BI5302 Dealing with spatial non-independence exercise

Exercise: Spatial GLS - Ungulate species distribution models

0. This practical expands on the spatial data analysis lecture. It will allow you to revisit the analysis of environmental drivers of the wild boar distribution in Poland, using a larger set of environmental variables, and of the distribution of other species if you wish. The idea is to practice doing model selection while accounting for non-independence in the residuals by modelling their spatial covariance. More info on the data for this exercise in the published paper: http://bit.do/UngulatesDistributionPaper. Variables:

- `Reddeer`: Red deer abundance index
- `Roedeer`: Roe deer abundance index
- `Wildboar`: Wild boar abundance index
- `Moose`: Moose abundance index
- `Total.arable`: proportion of arable land (all 3 categories above pooled)
- `Pastures`: proportion of pasture land
- `BroadLForest`: proportion of broad leaved forest
- `ConForest`: proportion of coniferous forest
- `MixForest`: proportion of mixed forest
- `Shrub`: proportion of shrubland
- `Total.forest`: proportion of forests (all 4 above pooled)
- `DeciMixFor`: proportion of broad-leaved and mixed forests
- `Marsh`: proportion of marsh land
- `Waterbodies`: proportion of water bodies
- `Open`: proportion of open habitat (non-forested)
- `SnowDays`: mean number of snow days
- `JanTemp`: mean January temperature
- `LONG`: Longitude
- `LAT`: Latitude
- `Quadrant`: Poland divided in 4 blocks: NE / NW / SW / SE

1. Import the data file ‘ungulates.csv’ into R. Before doing any data exploration, take a look at the structure of this dataframe, and check the (rather large!) list of predictors available to you.

Take a moment to think if all are plausible predictors of ungulate abundance, what they represent, and how they may differ (or perhaps some are partly redundant). Are there any that seem more relevant to you or some which you may wish to exclude a priori? If you refuse to make choices at this stage they might impose themselves to you and maybe complicate your life later on!

Please run the code below to load the data, the R libraries and a couple of extra optional but useful functions.

```
library(lattice)
library(nlme)
library(effects)
library(knitr)

# load extra panel function for pair plots
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs2<- function (x) pairs(x, lower.panel = panel.smooth, upper.panel = panel.cor)
```

2. Start with an initial data exploration of the predictors you have chosen to try and explain the abundance of wild boars. If you have too many predictors, consider making smaller subsets to make the exploration easier. Look at any collinearity between them, and if needed, narrow down the set of candidate predictors you want to work with and make any transformation you feel would be necessary.

3. Using your set of candidate predictors, do an extra piece of data exploration to look relationships with wild boars abundance index. Do you need to transform the response? Any other transformation required?

Hint: my set of candidate predictors includes `c("DeciMixFor", "ConForest", "LogPastures", "LogMarsh", "LogWaterbodies", "JanTemp")`, but there is absolutely no requirement for you to use the same ones. The most important is that you have a rationale for whichever candidate predictors you decide to use.

4. (Optional) Create maps of the candidate predictors. Hint: you could use the 'symbols' function (see lecture slides), remembering to set the maximum symbol size with 'inches='. Which of these variables vary at the broadest and smallest spatial scales?

5. Fit a GLS model with the predictors of interest (include only additive effects unless you have a theory for why a particular interaction would be relevant to include). Are all terms significant? Plot the empirical variogram of the residuals of the model. Based on lecture slide #18, what might be a sensible variogram model for fitting this empirical variogram? What might be sensible starting values for the range and the nugget?

6. Re-fit the GLS model, including potential spatial correlation structure. Hint: Remember to use the correct fitting method for selecting covariance strucures, and to check that estimated range and nugget are a sensible fit to the data. What structure is best according to AIC? **If model fitting is too slow on your computer, use 'form= ~ LONG + LAT | Quadrant' to speed things up.** (Note that this will probably affect the range estimate, as the 4 regions over which the covariance is calculated will be smaller: this is an approximation in order to gain some speed)

7. Apply model selection (using AIC?) to the best GLS model above. Remember to select the appropriate model fitting method for the fixed part of the model.

8. (Optional, if you have plenty of time left) Re-fit your minimum adequate model with REML and test alternative covariance structures, to be sure that the change in the fixed part of the model hasn't changed the residuals structure. Make any required adjustment to the random or fixed parts.

9. Validate your minimum adequate model, using the usual graphs and maps of the residuals as in the lecture. For which plot should you use which type of residuals? Is everything looking good?

10. Interpret your minimum adequate model using graphs of the fitted effects.

11. (Optional) compare the minimum adequate model obtained using GLS with spatial autocorrelation to a MAM obtained by model selection without accounting for spatial non-independence. What difference does it make?

12. (Optional) There are a few more species which you could use as response variables, if you fancy some practice…

End of the spatial autocorrelation exrcise