UCL DEPARTMENT OF COMPUTER SCIENCE

Research Note
RN/14/08

Mining Mobile Phone Data to Evaluate Urban Crime Theories at Scale

17.06.2014

Martin Traunmueller

Giovanni Quattrone

Licia Capra

Abstract

Prior work in architectural and urban studies suggests that there is a strong correlation between people dynamics and crime activities in an urban environment. These studies have been conducted primarily using qualitative evaluation methods, and as such are limited in terms of the geographic area they cover, the number of respondents they reach out to, and the temporal frequency with which they can be repeated. As cities are rapidly growing and evolving complex entities, complementary approaches that afford social scientists the ability to evaluate urban crime theories at scale are required. In this paper, we propose a new method whereby we mine telecommunication data and open crime data to quantitatively validate these theories. More precisely, we analyse footfall counts as recorded by telecommunication data, and extract metrics that act as proxies of urban crime theories. Using correlation analysis between such proxies and crime activity derived from open crime data records, we can reveal to what extent different theories of urban crime hold, and where. We apply this approach to the metropolitan area of London, UK and find significant correlations between crime and metrics derived from theories by Jacobs [11] (e.g., population diver- sity in terms of age) and by Felson and Clarke [7] (e.g., ratio of females and of young people). We conclude the paper with a discussion of the implications of this work on social science research practices.

Mining Mobile Phone Data to Evaluate Urban Crime Theories at Scale

ABSTRACT

Prior work in architectural and urban studies suggests that there is a strong correlation between people dynamics and crime activities in an urban environment. These studies have been conducted primarily using qualitative evaluation methods, and as such are limited in terms of the geographic area they cover, the number of respondents they reach out to, and the temporal frequency with which they can be repeated. As cities are rapidly growing and evolving complex entities, complementary approaches that afford social scientists the ability to evaluate urban crime theories at scale are required. In this paper, we propose a new method whereby we mine telecommunication data and open crime data to quantitatively validate these theories. More precisely, we analyse footfall counts as recorded by telecommunication data, and extract metrics that act as proxies of urban crime theories. Using correlation analysis between such proxies and crime activity derived from open crime data records, we can reveal to what extent different theories of urban crime hold, and where. We apply this approach to the metropolitan area of London, UK and find significant correlations between crime and metrics derived from theories by Jacobs [11] (e.g., population diversity in terms of age) and by Felson and Clarke [7] (e.g., ratio of females and of young people). We conclude the paper with a discussion of the implications of this work on social science research practices.

Author Keywords

Urban crime, telecommunication data, open data, data mining

ACM Classification Keywords

H.2.8 Database Management: Database Applications—*data mining*; K.4.1 Computers and Society: Human safety

INTRODUCTION

In modern society we are experiencing two phenomena: on one hand, there is a rapid population shift of people moving from rural areas into urban environments with an annual growth of 60 million new city dwellers every year [24]. On the other hand, we experience, as in the UK for example, a steady rise in crime activities over the last couple of years [3], focusing especially on densely populated areas [12]. Being able to understand and quantify the relationship between people's presence and crime activity in an area has thus become an important concern, for both citizens, planners and administrators.

The relationship between *people dynamics* and *crime* in urban environments has been researched extensively in architectural and urban studies over the last decades, with theories that sometimes appear to conflict with each other. Most influential theories lead back to the 1960's and 1970's: Jacobs [11] suggests that population diversity, activity and a high mix of functions lead to less crime for an area, whereas Newman [14] hypothesizes the opposite, supporting clear separation of public, semi-public and private areas towards urban safety. Each theory has been evaluated, and indeed supported, by means of qualitative research methods that enable in-depth investigations into the reasons behind certain phenomena. However, such methods are very expensive and time-consuming to run, so that studies are usually restricted to a rather small number of people (relative to the overall urban population) and constrained geographic areas (e.g., a neighbourhood); furthermore, they are almost never repeated over time, to observe potential changes. It becomes thus very difficult to collect sufficient evidence to explain under what conditions a certain theory holds.

To address this issue, in this paper we propose a new method to quantitatively evaluate urban crime theories at scale. The method leverages recent trends that have seen the public release of crime data records, as well as anonymised mobile telecommunication data. From the former, we extract quantitative information about what type of crime activity happens across different urban areas of very fine spatial granularity. From the latter, we extract metrics that act as proxies for previously developed urban crime theories that link people's presence with crime. We can do so as mobile telecommunication data provides a demographic breakdown (by age, gender and type – residents, workers or visitors) of how many people are present in a given area at a given time. As the penetration of mobile phones in cities of developed countries is very high, and as mobile phones are personal devices usually carried by people all the time, we expect such data (and the derived metrics) to offer a rather accurate and fine-grained image of the urban area under exam. We then use correlation analysis between crime activity and our defined metrics to validate urban crime theories at scale. We apply this method to crime and telecommunication data obtained for the city of London, UK. Our findings support the validity (in London) of Jacobs' theory of 'natural surveillance' [11]: we discover that age diversity, as well as the ratio of visitors in a given area, are significant and negatively correlated with crime activities; Felson and Clarke theory [7] that links a higher presence of young people with higher crime is also confirmed; however,

their association between a higher presence of women with lower crime rates is not supported. We believe the proposed method to be a powerful tool in the hands of social science researchers developing urban crime theories, as they can now complement qualitative investigations with quantitative ones: while the former affords them deep insights into the causality of certain phenomena, the latter affords them the ability to scale up findings in terms of population reach, geographical spread, and temporal evolution.

The remainder of the paper is structured as follows: we first provide a brief overview on background theories from architectural and criminological studies, and state-of-the-art follow-up research that has been grounded on them. We then present our method, in terms of the datasets we leverage, the pre-processing and data manipulation we have conducted, and the metrics we have extracted as proxies for urban crime theories. We discuss the results obtained when applying our method to data for the city of London, UK, and finally conclude the paper by discussing implications, limitations and future steps.

RELATED WORK

Background

Most well known architectural theories about the relationship between people dynamics, the urban environment and crime lead back to the studies of [11] and [14]. Following the two theories, we can split up the discussion in two different schools of thought. Jacobs [11] defines urban population as 'eyes on the street', a natural policy mechanism that supports urban safety through 'natural surveillance'. An open and mixed use environment supports this concept by enabling diversity and activity within the population using the area at different times. While Jacobs suggests that a high diversity among the population and a high ratio of visitors are contributing to an area's safety, Newman [14] argues the opposite. According to his theory, diversity and a high mix of people create the anonymity it needs for crime to take place. Newman suggests that a clear definition of public, semi-public and private space in a low dense and single use urban environment creates a 'defensible space' that is needed to support safety. Newman further argues that low population diversity, low visitor ratio and a high ratio of residents are contributing to an area's safety. Follow-up studies have tried to shed light onto these apparently conflicting theories. For instance, Felson and Clarke [7] have proposed the 'Routine Activity Theory', that studies people dynamics and crime in relation to specific points of interest; they have found that venues such as bars and pubs attract crime by pulling strangers into an area; the presence of middle aged women on the streets detracts crime instead.

These theories suggest different ways to design the built environment so to take advantage of the resulting social control of crime. But which one applies *where*, and also *when*? How do we know that theories developed in the '60s and '70s are still valid fifty years afterwards? To gain a deeper understanding of the context within which a certain theory holds, social science research needs a novel way to validate urban crime theories, that scales up in terms of the geographic urban areas under exam, the population sample captured, and the frequency with which studies can be repeated.

Computational Science & Crime

In recent years, open data movements have made available large repositories of crime data to the public. These circumstances have been useful to start studying crime in a more systematic manner. Data mining has become a popular method for crime research to detect crime patterns in an urban environment. Recorded crime data has been extensively mined to identify crime hot spots within a city [15, 23]. A crime hotspot is defined as an area with high criminal activity in immediately surrounding areas. Hot spots provide researchers with a pattern for crime distribution, shape and orientation [1, 6], and can be even used for crime predictions [2]. However, by focusing on crime density only, these methods do not put crime in any relationship to its environment: they are capable of signaling where crime will happen, without shedding light into possible *reasons* for incidents. According to Jacobs and Newman, the reasons for crime to happen are to be found in the built environment and the population that inhabits and uses it; different methods are required to quantitatively validate such theories.

Recent architectural and urban design research has attempted to describe the relationship between the built environment and crime. Wolfe and Mennis [25] discuss the influence of green space in relation to crime by using satellite images to detect green urban spaces and compare them to recorded crime data. Findings show clearly that well maintained green spaces contribute to less crime through an increased community activity and supervision, as also originally suggested by Jacobs. Hillier and Shabaz[1] discuss Jacobs' and Newman's theories using detailed spatial data about accessibility of the street network in a London borough, to evaluate correlations with recorded crime numbers. Findings show, for instance, that local movement within an area is beneficial to safety, while global movement from outside into an area is not. Furthermore the study supports the theory that a high mix of use is beneficial to safety.

In a follow-up study [16] the same researchers incorporate 'Routine Activities' theories, and explore them using space syntax measurements [9]. The work investigates the relationship between street crime occurrences (categorized as 'snatch', 'thread' and 'attack' crime) and the spatial layout of the street network for a London borough. Findings show an overall higher crime distribution along main roads compared to side roads, with the ratios changing throughout the day: the accumulation of 'snatch' crimes increases in the morning and evening hours dramatically, showing that up to 95% of all incidents happen at on main roads during these hours.

These works show that there is a strong relationship between the built environment and location of crime. However, the findings above also point to the fact that there is a third and

[1] **http://spacesyntax.com/wp-content/uploads/2011/11/Hillier-Sahbaz_An-evidence-based-approach_010408.pdf**. June, 2014

important dimension to the problem: people's dynamics. The very same built environment is appropriated and used by different people for different purposes and in different ways throughout the day. People's dynamics thus need be quantitatively explored in relation to crime too.

When it comes to analysing crime in relation to people, social and criminological research often uses census data. For instance, Tan and Haining in [20] use spatial data of crime and census data to explore the impact of crime on population's health for the city of Sheffield/UK. Findings show significant correlations between health deprivation and crime clusters of an area. Song and Daqian in [19] explored relationships between spatial patterns of property crime and characteristics of neighbourhoods, using census data. Results show significant correlations between property crime activity and socio-economical variables of a neighbourhood. Christens and Speer in [4] use census data to explore the relationship between crime and population density, following Jacob's hypothesis that high population density would predict reduced violent crime. They found the hypothesis to be true for densely populated urban areas, but failed in suburban areas where the population is less dense.

While shedding light into some important relationships between crime and demographics, census data is limited, in that it only offers a static image of the city (i.e., where people residence is), without disclosing where people actually spend time throughout the day. Furthermore, census data is only collected every few years, so the information it provides may become quickly stale, especially for areas undergoing massive urbanisation processes. According to Jacobs and Newman, it is these people dynamics that have great impact on the crime activities of a place; furthermore, they change steadily over time and space, so that we cannot use census data to analyse them.

To study theories of people dynamics in relation to crime, we need to know how this image changes over time and we need to quantify where people live, work and spend their time. To add a dynamic view over census data, surveys are usually conducted, asking people where they work, where they spend their time for leisure, and so on. These data collection methods are however limited, both in terms of the number of respondents they can reach, and in the coarse space/time granularity of the answers obtained. Both limitations can be overcome if we use mobile phone records, collected by telecommunication operators, as a source of information from which to extract people's dynamics at a fine spatio-temporal granularity. Previous work has shown how telecommunication data can be used to understand cities and even whole countries. For instance, Smith Clarke and Mashhadi in [18] use mobile phone data to extract features that can be used as proxy indicators for deprivation at the example of two developing countries. Derived proxies include network activity and traffic between different areas and were correlated with census data to validate the outcome. Correlations showed that discussed variables were clearly meaningful and could be used to evaluate poverty rates in a developing country. In [5] Eagle and Macy extract a measure of communication diversity (e.g. the geographical distribution of a person's social connections) from phone calls in England and found that a higher diversity correlates with socio–economic deprivation. These examples show there are opportunities of using telecommunication data to describe economic poverty. In the next section we will illustrate how this can be extended into the domain of urban crime.

METHOD

In this section, we describe the method we propose to quantitatively validate previous architectural theories of urban crime. We start with a brief description of our datasets; we then present the pre-processing steps these datasets underwent, and finally elaborate on the metrics we extracted from them as proxies for urban crime theories.

Dataset Description

The method we propose requires access to two types of datasets: one that provides information about people's dynamics, and and one that provides information about crimes.

For the former, we use anonymised and aggregated data collected and made available by a mobile telecommunication provider with a 25% penetration in the UK. The dataset contains 12,150,116 footfall count entries for the Metropolitan Area of London for the course of 3 weeks in December 2012/January 2013. The geographic area is divided by the data provider itself into 23,164 grid cells of varying size: for the more densely populated areas within inner London, a grid size is about by 210×210 meters, while the for the less densely areas of Greater London the grid size increases to about 425×425 meters. For each cell, footfall counts are given on a per hour basis over the three week period, not just as total number of people present in that cell at that time slot, but also further broken down by gender (number of males/females), by type (number of residents, workers, visitors) and by age group. In Table 1 we show a sample of our mobile phone dataset.

For the latter, we use open crime data records, which, for the area of Greater London, are made available by two authorities: the Metropolitan Police and the City of London Police. These records provide information about the reporting police district, the exact location (longitude and latitude) of the crime, the name and area code of the crime, and the crime type (which the UK police differentiates into 10 categories: antisocial behaviour, burglary, criminal damage and arson, drugs, other crime, other theft, robbery, shoplifting, vehicle crime and violent crime). Unfortunately, no timestamp is given of when the crime took place/was reported, and the only temporal information we have is the month during which it took place. We thus collected crime data for the months of December 2012 and January 2013 (to temporally match our mobile phone data), and retrieved 83,526 recorded crimes in total. In Table 2 we show a sample of our crime data set. Figure 1 also shows an overview of the spatial distribution of crime throughout the city (the darker the shade of blue, the higher the concentration of crimes). As shown, crime is concentrated in the centre of London, with some other hotspots spread out all over the city.

Date	*Time*	*Grid ID*	*Total*	*Home*	*Work*	*Visit*	*Male*	*Female*	*0–20*	*21–30*	*31–40*	*41–50*	*51–60*	*60+*
10/12/2012	9:00:00	1122...	430	110	290	30	240	190	0	80	90	120	100	40
10/12/2012	10:00:00	2412...	910	210	160	540	520	390	0	180	180	260	170	120
10/12/2012	11:00:00	1092...	900	570	250	80	520	380	10	160	190	250	210	80
10/12/2012	12:00:00	2124...	690	80	120	490	410	280	10	120	150	190	140	80

Table 1. Record sample of mobile phone data, showing the number of people per area, per hour

Crime ID	*Month*	*Reported by*	*Lon*	*Lat*	*Location*	*LSOA Code*	*Crime Type*
df0c4...	2012-12	Met Police	-0.219	51.568	near Clitterhouse Rd	E010...	Burglary
0f9a5...	2012-12	Met Police	-0.217	51.565	near Caney Mews	E010...	Burglary
62235...	2012-12	CoL Police	-0.221	51.570	near Claremont Way	E010...	Crim. damage & arson
194ed...	2012-12	CoL Police	-0.222	51.563	near Petrol Stn	E010...	Crim. damage & arson

Table 2. Record sample of open crime data, showing crime incidents including geo location and crime type

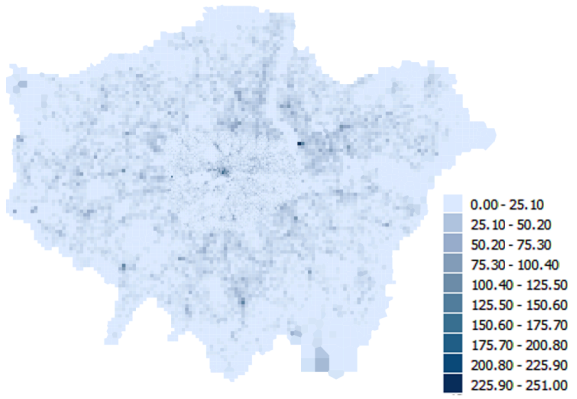

Figure 1. Choropleth showing total crime distribution in London for Dec 2012-Jan 2013, where dark blue indicates areas with a higher crime distribution

Data Pre-Processing

We first cleansed the telecommunication data, so to remove inconsistent entries (i.e., footfall count per area different from the sum of footfall counts broken down by gender, type or age). We further pruned grid cells that fell outside the Greater London area. This caused 1.8% of the raw telecommunication data to be removed.

In order to correlate people dynamics and crime data within an urban environment over time, we then needed to define a common spatio-temporal unit of analysis for both datasets.

In terms of *spatial* unit of analysis, we operated at the level of grid cells defined by the telecomm operator. As mentioned before, these are rather fine-grained cells, varying from 210 $\times$ 210 meters for inner London, to 425 $\times$ 425 meters for outer London. As crime data is recorded in terms of latitude/longitude coordinates, the spatial association of crime data to grid cells was straightforward. For each grid cell, we can thus count the total number of crimes that took place there; we also break down such counter by crime type, distinguishing *street crime*, covering crime most likely happening on the streets (e.g., antisocial behavior, drugs, robbery and violent crime – a total of 47,238 entries), and *home crime*, including crime types happening most likely indoors (e.g., on burglary, criminal damage and arson, other theft and shoplifting – a total of 36,288 entries).

In terms of *temporal* unit of analysis, we needed to align telecomm data, captured at hour-level unit of analysis, with crime data, captured at month-level unit of analysis. To do so, we computed average footfall counts per area per month; to reduce variance, we aggregated separately day-time hour slots (8AM-8PM) and night-time hour slots (8PM-8AM), as well as weekdays vs weekends. For each area, we thus ended up with four footfall count averages. As subsequent correlation analysis results did not show significant differences across these four aggregation values, we will report results on the average case only.

Having cleansed the data and defined a common spatial and temporal unit for analysis, we are now able to define the metrics we will use in our quantitative analysis.

Defining Hypotheses & Metrics

Crime activity

To begin with, we need to quantify crime activity per spatio-temporal unit of analysis. To this end, we cannot simply use a raw count of crimes, as this number alone does not give us information about which area is safe and which is not. Consider two areas with the same number of crimes but very different population density: the probability of being victim of a crime in the two cases is very different. To capture this, we use a metric that we called Crime Activity $CA(i)$; this metric takes into account how safe an area i is, by considering the number of crimes $c(i)$ over the estimated population $p(i)$ present in the area i. The number of crimes (total/home/street) is ready available in our pre-processed crime dataset; as for the number of people present in the area, we considered all people present in area i in the 3 weeks covered by our phone call dataset. Since the crime dataset and telecommunication dataset covered different timespans (8 weeks for the former, 3 weeks for the latter), we multiplied by 3/8 so to have the average number of crimes per person in one week:

$$CA(i) = 3/8 \cdot \frac{c(i)}{p(i)}$$

Figure 2 shows an overview of the Crime Activity $CA(i)$ throughout Greater London: the darker the shade of blue, the higher the crime activity in that area. By comparing this with Figure 1, we can observe that, while areas in the center of London have higher crime counts (Figure 1), the risk of being victim of a crime, that is, crime rate normalised by people present in that area (as captured by our Crime Activity $CA(i)$ metric), is much higher outside inner London (Figure 2).

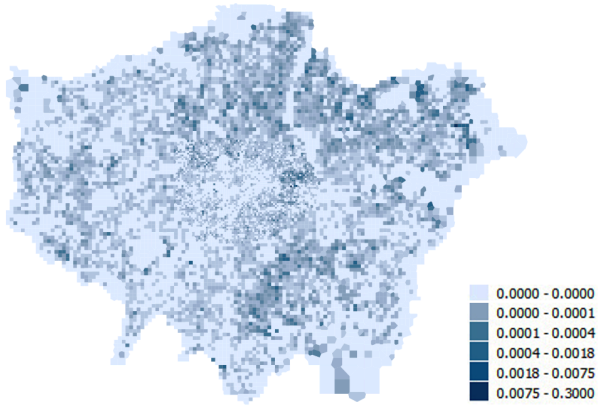

Figure 2. Choropleth map showing total crime activity in London for Dec 2012-Jan 2013, whereas dark blue indicates areas with a higher crime activity

Having defined a metric that captures crime per spatio-temporal unit of analysis, we next define metrics that act as proxies for urban crime theories linking people dynamics with crime activity. We have a total of six metrics and associated hypotheses ($H1$ to $H6$).

H1 - Diversity of people

According to Jacobs, diversity of functions in an area supports the area's safety, as it attracts a greater diversity of people at different times that collectively act as 'eyes on the street'. Jacobs points out in her examples the importance of age diversity. Newman, on the contrary, suggests that high diversity of people in an area provides opportunities for crime to happen through anonymity. However, the two theories do not describe the term 'diversity' further in detail. From our telecommunication dataset, we are able to extract one metric of diversity, relative to age. For each area under exam, we have a footfall count breakdown relative to age in terms of these age groups: 0–20, 21–30, 31–40, 41–50, 51–60, 60+. We thus computed age diversity Da as the Shannon-Wiener diversity index[2] over these counts. When correlating this metric with crime activity, according to Jacobs we would expect areas with high age diversity to be safer than others, while following Newman's theory we would expect the opposite.

H2 - Ratio of visitors

According to our reviewed theories, there are opposite opinions about the contribution towards crime of a high ratio of visitors for an area. Jacobs points out their importance for 'eyes on the streets', while Newman and follow up criminological research suggests that a high ratio of visitors actually brings crime to an area as a result of anonymity. To explore these apparently contrasting theories, we quantify the ratio of visitors Rv (relative to total footfall count) per area, and will then correlate these values with crime activity per area. Following Jacobs, we would expect to have less crime where there are more visitors, whereas following Newman we would expect the opposite.

[2]The Shannon diversity index is a measure that reflects how many different entries there are in a data set and the value is maximised when all entries are equally high [17].

H3 - Ratio of residents

A high number of residents in an area is strongly supported by Newman's territorial approach of 'defensible space' to reduce crime. Jacobs mentions residents as a less important factor for the 'natural surveillance' theory compared to shopkeepers, as residents provide less attention for street level activities. To validate Newman's theory, we compute the ratio of residents Rr compared to the overall population for each area, and correlate them with crime activities. According to Newman, we would expect a high ratio of residents in an area to correlate with less crime.

H4 - Ratio of workers

Jacobs suggests that a high variety of functions in an area supports urban safety, pointing out the importance of shops in an area, as shop keepers and people who work in an area provide 'natural surveillance'. We will validate the statement by computing the ratio of workers Rw compared to the area's overall population for each area, and compute correlations with crime activities. According to Jacobs' theory, we would expect in areas with a high ratio of workers to have less crime, compared to areas with a low ratio of workers.

H5 - Ratio of female population

Felson and Clarke suggest that a high ratio of women on the street is a positive sign towards urban safety, as they act as 'crime detractors'. To validate this statement, we will compute the ratio of female population Rf compared to the overall population for each area, and correlate the values with crime activity. Even though our proxy Rf is not precise enough to represent ratio female population on the streets only, but the overall ratio of female population for an area instead, we would expect a lower crime activity in areas with high ratio of females according to the theory.

H6 - Ratio of young people

According to Felson and Clarke, a higher ratio of young people leads to more criminal incidents in an area, as they show a higher aggression potential compared to elder people. We defined our young population group as those falling in the 0–20 and 21–30 age groups in our telecommunication dataset. We then compute the ratio of young (Ry) population relative to the area's overall population, and correlate it with the crime activity. In this case, the hypothesis is that areas with a higher ratio of young people also have higher crime rates.

Summary of Metrics

In Figure 3 we show diversity and ratio distributions for our six metrics across Greater London as choropleth maps. The darker the shade of blue, the higher the value of the metric in that area. We observe that population's age diversity (Figure 3(a)) is generally low for Inner London, while it increases towards the edges, especially in the south east. As expected, a high ratio of visitors is found in the centre of London (Figure 3(b)), which offers most points of interest as attractions and retail. Some parts of the edges show also a high visitor ratio, focusing towards the north and the east. While visitors concentrate in the centre of London, residents show an opposite picture: Figure 3(c) reveals that areas with high residents' ratio are to be found outside of the city and spread all

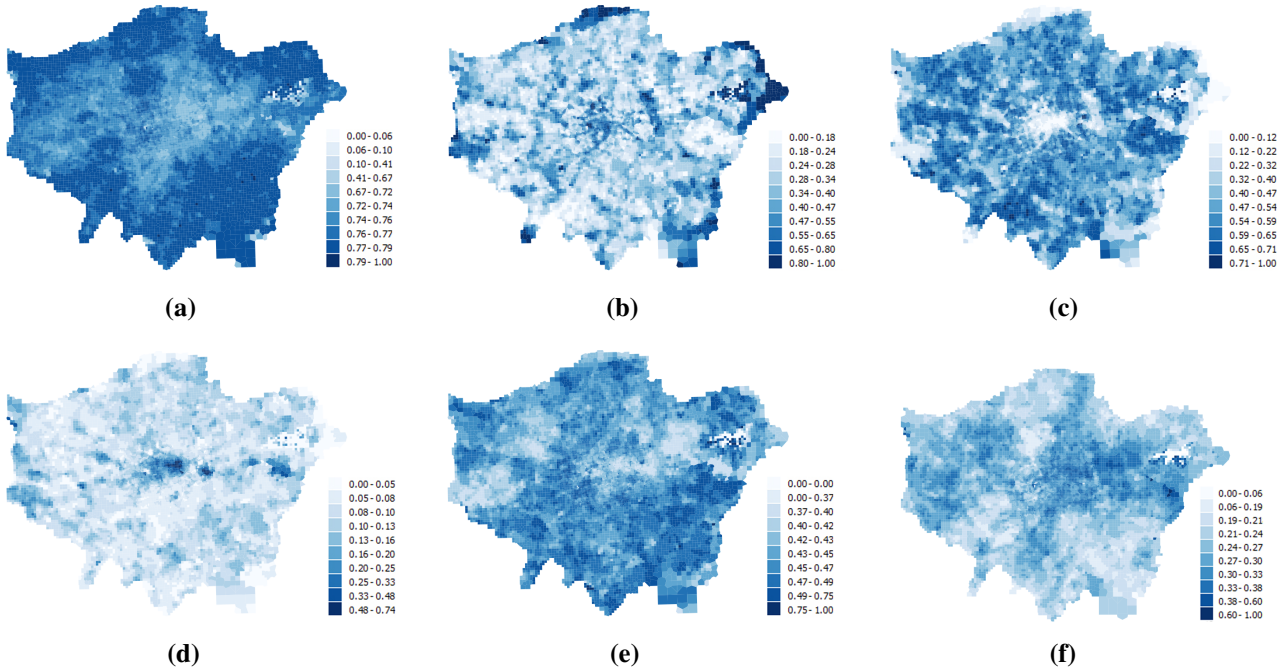

Figure 3. Choropleth map of age diversity (a), ratio visitors (b), ratio residents (c), ratio workers (d), ratio female (e), ratio young (f)

over Greater London. Ratio of workers (Figure 3(d)) is concentrated to the centre and sparsely distributed outside of Inner London, with isolated sub-centres. In Figure 3(e) we observe generally a higher female population ratio for the south of London, compared to the north. Furthermore the pattern reveals for districts of Inner London a higher female ratio towards the west, compared to the east of London. Finally, Figure 3(f) shows a higher concentration of young population in the centre of London spreading out towards the east, which is known to be popular among young people.

Correlation Analysis

In the previous sections, we have defined metrics both for crime activity, and for the six proxies we will use in relation to the selected urban crime theories under exam. The next step in our method is to correlate these metrics. To this purpose, the major challenge of our approach was to manage the spatial autocorrelation present in our datasets. Spatial autocorrelation is rather common when studying spatial processes, whereby observations captured at close geographic proximity appear to be correlated with each other, either positively or negatively, more than observations of the same properties at further distance [13, 22]. Spatial autocorrelation violates the assumption of standard statistical techniques that observations are independent; as such, common correlation analysis techniques that use Pearson, Spearman or Kendall coefficients to explore relationships between variables cannot be applied. To address this issue, we will use the Tjostheim correlation index [21, 10] instead; this index can be seen as an extension to Spearman and Kendall coefficients, so to explicitly account for spatial properties in our data. All results presented in the next section are thus to interpreted as correlations r_t computed between crime activity CA_i and the six metrics $H1 - H6$, using the Tjostheim correlation index.

Hypothesis	Variable	*Total Crime*	*Street Crime*	*Home Crime*
H1: diversity of people	Da	**-0.12**	**-0.14**	-0.10
H2: ratio of visitors	Rv	**-0.28**	**-0.26**	**-0.23**
H3: ratio of residents	Rr	**0.27**	**0.26**	**0.21**
H4: ratio of workers	Rw	0.02	0.02	0.03
H5: ratio of females	Rf	**0.16**	**0.14**	**0.16**
H6: ration of young	Ry	**0.13**	**0.17**	**0.10**

Table 3. Tjostheim Correlations r between Total, Street, Home Crime Activity and individual variables (shown in bold are statistically significant results with p-value < 0.01)

RESULTS

Greater London

Table 3 presents the Tjostheim correlation coefficients between each variable introduced in the previous section and the crime activity, for the total number of crimes as well as the two crime sub-groups street/home.

H1: Diversity of people

We find significant negative correlations between diversity of age and crime, both for total crime ($r_t = -0.12$) and for street crime ($r_t = -0.14$); for home crime, $r_t = -0.10$ but the p-value was greater than 0.01 so the result is not statistically significant. These findings seem to support Jacob's theory of 'natural surveillance', where she linked different age groups in the same area to a variety of activities taking place in the same space, and this was further associated to less crime.

H2: Ratio of visitors

Next we have computed the ratios of visitors (Rv) to the total population for each area and found a significant negative correlation with crime activity. For total crime, we found $r_t = -0.28$, for street crime $r_t = -0.26$ for home crime $r_t = -0.23$ (second row of Table 3). In all three cases, a higher ratio of visitors is linked to lower crime activity.

These findings again support Jacobs theory, that states how visitors lead to increased activity in an area, supporting the 'eyes on the street' theory and consequent increase in the levels of safety of an area.

H3: Ratio of residents

If we now focus on residents, we found a positive correlation between the ratio of residential population (Rr) in an area and crime. Newman's theory of 'defensible space' suggests that an increased ratio of residents is linked to urban safety, by clearly separating spaces for visitors from spaces for residents. However, our findings do not seem to support Newman's theory. In fact, results show that a high ratio of residents is statistically correlated with crime. We found $r_t = 0.27$ for total crime, $r_t = 0.26$ for street crime and $r_t = 0.21$ for home crime (third row of Table 3).

H4: Ratio of workers

Contrary to Newman, Jacobs suggests that residents are less involved with natural surveillance compared to, for example, shopkeepers, as they provide less attention to what is taking place around. Jacobs suggests to look at the relationship between the ratio of working people (Rw) in an area and crime instead. In particular, she posits that a high number of functions, especially shops, leads to increased safety as they attract people and support 'natural surveillance'. Unfortunately, our results do not help shed light into this controversy, as results are not statistically significant (fourth row of Table 3).

H5: Ratio of female population

A surprising result is found in the positive correlation between the female population (Rf) and crime in an area ($r_t = 0.16$ for total crime, $r_t = 0.14$ for street crime and $r_t = 0.16$ for home crime – fifth row of Table 3). This result shows the opposite of Felson and Clark's theory, suggesting that a higher ratio of female population in London is actually statistically correlated to a higher crime activity in an area. However, as we have stated above, Rf represents the overall ratio of female population for an area and not only the ratio of female population on the streets, so this result could have been affected by the poor metric.

H6: Ratio of elder population

Finally we have computed the ratio of the young people (Ry) per area and we have correlated it with crime. Findings show a light positive correlation between the younger population and crime ($r_t = 0.13$ for total crime, $r_t = 0.17$ for street crime and $r_t = 0.10$ for home crime – last row of Table 3). This result would support Felson and Clarke's statement, that a higher proportion of young population ratio is associated with more crime in an area.

Zooming in at Borough Level

We have shown how one may use our proposed methodology to quantitatively study the validity of certain urban crime theories at scale. However, one may wonder whether the chosen scale (that is, the whole metropolitan area of London) is appropriate for this type of investigations. London is a very large and complex city, composed of many different neighbourhoods, each with its own distinguishing characteristics

Variable	*Min*		1st Qu.		*Median*		*3rd Qu.*		*Max*
Da	-0.41	***	-0.19	***	-0.11		0.01	*	0.45
Rv	-0.57	***	-0.34	**	-0.27	***	-0.18	**	-0.03
Rr	-0.04	***	0.20	**	0.26	***	0.34	**	0.61
Rw	-0.32	***	-0.08		0.02	*	0.11	**	0.39
Rf	-0.18		0.02	*	0.15	***	0.25	**	0.47
Ry	-0.41	*	0.01		0.08	**	0.22	**	0.45

Table 4. Summary statistics of the Tjostheim correlations between total crime activity and each individual variable on the 32 London boroughs. Stars indicate the percentage of Tjostheim correlations that are statistically significant in each quartile (p-values < 0.01): 0% ' ' 25% '*' 50% '' 75% '***' 100%**

in terms of built environment, demographics, and possibly people dynamics. Choosing the whole of London as a single context to study urban theories may thus hide the fact that, in practice, different theories and correlations may hold in different London neighbourhoods. Indeed, theories by Jacobs and Newman had been previously investigated only at neighbourhood level, never at such a big geographic scale.

Our proposed methodology does not prescribe the size of the geographic area at which it should be applied, and it can thus be easily and automatically re-applied to separately study smaller areas. We have repeated our analysis, this time separately considering the 32 administrative boroughs in which London is divided. Table 4 shows summary statistics of the correlations between total crime activity and each variable previously defined, as they vary across boroughs. By looking at these new results, and by comparing them with those in Table 3, we note that all the individual variables that were (positively or negatively) correlated to crime activity in the whole city of London, now show considerably higher (in positive or in negative) correlations in at least half of the 32 London boroughs. This indeed suggests that this smaller unit of analysis can be more appropriate to investigate the validity of urban crime theories. For those metrics for which we did not find significant statistical results when considering the whole of London, we now find significance in certain areas. For instance, our findings reveal that a quarter of London boroughs show a significant negative correlation between the ratio of working population (Rw) and total crime ($-0.32 > r_w > -0.08$), whereas for Greater London correlations of the same variable were found not to be significant ($r_w = 0.02$). However, the results at borough level also show that, for another quarter of London boroughs, Rw is actually significantly and positively correlated with total crime instead ($0.11 > r_w > 0.32$). These findings suggest that different, possibly conflicting theories may hold in different parts of the same metropolitan city; using our method, it is possible to investigate whether a theory hold at the full city scale or not. If not, the method also helps social science researchers identify the sub-areas that require further qualitative investigation.

Building a Model of Crime

The quantitative evaluation method proposed before is not only useful for social scientists to validate urban crime theories at scale. Indeed, it can be used to also build predictive models of crime, to the benefit of city administrators and planners. To illustrate how, we selected the five boroughs within which the previously studied crime theories exhibited strongest correlation; these are shown in Figure 4. For each

of the five selected boroughs, we then built an Ordinary Least Square (OLS) regression model, obtaining an adjusted R-squared ranging between 0.20 and 0.30. It is worth pointing out that, although these results are not extraordinarily high, they are very promising since they show that a very simple linear model, which considers just the individual variables listed previously, is able to 'explain' up to the 30% of variation of crime activity. A complete model of crime should also include other metrics, for instance, from census data for socio–economic factors, and from the built environment for the city's physical properties. Here we show that, even by just looking at metrics of people dynamics obtained from mobile phone data, we can gain a good insight into urban crime and we can explain up to 30% of its variance in the selected boroughs.

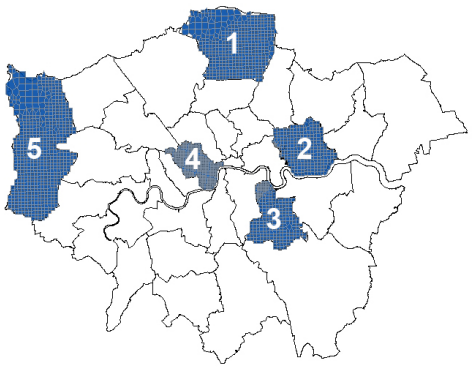

Figure 4. The five London boroughs where the Tjostheim correlations between total crime activity and the individual variables showed the strongest results, numbered from 1 (strongest) to 5 (least strongest)

DISCUSSION & FUTURE WORK

In this paper, we have presented a method to validate architectural theories on urban crime and people dynamics in a quantitative way. The method required access to two sources of information: crime data records and records about people presence in the built environment. From the former, we extracted a metric called Crime Activity $CA(i)$, that captures the concentration of crime relative to population in that area. From the latter, we extracted metrics that act as proxies for urban crime theories. Using correlation analysis, we have shown it is now possible to quantitatively validate urban crime theories at large geographic scale and frequent intervals, at almost no cost.

One may wonder how widely applicable our method is, in the sense of getting access to such data sources. Supported by the ongoing open data movement, an increasing amount of crime data for cities in different parts of the world is freely available and can be used for our purposes. Telecommunication data on the other hand is more difficult to access, but a variety of data mining challenges, such as D4D – Data for Development challenge[3] and the Big Data Challenge,[4] show a clear trend of mobile phone providers towards making their data available to the public. This development suggests that the proposed methodology will become increasingly applicable in the next years.

[3]D4D – Data for Development, by Orange: **http://www.d4d.orange.com/en/home**. June, 2014

[4]Big Data Challenge, by Telecom Italia: **http://www.telecomitalia.com/tit/en/bigdatachallenge.html**. June, 2014

Implications

The method we have proposed has both practical and theoretical implications. From a practical standpoint, tools can be built on top of it, to the benefit of different stakeholders. For example, citizens may appreciate predictive crime tools they can use to decide what areas of a city to explore safely and which to avoid; administrators may use tools that highlight time variations in the model, to monitor the impact of processes such as urbanisation and gentrification on area's dynamics and crime activity; and city planners may use tools that highlight crime model similarities and differences across different city neighbourhoods.

From a theoretical standpoint, the method offers social science researchers a new way to validate past crime theories, as well as develop new ones. We have shown how to use the method to validate past theories for the city of London. The same method could be used for a multitude of cities around the world, so to advance knowledge in terms of the contexts within which past theories hold. The method can also be re-applied over time, on newly available data streams, to detect possible changes that call for social scientists to refine past theories or develop new ones. Even when looking at the single city of London in a single period, we have shown that some theories do not hold across all boroughs, thus calling for deeper qualitative investigations. Indeed, we foresee the proposed quantitative method to be used in conjunction with qualitative methods, during alternate phases of theory development and evaluation.

Limitations

Our work suffers from a number of limitations. First, the temporal unit of analysis used in the two datasets at hand was different (i.e., crime data was recoded on a monthly basis, while footcounts were recorded on a hourly basis). This required a data-processing step that forces us to operate at the coarses level of granularity. This inevitably kept interesting questions unanswered. As previous studies suggest, different crime types follow different spatial and temporal patterns [8]; had we had access to crime timestamps, we would have been able to explore the relationship between people dynamics and crime activity in a more fine grained manner.

Furthermore, our findings are based on mobile phone data collected by a single mobile phone provider. Being one of the major mobile phone providers in the UK with almost 25% market share in 2013, our data set covers a high number and variety of people, but leaves a grey space for people using other providers or pay–as–you–go options that are excluded from the data. As general mobile phone usage in the UK is quite high – in 2013 94% of all adults use mobile phones and 61% run on contract – the choice of data source for our purposes seems over all promising, but could be improved by access also to other provider's data. We would also expect further improvement of the model by including data of

pay–as–you–go customers, as the socio–demographic characteristics of these users (e.g., young, income deprived) may be under represented in our dataset.

Finally, using recorded open source crime data as benchmark to test our hypotheses brought up limitations in terms of completeness. Accordingo the British Crime Survey [12] the number of non-recorded crime data – the 'dark figures' – includes about half of all UK crime activities, which therefore our study does not take into account. These circumstances may lead to a blurred outcome and show that urban safety is not strictly determined by where crime is being reported.

Future Work

Our ongoing and future work spans two main directions: on one hand, we aim to expand the model, so to incorporate properties of people dynamics, the built environment, and census within a single framework. In so doing, we expect not only to predict crime activity with greater accuracy, but also to understand the dependencies between all such variables in relation to crime. On the other hand, we aim to apply the model to data from multiple cities in the world. In the last year, telecommunication data has been released both for cities in Europe (e.g., Milan) and in Africa (e.g., Dakar); we wish to apply the method presented in this paper in these very different settings, so to understand in what contexts certain theories hold, thus advancing knowledge in the area of urban crime.

REFERENCES

1. Chainey, S., Reid, S., and Stuart, N. When is a hotspot a hotspot? a procedure for creating statistically robust hotspot maps of crime. *Innovations in GIS 9 Socio-economic Applications of Geographic Information Science* (2002).
2. Chainey, S., Tompson, L., and Uhlig, S. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21 (1-2) (2008), 4–28.
3. Chaplin, R., Flatley, J., and Smith, K. Home office statistical bulletin: Crime in england and wales 2010/11. *Home Office Statistical Bulletin* (2011).
4. Christens, B., and Speer, P. Predicting violent crime using urban and suburban densities. *Behavior and Social Issues*, 14 (2005), 113–127.
5. Eagle, N., and Macy, M. Network diversity and economic development. *Science*, 1029 (2010).
6. Eck, J., Chainey, S., Cameron, J., Leitner, M., and Wilson, R. Mapping crime: Understanding hot spots. *Special Report NIJ* (2005).
7. Felson, M., and Clarke, R. *Opportunity Makes the Thief: Practical theory of crime prevention*. Home Office, 1998.
8. Felson, M., and Poulsen, E. Simple indicators of crime by time of day. *International Journal of Forecasting*, 19 (2003), 595–601.
9. Hillier, B., and Hanson, J. *The Social Logic of Space*. Cambridge University Press, 1984.
10. Hubert, L. J., and Golledge, R. G. Measuring association between spatially defined variables: Tjostheim's index and some extensions. *Geographical Analysis*, 14 (1982), 273–278.
11. Jacobs, J. *The Death and Life of Great American Cities*. Random House Inc., 1961.
12. Jansson, K. *British Crime Survey: Measuring crime for 25 years*. 2006.
13. Legendre, P. Spatial autocorrelation: Trouble or new paradigm? *Ecology 74*, 6 (1993), 1659–1673.
14. Newman, P. *Defensible Space: Crime Prevention Through Urban Design*. Macmillian Pub Co, 1972.
15. Paynich, R. Identifying high crime areas. *International Association of Crime Analysts*, 2 (2013).
16. Sahbaz, O., and Hiller, B. The story of the crime: functional, temporal and spatial tendencies in street robbery. In *Proc of 6th International Space Syntax Symposium, Istanbul* (2007), 04–14.
17. Shannon, C. A mathematical theory of communication. *The Bell System Technical Journal 27* (1948), 379–423 and 623–656.
18. Smith Clarke, C., Mashhadi, A., and Capra, L. Poverty on the cheap: estimating poverty maps using aggregated mobile communication. In *Proc of CHI'14* (2014), 511–520.
19. Song, W., and Daqian, L. Exploring spatial patterns of property crime risks in changchun, china. *International Journal of Applied Geospatial Research 4*, 3 (2013), 80–100.
20. Tan, S., and Haining, R. An urban study of crime and health using an exploratory spatial data analysis approach. In *Proc of ICCSA'09*, vol. 1 (2009), 269–284.
21. Tjostheim, D. A measure of association for spatial variables. *Biometrika*, 65,1 (1978), 109–114.
22. Tobler, W. A computer movie simulating urban growth in the detroit region. *Economic Geography 46* (1970), 234–240.
23. Wang, D., Ding, W., Lo, H., Stepinski, T., Salazar, J., and Morabito, M. Crime hotspot mapping using the crime related factors a spatial data mining approach. *Applied Intelligence 39*, 4 (2006), 772–781.
24. WHO, U. H. *Hidden cities: unmasking and overcoming health inequities in urban settings*. WHO, Library Cataloguing-in-Publication Data, 2010.
25. Wolfe, M., and Mennis, J. Does vegetation encourage or suppress urban crime? Evidence from Philadelphia, PA. *Landscape and Urban Planning 108* (2012), 112–122.