Title

Using Social Media to Measure Temporal Ambient Population: Does it Help Explain Local Crime Rates?

Permalink

https://escholarship.org/uc/item/9nn8z06k

Journal

Justice Quarterly, 36(4)

ISSN

0741-8825

Authors

Hipp, John R
Bates, Christopher
Lichman, Moshe
et al.

Publication Date

2019-06-07

DOI

10.1080/07418825.2018.1445276

Peer reviewed

Using Social Media to Measure Temporal Ambient Population: Does it Help Explain Local Crime Rates?

John R. Hipp*

Christopher Bates

Moshe Lichman

Padhraic Smyth

February 14, 2018

Post-print. Published in Justice Quarterly:

Word count: 8,785

Word count (including references): 10,218

Running Head: "Tweets and crime"

* Department of Criminology, Law and Society and Department of Sociology, University of California, Irvine. Address correspondence to John R. Hipp, Department of Criminology, Law and Society, University of California, Irvine, 3311 Social Ecology II, Irvine, CA 92697; email: john.hipp@UCI.edu. This research is supported in part by the National Institute of Justice under Grant 2012-R2-CX-0010 and the National Science Foundation under Grant No. IIS 132057.

Using Social Media to Measure Temporal Ambient Population:

Does it Help Explain Local Crime Rates?

Abstract

A challenge for studies assessing routine activities theory is accounting for the spatial and temporal confluence of offenders and targets given that people move about during the daytime and nighttime. We propose exploiting social media (Twitter) data to construct estimates of the population at various locations at different times of day, and assess whether these estimates help predict the amount of crime during two-hour time periods over the course of the day. We address these questions using crime data for 97,428 blocks in the Southern California region, along with geocoded information on tweets in the region over an eight month period. The results show that this measure of the temporal ambient population helps explain the level of crime in blocks during particular time periods. The use of social media data appear promising for testing various implications of routine activities and crime patterning theories, given their explicit spatial and temporal nature.

Keywords: spatial crime, routine activities, social media, ambient population.

Bio

John R. Hipp is a Professor in the departments of Criminology, Law and Society, and Sociology, at the University of California Irvine. His research interests focus on how neighborhoods change over time, how that change both affects and is affected by neighborhood crime, and the role networks and institutions play in that change. He approaches these questions using quantitative methods as well as social network analysis.

Christopher Bates is a Ph.D. student in the department of Criminology, Law and Society, at the University of California, Irvine.

Padhraic Smyth is a Professor in the Department of Computer Science, with a joint appointment in the Department of Statistics, at the University of California, Irvine. He is also the director since 2014 of the UCI Data Science Initiative. His research interests include machine learning, data mining, pattern recognition, and applied statistics and he has published over 160 papers on these topics.

Moshe Lichman is a Ph.D. student in the Department of Computer Science at the University of California, Irvine.

Using Social Media to Measure Temporal Ambient Population: Does it Help Explain Local Crime Rates?

A bedrock insight of research on the location of crime at places is that the number of crime events at locations is a function of the presence of motivated offenders, suitable targets, and the absence of capable guardians. The role of crime opportunities in explaining spatial and temporal crime concentration originates from routine activities theory (Cohen and Felson 1979) along with the insights of crime patterning theory (Brantingham and Brantingham 2008). The body of research exploring the micro-location of crime typically faces challenges related to precise measurement, such as measuring the presence of people at locations at various times of day, much less distinguishing between offenders, targets, and guardians. Although research has shown crime tends to cluster at micro-spatial locations within cities (Sherman 1995; Sherman, Gartin, and Buerger 1989; Weisburd 2015) and crime tends to cluster temporally at various times of day and days of week (Andresen and Malleson 2015; Ashby and Bowers 2013; Ceccato and Uittenbogaard 2014; Haberman and Ratcliffe 2015), measuring the spatial and temporal determinants of these crime patterns is more challenging.

The number of targets and guardians in a location can be approximated by simply measuring the number of people in the area (Gove, Hughes, and Galle 1979; Harries 2006). The U.S. Census provides a straightforward measure of the residential population that is easily obtainable by researchers. However, a limitation of the residential population measure is that people do not remain in their homes all hours of the day and night, but rather engage in routine travel behavior. The population at a place, including residences and businesses, is not constant but rather shifts dynamically with time. Boggs (1965) first stressed the importance of measuring the ambient population at risk for crime victimization, rather than simply the residential population when computing crime rates. The ambient population is the population at a place at a specific time period. Indeed, a key insight of the routine activities theory is that opportunities for crime are dependent on both space and time. Thus, to fully understand where crime clusters spatially and promote effective crime prevention strategies, a reasonable estimate of the number of offenders and guardians present at a location at a particular point in time needs to consider where residents go during the daytime, evening, and nighttime.

Researchers have used various measures related to land use and employment to approximate the average ambient population in a location. For example, researchers (Bernasco and Block 2011; McCord and Ratcliffe 2009) utilized land use characteristics with the premise that certain types of land use—i.e., commercial, industrial, office buildings, parks, etc.—will be more likely to attract people during the daytime or nighttime compared to other land uses. However, land use information is a quite crude proxy of the actual number of individuals in an area. Another approach measures the number of employees in an area (Hipp 2007; Wo, Hipp, and Boessen 2016). The employee strategy is arguably a somewhat better proxy for the number of people. The number of employees of white collar and industrial jobs captures the actual number of workers in an area, whereas the number of retail jobs serves as both a measure of the number of workers as well as a proxy for the presence of some number of patrons of these stores, since retail stores must attract patrons if they are to remain in business. However, the ratio of patrons to employees at businesses is unclear and will vary among types of businesses. Furthermore, routine activities theory highlights that crime events will be more likely with the convergence *in both space and time* of offenders and targets along with a lack of guardians. Nearly all available proxies for the number of people do not provide information on the number of people *in a location at different times of day or days of the week*, failing to contribute information on temporal shifts of the population at risk of victimization.

In this study, we use a relatively new, and cost effective, measure of the number of people in an area by utilizing information from publicly available social media data. We measure the number of tweets emitted by individuals in an urban environment during various hours of the day and days of the week at specific locations. Although this is certainly an imperfect measure of the number of people at a location during particular time periods, we explore here whether it nonetheless provides some analytical purchase to understanding crime patterns. There is a growing body of research that utilizes social media to capture the presence of people throughout diverse locations of the city (Frias-Martinez and Frias-Martinez 2014; Lee, Davis, and Goulias 2016; Malleson and Birkin 2014). For example, a number of recent studies have used geolocated information from tweets, along with information on the home location of persons, to create alternative measures of "neighborhoods" that are based on the actual activity patterns of residents, rather than any other type of boundaries (Anselin and Williams 2015; Cranshaw, Schwartz, Hong, and Sadeh 2012; Shelton, Poorthuis, and Zook 2015). There have also been some studies that have used information on tweets to assess their relationship with the number of crimes at various small scale locations (Malleson and Andresen 2015a; Wang and Gerber 2015; Wang, Brown, and Gerber 2012), and one study analyzed temporal/spatial hot spots based on social media (Malleson and Andresen 2015b). However, we will argue below that although this nascent research provides exciting possibilities, the literature has not fully utilized social media to capture the spatial and temporal patterns of ambient populations and its association with crime rates at various times of day. Specifically, we use information on the typical spatial/temporal patterns of people, along with information on the date and time of crime events, to assess the relationship between this novel measure of the ambient population and crime events during different times of day.

The paper proceeds by first introducing the routine activity perspective and crime patterning theory to explain how the convergence of offenders, targets, and guardians influences crime concentration. Next, we review the previous literature on measuring the population at victimization risk throughout varying times of the day. Following that, we introduce research that utilized geocoded social media data, specifically Twitter, to measure ambient population. We then describe our data and analytic methods, present the results, and conclude by discussing the implications of our findings.

Literature Review

Routine activities theory and crime pattern theory

Routine activities theory posits that crime events are more likely to occur when there is a spatial and temporal confluence of motivated offenders and suitable targets, along with the absence of capable guardians (Felson and Boba 2010). Crime patterning theory (Brantingham and Brantingham 2008) has built upon routine activity theory by asserting that exposures to crime opportunities are governed by the spatial layout of the city and individual travel patterns.

People typically exhibit routine daily travel patterns in where they go for activities, and when they travel to these activity locations (Brantingham and Brantingham 2008). A consequence of these routine travel patterns is that there will likely be a relatively consistent pattern of the number of people converging at certain locations at specific times of day and days of the week. For example, people tend to sleep at home at night, wake up in the morning and then go to school or work. After the work day is over, late in the afternoon or early in the evening people either return home or else engage in various entertainment activities, such as dining out, shopping, or visiting friends. Eventually, people return home at night to sleep until morning again. The various repetitious spatial and temporal traveling patterns due to routine activities are posited to affect the patterning of crime events (Groff 2008).

The travel patterns and activity spaces of individuals on weekdays will differ on the weekend as less time is spent in activity locations for work, and more time is spent in home and entertainment areas (Brantingham and Brantingham 2008). Previous research demonstrates crime concentrates in different places on weekdays versus weekends (Andresen and Malleson 2015; Ceccato and Uittenbogaard 2014). Spatial and temporal patterns in crime clustering are due to daily routine activities that bring a convergence of potential populations of offenders, targets, and guardians. Researchers must construct better estimates of measuring the spatial and temporal patterns of where offenders and targets go throughout the day and night to fully understand where crime will occur and concentrate.

A further challenge is that this confluence of offenders and targets is not only a spatial one but also a temporal one. Crime and place research has focused on obtaining more precise spatial measures of crime events along with various characteristics of the physical and social environment and recognized that these features of the environment implicate temporal crime patterns (McCord and Ratcliffe 2009; Tompson and Townsley 2010). Recent research is just beginning to explore the temporal flows of people and the spatial concentration of offenses (Felson and Boivin 2015; Stults and Hasbrouck 2015). In part, this is likely due to the challenges of obtaining information on where people are at during specific times. On the one hand, advances in the electronic storage of data have made it more feasible to get information on the time and location of crime events. Some recent studies have explored the temporal patterns of crime events (Boessen 2014; Haberman and Ratcliffe 2015). Such studies sometimes combine information on the physical environment of a location along with the timing of crime incidents to detect areas that may be more vulnerable to crime events at certain times of day (Haberman and Ratcliffe 2015). Although such studies are informative, they are hampered by their reliance on crude proxies for the number of people that are actually in a location during certain times of day that often only distinguish between “daytime” and “nighttime” population.

There are only a handful of recent studies that have attempted to measure the number of people at a location (Andresen 2006; Andresen and Jenion 2010; Boivin 2013). One way to measure the ambient population at a location is to conduct a survey on where residents are actually spending their time throughout the day. Surveys on time use and transportation have been used by researchers to study crime opportunities based on the average temporal flows of people throughout space (Felson and Boivin 2015; Stults and Hasbrouck 2015). However, survey data is often aggregated to larger spatial units (e.g., census tracts) for anonymity, which hampers the effectiveness of survey data in studying temporal changes in population at a micro-spatial scale, such as a street segment.

Another way to proxy the number of people at a location measures the average ambient population with administrative data such as Census and land use. For example, one study used data from LandScan to capture ambient population (Andresen 2011). The LandScan project uses information on residential population, land use characteristics, and the locations of businesses to create estimates of the daytime ambient population (based on the locations of businesses, amenities, etc.) and the nighttime ambient population (this mainly comes from the Census, which captures residential population). While an estimate of overall ambient population is useful, the measure still relies upon proxies for creating estimates. Furthermore, the Andresen (2011) study only used a smoothed estimate of the number of people in a location, on average, over all hours of the day, and did not account for the actual population in a location at a particular time period nor did it account for the time of day of the crime events. Haberman and Ratcliffe (2015) measured various crime attractors and assessed the correlation between them and crime at specific times of the day. Whereas this study could test the relationship between certain land use categories and crime rates at various times of day, the static land use measures did not provide temporal information on the population present at different times of day or days of the week. Boessen (2014) used information on the nighttime population (from the Census), and the location of jobs and schools to create estimates of the daytime population. The study computed the daytime population by subtracting out the number of school-aged children and the number of employees who worked outside the local census block and adding in the number of children in the area if a school was located in a block and the number of employees in the block. The jobs and schools approach provided a reasonable proxy for the daytime and nighttime population; however, it does not make more fine-grained temporal distinctions.

Social media as a measure of ambient population

There is a clearly a need for alternative measures capable of capturing the population that is present in a location at various times of the day. We propose using social media as a way of capturing this construct. There is a small but growing literature that uses social media to measure the presence of people in the landscape (Malleson and Andresen 2016). The advantage of using social media data is that it provides an unobtrusive measure capturing where people actually go during the day rather than relying on retrospective survey responses.

Among the various social media platforms that provide geolocated user information (e.g., Foursquare, Instagram, Facebook), the Twitter platform is uniquely suited to measuring ambient population. When a person tweets, for individuals who opt to make their geolocation information public when they tweet, the tweet will contain their present latitude/longitude, as well as a date/time stamp. Additionally, there is a user ID associated with each tweet. Multiple tweets from the same person over a short period of time can be removed through the user ID. Therefore, researchers can more accurately construct approximate estimates of the number of *unique* persons at a location during a particular period of time. For all of these reasons, a growing body of scholarship has used Twitter information to capture the spatial/temporal patterns of populations (Leetaru, Wang, Cao, Padmanabhan, and Shook 2013; Lenormand, Picornell, Cantu-Ros, Tugores, Louail, Herranz, Barthelemy, Frias-Martinez, and Ramasco 2014; Steiger, de Albuquerque, and Zipf 2015)

A growing body of literature has used tweets in various research designs and demonstrated promising results. Research has demonstrated that the concentration of geolocated tweets is correlated as expected with land use characteristics. For example, studies have assessed the ability of geolocated tweets to measure the land use of areas, by comparing the number of tweets at varying times to the land use in locations (Frias-Martinez and Frias-Martinez 2014; Lee, Davis, and Goulias 2016; Malleson and Birkin 2014). A study collected geocoded tweets over seven weeks in London, Manhattan, and Madrid and found that the spatial and temporal pattern of tweets provided patterns consistent with the land use data collected in these cities (Frias-Martinez and Frias-Martinez 2014). Other research has determined characteristic words that are used near one’s home, and words that are used near work or other activities, in an attempt to measure the spatial activity patterns of people (Malleson and Birkin 2014).

Additionally, a body of research has used geolocated tweets and found them to be useful in constructing patterns of where people go, and then using this information to construct novel measures of “neighborhood”. For example, one study used the locations of tweets, and where tweeters reside, along with clustering routines, to construct “neighborhoods” (Anselin and Williams 2015). Other research has used the geographic information of tweets and their content to capture the language spoken in various parts of the city as another way to measure specific neighborhoods based on these linguistic patterns (Birkin, Harland, and Malleson 2013). Another study used social media information on how people moved throughout the day to create location-based social networks to capture social interactions between geographic areas, and then constructed “neighborhoods” based on these patterns (Wakamiya, Lee, and Sumiya 2013). Shelton et al. (2015) examined over 2 years of Twitter geolocated data and tweet content to create aggregate activity spaces of neighborhoods. Another research group used a broadly similar approach based on check-ins to the social media site Foursquare, and constructed neighborhoods based on activity patterns that they termed “livehoods” (Cranshaw, Schwartz, Hong, and Sadeh 2012). In this approach, persons “check-in” to a location by noting when they are at a particular location. Although this does not provide information on people who are not at one of the predefined locations in the service, it still can be quite useful information for understanding general spatial patterns of residents.

Despite the promise of using geocoded tweets to capture the presence of persons at various locations at various times to predict crime rates in those locations, only a few recent studies have utilized Twitter information in this way. For example, one study used information on tweets, including tweets by daily news sources and the content of the tweets, to assess their relationship with burglary events at specific times over six months in the city of Charlottesville (Wang, Brown, and Gerber 2012). Although there may be endogeneity issues given that the content of daily news tweets are often in response to the crime incident rather than predictive of it, this still highlights the potential utility of the approach. Another study used tweets to capture the sentiment of nearby persons *after* a crime event, which also raises the same endogeneity concerns (Kounadi, Lampoltshammer, Groff, Sitko, and Leitner 2015). Another approach is to use tweets to capture the presence of population in an area in an effort to obtain a better estimate of the population "exposure" for crime events, regardless of the timing of the offense (Malleson and Andresen 2015a). Although the use of Twitter for the average population-at-risk provides a better estimate of the ambient population, on average, Malleson & Andresen (2015a) did not attempt to assess the population or crime at specific times of day. Furthermore, a recent study by Malleson and Andresen (2016) examining various measures for the average population-at-risk, including census and social media, found Census workday population to be the most appropriate measure. Another study by Malleson and Andresen (2015b) did incorporate time of day information with Twitter data, although their goal was to create a handful of spatial/temporal hot spots and compare them to the spatial/temporal patterning of crime events 10 years earlier.

A handful of studies have used Twitter information in a predictive framework with local crime. These studies illustrate the potential of this approach, although they typically focus on data over a relatively short period of time. For example, one study used the content of tweets to predict the likely next venue someone would go to, and then assessed the correlation between crimes and tweets in Chicago over a one month period (Wang and Gerber 2015). Although it showed promising results, this study did not take into account time of day of crime events and limited the sample to only those who posted at least 20 tweets in the month. Another study used data from San Francisco and assessed the correlation between the number of tweets at locations and the number of crime events over a three month period, although this study also did not account for time of day (Bendler, Brandt, Wagner, and Neumann 2014).

Tweets and crime

Summary

Given that recent research has shown the geolocated tweets may be quite useful for tracking the spatial and temporal location of the population, we explore here whether this measure is systematically related to spatial and temporal levels of crime. In this study we use social media information based on geolocated tweets to capture the relative patterns of people as they move about during the day and night. Thus, we are not attempting to capture the actual number of people at a location at a particular point in time, but rather we use geolocated tweets over a 7-month period to construct estimates of the relative spatio-temporal pattern of persons across the study site. We then test the relationship between these spatio-temporal patterns of geolocated tweets and crime events in small geographic units (blocks) during 2-hour time periods of the day over a one-year period. We also take into account the possible difference in patterns that occurs on weekdays versus on weekends. We test whether tweets operate as a measure of the relative population in a location while controlling for the residential population and the socio-demographics of that population.

Data and methods

Data

The crime data come from the Southern California Crime Study (SCCS). In that study, the researchers made an effort to contact each police agency in the Southern California region[1] and request address-level incident crime data; many of the agencies were willing to share their data. We use data from the year 2012 given that this year is closest to our social media data and provides us with the largest number of cities with crime data. The crime data covers 76.5

[1] The region is defined as including five counties: San Bernardino, Riverside, Los Angeles, Orange and San Diego.

percent of the region's population. The data come from crime reports officially coded and reported by the police departments. We classified crime events into six Uniform Crime Report (UCR) categories: homicide, aggravated assault, robbery, burglary, motor vehicle theft, and larceny. Crime varies by time of day, and the category of time used needs careful consideration for statistical analysis (Felson and Poulsen 2003). We use two-hour time periods throughout the week (given that slicing to narrower time windows provided little analytical gain while making the data too sparse given the rarity of crime events) divided into 1) weekday (Sunday midnight to Friday 6pm) and 2) weekend (Friday 6pm to Sunday midnight).

Dependent variables

Crime events were geocoded to latitude–longitude point locations using a geographic information system (ArcGIS 10.2) and placed into the proper Census block. For almost all cities the geocoding match rate was above 95%, and the average across cities was 97.2%, suggesting an excellent match rate. Given that we know the day and time that the crime event was reported, we are able to code it to a particular 2-hour window on the day on which it occurred, and we make a distinction between weekdays (Sunday midnight-Friday 6pm) and weekends (Friday 6pm-Sunday midnight) given that these likely have different spatial and temporal patterns of where the ambient population is located. We classified crime events into five Uniform Crime Report (UCR) crime types: aggravated assault, robbery, burglary, motor vehicle theft, and larceny. We do not use homicides, as they are too rare at this micro temporal and spatial disaggregation. We do not use sexual assaults given the well-known reporting problems with this crime type. Given possible seasonality effects of crime events (Hipp, Bauer, Curran, and Bollen 2004; Sorg and Taylor 2011), we estimated models with crime measures taken only from the May to December period to match the time period of the Twitter data (models estimated using crime data from the entire year yielded a similar pattern of results).

Social media data

We collected data using the Twitter Streaming API over an 8 month period (from May 2015 to December 2015).[2] The Twitter Streaming API provides a push of tweets in real-time for a given search criteria, compared to the Twitter Search API which provides a pull of the tweets from the past for a given search criteria and a hard cap on query results. The Twitter Streaming API provides up to maximum of 1% to 40% of the full firehose of every single concurrent Tweet, due to data and infrastructure limitations. However, we bounded our query of the Twitter Streaming API to tweets with geolocation coordinates (GPS) located within the geographic boundaries of Southern California. Given that we limited our data gathering to tweets with geolocation information within Southern California, we remained safely under the Twitter Streaming API maximum limit of 1% of all tweets. Our dataset confidently contains all available geolocated tweets within Southern California, during the time period. (Driscoll and Walker 2014). Note that the fact that this collection period does not exactly match with the crime data is not problematic given that we are only trying to capture general relative patterns of when and where persons are at locations—that is, the *routine activities* of persons as identified by routine activities theory—and not attempting to place persons at a particular location at a particular day and time. By collecting Twitter georeferenced data over a multi-month period we obtain a more accurate picture of the general spatial/temporal patterns of people.

The metadata attached to a geolocated tweet in addition to the tweet message content include the time at which the tweet occurred, the location of the person based on

[2] https://dev.Twitter.com/streaming/

latitude/longitude, and the ID of the person. Geolocated tweets were placed into census blocks using the latitude/longitude data. Duplicate tweets by a user ID that occurred during a 2-hour period in the same block were eliminated. The geolocated tweets provide a count of the *unique* individuals known to be at a particular block during a particular two-hour window, and is what we use as our proxy for *temporal ambient population*. However, if the person tweets again at the same location during the next two-hour window, we would again count them as being present at the location during that next period. We then sum up the number of unique individuals at a particular block during a particular two-hour window over the 8 month period. To account for extreme values we log transformed this value (after adding 1); models with the untransformed measure always demonstrated worse fit (based on pseudo r-square values), and therefore we utilized the logged version of this measure.

Additional independent variables

We also account for a set of variables that are typically included in models estimating the geographic location of crime. Most of these measures are obtained from the U.S. Census or the American Community Survey 2008-12 5-year estimates. These are socio-demographic variables measured at both the block level and the ½ mile buffer surrounding the block (with an inverse distance decay). We measured *concentrated disadvantage* by combining four variables with a factor analysis and computing factor scores: percent at or below 125% of the poverty level; average household income; percent with at least a bachelor's degree; percent single parent households.[3] We measured *residential stability* by combining three variables in a factor analysis and computing factor scores: average length of residence; percent at least 5 years in residence;

[3] Given that only the percent single-parent households variable is available for blocks, we use synthetic estimation for ecological inference to impute the other variables (Cohen and Zhang 1988; Steinberg 1979). The imputation models use the following variables: racial composition, percent divorced households, percent households with children, percent owners, percent vacant units, population density, and age structure (percent aged: 0-4, 5-14, 20-24, 25-29, 30-44, 45-64, 65 and up, with percent 15-19 the reference category).

percent homeowners. We measured racial composition with measures of *percent black*, *percent Latino*, *percent Asian*, and a Herfindahl index of *racial/ethnic heterogeneity* (Gibbs and Martin 1962: 670) of five racial/ethnic groupings (percent white, black, Latino, Asian, and other race), which takes the following form:

(1) $$H = 1 - \sum_{j=1}^{J} G_j^{\ 2}$$

where G represents the proportion of the population of ethnic group *j* out of *J* ethnic groups, and subtracting from 1 makes this a measure of heterogeneity. Given that they can be crime attractors, we computed the *percent vacant units*. We also computed the *percent vacant lots*, as they may have a different impact on crime opportunities than vacant units (Raleigh and Galster 2015).[4] We computed the *percent aged 16 to 29* as this is the prime age of crime-prone population. Finally, we included a measure of the *residential population (logged)*.

In a second set of analyses, we assessed the ability of the tweets measure to provide unique information beyond other measures that are sometimes used as proxies for the ambient population. Specifically, we constructed a measure of *total employees (logged)* to account for workers in the area, and a measure of *retail employees (logged)*, to account for locations attracting patrons. These data come from Reference USA for 2010 businesses.

Methods

The outcome variable is whether or not a crime event occurred (for the particular type of crime) in that census block during that two-hour period during May to December of 2012. Given that there are almost no time periods in which more than one crime event occurred, we modeled the probability of occurrence of at least 1 crime event using logistic regression models. We corrected the standard errors to account for the fact that the data have repeated observations for

[4] The land use data was obtained from the Southern California Association of Governments (SCAG).

each census block. All models were estimated in Stata 13 with a maximum likelihood estimator. We assessed the relationship between our measure of temporal ambient population and crime at each time point separately to assess the efficacy of the measure at different hours of the day.

For all time periods, we estimated the relationship between the number of tweets (logged) and crime, but then also estimated subsequent models that included 1) quadratic and 2) cubic forms of tweets (logged). The linear version adequately captures the relationship.

Results

We begin by showing the relative number of tweets during various time periods, to get a sense of how well this measure is capturing the ambient population during these periods. Figure 1 shows that the fewest tweets tend to occur between 2-6am on either weekdays or weekends, with 6-8am being the next slowest period. Compared to the 4-6am time period, there are about 7 to 8 times as many tweets between noon and 8pm; tweets are particularly prevalent on weekends.

<<<Figure 1 about here>>>

The results from the estimated models are shown in Table 1; this table only displays the temporal ambient population variable results and suppresses the results for the control variables. Each cell represents a separate model (estimated on a particular time period). To explicitly compare the relative strength of this measure at different time periods, we plot the coefficient values for aggravated assault for weekdays over the hours of the day in Figure 2a (along with 95% confidence intervals). For aggravated assaults on weekdays we see that the strongest effects are detected when the temporal ambient population measure captures more people in a location between midnight and 4am. A 10 percent increase in tweets in a location results in a 9.5% log odds increase in aggravated assaults between midnight and 2am (exp(.9067/10)=1.095) and a 9.7% increase between 2 and 4am (exp(.9243/10)=1.097).[5] The strength of the relationship is weaker during other hours of the day as a 10 percent increase in tweets in a location results in from 4.1% to 7.2% increased log odds of aggravated assaults during other times of the day when the relationship is statistically significant.

<<<Table 1 about here>>>

<<<Figures 2a and 2b about here>>>

Turning to the relationship between the temporal ambient population and aggravated assaults on weekends, Figure 2b plots the coefficient estimates over the various hours of the day. We again see that the relationship is strongest with aggravated assaults between midnight and 4am, as a 10 percent increase in tweets is associated with a 10.7% to 11.3% increase in logged odds of aggravated assaults during this period. During the other time periods the relationship is similar to weekdays, as when significant, the log odds increase between 5.7% and 7.9% for a 10 percent increase in tweets.

Regarding robberies, we find a very robust relationship between our measure of temporal ambient population and robberies as the relationship is statistically significant in virtually all time periods (except 4-6am on weekdays and 8-10am on weekends). On weekdays, a 10 percent increase in tweets in a block during a time period increases the log odds of experiencing a robbery between 5.2% from 8-10pm and 8.8% from 10am-noon as shown in Figure 3a. On weekends the relationship is even stronger, ranging from an increase of 6% in log odds from 8-10pm to an increase of 12.4% in log odds from 4-6am as shown in Figure 3b. We see that the relationship with robberies is strongest from midnight to 6am on weekends. Thus, it appears that the presence of more temporal ambient population has a stronger relationship with violent crimes

[5] Given that the independent variable is log transformed, a .10 change in this logged variable represents an approximate 10% change in the number of tweets.

in the very early morning hours after midnight, which may represent an increase in potential targets in a context of few guardians given the limited number of people about at this time.

<<<Figures 3a and 3b about here>>>

For the three types of property crime, we find that the temporal ambient population remains a robust predictor. The presence of more temporal ambient population in a block is statistically significantly associated with more burglaries during most time periods on weekdays and weekends. On weekdays, a 10 percent increase in tweets during a time period is associated with increased log odds of burglaries between 3.6% from 8am-noon and 7.9% from 4-6am as shown in Figure 4a. On weekends, Figure 4b shows that the significant coefficients range between 4.2% for 8-10pm to 6.7% for 10am-noon. We generally do not see evidence that the presence of more temporal ambient population is associated with markedly more burglaries in the early morning hours compared to other hours of the day, in contrast to the violent crimes (the one exception is from 4-6am on weekdays). Instead, the presence of more temporal ambient population has a relatively consistent positive relationship with burglary rates regardless of the time of day, or whether it is a weekend or weekday.

<<<Figures 4a and 4b about here>>>

The relationship between temporal ambient population and motor vehicle thefts typically appears strongest in the afternoon and into the early evening. The relationship with motor vehicle theft on weekdays is positive and statistically significant for all time periods except 4-8am. A 10 percent increase in tweets on a weekday is associated with an increase in log odds of motor vehicle thefts of between 3.4% from 8-10pm and over 8% between noon and 6pm. In contrast to the violent crime results, we observe that the relationship between temporal ambient population and motor vehicle thefts is actually strongest in the afternoon hours. On both weekdays and weekends the positive relationship is strongest between noon and 6pm.

<<<Figures 5a and 5b about here>>>

Finally, we find that the relationship between temporal ambient population and larcenies is relatively robust. This relationship is positive and statistically significant in virtually all time periods on weekdays and all but 4-6am on weekends. The size of the effect for a 10 percent increase in tweets on weekdays ranges from a log odds increase in larcenies of 4.9% between 6-8am to between 8-9% between 10am and 6pm. On weekends the range is from 4.7% from 8-10am to 9.6% from 10am- noon. The relationship between temporal ambient population and larcenies is consistently strongest during the daytime from about 10am to 6pm on both weekdays and weekends.

<<<Figures 6a and 6b about here>>>

Results when including measures of employees

The results just discussed demonstrated that using tweets is a promising way to capture the ambient population at different times of day. For our final set of analyses we asked whether tweets can provide information *above and beyond* what is obtained by using a different proxy of ambient population that has been employed in the literature: information on the number of total and retail employees in the area. In these models, we simultaneously account for the number of total employees and retail employees in both the block and the surrounding ½ mile (with a distance decay), along with our regular set of control variables. These models are asking whether tweets provide additional information that is of use to criminologists.

The results from the estimated models are shown in Table 2, and whereas we observed the expected consistently weaker results compared to Table 1, there nonetheless are relatively strong relationships with crime at many time periods. The largest reductions for the size of the tweets coefficients are observed in the robbery and larceny models, suggesting that the number of employees is a proxy for ambient population for these crime types. The smallest coefficient reductions are observed for the aggravated assault models. In the first two rows of Table 2 we see that blocks with more tweets during most time periods experience more aggravated assaults, even controlling for the number of total and retail employees. A larger temporal ambient population is associated with more aggravated assaults on weekends between 10am and 4am; on weekdays this positive relationship is observed in the early part of the day (6am to 2pm) and in the evening (6pm to 4am).

<<<Table 2 about here>>>

The pattern is similar for the other crime types. Although the relationship between ambient population and robberies is reduced when accounting for the presence of total and retail employees, there nonetheless remain some robust effects. The tweets measure has a particularly strong relationship in some of the overnight hours, and shows a relatively consistent significant relationship in the afternoons of both weekdays and weekends. Although tweets have a relatively weak relationship in the overnight hours with burglaries and motor vehicle thefts, during the daytime starting at 10am into the evening the relationship is relatively consistent. Finally, whereas the relationship between ambient population and larcenies is weakened when accounting for total and retail employees, it nonetheless remains a consistent positive relationship from 10am to 4am.

Ancillary models

We also estimated ancillary models to assess the robustness of our results. First, we assessed any additional nonlinearity in the relationship between the number of tweets and crimes by estimating models including the quadratic of logged tweets, and models that also included the cubic of logged tweets. These models never indicated an improvement in model fit, indicating that the logarithmic relationship best captures the relationship between tweets and crime.

Second, a limitation of crime data is that the time of crime events is not always perfectly known. This is particularly the case for property crimes, as the time recorded for such events is typically the time the crime was detected and hence reported. For a burglary or motor vehicle theft, there can be considerable uncertainty about when the crime actually occurred. In contrast, the time that violent crimes such as aggravated assaults or robberies occurred is typically much better known and reported. This is a challenging problem to address, but as one way to assess whether this impacts our results, we created alternative measures of tweets that summed the number of tweets in: 1) the previous 4 hours; 2) the previous 6 hours; 3) the previous 8 hours. We then estimated ancillary models using these alternative measures that capture the number of tweets over a longer time period in an attempt to match the possible temporal uncertainty of the property crime events. There was no evidence that model fit was improved noticeably in the property crime models, as would be expected based on the temporal uncertainty of these crime types. The results were often similar, and whereas there were a few time periods in which the model fit slightly improved using these longer time lags for the property crimes, there were just as many time periods in which the model fit slightly improved for the violent crime types. So despite the temporal uncertainty for property crime events, using a larger temporal catchment period for tweets does not change the results.

Discussion

This study has employed a relatively new approach to measuring the ambient population in small geographic units—the presence of more geolocated Twitter activity. An advantage of this method is that it provides an estimate of the typical number of people in a location at a particular time of day. The spatial and temporal precision is desirable for linking with the predictions of routine activities theory and crime patterns theory, as these theories posit a spatial and temporal confluence of offenders and targets that is otherwise difficult to measure. By combining this social media information with the day and time of crime events, we have shown that despite the imperfections of this measure of general spatial and temporal patterns, it is a reasonable proxy for the ambient population in a location at a particular time, which is then related to the level of crime at that location during that period of time.

A challenge with assessing routine activities theory is not only assessing the ambient population in a location at a point in time, but assessing the number of offenders, targets, and potential guardians that make up that ambient population. Hipp (2016) discussed these challenges along with the additional challenge once collecting measures at a place at a point in time of offenders, targets, and guardians of estimating their possibly nonlinear relationship with crime. Whereas his approach adopted a strategy of estimating their relative presence based on general spatial patterns, we used Twitter data as another way to assess this. The measure of geolocated tweets appears to be a useful alternative to other possible proxies for the number of people in an area, such as the number of total or retail employees. When controlling for the standard socio-demographic measures that are included in ecological studies of crime the number of tweets in an area was strongly predictive of crime events in nearly all time periods of the day. The sizes of these effects were quite substantial, suggesting that despite the imperfections of tweets as representative of the population of people in the area, they nonetheless operate as a reasonable proxy to capture the relative size of the ambient population.

Tweets and crime

Although measuring the number of tweets that typically occur at a place at a point in time does not directly measure the presence of offenders, targets, and guardians, the parametric form of the relationship between the number of tweets and crime does provide some insights. For the violent crimes of aggravated assault and robbery, ambient population likely indicates the presence of all three (offenders, targets, and guardians). Given this, one possibility is that with the presence of a large enough ambient population the effect of more guardians would overwhelm the effects of more targets and offenders: however, this was not the case, as we tested for a quadratic effect and it was not significant. Instead, given the logarithmic relationship we detected between tweets and crime (rather than a simple linear one) it appears that the number of violent events increases at a slowing rate at high levels of ambient population. This implies that the effect of guardians is stronger than that of offenders and targets, but not enough to actually start reducing the probability of crime events. In the case of aggravated assaults, one might presume that the presence of more ambient population later at night might reflect the disproportionate presence of offenders, especially if more alcohol is involved: indeed, the relationship between tweets and aggravated assault was particularly strong between midnight and 4am, which is consistent with this idea.

On the other hand, for the property crimes of burglary and motor vehicle theft the presence of a larger ambient population may increase the number of offenders and guardians, but be less likely to increase targets for these property crimes (particularly if those tweeting are inside their homes, therefore reducing the probability of a burglary at that site). Indeed we detected a weaker relationship between the number of tweets and these two property crimes compared to the violent crimes. The fact that the relationship between tweets and motor vehicle thefts was strongest in the middle of the day may indicate the presence of more targets if the ambient population has driven to a location. Future research combining the presence of the ambient population with the land use characteristics of a location may be better able to test this possibility. Although one might likewise presume that the presence of more ambient population does not increase the number of targets for larceny, the relationship between the ambient population and larcenies was stronger than for the other two property crime types. This may indicate the presence of more offenders: one useful direction for future research is to use information on tweet patterns in an attempt to estimate the age of tweeters, and hence the age pattern of the ambient population in an area in an effort to determine locations with a possibly higher proportion of offenders.

An advantage of using tweets is that they provide time-specific information on the presence of persons in a location. Other proxies of the ambient population often do not contain such temporal specificity. Our results suggest that this social media measure is useful in that it provides unique information on the number of people in an area, and hence the increased crime risk of the area. Furthermore, additional models that also included measures of total and retail employees demonstrated that our measure of temporal ambient population provides additional unique information for crime events during many time periods. These results suggest that this social media measure is useful in that it provides unique information on the number of people in an area, and hence the increased crime risk of the area, beyond that captured by measures of the number of employees in a location. Similar to the findings of Malleson and Andresen (2016) the number of retail employees is a useful proxy that is attempting to capture the number of patrons of stores and hence the number of people in an area. However, in contrast to Malleson and Andresen (2016) the measure of the number of tweets is clearly providing additional unique information beyond employees that is useful, especially when temporal crime clustering is considered.

There are also theoretical insights to be gleaned from observing for which crime types the tweets measure is most reduced when including measures of the number of retail or total employees. For example, the fact that the effect of our ambient population measure was most reduced in the robbery and larceny models implies that the number of employees is a reasonable proxy for ambient population for these crime types for most times of day. It was still the case that the ambient population measure was strongly associated with more robberies overnight even when controlling for employees, which indicates that tweets likely capture the presence of more targets during these hours (and there are simply too few people present to serve as guardians). On the other hand, the effect of our ambient population measure was least impacted in the aggravated assault models. This may indicate that our ambient population measure is capturing the presence of more possible offenders or targets, particularly if it is younger persons who are more likely to tweet.

This study has focused on the geolocation and temporal information provided by tweets and shown that this is beneficial for understanding crime patterns, and this opens further exciting possibilities for other strategies that make use of social media data. For example, we did not use the content of the tweets at all. Some recent research has tried to use the content of tweets to characterize the type of activity at a location (Malleson and Birkin 2014; Wang and Gerber 2015; Williams, Burnap, and Sloan 2016). Analyzing the content of tweets may well extend the possibilities of Twitter information beyond what we have demonstrated here. Ambient population may further be estimated by applying textual analysis of a non geolocated tweet to provide an approximate location. Furthermore, textual analysis of tweets may reveal insights for other criminological theories. For example, the content of the tweets may be useful in characterizing the collective efficacy for informal social control action that exists in a neighborhood, which could be helpful for modifying the models presented here (Sampson, Raudenbush, and Earls 1997; Sampson and Groves 1989). In addition to Twitter, other types of social media platforms may be useful for understand the population at risk, and perceptions of place. The use of check-in data from Foursquare may be another way to capture the spatial and temporal patterns of people (Cranshaw, Schwartz, Hong, and Sadeh 2012). Alternatively, some research has utilized cell phone data as a way to capture the spatial patterns of people (Bogomolov, Lepri, Staiano, Oliver, Pianesi, and Pentland 2014). Future research on ambient populations could seek out data from other applications that utilizes a smartphones GPS, such as Strava and Uber. These various alternative data sources provide exciting options for measuring constructs of interest to criminologists. However, these datasets may be more heavily skewed towards particular users, and it may be difficult to make valid inferences to the population. We recommend social scientists consider exploring combining multiple sources of data (e.g. Twitter, Foursquare, Strava, Uber) to measure ambient population.

We acknowledge some limitations to this study. Despite the fact that tweets are a potentially useful measure of the potential population in a location at a particular time, there are nonetheless validity issues that must be acknowledged. First, of their approximately 974 million Twitter users, only 23% tweeted in the last 30 days (Koh 2014). Twitter does not provide information from a large proportion of the population. Second, the population that actually does tweet differs from the general population. People who tweet tend to be younger, more educated, and urban compared to the general population (Hecht and Stephens 2014; Shelton, Poorthuis, and Zook 2015; Sloan, Morgan, Burnap, and Williams 2015). Although this bias is not ideal, it is worth highlighting that younger persons are disproportionately likely to be offenders or targets—in the NIBRS data for 2013, whereas 20.5 percent of the general population in the represented cities was between 14 and 29 years of age, 69.1 percent of robbery offenders and 46.8 percent of victims were in this age range (the values were 46 and 41.9 percent for aggravated assault offenders and victims, respectively) (authors' calculations). Furthermore, the bias towards highly educated persons—27% of online users with a bachelor's degree use Twitter whereas 19% of those with less than a high school degree do so[6]—provides a downward bias to the estimate of the relationship between tweets and crime given that crime tends to concentrate in lower income (and hence lower education) neighborhoods. Third, there is variability in the frequency that different people tweet. Although the ID variable allows accounting for multiple tweets from a single person at a specific location, it is still the case that the measure will provide more *accurate* information for people who tweet frequently (given that we therefore frequently know where they are located), but relatively poorer information for people who tweet less often (since we do not know their location at many points in time). Fourth, people may not be equally likely to tweet during all activities. Thus, if people are disproportionately likely to tweet during certain types of activities (e.g., at a music concert) compared to when they are engaged in other types of activities, our measure will be upwardly biased in locations that contain venues with these certain types of activities. We are not currently aware of systematic research that would provide information on the types of locations at which tweeting is relatively more likely. Fifth, only about 4.2% percent of tweeters provide their geographic reference location (Leetaru et al. 2013). If there is a systematic difference in the types of people who provide geolocation information compared to those who do not, this will be a potential source of bias.

[6] http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/

We also acknowledge that tweets act as a proxy for the ambient population but fail to account for the disaggregate population of guardians, offenders, and victims. We currently did not examine if the tweeting population in the blocks are residents or visitors. Although the majority of crime victimization does happen close to the victim's residence, a significant proportion occurs away from the victim's residence (Tita and Griffiths 2005). Felson and Boivin's (2015) analysis of retrospective travel surveys demonstrated that crime concentration at place is generated by large influx of non-resident visitors. A follow up study could infer home locations of Twitter users and then distinguish between residential and visitor populations. Additionally, Twitter data may be combined with demographic data to determine resident location. For example, Barbera (2016) used a novel technique to match Twitter user name to data from voter rolls to obtain a residential home location. It should be noted that we currently only examined tweets for an eight month period; a longer period would increase statistical precision of the estimates of the ambient population. Lastly, there is temporal uncertainty in the reporting of crime, particularly property crimes, which could also explain the typically weaker relationship between tweets and property crimes.

Despite these limitations, we believe this technique utilizing a source that some might refer to as "big data" shows promise for the field of criminology. Although there are cautions about blindly using "big data" (Lazer, Kennedy, King, and Vespignani 2014), we showed that the general pattern of tweets at a small geographic unit of a block over particular hours of the day and the days of the week may provide a relatively robust predictor of various types of crime at such small units during those time periods. Geographic referenced social data including geocoded tweets allows researchers to more precisely measure the presence of people in the environment at a particular time point. Furthermore, measuring ambient population with Twitter data is cost-effective approach to cover a large area (i.e., Southern California) allowing for a more appropriate test of the routine activities and crime pattern theories that emphasize the spatial and temporal confluence of offenders and targets.

References

Andresen, Martin A. 2006. "A spatial analysis of crime in Vancouver, British Columbia: a synthesis of social disorganization and routine activity theory." *The Canadian Geographer/Le Géographe canadien* 50:487-502.

Andresen, Martin A and Greg W Jenion. 2010. "Ambient populations and the calculation of crime rates and risk." *Security Journal* 23:114-133.

Andresen, Martin A. 2011. "The Ambient Population and Crime Analysis." *The Professional Geographer* 63:193-212.

Andresen, Martin A. and Nick Malleson. 2015. "Intra-week spatial-temporal patterns of crime." *Crime Science* 4:1-11.

Anselin, Luc and Sarah Williams. 2015. "Digital neighborhoods." *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*:1-24.

Ashby, Matthew PJ and Kate J Bowers. 2013. "A comparison of methods for temporal analysis of aoristic crime." *Crime Science* 2:1-16.

Barberá, Pablo. 2016. "Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data.".

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

Bernasco, Wim and Richard L. Block. 2011. "Robberies in Chicago: A Block-Level Analysis of the Influence of Crime Generators, Crime Attractors, and Offender Anchor Points." *Journal of Research in Crime and Delinquency* 48:33-57.

Birkin, M, K Harland, and N Malleson. 2013. "The classification of space-time behaviour patterns in a British city from crowd-sourced data." *Computational Science and Its …*.

Boessen, Adam. 2014. *Geographic Space and Time: The Consequences of the Spatial Footprint for Neighborhood Crime*. Irvine, CA: Unpublished dissertation.

Boggs, Sarah L. 1965. "Urban Crime Patterns." *American Sociological Review* 30:899-908.

Bogomolov, Andrey, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. 2014. "Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data." *arXiv.org*.

Boivin, Rémi. 2013. "On the Use of Crime Rates 1." *Canadian Journal of Criminology and Criminal Justice* 55:263-277.

Brantingham, P. J. and P. L. Brantingham. 2008. "Crime pattern theory." Pp. 102-118 in *Environmental Criminology and Crime Analysis,* edited by R. Wortley and L. Mazerolle. New York, NY: Routledge.

Ceccato, Vania and Adriaan Cornelis Uittenbogaard. 2014. "Space--time dynamics of crime in transport nodes." *Annals of the Association of American Geographers* 104:131-150.

Cohen, Lawrence E and Marcus Felson. 1979. "Social change and crime rate trends: A routine activity approach." *American sociological review*:588-608.

Cohen, Michael Lee and Xiao Di Zhang. 1988. "The Difficulty of Improving Statistical Synthetic Estimation." Bureau of the Census, Washington, D.C.

Cranshaw, Justin, Raz Schwartz, Jason I Hong, and Norman Sadeh. 2012. "The livehoods project: Utilizing social media to understand the dynamics of a city." Pp. 58 in *International AAAI Conference on Weblogs and Social Media*.

Driscoll, Kevin and Shawn Walker. 2014. "Big data, big questions| working within a black box: Transparency in the collection and production of big twitter data." *International Journal of Communication* 8:20.

Felson, Marcus and Rachel L Boba. 2010. *Crime and everyday life*: Sage.

Felson, Marcus and Rémi Boivin. 2015. "Daily crime flows within a city." *Crime Science* 4:1-10.

Felson, Marcus and Erika Poulsen. 2003. "Simple indicators of crime by time of day." *International Journal of Forecasting* 19:595-601.

Frias-Martinez, Vanessa and Enrique Frias-Martinez. 2014. "Spectral clustering for sensing urban land use using Twitter activity." *Engineering Applications of Artificial Intelligence* 35:237-245.

Gibbs, Jack P. and Walter T. Martin. 1962. "Urbanization, Technology, and the Division of Labor: International Patterns." *American Sociological Review* 27:667-677.

Gove, Walter R, Michael Hughes, and Omer R Galle. 1979. "Overcrowding in the home: An empirical investigation of its possible pathological consequences." *American sociological review*:59-80.

Groff, Elizabeth R. 2008. "Adding the temporal and spatial aspects of routine activities: A further test of routine activity theory." *Security Journal* 21:95-116.

Haberman, Cory P. and Jerry H. Ratcliffe. 2015. "Testing for Temporally Differentiated Relationships Among Potentially Criminogenic Places and Census Block Street Robbery Counts." *Criminology* 53:457-483.

Harries, Keith. 2006. "Property crimes and violence in United States: an analysis of the influence of population density." *International Journal of Criminal Justice Sciences* 1:24-34.

Hecht, Brent and Monica Stephens. 2014. "A Tale of Cities: Urban Biases in Volunteered Geographic Information." *ICWSM* 14:197-205.

Hipp, John R. 2007. "Income Inequality, Race, and Place: Does the Distribution of Race and Class within Neighborhoods Affect Crime Rates?" *Criminology* 45:665-697.

—. 2016. "General theory of spatial crime patterns." *Criminology* 54:653-679.

Hipp, John R., Daniel J. Bauer, Patrick J. Curran, and Kenneth A. Bollen. 2004. "Crimes of Opportunity or Crimes of Emotion: Testing Two Explanations of Seasonal Change in Crime." *Social Forces* 82:1333-1372.

Koh, Yoree. 2014. "Only 11% of New Twitter Users in 2012 Are Still Tweeting." in *The Wall Street Journal*.

Kounadi, O., T. J. Lampoltshammer, E. Groff, I. Sitko, and M. Leitner. 2015. "Exploring Twitter to analyze the public's reaction patterns to recently reported homicides in London." *PLoS One* 10:e0121848.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343:1203-1205.

Lee, J H, A W Davis, and K G Goulias. 2016. "Activity Space Estimation with Longitudinal Observations of Social Media Data." *... for presentation at the 95th Annual ...*.

Leetaru, Kalev, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. 2013. "Mapping the global Twitter heartbeat: The geography of Twitter." *First Monday* 18.

Lenormand, M., M. Picornell, O. G. Cantu-Ros, A. Tugores, T. Louail, R. Herranz, M. Barthelemy, E. Frias-Martinez, and J. J. Ramasco. 2014. "Cross-checking different sources of mobility information." *PLoS One* 9:e105184.

Malleson, N and M Birkin. 2014. "New Insights into Individual Activity Spaces using Crowd-Sourced Big Data."

Malleson, Nick and Martin A Andresen. 2015a. "The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns." *Cartography and Geographic Information Science* 42:112-121.

—. 2016. "Exploring the impact of ambient population measures on London crime hotspots." *Journal of Criminal Justice* 46:52-63.

Malleson, Nick and Martin A. Andresen. 2015b. "Spatio-temporal crime hotspots and the ambient population." *Crime Science* 4.

McCord, Eric S and Jerry H Ratcliffe. 2009. "Intensity value analysis and the criminogenic effects of land use features on local crime patterns." *Crime Patterns and Analysis* 2:17-30.

Raleigh, Erica and George Galster. 2015. "Neighborhood Disinvestment, Abandonment, and Crime Dynamics." *Journal of Urban Affairs* 37:367-396.

Sampson, R. J., S. W. Raudenbush, and F. Earls. 1997. "Neighborhoods and violent crime: a multilevel study of collective efficacy." *Science* 277:918-24.

Sampson, Robert J and W Byron Groves. 1989. "Community structure and crime: Testing social-disorganization theory." *American journal of sociology*:774-802.

Shelton, Taylor, Ate Poorthuis, and Matthew Zook. 2015. "Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information." *Landscape and Urban Planning* 142:198-211.

Sherman, Lawrence W. 1995. "Hot spots of crime and criminal careers of places." *Crime and place* 4:35-52.

Sherman, Lawrence W, Patrick R Gartin, and Michael E Buerger. 1989. "Hot spots of predatory crime: Routine activities and the criminology of place*." *Criminology* 27:27-56.

Sloan, L., J. Morgan, P. Burnap, and M. Williams. 2015. "Who tweets? Deriving the demographic characteristics of age, occupation and social class from twitter user meta-data." *PLoS One* 10:e0115545.

Sorg, Evan T. and Ralph B. Taylor. 2011. "Community-level impacts of temperature on urban street robbery." *Journal of Criminal Justice* Forthcoming.

Steiger, Enrico, João Porto de Albuquerque, and Alexander Zipf. 2015. "An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data." *Transactions in GIS* 19:809-834.

Steinberg, Joseph. 1979. "Synthetic Estimates for Small Areas: Statistical Workshop Papers and Discussion." Pp. 282 in *National Institute on Drug Abuse Research Monograph Series*, vol. 24. Washington, D.C.: National Institute on Drug Abuse.

Stults, Brian J. and Matthew Hasbrouck. 2015. "The Effect of Commuting on City-Level Crime Rates." *Journal of Quantitative Criminology* 31:331-350.

Tita, George and Elizabeth Griffiths. 2005. "Traveling to violence: The case for a mobility-based spatial typology of homicide." *Journal of Research in Crime and Delinquency* 42:275-308.

Tompson, Lisa and Michael Townsley. 2010. "(Looking) Back to the Future: using space–time patterns to better predict the location of street crime." *International Journal of Police Science & Management* 12:23-40.

Wakamiya, Shoko, Ryong Lee, and Kazutoshi Sumiya. 2013. "LNCS 8238 - Social-Urban Neighborhood Search Based on Crowd Footprints Network."1-14.

Wang, Mingjun and Matthew S Gerber. 2015. "Using Twitter for Next-Place Prediction, with an Application to Crime Prediction." Pp. 941-948 in *Computational Intelligence, 2015 IEEE Symposium Series on*: IEEE.

Wang, Xiaofeng, Donald E Brown, and Matthew S Gerber. 2012. *Spatio-temporal modeling of criminal incidents using geographic, demographic, and twitter-derived information*: IEEE.

Weisburd, David. 2015. "The Law of Crime Concentration and the Criminology of Place*." *Criminology* 53:133-157.

Williams, Matthew L, Pete Burnap, and Luke Sloan. 2016. "Crime Sensing with Big Data: The Affordances and Limitations of using Open Source Communications to Estimate Crime Patterns." *British Journal of Criminology*:azw031.

Wo, James C., John R. Hipp, and Adam Boessen. 2016. "Voluntary Organizations and Neighborhood Crime: A Dynamic Perspective*." *Criminology* 54:212-241.

Tables and Figures

Tweets and crime

Table 1. Logistic regression results showing relationship between number of tweets during a 2 hour period and crime events. Models control for socio-demographic measures. Each time period represents a separate model. Using crime data from May to December to match twitter data.

	Midnight-2am	2-4am	4-6am	6-8am	8-10am	10-noon	Noon-2pm	2-4pm	4-6pm	6-8pm	8-10pm	10-midnight
Aggravated assault - weekdays												
Tweets (logged)	0.9067 **	0.9243 **	-0.0925	0.6164 **	0.5879 **	0.6807 **	0.6917 **	0.4997 **	0.4058 **	0.5008 **	0.4518 **	0.6116 **
	(9.98)	(6.87)	-(0.16)	(4.66)	(6.15)	(9.73)	(10.79)	(6.51)	(5.28)	(6.09)	(6.14)	(5.82)
Aggravated assault - weekends												
Tweets (logged)	1.0139 **	1.0735 **	0.3476	0.6007 *	0.2677	0.7224 **	0.6225 **	0.5583 **	0.6303 **	0.616 **	0.6842 **	0.7628 **
	(13.69)	(9.17)	(1.25)	(2.36)	(1.43)	(7.04)	(6.30)	(6.10)	(7.25)	(9.25)	(10.14)	(9.62)
Robbery - weekdays												
Tweets (logged)	0.6011 **	1.1412 **	0.5541	0.7936 **	0.7456 **	0.8391 **	0.7582 **	0.7025 **	0.7753 **	0.7508 **	0.5101 **	0.605 **
	(3.94)	(8.44)	(1.23)	(4.87)	(6.13)	(11.32)	(8.89)	(9.51)	(10.40)	(9.61)	(5.55)	(5.53)
Robbery - weekends												
Tweets (logged)	1.0089 **	0.6163 *	1.1719 **	0.6682 *	0.3634	0.7218 **	0.7993 **	0.8599 **	0.9179 **	0.6414 **	0.5858 **	0.6656 **
	(10.24)	(2.38)	(4.81)	(2.42)	(0.83)	(4.47)	(6.55)	(7.65)	(9.12)	(6.67)	(6.11)	(6.50)
Burglary - weekdays												
Tweets (logged)	0.3666 **	0.3952 **	0.7566 **	0.2564 †	0.3567 **	0.3584 **	0.4853 **	0.5739 **	0.5879 **	0.6173 **	0.4939 **	0.4571 **
	(4.30)	(3.29)	(5.67)	(1.83)	(5.44)	(6.51)	(9.70)	(10.67)	(11.49)	(11.19)	(7.30)	(6.13)
Burglary - weekends												
Tweets (logged)	0.6167 **	0.2634	0.3537	0.3901	0.2345 †	0.6439 **	0.5898 **	0.4751 **	0.6425 **	0.5675 **	0.4069 **	0.4566 **
	(6.80)	(1.51)	(1.32)	(1.59)	(1.65)	(8.62)	(8.15)	(6.56)	(8.77)	(9.94)	(5.49)	(6.14)
Motor vehicle theft - weekdays												
Tweets (logged)	0.254 *	0.3777 **	0.0175	0.2263 †	0.4879 **	0.7209 **	0.7205 **	0.6335 **	0.5996 **	0.3498 **	0.2206 **	0.333 **
	(2.56)	(2.65)	(0.05)	(1.67)	(7.12)	(12.64)	(12.56)	(10.46)	(8.99)	(4.68)	(2.67)	(4.25)
Motor vehicle theft - weekends												
Tweets (logged)	0.4232 **	0.0847	0.6712 **	0.3525 †	0.2897 †	0.6563 **	0.778 **	0.7959 **	0.804 **	0.6091 **	0.3334 **	0.5516 **
	(3.48)	(0.39)	(3.87)	(1.78)	(1.91)	(8.15)	(11.54)	(11.09)	(12.04)	(9.21)	(3.58)	(6.78)
Larceny - weekdays												
Tweets (logged)	0.5044 **	0.5233 **	0.5294 **	0.481 **	0.5793 **	0.7841 **	0.833 **	0.7899 **	0.8454 **	0.7242 **	0.5589 **	0.5393 **
	(7.53)	(6.04)	(4.24)	(6.25)	(12.40)	(19.45)	(23.12)	(19.60)	(22.05)	(17.71)	(12.02)	(10.31)
Larceny - weekends												
Tweets (logged)	0.6316 **	0.7714 **	0.3078	0.6388 **	0.4609 **	0.9133 **	0.79 **	0.8859 **	0.8533 **	0.7688 **	0.5975 **	0.6613 **
	(7.81)	(8.05)	(1.64)	(6.15)	(6.09)	(19.11)	(17.30)	(20.71)	(19.32)	(17.65)	(12.60)	(11.28)

Note: ** p < .01(two-tail test), * p < .05 (two-tail test), † p < .05 (one-tail test). Each cell represents an estimated coefficient from a single model, with the T-value in parentheses below the coefficient. Each model includes the following variables at the block level: concentrated disadvantage, residential stability, racial/ethnic heterogeneity, percent black, percent Latino percent Asian, percent occupied unit, percent aged 16 to 29, percent vacant lots. Each model includes the following 1/2 mile inverse distance decay spatial buffer variables: concentrated disadvantage, residential stability, racial/ethnic heterogeneity, percent black, percent Latino percent Asian, percent occupied unit, percent aged 16 to 29, population, percent vacant lots.

Tweets and crime

Table 2. Logistic regression results showing relationship between number of tweets during a 2 hour period and crime events, controlling for number of total and retail employees in area. Models control for socio-demographic measures, as well as total and retail employees (logged). Using crime data from May to December to match twitter data.

	Midnight-2am	2-4am	4-6am	6-8am	8-10am	10-noon	Noon-2pm	2-4pm	4-6pm	6-8pm	8-10pm	10-midnight
Aggravated assault - weekdays												
Tweets (logged)	0.7509 **	0.7176 **	-0.2571	0.332 *	0.2927 **	0.3741 **	0.3261 **	0.1158	0.0796	0.2784 **	0.246 **	0.4849 **
	(8.96)	(5.10)	-(0.42)	(2.26)	(2.84)	(4.87)	(4.37)	(1.36)	(0.96)	(3.15)	(3.14)	(4.59)
Aggravated assault - weekends												
Tweets (logged)	0.7623 **	0.8578 **	0.1552	0.39	-0.0173	0.5066 **	0.4027 **	0.322 **	0.3629 **	0.4217 **	0.5058 **	0.5763 **
	(10.29)	(7.47)	(0.50)	(1.43)	-(0.09)	(4.64)	(3.71)	(3.18)	(3.76)	(6.16)	(7.42)	(7.22)
Robbery - weekdays												
Tweets (logged)	0.2984	0.7919 **	0.0986	0.3883 *	0.2089	0.2776 **	0.1984 †	0.1748 *	0.2323 **	0.2664 **	0.0965	0.2952 *
	(1.63)	(5.40)	(0.18)	(2.02)	(1.46)	(3.00)	(1.93)	(1.99)	(2.61)	(2.90)	(0.92)	(2.39)
Robbery - weekends												
Tweets (logged)	0.7213 **	0.2769	0.8951 **	0.1697	-0.3511	0.1733	0.2719 †	0.3499 **	0.3712 **	0.0637	0.1551	0.3202 **
	(6.73)	(0.95)	(3.78)	(0.48)	-(0.65)	(0.87)	(1.85)	(2.58)	(2.93)	(0.52)	(1.35)	(2.77)
Burglary - weekdays												
Tweets (logged)	0.0552	0.0046	0.3592 *	0.0152	0.1089	0.1194 *	0.1978 **	0.2196 **	0.1819 **	0.2632 **	0.1584 *	0.1741 *
	(0.58)	(0.03)	(2.32)	(0.10)	(1.52)	(1.98)	(3.81)	(3.79)	(3.11)	(4.37)	(2.24)	(2.09)
Burglary - weekends												
Tweets (logged)	0.3401 **	-0.164	-0.1533	-0.065	-0.0749	0.2891 **	0.2271 **	0.12	0.3135 **	0.2259 **	0.1049	0.2013 **
	(3.49)	-(0.81)	-(0.44)	-(0.22)	-(0.49)	(3.51)	(2.76)	(1.48)	(3.98)	(3.46)	(1.31)	(2.58)
Motor vehicle theft - weekdays												
Tweets (logged)	0.1107	0.2493	-0.1347	-0.073	0.1528 *	0.3489 **	0.2864 **	0.2114 **	0.217 **	-0.009	-0.06	0.1292
	(1.06)	(1.62)	-(0.40)	-(0.47)	(2.02)	(5.57)	(4.41)	(3.10)	(3.01)	-(0.11)	-(0.70)	(1.54)
Motor vehicle theft - weekends												
Tweets (logged)	0.2095 †	-0.094	0.5594 **	0.1368	-0.082	0.3346 **	0.3648 **	0.4472 **	0.4425 **	0.2634 **	0.042	0.3154 **
	(1.67)	-(0.41)	(3.19)	(0.62)	-(0.49)	(4.01)	(4.89)	(5.83)	(6.08)	(3.77)	(0.44)	(3.71)
Larceny - weekdays												
Tweets (logged)	0.2965 **	0.315 **	0.2395 †	0.0506	0.0531	0.2151 **	0.2544 **	0.2021 **	0.283 **	0.2074 **	0.1329 **	0.2199 **
	(4.31)	(3.45)	(1.78)	(0.57)	(0.97)	(5.08)	(6.59)	(4.68)	(6.94)	(4.60)	(2.73)	(4.02)
Larceny - weekends												
Tweets (logged)	0.3214 **	0.5658 **	0.0615	0.3435 **	-0.0475	0.4276 **	0.2658 **	0.3681 **	0.3402 **	0.2596 **	0.1743 **	0.3279 **
	(3.69)	(5.75)	(0.29)	(3.06)	-(0.54)	(7.88)	(5.55)	(8.12)	(6.78)	(5.27)	(3.37)	(5.17)

*Note: ** p < .01(two-tail test), * p < .05 (two-tail test), † p < .05 (one-tail test). Each cell represents an estimated coefficient from a single model, with the T-value in parentheses below the coefficient. Each model includes the following variables at the block level: concentrated disadvantage, residential stability, racial/ethnic heterogeneity, percent black, percent Latino percent Asian, percent occupied unit, percent aged 16 to 29, percent vacant lots, total employees (logged), retail employees (logged). Each model includes the following 1/2 mile inverse distance decay spatial buffer variables: concentrated disadvantage, residential stability, racial/ethnic heterogeneity, percent black, percent Latino percent Asian, percent occupied unit, percent aged 16 to 29, population, percent vacant lots, total employees (logged), retail employees (logged).*

Figure 1. Average number of tweets per 2-hour time period in a block

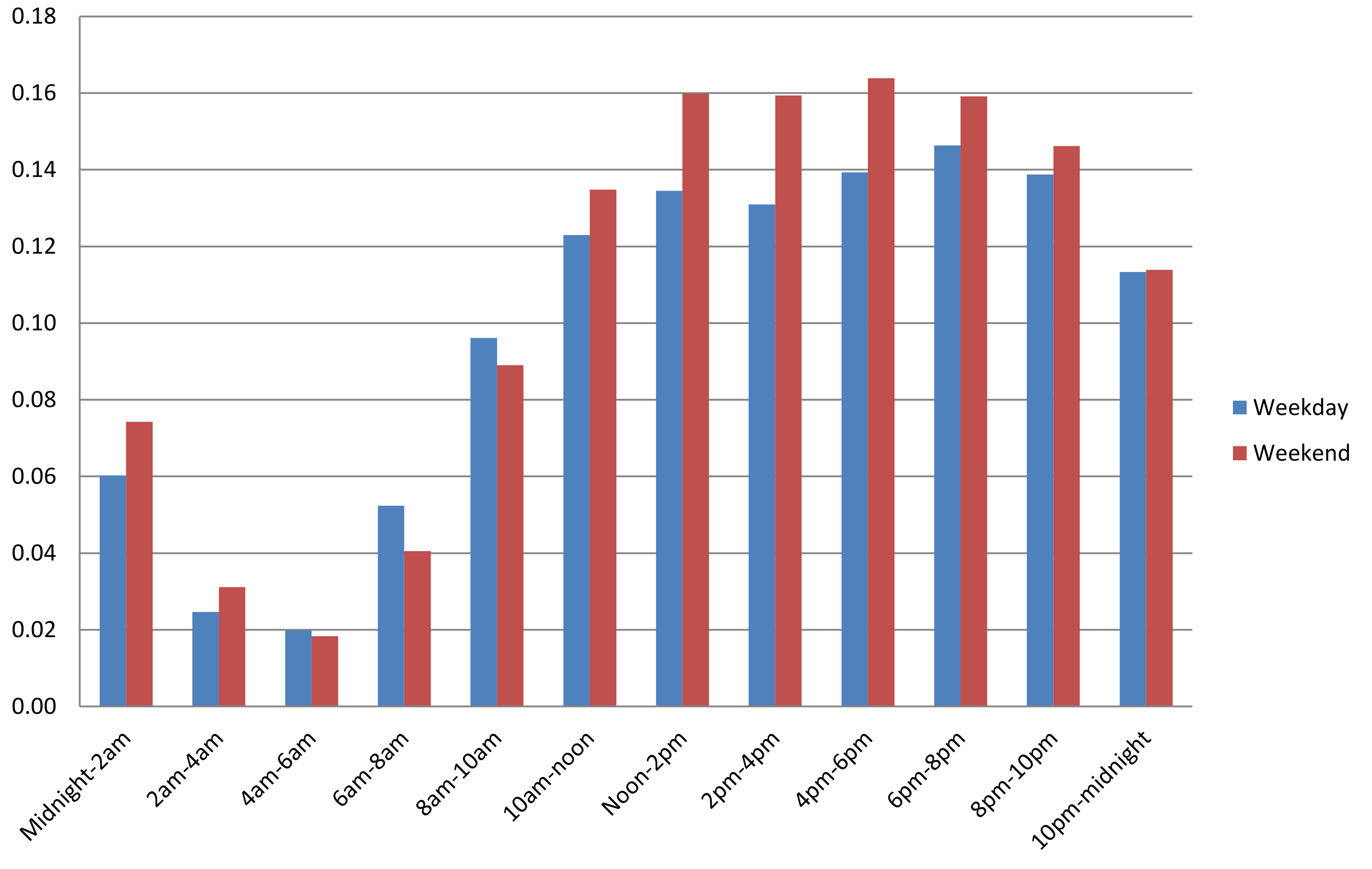


Figure 2a. Coefficient estimates: Aggravated assault - weekdays

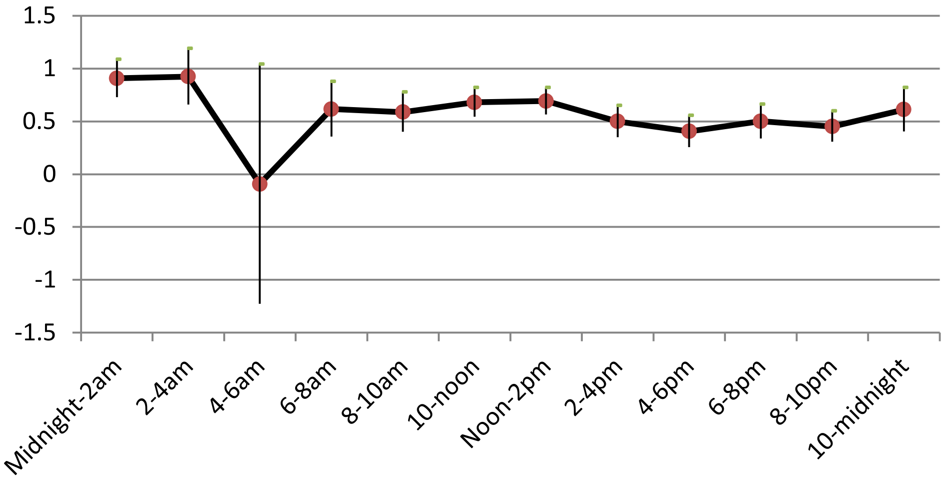


Figure 2b. Coefficient estimates: Aggravated assault - weekends

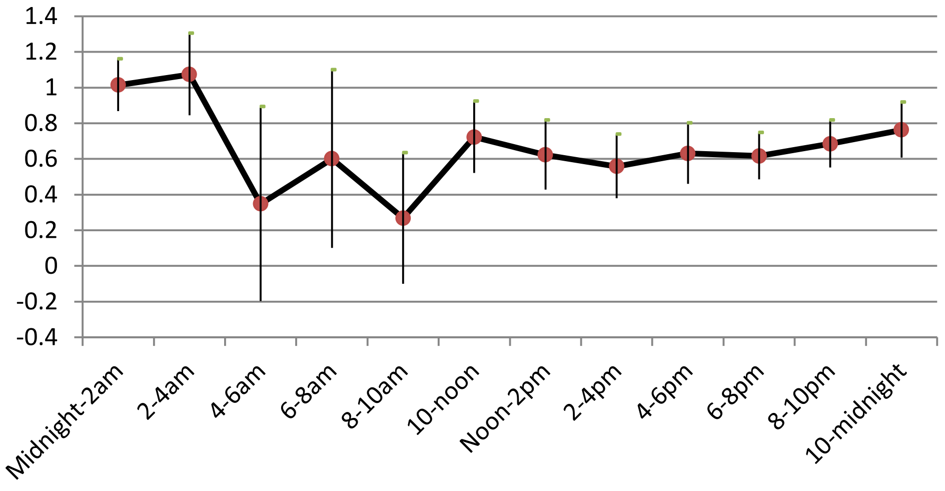


Figure 3a. Coefficient estimates: Robbery - weekdays

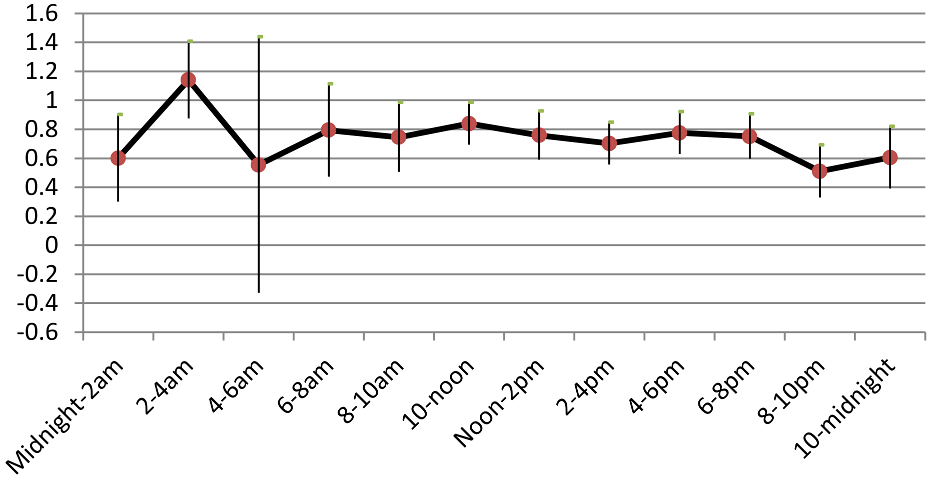


Figure 3b. Coefficient estimates: Robbery - weekends

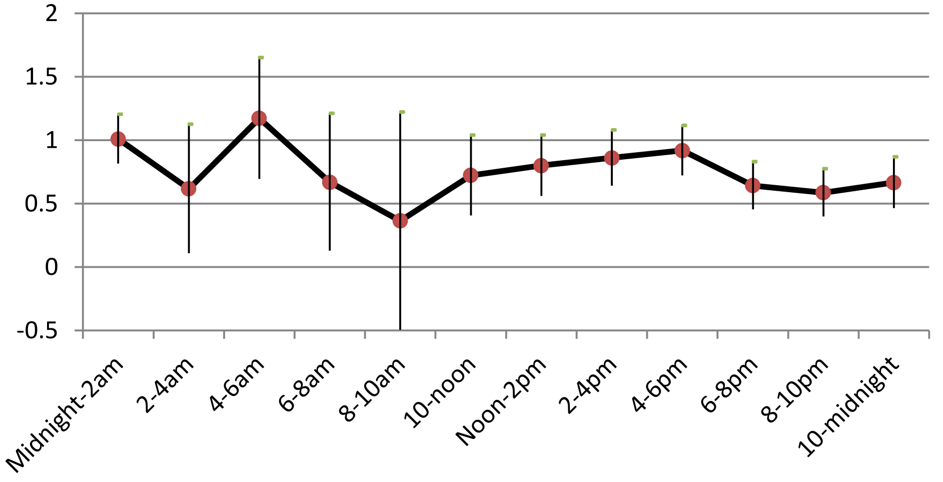


Figure 4a. Coefficient estimates: Burglary - weekdays

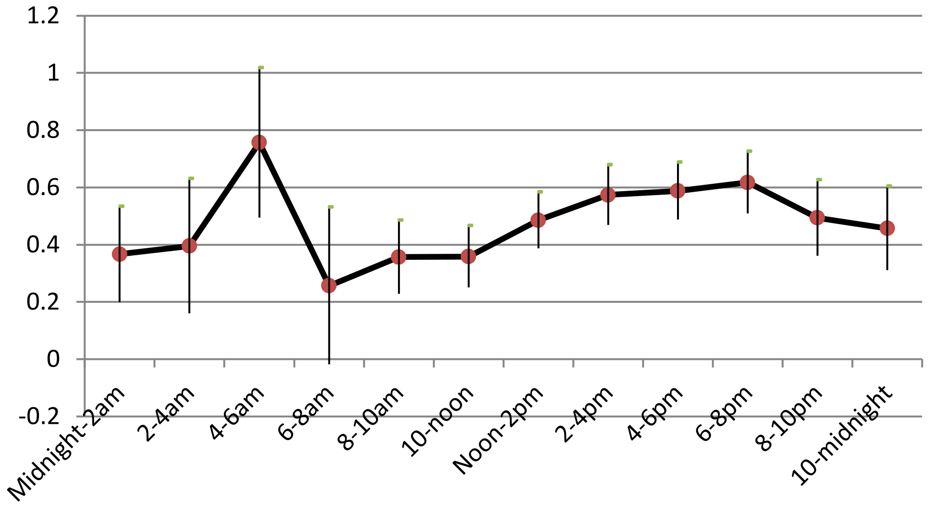


Figure 4b. Coefficient estimates: Burglary - weekends

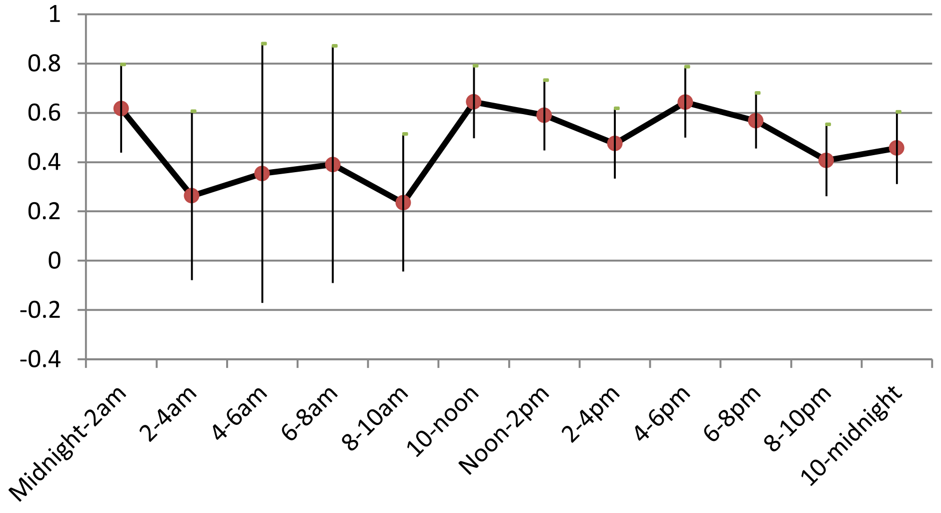


Figure 5a. Coefficient estimates: Motor vehicle theft - weekdays

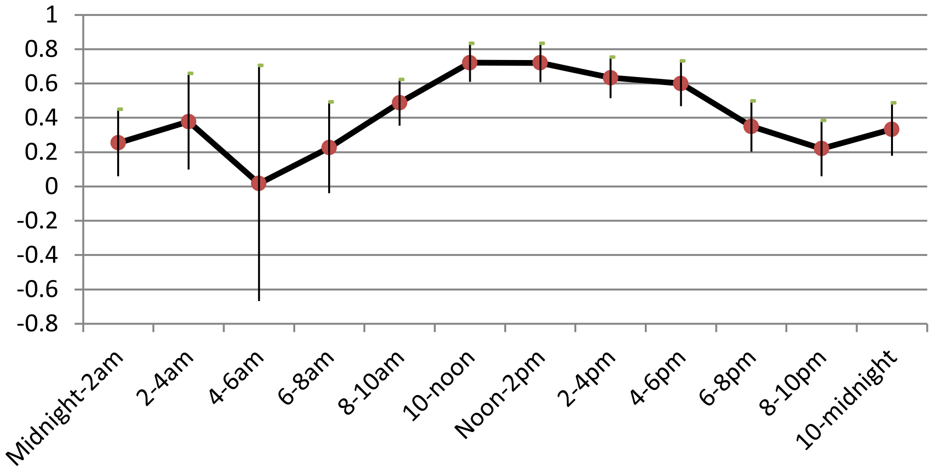


Figure 5b. Coefficient estimates: Motor vehicle theft - weekends

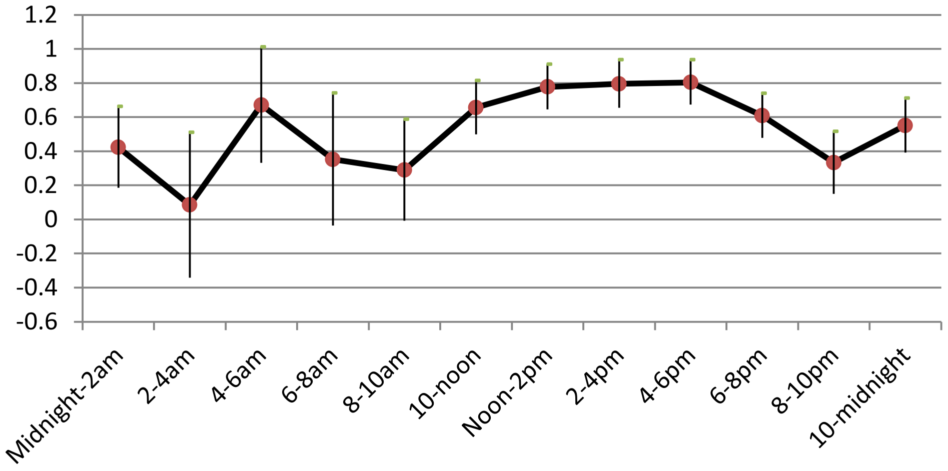


Figure 6a. Coefficient estimates: Larceny - weekdays

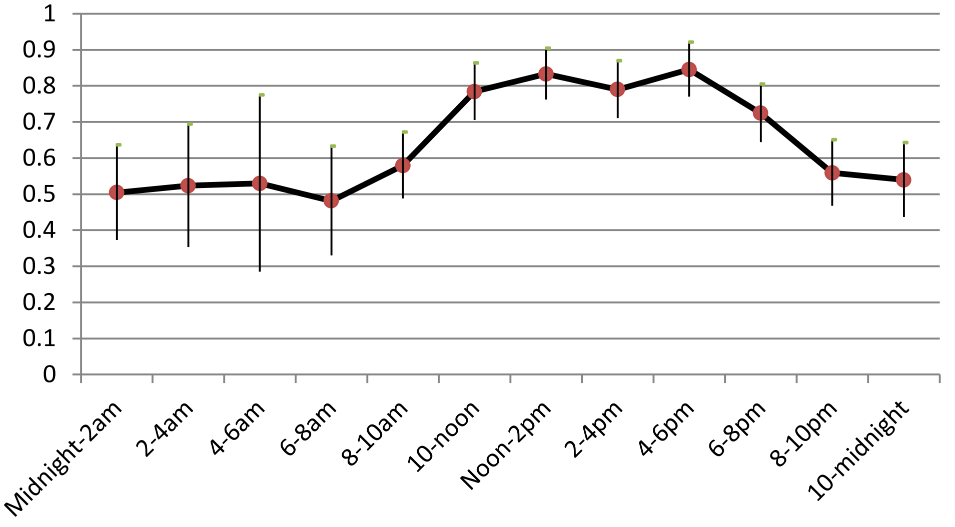
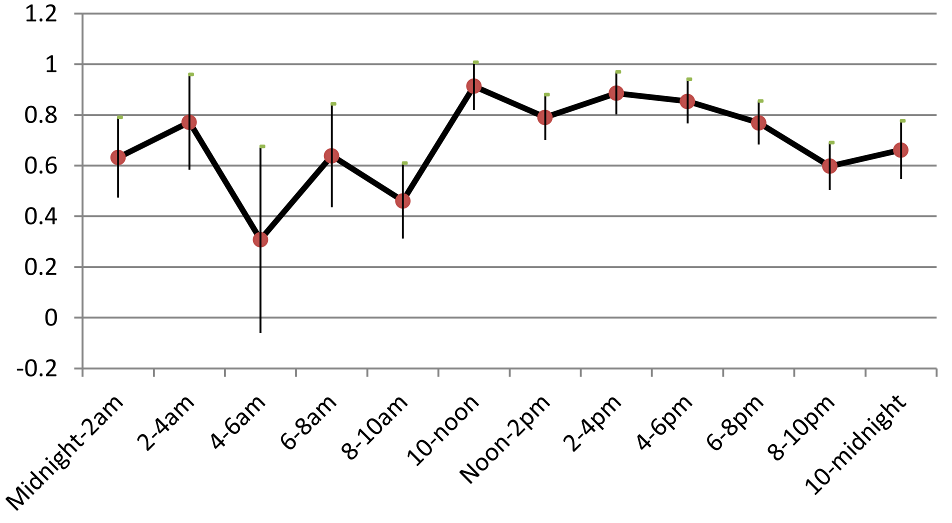


Figure 6b. Coefficient estimates: Larceny - weekends


Midnight-2am
2-4am
4-6am
6-8am
8-10am
10-noon
Noon-2pm
2-4pm
4-6pm
6-8pm
8-10pm
10-midnight