Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
January 17th, 2024

Problem 4: Modeling PM10 concentrations in Lombardia

Particulate Matter (PM) refers to tiny particles or droplets in the air that can be inhaled into the lungs and may cause serious health problems. PM10 specifically refers to inhalable particles with diameters that are generally 10 micrometers and smaller. These particles can come from various sources, including vehicle exhaust, industrial emissions, construction activities, and natural sources such as dust storms and wildfires.

The Lombardia region is interested in modeling the distribution of PM10 concentration over its entire territory to assess air quality and potential health risks.

The file `pm10.txt` contains measurements of daily maximum concentration z of PM10 $[\mu g/m^3]$ recorded on July 14th, 2023, within 160 locations in Lombardia. The locations s_k $(k = 1, ..., 160)$ are expressed in UTM coordinates (variables `x` and `y`). Additionally, an "urbanization measure" of the area surrounding the location through the *numerical* variable `urban` ranging from 0 to 3 is provided. The variable `urban` is built as follow: `urban` $= 3$ inside an agglomeration, `urban` $= 2$ in an urbanized plain, `urban` $= 1$ in an low-urbanized plain and `urban` $= 0$ in an a mountain or forest area.

For all questions, $\delta(s_k)$ is a 2^{nd} order stationary residual with an Exponential model *with* Nugget for the spatial dependence structure.

a) Consider for the PM10 concentration $z(s_k)$ the following stationary model:

$$z(s_k) = \beta_0 + \delta(s_k) \quad \text{for} \quad k = 1, ..., 180,$$

Fitting the model via Generalized Least Squares, estimate the range and the sill of $\delta(s_k)$, along with β_0.

b) Now, consider the following modification of the previous model:

$$z(s_k) = \beta_0 + \beta_1 \, \texttt{urban}_k + \delta(s_k) \quad \text{for} \quad k = 1, ..., 180,$$

Fitting the model via Generalized Least Squares, estimate the range and the sill of $\delta(s_k)$, along with β_0 and β_1.

c) Finally, consider the model below:

$$z(s_k) = \beta_{0,j} + \delta(s_k) \quad \text{for} \quad k = 1, ..., 180,$$

where $j \in [0 \dots 3]$ is the grouping variable induced by the type of surrounding area (as defined in the variable `urban`). Fitting the model via Generalized Least Squares, estimate the range and the sill of $\delta(s_k)$, along with $\beta_{0,\cdot}$

d) Which of the three models would you keep? Justify and comment on your choice.

e) Using the selected model, provide a point prediction for the maximum concentration of PM10 on July 14th, 2023 at the location `(x=514961, y=5034538)`, corresponding to the Duomo di Milano.

Upload your results here: `https://forms.office.com/e/1a01FLwRga`