Toward Predictive Crime Analysis via Social Media, Big Data, and GIS

Anthony J. Corso, Claremont Graduate University
Gondy Leroy, University of Arizona, Tucson
Abdulkareem Alsusdais, Claremont Graduate University

Abstract
To support a dissertation proposal a link between social media, incident-based crime data, and data of public domain needed to be verified. A predictive crime-based artifact utilizing data mining and natural language processing techniques commingled with graphical information system architecture is complex. With respect to social media, an attempt was made to observe such an artifact's data flexibility, process control, and predictive capabilities. Data and their capabilities were observed when preprocessing social media's noisy data, government-based structured data, and obscurely collected field data for use in a predictive GIS artifact. To support project goals the approach for artifact design, data collection, and discussion of results was couched as an exploratory study. Results indicate a link between social media data and domain specific datasets exist. Questions for further observation and research deal with processing the subtle differences between structured and noisy data, weighted social media input layers, and time-series analysis.
Keywords: Social Media; Crime; Predictive Analysis; GIS; Spatial Autocorrelation
Citation: Corso, A.J., Leroy, G., Alsusdais, A. (2015). Toward Predictive Crime Analysis via Social Media, Big Data, and GIS Spatial Correlation. In iConference 2015 Proceedings.

Research Data: In case you want to publish research data please contact the editor.
Contact: corsoa@cgu.edu

1 Introduction

In the 21st century, social media data and domain specific datasets, such as those on crime or healthcare, provide much information useful to predict human behaviors. These data approximately double in size every 20 months (Witten, Frank, & Hall, 2011). Our focus here is on crime and law enforcement data. Within this purview, demographic and crime data analysis have been incident-pattern-based and statistical. For example, a key element of crime analysis is composed of either an immediate, sort-term, or long-term problem where a cluster of crimes (e.g., robbery or assault) is based on victim pattern or crime events that share one or more distinct systematic opportunities (e.g., frequency or location) (Santos, 2012). By combining these data, social media, public crime, and census data via graphical information system analysis, we gain insight into social misconduct from a spatial-temporal context. With these modern data, analysis can be expanded beyond traditional risk terrain modeling (RTM) and focus on social media based RTM risk layers, where currently there is a dearth of research (Kennedy, Caplan, & Piza, 2011).

Today's social media enables real-time communication and reporting of events. Some of this information is relevant to law enforcement. For example, a bystander may have first-hand knowledge of a crime or a potential crime and publish critical information using social media. If the social media stream is timely and correctly processed, corrective action can be taken. Thus, it is plausible that with the construction of a proper data pipeline one can extract latent indicators and use them for predicting societal mishap (Manoochehri, 2014). Also, analyzing geospatial data and various other demographic data of public record, the latent indicators can be enhanced and subsequently used in reporting real-time crime (Featherstone, 2013). Such analysis is useful in many applications:

- **Event Recognition:** Spatial correlation as a record of observation by an eyewitness;
- **GeoEvent Processing:** Confirmation of trend mishap as per real-time event processing;
- **Crime Prediction:** Real-time trend prediction of criminal mishap; and
- **Enhanced Recommendation:** Casualty-based trend analysis via geo-mishap linguistic decay.

This work develops a comprehensive predictive crime-based artifact using social media. We focus on a Twitter corpus and intend to overcome a tweet's data sparsity to discover latent crime information embedded in the microblog. This will result in an information system artifact yielding predictive social media crime analysis via GIS trend-based dataset amalgamation. Therefore, given the exploratory nature of this work, its need for real-time data collection, data assimilation, and social media linguistic processing, the explicit focus here is towards identifying an affirmative link between a social media corpus and various data of public domain. The remainder of this work consists of section 2, a

review of literature. Section 3, problem definition, data selection, preprocessing, and software; with section 4 describing our procedure and results. Last, section 5 concludes and provides suggestions for future extension.

2 Literature Review

Risk Terrain Modeling and Big Data are of particular interest with respect to GIS artifacts. The first requires in-depth knowledge of environmental criminology and its correlation with social media and risk terrain model input layers; perhaps, to allow for real-world real-time crime-based social media event prediction (Brantingham & Brantingham, 1981) (J.M. Caplan & Kennedy, 2010) (Kennedy et al., 2011). The second addresses combining well-structured law enforcement data with a noisy Big Data corpus (Barbosa & Feng, 2010). The last considers methodological approaches for statistical processing of an RTM data pipe being processed by a GIS using Spatial Autocorrelation (Global Moran's I)[1] techniques.

In the design of a risk terrain model the operator, selects appropriate input layers. However, research suggests each layer of the RTM is not of equal value ("weight") for a particular solution. A weighting technique applied to RTM input layers would improve dataset processing and identify the best way to use a social media input layer. Also, if the operator is not familiar with an input layer, e.g., a tweet-based social media layer, they tend not to use it. Evaluation methods for weighting input layers is a prerequisite in producing better RTM outcomes and resolve the former issues, indeed, Kennedy et al. (2011) suggest enhancing risk terrain models by developing a reliable method for weighting the factors of each layer relative to one another. The work of Kennedy et al. (2011) inspired comments from Caplan et al. (2011) and help direct the community to extend RTMs in two ways: use contextual layer information like bars, strip clubs, or pawn shops, to estimate a layer's risk, second, test the predictive power of the artifact against retrospective hot spot maps. None of this work addresses the subtle aspect of weighting social media as an RTM input layer.

Despite the small amount of data given in a social media microtext if a tweet's sentiment-based classification label is used as an RTM input layer the artifact will gain predictive capabilities. Barbosa and Feng (2010) address automatic sentiment analysis of a tweet corpus and Go, Bhayani, and Huang (2009) classify sentiment of tweets based on polarity (positive and negative). The latter test four classifiers, and show Maximum Entropy and Naïve Bayes models using word-unigrams and word-bigrams outperform keyword lookup and Support Vector Machines (SVM). However, their classification work was conducted in a single-step classification process. Barbosa and Feng (2010) extend the polarity concept and divide classification into a two-step process. First, they classify tweets into a subjective or objective category. Then, with an SVM they classify polarity of the subjective tweets only, mapping part-of-speech and tweet metadata to achieve an 18.7% error rate. Agarwal, Xie, Vovsha, Rambow, and Passonneau (2011) likewise apply an SVM with polarity (positive, negative) but also extend polarity to a three class setup (positive, negative, neutral). Both projects return 75.39% and 60.83% accuracy for polarity (positive, negative) and polarity (positive, negative, and neutral). Once more, this community does not observe social media as a weighted RTM input layer.

Last, other researchers are developing technically feasible approaches to support weighted social media RTM input layers. To predict location Piza (2012) built a regression model to link location and crime, demonstrating the importance and necessity of automatic and statistical spatial-temporal crime analysis. J.M. Caplan and Kennedy (2010) are well known for this view and are currently the leading experts in layer-based crime analysis research. The work of Alsudais, Leroy, and Corso (2014) apply a random forest classifier to identify the type of location a tweet originated from; these results can be used to assign risk to a social media RTM input layer. As such, the views of Ratcliffe (2004), Kennedy et al. (2011), Piza (2012), and Alsudais et al. (2014) are in terms of crime risk analysis being applied to a GIS RTM layer. Still, many other risk terrain models exist and are considered in terms of place-based interventions and the theory of sparsity, nonetheless a social media corpus being unequivocally used as a weighted input layer remains unobserved.

3 Problem Definition

Risk terrain model input layers originating from a social media corpus are infrequently considered, must be heavily processed, must be properly weighted, and need to be statistically significant for a given solution. First, using natural language processing techniques and analyzing and interpreting results in real-time exhibits great difficulty and expense in terms of human and machine resources. Second,

[1] http://resources.arcgis.com/en/help/main/10.1/index.html#/How_Spatial_Autocorrelation_Global_Moran_s_I_works/005p0000000t000000/

although single domain Big Data analysis is successful, multi-domain dataset merging does not present the same ease of use. In addition, implementing a multi-domain dataset with layer weighting features would better amalgamate the disparate RTM input layers. Third, an artifact's framework needs to support data analysis including overt mathematical techniques, e.g., probability or regression calculations of constituent data streams. Such an artifact must consider data and data relationships processed via GIS and its corresponding spatial autocorrelation tools. This problem has existed since social media was recognized as an RTM input layer. A GIS artifact achieving this level of processing would considerably increase an RTM's predictive accuracy and increase the use of social media as a statically significant predictive input layer.

3.1 Data Selection and Software

Data Sets:

Although ad hoc single domain Big Data inquiry is successful we collected four data sets which will be combined and observed in a multi-domain GIS solution, see figure 1. The data are described as follows: first, from April 14, 2013—May 13, 2014 approximately 2,250,000 geo-coded tweets were collected from the Phoenix, AZ area. Collecting only geo-coded tweets was done by applying a latitude/longitude polygon bounding box within the tweet collection code. The polygon had a Southwest (bottom left) corner of 33.137051, -112.511466 and a Northeast (top right) corner of 33.767319, -111.531636. The tweets were collected using the Tweepy Python library.

Second, to reduce crime by identifying crime patterns, in 2012 SpotCrime released 15 million crime records ("SpotCrime Makes Database of 15 Million Crime Records Available to the Public," 2012). The records were made available to the public free of charge by accepting the terms of their database download disclaimer. The aggregation of SpotCrime's dataset is extracted from a number of different sources, for example, police departments, news reports, and user-generated content submissions ("SpotCrime," 2015).

Third, 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 and families access to eligible food items within a nationwide network of greater than 250,000 locations. Last, census data were obtained via ArcGIS and its access to the ESRI 2010 census CD-ROM dataset ("Demographic, Consumer, and Business Data," 2015).

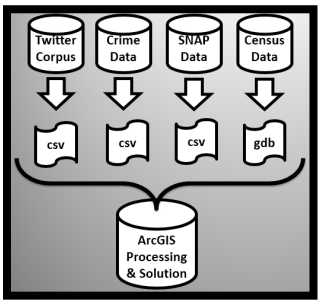

Figure 1. Data Sets Collected

Preprocessing:

GIS artifact design for this type of data analysis must provide visualization and statistical analysis tools to enable large-scale dataset evaluation in a meaningful and myopic way. Five surface maps of public domain were obtained and their corresponding shapefiles were: AZ State, AZ county lines, AZ census blocks, AZ street maps, and AZ transportation lines. The files were downloaded from the Data.Gov website, state of AZ website, and MapCruzin.com. ArcGIS 10.1, Java, Python, and the Microsoft Office suite of applications were used as development tools for the project. Python's NLTK processing components were used to create a tweet corpus consisting of the tweet's geo-code coordinates, its content, and its polarity. Given the laptop computer used for this project, random selection was used and the 2,250,000 tweet corpus was reduce to a 90,000 tweet corpus. The reduced corpus was used as a layer in ArcGIS and is shown in figure 2. The downloaded SpotCrime dataset was filtered to represent similar polygon coordinates to that of the tweet corpus and the crime dataset was reduce to 20,000 records. Crime attributes were, latitude, longitude, and crime type. It was loaded as a layer, see figure 3. The SNAP dataset was filtered to fit the AZ polygon and consisted of 1,940 locations. The SNAP location file was saved and loaded into ArcGIS and is displayed in figure 4. Last, income per census block was loaded from the ESRI CD-ROM dataset. Figure 5 and 6 map the tweet corpus, crime data, and SNAP locations against income levels where light green are low income areas and dark green

are high income areas. Also, the large scale map is at a scale of 1:800,000 and the small scale map is at a scale of 1:4,000,000. The processing for this work was completed on a Dell laptop in the research lab at Claremont Graduate University. The computer used was a Dell Latitude E6520, 2.60 GHz CPU with 8.00 GB RAM.

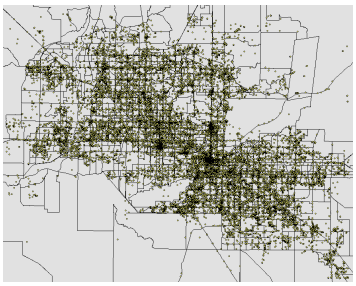

Figure 2. Tweet Corpus

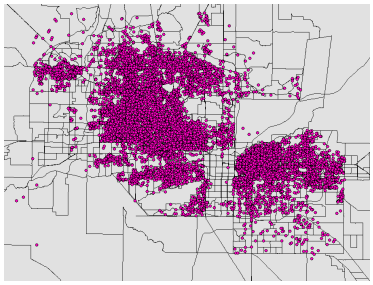

Figure 3. Crime Data

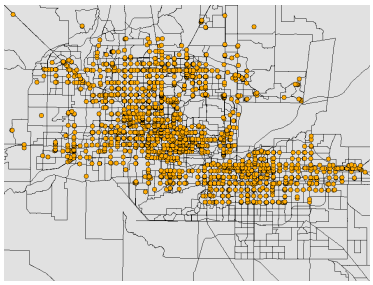

Figure 4. SNAP Locations

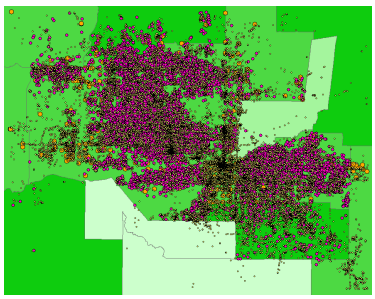

Figure 5. Data Large Scale

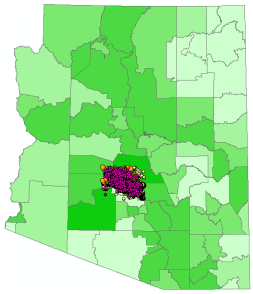

Figure 6. Data Small Scale

4 Exploratory Analysis

Although independently the tweet corpus, crime data, SNAP locations, and census data produce presumably simplistic maps, when spatial autocorrelation is conducted and a hot spot analysis is applied a complex and robust solution is created. The link between these data sets will be observed via exploratory analysis and results will help generate hypotheses for further research. An application for this type of data analysis must provide visualization techniques and overt statistical analysis tools, e.g., regression analysis, for proper evaluation. Therefore, to enable large-scale dataset evaluation to be visualized in an evocative way we present the following:

Procedure:

1. Create ArcGIS ArcMap solution file with basemap
2. Load each data layer checking that it is projected data
3. Aggregate incident data
   - Cluster crime, SNAP, and census layers
   - Identify scale of analysis
     - Run Incremental Spatial Autocorrelation
4. Run Hot Spot Analysis

After creating a new ArcMap solution file each dataset was loaded and its layer configured. Next, each layer was checked for accuracy and projected using the NAD_1983_UTM_Zone_12N projected coordinate system. Third, the artifact considered ArcGIS's Collect Event, Spatial Autocorrelation, and Spatial Joins tools. Crime data and SNAP locations were joined, using the Spatial Join tool, to create a new ArcMap layer with both sets of data represented. We then joined that layer to the census layer. The social media layer was clustered using the Collect Event tool and then joined to the crime, SNAP, census layer. Spatial Autocorrelation was used to find a distance measurement between locations for the social media and crime layer. Its use produced a distance peak value of 100 and 400 feet for the social media

and crime layer, respectively. A hot spot analysis was run with and without the social media in the joined data layer and a tolerance of 400 feet was set.

Results:

Figure 7 represents a hot spot analysis map for the final solution. The figure consists of a map showing significant hot spots with respect to social media and domain specific datasets. Also, this figure represents a multi-domain solution, i.e., social media, crime, SNAP, and census data are combined and processed by the GIS artifact. Figure 8 represents a hot spot analysis without the social media being considered within the multi-domain solution. Although the hot spots in figure 7 merely appear larger than in figure 8, the addition of social media data centers each hot spot differently such that the maps do not exactly overlap. Without further analysis we do not know if the solution predicts crime any better; however, the presence of social media does impact hot spot analysis outcome representation and suggests a correlation between social media and the other data does exist.

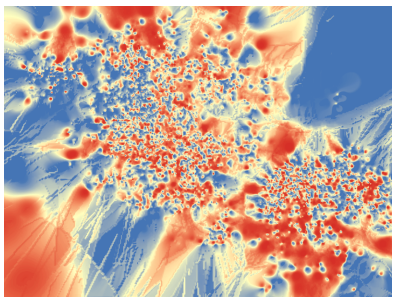

Figure 7. Hot Spot Social Media

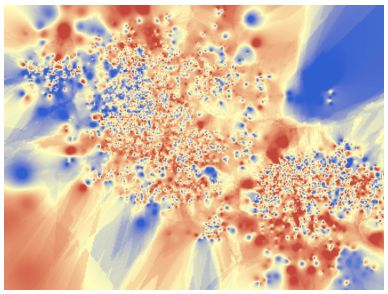

Figure 8. Hot Spot No Social Media

5 Conclusion

The results of the analysis represent overall better outcomes than a baseline solution of guessing whether a correlation exists. Construction of the ArcGIS solution calculates overall crime risk via hot spot analysis better than if the artifact was not created. Also, the new artifact yields significant insight with respect to a multi-domain solution using social media as an input layer. Social media, crime, and domain specific datasets helped provide visual representation of data and produced results suggesting subsequent inquiry should be considered. Therefore, combining multi-domain data in a lucid and orderly way produced a cohesively visual analysis.

Although integrating public domain and crime data are not novel, learning how to comingle them with ad hoc social media data streams in predictive ways via GIS is. As such, future work considering a weighting or value-based risk number for the social media layer is important. The ability to import a social media RTM risk layer with a risk value assigned to it will foster growth in the RTM community. Other possible extensions of work include application of Natural Language Processing (NLP) and data mining techniques applied to a Big Data social media corpus in order to project and evaluate predictive crime-based incident interactions in terms of linguistic features. NLP of a social media corpus is difficult, yet, recent progress in programming API like Python has made this a viable area for RTM social media research. Furthermore, time-series exploratory research is an approach to measure predictive results in a multi-domain artifact. This task is not trivial since all data sets must overlap in both time and place. Last, development of support tools and their evaluation will produce robust research alternatives for the community and opens the discussion of modeling a social media weighting solution.

6 References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. *Proceedings of the Workshop on Languages in Social Media*, 30-38.

Alsudais, A., Leroy, G., & Corso, A. (2014). We Know Where You Are Tweeting From: Assigning a Type of Place to Tweets Using Natural Language Processing and Random Forests. *Big Data (BigData Congress), IEEE International Congress on June 27, 2014-July 2, 2014*, 594-600. doi: 10.1109/BigData.Congress.2014.91

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

Brantingham, P. J., & Brantingham, P. L. (1981). *Environmental Criminology*. Beverly Hills, CA: Sage Publications.

Caplan, J. M., & Kennedy, L. W. (2010). *Risk Terrain Modeling Manual.* Newark, NJ: Rutgers Center on Public Security.

Caplan, J. M., Kennedy, L. W., & Miller, J. (2011). Risk Terrain Modeling: Brokering Criminological Theory and GIS Methods for Crime Forecasting. *Justice Quarterly, 28*(2), 360-381. doi: 10.1080/07418825.2010.486037

Demographic, Consumer, and Business Data. (2015). Retrieved January 2, 2015, from http://www.esri.com/data/esri_data/demographic-overview/census-overview/census2010

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

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1-12.

Kennedy, L. W., Caplan, J. M., & 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), 339-362. doi: 10.1007/s10940-010-9126-2

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

Piza, E. L. (2012). Using Poisson and Negative Binomial Regression Models to Measure the Influence of Risk on Crime Incident Counts. *Rutgers Center on Public Security*.

Ratcliffe, J. H. (2004). The hotspot matrix: A framework for the spatio-temporal targeting of crime reduction. *Police Practice and Research, 5*(1), 5-23.

Santos, R. B. (2012). *Crime analysis with crime mapping*. Los Angeles, CA: Sage.

SpotCrime. (2015). Retrieved January 2, 2015, from http://www.spotcrime.com/

SpotCrime Makes Database of 15 Million Crime Records Available to the Public. (2012, January 11). Retrieved January 2, 2015, from http://www.businesswire.com/news/home/20120111005878/en/SpotCrime-Database-15-Million-Crime-Records-Public#.VKcCHSvF-So

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (Third Edition ed.). San Francisco, CA: Elsevier.