Spatial Statistics

Computer lab – Session 4

Madlene Nussbaum Andreas Papritz

21 Oct 2024

Table of contents

1 Geostatistical analysis of lognormal data with anisotropy

1.1 Exploratory analysis

Copper content in soils around metal smelter in Dornach

Data:

- dornach.txt
- dornach_grid.txt

Read the data file dornach.txt into R. The data set contains the following variables:

- `x`, `y`: anonymized coordinates of sites where soil samples were taken, the origin of the coordinate system is the main stack of the smelter,
- `survey`: factor coding the survey in which the data was collected,

- `forest`: factor coding whether the site is in a forest,
- `built.up`: factor coding when a site has been built up,
- `geology`: factor coding the geologic parent material at a site,
- `cu`, `cd`, `zn`: heavy metal concentration (mg/kg) in topsoil (0–20 cm depth).

We will predict topsoil copper content `cu` to evaluate the pollution situation over the study are.

> Task 1
>
> Plot the spatial distribution of the sites with observed `cu` and its value, use e.g. a bubble plot.

> Task 2
>
> Add the distance to the smelter as a new column to the data. Explore the distribution of `cu`. Find a suitable transformation for both `cu` and the distance to approximately linearize the relationship.

> Task 3
>
> Explore the directional relationship. Compute the angle to the smelter by `atan2(x,y)` and convert the angle to a factor using `cut()`, e.g. differentiate North/East/South/West. Plot the angles with different colors per directional group.

1.2 Fit a trend model

> Task 1
>
> Fit a ordinary least squares regression using distance and angle as explored above. Use `cos()` and `sin()` to transform the angle in continuous covariates (results in a periodic function).

> Task 2
>
> Extract the residuals from this regression model and check whether `cu` further depends on the land use (variables `forest`, `built.up`), on the parent material (`geology`) and on the origin of the data (`survey`)? Use for example boxplots with notches. If necessary update the regression model by adding those covariates that seem to influence `cu`.

Task 3

Assess the fit of the regression model by the usual residual diagnostics plots and display the spatial distribution of the residuals by a bubble plot.

1.3 Explore variogram and fit REML

Task 1

Compute the sample variogram of the residuals of the regression model and fit an exponential variogram model to the sample variogram.

Task 2

Estimate the coefficients of the trend model fitted above and the parameters of the exponential variogram model now by Restricted Maximum Likelihood (REML) using the function `georob()`.

Task 3

Fit an anisotropic variogram model with `georob()`. Fit `f1` and `omega` and keep the other parameters fixed. Check example on the help page.

Hint

Try to find reasonable starting values: According to above directional plot, `cu` values in Western direction seem to be smaller, i.e. the decrease is larger than in the other directions. Therefore, it is likely `f1` $<$ `f2`. For `omega` we would need to do more plotting or try to fit with `georob()` and check if the numerical solving of the model converged. If not, we would need to find better starting values.

Task 4

Assess the fit of the model by residual diagnostic plots.

Note: we could now to further model selection and trying to improve the model also by adding interactions of the covariates. We could do some plotting and use `georob::step()`.

1.4 Predictions with random forest

Task 1

Fit a random forest model with `ranger()` using all available covariates. Since `cu` is strongly positively skewed use `log(cu)`. Maybe plot importance of covariates.

1.5 Model assessment

Task 1

Compute cross-validation for the geostatistical models fited with REML above (isotropic and anisotropic). Use the function `cv()` with `lgn = TRUE` to directly include unbiased backtransformation of the lognormal response.
Would you use `method = "random"` or `method = "block"` for spatial cross-validation?

Task 2

Compute cross-validation for random forest model. Use the same cross-validation subsets as above. You can access them from the `georob` cross-validation object by `yourObject$pred$subset`.

Task 3

Compute scatterplots with predicted vs. observed. Add a lowess smoother and a 1:1-line. Create these plots for the log-transformed results and the back-transformed results.
For random forest use backtransformation without bias correction.

Task 4

Compute meaningful validation metrics and compare the model performance.

1.6 Prediction for `dornach_grid`

> Task 1
>
> Compute kriging predictions with the best performing REML fit above. For this use `predict(..., control=control.predict.georob(extended.output=TRUE)`, then use the function `lgnpp()` to obtain the unbiased backtransformation of the lognormal predictions.

> Task 2
>
> Compute random forest predictions. Backtransform by `exp()`.

> Task 3
>
> Create maps of predictions and kriging standard errors. Compare the maps.

Note: for random forest we could compute the full predictive distribution using `quantreg = T` for the model fit and `predict(..., type = "quantiles", quantiles = ...)`. From this distribution, we could form the required quantity at each pixel.

2 Model assessment pH predictions Berne data set

In lab session 3 you computed predictions for topsoil pH in the Berne study area. We did not yet fully inspect the model performance.

Find your saved CSV file from session 3 or download the example CSV. CSV

> Task 1
>
> Create plots and metrics as you see fit. What is the model performance? Are there relevant problems?