Integration of Social Media in Spatial Crime Analysis and Prediction Models for Events

Alina Ristea
University of Salzburg,
Schillerstraße 30
Salzburg, Austria
mihaela.ristea@stud.sbg.ac.at

Michael Leitner
Louisiana State University,
E-104 Howe-Russell-Kniffen
Geoscience Complex,
Baton Rouge, LA, USA
mleitne@lsu.edu

Abstract

The last decade has been the most productive in respect to social media data exploration and possible uses in crime prediction. This area is thus a rapidly evolving and growing field. This PhD research aims to find and evaluate spatial relationships between crime occurrences and nearby social media activity for events areas and estimating the possible influence of this activity for crime prediction models. Overall, the thesis will focus on geospatial crime prediction concerning planned and emerging events through the exploration of social media data, and other information including demographic, economic and safety risk factors.

The thesis will utilize methods and tools from various fields including: social media text mining and classification from machine learning; spatial statistics together with forecasting models from crime prediction. Outcomes will be a valuable basis for defining new research areas, helping to understand further spatial crime analysis and prediction models that include secondary data sources, such as social media, on the basis of event exploration.

Keywords: spatial crime analysis, social media, spatial prediction, crowd based events.

1 Introduction

To date, crime prediction models in conjunction with social media data have been able to achieve a significantly high rate of success, for certain types of crime, complementing traditional crime prediction models (Corso, 2015; Gerber, 2014; Wang & Gerber, 2015; Wang et al, 2012).

Most of the crime prediction techniques are used for crime retrospective forecasting, which consider the existence of historical crime data. For this approach, quantitative methods were developed to categorize crime data in objective ways and to find characteristics such as the type of crime, typology of offender, result of investigation, confidential information using geospatial and statistical techniques, such as hot spot analysis (Eck et al, 2005), regression, cluster determination or spatiotemporal pattern recognition.

In recent period the crime predictive analytics are getting more interdisciplinary. This is also related to the "big data" growth, the last decade being the most productive in respect to social media data exploration. Researchers from informatics, computer science, mathematics and statistics are collaborating with criminologists, sociologists and others in developing new prediction models. Moreover, the high evolution of the technology is being a very important process in crime analytics as well as in social media and it opens up a plethora of research that can be done in different fields of interest.

Machine learning techniques together with linear and logistic modeling (Alruily, 2012; Burnap & Williams, 2015; Wang & Gerber, 2015; Wang et al, 2012), density based models (Bendler et al, 2014a; Cheng & Smyth, 2015; Featherstone, 2013a; b), risk terrain modeling (Perry, 2013) or Geographically Weighted Regression (Bendler et al, 2014b) have been used to predict crime occurrences using geotagged tweets or, in more detail, text mining from tweets. The algorithms have highly ranked results; however there are not many explanations about why the accuracy is changing for different crime or social media datasets. As for our knowledge, very few previous works are considering the effect of events on spatial crime distribution while using social media in prediction.

There is an important body of literature focusing on spatial crime distribution from the events mirror and on social media during events, such as big or mega events, sporting events, natural disasters. However, not so much research attempt has been done before specifically for predicting planned events considering social media and crime data, at a specific location or at a venue spot and also including environmental explanatory variables

in the models. Population trajectories and their impact on crime likelihood are different according to the environmental factors.

Finding attributes from social media that can give a boost in crime prediction models and their implementation along with the crime data for a better prediction is the core part of the PhD, with a main focus on public events. Three main elements are the base of this PhD research: crime occurrences, social media (mostly Twitter data) and events (planned events and emerging events).

An event can be defined as a matter that happens in a place, especially one of importance, such as a planned public and social occasion or particular contests making up a sports competition. The planned events are the ones for which their main parameters are defined, such as the location or the public attendance. The emerging events refer to the ones from which basic elements have the ability to develop novel relations and identities designed into higher-level elements.

Overall, the spatiotemporal analysis is the base of this PhD study, managed along with spatial relationships such as distance, connectivity, distribution, form, and space between spatial units. The study cases will be carefully chosen and discussed particularly, following a final comparison where an adapted and robust crime prediction model for events will be defined.

This PhD research aims at filling this gap of the social media integration in spatial crime prediction for different event occurrences. During the PhD study I will use the tools to extract, quantify and normalize the social media data and attributes that can lead to better results in geospatial crime prediction analytics models for different events.

Therewith, this research will aim at paving the way for the usage of multidisciplinary tools and integration of the results in geospatial prediction models that can answer spatial and temporal patterns in crime analysis. Spatial criminology theories will be support of the developed analyses during my PhD studies.

2 Related work

Crime presents an increased strategic complexity and interaction with other networks that are not necessarily connected. The main categories of prediction models applied in crime applications include hot spot analysis, regression methods, data mining and machine learning algorithms, near-repeat concept, spatiotemporal analysis and risk terrain analysis (Perry, 2013). For a better prediction algorithms are selected accordingly with the research approach.

Crowd based events (high attendance events) are considered attractors and generators of crime. There are studies emphasizing potential implications of theories like the routine activity, involved in the hooliganism and violence crime and the crime pattern theory, related to crime increase in specific areas for events such as sporting events (Kurland et al, 2014).

The analyses of crime patterns are the base of determining crime displacement, spatially and temporally. However, there is not a lot of focus on specific events in the growing field of spatial crime predictive analytics. This research aims to adapt and use the already mentioned crime prediction methods for events. The social media data processing for event analysis and the integration of the outcomes in the crime prediction models may improve the final results.

The opportunities offered by social media require the establishment of research methodology for drawing insights into extraction of information that can be helpful in many fields, as crime analysis. There is a huge volume of data that social media networks offer and it is analyzed in branches like social sciences, economics, GiScience, computer science, psychology or philosophy.

Key techniques go beyond text analytics to include opinion mining, entity extraction, event recognition, sentiment analysis, topic modeling, social network analysis, trend analysis, and visual analytics. The density of words and their consistency from a lexicon (dictionary) have the likelihood to define relationships between the data. Therefore, it is still an open field of research because of the noisy, unstructured and highly diverse social media data. The analysis of social data parameters, not considering the "spatial" component, was performed mostly from a computer and data science point of view.

The implementation of social media data in crime prediction models started just recently. However, crime prediction algorithms were tested in details through studies in the last five years, the

same can be confirmed for prediction algorithms for social media.

One approach for combining social media and crime data is developed through topic extraction and the connections with crime occurrences. The 2012 was the first time of bringing the social media and crime together in order to make a prediction (Wang et al, 2012). Automatic semantic analysis and NLP of Twitter data, dimensionality reduction through LDA and prediction with linear modeling for hit-and-run crimes in Charlottesville, Virginia represented the earliest research on this topic. Another study investigated the possible integration of rich textual content to predict users spatial trajectories, followed by the correlation with crime occurrences in Chicago, IL (Wang & Gerber, 2015).

A second approach points out the importance of the social media density. If the social media usage is sufficient in an area of study, it may establish a higher predictive value (Featherstone, 2013a; b). Researchers implemented Twitter data as predictors along with archived crime data, which resulted in an increase in the prediction for burglaries and robberies (Bendler et al, 2014a). However, the analysis considered just the number of the tweets and the number and crime type.

Twitter data is considered a proxy for ambient population used in crime rate calculations, showing impact on crime hotspots (Malleson & Andresen, 2015; 2016). Moreover, other datasets can be supportive for ambient population calculations. Considering social media as a dynamic variable, it is important to create also a dynamic population variable (ambient), challenge that would be tested during my PhD development (Kounadi et al, 2017).

Topic modeling and linguistic analysis of spatiotemporal tagged tweets added to crime data in kernel density estimation at neighborhoods level resulted in good predictions for the City of Chicago, IL (Gerber, 2014). Through this research, it was shown that Twitter-derived attributes improve prediction in 19 from 25 crime types. Acknowledging the importance of the study, the temporal patterns might be different for a longer period of time than the three months dataset used. Also the seasonality of crime can affect the prediction accuracy.

An additional innovative attempt considers the implication of sentiment analysis by applying lexicon-based methods and of weather parameters, combine with crime data in a kernel density algorithm (Cheng & Smyth, 2015). For the same city, researchers calculated user ranking for the concept of user credibility and then captured predictive context hidden variables to test in crime rate trend prediction.

Past research has already confirmed that crime types distribution show some similarities throughout different cultures, religions, languages, and socio-economic statuses. However, no research attempt has ever been done before specifically for predicting planned and emerging events considering social media and crime data, at different locations and also at a venue spot.

Besides the crime occurrences connected with sport events, research shows results in detecting sport events on Twitter, the public's overall perception of highly ranked events such as the SuperBowl, and crowd activities related to sport events. Moreover, some researchers are interested in crowd events such as festivals, concerts, political summits, expos, city traffic, etc.

Another important type of event considered in crime research is protests, which can lead to high crime displacement. Recent theoretical background argues that social media may increase the occurrence of emerging events, such as protests. The spatiotemporal variation in the event intensity can be connected with social media activity. On the other hand, the coordination and management of the protest activity might be done on social media, and also the social pressure might be developed through online announcements. The limited existing research in this field considers crowd activities related to events as a proxy for crime analysis and prediction.

As discussed before, there is a growing literature that investigates the impact on crime from events (sporting events, for example), as well as a growing literature that shows how peoples' behavior on social media changes during (sporting) events. However, there is limited research that investigates the relationship, if present, between events, social media activity, and criminal events.

3 Objectives. Research description

Prediction of crime incidents can benefit from social media implementation as an exogenous predictor and for possibly improving the precision of results. The innovative aspect of this research project will be the integration of social media analysis into crime prediction models for specific events and the evaluation of the quality of such predictions. Three main objectives followed by research questions and shortly presented data and methods are in the following rows:

- Objective 1: examine the relationship between the distribution of crime and social media at regularly occurring events

RQ1: What is the relationship between specific types of events and crime types?

RQ2: How can social media predict the diffusion of crimes related to the end of events?

Datasets: crime, tweets, points of interest, residential population, Landscan population.

Methods: topic extraction, text classification by finding "violent tweets"; heat maps, point pattern analyses, hierarchical clustering (KNN), logistic regression.

- Objective 2: investigate the relationship between crime occurrences at a venue and various event types

RQ1: How does the event type affect crime prediction at a venue?

RQ2: How are social media and the number of crimes correlated?

Datasets: crime, tweets, points of interest.

Methods: topic extraction, opinion mining (using Naïve Bayes); Gi* (clusters of points with values higher in magnitude than expected in randomize distributions), Moran's Index I (clustering likelihood), negative binomial logistic regression, evaluation using Area under the Curve (AUC).

- Objective 3: explore the adaptability of spatiotemporal techniques in the evaluation of emerging events (protests, riots)

RQ1: How may a spatiotemporal analysis of social media help identify emerging events influencing crime?

RQ2: How may social media predict crime related to the spatial displacement of an emerging event?

Datasets: crime, tweets, points of interest, old protest data, and socio-economic information

Methods: topic extraction, exponential dispersion models, logistic regression, crime displacement methods, trajectory analysis.

4 Discussion

Overall, this dissertation will focus on geospatial crime predictive analysis concerning planned and emerging events analysis through the exploration of the complex parameters of social media data. Moreover, the study will explore historical crime data and analyze the correlation between crime occurrences and social media data parameters (topic, term frequency, emotions). According to research, there is a tendency of crime prevention initiatives to displace crime or diffuse crime reduction benefits. The analysis will identify information from social media that may help predict crime related to spatial displacement regarding the occurrence of an event. Also other possible risk factors will be considered. Population data is very important in determining crime rates, so determining population at crime risk will be an additional risk factor into the crime prediction models.

The distinctive characteristic of this approach lies in the use of the three data elements in combination with some other information, such as demographic, to provide a new interpretation of social media integration in spatial crime prediction for different event occurrences.

Several spatial statistical models will be applied, including, spatial regression analysis for finding spatial relationships among crime and social data variables, geographically weighted regression for point data validation; linear and logistic regression; global spatial autocorrelation for finding the degree of dependency among the occurrences in the same geographic space.

The above listed methods will help the evaluation and integration of social media information in crime analysis and predictive analytics for event based occurrences. There are limitations in respect to the location of social media data. Because of the rather small percentage of the people who use geo-tagging, algorithms to improve the locational quality through text mining (the location is extracted from the text) were

developed. Other limitation may also be the quality of the crime data. We have to remember that these data are collected by humans, so it is very difficult to eliminate the bias included in all datasets used in research.

As a follow up application of this PhD, the results may be used for a higher effectiveness of police patrols allocation in a larger area of influence, not just on the event location vicinity, and also in monitoring emerging events for negative effects. This would ideally increase policing efficiency, and prevent damages to public property.

5 References

Alruily, M. (2012) Using text mining to identify crime patterns from arabic crime news report corpus.

Bendler, J., Brandt, T., Wagner, S. & Neumann, D. (2014a) Investigating crime-to-twitter relationships in urban environments-facilitating a virtual neighborhood watch.

Bendler, J., Ratku, A. & Neumann, D. (2014b) Crime Mapping through Geo-Spatial Social Media Activity.

Burnap, P. & Williams, M. L. (2015) Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*.

Cheng, Z. & Smyth, R. (2015) Crime Victimization, Neighbourhood Safety and Happiness in China.

Corso, A. J. (2015) Toward Predictive Crime Analysis via Social Media, Big Data, and GIS Spatial Correlation. *iConference 2015 Proceedings*.

Eck, J., Chainey, S., Cameron, J. & Wilson, R. (2005) Mapping crime: Understanding hotspots.

Featherstone, C. (2013a) Identifying vehicle descriptions in microblogging text with the aim of reducing or predicting crime, *Adaptive Science and Technology (ICAST), 2013 International Conference on*. IEEE.

Featherstone, C. (2013b) The relevance of social media as it applies in South Africa to crime prediction, *IST-Africa Conference and Exhibition (IST-Africa), 2013*. IEEE.

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

Kounadi, O., Ristea, A., Leitner, M. & Langford, C. (2017) Population at risk: using areal interpolation and Twitter messages to create population models for burglaries and robberies. *Cartography and Geographic Information Science*, 1-15.

Kurland, J., Tilley, N. & Johnson, S. D. (2014) The Football 'Hotspot'Matrix. *Football Hooliganism, Fan Behaviour and Crime: Contemporary Issues*, 21.

Malleson, N. & Andresen, M. A. (2015) The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, 42(2), 112-121.

Malleson, N. & Andresen, M. A. (2016) Exploring the impact of ambient population measures on London crime hotspots. *Journal of Criminal Justice*, 46, 52-63.

Perry, W. L. (2013) *Predictive policing: The role of crime forecasting in law enforcement operations*Rand Corporation.

Wang, M. & Gerber, M. S. (2015) Using Twitter for Next-Place Prediction, with an Application to Crime Prediction, *Computational Intelligence, 2015 IEEE Symposium Series on*. IEEE.

Wang, X., Gerber, M. S. & Brown, D. E. (2012) Automatic crime prediction using events extracted from twitter posts, *Social Computing, Behavioral-Cultural Modeling and Prediction*Springer, 231-238.