


Article

# Artificial Neural Network Model Development to Predict Theft Types in Consideration of Environmental Factors

Eunseo Kwon <sup>1</sup>, Sungwon Jung <sup>1,\*</sup> and Jaewook Lee <sup>2</sup> 

<sup>1</sup> Department of Architecture, Sejong University, Seoul 05006, Korea; kes2526kr@naver.com

<sup>2</sup> Department of Architectural Engineering, Sejong University, Seoul 05006, Korea; jaewook@sejong.ac.kr

\* Correspondence: swjung@sejong.ac.kr; Tel.: +82-02-3408-3289

**Abstract:** Crime prediction research using AI has been actively conducted to predict potential crimes—generally, crime locations or time series flows. It is possible to predict these potential crimes in detail if crime characteristics, such as detailed techniques, targets, and environmental factors affecting the crime’s occurrence, are considered simultaneously. Therefore, this study aims to categorize theft by performing k-modes clustering using crime-related characteristics as variables and to propose an ANN model that predicts the derived categorizations. As the prediction of theft types allows people to estimate the features of the possibly most frequent thefts in random areas in advance, it enables the efficient deployment of police and the most appropriate tactical measures. Dongjak District was selected as the target area for analysis; thefts in the district showed four types of clusters. Environmental factors, representative elements affecting theft occurrence, were used as input data for a prediction model, while the factors affecting each cluster were derived through multiple linear regression analysis. Based on the results, input variables were selected for the ANN model training per cluster, and the model was implemented to predict theft type based on environmental factors. This study is significant for providing diversity to prediction methods using ANN.

**Keywords:** artificial neural network; k-modes clustering; crime prediction; smart city; urban security



**Citation:** Kwon, E.; Jung, S.; Lee, J. Artificial Neural Network Model Development to Predict Theft Types in Consideration of Environmental Factors. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 99. <https://doi.org/10.3390/ijgi10020099>

Academic Editors: Wolfgang Kainz and Giuseppe Borruso

Received: 29 December 2020

Accepted: 19 February 2021

Published: 22 February 2021

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

As society and science technology develop, the technology of crime is becoming more sophisticated and the damage caused by crime is increasing. However, it is not easy to prevent or react to crime in a situation where the population is concentrated in a specific area and the city is changing rapidly. Therefore, various measures for crime prevention have been presented in the urban planning field as part of a smart city plan. In Korea, various attempts are being made; for example, a linkage system is being established and implemented to provide CCTV image information to the Smart City Center and the location center of the Ministry of Justice, and a smart city safety net linked to the national disaster safety system is also being promoted as a smart city integrated platform. In addition, recent studies on crime prediction for more efficient crime prevention are actively underway. Crime is a social phenomenon in which many factors such as physical, social, and economic elements function in combination; since there are diverse criminogenic factors, the correlation between factors and crimes is beyond the range of human perception. To overcome these limitations, artificial intelligence is being actively used in crime research, in particular in crime prediction activities using data mining. Data mining is a technique that excludes user experiences or subjectivity in a large amount of data, and identifies statistical rules, patterns, and correlations entirely based on data. It is a powerful tool that enables crime investigators to quickly and efficiently navigate large databases [1].

A classification method is used to predict potential crime events in specific times and areas, through clustering, which is a type of data mining technique. Clustering refers to grouping data with similar properties and classifying the entire dataset into several

clusters, that is, categorization. If crimes are categorized by simultaneously considering several subdivided categories such as location, time, and method of crime, it is possible to understand the detailed characteristics of each crime. More accurate crime prediction is feasible if crime investigators can predict certain types of crime events in particular areas and at particular times. However, while human-led categorization has limits in terms of simultaneously considering several factors, objective classification is impossible because of the researchers' subjective interpretation. On the other hand, clustering enables researchers to objectively classify several factors, to use them as variables, and eventually to objectively categorize crimes. In crime-related research, clustering is mainly used to analyze crime hotspots, largely by clustering crime locations or time series flows [2,3], while the derived patterns of clustering are used to identify crime trends [4].

Crime categorization based on crime locations or time series flows is a basic measure to prevent crime, and it can be applied to various additional prediction methods. Crime characteristics, such as detailed techniques, targets, and environmental factors affecting its occurrence, reveal different crime patterns of crime targets, crime plans, and escape plans, even for the same types of crime. Therefore, in terms of crime prediction, it is possible to predict crime patterns in more detail and clarity based on the locations, dates, and times of crime events, and the characteristics of each crime.

This study proposes an artificial neural network model that predicts the actual crime types by categorizing specific crimes via clustering. Besides predicting crime locations and time series flows, the prediction of specific crime types is suggested as a method for crime prediction. It is possible to develop a model that can predict a crime type if crimes are categorized after crime-related elements are used as variables and clustered, and if the derived crime type data are learned in an artificial neural network model. As crime type prediction allows people to determine the features of the most frequent crimes in random areas in advance, it enables the efficient deployment of police and the most appropriate tactical measures.

To predict crime types, it is necessary to find the factors influencing crime occurrence. Since Oscar Newman's "Defensible Space" [5] presented the possibility that crime can be prevented through urban design, there has been increased research on environmental criminology focusing on environmental factors in communities, not offenders, in finding criminogenic factors. In particular, environmental criminology mainly focuses on a spatial environment among the environmental factors affecting crime events. Crime pattern theory, one of the theoretical foundations of environmental criminology, concerns the interactions between the surrounding environment and offenders' crime location decisions. This theory emphasizes the importance of land use and facilities, understanding crime patterns, and identifying the facilities that influence crime occurrence [6]. In addition, in the theory of Crime Prevention Through Environmental Design asserted by Jeffery [7] according to environmental criminology, streetlights, and CCTV are important natural surveillance factors that reduce the incidence of crime. Previous studies that analyzed the effects of local environmental factors on crime incidence rates based on such theories showed that certain spatial properties and neighborhood environments (facilities, CCTV, etc.) affect crime incidence rates [8].

This study targeted theft among five types of crime because this type is most affected by neighboring environments. Crimes such as murder and assault are highly influenced by personal feelings between offenders and victims because the targets are specific individuals. On the other hand, theft is more influenced by neighboring environments and behavioral characteristics of offenders rather than personal feelings because the crime targets are specific buildings or objects [9]. Nevertheless, there is still insufficient spatial analysis studies on specific crimes such as theft, and it is also necessary to analyze crimes in the South Korean environment. Accordingly, this study attempts to estimate regional differences in theft locations in a spatial context, and as this research focuses on environmental factors, it is considered to adopt an environmental criminology approach toward theft.

Finally, this study categorizes theft through clustering, in consideration of various characteristics of the crime simultaneously; based on local environmental factors that affect theft, this study suggests an ANN model that can predict the types of theft that are most likely to occur in random areas. Rather than predicting crimes by simply using environmental factors, it categorizes specific crimes and subdivides the relationships between the derived categorizations and environmental factors, thereby giving diversity to prediction methods using ANN. This analysis enables policy makers to derive types of theft that are most likely to occur in random areas, to efficiently deploy police personnel, and to take the most appropriate tactical measures. In addition, if local characteristics influencing thefts are identified, it is possible to propose policy alternatives to secure the safety of the area, contributing to the safety of its residents.

## 2. Theoretical Background

### 2.1. Crime and Environment

Environmental criminology is a theory in which crime is caused by four factors: “laws, objects, offenders, and locations”. In particular, it considers “locations” as the crucial factor [10]. Crime pattern theory, one of the theoretical foundations of environmental criminology, defines crime locations as “places that provide environmental clues that meet the criteria for offenders who have learned characteristics of appropriate crime locations.” Factors in causing and deterring crime and target selection factors are categorized in line with area types, street networks, building locations, economic levels, and surveillability to identify crime occurrence patterns. The theory describes how human behaviors interact with the environment to create crime event patterns and generally emphasizes the importance of land use types and facilities. Certain land uses and facilities are associated with people’s daily activities, and the numbers and types of those who use them affect crime occurrence. Therefore, the crime levels in the surroundings can theoretically be predicted through the status of land use types and facilities [11].

Land use is an umbrella term for land use activities by humans. It is related to the space selection for people’s daily activities, and it can be regarded as people’s various activities on land with certain physical structures or as certain forms of use such as types, purposes, and densities to accommodate the activities [12]. Since criminal activities occur in land uses, it is necessary to analyze the characteristics of crime occurrences related to land use. Previous studies on the relationship between crime and land use based on environmental criminology [13] analyzed the association and found that specific land uses affect crime occurrences. In particular, commercial land and high-density residential areas were highly related to crime occurrences, whereas cemeteries, rivers, and factory sites were less related to the occurrences; this indicates that not all non-residential land uses are highly connected to crime. Studies on the relationship in South Korea are also mostly based on environmental criminology theory. In a study by Kim, Yoon, and Ahn [14], four districts in Seoul were targeted to analyze the correlation between land use and crime density, confirming that there were different possibilities of crime occurrence based on land use. In addition, their results revealed that crime density was relatively lower when the area showed one specific type of land use, whereas the density was higher when there were mixed land uses. However, there was a limitation in that this study was limited to correlation analysis. Lee [15] revealed the spatial relationship between land use and crime frequencies through a cluster analysis of bivariate correlation indices. The results confirmed that spatial factors had a substantial influence on the correlation between land use and crime and that the correlation of crime occurrences was higher with residential areas with apartment houses than with other land uses. Although it is known that crime rates are higher in commercial sites, it was found that if crimes begin to occur in residential areas with apartment houses, there is a high possibility that the crimes are clustered while their occurrences expand. However, the study is limited by employing the number of crime occurrences as a variable without classifying crime types.

Facilities, like land uses, are physical structures that provide people with space for daily life, and the numbers and types of users affect crime occurrences. Previous studies on the association have shown that many facilities affect crime occurrences. In particular, as local residents and many outsiders use commercial facilities, it is easy for offenders to approach targets without being spotted due to a large floating population [13]. There are several ongoing studies on the effects of schools, a representative educational and research facility, on crime occurrence, revealing that they mostly have a strong influence. Gouvis Roman [16] found that areas near schools showed higher crime rates than other residential areas, while Kaut and Roncek [17] also revealed that the existence of schools affects crime frequency in the neighborhood. Engstad [18] compared the number of crimes in an area with hotels to the number of the same crime types in an adjacent area without hotels. The number of crimes was standardized depending on the number of residents in the area, and the results showed that the existence of hotels affected the crime occurrence. Public transportation stops are places by which local residents and many outsiders pass, attracting diverse people to the surrounding area. Therefore, the places are claimed as criminogenic facilities as they attract a large number of drunk people or those who are unfamiliar with the surroundings, thus becoming easy targets for offenders [6]. Block and Davis [19] determined the concentrations of crime in the streets within a block and a half from Chicago subway stations; Block and Block [20] also found similar patterns around the Bronx subway station. The aging level of these facilities, that is, the age groups of buildings, also affects crime occurrences. Areas with higher age groups of buildings generally have poor aesthetics around their streets and buildings. Such housing deterioration increases crime fear among local residents and becomes a major factor in crime occurrence [21].

As facilities that increase crime rates, and crime prevention facilities that are installed to prevent crime, belong to local environmental factors, this study attempts to examine whether they affect crime occurrences in reality. CCTV can catch crime scenes and provide an opportunity to properly deploy security personnel or police. Many studies have proven the practical effects of CCTV. Brandon C. Welsh [22] analyzed the effects of CCTV on crime in public places and found that it contributed to lower crime rates in the experimental areas. The effective CCTV installation in parking lots resulted in a lower crime rate of 51%; 23% from the installation in public transportation facilities; and 7% in apartment houses. Phillips [23] examined whether CCTV was effective in reducing the crime incidence and found that there was less theft in specific areas where CCTV was installed. Streetlights are considered a necessary element for crime prevention because they are an economical and efficient method to reduce crime. In effect, Xu et al. [24] verified the relationship between streetlight density and crime incidence, confirming that the higher the former, the lower the latter.

## 2.2. Categorization of Crimes

This study examined previous relevant studies to ascertain whether crime categorization is practically useful in preventing crime. Categorization is a data mining technique used to analyze crime patterns, as well as one of the most commonly used major techniques. It is possible to extract meaningful information from large-scale data that can be effectively used to predict undefined types of crime. Such potential and effectiveness contribute to more research on crime categorization, and most studies have focused on the analysis of temporal and spatial changes of crime through categorization. Nakaya and Yano [3] confirmed that the geographic extent and duration of theft and escape crimes in Kyoto City between 2003 and 2004 can be simultaneously visualized through clustering. Furthermore, the correlations between temporal clusters were found by using the Spatio-Temporal Kernel Density Estimation (STKDE) and the Space-Time Scan Statistics (STSS) in parallel. As a result, there were transient clusters that were alternately generated between two clusters, which means that there was a displacement phenomenon in which offenders selected the areas with crime occurrences as their crime target areas. Based on these results, the study concluded that the space-time data analysis of clusters was valuable in “extracting new

knowledge of crime epidemiology.” Park [25] employed the K-means clustering technique as a method to analyze spatial patterns of crime data and spatial distribution of crime hotspots. As a result of extracting 10 clusters for each year from 2000 to 2002, it was found that the sizes and patterns of the clusters showed high similarity and that the crime locations indicated similar spatial patterns over the specified period. The resultant factors were digitized and, finally, a crime prediction map was implemented. The map was expected to help prevent crime along with the efficient distribution of the police force, with the implication that if there is a comparative analysis with the collected crime data in the future, the map’s accuracy will be improved. Although few studies have attempted to categorize crime in South Korea, some research has analyzed the influence of the physical environment in association with crime types and crime occurrence time. Kwak et al. [26] predicted that environmental factors were expected to have different impacts on crime occurrence depending on crime types and crime occurrence time, such as day and night. They analyzed those impacts based on those factors and found that the impact of the density of each building use type on crime occurrence was greater at night than during the daytime. In particular, the areas with concentrated factory facilities did not show significance with regard to crime occurrences during the daytime, but such significance was shown at night. This was found to have a greater impact only on thefts, not on all crimes. As such, it was found that urban environmental factors had different influential relationships by crime type and crime occurrence time. Some studies have estimated crimes by categorizing them into crime types or time zones but few have categorized crimes while considering these factors in combination.

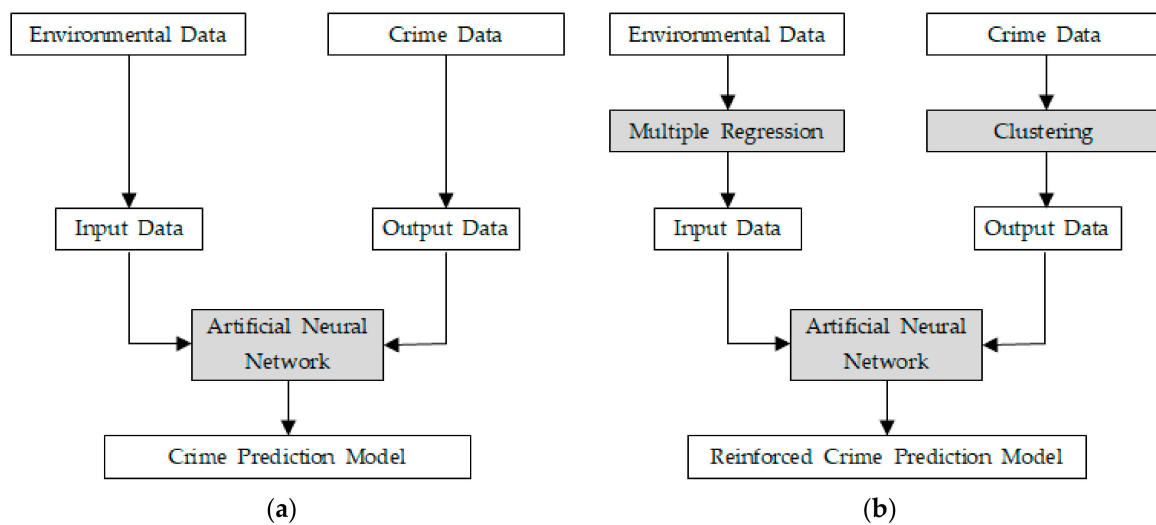
### 2.3. Crime Prediction Utilizing ANN

The artificial neural network (ANN) is a unique data processing method that has been continuously developed and is a statistical learning algorithm that approximates the mapping relation between input and output signals using a model inspired by how brains biologically respond to sensory stimuli. ANN has been used as a crime prediction model in crime research and showed high accuracy. Mahmud et al.’s [27] study introduced Crimecast, a crime prediction and strategy direction service that adopted an ANN model using variables such as crime rates, crime locations, date of crime, type of crime, etc., and proved that the model has high accuracy. Kang and Kang [28] also developed a crime prediction model by utilizing environmental factors based on an ANN model and demonstrated the accuracy of an ANN using environmental factors. These previous studies showed that an ANN is an appropriate prediction model for crimes.

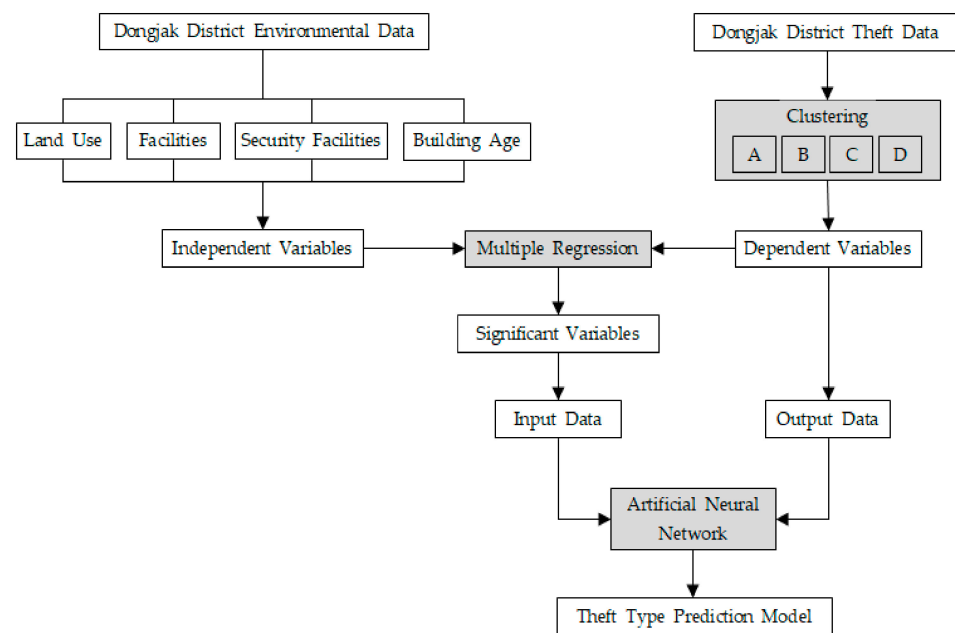
## 3. Research Methods and Research Design

### 3.1. Research Methods

In this study, rather than predicting crimes simply using environmental factors, specific crimes are categorized and the relationships between the derived categorizations and environmental factors are subdivided, thereby diversifying prediction methods using ANN (Figure 1). Figure 2 depicts the flow of this study. First, after pre-processing the theft data in Dongjak District, clusters were derived using k-modes clustering. After performing multiple linear regression analysis using the derived clusters and environmental factor data of the district, the environmental factors that affected each cluster were derived. An ANN model was constructed using the influential factors in each cluster as input data, and a model that predicted clusters was implemented. Finally, this study compared the performances of the model using all environmental factors as input data to those of the model using environmental factors with significance to each cluster as input data. Each regression analysis and prediction model construction was performed for each cluster to derive accurate factors and construct prediction models.



**Figure 1.** Differences in the development process of the ANN crime prediction model; (a) The development process of the general crime prediction model; (b) The development process of the crime prediction model proposed in this study.



**Figure 2.** Research process.

### 3.2. Research Target Area

One area in Seoul, South Korea, was selected as the target site for analysis. Seoul has a complex physical environment and a higher population density than other regions. The value of big data increases when the data features are more diverse, complex, and vast; more sophisticated research results can be expected by analyzing the physical environment of Seoul. This study selected Dongjak District as the analysis site among many areas in Seoul. The analysis period ranged from 2004 to 2015, where there are suitable data structures for the study. In 2015, Dongjak District was ranked 18th out of 25 districts in Seoul in terms of safety. Lee and Kim [29] analyzed the frequencies of five violent crimes recorded 792,260 times in Seoul from 2005 to 2011, calculating the crime risk rankings of 25 districts in Seoul using their self-developed “Crime Hot Spot Index”. The term “hot spot” in the study referred to an index that analyzes the frequencies of the five crimes per 1 km<sup>2</sup> within the area; if there are more than 105 assaults, 0.3 murders, 1.6 robberies, 69.7 instances of theft, or 0.6 sex crimes per year, the area is classified as a hotspot, and each crime sector

is allocated one point; the higher the score, the more dangerous the area. Autonomous districts with lower safety levels were mostly entertainment-concentrated areas. On the other hand, Dongjak District is a typical residential dense area with a commercial area (2.1%), ranked 24th among 25 autonomous districts in Seoul, but it showed higher crime rates than other districts. In particular, according to the trend of crime occurrences in the district for the three years from 2011 to 2014, theft indicated the highest increase among the five violent crimes.

### 3.3. Clustering

#### 3.3.1. Definition

Clustering is one type of data mining technique used in various fields such as media and marketing. The most representative clustering technique is k-means clustering, which was proposed by Macqueen in 1967. This technique is a clustering method in which data with similar features are clustered, finding the most appropriate central values and creating groupings with similar characteristics. However, this method is only applicable for continuous data with quantitative attribute values; as this study performed clustering using crime-related data, k-means clustering was not a suitable method. Most large-scale data consists of quantitative and categorical data. Therefore, it is essential to take a clustering method targeting categorical data to cluster large amounts of data [30]. Huang [31] suggested k-modes clustering as a clustering method targeting categorical data. This method is applicable for categorical data; that is, data consisting of nominal variables while maintaining the basic structure of k-means clustering.

K-modes clustering calculates similarities of categorical data consisting of a “distance function” that measures the distance between objects and a “cost function” that optimizes analysis. The set of object  $X$  consisting of nominal variables is  $D$ , while the number of  $X$  is  $n$ ; the equation is as follows:  $D = [X_1, X_2, \dots, X_n] (1 \leq i \leq n)$ . If the object  $X_1$  has  $m$  (number) nominal variables, the equation is followed as:  $X_i = [x_{i,1}, x_{i,2}, \dots, x_{i,m}] (1 \leq j \leq m)$ . The equation  $Z = [Z_1, Z_2, \dots, Z_k] (1 \leq j \leq k)$  indicates a set of  $k$  (number) cluster centers. If  $Z_1$ , the center of  $C_1$  cluster, has  $m$  (number) nominal variables, the equation is as follows:  $Z_l = [z_{l,1}, z_{l,2}, \dots, z_{l,m}] (1 \leq j \leq m)$

$$d(x_{i,j}, z_{l,j}) = \begin{cases} 0, & x_{i,j} = z_{l,j} \\ 1, & x_{i,j} \neq z_{l,j} \end{cases} \quad (1)$$

First, Equation (1) is an equation that calculates the distance between the  $C_l$  cluster's center,  $Z_l = [z_{l,1}, z_{l,2}, \dots, z_{l,m}]$  and the object  $X_i = [x_{i,1}, x_{i,2}, \dots, x_{i,n}]$  through the “distance function”. The function  $d(x_{i,j}, z_{l,j})$ , which measures the distance between  $X_i$  and  $Z_l$ , has the value of 1 when two values do not match and the value of 0 when matched; it calculates the distance between two objects through comparisons of the  $j$  number variables. Through the distance function equation, the distances between each object  $X$  and the center of the cluster  $Z$  are calculated while optimizing the model in the direction in which the result value of Equation (2) is minimized.

$$P(U, Z) = \sum_{l=1}^k \sum_{i=1}^n \sum_{j=1}^m u_{i,l} d(x_{i,j}, z_{l,j}) \quad (2)$$

$U = [u_{i,l}]$  is a matrix consisting of 0 and 1 in the form of  $n \times k$ , the equation  $u_{i,l} = 1$  represents the object,  $X_i$  is assigned to the nearest cluster  $C_l$ . The model is optimized and the cluster analysis result is derived by resetting the cluster centers and repeatedly analyzing them and calculating the minimum value of the cost function  $P(U, Z)$ .

#### 3.3.2. Analysis Environment Establishment

This study analyzed theft data that were obtained from district police stations; the data included 45,888 crime cases that occurred in Dongjak District from 2004 to 2015. The

data were classified into a total of 29 items; the items of occurrence time zone, occurrence address, location classification, victim gender, and crime methods were selected and used for the k-modes clustering analysis. After cleaning the data with inaccurate time zones or addresses, the data on 4464 theft cases were eventually used. Datasets were then prepared by organizing the 4464 cases into a total of four sections. The crime occurrence time was set by dividing it into morning (6–12), afternoon (12–18), evening (18–0), and dawn (0–6); in the case of items of location classification and crime methods, the classification was conducted based on theft data in Dongjak District. The location classification was divided into 16 categories: accommodation, church, construction site, educational institution, entertainment, hospital, financial institution, office, park, parking lot, public toilet, residence, store, street, vehicles, etc. Crime methods were classified into intrusion theft, hitting theft, vehicle-related theft, trick theft, etc.; “etc.” covered the group labeled as other theft in the theft data in Dongjak District or crime methods with too few data points to classify into individual items, such as theft when dozing, shoplifting, and locker room theft. The victim gender was set to “Y” for women and “N” for men. Therefore, 4464 cases of theft data in the district were classified into four items, and the dataset was constructed as shown in Table 1. The analysis environment was then built for k-modes clustering using Python 3.7.6.

**Table 1.** Dataset for K-modes clustering.

No.	Time of Crime	Location Classification	Method of Crime	Gender of the Victim
1	Evening	Residence	Intrusion	Y
2	Dawn	Store	Trick	N
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
4464	other	other	other	other

### 3.3.3. Setting the Optimal K-Value

K-modes clustering returns different results depending on the value of k, and it is then necessary to repeat the analysis for each k to find an optimized k whose cost function has the minimum value. In this study, an elbow method was used to attain k values, and each k value analysis was repeated a thousand times. K-modes clustering was executed in three steps as follows:

1. Define k random clusters.
2. Calculate the distance between the cluster center and each object and assign each object to the nearest cluster center. All objects are allocated to clusters, and the cluster center is moved in the direction where the distance between each object and the cluster center becomes closer.
3. If each moved cluster center has a different result value from the previous center value, you should return to step 2, run the analysis again, and repeat this step. If its result value is the same as the previous center value, the analysis is halted.

### 3.4. Multiple Linear Regression Analysis

#### 3.4.1. Definition

A multiple linear regression analysis was conducted to derive significant environmental factors for each cluster to confirm the individual influences of the factors for each derived cluster. The SPSS 18.0 statistical program was used for multiple linear regression analysis to derive significant environmental factors.

$$Y_k = a_n + b_1x_1 + b_2x_2 + \dots + b_nx_n \quad (3)$$

In the multiple linear regression analysis, the theft type (cluster) derived by clustering, as shown in Equation (3), was applied to the dependent variable ( $Y_k$ ); local environmental factors were input to the independent variable ( $X_n$ ), then the coefficient ( $b_n$ ) of each variable



and the constant term ( $a_n$ ) of the model were estimated. The terms  $k$  and  $n$  represent unknown variables. The hypothesis about the effects of the independent variables set in this study on clustering is as follows. The impact of environmental factors on each clustering would vary if there were differences in environmental factors related to each cluster and the related environmental factors have an impact on the results of clustering.

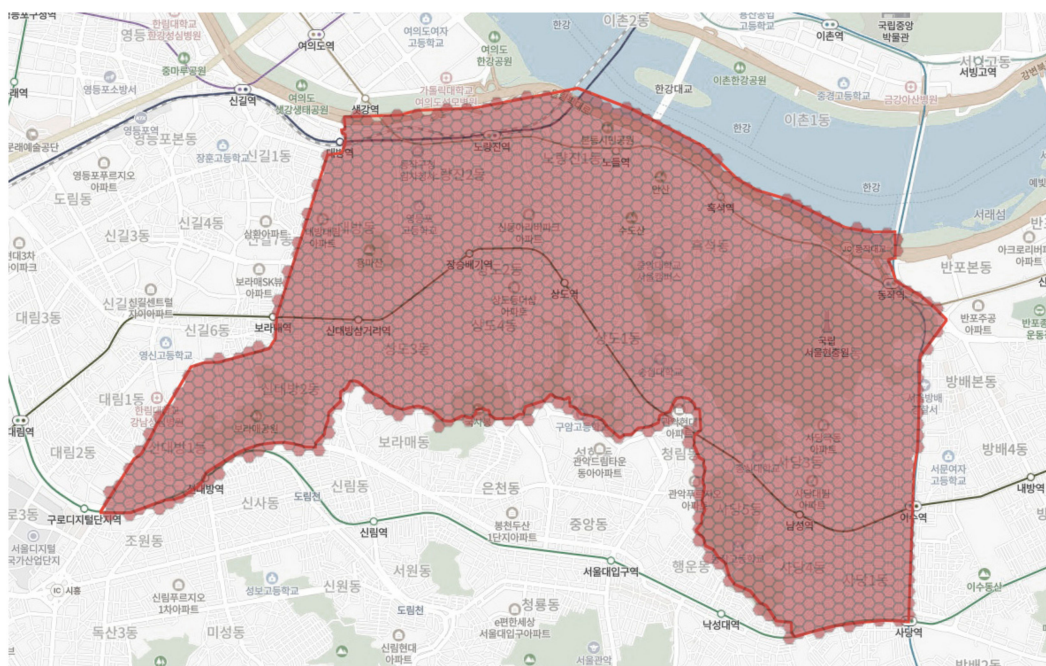
#### 3.4.2. Variable Setting

The data for land uses, facilities, and building ages were obtained from the National Spatial Data Infrastructure Portal. Land use data had the following classifications: Class 1 general residential area, Class 2 general residential area, Class 3 general residential area, Quasi-residential area, Neighboring commercial area, Distribution commercial area, General commercial area, Natural green area, Road, and Other areas. Class 1 general residential areas consist of low-rise housing areas and need to create a convenient residential environment; Class 2 general residential areas comprise middle-level housing and need to create a convenient residential environment; Class 3 general residential areas include mid- and high-rise buildings for a convenient residential setting; and Quasi-residential areas refer to places necessary for the well-being of residents and the preservation of a sound living environment. Neighboring commercial areas are for the supply of daily necessities and services in the neighborhood; distribution commercial areas are designed to improve distribution within and between cities; and general commercial areas are there for general commercial and business functions. Facility data consisted of a total of 28 major categories, and the following result of reclassification was based on the actual purpose of use, that is, 14 categories: detached houses, apartment houses, commercial facilities, educational and research facilities, public facilities, business facilities, medical facilities, lodging facilities, elderly facilities, factories, warehouses, cemetery-related facilities, hazardous material storage, treatment facilities, and parking lots. Regarding data on bus stops and subway stations, ranges were set to determine the existence of crimes within the scopes so as to determine their precise impacts. The range was set to 100m for bus stops and 200m for subway stations, based on previous studies [32,33]. Data on streetlights and CCTV were obtained from the Seoul Open Data Plaza.

#### 3.4.3. Grid Analysis Unit

Representative statistical spatial units consist of grid, census output area, administrative district, and so on, and, currently, census output area and administrative district are widely used. However, when dividing space into census output area or administrative districts, the resulting areas' shape and size are likely to be altered because they are irregular and boundaries are adjusted in line with temporal changes. On the other hand, since a grid has a constant shape and size, unlike those two units, statistical information can be objectively estimated and flexibility is applicable to changes in map scales [34]. The research related to crime analysis mainly employs administrative district units to analyze macroscopic crime scenes. Therefore, as it has limitations that cannot reflect micro-environments, this study attempts to analyze the relationship between thefts and environmental factors from a microscopic perspective.

This study proposed using hexagonal grids, similar to circles, for the following reasons. First, if the urban environment is divided into hexagons similar to circles, there is the least risk of data loss. Second, through a comparative analysis of land uses with the same area, there are improvements in data consistency, accuracy, and efficiency. Third, hexagonal grids show visually clearer outcomes than do lattice grids [35]. A radius of 25 and 50 m was used, respectively, for regression analysis to decide the size of the hexagon grid, and a radius of 50 m showed clearer results. Therefore, this study, using Geographic Information System information, split the target areas using a hexagonal grid with a radius of 50 m and then performed the analysis by inputting the environmental factor data into each cell (Figure 3).



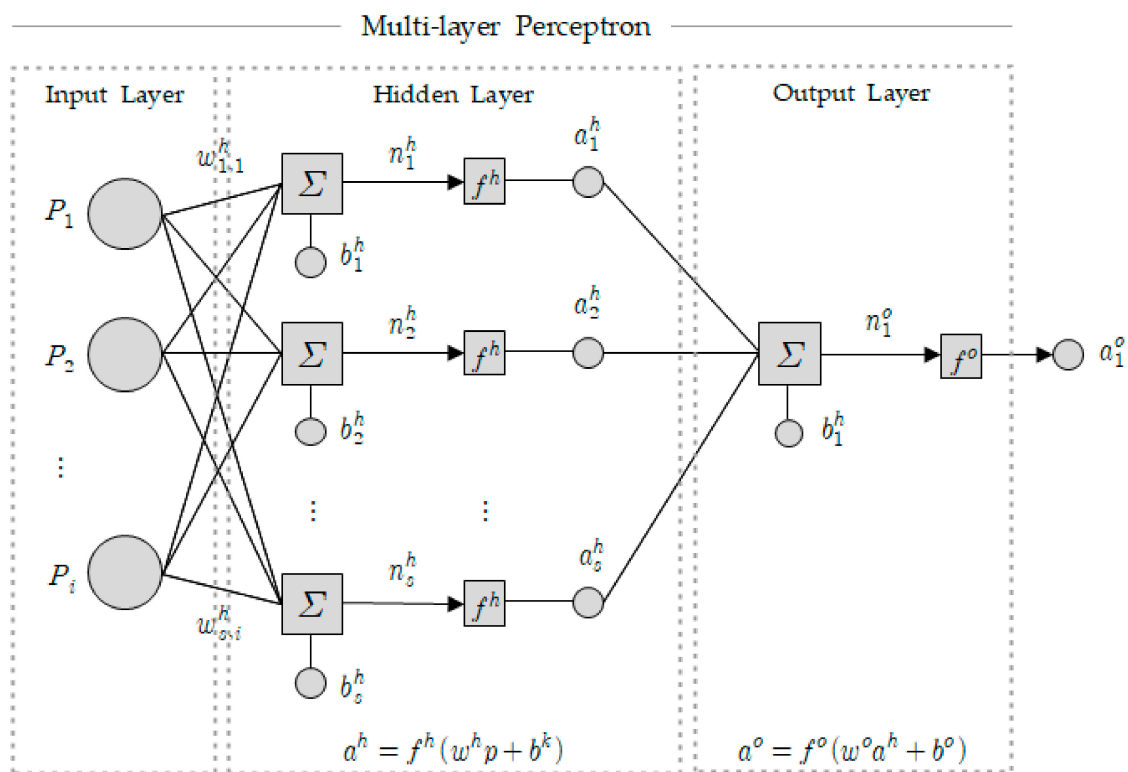
**Figure 3.** A map of Dongjak District, Seoul in a hexagon grid with a 50 m radius.

### 3.5. Artificial Neural Network

An Artificial Neural Network (ANN) is a machine learning algorithm that was first proposed by Warren McCulloch and Walter Pitts. They presented a simple computational model of how biological neurons interact for complex computations using propositional logic. Contrary to people's expectations, ANN could not solve the XOR problem, and, in the 1990s, ANN did not attract attention as the Support Vector Machine (SVM) and other machine learning algorithms with good performance emerged. However, with the development of the Deep Neural Network (DNN) model, which is an artificial neural network model composed of several hidden layers between input layers and output layers, it became possible to handle a vast amount of data that had not been used for analysis; it thus became possible to implement a prediction model that derives outputs through the weights of each node [36].

$$n_k^h = \sum_{j=1}^R w_{kj}^h p_j + b_k^h, k = 1 \text{ to } S \quad (4)$$

Figure 4 depicts the basic structure of ANN, and (4) shows its equation.  $R$  is the number of input variables,  $S$  is the number of hidden neurons,  $p$  is the input variable,  $b$  is the hidden layer, and  $w$  is the weight. The weight of each calculated element is used as an input of the activation function. The output is derived through the sum of these weighted values. In general, previous studies used the sigmoid function as an activation function. However, the function exhibits the gradient vanishing phenomenon in which existing information converges to zero as the neural network expands. In addition, the sigmoid function requires additional computing time being an exponential function. In an effort to address this challenge, the nonlinear ReLU function was proposed. This function does not lose information because it outputs the input value without any modification when it exceeds the threshold value, and the calculation speed is fast with a simple gradient value of 0 or 1. As a result, the performance of ANN increases remarkably with the ReLU function, and this approach was employed in numerous studies and is also used in the present work [37].



**Figure 4.** Artificial neural network (ANN) model structure.

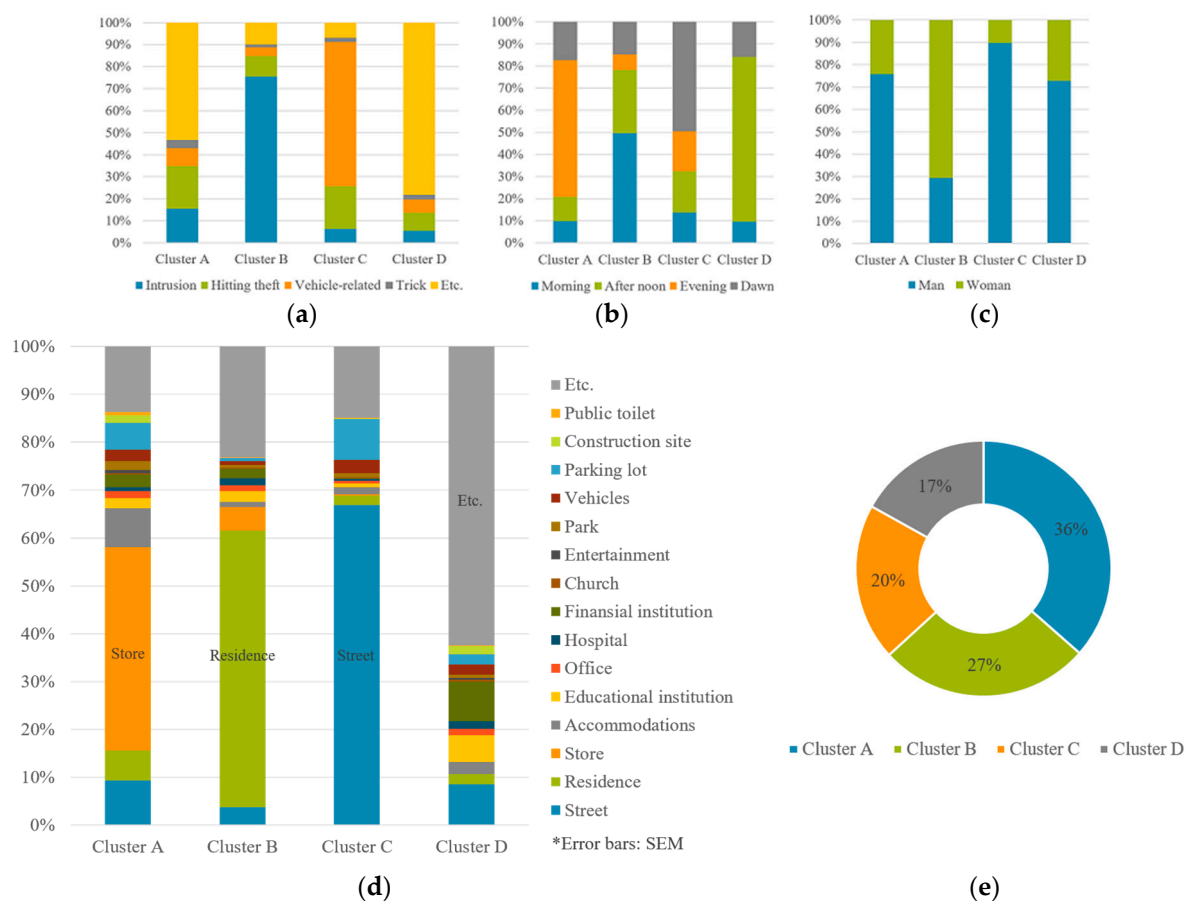
It is crucial to adjust the configuration of hidden nodes and layers to create an optimal ANN prediction model, but there are no clear criteria or methodology for this process. Therefore, it is necessary to find the optimal model with the lowest MSE value through as many attempts as possible. The minimum number of hidden nodes and layers should be more than the number of input variables, and the maximum number should not exceed  $2n+1$  for model training [38];  $n$  refers to the number of input variables. In this study, models with a minimum of one to a maximum of seven layers were examined. Along with the number of layers, the minimum, median, and maximum numbers of hidden nodes were to be processed in three to five cases, examining the optimal model with the smallest MSE value. In addition, to verify the model performance, this study attempts to compare each performance by configuring the environmental factors, which are input data, in different ways. This study seeks to compare the performances of the model using all environmental factors as input data and of the model using only environmental factors that influence each cluster as input data through multiple linear regression analysis.

## 4. Results and Discussion

### 4.1. Crime Categorization

As a result of k-modes clustering, it was found that there are typically four types of theft in Dongjak District. The figures for each category are visualized in Figure 5 to define theft types. Cluster A represents other theft (53%) in commercial facilities (43%) mainly targeting men (76%) in the evening (62%). Crime methods include intrusion theft (16%), hitting theft (19%), vehicle-related theft (8%), trick theft (4%), and other theft (53%); the ratios of other theft and hitting theft were recorded at higher rates than those in other clusters; crimes in the evening (62%) occurred three times more frequently than at other times. This can be interpreted to mean that there were thefts during congestion because there was a large floating population during the evening hours with a higher use of commercial facilities. Cluster B indicates home (58%) intrusion theft (76%) mainly targeting women (71%) during the morning (50%). It should be noted that crimes occurred most frequently during the morning and that intrusion theft (76%) was more than eight times

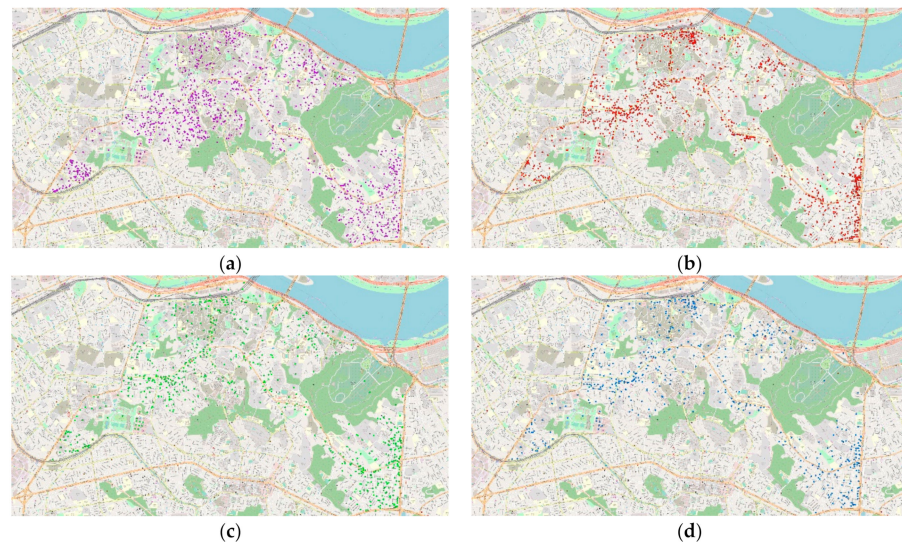
higher than other crime methods. Considering these results, it can be interpreted that, due to the three universities in Dongjak District along with many academies and reading rooms, many theft offenders mainly targeted empty houses after people went to academies, reading rooms, or universities in the morning. In the case of the gender of victims, there were twice as many female victims as male victims. Hwang [39] found that intruders often selected female victims on purpose. Among the reasons for this selection, “ease of control” was the most common reason, indicating that they tended to think that women were easy targets to control even if they encountered victims during crimes. These results imply that the gender of residents also contributed to the crime target selections of intruders. Cluster C represents vehicle-related theft (65%) and street theft (67%) mainly targeting men (90%) at dawn (49%). Among vehicle-related thefts, motorcycle theft occurred the most frequently (24%); considering the result that the ratios of “dawn” and “street theft” were high, there were more often crimes targeting motorcycles parked on the street at dawn, where it was difficult to conduct natural surveillance due to few passersby. There were overwhelmingly more male victims (90%) than female victims in such crimes because males hold more motorcycle licenses than females. According to the 2015 statistics of driver’s license holders in Seoul provided by the Seoul Metropolitan Police Agency, 90,018 males and 1353 females held a license for small vehicles. Cluster D shows other theft (78%) mostly targeting men (73%) in the afternoon (74%).



**Figure 5.** Values by clustering category item; (a) Method of crime; (b) Time of crime; (c) Gender of the victim; (d) Location classification; (e) Crime rates per clustering.

Just as the figures for each categorized item were different for each cluster, it was confirmed that the spatial distribution of clusters (Figure 6) also had different patterns. Therefore, through regression analysis, the relationship between clusters and regional

environmental factors was able to be estimated in more detail, deriving significant environmental factors for each cluster.



**Figure 6.** Spatial distribution of clusters; (a) Distribution of cluster A; (b) Distribution of cluster B; (c) Distribution of cluster C; (d) Distribution of cluster D.

#### 4.2. Derivation of Significant Environmental Factors

Table 2 shows the multiple linear regression analysis that was conducted to confirm the effects of environmental factor variables on the four theft types derived by k-modes clustering. The total number of cases was 2965, and the summary of the multiple linear regression analysis model is shown in Table 3. All four regression models are statistically fit because the Durbin–Watson value was close to 2 and not close to 0 or 4, so there is no correlation between the residuals. The explanatory power for each model was 27.3% (A), 22.3% (B), 25.3% (C), and 16.2% (D). Cohen’s  $f^2$ , which is generally used in statistical analysis in the social science field, is a common measure of calculating various types of effect size when using F-test for ANOVA and multiple regression analysis. Cohen [40] suggested that  $R^2 = 0.02$  be considered a small effect size, 0.13 represents a medium effect size, and 0.26 a large effect size. Thus, the multiple regression analysis models in this study all have values between 0.16 and 0.28, which shows the models are adequately validated.

The variables consisted of 10 land use variables, 17 facility variables, 2 security facility variables, and building age variables. In the case of nominal variables, regression analysis was performed through dummy coding. Before the analysis, a multicollinearity test was conducted to confirm the multicollinearity that occurs when the correlation between independent variables is high. As a result, as the VIF values of all variables were less than 10, there was no collinearity between variables.

As a result of multiple linear regression analysis, the environmental factors affecting each cluster are shown in Table 4. The detailed analysis results can be found in the Table A1. The analysis results showed that the effects of environmental factors on each cluster varied according to the hypothesis established earlier. If the significance level of the  $p$ -value is less than 0.10, the corresponding confidence level is 90%, if less than 0.05, it is 95%, and if less than 0.01, it is 99%. Regarding continuous variables, when they are increased by 1, they are assumed to affect the cluster creation by the value of the non-standardization coefficient (B); the dummy-coded variables are analyzed in comparison to the reference group.

**Table 2.** Data construction, descriptive statistics, and Variance Inflation Factor (VIF) analysis results.

Variable	Category	Mean	Standard Deviation	Common Difference	VIF	
Land use	Class 1 general residential area	Numeric	1673.914	2875.871	0.686	1.457
	Class 2 general residential area	Numeric	2718.752	3153.959	0.542	1.846
	Class 3 general residential area	Numeric	1540.355	2559.198	0.845	1.184
	Quasi-residential area	Numeric	131.554	758.575	0.890	1.123
	Neighboring commercial area	Numeric	16.044	219.330	0.817	1.224
	Distribution commercial area	Numeric	26.354	395.758	0.965	1.036
	General commercial area	Numeric	134.135	765.147	0.797	1.255
	Natural green area	Numeric	1976.576	3298.413	0.420	2.378
	Road	Numeric	1190.231	1294.794	0.448	2.231
Other areas	Numeric	226.810	1276.379	0.781	1.280	
Facility	Detached houses	Numeric	6.840	10.814	0.599	1.669
	Apartment houses	Numeric	2.380	3.715	0.671	1.491
	Commercial facilities	Numeric	0.740	1.609	0.629	1.591
	Educational and research facilities	Numeric	0.150	0.495	0.886	1.128
	Public facilities	Numeric	0.020	0.128	0.958	1.043
	Business facilities	Numeric	0.270	0.781	0.704	1.420
	Medical facilities	Numeric	0.040	0.230	0.893	1.120
	Lodging facilities	Numeric	0.010	0.890	0.824	1.214
	Religious facilities	Numeric	0.070	0.314	0.972	1.029
	Elderly facilities	Numeric	0.060	0.245	0.944	1.059
	Factories	Numeric	0.000	0.048	0.857	1.167
	Warehouses	Numeric	0.010	0.104	0.981	1.019
	Cemetery-related facilities	Numeric	0.010	0.101	0.942	1.062
	Hazardous material storage and treatment facilities	Numeric	0.000	0.058	0.915	1.093
	Parking lots	Numeric	0.000	0.034	0.995	1.005
	Bus stops	1 within 100 m	0.060	0.490	0.729	1.372
		2 none within 100 m	-	-	-	-
Subway stations	1 within 200 m	0.130	0.338	0.801	1.248	
	2 none within 200 m	-	-	-	-	
security facility	Streetlight	1 O	1.380	2.441	0.491	2.038
		2 X	-	-	-	-
	CCTV	1 O	0.300	0.577	0.816	1.225
		2 X	-	-	-	-
Building age range	Numeric	19.403	14.300	0.507	1.973	

**Table 3.** Model explanatory power and suitability analysis results.

	R	R <sup>2</sup>	Revised R <sup>2</sup>	Durbin-Watson	F	Significant Probability
Cluster A	0.523	0.273	0.261	1.759	21.647	0.000
Cluster B	0.473	0.223	0.210	1.962	16.573	0.000
Cluster C	0.507	0.257	0.245	1.957	19.969	0.000
Cluster D	0.403	0.162	0.148	1.945	11.152	0.000

**Table 4.** List of significant variables for clustering.

Variable	Category	Cluster A	Cluster B	Cluster C	Cluster D
Land use	Class 1 general residential area		✓		
	Class 2 general residential area				
	Class 3 general residential area	✓	✓		
	Quasi-residential area			✓	
	Neighboring commercial area		✓		
	Distribution commercial area				
	General commercial area	✓	✓	✓	✓
	Natural green area				
	Road		✓		
Other areas					
Facility	Detached houses	✓	✓	✓	
	Apartment houses		✓		
	Commercial facilities	✓	✓	✓	✓
	Educational and research facilities				✓
	Public facilities				
	Business facilities	✓		✓	
	Medical facilities	✓	✓	✓	✓
	Lodging facilities			✓	
	Religious facilities				
	Elderly facilities		✓		
	Factories				
	Warehouses				
	Cemetery-related facilities				
	Hazardous material storage and treatment facilities				
	Parking lots			✓	
	Bus stops			✓	
	Subway stations	✓			
security facility	Streetlight				
	CCTV				
	Building age range	✓		✓	✓
	Total	8	12	8	5

In total, three environmental factors were the highest influences in each cluster. In Cluster A, commercial facilities (13.277) showed the highest influence, followed by general commercial area (5.630), and Class 3 general residential area (4.089); in Cluster B, detached houses (9.538) had the highest influence, followed by apartment houses (5.095), and commercial facilities (4.419); in Clusters C and D, commercial facilities (8.059, 7.042) were the highest, followed by general commercial area (7.880, 6.511), and medical facilities (5.514, 3.143). General commercial area and commercial and medical facilities showed influences in all clusters. General commercial area and commercial facilities showed great influences in all clusters, which indicates that even though Dongjak District had fewer commercial areas (2.1%) than residential areas, there were more crime events concentrated in commercial areas. Areas with concentrated commercial facilities have a floating population until late in the day, and, accordingly, a certain degree of anonymity is guaranteed, so crime is prone to occur. Furthermore, commercial areas in Dongjak District had many buildings where commercial and residential facilities were mixed; Kim et al. [14] revealed that when commercial facilities and other uses were mixed, there was a higher crime frequency. Stucky and Ottensmann [13] found that there were more possibilities of major criminal offenses in medical facilities because that is where strangers gather, which supports the following outcome: These facilities had the highest impact in Cluster C (5.514).

#### 4.3. Artificial Neural Network Model

This study constructed an ANN model for four clusters and set the environmental factors with impacts in each cluster as input variables for analysis. Regarding nominal variables, the ANN model was constructed by including the reference group in the input variable after dummy coding. As for the input variables of each cluster model based on

significant environmental factors, Cluster A had 9 input variables, Cluster B had 12, Cluster C had 8, and Cluster D had 5. In total, 1763 data points were used for analysis: 1234 (70%) data for training, 264 (15%) data for validation, and 264 (15%) data for testing.

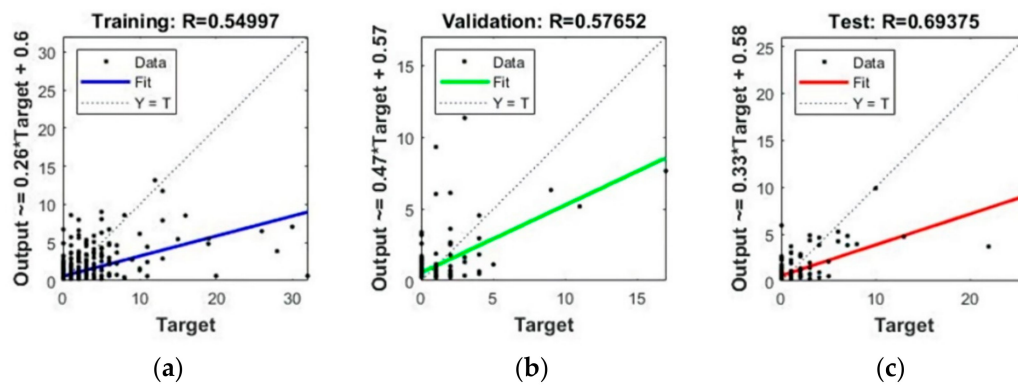
Table 5 indicates the case in which the model showed the best performance among the node compositions in line with the layer number of the prediction model for each cluster. Min and Max mean the minimum and maximum number of the sum of hidden layers and hidden nodes, and layer and neuron indicate the number of hidden layers and hidden nodes, respectively. In a case where there were more layers than were suggested in the table, this is indicative of a gradient vanishing problem, which means it cannot be used as a prediction model. The gradient vanishing problem refers to a phenomenon in which the gradient vanishes because the error is significantly diminished after passing through hidden layers many times, which leads to the inability to train the model.

**Table 5.** Performance of a model to predict theft types by hidden layers.

	Min	Max	Layer	Neuron	R Value of Training	R Value of Validation	R Value of Test	MSE	Terminated Epoch
Cluster A	9	19	1	10	0.670	0.445	0.281	4.0049	3th
				15	0.691	0.372	0.355	7.3922	5th
				19	0.415	0.035	0.106	5.4744	2th
			2	5	0.797	0.574	0.289	7.5817	13th
				7	0.549	0.576	0.693	1.8745	6th
				9	0.841	0.380	0.508	8.5441	12th
			3	4	0.623	0.678	0.384	4.8175	23th
				5	0.542	0.278	0.585	4.2803	3th
				6	0.492	0.445	0.548	2.8829	3th
			4	3	0.543	0.547	0.397	4.2073	24th
				4	0.358	0.131	0.306	6.0464	6th
				4	0.358	0.131	0.306	6.0464	6th
Cluster B	12	25	1	12	0.515	0.536	0.353	0.9840	5th
				17	0.452	0.350	0.523	1.4372	2th
				23	0.407	0.375	0.448	1.2945	5th
			2	8	0.598	0.359	0.503	1.4261	10th
				9	0.523	0.499	0.334	0.9553	9th
				12	0.403	0.429	0.387	1.3008	7th
			3	4	0.491	0.417	0.447	1.4915	7th
				6	0.457	0.472	0.484	1.4075	6th
				7	0.553	0.342	0.363	2.1231	8th
			4	3	0.480	0.449	0.411	1.7526	12th
				4	0.488	0.513	0.349	1.0793	4th
				6	0.427	0.487	0.360	0.9120	8th
			5	2	0.443	0.402	0.517	1.6537	15th
				3	0.490	0.412	0.422	2.0418	12th
				5	0.492	0.357	0.393	1.2083	13th
			6	2	0.468	0.393	0.520	1.4834	6th
				3	0.461	0.489	0.354	1.1160	5th
				4	0.473	0.416	0.474	1.7732	7th
Cluster C	8	17	1	9	0.343	0.139	0.181	0.7130	6th
				13	0.560	0.572	0.337	0.6818	6th
				17	0.599	0.506	0.429	3.4085	2th
			2	5	0.539	0.351	0.340	0.6894	7th
				7	0.593	0.508	0.404	0.6586	9th
				8	0.540	0.590	0.727	1.7023	5th
			3	3	0.477	0.514	0.431	0.7625	12th
				5	0.544	0.395	0.503	1.4863	5th
				6	0.536	0.483	0.527	0.7203	2th
Cluster D	5	11	1	6	0.322	0.099	0.165	2.7068	11th
				9	0.399	0.446	0.380	0.3960	3th
				11	0.357	0.341	0.325	0.6758	4th
			2	3	0.276	0.465	0.227	0.6714	10th
				4	0.351	0.368	0.448	1.3609	4th
				5	0.394	0.450	0.424	1.0653	12th
			3	2	0.377	0.422	0.451	0.8053	7th
				3	0.332	0.253	0.028	0.9770	4th
				4	0.395	0.398	0.347	1.5418	5th

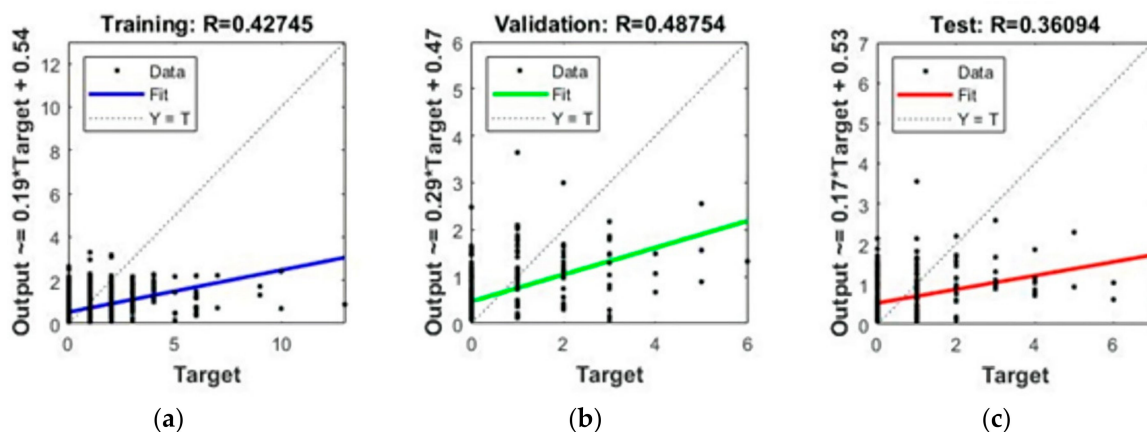
Figure 7 shows the R-values of training, validation, and test of the model with the best performance among Cluster A prediction models. Training had a value of 0.54997, validation had a value of 0.57652, and test had a value of 0.69375. The MSE value was  $10^6 \times 1.8745$ , and the DNN model showed the best performance with two layers and seven nodes in the model.





**Figure 7.** Cluster A model for crime occurrence prediction (layer: 2, node: 7); (a) Training data regression result; (b) Validation data regression; (c) Test data regression result.

Figure 8 shows the R-values of training, validation, and test of the model with the best performance among Cluster B prediction models. Training had a value of 0.42745, validation had a value of 0.48754, and test had a value of 0.36094. The MSE value was  $10^6 \times 0.9120$ , while the DNN model showed the best performance with four layers and six nodes in the model.



**Figure 8.** Cluster B model for crime occurrence prediction (layer: 4, node: 6); (a) Training data regression result; (b) Validation data regression; (c) Test data regression result.

Figure 9 shows the R-values of training, validation, and test of the model with the best performance among Cluster C prediction models. Training had a value of 0.59363, validation had a value of 0.50849, and test had a value of 0.40477. The MSE value was  $10^6 \times 0.6586$ , while the DNN model showed the best performance with two layers and seven nodes in the model.

Figure 10 shows the R-values of training, validation, and test of the model with the best performance among Cluster D prediction models. Training had a value of 0.39989, validation had a value of 0.44619, and test had a value of 0.38010. The MSE value was  $10^6 \times 0.3960$ , while the ANN model showed the best performance with one layer and nine nodes in the model.

Table 6 shows the performance comparisons of the models by dividing the attributes of input data into two types. Type I is a model that uses all environmental factor data as input data, while Type S is a model that uses only environmental factors that affect each cluster as input data through multiple linear regression analysis. The numbers of layers and neurons were set as the same for all types to control influences other than the input data. As a result of comparing the two types of model, the Type S model showed a higher accuracy in which only environmental factors affecting each cluster were used as

input data through multiple linear regression analysis. Type I encountered an overfitting problem because there was a large difference in prediction rates between the training and test models. Based on the comparative analysis of the two types of model, when a prediction is performed with crime-related environmental factors as variables, multiple linear regression analysis should proceed to derive significant variables, which will be used as input data; this method appears to improve the model performance.

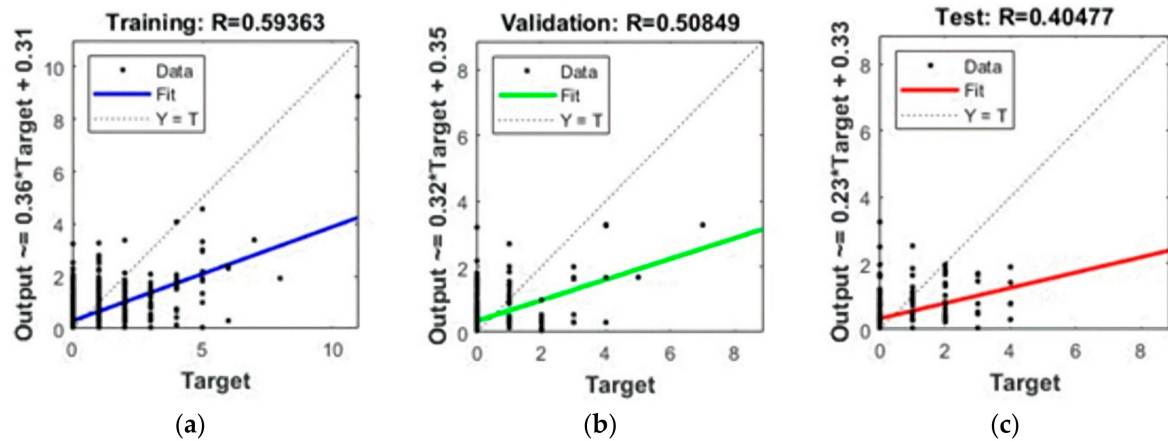


Figure 9. Cluster C model for crime occurrence prediction (layer: 2, node: 7); (a) Training data regression result; (b) Validation data regression; (c) Test data regression result.

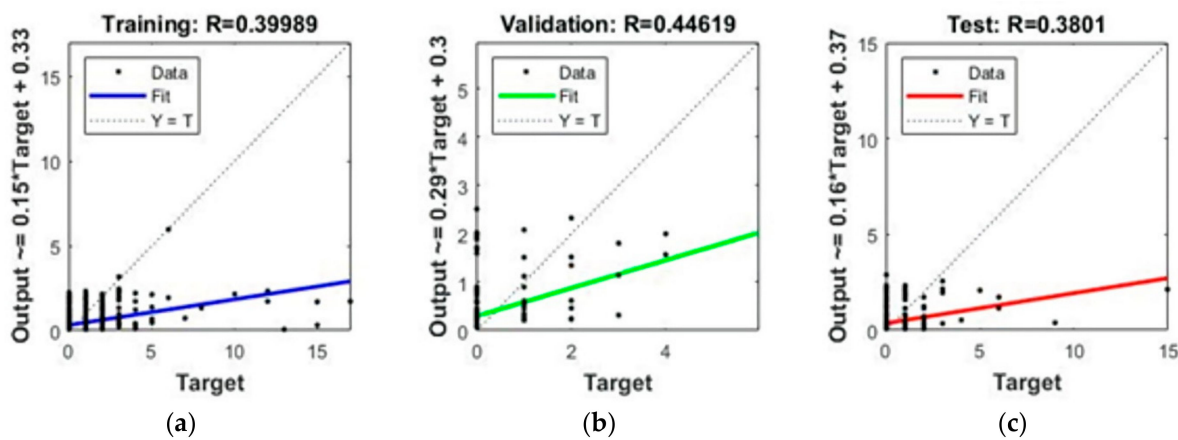


Figure 10. Cluster D model for crime occurrence prediction (layer: 1, node: 9); (a) Training data regression result; (b) Validation data regression; (c) Test data regression result.

Table 6. Performances of prediction models according to the input data type.

Cluster	Input Type	Layer	Neuron	R Value of Training	R Value of Validation	R Value of Test	MSE	Terminated Epoch
A	I	2	7	0.812	0.628	0.509	3.2180	14th
	S	2	7	0.549	0.576	0.693	1.8745	6th
B	I	4	6	0.535	0.456	0.275	1.0701	6th
	S	4	6	0.427	0.487	0.360	0.9120	8th
C	I	2	7	0.477	0.514	0.431	0.7626	12th
	S	2	7	0.593	0.508	0.404	0.6586	6th
D	I	1	9	0.178	0.346	0.288	0.5590	5th
	S	1	9	0.399	0.446	0.380	0.3960	3th

## 5. Conclusions

This study employed the data of theft in Dongjak District from 2004 to 2015 and then categorized that crime; based on the local environmental factors that influenced the crime, an artificial neural network model was proposed to predict theft types that are most likely to occur in random areas. First, using k-modes clustering, crime theft in Dongjak District was categorized, and a total of four types were derived. In addition, it was confirmed that each categorization had different corresponding environmental factors by using multiple linear regression analysis; it was identified that the environmental factors had different impacts and importance depending on theft types. The same environmental factors showed different levels of impact based on theft types; these results can be utilized to take the most appropriate tactical measures for each crime type when preventing theft. The prediction model developed in the study's crime prevention stage would allow us to predict the types of theft crime that are most likely to occur in a specific area and identify the environmental factors that affect the occurrence of the crime type. Thus, an effective crime prevention system could be established as the model identifies an environmental factor that needs to be improved primarily for each type of crime. Areas with a high probability of crime occurrence in Cluster A would need preventive measures focusing on commercial sites that become active during the evening hours, while areas with a high chance of crime in Cluster B should build preventive measures aimed at mostly protecting vacant homes in housing complexes against intrusions in the morning. Areas with a high probability of crime in Cluster C should prepare measures to prevent vehicle-related theft that occurs on the streets around commercial districts at dawn, while areas with a high likelihood of crime in Cluster D would need preventative measures, particularly for commercial areas that become active in the afternoon.

This study is meaningful in that it can contribute to creating areas that are safe from crime by presenting, such as an effective countermeasure method for crime prevention; in terms of environmental criminology, it empirically analyzed the influence of environmental factors on theft occurrence. The application of spatial statistics techniques contributes to helping policymakers more systematically understand the entire spatial distribution patterns of thefts at the local level. As such, the significant correlation between thefts and local environmental factors shows the necessity for an environmental criminology approach to theft prevention.

Regarding the theft type prediction model, the model fitness was evaluated through a comparison of two types of model. The Type I model used the entire environmental data as input data, while Type S used only environmental factors that affected each cluster as input data by performing multiple linear regression analysis. As a result of comparing the fitness of the models, it was confirmed that the S type model, which used only environmental factors that affected each cluster as input data by performing multiple linear regression analysis, showed a higher fitness. In general, it is expected that the more variables, the higher the model prediction rates. Although it is expected that more variables will result in higher correlations between variables, there is a high probability of multicollinearity problems; as the model becomes complicated, it is highly likely to have overfitting problems. On that basis, when a crime prediction model is implemented using environmental factors in the future, if the factors are applied to the model after verifying the influence of individual environmental factors through multiple linear regression analysis, the prediction model with higher performance can be implemented. In the case of the prediction model composition in this study, since the influential factors can be identified, it is possible to provide useful information, not only for simple crime type prediction but also for the preparation of plans to reduce crime. There were also categorizations of specific crimes rather than predictions of crime by simply using environmental factors, as well as a subdivision of the relationship between derived types and environmental factors; diversity was eventually given to prediction methods through an artificial neural network.

Since crime is a social phenomenon in which many factors such as physical, social, and economic elements function in combination, this study has limitations as it analyzed crime while focusing solely on local environmental factors. This study only considered environmental factor variables because it is a basic study to predict theft types by confirming

the relationship between this type and environmental factors and by using the factors as variables. It is meaningful, as it has determined the possibility that future studies could solidify, enabling them to derive clearer and more empirical results by considering diverse crime-related factors as variables. This study can be used as a basic study when conducting more sophisticated crime prediction studies by integrating all of the physical, social, and economic factors related to crime in the future.

**Author Contributions:** Conceptualization, Eunseo Kwon, and Sungwon Jung; Data curation Eunseo Kwon; Formal analysis, Eunseo Kwon; Methodology, Eunseo Kwon and Sungwon Jung; Project administration, Eunseo Kwon; Software, Eunseo Kwon; Supervision, Sungwon Jung and Jaewook Lee; Visualization, Eunseo Kwon; Writing—Original draft preparation, Eunseo Kwon; Writing—Review and editing, Eunseo Kwon, Sungwon Jung, and Jaewook Lee All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2018R1A2B2005528).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

**Table A1.** Results of deriving elements that affect clustering occurrence.

	Cluster A					Cluster B				
	Unstandardized Coefficient		Standardized Coefficient	t	p	Unstandardized Coefficient		Standardized Coefficient	t	p
	B	Standard Error				B	Standard Error			
tem	0.118	0.180	−0.046	0.655	0.513	0.182	0.096		1.903	0.057
Class 1 general residential area	$-3.829 \times 10^{-5}$	0.000	−0.046	−1.844	0.065	$-3.222 \times 10^{-5}$	0.000	−0.075	−2.925	0.003
Class 2 general residential area	$-1.692 \times 10^{-5}$	0.000	−0.022	−0.794	0.427	$-2.077 \times 10^{-6}$	0.000	−0.005	−0.184	0.854
Class 3 general residential area	$8.600 \times 10^{-5}$	0.000	0.091	4.089	0.000	$2.626 \times 10^{-5}$	0.000	0.054	2.353	0.019
Quasi-residential area	$1.837 \times 10^{-5}$	0.000	0.006	0.266	0.790	$4.279 \times 10^{-5}$	0.000	0.026	1.167	0.243
Neighboring commercial area	0.000	0.000	0.039	1.730	0.084	0.000	0.000	0.053	2.280	0.023
Distribution commercial area	$-4.013 \times 10^{-5}$	0.000	−0.007	−0.315	0.753	$-3.534 \times 10^{-5}$	0.000	−0.011	−0.523	0.601
General commercial area	0.000	0.000	0.129	5.630	0.000	$8.624 \times 10^{-5}$	0.000	0.053	2.243	0.025
Natural green area	$-1.012 \times 10^{-7}$	0.000	0.000	−0.004	0.997	$-7.865 \times 10^{-6}$	0.000	−0.021	−0.641	0.522
Road	$-9.227 \times 10^{-5}$	0.000	−0.050	−1.617	0.106	$-6.987 \times 10^{-5}$	0.000	−0.073	−2.307	0.021
Other areas	$-8.039 \times 10^{-5}$	0.000	−0.043	−1.833	0.067	$-2.401 \times 10^{-5}$	0.000	−0.025	−1.032	0.302
Detached houses	−0.020	0.006	−0.088	−3.338	0.001	0.030	0.003	0.261	9.538	0.000
Apartment houses	0.003	0.016	0.004	0.177	0.860	0.044	0.009	0.132	5.095	0.000
Commercial facilities	0.515	0.039	0.343	13.277	0.000	0.091	0.021	0.118	4.419	0.000
Educational and research facilities	0.189	0.106	0.039	1.778	0.076	0.060	0.056	0.024	1.064	0.288
Public facilities	−0.589	0.396	−0.031	−1.486	0.138	−0.122	0.210	−0.013	−0.581	0.561
Business facilities	0.221	0.076	0.071	2.923	0.004	−0.023	0.040	−0.014	−0.567	0.571
Medical facilities	0.832	0.228	0.079	3.657	0.000	0.327	0.121	0.061	2.709	0.007
Lodging facilities	0.521	0.613	0.019	0.850	0.395	−0.132	0.325	−0.009	−0.406	0.685
Religious facilities	−0.020	0.160	−0.003	−0.126	0.900	0.109	0.085	0.028	1.281	0.200
Elderly facilities	−0.121	0.208	−0.012	−0.584	0.559	0.456	0.110	0.090	4.129	0.000
Factories	−1.571	1.122	−0.031	−1.401	0.161	0.243	0.595	0.009	0.408	0.683
Warehouses	−0.178	0.481	−0.008	−0.371	0.711	−0.327	0.255	−0.027	−1.280	0.201
Cemetery-related facilities	−0.220	0.506	−0.009	−0.435	0.664	−0.072	0.268	−0.006	−0.269	0.788
Hazardous material storage and treatment facilities	0.718	0.887	0.017	0.809	0.418	−0.119	0.471	−0.006	−0.252	0.801
Parking lots	0.054	1.471	0.001	0.037	0.971	−0.529	0.781	−0.014	−0.678	0.498
Bus stops	−0.006	0.029	−0.006	−0.218	0.827	0.018	0.015	0.036	1.199	0.231
Subway stations	−0.141	0.095	0.034	1.486	0.137	−0.020	0.050	−0.009	−0.397	0.691
Streetlight	0.030	0.118	0.006	0.254	0.800	0.139	0.063	0.055	2.211	0.027
CCTV	0.425	0.163	0.060	2.600	0.009	−0.033	0.087	−0.009	−0.378	0.705
Building age range	0.013	0.005	0.075	2.611	0.009	0.002	0.003	0.027	0.903	0.367

Table A1. Cont.

	Cluster C					Cluster D				
	Unstandardized Coefficient		Standardized Coefficient	t	P	Unstandardized Coefficient		Standardized Coefficient	t	P
	B	Standard Error				B	Standard Error			
tem	0.110	0.073		1.508	0.132	0.106	0.098		1.079	0.281
Class 1 general residential area	$-1.293 \times 10^{-5}$	0.000	-0.038	-1.535	0.125	$-1.721 \times 10^{-5}$	0.000	-0.041	-1.525	0.127
Class 2 general residential area	$4.308 \times 10^{-6}$	0.000	0.014	0.498	0.618	$7.694 \times 10^{-6}$	0.000	0.020	0.664	0.507
Class 3 general residential area	$2.347 \times 10^{-6}$	0.000	0.006	0.275	0.783	$1.975 \times 10^{-5}$	0.000	0.041	1.727	0.084
Quasi-residential area	$7.140 \times 10^{-5}$	0.000	0.056	2.547	0.011	$4.208 \times 10^{-5}$	0.000	0.026	1.120	0.263
Neighboring commercial area	0.000	0.000	0.027	1.174	0.240	0.000	0.000	0.040	1.654	0.098
Distribution commercial area	$4.488 \times 10^{-6}$	0.000	0.002	0.087	0.931	$-4.236 \times 10^{-5}$	0.000	-0.014	-0.612	0.540
General commercial area	0.000	0.000	0.183	7.880	0.000	0.000	0.000	0.161	6.511	0.000
Natural green area	$-7.889 \times 10^{-6}$	0.000	-0.027	-0.841	0.401	$-7.895 \times 10^{-6}$	0.000	-0.021	-0.628	0.530
Road	$8.728 \times 10^{-7}$	0.000	0.001	0.038	0.970	$-4.961 \times 10^{-5}$	0.000	-0.053	-1.599	0.110
Other areas	$-2.532 \times 10^{-5}$	0.000	-0.033	-1.424	0.155	$-3.110 \times 10^{-5}$	0.000	-0.033	-1.304	0.192
Detached houses	0.005	0.002	0.055	2.060	0.040	-0.005	0.003	-0.043	-1.497	0.135
Apartment houses	0.009	0.007	0.033	1.321	0.187	-0.008	0.009	-0.025	-0.927	0.354
Commercial facilities	0.127	0.016	0.211	8.059	0.000	0.148	0.021	0.196	7.042	0.000
Educational and research facilities	-0.017	0.043	-0.009	-0.395	0.693	0.179	0.058	0.073	3.112	0.002
Public facilities	-0.050	0.161	-0.007	-0.312	0.755	0.108	0.215	0.011	0.501	0.616
Business facilities	0.073	0.031	0.059	2.390	0.017	0.079	0.041	0.050	1.921	0.055
Medical facilities	0.509	0.092	0.121	5.514	0.000	0.389	0.124	0.073	3.143	0.002
Lodging facilities	0.836	0.248	0.077	3.365	0.001	-0.615	0.333	-0.045	-1.847	0.065
Religious facilities	0.060	0.065	0.019	0.926	0.355	0.018	0.087	0.005	0.207	0.836
Elderly facilities	-0.096	0.084	-0.024	-1.132	0.258	-0.089	0.113	-0.018	-0.788	0.431
Factories	0.510	0.455	0.025	1.121	0.262	0.565	0.610	0.022	0.926	0.354
Warehouses	-0.116	0.195	-0.012	-0.594	0.552	0.261	0.262	0.022	0.996	0.319
Cemetery-related facilities	-0.119	0.205	-0.012	-0.582	0.560	0.072	0.275	0.006	0.263	0.793
Hazardous material storage and treatment facilities	0.684	0.360	0.041	1.902	0.057	0.506	0.482	0.024	1.050	0.294
Parking lots	-0.477	0.597	-0.017	-0.800	0.424	-0.291	0.800	-0.008	-0.364	0.716
Bus stops	0.011	0.012	0.028	0.936	0.349	0.020	0.016	0.041	1.300	0.194
Subway stations	0.036	0.038	0.021	0.925	0.355	0.050	0.052	0.024	0.970	0.332
Streetlight	-0.009	0.048	-0.004	-0.182	0.856	-0.006	0.064	-0.002	-0.093	0.926
CCTV	0.063	0.066	0.022	0.946	0.344	0.048	0.089	0.013	0.539	0.590
Building age range	0.004	0.002	0.058	1.983	0.048	0.006	0.003	0.070	2.272	0.023

## References

- Chen, H.; Chung, W.; Xu, J.J.; Wang, G.; Qin, Y.; Chau, M. Crime data mining: A general framework and some examples. *Computer* **2004**, *37*, 50–56. [\[CrossRef\]](#)
- Liu, H.; Brown, D.E. Criminal incident prediction using a point-pattern-based density model. *Int. J. Forecast.* **2003**, *19*, 603–622. [\[CrossRef\]](#)
- Nakaya, T.; Yano, K. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions* **2010**, *14*, 223–239. [\[CrossRef\]](#)
- Chandra, B.; Gupta, M.; Gupta, M.P. A multivariate time series clustering approach for crime trends prediction. In Proceedings of the 2008 IEEE International Conference on Systems, Man and Cybernetics, Singapore, 12–15 October 2008; pp. 892–896.
- Newman, O. *Defensible Space: Architectural Design for Crime Prevention*; US Department of Justice Law Enforcement Assistance Administration National Criminal Justice Reference Service: Washington, DC, USA, 1973; p. 53.
- Brantingham, P.; Brantingham, P. Criminality of place. *Eur. J. Crim. Pol. Res.* **1995**, *3*, 5–26. [\[CrossRef\]](#)
- Jeffery, C.R. *Crime Prevention Through Environmental Design*; Sage Publications: Beverly Hills, CA, USA, 1971; Volume 91.
- Cullen, J.B.; Levitt, S.D. *Crime, Urban Flight, and the Consequences for Cities*; National Bureau of Economic Research: Cambridge, MA, USA, 1996.
- Eck, J.; Weisburd, D.L. Crime places in crime theory. *Crime Place Crime Prev. Stud.* **2015**, *4*, 1–33.
- Brantingham, P.J.; Brantingham, P.L. *Environmental Criminology*; Sage Publications: Beverly Hills, CA, USA, 1981; pp. 27–54.
- McCord, E.S.; Jerry, H.R. Intensity value analysis and the criminogenic effects of land use features on local crime patterns. *Crime Patterns Anal.* **2009**, *2*, 17–30.
- Bae, W.K.; Kim, H.J.; Kwon, G.O. A study on the relationship between land use patterns and crime rates-Focused on the Bundang Newtown in 2006. *J. Urban Des. Inst. Korea Urban Des.* **2009**, *10*, 5–20.

13. Stucky, T.D.; Ottensmann, J.R. Land use and violent crime. *Criminology* **2009**, *47*, 1223–1264. [[CrossRef](#)]
14. Kim, D.K.; Yoon, Y.J.; Ahn, K.H. A study on urban crime in relation to land use patterns. *J. Korea Plan. Assoc.* **2007**, *42*, 155–168.
15. Lee, K.J.; Kim, Y.J.; Hong, S.J. An empirical study on exploration of spatial association between crime and land use. *J. Korean Urban Manag. Assoc.* **2015**, *28*, 245–267.
16. Gouvis, R.C. *Schools as Generators of Crime: Routine Activities and the Sociology of Place*; American University: Washington, DC, USA, 2002.
17. Kautt, P.M.; Dennis, W.R. Schools as criminal hot spots primary, secondary, and beyond. *Crim. Justice Rev.* **2007**, *32*, 339–357. [[CrossRef](#)]
18. Engstad, P.A. Environmental Opportunities and the Ecology of Crime. In *Crime in Canadian Society*; Silverman, R.A., Teevan, J.J., Eds.; Butterworths: Toronto, ON, Canada, 1975.
19. Block, R.L.; Davis, S. The Environs of Rapid Transit Stations: A Focus for Street Crime or Just Another Risky Place? In *Preventing Mass Transit. Crime*; Clarke, R.V., Ed.; Criminal Justice Press: Monsey, NY, USA, 1996; pp. 237–257.
20. Block, R.L.; Block, C.R. The Bronx and Chicago: Street Robbery in the Environs of Rapid Transit Stations. In *Analyzing Crime Patterns: Frontiers in Practice*; Goldsmith, V., McGuire, P.G., Mollenkopf, J.H., Eds.; Sage Publications: Thousand Oaks, CA, USA, 2000; pp. 137–152.
21. Greenberg, S.W. *Fear and its Relationship to Crime, Neighborhood Deterioration, and Informal Social Control. The Social Ecology of Crime*; Springer: New York, NY, USA, 1986; pp. 47–62.
22. Welsh, B.C.; Farrington, D.P. Public area CCTV and crime prevention: An updated systematic review and meta-analysis. *Justice Quart.* **2009**, *26*, 716–745. [[CrossRef](#)]
23. Phillips, C. A review of CCTV evaluations: Crime reduction effects and attitudes towards its use. *Crime Prevent. Stud.* **1999**, *10*, 123–155.
24. Xu, Y.; Fu, C.; Kennedy, E.; Jiang, S.; Owusu-Agyemang, S. The impact of street lights on spatial-temporal patterns of crime in Detroit, Michigan. *Cities* **2018**, *79*, 45–52. [[CrossRef](#)]
25. Park, M.K. *Implementation of Crime Prediction Map Using Space Analysis of GIS: A Case Study of Seongbuk-gu*; Kyung Hee University: Seoul, Korea, 2003.
26. Kwak, M.S.; Kwon, J.J.; Sung, H.G. Impacts of urban physical environment on crime incidence by its type and time. *J. Korea Plan. Assoc.* **2017**, *52*, 225–236. [[CrossRef](#)]
27. Mahmud, N.; Zinnah, K.I.; Rahman, Y.A.; Ahmed, N. Crimecast: A crime prediction and strategy direction service. In Proceedings of the 2016 19th International Conference on Computer and Information Technology (ICIT), Dhaka, Bangladesh, 18–20 December 2016; pp. 414–418.
28. Kang, H.W.; Kang, H.B. Prediction of crime occurrence from multi-modal data using deep learning. *PLoS ONE* **2017**, *12*, e0176244. [[CrossRef](#)] [[PubMed](#)]
29. Lee, H.I.; Kim, K.M. Spatial pattern analysis of urban crime in Seoul Metropolitan Area. *Korean J. Public Saf. Crim. Just.* **2013**, *22*, 217–246.
30. Kim, B.H.; Kim, K.S. General research papers: Improvements of K-modes algorithm and ROCK algorithm. *Korean J. Appl. Stat.* **2002**, *15*, 381–393.
31. Huang, Z. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. *Res. Issues Data Min. Knowl. Discov.* **1997**, *3*, 281–297.
32. Loukaitou, S.A. Hot spots of bus stop crime: The importance of environmental attributes. *J. Am. Plan. Assoc.* **1999**, *65*, 395–411. [[CrossRef](#)]
33. Groff, E.R.; Lockwood, B. Criminogenic facilities and crime across street segments in Philadelphia: Uncovering evidence about the spatial extent of facility influence. *J. Res. Crime Delinq.* **2014**, *51*, 277–314. [[CrossRef](#)]
34. Kim, M.S.; Lee, J.Y. A data transformation method for visualizing the statistical information based on the grid. *J. Korea Spat. Inform. Soc.* **2015**, *23*, 31–40.
35. Birch, C.P.; Oom, S.P.; Beecham, J.A. Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecol. Model.* **2007**, *206*, 347–359. [[CrossRef](#)]
36. Salakhutdinov, R.; Mnih, A.; Hinton, G. Restricted Boltzmann machines for collaborative filtering. In *2007 Proceeding, Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, USA, 20 June 2007*; Association for Computing Machinery: New York, NY, USA, 2007; pp. 791–798.
37. Lee, S.; Jung, S.; Lee, J. Prediction model based on an artificial neural network for user-based building energy consumption in South Korea. *Energies* **2019**, *12*, 608. [[CrossRef](#)]
38. Huang, W.; Foo, S. Neural network modeling of salinity variation in Apalachicola River. *Water Res.* **2002**, *36*, 356–362. [[CrossRef](#)]
39. Hwang, J.T. A study on target selection of burglars, robbers and thieves. *Korean Inst. Criminol.* **2004**, *04-26*, 217–247.
40. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Earlbaum Associates: Hillsdale, NJ, USA, 1988.