Assignment 2

Geospatial analysis in R
KDI School Fall 2024

Due date: Tuesday, November 26th before class

This assignment is designed to practice with rasters and spatial interpolation. I will also guide you through cross validation as a way to evaluate the performance of your interpolation.

There are two key files we will use: - `krpollution.csv`: this file includes information on pollution (PM 10) for a single day in November of 2024. The data is at the station level. - `koreashape.shp`: this is a shapefile for Korea.

Here are your tasks:

- First, I would like you to create three separate maps, all in one figure:
  - A map of log PM10 for each shapefile feature (i.e. each dong) using the nearest neighbor method.
  - A map of log PM10 for each shapefile feature (i.e. each dong) using inverse distance weighting, with a weighting parameter of 2.
  - A map of log PM10 for each shapefile feature (i.e. each dong) using simple kriging (no predictors).
  - Please describe how the maps differ and why you think that is.
- Second, we are going to see which of these three methods seems to best predict pollution levels. We are going to do this using cross validation. In the pollution dataset, you will see a variable called `folds`, that ranges from 1 to 5. These folds have been *randomly assigned* to each station. You are going to estimate each of the three methods above using the data from only four folds at a time, completely leaving out the last fold. You will then predict the pollution levels for the stations in the last fold. You will do this for each fold, and then calculate the mean squared error (MSE). Here are the exact steps:
  - First, estimate all three methods using the data from folds 2, 3, 4, and 5. Predict the pollution levels for the stations in fold 1. In other words, the `newdata` with be fold 1, while the `data` will be the other four folds. For each prediction, calculate the

mean squared error as the mean of the squared differences between the predicted and actual pollution levels. In other words:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the actual pollution level, $\hat{y}_i$ is the predicted pollution level, and n is the number of stations.[1]

  - You will repeat this four more times, leaving out one fold at a time. In total, you will estimate 15 separate predictions and calculate 15 separate MSEs.
  - Create a table that has all 15 MSEs, with the five folds as the rows and the three methods as the columns.

- For the final step, you will be creating a raster using the model above that gives the best (i.e. lowest) MSE across all folds. This will be a bit difficult, but here are some instructions and hints:
  - You should create a raster that covers the extent of the Korea shapefile with a resolution of 0.05 degrees.
  - You will then turn this raster into a grid using `as.polygons` (we covered this in class).
  - You will then predict pollution into each of these grid cells, using the centroids (again, using the method from above that gives the lowest MSE).
  - When you have the polygon grid with pollution estimates, you will need to figure out how to turn it back into a raster. Feel free to google it! Please plot the *raster* using `geom_spatraster`.

Along with the figures/tables, I'd like you to include a short write-up describing the results. All of this should be contained in a single document (i.e. there should be a single document with figures, tables, and a short write-up describing the results).

For full credit, you must turn in the following:

- R Markdown script
- pdf output of the R Markdown script
- R script if you did not directly do everything in the R Markdown script (I'll leave this choice up to you)

[1] Note that you can just use the `mean()` function here.