CEGEG076: Spatio-temporal Analysis and Data Mining

STDM Coursework 2016

Introduction

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

- Exploratory spatio-temporal analysis and visualisation
- Autocorrelation analysis
- Clustering (Scan Statistics and k-means)
- Statistical space-time modelling (ARIMA and STARIMA)
- 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. **Two of you should work as a team and deliver a joint report**. The mark for each individual of a joint report might be different due to the quality of the analysis conducted by the individual (refer to Task 3.3). The deadline for submission is **Friday the 1st of April, 2016 at 5pm.** Reports should be placed in the box by the general office. You have two options:

Option 1: Forecasting of urban travel times

In this task, you will use the skills you have gained to evaluate The data are travel times (TTs) collected using automatic number plate recognition (ANPR) technology on London's road network in 2011, as part of Transport for London's (TfL's) London Congestion Analysis Project (LCAP). The cameras operate in pairs called links: as a vehicle passes the first camera of the link its license plate is read and the time is recorded. The vehicle then traverses the link and the time is recorded again when it passes the second camera. Individual TTs are aggregated at 5 minute intervals to give 288 observations per day. The network is shown in figure 1. The data here comprise data collected between 6am and 9pm (180 observations per day). You have been provided with a subset of **30 days of data** for this tutorial, across **256 road links**. The data have been pre-processed into unit travel times (seconds/metre) and are stored in an R workspace called `UJTWorkSpace` on Moodle. Contained in the workspace are:

- **UJT:** The data matrix, with one road link per column. The column names are the road link IDs.
- **dates:** The dates of data collection. The first column is the date and the second column is the time of day, from 1 (6am) to 180 (9pm). This is the order in which the data appear in UJT.
- **LCAPAdj:** The adjacency matrix, where 1 indicates two links are adjacent. Note that the network is not fully connected as some links have been removed due to poor data quality.
- **LCAPShp:** A shapefile of the data, which can be quickly visualised using `plot(LCAPShp)`. This shapefile contains all links (1402) so you may want to extract/identify those links for which you have data. The IDs correspond to the `LCAP_ID` field in `LCAPShp@data`.

In your pairs, select and extract a subset of the network (e.g. ~15-30 road links) that you would like to work on. This can be the area around UCL, your home location, or perhaps a road that you

know to have problems with congestion. How you select the links is up to you (i.e. manually or by bounding box).

Task 1 (Joint task) – Exploratory spatio-temporal data analysis

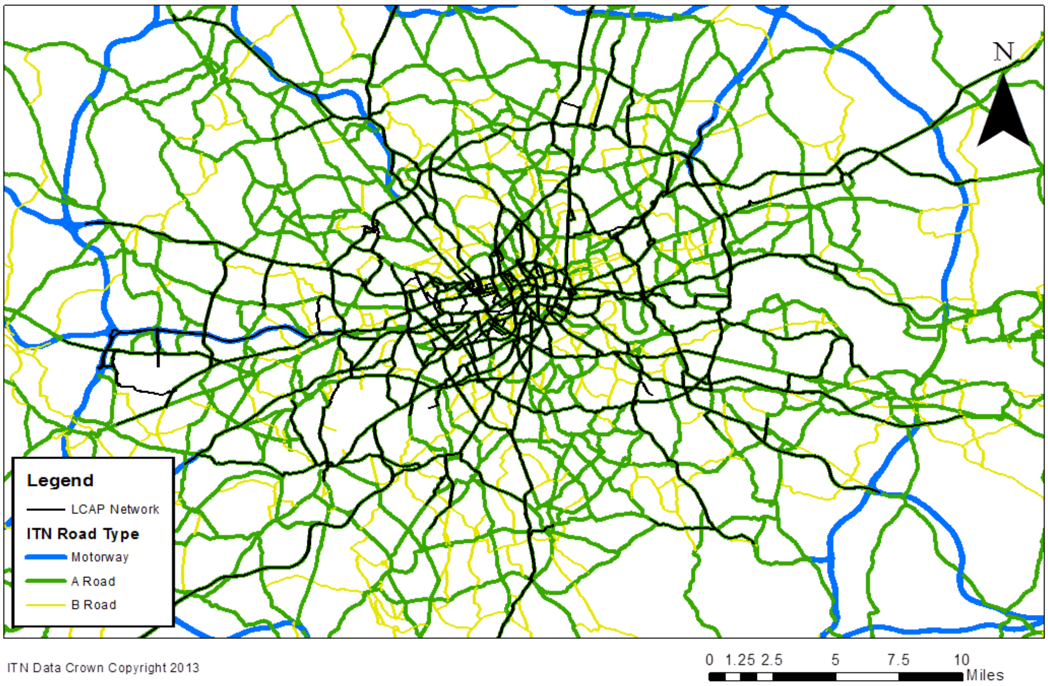
Use the space-time autocorrelation function and partial autocorrelation and/or other relevant techniques to investigate the spatio-temporal patterns in the data. Describe any patterns that you find and present them using some of the visualization methods that you have learned.

Task 2 (Individual task) – Space-time forecasting

Each of you should choose one of the forecasting algorithms you have learned (make sure you don't choose the same one). Alternatively, **you may choose another forecasting algorithm** that can be accessed from within *caret*, provided you can justify your choice. For the two chosen methods, your task is to use the first 23 days of the training data to forecast the final 7 days. **You do not have to use all of the training data.** You may want to*:*

- Try a number of parameter combinations to ensure you find a good model.
- Experiment with different training data lengths to see what length provides a good balance of training time vs. accuracy.
- Try different ways of incorporating spatio-temporal information in the model
- Try transforming the data (e.g. differencing) to see if it has an effect.

How you make use of the training data is up to you, but you must justify the decisions you have made. Remember that if you use STARIMA, you must first attempt to **transform the data to stationarity** and document how you achieved it.



ITN Data Crown Copyright 2013

Figure 1 – Map of the LCAP network in London

Task 3 (Joint task) - Report

CEGEG076: Spatio-temporal Analysis and Data Mining

You will be assessed based on a 4000-5000 word report of your experiments. This should include the following sections:

1. Introduction and data description (Joint – 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 (Joint – 20 marks) – Present the results of task 1 including:
   - Appropriate ST-ACF, ST-PACF or other plots.
   - Explanations of the patterns that you discover.
3. Methodology and results (Individual – 40 marks) – This part should contain two distinct sections, one for each method, which are marked individually, containing:
   - A brief description of the method.
   - 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. You may use the root mean squared error or r squared index or other measures you deem appropriate to measure the accuracy of the results. Make sure you are consistent.
4. Discussion and conclusions (Joint – 20 marks) – Compare the performance of the two models and discuss the results.
   - Did one model perform better than the other? If so, why might this be the case?
   - What are the relative merits of each of the models in terms of interpretability and ease of implementation, running time etc.?
   - How did the performance of the models vary across the study area?
   - How could the methods be improved?

A further 10 marks are available for quality of visualisation and presentation of data and results. You may use any of the R packages that have been introduced in the practical sessions. Feel free to use any other R packages that you may find if you feel they will improve the quality of your presentation.

Option 2: Define your own project

The second option is to define your own project. If you decide to do this task, you must source your own spatio-temporal dataset (online or elsewhere) and design your own experiment. You can use any of the techniques you have learned, depending on the dataset you choose. For example, if you would like to base your analysis on clustering, you could find a crime dataset such as this: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2. Working in pairs, you will each choose a technique to analyse the dataset. Your task is to produce a report with the following sections:

1. Introduction and data description (Joint – 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 (Joint – 20 marks) – Use some of the methods you have learned to analyse the spatio-temporal patterns in the data.

3. Methodology and results (Individual – 40 marks) – This part should contain two distinct sections, one for each method, which are marked individually, containing:
   - A brief description of the method.
   - 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 (Joint – 20 marks) – Compare the results of the two models and discuss the results.
   - Did one model perform better than the other? If so, why might this be the case?
   - What are the relative merits of each of the models in terms of interpretability and ease of implementation, running time etc.?
   - How did the performance of the models vary across the study area?
   - How could the methods be improved?

If you choose this option, you must submit the data you used via Moodle for assessment.