CSIS Discussion Paper No. 149

Descriptive measures of point distributions summarized over the entire spatial scale

Yukio Sadahiro

July 2017

Center for Spatial Information Science, The University of Tokyo
5-1-5, Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan
Phone: +81-471-36-4310
Fax: +81-3-5841-8521
sada@csis.u-tokyo.ac.jp

Abstract

Keywords: point distributions, spatial scale, summary measures, visualization

Visualization is a first step of exploratory point pattern analysis. It is often recommended to visualize point distributions at various spatial scales from local to global, because it permits us to detect spatial patterns of different scales that are useful for building research hypotheses. Visual analysis, however, takes a considerable amount of time especially when numerous maps are explored. Five measures are developed in this paper that summarize the properties of point distributions over the entire range of spatial scales. The maps of the measures help us to capture efficiently the overall spatial pattern of point distributions. Numerical experiments and applications to real data analysis are performed to test the validity of the proposed measures. The results reveal the effectiveness of the measures as well as their shortcomings to be revolved in future research.

1. Introduction

Visualization is a first step of exploratory point pattern analysis (Kovalerchuk and Schwing (2005); Oyana and Margai (2015)). Dot maps indicate the exact location of points using map symbols. Quadrat method generates lattices, counts the number of points in each cell, and visualizes them as grid maps. Kernel smoothing converts point distributions into smooth surfaces represented by hill-shaded and contour maps (Silverman (1986); Scott (2015)). These maps provide us a lot of useful information on the spatial pattern of point distributions that are helpful for building research hypotheses.

Spatial scale plays a key role in quadrat method and kernel smoothing. Large cells and windows yield smooth maps that reveal the global structure of point distributions. Small ones are appropriate for exploring the local pattern since they emphasize the details of point distributions. Analysts are recommended to try various spatial scales from local to global to capture spatial patterns of a wide variety of scales.

Recent GIS software permit us to easily visualize point distributions with changing the spatial scale of visualization. They generate grid maps and kernel surfaces of different scales very rapidly. Visual analysis, however, takes a considerable amount of time and is inevitably subjective when treating numerous maps generated at various scales. It is not easy to evaluate and memorize many spatial patterns in a consistent and objective way. The difficulty increases when many distributions are analyzed simultaneously as is often seen in ecology and biology.

One approach to this problem is to choose an optimal scale in visual analysis. The optimal window width is proposed in kernel smoothing that minimizes the difference between point distribution and kernel surface, or more precisely, approximately minimizes the mean integrated square error in density estimation (Silverman (1986); Sheather (2004); Scott (2015), Sheather (2004)). Spatial scan statistics define the optimal window as maximizing the difference in point density between inside and outside the window (Kulldorff (1997); Tango and Takahashi (2005)). Visualization at an optimal scale, unfortunately, inevitably conceals the spatial patterns observed at the other scales. Point clusters observed at a certain scale may not be visualized at the optimal scale. Moreover, the optimal scale depends on the spatial pattern of point distributions. Different distributions have different optimal scales so that the maps generated at their own optimal scales are not comparable with each other.

Summary measures of point distributions are an alternative to optimal scaling that also reduce the time of exploratory analysis. χ^2-statistic used in quadrat method evaluates the nonuniformity of point distributions. *K*-function evaluates the clustering degree of points (Ripley 2005). Spatial median and the number of peaks are basic measures in kernel smoothing. These measures summarize the properties of point distributions at a certain scale represented as cell size, circle radius, and window width. Measures calculated at various scales are visualized as graphs whose x-axis is spatial scale that permit us to grasp efficiently the properties of the distributions over different scales.

One of the drawbacks of summary measures is that they do not convey directly the spatial information on point distributions. For instance, summary measures do not indicate the location, size, and shape of point clusters. They are not effective to discuss the spatial relationship between point distributions. Some measures do not distinguish the same spatial pattern at different location since they are insensitive to the congruent transformation. To complement these measures, this paper proposes a new method of visualizing point distributions for exploratory spatial analysis. The method calculates summary measures that outline the properties of point distributions over the entire range of scales at each location and visualizes their spatial distributions as maps. These measures and the aforementioned ones are complementary to each other, i.e., the former encapsulate the properties of point distributions over the entire range of scales at each location, while the latter summarize the properties over the whole area at each spatial scale. Section 2 defines summary measures of point distributions. Measures for evaluating a single distribution of points are proposed, which are further extended to treat multiple distributions. Section 3 performs numerical experiments to test the validity of the proposed measures. Section 4 applies the measures to the analysis of real datasets. Section 5 summarizes the conclusions with discussion.

2. Methods

2.1 Evaluation of a single distribution of points

Suppose there are *n* points Π={P_1, P_2, ..., P_n} in region *S*. Let $\mathbf{z}_i$ be the location of *i*th point. Function $F_0(\mathbf{x}, w_0; \Pi)$ is the density of points within distance w_0 at location **x**:

$$F_0\left(\mathbf{x}, w_0; \Pi\right) = \frac{1}{\pi w_0^2} \sum_i \delta\left(\left|\mathbf{x} - \mathbf{z}_i\right|, w_0\right), \tag{1}$$

where δ(|**x-z**|, w_0) is a binary function given by

$$\delta\left(\left|\mathbf{x} - \mathbf{z}_i\right|, w_0\right) = \begin{cases} 1 & \text{if } \left|\mathbf{x} - \mathbf{z}_i\right| \le w_0 \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

Function $F_0(\mathbf{x}, w_0; \Pi)$ becomes large where points are clustered, while it is small where points are sparse. Distance w_0 works as a scale parameter at which $F_0(\mathbf{x}, w_0; \Pi)$ is evaluated. A large w_0 considers points in a wide area, and consequently, gives a global view on point distributions. A small w_0 focuses on the local variation of point distribution. Function $F_0(\mathbf{x}, w_0; \Pi)$ becomes smooth if w_0 is large while a small w_0 yields a rough function.

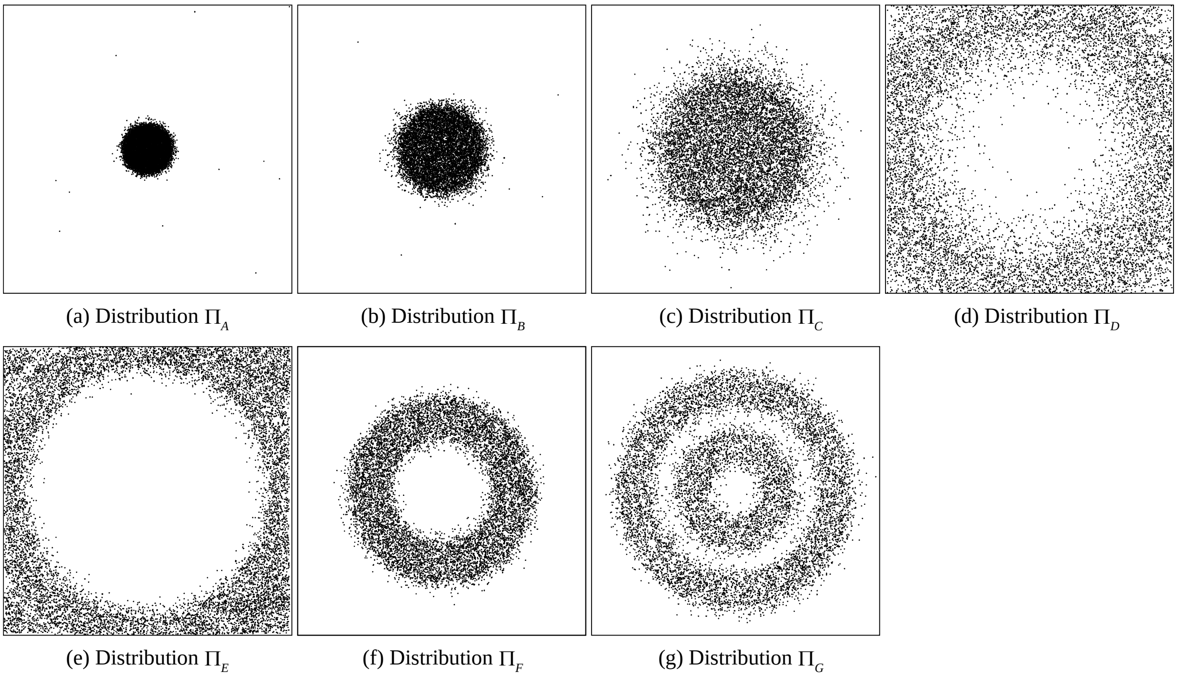
Standardized form of $F_0(\mathbf{x}, w_0; \Pi)$ is defined as

$$\begin{aligned} f_0\left(\mathbf{x}, w_0; \Pi\right) &= \frac{F_0\left(\mathbf{x}, w_0; \Pi\right)}{\int_{\mathbf{x}\in S} F_0\left(\mathbf{x}, w_0; \Pi\right) \mathrm{d}\mathbf{x}} \\ &= \frac{\sum_i \delta\left(\left|\mathbf{x} - \mathbf{z}_i\right|, w_0\right)}{\int_{\mathbf{x}\in S} \sum_i \delta\left(\left|\mathbf{x} - \mathbf{z}_i\right|, w_0\right) \mathrm{d}\mathbf{x}} \end{aligned}.$$

Function $f_0(\mathbf{x}, w_0; \Pi)$ indicates the relative density of points while $F_0(\mathbf{x}, w_0; \Pi)$ is the absolute density. The former is useful for comparing the spatial patterns between different distributions.

Both $F_0(\mathbf{x}, w_0; \Pi)$ and $f_0(\mathbf{x}, w_0; \Pi)$ are the functions of location $\mathbf{x}$ and scale w_0. Given a certain w_0, we can create the maps of these functions. If location $\mathbf{x}$ is specified, we can visualize the functions as graphs whose x-axis is w_0. Figure 1 shows seven distributions of points labelled Π_A-Π_G in square region S. Figure 2a illustrates the graphs of $F_0(\mathbf{x}, w_0; \Pi)$ as functions of w_0, where $\mathbf{x}$ is the center of the region. Points Π_A-Π_C clustered at the center in Figure 1 yield monotonically decreasing functions in Figure 2a. Points in Π_D and Π_E are distributed close to the boundaries, and consequently, $F_0(\mathbf{x}, w_0; \Pi)$ monotonically increases with w_0. Donut-shaped distributions of Π_F and Π_G generate unimodal and bimodal functions in Figure 2a, respectively.

Figure 2b shows $f_0(\mathbf{x}, w_0; \Pi)$ of the distributions in Figure 1. The functions almost inherit the topological properties of $F_0(\mathbf{x}, w_0; \Pi)$, though the two figures may look quite different. Points in Π_A-Π_C have monotonically decreasing functions in both in Figures 2a and 2b. Points in Π_F and Π_G have one and two peaks, respectively, in both figures.



(a) Distribution Π_A (b) Distribution Π_B (c) Distribution Π_C (d) Distribution Π_D

(e) Distribution Π_E (f) Distribution Π_F (g) Distribution Π_G

Figure 1 Seven distributions of points.

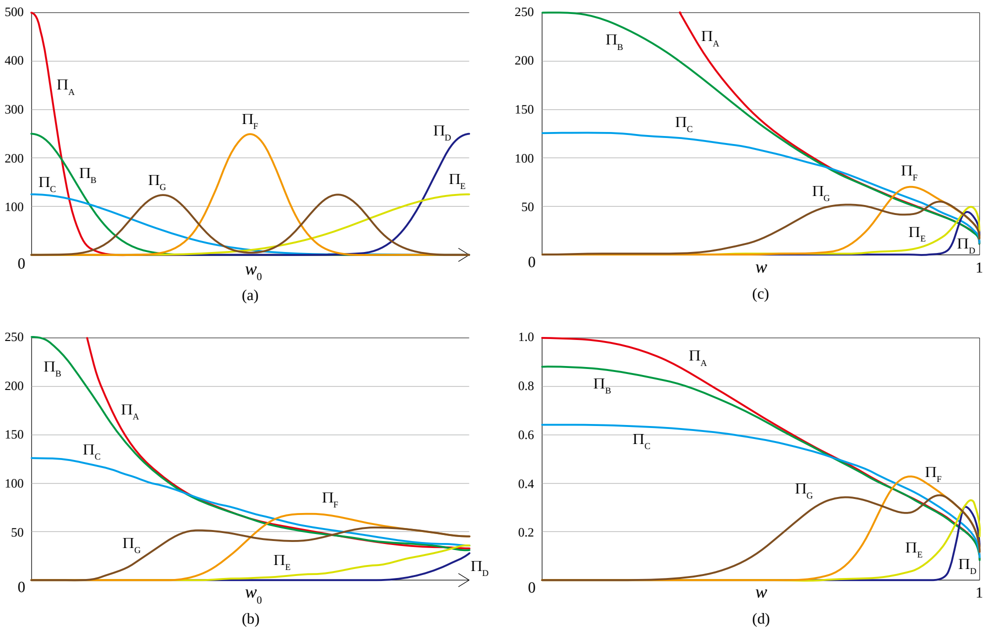


Figure 2 Density functions of points in Figure 1. (a) F_0(**x**, w_0; Π)**,** (b) f_0(**x**, w_0; Π), (c) *f*(**x**, *w*; Π), (d) function *f'*(**x**, *w*; Π).

Functions F_0(**x**, w_0; Π) and f_0(**x**, w_0; Π) indicate the degree of point clustering around **x** at various scales from local to global. Summary measures of these functions with respect to w_0 are useful in exploratory analysis since we can visualize the measures as maps. The maps permit us to grasp the spatial pattern of point distributions efficiently, and consequently, save our time of visual analysis with continuously changing the spatial scale. Suppose, for instance, the mean of f_0(**x**, w_0; Π) with respect to w_0. It becomes large where point clusters are observed consistently over a wide range of scales. Detection of such clusters is important and often critical in point pattern analysis and spatial statistics (Ord and Getis (1995); Anselin et al. (2008)).

One difficulty in the calculation of summary measures lies in the infinite range of scale parameter w_0, from zero to infinity. This prohibits us from calculating the mean of f_0(**x**, w_0; Π) in section [0, ∞) (details are discussed later in Appendix). We thus introduce a new parameter *w* ranging from zero to one that has a one-to-one correspondence with w_0:

$$\begin{aligned} w &= \frac{e^{w_0}-1}{e^{w_0}} \\ &= 1-e^{-w_0} \end{aligned} . \tag{3}$$

Parameter w_0 is given by

$$w_0 = \log(1-w).$$

(4)

Functions $F_0(\mathbf{x}, w_0; \Pi)$ and $f_0(\mathbf{x}, w_0; \Pi)$ become

$$F\left(\mathbf{x}, w; \Pi\right) = \frac{1}{\pi w_0^2} \sum_i \rho\left(\left|\mathbf{x} - \mathbf{z}_i\right|, \log\left(1 - w\right)\right)$$

(5)

and

$$\begin{aligned} f\left(\mathbf{x}, w; \Pi\right) &= \frac{F\left(\mathbf{x}, w; \Pi\right)}{\int_{\mathbf{x}\in S} F\left(\mathbf{x}, w; \Pi\right) \mathrm{d}\mathbf{x}} \\ &= \frac{\sum_i \rho\left(\left|\mathbf{x} - \mathbf{z}_i\right|, \log\left(1 - w\right)\right)}{\int_{\mathbf{x}\in S} \sum_i \rho\left(\left|\mathbf{x} - \mathbf{z}_i\right|, \log\left(1 - w\right)\right) \mathrm{d}\mathbf{x}} \end{aligned},$$

(6)

respectively. Figure 2c shows $f(\mathbf{x}, w; \Pi)$ of the distributions in Figure 1. The function still keeps the topological properties of $F_0(\mathbf{x}, w_0; \Pi)$ with the finite range of parameter w.

Another difficulty in the calculation of summary measures is caused by the undefinability of $F(\mathbf{x}, w; \Pi)$ and $f(\mathbf{x}, w; \Pi)$. They are not defined at the locations of points, where the numerator of Equation (5) is positive while the denominator is equal to zero. Integral of $f(\mathbf{x}, w; \Pi)$ with respect to w is incomputable, on which basic measures including the mean and variance are defined. One method to resolve the problem is to transform $f(\mathbf{x}, w; \Pi)$ into another function of a finite range:

$$f'\left(\mathbf{x}, w; \Pi\right) = 1 - e^{-f\left(\mathbf{x}, w; \Pi\right)}.$$

(7)

Figure 2d shows $f'(\mathbf{x}, w; \Pi)$ that ranges from zero to one. The *mean* of $f'(\mathbf{x}, w; \Pi)$ is given by

$$\mu\left(\mathbf{x}; \Pi\right) = \int_{w=0}^{1} f'\left(\mathbf{x}, w; \Pi\right) \mathrm{d}w.$$

(8)

The mean indicates the average degree of point clustering around location $\mathbf{x}$. The *variance* of $f'(\mathbf{x}, w; \Pi)$ is also a basic statistic:

$$\sigma^2\left(\mathbf{x}; \Pi\right) = \int_{w=0}^{1} \left\{ f'\left(\mathbf{x}, w; \Pi\right) - \mu \right\}^2 \mathrm{d}w.$$

(9)

Variance evaluates the stability in the degree of point clustering. If $\mu(\mathbf{x}; \Pi)$ is large and $\sigma^2(\mathbf{x}; \Pi)$ is small, it is highly plausible that a large and dense cluster of points exist around $\mathbf{x}$. Small $\mu(\mathbf{x}; \Pi)$ and $\sigma^2(\mathbf{x}; \Pi)$ indicate that few points exist around $\mathbf{x}$.

Both mean and variance concern the degree of point clustering at each location. In contrast, *range* represents the spatial extent of point distribution around each location. It is defined as the average of w weighted by $f'(\mathbf{x}, w; \Pi)$:

$$r\left(\mathbf{x};\Pi\right)=\frac{\int_{w=0}^{1} wf'\left(\mathbf{x},w;\Pi\right)\mathrm{d}w}{\int_{w=0}^{1} f'\left(\mathbf{x},w;\Pi\right)\mathrm{d}w}. \tag{10}$$

Range $r(\mathbf{x}; \Pi)$ shows a large value when points are located away from **x** such as Π_D and Π_E in Figure 1. Distributions Π_A and Π_B have small $r(\mathbf{x}; \Pi)$ because points are clustered around **x**.

The above three measures are defined based on $f'(\mathbf{x}, w; \Pi)$ that was introduced to avoid the undefinability of $f(\mathbf{x}, w; \Pi)$. An alternative to resolve the problem is to define summary measures that can be calculated directly from $f(\mathbf{x}, w; \Pi)$. The *median* of $f(\mathbf{x}, w; \Pi)$, which is denoted by $m(\mathbf{x}; \Pi)$, substitutes for the mean of $f'(\mathbf{x}, w; \Pi)$. It is defined in an implicit form as

$$\int_{w=0}^{1} \xi\left(f\left(\mathbf{x},w;\Pi\right),m\left(\mathbf{x};\Pi\right)\right)\mathrm{d}w=\frac{1}{2}, \tag{11}$$

where $\xi(f(\mathbf{x}, w; \Pi), s)$ is a binary function:

$$\xi\left(f\left(\mathbf{x},w;\Pi\right),s\right)=\begin{cases}1 & \text{if } f\left(\mathbf{x},w;\Pi\right)\geq s\\ 0 & \text{otherwise}\end{cases}. \tag{12}$$

Median $m(\mathbf{x}; \Pi)$ ranges from zero to infinity. Similar to mean $\mu(\mathbf{x}; \Pi)$, $m(\mathbf{x}; \Pi)$ indicates the average degree of point clustering around location **x**. It becomes large where clusters are observed consistently across different scales.

Extent $e(\mathbf{x}; \Pi)$ is an alternative to range $r(\mathbf{x}; \Pi)$ defined based on $f(\mathbf{x}, w; \Pi)$. This measure evaluates the relationship between w and $f(\mathbf{x}, w; \Pi)$ by extending the Spearman's rank correlation coefficient. Suppose we divide section [0, 1] into m subsections, in each of which a representative w is defined at the center:

$$W_m=\left\{\frac{1}{2m}w,\frac{3}{2m}w,\frac{5}{2m}w,\ldots,\frac{2m-1}{2m}w\right\}. \tag{13}$$

The set of $f(\mathbf{x}, w; \Pi)$ values corresponding set W_m is given by

$$\begin{aligned}\Psi_m&=\left\{\psi_1,\psi_2,\psi_3,\ldots,\psi_m\right\}\\&=\left\{f\left(\mathbf{x},\frac{1}{2m}w;\Pi\right),f\left(\mathbf{x},\frac{3}{2m}w;\Pi\right),f\left(\mathbf{x},\frac{5}{2m}w;\Pi\right),\ldots,f\left(\mathbf{x},\frac{2m-1}{2m}w;\Pi\right)\right\}.\end{aligned} \tag{14}$$

Let $R(\psi_i; \Psi_m)$ be the rank of ψ_i in set Ψ_m in ascending order. The Spearman's rank correlation coefficient between W_m and Ψ_m is

$$\rho\left(T_m,\Psi_m\right)=1-\frac{6\sum_{i=1}^{m}\left\{R\left(\psi_i;\Psi_m\right)-i\right\}^2}{m^3-m}.$$

(15)

Increasing *m* to infinity, we define *e*(**x**; Π):

$$\begin{aligned}e\left(\mathbf{x};\Pi\right)&=-\lim_{m\to\infty}\rho\left(T_m,\Psi_m\right)\\&=\lim_{m\to\infty}\frac{6\sum_{i=1}^{m}\left\{R\left(\psi_i;\Psi_m\right)-i\right\}^2}{m^3-m}-1\end{aligned}.$$

(16)

Similar to the Spearman's rank correlation coefficient, *e*(**x**; Π) ranges from -1 to 1. Extent *e*(**x**; Π) shows a large value when points decrease with an increase of *w* around **x** as seen in Π_A and Π_B in Figure 1. Points in Π_D and Π_E increase with *w*, and consequently, *e*(**x**; Π) takes a small value.

The five measures proposed above evaluate the degree of point clustering from different perspectives. Mean μ(**x**; Π) and median *m*(**x**; Π) focus on the average degree of point clustering across different scales. Range *r*(**x**; Π) and extent *e*(**x**; Π) concern the spatial extent of point distribution at each location. Variance σ^2(**x**; Π) indicates the stability in the degree of point clustering with respect to the scale of analysis. The maps of these measures reveal the spatial patterns of point distributions evaluated across different scales, including important point clusters observed consistently over the entire range of scales.

2.2 Analysis of the relationship between multiple distributions of points

This subsection extends the proposed measures for the analysis of the relationship between two and more distributions of points. The former consists of the evaluation of the change of a single distribution and the comparison of two different distributions.

We first discuss the change of point distribution from Π to Π'. Our primary interest lies in the areas where points significantly increase or decrease. The difference between the distributions is represented by

$$f_\Delta\left(\mathbf{x},w;\Pi,\Pi'\right)=\frac{F\left(\mathbf{x},w;\Pi'\right)-F\left(\mathbf{x},w;\Pi\right)}{\int_{\mathbf{x}\in S}F\left(\mathbf{x},w;\Pi'\right)\mathrm{d}\mathbf{x}+\int_{\mathbf{x}\in S}F\left(\mathbf{x},w;\Pi\right)\mathrm{d}\mathbf{x}}.$$

(17)

A variant with a finite range is defined as

$$f'_\Delta\left(\mathbf{x},w;\Pi,\Pi'\right)=1-e^{-f_\Delta\left(\mathbf{x},w;\Pi,\Pi'\right)}.$$

(18)

These functions indicate the increase in point density from Π to Π'. They permit us to evaluate the change of point distribution by extending the five measures proposed in the previous subsection. The mean, variance, and range of $f'_{\Delta}(\mathbf{x}, w; \Pi, \Pi')$ are given by

$$\mu_{\Delta}\left(\mathbf{x};\Pi,\Pi'\right)=\int_{w=0}^{1} f'_{\Delta}\left(\mathbf{x},w;\Pi,\Pi'\right)\mathrm{d}w, \tag{19}$$

$$\sigma^{2}{}_{\Delta}\left(\mathbf{x};\Pi,\Pi'\right)=\int_{w=0}^{1}\left\{f'_{\Delta}\left(\mathbf{x},w;\Pi,\Pi'\right)-\mu\right\}^{2}\mathrm{d}w, \tag{20}$$

and

$$r_{\Delta}\left(\mathbf{x};\Pi,\Pi'\right)=\frac{\int_{w=0}^{1} wf'_{\Delta}\left(\mathbf{x},w;\Pi,\Pi'\right)\mathrm{d}w}{\int_{w=0}^{1} f'_{\Delta}\left(\mathbf{x},w;\Pi,\Pi'\right)\mathrm{d}w}, \tag{21}$$

respectively. Median $m_{\Delta}(\mathbf{x}; \Pi, \Pi')$ is defined based on $f_{\Delta}(\mathbf{x}, w; \Pi, \Pi')$ as

$$\int_{w=0}^{1}\xi\left(f_{\Delta}\left(\mathbf{x},w;\Pi,\Pi'\right),m\left(\mathbf{x};\Pi,\Pi'\right)\right)\mathrm{d}w=\frac{1}{2}. \tag{22}$$

Extent $e_{\Delta}(\mathbf{x}; \Pi, \Pi')$ is defined similarly according to Equation (16). Mean $\mu_{\Delta}(\mathbf{x}; \Pi, \Pi')$ and median $m_{\Delta}(\mathbf{x}; \Pi, \Pi')$ show large positive values where points greatly increase, while the decrease of points yields negative values. Range $r_{\Delta}(\mathbf{x}; \Pi, \Pi')$ and extent $e_{\Delta}(\mathbf{x}; \Pi, \Pi')$ behave in the opposite way, i.e., they become negative where points increase. Variance $\sigma^{2}{}_{\Delta}(\mathbf{x}; \Pi, \Pi')$ is always positive and behave independently of the other measures.

Comparison of two different types of points Π_1 and Π_2 can be performed in two different ways. If a focus is on the absolute difference between Π_1 and Π_2, we can follow the above procedure by replacing Π and Π' with Π_1 and Π_2. If an interest lies in the difference in the relative spatial patterns of Π_1 and Π_2, we need to standardize each density function before comparison. The difference between Π_1 and Π_2 is represented as

$$\begin{aligned} f_{R\Delta}\left(\mathbf{x},w;\Pi_1,\Pi_2\right) &= f\left(\mathbf{x},w;\Pi_2\right)-f\left(\mathbf{x},w;\Pi_1\right) \\ &=\frac{F\left(\mathbf{x},w;\Pi_2\right)}{\int_{\mathbf{x}\in S}F\left(\mathbf{x},w;\Pi_2\right)\mathrm{d}\mathbf{x}}-\frac{F\left(\mathbf{x},w;\Pi_1\right)}{\int_{\mathbf{x}\in S}F\left(\mathbf{x},w;\Pi_1\right)\mathrm{d}\mathbf{x}} \end{aligned} \tag{23}$$

and

$$f'_{R\Delta}\left(\mathbf{x},w;\Pi_1,\Pi_2\right)=1-e^{-f_{R\Delta}\left(\mathbf{x},w;\Pi_1,\Pi_2\right)}.$$

(24)

Summary measures are defined based on these functions. Mean, for instance, is given by

$$\mu_{R\Delta}\left(\mathbf{x};\Pi_1,\Pi_2\right)=\int_{w=0}^{1} f\,'_{R\Delta}\left(\mathbf{x},w;\Pi_1,\Pi_2\right)\mathrm{d}w\,.$$

(25)

The above measures are applicable to the analysis of the relationship between more than two distributions of points. If we have numerous distributions, classification is also effective because we can focus on the comparison between a smaller number of groups rather than the comparison between individual distributions. Classification utilizes the integrals of the above measures over the whole area. For instance, the integral of $\mu_{\Delta}(\mathbf{x}; \Pi_1, \Pi_2)$ is given by

$$d_{\mu}\left(\Pi_1,\Pi_2\right)=\int_{\mathbf{x}\in S}\mu_{R\Delta}\left(\mathbf{x};\Pi_1,\Pi_2\right)\mathrm{d}\mathbf{x}\,.$$

(26)

Measure $d_{\mu}(\Pi_1, \Pi_2)$ works as a distance measure since it becomes large if Π_1 and Π_2 have different spatial patterns. Calculating $d_{\mu}(\Pi_1, \Pi_2)$ between every pair of distributions, we obtain a distance matrix. This gives us a basis for classifying the distributions by using cluster analysis methods (Everitt et al. (2011); Hennig et al. (2015)).

3. Numerical experiments

This section tests the performance of the proposed measures through numerical experiments. Evaluation is based on whether they visualize the properties of point distributions clearly and appropriately, especially whether they are helpful for the detection of point clusters. Subsections 3.1 and 3.2 treat a single distribution of points and the changes in a point distribution, respectively. Points are distributed in a square region of side 1.0 in all the experiments. Every distribution consists of point clusters, each of which follows a two-dimensional Gaussian distribution. We call the center of Gaussian distribution a *seed*. Variables n and ρ denote the number of points and the standard deviation of Gaussian distribution, respectively.

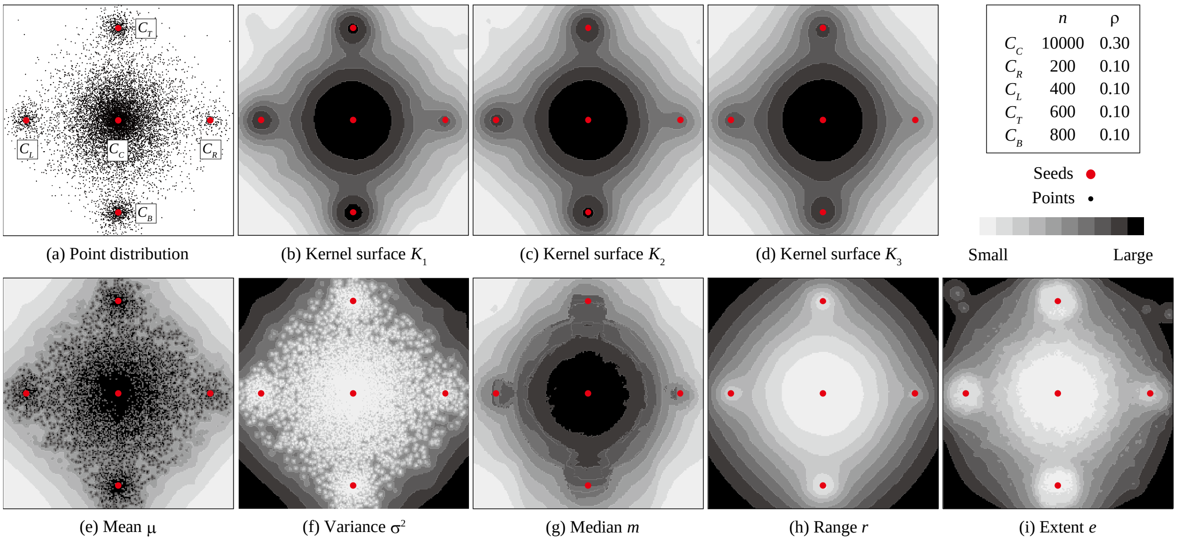
3.1 Visualization of a single distribution

Figure 3a shows a point distribution consisting of one large (C_C) and four small (C_R, C_L, C_T, C_B) clusters. The small clusters share the same spatial extent (ρ) with different densities (n/ρ^2). Figure 3b-d show the kernel density distributions of different window widths generated from the point distribution.

Figure 3e-i show the distributions the proposed measures. Mean μ and median m are large where points are clustered while variance σ^2, range r, and extent e are small in those areas. Measures μ and m have rough surfaces in Figures 3e and 3f where the global pattern and point clusters are not clear. Measures m and e, on the other hand, successfully displays all the small clusters. Median m is close to kernel surface

K_1, though the peaks of small clusters are located a little inside of their seeds. Range *r* looks similar to K_3, in both of which only the smallest cluster C_R is not clearly recognizable. Extent *e* is close to K_1 except its rough contour lines.

Measures μ and *m* aim to indicate the density of points, while *r* and *e* intend to represent the size of point clusters. Figure 3g shows that *m* successfully visualizes the difference in point density among the small clusters. Measures *r* and *e*, on the other hand, do not fully achieve their purpose since they visualize denser clusters as larger ones. These measures are affected by point density at least to some extent.



	n	ρ
C_C	10000	0.30
C_R	200	0.10
C_L	400	0.10
C_T	600	0.10
C_B	800	0.10

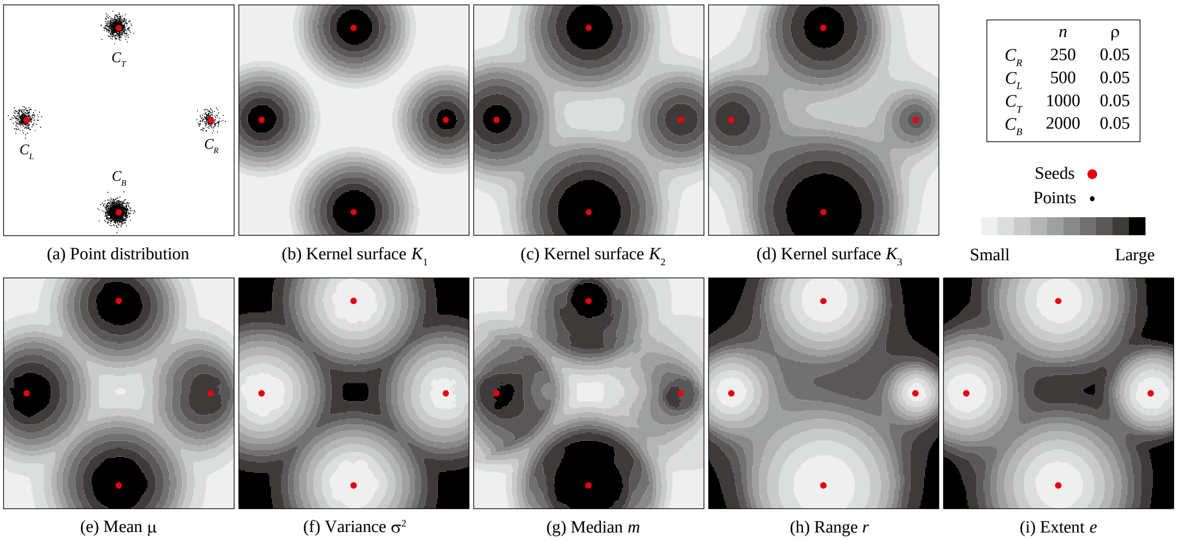
(a) Point distribution (b) Kernel surface K_1 (c) Kernel surface K_2 (d) Kernel surface K_3

(e) Mean μ (f) Variance σ^2 (g) Median *m* (h) Range *r* (i) Extent *e*

Figure 3 Point distribution and its summary measures.

Figures 4 and 5 focus on the effect of the size and density of point clusters on the performance of the proposed measures. Four clusters in Figure 4a share the same size (ρ) but differ in point density (n/ρ^2). Surfaces of μ and σ^2 are both smoother than those in Figure 3 because the points are more tightly clustered. Measures μ, σ^2 and *m* successfully visualize the difference in point density among the clusters. Measures *r* and *e* are expected to visualize all the clusters in the same size, which is not fully attained. Though *e* yields better result than *r*, both visualize the clusters in different sizes.

Four clusters in Figure 5a consist of the same number of points with different spatial extents, and consequently, they are different in both size and point density. Measures μ, σ^2, and *m* correctly visualize the difference in point density among the clusters, and *e* indicates the difference in cluster size. Range *r* visualizes the four clusters almost in the same way, which implies that *r* tends to represent the number of points rather than the spatial extent of point clusters.

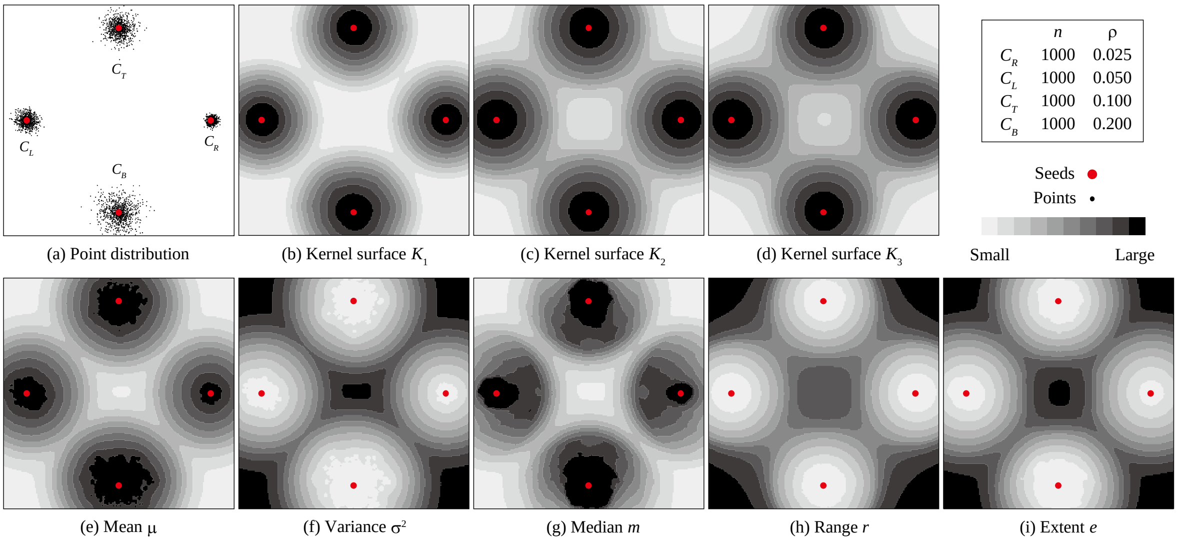


	n	ρ
C_R	250	0.05
C_L	500	0.05
C_T	1000	0.05
C_B	2000	0.05

(a) Point distribution (b) Kernel surface K_1 (c) Kernel surface K_2 (d) Kernel surface K_3

(e) Mean μ (f) Variance σ^2 (g) Median *m* (h) Range *r* (i) Extent *e*

Figure 4 Point distribution and its summary measures.



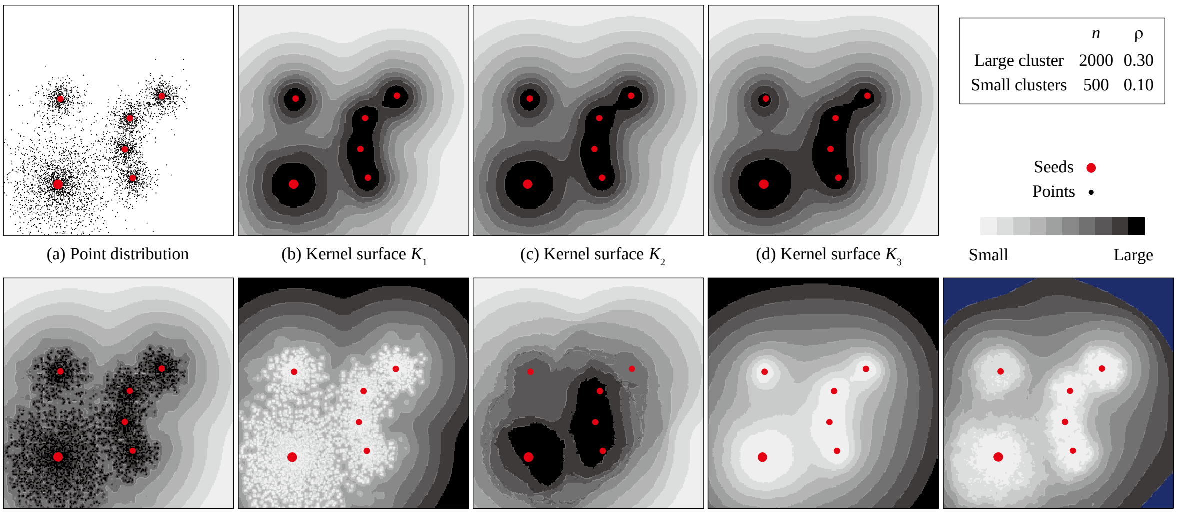
	n	ρ
C_R	1000	0.025
C_L	1000	0.050
C_T	1000	0.100
C_B	1000	0.200

(a) Point distribution (b) Kernel surface K_1 (c) Kernel surface K_2 (d) Kernel surface K_3

(e) Mean μ (f) Variance σ^2 (g) Median *m* (h) Range *r* (i) Extent *e*

Figure 5 Point distribution and its summary measures.

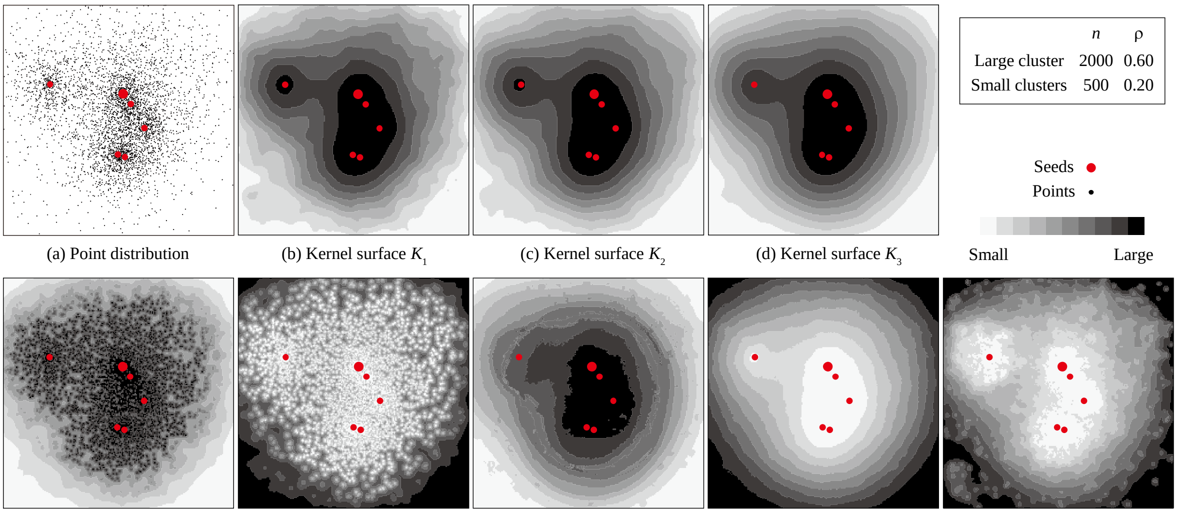
Figures 6 and 7 consider the cases where one large and five small clusters are arranged rather irregularly. The small clusters share the same size and density in each figure. Points are more tightly clustered in Figure 6a than those in Figure 7a so that point clusters are more easily detectable in the former. Mean μ and variance σ^2 have rough surfaces where clusters are rather undistinguishable. Median *m* also fails to visualize the clusters clearly in both figures; some are combined into one while others are covered by the larger cluster. Measures *r* and *e* perform better than the other measures. Extent *e* displays small clusters more separately than *r*, especially in Figure 6.



(a) Point distribution (b) Kernel surface K_1 (c) Kernel surface K_2 (d) Kernel surface K_3

(e) Mean μ (f) Variance σ^2 (g) Median *m* (h) Range *r* (i) Extent *e*

Figure 6 Point distribution and its summary measures.



(a) Point distribution (b) Kernel surface K_1 (c) Kernel surface K_2 (d) Kernel surface K_3

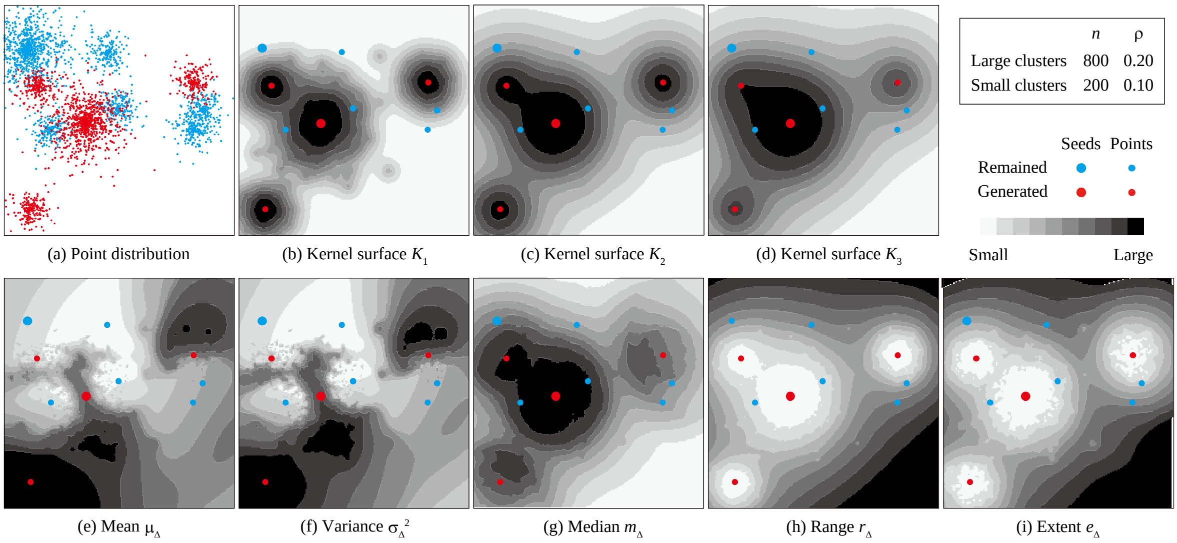
(e) Mean μ (f) Variance σ^2 (g) Median *m* (h) Range *r* (i) Extent *e*

Figure 7 Point distribution and its summary measures.

3.2 Visualization of changes in a point distribution

This subsection discusses the changes in a point distribution. We assume unmovable points, i.e., their change is limited to generation and disappearance. Other dynamic changes such as movement and integration are out of scope of this paper.

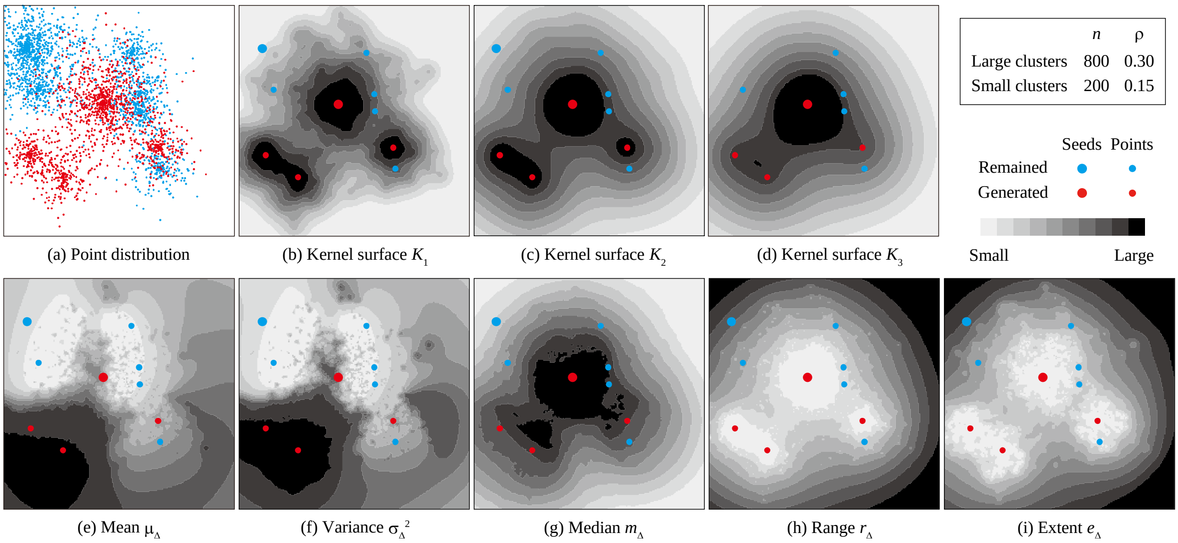
Figures 8 and 9 treat the cases where only the cluster generation is observed. One large and three small clusters are generated (red circles) where one large and five small clusters already exist (blue circles). Measures m_Δ, r_Δ, and e_Δ successfully detect the generation of large cluster in both figures, while it is not clear in μ_Δ and σ_Δ^2. Small clusters are also clear in e_Δ and r_Δ, although the latter to a lesser extent in Figure 9. Small clusters in m_Δ are detectable but their location is different from their actual ones. Measures μ_Δ and σ_Δ^2 fail to visualize the small clusters clearly except the one at the lower-left corner in Figure 8.



(a) Point distribution (b) Kernel surface K_1 (c) Kernel surface K_2 (d) Kernel surface K_3

(e) Mean μ_Δ (f) Variance σ_Δ^2 (g) Median m_Δ (h) Range r_Δ (i) Extent e_Δ

Figure 8 Cluster generation in a point distribution. The distribution originally consists of one large and five small clusters (blue circles). One large and three small clusters are newly generated (red circles).

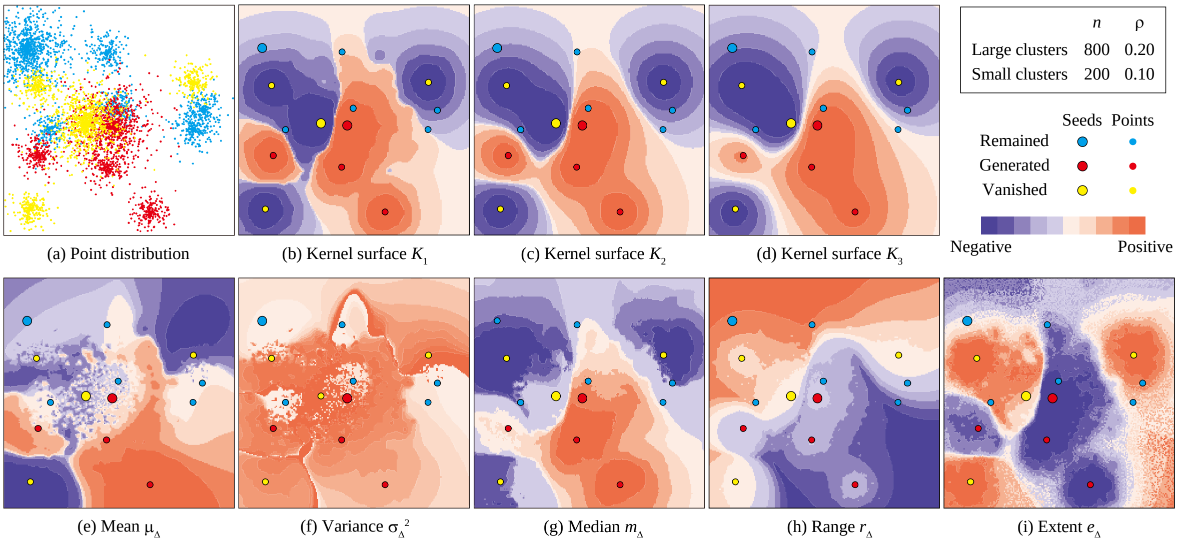


(a) Point distribution (b) Kernel surface K_1 (c) Kernel surface K_2 (d) Kernel surface K_3

(e) Mean μ_Δ (f) Variance σ_Δ^2 (g) Median m_Δ (h) Range r_Δ (i) Extent e_Δ

Figure 9 Cluster generation in a point distribution. The distribution originally consists of one large and five small clusters (blue circles). One large and three small clusters are newly generated (red circles).

Figures 10 and 11 treat the cases where both cluster generation and disappearance occur. A point distribution initially consists of two large and eight small clusters. One large and three small clusters disappear, while the same number of clusters are generated. Measures μ_Δ and m_Δ become positive (red shades) while r_Δ and e_Δ become negative (navy shades) where clusters are generated. Measures r_Δ and e_Δ detect point clusters almost successfully except the cases where cluster generation and disappearance closely occur as observed at the center in both figures. Median m_Δ is similarly effective in Figure 10, but tends to locate point clusters inward in Figure 11 where points are more dispersed in each cluster. Measures μ_Δ and σ_Δ^2 are not effective due to their rough distributions.



(a) Point distribution

(b) Kernel surface K_1

(c) Kernel surface K_2

(d) Kernel surface K_3

(e) Mean μ_Δ

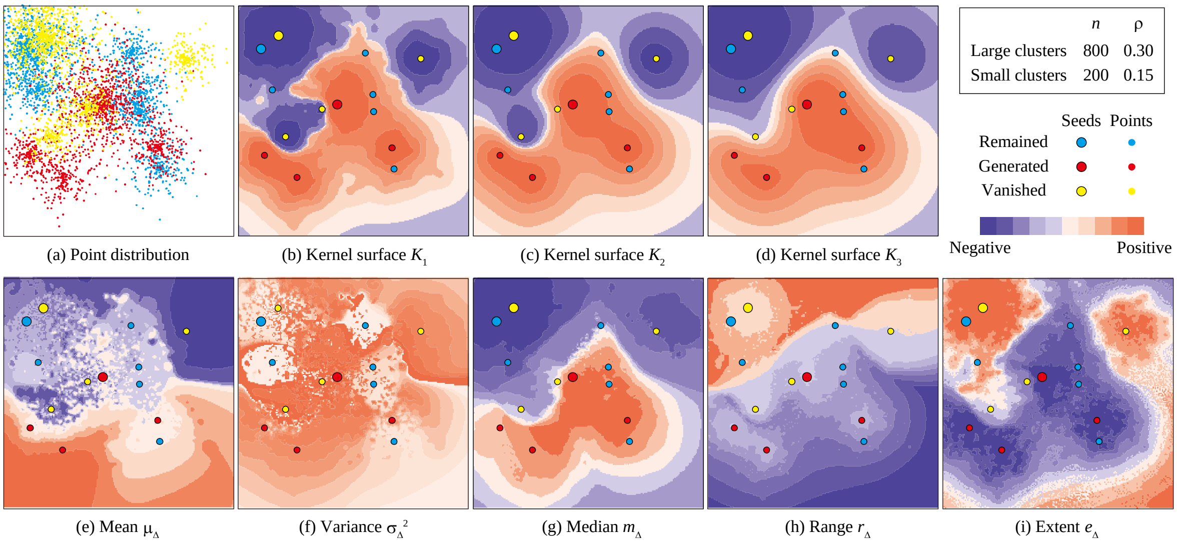
(f) Variance σ_Δ^2

(g) Median m_Δ

(h) Range r_Δ

(i) Extent e_Δ

Figure 10 Cluster generation and disappearance in a point distribution. The distribution originally consists of two large and eight small clusters (blue and yellow circles). One large and three small clusters disappear (yellow circles), while one large and three small clusters are generated (red circles).



(a) Point distribution (b) Kernel surface K_1 (c) Kernel surface K_2 (d) Kernel surface K_3

(e) Mean μ_Δ (f) Variance σ_Δ^2 (g) Median m_Δ (h) Range r_Δ (i) Extent e_Δ

Figure 11 Cluster generation and disappearance in a point distribution. The distribution originally consists of two large and eight small clusters (blue and yellow circles). One large and three small clusters disappear (yellow circles), while one large and three small clusters are generated (red circles).

4. Application to real data

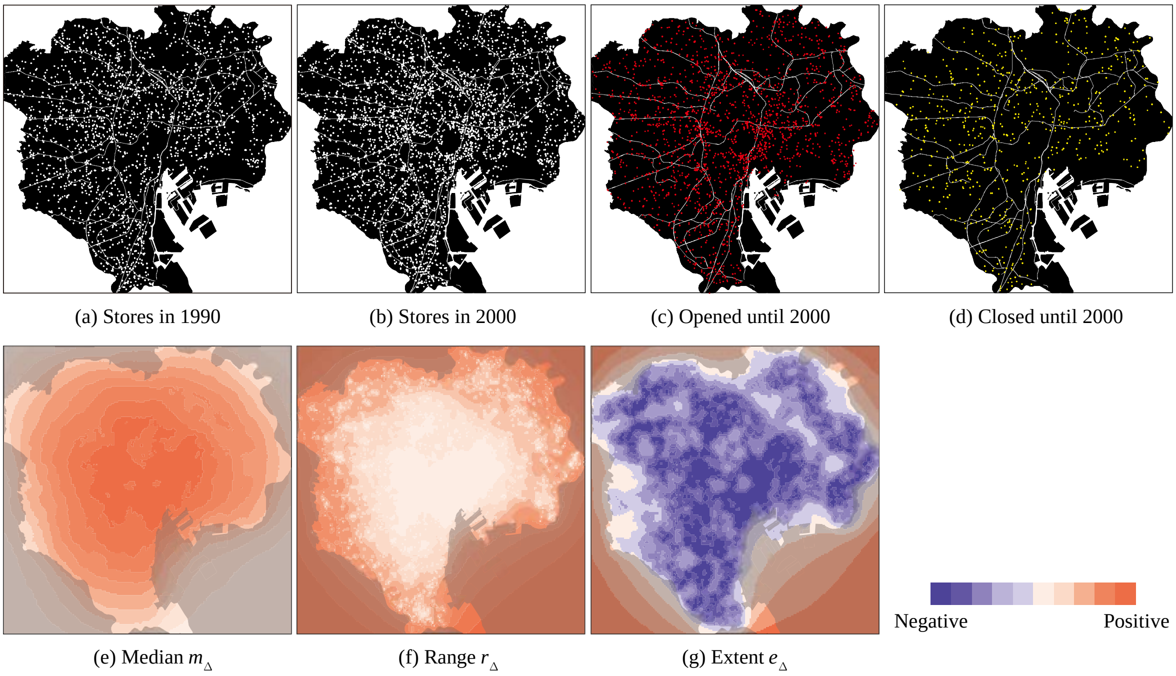
This section applies the proposed measures to the analysis of real data. Subsection 4.1 evaluates the change of the distribution of convenience stores in Tokyo, Japan. Subsection 4.2 analyzes the change in the number of children in the Greater Tokyo Area by using spatially-aggregated census data. Subsection 4.3 classifies the distributions of commercial facilities in Chiba, Japan, by using the distance measures proposed in Subsection 2.2.

4.1 Distribution of convenience stores in Tokyo

Figures 12a and 12b show the distributions of convenience stores in Tokyo listed in the telephone directory of the NTT TownPage cooperation. Stores had increased from 2070 in 1990 to 3382 in 2000, during which period 1949 new stores opened (Figure 12c) and 637 stores were closed (Figure 12d). Closed stores are distributed rather uniformly while new stores are clustered in some areas.

Figure 12e-Figure 12g show the distribution of calculated measures. We omit μ_Δ and σ^2_Δ due to the space limitations. Measures m_Δ and r_Δ are positive while e_Δ is negative over the whole area, because stores had increased as a total during this period. Measures m_Δ and r_Δ clearly indicate the global pattern of new stores, i.e., new stores opened more in the central area than in the surrounding area. Extent e_Δ reveals smaller clusters of new stores that are not easily recognizable in Figure 12c. Figure 13 shows the distributions of new and closed stores overlaid on the distribution of e_Δ. A close relationship exists between e_Δ and new stores in Figure 13a. The relationship, on the other hand, is not clear in Figure 13b.

One reason for this is that closed stores are fewer than new ones. The spatial pattern of new stores masks that of closed stores. Another reason is that closed stores are distributed rather uniformly as seen in Figure 12d. They do not contain large clusters that greatly reduce the value of e_Δ.


(a) Stores in 1990

(b) Stores in 2000

(c) Opened until 2000

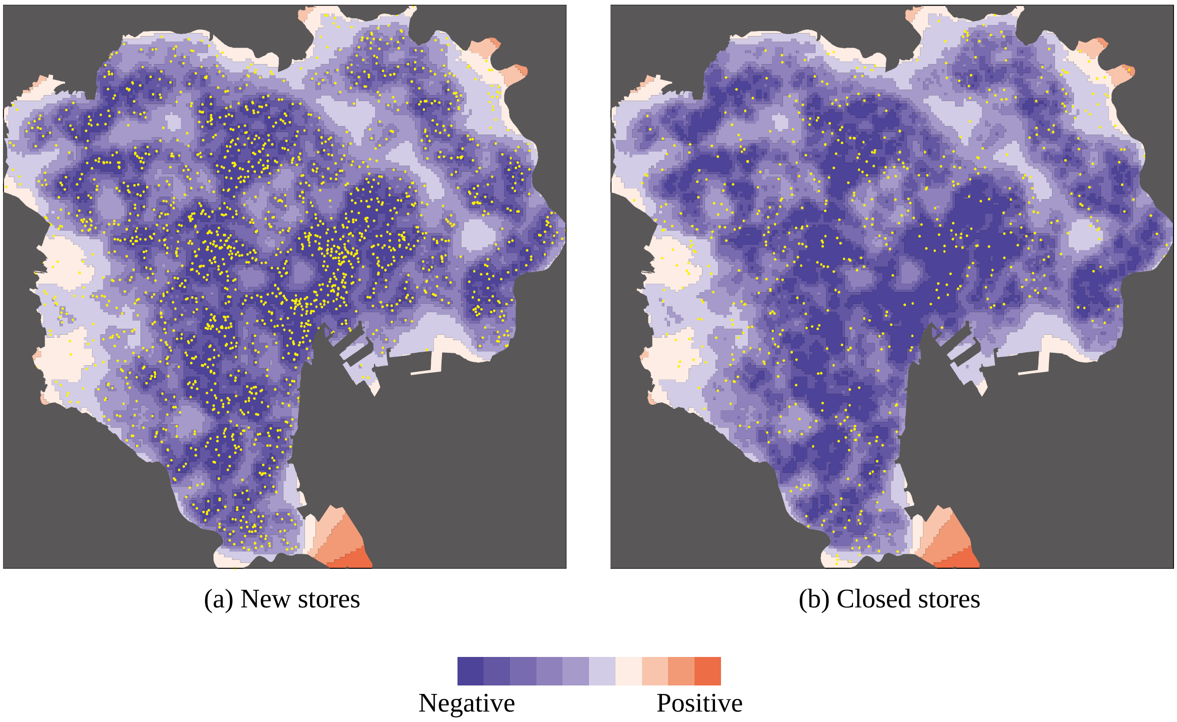
(d) Closed until 2000

(e) Median m_Δ

(f) Range r_Δ

(g) Extent e_Δ

Figure 12 Convenience stores in Tokyo, Japan, in 1990 and 2000.


(a) New stores

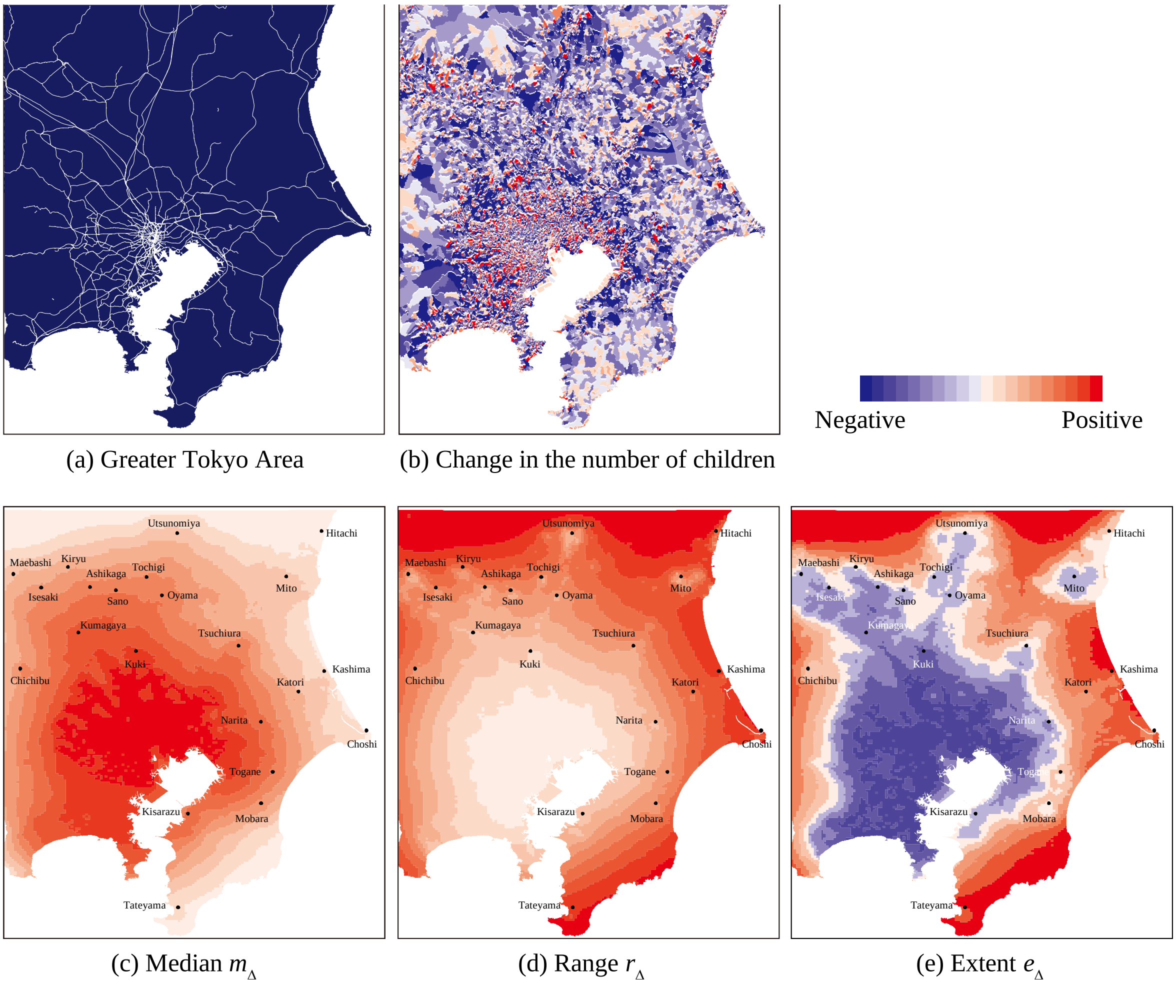
(b) Closed stores

Figure 13 New and closed convenience stores in Tokyo, Japan, between 1990 and 2000.

4.2 Distribution of children in the Greater Tokyo Area

The proposed measures are applicable to point data aggregated by spatial units such as census tracts and zip codes. This subsection analyzes the change in the number of children in the Greater Tokyo Area, Japan (Figure 14a). Population had increased continuously from 1995 to 2005 in this area due to centralization, while the birth rate had gradually decreased. Our interest lies in the spatial pattern of the change in the distribution of children.

Figure 14b shows the change in the number of children from 1995 to 2005 aggregated at the city level in the census data. Though we can see that children had increased in the central and some surrounding areas, local variation of the distribution prevents us of grasping the overall pattern of the change. Figure 14c-e show the distribution of calculated measures. Similar to Figure 12, m_Δ and r_Δ are positive while e_Δ is negative over the whole area in Figure 14, which implies that children had increased as a whole during this period. The scale of spatial pattern revealed by m_Δ, r_Δ and e_Δ changes from global to local in this order. Median m_Δ indicates that children had increased more in the central area than in its surroundings. Range r_Δ adds local clusters in the north area, represented as cities such as Maebashi, Utsunomiya and Mito. Extent e_Δ visualizes the local clusters more clearly, and detects even smaller clusters including Oyama, Tsuchiura, Choshi and Mobara. Extent e_Δ also reveals the detailed spatial pattern of children increase in the central area of Tokyo.



(a) Greater Tokyo Area

(b) Change in the number of children

(c) Median m_Δ

(d) Range r_Δ

(e) Extent e_Δ

Figure 14 Change in the number of children between 1995 and 2005 in the Greater Tokyo Area. White lines in Figure 14a indicate railway lines.

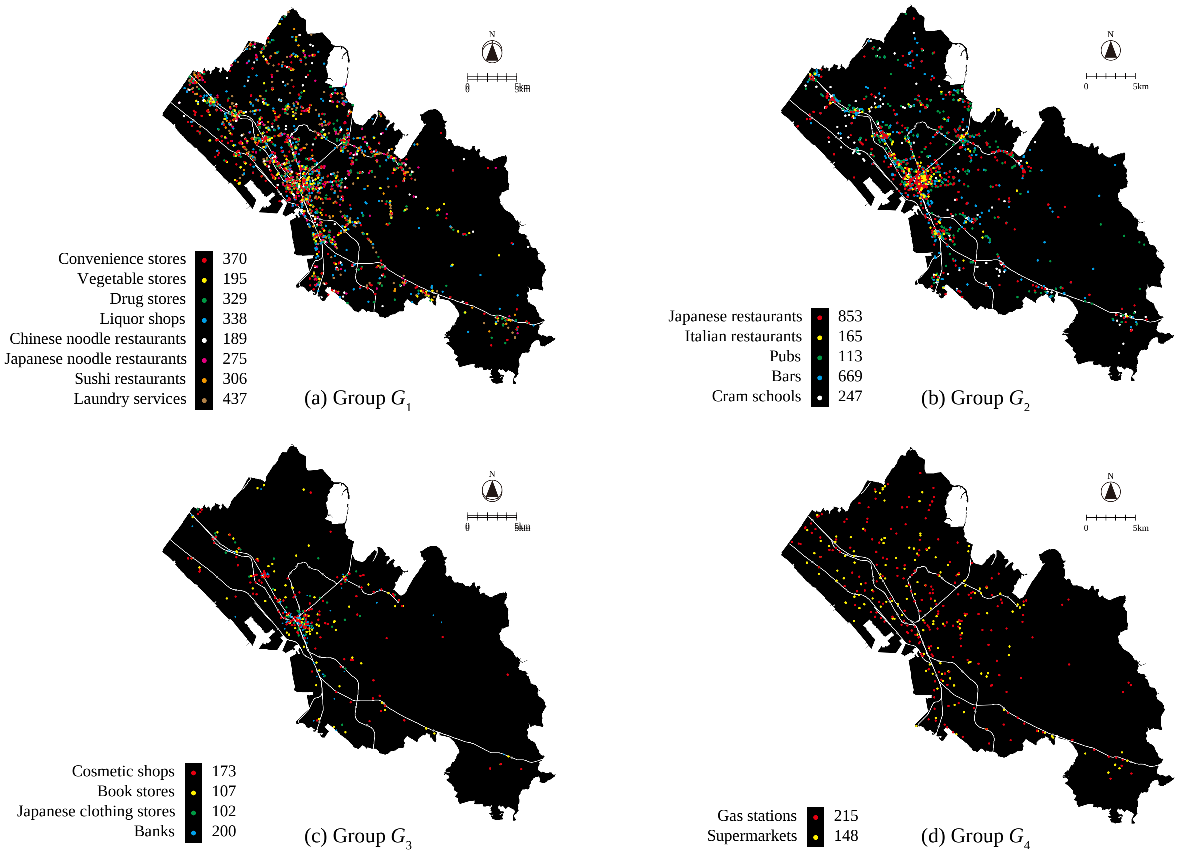
4.3 Classification of the distributions of commercial facilities

This subsection applies the proposed measures to the classification of the distributions of commercial facilities in Chiba, Japan. The data are based on the list of shops and restaurants provided by the NTT TownPage cooperation. We classify nineteen types of commercial facilities by their spatial patterns. The *k*-medoids method (Everitt et al. (2011); Hennig et al. (2015)) uses d_m, d_r, and d_e as distance measures in classification.

Figure 15 shows the distributions of commercial facilities classified based on d_e where k=4. We omit the other results since d_m and d_r yielded similar classifications. Group G_1 is the largest group containing convenience stores, vegetable stores and noodle restaurants that provide commodities and reasonable food. Shops are widely spread in both residential and commercial areas to satisfy the frequent demand of residents and workers. Facilities in G_2 are rather clustered around railway stations. Group G_3

is smaller and more tightly gathered around railway stations. People visit G_2 and G_3 stores less frequently such as pubs, bars, cosmetic shops and book stores. The stores form clusters around railway stations to offer opportunities for comparison and multipurpose shopping. Group G_4 consists of gas stations and supermarkets that are dispersed over the whole area. Unlike other commercial facilities, gas stations and supermarkets are not found around railway stations. They do not need to be close to railway stations where the land price is relatively high because people usually visit these facilities by car.

The above classification looks generally reasonable. Each group consists of commercial facilities that share similar spatial patterns as seen in Figure 15, which reflects the properties of goods and services provided by the facilities.



(a) Group G_1

(b) Group G_2

(c) Group G_3

(d) Group G_4

Figure 15 Classification of commercial facilities in Chiba, Japan.

5. Concluding discussion

A new method of visualizing point distributions was proposed. Measures μ, σ^2, m, r, and e and their extensions summarize the properties of point distributions evaluated over the entire range of scales. The measures were developed for exploratory purposes, i.e., they are expected to help us building research hypotheses. Visual analysis of these measures should be followed by confirmatory analysis that tests the significance of the hypotheses, which is out of scope of the paper.

Measures μ and σ^2 are extensions of descriptive measures widely used in general statistics. Numerical experiments, unfortunately, showed that μ and σ^2 are not necessarily effective for visual analysis. Rough surfaces of these measures often too much emphasize the local variation of point distributions so that the global pattern and point clusters are not easily recognizable. The strength of μ and σ^2, however, should not be underestimated. Their definition is clear and interpretation is straightforward because they are based on popular measures in general statistics. We should resolve the shortcomings of these measures to improve their performance.

The other three measures showed good results in both numerical experiments and real applications. They permit us to grasp the overall spatial pattern as well as detect point clusters and changes in point distributions. This saves our time of visual analysis with continuously changing the spatial scale. The maps of the three measures visualize the spatial patterns of different scales. Median *m* outlines the global pattern of point distributions, while *r* and *e* rather focus on the local details.

A weakness of *m* is that it tends to locate point clusters inward from their true locations as seen in Figures 8 and 9. It occurs primarily due to the edge effect, i.e., calculation of $f(\mathbf{x}, w; \Pi)$ and $f'(\mathbf{x}, w; \Pi)$ does not consider the points outside the study region that can potentially exist. All the measures suffer from the edge effect, though it was not evident in *r* and *e* in numerical experiments. We should try edge correction methods proposed in spatial statistics and image processing such as periodic edge correction and reflection correction methods in future research (Stoyan and Stoyan (1994); Wiegand and Moloney (2013)).

Extent *e* was found to be effective for the detection of point clusters. The primary objective of *e*, however, is to represent the spatial extent of point distribution around each location. This has not been fully attained because *e* is affected by the relative density of points in the neighborhood. Further improvement is necessary for *e* to be more independent of point density.

This paper assessed the performance of the proposed measures by visually comparing their spatial pattern with point distributions. We chose this approach to grasp the properties of the measures from various perspectives. Further evaluation includes cartographic experiments, questionnaire survey, and quantitative assessment. Experiments should test whether analysts can correctly understand the spatial pattern of point distributions and detect point clusters in the maps of the proposed measures. The performance of the measures can be assessed through questionnaire survey and the quantitative analysis of the results.

Dynamic changes such as movement, integration, and division of points were out of scope of this paper. However, recent development of tracking technology such as GPS, laser and video tracking systems permits us to obtain detailed point data that changes continuously. We should test the performance of the proposed measures in the analysis of dynamic data as well as develop further measures that can capture the properties of dynamic changes more appropriately.

References

Anselin L, Griffiths E, Tita G (2008) Crime mapping and hot spot analysis. *Environmental criminology and crime analysis***:** 97-116

Everitt BS, Landau S, Leese M, Stahl D (2011) *Cluster analysis, 5th edition*. Wiley,

Hennig C, Meila M, Murtagh F, Rocci R (2015) *Handbook of cluster analysis*. Chapman and Hall/CRC,

Kovalerchuk B, Schwing J (2005) *Visual and spatial analysis: Advances in data mining, reasoning, and problem solving*. Springer Science & Business Media,

Kulldorff M (1997) A spatial scan statistic. *Communications in Statistics - Theory and Methods* 26**:** 1481-1496

Ord JK, Getis A (1995) Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* 27**:** 286-306

Oyana TJ, Margai F (2015) *Spatial analysis: Statistics, visualization, and computational methods*. CRC Press,

Ripley BD (2005) *Spatial statistics*. John Wiley & Sons, New York

Scott DW (2015) *Multivariate density estimation: Theory, practice, and visualization*. John Wiley & Sons,

Sheather SJ (2004) Density estimation. *Statistical Science* 19**:** 588-597

Silverman BW (1986) *Density estimation for statistics and data analysis*. CRC Press, Boca Raton

Stoyan D, Stoyan H (1994) Fractals, random shapes and point fields. Methods of geometrical statistics. Chichester: J. Wiley & Sons,

Tango T, Takahashi K (2005) A flexibly shaped spatial scan statistic for detecting clusters. *Int J Health Geogr* 4**:** 1

Wiegand T, Moloney KA (2013) *Handbook of spatial point-pattern analysis in ecology*. CRC Press,

Yao Z, Tang J, Zhan F (2011) Detection of arbitrarily-shaped clusters using a neighbor-expanding approach: A case study on murine typhus in south texas. *Int J Health Geogr* 10**:** 1

Appendix

Mean and variance are basic measures of continuous functions. The mean of density function $f(\mathbf{x}, w_0; \Pi)$ is defined as

$$\mathrm{Mean}\left[f\left(\mathbf{x}, w_0; \Pi\right)\right] = \lim_{w_1 \to \infty} \frac{1}{w_1} \int_{w_0=0}^{w_1} f\left(\mathbf{x}, w_0; \Pi\right) \mathrm{d}w_0 . \tag{A1}$$

The mean defined above, however, is not meaningful since it is equal to one almost everywhere and undefined at the locations of points. The former occurs because $f(\mathbf{x}, w_0; \Pi)$ is almost independent of location $\mathbf{x}$ when w_0 is large enough, i.e.,

$$f\left(\mathbf{x}, w_0; \Pi\right) \approx \frac{1}{w_0} . \tag{A2}$$

Substituting this approximation into Equation A1, we obtain

$$\begin{aligned}\mathrm{Mean}\left[f\left(\mathbf{x},w_0;\Pi\right)\right] &= \lim_{w_1\to\infty}\frac{1}{w_1}\int_{w_0=0}^{w_1} f\left(\mathbf{x},w_0;\Pi\right)\mathrm{d}w_0 \\ &\approx \lim_{w_1\to\infty}\frac{1}{w_1}\lim_{\varepsilon\to 0+}\int_{w_0=\varepsilon}^{w_1}\frac{1}{w_0}\mathrm{d}w_0 \\ &= \lim_{w_1\to\infty}\frac{\log w_1}{w_1} \\ &=1\end{aligned} .$$

(A3)

The second problem occurs because *f*(**x**, 0; Π) is the delta functions divided by *n* at the locations of points, i.e.,

$$f\left(\mathbf{x},0;\Pi\right)=0\left(\forall i:\mathbf{x}\neq\mathbf{z}_i\right)$$

(A4)

and

$$\int_{\mathbf{x}} f\left(\mathbf{x},0;\Pi\right)d\mathbf{x}=1$$

(A5)

Density function *f*(**x**, w_0; Π) is not integrable at w_0=0 and thus its mean is not definable at the locations of points.

The variance of density function *f*(**x**, w_0; Π) is defined as

$$\begin{aligned}\mathrm{Var}\left[f\left(\mathbf{x},w_0;\Pi\right)\right] &= \lim_{w_1\to\infty}\frac{1}{w_1}\int_{w_0=0}^{w_1}\left\{f\left(\mathbf{x},w_0;\Pi\right)\text{-}\overline{f}\left(\mathbf{x};\Pi\right)\right\}^2\mathrm{d}w_0 \\ &= \lim_{w_1\to\infty}\frac{1}{w_1}\int_{w_0=0}^{w_1}\left\{f\left(\mathbf{x},w_0;\Pi\right)\right\}^2\mathrm{d}w_0-\left\{\overline{f}\left(\mathbf{x};\Pi\right)\right\}^2\end{aligned} .$$

(A6)

Substituting approximations A2 and A3 into the first term of the above equation, we obtain

$$\begin{aligned}\mathrm{Var}\left[f\left(\mathbf{x},w_0;\Pi\right)\right] &= \lim_{w_1\to\infty}\frac{1}{w_1}\int_{w_0=0}^{w_1}\left\{f\left(\mathbf{x},w_0;\Pi\right)\right\}^2\mathrm{d}w_0-\left\{\overline{f}\left(\mathbf{x};\Pi\right)\right\}^2 \\ &= \lim_{w_1\to\infty}\frac{1}{w_1}\int_{w_0=0}^{w_1}\frac{1}{w_0^2}\mathrm{d}w_0-1 \\ &= -\lim_{w_1\to\infty}\frac{1}{w_1}\lim_{\varepsilon\to 0+}\left[\frac{1}{w_0}\right]_{\varepsilon}^{w_1}-1 \\ &= \lim_{w_1\to\infty}\lim_{\varepsilon\to 0+}\frac{1}{w_1\varepsilon}-1\end{aligned} .$$

(A7)

The first term of the right hand is undefinable, and consequently, Var[$f(\mathbf{x}, w_0; \Pi)$] is undefinable almost everywhere.