University of California, Los Angeles
Department of Statistics

Statistics C173/C273 **Instructor: Nicolas Christou**

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 through out the course will be denoted with $s_1, s_2, \cdots, s_n$ and the spatial data collected at these locations will be denoted with $z(s_1), z(s_2), \cdots, z(s_n)$. 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:

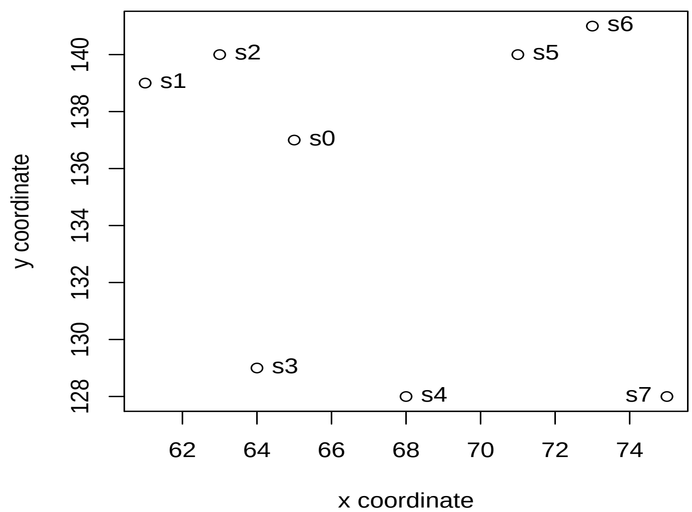
  $$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

  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$z(s_1), z(s_2), \cdots, z(s_n)$ (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.

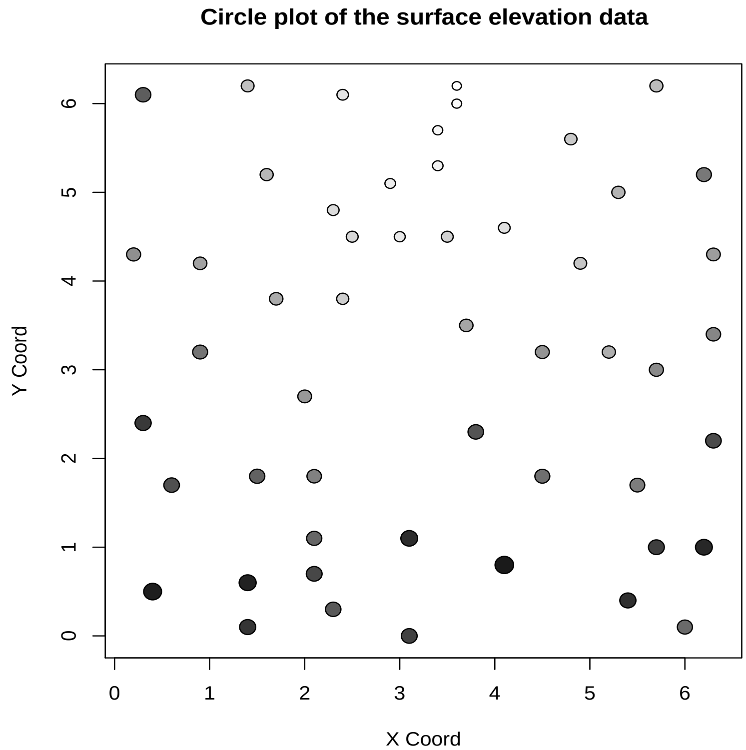
- A random function $Z(s)$ can be seen as a set of random variables $Z(s_i)$ defined at each point s_i of the random field $D : Z(s) = Z(s_i), \forall s_i \in D$. These random variables are correlated and this correlation depends on the vector h that separates two points s and $s + h$, the direction (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

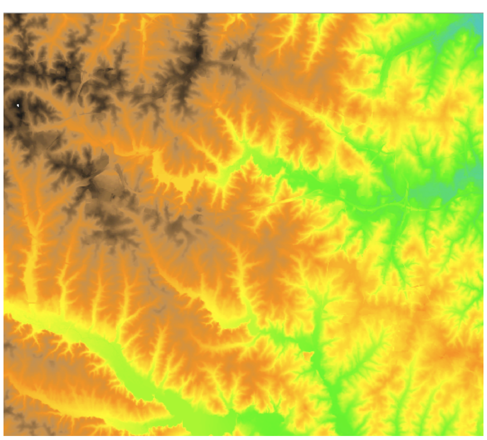
Example 1:

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



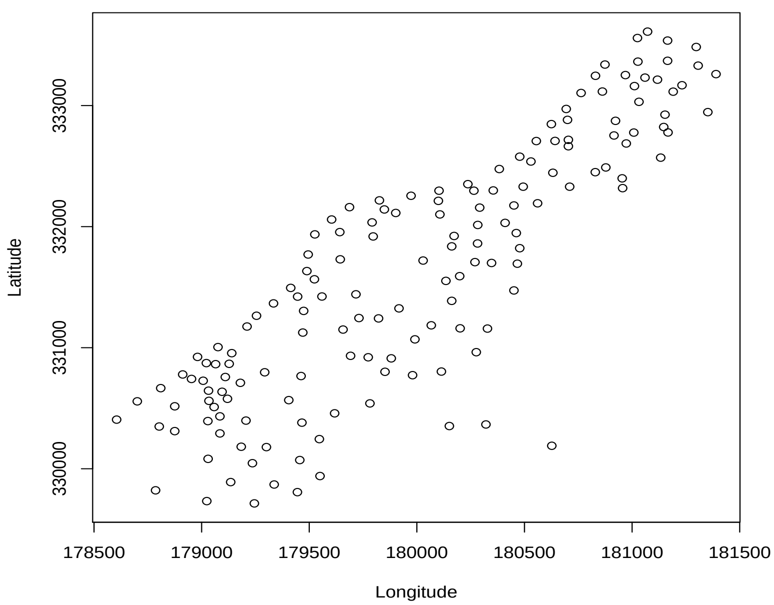
The circles have centers at the sampling locations given by the coordinates and the radius of each circle is determined by a linear transformation of the elevations. Also observed that the circles are filled with grey shades.

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.

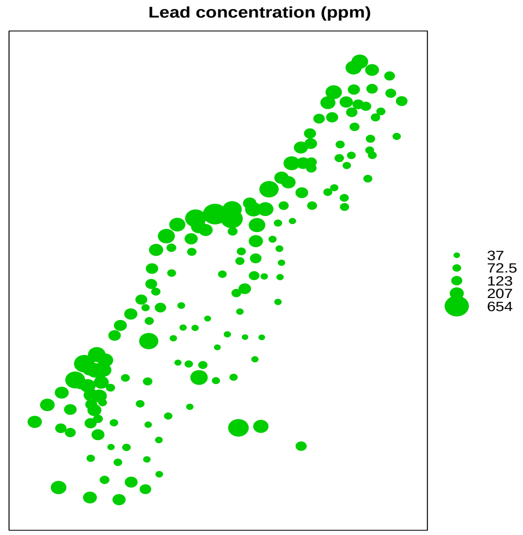


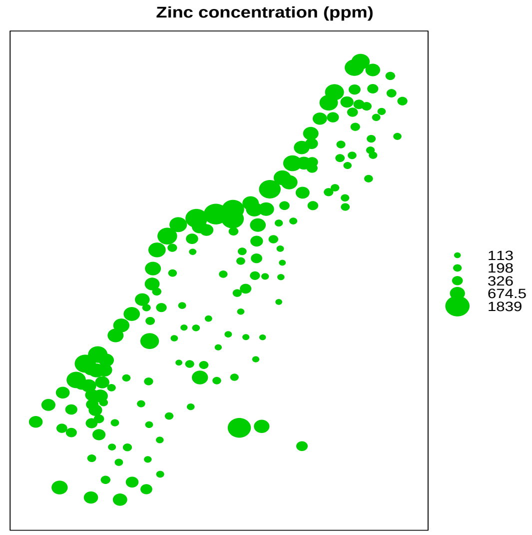
Example 2:
The data below were collected from the flooded banks of the Meuse river (in Dutch Maas river).

The data points:



Concentration of lead and zinc:

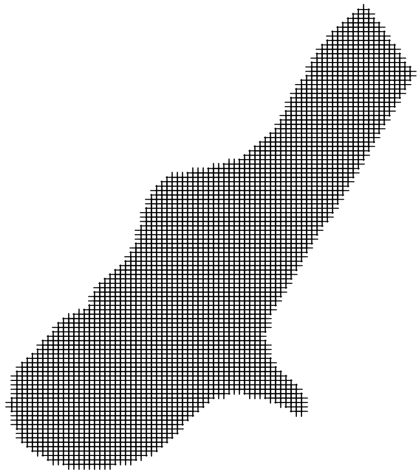




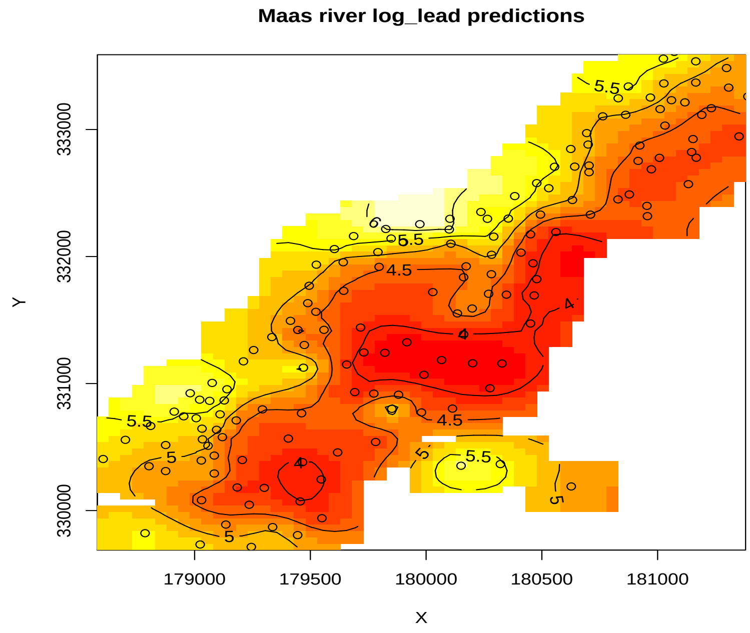
According to the Unites States Environmental Protection Agency (US EPA) the level of risk for surface soil based on lead concentration in *ppm* is given on the table below:

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:



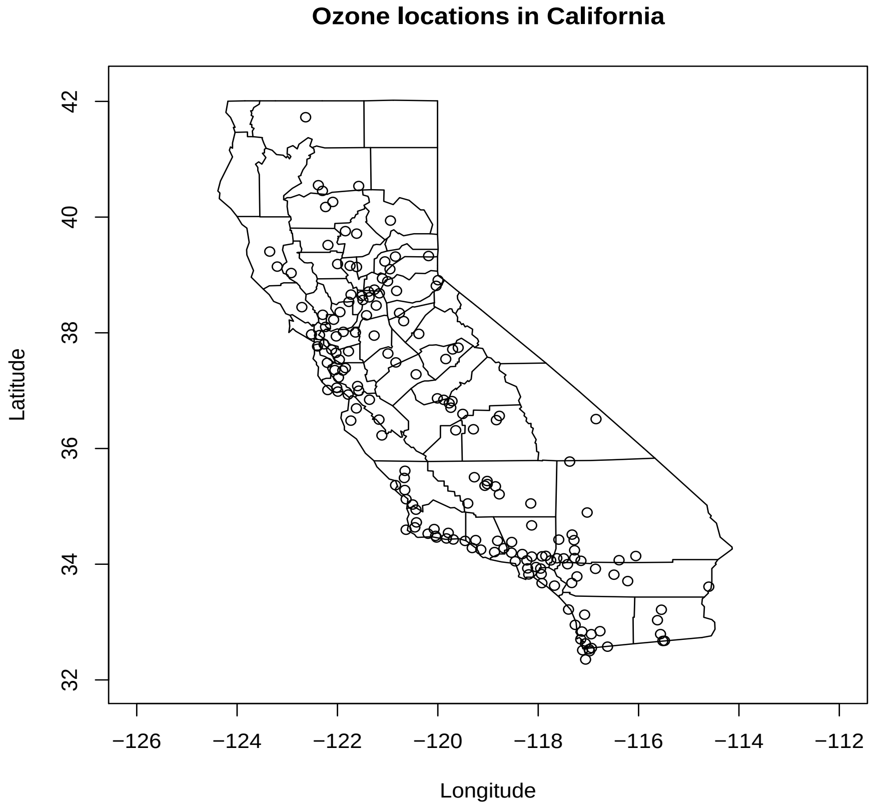
Few R commands:

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:

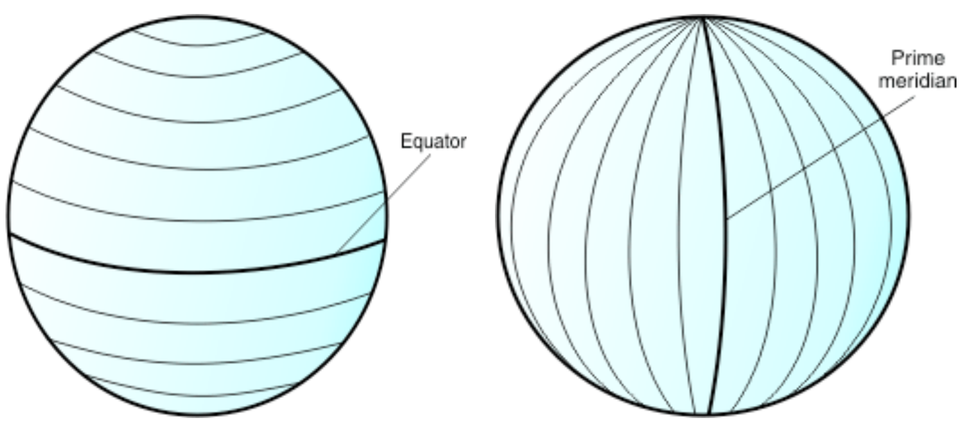
```
a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics_c173_c273/o3.txt",
     header=T)

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))
```

Coordinate systems

Geographic coordinate system (latitude-longitude:)

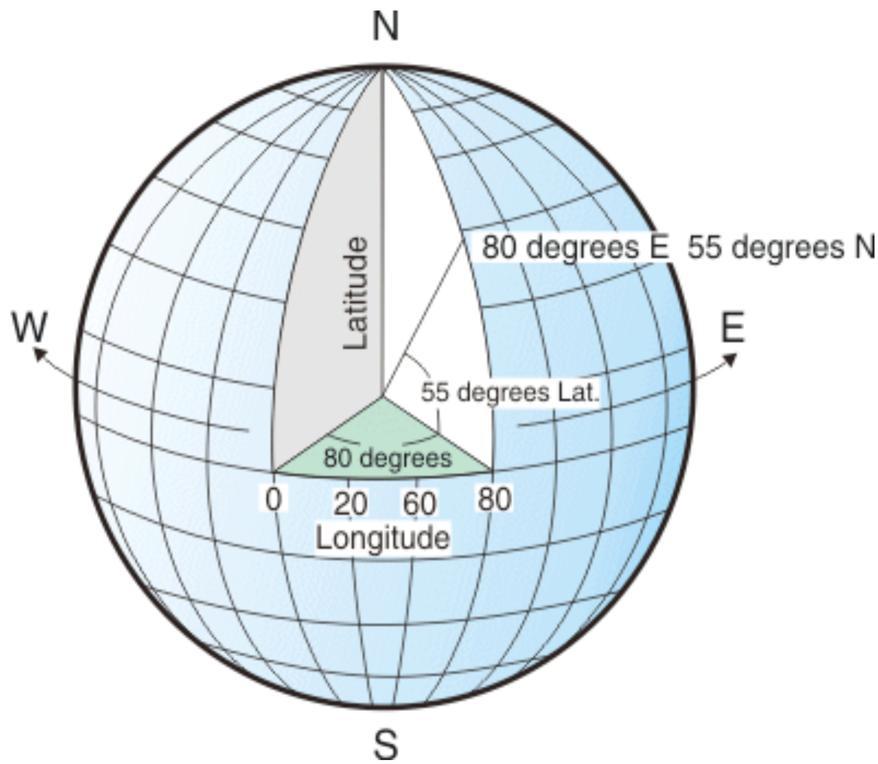
This is the most commonly used system. This coordinate system consists of **parallels** and **meridians** (see figure below). The parallels are parallel to the equator (and to each other) and they circle the globe from east to west. At the equator we assign the value zero. The angular distance from the equator to either pole is equal to one-fourth of a circle, therefore 90 degrees. Angular measurements north or south of the equator are called it latitude. For example, the location of Westwood Blvd. and Le Conte Ave. is 34 degrees 3 minutes and 49.4 seconds north of the equator (denoted with $N\ 34^o\ 3'\ 49.4''$). Meridians are drawn from south to north pole. The starting point passes through Greenwich, England (prime meridian). This prime meridian gets the value zero and angular measurements east and west are called *longitude*. For example, the location of Westwood Blvd. and Le Conte Ave. is 118 degrees 26 minutes and 43.5 seconds west of Greenwich (denoted with $W\ 118^o\ 26'\ 43.5''$). In decimal form the above location is defined as (34.063709, -118.445413). How do we convert from sexagesimal degrees to decimal degrees? Keep the degrees value, and then divide the minutes by 60 and add to this result the seconds divided by 3600. For our example: $N\ 34^o\ 3'\ 49.4''$ is converted to $34+\frac{3}{60}+\frac{49.4}{3600}=34.0637$. One degree of latitude approximately equals to 69 miles ($\sim$ 111 km), 1 minute is approximately 1.15 miles, and 1 second is equal to 0.02 miles. However the distance of a degree of longitude varies because as we move to the poles the distance between meridians gets smaller. But at the equator one degree of longitude is approximately equal to 69 miles (same as with the latitude degree).

Latitude and longitude lines:



The figure below shows how to measure latitude and longitude.

Measuring latitude and longitude:



Some numbers ...

- Earth radius distance: 6357 km to 6378 km $(3950 - 3963)$ miles.
- Length of equator, $L = 24901.5$ miles (40075.0 km).
- Length of parallel at latitude θ is equal to $cos(\theta) \times L$.
- At the equator, 1^o of latitude is 68.7 miles.
- At the poles, 1^o of latitude is 69.4 miles.
- At the equator, 1^o of longitude is 69.172 miles.
- Above or below the equator 1^o of longitude is equal to $cos(\theta) \times 69.172$.
- The great circle distance D between two points A and B on the earth's surface is computed as follows:

  $$D = 69.172 \times cos^{-1}\left[sin(a) \times sin(b) + cos(a) \times cos(b) \times cos(|d|)\right]$$

  where, a, b are the latitude of points A, B, and d is the absolute value of the difference between the longitudes of A, B.
- Example:
  Calculate the great circle distance between
  Los Angeles, CA (N 34^o $03'$ $15''$, W 118^o $14'$ $28''$) and
  New York City, NY (N 40^o $45'$ $6''$, W 73^o $59'$ $39''$)
- Using the formula above the distance is 2444.17 miles (3933.51 km).