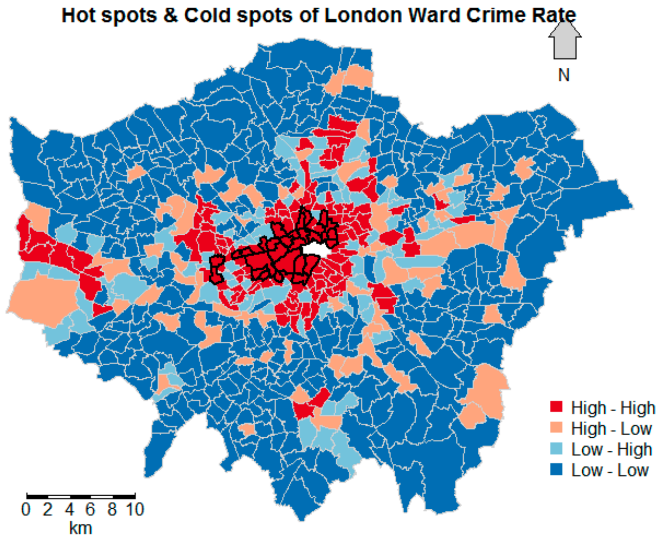
OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Predicting Urban Crime with Minimal Spatiotemporal Features Using Machine Learning

Zoe Munoz
zoe.zmg901@gmail.com

ABSTRACT

This study evaluates the efficiency of using borough and time to predict long-term crime patterns in London. The data was taken from Metropolitan Police Service records from 2008-2016. Furthermore the crime counts were aggregated annually at borough level and modeled as a function of borough and year. The aim of the study was to compare different classical machine learning, neural networks and ensemble approaches, all while testing the stability of spatial crime patterns. Linear regression, Ridge regression, Decision Tree, Random Forest, K-Nearest Neighbors, a Multilayer Perceptron, and a TensorFlow neural network were evaluated using MSE, MAE, and MAPE, alongside a mean-value baseline predictor. The results suggest that nonlinear models, particularly Decision Trees and ensemble methods out perform the baseline. This indicates structural differences between boroughs and gradual temporal change. Essentially this indicates that borough-level crime variation can be explained by minimal spatiotemporal information alone. This illustrates the predictive capacity of simple spatial-temporal structure. Furthermore the ethical need to interpret forecasts as recorded crime patterns instead of inherent characteristics of each borough.



INTRODUCTION

Crime prevention continues to be a priority in our communities, by reducing crime governments additionally improve economic growth and quality of life. Additionally studies in criminal psychology state that crime does not occur in random locations, instead happening in places where there is opportunity for crime. (Brantingham & Brantingham, 1984). However as cities grow, and socio-economic conditions evolve, forecasting these patterns has become increasingly important for public safety planning. As a result, increased availability of crime records has allowed for the opportunity to study crime trends using machine learning and deep learning methodology. This has made it possible for crime forecasting to become a valuable tool for crime worldwide. However this results in police stations spending a lot of time and resources in building effective crime prediction models. However the recent increase in digitised records have allowed for more research to be done using crime trends using machine learning and deep learning methodology. Systematic reviews report that these methods model complex spatial and temporal relationships in crime data often produce better performances than traditional statistical techniques. (Dakalbab et al., 2022) Recent studies have shown these methods provide more effective crime prediction models, requiring less time and resources and still providing accurate results.

This study focuses on a simpler version of the problem: can borough identity and year alone predict borough-level crime rate patterns over time? If this provides accurate predictions, this would suggest that borough-level crime patterns have firstly persistent differences between boroughs but also slow-moving trends. However if the model can not return accurate predictions, it suggests that more information for example population change, deprivation, and policing activity could be required to better model prediction accuracy and for interpreting any correlation claims. By testing stability in crime rates, this study explores whether crime is affected by enduring structural and environmental influences or if they are more dynamically driven by changing socio-economic conditions. While research has incorporated extensive socio-economic and demographic variables .(Zhou et al., 2023) fewer studies demonstrate how accurate predictions can be when only providing minimal spatiotemporal information such as location identity and time. By fully examining this baseline, we can further understand which crimes are based on structural crime patterns and which require additional complex explanatory variables.

Ultimately, the aim of this study is to compare Machine learning models in how well they predict borough-level crime patterns over time. This study hypothesizes that Borough-level crime patterns show long-term stability, so the distribution of crime across London boroughs will remain stable over time. Models trained on borough identity and year will hence make accurate predictions. (Zhou et al., 2023) This is an observational forecasting analysis, not a decision-support tool for policing deployment. The results will be a methodological comparison and pattern analysis rather than operational policing recommendations. Ultimately the study concludes that borough-level crime rates can be best modeled by non-linear models.

In this study it is important to note that predictive modeling of crime raises important ethical considerations. Crime is known to cluster spatially in specific hotspots instead of occurring randomly, a principle which is well established in criminology. These patterns greatly help the development of

predictive policing however, recent studies have indicated that there is a tendency for predictive crime algorithms to develop an unfair bias towards ethnic minorities. This is due to the data which police collect being implicitly or explicitly considering race when choosing which neighborhoods to patrol. Police often focus their attention on certain ethnic groups in certain boroughs. This causes an over representation of those groups in police records, the same records used to train and test these models. This results in algorithms which do not simply model crime rates but instead model the complex relationship between crime, policing strategy and racism.(Lum & Isaac 2016). Moreover scholars warn that the overrepresentation of crime in minority communities create algorithms which only perpetuate the same structural inequalities in police strategy. (Richardson et al., 2019)

The model in this study is similarly trained on policing data which undoubtedly has been affected by these biases. Although it does not use any demographic variables such as ethnicity or immigration status, the identity of each borough carries its own immigration status, ethnicity make up and wealth distribution. This will have caused a different policing strategy for each borough, creating the bias as highlighted in prior research. Therefore, the predictions generated in this study should be interpreted as forecasts of recorded crime rates, instead of direct measures of criminal activity. This factor is imperative when interpreting results to avoid overstating the objectivity of the data or implying that predictions represent inherent characteristics of specific boroughs.

During this study a dataset of crime in London from 2008-2016 was chosen. It provides a valuable case study due to its socio-economic diversity across boroughs, long term digital crime records and great variation in policing strategy across the 32 boroughs of London. This makes it preferable for testing the stability of borough-level crime structures. The models trained, tested and validated were various baseline models as well as a tensorflow model, a KNN regressor and a MLP neural network. These models were then evaluated using mean squared error, mean absolute error and mean absolute percentage error. Finally an ensemble model was created.

The crime data was released through the Metropolitan police service (MPS) recorded crime geographic breakdown dataset, which was made available via London Datastore. The dataset contains records of offences across London's 32 boroughs and the City of London, from 2008 to 2016. Additionally the data is organised by its Lower Layer Super Output Areas (LSOA) number, This is then classed by crime type, major and minor categories, and time period. Each record contains the borough, LSOA code, time variables (year and month), and the recorded crime count named “value”. In this study, the records were aggregated to their borough and year level. Crime counts were summed together across all LSOAs and months within each borough-year combination, thus producing total annual recorded crime per borough. This is used to model temporal and geographic variation in crime levels across London. The dataset has over 1,000,000 datapoints allowing for sufficient observations per borough-year. Additionally the large dataset can make performance differences more reliable as well as allowing for more stable parameter estimation. However it can also reduce sparsity due to the aggregation applied on the dataset. Moreover, there is a risk of recorded crime bias caused by shifts in recording standards or reflections of policing strategies. Finally, the optimisation of MSE penalises large errors heavily, so the performance may be influenced by hotspots of crime or structural breaks.

The aim of this study is to investigate whether borough-level crime patterns are stable over time, while additionally comparing various modeling methods that predict this stability. Crime rate is modeled as a function of spatial identity, which is encoded categorically, and time. Borough-level aggregation was chosen to reduce sparsity and improve model stability. LSOA- level data was not used because of its high dimensionality, unstable targets and extreme sparsity. This would require additional variables for instance population density, poverty. Furthermore, focusing on long-term trends through yearly aggregation reduces the influence of seasonal variation. The dependent variable is defined as the crime counts in each borough. To allow for the assessment of crime stability over time, the data was split by year. This allows for valid temporal evaluation by having the model test and validate on the later years. The validation data was the crime in 2015 and the testing data was the crime from 2016.

The dataset was imported into pandas and encoded with cp1252 to ensure correct interpretation of the characters. The borough variable was treated as a categorical predictor. The machine learning models require numerical inputs however direct numerical inputs would create artificial order. One-hot encoding ensures that there is no inherent order to the boroughs, while allowing models to learn borough-specific intercept shifts. The year therefore, was kept numeric and was not encoded because there is a natural order to time. By only one-hot encoding the borough and not the year, the pattern analysed by the model will be affected by the order of time and not any order by borough. (James et al., 2013)

A standardisation was implemented to scale the features between year and one-hot encoding numerical values. By putting the variables on comparable scales, it is preventing the models from giving more weight to large scale variables because of mathematical scale differences. It transforms each input feature x into a z-score using: $z = \frac{x-\mu}{\sigma}$

This is where μ and σ represent the mean and standard deviation of the feature calculated from the training data. The aim of the transformation is for the mean to be 0 and the standard deviation to be 1. In this study the models which were scaled were KNN, MLPregressor and TensorFlow Neural Network. This is due to these specific models depending on distance between points or gradient based optimisation. However, the rest of the models do not require this because they are based on thresholds which are not affected by variables which have not been standardised. To prevent leakage the scaler was fit only on training data. (scikit-learn, 2019)

There are a variety of classical machine learning models used as well as MLP, KNN and Tensorflow neural networks. There is a combination of linear and non-linear models which each reflect a different perspective on crime behavior. This allows them to cover a wide range of strategies to capture the behavior of crime patterns, including stable structural trends, nonlinear spatial interactions and local similarity patterns. Additionally a baseline predictor was implemented for further comparison.

First, a linear statistical model was tested as a baseline approach. This model assumes that crime levels can be predicted by spatial and temporal information. If it performs well the model would suggest that boroughs have consistent baseline crime, specifically reflecting slow changes. Conversely, if it performs poorly it suggests linear assumptions are insufficient in modeling borough crime. Therefore, this model evaluates whether spatial and temporal data is enough to explain crime patterns. To improve the stability

of coefficient estimates, Ridge regression was also implemented. The coefficients are shrinked to reduce variance and minimise over fitting due to the large number of borough indicator variables. This is done through L2 regularisation. Ridge still keeps the same linear assumptions, so by running this model, its regularisation improves the model's ability to predict stable and temporal crime patterns.

To contrast this method, the Decision Tree model was trained as it evaluates non-linear relationships. Instead it analyses and splits the data on thresholds, this is very useful in capturing relationships between borough and year. It's a non-parametric supervised learning method. This matches real crime patterns by capturing the structural complexity and not just the trends over time. If this model performs well it would suggest crime behavior is non-linear and that there are complex spatial interactions instead of only linear trends. However decision tree models can be prone to overfitting, that is why a random forest model was also produced.

A random forest model is an ensemble of decision trees. This allows it to capture non-linear relationships and model spatial structure relationships while reducing the risk of over-fitting. With the other non-linear models, it can evaluate whether borough-level stability still holds under nonlinear modeling, or if hidden complexities dominate.

Next neural networks were used to evaluate the nonlinear relationships and interactions without hard splits like trees. A multilayer perceptron is the simplest neural network evaluated, the model has many stacked hidden layers of "neurons" which can capture complex but smooth spatiotemporal structure patterns in crime. Neural networks are dependent on gradient-based optimisation, the variable features were standardised. Model hyperparameters were chosen using RandomisedSearchCV with 3-fold cross-validation and negative mean squared error as the optimisation criterion. Ultimately, MLP provides a flexible nonlinear standard to assess the extent of borough-level crime stability.

For a different approach, K-nearest neighbor (KNN) was trained as it is a distance based regressor model. Specifically it predicts crime based on similar borough-year observations. This reflects an ideology in criminology which is that similar places at similar times will exhibit comparable crime counts. This clustering behavior is well represented in KNN models. Because the model depends on distance between data points, standardisation was necessary in part because of this model as it allows for fair distance measurement. A KNN model therefore evaluates whether borough-level crime stability can be explained through local similarity structure instead of global linear trends or very complex nonlinear functions.

Finally a deep forward neural network was implemented in TensorFlow, it has multiple hidden layers therefore it has high modeling flexibility. More specifically, the neural network has the capacity to capture highly nonlinear relationships, interactions between spatial/temporal relationships and approximate complex functional forms with predefined rules. It is essentially a general function approximator making it a valuable model for modeling complex crime systems which are influenced by many interacting factors. In fact, the model tests whether crime patterns are so complex that deep linear modeling provides additional predictive power beyond classical ML. Due to the fact it relies on gradient descent optimization, standardization was essential to stabilize training and allow more balanced learning across features. Additionally, Dropout, L2 regularisation and Early sopping were applied with the intention of

minimising overfitting. It differs from sklearnMLP because it allows deeper architectures, offers greater modeling flexibility and ultimately tests whether increased complexity helps model borough-level crime patterns.

After testing the performance of each individual model, it is clear that each model is able to detect different patterns in the data, for example linear vs. non-linear. Crime systems are very complex so no single model will be able to capture all the relationships in crime patterns. This approach integrates complementary modeling assumptions, for instance linear trends, nonlinear spatial structures and local similarity patterns. The ensemble model can represent multi-mechanism crime behavior. Two ensemble models were tested, an average ensemble model and a Weighted ensemble model. First, the average ensemble involves combining the predictions of each model by taking the mean of all model outputs, in the end each model contributes equally. The average ensemble model is able to capture a balanced representation of all modeling assumptions and avoids relying too heavily on one model type. This provides a stable reference model. Secondly, a weighted ensemble model combines predictions from all the individual models previously mentioned. It then uses performance based weights, rather than equal averaging as used in the average ensemble model. The weights were derived from inverse validation mean squared error (MSE), allowing the models with stronger generalisation performance to have a bigger contribution in the final prediction. This approach evaluates whether different modelling assumptions capture complementary aspects of crime patterns. Conversely the average ensemble model evaluates whether combining different modelling perspectives without optimisation improves prediction stability.

$$w_i = \frac{1/MSE_i}{\sum(1/MSE)}$$

Dropout is a regularization technique which works by taking a portion of neurons and randomly deactivating them during each training iteration. This prevents the network from becoming too reliant on those neurons and forcing the other neurons to distribute feature learning. This ultimately forces network redundancy which gives better generalisation. However excessive dropout runs the risk of underfitting while the opposite can cause overfitting.

If there is not enough dropout there will be not enough patterns to train on and it will be underfitting the data. Regularisation is a method that is applied on neural networks to prevent overfitting, so it applies large weights in the neural network. L1 which is the final score/ accuracy minus the sum of the weights. L2 is the same thing minus the sum of the weights squared. This one will penalise the big weights. L2 is generally used more, however you don't know which one is better until you test it. Early stopping is a method to prevent overfitting by stopping the training when certain conditions are met.

There are three evaluation metrics that assess predictive performance: Mean squared error (MSE), Mean absolute error(MAE), and Mean absolute percentage error (MAPE). Mean Squared Error (MSE) is the average difference between observed and predicted values. $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$ This metric

measures the training loss for each model and heavily penalises large errors, which is relevant as crime counts exhibit occasional spikes. Mean Absolute Error (MAE) calculates the average absolute prediction error, $MAE = \frac{1}{n}\sum\left|y_i - \widehat{y_i}\right|$. MAE provides a more interpretable measure of typical prediction error, as the average given directly corresponds with proportionate data in practice. Additionally it is less sensitive to extreme outliers than MSE, offering a complementary perspective on model accuracy. Mean Absolute Percentage Error (MAPE), evaluates the average percentage deviation value. $MAPE = \frac{100}{n}\sum\left|\frac{y_i - \widehat{y_i}}{y_i}\right|$. This allows for a comparable analysis across the boroughs with different crime levels. A "safe" version of MAPE was implemented to avoid division by zero. This is achieved by replacing very small denominators with a minimal constant.

Along with MSE and MAE, a well-rounded evaluation of each model's performance is made, in each model as well as the ensemble evaluation. It's important to note that during hyperparameter tuning, MSE was minimised, using scikit-learn's "neg_mean_squared_error" scoring.

RESULTS

A baseline predictor was built and the results suggest that Machine learning models produce meaningful results.

MAE Test: 426.971
MSE Test: 318861.322

Best Parameters

Linear cvs	Linear best params: {'fit_intercept': False, 'positive': True}
Ridge	Ridge best params: {'alpha': 0.01, 'fit_intercept': False}
Dt tree	DT best params: {'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': None, 'max_depth': None}
Random forest	RF best params: {'n_estimators': 300, 'min_samples_split': 2, 'min_samples_leaf': 2,

	'max_features': 'log2', 'max_depth': None, 'bootstrap': False}
MLP	MLP best params: {'mlp__max_iter': 300, 'mlp__learning_rate_init': 0.001, 'mlp__hidden_layer_sizes': (128, 128), 'mlp__batch_size': 32, 'mlp__alpha': 1e-05} /usr/local/lib/python3.12/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:691: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (300) reached and the optimization hasn't converged yet. warnings.warn(
KNN	KNN best params: {'knn__weights': 'distance', 'knn__p': 1, 'knn__n_neighbors': 3, 'knn__metric': 'minkowski', 'knn__leaf_size': 30}
TensorFlow	🔹 TensorFlow NN (TRAIN) MSE: 18242.78125 MAE: 92.75859832763672 MAPE (%): 1575557600.0 🔹 TensorFlow NN (VALIDATION) MSE: 157651.515625 MAE: 161.5897674560547 MAPE (%): 7.8888083 🔹 TensorFlow NN (TEST) MSE: 17632.3125 MAE: 104.15953826904297 MAPE (%): 7.772748
Average ensemble	Average Ensemble - Test MSE: 12934.135549211414 Average Ensemble - Test MAE: 88.16726778505887 Average Ensemble - Test MAPE: 8.030046

Model	train_mse	val_mse	test_mse	train_mae	val_mae	test_mae	train_mape	val_mape	test_mape
DecisionTree	8157.636076	106277.200102	12293.288944	63.486920	149.916111	85.846667	3.164557e+07	7.678824	7.162026
RandomForest	6148.244609	93666.917162	12781.505701	54.476229	140.146469	83.132630	1.234007e+09	14.427736	14.883016
KNN	0.000000	124347.313551	12828.433774	0.000000	153.068307	88.091854	0.000000e+00	7.282199	7.272626
MLPRegressor	13463.647723	148944.132889	14905.427316	78.793548	144.646918	92.336689	1.423828e+09	7.278298	7.221871
TensorFlowNN	18242.781250	157651.515625	17632.312500	92.758598	161.589767	104.159538	1.575558e+09	7.888808	7.772748
LinearRegression	21598.122434	153444.118189	17665.602304	100.046768	157.527361	107.668370	7.891305e+08	8.151466	8.755205
Ridge	21593.634754	154000.803570	17671.599380	100.344946	158.819132	107.695682	8.063759e+08	7.006209	7.671861

DISCUSSION AND CONCLUSION

Overall the aim of the experiment was to predict crime rates across the 32 boroughs of London just based off of past crime records. According to the results, the decision tree model returns the best performance, with a MAPE score of 7.162026. Using a baseline model as a comparison with 426.971 MAE, it is

evident that the Decision tree is producing meaningful results. However, in this model there is a gap between training MSE (8,157) and testing MSE (106,277), suggesting overfitting. With the criteria in the early stopping the model was able to reach this score. This suggests Decision trees accurately predict crime based on temporal and spatial data. Additionally when comparing the best nonlinear and linear models, non-linear models outperform linear by 20% when comparing the MAE scores (85.35 and 107.70 respectively). This suggests that temporal crime rates are better modeled by non-linear approaches. The results support the use of AI in modeling crime rates across big cities such as London.

An analysis of mean prediction error was done to assess bias in each model. Decision Tree and MLP neural network models showed minimal bias with overprediction and underprediction rates of ~53% and ~46% respectively. However, Random forest showed an overprediction rate of ~73% indicating a tendency to overpredict the amount of crime in boroughs. Yet, the magnitude of the prediction error still was comparable to the MAE of other models, suggesting this tendency doesn't undermine overall predictive ability. Moreover, a cross-model analysis shows that the mean standard deviation across models was around 39 crimes per borough-year, this represents 3% of the average crime count. The performance gap was not driven by extreme outliers with the maximum difference in predictions being below 8% of the mean crime level. This suggests that predictions converge in different model types, consistent with structural stability in borough-level crime patterns.

Overall, the study could be taken further by applying different parameters for example smaller area boundaries like LSAO codes. Additionally, Specific features of each borough could be included such as wealth disparity, and policing strategy. Furthermore, this method could be applied to other cities with different geographical features, to see if spatial temporal crime patterns are predictable in other countries. Finally, while this research reveals how nonlinear models consistently out perform linear approaches, a further investigation of model behavior such as examining borough-level error distributions in more detail.

REFERENCES

Brantingham, P. L., & Brantingham, P. J. (1984). *Patterns in crime*. Macmillan.

Chainey, S., & Ratcliffe, J. (2005). *GIS and crime mapping*. Wiley.
(Classic text on spatial crime analysis — fits your borough mapping approach.)

Dakalbab, F., Abu Talib, M., Abu Waraga, O., Bou Nassif, A., Abbas, S., & Nasir, Q. (2022). Artificial intelligence and crime prediction: A systematic literature review. *Social Sciences & Humanities Open, 6*(1), 100342. https://doi.org/10.1016/j.ssaho.2022.100342

Eck, J. E., Clarke, R. V., & Guerette, R. T. (2007). Risky facilities: Crime concentration in homogeneous sets of establishments and facilities. *Crime Prevention Studies, 21*, 225–264.
(Supports idea that crime concentrates structurally.)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://www.stat.berkeley.edu/~rabbee/s154/ISLR_First_Printing.pdf

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance, 13*(5), 14–19.

Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime prediction using machine learning and deep learning: A systematic review and future directions. *IEEE Access, 11*, 60153–60170. https://doi.org/10.1109/ACCESS.2023.3286344

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association, 106*(493), 100–108.
(Major paper showing crime has temporal-spatial dependency.)

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. RAND Corporation.
(Key foundational predictive policing report.)

Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online, 94*, 192–233.

Scikit-learn. (2019). *StandardScaler*. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Weisburd, D. (2015). The law of crime concentration and the criminology of place. *Criminology, 53*(2), 133–157.
(Perfect support for "crime clusters in places" argument.)

Zhou, Y., Wang, F., & Zhou, S. (2023). The spatial patterns of the crime rate in London and its socio-economic influence factors. *Social Sciences, 12*(6), 340. https://doi.org/10.3390/socsci12060340