Journal of Intelligent Information Systems (2020) 55:535–559
https://doi.org/10.1007/s10844-019-00587-4

Analysis of street crime predictors in web open data

Yihong Zhang[1] · Panote Siriaraya[2] · Yukiko Kawai[2] · Adam Jatowt[1]

Received: 20 August 2019 / Revised: 1 November 2019 / Accepted: 1 November 2019 /
Published online: 21 November 2020

Abstract

Crime predictors have been sought after by governments and citizens alike for preventing or avoiding crimes. In this paper, we attempt to thoroughly analyze crime predictors from three Web open data sources: Google Street View (GSV), Twitter, and Foursquare, which provides visual, textual, and human behavioral data respectively. In contrast to existing works that attempt crime prediction at zip-code level or coarser granularity, we focus on street-level crime prediction. We transform data assigned to street-segments, and extract and determine strong predictors correlated with crime. Particularly, we are the first to discover visual clues on street outlooks that are predictive for crime. We focus on the city of San Francisco, and our extensive experiments show the effectiveness of predictors in a range of tests. We show that by analyzing and selecting strong predictors in Web open data, one could achieve significantly better crime prediction accuracy, comparing to traditional demographic data-based prediction.

Keywords Crime prediction · Web open data · Image and text analysis

1 Introduction

Crime creates negative impacts on people's lives, and therefore the ability to predict crime has been sought after by governments, business owners, and citizens alike. By predicting the likely number of crimes on a street, governments can design police patrolling more

✉ Yihong Zhang
yhzhang7@gmail.com

Panote Siriaraya
spanote@gmail.com

Yukiko Kawai
kawai@cc.kyoto-su.ac.jp

Adam Jatowt
adam@dl.kuis.kyoto-u.ac.jp

[1] Department of Social Informatics, Graduate School of Informatics, Kyoto University, Kyoto, Japan

[2] Division of Frontier Informatics, Kyoto Sangyo University, Kyoto, Japan

effectively (Camacho-Collados and Liberatore 2015), business owners can choose better business locations, and citizens can plan safer travel routes (Kim et al. 2014). We follow the hypothesis of geographical crime analysis research that seek to understand the environmental and local community features related to crime (Taylor et al. 1985; Graif et al. 2014). For instance, the *broken windows theory* states that the relationship between crime and environment such as visible signs of disorder and lack of maintenance (hence "broken windows") encourage further crime including serious ones (Wilson and Kelling 1982). Given the emergence of various geographical Web open data in recent years, geographical crime prediction using Web open data has attracted a number of research efforts (Gerber 2014; Wang et al. 2016; Zhao and Tang 2017; Yang et al. 2017). Essentially, these research aim to predict crime occurrence from open data, often open government data. Existing works mostly use zip-code level geographical unit or grids of square kilometers (Zhao and Tang 2017; Wang et al. 2016). However, it has been shown in literature that crime rates can differ significantly from street to street, even in the same neighborhood (Eck et al. 2007; Weisburd 2015; Gill et al. 2017). For example, with the "law of crime concentration at place", Weisburd states that, *there is strong street-by-street variability in crime within cities*, and that there are crime hot spots in good neighborhoods, while bad neighborhoods can have streets free of crime (Weisburd 2015). This situation is particular critical, for example, in a safety-based routing system to help users avoid dangerous streets (Utamima and Djunaidy 2017; Kim et al. 2014). Therefore, we extend the existing literature by analyzing crime predictors at a street level.

Specifically, in our analysis we provide answers to the following important research questions that could inform the design of effective predictive applications and planning of further research:

1. *What crime type can be predicted using which Web open data?*
2. *What particular elements in the data are predictive for the crime?*
3. *Can we use models trained with areas where crime reports are available to predict in an area where crime reports are not available?*
4. *Can image data about street outlook help in crime prediction?*
5. *Are Web open data more useful than demographic features?*

With these research questions, we attempt to address the challenges of incorporating novel data sources, particularly street images, in crime prediction, which enables us to conduct the analysis based on a fine location granularity. Although in recent years, crime data has been published online by governmental efforts, fine-grained crime records are still available only for a handful of cities. When analyzing 20 large cities around the world, only a few turned out to have block-level or finer crime data available, including New York, Chicago, San Francisco, and London. None of the Asian cities investigated has published fine-grained crime data. On the contrast, Web open data sources that we use have vast coverage of a large number of cities around the world with street-level details. Google Street View covered 39 countries and about 3,000 cities in 2012.[1] Twitter has more than 300 million monthly-active users distributed evenly in Asia Pacific, Europe and America.[2] About 1% of the 500 million tweets sent per day have geo-location attached,[3] making it a dense geographical data. Soon after its launch in 2009, Foursquare has covered 100 cities in America, Europe, and

[1] https://en.wikipedia.org/wiki/Google_Street_View

[2] https://www.statista.com/statistics/303684/regional-twitter-user-distribution/

[3] http://www.internetlivestats.com/twitter-statistics/

Asia Pacific,[4] and has since then reached many more cities. Hence it will be beneficial if we can discover predictors in Web open data for crime prediction, so they can be used in various cities and regions that do not have crime data available. Because such data sources have become available only recently, there is only a limited number of previous studies that make use of them, especially for the purpose of crime prediction.

The three data sources, Google Street View (GSV) images, Twitter tweets, and Foursquare venues provide us with visual, textual, and human behavioral geographical data. GSV images are 360-degree panorama street images captured mostly using cameras installed on cars. They are good representations of street outlooks. Previous works have used them for street-based studies such as street perception analysis (De Nadai et al. 2016). To the best of our knowledge, our work is the first to consider GSV images for street crime prediction, and we believe some visual clues contained in these images can be related to street crimes. *Twitter* is a micro-blogging platform that allows users to share short messages up to 140 characters. Tweets can include a wide range of topics about personal lives, news, opinions, and may also reflect the character of locations they are sent from. In the particular objective of crime prediction, tweets have been used as a prediction signal through a text-to-topic transformation - Latent Dirichlet Allocation (LDA) (Yang et al. 2017; Wang et al. 2012; Gerber 2014). In this work, besides LDA we also investigate the predictive potential of two other transformations of tweet content, TFIDF and word embeddings. Word embeddings capture the semantics of the text, which can contain signals related crime at a higher level (Pennington et al. 2014). *Foursquare* is a location-based online service that serves as a business directory as well as an activity record platform. A person can *check-in* to a venue (i.e., bars, restaurants, museums, etc.) listed in Foursquare, and the venue will keep a record of the checked-in people. In the crime prediction literature, venues and check-in activities have been used as prediction signals (Kadar et al. 2016; Zhao and Tang 2017). Note however that Foursquare venue categories are arranged at different levels. Different from existing works, which use only selected high-level categories from Foursquare (i.e., 9 categories such as food, professional places or shops), we analyze the impact of all categories including also the lowest possible level categories for venues (e.g., Chinese restaurant, factory or grocery store), which allow us to capture detailed information about specific venues that correlate with crime. We note that all the three above-discussed data sources provide accurate geo-location information, which can be used for street-level analysis.

To summarize, our main contributions are the following:

- We propose to use fine-grained geographical Web open data for street-level crime prediction. Preprocessing and vectorization of visual, textual, and human behavioral data are discussed in detail and can be easily replicated. To the best of our knowledge, our work is the first to study Web open data that contains both images and text for street-level crime prediction. Our findings therefore can benefit greatly the fine-grained crime prediction, especially when visual data is involved in the prediction process.
- We thoroughly analyze and identify important features in Web open data that are correlated with crime. These features include visual clues as well as text semantics and business venue distributions. As suggested by the “broken windows” theory, such analysis provides valuable insight for crime prevention and city planning.
- With extensive experiments, we evaluate different feature selections based on the feature importance we identify. We also compare Web open data features to traditional

[4] https://mashable.com/2009/11/19/foursquare-50-more-cities/

demographic features, and show that Web open data can serve as better crime predictors than demographic data, evident in significantly higher prediction accuracy.

The remainder of this paper is organized as the following: in Section 2, we discuss related work. In Section 3, we present our method to collect, pre-process, and transform relevant Web open data. In Section 4, we present our analysis of crime predictor in Web open data. Section 5 presents our experimental evaluation while Section 6 offers some insights and remarks obtained from this work. Finally, Section 7 concludes this paper.

2 Related work

Crime prediction using Web data has recently started attracting attention in the research community, following the availability of fine-grained crime reports. Wang et al. propose incorporating tweets into a crime prediction model (Wang et al. 2012). The authors introduce spatio-temporal generalized additive modeling (STGAM) that incorporates LDA-transformed tweets. They find that adding tweets into the prediction model, which is based on demographic and geographic data, improves the prediction accuracy. However, their tweets are limited to a certain news accounts, and they study only break and enter crimes. Gerber later makes another study on incorporating tweets into crime prediction models (Gerber 2014). Similarly, he transforms tweets using LDA and adds them to an existing model based on kernel density estimation (KDE). The experiments with 25 different crime types show that incorporating tweets improves prediction accuracy for 19 crimes types, but for a few other crime types the accuracy decreases. In his work, the target city, Chicago, is divided into 1km × 1km grid for data assignment. Kadar et al. introduce the idea of incorporating Foursquare data into prediction model, including category entropy, user dynamics, and check-in numbers (Kadar et al. 2016). The authors focus on feature selection, and the experiments on block-level prediction show the effectiveness of selected features. Ristea et al. make a study of correlation between crime and tweets given events around a football stadium (Ristea et al. 2018). While their study shows interesting connection between crime and tweets, their study is limited to certain places (e.g., football stadium) and uses 0.5km × 0.5km block as the unit of study. Aghababaei and Makrehchi propose another method to predict crime with tweets, using bag-of-words representation and a SVR classifier (Aghababaei and Makrehchi 2016). They attempt to predict whether crime trend is going up or down at a city level.

Wang et al. seem to be the first to use multiple Web open data sources in a crime prediction model (Wang et al. 2016). The data they use include Foursquare and taxi flow data. The authors propose to apply Negative Binomial Regression (NBR) for predicting the crime in an area, assuming crime data is missing for some areas. They use the total crime number provided by Chicago administration and divide the city by community areas as the unit of study. Experimental results show that when using all features generated from Web, demographic and geographic data, the prediction accuracy is the highest, and NBR is more effective than simple linear regression. Zhao and Tang propose to incorporate crime complaint records, weather, Foursquare, and taxi flow data into a prediction model (Zhao and Tang 2017), targeting the city of New York. This seems to be the first attempt to completely replace crime data with open data to build a prediction model, although data such as crime records are not available in every city. The authors also show that by considering spatial influences, prediction can be more accurate than using a simple linear regression on single areas. Their unit of study is a 2km × 2km grid. Yang et al. build a crime prediction model incorporating both tweets and Foursquare data (Yang et al. 2017). Their task is a binary classification problem, for which they divide the city into large grids and generate negative samples in even spaces. The authors experiment with different classifiers, including Naive Bayes, SVM, and Random Forest. They predict for different crime types separately and report the average accuracy in experimental evaluations. The authors show that incorporating tweets and Foursquare data increases the accuracy of a model based on KDE with past crimes. Kang and Kang propose a method to predict crime using a simple deep neural network trained on multi-modal data, including demographical data, street view images, and weather data (Kang and Kang 2017). Their study is based on crimes in each zip-code area of Chicago. Another interesting study is made by Chen et al., who propose a crime prediction method based on tweets and weather (Chen et al. 2015). They examine the effect of weather features by logistic regression with 1 × 1km geographical units.

While several different methods have been introduced in the past for predicting city crime, the community lacks thorough and comparative analysis for understanding the predictive potential of diverse kinds of features. In this paper we aim to fill in this gap and conduct in-depth analysis of the potential of Web open data for crime prediction. Importantly, our analysis is fine-grained as it is based on street segment units. We found that in the existing works, the unit for prediction is a block or a grid cell measured in kilometers (Kang and Kang 2017; Zhao and Tang 2017), and we are not aware of any existing work that deals with street-level crime prediction, for which the subject of study would be street segments with median length around 50 meters. Yet to be useful in real-world scenarios (e.g., safety-based route navigation, police force distribution) crime prediction should be based on fine-grained and natural units, i.e., street segments. Furthermore, we also incorporate novel data into our analysis that was not harnessed before for crime prediction. Specifically, we use GSV images and low-level Foursquare categories. With the benefit of the fine-grained data, we are able to achieve better prediction accuracies compared to the existing methods.

Although not directly related to our work, there is another group of existing works that deal with crime suspects instead of geographical regions, and are worth mentioning. For example, Liao et al. propose a Bayesian model that infers the next location for a series of crime by the same suspect (Liao et al. 2010). Du et al. propose a model to detect pick-pocket suspects by examining transit records, assuming that pick-pockets have different travel patterns than ordinary passengers (Du et al. 2016).

3 Web open data

As mentioned before, we use data from three sources to represent street segments, Google Street View (GSV), Twitter, and Foursquare, for the target city of San Francisco. More specifically, we collect data for the area of San Francisco defined by a bounding box with longitude/latitude coordinates (-122.523057, 37.813163) and (-122.354814, 37.708275). Crime record data is collected for training models. We also use OpenStreetMap (OSM) to obtain the geographical information of streets, which allows us to assign data to street segments. After grouping data to street segments, we convert data into vectors for further processing. For comparison purposes, we also collect demographic data, which we will present in Section 5. In this section, we discuss methods for collecting and cleaning the Web open data, as well as methods for transforming the data into vectors.

3.1 Data collection and cleaning

Data collection.[5] We first identify street segments in San Francisco using OpenStreetMap (OSM) data (Haklay and Weber 2008). In 2017, OSM contained over 4 billion nodes over the world, where each node represents a geographical point of interest (POI).[6] A street segment (also referred to in this paper as an *edge*) is defined in OSM as a series of points (particularly, it can have a high number of points if the street segment is not straight). OSM also provides the starting coordinates, ending coordinates, length and the name of the street that the segment belongs to. We collect all edges and their data in San Francisco city using OSM public API.[7] In total, the collected data contains 252,537 edges. The median length, the first quartile and the third quartile are 50, 20, and 96 meters, respectively. We then define the distance between a geographical point and a edge as the shortest distance between the point and the edge. If the edge is straight, the distance is the length of the perpendicular line from the point to the edge. We use the QGIS software[8] to perform this calculation. This measurement of distance is our basis for assigning geographical data to edges.

We obtained crime data for San Francisco from the "Police Department Incidents Data Set"[9] provided by DataSF.[10] The dataset includes information about 2,086,841 incidents of criminal activity (categorized into 39 different crime types) from the period between Jan 1, 2003 and Jun 25, 2017. Each crime incident has coordinates indicating the location of the crime, and is assigned by us to the nearest edge.

We collect GSV images using the Google Street View Image API[11] within the specified bounding box. For each data point location, there are three images representing front, rear-left, rear-right views. We stitch the three images horizontally to form a 360-degree panorama view. A total of 98,874 panorama view images were collected. We then assign an image to an edge if it is within 10 meters from the edge. Note that an image can be assigned to multiple edges. Tweets are collected by monitoring live stream using the Twitter Filter API[12] from May 2016 to April 2017, resulting in 751,628 geo-tagged tweets. We then collect Foursquare data using Venue Search API.[13] The information includes venue category and check-in count. The check-in count indicates the total number of check-ins since the venue was registered in Foursquare. We gather information for 41,515 venues in SF. Tweets and venues are then assigned to edges that are within 20 meters from them. Multiple edges can be assigned to one data point. Table 1 shows a summary of key statistics about the collected data, including the number of edges containing at least one data point, and the mean and median count of data points per edge.

Figure 1 shows Cumulative Distribution Function (cdf) of the data point counts, normalized by the maximum counts. We can see that, for tweets and Foursquare venues, the distribution of data points is more even. For GSV images, most edges have a few images while there are some rare edges containing a large number of images.

[5] The entire data used for the analysis is available upon request.

[6] https://wiki.openstreetmap.org/wiki/Stats

[7] https://wiki.openstreetmap.org/wiki/API

[8] https://www.qgis.org/en/site/

[9] https://catalog.data.gov/dataset/sfpd-incidents-from-1-january-2014

[10] https://datasf.org/opendata/

[11] https://developers.google.com/maps/documentation/streetview/

[12] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

[13] https://developer.foursquare.com/docs/venues/search

Table 1 Summary of key statistics about the data

	edges containing	mean count	median count
crime	18,569	185.4	58
GSV	33,669	4.28	3
tweets	12,452	42.29	3
venues	22,845	4.57	2

Edge containing is the number of edges containing at least one data point. Mean count and median count are the mean and median of the number of data points per edge

Data cleaning We have found some abnormal data points in the datasets. For example, a location next to the Hall of Justice has an abnormal number of crimes (63,409). It seems this is the default location for crimes without a specified location, and does not mean the actual location of these crimes. Another location near the Market Street has an abnormal number of tweets (94,069). It seems most of tweets from this location are sent by an automatic program tweeting about job advertisements. Considering the existence of such abnormal data points, we apply the outlier removal technique based on interquartile range (IQR), which defines outliers as data points outside the boundary of $Q1 - 3 \times IQ$ and $Q3 + 3 \times IQ$, where $Q1$ and $Q3$ is the first and the third quartiles of the data, respectively, and $IQ = Q3 - Q1$. Based on IQR, we remove 1,243 edges that have more than 568 crimes and 1,336 edges that have more than 57 tweets, and no edges are removed due to lower bound. After outlier removal, we have 5,936 edges with at least one crime record, one GSV image, one tweet, and one Foursquare venue.

3.2 Data transformation

GSV to inception vector Traditionally, images are processed based on color, textual, and scene detection (e.g., indoor or outdoor) (Khan et al. 2013; Ojala et al. 2002; Oliva and Torralba 2001). In recent years, deep neural networks have been used to convert image into

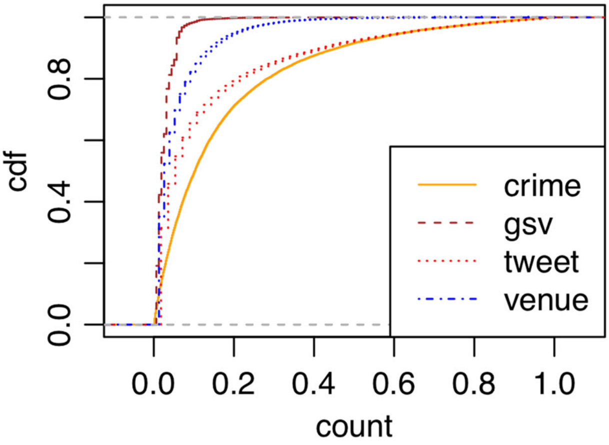


Fig. 1 CDF of data point counts by edge

vectors. Typically a complex neural network is trained with particular objective functions such as object classification (Krizhevsky et al. 2012) or sentiment identification (Chen et al. 2014), and individual layers in such network can be taken as an abstract representation of the image. Among the existing works, the most common network used is the "AlexNet" trained on 1.3 million images, the winner of the 2012 ImageNet challenge (Krizhevsky et al. 2012; Khosla et al. 2014). In recent years, more effective networks have been proposed. For instance, Szegedy et al. have released a network called Inception (Szegedy et al. 2016). It is deeper and wider than AlexNet, and involves asymmetric convolutions. The latest version, Inception-v3, has 42 deep layers, and is publicly available.[14] In the experimental evaluation for object classification, Inception-v3 reached a top-5 error rate of 3.46%, compared to 15.3% reached by AlexNet, and 6.67% by the original Inception network. Therefore, we choose Inception-v3 instead of AlexNet as the network for transforming images.

To use the Inception-v3 Network, we add a short code to the Inception-v3 program to extract the output of the third pooling layer, which is a vector of 2,048 length representing the semantics of the input image. We then select the centroid image that has the highest average cosine similarity to all other images in the edge.

Tweet to TFIDF Our first tweet-based vector transformation is term frequency inversed document frequency (TFIDF). We consider all tweets grouped to an edge as one virtual document. First we generate a bag of words by picking frequent terms from tweet data. Then for each edge, the TFIDF score of each vector element representing term t is calculated as $tf \cdot idf(t, d, D) = tf(t, d) \cdot \log \frac{|D|}{|d \in D : t \in d|}$, where d is the virtual document representing the edge, $tf(t, d)$ is the frequency of term t occurring in d, and D is the collection of all virtual documents.

However, a keyword-based method has the problem of localized association. For example, several tweets mentioning a restaurant that opened in a high crime street will cause words from the name of the restaurant to be associated with the crime, which is undesirable, as these words are characterized by low generality. Inception representation of images and Foursquare category count do not have this problem because features are generalized into kernels and categories. To mitigate this problem,we do several things. First, we remove repeated tweets for an edge that differ only by a link. Second, we discard terms with capitalized first letter as they often indicate a part of a name, unless the words are at the beginning of the tweet. Third, we remove terms that are mostly associated with only one edge. We calculate a term-edge focus as $focus = maxEdge/All$ where $maxEdge$ is the maximum frequency with which a term occurs in a single edge, and All is the frequency of the term in all edges. Terms with a $focus$ higher than 0.95 are discarded.

Tweet to LDA The Latent Dirichlet Allocation (LDA) proposed by Blei et al. has provided another way to represent text documents (Blei et al. 2003). We use LDA as our second vector transformation to generalize tweets into vectors representing topic distributions. We consider all tweets assigned to an edge as one document, and we set the number of topics as 100. This will result in a vector of 100 features for an edge where each feature is the topic probability of tweets assigned to the edge.

Tweet to GloVe Techniques such as word2vec create distributed representation of words through continuous bag-of-words or skip-gram (Mikolov et al. 2013). Such representations

[14] https://www.tensorflow.org/tutorials/image_recognition

capture the context of the words and by this represents their semantics. Among several implemented word embeddings available online, we use GloVe[15] for tweet processing as it considers both global statistics and contextual dependencies, and has proven effective in analogy and word similarity tests. The GloVe 100D word vector table we use is trained on Wikipedia 2014 and Gigaword 5, and contains 400k large vocabulary, with each word represented by a vector of length of 100. We first use GloVe to transform tweets into vectors, by averaging the vectors of the words in a tweet. Words not in the vocabulary are ignored. Then we produce a vector of the same length per edge by averaging the vectors of the tweets assigned to the edge. Each vector of an edge thus represents the aggregated semantic meaning of tweets assigned to the edge.

Foursquare to venue count The first feature set with Foursquare data is the venue category count, $VC = \{vc_1, ..., vc_l\}$, where $vc_i \in VC$ is the number of venues of category i that are present in the street segment. We use the lowest possible categories for venues. The total number of venue categories $|VC|$ is 528.

Foursquare to check-in count The second feature set with Foursquare data is the venue category check-in count, $CC = \{cc_1, ..., cc_l\}$, where $cc_i \in CC$ is the sum of check-ins for venues of category i. We add one extra feature that is the check-in count of all venue categories $\sum_i cc_i$ at the same street, as previous works have claimed that it can be used as the proxy for street popularity (Kadar et al. 2016). We again use the lowest possible categories for venues, and the total number of features for this feature set is 529.

4 Study of the correlation between crime and web open data

Given the crime data and the transformed features, we study the correlation between the features and crime number at a street segment level, which is an important step towards crime prediction. A fundamental assumption of this study is that, in location-specific social media messages, venues and human activities in streets, and street outlooks, there exist certain clues that can be associated with crime incidents. If such association exists, then using Web open data to predict crime is feasible and may even be superior to previously used predictors such as demographical information. Therefore, we aim to find the strong predictors that exist in the Web open data, represented as features. These predictors can be useful as an insight for better understanding crimes, and as a component in prediction models. In this section we present our crime type grouping, feature selection methods, and finally the determined strong predictors.

4.1 Crime types

For this study, we group specific crime types in the data into four major crime categories, namely, *offense against the person*, *offense against property*, *victimless crime*, and *theft*. Table 2 shows specific crime types in each group. The crime categories and assignments are based on online sources.[16,17] We set theft as a major crime category as it is the most

[15] https://nlp.stanford.edu/projects/glove/

[16] https://law.justia.com/codes/massachusetts/2010/partiv/titlei/

[17] http://www.cs.otago.ac.nz/staffpriv/ok/victimless.htm

Table 2 Specific crime types in four crime groups

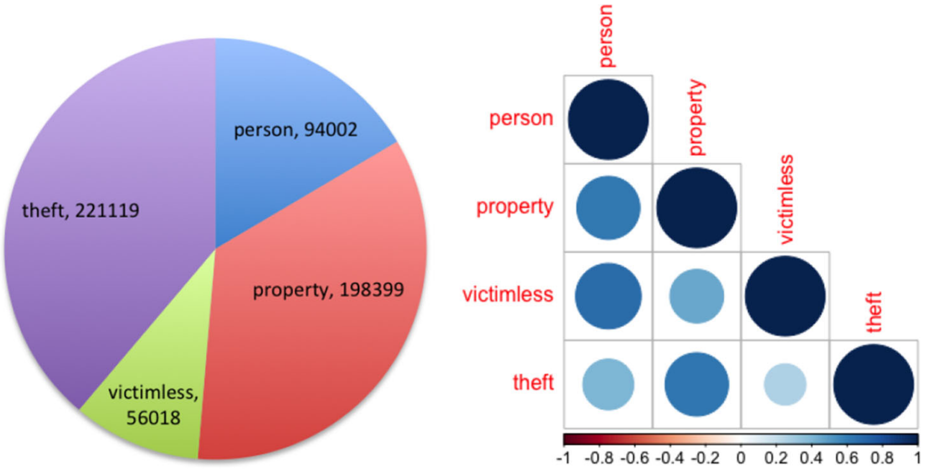
person offense	Sex Offenses Non-forcible, Sex Offenses forcible, Assault, Robbery, Kidnapping
property offense	Fraud, Embezzlement, Bad checks, Forgery Counterfeiting, Arson, Vehicle Theft, Extortion, Trespass, Bribery, Stolen Property, Burglary, Vandalism
victimless crime	Secondary Codes, Liquor Laws, Drug Narcotic, Weapon Laws, Prostitution, Driving Under the Influence, Disorderly Conduct, Gambling, Drunkenness
theft	Theft

common crime present in the data, with more than 220k occurrences, over a third of all crimes, and also because previously theft has been identified as the most common crime against tourists (Barker et al. 2002). Property crimes contain other types of property crimes except theft.

Figure 2a shows the total number and distribution of major crime categories. We can see the distribution is not even, with theft being the most common crime category, five times more than victimless crime. Figure 2b shows the correlation among crime categories. Different types of crime have some degree of correlation. Particularly, victimless crimes are more correlated with person crimes, while theft is highly correlated with property crimes.

4.2 Correlation between crime and data

We first investigate the correlation between crime number and a particular data source (GSV, Tweet, or Foursquare). We use an *adjusted multiple correlation coefficient* called *adjusted*



(a) Crime count of major categories

(b) Correlation of major categories

Fig. 2 Major crime categories and their correlations

Table 3 Adjusted R^2 for crime type and open data features

	gsv	tweet	foursquare
person offense	0.103	0.292	0.244
property offense	0.138	0.242	0.222
victimless crime	0.078	0.367	0.235
theft	0.155	0.271	0.190

R^2 to measure the correlation strength. Practically, R^2 is the correlation between the variable's values and the predictions using a linear model. A problem with R^2 is that its value always increases with more explanatory variables. As we have feature vectors of different sizes, the larger feature vectors will unfairly achieve higher R^2. The adjusted R^2 mitigates this problem by giving penalty to groups with higher variable size, and thus was chosen as a more suitable measurement for our investigation.

Note that here we wanted to answer the research question (1): *What crime type can be predicted using which Web open data?* We fit a linear model for each crime type with features for three data sources as inputs. The adjusted R^2 for each crime type vs. features is shown in Table 3.

From the table we can see that GSV data is particularly correlated with theft and less correlated with victimless crime. Tweets are correlated with victimless crime but less correlated with property crime. Finally, Foursquare data are correlated with person crime, but less correlated with theft. Among the three data sources, tweets generally have higher correlation with crime than the other two. The absolute R^2 values are not very high. We suspect that many non-predictive features could add noise to the prediction, and thus feature selection is required for better prediction.

4.3 Feature level analysis

Aiming at answer research question (2), we investigate which features are important signals correlated with crime. We consider three feature ranking methods, Pearson correlation, Spearman correlation, and Mutual Information. Mutual information is a method commonly used in feature selection works (Peng et al. 2005); it calculates the mutual dependence between two variables. For discrete random variables, the mutual information is computed as:

$$mi(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \tag{1}$$

where in our case, X and Y are the feature value and crime number, listed over edges, and $p(x, y)$ is the joint probability mass of the feature value and crime number in a particular edge.

Pearson correlation is drawn directly from the linear relationship between two variables. However, sensitivity to outliers is its weakness. A way to mitigate this problem is to consider Spearman correlation, which is less sensitive to outliers. However, Spearman correlation does not consider relative strength of values. Therefore we apply mutual information as the third ranking method.

We select the ranking method based on the following guideline. For a feature set, if Pearson and Spearman correlation agree, then we use Pearson correlation as the ranking method. If Pearson and Spearman correlations do not agree, then it is possible that Pearson correlation picks up outliers. In this case, if mutual information provides an extra degree

Table 4 Agreement analysis of feature ranking methods

		GSV	tfidf	lda	glove	venue	checkin
person offense	$f(\rho_p, \rho_s)$	0.965	0.863	0.615	0.894	0.920	0.740
	Δs	-0.002	0.041	-0.163	0.006	0.030	0.145
property offense	$f(\rho_p, \rho_s)$	0.973	0.907	0.575	0.888	0.966	0.753
	Δs	-0.006	0.019	-0.193	-0.023	0.002	0.129
victimless crime	$f(\rho_p, \rho_s)$	0.925	0.797	0.536	0.915	0.851	0.656
	Δs	0.006	0.040	-0.182	-0.008	0.065	0.187
theft	$f(\rho_p, \rho_s)$	0.854	0.844	0.734	0.698	0.903	0.750
	Δs	0.059	0.048	-0.135	-0.003	0.059	0.147

of correlation compared to Pearson correlation, we use mutual information as the ranking method. More specifically, we first calculate the Pearson correlation ρ_p, Spearman correlation ρ_s, and mutual information mi between each feature and the crime count. Then we measure agreement between methods as the Spearman ranking correlation. Note that here Spearman ranking correlation is used for a second time to measure agreements betweet correlations. We calculate the agreement between ρ_p and ρ_s as $f(\rho_p, \rho_s)$, where $f(.)$ is the Spearman rank correlation coefficient. We then calculate an agreement gain:

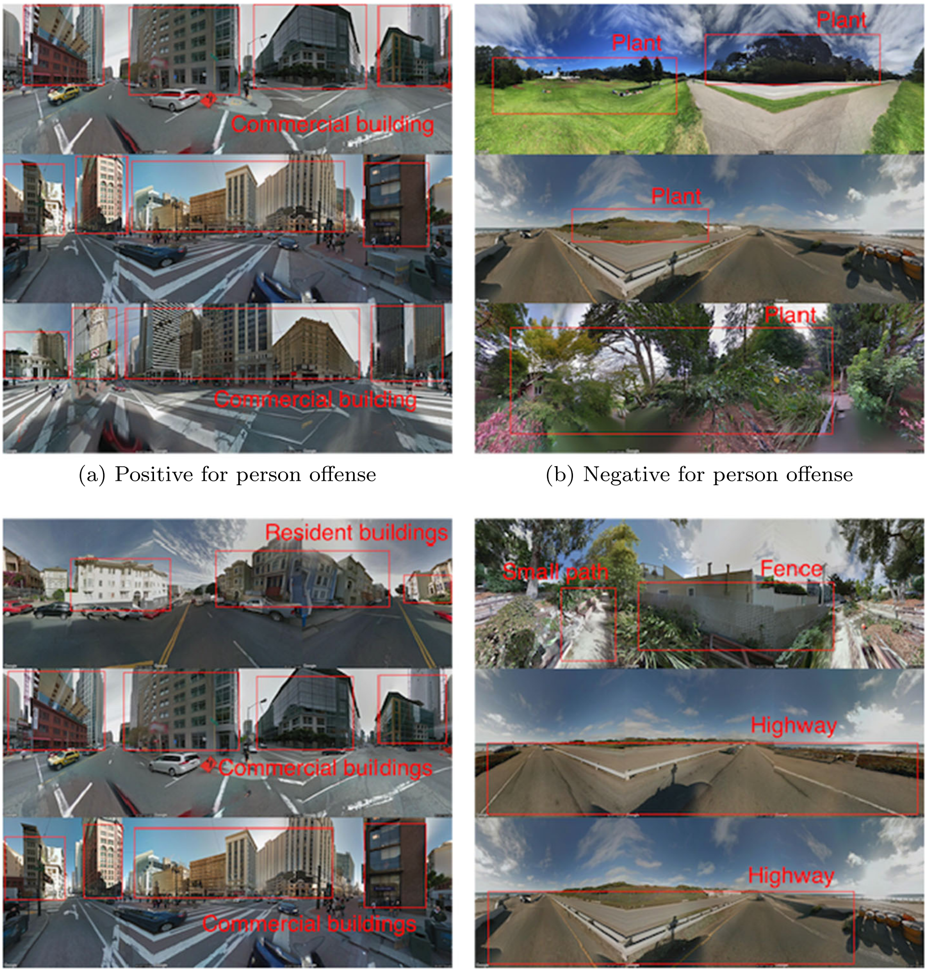
$$\Delta s = f(mi, \rho_p + \rho_s) - f(mi, \rho_p) \tag{2}$$

As the result, Δs indicates whether using mi is more preferable than using ρ_p. Table 4 shows results.

From Table 4 we can see that GSV and venue features have high agreement between ρ_p and ρ_s. TFIDF and Check-in features have relatively low agreement between ρ_p and ρ_s, but the mutual information shows that it gains positive correlation by considering both ρ_p and ρ_s. LDA and GloVe features have also relatively low ρ_p and ρ_s agreement, but mutual information does not gain positive correlation by considering both ρ_p and ρ_s. Therefore, we select the following ranking methods for different feature sets. For GSV, venue, LDA, and GloVe features, we use Pearson correlation. For TFIDF and check-in features, we use mutual information.

GSV features We first find the highly correlated features in GSV. Since it is difficult to interpret the value of an individual feature represented as CNN layers, we use the following method to reveal the meaning of features that are correlated with crimes. First, we select the top-3 GSV features correlated with a given crime type. Second, we find three images from all the images in our dataset that have the highest and the lowest sum of these features. Images with the highest sum of the top features are considered positive for the crime type, while images with the lowest sum are considered negative. The images we found are shown in Figs. 3 and 4. Note that the annotations are added manually.

From these images, we can capture some common visual clues related to crime. For person offense, the common positive clues are dense building blocks in commercial areas, while the negative clues are parks and plants with few buildings around. For property offense, the common positive clues are busy commercial areas, while the negative clues are highways and rural looking areas. For victimless crime, the common positive clues are residential buildings, while the negative clues are not so obvious but they include parks and open areas. For theft, the positive clues are dense buildings blocks, while the negative clues include gardens, woods, and small houses. It can be said from these clues, that person and



(a) Positive for person offense

(b) Negative for person offense

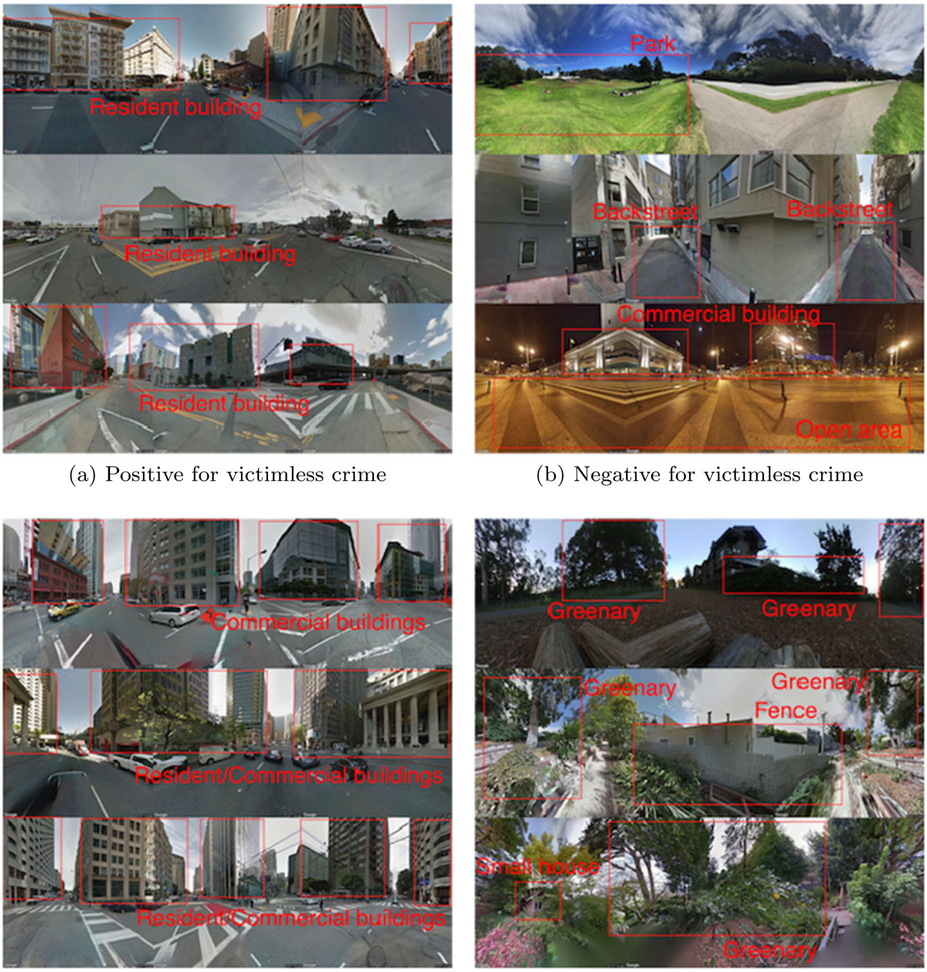
(c) Positive for property offense

(d) Negative for property offense

Fig. 3 GSV images that are positive and negative for different crime types

property offenses both tend to happen in dense building areas instead of areas with open views. Victimless crime is harder to distinguish using visual clues, which is justified by the low value of R^2 which GSV achieves for victimless crimes (see Table 3). Theft happens often in areas with large buildings. On the other hand, it does not occur often in areas with trees and small houses, where it might be more difficult to find a crime target.

Tweet features We next look at tweet features. Tables 5 and 6 show the top terms generated from TFIDF and LDA features, respectively. For LDA terms, we use the top terms assigned to the most correlated topic. By comparing with different crime types, we can see that busy traffic is an important verbal clue to all crime types. Also, all crime types seem to be more or less related to streets with drinking and night life activity. For theft, we see that it is



(a) Positive for victimless crime

(b) Negative for victimless crime

(c) Positive for theft

(d) Negative for theft

Fig. 4 GSV images that are positive and negative for different crime types (cont.)

exclusively related to terms that are about photographic and tourist activities. This supports the previous report that theft is the most common crime experienced by tourists (Barker et al. 2002).

Foursquare features Next, we look at Foursquare features, with the top features shown in Tables 7 and 8. By comparing different crime types, we find that Mexican Restaurants are correlated with person offense and victimless crime. Residential Buildings are commonly correlated with property offense and theft. In check-in data, Laundry Services is the strongest signal for all the crime types, indicating certain correlation between crime and places with laundry services. Other clues can also be found from these results, such as the correlation between Convenient Store and person offense, between Tech Startup and property crime, and between Outdoor Recreation and victimless crime.

Table 5 Top terms correlated with crime by TFIDF

person offense	traffic, stopped, great, drinking, night, mins, good, blocked, delay, day, dinner, latest, work, love, opening, time, today, tonight, last, happy
property offense	traffic, stopped, blocked, day, great, time, drinking, mins, work, delay, latest, last, dinner, today, night, love, opening, accident, lane, tonight
victimless crime	stopped, delay, night, day, mins, time, blocked, great, love, dinner, tonight, lane, good, last, today, amazing, accident, latest, work, best
theft	day, today, time, stopped, traffic, work, drinking, latest, blocked, mins, good, love, posted, delay, dinner, photo, last, opening, night, happy

5 Crime prediction

We next test the predictors in the actual prediction problems. In this paper, we focus on the missing data prediction problem, as asked in research question (3). The problem is that, if we train a prediction model on areas where crime data is available, can it be used to predict crime numbers in areas where crime data is not available? To address the problem with our edge-based data, we use 10-fold cross validation assuming that one-tenth of the edges do not have crime data. The tests we conduct include the effectiveness of different feature selections, different prediction models, and individual vs. combined features. In

Table 6 Top terms in the most correlated topic by LDA

person offense	time, lunch, style, burger, breakfast, finally, double, eating, drinks, animal, yummy, dinner, early, meal, chowder, pre, free, place, day, final
property offense	tonight, night, last, #repost, come, tomorrow, party, show, week, great, best, happy, omg, time, tasting, today, next, repostapp, happening, event
victimless crime	drinking, ipa, beer, flavor, time, nice, smooth, great, good, light, taste, sour, little, slight, beers, happy, hot, day, stop, love
theft	#sanfrancisco, #sf, #california, city, beach, posted, photo, ocean, #oceanbeach, beautiful, walk, #sunset, #nofilter, sunset, shot, view, #sanfran, boulevard, kirkham, #travel

Table 7 Top venue correlated with crime

person offense	Moving Target, Bus Station, Mexican Restaurant, Latin American Restaurant, Convenience Store, Liquor Store, Bakery, Pharmacy, Salon Barbershop, Pizza Place
property offense	Residential Building Apartment Condo, Speakeasy, Laundry Service, Salon Barbershop, Building, Moving Target, Office, Deli Bodega, Bus Line, Tech Startup
victimless crime	Mexican Restaurant, Halal Restaurant, Art Gallery, Moving Target, Liquor Store, German Restaurant, Pizza Place, Outdoors Recreation, Latin American Restaurant, Thrift Vintage Store
theft	Office, Moving Target, Tech Startup, Building, Hotel, Bus Station, Residential Building, Nightclub, Coffee Shop, Coworking Space

particular, we compare the prediction accuracies by using Web open data and demographical information, and show that Web open data can be more useful in crime prediction.

5.1 Prediction method and evaluation metrics

For prediction, we use linear regression (LR) and support vector regression (SVR). Particularly, we use ϵ-SVR (Smola and Schölkopf 2004), for which support vectors are selected if their error is within the error margin ϵ, while those producing higher error are disregarded. Although we have also tested other regression methods such as random forest and negative

Table 8 Top check-in activities correlated with crime

person offense	Laundry Service, Bus Station, Coffee Shop, Bus Line, Bank, Deli Bodega, Café, Speakeasy, Sandwich Place, Salon Barbershop
property offense	Laundry Service, Bus Station, Coffee Shop, Tech Startup, Speakeasy, Bank, Office, Deli Bodega, Café, Bus Line
victimless crime	Laundry Service, Bus Station, Bar, Coffee Shop, Speakeasy, Deli Bodega, Bus Line, Sandwich Place, Salon Barbershop, Café, Bank
theft	Laundry Service, Bus Station, Office, Tech Startup, Coffee Shop, Bus Line, Deli Bodega, Bank, Speakeasy

binomial regression as suggested in Wang et al. (2016), we found that LR and SVR achieve superior results. We use the LR and SVR implementations in R package e1071.[18]

We determine prediction accuracy using four measurements: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Symmetric Mean Absolute Percent Error (SMAPE), and Spearman Rank Correlation Coefficient (ρ_s). RMSE and MAE measure the prediction error in terms of absolute crime numbers. They are calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)}{N}} \tag{3}$$

$$MAE = \frac{\sum_{i=1}^{N}|\hat{y}_i - y_i|}{N} \tag{4}$$

where y_i is the actual crime number for edge i and $\hat{y}_i$ is the predicted crime number. RMSE and MAE tells us how many crimes are mistakenly predicted. SMAPE measures the error ratio of prediction, with regard to both prediction and actual values. It is calculated as:

$$SMAPE = 100\% \sum_i \frac{|F_i - A_i|}{|A_i| + |F_i|} \tag{5}$$

where F_i is the prediction for edge i, and A_i is the actual crime number. SMAPE is more meaningful when comparing prediction accuracy for different crime types, as the crime number in different crime types can vary significantly. ρ_s measures the ranking similarity of the prediction and actual crime number, which is important in some scenarios such as route selection. Lower RMSE, MAE, SMAPE, and higher ρ_s indicate higher accuracy. If our assumption is valid, we expect to see that Web open data, especially GSV images, achieve higher accuracy.

5.2 Determining feature set size

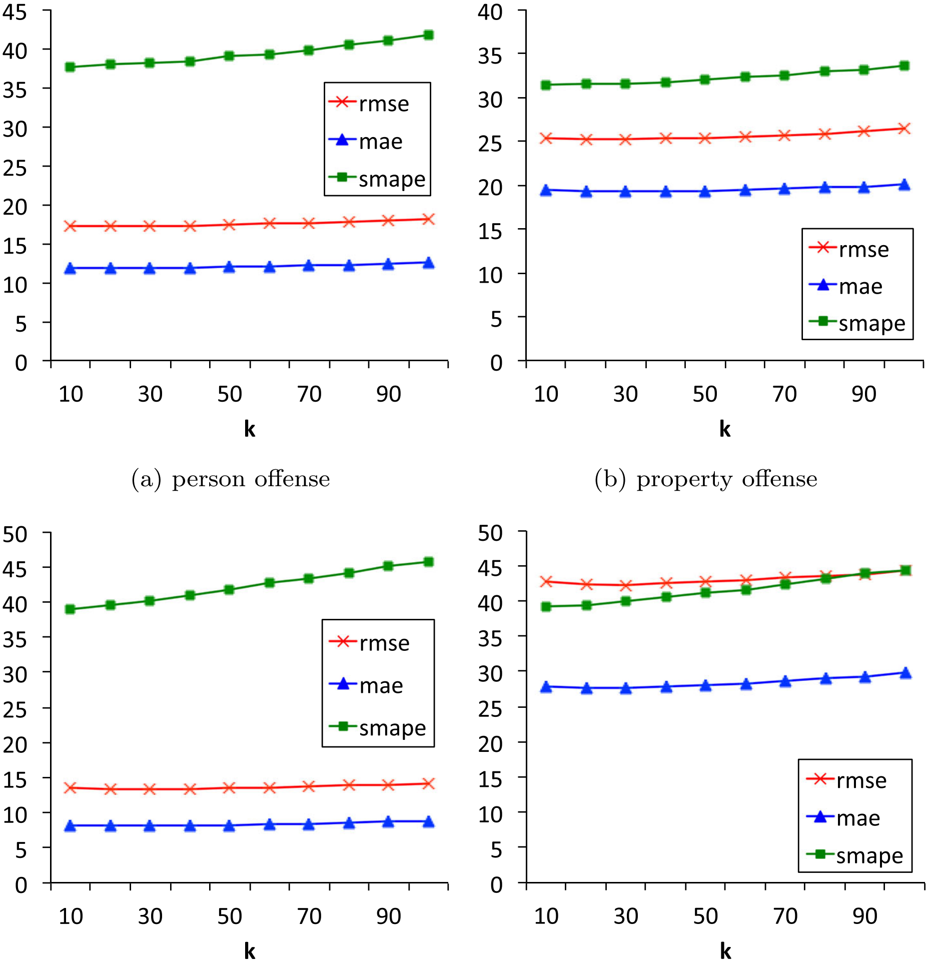
Given the feature importance ranking presented in Section 4, we test the effect of different feature sizes on prediction accuracy. We take a feature size parameter k. We use top-$3 * k$ features for GSV feature set, and top-k features for each of the other five feature sets. For instance, if we set $k = 20$, the selected feature set will consist of 160 features, including top-60 features for GSV, and top-20 features from other five feature sets. As the first test, we run linear regression and measure prediction error as RMSE, MAE, and SMAPE, using 10-fold cross validation, for a number of different ks. The results are shown in Fig. 5.

As we can observe from the figure, the prediction error continues to rise with k for $k > 30$ consistently in all the crime groups. It means that adding more features to the feature set after a certain size will not help improve prediction accuracy. Thus in subsequent studies, we set $k = 20$, which tends to provide the best prediction accuracy.

5.3 Prediction accuracy

In this section, we present the evaluation of different predictions including predictions with different feature selection approaches, individual feature sets and a comparison with demographic features, coarse versus fine granularity data, different image selection approaches, and a comparison of day and night crimes.

[18] https://cran.r-project.org/web/packages/e1071/index.html



(a) person offense

(b) property offense

(c) victimless crime

(d) theft

Fig. 5 Prediction error with linear regression given different feature size. Generally, for feature size > 30, larger feature sizes will decrease prediction accuracy

5.3.1 Prediction with web open data using different feature selections

Table 9 shows prediction accuracy of different feature selection strategies on Web open data, with LR and SVR and 10-fold cross validation. We can see that the feature selection method (discussed in Section 4.3) that mixes Pearson correlation and mutual information generally performs better than using entirely Pearson correlation or mutual information. Although Pearson-ranked features provide higher accuracy for property offense and theft with linear regression, they achieve the poorest results for person offense and victimless crime, with significantly higher RMSE than other methods, possibly due to strong localized feature association of these crimes. Model-wise, we see that linear regression is able to

Table 9 Performance of different feature selection approaches measured in prediction errors and ρ_s

		MI		Pearson		mixed	
		LR	SVR	LR	SVR	LR	SVR
person offense	RMSE	17.46	17.69	23.97	17.59	**17.13**	17.47
	MAE	12.00	11.17	12.05	11.03	11.84	**10.99**
	SMAPE	37.77	37.27	37.57	36.90	37.83	**36.66**
	ρ_s	0.370	0.398	0.388	0.410	0.382	**0.417**
property offense	RMSE	25.44	25.73	**25.08**	25.33	25.17	25.39
	MAE	19.50	18.93	19.16	**18.56**	19.26	18.63
	SMAPE	31.64	31.70	31.37	**31.14**	31.44	31.22
	ρ_s	0.378	0.395	0.399	**0.429**	0.394	0.425
victimless crime	RMSE	13.60	13.61	18.03	13.60	**13.41**	13.54
	MAE	8.28	**7.22**	8.27	7.25	8.18	7.24
	SMAPE	39.73	**36.75**	38.76	37.38	39.57	37.07
	ρ_s	0.312	**0.388**	0.317	0.360	0.305	0.380
theft	RMSE	43.65	43.97	**41.51**	42.26	42.07	42.63
	MAE	28.41	25.68	27.23	**25.09**	27.59	25.17
	SMAPE	39.43	37.26	39.21	36.81	39.48	**36.77**
	ρ_s	0.353	0.416	0.397	0.436	0.390	**0.442**

For each test, predictions using LR and SVR are measured separately

The bold font is used to indicate the best result

achieve lower RMSE, while SVR tends to have better SMAPE and ρ_s results. In subsequent tests, we use linear regression model for obtaining RMSE, and SVR for ρ_s.

5.3.2 Prediction with individual features and a comparison with demographic features

We next test the predictiveness of individual feature sets, including demographic features. With this test we reveal the effectiveness of road image data and demographic features as asked in research question (4) and (5). We use demographic features proposed in Wang et al. (2016), which include the following demographic features effective for crime prediction: total population, population density, poverty, disadvantage index, residential stability, ethnic diversity, race distribution. We obtain these demographic information from a government website,[19] and assign them to street segments. Prediction accuracy measured as RMSE using linear regression and ρ_s using SVR are shown in Tables 10 and 11, respectively. For this experiment, we also measure the significance of improvements from using Web data and combined data over demographic data, by performing the Diebold-Mariano test, which tests whether two series of prediction differ significantly (Diebold and Mariano 2002). The p-value of Diebold-Mariano test on the errors of tested methods is shown in the last two rows of Table 10.

We see that among individual features, venue features provide the best accuracy, with lower RMSE and higher ρ_s across all crime categories. To answer research question (4), we

[19] https://www.census.gov/

Table 10 Linear Regression RMSE of individual and combined features, including demographical features

	person offense	property offense	victimless crime	theft
GSV	18.07	26.49	13.85	44.00
tfidf	18.42	27.27	14.00	45.26
LDA	18.57	27.46	14.12	45.62
GloVe	18.43	27.42	14.00	45.63
venue	17.38	25.69	13.53	43.35
check-in	18.16	27.02	13.98	45.33
Web data	17.13	25.17	13.41	42.07
demographic	18.42	27.31	13.88	44.84
combined	**16.86**	**25.04**	**13.02**	**41.66**
Significance of improvement over demographic data (p-value)				
Web data	1.198e-25	3.314e-44	9.128e-18	3.294e-23
combined	3.044e-29	6.776e-46	6.271e-17	2.220e-28

A significant test for combined features is also shown

The bold font is used to indicate the best result

see that GSV is a strong predictor comparing to other data, with high accuracy for all crime types, only slightly worse than venue features. LDA topic distribution features, which have been used by several previous works, turn out to be the worst predictors. Importantly, the combination of all Web data features achieves better prediction than any individual feature. ρ_s increases by more than 10% compared to the best individual feature when using all the Web data features. To answer research question (5), we also see that the combined Web data features generally provide better accuracy than demographic features. Finally, combining Web data and demographic data achieves higher accuracy than either of them individually. This suggests that Web data and demographic data are good complements to each other for crime prediction. Based on the p-value, which are much smaller than 0.001 in all cases, we can tell that using Web data alone and combining them with demographic features both improve the prediction significantly.

Table 11 SVR ρ_s of individual and combined features, including demographical features

	person offense	property offense	victimless crime	theft
GSV	0.278	0.317	0.227	0.326
tfidf	0.219	0.158	0.184	0.216
LDA	0.178	0.157	0.148	0.192
GloVe	0.289	0.245	0.244	0.254
venue	0.320	0.317	0.272	0.314
check-in	0.309	0.249	0.271	0.275
Web data	0.417	0.425	0.380	0.442
demographic	0.219	0.198	0.293	0.333
combined	**0.448**	**0.440**	**0.443**	**0.466**

The bold font is used to indicate the best result

5.3.3 Comparison with block-level prediction

In order to demonstrate the advantage of street-level prediction, we compare it with the performance of block-level prediction. To make a valid comparison with street-level prediction presented, we create block-level data from street-level data and map block-level prediction to street-level prediction as follows. First we define a block size, and assign streets to the block it lays in. Then we create block-level data, with features as the average feature of all streets, and crime number as the sum of crime number of all streets. Then we run block-level 10-fold cross-validation and generate prediction for each block. After obtaining crime number prediction for each block, the crime number is assigned to streets contained in the block based on the proportion of their lengths to the total street length in the block. The features we use include 160 Web open data features and 5 demographic features. We use 0.5km × 0.5km blocks as we find that larger blocks can not generate meaningful predictions due to the number of blocks being smaller than feature count. We use SVR as the prediction method after we found that it produces much higher accuracy than LR at block-level.

Table 12 shows accuracy comparison between block-level and street-level prediction. Block-level prediction is transformed into street-level prediction using the method mentioned above, and the same measurement is applied. We can see that the street-level prediction using street-level Web open data achieves much higher accuracy than the block-level prediction using block-level data, with the RMSE reduced more than 20%. Block-level prediction is unable to produce meaningful ordering of streets, particularly for victimless crime. The best ordering it produces is for property crime, but the Spearman rank correlation coefficient is still 15% less than street-level prediction.

5.3.4 Prediction for day and night crimes

We finally investigate day crime and night crime separately. Intuitively, places where day crime are likely to happen (e.g., office building) may be different from places where night crime are likely to happen (e.g., bars and restaurants). In some application, it is also desirable to treat day and night separately, for example, in safety-based routing for pedestrians wishing to walk at night. We define a day crime as a crime that happened between 6am and 6pm, and night crime as otherwise. Thus previous crime category counts for each street are divided now into day and night crime counts. In total, we have 273,267 day crimes and

Table 12 Accuracy comparison of block-level and street-level prediction

		RMSE	MAE	SMAPE	ρ_s
person offense	block	21.68	14.32	44.08	0.167
	street	**17.23**	**10.84**	**36.37**	**0.450**
property offense	block	33.71	23.76	37.76	0.288
	street	**25.26**	**18.49**	**30.97**	**0.437**
victimless crime	block	16.47	9.512	48.61	0.043
	street	**13.29**	**7.096**	**36.49**	**0.437**
theft	block	56.51	33.86	47.39	0.123
	street	**42.39**	**24.94**	**36.61**	**0.462**

Block level prediction is convert to street-level prediction based on the length of the street

The bold font is used to indicate the best result

Table 13 SMAPE for day and night crime predictions

	person offense	property offense	victimless crime	theft
day	36.60	31.87	37.85	40.15
night	**36.37**	**30.97**	**37.73**	**38.92**

Day crime include crimes occur between 6am and 6pm, while nigh crimes include crimes occur between 6pm and 6am

The bold font is used to indicate the best result

295,584 night crimes. Running 10-fold prediction using selected features (k=20), we obtain the results shown in Tables 13 and 14. As day and night crime numbers are different, we use SMAPE instead of RMSE. From the results we can see that night crimes are easier to be predicted than day crimes as indicated by lower SMAPE and higher ρ_s in almost all the crime categories except theft. This may be due to night crimes occurring more often in places that can be represented by visual signals and text, such as bars and restaurants, and day crimes happening more often in more obscure places such as workplaces.

6 Discussions

Our study reveals several interesting insights between Web open data and street-level crimes. For example, web data is in general useful for crime prediction and can be even better than demographics data, although different types of data have different predictiveness for different crime types. Theft is most predictable using street view visual clues. Foursquare data is most suitable for predicting person offense. Victimless crime is more predictable using tweets and less predictable using GSV. Street panorama images are useful for capturing the characters of streets and complementing the street descriptions derived from user generated content. This is also confirmed by example observation that the visual clues of residential and office buildings seem to correlate with all types of crimes, while parks and open areas are associated with low crimes rates. Other interesting insights are that busy traffic conditions seems to correlate with all crime types, while attractive places with photo opportunities seem to correlate with theft. Bus stations and frequently visited laundry services appear to be correlated with high crime rates for all types of crimes. While these correlations may not present sufficient evidence to associate crime to a particular visual clue, tweet, or business, they can motivate future investigation towards revealing solid factors behind crimes, which can support decision making in urban planning.

Our study provides also a number of guidelines for future research and studies of crime prediction using Web open data listed below. Based on these guidelines, we believe that practitioners (e.g., city administrations) could build simple and low cost systems to support

Table 14 SVR ρ_s for day and night crime predictions

	person offense	property offense	victimless crime	theft
day	0.372	0.383	0.317	**0.434**
night	**0.410**	**0.444**	**0.359**	0.426

The bold font is used to indicate the best result

crime prediction and characterization. We also expect that future works aiming for crime prediction with Web open data could further improve prediction accuracy. For example, based on our investigation of the effect of individual features, we can claim that street view images can be a useful signal for crime prediction, complementing other signals. When other features are not available, or when a single data source is required, it is better to use Foursquare venue data as these provide the best prediction accuracy for all crime types. When using tweets as predictors, it is better to use word embedding as the feature transformation method, as it generally provides better prediction accuracy. Similarly, considering prediction accuracy, it is better to avoid using LDA as the feature transformation method. When it is possible to have street-level data, it is better to make street-level prediction instead of block-level or coarser granularity prediction, as the former achieves better accuracy. Since night crimes are easier to be predicted than day crimes, when one needs to consider the time of the day in the prediction, it is better to make separate predictions for night and day crimes. The choice of learning model depends on the desired measurements. Linear regression can provide better RMSE, while SVR is more effective for prediction ranking, with better Spearman rank correlation coefficients. Finally, it is not the case that more data always provide better accuracy. But a set of carefully selected features generated from Web data can effectively complement demographic information as the crime predictor.

We also acknowledge a number of limitations in this work which can be addressed in future studies. First, our study covers only a single city of San Francisco. The feature analysis and conclusions we draw may reflect some peculiar properties of the city, different from other cities. For example, as a popular tourist destination, San Francisco has many visual elements and commercial venues related to tourism, which a smaller, less touristic town may not have. Second, the data we examine are not aligned to the same period. The crime data consists of 14 years of crime records since 2003, while the tweets are collected for two years in 2016 and 2017. GSV images have been taken during various times, mostly from 2017. Foursquare venues are venues that are still open at the time of data collection, with unknown open dates. While in this work the data are considered as the geographical context of streets, and the temporal factor may not be so relevant, not aligning data period may nevertheless have some negative effects on the prediction, and future works may need to address this. Third, too few tweets make the textual representation of streets unreliable. With median number of 3, there may be too few tweets to semantically represent a street. One reason for the shortage of tweets is the small portion of geo-tagged tweets, which are less than 1% in all tweets. Future works may need to consider other localized social media sources or automatic geo-tag inference.

7 Conclusions

The emergence of fine-grained geographical Web open data has provided an opportunity to study the correlation between crime and visual, textual, and behavioral information. In this paper, we analyze three Web open data sources, namely, GSV, Twitter, and Foursquare, in relation to crimes in the city of San Francisco. Our settings is novel as the crime analysis and prediction are conducted not on a block or area level but on a street level, which is the most intuitive granularity. We extract and select features from the afore-mentioned Web open data types and study their predictiveness for different types of crimes at a street-segment level. Particularly, we are the first to propose and to investigate visual features from GSV images as a crime predictor. Our extensive experiments show the effectiveness of different feature selections as well as individual and combined feature predictiveness, among other tests.

Our work can be considered as an important step towards understanding how crime prediction systems can be enhanced using Web open data. In the future, we plan to consider more open data for analysis, such as Flickr images, Tripadvisor and Yelp reviews as well investigate other cities. We also plan to solve the online prediction issues and apply our prediction methods in real systems such as ones for safety-based route navigation.

References

Aghababaei, S., & Makrehchi, M. (2016). Mining social media content for crime prediction. In *2016 IEEE/WIC/ACM international conference on web intelligence (WI)* (pp. 526–531): IEEE.

Barker, M., Page, S.J., Meyer, D. (2002). Modeling tourism crime: the 2000 America's cup. *Annals of Tourism Research*, *29*(3), 762–782.

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

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

Chen, T., Borth, D., Darrell, T., Chang, S.F. (2014). Deepsentibank: visual sentiment concept classification with deep convolutional neural networks. arXiv:1410.8586.

Chen, X., Cho, Y., Jang, S.Y. (2015). Crime prediction using twitter sentiment and weather. In *Systems and information engineering design symposium (SIEDS), 2015* (pp. 63–68): IEEE.

De Nadai, M., Vieriu, R.L., Zen, G., Dragicevic, S., Naik, N., Caraviello, M., Hidalgo, C.A., Sebe, N., Lepri, B. (2016). Are safer looking neighborhoods more lively?: a multimodal investigation into urban life. In *Proceedings of the international multimedia conference* (pp. 1127–1135).

Diebold, F.X., & Mariano, R.S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.

Du, B., Liu, C., Zhou, W., Hou, Z., Xiong, H. (2016). Catch me if you can: detecting pickpocket suspects from large-scale transit records. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 87–96): ACM.

Eck, J.E., Clarke, R.V., Guerette, R.T. (2007). Risky facilities: crime concentration in homogeneous sets of establishments and facilities. *Crime Prevention Studies*, *21*, 225.

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

Gill, C., Wooditch, A., Weisburd, D. (2017). Testing the law of crime concentration at place in a suburban setting: implications for research and practice. *Journal of Quantitative Criminology*, *33*(3), 519–545.

Graif, C., Gladfelter, A.S., Matthews, S.A. (2014). Urban poverty and neighborhood effects on crime: incorporating spatial and network perspectives. *Sociology Compass*, *8*(9), 1140–1155.

Haklay, M., & Weber, P. (2008). OpenStreetMap: user-generated street maps. *IEEE Pervasive Computing*, *7*(4), 12–18.

Kadar, C., Iria, J., Cvijikj, I.P. (2016). Exploring foursquare-derived features for crime prediction in new york city. In *The international workshop on urban computing*.

Kang, H.W., & Kang, H.B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. *PloS one*, *12*(4), e0176244.

Khan, R., Van de Weijer, J., Khan, F.S., Muselet, D., Ducottet, C., Barat, C. (2013). Discriminative color descriptors. In *2013 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2866–2873): IEEE.

Khosla, A., Das Sarma, A., Hamid, R. (2014). What makes an image popular? In *Proceedings of the international conference on world wide web* (pp. 867–876): ACM.

Kim, J., Cha, M., Sandholm, T. (2014). SocRoutes: safe routes based on tweet sentiments. In *Proceedings of the international conference on world wide web* (pp. 179–182): ACM.

Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Liao, R., Wang, X., Li, L., Qin, Z. (2010). A novel serial crime prediction model based on bayesian learning theory. In *2010 international conference on machine learning and cybernetics (ICMLC)*, (Vol. 4 pp. 1757–1762): IEEE.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Ojala, T., Pietikainen, M., Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Peng, H., Long, F., Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238.

Pennington, J., Socher, R., Manning, C. (2014). Glove: global vectors for word representation. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1532–1543).

Ristea, A., Kurland, J., Resch, B., Leitner, M., Langford, C. (2018). Estimating the spatial distribution of crime events around a football stadium from georeferenced tweets. *ISPRS International Journal of Geo-Information*, *7*(2), 43.

Smola, A.J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Taylor, R.B., Shumaker, S.A., Gottfredson, S.D. (1985). Neighborhood-level links between physical features and local sentiments: deterioration, fear of crime, and confidence. *Journal of Architectural and Planning Research*, *2*(4), 261–275.

Utamima, A., & Djunaidy, A. (2017). Be-safe travel, a web-based geographic application to explore safe-route in an area. In *AIP conference proceedings*, (Vol. 1867 p. 020023): AIP Publishing.

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): ACM.

Wang, X., Brown, D.E., Gerber, M.S. (2012). Spatio-temporal modeling of criminal incidents using geographic, demographic, and twitter-derived information. In *Proceedings of the IEEE international conference on intelligence and security informatics* (pp. 36–41): IEEE.

Weisburd, D. (2015). The law of crime concentration and the criminology of place. *Criminology*, *53*(2), 133–157.

Wilson, J.Q., & Kelling, G.L. (1982). Broken windows. *Atlantic Monthly*, *249*(3), 29–38.

Yang, D., Heaney, T., Tonon, A., Wang, L., Cudré-Mauroux, P. (2017). Crimetelescope: crime hotspot prediction based on urban and social media data fusion. *World Wide Web*: 1–25.

Zhao, X., & Tang, J. (2017). Modeling temporal-spatial correlations for crime prediction. In *Proceedings of the international conference on information and knowledge management* (pp. 497–506).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.