University of Tartu
Faculty of Science and Technology
Institute of Ecology and Earth Sciences
Department of Geography

Master's thesis in Geoinformatics for Urbanised Society (30 ECTS)

Analyzing the relationships between crime and socioeconomic and spatial factors using random forest: a case study of Tallinn

Cheng-Wei Yu

Supervisors:
Prof. Evelyn Uuemaa
PhD Kadi Kalm
PhD Alexander Kmoch
MSc Jānis Zālīte

Tartu 2024

Abstrakt

Kuritegevuse ja sotsiaalmajanduslike ning ruumiliste tegurite vaheliste seoste analüüs otsustusmetsa abil: Tallinna juhtum

Kuritegevuse ruumilised mustrid ja selle sotsiaalmajanduslik taust on olulised teemad kuritegevuse uurimises. Selle töö eesmärk on ennustada otsustusmetsa mudeli abil kuritegude arvu igas 500m-ruudus Tallinnas ning tuvastada peamised kuritegusid soodustavad tegurid. Masinõppe mudelid ei seleta muutujate vahelisi põhjuslikke seoseid, kuid tõstab esile võimalikud korrelatsioonid, mistõttu tuleb kuritegevuse tegureid käsitleda Tallinna kontekstis. Töö tulemustest selgus, et kuritegevuste arvuga on kõige tugevam seos kaubanduslikel huvipunktidel. Need peegeldavad majandustegevuse ja inimeste koondumist, mis on võivad kuritegevust soodustada. Teiseks selgus, et avaliku korra rikkumiste arvu suurenemine oli seotud üüripindade arvu ja madalamasse sotsiaalmajanduslikku gruppi kuuluvate elanike arvuga.

Märksõnad: otsustusmetsa mudel, masinõpe, kuritegevuse ennustamine, mitme muutujaga analüüs

S230 Sotsiaalne geograafia

Abstract

Analyzing the relationships between crime and socio-economic and spatial factors using random forest: a case study of Tallinn

The spatial factors of crime and its socioeconomic background are important topics in crime research. This study uses a grid framework to represent various spatial, environmental, and socioeconomic factors across Tallinn in 500-meter grids. The study aims to predict the number of crimes in each grid cell through a random forest machine learning model and identify the main contributing factors. Machine learning models do not explain causal relationships between variables but highlight possible correlations, so crime factors need to be discussed within Tallinn's context. Among various types of crime, the factor of commercial locations shows the strongest relationship with the number of crimes. These reflect the concentration of economic activities, assets, and the gathering of people, which are important conditions for crime motivations. Secondly, factors such as the number of renters and the population with low socioeconomic status are associated with the number of crimes against public order.

Keywords: random forest model, machine learning, crime prediction, multivariable analysis
CERCS code: S230 - Social geography

Table of contents

1. Introduction

Urban crime is an important topic in many domains. This topic reflects how people constitute the essential values and understanding for the operation of urbanized societies, and crime is considered an obstacle to effective operation. Research on the topic of crime is multifaceted, covering a variety of aspects. With the element of opportunity and rational choice, crimes are committed after considering potential gains and costs (Iliyasu et al., 2022). This aspect also facilitated discussions in crime prevention by effectively utilizing environmental designs, penalties, and law enforcement to reduce opportunities and increase the risks of committing a crime. (Iliyasu et al., 2022; Jones & Fanek, 1997; Rand, 1982). The studies with social aspects are also crucial in exploring the structural factors that influence the formation of crime, e.g., the disadvantaged status based on poverty, inequality, segregation, and discrimination (Rand, 1982). Besides, urban morphology pays attention to the patterns and trends of crime's temporal and spatial distribution and the factors associated with crime. It explains how these influential factors are connected (Iliyasu et al., 2022). The impact of the unique characteristics of the Estonian capital city, Tallinn, on the spatial distribution of crime remains to be explored, particularly its overlapping segregation in terms of ethnicity, economy, occupation, and housing types that have persisted since the Soviet era. Even after regaining independence, the transformation has driven it towards a more polarized society.

Different hypotheses have been proposed to answer the relationship between spatial, environmental, and socioeconomic factors and crime, resulting in a diverse and thriving research landscape for analyzing crime, which also influences the development of crime prediction methods (Massey, 1995; Krivo et al., 2015; Cozens, 2011; Vogel & South, 2016; Tokey, 2023). Initially, crime prediction required significant effort and costs for data collection. Still, with the increasing availability of government open data and crowd-sourcing data, researchers can more easily conduct analyses based on local contexts, enriching research outcomes.

Research methods have also seen significant development in recent years. Kounadi et al. (2020) reviewed methods for predicting crime occurrence, e.g., Hotspot analysis using binary classifiers, which compare data layers to illustrate the correlation between them; Kernel-based estimation, which constructs curves extending outward from cores with decreasing intensity as distance increases, representing the influence and range of attributes of the sample points. This method involves selecting different numbers of core points, attribute weighting, and basic strategies. Machine learning is a popular approach that involves models learning features from sample data. The models then validate the results and determine optimal features and parameters through repeated iterations. Deep learning, on the other hand, is used for handling large datasets; it involves extracting deep features from sample data, followed by validation and model building (Kounadi et al., 2020)

The advantage of machine learning in handling multivariate data is that it allows researchers to explore the relationships between variables and adjust learning strategies to improve prediction accuracy. Also, interpretability remains a challenge in machine learning. Models can reveal possible correlations between data features, but translating these relationships into actionable crime prevention strategies still requires discerning the nature of these relationships (Mandalapu et al., 2023).

The study explores the relationships between crime events and socioeconomic and spatial factors in Tallinn City. Here posed two research questions:

1. What spatial and socioeconomic characteristics are most likely associated with crime incidents? Is there a difference when considering the various types of crime?
2. How well can the random forest model predict the spatial distribution of crime? How can the connections among the factors in the context of Tallinn be explained?

2. Theoretical overview

The nature of crime is regarded as a violation of the common rules, undermines people's sense of safety, and prevents them from engaging in productive social activities (Iliyasu et al., 2022). Most people may not experience true crime themselves. Still, the loss of public awareness might be more influential when the fear of crime associates public places with potential crime sites and stops citizens from engaging in their public and private lives (Rand, 1982). Several perspectives help to think about the causes of crime and the possible factors involved in crime. Social disorganization theory notes the weak links between individuals, institutions, and the communities in which they live, reducing the degree to which individuals are constrained by society. The loss of social connections could make individuals involved in criminal activities or make them vulnerable enough to become victims. The theory of daily routine suggests that crimes are always present in society, and the level of opportunities for crime occurrence changes dynamically along with profitability and costs (Graif et al., 2014; Ceccato, 2009).

2.1 Spatial factors influencing crime

Suppose we view crimes as a deviation from the expected well-functioning urban environment or as areas where social norms are not met. In that case, it may prompt us to consider ways to strengthen the urban environment or address any deficiencies in creating a high-quality urban environment (Iliyasu et al., 2022; Jones & Fanek, 1997) and result in the following discussions: (1) In the field of urban planning: How land use and parcels are connected to crime activities; (2) In the field of urban design and architecture: How public spaces and human constructions are linked to criminal activities; (3) in the field of geography: How spatial patterns of crime are embedded in society and time (Iliyasu et al., 2022).

Land use is one of the most essential ways to construct the urban form and is fundamentally intertwined with all aspects of daily life. In recent decades, the New Urbanist narratives encourage urban planners and city governments to consider high-dense and mixed land use to accommodate people's city life better and utilize public resources in a more concentrated and efficient manner. Jane Jacob's 'Eyes on the street' depicted the ideal urban social life in that people attend to one another with a strong sense of ownership of the street by the people and keep crimes away from communities (Cozens, 2011; Johnson & Bowers, 2010). Jacob's concept is welcomed by many city governments not only for the ways of land use but also for street design for seeking more walkable and permeable neighborhoods (from Johnson and Bowers, permeability is 'a construct devised to reflect and summarize how the street network influences pedestrian and vehicular movement') (Cozens, 2011; Johnson & Bowers, 2010).

However, Cozens (2011) examined that good practices in urban planning may not always be good practices in crime prevention, such as a permeable neighborhood is good for inhabitants to commute and access the nearby public amenities, but it also leads to higher chances of being penetrated by burglary; mixed-use brings commercial attractions to residential areas and exposes valuable private properties to the potential offenders; high density in population means high alienation and complexity and bystanders' passive interventions which makes crime occurring more easily. The above may not completely dismiss the value of New Urbanism. However, it is still hard to tell what types of land use could be resistible to crime occurrence without further inspection.

Rand's study (1982) gave a possible explanation that the area primarily devoted to family housing has relatively lower crime rates than mixed-use areas. This is because neighboring families tend to share the same interests and needs in communities and are more likely to attend to their community. For the same reason, high-rise housing and large buildings have higher crime rates due to the lack of familiarity between residents (Rand, 1982). Considering the population and urban fabric differences, Tokey (2023) argued that single-residential areas might be safer in some cities than mixed-use areas. However, in some cities, high-density residential areas may be more hazardous. Favarin (2018) found that the social disorganization theory effectively accounts for the concentration trends of burglary and robbery in Milan, Italy. The study revealed that these types of crimes are significantly reduced in street segments with mixed-use properties. This can be attributed to the fact that burglaries are more likely to take place in residential areas and that mixed-use street segments benefit from greater informal social control. These observations largely agree that maintaining adequate social interaction contributes to community safety. However, there is no premise for what correlation between population size and neighborhood scale that can support lower crime rates.

Another aspect of analyzing urban crime is how offenders physically come into contact with the potential targets or how offenders can escape from the crime scenes as fast as possible. The issue of street permeability has become central to considering how to deliver less crime and more sustainable urban design (Cozens, 2011). The dispute in permeability revealed two different perspectives in terms of crime prevention. The social control based on the 'Eyes on the street' defers crime commitment; another is blocking potential risks outside the community can reduce crime effectively (Cozens, 2011; Johnson & Bowers, 2009). Conversely, Johnson and Bowers' study (2009) argued that increased permeability is associated with elevated burglary crimes. Their study analyzed the street configurations and took an approach to measure the types of roads and their intended functionalities. The analysis indicated that major roads were more associated with burglary, and private roadways and cul-de-sacs were less influenced by criminal activities (Johnson & Bowers, 2009).

In Tallinn, most robberies occur in the center and non-residential areas; they are also observed outside the city center, especially along the major roads, stations, and local centers in mixed-residential land use areas. Also, higher robbery rates are found in census tracts with lower population density and weaker social control than in denser areas. (Ceccato and Oberwitter, 2008; Ceccato, 2009).

Researchers have different observations on how spatial and environmental factors influence crime. Generally, by examining the types of roads and land use, studies identify specific locations' value, chances for offenders to arrive and leave, chances for offenders to encounter targets, and the conditions under which offenders are detected or ignored, thereby establishing connections with criminal incidents.

2.2 Socioeconomic factors influencing crime

On the other hand, some studies tend to place crime within the social context of age, race, income, and privilege or deprivation (Ceccato, 2009; Vogel & South, 2016; Hirschfield et al., 2013). In the United States, Massey (1995) conducted a study on the high rates of crime, poverty, and segregation faced by Black ethnic groups. The study highlighted a series of interconnected factors that contributed to increased violence and stated that the social process of Black segregation was due to the exclusion from the White dominant neighborhoods and estate markets since the Black groups had no way to escape from the poverty and crimes in the segregated neighborhoods, adapting the niche of violence became necessary (Massey, 1995). Krivo (2015) studied the cities in the United States and pointed out some connections between crime and socioeconomic characteristics, e.g., segregated and disadvantaged neighborhoods have higher crime rates. Minorities living in unstable residential conditions also experience higher crime rates. Immigration also has a stronger relationship with property crime than other factors (Krivo et al., 2015). Vogel and South's study (2016) focused on neighboring areas other than young offenders' neighborhoods. They found that more crimes were reported in the wealthier neighborhoods if they were located near the neighborhoods with lower socioeconomic conditions. Considering the incentives of offenders, Rand (1982) similarly discussed that adolescents tend to commit affective-driven crimes (less profit-oriented or less considered), which often happen near their homes. Conversely, crimes committed through rational calculation usually maintain a certain distance from the offender's residence. Livingston et al.'s (2013) study in Glasgow, Scotland, investigated the effects of diversifying housing tenure. Diversity is generally seen as a favorable policy that promotes local connections, prevents polarization, and encourages community integration across various aspects such as age, ethnicity, economics, and housing tenure. Livingston et al. (2013) argued that having a mix of different housing tenures can lead to lower crime rates. This argument is supported by the idea that communities with diverse housing tenures are more likely to share internal resources, leading to positive development for different social groups and an equal and effective distribution of external resources. However, the downside is that too much diversity might reduce community integration, potentially leading to conditions that favor social disorganization.

In the case of Estonia, Ceccato and Oberwittler (2008) compared Tallinn with the Western European city Cologne; there are not too many fundamental differences in crime spatial patterns between them. However, the research provided a different finding compared to earlier research. Without considering land use, the importance of social disorganization (referring here to the proportion of different ethnic groups) for predicting robbery is moderate; after incorporating land use variables (referring here to the locations of pubs and bars), the inferred importance of social disorganization increases significantly (Ceccato, 2009). Compared to the United States and Western Europe, where ethnic segregation is often closely associated with crime, the history of segregation in Tallinn differs from that in Western Europe. In contrast to other European cities, where ethnic minorities mainly immigrated recently from non-European countries, the ethnic minorities in Tallinn are relatively homogeneous and have been living in Estonia since the Soviet era (Tammaru et al., 2016). These backgrounds make the connection between crime and segregation in Tallinn unique and worthy of further analysis.

Socioeconomic factors are closely related to local contexts, such as population composition, age, ethnicity, and neighborhood or community characteristics, and are often observed in crime studies. Factors like poverty, disadvantage, and segregation are often positively correlated with crime. Moreover, socioeconomic factors may not exist solely but can influence each other, affecting social ties that either enhance or deter opportunities for crime.

2.3 Spatial modeling of crime

Early research in criminology has already stated that crime does not occur randomly, and various studies, from the environment, society, and other perspectives, have all tried to explore crime patterns. Due to the development of analysis techniques and methods, establishing predictable and verifiable models has become an important part of crime research (Kounadi et al., 2020).

Ceccato (2009) used aspects of social disorganization and daily routine to analyze the distribution of various types of crime in Tallinn. The study employed standardized offense ratios as the dependent variables of an ordinary least regression model and identified the potential incentives for multiple types of crimes and the spatial trends of where these crimes occur.

Some machine learning methods have been implemented in crime prediction. Random forest is also a widely used machine-learning model in crime prediction (Kounadi et al., 2020). It is a non-parametric machine learning method that is well-known for its efficacy and accuracy. Random forest is an ensemble of trees algorithm that predicts the results by extracting the most influential features and splitting criteria. It uses the bagging method for random sampling to gradually increase the accuracy and generalization ability of the model (Breiman, 2001); it further benefits from two of its features: (1) the randomness denotes the model randomly selecting subsets among predictors that prevent overfitting, and (2) forest denotes the model growing many trees to identify which parts are overlapped and impactful (Reade et al., 2019). Compared with Linear-regression-based models, which usually assume data has a pattern of similar residual distribution and data value distribution and is sensitive to multicollinearity, Tree-based algorithms are non-parametric methods without making assumptions; they could contribute to good prediction even with a ‘small perturbation’ on the dataset and helpful in handling data types with nonlinear relationships. (Alves et al., 2018; Lotfata, 2023).

The study of Xia et al. (2021) utilized the random forest classification to predict the drag crime hotspots. It used socioeconomic and built environment factors with spatial and temporal variables for prediction and achieved remarkable accuracy (90.7%) in prediction; it also identified some important factors, e.g., vacant lots, vacant buildings, low-wage workers, the share of high-level education, and the share of the population in working ages. Wheeler and Steenbeek (2020) implemented a random forest model to identify risky terrains associated with crimes. They explained the methods for interpreting the interaction of variables and threshold effects by examining the local effect from a single factor on the model. Random forest can be further incorporated with spatial features. Random Forest Spatial Interpolation (RFSI) incorporates factors such as the values and distances of the nearest training data in the prediction results (Sekulić, 2020). The two-point ML model extracts the spatial information of variables. Each attribute value of each sample is bidirectionally calculated with every other sample's attribute value (as a two-point distance), establishing a semivariance matrix that is then segmented into training and testing data inputs for the model (Gao et al., 2022). Geographic Random Forest (GRF) transforms neighboring sample values into a spatial weighted matrix and iteratively tests to find the optimal number of neighbors (Lotfata et al., 2023).

3. Data and methods

3.1 Study area

The study area of Tallinn is the capital city of Estonia, with 458,398 inhabitants in an area of 159.39 square kilometers (Tallinn city government, 2023). This study categorized the eight administrative districts of Tallinn into four conceptual zones (Figure 1) based on previous research on the distribution of ethnicities and building types (Tammaru et al., 2016; Leetmaa et al., 2018).

In the city center and inner city, including Kesklinn, Kristiine, and Põhja-Tallinn, intense gentrification is occurring in some neighborhoods, with high socioeconomic status residents replacing those of lower socioeconomic status (Leetmaa et al., 2022).

Historically, the city center experienced community decline, leading to population outflow to the periphery areas; however, in recent decades, affluent groups have returned to the city center and areas previously occupied by factories and military facilities, introducing new high-priced developments (Tammaru et al., 2016; 2021). This zone is also a political and commercial center and a popular place for tourism (Tallinn city government, 2022).

In the zone of panel housing estates, these areas include Mustamäe, Väike-Õismäe in Haabersti, Lasnamäe, and Pelguranna and Kopli in Põhja-Tallinn. Some areas face decline, while some areas, like Mustamäe, are more successful in attracting new developments and residents with high socioeconomic status (Leetmaa et al., 2018). Lasnamäe is the district with the largest population and is also a district where the Russian-speaking population is significantly overrepresented (Tallinn city government, 2022). These areas have recently shown higher degrees of aging.

Outer city zones, including Pirita, Nõmme, and Haabersti, traditionally house residents of high socioeconomic status due to their extensive green spaces and low-density housing (Tammaru et al., 2016).

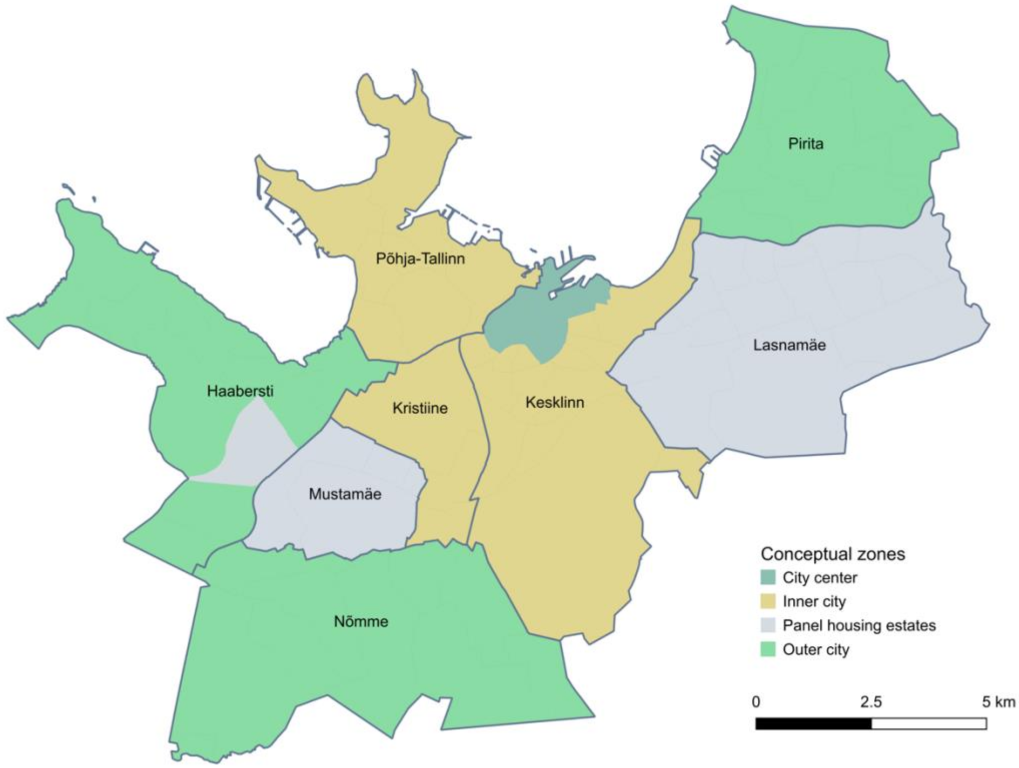


Figure 1. Tallinn and administrative districts and conceptual zones.

Tallinn constantly faces new opportunities and challenges as it struggles with its historical legacy. The city is vibrant, with businesses connected to other European countries (Tallinn city government, 2022). New commercial and residential development projects have emerged, driven by profit incentives (Tammaru et al., 2016). This creates valuable job opportunities, commerce, and entertainment in the city center and revitalizes previously old, underused, or abandoned areas (Ruoppila, 2007; Tammaru et al., 2016). Meanwhile, non-Estonian groups, particularly Russian-speaking groups, which are ca. 31% of total inhabitants in 2023 (Tallinn city government, 2023), have become increasingly concentrated in certain Soviet-era block housing, and their socioeconomic status has also declined (Leetmaa et al., 2018).

The segregation in Tallinn originated during the Soviet era when the population and urbanization were increasing. Block housing, public services, and education were built to meet industrial or governmental needs. Russian-speaking immigrants from other Soviet countries settled together to achieve economies of scale. Despite numerous new block housing developments, they were insufficient to accommodate the rapidly growing population. As a result, Estonians were often neglected by block housing supply and generally preferred traditional family homes in older neighborhoods (Tammaru et al., 2016). Segregation persisted even after the end of the Soviet era. Eastern European countries have generally undergone social, political, and economic transformations, leading to reformation in all aspects. During the transition from a state-planned economy to marketization, the relatively equal but static social and economic activities in the Soviet period began to be led by highly educated and competitive social groups. This shift results in polarization of the occupations, lifestyles, and housing patterns. Many European cities, including Tallinn, have consequently faced rapidly increasing levels of segregation (Tammaru et al., 2021).

The restitution and privatization of housing in Estonia led to significant changes in property values, with large Soviet-era apartment complexes decreasing in value. At the same time, single-family homes saw an increase in estate value due to the recent market-oriented economy (Tammaru et al., 2016). In 2000, the privatization of housing ownership reached 98%, and only 2% was owned by governments; state-led urban planning and housing policies were no longer effective, giving way to profit-oriented development plans by private enterprises (Tammaru et al., 2016). In privatization, houses became commodities that could be exchanged on the market. At the same time, due to economic growth, new estate projects targeting residents with high socioeconomic status became popular (Tammaru et al., 2021). The drastic changes in ownership also altered the responsibilities regarding housing management. Private house owners bear greater management and maintenance pressures (Leetmaa et al., 2018). After 2004, protective measures during the privatization transition phase were lifted, and the subsequent economic crisis from 2008 to 2010 reduced households' ability to afford housing expenses. During this period, those people entering the housing market were younger generations born in the late Soviet era or after independence, who did not benefit from the previous wave of privatization and could not afford high-priced housing. As a result, they could only choose Soviet-era apartments or become renters (Tammaru et al., 2016). All these economic and ethnic inequalities shape the unique urban fabric of Tallinn.

3.2 Data

The crime data from the Estonian Police and Border Guard Board was classified and used as the dependent variables. The covariates used the roads, land use, points of interest (POIs), green spaces, housing, and demographic and socioeconomic data.

Crime data

Crime data was acquired from the Estonian Police and Border Guard Board's (2023) crime incident statistics. The data consists of the jurisdiction of the incidents, types, violated places, violated laws, loss, prevention methods if they exist, and the date and time of the incidents from 2018 to 2022. Crime data was exported from the police procedural information system and updated weekly. This study excluded data from 2023 due to changes in data collection methods and incomplete data when the study began.

The location of each incident was anonymously recorded within a 2,500 square meter range (500-meter grid) in urban areas and a one square kilometer range (1000-meter grid) in city outskirts, suburban, or rural areas. To avoid the complexity of the model, crimes were classified into two types: crimes against public order and crimes against property (Table 1).

Solely considering the number of all crime incidents, crimes against property are the most numerous, so the spatial distribution of this type is similar to that of all crimes (Figures 2 and 3), where hotspots in the inner city and the housing estates can be identified. The overall number of crimes against public order is relatively small. This type of crime can be found in the city center and spreads to the surrounding areas, forming a belt-shaped hotspot covering the inner city and the housing estates (Figure 4).

Table 1. Crime variables.

	Crimes against public order	Crimes against property	Total number of crime incidents
Original types	• Vandalism • Public gathering • Violate public order • Conflict / arbitrary • Litter • Drug case	• Pocket robbery • Robbery • Bike theft • Keep / sell stolen property • Metal theft • Theft • Forestry theft • Motor theft	Incl. crimes against public order and crimes against property.
Number	6,104	47,548	53,652

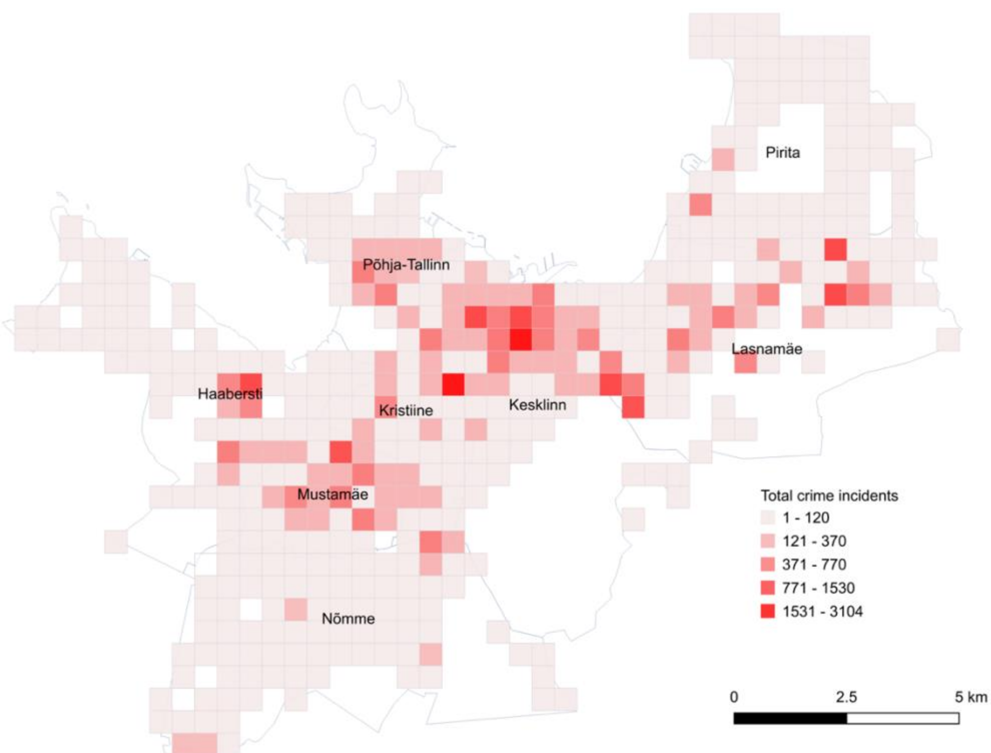


Figure 2. The spatial distribution of all crime incidents.

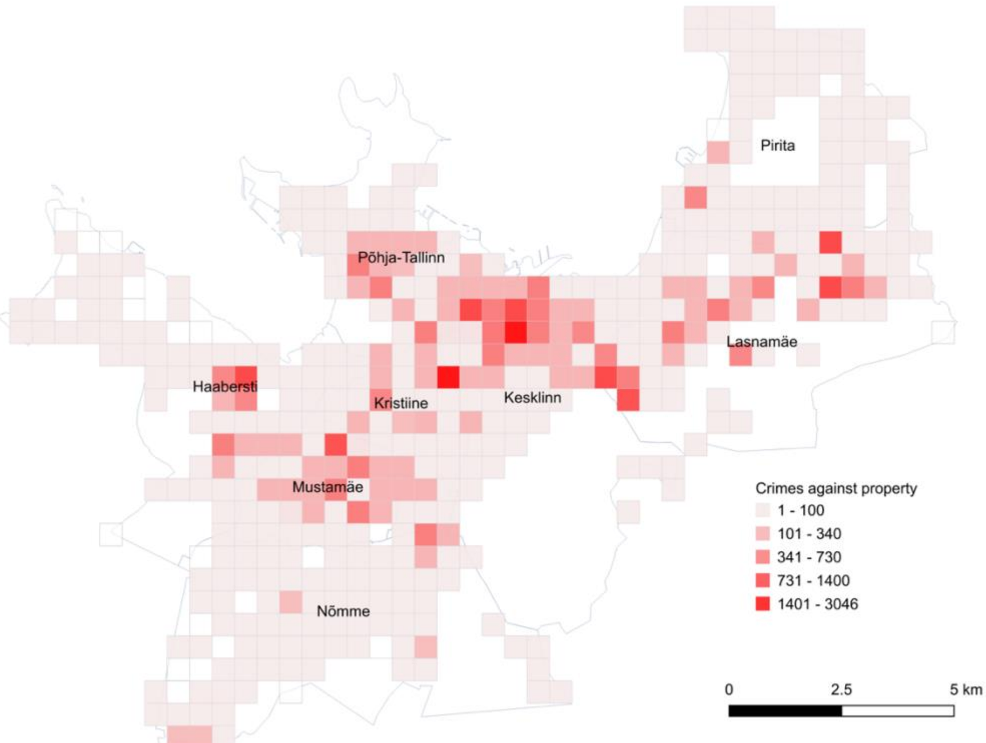


Figure 3. The spatial distribution of the crimes against property.

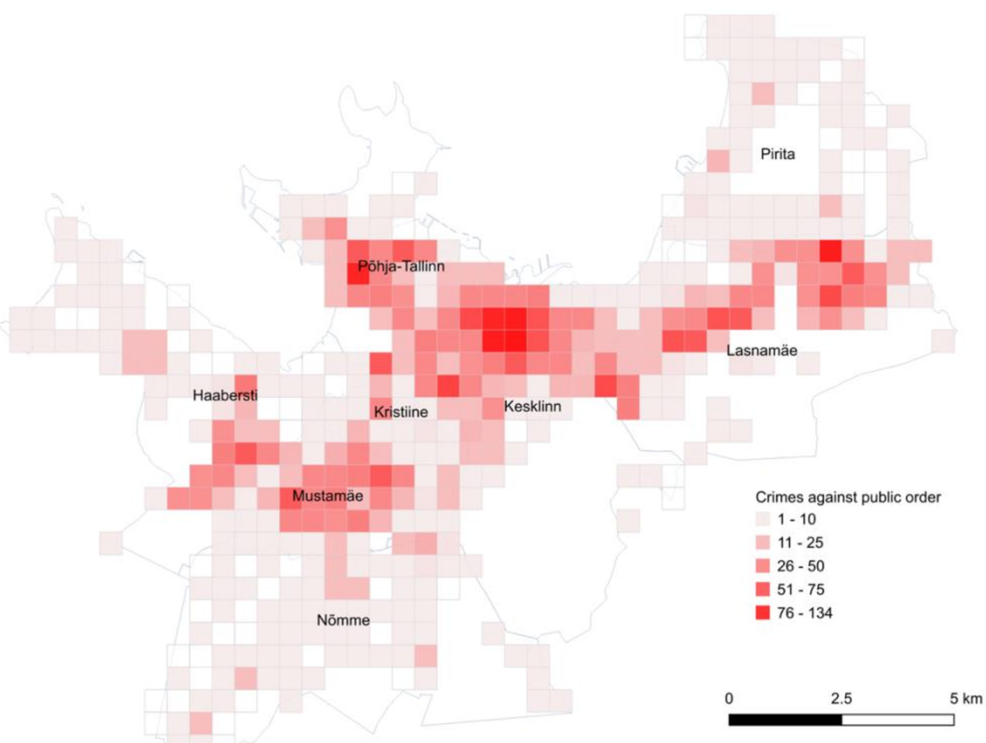


Figure 4. The spatial distribution of the crimes against public order.

Roads

The Estonian Topographic Database (ETAK) maintained the road data and released it on January 2, 2024 (Estonian Land Board, 2024). This study referred to Johnson and Bowers' (2009) study and applied a simplified classification (Table 2). Since no information was available on private roads in the data, this class in the original research was disregarded in this study. Additionally, instead of weighting based on the class of roads and the number of connections to other streets, the length of each type of road was calculated as an independent spatial variable.

Table 2. Classification of Roads.

	Trail	Neighborhood level	City level
Consist of	• Track and path • Local road	• Pedestrian and bicycle route • Street	• Main road • Basic road • Secondary road • Connecting road
Description	Internal connectivity	External connectivity	External and city-wise connectivity

Land use

Land use data was acquired from Urban Atlas Land Cover/Land Use 2018 (The Copernicus Land Monitoring Service, 2018). It is maintained by the Copernicus Land Monitoring Service and contains 788 functional urban areas with more than 50,000 inhabitants. The map layer consists of 17 types of land use and cover. In this study, four major urban atlas types with the highest proportion in the study area were selected, and their respective area was calculated in square meters for each grid cell (Table 3). These different types of land use help to identify the primary forms of commercial, residential, road, and other uses on the plot. Some of these types indicate the proportion of buildings on the plot and the ratio of primary and other land uses within the plot (The Copernicus Land Monitoring Service, 2023).

Table 3. The chosen types of urban atlas.

	Discontinued dense urban fabric	**Discontinued medium dense urban fabric**	**Industrial, commercial, public, military and private units**	**Continuous urban fabric**
Description	Residential buildings, roads and other artificially surfaced areas.	Residential buildings, roads and other artificially surfaced areas. The vegetated areas are predominant, but the land is not dedicated to forestry or agriculture.	At least 30% of the ground is covered by artificial surfaces. More than 50% of those artificial surfaces are occupied by buildings and / or artificial structures with non-residential use, i.e. industrial, commercial or transport related uses are dominant.	Predominant residential areas with a high degree (more than 80%) of soil sealing, incl. residential areas in city centers, and central business districts.

Green spaces

Green space data was maintained by the Estonian Topographic Database (ETAK) and released on January 2, 2024 (Estonian Land Board, 2024). Data was extracted from the following map layers: (1) green area and wasteland (wasteland was excluded from the further process for its relatively small portion in the study area), and (2) forest and shrub. The area in square meters in each grid cell was then calculated (Figure 5).

Figure 5. The distribution of green space.

POIs

Points of Interest (POIs) data was produced by the Open Street Map project. The POIs dataset (the version released on January 2, 2024) used in this study was acquired from a branch repository of Geofabrik (2024). This study reclassified POIs into four categories to improve interpretation (Table 4). A detailed list of POI types can be found in Appendix 1.

The commercial, public service, and camera surveillance POIs are all prominently located in Kesklinn (Figures 6, 7, and 9), especially the locations of surveillance cameras, most of which are in Kesklinn, with a few scattered in other districts. In contrast, recreational POIs are distributed more dispersed with a lower concentration and presence in the outer city zone (Figure 8).

Table 4. Classification of POIs.

	Commercial	Recreational	Public service	Camera surveillance
Number of POIs	2,563	405	2,581	3,912
Description	Shops restaurants, or accommodations on payment.	Places for tourism, cultural, outdoor activities, and sport.	Public services, e.g., schools, hospitals, art and cultural sites.	Locations setup with surveillance devices

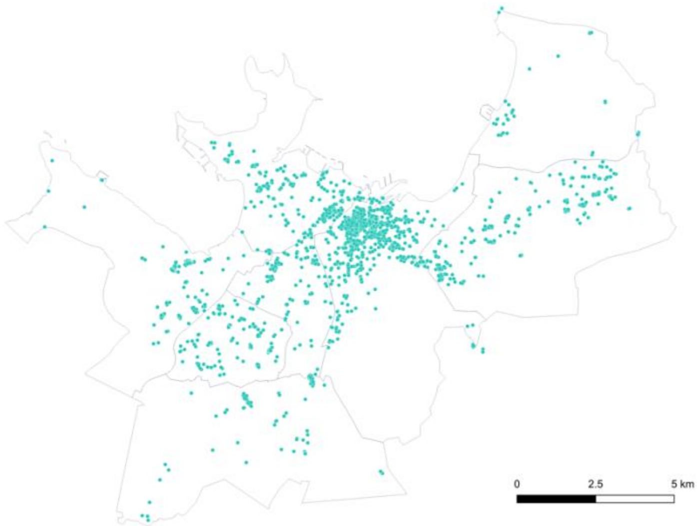


Figure 6. The spatial distribution of the commercial POIs.

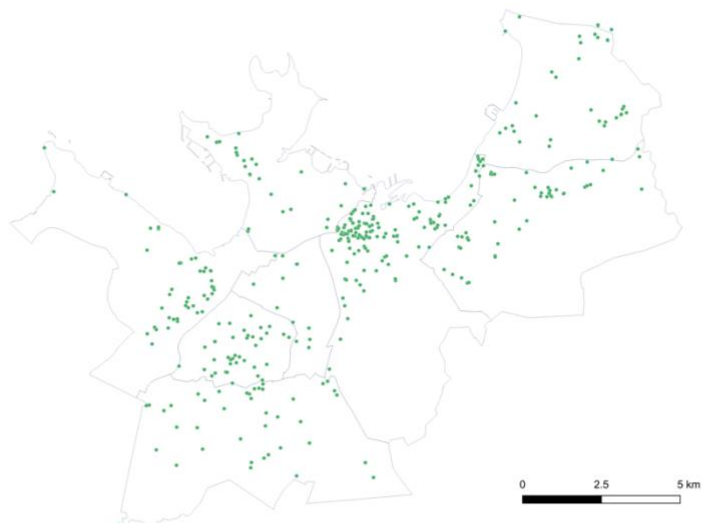


Figure 7. The spatial distribution of the recreational POIs.

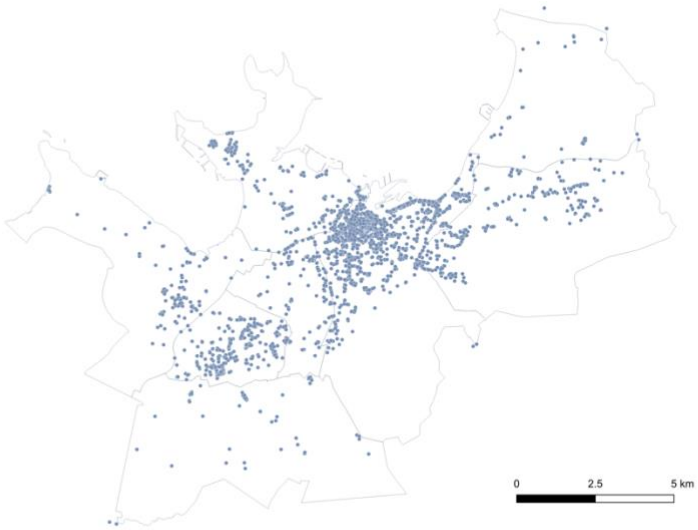


Figure 8. The spatial distribution of the public service POIs.

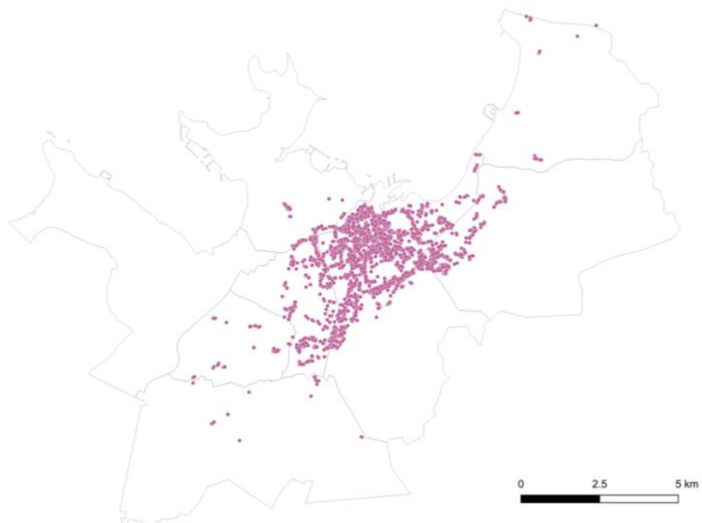


Figure 9. The spatial distribution of camera surveillance points.

Housing types

The Estonian Land Board maintained housing type data. It integrated data from the Estonian topographic database (ETAK), address database (Aadressiandmete süsteemi), and housing registration data (Ehitisregistri). The data was acquired in December 2023 and includes vector layers, housing types, and building codes for residential properties (Estonian Land Board, 2023). Data layers were further classified based on the common architectural forms in Tallinn, such as standalone family houses (Figure 12), apartment houses (further divided into small-scale and large-scale according to the number of units) (Figures 10 and 11), auxiliary houses (Figure 13), and other types (Figure 14), as shown in Table 5.

Table 5. Classification of housing types.

	Small-scale apartment house	**Large-scale apartment house**	**Standalone house**	**Auxiliary house**	**Others**
Consist of	• One-apartment residential buildings • Residential buildings with two apartments	• Residential buildings with two or more apartments • Residential buildings with three or more apartments • Other residential building with three or more apartments	• Terraced house • Terraced house or semi-detached house section • Detached house	• Cottage, garden house	• All except the left types
Number of buildings	941	8,817	16,628	922	2,878

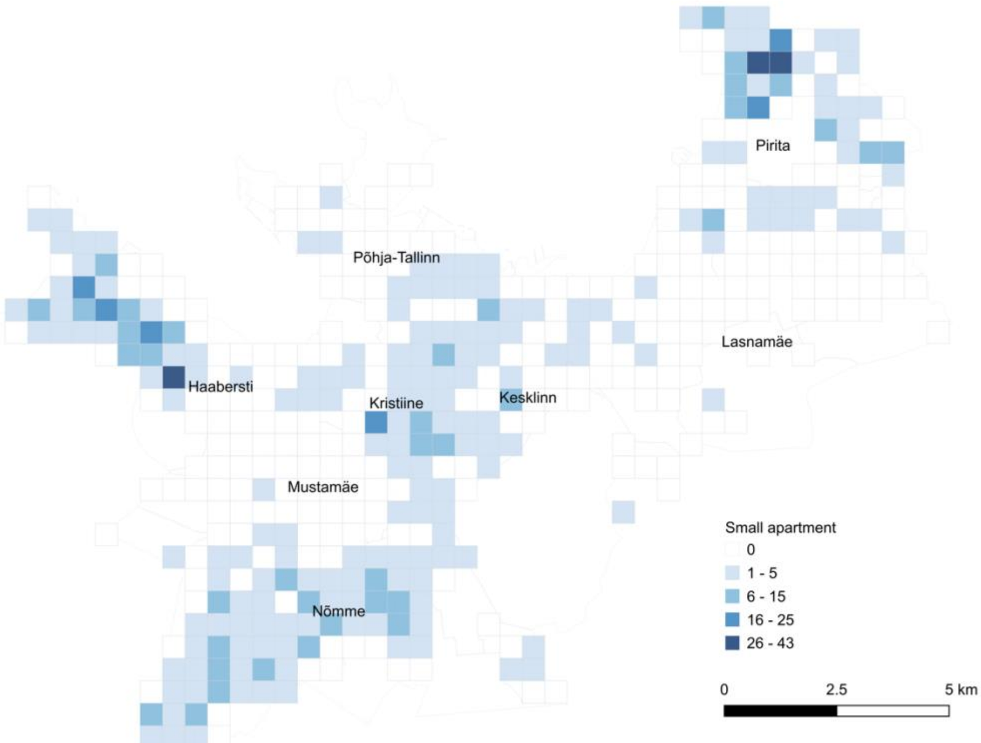


Figure 10. Small-scale apartment houses are distributed mainly in the inner city and outer city zones.

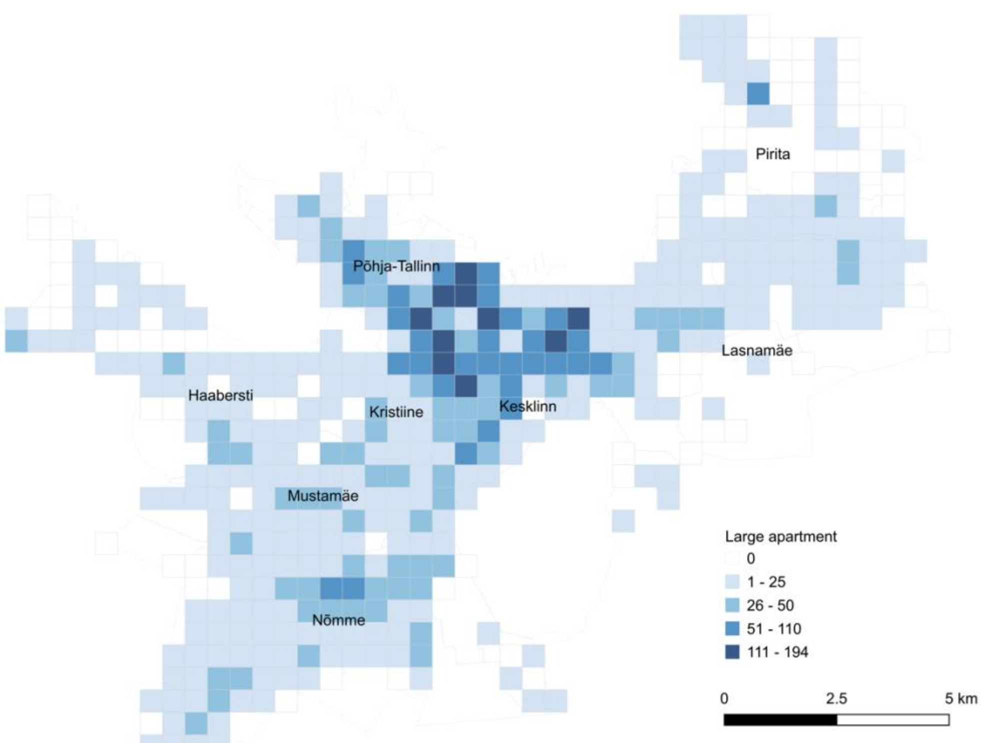


Figure 11. Large-scale apartment houses distributed in the inner city significantly and secondly in Nõmme.

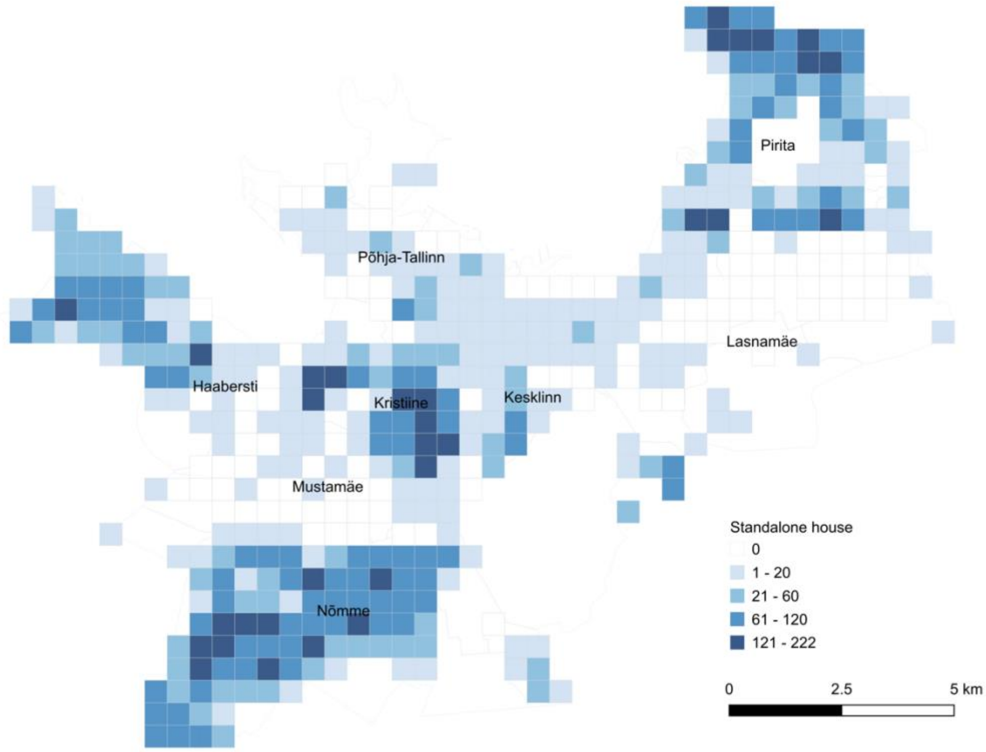


Figure 12. Standalone houses are in the outer city zone.

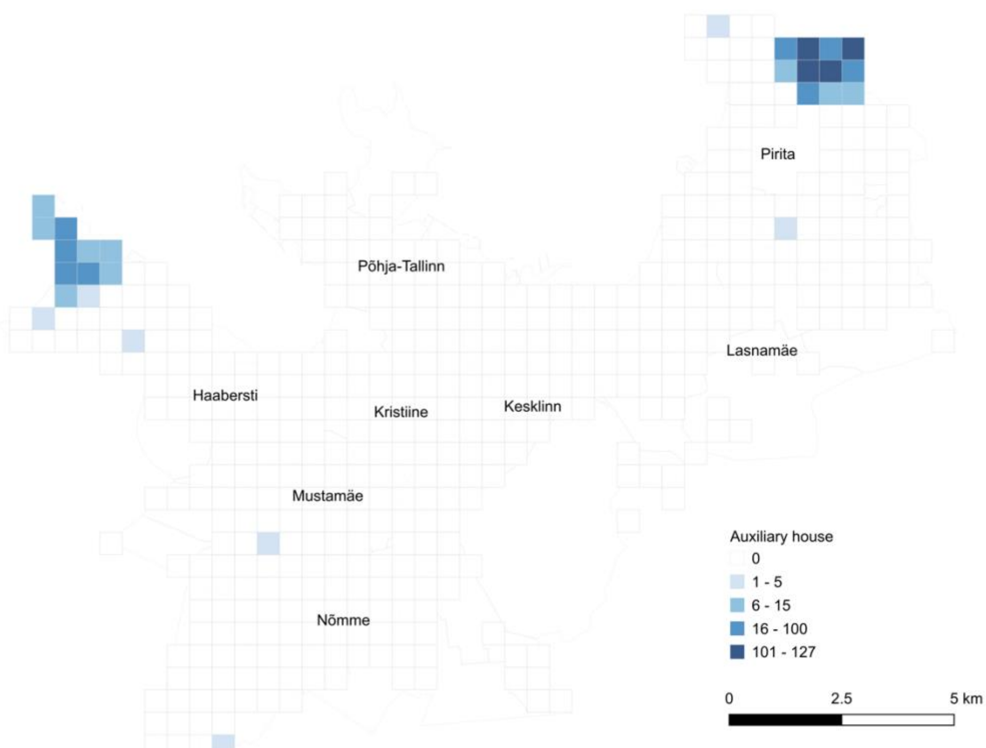


Figure 13. Auxiliary houses are concentrated in some areas in Pirita and Haabersti.

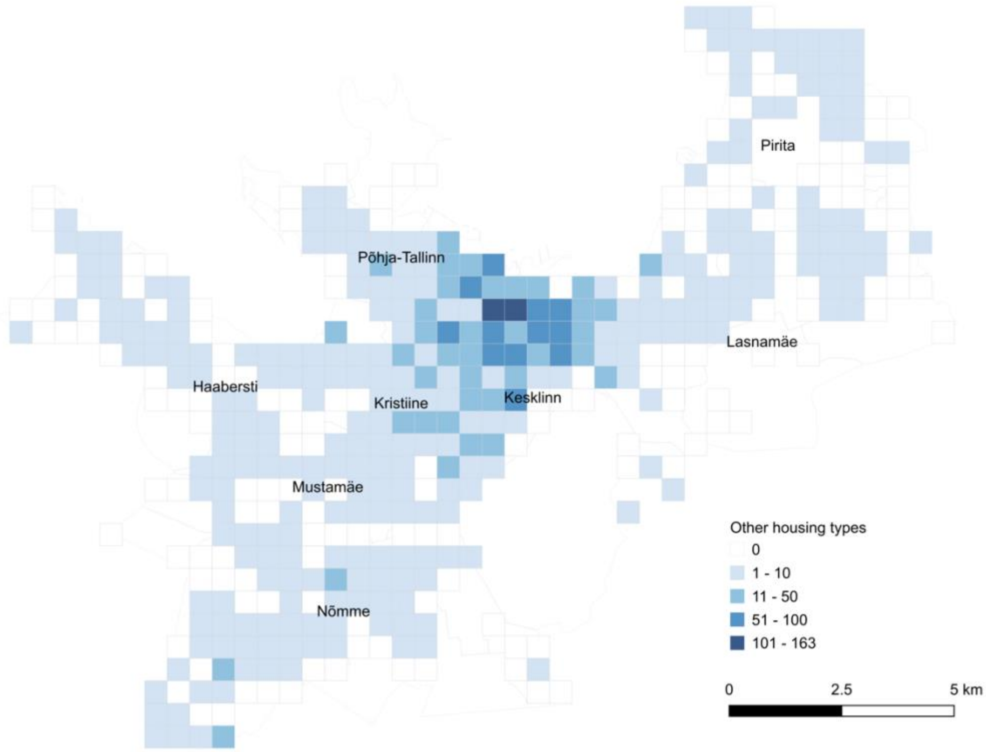


Figure 14. Other types show that the trend is centered in Kesklinn and scattered in different districts.

Demographic, socioeconomic, and housing data

The demographic, socioeconomic, and housing data were collected from the National Population and Housing Census 2021. The census provides aggregated data on population by age group (in 10-year intervals from 0 to over 90), population by ethnic group (Estonian and Russian native speakers), population by socioeconomic status (categorized as low, medium, and high according to the International Standard Classification of Occupations), population by residential building construction year (1945 and earlier, 1946 to 1990, 1991 and later), and population by housing tenure (house owner or renter) for each by a 500-meter by 500-meter census tract. This dataset was obtained through the Center for Migration and Urban Studies at the University of Tartu.

The distribution of high-status and low-status rates is complementary. High-status populations are overrepresented in both the inner city and the outer city (Figure 15). In contrast, low-status populations are overrepresented in the panel housing estate zone and some parts of Põhja-Tallinn of the inner city (Figure 16). The distribution of house owners' and house renters' rates are also complementary. Owner-occupied housing is prevalent in most parts of the city, especially in the outer areas, followed by panel housing estates (Figure 17). The rates of house renters in the inner city and sporadically in panel housing estates are higher. Some areas within the panel housing estates have exceptionally high renter rates (Figure 18).

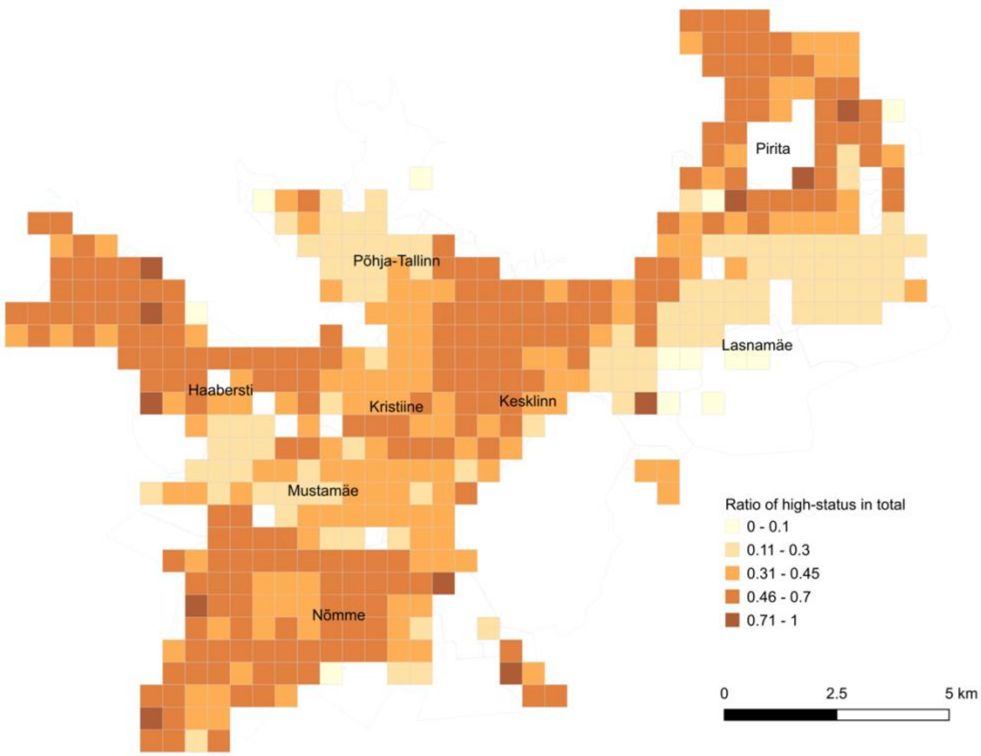


Figure 15. The distribution of the high-status population rate across all census tracks.

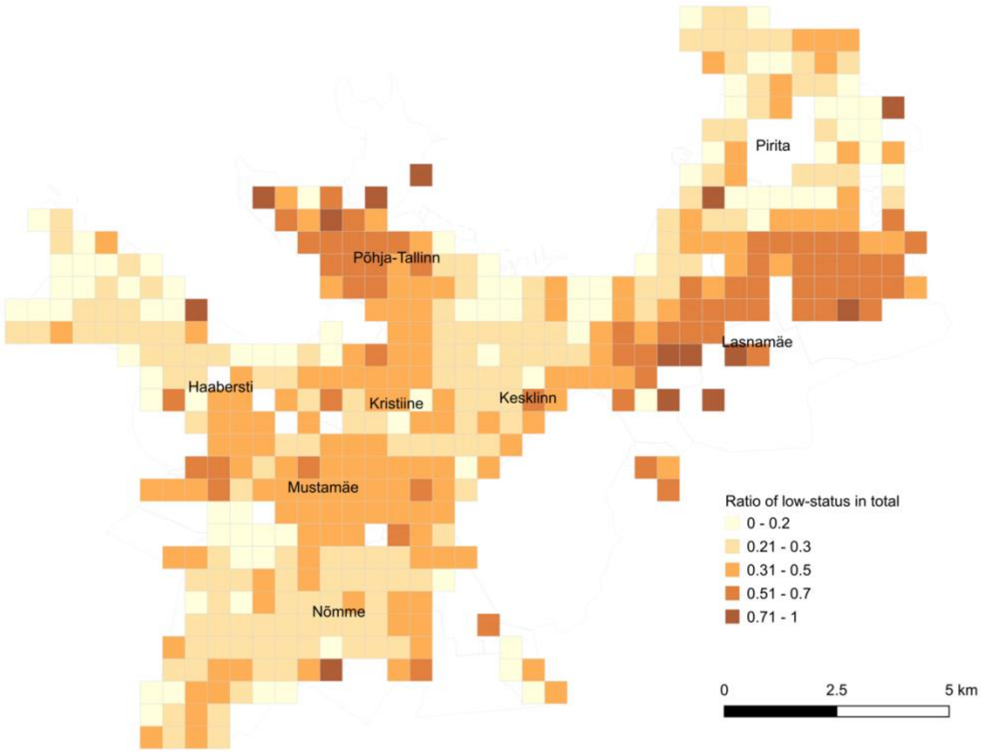


Figure 16. The distribution of the low-status population rate across all census tracks.

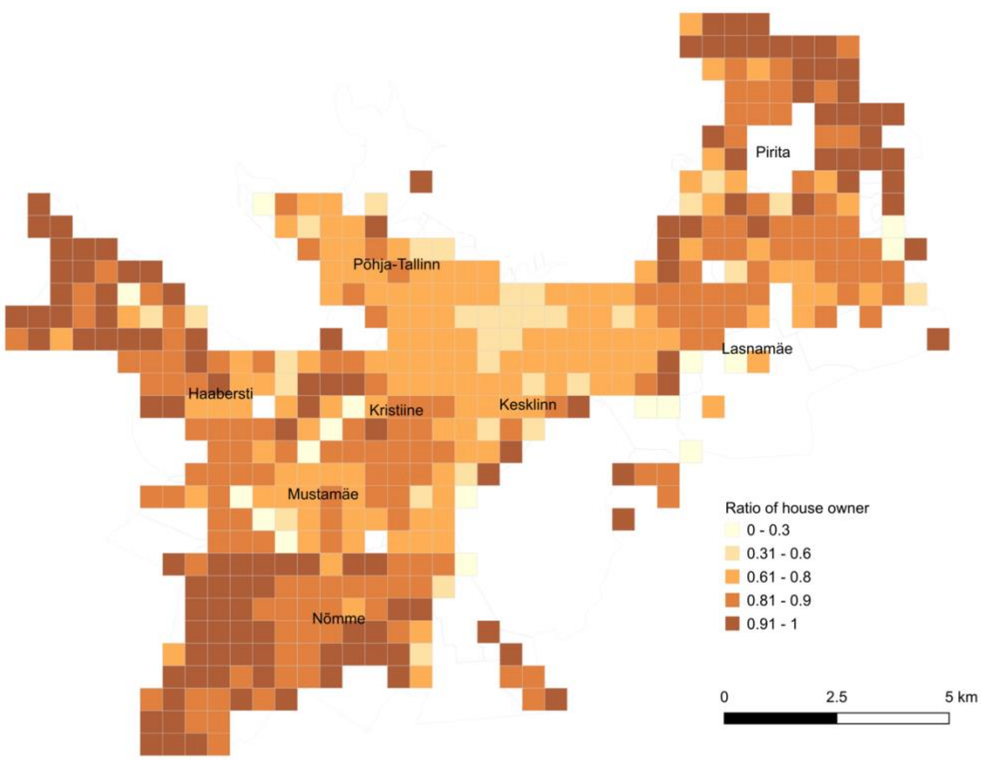


Figure 17. The distribution of the ratio of house owners across all census tracks.

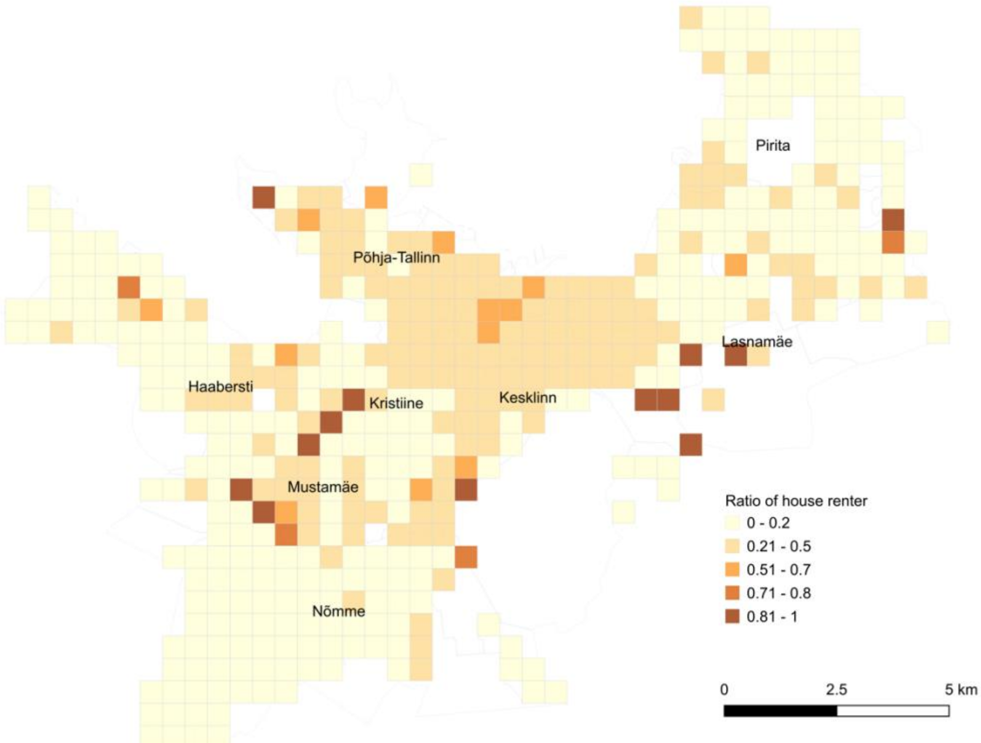


Figure 18. The distribution of the ratio of house renters across all census tracks.

Spatial covariates

To capture spatial relationships, the X and Y coordinates of the centroid of each grid were extracted as covariates.

3.3 Data pre-processing

There are two datasets with different grid systems. Estonian census data uses a 500m grid. Crime data uses a 500m grid in urban areas and a 1000-meter grid for the outskirts of the city and areas outside the metropolitan region. This study uses the 500-meter grid for further analysis. To combine data from two grid systems, the numbers of crimes in the 29 1000-meter grid cells were evenly distributed among the intersecting census tracts. (Figure 19). The categorical data, including land use, POIs, road, and housing types, were extracted and aggregated to align with the grid cells. Numerical data, such as green space, demographic, socioeconomic, housing, and crime data, were aggregated based on the quantities of each class within the grid cells.

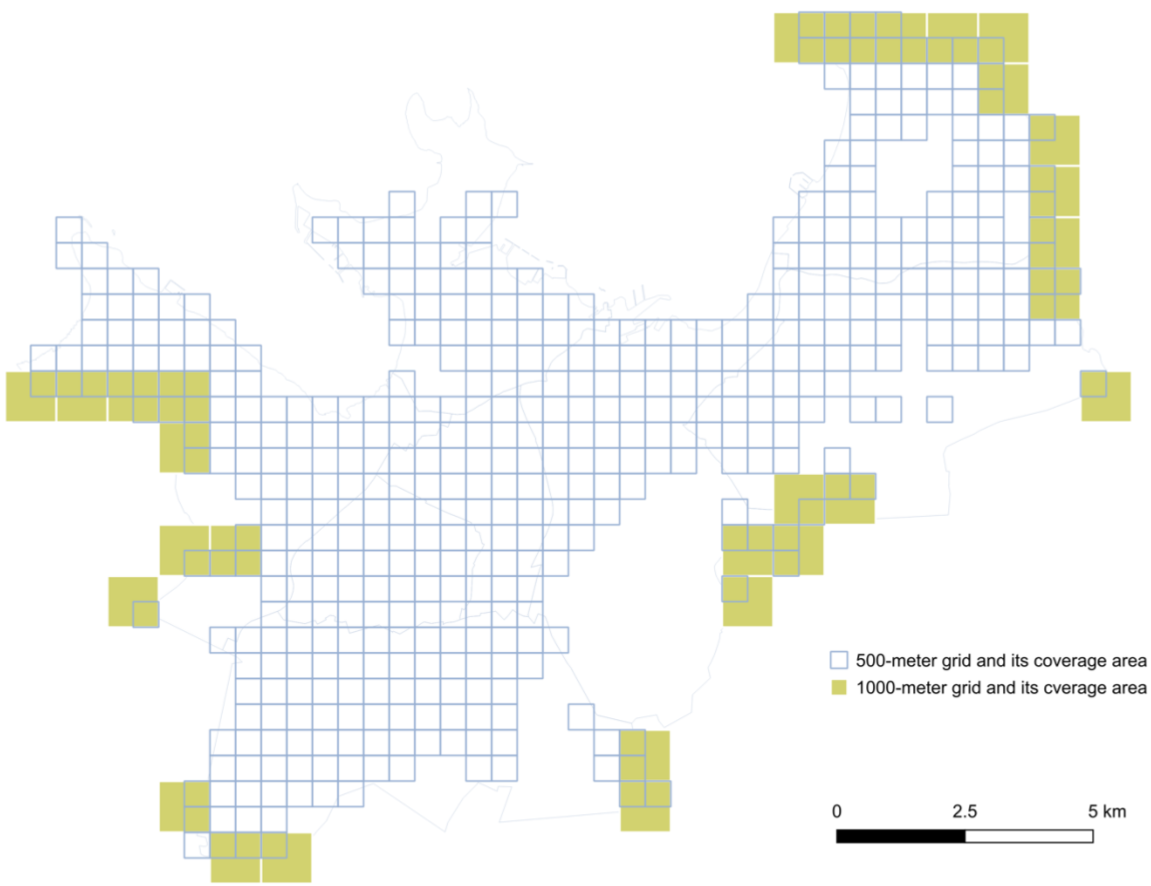


Figure 19. The intersection of census tract and crime grid systems.

In total, 40 covariates were used as predictors, and three dependent variables of the crime were used as targets to be predicted (Table 6). The detailed descriptive statistics can be found in the Appendix 2.

Table 6. All variables used for modeling.

Dataset	Variable	Unit
Crime	• Crimes against public order • Crimes against property • Total crime incidents	Number of incidents per grid cell
Road	• Trail • Neighborhood level • City level	Length (meters) per grid cell
Land use	• Discontinued dense urban fabric • Discontinued medium dense urban fabric • Industrial, commercial, public, military, and private units • Continuous urban fabric	Area (square meters) per grid cell
Green space	• Area of green space	Area (square meters) per grid cell
Diversity of land use	• Number of land use and green space types	Number of land use types per grid cell
Point of Interest	• Commercial • Recreational • Public services • Camera surveillance	Number of points per grid cell
Housing type	• Small-scale apartment house • Standalone house • Large-scale apartment house • Auxiliary house • Others	Number of houses per grid cell
Demographic, socioeconomic, and housing	• Age groups (10) • Ethnicity (2) • Socioeconomic status (3) • House conditions (3) • House ownership (2)	Number of people per grid cell
Spatial	• X and Y coordinates	X and Y coordinate of the grid cell centroid

To process data into a format that a machine learning model can handle, the datasets were processed into a tabular form, with each column being a covariable or a dependent variable. Each row represented a grid cell containing features described by the columns. The structured tabular form simplified the relationship between covariables, especially when all cells were lined up as a one-dimensional array, and the original spatial relationship among the covariables might be removed. The X and Y coordinates were added to maintain the spatial relationship. The data pre-processing was conceptualized in Figure 20.

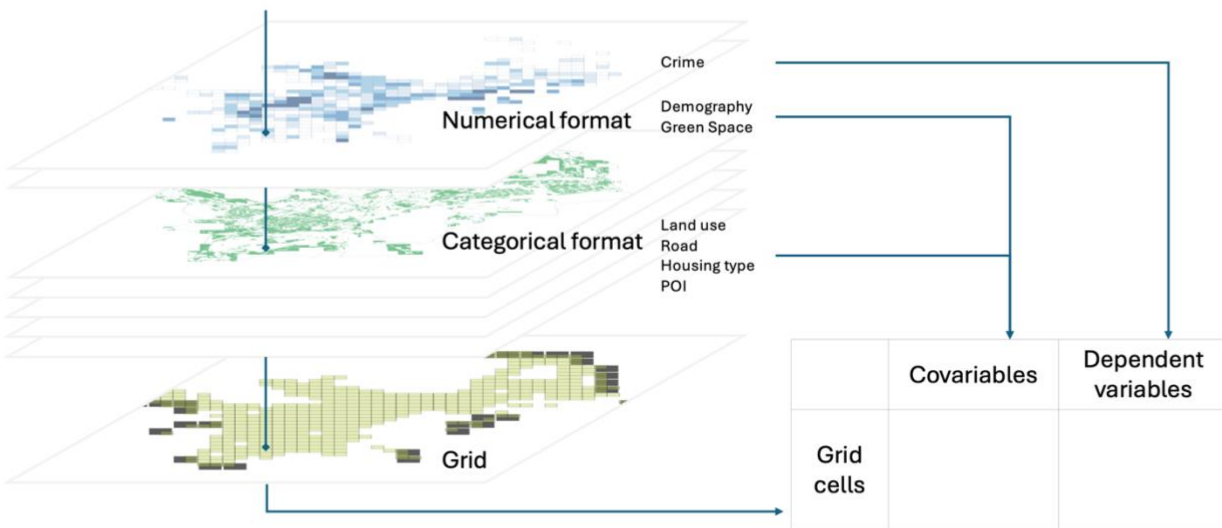


Figure 20. The conceptual workflow of pre-processing.

3.4 Modeling

This study used the random forest regression model. Random forest utilizes a process of extensive random sampling to extract features from the data. This creates a set of conditions used to determine covariates and predict response targets. The method includes mechanisms for computing generalized errors, impurity, and predictive strength. The various conditions can be further manipulated and compared by determining the number of input samples and features to underline the relationships between the features. Generally, random forest can produce good predictive results, but this still depends on the quantity of the data (Breiman, 2001).

This study adjusted six hyperparameters and evaluated them using the out-of-bagging (OOB) error rate metrics and the coefficient of determination (R^2). These hyperparameters were (1) the number of trees, which determines the number of resampling subsets; (2) the depth of trees, which denotes to what degree the deep features can be extracted; (3) the maximum feature methods, which impacted the complexity of the model by regulating the size of features in growing trees; (4) the minimum number of samples required to split a node; (5) the minimum number of samples required to form a leaf; the former two hyperparameters helped the model to learn from important features and prevented the model from over-fitting; and (6) the maximum leaf node, which selected the nodes with the least impurities to represent a tree. These hyperparameters helped to adjust sampling strategies when growing trees and splitting nodes. The OOB error rate is based on data points not included in the bootstrapped training set. OOB aggregated the errors between the predicted values and these out-of-bag data points to estimate the model's accuracy. R^2 indicates to what degree a variance of one variable can explain the variance of another variable.

To find the best hyperparameters, the study ran models with the number of trees, depth, and the other hyperparameters, respectively. A 10-fold cross-validation was repeatedly run until all possible combinations of the 9-fold data for training and one-fold for validation were completed. The scores were averaged as the final validation result. The detailed scores for the models' hyperparameters are in Appendix 3, 4, and 5. Based on the hyperparameter testing, this study eventually used the hyperparameters for three models (Table 7) and fixed a value of 2024 for the random state (it helped to get the same results with the same hyperparameters).

Table 7. Hyperparameters of all models.

Model	Total number of Crime incidents	Number of crimes against property	Number of crimes against public order
Train-Test split	70-30	70-30	60-40
Number of trees	70	100	70
Depth of trees	13	10	7
Maximum leaves in a node	6	6	5
Minimum samples of splitting	12	16	10
Minimum samples in a leaf	18	16	10
Maximum features method	Auto	Auto	Auto

3.5 Data and code availability

The random forest model was established using the Python library scikit-learn. Additionally, the SHAP library was utilized for assessment. The data and code are available at https://github.com/ansel-yu/RF_Tallinn_crime_prediction.

3.6 General model accuracy

In all three models of this study (Table 8), the performance concerning the total number of crime incidents and crimes against property is reasonably good. From the evaluation metrics perspective, both models have decent R^2 values during the training phase (R^2 scores 0.47 and 0.46). The RMSE metric shows a significant range of errors that noticeably increased during the testing phase (respectively, from 187.93 to 232.18 and 182.13 to 225.62). The model for predicting crimes against public order performs better (R^2 scores 0.79 in the training phase and 0.62 in the testing phase).

Table 8. Model comparison.

Model		Total number of crime incidents	Number of crimes against property	Number of crimes against public order
Training	R^2	0.47	0.46	0.79
	RMSE	187.93	182.13	8.55
	MAE	65.55	69.10	4.57
Testing	R^2	0.32	0.28	0.62
	RMSE	232.18	225.62	11.68
	MAE	97.36	96.31	5.69

In addition, this study used AI tools (Grammarly, 2024; OpenAI, 2024) to improve English writing and correct language errors. AI tools helped test sentence clarity and provide suggestions for improvement. The authors evaluated AI-revised content for alignment with the intended purpose.

4. Results

4.1 Feature importance for all crime incidents

From the SHAP plot, the contribution of each variable to the model can be examined. Positive SHAP values indicate a positive contribution to the model, while negative values indicate a negative impact, with colors representing changes in factor values. Figure 21 presents the ten most significant factors influencing the prediction of total crime incidents, revealing that only the number of commercial POIs significantly contributes to the model. Most factors have a negative impact on the model when their values are small. Figure 22, the partial dependent plot illustrates the effect of the number of commercial POIs on the prediction target while keeping other variables constant. This shows the positive influence of the given covariable on the model's predictions. The detailed effects of other important features on the models are in Appendix 6.

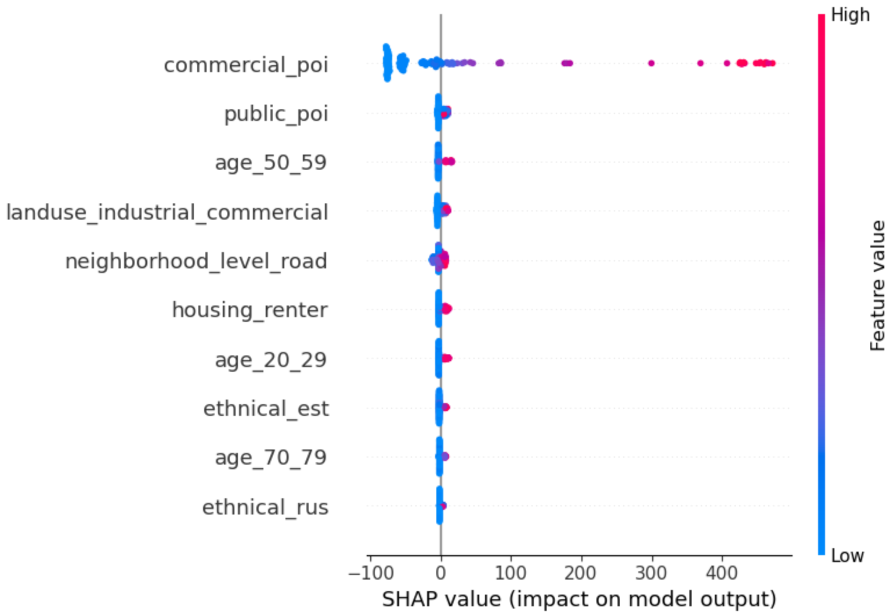


Figure 21. The SHAP values and important features of the total number of crime incidents.

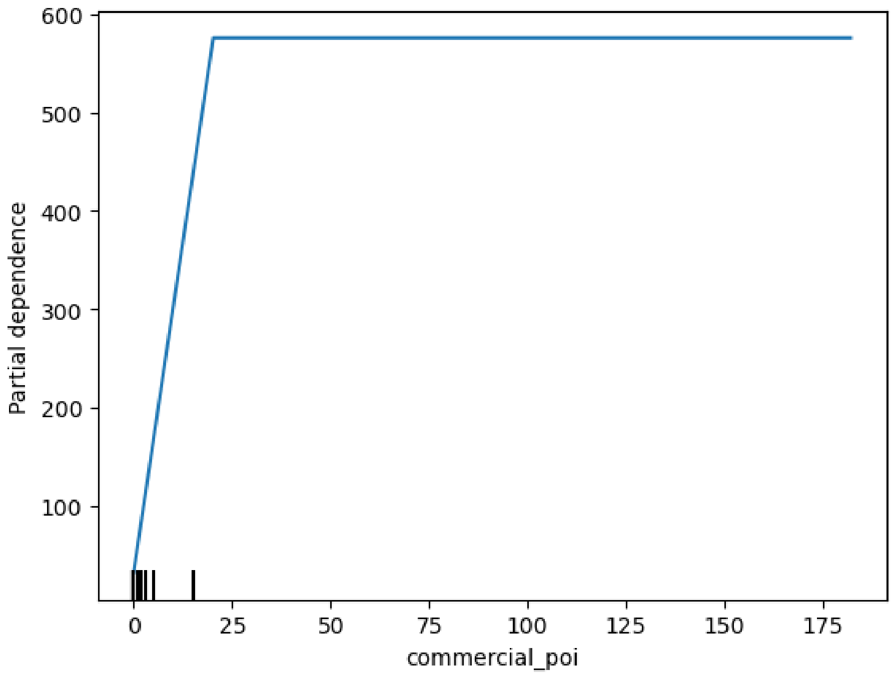


Figure 22. Partial dependent plot of the factor of commercial POIs.

4.2 Feature importance for crimes against property

This model shows a similar pattern of the important factor (Figure 23). Only the number of commercial POIs has a significant contribution to the prediction (Figure 24). The detailed effects of other important features on the models are in Appendix 7.

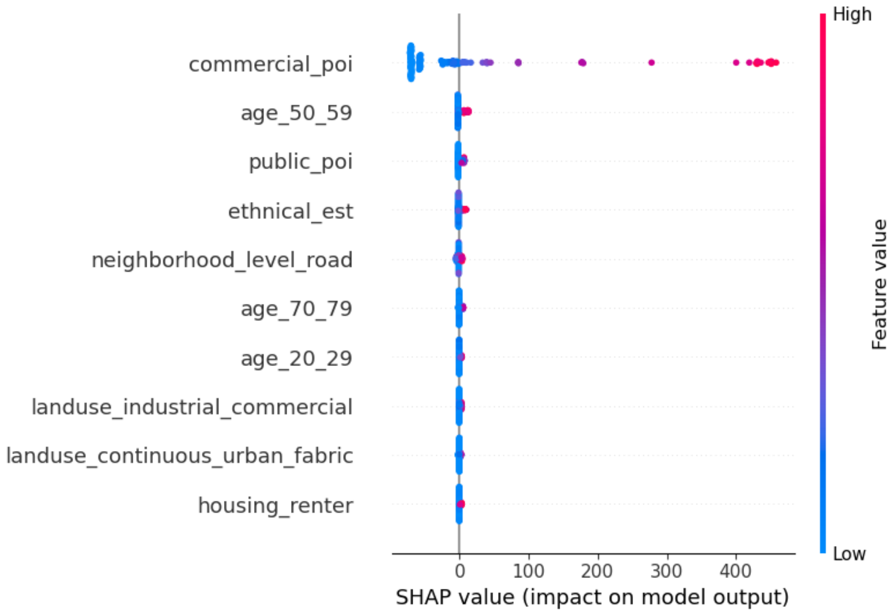


Figure 23. The SHAP values and important features of the number of crimes against property.

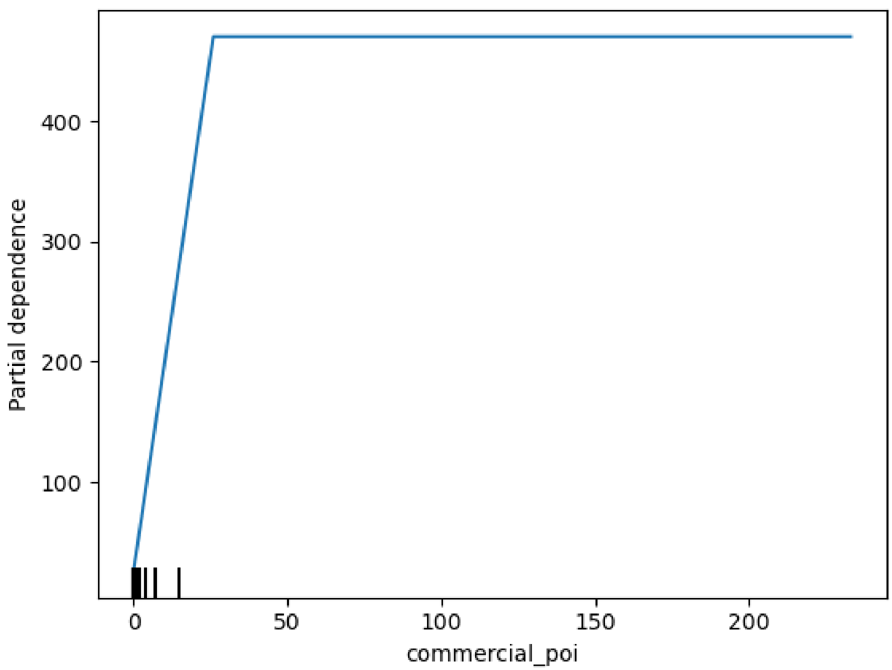


Figure 24. Partial dependent plot of the factor of the commercial POIs.

4.3 Feature importance for crimes against public order

This model demonstrates the highest predictive capability in this study, as evidenced by evaluation metrics (R^2 score 0.62) and the diverse factors contributing to the model's predictions (Figure 25). These factors include the age group between 20 and 29, commercial POIs, the number of house renters, and the age group between 50 and 59 (Figures 26, 27, 28, and 29). The detailed effects of other important features on the models are in Appendix 8.

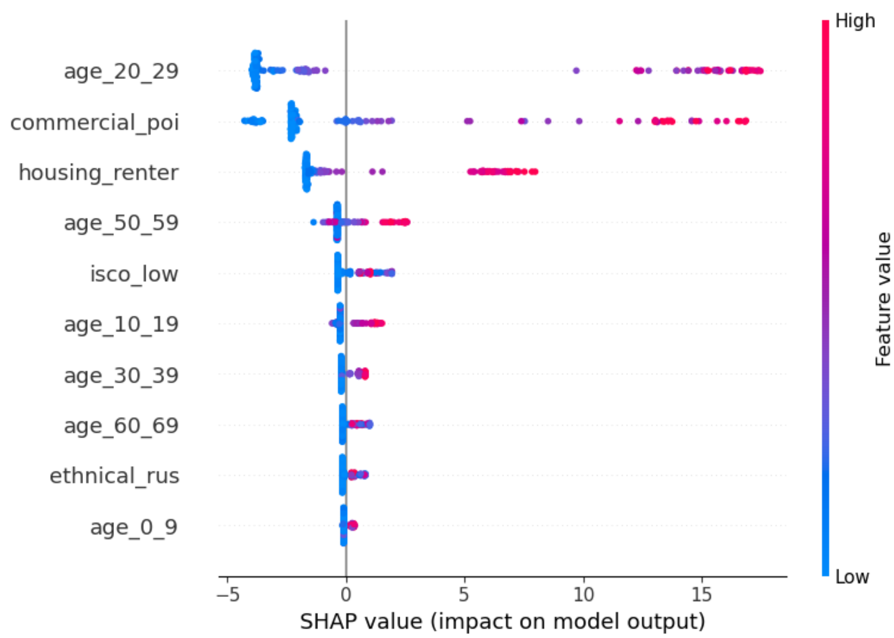


Figure 25. The SHAP values and important features of the number of crimes against public order.

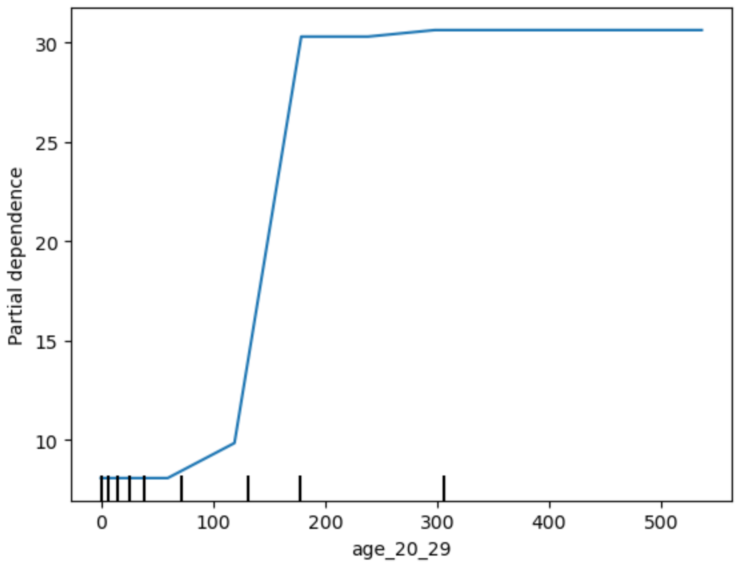


Figure 26. Partial dependent plot of the factor of the age group between 20 and 29.

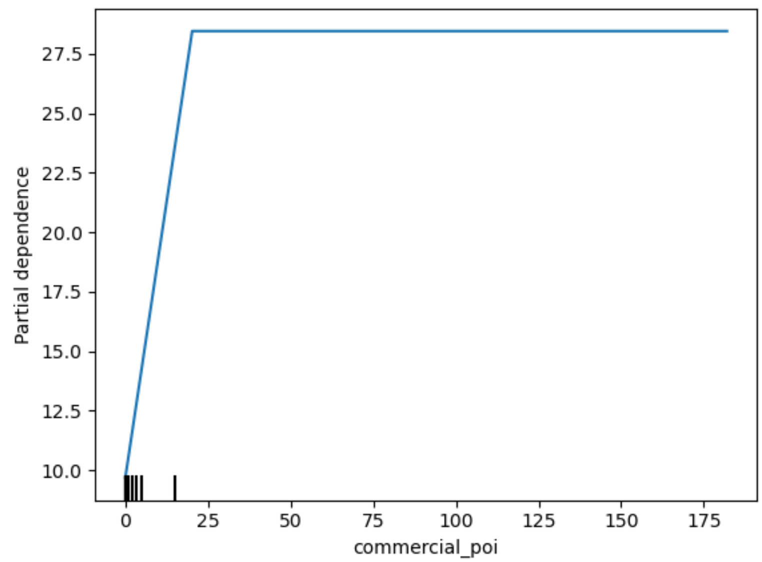


Figure 27. Partial dependent plot of the factor of the commercial POIs.

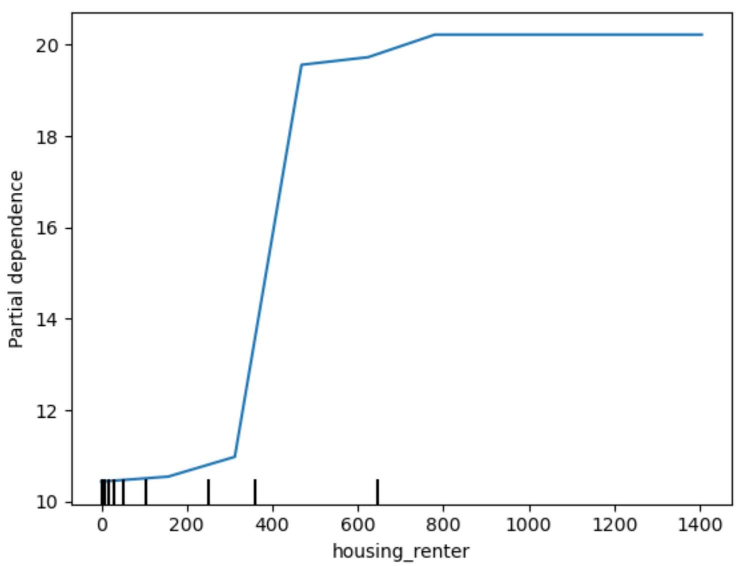


Figure 28. Partial dependent plot of the factor of the number if the house renters.

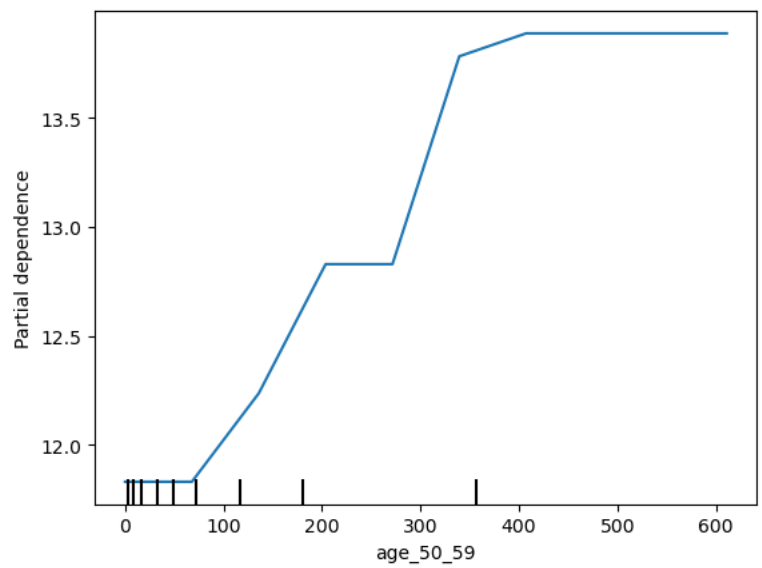


Figure 29. Partial dependent plot of the factor of the e group between 50 and 59.

4.4 Spatial distribution of the residuals of the models

The residual values of the model help to identify the difference between the predicted and actual values and whether there are patterns in the spatial distribution of these differences. The residual distributions of the models of the total number of crime incidents (Figure 30) and crimes against property (Figure 31) show significant residual fluctuation in the inner city and housing estate areas. Meanwhile, the residual distributions of these two models are similar.

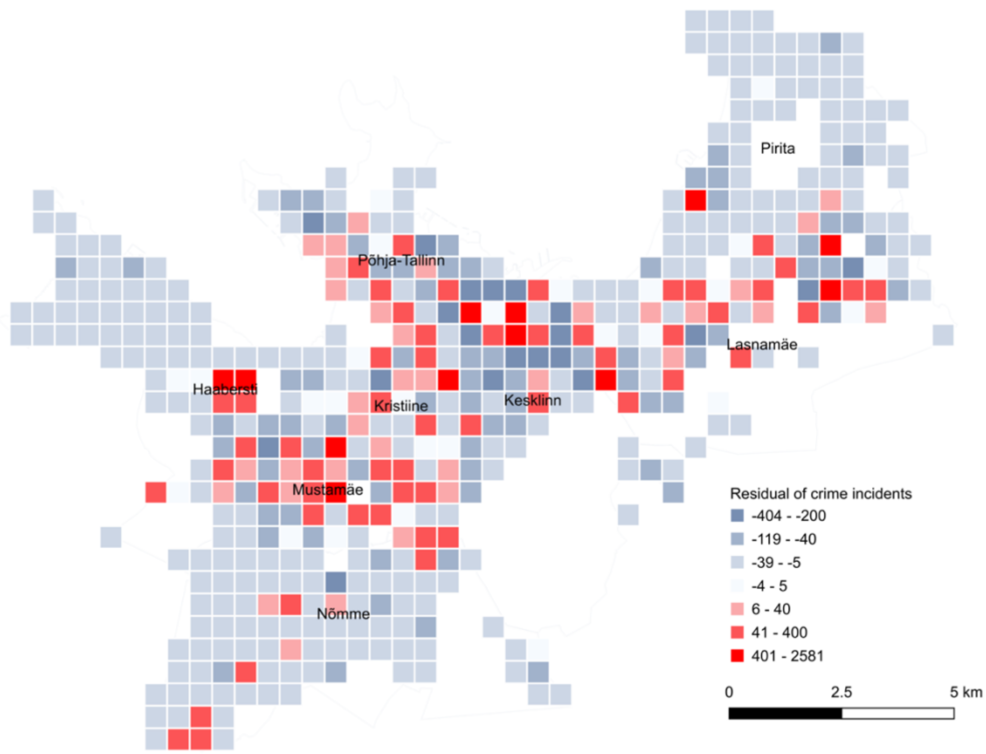


Figure 30. Spatial distribution of residual for all crime incidents.

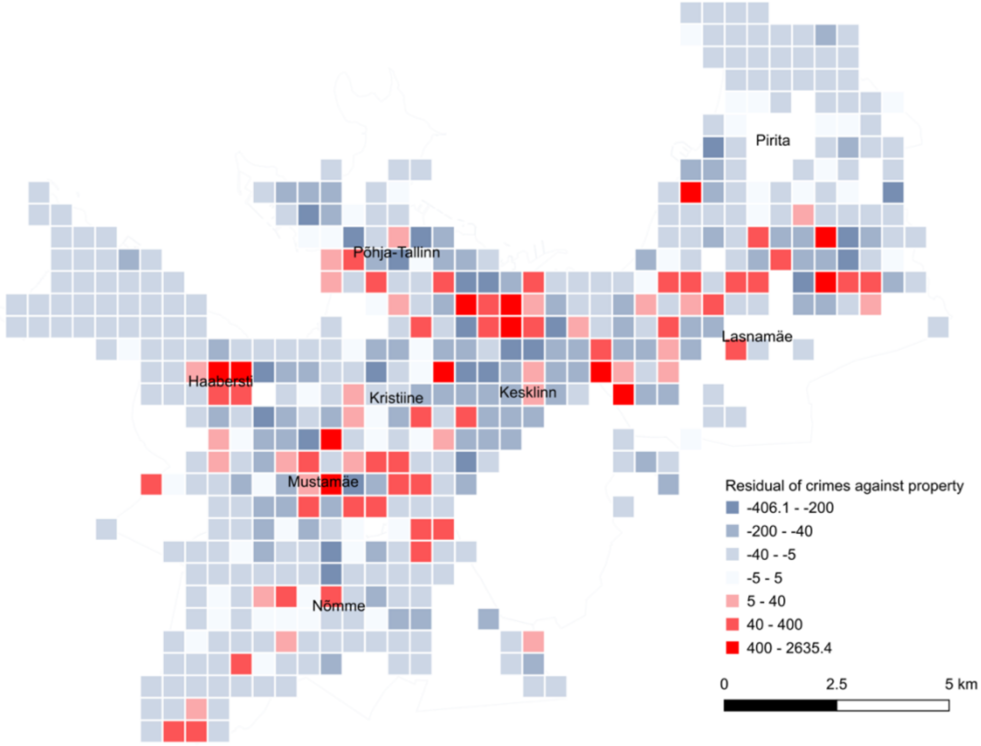


Figure 31. Spatial distribution of residual for crimes against property.

For the crimes against public order, Figure 32 shows that the residuals are more centered on the inner city. Although there is still a concentration trend towards the city center, it is less pronounced than before, and a few fluctuations in the panel housing estates are also identifiable. Additionally, due to the model's better performance, there is less variation in residual values compared to other models.

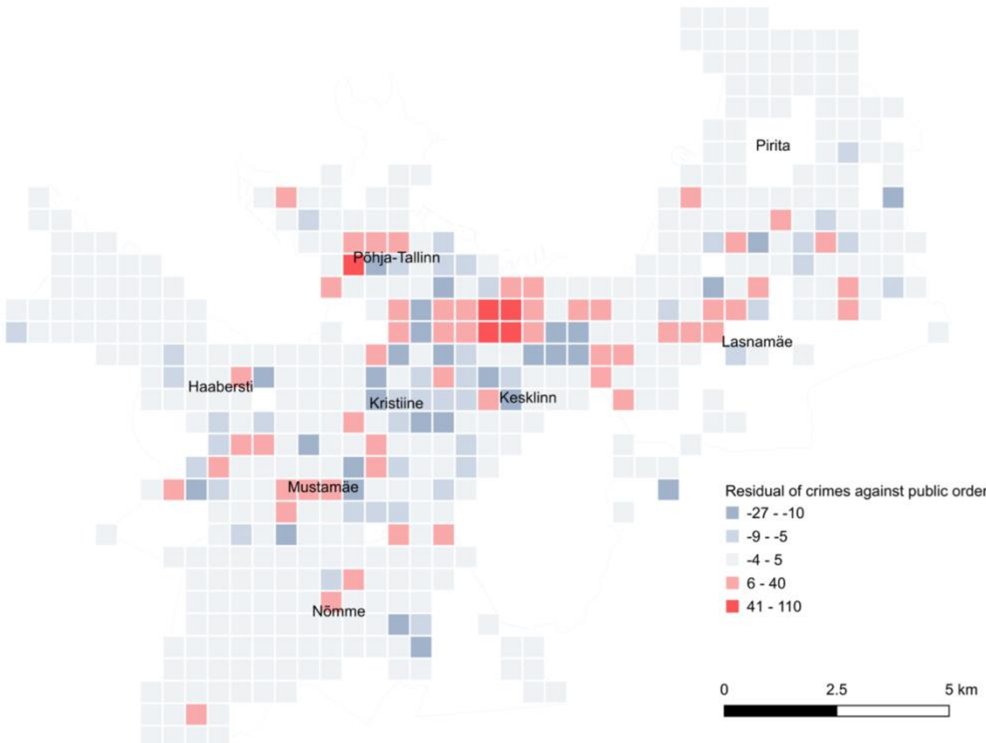


Figure 32. Spatial distribution of residual for crimes against public order.

5. Discussion

This study explores the spatial and socioeconomic characteristics most likely associated with crime incidents in Tallinn and their contextual connections. With the help of the random forest models, factors are identified. The most influential five are (1) the number of residents aged between 20 and 29, (2) the number of commercial POIs, (3) the number of house renters, (4) the number of residents aged between 50 and 59, and (5) the number of residents with low socioeconomic status. Among these factors, the number of commercial POIs is effective in predicting all types of crime, and other factors contribute only to the model of crimes against public order prediction.

5.1 The factor of commercial POIs

The factor of commercial POIs exhibits the broadest explanatory power among all factors in this study. It indicates the pattern of crowd gathering and commercial activities, which could be linked to the distribution of potential public order violators or property crime victims.

This discovery is similar to Ceccato's (2009) research, which focused on bars and pubs as key locations (the study classified bars and pubs as part of the commercial POIs category) and found a notable positive correlation between these locations and nearly all types of crime, with the exception of drug-related offenses. Examining the spatial distribution of this factor, the inner city is the most densely packed commercial location, followed by housing estates (Figure 33). The distribution of commercial POIs extends from the city center outward to the entire inner city, while the sub-centers of each district can also be identified. In the outer city, commercial activity is noticeably sparse and concentrated in a few locations.

Surprisingly, apart from commercial POIs, other factors contribute very little to predicting the number of crimes against property and overall crimes. This observation slightly differs from previous studies. Ceccato (2009) highlighted the impact of land use and transportation on crime, noting that in areas outside the city center, mixed-use land and transportation hubs reflect the density and complexity of people's activities, significantly influencing various types of crime to some degree. However, in this study, the land use and road factors did not impact the three models. This suggests that there may be issues with how data is categorized in this study, leading to uncertain predictive connections. Favarin's (2018) research in Milan may explain the uncertainty mentioned above.

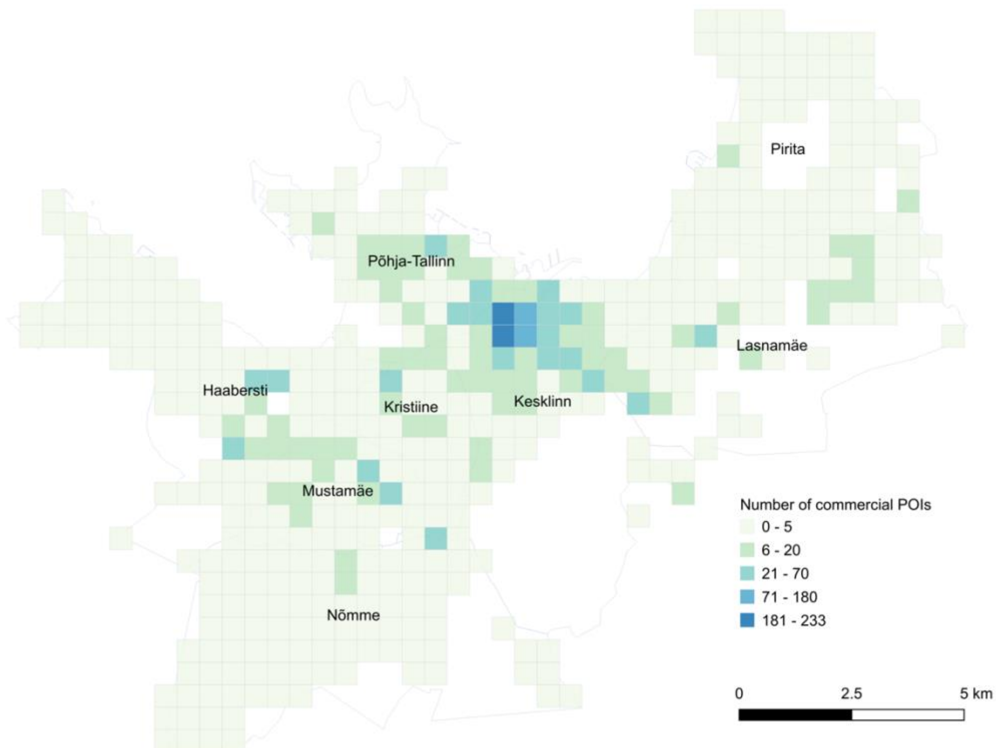


Figure 33. The spatial distribution of commercial POIs.

Favarin (2018) analyzed burglary or robbery and found that burglary is more related to residential factors, while robbery is linked to locations with goods and properties exhibiting different spatial distribution characteristics. Different types of property also exhibit spatial distribution differences, directly impacting the locations where crimes occur. However, in this study, the crime data do not include burglary types (which may have been classified under robbery), potentially resulting in inconsistent spatial characteristics. This inconsistency might explain why none of the land use and road factors contribute to the models.

5.2 The factors of housing tenure

The number of residents with low socioeconomic status and house renters positively impacts the prediction of crimes against public order.

Ceccato's study (2009) found a strong positive correlation between high unemployment rates and vandalism and a significant negative correlation between housing ownership and vandalism in the inner city and suggested that these factors are related to the degree of social disorganization. The case of Glasgow explained this connection, showing that neighborhoods primarily consisting of owner-occupied housing have the lowest crime rates, while those with mainly social housing renters have the highest crime rates (Livingston et al., 2013).

At first glance, this assertion appears similar to the findings of this study, as the spatial distribution of renters aligns with the spatial distribution of crime (Figures 2, 3, and 4), with both overlapping in the inner city and housing estates (Figure 34). However, unlike the previous study (Ceccato, 2009), the owner-occupancy factor did not exhibit a crime-inhibiting effect.

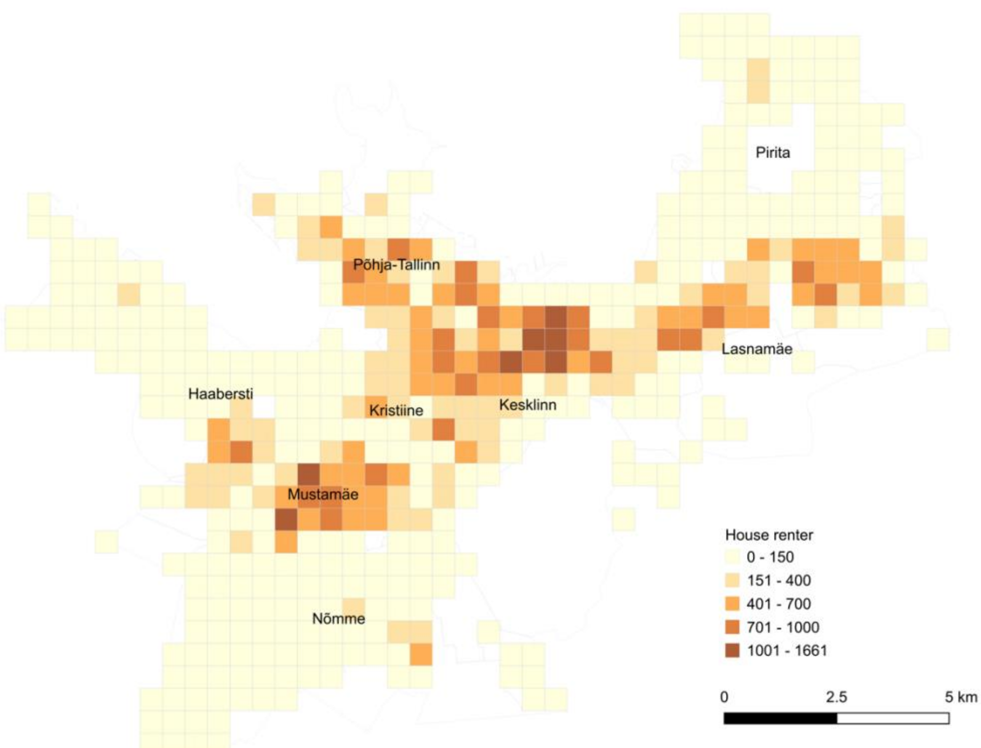


Figure 34. The population distribution of individuals who rent houses.

In Tallinn, housing renting is linked to lower economic capacity and younger groups (Tammaru, 2016). Additionally, it correlates with household structure, as renters are typically associated with single-person households (Statistikaamet, n.d.). This might clarify why the factor of residents aged 20 to 29 (Figure 35) significantly contributes to the prediction of crimes against public order and also clarify the spatial proximity of this age group, renters, and crimes against public order.

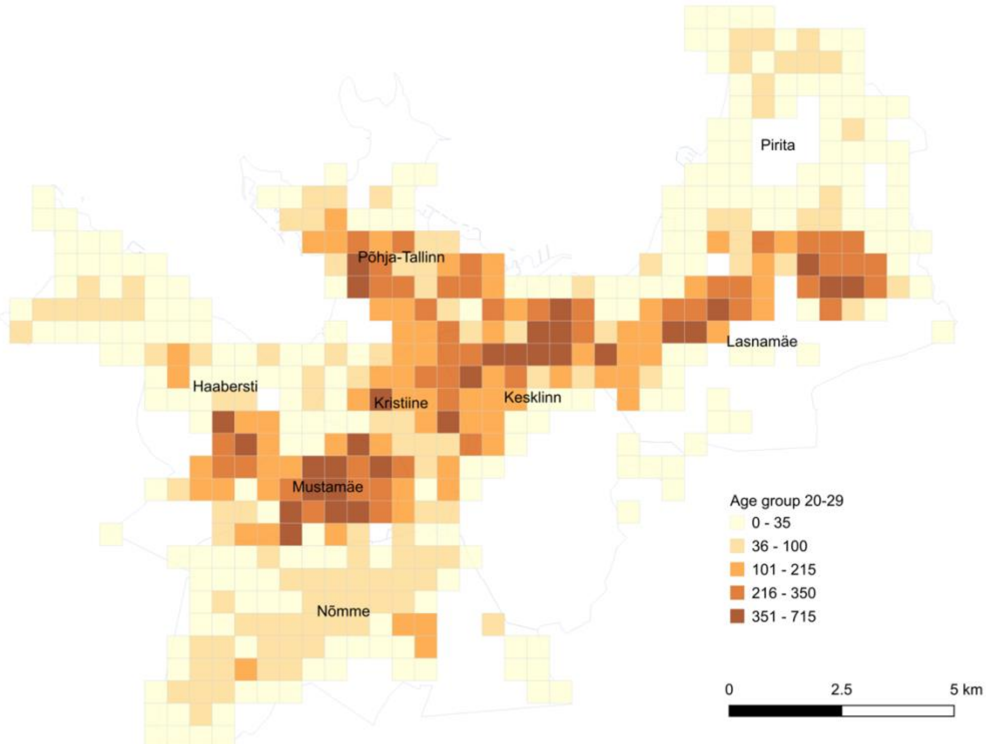


Figure 35. The population distribution of individuals aged 20 to 29.

5.3 The factor of socioeconomic status

The factor of low socioeconomic status contributes to the prediction of crimes against public order. This finding aligns with most research observations that poverty, unemployment, or low socioeconomic status positively correlate with crimes (Krivo et al., 2015; Massey, 1995; Tokey, 2023; Vogel & South, 2016).

The spatial distribution of low-status residents has been influenced by trajectories of urban development, such as varying preferences for residential locations and housing types, privatization and property restitution, as well as recent gentrification trends. Tammaru et al. (2016) identified two notable distribution patterns among high socioeconomic status groups. One pattern is observed in the outer city regions of Pirita, Nõmme, and Haabersti, which have been favored by affluent Estonians. The second pattern is evident in the downtown and inner-city neighborhoods, where gentrification has resulted in affluent residents progressively displacing lower-income residents. Simultaneously, disadvantaged groups have relocated to housing estates or peripheral places in the inner city. Gentrification has also resulted in the coexistence of high-status and low-status residents in certain inner-city neighborhoods. Overall, low-status residents are found in inner-city areas, which are also more concentrated in the housing estates (Figure 36). However, the inner city, Lasnamäe, and Mustamäe exhibit stronger residual fluctuations, suggesting that other crime-influencing factors were not fully included in the model.

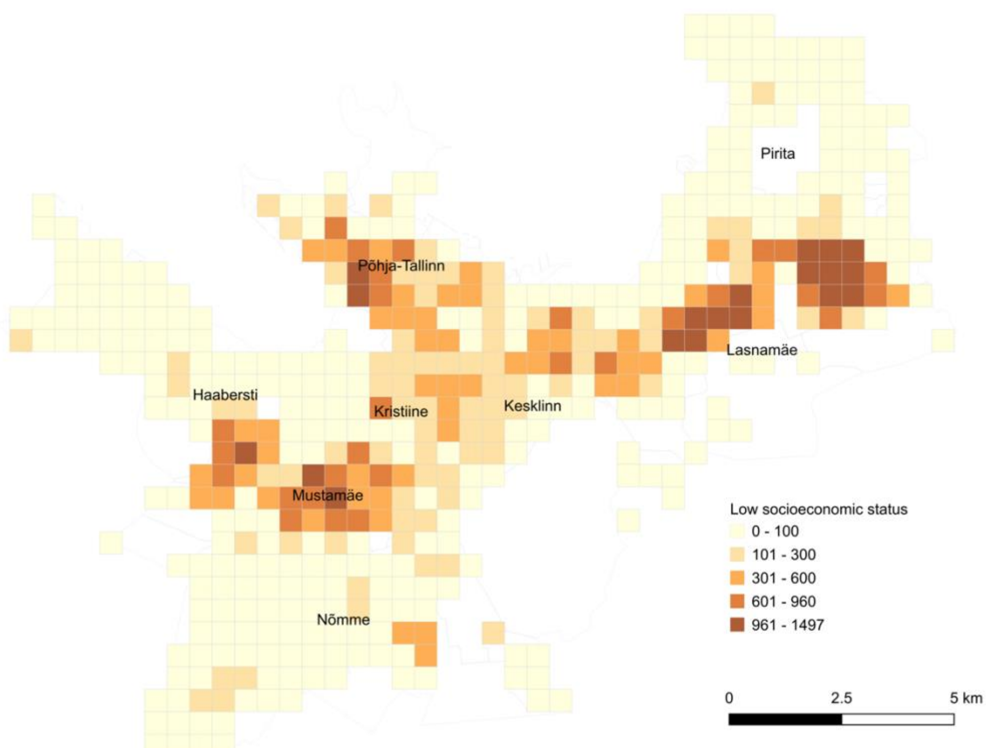


Figure 36. The population distribution of individuals classified as having low socioeconomic status.

5.3 Limitations

Differences in the data years might have excluded real-world changes. The data for this study ranges from 2018 (land use), 2021 (census), to 2024 (POI, green spaces, or roads), while crime incidents range from 2018 to 2022. Although crime types related to the pandemic were removed, temporal and spatial changes during this period might have affected crime occurrence.

Additionally, almost all significant predictors exhibit similar spatial distributions, concentrated in the inner city and panel housing estate areas (Figures 34, 35, 36, and 37), except commercial POIs, which are mainly concentrated in the inner city (Figure 33). The study reveals positive correlations between the number of individuals in two age groups and the number of crimes. However, the exact nature of these correlations is difficult to explain fully.

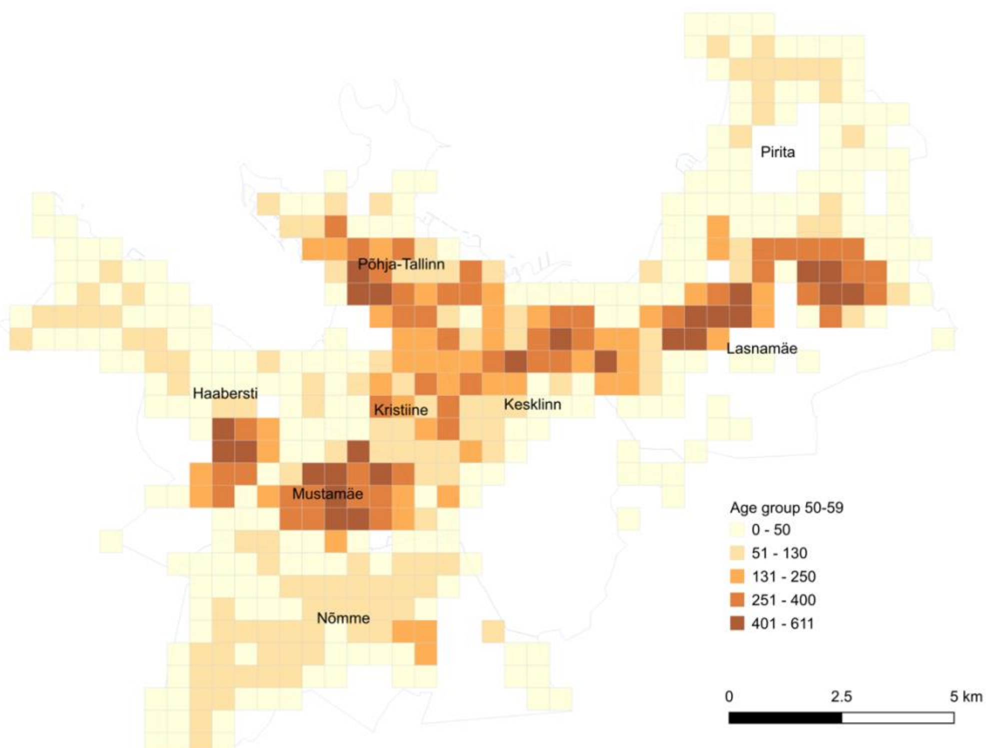


Figure 37 The population distribution of individuals aged 50 to 59.

Lastly, within the 483 grid cells of the study area, 28% of all crimes occurred in the same ten grid cells. 30% of crimes against property occurred in ten grid cells, while 15% of crimes against public order occurred in ten grid cells. The concentration of crimes indicates spatial autocorrelation exists.

The distribution of the models' residuals shows that locations with high crime cases are where the model's learning ability is weaker. This could mean that some important factors were not included in this study or that the data needs to be standardized to remove the influence of outliers.

6. Conclusion

This study employs a random forest machine learning model to identify the spatial, demographic, and socioeconomic factors related to crime in Tallinn and predict the number of crimes based on these factors. This study aims to answer the questions of what spatial and socioeconomic characteristics are most likely associated with crime incidents and how the connections can be explained within the context of Tallinn. The crime data is categorized based on the type of crime, including crimes against public order, crimes against property, and the total of the two. The other data used as covariables include information on road types and lengths, land use, green areas, points of interest (POIs), housing types, ownership or rental status, year of construction, resident age, and occupational income conditions. The study area of Tallinn is divided into 483 grid cells, each measuring 500 meters by 500 meters. Numeric data is aggregated within the grid cells, while categorical data is aggregated after classification. Coordinates are also included to account for spatial autocorrelation. A total of 40 covariables are utilized for prediction.

The study results present the model for predicting crimes against public order, which has achieved good results (R2 scores of 0.79 in training and 0.62 in testing). The model for predicting crimes against property has reasonably good results (R2 scores are 0.46 in training and 0.28 in testing). The third model combines the two types of crimes to predict overall crime numbers and has reasonably good results (R2 scores are 0.47 in training and 0.32 in testing). The most influential factors are (1) the number of residents aged between 20 and 29, (2) the number of commercial POIs, (3) the number of house renters, (4) the number of residents aged between 50 and 59, and (5) the number of residents with low socioeconomic status.

Commercial POIs indicate the pattern of crowd gathering and commercial activities, which could be linked to the distribution of potential public order violators or property crime targets. This factor is mainly concentrated in the inner city. The factors of rental housing and socioeconomic status are associated with the trajectories of urban development. These factors coincide with the spatial distribution of crimes against public order. These issues will be significant challenges for achieving equity and inclusive development in the city.

Analyzing the relationships between crime and socio-economic and spatial factors using random forest: a case study of Tallinn

Cheng-Wei Yu

Summary

This study employs a random forest machine learning model to identify the spatial, environmental, and socioeconomic factors related to crime in Tallinn and predict the number of crimes based on these factors. This study aims to answer the questions of what spatial and socioeconomic characteristics are most likely associated with crime incidents and how the connections can be explained within the context of Tallinn.

The topic of urban crime reflects how people constitute the essential values and understanding for the operation of urbanized societies and covers a variety of aspects from the social context of age, race, income, and privilege or deprivation to land use, infrastructure, the function of built-up environment, and other spatial features. The random forest model helps handle these multivariate data and their relationships and utilizes a process of extensive random sampling to extract features from them.

The crime data is categorized based on the type of crime, including crimes against public order, crimes against property, and the total of the two. The other data used as covariables include information on road types and lengths, land use, green areas, points of interest (POIs), housing types, ownership or rental status, year of construction, resident age, and occupational income conditions. The study area of Tallinn is divided into 483 grid cells, each measuring 500 meters by 500 meters. Numeric data is aggregated within the grid cells, while categorical data is aggregated after classification. Coordinates are also included to account for spatial autocorrelation. A total of 40 covariables are utilized for prediction.

The study results present the model for predicting crimes against public order, which has achieved good results (R2 scores of 0.79 in training and 0.62 in testing). The top three main factors are the

population of residents aged 20 to 29, the number of commercial POIs, and the number of renters. Other moderate factors are the population of residents aged 50 to 59, low socio-economic status, and the population of residents aged 10 to 19. The model for predicting crimes against property has reasonably good results (R2 scores are 0.46 in training and 0.28 in testing). The number of commercial POIs is the primary factor. Other factors have a limited impact on the model. The third model combines the two types of crimes to predict overall crime numbers and has reasonably good results (R2 scores are 0.47 in training and 0.32 in testing), with commercial POIs being the most important factor as well.

In conclusion, socioeconomic factors contribute more to the crime prediction models than spatial and environmental factors at the scale of 500-meter grids. The definitions and classifications of crime and various covariables also affect the prediction model's performance. In this study, the factors that influence crimes against public order largely align with major crime theories' explanations of socioeconomic structural factors. Commercial POIs indicate the pattern of crowd gathering and commercial activities, which could be linked to the distribution of potential public order violators or property crime victims. Additionally, the crime prediction reflects the increasingly severe trends of economic and residential segregation in Tallinn.

Kuritegevuse ja sotsiaalmajanduslike ning ruumiliste tegurite vaheliste seoste analüüs otsustusmetsa abil: Tallinna juhtum

Cheng-Wei Yu

Kokkuvõte

Käesolev uurimistöö kasutab masinõppe otsustusmetsa mudelit, et tuvastada Tallinna kuritegevusega seotud ruumilised, keskkonna ja sotsiaalmajanduslikud tegurid ning ennustada kuritegude arvu nende tegurite põhjal. Uurimuse eesmärk on vastata küsimusele, millised ruumilised ja sotsiaalmajanduslikud tunnused on kõige suurema tõenäosusega seotud kuritegude juhtumitega ning kuidas neid seoseid Tallinna kontekstis kirjeldada.

Linna kuritegevuse teema peegeldab seda, kuidas inimesed loovad peamisi väärtusi ja arusaama linnastunud ühiskondade toimimiseks ning hõlmab mitmeid aspekte alates vanuse sotsiaalsest kontekstis, rassist, sissetulekust, privileegidest või puudustest kuni maakasutuse, infrastruktuuri, hoonestatud keskkonna ja teise ruumiliste funktsioonideni. Otsustusmetsa mudel aitab käsitleta neid mitme muutujaga andmeid ja nende omavahelisi keerukaid suhteid.

Kuritegude andmed klassifitseeriti kuriteo liigi aluselavaliku korra rikkumisteks, varavastasteks kuritegudeks, ja nende kahe kogusummaks. Muude tunnustenakasutatiteavet teetüüpide ja nende pikkuste, maakasutuse, roheliste alade, huvipunktide (POI-andmete), eluasemetüüpide, omandivormi, ehituaastate, elanike vanuste ja ametialase sissetuleku tingimuste kohta. Tallinna uurimisala on jaotatud 483 ruuduks, millest iga üks on 500 meetrit korda 500 meetrit. Ruumilise autokorrelatsiooniga arvestamiseks on kaasatud ka koordinaadid. Ennustamiseks kasutatakse kokku 40 muutujat.

Avaliku korra rikkumiste kohta loodud mudeli treeningu R^2 oli 0.79 ja valideerimise R^2 oli 0.62 . Peamised kolm avaliku korra rikkumisi mõjutavad tegurit on elanike arv vanusevahemikus 20-29, kaubandusasutuste arv ja üüritavate pindade arv. Teised vähem olulised tegurid on elanike arv vanusevahemikus 50-59, madal sotsiaalmajanduslik staatus ja elanike arv vanusevahemikus 10-19.Varavastaste kuritegude prognoosimiseks loodud mudeli treening R^2 oli 0.46 väljaõppel ja valideerimise R^2 0.28. Peamine mõjutegur on kaubanduslike huvipunktide arv. Ülejäänud tegurid olulist mõju ei omanud.

Kolmas mudel kombineerib kaht tüüpi kuritegevusi, et ennustada üleüldist kuritegude arvu ning sellel on võrdlemisi head tulemused, treeninguR^2 0.47 ja valideerimisel vastavalt 0.32 ningkõige olulisemaks teguriks osutusid samuti kaubandusasutused.

Kokkuvõtteks, sotsiaalmajanduslikud tunnused seletavad kuritegevuse arvu rohkem kui ruumilised ja keskkondlikud tunnused. Kuritegevuse ja mitmete muutujate definitsioon ja klassifikatsioon mõjutavad mudelite täpsust. Selles töös on langevad avaliku korra rikkumisi mõjutavad tunnused suuresti kokku peamiste kuritegevuse teooriate seletusega sotsiaalmajanduslike naabruskonda iseloomustavate tunnuste kohta. Kaubandusasutused näitavad tihti kaudseltrahva kogunemise ja äritegevuse kohti, mida saab seostada potentsiaalsete avaliku korra rikkujate või varavastase kuritegude ohvrite asukohaga. Lisaks peegeldab kuritegevuse ennustamine Tallinnas majandusliku ja elukohapõhise segregatsiooni suurenevat trendi.

Acknowledgments

I thank Professor Evelyn Uuemaa and PhD. Alexander Kmoch for their guidance in data processing and modeling and their feedback on this thesis. I also thank the support of PhD. Kadi Kalm and MSc. Jānis Zālīte for their expertise in demographic and socioeconomic perspectives related to the study area.

References

Alves, L. G. A., Ribeiro, H. V., & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica. A*, *505*, 435–443. https://doi.org/10.1016/j.physa.2018.03.084

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Ceccato, V. (2009). Crime in a city in transition: the case of Tallinn, Estonia. *Urban Studies*, *46*(8), 1611–1638. https://doi.org/10.1177/0042098009105501

Ceccato, V., & Oberwittler, D. (2008). Comparing spatial patterns of robbery: Evidence from a Western and an Eastern European city. *Cities*, *25*(4), 185–196. https://doi.org/10.1016/j.cities.2008.04.002

Cozens, P. (2011). Urban Planning and Environmental Criminology: Towards a New Perspective for Safer Cities. *Planning Practice + Research/Planning Practice and Research*, *26*(4), 481–508. https://doi.org/10.1080/02697459.2011.582357

Estonian Land Board. (2023). *Residential buildings from address data* [Dataset]. https://metadata.geoportaal.ee/geonetwork/srv/eng/catalog.search#/metadata/6f2cffe2-65d9-4f9b-b667-ba7716e6f966

Estonian Land Board. (2024). *Estonian Topographic Database* [Dataset]. https://geoportaal.maaamet.ee/index.php?lang_id=2&page_id=618

Favarin, S. (2018). This must be the place (to commit a crime). Testing the law of crime concentration in Milan, Italy. *European Journal of Criminology*, *15*(6), 702–729. https://doi.org/10.1177/1477370818757700

Gao, B., Stein, A., & Wang, J. (2022). A two-point machine learning method for the spatial prediction of soil pollution. *International Journal of Applied Earth Observation and Geoinformation, 108*, 102742. https://doi.org/10.1016/j.jag.2022.102742

Geofabrik. (2024). *OpenStreetMap data for this region: Estonia* [Dataset]. https://download.geofabrik.de/europe/estonia.html

Graif, C., Gladfelter, A. S., & Matthews, S. A. (2014). Urban Poverty and Neighborhood Effects on Crime: Incorporating spatial and network perspectives. *Sociology Compass*, *8*(9), 1140–1155. https://doi.org/10.1111/soc4.12199

Grammarly. (2024). Grammarly (May 27 version) [Large language model]. https://app.grammarly.com/

Hirschfield, A., Birkin, M., Brunsdon, C., Malleson, N., & Newton, A. (2013). How places influence crime: The impact of surrounding areas on neighbourhood burglary rates in a British city. *Urban Studies, 51*(5), 1057–1072. https://doi.org/10.1177/0042098013492232

Iliyasu, I. I., Abdullah, A., & Marzbali, M. H. (2022). Urban morphology and crime patterns in urban areas: a review of the literature. *Malaysian Journal of Sustainable Environment/Malaysia Journal of Sustainable Environment*, *9*(1), 213. https://doi.org/10.24191/myse.v9i1.17301

Johnson, S. D., & Bowers, K. (2009). Permeability and burglary risk: Are Cul-de-Sacs safer? *Journal of Quantitative Criminology*, *26*(1), 89–111. https://doi.org/10.1007/s10940-009-9084-8

Jones, M. A., & Fanek, M. F. (1997). *Crime and the urban environment* (Vol. 2) [Space Syntax First International Symposium].

Kounadi, O., Ristea, A., De Araujo, A., & Leitner, M. (2020). A systematic review on spatial crime forecasting. *Crime Science*, *9*(1). https://doi.org/10.1186/s40163-020-00116-7

Krivo, L. J., Byron, R. A., Calder, C. A., Peterson, R. D., Browning, C. R., Kwan, M., & Lee, J. Y. (2015). Patterns of local segregation: Do they matter for neighborhood crime? *Social Science Research*, *54*, 303–318. https://doi.org/10.1016/j.ssresearch.2015.08.005

Leetmaa, K., Holvandus, J., Mägi, K., & Kährik, A. (2018). Population shifts and urban policies in housing estates of Tallinn, Estonia. In *The urban book series* (pp. 389–412). https://doi.org/10.1007/978-3-319-92813-5_17

Leetmaa, K., Maarja, E., Kalm, K., Pastak, I., & Tammaru, T. (2022). Gentrification as a micro-segregation phenomenon: social and spatial layers of Tallinn inner city. In *Vertical Cities* (pp. 219–236). https://doi.org/10.4337/9781800886391.00025

Livingston, M., Kearns, A., & Bannister, J. (2013). Neighbourhood Structures and Crime: The Influence of Tenure Mix and Other Structural Factors upon Local Crime Rates. *Housing Studies*, *29*(1), 1–25. https://doi.org/10.1080/02673037.2014.848267

Lotfata, A., Grekousis, G., & Wang, R. (2023). Using geographical random forest models to explore spatial patterns in the neighborhood determinants of hypertension prevalence across chicago, illinois, USA. *Environment and Planning. B, Urban Analytics and City Science/Environment & Planning. B, Urban Analytics and City Science*, *50*(9), 2376–2393. https://doi.org/10.1177/23998083231153401

Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime Prediction using Machine Learning and Deep Learning: A Systematic review and Future Directions. *IEEE Access*, *11*, 60153–60170. https://doi.org/10.1109/access.2023.3286344

Massey, D. S. (1995). Getting Away with Murder: Segregation and Violent Crime in Urban America. *University of Pennsylvania Law Review*, *143*(5), 1203. https://doi.org/10.2307/3312474

OpenAI. (2024). ChatGPT (May 27 version) [large language model]. https://chat.openai.com/

Police and Border Guard Board. (2023). *Open data of offences 2015-2022* [Dataset]. https://opendata.smit.ee/ppa/csv/vara_2.csv

Rand, G. (1982). Crime and environment: A review of the literature and its implications for urban architecture and planning. *Journal of Architectural and Planning Research*, *1*(1), 17. https://www.jstor.org/stable/43028670

Reades, J., De Souza, J., & Hubbard, P. (2018). Understanding urban gentrification through machine learning. *Urban Studies*, *56*(5), 922–942. https://doi.org/10.1177/0042098018789054

Ruoppila, S. (2007). Establishing a Market-orientated Urban Planning System after State Socialism: The Case of Tallinn. *European Planning Studies*, *15*(3), 405–427. https://doi.org/10.1080/09654310601017117

scikit-learn: machine learning in Python — scikit-learn 1.4.2 documentation. (n.d.). https://scikit-learn.org/stable/

Sekulić, A., Kilibarda, M., Heuvelink, G., Nikolić, M., & Bajat, B. (2020). Random forest spatial interpolation. *Remote Sensing*, *12*(10), 1687. https://doi.org/10.3390/rs12101687

Statistics Estonia. (2023). *Population and Housing Census 2021*. https://rahvaloendus.ee/en/census-2021/population-and-housing-census-2021

Statistikaamet. (n.d.). *Housing arrangements*. https://rahvaloendus.ee/en/results/housing-arrangements

Tallinn city government. (2022). *Statistics and yearbooks*. Tallinn. https://www.tallinn.ee/en/statistics-and-yearbooks

Tallinn city government. (2023). *Facts about Tallinn*. Tallinn. https://www.tallinn.ee/en/ettevotjale/facts-about-tallinn

Tammaru, T., Kährik, A., Mägi, K., Novák, J., & Leetmaa, K. (2016). The 'Market Experiment': Increasing socio-economic segregation in the inherited bi-ethnic context of Tallinn. In *Socio-Economic Segregation in European Capital Cities: East meets West* (pp. 333–357). New York: Routledge. https://doi.org/10.4324/9781315758879-22

Tammaru, T., Leetmaa, K., Silm, S., & Ahas, R. (2009). Temporal and Spatial Dynamics of the New Residential Areas around Tallinn. *European Planning Studies, 17*(3), 423–439. https://doi.org/10.1080/09654310802618077

Tammaru, T., Sinitsyna, A., Akhavizadegan, A., Van Ham, M., Marcińczak, S., & Musterd, S. (2021). Income inequality and residential segregation in European cities. In *The urban book series* (pp. 39–54). https://doi.org/10.1007/978-3-030-74544-8_3

The Copernicus Land Monitoring Service. (2018). *Urban Atlas Land Cover/Land Use 2018 (vector), Europe, 6-yearly* [Dataset]. https://land.copernicus.eu/en/products/urban-atlas/urban-atlas-2018

The Copernicus Land Monitoring Service. (2023). Mapping guide for a European urban atlas. *Urban Atlas Land Cover/Land Use and Street Tree Layer 2012 and 2018, v6.3*. https://land.copernicus.eu/en/technical-library/urban_atlas_2012_2018_mapping_guide/

Tokey, A. I. (2023). Property crime and violent crime in Detroit: spatial association with built environment before and during COVID-19. *Bulletin of Geography. Socio-Economic Series, 59*, 131–150. https://doi.org/10.12775/bgss-2023-0009

Vogel, M., & South, S. J. (2016). Spatial dimensions of the effect of neighborhood disadvantage on delinquency. *Criminology*, *54*(3), 434–458. https://doi.org/10.1111/1745-9125.12110

Welcome to the SHAP documentation — SHAP latest documentation. (n.d.). https://shap.readthedocs.io/

Wheeler, A., & Steenbeek, W. (2020). Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology*, *37*(2), 445–480. https://doi.org/10.1007/s10940-020-09457-7

Xia, Z., Stewart, K., & Fan, J. (2021). Incorporating space and time into random forest models for analyzing geospatial patterns of drug-related crime incidents in a major U.S. metropolitan area. *Computers, Environment and Urban Systems*, *87*, 101599. https://doi.org/10.1016/j.compenvurbsys.2021.101599

Appendix

Appendix 1. The classification of POIs and their original types

	Commercial	**Recreational**	**Public**	**Camera surveillance**
Original types	atm bank beauty_shop bookshop bicycle_rental bicycle_shop car_dealership car_rental car_wash chemist cinema clothes computer_shop convenience department_store florist furniture_shop gift_shop greengrocer guesthouse hairdresser hostel hotel ice_rink jeweller kiosk laundry mall market_place mobile_phone_shop motel nightclub optician outdoor_shop pharmacy pub shoe_shop sports_shop stationery supermarket toy_shop vending_any vending_machine bakery bar beverages butcher cafe fast_food restaurant	archaeological dog_park attraction battlefield camp_site garden_centre memorial monument park picnic_site pitch playground ruins sports_centre swimming_pool tourist_info track viewpoint windmill zoo	clinic college community_centre arts_centre artwork dentist doctors fire_station kindergarten library museum police post_box post_office school shelter theatre toilet town_hall university veterinary waste_basket	camera_surveillance

Appendix 2. The descriptive statistics of all variables

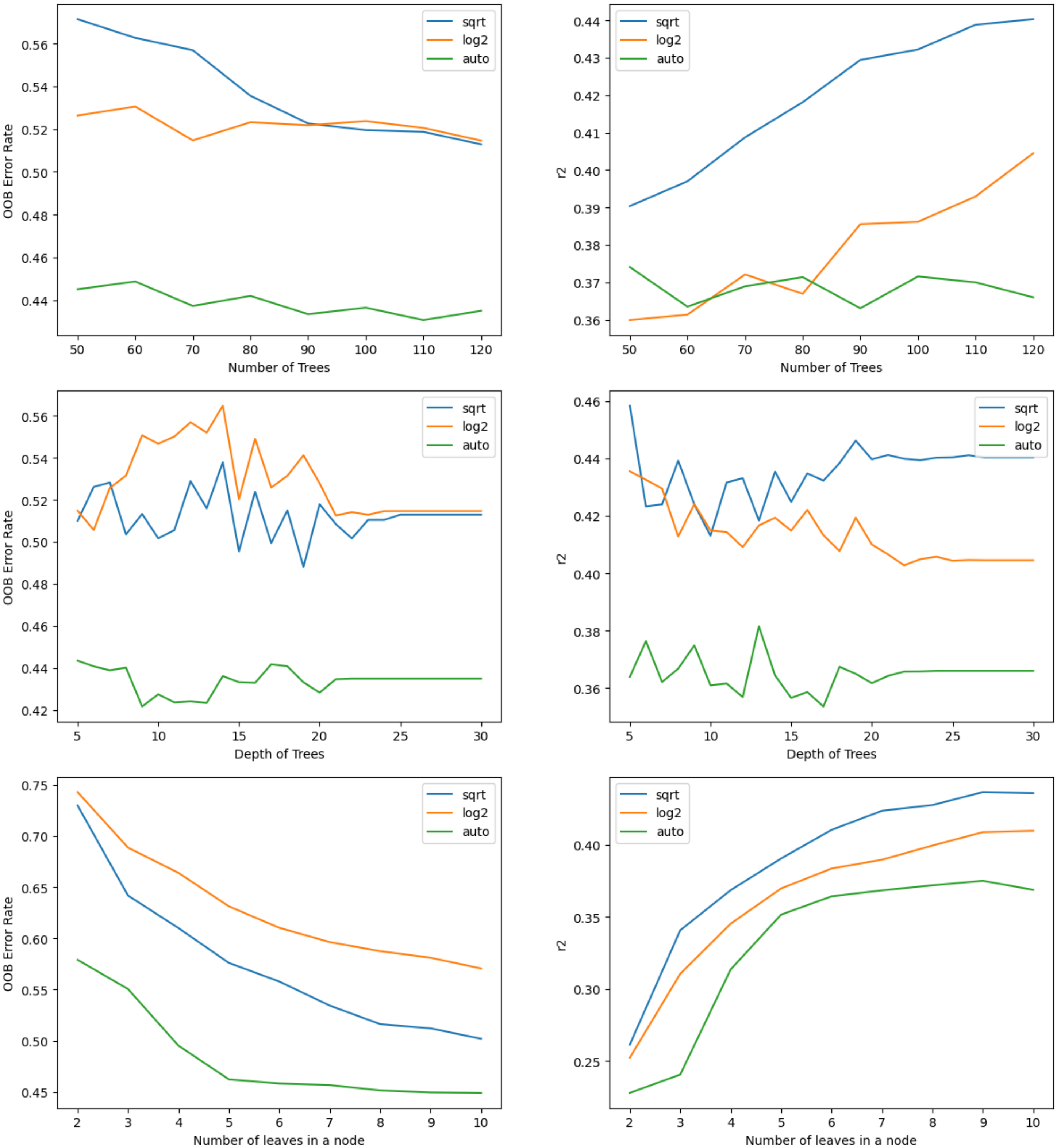
Descriptive Statistics

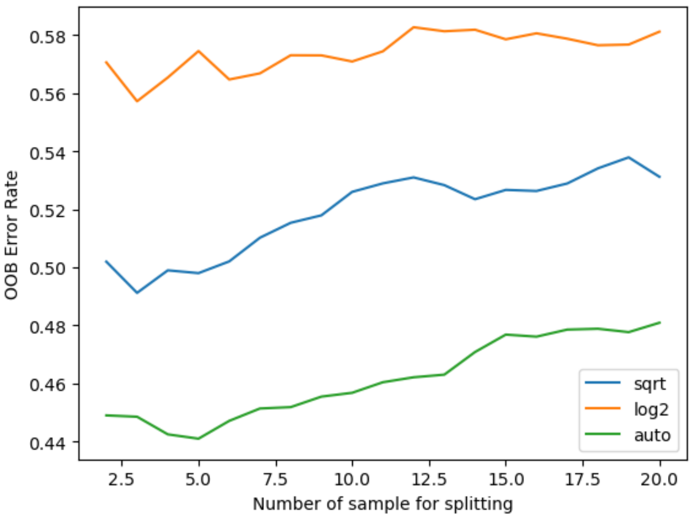
	Median	Mean	Std. Deviation	Minimum	Maximum	25th percentile	50th percentile	75th percentile
crimes_against_public_order	4.000	12.638	18.818	0.000	134.000	1.000	4.000	18.000
crimes_against_property	15.000	98.443	251.517	0.000	3046.000	3.000	15.000	76.500
total_crime_incidents	20.000	111.081	264.803	0.000	3102.000	4.750	20.000	95.500
trail	928.798	1319.858	1263.243	0.000	6486.824	348.982	928.798	1930.453
neighborhood_level_road	3272.395	3326.357	1614.852	0.000	8391.745	2159.814	3272.395	4310.063
city_level_road	38.260	274.173	425.750	0.000	2585.270	0.000	38.260	452.056
landuse_dis_dense_urban	32662.369	47333.805	49289.457	0.000	216129.442	2921.570	32662.369	77876.901
landuse_dis_medium_dense_urban	8891.326	29831.141	43253.475	0.000	213996.986	0.000	8891.326	42707.932
landuse_industrial_commercial	15545.173	37888.070	53045.610	0.000	230309.119	0.000	15545.173	56366.240
landuse_continuous_urban_fabric	0.000	12479.008	25723.346	0.000	166442.333	0.000	0.000	13504.292
landuse_green_area	70689.532	79476.687	56049.166	0.000	234078.615	31831.507	70689.532	116603.490
landuse_diversity	4.000	3.652	0.989	1.000	5.000	3.000	4.000	4.000
commercial_poi	0.000	5.306	18.251	0.000	233.000	0.000	0.000	4.000
recreation_poi	0.000	0.839	1.864	0.000	18.000	0.000	0.000	1.000
public_poi	1.000	5.344	13.277	0.000	146.000	0.000	1.000	5.000
camera_surveillance_poi	0.000	8.099	23.709	0.000	197.000	0.000	0.000	1.000
housing_small_apartment	0.000	1.948	4.201	0.000	43.000	0.000	0.000	2.000
housing_large_apartment	8.000	18.255	27.349	0.000	194.000	1.000	8.000	24.000
housing_standalone_house	10.000	34.422	45.358	0.000	222.000	1.000	10.000	60.500
housing_auxiliary_house	0.000	1.907	12.687	0.000	127.000	0.000	0.000	0.000
housing_others	2.000	5.959	15.194	0.000	163.000	0.000	2.000	5.000
age_0_9	41.000	95.033	120.663	0.000	562.000	11.000	41.000	134.000
age_10_19	48.000	88.861	110.851	0.000	791.000	11.000	48.000	118.000
age_20_29	40.000	95.961	126.066	0.000	715.000	11.000	40.000	138.500
age_30_39	54.000	162.106	214.500	0.000	907.000	14.500	54.000	245.000
age_40_49	62.000	133.333	170.114	0.000	817.000	15.000	62.000	181.000
age_50_59	51.000	106.257	136.272	0.000	611.000	12.500	51.000	138.500
age_60_69	41.000	102.994	153.507	0.000	843.000	10.000	41.000	113.500
age_70_79	26.000	70.565	108.785	0.000	701.000	5.000	26.000	80.000
age_80_89	13.000	43.524	73.054	0.000	426.000	3.000	13.000	46.500

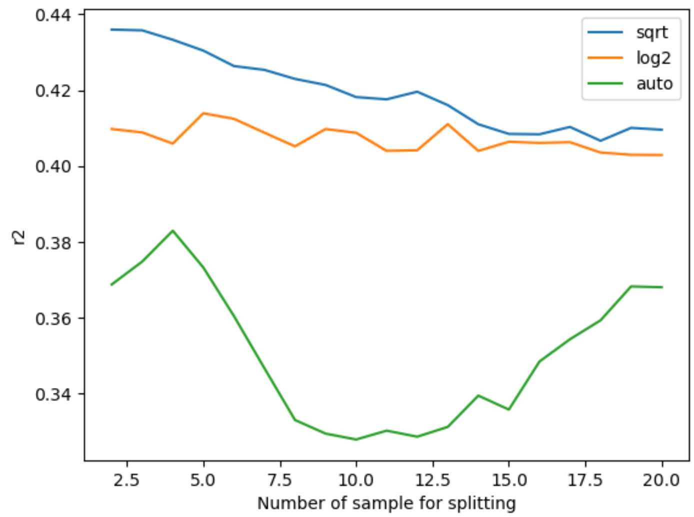
Descriptive Statistics

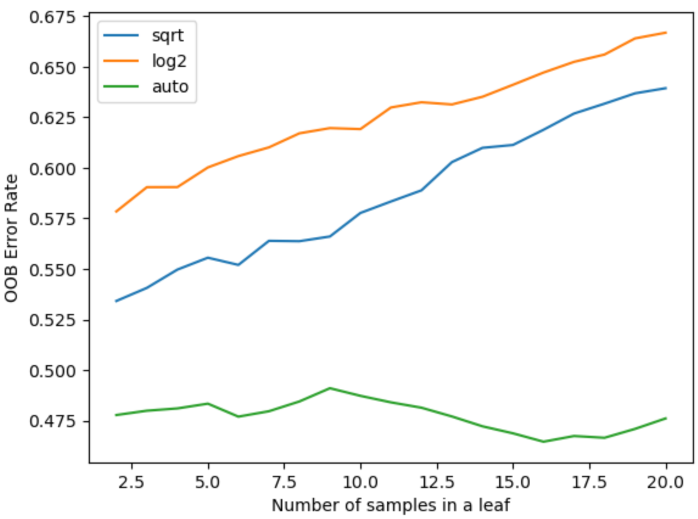
	Median	Mean	Std. Deviation	Minimum	Maximum	25th percentile	50th percentile	75th percentile
age_90_plus	3.000	7.698	12.378	0.000	69.000	0.000	3.000	9.000
ethnical_est	287.000	483.435	569.140	0.000	2786.000	71.000	287.000	688.000
ethnical_rus	54.000	310.288	575.961	0.000	3390.000	11.000	54.000	287.000
isco_high	93.000	166.004	199.542	0.000	1119.000	27.500	93.000	230.500
isco_medium	48.000	109.126	136.403	0.000	593.000	14.000	48.000	156.000
isco_low	57.000	187.199	291.557	0.000	1497.000	16.000	57.000	215.000
housing_buld_age_1945_before	3.000	92.240	265.123	0.000	2157.000	0.000	3.000	35.000
housing_buld_age_1946_1990	137.000	610.389	1043.769	0.000	5459.000	9.000	137.000	627.000
housing_buld_age_1991_after	79.000	199.453	312.941	0.000	2712.000	13.500	79.000	235.500
housing_owner	300.000	626.919	816.284	0.000	4322.000	77.000	300.000	824.000
housing_renter	49.000	183.513	269.187	0.000	1661.000	11.500	49.000	272.000

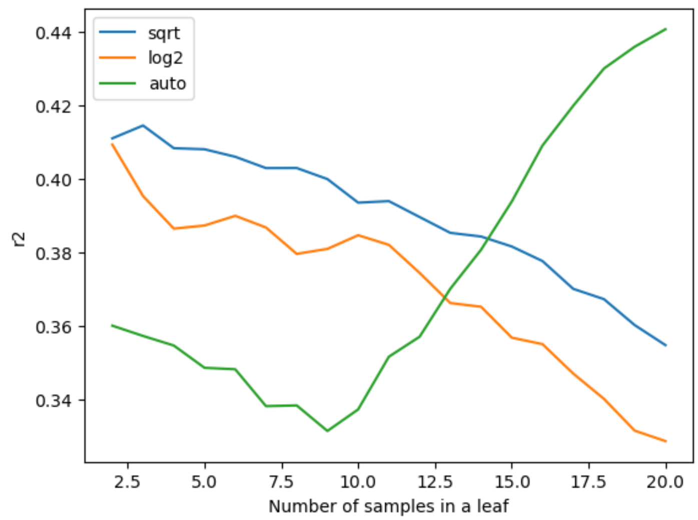
Appendix 3. The tuning results of the total number of crimes incident model



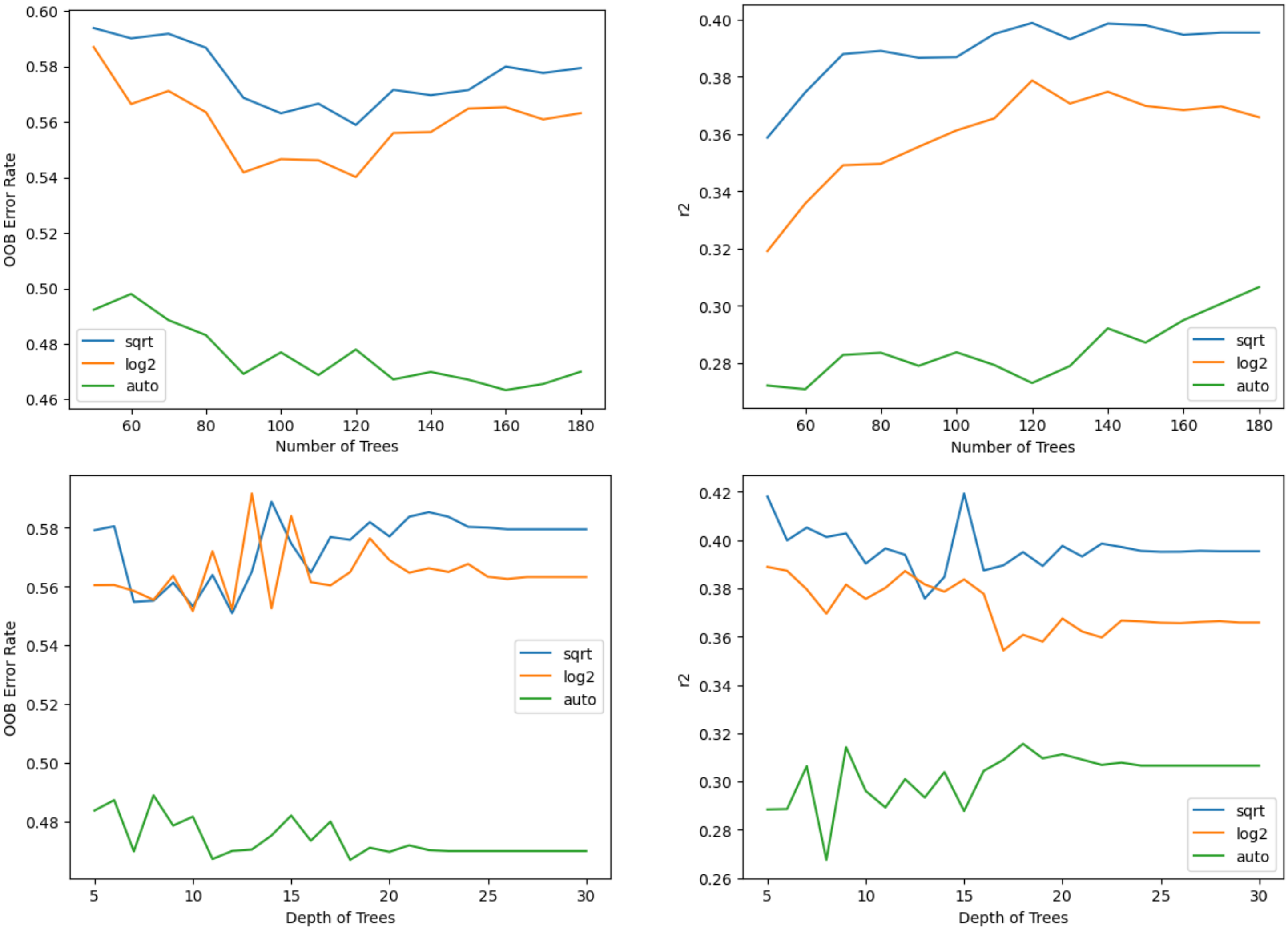

OOB Error Rate
Number of sample for splitting
sqrt
log2
auto

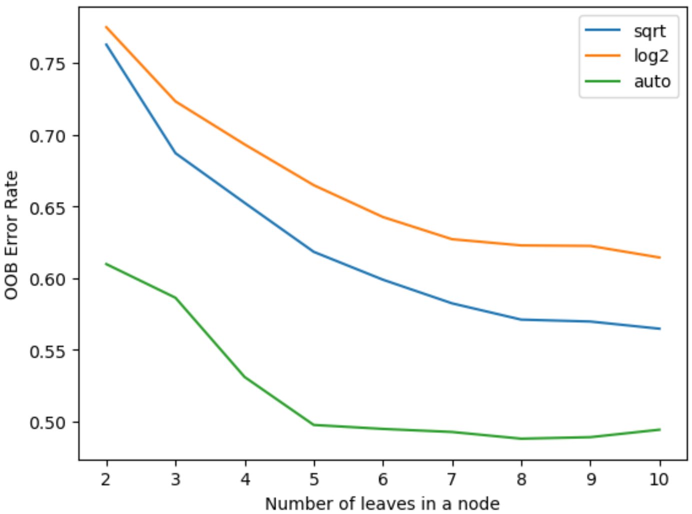

r2
Number of sample for splitting
sqrt
log2
auto

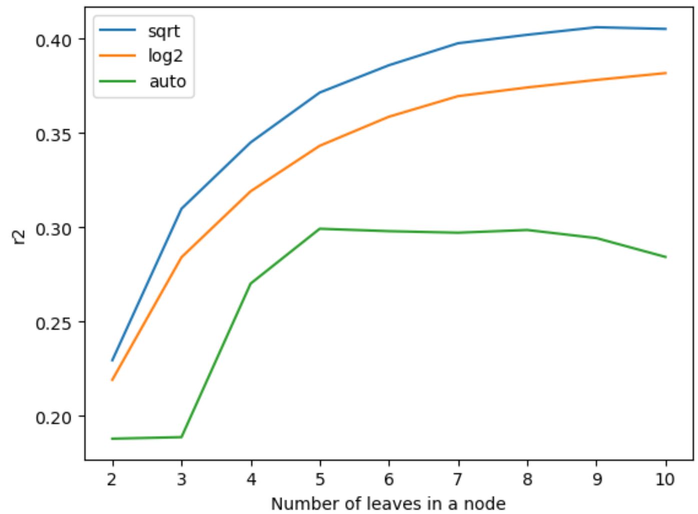

OOB Error Rate
Number of samples in a leaf
sqrt
log2
auto

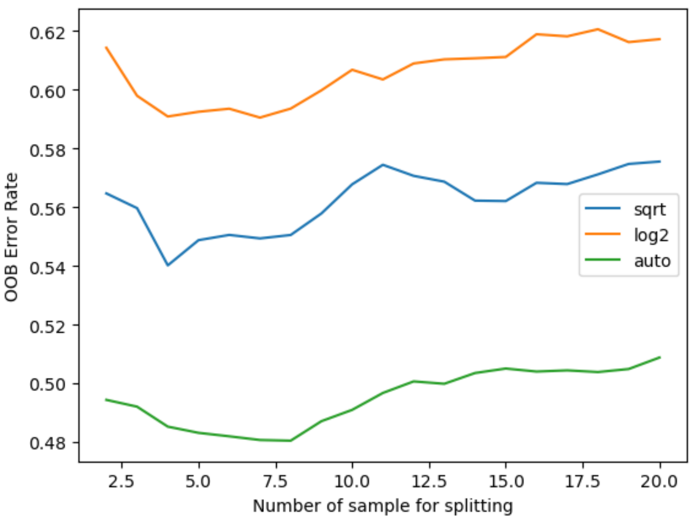

r2
Number of samples in a leaf
sqrt
log2
auto

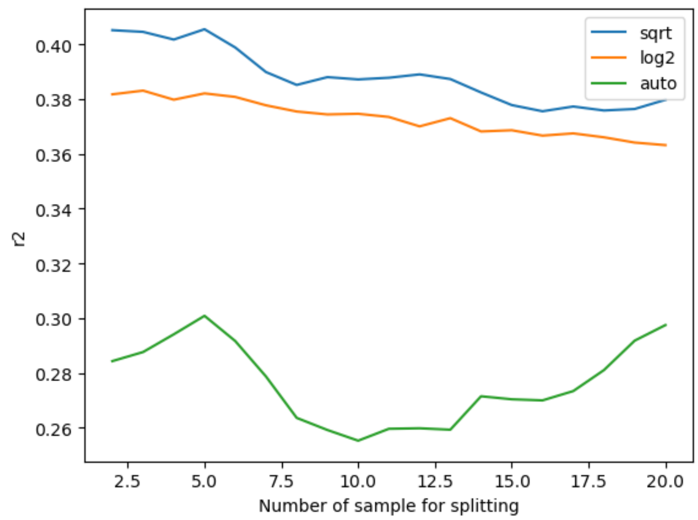
Appendix 4. The tuning results of the number of crimes against property model

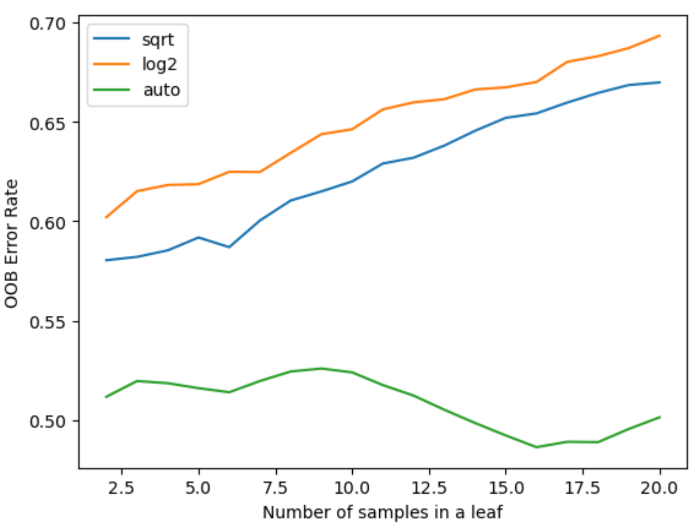


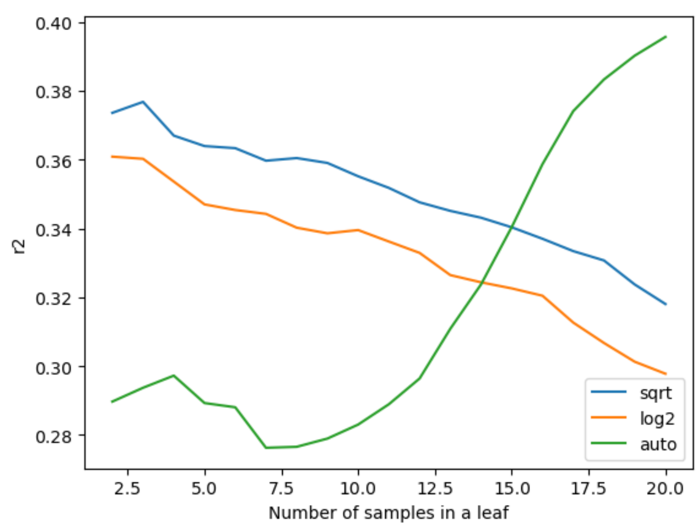

sqrt
log2
auto
OOB Error Rate
0.75
0.70
0.65
0.60
0.55
0.50
2 3 4 5 6 7 8 9 10
Number of leaves in a node


sqrt
log2
auto
r2
0.40
0.35
0.30
0.25
0.20
2 3 4 5 6 7 8 9 10
Number of leaves in a node

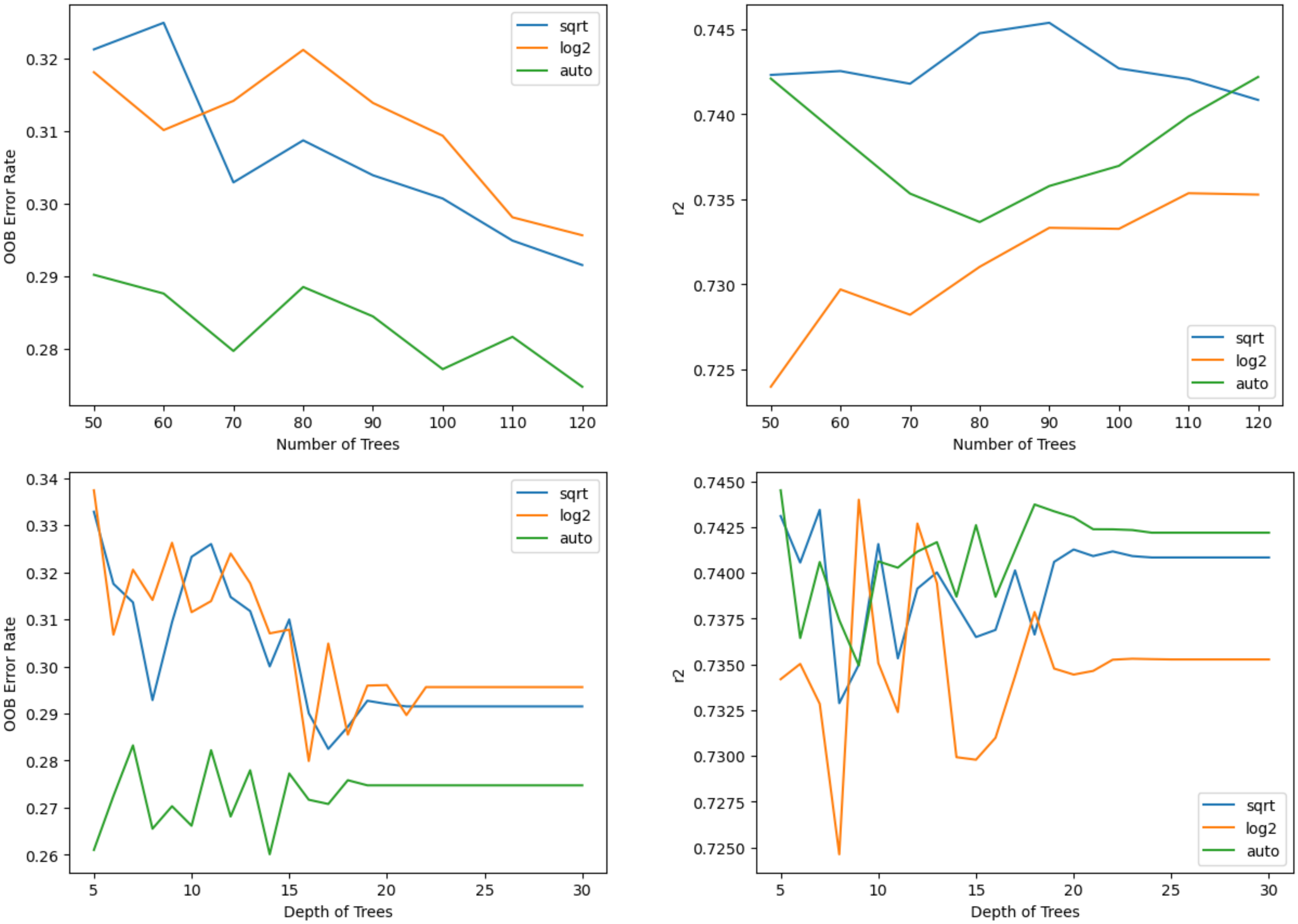

sqrt
log2
auto
OOB Error Rate
0.62
0.60
0.58
0.56
0.54
0.52
0.50
0.48
2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
Number of sample for splitting

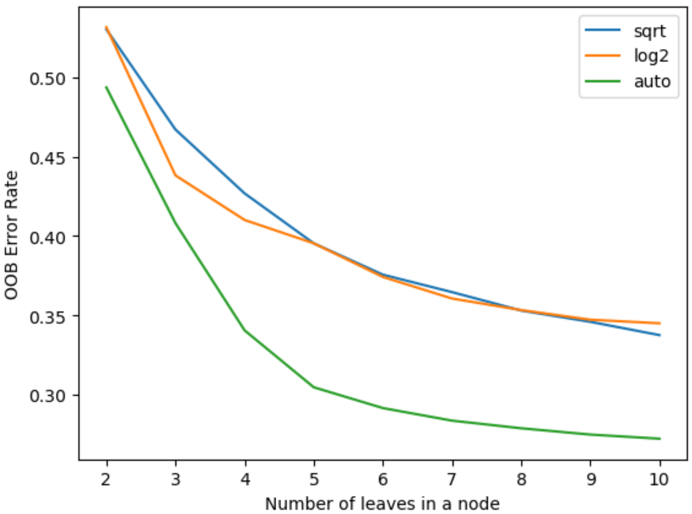

sqrt
log2
auto
r2
0.40
0.38
0.36
0.34
0.32
0.30
0.28
0.26
2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
Number of sample for splitting

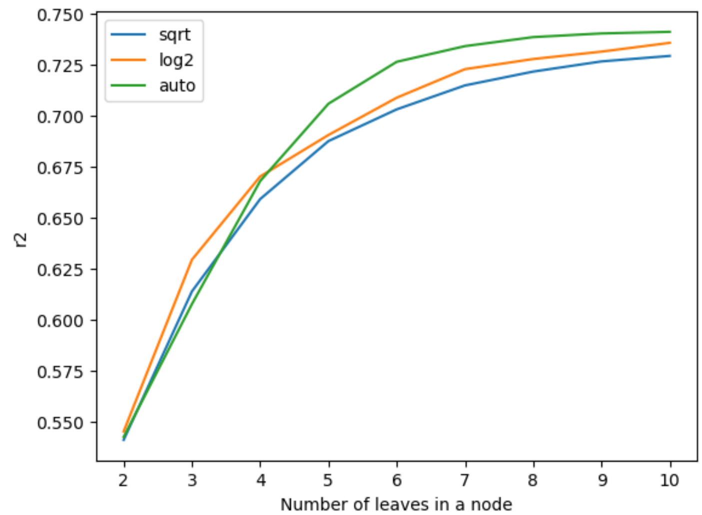

sqrt
log2
auto
OOB Error Rate
0.70
0.65
0.60
0.55
0.50
2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
Number of samples in a leaf

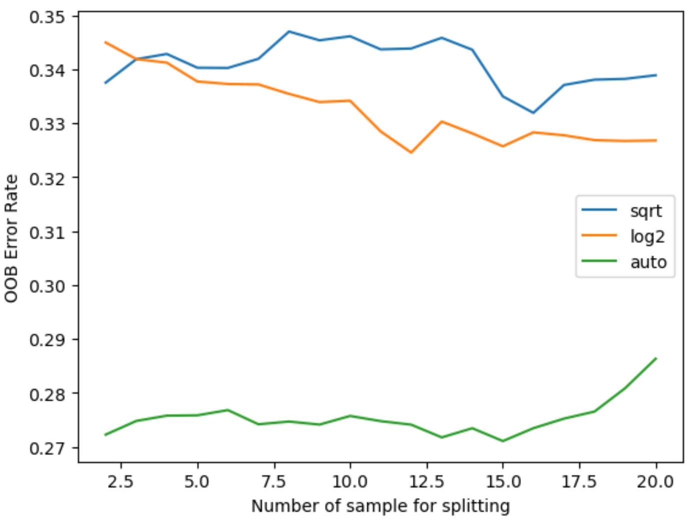

sqrt
log2
auto
r2
0.40
0.38
0.36
0.34
0.32
0.30
0.28
2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
Number of samples in a leaf

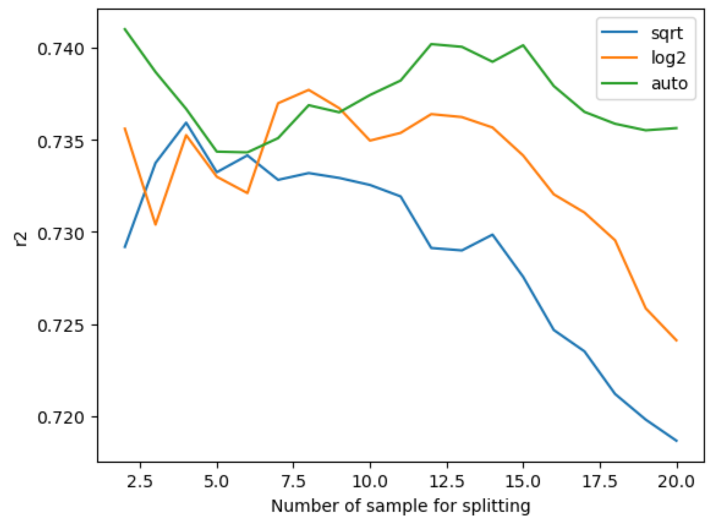
Appendix 5. The tuning results of the number of crimes against public order model

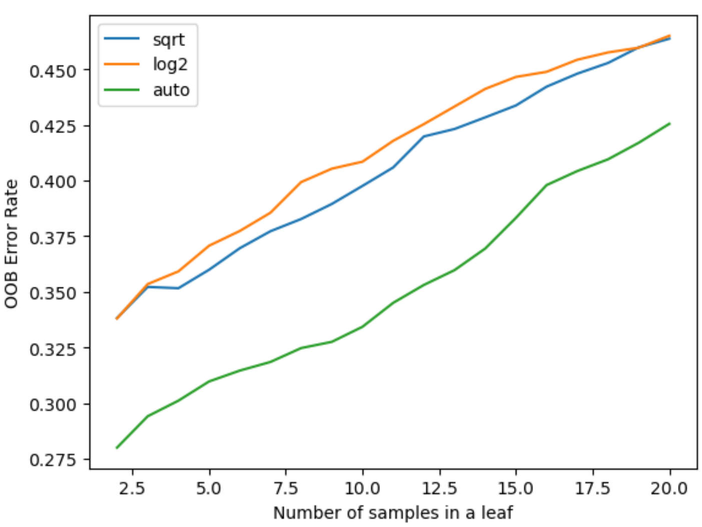


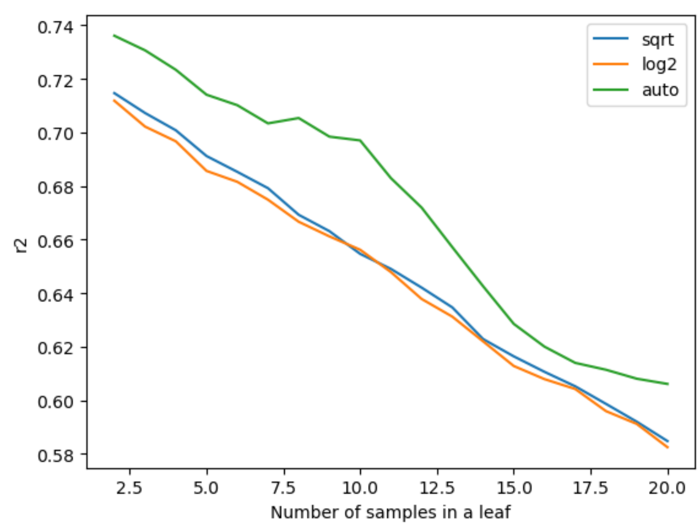

sqrt
log2
auto
OOB Error Rate
0.50
0.45
0.40
0.35
0.30
2 3 4 5 6 7 8 9 10
Number of leaves in a node

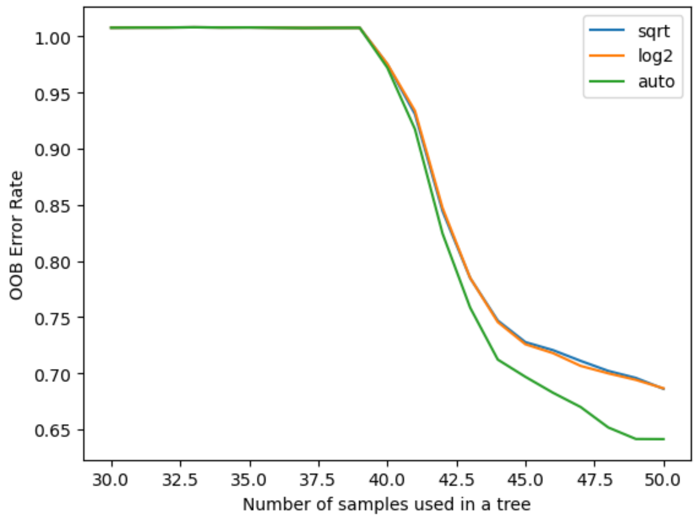

sqrt
log2
auto
r2
0.750
0.725
0.700
0.675
0.650
0.625
0.600
0.575
0.550
2 3 4 5 6 7 8 9 10
Number of leaves in a node

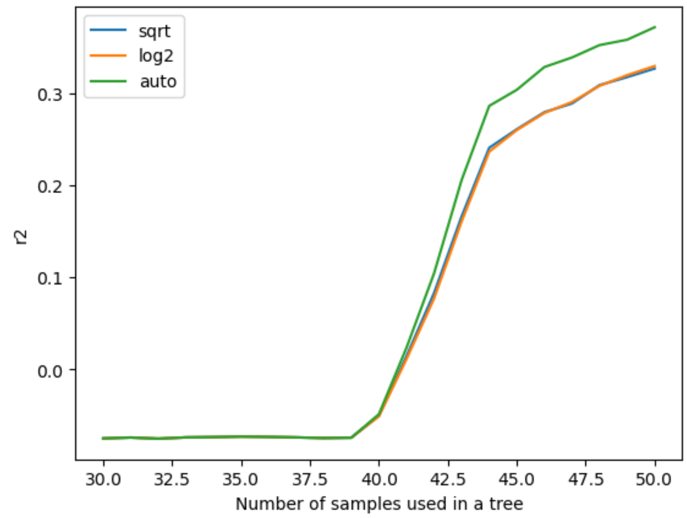

sqrt
log2
auto
OOB Error Rate
0.35
0.34
0.33
0.32
0.31
0.30
0.29
0.28
0.27
2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
Number of sample for splitting


sqrt
log2
auto
r2
0.740
0.735
0.730
0.725
0.720
2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
Number of sample for splitting

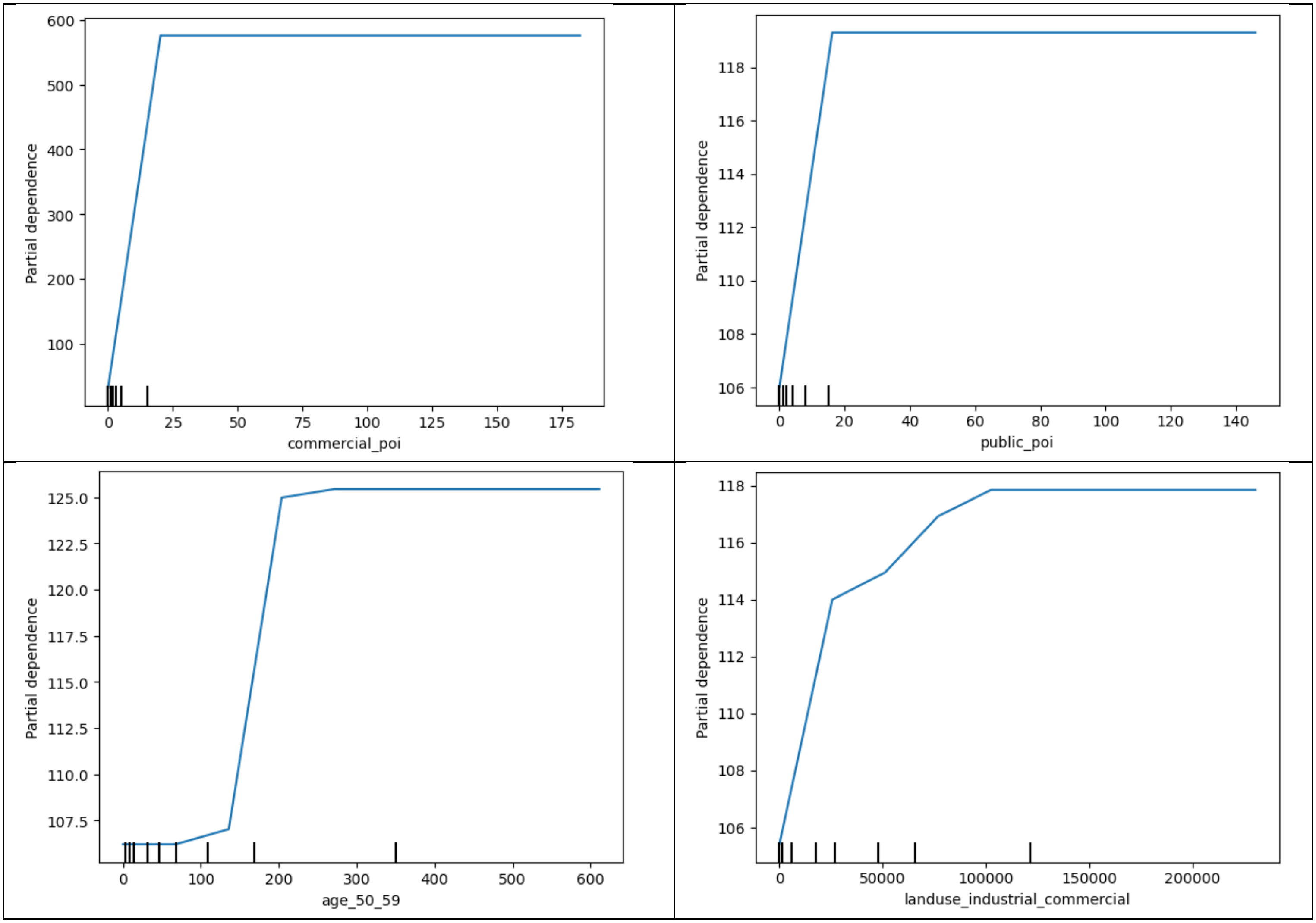

sqrt
log2
auto
OOB Error Rate
0.450
0.425
0.400
0.375
0.350
0.325
0.300
0.275
2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
Number of samples in a leaf

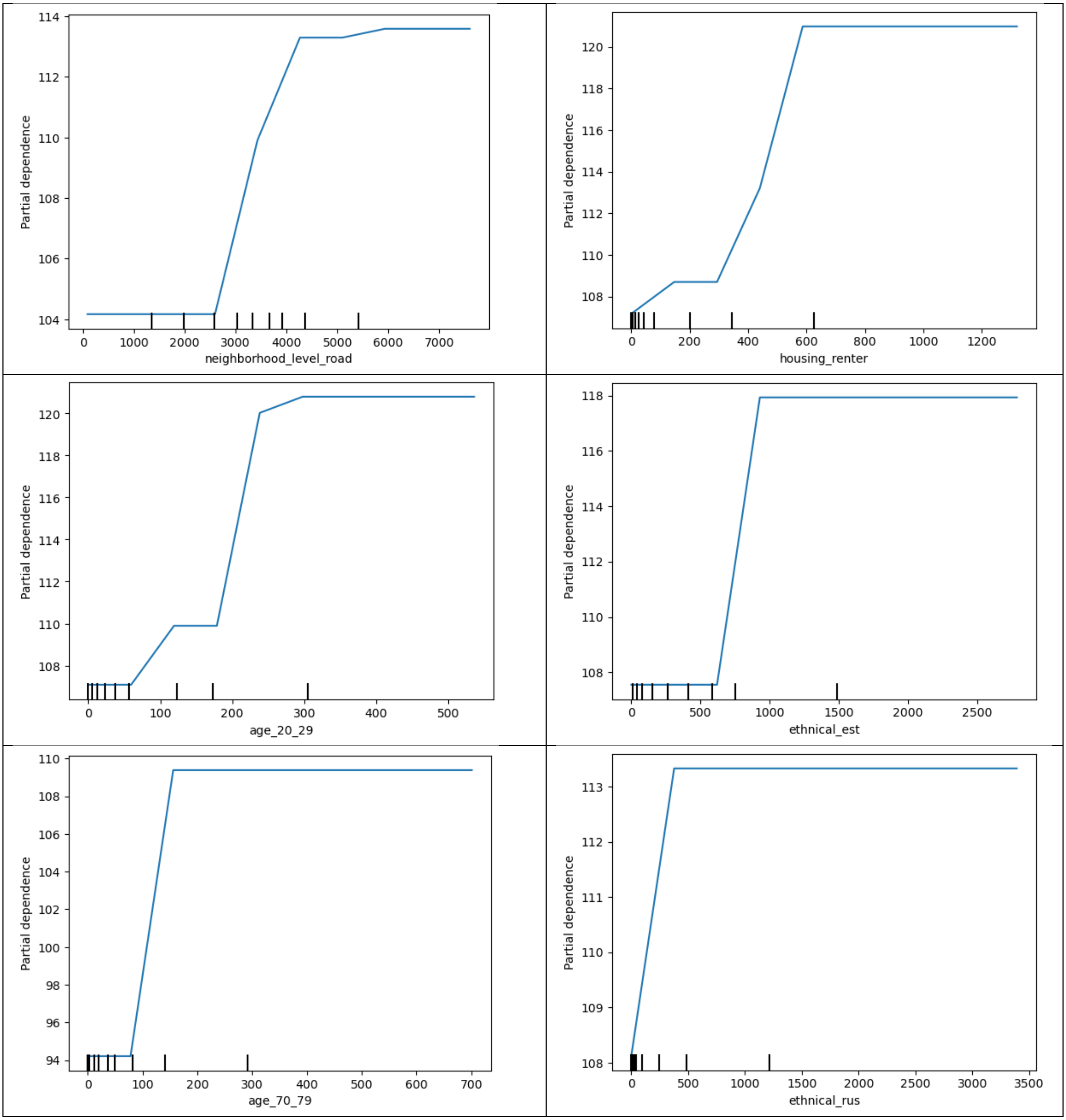

sqrt
log2
auto
r2
0.74
0.72
0.70
0.68
0.66
0.64
0.62
0.60
0.58
2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
Number of samples in a leaf


OOB Error Rate
Number of samples used in a tree
sqrt
log2
auto

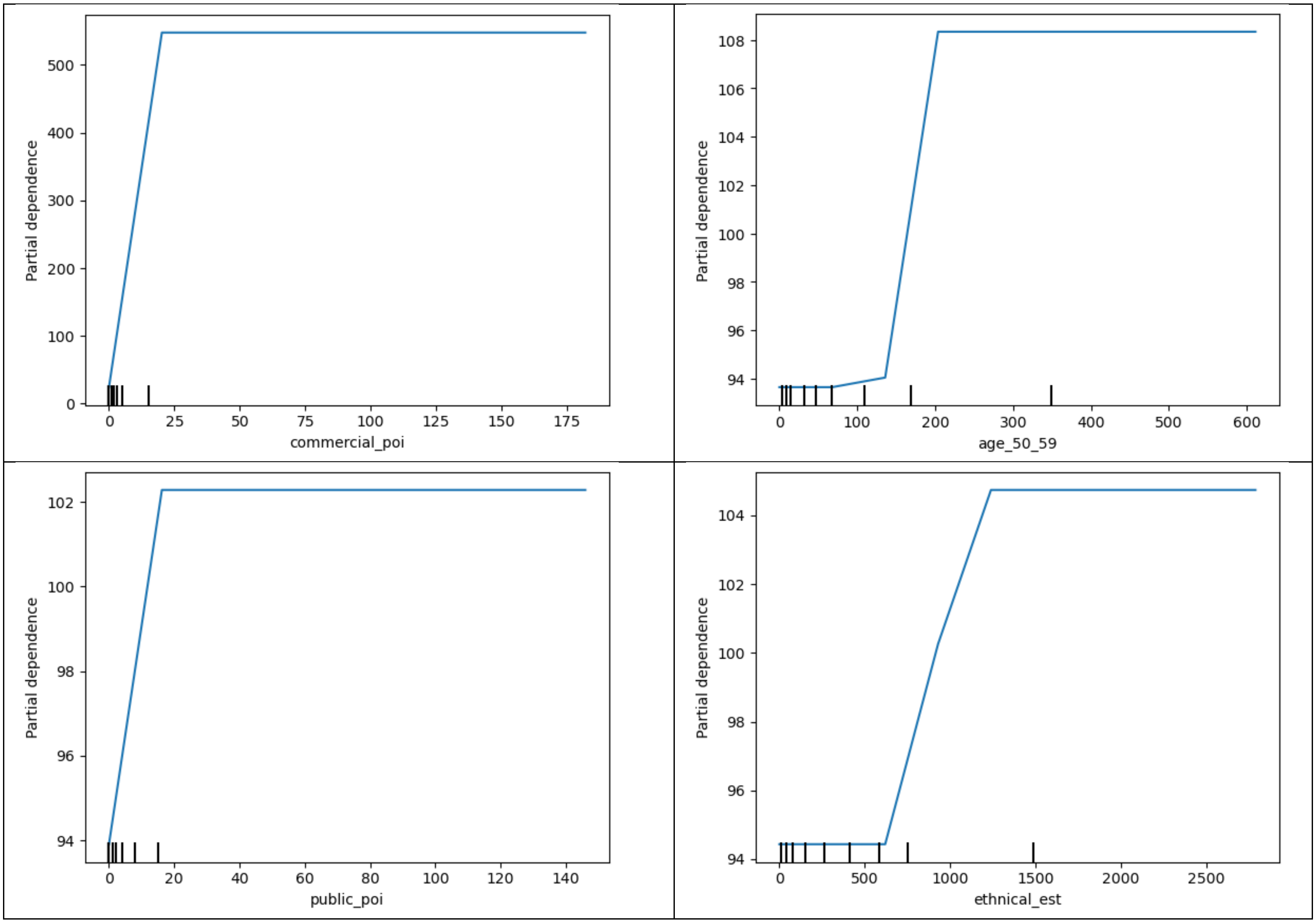

r2
Number of samples used in a tree
sqrt
log2
auto

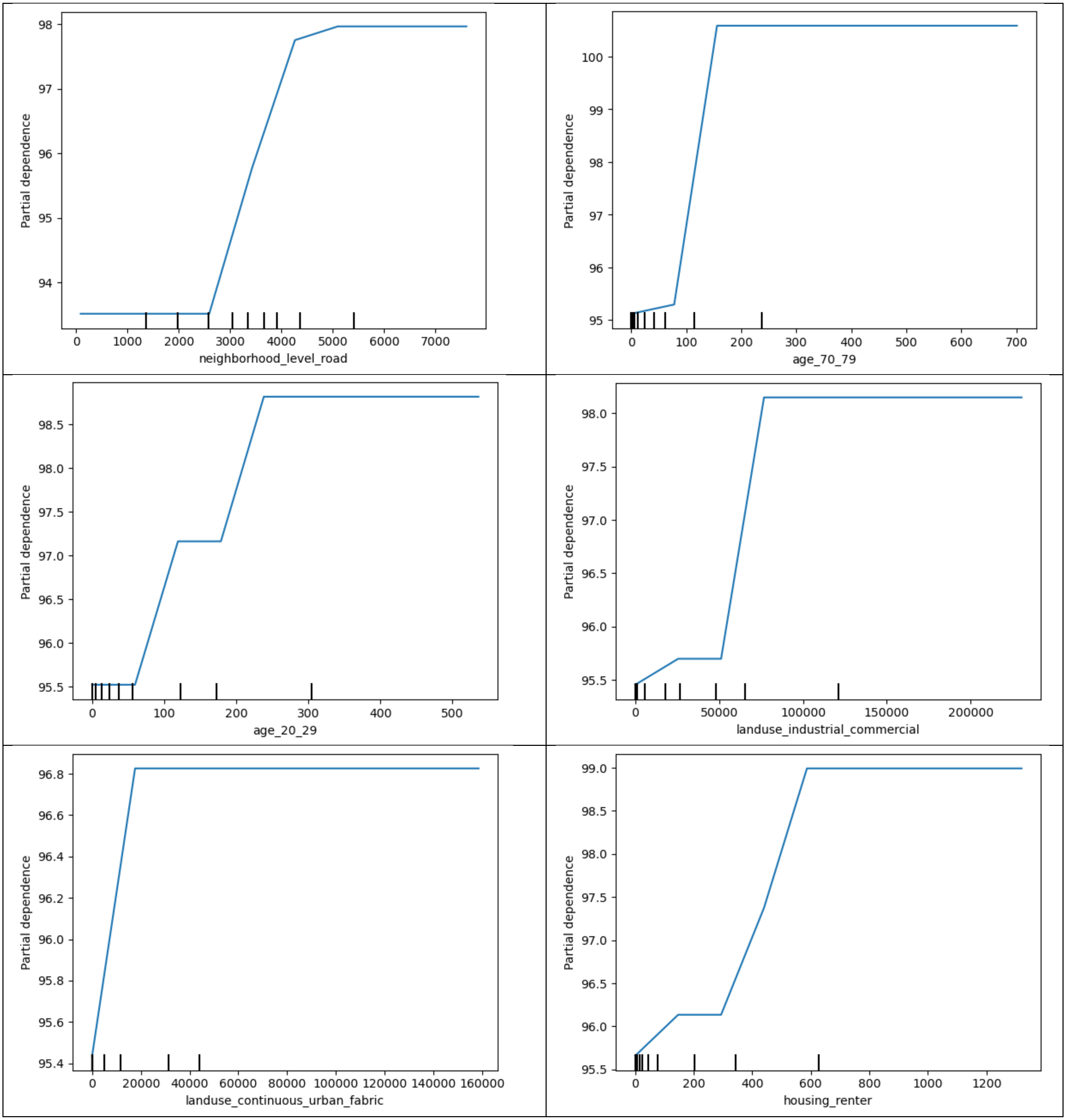
Appendix 6. The partial dependent plots of the top 10 important features for the number of total crime incidents model



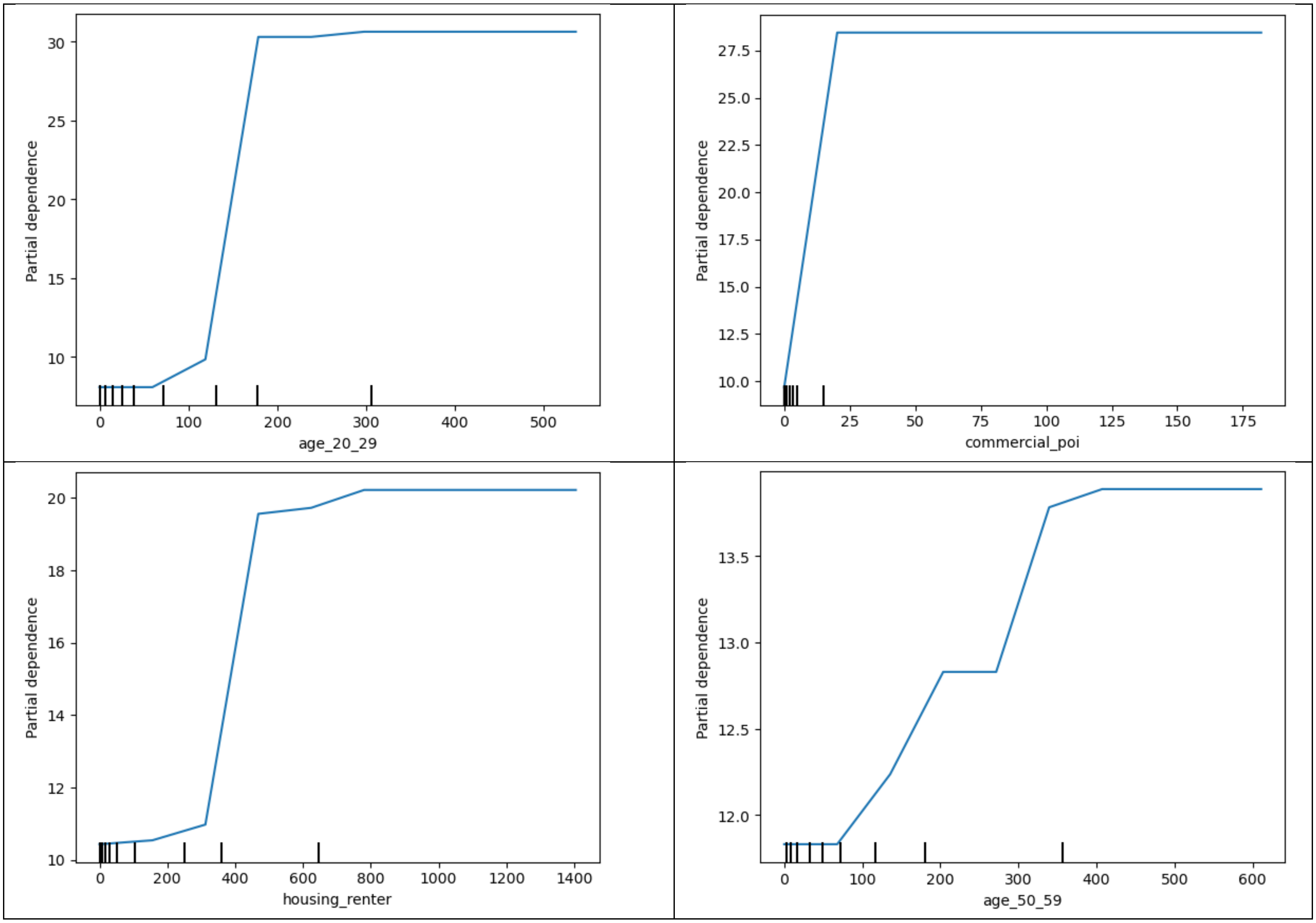

Partial dependence
neighborhood_level_road
Partial dependence
housing_renter
Partial dependence
age_20_29
Partial dependence
ethnical_est
Partial dependence
age_70_79
Partial dependence
ethnical_rus

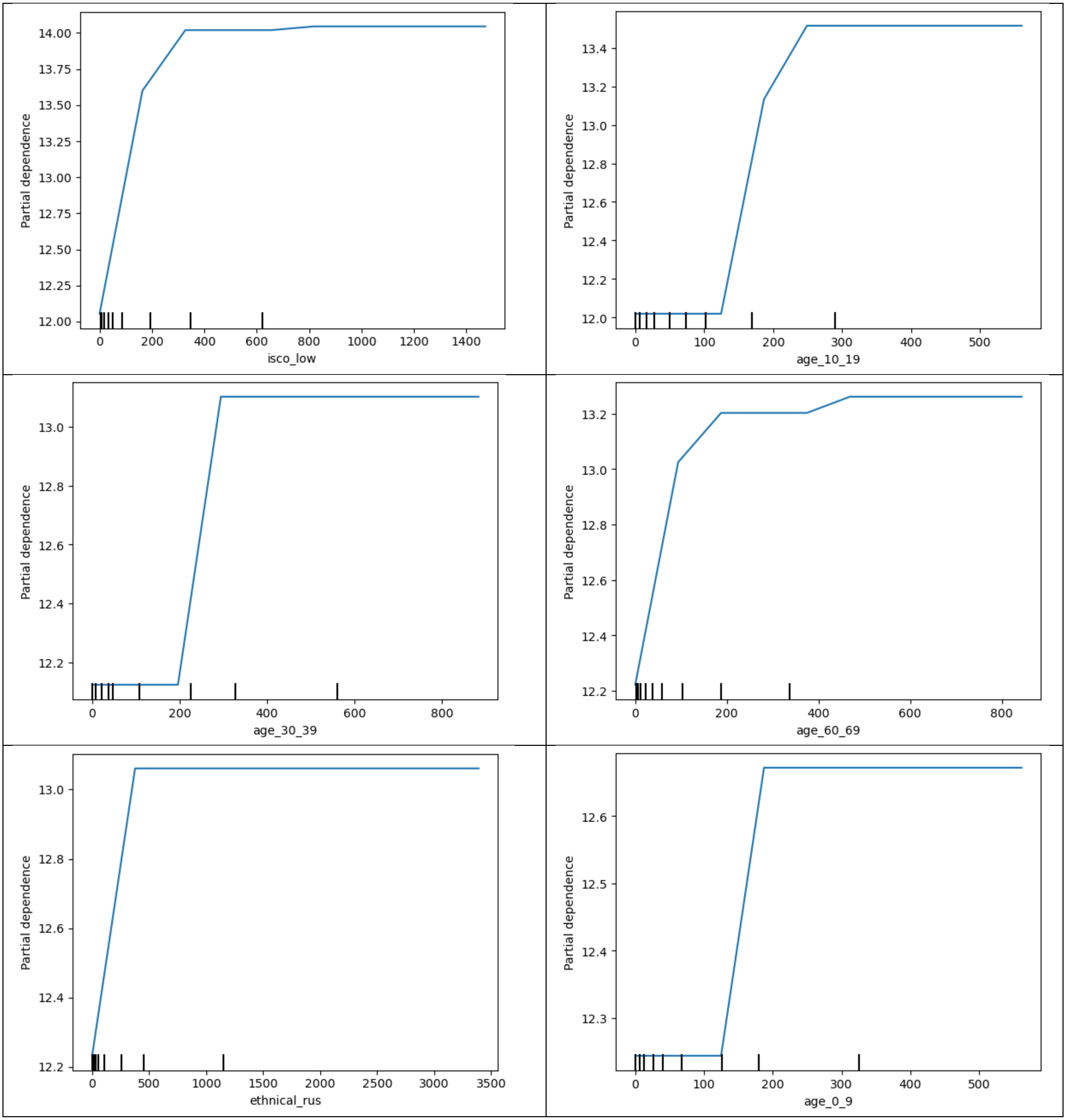
Appendix 7. The partial dependent plots of the top 10 important features for the number of crimes against property model




Partial dependence
98
97
96
95
94
0 1000 2000 3000 4000 5000 6000 7000
neighborhood_level_road
Partial dependence
100
99
98
97
96
95
0 100 200 300 400 500 600 700
age_70_79
Partial dependence
98.5
98.0
97.5
97.0
96.5
96.0
95.5
0 100 200 300 400 500
age_20_29
Partial dependence
98.0
97.5
97.0
96.5
96.0
95.5
0 50000 100000 150000 200000
landuse_industrial_commercial
Partial dependence
96.8
96.6
96.4
96.2
96.0
95.8
95.6
95.4
0 20000 40000 60000 80000 100000 120000 140000 160000
landuse_continuous_urban_fabric
Partial dependence
99.0
98.5
98.0
97.5
97.0
96.5
96.0
95.5
0 200 400 600 800 1000 1200
housing_renter

Appendix 8. The partial dependent plots of the top 10 important features for the number of crimes against public order model




Partial dependence
isco_low
age_10_19
age_30_39
age_60_69
ethnical_rus
age_0_9

The form for the Non-exclusive licence

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Cheng-Wei Yu,

1. grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

Analyzing the relationships between crime and socioeconomic and spatial factors using random forest: a case study of Tallinn,

supervised by Prof. Evelyn Uuemaa, PhD Kadi Kalm, PhD Alexander Kmoch, MSc Jānis Zālīte.

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Cheng-Wei Yu
29/05/2024