12 Linear regression model with spatial dependency for the Irish pH data

In this chapter we will apply a multiple linear regression model with a spatial dependency component using the SPDE approach on the Irish pH data. Preliminary analyses of this data set were presented in Chapters 2 and 4.

Prerequisite for this chapter: It is recommended that you have a conceptual understanding of the SPDE approach that was discussed in Chapter 11. However, if you skipped Chapter 11 you can still catch up because this chapter contains conceptual revisions.

12.1 Introduction

The Irish pH data were discussed at various points in previous chapters, and therefore we will keep this introduction short. Sampling took place at 257 locations (see Figure 2.1) along rivers in Ireland in 2003. We will use data from 210 locations in the analysis. The aim of the study is to model pH of the water as a function of SDI (sodium dominance index), altitude (log transformed; see Section 4.2), and whether a site is forested (categorical variable with the values yes and no).

We argued in Chapter 2 that spatial dependency between pH values at sites sampled close to one another is likely. To avoid pseudoreplication we need a model that allows for the spatial dependency.

Data exploration was (partly) carried out in Section 2.2 and is not repeated here.

12.2 Model formulation

In Chapter 2 we kept the models for this data set simple but here we apply the model with all main terms, two-way interactions, and the three-way interaction; see Equation (12.1).

$$
\begin{aligned}
& pH_i \sim N\left(\mu_i, \sigma^2\right) \\
& E\left(pH_i\right) = \mu_i \quad \text{and} \quad \text{var}\left(pH_i\right) = \sigma^2 \\
& \mu_i = \alpha + \beta_1 \times SDI_i + \beta_2 \times LogAltitude_i + \beta_3 \times Forested_i \\
& \qquad \beta_4 \times SDI_i \times LogAltitude_i + \beta_5 \times SDI_i \times Forested_i + \\
& \qquad \beta_6 \times LogAltitude_i \times Forested_i + \\
& \qquad \beta_7 \times SDI_i \times LogAltitude_i \times Forested_i
\end{aligned}
\tag{12.1}
$$

In the case of numerical estimations problems in R-INLA, the first thing to try for solving the problem is to standardise the continuous covariates. We did not do that here because working with a Gaussian distribution and a sample size of 210 observations makes us reasonably optimistic that we will not encounter numerical problems.

12.3 Linear regression results

The model in Equation (12.1) can be fitted in R-INLA with the following code. We import the data with the `read.table` function, define a categorical covariate fForested, log transform altitude, and then apply the linear regression model in R-INLA.

```
> iph <- read.table(file = "IrishPh.txt",
                    header = TRUE,
                    dec = ".")
> iph$fForested <- factor(iph$Forested,
                          levels = c(1, 2),
                          labels = c("Yes", "No"))
> iph$LogAlt <- log10(iph$Altitude)
> library(INLA)
> I1 <- inla(pH ~ LogAlt * SDI * fForested,
             family = "gaussian",
             control.predictor = list(compute=TRUE),
             data = iph)
```

The `compute = TRUE` for the `control.predictor` option ensures that R-INLA calculates the fitted values. The numerical output of the linear regression model is stored in the list `I1`, which has various objects that we can inspect. The first two objects are `I1$summary.fixed` and `I1$summary.hyperpar`.

```
> Beta1 <- I1$summary.fixed[, c("mean", "sd",
                     "0.025quant", "0.975quant")]
> print(Beta1, digits = 3)

                           mean    sd  0.025q 0.975q
(Intercept)              10.035 1.918  6.267 13.801
LogAlt                   -0.774 0.931 -2.604  1.053
SDI                      -0.036 0.033 -0.100  0.028
fForestedNo              -1.783 2.063 -5.839  2.268
LogAlt:SDI                0.005 0.016 -0.026  0.036
LogAlt:fForestedNo        0.880 1.009 -1.102  2.860
SDI:fForestedNo           0.008 0.037 -0.065  0.081
LogAlt:SDI:fForestedNo   -0.003 0.018 -0.039  0.032
```

The first column shows the posterior mean and the third and fourth columns the 95% credible intervals. The 95% credible interval for β_8 goes from –0.039 to 0.032. This means that there is a 95% probability that β_8 is in this interval. Because 0 is also in this interval we state that β_8 is 'not important' as compared to the frequentist phrase 'not significant'.