International Journal of
Geo-Information

Article

Estimating the Spatial Distribution of Crime Events around a Football Stadium from Georeferenced Tweets

Alina Ristea [1,*], Justin Kurland [2], Bernd Resch [1,3], Michael Leitner [1,4] and Chad Langford [1]

[1] Doctoral College GIScience, Department of Geoinformatics-Z_GIS, University of Salzburg, Schillerstraße 30, 5020 Salzburg, Austria; bernd.resch@sbg.ac.at (B.R.); mleitne@lsu.edu (M.L.); chad.langford@sbg.ac.at (C.L.)
[2] Computing and Mathematical Sciences—Institute for Security and Crime Science, University of Waikato, Gate 1 Knighton Road, Private Bag 3105, Hamilton 3240, New Zealand; justin.kurland@waikato.ac.nz
[3] Center for Geographic Analysis, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA
[4] Department of Geography and Anthropology, Louisiana State University, E-104 Howe-Russell-Kniffen Geoscience Complex, Baton Rouge, LA 70803, USA
* Correspondence: mihaela-alina.ristea@sbg.ac.at; Tel.: +43-662-8044-7557

Received: 30 November 2017; Accepted: 28 January 2018; Published: 31 January 2018

Abstract: Crowd-based events, such as football matches, are considered generators of crime. Criminological research on the influence of football matches has consistently uncovered differences in spatial crime patterns, particularly in the areas around stadia. At the same time, social media data mining research on football matches shows a high volume of data created during football events. This study seeks to build on these two research streams by exploring the spatial relationship between crime events and nearby Twitter activity around a football stadium, and estimating the possible influence of tweets for explaining the presence or absence of crime in the area around a football stadium on match days. Aggregated hourly crime data and geotagged tweets for the same area around the stadium are analysed using exploratory and inferential methods. Spatial clustering, spatial statistics, text mining as well as a hurdle negative binomial logistic regression for spatiotemporal explanations are utilized in our analysis. Findings indicate a statistically significant spatial relationship between three crime types (criminal damage, theft and handling, and violence against the person) and tweet patterns, and that such a relationship can be used to explain future incidents of crime.

Keywords: spatial crime analysis; Twitter mining; football related crime and disorder; spatial correlation; explanatory analytics

1. Introduction: Connecting Crime, Social Media, and Football Matches

Monitoring and analysing crowd-based events shows that such events play an important role as attractors and generators of crime. Recent research has found spatial correlations between crime and sporting events using various approaches. Additionally, it has been shown that social media platforms escalate the collective action of violent incidents after sporting events. This study aims to investigate the relationship between crime occurrences and geotagged Twitter activity in the proximity of a football stadium, and to test the inclusion of geotagged tweets in regression-based spatial crime explanatory models.

1.1. Sport and Spatial Crime Analysis

Football matches are responsible for producing both positive and negative externalities on society and local communities. On the negative side, literature has consistently demonstrated a relationship between football and hooliganism [1,2], disorderly fan behaviour [3,4], and a change in the count and distribution of crime events on home match days [4]. Positive externalities are highly emphasized by researchers in the framework of sporting events, including economic impact, such as subsequent changes in the stock market [5–7], mostly when the team is winning [8]; and the psychological aspect and social benefits [9–11]. This study focuses on those specific crimes that are influenced by conditions in and around a stadium on a football match day. More specifically, previous research conducted in the UK at Villa Park (the study area) has demonstrated a relationship between football matches and criminal damage, theft and handling, and violence against the person [4]. Consequently, this study utilizes these three crime categories.

In this paper, the problem of football-related crime is considered from an environmental criminological perspective. The approach is not a single theory, but instead refers to a suite of theories that share a common interest in criminal events. Routine activity theory (RAT), one of the various theories that contribute to the environmental criminological approach, is particularly suitable for attempting to understand the spatial-temporal connection between crime and football matches. More specifically, the theory suggests that there are three minimum elements required to converge in time and space to produce crime: A (1) motivated offender, (2) suitable target, and (3) the absence of a capable guardian. The lack of any one of these elements would be enough to prevent a crime event [12].

Specifically for football matches, the three minimum elements may, for instance, be present as follows: The large number of supporters attending matches is all potentially suitable targets; they themselves may be motivated offenders or may attract motivated offenders; and they may also constitute potential guardians. Additionally, the police are generally too few in number to provide substantial guardianship. Thus, conditions are favourable for the abovementioned crime types to be committed. It is worth noting that the RAT in the area around a football stadium would be different on a day that a match did not take place, which is why non-match days are identified for comparison [4,13,14].

1.2. Spatial-Temporal Variations in Crime Due to Football Matches

A growing number of recent studies on football-related crime have focused on changes in the spatial and temporal distribution of crime events in and around football stadia on both match and non-match days [15]. Propinquity to football stadia has been consistently identified as being significantly related to higher levels of crime and disorder. Indeed, [4] found that the distribution of theft and handling events, as well as violence against the person, significantly differed from on non-match equivalents across an area that extended three km from Wembley National Stadium in London. At the same time, various studies have explored changes in the temporal distribution of crime on football match days using hourly [16], daily [4,16] and seasonal [3,13] data. Using a similar methodology, the relationship between sporting events and crime incidents has been extended to demonstrate the temporal relationship, at the hourly level, between basketball games and robberies in Memphis, TN [17].

1.3. Data mining Social Media for Sporting Event Information

Incorporating social media data into predictive models for sporting events has become a reality only recently, owed largely to the increasing presence and availability of social media data. Before the global attraction of online social networks, researchers used data mining of videos or newspapers to extract information about football matches. A decision tree was created for football goal detection [18,19] with high precision (92.3%). Based on literature, Twitter data have been used for detecting goals, red cards, or penalties during football matches using semantic or "sentiment" analysis [20–22]. There are, however, still some relevant problems to be addressed with respect to semantic analysis. Real-time sport event detection and summarization in Twitter data is relatively new, but has been tested with American football games using Hidden Markov Chain (HMM) [23], and the

2010 World Cup using an unsupervised algorithm for generating a textual summary of events [24]. Summarizing each game using Twitter data and analysing crimes for the same time frame may show different crime patterns, but the approach is not utilized for this study.

Previous studies have shown it is possible to extract useful information from football related Twitter data via semantic analyses using simple subsets from specific dictionaries [25], naïve Bayes models, random forests, logistic regression, and support vector machine [26,27]. These earlier empirical studies motivated the research question in this paper. That is, whether the combination of historical crime data and a semantic analysis of tweets may help explain the spatial and temporal distribution of future crime events on match days.

1.4. Spatial Crime Prediction Using Social Media

A growing body of empirical evidence suggests a strong relationship between the spatiotemporal distribution of crime and social media. For studies focused exclusively on crime analysis, the spatial relationship between the density of social media (primarily tweets) and crime locations has been explored with no concern for the semantic message embedded in tweets [28–32]. If social media usage is sufficient for a particular study area, its usage may have some predictive value [29,30]. Researchers have implemented Twitter data as predictors together with archived crime data. This has, for example, resulted in an increase in the prediction for burglaries and robberies [28]. Such analyses has considered tweets and crime intensity and type, only, which can be useful for density and hot spot analyses, and spatial forecasting. Another example illustrates the attempt of weighing social media attributes for risk terrain modelling (RTM) analysis in conjunction with other types of datasets to improve crime prediction. The technical approach to RTM in crime forecasting is to identify possible factors that are spatially related to crime occurrences for which risk is being assessed. Essentially, RTM assigns weights to each of these criminogenic factor at every grid cell throughout the study area [33]. The RTM method for prediction was used by researchers in exploratory analysis with social media data as an input layer [34]. It shows that the integration of social media data in hot spot analysis had an impact on the visual outcome representation and it suggests correlation between datasets, however it does not show the implications of social media in prediction.

Another group of studies has examined the semantics of social media text and crime occurrences. Researchers included Twitter semantic analysis in combination with historical crime data in an attempt to improve crime prediction models [35]. More specifically, automatic semantic analysis of geotagged tweets, dimensionality reduction through latent dirichlet allocation (LDA), and crime prediction with linear or logistic models for different crime categories have all been used for crime forecasting [36–38]. The year 2012 also marks the beginning, when social media was introduced for the first time to forecast when and where crime would occur [35]. Automatic semantic analysis and Natural Language Processing of Twitter data, dimensionality reduction through LDA, and forecasting with linear modelling for hit-and-run crimes in Charlottesville, VA represented one of the earliest examples of this research. During the same year [36] text mining was used for finding crime patterns from crime reports that were written in the Arabic language. Another recent study investigated the possible integration of the rich textual content of Twitter data to automatically predict users' spatial trajectories with the so-called "next-place prediction model". In this same research, future spatial trajectories were also correlated with crime occurrences in Chicago, IL [37].

In addition to the academic research that has been conducted in this area, several companies (e.g., IBM, Google, Microsoft) have developed crime prediction software that utilizes machine-learning techniques with considerable success. For example, PREDPOL [39], bespoke software developed by mathematicians and behavioural scientists from UCLA and Santa Clara University, both located in the U.S., attempts to predict future crime by considering observations where certain crime types tend to cluster in time and space. Hitachi has combined live Twitter feeds and semantic analysis with historical crime data in order to predict crime occurrences [40].

1.5. Research Gap and Objectives

To date, crime prediction models that have taken advantage of social media data have had high predictive success for certain types of crime [38,41–45]. Despite this, many questions persist regarding how to best evaluate and integrate social media data for crime analysis. Our study utilizes this growing empirical base of tweets to test the explanatory function of these geotagged, violent tweets (i.e., tweets using violent words) under contrasting ecological conditions—match and non-match comparison days. Tweets are considered a proxy measure for ambient population [31], and they were temporally processed using the same methodology as the crime data. Other variables used in this analysis are time-insensitive (e.g., in all models the number of spatial features such as pubs are stable). Additionally, it attempts to evaluate social media data as a potentially significant factor for spatial crime explanatory models, and to test their influence in explanatory models. This study also incorporates information from the LandScan database for ambient population and the Geostat database for residential population, along with additional spatial features which prior research has considered representative in analysing crime occurrences around stadia [46].

To achieve this we use a bottom-up study design to address three specific hypotheses:

Hypothesis (H1). *Twitter activity correlates spatially (and temporally) with the density of crimes in and around a football stadium before, during, and after matches.*

Hypothesis (H2). *Violent tweet and football-related tweet densities are more highly correlated with crime occurrences than the total density of all tweets.*

Hypothesis (H3). *Geotagged tweets are a useful factor for explaining the presence or absence of crime in the area around a football stadium on match days.*

In the section that follows, we will discuss data collection and how data were cleaned and processed for the subsequent analysis. This is followed by a discussion of methods and analyses. Next, results are presented along with a discussion about how they contribute to the existing theoretical literature. We conclude with a discussion of the practical implication of findings and study limitations, and suggest future research that should be considered in light of our findings.

2. Data

Table 1 outlines all data sources used in this study. Each source is described in more detail in the following sections, along with how the data were pre-processed for analysis.

Police-recorded crime data were time-stamped and geocoded for five football seasons (2005–2010). Geocoding accuracy was tested (98 per cent occurred within 50 m of the appropriately assigned address-point) to ensure the data exceeded the minimum acceptable hit rate of 85 per cent [47] for the spatial analysis of police-recorded crime data [46].

Geotagged tweets were obtained using the Twitter Streaming API for 2012. Twitter data are highly considered in research, however several practical concerns may arise because these particular tweets represent approximately one [48,49] to ten percent of all online posted tweets [50,51].

Though we discuss this further in the limitations section below, it is important to note here that the police-recorded crime data and the Twitter data do not correspond temporally. That is, Twitter was already in place during the period corresponding with the police-recorded crime data (2006–2010), but it did not have a large volume of user activity. It was not until 2012 that the volume of Twitter usage was sufficient to extract the subset of geocoded tweets. Because of the importance of land use in the spatial distribution of crime [52] and the fact that land use change is very slow [53] we do not expect the spatial distribution of tweets to change, only their volume. In addition, we tested crime pattern stability using a Spatial Point Pattern test [54] where we considered the disaggregated crime types for game days and also for comparison days. Average results show ~80% similarity, which is within the threshold established by test designers [55,56]. Considering the amalgamated crime types, test results do not exceed the threshold (~65% similarity). As such, we expect the 2012 Twitter data to be a reasonable approximation of the spatial distribution of the population at risk for the slightly earlier period of the police-recorded crime data.

Table 1. Summary of datasets used in this study (** subsets of all collected geotagged tweets).

Supplier(s)	**Attribute(s)**	**Year(s)**	**Data Type**	**Purpose(s)**	**Sample Size**
West Midlands Police, British Transport Police	Geocoded, time-stamped police-recorded crime data	2005–2010	Vector (point data)	Used to analyse spatial and temporal crime patterns	2399 crimes
Twitter	Geotagged, time-stamped tweets	2012	Vector (point data)	Used to analyse relationships with crime patterns	6274 tweets
** Violent tweets					408 tweets
** Football-related tweets					1678 tweets
Online databases	Match dates, kick-off times, match locations	2005–2010 and 2012	Text	Used only for location and time of matches, not otherwise included in the analysis	240 days
LandScan^{TM}	Ambient population data	2008	Regular grid, ~1 km	Used as independent variable in explanatory models	486,443 persons
GEOSTAT	Residential population	2011	Regular grid, ~1 km	Used as independent variable in explanatory models	245,964 persons
Bars, pubs, and fast food restaurants	Location in stadium proximity	2010	Vector (point data)	Used as independent variable in explanatory models	131 points

Two datasets were used to estimate population in the explanatory models. There is an emerging discussion in the literature about which population is better to use for the population at risk, because some crime types are influenced by a dynamic population [57,58]. With this in mind, residential population from GEOSTAT database (2011) and ambient population from LandScanTM (2008) were used.

Population information from global LandScanTM (2008) was used as an independent variable for ambient population in the explanatory models. This dataset, created by Oakridge National Laboratory, is used to more accurately estimate on-street populations. LandScan data are collected annually using a number of different algorithms and techniques, such as remote sensing to estimate ambient population over the course of a 24-h period at a spatial resolution of approximately 1 km^2.

Population data from the GEOSTAT database (2011) is produced by Eurostat in cooperation with the European Forum for GeoStatistics (EFGS) and it is freely available online [59]. It represents grid data disaggregated from original Census data at ~1 km^2 resolutions.

Locations of bars, pubs, and fast food restaurants were extracted from OpenStreetMap. An online football database [60] was used for finding football match dates, and, in turn, for identifying a set of relative non-match comparison days for further analyses. Match days are considered the ones when the team plays at the home stadium.

The analysis focuses exclusively on three crime categories (criminal damage, theft and handling, and violence against the person) that took place on football match days between 2005 and 2010 and, for reasons of comparison, non-match days selected to be as similar as possible to match days to reduce potential confounders, such as day of the week or seasonal effects. For consistency, the same match and non-match day methodology was utilized to identify an appropriate sample from 2012 geotagged tweets.

Study Area

The study area is located within a three km radius around the Villa Park stadium, home to the Aston Villa Football Club in Birmingham, United Kingdom (Figure 1). This radius was selected in part because of the presence of key activity nodes that, according to environmental criminology, may affect the distribution of crime [46].

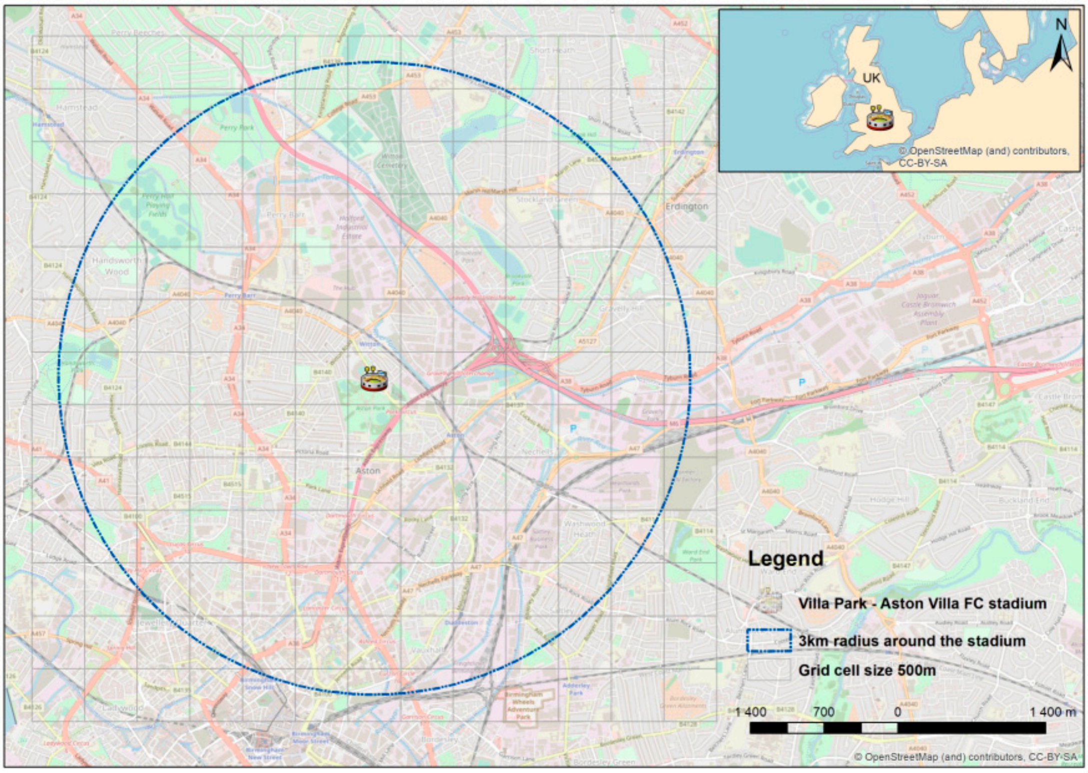

N
UK
© OpenStreetMap (and) contributors, CC-BY-SA
Erdington
Aston
Legend
Villa Park - Aston Villa FC stadium
3km radius around the stadium
Grid cell size 500m
1 400
700
0
1 400 m
© OpenStreetMap (and) contributors, CC-BY-SA

Figure 1. Study area of Aston Villa stadium, Birmingham, UK.

Crime and geotagged tweets are spatially aggregated to 500 m × 500 m grid cells (N = 169). The 250,000 m^2 cell size was identified in accordance with the recommendation for grid size calculation for crime hotspot analysis [61].

3. Methodology

The primary objective of this study is to uncover the relationship between crime events and subsets of geotagged tweets (violent tweets and football-related tweets). After data processing (see Figure 2 for details), spatial statistics are applied. First, the Moran's I (a descriptive spatial autocorrelation index) is calculated for all crime and tweet subsets in order to identify global clustering or dispersion in the spatial data distributions for game and comparison days. Next, the bivariate Moran's I is applied to find spatial relationships between dataset pairs, such as crimes and tweets for identified time frames. In this context, it is worth mentioning that correlation is not prove for causation. However, correlation can suggest a causal relationship [62]. After identifying the degree of correlation between crimes and tweets, the negative binomial logistic regression is applied to explore statistically significant dependent variables for crime subsets. The outcome from this regression model serves as the base for the predictive testing. Figure 2 details the analysis framework, which is explained further in subsequent sections. Data processing steps appear in Sections 3.1 and 3.2, applied algorithms are explained in Section 3.3, Section 3.4, Section 3.5, and results are discussed in Sections 4.1 and 4.2.

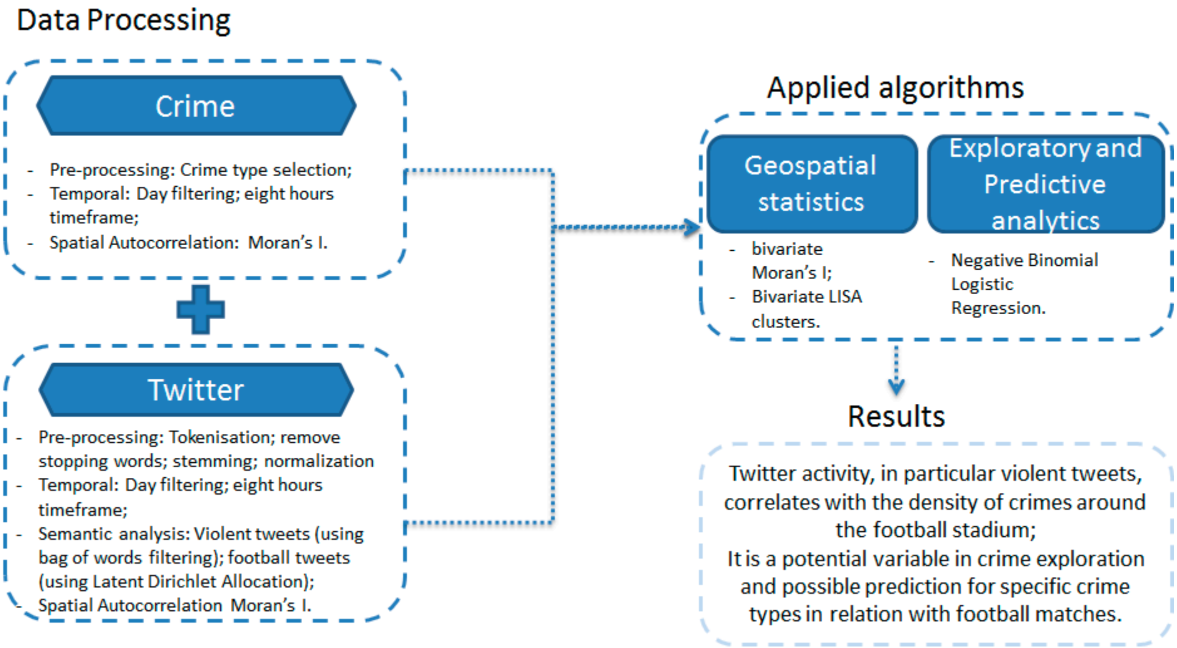

Data Processing
Crime
- Pre-processing: Crime type selection;
- Temporal: Day filtering; eight hours timeframe;
- Spatial Autocorrelation: Moran's I.
Twitter
- Pre-processing: Tokenisation; remove stopping words; stemming; normalization
- Temporal: Day filtering; eight hours timeframe;
- Semantic analysis: Violent tweets (using bag of words filtering); football tweets (using Latent Dirichlet Allocation);
- Spatial Autocorrelation Moran's I.
Applied algorithms
Geospatial statistics
- bivariate Moran's I;
- Bivariate LISA clusters.
Exploratory and Predictive analytics
- Negative Binomial Logistic Regression.
Results
Twitter activity, in particular violent tweets, correlates with the density of crimes around the football stadium;
It is a potential variable in crime exploration and possible prediction for specific crime types in relation with football matches.

Figure 2. Analysis framework.

3.1. Temporal Crime Event Processing

To collect a representative crime and tweets dataset, those days that the majority of matches took place were identified (Wednesday, Saturday, and Sunday), along with associated comparison days. Crimes were then recoded and aggregated into eight hourly bins in accordance to when they took place relative to match kick-offs. More specifically, they were recoded and aggregated into the four hours prior to match kick-off, and into the four hours after kick-off for both the match and comparison days. This process was done because of the variation in match kick-off times. For example, the majority of Wednesday matches began at 7:45 p.m., on Saturday between 3:00 and 5:30 p.m., and on Sunday between 12:00 and 4:00 p.m.. Thus, by recoding the time a crime took place, relative to when a match was played, the hourly distribution of crime intensity (crime counts normalized by percentage) for Wednesdays, Saturdays, and Sundays for both match and comparison days can be more readily compared (Figure 3).

Clear differences between when crime increases during the hours immediately around the kick-off time for match days can be visualized for each day relative to their respective comparison day hourly crime counts. Similar, temporal filtering was adopted for geotagged tweets. Geotagged tweets were extracted for match and comparison days, and hourly aggregated for the same eight-hour timeframe. The temporal distribution of match and comparison day tweets is visualized in Figure 4.

Interestingly, the distribution of crimes and tweets are similar, but shifted in time. More specifically, in the weekend the peaks for the three crime categories occur in the hour prior to kick-off, but the tweet peak coincides with the kick-off hour. One notable difference is on Wednesday, when the crime's peak is at kick-off time and the tweets three hours after kick-off.

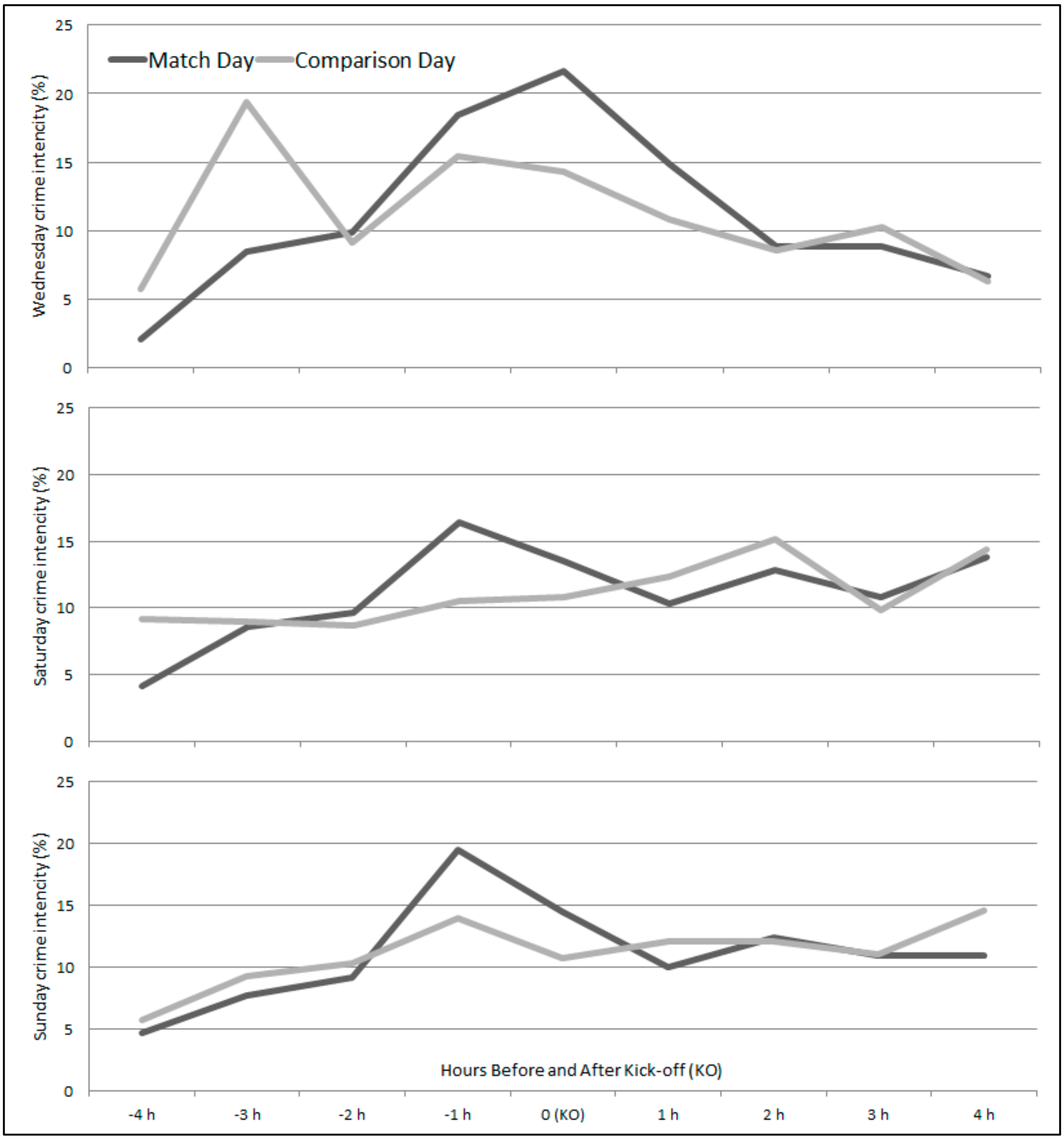


Figure 3. Crime intensity (normalized counts by percentage) per hour before and after the kick-off hour.

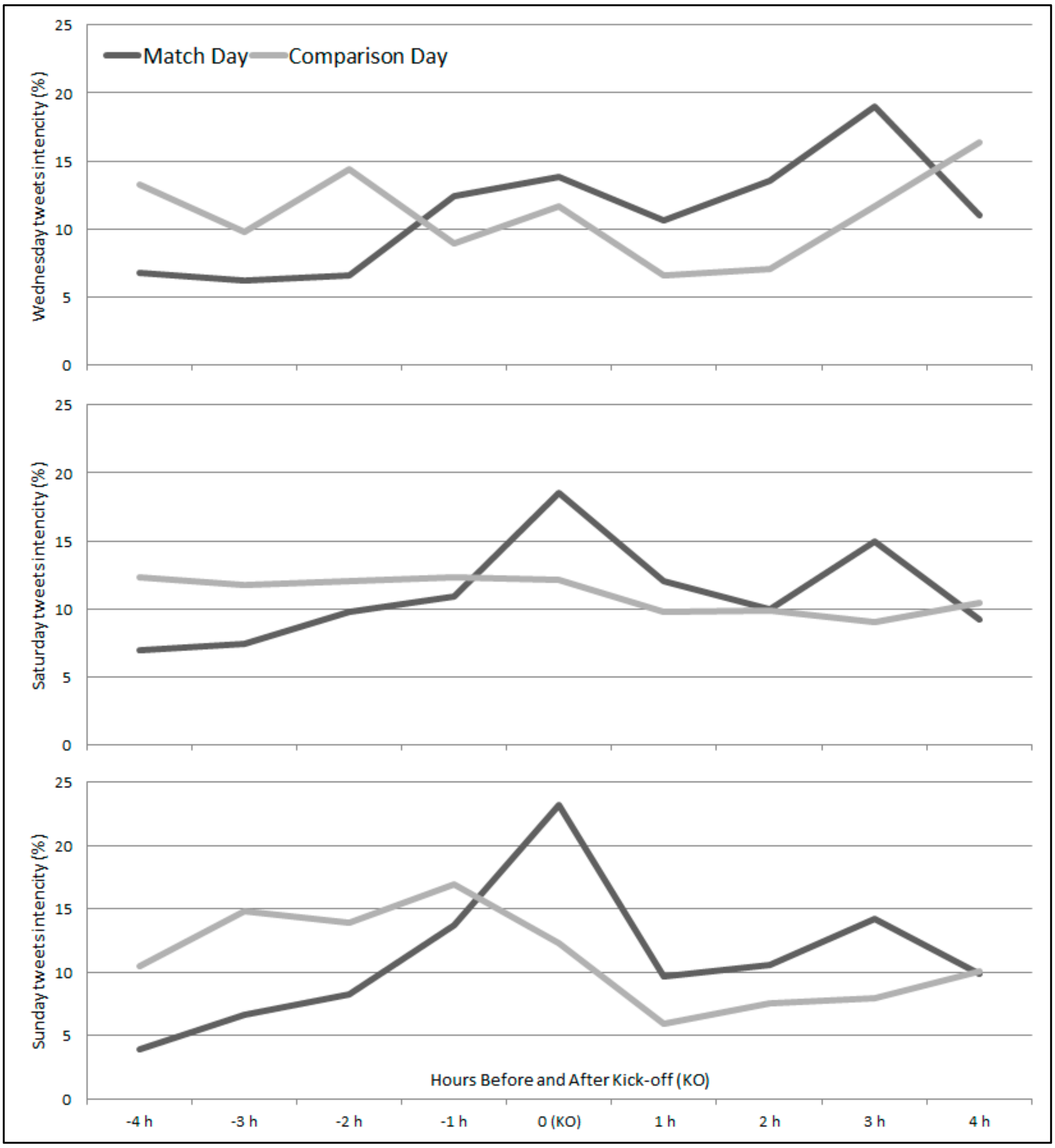


Figure 4. Geotagged tweets intensity (normalized counts by percentage) per hour before and after the kick-off hour.

3.2. Semantic Analysis

Two methods were used to aggregate geotagged tweets. First, we use latent dirichlet allocation (LDA) to extract semantic topics from tweets [63]. LDA assumes that each document (d) can be represented by a probabilistic distribution of topics (z) with the same dirichlet prior, and each topic includes a probabilistic distribution of words (w) [64]. Given a corpus that includes a number of words, LDA follows a generative process (Figure 5). The latent variable θ represents a multinomial distribution of topics in a document. α and β are corpus-level parameters and refer to the prior knowledge about the topic distribution (α), and the words distribution in a topic (β). A lower value of α leads to a higher concentration of topics. For the present study the value of α used equals 0.0001 [64].

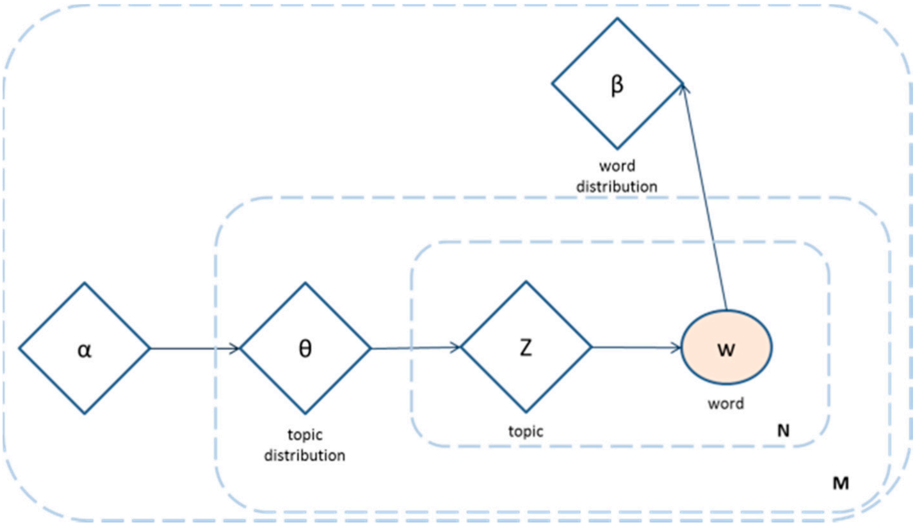


Figure 5. Latent dirichlet allocation (LDA) graphical model according to [63], α and β are corpus-level parameters, θ are document-level variables, and z and w are word-level variables.

Besides selecting the main parameters for the LDA, there are a few ways to find a proper number of topics (k) in a model. We calculate the maximum log-likelihood, given the data, from the harmonic mean example used by Martin Ponweiser [65]. The necessary control parameters for fitting the LDA model were chosen according to Gibbs sampling by Xuan-Hieu Phan and co-authors [66]. Considering the sampling, the best value to use is k = 12. Therefore, from the 12 sampled topics, our goal is to identify tweets, which are part of football topics. For example, the tweet "AVFCOfficial. Hoping for a villa win #keane to score the winner!" may be part of a football topic.

In total 1495 tweets were identified as being related to football, 183 of which were identified for comparison days. We defined topics through a threshold of a minimum of five terms related to football per topic from the 50 most used terms. Determining that words in a detected topic pertained to football was done manually. Even if choosing the threshold of terms was subjective, we argue that the selected topics have a clear connection with football events at Aston Villa's stadium. During match days for the eight-hour timeframe there are no other matches or similar events in the same location or in the immediate vicinity of the stadium, which leads us to conclude that the extracted football terms (Table 2) are most likely related to Aston Villa events. However, Birmingham City Football Club played during some comparison days selected for the tweets dataset, and its stadium's three km radius overlaps with the southern portion of the Aston Villa study area.

Table 2. LDA example of football topic vs other topic word probabilities (* avfc = Aston Villa Football Club).

Football Topic		Other Topic	
Word	**Probability**	**Word**	**Probability**
avfc *	0.021	lol	0.015
game	0.011	just	0.011
villa	0.010	one	0.009
team	0.006	look	0.007
player	0.004	new	0.006

Secondly, we considered an adaptive approach for the extraction of digital violence from tweets: We iterated through each tweet comparing it against a vector of 515 violent words identified from an online dictionary [67], and the name of all crime categories defined by the UK Home Office [68]. A violent word is chosen subjectively from this dictionary if it refers to offences, crime names, and hate crime elements (i.e., racism, xenophobia). Words such as "theft", "kill", "fight", "shooting" are considered in the dataset. 408 (247 Match, 161 Comparison) violent tweets were identified. For example, a tweet such as "Nottingham Forest can seriously burn. Cunting fans." is considered digital violence because it includes a violent word. The violence from tweets posted in specific locations can be a proxy for the mood of the crowd. In addition, those locations can show a more permissive attitude to disorder, where possible offenders are more motivated to commit a crime.

3.3. Spatial Correlation between Crime and Tweets

To reveal spatial clusters in crime data and tweets, and the association between crime and tweet clusters, local indicators of spatial association (LISA)-Local Moran's I, and bivariate Moran's I were calculated using GeoDa [69]. Local indices of spatial autocorrelation have been discussed since the 1990s [70–72] and LISA variables [72] can be used in descriptive statistics. First, we tested each dataset for spatial autocorrelation and subsequently calculated Local Moran's I values (Equations (1) and (2)). These values indicate, whether distributions are spatially random, clustered, or dispersed. Spatial autocorrelation relates to the degree of dependency between the spatial location and the variable measured at that location [73]. To control for the raw data skew and the Moran's I distribution, the "conditional permutation" method is applied. The advantage of this method is that it does not make assumptions about the data [72]. When deriving the Local Moran's I value, deviations between individual observations and their mean value are calculated (Equation (2)) [72]:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \, z_i^2} \tag{1}$$

where n is the number of observations (i.e., counts per polygons), z_i and z_j are deviations of individual observations from the mean, and the summation over j includes only neighbouring values $j \in J_i$

$$z_i = y_i - \overline{y}, z_j = y_j - \overline{y} \tag{2}$$

where $\overline{y}$ is the mean of the variable, y_i is the variable value at a particular location, y_j is the variable value at another location, and w_{ij} is a weight indexing location of i relative to j.

Second, the Bivariate Moran's I was applied to correlate crime occurrences with Twitter activity and to explore the spatial clustering of both variables together. Wartenberg (1985) [74] developed basic ideas about the spatial correlation matrix in a set of multivariate marked points, with the idea to extend the autocorrelation Moran's I to the multivariate case. Anselin et al. (2002) [75] also developed an extension of the Moran's I global index and local LISA indices for the bivariate case from these original ideas (Equation (3)).

$$I\,bv = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i^n \sum_j^n w_{ij} (x_i - \overline{x})(y_i - \overline{x})}{\sum_i^n (x_i - \overline{x})^2} \tag{3}$$

where x_i is the spatially lagged variable value at a particular location and $\overline{x}$ is the statistical mean of the lagged variable. The Queen Contiguity was used in defining the weight matrix and the models' significance was tested using different values of randomization. All other values are defined exactly as in Equation (1).

For the spatial correlation analyses, aggregated crime counts and geotagged tweets per grid cell, from the entire study period (2005–2010 and 2012) are used. In addition, violent and football-related tweets subsets are introduced. When running models in the Geoda software values for parameters are standardized [69,76]. The use of variables for explanatory models varies, which will be discussed in Section 3.5.

3.4. Spatial Analysis Units

For delineating clusters for spatial correlation between crimes and tweets, we use aggregated datasets for the 169 grid cells within the 3 km radius around the stadium. For the correlation analysis a total of 2399 crime occurrences, 6274 tweets, 408 violent tweets, and 1678 football-related tweets are used. Because the geotagged tweets dataset is available from 2012, crimes from the last football season are integrated into an exploratory regression model. The training data includes monthly crimes from August 2009 to April 2010, together with tweets from the same months in 2012. The testing and validating data include May 2010 and 2012, respectively.

When predicting crime, it is necessary to show meaningful results, and to ensure that results are not biased by the study area itself. Approximately 40 per cent of all crimes in the study area concentrate in the 25 grid cells directly surrounding the football stadium, while a higher proportion of grid cells contain no crime.

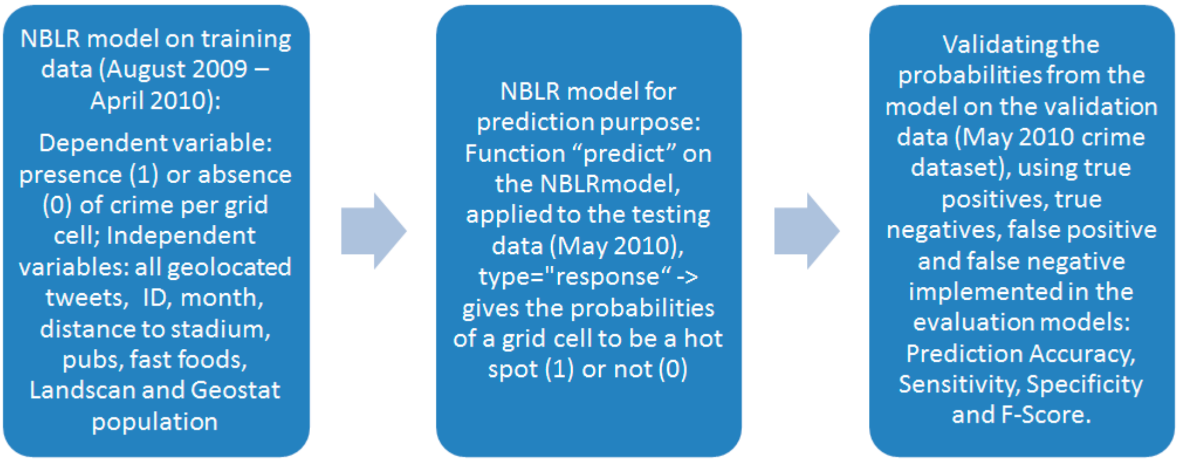
3.5. Explanatory Regression Models

Negative binomial logistic regression models are utilized to determine which variables are statistically significant features for each crime type. However, a feature can be a significant explanatory variable, but not a particularly useful element for prediction purposes [77]. Thus, the analysis is twofold; starting with explanatory models that are built to verify significant variables, then a basic prediction testing is implemented. First, a negative binomial logistic regression (NBLR) [78,79], is used to explain the probability of crime events occurring at the grid cell level. This type of model includes a random term explaining between-subject differences. Logistic regression was used instead of a linear regression because it is explaining or predicting a binary dependent variable and not a continuous one. The Euclidean distance to the stadium (measured in meters), Landscan population, Geostat population, pub density (number of pubs per grid cell), fast food density (number of fast food restaurants per grid cell), total tweet count, violent tweet count, and football-related tweet count for all grid cells were explored as potentially significant explicators or predictors of crime occurrences. It should be noted that Twitter subsets were included in the models one by one and they were never analysed together to avoid multicollinearity, resulting in 20 exploratory models. The same variables were also considered for individual crime types. In this type of explanatory models, the dependent variable is expressed as a binomial value, which indicates the presence (1) or absence (0) of the aggregated crime category. All available datasets are included in the explanatory models. We explore NBLR models based on logistic regression using variables from the aforementioned exploratory analysis, introducing one Twitter subset for each crime type separately for each model (Equation (4)). The NBLR combines the standard Poisson distribution with a gamma variable, defining the variation in the mean event counts (λ_i) of the dependent variable (crime subsets):

$$\mathrm{P}(\mathrm{Y_i} = \mathrm{y_i}) = \frac{\Gamma(\mathrm{y_i} + \varphi)}{\mathrm{y_i}!\Gamma(\varphi)} \frac{\varphi^{\varphi} \boldsymbol{\lambda}_{\mathrm{i}}^{\mathrm{y_i}}}{(\varphi + \boldsymbol{\lambda}_{\mathbf{i}})^{\varphi - \mathrm{y_i}}} \tag{4}$$

where $\mathrm{y_i}$ is the observed outcome (presence or absence of crime in a grid cell), Γ is the gamma function (a continuous version of the factorial function), and φ is the reciprocal of the residual variance of underlying mean counts, α [78].

Second, the same type of regression model is used in a basic prediction testing. Because of data limitation, nine months of data are used for training the prediction algorithm (August 2009 to April 2010) and one month (May 2010) for testing and validating. Since this is not a straightforward prediction additional information about the relationship with the explanatory analysis is provided below. The “stats” package in R is used [80], especially the function “predict”, and the type “response”, which calculates the predicted probabilities. Predictions are validated using real crime values from May 2010, including Prediction Accuracy, Sensitivity, Specificity, and F-Score. Figure 6 shows an example of the model, including geolocated tweets:



NBLR model on training data (August 2009 – April 2010):

Dependent variable: presence (1) or absence (0) of crime per grid cell; Independent variables: all geolocated tweets, ID, month, distance to stadium, pubs, fast foods, Landscan and Geostat population

NBLR model for prediction purpose: Function "predict" on the NBLRmodel, applied to the testing data (May 2010), type="response" -> gives the probabilities of a grid cell to be a hot spot (1) or not (0)

Validating the probabilities from the model on the validation data (May 2010 crime dataset), using true positives, true negatives, false positive and false negative implemented in the evaluation models: Prediction Accuracy, Sensitivity, Specificity and F-Score.

Figure 6. Example of negative binomial logistic regression (NBLR) usage for this study.

Given the spatial nature of the data, it is possible that there are issues of spatial dependency. This can be a problem as with most statistical tests, and when spatial autocorrelation is present between spatial residuals, it may lead to errors of statistical inference, if not corrected [79,81]. To test for this, the approach adopted by [82] is followed and the Moran's I statistic for negative binomial residuals is computed post-estimation.

4. Results

4.1. Geospatial Statistics

In this section, results are discussed with an emphasis on relationships between geotagged tweets and crime occurrences. When looking at the distribution of crime events around the Aston Villa stadium, different patterns for match days and comparison days can be observed (Figure 7). The same colour pallet was used to compare the spatial distribution of match and non-match day crime and tweets (Figure 8). As expected, crime occurs with greater frequency in the immediate vicinity of the stadium (a maximum of 80 crimes in the stadium grid cell) on match days, but also in the area northwest of the stadium, where Birmingham City University and a shopping mall are located. The distribution of crime events for comparison days appears similar with a hotspot in the area northwest of the stadium. Criminal damage is more pervasive in the area next to the stadium (a maximum of 14 crimes); the northwest area shows a high-density hotspot (38 crimes in total) for theft and handling. A large number of violence against the person occurred at the stadium location, and in the western area (Figure 7). Theft and violence against the person clustered around the stadium, while criminal damage clustered a little further away, which can be explained by temporal crime patterns. After the match, the tendency of security officers is to get the crowd away from the stadium which leads to a majority of football hooliganism, such as criminal damage, taking place after the match [83]. It could also be that fans go out for drinking after the match and then they may be prone for disruptive behaviour, destroying public goods and disturbing the peace.

For match days, Figure 8 shows an overall increase in the volume of tweets in the cell where the stadium is located (894 tweets). In contrast, comparison days show the highest densities of tweets in the south of the study area (138 tweets). There is a similarly high concentration of football-related tweets in the centre and in the south areas during match days, and no football-related tweets during comparison days (Figure 8). As expected, a higher number of violent tweets occur in close proximity to the stadium on match days (Figure 8). On comparison days, a high density of violent tweets is located in the southeast of the stadium, in Lozells area, which has a high population density, with several schools, restaurants, and shops.

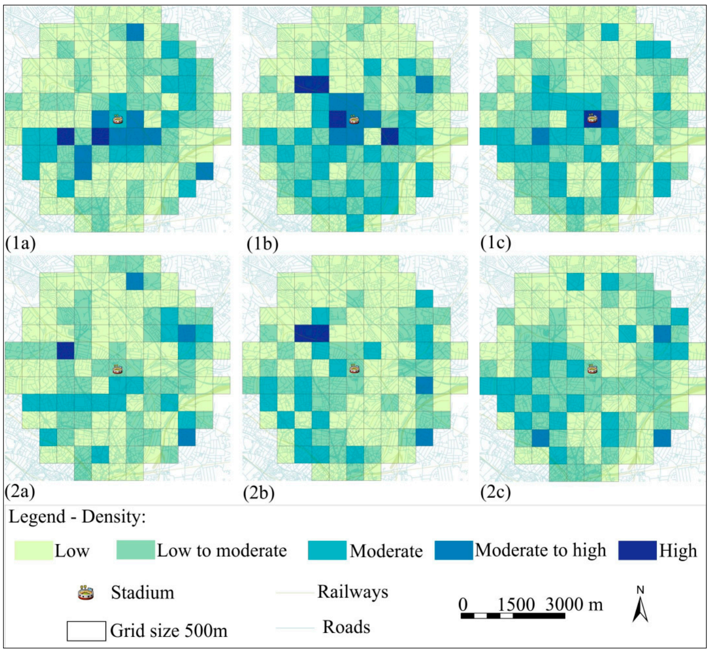


Figure 7. Density for each crime type (**a**) criminal damage, (**b**) theft and handling, and (**c**) violence against the person; (**1**) match days, (**2**) comparison days (using the Natural Jenks classification method).

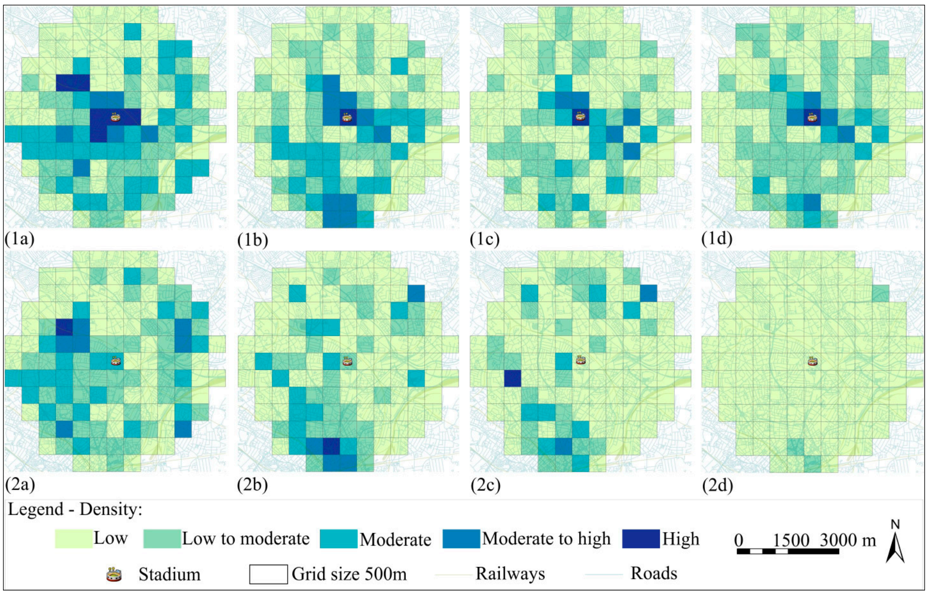


Figure 8. Density maps (**a**) amalgamated crimes, (**b**) geotagged tweets, (**c**) violent tweets, and (**d**) football-related tweets; (**1**) match days, (**2**) comparison days (using the Natural Jenks classification method).

The observed spatial density of crime and its possible relationship to Twitter activity is explored within the three km buffer around the stadium before, during, and after the match. It is important to point out that some cells have zero values, which may influence results of correlation statistics.

First, we consider the spatial autocorrelation of all variables (each individual crime type and each Twitter subset), calculating the univariate Moran's I value (Table 3).

Table 3. Spatial Autocorrelation Moran's I.

All Crimes		Criminal Damage		Theft and Handling		Violence against the Person		All Tweets		Violent		Football Topic	
M	C	M	C	M	C	M	C	M	C	M	C	M	C
0.45	0.36	0.35	0.21	0.34	0.21	0.26	0.27	0.14	0.44	0.17	0.16	0.11	0.21

All spatial autocorrelation values are positively correlated, indicating that neighbouring cells have similar tweet or crime counts. Results of the Moran's I analysis show the most clustered patterns for the amalgamated crime dataset (0.45) for match days, and for the all tweet dataset (0.44) for comparison days. Theft and handling has a moderate Moran's I value for match days (0.34) which may be explained by the greater number of opportunities for theft during match days in the broader area around the stadium. Football-related tweets show a low correlation during match days (0.11). Violent tweets have a Moran's I index of 0.17 during match days and 0.16 during comparison days. All values are calculated based on 999 permutations and are statistical significant (p's < 0.001).

Secondly, despite differences shown in the spatial autocorrelation for each variable, the bivariate spatial autocorrelation between crimes and each Twitter subset shows a higher correlation on match days (Table 4). This result confirms H1 insofar as the spatial distribution of crimes and tweets is more spatially clustered on match days, and that tweets are strongly correlated with crimes. All values are calculated based on 999 permutations and are statistical significant (p < 0.001).

Table 4. Bivariate spatial autocorrelation between crime density and tweets.

		All Crimes	Criminal Damage	Theft and Handling	Violence against the Person
Match Days	**All tweets**	0.26	0.24	0.24	0.18
	Violent tweets	0.29	0.25	0.27	0.19
	LDA football topic	0.25	0.23	0.22	0.16
Comparison Days	**All tweets**	0.21	0.12	0.16	0.19
	Violent tweets	0.14	0.09	0.10	0.15
	LDA football topic	0.17	0.12	0.12	0.16

Results from Table 4 and Figure 9 show that violent tweets are more spatially correlated with the amalgamated crime category than all tweets and football-related tweets. Conversely, the spatial relationship between violent tweets and crime on comparison days is low with bivariate LISA values ranging from 0.10 to 0.14. On this basis, it appears that the time in which a football match is played is related to both writing violent tweets and the occurrence of crime. Bivariate LISA clusters between crimes and tweets subsets are shown in Figure 8. For all crime and tweets subsets, spatial patterns vary between match and comparison days, showing higher correlation during match days. For match days, the highest densities of the crime and geotagged tweet datasets are found in the grid cell where the stadium is located.

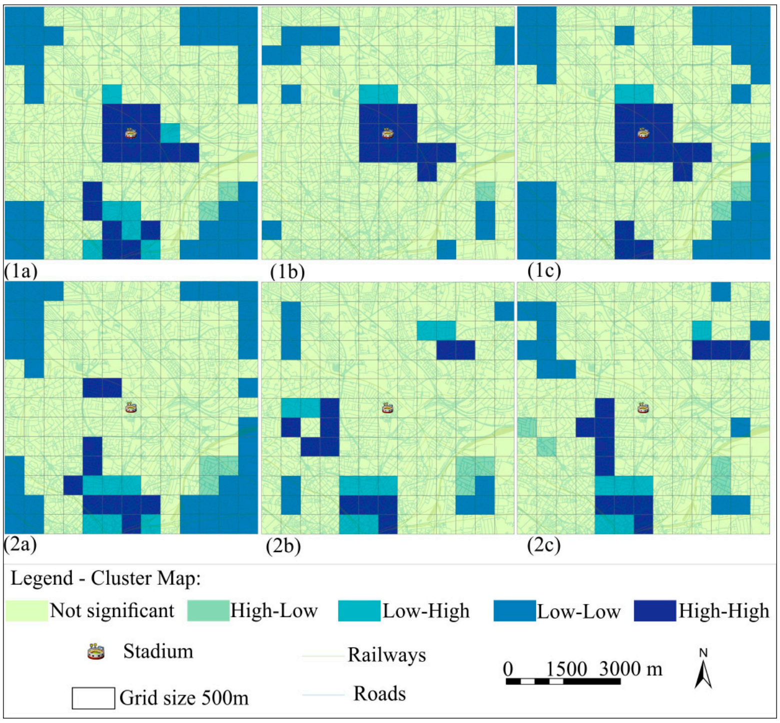


Figure 9. Bivariate LISA clusters between crime density and (**a**) tweets density, (**b**) violent tweets density, and (**c**) football topic tweets density; (**1**) match days, (**2**) comparison days.

4.2. Explanatory Regression Models

The third hypothesis posits that geotagged tweets would be a useful factor for crime explanatory models on match days. To test this, NBLR models were considered for each crime type and also for amalgamated crime types. To avoid multicollinearity, Twitter datasets are included in the regression models one by one, resulting in four models for each crime type, for a total of 20 models. Independent variables are the grid cell ID (representing the location), Euclidean distance to the stadium, Landscan and Geostat population density, pub density, fast food density, geotagged tweets count, violent tweet count, and football-related tweet count.

While testing the significance of the aforementioned explanatory variables, all models showed one main significant feature, the distance to the stadium ($p = 0.0001$). For the amalgamated crime types and theft and handling, the fast food and pubs were significant ($p \leq 0.05$) while for criminal damage the distance to the stadium is the only important factor ($p = 0.0001$). For aggregated crimes, the Landscan population is a significant variable ($p \leq 0.05$). Geolocated Twitter datasets were only significant for the violence against the person crime category ($p = 0.012$). The same is true for football-related tweets ($p = 0.064$), along with the grid cell ID representing the location ($p = 0.015$).

All models were evaluated using: the prediction accuracy (PA), by dividing the count of correctly determined crime locations (true positives) by the total number of observations; the sensitivity (recall), by dividing the true positives by true and false positives; the specificity (precision), by dividing the true negatives by true negatives and false positives; and the F-score, which is defined as the harmonic mean of precision and recall. The sensitivity shows the percentage chance that the model will correctly identify a location (grid cell) where crime is actually occurring, while the specificity shows the percentage chance that the model will correctly identify a location where there are no crimes.

Results show that after including all geolocated tweets in the amalgamated crime model, evaluation parameters are decreasing, and by introducing football-related and violent tweets, values remain unchanged (Tables 5 and 6). For criminal damage, the usage of violent tweets in the model helped increasing the precision (from 0.975 to 0.981), F-Score (from 0.970 to 0.972), and PA values (from

0.941 to 0.947). For theft and handling, inclusions of violent tweets in the model result in an increase in the PA value (from 0.911 to 0.917, when compared to the model without Twitter data). In addition, by only including the significant variables detected from the explanatory model for theft and handling (fast foods, pubs and distance to the stadium) the recall and F-Score increase slightly (from 0.957 to 0.969 compared to the no tweets model with 0.956 and 0.953) The addition of football-related tweets for violence against the person crime type increased the precision (0.994), F-score (0.969), and PA (0.941), while violent tweets have the same impact as football-related tweets for the precision score. However, there are no drastic changes in the models, which may open the path to a more detailed investigation considering more data. We analysed the models' residuals to test for the presence or absence of spatial autocorrelation to ensure that the underlying model assumption that errors are independent was met, and the residuals were not spatially correlated.

Table 5. NBLR model evaluation (* mentioned in Section 4.2).

		All Crimes	**Criminal Damage**	**Theft and Handling**	**Violence against the Person**
No tweets	Sensitivity (recall)	0.942	0.975	0.950	0.993
	Specificity (precision)	0.948	0.964	0.956	0.916
	F-measure	0.945	0.970	0.953	0.953
	PA	0.899	0.941	0.911	0.912
All geocoded tweets	Sensitivity (recall)	0.936	0.975	0.950	0.987
	Specificity (precision)	0.948	0.964	0.956	0.946
	F-measure	0.942	0.970	0.953	0.967
	PA	0.893	0.941	0.911	0.935
Football tweets	Sensitivity (recall)	0.942	0.975	0.944	0.994
	Specificity (precision)	0.948	0.964	0.956	0.946
	F-measure	0.945	0.970	0.950	0.969
	PA	0.899	0.941	0.905	0.941
Violent tweets	Sensitivity (recall)	0.942	0.981	0.950	0.994
	Specificity (precision)	0.948	0.964	0.962	0.940
	F-measure	0.945	0.972	0.956	0.966
	PA	0.899	0.947	0.917	0.935
Just significant variables *	Sensitivity (recall)	0.942	0.975	0.945	0.988
	Specificity (precision)	0.942	0.958	0.969	0.928
	F-measure	0.942	0.966	0.957	0.957
	PA	0.893	0.935	0.917	0.917

Table 6. NBLR models coefficients (p = 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1).

		Int.	**ID**	**Fast Food**	**Pub**	**Dist.**	**Pop Lands**	**Pop Geost**	**Tweets**	**Football Tweets**	**Violent Tweets**
All crimes	Estimate	−1.420	−0.001	0.164	0.146	−0.001	0.000	0.000	0.006	0.002	0.017
	Std. Error	0.464	0.002	0.060	0.076	0.000	0.000	0.000	0.005	0.011	0.090
	z-value	−3.063	−0.348	2.743	1.916	−6.631	1.968	0.133	1.116	0.201	0.188
	p-value	0.002	0.728	0.006	0.055	0.000	0.049	0.894	0.264	0.841	0.851
		**		**	.	***	*				
Criminal Damage	Estimate	−2.627	0.003	0.125	0.123	−0.001	0.000	0.000	−0.015	−0.031	−0.214
	Std. Error	0.890	0.004	0.121	0.174	0.000	0.000	0.000	0.019	0.039	0.242
	z-value	−2.951	0.841	1.033	0.705	−4.275	0.407	0.832	−0.816	−0.792	−0.885
	p-value	0.003	0.400	0.302	0.481	0.000	0.684	0.405	0.415	0.428	0.376
		**				***					
Theft and Handling	Estimate	−2.350	0.004	0.175	0.252	−0.001	0.000	0.000	−0.003	−0.023	−0.090
	Std. Error	0.684	0.003	0.085	0.105	0.000	0.000	0.000	0.009	0.024	0.143
	z-value	−3.434	1.157	2.055	2.395	−5.732	1.488	0.164	−0.292	−0.935	−0.632
	p-value	0.001	0.247	0.040	0.017	0.000	0.137	0.870	0.770	0.350	0.527
		***		*	*	***					
Violence against the Person	Estimate	−1.818	−0.010	0.120	0.078	−0.001	0.000	0.000	0.016	0.023	0.169
	Std. Error	0.812	0.004	0.103	0.131	0.000	0.000	0.000	0.006	0.012	0.118
	z-value	−2.240	−2.521	1.160	0.599	−4.017	1.448	0.372	2.497	1.849	1.436
	p-value	0.025	0.012	0.246	0.549	0.000	0.148	0.710	0.013	0.064	0.151
		*	*			***			*	.	

5. Discussion

5.1. Hypothesis 1—Correlation between Tweets and Crimes around a Football Stadium

The first hypothesis (H1) stated that Twitter activity is spatially correlated with the density of crimes within 3 km around the football stadium during a sporting match event. Our research highlights the discrepancies found when correlating crimes and tweets on match days and non-match days. For example, we found a positive correlation between crimes and tweets on match days while non-match days have almost no correlation. Given the co-occurrence in the spatial density of crimes and tweets and the correlation between them, an important result is the difference between fixed and shifting bivariate LISA clusters between match and comparison days.

The spatial distribution of crime events is routinely researched in various fields, including but not limited to criminology, and so the introduction of certain types of social media analyses may pay dividends in helping understand these patterns and their predictability. Researchers and criminal justice practitioners should be aware of the spatial relationship that exists between geotagged tweets and crimes, which in turn could be useful in finding elements of the urban landscape that may be crime attractors or generators (e.g., more violent tweets and crimes in and around areas with pubs). The strong spatial relationship identified between crime and tweets for football events is suitable for analysing the probability of disorder. Geotagged tweets may be considered a useful factor in determining an effective response when an event may increase the likelihood of crime being committed. However, this would be valid just for specific days when—such as in our case—a football match occurs. Our case study shows that during non-match days Twitter activity may have some minimal relationship as explanatory variable for determining crime density. One reason can be related to the amount of data; as in our study we selected comparison days with no events occurring in the relatively small study area. In contrast, previous literature highlights the utility of social media for criminological research in larger areas for connected periods of time, such as weeks or months. Another reason for this may be that crowd-based events are highly discussed in social media, and because the conditions for a crime to occur at these events are more likely to happen, since there are many possible offenders, targets, and insufficient guardianship. Therefore, the utility of the crime-tweet relationship appears to further enrich the picture of crime occurrences for events. However, at the present time it is still difficult to establish a definite cause and effect relationship, compared to establishing correlation. Correlation results shown in this analysis can be the base for further research into exploring whether a cause and effect relationship exists. The objective of this research is to identify the extent to which one variable is related to another variable, namely whether crime is related to tweets.

5.2. Hypothesis 2—Correlation between Violent Tweets and Football-Related Tweet Densities

The second hypothesis (H2) stated that violent tweets are more highly correlated with crime occurrences than all tweets. We have introduced an adapted framework for testing the relationship between crime and social media by extracting digital violence from geotagged tweets, creating a violent tweets subset. One advantage of this approach is the automated semantic analysis process. The relationship between violent tweets during match days and crime occurrences suggests that violent tweets may indeed be a suitable proxy for determining the crowd mood and the more permissive attitude to disorder. Bivariate spatial autocorrelation values are slightly higher by using the violent tweets subset compared with all geotagged tweets during match days, which might indicate a more negative texting pattern around the stadium.

An individual perceives space through vision, and this visual perception alters how the individual interacts within a given space. Tweets occur within the digital space of an event, which individuals may observe and interact with through technology. The condition of the digital space may be thought of as a symbol of the temporary perception individuals have of the physical space they occupy. Thus, it may be that the abovementioned correlations demonstrate that violent tweets increase the permissibility of attitudes on disorder in a physical space temporarily, and in turn this precipitates potential offenders to act out violently [84]. The possible effect on opportunities for crime is temporally sensitive, and appears to only last while digital violence occurs. In other words, once violent tweets stop being posted and shared during an event, the behaviour of each individual is expected to reflect the positive change, that is, crime occurrences decrease. We expect to extract more digital violent clusters within future research.

5.3. Hypothesis 3—Geotagged Tweets for Explaning the Presence or Absence of Crime around Stadia

The third hypothesis (H3) stated that geotagged tweets can be a useful variable in spatially explaining the presence or absence of crime around stadia when a football game is played. We argue that the spatial correlation between datasets is an important base in constructing an explanatory model, which in the future can be useful for prediction. We show that not all spatially correlated variables make a difference in the prediction accuracy evaluation measure. Moreover, research shows that variables uncorrelated with the dependent variable can be good predictors [73]. Through this basic explanatory step, we determine that creating a complex predictive model needs a larger number of independent variables and a larger volume of data.

An unexpected result shows that geotagged tweet activity does not increase the prediction accuracy of for all crime categories. Tweets can be highly correlated with crimes, but the estimated influence varies across crime categories. One explanation for this may be related to the spatial distribution of different crime types or with an unexpected distribution during the prediction days. For example, a higher probability for pickpocketing, a subclass of theft and handling, most likely occurs in crowded locations, where likely targets do not pay attention to their assets, such as bags or phones. People use their phone to tweet, which may provide a visual cue to motivated offenders about the presence of a potential target. Therefore, the area where tweets are posted may be a target for motivated offenders. This is one possible explanation for why models for theft and handling that include the violent tweets information introduce slightly higher prediction accuracy, than the simple model without tweets. Football-related tweets increase evaluation results for violence against the person crime type, which is an interesting aspect that should be further analysed. Football tweets are not the cause of crime, but they can be used to explore spatial crime hotspots in this case study.

5.4. Limitations

There are a few specific limitations associated with the current study. The first and most obvious limitation is that the incongruent temporal period for the available police-recorded crime data and geotagged tweets utilized. More recent police-recorded crime data are available from the UK police service. However, the 2012 data are aggregated by month and at a lower spatial resolution, making it untenable for use in this study. Despite this, and in light of recent criminological findings on the temporal stability of crime hotspots [85], we are confident that the results remain reflective of the likely pattern of crime around Villa Park. Moreover, as mentioned above, changes in the urban land use likely to influence the underlying spatial distribution of crime are slow to change [53].

The Twitter platform experienced quick growth starting from 2010 when there were ~30 million active users per month worldwide, which increased to ~320 million in 2016. The same Twitter platform from which the dataset of this present study was collected, had ~120 million users [86], thus containing a sufficiently large volume of geotagged tweets for a robust analysis. However, this particular limitation does rule out the possibility of a time-series analysis over the five football seasons.

In addition, we are aware of the biases introduced into our analysis because of comparing match days and comparison days from one year of geotagged tweets with five years of crime data. Future research needs to consider social media data falling in line with crime data. A further limitation is the distribution of crimes per month, including just match days with mostly two matches per month and a sixteen-hour timeframe analysed (eight hours for match days and eight hours for comparison days, so both couple match and comparison days sum up to a sixteen-hour timeframe). This limitation shows multiple cells with no crime incidents occurring, which may bias the output. Therefore, the predictive analysis of the study was applied to the centre of the study area where more crimes occurred. These limitations could potentially be mitigated by using alternative datasets with crime data collected for the same time period as geotagged tweets. In addition, geotagged tweets spanning over more than one year could be beneficial for understanding their seasonal distributions and temporal patterns in the relationship between the two data sources.

6. Conclusions

We provide strong evidence for increased crime-tweet correlation within a three km radius around the Aston Villa stadium during match days for an eight-hour timeframe. This research has introduced a framework for using geotagged tweets and specific subsets, such as violent tweets, to deduce the spatial-temporal relationship with crime events in and around a football stadium on match and comparison days. In addition, this study tested if geotagged tweets may be a useful explanatory variable in crime prediction models. The approach may be applicable to any other crime categories and can be tested and implemented for different event types.

While some gaps in knowledge regarding the use of geotagged tweets in crime analysis and prediction have been answered, many challenges remain. Acknowledging the shortcomings, we found positive spatial relationships between criminal damage, thefts, and violence against person incidents and three Twitter subsets (all tweets, violent tweets, and football-related tweets). In addition to showing similarities in the spatial distribution of crime and geotagged tweets, this study addressed the use of Twitter as a factor for crime prediction on football match days.

While the spatial correlation between violent tweets and crimes during match days was stronger than with other datasets, their reported interactions in explanatory and prediction models were not consistent with initial expectations. We believe that the utility of geotagged tweets and their subsets as independent predictors in a NBLR model is questionable and suggests a closer examination with a larger dataset that also includes more predictors.

This research illustrates the potential of using social media data to understand spatial crime patterns around a football stadium. Future research in this area should consider conducting seasonal crime pattern analyses, to extend what has been gleaned by using match and comparison days as the unit of analysis. Other research may seek to extend the boundary of the study area to encompass areas that may also experience a change in crime due to contrasting ecological conditions brought about by a football match or other large-scale event. Further, given the variability between volumes of crime on weekdays as opposed to weekends, future studies may seek to model these separately as dependent variables.

Acknowledgments: This research was funded by the Austrian Science Fund (FWF) through the Doctoral College GIScience at the University of Salzburg (DK W 1237-N23). We would like to express our gratitude to the Austrian Science Fund (FWF) for supporting the project "Urban Emotions", reference number I-3022.

Author Contributions: Alina Ristea.conceived and designed all experiments, analysed the data and wrote the paper. Justin Kurland helped designing the paper and wrote supplementary information. Bernd Resch gave technical support and helped with the conceptual framework. Bernd Resch and Michael Leitner supervised the work, gave new ideas, and edited the manuscript. Chad Langford helped design the initial ideas, and gave conceptual advice. All authors discussed the results and implications and commented on the manuscript at all stages.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Caruso, R.; Di Domizio, M. International hostility and aggressiveness on the soccer pitch: Evidence from european championships and world cups for the period 2000–2012. *Int. Area Stud. Rev.* **2013**, *16*, 262–273. [CrossRef]
2. Hopkins, M.; Treadwell, J. *Football Hooliganism, Fan Behaviour and Crime: Contemporary Issues*; Palgrave Macmillan: London, UK, 2014; ISBN 978-1-137-34797-8.

3. Montolio, D.; Planells, S. *Measuring the Negative Externalities of a Private Leisure Activity: Hooligans and Pickpockets around the Stadium*; Working Paper; Institut d'Economia de Barcelona: Barcelona, Spain, 2015.
4. Kurland, J.; Tilley, N.; Johnson, S.D. The football 'hotspot'matrix. In *Football Hooliganism, Fan Behaviour and Crime: Contemporary Issues*; Palgrave Macmillan: London, United Kingdom, 2014; pp. 21–48, ISBN 978-1-137-34797-8.
5. Gratton, C. The peculiar economics of english professional football. *Soccer Soc.* **2000**, *1*, 11–28. [CrossRef]
6. Santo, C. The economic impact of sports stadiums: Recasting the analysis in context. *J. Urban Aff.* **2005**, *27*, 177–192. [CrossRef]
7. Scholtens, B.; Peenstra, W. Scoring on the stock exchange? The effect of football matches on stock market returns: An event study. *Appl. Econ.* **2009**, *41*, 3231–3237. [CrossRef]
8. Bell, A.R.; Brooks, C.; Matthews, D.; Sutcliffe, C. Over the moon or sick as a parrot? The effects of football results on a club's share price. *Appl. Econ.* **2012**, *44*, 3435–3452. [CrossRef]
9. Königstorfer, J.; Uhrich, S. Riding a rollercoaster: The dynamics of sports fans' loyalty after promotion and relegation. *Mark. ZFP* **2009**, *31*, 71–84. [CrossRef]
10. Brent Ritchie, J. Assessing the impact of hallmark events: Conceptual and research issues. *J. Travel Res.* **1984**, *23*, 2–11. [CrossRef]
11. Kain, K.J.; Logan, T.D. Are sports betting markets prediction markets? Evidence from a new test. *J. Sports Econ.* **2014**, *15*, 45–63. [CrossRef]
12. Cohen, L.E.; Felson, M. Social change and crime rate trends: A routine activity approach. *Am. Sociol. Rev.* **1979**, *44*, 588–608. [CrossRef]
13. Marie, O. Police and thieves in the stadium: Measuring the (multiple) effects of football matches on crime. *J. R. Stat. Soc. Ser. A (Stat. Soc.)* **2016**, *179*, 273–292. [CrossRef]
14. Planells-Struse, S.; Montolio, D. *Should football teams be taxed? In Determining Crime Externalities from Football Matches*; Working Paper; Institut d'Economia de Barcelona: Barcelona, Spain, 2014.
15. Breetzke, G.; Cohn, E.G. Sporting events and the spatial patterning of crime in south africa: Local interpretations and international implications. *Can. J. Criminol. Crim. Just.* **2013**, *55*, 387–420. [CrossRef]
16. Kurland, J.; Johnson, S.; Tilley, N. Hotspotting and football violence: Current statistics and implications for prevention. In *The Wiley Handbook of Violence and Aggression*; John Wiley & Sons: Hoboken, NJ, USA, 2017; Volume 3, pp. 1–15.
17. Yu, Y.; Mckinney, C.N.; Caudill, S.B.; Mixon, F.G., Jr. Athletic contests and individual robberies: An analysis based on hourly crime data. *Appl. Econ.* **2016**, *48*, 723–730. [CrossRef]
18. Chen, S.-C.; Shyu, M.-L.; Zhang, C.; Luo, L.; Chen, M. Detection of soccer goal shots using joint multimedia features and classification rules. In Proceedings of the 4th International Workshop on Multimedia Data Mining (MDM/KDD'03), Washington, DC, USA, 24–27 August 2003.
19. Chen, S.-C.; Shyu, M.-L.; Chen, M.; Zhang, C. A decision tree-based multimodal data mining framework for soccer goal detection. In Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), Taipei, Taiwan, 27–30 June 2004; Volume 1, pp. 265–268. [CrossRef]
20. Zhao, S.; Zhong, L.; Wickramasuriya, J.; Vasudevan, V. *Human as Real-Time Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games*; Rice University and Motorola Labs: Houston, TX, USA, 2011.
21. Corney, D.; Martin, C.; Göker, A. Spot the ball: Detecting sports events on twitter. In *Advances in Information Retrieval*; Springer: Berlin, Germany, 2014; pp. 449–454, ISBN 3319060279.
22. Yu, Y.; Wang, X. World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans' tweets. *Comput. Hum. Behav.* **2015**, *48*, 392–400. [CrossRef]
23. Chakrabarti, D.; Punera, K. *Event Summarization Using Tweets*; ICWSM: Barcelona, Spain, 2011; pp. 66–73.
24. Nichols, J.; Mahmud, J.; Drews, C. Summarizing sporting events using twitter. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, Lisbon, Portugal, 14–17 Feruary 2012; pp. 189–198.
25. Ristea, A.; Langford, C.; Leitner, M. Relationships between crime and twitter activity around stadiums. In Proceedings of the 25th International Conference on Geoinformatics, Buffalo, NY, USA, 2–4 August 2017; pp. 1–5.
26. Kampakis, S.; Adamides, A. Using twitter to predict football outcomes. *arXiv* **2014**, arXiv:1411.1243.
27. Sinha, S.; Dyer, C.; Gimpel, K.; Smith, N.A. Predicting the nfl using twitter. *arXiv* **2013**, arXiv:1310.6998 .

28. Bendler, J.; Brandt, T.; Wagner, S.; Neumann, D. Investigating Crime-To-Twitter Relationships in Urban Environments-Facilitating a Virtual Neighborhood Watch. In Proceedings of the ECIS 2014, Tel Aviv, Israel, 9–11 June 2014.
29. Featherstone, C. The Relevance of Social Media as it Applies in South Africa to Crime Prediction. In Proceedings of the IST-Africa Conference and Exhibition, Nairobi, Kenya, 29–31 May 2013; pp. 1–7.
30. Featherstone, C. Identifying vehicle descriptions in microblogging text with the aim of reducing or predicting crime. In Proceedings of the International Conference on Adaptive Science and Technology (ICAST), Pretoria, South Africa, 25–27 November 2013; pp. 1–8.
31. Malleson, N.; Andresen, M.A. Exploring the impact of ambient population measures on london crime hotspots. *J. Crim. Just.* **2016**, *46*, 52–63. [CrossRef]
32. Malleson, N.; Andresen, M.A. The impact of using social media data in crime rate calculations: Shifting hot spots and changing spatial patterns. *Cartogr. Geogr. Inf. Sci.* **2015**, *42*, 112–121. [CrossRef]
33. Caplan, J.M.; Kennedy, L.W.; Miller, J. Risk terrain modeling: Brokering criminological theory and gis methods for crime forecasting. *Just. Q.* **2011**, *28*, 360–381. [CrossRef]
34. Corso, A.J. Toward predictive crime analysis via social media, big data, and gis spatial correlation. In *iConference 2015*; iSchools: Newport Beach, CA, USA, 2015.
35. Wang, X.; Gerber, M.S.; Brown, D.E. Automatic crime prediction using events extracted from twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*; Yang, S.J., Greenbeg, A.M., Endsley, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 231–238.
36. Alruily, M. *Using Text Mining to Identify Crime Patterns from Arabic Crime News Report Corpus*; DeMontfort University: Leicester, UK, 2012.
37. Wang, M.; Gerber, M.S. Using twitter for next-place prediction, with an application to crime prediction. In Proceedings of the 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, South Africa, 7–10 December 2015; pp. 941–948.
38. Gerber, M.S. Predicting crime using twitter and kernel density estimation. *Decis. Support Syst.* **2014**, *61*, 115–125. [CrossRef]
39. Perry, W.L. *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*; Rand Corporation: Santa Monica, CA, USA, 2013; ISBN 0833081551.
40. Vantara, H. Hitachi Data Systems Unveils New Advancements in Predictive Policing to Support Safer, Smarter Societies. 2017. Available online: https://www.hitachivantara.com/en-us/news-resources/press-releases/2015/gl150928.html (accesed on 7 August 2017).
41. Al Boni, M.; Gerber, M.S. Area-specific crime prediction models. In Proceedings of the 15th IEEE International Conference on Machine Learning And Applications (ICMLA), Orange, CA, USA, 18 December 2016; pp. 671–676.
42. Al Boni, M.; Gerber, M.S. Predicting crime with routine activity patterns inferred from social media. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, 2 August 2016.
43. Sathyadevan, S.; Devan, M.; Surya Gangadharan, S. Crime analysis and prediction using data mining. Proceedings of IEEE 1st International Conference on Networks and Soft Computing, Guntur, India, 19–20 August 2014; pp. 406–412.
44. Mookiah, L.; Eberle, W.; Siraj, A. Survey of crime analysis and prediction. In Proceedings of the International Conference of the Florida AI Research Society (FLAIRS), Hollywood, FL, USA, 6 April 2015.
45. Williams, M.L.; Burnap, P.; Sloan, L. Crime sensing with big data: The affordances and limitations of using open source communications to estimate crime patterns. *Br. J. Criminol.* **2016**, *57*, 320–340. [CrossRef]
46. Kurland, J. The Ecology of Football-Related Crime and Disorder. Ph.D. Thesis, University College London, London, UK, 2014.
47. Ratcliffe, J.H. Damned if you don't, damned if you do: Crime mapping and its implications in the real world. *Policing Soc.* **2002**, *12*, 211–225. [CrossRef]
48. Morstatter, F.; Pfeffer, J.; Liu, H.; Carley, K.M. Is the sample good enough? Comparing data from twitter's streaming api with twitter's firehose. In Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM, Cambridge, MA, USA, 8–11 July 2013; AAAI Press: Cambridge, MA, USA; pp. 400–408.

49. Li, L.; Goodchild, M.F.; Xu, B. Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 61–77. [CrossRef]
50. Zhang, Z.; Ni, M.; He, Q.; Gao, J. *Mining Transportation Information from Social Media for Planned and Unplanned Events*; University at Buffalo: Buffalo, NY, USA, 2016; pp. 1–68.
51. Anselin, L.; Williams, S. Digital neighborhoods. *J. Urban. Int. Res. Placemak. Urban Sustain.* **2016**, *9*, 305–328. [CrossRef]
52. Kinney, J.B.; Brantingham, P.L.; Wuschke, K.; Kirk, M.G.; Brantingham, P.J. Crime attractors, generators and detractors: Land use and urban crime opportunities. *Built Environ.* **2008**, 62–74. [CrossRef]
53. Wegener, M. Overview of land use transport models. In *Handbook of Transport Geography and Spatial Systems*; Hensher, D.A., Button, K.J., Haynes, K.E., Stopher, P.R., Eds.; Emerald Group Publishing Limited: Bingley, UK, 2004; pp. 127–146, ISBN 978-0-080-44108-5.
54. Andresen, M.A. An area-based nonparametric spatial point pattern test: The test, its applications, and the future. *Methodol. Innov.* **2016**, *9*. [CrossRef]
55. Andresen, M.A.; Malleson, N. Testing the stability of crime patterns: Implications for theory and policy. *J. Res. Crime Delinq.* **2010**. [CrossRef]
56. Andresen, M.A. Testing for similarity in area-based spatial patterns: A nonparametric monte carlo approach. *Appl. Geogr.* **2009**, *29*, 333–345. [CrossRef]
57. Andresen, M.A. The ambient population and crime analysis. *Prof. Geogr.* **2011**, *63*, 193–212. [CrossRef]
58. Kounadi, O.; Ristea, A.; Leitner, M.; Langford, C. Population at risk: Using areal interpolation and twitter messages to create population models for burglaries and robberies. *Cartogr. Geogr. Inf. Sci.* **2017**, 1–15. [CrossRef]
59. Eurostat. Geostat 2011. Available online: http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat (accesed on 15 May 2016).
60. FlashscoresEngland. Available online: http://www.flashscores.co.uk/football/england (accesed on 1 June 2016).
61. Eck, J.; Chainey, S.; Cameron, J.; Wilson, R. *Mapping Crime: Understanding Hotspots*; U.S. Department of Justice Office of Justice Programs: Washington, DC, USA, 2005; pp. 1–72.
62. Cliff, N. Some cautions concerning the application of causal modeling methods. *Multivariate Behav. Res.* **1983**, *18*, 115–126. [CrossRef] [PubMed]
63. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
64. Resch, B.; Usländer, F.; Havas, C. Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartogr. Geogr. Inf. Sci.* **2017**, 1–15. [CrossRef]
65. Latent Dirichlet Allocation in R. Available online: http://epub.wu.ac.at/3558/1/main.pdf (accesed on 15 May 2016).
66. Hornik, K.; Grün, B. Topicmodels: An r package for fitting topic models. *J. Stat. Softw.* **2011**, *40*, 1–30. [CrossRef]
67. University, V. Violence Vocabulary Word List. Available online: https://myvocabulary.com/word-list/violence-vocabulary/ (accesed on 1 June 2016).
68. User Guide to Home Office Crime Statistics. Available online: https://www.gov.uk/government/publications/user-guide-to-ho-crime-statistics (accesed on 1 June 2017).
69. Anselin, L.; Syabri, I.; Kho, Y. Geoda: An introduction to spatial data analysis. *Geogr. Anal.* **2006**, *38*, 5–22. [CrossRef]
70. Getis, A.; Ord, J.K. The analysis of spatial association by use of distance statistics. *Geogr. Anal.* **1992**, *24*, 189–206. [CrossRef]
71. Ord, J.K.; Getis, A. Local spatial autocorrelation statistics: Distributional issues and an application. *Geogr. Anal.* **1995**, *27*, 286–306. [CrossRef]
72. Anselin, L. Local indicators of spatial association-lisa. *Geogr. Anal.* **1995**, *27*, 93–115. [CrossRef]
73. Chainey, S.; Ratcliffe, J. *Gis and Crime Mapping*; John Wiley & Sons: Hoboken, NJ, USA, 2013; ISBN 1118685199.
74. Wartenberg, D. Multivariate spatial correlation: A method for exploratory geographical analysis. *Geogr. Anal.* **1985**, *17*, 263–283. [CrossRef]
75. Anselin, L. Under the hood issues in the specification and interpretation of spatial regression models. *Agric. Econ.* **2002**, *27*, 247–267. [CrossRef]

76. Anselin, L. Global Spatial Autocorrelation. Bivariate, Differential and eb Rate Moran Scatter Plot. 2017. Available online: https://geodacenter.github.io/workbook/5b_global_adv/lab5b.html (accesed on 5 November 2017).
77. Mac Nally, R. Regression and model-building in conservation biology, biogeography and ecology: The distinction between–and reconciliation of-'predictive'and 'explanatory' models. *Biodivers. Conserv.* **2000**, *9*, 655–671. [CrossRef]
78. Gardner, W.; Mulvey, E.P.; Shaw, E.C. Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial models. *Psychol. Bull.* **1995**, *118*, 392. [CrossRef] [PubMed]
79. Osgood, D.W. Poisson-based regression analysis of aggregate crime rates. *J. Quant. Criminol.* **2000**, *16*, 21–43. [CrossRef]
80. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
81. Anselin, L.; Kelejian, H.H. Testing for spatial error autocorrelation in the presence of endogenous regressors. *Int. Reg. Sci. Rev.* **1997**, *20*, 153–182. [CrossRef]
82. Bernasco, W.; Block, R. Robberies in chicago: A block-level analysis of the influence of crime generators, crime attractors, and offender anchor points. *J. Res. Crime and Delinq.* **2011**, *48*, 33–57. [CrossRef]
83. Frosdick, S.; Newton, R. The nature and extent of football hooliganism in england and wales. *Soccer Soc.* **2006**, *7*, 403–422. [CrossRef]
84. Wortley, R. A classification of techniques for controlling situational precipitators of crime. *Secur. J.* **2001**, *14*, 63–82. [CrossRef]
85. Groff, E.R.; Weisburd, D.; Yang, S.-M. Is it important to examine crime trends at a local "micro" level?: A longitudinal analysis of street to street variability in crime trajectories. *J. Quant. Criminol.* **2010**, *26*, 7–32. [CrossRef]
86. Statista. Number of Monthly Active Twitter Users Worldwide from 1st Quarter 2010 to 3rd Quarter 2017 (In Millions). Available online: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ (accesed on 21 February 2017).