Exploring Foursquare-derived features for crime prediction in New York City

Cristina Kadar
ETH Zurich
ckadar@ethz.ch

José Iria
Mobiliar
jose.iria@mobi.ch

Irena Pletikosa Cvijikj
ETH Zurich
ipletikosa@ethz.ch

ABSTRACT

Crime prediction based on traditional socio-demographic data is of limited value because it fails to capture the complexity and dynamicity of human activity in cities. With the rise of ubiquitous computing, there is the opportunity to improve crime prediction models with crowdsourced data that make for better proxies of human activity. In this paper, we propose the use of Foursquare data for crime prediction. We employ feature selection techniques to investigate the power of different features derived from Foursquare check-ins in predicting crime counts in New York over a period of 5 years. Our study shows that the number of venues (as a metric of neighborhood popularity) and the venues entropy (as a metric of neighborhood diversity) are the most discriminative features when considering all incidents. The number of users and their interactions with the venues in form of check-ins in specific types of venues (as proxies for the functional decomposition of the neighborhood) become relevant for certain types of incidents.

CCS Concepts

•Information systems → Data mining; Information systems applications; •Applied computing → *Law, social and behavioral sciences;*

Keywords

crime prediction; urban computing; open data; feature interpretation; feature selection; LBSN

1. INTRODUCTION

Crime prediction is inherently difficult. Crime is a complex social phenomenon driven by three forces: (1) the offender's motivation, (2) the victim's vulnerability, and (3) the absence of a capable guardianship [4] – or more generally said, the environment (seen as the time and place supporting the victimization) where the offender and victim come together [22]. This yields a highly dynamic and complex system, and scholars are still investigating various characteristics of the three forces for predictive power.

Traditionally, criminological studies have focused solely on socio-demographic attributes as factors correlating with victimization and have noticed that specific groups of people were facing higher risk of victimization compared with other groups [8]. But census data has an intrinsic limitation, in that it only offers a static and sometimes obsolete image of the city, without capturing the people dynamics over time and space. There is now the opportunity for non-conventional factors to be integrated in crime prediction models by tapping into novel data sources that reflect the structure and dynamics of our cities. With the emergence of mobile phones and other types of ubiquitous computing, a plethora of data sources can now offer better proxies for human activity, from mobile calls to geo-located Tweets or Foursquare check-ins. In particular, local-based social networks (LBSNs) like Foursquare offer a very vivid image of the city, being able to not only to provide time and location of human activity, but also the context (like traveling, shopping, working, going out, etc.) in which activities occur.

In this paper we assess the predictive power of crime prediction factors derived from information crowdsourced by Foursquare users. This is an initial but essential step towards deriving robust crime prediction models from data sources describing human dynamics in urban environments. To the best of our knowledge, our work constitutes the first attempt to:

1. use Foursquare data, beyond basic Points of Interest (POIs) data, as explaining features in a crime context;

2. perform extensive feature interpretation and selection on this data with the end goal of fine-grained crime prediction in mind.

The remainder of this paper is structured as follows. First, the related literature is surveyed in Section 2. This contains a brief overview of research in the area of urban computing which exploits similar data sources, followed by a survey of existing crime prediction models in the data mining community. Section 3 explains in detail the collected dataset, while Section 4 elaborates on the used methodology, derived potential crime correlates, and empirical results of the analysis. Finally, in Section 5, we discuss the implications of the results obtained, and conclude with a mention of future work.

KDD - Urban Computing WS '16 San Francisco, California USA

ACM ISBN 978-1-4503-2138-9.

DOI: 10.1145/1235

2. RELATED WORK

2.1 Urban Computing

Nowadays, sensing technologies and large-scale computing infrastructures have produced a variety of big data in urban spaces: geographical data, human mobility, traffic patterns, communication patterns, air quality, etc. The vision of urban computing, an emerging field coined by Zheng and collaborators [24], is to unlock the power of big and heterogeneous data collected in urban spaces and apply it to solve major issues our cities face today. They identify seven application areas of urban computing: urban planning, transportation systems, environmental issues, energy consumption, social applications, commercial applications, and public safety and security.

For example, within the urban planning and transportation domains, the authors in [23] attempt to infer the functions of different regions in the city by analyzing the spatial distribution of commercial activities and human mobility traces, while the authors in [3] mine different urban open data sources for optimal bike sharing station placement. Furthermore, for commercial purposes, researchers mine online location-based services for optimal retail store placement [9] or metro data for insights into the financial spending of public transport users [12].

Within the public safety and security sector, very few publicly available studies exploit human dynamics data. Scholars have just recently started to investigate the potential use of social media [6], of mobile data [2], and of taxi flow data [17] for the specific purpose of crime prediction.

2.2 Crime Prediction

A basic and widely applied model for understanding criminal patterns is the hot spot model [5]. It clusters past incidents into regions of high risk (the so-called hot spots) using statistical methods like kernel density estimation (KDE) or mixture models. In this case the past is prologue for the future: crime is likely to occur where crime has already occurred! These models exploit solely the historical crime records and do not integrate any further information. Their biggest disadvantage is that they cannot be generalized to areas without historical data.

The spatiotemporal generalized additive model (ST-GAM) [18] and the local spatiotemporal generalized additive model (LST-GAM) [19] start looking at socio-demographic data (like population density, unemployment rate, education level, net income, social aid, etc.), temporal data (like time of day/week/year, temporal proximity to special events such as football games, etc.), and spatial data (like spatial proximity to bus stations, governmental buildings, pawn shops, night life establishments, or target type such as household, store, park, etc.) describing a criminal incident. These models are extensions of regression models on grids, where the spatiotemporal features can be indexed by time. The real effect of such fetures on the level of crime is often not linear, and the generalized additive models can capture such non-linearities by modeling the link function of the dependent variable as a linear combination of unknown smooth functions of the independent variables.

In the same vein, authors in [16] exploit POIs from different sources to build classifiers of urban deprivation (a composite score of seven domains, with crime being one of them) at neighborhood level. Furthermore, authors in [15], assess the potential of metro flow data to identify areas of high deprivation in the city.

But only very recent research has started to incorporate human dynamics data into crime prediction models. Gerber and his co-authors [6, 21, 20] were the first to investigate the potential of social media for criminal forecasting. First, they show that an ST-GAM model incorporating Twitter-derived information performs better than the basic ST-GAM. They also show that combining topics derived from the Twitter stream with the historical crime density delivered by a standard KDE under a logistic regression model leads to an increase in the prediction performance versus the standard KDE approach for most of the tested crime types. Combining for the first time demographic data and aggregated and anonymized human behavioral data derived from mobile made public by Telefonica as part of a hackathon, Bogomolov and colleagues were able to obtain an accuracy of almost 70% when predicting whether a specific area in the city will be a crime hotspot or not [2]. Finally, researchers in [17] craft nodal features (using demographics and POIs data) and edge features (using geographical proximity and taxi flow data) to explain crime rate of a neighborhood given information of all other neighborhoods. By employing simple linear and negative-binomial regressions, they show that the addition of POI and taxi flow data reduces the prediction error by up to 17.6%.

3. DATASET

New York City (NYC) is a city that has experienced crime across time, though the levels have dropped since the 1990s [11], due to a series of factors including new policing tactics and the end of the crack epidemic [1]. Furthermore, as part of an initiative to improve the accessibility, transparency, and accountability of the city government, the NYC OpenData catalog has made over 1300 datasets available online as of May 2016[1]. The repository provides data in machine-readable formats on buildings, streets, infrastructure, businesses, and other entities within the city, including permits, licenses, crime related data, and 311 complaints.

More importantly, its 8.5 million inhabitants[2], leave rich digital footprints of their daily activity in various local-based online services, NYC being the most popular city on Foursquare[3] with about 132 million checkins as of May 2016[4].

3.1 NYC OpenData crime data

The dataset contains crime data across seven felony types: grand larceny (which is the theft of another's property – including money – over a certain value), robbery, burglary, felony assault, grand larceny of motor vehicle, rape, and murder and non-negligent manslaughter and has been downloaded from the NYC OpenData platform on 09.03.2016. For anonymization reasons, in case the offense has not occurred at an intersection, the New York Police Department (NYPD) projects the location of the incident to the center of the block (street segment). Furthermore, crime complaints which involve multiple offenses are classified according to

[1]https://nycopendata.socrata.com/
[2]http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk
[3]http://www.foursquare.com/
[4]http://www.4sqstat.com/

the most serious offense[5].

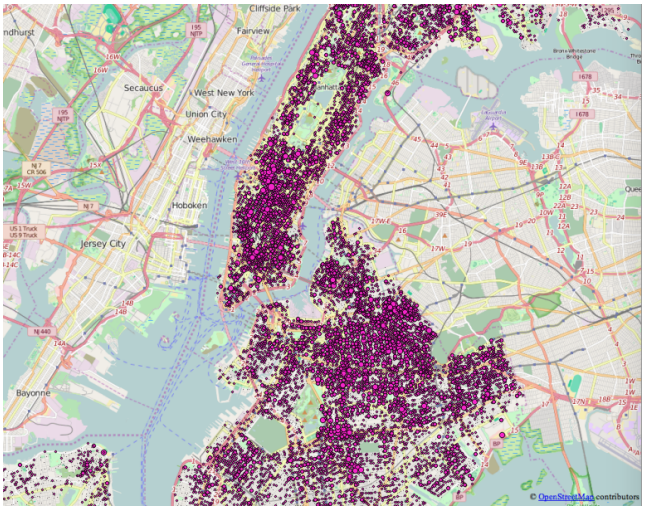

Figure 1: Burglaries locations between 2011 and 2015, weighted by number of burglary cases.

To capture as much diversity as possible, we do not limit space-wise our analysis to Manhattan only, but extend the area to three other boroughs: Bronx, Brooklyn, and Staten Island. Time-wise, we keep for analysis the data of the last 5 complete years (2011 throughout 2015), which is a long enough timeframe to aggregate crime patterns, and for which extreme values have not been observed in the crime statistics.

These preprocessing steps yield a dataset of 423384 incidents across seven types, four boroughs and five years. For exemplification, Figure 1 presents an overview of the burglary incidents within the 5 years under analysis.

3.2 Foursquare venues data

The Foursquare dataset was collected in May 2016 via the Foursquare API, using the venues search endpoint[6]. The Foursquare API has been serving both the Foursquare 8.0 and the Swarm apps since the 2014 split of the original Foursquare app[7]. While Foursquare continues to provide a local search-and-discovery service for places near a user's current location, Swarm lets the user share their location with friends at different precision levels (at city and neighborhood levels, or by checking-in to a specific venue).

The collected data consists of NYC venues with compact metadata like id, name, location, checkins count (total checkins ever done in that venue), users count (total users who have ever checked in), tip count (total number of tips written by users), associated categories, menu, etc. The Foursquare categories span a broad ontology, with the following 10 categories on the first level: (1) Arts and Entertainment, (2) College and University, (3) Event, (4) Food, (5) Nightlife Spot, (6) Outdoors and Recreation, (7) Professional and Other Places, (8) Residence, (9) Shop and Service, (10) Travel and Transport.

We have queried the API by searching for venues in the proximity of every incident location described previously, and this resulted into an extensive database of 245102 different venues. In total, these venues have experienced over 120 million checkins, distributed unevenly across the aforementioned ten top-categories.

[5]https://data.cityofnewyork.us/Public-Safety/NYPD-7-Major-Felony-Incidents/hyij-8hr7
[6]https://developer.foursquare.com/docs/venues/search
[7]https://developer.foursquare.com/docs/2014update

4. FEATURES FOR CRIME PREDICTION

4.1 Problem Definition

We cast the problem as a regression task on the counts of crimes on each block. For each block present in the NYC crime dataset and in the boroughs of interest, we aggregate all crime incidents (total and per crime type) occurring over the time period of interest. For each type, we keep entries with at least one count and let this be the dependent variable y. Table 1 presents the descriptive statistics of the aggregated counts per crime type: number of locations in the dataset, minimum, first quartile, median, mean, third quartile, and maximum of the respective counts.

Incident type	Locations	Min	Q1	Median	Mean	Q3	Max
all	50296	1	2	4	8.37	9	1380
grand larceny	36214	1	1	2	4.76	5	1145
burglaries	22882	1	1	2	2.81	3	213
assault	23383	1	1	2	3.33	4	167
vehicle larceny	17345	1	1	1	1.53	2	35
rape	61	11	48	73	82.03	109	275
murder	1321	1	1	1	1.16	1	8

Table 1: Descriptive statistics of the crime data.

Figure 2 is depicting the histogram of the incidents counts from 2011 to 2015 in NYC – note that the long tail of the distribution continues until 1380 and was cut from the graph for practical reasons. We can observe that the distribution of the data is positively skewed with many observations having low count values. The incident types expose similar power law distributions.

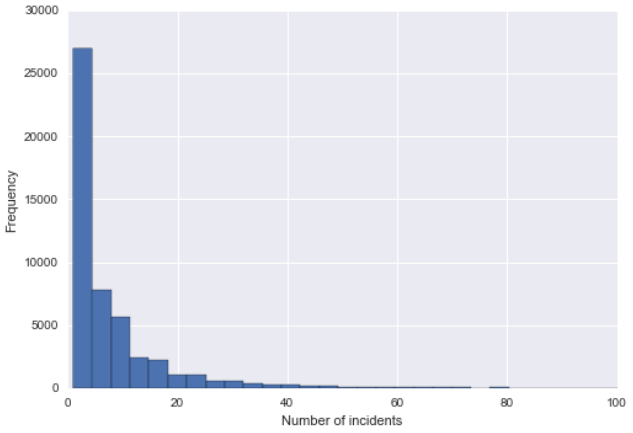

Figure 2: Incidents counts per block from 2011 to 2015.

In what concerns the independent variables $\boldsymbol{x}$, we craft a set of geographic and human dynamics features based on our raw Foursquare dataset, as explained in the following subsection. We then measure their predictive power by running different feature selection techniques and consolidating their results.

4.2 Prediction Features

We proceed by introducing the features we have derived based on the Foursquare dataset of NYC. Each feature represents a numeric score that assesses the area around a given incident location. We classify them into two overarching categories: (1) spatial features which exploit solely the static information about the venues (like location and category), and (2) human dynamics features which integrate knowledge about the way the population interacts with these venues (like check-ins).

Spatial Features

This category of features describes the urban environment around the place of interest, as captured by the spatial distribution of the Foursquare venues set, denoted as V. Specifically, we measure the density and heterogeneity of all Foursquare places that lie in a disk of varying radius r around a given incidents location l, defined as $\{v \in V | dist(v,l) < r\}$ whereby *dist* is the Euclidean distance between two locations in the local cartesian coordinate system NAD83 / New York Long Island (ftUS)[8]. As exemplification of this unit of analyis, Figure 3 depicts the area covered by a disk of 200m around a central block in the city downtown.

Figure 3: Area of radius $r = 200m$ around the street segment on 5th Avenue between Rockefeller Plaza and the Saks Fifth Avenue mall. In violet: incidents locations, weighted by total number of incidents. In blue: Foursquare venue locations, weighted by checkins counts.

Total number of venues: measures the density of venues around a location and is a static popularity metric of that area. Formally:

$$f(l,r) = |\{v \in V | dist(v,l) < r\}|$$

describes the number of venues v around the location l within a radius r, which can also be denoted as $N(l,r)$.

Venues entropy: measures the diversity of an area as captured by the categories of the venues within that area. Inspired by [9], we use the entropy measurement from information theory [14] as a diversity index. Intuitively, the entropy quantifies the uncertainty in predicting the category of a venue that is taken at random from the area. For a given location l, we denote the count of neighboring venues of category c within a radius r as $N_c(l,r)$. Formally, the entropy measures how many bits are needed to encode the corresponding vector of category counters $\{N_c(l,r)|c \in C\}$, with C being the set of the top 10 categories introduced previously, and is defined as follows:

$$f(l,r) = -\sum_{c \in C} \frac{N_c(l,r)}{N(l,r)} \times \log \frac{N_c(l,r)}{N(l,r)}$$

The higher this measure, the more heterogeneous the area is in terms of types of places, and following that, in terms of functions and activities of the neighborhood, whereas a least entropic area would indicate an area with a dominant function. For example, an area dominated by venues from the Professional and Other Places category, would indicate a part of the city where people primarly work.

Human Dynamics Features

In this section, we show how information on how the users engage with the Foursquare venues in an area, can be exploited to derive metrics of human activity in that area.

Total number of checkins: measures the popularity of the area. The total number of empirically observed checkins by Foursquare can be used as another proxy for the relative popularity of that area in the city, and is computed as follows:

$$f(l,r) = \sum_{\{v \in V | dist(v,l) < r\}} v_{checkins}$$

whereby $v_{checkins}$ denotes the total number of checkins experienced by venue v.

Total number of users: measures popularity and heterogeneity of the area, and can be regarded as a more accurate measure of human activity than the traditional population density statistics from the census:

$$f(l,r) = \sum_{\{v \in V | dist(v,l) < r\}} v_{users}$$

whereby v_{users} denotes the total number of users who have ever checked in venue v.

Total number of tips: measures popularity and quality of a given area, by looking at the involvement of the users with the venues:

$$f(l,r) = \sum_{\{v \in V | dist(v,l) < r\}} v_{tips}$$

whereby v_{tips} counts the total number of tips users have ever written about venue v.

Number of checkins per category: measures the intensity of the different activity contexts in which the Foursquare users engage, and is an empirical metric for the functional decomposition of that particular area in the city. For instance, an area with many Residence and Outdoors and Recreation checkins would correspond to a residential neighborhood, which is very different to an entertainment district, that would in turn be characterized by a high number of checkins in the Food, Nightlife Spot, and Shop and Service categories, e.g. For a given incident location l, radius r and

[8]http://www.spatialreference.org/ref/epsg/2263/

category c, this feature is calculated as follows:

$$f(l, r, c) = \sum_{\{v \in V | dist(v,l) < r \ \& \ v_{category} = c\}} v_{checkins}$$

with $v_{category}$ being the primary category of venue v.

4.3 Feature Selection Techniques

Feature selection techniques are meant to lead to better performing prediction models and to a better understanding of the underlying phenomena and the structure of the data. There are mainly two reasons why feature selection is used in supervised machine learning:

1. To reduce overfitting and improve the generalization of the subsequent machine learning models by reducing the number of features. This implicitly leads to shorter training times and models easier to interpret.

2. To gain a better understanding of the features and their relationship to the dependent variable.

These two goals are often contradictory and techniques that work well with one do not necessarily work well with the other. Especially, methods more suitable for (1) are indiscriminately applied for achieving (2). In the following we explore a set of feature selection techniques with the second goal in mind: which of the Foursquare-derived features above are discriminative for predicting the number of criminal incidents in a specific NYC location.

A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets to be used for learning, along with an evaluation measure which scores the different feature subsets. The feature selection methods are typically structured in three classes based on how they combine the subset selection and the model building: filter methods, embedded methods, and wrapper methods [7].

Filter methods are basic methods and use a proxy measure, a correlation coefficient, instead of the error rate to score a feature subset. These statistical methods are usually not computationally intensive, but produce a feature set which is not tuned to a specific type of predictive model. In what follows, we employ two popular filter methods which perform univariate feature selection. These are Pearson correlation and mutual information (MI). MI is not a metric and not normalized, instead the maximal information coefficient (MIC) [10] searches for optimal binning and turns mutual information score into a metric that lies in the range [0;1]. These correlation methods have a major common drawback: they do not take into account feature interactions. Furthermore, Pearson correlation specifically is unable to recognize non-linear relationships between the features and the dependent variable.

Embedded methods consist of machine learning techniques that automatically perform feature selection because of some inherent internal ranking of the features. One class of such methods are regularized regression models. Regularization is a technique that adds a penalty factor to the optimization function, with the goal of preventing overfitting and improving the generalization of the model. The L1 or LASSO regularization for regression adds the L1-norm of the coefficients vector $\boldsymbol{w}$ to the loss function that needs to be optimized: $\alpha \sum_i |\boldsymbol{w}_i|$ This forces weak features to have zero $\boldsymbol{w}_i$ values and yields a sparse model, thus performing automatic feature selection.

On the other hand, the L2 or Ridge regularization for regression adds the L2-norm penalty to the optimization objective, which forces the model coefficients $\boldsymbol{w}_i$ to have lower values and be spread out more evenly: $\alpha \sum_i {\boldsymbol{w}_i}^2$. The LASSO regression is unstable, meaning that the coefficients (and thus feature ranks) can vary significantly even on small data changes when there are correlated features in the data. In contrast, Ridge regression is not that volatile and, while it does not perform feature selection the same way LASSO does, it is more useful for feature interpretation: a predictive feature will get a non-zero coefficient that has a magnitude related to feature's importance.

Another popular embedded methods are decision trees and their ensemble extension, random forests. Random forests are very popular in practice, as they are are easy to use, robust, and yield relatively good accuracy on many different tasks. At every node in the decision trees of the random forest, a test is done on a single feature, with the goal to split the dataset into two subsets, so that instances with similar responses to the test end up in the same subset. The measure based on which the optimal condition is computed is called impurity, and in case of regression trees, it is variance. While training, one can compute how much each feature decreases the impurity in a tree and this can be used for ranking features. In a forest, the impurity decrease per feature can be averaged across all trees in the forest.

Finally, **wrapper methods** build on top of other selection methods, generating models on different subsets of the data and extracting the ranking from the aggregates. Stability selection is a relatively new technique which repeatedly applies a feature selection algorithm on different subsets of the data and using different subsets of the features. In the aggregation step, one can check how often a specific feature ended up being selected as important when it was present in the initial subset of features. The most popular form of feature selection in traditional statistics is recursive feature elimination (RFE). It is a greedy algorithm that adds the best feature (or removes the worst feature) at each round. At the end of the process, features are ranked according to when they were eliminated.

We let each of the above methods generate a ranked list of all the available features and then compare and contrast the results in the coming section.

4.4 Experimental Results

As some of the used methods below such as the regularized models are sensitive to the magnitude of the features, we first standardize all features by removing the mean and scaling to unit variance (by applying a standard scaler). Furthermore, we set the lookup radius r to a value of $200m$, a choice that is in agreement with what the urban community considers as the optimal neighborhood size [13], and has been used before in the urban computing literature [9]. In case the algorithms needed hyper-parameter tuning, like in the case of the regularization parameter α for regularized regressions or the number of trees n in the case of random forests, these were determined by cross-validation.

We first use the opportunity to look at the correlations matrix of the features themselves presented in Figure 4, as we expect multi-collinearity. Indeed, the total counts of venues in the proximity correlates with the total amount

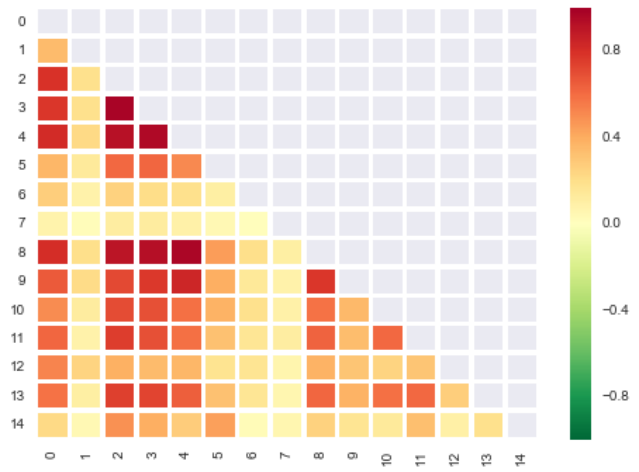

Figure 4: Features correlation matrix.

of checkins, users, and tips generated in the same area, and, furthermore, with the specific checkins in the food establishments. The venues entropy metric is uncorrelated to any of the other metrics, making it a promising discriminant. Also the specific checkins in the domains education and events seem uncorrelated to the rest of the features set.

We are now ready to run each of the above listed methods on the dataset and normalize the scores so that that are between 0 (for lowest ranked feature) and 1 (for the highest feature). By convention, for RFE, the top five feature will all get a score 1, with the rest of the ranks spaced equally between 0 and 1 according to their rank.

The following subsections describe for each method the ranks achieved by each feature, on the dataset of all incidents. Table 2 summarizes then the results.

4.4.1 *Pearson Correlation*

With Pearson correlation, each feature is evaluated independently, so the scores for the aggregated human dynamics features `checkins-total200m`, `users-total200m`, `tips-total200m` are very similar and relatively high. From the per-category checkins, the food specific checkins are the only ones having a score in the same value range. Still, the geographic features `venues-total200m` and `venues-entropy200m` are rendered as most discriminative by this method!

4.4.2 *Maximal Information Coefficient (MIC)*

MIC is similar to the correlation coefficient in identifying the same features as relevant. Additionally, it identifies `venues-residence200m` and `venues-shops200m` as similarly important.

4.4.3 *LASSO Regression*

The LASSO regression renders seven out of the 15 features as irrelevant. From the remaining ones, only the pure geographic features have considerably higher ranks.

4.4.4 *Ridge Regression*

In a similar way, the pure geographic features score highest also based on the L2 penalty in Ridge Regression. From the collinear factors of the total venues count, the only remaining feature with higher rank remains `tips-total200m`. The total number of check-ins and users receive lower weights.

4.4.5 *Random Forest*

The Random Forest experiment delivers a slightly different picture, with `venues-total200m`, `tips-total200m`, and `checkins-residence200m` scoring highest. We can see that random forest's impurity based ranking is typically aggressive in the sense that there is a sharp drop-off of scores after the first few top ones.

4.4.6 *Recursive Feature Elimination (RFE)*

Finally, RFE results seem to consist of the superset of features identified by the other methods: `venues-total200m`, `venues-entropy200m`, `tips-total200m`, `checkins-food200m`, and `checkins-residence200m`. `users-total200m` follows closely.

4.4.7 *Mean*

To summarize, the features scoring high consistently across all methods are `venues-total200m`, `venues-entropy200m`, `tips-total200m`, `checkins-residence200m`, and `users-total200m` (in this order). So, for the aggregated incident counts, there is a strong preference towards the static features based solely on location and category, seconded by aggregated check-ins values and residence checkins. This will not necessarily hold for each crime type, as we show in the following.

We continue by repeating the set of experiments for the subsets of instances of each type of crimes (see Section 3.1) and report in Table 3 only the resulting mean ranks of each feature for each crime type. While the burglary and vehicle larceny results show similar patterns to when considering all incidents, the results for the other types of incidents show particularities. In the case of rape cases the methods identify that regions with higher number of checkouts in the categories arts and entertainment, shop, and food are at higher risk, while the highest ranking features for grand larceny cases are the number of checkins at home and in food establishments. For assaults, the venues entropy is followed by the activity in the food, shop and travel venues. Finally, in the case of murder cases, all of the aggregated values of venues and checkins score high, together with the food, outdoors and residence checkins. Event check-ins score low across all crime types, with the exception of rapes.

4.5 Supervised Learning for Crime Counts

As an initial validation of the potential of these novel features for crime prediction, we briefly introduce in this section the first results in a supervised learning setting. We therefore train a LASSO linear regressor, a Ridge linear regressor, and a Random Forest regressor, all optimizing the root mean squared logarithmic error (RMSLE)[9] on the total number of incidents y. We choose this metric instead of the traditional root mean squared error (RMSE) to account for the positively skewed distribution of the dependent variable. The regressors use the complete set of features $\boldsymbol{x}$ introduced earlier. The hyper-parameters of the algorithms (regularization parameter *alpha* for the regularized regressions and number of trees n for the Random Forest) have been identified by a 5-fold cross-validation.

Table 4 summarizes the results of the three regressors. The Random Forests yield the lowest error in both cases, with RMSLE = 0.65 on the geographic features only, and RMSLE = 0.47 on the whole set of features. Also, we can observe that the set of human dynamics features was not able

[9]https://www.kaggle.com/wiki/RootMeanSquaredLogarithmicError

Feature ID	Feature Name	Pearson Corr.	MIC	LASSO Reg.	Ridge Reg.	Random Forest	RFE	Mean
Geographic Features								
0	venues-total200m	0.79	1.0	0.82	1.0	1.0	1.0	0.93
1	venues-entropy200m	1.0	0.93	1.0	0.93	0.12	1.0	0.83
Human Dynamics Features								
2	checkins-total200m	0.44	0.91	0.0	0.0	0.11	0.2	0.28
3	users-total200m	0.41	0.94	0.01	0.09	0.32	0.9	0.45
4	tips-total200m	0.45	0.89	0.13	0.31	0.63	1.0	0.57
5	checkins-arts200m	0.22	0.57	0.0	0.02	0.04	0.4	0.21
6	checkins-college200m	0.13	0.53	0.01	0.06	0.02	0.6	0.23
7	checkins-event200m	0.0	0.0	0.0	0.03	0.0	0.3	0.05
8	checkins-food200m	0.44	0.93	0.0	0.06	0.34	1.0	0.46
9	checkins-nightlife200m	0.35	0.61	0.11	0.1	0.02	0.5	0.28
10	checkins-outdoors200m	0.29	0.58	0.0	0.0	0.05	0.1	0.17
11	checkins-professional200m	0.31	0.85	0.01	0.1	0.43	0.8	0.42
12	checkins-residence200m	0.38	0.94	0.05	0.13	0.63	1.0	0.52
13	checkins-shop200m	0.33	0.93	0.0	0.08	0.28	0.7	0.39
14	checkins-travel200m	0.12	0.71	0.0	0.01	0.33	0.0	0.19

Table 2: Total incidents: the ranks of the individual features according to each selection criterion and the mean rank across all criteria.

Feature ID	Feature Name	All	Burglary	Grand Larceny	Assault	Vehicle Larceny	Rape	Murder
Geographic Features								
0	venues-total200m	0.93	0.73	0.6	0.36	0.95	0.56	0.62
1	venues-entropy200m	0.83	0.68	0.62	0.83	0.9	0.61	0.72
Human Dynamics Features								
2	checkins-total200m	0.28	0.5	0.63	0.49	0.68	0.29	0.53
3	users-total200m	0.45	0.5	0.59	0.48	0.66	0.23	0.65
4	tips-total200m	0.57	0.47	0.42	0.49	0.77	0.44	0.59
5	checkins-arts200m	0.21	0.23	0.29	0.27	0.23	0.64	0.3
6	checkins-college200m	0.23	0.24	0.23	0.33	0.26	0.48	0.25
7	checkins-event200m	0.05	0.03	0.08	0.0	0.03	0.34	0.01
8	checkins-food200m	0.46	0.48	0.57	0.54	0.66	0.55	0.59
9	checkins-nightlife200m	0.28	0.32	0.52	0.4	0.38	0.47	0.38
10	checkins-outdoors200m	0.17	0.28	0.49	0.29	0.34	0.36	0.43
11	checkins-professional200m	0.42	0.46	0.39	0.47	0.46	0.29	0.39
12	checkins-residence200m	0.52	0.45	0.72	0.32	0.52	0.3	0.43
13	checkins-shop200m	0.39	0.29	0.49	0.49	0.46	0.64	0.41
14	checkins-travel200m	0.19	0.23	0.45	0.49	0.27	0.53	0.38

Table 3: The ranks of the individual features for each of the crime types.

Method	Geographic Features		All Features	
	Hyper-parameter	RMSLE	Hyper-parameter	RMSLE
LASSO Regression	alpha=0.02	0.96	alpha=2.3	1.01
Ridge Regression	alpha=50000	0.97	alpha=25000	0.98
Random Forest	n=50	0.65	n=50	0.47

Table 4: Results of the regressors on the total incidents counts.

to decrease the error of the regularized regression models. While the ensemble method manages to optimally leverage the set of all features due to its non-parametric nature, the simpler linear models perform better on the subset of the top discriminative features identified by the selection methods in the previous section.

5. CONCLUSIONS

In this paper, we have presented first results on the investigation of LBSNs as data sources of static and dynamic features for crime prediction models. Crime is investigated at a fine-grained level: data has been segmented at city block scale and across several crime categories. The experiments have shown that from all features derived from Foursquare data, the best performing ones across all types of crimes are the venues counts and entropy, as proxies for the neighborhoods popularity and diversity. The feature selection algorithms prefer next different types of check-ins for specific types of crimes: e.g. residential and food check-ins for grand larcenies, and entertainment, shop, and food, for the rape incidents.

As future work, we plan to extend the Foursquare dataset along the time dimension, by segmenting the dynamics features across days of the week and opening hours. The same methodology could then be applied on data from other major cities around the globe where crime statistics are available and Foursquare is widely used, such that we can make more general claims about the predictive power of such factors globally. Furthermore, the same feature selection and interpretation techniques above could be applied to further ubiquitous data sources describing the fabric and pulse of our cities, like additional POIs, public transport turnstile data, or 311 calls.

6. REFERENCES

[1] A. Blumstein and J. Wallman. The Rise and Decline of Hard Drugs, Drug Markets, and Violence in Inner-City New York. In *The Crime Drop in America*, pages 164–206. Cambridge University Press, 2000.

[2] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland. Once upon a crime: Towards crime prediction from demographics and mobile data. In *ICMI '14*, pages 427–434, 2014.

[3] L. Chen, D. Zhang, G. Pan, X. Ma, D. Yang, K. Kushlev, W. Zhang, and S. Li. Bike sharing station placement leveraging heterogeneous urban open data. In *UbiComp '15*, pages 571–575, New York, New York, USA, 2015. ACM Press.

[4] L. E. Cohen and M. Felson. Social Change and Crime

Rate Trends: A Routine Activity Approach. *American Sociological Review*, 44(4):588, 1979.

[5] J. Eck, S. Chainey, J. Cameron, and R. Wilson. Mapping crime: Understanding hotspots. Technical report, U.S. Department of Justice, 2005.

[6] M. S. Gerber. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61(1):115–125, 2014.

[7] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 2003.

[8] M. J. Hindelang, A. D. Biderman, M. R. Gottfredson, and J. Garofalo. *Victims of Personal Crime: An Empirical Foundation for a Theory of Personal Victimization.*, volume 11. 1982.

[9] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo. Geo-spotting: Mining Online Location-based Services for Optimal Retail Store Placement. In *KDD '13*, page 793, New York, New York, USA, 2013. ACM Press.

[10] J. B. Kinney and G. S. Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

[11] P. A. Langan and M. R. Durose. The Remarkable Drop in Crime in New York City. In *International Conference on Crime*, 2003.

[12] N. Lathia and L. Capra. Mining mobility data to minimise travellers' spending on public transport. In *KDD '11*, page 1181, New York, New York, USA, 2011. ACM Press.

[13] M. Mehaffy, S. Porta, Y. Rofè, and N. Salingaros. Urban nuclei and the geometry of streets: The 'emergent neighborhoods' model. *Urban Design International*, 15(1):22–46, 2010.

[14] C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[15] C. Smith, D. Quercia, and L. Capra. Finger on the pulse: Identifying deprivation using transit flow analysis. In *CSCW'13*, pages 683–692, New York, New York, USA, 2013. ACM Press.

[16] A. Venerandi, G. Quattrone, L. Capra, D. Quercia, and D. Saez-Trumper. Measuring Urban Deprivation from User Generated Content. In *CSCW'15*, Vancouver, BC, Canada, 2015.

[17] H. Wang, D. Kifer, C. Graif, and Z. Li. Crime Rate Inference with Big Data. In *KDD'16*, San Francisco, California, USA, 2016.

[18] X. Wang and D. E. Brown. The spatio-temporal generalized additive model for criminal incidents. In *ISI '11*, pages 42–47, 2011.

[19] X. Wang and D. E. Brown. The spatio-temporal modeling for criminal incidents. *Security Informatics*, 1(1):2, 2012.

[20] X. Wang, D. E. Brown, and M. S. Gerber. Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In *ISI '12*, pages 36–41, 2012.

[21] X. Wang, M. Gerber, and D. Brown. Automatic Crime Prediction using Events Extracted from Twitter Posts. In *SBP '12*, volume 7227, pages 231–238, 2012.

[22] P. Wilcox. Theories of Victimization. In L. Grove and G. Farrell, editors, *Encyclopedia of victimiology and crime prevention*, pages 978–986. Sage Publications, Inc., 2010.

[23] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and POIs. In *KDD '12*, page 186, New York, New York, USA, 2012. ACM Press.

[24] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban Computing: Concepts, Methodologies, and Applications. *ACM Transaction on Intelligent Systems and Technology*, 5(38), 2014.