Risk discovery of car-related crimes from urban spatial attributes using emerging patterns

Atsushi Takizawa\textsuperscript{a,}\textsuperscript{*}, Fumie Kawaguchi\textsuperscript{a}, Naoki Katoh\textsuperscript{a}, Kenji Mori\textsuperscript{b} and Kazuo Yoshida\textsuperscript{b}

\textsuperscript{a}Department of Architecture and Architectural Engineering, Graduate School of Engineering, Kyoto University, Kyoto University Katsura Campus, Nishikyo-ku, Kyoto 615-8540, Japan
\textsuperscript{b}Crime Analysis Office, Kyoto Prefecture Police, Kamazadorishimotateuri, Kamigyo-ku, Kyoto 602-8550, Japan

Abstract. Using spatial data mining techniques, we analyze car-related crimes such as auto theft, auto parts theft, and breaking into a car in the area of Nishikyo-ku, Kyoto City. The strategy of natural surveillance proposed by Crime Prevention Through Environmental Design (CPTED) is taken into consideration as visibility attributes. From the viewpoint of risk discovery, we do not employ ordinary association rule but a new data mining technique called Emerging Patterns (EPs). EP is defined as an itemset whose support increases significantly from one dataset to another. Since a large number of EPs are generated in general as in association rule, it is difficult to identify the critical factors which affect crime occurrences. Therefore, we will introduce two new ideas; (a) appropriately aggregating several EPs with high growth-rate and (b) identifying a pair of similar patterns A and B such that A is not associated with high crime occurrences while B is highly correlated with crime occurrences. Finding such similar patterns reveals that the attribute value which is in B but not in A is then identified as a critical factor which arouses crimes when combined with certain factors.

1. Introduction

Recently, in Japan, atrocious crimes out of doors are standing out such that elementary school students were killed on the way home from school one after the other. Not only these atrocious crimes but also street crimes can not been overlooked. Street crimes such as auto theft or bag snatching dominate about half of crime occurrences in Japan.

Street crimes were considered to be related to the structure of urban space from 1960s of USA [1]. Then, the concept called Crime Prevention Through Environmental Design (CPTED) was formulated by criminologist Jeffery [2]. A more limited approach, termed defensible space, was developed concurrently by architect Newman [3]. The built environment strategy of CPTED has been modified and summarized nowadays in the following; natural surveillance, natural access control and natural territorial reinforcement.

Natural surveillance and access control strategies limit the opportunity for crime. Natural territorial reinforcement promotes social control. The strategy of CPTED is now drawing attention of urban planners and architects in many countries because of the rapid increase of crime or terrorism in urban areas.

The effectiveness of these strategies has been known empirically. It has not been evaluated, however, on the basis of objective data yet. The reason seems to be that a city is very complex and a lot of information is needed for analysis. Unlike the general tabular form data, spatial data is highly structured and difficult to be analyzed. These have been a barrier against deep and objective analysis.

Development of GIS of late years has increased possibility to meet such a demand. Current main contribution of GIS to crime analysis is crime mapping. Prior task of crime mapping is simply to plot the data of
crime occurrence points by using GIS, but this task may lead to some useful supports for crime deterrent and criminal investigation. One of them is to identify the area, where crimes occur often, called “hot spot” [4]. Hot spot can be obtained by spatial statistical analysis such as Kernel Density Estimation (KDE) [5] or spatial auto correlation. KDE is generally used for obtaining the hot spot because it can visualize the crime occurring status of an area without visualizing actual crime occurring points. Considering the privacy of victims, this feature of KDE is very important for showing the crime map to the public.

However, KDE considers only the density of points but not take into consideration the relationship with building and the environmental factors. On the contrary, spatial auto correlation can deal with the relationship of urban objects. However, since spatial entity is abstracted as a point form, the information concerning building forms is entirely neglected. Therefore, current crime spatial analysis methods are not adequate to our need concerning more concrete spatial factor.

Instead of the above methods, we are interested in spatial data mining technique [6] which can deal with the complexity of urban space. Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns in the GIS database. Spatial relationships have been classified into several types including distance relations, direction ones and topological ones. The purpose of spatial data mining is roughly divided into classification, clustering, and mining association rules. Among them mining spatial association rule has been massively studied [7–9].

An association rule is defined on transaction-based databases. It is expressed in the form of “W \rightarrow B (c\%)”, which says “if a pattern W appears in a transaction, there is c\% possibility (confidence) that the pattern B holds”, where W and B are a set of attribute values. Moreover, to ensure that such rules are interesting enough to cover frequently encountered patterns in a database, the concept of the support of a rule “W \rightarrow B” has been commonly used, which is defined as the ratio of B in the patterns of W. Though an association rule is simple and easy to understand, it has one serious drawback for practical use in that a lot of rules tend to be generated and most of them are well-known and rather trivial. The reason is that the association rule having higher support value is more evaluated.

From the viewpoint of risk discovery, we have to pay attention to the spatial patterns that have not been well known until now but are very important for crime occurrence. From this reason, instead of association rule we apply the concept of emerging patterns (EPs) [10], EP is defined as an itemset whose support increases significantly from one dataset to another. EPs can find itemsets which are typical to the class concerned but not to the other class. EPs can capture emerging trends in data of many domains, or useful contrasts between data classes. This feature of EPs is thought to be adequate for finding minor but important spatial patterns which we really want to know.

In this paper, car-related crime data of Kyoto-shi Nishikiyo-ku where is the suburb of Kyoto City is analyzed. By using EPs, a lot of interesting patterns are found. Since a large number of EPs are generated in general as in association rule, it is difficult to identify the critical factors which affect crime occurrences. In our analysis, we will introduce two ideas; the first is to appropriately aggregate several EPs while maintaining high growth-rate in order to increase the readability of patterns generated, and the second is to identify a pair of patterns A and B such that both consist of the same set of attribute values except one such that while the former pattern A is not associated with high crime occurrences while the pattern B is highly correlated with crime occurrences although A and B are very much similar. This reveals that the attribute value which is in B but not in A is then identified as a critical factor which arouses crimes when combined with certain factors.

The rest of this paper is organized as follows: In the next section, we present the outline of EP. In section 3, spatial attributes used as explaining variables are described. In section 4, crime and geographic data used for our analysis are explained. In section 5, results of EP mining are presented and some are discussed. Section 6 concludes our paper.

2. Emerging patterns

We assume that a database is composed of a set of attributes and that the original dataset denoted by T (i.e., a set of transactions) in a database have attribute set denoted by A. Each transaction is associated with a class label. Here we assume there are two labels, P and N which represents crime occurrence or non crime occurrence respectively in our analysis. According to class labels, T is partitioned into two datasets \( D_P \) and \( D_N \). A transaction is expressed as \( \{(A_i, v_i^t) | A_i \in A\} \), where \( v_i^t \) is the value of attribute \( A_i \) in transaction \( t \in T \). Each pair (attribute, value) is called an item, and an itemset is a set of items.

Let \( I = \{i_1, i_2, \ldots, i_N\} \) be a set of all possible items. A transaction can be identified with a subset of
A subset $X$ of $I$ is called a $k$-itemset when $k = |X|$. We say a transaction $t$ contains an itemset $X$, if $X \subseteq T$. The support of an itemset $X$ in a dataset $D$ denoted by $\text{supp}_D(X)$ is derived from $\#D(X)/|D|$. Here, $\#D(X)$ denotes the number of transactions containing $X$ in $D$. Given a positive number $\sigma$, we say an itemset $X$ is $\sigma$-large in $D$ if $\text{supp}_D(X) \geq \sigma$, and $X$ is a $\sigma$-small in $D$ otherwise. Let $\text{Large}_\sigma(X)$ (resp. $\text{Small}_\sigma(X)$) denote the collection of all $\sigma$-large (resp. $\sigma$-small) itemsets.

For a given ordered pair of datasets $D_P$ and $D_N$, the growth-rate of an itemset $X$ from $D_N$ to $D_P$ denoted by $\text{Growthrate}(X)$ is defined as:

$$\text{Growthrate}(X) = \begin{cases} 0, & \text{if } \text{supp}_{D_P}(X) = 0 \& \text{supp}_{D_N}(X) = 0, \\
\infty, & \text{if } \text{supp}_{D_P}(X) \neq 0 \& \text{supp}_{D_N}(X) = 0, \\
\frac{\text{supp}_{D_P}(X)}{\text{supp}_{D_N}(X)}, & \text{otherwise.} \end{cases}$$

Given $\rho > 1$ as a growth-rate threshold, an itemset $X$ is said to be an $\rho$-emerging pattern ($\rho$-EP or simply EP) from $D_N$ to $D_P$ if $\text{Growthrate}(X) \geq \rho$.

The above defined $\text{Growthrate}(X)$ will be often denoted by $\text{Growthrate}_P(X)$ in what follows. $\text{Growthrate}(X)$ can be defined by interchanging the role of $D_P$ and $D_N$ which is denoted by $\text{Growthrate}_D(X)$.

### 3. Attributes

#### 3.1. Visibility

As described in Section 1, the natural surveillance is a central principal of CPTED. The degree of natural surveillance is highly correlated with the visibility of the building environment. Benedikt [11] first proposed the concept “isovist” which represents visible space from a single view point.

In this study, in order to quantify the visibility from a given point, we will do as follows. In order to define the visibility in the 3D space, a half-world dome is introduced for modeling 3D isovist area. We call this dome “visibility dome”. Figure 2 shows the concept of a visibility dome. Visibility dome consists of three parameters; $R$ (meters), $H$ (degree) and $V$ (degree), where $R$ denotes the radius of the dome, $H$ denotes the angle of adjacent two visible axes on the horizontal surface around the view point, and $V$ denotes the angle of adjacent two visible axes on the vertical surface. Since natural surveillance is influenced by an individual sense, the type of building is considered to be important. Generally a residence has high surveillance performance throughout the day, whereas other buildings such as office or factory has less performance and their efficiency is limited to the hour of operation.

The visibility of the viewpoint $p = 1, 2, \ldots, P$ is obtained as follows where $P$ is the number of viewpoint points. Let $a^i_P, i = 1, 2, \ldots, N$ denote the $i$-th ray from $p$, where $N$ is the total number of rays. The type of building which $a^i_P$ first hits is denoted by $l(a^i_P)$, where $l(a^i_P) \in \{\text{detached house (DH), multi-dwelling (MD), commercial and corporate building (CB), public building (PB) and non-wall building (NW)}\}$. And let $l(a^i_P)$ with $0 \leq l(a^i_P) \leq r$ denote the length of the $i$-th ray from $p$ to the building. If there is no building within $r$ along to the direction $a^i_P$, let $l(a^i_P) = \text{null}$.

Various kinds of indices for visibility can be defined by using $l(a^i_P)$. In this study, we introduce two indices of visibility at $p$; $\nu_p$ which is the ratio of visibility space and $c^i_P$ which is the number of visual axes running into the building whose type is $j$. They are formulated as follows:
The visibility condition of a visible dome is highly dependent on the space. Especially, radius of the dome may much influence on these indices. Parameters are determined in advance by using actual data. They are as follows; $R = 100$, $V = 10$ and $H = 9$.

3.2. Other attributes

Needless to say, the road condition is considered to be an important factor on car-related crimes. Therefore, we introduce the following two attributes; $sd_p$ which represents the Euclidean distance to a main road from $p$, and $tl_p$ which represents the total length of roads within the circle of radius $CR$ (meters) centred at $p$. In our analysis main roads to define $sd_p$ correspond to national roads, prefectural roads, and other roads defined by the Zenrin Housing Map. (Zenrin is one of Japanese major map companies.) $tl_p$ expresses the degree of density of roads. If $tl_p$ is large, the area can be assumed to be intricate. In the experiment $CR$ is set to 100. At the end, we use $lu_p$ which expresses the main land-use around $p$ (see Table 2 for the types of land-use adopted in our analysis).

4. Sample data

The Nishikyo-ku area of Kyoto City is taken up as an example area. This area is located in the west suburb of Kyoto City. The reason why this area is taken up is that it has various spatial and land use characteristics, and our laboratory is in this area. The time period of the car-related crime data used for our data mining analysis is from Oct. 2002 to Jun. 2005.
spectively. As mentioned before, which are represented as \( P \) and \( N \), respectively but they are unified simply to car-related crimes shown in Table 1 are not considered separately but they are unified simply to car-related crimes hereafter.

Figure 3 shows the map of Nishikyo-ku with crime occurring points in the period. Table 1 shows the number of car-related crimes occurred in the area during the time period which were derived from quick estimation. About 60% of car-related crimes occurred in auto parks and the rest occurred in places such as roadsides. Some auto parks have monitor cameras or guards, which exhibits artificially stronger surveillance performance than other spaces. Therefore, we organized two kinds of analyses considering this influence: the one focuses on only auto parks, and the other focuses on the whole area. In this paper we describe the second analysis which aims at the general relationship between car-related crimes and spatial characteristic. The example area shown in Fig. 4 is divided into square grid whose unit size is 30 m. A center of each square is taken up as a sampling point of a grid.

We excluded from experiments the grids whose center point is inside a building. In consequence, 928 crime occurrence points and 12,148 non-crime occurrence points are extracted. In addition, the types of car-related crime occurrence points and 12,148 non-crime occurrence points are extracted. In addition, the types of car-related crimes shown in Table 1 are not considered separately but they are unified simply to car-related crimes hereafter.

Another data used for our analysis are basic 2D vector data containing geometric and building feature lines, 2D land-use mesh data, the vector data of 3D building model, and the housing map containing the usage of buildings. Table 3 shows attributes and their values at an observed point \( p \). These attributes are used for our data as explanatory variables. In order to apply EP analysis, numerical attributes; \( sd_p \), \( tl_p \), \( v_p \), and \( c_{p}^{lu} \) are discretized based on Fayyad and Irani’s MDL method [12]. The target variable (class) is occurrence or non-occurrence of car-related crimes, and as mentioned before, which are represented as \( P \) and \( N \), respectively.

### Table 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( lu_p )</td>
<td>{1: forest or waste land, 2: paddy, 3: farm, 4: developing land, 5: vacancy, 6: industrial area, 7: general low-rise residential area, 8: dense low-rise residential area, 9: middle or high-rise apartment area, 10: commercial and business district, 11: road, 12: park and green space, 13: public facility zone, 14: river and lake}</td>
</tr>
<tr>
<td>( sd_p )</td>
<td>{1: 0–5 meters, 2: 6–30, 3: 31–348, 4: 349–531, 5: 552–}</td>
</tr>
<tr>
<td>( tl_p )</td>
<td>{1: 0–91 meters, 2: 92–}</td>
</tr>
<tr>
<td>( v_p )</td>
<td>{1: 0–0.040, 2: 0.041–0.252, 3: 0.253–0.896, 4: 0.897–0.987, 5: 0.988–1.0}</td>
</tr>
<tr>
<td>( c_{MD}^{lu} )</td>
<td>{1: 0 pieces, 2: 1–27, 3: 28–97, 4: 98–245, 5: 246–291, 6: 292–}</td>
</tr>
<tr>
<td>( c_{CB}^{lu} )</td>
<td>{1: 0 pieces, 2: 1–4, 3: 5–14, 4: 14–}</td>
</tr>
<tr>
<td>( c_{NR}^{lu} )</td>
<td>{1: 0 pieces, 2: 1–2, 3: 3–}</td>
</tr>
<tr>
<td>( c_{MD}^{NW} )</td>
<td>{1: 0 pieces, 2: }</td>
</tr>
<tr>
<td>( c_{CB}^{NW} )</td>
<td>{1: 0 pieces, 2: 1–}</td>
</tr>
<tr>
<td>( c_{NR}^{NW} )</td>
<td>{1: 0–27 pieces, 2: 28–176, 3: 177–282, 4: 283–}</td>
</tr>
</tbody>
</table>

5. Extraction of EPs and interpretation

5.1. One dimensional EPs

In order to understand to what degree each attribute contributes to crime occurrence we have first computed \( Growthrate_{p}(X) \) for \(|X| = 1\). The result is summarized in Table 3. In the column of \( Growthrate_{p}(X) \), the values which are larger than or equal to one are placed on the left while the others are placed on the right, and the values written in boldface are the five largest or the five smallest ones. We now explain the results in more details. In the attribute \( lu_p \) (land-use), \( Growthrate_{p}(X) \) is large for the value 5 (vacancy), 9 (middle or high-rise apartment) or 10 (commercial and business area).

In the attribute \( sd_p \) (i.e. distance to a main road), \( Growthrate_{p}(X) \) takes the largest value when \( sd_p = 2 \) (the distance is relatively small, i.e. from 6 to 30 m).

In the attribute \( tl_p \), \( Growthrate_{p}(X) \) is extremely small for \( tl_p = 1 \) (i.e. the concentration rate of roads is small). In the attribute \( v_p \) (visibility), \( Growthrate_{p}(X) \) is small when \( v_p = 1 \) or 5 while it is high for \( v_p = 2 \) (visibility is relatively bad). In the attribute \( c_{MD}^{lu} \) (i.e. the number of rays which hit detached houses), we do not see any particular characteristics. In the attribute \( c_{CB}^{MD} \) (i.e. the number of rays which hit multi-dwellings), \( Growthrate_{p}(X) \) increases as \( c_{CB}^{MD} \) gets large, and in particular it takes the second largest value when \( c_{CB}^{MD} = 4 \). In the attribute \( c_{CB}^{CB} \) (i.e. the number of rays which hit commercial and business buildings), \( Growthrate_{p}(X) \) increases as \( c_{CB}^{CB} \) gets large, and it take the high values for \( c_{CB}^{CB} = 2 \) or 3. In the attribute \( c_{CB}^{NR} \) (i.e. the number of rays which hit public buildings), \( Growthrate_{p}(X) \) decreases as \( c_{CB}^{NR} \) gets large. In the attribute \( c_{NR}^{NW} \) (i.e. the number of rays which hit non-wall buildings), \( Growthrate_{p}(X) \) increases as \( c_{NR}^{NW} \) becomes large, and
takes the highest value for $c_{p}^{NW} = 3$, but it drops down to zero when $c_{p}^{NW} = 4$.

5.2. Derivation and ranking of EPs based on all possible combinations of attributes

In order to extract attribute patterns which are highly correlated with crime occurrence, we have computed EPs for all possible itemsets $X$ ($|X| \leq 10$) which consist of seven groups. In each group, the first one has the largest itemset which is a superset of the itemset of the other EPs in the same group. Since it is difficult to intuitively understand EPs in the table, we aggregate EPs in each group by extracting the items commonly appearing in EPs of the same group. The result is shown in Table 5. Although $Growthrate_P(X)$ slightly decreases, it captures the characteristics of those in Table 4. As will be explained below, it turns out that these EPs exhibit important features related to crime occurrences. Each pattern can be interpreted as follows.

#1 EPs say that if there is no dwelling house and there are many non-wall buildings nearby, $Growthrate_P(X)$ becomes high. Although the single attribute that there are many non-wall buildings nearby exhibits the highest $Growthrate_P(X)$ among itemsets $X$ with $|X| = 1$, the $Growthrate_P(X)$ further increases when combined with the condition that there is no dwelling house. It can be understood since the existence of a dwelling house can be regarded as one of the factors to improve the natural surveillance.

#2 EPs are the ones obtained from #1 EPs by adding the one such that the condition that the visibility is small. Since the condition of small visibility tends to decrease the natural surveillance, this pattern is reasonable.

#3 EPs are the ones obtained from #1 EPs by adding the condition that there are no public buildings nearby. It implies that the existence of public buildings helps to prevent crime occurrence.

#4 EPs are the ones obtained from #3 EPs by adding the condition that the visibility is small.

#5 EPs are the ones obtained from #1 EPs by adding the condition that the distance to a main road medium-range. Although the condition that the distance to a main road medium-range alone does not exhibit the high risk of crime occurrence, it contributes to the increase of the crime risk when combined with other conditions.

#6 EPs are the ones obtained from #1 EPs by adding the condition that the distance to a main road medium-range. Although the condition that the distance to a main road medium-range alone does not exhibit the high risk of crime occurrence, it contributes to the increase of the crime risk when combined with other conditions.

Table 3

| Variable | Level | $|DP|_1$ | $|DN|_1$ | $Growthrate_P(X)$ |
|----------|-------|---------|---------|-------------------|
| 1        | 41    | 950     |         | 0.565             |
| 2        | 77    | 1,047   |         | 0.963             |
| 3        | 36    | 545     |         | 0.865             |
| 4        | 4     | 271     |         | 0.193             |
| 5        | 125   | 1,131   | 1.447   | 0.882             |
| 6        | 13    | 193     |         | 0.793             |
| 7        | 164   | 2,707   |         | 0.583             |
| 8        | 37    | 568     |         | 0.332             |
| 9        | 149   | 587     |         | 0.226             |
| 10       | 103   | 593     | 2.274   | 0.291             |
| 11       | 123   | 1,720   |         | 0.936             |
| 12       | 11    | 637     |         | 0.226             |
| 13       | 40    | 974     |         | 0.538             |
| 14       | 5     | 225     |         | 0.759             |

Table 4 lists EPs such that $|DP| > |DN|$ is greater than or equal to 10 which consist of seven groups. In each group, the first one has the largest itemset which is a superset of the itemset of the other EPs in the same group. Since it is difficult to intuitively understand EPs in the table, we aggregate EPs in each group by extracting the items commonly appearing in EPs of the same group. The result is shown in Table 5. Although $Growthrate_P(X)$ slightly decreases, it captures the characteristics of those in Table 4. As will be explained below, it turns out that these EPs exhibit important features related to crime occurrences. Each pattern can be interpreted as follows.

#1 EPs say that if there is no dwelling house and there are many non-wall buildings nearby, $Growthrate_P(X)$ becomes high. Although the single attribute that there are many non-wall buildings nearby exhibits the highest $Growthrate_P(X)$ among itemsets $X$ with $|X| = 1$, the $Growthrate_P(X)$ further increases when combined with the condition that there is no dwelling house. It can be understood since the existence of a dwelling house can be regarded as one of the factors to improve the natural surveillance.

#2 EPs are the ones obtained from #1 EPs by adding the one such that the condition that the visibility is small. Since the condition of small visibility tends to decrease the natural surveillance, this pattern is reasonable.

#3 EPs are the ones obtained from #1 EPs by adding the condition that there are no public buildings nearby. It implies that the existence of public buildings helps to prevent crime occurrence.

#4 EPs are the ones obtained from #3 EPs by adding the condition that the visibility is small.

#5 EPs are the ones obtained from #1 EPs by adding the condition that the distance to a main road medium-range. Although the condition that the distance to a main road medium-range alone does not exhibit the high risk of crime occurrence, it contributes to the increase of the crime risk when combined with other conditions.

#6 EPs are the ones obtained from #1 EPs by adding the condition that the distance to a main road medium-range. Although the condition that the distance to a main road medium-range alone does not exhibit the high risk of crime occurrence, it contributes to the increase of the crime risk when combined with other conditions.
Fig. 4. 2D grid on the map.

From these observations, it turns out that the combination of spatial attributes such as the existence of non-wall buildings, the small visibility and the land use for middle or high-rise apartments contributes to the formation of areas with high crime risk.

5.3. The difference of growth-rates in two similar itemsets

Contrary to association rules, the subset of EP with high growth-rate does not always exhibit high growth-rate. Therefore, it may happen that between similar itemsets their growth-rates may greatly differ. If we can find such pair of itemsets, it will help to identify the discovery of risk factors which may often be overlooked since the slight difference of the conditions greatly changes the crime risk. From this viewpoint, the difference of Growthrate_{a}(X) for two itemsets X = A, B and for class a (= P or N) is defined by the ratio of Growthrate_{a}(A) and Growthrate_{a}(B), which is denoted by Difff(A, B, a). Since Growthrate_{a}(X) may take the value of infinity, its definition is slightly changed by replacing the zero value of the support by a very small positive value $10^{-\alpha}$ where $\alpha$ is set to 10. The rigorous definition of is given below where $b(= P \text{ or } N), b \neq a$.

$$
\text{Diff}(A, B, a) = \frac{\text{Growthrate}_{a}(A)}{\text{Growthrate}_{a}(B)} = \text{Growthrate}_{a}(A) \times \text{Growthrate}_{a}(B) \\
\approx \frac{\text{supp}_{D_{a}}(A)'}{\text{supp}_{D_{a}}(A)'} \times \frac{\text{supp}_{D_{a}}(B)'}{\text{supp}_{D_{a}}(B)'}
$$

where

$$
\text{supp}_{D_{a}}(C)' = \begin{cases} 10^{-\alpha}, & \text{if } \text{supp}_{D_{a}}(C) = 0, \\ \text{supp}_{D_{a}}(C), & \text{otherwise}. \end{cases}
$$

We compute $\text{Diff}(A, B, a)$ for a = P and N for various pairs of A and B such that Case 1: $B \subset A$ and $|A - B| = 1$, and Case 2: $|A| = |B|$ and $|A - B| = |B - A| = 1$. The results are summarized by computing DiffMax(A, B) which is denoted by

$$
\text{DiffMax}(A, B) = \max\{\text{Diff}(A, B, P), \text{Diff}(A, B, N)\}.
$$
Table 4
Top 7 EPs whose target class is P. They are ranked in lexicographically descending order of \( \text{Growthrate}_P(X) \) and \( |D_P| \). ‘+’ denotes the attribute used in the corresponding EP.

| # | \( \Delta \) | \( |\Delta| \) | \( \text{Growthrate}_P(X) \) | sub# | EP |
|---|---|---|---|---|---|
| 1 | 16 | 0 | \( \infty \) | 1 | \( h_{2,2} \) |
| 2 | 15 | 0 | \( \infty \) | 1 | \( v_{2,2} \) |
| 3 | 15 | 0 | \( \infty \) | 1 | \( \gamma_{2,2} \) |
| 4 | 14 | 0 | \( \infty \) | 1 | \( l_{9,9} \) |
| 5 | 12 | 0 | \( \infty \) | 1 | \( m_{3,3} \) |
| 6 | 11 | 0 | \( \infty \) | 1 | \( r_{1,1} \) |

5.3.1. Case 1

Itemsets \( A \) and \( B \) of 15 largest DiffMax(\( A, B \)) are given in Table 6. Attributes in the shaded region are added ones. In the pairs #1 through #10 and #13 through #15, the crime risk is high under the condition \( B \). This means that the addition of one extra item dra-
matically decreases the crime risk. Such item appended to $A$ seems to possess crime prevention effect.

Most of such items are $v_{p,5}$; the visibility is highest. For example, in the area corresponding to the itemset $B$ in #1 the road density is high, there are not many dwelling houses or public buildings, and thus the potential surveillance is not high. However, if the visibility is high, the crime risk dramatically decreases.

In the pairs #11 and #12, in the area corresponding to the itemset $B$ the distance to a main road is large, and there are not many dwelling houses or public buildings, and hence it is relatively safe, but when the condition that the visibility is low is added, the crime risk gets high.

In summary, we observed that the natural surveillance is a critical factor for the risk of car-related crimes.

5.3.2. Case 2

Itemsets $A$ and $B$ of 15 largest DiffMax$(A, B)$ are given in Table 7. Attributes in the shaded region differ each other. In each of those pairs the itemset $A$ corresponds to the area which is relatively safe, while $B$
corresponds to the area which is relatively dangerous. In the itemsets of $A$, we found that the item $v_{p,5}$ which means high visibility is most frequently used, followed by the item $lu_{p,5}$ which means the land use is vacant. In the itemsets of $B$, the item $c_{p,1}^{NW}$ which means there are many non-wall buildings is most frequently used, followed by $lu_{p,9}$ which means the land use as middle or high-rise apartment. For instance, the items common to $A$ and $B$ are $lu_{p,9}$ and $c_{p,1}^{DH}$ which says there are many roads and there does not exist dwelling house nearby.

In summary, visibility, and the existence or non-existence of urban activities, inhabitants, and middle or high-rise apartments greatly affects the risk of car-related crimes.

### 6. Concluding remarks

In this paper, car-related crime data of Kyoto-shi Nishikyo-ku was analyzed. By using EP, a lot of interesting patterns were found. Since a large number of EPs were generated, we introduced two ideas for simplifying them; the first was to appropriately aggregate several EPs with high growth-rate in order to increase the readability of the patterns generated, and the second was to identify a pair of patterns $A$ and $B$ such that both consist of the same set of attribute values except one such that while the former pattern $A$ is not associated with high crime occurrences while the pattern $B$ is highly correlated with crime occurrences although $A$ and $B$ are very much similar.

We conclude that the natural surveillance is a critical factor to reduce the risk of car-related crimes. Moreover, visibility, and the existence or non-existence of urban activities, inhabitants and middle or high-rise apartments greatly affects the risk of car-related crimes.

### Acknowledgements

We would like to thank to members of GIS workshop hosted by Crime Analysis Office of Kyoto Prefec-
tural Police, and Katoh laboratory in Kyoto University Faculty of Engineering. This study was supported by Grant-in-Aid for Young Scientists (B) (No.18760460) of Ministry of Education, Culture, Sports, Science and Technology, in Japan, and Grant-in-Aid for Scientific Research (C) (No.17500007) of Japan Society for the Promotion of Science.

References