Coursework Report Template for Module INM433 “Visual Analytics”

Zacharias Detorakis

**Abstract**—Put here a brief summary of your work: analysis task, data, approach, main findings. Length: up to 200 words.

# Problem Statement

The objective of the analysis presented in this paper is to study the crime rates in London (i.e. by ward and by borough) overtime and try to see if there is any correlation with other characteristics of census data within the same geographical areas. Therefore, the key questions are:

* Is there a pattern in the increase/decrease of the crime rate overtime?
* Do these patterns, if they exist, have a spatial collocation (i.e. neighbouring boroughs following the same patterns of increase or decrease)
* Are crime rates correlated with other indices of the studied areas (i.e. demographics)?

The study spans a period of 9 years (2010-2018). In April 2010 there was a change in the ward code boundaries therefore the data available before 2010 are not considered to avoid skewing the results. The key socio-economic characteristics used for the analysis against the crime rates are:

* Gender: i.e. the makeup of the areas in terms of male, female population
* Employment: i.e. the employment status of the residents in those areas also broken down by gender and type of employment

The data collected for the indices that will be accessed for their impact in the crime rate are at borough rather than ward level so human reasoning supported by visual representation will be required to see if it is meaningful to group wards by borough for the purposes of this analysis.

# State of the Art

First paragraph...

Following paragraphs...

*<500 words*

# Properties of the Data

The studied area in this research is Greater London which is a ceremonial county of England that makes up the majority of the London region. Greater London is divided into 32 local authority districts henceforth mentioned as *boroughs* each of which comprises multiple smaller subdivision henceforth mentioned as *wards*.

The data that is used in this analysis have been collected for the Greater London region over the course of the 9 years (2010-2018) and comprise the following datasets

**Crime Data**: Crime data was sourced from the metropolitan police and have been geographically aggregated at ward level (606) and subsequently at borough level (32). The timeseries have also been aggregated into monthly figures 105 instead of (12x9 = 108) months given that the oldest datapoints available are from April-10 instead of Jan-10. The crime records for each ward/month are broken down further by a major and minor classification of the crime as defined by the Home Office. There are 9 major crime classifications (Burglary, Criminal Damage, Drugs, Fraud or Forgery, Other Notifiable Offences, Robbery, Sexual Offences, Theft and Handling, Violence Against The Person) and 32 minor categories. Although data is not available at a lower grain (i.e. individual records per crime with locale and datetime information) the aggregation is still at a level that is useful for the analysis

**Census Data**: Census data have been gathered from the annual population survey dataset conducted by Nomis, a service provided by the Office for National Statistics. The data covers the same time period of 8 years and geography of greater London but this time the aggregation is at a higher level. The spatial grain is now at borough (32) instead of ward level and the timeseries are aggregated by year (9) instead of by month. The spatial aggregation may hinder local effects of census data and human reasoning is necessary to decide which boroughs to study (i.e. those with similar criminal patterns among their wards). The variables source from this dataset are:

* Unemployment Rates by gender and by age group (16-19, 20-24, 25-34, 35-49 and 50+)
* Population by nationality and ethnic minority: UK national vs non-UK national, white vs ethnic minority
* Educational Level: no qualifications, other qualifications, GCSE grades A-C or equivalent, GCE A level or equivalent, higher education below degree level, degree or equivalent and above

Given that the datasets are from official sources that use the same information to inform operational and other KPI reports the datasets are relatively clean and complete.

INSERT FIGURE

However, in order to gain useful information as part of the analysis we had to merge the two datasets. This was done by initially aggregating the crime data to the same grain per borough / per year as the census data and then using that borough+year combination as the unique key between the two.

Created additional data for increase or decrease

# Analysis

## Approach

As part of the analysis we need to combine multiple datasets (i.e. crime and census data), perform the analysis on temporal and spatial dimensions, aggregate data based on the different grains of the sets and arrive to conclusions on crime patterns and their correlation to census data. The following diagram illustrates the steps of this approach.

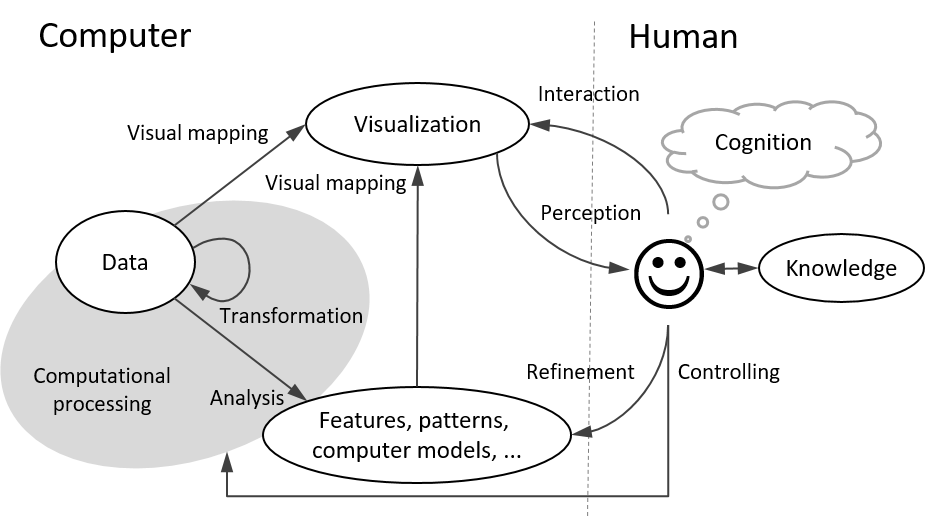


Fig. 1. An example of including a diagram in the document.

Crime Data Preparation: This step includes all data processing needed to transform the data into a format that can be used for downstream analysis and visualisation. We need to merge the dataset with the crime rates with another dataset with the population in the same grain (spatial and temporal). Following the merge, we need to address any missing values by either excluding records from the set or through data imputation.

Visualise Temporal: We will aggregate data over the entire region and observe the temporal effect on the dataset. That is expected to give us some useful insights on the overall trends of the various major crime categories overtime and perhaps inform a decision of which categories we can dive into more details with.

Visualise Spatial: We will aggregate the data over a selected period(s) to identify any spatial patterns of the selected crime categories. This is important since the census data is at a higher grain therefore, we need to only consider boroughs which comprise wards with similar crime characteristics. Selecting wards with high variance between the wards might result in masking of specific correlations. CLUSTERING???

Universe Selection: At this point, added by the visual artefacts created as part of the analysis we will confirm the selection of the study universe. In particular:

* Which major crime categories to consider overtime and
* Which areas (i.e. boroughs) to consider

Census Data Preparation: Similar to the first step in the process but this time with the census dataset. In order to merge census and crime data we need to group the later by year and by borough since this is the least common denominator in the two datasets. We will also need to the merged set to the area and crime categories decided in the previous section before proceeding with the analysis

Correlation Between Crime and Census Data: Finally, we will need to visualise the correlation between the different types of census indicators to see if there is a relation between them and the increase or decrease of the crime rates overtime. Visualization will help highlight where a relationship between an indicator and a crime rate is strong or the two are relatively independent.

## Process

First paragraph...

Following paragraphs...

*<1500 words, <=7 images*

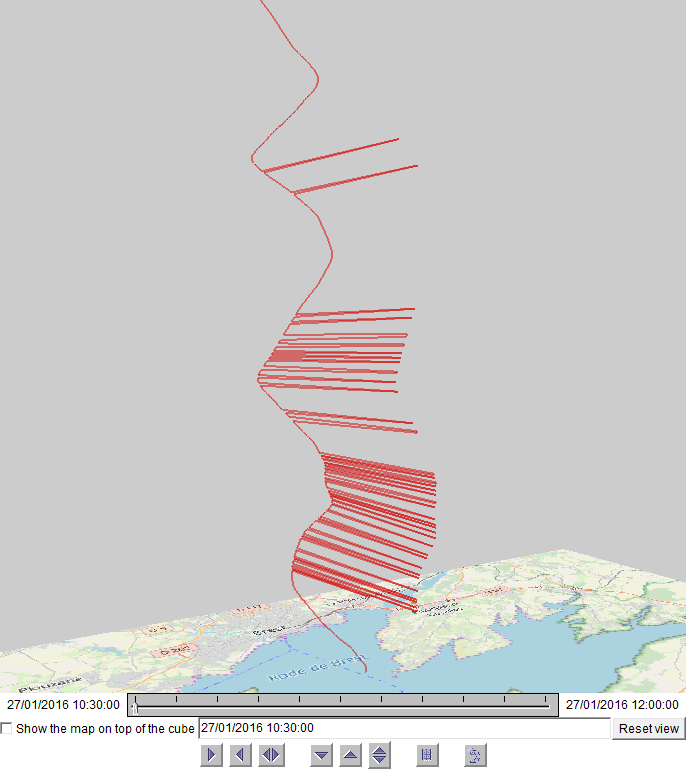


Fig. 2. An example of including a screenshot in the document.

## Results

First paragraph...

Following paragraphs...

*<200 words, <=2 images*

# Critical reflection

First paragraph...

Following paragraphs...

*<500 words*

Table of word counts

|  |  |
| --- | --- |
| Problem statement | 220/250 |
| State of the art | 500 |
| Properties of the data | 500 |
| Analysis: Approach | 500 |
| Analysis: Process | 1500 |
| Analysis: Results | 200 |
| Critical reflection | 500 |

References

The list below provides examples of formatting references.

1. M.Ankerst, M.Breunig, H.-P.Kriegel, J.Sander. OPTICS: Ordering points to identify the clustering structure. In: *ACM SIGMOD* 1999, pp. 49–60.
2. M. Bögl, W. Aigner, P. Filzmoser, T. Lammarsch, S. Miksch, and A. Rind. Visual Analytics for Model Selection in Time Series Analysis, *IEEE Trans. Visualization and Computer Graphics*, 19(12): 2237-2246, 2013.
3. T.F. Cox, M.A.A. Cox. *Multidimensional Scaling*. Chapman and Hall, 2001.
4. S. van den Elzen and J.J. van Wijk. BaobabView: Interactive construction and analysis of decision trees, *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST’11)*, pp. 151-160, 2011.
5. M.Harrower, C.A.Brewer: Colorbrewer.org: An online tool for selecting color schemes for maps. *The Cartographic Journal* 40(1): 27–37, 2003.
6. T. Mühlbacher and H. Piringer. A Partition-Based Framework for Building and Validating Regression Models, *IEEE Trans. Visualization and Computer Graphics*, 19(12): 1962-1971, 2013.
7. D. Phan, L. Xiao, R. Yeh, P. Hanrahan, T. Winograd. Flow map layout. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis 2005)*, pp.219-224, Oct. 2005.
8. J.W. Sammon. A nonlinear mapping for data structure analysis. IEEE Transactions on Computers, 18(5): 401–409, May 1969.
9. J. Sewall, D. Wilkie, and M.C. Lin. Interactive Hybrid Simulation of Large-Scale Traffic, *ACM Transactions on Graphics*, 30(6), Article 135.
10. N. Willems, H. van de Wetering, J.J. van Wijk. Visualization of vessel movements. *Computer Graphics Forum*, 28(3): 959-966, Jun. 2009.