THEFT CRIME PREDICTION BASED ON SPATIAL AND TEMPORAL CRIME HOTSPOTS

Kevin Ke, Angela Chiang, Kira Liang, Jack Liang
Institute for Information Industry, Taipei, Taiwan (ROC)

ABSTRACT

This study aims to find the spatial and temporal theft hotspots in Taipei city by analyzing real-world theft datasets through a statistical analysis and the utility of hotspot mapping.
First, we collected several environmental factors such as population structure, salary level, house price that corresponded to 456 villages in Taipei city. Then, by conducting the Random Forest algorithm, we ruled out the less important attributes and continued the rest on k-means clustering. We intend to group the village with similar environmental factors in the same cluster to avoid excessive disorganized rules while implementing the Apriori algorithm. Then, the study clarifies how we conducted Apriori algorithm to produce interesting frequent patterns for crime hotspots. The results of this solution could be used to raise people's awareness regarding the dangerous locations and to help law enforcement to predict future crimes in a specific location within a particular time-frame.

KEYWORDS

Data mining, crime prediction, crime frequent patterns, Taipei theft hotspots

INTRODUCTION

Data analysis can be applied to several business situations, such as e-commerce, retailing and advertising. However, despite those commercial values we hope to conduct a report that is closely related to our daily life. Thus, Maslow's Hierarchy of Needs came to our mind --- one of the basic needs of the theory for living is safety. Crime can cause social unrest and make us in fear and it is also the most common social problem affecting the quality of life and the economic growth of a society [1]. Reducing crime is one of the efforts that are being taken seriously by the Taiwanese Government. Based on the 2016 criminal statistics report published by the Law Enforcement Agency, theft is the most common crime and burglary rate has also increased among the Taiwanese youth. Thus, this study will focus on the theft crime and predictive mapping in crime prevention, it can capture the information from the patterns to predict the possible crime occurrence for a better crime prediction solution and support in the distribution of police at most likely crime places for any given time, to grant an efficient usage of police resources [2].

DATASETS

To construct our data mining model, we mainly focused on two tables. Our study is based on the theft crime in Taipei city, and we hope to find patterns of the occurrence time for a specific location. To begin with, we collected several environmental factors that corresponded to each village in Taipei. As indicated in Table 1, we classified the villages into different cluster based on those environmental factors before we conducted the Apriori algorithm. This step was taken to reduce the noise and to avoid excessive rules.

Table 1. Village attributes table[3]

<table>
<tr><th>Column</th><th colspan="3">Data Type</th><th>Source</th></tr>
<tr><td><b>Village</b></td><td colspan="3">Nominal</td><td>Department of Civil Affairs</td></tr>
<tr><td><b>Number of theft crime</b></td><td colspan="3">Numerical</td><td>Taipei open data</td></tr>
<tr><td><b>Number of monitors</b></td><td colspan="3">Numerical</td><td>Taipei open data</td></tr>
<tr><td><b>Number of police stations</b></td><td colspan="3">Numerical</td><td>Taipei open data</td></tr>
<tr><td><b>Number of parks</b></td><td colspan="3">Numerical</td><td>Taipei open data</td></tr>
<tr><td><b>Park area</b></td><td colspan="3">Numerical</td><td>Taipei open data</td></tr>
<tr><td><b>Metro population flow</b></td><td colspan="3">Numerical</td><td>Taipei MRT</td></tr>
<tr><td><b>House price</b></td><td colspan="3">Numerical</td><td>Ministry of the Interior</td></tr>
<tr><td><b>Salary level</b></td><td colspan="3">Numerical</td><td>Ministry of Finance</td></tr>
<tr><td rowspan="2"><b>Population structure</b></td><th>Data Type</th><th colspan="2">Categories</th><td rowspan="2">Taipei open data</td></tr>
<tr><td>Numerical</td><td>Male
Female</td><td>Youth
Adult
Senior</td></tr>
<tr><td><b>Education level</b></td><td>Numerical</td><td colspan="2">High
Medium
Low</td><td>Taipei open data</td></tr>
</table>

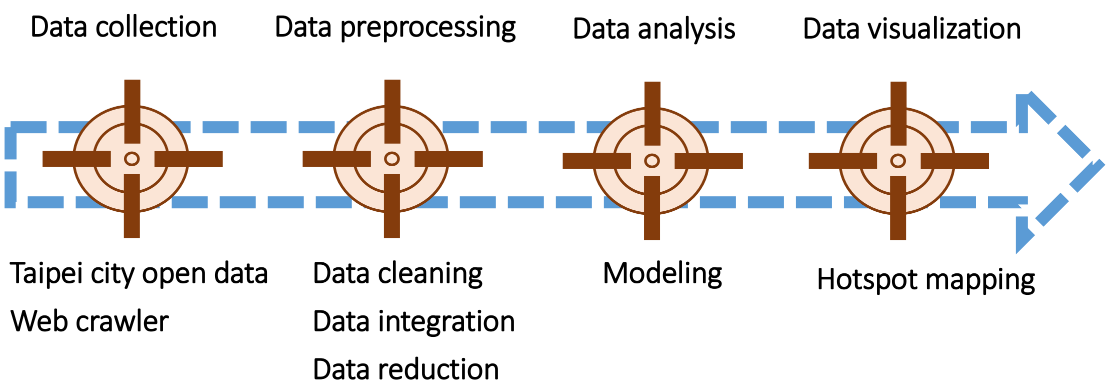
As for the Apriori analysis, we focused on the Taipei city theft crime from previous two years to current year (2015/1-2017/2). The dataset information is based on the Taipei Open Data. The original dataset gives the exact occurrence time of the crime along with the address. We extracted the village from the address and corresponded each theft crime to the village. Furthermore, we also transformed the exact occurrence time to a two hour time interval. The following table shows the crime attributes and its content values (Table 2).

Table 2. Crime attributes table

Column	Data Type	Number of Distinct Values	Values
Crime type	Nominal	4 categories	Burglary Automobile theft Bike theft Shoplifting
Date	Date	unlimited	2016/9/15
Weekday	Day	7 categories	Monday Tuesday Wednesday …
Occurrence Time	Time	12 time intervals	24:00-02:00 02:00-04:00 04:00-06:00 …
Village	Nominal	456 names	Furen village

METHODOLOGY

In this section, we explain how we prepared our datasets. After that, we provide how we analyzed the data using some statistical analysis. Then, we introduce how we constructed our data-mining models to achieve our purpose.



Data preprocessing

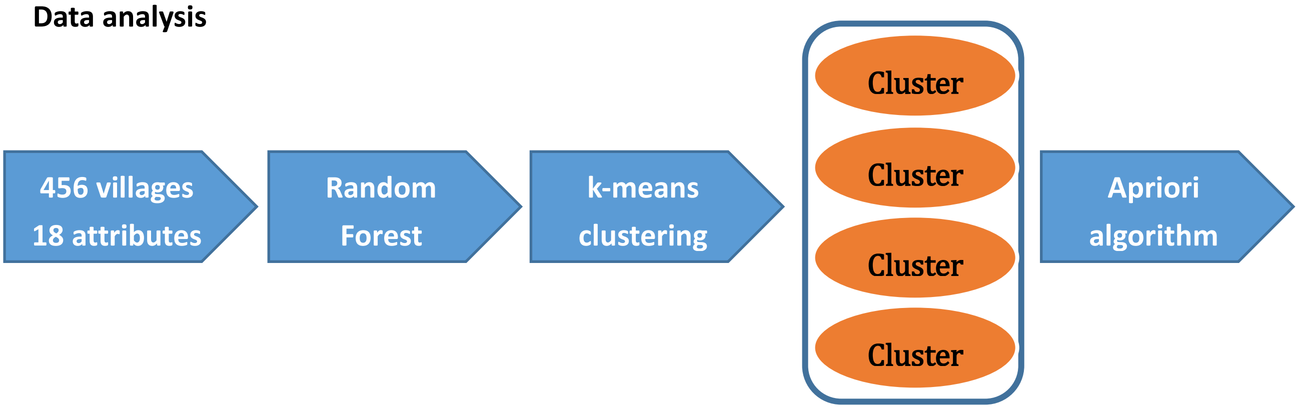
There are few missing values in some attributes such as house price or education level of some villages. We compared to the nearby village and calculated the average value to fill in the blank. Moreover, we eliminated the outliers to maintain the objectivity of the analysis. Finally, we normalized every numerical value before constructing the data mining model.

The normalized formula:

$$X_n = \frac{(X - Xmin)}{(Xmax - Xmin)}$$

We also found out that some village names appeared in different character written in the tables, we performed several steps to unify the village names and key attributes.

Data analysis



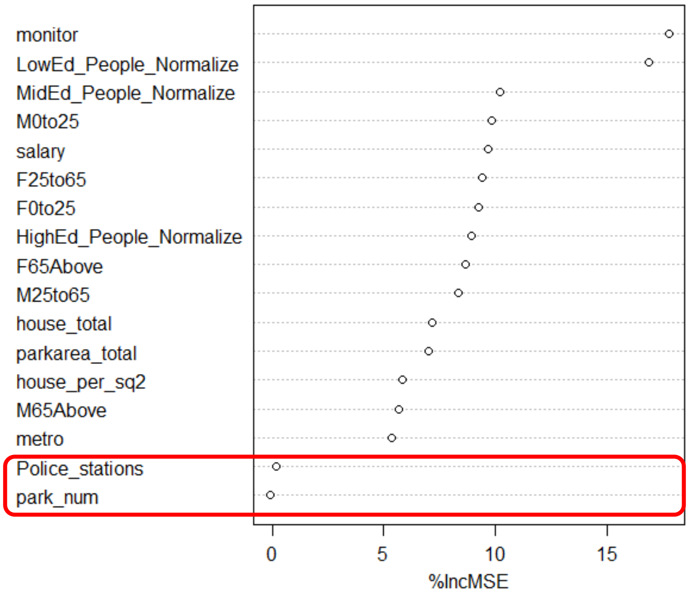
Before conducting the modeling method, we sorted out 18 attributes from our Village attributes table (Table 1):

1. *Number of theft crime*
2. *Number of monitors*
3. *Number of police stations*
4. *Number of parks*
5. *Park area*
6. *Metro population flow*
7. *Total house price*
8. *House price per square meter*
9. *Salary level*
10. *Low-education population*
11. *Mid-education population*
12. *High-education population*
13. *Male youth population (0-15 years old)*
14. *Male adult population (15-65 years old)*
15. *Male senior population (65 and above)*
16. *Female youth population (0-15 years old)*
17. *Female adult population (15-65 years old)*
18. *Female senior population (65 and above)*

Random Forest Classifier

Random Forest Classifier is an unsupervised learning algorithm, which is operated by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification). In order to determine which attribute is important to our clustering model, we constructed this model using "randomForest" that provides a set of open source data-mining tools for R. We set the number of theft crime as our key attribute in Table 1 and computed the model mean squared error (MSE). For each attribute in the model we permuted the attribute and calculated new model MSE according to variable permutation, then we collected the results in a list and ranked attributes' importance according to the value of the percentage of increasing mean squared error (%IncMSE). The greater the value the better it is (Figure 1).

Figure 1. Attribute importance



We decided to remove number of police stations and number of parks since %IncMSE is approximately close to zero and continued the other 16 attributes to k-means clustering.

k-means clustering

k-means clustering is also a type of unsupervised learning algorithm which can be applied while having data without defined categories or groups [4]. The goal of this method is to find k points (known as cluster centers or prototypes) of a dataset that can minimize the within-cluster sum of squares (WCSS). After obtaining these cluster centers, we can use these cluster centers for data classification such that the influence from noisy data is reduced.

The formula of k-means clustering is defined as:

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - u_k||^2$$

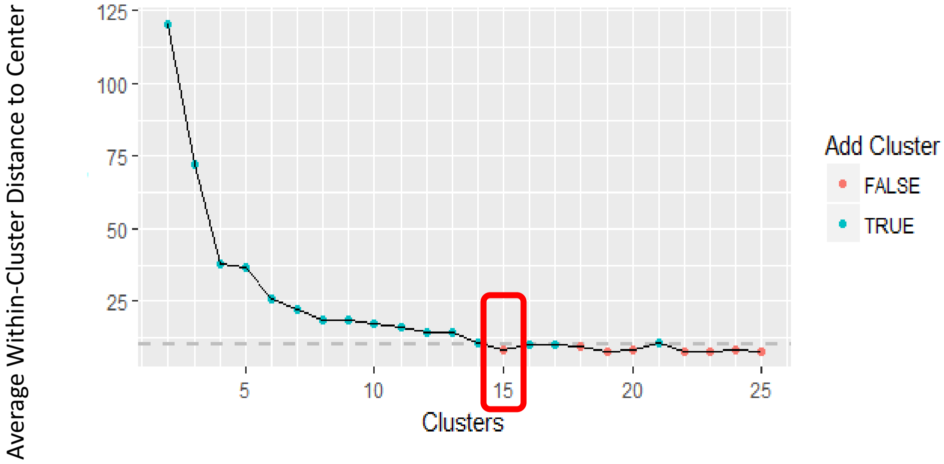
'$||x_i - u_K||$' is the Euclidean distance between x_i and u_k.

'c_k' is the number of data points in i^{th} cluster.

'k' is the number of cluster centers

To find the numbers of clusters in the data, we compared the mean distance between data points and their cluster centers across different values of K. This method is plotted in figure 2 and the "elbow point" is where the rate of decrease shifts, that can be used to roughly determine K [5].

Figure 2. The number of clusters



Hence, we run the k-means clustering setting K=15, plotted number of villages in each cluster (Figure 3) and calculated the average number of theft crime (Figure 4).

Figure 3. Number of villages

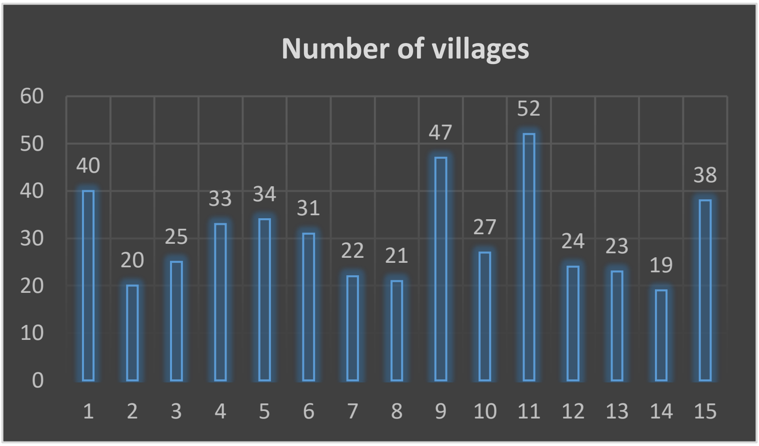
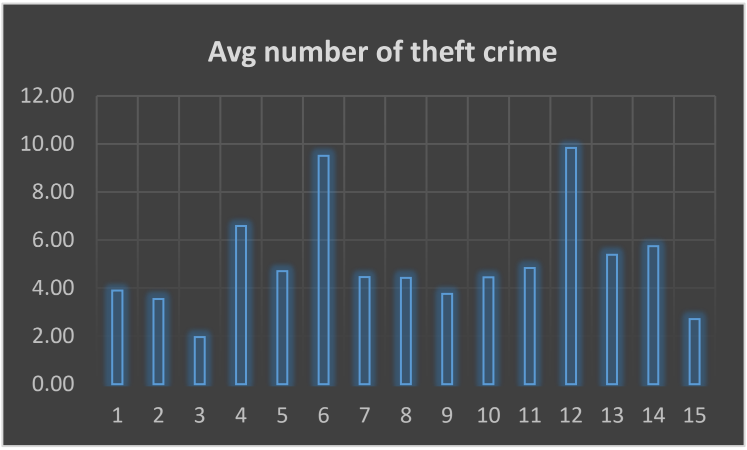
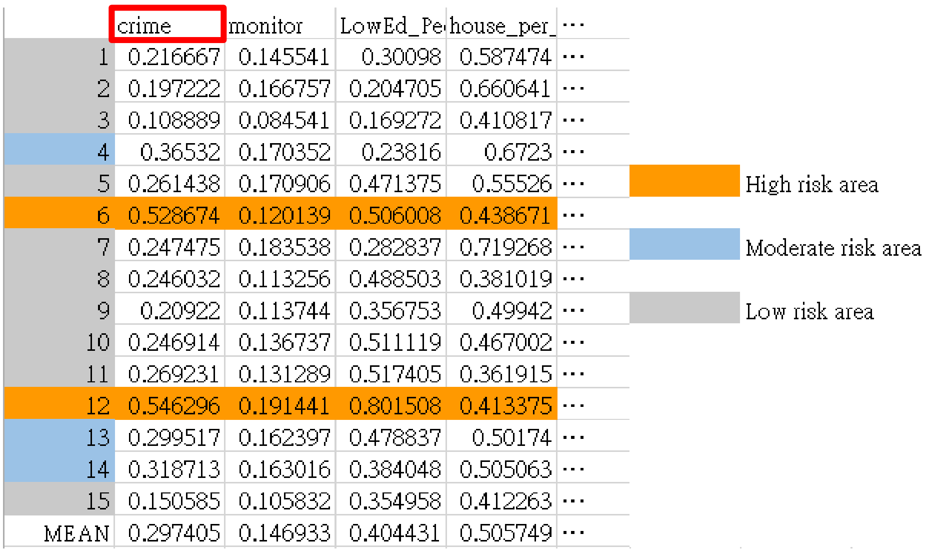


Figure 4. Average number of theft crime



The number of villages of each cluster falls between 19 and 52, and we can see that cluster 6 and cluster 12 have higher theft crime rate. Thus, we examined the k-means clustering centers for each cluster and all their attributes (Figure 5), each value represents the center value of the attribute for its corresponding cluster and the value scores from 0 to 1, since we conducted the normalized formula in the data preprocessing session. We found out that crime attribute (number of theft crime) in cluster 6 and cluster 12 have far exceeded other cluster, hence we classified those two cluster into high risk area, for those whose crime attribute is higher than average value, we categorized them into moderate risk area and the rest of the clusters are referred to the low risk area. By comparing other attributes of the high-risk area, we also found out that those villages tend to have less surveillance monitors, higher rate of lower-educated people and lower house price.

Figure 5. k-means clustering centers



	crime	monitor	LowEd_Pe	house_per_	⋯
1	0.216667	0.145541	0.30098	0.587474	⋯
2	0.197222	0.166757	0.204705	0.660641	⋯
3	0.108889	0.084541	0.169272	0.410817	⋯
4	0.36532	0.170352	0.23816	0.6723	⋯
5	0.261438	0.170906	0.471375	0.55526	⋯
6	0.528674	0.120139	0.506008	0.438671	⋯
7	0.247475	0.183538	0.282837	0.719268	⋯
8	0.246032	0.113256	0.488503	0.381019	⋯
9	0.20922	0.113744	0.356753	0.49942	⋯
10	0.246914	0.136737	0.511119	0.467002	⋯
11	0.269231	0.131289	0.517405	0.361915	⋯
12	0.546296	0.191441	0.801508	0.413375	⋯
13	0.299517	0.162397	0.478837	0.50174	⋯
14	0.318713	0.163016	0.384048	0.505063	⋯
15	0.150585	0.105832	0.354958	0.412263	⋯
MEAN	0.297405	0.146933	0.404431	0.505749	⋯

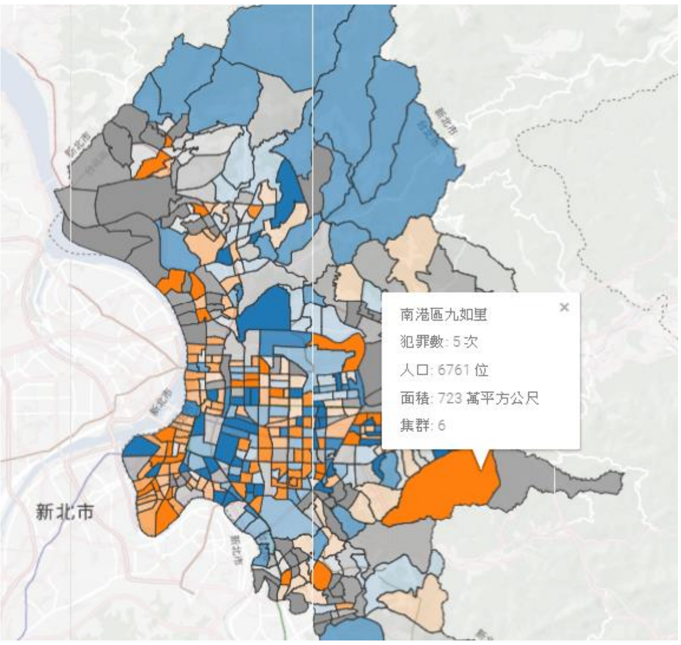
🟧 High risk area

🟦 Moderate risk area

⬜ Low risk area

To visualize our clustering result, we collected latitude and longitude coordinates of village boundaries and created several polygons with Google Maps API. We filled the village with different color based on the risk of the area. While clicking on each village, the user can see more detailed information of the location such as the number of theft crime, population, area and which cluster it belongs to (Figure 6).

Figure 6. Risk area in Taipei city


南港區九如里
犯罪數: 5 次
人口: 6761 位
面積: 723 萬平方公尺
集群: 6

Apriori algorithm

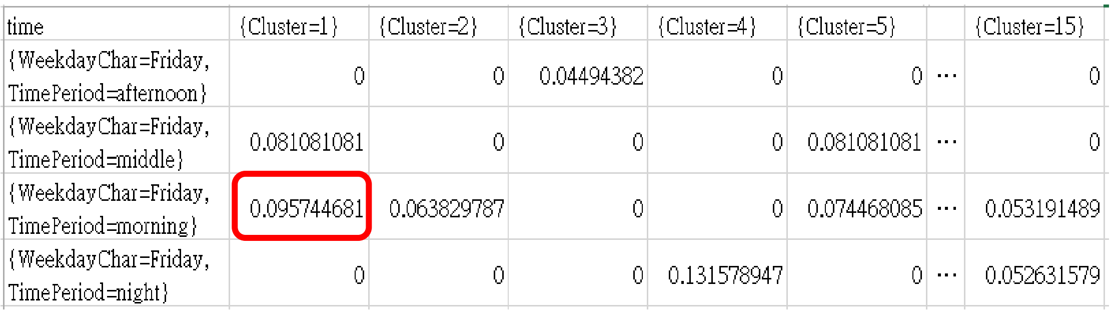
Apriori is one of the basic algorithms in association rule. It is a common way for mining frequent patterns in the dataset [6]. Our goal for using this model is to find all possible crime hotspots along with its related frequent time. After conducting the k-means clustering, we grouped 456 villages into 15 clusters. Then, we added an extra column for indicating the cluster behind the village column in Table 2.

Apriori algorithm has its support value, confidence value and lift. The higher the confidence value, the stronger the rule [7]. Lift value shows that the existence of the rule is not just a random occurrence while it is bigger than 1. We implemented this model using Table 2 and conducted multiple experiments using different minimum support values, then we selected the optimal choice. Furthermore, we also applied constraint-based mining by restricting the extraction process on the frequent patterns having this formula of three specific itemsets (Weekday, Time period and cluster). Finally, the minimum support value for our datasets was 0.001 and we obtained 188 absolute frequencies after the filtering criteria.

Crime prediction

The goal of our study was finding spatial and temporal criminal hotspots. We achieved this goal by conducting the Apriori algorithm. Table 3 indicates that the occurrence rate of the theft crime in a specific time of each cluster. For example, there's a 9.57% chance of threat for the villages in cluster 1 to occur the theft crime in Friday morning.

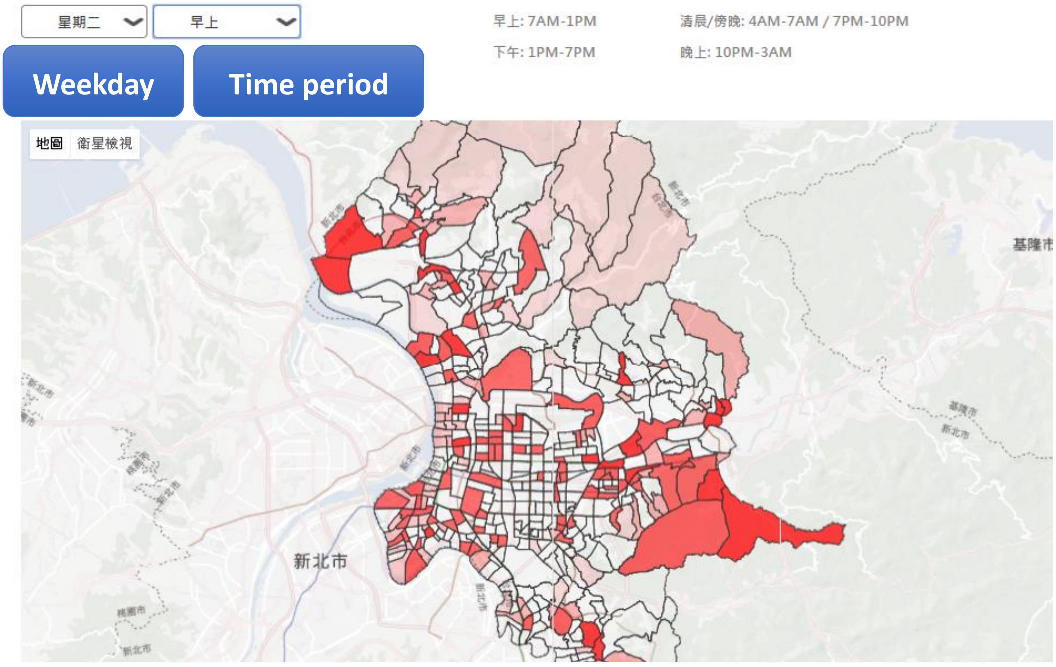
Table 3. Apriori results for Taipei theft crime prediction



time	{Cluster=1}	{Cluster=2}	{Cluster=3}	{Cluster=4}	{Cluster=5}		{Cluster=15}
{WeekdayChar=Friday, TimePeriod=afternoon}	0	0	0.04494382	0	0	…	0
{WeekdayChar=Friday, TimePeriod=middle}	0.081081081	0	0	0	0.081081081	…	0
{WeekdayChar=Friday, TimePeriod=morning}	0.095744681	0.063829787	0	0	0.074468085	…	0.053191489
{WeekdayChar=Friday, TimePeriod=night}	0	0	0	0.131578947	0	…	0.052631579

After obtaining the threat of each cluster in different time period, we visualized our results using Google Maps (Figure 7). With two pull-down menu on the top, users can choose the day and the time period easily, then the map will show the corresponding color. While the brighter the color is, the higher probability will the theft crime take place.

Figure 7. Visualization for crime prediction



Conclusion and future work

By analyzing the data from the past, we can identify where crime most densely concentrates and by conducting the predictive modeling, we can start a decision-making process that considers where to deploy the police force and to target the prevention resources.

As a future extension of our work, we consider adding more categories of crimes and expand our analysis to other cities in Taiwan. We hope by studying other crimes datasets from new cities along with their demographics datasets can enhance the performance of our model. Last but not least, we hope our study can benefit the appropriate party or law enforcements to combat the crime in our country and keep our community safer for everyone.

References

[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi and A. Pentland, 'Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data', CoRR, vol. 14092983, 2014.

[2] S. Nath, 'Crime Pattern Detection Using Data Mining', in Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on, 2006, pp. 41,44.

[3] Taipei Open Data [Online]. Available: http://data.taipei/

[4] J. A. Hartigan and M. A. Wong (1979) "A K-Means Clustering Algorithm", Applied Statistics, Vol. 28, No. 1, p100-108.

[5]DataScience: Introduction to k-means clustering [Online]. Available: https://www.datascience.com/blog/introduction-to-k-means-clustering-algorithm-learn-data-science-tutorials

[6]Buczak, A. L. and Gifford C. M. (2010) Fuzzy association rule mining for community crime pattern discovery. ACM SIGKDD Workshop on Intelligence and Security Informatics. ACM.

[7] Herawan, T. and Deris, M. M. (2011) A soft set approach for association rules mining. *Knowledge-Based Systems*