ST5226: Spatial Statistics

Project Assignment, AY 2025/2026

Due on Friday, 21 November, 2025

Instructions:

1. This project assignment consists of two parts. The total marks of this assignment is **60**.
2. Please upload your work in **1 single PDF file** to the Canvas assignment **"Project"**. As this is a group assignment, **only one submission is required per group**. The size of your file should be no more than **10Mb**. The deadline for submission is **11:59pm, Friday, 21 November, 2025**.
3. If you compile the files for Part I and Part II separately, you should merge two PDF files into one and submit one single PDF file.
4. If your group number on Canvas is XXX, please name your file

   Group XXX.pdf

   You can submit multiple times in Canvas. However, only the last submission will be marked.
5. Please write the **student numbers and names of all group members** on the first page of your file.
6. You should use either Markdown, Knitr, or LaTeX to compile your final file. You must ensure that your codes can be **copied and pasted** from the pdf file. **Screenshots/Pictures of programming codes are unacceptable and will be regarded as incomplete work.** If your codes do not produce the correct output, then your solution will be subject to mark deduction.
7. **No hard copy** will be accepted.
8. **No late submission** will be accepted (i.e., marks for your project assignment = zero).
9. You are encouraged to discuss with classmates or me if you have any questions. However, copying homework solutions is **strictly prohibited**.

Part I (30 marks): Analysis of Lip Cancer Data in Scotland

This part consists of several problems related to the data file `Scotland.rds`. The data set contains an `sf` object that consists of the following variables:

- `cancer`: number of lip cancer cases in the district
- `expected`: expected number
- `logratio`: log transformed ratio = log(Cancer/expected). The last two records have used a correction with Cancer = 0.5 because the recorded cancer counts are zero.
- `varlogratio`: the variance of the logratio, depends on the number of cases and the expected count
- `northkm`: Northing, in km, approximately centered so 0 is the location of Stirling
- `eastkm`: Easting, in km, approximately centered so 0 is the location of Stirling
- `percentAFF`: percent of the population in Agriculture, Forestry, and Fisheries.

Why is Stirling used as the center? Historically, Stirling served as the medieval capital of Scotland and lies close to the geographical center of the country. It's perhaps most renowned for the Battle of Stirling Bridge, where the Scottish forces led by William Wallace and Andrew Moray defeated the English army. This battle was famously depicted in Mel Gibson's 1995 film *Braveheart*, though, interestingly, the movie omits the bridge that played a critical role in the actual battle.

1. (15 marks) Explore the spatial autocorrelation.

   (a) (3 marks) Create three plots for the k-nearest neighbors of all regions, for the values of $k = 1, 2, 3$. For each of the three plots, are all the regions connected in the graph?

   (b) (3 marks) Consider the queen-style and rook-style neighbors based on contiguity. Identify all the regions whose queen-style neighbors and rook-style neighbors are different, by giving the names of these regions. Highlight these regions and their neighboring regions in the map. You should fill these regions with one color and their neighboring regions with another color.

   (c) (3 marks) Consider the count of lip cancers in the variable `cancer`. Compute and report the Moran's I statistic using the rook-style neighborhood and B-style weights. Then use the Moran's I to determine whether there exists positive spatial dependence for `cancer`, in each of the three tests: (i) normal test (without skewness correction); (ii) permutation test; (iii) Monte Carlo test by assuming that the count in each region follows a Poisson distribution with expectation given in the variable `expected`. Comment on your results.

   (d) (3 marks) Repeat Part (c) with Geary's c statistic. Do you see any difference in the results and your conclusion?

(e) (3 marks) Consider the variable `logratio`. Compute the local Moran' I statistics for all districts of Scotland, using the rook-style neighborhood and B-style weights. Plot a map of all districts in Scotland, and highlight those districts with significant local Moran'I on the map at the significance level 0.05. You should use different colors for the four types of significant districts: High-High, High-Low, Low-High, and Low-Low.

Liying

2. (15 marks) Modeling the areal data. One question of interest is the association between working outdoors and lip cancer (measured by `logratio`). The `percentAFF` variable quantifies the percent of the district population working in Agriculture, Forestry, or Fishing. All three occupations require large amounts of outdoor time. In addition, we also want to know whether there are any spatial trends in the north-south and east-west directions.

   (a) (3 marks) Fit a simple linear regression model, using `logratio` as the response variable, and using `percentAFF`, `eastkm`, and `northkm` as the predictors. Report the summary of model fit and determine which of these predictors are significant. Drop any insignificant predictor(s) from the model and refit the model with only significant predictors. Use plots to check if there is any violation of homogeneous variance assumption or the normality assumption.

   (b) (3 marks) Following the last model you have fitted in (a), refit it using weighted least squares by assuming that the individual variance is proportional to the reciprocal of the variable `expected`. Report the summary of model fit. Compare the results with those in (a). Which model is better?

   (c) (3 marks) Following (b), for the better model you have determined in (b), we check if it is necessary to fit a further model to account for spatial dependence in the residuals. Use `gls()` to fit a model such that the residuals are modeled by an exponential semivariogram. Report the summary of model fit. Compare the new model with those in (a) and (b). Determine which model gives the best fit to the data.

   (d) (3 marks) Following (a), for only the significant predictors in (a), fit the one-parameter SAR model using the B-style weights with rook-style neighbors and constant error variance. Report the summary of model fit. Use the permutation test based on Moran's I to determine if there exists positive spatial dependence in the residuals. Does the conclusion from this test agree with the test for spatial dependence in the summary of SAR model?

   (e) (3 marks) Repeat (d) for the one-parameter CAR model. Does your conclusion change regarding the spatial dependence in the residuals?

Part II (30 marks): Course Project

You are required to download a real-world spatial dataset and write a report, 10-15 pages in length, that includes a statistical analysis of the data. For groups with multiple members, **only one of you** needs to submit the report in Canvas. The detailed instructions are as follows.

(a) Your data must be real-world **spatial** data, specifically either **geostatistical data** or **areal data**. Using any other type of data will result in a score of zero for Part II of this assignment.

(b) Your dataset must fulfill the following requirements:

- It contains at least 1 response variable and at least 1 predictor variable. The response variable(s) must be spatially varying variable(s);
- It contains real geographical information (locations, or shapes of regions, etc.);
- It contains at least 50 observations, but no more than 1 million observations.

You are allowed to combine several real-data datasets into one dataset. For example, one dataset contains only the precipitation in different areas of Singapore, while another dataset contains the geographical boundaries of all districts in Singapore (such as the data used for Homework 1). You can combine them into one dataset and use it for subsequent analysis. **You do not need to submit your data.**

(c) **The first section** of your report must include a brief overview of the dataset, including the following information:

- The source of your data, and the type of the dataset (geostatistical or areal);
- The definition of all used variables in your dataset, the total number of observations, whether there are missing values, etc.
- The problem you want to investigate. You may want to check some background information of your data if necessary.
- An overview of statistical methods that you will use in the analysis.

(d) **The main body** of your report should include a comprehensive statistical analysis of the data. Your need to include the following:

- Simple descriptive analysis, such as maps showing the original variables in the dataset;
- Statistical models you want to apply. For regression or kriging models, please indicate clearly which variable is the response and which are the predictors. Your analysis should account for possible spatial dependence. You may use math formulas occasionally for clarity, such as an equation for some regression model. However, long math derivation will inevitably incur mark deduction because this is an applied project.
- Codes and outputs for the statistical analysis; your interpretation and discussion.

You should ensure that your analysis is logically sound and that every statistical model you propose is well justified. You may refer to the sample analyses in our lecture notes, tutorials, and homework assignments.

(e) You are strongly encouraged to use the various methods and techniques we have introduced in this course so far. You may also use other spatial models beyond our course materials that are suitable for your data, as long as you provide **sufficient justification**. For example, they significantly improve the model fitting or prediction performance compared to the baseline models in our lecture notes. Your work will be subject to mark deduction if such justification is missing.

(f) You will be heavily penalized for using or unnecessarily discussing any methods, models, or algorithms that are unrelated to spatial statistics.

(g) **The last section** of your report should be a brief discussion or conclusion that summarizes your findings.

(h) After the last section, please include the following **two short paragraphs**:

- **Members' Contributions**: 1-2 paragraphs describing the individual contributions of all group members. For example,

  "For Part I, Alice prepared the solutions for Question 1, and Bob prepared those for Question 2. All members took part in discussing the questions and checking the solutions.
  For Part II, Alice conceived the project and designed the methodology. Bob collected and preprocessed the data. Carol conducted the data analysis and produced the figures. David reviewed and verified the technical details. All members contributed to interpreting the results, discussing the findings, and writing the report."

  Since this project assignment consists of Part I and Part II, the contributions for both parts should be stated clearly. These paragraphs should accurately describe the specific roles and contributions of each team member. Please note that the level of detail in this statement will **not** affect individual scores. As a rule, all group members will receive the same mark unless a member has not made a fair contribution. In such cases, this should be stated clearly in the paragraph. If you disagree with the statement after the report has been submitted, please email me at `stalic@nus.edu.sg` with your reasons. I will then verify the details with all group members.

- **AI use disclaimer**: A brief statement disclosing any use of AI tools in the project.. For example, "ChatGPT-5 was used to assist in writing programming code and improving the structure and clarity of the report."

(i) After these two paragraphs, please include a list of all references (papers or books) cited in the main text, if any.

(j) **Report Format Requirement:**

- The report must be written in either Markdown, Knitr, or LaTeX, where the font size of your main text should be either 11 or 12;
- The total length of your report should be **from 10 to 15 pages** including everything. A report exceeding 15 pages will be subject to mark deduction. **Note**: A longer report within this range will not necessarily receive a higher mark.
- The report should include approximately 5 to 10 figures.

- The total length of code should be about 50 to 200 lines, including all headers such as `library()`, but excluding blank lines and spaces.

(k) Here are some useful sources for downloading spatial datasets:

- R CRAN Task View: Analysis of Spatial Data. Some R packages contain spatial datasets, such as `rnaturalearth`, `spData`, `SpatialDatasets`, etc. You can google and find more.
- GeoDa Data and Lab
- Singapore's open data portal
- NUS GIS webpage
- COVID-19 data, such as COVID tracking in US, COVID-19 Data Hub, etc.

You can also find spatial datasets from research papers. Please make sure that you cite the data source and the paper (if any) in the first section of your report.