Practicum on Areal Data Analysis

Alexandra M. Schmidt and Carlo Zaccardi

03-06, June 2025

Question 1

In this question you will use the 2012 North Carolina Presidential election data. The response variable is the second column, the percent of the votes for Obama. The remaining columns (obtained from the 2010 census and the NC county health site) can be used as predictors. The available files are: `Obama2012.csv`, `NCCentroids.csv` (contains the coordinates of the centroids of the counties and population size of each county) and `NCADJ.csv` (contains the 0-1 adjacency matrix in the same order as shown in the other files).

1. Identify a subset of the covariates to be used in the mean using a linear regression model with independent errors;
2. Plot the response and residuals on a map of NC divided by its counties;
3. Conduct Moran's test of autocorrelation using both the raw data and residuals using at least two choices of adjacency definitions (see function `knearneigh` in the `spdep` package);
4. Fit a CAR model for each weight definition. Compare the models using DIC and WAIC and for the "best" weights, compute the estimates and standard errors of the regression coefficients and compare these results with the OLS results;
5. Summarize your findings in a concise paragraph.

Question 2

The folder `DengueRio` contains the files to run an analysis of the number of cases of dengue across the neighborhoods of the city of Rio de Janeiro, Brazil. Rio de Janeiro has some islands and the code provides an example on how to handle the islands and build the neighborhood matrix. Run the file `Denguefit.R`.

Question 3 (Organized with the help of Victor Nogueira, UFMA, Brazil)

This question involves investigating if there is a spatial pattern on the distribution of the proportion of obese children (under 5 years old) across the microregions of the state of Bahia in the Northeast of Brazil. The available sample is composed of children whose families spontaneously sought the public health service closest to their homes and had their anthropometric measurements taken in 2018. The file `child_obesity_Bahia.txt` contains the following variables:

- Proportion of individuals in the microregion benefiting from the Bolsa Família Program (`bfp`), a public income transfer policy.Bolsa Família provides financial aid to poor Brazilian families. In order to be eligible, families had to ensure that children attend school and get vaccinated.

- Number of family health teams (`health teams`) in the microregion. These are groups of individuals with multidisciplinary training who work directly with the population.

- Proportion of illiteracy (`illiteracy`) in the microregion.

- Average per capita family income (`avg_per_capita_income`) in the microregion.

File `CreatingAdjMatrix.R` contains the commands to build a 0-1 neighborhood matrix using the shape file of Bahia through the package `geobr`.

- Carry out Moran's test for spatial autocorrelation using adjacency weights. Is there evidence of spatial autocorrelation in the proportion of obesity across the microregions? Does your conclusion change if you change the neighborhood matrix?

- Assume that the number of cases of each cancer follows a binomial distribution, that is,

$$\begin{aligned} Y_i &\sim Binomial(n_i, p_i) \\ logit(p_i) &= \alpha + \sum_{j=1}^{p} X_{ij}\beta_j + S_i, \end{aligned}$$

  where S_i is the random effect of microregion $i = 1, 2, \cdots, n$, $\mathbf{X}_{i.} = (X_{i1}, \cdots, X_{ip})'$ is a $p-$dimensional vector containing the covariates you decide to include in the model. Before fitting the models below, check the distribution of the covariates, consider centering them and transforming those which are skewed. You do not need to include all the available covariates in the model but you need to justify the ones that are included. If *you wish* to perform variable selection, one alternative is to use a spike-and-slab prior for the coefficients β_j. See Section 5.3 of the book *Bayesian Statistical Methods* by Brian Reich and Sujit Ghosh. Another alternative is to use a LASSO prior (see Section 4.2.3 of Reich and Ghosh (2019) on how to perform a LASSO regression under the Bayesian framework). Discuss which variables you decided to include in the model. Note that the spike-and-slab prior cannot be used in `Stan`, as it is not able to estimate discrete parameters. Once you have decided which covariates to include, fit the following models:

  1. independent random effect, $S_i \sim N(0, \sigma^2)$;
  2. CAR prior (with 0-1 neighborhood structure), $S_i \sim CAR(\tau^2)$;
  3. BYM model, such that $S_i = U_i + V_i$ with $V_i \sim CAR(\tau^2)$ (with 0-1 neighborhood structure), and $U_i \sim N(0, \sigma^2)$;
  4. BYM2 model, such that $S_i = \frac{1}{\tau}\left(\sqrt{1-\phi}U_i + \sqrt{\phi}S_i^*\right)$, where S^* is a scaled spatially structured component.

  Use WAIC and DIC to compare the fitted models and summarize your findings in a concise paragraph.