot, rather than the table:

Q28 :

1. *How many point-pairs are in the bin with the closest separation?*
2. *What is the average separation of all the point-pairs in this bin? (You will have to estimate by eye from the graph)*
3. *What is the average semivariance of all the point-pairs in this bin?(You will have to estimate by eye from the graph)*

Jump to A28 •

Features of the empirical variogram

Later we will look at fitting a **theoretical model** to the **empirical variogram**; but even without a model we can notice some features which characterize the **spatial dependence**, which we define here only qualitatively:

- **Sill**: maximum semi-variance
  - represents variability in the absence of spatial dependence
- **Range**: separation between point-pairs at which the sill is reached
  - distance at which there is no evidence of spatial dependence
- **Nugget**: semi-variance as the separation approaches zero
  - represents variability at a point that can't be explained by spatial structure

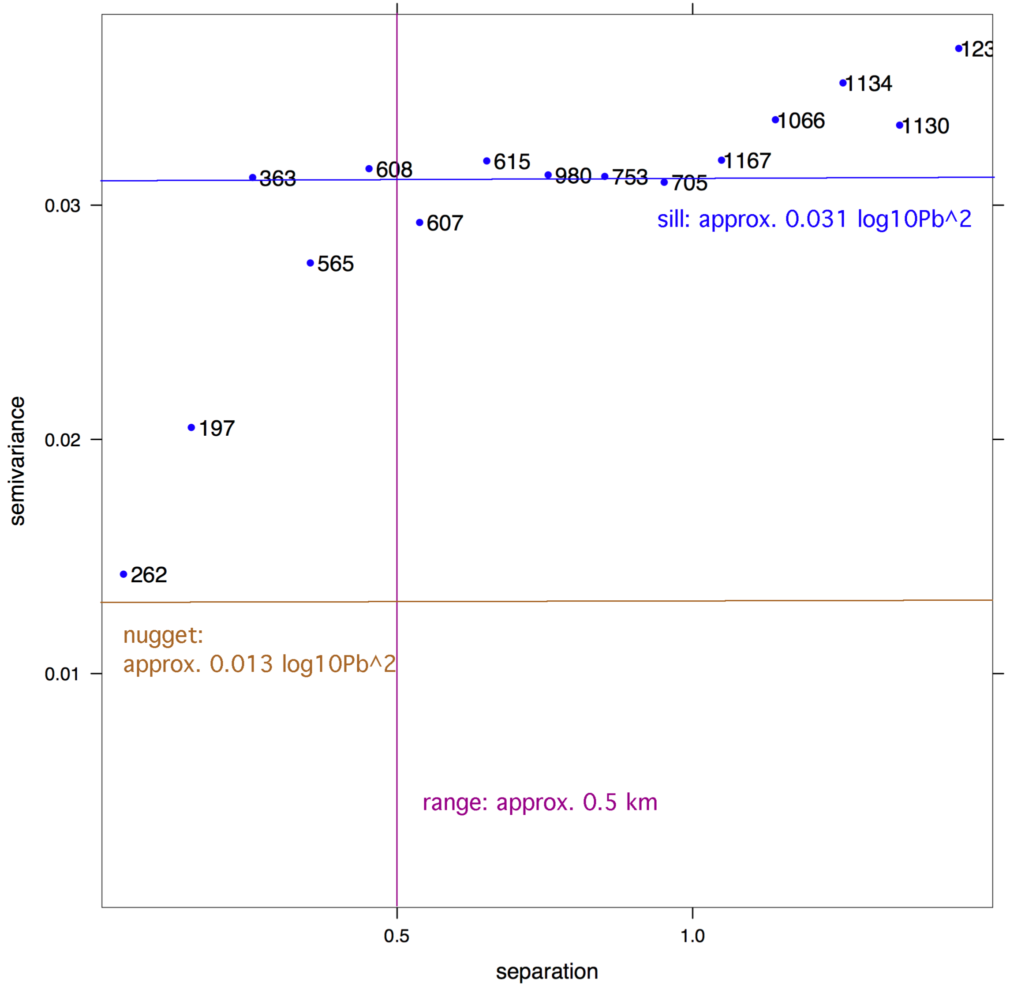
To check your understanding . . .

Q29 :

1. *What is the approximate* **sill** *of the empirical variogram for* $log_{10}Pb$ *from the Jura soil samples (previous graph)?*
2. *What is the approximate* **range** *of this empirical variogram?*
3. *What is the approximate* **nugget** *of this empirical variogram?*

Jump to A29 •

Empirical variogram of $\log_{10}$Pb, Jura soil samples annotated with approximate range, sill, nugget

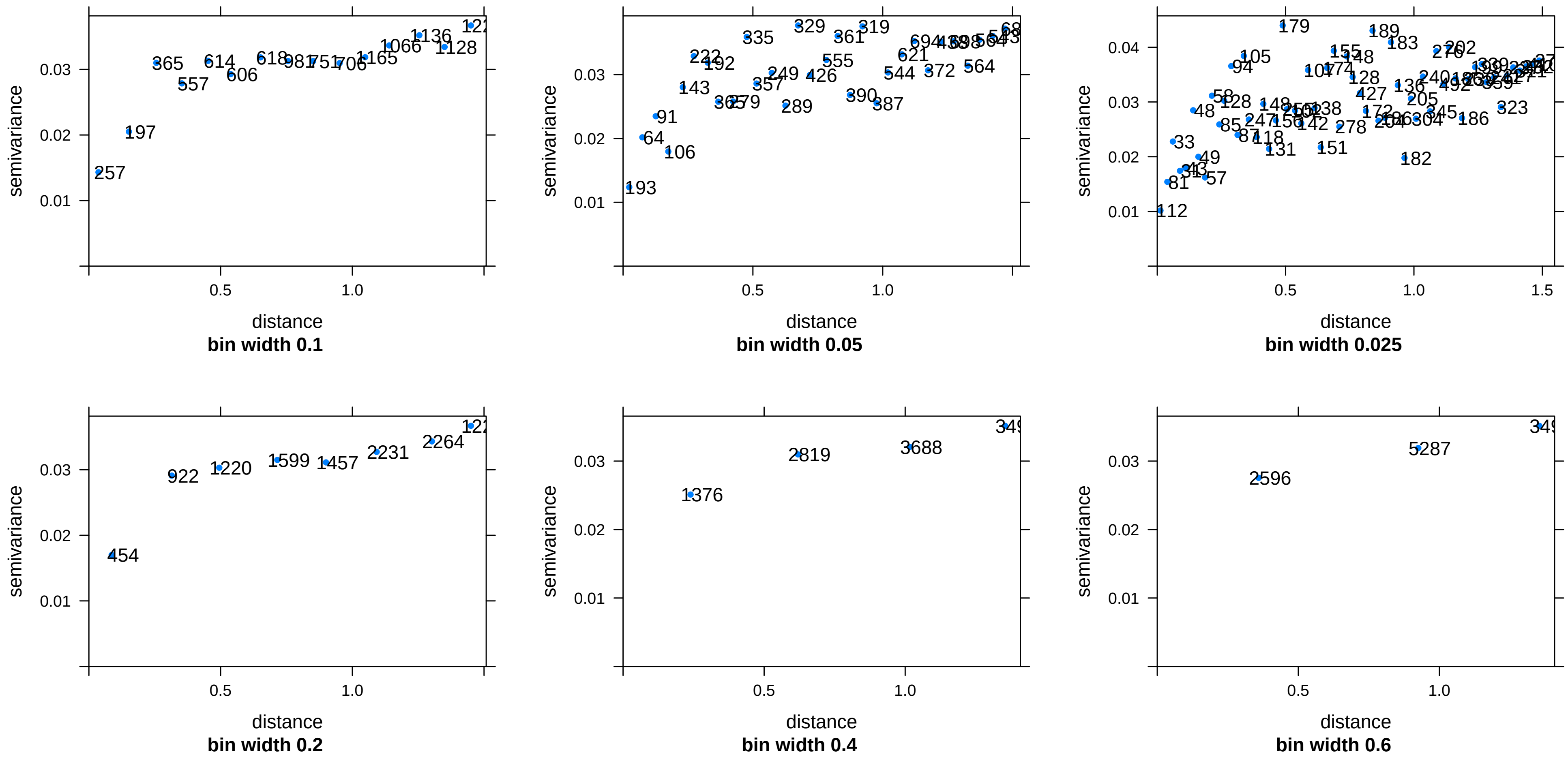


Effect of bin width

- The **same** set of points can be displayed with many **bin widths**
- This has the same effect as different bin widths in a univariate histogram: **same data, different visualization**
- In addition, **visual** and especially **automatic** variogram **fitting** is affected
- **Wider** (fewer) bins → **less** detail, also less noise
- **Narrower** (more) bins → **more** detail, but also more noise
- General rule:

  1. as narrow as possible (detail) without "too much" noise;
  2. and with **sufficient point-pairs** per bin ($>$ 100, preferably $>$ 200)

Variograms of Log(Cd), Meuse soil pollution with different bin widths



bin width 0.1

bin width 0.05

bin width 0.025

bin width 0.2

bin width 0.4

bin width 0.6

To check your understanding . . .

Q30 : *Which bin width gives the "best" summary of this empirical variogram? Which ones give unhelpful views?*

Jump to A30 •

Evidence of spatial dependence

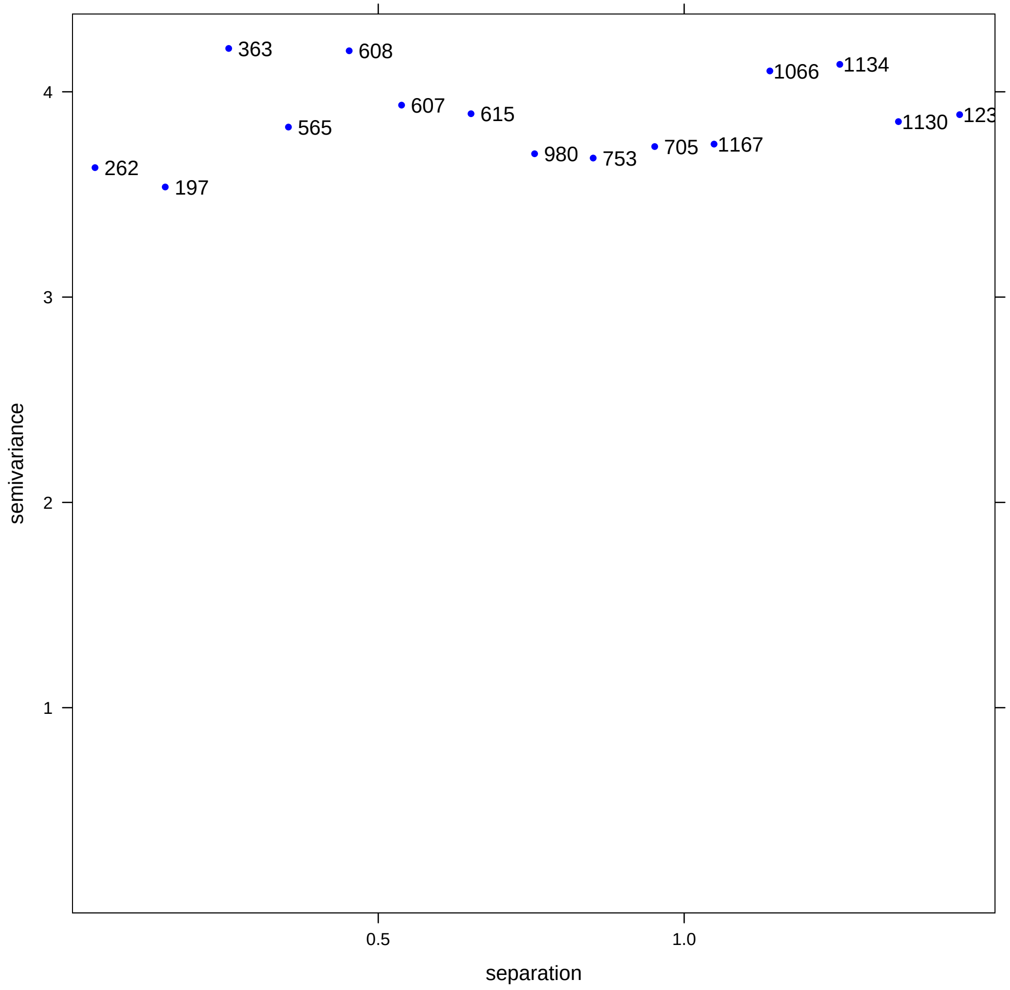
The empirical variogram provides **evidence** that there is **local spatial dependence**.

- The **variability** 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 dependence.
- The **relative magnitude** of the **total sill** and **nugget** give the **strength** of the local spatial dependence; the **nugget** represents completely **unexplained** variability.

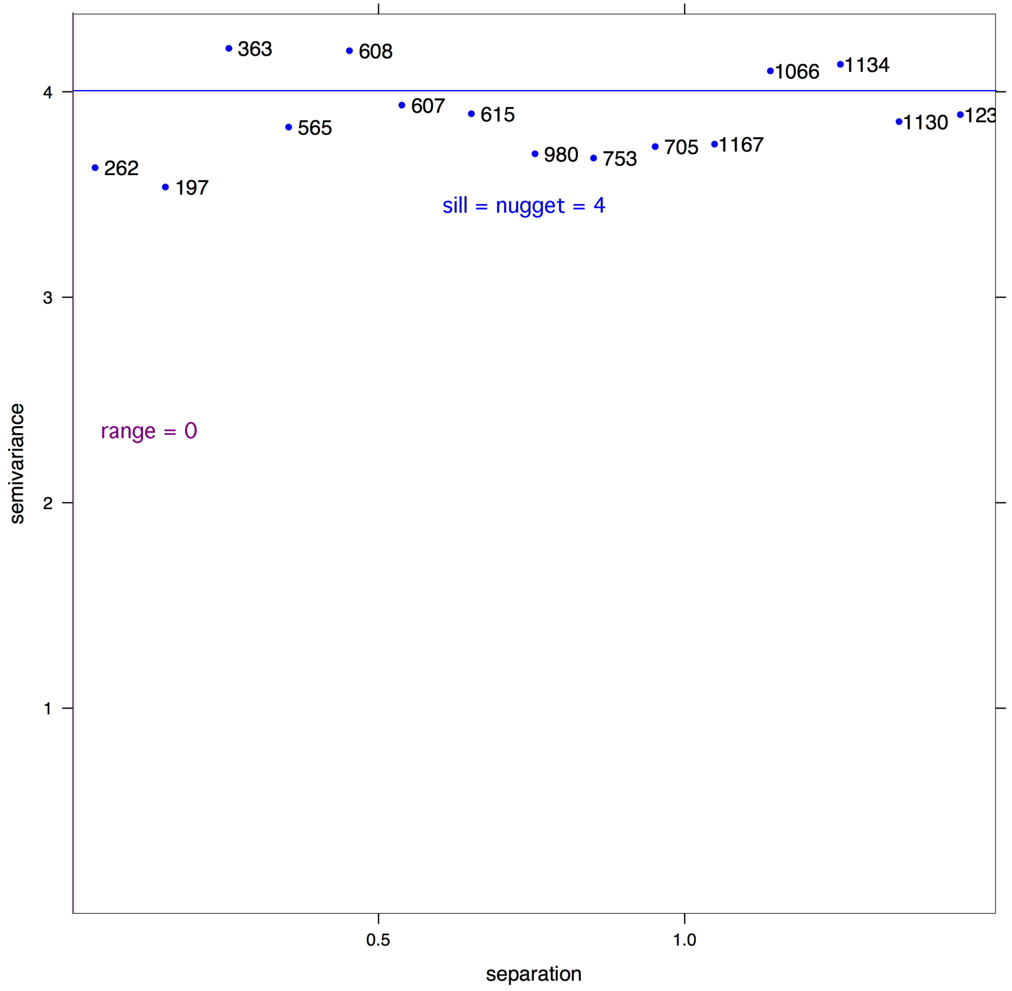
There are of course variables for which there is **no spatial dependence**, in which case the empirical variogram has the **sill equal to the nugget**; this is called a **pure nugget effect**

The next graph shows an example.

Empirical variogram of a variable with no spatial dependence



Empirical variogram of a variable with no spatial dependence annotated with range, sill, nugget



Topic 4: Visualizing anisotropy

1. Anisotropy
2. Variogram surfaces
3. Directional variograms

Commentary

We have been considering spatial dependence as if it is the same in all directions from a point (**isotropic** or **omnidirectional**).

For example, if I want to know the weather at a point where there is no station, I can equally consider stations at some distance from my location, no matter whether they are N, S, E or W.

But this is self-evidently not always true! In this example, suppose the winds almost always blow from the North. Then the temperatures recorded at stations 100 km to the N or S of me will likely be closer to the temperature at my station than temperatures recorded at stations 100 km to the E or W.

We now see how to detect anisotropy.

Anisotropy

- Greek *"Iso"* + *"tropic"* = English "same" + "trend"; Greek *"an-"* = English "not-"
- Variation may depend on **direction**, not just distance
- This is why we refer to the separation **vector**; up till now this has just meant distance, but now it includes direction
- Case 1: same sill, different ranges in different directions (**geometric**, also called **affine**, anisotropy)
- Case 2: same range, sill varies with direction (**zonal** anisotropy)

How can anisotropy arise?

- Directional **process**
  - Example: sand content in a narrow flood plain: much greater spatial dependence along the axis parallel to the river
  - Example: secondary mineralization near an intrusive dyke along a fault
  - Example: population density in a hilly terrain with long, linear valleys

- Note that the **nugget** must logically be **iso**tropic: it is variation *at* a point (which has no direction)

To check your understanding . . .

Q31 : *What is the* **physical** *reason you would expect* **greater spatial dependence** *(i.e. more similarity in values) of the sand content along the axis* **parallel** *to the river than in the axis* **perpendicular** *to it? Jump to A31* •

How do we detect anisotropy?

1. Looking for directional patterns in the **post-plot**;
2. With a **variogram surface**, sometimes called a **variogram map**;
3. Computing **directional variograms**, where we only consider points separated by a **distance** but also in a given hldirection from each other.

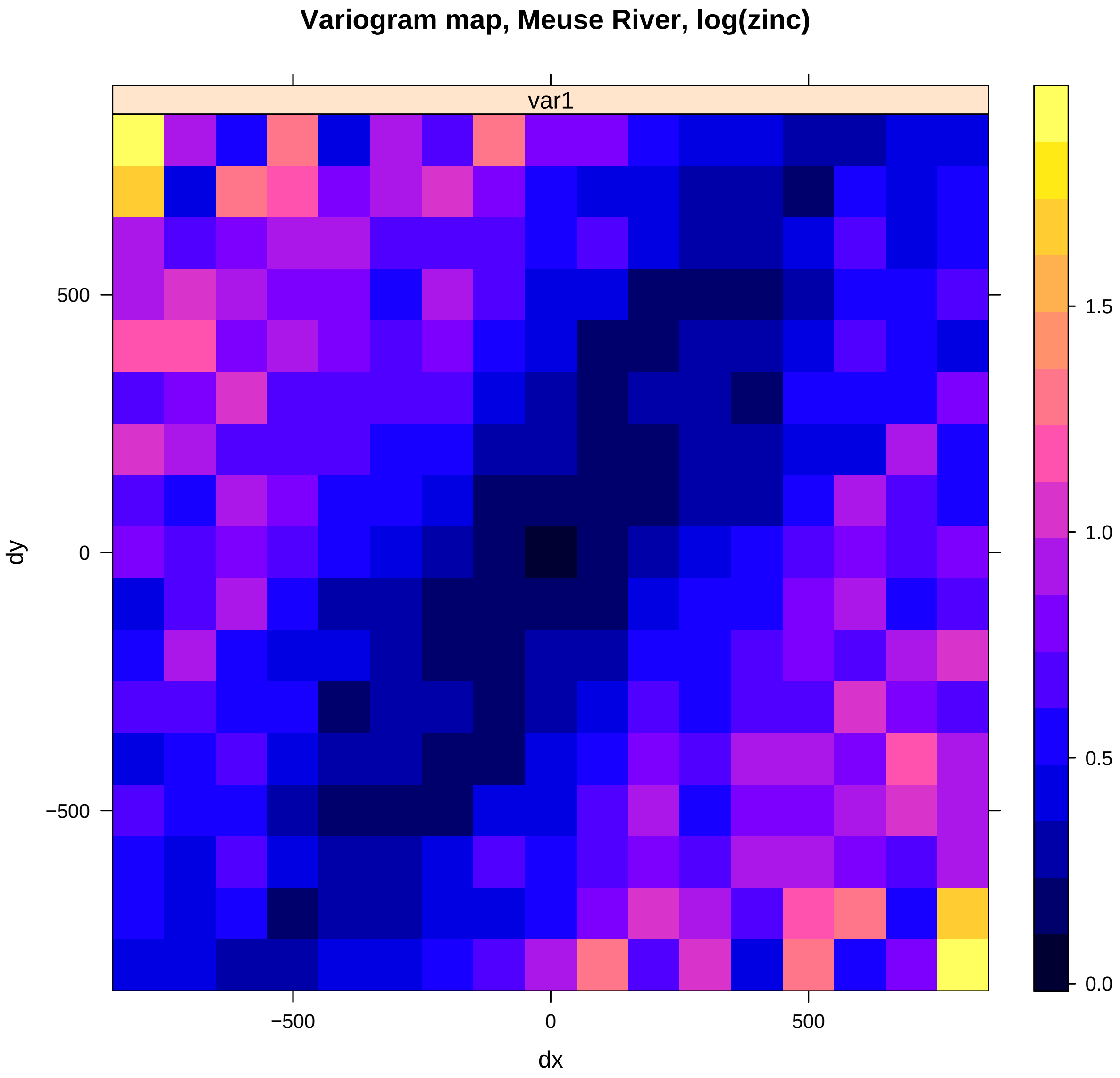
We can compute different directional variograms and see if they have different structure.

Detecting anisotropy with a variogram surface

One way to see anistropy is with a **variogram surface**, sometimes called a **variogram map**.

- This is *not* a map! but rather a plot of semivariances vs. distance and direction (the **separation vector**)
- Each grid cell shows the **semivariance** at a given **distance and direction** separation (lag)
- Symmetric by definition, can be read in either direction
- A **transect** from the origin to the margin gives a **directional variogram** (next visualization technique)

Variogram surface showing anisotropy



To check your understanding . . .

Looking at the variogram surface of the previous slide:

Q32 : *What is the approximate* **semivariance** *at a separation of* **distance** *500 m,* **direction** *due E (or W)?* *Jump to A32* •

Q33 : *What is the approximate separation* **distance** *for the cell at 300 m E, 200 m S?* *Jump to A33* •

Q34 : *What is the approximate separation* **azimuth** *(direction from N) in this cell?* *Jump to A34* •

Q35 : *What is the approximate* **semivariance** *at this separation?* *Jump to A35* •

Interpreting the variogram surface

The principal use is to find the **direction of maximum spatial dependence**, i.e. the lowest semivariances at a given distance.

To do this, start at the centre and look for the direction where the colour stays the same (or similar).

To check your understanding . . .

Looking at the variogram surface of the previous slide:

Q36 : *Which direction (as an azimuth from N) shows the* **strongest spatial dependence***, i.e. where the semivariance stays low over the farthest distance?* *Jump to A36* •

Q37 : *Does the* **orthogonal** *axis, i.e. 90° rotated from the principal axis of spatial dependence, appear to have the* **weakest spatial dependence***, i.e. where the semivariance increases most rapidly away from the centre of the map?* *Jump to A37* •

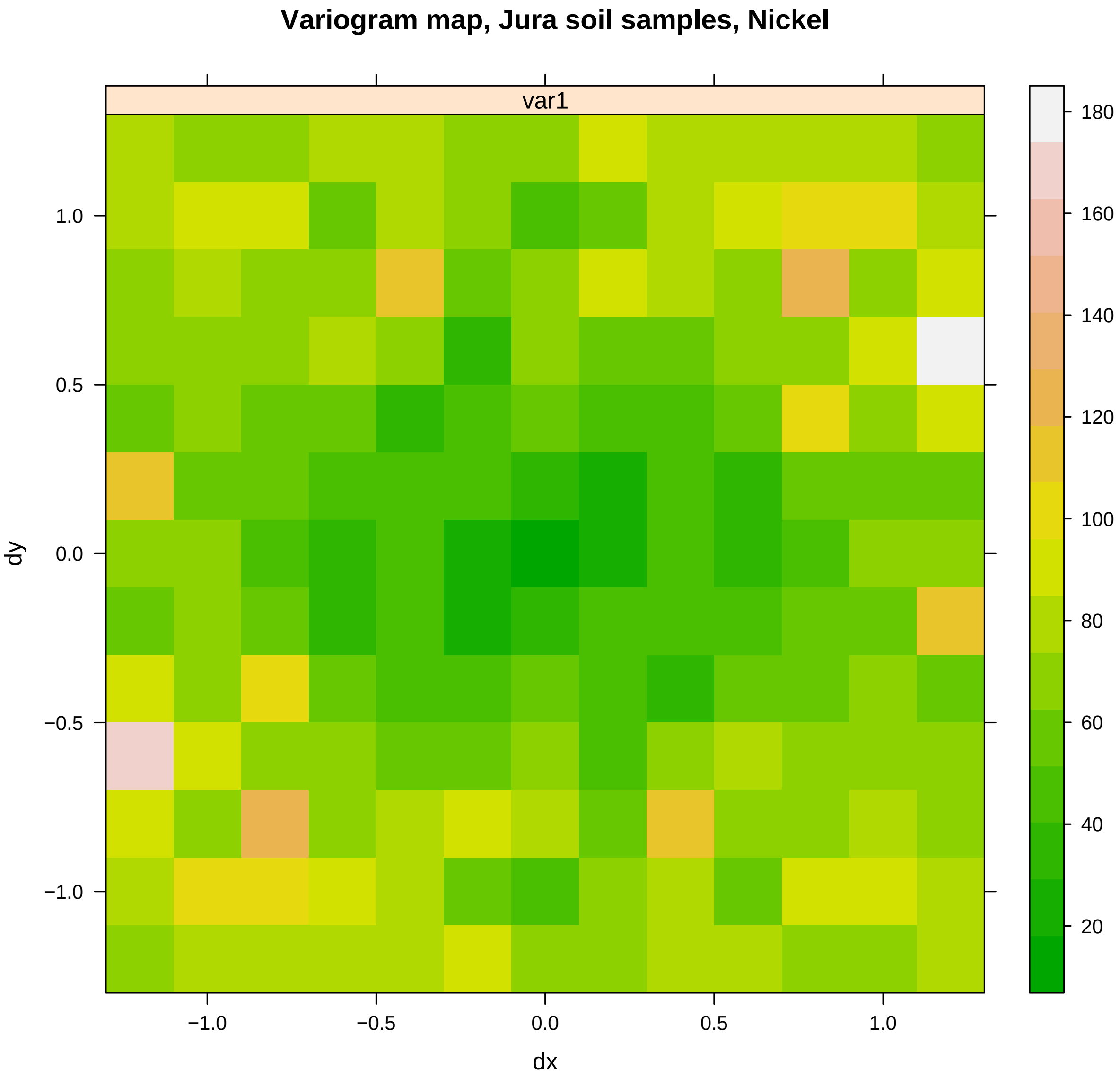
Commentary

We saw above the “pure nugget” variogram, showing that not all variables have spatial dependence.

Similarly, not all variables show anisotropy; in fact many do not. The following graph shows a variable with **isotropic** (same in all directions) spatial dependence.

Note: variogram maps are often “irregular” in appearance (“speckled”) because in most datasets there are **few point-pairs** to estimate the semivariance at a given distance and direction (separation) bin.

Variogram surface showing isotropy



Detecting anisotropy with a directional variogram

- A **directional variogram** only considers point pairs separated by a certain direction

- These are put in bins defined by distance classes, within a certain directional range, as with an omnidirectional variogram

- These parameters must be specified:

1. **Maximum distance** (cutoff) and **lag spacing** (width of variogram bins) as for omnidirectional variograms;
2. **Direction** of the major axis expressed as azimuths from 0° N to <180° N; implicitly specifies perpendicular ($+90^\circ$) as the minor axis;
3. **Tolerance**: Degrees on either side which are considered to have the 'same' angle;
4. **Band width**: Limit the sector to a certain width; this keeps the sector from taking in too many far-away points. Note: `gstat` does not implement this.

Parameters of the directional variogram

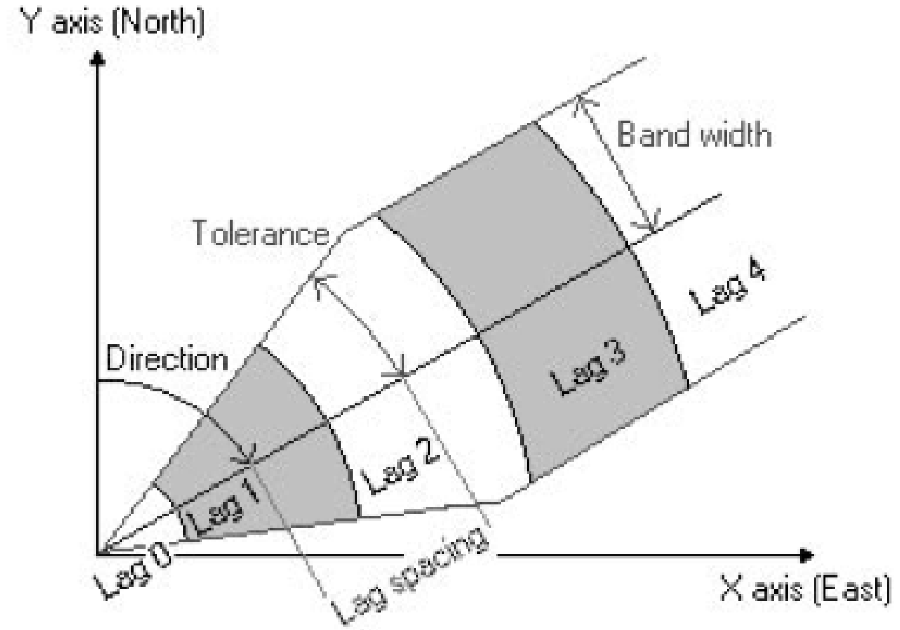
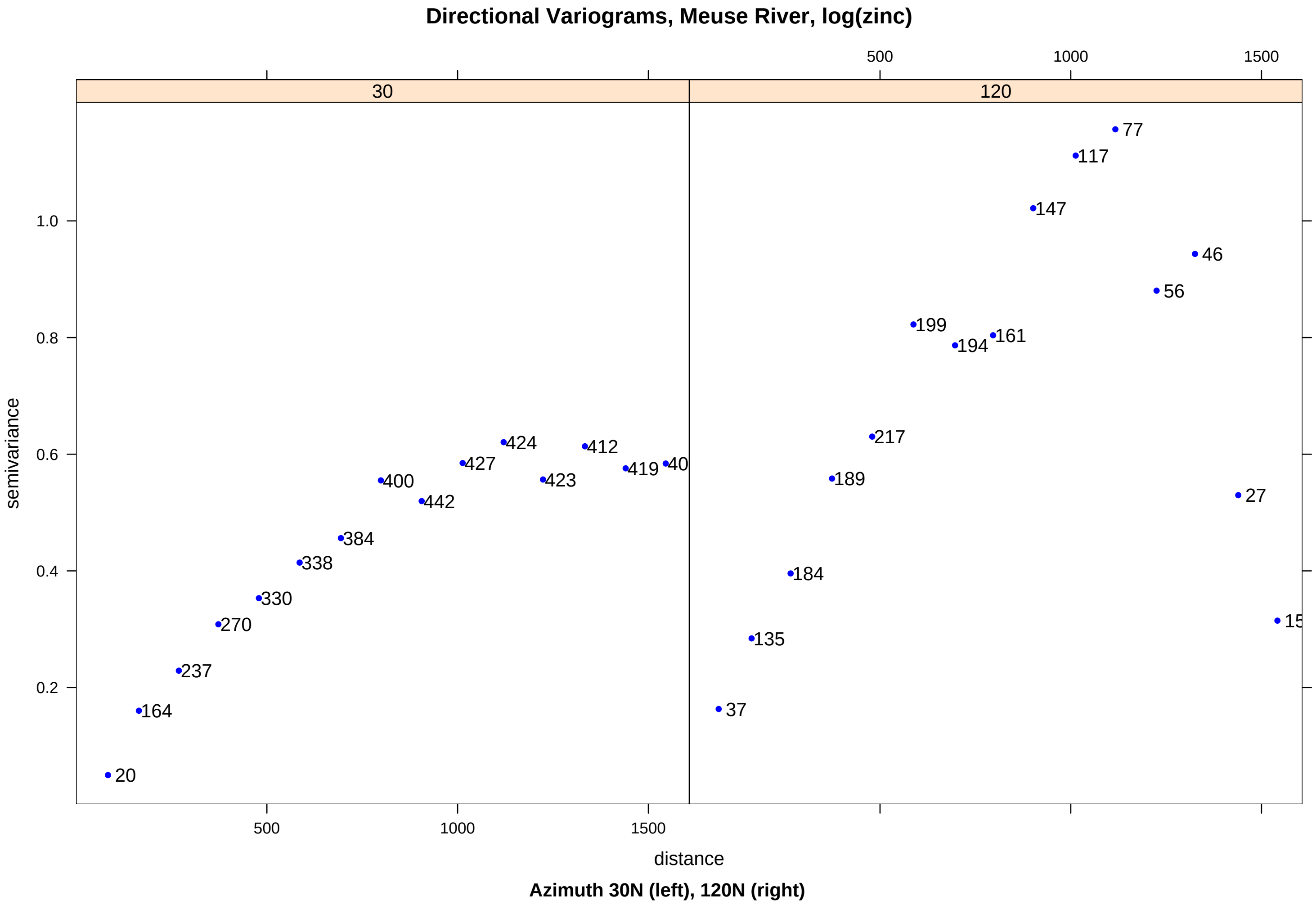


Figure 11.5: Schematic explanation of the input parameters for Spatial Correlation bidirectional method: lag spacing, nr of lags, direction, tolerance and band-width.

Source: Unit Geo Software Development. (2001). ILWIS 3.0 Academic user's guide. Enschede: ITC.

Directional variograms of Log(Zn), Meuse soil samples



Azimuth 30N

Azimuth 120N

To check your understanding . . .

Looking at the two directional variograms:

Q38 : *Do the two directions have similar variograms? (Consider sill, range, nugget)* *Jump to A38* •

Q39 : *In which of the two perpendicular axes is the spatial dependence stronger (longer range, lower nugget to sill ratio)?* *Jump to A39* •

Q40 : *How is this evidence that the spatial* **process** *by which the metal (Zn) was distributed over the area is directional?* *Jump to A40* •

Exercise

At this point you should complete **Exercise 2: Spatial visualisation** which is provided on the module CD.

This should take several hours.

1. Visualising spatial structure: postplots, quantile plots;
2. Visualizing regional trends
3. Visualizing spatial dependence: h-scatterplots, variogram cloud, experimental variogram
4. Visualizing anisotropy: variogram surfaces, directional variograms

In all of these 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.

Answers

Q1 : *Do the points cover the entire 5x5 km square? What area do they cover?* •

A1 : *They only cover one part; this is the* **study area**. *Unfortunately we do not have a map of boundary of the study area.* *Return to Q1* •

Q2 : *Within the* **convex hull** *of the points (i.e. the area bounded by the outermost points), is the distribution:*

1. *Regular or irregular?*
2. *Clustered or dispersed?*

•

A2 : *Irregular (i.e. not in a regular pattern); some small clusters but mostly well-distributed over the study area.* *Return to Q2* •

Answers

Q3 : *Do the points cover the entire 1x1 km square?* •

A3 : *No, they do not cover the NE or SE corners; however there are points outside the area (not visible in this plot) that are near to these.* *Return to Q3* •

Q4 : *Within the* **convex hull** *of the points (i.e. the area bounded by the outermost points), is the distribution:*

1. *Regular or irregular?*
2. *Clustered or dispersed?*

•

A4 : *Irregular; mostly clustered.* *Return to Q4* •

Answers

Q5 : *Which of this two graphs best shows the highest values ("hot spots")?* •

A5 : *The plot where the radius is proportional to the square root of the value; the points with the highest values are much large than the others and clearly highlighted.* *Return to Q5* •

Q6 : *Which of this two graphs best shows the distribution of the values in feature space?* •

A6 : *The plot where the radius is proportional to the value; taking the square root makes it difficult to see the difference between most of the values, i.e. only the hots spots stand out.* *Return to Q6* •

Answers

Q7 : *Which of the two ramps (blue/purple/yellow and cyan/magenta) do you think better highlights the hot spots?* •

A7 : *No correct answer here; it depends on the psychology of the viewer. The CM shows the highest values in bright pink but the colours are somewhat pastel and difficult to distinguish; the BPY shows the highest values in yellow against a mostly blue background.* *Return to Q7* •

Q8 : *Do you prefer the size or colour to show the feature-space value?* •

A8 : *Personal preference.* *Return to Q8* •

Answers

Q9 : *Are any of the deciles concentrated in parts of the area?* •

A9 : *The first decile appears to be concentrated in the N, other deciles are distributed over the entire area.*

Return to Q9 •

Q10 : *Are points in any or all of the deciles clustered?* •

A10 : *Yes, points in all deciles form small clusters.* *Return to Q10* •

Answers

Q11 : *What is the* **classifying** *attribute in this postplot?* •

A11 : *Soil type (four classes)* *Return to Q11* •

Q12 : *What is the* **continuous** *attribute in this postplot?* •

A12 : *Topsoil clay content, %* *Return to Q12* •

Q13 : *Does the clay content appear to depend, at least to some extent, on the soil type?* •

A13 : *Yes, the "red" class seems to have the highest values, the "blue" class the lowest, and the other two intermediate. There is, however, quite some variability within each soil type.* *Return to Q13* •

Answers

Q14 : *What* **order of trend surface** *would you* **expect** *from these situations:*

1. *Mean annual precipitation in a coastal region where there is a low coastal plain, bounded inland by a mountain range; both of these are more or less linear in a given direction;*
2. *The depth to a gas deposit trapped in a salt dome;*
3. *Clay content in a soil developed from a sedimentary series of shales (claystones), siltstones and sandstones, outcropping as parallel strata?*

•

A14 :

1. *First-order (linear, planar), increasing from the coast to the mountain (orographic effect on on-shore winds);*
2. *Second-order; a dome (shallowest in middle);*
3. *Higher-order (many parallel 'troughs' and 'ridges')*

Return to Q14 •

Answers

Q15 : *Describe the regional pattern of the clay content shown in the previous graph.* •

A15 : *It is lowest in the NW corner and increases towards the E and S.* Return to Q15 •

Q16 : *How much of the variability in clay content explained by the trend? (No computation, just your impression of the plot.)* •

A16 : *The big differences are explained, but within each of sample clusters there is some variability. An anomaly is found right in the middle of the plot near (680000, 325000) where a cluster of small values is found. Qualitatively, "much" of the variability is explained by the trend.* Return to Q16 •

Answers

Q17 : *Which of these two trend surfaces best fits the sample points? (Compare the overprinted post-plot with the surface).* •

A17 : *The second-order surface fits much better, because it matches all three clusters of high clay contents in the E, S, and NE of the area. The first-order surface under-estimates especially the NE cluster.* *Return to Q17* •

Q18 : *Is the second-order surface a* **bowl** *or* **dome***?* •

A18 : *It is a (elliptically-shaped) bowl; the lowest values are found in the NW corner and increase in all directions from there.* *Return to Q18* •

Answers

Q19 : *The Meuse data set has 155 sample points. How many* **point-pairs** *can be formed from these?* •

A19 : $(155 * 154)/2 = 11,935$ *Return to Q19* •

Q20 : *Are you surprised by the size of this number?* •

A20 : *Most people are surprised by the large number of pairs. Human intuition is not good at estimating combinations. An example is the birthday paradox: in any group of 23 or more people it is more likely that two of them share a birthday than not. This number seems very low intuitively, yet it is correct. See* `http://mathworld.wolfram.com/BirthdayProblem.html` *for the computation.* *Return to Q20* •

Answers

Q21 : *Describe the "evolution" of the cloud of point pairs as the separation distance increases.* •

A21 : *It becomes more diffuse, i.e. further from the 1:1 line.* *Return to Q21* •

Answers

Q22 : *What are the* **units of measure** *for the semivariance?* •

A22 : *The* **square** *of the units of measure of the variable.* *Return to Q22* •

Answers

Q23 : *Here are the first two points of Jura soil sample dataset:*

	coordinates	Rock	Land	Cd	Cu	Pb	Co	Cr	Ni	Zn
1	(2.386, 3.077)	Sequanian	Meadow	1.740	25.72	77.36	9.32	38.32	21.32	92.56
2	(2.544, 1.972)	Kimmeridgian	Pasture	1.335	24.76	77.88	10.00	40.20	29.72	73.56

For this point-pair, compute:

1. *The Euclidean* **distance** *between the points;*
2. *The* **difference** *between the Pb values;*
3. *The* **semivariance** *between the Pb values.* •

A23 :

1. $\sqrt{(2.386 - 2.544)^2 + (3.077 - 1.972)^2} = 1.1162$ *km*
2. $77.36 - 77.88 = -0.52$ *mg kg-1*
3. $0.5 \cdot (77.36 - 77.88)^2 = 0.1352$ *(mg kg-1)2*

Return to Q23 •

Answers

Q24 : *Can you see a trend in the semi-variances as the separation distance increases?* •

A24 : *As separation increases, so does semi-variance.* *Return to Q24* •

Q25 : *What is the difficulty with interpreting this graph?* •

A25 : *This is quite difficult to see because of the large number of low semi-variances; in the left graph there are hundreds of points almost on top of each other, making it very hard to get a sense of the average.*

In the right graph this is a bit clearer, but this only applies to the closest point-pairs. *Return to Q25* •

Answers

Q26 :

1. *What is the* **minimum and maximum separation** *for bin 2?*
2. **How many point-pairs** *are in this bin 2?*
3. *What is the* **average separation** *of all the point-pairs in bin 2?*
4. *What is the* **average semivariance** *of all the point-pairs in bin 2?*

•

A26 *:*

1. *0.1, 0.2 km (100 to 200 m); we specified are 15 bins, equally dividing 1.5 km.*
2. *197*
3. *0.151837 km (152 m); the middle of the range is 150 m; there is no reason why this set of 197 point-pairs, separated by 100 to 200 m, has to average this.*
4. *0.020510* $(\log_{10}(\text{mg kg}^{-1}))^2$

Return to Q26 •

Q27 : *What is the trend in the* **average** *semi-variances as the* **average** *separation distance increases?* •

A27 : *There is a definite increase: at closer separations the semi-variance is less. There is some "noise", the trend is not monotonic.* *Return to Q27* •

Answers

Now looking at the variogram plot, rather than the table:

Q28 :

1. **How many point-pairs** *are in bin with the* **closest** *separation?*
2. *What is the* **average separation** *of all the point-pairs in this bin? (You will have to estimate by eye from the graph)*
3. *What is the* **average semivariance** *of all the point-pairs in this bin?(You will have to estimate by eye from the graph)*

•

A28 *:*

1. *262*
2. $\approx 0.04\text{km}$
3. $\approx 0.013(\log_{10} \text{mg kg}^{-1})^2$

Return to Q28 •

Answers

Q29 :

1. *What is the approximate* **sill** *of the empirical variogram for* log_{10}*Pb from the Jura soil samples (previous graph)?*
2. *What is the approximate* **range** *of this empirical variogram?*
3. *What is the approximate* **nugget** *of this empirical variogram?*

•

A29 :

1. **Sill**: 0.031; *although the semivariance increases again after 1 km (see next)*
2. **Range**: *0.5 km for the above sill; although the semivariance increases again after 1 km, so there may be a double structure (next lecture)*
3. **Nugget**: 0.013; *extrapolate by eye to the y-axis.*

Don't worry if your answers are not exactly these; we will see how to get better values when we use the **empirical variogram** *to fit a* **variogram model**.

Return to Q29 •

Answers

Q30 : *Which bin width gives the "best" summary of this empirical variogram? Which ones give unhelpful views?* •

A30 : *In this case the default bin width (0.1 km) seems best: we can see the sill, nugget and range without too much noise. There are sufficient point-pairs in each bin.*

At smaller widths there is increasing noise, for a width of 0.025 km the variogram is almost unreadable. Even at a width of 0.05 km the short-range bins do not have enough point-pairs for reliable estimation.

At larger widths the detail is obscured. Even for 0.2 km there is only one bin that gives an idea of the nugget and range; at 0.6 km there is no information about model shape – it's impossible to guess how the spatial dependence is at close range ($<$ 0.4 km). There are very many point-pairs per bin, this will do no harm but is more than needed for reliable estimation.

Return to Q30 •

Answers

Q31 : *What is the* **physical** *reason you would expect* **greater spatial dependence** *(i.e. more similarity in values) of the sand content along the axis* **parallel** *to the river than in the axis* **perpendicular** *to it?* •

A31 : *The energy of the river is along its axis, so that when it floods the momentum of the floodwater keeps it flowing more-or-less along this axis. Also, topographic barriers to floodwater, such as river terraces, also tend to be along the main river axis.* *Return to Q31* •

Answers

Q32 : *What is the approximate semivariance at a separation of* **distance** *500 m,* **direction** *due E (or W)?* •

A32 : ≈ 0.8; *compare the purple colour with the legend at the right of the figure.* *Return to Q32* •

Q33 : *What is the approximate separation* **distance** *for the cell at 300 m E, 200 m S?* •

A33 : $\sqrt{300^2 + 200^2} \approx 360$ *Return to Q33* •

Q34 : *What is the approximate separation* **azimuth** *(direction from N) in this cell?* •

A34 : $\arctan(200/300) + \pi/2 \approx 2.16$ *radians from North; this is* $2.16 \cdot (180/\pi) \approx 124\,\mathrm{deg}$ *Return to Q34* •

Q35 : *What is the approximate semivariance at this separation?* •

A35 : *The cell at dx = +300, dy = -200 has a blue colour that corresponds to 0.5.* *Return to Q35* •

Answers

Q36 : *Which direction (as an azimuth from N) shows the* **strongest spatial dependence***, i.e. where the semivariance stays low over the farthest distance?* •

A36 : *Approximately 30 degrees from N. The dark blue colours form a clear band in the NNE - SSW axis.* *Return to Q36* •

Q37 : *Does the* **orthogonal** *axis, i.e. 90 degrees rotated from the principal axis of spatial dependence, appear to have the* **weakest spatial dependence***, i.e. where the semivariance increases most rapidly away from the centre of the map?* •

A37 : *Yes, the axis at approximately 120 degrees from N (30 + 90) does appear to be the direction in which semivariance increases most rapidly.* *Return to Q37* •

Answers

Q38 : *Do the two directions have similar variograms? (Consider sill, range, nugget)* •

A38 : *They are quite different. The 30N variogram has a very regular form, with almost zero nugget, range about 1100 m, and sill near 0.6. The 120N variogram is irregular (partly because of the smaller number of point-pairs in that direction), with a nugget near 0.1, a range that is quite difficult to estimate but which may be placed near 800 m where the first sill of about 0.8 is reached.*

Return to Q38 •

Q39 : *In which of the two perpendicular axes is the spatial dependence stronger (longer range, lower nugget to sill ratio)?* •

A39 : *30N.*

Return to Q39 •

Q40 : *How is this evidence that the spatial* **process** *by which the metal (Zn) was distributed over the area is directional?* •

A40 : *The longer range and lower variability in the 30N direction shows that whatever process distributed the Zn is oriented along that axis.* *Return to Q40* •