Crime Mapping through Geo-Spatial Social Media Activity

Completed Research Paper

Bendler, Johannes
University of Freiburg, Chair for Information Systems Research
79098 Freiburg, Germany
johannes.bendler@is.uni-freiburg.de

Ratku, Antal
University of Freiburg, Chair for Information Systems Research
79098 Freiburg, Germany
antal.ratku@hotmail.com

Neumann, Dirk
University of Freiburg, Chair for Information Systems Research
79098 Freiburg, Germany
dirk.neumann@is.uni-freiburg.de

Abstract

The presence of crime is one of the major challenges for societies all over the World, especially in metropolitan areas. As indicated by prior research, Information Systems can contribute greatly to cope with the complex factors that influence the emergence and location of delinquencies. In this work, we combine commonly used approaches of static environmental characteristics with Social Media. We expect that blending in such dynamic information of public behavior is a valuable addition to explain and predict criminal activity. Consequently, we employ Zero-Inflated Poisson Regressions and Geographically Weighted Regressions to examine how suitable Social Media data actually is for this purpose. Our results unveil geographic variation of explanatory power throughout a metropolitan area. Furthermore, we find that Social Media works exceptionally well for description of certain crime types and thus is also likely to enhance the accuracy of delinquency prediction.

Keywords: Crime Analysis, Spatial Statistics, Social Media, Business Intelligence, Big Data
Introduction

Society faces criminal activities of individuals or syndicates everywhere. Crime from petty theft to organized cartels is one of the scourges for mankind and a central challenge for communities around the globe. Most likely, there is no neighborhood in today’s World that does not have to cope with the presence of crime and the resulting effects. Many individuals have been exposed to crimes in various forms or have been subject to an immediate situation of threat. In our societies, crime happens in many forms at any time and invokes different consequences at varying impact.

In the course of time, studies have been carried out attempting to explain the origin of criminal intent, classifying people by their ethical or demographical background, or developing methods for prevention, prediction, and ‘efficient’ punishment. Governments all over the World spend large amounts for crime prevention, law enforcement, and information of citizens. As a result, residents have noticed the increasing amount of surveillance cameras staring onto the streets of all major cities in recent years, which is sensed and discussed controversially. However, tracking of crimes is nowadays facilitated by online platforms and services that deliver admission to virtualized neighborhood watch and official crime reports. For instance, the online portals CrimeReports, Crimemapping, and SpotCrime provide public access to reported criminal incidents in the US (CrimeReports 2014; Crimemapping 2014; SpotCrime 2014). Nextdoor is a company that freely distributes a mobile application with the same name allowing residents to share observations and incidents with users from the vicinity (Nextdoor 2014). All of these exemplary named services attempt to increase safety of communities by creating awareness among residents. To some extent, information and discussion on criminal incidents move to the social web.

The social web, consisting of thousands of mobile applications and online services, yields enormous amounts of user-generated data day after day. Many people share their thoughts, feelings, and desires publicly with their friends and followers. Travelers tag their photographs with keywords and geo-locations, residents share their visit to restaurants and take photographs of their food and drinks. Newlyweds share their wedding pictures, parents create social profiles for their newborn babies. Independently from the certain content and information delivered, all these examples of social postings have one major common ground: they reveal the activity or location of users. From such Social Media data streams we can sense the heartbeat of metropolitan areas; we can determine popular places and read the city’s public social life in real-time. Even traveled routes and routine behaviors of individuals can be observed over time. Therefore, Social Media reflects social presence and activity. Various online services aim at exploiting this valuable information. For instance, location-based services and recommender systems are often built upon crowd-sourced data to provide online decision support. Services such as Foursquare work exceptionally well without relying on official data or ratings by experts (Foursquare 2014). Still, we suppose that online social services available today are far from tapping the full potential of Social Media data.

With such powerful Social Media data streams at hand, the question arises why should we not exploit the underlying potential for our own safety and security? The Social Media data analysis may eventually result in a virtual neighborhood watch that allows for a predictive policing by referring to the wisdom of crowds expressed in the Social Media feeds. Obviously, such an application contributes to a better world by increasing awareness for conditions in social presence and activity that support different crime types. This can eventually lead to a better prediction of crimes in neighborhoods.

Our starting point of analysis is that certain environmental conditions may promote or suppress different forms of crime. We premise our approach on the theory of repeat victimization, which states that places that have been victimized once have a much higher likelihood of being victimized again (Farrell and Sousa 2001). These environmental conditions, on the one hand, are related to the structural characteristics of the environment. Examples for such conditions are buildings that provide cover from sight, points of interest that are required for a certain crime type (e.g. a bank for a bank robbery), or a dedication of the urban district (e.g. residential areas is where people mostly sleep at night). On the other hand, the conditions relate to the social presence and activity. Where there are no active people, there is no petty theft, battery, or disturbance of peace. Contrastingly, the intention to break into a vehicle rather requires absence of others. These considerations already suggest a relation between the social activity and different crime types to be promoted or suppressed. Consequently, we formulate the research question for this work. Supplementary to answering the stated problem, two fundamental aspects have to be clarified beforehand. We need to assess whether Social Media can serve as a proxy for the social presence and activity in an urban
Crime Mapping through Social Media Activity

The relationship between characteristics of Social Media and criminal incidents need to be explored in depth.

- **Research Question 1**
  Can Social Media data help to better understand characteristics of crime incidents?

- **Research Question 2**
  Do the Social Media data that are recorded from the vicinity of certain points of interest mediate or increase crime incidents?

- **Research Question 3**
  Do the spatio-temporal dimensions of Social Media data explain incidents of different crime types at different times and locations?

Clarification of the research question is beneficial to the safety of communities. In the first place, the findings can create awareness of general patterns in terms of social activity that support or suppress emergence of certain crime types. Consequently, in addition to the structural characteristics of the environment, our results can increase the accuracy of hot-spot determination for various forms of crime.

The remainder of this work is structured as follows. In the first section, we provide an overview of recent work in related research areas. From the literature review, we are able to derive the research gap that is attempted by this work. Subsequently, we provide objective and outline of our research according to the identified gap in the second section. The applicability of Zero-Inflated Poisson Regression models to assess the explanatory power of Social Media is judged in the third section. The shortcomings of the proposed method in terms of geographical correlation are discussed and addressed in the fourth section by application of Geographically Weighted Regressions. The fifth section states the findings of our research. This work closes with a conclusion in the sixth section and gives an outlook on future research.

**Related Work**

This research assesses the explanatory power of Social Media data on occurrence of everyday events. With regard to coordinates in both temporal and spatial dimension, we intend to find statistical correlation between official reports of criminal incidents and user-generated data, revealing the positions of people at certain times. Our proposed approach is multi-disciplinary, as it includes the research areas of Social Media behavior and crime prediction with a link to Geographical Information Science. This section points out relevant research from the adjacent fields and identifies the research gap that is cleared in this work.

**Behavioral Patterns and Social Media**

Analysis and extraction of user mobility pattern, trajectories, and general behavior is performed in various research articles with focus on different characteristics. González et al. (2008) calculate a spatial probability distribution of humans by assessing regularities in both temporal and spatial dimension. In their work, the authors rely on spatial data gathered from mobile phone users, where the approximate location of each user can be inferred from the mobile phone towers their device is registered at. Arase et al. (2010) focus on users' trip data in order to place suggestions on travel routes based on extracted patterns. Similarly, Scellato et al. (2011) predict user locations based on patterns uncovered by non-linear time series of tracking data. The authors test their method on various data sources (e.g. GPS tracks, WiFi access points) and are able to greatly improve the prediction accuracy over spatio-temporal Markov predictors. Cho et al. (2011) show that the human mobility patterns are highly periodic, but when it comes to long-term travels partially relate to the social network of the observed individual. Backstrom et al. (2010) go even further and show that the geographical location of a person can be predicted only by the locations of individuals from the respective social network. Towards general prediction of user movement pattern, Taniar and Goh (2007) propose an approach of Apriori-like Movement Pattern and Movement Trees. Consequently, the authors are able to spot the locations a certain user visits recurrently. These periodically visited locations are sometimes referred to as the mobility profile of an individual. Ghosh et al. (2006) perform location predictions based on such mobility profiles and state that their prediction is more accurate than general statistical predictions based on location hubs.
In addition to the probabilistic location prediction of residents, the location-based recommendation is an active field of research. Location-based recommendation systems are relevant to the scope of this research, because many of them intend to understand the patterns users follow when choosing the venues they visit. In many cases based on the location history of users, online services and recommender systems are able to find and propose other locations that fit to the user’s preferences. Bao et al. (2012) employ user preferences and location histories from geo-social data in order to recommend new venues. The geographic information system proposed by Ballatore et al. (2010) yields personalized recommendations by monitoring social interaction and context. Liu et al. (2013) present a framework capable of learning a user’s check-in behavior according to points of interest. Their framework recommends venues to a user not solely based on other who have visited the same location, but additionally by incorporation of the user’s historic behavior. From the studies and frameworks that have been proposed in the sector of recommendation systems and human mobility estimation, we infer that behavior in urban environments is similar between different individuals, with large parts of it being routine.

**Spatial and Temporal Explanation and Prediction of Crimes**

The explanation and prediction of crimes is an active research area. Already in 1976, Pyle (1976) carried out analyses on spatial and temporal aspects of crime. The author regards architectural characteristics of housing units, as well as features from populations. Lately, crime prediction gained increased interest driven by the public availability of large amounts of data, covering Social Media records and official observations. The assessment of Security and Public Safety with particular focus on crime mapping and analysis has been outlined by Chen et al. (2012) as one of the emerging Business Intelligence and Analytics perspectives.

The forecasting of crimes based on the preference structure of criminals is carried out by Liu and Brown (2003). The authors build a point-pattern model based on characteristics of past crimes in order to estimate and predict the occurrence of crimes in Richmond, VA. Wang and Brown (2011) propose a generalized model to discover causality of crimes and underlying factors related to them, based on spatial, temporal, geographic, and demographic data. Furthermore, the presented approach can be utilized for crime prediction. The report by Eck et al. (2005) assesses the characteristics of crime hot spots and proposes a broad variety of tools and methods for hot spot detection and crime mapping. The authors present various theories of hot spot definition, such as mapping them to single places, to streets (or sections of streets), or even to entire neighborhoods. In subsequent chapters, techniques for better hot spot understanding, as well as statistical methods for geographical auto-correlation and dispersion measures are outlined. Furthermore, the authors provide a brief introduction into spatial analysis tools that are appropriate for crime hot spot detection, mapping, and interpretation. Bachner (2013) presents an approach for crime prevention by using Big Data and appropriate analysis tools. The reports is a collection of theory and methods, and it additionally gives actual implementation recommendations for law enforcing agencies engaged in ‘Predictive Policing’. Methods for crime prediction and analysis tools for data assessment in temporal and spatial dimensions are proposed.

In contrast to models that rely on historical crime data for forecast, Wang et al. (2012) focus on Twitter-based prediction by automatic topic extraction using Latent Dirichlet Allocation and linear modeling. The highly topical research of Gerber (2014) further investigates on content analysis of Twitter messages by implementing a combined approach of Latent Dirichlet Allocation with Kernel Density Estimation.

**Research Gap**

Comprehensive research is present in both related areas, human behavior pattern and crime explanation and prediction. However, most of the work employs contentual analysis of Social Media or actual identification of crime hot spots. To our best knowledge, no research is present that combines the public presence and activity (indicated by geo-located Twitter messages as a proxy) with the observation of criminal incidents. We aim to examine the relationship between social activity in an urban environment and location and likeliness of different crime types to emerge in proximity.
Research Objective and Outline

In this research, we assess the explanatory power of Social Media data in regard to the emergence of various crime types. Committing a crime depends on many factors that are impossible to be collected and analyzed to the full extent. Some communities may impute stereotyped ethnicities to be more likely to commit crimes, others fear criminal intent from certain demographic conditions. Furthermore, the structural environment, such as yards covered from sight, or secluded areas may have an impact on the origination of criminal incidents. And, in addition, the presence or absence of unrelated people in proximity is also likely to evoke situations that seem appealing or repellent for criminal-minded. Delinquents may also act in impulse when they face a promising opportunity, or they follow their own preference in terms of location and time of day. The actual combination of these exemplary factors is complex and possibly unfeasible to be determined. However, the human mind is in fact able to assess the exposure of many situations, which can result in avoidance of certain districts at certain times. Thus, a key factor towards estimation of the risk potential at given environmental conditions is the behavior of community members. With Social Media as a proxy for social presence in an urban setting, we aim to further explore the conditions that drive emergence of crime.

In the first place, we assess the occurrence of a certain event (e.g. a criminal incident) by presence of certain environmental conditions. For this purpose, we join data from different sources to explore dependencies in the geographical and temporal dimension. Figure 1 depicts the process outline relevant to our approach. The first data source (‘Environment’) describes the spatial conditions in terms of points of interest. The second data source (‘Social Activity’) covers the online activity in geo-located social networks, such as Twitter, Facebook, or Foursquare. The third data source (‘Event Observation’) relates to the crime data we seek to explain. Combining these three data sources enables us to assess their dependencies in a geo-spatial context using current techniques from the area of geographic information systems. Consequently, we aim to explain the occurrences of observed events based on spatial and temporal characteristics.

Data Source Characteristics

The most crucial aspect towards interpretable analyses is reliable data from authentic and credible sources. Our dataset contains events represented as point data, mapping each observation or venue to a certain geographic location. We rely on points of interest to represent the environmental data. The collection covers venues in more than 90 different categories, for instance restaurants, bars, banks, parks, or churches. Social activity is assessed by more than 600,000 geo-tagged Twitter status messages, each of them providing both a precise geographic location and an exact point in time. The observed events are more than 32,000 criminal incidents in 14 different categories, such as burglary, theft, fraud, or disturbing the peace. All data at hand refers to the months August through October in 2013 and is related to the city of San Francisco.

Environment Data

The set of Environmental Data covers information on structural characteristics of environments. Beside others, it represents public places, locations of shops and services, and landmarks. Consisting of many different kinds of venues, this data source is implicitly able to represent the major purpose of city districts. For instance, from an increased density of shops, we can infer that the corresponding area is likely to be a pedestrian zone or a shopping mall. Restaurants, and shops in proximity to landmarks can be tourist hot spots, and playgrounds along with a low density of other points of interest may suggest a residential area.
Since these spots and establishments are rather unlikely to change at a high frequency, we only rely on the associated geo-spatial information and omit the temporal dimension of this data source.

For our analyses, we choose to rely on point of interests from Google's Maps API, since they provide a comprehensive source of data. Furthermore, the fact that these records are audited by employees greatly reduces the likeliness to find uncertainty in geo-spatial mapping or maliciously modified data. However, the 91 categories that are used for classification of points of interest are not disjoint. Since single venues can be assigned to multiple categories at the same time, these categories may show a considerable degree of multicollinearity. The presence of multicollinearity can result in wrong estimations in statistical analyses. Contrastingly, multicollinearity may as well provide information relevant to the outcome of calculations, especially in geo-spatial analyses. According to O'Brien (2007), attempts to completely remove multicollinearity in regression models may even result in more damage than was originally caused. In order to deal with this two-faced complex of problems, we only remove obvious multicollinearity between categories of points of interest and then aggregate them to only few disjoint major categories, all of which represent a certain dedication. The panels (a) – (g) of Figure 2 show the resulting seven major categories and their geographical distribution across the city of San Francisco.

![Figure 2. Spatial Distribution of Seven Major Categories of Points of Interest](image)

Social Activity Data

The quick and intense rise of Social Media platforms and services has led to immense amounts of user-generated data being publicly available. There is no data source that better reflects the interests, activities, thoughts, and desires of an entire community. Nevertheless, the frequency, quality, consistency, and especially availability of data varies from platform to platform. While all Social Media platforms together surely reflect the society’s characteristics, collecting and aggregating their data is a challenging task. The Social Activity Data in our research approach thus initially only relies on one data source, that provides essential information. We choose Twitter status messages for our analyses, because their exact time of publication can be retraced and a share of published messages also provides an accurate geographical location. Social Media data that provides coordinates in both spatial and temporal dimensions is superior since behavior and motion of masses can be followed without contentual evaluation. Regarding all Twitter messages from one city that are published within a single hour provides a snapshot of the people and their activities. If we consider several subsequent snapshots, we can unveil the paths of people streaming through the urban environment. These observations allow calculating the density of people in certain areas at
specific times. Furthermore, the spatio-temporal monitoring enables calculation of patterns by comparing a specific hour of day with the corresponding hours in subsequent days and weeks. Thus, we can record the routine behavior of individuals and crowds. This vast potential allows us to determine areas that are popular or, contrastingly, mostly avoided based on hour of day and day of week. The small fraction of geo-tagged Twitter status messages, which makes roughly one per cent of all Twitter messages published, can be sufficient to represent social presence and activity.

Even though Twitter messages contain a lot more information, we only rely on the geographical location and the point in time they were published. We interpret Twitter messages as a beacon that informs us about the fact that someone is at a certain location at the current time. Receiving and recording these signals over time, we can aggregate them by hour of day and day of week and so create a stable Social Media basis for our analyses. The plots given in Figure 3 show the cumulative Twitter activity for 1 a.m. – 2 a.m. (panel (a)) and 9 p.m. – 10 p.m. (panel (b)) during the observation period in the city area of San Francisco.

![Figure 3. Spatial Distribution of Cumulative Twitter Activity for Different Time Windows](image)

**Event Observation Data**

Many different everyday occurrences can serve as Event Observation Data to be explained by our proposed method. A broad variety of different observations can be extracted from Social Media platforms, most of which are societal events. For instance, planned weddings or other ceremonies, visitation of venues, and holiday trips can be monitored. Other online services collect and provide housing offers and prices or openings of new venues. In the past few years, an increasing amount of authorities initiated online platforms that provide access to official data and statistics. For example, the city of San Francisco launched an own website (https://data.sfgov.org/) that is dedicated to providing data collections, charts, maps, and other valuable information about the city.

![Figure 4. Spatial Distribution of Different Crimes](image)
In our belief both explanation and possibly resulting reduction of criminal incidents is highly important and valuable to society. In order to efficiently explain and investigate crime data, we need to rely on both spatial and temporal coordinates. Each crime observation is accurately located in time and space. Furthermore, the visual inspection of Figure 4 provides an idea of the geo-spatial variation depending on the certain crime type. The observed categories of crime are Assault, Burglary, Disturbing the Peace, Drugs/Alcohol Violations, Driving Under the Influence, Fraud, Homicide, Motor Vehicle Theft, Robbery, Sex Crimes, Theft/Larceny, Vandalism, Vehicle Break-In/Theft, and Weapons.

The Explanatory Power of Social Media

This research intends to explore what we can learn from Social Media activity when explaining the emergence of crimes. Therefore, we assume Twitter media to represent the public presence and activity to some extent. Hence, we expect Twitter messages to show a significant explanatory value for several, but not for all observed crime types.

For our analyses, we divide the city area of San Francisco into a grid of 50 by 50 cells. The collected data is assigned to geographically corresponding cells, each representing a bin to collect and count the observations. Concerning Twitter and crime data, which provide an additional temporal component, we expect pattern to occur on a weekly basis. For example, disturbing the peace may be more present on Friday and Saturday evenings, while fraud probably takes place during daytime on workdays. Similarly, we expect people to show deviation in Twitter usage behavior depending on time of day and day of week, as well. Since we seek to explore everyday occurrences instead of single large events, these considerations allow us to aggregate temporally dependent data to a shorter time period. The time span of one week split into hourly slots yields $24 \times 7 = 168$ temporal bins. Point of interest data is assumed not to change over time and thus remains the same in every time slot for each individual cell. With 168 time slots and the grid of 50 by 50 cells, we end up with $50 \times 50 \times 168 = 420,000$ observations that are available for regression analysis.

As a preparation for our regression analysis and with regard to expected temporal patterns, we are required to introduce dummy variables for the temporal dimension. Adding a dummy for every hour would probably result in overestimation, and logical links between time slots would get lost. Thus, we would like to differentiate between time-slots that correspond to a working day’s daytime, a weekend, a night, and a weekend-night. Using the regular daytime as the default assumption, we only need to include three temporal dummy variables into our model, representing weekend, night, and weekend-night. All in all, we end up in eleven explanatory variables, containing seven categories of points of interest, three temporal dummies, and the amount of Twitter messages.

Zero-Inflated Poisson Regression

The nature of the crime data at hand suggests implementation of a Poisson regression, because the amount of crimes in a grid cell at a certain time can only take non-negative integer values. Moreover, owed to the short observation period of no more than three months, the maximum count of the same crime type in one cell was 8. Thus, our dataset is naturally restricted to small positive integer values, which makes a Poisson regression the optimal choice.

Furthermore, we detected an excess amount of zeroes in the cells regarding our data on crime incidents. To some extent, this can surely be explained by the geographical characteristics of the area, because some of the grid cells map areas of nothing else but water. Even though most cells cover ground area, the geographical factors can render emergence of a criminal incidents impossible. Additionally, some grid cells do not contain any crime at all, though their geographical and structural characteristics in general would permit it. In order to account for this problem, we decide to rely on a Zero-Inflated Poisson Regression, which on the one hand employs a Logit Regression to deal with the excess zeroes, and on the other hand applies a Poisson Regression to estimate the effects of the independent, explanatory variables on the dependent variable.

The regression analysis as described in this section basically consists of four main steps. We align our procedure at the flow proposed by UCLA: Statistical Consulting Group (2014). First of all, we run a Zero-Inflated Poisson Regression for every crime type separately. The second step is to compare the resulting models to so-called null-models (i.e. regressing the dependent variable only on an intercept). Third, we
carry out Vuong’s closeness test to assist in comparison of the zero-inflated models with an ordinary Poisson regression. Fourth and finally, our models are compared to adequate regressions that do not contain Tweets as independent variable.

**Step 1.** The Zero-Inflated Poisson Regression is carried out for each of the fourteen different crime types using the same explanatory variables, as shown in Eq. 1. Besides the intercept $\beta_0$, we include the amount of Tweets and points of interest from each of the seven major groups into our regression model. Additionally, we add three temporal dummy variables as discussed before, namely weekend, night, and weekend-night.

$$c \sim \beta_0 + \beta_1 \cdot \text{Tweets} + \beta_2 \cdot \text{Retail\&Services} + \beta_3 \cdot \text{Food\&Drinks} + \beta_4 \cdot \text{Social\&Religion} + \beta_5 \cdot \text{Authorities} + \beta_6 \cdot \text{Finance\&Law} + \beta_7 \cdot \text{Transportation} + \beta_8 \cdot \text{Entertainment} + \beta_9 \cdot \text{Weekend} + \beta_{10} \cdot \text{Night} + \beta_{11} \cdot \text{WeekendNight}$$

with $c$ the crime type

( Assault, Burglary, Driving Under the Influence, Disturbing the Peace, Drugs/Alcohol Violations, Fraud, Homicide, Motor Vehicle Theft, Robbery, Sex Crimes, Theft/Larceny, Vandalism, Vehicle Break-In/ Theft, Weapons)

$\beta_i$ the parameter estimates

**Step 2.** Null-model tests are performed in order to find out whether the Zero-Inflated Poisson Regression model is significant as a whole. Thus, comparing our proposed models to the corresponding null-models, we are able to state whether our method actually explains something that is not intrinsic to the data itself. In regression analyses, null-models are obtained by regressing the dependent variable only on an intercept, omitting all explanatory variables. In almost all cases, our proposed model exhibits statistical significance on a 0.1 per cent level. Efficiently, this means that our crime data is better explained by the Zero-Inflated Poisson Regression that contains Twitter data and points of interest, than by the model consisting only of the intercept. The only exception is given by the crime type ‘Homicide’, which shows no statistical significance. This is most likely caused by the (fortunately) very low amount of homicide incidents in the observation period (our data shows 9 homicides).

**Step 3.** Vuong’s closeness test is a method for statistical model selection that compares two models by their probabilistic distance to the actual model (Vuong 1989). This test is applicable and quite common when comparing zero-inflated regression models to ordinary regressions. The result gives evidence whether the two compared models are equally close to the actual one or not (i.e. the null hypothesis is that the models are indistinguishable). In all of our cases, except for homicide, Vuong test results show that the zero-inflated model is closer to the actual model than the ordinary Poisson Regression model. Discussion of test results in detail is omitted here, because all positive results are highly significant on a 0.1 per cent level. These results confirm the selection of a zero-inflated model over an ordinary regression given our data and setup.

**Step 4.** Finally, we need to ascertain whether application of the Twitter data as an independent variable increases the overall performance of the model. Consequently, we compare our proposed models that include Twitter as explanatory variable to equal models (i.e. using the same data set and the same variables) that omit the Twitter variable. From the results, we can examine whether Twitter supports explaining the occurrence of crimes for each crime type individually. In order to compare two models to each other, we rely on the results from Vuong tests, as well as the commonly known goodness-of-fit measures AIC and BIC. In contrast to the Vuong test that measures the probabilistic distance between the tested models and the actual one, AIC and BIC rather show which of the two models is preferable with respect to the tradeoff between goodness-of-fit and model complexity.

The results of our Zero-Inflated Poisson Regression models including Twitter are provided in Table 1. For each crime type, the Twitter parameter estimate is given along with the corresponding z-statistics and significance level. Furthermore, the last column shows whether a Zero-Inflated Poisson Regression model that includes Twitter is superior to one that does not, according to Vuong test results, AIC, and BIC comparison. The recommendation to use Twitter is only given if all three measures evaluate the Twitter model to be preferable. Concerning the Vuong test, we require the Twitter model to be closer to the actual model on 5 per cent significance level. The results of our comparison using AIC, BIC, and Vuong test show
that Twitter is valuable in explaining Burglary, Motor Vehicle Theft, Robbery, and Theft/Larceny (highlighted rows).

The first thing to notice is the variation in the Twitter parameter estimate among different crime types. We can observe Twitter to be highly significant for some of the crime types while not being significant at all for various others. Since our regression analysis only relies on geographic and temporal coordinates of Tweets instead of the written content, we can interpret the significant positive or negative influence of Twitter on the respective crime as being related to the public presence. We do not suspect people to send Tweets because of the crime, but rather argue that in case of many people in the streets, some crime types are more or less likely to happen due to the corresponding public attention. Accordingly, crimes where Twitter does not show a significant influence thus probably happen mostly independent from public presence and perception. Burglary, for example, requires mostly absence of public perception and thus shows a significant negative influence of Twitter. Contrastingly, Motor Vehicle Theft seems to be least noticeable if many people are around and nobody has any suspicion.

<table>
<thead>
<tr>
<th>Crime Type</th>
<th>Twitter Parameter Estimate</th>
<th>Use of Twitter Recommended for Explanation?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assault</td>
<td>0.0028 * (2.188)</td>
<td>—</td>
</tr>
<tr>
<td>Burglary</td>
<td>-0.0079 ** (-3.215)</td>
<td>✓</td>
</tr>
<tr>
<td>Disturbing the Peace</td>
<td>-0.0014 (-1.375)</td>
<td>—</td>
</tr>
<tr>
<td>Drugs/Alcohol Violations</td>
<td>0.0043 * (2.073)</td>
<td>—</td>
</tr>
<tr>
<td>Driving Under the Influence</td>
<td>0.0050 (1.333)</td>
<td>—</td>
</tr>
<tr>
<td>Fraud</td>
<td>-0.0050 (-1.670)</td>
<td>—</td>
</tr>
<tr>
<td>Homicide</td>
<td>0.5585 (1.454)</td>
<td>—</td>
</tr>
<tr>
<td>Motor Vehicle Theft</td>
<td>-0.0162 *** (-4.679)</td>
<td>✓</td>
</tr>
<tr>
<td>Robbery</td>
<td>0.0065 *** (3.922)</td>
<td>✓</td>
</tr>
<tr>
<td>Sex Crimes</td>
<td>-0.0073 * (-2.096)</td>
<td>—</td>
</tr>
<tr>
<td>Theft/Larceny</td>
<td>0.0062 *** (6.233)</td>
<td>✓</td>
</tr>
<tr>
<td>Vandalism</td>
<td>0.0004 (0.193)</td>
<td>—</td>
</tr>
<tr>
<td>Vehicle Break-In/Theft</td>
<td>0.0011 (0.883)</td>
<td>—</td>
</tr>
<tr>
<td>Weapons</td>
<td>0.0017 (0.746)</td>
<td>—</td>
</tr>
</tbody>
</table>

Stated: MLE coefficients, z-statistics in parentheses; Comparison between models based on Vuong test on 5% significance level, AIC, and BIC; Observations: 420000 Significance: * 0.05 ** 0.01 *** 0.001
From these preliminary results, we suggest further investigation of the crime types Motor Vehicle Theft, Theft/Larceny and Driving under the Influence. We select these models, because each of them represents a different impact of the Twitter parameter estimate. For Motor Vehicle Theft and Theft/Larceny, the Twitter coefficient is statistically highly significant; for the former negatively and for the latter positively. Regarding the crime type Driving under the Influence, Twitter data shows no statistical significance. Table 2 summarizes the Poisson Regression results of the Zero-Inflated Regression model for these three crime categories and allows for interpretation of the coefficient estimates of the independent variables.

First of all, even though we find many highly significant variables for the three selected models, we have to be cautious when interpreting the results shown in Table 2. From the independent variables we can only be certain about the results of Twitter and the three temporal dummies Weekend, Night, and WeekendNight. The seven point of interest groups show significance to some extent, but they might be geographically correlated and thus require us to pay attention in interpretation.

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Motor Vehicle Theft</th>
<th>Theft/Larceny</th>
<th>Driving Under the Influence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without Twitter</td>
<td>With Twitter</td>
<td>Without Twitter</td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.4363 ***</td>
<td>-1.4363 ***</td>
<td>-1.4397 ***</td>
</tr>
<tr>
<td></td>
<td>(-26.348)</td>
<td>(-26.263)</td>
<td>(-26.294)</td>
</tr>
<tr>
<td>Twitter</td>
<td></td>
<td>-0.0162 ***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-4.679)</td>
<td></td>
</tr>
<tr>
<td>Weekend</td>
<td>0.0033 (0.068)</td>
<td>0.0050 (0.105)</td>
<td>-0.1395 *** (3.735)</td>
</tr>
<tr>
<td>Night</td>
<td>-0.7195 *** (10.893)</td>
<td>-0.6760 *** (10.148)</td>
<td>-0.7848 *** (16.079)</td>
</tr>
<tr>
<td>WeekendNight</td>
<td>-0.1910 * (-1.972)</td>
<td>-0.1816 * (-1.874)</td>
<td>-0.2013 * (2.664)</td>
</tr>
<tr>
<td>Retail &amp; Services</td>
<td>-0.0092 ** (-3.268)</td>
<td>-0.0089 ** (-3.115)</td>
<td>0.0067 *** (6.197)</td>
</tr>
<tr>
<td>Food &amp; Drinks</td>
<td>-0.0082 ** (-2.685)</td>
<td>-0.0062 * (-2.048)</td>
<td>0.0100 *** (6.913)</td>
</tr>
<tr>
<td>Social &amp; Religion</td>
<td>-0.0150 *** (-3.524)</td>
<td>-0.0151 *** (-3.562)</td>
<td>0.0041 (1.728)</td>
</tr>
<tr>
<td>Authorities</td>
<td>0.0135 (1.201)</td>
<td>0.0168 (1.501)</td>
<td>0.0398 *** (5.325)</td>
</tr>
<tr>
<td>Finance &amp; Law</td>
<td>0.0021 (0.739)</td>
<td>0.0017 (0.586)</td>
<td>-0.0060 *** (-4.421)</td>
</tr>
<tr>
<td>Transportation</td>
<td>-0.0300 ** (-2.578)</td>
<td>-0.0268 * (-2.291)</td>
<td>0.0140 * (1.968)</td>
</tr>
<tr>
<td>Entertainment</td>
<td>-0.0097 (-1.189)</td>
<td>-0.0018 (-0.216)</td>
<td>0.0079 * (2.273)</td>
</tr>
<tr>
<td>Pseudo R² (McFadden)</td>
<td>0.4695 0.4704 0.5216</td>
<td>0.5225</td>
<td>0.3819</td>
</tr>
<tr>
<td>AIC</td>
<td>15674 15648 22280 22242</td>
<td>4666</td>
<td>4666</td>
</tr>
<tr>
<td>BIC</td>
<td>15710.2325 15687.8557 22316.2325 22281.8557</td>
<td>4702.2325</td>
<td>4705.8557</td>
</tr>
</tbody>
</table>

Stated: MLE coefficients, z-statistics in parentheses; Observations: 420000 Significance: * 0.05 ** 0.01 *** 0.001

Crime Mapping through Social Media Activity
Comparing the Twitter parameter estimates for the three selected crime types Motor Vehicle Theft, Theft/Larceny, and Driving under the Influence, our expectations posed at the outset are statistically confirmed. The occurrence of Motor Vehicle Theft highly corresponds to the absence of Twitter messages, effectively meaning that vehicles are mainly stolen when no public presence and activity is given in proximity. The exact opposite is unveiled for Theft/Larceny crimes. They are more likely to happen where many people are present. This confirms what many police departments alert: where many people are crowded, for example in city centers or on annual fairs, pickpockets are likely to take the opportunity. The Twitter parameter estimate for Driving under the Influence (DUI) is not statistically significant, ultimately meaning that people are caught driving under influence of drugs independently from the social activity hot spots. Besides the Twitter statistics, the results for the Night dummy are very interesting. Its values reveal that most thefts are performed during daytime on working days (both vehicle thefts, as well as regular theft and larceny), while DUI can mostly be observed in the night hours.

The comparative values of McFadden Pseudo R², AIC, and BIC shown in the lower rows of Table 2 state the difference between each model with and without Twitter as an explanatory variable. For Motor Vehicle Theft and Theft/Larceny we can observe an increase in the Pseudo R². Even though the McFadden Pseudo R² does not explicitly state the exact degree explanation in per cent as other R² values do, an increase also suggests an increase in the explanatory power. However, this increase can also be due to the inclusion of one more explanatory variable. Therefore, we also perform the AIC and BIC tests. For both Motor Vehicle Theft and Theft/Larceny, AIC and BIC are reduced by incorporation of Twitter, which indicates improvement of the model. Regarding Driving under the Influence, the Pseudo R² is slightly increased, but AIC and BIC reveal that the proposed Twitter model is not stronger than the model without Twitter.

**Limitations of the Method**

The applied method of Zero-Inflated Poisson Regression clearly indicates a global relationship between Twitter activity and crimes for several crime types. Especially the temporal dependency between these two measures can be assessed by the proposed method. Nonetheless, the methodology is tainted with the major drawback that adjacent cells are disregarded when regressing the observations of a single area. A general regression analysis does not take spatial dependency into account.

Spatial dependency denotes the characteristic of values from adjacent regions to appear correlated. This correlation inevitably yields multi-collinearity in general statistical analyses, such as regressions. With spatial dependent data, regression results may get biased and parameter estimates can become unreliable. However, variables that show significant impact in a regular regression would remain significant even if multi-collinearity was entirely removed, rendering the parameter estimates comparable and interpretable again. In the context of geographical analyses, multicollinearity cannot be avoided that easily, since it is part of the data itself. Especially geo-spatial data from urban areas almost always contains spatial dependency, because it reflects the city as such. This fact is owed to the nature of cities and man-made structures, which always tend to contain intrinsic geographical clustering. Establishments of various kinds are located in proximity of each other. For example, shoe stores and clothing stores can often be found in the same areas, such as shopping malls or pedestrian zones. This observation holds for many other points of interest as well, for instance banks and ATMs, or restaurants, cafes, and other venues dedicated to food and drinks, which are likely to appear in contiguity. The same problem can be observed among different data sources. Where there is an increased amount of points of interest, the amount of posted Twitter messages is also likely to be above average.

Moran (1950) formulated one of the first and most common measures for spatial autocorrelation. Moran’s Index I is a global indicator of spatial correlation, which can be employed to assess the clustering likelihood of data in a lattice. Concerning our analyses, Moran’s I would be applicable to each single data source and can yield an estimate on how geographically dispersed the observations are. But, due to spatial dependency resulting from the structure of cities, it obviously yields similar values for every data source. In this special case, Moran’s I will not help in assessing correlation within data, because similar spatial densities are likely to occur between different data sources, not only among observations of a single source.

Consequently, we are not able to remove spatial dependency from our data, neither among different data sources, nor within each separate source. The geographical correlation is an essential part of our data and to some extent drives the explanatory power of the methodology. It carries the information we seek to
analyze and therefore must not be removed in advance. Instead, we are required to employ an analysis method that is able to compensate spatial dependency by design. One of such methods are Geographically Weighted Regressions.

**Exploration of Geo-Spatial Correlations**

The interpretation of the Zero-Inflated Poisson Regression uncovered a major drawback of discovering relationship between Social Media data and criminal incidents from a geographically global perspective. Though we actually find a significant influence of Twitter activity, it remains uncertain whether this observation is mainly driven by a small geographic area or is valid over the entire grid. Further analyses are required that incorporate the spatial proximity, ultimately resulting in geographically varying significance.

An approach that is capable of handling geographical dependencies is proposed by Brunsdon et al. (1998) and called Geographically Weighted Regression (GWR). We employ a GWR to our data in order to test for spatial non-stationarity and deepen our analyses in the course of this section.

**Geographically Weighted Regression**

The methodological concept of Geographically Weighted Regressions is designed to explore non-stationarity in geographic parameters. In general, a GWR estimates a regular regression, such as OLS, for each spatial feature (i.e. for every single grid cell). Depending on a chosen bandwidth or amount of nearest neighbors, the estimates are then put into spatial relationship by employing a spatial kernel function (which can be, for example, a Gaussian). The method is designed to handle the exact difficulty we face in this research. Even though the approach sounds promising, it requires cautiousness. Recent studies have unveiled its weaknesses of increased amount of false-positives and faulty recognition of spatial non-stationary (Páez et al. 2011). Hence, a fitted model is usually required before a GWR is performed, which in turn only explores spatial non-stationarity of the already identified parameters.

The coefficient estimates from the Geographically Weighted Regression generally confirm the results of the Zero-Inflated Poisson Regressions for tested crime types. Figure 5 depicts the geographically varying coefficients for the three crime types (a) Motor Vehicle Theft, (b) Theft/Larceny, and (c) Driving under the Influence. The crime types corresponding to panels (a) and (b) show statistical significance of Twitter in the Poisson Regression run, (a) with negative sign, (b) with a positive one. DUI, shown in panel (c), cannot be significantly explained by Twitter. All three plot panels illustrate quartile plots of the corresponding Twitter estimate resulting from the Geographically Weighted Regression. The plots’ colors relate to the absolute value of the estimates. Thus, lower negative values in panel (a) yield a more intense color, while on panel

![Figure 5. Geographically Varying Twitter Coefficient Estimates for Three Different Crime Types](image)
(b) and (c) colors are increasingly intense with positively ascending values. This facilitates determination of areas where the explanatory value of Twitter is high.

Besides the geographical variation among grid cells that cover ground area, also water cells appear to sometimes yield coefficients from the upmost quartile. Due to the fact that no crimes and only very few Tweets are recorded on water cells, the Geographically Weighted Regression pushes estimates to neighboring cells based on the selected bandwidth. Hence, upper-quartile estimates for cells without observations may be generated by fault. Regarding the cells that cover ground area, we can clearly identify districts where the explanatory impact of Twitter is high for the crime types shown in the three panels (a)-(c) in Figure 5. As indicated by the legend, quartiles for (c) are close to each other and generally closer to zero, ultimately illustrating the random explanatory power of Twitter on DUI crimes. For panels (a) and (b), dark areas outline districts where the amount of Tweets is significantly tied to the committed crimes.

**Findings**

In this research, we examine whether publicly available Social Media data can be employed to better understand and explain the location of certain types of criminal incidents. Our Zero-Inflated Poisson Regressions indicate that analyzing the spatio-temporal intensity of Social Media usage is valuable in explaining Burglary, Motor Vehicle Theft, Robbery, and Theft/Larceny. For the other tested crime types the Zero-Inflated Poisson Regression models did not show an improvement when including Social Media measures. These findings coincide with our expectations about the occurrence of the different delinquencies. Initially, we suggested that Robbery and Theft/Larceny are more likely to occur in proximity of increased social activity, while Burglary and Motor Vehicle Theft are less frequent in such areas and times. We can confirm our expectation by examination of the coefficient estimate that corresponds to the Social Media variable for these four crime types. For those delinquencies where the amount of Twitter messages did not prove to be significant, we can deduce that the criminal events are not closely related to social presence and activity in the vicinity. For example, Driving under the Influence is likely to occur uniformly distributed throughout the city instead of cumulating in socially active areas. Hence, our results clearly indicate that Social Media reflects the social presence within a city and is moreover a valuable factor to explain the spatio-temporal appearance of different crime types on city level.

In order to account for the shortcomings of a Zero-Inflated Poisson Regression in analyzing geo-spatial relations, we additionally execute Geographically Weighted Regressions. These regressions indicate varying coefficient estimates of Social Media across the city area. From the corresponding results, we can argue that places with a high Social Media activity yield considerably negative Twitter coefficient estimates for Motor Vehicle Theft and positive estimates for Theft/Larceny. No such connection can be established for Driving under the Influence. These findings are consistent with the Zero-Inflated Poisson Regression results, but as the Geographically Weighted Regression also accounts for the geo-spatial dependencies, the corresponding findings may help to explain the occurrence of crimes more precisely.

Consequently, our findings indicate that Social Media data should be used to better understand the spatio-temporal conditions for crimes to occur. It is most valuable in explaining those types of crime that are closely connected to social presence and activity, either in a positive or negative direction.

**Conclusion**

In our societies, crime happens in many forms at any time and invokes different societal consequences at varying impact. Studies were carried out that were targeted at explanation, estimation, and prediction of criminal incidents. We suggest that the vast potential inherent to Social Media data can be exploited to explain the emergence of crime, ultimately resulting in higher awareness and eventually even in reduction of criminal incidents or situations of actual threat. In this research, we propose a methodology of data calculation, preparation, and analysis using two state-of-the-art techniques. The Zero-Inflated Poisson Regression serves well for assessment of the actual explanatory value of Twitter data in urban environments, whereas the Geographically Weighted Regression gives evidence on the spatial dependencies and variations. Both methods have their limitations according to our posed challenge, but regarding their certain domains, each of them also yields valuable and interpretable results. The insights learned from this research work open the door for a broad area of subsequent work. Crime is a common problem to society
and our findings propose a path for better explanation and understanding, at some point eventually resulting in actual reduction of incidents.

With respect to the first research question, we state that Social Media is a valuable addition when investigating the environmental and structural characteristics of crime spots. Even though the statistical significance varies among different crime types, Social Media data increases the explanatory power of the corresponding models. According to the second and third research question, we find evidence that different crime types are best explained by different sets of variables. Potentially, a separate table could be designed for each type of crime in order to further examine the different conditions. Some crimes are most likely to occur in densely populated areas (e.g. Theft/Larceny), while other crimes are most likely to occur in rather secluded areas (e.g. Motor Vehicle Theft). The presence of points of interest thus is likely to be a strong additional predictor of Theft/Larceny, but not of Motor Vehicle Theft. Our findings can greatly support the accuracy of hot spot determination and thus help in crime prevention by increasing awareness for conditions in social presence and activity that support emergence of various crime types.

Our work is mainly limited by the shortcomings of the methodology. The Zero–Inflated Poisson Regression does not account for spatial dependency among observations, while the Geographically Weighted Regression does not yet provide a Zero–Inflated Poisson model. Especially the GWR is a constantly developed tool and additional regression methods are added frequently. Besides the methodological drawbacks, our research is limited by the relatively short observation period. Twitter messages and crime records originate from a three-month period (August through September, 2013). We have shown evidence of a stable spatio-temporal relationship between Social Media and criminal incidents, but to obtain more reliable and robust results, another data set should be consulted.

In future research, we aim to prepare the data using Kernel Density Estimation and Inverse Distance Weighting in order to smoothen out outliers and to represent spatial conditions more accurately. Even though the grid-based approach works well and yields good results, different area boundaries should be regarded. We plan to add administrative districts as dummy variables to both the Zero-Inflated Poisson Regression, and the Geographically Weighted Regression to account for their potential impact. Furthermore, since both crime data and Twitter data contain a temporal component, we aim to perform Geographically Weighted Panel Regressions as proposed by Bruna and Yu (2013). Combined with research from the area of Risk Terrain Mapping, our approach of blending in Social Media data can greatly improve the results and probably even lead to stronger predictors. Moreover, we will conduct textual analysis (e.g. sentiment analysis and automatic topic extraction using Latent Dirichlet Allocation) to further refine the explanatory and predictive power of Social Media data.

References


