Opinion mining from Twitter and spatial crime distribution for hockey events in Vancouver

Alina Ristea
Doctoral College GIScience, Department of Geoinformatics Z_GIS, University of Salzburg, Salzburg, Austria
mihaela-alina.ristea@sbg.ac.at

Michael Leitner
Department of Geography and Anthropology, Louisiana State University
Baton Rouge, LA, USA
mleitne@lsu.edu

Martin A. Andresen
Institute for Canadian Urban Research Studies, School of Criminology,
Simon Fraser University
Vancouver, BC, Canada
andresen@sfu.ca

Abstract

Spatial analysis of crime during events shows an increased density pattern. In addition, social media activity is intensified during crowded events. Research from the last decade shows relationships between crime patterns and location of Twitter messages. This study is building upon previous literature by using sentiment analysis from Twitter data and applying it in spatial crime analysis regarding events of hockey games. Three crime types are aggregated in dissemination areas for a buffer around Rogers arena in the city of Vancouver, Canada. Landscape features with criminogenic spatial influence are identified, and together with the georeferenced tweets and their sentiment subsets are also aggregated in the spatial units. This approach uses spatial clustering, opinion mining and regression analysis, in order to find meaningful explanatory variables for crime occurrences. Results show the influence of social media text analysis in describing the geography of crime along with the importance of additional criminogenic factors. Regression models displays higher goodness of fit by introducing anticipation, joy, surprise, trust and positive tweets in the models. In addition, traffic signals, liquor businesses, street light poles and public roads were the most significant spatial features in this study for the three crime types.
Keywords: spatial crime patterns; sentiment analysis; regression techniques; Vancouver hockey games.

1 Introduction

The spatial component is an essential element from the environmental criminology perspective. Location of spatial features in different areas and their dispersed or clustered pattern are highly influential in crime analysis. In addition, when changes in the surroundings occur for specific events, the activity space is changing, thus the possibility of crime occurrences. Organized events are attracting a high volume of people on event location, on the route taken by the attendants and in places where they gather to celebrate or to watch the event. Research illustrates two criminology theories regarding events, namely: Routine Activity Theory (Cohen and Felson, 1979) that discusses the high risk of crime to occur at the connection between three main components (suitable targets, motivated offenders, no capable guardians); and theory regarding crime attractors and crime generators (Brantingham and Brantingham, 1984), which defines what makes a place more attractive for criminals. Based on the aforementioned theories, several studies support spatial correlations between sporting events and crime (Kurland et al., 2013, Marie, 2015).

Besides the increase in criminal behaviour, organized events show spikes in social media volume (Cheng and Wicks, 2014). Previous work investigates the social media and crime relationship by using density counts (Bendler et al., 2014) or more developed algorithms for text analysis, such as opinion mining or topic analysis (Al Boni and Gerber, 2016). Opinions are extracted using sentiment analysis techniques. Human feelings are, thus, often classified as positive and negative, for different scales and specific time frames (e.g. in order to detect crowds (Wakamiya et al., 2015). Although comparison between manual, automated and semi-automated sentiment analysis methods were tackled by researchers (Roberts et al., 2018), results show limitations for all of them.

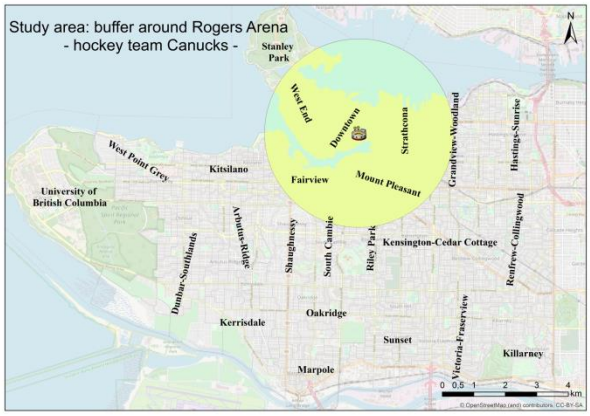
There is an emerging literature showing the connection between events and crime, mostly football games, and people's behaviour on social media during events. However, little research tries to find connections between these three elements. This work contributes to this literature as a case study around a hockey stadium that considers relationships between theft-from-vehicle, mischief and other theft, spatial criminogenic features and Twitter data. Furthermore, sentiment analysis using NRC lexicon was applied over tweet messages in order to extract polarity and eight sentiments. We hypothesize that specific subsets from social media enhances the spatial relationships with crime occurrences and produces improved explanatory models for future crimes. We define game days and comparison days to be able to detect non-routine behaviours during hockey games that can show the role of the stadium as crime attractor. Additionally, we test the significance of multiple spatial features, such as liquor stores or traffic signs, in global and local regression models.

2 Data and Methods

The study area is a buffer of 3km around Rogers arena, home to Canucks hockey team in Vancouver, Canada (Figure1). This study area was selected based on environmental criminology information about activity nodes, that may affect

crime distribution (Kurland, 2014, Ristea et al., 2018). The spatial unit of analysis is the Dissemination Area (DA), defined by Statistics Canada, 295 polygons intersecting the buffer area.

Figure 1: Study area in Vancouver, Canada.



2.1 Data

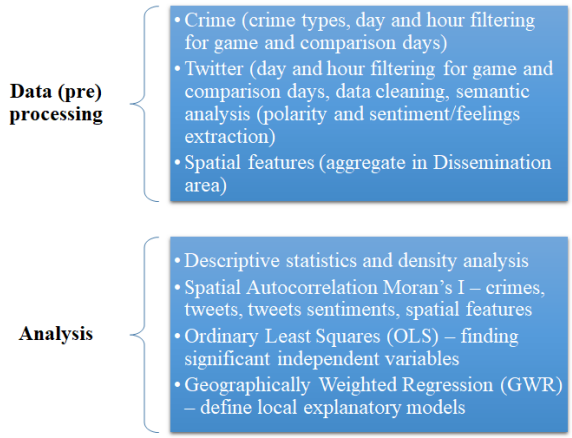
The period of the study includes two hockey seasons, from 2014-2015 and 2015-2016, considering only the home game days: 41 days per season. We also select comparison days, that are days with no sporting activity at the stadium, in close proximity to the game day. The datasets were pre-processed by defining a period of four hours before the start and four hours after the end of the hockey game, similarly for the comparison days.

This study integrates crime data, Twitter messages and spatial features as criminogenic explanatory variables. Firstly, crime data for Vancouver were downloaded from Vancouver Open Data Catalogue: theft from vehicle, mischief and other theft. Secondly, georeferenced tweets were obtained using the Twitter Streaming Application for 2014-2016. Several practical questions may arise because the particular tweets represent approximately one to ten per cent of all posted tweets and different age groups are underrepresented (Zhang et al., 2016, Mitchell et al., 2013, Resch et al., 2017). Lastly, in order to introduce information regarding the surrounding environment, we use the following features: population from census 2011, public roads (number of street segments in each Dissemination Area), parks, street parking, disability parking, motor vehicles parking, street light poles, rapid transit stations, traffic signals, public washrooms, and liquor businesses. All the mentioned variables were aggregated in the 295 DAs from the study area.

2.2 Methods

Figure 2 outlines the main steps of this study. During (pre)processing, we selected the data in the study area and aggregate it within the DA polygons. We also undertook a semantic approach for the Twitter data, by implementing opinion mining algorithm over the georeferenced messages. We used the NRC Emotion Lexicon (Emolex) (Mohammad and Turney, 2010, Mohammad and Turney, 2013) that includes 14,182 unigrams (words) and ~25,000 senses. The outcome of this algorithm involves calculating emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and polarity (negative and positive) for each tweet, and it defines a scale of association between them and the tweet text (not associated, weakly, moderately, or strongly associated). Tweets with an index value of zero are considered "not associated". However, due to the low sentiment value, the tweets classified as "weakly" connected with the specific polarity or feeling can be easily misclassified. In the following analysis we consider just the messages included in the "moderately" and "strongly associated" groups. The NRC algorithm was implemented using the Syuzhet package in R (Jockers, 2015). The lexicon was manually annotated using Mechanical Turk users.

Figure 2: Workflow.



With reference to the spatial analysis, we used Moran's I (Anselin and Kelejian, 1997). The results are divided in three elements, the Observed General G, Expected General G, z-score and p-value. When the p-value is statistically significant ($p \leq 0.05$), we can reject the null hypothesis. Then, when z-score is positive means that high values of the analysed data are more clustered in space than expected in a random situation. The Observed General G is larger than the Expected General G when z-score is positive. Furthered, we considered the three crime types as dependent variables, taken one by one, in regression models.

Firstly, our approach integrates Ordinary Least Squares (OLS), in order to determine significant independent variables for crime occurrences. All the OLS models include all spatial features mentioned in the Data section, while Twitter subsets were added one at a time to avoid multicollinearity, summing 66 OLS models (22 models for each crime type, from which 11 for game days and 11 for comparison days).

Secondly, we applied Geographical Weighted Regression (GWR) in order to identify any spatial influence from the explanatory variables (Fotheringham et al., 1998, Brunsdon et al., 1996). In the GWR models, we include only the significant values from the aforementioned OLS models. The bandwidth for GWR was selected for adaptive distance. It is important to mention that as many other analytic methods, GWR has limitations, including multiple hypothesis testing, local collinearity in coefficients, or bandwidth selection

(Wheeler and Tiefelsdorf, 2005, Wheeler and Waller, 2009). Researchers discuss these concerns; however, GWR is still a used tool in exploring local spatial relationships.

3 Results and Discussion

3.1 Spatial patterns

The three crime types display similar density patterns, with higher values in the centre of the study area, tending to fall off on the southern part (Figure3), in a similar situation with the tweets (Figure4). In the immediate vicinity of the stadium, all crime types have increased density when hockey games are played. It is important to mention that the Downtown area is located close to the stadium, which is in itself a criminogenic factor. Theft-from-vehicle and mischief subsume more crimes during game days, showing also a higher standard deviation, while other theft is more prone to occur during comparison days.

Figure 3: Density distribution of disaggregated crime types.

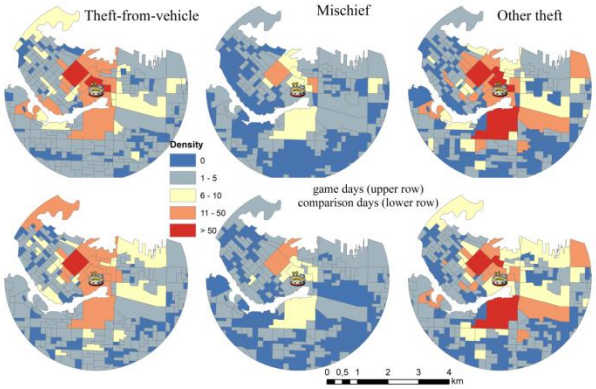
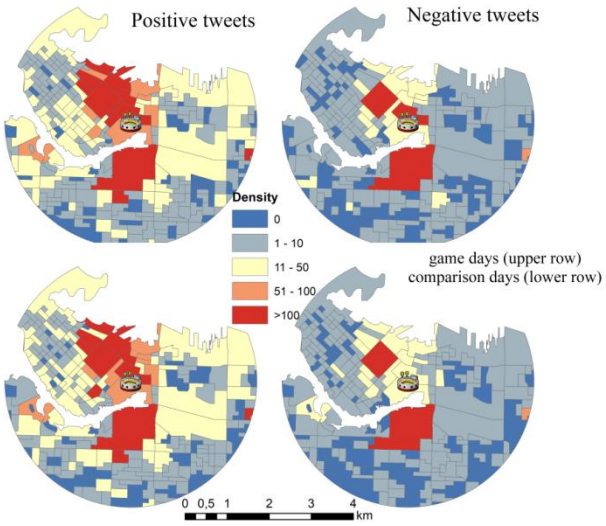


Figure 4: Density distribution of tweets polarity: positive and negative.



Supporting previous research, we find higher counts of tweets during game days. After the semantic analysis, tweet subsets for anger, disgust, fear, sadness and negative show less volume during game days, while anticipation, joy, surprise, trust and positive tweets increase. This may be an event-effect on social behaviour, mostly from the positive point.

Moran's I was calculated for the crime data and explanatory variables, to identify clustered, dispersed or random patterns. It is worth mentioning that the p-values refer to global clustering throughout the study area; as such, if there are no general clusters, they they exist at higher spatial scale. Parks, disable parking and washrooms are the only spatial features that are not clustered in the study area, having an insignificant p-value (Table1). Mischief is the only crime type with an increased clustering index during comparison days. Interestingly, the feelings of anger, disgust, fear and sadness have a random distribution during comparison days and a clustered one during game days (Table2), auxiliary to the count analyses.

Table 1: Spatial Autocorrelation (Global Moran's I) for crimes and spatial features

	Observed General G	z-score	p-value
Aggregated crimes	0,31	11,44	0,00
	0,30	11,36	0,00
Theft-from-vehicle	0,31	10,90	0,00
	0,22	8,69	0,00
Mischief	0,23	8,56	0,00
	0,29	10,60	0,00
Other theft	0,30	10,09	0,00
	0,21	8,56	0,00
Population	0,10	3,19	0,00
Parks	0,05	1,48	0,14
Disable parking	0,01	0,46	0,65
Street parking	0,37	11,47	0,00
Motor parking	0,17	5,58	0,00
Light poles	0,19	6,58	0,00
Rapid trans	0,12	3,99	0,00
Traffic signs	0,30	9,54	0,00
Washrooms	-0,01	-0,55	0,58
Liquor stores	0,24	7,74	0,00
Public roads	0,21	7,04	0,00

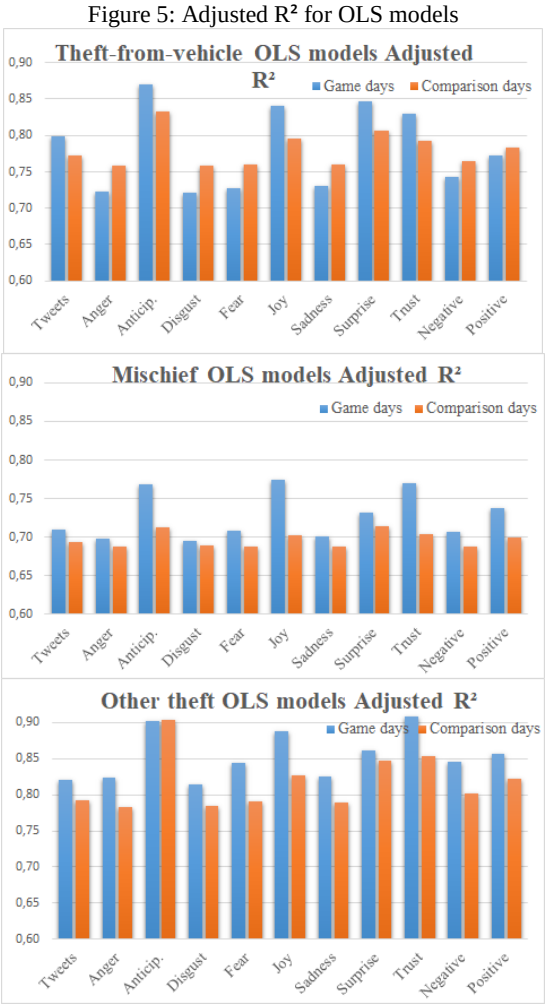
Table 2: Spatial Autocorrelation (Global Moran's I) for tweets and subsets from opinion mining

	Observed General G	z-score	p-value
Geolocate tweets	0,15	5,10	0,00
	0,17	5,71	0,00
Anger	0,02	1,24	0,00
	0,00	0,25	0,81
Anticipation	0,19	6,77	0,00
	0,21	7,58	0,00
Disgust	0,04	1,87	0,06
	0,02	1,32	0,19
Fear	0,04	1,68	0,09
	0,01	0,61	0,54
Joy	0,20	7,46	0,00
	0,18	6,28	0,00
Sadness	0,05	2,16	0,03
	0,02	1,18	0,24
Surprise	0,20	6,98	0,00
	0,25	9,26	0,00
Trust	0,23	8,16	0,00
	0,21	7,41	0,00
Negative	0,06	2,32	0,02
	0,04	2,03	0,04
Positive	0,24	8,03	0,00
	0,25	8,45	0,00

3.2 Explanatory models

OLS models were run for each crime type, including the spatial features and the tweets (one by one integrated in the regression), summing 66 models. Results show that some tweet subsets overtake the georeferenced tweets effect in the adjusted R^2 values. Figure 5 shows that during game days tweets, anticipation, joy, surprise and trust feelings display higher values of adjusted R^2 for theft-from-vehicle, while for mischief and other theft all the Twitter datasets in game days show an over adjusted R^2 than comparison. Practically this figure shows that by using tweets or some subsets of them would increase the power of the model. At a first glance, this result appears to be related with the density patterns, however tweets do not always have higher counts during game days. By looking at Moran's I values, we notice clustered patterns for anger, disgust, fear and sadness only during game days, that partially explains the results. The most significant variables were traffic signals, liquor businesses, street light poles, public roads, disability parking (not significant for mischief in comparison), and motor vehicles parking (not significant for other theft in comparison). For example, street light poles can define light and dark areas. Different case studies show arguments on improving street lightning for crime prevention (Painter, 1996, Suminski et al., 2005) and perception (Craig et al., 2002), while other support the fact that lights on the streets are not correlated with crime (Nair et al., 1993). Population from Census was not significant for any theft-from-vehicle model, washrooms not in game days for any crime and parks not significant in any of all models. This information is confirming literature regarding the questionable use of residential population in crime models for specific crime types (Malleson and Andresen, 2015, Kounadi et al., 2017).

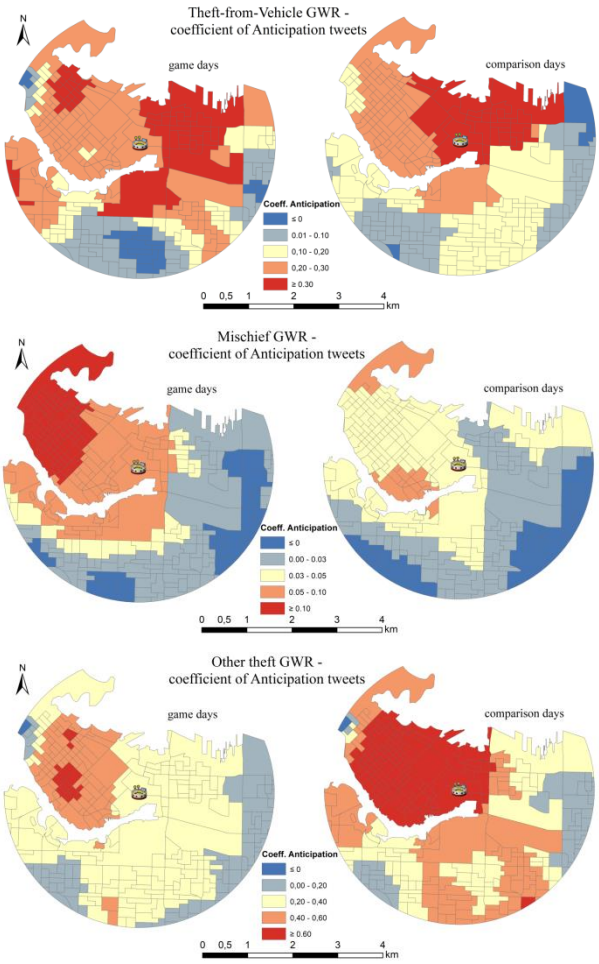
Figure 5: Adjusted R^2 for OLS models



In order to analyse local patterns, we apply GWR models considering only the significant variables from OLS. Supporting OLS, anticipation, joy and trust tweets increase the explanatory power of GWR compared with the models using the georeferenced tweets (Table3). In particular, models including anticipation tweets are stronger for all crime types (Figure6). This shows the importance of detecting a subset of tweets according to the work purpose, in this case finding relationship with crimes. R^2 values present the proportion of variation in the dependent variable (crime) given by the independent variables. The values are influenced by the

number of independent variables, which is why it is better to consider the adjusted R².

Figure 6: Coefficient of anticipation tweets in crime types GWR models



4 Conclusion and Future Work

This study introduces insights in showing spatial relationships between criminal activity, Twitter activity and other socio-demographic and economic variables, in the context of hockey game events around the stadium area. It shows that specific sentiments from tweets (anticipation, joy, trust and positive) can end up explaining and predicting crime better than all georeferenced tweets, in addition to other predictors.

This research was applied in the city of Vancouver, in a buffer area around Rogers stadium, home place for Canucks hockey team, for two seasons 2014-2016. We considered game days and comparison days and aggregated the spatial data of this study in dissemination areas. Spatial density of crime occurrences is more intense for game than comparison days in the buffer around the stadium, mostly for theft-from-vehicle and other theft, and their spatial pattern is clustered in both cases. OLS models show that anticipation, joy, surprise, trust and positive tweets lead to higher R^2 models for crimes rather than using all georeferenced tweets. In addition, GWR models including anticipation tweets are slightly stronger than the ones using georeferenced tweets. For the coefficients in GWR, anticipation tweets show higher values for theft-from-vehicle and mischief during games, while other theft has increased influence in comparison days. However, this study is not without limitations. Using the low volume of freely downloadable Twitter and considering the bias from the text data, this analysis may raise concerns. In addition, tracking emotions from text can be considered subjective and having a relative scale. Acknowledging the abovementioned, we believe the results of this study are relevant for crime prevention and spatial crime analysis, together with information about spatial distribution of social data.

Considering the results, a future direction can consider splitting the game days according to their result: win, lose or draw. This will give more insight in the clustering patterns of specific sentiments and their influence in the explanatory models. In addition, topic modelling using Latent Dirichlet Allocation (LDA) would improve the performance of the already analysed opinion mining, by adding information about

Table 3: R^2 and R^2 adjusted of GWR models, for game days and comparison days

	Game days		Comparison days		Game days		Comparison days		Game days		Comparison days	
	Theft-from-vehicle				Mischief				Other theft			
Independent	R2	AdjR2	R2	AdjR2	R2	AdjR2	R2	AdjR2	R2	AdjR2	R2	AdjR2
Tweets	0,95	0,94	0,94	0,91	0,88	0,83	0,82	0,78	0,95	0,93	.	.
Anger	.	.	.	.	.	.	.	.	0,96	0,94	0,96	0,94
Anticip.	0,95	0,94	0,95	0,92	0,89	0,85	0,84	0,79	0,96	0,94	0,94	0,93
Disgust	.	.	.	.	.	.	.	.	0,95	0,93	0,95	0,93
Fear	0,93	0,90	.	.	.	.	.	.	0,92	0,91	0,96	0,95
Joy	0,95	0,93	0,94	0,92	0,88	0,84	.	.	0,97	0,95	0,96	0,95
Sadness	0,93	0,91	.	.	.	.	.	.	0,95	0,93	0,95	0,93
Surprise	0,95	0,93	0,94	0,91	0,87	0,82	0,82	0,78	0,95	0,92	0,94	0,92
Trust	0,95	0,92	0,94	0,92	0,88	0,85	0,83	0,78	0,97	0,96	0,96	0,95
Negative	0,94	0,91	0,94	0,91	.	.	.	.	0,95	0,94	0,96	0,94
Positive	0,95	0,93	0,94	0,91	0,88	0,84	0,83	0,79	0,95	0,93	0,95	0,93

Note: the GWR models were run for the variables when OLS determined tweets or their subsets as significant

message topics. Thus, a future step includes adding the gathered information in a predictive machine learning technique by including training and testing data and analyse the prediction value of all georeferenced tweets, opinion mining subsets and topic subsets over crime occurrences. The implications of this study show that georeferenced tweets and their emotions, supplemented by additional information, can be helpful on portraying the geography of crime. The outcomes of this study support previous theories and add to them by integrating information about crime types which, for this case study, are correlated with social media text analysis and distribution.

Aknowledgement: This research was funded by the Austrian Science Fund (FWF) through the Doctoral College GIScience at the University of Salzburg (DK W 1237-N23).

5 References

Al Boni, M. & Gerber, M. S. (2016) Predicting crime with routine activity patterns inferred from social media. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*.), Budapest, Hungary.

Anselin, L. & Kelejian, H. H. (1997) Testing for spatial error autocorrelation in the presence of endogenous regressors. *International Regional Science Review* **20(1-2)**:153-182.

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

Brantingham, P. J. & Brantingham, P. L. (1984) *Patterns in crime.* Macmillan New York.

Brunsdon, C., Fotheringham, A. S. & Charlton, M. E. (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis* **28(4)**:281-298.

Cheng, T. & Wicks, T. (2014) Event detection using Twitter: a spatio-temporal approach. *PloS one* **9(6)**:e97807.

Cohen, L. E. & Felson, M. (1979) Social change and crime rate trends: A routine activity approach. *American sociological review*:588-608.

Craig, C. L., Brownson, R. C., Cragg, S. E. & Dunn, A. L. (2002) Exploring the effect of the environment on physical activity: a study examining walking to work. *American Journal of Preventive Medicine* **23(2)**:36-43.

Fotheringham, A. S., Charlton, M. E. & Brunsdon, C. (1998) Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and planning A* **30(11)**:1905-1927.
Jockers, M., L (2015) Syuzhet: Extract Sentiment and Plot Arcs from Text.).

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. (2014) The Ecology of Football-Related Crime and Disorder. In *Department of Security and Crime Science*.) University College London, vol. Ph.D., pp. 1-420.

Kurland, J., Johnson, S. D. & Tilley, N. (2013) Offenses around stadiums: A natural experiment on crime attraction and generation. *Journal of research in crime and delinquency*:0022427812471349.

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.

Marie, O. (2015) Police and thieves in the stadium: measuring the (multiple) effects of football matches on crime. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S. & Danforth, C. M. (2013) The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one* **8(5)**:e64417.

Mohammad, S. M. & Turney, P. D. (2010) Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*.) Association for Computational Linguistics, pp. 26-34.

Mohammad, S. M. & Turney, P. D. (2013) Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* **29(3)**:436-465.

Nair, G., Ditton, J. & Phillips, S. (1993) Environmental improvements and the fear of crime: the sad case of the 'Pond'area in Glasgow. *The British Journal of Criminology* **33(4)**:555-561.

Painter, K. (1996) The influence of street lighting improvements on crime, fear and pedestrian street use, after dark. *Landscape and Urban Planning* **35(2-3)**:193-201.

Resch, B., Usländer, F. & Havas, C. (2017) Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*:1-15.

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.

Roberts, H., Resch, B., Sadler, J., Chapman, L., Petutschnig, A. & Zimmer, S. (2018) Investigating the Emotional Responses of Individuals to Urban Green Space Using Twitter Data: A Critical Comparison of Three Different Methods of Sentiment Analysis. *Urban Planning* **3(1)**:21-33.

Suminski, R. R., Poston, W. S. C., Petosa, R. L., Stevens, E. & Katzenmoyer, L. M. (2005) Features of the neighborhood environment and walking by US adults. *American Journal of Preventive Medicine* **28(2)**:149-155.

Wakamiya, S., Belouaer, L., Brosset, D., Lee, R., Kawai, Y., Sumiya, K. & Claramunt, C. (2015) Measuring crowd mood in city space through twitter. In *International Symposium on Web and Wireless Geographical Information Systems*.) Springer, pp. 37-49.

Wheeler, D. & Tiefelsdorf, M. (2005) Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems* **7(2)**:161-187.

Wheeler, D. C. & Waller, L. A. (2009) Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. *Journal of Geographical Systems* **11(1)**:1-22.

Zhang, Z., Ni, M., He, Q. & Gao, J. (2016) *Mining Transportation Information from Social Media for Planned and Unplanned Events*. Buffalo, NY United States, pp. 1-68.