STAT447: Project Proposal

Team: Youjung Kim (38762639) & Zelalem Araya (92797935)

Repo Link: https://github.com/eunicekim919/STAT-447c-project.git

Possible Datasets (min 2):

1. Countrywide car accident dataset over 49 states of the USA from Feb 2016 to Mar 2023.

This datum provides detailed information about where the event occurred such as street number/side/city/county/state/zipcode and latitude/longitude coordinates. This enables the calculation of distances between different traffic event locations. We can utilize the county-level accident counts to estimate unobserved statistics within the geographic area of interest.
US-Accidents: A Countrywide Traffic Accident Dataset - Sobhan Moosavi

2. Air Quality Index (AQI)

This dataset contains the AQI of most of the countries in the world.
However, it lacks specific information regarding the cities(or coordinates) where these AQI values were observed. To incorporate this data, it's necessary to select a representative point within each country and calculate the distance from the area of interest.
AQI - Air Quality Index

3. Vancouver Crime Data

This dataset includes the police-reported data on crimes committed in Vancouver for years up until 2023. It contains the type of crime, year, month, day, hour, minute, address generalized to the block level, and the neighbourhood.
VPD OPEN DATA

Project Themes: Bayesian Kriging vs. Classical Kriging, Spatial Interpolation, Cross-Validation
Potential Approaches:
Comparison of Spatial Interpolation Methods: Bayesian Kriging vs Classical Kriging Using a Cross-Validation Approach

To compare these two methods, we plan to use K-fold Cross-Validation* (or leave one out, depending on the size of our data). Dividing the data set, then holding out one fold as the test and conducting Bayesian & Classical Kriging to interpolate the region we left out spatially, and repeating this process K times.
As a performance metric, we will likely use RMSE or/and MAE.
Then aggregate the results to find the average performance of both methods to determine which is the better spatial interpolation method for this data.

*needs to account for spatial autocorrelation (somehow)

Contribution Plan:

To ensure an even contribution between the two of us, we will define clearly our work and responsibilities (rough preliminary example below) & keep a timeline to ensure the project is completed promptly. We plan on setting up weekly meetings to keep track of tasks being done independently & being open about challenges we face along the way. We also hope to regularly review each other's work to improve & keep accountable.

Tasks to be Completed:
- Literature Review (Both)
- Data Collection and Preprocessing (Youjung)
- Cross-validation Setup (Zela)
- Implementation of Classical Kriging (Youjung)
- Implementation of Bayesian Kriging (Zela)
- Results Visualization (Both)
- Writing the Introduction and Conclusion sections (Zela)
- Writing the Methodology and Discussion sections (Youjung)
- Formatting and Proofreading the Paper (Both)