RESEARCH

Spatial analysis of outdoor indecent assault risk: a study using ambient population data

Hiroki M. Adachi[1]* and Tomoki Nakaya[1]

Abstract

Spatiotemporal data on ambient populations have recently become widely available. Although previous studies have indicated a link between the spatial patterns of crime occurrence and ambient population distribution, more detailed information, such as the population most likely to be victims by gender and age group, could better predict the risk of crime occurrence. Therefore, this study aimed to analyze the risk of indecent assault, a typical crime with a high number of young female victims, in southern Kyoto Prefecture. We utilized population distribution by gender and age group at different times of the day. After extracting daily patterns (factors) of the population using non-negative matrix factorization, we statistically modeled the risk of indecent assault using a spatial conditional autoregressive model. The results showed that the model, which considered a spatiotemporal ambient population, demonstrated superior performance during nighttime hours. Furthermore, by interpreting the factors significantly associated with the risk of crime occurrence, the findings provided valuable insights into local crime prevention measures that consider daily temporal changes in the gender and age-group composition of individuals present in a specific area.

Keywords Street crime, Ambient population, Mobile spatial statistics, Non-negative matrix factorization, Conditional autoregressive model

Introduction

Crimes occur at specific locations and times (Brunsdon et al., 2007; Newton & Felson, 2015; Sherman et al., 1989; Weisburd, 2015). If the time and space in which the risk of a crime increases can be identified, specific crime prevention measures can be recommended to law enforcement agencies and citizens, thereby reducing the number of crimes. The opportunity to commit a crime occurs when the offender and the target of the crime (target) come together in the same time and space. The routine activity approach considers the time and space in which "motivated offenders," "suitable targets," and the "absence

of capable guardians" converge to create the conditions for the establishment of a crime (Cohen & Felson, 1979). A capable guardian is some presence that intervenes in the crime opportunity by being there (Hollis et al., 2013), and is also associated with "eyes on the street" and natural surveillance as an effect of informal surveillance by a human presence (Jacobs, 1961; Newman, 1972).

Crime pattern theory, which incorporates the routine activity approach, focuses on templates of people's daily activities to explain the relationship between offenders' or targets' activities and criminal opportunities (Brantingham & Brantingham, 1984). Crime pattern theory emphasizes that people move between specific activity nodes as part of their routine activities (e.g., an activity pattern in which a person goes from home to work in the morning, leaves in the evening, drinks with colleagues at a particular downtown establishment at a specific restaurant, and returns home at a fixed time). It explains that criminal opportunities arise in a time–space where

*Correspondence:
Hiroki M. Adachi
hiroki.adachi.e4@tohoku.ac.jp
[1] Department of Frontier Science for Advanced Environment, Graduate School of Environmental Studies, Tohoku University, 468-1, Aoba, Aramaki, Aoba-ku, Sendai-shi, Miyagi 980-0845, Japan

potential victims' and offenders' lifestyles and daily activity patterns overlap. In addition, crime pattern theory elucidates the mechanism of crime hot spot occurrence through crime generators and crime attractors (Brantingham & Brantingham, 1995). Crime generators are specific areas where large numbers of people congregate, such as shopping precincts and entertainment districts. Crime attractors pertain to areas known for criminal opportunities, including bar districts and prostitution areas. Therefore, the risk of crime is considered higher around facilities used by people at certain times of the day because of the large number of interacting people.

For example, Haberman and Ratcliffe (2015) demonstrated that the relationship between the risk of robbery in Philadelphia and surrounding facilities varies with the time of day. They suggested that dynamic changes in the number of people using various facilities affect the risk of crime. Felson and Boivin (2015) examined the relationship between the number of property and assault crimes, respectively, and the number of visitors for various travel purposes (work, shopping, recreation, and education) and found a statistically significant correlation between these increases. Therefore, it is important to investigate the relationship between the amount of human presence and crime. Several studies have used residential population data obtained from the census to determine the relationship between static population and crime. However, the residential population does not reflect population changes due to patterns of human activity (Boggs, 1965).

In recent years, advances in information technology have made ambient population data available in various ways (Andresen, 2006, 2011). For example, location data from cell phones and social network services (SNS) can be examined to indirectly learn about the ambient population. Several studies have quantitatively assessed the association between ambient population data and crime risk (Amemiya et al., 2018; Andresen & Jenion, 2010; Bogomolov et al., 2015; Hanaoka, 2016; Harada, 2020; Hipp et al., 2019; Malleson & Andresen, 2015; Tucker et al., 2021). For example, Hipp et al. (2019) estimated the probability of a crime occurring in each census block every 2 h based on SNS tweet data. As a result, they could discuss crime risk in terms of suitable targets and capable guardians. Tucker et al. (2021) used geotagged Twitter data to study the correlation between ambient population and crime occurrence in four different time periods: weekday days, weekday nights, weekend days, and weekend nights. They found that the impact of commuters and tourists on crime varies by crime type and time of day. Hanaoka (2016) showed that the relationship between ambient population and snatch-and-run risk is reversed between daytime and nighttime through an analysis using ambient population data obtained from

mobile phone GPS location data. The usefulness of crime research focusing on human activity patterns through ambient population in crime analysis is beginning to be demonstrated (Newton et al., 2021).

Objectives

As noted above, many studies have suggested the utility of using ambient population in crime research. However, people's activity patterns are likely to vary by gender and age (Collia et al., 2003; Hamed & Mannering, 1993; Havet et al., 2021; Masso et al., 2019). Furthermore, differences exist in the number of victims based on gender and age (Morgan & Thompson, 2021). For example, victims of indecent assaults are primarily women in their teens and 20s (Cabinet Office, 2012; National Police Agency, 2020; Perkins, 1997). Given this fact and considering the routine activity approach and crime pattern theory, the ambient population by gender and age group (APGA) related to potential victims is likely to differ across space and time. Harada (2020), using "People Flow Project" data based on a Person trip (PT) survey, showed that the risk of snatch-and-run victimization in Tokyo is higher for older women and housewives, emphasizing attributes such as gender, age group, and means of transportation. The study also suggested that an analysis focusing on the attributes of victims is essential.

Bogomolov et al. (2015) conducted a crime prediction study using APGA obtained from cell phone data. They showed that ambient population and demographic data contributed significantly to prediction accuracy. Although their study demonstrated that APGA could improve the accuracy of crime prediction, an examination of the relationship between APGA and crime risk was beyond its scope, as the primary goal was to examine better crime prediction methods. However, not many crime studies based on APGA have been conducted. Thus, while existing research suggests that APGA is crucial information explaining criminal risk, few studies have used spatiotemporally detailed APGA.

Recognizing the need to select a crime category focusing on a specific gender and age group to demonstrate the applicability of APGA to crime analysis, this study selected indecent assault as its subject. Indecent assaults are index crimes used to monitor public safety in Japan. Although sex crimes, including indecent assaults, have been declining in recent years, the reduction is smaller than for street crimes (Kyoto Prefecture Police, 2020), and the severity of the damage they cause, known as the "murder of the soul" (National Police Agency, 2008), requires focused action.

To understand when and where these crimes are most likely to occur, it is crucial to examine the conditions

under which the "suitable targets" and "absence of capable guardians" of the routine activity approach are likely to be established. According to information on reported cases of indecent assaults in Japan, approximately 97% of victims are female, with approximately 80% in their teens to 20s (National Police Agency, 2020). The reasons why young women are more likely to be suitable targets for sexual offenses may be related to their low resistance and the sexual preferences of offenders, although a precise theoretical explanation is beyond the scope of this study. Because the number of these assaults increases from the evening to late at night, the distribution of capable guardians during this period should be considered.

Based on the above information, this study aimed to analyze the distribution of the population likely to be "suitable targets" and the "capable guardians" by using the distribution of the population at different times of the day by gender and age when examining the location of outdoor indecent assaults. Additionally, we perform spatial regression modeling of the number of reported crimes of indecent assaults that occurred outdoors in the southern part of Kyoto Prefecture, Japan, from 2015 to 2019 using APGA based on large-scale cell phone network information.

Materials and methods

Ambient population data

Mobile Spatial Statistics (DOCOMO InsightMarketing, Inc., Tokyo, Japan) comprise population statistics generated using operational data from the cell phone network of NTT DOCOMO (Okajima et al., 2013; Terada et al., 2013). Specifically, they provide hourly population distributions in small spatial units, approximately 500 m × 500 m grid cells, which are half the grid squares of the Standard Grid Square (Statistics Bureau of Japan), categorized by total number, gender, and age group. This study focuses on all areas south of Kyoto City in Kyoto Prefecture as the study area (Fig. 1). The average APGA on weekdays, excluding holidays (244 days), from January 1, 2019, to December 31, 2019, represents the general APGA of the study area. It is important to note that for Mobile Spatial Statistics, the data by gender and age group comprise two gender categories (men and women) and seven age categories (15–19, 20–29, 30–39, 40–49, 50–59, 60–69, and 70–79), respectively. A total of 336 columns of data representing the ambient population by time (24 columns), gender (two columns), and age (seven columns) were used. We excluded grid cells that contained zeros in any of the fields, signifying the absence of people at any time of day, resulting in 1,520 grid cells for analysis.

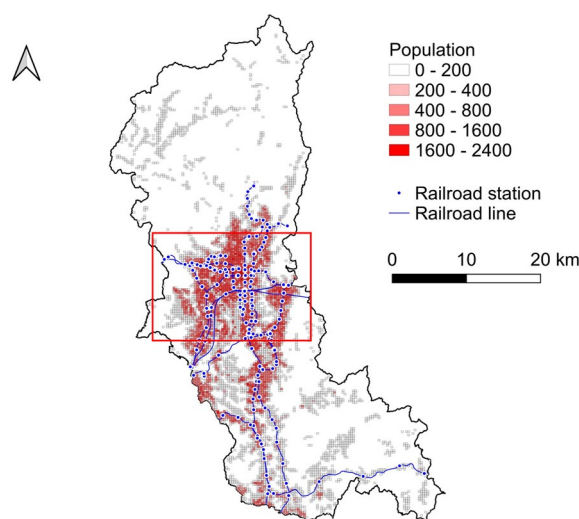Figure 2 presents the hourly changes in APGA for the grid cells containing the central station, downtown area,

Fig. 1 Survey area. The blue dots represent railroad stations and the blue lines railroad lines. The red rectangle is the area analyzed in this study; it is a relatively populated and central area
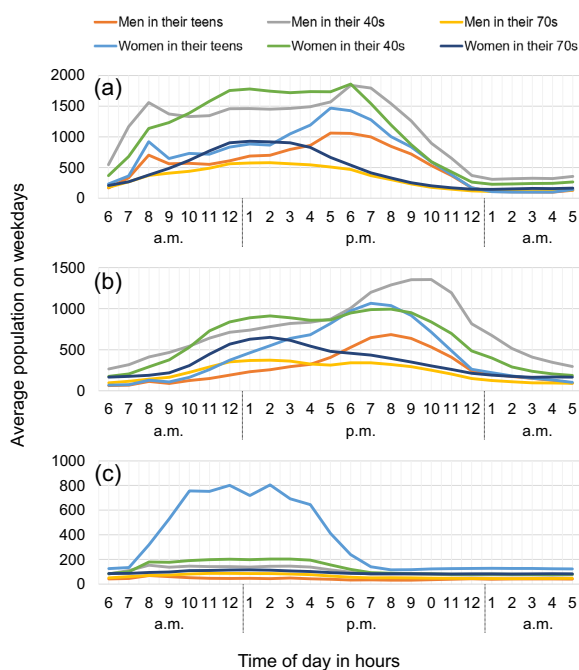
Fig. 2 Average APGA on weekdays. **a** The grid cell containing the central station (Kyoto station). **b** The grid cell containing the downtown area (Kawaramachi and Gion). **c** The grid cell containing the women's college

and women's college. In the grid containing the central station (Fig. 2a), individuals in their teens and men in their 40s exhibited small peaks from 6 a.m. to 9 a.m. and 5 p.m. to 6 p.m., indicating commuting patterns to work and school. On the contrary, women in their 40s and men

and women in their 70s showed an increase in the 9 a.m. to 4 p.m. period, indicating daytime trips, such as shopping. For the grid including the downtown area (Fig. 2b), the population of people in their teens and 40s increased after 5 p.m., indicating heightened activity in nightlife facilities. Conversely, the population in their 70s showed no increase at night. The grid including the women's college (Fig. 2c) demonstrated an increase in only women in their teens from 7 a.m. to 7 p.m., indicating gatherings of female students for classes and other activities.

Dimension reduction by non-negative matrix factorization

The APGA comprises 336 variables, some of which are strongly correlated with each other (e.g., the Pearson correlation coefficient of ambient population for men in their 20s and 30s in the grid cell containing the central station is 0.96). This situation poses the risk of multicollinearity during data analysis and makes interpretation difficult due to the large number of variables. To overcome this issue, we employed a dimension reduction technique known as non-negative matrix factorization (NMF).

NMF approximates a non-negative matrix as the product of non-negative matrices. It reconstructs each component of a non-negative matrix by adding up non-negative values, facilitating easy interpretation of decomposition results. NMF is often used for feature extraction from images, topic recovery, document classification, and audio source separation (Gillis, 2020). Given that ambient population values are positive, we selected NMF for this study.

Specifically, let X be a $n \times k$ non-negative matrix (data matrix), W a $n \times r$ non-negative matrix (basis matrix), and H a $r \times k$ non-negative matrix (coefficient matrix). Subsequently, NMF decomposes X as follows:

$$X \sim WH.$$

This is accomplished by minimizing the distance between X and WH:

$$\min D(X, WH),$$

where D is a distance function such as the Kullback–Leibler divergence (KL divergence) or Euclidean distance. r is the dimension (rank) after dimension reduction. We can choose any r, where $r \leq min(n, k)$.

NMF outcomes depend on the initial values of W and H. We utilized the non-negative double singular value decomposition (NNDSVD) method, which employs the non-negative component of the singular vector obtained by singular value decomposition as the initial value. The NNDSVD method uniquely determines initial values and enhances sparsity by replacing negative-valued

elements of singular vectors with zeros. Additionally, the non-smooth non-negative matrix factorization (nsNMF) algorithm (Pascual-Montano et al., 2006), using KL divergence and providing sparse results, was employed. The NMF package in R was used for NMF calculations.

Crime occurrence data

This study focuses on indecent assaults in Kyoto Prefecture between January 1, 2015, and December 31, 2019, as reported to the Kyoto Prefectural Police. It is important to note that several considerations apply to crime data recorded by the police. For example, not all crimes are recorded, and the number of reported crimes is skewed by region (Buil-Gil et al., 2021; Pina-Sánchez et al., 2023). The present study is also subject to such limitations. Furthermore, it should be noted that sex crimes, also covered in this study, are generally reported in lower numbers than non-sex crimes (Ministry of Justice, 2020).

Now, considering the 336 variables in the APGA, while analyzing the correspondence between the number of crimes matching each condition in the aggregate would be desirable, the number of crimes is too small for such a resolution. Therefore, this study aimed to extract the general typology of APGAs and examine their cumulative total in relation to crimes. Hanaoka (2016) successfully demonstrated the relationship between ambient population and snatch-and-run patterns by dividing the time period into daytime (6 a.m. to 6 p.m.) and nighttime (7 p.m. to 5 a.m.). The data analysis was conducted separately for each case, focusing on the distinct patterns of crimes during the day and night. Following Hanaoka's methodology, we divided the time of occurrence into daytime (6 a.m. to 5 p.m.) and nighttime (6 p.m. to 5 a.m.), where indecent assaults peaked between 6 p.m. and 2 p.m., differing from the daytime trend (Fig. 3). The number of indecent assaults recorded was 46 during the day and 300 during the night. We counted each crime per 500-m grid cell for statistical modeling. Notably, approximately 85% of victims aged 15 years and older were in their teens and 20s.

Geographic environmental factor data

This study incorporated variables related to people's daily activity patterns as control variables for spatial statistical modeling. First, we employed these variables based on the premise that environments with larger building areas or longer road lengths are more likely to be inhabited (built environment variables). Next, we considered facilities that are frequently visited by many people (facility variables). According to studies by Felson and Boivin (2015) and Boivin and Felson (2018), people's visitation purposes can be categorized as employment, shopping, recreation, and education, and their association with the
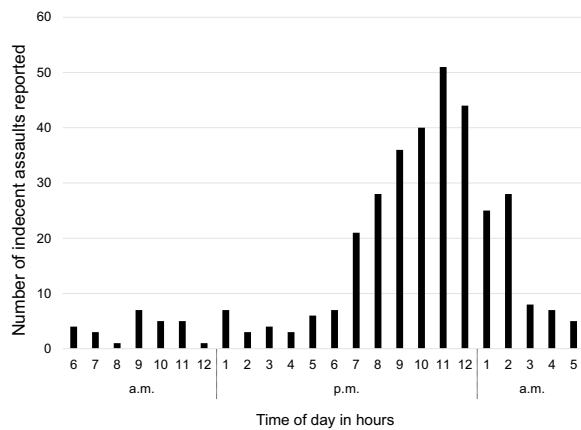
Fig. 3 Number of indecent assaults by the time of day in the target area on weekdays, from January 1, 2015, to December 31, 2019

risk of crime occurrence was examined. These purposes and the related facilities are considered crime generators (Brantingham & Brantingham, 1995).

We identified stations and bus stops as facilities commonly used by people, regardless of their purpose of travel. People engaged in work or educational activities are likely to concentrate around these facilities at specific times, such as commuting to work or returning home (Fig. 2a). We considered high schools, universities, and junior colleges as facilities for educational purposes. It is important to note that we excluded the number of elementary and junior high schools from consideration since the data for the ambient population aged 15 years and older were utilized. Facilities related to recreation and lifestyle were regarded as facilities for recreational purposes. Each facility's count in every grid cell served as an explanatory variable in our spatial statistical modeling. Table 1 provides details regarding each variable and its source.

Spatial statistical modeling

The number of crimes in each grid cell within the study area tends to be generally small, resembling a distribution that could follow a Poisson distribution, often applicable to crime data with rare probabilities of occurrence (MacDonald & Lattimore, 2010). In addition, geographic data commonly exhibit spatial autocorrelation. Given the inherent uncertainty in observational data such as crime, it is considered appropriate to perform modeling that considers the spatial autocorrelation between predicted and measured values. Therefore, we assume that the number of crimes follows a Poisson distribution and employ the conditional autoregressive (CAR) model, which accounts for spatial correlation in random

Table 1 Explanatory variables

Variables	Source	Count	Median	Min	Max
Target variable					
Indecent assault (Daytime)		46	0	0	2
Indecent assault (Nighttime)		300	0	0	8
Ambient population variables					
(Weight components of each factor by NMF)					
Built-environment variables					
Building area	Fundamental geospatial data (Geospatial Information Authority of Japan; obtained December 2020)	49,317	7		155,754
Total street length	ArcGIS Geo Suite Road Network 2021 (ESRI Japan Inc.)	4,675	0		13,053
Facility variables					
Number of train stations	Digital national land information (Published 2019)	0	0		6
Number of bus stops	Digital national land information (Published 2010)	0	0		26
Number of high schools	MapFan DB (GeoTechnologies Inc.)	0	0		2
Number of universities and junior colleges	MapFan DB (GeoTechnologies Inc.)	0	0		2
Number of convenience stores	MapFan DB (GeoTechnologies Inc.)	0	0		20
Number of wholesale trades and retail trades	Economic census (2016)	6	0		431
Number of life-related and entertainment services	Economic census (2016)	2	0		98

NMF: non-negative matrix factorization

variation caused by unknown factors. Specifically, we use the Leroux model (Lawson, 2021), which estimates the strength of smoothing in adjacent random effects. The Leroux model is expressed as follows:

$$y_i \sim Poisson(\lambda_i)$$

$$\ln(\lambda_i) = \boldsymbol{x}_i^T \beta + u_i$$

$$u_i|u_{-i} \sim Normal\left(\frac{\rho\sum_j u_j w_{ij}}{\rho\sum_j w_{ij} + 1 - \rho}, \frac{\tau_u^2}{\rho\sum_j w_{ij} + 1 - \rho}\right).$$

Here, y_i is the number of reported crimes, $\boldsymbol{x}_i$ the vector of explanatory variables, $\boldsymbol{\beta}$ the vector of coefficients, u_i the random effect with spatial correlation for the ith grid, u_{-i} the random effect of adjacent cells of the ith grid, w_{ij} the adjacency matrix, ρ the strength of the spatial correlation, and τ_u^2 the variance of the spatial random effect. The model allows flexible estimation of the strength of spatial correlation, ranging from no spatial correlation ($\rho = 0$) to solid spatial correlation ($\rho = 1$). Table 1 lists the explanatory variables used in this study. We used standardized variables (mean transformed to 0, standard deviation to 1) and a queen adjacency matrix. To avoid multicollinearity, we excluded convenience stores and wholesale and retail establishments with a variance inflation factor (VIF) of 8 or above. We used the CARBayes package in R for estimating the CAR model (Lee, 2013).

Results

Settings and results of NMF
We set the rank of NMF to 8, running the NMF 30 times with random initial values each time. High values of cophenetic correlation coefficients were observed, indicating a robust basis. The basis vectors (hereafter, factors) were standardized such that the sum of their elements equaled 1.

Interpretation of each factor obtained by dimension reduction
Table 2 presents the interpretation of each factor obtained using NMF.

Figure 4 illustrates the values and geographic distributions of each factor in the downtown area, which has a relatively large population distribution in the study area (Fig. 1). Factor 1 (Fig. 4a) exhibited high values at locations slightly away from the train station. The relatively higher values between 9 p.m. and 7 a.m. suggest many people staying at home during this time. In addition, the values generally decreased from 8 a.m. to 8 p.m., but the degree of decrease reduced with

Table 2 Interpretation of each factor obtained using non-negative matrix factorization (NMF)

Factor	Interpretation
1	Suburban residential area
2	Place of work and shopping with office and shopping areas
3	Areas with many students in or near high schools and universities
4	Nightlife in the downtown area, shopping destinations for young women and older women
5	Residential area with a large number of young people
6	Residential area with a stable population of people in their 30s–50s
7	Places of work for men, such as factories
8	Daytime activity places for young women and older people, such as hospitals and welfare facilities

increasing age. Therefore, we interpreted Factor 1 as a degree of the "suburban residential area." School-going and working individuals are likely to move to other areas during the day.

The area with a high Factor 2 (Fig. 4b) demonstrated high values near stations with exceptionally high ridership in Kyoto City, such as the Kyoto and Shijo Stations, which many people use daily. In this area, the population of individuals in their 20s to 50s, regardless of gender, increases from 8 a.m. to 8 p.m. but is small from 12 a.m. to 7 p.m. We interpreted the area as the degree of "place of work and shopping with office and shopping areas," as the area has many facilities used by numerous people during the daytime.

Factor 3 (Fig. 4c) showed an increase in the number of young people in their teens and 20s during the day, with a certain number remaining in the area during the night. Furthermore, since areas with high Factor 3 values have universities and high schools, we interpreted this as "areas with or around high schools and college students."

Factor 4 (Fig. 4d) demonstrated significantly higher values at central stations and downtown areas. The value of Factor 4 was zero for men between 4 a.m. and 5 p.m., while it was present for women in their 20s to 30s and 60s to 70s around 11 p.m. Therefore, the area with a high Factor 4 was considered an area with facilities that women visit during the daytime. In particular, those in their 20s and 30s were seen to remain in the area until 3 a.m. An area with this pattern implies a downtown area with many nightlife facilities such as pubs. Based on these results, we categorized the areas with high Factor 4 as "nightlife in the downtown area, shopping destinations for young and older women". This interpretation is reasonable because the number of men in their 70s and women in their 60s and 70s decreases after 6 p.m.

Factor 5 (Fig. 4e) exhibited high values around major stations, such as Kyoto and Shijo. However, it was present only from 8 p.m. to 9 a.m. for women in their 20s to 50s. Men also demonstrated the same trend but with non-zero values for the 10 a.m. to 7 p.m. periods. Therefore, we interpreted the area with high Factor 5 as a "residential area with a large number of young people."

Factor 6 (Fig. 4f) suggested that men in their 30s to 50s stayed only from 10 p.m. to 7 a.m. and likely worked during the day. However, the number of women in their 30s to 50s did not decrease significantly during the daytime, and they were consistently present. These women may be homemakers. Based on this trend, we interpreted the area with high Factor 6 as a "residential area with a stable population of people from their 30s to 50s," considering it as the residential area of the working generation.

Factor 7 (Fig. 4g) was rarely experienced by women. By contrast, the number of men in their 30s to 50s increased significantly from 7 a.m. to 6 p.m. This trend can be considered to represent an area with a large number of male workers. In addition, Factor 7 was higher in factories and central wholesale markets. Therefore, we interpreted the area with high Factor 7 as "places of work for men, such as factories."

Factor 8 (Fig. 4h) showed an increase from 8 a.m. to 4 p.m., especially among men and women in their 60s and 70s. It is also notable that women in their 20s were observed between 2 p.m. and 4 p.m. The geographic distribution of Factor 8 was higher in facilities such as hospitals and temples. Considering these factors, we interpreted the areas with a high Factor 8 as "daytime activity places for young women and older people, such as hospitals and welfare facilities."

Evaluation of spatial statistical modeling

Table 3 presents the estimation results of the spatial regression model. It is noteworthy that the absolute value of Geweke's diagnostics for all explanatory variables in the models of occurrence during daytime and nighttime is below 1.96. This suggests that the sampling of Markov chain Monte Carlo had converged for each coefficient.

For the weight components of each factor (hereafter called ambient population variables) that occurred during the daytime, all coefficients on 95% Bayesian credible intervals (CIs) included zero, making it unclear whether these variables are positively or negatively associated with the number of crime occurrences. A comparison of the deviance information criterion (DIC) for this model with and without ambient population variables revealed that the model without ambient population variables had a lower DIC and showed no utility from introducing ambient population variables (with: 378.4; without: 369.5).

In the model of occurrence during nighttime, a comparison of the DIC of the models with and without ambient population variables confirmed their usefulness. The results indicate that the DIC of the model with the ambient population variable was better (with: 1265.9; without: 1274.2; Table 3). Therefore, a better model was obtained for both crime types by considering ambient population variables. Further comparison of each crime-type model with and without spatially correlated random effects indicated that the model had a lower DIC when accounting for spatially correlated random effects (with: 1265.9; without: 1387.8; Table 3).

Explanation of each estimation

For ambient population variables, positive coefficient parameters were obtained for Factors 1, 4, 5, and 8 for nighttime indecent assaults (median coefficient parameter of Factor 1: 0.27; 95% CI [0.10, 0.46]; Factor 4: 0.13:95% CI [0.03, 0.23]; Factor 5: 0.13; 95% CI [0.00, 0.25]; Factor 8: 0.19; 95% CI [0.07, 0.31]).

Among the built environment variables (Table 3), the posterior median parameter of the building area (0.39; 95% CI [0.07, 0.73]) and total street length (0.50; 95% CI [0.25, 0.78]) indicated relatively statistically significant positive coefficient parameters. This finding suggests that these crimes may have occurred in well-developed environments. Since this study employed standardized explanatory variables, we could compare the size of each coefficient. The results indicated that the magnitudes of the coefficients for the ambient population variable were larger for Factors 1, 8, 4, and 5, in that order, and were relatively small compared to the built environment variables (building area and total street length). Nonetheless, those coefficients comprised approximately half of the coefficients for the built environment variables, which we judge to be of a magnitude that affects the risk of crime occurrence. The parameter estimation results showed that indecent assault may have a random effect with a relatively weak spatial correlation (0.24; 95% CI [0.02, 0.61]).

Figure 5 shows the number of reported crimes for each crime type and the regression analysis predictions.

(See figure on next page.)

Fig. 4 Hourly weight component values and geographic distribution of each factor (basis vector of NMF). **a** Factor 1, **b** Factor 2, **c** Factor 3, **d** Factor 4, **e** Factor 5, **f** Factor 6, **g** Factor 7, **h** Factor 8. The unit of value is 10e−3. The lines and triangles represent train lines and stations, respectively. The maps show only the central part of Kyoto City. See Appendix A in Additional file 1 for the distribution of the entire study area
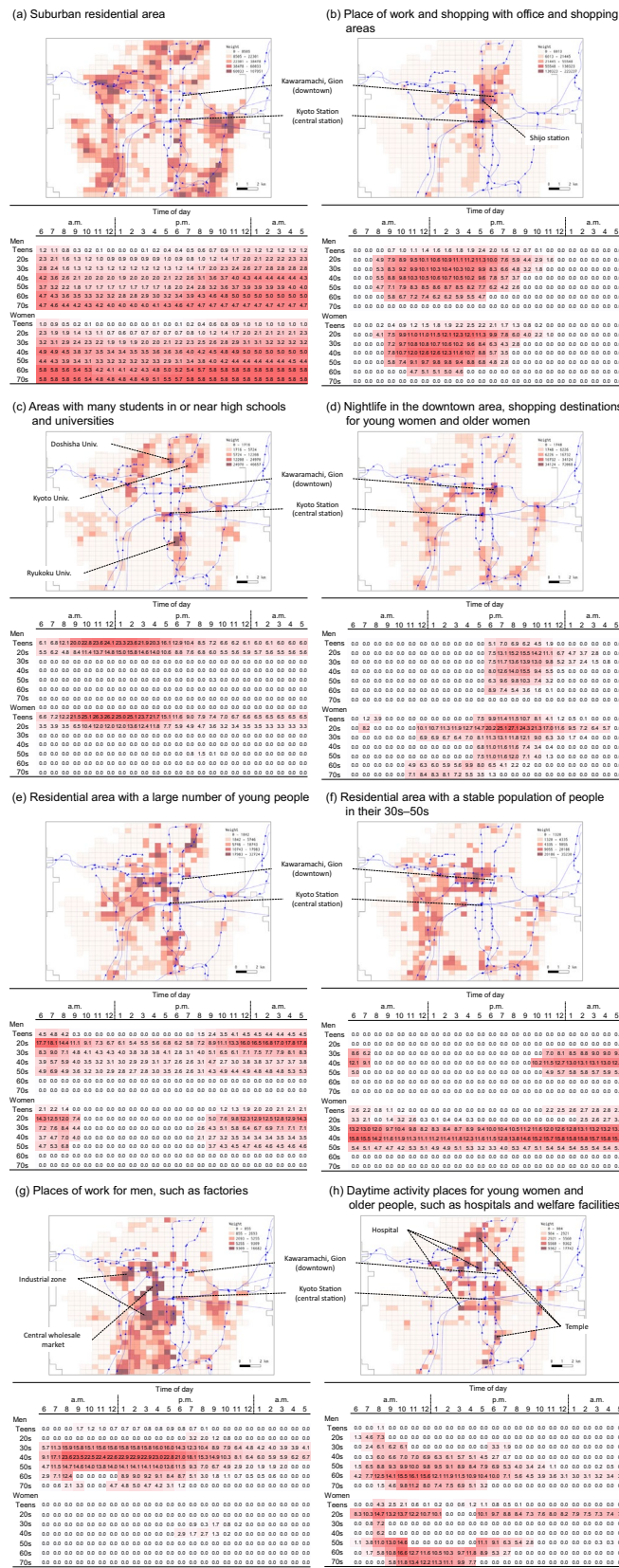
Fig. 4 (See legend on previous page.)

Table 3 Results of spatial regression

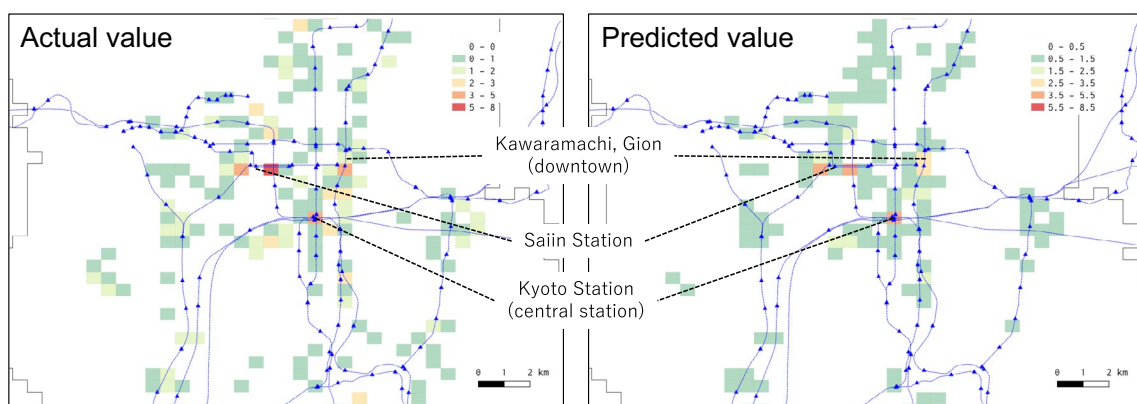|	Indecent assault			
	Daytime		Nighttime	
	Median	95% CI	Median	95% CI
Intercept	−4.23	[−4.72, −3.81]	−2.91	[−3.28, −2.61]
Ambient population variables				
Factor 1: Suburban residential area	0.13	[−0.27, 0.51]	**0.27**	**[0.10, 0.46]**
Factor 2: Place of work and shopping with office and shopping areas	−0.04	[−0.33, 0.23]	0.03	[−0.13, 0.18]
Factor 3: Areas with many students in or near high schools and universities	−0.01	[−0.37, 0.26]	0.10	[−0.05, 0.23]
Factor 4: Nightlife in the downtown area, shopping destinations for young and older women	0.12	[−0.02, 0.23]	**0.13**	**[0.03, 0.23]**
Factor 5: Residential area with a large number of young people	0.14	[−0.10, 0.34]	**0.13**	**[0.00, 0.25]**
Factor 6: Residential area with a stable population of people in their 30s–50s	−0.01	[−0.24, 0.19]	−0.06	[−0.20, 0.06]
Factor 7: Places of work for men, such as factories	−0.05	[−0.38, 0.22]	0.06	[−0.07, 0.19]
Factor 8: Daytime activity places for young women and older people, such as hospitals and welfare facilities	0.12	[−0.10, 0.33]	**0.19**	**[0.07, 0.31]**
Built-environment variables				
Building area	−0.10	[−0.85, 0.57]	**0.39**	**[0.07, 0.73]**
Total street length	**0.63**	**[0.14, 1.11]**	**0.50**	**[0.25, 0.78]**
Facility variables				
Number of train stations	**0.31**	**[0.12, 0.49]**	0.09	[−0.03, 0.21]
Number of bus stops	−0.05	[−0.28, 0.19]	0.05	[−0.08, 0.18]
Number of high schools	**0.26**	**[0.04, 0.45]**	0.10	[−0.03, 0.22]
Number of universities and junior colleges	−0.07	[−0.45, 0.16]	−0.09	[−0.25, 0.05]
Number of life-related and entertainment services	0.14	[−0.20, 0.45]	0.01	[−0.19, 0.21]
τ	**0.01**	**[0.00, 0.04]**	2.35	[1.00, 4.50]
ρ	**0.35**	**[0.02, 0.91]**	0.24	[0.02, 0.61]
DIC	378.4		1265.9	
DIC (without ambient population variables)	369.5		1274.2	
DIC (without random effect with spatial correlation)	378.8		1387.8	

Fig. 5 Geographic distributions of actual and predicted values of indecent assault during nighttime

The obtained model was able to predict high values for locations with a high number of indecent assaults (near downtown, Kyoto Station, Saiin Station, and along the railroad line). The Pearson correlation coefficient between the actual and predicted values for the study area was 0.843, a high positive correlation.
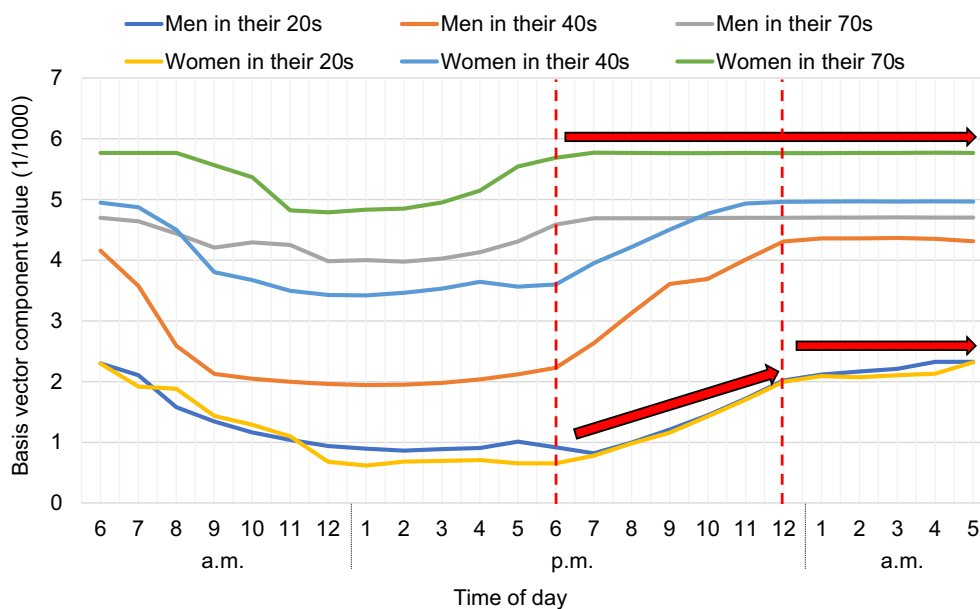
Fig. 6 Change of basis vector component values of Factor 1 over APGA

Discussion

This study applied NMF to APGA obtained from Mobile Spatial Statistics to extract hourly behavior pattern factors of people by considering gender and age groups. These factors enabled a comprehensive understanding of the temporal variation and geographic distribution of behavioral patterns. Additionally, this study employed spatial statistical modeling to analyze the relationship between these factors (ambient population variables) and crime. The results of the spatial statistical modeling showed that, for indecent assaults occurring during the daytime, for both models, coefficient parameters for ambient population variables could not be obtained, for which the 95% Bayesian credible interval of the posterior distribution of the parameter did not include zero. However, meaningful coefficient parameters for ambient population variables were estimated for nighttime indecent assaults. The model with the lowest DIC considered both spatial autocorrelated errors and ambient population variables. These results highlight the importance of considering both ambient population variables and residual spatial autocorrelation in crime risk assessment.

Figure 6 shows the hourly changes in the basis vector component values of Factor 1 by gender and age. After 12 a.m., almost all APGA remained unchanged, suggesting that people had completed their movements and stayed indoors. From 6 p.m. to 12 a.m., the working-age population increased, whereas the older population remained unchanged. Therefore, during this period, the increase in "suitable targets" and the "absence of capable guardians" of older people and other community members were

more likely. However, the degree of absence of monitors may be low because of the potentially large population in residential areas. We expected that areas characterized by Factor 1 would exhibit a pattern of victimization of people returning home. The behavioral patterns of sex offenders include selecting targets at facilities that many people use, such as train stations and convenience stores, and following them to commit crimes (National Police Agency, 2019). In areas characterized by Factor 1, it may be practical to alert people on their way home to the station nearest to their homes.

Factor 4 (Fig. 4d) shows an increase in the number of women in their 20s who are likely targets of indecent assaults between 12 a.m. and 3 a.m. This increasing trend is particularly pronounced between 6 p.m. and 10 p.m. Furthermore, both men and women in their 20s and 30s remain present between 3 a.m. and 4 a.m. the following day. Areas with a high Factor 4 are characterized by many nightlife establishments such as bars, where young men and women often amuse themselves, even late at night. Cohen et al. (1981) suggest that people with lifestyles similar to potential offenders are more likely to be victims because of increased contact with offenders (principle of homogamy). Schwartz and Pitts (1995) also indicate that women who are intoxicated are more likely to be targets of sexual crimes and approached in bars and restaurants, consistent with our results. Although we hesitate to label a male in his 20s as a potential offender, the absence of a capable guardian in this area, as it reacts strongly only to the presence of men and women in their 20s and 30s during nighttime, suggests a high opportunity for crime

Table 4 Interpretation of the vulnerability of areas with high factors in terms of the routine activity approach and proposed crime prevention measures

Factor	Viewpoint of routine activity approach		Possible occurrence timing	Proposed Crime prevention measures
	Suitable target	Absence of capable guardians		
1	✔		Going home	Alerts at train stations and convenience stores
4	✔	✔	At nighttime	Increased surveillance (e.g., install security cameras)
5	✔		Going home	Alerts at train stations and convenience stores
8	✔	✔	At nighttime/Going home	Increased surveillance (e.g., install security cameras)

because "suitable targets" and the "absence of capable guardians" are more likely to have coexisted. For example, in neighborhoods characterized by this factor, victimization is more likely to occur on the way home from a nightlife facility when capable guardians are absent. In such areas, it may be essential to evaluate potential operationally competent defenders and implement measures to reduce the risk of crime based on the results (Hollis et al., 2013).

Regarding Factor 5 (Fig. 4e), the ambient population of men and women in their 20s to 50 s increases from 9 p.m. to 12 a.m. This period also indicates an increase in "suitable targets." This result is similar to Factor 4, indicating a direct correlation with an increased target population. However, the factor's value remains constant from 9 p.m. to 6 a.m., implying that individuals in this demographic group tend to stay indoors during these hours. Factor 5 is likely to occur in areas with a significant population of individuals in their 20 s. Considering the potential increase in the number of women in their 20 s between 9:00 p.m. and 6:00 a.m., as with Factor 1, we presume that this is an area where individuals are more likely to be victims of crime on their way home.

Factor 8 (Fig. 4h) is characterized by an increase in the number of men in their 40s–70s and women in their 60s–70s from 8 a.m. to 6 p.m. After 5:00 p.m., women in their 20s are present, but after 10 p.m., only men in their 60s and women in their 20s remain, indicating the latent presence of women in their 20s. Therefore, we consider Factor 8 as the simultaneous establishment of "suitable targets" and the "absence of capable guardians." This area could be one where women in their 20s are known to reside during nighttime. For motivated offenders, the area may exhibit characteristics similar to those of crime attractors (Brantingham & Brantingham, 1995).

Based on the previous discussion, Factors 1 and 5 exhibit high values in residential areas, confirming an increase in the population of potential targets when returning home. This suggests a pattern where targets returning home near residential neighborhoods are likely

to be criminalized. Hanaoka (2016) analyzed the relationship between snatch-and-run farming and the ambient population in Osaka, Japan, and demonstrated that at night (defined as 7 p.m. to 5 a.m.), the risk increases with the ambient population density. This result is consistent with the characteristics observed in the patterns of Factors 1 and 5. However, Factors 4 and 8 show women in their 20s during the evening and late at night, but the presence of different genders and age populations is almost nonexistent. Furthermore, considering that outdoor indecent assaults are more likely to occur from evening to late at night, the time and space where these factors are high can be considered an environment where "suitable targets" and "absence of capable guardians" are more likely to be established. We believe that the installation of security cameras and other security measures to increase surveillance in these areas will be useful. Table 4 summarizes our discussion and shows that "suitable targets" and "absence of capable guardians" tend to exist differently in Factors 1 and 5 and in Factors 4 and 8. Appropriate crime prevention measures have been suggested for each case.

As discussed above, this study is the first to demonstrate the utility of using APGA to analyze the risk of crime occurrence; APGA allowed us to discuss in detail the spatio-temporal characteristics of crime formation along with a routine activity approach. Considering that crime risks differ by the individual attributes of victims, APGA may provide important clues for understanding the mechanisms of crime occurrence patterns in a city.

Limitations and future research

Although the present study reveals important findings, it has several limitations. It is desirable to use outdoor population data because street crime victims are likely to be located outdoors. However, Mobile Spatial Statistics cannot distinguish between indoor and outdoor populations. To accurately analyze crimes occurring on the street, the outdoor ambient population must be considered. In addition, the study used the average value of the time of day

calculated from the population data for 1 year. Therefore, changes in the population owing to events and seasons were not considered. The study did not consider the transportation modes of victims either, as victims are presumed to walk or bike (Harada, 2020), yet Mobile Spatial Statistics do not identify the mode of transportation. These issues should be addressed in future studies.

This study typified each area by APGA and analyzed the relationship between the number of crimes during the day and at night. It would be desirable to examine the relationship between the number of crimes per hour and the APGA analysis. However, the low number of crimes per hour poses challenges in obtaining stable statistical analysis results (Felson & Poulsen, 2003). In the future, efforts should be made to resolve this issue and elucidate the relationship between detailed hourly crime occurrences and APGA.

Although this study analyzed the main effects of ambient population variables, built environment variables, and facility variables, the risk of crime may increase where certain factors within each factor are mixed (Adams et al., 2015). Models that include interaction terms are more difficult to interpret; however, it is important to understand the interaction terms related to the risk of crime occurrence. Therefore, future research should adopt an expanded approach to address this issue.

Conclusions

This study evaluated the relationship between APGA obtained from Mobile Spatial Statistics and the number of reported indecent assaults. First, NMF was applied to APGA to extract factors related to temporal changes in gender and age groups, and regional characteristics were organized for each factor. A spatial statistical model of the number of reported crimes was created using these factors, allowing us to quantitatively identify the relationship between the regional characteristics characterized by APGA and the risk of crime. This analytical approach enables an analysis of how changes in the ambient population by gender and age in each area affect crime risk. The study discussed these results in light of the routine activity approach and proposed effective crime prevention measures for each area. For example, areas where young people congregate at night (e.g., downtown areas) were identified, and crime prevention measures suggesting alerts for young women returning home from those areas at night could be effective strategies to reduce crime.

Abbreviations

APGA   Ambient Population by Gender and Age Group
CAR   Conditional Auto Regressive

CI   Credible interval
DIC   Deviance information criterion
KL   Kullback–Leibler
NMF   Non-negative matrix factorization
NNDSVD Non-negative double singular value decomposition
nsNMF   Non-smooth non-negative matrix factorization
PT   Person trip
SNS   Social network service
VIF   Variance inflation factor

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40163-024-00205-x.

> **Additional file 1:** Geographic distributions of each factor (basis vector of NMF) in entire study area. **Fig. A1.** Factor1. **Fig. A2.** Factor2. **Fig. A3.** Factor 3. **Fig. A4.** Factor 4. **Fig. A5.** Factor 5. **Fig. A6.** Factor 6. **Fig. A7.** Factor 7. **Fig. A8.** Factor 8.

Declarations

Competing interests
The authors declare that they have no competing interests.

References
Adams, W., Herrmann, C. R., & Felson, M. (2015). Crime, transportation and malignant mixes. In V. Ceccato & A. Newton (Eds.), *Safety and security in transit environments. Crime prevention and security management* (pp. 181–195). Palgrave Macmillan. https://doi.org/10.1057/9781137457653_10

Amemiya, M., Ohyama, T., & Shimada, T. (2018). Analyzing sex crime victimization risk: Who, when, and where is riskiest? *Reports of the City Planning Institute of Japan, 16*, 297–301.

Andresen, M. A. (2006). Crime measures and the spatial analysis of criminal activity. *British Journal of Criminology, 46*(2), 258–285. https://doi.org/10.1093/bjc/azi054

Andresen, M. A. (2011). The ambient population and crime analysis. *The Professional Geographer, 63*(2), 193–212. https://doi.org/10.1080/00330124.2010.547151

Andresen, M. A., & Jenion, G. W. (2010). Ambient populations and the calculation of crime rates and risk. *Security Journal, 23*(2), 114–133. https://doi.org/10.1057/sj.2008.1

Boggs, S. L. (1965). Urban crime patterns. *American Sociological Review, 30*(6), 899–908. https://doi.org/10.2307/2090968

Bogomolov, A., Lepri, B., Staiano, J., Letouzé, E., Oliver, N., Pianesi, F., & Pentland, A. (2015). Moves on the street: Classifying crime hotspots using aggregated anonymized data on people dynamics. *Big Data, 3*(3), 148–158. https://doi.org/10.1089/big.2014.0054

Boivin, R., & Felson, M. (2018). Crimes by visitors versus crimes by residents: The influence of visitor inflows. *Journal of Quantitative Criminology, 34*, 465–480. https://doi.org/10.1007/s10940-017-9341-1

Brantingham, P. J., & Brantingham, P. L. (1984). *Patterns in crime*. MacMillan.

Brantingham, P., & Brantingham, P. (1995). Criminality of place: Crime generators and crime attractors. *European Journal on Criminal Policy and Research, 3*(3), 5–26. https://doi.org/10.1007/BF02242925

Brunsdon, C., Corcoran, J., & Higgs, G. (2007). Visualizing space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems, 31*(1), 52–75. https://doi.org/10.1016/j.compenvurbsys.2005.07.009

Buil-Gil, D., Medina, J., & Shlomo, N. (2021). Measuring the dark figure of crime in geographic areas: Small area estimation from the Crime Survey for England and Wales. *The British Journal of Criminology, 61*(2), 364–388. https://doi.org/10.1093/bjc/azaa067

Cabinet Office. (2012). Retrieved from https://www.gender.go.jp/kaigi/senmon/boryoku/houkoku/index_hbo07.html.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review, 44*(4), 588–608. https://doi.org/10.2307/2094589

Cohen, L. E., Kluegel, J. R., & Land, K. C. (1981). Social inequality and predatory criminal victimization: An exposition and test of a formal theory. *American Sociological Review, 46*(5), 505–524. https://doi.org/10.2307/2094935

Collia, D. V., Sharp, J., & Giesbrecht, L. (2003). The 2001 National Household Travel Survey: A look into the travel patterns of older Americans. *Journal of Safety Research, 34*(4), 461–470. https://doi.org/10.1016/j.jsr.2003.10.001

Felson, M., & Boivin, R. (2015). Daily crime flows within a city. *Crime Science, 4*, 31. https://doi.org/10.1186/s40163-015-0039-0

Felson, M., & Poulsen, E. (2003). Simple indicators of crime by time of day. *International Journal of Forecasting, 19*(4), 595–601. https://doi.org/10.1016/S0169-2070(03)00093-1

Gillis, N. (2020). *Nonnegative matrix factorization*. SIAM.

Haberman, C. P., & Ratcliffe, J. H. (2015). Testing for temporally differentiated relationships among potentially criminogenic places and census block street robbery counts. *Criminology, 53*(3), 457–483. https://doi.org/10.1111/1745-9125.12076

Hamed, M. M., & Mannering, F. L. (1993). Modelling traveler's postwork activity involvement: Toward a new methodology. *Transportation Science, 27*(4), 381–394. https://doi.org/10.1287/trsc.27.4.381

Hanaoka, K. (2016). New insights on relationships between street crimes and ambient population: Use of hourly population data estimated from mobile phone users' locations. *Environment and Planning B: Urban Analytics and City Science, 45*(2), 295–311. https://doi.org/10.1177/0265813516672454

Harada, Y. (2020). Estimating victimization risks of purse-snatching using a five-year data of Tokyo metropolitan areas. *Sociological Theory and Methods, 35*(1), 62–75. https://doi.org/10.11218/ojjams.35.62

Havet, N., Bayart, C., & Bonnel, P. (2021). Why do gender differences in daily mobility behaviours persist among workers? *Transportation Research Part a: Policy and Practice, 145*, 34–48. https://doi.org/10.1016/j.tra.2020.12.016

Hipp, J. R., Bates, C., Lichman, M., & Smyth, P. (2019). Using social media to measure temporal ambient population: Does it help explain local crime rates? *Justice Quarterly, 36*(4), 718–748. https://doi.org/10.1080/07418825.2018.1445276

Hollis, M. E., Felson, M., & Welsh, B. C. (2013). The capable guardian in routine activities theory: A theoretical and conceptual reappraisal. *Crime Prevention and Community Safety, 15*(1), 65–79. https://doi.org/10.1057/cpcs.2012.14

Jacobs, J. (1961). *The death and life of great American cities*. Random House.

Kyoto Prefecture Police. (2020). Retrieved from https://www.pref.kyoto.jp/fukei/anzen/toke/documents/r2_hanzaitokei.pdf.

Lawson, A. B. (2021). *Using R for Bayesian spatial and spatiotemporal health modeling*. Chapman & Hall/CRC.

Lee, D. (2013). CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software, 55*(13), 1–24. https://doi.org/10.18637/jss.v055.i13

MacDonald, J. M., & Lattimore, P. K. (2010). Count models in criminology. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 683–698). Springer.

Malleson, N., & Andresen, M. A. (2015). Spatio-temporal crime hotspots and the ambient population. *Crime Science, 4*(1), 10. https://doi.org/10.1186/s40163-015-0023-8

Masso, A., Silm, S., & Ahas, R. (2019). Generational differences in spatial mobility: A study with mobile phone data. *Population, Space and Place, 25*(2), e2210. https://doi.org/10.1002/psp.2210

Ministry of Justice (2020). Retrieved from https://www.moj.go.jp/content/001316208.pdf.

Morgan, R. E., & Thompson, A. (2021). *Criminal victimization, 2020. Washington, DC: National Crime Victimization Survey, Bureau of Justice Statistics*. Retrieved from https://bjs.ojp.gov/sites/g/files/xyckuh236/files/media/document/cv20.pdf.

National Police Agency. (2008). Retrieved from https://www.npa.go.jp/hanzaihigai/shien/handbook/html/2-2-4.html.

National Police Agency. (2019). White paper on crime 2019. Retrieved from https://hakusyo1.moj.go.jp/jp/66/nfm/mokuji.html.

National Police Agency. (2020). Retrieved from https://www.npa.go.jp/toukei/seianki/R01/r01keihouhantoukeisiryou.pdf.

Newman, O. (1972). *Defensible space: Crime prevention through urban design*. MacMillan.

Newton, A., & Felson, M. (2015). Editorial: Crime patterns in time and space: the dynamics of crime opportunities in urban areas. *Crime Science*. https://doi.org/10.1186/s40163-015-0025-6

Newton, A., Felson, M., & Bannister, J. (2021). Introduction to the special issue: Urban mobility and crime patterns. *European Journal on Criminal Policy and Research, 27*, 307–311. https://doi.org/10.1007/s10610-021-09501-7

Okajima, I., Tanaka, S., Terada, M., Ikeda, D., & Nagata, T. (2013). Supporting growth in society and industry using statistical data from mobile terminal networks – Overview of mobile spatial statistics. *NTT DOCOMO Technical Journal, 14*(3), 16–23.

Pascual-Montano, A., Carazo, J. M., Kochi, K., Lehmann, D., & Pascual-Marqui, R. D. (2006). Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(3), 403–415. https://doi.org/10.1109/TPAMI.2006.60

Perkins, C. A. (1997). Age patterns of victims of serious violent crime. Bureau of Justice Statistics. Retrieved from https://bjs.ojp.gov/content/pub/pdf/apvsvc.pdf.

Pina-Sánchez, J., Brunton-Smith, I., Buil-Gil, D., & Cernat, A. (2023). Exploring the impact of measurement error in police recorded crime rates through sensitivity analysis. *Crime Science*. https://doi.org/10.1186/s40163-023-00192-5

Schwartz, M. D., & Pitts, V. L. (1995). Exploring a feminist routine activities approach to explaining sexual assault. *Justice Quarterly, 12*(1), 9–31. https://doi.org/10.1080/07418829500092551

Sherman, L. W., Gartin, P. R., & Buerger, M. E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology, 27*(1), 27–56. https://doi.org/10.1111/j.1745-9125.1989.tb00862.x

Statistics Bureau of Japan. Retrieved from https://www.stat.go.jp/english/data/mesh/05.html.

Terada, M., Nagata, T., & Kobayashi, M. (2013). Population estimation technology for Mobile Spatial Statistics. *NTT DOCOMO Technical Journal, 14*(3), 10–15.

Tucker, R., O'Brien, D. T., Ciomek, A., Castro, E., Wang, Q., & Phillips, N. E. (2021). Who 'tweets' where and when, and how does it help understand crime rates at places? Measuring the presence of tourists and commuters in ambient populations. *Journal of Quantitative Criminology, 37*, 333–359. https://doi.org/10.1007/s10940-020-09487-1

Weisburd, D. (2015). The law of crime concentration and the criminology of place. *Criminology, 53*(2), 133–157. https://doi.org/10.1111/1745-9125.12070

Publisher's Note