Guns of Brixton: Which London neighborhoods host gang activity?

Guns of Brixton: which London neighborhoods host gang activity?

Alessandro Venerandi
University College London
Gower Street
London, UK
alessandro.venerandi.12@ucl.ac.uk

Giovanni Quattrone
University College London
Gower Street
London, UK
g.quattrone@cs.ucl.ac.uk

Licia Capra
University College London
Gower Street
London, UK
l.capra@ucl.ac.uk

ABSTRACT

Previous works in architecture and social science found that aspects of the built environment such as density, connectivity, and house typologies are related to crime. However, these studies are qualitative, and thus hardly repeatable at larger scales. In this work, we overcome this limitation by offering a quantitative approach that explores the relationship between the configuration of the built environment and the activity of criminal groups in city areas. The method extracts a wide set of metrics related to aspects of urban form from openly accessible datasets. We then input these metrics in a step-wise logistic linear model, using presence of gang activity as dependent variable, and obtain a parsimonious model with an excellent fit when applied to the metropolitan area of London, UK. We then use values and slopes of model coefficients to build a narrative of the typical city area characterized by gang activity, re-connecting to previous theories. Outcomes of this research can help policy makers and architects in better understanding the relationship between neighborhood design and criminal activity.

CCS Concepts

•Applied computing → Architecture (buildings);

Keywords

Urban Form; Quantitative Analysis; Open Data

1. INTRODUCTION

Concentration of criminal activity in certain areas is related to a wide range of different factors such as economic conditions [12, 28], social issues [27, 4], and level of education [17, 3]. The built environment also plays a role in this. Urbanists have investigated the relationship between urban form and crime, adopting different methodologies and techniques. On one side, there are those who used qualitative methods and whom research outcomes appear to disagree for certain aspects. This is the case, for example, of the theories postulated by Jane Jacobs and Oscar Newman in the 1960's and 1970's. Jacobs suggested that medium to high densities of people and buildings, pedestrian environments and mix of uses and activities are beneficial against crime [13]; Newman had an almost opposite view and advocated for a more sparse built form characterized by more segregated activities [24]. The main limitation of these works is that they are costly to reproduce at scale, in terms of time and resources, and thus findings are hard to generalize.

For these reasons and thanks to the advent of GIS science (i.e., the discipline that analyses geographic information through the use of computational techniques), researchers started to develop quantitative methods to study the link between urban form and crime. For example, Budd analyzed the British Crime Survey data and found that flats were the safest housing typologies [2]. Hillier studied the configuration of the street network and found that dead end roads do not necessarily attract criminal activity [10]. Other recent studies focused on the relation between high densities and presence of crime [8, 16]; however did not find any association. This set of works helps in providing more evidence-based findings on the relationship between urban form and crime; however, they are still limited, as single aspects of the built environment are studied separately, whereas it is their mutual interplay that might better explain the presence of criminal activity.

We therefore aim to provide, through a quantitative method, a more comprehensive characterization of what aspects of urban form are related to an aspect of crime (i.e., gang activity) by using open data as input and by taking the metropolitan area of London (i.e., Greater London) as case study. More precisely, we identify a set of descriptors which are able to capture multiple aspects of the built environment such as build period, type and fiscal band of dwellings, place accessibility, and street network connectivity. We extract measures to describe the built environment from Ordnance Survey (OS) VectorMap District and the London Datastore (i.e., the official web portal for statistical information on Greater London). We derive the areas where gangs currently operate from a geodataset made publicly available on an online article[1] by the Independent (i.e., a well known British national newspaper). The information contained in that dataset was gathered through an interview which saw the involvement of an ex-sergeant of the London Metropolitan Police. Having computed the measures and assigned presence or absence of gang activity to the different areas of London, we perform step-wise logistic regression to obtain a parsimonious set of descriptors of the urban environment which are significantly associated with the activity of criminal groups. Such model shows an excellent overall performance (i.e., McFadden's pseudo R^2 of 0.24). Moreover, we use signs and values of the regression coeffi-

Urb-IoT'16, May 24-25, 2016, Tokyo, Japan

DOI: http://dx.doi.org/10.1145/2962735.2962750

[1] http://i100.independent.co.uk/article/these-are-londons-gang-territories-in-a-single-map--Z1oinQab_g

cients to provide a description of the typical city area where gangs are likely to operate and to reflect back on past theories that confirm our findings. The outcomes of this research can be helpful to city administrators and urban designer as they can inform the debate on how to design safer neighborhoods.

The remainder of this paper is structured as follow: we first present previous works which address related topics. We then provide a description of the datasets used and metrics. We follow with the methodology implemented and results obtained. Finally, we conclude the paper with a discussion of the implications and limitations of our study.

2. RELATED WORK

Many works have been carried out to study the relationship between built form and presence of crime. To study this topic, researchers from the most varied set of backgrounds (e.g., architecture, journalism, data science) implemented qualitative and quantitative approaches. Journalist and activist Jane Jacobs belongs to the former group and suggested that a tendentially crime-free urban environment should be characterized by medium to high densities, and by a well connected street network, where the presence of different activities, as well as pedestrians and cyclists can provide enough "eyes on the street" to prevent crimes [13]. City planner and architect Oscar Newman used a similar method; however, he reached different conclusions. Newman, in fact, was in favor of lower densities and less connected places as reduced accessibility decreases the chance of escape opportunities to criminals and, at the same time, raises the level of street control that residents can exert as there are less passersby [24]. The main limitation of these works is that they are hardly generalizable and replicable as they are based on personal observations or on spatial samples of small size.

More recently, other researchers tried to overcome this methodological issue by implementing quantitative techniques. Budd performed a multivariate analysis of official crime data controlling for social and economic factors and suggested that flats were the safest house type, followed by terraced and detached houses [2]. Hillier studied the relationship between the presence of cul-de-sacs in neighborhoods, through the Space Syntax approach [9], and found that cul-de-sacs can be safe if located into an urban fabric characterized by high inter-visibility among properties and a through-street pattern [10]. In a more recent study, Hillier and Sahbaz [11] provided separate answers to a set of questions related to urban form and crime; the main new findings relate to building density, movement of people in streets, and places with mix use. Densities of people and dwellings generally increase safety; however, these densities work better if they are located at the ground level rather than at higher ones. Movement of people was found to be a beneficial factor against crime in local streets but not in major roads. Mixed use streets can be considered safe when they have a good number of residents, more vulnerable otherwise. Although these works provide relevant scientific grounded findings, they are limited by the fact that urban features are studied separately. Urban form, in fact, is a complex entity composed by multiple features which quite plausibly interact with one another rather then acting in isolation. For this reason, we argue that different urban features should be studied in conjunct rather than separately.

Other quantitative works in the field of computational social science tried to quantify the relationship between urban form and perception of safety. More precisely, Naik *et al.* [23] built an online survey where participants were asked to assign a vote to different streetscapes based on their perceptions of safety. After having collected data from more than 7,000 participants, they built an algorithm able to predict the perceived safety of around 20 cities of the United States. Other researchers [30] tried to validate urban theories at-scale by performing correlation analysis between a set of metrics extracted from telecommunication data representing the presence of specific kinds of people in space and levels of crimes for the metropolitan area of London, UK. The main findings suggest that diversity of people is linked to less criminal activity while the above-average presence of young people to more recorded crime.

In this work, we propose a quantitative method to study the relationship between a set of different descriptors of urban form extracted from open datasets and gang activity for the areas of Greater London, UK. We present the datasets used in this study next.

3. DATASETS

To carry out this study, we had to access datasets with information concerning the urban form of a city, and the areas where criminal groups operate within such a city. We chose Greater London as case study both because we had easy access to these datasets, and because it is a large metropolitan city for which we expect relevant patterns between urban form and crime to emerge. We extracted measures to quantify the built environment from OS VectorMap District, descriptors of the housing stock from the London Datastore, and London's gang territories from an openly accessible map created through Google My Maps.

3.1 OS VectorMap District

The OS VectorMap District provides a reliable and detailed map of the UK in vectorial format. It is provided and kept updated by the UK official mapping agency (i.e., Ordnance Survey) and it is openly accessible via a dedicated institutional website.[2] This dataset provides information on multiple geographic features such as streets, building footprints, natural resources in TIFF and GIS shapefile formats. OS VectorMap District was released for the first time for open access in April 2010. Since this geodataset is provided in 100km by 100km tiles, we downloaded those tiles which contained the metropolitan area of London and obtained 165,921 street segments and 102,187 polygons representing building footprints. The geographic data contained in this dataset dates back to September 2015.

3.2 London Datastore

Launched in January 2010 as part of the Mayor's Smart London Plan, the London Datastore is an official web portal curated by the Greater London Authority (GLA) which contains a wide range of freely accessible statistical information on London (e.g., employment, pollution levels, health).[3] The data collected on this website is provided and kept updated by different institutional bodies such as the Valuation Office Agency, Transport for London, Department of Health.

For the purpose of this work, we accessed two datasets containing information on the London housing stock: one with property build periods and house types (i.e., the Dwellings by Property Build Period and Type),[4] and another with the council tax bands for all London's properties (i.e., the Dwellings by Council Tax Band).[5] The former dates back to 2014 while the latter to 2015.

[2]https://www.ordnancesurvey.co.uk/opendatadownload/products.html

[3]http://data.london.gov.uk/

[4]http://data.london.gov.uk/dataset/property-build-period-lsoa/resource/0198b6ca-9297-418a-97d2-d3bd9ef6f3ff

[5]http://data.london.gov.uk/dataset/property-build-period-lsoa/resource/13154d31-7264-41ce-9611-8736dd0630ed

3.3 London's gang territories

In an online article published in 2015,[6] a journalist from the Independent interviewed an ex-sergeant of the Metropolitan Police on the issue of criminal groups in London. Beside giving general information on the phenomenon, the ex-sergeant was able to provide names of London's criminal groups and places where they currently operate, identifying a total of 190 areas; the journalist included this information in the article by creating an openly accessible interactive map built through Google My Maps.[7] After having downloaded the KML file associated with this map, we imported it in a GIS software and obtained the 190 polygons representing the areas where London's gangs operate. A map of these areas is provided in Figure 1.

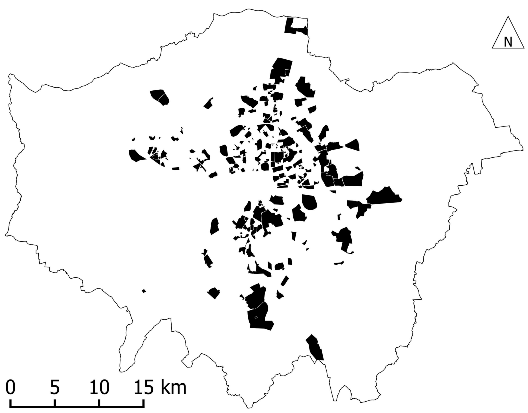

Figure 1: **Areas where London's gangs operate.**

4. METRICS

Our hypothesis is that a specific configuration of the urban form and certain characteristics of dwellings are related to gang activity. To test this assumption, we subdivided London in different areas – which the next section will describe in detail – then, for each of these areas, we computed a total of thirteen measures from the datasets presented in the previous section: six of them measure the configuration of the built environment while the other seven measure housing stock characteristics.

4.1 Configuration of the built environment

After having explored literature in urban studies, we identified the following measures related to the configuration of the built environment:

- *Dead-end density* (DEDEN). Dead-end roads (or cul-de-sacs) are streets which end with no intersections. These were a common design feature in most of the urban and suburban developments planned after World War II. Jane Jacobs argued against this design approach as, in her view, cul-de-sacs are detrimental to urban safety as they decrease street network connectivity. This, in return, reduces the amount of pedestrian passage and activities at the ground floor which guarantee the natural surveillance of streets [13]. Inspired by this theory, we propose a variable which is able to capture how much a city area is characterized by the presence of cul-de-sacs by calculating the ratio between the number of cul-de-sacs and the areal unit.

- *Regularity* (REG). A street network can have a more regular configuration and, therefore, tending towards a system of interconnected perpendicular streets (i.e., grid) or it might show a more organic structure which resembles the lymphatic vases of a leaf. Jacobs emphasized the perks of the grid street pattern as it promotes pedestrian presence and more intervisibility between sidewalks which, in turn, provide natural surveillance [13]. Based on this theory, we propose a metric which quantifies the degree of geometrical regularity of a street network; this is calculated as the standard deviation of node degrees (i.e., number of streets connected to an intersection) for areal unit normalized on the average node degree.

- *Connected node ratio* (CNR). This measures the degree of connectivity of a street network. This has been deemed [13] to be a fundamental characteristic for crime-free city areas as more connectivity also promotes pedestrian passage and thus more natural surveillance (i.e., "eyes on the street"). We derived the way to calculate this variable from previous work carried out by Garrick and Marshall [18]; it is computed as the ratio between the number of real intersections (i.e., not cul-de-sacs) and the total number of intersections per areal unit, normalized on the extension of the areal unit.

- *Intersection density* (IDEN). This represents a measure of street network density. Recent studies seem not to have found significant relationship between density and presence of crime [8, 16]; however, Jacobs suggested that medium to high densities are beneficial as they increase the number of people in the urban environment and consequentially the "eyes on the street" effect [13]. Following this theory, we decided to calculate *intersection density* in the way Garrick and Marshall defined it in their work [18], that is the ratio between the number of real intersections and the total number of intersections per areal unit.

- *Percentage of open space* (OSPERC). More or less open space determines how dense or sparse is the urban fabric. As illustrated in the previous point, density seems to be an important factor in determining the presence or absence of crime in urban areas. Thus, this measure differs from the previous one as density, in this case, is measured as an aspect of the urban fabric (i.e., how much space is left unbuilt) rather than one of the street network. We decided to measure the amount of open space by using a method previously implemented by Banister *et al.* [1]. This variable is calculated as percentage ratio between the extension of the areal unit minus its built up part, divided by the extension of the areal unit.

- *Betweenness* (BET). It quantifies the accessibilty of a street considering all the possible paths in a street network. Spatial accessibility is another important feature of the urban environment which has showed to be associated with levels of crime. Hillier and Sahbaz, for example, found that accessible major roads, which involve movement at a large scale, tend to attract crime; conversely, accessible local streets tend to be safer [11]. We therefore included *betweenness* in our study as calculated by Porta et al. [25]. We refer to their work for the details of the formula.

[6] http://i100.independent.co.uk/article/these-are-londons-gang-territories-in-a-single-map--Z1oinQab_g
[7] https://www.google.com/maps/d/u/0/viewer?mid=z0TiisceV2FI.kS6rmTEYXY4A

4.2 Housing stock characteristics

As mentioned in Section 2 of this paper, previous studies have revealed relationships between dwellings' characteristics and crime. We thus chose a metric that aims to capture the distinctiveness of each city area, as related to all the others, in terms of its housing stock characteristics. To do so, we computed the *Offering Advantage* (OA) metric. OA quantifies whether a city area offers more of a specific kind of house compared to the average offering of that kind of house for the whole city under study. More specifically

$$OA(h_i, a_k) = \frac{count(h_i, a_k)}{\sum_{j=1}^{N} count(h_j, a_k)} \cdot \frac{\sum_{j=1}^{N} count(h_j)}{count(h_i)}$$

where $OA(h_i, a_k)$ represents the OA of a kind of house h_i in the city area a_k; $count(h_i, a_k)$ counts how many properties of type h_i are present in the city area a_k; N is the total number of the different kinds of houses; finally, $count(h_i)$ counts how many properties of kind h_i are present in the whole city. This metric has already been implemented in other studies and showed to be very effective in profiling city areas in terms of what features characterize them [31, 26]. We computed OA for all the variables presented below.

We extracted five variables related to dwelling typologies from the Dwellings by Property Build Period and Type dataset. These are:

- *OA of flats* (OA-FLATS);
- *OA of semi-detached houses* (OA-SEMI);
- *OA of detaced houses* (OA-DET);
- *OA of terraced houses* (OA-TERR);
- *OA of bungalows* (OA-BUN).

From the same dataset, we extracted another variable, this time, related to the build periods of London's properties. The dataset providing this information has a fine temporal resolution (i.e., thirteen temporal thresholds of about nine years each). For the purpose of our work, though, such fine temporal resolution was not necessary as architectural styles and, consequently, the urban form they generate change at a much slower pace (e.g., the Gothic style lasted three centuries). We thus identified a temporal threshold which was sensitive to this matter. For the particular period of time framed by the dataset considered (i.e., 1900 - 2015), we recognized two phases: before and after 1928. This year corresponds to an event, the Congres Internationaux d'Architecture Moderne (CIAM), or International Congress of Modern Architecture, which brought important changes in the planning practice worldwide. From that time on, in fact, a relevant number of architecture firms and governmental planning departments embraced the modernist approach in city planning. This saw the abandonment of design approaches based on density and pedestrian areas, in favor of one characterized by sparse residential towers in a more car-oriented type of city [6]. After having defined this temporal threshold, we grouped together the temporal bins referring to time periods pre-1928 (i.e., pre-1900, 1900 - 1918, 1919 - 1929) and calculated OA for this group. The variable thus obtained is:

- *OA of pre-CIAM properties* (OA-PRECIAM).

Finally, we identified a variable which provides information on the council tax bands of London's dwelling. The dataset from which we extracted this information (i.e., Dwellings by Council Tax Band) counts the number of properties for each council tax band starting from A, which correspond to dwellings with the cheapest fiscal band, to H, which represents properties with the most expensive ones. After having performed a preliminary analysis, we noted that the eight council tax bands showed a natural split more or less in the middle of the scale (i.e., between band C and D). We therefore grouped the first three bands (i.e., band A, B, C) and computed OA for this group. The variable thus obtained is:

- *OA of dwellings with cheap council tax bands* (OA-CHEAPCTB).

5. METHOD

The aim of this work is to study the relationship between a set of descriptors of the urban environment and the activity of criminal groups in London neighborhoods. To do so, we needed to first define the spatial and temporal units of analysis adopted in our study. Then, given the binary nature of our dependent variable (i.e., presence or absence of gang activity), we opted for a binary logistic regression model as it is appropriate to identify what the most influential variables are in affecting the probability of having active criminal groups in areas. This technique requires the absolute independence of observations; however, our variables may have shown spatial dependencies. We thus tackled this issue by performing a spatial auto-correlation test.

This section is structured as follows: we first present the units of analysis adopted in this study (Section 5.1). Then, since our variables may have had different shapes and magnitudes – and since this would have made it difficult to compare the relative role that each of them played in affecting the probability of having presence of gang activity – we transformed them through a normalization and standardization procedure. We illustrate this step in Section 5.2. We conclude by presenting the method used to tackle possible spatial auto-correlation in the independent variables (Section 5.3) and by illustrating the adopted logistic regression model (Section 5.4).

5.1 Units of analysis

The extensions of the gangs' territories identified in the article mentioned above vary quite substantially, from a minimum of 1.3 hectares (i.e., around the size of a football ground) to a maximum of 588.3 hectares with most of the areas being smaller than 30 hectares. To keep the fine spatial granularity of this information, we selected the most fine grained official statistical area available for the UK – the Lower Super Output Area (LSOA) – as the spatial unit of analysis. LSOAs are defined by the UK Office for National Statistics[8] and were designated to be as consistent in population size as possible with an average value of around 1,500 people. There are 4,835 LSOAs within the boundary of Greater London, with a minimum extension of 1.7 hectares and a mean value of around 30 hectares.

To match the information concerning the location of the gangs' territories with the chosen areal unit of analysis, we marked LSOAs as *containing gang* if more than 50% of their surfaces were overlapping with any gang territory. Conversely, we marked them as *not containing gang* in the opposite case. As for the variables presented in Section 4, these were computed at LSOA level.

For what concerns the temporal aspect, the datasets used in this work are aligned (i.e., 2015) expect for the Dwellings by Property Build Period and Type one which dates back to 2014. We assume that this would not affect the robustness of our results as urban form – the subject of this work – change at a relatively slow pace.

[8]http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/super-output-areas--soas-/index.html

5.2 Normalization and standardization

Having identified the unit of analysis, we calculated the thirteen variables presented in the previous section for the 4,835 London areas. As we introduced before, this work involves the interpretation of regression coefficients which, for being comparable to one another, need to be on the same scale of values. To this end, we transformed our variables in two subsequent steps. Firstly, we normalized them through exponentiation. Secondly, we computed their relative z scores.

5.3 Spatial auto-correlation test

As we mentioned above, the regression technique used in this work requires the absolute independence of observations; however, our variables might show spatial auto-correlation (i.e., spatial dependency). This is a common phenomenon in geographic data and it can be interpreted as direct demonstration of Tobler's First Law of Geography: "everything is related to everything else, but near things are more related than distant things" [29]. In practice, spatial auto-correlation is the tendency of nearby observations to be correlated to one another. To exclude the variables which presented this issue, we performed a well-known geostatistical test – the Moran's test [5]. This technique determines whether spatial dependencies are present in data, by assessing the relationships of its relative spatial matrix. The Moran's test falls in the category of null hypothesis testing, whereby a p-value greater than 0.05 (at 95% confidence level) implies weak evidence that data shows auto-correlation; on the other hand, a p-value smaller than 0.05 implies strong evidence that data does show spatial dependency. This test also outputs an index which can be read as a Pearson's r and which indicates the strength of the spatial dependency of the tested variable.

5.4 Multicollinearity and logistic regression

After having performed the test illustrated in the previous subsection, we checked whether our variables present some form of multicollinearity through cross-correlation analysis. Although multicollinearity is not problematic per se because it does not invalidate a logistic regression model, it could be still an issue if this is phenomenon is accentuated as it can increase the variance of the coefficient estimates thus making it difficult to elaborate interpretations. Depending on whether variables showed inter-dependencies or not, two further steps were possible. If the variables were not to show cross correlations, we would have followed with simple logistic regression. On the other hand, if they were to show dependencies, we would have proceeded with step-wise logistic regression with backward elimination [7]. This latter technique handles multicollinearity of variables by performing a series of iterative tests on all candidate variables until an optimal model is reached.

We evaluated model performance and legitimacy of coefficients through the computation of the McFadden's pseudo R^2 [20]

$$R^2 = 1 - \frac{\ln(L_{full})}{\ln(L_{null})}$$

where L_{full} indicates the estimate likelihood of the adopted logistic regression model with predictors; conversely, L_{null} denotes the estimate likelihood of a null model – that is the model with only an intercept and no covariates. Since a likelihood falls between 0 and 1, the logarithm of this value is less than or equal to zero. If a model shows a very high fit, then its likelihood is almost 1, and therefore the log-likelihood value $\ln(L_{full})$ will be close to 0 and the McFadden's pseudo R^2 will be close to 1. Conversely, if a model does not present a good fit, the ratio $\frac{\ln(L_{full})}{\ln(L_{null})}$ will be close to 1 and the McFadden's pseudo R^2 will be close to 0. Due to the logarithm transformation of the estimate likelihood, McFadden's pseudo R^2 values ranging between 0.2 to 0.4 represent an excellent model fit [21].

Finally, as for the interpretation of our logistic regression model, this work includes the interpretation of regression coefficients, or β coefficients. These indicate how much one unit change in the independent variables affects the predicted logistic odds of the dependent variable, where the logistic transformation of the odds p is

$$\frac{p}{(1-p)}.$$

Negative values indicate an inverse relationship between the independent variable and the probability of "success" of the dependent one; positive values represent, instead, a positive relationship.

6. RESULTS

In this section, we illustrate the outcomes of the Moran's test performed on the thirteen variables; we then present the results of the cross correlation analysis and the model obtained from the logistic regression; finally, we interpret the outcomes.

6.1 Spatial auto-correlation test

Outcomes from the Moran's test (Table 1) showed that some of the variables presented spatial auto-correlation. These are: *connected node ratio* (CNR), *percentage of open space* (OSPERC), *betweenness* (BET), and *OA of flats* (OA-FLAT). These all showed a significant Moran's index of around 0.10 and a p-value smaller than 0.05. We therefore had to withdraw these variables from our study and continue only with the ones denoted with 'pass' equal to Y (i.e., yes) in Table 1.

Table 1: **Results of the Moran's test.**

Category	*Variable*	*Moran's index*	*p-value*	*Pass*
Built environment	$DEDEN$	0.04	0.21	Y
	REG	0.04	0.23	Y
	CNR	0.10	0.02	N
	$IDEN$	0.08	0.07	Y
	$OSPERC$	0.09	0.04	N
	BET	0.09	0.04	N
Housing	$OA-FLAT$	0.11	0.02	N
	$OA-SEMI$	0.08	0.07	Y
	$OA-DET$	0.03	0.24	Y
	$OA-TERR$	0.06	0.13	Y
	$OA-BUN$	0.08	0.05	Y
	$OA-PRECIAM$	0.08	0.05	Y
	$OA-CHEAPCTB$	0.08	0.06	Y

6.2 Cross-correlation analysis

Having excluded the above mentioned variables, we then ascertained whether the remaining ones showed inter-dependencies. Results from the cross correlation analysis confirmed the presence of patterns of dependencies (Figure 2). The OA values of suburban housing typologies (i.e., bungalows, semi and detached houses) are negatively correlated with *intersection density* and *OA of dwellings with cheap council tax bands*. The same suburban house types also seem to be positively correlated among them. Moreover, less *regularity* of the street network is positively associated with presence of dead-end roads. Finally, the OA of properties built before the CIAM (1928) shows moderate negative correlations with *OA of bungalows*, *regularity* and *dead end density*. Given the presence of cross correlations among variables, we opted for the step-wise logistic regression whose results are presented next.

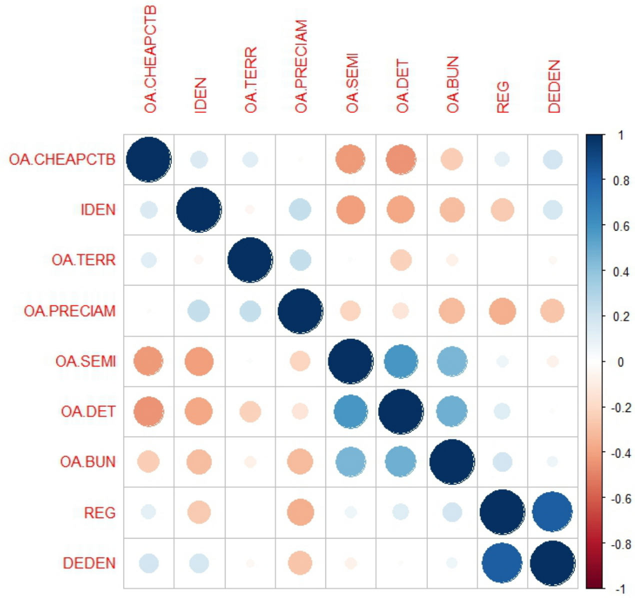

Figure 2: **Cross-correlation matrix.**

6.3 Logistic regression

We input our variables in a step-wise logistic regression and obtain a parsimonious model of seven variables (Table 2) with an overall McFadden's pseudo R^2 of 0.24, indicating good model fit. All regression coefficients showed high level of significance (i.e., p-value ≤ 0.001) with the most relevant ones being *OA of dwellings with cheap council tax bands* (β = 0.82), *OA of semi-detached houses* (β = -0.44), and *OA of detached houses* (β = -0.36). For what concerns the variables related to the configuration of the built environment, *dead-end density* showed the highest value with a β of 0.26.

In the next section, we interpret these results by providing *(i)* a remark concerning the distribution of housing in London and *(ii)* a narrative of the typical area where gangs are likely to operate based on strength and slope of the regression coefficients.

Table 2: **Logistic regression model. All variables are significant with p-values ≤ 0.001.**

Category	*Indep. variable*	β
Built environment	$DEDEN$	0.26
	$IDEN$	0.21
Housing	$OA-SEMI$	-0.44
	$OA-DET$	-0.36
	$OA-BUN$	-0.26
	$OA-PRECIAM$	-0.29
	$OA-CHEAPCTB$	0.82

McFadden's pseudo $R^2 = 0.24$

6.4 Interpretation

The cross correlation analysis (see Figure 2) highlighted inter dependencies especially among variables related to house types. We observed how the suburban typologies (i.e., bungalow, semi and detached houses) tend to cluster together and exclude the presence of dwellings with cheap council tax bands, which are often flats. Moreover, the same house types seem to be located in areas with low values of street network density. We suggest the following explanations for this neat division. One relates to density. We hypothesize that it is quite unlikely that city areas characterized by high density and flats host suburban dwelling types (e.g., detached houses). The other concerns socio-economic factors: more affluent areas tend to be separated from the less advantaged ones. This observation is backed up by previous works in the fields of sociology [14], urban geography [22], and demography [19].

We now interpret the regression coefficients of the model presented in the previous section (see Table 2). London areas which host active criminal groups seem to be predominantly characterized by the presence of dwellings in the least expensive council tax bands and by the absence of suburban house types (i.e., bungalows, detached and semi detached houses) and historic housing stock (i.e., properties built before the CIAM). Houses associated with the least expensive tax bands tends to represent cheap properties which are usually inhabited by less advantaged people. As we noted before, and as the negative β coefficients of the suburban house types highlight, these kinds of properties tend also to be separated from more affluent areas. We hypothesize that spatial segregation and concentrated poverty might be the causes for having people more prone to join criminal groups and take active roles in their illicit operations. This consideration is in line with a recent work by Kang [15] who found that a specific measure of inequality – poverty concentration – is strongly linked with crime. The other part of the story is told by the configuration of the built environment: regression coefficients indicate that the probability of having gang activity in neighborhoods is linked with concentration of cul-de-sacs and a dense urban fabric (i.e., positive values of *intersection density*). This finding is backed up by Jacobs who argued that, in a dense urban context, a well connected street network would be beneficial against crime as the passage of people would provide the necessary level of natural surveillance [13].

7. CONCLUSIONS

In this paper, we used a quantitative method to study the relationship between urban form and gang activity for Greater London. We did so, by extracting a set of descriptors of the urban environment from a mix of open datasets. Through logistic regression we were able to identify a small number of significant features which we then interpreted to provide a plausible description of the typical London area which is likely to be characterized by the activity of criminal groups. Results seem to point at a dense urban fabric, mostly characterized by dwellings in the least expensive council tax bands, by the presence of cul-de-sacs and absence of suburban house types as well as historic housing stock. Although these findings do not imply any causation, they still might be used by urban planners and administrators to inform the debate on how to design safer neighborhoods.

We ought to acknowledge some limitations for this work. The first concerns the accuracy of the geodataset of gang activity. Although the information was collected from a reliable source (i.e., an ex-sergeant of the London Metropolitan Police) and published on a trustworthy British newspaper (i.e., the Independent), the methodology used to gather data (i.e., interview with a single person) may well introduce small inaccuracies which may affect results.

Another limitation lies in the methodology used to assign to LSOAs presence or absence of gang activity. As we illustrated in the Method section, our technique assigns a true flag (i.e., presence of gang activity) to a LSOA if the majority of its surface (i.e., more than 50%) is overlapped by any gang territory. This threshold was adopted as it seemed to be the most intuitive one; however, other thresholds could have been experimented and results might have slightly changed. We argue, though, that this limitation is mitigated as only very few LSOAs are partially overlapped by gang territories.

A further limitation concerns the link proposed by the authors between properties in the least expensive tax bands and cheap houses. Since tax bands in England were assigned to properties by considering their 1991 values,[9] it is plausible that some prices have significantly increased over time and that, therefore, their council tax bands do not reflect their respective values anymore. We argue, though, that this might hold for a minority of houses mainly located in very economically active neighborhoods (e.g., Shoreditch in East London).

The last limitation is that findings apply to Greater London only. As future work, we plan to apply the very same methodology used in this work to other urban areas, and to then compare findings so to be able to generalize, if those were found to be consistent.

8. REFERENCES

[1] Banister, D., Watson, S., and Wood, C. Sustainable cities: transport, energy, and urban form. *Environment and Planning B*, 24:125–144, 1997.

[2] Budd, T. Burglary of domestic dwellings: findings from the british crime survey. 1999.

[3] Buonanno, P. and Leonida, L. Education and crime: evidence from Italian regions. *Applied Economics Letters*, 13(11):709–713, 2006.

[4] Chilton, R. J. and Markle, G. E. Family disruption, delinquent conduct and the effect of subclassification. *American Sociological Review*, pages 93–99, 1972.

[5] Cliff, A. D. and Ord, J. K. *Spatial processes: models & applications*, volume 44. Pion London, 1981.

[6] Corbusier, L. *Manière de penser l'urbanisme: urbanisme des CIAM*, volume 1. Éditions de l'architecture d'aujourd'hui, 1946.

[7] Derksen, S. and Keselman, H. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265–282, 1992.

[8] Harries, K. Property crimes and violence in United States: an analysis of the influence of population density. *International Journal of Criminal Justice Sciences*, 1(2):24–34, 2006.

[9] Hillier, B. *Space is the Machine: A Configurational Theory of Architecture*. Cambridge University Press, Cambridge, UK, 1996.

[10] Hillier, B. Can streets be made safe? *Urban Design International*, 9(1):31–45, 2004.

[11] Hillier, B. and Sahbaz, O. An evidence based approach to crime and urban design, or, can we have vitality, sustainability and security all at once. *R. Cooper, G. Evans, C. Boyko, C.(eds.), Designing Sustainable Cities: Decision-making Tools and Resources for Design, Wiley Blackwell*, pages 163–186, 2009.

[12] Hsieh, C.-C. and Pugh, M. D. Poverty, income inequality, and violent crime: a meta-analysis of recent aggregate data studies. *Criminal Justice Review*, 18(2):182–202, 1993.

[13] Jacobs, J. *The life and death of great American cities*. Random House, New York, 1961.

[14] Jargowsky, P. A. Take the money and run: economic segregation in US metropolitan areas. *American sociological review*, pages 984–998, 1996.

[15] Kang, S. Inequality and crime revisited: effects of local inequality and economic segregation on crime. *Journal of Population Economics*, pages 1–34, 2015.

[16] Li, J. and Rainwater, J. The real picture of land-use density and crime: a gis application. ESRI, May 2006.

[17] Lochner, L. and Moretti, E. The effect of education on crime: evidence from prison inmates, arrests, and self-reports. Technical report, National Bureau of Economic Research, 2001.

[18] Marshall, W. and Garrick, N. The shape of sustainable street networks for neighborhoods and cities. In *Congress for the New Urbanism XVII, Denver, CO*, 2009.

[19] Massey, D. S. The age of extremes: concentrated affluence and poverty in the twenty-first century. *Demography*, 33(4):395–412, 1996.

[20] McFadden, D. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, pages 105–142. Academic Press, 1974.

[21] McFadden, D. Quantitative methods for analyzing travel behaviour of individuals: some recent developments. In *Bahvioural Travel Modelling*. Croom Helm London, 1979.

[22] Musterd, S. and Ostendorf, W. *Urban segregation and the welfare state: inequality and exclusion in western cities*. Routledge, 2013.

[23] Naik, N., Philipoom, J., Raskar, R., and Hidalgo, C. Streetscore–predicting the perceived safety of one million streetscapes. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 793–799. IEEE, 2014.

[24] Newman, O. *Defensible space: crime prevention through urban design*. Collier books, New York, 1973.

[25] Porta, S., Crucitti, P., and Latora, V. The network analysis of urban streets: a primal approach. *ENVIRONMENT AND PLANNING B PLANNING AND DESIGN*, 33(5):705, 2006.

[26] Quercia, D. and Saez, D. Mining urban deprivation from foursquare: implicit crowdsourcing of city land use. *Pervasive Computing, IEEE*, 13(2):30–36, 2014.

[27] Sampson, R. J. and Groves, W. B. Community structure and crime: testing social-disorganization theory. *American journal of sociology*, pages 774–802, 1989.

[28] Sociales, P. Crime as a social cost of poverty and inequality: a review focusing on developing countries. *Facets of Globalization*, page 171, 2001.

[29] Tobler, W. R. A computer movie simulating urban growth in the detroit region. *Economic geography*, pages 234–240, 1970.

[30] Traunmueller, M., Quattrone, G., and Capra, L. Mining mobile phone data to investigate urban crime theories at scale. In *Social Informatics*, pages 396–411. Springer, 2014.

[31] Venerandi, A., Quattrone, G., Capra, L., Quercia, D., and Saez-Trumper, D. Measuring urban deprivation from user generated content. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 254–264. ACM, 2015.

[9] https://www.gov.uk/guidance/understand-how-council-tax-bands-are-assessed