Modeling Crime Rates Through Sociodemographic Similarities

Matteo Roda, Michael Ladaa, Muhammad Bin Arshad, Yash Patel

Bocconi University, Via Roberto Sarfatti, Milan, 20136, MI, Italy

Target Journal: Crime Science.

Contributing authors: matteo.roda@studbocconi.it; michael.ladaa@studbocconi.it; muhammad.arshad2@studbocconi.it; yash.patel2@studbocconi.it;

Abstract

This study examines the relationship between crime rates and sociodemographic factors across London boroughs using a mixed-methods approach that integrates statistical analysis and network-based modeling. Drawing on a two-decade dataset, we construct a similarity network based on variables such as unemployment, income inequality, immigration, population density, and especially disaggregated employment rates of minority and non-minority groups.

To model crime reduction diffusion, we apply a Susceptible-Infected (SI) framework, treating safe boroughs as "infected" and modeling the spread of safety through sociodemographic similarity and geographic proximity. Using grid search and maximum likelihood estimation, we test 220 model configurations to identify the optimal network and contagion parameter ($\boldsymbol{\beta}$).

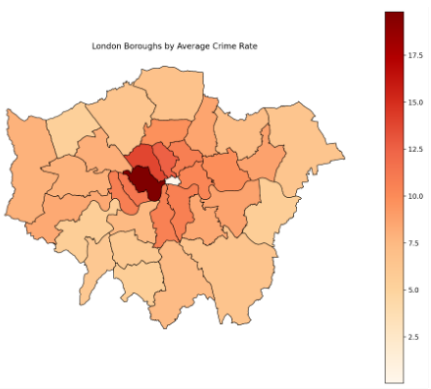
Our findings show that emphasizing minority employment yields the best model fit, with $\boldsymbol{\beta} = \mathbf{0.0158}$ accurately reproducing the shift from dangerous to safe boroughs (2004–2018). This underscores minority employment disparities as key predictors of inter-borough crime dynamics and highlights the need for targeted, sustained interventions to address economic exclusion and promote long-term safety.

Keywords: Crime, Networks, Sociodemographics, Agent Based Modeling

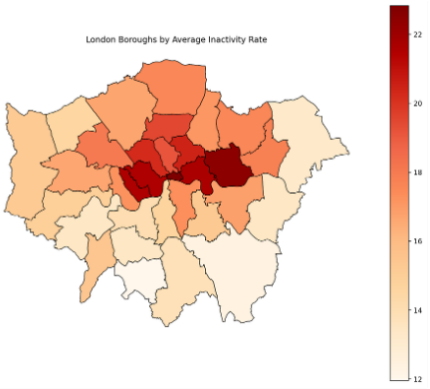
1 Introduction

Crime is a complex social phenomenon influenced by a wide range of economic, cultural, and demographic variables. Understanding the causes and correlations of crime is crucial not only for policy-making but also for creating more resilient and equitable societies. In this paper, we aim to explore the relationship between crime rates and various sociodemographic factors such as unemployment, income inequality, immigration, population density, and urban-suburban distinctions.

We approach this issue through a mixed-method framework, combining empirical data, statistical correlations, and network-based models to uncover potential patterns. As a preliminary insight, we examine data from London, where notable spatial variation in both crime rates and inactivity rates can be observed. Figure 1a shows the distribution of crime rates across boroughs in London, while Figure 1b displays corresponding inactivity rates.
The visual similarities brought us to our research question: To what extent can sociodemographic similarities predict or explain crime?



(a) London Boroughs by Average Crime Rate



(b) London Boroughs by Average Inactivity Rate

Fig. 1: Comparison of Crime and Inactivity Rates across London Boroughs

The structure of this paper is as follows: The next section provides a detailed background and summarizes key existing research on the topic. This is followed by a description of our data collection and preprocessing methods. Subsequently, we outline our modeling approaches, which range from network-based models to agent-based models, along with their parametrization. Finally, we present and analyze the results before concluding with a summary of our key findings.

2 Theoretical Background

Several sociodemographic variables have long been associated with crime rates. One of the most frequently discussed correlations is that between unemployment (or inactivity) and crime. According to the study by Belitto and Coccia [1], there is a positive association between higher unemployment rates and increased criminal activity. In this theory, unemployment is often interpreted as experiencing financial hardship, which leads to criminal behavior as a coping mechanism or as an alternative means of income.

This relationship is especially relevant in urban contexts. In Figure 1b, we present the inactivity rate in London, which mirrors certain geographical patterns observed in Figure 1a with the crime rate. While correlation does not imply causation, such visual and statistical overlaps support the hypothesis that socioeconomic deprivation may be a driving factor behind elevated crime levels. Therefore, another important factor is urban density. Research shows that violent crime rates are higher in urban areas than in suburban or rural ones—48, 37, and 28 incidents per 1,000 people, respectively [2]. Moreover, the nature of urban crime differs: urban crimes are more likely to involve strangers, and population density adds complexity to the social environment. However, this relationship appears to be non-linear, and traditional linear models may not capture the full picture. [2] Therefore, we incorporate agent-based modelling and network analysis to simulate the spatial and social structures of urban environments and their influence on crime rates.

Income inequality is another well-documented predictor of crime. According to de Coccia [3], regions with higher income disparity tend to exhibit higher levels of crime, especially violent offenses. Inequality can fuel social tension and undermine social cohesion, leading to environments where crime is more likely to occur.

Immigration has also been scrutinized in this context. Borjas et al. [4] found that immigration may indirectly influence crime by low wages and reducing employment opportunities, particularly among low-skilled native workers. Specifically, their findings show a negative correlation between immigration and wages, and a positive correlation with incarceration rates. Yet, this perspective is far from conclusive. Other research suggests that the spatial distribution of immigrant communities plays a more nuanced role. Feldmeyer et al. [5] emphasize that while the size of immigrant populations has been widely studied, their geographic segregation within cities may have stronger implications for local violence rates.

This suggests a need to look beyond aggregate numbers and examine how different groups are distributed within urban landscapes. To address this gap, our analysis incorporates spatial clustering and neighbourhood-level data to evaluate how the distribution of demographic groups relates to crime patterns.

In summary, this background highlights the complexity of the relationship between crime and sociodemographic factors. Unemployment, urban density, inequality, and

immigration all interact with one another and with broader systemic factors. Therefore, in the following sections, we aim to combine empirical data with computational modeling to gain deeper insights into these dynamics.

3 Data

Our analysis is based on a multi-source dataset that integrates various sociodemographic and crime-related indicators for all London boroughs over time. We began by obtaining two main datasets from Kaggle[1], which include both yearly and monthly data. The yearly dataset contains borough-level information such as mean and median salary, housing stock, and area size. The monthly dataset includes dynamic indicators such as the average real estate prices, number of houses sold, and monthly crime counts. We merged these datasets based on the borough name and corresponding time variables, allowing us to build a time series at the borough level that combines economic, housing, and criminal activity indicators.

To enable spatial visualization and geographic analysis, we downloaded polygon shapefiles for each London borough from the official London Datastore[2]. These were merged with our existing dataset by borough name, making it possible to plot maps and integrate spatial structure into our models.

Finally, we supplemented our dataset with socio-economic labor market indicators, specifically the Economic Activity Rate, Employment Rate, and Unemployment Rate, disaggregated by both ethnic group and nationality. This data was sourced from the London Datastore as well[3], and provided critical differences on the demographic dimension of economic exclusion and its potential correlation with crime. To harmonize this data with our existing time series, we cleaned and reformatted the raw files, addressing missing values through interpolation and aligning variable names and structures. Incorporating this level of disaggregation allowed us to explore how economic exclusion manifests across different communities and how such disparities may be linked to spatial variations in crime.
The final dataset spans nearly two decades and includes detailed temporal and spatial coverage of London's boroughs, providing a comprehensive basis for our subsequent analysis.

4 Model

4.1 Crime Rate as Outcome Variable

A key methodological decision in our analysis is the use of the crime rate, which is calculated by dividing the total number of crimes per 1000 inhabitants, rather than the absolute number of crimes. This choice is motivated by several considerations. First, using crime rates allows for meaningful comparisons between boroughs of varying

[1] https://www.kaggle.com/datasets/justinas/housing-in-london
[2] https://data.london.gov.uk/dataset/london_boroughs
[3] https://data.london.gov.uk/dataset/economic-activity-rate-employment-rate-and-unemployment-rate-ethnic-group-national

population sizes, thereby controlling for the confounding effect of population density. As highlighted in recent research [6], absolute crime counts can be misleading in urban studies, as larger boroughs naturally report more incidents simply due to their size, obscuring true differences in risk and exposure. By normalizing crime data, we ensure that our analysis reflects relative safety and risk, making our findings both more interpretable and more relevant for policy and intervention (see Appendix A3).

4.2 Network

The construction of our model was strongly informed by the core objective of our research. Aiming to capture the evolution of crime rates within an urban setting, we grounded the connections between London's neighbourhoods primarily on socio-economic similarity. Furthermore, considering the urban reality in which local administrative areas are often strongly interconnected, we also incorporated geographical proximity specifically, the sharing of administrative borders, into our model as a critical determinant of inter-neighbourhood influence. Thus, our model was constructed by integrating both socioeconomic indicators and geographical data as proxies for potential influence among the boroughs.
We divided the time into three periods, Consequently, the periods under consideration were 2001–2006, 2007–2012, and 2013–2018. From this point onward, we describe the operations applied to one of these three datasets, with the understanding that the same procedures were replicated for the other two periods.

The first challenge we faced was transforming the continuous variables we had into discrete clusters, so that we could effectively represent neighbourhood similarity for each of these metrics. The employment rate needed a two-faced approach, separating between minorities and non-minorities. We began by analysing employment rates for white individuals. According to historical data, the UK's overall employment rate has typically remained above 70%, with an employment target of 80% set by the UK government, placing the UK among the highest in Europe[7, 8] . Thus, we defined three clusters: underperforming (<70%), average (70–80%), and overperforming (>80%).

Regarding the clustering based on minority employment rates, we again referred to historical data and relevant thresholds. Values below 50% are relatively rare and often associated with specific ethnic groups [9]. The second cluster included employment levels between 50% and 60%, slightly above what the OECD classifies as a low employment threshold. The third cluster comprised boroughs with minority employment rates between 60% and 70%, consistent with broader national averages. By contrast, the fourth cluster consisted of boroughs whose minority employment levels exceeded 70%, representing a significant outperformance relative to the UK average.
This clustering ensured a balanced representation of both extremes and the average boroughs.
Turning to the inactivity rate, using historical UK inactivity rates (21–23%), we set a 22% threshold to classify boroughs into low (Type 1) and high inactivity (Type 2) [8].
Salary thresholds were determined by accounting for wage growth over time. For the first period, we used £25,000, the overall median salary across boroughs as the

boundary between average and low income. Applying the observed increases—20% from 2001–2006 to 2007–2012 and 5% thereafter, we adjusted the thresholds to £30,000 and £32,000 for the subsequent periods. In each period, a third category captured boroughs with exceptionally high median salaries, considered as outliers.

The outcome of this initial phase was the construction of a similarity matrix (33×33, corresponding to the number of London boroughs) for each of the three time periods under consideration. Each cell represented the socio-economic similarity between two boroughs as a percentage. This led to a need for further research, as we wanted to find the variables, which contribute most significantly to the similarity score.
To avoid arbitrary assumptions, we decided to explore this question empirically by defining 11 distinct configurations to be used as the basis for simulating the contagion process in our SI model (see later sections). First, we included an equal configuration in which all six socio-economic variables were assigned the same weight, implying no dominance of one dimension over another. Next, we constructed six one-hot configurations, each assigning a 100% weight to a single variable while setting all others to zero. These configurations allowed us to isolate the influence of each individual feature.

Finally, we introduced four focus configurations designed to emphasize one variable or group of variables more heavily, without fully excluding the others. In the first focus scenario, the weight of median salary was tripled relative to the other variables. In the second, the inactivity rate was similarly emphasized. The third and fourth focus configurations concentrated on employment-related metrics: one set emphasizing employment rates among minority populations, and the other set focusing on white populations. Since each of these demographic categories included two sub-variables (UK-born and non-UK-born), their weights were doubled rather than tripled, to avoid making the resulting similarity matrices too similar to the corresponding one-hot scenarios.
On top of these socio-economically driven connections, we systematically incorporated geographic adjacency into the network structure. Boroughs sharing borders were always connected in the network, ensuring consistent inclusion of spatial proximity across all configurations. This ensured that spatial proximity, reflecting the intuitive notion that neighbouring areas exert stronger mutual influence, was consistently integrated into our model.

4.3 SI Model

In this phase of the model, our focus shifted to the dependent variable and the development of an SI model capable of accounting for changes in the borough network across the three-time intervals under study. One immediately observable pattern was the general downward trend in crime rates from 2001 to 2018. However, since crime rate is a continuous variable and therefore not directly compatible with a binary-state SI model, we were required to establish a threshold to determine whether, at a given point in time, a borough should be classified as a "dangerous borough" or a "safe borough."
To do so, we examined the most used quantiles of the distribution as well as the

aggregate mean crime rate. For the sake of operational simplicity, we selected the median crime rate as the threshold: boroughs with crime rates above the median were classified as dangerous, while those below were deemed safe.
Upon plotting the total number of safe and dangerous boroughs over time we immediately noticed the noisy nature of the resulting graph, likely due to short-term shocks distorting the true trend. To address this issue, we applied a moving average with a window of 3 data points per year. This yielded a smoother approximation of the underlying dynamics, clearly showing an increase in the number of safe boroughs and a corresponding decrease in dangerous ones (see Figure 2). Given the SI model's reliance on a monotonic infection process, and the observation that the trends between 2001–2004 and 2015–2018 did not conform to the expected dynamics of an SI system, we decided to focus exclusively on the intermediate years, during which the number of safe boroughs consistently increased. The annual figures derived from the smoothed moving average were used as the empirical basis for validating the SI model on the socio-economic similarity network.

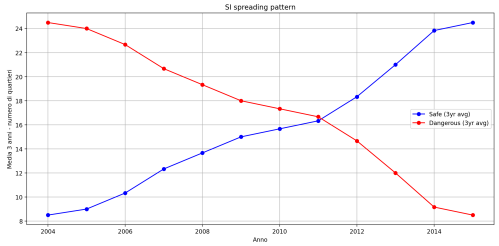


Fig. 2: Spreading Pattern with Moving Average

As for the SI model implementation itself, the main methodological nuance involved handling the network's evolution over time. Specifically, we introduced a temporal control mechanism allowing the model to switch the underlying network at the appropriate year mark. This mechanism preserved the status of each node across network transitions by mapping each node's state from the previous network onto the corresponding node in the new one. Since we adopted an SI modeling framework, the central parameter requiring calibration was the infection rate, denoted as β (beta), which governs the likelihood of a borough transitioning from dangerous to safe due to influence from its connected peers.
In this SI model, the safe boroughs are conceptually treated as the "infected" nodes in the classical epidemiological framework, while the dangerous boroughs represent the "susceptible" population. Accordingly, the infection process corresponds to the spread of positive conditions, thus the adoption or transmission of socio-economic or policy improvements conducive to crime reduction.
In this context, the β parameter, traditionally referred to as the infection rate, can be more appropriately reinterpreted as an imitative factor. It symbolizes a borough's

propensity to adopt safer conditions because of its exposure to neighbouring safe boroughs.

4.4 Parametrization and choice of the Network

With the contagion model and the network structure of boroughs in place, the final task was to identify the combination of network configuration and β value that best replicated the empirical data, namely the 3-year moving average of the number of safe and dangerous boroughs in London.

We began by simulating the model on the equal network configuration, where all variables contributing to the similarity score were equally weighted, using an initial β value of 0.1. This simulation immediately revealed a rapid increase in the number of safe boroughs, which diverged significantly from the observed trend. In reality, the number of safe boroughs rose more gradually, from approximately 8 in 2005 to about 25 by 2018. This discrepancy indicated the need for a more systematic calibration of the model parameters.

To address this, we implemented a grid search over 20 beta values evenly spaced between 0.0 and 0.1, combined with the 11 network configurations previously defined. This resulted in 220 distinct model combinations. For each of these, we conducted 200 independent simulations. The relatively short time span of interest (approximately 10 years) allowed us to perform this computationally intensive procedure without excessive resource demands.

The initial condition for each simulation was fixed at 20% of the boroughs in the "safe" state, reflecting the actual proportion observed in 2005. At each time step, the model simulated the spread of safety through the network based on the chosen similarity configuration and value of β.

To evaluate model performance, we employed the maximum likelihood estimation (MLE) framework. The residuals were computed as the difference between the observed and simulated number of safe boroughs for each year t. Across the 200 simulations, we calculated the mean predicted value at each time point and derived the residuals $r_t = \hat{y}_t - y_t$, where $\hat{y}_t$ is the average simulated value and y_t the observed value.

We then estimated the residual variance as:

$$\sigma^2_{\text{MLE}} = \frac{1}{n} \sum_{t=1}^{n} r_t^2$$

and computed the log-likelihood under the assumption of i.i.d. Gaussian errors:

$$\log \mathcal{L} = -\frac{n}{2} \log(2\pi\sigma^2_{\text{MLE}}) - \frac{1}{2\sigma^2_{\text{MLE}}} \sum_{t=1}^{n} r_t^2$$

This likelihood function penalizes both systematic deviations and high variance in model predictions. By maximizing the log-likelihood over all 220 configurations, we identified the optimal pair of parameters (network structure and β value) that best captured the observed dynamics of crime reduction across London boroughs.

5 Results

5.1 Network Configuration Performance and Model Selection

The systematic evaluation of all 11 network configurations using maximum likelihood estimation revealed that the minority focus network achieved the best model fit. With a log-likelihood of -16, this configuration attained the highest value, outperforming the other variables. This finding indicates that minority employment rates provide the strongest predictive power for similarity patterns between London boroughs, suggesting that employment disparities among minority groups serve as a critical determinant of inter-borough crime dynamics. The network visualizations demonstrate distinct connectivity patterns across different sociodemographic weightings. While the inactivity focus (see Appendix A1) configuration shows dense similarity connections, and the salary focus (see Appendix A2) network exhibits sparser relationships, the minority focus network reveals an optimal balance of connectivity that best captures the underlying crime reduction patterns observed in the empirical data.

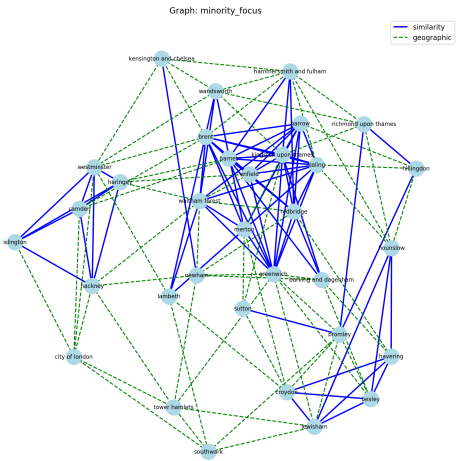


Fig. 3: Minority Network

5.2 Optimal Parameter Estimation

The comprehensive grid search across 220 combinations of network configurations and β values identified $\beta = 0.0158$ as the optimal infection parameter. This parameter combination of the minority focus network with $\beta = 0.0158$ achieved a log-likelihood of -16, representing the best overall model performance. The relatively low β value suggests that safety-promoting conditions spread gradually between similar boroughs rather than through rapid contagion, reflecting realistic timescales for policy implementation and socioeconomic change.

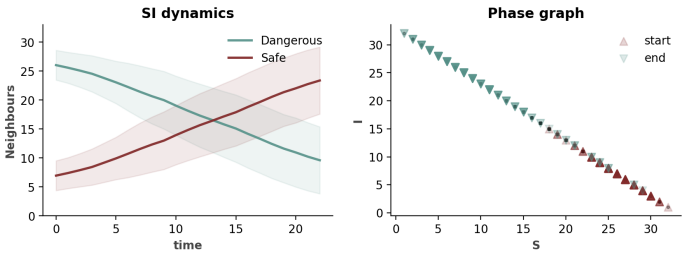


Fig. 4: SI with optimal Beta

5.3 SI Model Dynamics and Empirical Validation

The optimal model successfully replicates the empirical trend of crime reduction across London boroughs from 2005-2018. The SI dynamics panel demonstrates the model's ability to capture the observed decrease in dangerous boroughs (from approximately 26 to 9) and the corresponding increase in safe boroughs (from about 7 to 24). The phase diagram illustrates a systematic transition from a state dominated by dangerous boroughs to one where safe boroughs predominate, with consistent spacing between time points indicating stable model dynamics.

6 Conclusion

In conclusion, our analysis demonstrates that sociodemographic similarities, particularly minority employment rates, are highly predictive of crime rate patterns across London boroughs. By integrating both empirical data and network-based modeling, we found that the minority focus networks most accurately captured the observed dynamics of crime reduction between 2005 and 2018. This aligns with previous research highlighting the significance of employment disparities among minority groups as a key driver of spatial crime variation and inter-borough contagion effects [3].

These findings carry several important implications for policy and future research. First, interventions aimed at improving employment opportunities and reducing economic exclusion among minority populations may yield the most substantial benefits in terms of crime reduction. Policymakers should prioritize targeted labour market programs and anti-discrimination measures in boroughs where minority employment rates lag behind the city average. Additionally, the gradual nature of the observed contagion process suggests that sustained, long-term investment is necessary to achieve meaningful and lasting change. From a methodological perspective, our results underscore the value of network-based models that account for both sociodemographic similarity and geographic proximity, offering a powerful framework for simulating urban social dynamics and informing place-based policy design.

Looking ahead, further research could build on our work by incorporating additional variables such as educational attainment, housing quality, or access to social services, which may also influence crime dynamics. Moreover, extending the analysis to other metropolitan areas could test the generalizability of our findings and refine the model for broader application.

Appendix A Section title of first appendix

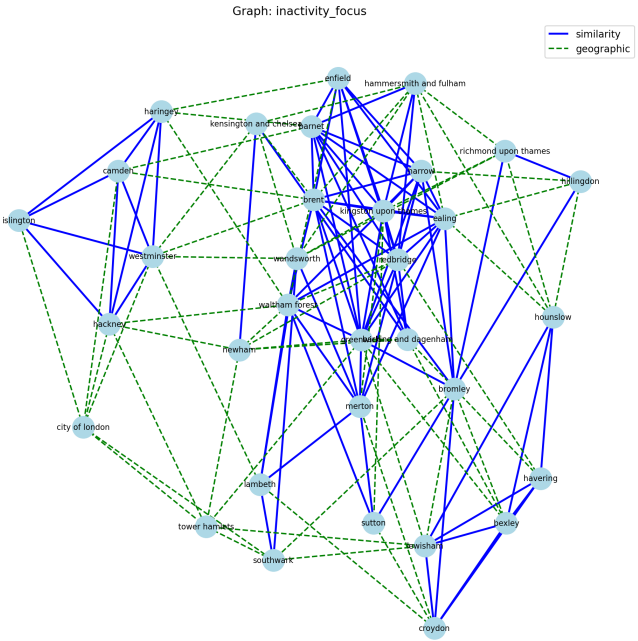


Fig. A1: Inactivity Network

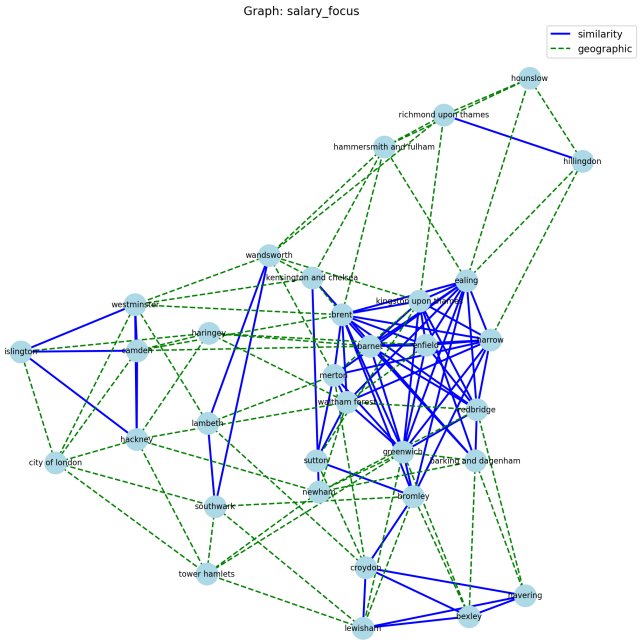


Fig. A2: Salary Network

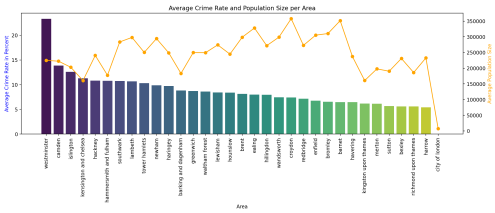


Fig. A3: Average Crime Rate and Population Size per Area

References

[1] Matteo Bellitto and Mario Coccia. "Interrelationships between violent crime, demographic and socioeconomic factors: a preliminary analysis between Central-Northern European countries and Mediterranean countries". In: *Journal of Economic and Social Thought (JEST)* 5.3 (2018). Available at SSRN: https://ssrn.com/abstract=3275310, pp. 230–.

[2] Brian Christens and Paul Speer. “Predicting Violent Crime Using Urban and Suburban Densities”. In: *Behavior and Social Issues* 14 (Mar. 2006), p. 113. DOI: 10.5210/bsi.v14i2.334.

[3] Mario Coccia. “A Theory of general causes of violent crime: Homicides, income inequality and deficiencies of the heat hypothesis and of the model of CLASH”. In: *Aggression and Violent Behavior* 37 (2017), pp. 190–200. ISSN: 1359-1789. DOI: https://doi.org/10.1016/j.avb.2017.10.005. URL: https://www.sciencedirect.com/science/article/pii/S1359178916302105.

[4] George J. Borjas, Jeffrey Grogger, and Gordon H. Hanson. “Immigration and the economic status of African-American men”. In: *The Economic Journal* 120.534 (2010), F77–F92. DOI: 10.1111/j.1468-0335.2009.00803.x.

[5] Ben Feldmeyer, Casey T. Harris, and Jennifer Scroggins. “Enclaves of opportunity or “ghettos of last resort?” Assessing the effects of immigrant segregation on violent crime rates”. In: *Social Science Research* 52 (2015), pp. 1–17. ISSN: 0049-089X. DOI: https://doi.org/10.1016/j.ssresearch.2015.01.003. URL: https://www.sciencedirect.com/science/article/pii/S0049089X15000204.

[6] Yunqi Zhou, Fengwei Wang, and Shijian Zhou. “The Spatial Patterns of the Crime Rate in London and Its Socio-Economic Influence Factors”. In: *Social Sciences* 12.6 (2023). ISSN: 2076-0760. DOI: 10.3390/socsci12060340. URL: https://www.mdpi.com/2076-0760/12/6/340.

[7] Wikipedia contributors. *Tasso di occupazione*. https://it.wikipedia.org/wiki/Tasso_di_occupazione. Abgerufen am 26. Mai 2025. n.d.

[8] Institute for Fiscal Studies. *The government’s 80% employment rate target: Lessons from history and abroad*. https://ifs.org.uk/articles/governments-80-employment-rate-target-lessons-history-and-abroad. Abgerufen am 26. Mai 2025. 2022.

[9] Office for National Statistics. *Diversity in the labour market, England and Wales*. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/articles/diversityinthelabourmarketenglandandwales/cens. Abgerufen am 26. Mai 2025. 2021.