Mining Mobile Phone Data to Investigate Urban Crime Theories at Scale

Martin Traunmueller[1], Giovanni Quattrone[2], and Licia Capra[1]

[1] ICRI Cities, Dept. of Computer Science, University College London
[2] Dept. of Computer Science, University College London
Gower Street, WC1E 6BT, London, UK
{martin.traunmueller.11,g.quattrone,l.capra}@ucl.ac.uk

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 observe 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 (e.g., population diversity) and by Felson and Clarke (e.g., ratio of young people). We conclude the paper with a discussion of the implications of this work on social science research practices.

Keywords: Urban crime, telecommunication data, open data, data mining.

1 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 [29]. On the other hand, crime activities are on the rise (e.g., [5]), especially in densely populated areas [13]. Being able to understand and quantify the relationship between people presence and crime activity in an area has thus become an important concern, for both citizens, urban planners and city 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 [12] suggests that population diversity, activity

L.M. Aiello and D. McFarland (Eds.): SocInfo 2014, LNCS 8851, pp. 396–411, 2014.

and a high mix of functions lead to less crime for an area, whereas Newman [15] 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.

In this paper we propose a new method to quantitatively investigate urban crime theories at scale, using open crime data records and anonymised mobile telecommunication data. From the former, we extract quantitative information about crime activity, as it 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 presence in an area with crime. We then use correlation analysis between crime data and our defined metrics to validate urban crime theories at scale. We apply this method to data obtained for the city of London, UK, and find that, in this city and at the present time, Jacobs' theory of 'natural surveillance' [12] holds: we discover that age diversity, as well as the ratio of visitors in a given area, are significant and negatively correlated with crime activities; furthermore, Felson and Clarke theory [9] that links a higher presence of young people with higher crime is also confirmed. 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 afford them deep insights into the causality of certain phenomena, the latter afford 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 by discussing implications, limitations and future steps.

2 Related Work

2.1 Background

Most well known architectural theories about the relationship between people dynamics, the urban environment and crime lead back to the studies of Jacobs [12] and Newman [15], with two different schools of thought. Jacobs [12] 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 [15] 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 [9] 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 area under exam, the population sample captured, and the frequency with which studies can be repeated.

2.2 Computational Science and 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 tool for crime research to detect crime patterns in an urban environment. Recorded crime data has been extensively mined to identify crime hotspots within a city [16,27,3,8], and can even be used for crime predictions [4]. These methods are capable of signaling where crime will happen; however, they do not shed light into possible *reasons* for incidents.

Recent architectural and urban design research has attempted to describe the relationship between the built environment and crime. Wolfe and Mennis [30] 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 [18] investigate the relationship between street crime occurrences 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. 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 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 dynamics thus need to 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 [23] use spatial data of crime and census data to explore the impact of crime on population health for the

city of Sheffield, UK. Song and Daqian [22] explored relationships between spatial patterns of property crime and socio-economic variables of a neighbourhood. Christens and Speer [6] 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 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 reside), 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 urbanization processes. According to Jacobs and Newman, it is these people dynamics that have great impact on the crime activities of a place which change steadily over time and space, so that we cannot use census data to analyse them.

People dynamics have started to be inferred from geo-located social networks, and used for different purposes. For instance, Prasetyo et. al. [17] use Twitter and Foursquare data to analyse the impact of major natural disasters on people; they do so for haze events in Singapore, and discuss how their approach can help both the private and public sector to better prepare themselves to similar future events. Wakamiya et. al. [26] use geo-located Twitter data to examine crowd interactions, from which social neighbourhood boundaries are defined, thus expanding upon the traditional concept of spatial, administratively-defined neighbourhoods. Discussing crime, Wang et. al. [28] use sentiment analysis to relate the content of Twitter messages to hit-and-run crime activity and demonstrate a high usability for crime prediction. Social media is a rich data source from which to derive information about people dynamics; however, it is also unrepresentative of the whole urban population, because of high bias in its adoption [2]. An alternative data source that can be used to mine people dynamics in urban areas, and that is subject to significantly lower bias than social media, is telecommunication data.

Telecommunication data has been recently used to understand the relationship between cities (and even whole countries) and socio-economic deprivation, both in the developed world [7] and in developing countries [20]. In relation to crime, recent work [1] uses a similar mobile phone data set as used in this paper in combination with census data to predict crime activity for urban areas of London. As results show the importance of variables extracted from the mobile phone data set predicting almost 70% of the cases when included, they underline the importance of people diversity in relation to crime activity in an area as described by Jacobs [12]. Focusing less on the predictive and more on the descriptive aspect, we believe the same data can be used to understand other established theories as well, as we will show next.

3 Method

In this section, we describe the method we propose to quantitatively explore 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.

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

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

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

3.1 Dataset Description

The method we propose requires access to two types of datasets: one providing information about people dynamics, and with information about crimes. For the purpose of this study, we chose datasets that cover the city of Greater London, UK. We did so as London represents a large and complex metropolitan city, composed of many different neighbourhoods, each with its own distinguishing characteristics in terms of built environment, demographics, and people dynamics. It thus represents a case where qualitative approaches to investigate urban crime theories would not scale, both because of the geographic span of the areas to study, and because of the time frequency with which one may wish to repeat these studies (e.g., to observe changes in relation to ongoing immigration processes [21]).

People dynamics. We use anonymised and aggregated data collected and made available by a mobile telecommunication provider in context of a data mining challenge 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 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, further broken down by gender (number of males/females), by type (number of residents, workers, visitors) and by age group. Table 1 shows a sample of our mobile phone dataset.

Crime data. We use open crime data records[1], 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: i.e. burglary, drugs, robbery, shoplifting, etc.). Unfortunately, no timestamp is given of when the crime took place/was reported, and the only temporal information we have is the month

[1] Open–source crime data: `http://data.police.uk`. June, 2014

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. Table 2 shows a sample of our crime data set.

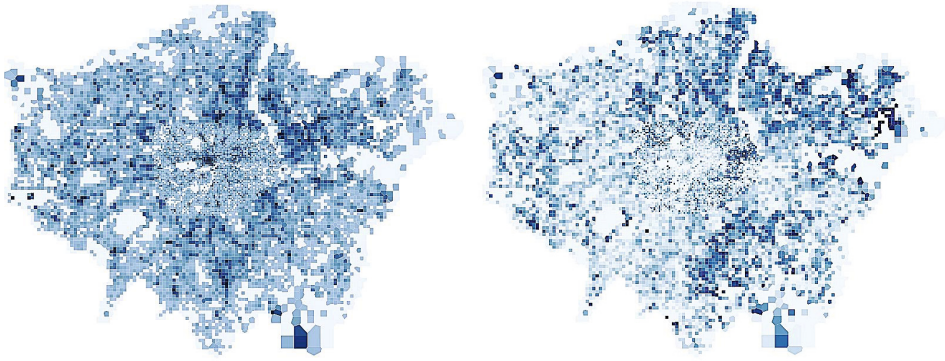
3.2 Data Pre-Processing

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 grid 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 for the weekday/daytime 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.

3.3 Hypotheses and Metrics

Crime Count and Crime Activity. To begin with, we need to quantify crime per spatio-temporal unit of analysis. For each area i, we consider two complimentary metrics: crime count $CC(i)$, and crime activity $CA(i)$. The former simply counts the number of crimes that have taken place in area i; since most of the areas under study have comparable size, we may consider $CC(i)$ as a way of measuring crime normalized by area size. Areas have similar sizes, but not similar population density. To investigate possible differences caused by population density, we use $CA(i)$ to quantify crime normalized by population density instead; we can consider this metric as an indicator of the probability of being victim of a crime. We can compute crime activity $CA(i)$ by dividing the number of crimes in an area $CC(i)$ by the estimated population $P(i)$



(a) crime count **(b)** crime activity

Fig. 1. Choropleth maps showing crime count CC (left) and crime activity CA (right) all over Greater London for Dec 2012-2013, where the darker the shade of blue, the higher the crime rate in that area

present in area i. The number of crimes per area $CC(i)$ 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:

$$CA(i) = 3/8 \cdot \frac{CC(i)}{P(i)}$$

Figure 1 shows the spatial distribution of crime count and crime activity over Greater London (the darker the shade of blue, the higher the $CC(i)$ and $CA(i)$ values). As shown, crime count $CC(i)$ is found to be higher in the centre of London, with some other hotspots spread out all over the city (Figure 1a), whereas crime activity $CA(i)$ (that is, crime count normalised by people present in that area) is much higher outside inner London (Figure 1b). 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 count and 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' in further 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 D_a as the Shannon-Wiener diversity index[2] over these counts. When correlating this metric with crime, according to Jacobs we would expect areas with higher 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 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 R_v (relative to total footfall count) per area, and will then correlate these values with crime metrics. Following Jacobs, we would expect to have less crime where there are more visitors, whereas following Newman we would expect the opposite.

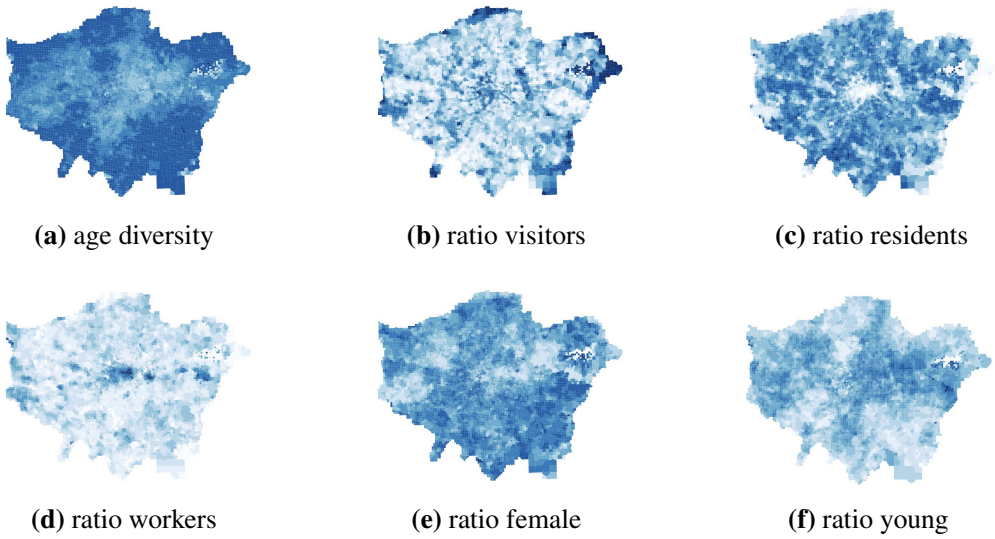
H3 - Ratio of Residents. A high number of residents in an area is strongly supported by Newman's territorial approach of 'defensible space' to reduce crime. Jacobs mentions residents as a less important factor for the 'natural surveillance' theory compared to shopkeepers, as residents provide less attention for street level activities. To validate Newman's theory, we compute the ratio of residents R_r compared to the overall population, and correlate them with crime metrics. 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 R_w compared to the area's overall population for each area, and compute correlations with crime metrics. According to Jacobs' theory, we would expect to have less crime in areas with a higher 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, we will compute the ratio of female population R_f compared to the overall population for each area, and correlate the values with crime metrics. We would expect a lower crime activity in areas with a higher 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 (R_y) 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.

[2] The Shannon diversity index is a measure that reflects how many different entries there are in a data set and the value is maximized when all entries are equally high [19].



(a) age diversity **(b)** ratio visitors **(c)** ratio residents

(d) ratio workers **(e)** ratio female **(f)** ratio young

Fig. 2. Choropleth maps of our six metrics, where the darker the shade of blue, the higher the value of the metric

Summary of Metrics. Figure 2 illustrates the distributions of our six metrics across Greater London as choropleth maps. We observe that population's age diversity (Figure 2(a)) is generally low for Inner London, while it increases towards the edges. A high ratio of visitors is found in the centre of London (Figure 2(b)), which offers most points of interest as attractions and retail, and in some parts of the edges towards the north and the east. Ratios of residents (Figure 2(c)) and workers (Figure 2(d)) show a clear opposite picture between them: while workers concentrate in the central business districts, residents are found to be more widespread in less central boroughs. In Figure 2(e) we observe generally a higher female population ratio for the south of London, compared to the north. Finally, Figure 2(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.

3.4 Correlation Analysis

Having defined metrics for crime count, crime activity and the six proxies relating to selected urban crime theories, the next step is to correlate these metrics. 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 [14]. This is the direct quantitative demonstration of Tobler's First Law of Geography, which states that everything is related to everything else, but near things are more related than distant things [25]. Spatial autocorrelation violates the assumption that observations are independent; as such, common correlation analysis techniques that use Pearson, Spearman or Kendall coefficients to explore relationships

Table 3. Tjostheim Correlations r between crime metrics (crime count and crime activity) and individual variables; shown in bold are statistically significant results with p-value < 0.01

		crime count $CC(i)$			crime activity $CA(i)$		
Hypothesis	*Variable*	*Total Crime*	*Street Crime*	*Home Crime*	*Total Crime*	*Street Crime*	*Home Crime*
H1: diversity of people	D_a	**-0.27**	**-0.26**	**-0.23**	**-0.12**	**-0.14**	-0.10
H2: ratio of visitors	R_v	**-0.20**	**-0.20**	**-0.17**	**-0.28**	**-0.26**	**-0.23**
H3: ratio of residents	R_r	**0.17**	**0.19**	**0.14**	**0.27**	**0.26**	**0.21**
H4: ratio of workers	R_w	0.09	0.07	0.09	0.02	0.02	0.03
H5: ratio of females	R_f	-0.02	-0.02	-0.01	**0.16**	**0.14**	**0.16**
H6: ration of young	R_y	**0.31**	**0.31**	**0.25**	**0.13**	**0.17**	**0.10**

between variables cannot be applied. To address this issue, we will use the Tjostheim correlation index instead [24,11]; 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 be interpreted as correlations r_t computed between crime count $CC(i)$, crime activity $CA(i)$ and the six metrics $H1 - H6$, using the Tjostheim correlation index.

4 Results

4.1 Correlation Results for Greater London

Table 3 presents the Tjostheim correlation coefficients between our two crime metrics ($CC(i)$ and $CA(i)$) and each variable introduced in the previous section. Note that the same correlation signs were found both when using crime count and crime activity, with only relatively small changes in actual correlation values. We interpret this as an indication of the robustness of our proposed metrics. The findings discussed below apply to both crime metrics used.

H1: Diversity of People. We find significant negative correlations between diversity of age and crime, both for total crime ($r_t = -0.27$ for CC and $r_t = -0.12$ for CA) and for street crime ($r_t = -0.26$ for CC and $r_t = -0.14$ for CA); for home crime, we found significant results only for the correlations with CC ($r_t = -0.23$) whereas for CA the p-value was found to be 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. We found a significant negative correlation between the ratios of visitors (R_v) of an area and crime. For total crime, we found $r_t = -0.20$ for CC and $r_t = -0.28$ for CA; for street crime, $r_t = -0.20$ and $r_t = -0.26$ respectively; and for home crime $r_t = -0.17$ and $r_t = -0.23$ (second row of Table 3). In all three cases, a higher ratio of visitors is linked to lower crime. These findings again support Jacobs' theory of 'eyes on the street', with consequent increase in the levels of safety of an area where visitors concentrate.

H3: Ratio of Residents. If we now focus on residents, we found a positive correlation between the ratio of residential population (R_r) 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 this. In fact, results show that a high ratio of residents is statistically correlated with crime (from a minimum of $r_t = 0.14$ for home crime correlated with crime count CC, to a maximum of $r_t = 0.26$ for street crime and crime activity CA (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 (R_w) 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 they 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 (R_f) and crime activity CA 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), though correlations with crime count CC were found not significant . 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, we should note a limitation of our metric in this case: in fact, R_f represents the overall ratio of female population for an area (residents, workers, or visiting), and not only the ratio of female population on the streets, so this result could have been affected by a relatively poor metric.

H6: Ratio of Younger Population. Finally, we have computed the ratio of young people (R_y) per area and we have correlated it with crime. Findings show a positive correlation between the younger population and crime (from a minimum of $r_t = 0.10$ for home crime and crime activity CA, to a maximum of $r_t = 0.31$ for total/street crime and crime count CC – last row of Table 3). This result would support Felson and Clarke's theory that a higher proportion of young population ratio is associated with more crime in an area.

4.2 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. As mentioned before, London is a very large and complex city, composed of many different neighbourhoods. Choosing the whole of London as a single context to study urban theories may thus hide the fact that, in practice, different

Table 4. Summary statistics of the Tjostheim correlations between total crime count *CC* 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%

Variable	*Min*		1st Qu.		*Median*		*3rd Qu.*		*Max*
D_a	-0.51	**	-0.27	***	-0.20	**	-0.12		0.23
R_v	-0.53	**	-0.30	***	-0.20	***	0.00	*	0.18
R_r	-0.16	**	-0.04	***	0.17	***	0.31	**	0.60
R_w	-0.28	***	-0.02	**	0.09	*	0.17	*	0.44
R_f	-0.28	*	-0.08	***	0.03	*	0.17	*	0.47
R_y	-0.18		0.18		0.24	***	0.40	**	0.54

Table 5. Summary statistics of the Tjostheim correlations between total crime activity *CA* 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%

Variable	*Min*		1st Qu.		*Median*		*3rd Qu.*		*Max*
D_a	-0.41	***	-0.19	***	-0.11		0.01	*	0.45
R_v	-0.57	***	-0.34	**	-0.27	***	-0.18	**	-0.03
R_r	-0.04	***	0.20	**	0.26	***	0.34	**	0.61
R_w	-0.32	***	-0.08		0.02	*	0.11	**	0.39
R_f	-0.18		0.02	*	0.15	***	0.25	**	0.47
R_y	-0.41	*	0.01		0.08	**	0.22	**	0.45

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.

As our proposed methodology is not prescribed to a size of geographic area, we have repeated our analysis, this time separately considering the 32 administrative boroughs in which London is divided. We assigned grid cells to boroughs boundaries according to their centroids. Table 4 shows summary statistics of the correlations between crime count CC and each variable previously defined, as they vary across boroughs; Table 5 shows results obtained when using crime activity CA instead. 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 have a significant negative correlation between the ratio of working population (R_w), and both crime count CC ($-0.28 > r_w > -0.02$) and crime activity CA ($-0.32 > r_w > -0.08$), whereas for Greater London correlations of the same variable were found not to be significant (CA: $r_w = 0.02$, CC: $r_w = 0.09$). Interestingly, the results at borough level also show that, for another quarter of London boroughs, R_w is actually significantly and positively correlated with crime activity CA ($0.11 > r_w > 0.39$) and crime count CC ($0.17 > r_w > 0.44$) instead. 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 holds 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.

5 Discussion, Limitations and Future Work

Summary. In this paper, we have presented a method to investigate architectural theories of urban crime and people dynamics in a quantitative way. The method requires access to two sources of information: crime data records and records about people presence in the built environment. From the former, we extracted two metrics of crime, crime count $CC(i)$ and crime activity $CA(i)$. 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 investigate urban crime theories at large geographic scale and frequent intervals, at almost no cost.

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 the 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.

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, as citizens, administrators and city planners. To illustrate what such a tool would look like, we built an Ordinary Least Square (OLS) regression model for each of the 33 boroughs in Greater London separately, as well as for the whole of London. For each such regression model, we analysed the adjusted R^2 value, to understand the extent to which the built model was capable of 'explaining' crime variance. We found that, for a model that considers Greater London as a whole, the adjusted R^2 value is 0.12. However, when we build such model per borough, we are capable of reaching an adjusted R^2 between 0.20 and 0.30 for a quarter of the boroughs. We believe these results are quite promising, considering that we used a rather simple linear model, with just 'people dynamics' variables, as listed previously. 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.

[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

From a theoretical standpoint, the method offers social science researchers a new way to investigate past crime theories, as well as develop new ones. We have shown how to use the method to explore 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 in selected areas. 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 recorded on a monthly basis, while foot-counts were recorded on a hourly basis). This required a data-processing step that forces us to operate at the coarser level of granularity. This inevitably kept interesting questions unanswered. As previous studies suggest, different crime types follow different spatial and temporal patterns [10]; had we had access to crime timestamps, we would have been able to explore the relationship between people dynamics and crime 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 dataset covers a high number and variety of people, but leaves a grey space for people using other providers or PayAsYouGo options that are excluded from the data. For those people covered from our dataset it stays unclear how the provider categorized them as resident, worker or visitor which could provide a more detailed insight. By including additional datasources, as for instance urban topology data, the ratio of workers could be discussed in more detail. Note that these limitations pertain the datasets used, and not the method proposed. As such, while actual results on the validity of the reviewed urban crime theories for the case study of Greater London would have to be revisited should more accurate and complete datasets become available, we believe the validity of the method withstands.

Future Work. Our 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. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., Pentland, A.: Once upon a crime: Towards crime prediction from demographics and mobile data. In: ICMI (2014)
2. Boyd, D., Crawford, K.: Critical questions for big data. Information, Communication and Society 15(5), 662–679 (2012)
3. Chainey, S., Reid, S., 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)
4. Chainey, S.P., Tompson, L., Uhlig, S.: The utility of hotspot mapping for predicting spatial patterns of crime. Security Journal 21(1-2), 4–28 (2008)
5. Chaplin, R., Flatley, J., Smith, K.: Home office statistical bulletin: Crime in england and wales 2010/11. Home Office Statistical Bulletin (2011)
6. Christens, B., Speer, P.W.: Predicting violent crime using urban and suburban densities. Behavior and Social Issues (14), 113–127 (2005)
7. Eagle, N., Macy, M.: Network diversity and economic development. Science (1029) (2010)
8. Eck, J., Chainey, S., Cameron, J., Leitner, M., Wilson, R.: Mapping crime: Understanding hot spots. Special Report NIJ (2005)
9. Felson, M., Clarke, R.: Opportunity Makes the Thief: Practical theory of crime prevention. Home Office (1998)
10. Felson, M., Poulsen, E.: Simple indicators of crime by time of day. International Journal of Forecasting (19), 595–601 (2003)
11. Hubert, L.J., Golledge, R.G.: Measuring association between spatially defined variables: Tjostheim's index and some extensions. Geographical Analysis (14), 273–278 (1982)
12. Jacobs, J.: The Death and Life of Great American Cities. Random House Inc. (1961)
13. Jansson, K.: British Crime Survey: Measuring crime for 25 years (2006)
14. Legendre, P.: Spatial autocorrelation: Trouble or new paradigm? Ecology 74(6), 1659–1673 (1993)
15. Newman, P.: Defensible Space: Crime Prevention Through Urban Design. Macmillian Pub. Co. (1972)
16. Paynich, R.: Identifying high crime areas. International Association of Crime Analysts (2) (2013)
17. Prasetyo, P.K., Gao, M., Lim, E.P., Scollon, C.N.: Social sensing for urban crisis management: The case of singapore haze. In: Proc of SocInfo 2013, pp. 478–491 (2013)
18. Sahbaz, O., Hiller, B.: The story of the crime: functional, temporal and spatial tendencies in street robbery. In: Proc of 6th International Space Syntax Symposium, Istanbul, pp. 4–14 (2007)
19. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal 27, 379–423, 623–656 (1948)
20. Clarke, C.S., Mashhadi, A., Capra, L.: Poverty on the cheap: estimating poverty maps using aggregated mobile communication. In: Proc of CHI 2014, pp. 511–520 (2014)
21. Snyder, M.: The impact of recent immigration on the london economy. Technical report, London School of Economics and Political Science (2007)
22. Song, W., Daqian, L.: Exploring spatial patterns of property crime risks in changchun, china. International Journal of Applied Geospatial Research 4(3), 80–100 (2013)
23. Tan, S.-Y., Haining, R.: An urban study of crime and health using an exploratory spatial data analysis approach. In: Gervasi, O., Taniar, D., Murgante, B., Laganà, A., Mun, Y., Gavrilova, M.L. (eds.) ICCSA 2009, Part I. LNCS, vol. 5592, pp. 269–284. Springer, Heidelberg (2009)
24. Tjostheim, D.: A measure of association for spatial variables. Biometrika (65,1), 109–114 (1978)

25. Tobler, W.R.: A computer movie simulating urban growth in the detroit region. Economic Geography 46, 234–240 (1970)
26. Wakamiya, S., Lee, R., Sumiya, K.: Social-urban neighborhood search based on crowd footprints network. In: Jatowt, A., Lim, E.-P., Ding, Y., Miura, A., Tezuka, T., Dias, G., Tanaka, K., Flanagin, A., Dai, B.T. (eds.) SocInfo 2013. LNCS, vol. 8238, pp. 429–442. Springer, Heidelberg (2013)
27. Wang, D., Ding, W., Lo, H., Stepinski, T., Salazar, J., Morabito, M.: Crime hotspot mapping using the crime related factors - a spatial data mining approach. Applied Intelligence 39(4), 772–781 (2006)
28. Wang, X., Gerber, M.S., Brown, D.E.: Automatic crime prediction using events extracted from twitter posts. In: Yang, S.J., Greenberg, A.M., Endsley, M. (eds.) SBP 2012. LNCS, vol. 7227, pp. 231–238. Springer, Heidelberg (2012)
29. U. H. WHO. Hidden cities: unmasking and overcoming health inequities in urban settings. WHO, Library Cataloguing-in-Publication Data (2010)
30. Wolfe, M.K., Mennis, J.: Does vegetation encourage or suppress urban crime? Evidence from Philadelphia, PA. Landscape and Urban Planning 108, 112–122 (2012)