MULTI-FACTORS EVALUATION OF THE IMPACT OF THE STREET-LEVEL ON SEXUAL CRIME OCCURRENCES USING COMPUTER VISION AND BIG DATA: A CASE STUDY OF MANHATTAN (1120)

Yuxuan Cai[1], Hongqian Li[2], Shuyao Dai[3*]

[1] University of California, Berkeley, United States

[2] Tongji University, China

[3] University of Sydney, Australia; shuyao.dai.18@ucl.ac.uk

Abstract. Manhattan, a central US hub for economics and entertainment, consistently battles high crime rates. Despite existing research examining correlations between overall crime, spatial and sociodemographic factors, or urban architecture's influence on specific crimes, there's a dearth of empirical studies integrating street-level urban features and sociodemographic contexts in sexual crime analysis. This study employs computer vision, machine learning, and big data to investigate associations between Manhattan's urban environment and street-level sex crimes. Two initial Ordinary Least Squares (OLS) models examine the distribution of these crimes from macro-urban and micro-environmental perspectives. A Geographically Weighted Regression (GWR) model further explores local sex crime correlations with different spatial-scale variables. The results suggest a model incorporating multi-dimensional microenvironmental characteristics more effectively explains the incidence of sexual crimes.

Keywords: Sexual Violence, Semantic Segmentation, Space Syntax, Least Square Regression (OLS), Geographic Weighted Regression (GWR).

1. Introduction

In various public domains, incidents of sexual harassment and various forms of sexual assault against women and girls is prevalent globally. The city of New York presents a formidable challenge in guaranteeing the safety and security of people in public spaces (Syeda et al., 2020). Building on the theory that variations in environmental features can influence specific types of criminal activity (Taylor and Gottfredson, 1986), numerous studies have drawn connections between the physical urban environment and crime. Elements such as car parks, commercial buildings, street lights, and vegetation coverage (Ye et al., 2018) have been examined, revealing that the characteristics of crime hotspots significantly explain criminal behavior (Law and Chan, 2012).From a traditional sociological perspective, factors indicative of social inequality like demography, foreclosure, poverty, unemployment, educational level, and racial heterogeneity have also been instrumental in investigating the impact on crime (Arnio and Baumer, 2012;

Raphael and Winter-Ebmer, 2001). Since the 1960s, Geographic Information Systems (GIS) have been utilized in numerous crime-related studies. However, recent developments in AI, such as neural networks, machine learning, and random forests, have allowed for a more nuanced understanding of environmental and economic features how to affect crime. Combining these technologies with GIS enables a more detailed discussion of how varying physical environmental features can influence crime (McClendon and Meghanathan, 2015; Alves et al., 2018; Shah et al., 2021).

Inspired by previous studies and based on the hypothesis correlating socioeconomic and physical environmental features at the street scale in Manhattan with sexual crimes, this study adopts a multifactorial analytical framework. Firstly, the results of computer vision (CV) and spatial syntax analyses are used as variables to objectively evaluate the microscopic architectural environment of Manhattan. Secondly, socioeconomic and land use factors are considered to explore the macro environment. Lastly, two OLS models and a GWR model are employed to explore the correlation of these variables with incidents of sexual crime globally and at different local spatial scales, while investigating their spatio-temporal distribution.

2. Literature Review

Previous literature believes that the spatial characteristics of urban environments may directly impact criminal activity(Quick et al., 2018). As such, there is a need to understand their relationship in order to shape safe streets and gender-inclusive communities (Miranda et al., 2020). The recent developments in emerging technologies such as computer vision, machine learning, and big data make it possible to quantify the impact of street-level built environments on crime. Multi-scale analysis of street images and urban street information can effectively quantify the impact of the built environment on crime occurrence (Jing et al., 2021; Khorshidi et al., 2021; Su et al., 2022). However, previous research has concentrated mostly on the specific spatial location or economic features of total crime (He et al., 2022). Or concentrate on the impact of urban physical characteristics on specific crimes (Summers and Johnson, 2017), but little empirical research exists on the street-level social and environmental features and urban spatiotemporal scale associated with sexual crime (Lee et al., 2023).

2.1. Macroscopic - Urban Socioeconomic Factors with Crime

Quantifying socioeconomic features through census data is a fundamental approach to exploring the correlation between relevant variables and crime rates. Numerous studies have reported the impact of macro-scale urban, demographic, socio-economic, and land use factors on crime rates (Socia, 2016; Wang et al., 2019). Open data platforms for various cities and platforms like OpenStreetMap provide data about urban infrastructure distribution and various urban features, enabling the exploration of linear and spatial relationships between features of different regions and the occurrence of crimes (Lee and Lee, 2020).

2.2. Micro-Urban Street Assessment and CPTED Theory

Street feature assessments have traditionally been conducted through designing and distributing surveys to obtain subjective community evaluations. However, these subjective evaluations based on surveys are somewhat limited in their interpretation of human perceptions. Alternatives, such as exploring street emotional features through social media data (Tang et al., 2022), suffer from low accuracy due to data precision and inherent biases in social media data.

In recent years, environmental feature frameworks, developed from criminological theories like Crime Prevention Through Environmental Design (CPTED), have emphasized the impact of spatial factors on criminal behavior and victims. This allows visual and infrastructure elements of public spaces to be better utilized for quantifying the relationship between street features and criminal incidents (He et al., 2022). With the diversification and convenience of open data platforms like Google Street View, remote sensing images and machine learning technologies are used to define street features and assess environmental street characteristics through computer vision (CV) .

Advanced convolutional neural network technologies like Pyramid Scene Parsing Network (PSP-Net) and Mask R-CNN allow for more precise separation and evaluation of environmental features in street view images. In terms of objective street environment evaluation, spatial syntax serves as a concept and analytical method for understanding the morphological logic of urban grids from urban design and architectural perspectives (Hillier, 2007). Combined with urban morphology and socioeconomic factors, it enables a more comprehensive measurement of urban space (Alabi, 2021).

Spatial syntax also provides an integrated framework for describing the shape and structure of environments, allowing for the quantitative formulation and testing of theories about environmental behavior and emotional responses (Franz and Wiener, 2008). The combination of these technologies enables a more comprehensive assessment of the street environment and facilitates the exploration of correlations between urban environments and crime rates.

3. Method and Data

We propose a method that combines different analysis techniques (corresponding to different city scale levels) and successfully unifies them at the street level. This allows the results of different analyses to be compared and studied at the same level. Further correlation studies on these different analytical data allow us to reveal the potential relationship that may exist between criminal activities and urban space from a data science perspective. Based on these results, we further propose strategies and suggestions on how to intervene and propose safer urban space design.

This study primarily utilized multiple open databases, including US Census, POI data, 21,073 Google Street View images, 100,252 total crime reports, and 14,305 sex crime reports. Through techniques such as spatial syntax (Hillier, 2007), Geographic Information Systems (GIS), and the Pyramid Scene Parsing Network (PSPNet) (Zhao et al., 2017), These techniques allow us to study and process data, mainly from multiple dimensions, in both the physical and the social environments. Specifically, guided by the theory of crime prevention through environmental design (CPTED; Jeffery, 1971), this study considers 38 physical and 10 social environmental variables.

In the physical environmental assessment is focusing on using the PSPNet model based on the ADE20K dataset to extract streetscape pixels, performing physical feature classification based on visual index values and pixel percentages, quantifying the results on a street scale using a GIS system, on the side to evaluating street-scale environmental connectivity, integration, mean depth using the spatial syntax.

Part of the social environmental assessment focuses on using the extraction of land use and POI data, as well as American Community Survey (ACS) data, to quantify Manhattan's social environmental characteristics at the street level using GIS, and analyzing the spatial and temporal distribution of crime using NYPD crime report data to build a spatial-temporal cube and analyze distribution patterns of crime at the street level. Finally, based on the aforementioned. This study two Ordinary Least Squares (OLS) models are introduced for the analysis, exploring the distribution patterns of street-level sex crimes from both macro urban features and micro-environmental characteristics perspectives. This provides recommendations for safer city planning and crime intervention. Ultimately, by selecting variables, a Geographically Weighted Regression (GWR) model is constructed to locally explore the correlation between sex crimes and variables at different spatial scales. The fit results of each model are compared to determine the optimal explanatory model. Help to suggesting recommendations for spatial matching of urban resources and policies.

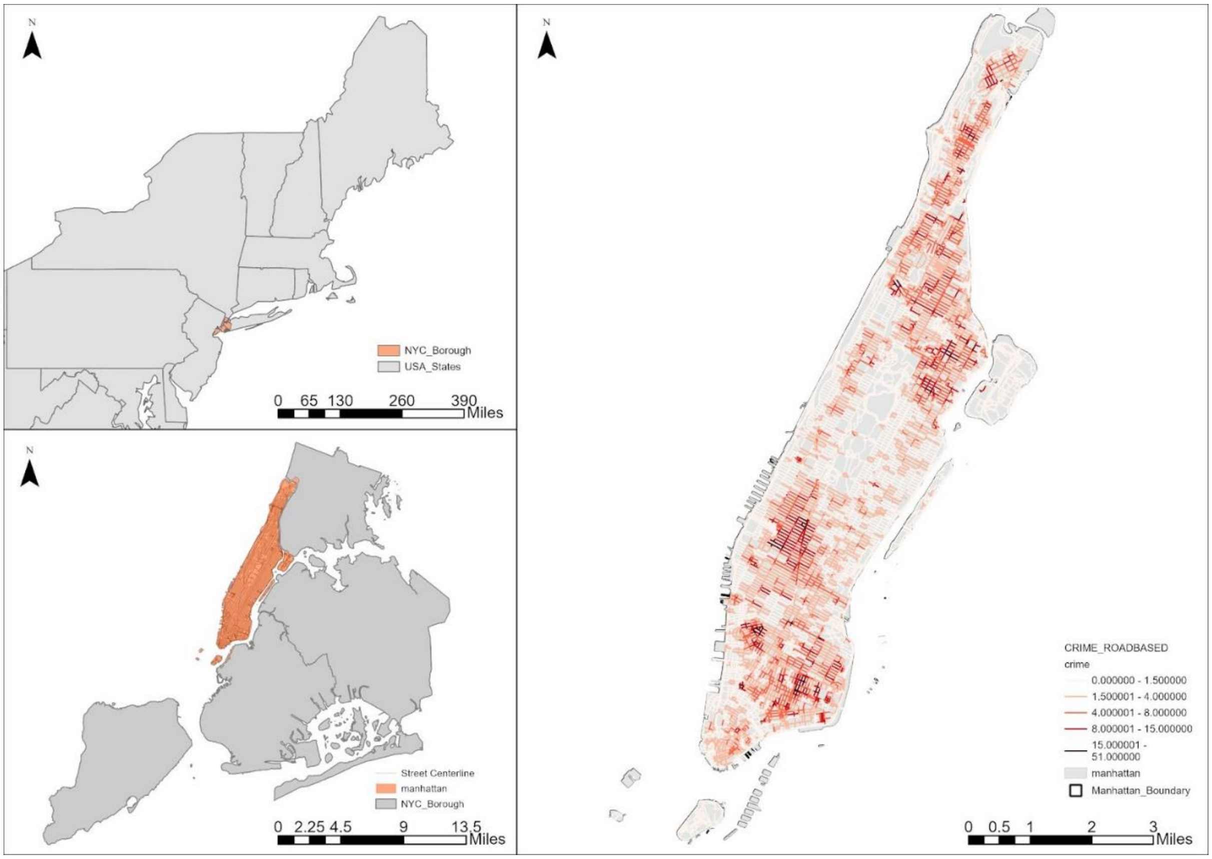

Figure 1. Study area: Manhattan & sexual crime reports

3.1. Study Area

Manhattan, the core urban part of New York City, has a population of approximately 1.6 million people. The city encompasses a variety of land uses, including residential, commercial, and industrial. With its high population density and advanced economic and cultural development, it also has a high crime rate. This study aims to identify correlations between crime occurrence rates and urban features within a multi-factorial framework, centering on specific target locations. The case study area will primarily focus on the urban street level. The data for the entire area comprises polylines of 13,149 street units (Figure 1).

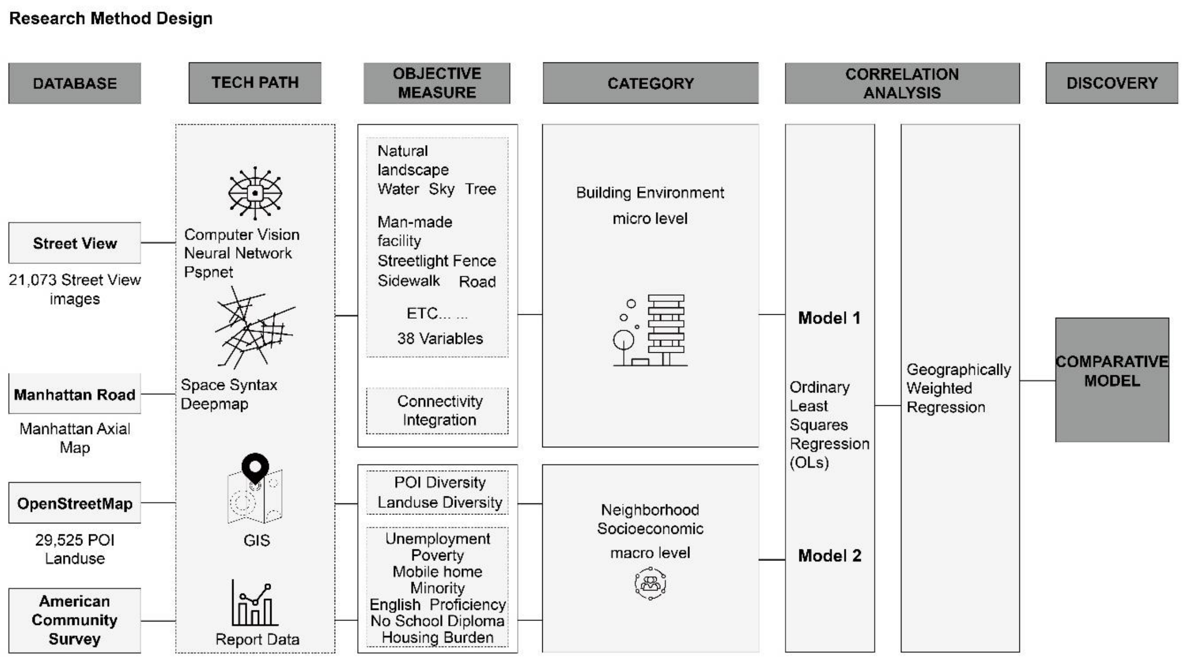

Figure 2. Research Method Design

3.2. Study Design

The research framework for our investigation is delineated in four steps, as illustrated in (Figure 2). This study is designed to address the pivotal issue, exploring the key factors in assessing criminal activities at both the micro level - architectural physical environment, and the macro level - neighborhood socioeconomic characteristics.

In terms of evaluating the architectural physical environment, we first employ SVI, and CV, for objective assessments. The focus lies on extracting streetscape pixels using the PSPNet model based on the ADE20K dataset. The physical features are classified based on visual index values and pixel percentages, segmenting the image content into 38 variables. Furthermore, by employing spatial syntax, two variables related to street-scale environmental connectivity and integration are incorporated into this system for evaluation. Subsequently, on the macro level of Neighborhood Socioeconomic characteristics, the assessment is conducted by selecting seven principal factors related to social vulnerability, encompassing socioeconomic and educational elements, as well as the diversity of points of interest (POIs) as the main variables. Thirdly, these factors are quantified at the street level through the application of The Geographic Information System (GIS). Lastly, using the aforementioned data, this study establishes an Ordinary Least Squares (OLS) regression model, a space-time cube hotspot distribution, as well as a Multiscale Geographically Weighted Regression (GWR) model. These are used to elucidate the relationship between criminal activities in Manhattan and urban characteristics, as well as their spatiotemporal distribution features.

3.3. Dependent Variable

As we delve into the relationship between the environmental backdrop of Manhattan and sexual crimes, a comprehensive dataset was necessitated. Thus, we sifted through the NYPD Complaint Data, procured from the NYC open data platform, isolating 14,305 reports of sexual crimes from a total of 100,252 reports, which were then collated to serve as the dependent variable.

The NYC open data platform is a public resource released by New York City agencies and other partners, that includes NYPD Complaint Data, regularly updated by the NYPD. Detailed within the annual historical crime data reports for Manhattan, from January to December 2022, each crime record is inclusive of the precise date of occurrence, the coordinates of the incident (longitude and latitude details), offense classification code, and level of offense. This invaluable resource proved instrumental in facilitating our study and its objectives.

3.4. Independent Variables

To foster a more nuanced understanding of the dynamics behind sexual crimes in Manhattan, our study proposes a bifocal perspective, dissecting the independent variables into two conceptual frameworks: Microscopic and Macroscopic.

In the Microscopic context, we studied:

1) Street Image Objective Features Measure: With the help of computer vision technology, we segmented and analyzed visual elements within the environment, with the proportional representation of different elements serving as variables.
2) Space Syntax: DeepMap software was used to analyze the structural layout of Manhattan's streets, with selected outcomes used as explanatory variables.

Switching to the Macroscopic lens, our study incorporated:

1) Point of Interest (POI) Diversity & 2) Land-use Diversity: The Shannon Index formula was employed to evaluate the diversity of land use and POI attributes within the buffer zone of the streets.

2) Socioeconomic Characteristics: Drawing from the CDC's Social Vulnerability Index (SVI) 2020 project, itself reliant on the American Community Survey (ACS) five-year estimates from 2016 to 2020, we selected seven facets of socioeconomic characteristics as integral parts of our inquiry.

By employing this dual framework, our study aims to draw a more comprehensive picture of the interplay between environmental factors and the incidence of sexual crimes in Manhattan.

3.4.1. Microscopic - Google Street Image Sampling and Objective Features Measure

Google Street View (GSV) data was crawled from the Google Maps Developer platform (https://developers.google.com/maps). In this study, sample points were generated within the urban range of Manhattan, with each sample point yielding a GSV image. Each image measures 800x400 pixels, with a vertical angle of 0◦ (pitch = 0, fov ='90'). Street View Images (SVI) for each sampling point can be downloaded by providing unique coordinates via the Google Street View (GSV) API and visualized in GIS. Ultimately, 21,073 street-view images of Manhattan were collected. To better segment the proportion of pixels for characteristic visual elements in the images, such as buildings, trees, sidewalks, and other typical elements, the Pyramid Scene Parsing Network (PSPNet) was used. This method automates the collection of this information for visualization, allowing for the classification and distinction of different visual elements in the image, and calculating the percentage of each feature in the total image pixels. This data serves as variables to measure the relationship between spatial features and sexual crimes (Figure 3).

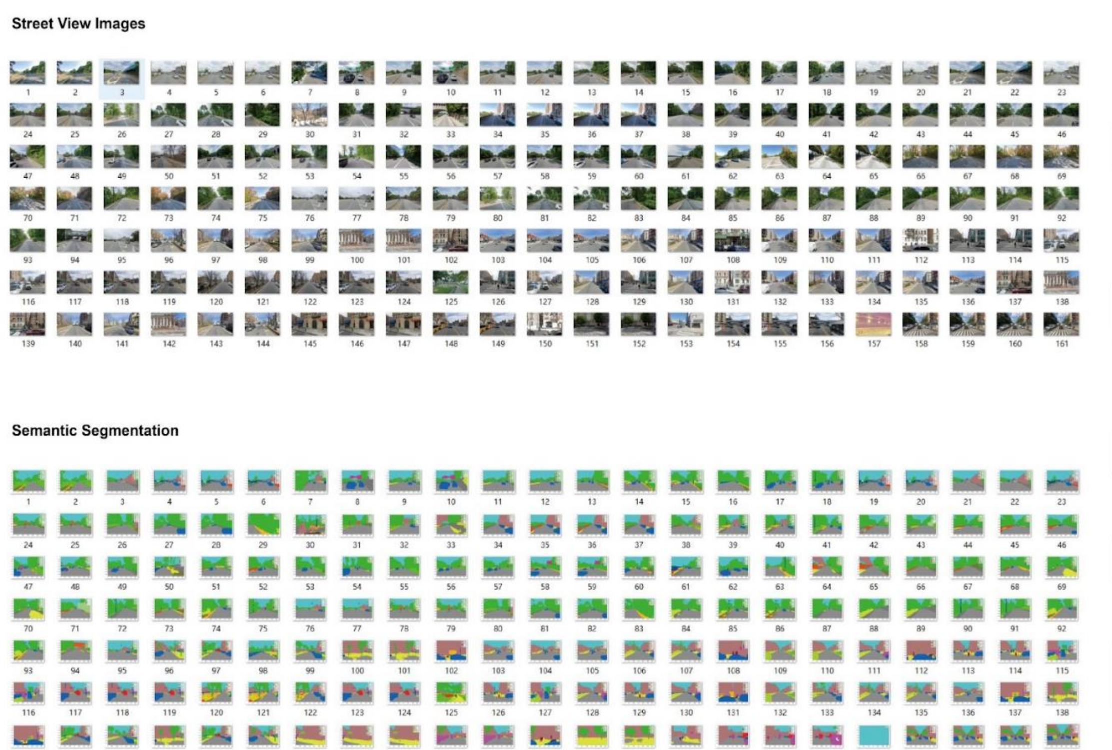

Figure 3. SVI Semantic Segmentation

3.4.2. Microscopic - Space Syntax Features Measure

Space syntax can strictly represent and analyze the urban spatial structure. An analysis of Manhattan streets was conducted using space syntax analysis software, with 'integration' serving as the most effective standard for measuring site accessibility in the results of space syntax analysis, and 'connectivity' is an important indicator for measuring site street connectivity (Talavera-Garcia, 2012). Therefore, 'integration' and 'connectivity' were chosen as two variables for measuring street scale (Figure 4).

3.4.3. Macroscopic - Point of Interest(POI) Diversity

POI attributes are extremely useful for crime prediction, allowing for accurate differentiation between high-risk and low-risk areas (Cichosz, 2020). POI diversity also serves as an important indicator of urban vitality (Tu et al., 2020). Therefore, 29,525 POI data points within the urban range of Manhattan were obtained through the OpenStreetMap platform (https://www.openstreetmap.org/), and the results were classified and mapped at the street level. In the formula, SWD represents the diversity index, 'pi' denotes the proportion of the number of points of interest (POI) of type 'i' within the grid, relative to the total number of POIs of all types. 'm' refers to the total number of unique POI types within the grid.

$$SWD(POI) = -\sum_{i=1}^{m} pi \times ln\, p\, i$$

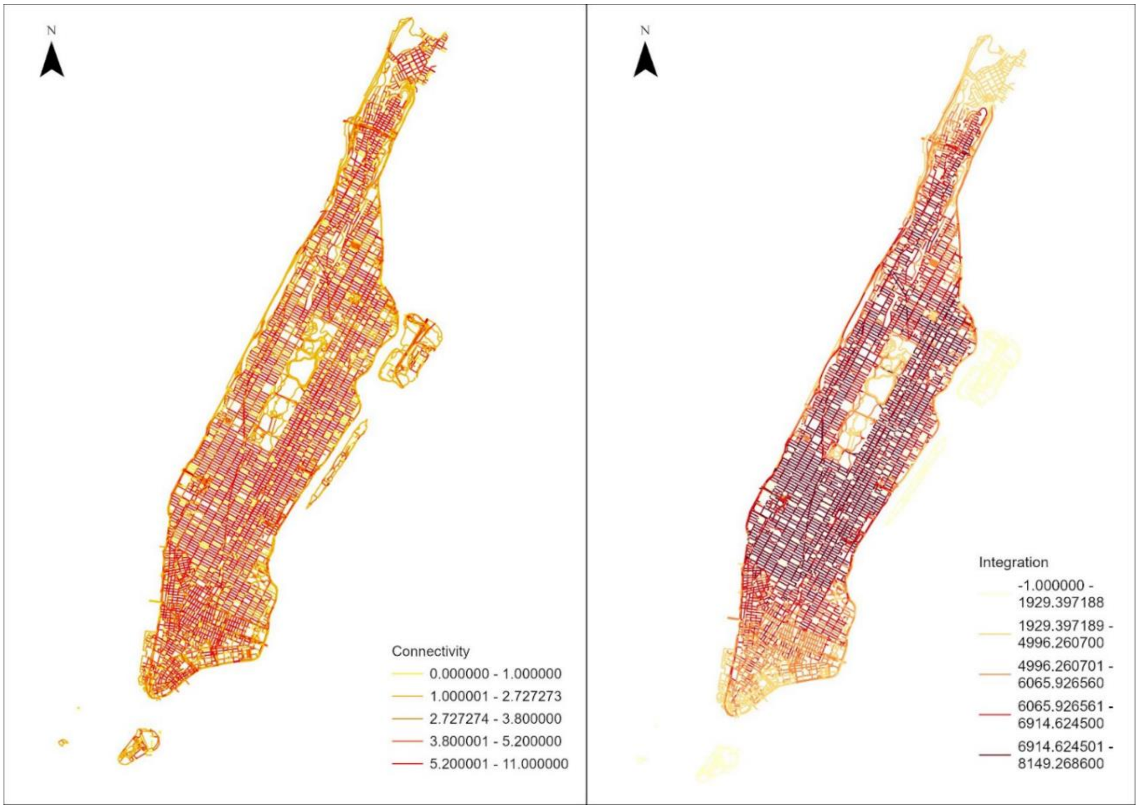

Figure 4. Space syntax distribution

3.3.4. Macroscopic -Socioeconomic Characteristics of Manhattan

Seven variables related to economy, education, and housing were obtained from the Social Vulnerability Index (SVI) database released by the CDC in 2020, serving as the main socioeconomic variables. The SVI encompasses diverse data on social, economic, and residential environments and can effectively predict community crime risk (Polcari et al., 2022).

3.3.5. Macroscopic -Landuse Diversity

In this study, land use data was collected from the MapPLUTO project, a public repository maintained by the New York City Department of City Planning. Our methodology involved constructing a buffer zone extending 100 meters from all roadways in the city, encompassing the variety of land use types within this buffer zone. These types were subsequently quantified as percentages of the total land use within the defined area. To assess diversity in land use, we utilized the Shannon diversity index (SDI), a quantitative measure widely used in environmental and ecological studies to capture the species diversity in a community. Here, 'species' is metaphorically used to refer to different types of land use within our buffer zone.

In the formula for SDI, denoted by SWD, 'pi' represents the proportion of the area occupied by the ith land use type within the grid, and 'm' refers to the total number of distinct land use types. Specifically, SDI is calculated as:

$$\mathrm{SWD(LANDUSE)} = -\sum_{i=1}^{m} \mathrm{pi} \times \ln \mathrm{p\,i}$$

where 'i' ranges from 1 to 'm', and 'pi' represents the proportion of each land use type 'i'. The SDI achieves its maximum value when all land use types cover an equal proportion of the area within the buffer zone. This indicates maximum diversity, or equality in the distribution of different land use types. Conversely, the SDI value becomes zero when only one land use type is present within the area, representing a lack of diversity in land use.

This quantitative analysis of land use in New York City provides a crucial framework for understanding urban dynamics in major cities, with potential implications for urban planning and policy.

Table 1. Primary Variables

Variables	Data Source	Data Description and Extraction	Reference
Macroscopic Sociodemographic Factors			
Poverty	2016-2020 ACS	Persons below 150% poverty estimate	Centers for Disease Control and Prevention/ Agency for Toxic Substances and Disease Registry/
Unemployed		Civilian (age 16+) unemployed estimate	
Housing Cost Burden		Housing cost-burdened occupied housing units with annual income less than $75,000 (30%+ of income spent on housing costs) estimate,	

No High School Diploma		Persons (age 25+) with no high school diploma estimate	Geospatial Research, Analysis, and Services Program. CDC/ATSDR Social Vulnerability Index 2020 Database NYC.
English Proficiency		Persons (age 5+) who speak English "less than well" estimate	
minority		People of color	
mobile home		Mobile homes estimate	
POI diversity	Open Street Map- 29,525 POI	Number of Poi species	openstreetmap.org
Landuse Diversity	Open Street Map	Number of *Landuse* species	openstreetmap.org
Microscopic- Built Environment			
sky, tree, road, grass, person, earth, mountain, car, fence, water, street light, wall, building, plant, signboard, sidewalk, chair, awning, railing, skyscraper, bridge, bicycle, van, ceiling, column, windowpane, ashcan, sculpture, minibike, pier,	21,073 Google Street Image	PSP Net semantic segmentation framework was used in Python.	

fountain, bulletin board, lamp, booth, lake, sofa, desk, glass.			
Connectivity, Integration	NYC Open Data- NYC Street Centerline	Space Syntax	URL:https://data.cityofnewyork.us/d/exjm-f27b?category=City-Government&view_name=NYC-Street-Centerline-CSCL-

4.1. Descriptive Statistics of the Segmentation

We computed the view indices for thirty-nine physical attributes using training images with the aid of a PSPNet pre-trained semantic segmentation algorithm, based on a standard formula. According to the Crime Prevention Through Environmental Design (CPTED) theory, there is a significant correlation between physical environmental factors and urban crime, as shown in Table 2.

Table 2. Descriptive Statistics of the Segmentation

	sky	**tree**	**road**	**grass**	**person**	**earth**	**mountain**	**car**
mean	0.147958	0.141806	0.213156	0.021477	0.005228	0.018091	0.002029	0.074557
std	0.142800	0.103547	0.060962	0.044363	0.008684	0.031900	0.012527	0.045819
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	0.997859	0.647649	0.422518	0.440359	0.164002	0.350594	0.233836	0.259734
	water	**fence**	**streetligh**	**wall**	**building**	**plant**	**signboard**	**sidewalk**
mean	0.007074	0.027355	0.001329	0.042803	0.292240	0.016421	0.004970	0.053154
std	0.032149	0.039137	0.001888	0.067739	0.148743	0.025722	0.005993	0.038776
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	0.442109	0.371488	0.023824	0.838697	0.741828	0.319149	0.136176	0.329730
	chair	**awning**	**railing**	**skyscraper**	**bridge**	**bicycle**	**van**	**ceiling**
mean	0.000439	0.003554	0.011594	0.004227	0.014483	0.002877	0.015880	0.034981
std	0.002633	0.007390	0.018556	0.013693	0.034754	0.008893	0.028185	0.075380

min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	0.047887	0.109641	0.275492	0.151884	0.343323	0.124076	0.357352	0.444954
	column	**windowpane**	**ashcan**	**sculpture**	**minibike**	**pier**	**fountain**	**bulletin b**
mean	0.003199	0.002470	0.001642	0.000310	0.001437	0.000513	0.000499	0.000218
std	0.011466	0.019797	0.003775	0.002535	0.003959	0.003488	0.003759	0.004687
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	0.173672	0.702074	0.058778	0.039414	0.047719	0.142441	0.061766	0.273859
	lamp	**booth**	**lake**	**sofa**	**desk**	**glass**		
mean	0.000007	0.000010	0.000139	0.000007	0.000038	0.000013		
std	0.000095	0.000271	0.003387	0.000215	0.000991	0.000326		
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000		
max	0.002248	0.011605	0.082871	0.008246	0.029301	0.010043		

4.2. Spatial and Temporal Distribution

Utilizing GIS PRO, the study uses the "CMPLNT_FR_DT" field (exact date of the reported event or the starting date of occurrence, if exists) in the sexual crime database as the time range for statistical analysis, aligning with the end time as the Time Step Alignment.

The data was aggregated in square grids of 200 meters on each side at hourly intervals (Time Step Interval).

According to the distribution results, most regions did not exhibit significant trends. However, certain areas such as Downtown Manhattan, Midtown Manhattan, and parts of the Upper West Side showed an increasing trend, with crime rates escalating towards the evening. Contrarily, some areas like parts of the Upper East Side displayed an opposite pattern. Sexual crimes appeared to have some 'sporadic hot spots' in areas like Downtown Manhattan, Midtown Manhattan, and Harlem (Figure 5).

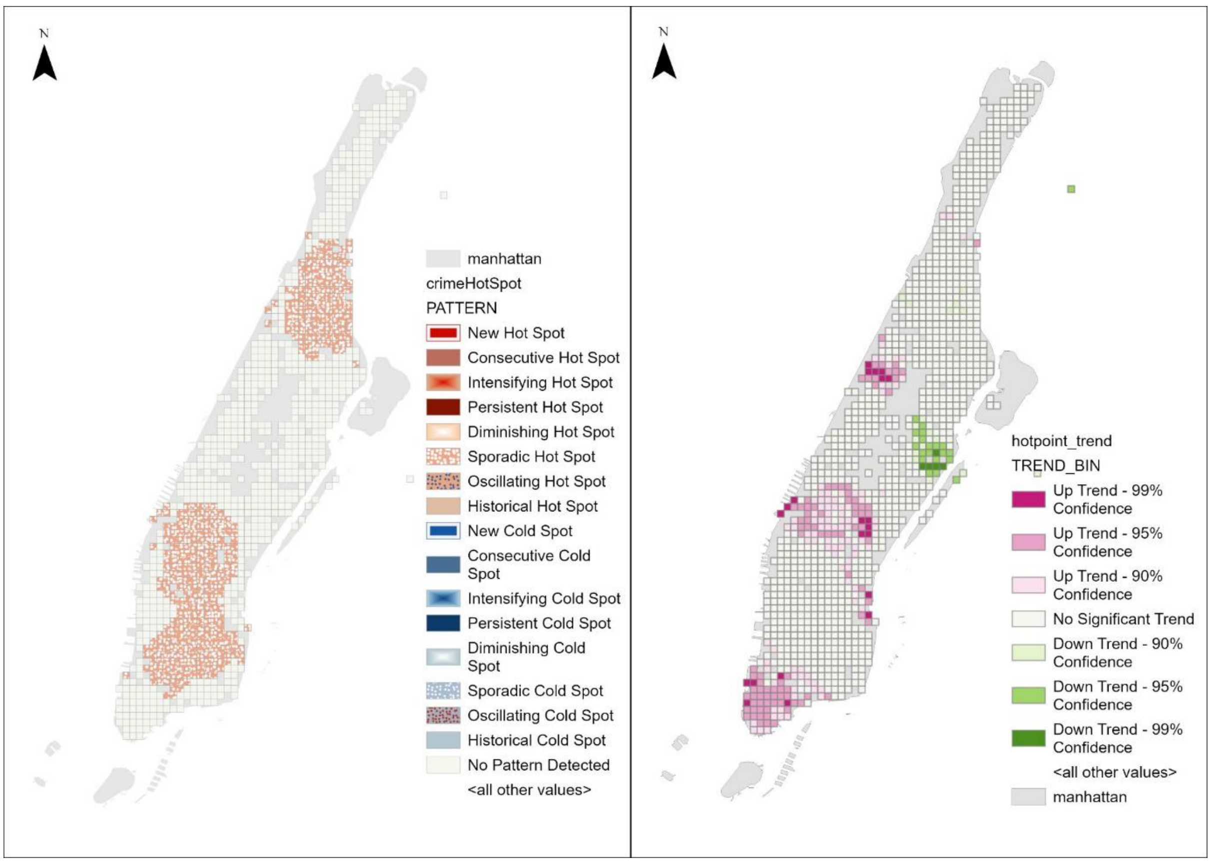

Figure 5. Temporal-spatial hotspot trends

4.3. Correlation Analysis of Independent Variables

To ascertain the presence of any relationships between criminal behavior and other scale variables, an analysis is conducted through the parsons correlation analysis.

4.3.1 Correlation Analysis of Physical Environment Variables *and Sexual Crime Rate*

Based on Parsons correlation analysis, it is evident that the crime rate is significantly negatively correlated ($p<0.01$) with indicators such as sky, tree, grass, earth, mountain, water, fence, streetlight, wall, plant, railing, skyscraper, bridge, ceiling, column, ashcan, and pier. It is also negatively correlated ($p<0.05$) with road and sculpture indicators. This implies that as the presence of sky, tree, grass, earth, mountain, water, fence, streetlight, wall, plant, railing, skyscraper, bridge, ceiling, column, ashcan, pier, road, and sculpture increases, the incidence of crime decreases. The increase in elements associated with the natural environment, such as sky, tree, grass, earth, mountain, and water, leads to a lower crime rate. Similarly, as monitoring elements like fence, streetlight, and wall increase, the occurrence of crime decreases, aligning with the majority of research findings.The crime rate shows a significant positive correlation ($p<0.01$) with indicators such as person, car, building, awning, bicycle, van, minibike, fountain, lake, connectivity, integration, land use diversity, and poi diversity. This suggests that as the presence of person, car, building, awning, bicycle, van, minibike, fountain, lake, connectivity, integration, land use diversity, and poi diversity increases, the crime rate tends to be higher. Some of these correlations align with the conclusions drawn in most previous studies. However, certain results deviate from the findings of previous research. For instance, indicators representing surveillance measures such as person, car, bicycle, van, and minibike contradict the notion proposed by the natural surveillance theory that higher human presence leads to lower crime rates. Furthermore, while it would be expected that an increase in certain natural environmental indicators such as fountain and lake would result in a lower crime rate, this study found that a higher presence of these indicators is associated with a higher incidence of crime. This suggests that sexual crimes primarily concentrate in areas with high population density, complex land use, and high accessibility, where the presence of others does not necessarily prevent sexual offenses. The probability of crime occurrence is comparatively higher in environments such as fountains and lakes.

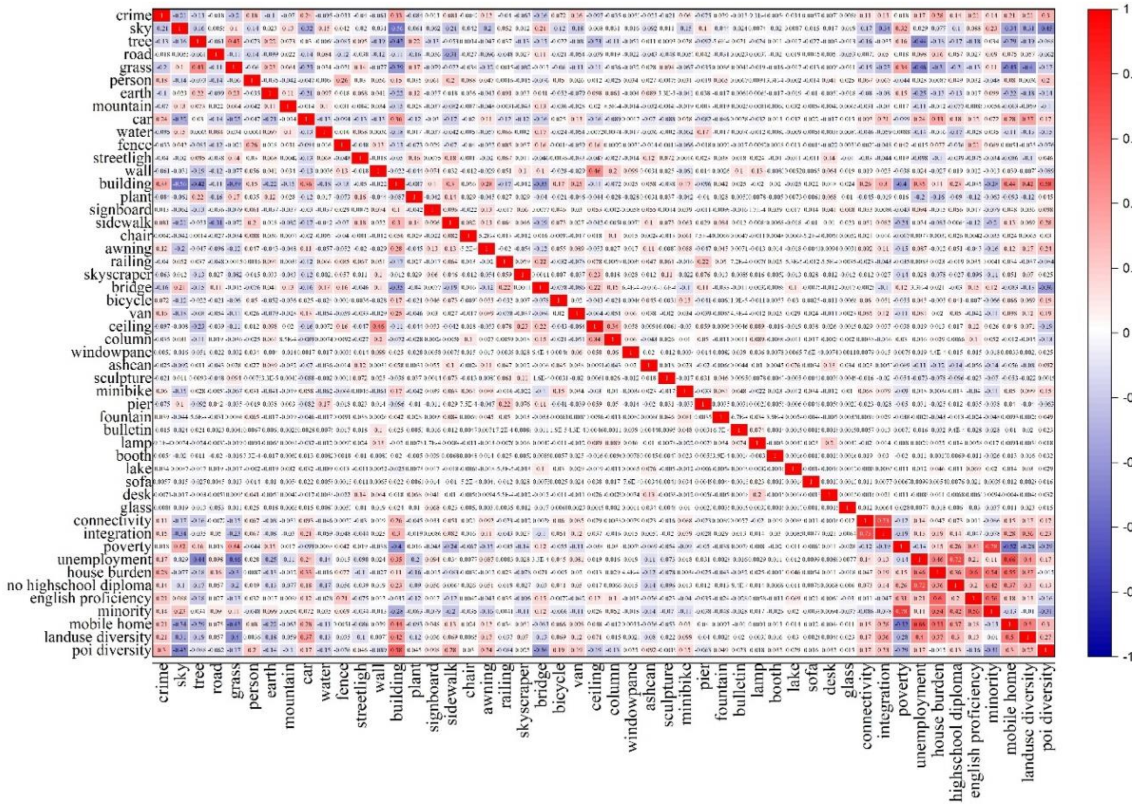

Figure 6. Correlations

4.3.2 Correlation Analysis of Socioeconomic Variable Indicators and Sexual Crime Rate

The crime rate exhibits a significant positive correlation with indicators such as unemployment, house burden, no high school diploma, English proficiency, minority, and mobile home ($p<0.01$). It is also positively correlated with poverty. This indicates that as unemployment, house burden, no high school diploma, English proficiency, minority, mobile home, and poverty increase, the crime rate tends to be higher, aligning with the conclusions of most research papers (Figure 6).

4.4. Regression Results

Two Ordinary Least Squares (OLS) models are instrumental in interpreting the global correlation between macro and micro variables and sexual crime. It provides a broad perspective on the relationship by utilizing data from the entire study area, offering a comprehensive understanding of the interplay between various independent variables and sexual crime on a large scale. However, while the OLS model provides valuable insights at the global level, it may not adequately capture the nuances of local spatial variations and their impact on crime rates. The GWR model complements the global perspective provided by the OLS model by examining the correlation at a local level. It does this by focusing on spatial units within Manhattan, analyzing the local spatial variations and the influence they have on the correlation with sexual crime. This allows for an exploration of how relationships between variables may change in different locations within the same city. These offer a more holistic perspective, allowing us to discuss the correlation between sexual crimes and various variables from both a global and local viewpoint. This combination results in a nuanced and comprehensive understanding of the intricate relationships at play, which can inform more effective strategies for addressing and preventing sexual crimes in different areas.

4.4.1 OLS Regression Result of Physical Built Environment Indicators and Sexual Crime Rate

Through the construction and analysis of an Ordinary Least Squares (OLS) regression model, the following outcomes were obtained (Table 3): Variance Inflation Factor (VIF) values for indicators were less than 7.5, the model's R-squared value was 0.28, and the Akaike's Information Criterion (AIC) was 65525.93. Among the physical built environment indicators, trees, population count, cars, buildings, awnings, railings, trucks, posts, fountains, light fixtures, lakes, integration, and POI diversity demonstrated significant positive correlations with the sexual crime rate. These findings align with the majority of existing literature. However, a divergence was observed regarding the correlation of people and mixed land use, where urbanists commonly perceive mixed land use and density as beneficial for increasing "eyes on the street" and facilitating natural surveillance of the community (Grant, 2002; Jacobs, 1961). This study found that POI diversity and population indicators were positively correlated with crime, suggesting that the surveillance effect of street users might have been overstated (Kitchen & Schneider, 2007).

On the other hand, sky, grass, ground, water, fences, vegetation, chairs, bridges, ceilings, window glass, trash cans, docks, display stands, desks, and glass showed significant negative correlations with crime rates. Most of these outcomes are in line with existing theories. For example, the study found that the window glass index was negatively correlated with the crime rate, suggesting that the presence of windows might deter potential criminals due to the risk of being observed. This perspective is supported by the Crime Prevention Through Environmental Design (CPTED) theory, which proposes that windows provide opportunities for natural surveillance. Likewise, the negative correlation between the fence index and the crime rate supports the CPTED conclusion that fences promote territoriality (Cozens et al., 2005). Natural landscape elements such as grass, sky, ground, and vegetation showed negative correlations with crime rates, potentially due to the calming effects of green spaces reducing anger and impulsive criminal behavior in potential offenders (Kuo & Sullivan, 2001; Markevych et al., 2017).

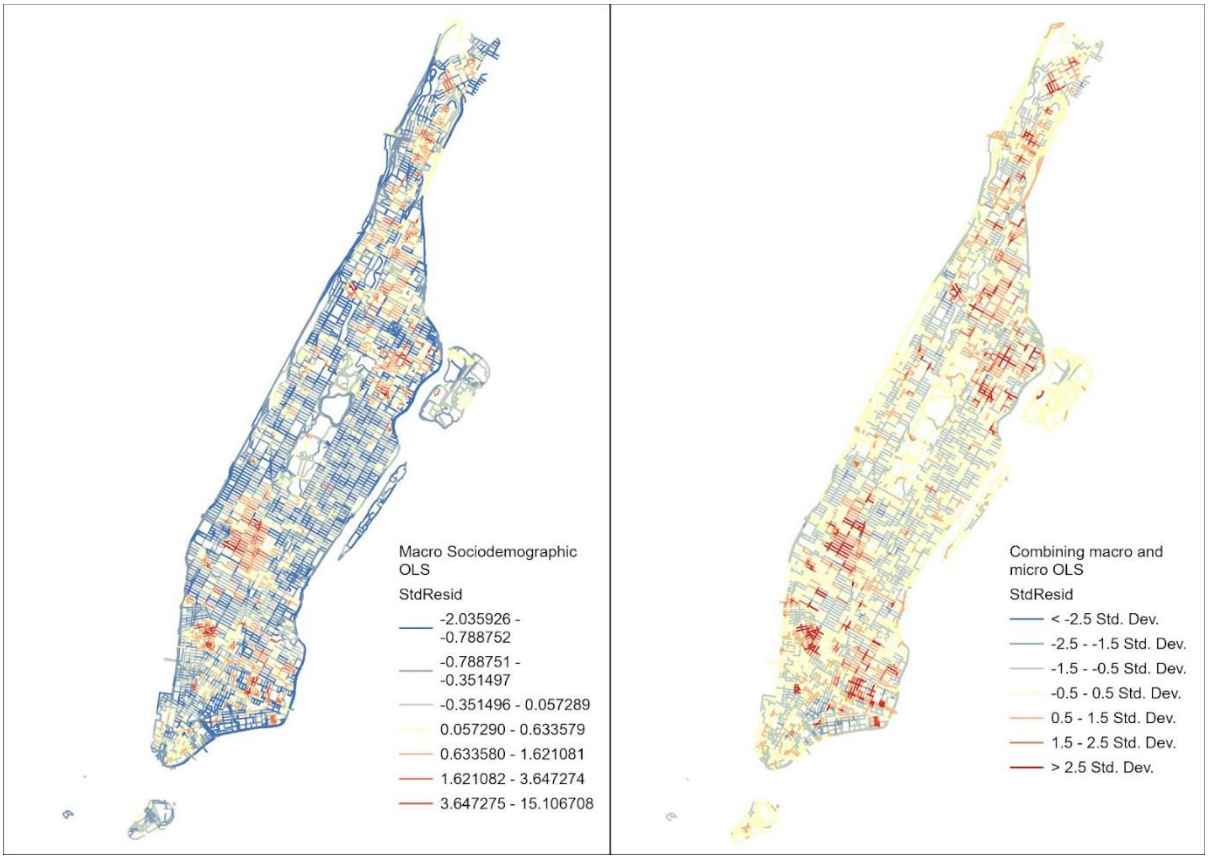

Figure 7. OLS StdResid

4.4.2 OLS Regression Result of Socioeconomic Factors Indicators and Sexual Crime Rate

In socio-economic factors, poverty, unemployment, housing burden, English proficiency, and minority populations demonstrated a significant positive correlation with the crime rate. Among the socio-economic indicators, a high school diploma showed a significant negative correlation with the crime rate, with a regression coefficient of -0.003. This aligns with previous literature endorsing the idea that social instability and structural defects could precipitate criminal activity (Hipp, 2007; Merton, 1938).

Table 3. Combining Macro and Micro Variables OLS

Variable	Coefficient [a]	Standard Deviation	T-statistic	Probability [b]	Robust_SE	Robust_T	Robust_Pr [b]	VIF [c]
intercept	-2.651756	0.321006	-8.260768	0.000000*	0.276944	-9.575073	0.000000*	--------
SKY	-0.764395	0.31616	-2.417752	0.015616*	0.269252	-2.838961	0.004538*	3.146977
TREE	0.794407	0.44803	1.77311	0.076239	0.391417	2.029568	0.042410*	3.322918
ROAD	-0.380168	0.538673	-0.70575	0.48035	0.436894	-0.87016	0.384214	1.66495
GRASS	-5.59241	0.77856	-7.183021	0.000000*	0.536225	-10.429227	0.000000*	1.841808
PERSON	53.911491	3.241885	16.629674	0.000000*	12.942742	4.165384	0.000037*	1.223669
EARTH	-1.587011	0.889425	-1.784311	0.074401	0.720655	-2.202179	0.027654*	1.242874
MOUNTAIN	0.515269	2.126655	0.242291	0.808561	1.316906	0.391272	0.695618	1.095715
CAR	3.318511	0.799273	4.151912	0.000039*	0.738374	4.494348	0.000009*	2.070646

WATER	-2.851314	0.856781	-3.327937	0.000894*	0.541949	-5.261222	0.000000*	1.171425
FENCE	-5.971515	0.746009	-8.004619	0.000000*	1.043064	-5.724976	0.000000*	1.316076
STREETLIGH	1.391336	14.320926	0.097154	0.922588	13.033237	0.106753	0.91497	1.128374
WALL	-0.643753	0.444765	-1.447401	0.147824	0.415351	-1.549901	0.121205	1.401424
BUILDING	4.444403	0.386985	11.484694	0.000000*	0.362907	12.246682	0.000000*	5.115516
PLANT	-5.641156	1.077079	-5.23746	0.000000*	1.065421	-5.294764	0.000000*	1.185053
SIGNBOARD	-9.276641	4.418061	-2.099709	0.035760*	5.16918	-1.794606	0.072744	1.082366
SIDEWALK	-0.18069	0.845521	-0.213703	0.830779	0.854076	-0.211562	0.832449	1.659595
CHAIR	-26.721652	9.927091	-2.691791	0.007114*	8.512782	-3.139003	0.001714*	1.054951
AWNING	14.129977	3.719611	3.798778	0.000157*	4.33014	3.263169	0.001121*	1.166626
RAILING	6.648851	1.495067	4.447191	0.000012*	1.3054	5.093344	0.000001*	1.188267

SKYSCRAPER	-1.921334	2.003828	-0.958832	0.337647	1.533459	-1.252941	0.210255	1.162421
BRIDGE	-3.325639	0.910838	-3.651184	0.000276*	0.645964	-5.148337	0.000001*	1.547105
BICYCLE	0.173516	2.965109	0.058519	0.95332	2.839892	0.061099	0.951265	1.073493
VAN	9.477358	0.949195	9.984625	0.000000*	1.308899	7.240709	0.000000*	1.105024
CEILING	-1.207572	0.447653	-2.697563	0.006992*	0.347941	-3.470621	0.000537*	1.758013
COLUMN	4.467661	2.445834	1.826642	0.067779	1.489489	2.99946	0.002721*	1.214216
WINDOWPANE	-2.9697	1.301632	-2.281521	0.022517*	0.728679	-4.075459	0.000053*	1.025148
ASHCAN	-33.1705	7.163687	-4.630367	0.000005*	5.343961	-6.2071	0.000000*	1.128836
SCULPTURE	12.627669	10.368085	1.217937	0.223273	6.580338	1.919	0.055002	1.066253
MINIBIKE	-0.962461	6.729257	-0.143026	0.886258	7.199626	-0.133682	0.893642	1.095992
PIER	-18.108469	7.807524	-2.319361	0.020374*	4.068016	-4.451425	0.000011*	1.145227

FOUNTAIN	24.142161	6.886839	3.50555	0.000473*	11.719357	2.060024	0.039404*	1.034422
LAMP	25.040477	278.390805	0.089947	0.928313	191.368957	0.130849	0.895882	1.074156
BOOTH	-191.259368	94.328423	-2.02759	0.042612*	45.755077	-4.180069	0.000035*	1.007448
LAKE	20.366474	7.678675	2.652342	0.007999*	6.864852	2.966776	0.003025*	1.044417
SOFA	-61.392527	118.730752	-0.517074	0.605125	52.635641	-1.166368	0.243486	1.008577
DESK	-52.139247	26.873655	-1.940162	0.052376	14.947503	-3.488158	0.000504*	1.095096
GLASS	-59.918615	78.588649	-0.762433	0.445804	22.27023	-2.690525	0.007141*	1.014534
CONNECTIVI	-0.001431	0.02152	-0.066515	0.946952	0.020886	-0.068534	0.945345	2.646905
INTEGRATIO	0.000034	0.000018	1.904941	0.056807	0.000017	1.975222	0.048257*	3.151839
POVERTY	0.007534	0.002633	2.860943	0.004236*	0.002501	3.012429	0.002608*	6.588402
UNEMPLOYME	0.000662	0.00024	2.762894	0.005738*	0.000155	4.260338	0.000025*	5.975327
HOUSE_BURD	0.011806	0.003503	3.370224	0.000770*	0.003381	3.492036	0.000497*	4.562402

NO_HIGH_SC	-0.003156	0.000287	-10.993428	0.000000*	0.000228	-13.840033	0.000000*	4.928765
ENGLISH	0.035307	0.003828	9.223058	0.000000*	0.004455	7.925624	0.000000*	2.252184
MINORITY	0.025275	0.002232	11.324862	0.000000*	0.002078	12.164131	0.000000*	6.641822
MOBILE_HOM	0.000153	0.000202	0.757425	0.448798	0.000132	1.163984	0.24445	5.445775
LANDUSE_Diversity	0.005251	0.068942	0.076166	0.939271	0.068156	0.077044	0.938572	1.800504
POI_Diversity	0.6707	0.037388	17.938823	0.000000*	0.037463	17.902851	0.000000*	2.001219

4.4.3. Geographically Weighted Regression (GWR)

In the occurrence recognized as spatial non-stationarity, a Geographically Weighted Regression (GWR) model is employed to elucidate spatial disparities that global Ordinary Least Squares (OLS) models inadequately explain (Iyanda et al., 2022). The preliminary global dataset consisted of 49 variables. In an effort to prevent multicollinearity complications, variables displaying high Variance Inflation Factors (VIF) and prominent mutual correlations were omitted following correlation and OLS ordering procedures. This step led to the removal of 33 variables, leaving 19 within the macro and micro to operate the GWR model (Table 4).

With regard to the distribution of local R^2, areas in Manhattan where the GWR model demonstrated superior explanatory power include the southwestern region of Downtown Manhattan, Midtown Manhattan, the northern part of Upper East Side, and Harlem. Conversely, the model's explanatory capacity appeared to be weaker in central Downtown Manhattan and the Upper West Side (Figure 8).

Figure 8. GWR Local R^2 Distribution in Manhattan

As per the Geographically Weighted Regression (GWR) results, the regression fit between each variable and sexual crime indicated that primary variables with a higher R^2 included building, Point of Interest (POI) diversity, house cost, and car. Moreover, the distribution of these coefficients broadly corresponds with the overall model R^2 results. These variables provide a more comprehensive explanation for the occurrence of sexual crimes (Figure 9). In the downtown Manhattan area, Points of Interest (POIs), rent proportion, and the number of vehicles show a strong explanatory power. Apart from a region in the west where POIs are highly correlated, the distribution is highly consistent within other ranges. Buildings, however, exhibit weaker explanatory power in this area. Downtown Manhattan is the commercial, cultural, and governmental center of New York City, with a dense distribution of POIs, high rent, and a large number of vehicles required for transportation. These factors make it more likely to correlate with crime. The weaker correlation with buildings may be due to the substantial proportion of parks and green spaces in the central area. Another region with high explanatory power is the eastern part of Harlem, specifically East Harlem. This area has long been plagued by problems such as poverty, crime, and poor housing conditions, making it one of the higher-risk communities. POIs, buildings, and cars show a high explanatory power, and their distribution is roughly the same. The factor of rent proportion, however, primarily exhibits strong explanatory power in the northern part.

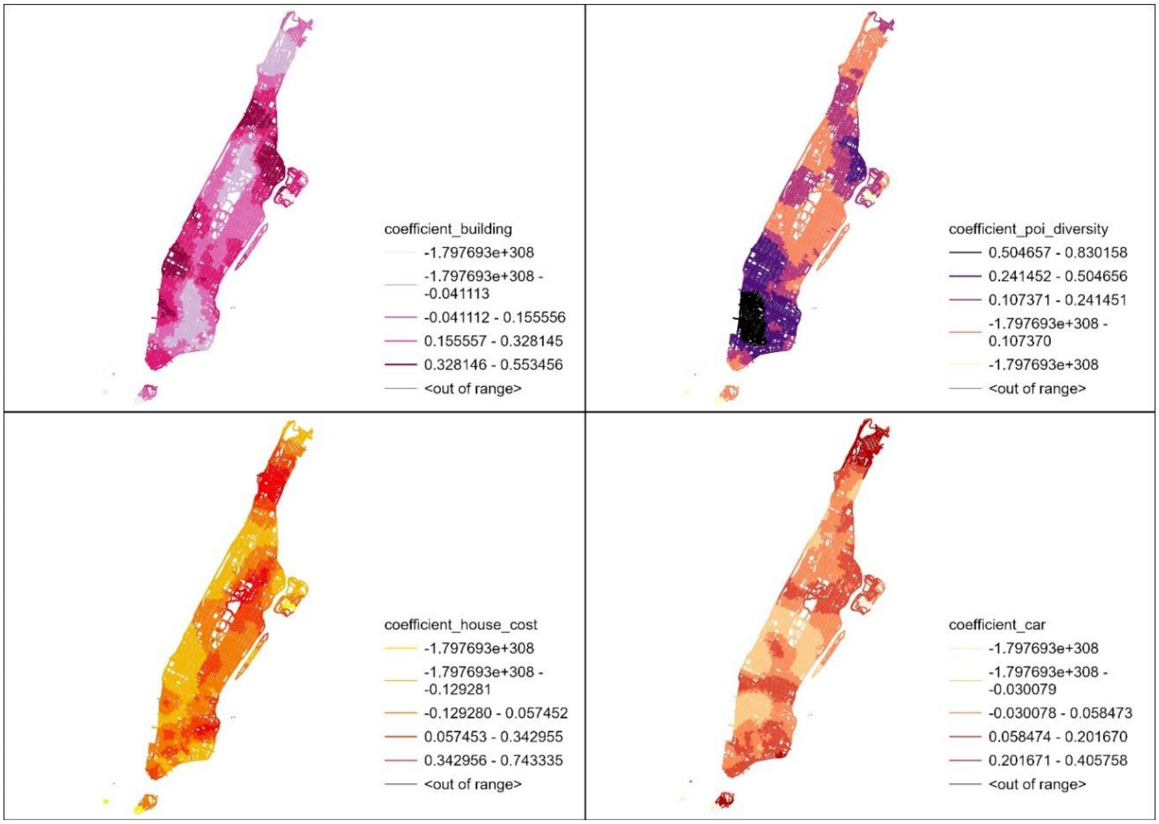

Figure 9. GWR Coefficient Distribution in Manhattan

Table 4. GWR result and variables

Variables	**Parameter**
Residual Squares	1336.949931
Effective Number	186.374001
Sigma	0.570509
AICc	7483.170824
R2	0.543554

R2Adjusted	0.522955
Dependent Variables	
sexual crime	
Independent Variables	
Physical Environment Variables	**Socioeconomic Variable**
Building, person, car, van	Housing Cost Burden
Fence, grass, plant, tree	Poi diversity
Fountain, ashcan, water	Minority
Awning, ceiling, railing	No High School Diploma
	English Proficiency

5. Discussion and Conclusion

The results show that, compared to the model that only considers socioeconomic factors, the OLS model that integrates a variety of microscopic environmental indicators has a decrease of 1066.731728 in the AICc and an increase of 0.079369 in the Adjusted R-Squared value compared to the previous model. The research findings indicate that the latter model, which integrates microscopic environmental variables, offers superior predictive power and explanation for urban sexual crimes, thereby emphasizing the crucial role these variables play in such investigations. Lastly, based on the Geographically Weighted Regression (GWR) model, the occurrence of sexual crimes is explained across different spatial scales. The model exhibits an Adjusted R-Squared of 0.522955 and an Akaike's Information Criterion (AICc) of 7483.170824, demonstrating superior statistical efficiency and explanatory power. This provides valuable insights for policy interventions and resource allocation on a spatial scale. Furthermore, the Geographically Weighted Regression (GWR) model was used to explain the occurrence of sexual crimes across different spatial scales, showcasing an impressive statistical efficiency and explanatory capacity, which offers important implications for spatially focused policy interventions and resource allocation.

Nevertheless, Given the complexity of urban dynamical systems, wherein concurrent dynamic influences occur across temporal and spatial scales between systems, gauging the effects of different factors on sexual crime at the scale of street layouts pose a considerable challenge. Due to data limitations, we could not delve deeper into the correlation between daily temporal variations and macro and microscopic urban characteristics. Also, according to research conducted (Hoffman et al.,2022), there is a significant correlation between reported cases of sexual crimes, income levels, and racial heterogeneity. However, it's important to note that inherent biases may exist within the data itself. We recommend that future research explicates this correlation, considering factors across various spatial scales and basic temporal distributions.

The multifactorial analytical approach employed in this study may offer valuable insights for future criminal analysis research. Additionally, combining analytical methods like spatial syntax and computer vision could supplement and provide powerful tools for urban informatics researchers exploring and addressing urban crime issues. Two comparative multifactorial OLS analytic frameworks are established, investigating the efficacy of an OLS model solely encompassing socioeconomic factors versus one incorporating several dimensions of microscopic environmental indicators in explaining and predicting the incidence of sexual crimes in Manhattan. And a GWR model to explain the crime in the local scale, and finally explore the spatial and temporal distribution. The methods discussed in this study demonstrate significant potential and are expected to offer more comprehensive and accurate insights into understanding and predicting urban sexual crime behavior.

References

Andresen, M.A., 2006a. A spatial analysis of crime in Vancouver, British Columbia: a synthesis of social disorganization and routine activity theory. Canadian Geographies / Les géographies canadiennes 50, 487–502. https://doi.org/10.1111/j.1541-0064.2006.00159.x

Andresen, M.A., 2006b. Crime Measures and the Spatial Analysis of Criminal Activity. The British Journal of Criminology 46, 258–285. https://doi.org/10.1093/bjc/azi054

Ceccato, V., Paz, Y., 2017. Crime in São Paulo's metro system: sexual crimes against women. Crime Prev Community Saf 19, 211–226. https://doi.org/10.1057/s41300-017-0027-2

Costanza, S.E., Bankston, W.B., Shihadeh, E., 2001. Alcohol Availability and Violent Crime Rates: A Spatial Analysis. Journal of Crime and Justice 24, 71–83. https://doi.org/10.1080/0735648X.2001.9721617

Cozens, P., 2013. Crime prevention through environmental design. In Environmental criminology and crime analysis (pp. 175-199). Willan.

Curtis-Ham, S., Walton, D., 2017. Mapping crime harm and priority locations in New

Zealand: A comparison of spatial analysis methods. Applied Geography 86, 245–254. https://doi.org/10.1016/j.apgeog.2017.06.008

Cichosz, P., 2020. Urban Crime Risk Prediction Using Point of Interest Data. ISPRS International Journal of Geo-Information 9, 459. https://doi.org/10.3390/ijgi9070459

Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A., 2016. Deep Learning the City: Quantifying Urban Perception at a Global Scale, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, 计算机科学系列讲义. Springer International Publishing, Cham, pp. 196–212. https://doi.org/10.1007/978-3-319-46448-0_12

Fotheringham, A.S., Yang, W., Kang, W., 2017. Multiscale Geographically Weighted Regression (MGWR). Annals of the American Association of Geographers 107, 1247–1265. https://doi.org/10.1080/24694452.2017.1352480

Grant, J., 2002. Mixed use in theory and practice: Canadian experience with implementing a planning principle. Journal of the American planning association, 68(1), pp.71-84.

He, Z., Wang, Z., Xie, Z., Wu, L., Chen, Z., 2022a. Multiscale analysis of the influence of street built environment on crime occurrence using street-view images. Computers, Environment and Urban Systems 97, 101865. https://doi.org/10.1016/j.compenvurbsys.2022.101865

Hipp, J.R., 2007. Income inequality, race, and place: Does the distribution of race and class within neighborhoods affect crime rates?. Criminology, 45(3), pp.665-697.

Merton, R.K., 1938. Social structure and anomie. American sociological review, 3(5), pp.672-682.

Hipp, J.R., Kim, Y.-A., Wo, J.C., 2021. Micro-Scale, Meso-Scale, Macro-Scale, and Temporal Scale: Comparing the Relative Importance for Robbery Risk in New York City. Justice Quarterly 38, 767–791. https://doi.org/10.1080/07418825.2020.1730423

Hillier, B., 2007. Space is the machine: a configurational theory of architecture, Space Syntax: London, UK. (2007). Space Syntax, London, UK.

Iyanda, A.E., Adeleke, R., Boakye, K.A., Adeusi, T.J., Lu, Y., 2022. Underage tobacco sales violations and neighborhood crime arrest in Philadephia: a multiscale GIS-based analysis. GeoJournal 87, 4455–4474. https://doi.org/10.1007/s10708-021-10507-1

Jeffery, C.R., 1971. Crime Prevention Through Environmental Design. American Behavioral Scientist 14, 598–598. https://doi.org/10.1177/000276427101400409

Jacobs, J., 2016. The death and life of great American cities. Vintage.

Kitchen, T. and Schneider, R.H., 2007. Crime prevention and the built environment. Routledge.

Kim, H.J., Choi, Y., 2016. A Study on the Influence of the Space Syntax and the Urban Characteristics on the Incidence of Crime Using Negative Binomial Regression. KSCE Journal of Civil and Environmental Engineering Research 36, 333–340. https://doi.org/10.12652/Ksce.2016.36.2.0333

Kounadi, O., Ristea, A., Araujo, A., Leitner, M., 2020. A systematic review on spatial crime forecasting. Crime Science 9, 7. https://doi.org/10.1186/s40163-020-00116-7

Lee, D.W., Lee, D.S., 2020. Analysis of Influential Factors of Violent Crimes and Building a Spatial Cluster in South Korea. Appl. Spatial Analysis 13, 759–776. https://doi.org/10.1007/s12061-019-09327-1

Lee, S., Park, S., Kim, J., & Lee, J. (2023). Street environments and crime around low-income and minority schools: Adopting an environmental audit tool to assess crime prevention through environmental design (CPTED). Landscape and Urban Planning, 232, 104676. https://doi.org/10.1016/j.landurbplan.2022.104676

McIlhatton, D., McGreal, W., Taltavul de la Paz, P., Adair, A., 2016. Impact of crime on spatial analysis of house prices: evidence from a UK city. International Journal of Housing Markets and Analysis 9, 627–647. https://doi.org/10.1108/IJHMA-10-2015-0065

Markevych, I., Schoierer, J., Hartig, T., Chudnovsky, A., Hystad, P., Dzhambov, A.M., De Vries, S., Triguero-Mas, M., Brauer, M., Nieuwenhuijsen, M.J. and Lupp, G., 2017. Exploring pathways linking greenspace to health: Theoretical and methodological guidance. Environmental research, 158, pp.301-317.

Miranda, J. V., & van Nes, A. (2020). Sexual violence in the city: Space, gender, and the occurrence of sexual violence in Rotterdam. Sustainability, 12(18), 7609. https://doi.org/10.3390/su12187609

NYC Open Data. (n.d.). https://opendata.cityofnewyork.us/

OpenStreetMap. (n.d.). https://www.openstreetmap.org/about

Polcari, A.M., Slidell, M.B., Nordgren, R., Hoefer, L.E., Zakrison, T.L., Benjamin, A.J., 2022. Social Vulnerability Index Predicts Violent Crime in Chicago. Journal of the American College of Surgeons 235, S296. https://doi.org/10.1097/01.XCS.0000895336.61680.d6

Quick, M., Li, G., Brunton-Smith, I., 2018. Crime-general and crime-specific spatial patterns: A multivariate spatial analysis of four crime types at the small-area scale. Journal of Criminal Justice 58, 22–32. https://doi.org/10.1016/j.jcrimjus.2018.06.003

Sohn, D.-W., 2016. Residential crimes and neighbourhood built environment: Assessing the effectiveness of crime prevention through environmental design (CPTED). Cities 52, 86–93. https://doi.org/10.1016/j.cities.2015.11.023

Socia, K.M., 2016. Examining the Concentration of Registered Sex Offenders in Upstate New York Census Tracts. Crime & Delinquency 62, 748–776. https://doi.org/10.1177/0011128714526563

Su, N., Li, W., Qiu, W., 2022. Measuring The Associations Between Eye-Level Urban Design Quality and On-Street Crime Density Around Subway Entrances in Manhattan. https://doi.org/10.13140/RG.2.2.32289.28002

Summers, L., Johnson, S.D., 2017. Does the Configuration of the Street Network Influence Where Outdoor Serious Violence Takes Place? Using Space Syntax to Test Crime Pattern Theory. J Quant Criminol 33, 397–420. https://doi.org/10.1007/s10940-016-9306-9

Syeda, H. (2020). Public sexual harassment mayhem on public transport in megacities-Karachi and London: A comparative review. Aggression and Violent Behavior, 52, 101420. https://doi.org/10.1016/j.avb.2020.101420

Talavera-Garcia, R., 2012. IMPROVING PEDESTRIAN ACCESSIBILITY TO PUBLIC SPACE THROUGH SPACE SYNTAX ANALYSIS.

Tu, W., Zhu, T., Xia, J., Zhou, Y., Lai, Y., Jiang, J., Li, Q., 2020. Portraying the spatial dynamics of urban vibrancy using multisource urban big data. Computers, Environment and Urban Systems 80, 101428. https://doi.org/10.1016/j.compenvurbsys.2019.101428

Zhanjun, H. E., Li, J., Wang, X., Ma, H., & Chen, J. (2022). Multiscale analysis of the influence of street built environment on crime occurrence using street-view images. Computers, Environment and Urban Systems, 97, 101865. https://doi.org/10.1016/j.compenvurbsys.2022.101865

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. https://doi.org/10.48550/arXiv.1612.01105

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A. and Torralba, A., 2019. Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision, 127,302-321. https://doi.org/10.1007/s11263-018-1140-0