Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
February 15th, 2023

Problem n.2

The shapefile `Milan` contains the current NILs (Nuclei d'Identità Locale, i.e. the neighbourhoods) in the city of Milano (source: Comune di Milano) The neighborhoods are univocally defined by an `ID_NIL` (a number, to be used as ID variable), or `NIL` (the extended neighborhood name). For each NIL, the number `n` of public establishments such as bars, restaurants, pubs, etc., where serving is the primary activity (source: Comune di Milano) is reported. Moreover, the standardized variable `n_per_area` was created dividing, for each NIL, the variable `n` by the variable `Shape_Area` measuring the area.

a) By making use of a contiguity-based spatial weight of order 1 (**W1**), able to consider both common sides and common vertices, report the *minimum* and *maximum* number of neighborhoods and the *sparsity percentage* (i.e., the percentage of non-zero values).

b) According to **W1**, what is the number of Neighbors associated to the highest frequency? Report both this number as well as the associated highest frequency.

c) Based on **W1**, focus now on the variable `n_per_area` and build the LISA Cluster Map and Significance Map. Comment on them.

   Moreover, report the `NIL`s identified as outliers for a p-value≤ 0.05 when testing spatial autocorrelation. How would you explain their outlyingness?

d) Report the *Moran's I* and, among the `NIL`s with a positive spatial autocorrelation, report the one with the highest $z_j - \overline{z}$.

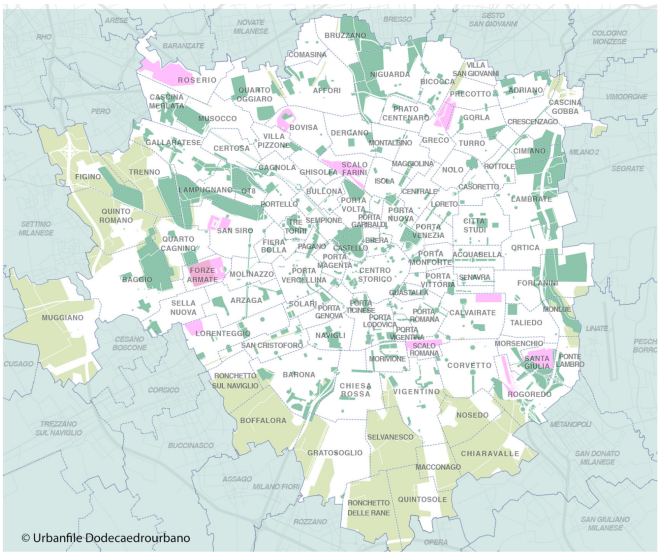

Figure 1: *Dark green areas represent usable public green spaces, light green areas represent non-usable green spaces, while pink areas indicate developing zones.*

In addition, the file `revenues.txt` collects the average daily revenues (`revenue` [k€]) in 2023 of 70 minimarkets located in Milan. For each minimarket i, the dataset also reports its UTM coordinates s_i and the resident population in the neighborhood (`population`(s_i)). For $i = 1, ..., 70$, consider the following model:

$$\texttt{revenue}(s_i) = a_0 + a_1 \cdot \texttt{population}(s_i) + \delta(s_i)$$

with $\delta(s_i)$ a stationary residual (spherical without nugget, with initial parameters 500 and 1000 for the sill and range, respectively).

g) Estimate via generalized least squares the parameters a_0, a_1 of the model, and briefly detail the implementation choices and procedure, reporting also the relevant R code. Report the estimated values for the variogram.

h) Provide a kriging prediction `revenue`$^*(s_0)$ of the revenues at a shop located in the Brera district at location $s_0 = (514703.8, 5035569.3)$, for which `population`$(s_0) = 6054.468$, briefly detailing the implementation choices.

Upload your solution **https://forms.office.com/e/sdXRu9xQmc**