CEGE0042: Spatial-temporal Data Analysis and Data Mining

STDM Coursework 2023/24

Date set: w/c 09/01/2024

During this course, you learn how to use R Studio and a number of other software packages to explore, visualise, model, cluster, classify and forecast spatial, temporal and spatio-temporal data, using a variety of techniques including:

- Exploratory spatio-temporal analysis, visualisation and data processing
- Spatio-temporal autocorrelation analysis
- Clustering
- Statistical space-time modelling
- Machine Learning (Kernel Methods (SVMs), Artificial Neural Networks, Random Forests)
- Agent based simulation

In this coursework, you will use the skills you have gained to analyse and model a new dataset. The deadline for submission is **Friday the 22nd March, 2024 at 5pm.** Reports should be submitted **online via Moodle**.

Your task is to source and analyse a spatio-temporal dataset using the methods you have learned during the course. Depending on which dataset you choose, you may use different methods to analyse it. The requirement for the dataset is that it is geolocated and time-stamped.

Some examples:

Crime Location Data

Crime locations are usually recorded as point (event) data. Some options for analysing these data include:

1. Identifying crime clusters/hotspots using different clustering methods
2. Aggregating crimes into spatial units (e.g. census geography, postal units) and predicting crimes at the level of the spatial unit.

Example data source: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Road Traffic Data

Traffic data (flows, travel times etc.) are usually recorded on road segments, which form a spatial network. Using the adjacency of the network, carry our short-term prediction of traffic flows. Some options to explore:

1. Do machine learning or statistical methods perform better at short term traffic forecasting? You could test this by comparing 2 or more methods.

Example data source: https://dot.ca.gov/programs/traffic-operations/mpr/pems-source

Covid-19 pandemic data

CEGE0042: Spatial-temporal Data Analysis and Data Mining

Carry out spatio-temporal analysis or forecasting of Covid-19 pandemic spread. Covid-19 data is available from a range of places including:

- Kaggle: https://www.kaggle.com/datasets?search=covid

Example projects

1. **Use a data driven approach to predict case numbers by country/region**
2. **Construct an agent-based model of spread**

The datasets and tasks suggested here are just examples. You are encouraged to search for datasets and choose a topic you are interested in. There are various places you may find data such as government websites and repositories such as Kaggle (https://www.kaggle.com/datasets).

Your task is to produce a report with the following sections:

1. Introduction and data description (10 marks) – Provide an outline of the experiment, including a brief literature review of the methods being used. Describe the data and visualise it using some of the methods you have learned.
2. Exploratory spatio-temporal data analysis (20 marks) – Use some of the methods you have learned to analyse the spatio-temporal patterns in the data. This could include autocorrelation analysis, density estimation etc. depending on the nature of the data.
3. Methodology and results (40 marks) – This part should contain:
    - A brief description of the method used to analyse the dataset.
    - A detailed explanation of the experimental setup (e.g. the way the data were divided, the parameters that were used, the transformations that were used, i.e. differencing).
    - Presentation of the results with appropriate graphs and/or maps.
    - An assessment of the performance of the method (with error indices or other appropriate measures).
4. Discussion and conclusions (20 marks) – Discuss/compare the results of your models.
    - If you used multiple models, did one model perform better than the other? If so, why might this be the case? What are the strengths the model(s) in terms of interpretability and ease of implementation, running time etc.?
    - How did the performance of the model vary across the study area?
    - What were the limitations of the method(s) used?
    - How could the method(s) be improved?
5. Reproducible code (10 marks) – Your code should run the entire workflow and reproduce the results in your report. You should make your code available for testing and provide instructions on how to test it.

Report Length

The report does not have a word limit but is limited to 6 pages A4 with Arial font size 11, including tables and figures but **excluding** references. This is a common requirement when writing short papers, e.g. for an academic conference or journal. You should divide the content of your report among the sections according to the proportion of marks available for each one.