An Exploration of Crime Prediction Using Data Mining on Open Data

Article *in* International Journal of Information Technology and Decision Making · May 2017

DOI: 10.1142/S0219622017500250

2 authors:

Ginger Saltos
University of Portsmouth
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Mihaela Cocea
University of Portsmouth
119 PUBLICATIONS **800** CITATIONS
SEE PROFILE

International Journal of Information Technology & Decision Making

AN EXPLORATION OF CRIME PREDICTION USING DATA MINING ON OPEN DATA

FIRST AUTHOR
University Department, University Name, Address
City, State ZIP/Zone, Country

SECOND AUTHOR
Group, Laboratory, Address
City, State ZIP/Zone, Country

Received Day Month Year
Revised Day Month Year

The increase in crime data recording coupled with data analytics resulted in the growth of research approaches aimed at extracting knowledge from crime records to better understand criminal behaviour and ultimately prevent future crimes. While many of these approaches make use of clustering and association rule mining techniques, there are fewer approaches focusing on predictive models of crime. In this paper we explore models for predicting the frequency of several types of crimes by LSOA code (Lower Layer Super Output Areas – an administrative system of areas used by the UK police) and the frequency of anti-social behaviour crimes. Three algorithms are used from different categories of approaches: instance-based learning, regression and decision trees. The data are from the UK police and contain over 600,000 records before preprocessing. The results, looking at predictive performance as well as processing time, indicate that decision trees (M5P algorithm) can be used to reliably predict crime frequency in general, as well as anti-social behaviour frequency.

Keywords: Crime prediction; Data mining; Open data; Regression; Decision trees; Instance-based learning

1. Introduction

Crime data has been systematically recorded by the police for many years and in the last decade there has been a surge of Open Crime Data[1] and of apps and/or web-based applications displaying crime statistics on maps, both by official sources, such as from Police UK[2], and other sources using the same official data. For example, on the `data.gov.uk` website, there are 45 apps listed from a variety of sources which give statistics and maps about crime in the UK[a].

These data can be used to support decision making by the police, marketing agencies and the government. Some basic statistics are already in use as part of

[a]This information was retrieved on 7 May 2015; the number of apps may have changed since then.

Geographical Information Systems (GIS), which can display, for example, the number of crimes in an area, a breakdown by types of crime and the location(s) of the crimes. All the apps mentioned above make use and/or display these statistics. A particular approach has been found to be useful by the police, which is the identification of crime 'hot spots'[3], which indicate areas with a high concentration of crime. The main argument for identifying hot spots is that particular areas have disproportionate numbers of crimes[4], an aspect which has been repeatedly supported by research evidence[5,6,7].

In addition, research evidence about the risk factors for a variety of crimes[8,9,10,11], about resilience and protective factors[12], as well as economical factors[13,14], have put a greater emphasis on crime prevention[15].

Predictive modelling can support decision making for resources allocation in terms of prevention strategies as part of the wider strategy, as well as in terms of management of perceived risk in the communities[16].

Most GIS systems facilitate spacial analysis through visualisation and the use of map layers (of which hot spots are one), but provide limited tools for analysis[17]. The limited predictive capabilities of the GIS tools are methods from earth sciences and economics[17], which may not be the most useful in creating predictive models of crime.

In this paper we investigate several predictive models of crime and discuss their applicability. Building on the idea of hot spots, we investigate the prediction of frequencies of different crime types per month and per LSOA code (Lower Layer Super Output Areas), which is an administrative system of areas used by the UK police. The choice to focus on the LSOA codes was made to facilitate decision making, since these are the administrative areas the police already work with. We also explore if information about the postcode of a location has an influence on prediction, and whether building separate models for particular crimes that are frequent, e.g. anti-social behaviour, leads to better predictions.

The rest of the paper is organised as in the following. Section 2 outlines previous research on crime data analysis, including predictive modelling. Section 3 describes the data used in our research and outlines the our methodological approach. The experiments and results are presented in Section 4, while Section 5 discusses the results and their wider implications, and concludes the paper.

2. Crime Data Analysis

Spatial analysis of crime has grown in the last decade. One of the most popular approaches is hot spot analysis, e.g.[18,19,20,21]. Some of the most popular approaches used for this purpose are point pattern analysis[22,23] and clustering/ distance statistics[24,25,26]. Another popular approach is the discovery of patterns or trends through various techniques from data mining and knowledge discovery research[27], such as association rule mining[28], text mining and spatial analysis[22], and self-organising maps[29].

An overview of research in this area is given in Table 1, which includes the authors, the techniques used, information about the data (if provided) and a brief description of the research conducted.

Table 1: Crime data analysis research

Authors	Techniques	Data	Description
Andersen and Malleson[23]	Spatial point analysis	Records of over 3 years	Investigate crime displacement by identifying changes in the spatial patterns/ distribution of crime
Bachner[30]	Clustering, Social Network Analysis	Not Applicable	Overview of predictive modelling
Brown and Hagen[28]	Association rule mining	39 records (cases)	Tool for discovering associations between different crimes
Chen et al.[31]	Co-occurrence, hierarchical clustering	120 records for identity detection; 272 records for network analysis	Deceptive identity detection, criminal network analysis
Dahbur and Muscarello[32]	Kohonen neural networks and heuristics	Not mentioned	Discover patterns of serial crimes; case study on armed robberies
Grubesic[25]	Fuzzy clustering	613 records	Hot spot detection
Helbich et al.[22]	Text mining and Spatial point analysis	200 individual information packages (i.e. emails, transcribed interviews and phone calls)	Text mining and spatial analysis to discover new patterns and relationships
Li et al.[29]	Fuzzy self-organising map, Rule extraction	6720 records; 14 crime types	Data analysis to support decision making; 4 crime trends: typical, gradual increase, sharp increase and wintertime
Lin and Brown[33]	Clustering and outlier-based approach	170 records	Association of incidents for identification of crimes committed by the same individual; case study on robbery data

Malathi and Baboo[34]	DBScan clustering (density based), k-means clustering, Decision trees (C4.5)	8 years of crime data	Clustering of crime data by crime type and prediction of crime frequency for the following year
Murray and Grubesic[26]	Non-hierarchical clustering using spatial lag	848 records	Hot spot detection
Nath[35]	k-means clustering	309 records	Patterns of crime (6 types)
Oatley and Ewart[24]	Logistic regression, neural networks, Bayesian Network	70,000 records	Analysis and prediction of burglary data
Phillips and Lee[36]	Graph similarity	Not mentioned	Identification and description of crime patterns
Wang et al.[37]	Series Finder (supervised learning for detecting patterns)	4855 records	Identification of patterns in housebreaking crimes (from 51 patterns)
Xue and Brown[17]	Discrete Choice Theory and Clustering	Over 1200 records	Analyse and predict spatial choice of criminals for residential breaking and entering crimes
Yu at al.[38]	k-NN, Decision trees (J48), SVM, Neural Network, Nave Bayes, ensemble learning	Not mentioned	Prediction of burglary data with different levels of aggregation of historical data (1 month to 10 months)
Zubi and Mahmud[39]	k-means clustering, association rules	350 records	Crime analysis of Libyan crime data

A more recent research trend is the development of predictive models due to the emphasis on crime prevention[15], which is also the aim of the research presented in this paper. Consequently, the research works in this area are described in detail in the following.

Malathi and Baboo[34] used a classification technique (decision trees) to predict crime trends (out of 4 options) for the following year. They also describe the prediction of the numbers of crimes for a particular year using data from the previous 8 years, although it is not clear what method was used for this numeric prediction. In terms of the data used, no number of records is given; they mention that the data covers 9 years of crime information. Due to the very brief description, it is not clear how this work can be replicated by other researchers. Most notably: (a) the data is not described in detail, i.e. the features/attributes used are not listed and the number of records is not specified, and (b) it is not clear how they converted a categorical output from a decision tree classification algorithm to a numerical one. Unlike this approach, we use numerical prediction models and the data we used is described in detail, both in terms of features and number of records.

Another approach by Oatley and Ewart[24] focused on the prediction of likelihood of repeated burglary for a particular property. For this purpose, they used a Bayesian belief network, using the following features or attributes: offender features; modus operandi[b] features; property stolen; premise crime history; prevalence, incidence and concentration, which are numeric indicators of the distribution of crimes over an area. They used 70,000 records of burglary-related crimes, including motor vehicle theft, street robbery and burglary from dwelling houses. The focus of this research was the development of the software and the paper does not describe any evaluation of the proposed approach. In terms of the Bayesian belief network, the focus is on the interpretability of the output rather than the performance of the method, which is mentioned as part of future work. In contrast, our approach focuses on the evaluation of prediction models, both in terms of their predictive performance, as well as their complexity, as an important practical aspect that is relevant for large volumes of data.

Xue and Brown[17] developed an approach for the prediction of future crime locations based on discrete choice theory and clustering. The spatial choice model they developed combines the predicted probabilities for all clusters for an overall prediction in a particular area. In terms of data, they used over 1200 crime records. They compared their proposed approach with a traditional hot spot identification method and found that their models outperform the traditional ones. The authors argue against aggregating individual crime records; however, they do not discuss options for handling vast amounts of data without aggregation. Unlike this approach, we use a large volume of data, i.e. over 600 000 crime records, for which analysis without aggregation would take too long and would, thus, not be useful in practice. Moreover, we discuss practical aspects related to processing time for large amounts

[b]A set of habits that an offender follows.

of data.

A classification approach has been used by Yu at al.[38] to classify areas into hot spots and cold spots, and to predict if an area will be a hot spot for residential burglary. They defined a hot spot as an area with at least 1 crime. They experimented with different levels of aggregation of historical data (1 month to 10 months), and a variety of classification techniques: k-Nearest Neighbor (k-NN), Decision trees (J48 algorithm), Support Vector Machines (SVM), Neural Network, Naive Bayes and ensemble learning. They found that the best results were obtained with the 1-nearest neighbour and the neural network algorithms. They did not mention the number of records, but they trained the models on 11 months of data and tested them on 1 month of data. The authors acknowledge that predicting an increase/decrease would be more useful than hot/cold spots and include this in their future work, along with exploring other types of crimes. Our proposed numerical prediction models aim to address this limitation, as outlined below.

Unlike previous research, we focus on the prediction of crime frequency as a numeric value rather than as a label (hot/cold spot), because the definition of a hot spot may vary according to: (a) area – 1 crime in a low-crime area may constitute a hotspot, while 10 or more crimes may be considered as a hotspot in a high-crime area; (b) crime type – some crimes, e.g. anti-social behaviour, are much more frequent than others such as armed robbery, and thus, hotspots for different types of crimes need to be defined proportionately to their frequency. A numeric prediction would output a number (rather than a label), which can then be interpreted in context.

Our proposed approach also uses a large number of records and discusses the time required to build and test prediction models based on such large volumes of data – an aspect that has not been addressed in previous research, but is very important in today's context of large amounts of data available and the practical issues involved in their analysis.

3. Data and methodology

In this section the data that was used in our experiments is described in detail. The methodology is also described, including a brief outline of the algorithms employed in our experiments, as well as the evaluation metrics used.

3.1. *Data*

The data used in this research comes from `data.police.uk`, an website for open data about crime and policing in England, Wales and Northern Ireland. These data started to be released in December 2010 and are updated on a monthly basis. The data are originally reported by each police force and go through a rigorous quality control process before being published. This quality process involves format validation, automated testing, and manual verification and approval by two people.

Table 2. Description of the dataset's features.

Name	Type of Data	Description
Crime ID	Nominal	Id of Crime.
Month	Nominal	Date of the crime in the format yyyy-mm.
Reported by	Nominal	The force that provided the data.
Falls within	Nominal	Same as "Reported By".
Longitude	Interval	Anonymised longitude coordinate of the crime.
Latitude	Interval	Anonymised latitude coordinate of the crime.
Location	Nominal	Specific or near location of the crime.
LSOA code	Nominal	Code of the Lower Layer Super Output Area (LSOA) where the crime was committed.
LSOA name	Nominal	Name of the LSOA where the crime was committed.
Crime type	Nominal	16 types of crime according to Data.police.uk (n.d.).
Last outcome category	Nominal	A reference to whichever of the outcomes associated with the crime occurred most recently.
Context	Nominal	Additional data.

Furthermore, the UK Police Department also explains the known issues, and how they are solving them, such as location accuracy, court result matching, double counting of anti-social behaviour and crime, constantly changing data, and missing outcome data[40].

For the purpose of our experiments we focus on the Hampshire Constabulary and data from December 2010 to March 2014 (40 months). Table 2 describes the attributes/features of the dataset. On all our data the "Reported by" and "Falls within" attributes have the value "Hampshire Constabulary"; the documentation mentions that although these attributes are currently the same, the "Falls within" attribute will change in the near future. The "Location" attribute provides a description of the location of the crime in relation to a reference point, such as a road (e.g. A2030, Andover Way) or a point of interest (e.g. Shopping area, Supermarket, Parking area). The attributes related to LSOA refer to the Lower Layer Super Output Area (LSOA) that the anonymised point falls into, according to the LSOA boundaries provided by the UK Office for National Statistics. For Hampshire Constabulary there are 1454 unique LSOAs.

The crime type is one of the 16 categories used by the police, which are listed in Table 3 in descending order of frequency on the Hampshire Constabulary data used in our research. The "Last outcome category" has options such as: under investigation; unable to prosecute suspect; investigation complete – no suspect identified; offender given warning; offender fined, etc. The "Context" attribute is a textual description of the context of crime; on recently published data, this is always empty.

An instance is one data object or record, which is characterised by the attributes described above. The data is released in monthly datasets, where each row is an instance, i.e. one crime.

The first step in our data analysis was to aggregate the individual monthly datasests into one dataset, which was further used in our experiments (details are given in the next section). Table 4 describes the dimensionality of the data and

Table 3. Types of crime

No	Crime type	No of records
1	Anti-social behaviour	44,070
2	Burglary	22,081
3	Criminal damage and arson	21,333
4	Violent crime	19,673
5	Other theft	19,538
6	Vehicle crime	18,260
7	Other crime	14,684
8	Drugs	8,836
9	Shoplifting	8,318
10	Violence and sexual offences	5,956
11	Public disorder and weapons	5,266
12	Public order	2,658
13	Bicycle theft	2,323
14	Robbery	2,166
15	Theft from the person	799
16	Possession of weapons	459

the number of missing values for the aggregated dataset. It contains over 600,000 records and if there were no missing values, the total number of values would be 7,313,016 (number of records multiplied by the number of attributes). Most of the missing values are from the last two attributes (i.e. "Last outcome category" and "Context"), and some from the "Crime ID" attribute. A small number of instances, i.e. 46, also have missing values for the "LSOA code" and "LSOA name" attributes, meaning a total of 92 missing values.

Table 4. Summary of the dataset.

Data Objects:	609,418
Attributes:	12
Values:	5,899,452
Missing Values:	1,413,564
% of Missing Values:	19%

3.2. *Methodology*

In our experiments, we used a well-known data mining methodology called CRISP-DM (Cross Industry Standard Process for Data Mining)[41], which was found in a comparison of data mining methodologies to be well suited for predictive tasks for crime data[42].

The CRISP-DM methodology involves six phases, which are briefly described in the following[41]:

1) *Business understanding* involves understanding the objectives from a business perspective and defining data mining problems for achieving the objectives;

2) *Data understanding* entails a process of familiarisation with the data, including spotting quality issues and noticing properties of the data that may be useful for the modelling phase;
3) *Data preparation* refers to the process of formatting the data that is needed in the modelling process; this could include for example, selection of attributes, transformation/creation of attributes and removal of noisy data;
4) *Modelling* involves the application of several modelling techniques or algorithms;
5) *Evaluation* refers to the assessment of the quality of the models developed at the previous phase;
6) *Deployment* depends on the objectives from the first phase; it could vary from a simple report with the results to a complex implementation based on the developed models.

For the purpose of our experiments, the first and last phases are the same. More specifically, the first phase is related to the objective of predicting crime; consequently, the aim is to investigate several predictive models. The last phase involves the reporting of the results of this investigation. The variations in the other phases for each experiment are presented in Section 4.

In relation to the modelling phase, it typically involves building a model using some of the data available; this data is referred to as the training set. The remaining data is used for evaluating the performance of the model (in the evaluation phase), and it is referred to as the test set. The model itself can be built using a variety of algorithms. The following subsection describes the algorithms used in our experiments.

3.2.1. *Algorithms*

For our experiment, we chose three algorithms from three categories of approaches[43]: instance-based learning, regression and decision trees. Details of each category and the particular algorithms used in our experiments, i.e. Locally Weighted Learning (LWL), linear regression (LR) and M5P, are given in the following.

1) *Instance-based learning* is a form of lazy learning characterised by deferring the processing of training the data until a query needs to be answered (i.e. to classify or predict the variable of interest for a particular instance) rather than building an explicit model. Typically this involves the storage of the training data in memory and finding the relevant data for answering a particular query; consequently, this type of learning is also referred to as memory-based learning. To assess the relevance of data for answering the query, a distance function is often used, with closer points having more relevance than further points. The closest points are then combined (e.g. averaged) to answer the query.

   The algorithms in this category for numerical prediction can be divided into two types: (a) similarity-based, e.g. Euclidean (IBk) or entropy-based (KStar) and (b) regression-based, e.g. LWL. Since regression is one of the most popular

methods for numerical prediction, a regression-based algorithm was chosen.

The Locally Weighted Learning (LWL) algorithm "uses locally weighted training to average, interpolate between, extrapolate from, or otherwise combine training data"[44]. More specifically, for prediction, the LWL algorithm uses regression to provide an answer to the query; in particular, it fits a surface to nearby points using distance weighted regression[45]. This fitting is done through multivariate smoothing, i.e. the dependent variable (the one that is being predicted) is smoothed as a function of the independent variables (the other variables/attributes involved in the prediction) in a moving manner which is similar to how a moving average is calculated for time series[46].

Despite the simplicity and naivety of this approach, instance-based algorithms are often competitive in terms of prediction accuracy[47]. The main disadvantages of this class of algorithms are the storage needs (because all the data needs to be stored) and computational complexity (because of the time required to search the closest points for each query). In the context of big data, these are major disadvantages; however, techniques have been developed to reduce the storage need and computational cost by selecting instances that are likely to be most relevant for the query, e.g.[48,49,50].

2) *Linear Regression* is a simple method for numeric prediction which has been widely used in statistics[51]. It involves finding a relationship between a variable of interest (the dependent variable) and one or more explanatory factors (the independent variables). For this purpose, linear functions are used, for which the unknown parameters, i.e. weights of the independent variables, are estimated from the training data[52]. These can then be used to predict the values of the dependent variable for new instances. To estimate the parameters, several methods can be used, of which one of the most popular is the least mean squares[51].

   Linear regression algorithms for prediction include simple regression (only one independent variable/predictor), multiple regression (two or more predictors) and pace regression[53], which is suitable for data of high dimensionality and only accepts binary nominal attributes. Our data has nominal attributes that are not binary, and the prediction involves more than two predictors; consequently, multiple regression was used.

   The main disadvantage of linear regression is its linearity. If the data has non-linear dependencies, a linear regression model will output the best-fitting line (as in the least mean-squared difference), which may not fit very well. In addition, regression can be computationally intensive when applied to high-dimensional data[54].

3) *Decision trees* can be used for both classification and prediction. For classification purposes, a function can be learned that is constant in intervals defined by splits on the individual attributes values[55]. The internal nodes of the tree represent the split decisions based on information gain or impurity metrics defined in terms of the class distribution of records before and after splitting. The leaf nodes of the tree are assigned a specific class attribute value (i.e. class label).

For prediction purposes, the decision trees algorithms for classification have been adapted to output a numerical value[51]. The main difference is that the leaves of the tree have numerical values, unlike classification trees which have class labels. Moreover, we can distinguish between regression trees and regression model trees[56]. In the first ones (e.g. REPTree), the leaves have a single value corresponding to the average of values that reach the leaf, while the second ones (e.g. M5P) use linear regression models to calculate the value of the leaves. The second category has the advantage of being more compact and delivering better prediction accuracy[56]; hence, this was used in our analysis.

The M5P algorithm[57] allows the prediction of continuous variables. It improved the M5 algorithm[58] by handling enumerated attributes and attribute missing values.

Decision trees have several advantages[51,47]: (a) they have an intuitive representation of the knowledge domain they are mapping; (b) they are non-parametric, which makes them especially suited for datasets where there is no prior knowledge of the probability distribution of attributes; (c) they are relatively fast and computationally inexpensive to construct, and the resulting model can be stored in a compact form.

A disadvantage of the decision trees algorithm is that they may include irrelevant attributes in the tree, and, consequently, produce trees that are larger than necessary. To address this disadvantage, pruning[59] is used, with the aim to simplify the tree structure. The M5P algorithm includes pruning.

3.2.2. *Evaluation Metrics*

The evaluation of a particular algorithm is typically done by evaluation metrics used on a test set, i.e. a set of data that was not used in building the model; the data set used for building the model is called the training set. One of the most popular methods is 10-fold cross-validation, which is also used in this research.

Cross-validation uses a number of folds or sets, which are repeatedly split into training and testing. The most popular is 10-fold cross-validation, which involves splitting the data into 10 parts. Each part is held out in turn and training is done on the remaining 9 parts; the evaluation metrics are calculated on the holdout set (i.e. the test set). This procedure is repeated 10 times such that each of the 10 parts is used as the test set. To evaluate the performance, the 10 evaluation metrics are averaged to give an overall performance estimate.

For the evaluation of numeric prediction there are several evaluation metrics that could be used. Three of the most popular metrics, which are also used in this research, are[51]: *mean absolute error*, *(root) mean-squared error* and *correlation coefficient* – their formulas are given below.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{1}$$

$$MSE = \frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n} \tag{2}$$

$$r = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{\left[n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2\right] * \left[n \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2\right]}} \tag{3}$$

where x_i are the values given by the prediction model, y_i are the *truth values* from the test set, n is the number of instances in the test set and i is the instance index taking values from 1 to n.

The *mean absolute error* (MAE) averages the magnitude of the individual errors. The *mean-squared error* (MSE) is often used because it tends to be easier to manipulate mathematically[51]; however, it is difficult to interpret – for this reason, the *root mean squared error* (RMSE) is used because it gives values in the same range as the predicted value itself, thus making the interpretation of the results easier. The *(root) mean-squared error* also is sensitive to outliers and exaggerates their effect, unlike the *mean absolute error*. A good performance is indicated by low error values.

The *correlation coefficient* measures the statistical correlation between the actual and the predicted values from the test set. It ranges from 1, which represents a perfect correlation to 0, when there is no correlation, to -1 where the values are perfectly inversely correlated. For prediction methods, negative values should not occur. A good performance is indicated by large values, i.e. the closer to 1 the better.

A model is judged by looking at the error metrics, as well as the coefficient values. Interpreting the coefficient value independently from the error metrics can lead to the wrong conclusions; however, when the error metrics are similar, the coefficient value can give an indication of which model performs better[51].

In statistical modelling, the PRESS (predicted residual sum of squares) statistic[60] is often used as a metric to compare the predictive value of several models. This metric is the sum of squared errors for the test set and it is equivalent to MSE multiplied with n, where n is the size of the test set. For the reasons outlined above, RSME is preferred to MSE. In conclusion, when using the RSME metric, the same ranking of models would result as when using the PRESS metric.

4. Experiments and Results

In this section, we present three experiments conducted on the data described in Section 3.1 and their results. The experiments investigate: (a) crime frequency prediction by LSOA code; (b) crime frequency prediction using postcode information; and (c) anti-social behaviour frequency prediction. In addition, we report and analyse the time required for the models to be built and tested, which has a practical implication on the use of the algorithms in the context of large volumes of data.

The experiments were conducted using the XXXXX High Performance Computer Cluster at the University of XXXXXXXXXX and the Weka software[51].

4.1. *Experiment 1: crime frequency prediction by LSOA code*

This experiment investigates the prediction of crime frequency by LSOA code. Fig. 1 illustrates the procedure used in the experiment, outlining the four middle steps of the CRISP-DM methodology, i.e. data understanding, data preparation, modelling and evaluation.

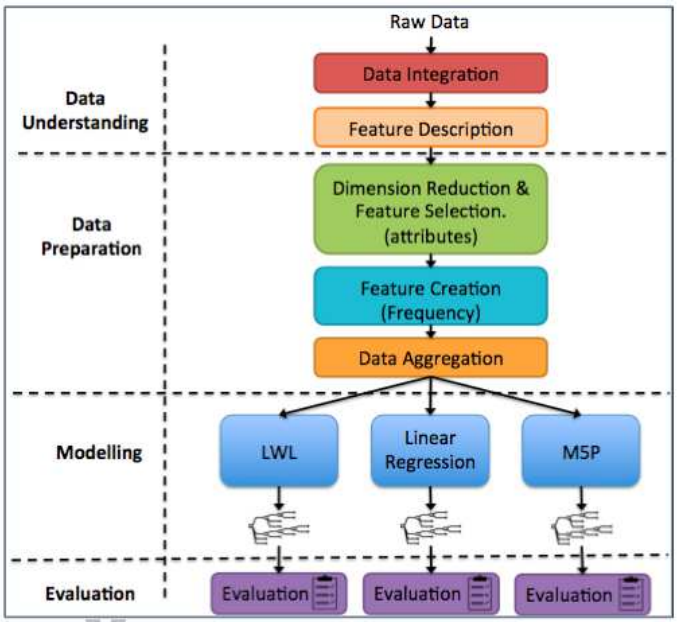

Fig. 1. Procedure for Experiment 1.

The data described in Section 3.1 is released in monthly files; consequently, the first stage in the Data Understanding step was to integrate the data into one dataset by aggregating the monthly files. This led to the dataset described in Section 3.1. All experiments started with this dataset.

For the purpose of this experiment, the frequency per month, per LSOA code, for each crime type was computed, i.e. an instance represents the frequency of crime for a particular month, LSOA code and crime type. This aggregation meant that some of the attributes that were relevant for individual crimes became irrelevant for a monthly record of crime frequency. These attributes are: Crime ID, Reported by, Longitude, Latitude, Location, Last outcome category and Context. Consequently, the dataset for this experiment included 5 attributes: Month, LSOA code, LSOA

name, Crime type and Frequency. The 46 instances from the original dataset that had missing values for the LSOA code and LSOA name were excluded.

The summary for the dataset used in this experiment is given in Table 5. We notice that the number of instances has been reduced by an order of 3, indicating that, on average, across all crimes, all LSOA codes and all 40 months, approximately 3 crimes occur per month.

Table 5. Summary of the dataset for Experiment 1.

Data Objects:	196,374
Attributes:	5
Values:	981,870
Missing Values:	0
% of Missing Values:	0%

Table 6 presents the statistical characteristics of the Frequency attribute, i.e. minimum (min) and maximum (max) values, the mean or average, and the standard deviation (StdDev). The mean of 3.12 confirms the average inferred above from the process of data aggregation. The minimum and maximum values clearly indicate that the frequency varies greatly, while the standard deviation indicates that most values are concentrated at the lower end.

Table 6. Statistics for the Frequency attribute in Experiment 1.

Min	Max	Mean	StdDev	25^{th} percentile	median	75^{th} percentile
1	233	3.12	4.51	1	2	3

To find more detailed information about the distribution of the frequency of crime, several categories were created, as displayed in Table 7. The majority of instances (150,186 corresponding to 76.46%) have a frequency of less or equal to 3. Moreover, 95.41% of instances have a value of 10 or less for the frequency of crime. There were only 2 instances with a frequency value of more than 200.

Table 7. Distribution of instances per frequency (f) categories.

$f <= 3$ 76.46%	$3 < f <= 5$ 10.45%	$5 < f <= 10$ 8.50%	$10 < f <= 15$ 2.58%	$10 < f <= 20$ 1.04%
$20 < f <= 50$ 0.87%	$50 < f <= 100$ 0.09%	$100 < f <= 200$ 0.01%	$f > 200$ 0.001%	

The three algorithms were applied and the results are presented in Table 8. The results are missing for the Linear Regression (LR) because the time to build the model was very long, and thus impractical. We stopped the building of the

model after 900 hours (more than 3 million seconds). The building and testing of the models using the LWL and M5P algorithms took approximately 6 and 33 hours, respectively. We noticed that the LWL algorithm is very quick in the training stage and takes longer in the testing stage – this is a characteristic of lazy learning algorithms, for which the answer to a query takes place in the testing stage (the training stage only involves loading the training data in memory). Unlike LWL, the M5P algorithm takes most of the time in the training stage, when an explicit model is built, while the testing is much quicker because it involves the use of the model on the test data. For more analysis on the time taken by the different algorithms, please see Section 4.4.

Table 8. Experiment 1 results.

Evaluation metric	LWL	LR	M5P
Mean Absolute Error (MAE)	2.04	–	1.32
Root Mean Squared Error (RMSE)	3.98	–	2.49
Correlation Coefficient	0.47	–	0.83
Time training (seconds)	0.04	3,280,680.00	117,668.67
Time testing (seconds)	19,935.54	–	6.83

In terms of the performance of the models, the LWL algorithm has a relatively low performance, with a correlation coefficient of 0.47, which indicates a medium strength relationship between the values predicted by the model and the real values. The M5P algorithm, on the other hand, has a correlation coefficient of 0.83, which indicates a strong relationship between the values predicted by the model and the real values. The error values, as expected, indicate higher values for the RMSE compared with MAE. The error values are relatively low; for the M5P algorithm, for example, the MAE value indicates that the predicted values are on average overestimated or underestimated by the value of 1, i.e. if the real value is 4, the predicted value could be 3 or 5.

Consequently, this experiments shows that the M5P algorithm can be reliably used to predict the frequency of crime per month, per crime type, per LSOA code. In the next experiment, we add the postcode as an attribute to investigate if the information about postcodes would improve the prediction performance.

4.2. *Experiment 2: crime frequency prediction using postcode*

The procedure for Experiment 2 is very similar to the one for Experiment 1, as illustrated in Fig. 2. The only difference is the addition of the postcode attribute, i.e. an instance represents the frequency of crime in a particular month, for a particular LSOA code and postcode, and for a particular crime type. To find the correspondence between LSOA codes and postcodes, the Office for National Statistics[61] website was used, which has a database listing postcodes and their different output areas, including LSOA codes. This database is from 2011.

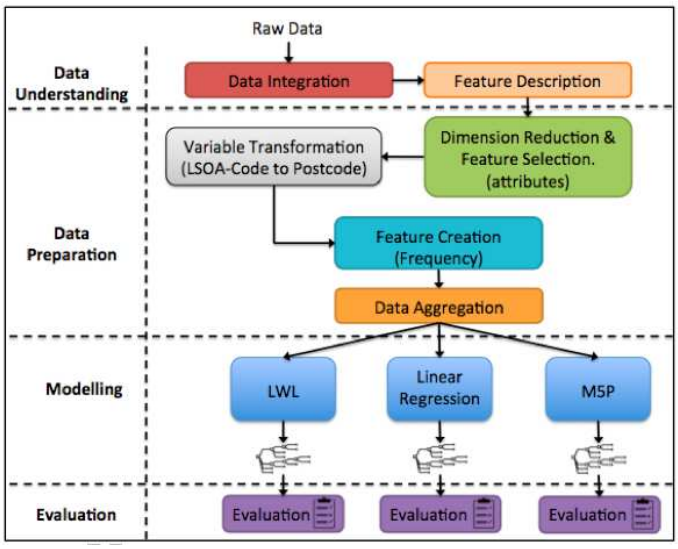

Fig. 2. Procedure for Experiment 2.

The summary of the dataset used in this experiment is given in Table 9. The lower number of data objects in this dataset compared with Experiment 1 is due to the lack of information on the equivalence between the LSOA codes and postcodes. Generally, postcodes areas cover several LSOA codes areas, thus capturing information at a different geographical level which could potentially improve predictions. From the 1494 LSOA codes, 1119 LSOA codes could me matched with one of the 129 Hampshire postcodes. The datasets for this experiment has 6 attributes – the same 5 as in Experiment 1, plus the postcode attribute obtained as described above.

Table 9. Summary of the dataset for Experiment 2.

Data Objects:	155,021
Attributes:	6
Values:	775,105
Missing Values:	0
% of Missing Values:	0%

Table 10 shows the minimum, maximum, mean and standard deviation of the Frequency attribute used in Experiment 2. These are different from Experiment 1 due to the change in the number of data objects. For example, we notice that the maximum value has changed dramatically, indicating that objects with high values were removed. The mean is the same, while the standard deviation is a bit lower, indicating that, similarly to Experiment 1, most instances have frequencies with

values at the lower end.

Table 10. Statistics for the Frequency attribute in Experiment 2.

Min	Max	Mean	StdDev	25^{th} percentile	median	75^{th} percentile
1	140	3.12	4.26	1	2	3

The results for the three algorithms are given in Table 11. In terms of the performance of the algorithm, LWL and LR have medium correlation coefficients, while the M5P algorithm has a strong correlation coefficient between the values predicted by the model and the real values. In terms of the error metrics, i.e. MAE and RMSE, as in Experiment 1 and as expected, the RMSE values are higher than the MAE values. The values are very similar to Experiment 1. For the LWL and M5P algorithms, we notice a small improvement, in both the correlation coefficient (2 to 3 %) and the error metrics compared with Experiment 1. This improvement could be due to the use of the postcode attribute and/or the reduction in data objects, and especially data objects with outlier values. As the dataset contained over 150 000 data objects, i.e. 79% of the dataset in Experiment 1, the size of the dataset is unlikely to have affected the results. Consequently, the improvement is likely to be due to the removal of objects with outlier values and/or the postcode attribute.

Table 11. Experiment 2 results.

Evaluation metric	LWL	LR	M5P
Mean Absolute Error (MAE)	2.00	1.95	1.30
Root Mean Squared Error (RMSE)	3.72	3.88	2.20
Correlation Coefficient	0.49	0.50	0.86
Time training (seconds)	0.04	389,816.58	61,159.30
Time testing (seconds)	15,182.61	7.35	5.05

Comparing the results of Experiment 1 and 2, we can conclude that the use of the postcode attribute does not lead to a considerable improvement in prediction performance. The improvement in performance in Experiment 2 was small, i.e. 2 to 3%, which indicates that the postcode attribute does not add much information compared with the LSOA attribute alone. As generally simpler models are better models[62], we believe the small improvement in performance does not justify the use of an additional attribute (which would increase the complexity of the models).

In terms of time, the Linear Regression algorithm is the slowest taking over 100 hours to build and test the model, and the LWL is the fastest, taking approximately 4 hours. The M5P algorithm took 17 hours to build and test the model.

Both Experiment 1 and 2 focused on a general models predicting the frequency of crime for all types of crime. As different types of crimes have different frequency patterns, building models for crime frequency for individual crimes may lead to

better performing models. To investigate this aspect, in Experiment 3, we explored models for the most frequent crime of the 16 crime types, i.e. anti-social behaviour.

4.3. *Experiment 3: anti-social behaviour frequency prediction*

The procedure for this experiment is illustrated in Fig. 3. Unlike the previous two experiments, we focus on only one type of crime, i.e. anti-social behaviour. We chose to focus on this crime because it is the most frequent of the 16 types of crime – see Table 3 in Section 3.1.

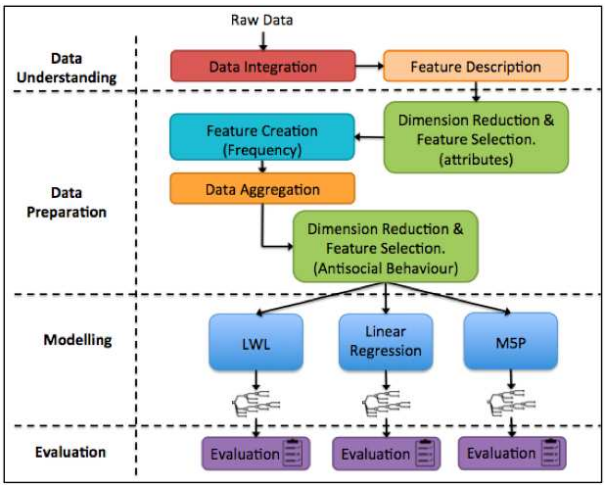

Fig. 3. Procedure for Experiment 3.

Consequently, for this experiment the dataset was much smaller, as illustrated in Table 12. It includes the same attributes as in Experiment 1 (except crime type), i.e. an instance represents the frequency of anti-social crime for a particular month and LSOA code. The minimum, maximum, mean and standard deviation for the Frequency attribute are given in Table 13. The maximum is 93, with an average of approximately 6 crimes and a standard deviation of approximately 6 as well, indicating that most instances would have values at the lower end (less than 20).

Table 12. Summary of the dataset for Experiment 3.

Data Objects:	44,053
Attributes:	4
Values:	176,280
Missing Values:	0
% of Missing Values:	0%

Table 13. Statistics for the Frequency attribute in Experiment 3.

Min	Max	Mean	StdDev	25^{th} percentile	median	75^{th} percentile
1	93	5.95	6.35	2	4	8

Similarly to Experiment 1, to further look into the distribution of instances according to crime frequency values, we created several categories, as displayed in Table 14. More than two thirds of the instances (30701 representing 69.70%) have frequencies less than or equal to the mean, i.e. 6. In addition, 97% of instances have frequencies of less than or equal to 20 crimes. There are only 3 instances with frequencies more than 90.

Table 14. Distribution of instances per frequency (f) categories.

$f <= 6$	$6 < f <= 10$	$10 < f <= 15$	$15 < f <= 20$	$20 < f <= 30$	$30 < f <= 40$
69.70%	15.72%	8.00%	3.58%	2.07%	0.51%
$40 < f <= 50$	$50 < f <= 60$	$60 < f <= 70$	$70 < f <= 80$	$f > 90$	
0.17%	0.14%	0.07%	0.04%	0.01%	

The results of the three algorithms are presented in Table 15. Both the LWL and the LR algorithms are performing better than in the previous experiments with correlation coefficients above 0.75, indicating that focusing on a particular crime may lead to better prediction models, at least for some algorithms. Moreover, the LR algorithm has a similar performance to the M5P algorithm, with a correlation coefficient of 0.85. In terms of the error metrics, the M5P is marginally better than the LR algorithm, which in turn, is better than the LWL algorithm.

Table 15. Third experiment results.

Evaluation metric	LWL	LR	M5P
Mean Absolute Error (MAE)	3.35	2.33	2.26
Root Mean Squared Error (RMSE)	4.32	3.39	3.33
Correlation Coefficient	0.75	0.85	0.85
Time training (seconds)	0.02	294677.50	46512.61
Time testing (seconds)	486.84	2.19	1.52

In terms of time taken to build and test the models, the LWL algorithm takes about 9 minutes, the LR algorithm takes about 81 hours and the M5P algorithm takes about 13 hours. The following section analyses the time required for training and testing across the 3 algorithms and the 3 experiments, and discusses the practical implications involved.

The Linear Regression model is illustrated in Equation (1) and the M5P model in Fig. 4. The LWL algorithm does not create a model, as pointed out previously as a characteristic of lazy learning algorithms (i.e. the prediction outputs are based

on instances stored in memory). In Equation (1), MC stands for Month Category, while LC stands for LSOA Category. These categories are subsets of the values of the Month and LSOA attributes. In Figure 4, the leaves of the tree are Linear Models (LM), which have the same form as Equation (1).

$$\begin{aligned} Frequency = & 0.2876 * MC_1 + 0.7129 * MC_2 + \ldots + \\ & 1.0284 * LC_1 - 1.4938 * LC_2 + \ldots + \\ & - 1.3642 \end{aligned} \tag{4}$$

4.4. *Time analysis*

The time required to build and test the prediction models has practical implications on the use of the different algorithms, especially when new data becomes available on a regular basis and updating the models could lead to better results.

Table 16 displays the number of instances per experiment and the time required for the 3 algorithms to build and test models. Fig. 5 displays the same information for an easier visual comparison. As pointed out in Experiment 1, unlike the LR and the M5P algorithms which take most of the time in the training stage, the LWL algorithm has a very brief training stage and a long testing stage, due to the lack of an explicit model.

Fig. 6 shows the *total time* (the sum of training and testing times) for the 3 algorithms in the 3 experiments. We excluded from the graph the time for the LR algorithm in Experiment 1, i.e. over 3 million seconds, to keep a lower scale for the time axis, which would enable a better visual comparison between the three algorithms.

The graph shows that the fastest algorithm is LWL; however, this algorithm is the one with the lowest performance. The linear regression algorithm takes the

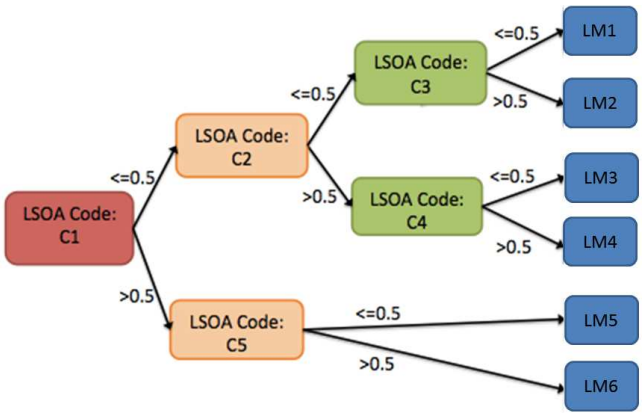

Fig. 4. Decision tree with M5P

Table 16. Number of instances and time for training and testing the models.

Experiment	No Instances	LWL		LR		M5P	
		Train	Test	Train	Test	Train	Test
Experiment 1	196,420	0.04	19,935.54	3,280,680.00	–	117,668.67	6.83
Experiment 2	155,021	0.04	15,182.61	389,816.58	7.35	61,159.30	5.05
Experiment 3	44,070	0.02	486.84	294,677.50	2.19	46512.61	1.52

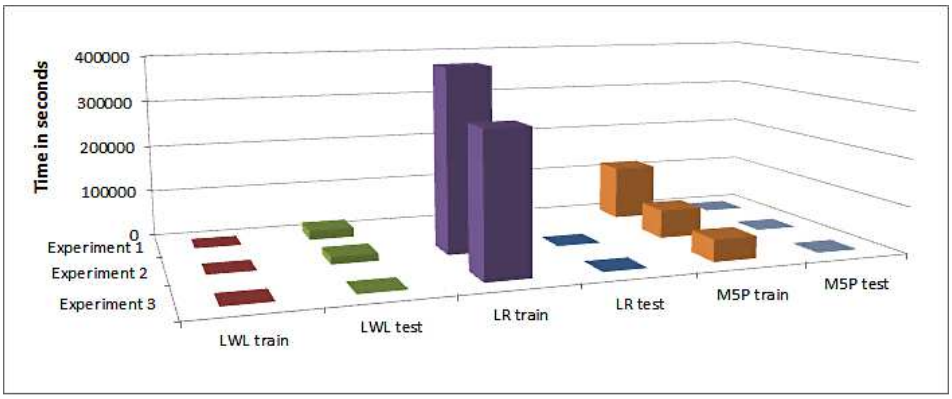

Fig. 5. Time in seconds for training and testing the 3 algorithms in the 3 experiments.

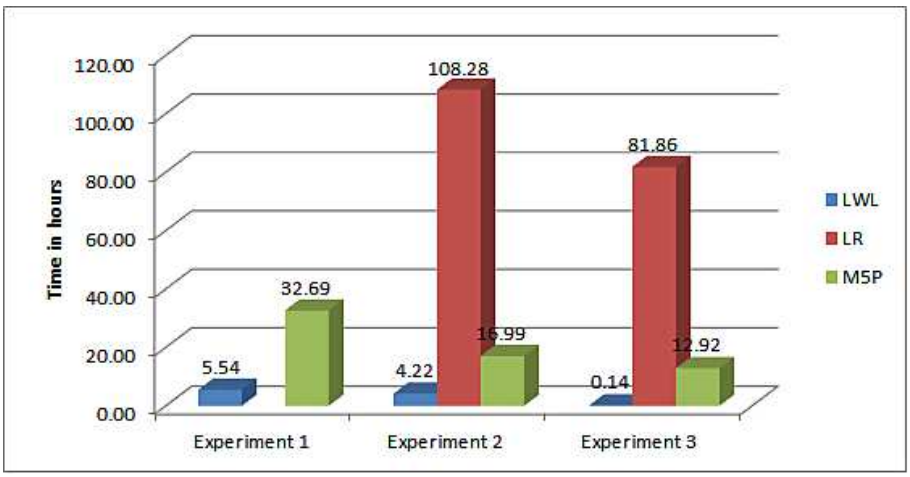

Fig. 6. Total time in hours for the 3 algorithms in the 3 experiments.

longest, while the M5P takes between 13 and 33 hours, depending on the size of the dataset. For all algorithms we see a correlation between the number of instances in the dataset and the time it takes to build and test a model, i.e. the more instances,

the longer the time needed.

From practical point of view, to decide which algorithm results in the best model, several aspects need to be considered: (a) the performance of the algorithms; (b) the time required for building and testing the model; (c) whether and how frequently the model would be updated; and (d) how important it is to understand the model, i.e. to understand how the prediction is calculated – an aspect also referred to as interpretability[63].

For example, if the model is only updated rarely, the time required to obtain the model is less important and priority may be given to the performance of the model, as well as its interpretability. If data becomes available on a regular basis, the model could be updated on a regular basis, in which case the time to produce the model is more relevant, and may take priority over performance (assuming it is above a reasonable threshold) or even interpretability.

5. Discussion and Conclusions

In this paper we focused on building predictive models for crime frequencies per crime type, per month and per LSOA code, from official data released by the UK Police. The main goal of the analysis was to investigate the level of performance that could be obtained from 40 months of crime records. The other goal was to investigate how a global model including all crime types compares with a specialised model for a particular crime type. For the specialised model, we chose the anti-social behaviour crime type due to it being the most frequent crime out of the 16 crime types.

We focused on LSOA code areas because these are the units used by the police. Consequently, working with predictions about frequency of crimes per LSOA code enables management of resources at that level. Moreover, the predictions can be easily aggregated across several LSOA codes, and indeed, the entire county, if that would be of interest. This allows identification of hot spots per LSOA code or wider areas (by aggregating the data across several LSOA code areas).

We also explored predictions for an individual crime type that allows identification of hot spots by crime. Similarly to the LSOA code, the data can be aggregated to give information across several types of crime. This may lead to finding that certain areas are hot spots for a variety of crimes, while others are hot spots only for particular types of crimes.

With regards to the hot spots definition, this can be defined differently per crime type or per area. This is one of the reasons we defined the prediction problem as a numeric prediction problem rather than a classification problem where for each instance the output would be a label indicating either a hot spot or a cold spot. A number can be better interpreted in context than a label.

In terms of time frame, we focused at month level because the data released by the police is on a monthly basis, indicating that this is the unit of time they are working with. Also, as this was an exploratory study, the focus was more on the feasibility of obtaining prediction models from the data, rather than detailed

consideration of the time frame. Given that the feasibility has now been established, further studies can be conducted to investigate predictions over different time frames.

Three categories of algorithms were used in our experiments, each with advantages and disadvantages. LWL, an instance-based learning algorithm, is quick overall, but has the disadvantage of not producing an explicit model. We found that this algorithm leads to a relatively poor performance and that it may be more suitable for creating specialised models, as its best performance was obtained in Experiment 3, which focused on anti-social behaviour.

Linear Regression is a well-researched algorithm, which is know to perform well on linear numeric predictions. Our experiments indicated that, similarly to LWL, the LR algorithm may be more suitable for specialised models. This could be explained by the fact that when building a global model, the data is likely to be less linear than for a specialised model. The disadvantage of this algorithm for large volumes of data is that it takes a long time to train and test a model. Indeed, in our experiments, this algorithm was the slowest. If in practice a model would need to be updated frequently over large amounts of data, this algorithm may not be suitable.

The M5P algorithm is part of a category of algorithms called decision trees, which are also know to perform well on a variety of prediction problems. In our experiments, this was consistently the best performing algorithm. Interestingly, for the global model, the performance, i.e. an RSME value of 2.49 and a correlation coefficient of 0.83, was only marginally lower than for the specialist model, i.e. an RSME value of 3.33 and a correlation coefficient of 0.85[c]. In terms of time, depending on the amount of data used, in our experiments, this algorithm took between 13 and 33 hours for training and testing a model.

We chose the three algorithms mentioned above as representatives of different categories of learning approaches, i.e. instance-based learning, regression and decision trees. In further research we will investigate other algorithms and their predictive performance.

Given the time to build the models and the other relevant criteria to be considered when building/updating models which were mentioned in Section 4.4, an interesting research direction would be the use of multiple criteria decision making approaches to identify the most appropriate algorithm for the task at hand. Such approaches have been proposed for classification algorithms[64,65], which could be extended to numeric prediction algorithms.

In our experiments, we used a relatively low number of attributes, e.g. 5 attributes for Experiments 1 and 6 attributes for Experiment 2. In addition, one of the used attributes is redundant, as the LSOA code and the LSOA name reflect the same information. In fact, when inspecting the LR and M5P models, we noticed

[c]Although the RSME value may seem better for the global model, the distribution of the data needs to be considered when comparing RSME values on different data; thus for the global model the mean values was around 3, while for the specialist model the mean value was around 6.

that the LSOA name was not included in the prediction models. The exclusion of the LSOA name attribute may also lead to a reduction in the time taken to train and test the models. While the number of attributes may seem low, it is in alignment with other research showing that simple models perform well, while also having the advantage of reduced computational complexity[62], which is an important aspect to consider when dealing with large volumes of data.

In conclusion, building prediction models related to frequencies of crime from large amounts of data is feasible, even when the information available is limited. Further experiments can be conducted to investigate other aspects such as: the time frame for prediction, the amount of data necessary for reliable prediction models, and predictive models for particular types of crime.

Such models of frequency prediction for all crimes in a particular area or just particular crimes can be used in decision-making processes for allocation of police resources. An increase in crime in a particular area would need additional resources for dealing with that increase. The prediction models can indicate in which areas crime will increase and in which areas it will decrease, thus allowing the transfer of resources from one area to another by ensuring that resources are reduced for areas with a high likelihood of crime decrease.

They can be also be used for exploring temporal patterns for particular areas, for example to identify particular months in which crime frequency increases or decreases regularly, which could facilitate the understanding of factors leading to such regular variance.

To explore the aspects mentioned above, the prediction models could be integrated with existing decision support systems, which would allow the production of reports based on the different aspects investigated, as well as filtering by location, time period and type of crime.

Acknowledgements

This research has been partly funded by the XXXXXXXXXXXXXXXXX, the University of XXXXXXXXXXX.

References

1. L. Tompson, S. Johnson, M. Ashby, C. Perkins, and P. Edwards, UK open source crime data: accuracy and possibilities for research, *Cartography and Geographic Information Science*, **42** (2) (2015) 97–111, `http://dx.doi.org/10.1080/15230406.2014.972456`
2. S. Chainey and L. Tompson, Engagement, empowerment and transparency: Publishing crime statistics using online crime mapping, *Policing*, **6** (3) (2012) 228–239, `http://policing.oxfordjournals.org/content/6/3/228.abstract`.
3. N. Levine, Hot spot analysis of zones, in *Quickguide to CrimeStat IV (CrimeStat IV: A Spatial Statistics Program for the Analysis of Crime Incident Locations, Version 4.0)* (2013), `https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=265044`.
4. R. B. Santos, Effectiveness of police in reducing crimes and the role of crime analysis, in *Crime Analysis with Crime Mapping*, ed. R. B. Santos (Sage, 2012), pp. 40–53.

5. S. D. Johnson, A brief history of the analysis of crime concentration, *European Journal of Applied Mathematics*, **21** (4-5) (2010) 349–370, http://journals.cambridge.org/article_S0956792510000082.
6. K. Pease and A. Tseloni, Crime concentration and its prevention, in *Using Modeling to Predict and Prevent Victimization*, ser. SpringerBriefs in Criminology (Springer International Publishing, 2014), vol. 13, pp. 17–27, http://dx.doi.org/10.1007/978-3-319-03185-9_2.
7. A. A. Braga and R. V. Clarke, Explaining high-risk concentrations of crime in the city: Social disorganization, crime opportunities, and important next steps, *Journal of Research in Crime and Delinquency*, 2014, http://jrc.sagepub.com/content/early/2014/01/29/0022427814521217.abstract.
8. D. W. Brook, J. S. Brook, Z. Rosen, M. D. la Rosa, I. D. Montoya, and M. Whiteman, Early risk factors for violence in Colombian adolescents, *American Journal of Psychiatry*, **160** (8) (2003) 1470–1478, pMID: 12900310. http://dx.doi.org/10.1176/appi.ajp.160.8.1470
9. M. Townsley, S. Reid, D. Reynald, J. Rynne, and B. Hutchins, Risky facilities: Analysis of crime concentration in high-rise buildings, *Trends and Issues in Crime and Criminal Justice*, **476** (2014) 1–7.
10. E. Raleigh and G. Galster, Neighborhood disinvestment, abandonment, and crime dynamics, *Journal of Urban Affairs* (2014), http://dx.doi.org/10.1111/juaf.12102.
11. J. J. Sloan and B. S. Fisher, Campus crime, in *The Encyclopedia of Criminology and Criminal Justice* (Blackwell Publishing Ltd, 2014), http://dx.doi.org/10.1002/9781118517383.wbeccj236.
12. F. Glowacz and M. Born, Away from delinquency and crime: Resilience and protective factors, in *The Development of Criminal and Antisocial Behavior. Theory, Research and Practical Applications*, eds. J. Morizot and L. Kazemian (Springer International Publishing, 2015), pp. 283–294.
13. E. Drake, S. Aos, and M. Miller, Evidence-based public policy options to reduce crime and criminal justice costs: Implications in washington state, *Victims and Offenders*, **4** (2) (2009) 170–196, http://www.scopus.com/inward/record.url?eid=2-s2.0-60449116279&partnerID=40&md5=59130d507826ed767fb03fddf92dcaac.
14. B. C. Welsh and D. P. Farrington, The benefits and costs of early prevention compared with imprisonment: Toward evidence-based policy, *The Prison Journal*, **91** (3) (2011) 120S–137S, http://tpj.sagepub.com/content/91/3_suppl/120S.abstract.
15. B. C. Welsh and D. P. Farrington, Science, politics, and crime prevention: Toward a new crime policy, *Journal of Criminal Justice*, **40** (2) (2012) 128–133, http://www.sciencedirect.com/science/article/pii/S0047235212000232.
16. R. Stein and C. Griffith, Community policing strategies need to take into account police and residents' different perceptions of neighborhood crime. *USApp–American Politics and Policy Blog* (2015).
17. Y. Xue and D. E. Brown, Spatial analysis with preference specification of latent decision makers for criminal event prediction, *Decision Support Systems*, **41** (3) (2006) 560–573, http://www.sciencedirect.com/science/article/pii/S0167923604001319.
18. M. B. Short, M. R. D'Orsogna, V. B. Pasour, G. E. Tita, P. J. Brantingham, A. L. Bertozzi, and L. B. Chayes, A statistical model of criminal behavior, *Mathematical Models and Methods in Applied Sciences*, **18** (supp01) (2008) 1249–1267.
19. G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, Self-exciting point process modeling of crime, *Journal of the American Statistical Association*, **106** (493) (2011) 100–108, http://dx.doi.org/10.1198/jasa.2011.ap09546.

20. M. Short, M. D'Orsogna, P. Brantingham, and G. Tita, Measuring and modeling repeat and near-repeat burglary effects, *Journal of Quantitative Criminology*, **25** (3) (2009) 325–339.
21. J. Eck, S. Chainey, J. Cameron, and R. Wilson, Mapping crime: Understanding hotspots, Technical Report (2005), `http://discovery.ucl.ac.uk/11291/1/11291.pdf`.
22. M. Helbich, J. Hagenauer, M. Leitner, and R. Edwards, Exploration of unstructured narrative crime reports: an unsupervised neural network and point pattern analysis approach, *Cartography and Geographic Information Science*, **40** (4) (2013) 326–336, `http://dx.doi.org/10.1080/15230406.2013.779780`.
23. M. A. Andresen and N. Malleson, Police foot patrol and crime displacement a local analysis, *Journal of Contemporary Criminal Justice*, **30** (2) (2014) 186–199.
24. G. C. Oatley and B. W. Ewart, Crimes analysis software:ins in maps clustering and bayes net prediction, *Expert Systems with Applications*, **25** (4) (2003) 569–588.
25. T. H. Grubesic, On the application of fuzzy clustering for crime hot spot detection, *Journal of Quantitative Criminology*, **22** (1) (2006) 77–105.
26. A. T. Murray and T. H. Grubesic, Exploring spatial patterns of crime using non-hierarchical cluster analysis, in *Crime modeling and mapping using geospatial technologies*, ed. M. Leitner (Springer, 2013), pp. 105–124.
27. Y. Peng, G. Kou, Y. Shi, and Z. Chen, A descriptive framework for the field of data mining and knowledge discovery, *International Journal of Information Technology & Decision Making*, **7** (4) (2008) 639–682.
28. D. E. Brown and S. Hagen, Data association methods with applications to law enforcement, *Decision Support Systems*, **34** (4) (2003) 369–378, `http://www.sciencedirect.com/science/article/pii/S0167923602000647`.
29. S.-T. Li, S.-C. Kuo, and F.-C. Tsai, An intelligent decision-support model using FSOM and rule extraction for crime prevention, *Expert Systems with Applications*, **37** (10) (2010) 7108–7119, `http://www.sciencedirect.com/science/article/pii/S0957417410001855`.
30. J. Bachner, *Predictive policing: Preventing crime with data and analytics* (IBM Center for The Business of Government, 2013).
31. H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, Crime data mining: a general framework and some examples, *IEE Computer*, **37** (4) (2004) 50–56.
32. K. Dahbur and T. Muscarello, Classification system for serial criminal patterns, *Artificial Intelligence and Law*, **11** (4) (2003) 251–269.
33. S. Lin and D. E. Brown, An outlier-based data association method for linking criminal incidents, *Decision Support Systems*, **41** (3) (2006) 604–615, `http://www.sciencedirect.com/science/article/pii/S0167923604001344`.
34. A. Malathi and S. S. Baboo, An enhanced algorithm to predict a future crime using data mining, *International Journal of Computer Applications*, **21** (1) (2011) 1–6 .
35. S. V. Nath, Crime pattern detection using data mining, in *Web Intelligence and Intelligent Agent Technology Workshops, 2006 IEEE/WIC/ACM International Conference on* (IEEE, 2006), pp. 41–44.
36. P. Phillips and I. Lee, Mining co-distribution patterns for large crime datasets, *Expert Systems with Applications*, **39** (14) (2012) 11 556–11 563, `http://www.sciencedirect.com/science/article/pii/S0957417412005945`.
37. T. Wang, C. Rudin, D. Wagner, and R. Sevieri, Learning to detect patterns of crime, *Machine Learning and Knowledge Discovery in Databases*, **8190** (2013) 515–530.
38. C.-H. Yu, M. W. Ward, M. Morabito, and W. Ding, Crime forecasting using data mining techniques, in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International*

Conference on (IEEE, 2011), pp. 779–786.
39. Z. S. Zubi and A. A. Mahmmud, Using Data Mining Techniques to analyze crime patterns in the Lybian National Crime Data, in *Recent Advances in Image, Audio and Signal Processing*, ed. S. Sergyan (WSEAS, 2013) pp. 79–85.
40. HO. Home Office UK. `http://data.police.uk/`
41. R. Wirth and J. Hipp, CRISP-DM: Towards a standard process model for data mining, in *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (2000), pp. 29–39.
42. F. Ozgul, C. Atzenbeck, A. Çelik, and Z. Erdem, Incorporating data sources and methodologies for crime data mining, in *Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on* (IEEE, 2011), pp. 176–180.
43. M. Kantardzic, *Data mining: concepts, models, methods, and algorithms* (John Wiley & Sons, 2011).
44. V. Vapnik and L. Bottou, Local algorithms for pattern recognition and dependencies estimation, *Neural Computation*, **5** (6) (1993) 893–909.
45. C. Atkeson, A. Moore, and S. Schaal, Locally weighted learning, *Artificial Intelligence Review*, **11** (1-5) (1997) 11–73, `http://dx.doi.org/10.1023/A\%3A1006559212014`.
46. W. S. Cleveland and S. J. Devlin, Locally weighted regression: an approach to regression analysis by local fitting, *Journal of the American Statistical Association*, **83** (403) (1988) 596–610.
47. F. Tzima, K. Karatzas, P. Mitkas, and S. Karathanasis, Using data-mining techniques for PM10 forecasting in the metropolitan area of Thessaloniki, Greece, in *Neural Networks, 2007. IJCNN 2007. International Joint Conference on* (2007), pp. 2752–2757.
48. D. Wilson and T. Martinez, Reduction techniques for instance-based learning algorithms, *Machine Learning*, **38** (3) (2000) 257–286, `http://dx.doi.org/10.1023/A\%3A1007626913721`.
49. H. Brighton and C. Mellish, Advances in instance selection for instance-based learning algorithms, *Data Mining and Knowledge Discovery*, **6** (2) (2002) 153–172, `http://dx.doi.org/10.1023/A\%3A1014043630878`.
50. C. C. Aggarwal, Instance-based learning: A survey, in *Data Classification: Algorithms and Applications*, ed. C. C. Aggarwal (CRC Press, 2014), pp. 157–185.
51. I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 3rd edn. (Morgan Kaufmann, 2011).
52. H. Späth, *Mathematical algorithms for linear regression* (Academic Press, 2014).
53. Y. Wang, A new approach to fitting linear models in high dimensional spaces, Ph.D. dissertation, The University of Waikato (2000).
54. J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and techniques*, 3rd edn. (Morgan Kaufmann, 2012).
55. S. K. Murthy, Automatic construction of decision trees from data: A multi-disciplinary survey, *Data mining and knowledge discovery*, **2** (4) (1998) 345–389.
56. M. Göndör and V. Bresfelean, Reptree and M5P for measuring fiscal policy influences on the Romanian capital market during 2003–2010, *International Journal of Mathematics and Computers in Stimulation*, **6** (4) (2012) 378–386.
57. Y. Wang and I. H. Witten, Inducing model trees for continuous classes, in *Proceedings of the Ninth European Conference on Machine Learning* (1997), pp. 128–137.
58. J. R. Quinlan, Learning with continuous classes, in *5th Australian joint conference on artificial intelligence*, vol 92 (1992), pp. 343–348.
59. F. Esposito, D. Malerba, G. Semeraro, and J. Kay, A comparative analysis of methods for pruning decision trees, *Pattern Analysis and Machine Intelligence, IEEE Trans-*

actions on, **19** (5) (1997) 476–491.
60. T. Tarpey, A note on the prediction sum of squares statistic for restricted least squares, *The American Statistician*, **54** (2) (2000) 116–118.
61. ONS. (2011) Postcodes (enumeration) (2011) to output areas (2011) to lower layer super output areas (2011) to middle layer super output areas (2011) to local authority districts (2011) e+w lookup. Office for National Statistics. `https://geoportal.statistics.gov.uk/geoportal/catalog/search/resource/details.page?uuid={18444B52-47C2-40FE-8003-92230C344598}`
62. M. Cocea and S. Weibelzahl, Log file analysis for disengagement detection in e-learning environments, *User Modeling and User-Adapted Interaction*, **19** (4) (2009) 341–385.
63. H. Liu, M. Cocea, and A. Gegov, Interpretability of Computational Models for Sentiment Analysis, in *Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence*, eds. W. Pedrycz and S. M. Chen, Studies in Computational Intelligence, vol.639 (Springer, 2016), pp. 199–220.
64. Y. Peng, G. Kou, G. Wang, and Y. Shi, FAMCDM: A fusion approach of mcdm methods to rank multiclass classification algorithms, *Omega*, **39** (6) (2011) 677–689.
65. G. Kou, Y. Lu, Y. Peng, and Y. Shi, Evaluation of classification algorithms using mcdm and rank correlation, *International Journal of Information Technology & Decision Making*, **11** (1) (2012) 197–225.