

## Introduction

- What is geostatistics?

Geostatistics is concerned with estimation and prediction for spatially continuous phenomena, using data obtained at a limited number of spatial locations. Here, with phenomena we mean the distribution in a two- or three-dimensional space of one or more random variables called *regionalized variables*. The phenomenon for which the regionalized variables are referred to it is called *regionalization*. For example, the distribution of mineral ore grades in the three-dimensional space. Or the distribution of ozone, etc.

- History: The term *geostatistics* was coined by Georges Matheron (1962). Matheron and his colleagues (at Fontainebleau, France) used this term in prediction for problems in the mining industry. The prefix “geo” concerns data related to earth.
- Today, geostatistical methods are applied in many areas beyond mining such as soil science, epidemiology, ecology, forestry, meteorology, astronomy, corps science, environmental sciences, and in general where data are collected at geographical locations (spatial locations).
- The spatial locations throughout the course will be denoted with  $s_1, s_2, \dots, s_n$  and the spatial data collected at these locations will be denoted with  $z(s_1), z(s_2), \dots, z(s_n)$ . Spatial locations are determined by their coordinates  $(x, y)$ . We will mainly focus in two-dimensional space data.
- Very important in the analysis of spatial data is the distance between the data points. We will use mostly Euclidean distances. Suppose data point  $s_i$  has coordinates  $(x_i, y_i)$  and data point  $s_j$  has coordinates  $(x_j, y_j)$ . The Euclidean distance between points  $s_i$  and  $s_j$  is given by:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

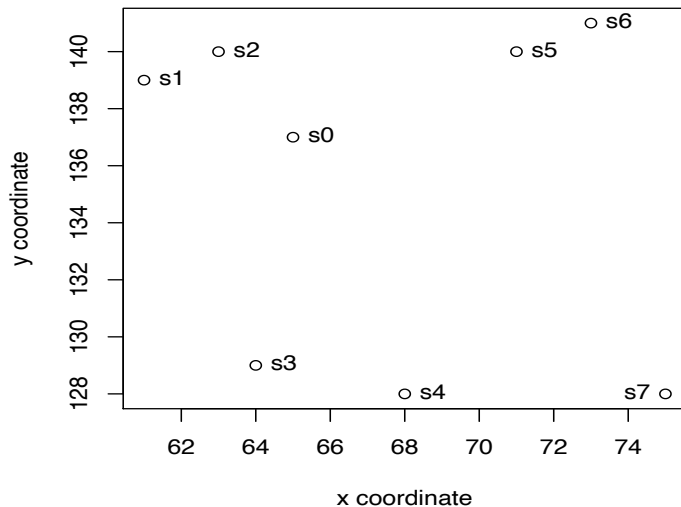
Other forms of distances can be used (great-circle distance, azimuth distance, travel distance from point to point, time needed to get from point to point, etc.).

- The problem:

- Present and explain the distribution of the random function

$$Z(s) : s \in D$$

- Predict the value of the function  $Z(s)$  at spatial location  $s_0$  (in other words the value  $z(s_0)$ ) using the observed data vector  $z(s_1), z(s_2), \dots, z(s_n)$  (see figure below).



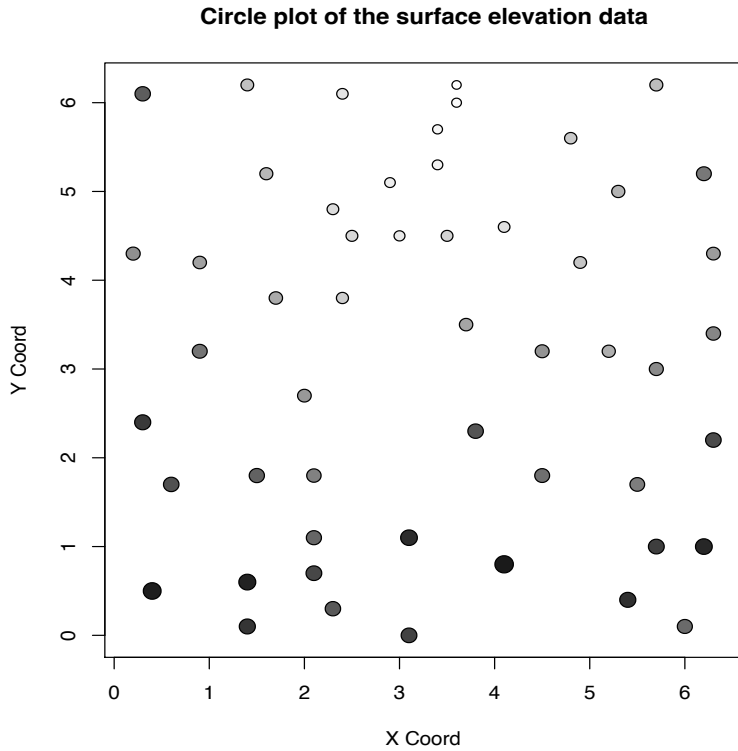
- Environmental protection agencies set maximum thresholds for harmful substances in the soil, atmosphere, and water. Therefore given the data we should also like to know the probabilities that the true values exceed these thresholds at unsampled locations.

- A random function  $Z(s)$  can be seen as a set of random variables  $Z(s_i)$  defined at each point  $s_i$  of the random field  $D : Z(s) = Z(s_i), \forall s_i \in D$ . These random variables are correlated and this correlation depends on the vector  $h$  that separates two points  $s$  and  $s + h$ , the direction (south-north, east-west, etc.), but also on the nature of the variables considered. The data can be thought as a realization of the function  $Z(s)$  with  $s$  varying continuously throughout the region  $D$ .
- Geostatistical theory is based on the assumption that the variability of regionalized variables follows a specific pattern. For example, the ozone level  $z(s)$  at location  $s$  is auto-correlated with the ozone level  $z(s+h)$  at location  $s+h$ . Intuitively, locations close to one another tend to have similar values, while locations farther apart differ more on average. Geostatistics quantifies this intuitive fact and uses it to make predictions.

## Motivating examples

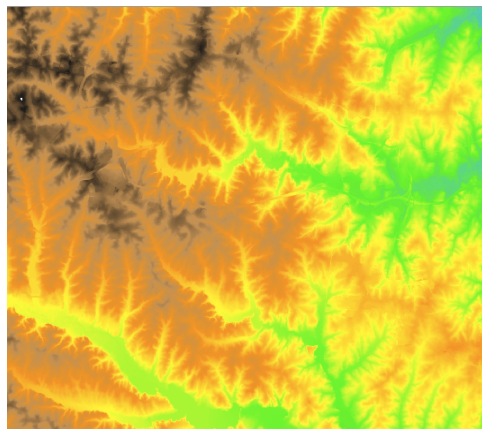
Example 1:

Surface elevations. For these data the coordinates  $x, y$  and elevation was recorded at 52 locations as shown below.



The circles have centers at the sampling locations given by the coordinates and the radius of each circle is determined by a linear transformation of the elevations. Also observed that the circles are filled with grey shades.

The objective in analyzing these data is to construct a continuous elevation map resulting in a raster map. The raster map below shows the elevation of an area in south-west Wake county in North Carolina, USA.

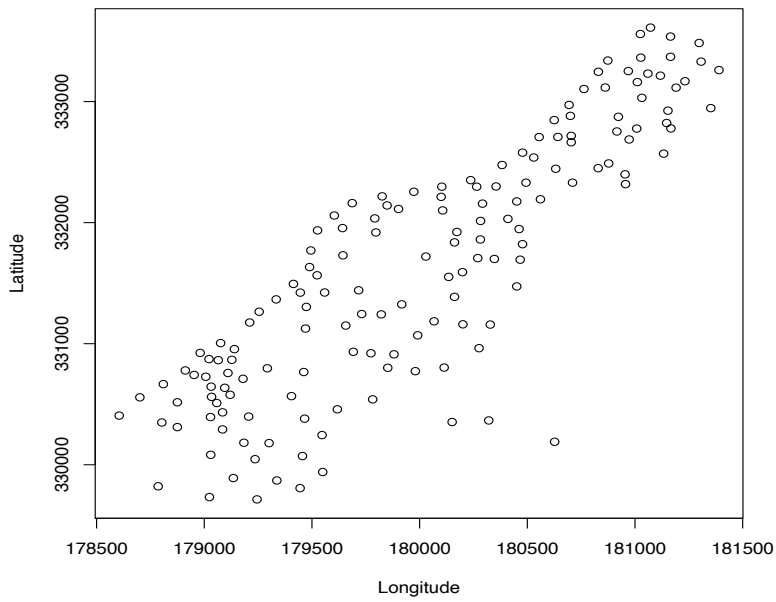


Example 2:

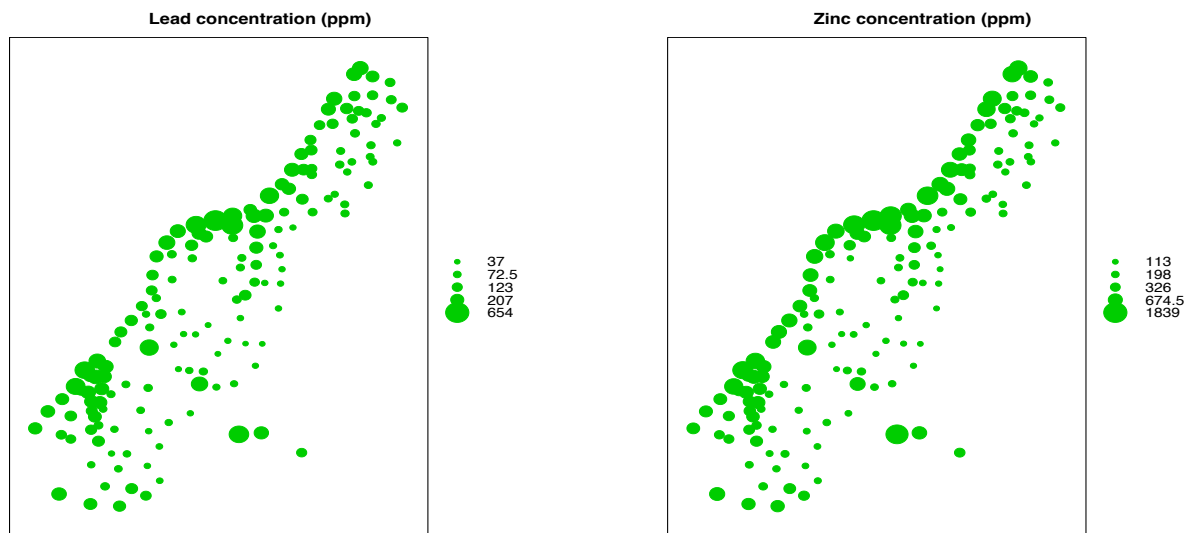
The data below were collected from the flooded banks of the Meuse river (in Dutch Maas river).



The data points:



## Concentration of lead and zinc:



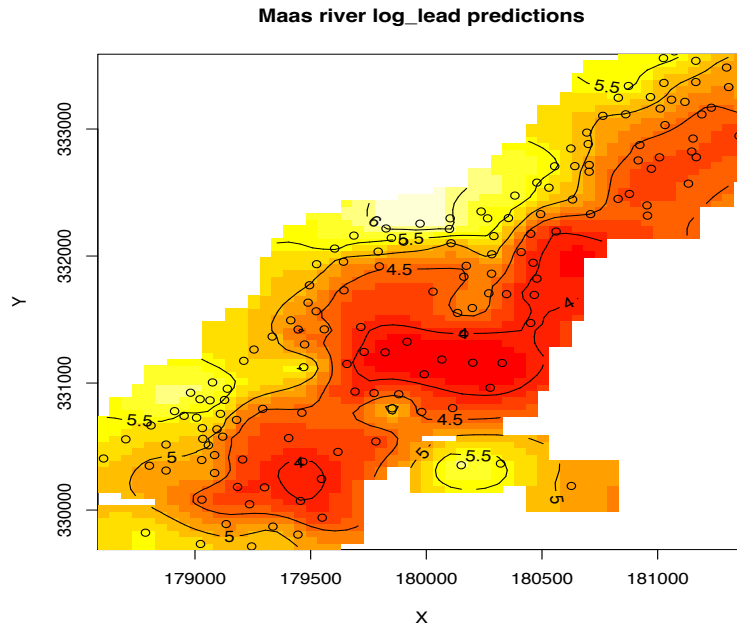
According to the United States Environmental Protection Agency (US EPA) the level of risk for surface soil based on lead concentration in *ppm* is given on the table below:

Mean concentration (ppm)	Level of risk
Below 150	Lead-free
Between 150-400	Lead-safe
Above 400	Significant environmental lead hazard

## Construction of a grid:



## Construction of a raster map:



### Few R commands:

Read the Maas data:

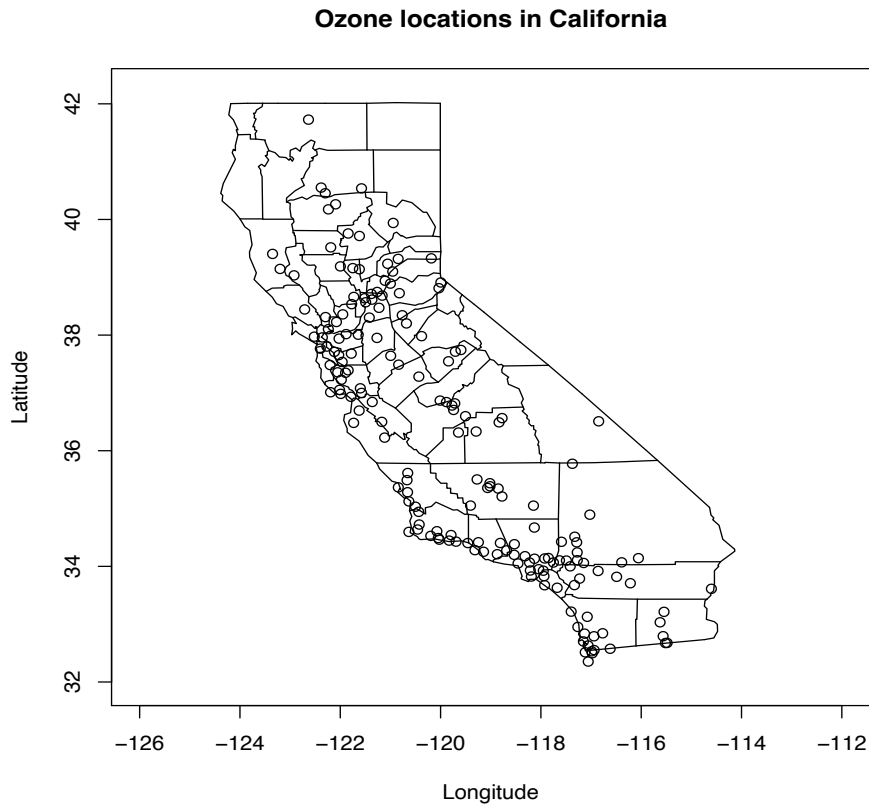
```
> a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics_c173_c273/
  soil.txt", header=TRUE)
> class(a)

> library(geoR)
> b <- as.geodata(a)
> class(b)
> points(b)
> plot(b)

> library(gstat)
> coordinates(a) <- ~x+y
> class(a)
> bubble(a, "lead", main="Lead concentration (ppm)")
> bubble(a, "zinc", main="Zinc concentration (ppm)")
```

## Another example:

The map below shows 175 ozone stations (08 August 2005 data):



## Try the following commands:

```
a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics_c173_c273/o3.txt",
  header=TRUE)

library(geoR)
library(gstat)

library(maps)

plot(a$lon,a$lat, xlim=c(-126,-112), ylim=c(32,42.2), xlab="Longitude",
  ylab="Latitude", main="Ozone locations in California")

map("county", "ca", add=TRUE)

#What do the following commands do?
aa <- as.data.frame(cbind(a$lon,a$lat,a$o3))
bb <- as.geodata(aa)
class(bb)
points(bb)

#How about these?
coordinates(a) <- ~ lon+lat
class(a)
bubble(a, "o3", xlab="Longitude", ylab="Latitude", maxsize=1.3, key.entries=0.02*(1:6))
```

## An example using the maps package

Data on ozone and other pollutants are collected on a regular basis. The data set for this example concerns 175 locations for ozone (ppm) in California on 08 August 2005. You can read more about smog-causing pollutants at

<http://www.nytimes.com/2010/01/08/science/earth/08smog.html?th&emc=th>

The data can be accessed here:

```
a <- read.table("http://www.stat.ucla.edu/~nchristo/
  statistics_c173_c273/o3.txt", header=TRUE)
```

The package maps in R can be loaded as follows:

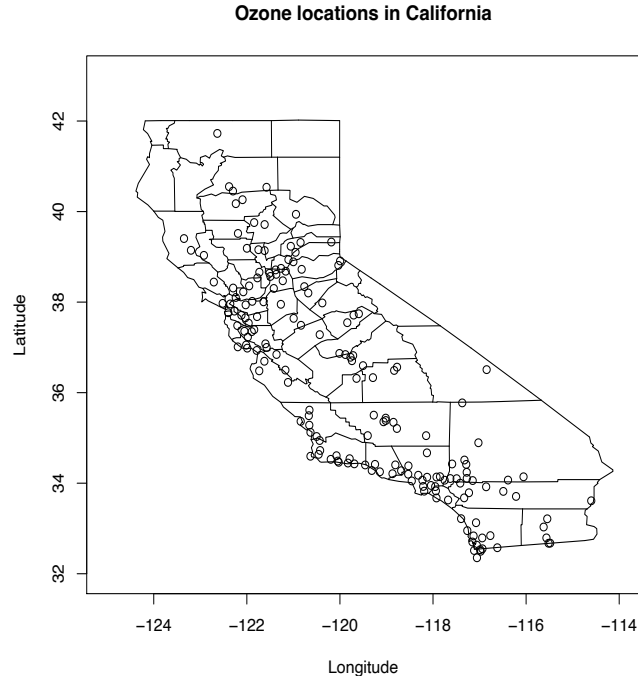
```
library(maps)
```

We can display the data points and the map using the following commands:

```
plot(a$lon,a$lat, xlim=c(-125,-114),ylim=c(32,43), xlab="Longitude",
  ylab="Latitude", main="Ozone locations in California")
```

```
map("county", "ca",add=TRUE)
```

Here is the plot:





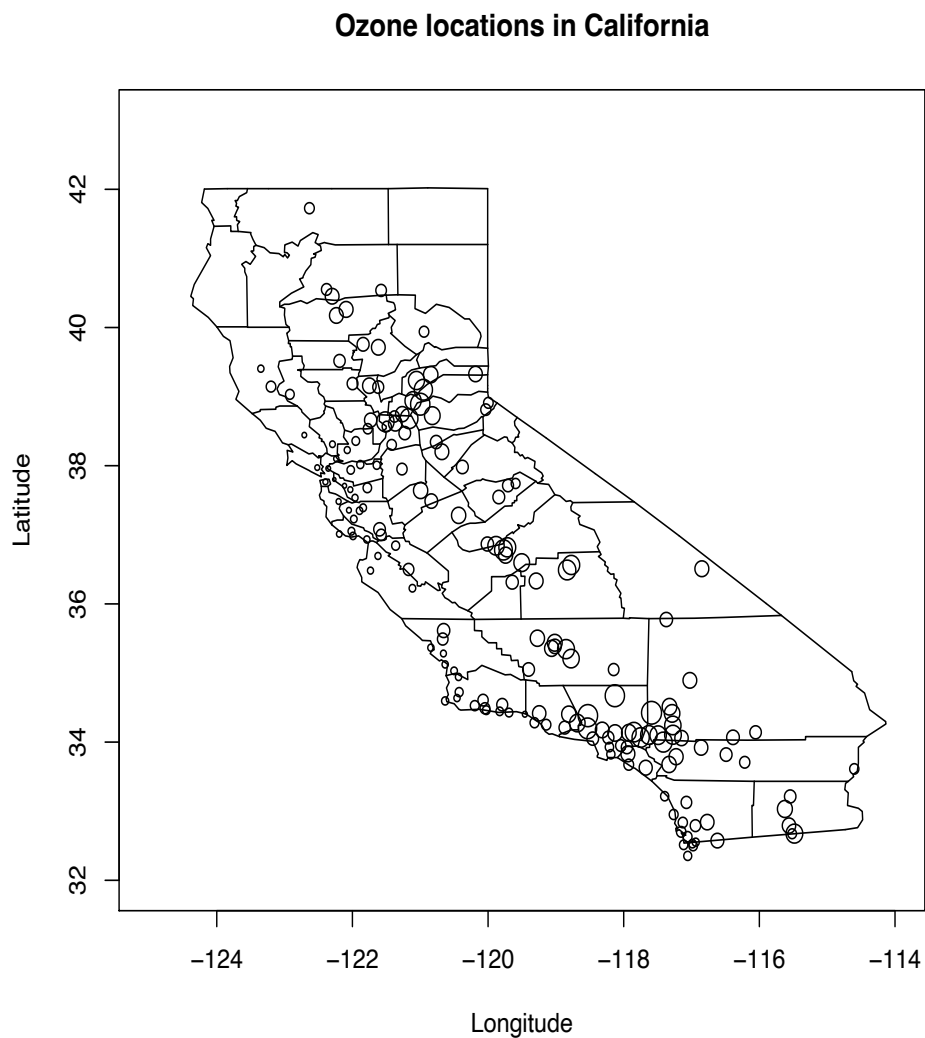
We can also plot the points relative to their value (larger values will be displayed with larger circles). Here are the commands:

```
plot(a$lon,a$lat, xlim=c(-125,-114),ylim=c(32,43), xlab="Longitude",  
ylab="Latitude", main="Ozone locations in California", "n")
```

```
map("county", "ca",add=TRUE)
```


```
points(a$x, a$y, cex=a$o3/0.06)
```

Here is the plot:



The following chart illustrates the health-related interpretation of the Ozone data in terms of the particulate (particles per million, ppm) recordings, according to the National Oceanic and Atmospheric Administration's (NOAA) Air Quality Index (AQI).

<http://www.noaa.gov/>

 <h2 style="display: inline;">Air Quality Index for Ozone</h2>		
Concentration Range (ppm)	Air Quality Description	Cautionary Statements for Ozone
0.00 – 0.060 ppm	Good	No health impacts are expected
0.061 – 0.075 ppm	Moderate	Unusually sensitive people should consider limited prolonged outdoor exertion
0.076 – 0.104 ppm	Unhealthy for Sensitive Groups	Active children and adults, and people with respiratory conditions (e.g., asthma) should <u>limit prolonged outdoor exertion</u>
0.105 – 0.115 ppm	Unhealthy	Active children and adults, and people with respiratory conditions (e.g., asthma) should <u>avoid prolonged outdoor exertion</u> . Everyone else, especially children and elderly, should limit prolonged outdoor exertion
0.116 – 0.374 ppm	Very Unhealthy	Active children and adults, and people with respiratory conditions (e.g., asthma) should <u>avoid all outdoor exertion</u> . Everyone else, especially children and elderly, should limit outdoor exertion

What is next?

Try to match data location with the Air Quality Index:

```
AQI_colors <- c("green", "yellow", "orange", "dark orange", "red")
AQI_levels <- cut(a$o3, c(0, 0.06, 0.075, 0.104, 0.115, 0.374))

as.numeric(AQI_levels)

plot(a$lon,a$lat, xlim=c(-125,-114),ylim=c(32,43), xlab="Longitude",
      ylab="Latitude", main="Ozone locations in California", "n")

map("county", "ca",add=TRUE)
points(a$lon,a$lat, cex=a$o3/mean(a$o3),
       col=AQI_colors[as.numeric(AQI_levels)], pch=19)
```

Here is the plot:

