sustainability

MDPI

Article

Theft Prediction Model Based on Spatial Clustering to Reflect Spatial Characteristics of Adjacent Lands

Dongyoung Kim, Sungwon Jung * and Yongwook Jeong

Department of Architecture, Sejong University, Seoul 05006, Korea; yosi91@naver.com (D.K.); yjeong@sejong.ac.kr (Y.J.)

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

Abstract: Previous studies have shown that when a crime occurs, the risk of crime in adjacent areas increases. To reflect this, previous grid-based crime prediction studies combined all the cells surrounding the event location to be predicted for use in model training. However, the actual land is continuous rather than a set of independent cells as in a geographic information system. Because the patterns that occur according to the detailed method of crime vary, it is necessary to reflect the spatial characteristics of the adjacent land in crime prediction. In this study, cells with similar spatial characteristics were classified using the Max-p region model (a spatial clustering technique), and the performance was compared to the existing method using random forest (a tree-based machine learning model). According to the results, the F1 score of the model using spatial clustering increased by approximately 2%. Accordingly, there are differences in the physical environmental factors influenced by the detailed method of crime. The findings reveal that crime involving the same offender is likely to occur around the area of the original crime, indicating that a repeated crime is likely in areas with similar spatial features to the area where the crime occurred.

Keywords: crime prediction; machine learning; spatial clustering; smart city; GIS

Citation: Kim, D.; Jung, S.; Jeong, Y. Theft Prediction Model Based on Spatial Clustering to Reflect Spatial Characteristics of Adjacent Lands. *Sustainability* **2021**, *13*, 7715. https://doi.org/10.3390/su13147715

Academic Editors: Pierfrancesco De Paola, Francesco Tajani, Marco Locurcio and Felicia Di Liddo

Received: 17 June 2021
Accepted: 8 July 2021
Published: 10 July 2021

1. Introduction

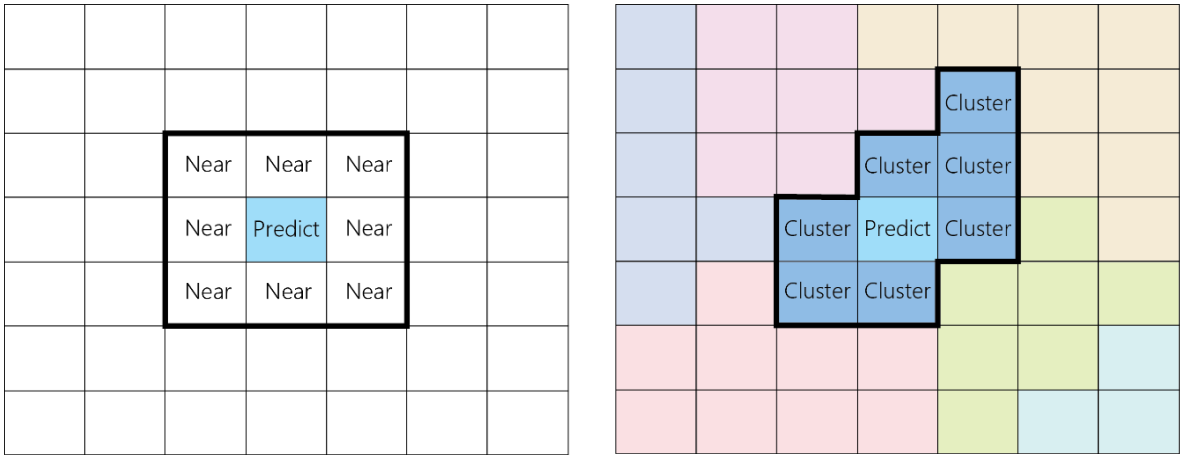
Today, with the enhancement of computer performance and data analysis techniques, it has become possible to process large amounts of data with ease. Pre-processed big data is used for prediction and analysis in various fields, such as stock price predictions [1,2] and financial analysis [3,4], using machine learning or artificial neural networks. In the field of crime prediction, various studies related to online crime detection [5] and the identification of crime hotspots [6] are being actively conducted.

Many government authorities across the world are already making efforts to prevent crime by applying crime prediction systems. In the case of PredPol, a crime prediction system for the Santa Cruz Police Department in the United States, the number of breaking and entering cases dropped by 27% from July 2010 to July 2011, when the system was in place, and fell by 25–29% in June and July 2013 compared to the same months of the previous year, which demonstrated the consistency in its effect. In Korea, GeoPros and CLUE are being used as part of the Smart City initiative. As a result of the GeoPros pilot run in 2013, robbery cases declined by 44.4%, while rape and theft decreased by 22.1% and 13.1%, respectively. Meanwhile, CLUE provides similar cases and investigation clues based on police investigation records, as well as crime prediction. Other crime prediction systems such as HunchLab and COMPStat, used in the Miami and New York Police Departments, are also contributing meaningfully to crime reduction, with reliable results.

In order to effectively carry out crime prevention activities through crime prediction, it is important to accurately set the prediction range, as well as making precise predictions, so that crime prevention resources such as CCTVs and police personnel can be properly allocated. Recently, researchers have actively studied machine learning-based crime prediction using grids as the units of analysis. In this regard, because grids have a uniform shape

and size compared to administrative districts or census output areas, statistical information can be objectively examined. Moreover, because grids can be flexibly applied to changes in map scale, microscopic analysis is also possible. Yu et al. [7] predicted residential burglaries by training an algorithm using the crime records of each cell, based on the crime's spatiotemporal concentration characteristics. To reflect the effects of crimes in an adjacent land and the physical environment during training, Lin et al. [8] predicted vehicle theft crimes by using 84 types of landmark data through Google API, along with crime information from adjacent cells for learning. Here, in the landmark data of Google API, the purposes and addresses of establishments such as schools, pubs, and restaurants are indicated, and these were used to reflect the geographical characteristics of the study site. As an extension of earlier research, the purpose of this study is to develop a crime prediction model that reflects the influence of surrounding areas and geographic characteristics on crime.

An actual land is continuous, rather than independently divided like a grid, in a geographic information system (GIS), and when a crime occurs, the risk of crime in the adjacent areas increases [9–11]. According to studies analyzing the relationship between spatial characteristics and crime, environmental factors such as patterns, establishments, and land use were found to be different, depending on the detailed modus operandi of crimes [12–15]. Even in the case of the same type of crime, the related factors were shown to be different depending on the detailed modus operandi [16]. Therefore, in micro-scale studies using a grid, there is concern regarding how to reflect the characteristics of the adjacent land. When training with crime information, this issue can be solved by combining all the cells adjacent to the point to be predicted and using it in the training. However, this solution requires a focus on a specific method of crime, and if the cell to be predicted and its adjacent cells vary greatly in spatial characteristics, the training may be negatively affected. Accordingly, to distinguish cells with similar spatial characteristics for use in training, this study proposes a crime prediction method that applies spatial clustering as shown in Figure 1. To perform spatial clustering, high weights are assigned between geographically adjacent cells, according to the distance between each instance in the vector space. Thus, cells adjacent to the cell to be analyzed and with similar attributes are included in the same cluster and, rather than training all the adjacent cells, only the cells with similar spatial characteristics can be used for training.



(**a**) Crime prediction combining adjacent cells (**b**) Crime prediction reflecting spatial similarity

Figure 1. Proposed crime prediction using spatial clustering.

The flow of this study is shown in Figure 2. Dongjak District, Seoul, the target analysis area, is divided into grids of 100 m × 100 m on GIS. After inserting data on the physical environment in each cell, such as crimes, facilities, and land use that occurred previously, the target sites are clustered according to spatial similarity, based on the physical

environment data. Cells in which no crime occurred during the analysis period are removed because they might negatively impact training. The remaining cells are then used for training. Crime data are imbalanced because there is less data on where a crime occurred than where it did not. Accordingly, resampling is used to solve the problems caused by the data imbalance. The preprocessed data are trained using a random forest, a tree-based machine learning algorithm, and the differences between the model with spatial clustering and the general model are compared.

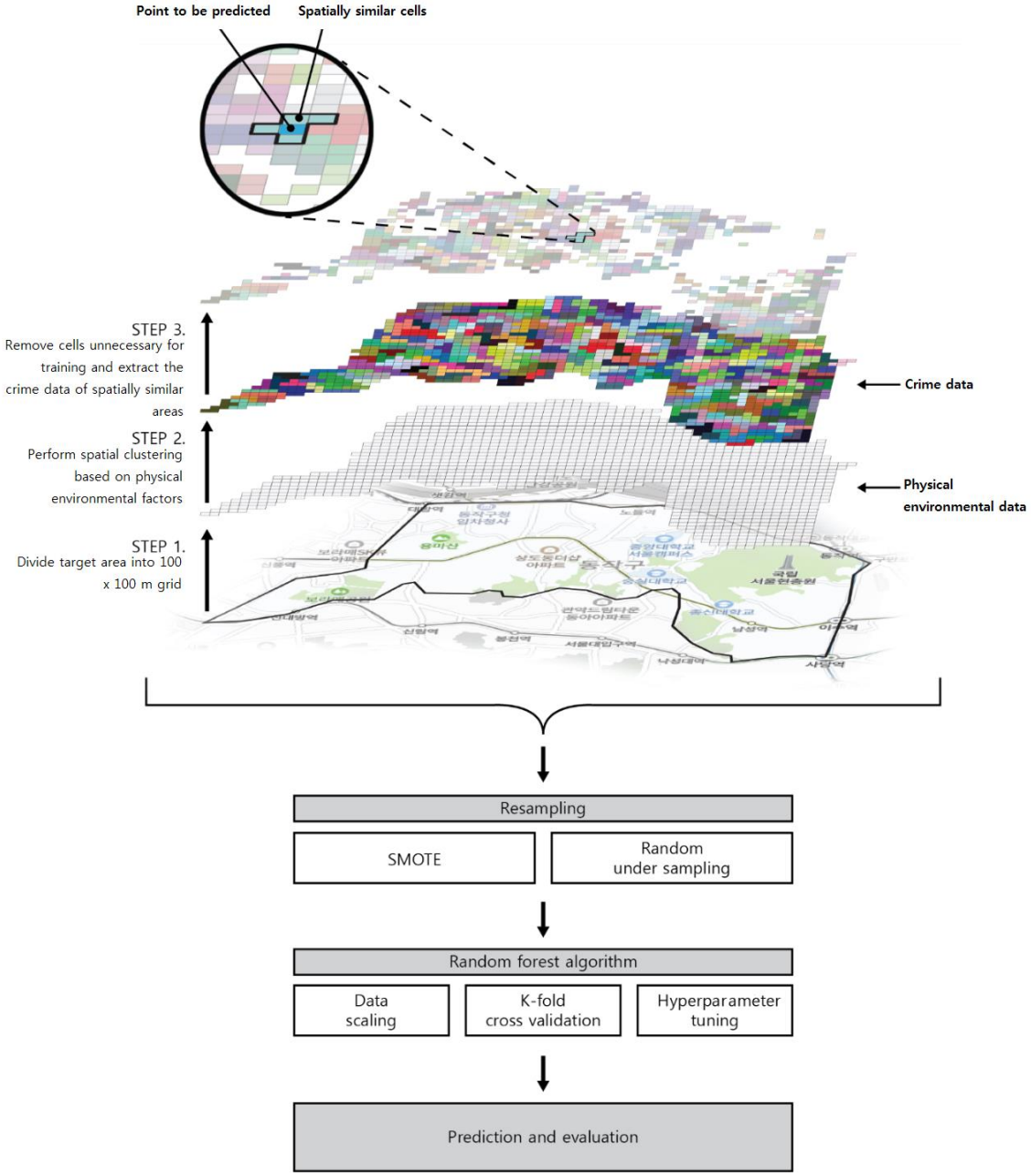


Figure 2. Research flow.

2. Theoretical Review

Efforts have been made to prevent crime by identifying and mitigating its causes. Environmental criminology seeks to explain the causes of crime using the surrounding environment. The main theories employed are routine activity theory (RAT) [17] and crime pattern theory (CPT) [18]. RAT states that a crime occurs when a motivated offender

(a suitable target) and the absence of a capable guardian simultaneously intersect in time and space. According to RAT, it is important to view individuals as motivated offenders and minimize their opportunities to commit the crime. The CPT states that people move in certain patterns because of their physical or social environment, such as their occupation. During their routine activities, motivated offenders identify the characteristics of these areas and suitable targets for crime, choose a suitable time and place, and then carry out the crime. Hence, crimes do not occur randomly; however, they are concentrated in certain locations, owing to specific factors, and are influenced by the surrounding environment and living patterns of individuals and their neighbors.

Wolfgang, Figlio, and Sellin [19], and Sherman, Gartin, and Buerger [20] found that approximately 50% of all crimes every year occur within 4% to 5% of the total street and explained that crime is spatially concentrated. Studies have also shown that the areas surrounding the place where a crime occurs are at risk of identical crimes, and the more recent the crime, the greater its influence [9–11]. Notably, Bernasco [10] reported that, among the crimes that occurred within 100 m of an earlier crime, 90% that occurred within seven days involved the same offender. Moreover, the same offender was involved in 64% of the crimes that occurred within 90 days, and 13% within nine years. This indicates that offenders familiar with the surrounding area are more likely to commit other crimes in neighboring locations.

Facilities and land use play a key role in the relationship between crime and the physical environment. These factors provide the purpose of people's movements and have a close relationship with individuals' living patterns. Brantingham and Brantingham [11] analyzed the correlation between commercial theft and facilities. In their results, blocks with supermarkets and department stores showed similar crime rates to blocks without these landmarks, whereas blocks with fast-food restaurants, traditional restaurants, and pubs had 2- to 2.5-times more commercial theft than blocks without these landmarks. Lee, Yoon, and Kim [13] analyzed the causes of crime according to crime type in specific cities in Korea. Visitor accommodation, restaurants, financial institutions, and homes in non-residential buildings were highly correlated with theft crimes. In the case of CCTV, it was found that the related factors were different depending on the type of crime, such as showing a significant correlation only with rape and violence. Studies analyzing the impact of land use on crime are also underway. As in the case of facilities, there were differences in the factors affected by crime and, in the case of commercial areas, it was discovered to be related to most crimes [14–16]. Stucky and Ottensmann [15] analyzed the relationship between land use and crimes such as violent crime, homicide, robbery, aggravated assault, and rape. The correlation between land use and crime type was shown to be different, showing a significance in crime, homicide, and aggravated assault. Kwon, Kwon, and Jung [16] examined the correlation between each crime type and land use by clustering the theft crimes into detailed types according to the victim's gender, the time of the occurrence, and the place of the occurrence. It was shown that the associated physical environment was different. As such, crimes do not occur randomly, but have factors influencing them; in environments where crimes can occur with ease, it is important to identify these related factors.

3. Data and Methodology

3.1. Research Area and Analysis Unit

Dongjak-gu, the research area, is one of the administrative districts of Seoul, the capital of South Korea. Its population density is 24,190/km^2, which is similar to that of Manhattan in the United States. Residential areas are high-density areas comprising 84% of the total population, and there are 8.5 cases of violent crime per 1000 people. Although this rate is ranked 17 out of the 25 administrative districts, it is rather high because most of the areas with high crime rates have a developed entertainment industry.

To effectively perform crime prevention activities using crime prediction, it is important to precisely set the analysis unit so that crime prevention resources can be allocated

to the appropriate locations. Accordingly, this study attempts to predict crime in the microscopic range through grid-level analysis. Compared to the administrative districts and census output areas, which are statistically spatial units used in existing statistical map services, grids have a uniform shape and size, allowing statistical information to be objectively examined. Moreover, the grids can be flexibly applied to changes in the map scale. This study used GIS to divide the target area into a grid of 100 × 100 m cells, then time- and space-related data were added to each cell to perform the analysis.

3.2. Crime Data

In the case of crime prediction, it is generally known that theft crimes are easier to predict than other types of crimes. Crimes such as murder and assault are highly influenced by ill feelings between the offender and the victim because the target is a specific individual. In contrast, since the target of theft is a specific building or object, it is influenced more by the surrounding environment and the behavioral characteristics of the criminal than by personal feelings [21]. The analysis of this study focuses on theft. With the cooperation of the police department with the relevant jurisdiction, data on incidents of theft in Dongjak-gu from 2013 to 2017 were used in the analysis. Figure 3 shows the monthly distribution of theft in Dongjak-gu. During this period, an average of 95 thefts occurred per month; the most occurred in March 2013 (199), and the fewest occurred in November 2016 (31). The theft data include the date, time, method, and exact location of the crime. Inaccurate data (such as cases with incorrect addresses and duplicate reports on the same date) were excluded. A total of 8023 theft cases were used for the analysis. Based on prior studies showing that the more recent the crime, the greater its influence on future crimes, in order to train the influence over time in a predictive model, this study calculated the average number of crimes that occurred in each cell over the periods of two weeks, one month, three months, six months, and one year, and used these values for the training. In the grid-level analysis, cells in which crime never occurred were mainly those areas (such as mountains or water) in which it was difficult for crime to occur. These data can easily lead the model to predict that no crime incidents occur in areas with no previous record of crime, which may negatively impact the predictions [7,8]. Therefore, before training, the cells in which no crime occurred from 2013 to 2016 were removed. The 2017 crime record was excluded, as it was used as the test set.

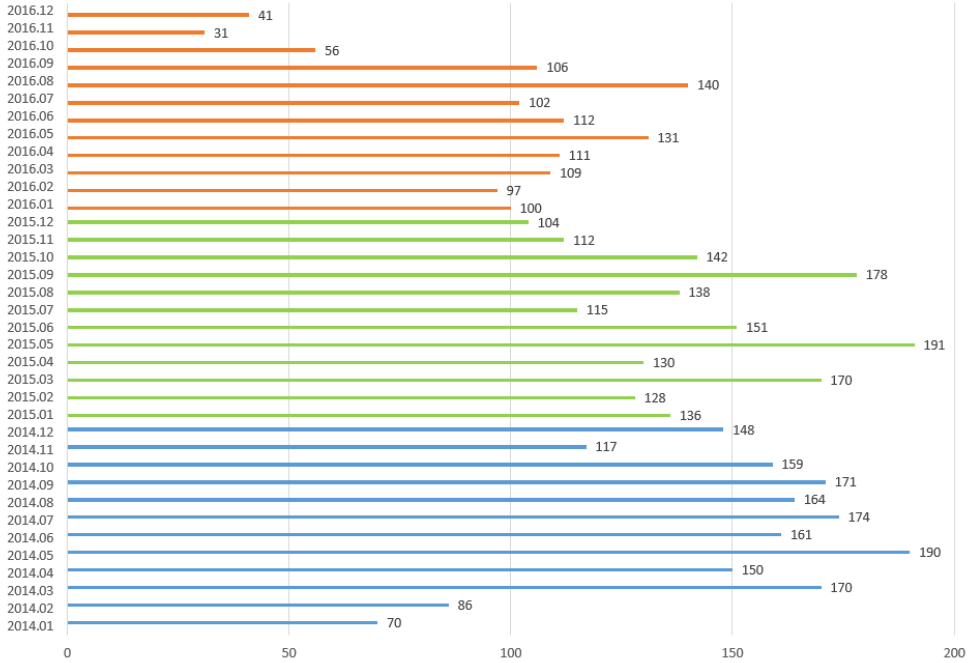


Figure 3. Distribution of theft, sorted by month, in Dongjak-gu.

3.3. Physical Environment Data

The data used in this study were provided by the National Spatial Data Infrastructure Portal (http://www.nsdi.go.kr, 15 January 2020) and the Open Data Plaza (https://data.seoul.go.kr/, 15 January 2020). The data on building usage comprise basic information on location, size, etc., and are categorized into 152 types according to building use. However, using data on all buildings for the training may degrade the model's performance while consuming extensive computing resources because the model would be trained on data with an insufficient correlation with crime. Consequently, considering the training, this study applied data on building use related to restaurants, pubs, accommodations, banks, and residences, which have been demonstrated by previous studies as being related to crime. Restaurants were categorized into general restaurants, where patrons stay for a long time to eat, and rest-area restaurants, which sell simple meals such as fast food. Residential buildings were categorized into single-family housing, multi-family housing, and apartments, depending on the type of residence. Additionally, CCTV and streetlights, which are factors influencing natural surveillance, and bus stops, known to induce crime because of crowding, were added to the facility variables for the training.

Regarding the information on land characteristics, data on land usage and officially assessed land prices (OALP) were used for the training. In South Korea, land use is divided into eight categories: general commercial areas, neighboring commercial areas, circulating commercial areas, first-class residential districts, second-class residential districts, third-class residential districts, semi-residential areas, and natural green belt zones. The facilities and allowable sizes that can be built according to each usage area have different legal regulations. To apply the land usage data for the training, the area occupied by the usage category in each cell of the grid was converted to a percentage. In addition to land use, the average OALP of each cell was calculated and used as a variable. This is used as an indicator to identify the geographic continuity in the spatial clustering analysis. Table 1 lists the variables used in the study. Finally, applying crime data from 2014 to that of 2016, with the training set and crime data from 2017 as the test set, the data were used in training, and the model's performance was evaluated.

Table 1. Feature selection.

Feature	Precision
Crime	Average number of crimes in each cell over the previous 1, 3, 6, 9, and 12 months
Adjacent crime	Average number of crimes within the same cluster over the previous 1, 3, 6, 9, and 12 months
Factors related to crime	CCTV, streetlight, bus stop
Facility-related variables	General restaurants, rest-area restaurants, pubs, accommodation, banks, multi-family housing, single-family housing, apartments
Land-related variables	General commercial area, neighboring commercial area, circulating commercial area, first-class residential district, second-class residential district, third-class residential district, semi-residential area, natural green belt zone, officially assessed land price (OALP)

3.4. Spatially Constrained Clustering Methods

Clustering is a data-mining technique that classifies the given data into multiple clusters, based on the similarity of their attributes. Because it is difficult for general clustering techniques to reflect the spatial continuity of data in a vector space such as GIS, researchers have been studying spatially constrained clustering methods [21,22] to solve this issue. One of them is the max-p regions model [23,24]; unlike the general clustering techniques that classify data into a limited number of clusters, this model aims to maximize the number of clusters that satisfy the minimum threshold of the constraint, while minimizing spatial heterogeneity in each cluster. This constraint is the minimum

value of the variables (population size, number of houses, etc.) included in each instance, or the minimum number of instances that must be included in each cluster. To cluster the cells that are spatially similar and adjacent in distance, this study sets the number of cells that can be included in each cluster as the constraint. As a feature of the max-p regions model, a specific cluster can be prevented from growing excessively larger than the other clusters, and the land can be uniformly clustered while maintaining spatial continuity. Thus, the model can be effectively used for microscale analysis.

The equation of the max-p regions model is as follows: first, $A = \{A_1, A_2, \cdots, A_n\}$, $(n = |A|)$ is defined as the set for the entire land area, and A is defined as the set divided into p regions, $P_p = \{R_1, R_2, \cdots, R_p\}$, $(1 \le p \le n)$. In this study, l_i is the attribute that must at least reach the minimum threshold in area A_i.

$$\begin{cases} |R_k| > 0 \; for \; k = 1, 2, \cdots, p \\ R_k \cap R_{k'} = \theta \; for \; k, k' = 1, 2, \cdots, p \wedge k \neq k' \\ \cup_{k=1}^{p} R_k = A \\ \sum_{A_i \in R_k} l_i \ge thershold > 0 \; for \; i = 1, 2, \cdots, n \; and \; k = 1, 2, \cdots, p \end{cases} \quad (1)$$

Here, all the divisible sets of A are defined as Π. Thereafter, the max-p algorithm can be defined as in Equation (2). $H(P_p)$ is the sum of the heterogeneity of space over all of $P_p \in \Pi$.

$$\begin{cases} P_p^* = max\left(\left|P_p^*\right| : P_p \in \Pi\right) \\ \nexists P_p \in \Pi : |P_p| = \left|P_p^*\right| \text{ AND } H(P_p) < H\left(P_p^*\right) \end{cases} \quad (2)$$

In this study, facility and land data were inserted into the grid-divided area and used as variables for the max-p regions model, through which cells with geographically similar characteristics were clustered. Based on this, in the machine learning step, crimes that occurred in the same cluster were used as a prediction variable to reflect the influence of crimes that occurred in the adjacent land during the training. Figure 4 shows an example of the max-p regions model, and Table 2 shows an example of average attributes for each cluster. To ensure that the cell to be predicted and cells that are physically far away do not belong to the same cluster, the number of cells n belonging to each cluster was set between 2 and 10.

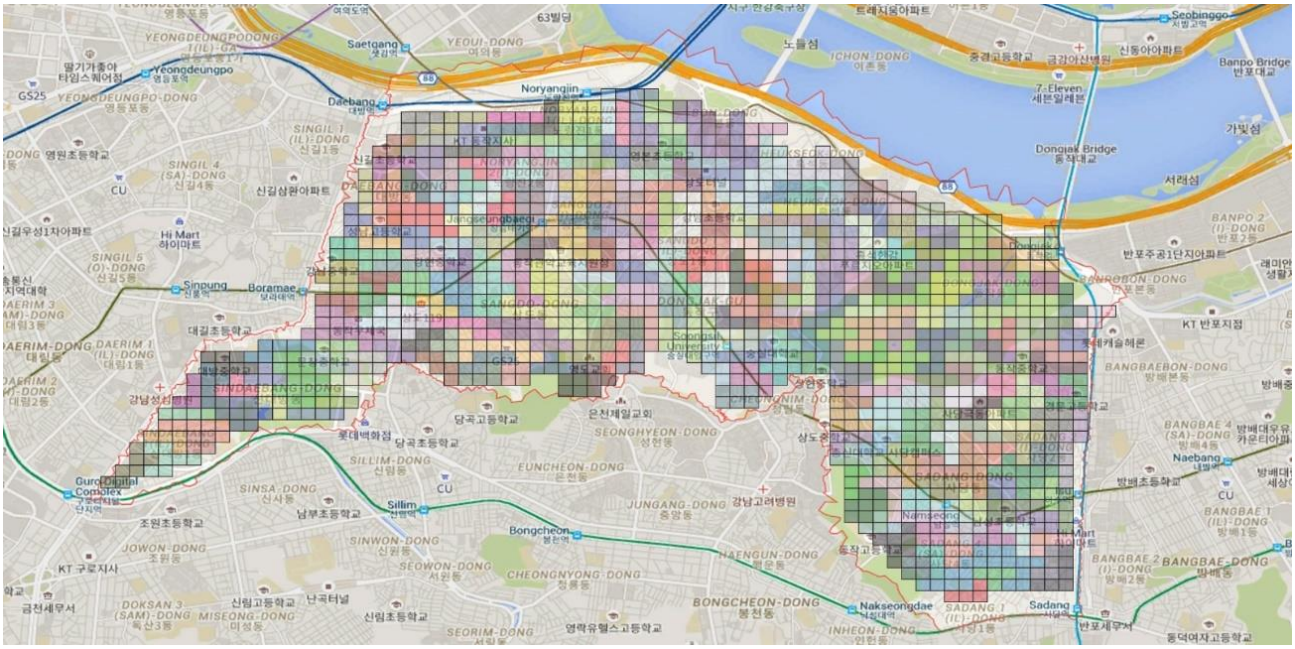


Figure 4. Example of a max-p regions model (n = 4).

Table 2. Example of average attributes for each cluster in the max-p regions model.

Cluster Number	General Restaurant	Alcohol	Multiple Housing	Single Housing	...	Type 2 Residential Area	Type 3 Residential Area	Natural Green Area
1	0	0	4.91	5.33		14.5	0.01	0
2	1.09	0.27	25.90	10.81		91.4	8.03	0
3	0	0	0	0		0	0	100
304	6.5	2	3.75	4.5	...	29.04	70.94	0
305	0	0	28.5	12.5		95.2	0.02	0
306	10.5	2.75	14.75	25.25		45.04	53.67	0

3.5. Resampling

Imbalanced data are those in which the distribution is overconcentrated in a specific class. In imbalanced data, the minority classes are recognized as noise in the training process, and the classification does not proceed correctly, which may adversely affect performance [25]. The theft data used in this study are also imbalanced; only approximately 10% of the total data correspond to the theft class. Accordingly, this study used random undersampling and the synthetic minority oversampling technique (SMOTE) to solve the problems due to data imbalance.

Random undersampling is a resampling technique that randomly deletes instances from the majority class to balance distribution with the minority class. When there are many training sets, it is possible to increase the learning speed and reduce the data capacity by decreasing the number of samples. However, because this technique involves deleting data, there is a risk of information loss. The SMOTE is an oversampling technique that interpolates data in the minority class to create new instances to balance the data. Whereas this results in a slower training speed than undersampling, there is no risk of data loss, and overfitting is less likely to occur than random oversampling, which randomly replicates the minority data.

3.6. Model Training and Evaluation

3.6.1. Model Training

The data preprocessed via the above procedure were trained using a random forest, a tree-based machine learning algorithm. Random forest, an ensemble technique widely used in general classification problems, creates multiple decision trees and combines the output of each decision tree. This study used the random forest technique to build a crime prediction model and then compared each model.

First, as the range of values for each variable differed, the values of the data were normalized using min-max scaling. The ratio between the training set and test set is generally set to between 7:3 and 8:2; nevertheless, this is flexible, depending on the amount of data and the research method. The purpose of crime prediction is to predict future crimes based on those crimes that occurred in the past. As such, the data from 2014 to 2016 were used as the training set, and those from 2017 were used as the test set. K-fold cross-validation was applied to each model in the training process to prevent the bias and overfitting that might occur when repeatedly performing the training using only the training and test sets [26,27]. In the k-fold cross-validation, the test set was divided into k-folds, and training and validation were performed sequentially. The K value typically ranges from five to ten. In this study, it was set to five. After the cross-validation, the parameters of each model were adjusted to obtain the optimal performance. In this study, the grid search CV of the Python scikit-learn library was used to adjust the parameters and found those parameters with optimal performance for each model.

3.6.2. Model Evaluation

Because the crime data used in the training were imbalanced, it was difficult to determine how well the minority class was predicted by evaluating the model with any accuracy, and this was a classification indicator for the entire dataset [28]. Therefore, suitable methods for evaluating imbalanced data must be considered. This study evaluated the performance of each model using a confusion matrix [29,30], which is primarily used when evaluating the performance of general algorithms and imbalanced data. The confusion matrix compares the results predicted by the model with the actual class in the data, and classifies them as TN, TP, FP, or FN. Using this, the precision and recall values were obtained and harmonized in order to calculate the F1 score. The accuracy and F1 score of each model were compared to evaluate prediction performance (Figure 5).

		Predicted	
		Cold Spot	Hot Spot
Observed	Cold Spot	TN (True Negative)	FP (False Positive)
	Hot Spot	FN (False Negative)	TP (True Positive)

Figure 5. Example of a confusion matrix.

4. Results

4.1. Model Prediction Results

Table 3 lists the model prediction results based on the difference between the clustering and resampling methods. The model using spatial clustering showed higher F1 scores than the calculation method that combined the adjacent cells. Accordingly, there are differences in the physical environmental factors influenced by the detailed method of crime. Based on the findings of previous studies, a crime involving the same offender is likely to occur around the area of the original crime, indicating that a repeat of the crime is more likely in areas with similar spatial features to the area where the crime occurred. For both the SMOTE and random undersampling techniques, when the minimum threshold for a cell was $n = 6$, the F1 score was the highest, at 33.85% and 34.90%, respectively, and the F1 score increased by approximately 2% compared to the method combining the adjacent cells. In the models using the max-p method, the SMOTE-based model showed a regular pattern in which the F1 score gradually decreased as the distance from $n = 6$ increased, whereas the F1 score in the random undersampling-based model showed an irregular pattern according to the *p*-value. The results show low stability because the random undersampling method randomly deletes the instances. The pattern of the F1 score in the SMOTE-based model indicates that the model's performance may decrease if the *p*-value is too small or too large, and that there is a value yielding the optimal performance.

Table 3. Model performance according to resampling method and minimum threshold.

Resampling	Threshold Value	Precision	Recall	Accuracy	F1 Score
SMOTE	Surrounding grid	33.5727	30.1127	87.074	31.7487
	$n = 2$	31.8462	33.3333	86.2279	32.5728
	$n = 4$	35.8736	31.0789	87.5723	33.3046
	$n = 6$	34.9315	32.8502	87.1865	33.8589
	$n = 8$	32.4734	34.4605	86.3023	33.4375
	$n = 10$	32.6466	33.1723	86.4952	32.9073

Table 3. *Cont.*

Resampling	Threshold Value	Precision	Recall	Accuracy	F1 Score
Random Undersampling	Surrounding grid	22.81	57.48	76.33	32.66
	$n = 2$	24.13	56.33	78.02	33.73
	$n = 4$	22.35	59.09	75.41	32.43
	$n = 6$	24.26	62.15	76.84	34.90
	$n = 8$	24.15	56.52	77.94	33.84
	$n = 10$	24.25	55.23	78.31	33.71

Comparing the average accuracy and F1 scores of the models according to the resampling method, the SMOTE and random undersampling methods showed accuracies of 86.81% and 77.14% and F1 scores of 32.97% and 33.54%, respectively. Therefore, the SMOTE method had a 10% higher accuracy and a 0.5% lower F1 score than the random undersampling method. The random undersampling-based model showed a recall of approximately 55% to 62%, predicting many crime classes out of the total data. However, the precision and accuracy values were generally lower than those of the SMOTE, showing that its ability to accurately predict crime was inadequate. Figure 6 shows the models' prediction results according to the resampling method using a confusion matrix ($n = 6$). The value in the second quadrant is the number of data that correctly predicted cold spots (i.e., where no crime occurred), and the value in the fourth quadrant is the number of data that correctly predicted hot spots (i.e., where the crime occurred). The value in the first quadrant is the number of data points that incorrectly predicted a hot spot where the actual data were cold spots. The value in the third quadrant is the reverse (i.e., points that incorrectly predicted a cold spot where the actual data were hot spots). Considering the SMOTE method, 204 of the 584 data predicted as crime classes were correctly predicted, and for the random undersampling method, 386 of the 1591 data were correctly predicted. Because random undersampling deletes data from the majority class among all the data, precise prediction is difficult because of information loss.

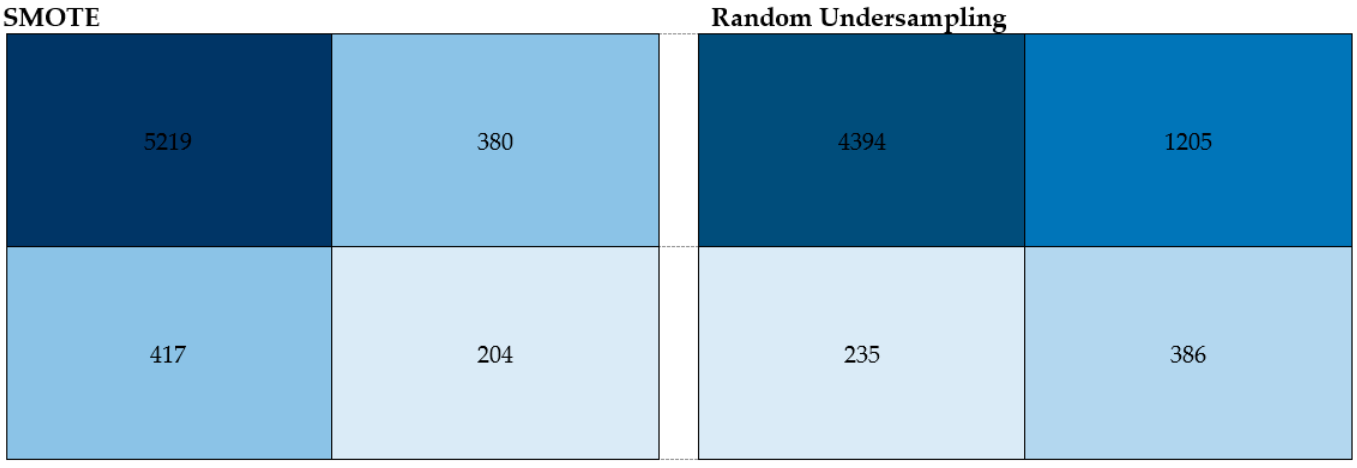
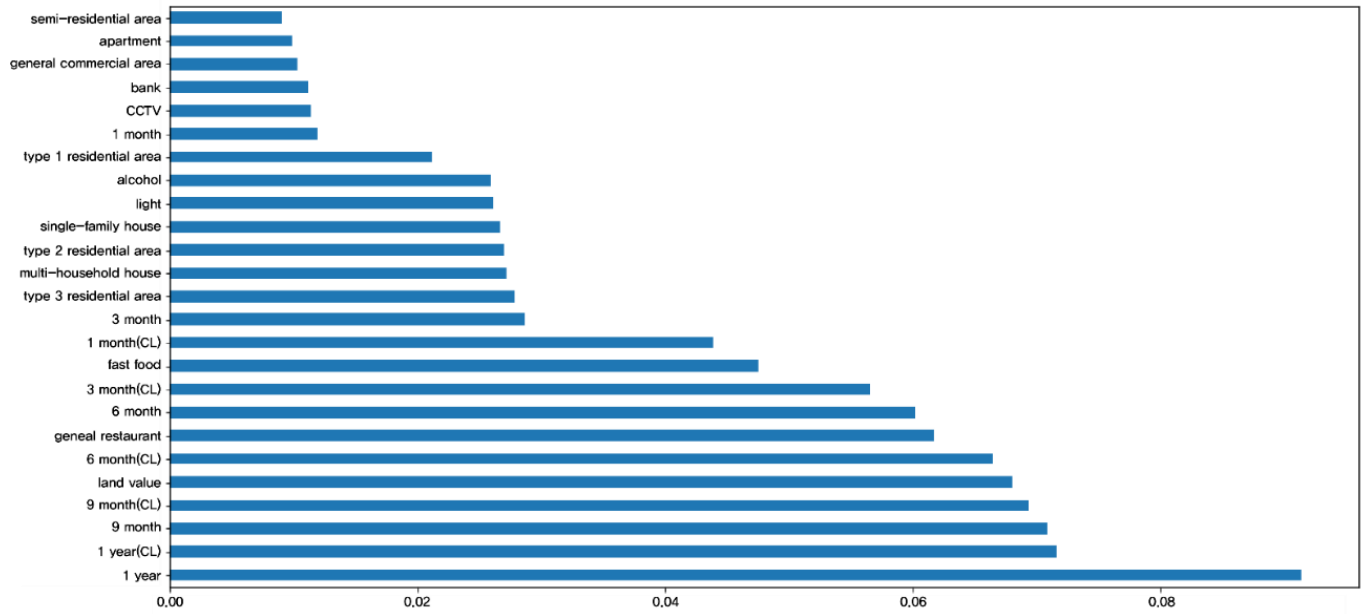


Figure 6. Confusion matrix results according to the resampling method ($n = 6$).

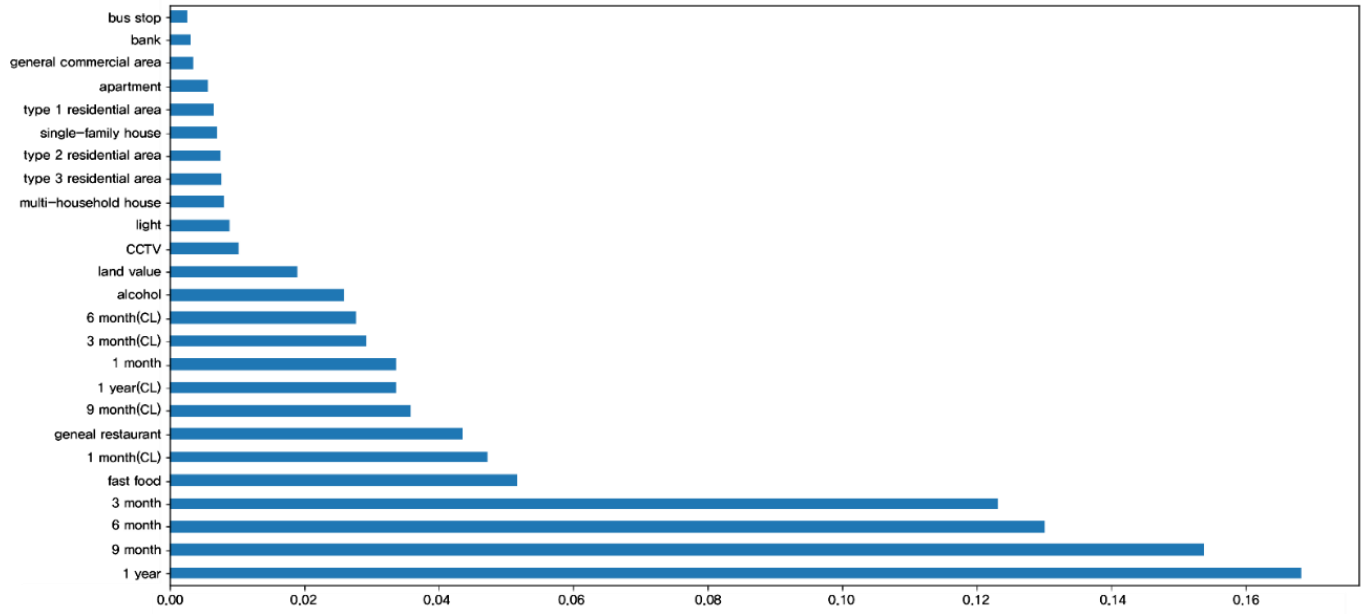
4.2. Feature Importance

In the case of the random forest algorithm, the feature importance function can be used to numerically express the influence of each variable for the prediction. Accordingly, this study analyzed the relative importance of each variable using this function. According to the analysis, the distribution of the feature importance varied with the resampling method (Figure 7). The feature importance was more evenly distributed under the random undersampling method than the SMOTE method. Because the random undersampling

method reduces the size of the entire dataset for training, the model is more sensitive to the features of the data with fewer samples.



(**a**) Random sampling



(**b**) SMOTE

Figure 7. Feature importance chart, following the resampling methods.

For both the random undersampling and SMOTE, time-related variables were the highest. Among these, the variable related to the average number of crimes that occurred in the cell over the previous year was the most important. In this regard, because crimes generally do not occur frequently, when the analysis period is shorter, less information can be learned from the variable. In contrast, crimes that occurred within the cluster showed different patterns according to the resampling method. Considering the random undersampling, the variables related to the average number of crimes during a particular period showed a higher importance as the period increased to six months, nine months, and 1 year, respectively. However, with regard to the SMOTE method, the influence of recent

crimes was high at three, six, and nine months. Because the SMOTE method generates new instances by interpolating the data, a large amount of data can be trained. Moreover, because the crime data created in the clustered instances are used together for the training, sufficient crime-related information can be obtained, even for short periods. While more recent crimes are known to generally have a greater influence on future crimes, in crime prediction research using machine learning it is important to appropriately configure the time-related variables, considering the training of the algorithm. Considering the physical environment-related variables, when using random undersampling, general restaurants showed the highest importance, followed by rest-area restaurants and pubs. When using the SMOTE method, the order of importance was rest-area restaurants, general restaurants, and pubs. However, the importance of residential buildings, banks, and CCTV-related facilities is relatively low. As is similar to the findings of previous studies on the influence of the surrounding environment, the likelihood of becoming a target of repeated crime is high when there are insufficient factors that can deter crime in places where people frequently engage in routine activities. Therefore, it is necessary to identify places where crime is spatiotemporally concentrated based on crowded spaces, predict where crime is likely to occur, and strengthen crime prevention activities in those places.

5. Conclusions

In previous grid-based crime prediction studies, information from all cells was combined and used for training the data on crimes in adjacent land. However, the actual land has a continuous flow, and the patterns and affected environmental factors vary with the method of crime. This study proposes a spatial clustering technique to solve this problem. The results showed that using reflecting spatial continuity to predict crime was effective in enhancing the model's performance. Moreover, by identifying the importance of each variable, it was found that there were places where crimes were spatiotemporally concentrated. With regard to the time-related variables, more recent crimes are known to have a greater influence on future crimes. However, it was difficult to significantly influence the model training if the period set as a variable was too short. Considering the physical environment-related variables, the feature importance of restaurants and pubs was high, suggesting that spaces frequented by people in their daily lives are more related to crime. Therefore, further in-depth analysis is required.

As part of the existing machine learning-based crime prediction research, this study is important because it provides guidelines for future related studies to apply spatial clustering in the crime prediction process and compare the results according to the cluster's configuration. Furthermore, this study attempted to predict the location of crimes more microscopically, using a grid unit in the analysis. Once this study is supplemented and practically applied in the future, it can help to improve the effectiveness of crime prevention by distributing crime prevention resources more efficiently.

Regarding this study's limitations, first, while the physical and environmental factors described in environmental criminology are highly diverse, this study used only some of them as variables. Second, because machine learning algorithms focus on prediction and classification using the given data rather than identifying the correlation between each variable, it is difficult to describe the correlation between each variable and the prediction result in detail. This study performed spatial clustering based on the entire target area; however, future studies can consider a method to derive cells with high similarity that is based on each cell. Finally, this study is expected to serve as a basis for further studies that use various variables and conduct both regression and statistical analyses.

Author Contributions: Conceptualization, D.K. and S.J.; methodology, D.K. and Y.J.; software, D.K.; validation, S.J.; formal analysis, D.K., Y.J. and S.J.; investigation, D.K. and Y.J.; resources, Y.J. and S.J.; data curation, D.K. and Y.J.; writing—review and editing, D.K and S.J.; visualization, D.K.; supervision, Y.J. and S.J.; project administration, S.J.; funding acquisition, S.J. 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).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Li, T.; Kou, G.; Peng, Y. Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods. *Inf. Syst.* **2020**, *91*, 101494. [CrossRef]
2. Sebastião, H.; Godinho, P. Forecasting and trading cryptocurrencies with machine learning under changing market conditions. *Financ. Innov.* **2021**, *7*, 1–30. [CrossRef]
3. Kou, G.; Peng, Y.; Wang, G. Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Inf. Sci.* **2014**, *275*, 1–12. [CrossRef]
4. Gupta, A.; Dengre, V.; Kheruwala, H.A.; Shah, M. Comprehensive review of text-mining applications in finance. *Financ. Innov.* **2020**, *6*, 1–25. [CrossRef]
5. Nayak, S.C.; Misra, B.B. Extreme learning with chemical reaction optimization for stock volatility prediction. *Financ. Innov.* **2020**, *6*, 1–23. [CrossRef]
6. Arietta, S.M.; Efros, A.A.; Ramamoorthi, R.; Agrawala, M. City forensics: Using visual elements to predict non-visual city attributes. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 2624–2633. [CrossRef] [PubMed]
7. Yu, C.-H.; Ward, M.W.; Morabito, M.; Ding, W. Crime forecasting using data mining techniques. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, 11–14 December 2011; pp. 779–786.
8. Lin, Y.-L.; Yen, M.-F.; Yu, L.-C. Grid-based crime prediction using geographical features. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 298. [CrossRef]
9. Johnson, S.D.; Bernasco, W.; Bowers, K.J.; Elffers, H.; Ratcliffe, J.; Rengert, G.; Townsley, M. Space–time patterns of risk: A cross national assessment of residential burglary victimization. *J. Quant. Criminol.* **2007**, *23*, 201–219. [CrossRef]
10. Bernasco, W. Them again? Same-offender involvement in repeat and near repeat burglaries. *Eur. J. Criminol.* **2008**, *5*, 411–431. [CrossRef]
11. Groff, E.R.; Weisburd, D.; Yang, S.-M. Is it important to examine crime trends at a local "micro" level?: A longitudinal analysis of street to street variability in crime trajectories. *J. Quant. Criminol.* **2010**, *26*, 7–32. [CrossRef]
12. Brantingham, P.L.; Brantingham, P.J. Mobility, notoriety and crime: A study of crime patterns in urban nodal points. *J. Environ. Syst.* **1982**, *11*, 89–99. [CrossRef]
13. Lee, D.; Yoon, S.; Kim, J. Analysis of the Crime Pattern and Influencing Factors by the Spatial Autocorrelation in Busan. *J. Korean Reg. Dev. Assoc.* **2015**, *27*, 259–276.
14. Lockwood, D. Mapping crime in Savannah: Social disadvantage, land use, and violent crimes reported to the police. *Soc. Sci. Comput. Rev.* **2007**, *25*, 194–209. [CrossRef]
15. Stucky, T.D.; Ottensmann, J.R. Land use and violent crime. *Criminology* **2009**, *47*, 1223–1264. [CrossRef]
16. Kwon, N.-Y.; Kwon, E.; Jung, S. A Study on the Classification of Theft using K-modes Clustering-Focused on Correlation between Land Use and Types of Theft. *J. Archit. Inst. Korea* **2020**, *36*, 81–90.
17. Cohen, L.E.; Felson, M. Social change and crime rate trends—A routine activity approach. *Am. Sociol. Rev.* **1979**, *44*, 588–608. [CrossRef]
18. Brantingham, P.L.; Brantingham, P.J. Environment, routine, and situation: Toward a pattern theory of crime. *Adv. Criminol. Theory* **1993**, *5*, 259–294.
19. Wolfgang, M.E.; Figlio, R.M.; Sellin, T. *Delinquency in A Birth Cohort*; University of Chicago Press: Chicago, IL, USA, 1987.
20. Sherman, L.W.; Gartin, P.R.; Buerger, M.E. Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology* **1989**, *27*, 27–56. [CrossRef]
21. Ferligoj, A.; Batagelj, V. Clustering with relational constraint. *Psychometrika* **1982**, *47*, 413–426. [CrossRef]
22. Wise, S.; Haining, R.; Ma, J. Regionalisation tools for the exploratory spatial analysis of health data. In *Recent Developments in Spatial Analysis*; Fischer, M.M., Getis, A., Eds.; Springer: Berlin/Heidelberg, Germany, 1997; pp. 83–100.
23. Hansen, P.; Jaumard, B.; Meyer, C.; Simeone, B.; Doring, V. Maximum split clustering under connectivity constraints. *J. Classif.* **2003**, *20*, 143–180. [CrossRef]
24. Duque, J.C.; Anselin, L.; Rey, S.J. The max-p-regions problem. *J. Reg. Sci.* **2012**, *52*, 397–419. [CrossRef]
25. Mani, I.; Zhang, I. kNN approach to unbalanced data distributions: A case study involving information extraction. In Proceedings of the International Conference on Machine Learning (ICML 2003), Workshop on Learning from Imbalanced Data Sets, Washington, DC, USA, 21 August 2003.
26. Zhang, Y.; Yang, Y. Cross-validation for selecting a model selection procedure. *J. Econom.* **2015**, *187*, 95–112. [CrossRef]

27. Santos, M.S.; Soares, J.P.; Abreu, P.H.; Araujo, H.; Santos, J. Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. *IEEE Comput. Intell. Mag.* **2018**, *13*, 59–76. [CrossRef]
28. Chawla, N.V. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*; Springer: Boston, MA, USA, 2009; pp. 875–886.
29. Bekkar, M.; Djemaa, H.K.; Alitouche, T.A. Evaluation measures for models assessment over imbalanced data sets. *J. Inf. Eng. Appl.* **2013**, *3*, 27–38.
30. Luque, A.; Carrasco, A.; Martín, A.; de las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **2019**, *91*, 216–231. [CrossRef]