Exercise 2

Load the river blindness data, `LiberiaRemoData.csv`. Consider the empirical logit transformation of prevalence, given by

$$\tilde{Y}_i = \log\left\{\frac{Y_i + 0.5}{n_i - Y_i + 0.5}\right\},$$

where Y_i and n_i are the number of positive cases and number of tested individuals at location x_i.

1. Consider a standard linear model that ignores spatial correlation, hence

   $$\tilde{Y}_i = \alpha + \beta \log\{d(x_i)\} + Z_i.$$

   Create a grid of prediction locations over which to carry out spatial prediction of prevalence. After creating a raster file for the predictions, generate a plot in R that shows the predicted prevalence over the grid.

2. Fit a linear geostatistical model to the empirical logit, i.e.

   $$\tilde{Y}_i = \alpha + S(x_i) + Z_i$$

   where $S(x_i)$ is a stationary Gaussian process with exponential correlation function and Z_i are i.i.d. Gaussian variables.

3. Use the model from the previous question to predict nodule prevalence across Liberia and display the exceedance probability for a 20% threshold.

4. Now repeat point 2 and 3, but adding the log-transformed elevation as a linear predictor, i.e.

   $$\tilde{Y}_i = \alpha + \beta \log\{d(x_i)\} + S(x_i) + Z_i$$

   where $d(x_i)$ is the elevation (in meters) a location x_i. What differences do you observe with the model that did not use elevation?

5. Now consider a Binomial geostatistical model for prevalence, with linear predictor

   $$\log\left\{\frac{p(x_i)}{1 - p(x_i)}\right\} = \alpha + \beta \log\{d(x_i)\} + S(x_i) + Z_i$$

   Obtain *i*) predictions of prevalence from this model and *ii*) the probabilities of exceeding a 20% prevalence. Finally, compare this with those you have obtained from the linear geostatistical model fitted to the empirical logit, where elevation was used a covariate.