Predicting crime at micro places: Comparing machine learning methods across European cities

Robin Khalfa, Thom Snaphaan, Alina Ristea, Ourania Kounadi and Wim Hardyns

Robin Khalfa (**corresponding author**)
Ph.D. Researcher
Department of Criminology, Criminal Law and Social Law, Ghent University
Universiteitstraat 4 – 9000 Ghent (Belgium)
Email: Robin.Khalfa@UGent.be
Phone: +32 (0) 9 264 67 03

Thom Snaphaan, Ph.D.
Senior researcher and lecturer
Department of Criminology, Criminal Law and Social Law, Ghent University
Universiteitstraat 4 – 9000 Ghent (Belgium)
Associate professor
Centre of Expertise Safe and Resilient Society, Avans University of Applied Sciences, Hugo de Grootlaan 37, 5223 LB s-Hertogenbosch, the Netherlands

Alina Ristea, Ph.D.
Assistant Professor
Department of Security and Crime Science, University College London
35 Tavistock Square – WC1H 9EZ London (United Kingdom)

Ourania Kounadi, Ph.D.
Assistant Professor
Department for Geography and Regional Research, University of Vienna
Universitätsstraße 7/B0518b – 1010 Vienna (Austria)

Wim Hardyns, Ph.D.
Full Professor
Department of Criminology, Criminal Law and Social Law, Ghent University
Universiteitstraat 4 – 9000 Ghent (Belgium)

Abstract The present study compares the performance of three different supervised machine learning methods, namely an Ensemble Neural Network algorithm (ENN), a Random Forest algorithm (RF), and a K-Nearest Neighbor algorithm (KNN), in predicting residential burglary hot spots across different cities in Europe, i.e., Brussels, Vienna and London. Crime and crime-supporting data are collected for the three cities, spanning the period 2014-2016. The data are (dis)aggregated to a 200x200m grid overlaying the study areas and monthly predictions are made for each month in 2016 using the so-called rolling window approach. For each method and city, four prediction performance measures are calculated and compared (i.e., direct hit rate, near hit rate, precision and F1-scores). The results indicate that the ENN and RF algorithm achieve comparable prediction performance when predicting a smaller number of high-risk grid cells, outperforming the KNN algorithm. This suggests that in general, law enforcement agencies wishing to apply a spatiotemporal crime prediction approach should be cognizant of the enhanced performance exhibited by ensemble machine learning models such as the ENN and RF compared to non-ensemble methods such as KNN. Moreover, although the three algorithms achieved more consistent performance measures for London, no substantial differences in performance were observed across cities, suggesting that the predictive modeling approach used in this study holds premise for cross-city and cross-country application. The implications and limitations of this study are furthermore discussed.

Keywords Crime prediction – Predictive modeling – Machine learning – Spatiotemporal crime analysis – Big data policing – Algorithm

1. Introduction and background

The increasing prevalence of big data analytics in crime prevention is facilitating the development and implementation of targeted policing approaches aimed at enhancing the safety and sustainability of environments (Allam et al. 2019; Zhang et al. 2022). This trend includes the growing global emphasis on leveraging advanced statistical methods, such as machine learning, to predict crime. It is argued that this focus is of particular value due to its potential to optimize the proactive deployment of police resources and foster a shift away from 'intuition-based' decision-making to more objective forms of decision-making (Ratcliffe 2014). One main denominator to this evolution is the use of advanced statistical (big data) methods to predict where and when a new crime event is most likely to occur (Caplan and Kennedy 2016; Chainey et al. 2008; Haberman 2017; Levine 2008; Ratcliffe 2016; Rosser et al. 2017; Rummens and Hardyns 2021; Wheeler and Steenbeek 2021). This is often referred to as place-based predictive policing, which generally involves "*the use of historical data to create a spatiotemporal forecast of areas of criminality or crime hot spots that will be the basis for police resource allocation decisions with the expectation that having officers at the proposed place and time will deter or detect criminal activity"* (Ratcliffe 2014, p. 4).

Spatiotemporal crime prediction approaches are rooted within environmental criminology, and spatiotemporal criminology more specifically, and mainly draw from the empirical observation that crime is not evenly distributed in place and time (Hardyns and Khalfa 2023; Kadar et al. 2019; Rummens et al. 2017; Weisburd 2015). Regarding the spatial distribution of crime, it is argued for decades that crime tends to concentrate at only a few small micro-places or crime hot spots (e.g., Sherman et al. 1989; Weisburd et al. 2015). However, although crime hot spots usually persist 'hot' over a longer period of time (i.e., so-called chronic hot spots), crime hot spots tend to fluctuate over shorter time periods or intervals (Johnson et al. 2008). This implies that, although the same areas might be responsible for a cumulative proportion of crime over the course of a longer period, it appears that these areas are usually only exposed to crime between certain time intervals and thus that crime hot spots may dynamically pop up during so-called 'burning times' (Brantingham et al. 2020; Tompson and Coupe 2018). In that regard, spatiotemporal crime prediction approaches, such as the use of machine learning methods (e.g., neural networks) (Kounadi et al. 2020), are considered a next step into the analysis of short-term crime hot spots (Hart 2020). According to Hart, these approaches represent "a shift away from retrospective and prospective mapping techniques" (p. 98) (e.g., retrospective methods such as Local Tests of Spatial Association or Point Pattern Analysis and prospective methods such as Kernel Density Estimation or Risk Terrain Modeling), as these prioritize recent events over distant events. Additionally, machine learning models enable to use new (big) data sources such as data derived from GPS-tracking systems of police units or mobile phone data, providing enhanced opportunities to employ police resources in or near real time (Dau et al. 2022; Rummens et al. 2021).

As a result of the rapid expansion of spatiotemporal crime prediction methods and approaches, as well as their promised benefits for intelligence-led policing practices, multiple cities of different countries are increasingly experimenting with different prediction methods (e.g., Gerstner 2018; Hardyns and Khalfa 2023; Kadar et al. 2019; Mali et al. 2017). In the US, for example, place- and/or time-based crime prediction methods and applications such as Geolitica (formerly Predpol) and ShotSpotter's ResourceRouter (formely Hunchlab) have been implemented in several police departments across the country (Hardyns and Rummens 2018). Although some of these US-based applications have been empirically evaluated against the backdrop of their prediction performance and the effects of their application in terms of deploying police resources more efficiently and effectively, there is still no empirical knowledge on how well spatiotemporal crime prediction methods or applications perform across different-sized settings, how contextual differences may influence prediction performance and how crime prediction methods should or can be tailored to the setting/contexts in which they are being employed. These issues also extend to the context of Europe, where researchers and practitioners are increasingly exploring the potential of spatiotemporal crime prediction methods and approaches for intelligence-led policing. In the Netherlands, for example, the *Criminaliteits Anticipatie Systeem* (CAS) has been developed, and even though a pilot study did not show a significant reduction in crime rates, the CAS system has been rolled out regionally across all Dutch local police departments (Mali et al. 2017). This is also the case in Germany, for example, where spatiotemporal crime prediction systems have been implemented across 16 federal states, with popular applications including PreMap and PRECOBS (Gerstner 2018; Hardyns and Khalfa 2023). Hence, given that most spatiotemporal crime prediction models were initially developed within or for a specific context, it is important to assess whether these models can be transferred to other cities. In other words, can we expect consistent performance from spatiotemporal crime prediction models across different cities and countries, or does it require a more tailored approach for each specific setting (such as a specific city) and context?

A recent study in Belgium has demonstrated that using a machine learning crime prediction model, specifically an ensemble neural network, yielded relatively higher and consistent performance measures in larger and denser urban cities, while for a smaller urban city, the model was observed to be overfitted (Hardyns and Khalfa 2023). These findings are more or less in line with the study of Kadar et al. (2019), which found that a machine learning model tends to perform better in an area with high population density (in Switzerland). Accordingly, it is argued that scale effects are important to consider when building and configuring crime prediction models across several contexts, as larger and more densely populated settings generally experience more crime, and more crime events result in more information for the predictive model to learn discriminative patterns in the data. However, differences between settings regarding the performance of spatiotemporal crime prediction models may also go beyond scale effects, pertaining to other differences as regards structural characteristics such as socio-demographic or socio-economic conditions, spatiotemporal crime patterns, law enforcement practices or even country-specific characteristics such as the registration of criminal offences and the availability of crime-supporting data. In that regard, besides the efforts of a limited number of studies to evaluate the performance of a place- and/or time-based crime prediction model across several cities in one country (e.g., Kadar et al. 2019; Hardyns and Khalfa 2023), to the best of our knowledge, no study has yet applied nor evaluated the performance of spatiotemporal crime prediction models across multiple cities in different countries and thus different settings and contexts.

In this study, we focus on applying different supervised machine learning methods to derive spatiotemporal crime prediction models. Unlike geospatial and near-repeat methods such as Risk Terrain Modeling and self-exciting process models, which often apply relatively simple statistical procedures, it is argued that supervised machine learning methods are more versatile and provide enhanced opportunities to make more dynamic predictions regarding spatiotemporal crime distributions (Rummens and Hardyns 2020). This is because supervised machine learning methods are capable of learning patterns in (training) data to anticipate future patterns and enable to incorporate additional spatial and/or temporal crime-related features into the modeling process, including spatiotemporal features extracted from different big data sources. In that regard, Rummens and Hardyns (2020) found that an ensemble neural network machine learning model outperformed a Risk Terrain and near-repeat model and suggested that machine learning models might be more successful in generating both short-term and long-term predictions compared to these alternative prediction methods. With regard to the latter, however, there is currently little empirical evidence on which machine learning method performs better and is more consistent for such a predictive modeling task. Kadar et al. (2019) suggest that this absence of evidence stems from varied focuses in empirical studies. They contend that most studies have predominantly scrutinized distinct prediction horizons, particularly in terms of time and place resolutions, input features, and prediction tasks (regression vs. classification). Additionally, differences in prediction contexts may have hindered consistent parameter comparisons.

This study aims to assess and compare the prediction performance of three different machine learning methods in deriving crime prediction models predicting crime at micro places across multiple European cities, specifically Brussels (Belgium), London (UK) and Vienna (Austria). The machine learning methods employed in this study include an Ensemble Neural Network (ENN), a Random Forest (RF), and a K-Nearest Neighbor (KNN) algorithm, which were selected based on prior research within the field of place- and time-based predictive crime modeling, suggesting these methods are most commonly applied and are effective and efficient in making spatiotemporal crime predictions (e.g., Jenga et al. 2023; Kounadi et al. 2020). Ensemble machine learning methods such as ENN and RF have demonstrated improved prediction performance in contrast to using non-ensemble methods (Rummens and Hardyns 2020), rendering them useful within specific and complex domains such as crime prediction. Nevertheless, in addition to these ensemble machine learning methods, KNN is employed as a baseline machine learning method, which is also considered to be an efficient machine learning method in deriving crime prediction models aimed at predicting crime in place and time. However, although important differences exist among the ENN, RF and KNN algorithms, KNN is a relatively simple machine learning method that can offer increased transparency and interpretability in predicting outcomes compared to more complex machine learning methods like the ENN algorithm or tree-based machine learning methods like the RF algorithm (Zhang et al. 2020). Nonetheless, like ENN and RF, KNN is also capable of modeling complex non-linear relationships without heavily relying on formal assumptions about the data distributions used in predictive crime modeling. Therefore, KNN can introduce increased transparency and interpretability, while still achieving comparable or even higher prediction performance compared to more complex machine learning methods (Kuhn and Johnson 2013).

The chapter is structured as follows. First, we provide a comprehensive description of the data and methods employed in this study. This entails presenting an overview of the features used and explaining the machine learning methods that were used. Second, we present the results of the study. We begin by discussing the prediction performance of each machine learning method and for each city. Subsequently, we compare the performance measures of the crime prediction models across the various cities and methods used. Finally, we conclude the chapter by summarizing the main contributions of the study and discussing its significance in the broader context of future research and policy.

2. Data

This study uses police-registered crime data on residential burglary collected from 2014 to 2016 from the Federal police department of Brussels, the Federal Criminal Police Office of Vienna and the Metropolitan Police of London.[1] The data were obtained through requests made to the involved police departments of each respective city.[2] A summary of the central characteristics of these cities is provided in Table 1.[3]

The crime data were spatially aggregated on a 200 by 200-meter grid overlaying the cities with a monthly temporal resolution. The choice of using grid cells as operationalization of micro-places aligns with prior research on the spatial concentration of crime in European settings. In this regard, it is argued that other operationalizations of micro-places, such as street segments, are more difficult to reconcile with the geographical structure and morphology of European settings, which typically lack the gridiron pattern commonly observed in US urban landscapes. As highlighted by Hardyns et al. (2019), the absence of a gridiron pattern complicates the use of alternative micro-place operationalizations, such as street segments, due to challenges in maintaining consistent street segment lengths and subsequent variations in crime concentrations. In contrast, employing grid cells offers flexibility in adjusting the level of analysis while ensuring defined and stable boundaries within a given geographical area.

[1] The Metropolitan Police of London provided the exact date and timestamps for the crime events. For Brussels and Vienna, the midpoint of the provided time range was used to estimate the correct date and time of the crime events. We excluded residential burglaries with a time range exceeding 31 days and we also excluded residential burglaries that could not be linked to a specific grid cell

[2] The national definitions of residential burglary from each country (Austria, Belgium, and the UK) were compared to each other to ensure that data collection was equally comparable across the three cities. The main criteria was to include all thefts committed in a residence (house, flat, garage, etc.) by breaking, entering, or using false keys, with or without violence.

[3] In this study, Brussels is defined as an agglomeration of 19 municipalities that encompasses six different police departments, including Brussels city, which is the capital of Belgium and the centre of administration of the European Union.

Moreover, adopting a grid size of 200 by 200 meters as the spatial resolution and a monthly prediction window as the temporal resolution adheres to established methodologies in predictive crime modeling research, also within European contexts. This approach is underscored by previous studies such as those by Kadar et al. (2019), Hardyns and Khalfa (2023), and Rummens and Hardyns (2020), which have demonstrated the efficacy of this spatial-temporal aggregation framework within the context of crime prediction. Specifically, Rummens and Hardyns (2020) have shown that monthly predictions at a spatial grid resolution of 200 by 200-meters exhibit greater statistical reliability compared to alternative spatiotemporal resolutions.[4]

Table 1 Central characteristics of the three cities.[5]

	Brussels, Belgium	London, UK	Vienna, Austria
Population (2016)	1,187,890	8,798,957	1,840,226
Area (2016)	161 km^2	1,569 km^2	415 km^2
Population density (2016)	7,315/km^2	5,608/km^2	4,434/km^2
Number of grid cells	4,284	40,428	10,773
Number of residential burglaries (2016)	7,371	41,428	6,159

Furthermore, crime-supporting data were collected from various structural indicators for each city, including socio-economic, demographic, and environmental indicators, sourced from all national censuses of 2011 and OpenStreetMap. The decision to collect additional crime-supporting data and include structural indicators is grounded in both theoretical and empirical research. First of all, the structural indicators serve as proxies to account for underlying social disorganization and opportunity structures that have been theoretically and empirically linked to variations in spatiotemporal crime patterns (e.g., Hardyns et al. 2015; Jones and Pridemore 2019). Second, studies have shown that including additional structural indicators may increase the performance of spatiotemporal crime prediction models in general, and machine learning models more specifically (e.g., Kadar et al. 2019). Only openly and freely available data were collected for each of the three cities and in doing so, we aimed to retrieve data on structural indicators that were more or less comparable across the three cities. However, since not all indicators were available for each city, the amount and type of data collected is restricted to those indicators that were available across all three cities and were deemed most relevant for integration within the prediction models. The socio-economic and demographic indicators were only available at higher aggregated spatial scales, respectively at the statistical sector and Lower Super Output Area for Brussels and London, which are comparable to the neighborhood level, and at the district level for Vienna. Therefore, these data were disaggregated on the 200 by 200-meter grid using simple areal weighted interpolation.

[4] After data cleaning, 1,55% of the total number of residential burglaries for Brussels and 3,15% of the total number of residential burglaries for Vienna were excluded, whereas for London, only 0,09% of the total number of residential burglaries were excluded.

[5] See also Appendix 2 for an overview of the monthly numbers of residential burglary incidents spanning the period 2014-2016 in the three cities.

Environmental features, on the other hand, were available at point level and were aggregated on the 200 by 200-meter grid.

Based on this data collection, as detailed in Appendix 1, several features were created and modelled through different machine learning methods to predict residential burglary hot spots across the three cities. These features include crime history features to account for temporal crime patterns or (near) repeat victimization patterns, as crime events are more likely to occur at locations where crime has previously taken place. This is due to some places having higher crime concentrations because of more prevalent opportunities to commit certain types of crime, which are in turn related to the distribution of opportunity and structural characteristics (Brantingham and Brantingham 1995; Glasner et al. 2018). Alongside the crime history features, the models thus also include socio-economic, demographic, environmental and seasonal features to account for (the complex interplay between) social, economic, and environmental factors that are known to influence spatiotemporal crime patterns within urban areas. It is important to note, however, that the present study's focus is on comparing the methodological performance of different machine learning methods in predicting crime spatiotemporally across disparate urban settings, rather than delving into the relative importance of individual features within the models. This delineation underscores the study's scope, which precluded a comprehensive examination of feature importance—an endeavor that nevertheless requires dedicated attention in future research.

3. Methods

3.1. Predictive modeling approach

Following the data collection and construction of the input features, the data are modelled to predict residential burglary hot spots for each of the three cities comparing the three different machine learning algorithms as mentioned above and as explained in more detail below.[6] The predictive modeling of residential burglary is done using a binary target variable. The binary target variable signifies the presence or absence of residential burglaries within each grid cell on a monthly basis. The use of a binary target variable thus encapsulates a classification task, where the outcome is dichotomized into two classes: no crime (0) or crime (1). The algorithms employed in the analysis therefore predict the probability of each grid cell's classification into one of these two target classes, representing the likelihood of crime occurrence within a given spatial unit (200 by 200-meter grid cells) over the specified time frame (one month); thus, the models produce continuous probabilistic valued predictions.[7] Using such a probabilistic classification approach allows for a more conservative model performance evaluation and is very useful for determining the modeling confidence with regard to the predicted classification output (Kuhn and Johnson 2013). In this regard, it is imperative to underscore that our predictive framework does not aim to predict absolute crime counts within grid cells, as this

[6] All analyses were conducted using R statistical software (version 4.3.3.).

[7] For instance, a predicted probability of 0.80 indicates an 80% likelihood of the grid cell being classified as belonging to the 'crime' class, signifying a higher probability of burglary occurrence compared to a grid cell with a predicted probability of, say, 0.30 or 30%.

would entail a regression problem. Instead, we employ a classification approach, which offers a more straightforward method for predicting crime occurrences and mitigates the impact of low crime count scenarios.

A rolling window approach is used to model the predictions, wherein monthly predictions are made for each month of 2016 (Rummens et al. 2017). This implies that, for any given prediction month, data from the preceding two years leading up to that month are used as the training dataset, with 10-fold cross-validation being performed to determine and optimize the tuning of the hyperparameters (for an overview of K-fold cross-validation see Anguita et al. 2012).[8] Cross-validation here refers to randomly partitioning the training dataset into ten equally sized subsets, or 'folds.' The model is then trained and evaluated ten times, each time using a different fold as the validation set and the remaining nine folds as the training set.[9] The remaining data are used as test data to evaluate the performance of the model in predicting burglary hot spots.

The final training datasets contained 102,816 spatiotemporal observations for Brussels (4,284 grid cells x 24 months), 969,888 observations for London (40,412 grid cells X 24 months) and 258,576 observations for Vienna (10,774 grid cells X 24 months). However, since the binary target features in each of the respective datasets were facing class imbalance, i.e., the number of spatiotemporal observations of one specific class (in this case, no crime) outnumbered the number of observations from the other class (in this case, crime), random under-sampling was employed on the training datasets. This entails randomly selecting a subset of instances from the majority class, which is in this case the 'no-crime class', such that its size matches the size of the minority class, which is typically the class of interest (in this case the 'crime' class). This subset is then combined with all instances of the minority class to form the balanced training dataset used to train the predictive model. Exploratory analysis revealed that random under-sampling of the training datasets resulted in enhanced prediction performance of the machine learning models while concurrently reducing computational runtime. These findings are in line with the study of Kadar et al. (2019).

A fixed number of burglary hot spots is determined for each city and prediction month based on the probabilities predicted by the machine learning models. This involves ranking the predicted probabilities for each grid cell in the test dataset in descending order and defining a fixed number of grid cells as hot spots according to a predetermined area coverage percentage or level (see infra). This approach allows for a better comparison of the model's performance across the different cities and prediction months, and it is also a more pragmatic approach that can be easily adopted by police organizations, because in this way, police organizations

[8] Automatic hyperparameter tuning was applied in order to standardize the methodology employed and to increase the prediction performance obtained from applying the three machine learning algorithms. This involved an automatic search for the optimal hyperparameters of the machine learning models, aiming to enhance their performance without manually tuning the hyperparameters across the different iterations (different settings, different periods).

[9] Explorative predictive analyses showed that a 10-fold cross validation generally resulted in higher prediction performance across the three cities.

are able to better align the number of hot spots with the actual police resources available (Rummens and Hardyns 2020).

Machine learning methods

First of all, an Ensemble Neural Network (ENN) is applied, which is generally considered a robust machine learning method that involves training an ensemble of single neural networks by repeatedly fitting a single neural network with alternating random number seeds (Kuhn and Johnson 2013; Ripley 1996). This approach allows for the averaging of model scores from the individual neural networks before converting them into predicted class probabilities, which helps to mitigate overfitting and produce a more reliable model. The ENN is based on a simple artificial neural network approach, which draws from the internal representations of the human brain, enabling it to learn and identify patterns in historical data to find the function f (Bishop 1995; Ripley 1996). Neural networks typically consist of an input layer, output layer, and one or more hidden layers in between, with each layer comprising one or more nodes or neurons. The input layer encompasses all data points from the input features used in the modeling process, while the output layer presents the neural network's results as output vectors or nodes. The hidden layer(s) also typically consists of one or more neurons/nodes, each of them receiving inputs from the nodes in the previous layer (input layer or the preceding hidden layer(s)), following a mathematical operation that is performed on the inputs from which the results are passed onto the next layer. This mathematical operation involves the application of an activation function inputted by the weighted sum of the input features that are part from the preceding layer. These initial random weights are updated and fitted during the training process in order to decrease and minimize the predicted and observed outputs, often referred to as the backpropagation process. The neurons/nodes of the hidden layer are thus trained to output the classification boundaries, and these results are then combined in the output layer (Kigerl et al. 2022; Kuhn and Johnson 2013). In case of the ENN, the final result is derived from averaging the different model scores. The hyperparameters that are tuned in the modeling of the ENN are the size of the network, referring to the number of neurons in the hidden layer(s), the number of hidden layers and decay, referring to the decay rate of the weights that are updated during the training process.

In addition, the Random Forests (RF) algorithm is also an ensemble machine learning method combining the outputs of multiple decision trees to improve prediction accuracy and reduce overfitting. Unlike the ENN, which uses a neural network as a baseline method for classification, a RF aggregates the predicted outputs of different decision trees. In this regard, the RF follows a tree-based modeling logic, implying that the data are partitioned into increasingly smaller subsamples through a process initiated by nested 'if-then-statements'. As Wheeler and Steenbeek (2021) mention, these statements can be visualized as trees: “starting at the top, the decisions branch out downwards until the final prediction is reached” (p.451). In that regard, tree-based methods split the data into several terminal nodes or leaves of the tree and for each new subsample of the data, a prediction is obtained following the ‘if-then-statements’ while using the values of the predictors until a terminal node is reached. The predictions are then generated by using the model formula in this terminal node. However, unlike basic single classification trees, RF classifiers go one step further by aggregating the results of the different trees that were trained. The training of the trees is done using bootstrapped samples of the training datasets, implying that the different subsamples drawn from the original training dataset are randomly resampled to decrease tree correlation and further reduce variance (Berk 2013; Berk and Bleich 2013; Kigerl et al. 2022; Kuhn and Johnson 2013; Mohler and Porter 2018). The hyperparameters that are tuned for the RF model are the number of trees, the number of features randomly sampled at each split, the number of samples that need to be drawn, the number of data points in each terminal node and eventually, the maximum number of nodes.

Finally, the use of a K-Nearest Neighbors algorithm (KNN), for classification tasks differs from using the ENN or the RF. KNN uses a sample's 'geographic neighborhood' to predict classes or class probabilities (Kigerl et al. 2022). In this regard, KNN predicts samples by using the K-closest samples from the training data, based on specific distance metrics such as the Euclidean distance, used to determine the distance between samples. In other words, the classification of a sample is based on the classification of its K-nearest neighbors in the training data, and the new predicted class equals the class with the highest estimated probability. The number of Ks, referring to the number of K-nearest neighbor samples selected, largely determines the model fit (Zhang et al. 2020). As a result, it is crucial to carefully select the number of Ks to avoid under- and overfitting of the data. In this study, a weighted KNN method is applied, implying that additional kernel functions are used to weight the KNNs based on their distances and proximity to each other. The tuning hyperparameter that is set is the number of K-neighbors used in the modeling process, and this hyperparameter determines the model's complexity and generalization performance.

3.2. Prediction performance measures

The prediction performance of the derived machine learning models is assessed based on four specific prediction performance measures, including the direct hit rate or recall, the near hit rate, precision, and the F1 score (Hardyns and Khalfa 2023; Rummens and Hardyns 2020). The direct hit rate or recall measures the proportion of correctly predicted crime hot spots relative to the actual number of areas in which an actual criminal event occurred. The near hit rate is a less restrictive measure as it considers crime events that occurred in grid cells adjacent to the predicted burglary hot spots. Both the direct hit and near hit rates are commonly referred to as measures of sensitivity in literature. Furthermore, precision reflects the proportion of correct predictions made relative to the total number of predictions made, indicating the model's efficiency. A good prediction model is characterized by high scores across all performance measures. However, the values of the first three indicators are dependent on the number of hot spots predicted, meaning that the direct and near hit rates will generally be higher with more predicted hot spots, while precision will be lower, and vice versa, for a lower number of predicted hot spots. Therefore, the F1 score is used to strike a balance between recall and precision. The F1 score is the harmonic average of the precision and the direct hit rate and ranges from 0 to 1, with higher values indicating a better balance and prediction performance equilibrium between recall and precision. By applying the same performance measures to compare the performance of the three machine learning model across the three cities, the aim is to establish a standardized basis for model evaluation. Although the performance measures should be interpreted independently of each other, as they each highlight different facets of model performance, collectively, however, they offer a more holistic assessment of model performance.

4. Results

4.1. The spatial concentration of residential burglary

Prior to presenting the predictive modeling results, we present descriptive statistics on the spatial concentration of residential burglaries in Brussels, London, and Vienna spanning the period 2014 to 2016. As recommended by Bernasco and Steenbeek (2017), we use the Gini coefficient for standardized crime concentration reporting. The Gini coefficient provides insights into cumulative crime distributions at micro-places. In general, the Gini coefficient quantifies the proportion of crime falling within a specific distribution of places. The Gini coefficient ranges between 0 and 1, with 0 indicating perfect equality and 1 signifying perfect inequality, where all crimes are concentrated in one place. The Gini coefficient (see equation 1) can be calculated using the following formula:

$$\boldsymbol{G} = \left(\frac{\boldsymbol{1}}{\boldsymbol{n}}\right)\ \left(\boldsymbol{2}\sum_{i=1}^{n} \boldsymbol{i}\boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{n} - \boldsymbol{1}\right) \quad \textbf{(1)}$$

where G is the Gini coefficient and n is the total number of places, y_i is the fraction of crimes that occur in place i, and i is the place's rank order when places are sorted by the number of crimes represented by y (Bernasco and Steenbeek 2017). However, due to having cases where there are more places than crimes, we present Generalized Gini coefficients, as proposed by Bernasco and Steenbeek (2017). Referring to the original Gini coefficient, the generalized Gini coefficient (G') can be calculated as a function of the original Gini coefficient (G), the number of places (n) and the number of crimes (c), using the following formula (see equation 2):

$$\boldsymbol{G'} = \mathbf{max}\left(\frac{\boldsymbol{n}}{\boldsymbol{c}}, \boldsymbol{1}\right)(\boldsymbol{G} - \boldsymbol{1}) + \boldsymbol{1} \quad \textbf{(2)}$$

Table 2 shows that residential burglaries are most concentrated in Vienna, with a small cumulative proportion of grid cells accounting for a large cumulative proportion of burglaries, especially over the three-year period (2014-2016). For individual years, slightly lower Gini coefficients are observed, but still indicate high burglary concentration. For London, the Gini coefficients highlight high burglary concentration, especially when examining yearly counts versus three-year aggregation. Interestingly, London's Gini coefficients exhibit greater stability from 2014 to 2016 compared to Vienna and Brussels. In this regard, lower residential burglary concentration is observed for Brussels, suggesting a less concentrated pattern with a larger number of contributing grid cells. In general, these descriptive statistics reveal differences in crime concentrations across the three cities, which may provide more context to the predictive modeling results.

Table 2 Residential burglary concentrations at the grid cell level (Gini coefficients).

	Brussels, Belgium	London, UK	Vienna, Austria
2014	0.676	0.764	0.791*
2016	0.692	0.773	0.761*
2016	0.716	0.775	0.735*
2014-2016	0.643	0.724	0.809

* Generalized Gini coefficient[10]

4.2. Prediction performance across machine learning methods within cities

Turning to the results of the predictive modeling, we first present the performance of the models derived for each of the three cities in Figure 1, comparing the performance of the different algorithms in deriving a spatiotemporal crime prediction model. The four performance measures are presented in accordance with the fixed number of hot spots that are predicted. As previously stated, the number of predicted burglary hot spots for each prediction month and city is fixed and is determined by sorting the predicted probabilities generated by the machine learning models in descending order. Specifically, the models predict the top 5, 10, 20, and 50% of grid cells with the highest predicted probabilities out of the total number of grid cells.[11] In order to present the findings more succinctly, the resulting means of the performance measures, averaged across the 12 prediction months, are presented visually.

Brussels

In the case of Brussels, both the ENN and RF algorithm demonstrate increased performance compared to the KNN algorithm in terms of direct hit rate. On average, they correctly predict approximately 20% of grid cells associated with residential burglaries, while KNN achieves a lower rate of around 15%. This advantage diminishes beyond predicting 50% of the total grid cells. Notably, the KNN algorithm performs worse than ENN and RF algorithms when predicting 5%, 10%, and 20% of grid cells. Regarding the near hit rates, KNN exhibits higher rates compared to ENN and RF algorithms, suggesting its ability to predict adjacent grid cells to actual burglary hot spots more accurately. However, predicting a larger number of hot spots somehow compromises model precision, leading to increased false positives. In this context, both ENN and RF algorithms outperform KNN, with ENN slightly surpassing RF on a monthly average. As the number of predictions increases, the precision of all three methods converges, making KNN less advantageous for predicting smaller numbers of hot spots. Similarly, F1-scores increase with more predicted hot spots, with ENN and RF outperforming KNN. F1-

[10] Here, generalized Gini coefficients are presented due to having cases where there are more places than crimes.

[11] For instance, consider a city encompassing a total number of 4000 grid cells. In this scenario, the top 5% of hot spots equals the 200 grid cells with the highest predicted probabilities. This method ensures a standardized approach to hot spot identification, as all grid cells maintain uniform dimensions of 200 by 200 meters.

scores of models derived from all three algorithms converge when predicting over 20% of total grid cells, with ENN showing the highest scores but with no substantial difference from RF.

London

For London, both the ENN and RF algorithm display similar patterns in direct hit rates, with ENN slightly outperforming RF. On average, they accurately predict around 25% of burglary hot spots when predicting the top 5%, while KNN performs worse in direct hit rates. A convergence in direct hit rates across all algorithms is observed when more hot spots are predicted. While there is a noticeable difference in direct hit rates, the disparity in near hit rates between the algorithms is less substantial. The three algorithms exhibit comparable near hit rates, with ENN and KNN slightly outperforming RF, particularly when fewer predictions are made. In terms of precision, both ENN and RF outperform KNN, especially when predicting fewer hot spots, indicating KNN introduces more false positives. ENN demonstrates higher efficiency in predicting the top 5% hot spots and slightly outperforms RF, with approximately 32% of ENN predictions being correct on average. Regarding F1-scores, the difference between ENN and RF is less pronounced, yet both outperform KNN when predicting the top 5% of burglary hot spots. F1-scores increase with more predicted hot spots across all models.

Vienna

In case of Vienna, the ENN and RF algorithms exhibit slightly better direct hit rates compared to the KNN algorithm, particularly noticeable for the top 5% of hot spots. However, as the proportion of predicted hot spots increases, the differences in direct hit rates among the algorithms diminish. Conversely, the KNN algorithm shows a marginal advantage over the ENN and RF algorithms in near hit rates for the top 5% of burglary hot spots. Although these differences are relatively minor, they suggest that when assessing model performance with less strict hit rate criteria, the efficacy of employing a specific algorithm for accurately predicting burglary hot spots may diminish. However, all three algorithms demonstrate low precision in modeling, especially evident when targeting the top 10% of hot spots. With respect to the F1-scores, the same conclusions can be drawn. The ENN algorithm outperforms the RF and KNN algorithm when predicting a smaller number of hot spots, yet the F1-scores remain below the 30% on a monthly average when predicting the top 5% of hot spots.

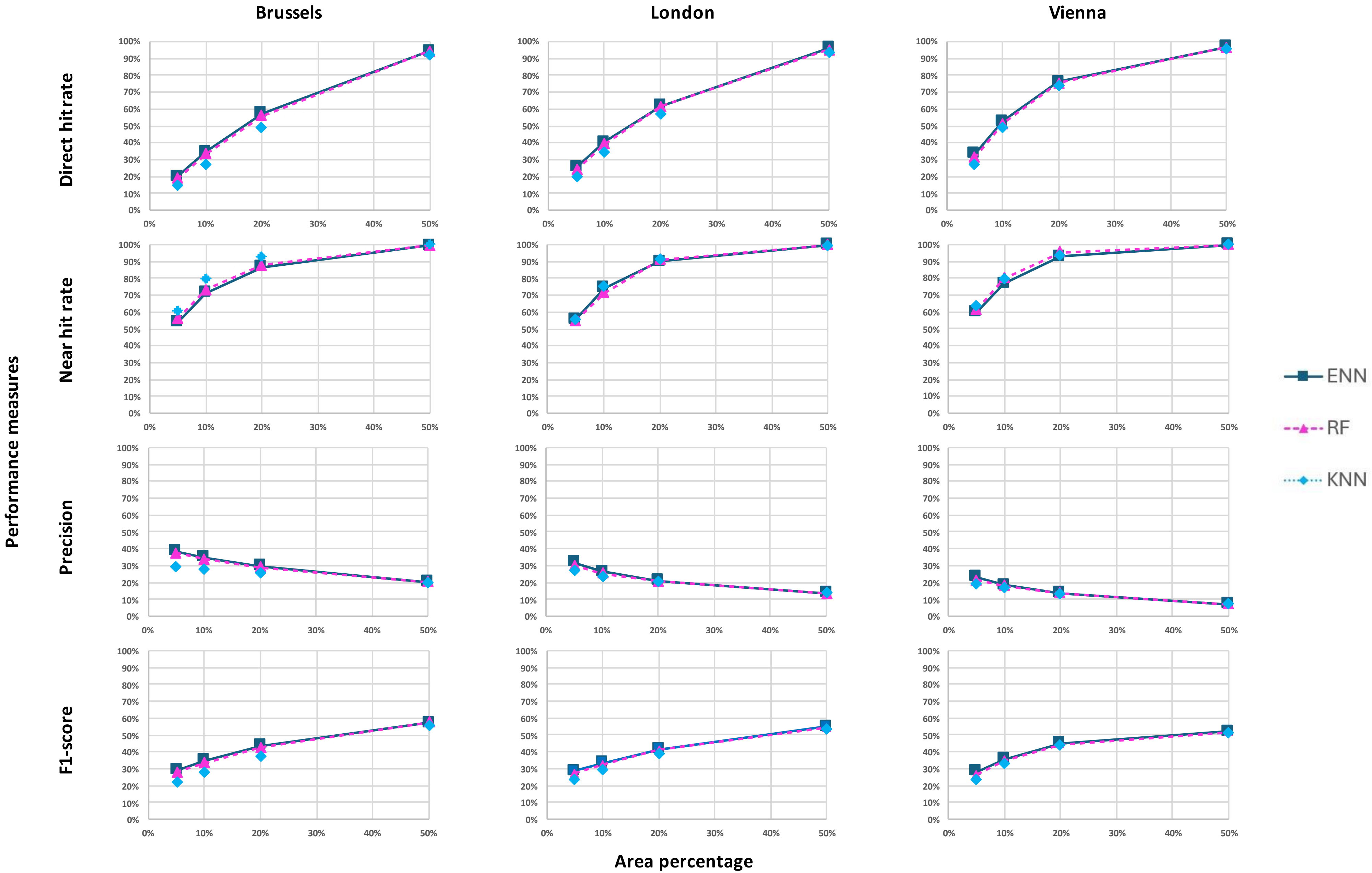


Figure 1. Prediction performance measures across cities and methods for different percentages of area coverage.

4.3. *Comparing prediction performance across machine learning methods and cities*

In addition to evaluating the prediction performance of various machine learning methods within each specific city, we now compare the prediction performance across three European cities for each machine learning method used. Instead of computing average performance measures over the prediction months, we visually present the prediction performance measures for each individual month in 2016. In doing so, each of the four prediction performance measures will be discussed separately. To enable a comparative analysis across the different cities, the number of burglary hot spots predicted for each prediction month and city is again fixed, but now the number of hot spots that are predicted is only based upon the number of grid cells accounting for 10% of the total number of grid cells, rank-ordered in descending order.

Direct hit rates

Figure 2 illustrates the direct hit rates obtained from applying the three machine learning algorithms across Vienna, Brussels, and London throughout 2016. The ENN algorithm consistently achieves higher direct hit rates for Vienna compared to Brussels and London. Over time, the direct hit rates gradually increase for London and Brussels, with more consistent rates observed for London, peaking in June and July. Conversely, Vienna's ENN models exhibit an upward trend in direct hit rates from April 2016, followed by a decline after July 2016. Similar patterns are observed for models generated using the RF algorithm, although initially yielding lower rates for Brussels and Vienna, which gradually increase towards the summer. In contrast, direct hit rates for London display greater consistency throughout 2016, with higher rates during the summer. However, the KNN algorithm falls short in direct hit rate performance, particularly for Brussels, where it struggles to predict more than 30% of burglary grid cells accurately. Nonetheless, it introduces a more consistent pattern of direct hit rates across the prediction months in 2016. For London and Vienna, although exhibiting slightly lower rates, the KNN algorithm follows similar patterns to the ENN and RF algorithms.

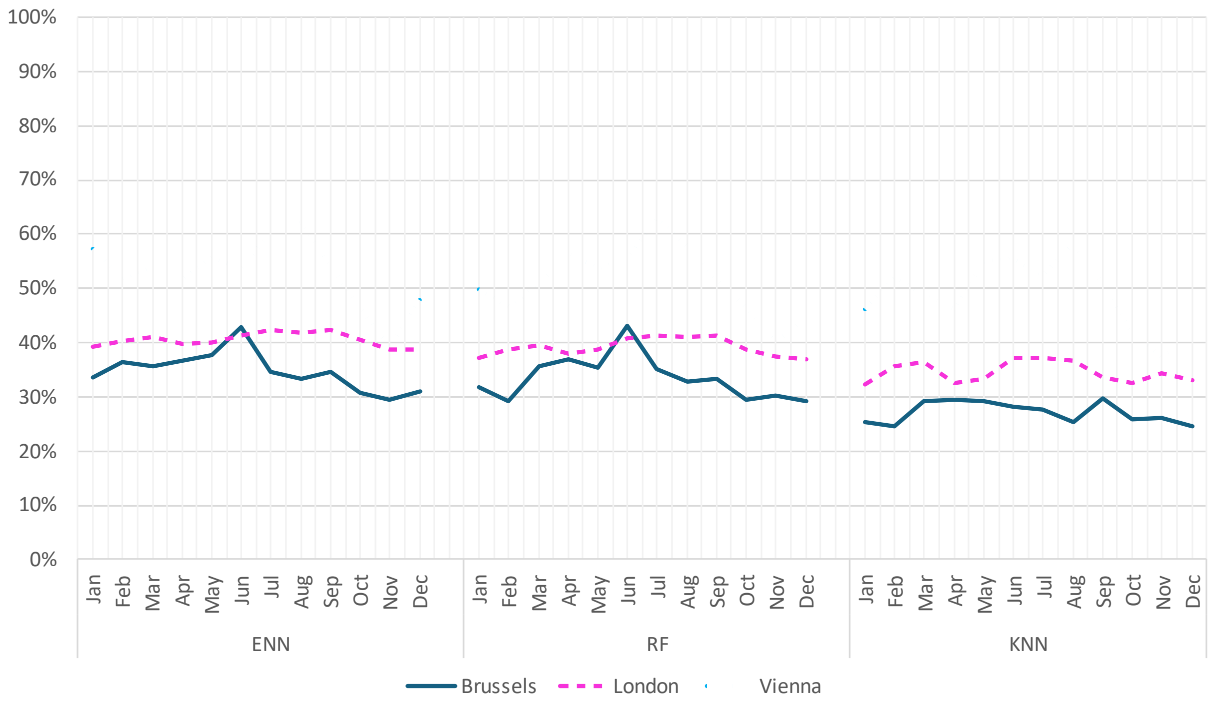


Fig. 2 Direct hit rates across machine learning methods and cities.

Near hit rates

Similar to the direct hit rates, the ENN algorithm consistently achieves more consistent near hit rates for London as can be seen from Figure 3, whereas for Brussels and Vienna, more fluctuations can be observed. Nevertheless, we clearly see that in general, despite the sudden drop in October 2016, higher near hit rates are achieved for Vienna. The same patterns are somehow observed for the near hit rates obtained from the RF algorithm compared to the ENN algorithm, although we clearly see that the near hit rates are higher for Vienna. For Brussels and London, there is almost no difference in near hit rates obtained from the ENN and RF algorithms, suggesting both algorithms produce very similar near hit rates throughout 2016. Interestingly, the KNN algorithm yields the highest hit rates throughout the year, particularly for Brussels. Between May and June 2016, the KNN algorithm accurately predicts nearly 85% of burglary grid cells for Brussels when considering adjacent grid cells as correctly predicted.

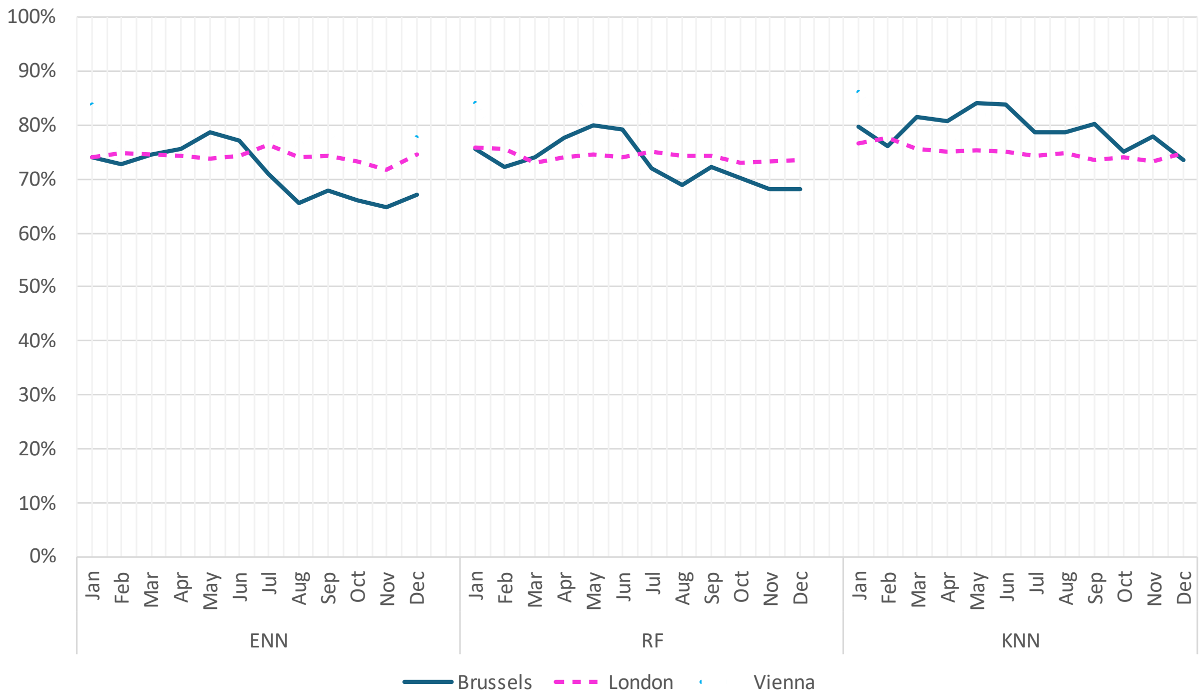


Fig. 3 Near hit rates across machine learning methods and cities.

Precision

Figure 4 shows that in terms of modeling precision, the ENN algorithm demonstrates higher precision scores for Brussels, indicating fewer false positives compared to London and Vienna. However, it is worth noting that Brussels exhibits more fluctuations in modeling precision compared to London. The same precision patterns are somehow produced by the RF algorithm: a higher precision is observed for Brussels with more consistent precision scores for London, and lower precision scores for Vienna. The precision scores for Vienna only tend to increase towards the end of 2016, which again shows that there are more false positive predictions for Vienna compared to the other cities. Moreover, based upon the precision scores produced by applying the KNN algorithm, we only observe significantly lower precision scores for Brussels compared to the ENN and RF algorithm, whereas for London and Vienna, more or less the same precision scores are observed compared to the ENN and RF algorithm.

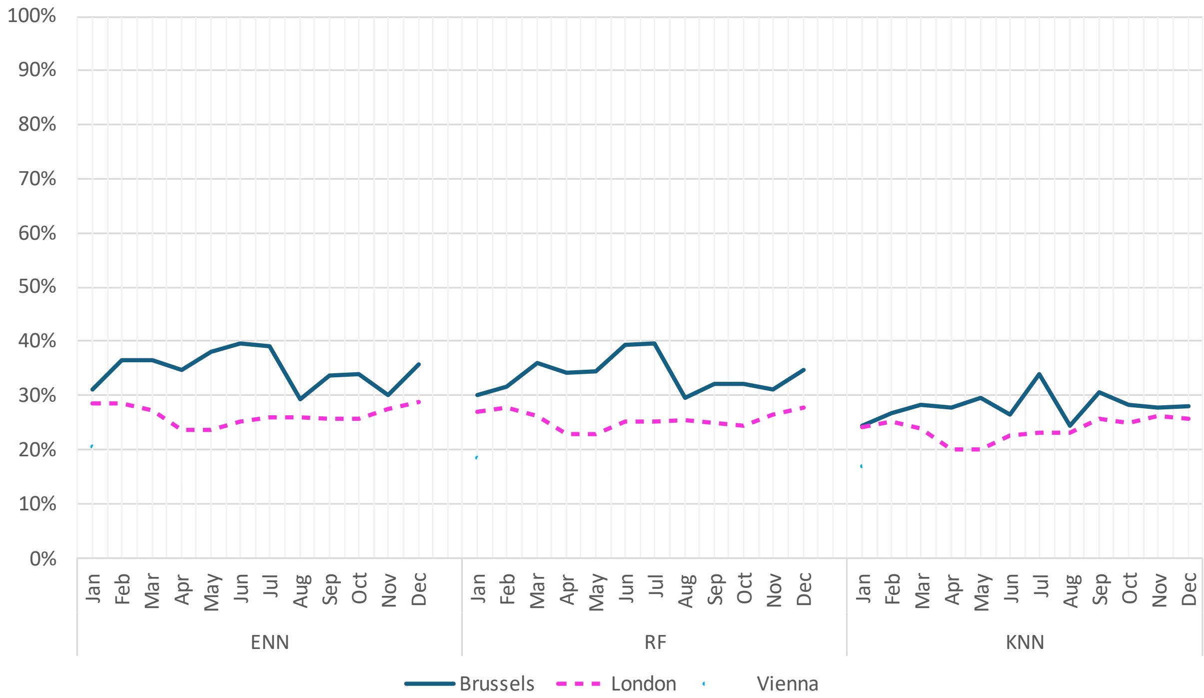


Fig. 4 Precision scores across machine learning methods and cities.

F1-scores

In terms of F1-scores, Figure 5 shows that the ENN algorithm performs quite similar across the three cities, yet with more consistent scores across the prediction months for London and higher fluctuations for Brussels and Vienna. As previously observed, the summer period demonstrates increased performance, followed by a slight decline in modeling performance towards the end of the year. Notably, London exhibits more consistent modeling performance throughout 2016, with fewer fluctuations overall. The same is true for the F1-scores derived from applying the RF algorithm. The RF algorithm produces a more comparable pattern of F1-scores for London, whereas for Brussels and Vienna, the F1-scores are slightly lower, especially after June 2016. However, it seems that for all cities, a better equilibrium between the direct hit rates and the precision scores is achieved throughout the summer period and in December 2016. In case of the KNN algorithm, a slightly different picture is observed. Overall, the F1 scores are slightly lower for Brussels and London, whereas for Vienna, more or less the same patterns are observed compared to the patterns obtained from the ENN and RF algorithm. However, compared to the scores produced by the ENN and the RF algorithm, fewer fluctuations in F1-scores are now observed for Brussels and London. Only in July and September 2016, the F1-scores are slightly higher for Brussels compared to the F1-scores for London.

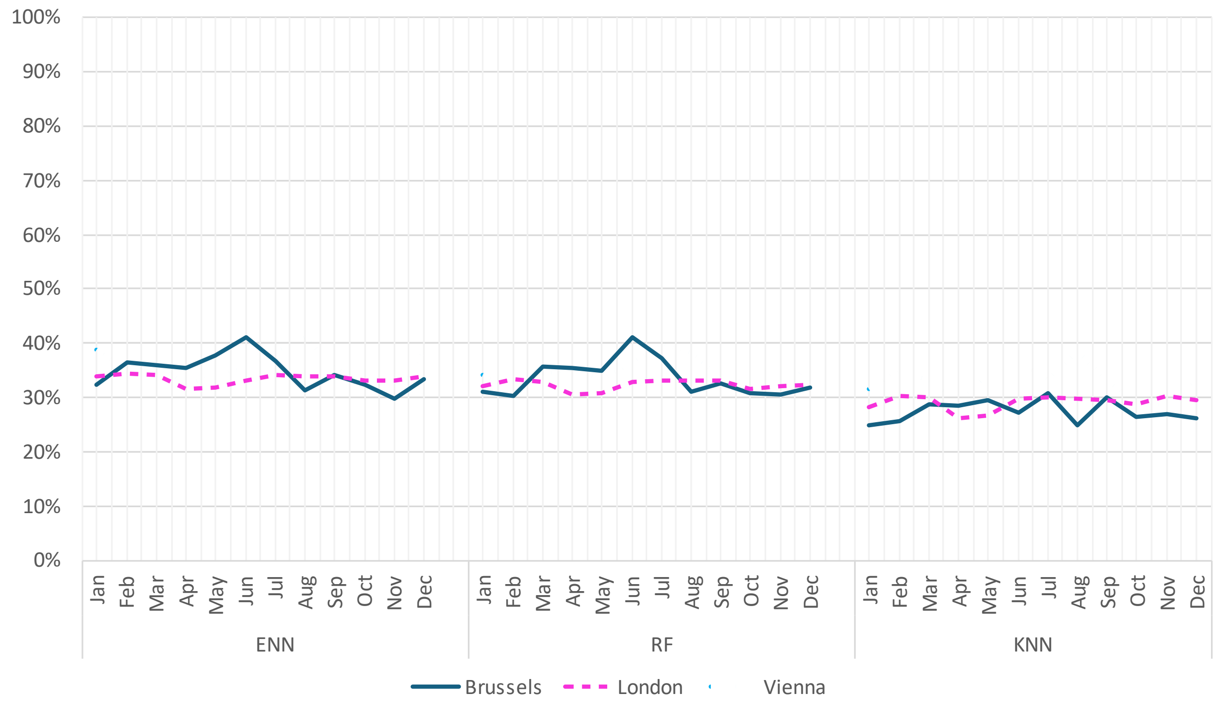


Fig. 5 F1-scores across machine learning methods and cities.

5. Discussion and conclusion

The present study uncovers important insights into the applicability and robustness of machine learning methods for predicting spatiotemporal crime patterns across three different European cities. It should be acknowledged already at this point, however, that while our findings offer significant contributions, their generalizability is limited solely to the specific settings, methodologies, and crime type (residential burglary) investigated throughout. Nevertheless, the crime prediction framework delineated within this study holds promise for informing forthcoming research endeavors and practical applications focused on deploying tailored machine learning techniques or algorithms to address spatiotemporal crime prediction challenges. Consequently, the findings of this investigation stand to provide valuable guidance to scholars and practitioners focusing on analogous urban contexts, where the prediction of spatiotemporal crime occurrences remains a pertinent concern.

Within each city, the ENN algorithm exhibited higher overall performance when predicting a lower number of burglary hot spots. Notably, when focusing on the top 5% of burglary hot spots, the ENN algorithm slightly outperformed the RF algorithm and significantly outperformed the KNN algorithm, particularly in terms of modeling precision. However, it is important to highlight that both the ENN and RF algorithms yielded highly comparable performance measures, indicating their enhanced capacity to accurately predict grid cells associated with monthly residential burglaries. An intriguing observation is the convergence of the three algorithms' modeling performance as more burglary hot spots are predicted, particularly for London and Vienna. This suggests that when generating a higher number of predictions, the algorithms exhibit similar abilities to identify burglary hot spots. This implies that the choice of a specific algorithm may become less critical when attempting to capture a broader range of potential burglary hot spots, especially on the long-term, e.g., within the context of defining strategic objectives. However, although predicting a larger number of potential hot spots will eventually increase the number of correctly predicted hot spots, it will nevertheless also increase the number of false positives. This trade-off underscores the significance of resource allocation. From a strategic standpoint, predicting a larger number of potential burglary hot spots may compromise the cost-effectiveness of deploying police resources on the basis of predicted spatiotemporal crime patterns: although a broader coverage of potential burglary areas can be achieved, it requires the allocation of police resources to a larger number of predicted hot spots, which could potentially result in a more dispersed resource allocation pattern. In this regard, we should also ask ourselves whether it is even meaningful/realistic to predict for example 50% of the areas. As we are aware of the fact that crime is highly concentrated, wherein a minority of micro-places, typically comprising around 4 to 5%, contribute to approximately 50% of criminal incidents, this would be a waste of time and efforts of police forces.

An important question that arises is why the KNN algorithm performs worse in each city compared to the ENN and RF algorithms. One plausible explanation may lie in the inherent differences between the algorithms in terms of their modeling logic. As the modeling of spatiotemporal crime patterns often involves capturing nonlinear relationships, it is possible that the KNN algorithm is less adept at learning such discriminative patterns in the data, leading to less accurate predictions on unseen test data, especially when a smaller number of predictions is required. KNNs operate based on the assumption of local similarity, meaning that they predict class probabilities of data points based on the classes of their nearest neighbors. This assumption works best when data points with similar characteristics are located in close proximity to each other in a feature place. However, when it comes to predicting complex, often nonlinear, spatiotemporal crime patterns, this assumption may not hold true due to the influence of important spatial and temporal factors that may extend beyond ‘local neighborhoods’ and thus similar data points. Another notable factor is that both the ENN and RF algorithms are ensemble classifiers, which often yield improved prediction performance compared to individual classifiers like KNN. By harnessing the collective predictive power of multiple individual models, both the ENN and RF algorithms could have captured a broader range of patterns, relationships, and interactions within the training data compared to the individual KNN model. After all, the KNN algorithm is often depicted as a ‘lazy learner’, implying that the algorithm postpones the processing of the training data until the prediction phase, using the entire dataset to make predictions by searching for the nearest neighboring data points. Consequently, the KNN algorithm may struggle to generalize well to unseen data and capture the complex relationships present when modeling spatiotemporal crime patterns (Kigerl et al. 2022).

Moreover, our study compared the prediction performance of the derived prediction models across the three cities. Intriguingly, both the ENN and RF algorithms exhibit markedly enhanced performance for Brussels during the entire summer period, implying their capacity to predict crime patterns more accurately within this temporal context. Conversely, for London, no substantial increase in prediction performance is observed throughout the summer period and more stable patterns are observed. Important to note is that in terms of modeling precision, both the ENN and the RF algorithm performed substantially better for Brussels, introducing a larger number of false positives for London and Vienna. In case of the KNN algorithm, however, lower performance is observed for Brussels, but no substantial differences are observed for London and Vienna, suggesting that the three algorithms perform relatively well on these datasets. Despite the lower precision, the three algorithms tend to perform very consistent for London, and no real fluctuations are observed across all performance measures compared to Brussels and Vienna, for which performance measures were more fluctuating over time. The consistency in the performance measures for London could indicate that the underlying crime patterns and distributions are relatively stable and somehow more predictable in London and that the three algorithms were able to successfully capture these underlying patterns when modeling the data. The latter is somehow reflected by the Gini coefficients in Table 2, suggesting that residential burglaries are quite concentrated in London between 2014-2016, showing less fluctuations and thus more stable yearly crime concentrations. The same could be true for Vienna and Brussels, as it is apparent that higher direct hit rates are achieved for Vienna, while lower direct hit rates were achieved for Brussels, which may be explained by the higher and lower clustering of residential burglaries at grid cells in Vienna and Brussels, respectively. However, even though differences in burglary concentrations provide some context for the observed modeling performance across the three cities, it is evident that additional structural factors may have also influenced the predictive modeling performance across these cities. As previously mentioned, for example, areas with higher population densities tend to experience more crime (so have higher crime counts), resulting in a larger number of positive instances (crime occurrences) and a wider distribution of features within the datasets. This provides increased opportunities for the algorithms to learn discriminative patterns and establish more accurate models. In this regards, it is important to acknowledge that there might be significant differences between the three cities in terms of how prediction performance is influenced by the features that are included in the predictive modeling. While the present study does not assess the impact of these features on modeling performance, Kadar et al. (2019) demonstrate that the importance of crime history features tends to increase in areas with higher population densities.

However, it is crucial to note that cities with lower population densities and crime counts are not necessarily unable to benefit from predictive crime models. If the frequency of a specific crime type is too low, alternative approaches can be employed to compensate for this limitation. For instance, the data can be aggregated to a higher spatial or temporal level, or related crime types can be grouped together into a larger category (Hardyns and Khalfa 2023). This increased scale can not only enhance prediction performance but may also facilitate the detection of cross-border crimes that transcend the boundaries of different police departments. Nevertheless, despite considering these factors, it should be acknowledged that scale and frequency alone do not provide a comprehensive understanding of the situation. For example, in the case of the present study, important differences arise regarding the level of aggregation for the collected crime-supporting data among the cities, which could have introduced an additional layer of complexity when modeling the data. Specifically, data for Vienna were only available at the higher aggregated level of political districts, in comparison to the neighborhood-level data collected for Brussels and London. This discrepancy in data quality and granularity could have affected the ability of the algorithms to effectively capture fine-grained variations and nuances in the data, due to a potential loss of localized information and lower variability when disaggregating the data to a lower spatial scale. Evidently, other factors that are associated with crime type and/or a combination of contextual factors such as variations in socioeconomic (e.g., income level, unemployment level etc.) and demographic characteristics (e.g., population composition) or cultural (e.g., community dynamics) and environmental characteristics (e.g., the presence of crime opportunity factors such as lower street connectivity) could also have contributed to differences in modeling performance.

5.1.1. Implications of this study for law enforcement and academia

The present study has implications for police decision-making processes within the context of intelligence-led and big data policing and contributes to the existing scientific knowledge regarding the reliability of spatiotemporal crime prediction methods when applied across various European cities. The findings may prove especially valuable on a European and international scale, providing insights for policymakers and law enforcement agencies worldwide seeking to adopt (big) data-driven approaches to crime prevention and resource allocation. The results will also contribute to the further development of predictive models and tools that will become increasingly interesting for law enforcement agencies in the future. Previous research (Hardyns and Khalfa 2023) has already shown that these models need to strongly consider different degrees of urbanization within the same country, hence the importance of international comparative research.

Our findings indicate that when law enforcement agencies seek to deploy police resources based on spatiotemporal crime predictions, they should be cognizant of the enhanced performance exhibited by ensemble machine learning models compared to individual models like KNN. Although we did not specifically test other individual models in relation to the ensemble methods employed, namely ENN and RF, our results clearly demonstrate that relying solely on individual machine learning methods may yield suboptimal outcomes in terms of prediction performance. By employing ensemble methods such as ENN and RF, law enforcement agencies can optimize resource allocation to address residential burglary hot spots, as these methods exhibit lower rates of false negatives and false positives. Consequently, this enables a more targeted approach, enhancing the probability of deterring potential offenders or detecting criminal activity more effectively.

Furthermore, although some differences in prediction performance are observed across cities, mainly depending on the prediction performance measure that is assessed, the differences are not that substantial. This suggests that the predictive modeling approach used in this study holds premise for cross-city and cross-country application. For law enforcement, this implies the feasibility of employing such a standardized predictive modeling approach across cities and even national borders, allowing the establishment of standardized resource allocation strategies, crime prevention efforts and collaboration. Moreover, the observation that the predictive modeling approach somehow maintains its efficacy across various cities presents a promising avenue for comparative research, introducing a transferable framework that may potentially contribute to a broader understanding of the applicability of machine learning approaches to predict crime spatiotemporally across cities and countries. However, while the core methodology may remain robust and transferable, it is still crucial to recognize that adopting context-specific and tailored approaches when applying machine learning methods across cities can be imperative. For example, areas characterized by lower population densities or lower crime concentrations and rates may potentially benefit from the inclusion of different or supplementary features in predictive models that are contextually relevant (Kadar et al. 2019). Conversely, for areas with higher population densities or elevated crime concentrations or rates, a model encompassing fewer features may suffice. For instance, such areas may derive greater benefit from incorporating solely crime history features into the predictive models resulting in a less intricate model that excludes sensitive features from a human rights perspective (e.g., percentage of immigrants). It is important to note, however, that while existing literature indicates that certain area characteristics explain a significant proportion of the variance in crime outcomes, the utilization of these features may surpass the scope of a crime prediction task.

Furthermore, with regard to the allocation of police resources to crime hot spots, it is our firm belief that law enforcement agencies should consider deploying police resources to a smaller number of predicted hot spots, based on their capacities, priorities and actual crime trends and patterns. While this may result in a higher number of false negatives, it enhances the precision of the predictive models, leading to a decrease in false positives. However, this approach lightens the burden on police officers while enabling a greater concentration of resources in the predicted hot spots, thereby maximizing efforts to reduce crime in those areas. To assess the cost-effectiveness of such a strategy, an intriguing tactic to explore is the deployment of police resources not only to predicted crime hot spots but also to the micro-places, such as grid cells, adjacent to those hot spots. The findings of this study indicate that while the direct hit rates substantially decrease when predicting a smaller number of hot spots, the near hit rates consistently remain higher. This suggests that incorporating micro-places adjacent to predicted crime hot spots in the calculation of hit rates results in fewer false negatives. Consequently, adopting a more focused approach by deploying police resources to fewer crime hot spots while maintaining a presence at the micro-places adjacent to the predicted hot spots could potentially enhance the effectiveness of crime prevention and detection. Although this approach entails allocating resources to additional micro-places, which may again reduce precision, it is possible that the time required for officers to travel from a predicted hot spot to its adjacent grid cells could be lower compared to the time needed to travel between predicted hot spots solely. Consequently, it is worth considering a comparative analysis of the (cost-)effectiveness achieved by deploying police resources to both predicted crime hot spots and their adjacent grid cells, in contrast to deploying resources to an increased number of predicted hot spots without considering adjacent grid cells. It is important to acknowledge, however, that the effectiveness of such an approach may depend on the concentration of crime. Furthermore, the evaluation of this approach would necessitate data collected within the framework of a rigorous evaluation study.

5.1.2. Limitations and avenues for future research

Although the present study contributes to the current scientific literature on the use of machine learning methods for predicting crime in place and time, it is important to acknowledge several limitations that may provide potential avenues for future work.

First, the focus of the present study is solely on residential burglary, excluding other crime types from the predictive modeling. While spatiotemporal prediction models are commonly applied to property crimes like residential burglary or car theft, previous research suggests that better and more stable prediction performance is often achieved when focusing on violent crime types, such as aggravated assault, or combination crime types, like aggressive theft (e.g., Hardyns and Khalfa 2023). Future research should explore the application of predictive models to a broader range of crime types to assess their performance and validity across different criminal activities.

Second, it is important to acknowledge that the number of features used in this study to predict crime in place and time is somewhat limited compared to previous studies within this area of research. This limitation mainly arises from the need to employ comparable features across the cities for a meaningful comparative analysis and the differences in data availability across the three cities. In this regard, it should be noted that is beyond the scope of the present study to estimate the relative importance of the features included in the machine learning models. Therefore, future studies should explore the question to what extent specific features add to the performance of machine learning models in predicting crime spatiotemporally across diverse settings and contexts. While the selected features were carefully chosen to capture relevant aspects related to crime patterns and contextual factors from a theoretical point of view, and although considerable levels of prediction performance were achieved for each city considering the limited number of features, the inclusion of additional or more granular features could provide further insights and potentially enhance the performance of the predictive models. For example, including additional features grounded in crime opportunity theories could have further improved the capabilities of the prediction models to account for important factors known to shape crime in place and time. With regard to the latter, the use of alternative types of data such as big data derived from electronic devices such as mobile phones, Automatic Number Plate Recognition systems, GPS devices from police patrol units, satellites etc. may prove valuable (Snaphaan and Hardyns, 2021). Drawing from Routine Activity Theory (RAT) (Cohen and Felson 1979), for instance, big data sources may enhance the measurement of the distribution of suitable targets, potential offenders and capable (formal and informal) guardians in particular places and during specific moments in time. In this regard, future studies could investigate using GPS data from police patrol units to generate measures of police presence (e.g., the frequency, duration, and intermittency of police presence) at microgeographic units as proxies of formal guardianship. In addition, Rummens and colleagues (2021) have shown that mobile phone data provide more accurate reflections of population mobility in terms of spatiotemporal resolution, enabling the extraction of specific ambient population features (e.g., the number of people that are 'most likely to live in a specific area'). Integrating diverse big data sources related to criminological concepts and crime characteristics could therefore facilitate a more nuanced and fine-grained analysis of spatiotemporal crime patterns, both from a theoretical and methodological point of view. As place-based predictive policing applications are mainly revolved around predicting short-term crime hot spots in or near real time, incorporating features at fine-grained spatiotemporal levels may prove valuable and increase variability, instead of only using static data sources that limit predictive power at such fine-grained spatiotemporal levels.

Third, only monthly predictions were made. In this regard, the present study did not explore to predict crime hot spots at fine-grained temporal resolutions such as bi-weekly, weekly, or daily resolutions. Future research should investigate the value and costs associated with predictions at different temporal (and spatial) resolutions, considering the trade-off between prediction performance, data quality, and costs. It is important to note that when predicting crime at higher temporal resolutions, such as daily or weekly predictions, the incorporation of real-time or near-real-time information may yield greater benefits compared to lower temporal resolutions.

Fourth, in interpreting the results of the present study, researchers and policymakers should keep in mind the so-called 'No Free Lunch Theorem' (Wolpert 1996). This theorem suggests that algorithms may perform differently across specific prediction (crime) problems (Kigerl et al. 2022). Although the present study focused on three specific algorithms commonly used to predict crime in place and time, it is possible that these algorithms may perform differently or be outperformed by other algorithms when applied to other crime types or crime-related problems. Therefore, future work should consider the comparative performance of various algorithms to identify the most effective approaches for different crime prediction tasks. In this regard, it could also be interesting to explore whether combining the outputs of different algorithms, such as the ones used in the present study, into a 'hyper ensemble model', yields a better prediction performance (Kadar et al. 2019). This would provide more insights into the complementary strengths of individual algorithms and their collective predictive power. Furthermore, it is important to acknowledge that in this study, specific methodological decisions were made with respect to e.g., data aggregation, spatiotemporal resolution of predictions, and the amount and type of data collected. While many of these decisions were informed by prior research, there remains a need for further investigation into how different configurations of these choices may impact the performance of spatiotemporal crime models. Despite this, it is worth emphasizing that extensive efforts were undertaken throughout this study to ensure a comprehensive and comparative perspective by maintaining consistency in the modeling of crime predictions through these methodological choices.

While the predictive modeling approach used in this study has potential for further optimization, its value lies in its practical implementation by law enforcement agencies. Merely deploying officers to predicted crime hot spots may be considered satisfactory, but further investigation is needed to determine the optimal conditions for police presence when allocating resources to daily, weekly, or monthly predicted crime hot spots. Hence, future research should also address strategic and tactical questions concerning what the police should do at crime hot spots and how to evaluate these approaches within the context of proactive and dynamic intelligence, while equally taking into ethical and legal risks associated with the development and implementation of crime prediction models within police organizations. In this regard, the comparative nature of this study may serve as a model for future international research endeavors seeking to elucidate the effectiveness of predictive modeling approaches across diverse urban contexts. In this regard, both scholars and practitioners may think of further comparing different crime prediction methods across different settings (e.g., cross-country, cross-region) or across different contexts within a setting (e.g., applying the crime prediction framework for individual police departments within a setting). The collaborative efforts between researchers and law enforcement agencies in implementing and refining predictive modeling strategies can contribute to the ongoing evolution of crime prevention practices, ultimately enhancing public safety and community well-being.

Acknowledgements

This work is funded/co-funded by the European Union (ERC-2022-COG, BIGDATPOL, 101088156 (project number)). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

The authors would like to express their gratitude to the Federal police department of Brussels, the Federal Criminal Police Office of Vienna and the Metropolitan Police of London, for providing the residential burglary data. Additionally, the authors would like to thank the editor(s) and the anonymous reviewers for their insightful comments and suggestions.

References

Allam Z, Tegally H, Thondoo M (2019) Redefining the use of big data in urban health for increased liveability in smart cities. Smart Cities 2(2): 259-268.

Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S (2012) The 'K' in K-fold Cross Validation. In: *ESANN* Vol. 102, pp 441-446.

Berk R (2013) Algorithmic criminology. Secur Inform 2:1-14.

Berk RA, Bleich J (2013) Statistical procedures for forecasting criminal behavior: A comparative assessment. Criminol Public Pol 12:513.

Bernasco W, & Steenbeek W (2017) More places than crimes: Implications for evaluating the law of crime concentration at place. J Quant Criminol, 33:451-467.

Bishop CM (1995) Neural networks for pattern recognition. Oxford university press.

Brantingham P, & Brantingham P (1995) Criminality of place: Crime generators and crime attractors. Eur J Crim Policy Re 3:5-26.

Brantingham PJ, Brantingham PL, Song J, Spicer V (2020) Crime Hot Spots, Crime Corridors and the Journey to Crime: An Expanded Theoretical Model of the Generation of Crime Concentrations. In: Lersch K, Chakraborty J (eds), Geographies of behavioural health, crime, and disorder: The intersection of social problems and place. Springer, Cham, Switzerland, pp. 61-86.

Caplan JM, Kennedy LW (2016) Risk terrain modeling: Crime prediction and risk reduction. University of California Press.

Chainey S, Tompson L, Uhlig S (2008) The utility of hotspot mapping for predicting spatial patterns of crime. Secur J 21:4-28.

Cohen LE, Felson M (1979) Social change and crime rate trends: A routine activity approach. Am Sociol Rev, 44(4):588-608

Dau PM, Dewinter M, Witlox F, Van Der Beken T, Vandeviver C (2022) How concentrated are police on crime? A spatiotemporal analysis of the concentration of police presence and crime. Camb J Evid.-Based Polic 6(3-4):109-133.

Gerstner D (2018) Predictive Policing in the Context of Residential Burglary: An Empirical Illustration on the Basis of a Pilot Project in Baden-Württemberg, Germany. Eur J Int Secur 3:115-38

Glasner P, Johnson SD, Leitner M (2018) A comparative analysis to forecast apartment burglaries in Vienna, Austria, based on repeat and near repeat victimization. Crime Sci 7:1-13.

Haberman CP (2017) Overlapping hot spots? Examination of the spatial heterogeneity of hot spots of different crime types. Criminol Public Pol 16(2):633-660.

Hardyns W, Khalfa R (2023) Predicting crime across cities and regions: a comparative analysis of predictive modeling in three Belgian settings. Appl Spat Anal Polic 16(1):485-508.

Hardyns W, Rummens A (2018) Predictive policing as a new tool for law enforcement? Recent developments and challenges. Eur J Crim Policy Re 24:201-218.

Hardyns W, Snaphaan T, Pauwels LJR (2019) Crime concentrations and micro places: An empirical test of the "law of crime concentration at places" in Belgium. Aust N Z J Criminol 52(3):390-410. https://doi.org/10.1177/0004865818807243

Hardyns W, Vyncke V, Pauwels LJR, Willems S (2015) Study protocol: SWING–social capital and well-being in neighborhoods in Ghent. Int J Equity Health 14:1-10.

Hart TC (2020) Hot Spots of Crime: Methods and Predictive Analytics. In: Lersch K, Chakraborty J (eds), Geographies of behavioural health, crime, and disorder: The intersection of social problems and place. Springer, Cham, Switzerland, pp. 87-103.

Jenga K, Catal C, Kar G (2023) Machine learning in crime prediction. J. Ambient Intell Humaniz Comput 14(3):2887-2913.

Johnson SD, Bowers KJ (2008) Stable and fluid hot spots of crime: Differentiation and identification. Built Environ 34:32–45.

Jones RW, Pridemore WA (2019) Toward an integrated multilevel theory of crime at place: Routine activities, social disorganization, and the law of crime concentration. J Quant Criminol 35:543-572.

Kadar C, Maculan R, Feuerriegel S (2019) Public decision support for low population density areas: An imbalance-aware hyper-ensemble for spatio-temporal crime prediction. Decis Support Syst 119:107-117.

Kigerl A, Hamilton Z, Kowalski M, Mei X (2022) The great methods bake-off: Comparing performance of machine learning algorithms. J Crim Justice 82:101946.

Kounadi O, Ristea A, Araujo A, Leitner M (2020) A systematic review on spatial crime forecasting. Crime Sci 9:1-22.

Kuhn M, Johnson K (2013) *Applied predictive modeling*. Springer, New York.

Levine N (2008) The "Hottest" part of a hotspot: comments on "The utility of hotspot mapping for predicting spatial patterns of crime". Secur J 21(4):295-302.

Mali B, Bronkhorst-Giesen C, den Hengst M (2017) Predictive policing: lessen voor de toekomst. Politieacademie Apeldoorn.

Mohler G, Porter MD (2018) Rotational grid, PAI-maximizing crime forecasts. Stat Anal Data Min 11(5):227-236.

Ratclife JH (2014) What is the future… of predictive policing. Pract 6(2):151–166

Ratcliffe JH (2016) Intelligence-led policing. Routledge, London.

Ripley BD (1996) Pattern recognition and neural networks. Cambridge university press.

Rosser G, Davies T, Bowers KJ, Johnson SD, Cheng T (2017) Predictive crime mapping: Arbitrary grids or street networks?. J Quant Criminol 33:569-594.

Rummens A, Hardyns, W (2020) Comparison of near-repeat, machine learning and risk terrain modeling for making spatiotemporal predictions of crime. Appl Spat Anal Polic 13(4):1035-1053.

Rummens A, Hardyns W (2021) The effect of spatiotemporal resolution on predictive policing model performance. Int J Forecast 37(1):125-133.

Rummens A, Hardyns W, Pauwels L (2017) The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. Appl Geogr 86:255-261.

Rummens A, Snaphaan T, Van de Weghe N, Van den Poel D, Pauwels LJR, Hardyns W (2021) Do mobile phone data provide a better denominator in crime rates and improve spatiotemporal predictions of crime?. ISPRS Int J Geo-Inf 10(6):369.

Sherman LW, Gartin PR, & Buerger ME (1989) Hot spots of predatory crime: Routine activities and the criminology of place. Criminol 27(1):27-56.

Snaphaan T, Hardyns W (2021) Environmental criminology in the big data era. Eur J Criminol 18(5):713-734.

Tompson L, Coupe T (2018) Time and opportunity. In: Bruinsma GJN, Johnson SD (eds) The oxford handbook of environmental criminology. Oxford University Press, pp 716–731.

Weisburd D (2015) The law of crime concentration and the criminology of place. Criminol 53(2):133-157.

Wheeler AP, Steenbeek W (2021) Mapping the risk terrain for crime using machine learning. J Quant Criminol 37:445-480.

Wolpert DH (1996) The existence of a priori distinctions between learning algorithms. Neural Comput 8(7):1391-1420.

Zhang X, Liu L, Xiao L, Ji J (2020) Comparison of machine learning algorithms for predicting crime hot spots. IEEE Access 8:181302-181310.

Zhang D, Pee LG, Pan SL, Cui L (2022) Big data analytics, resource orchestration, and digital sustainability: A case study of smart city development. Gov Inf Q 39(1):101626.

Appendices

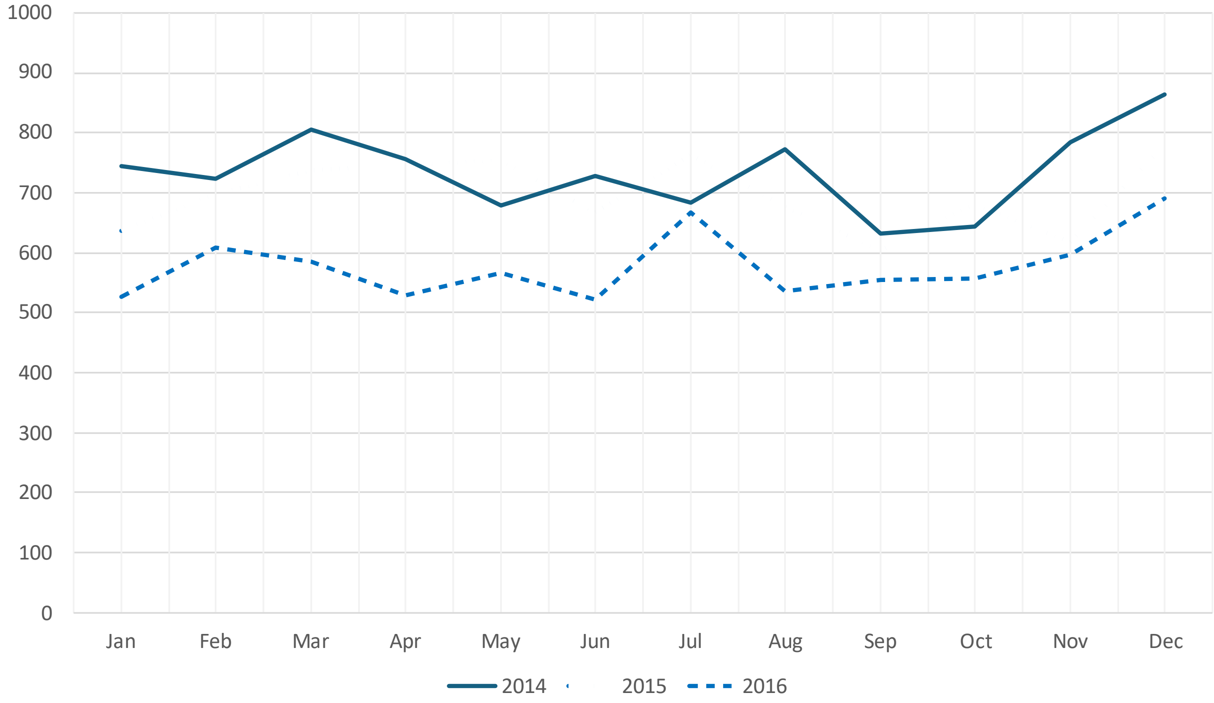
Appendix 1: Overview of the included features

(1) Crime history features
Number of crime events in the previous month
Time since last crime event (in months)
Number of crime events in the past year
Number of crimes in the previous period in the neighbourhood*
Number of crimes in the same month last year
(2) Demographic features
Total residential population*
Percentage youth*
Percentage of non-domestic inhabitants*
Percentage single households*
(3) Socio-economic features
Unemployment rate*
Percentage of houses occupied by homeowners*
Dwelling stock*
(4) Environmental features
Number of shops
Number of bars/cafés
Number of restaurants
Number of snack bars
Presence of green space (Boolean)
(5) Proximity features
Distance to nearest station (m)
Distance to nearest highway (m)
Distance to nearest bus/metro/tram stop (m)
(6) Additional feature(s)
Seasonal indicator (winter, summer)

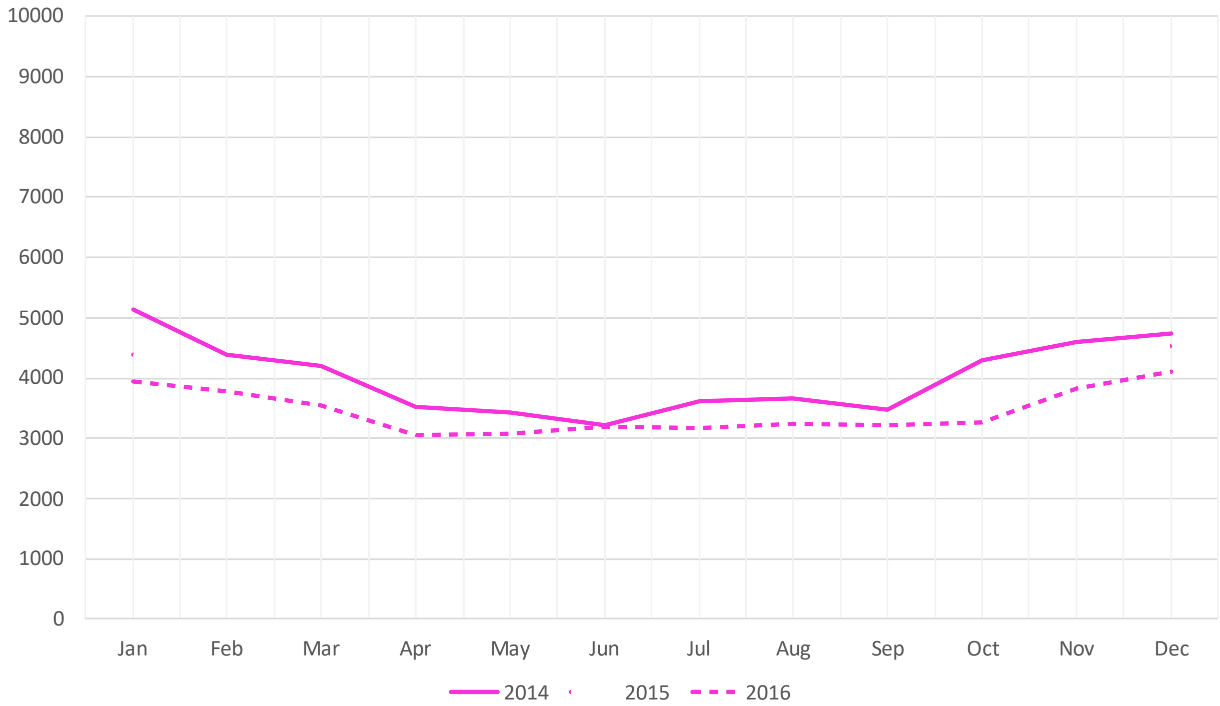
*Only available at the neighbourhood level for Brussels and London and at the district level for Vienna

Appendix 2: Monthly number of residential burglaries in each city (2014-2016)

Appendix 2.1. Monthly number of residential burglaries in Brussels (2014-2016).



Appendix 2.2. Monthly number of residential burglaries in London (2014-2016).



Appendix 2.3. Monthly number of residential burglaries in Vienna (2014-2016).