Geosocial Media Data as Predictors in a GWR Application to Forecast Crime Hotspots

Alina Ristea
Department of Geoinformatics – Z_GIS, Doctoral College GIScience, University of Salzburg, Austria
mihaela-alina.ristea@sbg.ac.at
https://orcid.org/0000-0003-2682-1416

Ourania Kounadi
Department of Geo-information Processing, Faculty of Geo-Information Science and Earth Observation, University of Twente, Enschede, Netherlands
o.kounadi@utwente.nl
https://orcid.org/0000-0002-5998-7343

Michael Leitner
Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA, USA
mleitne@lsu.edu
https://orcid.org/0000-0002-1204-0822

Abstract

In this paper we forecast hotspots of street crime in Portland, Oregon. Our approach uses geosocial media posts, which define the predictors in geographically weighted regression (GWR) models. We use two predictors that are both derived from Twitter data. The first one is the population at risk of being victim of street crime. The second one is the crime related tweets. These two predictors were used in GWR to create models that depict future street crime hotspots. The predicted hotspots enclosed more than 23% of the future street crimes in 1% of the study area and also outperformed the prediction efficiency of a baseline approach. Future work will focus on optimizing the prediction parameters and testing the applicability of this approach to other mobile crime types.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases spatial crime prediction, street crime, population at risk, geographically weighted regression, geosocial media

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.56

Category Short Paper

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

1 Introduction

Crime occurrences are complex phenomena studied from an interdisciplinary path, including criminology, law, psychology, geography, or economy. An important factor in understanding crime patterns relates to their spatial and temporal attributes. Some of the methods that have been used to explore these attributes in crime analysis are hot spot detection [7], spatial regression [8], retrospective forecasting, machine learning, near-repeat concept [10], and

10th International Conference on Geographic Information Science (GIScience 2018).
Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 56; pp. 56:1–56:7
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

risk terrain analysis [14]. However, many prediction models and their strategies are defined and modeled for places with disparate regional conditions. Also, crime types have different spatiotemporal distributions because they are affected by different factors. For example, robberies increase during nights and weekends [14], while assaults are frequent around liquor outlet areas on weekends [4]. The current study aims to forecast crime in three different future periods by considering GWR models from precedent similar periods. Additionally, we consider only street crimes and integrate information about their particular spatial and temporal patterns to predict areas where crimes are more likely to occur.

1.1 Predictors of crime & population at crime risk

Some elements of the build environment that are strongly correlated to crime and have being used for prediction include hospitals, schools, police stations, and population. Additionally, Twitter data have been integrated in crime analysis by considering their location [2], their topic [16, 9, 1], their sentiment [6], and by using a dictionary to select tweets that include specific words [15]. Regarding population information, census data are commonly used in calculating population at crime risk. However, population is not random in space and has varying patterns during working days and hours compared with home or leisure times. Hence, recent studies integrate dynamic population models. The ancillary dynamic information can be extracted from social media data [11, 12], mobile phone data [13], or spatial data and imagery analysis like LandScan Global Population Database, provided by Oak Ridge National Laboratory.

1.2 Research objective

The objective of this study is to integrate and test geosocial media data as variables in GWR crime prediction models. Geosocial media data are free and easier to obtain than authoritative data. In addition, they can be used to produce ambient models compared to the static nature of census data. Furthermore, there are at least two cases in which retrospective methods that require historical data of more than one past period cannot be used in prediction and thus regression-based approaches are promising alternatives. The first case is to estimate the crime occurrence in an area for which data are not available. An area with similar profile and availability of data can be used to deliver GWR regressions for the "unknown" area. The second case is when there are crime and predictor data for time t-1 and we want to estimate the crime prevalence in t by assuming that slight variations in time can be better represented in a generalized model of t-1 than the actual crimes of t-1.

2 A social media based GWR application for the prediction of crime

GWR is a modeling approach for spatially heterogeneous processes [3]. In the last decade, implementing GWR as a predictor increased substantially even if still controversial. GWR technique has the advantage of considering non-stationary variables and modeling local relationships between dependent and independent variables. Our strategy is to use recent past variables to create a model over the study area and predict crimes in the future. As for the parameters need to set for this tool, we used the "adaptive" bandwidth, as recommended in literature. For crime analysis, researchers are using a palette of interdisciplinary variables in regression models to understand crime distribution, not only spatially. We define explanatory variables from social data to examine, if being the only predictors in a model can favor the understanding of spatial crime distribution. Two types of variables are defined and tested:

Population at crime risk (PopCR) and crime-related tweets (CrimeTW). We used geolocated Twitter data in calculating PopCR, adapting the methodology described by Kounadi et al. 2017 for translating density of tweets into population density with the point-based areal interpolation method. Residential population is calculated for large geometries, such as census tracks or neighborhoods. Density-weighted interpolation disaggregates the data included in large geometries (source zones, i.e. residential population values) by using control point data (i.e. the spatial distribution of tweets) in order to obtain a new variable for target zones, which are smaller polygon geometries (i.e. grid cells). In addition, to define relevant geolocated tweets for our analysis, we first analyzed the temporal distribution of street crimes. Then, we chose the days and times of crime peaks and only for those timeframes, geolocated tweets were extracted and introduced in PopCR models. Second, we extracted CrimeTW by filtering the entire geolocated Twitter dataset for crime related terms. After preprocessing the data, we noticed four sources that post constantly about crimes: City of Portland 911 feed, City of Portland Fire/EMS feed, TTN POR traffic, Multnomah County Sherriff feed. The last source is an unofficial posting of the East County using a scanner feed from police information. Practically, we are using the intensity per polygon for these two independent variables, PopCR and CrimeTW, in order to explain the dependent crime counts. The analysis was performed for the three periods (one week, two months, three months), as well as for two cell sizes (0.006 km^2 small size called cell A and 0.023 km^2 large size called cell B).

3 Case study: Portland

The study area is in Portland, the largest city from the state of Oregon in the USA. The size of the study area is 382.6 km^2 and includes an estimated population of 640,000 people in 2016. Data for this case study contain crime occurrences in 2015 and 2016 from call-for-service data from the Portland Police Bureau, and Twitter data from 2015. We only consider street crime types that affect the mobile population, which include assault, disturbance, gang related crimes, robbery, shooting, stabbing, drugs, liquor, prostitution, and gambling. We tested three periods from the two years for which we downloaded the crime data: One week (1st week of March: 559 crimes in 2015 and 538 in 2016); two months (March to April: 5,129 crimes in 2015 and 5,386 in 2016); and three months (March to May: 7,987 crimes in 2015 and 8,417 in 2016). Twitter data were obtained using the Twitter API. We only used tweets that had the geolocation activated, so that we know the exact coordinates of the message. For the PopCR variable, we extracted the tweets from the three periods in 2015 and kept only those that showed a peak of crime events (e.g. weekend nights). Namely, we processed the temporal information from the tweets and we extracted the ones which have a corresponding time slot with crime at its peak (e.g. street crime type has temporal peaks during weekend nights, so we extracted the tweets from weekend nights to be control points for PopCR). For the second variable, CrimeTW, the entire filtered data set was used in all three periods (the tweets from the four aforementioned users).

4 Results

The analysis was performed for three time periods (one week, two months, three months), as well as for two cell sizes. The first size, called cell A, covers a rather small area of 0.006 km^2 (total number of cells: 66,841), while the second one, called cell B, covers an area of 0.023 km^2 (total number of cells: 16,753). We applied the analysis six times, one for each combination of time period and cell size, so that we can have a first exploration on the effects

Table 1 Evaluation of GWR models (three prediction periods and two cell sizes).

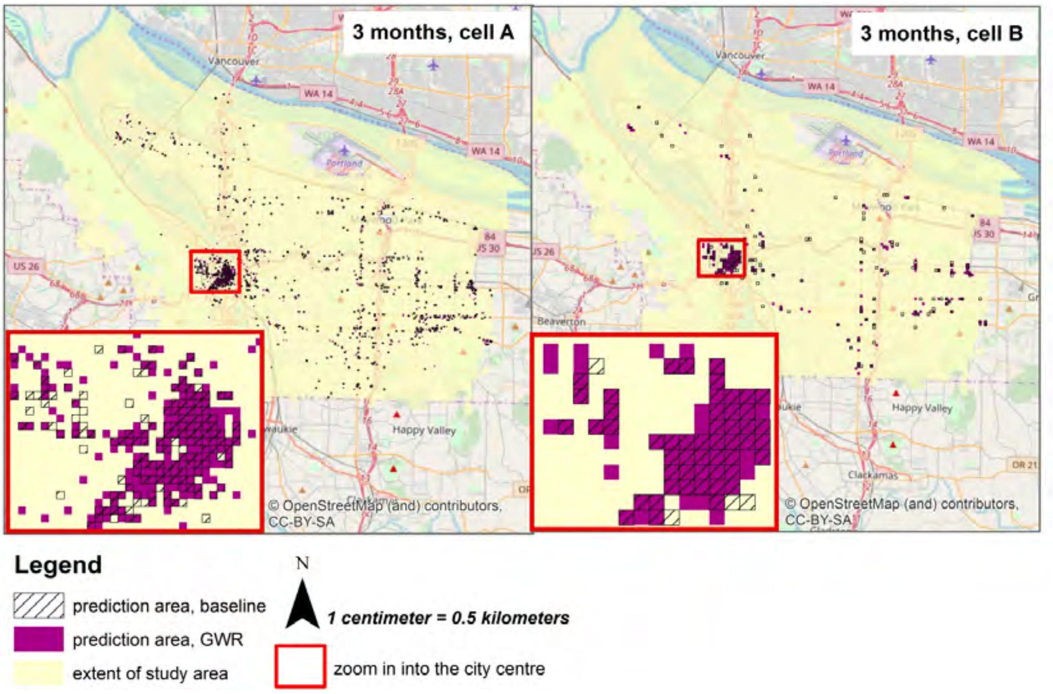
	Cell A			**Cell B**		
Prediction period	*1 week*	*2 months*	*3 months*	*1 week*	*2 months*	*3 months*
AICc	123,245.67	55,119.85	96,639.98	22,994.95	37,001.81	23,664.39
R-squared	0.13	0.44	0.49	0.54	0.67	0.61

of these parameters on the estimation models and prediction accuracy. Table 1 shows the Akaike Information Criteria (AICc) and R^2 values of the six GWR models. The AICc scored the lowest value for cell B and a three month period and R^2 has the highest value for cell B and a two month period. In general, we observe that larger cell sizes and longer time periods (two or three months) give considerably better results compared to smaller cell sizes and shorter prediction periods. The fact that different spatial aggregation chosen in analysis produce different results is one well-known issue in geography, named modifiable areal unit problem (MAUP), and in many cases the aggregation to a larger cell size can yield to better results. To identify hotspot areas in 2016 using data from 2015, we selected areas with high prediction values. Also, to compare the results among the models we standardized the size of the prediction area to approximately 1% of the total number of cells. The size of the total area is 382.6 km^2 and the size of the prediction area is 3.9 km^2. This amounts to 668 A cells and 168 B cells (selected based on their prediction values). Since the prediction area among models was the same and about 1, the denominator of the prediction accuracy index (PAI), which represents a common accuracy index in crime analysis [5], was canceled and thus we considered only the denominator, which is essentially the hit rate (i.e. success rate). Furthermore, we compared the GWR models with baseline models. As a baseline model we define a simple exploratory approach‘, where cells with the highest crime intensity in 2015 define hotspots in 2016. Again, the amount of cells was 668 A cells and 168 B cells so as to compare only hit rates. Figure 2 shows the results of one period for cells A and B and a zoomed-in section where we observe the divergence between baseline and GWR hotspots.

Street crimes in 2016 were used to calculate the hit rate. Although the prediction area was defined to be quite small, all GWR models resulted in a hit rate between 23.2% and 27.8% (Table 2). In particular, GWR models identified more than 20% of the crimes being located in 1% of the area. On the other hand the hit rate of the baseline models ranges between 12.4% and 28.8%. In Table 3, we provide a summary of the predictive efficiency analysis by calculating the mean hit rate by predictive period, cell size, and method. Although in Table 1 it is apparent that the larger cell size creates better GWR models, in terms of the prediction efficiency the smaller cell size predicts more crimes than the larger cell size (Table 3). Additionally, the larger the prediction period the better the results are. Finally, hotspot areas defined from GWR models predict a significantly higher number of crimes than areas defined from baseline models.

5 Discussion

The success of a prediction that employs spatial regression analysis depends on explanatory variables. In the current study we used variables from one main data source, Twitter, and we calculated how much of the spatial distribution of crimes is explained by tweets. We based our analysis on the 2015 information in order to get hotspots for 2016. Six GWR models were created using as dependent variables street crime data and independent variables from two tweet subsets (i.e. population at crime risk and crime related tweets) both from the year


3 months, cell A
3 months, cell B
Legend
prediction area, baseline
prediction area, GWR
extent of study area
N
1 centimeter = 0.5 kilometers
zoom in into the city centre

© OpenStreetMap (and) contributors, CC-BY-SA

Figure 1 Overlapping of prediction areas resulting from GWR and baseline methods for a period of three months and two different cell sizes (cell A = 0.006 km^2 and cell B = 0.023 km^2).

Table 2 Predictive efficiency grouped by period, cell size, and method (GWR vs Baseline).

Prediction period	Cell size	Method	Hit Rate
1 week	A	GWR	25.4
		Baseline	13.5
	B	GWR	23.4
		Baseline	12.4
2 months	A	GWR	27.8
		Baseline	26.3
	B	GWR	23.8
		Baseline	14.1
3 months	A	GWR	27.6
		Baseline	28.8
	B	GWR	23.2
		Baseline	23.2

Table 3 Average Hit Rate by the three parameters (predictive period, cell size, and method). * Indicates higher Hit Rate among comparisons.

Mean values of Hit Rate		
cell size	cell A*	24.9
	cell B	20.0
length of prediction period	1 week	18.7
	2 months	23.0
	3 months*	25.7
method	GWR*	25.2
	Baseline	19.7

2015. The predictive efficiency of GWR outcomes was higher than a baseline model that considered the past areas of high crime density as being the same for the next period. In order to account for effects of the spatial resolution and temporal differences, we selected three testing periods (one week, two months, and three months) and two different grid cell sizes, namely small and large. Results for the year 2015 show that by using the larger cell size the GWR models explain more variance of crime distribution patterns than by using the smaller cell size. However, when it comes to prediction efficiency the smaller cell size yielded higher accuracy than the larger one. This may be a characteristic of the crime type in question and possibly high number of repeats and/or near-repeats. Also, the accuracy varies by prediction period with the longest analyzed period (i.e. three months) having the highest prediction efficiency. The main limitations of our approach is the under-representativeness of Twitter sample data (not each person is tweeting; not all users are using geolocation actively), the possibility of having non-reported crime occurrences that we did not evaluate, multicollinearity issues for GWR and the MAUP, which is not sufficiently addressed by two different cell sizes. To compensate for these limitations, our future work on this topic will employ the next three additions. First, we want to use additional types of social media platforms (e.g. Foursquare or Flickr) for the development of predictors. Second, we will perform multiple case studies for which the accuracy and completeness of crime data will be tested. Last, we will extensively and empirically test the parameters of our approach, including but not limited to the spatial resolution of the models and the temporal resolution of the prediction periods.

References

1 Meshrif Alruily. *Using text mining to identify crime patterns from arabic crime news report corpus*. Thesis, DeMontfort University, 2012. URL: http://hdl.handle.net/2086/7584.

2 Johannes Bendler, Tobias Brandt, Sebastian Wagner, and Dirk Neumann. Investigating crime-to-twitter relationships in urban environments-facilitating a virtual neighborhood watch. In *ECIS 2014*, 2014. URL: http://aisel.aisnet.org/ecis2014/proceedings/track11/10/.

3 Chris Brunsdon, A Stewart Fotheringham, and Martin E Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4):281–298, 1996.

4 Joel M Caplan and Leslie W Kennedy. Risk terrain modeling compendium. *Rutgers Center on Public Security, Newark*, 2011.

5 Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1):4–28, 2008.

6 Xinyu Chen, Youngwoon Cho, and Suk young Jang. Crime prediction using twitter sentiment and weather. In *Systems and Information Engineering Design Symposium (SIEDS), 2015*, pages 63–68. IEEE, 2015.

7 JE Eck, S Chainey, JG Cameron, M Leitner, and RE Wilson. Mapping crime: Understanding hot spots. national institute of justice. *Washington, DC*, 2005.

8 A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

9 Matthew S Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014. doi:10.1016/j.dss.2014.02.003.

10 Philip Glasner and Michael Leitner. Evaluating the impact the weekday has on near-repeat victimization: A spatio-temporal analysis of street robberies in the city of vienna, austria. *ISPRS International Journal of Geo-Information*, 6(1):3, 2016.

11 Ourania Kounadi, Alina Ristea, Michael Leitner, and Chad Langford. Population at risk: using areal interpolation and twitter messages to create population models for burglaries and robberies. *Cartography and Geographic Information Science*, pages 1–15, 2017. doi:15230406.2017.1304243.

12 Nick Malleson and Martin A Andresen. The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, 42(2):112–121, 2015. doi:10.1080/15230406.2014.905756.

13 Nick Malleson and Martin A Andresen. Exploring the impact of ambient population measures on london crime hotspots. *Journal of Criminal Justice*, 46:52–63, 2016.

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

15 Alina Ristea, Justin Kurland, Bernd Resch, Michael Leitner, and Chad Langford. Estimating the spatial distribution of crime events around a football stadium from georeferenced tweets. *ISPRS International Journal of Geo-Information*, 7(2):43, 2018. URL: `http://www.mdpi.com/2220-9964/7/2/43`.

16 Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 231–238. Springer, 2012.