School of
Information Systems

SMU
SINGAPORE MANAGEMENT UNIVERSITY

IS415 Geospatial Analytics and Applications
Take-Home Exercise (15%)
Spatial Clustering of Geographically Referenced Attribute: Singapore Public Bus Commuting Patterns Study

Overview

Since 2013, Singapore has an active Open Data initiative. It aims to enhance transparency, public participation, and collaboration in the nation. The nation has big data ambitions and believes that the bountiful pool of available data should be used to gain new insight that will improve economic welfare. With the inception of The Smart Nation vision in 2017, Open Data is seen as a necessary component of this initiative, especially in the promotion of public-private collaborations (co-innovation). However, most of the effort to date tend to focus on hosting online open data portal by various government agency such as data.gov.sg, SLA OneMap, URA SPACE and LTA LTA DataMall , just to mention a few of them. There are very little work on how to integrate information shared by these agencies to gain better understanding on national development or public services issues.

Objectives

In view of this, we are going to conduct a use-case to demonstrate the potential contribution of geospatial analytics in R to integrate, analyse and communicate the analysis results by using open data provided by different government agencies. The specific objectives of the study are as follow:

- Calibrating a simple linear regression to reveal the relation between public bus commuters' flows (i.e. tap-in or tap-out) data and residential population at the planning sub-zone level.
- Performing spatial autocorrelation analysis on the residual of the regression model to test if the model conforms to the randomization assumption.
- Performing localized geospatial statistics analysis by using commuters' tap-in and tap-out data to identify geographical clustering.

The Data

For the purpose of this study, the *passenger volume by busstop* data set of Land Transport Authority (LTA) will be provided. This data set is extracted using the dynamic API provided at LTA DataMall. You are required to obtain the remaining data from the government open data portal.

Grading Criteria

This exercise will be graded by using the following criteria:

1. **Geospatial Data Wrangling**: This is an important aspect of geospatial analytics. You will be assessed on your ability to employ appropriate R functions from various R packages specifically designed for modern data science such as readr, tidyr, dplyr, sf just to mention a few of them, to perform the entire geospatial data wrangling processes, including. This is not limited to data import, data extraction, data cleaning and data transformation. Besides assessing your ability to use the R functions, this criterion also includes your ability to clean and derive appropriate variables to meet the analysis need. (**Warning:** All data are like vast grassland full of land mines, your job is to clear those mines and not to step on them). (30 marks)

2. **Geospatial Analysis**: In this exercise, you are expected to use the spatial autocorrelation and localized spatial statistics methods and R functions introduced in class to analysis the geospatial data prepared. You will be assessed on your ability:
    a. to compute different spatial weight matrix and to understand their usage;
    b. to conduct spatial autocorrelation test on the residual of the simple regression model and interpret the test result;
    c. to calculate the local spatial clustering statistics and to interpret their analysis results; and
    d. to discuss the analysis results. (30 marks)

3. **Geovisualisation**: In this section, you will be assessed on your ability to communicate the complex spatial statistics results in business friendly visual representations. This course is geospatial centric, hence, it is important for you to demonstrate your competency in using appropriate geovisualisation techniques to reveal and communicate the findings of your analysis. (30 marks)

4. Bonus: Demonstrate your ability to employ methods beyond what you had learned in class to gain insights from the data. The methods used must be geospatial in nature. (10 marks)

Deliverables

- The project folder in a single zip file format. The project folder should consist of the followings:
  - The project sandbox in data sub-folder (all raw, intermediate and final data files)
  - An R Markdown file contains all code chunks used and the written statements.
  - A html report knitted from the R Markdown document.
- A published version of the report on RPubs.

Submission Instructions

- The final deliverable (e.g. R Markdown file, project sandbox and report) must be zipped in a single zip file format.
- Name the zip file according to the course code and assignment, for example: ***IS415_Take-home_Ex01***.
- The deliverable is to be submitted in softcopy. You are required to upload the zipped into the Dropbox of LMS before the stated assignment due date. Late work will be severely penalised. Students must check and confirm on LMS the assignment due date.

Due Date

10th May 2020 (Sunday), 11.59pm (midnight).