Measuring ambient population from location-based social networks to describe urban crime

Cristina Kadar[*], Raquel Rosés Brüngger, and Irena Pletikosa

Information Management Chair
D-MTEC, ETH Zurich, Switzerland
{ckadar@ethz.ch,rroses@ethz.ch,ipletikosa@ethz.ch}

Abstract. Recently, a lot of attention has been given to crime prediction, both by the general public and by the research community. Most of the latest work has concentrated on showing the potential of novel data sources like social media, mobile phone data, points of interest, or transportation data for the crime prediction task and researchers have focused mostly on techniques from supervised machine learning to show their predictive potential. Yet, the question remains if indeed this data can be used to better describe urban crime. In this paper, we investigate the potential of data harvested from location-based social networks (specifically Foursquare) to describe urban crime. Towards this end, we apply techniques from spatial econometrics. We show that this data, seen as a measurement for the ambient population of a neighborhood, is able to further describe crime levels in comparison to models built solely on census data, seen as measurement for the resident population of a neighborhood. In an analysis of crime on census tract level in New York City, the total number of incidents can be described by our models with up to $R^2 = 56\%$, while the best model for the different crime subtypes is achieved for larcenies with roughly 67% of the variance explained.

Keywords: urban computing, social computing, computational social science, crime analysis, spatial econometrics, location-based social networks

1 Introduction and Related Work

Many past criminological studies have already highlighted the relationship between urban crime and various characteristics of the *resident population* in the area, like e.g. ethnicity [12], income level [12], and residential stability [26]. The theoretical underpinning of these studies lies in the *social disorganization theory* [24] which links the ecological attributes of a neighborhood to its crime levels. In these empirical studies, scholars have exploited traditional census data to measure the population at risk.

[*] Corresponding author.

Yet, with the advent of internet-enabled mobile devices, citizens have become sensors [16] that produce rich data revealing the intensity and nature of human activity in cities. Specifically, location-based social networks (LBSNs) like Foursquare bridge the physical and digital worlds by allowing their users to share their location when visiting different spots in a city. Such services expose information on the location, time, and nature of the activities their users engage in (like shopping, eating out, commuting, being at home, etc.). Moreover, the users can be seen as exponents of the *ambient population* in an area, a more loyal measure of the population at risk expected in that area at any given time. We argue that characteristics of such data can be integrated in models of urban crime.

The relationship between human dynamics and crime in urban environments has been loosely captured by criminologists under the umbrella of the *routine activity theory* [11] or hypothesized by urban planners in quantitative studies (e.g. *eyes on the street* in [18]). Under the routine activity theory paradigm, crime is seen as occurring when a motivated offender meets an unguarded target at a suitable point in time and space [11]. Even further, Brantingham and Brantingham [7] argue that one common way offenders encounter their targets is through overlapping or shared activity spaces, like the offenders home, work, school, and places of recreation as nodes. Furthermore, the same authors go on in [8] and classify some urban hot spots as *crime attractors* (particular places where strongly motivated offenders are attracted due to the known criminal opportunities, like bar and prostitution areas), and others as *crime generators* (particular areas to which large numbers of people are attracted for reasons unrelated to any particular level of criminal motivation, like shopping and entertainments areas). On the other hand, the visionary author and activist Jane Jacobs postulated in her 1961 book *The Death and Life of Great American Cities*, that higher densities and diversities of people and human activities would act as crime deterrents [18].

Very recently, computational social scientists have started to test such theories at scale in descriptive studies by leveraging human dynamics data such as mobile phone data [30]. The authors find significant negative correlations between crime and the diversity of age and ratio of visitors and positive correlations between crime and the ration of residents in an area, as computed using footfall counts extracted from telecommunication data. Finally, latest literature from the data mining community has concentrated on showing the potential of novel data sources of human dynamics like social media [15], mobile phone data [6], or transportation data [31] in a *crime prediction setup.* They prove that machine learning techniques can achieve competitive predictive scores on features mined from such alternative data sources.

In this work, we focus on showing the novelty of the LBSN data in a *crime description setup.* Shmueli outlays in [29] the three different scenarios in which statistical modeling can be used to develop and test theories: causal explanation, prediction, and description. In a descriptive setup, a model is used for capturing the association between the dependent and the the independent variables, rather than for causal inference or for predictive modeling [29] . In this work, we are applying linear models from spatial econometrics that are able to produce an interpretable model. The contribution of the new factors to the dependent variable is precisely quantified in a multivariate setup (as compared to the correlation analyses in [30]) and tested for statistical significance (as compared to the non-parametric machine learning models in [6], e.g.).

In the remainder of the paper we: present the leveraged datasets and derived factors in Section 2, elaborate on the methods and results of the analysis in Section 3, and summarize the conclusions, implications and limitations in Section 4.

2 Datasets and Factors

In our empirical study, we use data from New York City (NYC), a city that is sufficiently large, diverse, and high-tech savvy to assure a high degree of penetration for location-sharing services.

The dependent variable are incident counts from years 2011 through 2015 at census tract level: $N = 2,167$ census tracts. A census tract is a stable geographical unit for the presentation of statistical data with a population size between 1,200 and 8,000 people, with an optimum size of 4,000[1]. To account for the heterogeneity of the unit of analysis, we include the census tract's **area** as a control variable. In terms of independent variables, we will be crafting two sets of variables: one set describing the resident population and one set describing the ambient population. In trying to keep the variables count low, we will be relying heavily on aggregate metrics like fractions and diversity indexes. Even more, we will attempt to craft suitable counterparts of the established resident population metrics when using proxy data of the ambient population.

2.1 Crime

The full crime dataset was downloaded from the NYC Open Data platform[2] and all incidents from 2011-2015 were mapped to the census tracts of the city. Each incident belongs to one of the following sub-types: grand larceny, robbery, burglary, felony assault, and grand larceny of motor vehicle, rape, or murder. In the following, we will be analyzing the total number of incidents, as well as the specific sub-types of crimes that can be described by population characteristics: grand larcenies, robberies, burglaries and assaults. First, to address the skewed distribution of the count data – see Figure 1 (upper left) for the raw counts –, we apply a log-transformation. This operation will yield the dependent variable in the models.

[1] `https://www.census.gov/geo/reference/gtc/gtc_ct.html`

[2] `https://nycopendata.socrata.com/`

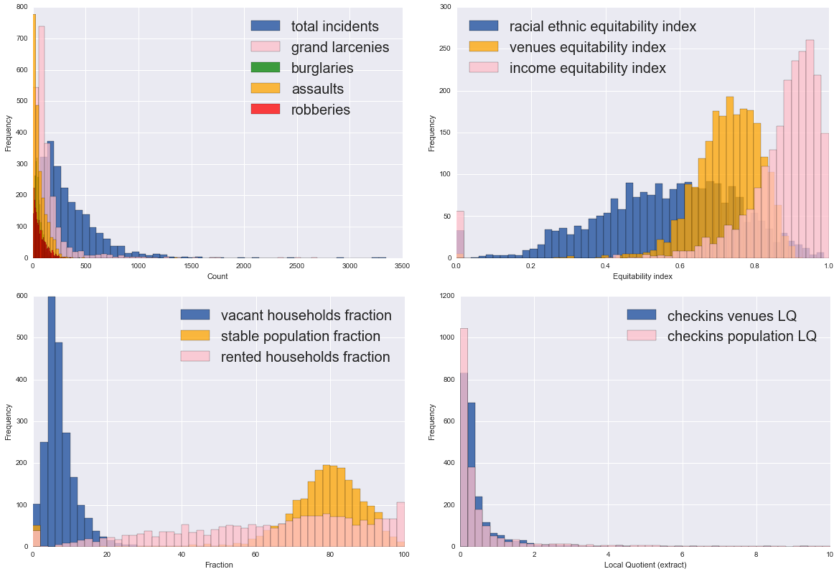

Fig. 1. Histograms of raw variables: crime counts (upper left), equitability indexes (upper right), fractions (lower left), and local quotients (lower right).

2.2 Residential population

The census data for NYC have been derived from the 2010 Decennial Census and the 2010-2014 American Community Survey[3] and it has been employed to derive the crime correlates as per the *social disorganization theory*.

The first variable is the **total population** and it is a measure of the resident population at risk. The population diversity has been shown to play a role in the crime phenomenon [17] so we compute two diversity indexes based on the socio-demographic and economical information: a **racial-ethnic equitability index** and an **income equitability index**. The racial-ethnic index measures the presence of multiple ethnic and racial groups within a certain area and is computed based on five exhaustive and mutually exclusive aggregates (non-Hispanic whites, non-Hispanic blacks, Hispanics of any race, Asians, and others – Native Americans, members of other races, and multi-racial persons), and the income index measures the variance in household income across three main income levels (low, medium, and high-income households). To compute these indexes, we utilize the Shannon diversity index [27], initially developed in information theory, and later used in ecology to summarize the diversity of species [23]. Finally, the Shannon equitability index [28] is simply the Shannon diversity index divided by the maximum diversity, yielding a normalized value

[3] `http://www.census.gov/`

within $[0,1]$:

$$-\sum_{i=1}^{k} p_i \ln p_i / \ln k$$

where p_i is denoting the proportion of the population in group i and k is the total number of groups. Note that lower values indicate the relative abundance of a given group, while higher values indicate equiprobability of all groups. We plot the histograms of all equitability indexes in Figure 1 (upper right). Finally, we calculate the **fraction of vacant households**, the **fraction of rented households** from the occupied ones, and the **fraction of stable population** (individuals who moved in prior to 2010) as measures for the neighborhood instability which has been shown to be associated with violence [26] – see Figure 1 (lower left).

2.3 Ambient population

The factors describing the ambient population were derived based on a dataset collected over the Foursquare API[4] in May-June 2016, consisting of 250,926 venues covering the whole area of NYC, and spanning following 10 broad categories: Arts and Entertainment (11,794 venues), College and University (7,082), Event (84), Food (47,590), Nightlife Spot (11,140), Outdoors and Recreation (18,011), Professional and Other Places (64,055), Residence (14,632), Shop and Service (62,627), Travel and Transport (13,911). These venues have experienced in total almost 122 million checkins since their creation in the app.

The **number of total venues**, **number of total checkins** and in typical **week and weekend afternoons** within a census tract reflect the popularity of that area [5] are all potential metrics of the ambient population at risk, similarly to how the census's total population is a measure of the resident population at risk.

The **venues equitability index**, computed by an analog equitability formula to the one introduced for the residential population, is then a metric capturing the functional decomposition of a neighborhood. Previous work in urban computing has used similar metrics of neighborhood diversity based on LBSN data [19] or on mobile phone data [13, 22], as pioneered in [14]. The higher the equitability index, the more heterogeneous the area is in terms of types of places, and following that, in terms of functions and activities of the neighborhood. On the other hand, a least entropic area would indicate an area with a dominant function. For example, an area dominated by College and University venues would indicate an area where people primarily study, like an university campus.

Finally, inspired by recent work on digital neighborhoods [4], we compute **local quotients** of (digital) social activity within an area as concentrations of checkins relative to the number of businesses and to the reference census population:

$$\frac{1 + C(t_i)}{total\_checkins} \times \frac{total\_venues}{1 + V(t_i)}$$

[4] `https://developer.foursquare.com/`

$$\frac{1 + C(t_i)}{total\_checkins} \times \frac{total\_population}{1 + P(t_i)}$$

where $P(t_i)$ is the total population count within a census tract, $C(t_i)$ is the total number of checkins, and $V(t_i)$ is the total number of venues. Neighborhoods with local quotients >> 1 can be considered (digital) hot spots, while neighborhoods with local quotients << 1 can be considered (digital) deserts. A zoom-in on the $[0, 10]$ interval of the local quotients distributions is plotted in Figure 1 (lower right).

3 Analysis

3.1 Transformations

As seen in Section 2, many of the raw explanatory variables exhibit skewed distributions. Therefore we first transform them towards a normal distribution by using the *Box-Cox method* [9], making them more suitable for linear regression and correlation analysis. As in this work we aim at interpreting and comparing the different regression coefficients, the values of the explanatory variables need to be on the same numerical scale. Towards this end, we apply a second transformation by computing their *z-values* (subtracting the mean μ and normalizing by standard deviation σ).

ID	Factor	Pearson Corr.
0	area	-0.1184^{***}
1	population	$+0.5046^{***}$
2	racial_ethnic_div_index	$+0.1410^{***}$
3	income_div_index	-0.1024^{***}
4	vacant_fraction	$+0.1256^{***}$
5	rented_fraction	$+0.5516^{***}$
6	stable_fraction	-0.1217^{***}
7	venues	$+0.5875^{***}$
8	checkins	$+0.4679^{***}$
9	ven_pop_we_afternoon	$+0.4012^{***}$
10	ven_pop_week_afternoon	$+0.4406^{***}$
11	venues_div_index	$+0.2516^{***}$
12	checkins_venues_lq	$+0.2162^{***}$
13	checkins_population_lq	$+0.1764^{***}$

Table 1. Pearson correlation between all considered factors and the dependent variable: total number of incidents in a census tract (significance levels: $^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$).

3.2 Correlation Analysis

Before we delve into building the explanatory models, we run a series of tests first. First, Table 1 shows the Pearson correlation coefficients between the different factors and the log-transformed number of total crime incidents within a census tract. We observe that all correlations are significant at 0.1%. Furthermore, most of the factors are positively correlated with the crime levels, with the exception of the census tract's area, income diversity index, and fraction of stable population.

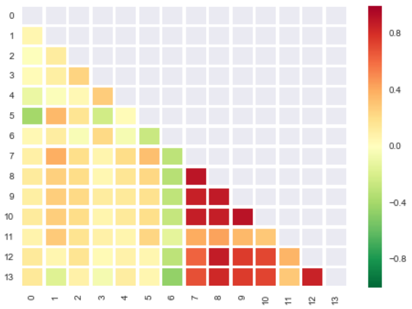

Fig. 2. Correlation matrix of all factors.

ID	Factor	VIF (all factors)	VIF (remaining factors)
0	area	1.4248^*	1.3987^*
1	population	3.4659^*	2.4226^*
2	racial_ethnic_div_index	1.1999^*	1.1599^*
3	income_div_index	1.4420^*	1.3985^*
4	vacant_fraction	1.2127^*	1.1986^*
5	rented_fraction	2.0193^*	1.9871^*
6	stable_fraction	1.4591^*	1.4241^*
7	venues	40.0859	4.5949^*
8	checkins	64.3999	-
9	ven_pop_we_afternoon	9.1538^*	-
10	ven_pop_week_afternoon	8.3502^*	-
11	venues_div_index	1.3938^*	1.3065^*
12	checkins_venues_lq	30.0767	-
13	checkins_population_lq	12.8960	4.0846^*

Table 2. Variance Inflation Factor for all factors and for the remaining factors (* marks an accepted value under 10).

We proceed by looking at the correlation matrix of variables defined above to identify potentially correlated factors – depicted in Figure 2. While multicollinearity does not reduce the reliability of the whole model within the sample set, it is a problem if we are interested in the effects of individual factors on the outcome, since we cannot separate out their individual contributions. As expected from the way they were constructed, the number of venues, checkins, and popular venues are correlated between each other. Also, the two local quotients values are highly correlated. Furthermore, we compute the variance inflation factor (VIF) which quantifies the severity of multicollinearity in an ordinary least squares regression analysis. The lower the VIF value, the better, while an upper limit value of 10 is accepted in the literature [20]. The first column in Table 2 lists the VIF scores considering a specification with all factors. Based on the results from the correlation matrix and the VIF analysis, we keep one variable per group of correlated variables: the number of venues, and the local quotient relative to the population, respectively. We recompute the VIF values in a specification with the remaining 9 factors – see second column in Table 2 –, and conclude that all kept factors have VIF values smaller than 5, which is well below the accepted threshold.

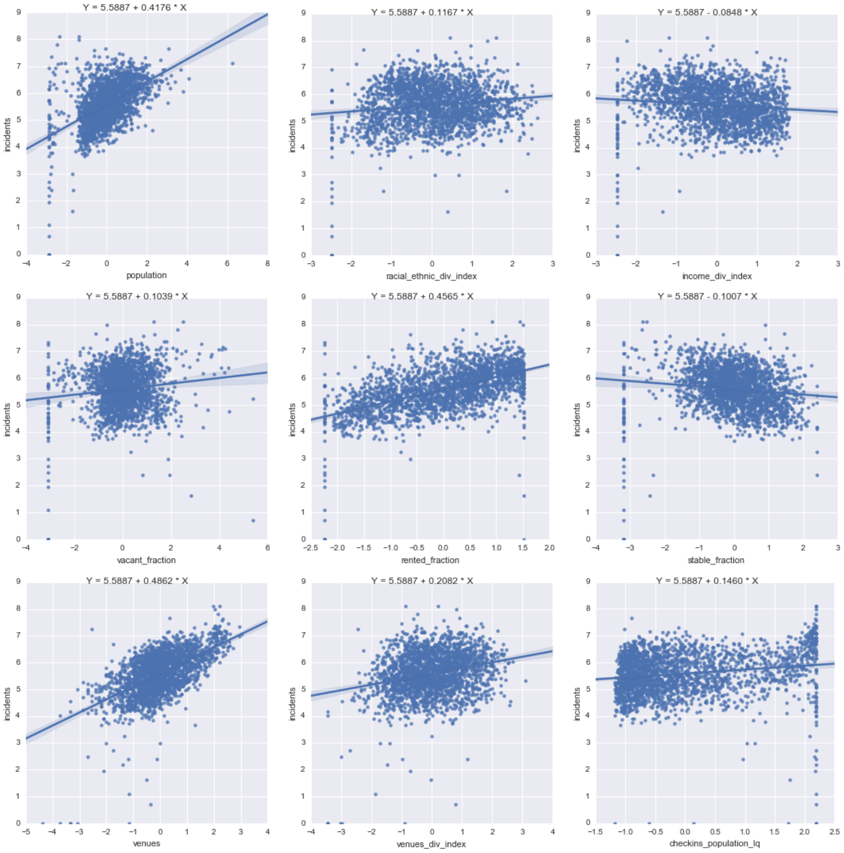

Fig. 3. Total incident counts regressed on the final resident and ambient population metrics. The dependent variable is log-transformed. All independent variables are Box-Cox-transformed and normalized.

Finally, Figure 3 presents the linear relations between the crime counts and the the remaining variables. These initial results support our assumption that specific old and novel attributes of the resident and ambient population of a neighborhood are related to the crime levels. We keep all 9 factors for further analysis.

In addition, we perform a Moran's test [10] to test wether spatial dependencies are present in the crime data. As we obtain a significant global Moran's Index of 0.5552 ($^{***}p \leq 0.001$), we conclude that the spatial distribution of high values and/or low values in the dataset is more spatially clustered than would be expected if underlying spatial processes were random. This result confirms the choice of the spatial lag model. Even more, in addition to the global autocorrelation statistics, we calculate a local indicator of spatial association (LISA) [2], which can help identify and visualize local hot-spots and cold-spots of crime – see Figure 4.

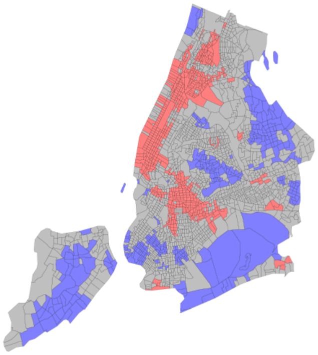

Fig. 4. Crime hot-spots and cold-spots.

3.3 Multivariate Explanatory Models

To explain the relationship between the descriptors of the resident and ambient populations and crime levels, we opt for regression models. In a first attempt, which serves as a benchmark, we build a **linear model**:

$$\boldsymbol{C} = \alpha + \beta_0 \boldsymbol{A} + \beta_1 \boldsymbol{RP} + \beta_2 \boldsymbol{AP} + \epsilon$$

where $\boldsymbol{C}$ is the level of crime in a neighborhood, α is the intercept term, A is the area of the neighborhood, $\boldsymbol{RP}$ is the set of variables relating to the resident population in a neighborhood, $\boldsymbol{AP}$ is a set of variables describing the different characteristics of the ambient population, and ϵ is the error term. The parameters β_0, β_1, and β_2 are capturing the effect of these metrics on the crime levels at an intra-urban level.

This technique requires the independence of the observations, yet the distribution of crime across NYC is likely to have a marked spatial dimension. If so, failing to account for the spatial correlation of the dependent variable in an econometric model leads to a biased model [1]. For this reason, we expand the baseline linear model by including the so-called *spatial lag* – an explanatory variable that captures the values of the dependent variable in the surrounding neighborhood and obtain following **spatial lag model** [1]:

$$\boldsymbol{C} = \alpha + \rho \boldsymbol{W} \times \boldsymbol{C} + \beta_0 \boldsymbol{A} + \beta_1 \boldsymbol{RP} + \beta_2 \boldsymbol{AP} + \epsilon$$

where the new term $\boldsymbol{W}$ is a spatial weights matrix encoding the spatial relationships between the census tracts in the dataset (we use a queens contiguity matrix which considers as neighbors any pair of cells that share a vertex) and ρ is the spatial autoregressive parameter.

Total incidents Table 3 presents the regression results, reporting the R^2 measure for model fit, as well as the value and sign of the coefficients and whether they are significant or not. Specifically, we use the PySAL library in Python [25] and for the linear models estimated using ordinary least squares we report the adjusted R^2, while for the spatial lag models estimated using a generalized spatial two-stage least squares we report the spatial pseudo R^2. Firstly, in Model *(1)* the regression was run with only the area and the independent variables describing the resident population. Secondly, in Model *(2)*, the crime levels are described only in terms of the ambient population variables. Finally, Model *(3)* makes use of the whole set of descriptors.

Although results are comparable across the standard and spatial models some details do change when introducing the spatial effects. Firstly, the size of the significant coefficients is smaller in the spatial model – a known effect of ignoring positive spatial autocorrelation. When properly accounting for the spatial effect, variation is absorbed by the spatial lag term and the other coefficients display more conservative values. The presence of relevant spatial auto-correlation is further confirmed by the significance and large size of the spatial parameter w_incidents (ρ) in all three models.

We now turn at comparing the three model specifications. Overall, in terms of significance, we observe that the racial-ethnic index and the census tract area lose significance when moving from the resident population only model to the full model, while the stable fraction gains significance. The ambient population factors remain significant throughout the models, with the exception of the local quotient of digital activity, which loses significance in some of the specifications. The sign of the significant variables remain stable for the control variable, resident and ambient population variables.

Most importantly, overall, the two models in *(3)* achieve best explanatory performance, with roughly 56% explained variance! This confirms the hypothesis that, the novel factors contribute towards more performant descriptive models of crime. By looking at the spatial-lag formulation of the full model, we interpret that, in the multivariate setup defined previously:

- the population of an area, the fraction of vacant and rented households, the fraction of stable population, the number of venues in the area, and the crime

	(1)		*(2)*		*(3)*	
	Linear	Spatial Lag	Linear	Spatial Lag	Linear	Spatial Lag
Adj./ spatial pseudo R^2	43.90%	43.93%	46.88%	46.91%	56.15%	56.33%
constant	$+5.5887^{***}$	$+2.5020^{***}$	$+5.5887^{***}$	$+3.7853^{***}$	$+5.5887^{***}$	$+3.8228^{***}$
area	$+0.0336^{*}$	$+0.1005^{***}$	-0.1236^{***}	-0.0454^{**}	-0.0271	$+0.0249$
population	$+0.3109^{***}$	$+0.2359^{***}$			$+0.1152^{***}$	$+0.0919^{***}$
racial_ethnic_div_index	$+0.0404^{**}$	$+0.0321^{**}$			$+0.0062$	$+0.0082$
income_div_index	-0.0813^{***}	-0.0240			-0.1084^{***}	-0.0704^{***}
vacant_fraction	$+0.1240^{***}$	$+0.0944^{***}$			$+0.0447^{***}$	$+0.0384^{***}$
rented_fraction	$+0.3245^{***}$	$+0.2036^{***}$			$+0.2355^{***}$	$+0.1809^{***}$
stable_fraction	-0.0269	-0.0205			$+0.0462^{***}$	$+0.0545^{***}$
venues			$+0.7296^{***}$	$+0.6342^{***}$	$+0.5104^{***}$	$+0.4602^{***}$
venues_div_index			$+0.0284^{*}$	$+0.0007$	-0.0174	-0.0355^{**}
checkins_population_lq			-0.3498^{***}	-0.3294^{***}	-0.1832^{***}	-0.1864^{***}
w_incidents		$+0.5478^{***}$		$+0.3201^{***}$		$+0.3134^{***}$

Table 3. Results of the three regression models. Dependent variable: total number of incidents in a census tract. (significance levels: $^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$).

in the surrounding areas are positively and significantly related to the area's crime levels.

- the income diversity of the population in an area, as well as the venues diversity and the level of (digital) social activity in that area are negatively and significantly related to the area's crime levels;
- the racial-ethnic diversity of a neighborhood, as well as the neighborhood area are not significantly related to the overall crime levels in that neighborhood.

In terms of implications, the resident population factors derived from the *social disorganization theory* have therefore found strong support in our empirical results. The one exception is the racial-ethnic diversity of a neighborhood, which is actually a pleasant finding: the expectation is that the model will perform similarly well when leaving out this race-based factor. Most importantly, the activity of the ambient population in terms of its diversity and intensity have been found to be statistically significant and negatively related to the crime counts of a neighborhood – result which is in line with Jacob's *Eyes on the Street* theory. Finally, the other factor derived from LBSN data, the number of venues in a neighborhood is found to be statistically significant and positively related to crime – result which supports the *criminogenic places* theory of Brantingham and Brantingham.

Finally, comparing the actual number of incidents versus the estimated number of incidents we obtain a Pearson's $r = 0.6779$ ($^{***}p \leq 0.001$) and $MSE = 0.2434$ on the whole dataset.

Crime types So far, the analysis investigated the total number of incidents. Yet, some types of crime can be better described by the attributes of the resident and ambient population than others. We proceed by employing model *(3)* for all crime sub-types. Larcenies, as they include all types of thefts, including e.g. pick-pocketing, seem to be best described by the complete set of population factors with a spatial pseudo R^2 of 66.98%. The models for burglaries and assaults are also competitive with metrics of 45.09% and 45.29%, respectively. Finally, the model for robberies managed to explain 34.56% of the variance, while the remaining types of incidents could not be well described by the population factors.

	total incidents	*grand larcenies*	*robberies*	*burglaries*	*assaults*
Adj./ spatial pseudo R^2	56.33%	66.98%	34.56%	45.09%	45.29%
area	0.0249 🟥	0.0178 🟦	0.0604 🟥	0.0265 🟥	0.1168 🟥
population	0.0919 🟥	0.1150 🟥	0.0165 🟥	0.1054 🟥	0.0722 🟥
racial_ethnic_div_index	0.0082 🟥	0.0210 🟦	0.0312 🟥	0.0028 🟥	0.0205 🟥
income_div_index	0.0704 🟦	0.0649 🟦	0.0202 🟦	0.0113 🟥	0.0578 🟦
vacant_fraction	0.0384 🟥	0.0496 🟥	0.0375 🟦	0.0796 🟥	0.0266 🟦
rented_fraction	0.1809 🟥	0.1123 🟥	0.1553 🟥	0.1519 🟥	0.2774 🟥
stable_fraction	0.0545 🟥	0.0620 🟥	0.0413 🟥	0.0007 🟦	0.0414 🟥
venues	0.4602 🟥	0.5656 🟥	0.3876 🟥	0.3804 🟥	0.3520 🟥
venues_div_index	0.0355 🟦	0.0328 🟦	0.0706 🟦	0.0406 🟦	0.0703 🟦
checkins_venues_lq	0.1864 🟦	0.0331 🟦	0.2264 🟦	0.2468 🟦	0.2654 🟦
w_incidents	0.3134 🟥	0.1994 🟥	0.7580 🟥	0.0966 🟥	0.6020 🟥

Table 4. Contribution of the different explanatory variables across different crime types (red – positive coefficient, blue – negative coefficient).

Table 4 visualizes the contribution of each factor in the explanatory models of the different crime types. We can observe that, in general, the sign and relative size of the coefficients stays similar across the models, with some notable exceptions. The geographical influence of crime in the neighboring areas plays a much greater role in the case of robberies and assaults. On the other hand, for grand larcenies and burglaries the influence of the surrounding areas is smaller in comparison to the general model of total incidents. For the category of grand larcenies, which include cases of street thefts, the number of venues in the areas has a higher contribution in comparison to the general model, while the other two ambient population factors have lower negative coefficients. For the category of assaults, which include more violent types of crimes, the neighborhood's fraction of rented house units and the neighborhood's area have higher positive coefficients in comparison to the general model.

4 Conclusions and Discussion

In this paper, inspired by literature in criminology and urban computing, we have leveraged a location-based online service to craft a series of factors describing the ambient population and used these as describing factors for urban crime. First, we proved that these novel factors are significantly related to the crime levels in an area. We then built linear and spatial econometric models of crime and concluded that these factors and the geographical influence improve the explanation models based only on factors describing the resident population. Specifically, we found support for both Jacobs *Eyes on the Street* and Brantingham and Brantinghams *criminogenic places* theories, as the diversity and intensity of the ambient population's activities was found to be negatively related to the crime counts, while the number of venues within the neighborhood was positively related to the crime counts. We have repeated the analysis and built models for the various criminal incident types and found that the results hold across the specific models. In the case of street thefts, the positive influence of the number of venues was more substantial, while in the case of more violent types of incidents like assaults and robberies, the geographical influence was found more pronounced.

Limitations of this work reside in the geographical and demographic biases of the users active in such services, as well as in limiting the analysis to one city. Future work will address these points, by including further ubiquitous data sources in the models and by analyzing other urban areas. We ought also to stress the fact, that this is an observational study and not a controlled experiment, therefore the results should be seen as correlation and not causality between the observed variables and the target variable.

Note: We have tested alternative specifications that included further explanatory variables mentioned in the literature, like household diversity index, age diversity index, and venues offering advantages. We do not include these because of space limitations and also because they do not contribute fundamentally new insights or improve R^2 significantly anymore. Furthermore, we believe the true interesting result of this work is the systematic comparison of the two sets of resident and ambient population variables. Details are, however, available from the authors.

Data accessibility: All raw data used in this study is available from the public sources listed in the text (NYC OpenData platform for the crime data, FTP pages of the US Census Bureau for the census data, and Foursquare API for the LBSN data). In case of interest, readers are welcome to contact the authors for the processed data and used source code.

References

1. Anselin, L. *Spatial Econometrics: Methods and Models*, vol. 4 of Studies in Operational Regional Science. Springer Netherlands, Dordrecht, (1988).
2. Anselin, L. Local Indicators of Spatial AssociationLISA. *Geographical Analysis 27*, 2, (1995), pp. 93-115.
3. Anselin, L., Cohen, J., Cook, D., Gorr, W., and Tita, G. Spatial Analyses of Crime.
4. Anselin, L., and Williams, S. Digital Neighborhoods. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, 9175, (2015), 24.
5. Arribas-Bel, D., Kourtit, K., and Nijkamp, P. The sociocultural sources of urban buzz. *Environment and Planning C: Government and Policy*, 34, 1, (2016), pp. 188-204.

6. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A. Once upon a crime: Towards crime prediction from demographics and mobile data. In *ICMI '14* (2014).
7. Brantingham, P. J., and Brantingham, P. L. Nodes, paths, and edges: Consideration on the complexity of crime and the physical environment. *Journal of Environmental Psychology*, 13, (1993), pp. 328.
8. Brantingham, P. L., and Brantingham, P. J. Criminality of place: Crime generators and crime attractors. *European Journal on Criminal Policy and Research*, 3, (1995), pp. 526.
9. Box, G. E. P., and Cox, D. R. An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological) Vol. 26, No. 2 (1964), pp. 211-252.
10. Cliff, A. D. A. D., Ord, J. K., and Cliff, A. D. A. D. *Spatial processes : models & applications.* Pion, 1981.
11. Cohen, L. E., and Felson, M. Social Change and Crime Rate Trends: A Routine Activity Approach. American Sociological Review, 44(4), (1979), pp. 588.
12. Cohen, L. E., Kluegel, J. R., and Land, K. C. Social Inequality and Predatory Criminal Victimization: An Exposition and Test of a Formal Social Inequality and predatory criminal victimization: an exposition and test of a formal theory *American Sociological Review*, 5, (1981), pp. 505-524.
13. De Nadai, M., and Staiano, J., and Larcher, R. and Sebe, N., and Quercia, D. and Lepri, B. The Death and Life of Great Italian Cities In *WWW '16* (2016)
14. Eagle, N., Macy, M., and Claxton, R. Network diversity and economic development *Science*, 328(5981), (2010), pp. 10291031.
15. Gerber, M. S. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems 61*, 1 (2014), pp. 115-125.
16. Goodchild, M. F. Citizens as sensors: The world of volunteered geography, 2007.
17. Graif, C., and Sampson, R. J. Spatial Heterogeneity in the Effects of Immigration and Diversity on Neighborhood Homicide Rates. *Homicide Studies 13*, 3, (2009), pp. 242-260.
18. Jacobs, J. The death and life of great American cities. Vintage Books, 1961.
19. Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., and Mascolo, C. Geo-spotting: Mining Online Location-based Services for Optimal Retail Store Placement. In *KDD '13* (2013)
20. Kutner, M. H., Nachtsheim, C. J., and Neter, J. Applied Linear Regression Models (4th ed.). McGraw-Hill Irwin. (2004)
21. Lee, Barrett A, Iceland, John, and Sharp, Gregory Racial and Ethnic Diversity Goes Local: Charting Change in American Communities Over Three Decades Key Findings (2012)
22. Pappalardo, L., Vanhoof, M., Gabrielli, L., Smoreda, Z., Pedreschi, D., and Giannotti, F. An analytical framework to nowcast well-being using mobile phone data *International Journal of Data Science and Analytics*, 2(12), (2016), pp. 7592.
23. Peet, R. K. The Measurement of Species Diversity. *Annual Review of Ecology and Systematics*, 5(1), (1974), pp. 285307.
24. Pratt, T. C., and Cullen, F. T. Assessing Macro-Level Predictors and Theories of Crime: A Meta-Analysis. *Crime and Justice*, 32, (2005), pp. 373450.
25. Rey, S., and Anselin, L. PySAL: A Python library of spatial analytical methods. In *Handbook of Applied Spatial Analysis*, vol. 37. Springer Berlin Heidelberg, Berlin, Heidelberg, (2009), pp. 175–193.

26. Sampson, R., Rauenbush, S., and Earls, F. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science 277* (1997), 918–924.
27. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, 27, (1948), pp. 379423.
28. Sheldon, A. L. Equitability indices: dependence on the species count *Ecology*, 50(3), (1969), pp. 466467.
29. Shmueli, G To explain or to predict? *Statistical Science*, 25(3), (2010), pp. 289310.
30. Traunmueller, M., Quattrone, G., and Capra, L. Mining mobile phone data to investigate urban crime theories at scale. In *SocInfo '14* (2014).
31. Wang, H., Kifer, D., Graif, C., and Li, Z. Crime Rate Inference with Big Data. In *KDD'16* (2016).