Applied geostatistics

Lecture 1
What is geostatistics; Geostatistical computing

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

January 7, 2014

Topics for this lecture

1. What is geostatistics?
2. The added value of geostatistics
3. Feature and geographic spaces
4. Geostatistical computing: inventory of packages
5. The R Project for Statistical Computing: what and why?
6. Exercise: Introduction to the R environment and S language
7. Appendix: learning resources

Topic 1: What is "geostatistics"?

1. What is **"statistics"**?

2. What then is **"geo"-statistics**?

Commentary

We start this lecture series by being clear on what we mean by **geostatistics**. First we have to define **statistics**, and then see what it means when we add the **geo-**.

The term **"statistics"** has two common meanings, which we want to clearly separate: **descriptive** and **inferential** statistics.

But to understand the difference between descriptive and inferential statistics, we must first be clear on the difference between **populations** and **samples**.

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 the different tree species in a forest reserve. Are all the trees in this forest reserve a* **population** *or* **sample***?* *Jump to A1* •

Q2 : *If we make a transect from one side of the forest to the other, and identify the species of all the trees within 10 m of the centre line, is this a* **population** *or* **sample** *of the trees in the forest reserve?* *Jump to A2* •

What do we mean by "statistics"?

Two common use of the word:

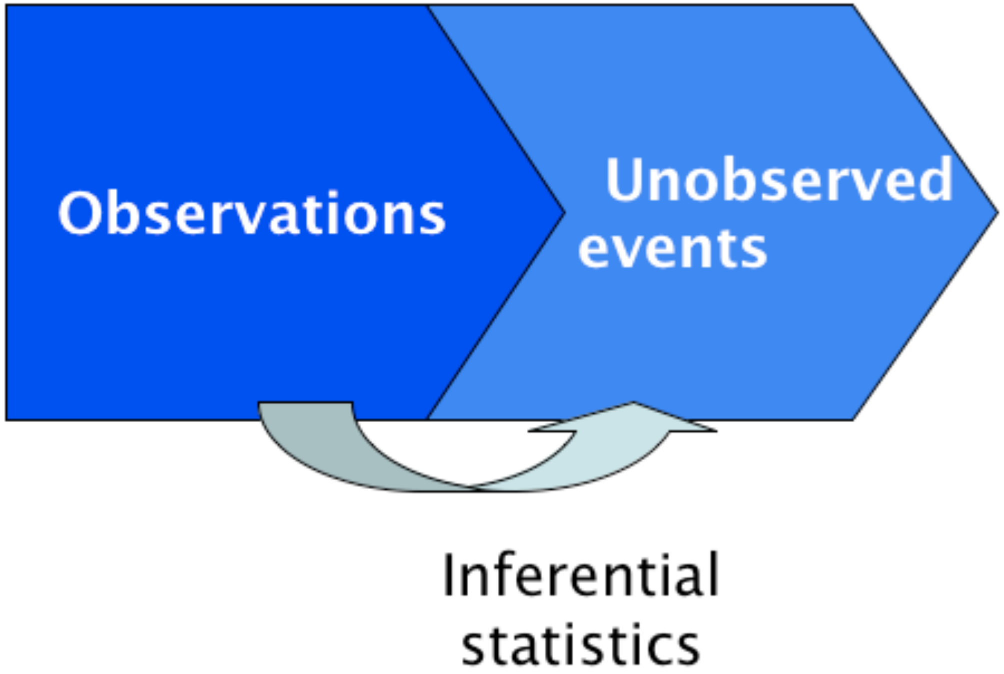
1. **Descriptive** statistics: numerical summaries of **samples**;
   - (what was observed)
   - Note the 'sample' may be **exhaustive**, i.e., identical to the population
2. **Inferential** statistics: from **samples** to **populations**.
   - (what could have been or will be observed in a larger **population**)

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."

Inference



To check your understanding . . .

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?* *Jump to A3* •

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?* *Jump to A4* •

A concise definition of inferential statistics

Statistics: "The determination of the **probable** from the **possible**"

– Davis, *Statistics and data analysis in geology*, p. 6

. . . which implies the **rigorous definition** and then **quantification** of "probable".

- Probable **causes** of **past** events or observations
- Probable **occurrence** of **future events** or observations

This is a definition of **inferential** statistics:

$$\textbf{Observations} \Longrightarrow \textbf{Inferences}$$

Commentary

As humans, we infer constantly from the evidence around us. For example, I **observe** a person who can not walk a straight line, whose speech is slurred, and who smells strongly of alcohol – these are observable facts. I **infer** that the person is drunk – yet I didn't see the person drink alcohol, and I haven't analyzed the person's blood alcohol level. Often we do not realize the line between observable facts and inferences.

In statistical inference we **quantify** the degree of **probability** that our inference is true.

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

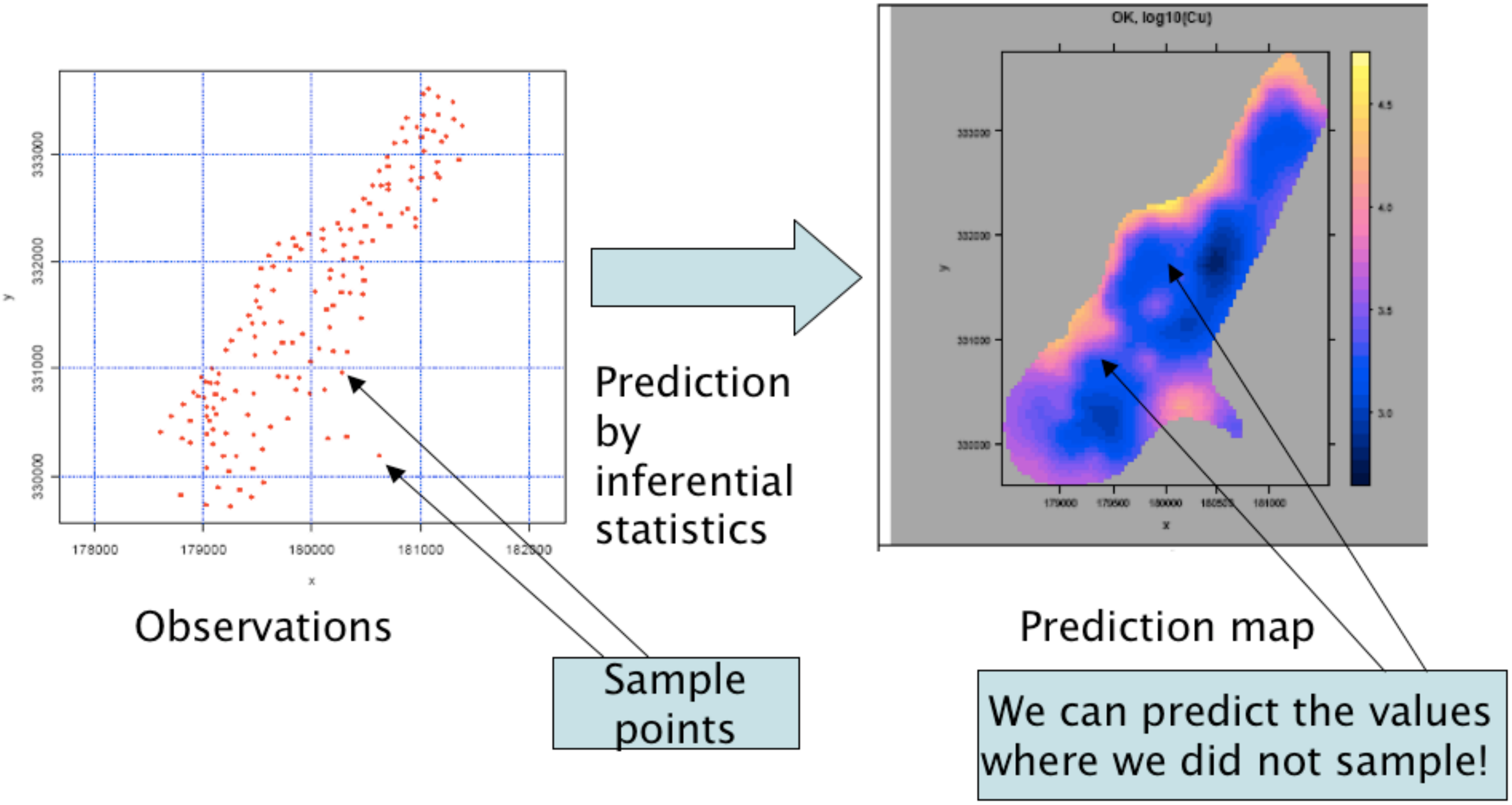
   - So we need to **model** it, i.e. make a conceptual or mathematical representation, from which we **infer** the process.
   - But how do we know if the model is "correct"?
     - Are we imagining relations where there are none?
     - Are there true relations we haven't found?
   - Statistical analysis 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 the Meuse river floodplain in the southern Netherlands. 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).

Inference


OK, log10(Cu)
Prediction by inferential statistics
Observations
Sample points
Prediction map
We can predict the values where we did not sample!
178000
179000
180000
181000
182000
x
y
330000
331000
332000
333000
4.5
4.0
3.5
3.0

What is "geo"-statistics?

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

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. **coördinates** (**where** it is located); and
2. **attributes** (**what** it is).

Commentary

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   30.8
                4   56.4
                5   66.4
                6   72.4
                7     60
                8    141
                9   52.4
               10   41.6
               11     46
               12   56.4
```

The column `Pb` is the **attribute** of interest.

To check your understanding . . .

Q5 : *Can we determine the median, maximum and minimum of this set of samples?* *Jump to A5* •

Q6 : *Can we make a map of the sample points with their Pb values?* *Jump to A6* •

Commentary

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

These are soil samples taken in the Jura mountains of Switzerland, and their lead content; but this time with their coördinates. First let's look at the tabular form:

Observation_ID	E	N	Pb
1	2.386	3.077	77.36
2	2.544	1.972	77.88
3	2.807	3.347	30.80
4	4.308	1.933	56.40
5	4.383	1.081	66.40
6	3.244	4.519	72.40
7	3.925	3.785	60.00
8	2.116	3.498	141.00
9	1.842	0.989	52.40
10	1.709	1.843	41.60
11	3.800	4.578	46.00
12	2.699	1.199	56.40

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

To check your understanding . . .

Q7 : *Comparing this to the non-geostatistical list of soil samples and their lead contents (above), what new information is added here?* *Jump to A7* •

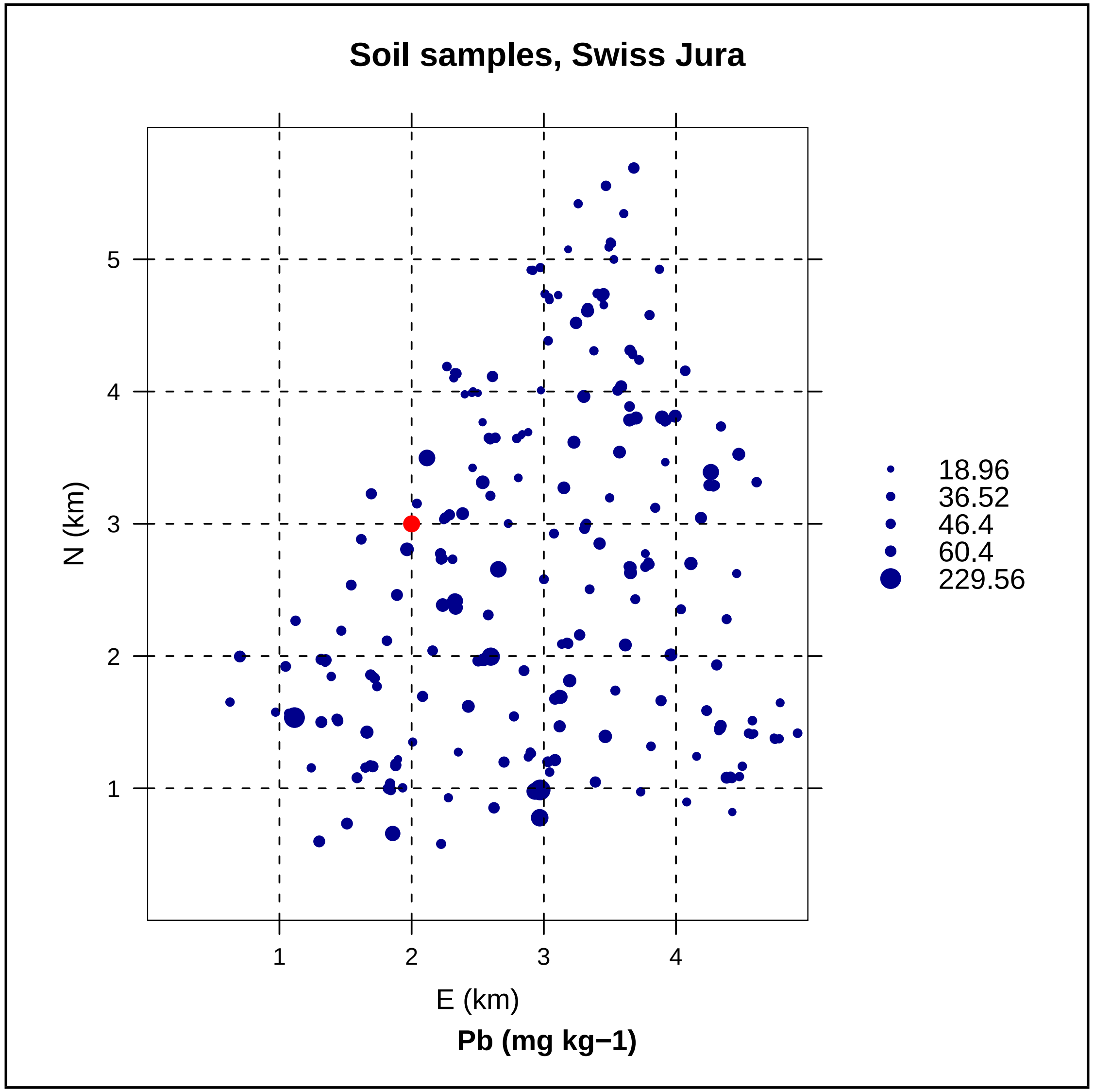
Commentary

On the figure (next slide) you will see:

1. A **coördinate 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).

Post-plot of Pb values, Swiss Jura



To check your understanding . . .

Q8 : *In the figure, how can you determine the coördinates of each sample point?* *Jump to A8* •

Q9 : *What are the coördinates of the sample point displayed as a red symbol?* *Jump to A9* •

Q10 : *What is the* **mathematical origin** *(in the sense of Cartesian or analytic geometry) of this coördinate system?* *Jump to A10* •

Q11 : *How could these coördinates be related to some common system such as UTM?* *Jump to A11* •

To check your understanding . . .

Q12 : *Suppose we have a satellite image that has not been geo-referenced. Can we speak of geostatistics on the pixel values?* *Jump to A12* •

Q13 : *In this case, what are the* **coördinates** *and what are the* **attributes**? *Jump to A13* •

Q14 : *Suppose now the images has been geo-referenced. What are now the coördinates?* *Jump to A14* •

Commentary

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?

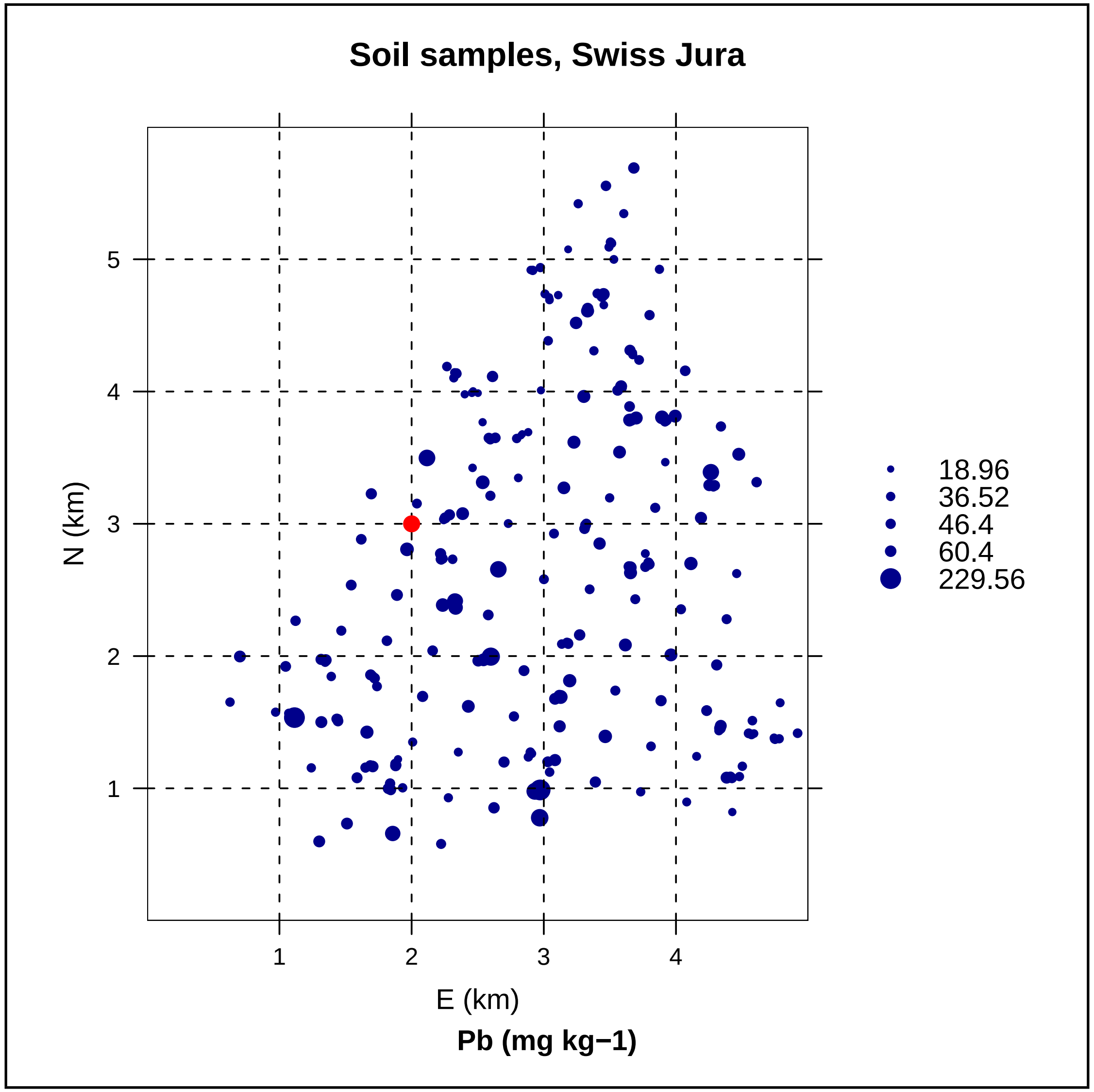
Topic 2: The added value of geostatistics

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**

Commentary

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.

Post-plot of Pb values, Swiss Jura



To check your understanding . . .

Q15 : *Do large circles (representing high Pb concentrations) seem to form clusters?* *Jump to A15* •

Q16 : *Do small circles (representing low Pb concentrations) seem to form clusters?* *Jump to A16* •

Q17 : *What is the approximate radius the clusters?* *Jump to A17* •

Topic 3: Feature and geographic spaces

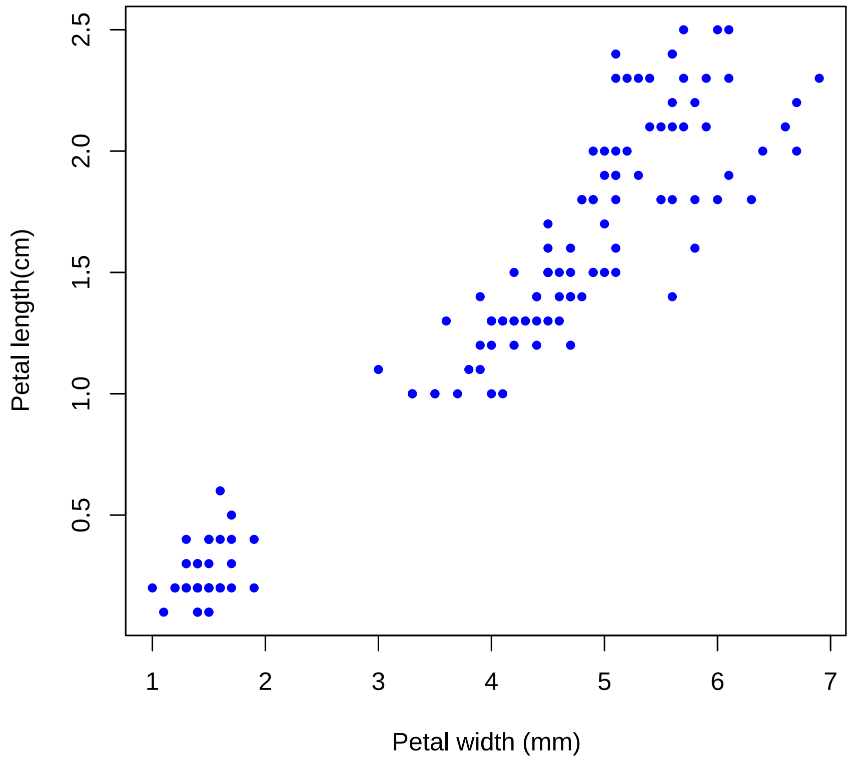
- The word **space** is used in mathematics to refer to any set of variables that form metric axes and which therefore allow us to compute a **distance** between points in that space.
- If these variables represent geographic **coördinates**, we have a **geographic** space.
- If these variables represent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

Commentary

It is important to understand these two uses of the word **space**, because we often want to contrast an analysis in **feature** space (not taking spatial position into account) with an analysis in **geographic** space (considering spatial position as the key element in the analysis).

Let's see an example of a feature space. This concept should be familiar from non-spatial statistics, although the term "feature space" may be new to you.

Scatterplot of a 2D feature space



This is a **visualisation** of a 2D feature space using a **scatterplot**. The points are individual iris flowers measured by Edgar Anderson in 1935 and published in *The irises of the Gaspé Peninsula*, Bulletin of the American Iris Society, 59, 2-5.

To check your understanding . . .

Q18 : *What are the two dimensions of this feature space, and their units of measure?* *Jump to A18* •

Q19 : *Does there appear to be a correlation between the two dimensions for this set of observations?* *Jump to A19* •

Feature space

This "space" is not geographic space, but rather a **mathematical space** formed by any set of variables:

- **Axes** are the range of each variable
- **Coördinates** are values of variables, possibly transformed or combined
- The observations are related in this 'space', e.g. the "distance" between them can be calculated.
- We often plot variables in this space, e.g. **scatterplots** in 2D or 3D.

Note: **Feature space** is sometimes referred to as **attribute space**.

Commentary

You are probably quite familiar with feature space from your study of non-spatial statistics.

Even with **one** variable, we have a unit of measure; this forms a 1D or **univariate** feature space.

Most common are **two** variables which we want to relate with **correlation** or **regression** analysis; this is a **bivariate** feature space.

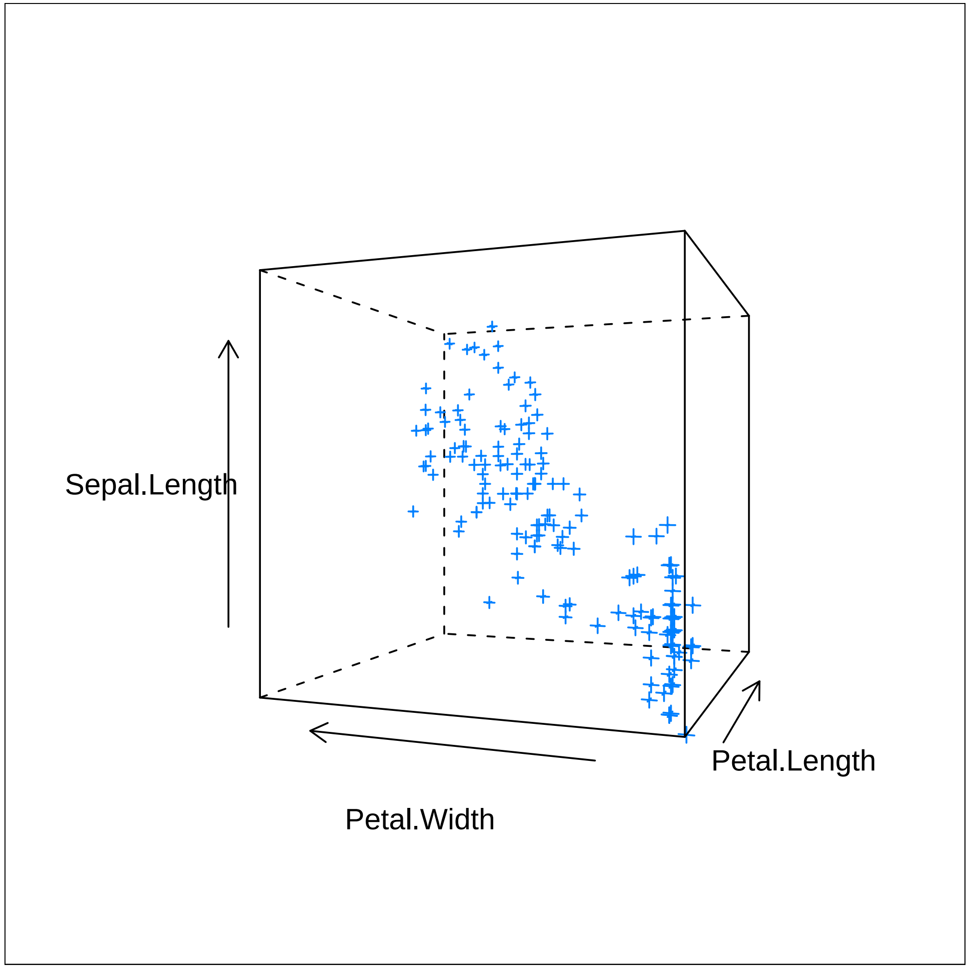
In **multivariate** analysis the feature space has more than two dimensions.

Commentary

Multivariate feature spaces can have many dimensions; we can only see three at a time.

The next slide is a visualisation of a **3D** feature space using a 3D-scatterplot. Again the observations are from the Anderson *Iris* data.

Scatterplot of a 3D feature space



Anderson *Iris* data.

To check your understanding . . .

Q20 : *What are the three dimensions of this feature space?* *Jump to A20* •

Q21 : *Does there appear to be a correlation between the three dimensions for this set of observations?* *Jump to A21* •

Commentary

So, feature space is perhaps a new term but not a new concept if you've followed a statistics course with univariate, bivariate and multivariate analysis.

What then is **geographic** space? Simply put, it is a mathematical space where the axes are **map coördinates** that relate points to some reference location on or in the Earth (or another physical body).

These coördinates are often in some **geographic coördinate system** that was designed to give each location on (part of) the Earth a unique identification; a common example is the Universal Transmercator (UTM) grid.

However, a **local coördinate system** can be used, as long as there is a clear relation between locations and coördinates.

Geographic space

- Axes are 1D **lines**; they almost always have the **same units of measure** (e.g. metres, kilometres . . .)
- **One-dimensional**: coördinates are on a line with respect to some origin (0): $(x_1) = x$
- **Two-dimensional**: coördinates are on a grid with respect to some origin $(0, 0)$: $(x_1, x_2) = (x, y) = (E, N)$
- **Three-dimensional**: coördinates are grid and elevation from a reference elevation: $(x_1, x_2, x_3) = (x, y, z) = (E, N, H)$
- Note: **latitude-longitude** coördinates do not have equal distances in the two dimensions; they should be transformed to metric (grid) coördinates for geo-statistical analysis.

To check your understanding . . .

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

Jump to A22 •

Topic 4: Geostatistical computing

1. Why geostatistical computing?

2. Geostatistical computing programs

Why geostatistical computing?

- **Visualization**: look for patterns in geographic and feature space; these suggest possible analyses
  * Is there a **trend** in the attributes with geographic position? E.g. rainfall decreasing away from an ocean
  * Is there local **spatial dependence**, i.e. values of points or polygons close by are more similar than those further apart?
  * Is there a **spatial pattern** to the sample points?

- **Computation**: large numerical systems (e.g. the kriging system) that are practically impossible to solve by hand if there are more than a few points.

Commentary

It is impossible to consider geostatistical analysis without modern computing facilities. Here we list the many possibilities for geostatistical computing, along with a list of resources.

In this course we will use the R software environment for statistical computing and graphics, but we want to list the many alternatives and reasons you might choose one of them.

Commentary

There is a bewildering variety of software that deals with geostatistics. Some are commercial, some free. Some are part of a **larger system** (e.g. a GIS or a statistical computing environment), some **stand-alone**. Some only run on one operating system, some on many. Some are more comprehensive (more functionality) than others. What follows is only a partial list.

If you are connected to the internet, **all the web links in these notes are "live"**; by clicking on the URL your browser will go to that site.

What programs are available?

- Lists
  - **AI-GEOSTATS**: A Web Resource for Geostatistic and Spatial Statistics
    `http://www.ai-geostats.org/`
    Freeware used for spatial statistics and geostatistics are listed under the "Software" button.

- As a module of commercial GIS
  - **ArcGIS Geostatistical Analyst**: `http://www.esri.com/software/arcgis/extensions/geostatistical/index.html`
    An extension to the commercial ArcGIS program. A wide variety of procedures but weak documentation.
  - **IDRISI**: `http://www.clarklabs.org/products/`
    From Clark Labs (US); a medium-cost GIS with good geostatistical functions.
  - **ILWIS**: `http://www.itc.nl/ilwis/`
    From ITC (NL); almost free; with some geostatistical functions and excellent documentation (User's Guide Ch. 11, detailed on-line help).

- Stand-alone programs
  - **gstat**: `http://www.gstat.org/`
    Developed by Edzer Pebesma at Utrecht University (NL); we will use this as our primary tool but in the form of an R library.
  - **gslib**: `http://www.gslib.com/`
    This is sophisticated code used by many advanced practioners.
  - **GeoEAS**: `http://www.epa.gov/ada/csmos/models/geoeas.html`
    This was developed by the US Environmental Protection Agency and used for many regulatory studies.
  - **SURFER**:
    `http://www.goldensoftware.com/products/surfer/surfer.shtml`
    This commerical program has a wide variety of methods for making smooth surfaces or contour maps from point data.
  - **VESPER**:
    `http://sydney.edu.au/agriculture/pal/software/vesper.shtml`
    From the University of Syndey (AU), especially useful for interactive variogram modelling
  - **FRAGSTATS**:
    `http://www.umass.edu/landeco/research/fragstats/fragstats.html`
    Landscape analysis

- Spreadsheets

  These are designed for data manipulation; but they can also do matrix computations. Simple statistics are built-in as functions; and it is possible to program some geostatistical analysis with the **matrix** operations.

  * **MS Excel** (commerical): `http://www.microsoft.com/excel/`
  * **OpenOffice** (open-source, free): `http://www.openoffice.org/`

- As part of a **statistical computing environment**
  - **R**: `http://www.r-project.org/`: Open-source environment for statistical computing and visualisation; includes several relevant **packages**, including (among others):
    - `gstat`: variogram modelling; simple, ordinary and universal point or block (co)kriging, sequential Gaussian or indicator (co)simulation
    - `spatial`: Functions for kriging and point pattern analysis
    - `geoR`: Geostatistical analysis including traditional, likelihood-based and Bayesian methods
    - `spdep`: Spatial dependence: weighting schemes, statistics and models
    - `spatstat`: point pattern analysis
    - `sp`, underlying spatial data structures
    - `DCluster`: functions for the detection of spatial clusters of diseases
  - **S-PLUS**: A commerical implementation of the S language with a comprehensive GUI. Has many of the same libraries as R. It has recently been aquired by the TIBCO data analytics company and re-branded as Spotfire `http://spotfire.tibco.com/discover-spotfire`.
  - **GenStat**: `http://www.vsni.co.uk/software/genstat`; now commerical but originally conceived and developed at the Rothamsted Experimental Station (UK)

Commentary

So, which to choose?

1. The program must have the functionality you want
2. It must be able to read and write data in the formats you have, or to which you can convert
3. It must be computationally-correct
4. It must give you sufficient control of the parameters and sufficient understanding of what it is doing
5. It must be possible to integrate it with your other tools

The arguments for **open-source, multiple-platform** programs (like R) are:

1. You can see exactly what the code is doing if you wish
2. It is free
3. You can contribute new methods if you reach that level of skill
4. You are not limited to one vendor's operating system

There is nothing wrong with combining programs as part of a **toolkit**.

Topic 5: The R Project for Statistical Computing

1. What is it?

2. Why do we use it?

3. Structure

4. Introduction to using R

Commentary

From the many geostatistical computation programs reviewed in the previous topic, we have chosen to use the R Project for Statistical Computing in this course/module.

In this topic we introduce R, and explain its advantages and disadvantages for (geo)statisical computing.

What is R?

- **R** is an **open-source** environment for **statistical computing and visualisation**
- It is based on the **S language** developed by John Chambers at Bell Laboratories in the 1980's (the same group that developed C and UNIX©)
- It is the product of an active movement among **statisticians** for a **powerful**, **programmable**, **portable**, and **open** computing environment, applicable to the most complex and sophsticated problems, as well as "routine" analysis.
- There are **no restrictions** on access or use.
- Statisticians have implemented hundreds of **specialised statistical procedures** for a wide variety of applications as **contributed packages**, which are also freely-available and which integrate directly into R.

Advantages of R

1. It is **completely free** and will always be so, since it is issued under the GNU Public License;

2. It is **freely-available over the internet**
   - R Project home page `http://www.r-project.org/`
   - software download `http://cran.r-project.org/`

3. It runs on almost **all operating systems** Unix© and derivatives including Darwin, Mac OS X, Linux, FreeBSD, and Solaris; most flavours of Microsoft Windows; etc.;

4. It is the product of **international collaboration** between **top computational statisticians** and computer language designers;

5. It allows **statistical analysis** and **visualisation** of **unlimited sophistication** with many alternative methods of analysis;

Advantages of R (2)

6. It can work on **objects of unlimited size and complexity**;

7. It can **exchange data** in MS-Excel, text, fixed and delineated formats (e.g. CSV), so that existing datasets are easily imported, and results computed in R are easily exported;

8. It is supported by comprehensive **technical documentation** and user-contributed tutorials. There are also several good textbooks on statistical methods that use R for illustration;

9. Every computational step is **recorded**, and this history can be saved for later use or documentation;

10. It stimulates **critical thinking** about problem-solving rather than a "push the button" mentality.

Advantages of R (3)

11. It is **fully programmable**, with its own sophisticated computer language, named **S**;

12. Repetitive procedures can easily be automated by user-written **scripts** or **functions**;

13. All **source code** is published, so you can see the exact algorithms being used; also, expert statisticians can make sure the code is correct.

Disadvantages of R

"Every disadvantage has its advantage" – Johann Cruiff, Dutch footballer

1. The default Windows and Mac OS X **graphical user interface** (GUI) is limited to simple system interaction and **does not include statistical procedures**. The user must **type commands** to enter data, do analyses, and plot graphs.

   But . . . this has the **advantage** that you have complete control over the system.

   Note: The `Rcmdr` add-on package provides a reasonable GUI for common tasks.

2. The user must decide on the sequence of analyses and execute them step-by-step. However, it is easy to create **scripts** with all the steps in an analysis, and run the script from the command line or menus.

   But . . . this has the **advantage** that you can **save the processing log** of all your analysis steps and their results for inclusion in reports or re-use.

Disadvantages of R (2)

3. The user must learn a **new way of thinking about data**, as **objects** each with its **class**, which in turn supports a set of **methods**.

   But . . . this has the **advantage** that you can only operate on an object according to methods that make sense for it.

4. The user must learn the **S language**, both for commands and the notation used to specify statistical models. However, this allows the user to specify models using a compact and consistent notation.

Commentary

Now we begin to learn R. For the purposes of this course/module we only need to learn a limited part of what R can offer; perhaps after the course/module you will be motivated to learn more.

R is a very complex program with unlimited possibilities. The best way to learn it is step-by-step, from the basics to the more complex. In this first lesson we will not do any **geo-**statistics; the computer exercise will illustrate the **basic operation** of R.

In later lessons we will examine some geostatistical functions.

Exercise

At this point you should complete **Exercise 1: Introduction to R** which is provided on the module CD.

This should take several hours.

1. R basics
2. Reading and examining a data set
3. Exploratory graphics
4. Descriptive statistics

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 : *Suppose we are studying the distribution of the different tree species in a forest reserve. Are all the trees in this forest reserve a* **population** *or* **sample***?* •

A1 : *This is the* **population***; it includes all the objects of interest for the study.* *Return to Q1* •

Q2 : *If we make a transect from one side of the forest to the other, and identify the species of all the trees within 10 m of the centre line, is this a* **population** *or* **sample** *of the trees in the forest reserve?* •

A2 : *This is the* **sample***; it is a defined subset of the population.* *Return to Q2* •

Answers

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.* *Return to Q3* •

Answers

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 : *This is an* **inferential** *statistic.*

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.* *Return to Q4* •

Answers

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:

```
30.8  41.6  46.0  52.4  56.4  56.4 60.0  66.4  72.4  77.36  77.88 141.0
```

The point is that this is a list of values and we can compute descriptive statistics on it. There is no geographical context. *Return to Q5* •

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* **coördinates***.* *Return to Q6* •

Answers

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)* **coördinates** *for each sample.*

Return to Q7 •

Answers

Q8 : *In the figure, how can you determine the coördinates of each sample point?* •

A8 : *We can estimate them from the overprinted* **grid**. *Return to Q8* •

Q9 : *What are the coördinates of the sample point displayed as a red symbol?* •

A9 : *2 km E, 3 km N (right on a grid intersection)* *Return to Q9* •

Answers

Q10 : *What is the* **origin** *of this coördinate system?* •

A10 : *(0,0) at the lower-left corner of the study area.* *Return to Q10* •

Q11 : *How could these coördinates be related to some common system such as UTM?* •

A11 : *If we can find out the UTM coördinates of the origin, i.e. the (0,0) of the local system, we can add this to the local coördinates to get the UTM coördindate. Then, if the local coördinate system is already North oriented, we just add this UTM origin to all the local coördinates; if not, we need the UTM coördinate of one other point, and then we can apply a transformation.* *Return to Q11* •

Answers

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.* *Return to Q12* •

Q13 : *In this case, what are the* **coördinates** *and what are the* **attributes***?* •

A13 : *The* **row** *and* **column** *in the image is a coördinate; 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.* *Return to Q13* •

Answers

Q14 : *Suppose now the images has been geo-referenced. What are now the coördinates?* •

A14 : *Whichever coördinate system that was used for geo-referencing.* *Return to Q14* •

Answers

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.*
Return to Q15 •

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

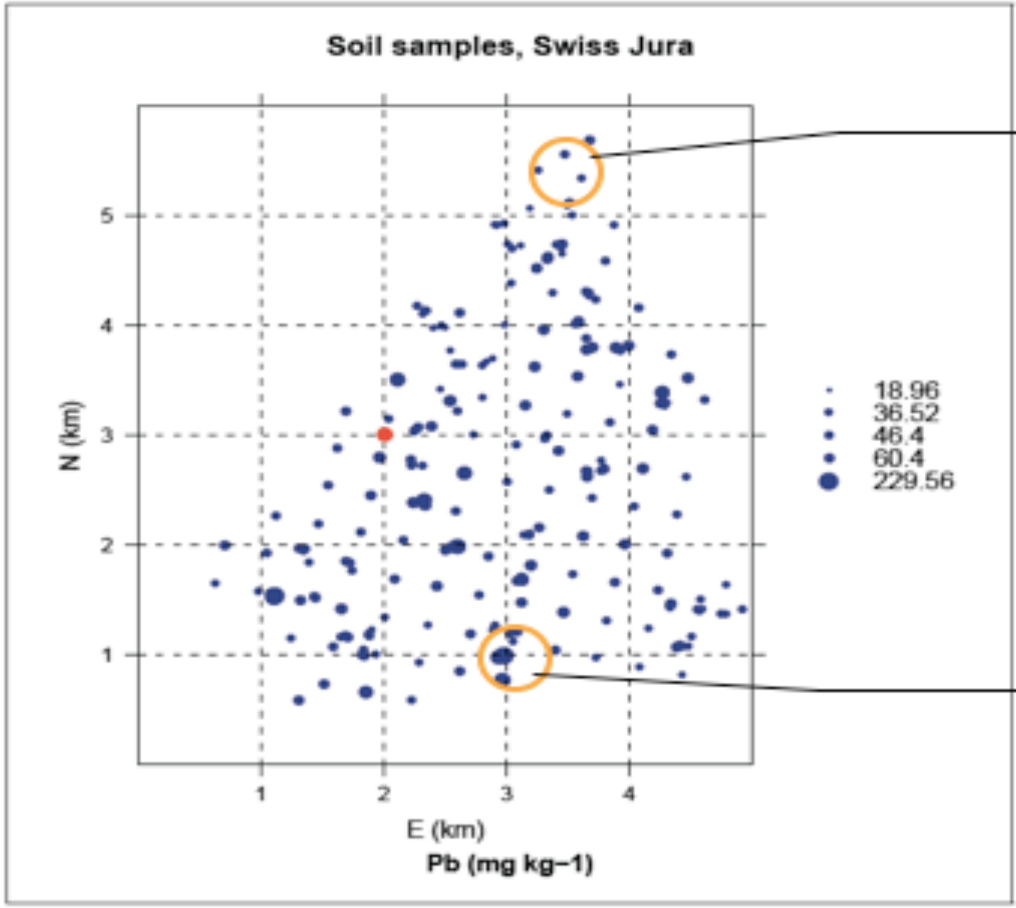
A16 : *Yes; for example at the top of the map near* `(3.5, 5.5)`*; see the figure on the next page.* *Return to Q16* •

Q17 : *What is the approximate radius the clusters?* •

A17 : *They are not so big, approximately 0.5 km radius.* *Return to Q17* •

Clustering of attribute values in geographic space

Post-plot of Pb values, Swiss Jura



(E,N) = (3.5, 5.5)

Small circles= low values

(E,N) = (3, 1)

"hot spot" (big circles= high values)

study area

Answers

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

A18 : *Dimension 1: Petal width in mm; Dimension 2: Petal length in cm.* *Return to Q18* •

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.* *Return to Q19* •

Answers

Q20 : *What are the three dimensions of this feature space?* •

A20 : *Petal width, petal length, sepal length* *Return to Q20* •

(A sepal is a small leaf-like structure which lies directly under and supports the flower petal).

Q21 : *Does there appear to be a correlation between the three dimensions for this set of observations?* •

A21 : *Yes, in general all three attributes get bigger together; however there is considerable spread.* *Return to Q21* •

Answers

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

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

Return to Q22 •

Supplementary information: Learning resources

1. Geostatistics textbooks

2. Useful web pages

3. Resources for learning R

Resources for learning Geostatistics

This section is included for your reference in case you want to go beyond the material presented in this course.

There are **textbooks** for almost every mathematical level and application area; we list some of the ones we find most useful.

Web pages can be useful resources, but are no substitute for a carefully-written text.

Geostatistics texts: Mathematical

- Diggle, P. J. and Ribeiro Jr, P. J., 2007. *Model-based geostatistics*. Springer.
- Chilès, J.-P. and Delfiner, P., 1999. *Geostatistics: modeling spatial uncertainty*. Wiley series in probability and statistics. John Wiley & Sons, New York.
- Christakos, G., 2000. *Modern spatiotemporal geostatistics*. Oxford University Press, New York.
- Cressie, N., 1993. *Statistics for spatial data*. John Wiley & Sons, New York.
- Ripley, B.D., 1981. *Spatial statistics*. John Wiley & Sons, New York.

Geostatistics texts: In the context of a particular application field

- Davis, J.C., 2002. *Statistics and data analysis in geology*. John Wiley & Sons, New York.

- Fotheringham, A.S., Brunsdon, C. and Charlton, M., 2000. *Quantitative geography : perspectives on spatial data analysis*. Sage Publications, London ; Thousand Oaks, Calif.

- Stein, A., Meer, F.v.d. and Gorte, B.G.F. (Editors), 1999. *Spatial statistics for remote sensing*. Kluwer Academic, Dordrecht.

- Kitanidis, P.K., 1997. *Introduction to geostatistics : applications to hydrogeology*. Cambridge University Press, Cambridge, England.

Geostatistics texts: Application-oriented but mathematical

- Webster, R., and Oliver, M. A., 2001. *Geostatistics for environmental scientists*. Wiley & Sons, Chichester.
- Goovaerts, P., 1997. *Geostatistics for natural resources evaluation*. Oxford University Press, Oxford and New York.
- Isaaks, E.H. and Srivastava, R.M., 1990. *An introduction to applied geostatistics*. Oxford University Press, New York.

Geostatistics texts: Emphasis on computational methods

- Venables, W.N. & Ripley, B.D., 2002. *Modern applied statistics with S*, 4th edition. Springer-Verlag, New York.

  For the serious R user. This covers a wide variety of modern statistical methods, including spatial statistics.

- Deutsch, C. V., & Journel, A. G., 1992. *GSLIB: Geostatistical software library and user's guide*. Oxford University Press, Oxford.

Web pages

- Statistics
  - **Electronic Statistics Textbook** from StatSoft:
    `http://www.statsoft.com/textbook/stathome.html`
  - **NIST/SEMATECH e-Handbook of Statistical Methods**:
    `http://www.itl.nist.gov/div898/handbook/`

- Geostatistics
  - **R task view: Analysis of Spatial Data**
    `http://cran.r-project.org/web/views/Spatial.html`
  - **Geostatistical analysis tutor**, mainly aimed at mining applications:
    `http://www.uncert.com/tutor/`

Resources for learning R

These are not necessary for this module; they are listed for your reference in case you decide to continue with R.

General introductions

- Venables, W. N. ; Smith, D. M. ; R Development Core Team, 2013. *An Introduction to R* (Notes on R: A Programming Environment for Data Analysis and Graphics), Version 3.0.2 (2013-09-25).

  `http://www.cran.r-project.org`; also included with R distribution

  The standard introduction; this is updated with each release and included in the standard download, placed in the `doc/manual` directory of the installation.

- Rossiter, D.G., 2012. *Introduction to the R Project for Statistical Computing for use at ITC*. International Institute for Geo-information Science & Earth Observation (ITC), Enschede (NL), 136 pp.

  `http://www.itc.nl/personal/rossiter/teach/R/RIntro_ITC.pdf`

  My introduction; a bit less technical and somewhat slower-paced than the standard; with a long section on the S language.

Textbooks using R

More and more texts are using R code to illustrate their statistical analyses.

- Dalgaard, P. 2002. *Introductory Statistics with R*. Springer Verlag.

  This is a clearly-written introduction to statistics, using R in all examples.

- Venables, W. N. & Ripley, B. D. 2002. *Modern applied statistics with S*. New York: Springer-Verlag, 4th edition; `http://www.stats.ox.ac.uk/pub/MASS4/`

  Presents a wide variety of up-to-date statistical methods (including spatial statistics) with algorithms coded in S; includes an introduction to R, R programming, and R graphics.

- Fox, J. and Weisberg, S. 2011. *An R companion to applied regression, 2nd edition*. Thousand Oaks, CA: Sage.

  A social scientists explains how to use R for regression analysis, including advanced techniques; this is a companion to his text: Fox, J. 2008. *Applied regression, linear models, and related methods, 2nd edition*. Los Angeles: Sage; **Highly-recommended**

  This is accompanied by an on-line Appendix `http://cran.r-project.org/doc/contrib/Fox-Companion/appendix.html` with excellent tutorials, with R code, on advanced topics such as non-parametric regression, non-linear regression, and mixed-effects models.

Technical Notes using R

D G Rossiter has written a number of technical notes showing how to accomplish some statistical tasks with R; the full list is at

`http://www.itc.nl/personal/rossiter/pubs/list.html#pubs_m_R`.

These include general data analysis, logistic regression, confusion matrices, co-kriging, partioning transects, and fitting rational functions.