Supplements

Appendix A

A significance test for discovery of global co-location patterns based on the natural neighborhood

The non-parametric significance test developed by Deng *et al*. (2017) can be extended to identify statistically significant global co-location patterns based on the natural neighborhood. To construct the significance test, the null model should be defined firstly. The null model assumes that the distributions of different spatial features are mutually independent. To construct this null model, two hypotheses should be satisfied: (i) similar spatial distribution of each spatial feature as that of the observed dataset; (iii) same number of instances for each feature as in the observed data. To completely describe the spatial distribution of certain spatial feature, three summary statistics were selected, i.e., the pair-correlation function $g(r)$ (a second-order summary statistic), the nearest-neighbor distribution function $D(r)$ (a nearest-neighbor summary statistic) and the spherical contact distribution function $Hs(r)$ (a morphological summary statistic) (Diggle 2003). Under the null model, for each spatial feature, a number of simulated datasets are generated by using a pattern reconstruction method that produces stochastic replicates of the observed dataset that closely approximate the three summary statistics of the observed dataset (Wiegand *et al*. 2013).

By using the pattern reconstruction method, for each spatial feature, N simulated datasets are obtained. The natural neighborhood is used to construct the neighbor relationships among instances of different spatial features for both observed and simulated datasets. The participation index (PI) is used as the test statistic. The p-value of a candidate co-location pattern C_i can be calculated as follows:

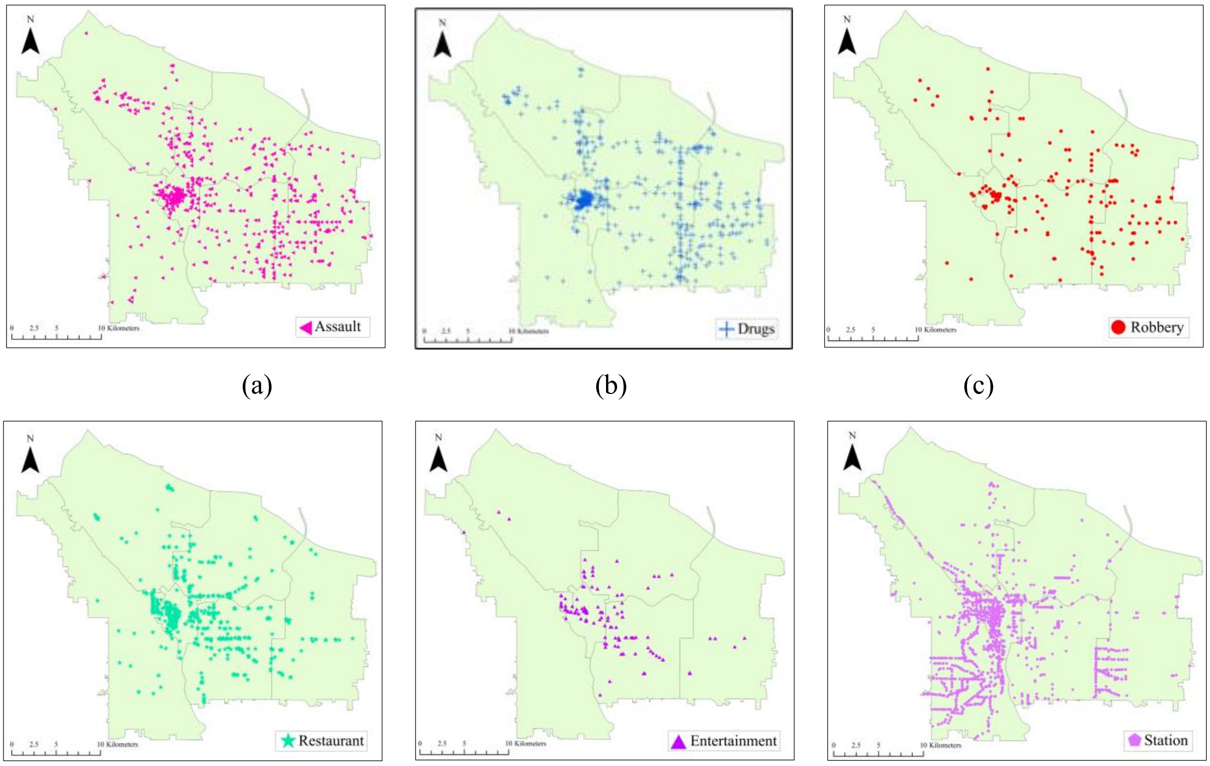
$$p - value(C_i) = \frac{\#(PI_{null}^{n}(C_i) \geq PI_{obs}(C_i)) + 1}{N+1}, n = 1,2,\ldots,N \tag{1}$$

Where PI_{null}^{n} represents the value of participation index calculated for simulated datasets, PI_{obs} represents the value of participation index calculated for observed dataset, "#" represents the number of simulations in which PI_{null}^{n} is not less than PI_{obs}. Given a significance level a, if *p-value*(C_i) ≤ a, then the null hypothesis is rejected, and the pattern C_i is identified as a statistically significant global co-location pattern. Experiments on both simulated and real-life datasets used in this study show that the global co-location patterns detected by setting prevalence threshold T to 0.5 are identical to that discovered by the significance test (a= 0.05).

Appendix B

Experiments on the Portland datasets

The Portland datasets contain crimes and urban facilities collected from January to March 2014 in the city of Portland. These datasets were obtained from the CivicApps Data Catalog (http://www.civicapps.org/datasets). It has been known that the spatial distribution of crime occurrence can be shaped by the presence or absence of urban facilities (He *et al*. 2020). Therefore, discovery of co-location relationships between crimes and facilities may be helpful for crime prevention and urban planning. In this study, three types of crimes were selected: Assault (740 instances), Drugs (723 instances) and Robbery (188 instances). Three types of urban facilities were selected: Restaurant (715 instances), Entertainment (131 instances) and Station (1764 instances). The spatial distributions of the crimes and urban facilities are displayed in Figure S.1.



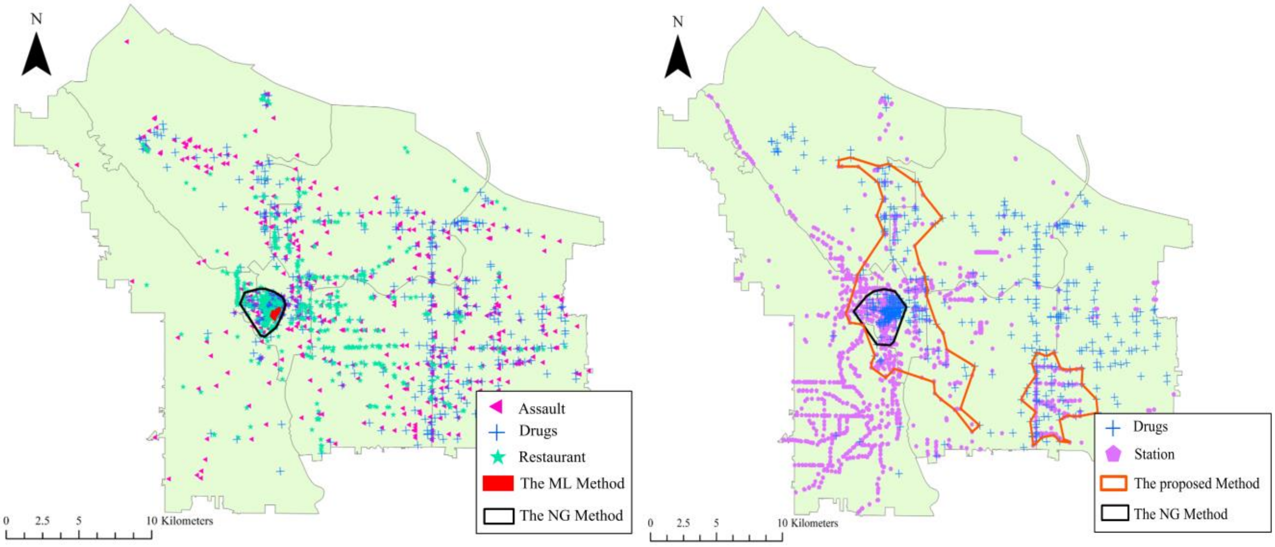
(a) (b) (c)

(d) (e) (f)

Figure S.1. Spatial distributions of crimes and facilities: (a) Assault; (b) Drugs; (c) Robbery; (d) Restaurant (Misc.); (e) Entertainment; (f) Station

For the ML and NG method, the distance threshold was set to 400m, suggested by He *et al*. (2020). In Table S.1, the multilevel co-location patterns discovered by the NG, ML, and the proposed method are listed. The proposed method discovers 16 global co-location patterns and 26 local co-location patterns. We only show the ten most prevalent global and local co-location patterns in Table S.1. All the detected multilevel co-location patterns were evaluated based on the rules given in Section 6.2. Table S.2 summarizes the assessments of precision and recall of co-location patterns discovered by the three methods. It can be found that many global co-location patterns are missed or wrongly identified as local co-location patterns by the ML and NG method. At the same time, some local co-location patterns are neglected by the ML and NG method. In Figure S.2, some co-location patterns discovered by the three methods are shown. In Figure S.2(a), we can find that the ML and NG method only identify the co-location relationship among Assault, Drugs and Restaurant in the downtown; however, these three features are frequently co-located throughout the city. In Figure S.2(b), the ML method does not discover the local co-location pattern {Drugs, Station}, while the NG method only discovers one locality of {Drugs, Station} in the downtown. We can infer that the ML and NG method do not construct instances of candidate co-location patterns in sparse regions. Therefore, some co-location patterns were missed or wrongly identified.

Based on the above analysis, the advantage of the proposed method can be well illustrated. The discovered co-location patterns make a positive contribution to crime prevention and urban planning, and are useful for investigating the association between criminal behavior and the environment.



(a) (b)

Figure S.2. A comparison of the co-location patterns discovered by the NG, ML and the proposed method: (a) Global co-location pattern {Assault, Drugs, Restaurant} neglected by NG and ML; (b) Localities of {Drugs, Station} neglected by the ML and NG.

Table. S.1. The multilevel co-location patterns discovered by the NG, ML and the proposed method

Multi-level co-location patterns	The Proposed Method					The ML method					The NG method		
	Level	PI_{global}	PI_{max}	PI_{min}	$N_{locality}$	Level	PI_{global}	PI_{max}	PI_{min}	$N_{locality}$	RPI_{max}	RPI_{min}	$N_{locality}$
{ Restaurant, Robbery}	Global	0.66	-	-	1	Local	0.35	0.76	0.76	1	0.13	0.13	1
{Entertainment, Robbery}	Global	0.61	-	-	1	Local	0.22	1.00	1.00	1	-	-	-
{Assault, Station}	Global	0.61	-	-	1	Local	0.42	0.63	0.63	1	0.15	0.15	1
{Assault, Restaurant}	Global	0.60	-	-	1	Global	0.52	-	-	1	0.23	0.23	1
{Drugs, Restaurant}	Global	0.56	-	-	1	Global	0.65	-	-	1	0.34	0.34	1
{Drugs, Entertainment}	Global	0.56	-	-	1	Local	0.46	0.69	0.69	1	0.24	0.24	1
{Drugs, Entertainment, Robbery}	Global	0.54	-	-	1	-	-	-	-	-	-	-	-
{Assault, Drugs, Restaurant}	Global	0.54	-	-	1	Local	0.40	0.78	0.61	2	0.22	0.22	1
{Assault, Restaurant, Robbery}	Global	0.54	-	-	1	-	-	-	-	-	0.13	0.13	1
{Assault, Entertainment}	Global	0.54	-	-	1	Local	0.31	0.91	0.91	1	0.20	0.20	1
{Drugs, Restaurant, Station}	Local	0.34	0.97	0.62	2	-	-	-	-	-	0.12	0.12	1
{Restaurant, Robbery, Station}	Local	0.49	0.94	0.59	2	Local	0.14	1.00	1.00	1	-	-	-
{Drugs, Restaurant, Robbery, station}	Local	0.27	0.92	0.53	2	-	-	-	-	-	-	-	-
{Drugs, Station}	Local	0.46	0.92	0.77	2	-	-	-	-	-	0.14	0.14	1
{Drugs, Robbery, Station}	Local	0.38	0.89	0.68	2	-	-	-	-	-	-	-	-
{Assault, Restaurant, Station}	Local	0.45	0.86	0.77	2	-	-	-	-	-	0.14	0.14	1
{Assault, Restaurant, Robbery, Station}	Local	0.33	0.86	0.57	2	-	-	-	-	-	-	-	-
{Assault, Drugs, Station}	Local	0.41	0.84	0.73	2	-	-	-	-	-	0.11	0.11	1
{Assault, Drugs, Robbery, Station}	Local	0.34	0.83	0.62	2	-	-	-	-	-	-	-	-
{Assault, Drugs, Restaurant, Station}	Local	0.30	0.82	0.58	2	-	-	-	-	-	0.11	0.11	1

PI_{global}: participation index calculated at global level; PI_{min}: minimal participation index calculated at local level; PI_{max}: maximal participation index calculated at local level; $N_{locality}$: number of localities of a co-location pattern; RPI_{min}: minimal regional participation index; RPI_{max}: maximal regional participation index

Table. S.2. Evaluation of the multilevel co-location patterns discovered by the NG, ML and the proposed method

Multi-level co-location patterns	The Proposed Method			The ML method			The NG method	
	Level	Precision	Recall	Level	Precision	Recall	Precision	Recall
{Robbery, Restaurant}	Global	1	1	Local	0	0	0	0
{Entertainment, Robbery}	Global	1	1	Local	0	0	-	0
{Assault, Station}	Global	1	1	Local	0	0	0	0
{Assault, Restaurant}	Global	1	1	Global	1	1	0	0
{Drugs, Restaurant}	Global	1	1	Global	1	1	0	0
{Drugs, Entertainment}	Global	1	1	Local	0	0	0	0
{Drugs, Entertainment, Robbery}	Global	1	1	-	-	0	-	0
{Assault, Drugs, Restaurant}	Global	1	1	Local	0	0	0	0
{Assault, Restaurant, Robbery}	Global	1	1	-	-	0	0	0
{Assault, Entertainment}	Global	1	1	Local	0	0	0	0
{Drugs, Restaurant, Station}	Local	1	1	-	-	0	1	0.5
{Restaurant, Robbery, Station}	Local	1	1	Local	1	0.5	-	0
{Drugs, Restaurant, Robbery, station}	Local	1	1	-	-	0	-	0
{Drugs, Station}	Local	1	1	-	-	0	1	0.5
{Drugs, Robbery, Station}	Local	1	1	-	-	0	-	0
{Assault, Restaurant, Station}	Local	1	1	-	-	0	1	0.5
{Assault, Restaurant, Robbery, Station}	Local	1	1	-	-	0	-	0
{Assault, Drugs, Station}	Local	1	1	-	-	0	1	0.5
{Assault, Drugs, Robbery, Station}	Local	1	1	-	-	0	-	0
{Assault, Drugs, Restaurant, Station}	Local	1	1	-	-	0	1	0.5

Reference

Barua, S. and Sander, J., 2014. Mining statistically significant co-location and segregation patterns. *IEEE Transactions on Knowledge and Data Engineering*, 26(5), 1185-1199.

Deng, M., *et al*., 2017. Multi-scale approach to mining significant spatial co-location patterns. *Transactions in GIS*, 22(5), 1023-1039.

Diggle, P.J., 2003. *Statistical analysis of spatial point patterns*. 2nd ed. London: Edward Arnold.

He, Z.J., *et al*., 2020. Discovering the joint influence of urban facilities on crime occurrence using spatial co-location pattern mining. *Cities*, 99, 102612.

Wiegand, T., *et al*., 2013. A systematic comparison of summary characteristics for quantifying point patterns in ecology. *Ecography*, 36, 92-103