EMERGING PATTERN BASED STREET CRIME ANALYSIS
Street-level spatial analysis of crime location associated with built environment in Fushimi ward, Kyoto city

Atsushi TAKIZAWA*  

We analyze the relationship between bag snatching and spatial relationships in Kyoto City. In particular, we model natural surveillance with various façade components. We then analyze the data as a classification problem that divides space into areas in which crimes occur and areas in which crimes do not occur. Since there exists fuzziness in the categorization of space, we propose a new approach that combines clustering and classification based on semi-supervised learning, which determines the categorization so as to improve classification accuracy. In addition, we use an emerging pattern based classifier and obtain spatial features related to crime.

Keywords: street crime, bag snatching, natural surveillance, wall component, classification by aggregating emerging patterns, semi-supervised clustering

1. Introduction

Crime prevention through environmental design (CPTED)\textsuperscript{1,2} considers the relationship between crime occurrence and environment and attempts to prevent crime by modifying the environment. CPTED has three primary strategies for existing environments: natural surveillance, natural access control, and natural territorial reinforcement. Among these strategies, preservation of natural surveillance is important. Natural surveillance is a geometric character of space by which the space is naturally within view of people. Desyllas et al.\textsuperscript{3} pointed out that Jacobs\textsuperscript{4} had classified natural surveillance into two types: that by occupants of buildings looking outside onto a street (type-1) and that by pedestrians along the street (type-2). However, since the strategy of CPTED was derived experientially, the strategy should be verified quantitatively using real data to determine its effectiveness with respect to crime prevention.

For example, Hiller\textsuperscript{5} studied the relationship between crimes (burglary and car crime) and the structure of streets using the methodology of space syntax\textsuperscript{6}. The study deals with type-2 natural surveillance. Ishikawa and Suzuki\textsuperscript{7} also examined the type-2 natural surveillance against bag-snatching in Osaka City. On the other hand, Desyllas et al.\textsuperscript{8} modeled type-1 natural surveillance using a visibility graph\textsuperscript{9} and evaluated the visibility of streets from building entrances. Nes and López\textsuperscript{10} also investigated inter-visibility of doors between multiple buildings and their inter-relationship with street segments. The main consideration in the previous studies was the entrances; the effects of other components, such as windows, have not yet been given sufficient consideration. In addition, the buildings in the studied area have various uses, such as shops, residences, and factories. We pointed out in a previous study\textsuperscript{11} the existence of a correlation between building usage and street crime. Therefore, the data for analysis tends to become large. In previous studies, the databases that have a relatively small number of variables were analyzed by simple analysis methods, such as correlation analysis or linear regression. However, such simple methods appear to be insufficient for treating large databases.

Based on the above considerations, in the present study, we analyze the relationship between a number of spatial attributes and bag snatching incidents that occurred in Kyoto city in Japan. As spatial attributes, we primarily model type-1 surveillance with various façade components (not only entrances, but also windows and shutters) as “wall visibility”, and the number of pedestrians on streets by simple random walks. We attempt to clarify the relationship between crime and street-level spatial characteristics as a classification problem using a data-mining method for large database analysis. Here, the purpose of classification is to classify each actual crime location (ACL) and sampling point (see Section 4.1) into two classes: (actual or) potential crime location (PCL) and crime-free location (CFL). The distinction between an actual crime location and a potential crime location stems from the question as to whether we may classify only actual crime locations as crime locations. Criminals involved in bag snatching
select victims from a distance, shadow the victim, and commit the crime when an opportunity arises. The timing of the offense may be influenced by several incidental factors, such as the presence of witnesses and the difference of moving speed between the criminal and victim. Therefore, in the present study, we also regard the neighboring points of actual crime locations as crime locations and decide their area as a type of clustering problem. This clustering problem is solved in order to maximize the classification accuracy. In a broad sense, such clustering with additional information is referred to as supervised clustering. In particular, the method of handling unlabeled data is referred to as semi-supervised clustering. In the analysis method of the present study, unlabeled data corresponds to the neighboring area of each actual crime location. Thus, the clustering method can be categorized as a type of semi-supervised clustering.

Although data-mining analysis for criminal data has been conducted previously, the primary purpose of this previous analysis appears to be to analyze large investigative databases maintained by the police and to offer criminal-investigation support to the police; however, the previous analysis does not sufficiently explore the relationship between space and crime. Data-mining methods applied to spatial data are referred to as spatial data-mining methods. Although the basic methodology of them follows the general framework of data mining, some spatial data-mining methods are extended to deal with spatial problems such as spatial association rules. In the present study, we use a general data-mining method of classifiers and add a semi-supervised clustering to deal with the spatial problem of the present study.

Various classification models have been proposed in data-mining communities. Among them, we apply classification by aggregating emerging patterns (CAEP). Classification by aggregating emerging patterns performs classification using emerging patterns (EPs), which tend to appear in a specific class in frequent itemsets. Emerging patterns and CAEP are categorized by the method of contrast data mining, which is the mining of patterns and models contrasting two or more classes/conditions. In particular, its ability to classify rare events is important when analyzing criminal data in Japan, where the frequency of criminal incidents is lower compared with other countries. CAEP can classify such imbalanced data with high accuracy and output useful classification rules as EPs. We previously applied these methods to analyze car-related crimes and bag snatchings in Kyoto City. The latter study is our most recent previous study. The primary differences from that previous study are that we herein consider the wall visibility using wall components, the use of a random walk model for pedestrian estimation, and a different analysis method. Nakaya and Yano extended kernel density estimation to visualize the spatiotemporal distribution of street crime points in Kyoto City. Satoh and Okabe also extended kernel density estimation to deal with line and polygon segments and applied the extended kernel density estimation to analyze the distribution of street crimes on a road network in Kyoto City.

The remainder of the present paper is organized as follows. In the next section, we explain in detail the studied area and the database used in this article. Section 3 explains the attributes used for these analyses, and Section 4 presents the analysis method. In Section 5, we explain how the analysis is carried out and discuss some findings. Section 6 explains the results of the detailed analysis. Section 7 concludes the article.

2. Studied area and databases

According to Statistics Kyoto City, the number of non-home-invasion crime including bag snatching in Kyoto City in 2005 was 13,850, 2,829 of them occurred in Fushimi Ward and this is the largest number in all wards of Kyoto City. Fushimi Ward is located in the suburbs of Kyoto City and its street network is rather irregular unlike that of the city center. In addition, land uses there are more mixed than those of city center. From these reasons, we analyze Fushimi Ward. Table 1 lists the databases used in the present study. Database #1 includes data on bag snatching incidents. This data is not latest, but we think it precise and valuable since gotten from Kyoto Prefecture Police. From January 2004 to December 2005, 342 bag snatching incidents were recorded in the Fushimi Ward. The mean Euclidian distance from those crime locations to their nearest stations is about 660 meters. On the other hand, the mean Euclidian distance from 723 gravity center points of small areas in population census data of Fushimi Ward (see Database #5 in Table 1) to their nearest stations is about 991 meters. Therefore, it can be said that the bag snatching incidents in this data tend to occur around stations.

Since we consider wall components in Section 3.1 and measuring them takes a lot of time and costs, we limit the studied area as shown in Fig.1. The area studied herein is a rectangle of approximately 0.8 km by 1 km and includes three train stations. We consider four more stations when the number of pedestrians is estimated by a random walk, as described in Section 3.2. The east and south sides of the area are divided with rail ways. Shintakase River is flowing along the west side boundary of the area. The prefecture road No.35, which is a relatively wide road in this region, runs...
The gradation shown in gray represents the kernel density estimation of bag snatching. 57 of bag snatching incidents took place in the studied area. Five occurred in building areas and are omitted from the present analysis. Where the density of bag snatching incidents in whole Fushimi Ward is about 7.1/km², that of the studied area is increased to 57.7/km². Whereas the density of bag snatching incidents in whole Fushimi Ward is about 7.1/km², that of the studied area is increased to 57.7/km².

The author thinks the degree of the natural surveillance will be primarily affected by the distance between a criminal and observers. In the case of type-1 natural surveillance, we also have to consider the area of walls and wall components. This is indexed by a visible angle from a viewpoint. Furthermore, the effect of distance and area can be merged into a single index of an angle. However, we cannot distinguish both effects if the gradation shown in gray represents the kernel density estimation of bag snatching. 57 of bag snatching incidents took place in the studied area. Five occurred in building areas and are omitted from the present analysis. Where the density of bag snatching incidents in whole Fushimi Ward is about 7.1/km², that of the studied area is increased to 57.7/km².

The illuminance on the street is thought to have an important impact on crime. We measured the illuminance from 8 p.m. on September 9, 2008, to 3 a.m. the following day. The author wore a helmet with an illuminance meter and a GPS logger and traveled along the street on a motorbike. The illuminance meter was approximately 1.5 m above the ground. The illuminance measurements from the meter and the GPS positions from the logger were subsequently matched by time. The GPS data sometimes had considerable errors; outliers were replaced by reasonable values using an appropriate interpolation method. Then, illuminance data was then created (Eq. 8). The length of the studied street is approximately 20 km. In order to reduce the difference in the illumination caused by the difference in measurement time for each area, the route was planned not to measure a specific area at a time but to go around a wide range. In addition, we do not perform any temporal adjustment of the illuminance.

Table 2 lists attributes of 57 victims on incidents. These are summarized from Database #1. Females in the gender, 20s, 30s, and 40s in the age, and workers in the job occupy the majority of victims. Bag snatching incidents tend to occur from evening to early morning. Many victims were in the middle of commuting (or going home) when they encountered crimes.

### Table 1 Databases used in the present study.

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Source</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bag snatching data in Fushimi Ward, Kyoto City, from January 2004 to December 2005.</td>
<td>Kyoto Prefectural Police.</td>
<td>Point data of actual crime locations with a high degree of accuracy.</td>
</tr>
<tr>
<td>5</td>
<td>Population Census, Kyoto, 2005.</td>
<td>Ministry of Internal Affairs and Communications, Japan.</td>
<td>Population census data tabulated by small areas as subdivision of municipalities</td>
</tr>
<tr>
<td>7</td>
<td>Location View, Kyoto, 2006.</td>
<td>Asia Air Survey Co., Ltd.</td>
<td>Omnidirectional images of the street.</td>
</tr>
</tbody>
</table>

### Table 2 Attributes of 57 victims on incidents.

<table>
<thead>
<tr>
<th>Category</th>
<th>Item (Number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male (10), Female (47)</td>
</tr>
<tr>
<td>Age</td>
<td>10-19 (3), 20-29 (14), 30-39 (3), 40-49 (5), 50-59 (15), 60-69 (14), 70-79 (2), 80-89 (1)</td>
</tr>
<tr>
<td>Job</td>
<td>Worker (33), Inoccupation (11), Student (7), Homemaker (6)</td>
</tr>
<tr>
<td>Time</td>
<td>0-3 (9), 4-7 (10), 8-11(2), 12-15 (3), 16-19(15), 20-23(19)</td>
</tr>
<tr>
<td>Behavior</td>
<td>Commuting (23), Commuting to school (2), Shopping (7), Sightseeing(1), Working (1), Jogging (1), Other (22)</td>
</tr>
</tbody>
</table>

3. Attributes

### 3.1 Wall visibility

Space visibility is a central concept of natural surveillance. The visibility analysis proposed herein investigates the openness of a two-dimensional plane and can be executed at any point outside building polygons. There are two methods for measuring space visibility: Isovista (22), which is a simple approximate method, and the visibility-graph-based method, which, although more precise, is rather complex. Since we consider the effect of wall components, we need a method that precisely detects the building walls (stored in database #4); see Fig.2.

This is a simple visibility-graph-based method that detects both endpoints of a line segment \( w \in W_p \) that indicates the surface of a wall within a radius of \( r \) m from viewpoint \( v \). Here, \( W_p \) represents the set of line segments that is visible from \( v \), and the radius \( r \) limits the visible range. We set \( r = 40 \)m. Yoshida et al. (20) concluded that the distance of outer space at which a dweller in a house with transparent windowpanes feels that his/her privacy is breached is approximately 30 m. Following this result, Tanaka et al. (26) evaluated the degree of natural surveillance of a street using laser scanner data. While, since a pedestrian along a street can foresee far more, we add more 10 m hypothetically to Yoshida’s conclusion for calculating the wall visibility. When neither of the corner points can be seen, the original line segment is divided according to radius \( r \). After detecting a visible line segment \( w \), we calculate the Euclidean distance to the line segment \( d(v,w) \), the angle of spread of the line segment \( \theta(v,w) \), and the length of the visible line segment \( l(v,w) \) as basic geometric information for the viewpoint and the visible wall (see Fig.3).

The author thinks the degree of the natural surveillance will be primarily affected by the distance between a criminal and observers. In the case of type-1 natural surveillance, we also have to consider the area of walls and wall components. This is indexed by a visible angle from a viewpoint. Furthermore, the effect of distance and area can be merged into a single index of an angle. However, we cannot distinguish both effects if...
they are merged. The effect of a near small wall (component) should be weighted more heavily than that of a far large wall (component).

From these reasons, we consider the effect of the distance and the angle separately when quantifying the effect of the wall visibility. The value of the wall visibility of a wall $w$ from a viewpoint $v$ is given as follows:

$$ns(v, w) = \frac{\theta(v, w)}{\pi} \frac{r - d(v, w)}{r}. \quad (1)$$

Roughly speaking, this formula firstly calculates the distance effect of components, where $r$ is the distance from the viewpoint to the wall, and normalized it. This formula returns a larger value when the angle, and normalized it. This formula returns a larger value when the distance is smaller and the spread of the wall is larger.

Next, we consider the effect of wall components. We consider seven components $k \in WC$ listed in Table 4. The area of the components of all walls that are visible from the studied roads was measured manually using the 360° street image data of database #7 and the measurement function of the image-viewer software, LV Local Viewer by Asia Air Survey Co., Ltd. Since measuring the wall components of higher floors was difficult, only those of the first floor were measured. Then, the total area of each component $k$ in wall $w$ is $ac(w, k)$.

In the present study, the effect of a wall component on natural surveillance is the ratio of the total area of the wall components to the area of the wall for the first floor. The height of the first floor for all walls is $h$ m. We set $h = 3.5$ m. The wall visibility $ns_c(v, w, k)$ that $w$ gives to $v$ given wall component $k$ is

$$ns_c(v, w, k) = \frac{ac(w, k)}{\text{area}(v, w)} \cdot ns(v, w). \quad (2)$$

In addition, two different building categories are defined in database #4. We combine them as listed in Table 4.

We now define some indices of type-1 natural surveillance. The first is a simple index that is based on only the geometric features of walls:

$$w1_l(v) = \sum_{w \in WC} ns(v, w). \quad (3)$$

The second index is based on the specific building usage. Let $bt \in Btype$ denote a building category listed in Table 4 and $W_b(bt) \subset WC$ denote the set containing the line segments that are members of $bt$. Then,

$$w2_{bt}(v) = \sum_{w \in W_b(bt)} ns(v, w). \quad (4)$$

The third index is based on wall components:

$$w3_k(v) = \sum_{w \in WC} ns_c(v, w, k). \quad (5)$$

The fourth and final index is based on both building usage and wall components:

$$w4_{bt, k}(v) = \sum_{w \in W_b(bt)} ns_c(v, w, k). \quad (6)$$

We take into account not only wall visibility but also the openness of the space. Let $non\text{-}vis(v)$ denote the area that is not visible from viewpoint $v$ within radius $r$ (see Fig.4). This is a kind of type-2 natural surveillance indices.

### 3.2 Number of pedestrians on the street

Since most bag snatching incidents occur on the street, it is important to know the pedestrian flow. However, the pedestrian flow is difficult to determine unless a large-scale investment is conducted because several complicated factors affect pedestrian movement. However, Li showed that the accessibility of streets can be roughly estimated by the random-walk 25). As mentioned in Section 2, bag snatching incidents in Fushimi Ward tend to occur near stations. There exist some stations inside/outside the area studied, and the density of bag-snatching of this area is relatively high in the whole Fushimi Ward. Besides, we found from Table 2 that bag snatching incidents in Fushimi Ward often occurred as an individual was commuting (returning home from work). Since many people are thought to use the train to commute, the paths from the station to their homes appears to be used more frequently than other paths. Based on this assumption, we attempt to model the pedestrian flow from each station and estimate the number of pedestrians on the street by using the random-walk simulation. The estimation process is described below.

First, the central line of the road that runs inside and outside the studied area is subdivided every $rwp$ m, which is a length parameter of subdivision, to create a road network for the random-walk simulation, in which the vertices are represented as $v \in RV$, where $RV$ denotes the set of vertices. The movement range from a station is set to $rwp_{max}$ km by the shortest distance of the network. The seven stations shown in Fig.1 that are within the movement range are considered as starting points for a pedestrian agent.

We set $rwp = 30$ m and $rwp_{max} = 1.2$ km. The seven stations listed in Table are denoted as $st \in ST$. Initially, a pedestrian agent is placed at one of the stations. The agent is then moved randomly to a new vertex connected to the current vertex. The random walk is repeated until the agent moves beyond the movement range of $st$. This procedure constitutes one trial. Let $rwp_{st}(rv)$ denote the cumulative number of pedestrians reaching vertex $rv$ from station $st$. Whenever an agent from $st$ passes $rv$, $rwp_{st}(rv)$ is incremented. The trial is repeated $rwp_{max}$ times, and the average $rwp_{st}(rv)$ of the trials is the estimated number of pedestrians reaching $rv$ from $st$, which is denoted as $rwp_{st}(rv)$. We set $rwp_{max} = 100,000$.

We also define the expected value of pedestrians at node $n$ from the seven stations by considering the number of passengers using each station per day. The reason why we consider the number of passengers is that
The area north of Keihan Chushojima Station contains several
restaurants and bars. Furthermore, there is a long shopping mall to the west
of Kintetsu Momoyamagoryomae Station. Since numerous pedestrians use
these streets, bag snatching might occur more frequently in this vicinity.
Therefore, the density of stores and bars might be a good indicator of
criminal activity. Using information from database #6, we divide the

The flow of the analysis method is as follows. In Section 4.1, the spatial
unit is defined, and the analysis data is created. In Section 4.2, a clustering
method is proposed and applied in order to divide the data into two classes.
A classification is performed using this data. The details of the classifier
used in the present study are shown in Section 4.3. Section 4.4 explains the
classification accuracy of the classifier. The accuracy is used to optimize
the clustered area, as explained in Section 4.5. A series of these procedures

*1 Kyoto Prefecture statistics
http://www.pref.kyoto.jp/tokei/yearly/tokeisyosiyo/tokeisyotop.html
*2 Kintetsu Corporation website
http://www.kintetsu.jp/kouhou/corporation/koutsu/i.html

\[ rw_{st}(rv) = \text{the result of the random walk with one passenger.} \]

Table 3 Number of passengers using various stations per day.

<table>
<thead>
<tr>
<th>Station</th>
<th>Number of passengers per day (year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keihan Kangetsukyo</td>
<td>2,890 (2006)*1</td>
</tr>
<tr>
<td>Keihan Tambabashi</td>
<td>29,145 (2006)*1</td>
</tr>
<tr>
<td>Keihan Chushojima</td>
<td>6,134 (2006)*1</td>
</tr>
<tr>
<td>Kintetsu Pushimimoyama</td>
<td>5,836 (2006)*1</td>
</tr>
<tr>
<td>Kintetsu Tambabashi</td>
<td>25,568 (2008)*2</td>
</tr>
<tr>
<td>Kintetsu Momoyamagoryomae</td>
<td>7,634 (2008)*2</td>
</tr>
<tr>
<td>JR Momoyama</td>
<td>1,732 (2006)*1</td>
</tr>
</tbody>
</table>

*1 Kyoto Prefecture statistics
http://www.pref.kyoto.jp/tokei/yearly/tokeisyosiyo/tokeisyotop.html
*2 Kintetsu Corporation website
http://www.kintetsu.jp/kouhou/corporation/koutsu/i.html

\[ rw_{TTL}(rv) = \sum_{st \in ST(rv)} ps(st) \cdot rw_{st}(rv). \quad (7) \]

Table 4 Symbols used in the present study.

<table>
<thead>
<tr>
<th>Category (name of the set)</th>
<th>Element (symbol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building type (Btype)</td>
<td>General building (gb), landmark building (lb), non-wall building (rw), public building (pb), apartment building (ab), individual house (ih), business institution (bi), other building (ob)</td>
</tr>
<tr>
<td>Wall component (WC)</td>
<td>Door (dr), normal window (wn), grilled window (wg), Kyoto-style grilled window (wk), window with a shutter (ws), column or non-wall façade (cn), shutter (sh)</td>
</tr>
<tr>
<td>Station (ST)</td>
<td>Keihan Fushimimoyama (KFM), Keihan Chushojima (KCJ), Keihan Tambabashi (KTB), Keihan Kangetsukyo (KKG), Kintetsu Momoyamagoryomae (CMG), Kintetsu Tambabashi (CTB), JR Momoyama (MY)</td>
</tr>
<tr>
<td>Land use (LU)</td>
<td>See the legend of Fig.5</td>
</tr>
</tbody>
</table>

Table 5 Attributes used in the present study.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>w1</td>
<td>Type-1 natural surveillance index considering only wall area</td>
</tr>
<tr>
<td>w2, k</td>
<td>Type-1 natural surveillance index of k ∈ Btype</td>
</tr>
<tr>
<td>w3, k</td>
<td>Type-1 natural surveillance index of k ∈ WC</td>
</tr>
<tr>
<td>w4, k</td>
<td>Type-1 natural surveillance index of k ∈ Btype and k ∈ WC</td>
</tr>
<tr>
<td>non_vis</td>
<td>Non visible area within the radius r (type-2 natural surveillance)</td>
</tr>
<tr>
<td>rw_st</td>
<td>Estimated number of pedestrians from st ∈ ST</td>
</tr>
<tr>
<td>rw_TTL</td>
<td>Total estimated number of pedestrians from all stations</td>
</tr>
<tr>
<td>sp_pr_rd</td>
<td>Number of restaurants and stores per length of the road</td>
</tr>
<tr>
<td>cl_pr_rd</td>
<td>Number of clerks per length of the road</td>
</tr>
<tr>
<td>dist_nst</td>
<td>Distance from the nearest station.</td>
</tr>
<tr>
<td>pop_dns</td>
<td>Population density</td>
</tr>
<tr>
<td>illum</td>
<td>Street illuminance</td>
</tr>
<tr>
<td>lu_x</td>
<td>Land use x ∈ LU</td>
</tr>
</tbody>
</table>

Fig.5 Land-use of the area studied from Database #3.
the sum of the distances of each attribute between records. On the other hand, the purpose of the present clustering is to decide the area of potential crime location around actual crime location. This implies that the area should be clustered to the range by which a classification model can classify the crime location with the highest accuracy. However, such significant attributes for good classification cannot be determined before classification. This makes the problem complex and difficult to solve. Therefore, we attempt to find the best results though a trial-and-error process.

4.1 Data preparation

The unit of analysis is a point. In other words, points are sampled discretely, as in the previous study. Since pedestrians tend to walk on the edge of the road, the sampling points are placed 1 m inward from the boundary of the road (see Fig.6). The interval of sampling points on the road is 10 m. We set the interval at 10 m. In this way, 2,769 sampling points are generated. Each sampling point has attributes described in Section 3. Indices in 3.1 are calculated just at the sampling point, and the value of the nearest observation point from the sampling point is assigned for other attributes. We discretize the values of the numeric attributes of each sampling point to apply CAEP to the datasets. We divide each attribute into three subintervals so that the number of records in each subinterval is approximately the same. The discretized attribute is expressed as “attribute name = (-inf, value1], [value1, value2], [value3, inf)” in increasing order, where “inf” represents infinity.

4.2 Clustering method

Let \( c \in \{ c_1, c_2, \ldots, c_{c_3} \} = C \) denote each actual crime location, and let \( s \in S \) denote the sampling points, where \( | \cdot | \) denotes the number of elements of a set. We define class labels P and N which represent (potential) crime location and crime-free location, respectively. Class label P is assigned to all \( c \), and either P or N is assigned to each \( s \). The method by which class label P is assigned to \( s \) around \( C \) is described below.

Sampling points that are visible from \( c \) and are less than \( t_{max�i} \) in Euclidean distance from \( c \) are selected as candidate points for class P clustered by \( c \) (constraint I). Sampling points within a circle of \( t_{min�i} \) radius from \( c \) are always labeled P (constraint II). Let \( s^{(0)} \in \{ s_1^{(0)}, s_2^{(0)}, \ldots, s_{c_3}^{(0)} \} = S_c \subset S \) denote the candidate points for class P sorted in ascending order of Euclidean distance from \( c \). In \( S_c \), the candidate points from \( s_1^{(0)} \) to \( s_{u_{c_i}}^{(0)} \) are labeled P, where \( u_c \) represents the upper index of \( s^{(0)} \). The range of \( u_c \) is \( 0 \leq u_{min_c} \leq u_c \leq |S_c| \), where \( u_{min_c} \) represents the minimal index of \( s^{(0)} \) that satisfies constraint II. If the

Table 6 Confusion matrix of the two-class classification problem.

<table>
<thead>
<tr>
<th>Actual class</th>
<th>Classified into</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>TP</td>
</tr>
<tr>
<td>N</td>
<td>FN</td>
</tr>
</tbody>
</table>

The unit of analysis is a point. In other words, points are sampled discretely, as in the previous study. Since pedestrians tend to walk on the edge of the road, the sampling points are placed 1 m inward from the boundary of the road (see Fig.6). The interval of sampling points on the road is 10 m. We set the interval at 10 m. In this way, 2,769 sampling points are generated. Each sampling point has attributes described in Section 3. Indices in 3.1 are calculated just at the sampling point, and the value of the nearest observation point from the sampling point is assigned for other attributes. We discretize the values of the numeric attributes of each sampling point to apply CAEP to the datasets. We divide each attribute into three subintervals so that the number of records in each subinterval is approximately the same. The discretized attribute is expressed as “attribute name = (-inf, value1], [value1, value2], [value3, inf)” in increasing order, where “inf” represents infinity.

4.2 Clustering method

Let \( c \in \{ c_1, c_2, \ldots, c_{c_3} \} = C \) denote each actual crime location, and let \( s \in S \) denote the sampling points, where \( | \cdot | \) denotes the number of elements of a set. We define class labels P and N which represent (potential) crime location and crime-free location, respectively. Class label P is assigned to all \( c \), and either P or N is assigned to each \( s \). The method by which class label P is assigned to \( s \) around \( C \) is described below.

Sampling points that are visible from \( c \) and are less than \( t_{max�i} \) in Euclidean distance from \( c \) are selected as candidate points for class P clustered by \( c \) (constraint I). Sampling points within a circle of \( t_{min�i} \) radius from \( c \) are always labeled P (constraint II). Let \( s^{(0)} \in \{ s_1^{(0)}, s_2^{(0)}, \ldots, s_{c_3}^{(0)} \} = S_c \subset S \) denote the candidate points for class P sorted in ascending order of Euclidean distance from \( c \). In \( S_c \), the candidate points from \( s_1^{(0)} \) to \( s_{u_{c_i}}^{(0)} \) are labeled P, where \( u_c \) represents the upper index of \( s^{(0)} \). The range of \( u_c \) is \( 0 \leq u_{min_c} \leq u_c \leq |S_c| \), where \( u_{min_c} \) represents the minimal index of \( s^{(0)} \) that satisfies constraint II. If the distance from the nearest sampling point exceeds \( t_{min�i} \), then \( u_{min_c} = 0 \). We let the vector of the upper indices be \( u = (u_{c_1}, u_{c_2}, \ldots, u_{c_{c_3}}) \). The original dataset \( D \) is divided into two datasets as \( D^{(0)} = D_p^{(0)} \cup D_n^{(0)} \), where \( D_p^{(0)} \) and \( D_n^{(0)} \) represent the datasets of classes P and N divided by the upper indices \( u \). In addition, a sampling point that is clustered as P by more than two actual crime locations is copied in each case in order to weight the record of the sampling point as a place of frequent criminal activity. Fig.7 illustrates the concept of the clustering method. The evaluation major of a clustering result is defined later herein.

4.3 Classification method

As described in Section 1, we use CAEP\(^{26}\), which internally uses EPI\(^{26}\) as a classifier. The brief summary of them are described in our article\(^{18}\). In our CAEP implementation, an apriori algorithm\(^{26}\) is used to extract itemsets for which the support is at least \( min\_sup \), and the maximum number of items contained in an itemset is limited to \( max\_dim \). Emerging patterns having a growth rate of at least \( min\_gr \) are extracted from these itemsets. Here, \( min\_gr \) represents the minimum growth-rate threshold, which determines whether an EP is used for classification.

4.4 Classification accuracy

We now discuss the classification accuracy. In the general statistical approach like logistic regression, a likelihood-based measure by multinomial distribution is used for model evaluation. Since such a likelihood-based measure evaluates the overall accuracy rate like equation (8) described later, if the datasets for the classes are very different in size, the accuracy of the minor class could not be improved sufficiently. Our dataset exactly corresponds to this type since criminal activity is limited to a relatively small area. Moreover, the motivation for using a likelihood-based major is to estimate the optimum model parameters numerically. However, CAEP is a pattern-based classifier and does not estimate model parameters unlike other classifiers.

Therefore, we use another simple method with a confusion matrix for model evaluation. Table 6 lists a confusion matrix of the two-class classification problem. A confusion matrix is used for visualization of the performance of a classifier in machine learning. Each cell gives the number
of classified records according to the actual class and the predicted class. For example, \( TP \) is the number of records for which the actual and predicted classes are both P. The “Not classified” column represents the number of records that are not classified into either class. In most general classifiers, every record is classified into one of the classes. However, we use CAEP, and occasionally, when no EP is found in any dataset, the record cannot be classified into any class.

The general overall classification accuracy of dataset \( D^{(a)} \) is

\[
    \text{precision}(D^{(a)}) = \frac{TP + TN}{|D^{(a)}|} \tag{8}
\]

However, this index is not appropriate for the smaller class if the datasets for the two classes are very different in size. It is preferable to consider the accuracies of the two classes separately. Let \( TPrate(D^{(a)}) \) and \( TNrate(D^{(a)}) \) denote the accuracies of classes P and N, respectively, of dataset \( D^{(a)}:\)

\[
    TPrate(D^{(a)}) = \frac{TP}{|D^{(N)}|}, \quad TNrate(D^{(a)}) = \frac{TN}{|D^{(N)}|} \tag{9}
\]

The \( m \)-fold cross-validation is used to estimate the practical accuracy of a classifier. Here, multiple \( m \)-fold cross-validations with different combinations of training and test datasets are repeated \( m \) times. This provides more realistic results than can be obtained by a single \( m \)-fold cross-validation. Let \( D^{(a)}(r) \) denote the datasets for the \( m \)-fold cross-validation generated from \( D^{(a)} \) by a series of random numbers \( r \in \{r_1, r_2, \ldots, r_{tm}\} = R \). The evaluation of the classification accuracy for \( D^{(a)}(r) \) is

\[
    ac(D^{(a)}(r)) = \min(TPrate(D^{(a)}(r)), TNrate(D^{(a)}(r))) \tag{10}
\]

The final classification accuracy for all \( r \in R \) is

\[
    fac(D^{(a)}, R) = \min(ac(D^{(a)}(r_1)), \ldots, ac(D^{(a)}(r_{tm}))) \tag{11}
\]

In other words, the minimum \( ac(D^{(a)}(r)) \) for all \( r \in R \) is the final classification accuracy, and the value of the above equation is used in order to evaluate clustering. The clustering problem is to find the \( u \) which maximizes equation (11). Since this framework is general, any classifier and many optimization solvers can be applied. The implementation used in the present study is described in the following.

4.5 Optimization method

Since this clustering problem is a combinatorial optimization one and it is difficult to find the exact solution, we use meta-heuristics to approximately solve the problem. Specifically, we use differential evolution (DE)\(^{(27)}\), which is a simple evolutionary algorithm that performs well on many optimization problems. Differential evolution is intended for multidimensional real-valued function optimization but can be used to solve discrete problems if the solution variables are rounded to integers. There are various forms of DE, and we use the Joker method\(^{(28)}\), which can find precise solutions to a range of problems. Let \( x = (x_1, x_2, \ldots, x_n) \) be the real-valued solution vector (found by DE) of an optimization problem. Here, \( n \) is the dimension of the vector. In order to apply DE to the clustering problem, the upper-limit vector \( u \) described in Section 4.2 is found from \( x \). Since each element of \( u \) is an integer, \( u \) is found by rounding the values of \( x \). Moreover, given constraints \( I \) and \( II \) on \( u \), the range of each element of \( x \) is

\[
    \min u_i \leq x_i \leq [c^{(a)}] + 0.999999. \tag{12}
\]

\begin{table}[h]
\centering
\caption{Size of records.}
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
\( opt_u \) & ud=5 m & 10 m & 20 m & 30 m & 40 m & 50 m \\
\hline
\( P \) & 505 & 118 & 214 & 430 & 630 & 829 & 1,011 \\
\hline
\( N \) & 2,380 & 2,704 & 2,619 & 2,432 & 2,274 & 2,130 & 2,019 \\
\hline
\hline
Total & 2,885 & 2,822 & 2,833 & 2,862 & 2,904 & 2,959 & 3,030 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Classification accuracy of CAEP.}
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
\( opt_u \) & ud=5 m & 10 m & 20 m & 30 m & 40 m & 50 m \\
\hline
\( TPrate \) & 0.857 & 0.534 & 0.743 & 0.767 & 0.808 & 0.820 & 0.809 \\
\hline
\( TNrate \) & 0.857 & 0.921 & 0.858 & 0.83 & 0.813 & 0.825 & 0.829 \\
\hline
\( Precision \) & 0.857 & 0.905 & 0.849 & 0.821 & 0.812 & 0.824 & 0.822 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Classification accuracy of logistic regression and LADTree.}
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
\( Logistic \) regression & \( opt_u \) & ud=40 m \\
\hline
\( TPrate \) & 0.752 & 0.754 & 0.848 & 0.793 \\
\hline
\( TNrate \) & 0.795 & 0.711 & 0.756 & 0.686 \\
\hline
\( Precision \) & 0.788 & 0.723 & 0.772 & 0.716 \\
\hline
\end{tabular}
\end{table}

In this formula, the value 0.9999999 is a numerical approximation of the open interval of the upper limit for the rounding of \( x \).

5. Solving the clustering problem

5.1 Setup

The parameter values for clustering and classification are as follows:

\( n = 52 \), \( t_{\text{min, dist}} = 3 \text{ m} \), \( t_{\text{max, dist}} = 50 \text{ m} \), \( t_n = 10 \), and \( t_m = 3 \). The CAEP parameter values are as follows: \( t_{\text{min, gr}} = 3 \), \( t_{\text{min, sup}} = 0.01 \), and \( \text{max, dim} = 2 \). A higher maximum dimension improves the classification accuracy. However, as the dimension increases, the calculation time increases exponentially, and we therefore set the maximum dimension to 2. The DE parameter values are as follows: convergence parameter \( F = 0.75 \), crossover rate \( CR = 0.9 \), population size \( = 500 \), and maximum generation \( = 40 \).

5.2 Results

The classification accuracy (that is \( fac(D^{(a)}, R) \) ) in the initial generation was approximately 0.788 and increased to 0.857. Differential evolution improved the classification accuracy. In the following analysis, we denote the optimal \( u \) as \( opt_u \) and analyze the results of dataset \( D^{(opt, u)} \).

We now compare the classification accuracy of \( opt_u \) with that of datasets derived by changing the value of each element of \( u \) uniformly rather than individually. In other words, \( u_1 = u_2 = \ldots = u_n = u_d \). where \( u_d \) is a distance parameter. We gradually increase \( u_d \) from 5 m to 50 m, as shown in Tables 7 and 8. Table 7 lists the size of the records in each class and the total. As \( u_d \) increases, the number of P records also increases. The total number increases as \( u_d \) increases because the record of a sampling point clustered as P by more than two actual crime locations is copied in each case. Table 8 lists the classification accuracy of \( t_n \)-fold cross-validation. Except for the \( opt_u \) result, the maximum dimension of the EPs in CAEP is 3. When \( u_d \) is small, since the record for class P is so small, the accuracy of P is low. When \( u_d \) increases to 40 m, all accuracies are high. However, \( opt_u \) has the best accuracy among all of the results.
LADTree\textsuperscript{29} which is a recent variant of decision trees with the data of \textit{opt\_u} and \textit{u}=40m by using data mining software; Weka 3.6.8 (http://www.cs.waikato.ac.nz/ml/weka/). The cost sensitive classification technique, which the miss-classification cost of Class P is \(\frac{|D_P^{(0)}|}{|D_N^{(0)}|}\) times larger than that of Class N, is applied to both classifiers. Table 9 lists their classification accuracies. Compared with Tables 8 and 9, we can find that CAEP exhibits the best accuracy among all of them.

6. Detailed analysis of the \textit{opt\_u} results

We analyze the results of \textit{opt\_u} in detail. In order to simplify the analysis, we do not consider the classification results of the \textit{tn}(=10)-fold cross-validation, but rather the results for all of the training data. In this case, \(TP\text{rate} = 0.883\), \(TN\text{rate} = 0.879\), and \(Precision = 0.879\).

Fig.8 shows the classification results for each sampling point and actual crime location. Large circles represent the areas of the potential crime location for each actual crime location. Although a circle does not represent an area on network space exactly, we use the circle expression in the present study since a circle is visually intelligible and it can be drawn easily by the function of GIS. The circles vary in size. The actual crime locations with a small circle tend not to be classified correctly. This is due to the optimization results of \textit{u} such that the upper limit of the clustering area of an actual crime location for which the classification accuracy with clustered sampling points was not so high contracted as little as possible. The sampling points in larger circles are usually correctly classified as class P. The difference in the size of each circle indicates the extent in which spatial attributes on the criminal occurrence resemble each other. This can be useful information when determining the areas where some crime prevention measures are enforced.

Gray circles, which represent misclassification, are conspicuous in the central area, where there are several actual crime locations. Since the spatial features of neighborhoods tend to be similar, it is difficult to completely eliminate these misclassifications. Furthermore, the classification of several actual crime locations located to the far west of Fushimimomoyama Station was not complete. As explained later, in the area far away from a station, since there are few pedestrians, it is difficult to classify sampling points correctly. The clustering problem is defined so as to achieve the maximum possible increase in classification accuracy of each class. We do not regard the area of crime-free location where sampling points are misclassified into class P as potential crime location since we first confirm whether our clustering-classification method functions well in the present study. This problem is categorized as a crime prediction problem and future work.

Thirty six out of the 52 actual crime locations have been correctly classified. This means that 36 actual crime locations can be classified using the attributes adopted in the present study. The present method clearly distinguishes classifiable actual crime locations from those that are not classifiable based on the different circle sizes. This is useful for clarifying the latent attributes that might improve the accuracy of the misclassified area. It is also important to increase the classification accuracy of the points at which crimes originate. We intend to investigate this issue in the future.

CAEP found 1,452 EPs in class P and 2,015 EPs in class N. The EPs with the twenty highest contributions in Class P and N are listed in Tables 10 and 11, respectively. For class P, we can often see the combinations of itemsets such as \(w_2_{-nw}=(0.0135-\infty)\), which indicates that the degree of natural surveillance from non-wall buildings are relatively high, and \(rw\_{***}, w_2\_{**}, \text{ or } wd\_{**}=-\infty\) (inf-some value other than infinity). These EPs are #1, 3, 5, 8, 9, 10, 13 and 19. Multistory garages or simple store structures are classified into non-wall buildings. Since people do not reside in these buildings, non-wall buildings do not have the high degree of
the type-1 natural surveillance. Moreover, the level of the natural surveillance from other buildings paired with non-wall buildings in those EPs is low. Therefore, although the condition that the crime locations are far from some stations outside the studied area is sometimes added, the existence of non-wall buildings itself is thought to be an important factor on the occurrence of bag snatching incidents.

We focus on other EPs. EP #4 has rw_TTL=36(820.5-inf) and this indicates that the total number of pedestrians from the seven stations is large. However, another item (rw_KKG=(0.818-inf)) is also included in #4 and rw_TTL=36(820.5-inf) cannot be found in other EPs. This indicates that the large number of pedestrians from stations itself is not a critical factor on the occurrence of crimes. We also find the items (sp_pr_rd=(inf-0.0014)) in EP #12 which indicates the density of restaurants and shops are small, and (dis_nst=(398.3-577.6)) in EP #20 which indicates that the distance from the nearest station is middle. These items might evoke the crime location which has not a little pedestrian, is a little away from stations but quiet with little shops and restaurants.

Other interesting EPs in Class P are as follows. EPs #6, 7 and 8 have similar patterns that the number of pedestrians from Keihan Chushojima Station is large and the existence of the shutter or business instructions. As listed in Table 3, the number of pedestrians from Keihan Chushojima Station is not so many. However, there exist many multi-talent buildings exist and their operations might finish and shutters are closed in the night. Therefore, enough degree of type-1 natural surveillance may not be expected from such buildings. EPs #14, 15 and 18 also have similar patterns on individual houses. They have items w4 ih wk=(0.004-inf), which indicates, or w4 ih sh=(0.0072-inf), which indicates. They inhibit the natural surveillance from a window. In addition, EPs #14 and 18 has the item (w4 pb wn=(inf-0.000001)), which indicates the level of wall visibility from normal windows on individual houses is low. These EPs imply that the window with any shields of individual houses might decline the degree of type-1 natural surveillance.

In the case of class N, the pedestrian-related items have a noticeably strong influence. In particular, rw_KFM=(inf-0.7111), which indicates the number of pedestrians from Keihan Fushimimomyo Station located in the east of the area studied, is a little, has a strong influence on the non-occurrence of crimes through those twenty EPs. The number of pedestrians from Kinetsu Momoyamagoryoymae Station (#5), Keihan Tambabashi Station (#5) or Kinetsu Tambabashi Station (#6) are also a little. Especially, EP #4 has only rw_KFM=(inf-0.7111). On the other hand, there are various items paired with rw_KFM=(inf-0.7111). The common feature of them is that those items are w4 pb ws=(inf-some value other than infinity). Although individual houses are not included as the building usage of them, the factors of building usages and wall components for class N is not clear in comparison with class P. It can be said that in the case of class N, only the littleness of pedestrians from stations is a critical factor.

Meanwhile, for both classes, wall visibility without any distinction between building usages or wall components (w1), type-2 natural surveillance (non_vis), population density, street illuminance and land uses seem not to be so important factors in the studied area.

### Table 10 EPs with the twenty highest contributions in Class P.

<table>
<thead>
<tr>
<th>#</th>
<th>Contribution</th>
<th>Item1</th>
<th>Item2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.158</td>
<td>rw_MI=(inf-0.05)</td>
<td>w2 nw=(0.0135-inf)</td>
</tr>
<tr>
<td>2</td>
<td>0.155</td>
<td>rw_MI=(0.5805-inf)</td>
<td>w4 bs=en=(inf-0.000001)</td>
</tr>
<tr>
<td>3</td>
<td>0.154</td>
<td>rw_KTB=(inf-0.1014)</td>
<td>w2 nw=(0.0135-inf)</td>
</tr>
<tr>
<td>4</td>
<td>0.152</td>
<td>rw_KKG=(0.818-inf)</td>
<td>rw_TTL=36(820.5-inf)</td>
</tr>
<tr>
<td>5</td>
<td>0.152</td>
<td>rw_CTB=(inf-0.1471)</td>
<td>w2 nw=(0.0135-inf)</td>
</tr>
<tr>
<td>6</td>
<td>0.152</td>
<td>rw_KCG=(3.2267-inf)</td>
<td>w4 gb sh=(0.0332-inf)</td>
</tr>
<tr>
<td>7</td>
<td>0.151</td>
<td>rw_KCI=(3.2267-inf)</td>
<td>w3 sh=(0.0371-inf)</td>
</tr>
<tr>
<td>8</td>
<td>0.151</td>
<td>w2 nw=(0.0135-inf)</td>
<td>w2 pb=en=(inf-0.000004)</td>
</tr>
<tr>
<td>9</td>
<td>0.150</td>
<td>w2 nw=(0.0135-inf)</td>
<td>w4 ob=en=(inf-0.000001)</td>
</tr>
<tr>
<td>10</td>
<td>0.149</td>
<td>w2 fbn=en=(inf-0.000001)</td>
<td>w2 nw=(0.0135-inf)</td>
</tr>
<tr>
<td>11</td>
<td>0.144</td>
<td>rw_KCI=2.2267-inf</td>
<td>w2 fbn=(0.0579-inf)</td>
</tr>
<tr>
<td>12</td>
<td>0.144</td>
<td>sp_pr_rd=(inf-0.0014)</td>
<td>rw_KFM=(0.7111-1.9299)</td>
</tr>
<tr>
<td>13</td>
<td>0.142</td>
<td>w2 nw=(0.0135-inf)</td>
<td>w4 ob=en=(inf-0.000001)</td>
</tr>
<tr>
<td>14</td>
<td>0.139</td>
<td>w4 ih wn=(inf-0.000001)</td>
<td>w4 ih wk=(0.004-inf)</td>
</tr>
<tr>
<td>15</td>
<td>0.136</td>
<td>w4 ih dr=(0.0036-0.0424)</td>
<td>w4 ih wk=(0.004-inf)</td>
</tr>
<tr>
<td>16</td>
<td>0.133</td>
<td>w3 wkb=(0.0056-inf)</td>
<td>w2 pb=en=(inf-0.000001)</td>
</tr>
<tr>
<td>17</td>
<td>0.133</td>
<td>w4 gb wk=(0.0056-inf)</td>
<td>w2 pb=en=(inf-0.000001)</td>
</tr>
<tr>
<td>18</td>
<td>0.133</td>
<td>w4 ih wn=(inf-0.000001)</td>
<td>w4 ih sh=(0.0072-inf)</td>
</tr>
<tr>
<td>19</td>
<td>0.132</td>
<td>w2 nw=(0.0135-inf)</td>
<td>w4 pb=en=(inf-0.000001)</td>
</tr>
<tr>
<td>20</td>
<td>0.129</td>
<td>dis_nst=(398.3-577.6)</td>
<td>w4 bi wp=(0.000001-0.0016)</td>
</tr>
</tbody>
</table>

### Table 11 EPs with the twenty highest contributions in Class N.

<table>
<thead>
<tr>
<th>#</th>
<th>Contribution</th>
<th>Item1</th>
<th>Item2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.316</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 nw abv=(inf-0.000002)</td>
</tr>
<tr>
<td>2</td>
<td>0.315</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 nw drv=(inf-0.000001)</td>
</tr>
<tr>
<td>3</td>
<td>0.311</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>rw_CMG=(inf-0.1986)</td>
</tr>
<tr>
<td>4</td>
<td>0.310</td>
<td>rw_KFM=(inf-0.7111)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.309</td>
<td>rw_KTB=(inf-0.1014)</td>
<td>rw_KFM=(inf-0.7111)</td>
</tr>
<tr>
<td>6</td>
<td>0.309</td>
<td>rw_CTB=(inf-0.1471)</td>
<td>rw_KFM=(inf-0.7111)</td>
</tr>
<tr>
<td>7</td>
<td>0.309</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 ab wn=(inf-0.000001)</td>
</tr>
<tr>
<td>8</td>
<td>0.302</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 pb wn=(inf-0.000001)</td>
</tr>
<tr>
<td>9</td>
<td>0.302</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 gh wn=(inf-0.000001)</td>
</tr>
<tr>
<td>10</td>
<td>0.302</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 pb dw=(inf-0.000001)</td>
</tr>
<tr>
<td>11</td>
<td>0.301</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 pb ws=(inf-0.000002)</td>
</tr>
<tr>
<td>12</td>
<td>0.298</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 bt dw=(inf-0.000001)</td>
</tr>
<tr>
<td>13</td>
<td>0.297</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 pb dw=(inf-0.000001)</td>
</tr>
<tr>
<td>14</td>
<td>0.294</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 pb wn=(inf-0.000001)</td>
</tr>
<tr>
<td>15</td>
<td>0.294</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 pb wn=(inf-0.000001)</td>
</tr>
<tr>
<td>16</td>
<td>0.294</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 pb wn=(inf-0.000001)</td>
</tr>
<tr>
<td>17</td>
<td>0.292</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 bt wk=(inf-0.000001)</td>
</tr>
<tr>
<td>18</td>
<td>0.291</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 bt wk=(inf-0.000001)</td>
</tr>
<tr>
<td>19</td>
<td>0.291</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 bt wn=(inf-0.000001)</td>
</tr>
<tr>
<td>20</td>
<td>0.289</td>
<td>rw_KFM=(inf-0.7111)</td>
<td>w4 ob wn=(inf-0.000001)</td>
</tr>
</tbody>
</table>

Figs. 9 and 10 show two panoramic images at points A and B in Fig. 8, where the sum of the contributions of the EPs is the highest. At point A (for Class P), we see a concrete block wall and there are no residential buildings nearby. We can also see a multilevel parking tower, which is a type of non-wall building. Thus, many class-P patterns revealed by CAEP are included at point A. Point B (for Class N) is located at a Riverside to the west of the station and is not crowded with buildings, but rather single-family houses face the roadside. At point B, several patterns that correspond to class N can be observed, e.g., a great distance from Keihan Fushimimomyo Station.

7. Conclusion

We have proposed a clustering-classification hybrid method and have analyzed the relationship between bag snatching and spatial attributes in the Fushimi Ward of Kyoto City. We applied CAEP to the database with several attributes and discovered numerous spatial patterns. The effectiveness of
CAEP to the rare-class classification problem as mentioned in Section 1 was confirmed also in the present study. Compared with Tables 8 and 9, we can find that CAEP exhibit better accuracy than that of logistic regression and decision tree which are popular classifiers with cost sensitive classification. The patterns can be summarized as follows.

The structure of the area studied is mixed land-use of mainly dense business and residential districts and is suitable for the use of trains. Although the pedestrian modeling is not yet sufficient, the patterns discovered appear to be based on these characteristics. All EPs for Class P listed in Table 10 are composed of two items. The primal pattern is the combination of the littleness of pedestrians from some stations and existence of the non-wall building such as a multi-level car parking. Other patterns evoke the crime location which has not a little pedestrian, is a little away from stations but quiet place with shutters little shops and restaurants. We also got the finding that windows with any shields of individual houses might decline the degree of type-1 natural surveillance. On the other hand, EPs for Class N listed in Table 11 are simpler. That is, the factors of building usages and wall components for class N is not clear in comparison with class P. It can be said that in the case of class N, only the littleness of pedestrians from stations is a primal factor.

Meanwhile, the index of type-2 natural surveillance was not ranked in the upper EPs for both classes. This indicates that the openness of space is not a differentiate factor in the area studied, and has been already found in our previous study\(^{19}\). We will continue to study the effect of type-2 natural surveillance in other areas in the future.

The analysis method of the present study had the highest overall classification accuracy among the usage of CAEP, which indicates the validity of this method, in which unlabeled data are used for classification. In addition, it is interesting that whether an actual crime location and its neighbors can be classified is determined by the size of the clustering area. However, as explained in Section 6, since only approximately 2/3 of the points of the actual crime locations could be correctly classified, improving the classification accuracy of actual crime locations should be considered.

One solution is to extend the clustering problem to a multi-objective optimization problem with two objective functions: one to maximize the accuracy of all sampling points defined in this article and one to maximize the accuracy of actual crime locations. This extended problem can be solved by using multi-objective genetic algorithms, for example. At present, the computational time required to solve this problem by CAEP is excessive. Because of the apriori algorithm that extracts emerging patterns, solving the problem requires one week by a PC (Intel Core i7 960, 24 MB memory, Windows 7 64 bit professional, and Visual C++ 2008 professional compiler). In the future, we will use other fast itemset miner, such as linear time closed itemset miner\(^\text{30}\) for apriori algorithms.

Acknowledgements

The bag snatching data were provided by the Kyoto Prefectural Police. This study was supported by a Grant-in-Aid for Young Scientists (B) (No. 22760457). The author also thanks to the reviewers who gave valuable comments.

References

和文要約

1. 序
防災環境設計（CPTED）の考え方の基礎である空間の監視性に関し
て，建物間空間（type-1）と空間対空間（type-2）の区別があることが
指摘されている。GIS を用いて監視性と犯罪発生の関係を分析する試
みはこれまでにも多数なされているが，その多くがtype-2の監視性を
対象としたものであった。本研究では，京都市伏見区のひっくり犯
罪を対象としてtype-1の監視性等との関連を分析する。監視性の属性
を含めて様々なデータを分析するための手法として，コントラストデー
ータマイニングの分析で研究されていた既在パターン分析を用いて，犯
罪発生／非発生場所の分類と特性抽出を行う。本研究では，犯罪発生
地点の周囲の犯罪発生を起こしやすい傾向を有する場所と考え，その範囲
を，分類精度が最大となるような半教師型のクラスタリングの問題を定
義して決定するにあたりリテラリティがある。

2. 対象地域と使用データ
対象地域は，平成17年の京都市の統計で，ひっくりが含まれる
未発見犯罪の認知件数が最も高い京都市伏見区区圏内であり，東西
約9.8km南北約1kmの対象地域の内外に7つの鉄道駅が存在する。分
析対象道路の総延長は約20kmである。対象地域内で2004年1月から
2005年12月の間に計57件のひっくり発生を発見した。このポイントデー
タを分析用データとして用いる。GISのデータベースとして，空間デー
ータ基盤，住宅地図，土地利用，国勢調査，事業所統計などのデータ
を利用した。さらに，建物壁面の特徴抽出を計236度の全方位
画像を利用する。また，夜間照明を計測し，ポイントデータとして用
いる。

3. 属性
3.1 壁面監視性
Type-1の監視性をモデル化するためには，建物壁面の関階，ドア，シャッターなど7種類の構成要素の面積が，道路のある地点からの
程度を見えるかを平面上でモデル化し，指標を作成する。基本的な指
標は，式(1)に示すように，ある地点からある壁面までの距離，見えの
角度，見える壁面の大きさに単純なものとする。これを基礎として，
壁面の各構成要素の面積を考慮した式(2)を定義する。さらに建物用途
も考慮して，最終的な建物の監視度を式(3)～(6)の4種類定義する。

3.2 駅からの推定歩行者数
京都市伏見区では駅間でひっくりの発生頻度が高い。駅か
らの人の流れを，ランダムウォークとしてモード化する。駅から，
一人の歩行者を出発させて道路ネットワーク上をランダムウォーク
させ，各駅から1kmの範囲外に出るまでを一試行とし，それを
100,000回繰り返して各駅の位置で平均をとり，そこで各駅の一
日の乗車客数を反映させて，推定歩行者数とする。

3.3 その他の属性
上記以外に，最寄駅までの距離，土地利用，人口密度，店舗の従
業員密度，夜間照明などを属性として用いる。

4. 分析方法
4.1 データの準備
ひっくり犯人も発生を考慮して，道路脇に約1kmだけ内側の直
線上に10mおきにサンプリング点を配置した（図6）。その結果，2,769
点のサンプリング点が生成された。各サンプリング点には，犯罪発生
点を表すクラスターPもしくは，非発生点のクラスNのラベルが割り当て
られる。これにより値から示す方法で決定される。

4.2 クラスタリング方法
実際に犯罪が起こった場所をACL，サンプリング点の中でACLに
近い潜在的な犯罪発生点をPACLから遠いところをCFLとする。
ACLにはPのラベルが，CFLにはNのラベルが付与される。PCLに
関しては，Fig7に示すように，各ACLから遠方で，半径がある範囲
内のものにラベルPを，いずれのACLからの範囲外にあるもの
にはラベルNを付与する。この半径を以下に示す方法で最適化する。

4.3 分類方法
本研究では顕在パターンをベースとしたCAEPを用いて，犯罪発生
点とサンプリング点の離れ度を分類する。4.2で決定した
データセットの各点のクラスタラベルを判断する。

4.4 分類精度
4.3のモデルの分類精度を，PとNに関する真陽性率と真陰性率で
評価する。そして，4.2を含めたクラスタリング問題を，10回の交叉
検証によるこれらの精度の最小値を最大化する問題とする。

4.5 最適化方法
メソッドリスティクスの一種である差分進化法を用いて，ACLの
目的関数を最大化するよう，4.2～4.3まで何回も繰り返し，各ACL
からの最適なPCLの半径を決定する。

5. クラスタリング問題の解

5.1 設定
計画に必要なCAEPとDEの各種パラメータを設定して，4.2～4.5
クラスクラスタリング問題解く。

5.2 結果
問題を解いた結果，分類精度（真陽性率，真陰性率，加えて正答率）
がともに85.5の精度でPCLの範囲を確定することができた。この
ときの半径の集合をopt_uとする。Table 8に示すように半径を一辺に
変化させた場合の精度と比較すると，opt_uの結果が最も精度が高
かった。さらに，CAEP以外に，マトシティ回帰と決定木でも分類
を行ったがTable 9に示すようにCAEPを使った場合が最も精度が高
く，本研究での提案手法の効果が認められた。

6. opt_uの結果の詳細分析
前章で得られた結果を詳細に分析する。Fig 8は各サンプリング点
の分類結果である。図中の大きな円が最適化されたPCLの範囲を示
している。場所によってその大きさに差が生じることがある。大きな
円から，分類に有用な説明変数の同質性が確保されたことを示してい
る。また，CAEPを適用することで得られる顕在パターンを分析した結
果，犯罪発生場所では立体駐車場などの無駄舎といくつかの遠方の駅
からの推定歩行者数の大小の組み合わせが最も動的なパターンで
あった。一方犯罪が発生する場所では，対象地域内にある京成伏見競
馬場駅から遠いことが，単独で主要な因子として抽出された。

7. 結
また移動行動のモデリングは十分とはいえないが，半教師クラスタ
リングとCAEPによる手法を組み合わせることで，高い精度で犯
罪発生場所を分類し，特徴的な空間構成パターンを抽出することができた
ため，ACL自体の分類精度は20程度なので，精度の向上が必要
である。また，現在の実装ではモデルの最適化計算時間がかさむの
で，効率的なアルゴリズムを改良する必要がある。

(2012年8月10日受理後，2013年1月15日採用決定)