National College of Ireland

Identifying Theft Hotspots: A Machine Learning Approach to Aggregated Data Analysis

MSc Research Project
Data Analytics

Mark Meyler
Student ID: x21171254

School of Computing
National College of Ireland

Supervisor: Mohammed Hasanuzzaman

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Mark Meyler		
Student ID:	x21171254		
Programme:	MSc in Data Analytics	**Year:**	2025
Module:	Research Project		
Supervisor:	Mohammed Hasanuzzaman		
Submission Due Date:	28/01/2026		
Project Title:	Identifying Theft Hotspots: A Machine Learning Approach to Aggregated Data Analysis		
Word Count: 6566		**Page Count:** 20	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Mark Meyler

Date: 27/01/2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Identifying Theft Hotspots: A Machine Learning Approach to Aggregated Data Analysis

Mark Meyler
x21171254

Abstract

Crime prediction is a critical domain of research due to its potential economic benefits and improvements in quality of life. Theft crimes represent a significant proportion of criminal incidents and, therefore, the ability to identify areas of increased theft risk, i.e. a hotspot, is essential to enable mitigation measures to be implemented by crime prevention agencies, but also to inform all interested parties of localised hotspots to safeguard against potential personal risk. A widely discussed challenge of effectively achieving this is the distinct lack of openly available crime data at low resolutions. This study proposes a geographically weighted LightGBM developed using fully openly available data at a coarse aggregation that can be used to predict theft hotspots at a more granular level. The paper focusses on burglary counts at MSOA levels to generate predictions at the finer LSOA scale, which are used to disaggregate the MSOA counts. Three different combinations of regions in England were utilised to evaluate scalability. The results demonstrate the enhanced capabilities of classifying areas of higher theft risks, with an increase in precision of up to 11% and recall of up to 20% in comparison to assuming the MSOA rate across all LSOAs within. There is some validation in the proposal that weighting the predictions from the local models can achieve more accurate and stable results, but further research is required to be conclusive. The integration of SHAP facilitates improved interpretability of key features contributing to these predictions.

Keywords: Hotspot, Spatial disaggregation, coarse resolution, finer resolution, geographically weighted, spatial autocorrelation, interpretability

1 Introduction

1.1 Background

Crime represents a significant issue that inevitably causes negative impacts on its victims and on society as a whole. Crime prediction and forecasting is an area of research that has gained significant attention over the years given it's potential to provide meaningful benefits both to quality of life and to the economy. One of the most prominent crime types are theft related incidents, which averages around 40% of all crimes in Ireland over the past 20 years according to the Central Statistics Office, and therefore presents an interesting topic of research. Whilst the problem will very likely never be fully eradicated, preventative measures can certainly be put in place to limit its effects. In order to do so, the underlying patterns and locations of such activities need to be understood.

Crime prediction and forecasting has been thoroughly investigated in recent years given its potential benefit in enhancing preventative measures and negating the negative societal impacts. As continuous improvements are made to the capabilities of AI, opportunities are presented for advancing existing crime prediction methodologies. Du and Ding (2023) and Kwon *et al.* (2021) suggest there is gap in research conducted at the micro-level with the majority conducted at coarser resolutions such as census tract or districts. This is most likely

a result of lack of openly available granular level crime data, an issue widely discussed as one of the biggest challenges within the field of crime prediction (Huamantingo *et al.*,2025; Du and Ding; Liang *et al*.; Mandalapu *et al*.; Yin, 2023). This is one of three commonly identified complications with crime prediction strategies, which also include spatial heterogeneity and model interpretability. Spatial heterogeneity refers to the concept that patterns vary over geographical space, and previous literature finds that this is a critical characteristic of crime (Du and Ding, 2023; Cohen and Felson, 1979). Interpretability of machine learning (ML) techniques is consistently discussed as fundamental in general (Du and Ding, 2023), but it becomes even more essential when used for topics such as crime prediction where outcomes could have distinct consequences.

1.2 Research Question

These challenges present the research question of this paper:

How can machine learning be utilised to effectively and interpretably identify high-risk theft hotspots when combined with spatial disaggregation techniques in the absence of granular incident-level crime data?

The project approaches this by focussing on burglary incidents in England, and aims to develop a method of leveraging trends in the data at a coarse aggregation level to predict burglary counts at a finer resolution and, hence, enable more accurate identification of hotspots should granular data at that resolution not be available. A hotspot can be defined as an area with a higher concentration of an activity relative to overall distribution across the total study area (Hajela *et. al*, 2020), and hence, improved burglary hotspot classification in this project represents development of a model that can highlight zones of increased burglary frequency. Whilst this presents a classification problem, i.e. whether an area is a hotspot or not, the proposed method will initially treat it as a regression problem. The rationale is that there is potential that predictions for the finer resolution may be overstated in spatial disaggregation models if the model fails to understand the refined spatial autocorrelation, but by first estimating the crime counts and constraining them to actuals at the coarser level it should ensure that the model does not overestimate crime with respect to the overall study area. In other words, an area might appear as a hotspot within its broader area but not necessarily within the entire region, and the logic may help preserve the distinction. The proposed method involves constructing local LightGBM (LGBM) models alongside a global LGBM, i.e. a geographically weighted model, at Middle layer Super Output Area (MSOA) level to capture the spatial heterogeneity, estimating the counts for a Lower layer Super Output Area (LSOA) by weighting the predictions from the models using Gaussian kernel and re-proportioning the actual MSOA counts to the LSOA level based on those predictions. This represents downscaling of approximately five-fold, varying depending on the study region. The methodology will also utilise SHapley Additive exPlanations (SHAP), which will provide values that have been weighted in the same manner as the predictions, to provide insight into the features driving those predictions.

1.3 Structure of the Paper

Section 2 discusses some of the most relevant literature reviewed on the research topic, highlighting insights gained and potential limitations. Section 3 presents the general approach and techniques utilised in the methodology. Section 4 builds on Section 3 by outlining how these techniques were adapted and applied to the problem. Section 5 details how this

framework was implemented before Section 6 presents the findings. Finally, section 7 provides final conclusions on the research and proposes future directions.

2 Related Work

2.1 Crime Prediction with Machine Learning

As mentioned previously, there has been extensive research in recent years on how to best apply ML technologies to predict crime. Zhang *et al.* (2022) present an interpretable approach that combines an Extreme Gradient Boosting (XgBoost) model with SHAP for identifying whether relatively small grids have experienced theft crimes. The accuracy suggests positive performance, but the receiver operating characteristic (ROC) of 0.586 suggests that its classification ability is only marginally better than random chance (Géron, 2019). This could potentially be attributed to insufficient attention to the spatial effects of crime patterns. The research does positively support the use of SHAP for explainability, clearly outlining the features influencing the models' predictions.

A similar method is adopted by Kim *et al.* (2024), who investigate using tree-based ensemble algorithms in conjunction with SHAP to predict the location of crimes. Their findings demonstrate the predictive power of such techniques, suggesting that their proposed XgBoost significantly outperforms others. These results are perhaps misleading as it appears hyperparameters have been thoroughly optimized for their proposed method, but not for comparison models. There would also appear to be a lack of consideration given to spatial heterogeneity in the model development.

Addressing spatial heterogeneity is commonly discussed as major component of effective crime prediction models. He *et al.* (2020) exhibit the use of Moran's I for exploratory data analysis, which clearly highlights the spatial pattern of larceny theft in the study area, but it would appear they do not incorporate the findings nor any spatially derived variables in the model. Deng *et al.* (2023) account for spatial effects by including a spatiotemporal lag variable, derived using historical crime data. Whilst this variable demonstrated predictive power, its inclusion may reduce the explainability of crime patterns as reasons for the past crimes would need to be understood to infer what is driving current crime trends, as highlighted by Deng *et al*. themselves.

The most common method of accounting for spatial heterogeneity that appears be used in crime prediction is simply the inclusion of location as a feature. An example is in the works of Safat *et al.* (2021), who include latitude and longitude of the incidents. However, using these geolocation variables as features can lead to considerable overfitting due to their highly correlated nature (Meyer *et al*., 2019) and therefore careful engineering is required if they are to be included. The work of Santos *et al.* (2023) further illustrates this point, presenting correlation matrices suggesting that longitude and latitude are indeed highly correlated with each other, but also with temporal features. One method of transformation of the XY coordinates was shown by İlgün and Dener (2025). They used these coordinates to develop a 'Cluster' feature using K-Means clustering, and created 'rotated' coordinates features of varying degrees. Whilst these features seemed to demonstrate predictive capabilities, the paper does not seem to address the concern of multicollinearity.

2.2 Crime Prediction with Geographically Weighted Models

Another approach to tackling spatial heterogeneity is using geographically weighted models. These methods involve fitting individual local models at each data point in the

training data, with the observations closer to that data point weighted more heavily and hence having more influence on the model (Jordan, 2023). Yue and Chen (2025) utilise a geographically weighted random forest (GWRF) model for predicting the number of street theft crimes. The results indicated steady improvement when compared with the standard global random forest (RF) model. Their findings also present the relative feature derived using SHAP, indicating on-street and ambient population are the main drivers behind street theft. The distribution of these SHAP values is illustrated for individual factors calculated by each local model, but it is not clarified if these values have been weighted in the same manner as the data points were when fitting the local models. Without doing so, these values may not fully represent the true influence of the features on the predictions and potentially over or underestimate the importance of certain features.

Sun *et al.* (2024) developed a Python package named PyGRF, that adopts similar logic as the GWRF used by Yue and Chen (2025). The procedure produces a prediction based on a combination of a global RF and the local RF closest to the test data point being evaluated, which seems to offer improvement on using one or the other but means if a single local model is not efficiently constructed it can result in poor prediction of points in that area. Another limitation is the use of spatial matrices for the weighting of training data points for fitting models, which despite offering potential improvement on computational times require a significant amount of memory space to construct, potentially limiting the scalability to larger datasets. This Python package was used as the basis for the methodology developed in this project, while aiming to address some of the limitations raised.

2.3 Spatial Disaggregation

There would seem to be a considerable lack of research conducted in the area of disaggregating crime data to smaller areas. The only study found that represents a similar study to this paper was conducted by Lamari *et al.* (2020), who aimed to train a predictive model based on 13,897 U.S. Block Groups to estimate the crime counts at 217,840 contiguous U.S. Census Blocks, although it is not clear exactly from the paper what level of downscaling this represents. Their findings suggest that LGBM outperforms both their proposed method of Poisson Regression and Deep Learning techniques. Feature importance is illustrated using the in-built functionality in tree-based models, which suffers at correctly assigning relative importance in the presence of highly correlated features (Raschka and Mirjalili, 2019).

As highlighted by Atkins *et al.* (2025) the majority of spatial disaggregation studies focus on health risk or population density mapping. Congdon (2025) proposes a complex mathematical solution for drug-related problems that first estimates counts and weights for sub-regions based on the region age structure and the global relationship of age with the dependent variable. A relative risk is then calculated using regression, which is adjusted based on the estimated count and weight. This method assumes a strong relationship with a single variable in the initial weighting step, which may not exist in more complex problems such as crime prediction, and could lead to misallocation of the prediction where the relationship may be insignificant. Monteiro *et al.* (2019) present a hybrid technique that combines spatial disaggregation techniques with ML for population mapping, but their solution represents the opposite approach to this proposal. They utilise pycnophylactic interpolation to produce initial estimates at finer resolutions based on aggregated data, before refining these estimates by fitting a regression model with ancillary data at the refined level. This approach assumes uniform degree of spatial autocorrelation across the study area, a condition that is rare in crime data which is inherently noisy. If there is significant error in

initial estimates, it is unlikely that a regression model will achieve accurate refinement in the final predictions.

2.4 Research Gap

This paper aims to address a critical gap that appears to persist in crime prediction, despite the considerable research. Most papers reviewed seem to utilise relatively granular crime data when conducting their experiments, although not at the micro level as Du and Deng (2023) suggest, for the purpose of aiding crime prevention agencies at identifying areas to allocate resources. However, availability of such data is widely discussed as a major challenge in crime prediction, and crime prevention agencies are not the only interested party in crime statistics. Whilst incident level data may be available to the likes of police departments enabling utilisation of the methods discussed above, other stakeholders likely only have access to aggregated data, such as insurance companies aiming to detect high risk areas or real estate agencies that wish to highlight to potential customers that neighbourhoods are comparatively safe despite the wider area presenting higher crime rates. Therefore, this project presents a scalable ML method that leverages this aggregated data to enable prediction at a more refined resolution.

3 Research Methodology

3.1 Datasets

The scope of the research focussed on England, due to the openly available crime data and easily accessible Census data. Crime data is made available by the Single Online National Digital Team[1], generated using details reported by each individual Police Force. This data is provided at a street level, with a categorisation of crime type and the longitude and latitude of the incident. To avoid breaches of data protection legislation laws, these locations have been anonymised by mapping the incident to an approximate point, such as the centre point of a street or a commercial premise, rather than the exact location. It is worth noting that this anonymisation could potentially impact the true validity of predicting crimes to the incident level. Tompson *et al.* (2015) performed a review of this accuracy of this open crime data, finding that aggregated counts were accurate 90% of the time for MSOAs and LSOAs but demonstrating significant error at the lowest level of granularity, Output Areas. Therefore, this research focussed on MSOA and LSOA level counts, with MSOA data representing the training data with LSOA used for test. The boundary data for these aggregation levels, i.e. the geometries of their polygons, was sourced from the Office for National Statistics (ONS)[2]. The ancillary data was obtained from several sources. The majority of data came from the ONS[3] who provide openly available Census data. Data related to built-up areas, greenspace, rivers and roads were sourced from Ordnance Survey[4]. The annual average temperature and precipitation measures were extracted from the Met Office Climate Data Portal[5]. All data utilised in the project are made openly available under the Open Government Licence v3.0[6].

[1] Available at: https://data.police.uk/data/

[2] Available at: https://geoportal.statistics.gov.uk/datasets/

[3] Available at: https://www.ons.gov.uk/

[4] Available at: https://osdatahub.os.uk/data/downloads/open/

[5] Available at: https://climatedataportal.metoffice.gov.uk/search

[6] https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

3.2 Data Pre-Processing

There was a significant amount of data cleaning and pre-processing required to produce final working datasets. The crime data has several known issues as reported by Police.UK, such as duplicate records and constantly changing data. These duplicate records were handled by removing records where all data was matching a previous record. Where incidents were appearing more than once due to updates to reported details, an assumption was made that the month of the first record and all other details of the final record was the correct information. All records reported by the British Transport Police were also removed, given they specifically relate to crimes on railways and would likely experience different crime trends. Finally, any record that was missing Longitude or Latitude were removed, as they could not be placed within boundary areas.

Crime datasets at MSOA and LSOA levels were produced by counting burglary incidents within each boundary over the period of 2023 and 2024, with separate datasets developed to cover three different areas to investigate the scalability of the method. All datasets excluded Greater Manchester who currently do not possess an IT system to report individual incident data (Police.uk, n.d) and the Isle of Scilly and Isle of Wight, which are islands that may not be comparable given their disconnection from the mainland. The first area scope was the London region on its own, as it would be easier to visualise the hotspots at this level. A larger subset of the country, henceforth referred to as South England, consisting of East of England, London, South East and South West regions was then analysed followed by all regions in England excluding those noted.

3.3 Feature Engineering

3.3.1 Dependent Variables

As previously mentioned, the objective of this research presents both a regression and classification task meaning two sets of dependent variables. The dependent variable used for regression is the count of burglary incidents. For spatial disaggregation analysis, it is essential that all variables are brought to the same scale for both aggregation levels, so the counts were converted to the number of incidents per 1,000 people. A log transformation was then performed to produce a normal distribution. Although this step is not essential for ML techniques, doing so can aid in reducing the inherent noise in crime data. This dependent variable is referred to as 'LOG_CRIME' at stages of this report.

'LOG_CRIME' is then used to classify areas into five groups based on their risk of burglary, and are referred to as 'RISK_GROUPS' within the analysis. These groups and their descriptions are described in Table 1.

Table 1. Descriptions of the 5 RISK_GROUPs derived based on crime rates, and the colour scheme used in choropleth charts in this report

RISK_GROUP	Label	Colour	Description
1	Coldspot		More than 2 standard deviations below the mean
2	Low-Risk Area		Between 1 and 2 standard deviations below the mean
3	Average		Within 1 standard deviation of the mean
4	High-Risk Area		Between 1 and 2 standard deviations above the mean
5	Hotspot		More than 2 standard deviations above the mean

3.3.2 Independent Variables

For the most part, independent features obtained from census data were already aggregated to MSOA and LSOA boundaries but some, such as number of educational establishments, had to be aggregated manually. Precipitation and temperature data consists of monthly averages for 2km grids from 1991 to 2020. To convert these to administrative boundary levels, these monthly averages were first converted to an annual average per grid. The proportion that each grids overlapped with each boundary area was calculated, and an annual average for the boundary area was derived as the total of the 2km grid averages multiplied by the overlap proportion.

The majority of census data features such as the age, level of education, etc. were converted to a proportion of the population in the area. Features such as number of educational establishments were converted to the counts per km^2, i.e. the density. For the Ordnance Survey features, built up and greenspace areas have been calculated as the percentage of the boundary area classed as each respectively, while river and road data are represented in density. A full list of the independent variables and their descriptions can be found within the Configuration manual supplied with this report.

Feature scaling was performed using RobustScaler, which rescales the data according to the 1^{st} and 3^{rd} quantile (Raschka and Mirjalili, 2019). Whilst this step is again not necessarily required with tree-based ML algorithms, it can help when utilising small or noisy datasets by making outliers less pronounced (Raschka and Mirjalili, 2019). The set up of the methodology minimises potential data leakage, which refers to information about the test data being gained at pre-processing steps such as the training/test split and data scaling (Apicella *et al.*, 2025; Rosenblatt *et al.*, 2024; Kapoor and Narayanan, 2023), as no data splitting is required and both datasets are scaled independently of each other. By approaching the project initially as a regression problem rather than classification one, it negates the potential requirement for class balancing, given there is likely considerably less hotspots than not, and possibly introducing further data leakage.

3.4 Exploratory Data Analysis

3.4.1 Spatial Autocorrelation

Spatial autocorrelation describes the First Law of Geography that attribute values of a target variable are spatially related and the locations in closer proximity would share more similar attributes than those further apart (Gao *et al.*, 2023). It can be measured at both global level, i.e. the degree of spatial clustering in the whole dataset, and local level, which highlights localised area clusters that can aid in identifying hotspots (Jordan, 2023). Fig 1 shows the global spatial autocorrelation for London, with LOG_CRIME yielding a global Moran's I statistic of 0.56715 with a p-value of 0.001 suggesting that spatial autocorrelation exists in the crime data. Fig 2 furthers the relevance of these local patterns by grouping them into LISA clusters and calculating their significance, highlighting the areas where there are statistically significant clusters of high and low crime rates.

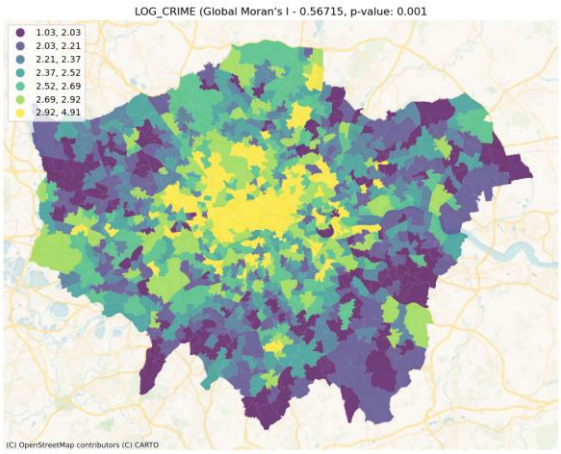

Fig 1. Global Moran's I statistic of the dependent variable 'LOG_CRIME' for the London region

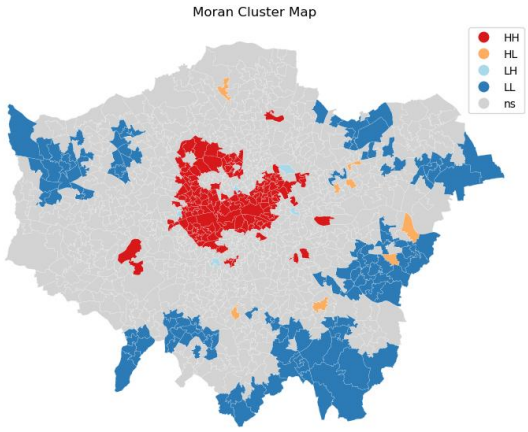

Fig 2. Statistically significant local Moran's I statistic clusters of the dependent variable 'LOG_CRIME' for the London region (HH = High rates clustered with High rates, HL = High rates with Low rates, LH = Low with High, LL = Low with Low and ns = Non-Significant

3.4.2 Feature Selection

Feature importance was investigated using three methods to identify the most suitable features for the model.

3.4.2.1 Correlation Coefficient Matrix

A Pearson correlation coefficient describes the strength and direction of the linear relationship between two continuous variables (Subramanian, 2025). Correlation coefficients were derived between each of the variables, both independent and dependent. These correlations were only used for validation of future feature selection techniques, as they ignore potential non-linear relationships and suggest considerable multicollinearity.

3.4.2.2 One Way GLMs

Generalized linear models (GLMs) possess the ability to model complex, non-linear relationships between independent and dependent variables (Yehoshua, 2025). To analyse potential feature importance accounting for non-linear relationships, a GLM was fitted with each independent variable separately and the R^2 value was calculated with the results shown in Fig 3.

3.4.2.3 Mutual Information with Variance Inflation Factor

Mutual information (MI) combined with variance inflation factor (VIF) monitoring is presented in the works of Cheng *et al.* (2022). They highlight that MI has an advantage over other methods as it calculates the shared information between variables based on any type of relationship. The inclusion of VIF is important to ensure that multicollinearity, i.e. strong relationships between independent variables, is not present in the model. Although the tree-based method used in this project is robust to multicollinearity when making predictions, highly correlated features do impact the interpretability (Khalifa, 2025; Raschka and Mirjalili, 2019). The process iteratively adds features with the highest MI score to the models features before subsequentially calculating the VIF, and removing that feature if any VIF

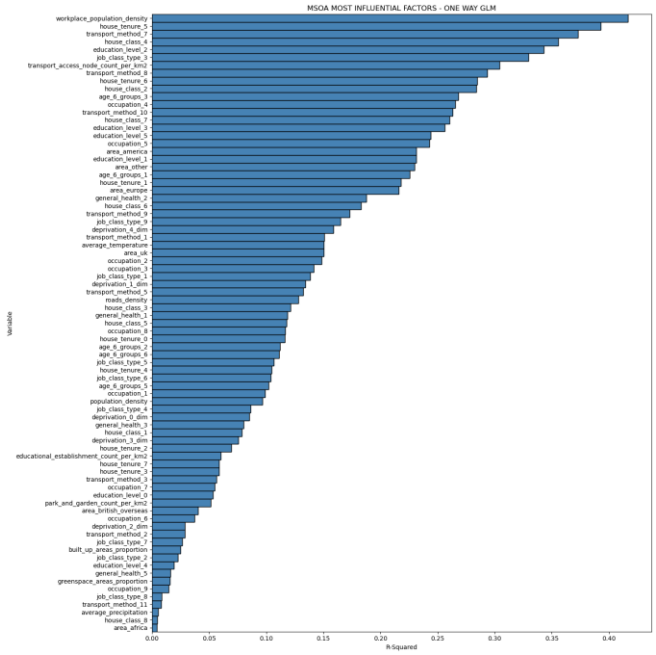

Fig 3. Suggested Feature Importance based on One-Way GLMs using London region data

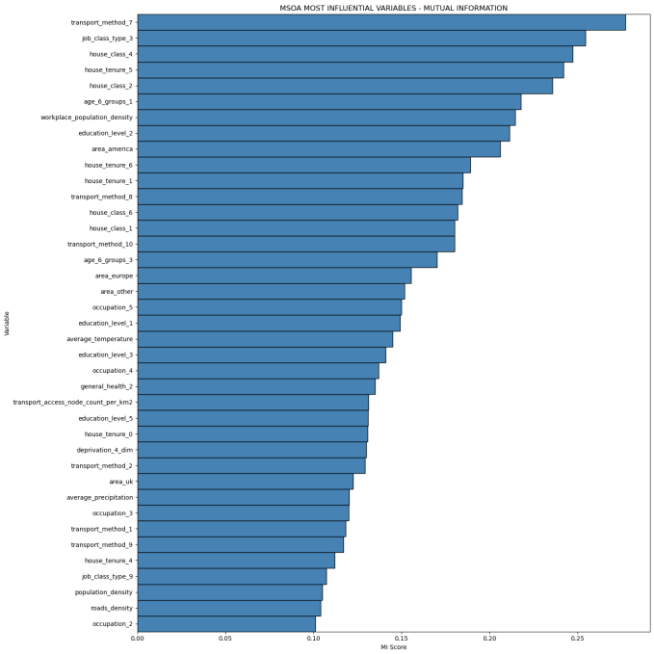

Fig 4. Suggested Feature Importance based on Mutual Information using London region data

breaches a threshold. Feature importance using MI without considering VIF is exhibited in Fig 4, suggesting similar features for selection as the one-way GLMs and correlation matrix.

3.5 Algorithms

3.5.1 Geographically Weighted LightGBM

The proposed method used in the project is a geographically weighted LightGBM (GW-LGBM). A LGBM is a tree-based ensemble algorithm that utilises gradient boosting. An ensemble algorithm is one that combines the predictions from two or more models into a single output (Delen, 2020) and generally tend to yield improved predictive power on any of the individual models while providing enhanced robustness against noise (Hearty, 2016). As crime data is generally noisy, ensemble methods present an ideal approach to crime prediction, as supported by findings by Zhang *et al.* (2022) and Khastri *et al.* (2021).

LGBMs can utilise both boosting and bagging enhancements on standard tree-based models. Bagging refers to the concept of individual trees being developed on resampled subsets of training data (Delen, 2020; Hearty, 2016). Boosting is a technique that aims to sequentially improve the performance of the ensemble by adjusting weak learners according learnings from the previous learner and, like XgBoost, LGBM does so by moving against a loss gradient to reduce the residuals of the previous weak learner (Hearty, 2016). The advantage of LGBM over other boosting techniques such as XgBoost is faster training speeds, which is important for the proposed geographically weighted method.

The GW-LGBM developed involves assigning data to each local model purely based on the distance between the points, and the weights for each are derived using a Gaussian kernel. Kernel weights are commonly used for deriving distance weights as they can provide more flexibility over other methods such as K-nearest neighbours by smoothing the influence of the weights (Rey, Arribas-Bel and Levi John Wolf, 2023). There seems to be a lack of research to suggest the most suitable kernel, but Du *et al.* (2020) suggest Gaussian is marginally better. An adaptive kernel has been implemented rather than fixed, which allows

for the bandwidth of the distances between points to vary by model (Jordan, 2023). This method seems more appropriate for this analysis given the irregular size and shape of the administrative boundaries.

3.5.2 Partial Pooling

Partial pooling is a technique often used when working with multilevel data, whereby the estimation at the lower level is influenced by the mean and variance of other lower level structures and of the coarser levels within the whole population (Gelman and Hill, 2006). Gelman and Hill (2006, pg.253) present this with the formula:

$$\hat{\alpha}_j^{multilevel} \approx \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\bar{y}_{all}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} \quad (1)$$

where n_j is the number of lower level units in the coarse level unit j, $\bar{y}_j$ is the weighted average of the lower units in j, y_{all} is the mean of all lower level units, σ_y^2 the within lower level variance and σ_α^2 is the variance amongst all coarser level units.

3.6 Evaluation Metrics

3.6.1 Regression

The ability of the model to estimate the crime rates per area are assessed using:
R^2: Measures how well the independent variables explain the dependent variable. Higher values indicates better performance.
Mean Squared Error (MSE): Average squared difference between the predicted and actual values. Lower values indicates better performance.
Mean Absolute Error (MAE): Represents the absolute difference between the predicted value and true value. Lower value indicates better performance.

3.6.2 Classification

The predicted crime rates from the regression model are classified into risk groups and the ability to predict each group is assessed using the below metrics. In all cases, higher values suggests better performance.
Precision: Proportion of correct predictions for a group out of all predictions for that group.
Recall: Proportion of actual instances of a group correctly predicted.
F1: The harmonic mean of precision and recall (Chandramouli *et al.*, 2018).

3.7 SHAP

To improve interpretability, the importance of features for each individual prediction is determined by calculating the SHAP values at the instance level. SHAP considers all possible feature combinations and their contributions to the prediction to attribute a portion of the output to each feature (Subramanian, 2025). Where predictions are made via GW-LGBM, the SHAP values are weighted in the same manner.

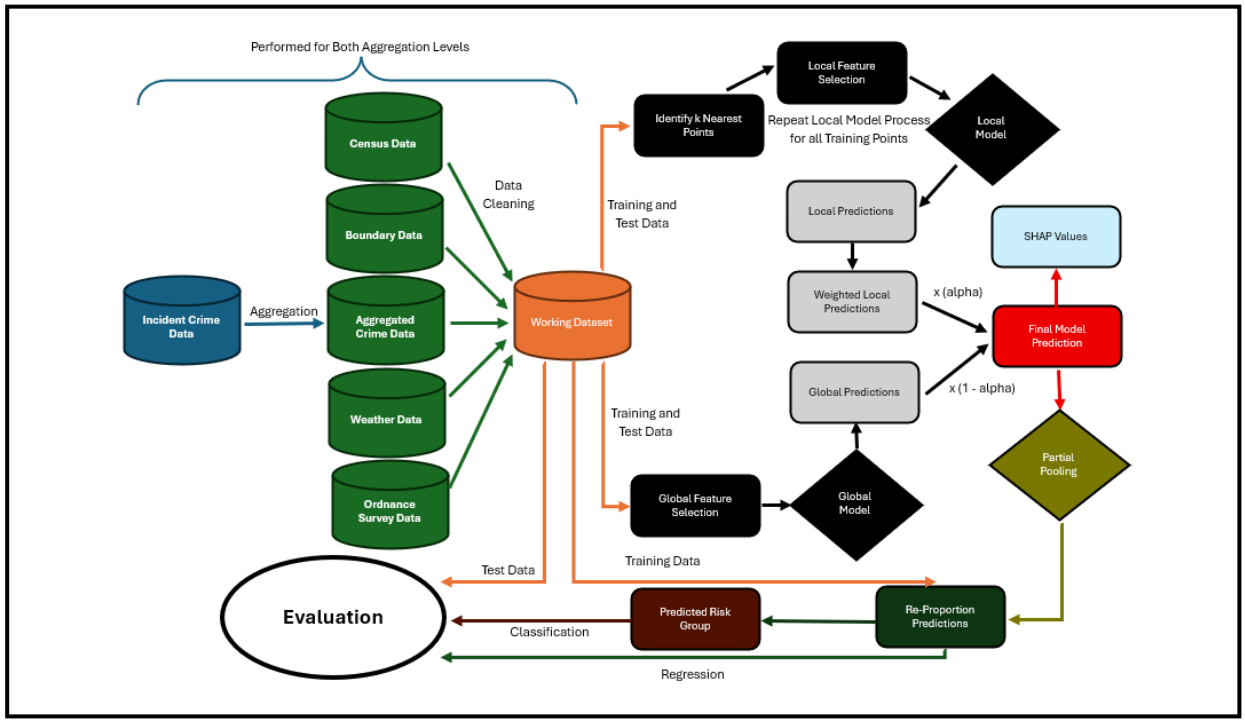

Fig 5. Design Specification Flow Chart

4 Design Specification

The full process design, excluding exploratory data analysis, is illustrated in Fig 5.

4.1 Data Cleaning

The process includes the capability to produce working datasets for any Region within England. The same process is carried out for the selected coarse and refined aggregation levels, which make up the training and test datasets respectively. All ancillary data is cleaned as described above to the requested Region boundaries.

4.2 Feature Selection

Feature selection is performed using MI and VIF as described above. The thresholds for both can be amended manually by the user, but have been set to 0.1 and 5.0 respectively with the aim of ensuring that only important features are included while avoiding introducing any multicollinearity. Research generally suggests that VIFs above 5 indicate multicollinearity.

4.3 Model Development

A global model is fitted incorporating all training data points. A local model is then fitted for each individual data point, with each model fitted with only the user defined nearest k neighbours. The importances of the nearest neighbours in the model are weighted using a Gaussian kernel, which is described by the formula (Raschka and Mirjalili, 2019):

$$k(x_i, x_j) = \exp\left(\frac{-\| x_i - x_j \|^2}{2\sigma^2}\right) \quad (2)$$

where $\|x_i - x_j\|$ represents the Euclidean distance between x_i and x_j and σ is the bandwidth parameter. This parameter has been set as the distance between the data point and its kth neighbour, i.e. the furthest distance, but can also be input as the mean of the nearest neighbour distances. Feature selection is carried out for each individual model based on the

training points selected, although this functionality can be disabled in which case the global model features are utilised throughout.

4.4 Model Predictions

Predictions are made by the global model and each local model for all of the data points. A weighted average prediction for each point is then derived by taking the predictions made by the nearest *k* models from the point and scaling the predictions using the Gaussian kernel. This *k* does not need to be same as the *k* used for fitting the models. A final prediction is calculated by combining the global and weighted local predictions, with the contribution of each defined by the value of *alpha*, i.e. *alpha* of 0 and 1 would use the full global and local predictions respectively.

4.5 SHAP Values

SHAP values can be calculated for each point based on the combined global and local predictions. These values are weighted in the same manner as the predictions, i.e. the same kernel adjustments are applied to the local SHAP values and *alpha* determines the strength of the global and local models in the final SHAP value.

4.6 Partial Pooling

The algorithm described in Section 3.4.2 has been adapted to this problem, working similarly to *alpha* in that a coefficient is calculated as to how much the prediction for the finer resolution is shrunk towards the prediction for the coarser area. Equation 3 describes the calculation of the partial pooling coefficient

$$pp\_coeff = \frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} \tag{3}$$

where the variables are as described in Section 3.4.2, and equation 4 shows how it's applied

$$pp_{pred} = pp_{coeff} * \ pred_y + (1 - \ pp_{coeff}) * pred_j \tag{4}$$

where $pred_y$ is the prediction for the finer resolution area and $pred_j$ the prediction for the coarse area it resides in.

4.7 Prediction Re-Proportioning

The final prediction at the finer resolution areas are converted back to raw counts and re-proportioned so that the sum of the predicted crime counts of the areas within a coarse aggregation matches back to the actual count for that coarse area. This reproportioning can be described by the equation:

$$y_i^* = \ y_i * \frac{Y_j}{\Sigma_{k \epsilon j} y_k} \tag{5}$$

where y_i^* represents the proportioned refined area count, y_i the predicted count for that area, $\Sigma_{k \in j} y_k$ the sum of predicted counts for refined areas in coarse area j and Y_j the actual coarse area count. The re-proportioned predictions are converted back to the log of the crime rate per 1,000 and assigned to risk groups for evaluation purposes. This step replaces standard dasymetric mapping, in that instead of directly using covariate data to proportion counts, the models' predictions are used as the basis for disaggregation, and addresses a common drawback of spatial disaggregation ML approaches whereby they tend to not account for mass-preserving.

5 Implementation

5.1 Benchmark Metrics

To evaluate the proposed methods performance, simple benchmark metrics are derived. These metrics are calculated based on the assumption that the same crime rate recorded in the coarse area applies to all finer areas within that coarse area. This assumption makes up the premise of the research, that in the absence of finer resolution data one would have to assume that crime rates are equal across all sub areas.

5.2 LightGBM Method

Six variations of the proposed design were developed and evaluated to assess different elements of the structure. Each scenario is shown in Table 2. All proposed methods include the re-proportioning of predictions step.

Table 2. Variations of LightGBM Models used

Method	Scenario
1	Global model only, i.e. non-geographically weighted, no partial pooling
2	Global model only, i.e. non-geographically weighted, with partial pooling
3	GW-LGBM with global features used for all local models (i.e. feature selection only performed for global model
4	GW-LGBM with feature selection performed for global and each local model and adaptive *k* for model evaluation set to 1
5	GW-LGBM with feature selection performed for global and each local model and adaptive k for model evaluation set to 30
6	GW-LGBM with Gaussian kernel bandwidth set to mean distance rather than max

5.3 Model Hyperparameters

The same hyperparameters shown in Table 3 were used across all methods. These parameters were manually set to adjusted standard values that might reduce potential overfitting, as per details from Yehoshua (2025). No parameter tuning was conducted as standard geospatial models present a challenge in doing so while avoiding data leakage through cross-validation, but the introduction of the multilevel element into the problem adds an extra layer of complexity and therefore it did not seem prudent to include parameter tuning without further investigation.

Table 3. Hyperparameters used for all LightGBM methods

Parameter	Selected Value
objective	regression
metric	rmse
boosting_type	gbdt
num_leaves	15
max_depth	3
learning_rate	0.05
n_estimators	300
min_data_in_leaf	5
feature_fraction	0.8
bagging_fraction	0.8
bagging_freq	1
lamda_l1	0.1
lamda_l2	0.1

6 Evaluation & Discussion

6.1 London Predictions

The results for the regression predictions for the London region only are displayed in Table 4, and suggest all methods provide reasonable improvement in estimating the crime rates in comparison to the benchmark method. The R_2 value for the GW-LGBM methods on the training and test data highlight the potential enhanced predictive power of geographically weighted models for geospatial data, although this increased benefit is diminished somewhat after the prediction re-proportioning. The more impressive indication of positive results can be seen in the classification evaluation metrics in Table 5. Precision and recall have both notably increased for the prediction of Risk Groups 4 and 5, demonstrating the methodologies capabilities of identifying areas of increased crime rates. While a precision of 55%-57% and recall of 62%-63% may seem unremarkable, the models ability to correctly classify the 78 out of the 123 of the highest risk areas taking the volatility of crime, downscaling of the study areas and severe misbalance with Risk Group 5 only making up 2.5% of the test areas, is extremely positive. The full picture of improved correct predictions, and class imbalance, is evident in the confusion matrices for the benchmark method and Method 5 in Figs 6 and 7 respectively. The predictions for Group 4 also convey improved efficiency, although there is still a significant amount of false positives to be addressed. The results for Groups 1 and 2 indicate no significant progress in capturing the reduced crime activity areas. These results also imply that the iterative feature selection when fitting each local model has merit in boosting performance, with some uplift in all metrics in comparison to method 3. Furthermore, the weighting of predictions from *k* nearest models yielded some additional uplift, although further analysis would need to be performed to confirm its true benefit. There would appear to be no additional benefit in applying partial pooling, suggest geographically weighted models offer a better solution to spatial disaggregation problems. The increased granularity of the hotspots are illustrated in Figs 8, 9 and 10 which chart the actual MSOA, actual LSOA and predicted LSOA crime rates respectively for the London area. These charts potentially highlight lessened efficiency in the models for areas previously identified as having non-significant spatial autocorrelation patterns in Fig 2.

Table 4. Regression results of each models ability to predict the crime rates in the London region

	Training			Test			Re-Proportioned Test		
Scenario	**R^2**	**MSE**	**MAE**	**R^2**	**MSE**	**MAE**	**R^2**	**MSE**	**MAE**
Benchmark	N/A	N/A	N/A	N/A	N/A	N/A	43.46%	0.22	0.36
Method 1	81.73%	0.04	0.15	28.84%	0.28	0.40	48.79%	0.20	0.34
Method 2	81.73%	0.04	0.15	28.84%	0.28	0.40	48.69%	0.20	0.34
Method 3	92.25%	0.02	0.10	31.88%	0.27	**0.39**	49.46%	0.20	0.34
Method 4	92.15%	0.02	0.10	32.84%	**0.26**	**0.39**	50.12%	0.20	**0.33**
Method 5	92.20%	0.02	0.10	33.38%	**0.26**	**0.39**	**50.47%**	**0.19**	**0.33**
Method 6	**92.94%**	**0.01**	0.10	**33.49%**	**0.26**	**0.39**	50.43%	**0.19**	**0.33**

Table 5. The Precision (Prec), Recall (Rec) and F1 Score of each models ability to classify Risk Groups based on Re-Proportioned predictions in the London region

	Risk Group 1			Risk Group 2			Risk Group 3			Risk Group 4			Risk Group 5		
Scenario	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Benchmark	27%	14%	19%	36%	38%	37%	80%	80%	80%	39%	39%	39%	46%	43%	45%
Method 1	**31%**	**16%**	**21%**	36%	41%	39%	81%	81%	81%	46%	42%	44%	50%	59%	54%
Method 2	29%	**16%**	**21%**	36%	40%	38%	81%	81%	81%	45%	43%	44%	52%	59%	56%
Method 3	29%	15%	20%	36%	41%	38%	81%	81%	81%	47%	44%	46%	52%	58%	55%
Method 4	29%	15%	20%	**37%**	41%	39%	81%	81%	81%	**48%**	**46%**	**47%**	**57%**	62%	**59%**
Method 5	29%	15%	20%	**37%**	41%	39%	81%	81%	81%	**48%**	45%	46%	55%	**63%**	**59%**
Method 6	29%	15%	20%	**37%**	41%	39%	81%	81%	81%	**48%**	45%	**47%**	55%	**63%**	**59%**

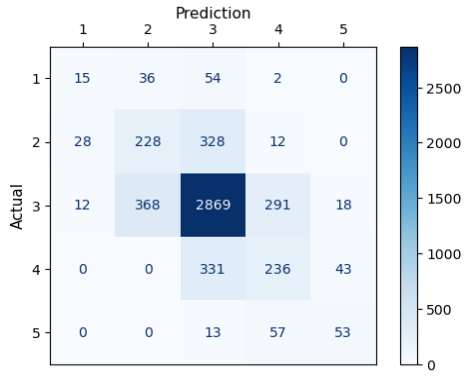

Fig 6. Confusion Matrix for categorical results of London region with Benchmark method

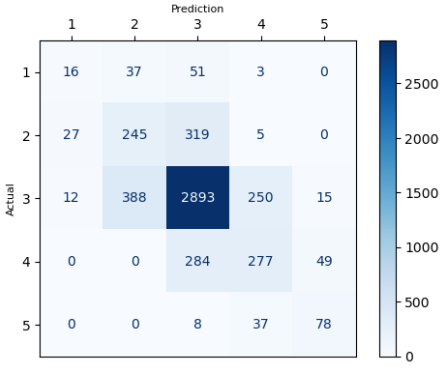

Fig 7. Confusion Matrix for categorical results of London region with Method 5

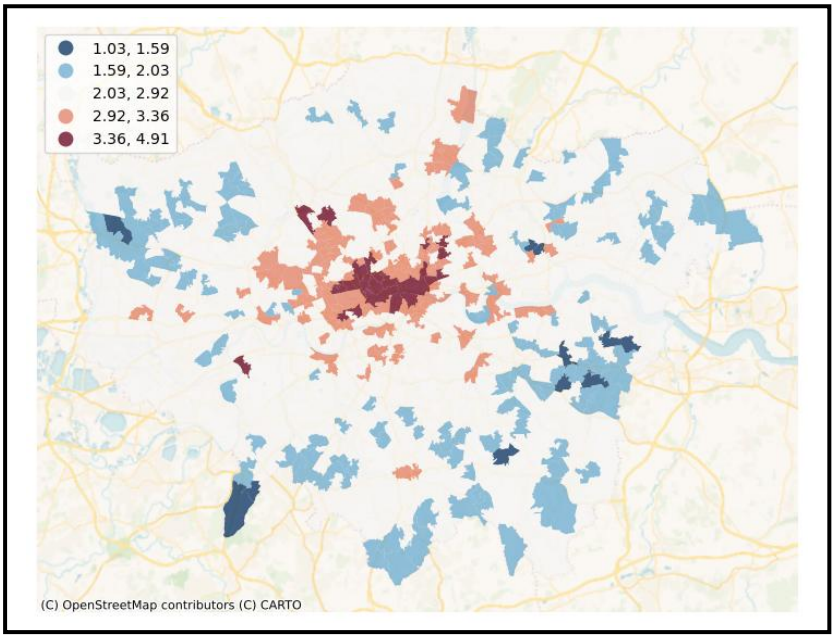

Fig 8. Actual crime rates at coarser MSOA level in London region (Colour scheme as per Table 1)

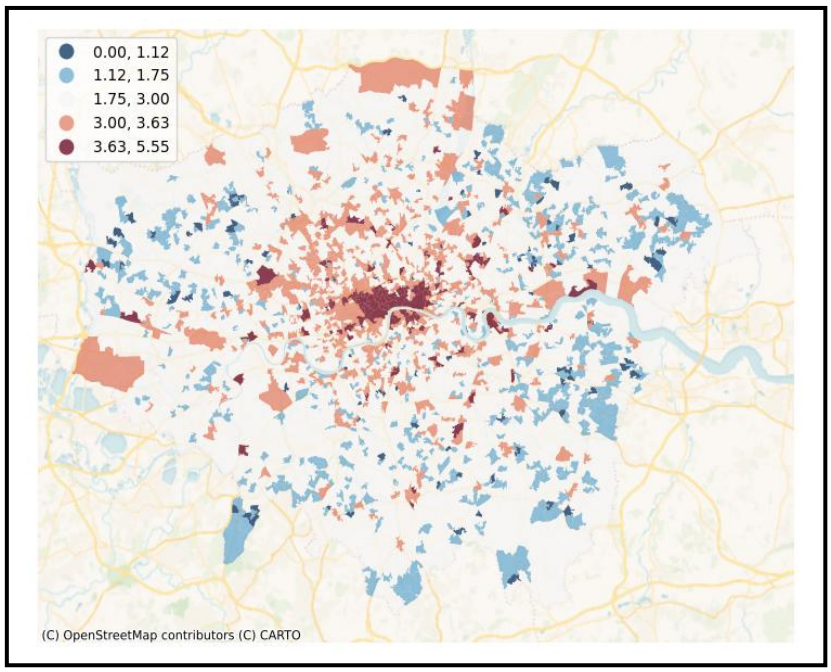

Fig 9. Actual crime rates at finer LSOA level in London region (Colour scheme as per Table 1)

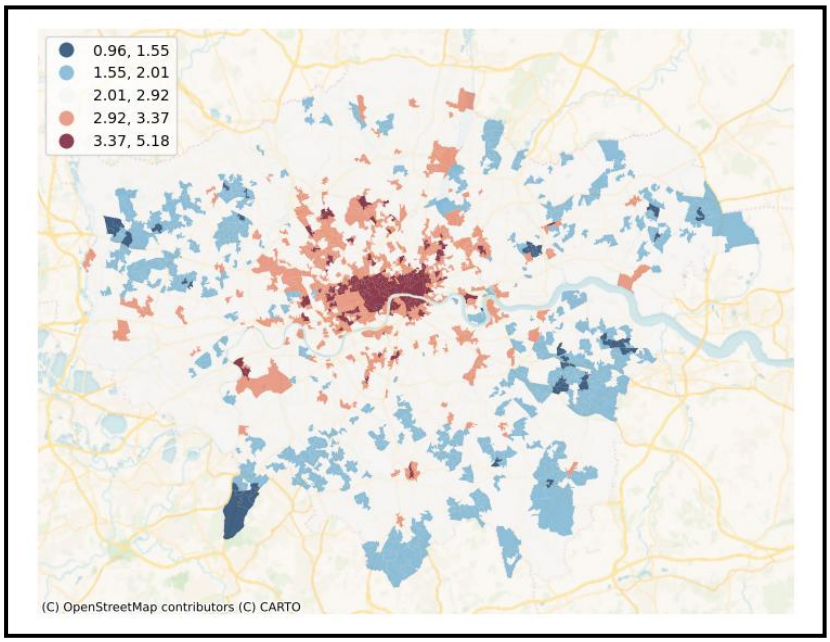

Fig 10. Predicted crime rates at finer LSOA level in London region (Colour scheme as per Table 1)

6.2 South England Predictions

The regression and classification results for 'South England' are displayed in Tables 6 and 7 respectively, again demonstrating the proposed methods potential in hotspot identification. The training and test regression results validate the use of geographically weighted models with relatively substantial improvement to R^2 metrics. However, this gain seems to be fully negated in the re-proportioning of the results. Nevertheless, a consistent R^2 of around 62% indicates valuable insights. The classification results again prove significant progress made by the method by correctly identifying up to 68% of the Group 5 areas. As before, considering the high instability in crime rates and that these Group 5 areas constitute just over 2% of all areas as shown in the confusion matrices in Figs 11 and 12, this represents

promising success. The matrix in Fig 12 also shows that only five Group 5 areas are not being identified as at least high risk Group 4. At this widened scope, the model seems to improve on capturing the cold spots as well, although still displays no particularly beneficial results.

Table 6. Regression results of each models ability to predict the crime rates in the South England region

	Training			Test			Re-Proportioned Test		
Scenario	**R^2**	**MSE**	**MAE**	**R^2**	**MSE**	**MAE**	**R^2**	**MSE**	**MAE**
Benchmark	N/A	N/A	N/A	N/A	N/A	N/A	53.08%	0.24	0.38
Method 1	76.37%	0.07	0.21	38.47%	0.31	0.43	62.01%	**0.19**	0.34
Method 2	76.37%	0.07	0.21	38.47%	0.31	0.43	61.92%	**0.19**	0.34
Method 3	**92.05%**	**0.02**	**0.12**	**45.08%**	**0.28**	**0.41**	**62.36%**	**0.19**	0.34
Method 4	90.81%	0.03	0.13	43.87%	0.29	0.41	61.33%	0.20	0.34
Method 5	91.00%	0.03	0.13	44.49%	**0.28**	**0.41**	61.83%	0.20	0.34
Method 6	91.72%	0.02	0.12	44.64%	**0.28**	**0.41**	61.80%	0.20	0.34

Table 7. The Precision (Prec), Recall (Rec) and F1 Score of each models ability to classify Risk Groups based on Re-Proportioned predictions in the South England region

	Risk Group 1			Risk Group 2			Risk Group 3			Risk Group 4			Risk Group 5		
Scenario	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Benchmark	30%	**21%**	**25%**	39%	43%	41%	81%	80%	81%	47%	48%	47%	43%	51%	47%
Method 1	**34%**	18%	24%	43%	**50%**	46%	84%	82%	83%	**54%**	53%	53%	**53%**	**68%**	**59%**
Method 2	**34%**	19%	24%	43%	49%	46%	83%	82%	83%	53%	53%	53%	**53%**	67%	**59%**
Method 3	**34%**	20%	**25%**	43%	49%	46%	84%	82%	83%	**54%**	**54%**	**54%**	**53%**	67%	**59%**
Method 4	33%	20%	**25%**	43%	49%	46%	83%	82%	83%	53%	52%	52%	52%	66%	58%
Method 5	**34%**	20%	24%	43%	49%	46%	83%	82%	83%	53%	53%	53%	52%	67%	**59%**
Method 6	33%	20%	24%	43%	49%	46%	83%	82%	83%	**54%**	53%	53%	52%	67%	58%

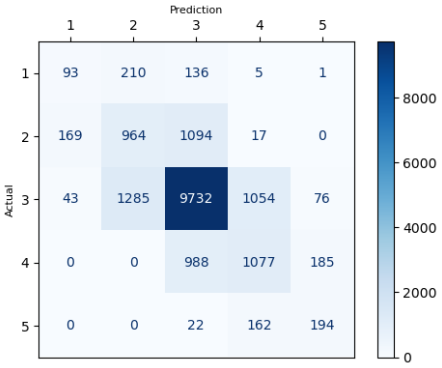

Fig 11. Confusion Matrix for categorical results of the South England region with Benchmark method

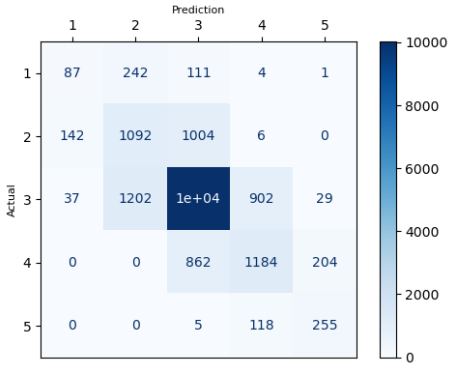

Fig 12. Confusion Matrix for categorical results of the South England region with Method 5

6.3 All Regions

The results when including all regions in England are presented in Tables 8 and 9 and Figs 13 and 14, and again show improved performance in the models over the benchmark underscoring the approaches scalability. The gains at this tier are perhaps less pronounced than the others but still illustrate positive progress. The training and test R^2 value also support the findings that a GW-LGBM offers advancements in predicting crime patterns over non-stationary ML methods.

Table 8. Regression results of each models ability to predict the crime rates in All Regions of England

	Training			Test			Re-Proportioned Test		
Scenario	**R^2**	**MSE**	**MAE**	**R^2**	**MSE**	**MAE**	**R^2**	**MSE**	**MAE**
Benchmark	N/A	N/A	N/A	N/A	N/A	N/A	54.37%	0.23	0.37
Method 1	58.05%	0.12	0.28	30.22%	0.36	0.47	59.37%	0.21	0.35
Method 2	58.05%	0.12	0.28	30.22%	0.36	0.47	59.36%	0.21	0.35
Method 3	86.25%	0.04	0.16	40.65%	0.30	0.43	60.44%	0.20	0.35
Method 4	85.84%	0.04	0.16	40.53%	0.30	0.43	60.38%	0.20	0.35
Method 5	86.01%	0.04	0.16	41.20%	0.30	0.43	**60.81%**	0.20	**0.34**
Method 6	**86.83%**	0.04	0.16	**41.39%**	0.30	0.42	60.80%	0.20	**0.34**

Table 9. The Precision (Prec), Recall (Rec) and F1 Score of each models ability to classify Risk Groups based on Re-Proportioned predictions in All Regions of England

	Risk Group 1			Risk Group 2			Risk Group 3			Risk Group 4			Risk Group 5		
Scenario	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Benchmark	38%	19%	26%	38%	45%	41%	81%	80%	81%	47%	47%	47%	39%	48%	43%
Method 1	40%	19%	26%	39%	46%	42%	83%	81%	82%	51%	51%	51%	46%	59%	51%
Method 2	40%	19%	26%	39%	46%	42%	83%	81%	82%	51%	51%	51%	46%	58%	52%
Method 3	41%	19%	26%	40%	48%	43%	83%	82%	82%	52%	52%	52%	47%	**61%**	53%
Method 4	41%	20%	27%	40%	47%	43%	83%	82%	82%	52%	51%	52%	47%	60%	53%
Method 5	41%	19%	26%	40%	47%	43%	83%	82%	82%	52%	52%	52%	47%	**61%**	53%
Method 6	41%	20%	26%	40%	47%	43%	83%	82%	82%	52%	52%	52%	**48%**	**61%**	53%

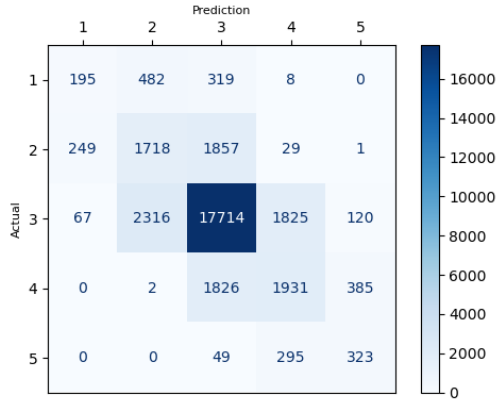

Fig 13. Confusion Matrix for categorical results of All Regions with Benchmark method

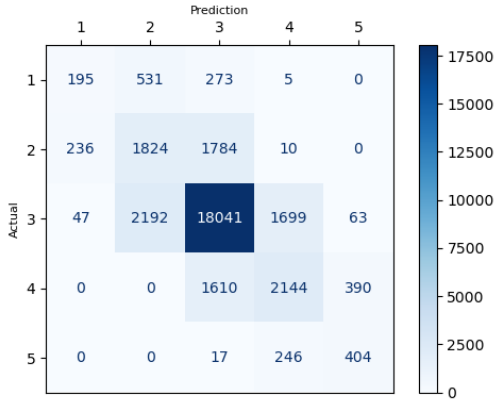

Fig 14. Confusion Matrix for categorical results of All Regions with Method 5

6.4 Model Explainability – SHAP

SHAP values calculated using Method 5 for London are presented in Fig 15 for the six features with the highest absolute average SHAP values, enhancing the models interpretability by clarifying the relative influence of each feature. In these charts red or orange depicts that the feature structure is contributing to an increase of LOG_CRIME while blue suggests a decrease. Whilst SHAP possesses the ability to derive if a positive or negative value of the features drives these values, that insight has not been included as further analysis would be needed on how to account for this in the weighting process, as well as potential ethics violations of making assumptions from them.

These results suggest the model has found 'transport_method_7' (proportion of people travelling to work by car or van), 'workplace_population_density', 'house_tenure_5' (proportion in socially rented accommodation) and 'house_class_4' (proportion in flats) are the four key influencers of increased burglary rates in London, and suggest that 'transport_method_10' (proportion that walk to work) is linked to lower burglary rates. The fact that the four most influential features are associated considerably with the increased

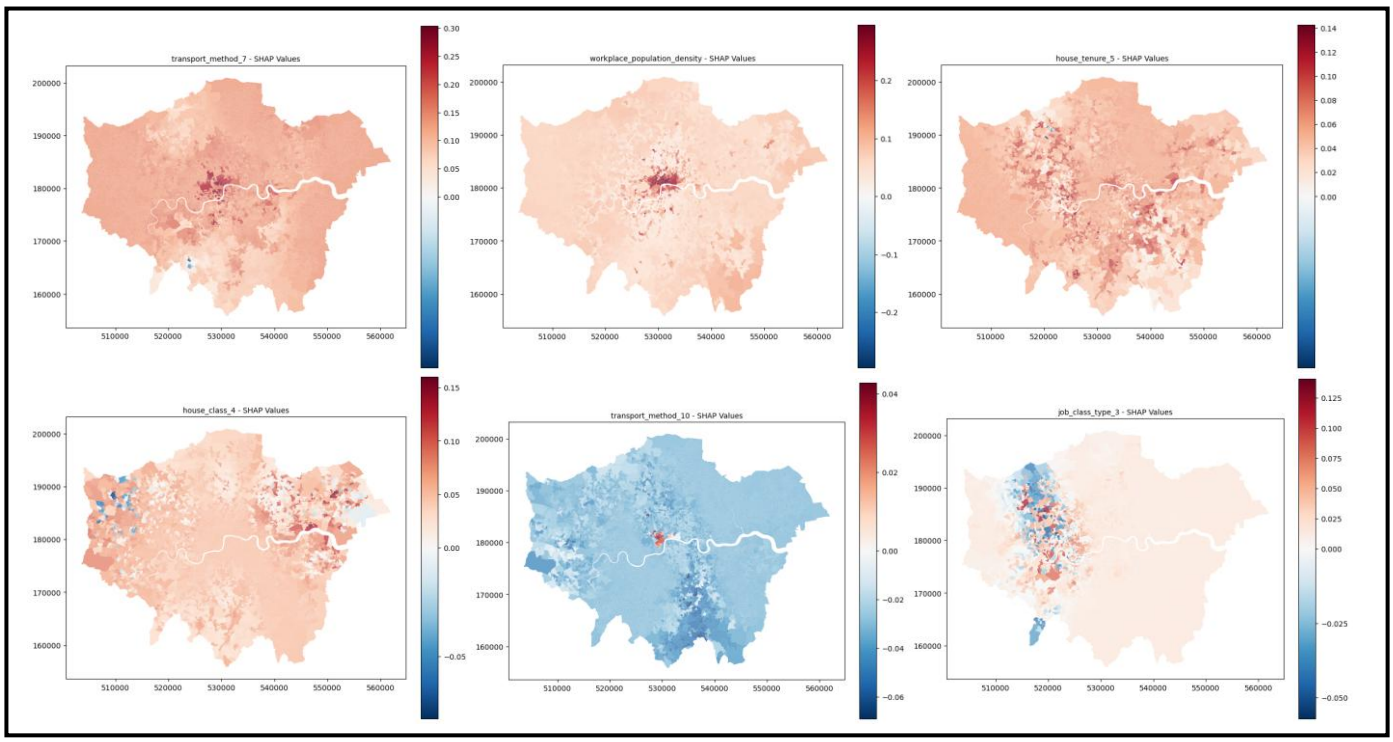

Fig. 15. Geographically weighted SHAP charts for the top 6 predictive features in London region using Method 5

crime rates magnifies the models enhanced ability to identify high risk areas and struggles to capture relationships between features and lower crime rates, or potentially highlight that these features have not been included in the analysis.

7 Conclusion and Future Work

This research aimed to investigate how ML could be utilised to accurately and interpretably identify high-risk theft hotspots when combined with spatial disaggregation techniques in the absence of granular incident-level crime data. To address the task, a geographically weighted LightGBM regression model was developed using openly available MSOA level burglary and ancillary data to predict rates at the finer LSOA scale at three different segments of England. These rates were used to classify LSOAs into risk groups, facilitating the identification of hotspots. The methodology demonstrated promising enhancements in detecting locations that represent the highest risk of burglary considering the volatility of crime rates and area downscaling, achieving 57% precision and 63% recall in the London region. SHAP values pinpoint the proportion travelling to work by car or van and workplace population density as the two main contributors to these predictions. The results demonstrate that the model offers little benefit in distinguishing lower burglary risk area, although this was not the goal of the research. The methods effectiveness seems to decrease with enlarged study areas, suggesting that it still struggles with increased noise in the data.

Future research would look to address the possible drawback of noisy data at the coarse aggregation level, potentially by employing techniques such as Bayesian smoothing. As discussed in Section 5, no work was performed on tuning either the hyperparameters of the LightGBM or the introduced parameters, such as the ideal adaptive k nearest neighbours for training and prediction, due to the concerns around cross validation in geospatial data. Some initial work regarding conducting cross validation based on spatial autocorrelation was conducted which would be continued, and potentially yielding improved results with the methods application to larger areas as k used in this method was relatively small. Another limitation of the method at increased area scopes is the potentially considerable times required for training and SHAP calculations. These could be mitigated somewhat by only fitting local models where necessary, i.e. where significant spatial autocorrelation is evident.

References

Apicella, A., Isgrò, F. and Prevete, R. (2025) 'Don't push the button! Exploring data leakage risks in machine learning and transfer learning', *Artificial Intelligence Review*, 58(11). https://doi.org/10.1007/s10462-025-11326-3

Atkins, R., Santos, R., Panagioti, M., Kontopantelis, E., Evans, J., Grigoroglou, C., Sutton, M. and Munford, L. (2025) 'Understanding the relationship between health and place: A systematic review of methods to disaggregate data to small areas', *Social Science & Medicine*, 367, p.117752. https://doi.org/10.1016/j.socscimed.2025.117752

Cohen, L.E. and Felson, M. (1979) 'Social change and crime rate trends: A routine activity approach', *American Sociological Review*, 44(4), pp.588–608. https://doi.org/10.2307/2094589

Chandramouli, S., Dutt, S. and Das, A.K. (2018) *Machine Learning*. New Delhi: Pearson Education India

Cheng, J., Sun, J., Yao, K., Xu, M. and Cao, Y. (2022), 'A variable selection method based on mutual information and variance inflation factor', *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 268, p.120652. https://doi.org/10.1016/j.saa.2021.120652

Congdon, P. (2025) 'Spatial disaggregation for relative risks, with an application to drug-related problems in North East England', *Journal of the Royal Statistical Society Series A: Statistics in Society*. https://doi.org/10.1093/jrsssa/qnaf190

Delen, D. (2020) *Predictive Analytics*. Upper Saddle River, NJ: FT Press.

Deng, Y., He, R. and Liu, Y. (2023) 'Crime risk prediction incorporating geographical spatiotemporal dependency into machine learning models', *Information Sciences*, 646, p.119414. https://doi.org/10.1016/j.ins.2023.119414

Du, Y. and Ding, N. (2023) 'A Systematic Review of Multi-Scale Spatio-Temporal Crime Prediction Methods', *ISPRS International Journal of Geo-Information*, 12(6), pp.209–209. https://doi.org/10.3390/ijgi12060209

Du, Z., Wang, Z., Wu, S., Zhang, F. and Liu, R. (2020) 'Geographically neural network weighted regression for the accurate estimation of spatial non-stationarity', *International Journal of Geographical Information Science*, 34(7), pp.1353–1377. https://doi.org/10.1080/13658816.2019.1707834

Gao, S., Hu, Y. and Li, W. (2023) *Handbook of Geospatial Artificial Intelligence*. CRC Press

Gelman, A.B. and Hill, J. (2006) *Data analysis using regression and multilevel/hierarchical models*. Pg. 253. Cambridge; New York: Cambridge University Press, p. 253

Géron, A. (2019) *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2nd edn. Sebastopol, CA: O'Reilly Media, Inc.

Hajela, G., Chawla, M. and Rasool, A. (2020), 'A Clustering Based Hotspot Identification Approach For Crime Prediction', *Procedia Computer Science*, 167, pp.1462–1470. https://doi.org/10.1016/j.procs.2020.03.357.

He, L., Páez, A., Jiao, J., An, P., Lu, C., Mao, W. and Long, D. (2020) 'Ambient Population and Larceny-Theft: A Spatial Analysis Using Mobile Phone Data', *ISPRS International Journal of Geo-Information*, 9(6), p.342. https://doi.org/10.3390/ijgi9060342

Hearty, J. (2016) *Advanced Machine Learning with Python*. Birmingham: Packt Publishing.

Huamantingo, R., Cano-Lengua, M. and Rodriguez, C. (2025) 'Machine Learning Crime Prediction Models and the Gap Between Research and Implementation: A Systematic Review', *Karbala International Journal of Modern Science*, 11(3). https://doi.org/10.33640/2405-609x.3419

İlgün, E.G. and Dener, M. (2025) 'Exploratory Data Analysis, Time Series Analysis, Crime Type Prediction, and Trend Forecasting in Crime Data using Machine Learning, Deep Learning, and Statistical Methods', *Neural Computing and Applications*, 37, pp. 11779-11798. https://doi.org/10.1007/s00521-025-11094-9

Jordan, D.S. (2023) *Applied Geospatial Data Science with Python*. Birmingham: Packt Publishing Ltd.

Kapoor, S. and Narayanan, A. (2023) 'Leakage and the reproducibility crisis in machine-learning-based science', *Patterns*, 4(9), pp.100804–100804. https://doi.org/10.1016/j.patter.2023.100804

Kim, G., Cho, Y., Lee, J.-H. and Lee, G. (2024), 'Correlation analysis between urban environment features and crime occurrence based on explainable artificial intelligence techniques', *Journal of Asian Architecture and Building Engineering*, pp.1–20. https://doi.org/10.1080/13467581.2024.2421260

Kim, J.H. (2019) ‘Multicollinearity and misleading statistical results’, *Korean Journal of Anesthesiology*, 72(6), pp.558–569. https://doi.org/10.4097/kja.19087

Khalfa, R., Theinert, N. and Hardyns, W. (2025) ‘Comparing XAI techniques for interpreting short-term burglary predictions at micro-places’, *Computational Urban Science*, 5(1). https://doi.org/10.1007/s43762-025-00185-x

Kshatri, S.S., Singh, D., Narain, B., Bhatia, S., Quasim, M.T. and Sinha, G.R. (2021), ‘An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach’. *IEEE Access*, 9, pp.67488–67500. https://doi.org/10.1109/ACCESS.2021.3075140

Kwon, E., Jung, S. and Lee, J. (2021) ‘Artificial Neural Network Model Development to Predict Theft Types in Consideration of Environmental Factors’, *ISPRS International Journal of Geo-Information*, 10(2), p.99. https://doi.org/10.3390/ijgi10020099

Lamari, Y., Freskura, B., Abdessamad, A., Eichberg, S. and de Bonviller, S. (2020) ‘Predicting Spatial Crime Occurrences through an Efficient Ensemble-Learning Model’, *ISPRS International Journal of Geo-Information*, 9(11), p.645. https://doi.org/10.3390/ijgi9110645

Liang, W., Cao, J., Chen, L., Wang, Y., Wu, J., Beheshti, A. and Tang, J. (2023) ‘Crime Prediction With Missing Data Via Spatiotemporal Regularized Tensor Decomposition’, *IEEE Transactions on Big Data*, 9(5), pp.1392–1407. https://doi.org/10.1109/tbdata.2023.3283098

Mandalapu, V., Elluri, L., Vyas, P. and Roy, N. (2023), ‘Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions’, *IEEE Access*, 11, pp.60153–60170. https://doi.org/10.1109/access.2023.3286344

Martin, O. (2024) *Bayesian Analysis with Python*. 3rd edn. Birmingham: Packt Publishing.

Meyer, H., Reudenbach, C., Wöllauer, S. and Nauss, T. (2019) ‘Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction’, *Ecological Modelling*, 411, p.108815. https://doi.org/10.1016/j.ecolmodel.2019.108815

Monteiro, J.M., Martins, B., Murrieta-Flores, P. and Pires, J.M. (2019) ‘Spatial disaggregation of historical census data leveraging multiple sources of ancillary information’, *ISPRS International Journal of Geo-Information*, 8(8), p. 327. 10.3390/ijgi8080327.

Patil, S., Pflugradt, N., Weinand, J.M., Stolten, D. and Kropp, J. (2024) ‘A systematic review of spatial disaggregation methods for climate action planning’. *Energy and AI*, 17, p.100386. doi:https://doi.org/10.1016/j.egyai.2024.100386.

Police.uk (n.d.) *Changelog*. Available at: https://data.police.uk/changelog/ [Accessed: 9 September 2025]

Raschka, S. and Mirjalili, V. (2019) *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2*. 3rd edn. Birmingham: Packt Publishing.

Rey, S.J., Arribas-Bel, D. and Wolf, L.J. (2023) *Geographic Data Science with Python*. CRC Press.

Rosenblatt, M., Link Tejavibulya, Jiang, R., Noble, S. and Scheinost, D. (2024) ‘Data leakage inflates prediction performance in connectome-based machine learning models’, *Nature Communications*, 15(1). https://doi.org/10.1038/s41467-024-46150-w.

Safat, W., Asghar, S. and Gillani, S.A. (2021), ‘Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques’, *IEEE Access*, 9, pp.1–1. https://doi.org/10.1109/access.2021.3078117

Santos, R. de V. dos, Coelho, J.V.V., Cacho, N.A.A. and de Araújo, D.S.A. (2024), ‘A Criminal Macrocause Classification Model: An enhancement for violent crime analysis considering an unbalanced dataset’, *Expert Systems with Applications*, 238, p.121702. https://doi.org/10.1016/j.eswa.2023.121702

Subramanian, V. (2025) *Applied Machine Learning for Data Science Practitioners*. Hoboken, NJ: John Wiley & Sons.

Sun, K., Zhou, R., Kim, J. and Hu, Y. (2024) ‘PyGRF: An Improved Python Geographical Random Forest Model and Case Studies in Public Health and Natural Disasters’, *Transactions in GIS*, 28(7), pp.2476–2491. https://doi.org/10.1111/tgis.13248

Tompson, L., Johnson, S., Ashby, M., Perkins, C. and Edwards, P. (2015) ‘UK Open Source Crime Data: Accuracy and Possibilities for Research’,
Cartography and Geographic Information Science, 42(2), pp.97–111. https://doi.org/10.1080/15230406.2014.972456.

Yehoshua, R. (2025) *Machine Learning Foundations, Volume 1: Supervised Learning*. Boston, MA: Addison-Wesley Professional.

Yue, H. and Chen, J. (2024) 'Interpretable Spatial Machine Learning for Understanding Spatial Heterogeneity in Factors Affecting Street Theft Crime', *Applied Geography*, 175, pp.103503–103503. https://doi.org/10.1016/j.apgeog.2024.103503

Yin, J. (2023) 'Crime Prediction Methods Based on Machine Learning: A Survey', *Computers, Materials & Continua*, 74(2), pp.4601–4629. https://doi.org/10.32604/cmc.2023.034190

Zhang, X., Liu, L., Lan, M., Song, G., Xiao, L. and Chen, J. (2022) 'Interpretable machine learning models for crime prediction', *Computers, Environment and Urban Systems*, 94, p.101789. https://doi.org/10.1016/j.compenvurbsys.2022.101789