Geo406   Data Analysis in Geology

Dr. Mohamed El Alfy

melalfy@ksu.edu.sa

1 Topics to be Covered

Topic	No. of Weeks	Contact hours
Sampling methods	1	(1+1)
Data distributions	1	(1+1)
Precision and accuracy	1	(1+1)
Confidence intervals	1	(1+1)
Least squares methods	2	(2+2)
Correlation	1	(1+1)
Time series analysis	1	(1+1)
Multivariate techniques	2	(2+2)
Cluster analysis	1	(1+1)
Principal component analysis	1	(1+1)
Kriging	1	(1+1)
Geologic modeling	2	(2+2)

Schedule of Assessment Tasks for Students During the Semester

Assessment	Assessment task (eg. essay, test, group project, examination etc.)	Week due	Proportion of Final Assessment
1	1st exam	6	15
2	lab reports	Every 3 weeks	5
3	Mid-term exam	12	15
4	1st lab exam	6	5
5	2nd lab exam	12	10
6	Final Practical exam	Last week	10
7	Final exam	As scheduled by the College	40

A simple definition

Statistics:

“The determination of the probable from the possible”

What is “statistical analysis”?

- This term refers to a wide range of techniques to. . .
- 1. (Describe)
- 2. Explore
- 3. Understand
- 4. Prove
- 5. Predict

Geostatistics

- Geostatistical analysis is distinct from other spatial models in the statistics in that it assumes the region of study is continuous.
  - Observations could be taken at any point within the study area
  - Interpolation at points in between observed locations makes sense

Why Geostatistics?

- Classic statistics is generally devoted to the analysis and interpretation of un- certainties caused by limited sampling of a property under study.

- Geostatistics however deviates from classic statistics in that Geostatistics is not tied to a population distribution model that assumes, for example, all samples of a population are normally distributed and independent from one another.

Why Geostatistics?

- Most of the earth science data (rock properties, contaminant concentrations) can be highly skewed and/or possess spatial correlation (data values from locations that are closer together tend to be more similar than data values from locations that are further apart).

- To most geologists, the fact that closely spaced samples tend to be similar is not surprising since such samples have been influenced by similar physical and chemical depositional/transport processes.

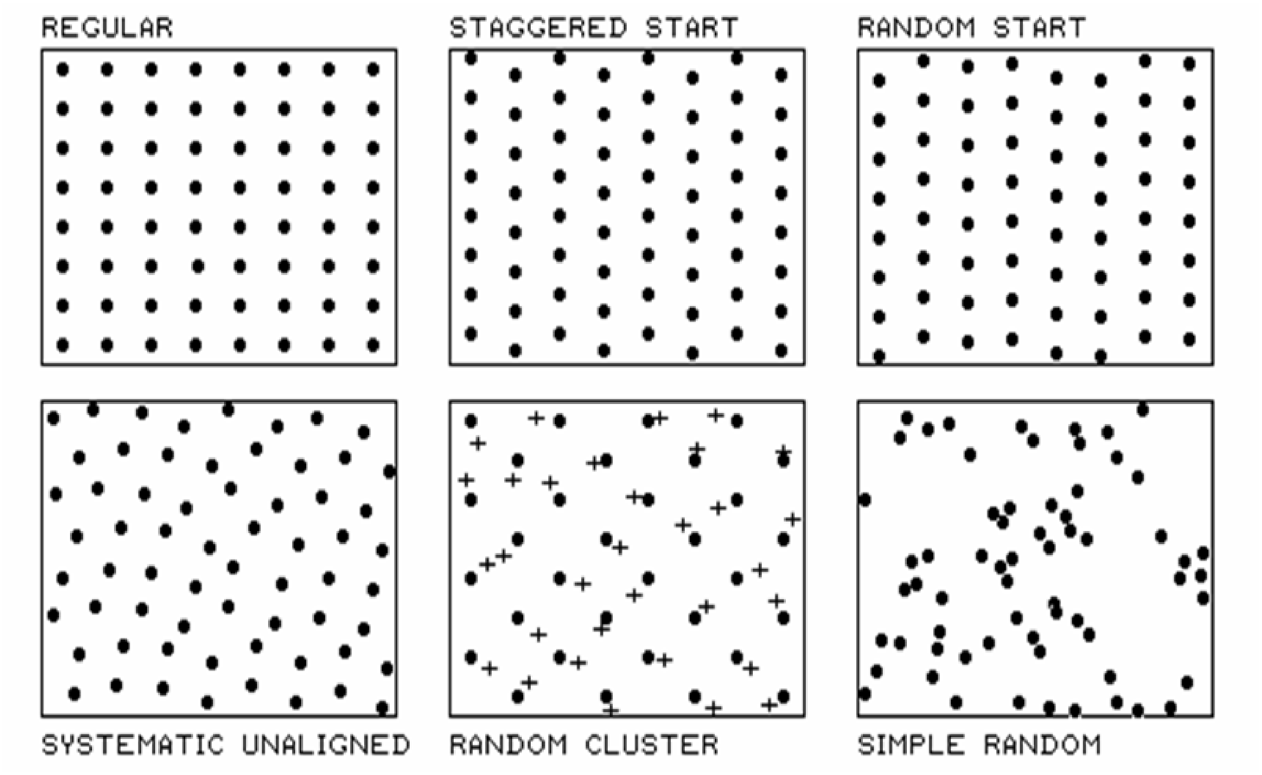
- Compared to the classic statistics which examine the statistical distribution of a set of sampled data, geostatistics incorporates both the statistical distribution of the sample data and the spatial correlation among the sample data

Spatial data exploration

- Sampling frameworks
  - Pure random sampling
  - Stratified random – by class/strata
  - Randomised within defined grids
  - Uniform
  - Uniform with randomised offsets
  - Sampling and declustering

Spatial data exploration

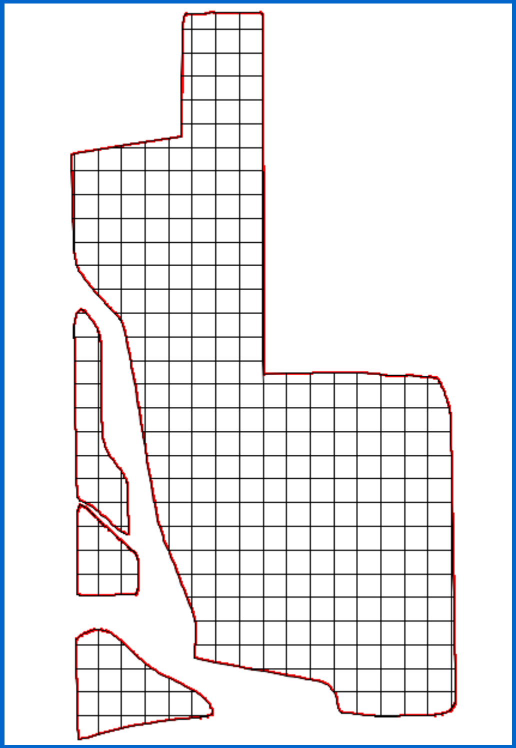
- Sampling frameworks – point sampling



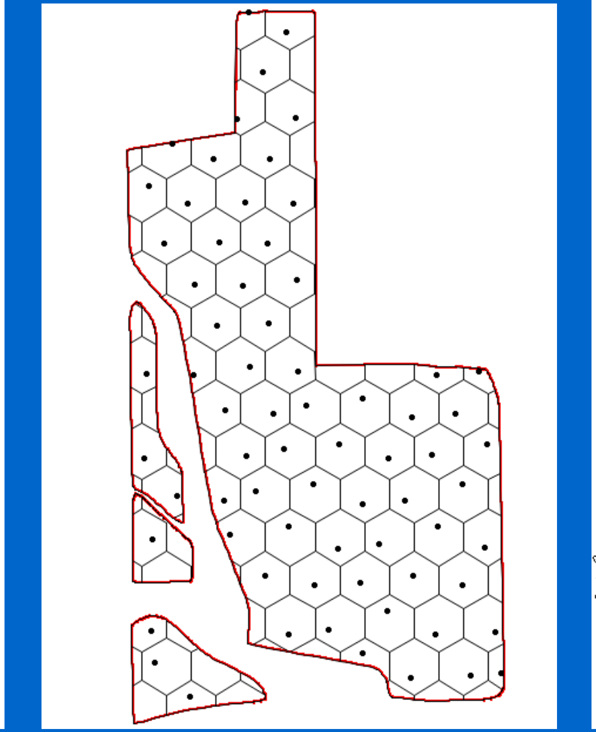
Spatial data exploration

– Sampling frameworks – within zones

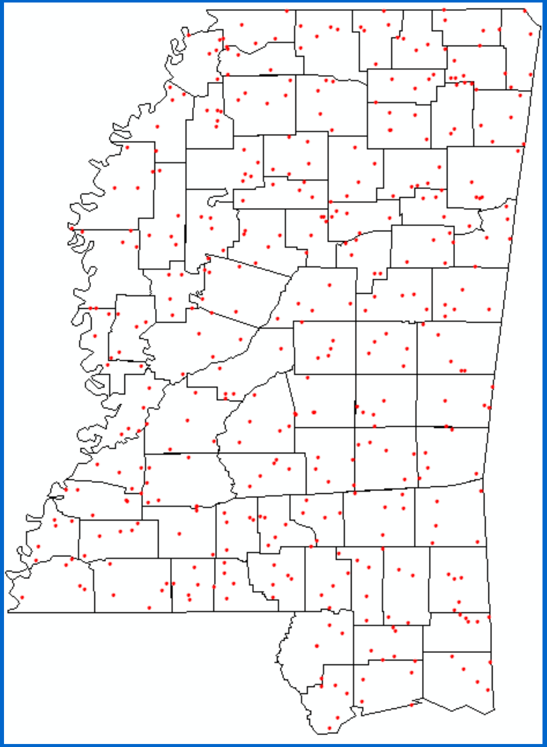
Grid generation - square grid within field boundaries



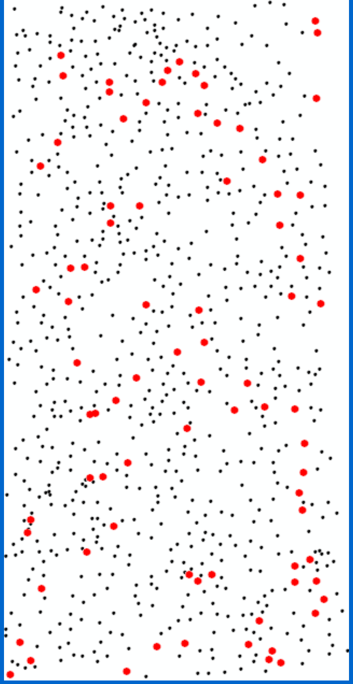
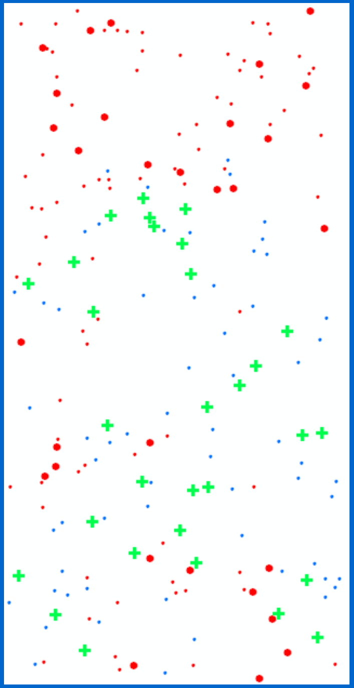
Grid generation (hexagonal) - selection of 1 point per cell, random offset from centre



Selection of 5 random points per zone

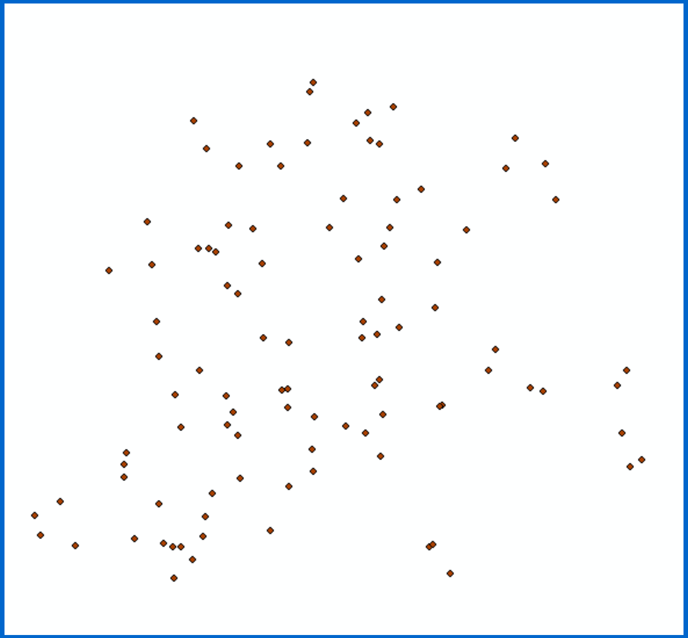


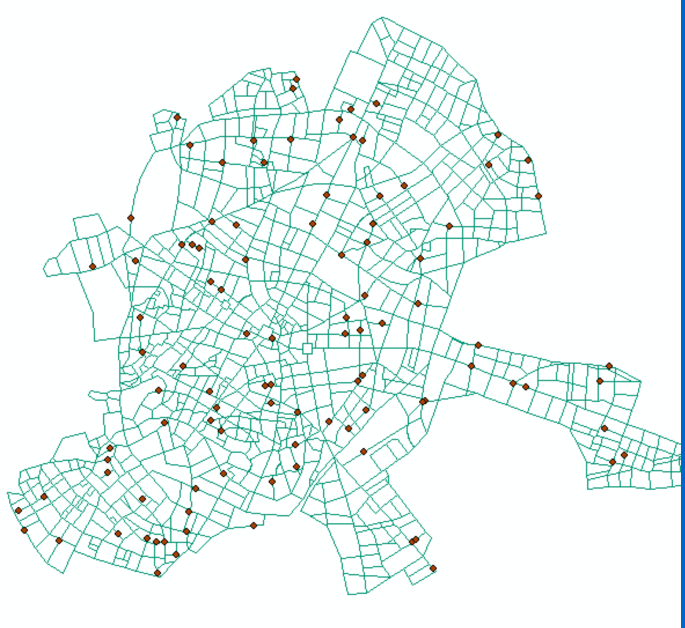
Spatial data exploration

A. 10% random sample from existing point set	B. Stratified random selection, 30% of each stratum
	
800 radio-activity monitoring sites. Random sample of 80 (red/large dots)	200 radio-activity monitoring sites. Random sample of 30 (red/large dots)<100 units of radiation and 30 (crosses)>=100 units of radiation

Spatial data exploration

- Random points on a network





Populations and samples

. A population is a set of well-defined objects.

1. We must be able to say, for every object, if it is in the population or not.

2. We must be able, in principle, to find every individual of the population.

A geographic example of a population is all pixels in a multi-spectral satellite image.

. A sample is some subset of a population.

1. We must be able to say, for every object in the population, if it is in the sample or not.

2. Sampling is the process of selecting a sample from a population.

Continuing the example, a sample from this population could be a set of pixels from known ground truth points.

To check your understanding . . .

- Q1 : Suppose we are studying the distribution of rare element species in an aquifer. Are all the element species in this aquifer reserve population or sample?
- A1 : This is the population; it includes all the objects of interest for the study.

To check your understanding . . .

- Q2 : If we make a transect from one side of the aquifer to the other, and measure all the element species within 10 km of the centre line, is this a population or sample of the element species in the aquifer reserve?
- A2 : This is the sample; it is a defined subset of the population.

What do we mean by “statistics”?

Two common use of the word:

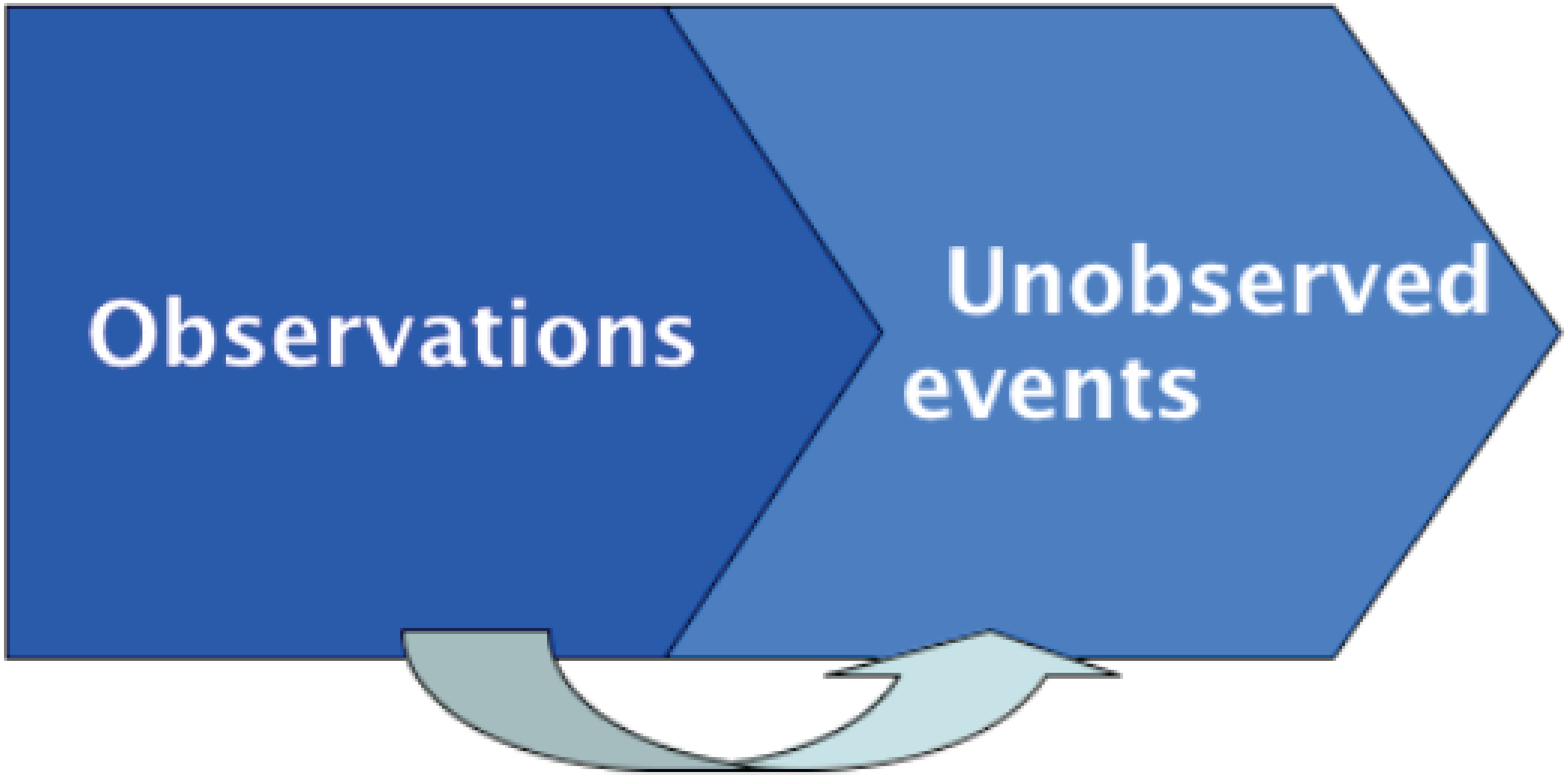
1. Descriptive statistics: numerical summaries of samples; (what was observed)

2. Inferential statistics: from samples to populations. (what could have been or will be observed in a larger population)

Example:

Descriptive “The adjustments of 14 GPS control points for this orthorectification ranged from 3.63 to 8.36 m with an arithmetic mean of 5.145 m”

Inferential “The mean adjustment for any set of GPS points taken under specified conditions and used for orthorectification is no less than 4.3 and no more than 6.1 m; this statement has a 5% probability of being wrong.”



Observations

Unobserved events

Inferential statistics

Q3 : Suppose we do a survey of all the computers in an organization, and we discover that, of the total 120 computers, 80 are running some version of Microsoft Windows operating system, 20 Mac OS X, and 20 Linux. If we now say that 2/3 of the computers in this organization are running Windows, is this a descriptive or inferential statistic?

A3 : This a descriptive statistic. It summarizes the entire population. Note that we counted every computer, so we have complete information. There is no need to infer.

Q4: Suppose we create a sampling frame (list) of all the businesses of a certain size in a city, we visit a random sample of these, and we count the operating systems on their computers. Again we count 80 Windows, 20 Mac OS X, and 20 Linux. If we now say that 2/3 of the computers used for business in this city are running Windows, is this a descriptive or inferential statistic?

A4 : We have summarized a sample (some of the businesses) that is representative of a larger population (all the business). We infer that, if we could do an exhaustive count (as in the previous example), we would find this proportion of each OS.

Why use statistical analysis?

1. Descriptive: we want to summarize some data in a shorter form

2. Inferential: We are trying to understand some process and maybe predict based on this understanding

. So we need to model it, i.e. make a conceptual or mathematical representation, from which we infer the process.

. But how do we know if the model is "correct" ?

* Are we imagining relations where there are none?

* Are there true relations we haven't found?

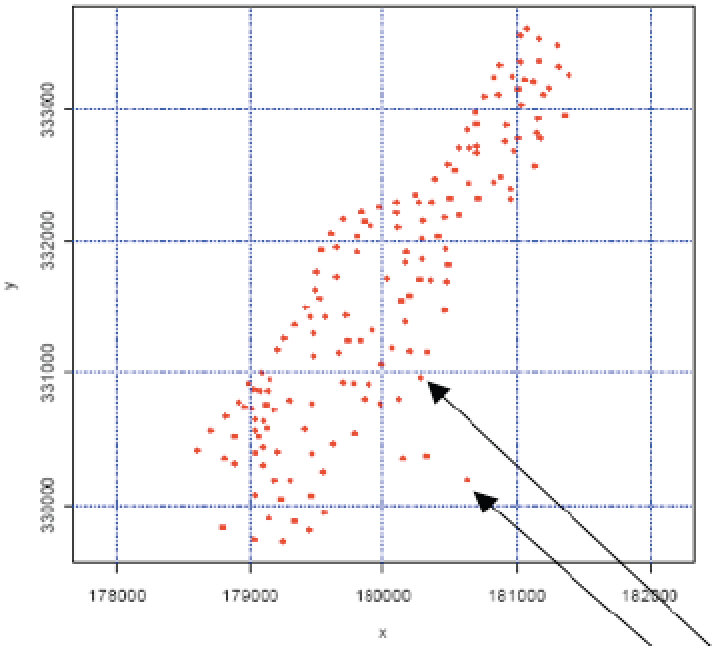
. Statistical analysis gives us a way to quantify the confidence we can have in our inferences.

Commentary

The most common example of geostatistical inference is the prediction of some attribute at an unsampled point, based on some set of sampled points.

In the next slide we show an example from Hail area, the copper (Cu) content of soil samples has been measured at 155 points (left figure);

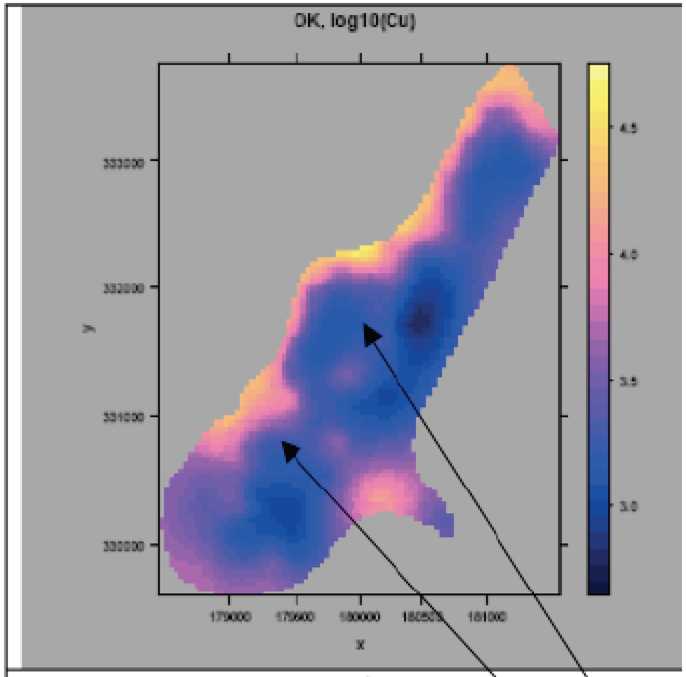
from this we can predict at all points in the area of interest (right figure).



Observations

Sample points

Prediction by inferential statistics



Prediction map

We can predict the values where we did not sample!

What is “geo”-statistics?

Geostatistics is statistics on a population with known location, i.e. coordinates:

1. In one dimension (along a line or curve)

2. In two dimensions (in a map or image)

3. In three dimensions (in a volume)

The most common application of geostatistics is in 2D (maps).

Key point: Every observation (sample point) has both:

1. coordinates (where it is located); and

2. attributes (what it is).

Let’s first look at a data set that is not geo-statistical.

It is a list of soil samples (without their locations) with the lead (Pb) concentration:

Observation ID	Pb
1	77.36
2	77.88
3	30.80
4	56.40
5	66.40
6	72.40
7	60.00
8	144.00
9	52.40
10	41.60
11	46.00
12	56.40

The column Pb is the attribute of interest.

Q5 : Can we determine the median, maximum and minimum of this set of samples?

A5 : Yes; the minimum is 30.8, the maximum 141, and the median 58.2 (half-way between the 6th and 7th sorted values).

We can see this better when the list is sorted:

٣٠,٨ ٤١,٦ ٤٦,٠ ٥٢,٤ ٥٦,٤ ٥٦,٤ ٦٠,٠ ٦٦,٤ ٧٢,٤ ٧٧,٣٦ ٧٧,٨٨ ١٤١,٠

The point is that this is a list of values and we can compute descriptive statistics on it. There is no geographical context.

Q6 : Can we make a map of the sample points with their Pb values?

A6 : No, we can’t make a map, because there are no coordinates'.

Now we look at a data set that is geo-statistical.

These are soil samples taken in Hail, and their lead content; but this time with their coordinates. First let’s look at the tabular form:

Observation ID	Latitude	Longtuide	Pb
1	26.51236	43.56245	77.36
2	26.42135	43.56235	77.88
3	26.65231	43.65842	30.80
4	26.45236	43.52345	56.40
5	26.52356	43.56564	66.40
6	26.52654	43.52655	72.40
7	26.45632	43.52355	60.00
8	26.26543	43.51245	144.00
9	26.62356	43.52243	52.40
10	26.63564	43.57245	41.60
11	26.63251	43.57248	46.00
12	26.55674	43.52244	56.40

The columns E and N are the coordinates, i.e. the spatial reference; the column Pb is the attribute.

Q7 : Comparing this to the non-geostatistical list of soil samples and their lead contents (above), what new information is added here?

A7 : In addition to the sample ID and the Pb content, we also have east (E) and north (N) coordinates for each sample.

On the figure (next slide) you will see:

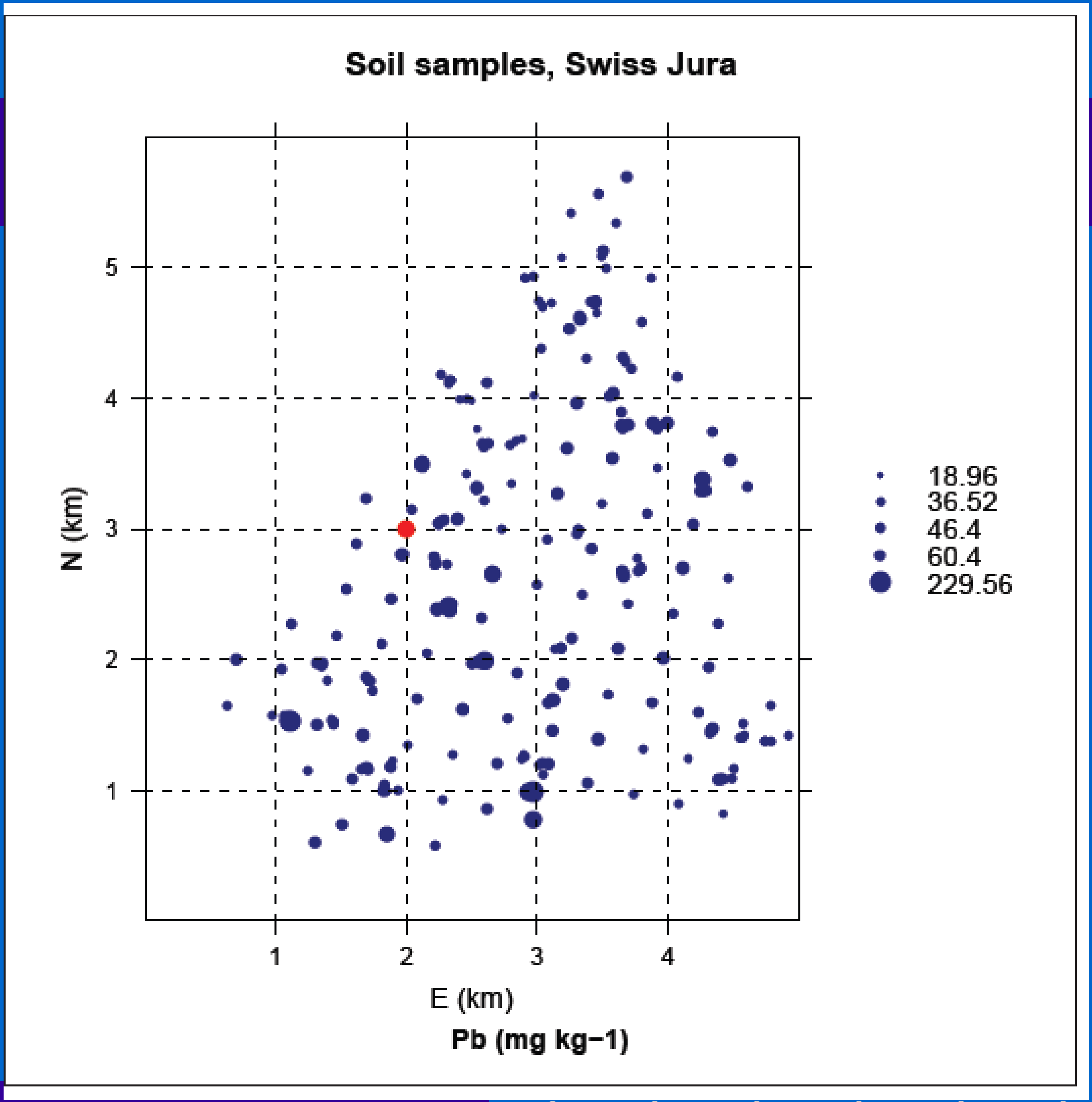
1. A coordinate system (shown by the over-printed grid lines)

2. The locations of 256 sample points – where a soil sample was taken

3. The attribute value at each sample point – symbolized by the relative size of the symbol at each point

– in this case the amount of lead (Pb) in the soil sample

This is called a post-plot (“posting”the value of each sample) or a bubble plot (the size of each“bubble”is proportional to its attribute value).


Soil samples, Swiss Jura
N (km)
5
4
3
2
1
1
2
3
4
E (km)
Pb (mg kg−1)
18.96
36.52
46.4
60.4
229.56

Q8 : In the figure, how can you determine the coordinates of each sample point?

A8 : We can estimate them from the overprinted grid.

Q9 : What are the coordinates of the sample point displayed as a red symbol?

A9 : 2 km E, 3 km N (right on a grid intersection)

Q10 : What is the origin of this coordinate system?

A10 : (0,0) at the lower-left corner of the study area.

Q11 : How could these coordinates be related to some common system such as UTM?

A11 : If we can find out the UTM coordinates of the origin, i.e. the (0,0) of the local system, we can add this to the local coordinates to get the UTM coordinate. Then, if the local coordinate system is already North oriented, we just add this UTM origin to all the local coordinates; if not, we need the UTM coordinate of one other point, and then we can apply a transformation.

Q12 : Suppose we have a satellite image that has not been geo-referenced. Can we speak of geostatistics on the pixel values?

A12 : Yes, because there is a spatial relation between pixels.

Q13 : In this case, what are the coordinates and what are the attributes?

A13 : The row and column in the image is a coordinate; the DN (digital number, reflectance) is the attribute.

Note that the image is not geo-referenced so we can’t do geostatistics in terms of position on the Earth; but we can speak of spatial relations within the image.

Q14 : Suppose now the images has been geo-referenced. What are now the coordinates?

A14 : Whichever coordinate system that was used for geo-referencing.

So, we know that each sample has a location.

What is so special about that?

After all, the attribute information is the same.

What is the value-added of knowing the location? What new possibilities for analysis does this imply?

The added value of geostatistics

1. The location of a sample is an intrinsic part of its definition.

2. All data sets from a given area are implicitly related by their coordinates

So they can be displayed and related in a GIS

3. Values at sample points can not be assumed to be independent: there is often evidence that nearby points tend to have similar values of attributes.

4. That is, there may be a spatial structure to the data

. Classical statistics assumes independence of samples

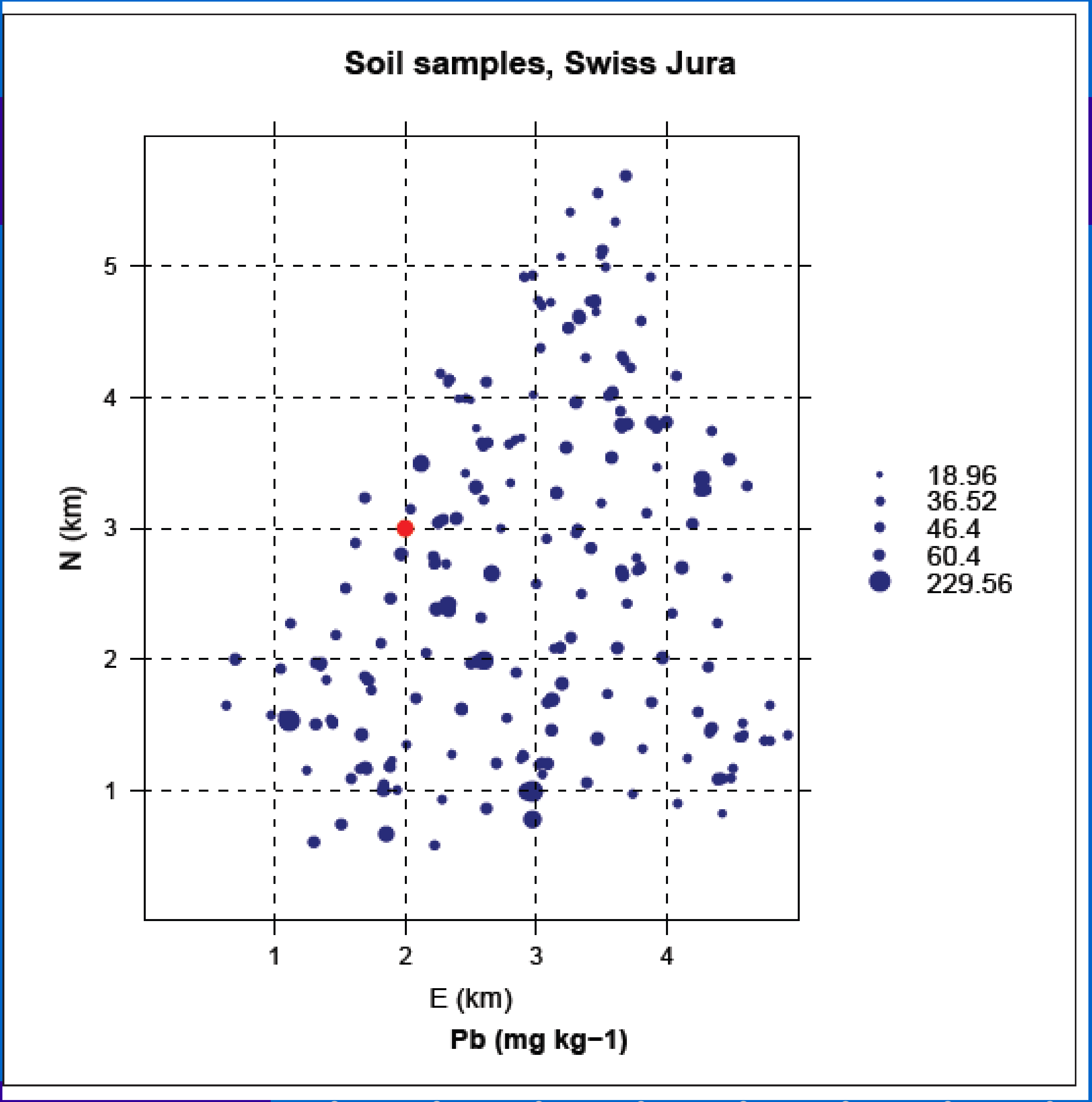
. But, if there is spatial structure, this is not true!

. This has major implications for sampling design and statistical inference

5. Data values may be related to their coordinates → spatial trend

- Let’s look again at the post-plot,
- this time to see if we can discover evidence of spatial dependence –

- that is,
- points that are close to each other have similar attribute values.



Q15 : Do large circles (representing high Pb concentrations) seem to form clusters?

A15 : Yes; for example there seems to be a“hot spot”around (3,1); see the figure on the next page.

Q16 : Do small circles (representing low Pb concentrations) seem to form clusters?

A16 : Yes; for example at the top of the map near (3.5, 5.5); see the figure on the next page.

Q17 : What is the approximate radius the clusters?

A17 : They are not so big, approximately 0.5 km radius.

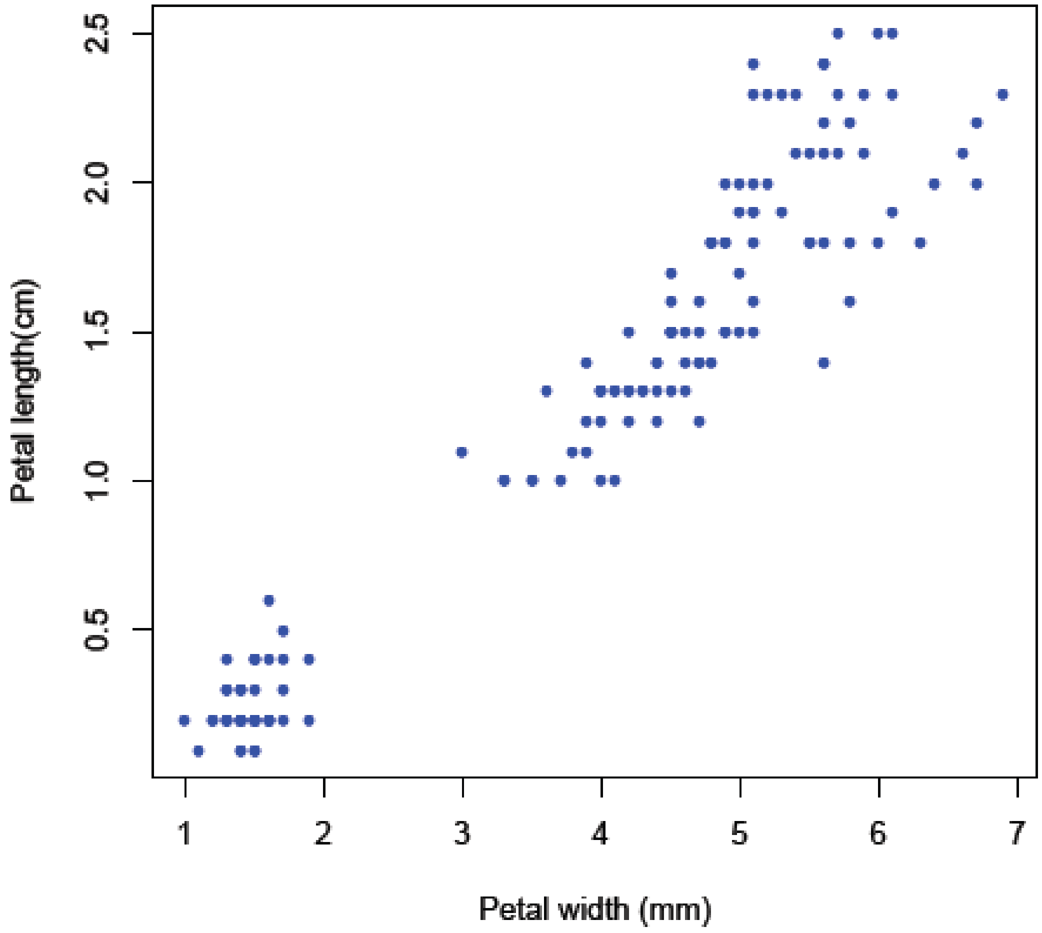
Feature and geographic spaces

- The word space is used in mathematics to refer to any set of variables that form metric axes and which therefore allow us to compute a distance between points in that space.
- If these variables represent geographic coordinates, we have a geographic space.
- If these variables represent attributes, we have a feature space.

Scatter plot of a 2D feature space

This is a visualization of a 2D feature space using a scatter plot.

The points are individual fossil measured



Q18 : What are the two dimensions of this feature space, and their units of measure?

A18 : Dimension 1: Fossil width in mm; Dimension 2: Fossil length in cm.

Q19 : Does there appear to be a correlation between the two dimensions for this set of observations?

A19 : Yes, there appears to be a strong positive correlation.

Geographic space

- **Axes are 1D lines; they almost always have the same units of measure (e.g. metres, kilometres . . .)**

- **One-dimensional: coordinates are on a line with respect to some origin (0): $(x1) = x$**

- **Two-dimensional: coordinates are on a grid with respect to some origin $(0, 0)$: $(x1, x2) = (x, y) = (E, N)$**

- **Three-dimensional: coordinates are grid and elevation from a reference elevation: $(x1, x2, x3) = (x, y, z) = (E, N, H)$**

- **Note: latitude-longitude coordinates do not have equal distances in the two dimensions; they should be transformed to metric (grid) coordinates for geo-statistical analysis.**

Q22 : What are the dimensions and units of measure of a geographic space defined by UTM coordinates?

A22 : There are two dimensions, UTM East and UTM North. These are both measured in meters from the zone origin.

Exploring and visualizing spatial data

Topics for this lecture

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

Visualizing spatial structure

1. Distribution in space

2. Postplots

3. Quantile plots

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 sub-regions?
- 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 coordinate axes, showing only the position of each sample.

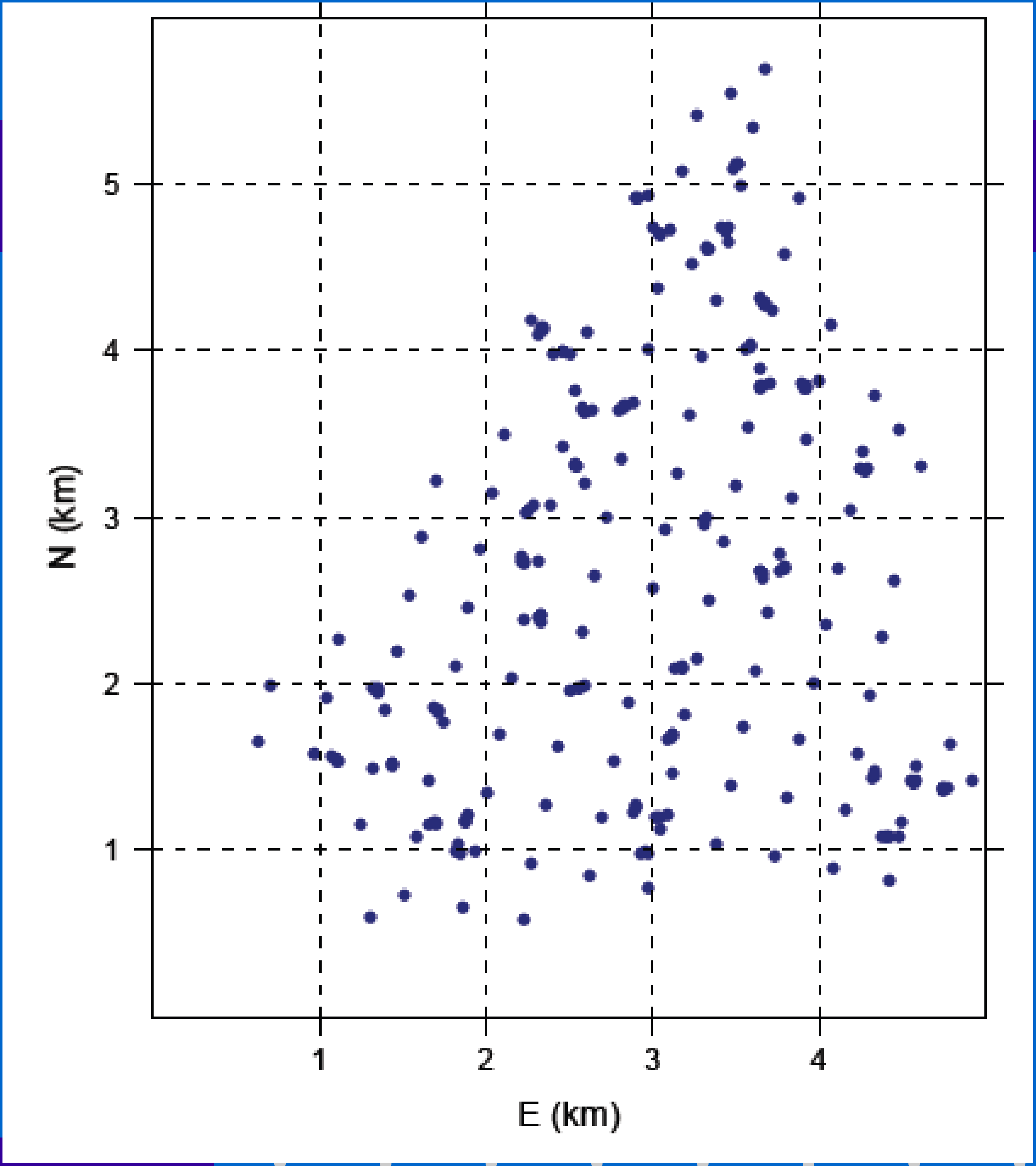
This can be in:

1. 1D : along a line or curve

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

3. 3D: in a volume of space

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



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.

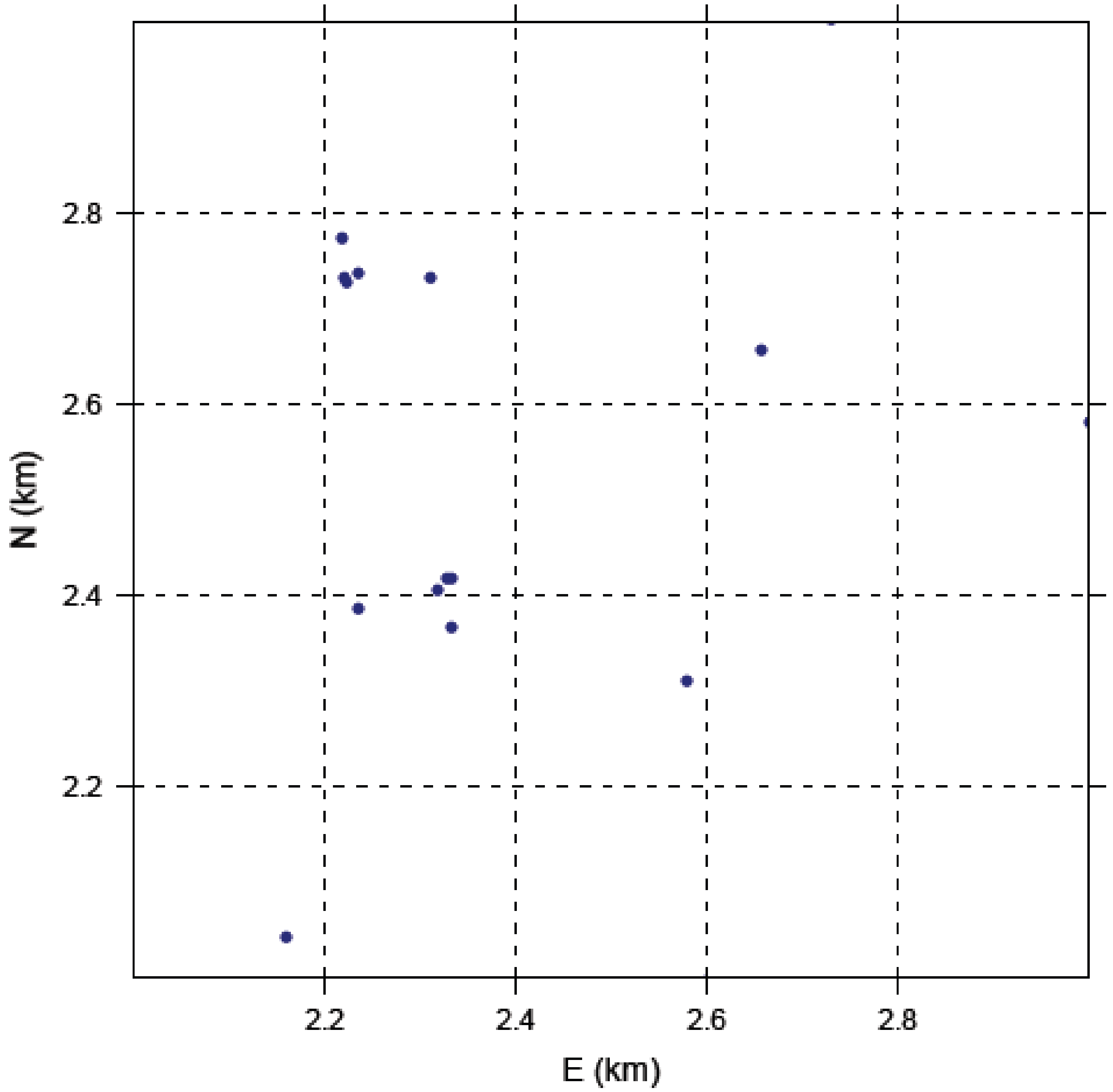
Q2 : Within the convex hull of the points (i.e. the area bounded by the outermost points), is the

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.


2.8
2.6
2.4
2.2
N (km)
2.2
2.4
2.6
2.8
E (km)

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.

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.

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. color, or

3. both

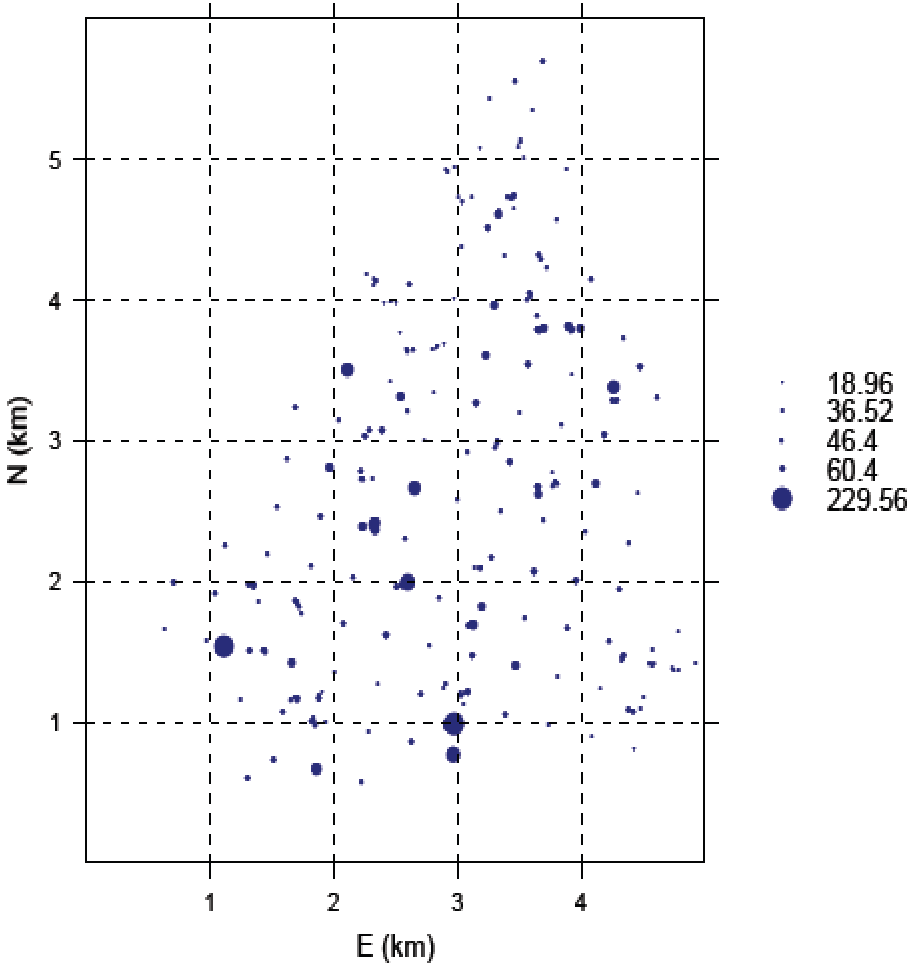
Showing feature-space values with symbol size

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

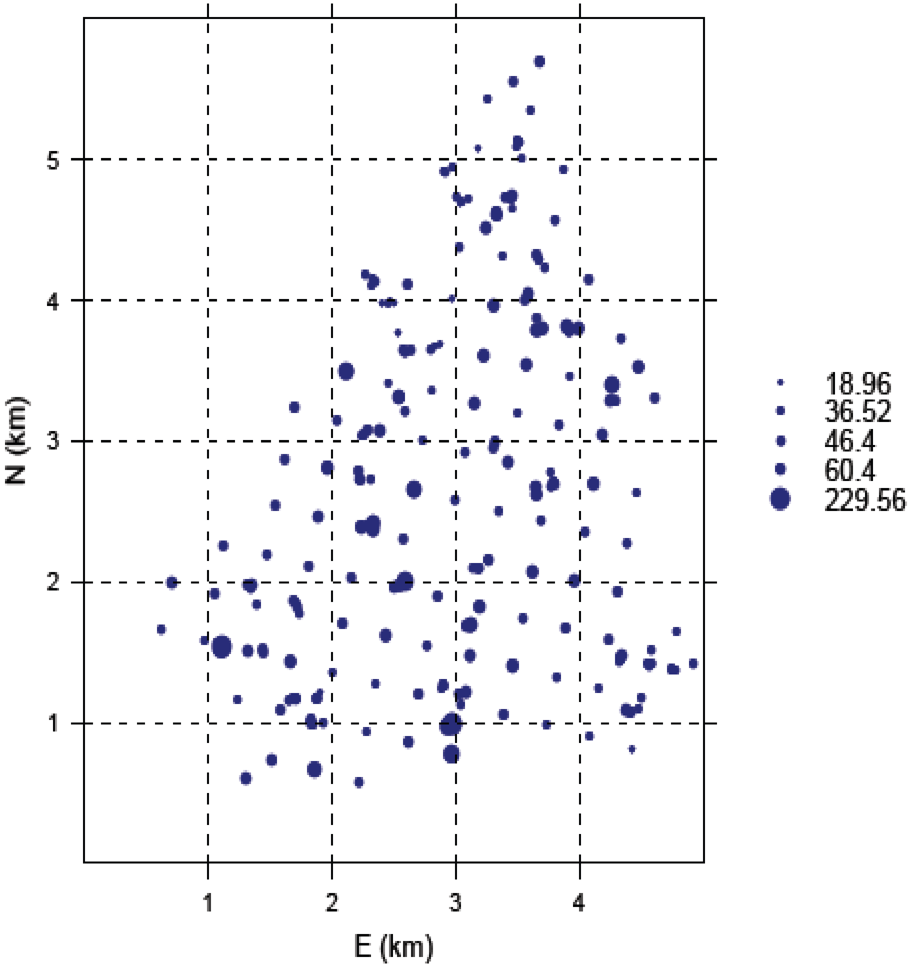
Radius proportional to value

Radius proportional to transformed value

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



Pb (mg kg−1), radius proportional to square root of value



Pb (mg kg−1), radius proportional to value

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.

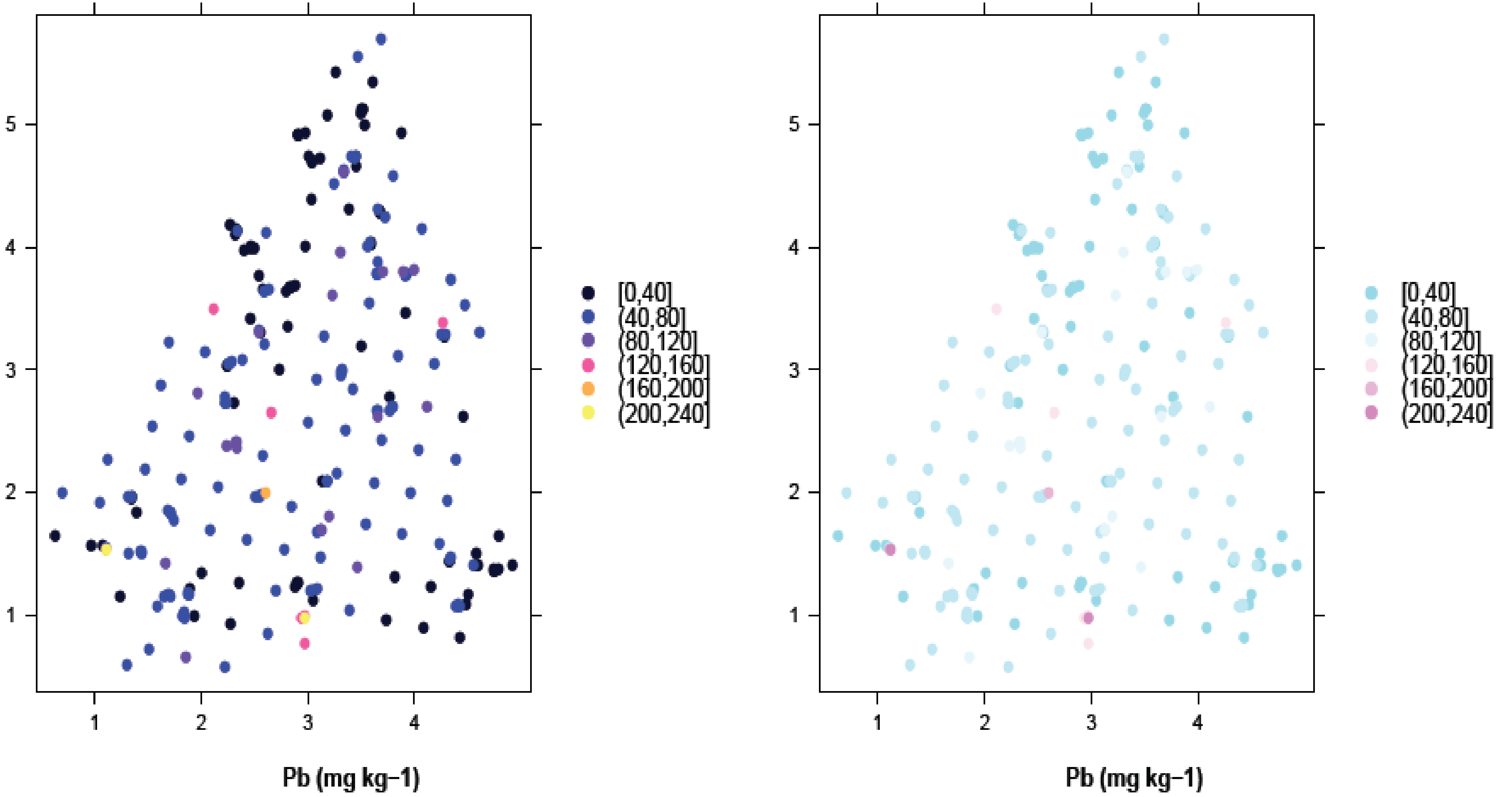
Q6 : Which of this two graphs best shows the distribution of the values in feature space?

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 hot spots stand out.

Showing feature-space values with a color ramp

- Another way to show the difference between feature-space values is with a color ramp, i.e. a sequence of colors from low to high values.
- Different color ramps give very different impressions of the same data, as we now illustrate.

Post-plot of Pb values, Qassim area, Colors



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

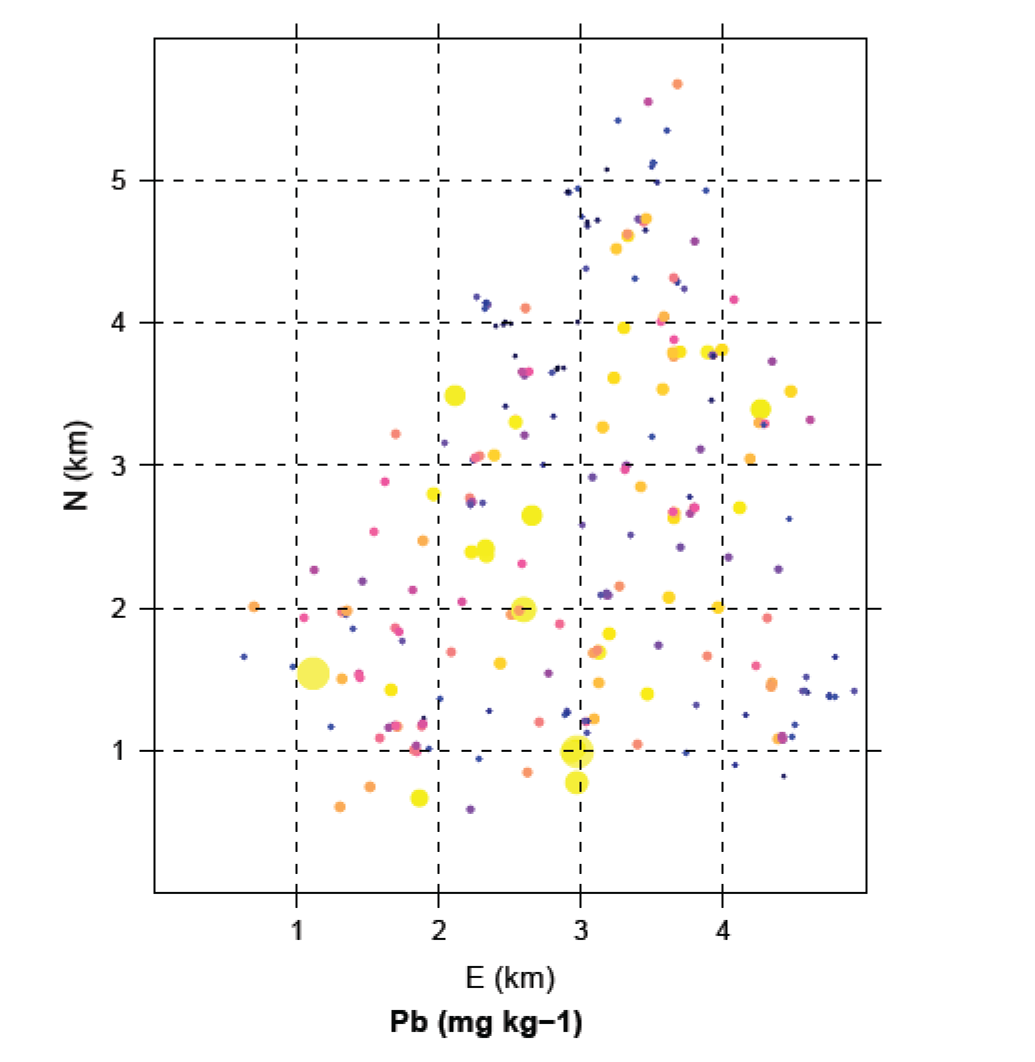
A7 : No correct answer here; it depends on the psychology of the viewer. It shows the highest values in bright pink but the colors are somewhat pastel and difficult to distinguish; the BPY shows the highest values in yellow against a mostly blue background.

Q8 : Do you prefer the size or color to show the feature-space value?

A8 : Personal preference.

-

We can combine both visualization techniques (Size and color in one graph.



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 empirical distribution of the variable.
- The quantiles can be:

quartiles (0-25%, 25-50% . . .);

deciles (0-10%, 10-20% . . .);

- any cutoff point interesting to the analyst, e.g 95% (i.e. highest 5%)

Examples of quantiles of the Pb contents of 259 soil samples:

Here are all the values, sorted ascending:

```
[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
```

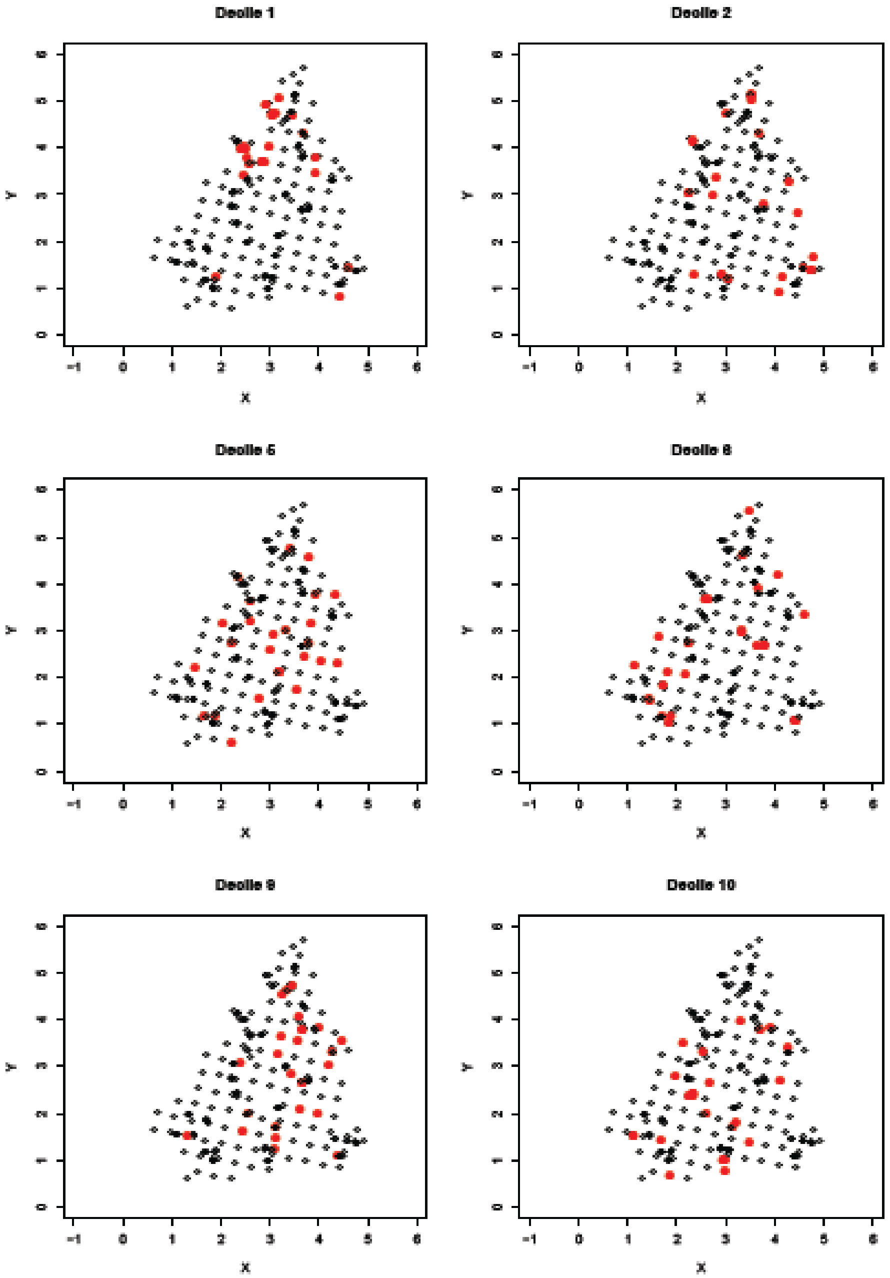
Here are some quantiles:

```
0% 25% 50% 75% 100%
18.96 36.52 46.40 60.40 229.56
```

```
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
```

95% 104.97

- 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

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

A9 : No, each decile is distributed over the entire area.

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

A10 : Yes, points in all deciles form clusters.

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

Orders of trends

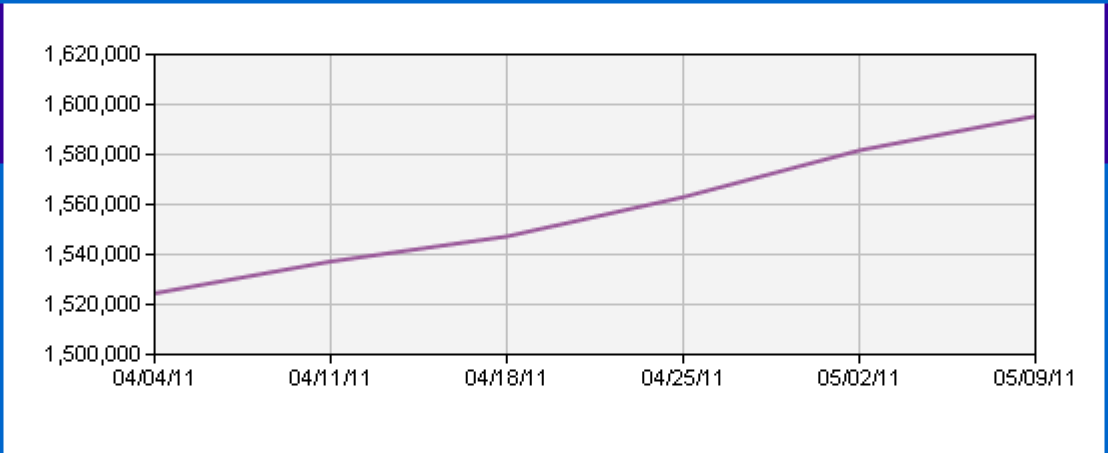
A systematic trend is an approximation to some mathematical surface;

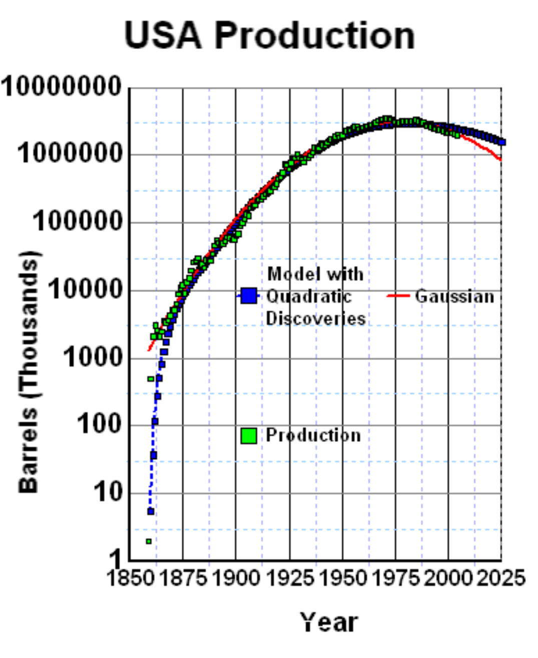
For now, we are concerned with the form of the surface:

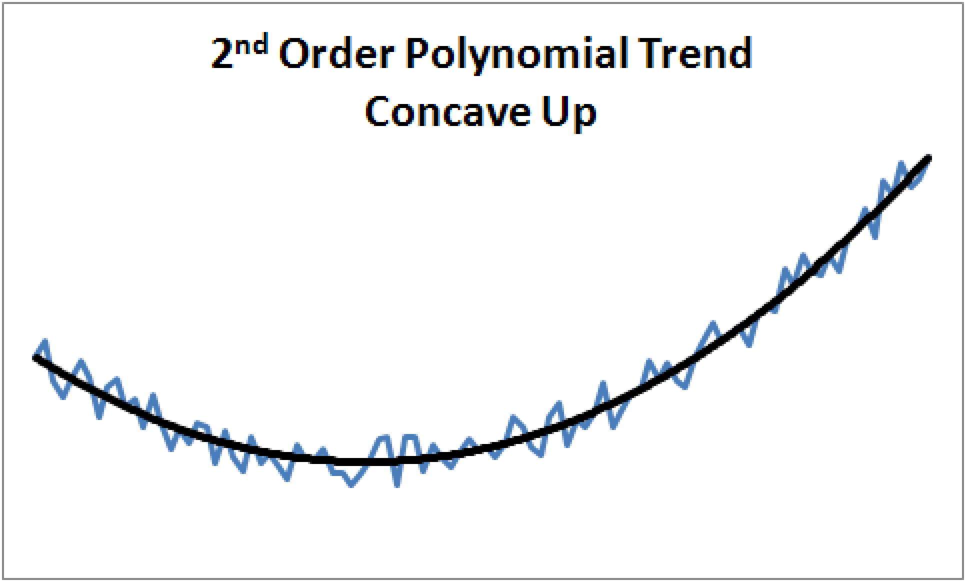
First-order, where the surface is a plane (also called linear): the attribute value changes by the same amount for a given change in distance away from an origin;

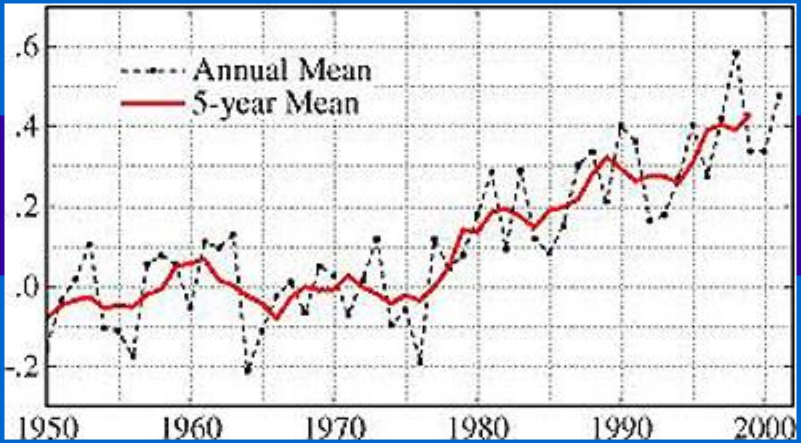
Second-order, where the surface is a paraboloid (2D version of a parabola), i.e. a bowl (lowest in the middle) or dome (highest in the middle)

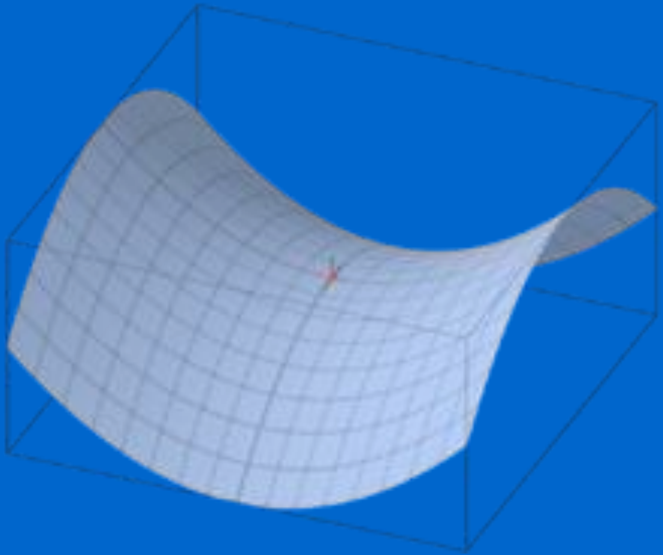
Higher-order, where the surface has saddle points or folds

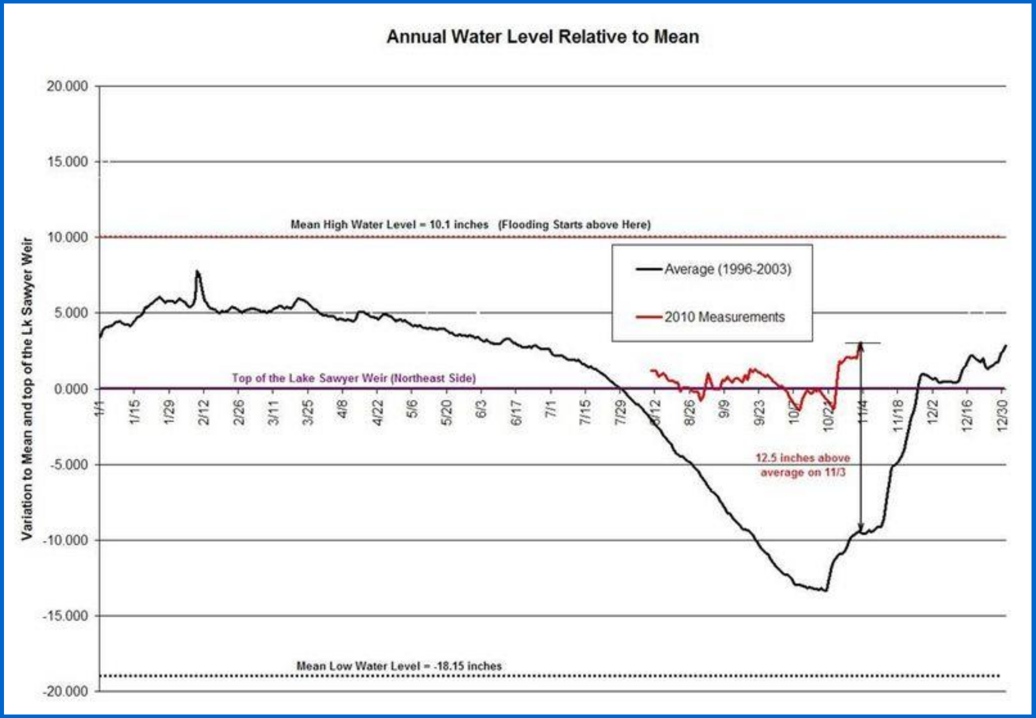

1,620,000
1,600,000
1,580,000
1,560,000
1,540,000
1,520,000
1,500,000
04/04/11
04/11/11
04/18/11
04/25/11
05/02/11
05/09/11


USA Production
10000000
1000000
100000
10000
1000
100
10
1
Barrels (Thousands)
1850
1875
1900
1925
1950
1975
2000
2025
Year
🟦 Model with Quadratic Discoveries
Gaussian
🟩 Production


2nd Order Polynomial Trend
Concave Up


Annual Mean
5-year Mean
.6
.4
.2
.0
-.2
1950
1960
1970
1980
1990
2000




Annual Water Level Relative to Mean
Variation to Mean and top of the Lk Sawyer Weir
20.000
15.000
10.000
5.000
0.000
-5.000
-10.000
-15.000
-20.000
Mean High Water Level = 10.1 inches (Flooding Starts above Here)
Average (1996-2003)
2010 Measurements
Top of the Lake Sawyer Weir (Northeast Side)
12.5 inches above average on 11/3
Mean Low Water Level = -18.15 inches
1/1
1/15
1/29
2/12
2/26
3/11
3/25
4/8
4/22
5/6
5/20
6/3
6/17
7/1
7/15
7/29
8/12
8/26
9/9
9/23
10/7
10/21
11/4
11/18
12/2
12/16
12/30

Q11 : 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?

A11 :

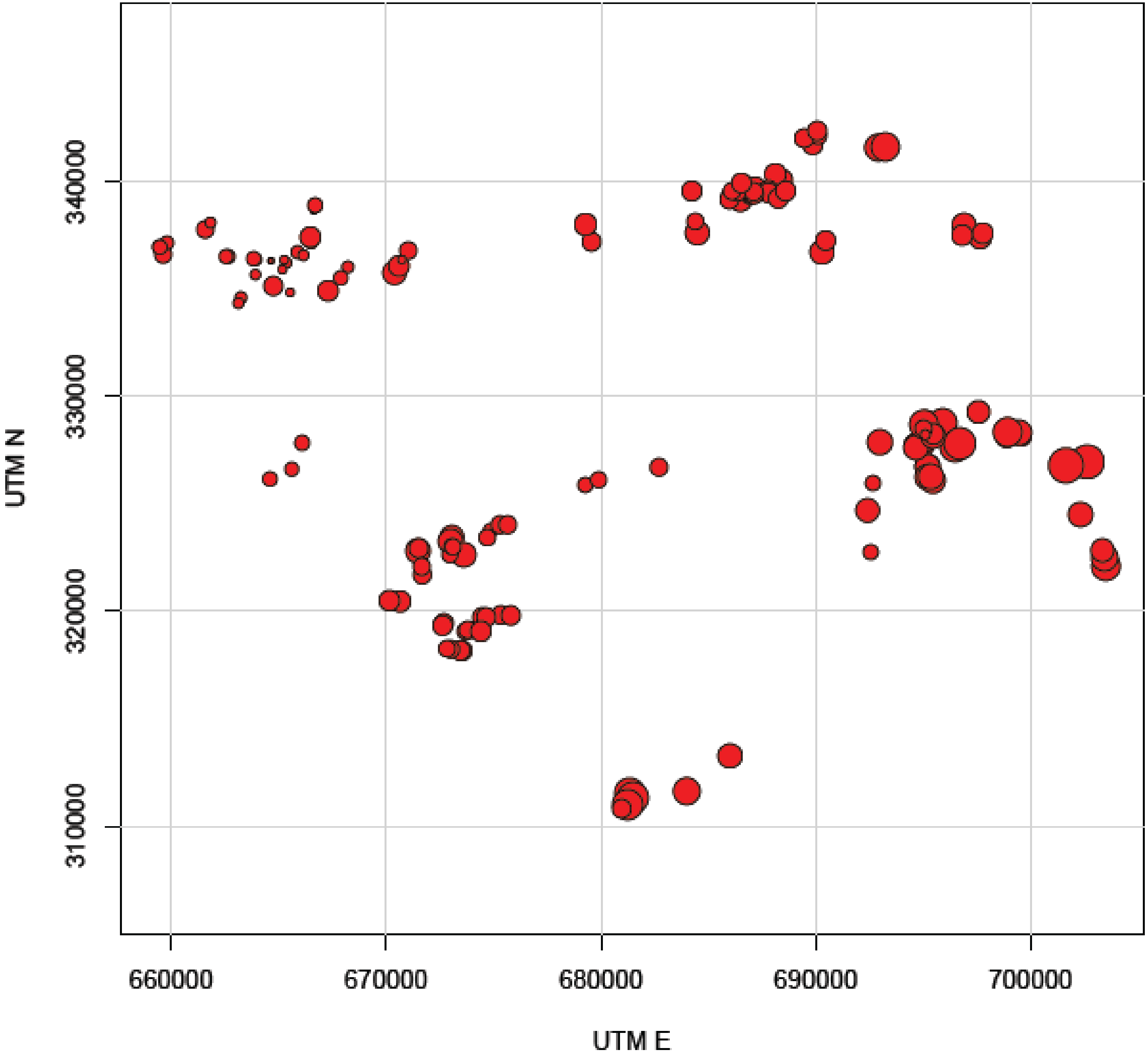
1. First-order (linear, planar), increasing from the coast to the mountain (orographic effect on on-shore winds);

2. Second-order; a dome (shallowest in middle);

3. Higher-order (many parallel ‘troughs’ and ‘ridges’)

- 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 color of the symbols (related to the attribute values) systematically change over the whole plot.


340000
330000
320000
310000
UTM N
660000
670000
680000
690000
700000
UTM E

Q12 : Describe the regional pattern of the clay content shown in the previous graph.

A12 : It is lowest in the NW corner and increases towards the E and S.

Q13 : How much of the variability in clay content explained by the trend?

A13 : 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.

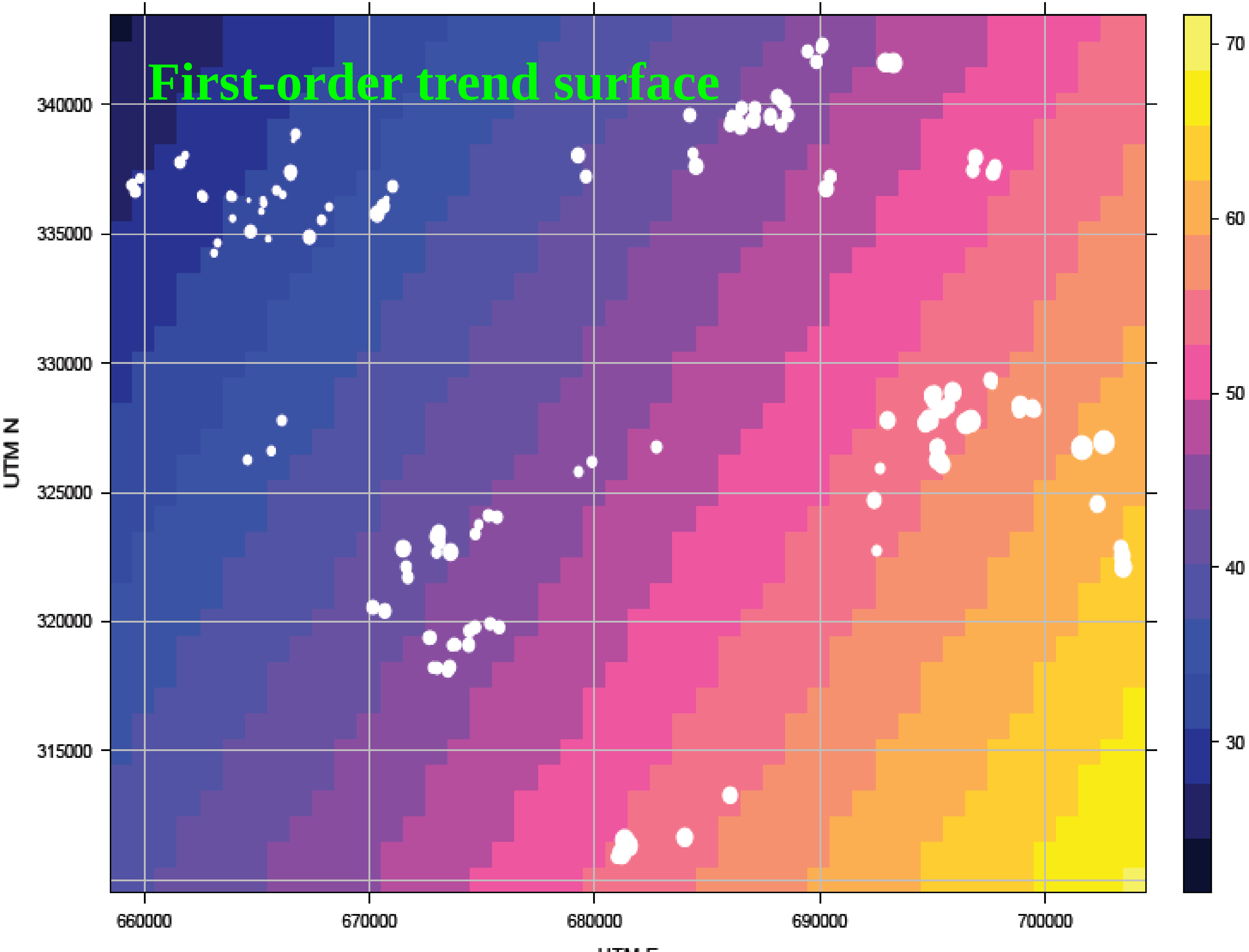
Visualizing a trend with a trend surface

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

z = f (x) where x is a coordinate vector, e.g. in 2D it might be:

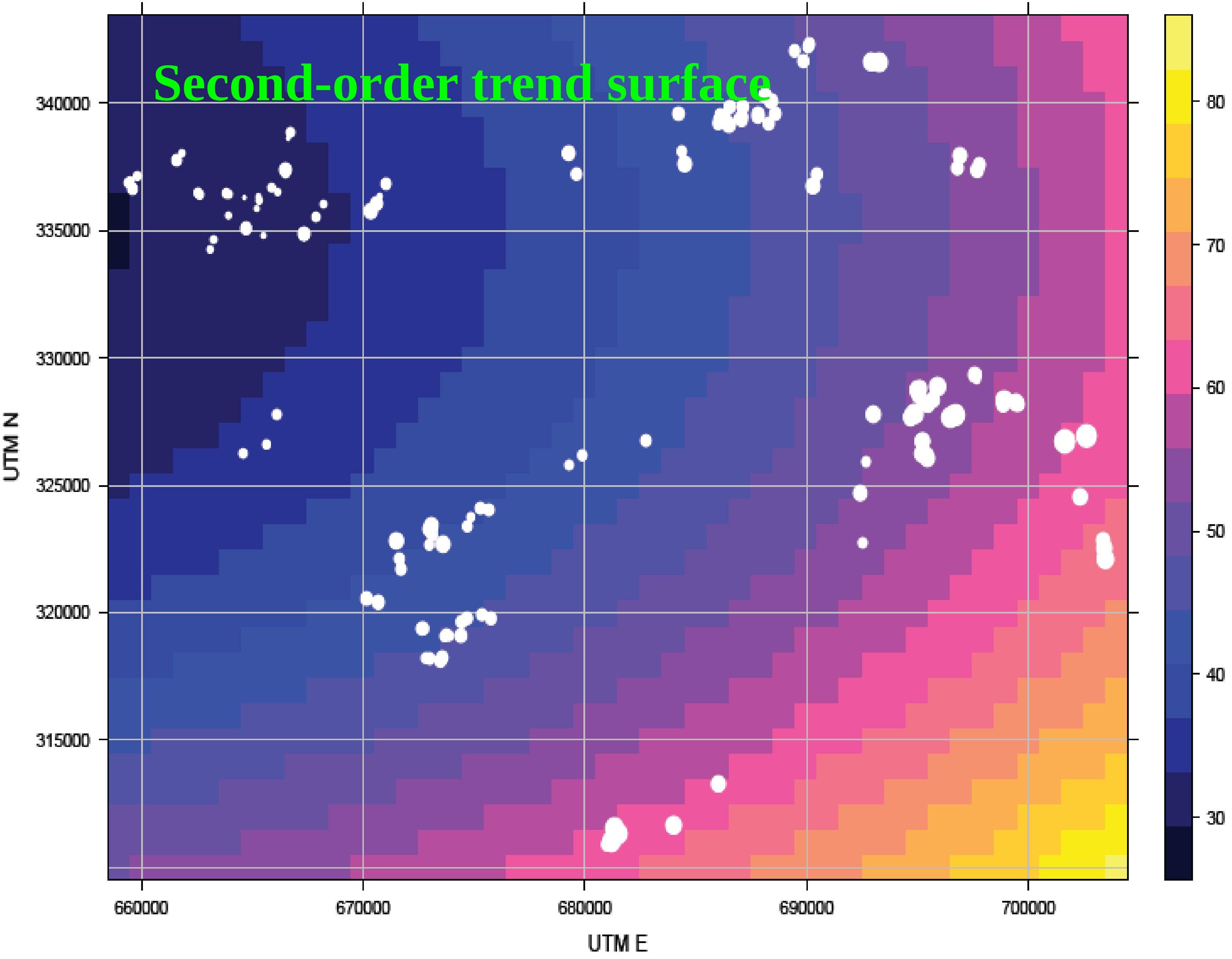
(x,y) = (UTM E,UTM N)

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.



Sample points overprinted as post−plot

Second-order trend surface



Sample points overprinted as post-plot

Q14 : Which of these two trend surfaces best fits the sample points? (Compare the overprinted post-plot with the surface).

A14 : 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.

Q15 : Is the second-order surface a bowl or dome? •

A15 : It is a (elliptically-shaped) bowl; the lowest values are found in the NW corner and increase in all directions from there.

The least squares regression line

- The least squares regression line is the line which produces the smallest value of the sum of the squares of the residuals.

- A residual is the vertical distance from a point on a scatter diagram to the line of best fit. Therefore the least squares regression line can be seen as the best line of best fit.

- The equation of the least squares regression line of y on x is:
- *Y-Y/= b(X-X/)*
-

Where b is: $$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{\frac{\sum x_i y_i}{n} - \overline{x}\,\overline{y}}{\frac{\sum x_i^2}{n} - \overline{x}^2}$$

-

As you can see there are 3 different ways to calculate the value for b, based on what information you are given in the question

- Example

- Calculate the least squared regression line of y on x from the following data:
- X: 20 30 40 50 60 70
- Y : 2.49 2.41 2.38 2.14 1.97 2.03

- Firstly draw up a table of the values you need and fill it out. In this example we'll use the third equation for b so we need all the values of x^2 and x_iy_i:

X	Y	x^2	x_iy_i
20	2.49	400	49.9
30	2.41	900	72.3
40	2.38	1600	95.2
50	2.14	2500	107
60	1.97	3600	118.2
70	2.03	4900	142.1
270	13.42	13900	584.7

- Next step is to calculate the means of the x and y values:

$$\overline{x} = \frac{270}{6} = 45$$

$$\overline{y} = \frac{13.42}{6} = 2.24$$

$$b = \frac{\frac{584.7}{6} - 45 \text{ x } 2.24}{\frac{13900}{6} - 45^2} = -0.01$$

- **Hence the equation is therefore:**

$$y-2.24=-0.01(x-45)$$

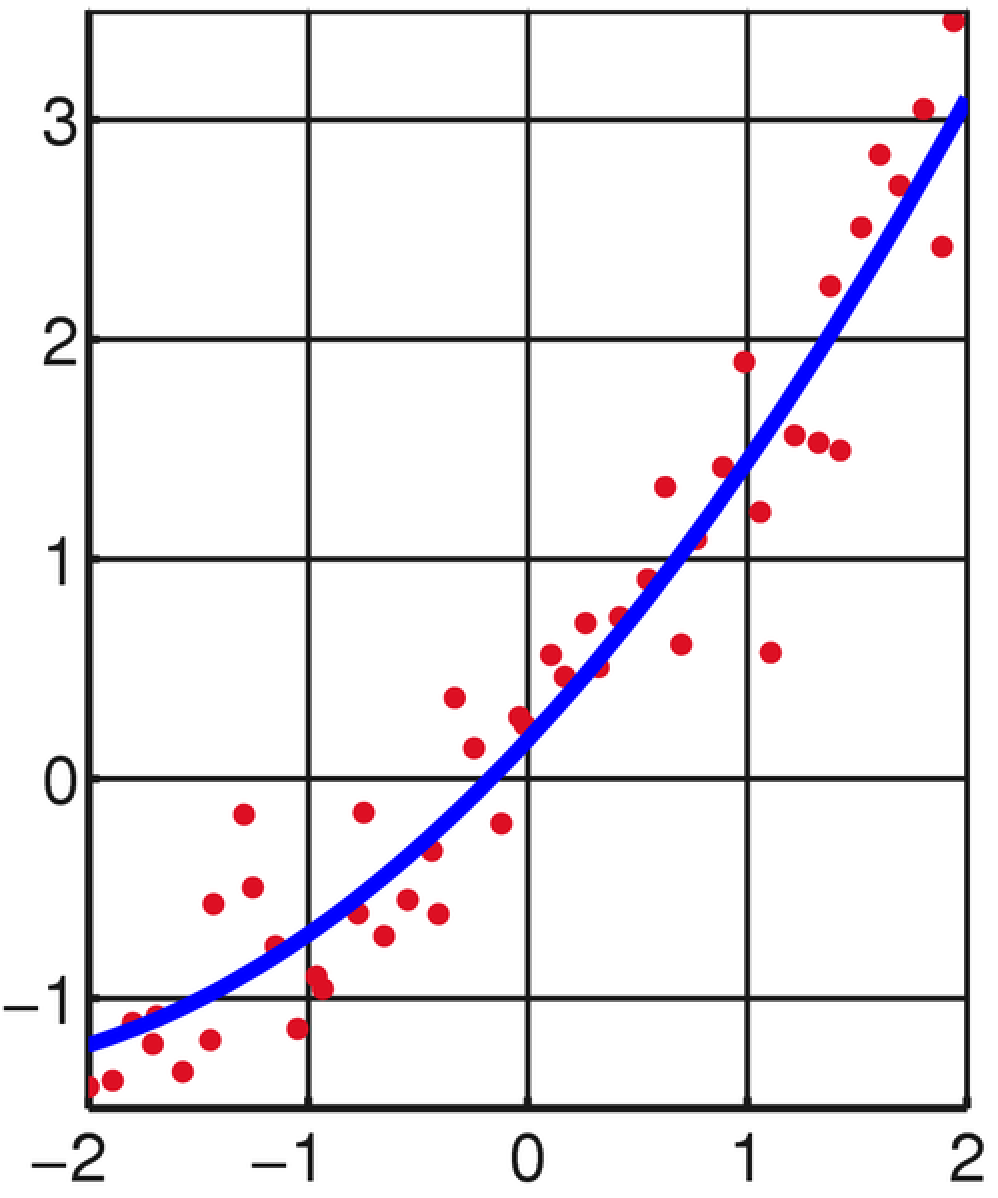
$$y=-0.01x+2.69$$

Computing a trend surface model by ordinary least squares regression

We use the Qassim soils data of; each observation has two UTM coordinates and two attributes:

	UTM_E	UTM_N	clay	pH
1	702638	326959	78	4.80
2	701659	326772	80	4.40
3	703488	322133	66	4.20
4	703421	322508	61	4.54
5	703358	322846	53	4.40
6	702334	324551	57	4.56

We use the linear models method; all statistical packages have an equivalent comment to compute a least-squares fit

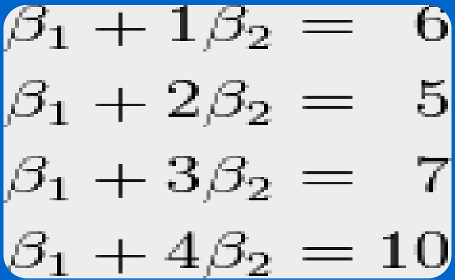

3
2
1
0
–1
–2
–1
0
1
2

As a result of an experiment, four (*x*,*y*) data points were obtained, (1,6), (2,5), (3,7), and (4,10)

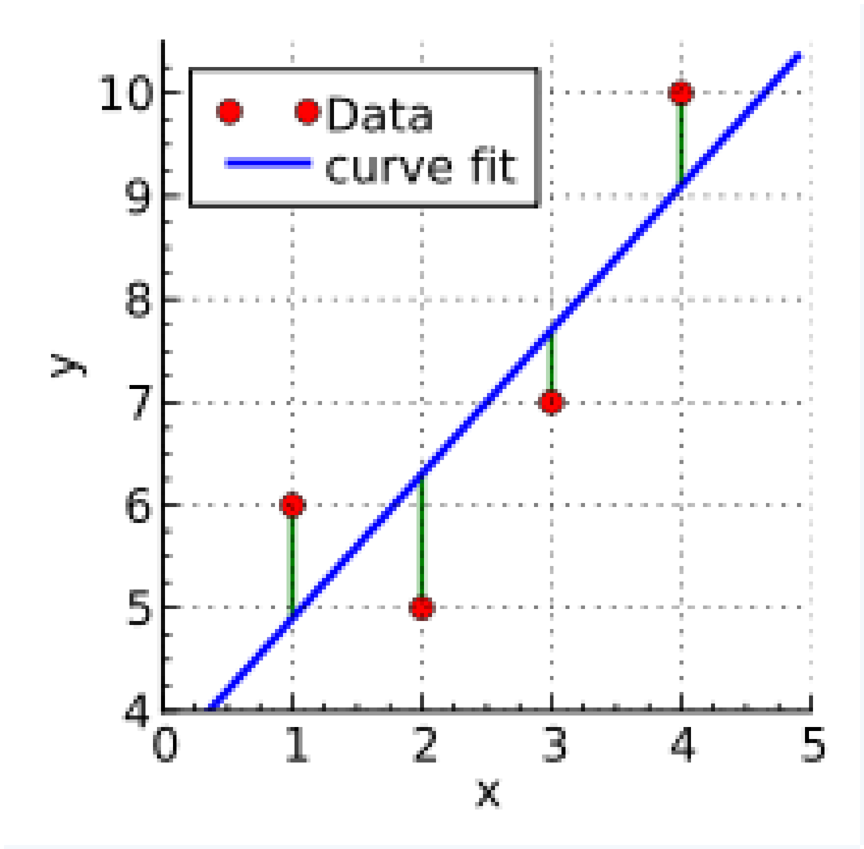
It is desired to find a line

$y = \beta_1 + \beta_2 x$

that fits "best" these four points.



$$\begin{aligned}\beta_1 + 1\beta_2 &= 6\\ \beta_1 + 2\beta_2 &= 5\\ \beta_1 + 3\beta_2 &= 7\\ \beta_1 + 4\beta_2 &= 10\end{aligned}$$



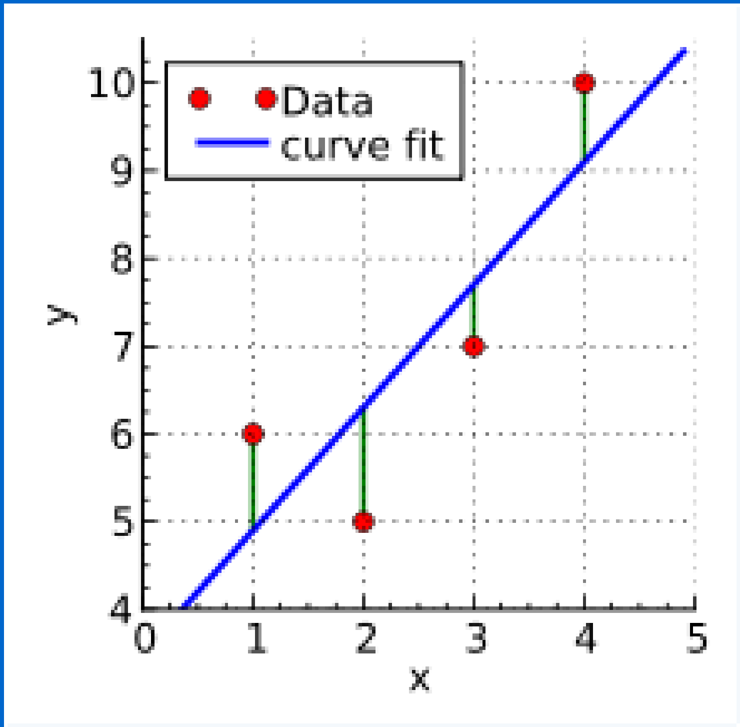
The least squares approach to solving this problem is to try to make as small as possible the sum of squares of "errors" between the right- and left-hand sides of these equations, that is, to find the minimum of the function

$$S(\beta_1, \beta_2) = [6 - (\beta_1 + 1\beta_2)]^2 + [5 - (\beta_1 + 2\beta_2)]^2 + [7 - (\beta_1 + 3\beta_2)]^2 + [10 - (\beta_1 + 4\beta_2)]^2 .$$

- The minimum is determined by calculating the partial derivatives of *S*(β_1,β_2) with respect to β_1 and β_2 and setting them to zero.
- This results in a system of two equations in two unknowns, called the normal equations, which give, when solved
- $\beta_1 = 3.5$ $\beta_2 = 1.4$
- and the equation $y = 3.5 + 1.4x$ of the line of best fit.

The residuals, that is, the discrepancies between the *y* values from the experiment and the *y* values calculated using the line of best fit are then found to be :

1.1, − 1.3, − 0.7, and 0.9



The minimum value of the sum of squares is :

$S(3.5,1.4) = 1.1^2 + (-1.3)^2 + (-0.7)^2 + 0.9^2 = 4.2.$

- http://www.youtube.com/watch?v=jEEJNz0RK4Q

Visualizing spatial dependence

1. Point-pairs

2. h-scatterplots

3. Variogram clouds

4. Experimental variograms

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.

Point-pairs

Q16 : The Hail data set has 155 sample points. How many point-pairs can be formed from these?

A16 : (155*154)/2 = 11, 935

Q17 : Are you surprised by the size of this number?

A17 : Most people are surprised by the large number of pairs. Human intuition is not good at estimating combinations.

Point-pairs

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.

Point-pairs

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 (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.

h-scatterplots

Interpreting the h-scatterplot

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

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

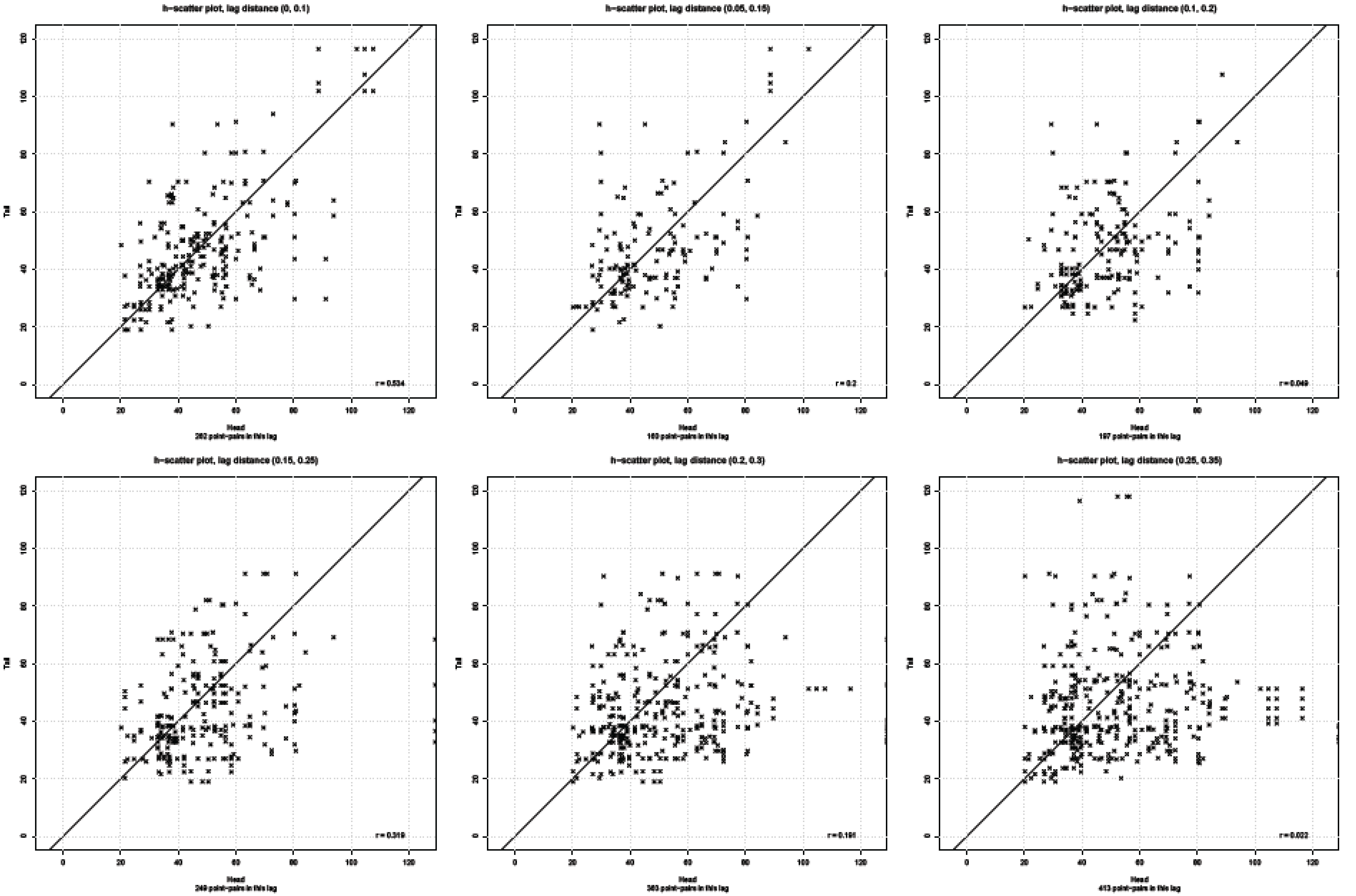
h-scatterplots

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

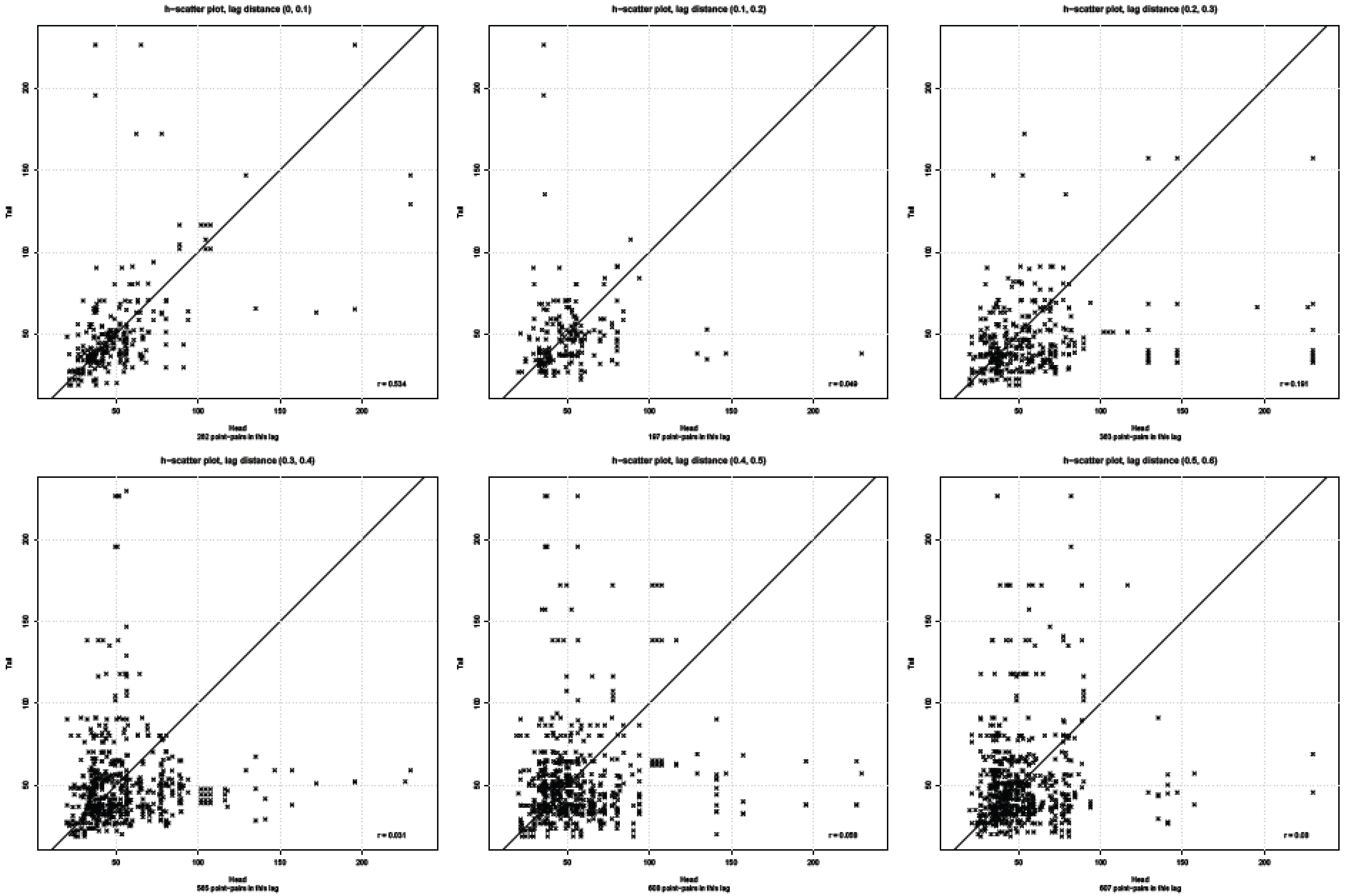
- Bins of 50 m width up to 300 m, i.e. (0, 50), (50, 100), . . . , (250, 300)
- Bins of 100 m width up to 600 m, i.e. (0, 100), (100, 200), . . . , (500, 600)

The 50 m bins only show the lower part of the Pb attribute value range.

Bins of 50 m width up to 300 m



Bins of 100 m width up to 600 m



h-scatterplots

- Q18 : Describe the “evolution” of the cloud of point pairs as the separation distance increases.
- A18 : It becomes more diffuse, i.e. further from the 1:1 line.

Correlation Coefficient

- The correlation coefficient computed from the sample data measures the strength and direction of a relationship between two variables.
- The range of the correlation coefficient is.

  - 1 to + 1 and is identified by *r.*

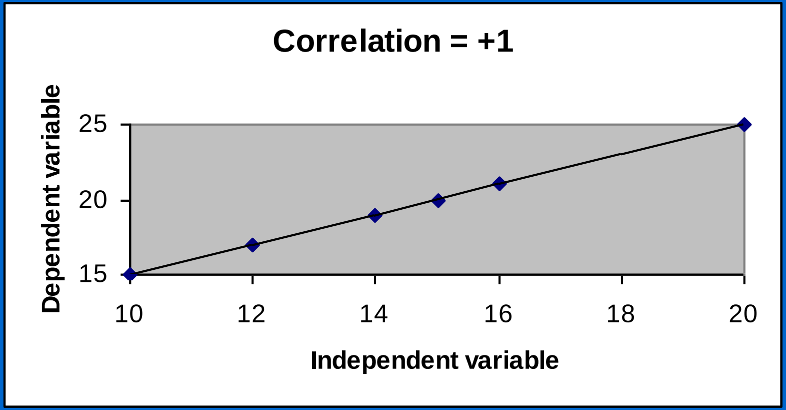
Positive and Negative Correlations

- A positive relationship exists when both variables increase or decrease at the same time. (Weight and height).

- A negative relationship exist when one variable increases and the other variable decreases or vice versa. (Strength and age).

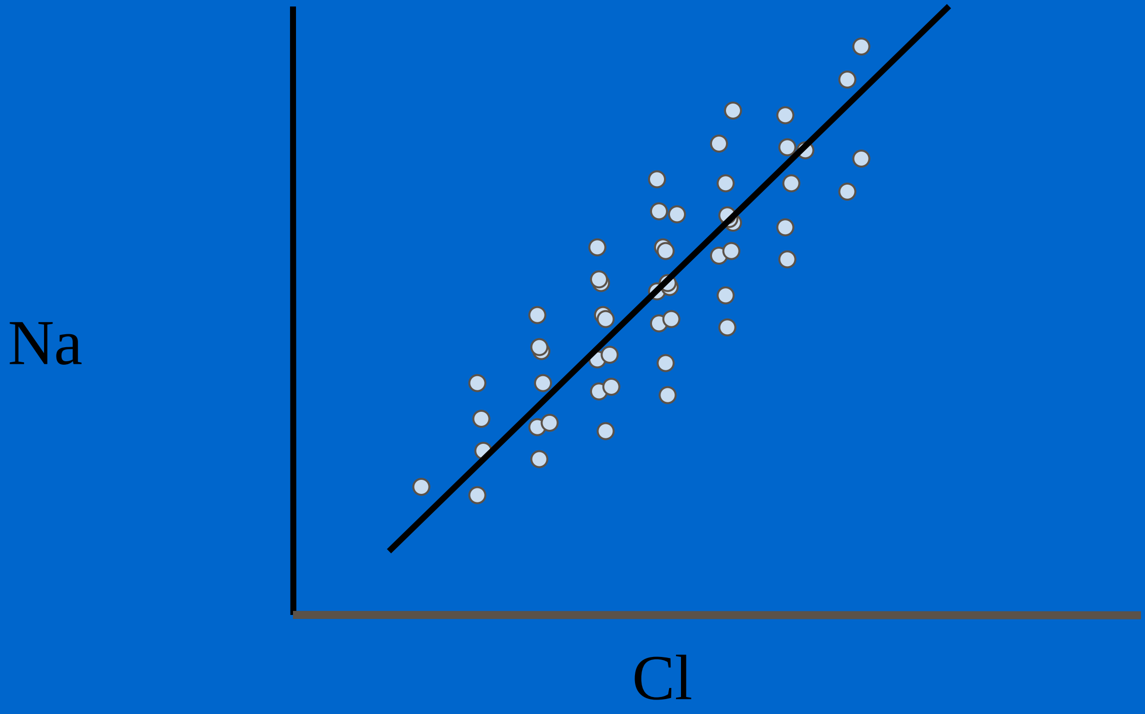
Range of correlation coefficient

- In case of exact positive linear relationship the value of r is +1.
- In case of a strong positive linear relationship, the value of *r* will be close to + 1.



High Degree of positive correlation

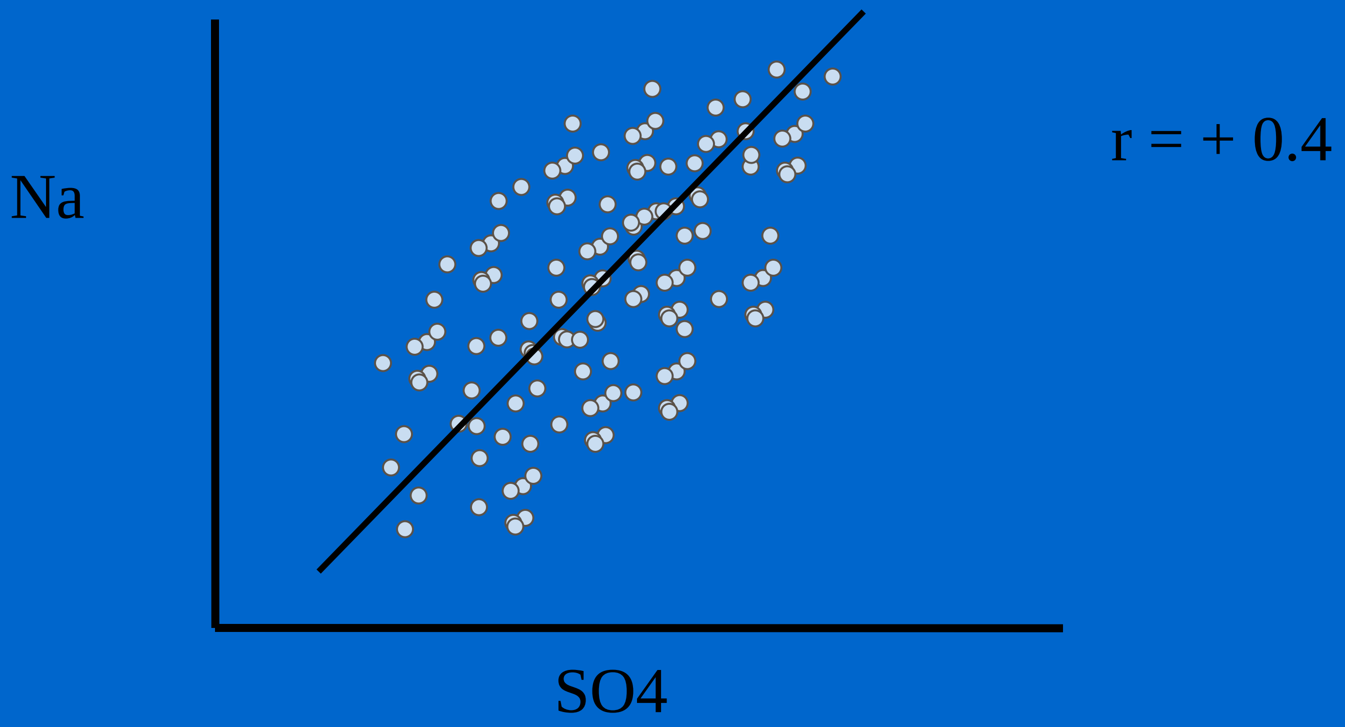
- Positive relationship



r = +.80

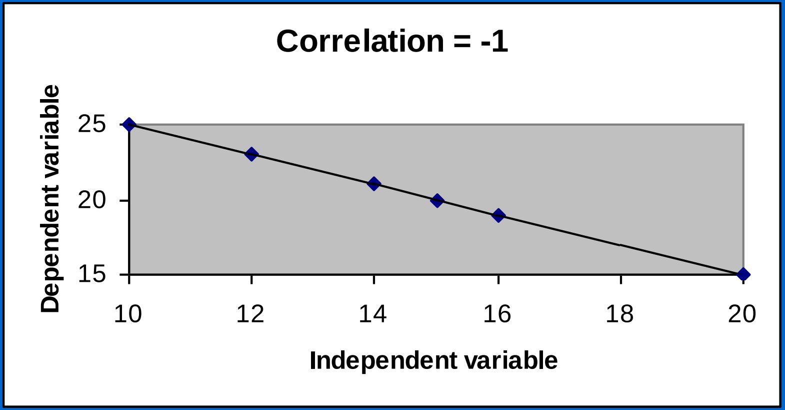
Degree of correlation

- **Moderate Positive Correlation**



Range of correlation coefficient

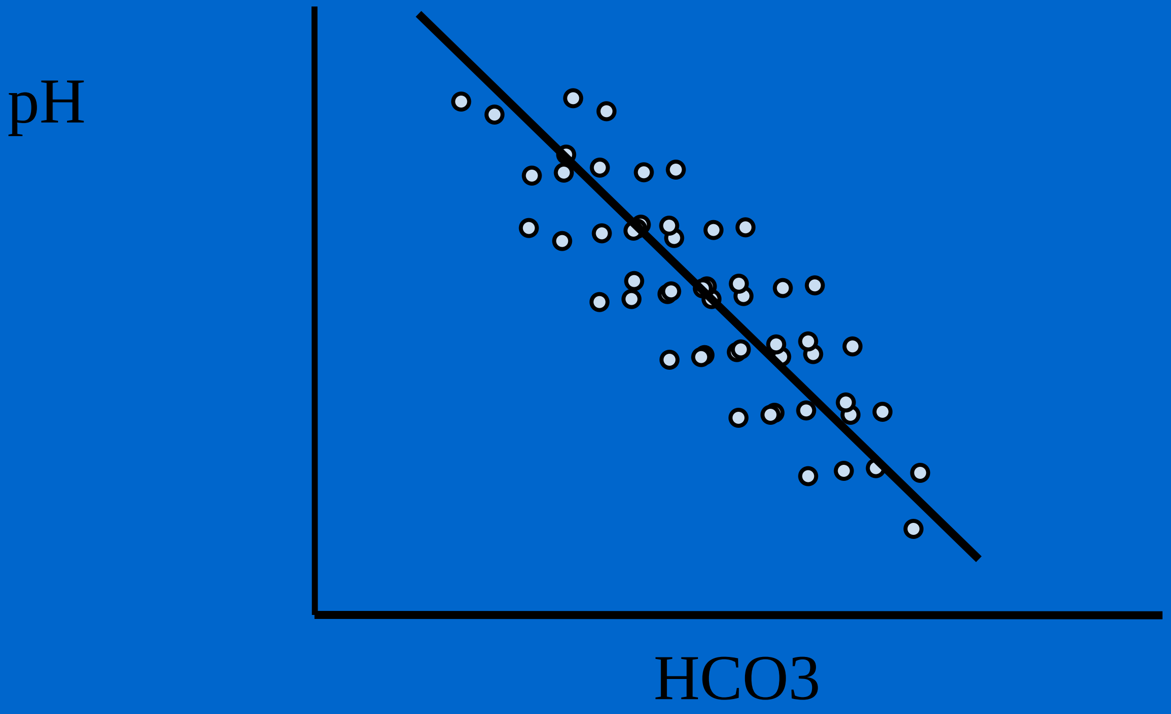
- In case of exact negative linear relationship the value of *r* is –1.
- In case of a strong negative linear relationship, the value of *r* will be close to – 1.



Degree of correlation

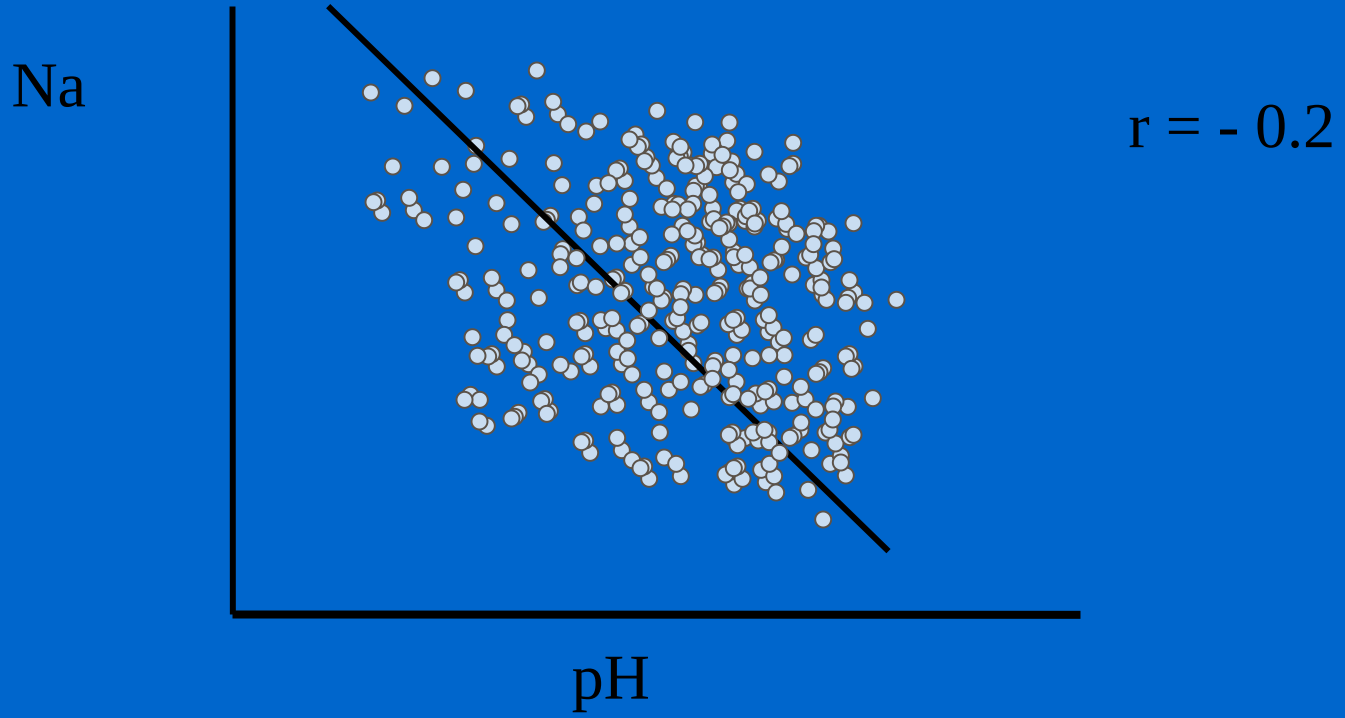
- **Moderate Negative Correlation**

r = -.80



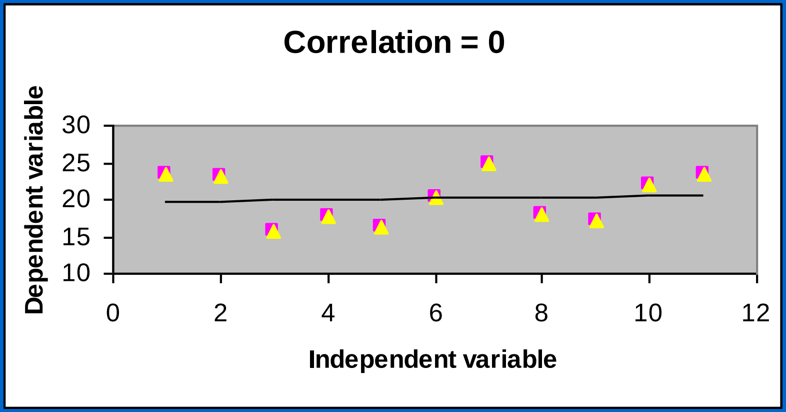
Degree of correlation

- **Weak negative Correlation**



Range of correlation coefficient

In case of a weak relationship the value of *r* will be close to 0.



$$r = [n(\sum xy) - (\sum x)(\sum y)] \,/\, \{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]\}^{0.5}$$

Sample	Na X	Cl Y	X*Y	X^2	Y^2
1	43	128	5504	1849	16384
2	48	120	5760	2304	14400
3	56	135	7560	3136	18225
4	61	143	8723	3721	20449
5	67	141	9447	4489	19881
6	70	152	10640	4900	23104
$\sum$	**345**	**819**	**47634**	**20399**	**112443**

- Substitute in the formula and solve for *r*:

  $r = \{(6*47634)-(345*819)\}/\{[(6*20399)-345^2][(6*112443)-819^2]\}^{0.5}$.

  $r = 0.897$.

- The correlation coefficient suggests a strong positive relationship between Na and Cl.
- http://www.youtube.com/watch?v=_lHnF5Rd8Wc

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:
- $\gamma\ (xi, xj) = 1/2[z(xi) - z(xj)]^2$
- where x is a geographic point and *z(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.

Q19 : What are the units of measure for the semivariance?

A19 : The square of the units of measure of the variable.

Q20 : Here are the first two points of Qassim soil sample dataset:

	coordinates	Rock	Cd	Cu	Pb	Co	Cr	Ni	Zn
1	(2.386, 3.077)	Sand stone	1.740	25.72	77.36	9.32	38.32	21.32	92.56
2	(2.544, 1.972)	Sandstone	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;

A20 :

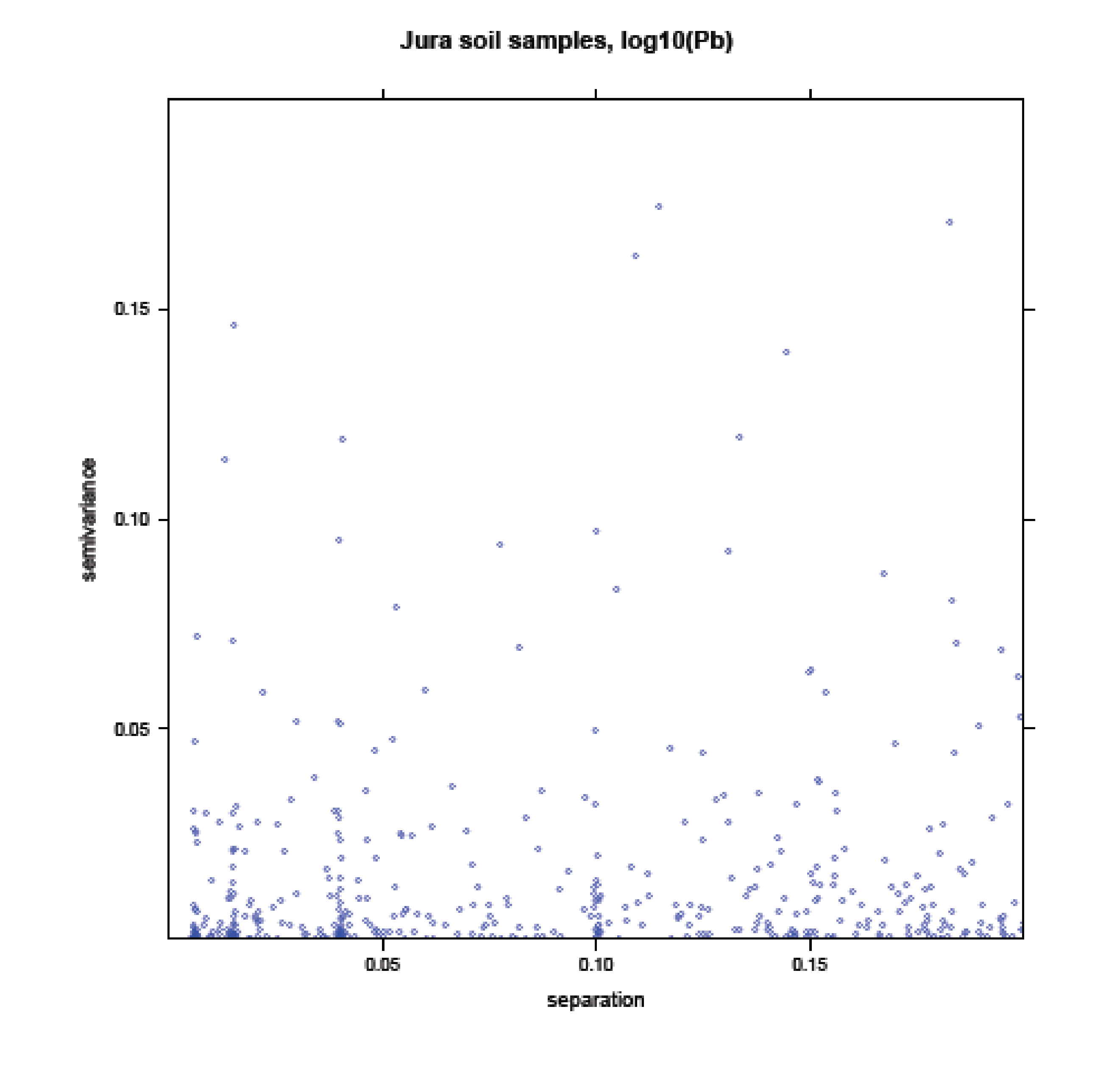
1. $\sqrt{(2.386 - 2.544)^2 + (3.077 - 1.972)^2} = 1.1162$ km
2. $77.36 - 77.88 = -0.52$ mg kg-1
3. $0.5 \cdot (77.36 - 77.88)^2 = 0.1352$ (mg kg-1)2

- 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 spatial dependence.

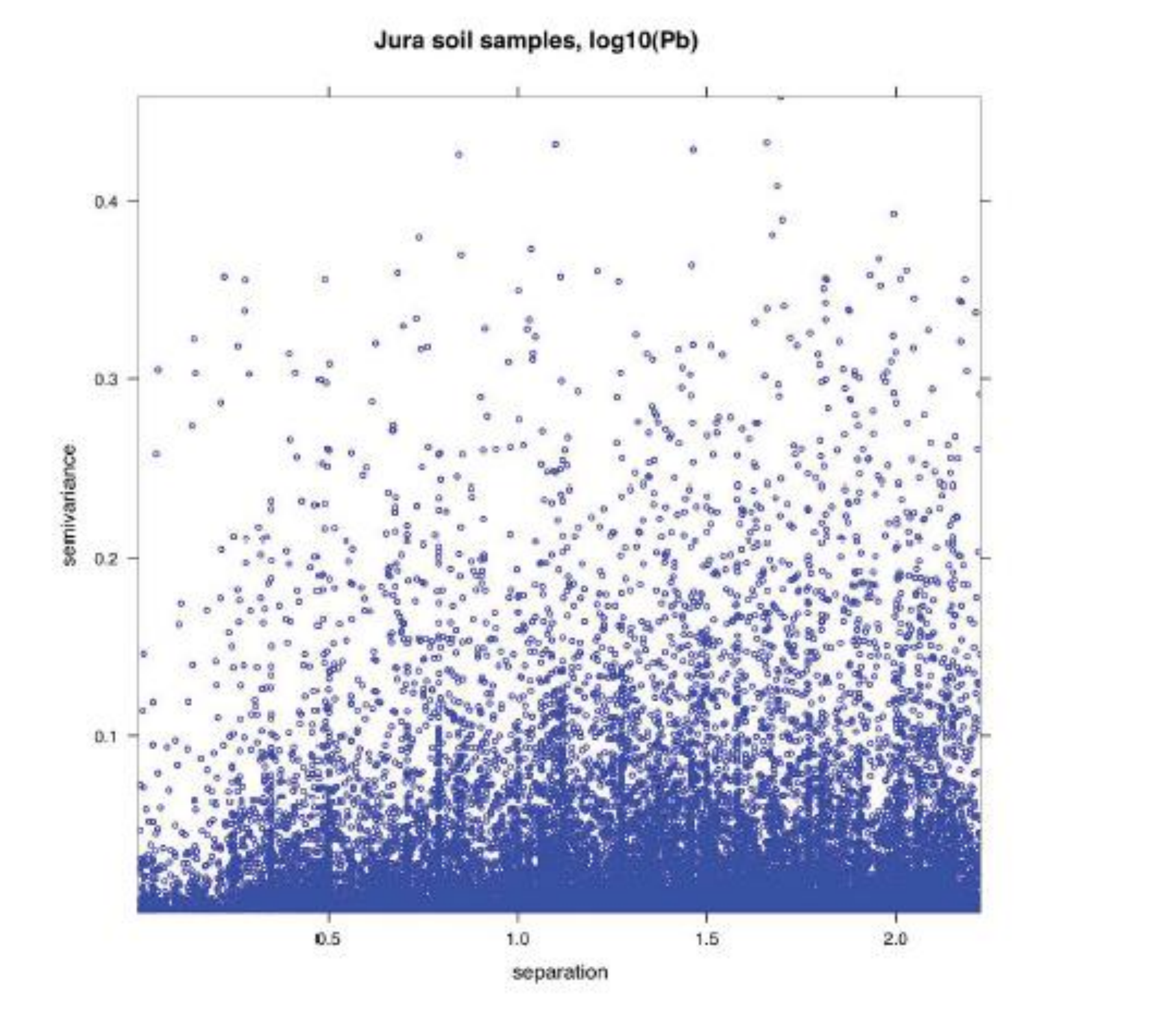
The variogram cloud

- 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 comparison 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

Separations to 200 m



Separations to 2 Km



Q21 : Can you see a trend in the semi-variances as the separation distance increases?

A21 : As separation increases, so does semi-variance.

Q22 : What is the difficulty with interpreting this graph?

A22 : This is quite difficult to see because of the large number of low semi-variances;

* The second 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 first graph this is a bit clearer, but this only applies to the closest point-pairs.

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

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

m(h) is the number of point pairs separated by vector h

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;
- The largest separation should not exceed half the longest separation in the dataset; in general it 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;
- It can use 1/3 of the longest separation, and divides this into 15 equal-width bins.

Numerical example of an experimental variogram

Here is an experimental variogram of log10Pb from the Qassim 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.

Q23 :

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?

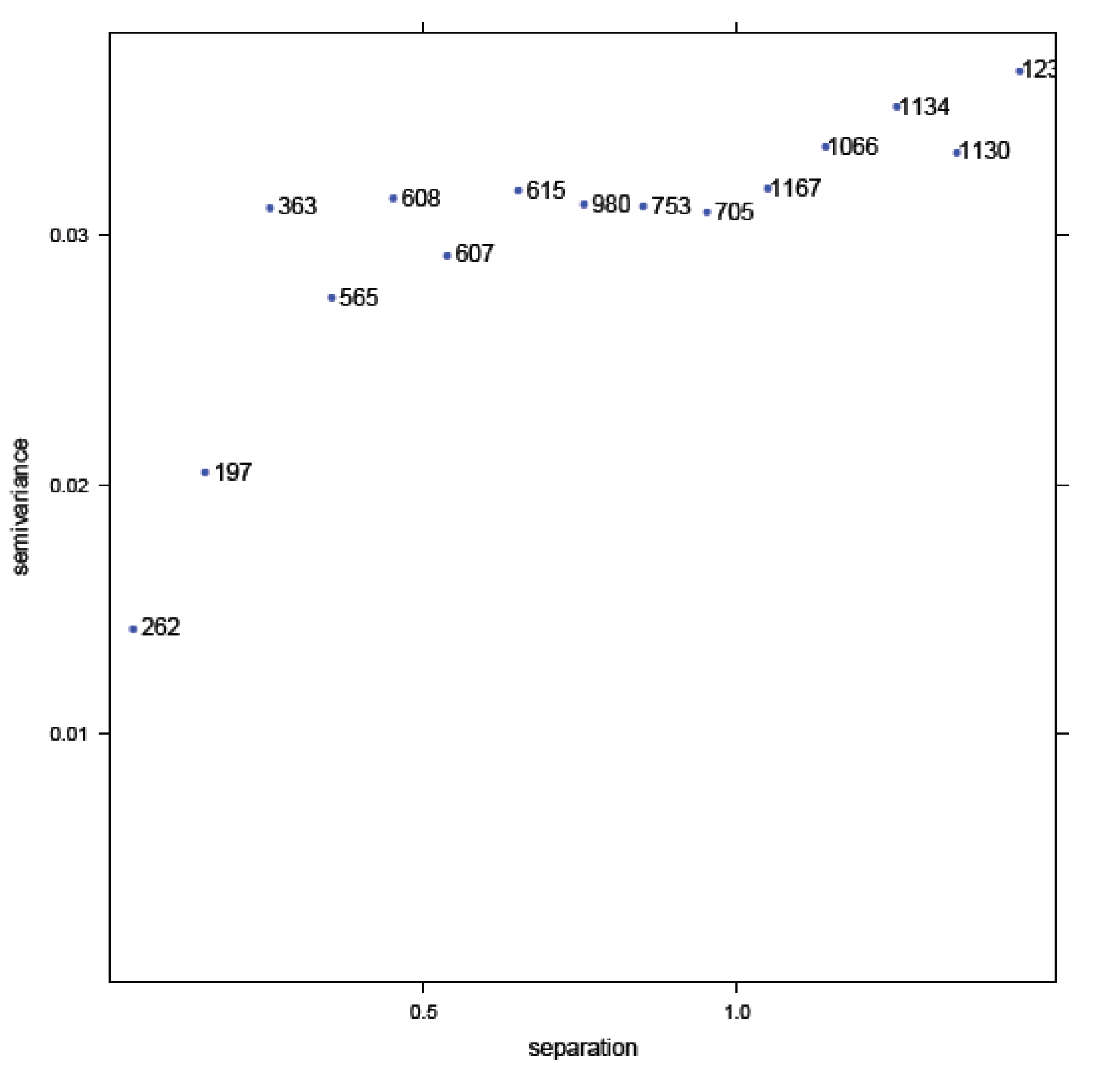
A23 :

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;
4. 0.020510

- Q24 : What is the trend in the average semi-variances as the average separation distance increases?
- A24 : There is a definite increase: at closer separations the semi-variance is less. There is some “noise”, the trend is not monotonic.

Plotting the experimental variogram

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



Experimental variogram of log10Pb, Qassim soil samples

Q25 :

1. How many point-pairs are in bin with the closest separation?

262

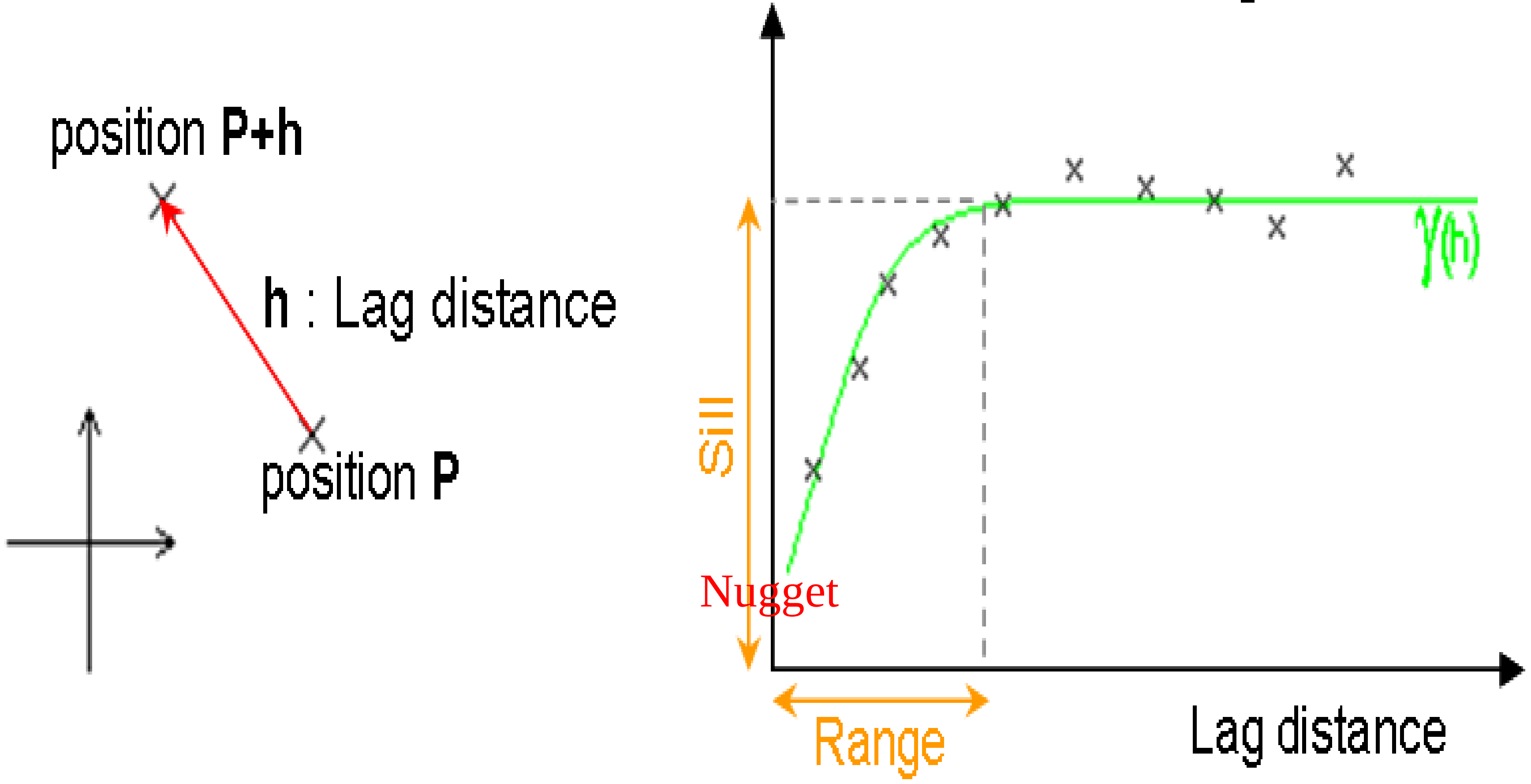
2. What is the average separation of all the point-pairs in this bin? (You will have to estimate by eye from the graph)

≈ 0.04km

3. What is the average semivariance of all the point-pairs in this bin?(You will have to estimate by eye from the graph)

≈ 0.013(log10 mg kg^{-1})^2

Semivariogram



Features of the experimental variogram

Later we will look at fitting a theoretical model to the experimental 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

Q26 :

1. What is the approximate sill of this experimental variogram?

   Sill: 0.031;

2. What is the approximate range of this experimental variogram?

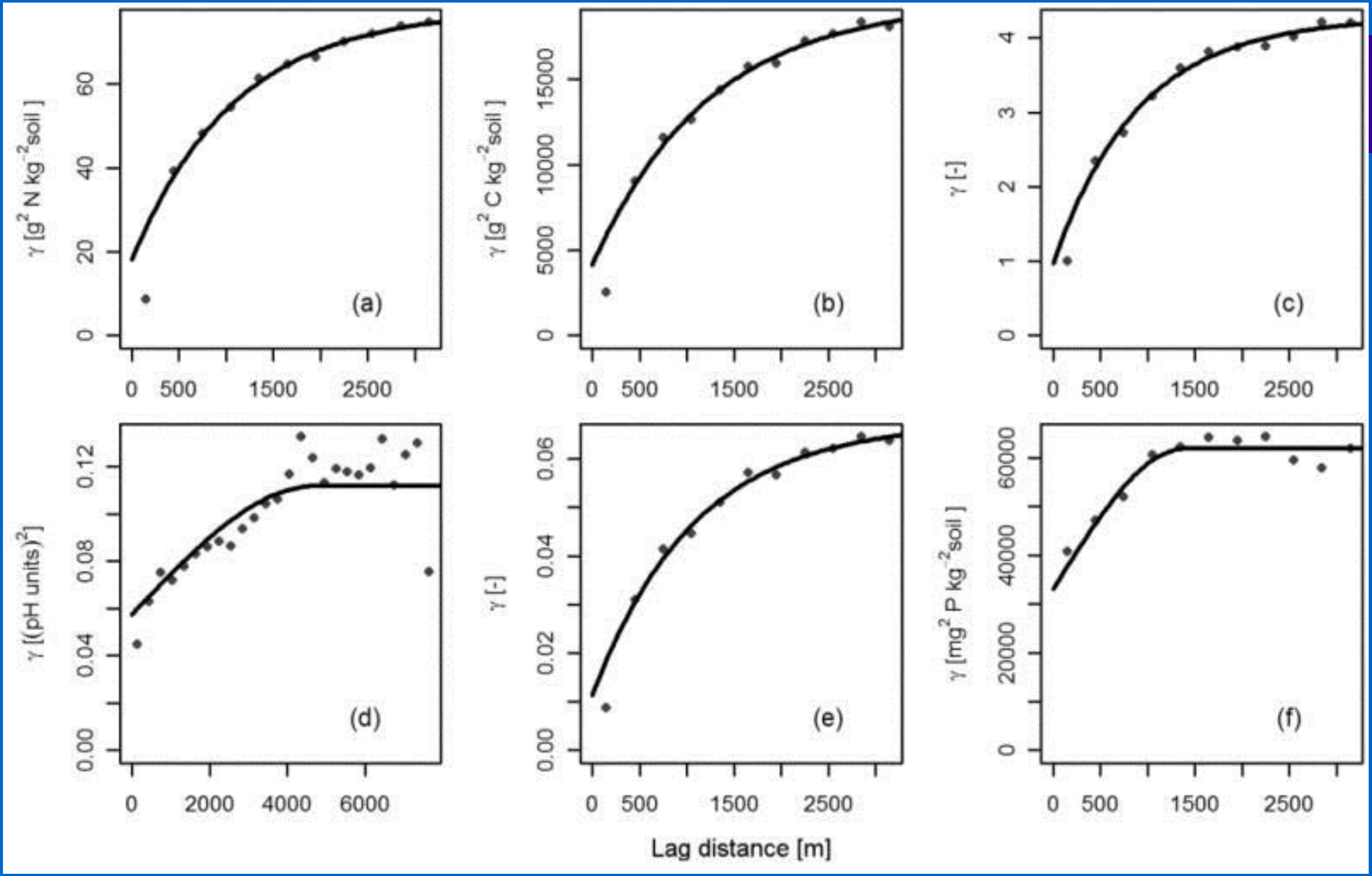
   Range: 0.5 km for the above sill

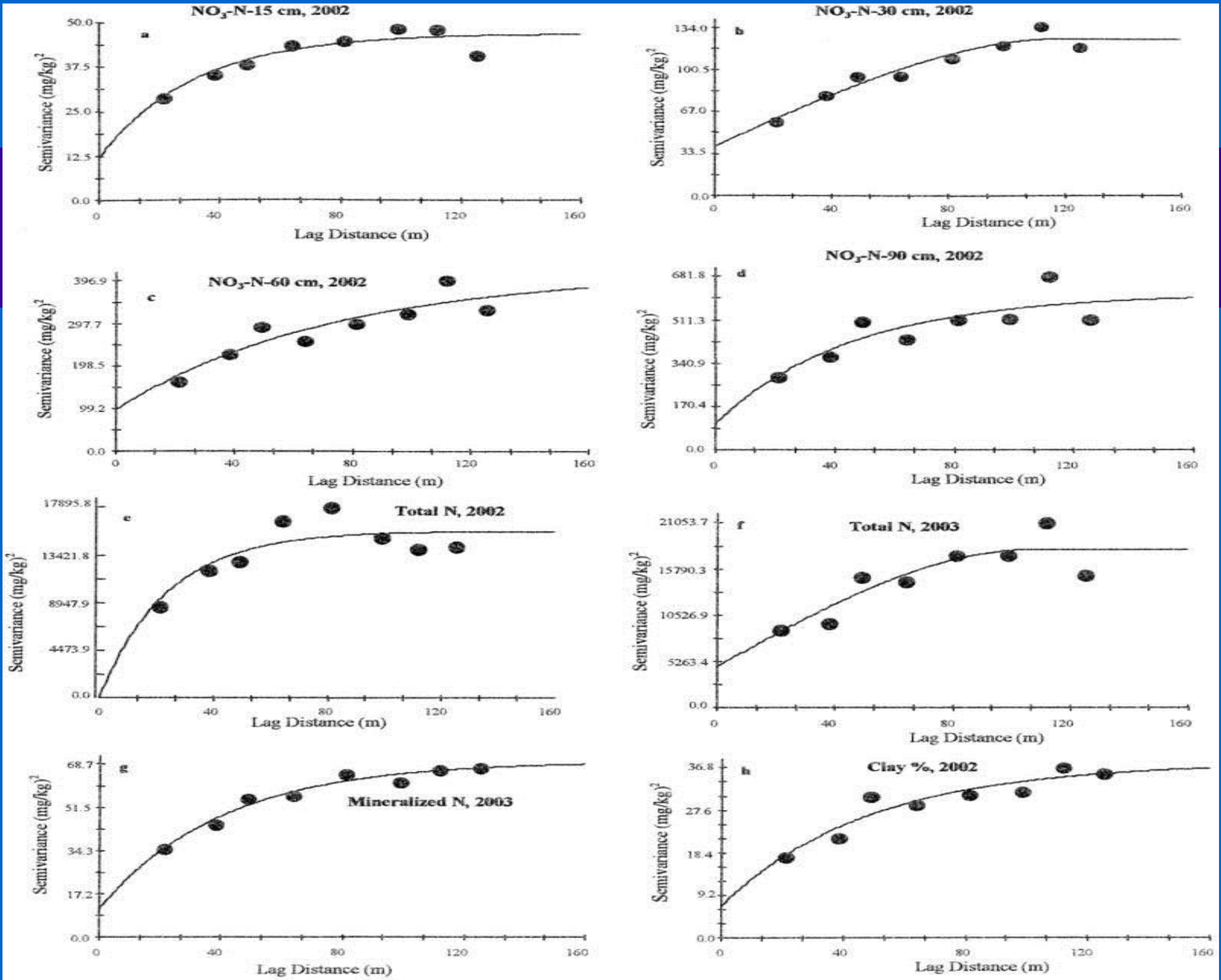
3. What is the approximate nugget of this experimental variogram?

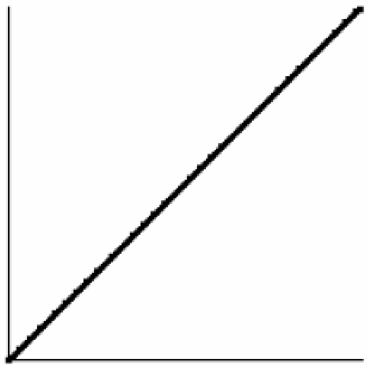
   Nugget: 0.013; extrapolate by eye to the y-axis.

Evidence of spatial dependence

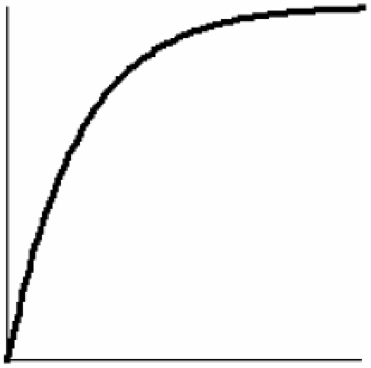
- The experimental variogram provides evidence that there is local spatial dependence.
- The variability between point-pairs is lower if they are closer to each other; i.e. the separation is small.
- There is some distance, the range where this effect is noted; beyond the range there is no dependence.
- 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 experimental variogram has the sill equal to the nugget; this is called a pure nugget effect
- The next graph shows an example.


γ [g^2 N kg^-2 soil]
0
20
40
60
0
500
1500
2500
(a)
γ [g^2 C kg^-2 soil]
0
5000
10000
15000
0
500
1500
2500
(b)
γ [-]
0
1
2
3
4
0
500
1500
2500
(c)
γ [(pH units)^2]
0.00
0.04
0.08
0.12
0
2000
4000
6000
(d)
γ [-]
0.00
0.02
0.04
0.06
0
500
1500
2500
(e)
γ [mg^2 P kg^-2 soil]
0
20000
40000
60000
0
500
1500
2500
(f)
Lag distance [m]

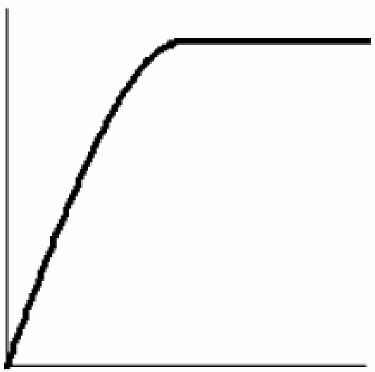




Linear Model

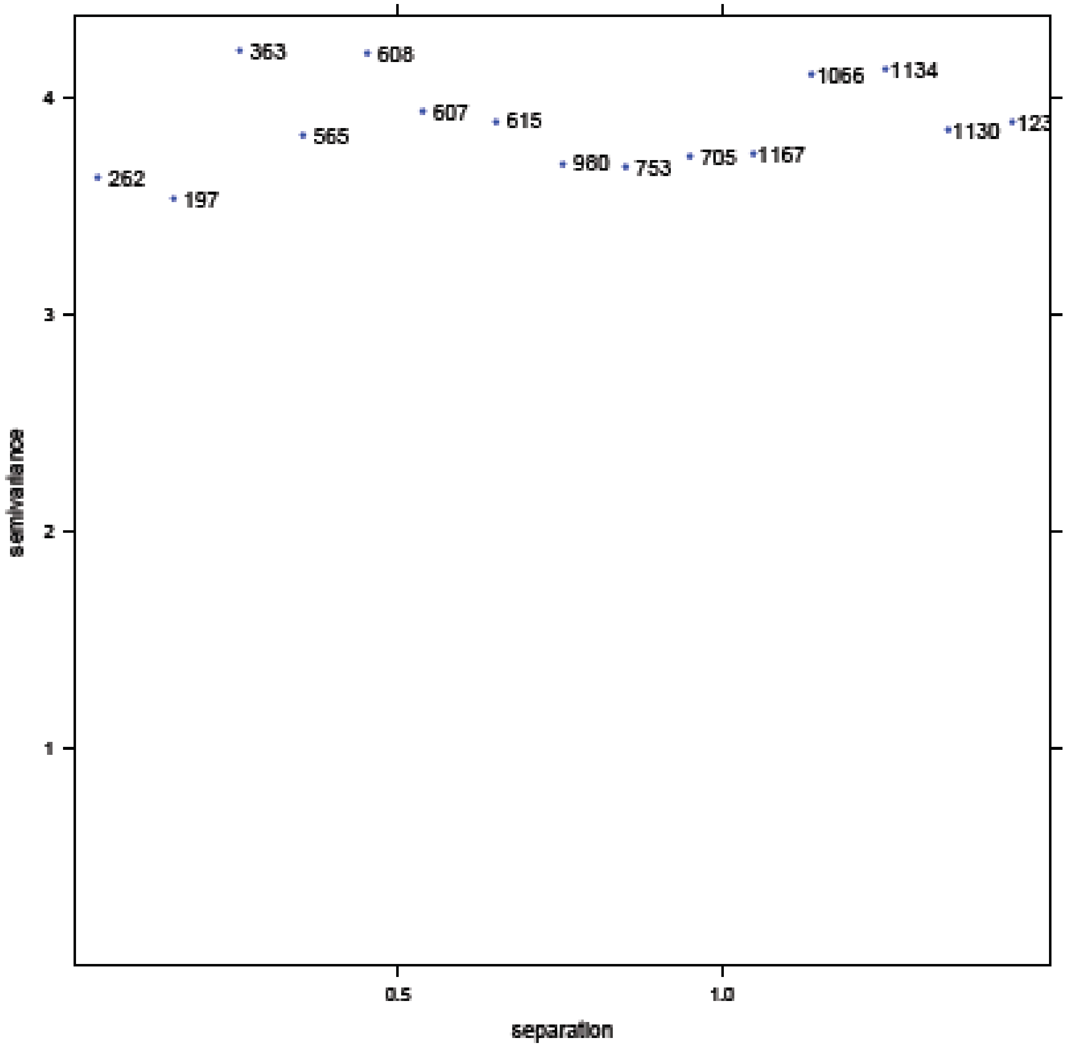


Exponential Model



Spherical Model

Experimental variogram of a variable with no spatial dependence



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 isotropic: it is variation at a point (which has no direction)

- Q27 : 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?
- A27 : 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.

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 direction 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)

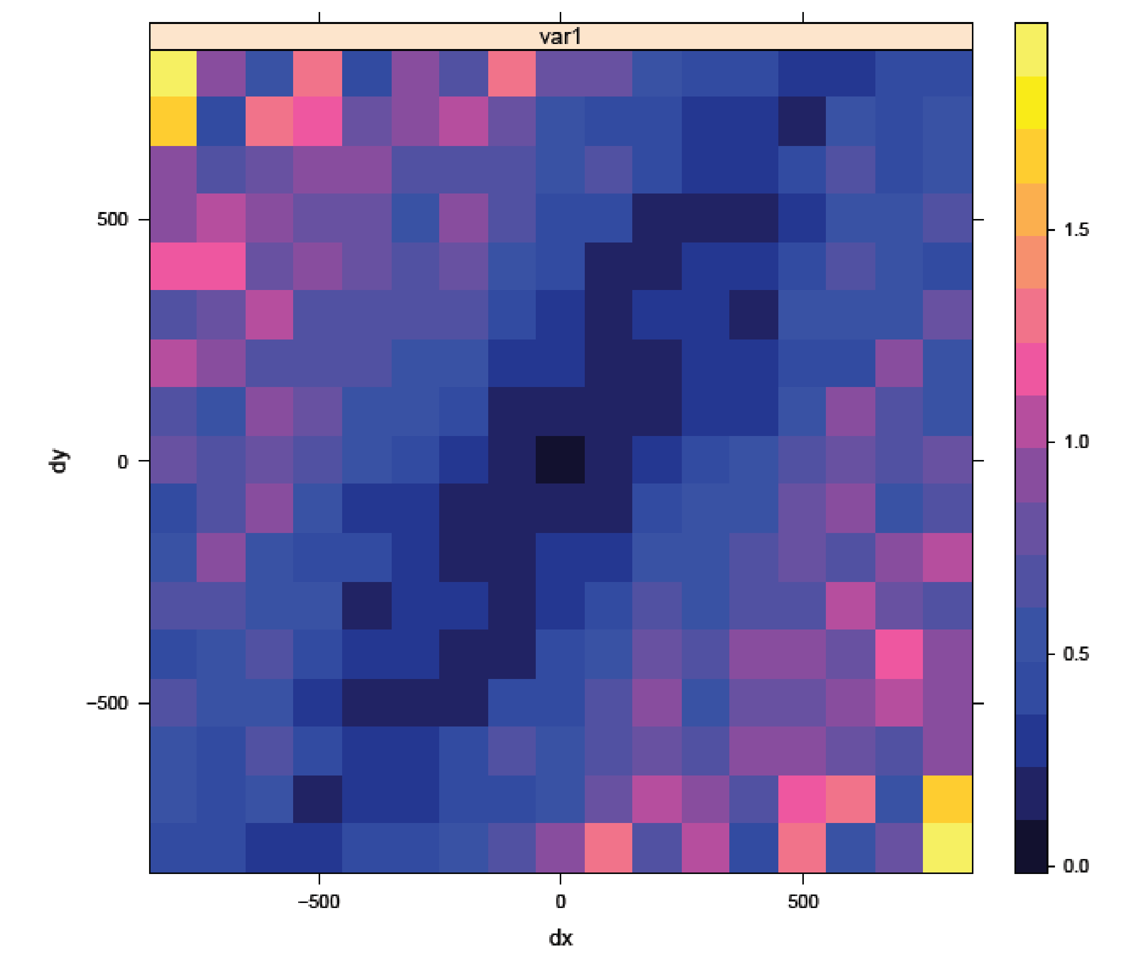
Detecting anisotropy with a variogram surface

- 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)

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



Variogram surface showing anisotropy

- **Q28 : What is the approximate semivariance at a separation of distance 500 m, direction due E (or W)?**

- **A28 : 0*.8; compare the purple colour with the legend at the right of the figure.***

- **Q29 : What is the approximate separation distance for the cell at 300 m E, 200 m S? •**

- **A29 :**

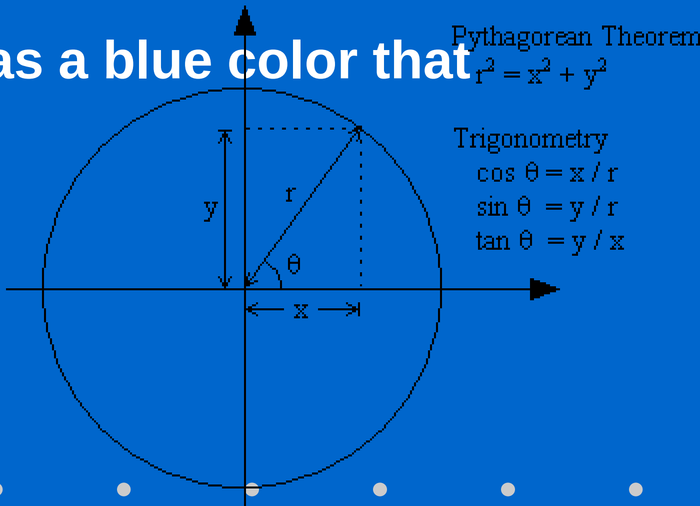
- **$((300)^2+(200)^2)^{1/2}$= 360**

- **Q30 : What is the approximate separation azimuth (direction from N) in this cell?**

- **A30 : arctan*(200/300) +* $\pi/2$ *=2.16 radians from North; this is 2.16 ** $(180/\pi)$ *=124 deg***

- **Q31 : What is the approximate semivariance at this separation?**

- **A31 : The cell at dx = +300, dy = -200 has a blue color that corresponds to 0.5.**



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 color stays the same (or similar).

Interpreting the variogram surface

- Q32 : Which direction (as an azimuth from N) shows the strongest spatial dependence, i.e. where the semivariance stays low over the farthest distance?
- A32 : Approximately 30 degrees from N. The dark blue colors form a clear band in the NNE - SSW axis.
- Q33 : 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?
- A33 : Yes, the axis at approximately 120 degrees from N (30 + 90) does appear to be the direction in which semivariance increases most rapidly.

Interpreting the variogram surface

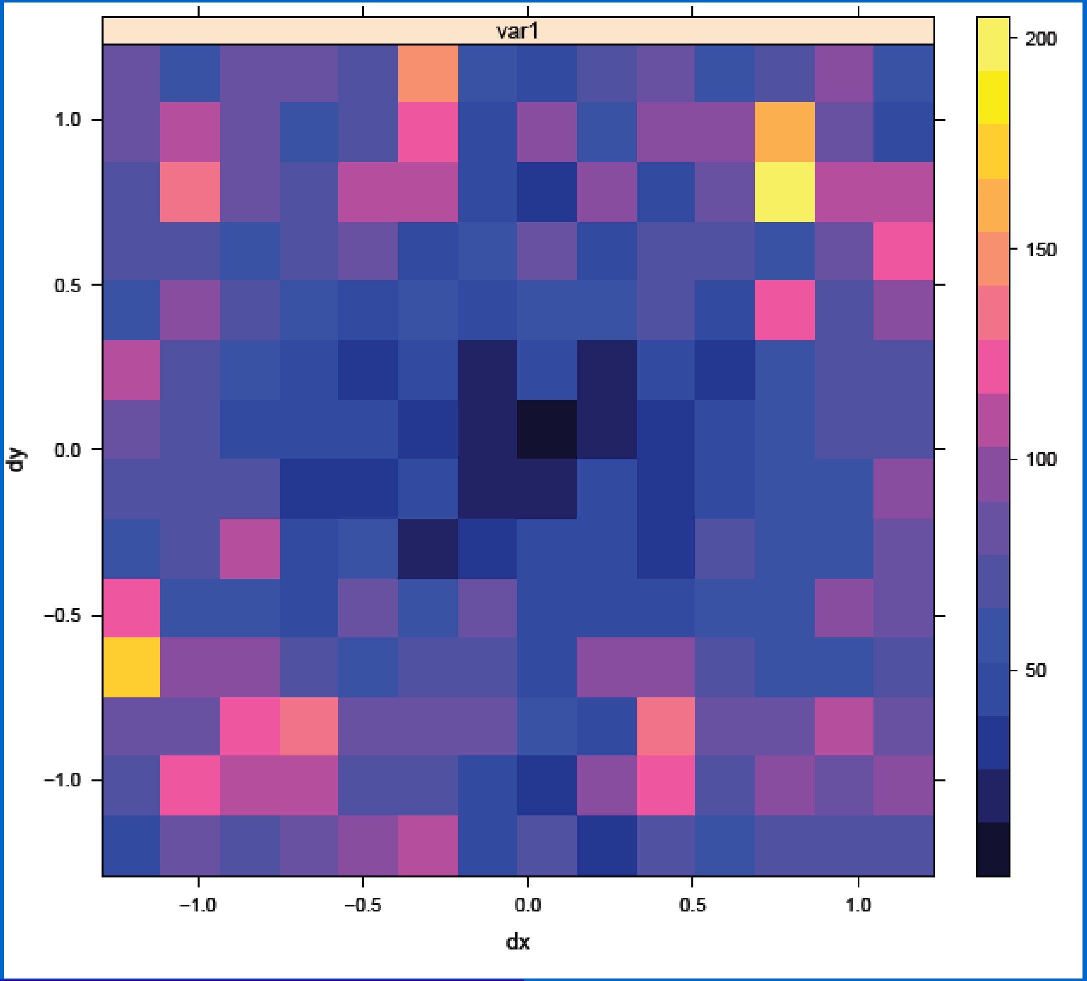
- We saw above the “pure nugget” variogram, showing that not all variables have spatial dependence.

- 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, width of bins as for omnidirectional variograms;

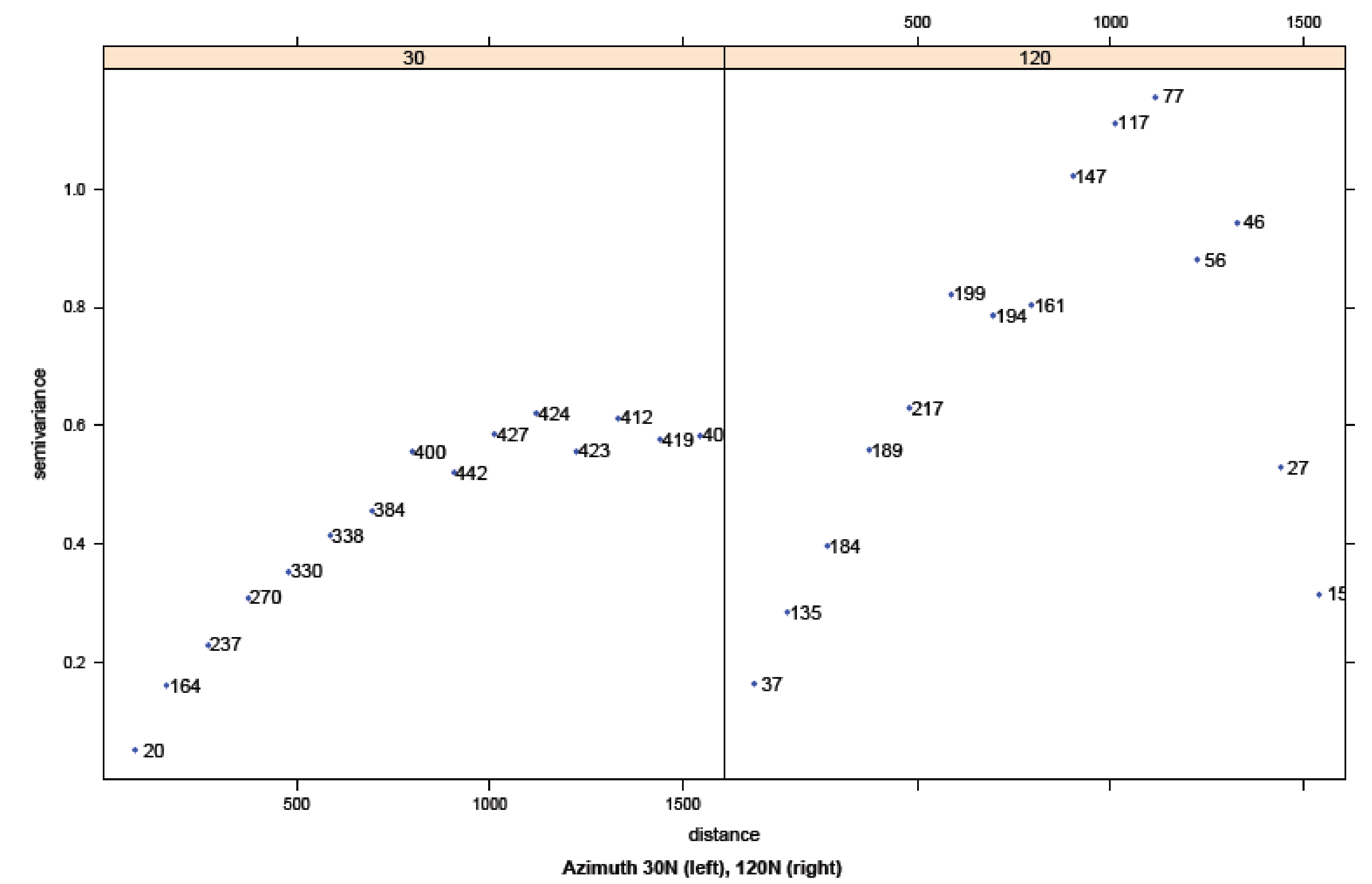
2. Direction of the major or minor axis in 1st quadrant; implicitly specifies

perpendicular as other axis;

3. Tolerance: Degrees on either side which are considered to have the 'same' angle;

4. Band width: Limit the bin to a certain width; this keeps the band from taking in too many far-away points.

Directional variograms of Log(Zn), Qassim soil samples



- Q34 : Do the two directions have similar variograms? (Consider sill, range, nugget)

- A34 : 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.

- Q35 : **In which of the two perpendicular axes is the spatial dependence stronger (longer range, lower nugget to sill ratio)?**

- A35 : **30N.**

- Q36 : **How is this evidence that the spatial process by which the metal (Zn) was distributed over the area is directional?**

- A36 : **The longer range and lower variability in the 30N direction shows that whatever process distributed the Zn is oriented along that axis.**