ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Vomfell, Lara; Härdle, Wolfgang Karl; Lessmann, Stefan

Working Paper

Improving Crime Count Forecasts Using Twitter and Taxi Data

IRTG 1792 Discussion Paper, No. 2018-013

Provided in Cooperation with:
Humboldt University Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series"

Suggested Citation: Vomfell, Lara; Härdle, Wolfgang Karl; Lessmann, Stefan (2018) : Improving Crime Count Forecasts Using Twitter and Taxi Data, IRTG 1792 Discussion Paper, No. 2018-013, Humboldt-Universität zu Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series", Berlin

This Version is available at:
https://hdl.handle.net/10419/230724

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

IRTG 1792 Discussion Paper 2018-013

Improving Crime Count Forecasts Using Twitter and Taxi Data

Lara Vomfell *
Wolfgang Karl Härdle *
Stefan Lessmann *

* Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

International Research Training Group 1792

Improving Crime Count Forecasts Using Twitter and Taxi Data

Lara Vomfell[a,*], Wolfgang Karl Härdle[a,b], Stefan Lessmann[a]

[a] *Faculty of Business and Economics, Humboldt University of Berlin, Unter-den-Linden 6, 10099 Berlin, Germany*
[b] *Singapore Management University, 50 Stamford Road, Singapore 178899*

Abstract

Data from social media has created opportunities to understand how and why people move through their urban environment and how this relates to criminal activity. To aid resource allocation decisions in the scope of predictive policing, the paper proposes an approach to predict weekly crime counts. The novel approach captures spatial dependency of criminal activity through approximating human dynamics. It integrates point of interest data in the form of Foursquare venues with Twitter activity and taxi trip data, and introduces a set of approaches to create features from these data sources. Empirical results demonstrate the explanatory and predictive power of the novel features. Analysis of a six-month period of real-world crime data for the city of New York evidences that both temporal and static features are necessary to effectively account for human dynamics and predict crime counts accurately. Furthermore, results provide new evidence into the underlying mechanisms of crime and give implications for crime analysis and intervention.

Keywords: Predictive Policing, Crime Forecasting, Social Media Data, Spatial Econometrics

*Corresponding author
Email addresses: l.vomfell@warwick.ac.uk (Lara Vomfell), haerdle@hu-berlin.de (Wolfgang Karl Härdle), stefan.lessmann@hu-berlin.de (Stefan Lessmann)

1. Introduction

Crime research has investigated how social structures in and between neighbourhoods influence criminal activity (Sampson et al., 1997; Mears & Bhati, 2006). Crime is facilitated or deterred not only by communal structures but also by human dynamics. Every day, people leave their neighbourhood to commute to work, shop in malls, or relax in museums or bars. Such travels create a social flow of both crime targets and perpetrators that connect areas beyond spatial distance, which influences criminal activity (Wikström et al., 2010). Exploitation of location-based data offers new perspectives on the mechanisms that influence different types of crime and can help predict their occurrence. Governmental institutions and especially police depend on adaptive, short-term crime predictions to anticipate changes and breaks in crime patterns and allocate scarce resources efficiently (e.g., Xue & Brown, 2006).

The objective of the paper is to establish the relationship between human dynamics and crime rates. In pursuing this goal, the paper proposes predictive models that extend conventional crime forecasts by incorporating spatial dependency using three sources of data: public venues, social media activity, and taxi flows. Joint consideration of these sources provides insights into the marginal relevance of different predictors and accounts for possible interaction effects. The paper suggests alternative ways to extract features from the data sources and examines the explanatory and predictive value of the features in an empirical study related to crime incidences in New York city using several statistical and machine learning models.

Empirical results confirm the relevance of the proposed features. Their inclusion in a forecasting model improves neighbourhood-level crime predictions for different types of crime significantly. The results also indicate interaction effects. Features from different data sources work best when used in combination. These findings add to a better understanding of the link between crime opportunities and incidents as well as the moderators of this link.

The paper is organised as follows: Section 2 and Section 3 discuss related

work and the spatial and non-spatial prediction models employed therein, respectively. Section 4 outlines the data sources and feature construction methods. Empirical results are presented in Section 5 and discussed in Section 6. Section 7 concludes the paper.

2. Related Work

Theories explaining the spatio-ecological dimension of crime include opportunity theory and social disorganisation theory. The former analyses crime events as the co-occurrence of an opportunity in form of a suitable target, a lack of supervision, and a motivated offender (Cohen & Felson, 1979). This view supports urban interventions to reduce opportunity such as blocking roads to cut off highway connections in neighbourhoods with high rates of drive-by shootings (Lasley, 1998). Social disorganisation theory considers neighbourhood characteristics that influence the likelihood of criminal activity among inhabitants. A lack of social control and social cohesion within a community combined with structural disadvantages give rise to criminal behaviour. Social disorganisation theory has been applied in a wide range of crime research such as domestic violence (Beyer et al., 2015).

Drawing on these theories, much research examines socio-economic predictors of criminal behaviour (on an individual level) and crime rates (on a neighbourhood level). Examples include residential stability, ethnic heterogeneity and population size and density (e.g. Land et al., 1990; Sampson et al., 1997; Kubrin, 2003). Spatio-temporal elements and aggregated human behavioural data have also been considered to capture geographic crime drivers and heterogeneity in crime rates between neighbourhoods. Brantingham & Brantingham (1991) observe that i) different crimes have different spatial distributions, ii) the spatial concentration of crime is influenced by location characteristics, and iii) spatial patterns are relatively stable over time.

More recent research studies how data on immediate human behaviour affects crime using geo-tagged social media data and data on human movements.

For example, prior work predicts crime from Twitter data using text mining techniques such as term frequency (Williams et al., 2017) or topic models (Gerber, 2014) for feature creation. Novel data sources and analysis methods are also studied in terrorism and organised crime detection as well as case association (e.g. Xu & Chen, 2004; Lin & Brown, 2006; Yang & Li, 2007).

Crime prediction has received less attention compared to the above explanatory studies. This causes a research gap in crime prediction regarding the synthesis of space, human dynamics, and effective modelling approaches. More specifically, studies that incorporate features related to human dynamics either disregard the spatial dependence between observation units or are purely explanatory and do not test the predictive ability of empirical models. However, predictive power is important if model-based forecasts inform decision-making, for example in the scope of predictive policing (Camacho-Collados & Liberatore, 2015). Surveying corresponding studies in the field, Table 1 illustrates the research gap, which we aim to overcome with this study.

Table 1 suggests that considering human activity patterns (i.e., dynamics) is popular in crime modelling. Traunmueller et al. (2014) examine correlations between people activity features, which they derive from mobile phone data, and monthly crime rates. A study by Williams et al. (2017) uses regression to examine the relationship between crime, Twitter activity, and Twitter mentions related to "broken window" theory (e.g., Hinkle & Yang, 2014). Bendler et al. (2014) use Twitter and local points of interest (POI) data to capture human activity and explore spatial dependence between crime locations. Their results indicate that only some crimes such as burglary benefit from the inclusion of Twitter activity. Interestingly, certain crime types such as theft are more likely to occur when many people are in the area whereas others (e.g., motor vehicle theft) occur when there is no activity around.

Wang et al. (2016) also consider POI data, which they integrate with taxi flow data to model yearly crime rates for a community in Chicago. They find a model using both types of information to outperform models using only POI or taxi data. Such synergy hints at an interdependence between the two sources,

Study	Prediction	Spatial	Dynamics	Machine Learning	Crime Type	City	Time Frame
Wang et al. (2016)		✓	✓		no differentiation between types	Chicago	yearly
Bogomolov et al. (2014)	✓		✓	✓	(hotspot classification)	London	monthly
Bendler et al. (2014)		✓	✓		assault, burglary, homicide, theft, vandalism, etc.	San Francisco	hourly
Traunmueller et al. (2014)			✓		street vs. indoor	London	monthly
Williams et al. (2017)			✓		burglary, theft, drug possession, criminal damage, violent crime	London	monthly
Xue & Brown (2006)	✓	✓		✓	burglary	Richmond, VA	monthly
This study	✓	✓	✓	✓	violent and property crime	New York	weekly

Table 1: Literature Overview

which has also been observed by Bendler et al. (2014).

Unlike the previous studies that concentrate on explaining criminal activity, only Bogomolov et al. (2014) and Xue & Brown (2006) explicitly consider crime prediction. Bogomolov et al. (2014) employ telecommunication records to devise features related to demographics and visitor number in an area. Using Random Forests, they classify areas in the city of London as high-/low-crime depending on whether the number of criminal incidents in the area is above/below the median. Xue & Brown (2006) model the spatial coordinates of crime as a locally optimal site picked by the offender from a set of spatial alternatives to commit the crime. Concentrating on burglary, they cluster locations on the basis of their characteristics (e.g., distance to highway) and predict both burglary incidents and hotspots in Richmond, VA.

This study contributes to prior literature through integrating data sources that proxy human activity patterns with geo-spatial data sources for crime prediction. We also consider a weekly forecasting horizon which complements previous findings obtained at a monthly/yearly or hourly level. A weekly horizon may be useful to inform predictive policing and to (re-)adjust tactical patrolling plans. With such applications in mind, we also consider a broad set of alternative prediction methods including techniques from spatial econometrics and machine learning. This allows us to shed light on the trade-off between model interpretability and forecasting accuracy in crime prediction.

3. Crime Modelling Methodology

Crime rates depend on the underlying population at risk, which need not correspond to the residential population in a geographic unit. Therefore, a common modelling approach, which we adopt in this study, is to model counts of crime incidents (Andresen, 2006; Malleson & Andresen, 2015). Our data forms a panel of crime counts and covariates per census tract for each week, indexed by i and t, respectively, over a period of six months.

Below, we describe models for this type of data including i) spatial linear

regression (Subsection 3.1); ii) Poisson Generalised Linear Model (GLM) and a GLM with spatial random effects (Subsection 3.2), and iii) machine learning methods (Subsection 3.4).

Before describing modelling approaches, we introduce some notation. Modelling crime counts in a city begins with the area: a specific, bounded two-dimensional area $D \subset \mathbb{R}^2$ where D denotes the surface area of the city. This fixed subset of irregular shape can be partitioned into a finite number of well-defined areal units, e.g. census tracts. More formally, let the simple partition $\{B_1, B_2, \ldots, B_N\}$ form the lattice of D such that $B_1 \cup B_2 \cup \ldots B_N = D$ and $B_i \cap B_j = \emptyset$ for $i \neq j$.

Let crime events be realisations of a point process that occurs at random locations in space. Let the realisations of this random spatial process be denoted by $\boldsymbol{S}$ with elements $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_k\}$. Given a point process, one can model the number of realised events in an areal unit as a count variable. Let this count variable be defined as $m(B) = \sum_{\boldsymbol{s}_i \in \boldsymbol{S}} 1(\boldsymbol{s}_i \in B)$ such that $m(B)$ gives the number of event points in set B. Let y denote the vector of count variables observed at the N areal units forming D such that $m(B_i) = y_i$ where $i = 1, \ldots, N$ indexes the areal unit. This notation is easily extended to a panel setting by specifying $m_t(B_i) = y_{it}$.

Spatial dependence between areas can take the form of a Markov random field, which defines a neighbourhood for each element in y. An areal unit j is a neighbour of areal unit i if the conditional distribution of y_i depends on y_j (Cressie, 1993). Let $N_i = \{j : j \text{ is a neighbour of } i\}$ be the neighbourhood of unit i. Note that N_i excludes unit i.

3.1. Linear Models

Consider the simple pooled linear panel regression model:

$$y = X\beta + e, \quad e \sim N(0, \sigma^2 I_{NT}), \tag{1}$$

where N denotes the number of spatial units and T the number of cross-sections. In the presence of spatial dependence, the error terms in (1) are no longer uncorrelated. Approaches to account for this include the simultaneous autoregressive (SAR) and the conditional autoregressive (CAR) model (Cressie, 1993).

The SAR model introduces spatial structure through a spatial lag:

$$y = (I_T \otimes \rho W)y + X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_{NT}), \tag{2}$$

where $\otimes$ denotes the Kronecker product, I_T denotes the identity matrix of order T, and W is a $N \times N$ binary matrix specifying which areas are spatially adjacent with $w_{ii} = 0$, and ρ is the parameters that specifies the magnitude of spatial dependence.

The inclusion of a spatial lag of the dependent variable accounts for spatial spillovers and a mismatch of the spatial scale with the spatial event (e.g., Cressie, 1993). Both effects occur in crime modelling since the contagion effect of crimes leads to a diffusion through space. In addition, economic and criminal features do not match perfectly with the spatial units. Therefore, a spatial lag SAR model is a convenient choice to account for these characteristics.

The CAR model introduces a spatial dependence parameter in the error term. Large-scale variation is captured in the regression parameters. Small-scale spatial variation is modelled in the error. This yields the following model:

$$\begin{aligned} y &= X\beta + \varepsilon, \\ \varepsilon &\sim N\left(0, \sigma^2\{I_T \otimes (I_N - \delta W)^{-1}\}\right), \end{aligned} \tag{3}$$

where W is again a $N \times N$ spatial adjacency matrix and δ denotes the magnitude of spatial dependence between neighbouring regions. The CAR model introduces spatial structure as a Markov random field such that the conditional distribution of each area depends on the neighbourhood. The distribution of y_{it}

conditional on all y_{jt} in N_i is given as

$$y_{it}|y_{jt} \sim N\left(X_{it}^{\top}\beta + \sum_{j} \delta w_{ij}(y_{jt} - X_{jt}^{\top}\beta), \sigma_i^2\right), \tag{4}$$

for $i \neq j$, where σ_i^2 denotes the conditional variance.

3.2. Count Models

Linear models offer a broad framework to include spatial structure but fail to accommodate the integer-valued and non-negative nature of crime counts. Small counts are better modelled by a Poisson regression. In the case of crime counts, the Poisson parameter λ represents the average incident count:

$$\lambda = E(y \,|\, X) = e^{X^{\top}\beta}. \tag{5}$$

The Poisson distribution assumes random variables to be independently identically distributed. Spatial dependence between crime counts violates this assumption. Poisson Generalised Linear Mixed Models (GLMM) account for spatial dependence by incorporating a random effect in the GLM predictor. GLMM model the expectation of Poisson distributed y as a linear combination of fixed effects X and random effects Z with a logarithmic link function (Agresti, 2007):

$$\log \lambda_{it} = X_{it}^{\top}\beta + Z_{it}\eta_i. \tag{6}$$

Here, the random effect $Z\eta$ is a spatial-specific random effect, which introduces the spatial structure of a Markov random field. Z is specified as an indicator matrix of the spatial units such that $Z\eta$ is a random intercept added to the conditional mean. More specifically, Z is a $NT \times N$ stacked matrix of matrix I_N stacked T times such that $Z = \textit{block diagonal}(I_T \otimes I_N)$. The distribution of the random vector η is assumed to be multivariate normal:

$$\eta \sim N(0, D), \quad D = \sigma^2 Q^{-1}. \tag{7}$$

Q is a symmetric spatial dependency matrix determined by the neighbourhood structure with entries:

$$Q_{ij} = \begin{cases} |N_i| & \text{if } i = j, \\ -1 & \text{if } j \in N_i \text{ and } i \neq j, \\ 0 & \text{if } j \notin N_i \text{ and } i \neq j, \end{cases} \tag{8}$$

where the $|N_i|$ entries on the diagonal denote the size of the neighbour set and neighbours are indicated by -1 (Leroux et al., 2000, p. 186). Unlike the non-spatial Poisson GLM, the variance of the model in (6) is not only unequal to the expectation, it accounts for both overdispersion of the variance and spatial dependence. The parameters in (6) and (7) are estimated using restricted maximum likelihood (REML) and Fisher Scoring (Kneib, 2003).

3.3. Predictors

The predictions for weekly crime counts are obtained by using the best linear unbiased predictor or its panel equivalent (Baltagi et al., 2011). Table 2 gives the predictors for the time period $T + S$ for the regression models. The SAR

Model	Predictor
LR	$\hat{y}_{T+S} = X_{T+S}\hat{\beta}$
SAR	$\hat{y}_{T+S} = (I_N - \rho W)^{-1} X_{T+S}\hat{\beta}$
	$+ (I_N - \rho W)^{-1}\hat{u}$
CAR	$\hat{y}_{i,T+S} = X_{i,T+S}^{\top}\hat{\beta} + \sum_j \delta w_{ij}$
	$\left(T^{-1}\sum_{t=1}^{T}(y_{jt} - X_{jt}^{\top}\hat{\beta})\right)$
GLM	$\hat{y}_{T+S} = \exp(X_{T+S}\hat{\beta})$
GLMM	$\hat{y}_{T+S} = \exp(X_{T+S}\hat{\beta} + Z_{T+S}\hat{\eta})$

Table 2: Predictors for the spatial linear regression models considered in the study.

predictor is obtained by spatially lagging the linear predictor and adding the spatially lagged error vector of the model. The CAR predictor is obtained by taking a time-averaged conditional expectation (Cressie et al., 1999). In

practice, the variance components and spatial parameters are unknown and replaced by their maximum likelihood estimates.

3.4. Machine Learning Approaches

Previous models make assumptions about the data-generating process and consider a linear additive relationship between crime counts and covariates. Machine learning techniques are more flexible and account for non-linearity in a data-driven manner (Kuzey et al., 2014). We concentrate on random forest (RF), gradient boosting machines (GBMs), and feed-forward artificial neural networks (ANNs), all of which have shown promising results in previous studies (e.g. Bhattacharyya et al., 2011; Delen, 2010).

RF develops an ensemble of size k through drawing k bootstrap samples from the training data. The base models in RF consist of individual decision trees, which are grown from the bootstrap samples. To increase randomness among the base models, RF determines the best split during tree growing among a randomly sampled subset of covariates (Breiman, 2001). The model prediction consists of the simple average calculated across the k base models.

GBMs embody the idea of additive modelling. The algorithm incrementally develops an ensemble through adding base models that are fitted to the residuals—more specifically the negative gradient of the loss function—of the current ensemble. GBM predictions are obtained by calculating a weighted average over base model forecasts, whereby the weights are determined during gradient descent (Friedman, 2002).

An ANN model consists of interconnected layers of processing units (neurons) with connection weights representing the model parameters. Estimating an ANN model involves minimising some loss function with respect to connection weights using gradient-based methods. ANNs calculate the output of a neuron as a non-linear transformation of the weighted sum over its input neurons. The transformations are called activation functions and allow an ANN to capture non-linear patterns in data (Kim & Kang, 2016).

Machine learning methods exhibit meta-parameters such as the number of

trees in RF. We consider candidate settings for each meta-parameter and determine the best setting (i.e., minimal MSE) on a validation sample. Interested readers find a comprehensive discussion of the above and other machine learning methods and practices in, e.g., Hastie et al. (2009).

4. Data Integration and Feature Construction

The paper models crime counts via features of census, POI data, spatial influence, taxi flow, and Twitter activity. The features come from different data sources with different time frames. The most recent complete overlap is June 1, 2015 to November 29, 2015. We chose this period for subsequent analysis and aggregate temporal data to weekly intervals, which begin uniformly on Monday. The final data set covers 26 weeks. We use the first 24 weeks of data for model estimation. Machine learning models require auxiliary data for meta-parameter tuning (e.g., Carneiro et al., 2017). For such models, we use the first 22 weeks for model training and weeks 23 and 24 for tuning. The last two weeks serve as out-of-sample prediction set. Such split-sample setup is common practice in comparisons of alternative forecasting methods (e.g., Sermpinis et al., 2012).

The spatial unit of analysis are census tracts as defined by the US Census Bureau. We integrate area-referenced census data with point-referenced data from other sources using geo-coordinate matching. For example, we use the coordinates of a tweet to identify the Census track in which it was posted.

The following subsections introduce the data sources. Developing features from these sources is non-trivial and leaves some degrees of freedom. Therefore, we elaborate on alternative options for feature engineering and how the final set of features has been selected.

4.1. Census

Demographic variables are taken from Summary File 1 of the 2010 census data (U.S. Census Bureau, 2017). Based on previous studies on violent and property crime (e.g., Wang et al., 2016), we select the following eight demographic census variables: the total population in the census tract, the median

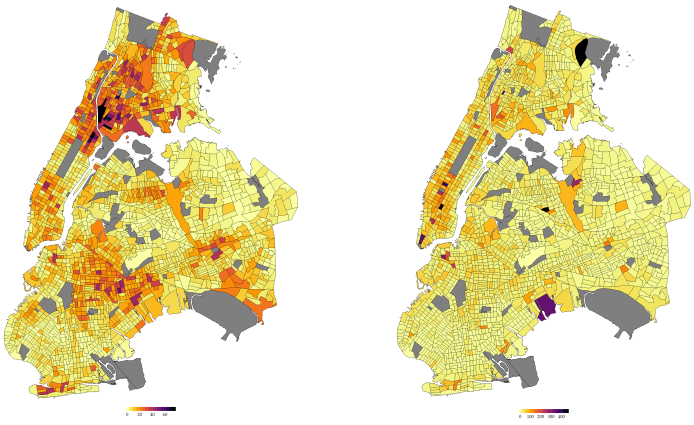
age of the population, the percentage of males, the percentages of the black, Asian, and Hispanic population, respectively, the percentage of female-headed family households, and the rate of vacant accommodation. The spatial structure in the CAR, SAR, and Poisson GLMM models is defined through the spatial neighbourhood matrix W (3). We define tracts as neighbours when they share a boundary. On average, a census tract is connected to six others.

We exclude census tracts with a residential population of less than 50 from the analysis because the estimates in these areas are not reliable. This concerns uninhabited areas such as cemeteries, but also Central Park and JFK Airport. We also exclude islands such as Staten Island and City Island because their spatial distance to the nearest large landmass exceeds 2 km. Regions separated by water but connected through a bridge shorter than 1 km are defined as neighbours as well. This is relevant when considering the connections between Manhattan and Brooklyn.

4.2. New York City Crime Data

Data on criminal incidents is provided by the New York City Police Department (New York City Police Department, 2016). Each report includes detailed information on crime date, type, and location. We focus on violent and property crime because they represent serious threats to public safety. In addition, the spatial distribution of crime incidents differs between these crimes, which facilitates examining the proposed features in a context of varying spatial dependence. Violent crime encompasses murder and non-negligent manslaughter, robbery, and aggravated assault (Federal Bureau of Investigation, 2014). Property crime comprises burglary, larceny-theft, motor vehicle theft, and arson. Using shapefiles of New York City by the Department of City Planning New York (2017), we map the coordinates of these crimes to census tracts. Since rape incidence are not geo-located in the NYPD dataset, we exclude this crime from the analysis.

Figures 1a and 1b show the spatial distribution of crime across census tracts for the analysis period of June to November 2015. Clearly, violent crime clusters



(a) Violent Crime (b) Property Crime

Figure 1: Number of crime incidents between June and November 2015. In the property crime map, the area around Penn Station (largest outlier with 2002 incidents) is excluded for more consistent colour scaling.

primarily in Harlem, the Bronx, and certain parts of Queens. Property crime shows a more even distribution. Notable exceptions include i) the downtown Manhattan area with a lot of transitory traffic and many tourists, ii) census tracts with shopping centres such as the Gateway Center or the Queens Center, which attract high numbers of shoplifting, and iii) cooperative housing project Co-op City in the Bronx with its own security force. The strength of spatial correlation between areas is tested using Moran's I. For both crime types and every time period, the null hypothesis of no spatial dependence is rejected with $p < .000$.

4.3. Foursquare

We gather POI data from Foursquare; a mobile app that recommends places based on a user's location and preferences. Through its API, Foursquare's location database facilitates extracting the coordinates of restaurants, schools, or museums. All venues are categorised along nine main dimensions: nightlife, food, arts & entertainment, residence, shops, travel, outdoors & recreation, college & education, and professional. We consider POI data a characterisation of the census tract. Different POI categories attract different groups of people at

different times of day. For example, one can expect that more nightlife venues attract drunken behaviour at night. In total, we obtain 47,113 POI in the geographic area of interest.

Three different ways of constructing the POI feature are considered: 1. The total counts of venues per category, 2. the share of categories on the total number of venues in the census tract, and 3. an entropy measure, typically used in ecology to quantify population diversity. Specifically, we use the Shannon index (Shannon, 1948), which is defined as

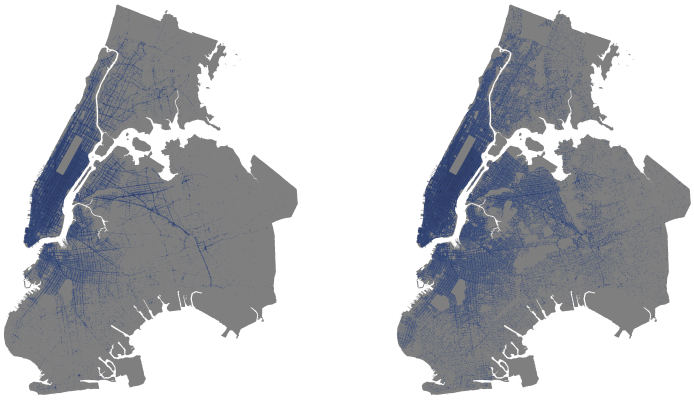
$$H = -\sum_{i=1}^{C} \frac{N_i}{N} \cdot \log\left(\frac{N_i}{N}\right) \tag{9}$$

where N_i denotes the total number of venues in venue category i, $N = \sum_i N_i$, and C is the total number of categories. H can be considered an extension of the second option to construct the POI feature in that the fraction $\frac{N_i}{N}$ equals option 2.

4.4. Taxi

The NYC Taxi & Limousine Commission (2016) provides the taxi flow data. Their dataset covers over 1.3 billion individual trips from January 2009 to June 2016, including start and end point of the trip, how many passengers entered, how many minutes the trip took, how much it cost, and how it was paid. Figure 2 illustrates the coordinates of taxi pickups and dropoffs. Comparing Figures 2a and 2b, it is clear that most taxi trips begin in Manhattan and Brooklyn and end all over the city. We argue that taxi flows provide a connection between different neighbourhoods beyond spatial proximity alone, which makes them a valuable source for crime modelling.

We consider all trips within New York City in the analysis time frame but exclude trips that start or end outside the analysis area. This gives 70,288,218 trips in the 26 weeks. We aggregate individuals trips to a weekly connection flow matrix F, with rows (columns) of F referring to the census tract where the trip started (ended). Hence, f_{ij} denotes the number of trips made from tract i



(a) Pickups (b) Dropoffs

Figure 2: Coordinates of complete taxi trips in New York City in week 46 in 2015.

to j for each time interval. In view of a substantial number of missing values in the TLC data on passenger counts, we use the number of trips in the analysis. The available records suggest an average of 1.6 passengers per trip.

The taxi flow feature is then constructed as $t = Fy$ such that neighbouring crime rates are weighted by the magnitude of flow F. It is crucial to note that the crime vector y is lagged by a week to prevent unintended implicit simultaneity of the response and its predictors. We refrain from incorporating a week index w as in $t_w = F_w y_{w-1}$ for ease of notation.

We propose three different ways to construct t and demonstrate the calculation for t_1, the feature of example region 1:

1. Raw multiplication: One can define t as the simple matrix multiplication of the flow matrix F and the crime count vector y:

$$t_1 = f_{11}y_1 + \ldots + f_{1n}y_n.$$

2. Normalised by source: The taxi flow arriving in each census tract is normalised by the total number of flows leaving the source census tract. For example, the flow leaving the second census tract towards the first tract

is normalised by all flows leaving from the second tract:

$$t_1 = \frac{f_{21}}{f_{21} + f_{23} + \ldots + f_{2n}} y_2 + \ldots + \frac{f_{n1}}{\sum_{i=1}^{n} f_{ni}} y_n.$$

3. Normalised by destination: The taxi flow arriving in each census tract is normalised by the total number of flows arriving in that census tract:

$$t_1 = \frac{f_{21}}{f_{21} + f_{31} + \ldots + f_{n1}} y_2 + \ldots + \frac{f_{n1}}{\sum_{i=1}^{n} f_{i1}} y_n.$$

Around 25% of all taxi trips end in a census tract that is not a neighbour of the tracts they started in. This means that the taxi feature captures connections between census tracts that go beyond spatial proximity.

4.5. Twitter

Downloading historic Twitter data is inherently limited as the Twitter GET search/tweets API provides access to tweets published in the last week only. However, given the unique tweet ID, the Twitter GET statuses/lookup API returns the full tweet from any point in time. The tweet IDs used here are sourced from Pfeffer & Morstatter (2016) who provide IDs to tweets published in the United States between June 1, 2015 to November 30, 2015.

After retrieving the data including the time stamp, text, language and coordinates, each tweet is mapped to the corresponding census tract. We aggregate the number of tweets per week and census tract, and implement six versions of the Twitter feature: 1. Using the full activity, 2. counting night-time tweets only, 3. using logged full activity, 4. using logged night-time activity. Any tweet sent out between 22pm and 6am contributed to the night-time feature. Taking the logarithm of the tweets served to reduce variation between census tracts.

4.6. Evaluation and Feature Selection

For each feature group, different variable definitions have been proposed out of which we select the best definition per feature category using a variable selection procedure. Our procedure involves estimating a CAR model on the first 22 weeks of data, and examining model performance in terms of mean-squared error (MSE) on the two consecutive weeks (the validation period). Variable

Table 3: MSE values for crime predictions from CAR models including POI data in the form of the total counts of venues per Foursquare category together with alternative definitions of the Twitter and taxi features.

(a) Property crime

Twitter	Taxi		
	Raw	Destination	Source
All	3.6890	3.6922	4.1847
Night	3.6682	3.6786	4.0864
log All	3.6246	3.6320	4.0081
log Night	3.6284	**3.6049**	4.0357

(b) Violent crime

Twitter	Taxi		
	Raw	Destination	Source
All	0.5220	0.5023	0.5193
Night	0.5218	0.5041	0.5200
log All	0.5129	**0.5020**	0.5096
log Night	0.5200	**0.5020**	0.5137

selection for prediction purposes where the underlying process is still of interest is notoriously difficult. Some variables may be important in explaining but not useful in predicting an outcome. While some machine learning techniques such as Random Forests entail variable importance rankings that can guide variable selection, they may pick up non-linear relationships that linear models cannot accommodate. This would give machine learning models an unfair advantage in subsequent comparisons. Therefore, we select feature groups through optimising predictions of a linear model. We chose the CAR model for its straightforward dependence structure.

Tables 3a and 3b show MSE values for property and violent crime. We present results for alternative definitions of the Twitter and Text features. For the POI feature, we find the total counts of venues per Foursquare category to perform better than the two alternatives, the venue share and the entropy measure H across both types of crime and all possible definitions of the Twitter and taxi features. In the following, we refer to this definition as non-normalised POI feature.

Overall, we observe the best results with the non-normalised POI feature, logged nightly tweet activity, and taxi data normalised by destination. For the taxi and Twitter feature, some form of normalisation, either by taking the logarithm or weighting the flows, improves predictions; presumably through reducing the range of the variables. For the POI feature, however, using total

Features	Settings							
	1	2	3	4	5	6	7	8
Census	✓	✓	✓	✓	✓	✓	✓	✓
POI		✓	✓	✓				✓
Taxi			✓		✓	✓		✓
Twitter				✓	✓		✓	✓

Table 4: Definition of experimental settings in terms of different groups of crime predictors

counts outperforms normalisation and the entropy measure. The counts preserve differences in the POI distribution across New York City which appears to be more important to crime prediction than the shares of categories. Note that logged total tweet activity yields the same MSE as the logged nightly tweet activity for violent crime. To preserve consistency in the analysis of property and violent crime, we choose logged nightly activity.

5. Results

We consider eight different combinations of the feature groups to shed light on cross-group interactions. The census data serves as baseline and is included in all settings. The other groups are added in all possible combinations. Table 4 documents the settings, which we number from 1 to 8 in the remainder.

We begin with examining the regression models to appraise the explanatory power of the individual features and their interactions. The black-box character of machine learning models conceals such information. We consider learning methods when testing the predictive power of crime forecasting models.

In view of the (still) large number of 2 types of crime $\times$ 5 models $\times$ 8 feature settings = 80 regressions, results in the form of regression coefficients are not reproduced in their entirety. Instead, Tables 5 and 6 show the estimates from all models for setting 8, which includes all groups of features. Regression results for the other settings are available from the authors upon request.

For violent crime, the parameter estimates for rate of the male population, of vacant homes, and of female-headed family households are substantial and

are associated with higher crime counts. The negative effect of the median age confirms earlier studies where younger people are more likely to commit crimes and more likely to be victims of crime (Cohen & Land, 1987).

Property crime is associated with different parameters. Most importantly, the rate of female-headed family households has a large, negative association with property crime counts. Interestingly, the reverse is true for another covariate of social cohesion, the vacancy rate, which is positively correlated with property crime counts.

Both crime types are positively associated with some POI venue types. Regression coefficients suggest that an increment of one venue in either the food or professional category is associated with a 2% increase in property and a 1% increase in violent crime counts.

A single additional residential venue, often elderly homes, is associated with a 3% decrease in property crime; an intuitive result when considering the higher presence of watchful neighbours. A similar result is observed for nightlife venues which are associated with a 3% decrease. Violent crime, however, is only weakly positively associated with nightlife venues. This may again refer to the presence of capable guardians where areas with lots of people around at night are less vulnerable to, e.g. burglary, whereas violent crime may rise in areas with alcohol outlets. Unsurprisingly, shopping venues are more strongly positively correlated with property than with violent crime and are associated with a 6–10% increase in property crime count.

The taxi feature is significant in all settings and is associated with a crime count increase between 5% and up to 25% for both crime types.

With respect to spatial dependence, we find that estimates of the corresponding parameter in the CAR model are considerably larger than in the SAR model. For the CAR model, the δ estimate throughout the eight settings and crime types is 0.1357. For the SAR model, we obtain an average ρ estimate of 0.0629. The CAR model implies local autocorrelation where crime counts depend mainly on their neighbours. The spatial structure implied by the SAR model is more global. In this regard, we find evidence for substantial local cor-

relation of crime counts (from the CAR model), whereas the comparatively low but also significant estimate in the SAR model suggests that global dependence is only weak.

To complement the previous explanatory analysis, Tables 7 and 8 report forecast accuracy in terms of MSE of alternative crime prediction models. With few exceptions, we observe a trend of MSE values being largest in setting 1, which uses census variables only, and decreasing upon adding novel features related to Twitter activity, taxi flow, or POI among the spatial econometric models. Among the machine learning models, incorporating novel features often improves accuracy. However, we observe a decrease of model performance compared to the baseline setting 1 more often than in the case of spatial econometric models.

With respect to the performance of different model families, machine learning models outperform spatial econometric models across both feature groups and crime types. This illustrates that crime is driven by relationships more complex than the ones identified through linear regression models. The random effects estimated in the Poisson GLMM further emphasise this view. The density of the random effects for violent crime is shown in Figure 3. While the distribution is normal for property crime, there is a distinct second mode in the left tails for violent crime. This "hump" persists for the exact same 138 spatial units across all settings. These census tracts are areas with very different, namely lower, violent crime counts than their surrounding areas. Therefore, the random effects appear to capture an omitted variable, which makes these areas different from their neighbours. No other linear model accounts for this very specific violent crime effect. This explains why the GLMM outperforms other linear models in terms of MSE with noticeable margin in forecasting violent crimes (see Table 8).

6. Discussion

Our mixed approach of explanatory analysis and prediction reflects the dual objective of police and policy makers. Predictive policing does not suffice to

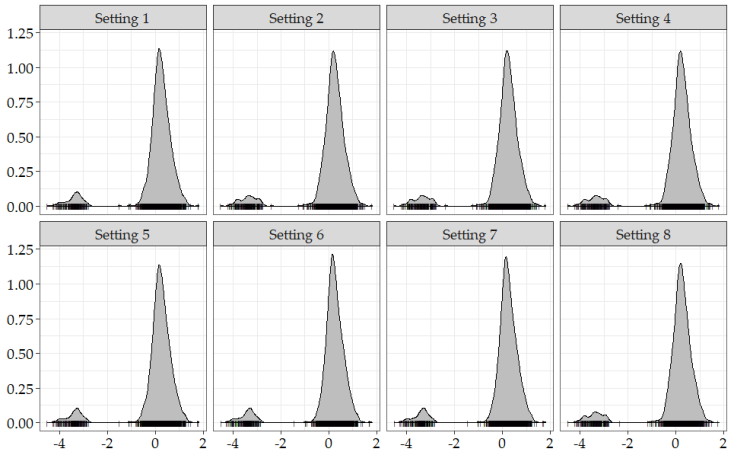


Figure 3: Density of random effects in Poisson GLMM across settings for violent crime

Parameter	CAR	GLM[1]	LR	GLMM[1]	SAR
Intercept	−0.4329	−0.2881***	−0.9876***	0.3750***	−0.9387***
	(0.2500)	(0.0809)	(0.2295)	(0.0000)	(0.2267)
Population	0.0001*	0.0001***	0.0001***	0.0001***	0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Median age	0.0120***	−0.0076***	−0.0003	−0.0076***	0.0011
	(0.0024)	(0.0008)	(0.0019)	(0.0002)	(0.0019)
Male	−1.2503**	−0.5132***	0.9174*	−0.5419***	0.4780
	(0.4012)	(0.1356)	(0.3833)	(0.0000)	(0.3787)
Black	0.6989***	0.3935***	0.2649***	0.4144***	0.1795**
	(0.0939)	(0.0305)	(0.0666)	(0.0001)	(0.0658)
Asian	0.4776***	0.1227***	0.2305**	0.0333***	0.1616*
	(0.1089)	(0.0329)	(0.0713)	(0.0000)	(0.0704)
Hispanic	0.9588***	0.4686***	0.2935***	0.4958***	0.2343**
	(0.1051)	(0.0346)	(0.0775)	(0.0001)	(0.0765)
Vacancy rate	2.3431***	0.6035***	2.3550***	0.5781***	2.0793***
	(0.2139)	(0.0528)	(0.1837)	(0.0000)	(0.1815)
Female-headed HH	−0.1444	−0.0275	1.5102***	−0.1374***	1.3984***
	(0.2468)	(0.0905)	(0.2123)	(0.0001)	(0.2097)
log night tweets	0.0971***	0.2341***	0.2018***	0.2138***	0.1197***
	(0.0101)	(0.0032)	(0.0090)	(0.0028)	(0.0089)
Entertainment POI	0.0150***	−0.0058***	0.0170***	−0.0045***	0.0157***
	(0.0037)	(0.0013)	(0.0037)	(0.0013)	(0.0037)
Uni POI	−0.0026	0.0001	0.0022	−0.0011	0.0011
	(0.0033)	(0.0013)	(0.0033)	(0.0013)	(0.0033)
Food POI	0.0545***	0.0219***	0.0443***	0.0200***	0.0514***
	(0.0046)	(0.0017)	(0.0047)	(0.0016)	(0.0046)
Professional POI	0.0187***	0.0239***	0.0220***	0.0239***	0.0161**
	(0.0056)	(0.0021)	(0.0057)	(0.0020)	(0.0056)
Nightlife POI	−0.0606***	−0.0336***	−0.0768***	−0.0336***	−0.0710***
	(0.0050)	(0.0017)	(0.0050)	(0.0017)	(0.0049)
Outdoors POI	0.0219***	0.0136***	0.0157**	0.0116***	0.0113
	(0.0060)	(0.0022)	(0.0060)	(0.0021)	(0.0059)
Shops POI	0.1444***	0.0586***	0.1221***	0.0620***	0.1247***
	(0.0050)	(0.0017)	(0.0050)	(0.0017)	(0.0050)
Travel POI	0.0351***	−0.0015	0.0328***	0.0032	0.0362***
	(0.0052)	(0.0018)	(0.0050)	(0.0017)	(0.0049)
Residential POI	−0.0642***	−0.0322***	−0.0467***	−0.0280***	−0.0475***
	(0.0053)	(0.0021)	(0.0051)	(0.0020)	(0.0051)
Taxi	0.1750***	0.0478***	0.2546***	0.0558***	0.2051***
	(0.0043)	(0.0007)	(0.0038)	(0.0007)	(0.0037)

[1] Coefficients are on the log scale.

Standard errors in parentheses. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 5: Estimates and standard errors for property crime in the full setting (i.e., setting 8).

Parameter	CAR	GLM[1]	LR	GLMM[1]	SAR
Intercept	−0.6406***	−3.2257***	−0.8597***	−11.3048***	−0.8026***
	(0.0883)	(0.1820)	(0.0803)	(0.0001)	(0.0798)
Population	0.0000***	0.0001***	0.001***	0.0002***	0.0000***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Median age	−0.0010	−0.0262***	−0.0021**	−0.0202***	−0.0014*
	(0.0008)	(0.0021)	(0.0007)	(0.0005)	(0.0007)
Male	0.7794***	2.0454***	1.1705***	−0.1671***	1.0551***
	(0.1417)	(0.2882)	(0.1341)	(0.0001)	(0.1332)
Black	0.2130***	1.3509***	0.1054***	3.8609***	0.0466*
	(0.0332)	(0.0605)	(0.0234)	(0.0005)	(0.0232)
Asian	0.0793*	0.6692***	−0.0086	3.8710***	−0.0133
	(0.0384)	(0.0775)	(0.0249)	(0.0001)	(0.0248)
Hispanic	0.3449***	1.2161***	0.1931***	4.8347***	0.1160***
	(0.0371)	(0.0658)	(0.0272)	(0.0003)	(0.0270)
Vacancy rate	0.3610***	1.3688***	0.5714***	0.4828***	0.4895***
	(0.0752)	(0.1407)	(0.0635)	(0.0001)	(0.0631)
Female-headed HH	0.8407***	1.7815***	1.6145***	−1.0915***	1.4370***
	(0.0872)	(0.1637)	(0.0742)	(0.0002)	(0.0738)
log night tweets	0.0103**	0.1092***	0.0156***	−0.0064	0.0102**
	(0.0036)	(0.0067)	(0.0030)	(0.0064)	(0.0030)
Entertainment POI	−0.0003	−0.0052	0.0028*	0.0076**	0.0010
	(0.0013)	(0.0032)	(0.0013)	(0.0029)	(0.0013)
Uni POI	0.0001	0.0067*	0.0020	0.0114***	0.0015
	(0.0012)	(0.0029)	(0.0012)	(0.0032)	(0.0012)
Food POI	0.0056***	0.0151***	0.0069***	0.0460***	0.0072***
	(0.0016)	(0.0036)	(0.0016)	(0.0036)	(0.0016)
Professional POI	0.0063**	0.0180***	0.0041*	−0.0028	0.0042*
	(0.0020)	(0.0046)	(0.0020)	(0.0044)	(0.0020)
Nightlife POI	0.0037*	0.0153***	0.0031	0.0469***	0.0028
	(0.0018)	(0.0037)	(0.0017)	(0.0036)	(0.0017)
Outdoor POI	0.0016	−0.0057	−0.0028	0.0497***	−0.0028
	(0.0021)	(0.0048)	(0.0021)	(0.0047)	(0.0021)
Shops POI	0.0036*	−0.0063	−0.0003	0.0811***	0.0002
	(0.0018)	(0.0040)	(0.0018)	(0.0039)	(0.0017)
Travel POI	0.0040*	−0.0115**	−0.0002	−0.0064	0.0017
	(0.0018)	(0.0040)	(0.0017)	(0.0039)	(0.0017)
Residential POI	−0.0051**	−0.0005	−0.0018	−0.0076	−0.0020
	(0.0019)	(0.0043)	(0.0018)	(0.0041)	(0.0018)
Taxi	0.0531***	0.1201***	0.1095***	0.0454***	0.0880***
	(0.0055)	(0.0060)	(0.0050)	(0.0880)	(0.0050)

[1] Coefficients are on the log scale.

Standard errors in parentheses. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 6: Estimates and standard errors for violent crime in the full setting (i.e., setting 8)

Model	Settings							
	1	2	3	4	5	6	7	8
LR	5.3526	4.9326	4.2713	4.6464	4.3067	4.3868	4.7538	4.2181
GLM	6.6332	6.3506	6.2617	6.1874	6.2118	6.3685	6.3065	6.1431
CAR	5.7746	5.5127	5.0709	5.3929	5.1239	5.2021	5.5228	5.0357
SAR	4.9995	4.7097	4.1641	4.4851	4.2183	4.2533	4.5534	4.1254
GLMM	6.8903	12.4732	14.0517	7.8622	8.3090	21.8949	6.8442	8.6743
RF	1.9405	1.8703	1.9334	1.9195	**1.8602**	1.8989	1.9400	1.9737
GBM	1.8840	1.8906	1.9048	1.9182	1.9301	1.9273	1.9276	1.9179
NN	2.1155	1.9027	1.9278	1.9638	2.0666	2.0722	2.0965	1.9696

Table 7: MSE values of different models for property crime predictions.

Model	Settings							
	1	2	3	4	5	6	7	8
LR	0.5010	0.4954	0.4944	0.4889	0.4896	0.4967	0.4919	0.4959
GLM	0.6064	0.6023	0.6011	0.6028	0.6056	0.6056	0.6066	0.6018
CAR	0.5172	0.5041	0.5056	0.5008	0.5109	0.5094	0.5134	0.5046
SAR	0.4970	0.4909	0.4944	0.4825	0.4881	0.4904	0.4868	0.4934
GLMM	0.4737	0.4732	0.4729	0.4759	0.4742	0.4729	0.4759	0.4742
RF	0.4681	0.4650	0.4737	0.4790	0.4807	0.4729	0.4807	0.4863
GBM	0.4529	0.4569	0.4514	0.4547	0.4559	**0.4509**	0.4567	0.4521
NN	0.4666	0.4656	0.4549	0.4625	0.4607	0.4633	0.4676	0.4564

Table 8: MSE values of different models for violent crime predictions.

reduce crime. Proactive policies addressing community resource deprivation and local effects are required to complement and amplify crime reduction efforts by police.

Points of intervention can be identified through regression analysis of the factors which co-occur with crime. The empirical results support routine activity and social disorganisation theory as drivers of criminal activity. In New York City, violent crime is taking place in neighbourhoods with poor social cohesion as evident by the positive association with vacant homes and female-headed family households. In line with disorganisation theory, social deprivation provides the context for delinquent, violent behaviour. Support for social disorganisation theory is supplemented by the fact that violent crime counts are not particularly sensitive to POI venues (cf. Table 6).

Property crime appears to be less related to the residential make up of the census tract where the crime takes place. Multiple demographic variables are insignificant, male and age among them. Rather than local deprivation, local opportunities matter as evident from the large positive coefficient for the vacancy rate variable. Overall, the coefficients for property crime capture an ambivalence between more opportunities or targets through more human activity, and more watchful eyes, preventing crime. This is illustrated in the negative association of property crime with nightlife and residential venues and the positive association with shopping venues and Twitter activity.

The notion that different circumstances drive property and violent crime differently is further supported by the rather low correlation between the crime types (Pearson's $r = 0.17$), indicating that the two crime types take place in very different areas.

Based on those results, crime prevention strategies need to account for this spatial and structural spread. Property crime is highly driven by localised opportunities, which means that interventions need to target those intersections of opportunity and offender. Violent crime, however, appears to emerge from social and structural context. Corresponding crime prevention programmes can be supported by more proactive, predictive policing strategies.

Crime forecasting accuracy is significantly improved by accounting for changing human behaviour. The relatively low reductions in MSE when adding POI variables to census-based crime forecast models suggest that static data does not suffice to forecast crime counts accurately (compare Setting 1 versus Setting 2 in Table 8 and Table 7). More fine-grained and detailed data such as Twitter and taxi data are preferable.

For violent crime, Table 8 reveals that out of the two feature categories, the taxi feature improves crime predictions more successfully than Twitter activity (compare results for setting 3 versus setting 4) and gives almost consistently lower MSE values. Joint use of the taxi and Twitter feature (setting 5) does not facilitate further improvements but consistently increases MSE compared the better of the settings where only one feature is included. Results for property crime do not show a clear trend whether taxi or Twitter data is preferable to capture human dynamics. However, we observe the same tendency as in the case of violent crime that a combination of POI with either taxi or Twitter data outperforms the baseline and POI only setting. This suggests that the POI feature develops full potential in combination with other features that provide a supplementary characterisation of human dynamics.

In view of the fact that the overall most accurate crime prediction in Table 8 and Table 7 are observed in settings that share the taxi feature, we suggest that a combination of node-specific population data in combination with edge-specific data on social taxi flow is the best combination of different data sources to predict crime rates. Clearly, criminals do not take a taxi to the scene of crime. However, the taxi feature proxies human dynamics between areas and how people proliferate crime through space. The spatial dependence matrix models only first-order dependence of immediate neighbours. Many taxi trips traverse multiple areas such that the taxi feature accounts for social connection and crime proliferation beyond just neighbouring sites.

7. Conclusion

This paper investigates the potential of new data sources on crime modelling and forecasting. It presents a multi-model solution to predicting the number of crime incidents in a census tract by combining demographic data with aggregated social media, venue, and flow data. In addition, it addresses the two-fold concerns of policy makers: preventing crime before it happens and as it happens. The linear models are crucial to understanding the social processes that generate crime within small spatial neighbourhoods. Variables in line with routine activity theory for property and social disorganisation theory for violent crime emerge as important explanatory variables. However, their predictive power is limited, which speaks to the fundamental difference between explanatory and predictive modelling.

The results from the previous section show that anonymous data on human behaviour is crucial to predicting crime. Already well-tuned machine learning models using baseline demographic data are still outperformed by models incorporating Twitter and taxi data, demonstrating the high relevance of the new feature domains. By not only relying on previous crime observations and quinquennial census data but rather on abundantly available behavioural data, the models can generalise to new areas or areas with poor reporting rates. It can also be easily adopted in other cities.

Following an applied perspective, the proposed approach can be employed to predict future problematic crime areas and improve police responsiveness and resource allocation. By analysing underlying mechanisms of different crime types, likely causes and areas for intervention have been identified. Different crime theories can explain different crime types. Further, night-time activity emerged as an important predictor as well as the destination of human movement.

This work represents an extension to existing approaches of crime predictions. A key element to future work on crime prediction will be accounting for human dynamics through a city. Beyond spatial distance alone, human movement has been shown to not only connect areas, but also propagate crime.

Future research exploiting these data should focus on this dynamic spatial crime proliferation.

References

Agresti, A. (2007). *An introduction to categorical data analysis*. (2nd ed.). Hoboken: John Wiley & Sons.

Andresen, M. A. (2006). Crime measures and the spatial analysis of criminal activity. *British Journal of Criminology*, *46*, 258–285.

Baltagi, B. H., Fingleton, B., & Pirotte, A. (2011). *Estimating and forecasting with a dynamic spatial panel data model*. Discussion Paper 95 Spatial Economics Research Centre.

Bendler, J., Ratku, A., & Neumann, D. (2014). Crime mapping through geospatial social media activity. In *Proceedings of the 35th International Conference on Information Systems* (pp. 1–16).

Beyer, K., Wallis, A. B., & Hamberger, L. K. (2015). Neighborhood environment and intimate partner violence: A systematic review. *Trauma, Violence, & Abuse*, *16*, 16–47.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*, 602–613.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014). Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th International Conference on Multimodal Interaction* (pp. 427–434).

Brantingham, P. J., & Brantingham, P. L. (1991). *Environmental criminology*. (2nd ed.). Prospect Heights: Waveland Press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Camacho-Collados, M., & Liberatore, F. (2015). A decision support system for predictive police patrolling. *Decision Support Systems*, *75*, 25–37.

Carneiro, N., Figueira, G., & Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail. *Decision Support Systems*, *95*, 91–101.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, *44*, 588–608.

Cohen, L. E., & Land, K. C. (1987). Age structure and crime: Symmetry versus asymmetry and the projection of crime rates through the 1990s. *American Sociological Review*, *52*, 170–183.

Cressie, N. (1993). *Statistics for Spatial Data*. Hoboken: John Wiley & Sons.

Cressie, N., Kaiser, M. S., Daniels, M. J., Aldworth, J., Lee, J., Lahiri, S. N., & Cox, L. H. (1999). Spatial analysis of particulate matter in an urban environment. In J. Gómez-Hernández, A. Soares, & R. Froidevaux (Eds.), *geoENV IIGeostatistics for Environmental Applications* (pp. 41–52). Dordrecht: Springer Netherlands.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*, 498–506.

Department of City Planning New York (2017). Census tracts 2010 (clipped to shoreline). URL: `http://www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page` retrieved 31/01/2017.

Federal Bureau of Investigation (2014). *Uniform Crime Reporting Handbook*. U.S. Department of Justice.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*, 367–378.

Gerber, M. S. (2014). Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, *61*, 115–125.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (2nd ed.). Berlin, Heidelberg: Springer.

Hinkle, J. C., & Yang, S.-M. (2014). A new look into broken windows: What shapes individuals perceptions of social disorder? *Journal of Criminal Justice*, *42*, 26 – 35.

Kim, J., & Kang, P. (2016). Late payment prediction models for fair allocation of customer contact lists to call center agents. *Decision Support Systems*, *85*, 84–101.

Kneib, T. (2003). Restricted maximum likelihood estimation of variance parameters in generalized linear mixed models. URL: `https://www.uni-goettingen.de/de/304966.html` retrieved 15/02/2017. Published on personal website:.

Kubrin, C. E. (2003). Structural covariates of homicide rates: Does type of homicide matter? *Journal of Research in Crime and Delinquency*, *40*, 139–170.

Kuzey, C., Uyar, A., & Delen, D. (2014). The impact of multinationality on firm value: A comparative analysis of machine learning techniques. *Decision Support Systems*, *59*, 127–142.

Land, K. C., McCall, P. L., & Cohen, L. E. (1990). Structural covariates of homicide rates: Are there any invariances across time and social space? *American Journal of Sociology*, *95*, 922–963.

Lasley, J. R. (1998). *"Designing Out" Gang Homicides and Street Assaults*. Technical Report National Institute of Justice, U.S. Department of Justice. Research in Brief.

Leroux, B. G., Lei, X., & Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In M. Halloran, &

D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (pp. 179–191). New York: Springer.

Lin, S., & Brown, D. E. (2006). An outlier-based data association method for linking criminal incidents. *Decision Support Systems*, *41*, 604–615.

Malleson, N., & Andresen, M. A. (2015). The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, *42*, 112–121.

Mears, D. P., & Bhati, A. S. (2006). No community is an island: The effects of resource deprivation on urban violence in spatially and socially proximate communities. *Criminology*, *44*, 509–548.

New York City Police Department (2016). NYPD complaint map (historic). URL: `https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Map-Historic-/57mv-nv28` retrieved 31/01/2017.

NYC Taxi & Limousine Commission (2016). TLC trip record data. URL: `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml` retrieved 02/01/2017.

Pfeffer, J., & Morstatter, F. (2016). Geotagged twitter posts from the united states: A tweet collection to investigate representativeness. doi:`http://doi.org/10.7802/1166` retrieved with permission 10/02/2017.

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, *277*, 918–924.

Sermpinis, G., Dunis, C., Laws, J., & Stasinakis, C. (2012). Forecasting and trading the EUR/USD exchange rate with stochastic neural network combination and time-varying leverage. *Decision Support Systems*, *54*, 316–329.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423. Reprint.

Traunmueller, M., Quattrone, G., & Capra, L. (2014). Mining mobile phone data to investigate urban crime theories at scale. In *International Conference on Social Informatics* (pp. 396–411).

U.S. Census Bureau (2017). Profile of general population and housing characteristics: 2010 census. URL: `https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_10_SF1_SF1DP1&prodType=table`.

Wang, H., Kifer, D., Graif, C., & Li, Z. (2016). Crime rate inference with big data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 635–644).

Wikström, P.-O. H., Ceccato, V., Hardie, B., & Treiber, K. (2010). Activity fields and the dynamics of crime. *Journal of Quantitative Criminology*, *26*, 55–87.

Williams, M. L., Burnap, P., & Sloan, L. (2017). Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. *The British Journal of Criminology*, *57*, 320–340.

Xu, J. J., & Chen, H. (2004). Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks. *Decision Support Systems*, *38*, 473–487.

Xue, Y., & Brown, D. E. (2006). Spatial analysis with preference specification of latent decision makers for criminal event prediction. *Decision support systems*, *41*, 560–573.

Yang, C. C., & Li, K. W. (2007). An associate constraint network approach to extract multi-lingual information for crime analysis. *Decision Support Systems*, *43*, 1348–1361.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

001 "Data Driven Value-at-Risk Forecasting using a SVR-GARCH-KDE Hybrid" by Marius Lux, Wolfgang Karl Härdle and Stefan Lessmann, January 2018.

002 "Nonparametric Variable Selection and Its Application to Additive Models" by Zheng-Hui Feng, Lu Lin, Ruo-Qing Zhu asnd Li-Xing Zhu, January 2018.

003 "Systemic Risk in Global Volatility Spillover Networks: Evidence from Option-implied Volatility Indices " by Zihui Yang and Yinggang Zhou, January 2018.

004 "Pricing Cryptocurrency options: the case of CRIX and Bitcoin" by Cathy YH Chen, Wolfgang Karl Härdle, Ai Jun Hou and Weining Wang, January 2018.

005 "Testing for bubbles in cryptocurrencies with time-varying volatility" by Christian M. Hafner, January 2018.

006 "A Note on Cryptocurrencies and Currency Competition" by Anna Almosova, January 2018.

007 "Knowing me, knowing you: inventor mobility and the formation of technology-oriented alliances" by Stefan Wagner and Martin C. Goossen, February 2018.

008 "A Monetary Model of Blockchain" by Anna Almosova, February 2018.

009 "Deregulated day-ahead electricity markets in Southeast Europe: Price forecasting and comparative structural analysis" by Antanina Hryshchuk, Stefan Lessmann, February 2018.

010 "How Sensitive are Tail-related Risk Measures in a Contamination Neighbourhood?" by Wolfgang Karl Härdle, Chengxiu Ling, February 2018.

011 "How to Measure a Performance of a Collaborative Research Centre" by Alona Zharova, Janine Tellinger-Rice, Wolfgang Karl Härdle, February 2018.

012 "Targeting customers for profit: An ensemble learning framework to support marketing decision making" by Stefan Lessmann, Kristof Coussement, Koen W. De Bock, Johannes Haupt, February 2018.

013 "Improving Crime Count Forecasts Using Twitter and Taxi Data" by Lara Vomfell, Wolfgang Karl Härdle, Stefan Lessmann, February 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.