Big Social Data and GIS: Visualize Predictive Crime

Full Papers

Anthony J. Corso
California Baptist University
acorso@calbaptist.edu

Abdulkareem Alsudais
Claremont Graduate University
abdulkareem.alsudais@cgu.edu

Brian Hilton
Claremont Graduate University
brian.hilton@cgu.edu

Abstract

Social media is a desirable Big Data source used to examine the relationship between crime and social behavior. Observation of this connection is enriched within a geographic information system (GIS) rooted in environmental criminology theory, and produces several different results to substantiate such a claim. This paper presents the construction and implementation of a GIS artifact producing visualization and statistical outcomes to develop evidence that supports predictive crime analysis. An information system research prototype guides inquiry and uses crime as the dependent variable and a social media tweet corpus, operationalized via natural language processing, as the independent variable. This inescapable realization of social media as a predictive crime variable is prudent; researchers and practitioners will better appreciate its capability. Inclusive visual and statistical results are novel, represent state-of-the-art predictive analysis, increase the baseline R^2 value by 7.26%, and support future predictive crime-based research when front-run with real-time social media.

Keywords

Big Data, GIS, Predictive Crime Analysis, Risk Terrain Modeling, Social Media, Spatial Correlation

Introduction

In the new millennium the emergence of geographic information systems and social media have successfully reinforced many aspects of crime analysis. In 2015, the Chicago Police Department's Superintendent, Garry McCarthy, reported official crime and murder numbers were at all-time record lows (2015). However, crime reporting between 2006 and 2010 by the Bureau of Justice Statistics, revealed that 3.4 million violent victimizations per year went unreported (Langton et al. 2012). Accuracy and precision of crime reporting are important features for police, city authorities, and citizens. Ostensibly, Big Data supports an abundance of crime-incident statistics that allow policy makers, law enforcement agencies, and public safety bureaucracies to allocate resources efficiently and effectively; yet, many statistics are predominately retrospective. Conversely, predictive crime analysis is forward-looking, opportune, and when combined with accurate social behavior, unequivocally gives citizens the ability to identify and circumnavigate high-risk topographies. A Big Data social media geographic information system (GIS), bootstrapped with historical crime data, establishes the state-of-the-art predictive crime paradigm. As a result, retrospective and predictive behavioral social data are used to analyze, evaluate, and predict crime.

Whereas formal investigation and analysis of crime is beneficial, many incidents go unreported by citizenry, witnesses, and victims. This action is self-justified because of concern that reporting criminal incident may reveal substantive clues about their identity and subsequent exposure to nefarious activity. Furthermore, fear, absence of police cooperation, and general lack of importance are common reasons for unreported violent crime (Langton et al. 2012). On one hand, crime may be decreasing; on the other hand, it is frankly not properly being reported. Nonetheless, this does not automatically indicate criminal incident is—in fact—unanalyzable. It means crime not reported to authorities via formal process must be collected using alternative means. This effects long-term GIS crime analysis with respect to bootstrapping the artifact with crime data. Thus, alternative methods to collect, report, and verify criminal incident are needed. Research suggests predictive crime methods developed using linguistic analysis of social media, risk terrain modeling, and individual behavior extracted from social media (Agarwal et al. 2014) are well suited to support such collection, reporting, and analysis.

Modern natural language processing procedures allow access to latent features within a social media corpus. Featherstone (2013) describes a social media crime-based collection and prediction system allowing such analysis; unlike crime-based analysis and prediction via data mining algorithms, their solution adds functionally by identifying the context of the environment. Gerber (2014) and Wang et al. (2012) indicate social media predictive crime analysis is possible. Also, given that a large amount of crime is unreported, alternative crime reporting and verification methods are critical; research from Derczynski et al. (2013) on tweet sparsity, Kaufmann et al. (2010) on addressing tweet normalization, and Bontcheva et al. (2013) on tweet information extraction build the support framework for social media to be used in place of formal crime reporting procedures. Together, an unconventional method is formed that provides new ways to construct a GIS risk terrain model (RTM) integrated with geospatial social media in order to significantly enhance crime prediction or expose previously unreported incidents.

Risk terrain models infrequently consider input risk layers originating from a social media corpus. A microblog corpus is complex and its preprocessing needs are considerable if it is to be consumed as a geographic information system RTM input risk layer. In addition, the corpus must be operationalized to expose its latent behavioral structure and intricate social context. Subsequently, evaluating social media as an RTM input risk layer within a geographic information system requires spatial pattern analysis and overt mathematical techniques such as probability or regression calculations. Whereas social media is orthogonal to the current RTM research model it is poised to add significant value as an emerging and disruptive RTM input risk layer. For example, Drawve (2014) found RTM crime-based technology artifacts can be compared to traditional spatial and temporal analysis of crime, density estimation, and various other retrospective crime analysis solutions. Social media realized as an input risk layer to increase an RTM's statistically significant predictive capability, crime-reporting, and domain-specific GIS crime analysis functionality will increase its acceptance as a social risk variable.

The revolutionary geographic information system artifact constructed implements social media's latent linguistic features to explore criminal incidents and their visual representations. Explicitly, it implements a Twitter corpus by overcoming a tweet's data sparsity and discovers the latent linguistic-based grammatical structures embedded within. The solution is supported by ArcGIS data analysis, fundamental linguistic experiments, visualizations, and regression calculations. It overcomes traditional GIS retrospective outcomes because the social media corpus is used as a proxy that yields an information system artifact with predictive crime analysis capabilities. While the solution is exploratory it exemplifies the need for real-time data collection, social media linguistic processing, and Big Data assimilation with respect to intelligence-based predictive theory. Consequently, the project unequivocally acknowledges the possibility of a relationship between a grammatically processed social media corpus and domain specific datasets combined within the confines of a GIS RTM solution. Beyond the introduction, the remainder of the work consists of a literature review. A section describing hypothesis development followed by the research methodology including data and its features. An analysis section describing the study and its quantitative and qualitative features. Last, the experiment conducted, conclusions drawn, limitations reviewed, and suggestions for future extension are provided.

Literature Review

Social media and its integration with crime data is of interest with respect to predictive crime analysis as observed in a GIS risk terrain modeling environment. The "Big Data" definition of Manoochehri (2014) would include social media and therefore a tweet corpus collected via Twitter's API. Many have considered such a corpus, e.g., in a compared algorithm research model Barbosa et al. (2010) set precedent for tweet corpus normalization and semantic detection, which, are now foundational components for big data tweet corpus artifact construction. The information system artifact is best formalized by Offermann et al. (2010), and the concurrent processing of social media's noisy corpus and law enforcement data is furthering big data crime analysis artifact construction. Early work in risk terrain modeling was concerned with in-depth knowledge of environmental criminology and its correlation with social behavior. Risk terrain modeling input layers were alluded to as early as Block et al. (1995). Subsequently, they were operationalized by Caplan et al. (2010) and Kennedy et al. (2011); standard RTM risk layers currently exist. Social media event prediction, crime prediction, and intelligence-based analysis are methodological approaches to consider when RTM risk layer guidelines for social media are operationalization. Noisy corpora are currently not operationalized for RTM risk layer input. This suggests a strong need for comparison of disparate artifacts using standard GIS spatial analysis techniques and statistical processing functions or development of social media RTM risk layers.

Natural language processing methodologies mitigate tweet noise, thus, normalizing a sparse tweet corpus and preparing it for examination is conceivable. Two customary methods include spell-checking as implemented by Choudhury et al. (2007) and Mays et al. (1991), and more currently normalization via noisy channel methods proposed by Cook et al. (2009). The latter scrutinizes a sparse corpus using an unsupervised text normalization approach, and on a set of 303 text messages they achieved 59% overall accuracy. A normalization method using statistical machine translation, most noted by Aw et al. (2006) and perfected by Han et al. (2011), extends traditional machine translation approaches and attempts to obtain analysis at the context or grammatical structure level of a corpus. As a result, they were able to normalize an SMS corpus and return only an 11% error rate. Mitigating tweet noise via NLP processing starts the operationalization of a tweet corpus, but it does not convert it to an RTM risk layer.

Several researchers address periphery issues of a sparse big data NLP tweet corpus and crime prediction. Wang et al. (2012) used an information system development methodology with automatic semantic analysis and linear prediction to study various aspects of social media crime prediction. In particular, they stress that traditional systems do not implement social media in the setting of a contiguous event. Gerber (2014) and Featherstone (2013) study predictive crime via latent Dirichlet and frequency-based topic analysis, respectively. They incorporate the initial operationalization of social media as an RTM risk layer, and its evaluation of the optimal impact of an artifact. In a work on geospatial crime mapping Bendler et al. (2014) validate the use of a normalized tweet corpus and suggest social media can enhance accuracy of specific crime type prediction. In a GIS artifact using geographically weighted and zero-inflated Poisson regression models to examine social media and crime, results were significant and supported accuracy of hot spot construction (Bendler et al. 2014). Bendler et al. (2014) clarify and extend the many ways early work in crime research relates to a tweet corpus.

When risk layers are selected and consumed by the GIS model many factors significantly influence artifact predictive accuracy and artifact comparability. To address this, traditional risk terrain models use quantifiable techniques such as predictive accuracy and recapture rate indices (Drawve, 2014) for assessment. Drawve (2014) tested performance comparison models and processes using, STAC, Nnh, KDE, and RTM artifacts. The predictive accuracy index (PAI) was used as a measure of accuracy with assessment being component-based variables consistent with the environmental criminology variables of time, place, and risk. In addition, the recapture rate index (RRI) suggests measurement for short-term and long-term crime prediction. Since it is assumed that social media, in a similar way to traditional risk layers, can be operationalized and subsequently compared via the same performance priorities of traditional layers, preliminary investigation and analysis to determine its scope of use is relevant.

Hypothesis Development

Prior studies consistently suggest social media corpora as a GIS crime analysis risk layer are associated with three specific dimensions of preprocessing. Social media corpus normalization, crime identification, and context of surrounding social behavior. Each needs examination to uncover quantifiable results and qualitative visual relationships found in both retrospective and RTM social media crime analysis artifacts. Although very different from environmental criminology's traditional GIS risk layer construct of time, place, and crime, a tweet corpus combined with Supplemental Nutrition Assistance Program locations produces similar outcomes and sophisticated crime maps. Such revolutionary crime layers are key proxies for predictive crime. The concept is analogous to John Snow's 1855 description of cholera and proximity to the Broad Street water pump, where quantitative and qualitative theory, experiment, and analysis produced results to generate hypotheses and discussion, which, beforehand were limited (Tufte 1983). Therefore, simple conjecture supports a social media GIS crime analysis artifact design that reveals a relationship between social media and crime when controlling for the effects of social behavior.

In view of existing hot spot crime mapping theory, it is expected that no metadata in addition to crime frequency is needed to represent the potential for future crime (Drawve 2014). In other words, hot spot mapping is retrospective, only predicting the future given the past. In Kennedy et al. (2011) retrospective or density mapping, is considered reactive and its crime risk or potential value to identify a subsequent high or low crime risk topography does not make it predictive. Further, considering the context of social media as an online communication forum sketched by Bendler et al. (2014), Rakesh et al. (2013), and more importantly Li et al. (2012), different levels of linguistic processing yield richness inherent of social and emotional structures in connection with users. The identification of hot spots created from a risk terrain model implementing a social media risk layer via grammatical structure captures greater levels of social behavior and therefore increases predictive richness just as greater frequency-based incident increases traditional density map constructs. It is expected that a tweet corpus heavily processed for grammatical constructs will be more predictive than a corpus simply linguistically tagged with parts of speech, or a retrospective crime analysis tweet corpus where no linguistic processing is introduced. Therefore, implementing a risk-based social media artifact, scrutinized in quantitative and qualitative ways; a hypothesis is considered.

Hypothesis (H1). Risk terrain model predictive capability will be highest in the grammatically tagged tweet corpus condition, when compared to a standard tagged or untagged social media corpus operationalized as a GIS risk layer.

Within the defined theoretical framework, the research model in Figure 1 is presented. It tests the relationship between social media and crime when controlling for multiple levels of social behavior. A favorable outcome will produce a GIS RTM artifact with better capability to predict crime and help observe social media risk variables.

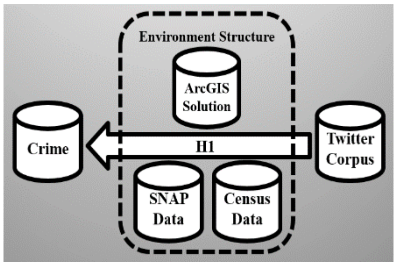


Figure 1. Research Model

Research Methodology

Four data sources, one primary and three secondary, are used to construct the artifact. The primary data consists of a tweet corpus that originated from approximately 512,000 tweets collected from the Twitter stream between August 2014 and December 2014. Each original tweet consists of three primary content lines: author, time, and content, with approximately 25 additional metadata fields. Collecting only geo-coded tweets is done by applying a latitude and longitude polygon bounding box in the tweet collection code. Thus, the primary tweet collection polygon is coded with a Southwest (bottom left) corner of 33.137051, -112.511466 and a Northeast (top right) corner of 33.767319, -111.531636. The source code is written in Java and uses standard JSON metadata collection. Due to a dearth of processing cycles and subsequent to initial collection, the tweet corpus was narrowed to approximately 25,000 tweets within the Chicago, IL area. Figure 2a displays the eight by six mile rectangle that represents the entire tweet population used to construct the tweet corpus.

Next, secondary crime data is collected from the City of Chicago and consists of approximately 270,000 crimes (2014a). These data originated between August 2014 and December 2014 and downloaded from the Chicago website December 31, 2014. Each record consists of 22 fields with primary type of crime, latitude, and longitude being examples, see Figure 2b.

Third, the Supplemental Nutrition Assistance Program (SNAP) is a nutrition assistance social program offered to low-income participants. Via Electronic Benefits Transfer (EBT) cards SNAP benefits provide low-income individuals or families access to eligible food items within a nationwide network of greater than 250,000 locations (2014b). There are eight fields for each record: ID, store name, latitude, longitude, address, city, state, zip code, and county. Figure 2c displays the 521 locations identified within the study area, which, overlaps tweet and crime data. Last, census data were obtained via ArcGIS and its access to the ESRI 2010 census CD-ROM dataset (2012).

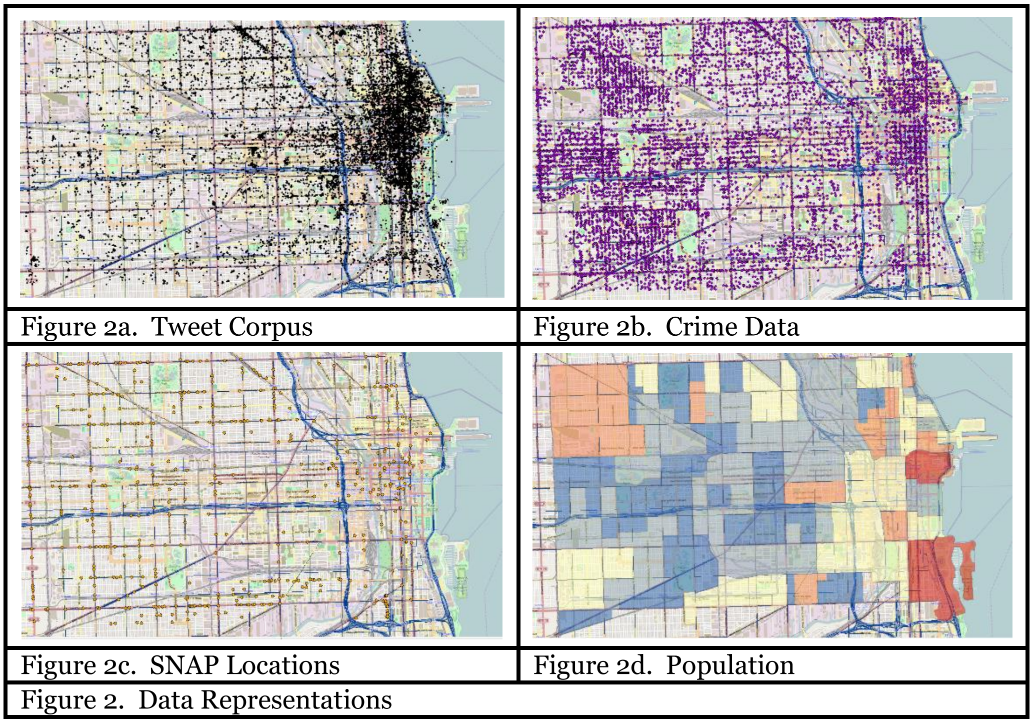


Figure 2a. Tweet Corpus

Figure 2b. Crime Data

Figure 2c. SNAP Locations

Figure 2d. Population

Figure 2. Data Representations

The features of each dataset represent social behavior of real-world locations in which they exist. For example, a grammatically tagged tweet represents an English-like structure, as applied by an NLP technique, of linguistic social behavior for a specific user at a specific location. Crime and SNAP features are the latitude and longitude of the actual crime incident and store location and represent the social environment in which they exist. The full assortment and counterpart of each, according to its variable assignment are as follows:

Dependent Variable. The variable to be measured with the measurement number recorded and used as an evaluation mechanism for the independent variable and is represented by R^2 value based on the OLS outcome. The R^2 value will be used for the dependent variable outcome because its calculation is deemed to be significantly influenced by the NLP process being used to operationalize the tweet RTM risk layer.

Independent Variables. The amount of NLP procedure used. Each sampling unit, i.e., a single tweet, may influence the outcome of the study. Two NLP techniques will be applied to the tweet corpus systematically. Each is deemed to influence the dependent variable.

Nuisance Variables. The geographical information system OLS model and mapping algorithm used for each different treatment to produce an R^2 value and visualization will be influenced by a number of nuisance variables. This variable includes different versions of the mapping solution, which, will be measured in terms of R^2 value of an Nnh and RTM artifacts.

- Nnh: Nearest Neighbor Clustering is based on threshold distance to which the crime incidents are compared to identify clusters, i.e., minimum number of points specified to make cluster decision.
- RTM: Risk Terrain Model. A common model used with point data to calculate risk assessment of a study area based upon physical but more importantly social factors. It results in higher risk values being associated with increased odds of a criminal event occurring in the future. Risk factors are commonly operationalized via proximity to or density of each factor across the overlaid grid of the study area. Each risk factor is expressed separately as a map layer because of differing spatial influences, and then combined to form an overall risk assessment. For example, high-density areas of alcohol outlets being within a block of a pawn shop could represent riskier areas for potential victims. These two factors have different operational spatial influences but can be spatially joined. It is the spatial overlap of risk factors that creates risk in the environment, i.e., where future crime would be expected to occur.
- Census and SNAP locations although act as an independent variable are a component of the solution being used to enforce a social environment as needed for a risk terrain model environment.

Sampling Units. This will be each tweet as collected from the data collection Twitter API data collection stream. This should be consistent with other like participants or sampling units used in informatics.

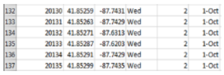
Random Assignment. Assignment of Sampling Units will be done via the default assignment of tweet ID as pulled from the data collection process. As each tweet is created a random ID is generated by Twitter. The tweet ID is sorted in ascending order; the top 550,000 were used as the corpus of this study.

The experiment was conducted in a controlled setting where artifacts were constructed in a usability laboratory at Claremont Graduate University. The study was designed as a two level factorial experiment, with crime as one factor and artifact construction the other, see Table 1. The artifact construction factor was extended to three sublevels, each level represents a tweet corpus NLP processing treatment. The three treatments are noNLP, medNLP, and hiNLP, and the experiment was conducted 3 times, one for each.

Table 1. Experiment Design

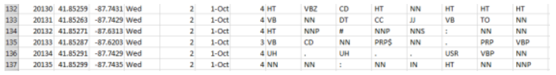
	Artifact Construction			
	noNLP	medNLP	hiNLP	Total Treatments
Crime Data	1	1	1	3

With the tweet corpus being the independent variable it was the basis for manipulation. It was selected after careful consideration because of the social behavior it represents. Geocoded tweets allow for point of reference and random solicitation of latent linguistic societal variables. As displayed in Table 1 under artifact construction and shown in Figure 3, various versions of the tweet corpus were constructed; a control corpus with no NLP processing (Figure 3a), i.e., a latitude and longitude point representation of a tweet's existence. Next, a nonspecific NLP tweet corpus was constructed using a General Architecture for Text Engineering (GATE) annotation pipe. This annotation pipe used A Nearly-New Information Extraction System (ANNIE), to process and apply part of speech tags to the corpus (Figure 3b). A third corpus using ANNIE; however, 29 Java Annotation Patterns Engine (JAPE) rules were written and used to convert the standard ANNIE part of speech tags into an English-like grammar tag (Figure 3c). Therefore, each version of artifact was composed of the secondary data and a specific tweet corpus, i.e., control tweet corpus (noNLP), nonspecific NLP (medNLP), and grammatically processed (hiNLP) corpus. In a 1 by 3 design with ordering of the conditions insignificant this inclusive experiment tested crime prediction via tweet corpus NLP manipulation.



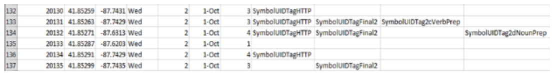
132	20130	41.85259	-87.7431	Wed	2	1-Oct
133	20131	41.85263	-87.7429	Wed	2	1-Oct
134	20132	41.85271	-87.6313	Wed	2	1-Oct
135	20133	41.85287	-87.6203	Wed	2	1-Oct
136	20134	41.85291	-87.7429	Wed	2	1-Oct
137	20135	41.85299	-87.7435	Wed	2	1-Oct

Figure 3a. noNLP



132	20130	41.85259	-87.7431	Wed	2	1-Oct	4	HT	VBZ	CD	HT	NN	HT	HT	HT
133	20131	41.85263	-87.7429	Wed	2	1-Oct	4	VB	NN	DT	CC	JJ	VB	TO	NN
134	20132	41.85271	-87.6313	Wed	2	1-Oct	4	HT	NNP	#	NNP	NNS	:	NN	NN
135	20133	41.85287	-87.6203	Wed	2	1-Oct	3	VB	CD	NN	PRP$	NN	.	PRP	VBP
136	20134	41.85291	-87.7429	Wed	2	1-Oct	4	UH	.	UH	.	.	USR	VBP	NN
137	20135	41.85299	-87.7435	Wed	2	1-Oct	4	NN	NN	:	NN	IN	HT	NN	NNP

Figure 3b. midNLP



132	20130	41.85259	-87.7431	Wed	2	1-Oct	3	SymbolUIDTagHTTP			
133	20131	41.85263	-87.7429	Wed	2	1-Oct	3	SymbolUIDTagHTTP	SymbolUIDTagFinal2	SymbolUIDTag2cVerbPrep	
134	20132	41.85271	-87.6313	Wed	2	1-Oct	4	SymbolUIDTagHTTP	SymbolUIDTagFinal2		SymbolUIDTag2dNounPrep
135	20133	41.85287	-87.6203	Wed	2	1-Oct	1				
136	20134	41.85291	-87.7429	Wed	2	1-Oct	4	SymbolUIDTagHTTP			
137	20135	41.85299	-87.7435	Wed	2	1-Oct	3		SymbolUIDTagFinal2		

Figure 3c. hiNLP

Figure 3. Corpus NLP Processing

Analysis of Methodology, Limitations, and Summary of Results

To analyze the qualitative and quantitative results a number of spatial statistic tools were employed. First, descriptive statistics in the form of spatial autocorrelation of feature location and value were considered. Next, a visual representation of the surface data was prepared. Last, a continuous surface to visualize results is helpful, although it does not reveal statistical analysis of the experiment. Therefore, ordinary least squares (OLS) regression analyses were executed.

Figure 4 provides spatial statistics and pattern evaluation for each data layer represented in the GIS artifact using the control tweet corpus. Figure 4a explains tweet corpus point locations, with a z-score of 2.16, the corpus is represented via spatial clusters with a less than 5% likelihood the clustered pattern is random chance. A similar outcome is true for SNAP data (Figure 4c) with a z-score of 3.86, and population data (Figure 4d) with a z-score of 3.10; both have less than 1% likelihood that their respective clustered patter is a result of random chance. Although the pattern for each prior dataset is spatially clustered, the crime data, Figure 4b, yields a z-score of .375, thus representing a pattern that does not appear to be significantly different than random.

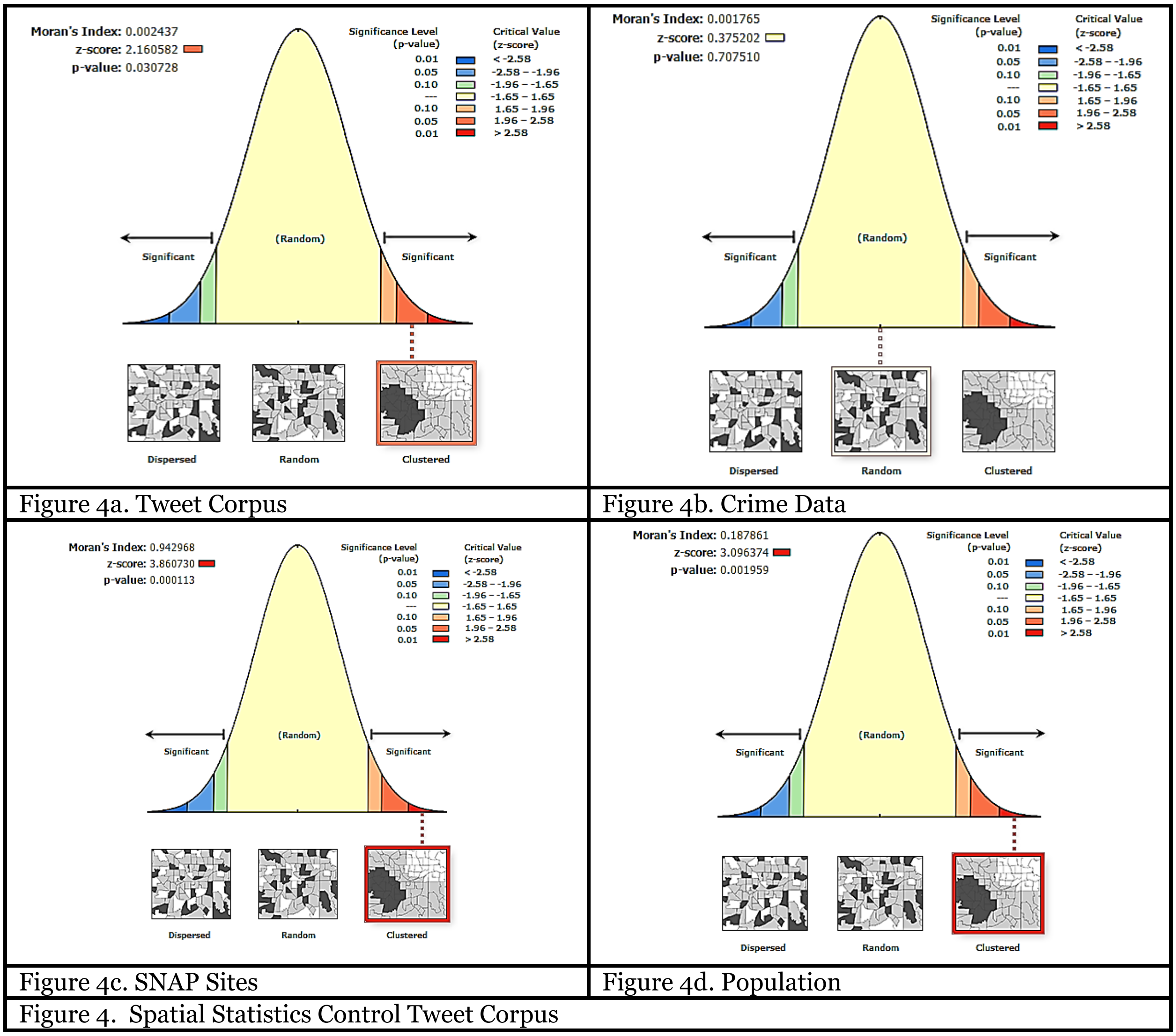


Figure 4a. Tweet Corpus

Figure 4b. Crime Data

Figure 4c. SNAP Sites

Figure 4d. Population

Figure 4. Spatial Statistics Control Tweet Corpus

To demonstrate the implications of NLP processing of the corpus, aggregation of crime, SNAP, and population data was necessary. A number of spatial joins were constructed and executed. First, crime data was spatially joined with population. Next, SNAP locations were joined with the crime and population data. Subsequently, tweet point locations, representing the noNLP tweet corpus, were joined with the crime, population, and SNAP data, Figure 5 illustrates the combined layers and represents the baseline environment for the study. The spatial join process was repeated for the midNLP and the hiNLP corpora.

With all the layers combined further qualitative and quantitative analysis in an ordinary least squares approach was employed. The ArcGIS OLS tool simultaneously tests feature structures and measures the models relational strength; it provides a more complete analysis of interrelationships with respect to outcomes than the individual layer spatial autocorrelation results represented in Figure 4. For each artifact construction NLP treatment, crime was selected as the dependent variable with population, SNAP, and tweets selected for explanatory variables. Figure 6 displays qualitative results of the OLS analysis for each tweet corpus NLP condition. That is, noNLP processing in Figure 6a, nonspecific midNLP in Figure 6b, and grammatically processed or hiNLP in Figure 6c. Considering the OLS resultant for each NLP treatment, the change in polygon shading visually illustrates NLP's influence on the dependent variable.

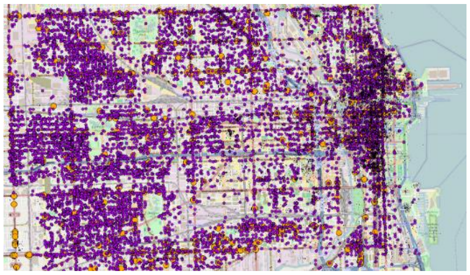


Figure 5. Joined GIS Layers noNLP

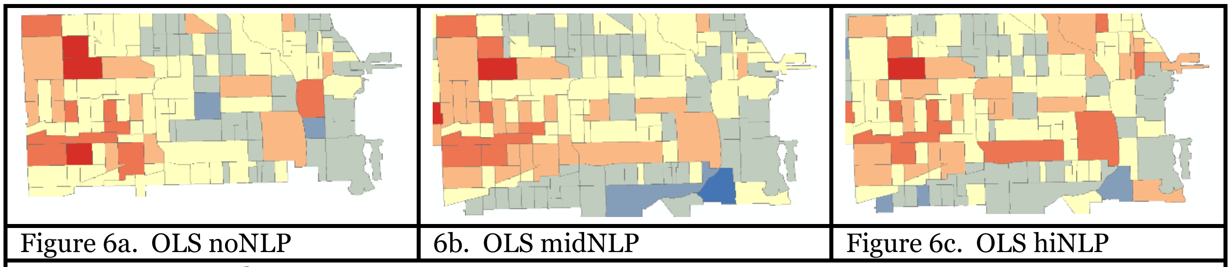


Figure 6a. OLS noNLP

6b. OLS midNLP

Figure 6c. OLS hiNLP

Figure 6. OLS Qualitative

Moreover, Figure 7 displays spatial autocorrelation resultants for each condition of the above OLS qualitative analysis. This spatial autocorrelation analysis was conducted to examine how dispersed, random, or clustered explanatory variables for each OLS model impacted the dependent variable. Thus, for z-scores of .48 (Figure 7a), .89 (Figure 7b), and 1.58 (Figure 7c) the noNLP, midNLP, and hiNLP tweet corpus layers, respectively, display that a random spatial patter exists for each.

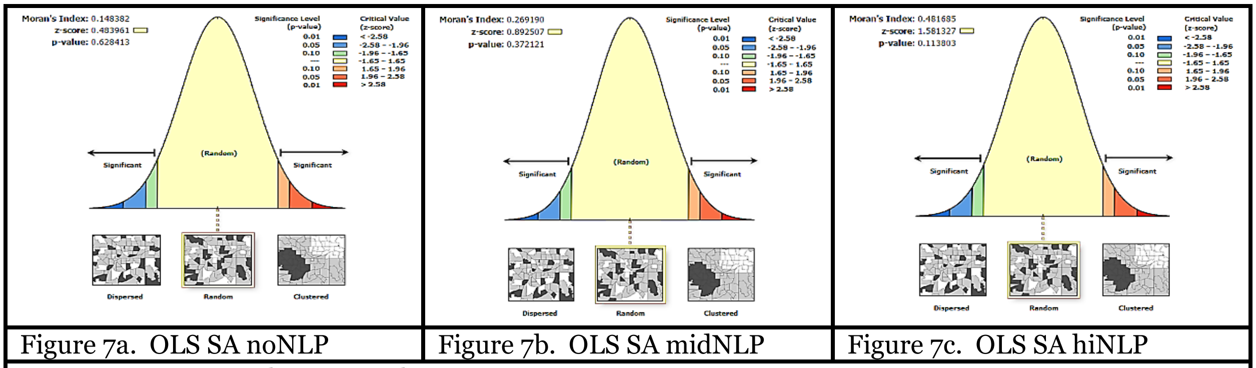


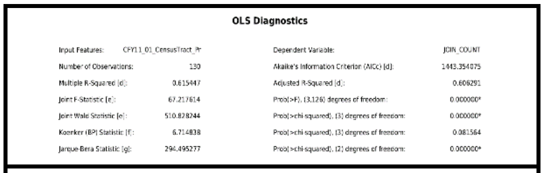
Figure 7a. OLS SA noNLP

Figure 7b. OLS SA midNLP

Figure 7c. OLS SA hiNLP

Figure 7. OLS Spatial Autocorrelation (SA)

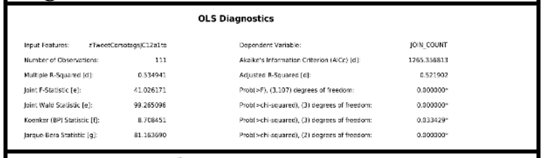
To accept the hypothesis, consideration of how much variation in the dependent variable (crime) has been explained by the model (in particular each NLP treatment layer). Figures 8, 9, and 10 present OLS diagnostic results. Figure 8 is associated with the crime, SNAP, population, and noNLP layers. It yields an R² value of .606, is the baseline, and control group for the project. Ensuing NLP tweet corpus layers with the ability to incite positive predictors will need to better this value to explain variations in the dependent variable. Next, Figure 9 represents the crime, SNAP, population, and midNLP layers. Yielding an R² of .522 its capability is worse than baseline. The dependent variable is less explained given a tweet corpus with a nonspecific NLP treatment. Last, crime, SNAP, population, and hiNLP layers are displayed in Figure 10. With an R² of .679 this solution best explains the dependent variable. The independent variable, hiNLP, did influence the dependent variable outcome. Therefore, the hypothesis—Risk terrain model predictive capability will be highest in the grammatically tagged tweet corpus condition, when compared to a standard tagged or untagged social media corpus operationalized as a GIS risk layer—is accepted.



OLS Diagnostics

Input Features:	CFY11_01_CensusTract_Pr	Dependent Variable:	JOIN_COUNT
Number of Observations:	130	Akaike's Information Criterion (AICc) [d]:	1443.354075
Multiple R-Squared [d]:	0.615447	Adjusted R-Squared [d]:	0.606291
Joint F-Statistic [e]:	67.217614	Prob(>F), (3,126) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	510.828244	Prob(>chi-squared), (3) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	6.714838	Prob(>chi-squared), (3) degrees of freedom:	0.081564
Jarque-Bera Statistic [g]:	294.495277	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

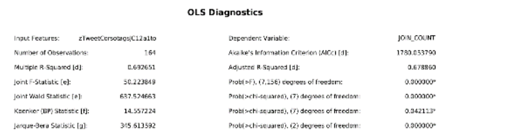
Figure 8. noNLP



OLS Diagnostics

Input Features:	zTweetCorsotags|C12a1to	Dependent Variable:	JOIN_COUNT
Number of Observations:	111	Akaike's Information Criterion (AICc) [d]:	1265.356813
Multiple R-Squared [d]:	0.534941	Adjusted R-Squared [d]:	0.521902
Joint F-Statistic [e]:	41.026171	Prob(>F), (3,107) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	99.265096	Prob(>chi-squared), (3) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	8.708451	Prob(>chi-squared), (3) degrees of freedom:	0.033429*
Jarque-Bera Statistic [g]:	81.163690	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

Figure 9. midNLP



OLS Diagnostics

Input Features:	zTweetCorsotags|C12a1to	Dependent Variable:	JOIN_COUNT
Number of Observations:	164	Akaike's Information Criterion (AICc) [d]:	1780.053790
Multiple R-Squared [d]:	0.692651	Adjusted R-Squared [d]:	0.678860
Joint F-Statistic [e]:	50.223849	Prob(>F), (7,156) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	637.524663	Prob(>chi-squared), (7) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	14.557224	Prob(>chi-squared), (7) degrees of freedom:	0.042113*
Jarque-Bera Statistic [g]:	345.613592	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

Figure 10. hiNLP

Limitations of the work fit into a number of specific areas. First, to the extent of methodological areas such as sample size, data, and prior investigation. Second, limitations with respect to researcher's longitudinal effects. Last, statistical validity in the area of meta-analysis of existing literature. Although three distinct areas are addressed, the limitations span the entirety of the problem under investigation.

In the area of corpus collection, random selection considering more than half a billion tweets sent via Twitter every single day. The issue becomes whether or not the corpus collected is a representative sample given this enormity. This view can be seen from two perspectives; one, time and location of tweet selection; two, API used to collect the tweets from the Twitter stream. The time-based error component does not influence overall results with the randomization process used, e.g., users post at different times.

Another limitation is the tokenization and part-of-speech tagging methods applied to the corpus. The NLP tokenization process used was break on space. More sophisticated techniques should be implemented in order to enhance results and overcome the known noise within the tweet tokenization process. Each word in a tweet was assigned a part-of-speech via Wordnet lookup; however, upon tweet chunking the rules applied only incorporated rudimentary chunk structure, e.g., (<NN><VBD><NN>)(<PRO><JJ><NN>). A more advanced process would increase identifying a causal relationship within a tweet; the inability to exploit such a novel relationship between tweet content and real-world context is present.

To best describe the meta-analysis limitation in detailed but concise terms the multi-data domain aspect of this work must be recognized. Each data domain has its own evaluation methodologies and measures. Formal NLP evaluation of the social media corpus does not provide satisfactory GIS artifact evaluation analysis. Likewise, overall GIS regression accuracy does not provide adequate NLP analysis. Hence, the meta-analysis of formal NLP, regression, and the predictive accuracy index (PAI) and recapture rate index (RRI) are used as a measure of accuracy with assessment being component-based variables with the environmental criminology variables of time, place, and risk. Impact to overall findings appears nominal and perceives to have little impact on conclusions of the study, but needs consideration.

Conclusion

The predictive crime capability of a social media GIS RTM risk layer was modeled, and through OLS regression and visualization was evaluated. Figures 8, 9, and 10 represent OLS outcomes with GIS processing of secondary data and implementation of a tweet corpus in three states. The noNLP corpus has tweet location only. NLP processing was applied to the midNLP and hiNLP corpora. Figure 7 highlights the quantitative R^2 values for the qualitative OLS regression analysis generated from the model shown in Figure 6. In other words, each proposed NLP solution was tested for the relationship between crime and a latent social behavior identified as the grammatical structure of the tweet. Tweets with greater English-like grammar were expected to provide greater predictive capability, social RTM structure, and be more trustworthy within the GIS model. With R^2 values of .606, .522, and .679 a relationship does in fact exist.

Future work is intended to extend the comparison of predictive spatial techniques, incorporate a social media GIS RTM risk layer, and use the predictive accuracy index (PAI) and recapture rate index (RRI) for evaluation of an artifact. With RTM increasing in popularity among researchers and practitioners, comparison is required to established spatial prediction techniques commonly used in hot spot mapping. The focus of prior work was to operationalize social media layers and use regression to evaluate outcomes; this study considers such an approach, with subsequent work to focus on the PAI and RRI indices to compare RTM to traditional hot spot techniques. Historically, crime not included in the pool of risk factors for the RTM model resulted in a model being constructed strictly from crime generators and crime attractors (CGAs). Given the notion that to predict where crime will occur—past crime has had to occur—allows RTM to different itself from such traditional hot spot techniques by excluding crime as a risk factor. Therefore, this effort is novel and its contribution is directed at constructing “foundational models” that integrate social media and past crime as risk layers with the intent on building an examination framework of how future research can compare “best models” without crime to “best models” with crime.

REFERENCES

2012. "Demographic, Consumer, and Business Data."

2014a. "City of Chicago."

2014b. "United States Department of Agriculture."

2015. "Chicago Crime Reached "Historic Low" in 2014: Police."

Agarwal, R., and Dhar, V. 2014. "Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research," *Information Systems Research* (25:3), pp 443-448.

Aw, A., Zhang, M., Xiao, J., and Su, J. Year. "A Phrase-based Statistical Model for SMS Text Normalization," COLING• ACL 20062006, p. 33.

Barbosa, L., and Feng, J. 2010. "Robust Sentiment Detection on Twitter from Biased and Noisy Data," *Proceedings of the 23rd International Conference on Computational Linguistics* (Poster Volume), pp 36-44.

Bendler, J., Ratku, A., and Neumann, D. 2014. "Crime Mapping through Geo-Spatial Social Media Activity,").

Block, R. L., and Block, C. R. 1995. "Space, place, and crime: Hot spot areas and hot places of liquor-related crime," *Crime and place: crime prevention studies, vol 4. Criminal Justice Press, Washington, DC*), pp 145-183.

Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D., and Aswani, N. Year. "TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text," RANLP2013, pp. 83-90.

Caplan, J. M., and Kennedy, L. W. 2010. "Risk Terrain Modeling Manual.," Newark, NJ: Rutgers Center on Public Security.

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., and Basu, A. 2007. "Investigation and modeling of the structure of texting language," *International Journal of Document Analysis and Recognition (IJDAR)* (10:3-4), pp 157-174.

Cook, P., and Stevenson, S. 2009. "An unsupervised model for text message normalization," in *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, Association for Computational Linguistics: Boulder, Colorado, pp. 71-78.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. Year. "Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data," Hissar, Bulgaria, 2013, pp. 198-206.

Drawve, G. 2014. "A Metric Comparison of Predictive Hot Spot Techniques and RTM," *Justice Quarterly*), pp 1-29.

Featherstone, C. 2013. "The relevance of social media as it applies in South Africa to crime prediction," *IST-Africa Conference and Exhibition (IST-Africa)*) 29-31 May 2013, pp 1-7.

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

Han, B., and Baldwin, T. Year. "Lexical normalisation of short text messages: Makn sens a# twitter," Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies2011, pp. 368-378.

Kaufmann, M., and Kalita, J. 2010. "Syntactic normalization of twitter messages," in *International conference on natural language processing.*: Kharagpur, India.

Kennedy, L. W., Caplan, J. M., and Piza, E. 2011. "Risk Clusters, Hotspots, and Spatial Intelligence: Risk Terrain Modeling as an Algorithm for Police Resource Allocation Strategies," *Journal of Quantitative Criminology* (27:3), pp 339-362.

Langton, L., Berzofsky, M., Krebs, C., and Smiley-McDonald, H. 2012. "Victimizations Not Reported to the Police, 2006-2010."

Li, H., Chen, Y., Ji, H., Muresan, S., and Zheng, D. 2012. "Combining social cognitive theories with linguistic features for multi-genre sentiment analysis," in *26th Pacific Asia Conference on Language, Information and Computation*: Bali, Indonesia, pp. 127-136.

Manoochehri, M. 2014. *Data Just Right: Introduction to Large-Scale Data and Analytics*, (Addison-Wesley: Upper Saddle River, NJ.

Mays, E., Damerau, F. J., and Mercer, R. L. 1991. "Context based spelling correction," *Information Processing & Management* (27:5) 1991/01/01, pp 517-522.

Offermann, P., Blom, S., Schönherr, M., and Bub, U. 2010. "Artifact Types in Information Systems Design Science – A Literature Review," in *Global Perspectives on Design Science Research,* R. Winter, J. L. Zhao and S. Aier (eds.), Springer Berlin Heidelberg, pp. 77-92.

Rakesh, V., Reddy, C. K., Singh, D., and Ramachandran, M. Year. "Location-specific tweet detection and topic summarization in twitter," Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on, IEEE2013, pp. 1441-1444.

Tufte, E. R. 1983. *The visual display of quantitative information*, (Graphics Press: Cheshire, Conn. (Box 430, Cheshire 06410).

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