Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
July 7th, 2023

Problem n.3

Crop yield estimation refers to the process of predicting or estimating the amount of agricultural produce that will be harvested from a specific crop or field. The file `crops.txt` collects the yearly crop yield (i.e., the amount of harvested production per unit of land) of 60 locations s_i of Tuscany, next to the Santa Luce Lake (Riserva Naturale Provinciale). The dataset reports the UTM coordinates (`x`, `y`) of the locations s_i, whether the soil type of the location is sandy or clay (`sandy` = {Yes, No}) and the distance from the water source (Santa Luce Lake) $d(s_i) = ||s_i - s_f||$, with $s_f = (623299.322, 4811132.941)$. Consider the following model for the crop yield estimation:

$$y(s_i) = \beta_{0,g} + \beta_{1,g} \cdot d(s_i) + \delta(s_i) \tag{1}$$

with $\delta(s_i)$ a stationary residual with *spherical* variogram *with nugget* and $g = 0, 1$ the grouping induced by the variable `sandy` ($g = 0$ for `sandy` = No, $g = 1$ for `sandy` = Yes)

a) In Eq. 1, consider $\beta_{1,g} = 0$ and $\beta_{0,g} = \beta_0$ for $g = 0, 1$. Estimate the parameter β_0 of the model via generalized least squares. Report the model estimated for $\delta(s_i)$ and discuss the model assumptions.

b) By using the model fitted at point a), consider the *Podere di San Martino* as a new location s_0 = (`x`=625182.088, `y`=4811723.685), knowing that `sandy`=No. By assuming independence from the rest of the field, perform a prediction for the yearly crop yield $y^*(s_0)$. Comment on the obtained result.

c) In Eq. 1, consider now $\beta_{1,g} \neq 0$ for $g = 0, 1$. Estimate the parameters of the model $\beta_{0,g}$ and $\beta_{1,g}$ for $g = 0, 1$ via generalized least squares.

d) By using the model fitted at point c), consider the *Podere di San Martino* again and, by assuming correlation with the rest of the field, perform a prediction for the yearly crop yield $y^*(s_0)$.

e) Which model, between options a) and c), do you deem more appropriate for describing the data? Provide a comment explaining your choice.

Upload your solution here