THE UNIVERSITY OF CHICAGO

Mapping Informality and Violence: Machine Learning Insights into Crime Patterns Across South African Police Jurisdictions

By

Yingyi Liang

May, 2025

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Professor Benjamin Lessing

Preceptor: Fabricio Vasselai

Acknowledgements

To my parents, for the longing.

To Ruge, for the wishing.

Abstract

This study examines the relationship between neighborhood informality and crime rates in South Africa, where rapid urbanization has led to criminal organizations stepping in to provide services within some of these settlements. Applying K-means clustering and Local Moran's I analysis to 1,164 police jurisdictions, we observe that the cluster with the highest level of informality also exhibited the lowest crime rates. Furthermore, in clusters representing moderately dense urban areas, linear regression reveals a statistically significant negative correlation between levels of informality and crime rates. Machine learning models, including KNN and random forest, show that crime can be predicted with low MSE, with k-complexity emerging as a key feature. Finally, a longitudinal case study on crime and growth data offers further insight into potential causal relationships, though the results are not statistically significant. These findings have implications for urban governance and public safety policies in developing countries.

Keywords: Urbanization; Crime; Informality; K-means clustering; Machine Learning; Granger Causality;

1 Introduction

The United Nations (2022) has projected that the most rapid urbanization over the next decade will take place in low-income developing countries. However, it is already evident that in many nations within the Global South, the development of infrastructure needed to accommodate this increasing population density is lagging. One significant outcome is the migration of individuals into informal settlements, such as slums, where competition for scarce resources becomes more intense. By the informality of neighborhoods in this paper, I refer to the lack of formal planning and regulation by governing authorities on properties, leading to inadequate access to roads, other infrastructure and services. In the Method section, I will detail how I use measures derived from building footprints to proxy informality and frame this analysis.

In these environments, non-governmental actors, including organized criminal groups and informal networks, frequently intervene to provide essential goods and services, stepping into roles traditionally served by formal institutions. This phenomenon prompts critical questions regarding its implications for violence within these communities. Specifically, what factors influence the degree of involvement of non-governmental actors in resource allocation, and how does this involvement affect the levels of violent conflict in different urban settings? Understanding these complex relationships is not only critical for managing the immediate challenges posed by urban expansion but will also inform future policies for fostering long-term stability and security in urban areas of low-income developing countries.

South Africa presents a compelling context for this study. Its tumultuous history has led to a weakened government system and a proliferation of vigilantes and criminal organizations across various industries (Vigneswaran, 2014). The country's legacy of apartheid has resulted in persistent economic inequalities that continue to the present day, contributing to a lack of governmental monopoly over violence. Criminal organizations have gained political control by leveraging influence over votes to further their business interests. Scholars have studied the phenomenon of "state capture" in South Africa, whereby state apparatus is appropriated for private gains (February, 2019). These criminal groups are not only entrenched in informal settlements and townships but also have a significant presence in transportation, construction, and the drug trade (Irish-Qhobosheane, 2022).

This study adds to ongoing discussions by examining the mediating role of informality in crime rates in developing countries with the presence of strong criminal organizations, and finds significantly lower crime rates in informal urban areas in South Africa. The next section provides a review of existing research on the conflicting relationship between gang presence and violence, and the dynamics of slums and gang activities in South Africa. Section 3 describes the data sources, selected variables, and pre-processing steps. The fourth section outlines the methodologies used for cluster identification, spatial autocorrelation and crime prediction. Section 5 presents the results and interpretation. The paper concludes with key findings, research limitations and future directions.

2 Literature Review

2.1 Slums in South Africa

According to Mike Davis in his book "Planet of Slums" (2006), urbanization in the Global South both mirrors and diverges from the patterns observed in 19th and early 20th-century Europe and North America. In contrast to the industrializing cities of the past, many cities in the Global South are experiencing growth without the accompanying industrialization. This phenomenon is largely driven by global forces that push people from the countryside into urban areas in search of better opportunities. A lack of infrastructure in said urban areas then forces migrants into informal settlements that develop haphazardly. These slums are typified by overcrowding, substandard housing, lack of access to safe water and sanitation, and insecurity of tenure. Millions of people flee impoverished rural areas with hopes of better opportunities, only to find themselves in urban landscapes that struggle to support their growing populations.

Cities are not merely physical spaces. They reflect complex social processes, as David Havery (1989) has claimed. Social processes are more fundamental than the cities they produce, yet they are mediated by the landscapes they create, sustain, and dissolve. Urbanization is a class phenomenon inseparable from its underlying capitalistic spatiotemporal relationships: South African cities remain deeply inadequate due to apartheid-era planning and inequitable market dynamics, which have entrenched inequality and exclusion that continues to dominate the country today, undermining efforts toward urban democratization and integration. Under this background, informal settlements in urban areas, typically arising from unauthorized land occupation and disregard land use and building regulations, occupy contested spaces in South African cities both physically, legally, and in public discourse (Huchzermeyer, 2009).

Although legal battles have led to the government's recognition of informal settlement residents to occupy the areas they have settled, the ensuing formal development often erases any trace of organic or community-driven efforts. Instead, these developments tend to replicate the segregated, low-density urban planning and typologies inherited from the apartheid era, as seen in areas like Hout Bay in Cape Town. With reference to the "Cities Without Slums" campaign by the Cities Alliance—a joint initiative of UN-Habitat and the World Bank—the government's campaign against informal land development and house construction by the poor portrays these activities as undesirable and criminal, using words such as "eradication", "elimination" and "zero tolerance" (Huchzermeyer, 2009, p. 61). This resulted in policies prioritizing slum removal over improving residents' lives. Relocation often worsens people's conditions due to the distance from schools and business, with harsh eviction methods, fractured social networks, even disrupt access to basic necessities like water and shelter. Given all these, Huchzermeyer argues that in-situ upgrades to these housing would be the optimal policy, and a mindset change is the most important factor required to push forward these policy. The current situation resonates with Wacquant's theory of "advanced marginality" (Wacquant, 2007), wherein the negative perception of stigmatized spaces, irrespective of their reality, is the critical factor. Such perceptions lead to marginalizing policies and attitudes that drive these spaces into a downward spiral. Consequently, these areas become increasingly alienated, resulting in the erosion of their supportive system. Examining the dynamics of these informal settlements is crucial for understanding the true underlying issues and will inform the development of targeted solutions to address the challenges faced by these communities.

2.2 Criminal Governance and Crime

Empirical research indicates that gang governance can produce varying impacts on crime rates and economic conditions. A study has shown that homicide mortality rates within favelas—informal urban settlements characterized by high population density and limited infrastructure—of Rio de Janeiro are equal to or lower than those in the rest of the city and that violence is typically driven by territorial disputes rather than the mere existence of favelas (Barcellos & Zaluar, 2014).

Similarly, Oliveira (2023) demonstrated that the governance established by the PCC in São Paulo substantially suppresses homicide rates and fosters economic development. The region is stabilized by the provision of governance and public goods, including the enforcement of property rights and the resolution of disputes. Conversely, Melnikov et al. (2022) show the negative economic consequences of gang territorial control through a natural experiment in El Salvador following the deportation of gang leaders from the U.S. Using survey and census data, they attribute these negative effects to restrictions on residents' mobility imposed by gangs in the presence of rival factions, which limit the residents' choices of jobs. Contrasting their findings with past studies suggesting more positive outcomes, the authors highlight the role of the urban environment, which facilitates government access to gang-controlled areas and the provision of public services. Consequently, it is not in the gangs' interest to function as agents of urban clientelism and contribute to the provision of public goods, therefore, their existence has exclusively negative impacts.

In addition to gangs profiting from extortion or illegal trades, informal governance of housing can also contribute to violence: insecure or informal tenure systems of property ownership exacerbate the scarcity of land in rapidly urbanizing regions, thereby intensifying conflicts and instability. Conflicts over informally developed land in the outskirts of the Mexican city of Xalapa had escalated into violent clashes. Parallel instances have also been documented in Juba and Nairobi. Such land conflicts can disrupt social cohesion among residents, as they may lead to renewed hostilities along pre-existing fault lines rooted in ethnic, national, and party political divisions in already marginalized areas (Lombard & Rakodi, 2016). In the context of developing countries in the Global South, various groups—such as governments, criminal organizations, and residents—often navigate unstable land regimes, potentially leading to violence. Recently, South Africa has been grappling with the emergence of the "construction mafia," organized groups that exploit the construction sector under the guise of radical economic transformation. These groups originated in the townships of Umlazi and KwaMashu in the KwaZulu-Natal province and have expanded their operations across the country since 2018 (2022). Exploiting the weak response from law enforcement agencies, these gangs engage in systematic extortion by demanding a share of profits or stakes in land development projects, often resorting to violence and site invasions to ensure their demands are met. Similar activities are prevalent in other sectors, such as the taxi business, where criminal organizations employ violence, including hiring armed personnel to target rival company employees to dominate the market (Dugard, n.d.). Gangs thrive in various industries that flourish in urban areas with insufficient governmental oversight, leveraging violence to secure and enhance their profits.

Violence can have profound implications on urbanization patterns and the organization of urban spaces. Research has demonstrated how urban environments, particularly in regions such as Latin America, are increasingly being reshaped in response to violence. This transformation involves various actors, including state authorities, criminal organizations, and private entities, and often results in heightened spatial segregation (Koonings & Kruijt, 2007). In another case, Villarreal (2021) provides an analysis of Monterrey, Mexico, illustrating the differential impact of urban violence on wealthy and impoverished communities. Specifically, this study demonstrates how the efforts by the upper class to transform the municipality of San Pedro into a fortified enclave has contributed to the fragmentation of urban space.

Benjamin Lessing (2021) proposed a framework for examining the “duopoly” of violence shared between state and criminal organizations, a defining characteristic that sets criminal governance apart from other forms of governance. State formation often relied on collaboration with criminal groups, while the structure of government control simultaneously influenced criminal activities. Criminal governance embeds itself in areas of state weakness, such as prisons and low-income neighborhoods, which can sometimes be penetrated by strong state power. By examining the dimensions of “who, what, and how” in criminal governance, Lessing aims to offer a framework for future scholars to comprehend the mechanisms and commonalities of such systems. Additionally, the author described potential pathways that lead to gang governance over civilians: in order to maximize revenue, gangs may provide security in exchange for taxation, reduce police attention, deter competitors, and increase political leverage. This phenomenon is particularly evident in urban peripheries, where criminal organizations may establish public order and stable property rights. This not only aids in revenue extraction but also decreases the state’s costs of neglecting marginalized areas and populations. Consequently, this fosters a symbiotic relationship between gangs and a weak government.

2.3 South African Gangs

Researchers have discovered that criminal governance networks in rapidly urbanizing South Africa adapt their organizational strategies to fill the gaps left by inadequate formal state authority. Lambrechts (2012) conducted qualitative research on the Cape Flats, a relatively informal area located southeast of the capital city, Cape Town. Utilizing interviews and participant observation, the study explored the interactions between residents, gang members, and the local government. The findings revealed that local gangs that profit from drug trade and protection money function as providers. They supplying jobs, investments, and goods to the community. Both community members and local police agreed that gangs often purchased everyday necessities, such as food, for the community. Moreover, these criminal organizations frequently covered expenses for water and electricity bills and provided clothing for their employees. Criminal bosses has also been said to engage in charity acts through community investment. Respondents described that, provided that they complied with the gangs, they were left in peace. In contrast, community members consistently perceive law enforcement responses to incidents as highly inadequate, and some police officers get paid by or are friends with the drug merchants. In this case, it is clear that gangs embed themselves within communities by offering services and a sense of identity in the absence of governmental institutions.

Similarly, in Johannesburg, Vigneswaran (2014) documents the emergence of innovative territorial power strategies that encompass both violence and adaptations in policing. In the postapartheid era, South African policing policies in Johannesburg transitioned towards community-focused law enforcement and the use of crime mapping to address high crime rates, while still grappling with the legacies of apartheid-era practices. Vigilante groups emerged as a response to perceived inadequacies in official policing, often employing aggressive tactics that blend historical methods with modern strategies, challenging the state's monopoly on legitimate violence. The manifestations of criminal governance in these South African cities exemplify varied forms of governance, each yielding distinct outcomes concerning public safety.

Why does gang control produce favorable outcomes in certain scenarios but unfavorable ones in others? This paper proposes that informality might be an important confounder in the urban areas of developing countries. Informality might mediate criminal violence in three ways. Firstly, due to the difficulty in police navigation and limited access to roads, informal areas may become desirable territories for gang control as they are easier to defend and regulate. This also means that violence tends to occur in more open domains or at the boundaries of regions managed by different actors, where these groups meet more frequently. Secondly, the informal nature of residences can constrain access to essential governmental resources such as electricity and transportation, compelling residents to seek out informal support networks. Consequently, gangs are incentivized to supply these public goods to garner support from the local community and potentially enhance the amount of extortion they collect. This also implies that with constrained resources, there may be reduced investment in combat. Gang are also more likely to foster a positive relationship with these communities. In contrast, in more established settlements, residents have access to state-provided services. The presence of gangs can only hinder this by strategically destroying certain infrastructure, thereby causing negative consequences. Finally, the unregulated land tenure system in the urban areas of these rapidly developing regions can lead to conflicts, particularly when multiple actors lay claim to the same territories. This situation is more likely to occur in formal areas where criminal organizations attempt to wrest control from the government or rival groups. Based on this reasoning, this study hypothesizes a negative correlation between informality and crime rates in urban areas marked by gang activity. The hypothesis will be tested using K-means clustering, local spatial autocorrelation methods, and crime prediction algorithms, which are discussed in the following sections. The measures of accessibility in these pathways, such as road length and altitude, could influence violence independently of informality. Therefore, they are included in the analysis as explanatory variables, which are discussed further in the next section.

3 Methods

3.1 Data Description

The principal variable of interest in this project is the crime rate across different types of urban settlements in South Africa. Crime statistics are reported annually by the South African Police Service (SAPS), categorized by crime type for each of the 1,164 police stations. For this analysis, the five-year crime data from 2018 to 2022 was utilized, which ensures alignment with the independent variables and helps mitigate anomalies in the dataset. The total number of crimes reported was used as an approximation for the distribution of all forms of violence instead of any single category. The five-year average of total crime for each station was calculated then geocoded by joining it to the corresponding point locations of police stations and respective jurisdictions published by SAPS website. Given that the police station area represents the spatial unit with the largest granularity in the dataset, it was selected as the unit of analysis.

The independent variables primarily originate from the Million Neighborhood Project (Bettencourt & Marchio, 2023), which offers the first comprehensive map of population and urban development at the street-block level across sub-Saharan Africa. For South Africa, 1,457,745 blocks are covered. From the MNP dataset, density variables covering population and building density are included. Both the total population and the number of buildings were aggregated from the block level to their corresponding police jurisdictions. Subsequently, logarithmic transformations of population and building density (measured as the number of buildings per km^2) were calculated for use in clustering analyses. Additionally, the MNP dataset provides an innovative quantification of informality, k-complexity, by examining the topology of building footprints within a block, defined via weak dual graphs of a parcel-level neighborhood map (Figure 1). This entails constructing successive graphs by connecting the center of each parcel that shares a boundary, and repeating this process for the next graph until the final geometry is a tree with no interior faces. Therefore, the k-complexity of a group of buildings is represented as a positive integer with a minimum value of one. A higher k-complexity indicates that a greater number of buildings must be traversed to access a road, starting from the most disadvantaged point within the group. An illustrative example is provided in Figure 2.

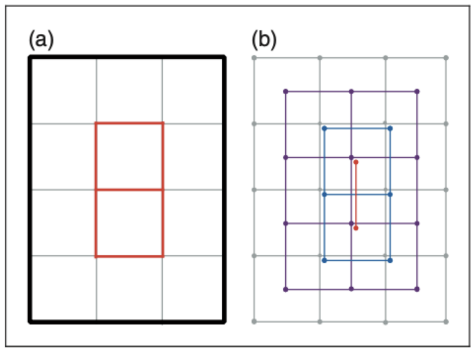

Figure 1: Sample calculation of k-complexity by constructing successive graphs (gray - purple - blue - red) until the graph (red) is a line with no interior faces. This process was repeated 3 times so k=3 for this geometry. (Figure from Brelsford et al., 2019).

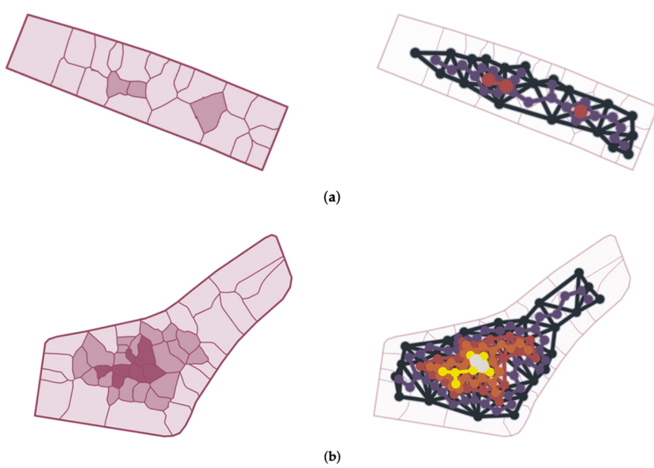

Figure 2: Two blocks in Freetown, Sierra Leone with different levels of inner parcels, corresponding to k=3 (a) and k=6 (b). (Figure from Soman et al., 2020).

Previous studies have utilized remote sensing techniques (Cinnamon & Noth, 2023) for the detection of informal settlements in South Africa. In particular, researchers employed satellite imagery from the USGS GloVis platform and high-resolution aerial photographs obtained from the City of Cape Town's open data portal. By implementing a supervised object-based classification methodology, they effectively identified and analyzed the temporal changes in slum areas within Cape Town over a twenty-year period, from 2000 to 2020. In this project, we attempt to utilize a more recent dataset with greater coverage to identify all the informal areas in South Africa.

Accessibility of different areas are also proxied by the total vehicular road length within each policing area. The data was obtained through the ohsome package in R, a tool designed for analyzing OpenStreetMap (OSM) data, which enables querying road lengths by specific types within given geometries. The R code used is included in the GitHub repository linked at the end of the paper in the Data and Code Availability statement for reference. Furthermore, the average altitude of each police station area was computed using the Copernicus GLO-90 Digital Elevation Model (DEM). Altitude is considered relevant to accessibility as mountainous areas generally exhibit higher altitudes and are more challenging to access.

3.2 Exploratory Data Analysis

The histograms of the six variables mentioned above are shown below (Figure 3). The distributions of building and population densities appear to resemble a normal distribution, with an additional peak to the right. This bimodal pattern potentially suggests the existence of two distinct types of cities in urban and suburban areas. In contrast, K-complexity, road length, and crime exhibit a right-skewed normal distribution, with more units on the lower side of the distribution and a long tail. Finally, the average altitude shows a highest peak around 0 - 40 meters and a smaller peak around 800 meters, corresponding to the plain and high central plateau in South Africa.

The population-weighted K-Complexity measure by police jurisdiction areas is mapped in Figure 4. We observe that the most informal areas are primarily concentrated in the southeast and scattered across the northwest. This pattern roughly aligns with the map of the poverty headcount published by Statistics South Africa (2016). To validate the use of the MNP K-Complexity measure as an identification tool for slums and as a measure of informality, we zoomed in on Cape Town (left of the figure), a well-studied area, to compare the data used in this study with official data from the City of Cape Town in 2015. This data was published as a joint effort from OpenUp and collaborating organizations including Ndifuna Ukwazi, Social Justice Coalition, and International Budget Partnership (2022).

The map in the upper left corner shows the plotted informal areas from City of Cape Town, with color intensity representing the population density for each informal settle-

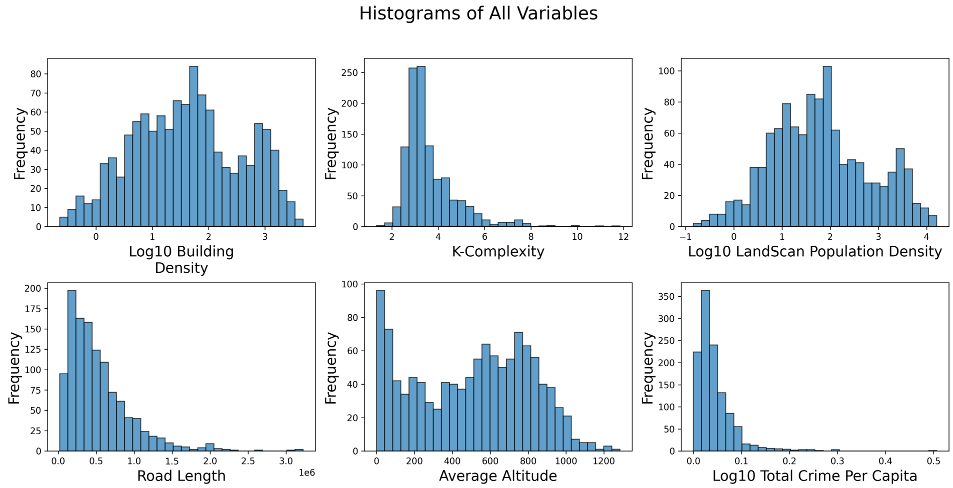

Figure 3: Descriptive histograms for all six variables used in this study.

ment area. And the lower left corner is the MNP data, colored by population weighted k-complexity. There is a clear match between the identified settlements lining the N2 highway, which runs southeast-northwest. Additionally, the MNP dataset appears to have captured the newly formed informal settlement pocket on the north of the road, located in the Driftsands Nature Reserve. Thousands of people occupied this area during the Covid lockdown in 2020 (VPUU, 2022). These observations validate the use of this parameter.

The slums manifest a dispersed character, often surrounded by formal settlements. Examples include the planned housing leftover from the apartheid period in Mitchell's Plain in the west and Delft in the north. According to Garrido (2021), segregation in urban areas with slums is characterized by the interspersion of slums and enclaves rather than the concentration of impoverished neighborhoods. This highlights a relational nature of segregation, emphasizing unequal interactions between residents of slums and those in enclaves, which leads to distinct social dynamics and the spatialization of group identities. Unfortunately, this nature of the informal settlements is somewhat removed in the aggregated data at the police jurisdiction level with lower granularity. However, it remains an important phenomenon for future studies.

The spatial maps of the remaining four independent variables—population density, building density, road length, and average altitude—offer additional information on the geographical and infrastructural landscape of the different areas. The building density and population density maps exhibit similar spatial distributions, as anticipated. Both variables highlight major urban centers such as Cape Town, Johannesburg, and Durban with higher densities. In contrast, the map of road length does not exhibit a strong correlation

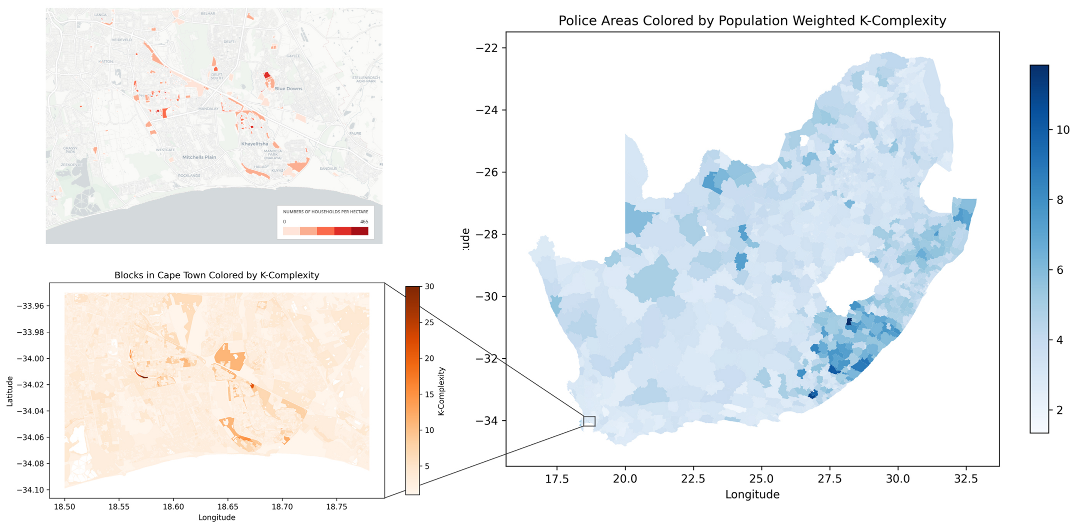

Figure 4: This map presents the population-weighted k-complexity derived from MNP data, by aggregated to police area across South Africa. The inset on the left provides a detailed view depicting the weighted k-complexity at the block level for Cape Town. The upper figure illustrates the informal settlements identified using official data from the City of Cape Town, while the lower figure is based on MNP data. A comparative analysis between the two figures demonstrates that using k-complexity as an indicator accurately identifies informal settlements present in the 2015 official data, as well as newly established slums that have emerged more recently.

with population or building density. This observation suggests that the development of road infrastructure is not solely dependent on the density of buildings or the concentration of population. Nonetheless, there appears to be a somewhat negative correlation between road length and altitude. The average altitude map reveals the topographical features of the country, showing the high central plateau known as the Highveld, characterized by altitudes exceeding 1000 meters. This elevated region stands in contrast to the lower altitudes observed along the coastal and peripheral areas.

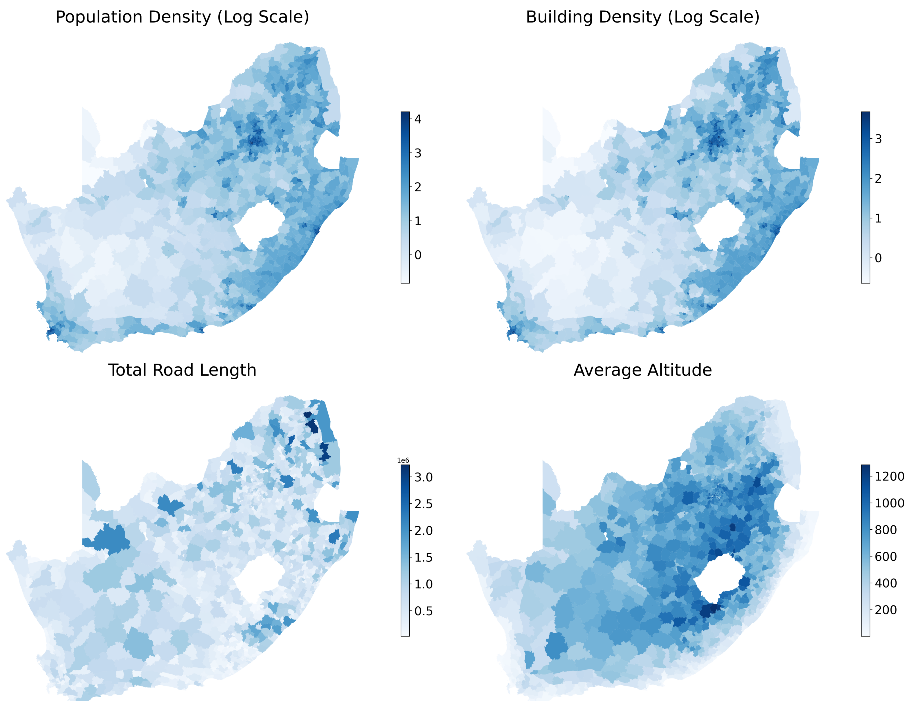

Figure 5: Descriptive maps for four dependent variables used in this study.

Figure 6 illustrates the log-transformed (base 10) per capita 5-year average crime rate across different regions of South Africa. The major metropolitan areas with the highest populations, Johannesburg and Cape Town, are zoomed in to display the smaller jurisdiction areas, highlighting a concentration of crime within these densely populated cities. Furthermore, other major coastal cities, including Durban on the eastern coast and Gqeberha on the southern coast, exhibit high crime rates within certain police precincts. Notably, the origin location of construction gangs mentioned before, the eThekwini District in KwaZulu-Natal, is among the top ten locations with the highest per capita crime rates. An analysis of categorized crime statistics reveals that this district has experienced above-average levels of “crime detected as a result of police action” and “drug-related crime” during the period from 2018 to 2022. Conversely, regions along the northern borderline and near the southeast coast display comparatively lower crime rates. Finally, three areas display missing data due to discrepancies between crime statistics and the police jurisdiction geometry as provided by SAPS.

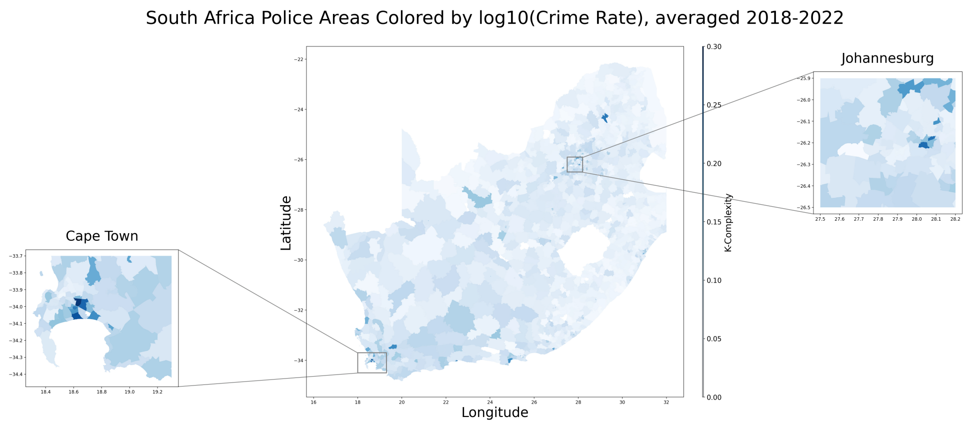

Figure 6: Log10 per capita crime rates for each precinct across South Africa. Insets for Cape Town and Johannesburg highlight the regions with higher crime rates within smaller police jurisdictions, providing a more granular view of high-crime areas within these major metropolitan cities.

In the subsequent machine learning analyses, we incorporate longitude and latitude as additional independent variables to better capture geospatial dynamics. Displayed in Figure 7, the correlation plot illustrates the relationships among all seven variables under consideration. The plot reveals a strong positive correlation between population and building density, which aligns with the spatial distribution patterns previously observed. Furthermore, there exists a positive correlation between density and crime rates, suggesting that densely populated regions may be more prone to criminal activities. Other access measures such as altitude and road length exhibit a negative correlation with crime, with k-complexity demonstrating the strongest relationship.

An intriguing inverse correlation also emerges between crime and geographical coordinates—latitude and longitude. This relationship may be influenced by the geographic distribution of criminal activities. The regions with highest per capita crime rates are concentrated in the southwestern region of the country, particularly radiating out from Cape Town. This spatial pattern likely skews the overall correlation with longitude and latitude variables.

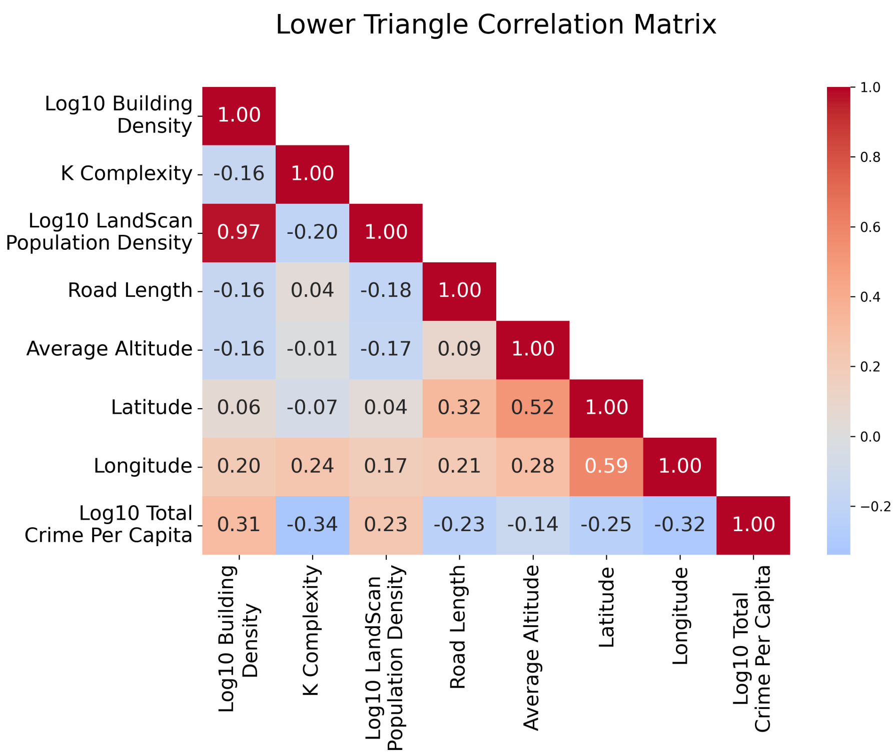

Figure 7: Correlation matrix displaying the relationships between all seven independent variables and the dependent variable, crime.

3.3 Cluster Analysis Methodology

Researchers have utilized clustering methods to analyze and predict crime patterns. One such method is the K-Nearest Neighbors (KNN) algorithm, which is a non-parametric technique used for classification and regression. The KNN algorithm operates by examining the proximity of data points within a feature space, and it determines the value of the dependent variable of an observation by considering that of its K nearest data points. A classification is made based on the majority class among these neighbors, demonstrating the use of supervised learning. Conversely, the K-means clustering method is employed to uncover intrinsic patterns within the data through unsupervised learning where there is no outcome variable to be predicted. This technique also groups data points into clusters based on their similarity, effectively reducing the complexity of high-dimensional relationships among variables into a lower-dimensional distance. Through this process, K-means clustering facilitates the identification of natural groupings within the data, thereby enhancing our understanding of underlying patterns and trends in crime data.

The dataset from the National Crime Records Bureau (NCRB) of India, which records crime as individual events, has been used to train KNN models for crime prediction. With temporal variables and location variables including hour of the day, day of year, and longitude and latitude information as features, the existence of different classes of crimes included as binary variables are used as dependent variables for six cities in Tamilnadu, India (Kumar et al., 2020; Sivaranjani & Sivakumari, 2016). In contrast to hard clustering methods that put objects in one and only one cluster, a study used fuzzy clustering method which classifies events in more than one clusters in a membership gradient. This algorithm was used to understand types of felonies from annual data from 2005 to 2006 in the Mexican city of Hermosillo, which allowed researchers to correlate the fuzzy clusters with urban structural patterns in an exploratory unsupervised study (López-Caloca et al., 2009).

The aforementioned studies all used individual criminal events as samples within their clustering analyses. In contrast, this study employs police districts as the units of analysis, driven by the considerations for data availability and the objective to comprehend criminal governance and violence at a broader geographical level, encompassing the formality of the surrounding area. In this research, the K-means clustering algorithm is utilized for unsupervised clustering of police jurisdictions. Subsequently, the relationship between independent variables and per capita crime rates is examined within each identified cluster. This method is well-suited for handling multidimensional data, where several variables can interact simultaneously, as have been shown in the exploratory analysis above with matching spatial patterns. For the implementation, we leverage the Python programming language and the sklearn package, ensuring robust and reproducible results.

The five variables utilized in the clustering analysis are as follows: the logarithm of building density, the logarithm of population density, the k-complexity measure derived from MNP and weighted by population, the total length of vehicular roads, and the average altitude, for each police jurisdiction. Informed by previous literature, the hypothesis is that police areas characterized by higher k-complexity, indicative of greater informality, are expected to exhibit lower levels of violence. This is attributed to the tendency of gangs in such areas to provide services and governance due to the vacuum of power from the state, fostering better relationships with residents and reducing their incentive to incite conflicts. Conversely, it is hypothesized that more formal areas, characterized by extensive land development projects and heightened police presence, are likely to experience elevated levels of violence. Moreover, the inclusion of other accessibility-related variables is anticipated to show that less accessible areas tend to have lower levels of violence. Finally, it is hypothesized that higher population and building densities do not necessarily correlate with increased levels of violence.

Initially, an elbow plot is generated to determine the optimal number of clusters by plotting the sum of squared distances of each data point to its nearest cluster center against the number of clusters (Figure 8). Seven clusters are chosen because the reduction in inertia diminishes when transitioning from 7 to 8 clusters compared to the transition from 6 to 7 clusters, and a relatively small number of clusters maintains the interpretability of each cluster's representation. Additionally, various k values are assessed to evaluate sensitivity. Subsequently, the resulting clusters are analyzed for differences in their crime rates and other characteristics, including their geographical locations. The findings are presented in the following section.

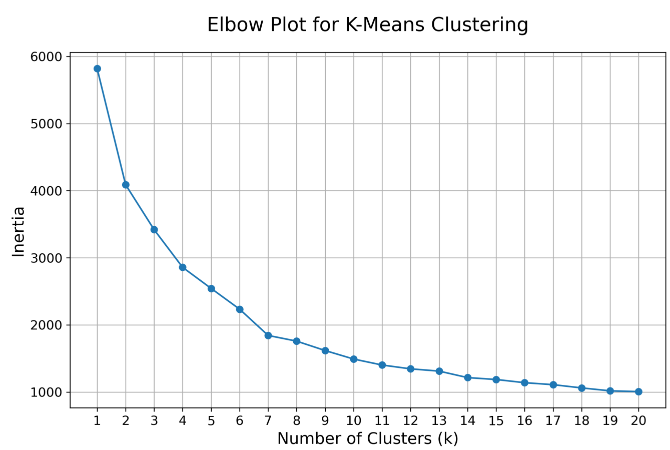

Figure 8: Elbow plot for K-Means clustering.

3.4 Local Spatial Autocorrelation

To further explore the spatial patterns within the dataset and evaluate local spatial autocorrelation, Local Moran's I is employed to identify clusters in the crime statistics. This method uses a spatial weights matrix to locate local patterns of correlation in a variable. These clusters can subsequently be compared with those generated using KNN from the previous step to identify crime hot spots. To ensure the robustness of the identified local clusters, three different spatial weight schemes were utilized: (1) rook contiguity, which considers shared edges between spatial units; (2) queen contiguity, which accounts for both shared edges and vertices; and (3) KNN weights, which consider all k nearest neighbors as connected, with k=5 being used in this study.

3.5 Crime Prediction with KNN and Random Forest

Building on the insights gained from the unsupervised K-means clustering analysis, we next employed supervised machine learning algorithms to predict crime levels using the same set of features, supplemented by location data. Specifically, we applied KNN and random forest models. KNN predicts crime rates by identifying patterns among the most similar observations within the dataset. In contrast, random forest constructs an ensemble of decision trees, where each tree is trained on a randomly selected subset of the data and features. A single decision tree operates by recursively splitting the data based on feature thresholds, creating branches that ultimately lead to a prediction. By aggregating the predictions of multiple trees through averaging, random forest mitigates overfitting, improves predictive accuracy, and captures complex, nonlinear relationships within the data. These algorithms are particularly well-suited to modeling crime, as they can account for intricate interactions among socioeconomic and spatial variables. Following model training, we evaluated their performance using Mean Squared Error (MSE) and analyzed feature importance to interpret the contributions of different predictors.

Similarly to the process involved in K-means clustering, hyperparameter tuning is needed when applying the K-Nearest Neighbors (KNN) algorithm, specifically choosing the number of neighbors. Figure 9 illustrates the relationship between the number of neighbors and the MSE of the predictions. Notably, the most significant reduction in error occurs between k=2 and k=3. Although the error continues to decline up to k=5, it eventually plateaus. Beyond k=5, the decrease in MSE is minimal, with only a 1.5% reduction observed between k=5 and k=7. Therefore, we opted for a smaller k value of 5 to achieve a balanced variance-bias trade-off. This choice allows the model to closely adapt to local patterns, including spatial variations, by capturing finer details while ensuring predictive stability.

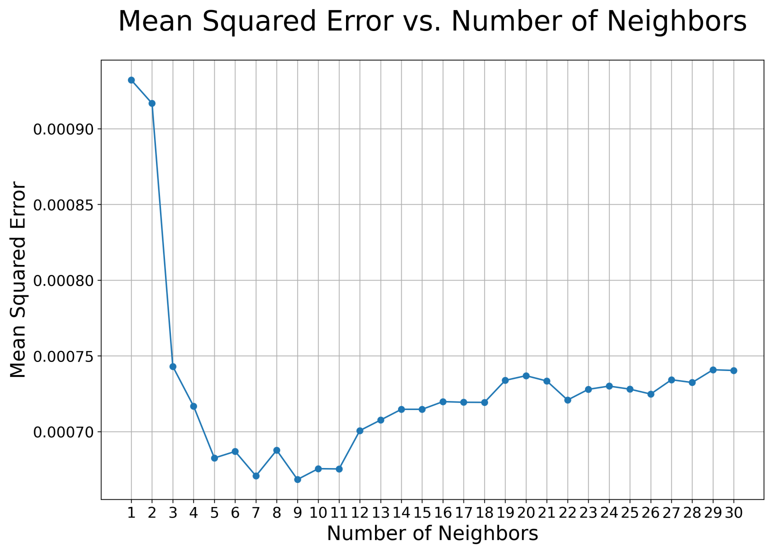

Figure 9: Impact of number of neighbors on the mean squared error in the KNN Algorithm, illustrating the characteristic U-shaped curve.

3.6 Case study: Crime and Growth in KwaZulu-Natal

To better understand the causal dynamics between growth and violence within South Africa, we will conduct a case study focusing on the province of KwaZulu-Natal, where the construction gangs emerged. By utilizing the Open Buildings 2.5D Temporal Dataset(Google Research, 2024), which provides longitudinal building footprint data complementary to the MNP dataset, we extract growth rates of building density across the years 2016 to 2023. This dataset allows us to examine changes in building density at the level of police jurisdictions, comparing growth rates in these areas to crime rates on an annual basis. In this analysis, density change serves as a proxy for urban growth, a practice commonly adopted in the literature on urban studies. Building density is frequently used because it generally correlates with population growth and can reflect the intentions of individuals or communities to settle in particular areas.

To explore potential causal relationships between these two variables, we apply the Granger causality test, originally introduced by econometrician Clive Granger((1969)). This test assesses whether fluctuations in one time series systematically precede changes in another, suggesting a direction of causality. We investigate both directions of causality: whether violent crimes affect growth, and conversely, whether growth influences violent crime rates. Due to the nature of this test, only two variables can be included, and the assumption is that all external factors remained relatively constant locally during the examined time frame. According to Irish-Qhobosheane (2022), invasions linked to local business forums affected nearly all construction sites in KwaZulu-Natal between 2016 and 2019. Our hypothesis posits that a stronger causal relationship exists from high growth to high crime, particularly as regions with housing projects are often targeted by construction mafias. Such activities could have then disrupted economic growth by halting or delaying construction projects in formal urban settings, which can potentially result in lower crime rates due to diminished targets for criminal activity. This suggests that we might observe fluctuations in crime that lag behind growth patterns.

4 Results

4.1 Clustering Results and Interpretation

The geographical distribution of the clusters is depicted on the left side of Figure 10, highlighting the spatial patterns. Although the K-means clustering method employed is not spatially constrained, most clusters exhibit contiguous areas. On the right side of Figure 5, a bar plot depicts the distribution of police jurisdictions across the clusters, revealing a relatively balanced distribution with no single cluster being disproportionately large or small. Figure 11 provides boxplots for each independent variable, in addition to the dependent variable, across the seven clusters.

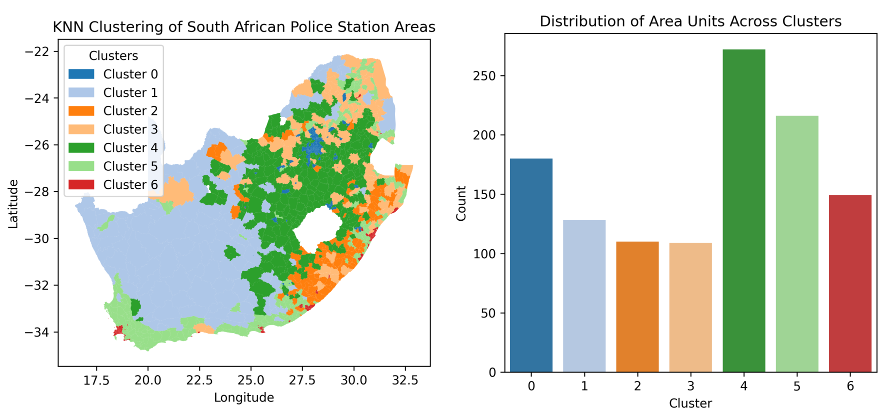

Figure 10: Map of seven clusters from K-means clustering, with the number of police jurisdictions in each cluster.

Clusters 0 and 6 exhibit the highest population and building densities, with the blue regions roughly corresponding to Johannesburg and the red regions to Cape Town and Durban, the three largest cities in South Africa. Despite differences in average altitude, both clusters with highest population density also are the ones with the highest average crime

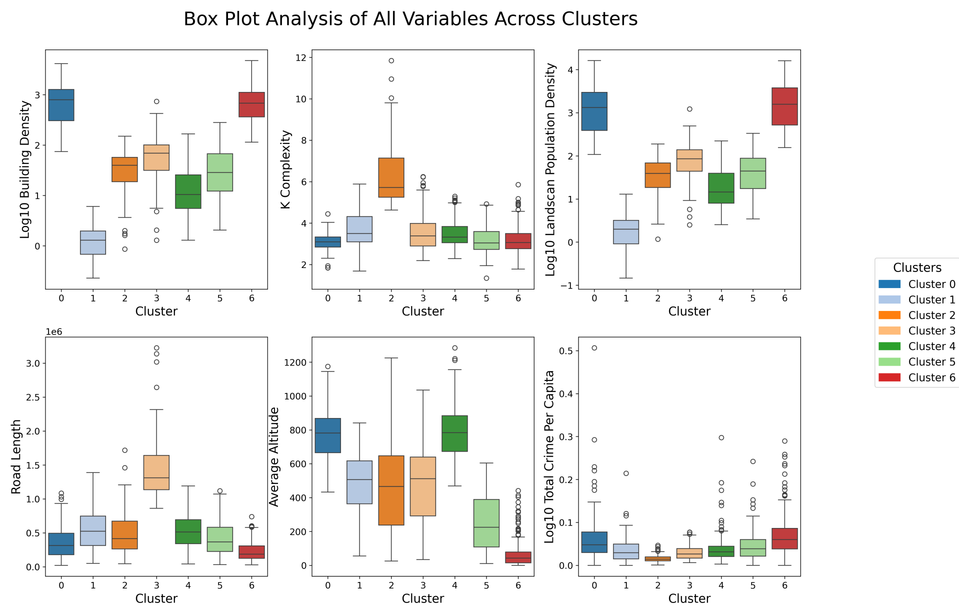

Figure 11: Box plot of all variables for each cluster.

rates. However, numerous confounding variables in larger cities, which were not included in this study, make it difficult to draw definitive conclusions regarding causation. In contrast, cluster 1, representing areas with the lowest population and building densities, is situated in the southwest region of the country, likely encompassing rural areas. Interestingly, this cluster does not exhibit the lowest per capita crime rate. Cluster 2 has the lowest crime rate, and this difference is statistically significant (see Appendix). This cluster also has the highest k-complexity. Clusters 2 to 5 have relatively similar population and building densities, with cluster 3 showing the highest vehicular road length, and cluster 4 the highest average altitude. Despite these differences, only cluster 2 demonstrates significantly lower crime rates.

A closer examination of the relationship between crime and k-complexity within each cluster reveals a consistent pattern: in nearly all clusters, there is a negative correlation between normalized k-complexity and per capita crime rates. For example, Cluster 2 exhibits a negative slope of -0.34 with an $R^2 = 0.11$. The strongest correlation is observed in Cluster 5 (slope $= -0.34$, $R^2 = 0.11$), a cluster that shares many characteristics with Cluster 2 aside from its lower k-complexity. Only Clusters 3 and 6 displays an ambiguous association between the two variables; detailed plots for all clusters are available in Figure 12.

To further investigate this pattern, we aggregated data from Clusters 2 through 5 in Figure 13. These clusters represent urban areas with sufficiently high density to be classified as urban, yet not among the most densely populated metropolitan zones. The combined

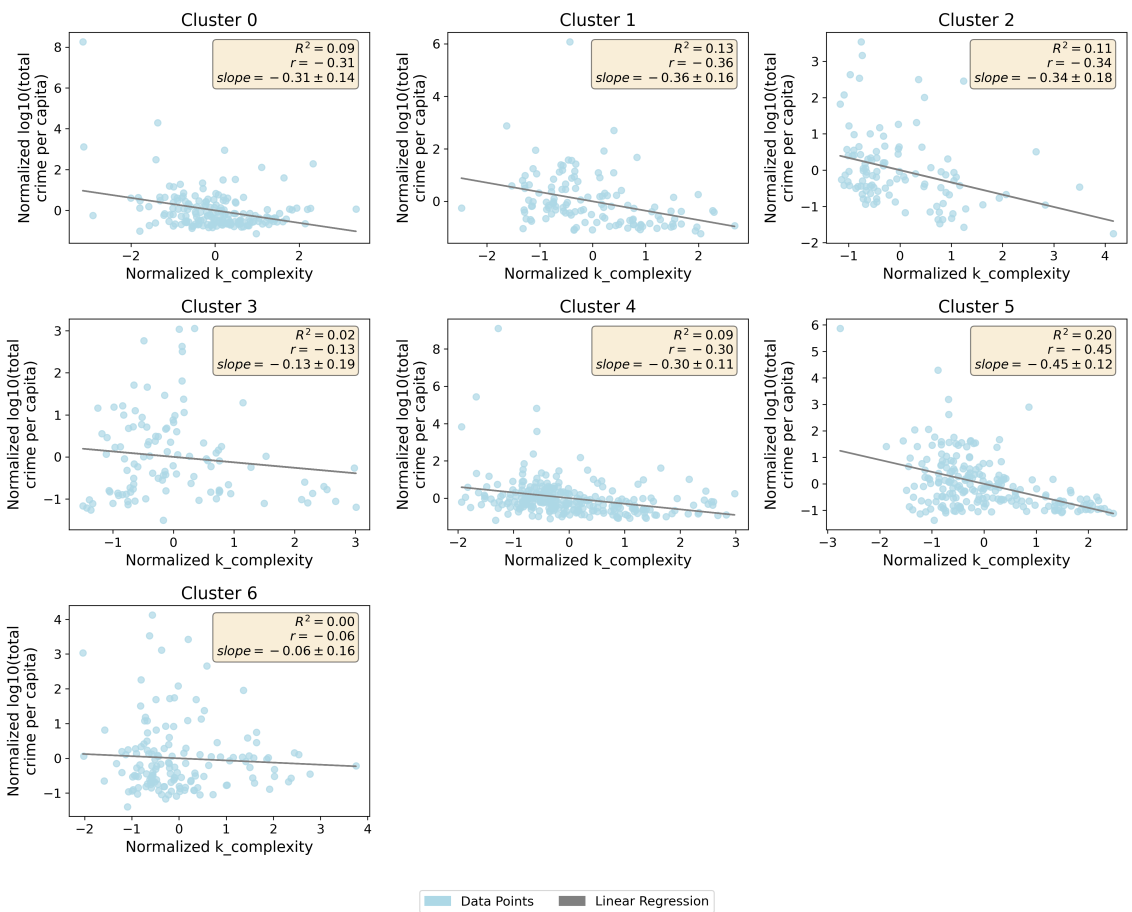

Figure 12: Regression of normalized k-complexity vs. log10 per capita crime across clusters (0-6), illustrating a predominantly negative correlation pattern in most clusters.

data continue to reflect a clear negative relationship. The R^2 value of 0.16 suggests that approximately 16% of the variance in the normalized crime rate can be explained by k-complexity—indicating a modest but meaningful association. The confidence interval for the slope further confirms that this negative relationship is statistically significant, though not particularly strong. The plot reveals a clear pattern: at low levels of informality, crime rates exhibit substantial variability. King Shaka International Airport and Van Reenen in KwaZulu-Natal are striking outliers, likely driven by the way per capita crime is calculated using residential population. In contrast, as informality increases (i.e., with greater k-complexity), crime rates tend to converge at consistently low levels. Cluster 2 lies at the far right of this distribution, characterized by both high k-complexity and low crime.

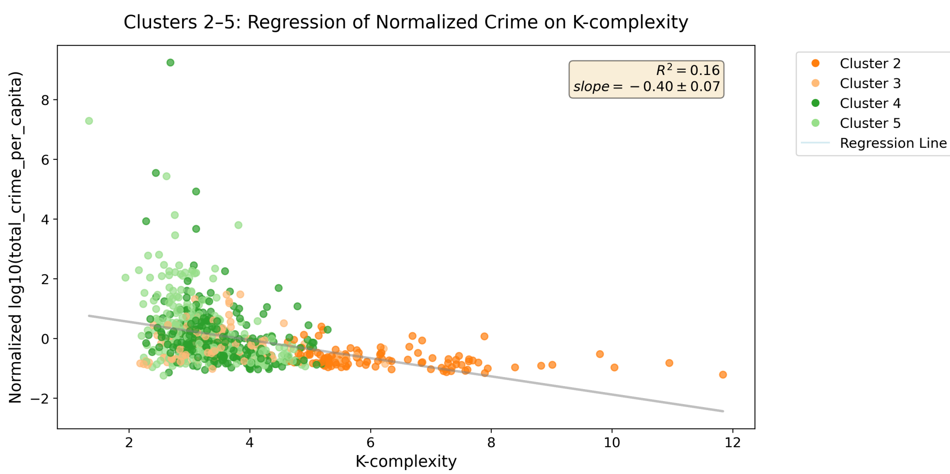

Figure 13: Linear regression results for normalized k-complexity and log10(crime per capita) of clusters 2-5 combined, showing negative correlation.

It is possible that this outcome is influenced by the location of police stations if they are more frequently situated near formal neighborhoods. To investigate this, the distance of each neighborhood to its corresponding police station is examined, with the results illustrated in Figure 14. The scatter plot shows no significant correlation, and the Pearson correlation coefficient between the k-complexity of a block and its distance to the police station is -0.123. This negligible negative relationship indicates that k-complexity and lower crime rates is most likely not due to reporting biases resulting from police station locations. Nevertheless, it is still possible that the correlation is influenced by under reporting associated with other factors correlated with informality.

Lastly, the average per capita crime rates by category for each cluster are compared, showing no significant decrease in any single category for cluster 2, with the decrease being similar across all categories. Comparable clustering results are obtained when k is set to

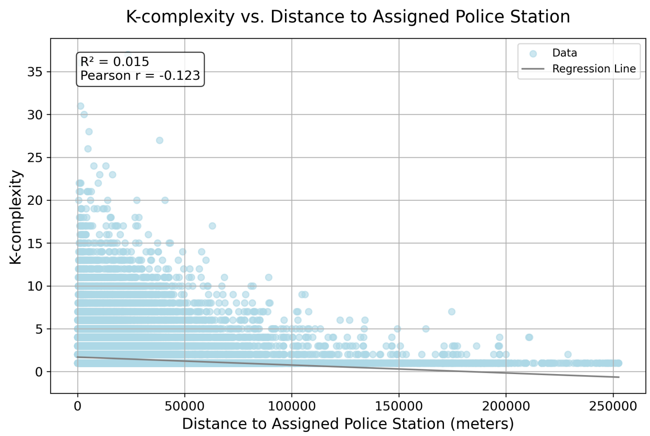

Figure 14: Scatter plot of neighborhood level k-complexity versus distance to police station for each block shows no significant positive correlation to explain lower crime rates in clusters with higher k-complexity.

6 and 8. With k = 6, the two highest density clusters combine into one cluster, spanning a large range in altitude. With k = 8, the informal cluster splits into two clusters with varying road lengths and altitudes. In both cases, similar spatial patterns emerge: cluster(s) with the highest population and building densities exhibit the highest crime rates, whereas cluster(s) with low k-complexity have the lowest crime rates. These additional findings are included in the Appendix.

4.2 Local Moran's I

Local Moran's I was calculated using three types of spatial weights: rook, queen, and KNN. The resulting four types of high-low clusters are illustrated in Figure 15, with the three distance metrics yielding similar outcomes. Most of the identified clusters are high-high or low-low clusters, indicating that cities with similar crime rates are geographically close to each other. High-high clusters represent adjacent areas with high crime rates, while low-low clusters represent areas with both low crime rates, which aligns with the expectation that geographically proximate cities tend to have similar crime rates.

Upon comparing these results with the clustering outcomes, it is evident that the high-high clusters most commonly correspond to cluster 6, characterized by high population and building density. There is also some correspondence with cluster 5 when local clusters are identified using KNN spatial weights. Conversely, the low-low clusters predominantly align with cluster 2 police areas, while also including some regions within clusters 1, 3, and 4. The overlap of cluster 2, characterized by high informality, with low-low clusters, further

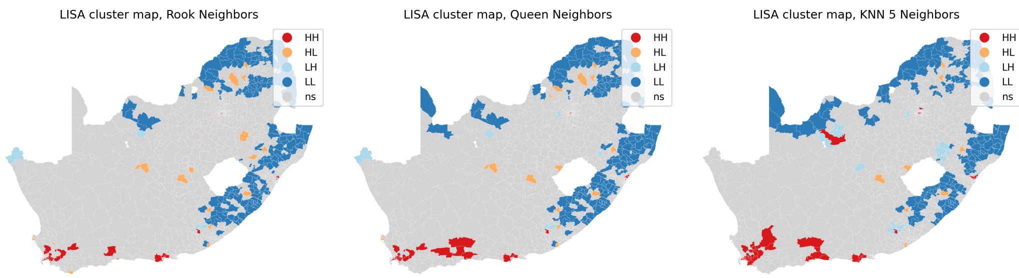

Figure 15: Scatter plot of k-complexity and distance to police station for each block, showing no significant patterns of correlation.

Table 1: KNN Performance Metrics

Metric	Value
Mean Squared Error	0.000683
R-squared	0.377438
Mean Absolute Error	0.017716

support the hypothesis that areas with higher levels of informality have lower crime rates.

4.3 KNN and Random Forest Models

To evaluate the ability of machine learning models to predict crime rates across South African police jurisdictions, we tested two approaches: K-Nearest Neighbors (KNN) regression and random forest regression. Both models achieved relatively low Mean Squared Errors (MSE) compared to the mean and standard deviation of the crime rate variable (mean = 0.044 and standard deviation = 0.040). Specifically, the KNN model achieved an MSE of 0.00068 and an R-squared of approximately 0.38 (Table 1), while the random forest model slightly outperformed it with an MSE of 0.00063 and an R-squared of approximately 0.43 (Table 3). These R-squared values indicate that although the models capture a substantial portion of the variance in crime rates, a considerable amount of unexplained variance remains, suggesting that additional factors not captured in the current features may also influence crime levels. Nevertheless, the low Mean Absolute Errors (MAE) in both models (around 0.017 for KNN and 0.016 for random forest) imply that the models predict crime rates with fairly small average deviations from the true values.

Feature importance results further illustrate that different models prioritize different predictors when estimating crime rates. In KNN, which relies heavily on proximity in feature

Table 2: KNN Feature Importances

Feature	Importance
$\log_{10}$(Building Density)	0.446760
$\log_{10}$(LandScan Population Density)	0.365491
Longitude	0.263004
k-Complexity	0.221077
Latitude	0.217999
Average Altitude	0.207529
Road Length	0.053363

Table 3: Random Forest Regression Metrics

Metric	Value
Mean Squared Error	0.000628
R-squared	0.427591
Mean Absolute Error	0.016673

Table 4: Random Forest Feature Importances

Feature	Importance
$\log_{10}$(Building Density)	0.237914
Road Length	0.195766
k-Complexity	0.149781
Latitude	0.148514
$\log_{10}$(LandScan Population Density)	0.122007
Longitude	0.089265
Average Altitude	0.056754

space, log of building and population density emerged as the most influential variables, with longitude and k-complexity also contributing significantly(Table 2). In contrast, the random forest model highlighted building density and road length as the top features, followed by k-complexity and latitude(Table 4).

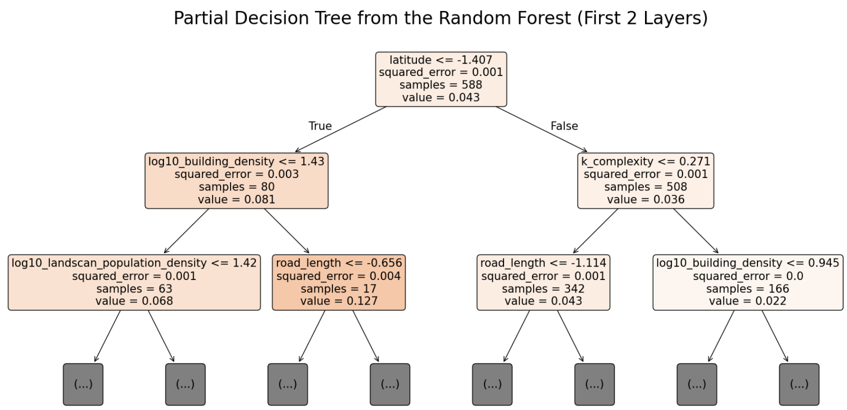

Figure 16: First 2 layers of one tree from the random forest algorithm.

Figure 16 illustrates the first two layers of a tree from the random forest algorithm. The initial split focuses on the latitude of the area, dividing the police jurisdictions into northern and southern regions. In the northern districts, which have lower latitudes, the tree then examines building density. Referring to the exploratory map for k-complexity in Figure 4, we observe that most variation in this variable occurs along the southern coast. Therefore, it is logical that for the northern regions, the algorithm considers other variables, such as building density and population density, when predicting crime. For the southern areas with higher latitudes, k-complexity becomes the next factor analyzed.

In both models, building density and population density consistently play central roles, suggesting that built environment intensity and human presence are strong determinants of crime rates. Informality also plays a significant role in these predictions. Overall, our models demonstrate that relatively simple spatial and infrastructural features can provide meaningful predictive power for crime, even in the complex and varied urban environments of South Africa.

4.4 Granger Causality Test in KwaZulu-Natal

Building on these findings, we further explore potential causal relationships between growth and crime through a Granger causality test in KwaZulu-Natal. Due to the limited temporal scope of the data, only a one-year lag is examined for each directional Granger causality test. Among the 186 police areas in KwaZulu-Natal, only four exhibited statistically significant results ($p < 0.05$) indicating that crime influenced changes in growth, while eight demonstrated significant results for growth influencing crime. Although these findings align with our hypothesis, suggesting a more likely causative relationship from growth to crime in this province, they are not sufficient to draw any conclusions.

Nonetheless, it is interesting to explore these areas with significant outcomes further. Figure 17 illustrates two regions where changes in growth significantly influenced crime, highlighted by their small p-values. Notably, both plots show a marked increase in growth in 2018, possibly attributable to the 2017 agreement between the South African Forum of Civil Engineering Contractors and the Federation for Radical Economic Transformation associated with the construction mafia. Despite this agreement, the construction mafia continued their activities, proliferating nationwide. In Kwambonambi, a more populous coastal area, building density growth persisted alongside elevated crime rates since 2018, with crime rates fluctuating in the opposite direction of growth. This shows that relatively higher growth corresponds to comparatively lower crime rates. Conversely, in the less populated Mid Illovo, growth peaked in 2018 but subsequently slowed, continuing to oscillate with violence with a one-year time lag. However, due to the short duration of the data coverage, it remains uncertain whether the 2018 growth indeed impacted crime rates or if the one-year lagged correlation is coincidental. Given that only eight out of 186 areas exhibit significant results, it is likely that these correlations are merely incidental.

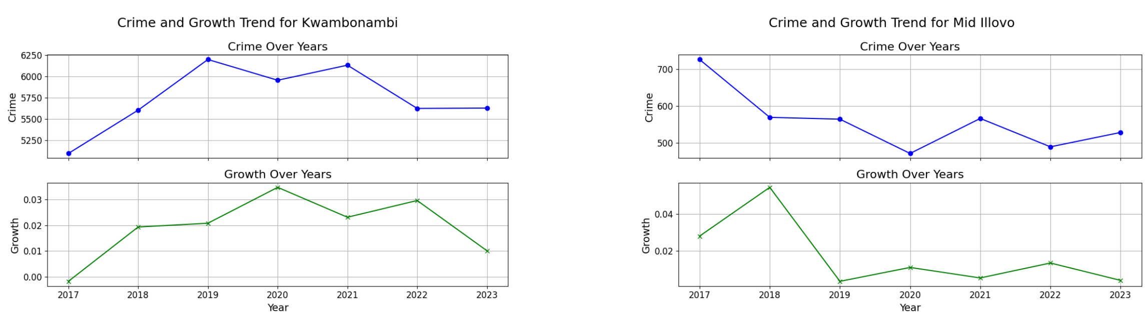

Figure 17: An analysis of crime rates and their growth in Kwambonambi and Mid Illovo, highlighting the time-lagged correlation between these two variables.

5 Discussion

The clustering analysis revealed that areas with the highest crime rates are often associated with the highest building and population densities, represented by clusters 0 and 6. Conversely, in less densely populated cities (clusters 2–5), the regression results indicate that informality serves as a stronger predictor of lower crime rates compared to factors such as average altitude and vehicular road length. Machine learning algorithms further emphasize informality as a key predictor of crime, alongside population and building densities. These findings support the hypothesis that informality acts as a mediating variable for crime, though they do not explicitly assess causal pathways.

To build upon these results, we propose several approaches. First, more informative measures of the local terrain, such as variation in altitude, could provide additional insights by approximating gang and government access to certain neighborhoods. Second, geographical factors such as plantation coverage, building type, and the location of gated communities can serve as informative factors that facilitate a better understanding of localized crime systematically. Local economy data with high granularity can capture patterns of interspersed segregation in housing, a prevalent phenomenon in the Global South. Such data could also serve as an important dependent variable.

The hypothesis in this study depends on the proposed pathway for crime reduction which emphasizes the presence of gangs. However, gang presence data at the police area level was not available and, thus, was not used in the K-means clustering. The assumption of universal gang presence needs to be made to support the hypothesis, which is reasonable in South Africa. However, a critical next step is to find localized areas of gang violence and use them to precisely examine the dynamics between urban informality and violence.

In the context of the KwaZulu-Natal province case study, this research underscores the potential of utilizing newly accessible longitudinal building footprint data to explore the causal relationship between growth and variables such as crime. However, the Granger causality test is limited to analyzing only two variables at a time while assuming others remain constant, and the current dataset lacks extensive temporal coverage. This limitation hinders the ability to discern between mere noise and actual causality, as well as to assess greater time lags regarding how growth shocks may translate into changes in crime rates. Consequently, no definitive conclusions can be drawn from this study. Nevertheless, the Google Open Buildings data offers remarkable granularity down to the household level, presenting exciting research opportunities in urbanization studies.

6 Conclusion

Previous studies have shown conflicting results on how organized crime impacts violence and economic outcomes in a region. This study proposes the use of novel quantitative informality measures to explain the variations and found significantly lower crime rates in police jurisdictions with high K-complexity in South Africa using the K-means clustering method. K-complexity has also been shown to be important in prediction of crime with KNN and random forest algorithms. Together, these analysis highlights the significant role of neighborhood informality and built environment characteristics in shaping crime patterns across South Africa's urban landscape. Our additional analysis using Granger causality tests suggests a potential relationship between urban growth and crime, though limited temporal data prevents definitive claims.

Understanding these relationships is crucial for informing more effective urbanization policies, especially those aimed at addressing housing crises and crime in developing countries within the Global South. These insights can also contribute to the development of more nuanced approaches to public safety, including community-centered policing and strategies for mitigating violence in urban settlements. Overall, these findings emphasize the need for in-situ urban policy interventions that address infrastructure gaps and informal settlement dynamics, particularly in rapidly urbanizing regions where traditional governance structures may be weak.

Data and Code Availability Statement

The data for this project is sourced from the South African Police Service, the Million Neighborhoods Project, OpenStreetMap, Google Open Buildings 2.5D Temporal Dataset and Copernicus DEM, all of which are open-source and available online. The full analysis code in Python is accessible through this Github repository.

References

Barcellos, C., & Zaluar, A. (2014). Homicídios e disputas territoriais nas favelas do Rio de Janeiro. *Revista de Saúde Pública*, *48*(1), 94–102. https://doi.org/10.1590/S0034-8910.2014048004822

Bettencourt, L. M. A., & Marchio, N. (2023). Million neighborhoods africa database.

Brelsford, C., Martin, T., & Bettencourt, L. M. (2019). Optimal reblocking as a practical tool for neighborhood development. *Environment and Planning B: Urban Analytics and City Science*, *46*(2), 303–321. https://doi.org/10.1177/2399808317712715

Cinnamon, J., & Noth, T. (2023). Spatiotemporal development of informal settlements in cape town, 2000 to 2020: An open data approach. *Habitat International*, *133*, 102753. https://doi.org/https://doi.org/10.1016/j.habitatint.2023.102753

Davis, M. (2006). *Planet of slums*. Verso.

Dugard, J. (n.d.). From Low Intensity War to Mafia War: Taxi violence in South Africa (1987 - 2000).

February, J. (2019). *State capture: An entirely new type of corruption* (tech. rep.). Institute for Security Studies. Retrieved April 18, 2025, from http://www.jstor.org/stable/resrep62992

Garrido, M. (2021). Reconceptualizing Segregation from the Global South. *City & Community*, *20*(1), 24–37. https://doi.org/10.1111/cico.12504

Google Research. (2024). Open buildings: A Global Building Dataset [Accessed: April].

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438. https://www.jstor.org/stable/1912791

Harvey, D. (1989). *The urban experience*. Johns Hopkins University Press.

Huchzermeyer, M. (2009). The struggle for *in situ* upgrading of informal settlements: A reflection on cases in Gauteng. *Development Southern Africa*, *26*(1), 59–73. https://doi.org/10.1080/03768350802640099

Irish-Qhobosheane, J. (2022). *Extortion or transformation? the construction mafia in south africa* (tech. rep.). Global Initiative against Transnational Organized Crime. Retrieved April 18, 2025, from https://globalinitiative.net/wp-content/uploads/2022/06/GITOC-Extortion-or-Transformation-The-construction-mafia-in-South-Africa.pdf

Koonings, K., & Kruijt, D. (Eds.). (2007). *Fractured cities: Social exclusion, urban violence and contested spaces in latin america* [Copyright © 2007 Kees Koonings and Dirk Kruijt]. Zed Books Ltd.

Kumar, A., Verma, A., Shinde, G., Sukhdeve, Y., & Lal, N. (2020). Crime Prediction Using K-Nearest Neighboring Algorithm. *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 1–4. https://doi.org/10.1109/ic-ETITE47903.2020.155

Lambrechts, D. (2012). The Impact of Organised Crime on State Social Control: Organised Criminal Groups and Local Governance on the Cape Flats, Cape Town, South Africa. *Journal of Southern African Studies*, *38*(4), 787–807. https://doi.org/10.1080/03057070.2012.749060

Lessing, B. (2021). Conceptualizing Criminal Governance. *Perspectives on Politics*, *19*(3), 854–873. https://doi.org/10.1017/S1537592720001243

Lombard, M., & Rakodi, C. (2016). Urban land conflict in the Global South: Towards an analytical framework. *Urban Studies*, *53*(13), 2683–2699. https://doi.org/10.1177/0042098016659616

López-Caloca, A., Martínez-Viveros, E., & Chapela-Castañares, J. I. (2009, April). Application of a clustering-remote sensing method in analyzing security patterns. In B. V. Dasarathy (Ed.). https://doi.org/10.1117/12.818911

Melnikov, N., Schmidt-Padilla, C., & Sviatschi, M. M. (2022). Gangs, Labor Mobility, and Devlopment. *NBER Working Paper*.

Oliveira, B. P. D. (2023). Descanse O Seu Gatilho: Illegal Markets' Hegemony and Its Peace Dividends.

OpenUp. (2022). Struggle for dignity in cape town's informal settlements: The facts [https://ismaps.org.za/desktop.html].

Sivaranjani, S., & Sivakumari, D. S. (2016). Crime Prediction and Forecasting in Tamilnadu using Clustering Approaches.

Soman, S., Beukes, A., Nederhood, C., Marchio, N., & Bettencourt, L. (2020). Worldwide Detection of Informal Settlements via Topological Analysis of Crowdsourced Digital Maps. *ISPRS International Journal of Geo-Information*, *9*(11), 685. https://doi.org/10.3390/ijgi9110685

UN-Habitat. (2022). World Cities Report 2022: Envisaging the Future of Cities. https://unhabitat.org/sites/default/files/2022/06/wcr_2022.pdf

Vigneswaran, D. (2014). The Contours of Disorder: Crime Maps and Territorial Policing in South Africa. *Environment and Planning D: Society and Space*, *32*(1), 91–107. https://doi.org/10.1068/d18311

Villarreal, A. (2021). Reconceptualizing urban violence from the global south. *City & Community*, *20*(1), 48–58. https://doi.org/10.1111/cico.12506

VPUU. (2022). Workshop in driftsands: Working with covid community leadership to stop land invasion in unsafe areas [Accessed: 2025-04-20]. http://vpuu.org.za/projects/

workshop-in-driftsands-working-with-covid-community-leadership-to-stop-land-invasion-in-unsafe-areas/

Wacquant, L. (2007). Territorial Stigmatization in the Age of Advanced Marginality. *Thesis Eleven*, *91*(1), 66–77. https://doi.org/10.1177/0725513607082003

Appendix

Table 5: T-Test Results: Cluster 2 vs Other Clusters

Comparison	t-Statistic	p-Value
Cluster 2 vs Cluster 1	-7.32113e+00	1.33058e-11
Cluster 2 vs Cluster 3	-8.03580e+00	1.53695e-13
Cluster 2 vs Cluster 0	-1.10065e+01	3.35744e-22
Cluster 2 vs Cluster 5	-1.23081e+01	7.31068e-28
Cluster 2 vs Cluster 6	-1.30054e+01	8.60419e-27
Cluster 2 vs Cluster 4	-1.08065e+01	8.91536e-24

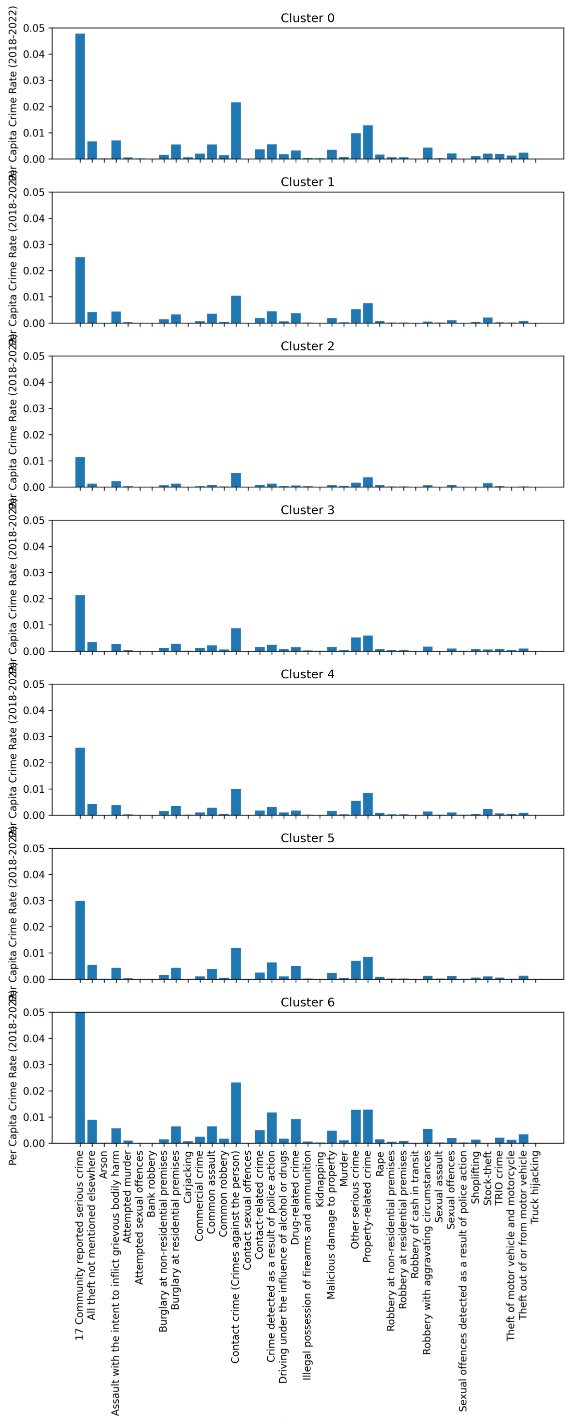

Figure 18: Crime by category by cluster.

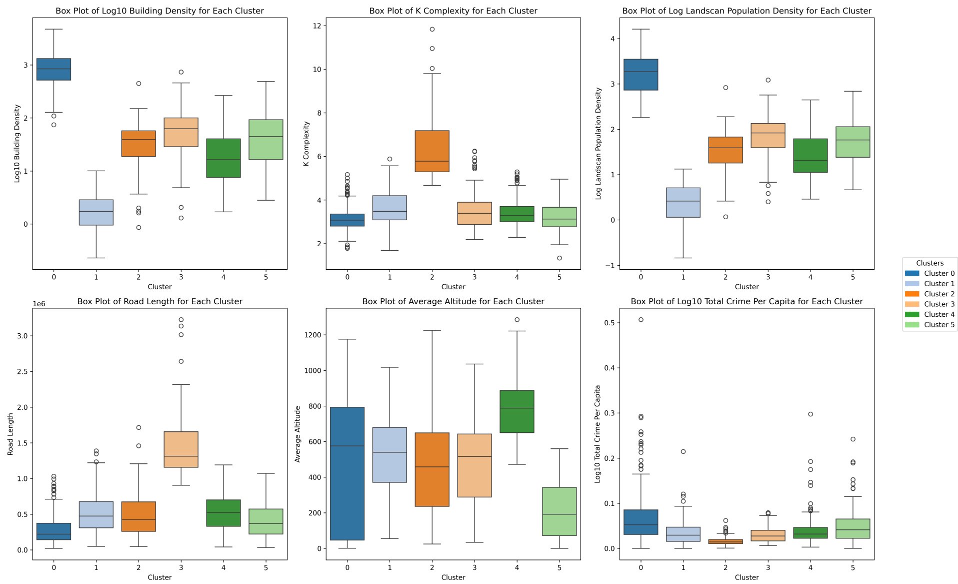

Figure 19: KNN results for k = 6.

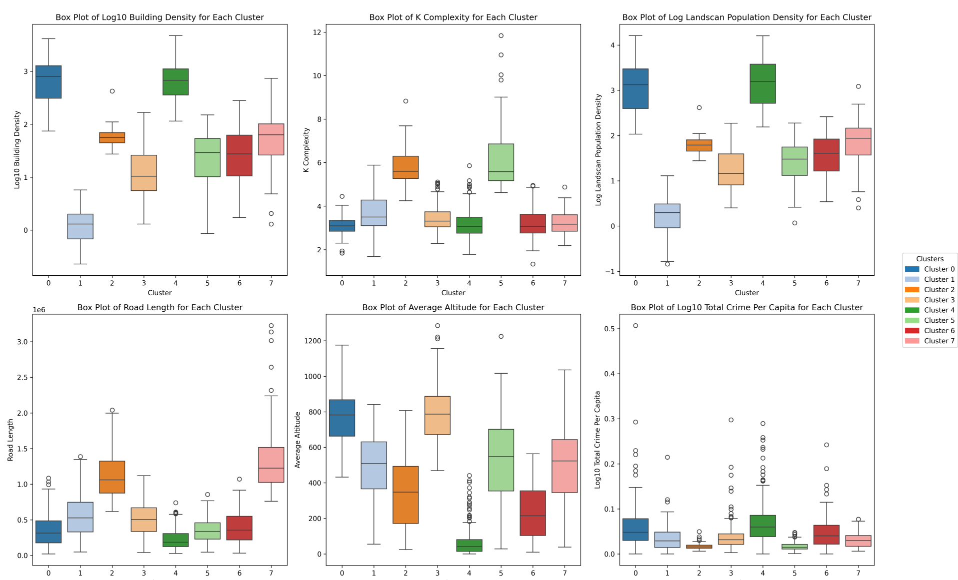

Figure 20: KNN results for k = 8.