Week 4 (part 2): Random fields and exploratory data analysis

This week we discuss the first class of models, which is used to describe raster data. Then, we use R for exploratory spatial data analysis. In particular, we consider

- Revision of basic definitions and properties of **random variables**.
- **Weakly stationary time series**.
- Introduction to **random fields**.
- **Positive definiteness**.
- Properties of **positive definite** functions.
- Plotting with **geoR.**
- Basic analysis with **sp.**
- **Estimating trends.**

RANDOM VARIABLES.

A **random variable** is a variable that takes a numerical value for each possible outcome of a statistical experiment. For simplicity we denote a random variable $X(\omega)$ by X.

If X is a random variable, then we cannot predict its value with certainty, but can assign probabilities to events such as $\{X = 1\}$ and $\{X > 2\}$ etc.

Discrete random variables

A random variable X is called **discrete** if all of its possible values can be written down in a list. The probability distribution of a discrete random variable X is a list of the possible values that X can take (put in increasing order), together with the probabilities that X takes each of these possible values.

Expected value of X

The **expected value** of X, denoted $E(X)$, is defined to be

$$E(X) = \sum_x xP(X = x).$$

This is a weighted average of the possible values of X where the weights are the corresponding probabilities. $E(X)$ is a measure of the **centre** of the probability distribution.

Continuous Random Variables

A random variable that can take on a continuum of possible values is called a **continuous** random variable.

We describe the probability properties of a continuous random variable Y by a function $f(y)$ called the **probability density function (pdf)**. Probabilities are given by the relevant area under the probability density function curve.

Expected value of a continuous random variable Y

Suppose that Y is a continuous random variable.
The expected value of Y, denoted $E(Y)$, is defined mathematically as

$$\int_{-\infty}^{\infty} yf(y)dy.$$

Variance of X

Suppose that X is a random variable. The variance of X, denoted $\text{Var}(X)$, is defined as

$$\text{Var}(X) = E\left[(X - E(X))^2\right].$$

The standard deviation is defined to be $\sqrt{\text{Var}(X)}$.

Properties of $E(X)$ and Var(X)

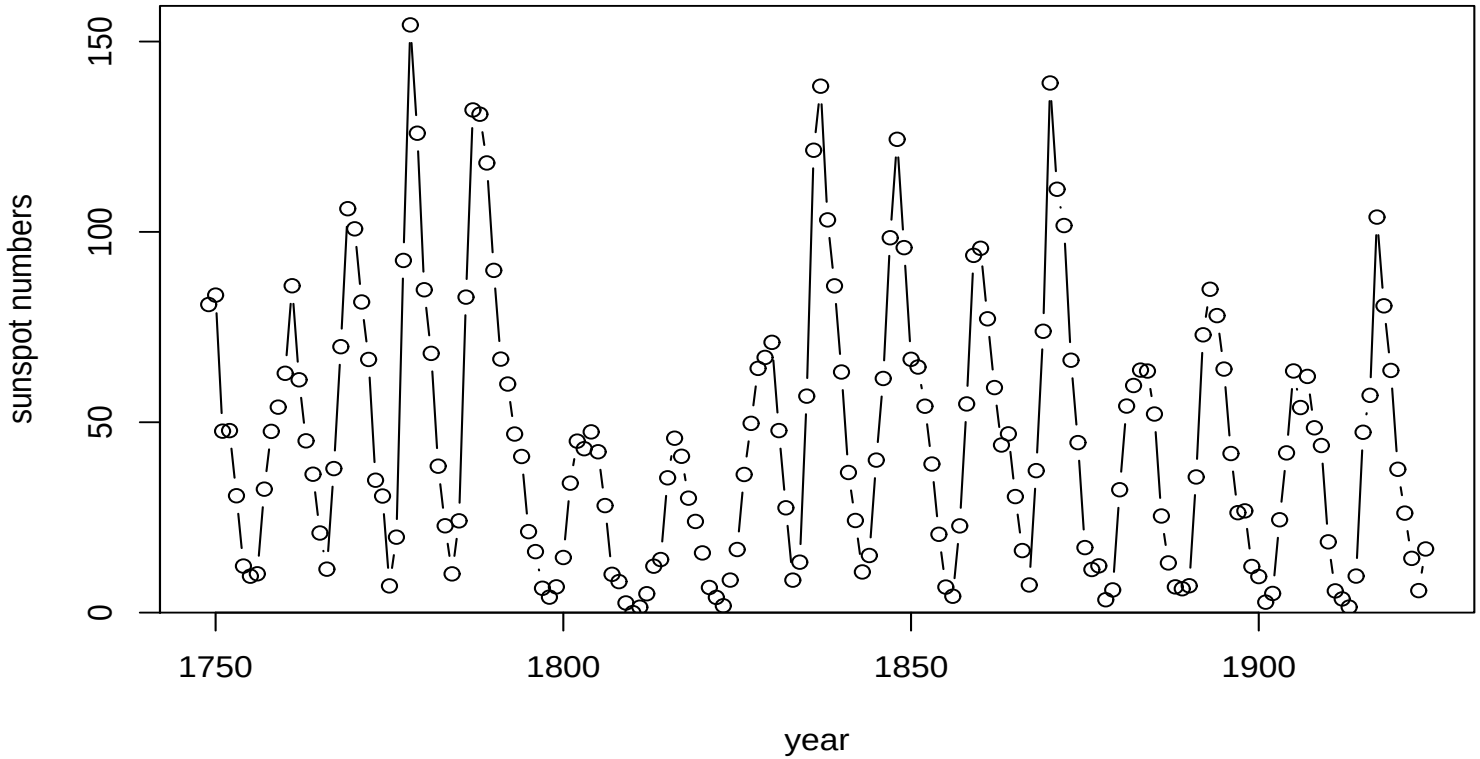
Suppose that X and Y are random variables. Also suppose that a and b are numbers. Then

- $E(a) = a$
- $E(bX) = bE(X)$
- $E(X + Y) = E(X) + E(Y)$
- $\text{Var}(aX + b) = a^2\text{Var}(X)$.

Time Series

Observations of random variables over time typically display dependence. It is this dependence that we model by using time series models.

We use the notation $\{X_t(\omega) : t \in T\}$ (for simplicity X_t) to denote a time series, where T (usually $\mathbb{N}, \mathbb{Z}$ or $\mathbb{R}$) is the index set.



Weakly stationary time series

Definition 1

A time series $\{X_t\}$ is said to be **weakly stationary** *if*

- $E(X_t)$ *does not depend on* t.
- *For every integer* s, $E\left(X_t X_{t-s}\right)$ *does not depend on* t.

For a weakly stationary time series $\{X_t\}$ we define:

- $m = E(X_t)$ (for simplicity we often use $m = 0$)
- The autocovariance function γ by

$$\gamma(s) = E\left(X_t X_{t-s}\right) \quad \text{for all integer } s.$$

- The autocorrelation function ρ by

$$\rho(s) = \frac{\gamma(s)}{\gamma(0)} \quad \text{for all integer } s.$$

Random Fields

A formal definition of a **random field** is:

Definition 2

Let a probability space $(\Omega, \mathcal{F}, P)$ and a parameter set T be given. A random field is a function $X(\mathbf{t}, \omega)$ which, for every fixed $\mathbf{t} \in T$, is a random variable of $\omega \in \Omega$.

For a fixed $\omega \in \Omega$, the function $X(\mathbf{t}, \omega)$ is a non-random function of $\mathbf{t}$. This deterministic function is usually called a **sample path (sample function)** or a **realization.**

For simplicity we denote $X(\mathbf{t}, \omega)$ by X_t.

A time series is a particular case of random fields, when T is a one-dimensional space.

Expectation and Covariance

The **expectation** of a random field equals

$$m(\mathbf{t}) = E\{X_{\mathbf{t}}\}.$$

The **(auto-) covariance function** is defined by

$$C(\mathbf{t},\mathbf{s}) \quad = \quad Cov\{X_{\mathbf{t}}, X_{\mathbf{s}}\} = E\{X_{\mathbf{t}}X_{\mathbf{s}}\} - m(\mathbf{t})m(\mathbf{s}),$$

whereas the *variance* is

$$\sigma^2(\mathbf{t}) = C(\mathbf{t},\mathbf{t}).$$

The **(auto-)correlation function** of a random field equals

$$\rho(\mathbf{t},\mathbf{s}) = Corr\{X_{\mathbf{t}}, X_{\mathbf{s}}\} = \frac{C(\mathbf{t},\mathbf{s})}{\sigma(\mathbf{t})\sigma(\mathbf{s})}.$$

Positive definiteness

Definition 3

Let n be a positive integer, and let $\mathbf{t}_k \in T$ and $c_k \in \mathbb{C}$ (or $\mathbb{R}$) for $k = 1, \ldots, n$. Then a function $B(\cdot, \cdot)$ is positive definite on T if

$$\sum_{k=1}^{n} \sum_{l=1}^{n} c_k \bar{c}_l B(\mathbf{t}_k, \mathbf{t}_l) \geq 0$$

for any $n, \{\mathbf{t}_1, \ldots, \mathbf{t}_n\}$, and $\{c_1, \ldots, c_n\}$ ($\bar{c}_k$ is a complex conjugate of c_k).

The concept of positive definiteness is fundamental in many areas.

Theorem 1

The class of covariance functions coincides with the class of positive definitive functions.

Properties of $\mathcal{P}_T$

Let $\mathcal{P}_T$ be the class of positive functions on T.

(1) $B(t,s) \in \mathcal{P}_T, \alpha \geq 0 \Rightarrow \alpha \cdot B(t,s) \in \mathcal{P}_T$.

(2) $B_1(t,s) \in \mathcal{P}_T, B_2(t,s) \in \mathcal{P}_T \Rightarrow B_1(t,s) + B_2(t,s) \in \mathcal{P}_T$.

(3) $\alpha_1 \geq 0, \ldots, \alpha_n \geq 0; B_1(t,s), \ldots, B_n(t,s) \in \mathcal{P}_T \Rightarrow$
$\sum_{k=1}^{n} \alpha_k B_k(t,s) \in \mathcal{P}_T$.

(4) $B_1(t,s) \in \mathcal{P}_T, B_2(t,s) \in \mathcal{P}_T \Rightarrow B_1(t,s) \cdot B_2(t,s) \in \mathcal{P}_T$.

(5) $B_n(t,s) \in \mathcal{P}_T \Rightarrow \lim_{n\to\infty} B_n(t,s) \in \mathcal{P}_T$.

Spatial Exploratory Data Analysis

Spatial exploratory data analysis starts with the plotting of maps with a measured variable and descriptive statistics.

geoR

The package geoR provides functions for geostatistical data analysis.

To start package and load its simulated data, type:

```
> library(geoR)
> data(s100)
```

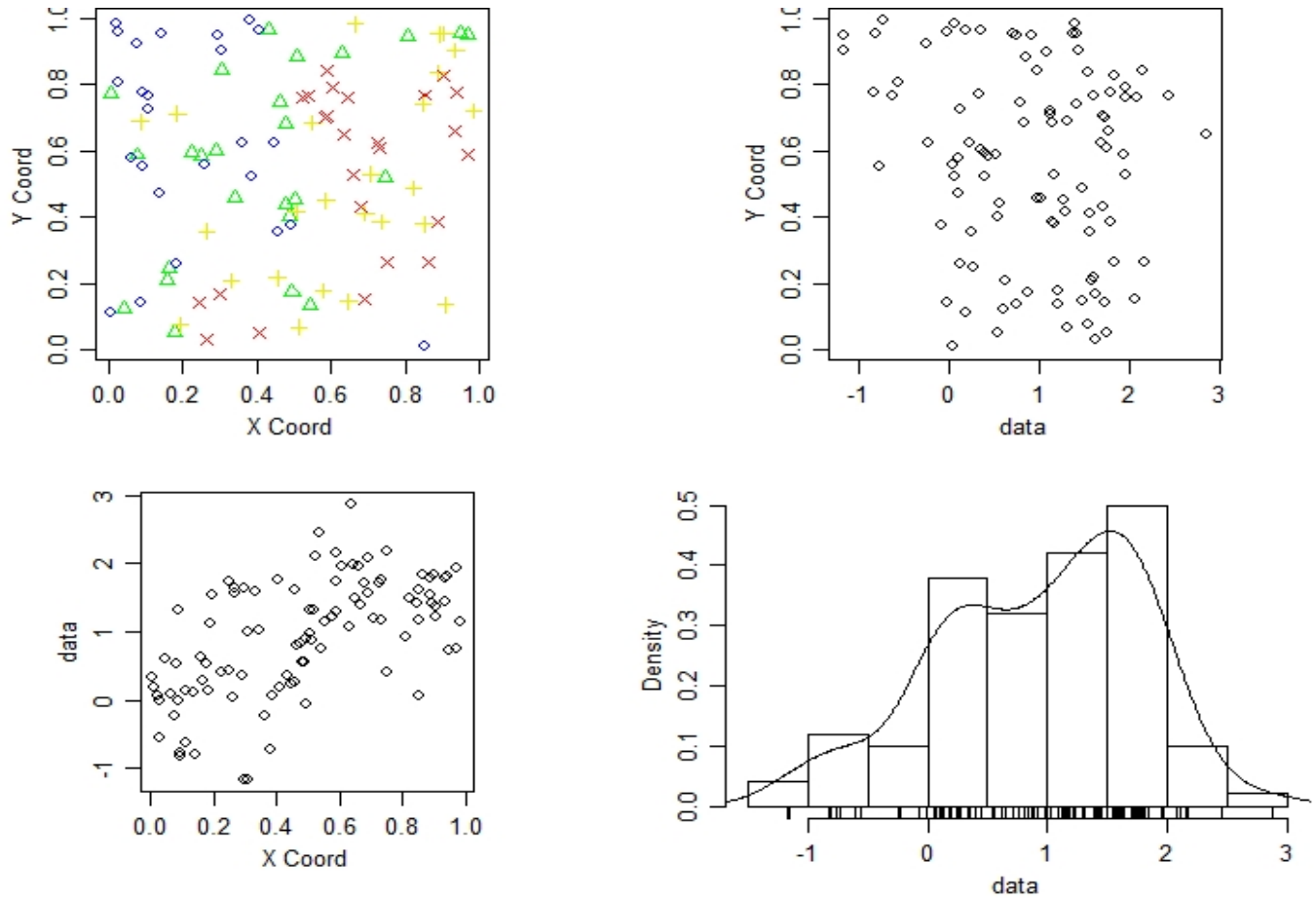
A quick summary of the data can be obtained typing

```
> summary(s100)
```

It will return basic information about the coordinates and data values.

The function **plot.geodata** shows a 2 x 2 display with data locations (top plots) and data versus coordinates (bottom plots):

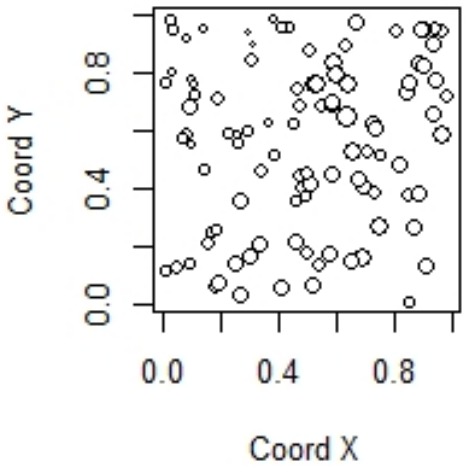
```
> plot(s100)
```

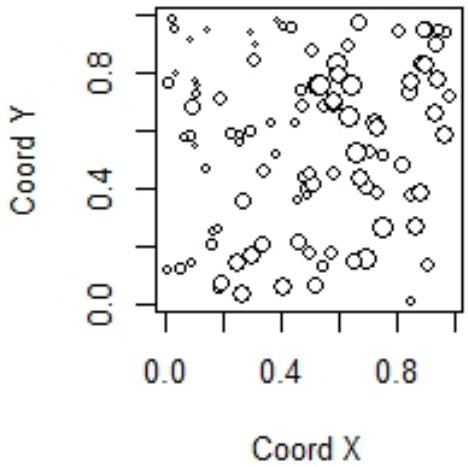


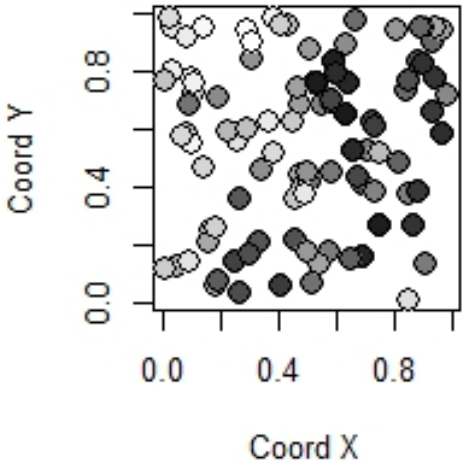
The function **points.geodata** produces a plot showing the data locations.

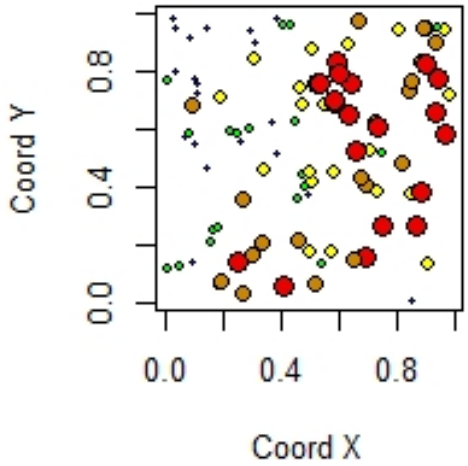
There are options to specify point sizes, patterns and colors, which can be set to be proportional to the data values or specified quantiles. Some examples of graphical outputs are illustrated by the commands and corresponding plots as shown below:

```
> par(mfrow = c(2,2))
> points(s100, xlab = "Coord X", ylab = "Coord Y")
> points(s100, xlab = "Coord X", ylab = "Coord Y",
+ pt.divide = "rank.prop")
> points(s100, xlab = "Coord X", ylab = "Coord Y",
+ cex.max = 1.7, col = gray(seq(1, 0.1, l=100)),
+ pt.divide = "equal")
> points(s100, pt.divide = "quintile", xlab = "Coord X",
+ ylab = "Coord Y")
```


Coord Y
0.0
0.4
0.8
0.0
0.4
0.8
Coord X


Coord Y
0.0
0.4
0.8
0.0
0.4
0.8
Coord X


Coord Y
0.0
0.4
0.8
0.0
0.4
0.8
Coord X


Coord Y
0.0
0.4
0.8
0.0
0.4
0.8
Coord X

Basic analysis with sp

We use MEUSE data:

```
> library(lattice)
> library(sp)
> data(meuse)
> ?meuse
```

This data set gives locations and top soil heavy metal concentrations (ppm), along with a number of soil and landscape variables, collected in a flood plain of the river Meuse in Belgium.

Heavy metal concentrations are bulk sampled from areas of approximately 15m x 15m.

This data frame contains the following columns in which we are interested:

- **x** a numeric vector; x-coordinate (m) in RDM (Dutch topographical map coordinates)
- **y** a numeric vector; y-coordinate (m) in RDM (Dutch topographical map coordinates)
- **zinc** topsoil zinc concentration
- **dist** distance to river Meuse; obtained from the nearest cell in meuse.grid, which in turn was derived by a spread (spatial distance) GIS operation, therefore it is accurate up to 20 metres; normalized $[0, 1]$

First we transform the data the Spatial class:

```
> coordinates(meuse) <- c("x", "y")
```

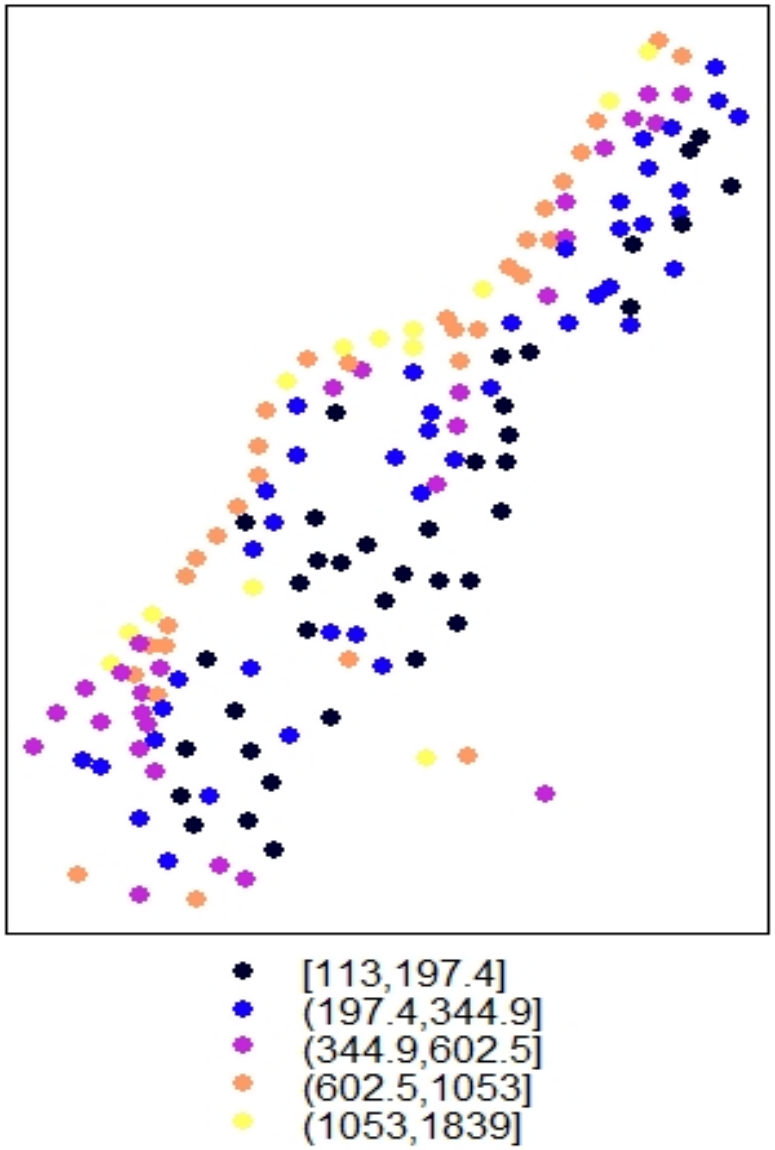
Then, to plot the observed value, we can use colour

```
> spplot(meuse, "zinc", do.log = T)
```

or symbol size:

```
> bubble(meuse, "zinc", do.log = T, key.space = "bottom")
```

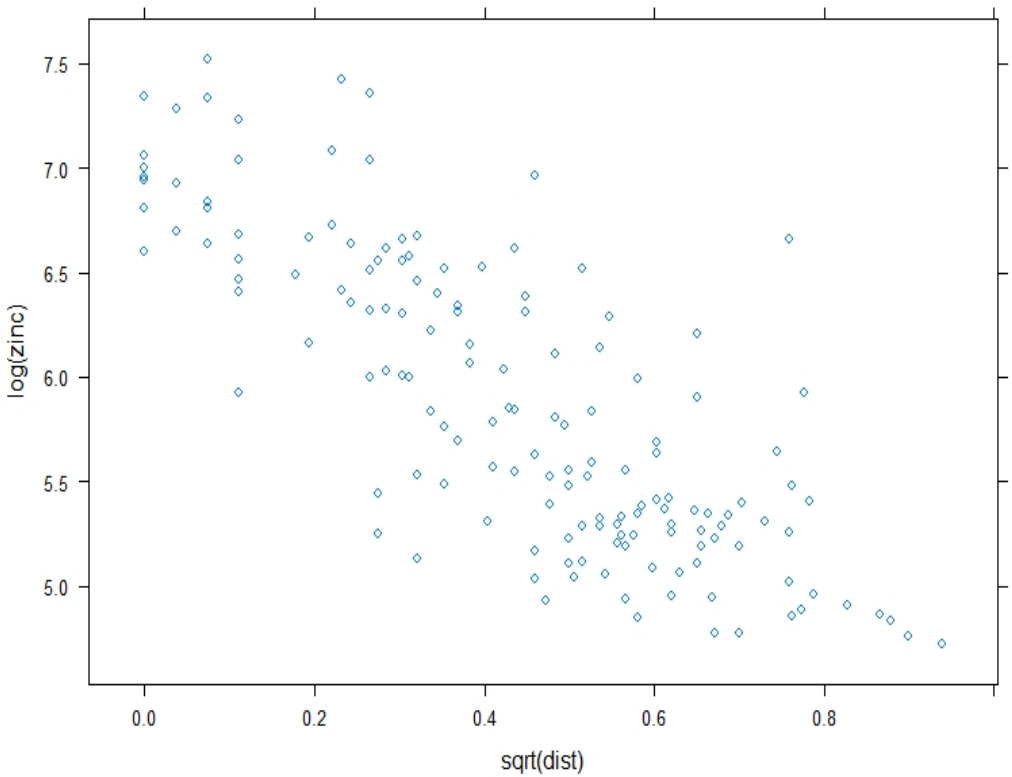
The evident structure here is that zinc concentration is larger close to the river Meuse banks.


[113,197.4]
(197.4,344.9]
(344.9,602.5]
(602.5,1053]
(1053,1839]


zinc
113
198
326
674.5
1839

There is an evident spatial trend, such as the relation between top soil zinc concentration and distance to the river. However this trend is non-linear. Applying appropriate transformations of the concentration and distance attributes it can be made approximately linear:

```
> xyplot(log(zinc) ~ sqrt(dist), as.data.frame(meuse))
```



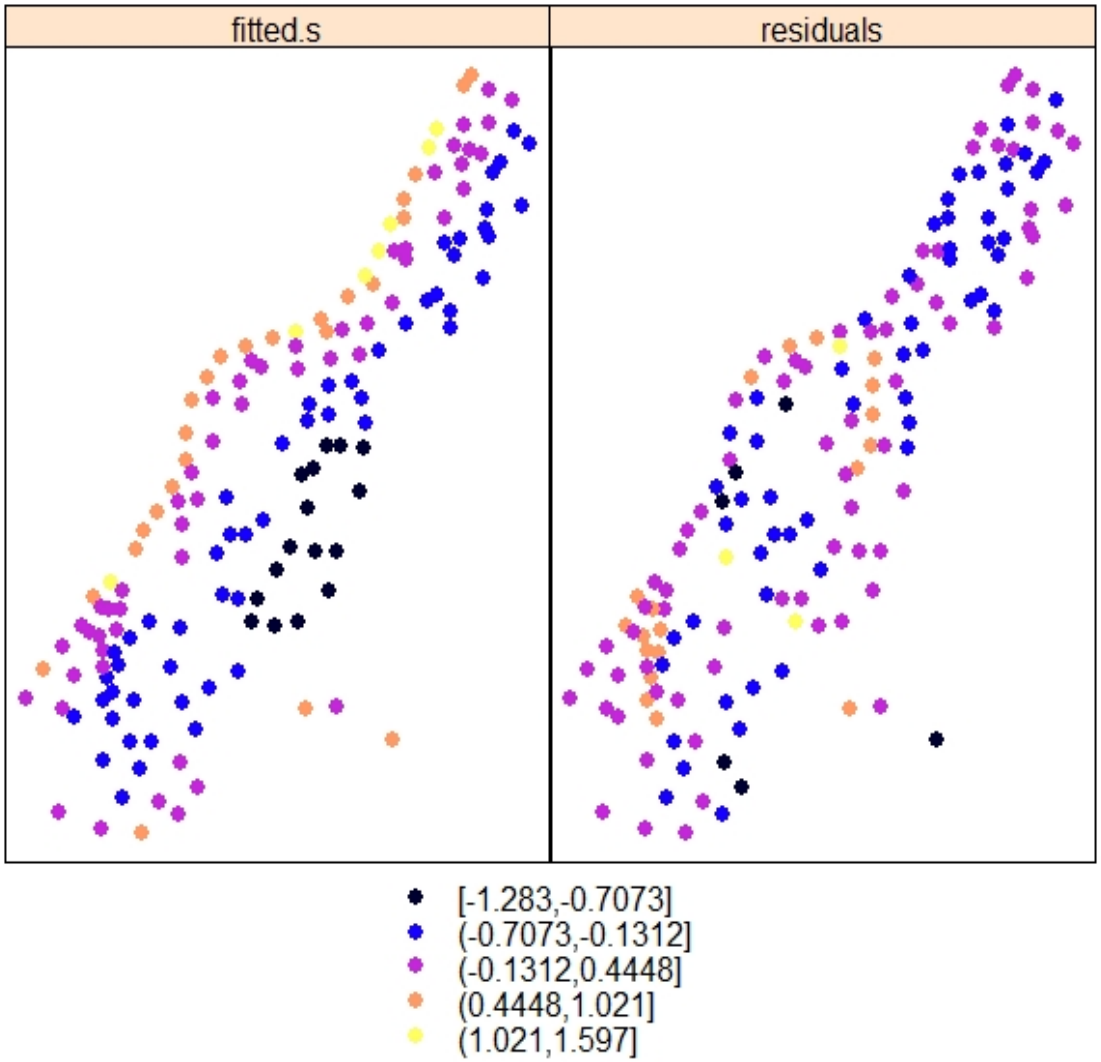
Therefore, we can also plot maps with fitted values and with residuals:

```
> zn.lm <- lm(log(zinc) ~ sqrt(dist), meuse)
> meuse$fitted.s <- predict(zn.lm,meuse)
- mean(predict(zn.lm,meuse))
> meuse$residuals <- residuals(zn.lm)
> spplot(meuse, c("fitted.s", "residuals"))
```

where the formula $y \sim x$ indicates dependency of y on x.

The resulting figure reveals that although the trend removes a large part of the variability, the residuals do not appear to behave as spatially unstructured: residuals with a similar value occur regularly close to another.

So, we need some random field model to describe the residuals and improve the prediction.


fitted.s
residuals
[-1.283,-0.7073]
(-0.7073,-0.1312]
(-0.1312,0.4448]
(0.4448,1.021]
(1.021,1.597]

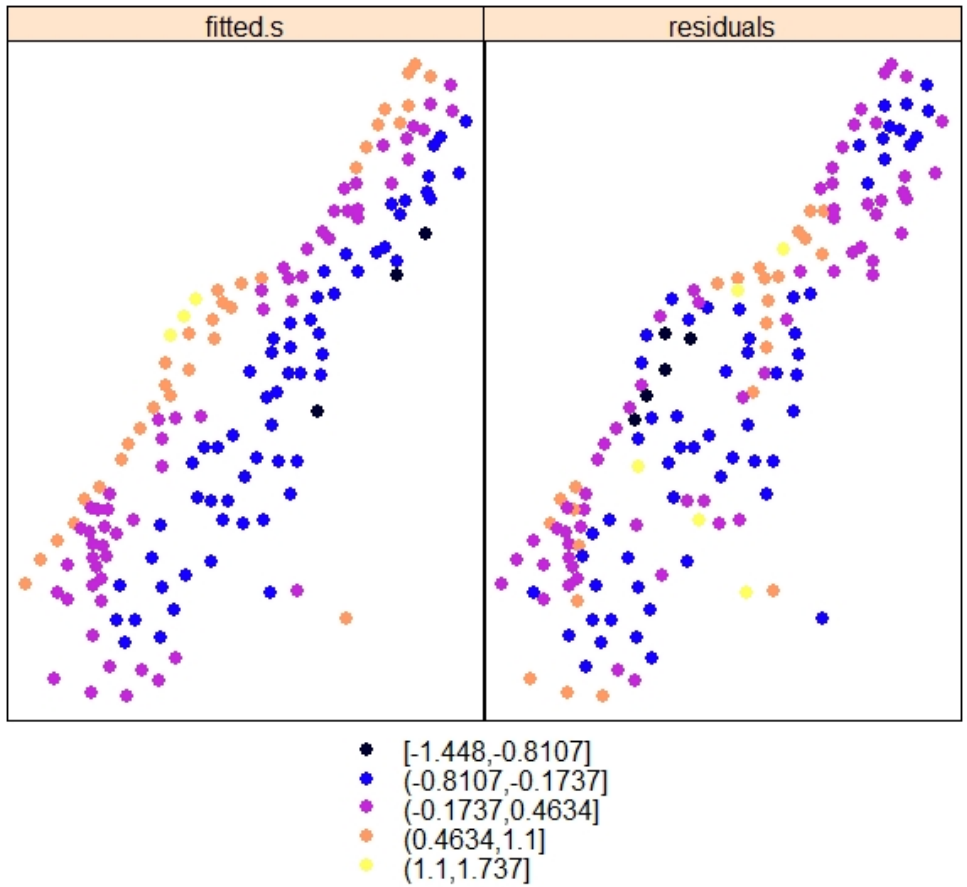
Estimating trends.

A special form of linear regression is obtained when polynomials of spatial coordinates are used for predictors, for example a second-order polynomial. This form is called trend surface analysis.
To use **lm** for trend surface analysis, for example, for the second order trend apply `I` (to treat powers and products as is):

```
> fit <- lm(log(zinc) ~I(x^2)+I(y^2)+I(x*y)+x+y,meuse)
> fit
Call:
lm(formula = log(zinc) ~ I(x^2) + I(y^2) + I(x * y) + x + y,
data = meuse)
Coefficients:
(Intercept)       I(x^2)       I(y^2)     I(x * y)
2.395e+04    8.575e-07    8.467e-07   -1.623e-06
x            y
2.279e-01   -2.684e-01
```

To plot maps with fitted values and residuals, use the commands:

```
> meuse$fitted.s <- predict(fit,meuse) - mean(predict(fit,meuse))
> meuse$residuals <- residuals(fit)
> spplot(meuse, c("fitted.s", "residuals"))
```



Key R commands

read.csv(x)	*reads a csv file and creates a data frame from it*
cbind(x)	*combines the sequence x by columns*
plot.geodata(x)	*produces a 2 x 2 display of geostatistical data*
points.geodata(x)	*produces a plot with points indicating the data locations*
bubble(x)	*creates a bubble plot of spatial data*
xyplot(x)	*produces a bivariate scatterplot*
lm(x)	*fits a linear model*
I(x)	*indicates that x should be treated "as is"*
predict(x)	*predicts using a fitted model*
residuals(x)	*extracts model's residuals*