Discovering Distinctive Spatial Patterns on Snatch in Kyoto City with CAEP

A. Takizawa[1], W. Koo[2], N. Katoh[3]

[1]Kyoto University, Graduate School of Engineering,
Kyoto University Katsura Campus C-cluster, Nishikyo-ku, Kyoto, 615-8540, Japan.
Telephone: +81-75-383-2941
Fax: +81-75-383-2941
Email: kukure@archi.kyoto-u.ac.jp

[2]Nanzan University, Graduate School of Mathematical Sciences and Information Engineering,
27 Seirei-cho, Seto, Aichi, 489-0863, Japan
Telephone: +81-561-89-2000
Email: oaiskoo@nanzan-u.ac.jp

[3]Kyoto University, Graduate School of Engineering,
Kyoto University Katsura Campus C-cluster, Nishikyo-ku, Kyoto, 615-8540, Japan.
Telephone: +81-75-383-2939
Fax: +81-75-383-2939
Email: naoki@archi.kyoto-u.ac.jp

1. Introduction

In Japan, street crimes dominate about half of crime occurrences. Street crimes have been considered to be closely related to the structure of urban space. Then, the concept called Crime Prevention through Environmental Design (CPTED) was formulated (Jeffery, 1971). Development of GIS has promoted the movement of the evidence-based crime prevention. Since street crimes occur in the complex context of urban spaces, analysis methods used in previous studies has limitation for deeper understanding of crime occurrences. We are interested in data mining techniques which can deal with the complicated data of urban space, and have studied the car-related street crimes in Nishikyo-ku of Kyoto City (Takizawa et al., 2007) with a data mining analysis.

In this study, we investigate the relation between snatch occurrences and spatial attributes in Fushimi-ku of Kyoto City. As spatial attributes, we consider demographic data, land-use, visibility of space and illuminance on the street. These attributes are analyzed with CAEP (Dong et al., 1999-2) which can classify a database with high precision and find useful patterns of itemsets in the database.

2. Targeted area and data

The targeted area is located at the central zone of Fushimi-ku which is the suburb of Kyoto City. The targeted area is a rectangle of about 2km by 1km. There are five railway stations. Snatch data was provided from Kyoto Prefectural Police. 343 snatches were occurred in the whole Fushimi-ku from Jan. 2004 to Dec. 2005, among which 96 snatches were occurred in the studied area. Other databases are spatial fundamental map, land-use map, house map, and census data. Since more than 10 types land usages are defined in the land-use map and they seem to be too fine for our analysis, we define land-use taxonomy (see fig. 1). We also measured illuminance on the street and used it for analysis.

Open Space (os)
-- Passable space (ps)
---- Non-road (nr)
------ Field (fi)
------ Vacant space (va)
------ Park (pa)
---- Road (ro)
-- Impassable space (= River(ri))
Closed space (cs)
-- Housing site (ho)
---- Low-rise housing site (lh)
---- Dense-low-rise housing site (dh)
---- Mid-to-high rise housing site (mh)
-- Non-housing site (nh)
---- Industrial site (in)
---- Business site (bu)
---- Public facility site (pu)

Figure 1. Taxonomy of land-use attributes.

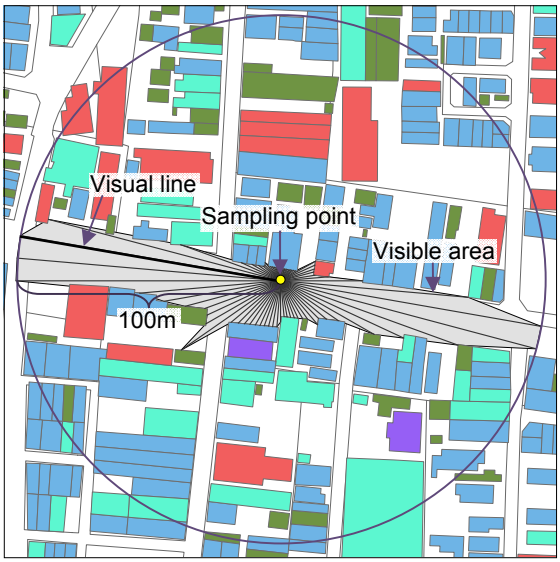


Figure 2. Visual lines and visible area.

3. Attributes

We created a database consisting of the following attributes at every sampling point.

Since a lot of people use trains for commuting, the path from a station to his/her house seems to be used more frequently than other streets. We estimate the number of pedestrians on the street (Pedestrian) by summing up the night population passing through it as the shortest path from their houses to the nearest station.

Visibility analysis is also performed. As shown in fig. 2, visual lines are drawn from a sampling point at every 5 degrees in a radial order. The maximum length of a visual line is limited to 100m. In our analysis we regard only building walls as obstacle objects against visibility. The following attributes are created: number of visible buildings (Visbsize), maximum, mean and minimum lengths of visual lines (Visln_max, Visln_mean and Visln_min), standard deviation of visual line lengths (Visln_std), the sum of visual lines for each different classification system of buildings (Visbtype1_XX, Visbtype2_YY) where XX and YY denote building types, the area of each land-use within a circle of radius 20m centred at the sample point (Land_ZZ) and the visible area (Visland_ZZ) where ZZ denotes a land-use defined in fig. 1.

The following attributes are also defined: distance from the nearest station (Distance), night population in the 100m mesh where the point is located (Population), and the weighted mean of illuminance levels of the nearest three points (Illuminance).

4. CAEP

CAEP is a classifier method which can classify a database with high precision. At first, we explain emerging pattern (EP) (Dong et al., 1999-1). EP is defined as an itemset whose support increases significantly from one dataset to another. Let D denote the database we are dealing with. Suppose an attribute A has two distinct classes C and $\overline{C}$,

and let us consider datasets D_c and $D_{\bar{c}}$ obtained by partitioning D according to the values of attribute A. Let $t \in D_c$ denote a record which belongs to C, and $e \subseteq t$ denote an itemset of t, respectively. Support which represents the ratio of containing t in D is defined as

$$sup_c(e) = \frac{|\, t \in D_c, e \subseteq t \,|}{|\, D_c \,|}.$$

The growth-rate of e from $D_{\bar{c}}$ to D_c denoted by $growth\_rate_c(e)$ is defined as

$$growth\_rate_c(e) = \begin{cases} \dfrac{sup_c(e)}{sup_{\bar{c}}(e)} & (sup_{\bar{c}}(e) \neq 0), \\ \infty & (sup_{\bar{c}}(e) = 0). \end{cases}$$

We call e whose growth-rate exceeds 1 as an EP. Then, let us give the definition of CAEP. The contribution of e to C is defined as

$$\alpha_c(e) = \frac{growth\_rate_c(e)}{growth\_rate_c(e)+1} \cdot sup_c(e).$$

Let $E(C)$ denote the set of EPs for C derived from training data. Aggregated score which represents the possibility that an instance s belongs to C is defined as

$$score(s, C) = \sum_{e \subseteq s, e \in E(C)} \alpha_c(e).$$

In order to compare aggregated scores between different classes, the value is normalized by mean of aggregate scores. Finally, s is classified as the class having the higher normalized aggregate score.

5. Analysis

5.1 Setup

Considering that pedestrians tend to walk on the edge of the street, sampling points are placed 1m inside of the road. The interval of sampling points on the road is set to be 10m. Then, class label either N or P which represents non-occurrence or occurrence of snatch respectively is assigned to each point. Sampling points within 20m radius of a snatch occurrence point is labelled as P. Furthermore, if the sampling point is contained in more than one neighbourhood of snatch occurrence points, we duplicate the sampling point at the same place with the same attribute values. Consequently, 7,071 sampling points are labelled as N and 978 points are labelled as P. Then, they are divided into two datasets D_N and D_P, respectively.

5.2 result

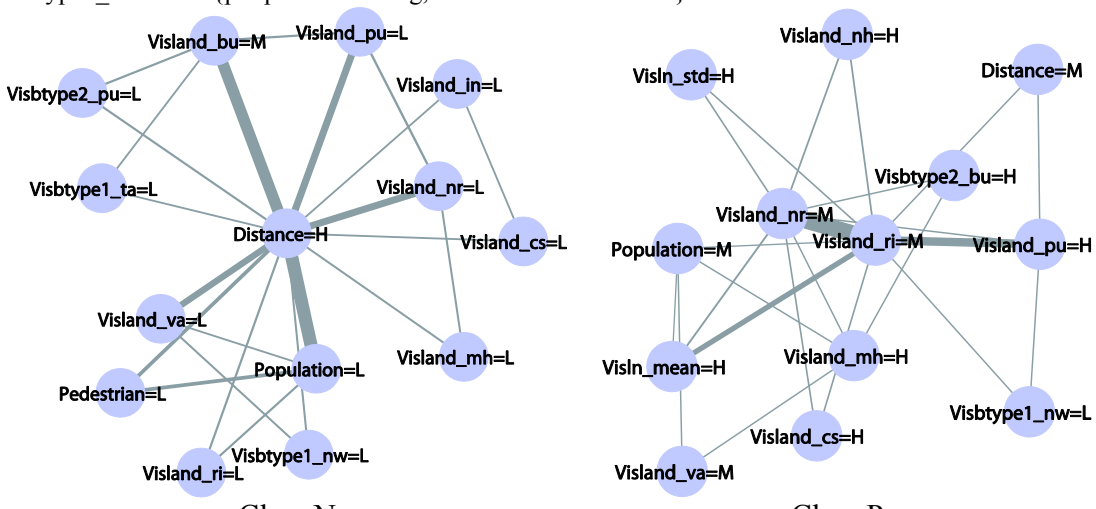
In order to apply CAEP to datasets, we divide each attribute into three subintervals so that the number of records in each subinterval is as equal as possible. According to the level of subinterval, each attribute are relabelled as L (Low), M (Middle) and H (High). Classification accuracy of each class and the whole dataset are obtained by 10-fold cross validation. After classification, we compared the classification accuracy of CAEP with those of other major classifiers: decision tree and logistic regression (see table 1). CAEP exhibits the best accuracy in all cases.

Classifier	Class N	Class P	Whole
CAEP	0.807	0.728	0.797
Logistic regression	0.660	0.708	0.666
Decision tree	0.739	0.719	0.737

Table 1. Comparison of classification accuracies.

CAEP found 27,562 EPs in class N and 6,457 EPs in class P. In order to understand the relationship of items used for EPs, they are visualized as graphs. Fig. 3 shows graph representations of items for both classes having ten highest supports. The weight of an edge is defined as the sum of supports of EPs to which the pair of items belongs. An edge with higher support is drawn thicker. From the observation of these graphs, we can say that snatch tends not to occur at the place where distance from stations is relatively large and the area of facility site, non-road site or vacant space are small. Conversely, snatch tends to occur at the place where visibility of river or non-road site is relatively middle, and visibility of public facility site or openness of space are relatively high.

Visbtype1_XX: XX={ta: target building, nw: non-wall building}
Visbtype2_YY: YY={pu: public building, bu: business institution}



Class N Class P

Figure 3. Graph representations of items of primal EPs on both classes.

Fig. 4 shows classification result of each sampling point. In the north of the targeted area where residential sites dominate, most points seem to be classified well. Meanwhile, misclassification that actual class is N but classified class is P outstands in the south of the targeted area where a lot of drinking places exist. This result might mean that in residential area the risk of the snatch occurrence tends to be high only in particular sites, meanwhile in downtown it can become high regardless of the site location.

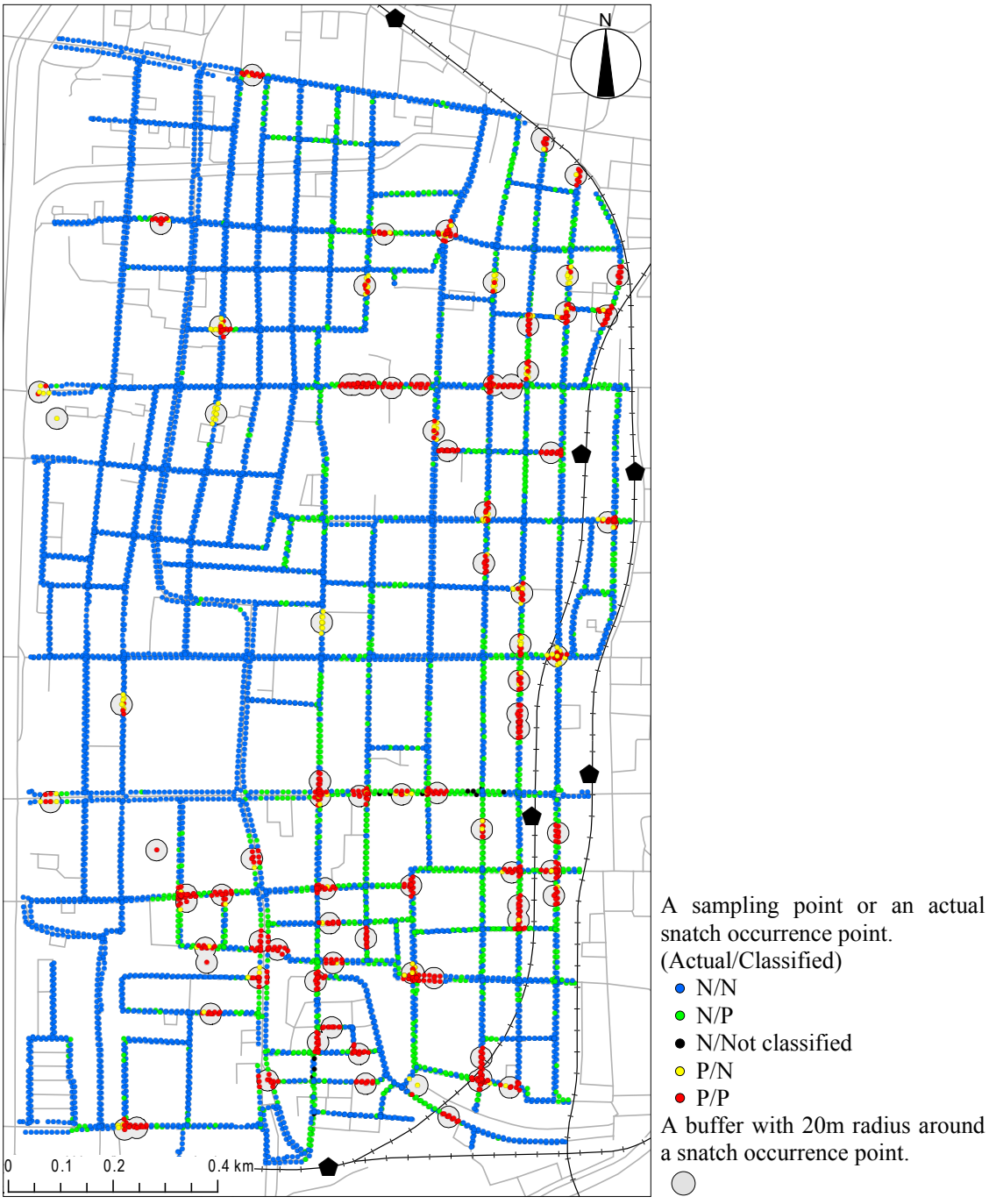


Figure 4. Classification result of snatch occurrences with CAEP

6. Conclusion

In this study, the relation between snatch occurrences and spatial attributes in Fushimi-ku of Kyoto City was analyzed by applying classification method with CAEP. We obtain the following conclusions.

i) Classification accuracy of CAEP is higher than those of logistic regression and decision tree. CAEP is suitable for analyzing spatial characteristics of crime hotspots whose area is much smaller than that of other crime non-occurrence sites.
ii) Primal spatial patterns of crime occurrence points extracted as the form of EP are that visibility of river or non-road site is relatively middle, and visibility of public facility site or openness of space are relatively high.
iii) In residential area, the risk of snatch occurrences tends to be high only in particular sites. Meanwhile in downtown, it can become high regardless of the site location.

7. Acknowledgements

The snatch data was offered by Kyoto Prefectural Police. This study is supported by Grant-in-Aid for Young Scientists (B) (No.20760405) of the Ministry of Education, Culture, Sports, Science and Technology-Japan.

8. References

Dong G and Li J, 1999, Efficient mining of emerging patterns: Discovering trends and differences, *Proc. of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA: ACM Press, 43-52.

Dong G, Zhang X, Wong L and Li J, 1999, CAEP: Classification by Aggregating Emerging Patterns, *Int'l Conference on Discovery Science*, 30-42.

Jeffery R, 1971, *Crime Prevention through Environmental Design*, Beverly Hills, CA: Sage Publications.

Takizawa A, Kawaguchi F, Katoh N, Mori K and Yoshida K, 2007, Risk Discovery of Car-Related Crimes from Urban Spatial Attributes Using Emerging Patterns, *Int. J. of Knowledge-based and Intelligent Engineering Systems*, 11(5): 301-311.