Introduction to Geostatistics

12. Looking forward: multivariate and geostatistics

Edzer J. Pebesma

edzer.pebesma@uni-muenster.de
Institute for Geoinformatics (**ifgi**)
University of Münster

summer semester 2007/8,
July 7, 2008

Regression prediction, prediction SE

We can make regression predictions, for specific conditions of the X variables, e.g. for simple linear regression at x_0:

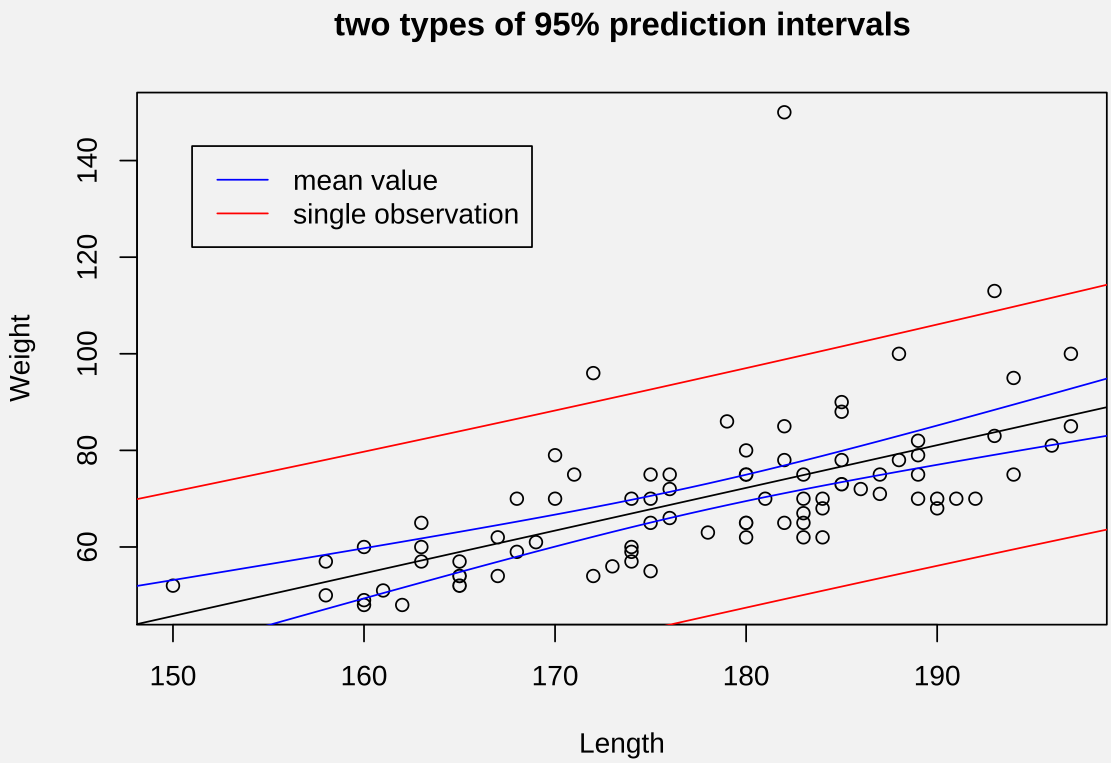
$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

This prediction has an error, the prediction standard error that we can retrieve from the analysis; for the mean value it is:

$$\mathrm{SE}_{\bar{y}_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

for a single observation it is:

$$\mathrm{SE}_{\hat{y}_0} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$



Looking forward

We could take this course forward in many ways, notably

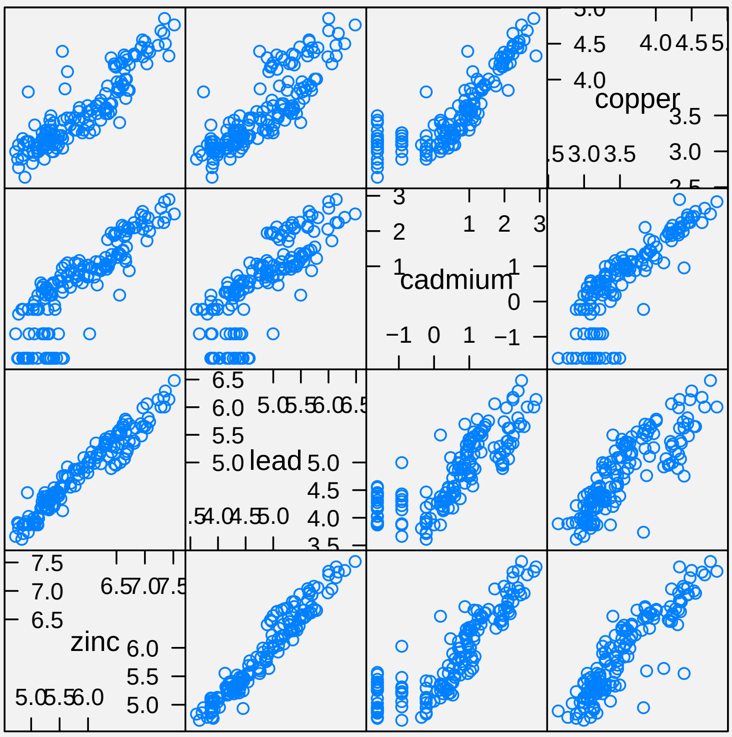
- considering more regression extensions,
- looking at multivariate data
- looking at spatial data, temporal data, and spatio-temporal data.

This will be done to some extent, probably in the analysis of

- species composition (ordination)
- remotes sensing data (multivariate data)
- analysis of e.g. chemical data, or sediment data
- analysis of spatio-temporal data

Multivariate data

- Looking at the world as either univariate or multivariate, everything we did starting from correlation and regression, maybe even the two-sample t-test, is multivariate
- In a more strict sense, or traditionally, multivariate data analysis looks at multiple *dependent* variables (i.e., *not* regression)
- Typical problems include
  1. exploration: what are the structures or patterns in the data?
  2. induction: can we test, and infer something about the population?
  3. prediction: can we predict (and map) the variable(s) under question?



Scatter Plot Matrix

Exploration: Directions

Exploratory methods usually looks at either ordination or clustering (finding groups).
Ordination is concerned with finding main directions of variability, ignoring minor directions:

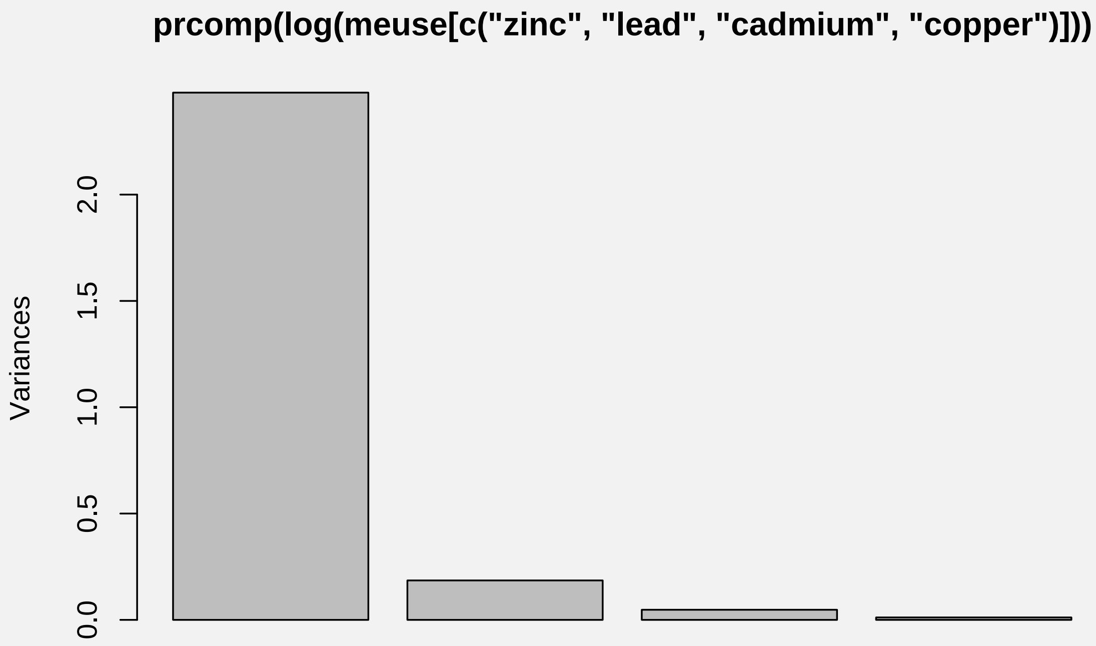
- Can we summarize or reduce the data set to a small (1-3) number of independent variables, with a minimum loss of information?

Principal component analysis

```
> prcomp(log(meuse[c("zinc", "lead", "cadmium", "copper")]))

Standard deviations:
[1] 1.5746879 0.4307029 0.2181036 0.1070498

Rotation:
               PC1        PC2         PC3         PC4
zinc    -0.4362357 -0.4786253  0.04790384  0.76047447
lead    -0.3876017 -0.5674372  0.40244994 -0.60482558
cadmium -0.7577083  0.6410163  0.11615542 -0.03852475
copper  -0.2921326 -0.1950152 -0.90677847 -0.23319613
```

$$PC_1 \approx -0.43\log(zn) - 0.38\log(pb) - 0.75\log(cd) - 0.29\log(cu)$$

$$PC_2 \approx -0.47\log(zn) - 0.56\log(pb) 0.64\log(cd) - 0.19\log(cu)$$



Exploration: Groups

Instead of finding *directions*, i.e. continuous variation, we could also look for discrete structure in the variable space, by ways of finding groups. Typically this is done by cluster analysis.

Question:

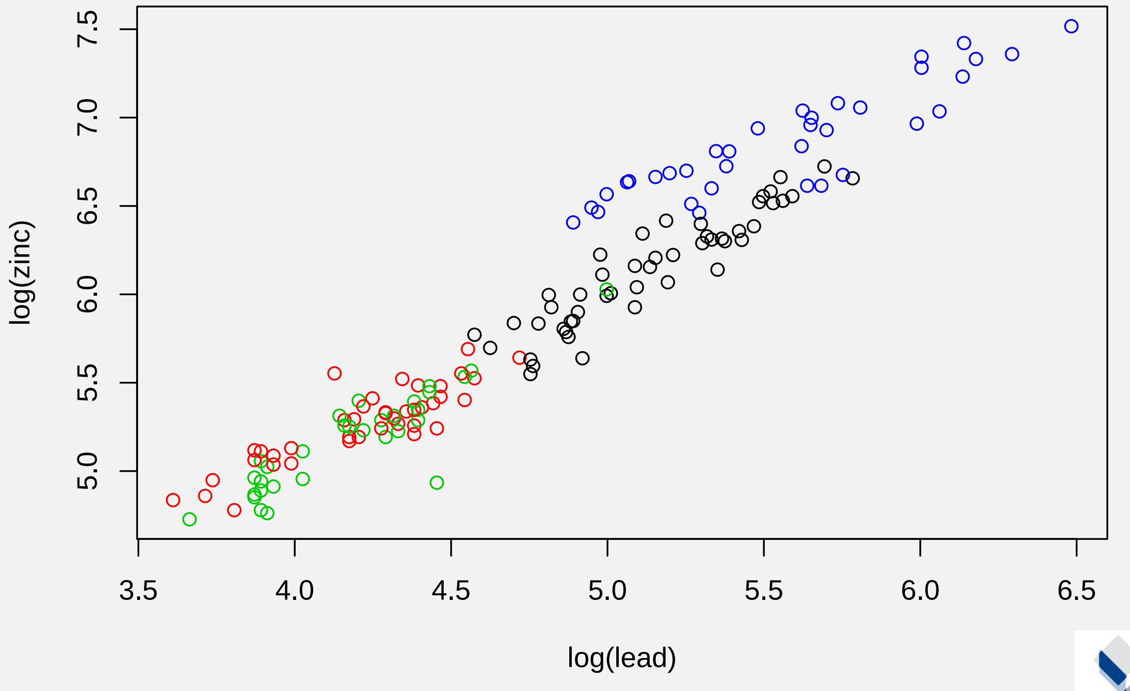
- Can we find a set of (n?) groups, that point to distinct behaviour?

Typical problems:

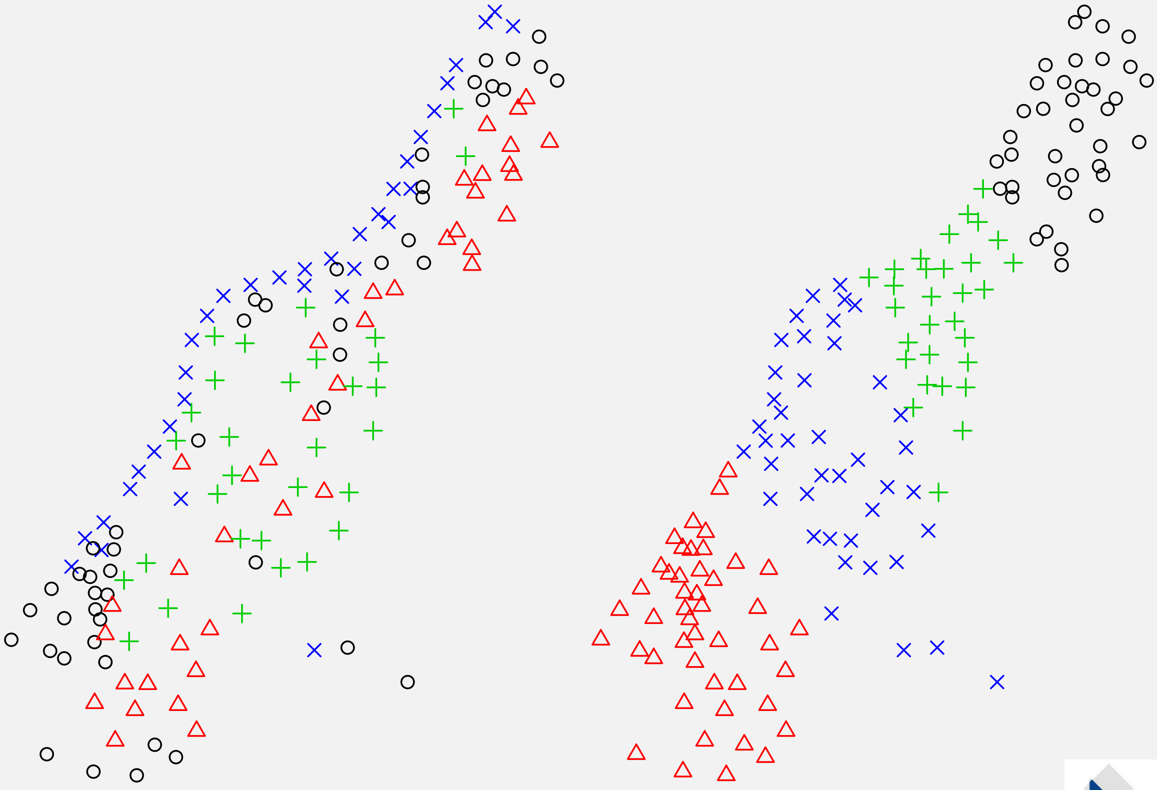
- How large should n be?
- How should we measure distances in feature space?
- What does it *mean*?

```
> kmeans(log(meuse[1:20, c("zinc", "lead", "cadmium", "copper")]),
+     4)

K-means clustering with 4 clusters of sizes 7, 2, 9, 2

Cluster means:
      zinc     lead   cadmium   copper
1 5.815262 4.918431 1.0097533 3.746102
2 5.656175 4.721123 0.6404669 3.257356
3 6.766799 5.402239 2.1974349 4.374412
4 5.225617 4.418187 0.4032379 3.198465

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 3  3  3  1  1  1  1  1  1  4  4  2  3  1  2  3  3  3  3  3

Within cluster sum of squares by cluster:
[1] 1.48979892 0.08556131 1.83742203 0.01288405

Available components:
[1] "cluster"  "centers"  "withinss" "size"
```

Clusters in feature space:



Clusters in geographical space:



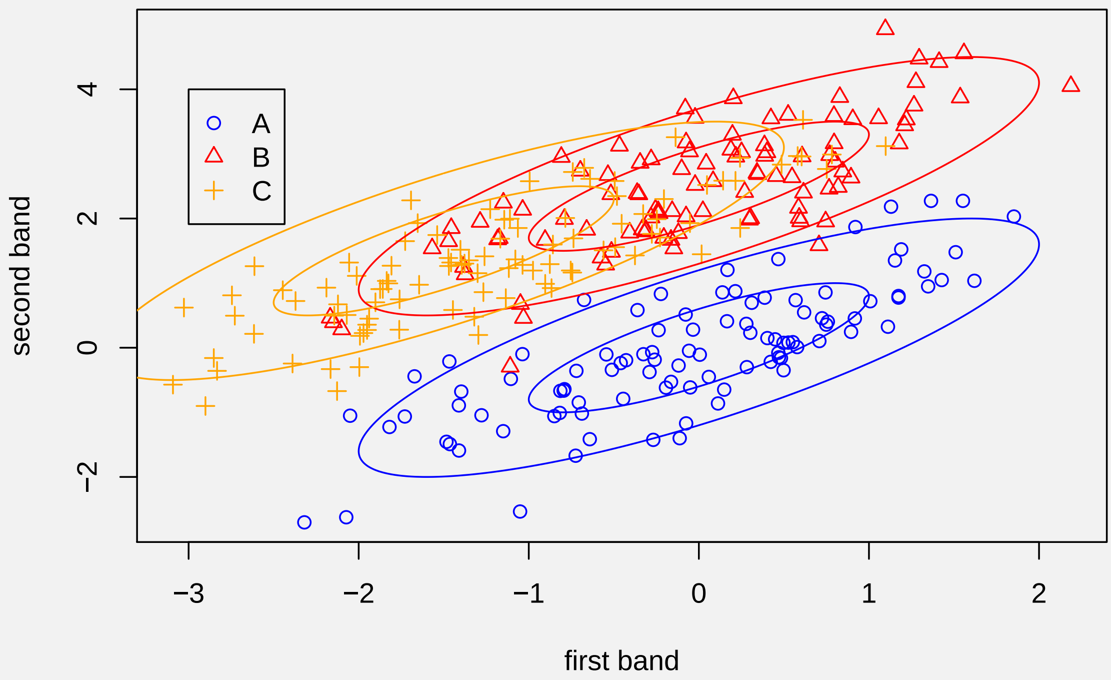
Left: based on (log-) heavy metals, right: based on spatial coordinates

Predicting a group variable

Suppose we have a dependent variable that is categorical, and a set of (discrete or continuous) independent variables, known as map, and we want to map the dependent variable. A prototypical example is land use classification:

- we have a set of images with reflectances, in different band widths (e.g. panchromatic, R, G, B) or e.g. Landsat (7 bands), Aster (15 bands), or hyperspectral ($>$ 100 bands)
- we have a set of *ground truth* observations, with known location and known category (land use).

Now, (i) model these data and (ii) predict land use at all locations (image pixels).



Discriminant analysis

(In Remote Sensing known as *maximum likelihood classification*)
General idea:

- elliptical contours are formed from group means and covariances, assuming normal distribution
- subsequent contours indicate the likelihood that we belong to a certain class
- a new pixel will be classified to the category for which its membership likelihood is maximized
- lines can be drawn where the class boundaries take place
- Here: ellipses have identical shapes and orientation, this can be generalized to group-dependent shape and orientation

Spatial statistics

Geostatistics in the narrow(er) sense considers

- to which extent observations are correlated in space (“observations near in space tend to be similar”)
- how we can best use this spatial correlation for spatial prediction (interpolation), and
- what is the interpolation error

Note that *independence* resulting from simple random sampling can coexist with the notion of *spatial dependence*:

1. spatial random sampling: $z(X)$, z non-random, X random
2. geostatistics: $Z(x)$, Z random, x non-random

However, model 1 is not of much use if we want to interpolate, because we do this at non-random locations.

Time series data

Time series analysis typically looks at two aspects:

- temporal correlation (small time lags typically show small variation)
- periodicities, because of the periodicity in nature (days, years) and human behaviour (weeks).

Questions addressed are:

- can we describe the temporal variability with a simple model?
- (how well) can we predict the future?

Looking back

Much of what we did follows from two questions:

- What kind of variable(s) are we interested in?
  - is it one, two, three? relations between them? prediction?
  - are we interested in location, variability, correlation?
- What is/are the measurement scale(s) of this/these variable(s)?

The test

Looking at older tests could help, but don't expect much—I did not look at them.
Multiple question, (hopefully) bi-lingual.
Simple calculator recommended.
No R commands will be asked, but statistical output (graphs, text) will be there.