Discussion Paper No. 70R

Statistical Methods for Analyzing the Distribution of Spatial Objects in Relation to a Surface

Yukio Sadahiro *

DECEMBER, 1997

*Research Center for Advanced Science and Technology, University of Tokyo
4-6-1, Komaba, Meguro-ku, Tokyo 153, Japan

December 22, 1997

Statistical Methods for Analyzing the Distribution of Spatial Objects in Relation to a Surface
(Spatial Distribution and Surface)

Yukio Sadahiro
Research Center for Advanced Science and Technology, University of Tokyo
4-6-1, Komaba, Meguro-ku, Tokyo 153, Japan

Phone: +81-3-3481-4541
Fax: +81-3-3481-4582
E-mail: sada@okabe.t.u-tokyo.ac.jp

Statistical Methods for Analyzing the Distribution of Spatial Objects in Relation to a Surface

Abstract

This paper develops statistical methods for analyzing the distribution of spatial objects -- points, convex polygons, and line segments -- in relation to a surface. We propose statistics for measuring the relationship between the distribution of these objects and a surface, and derive their expectations and variances under the null hypothesis that the objects are independently and randomly distributed. The statistics are approximately distributed according to the normal distribution under the null hypothesis, which enables us to test the significance of the spatial relationships statistically. Using the proposed methods, we empirically analyze the distribution of convenience stores in relation to the distribution of population in a suburb of Osaka, Japan. Some empirical findings are shown.

KEYWORDS: Statistical analysis, spatial objects, surface, distribution

1. INTRODUCTION

Analysis of the spatial relationship between the distributions of spatial objects is one of the most important subjects in GIS. Epidemiologists, for instance, analyze the distribution of disease cases in relation to the distribution of spatial objects such as sources of air pollution to detect the causes of the disease (Gatrell and Rowlingson, 1994; de Lepper *et al*., 1995; de Savigny and Wijeyaratne, 1995). Urban analysts are interested in the spatial interaction among spatial objects such as retail stores, streets, railway stations, and population, so they investigate the spatial relationship among their distributions.

There are four types of spatial objects used in GIS, namely, points, lines, polygons, and surfaces, and numerous statistical methods have been developed for analyzing the spatial relationship between the distributions of these objects. Let us briefly take a look at existing methods. The relationship between two distributions of points (points-points relationship) is often analyzed by the quadrat method which is one of the major statistical methods to treat this relationship. This method is based on the counts of points in quadrats, and the significance of the relationship can be examined by statistical tests such as Pearson's χ^2 goodness-of-fit test. Since the quadrat method is easily applicable, it is widely used in GIS and its related fields. The distance methods based on the nearest neighbor distance are also used in GIS. For instance, Pielou (1961), Lee (1979) and Okabe and Miki (1984) investigated the locational interdependence between two distributions of points using distance methods. Cuzick and Edwards (1990) developed a distance method for analyzing the distribution of point clusters in relation to the inhomogeneous point distribution. The points-lines and points-polygons relationships can be analyzed by the computational methods proposed by Okabe and Fujii (1984) and Okabe *et al*. (1988). For the relationship between two distributions of lines, statistical methods based on integral geometry are useful (see Koshizuka and Ooki, 1982, for instance).

As seen above, the spatial relationships between point, line, and polygon distributions can be statistically analyzed by the existing methods. The relationship between a surface and the distribution of spatial objects, however, is difficult to analyze because few methods have been developed in the literature. In GIS, a surface is an indispensable object for representing spatially continuous phenomena, and it often affects the distribution of discrete spatial objects. Variables continuously defined over a region, such as terrain elevation and atmospheric temperature, are treated and encoded as surfaces in GIS. Point distributions are sometimes aggregated into some spatial units to be surfaces representing the density of points. Given a surface and the distribution of spatial

objects, we often have a question whether any spatial relationship exists between them: are the spatial objects distributed where the surface values are large, or independent of the surface value? GIS permits us to consider these questions visually. The significance of spatial relationships, however, still remains unanswerable.

To answer the questions statistically, we propose test procedures to explore the distribution of spatial objects in relation to a surface. In this paper, we consider three types of spatial objects that are most frequently used in GIS: points, convex polygons, and line segments. In the following three sections, we successively develop methods for analyzing their distributions in relation to a surface. Using the methods, we empirically analyze the distribution of convenience stores in relation to the distribution of population in Section 5. Finally, we summarize the conclusions in Section 6.

2. DISTRIBUTION OF POINTS AND SURFACE

When analyzing a point distribution in relation to a surface, we can use the methods developed by Okabe and Sadahiro (1994). They dealt with three types of points-surface relationships, and proposed measures indicating the fitness of a point distribution and a surface. In this paper, however, they did not give the probability distributions of the measures in the case that the points are distributed independent of the surface. This is problematic because we cannot statistically test the significance of the spatial relationship. To solve this problem, we propose another method for analyzing the distribution of point objects in relation to a surface.

Assume that a surface whose value at $\mathbf{x}$ is denoted by $f(\mathbf{x})$ is defined in a region S_0 of area a_0 and perimeter l_0, and n points labelled by location vectors $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ are distributed in S_0 (Figure 1).

Figure 1. The distribution of point objects and a surface. The gray shade indicates the surface value.

Given the point distribution and the surface, we have the question whether the points are distributed where the surface values are large (small), or they are distributed independent of the surface value (Figure 2). To answer this question, we propose the statistic Z defined by

$$Z = \frac{1}{n}\sum_i f(\mathbf{x}_i). \tag{1}$$

If the points are distributed where the surface has large values, that is, the surface values are large at the locations of the points, Z shows a large value. If the point distribution is

spatially independent of the surface, Z shows a value close to the average of $f(\mathbf{x})$ in S_0.

Figure 2. Three relationships between the distribution of point objects and a surface. (a) Points are distributed where the surface values are large, (b) points are distributed where the surface values are small, (c) points are distributed independent of the surface value.

The significance of Z is examined by the statistical test where the null hypothesis is that the points are independently and randomly distributed in S_0. We reject the null hypothesis if Z is large enough. Under the null hypothesis, the expectation and the variance of Z are given by

$$\mathrm{E}[Z] = \frac{1}{a_0}\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x} \tag{2}$$

and

$$\mathrm{V}[Z] = \frac{1}{n}\left[\frac{1}{a_0}\int_{\mathbf{x}\in S_0}\{f(\mathbf{x})\}^2 d\mathbf{x} - \left\{\frac{1}{a_0}\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}\right\}^2\right], \tag{3}$$

respectively. One might think that the computation of these equations is difficult because they contain integral terms. The integrals of $f(\mathbf{x})$ and $\{f(\mathbf{x})\}^2$ over S_0, however, are easily computable because ordinary GIS provides a module for calculating the volume of three-dimensional objects defined by surfaces.

The central limit theorem guarantees that, if n is reasonably large, Z has approximately the normal distribution with mean E[Z] and variance V[Z] under the null hypothesis. Consequently, given a significance level, we can test whether the points are distributed where the surface values are large.

3. DISTRIBUTION OF CONVEX POLYGONS AND SURFACE

Suppose n congruent convex polygons S_1, S_2, ..., S_n intersecting a region S_0 of area a_0 and perimeter l_0. Note that polygons are allowed to intersect with each other, and that not all polygons completely lie in S_0 (see Figure 3). The area and perimeter of these polygons are denoted by a and l, respectively.

Figure 3. The distribution of convex polygons intersecting S_0 and a surface. The gray shade indicates the surface value.

We first consider the region S_P, in which the function $f(\mathbf{x})$ representing a surface is to be defined. Since some polygons may be located partly outside S_0, it is not satisfactory

to define $f(\mathbf{x})$ inside S_0. Let C_P be the smallest circle containing S_i, and C_P' be the circle centered at the origin whose radius is twice as large as that of C_P (Figure 4). Then we have

$$\begin{aligned} S_P &= S_0 \oplus C_P' \\ &= \left\{ \mathbf{x} + \mathbf{y} \middle| \mathbf{x} \in S_0, \mathbf{y} \in C_P' \right\}' \end{aligned} \tag{4}$$

where $\oplus$ is Minkowski addition operator. This equation indicates that S_P is the outer parallel region of S_0, which is generated by buffering operation in GIS. Using S_P, we define the region S_B as

$$S_B = S_P \setminus S_0. \tag{5}$$

Figure 5 shows an example of the regions S_0, S_P and S_B.

Figure 4. The circles C_P and C_P'.

Figure 5. The regions S_0, S_P and S_B.

Let us consider the question whether the polygons are distributed where the surface values are large, or they are distributed independent of the surface value. A statistical test procedure to answer this question occurs on the analogy of the points-surface method proposed in the preceding section.

Method 0

Let us define the statistic Z_P by

$$Z_P = \frac{1}{na} \sum_i \int_{\mathbf{x} \in S_i} f(\mathbf{x}) d\mathbf{x} \tag{6}$$

to measure the spatial relationship between the polygon distribution and the surface. Suppose the null hypothesis that the polygons are independently and randomly distributed so as to intersect S_0. Calculating the expectation and variance of Z_P under the null hypothesis, we can statistically test if the polygons are distributed where the surface values are large.

The expectation and the variance mentioned above, however, cannot be calculated explicitly because of the edge effect. This implies that we have to perform a costly calculation such as the spatial sampling or the Monte Carlo simulation to obtain these values. To reduce the computational cost, we develop new methods for analyzing the distribution of polygons and the surface in the following.

Method 1

This method is applicable when S_0 is convex and $f(\mathbf{x})$ is almost constant in S_B. In

this case, we assume:

Assumption 1: The function $f(\mathbf{x})$ is constant in S_B.

On this assumption, we define the function $f_1(\mathbf{x})$ to represent the surface by

$$f_1(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \mathbf{x} \in S_0, \\ f_0 & \mathbf{x} \in S_B. \end{cases} \tag{7}$$

Similar to Z_P, to measure the polygons-surface relationship, we define the statistic Z_{P1} as

$$\begin{aligned} Z_{P1} &= \frac{1}{na} \sum_i \int_{\mathbf{x} \in S_i} f_1(\mathbf{x}) d\mathbf{x} \\ &= \frac{1}{na} \sum_i \left\{ \int_{\mathbf{x} \in S_0} f(\mathbf{x}) g_i(\mathbf{x}) d\mathbf{x} + \left(a - \int_{\mathbf{x} \in S_0} g_i(\mathbf{x}) d\mathbf{x} \right) f_0 \right\}, \end{aligned} \tag{8}$$

where

$$g_i(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in S_i \cap S_0, \\ 0 & \mathbf{x} \notin S_i \cap S_0. \end{cases} \tag{9}$$

To derive the expectation and the variance of Z_{P1} under the null hypothesis mentioned earlier, we consider the case where a convex polygon S congruent to S_i is distributed randomly in such a way that it intersects S_0. In this case, we employ the statistic Z_{P1}' defined by

$$Z_{P1}' = \frac{1}{a} \int_{\mathbf{x} \in S_0} f(\mathbf{x}) g(\mathbf{x}) d\mathbf{x} + \left(1 - \frac{1}{a} \int_{\mathbf{x} \in S_0} g(\mathbf{x}) d\mathbf{x} \right) f_0, \tag{10}$$

where

$$g(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in S \cap S_0, \\ 0 & \mathbf{x} \notin S \cap S_0. \end{cases} \tag{11}$$

Since the polygons are assumed to be distributed independently, the below equations hold.

$$\mathrm{E}[Z_{P1}] = \mathrm{E}[Z_{P1}'] \tag{12}$$

$$\mathrm{V}[Z_{P1}] = \frac{\mathrm{V}[Z_{P1}']}{n}. \tag{13}$$

The expectation of Z_{P1}' is given by

$$\mathrm{E}[Z_{P1}'] = \frac{1}{a} \int_{\mathbf{x} \in S_0} f(\mathbf{x}) \mathrm{E}[g(\mathbf{x})] d\mathbf{x} + \left(1 - \frac{1}{a} \int_{\mathbf{x} \in S_0} \mathrm{E}[g(\mathbf{x})] d\mathbf{x} \right) f_0. \tag{14}$$

The expectation $\mathrm{E}[g(\mathbf{x})]$ indicates the probability that the point located at $\mathbf{x}$ is contained in S, and it does not depend on $\mathbf{x}$. Thus we have

$$\mathrm{E}[Z_{P1}'] = \frac{\mathrm{E}[g(\mathbf{x})]}{a} \left\{ \int_{\mathbf{x} \in S_0} f(\mathbf{x}) d\mathbf{x} - a_0 f_0 \right\} + f_0. \tag{15}$$

Let $m(S; \mathbf{x})$ and $m(S; S_0)$ be the measures of the set of all polygons congruent to S

containing the point at **x** and that intersecting S_0, respectively (for the formal definition of the measures, see Santaló, 1976). The expectation E[g(**x**)], the probability of the point located at **x** being contained in S, is then written as

$$\mathrm{E}\left[g(\mathbf{x})\right] = \frac{m(S;\mathbf{x})}{m(S;S_0)}. \tag{16}$$

The measure $m(S; \mathbf{x})$ is given by

$$m(S;\mathbf{x}) = 2\pi a. \tag{17}$$

If both S and S_0 are convex polygons, the measure $m(S; S_0)$ is given by

$$m(S;S_0) = 2\pi(a_0 + a) + l_0 l \tag{18}$$

(Santaló, 1976). Substituting equations (16), (17), and (18) into equation (15) yields

$$\mathrm{E}\left[Z_{P1}'\right] = \frac{1}{2\pi(a_0 + a) + l_0 l}\left\{2\pi\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x} + (2\pi a + l_0 l) f_0\right\}. \tag{19}$$

As mentioned in the preceding section, the integral in equation (19) can easily be computed with GIS.

The variance of Z_{P1}' is given by

$$\begin{aligned}\mathrm{V}\left[Z_{P1}'\right] = {} & f_0^2 + \frac{4\pi f_0}{2\pi(a_0 + a) + l_0 l}\left\{\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x} - a_0 f_0\right\} \\ & + \frac{1}{a^2}\mathrm{E}\left[\left\{\int_{\mathbf{x}\in S_0} f(\mathbf{x})g(\mathbf{x})d\mathbf{x}\right\}^2\right] + \frac{f_0^2}{a^2}\mathrm{E}\left[\left\{\int_{\mathbf{x}\in S_0} g(\mathbf{x})d\mathbf{x}\right\}^2\right]. \\ & - \frac{2 f_0}{a^2}\mathrm{E}\left[\int_{\mathbf{x}\in S_0} f(\mathbf{x})g(\mathbf{x})d\mathbf{x}\int_{\mathbf{x}\in S_0} g(\mathbf{x})d\mathbf{x}\right] - \left(\mathrm{E}\left[Z_{P1}'\right]\right)^2\end{aligned} \tag{20}$$

Using

$$\left\{\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}\right\}^2 = \int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0} f(\mathbf{x})f(\mathbf{t})d\mathbf{x}d\mathbf{t} \tag{21}$$

and

$$\int_{\mathbf{x}\in S_0} f(\mathbf{x})g(\mathbf{x})d\mathbf{x}\int_{\mathbf{x}\in S_0} g(\mathbf{x})d\mathbf{x} = \int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0} f(\mathbf{x})g(\mathbf{x})g(\mathbf{t})d\mathbf{x}d\mathbf{t}, \tag{22}$$

we obtain

$$\begin{aligned}\mathrm{V}\left[Z_{P1}'\right] = {} & \frac{1}{a^2}\int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0}\mathrm{E}\left[g(\mathbf{x})g(\mathbf{t})\right]\left\{f(\mathbf{x})f(\mathbf{t}) - 2 f_0 f(\mathbf{x}) + f_0^2\right\}d\mathbf{x}d\mathbf{t} \\ & - \left\{\frac{2\pi}{2\pi(a_0 + a) + l_0 l}\right\}^2\left\{\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x} - a_0 f_0\right\}^2\end{aligned}. \tag{23}$$

Similar to E[Z_{P1}'], the second term of equation (23) is computable with GIS. The first term can also be calculated as follows. The expectation E[g(**x**)g(**t**)] indicates the probability of the two points located at **x** and **t** being contained in S, which is given by

$$\mathrm{E}\left[g(\mathbf{x})g(\mathbf{t})\right] = \frac{m(S;\mathbf{x},\mathbf{t})}{2\pi(a_0 + a) + l_0 l}, \tag{24}$$

where $m(S; \mathbf{x}, \mathbf{t})$ is the measure of the set of all polygons congruent to S containing the points at $\mathbf{x}$ and $\mathbf{t}$. The measure $m(S; \mathbf{x}, \mathbf{t})$ has an explicit form, and its value is easily computable when S has a simple shape (see Appendix 1). Hence we can compute $\mathrm{V}[Z_{P1}']$ with equation (23) using numerical integration.

Under the null hypothesis that the polygons are independently and randomly distributed so as to intersect S_0, the probability distribution of Z_{P1} is approximated by the normal distribution with mean $\mathrm{E}[Z_{P1}]$ and variance $\mathrm{V}[Z_{P1}]$ if n is reasonably large. This enables us to perform the statistical test based on Z_{P1} whether the polygons are distributed where the surface values are large, or they are distributed independent of the surface value. The degree of approximation will be discussed in Section 5.

Method 2

If S_0 is a rectangle and the variability of $f(\mathbf{x})$ in S_B is similar to that in S_0, another method can be applied for analyzing the polygons-surface relationship. We first put an assumption as:

Assumption 2: Variability of $f(\mathbf{x})$ in S_B is a copy of that in S_0.

This assumption is often called the *periodic continuation* in spatial modelling (Stoyan and Stoyan, 1995), and its formal representation is as follows. Let $\mathbf{p}$ and $\mathbf{q}$, respectively, be the locational vectors of the lower-right and upper-left corners of S_0 with respect to the lower-left corner. Consider the translation $T_{jk}(\mathbf{x})$ written as

$$T_{jk}(\mathbf{x}) = \mathbf{x} + j\mathbf{p} + k\mathbf{q}, \tag{25}$$

where j and k are integers. We define $T(\mathbf{x})$ as

$$T(\mathbf{x}) = T_{jk}(\mathbf{x}) \tag{26}$$

by determining j and k so that $T_{jk}(\mathbf{x}) \in S_0$. Using $T(\mathbf{x})$, we define $f_2(\mathbf{x})$ to represent the surface by

$$f_2(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \mathbf{x} \in S_0, \\ f(T(\mathbf{x})) & \mathbf{x} \in S_B. \end{cases} \tag{27}$$

An example of Assumption 2 is depicted in Figure 6. Though one might think that this assumption is too strong, it is acceptable when the spatial variation of $f(\mathbf{x})$ in S_P is globally uniform.

Figure 6. The function $f_2(\mathbf{x})$ in S_0 and S_B.

We define the statistic Z_{P2} by

$$Z_{P2} = \frac{1}{na} \sum_i \int_{\mathbf{x} \in S_i} f_2(\mathbf{x}) d\mathbf{x}. \tag{28}$$

For convenience of further discussion, we rewrite this equation using the transformation of S_i into S_i' defined by

$$S_i' = \left\{ T(\mathbf{x}) \middle| \mathbf{x} \in S_i \right\}. \tag{29}$$

An example of this transformation is depicted in Figure 7. Equation (28) then becomes

$$Z_{P2} = \frac{1}{na} \sum_i \int_{\mathbf{x} \in S_0} f(\mathbf{x}) g_i(\mathbf{x}) d\mathbf{x}, \tag{30}$$

where

$$g_i(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in S_i', \\ 0 & \mathbf{x} \notin S_i'. \end{cases} \tag{31}$$

Figure 7. Transformation of S_i into S_i'. (a) The polygon S_i, (b) the polygon S_i' generated through the transformation.

The null hypothesis considered here is that the polygons are independently distributed so that all possible shapes and positions of S_i' appear randomly. This hypothesis is equivalent to the case where polygons are independently and randomly distributed at density n/S_0 over the unbounded region where the surface function $f_u(\mathbf{x})$ is defined by

$$f_u(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \mathbf{x} \in S_0, \\ f\left(T(\mathbf{x})\right) & \mathbf{x} \notin S_0 \end{cases} \tag{32}$$

(see Figure 8).

Figure 8. The distribution of polygons and the surface function $f_u(\mathbf{x})$.

Under the null hypothesis, the expectation of Z_{P2} is given by

$$\mathrm{E}\left[Z_{P2}\right] = \frac{1}{na} \sum_i \int_{\mathbf{x} \in S_0} f(\mathbf{x}) \mathrm{E}\left[g_i(\mathbf{x})\right] d\mathbf{x}. \tag{33}$$

The expectation $\mathrm{E}[g_i(\mathbf{x})]$ is the probability of the point located at $\mathbf{x}$ being contained in S_i', which is written as

$$\mathrm{E}\left[g(\mathbf{x})\right] = \frac{m(S;\mathbf{x})}{m\left(S_i';S\right)}, \tag{34}$$

where S is a convex polygon congruent to S_i, and $m(S_i'; S_0)$ is the measure of the set of all polygons congruent to S_i having differing the shape and position of S_i'. The measure $m(S_i'; S_0)$ is given by

$$m\left(S_i';S_0\right) = 2\pi a_0. \tag{35}$$

Substitution of equations (17), (34) and (35) into equation (33) yields

$$\mathrm{E}\left[Z_{P2}\right]=\frac{1}{a_0}\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}. \tag{36}$$

The variance of Z_{P2} is given by

$$\mathrm{V}\left[Z_{P2}\right]=\frac{1}{n^2a^2}\sum_{\mathrm{i}}\int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0} f(\mathbf{x})f(\mathbf{t})\mathrm{E}\left[g_i(\mathbf{x})g_i(\mathbf{t})\right]d\mathbf{x}d\mathbf{t}-\frac{1}{n}\left\{\frac{1}{a_0}\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}\right\}^2. \tag{37}$$

The expectation $\mathrm{E}[g_i(\mathbf{x})g_i(\mathbf{t})]$ indicates the probability that the points located at **x** and **t** are contained in S_i'. Hence it is given by

$$\mathrm{E}\left[g_i(\mathbf{x})g_i(\mathbf{t})\right]=\frac{1}{2\pi a_0}\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} m\left(S;\mathbf{x},T_{jk}(\mathbf{t})\right), \tag{38}$$

When S is very small compared to S_0, equation (38) can be rewritten as

$$\mathrm{E}\left[g_i(\mathbf{x})g_i(\mathbf{t})\right]=\frac{1}{2\pi a_0}\sum_{j=-1}^{1}\sum_{k=-1}^{1} m\left(S;\mathbf{x},T_{jk}(\mathbf{t})\right). \tag{39}$$

Substituting equation (38) into equation (37) yields

$$\mathrm{V}\left[Z_{P2}\right]=\frac{1}{2n\pi a_0a^2}\int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0} f(\mathbf{x})f(\mathbf{t})\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} m\left(S;\mathbf{x},T_{jk}(\mathbf{t})\right)d\mathbf{x}d\mathbf{t}-\frac{1}{n}\left\{\frac{1}{a_0}\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}\right\}^2. \tag{40}$$

Similar to Z_{P1}, Z_{P2} approaches the normal distribution with mean $\mathrm{E}[Z_P]$ and variance $\mathrm{V}[Z_P]$ as n increases. Consequently, using Z_{P2}, we can perform the statistical test based on Z_{P2} to examine the significance of the polygons-surface relationship.

Method 3

In the above two methods, we have explicitly taken account of the edge effect and put assumptions on $f(\mathbf{x})$ to develop tractable methods. Consequently, both Methods 1 and 2 are applicable to the distribution of polygons of any size. Conversely, if the polygons are reasonably small compared with S_0, we can neglect the edge effect so that no assumption on $f(\mathbf{x})$ is required. To deal with this case, we propose Method 3 for analyzing the distribution of small polygons in relation to a surface. This method is useful when S_0 is a convex polygon and the distributed polygons are fairly smaller than S_0.

Regarding the size of S_0 and S_i, we assume:

Assumption 3: The area a is negligible compared to the area a_0.

Mathematically, this assumption is represented as

$$\frac{a}{a_0}\approx 0. \tag{41}$$

Method 3 is obtained by considering Method 1 in the limit $a/a_0\to 0$. Using equation (8), we define statistic Z_{P3} by

$$
\begin{aligned}
Z_{P3} &= \lim_{a/a_0 \to 0} \frac{1}{na} \sum_i \left\{ \int_{\mathbf{x} \in S_0} f(\mathbf{x}) g_i(\mathbf{x}) d\mathbf{x} + \left(a - \int_{\mathbf{x} \in S_0} g_i(\mathbf{x}) d\mathbf{x} \right) f_0 \right\} \\
&= \frac{1}{na} \sum_i \int_{\mathbf{x} \in S_0} f(\mathbf{x}) g_i(\mathbf{x}) d\mathbf{x} + \lim_{a/a_0 \to 0} \frac{f_0}{na} \sum_i \left(a - \int_{\mathbf{x} \in S_0} g_i(\mathbf{x}) d\mathbf{x} \right)
\end{aligned}
\tag{42}
$$

Assumption 3 permits the approximation below.

$$
\int_{\mathbf{x} \in S_0} g_i(\mathbf{x}) d\mathbf{x} \approx a \tag{43}
$$

Substituting equation (43) into equation (42) yields

$$
Z_{P3} = \frac{1}{na} \sum_i \int_{\mathbf{x} \in S_0} f(\mathbf{x}) g_i(\mathbf{x}) d\mathbf{x}. \tag{44}
$$

The expectation and the variance of Z_{P3} under the null hypothesis are obtained from equations (19) and (23) in the limit $a/a_0 \to 0$.

$$
\mathrm{E}\left[Z_{P3} \right] = \frac{1}{a_0} \int_{\mathbf{x} \in S_0} f(\mathbf{x}) d\mathbf{x} \tag{45}
$$

$$
\begin{aligned}
\mathrm{V}\left[Z_{P3} \right] = & \frac{1}{2n\pi a_0 a^2} \int_{\mathbf{t} \in S_0} \int_{\mathbf{x} \in S_0} m(S; \mathbf{x}, \mathbf{t}) f(\mathbf{x}) f(\mathbf{t}) d\mathbf{x} d\mathbf{t} \\
& - \frac{1}{n} \left\{ \frac{1}{a_0} \int_{\mathbf{x} \in S_0} f(\mathbf{x}) d\mathbf{x} \right\}^2
\end{aligned}
\tag{46}
$$

The derivation of these equations is shown in Appendix 2. The significance of Z_{P3} can be tested in a way similar to the one used for Z_{P1}.

We should note that equations (45) and (46) do not include f_0. This confirms that the edge effect disappears when Assumption 3 is reasonable. We should also point out that these equations are identical with equations (36) and (40) when $a/a_0 \approx 0$. This implies that Methods 1 and 2 becomes equivalent as a/a_0 approaches zero.

In this section, we have proposed three methods for analyzing the polygons-surface relationship. The choice depends on the shape and size of the region S_0, the size of the distributed polygons, and the variability of the surface. Method 1 works if S_0 is convex and $f(\mathbf{x})$ is almost constant in S_B. Method 2 is applicable when S_0 is a rectangle and the spatial variation of $f(\mathbf{x})$ in S_B is similar to that in S_0. This implies that these methods work if $f(\mathbf{x})$ is globally constant in S_P. Note that the two methods do not require any assumptions on the sizes of S_0 and polygons. Unlike these methods, Method 3 assumes the sizes of S_0 and polygons instead of the variability of the surface. This method works if S_0 is convex and the polygons are very small compared with S_0. Since distributed polygons analyzed in GIS are usually small, Method 3 seems to be the most useful among three methods.

In the above discussion, we have assumed that all the distributed polygons are

congruent. The proposed methods, however, can be extended to the case where sets of congruent polygons are distributed if each set consists of a large number of elements. We now briefly describe this extension in the case of Method 3.

Let us suppose sets of congruent polygons intersecting S_0. Let Θ_i be the set of n_i congruent polygons S_{i1}, S_{i2}, ..., S_{in_i}, whose area and perimeter are denoted by a_i and l_i, respectively. The spatial relationship between the distribution of polygons belonging to Θ_i and the surface $f(\mathbf{x})$ can be measured by the statistic Z_{Si} which is defined by

$$Z_{Si} = \frac{1}{n_i a_i} \sum_j \int_{\mathbf{x} \in S_0} f(\mathbf{x}) g_{ij}(\mathbf{x}) d\mathbf{x}, \tag{47}$$

where

$$g_{ij}(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in S_{ij} \cap S_0, \\ 0 & \mathbf{x} \notin S_{ij} \cap S_0. \end{cases} \tag{48}$$

Regarding to the distribution of sets of polygons, we define the statistic Z_S as

$$\begin{aligned} Z_S &= \frac{1}{n} \sum_i n_i Z_{Si} \\ &= \frac{1}{n} \sum_i \frac{1}{a_i} \sum_j \int_{\mathbf{x} \in S_0} f(\mathbf{x}) g_{ij}(\mathbf{x}) d\mathbf{x} \end{aligned}, \tag{49}$$

where $n = \sum_i n_i$. Under the null hypothesis that all the polygons are independently and randomly distributed in such a way that they intersect S_0, $\mathrm{E}[Z_S]$ and $\mathrm{V}[Z_S]$ are given by, respectively,

$$\mathrm{E}[Z_S] = \frac{1}{a_0} \int_{\mathbf{x} \in S_0} f(\mathbf{x}) d\mathbf{x} \tag{50}$$

and

$$\begin{aligned} \mathrm{V}[Z_S] = {} & \frac{1}{2n^2 \pi a_0} \sum_i \frac{n_i}{a_i^2} \int_{\mathbf{t} \in S_0} \int_{\mathbf{x} \in S_0} m(S_i; \mathbf{x}, \mathbf{t}) f(\mathbf{x}) f(\mathbf{t}) d\mathbf{x} d\mathbf{t} \\ & - \frac{1}{n} \left\{ \frac{1}{a_0} \int_{\mathbf{x} \in S_0} f(\mathbf{x}) d\mathbf{x} \right\}^2 \end{aligned}, \tag{51}$$

where S_i is a polygon congruent to S_{ij}. The significance of Z_S can be statistically tested when n_i's are reasonably large.

4. DISTRIBUTION OF LINE SEGMENTS AND SURFACE

Using the methods proposed in the preceding section, we analyze the distribution of line segments in relation to a surface. Assume n line segments L_1, L_2, ..., L_n of length l intersecting a region S_0 of area a_0 and perimeter l_0. The function $f(\mathbf{x})$ representing a surface is defined in S_L which is given by

$$\begin{aligned} S_L &= S_0 \oplus C_L \\ &= \left\{ \mathbf{x} + \mathbf{y} \middle| \mathbf{x} \in S_0, \mathbf{y} \in C_L \right\}, \end{aligned} \tag{52}$$

where C_L is the circle with radius l centered at the origin.

We replace the line segments by rectangles R_1, R_2, ..., R_n of sides b and l ($b \le l$) so as to employ the polygons-surface methods, and then consider the limit $b \to 0$ (see Figure 9).

Figure 9. The distributions of line segments, their replacing rectangles, and a surface. The gray shade indicates the surface value.

Let us consider the case of Method 1. The spatial relationship between the distribution of rectangles and the surface is measured by the statistic Z_{P1} given by equation (8). Using Z_{P1}, we define the statistic Z_{L1} as

$$Z_{L1} = \lim_{b \to 0} Z_{P1}. \tag{53}$$

After a few steps of calculation (see Appendix 3 for details), we obtain

$$Z_{L1} = \frac{1}{nl} \sum_i \int_{\mathbf{x} \in L_i} f_1(\mathbf{x}) d\mathbf{x}. \tag{54}$$

This implies that the spatial relationship between the distribution of line segments and the surface can be measured by the curvilinear integral of the surface function on the line segments.

The expectation and the variance of Z_{L1} are given by, respectively,

$$\mathrm{E}[Z_{L1}] = \frac{1}{\pi a_0 + l_0 l} \left\{ \pi \int_{\mathbf{x} \in S_0} f(\mathbf{x}) d\mathbf{x} + l_0 l f_0 \right\} \tag{55}$$

and

$$\begin{aligned} \mathrm{V}[Z_{L1}] = {} & \frac{1}{nl^2(\pi a_0 + l_0 l)} \int_{\mathbf{t} \in S_0} \int_{\mathbf{x} \in S_0} m'(\mathbf{x}, \mathbf{t}) \left\{ f(\mathbf{x}) f(\mathbf{t}) - 2 f_0 f(\mathbf{x}) + f_0^2 \right\} d\mathbf{x} d\mathbf{t} \\ & - \frac{1}{n} \left\{ \frac{\pi}{\pi a_0 + l_0 l} \right\}^2 \left\{ \int_{\mathbf{x} \in S_0} f(\mathbf{x}) d\mathbf{x} - a_0 f_0 \right\}^2 \end{aligned}, \tag{56}$$

where

$$m'(\mathbf{x}, \mathbf{t}) = \begin{cases} \dfrac{l - |\mathbf{x} - \mathbf{t}|}{|\mathbf{x} - \mathbf{t}|} & (|\mathbf{x} - \mathbf{t}| \le l), \\ 0 & (l < |\mathbf{x} - \mathbf{t}|). \end{cases} \tag{57}$$

The derivation of these equations is shown in Appendix 3. Methods 2 and 3 can also be applied in a similar way, and the choice depends on the shape and size of S_0 and the variability of $f(\mathbf{x})$ as in the polygons-surface relationship analysis.

5. EMPIRICAL STUDY

In this section, using the methods proposed above, we empirically analyze the distribution of convenience stores in relation to the population distribution. We are concerned with how the population distribution affects the distribution of convenience stores. The study region was a 4.5 km × 3.0 km rectangular area in a suburb of Osaka, Japan. There were 33 convenience stores, and the total population was around 130,000. The locations of convenience stores were given by their coordinates, and the distribution of population was given in the square lattice data of side 10 meters.

In Japan, it is often said that the locations of convenience stores depend on the distribution of persons aged 20-29. Asano (1993), for instance, visually analyzed the distribution of convenience stores and found that they were located around universities and apartment houses for the students. To test this hypothesis, we classified the people by their age and examine the 10-19, 20-29, and 30-39 age groups in this analysis. The populations of these groups in the study region were around 18000, 22000, and 18000, respectively. The distributions of these age groups are shown in Figure 10. In this figure, the population distributions $f(\mathbf{x})$'s are standardized to allow a visual comparison among different age groups using the equation

$$g(\mathbf{x}) = \frac{f(\mathbf{x}) - \dfrac{1}{a_0}\displaystyle\int_{\mathbf{t}\in S_0} f(\mathbf{t})d\mathbf{t}}{\dfrac{1}{a_0}\displaystyle\int_{\mathbf{t}\in S_0}\{f(\mathbf{t})\}^2 d\mathbf{t} - \left\{\dfrac{1}{a_0}\displaystyle\int_{\mathbf{t}\in S_0} f(\mathbf{t})d\mathbf{t}\right\}^2}, \tag{58}$$

where S_0 is the study region and a is its area. We should note, however, that the analyzing methods themselves do not require such standardization of surface functions.

Figure 10. The distribution of convenience stores and the standardized population distribution $g(\mathbf{x})$. White dots indicate the convenience stores, and the circles centered at the stores are 500 meters buffer regions. Broken line indicates the study region. (a) The 10-19 age group, (b) the 20-29 age group, (c) the 30-39 age group.

Let us discuss the method for analyzing the distribution of convenience stores in relation to the population distribution. Since the former is a point distribution while the latter is given as a surface, one possible choice is to use the points-surface method proposed in Section 2. The points-surface method, however, seems to be inappropriate for this analysis because this method is meaningful when the point distribution is locally determined by the surface (similar discussion is found in Okabe and Sadahiro, 1994).

The customers of retail stores are generally distributed in fairly large regions around the stores, which suggests that the distribution of stores is affected by the global distribution of population rather than its local distribution. We hence employed the distribution of market areas of the convenience stores instead of the distribution of stores themselves, and analyzed it in relation to the population distribution using the points-surface method proposed in Section 3. This enables us to consider explicitly the spatial influence of the population distribution on the distribution of stores. We assume that the market areas are the circles of radius r centered at the convenience stores, and adopted Methods 1 and 2 because it does not seem that the market areas are so small compared with the study region.

Methods 1 and 2 are based on several assumptions as mentioned earlier. Both methods require the normal approximation of the probability distribution of the statistics, and assume the variability of $f(\mathbf{x})$ in S_B. To examine the validity of the normal approximation, we first performed the Monte Carlo simulation and investigated how the number and size of circles affect the degree of the approximation. The procedure of this simulation was as follows:

Step 1) Fix the method and age group.

Step 2) Fix n and r, the number and radius of circles.

Step 3) Give the location of circles following the uniform distribution.

Step 4) Calculate the value of the statistics.

We tried various numbers of n ranging from 1 to 30 and various values of r from 100 to 1000 meters for every method and age group (Step 2). Given the method, age group, n and r, we repeated Steps 3 and 4 (10,000 iterations). The constant value f_0 used in Method 1 was given by the average density of population in S_B, namely,

$$f_0 = \frac{\int_{\mathbf{x}\in S_B} f(\mathbf{x})d\mathbf{x}}{\int_{\mathbf{x}\in S_B} d\mathbf{x}}. \tag{59}$$

The obtained probability density distributions are depicted in Figure 11. The distributions were standardized so that the degree of the approximation can be easily understood. Since the results did not differ much with methods and age groups, we show only the results for Method 1 where the distribution of 20-29 age group was employed as $f(\mathbf{x})$.

Figure 11. The standardized probability density distributions of Z_{P1}. The distribution of 20-29 age group is used as $f(\mathbf{x})$. The radius of the circles is (a) 100 meters, (b) 300 meters, (c) 500 meters. The gray shade indicates the standard normal distribution.

Figure 11 shows that the probability density distributions of Z_{P1} approach to the normal distribution as the number and size of circles increase. For the circles of radius 100 meters, it is difficult to assume the normality if $n \leq 10$, but if $n \geq 20$, the normal approximation seems acceptable. For the circles more than 300 meters in radius, we cannot reject the normality assumption if $n \geq 10$. Since $n = 33$ in the present study, it is appropriate to employ the normal approximation.

We next investigated whether Assumptions 1 and 2 are acceptable in practice. To this end, we calculated the statistics based on $f(\mathbf{x})$, the true distribution of population, and compared the results with those obtained by Methods 1 and 2 in a standardized form. Since the expectations and variances of the statistics based on $f(\mathbf{x})$ are not given explicitly, we employed the Monte Carlo simulations (10000 iterations) to obtain the probability distributions of the statistics. As a result, we performed the analysis using the four methods shown below.

Method	Function representing the surface
Method 1	$f_1(\mathbf{x})$
Method 1'	$f(\mathbf{x})$
Method 2	$f_2(\mathbf{x})$
Method 2'	$f(\mathbf{x})$

For the radius r, we tried values from 300 to 700 meters incremented by 10 meters. This is because people living in Japan usually go to convenience stores by walk, and it is said that the stores have circular market areas of radius around 500 meters. In Method 2', we considered the null hypothesis that the convenience stores are distributed independently and randomly in the region S_0, which implies that the circles representing market areas are distributed in S_P' depicted in Figure 12. We divided the study region by the lattice of squares of side 10 meters to compute the variances of statistics by numerical integration.

Figure 12. The region S_P' where circles are distributed.

The results are shown in Figures 13 and 14. To allow a comparison across different methods and age groups, we standardized the statistics for each method and group using

$$z_{Pi} = \frac{Z_{Pi} - \mathrm{E}\left[Z_{Pi}\right]}{\sqrt{\mathrm{V}\left[Z_{Pi}\right]}} \quad (i = 1, 2). \tag{60}$$

Figure 13. The results of the analysis given by Methods 1 and 1'. Solid lines and dotted lines indicate the values of z_{P1} given by Method 1 and Method 1', respectively. Critical values of z_{P1} at significance levels of 1% and 5 % are shown as $z_{0.01}$ and $z_{0.05}$, respectively.

Figure 14. The results of the analysis given by Methods 2 and 2'. Solid lines and dotted lines indicate the values of z_{P2} given by Method 2 and Method 2', respectively. Critical values of z_{P2} at significance levels of 1% and 5 % are shown as $z_{0.01}$ and $z_{0.05}$, respectively.

Let us compare the results given by Methods 1 and 2 (solid lines) with those given by Methods 1' and 2' (dotted lines). In Figure 13, we notice that the distance between the solid and dotted lines increases as r becomes larger, which implies that the difference between the results given by Methods 1 and 1' increases. This is mainly because S_B, the region in which $f(\mathbf{x})$ and $f_1(\mathbf{x})$ have different values, expands as the circles become larger. When r is smaller than 500 meters, however, the difference between the two lines is not significant, hence we can say that Assumption 1 is reasonable when the circles are smaller than 500 meters in radius. Unlike to Figure 13, Figure 14 does not show such a significant difference between the solid and dotted lines (note that the vertical scales of Figures 13 and 14 are different). This suggests that Assumption 2 is allowable regardless of the size of polygons, and thus Method 2 seems to be more appropriate than Method 1 in this empirical study.

Using Figure 14, we now examine how the population distribution affects the distribution of convenience stores. In this figure, we notice that the 20-29 age group shows the largest values of z_{P2} among the three groups for all r, and they are larger than the critical value at a significance level of 1%. This suggests that the distribution of the 20-29 age group is the most influential on the distribution of convenience stores among the three groups. Compared with this group, the 30-39 age group is less influential though they also show large values of z_{P2}. The distribution of the 10-19 age group, on the other hand, does not seem to affect the distribution of convenience stores. The null hypothesis that the convenience stores are distributed independent of the distribution of the 10-19 age group cannot be rejected at a significance level of 5% for all r.

We finally discuss the size of market areas of the convenience stores. Regarding the 20-29 and 30-39 age groups, z_{P2} values are larger than the critical value at a significance level of 5% at r from 500 to 700 meters. This suggests that the convenience stores are located so that the population of 20-39 years living within a radius of 500-700 meters from the stores is large enough. Though this result is not inconsistent with the

hypothetical size of market areas (radius around 500 meters), it suggests that convenience stores may have larger market areas.

6. CONCLUDING DISCUSSION

In this paper, we have proposed statistical methods for analyzing the distribution of spatial objects in relation to a surface. Using these methods, we can analyze the distribution of points, convex polygons, and line segments which are most frequently used in GIS in relation to a surface.

Choice of analyzing methods depends, of course, on the type of spatial objects. The object type, however, can be converted if necessary, as shown in the empirical study. If the distribution of points is locally determined by a surface, the points-surface method is applicable. Conversely, if the distribution has global relationship with the surface, it might be better to convert the points into polygons through a transformation method such as the buffering operation and apply the polygons-surface method. If we wish to analyze a polygon distribution in relation to a surface by the points-surface method, we can replace the polygons by their representative points such as the gravity centers. For further details of the transformation methods, see Bohman-Carter (1994), for instance.

The GIS data treated as a surface includes not only the continuous variables but the discrete and binary variables distributed continuously over a region. Hence, by regarding polygons as the surface of a binary variable, we can analyze the distribution of convex polygons in relation to the distribution of polygons of different shapes using the polygons-surface method. Similarly, we can treat the relationship between the distribution of line segments and the distribution of incongruent polygons.

The methods proposed in this paper, however, have some limitations. We finally discuss them with possible modifications. First, the proposed methods are not applicable in a strict sense to the distribution of spatial objects interacting with each other, either repulsively (spatial avoidance) or attractively (spatial clustering). If we analyze the distribution of such objects, the obtained results may be erroneous to some extent because the null hypothesis considered in the statistical tests implicitly assume the independency of spatial objects. For instance, the null hypothesis may be rejected if spatial objects tend to gather though they are distributed independent of the surface. In usual empirical analyses, however, we do not know the spatial interaction among spatial objects a priori. Hence, if it seems that the objects are strongly interacting with each other, we should lower the significance level to avoid the error of the test. Second, the methods proposed in Sections 3 and 4 do not work for the distribution of spatial objects

which are not allowed to intersect with each other. This is because the inhibition of intersection involves a strong interaction among spatial objects, which violates the independency assumption mentioned above. A new method to treat this case should be developed in further researches.

References

Asano, A. (1993) *Time-Series Analysis of the Distribution of Convenience Stores in Hiratsuka*. Graduation Thesis, Department of Urban Engineering, University of Tokyo (in Japanese).

Bohman-Carter, G. F. (1994) *Geographic Information Systems for Geoscientists.* Pergamon, Oxford.

Cressie, N. (1993) *Statistics for Spatial Data.* John Wiley, New York.

Cuzick, J. and Edwards, R. (1990) Spatial Clustering for Inhomogeneous Populations. *Journal of the Royal Statistical Society, Series B*, **52**, 73-104.

de Lepper, M. J. C, Scholten, H. J. and Stern, R. M. (1995) *The Added Value of Geographical Information Systems in Public and Environmental Health.* Kluwer, Dordrecht.

de Savigny, D. and Wijeyaratne, P. (1995) *GIS for Health and the Environment.* International Development Research, Ottawa.

Gatrell, A. and Rowlingson, B. (1994) Spatial Point Process Modelling in a GIS Environment. In *Spatial Analysis and GIS* (eds. S. Fotheringham & P. Rogerson). Taylor and Francis, London pp.147-163.

Koshizuka, T. and Ooki, T. (1982) On the Relative Density of Bridges. *Papers of the Annual Conference of the City Planning Institute of Japan*, **17**, 91-96 (in Japanese).

Lee, Y. (1979) A Nearest-neighbor Spatial-association Measure for the Analysis of Firm Interdependency. *Environment and Planning A*, **11**, 169-176.

Okabe, A. and Fujii, A. (1984) The Statistical Analysis through a Computational Method of a Distribution of Points in Relation to its Surrounding Network. *Environment and Planning A*, **16**, 107-114.

Okabe, A. and Miki, F. (1984) A Conditional Nearest-neighbor Spatial-association Measure for the Analysis of Locational Interdependency. *Environment and Planning A*, **16**, 163-171.

Okabe, A., Yoshikawa, T., Fujii, A., and Oikawa, K. (1988) The Statistical Analysis of a Distribution of Activity Points in Relation to Surface-like Elements. *Environment and Planning A*, **20**, 609-620.

Okabe, A. and Sadahiro, Y. (1994) A Statistical Method for Analyzing the Spatial Relationship between the Distribution of Activity Points and the Distribution of

Activity Continuously Distributed over a Region. *Geographical Analysis*, **26**, 152-167.

Pielou, E. C. (1961) Segregation and Symmetry in Two-species Populations as Studied by Nearest Neighbor Relations. *Journal of Ecology*, **49**, 255-269.

Ripley, B. D. (1981) *Spatial Statistics*. John Wiley, New York.

Santaló, L. A. (1976) *Integral Geometry and Geometric Probability*. Addison-Wesley, London.

Stoyan, D. and Stoyan, H. (1994) *Fractals, Random Shapes and Point Fields*. John Wiley, New York.

APPENDIX 1. THE MEASURE OF ALL POSITIONS OF A CONVEX POLYGON CONTAINING TWO POINTS

Suppose a convex polygon S of area a and perimeter l. Let $m(S; \mathbf{x}, \mathbf{t})$ be the measure of all convex polygons congruent to S that contain the two points located at $\mathbf{x}$ and $\mathbf{t}$, respectively. This measure can easily be calculated if S has a simple shape. When S is a circle of radius r, the measure is directly given by the multiplication of 2π and the area of intersection of two circles of radius r centered at $\mathbf{x}$ and $\mathbf{t}$.

$$m(S;\mathbf{x},\mathbf{t}) = \begin{cases} 4\pi r^2 \arccos\left(\dfrac{|\mathbf{x}-\mathbf{t}|}{2r}\right) - \pi|\mathbf{x}-\mathbf{t}|\sqrt{4r^2-|\mathbf{x}-\mathbf{t}|^2} & (|\mathbf{x}-\mathbf{t}| \le 2r), \\ 0 & (|\mathbf{x}-\mathbf{t}| > 2r). \end{cases} \tag{A 1}$$

For a rectangle of sides b, c ($b \le c$), we have

$$m(S;\mathbf{x},\mathbf{t}) = \begin{cases} 2\pi bc - 4(b+c)|\mathbf{x}-\mathbf{t}| + 2|\mathbf{x}-\mathbf{t}|^2 & (|\mathbf{x}-\mathbf{t}| \le b), \\ 4c\sqrt{|\mathbf{x}-\mathbf{t}|^2 - b^2} - 4c|\mathbf{x}-\mathbf{t}| - 2b^2 + 4bc\arcsin\dfrac{b}{|\mathbf{x}-\mathbf{t}|} & (b < |\mathbf{x}-\mathbf{t}| \le c), \\ 4c\sqrt{|\mathbf{x}-\mathbf{t}|^2 - b^2} + 4b\sqrt{|\mathbf{x}-\mathbf{t}|^2 - c^2} - 2\left(b^2 + c^2 + |\mathbf{x}-\mathbf{t}|^2\right) & \\ \quad + 4bc\left(\arcsin\dfrac{c}{|\mathbf{x}-\mathbf{t}|} - \arccos\dfrac{b}{|\mathbf{x}-\mathbf{t}|}\right) & \left(c < |\mathbf{x}-\mathbf{t}| \le \sqrt{b^2+c^2}\right), \\ 0 & \left(\sqrt{b^2+c^2} < |\mathbf{x}-\mathbf{t}|\right). \end{cases} \tag{A 2}$$

Substituting b=c=d into equation (A 2) yields the measure of a square of side d:

$$m(S;\mathbf{x},\mathbf{t}) = \begin{cases} 2\pi d^2 - 8d|\mathbf{x}-\mathbf{t}| + 2|\mathbf{x}-\mathbf{t}|^2 & (|\mathbf{x}-\mathbf{t}| \le d), \\ 8d\sqrt{|\mathbf{x}-\mathbf{t}|^2 - d^2} + 2(\pi-2)d^2 - 2|\mathbf{x}-\mathbf{t}|^2 - 8d^2\arccos\dfrac{d}{|\mathbf{x}-\mathbf{t}|} & \left(d < |\mathbf{x}-\mathbf{t}| \le \sqrt{2}d\right), \\ 0 & \left(\sqrt{2}d < |\mathbf{x}-\mathbf{t}|\right). \end{cases} \tag{A 3}$$

For other convex polygons, a procedure of calculating $m(S; \mathbf{x}, \mathbf{t})$ is given by Santaló (1976).

APPENDIX 2. CALCULATION OF E[Z_{P3}] AND V[Z_{P3}]

The expectation of Z_{P3} is given as follows. Using equation (19), we have

$$\begin{aligned}\mathrm{E}\left[Z_{P3}\right] &= \lim_{a/a_0\to 0}\frac{a}{2\pi\left(a_0+a\right)+l_0 l}\left\{2\pi\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}+\left(2\pi a+l_0 l\right)f_0\right\} \\ &= \lim_{a/a_0\to 0}\frac{a}{2\pi\left(1+O\left(\frac{a}{a_0}\right)\right)+O\left(\sqrt{\frac{a}{a_0}}\right)}\left\{\frac{2\pi}{a_0}\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}+\left(2\pi O\left(\frac{a}{a_0}\right)+O\left(\sqrt{\frac{a}{a_0}}\right)\right)f_0\right\}.\end{aligned} \quad \text{(A 4)}$$

Substituting

$$\lim_{a/a_0\to 0} O\left(\frac{a}{a_0}\right)=\lim_{a/a_0\to 0} O\left(\sqrt{\frac{a}{a_0}}\right)=0, \quad \text{(A 5)}$$

we obtain

$$\mathrm{E}\left[Z_{P3}\right]=\frac{a}{a_0}\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}. \quad \text{(A 6)}$$

The variance of Z_{P3} is derived in the following way. From equations (13) and (23), we have

$$\begin{aligned}\mathrm{V}\left[Z_{P3}\right] = \lim_{a/a_0\to 0}\Bigg[&\frac{1}{na^2}\int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0}\mathrm{E}\left[g(\mathbf{x})g(\mathbf{t})\right]\left\{f(\mathbf{x})f(\mathbf{t})-2f_0 f(\mathbf{x})+f_0^2\right\}d\mathbf{x}d\mathbf{t} \\ &-\frac{1}{n}\left\{\frac{2\pi}{2\pi\left(a_0+a\right)+l_0 l}\right\}^2\left\{\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}-a_0 f_0\right\}^2\Bigg]\end{aligned}. \quad \text{(A 7)}$$

Substituting equation (24), we rewrite equation (A 7) as

$$\begin{aligned}\mathrm{V}\left[Z_{P3}\right] = \lim_{a/a_0\to 0}\Bigg[&\frac{1}{na^2\left\{2\pi\left(a_0+a\right)+l_0 l\right\}}\int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0} m(S;\mathbf{x},\mathbf{t})f(\mathbf{x})f(\mathbf{t})d\mathbf{x}d\mathbf{t} \\ &-\frac{1}{n}\left\{\frac{2\pi}{2\pi\left(a_0+a\right)+l_0 l}\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}\right\}^2+\frac{2f_0}{n}\delta_1(a)+\frac{f_0^2}{n}\delta_2(a)\Bigg]\end{aligned}, \quad \text{(A 8)}$$

where

$$\delta_1(a)=a_0\left\{\frac{2\pi}{2\pi\left(a_0+a\right)+l_0 l}\right\}^2\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}-\frac{1}{a^2}\int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0}\mathrm{E}\left[g(\mathbf{x})g(\mathbf{t})\right]f(\mathbf{x})d\mathbf{x}d\mathbf{t} \quad \text{(A 9)}$$

and

$$\delta_2(a)=\frac{1}{a^2}\int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0}\mathrm{E}\left[g(\mathbf{x})g(\mathbf{t})\right]d\mathbf{x}d\mathbf{t}-a_0^2\left\{\frac{2\pi}{2\pi\left(a_0+a\right)+l_0 l}\right\}^2. \quad \text{(A 10)}$$

In the limit $a/a_0\to 0$, $\delta_1(a)$ reduces to

$$\begin{aligned}\lim_{a/a_0\to 0}\delta_1(a) &= \lim_{a/a_0\to 0}\left[a_0\left\{\frac{2\pi}{2\pi\left(a_0+a\right)+l_0 l}\right\}^2\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}-\frac{1}{a^2}\int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0}\mathrm{E}\left[g(\mathbf{x})g(\mathbf{t})\right]f(\mathbf{x})d\mathbf{x}d\mathbf{t}\right] \\ &= \frac{1}{a_0}\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x}-\lim_{a/a_0\to 0}\frac{1}{a}\int_{\mathbf{x}\in S_0} f(\mathbf{x})\mathrm{E}\left[g(\mathbf{x})\right]d\mathbf{x} \\ &= 0\end{aligned}. \quad \text{(A 11)}$$

Similarly,

$$\lim_{a/a_0\to 0}\delta_2(a)=\lim_{a/a_0\to 0}\left[\frac{1}{a^2}\int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0}\mathrm{E}\left[g(\mathbf{x})g(\mathbf{t})\right]d\mathbf{x}d\mathbf{t}-a_0^2\left\{\frac{2\pi}{2\pi(a_0+a)+l_0 l}\right\}^2\right]$$
$$=\lim_{a/a_0\to 0}\frac{1}{a^2}\mathrm{E}\left[\left\{\int_{\mathbf{x}\in S_0}g(\mathbf{x})d\mathbf{x}\right\}^2\right]-1 \quad . \quad \text{(A 12)}$$
$$=0$$

Substituting these equations into equation (A 8), we obtain

$$\mathrm{V}\left[Z_{P3}\right]=\lim_{a/a_0\to 0}\left[\frac{1}{na^2\left\{2\pi(a_0+a)+l_0 l\right\}}\int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0}m(S;\mathbf{x},\mathbf{t})f(\mathbf{x})f(\mathbf{t})d\mathbf{x}d\mathbf{t}\right.$$
$$\left.-\frac{1}{n}\left\{\frac{2\pi}{2\pi(a_0+a)+l_0 l}\int_{\mathbf{x}\in S_0}f(\mathbf{x})d\mathbf{x}\right\}^2\right] \quad . \quad \text{(A 13)}$$
$$=\frac{1}{2n\pi a_0 a^2}\int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0}m(S;\mathbf{x},\mathbf{t})f(\mathbf{x})f(\mathbf{t})d\mathbf{x}d\mathbf{t}-\frac{1}{n}\left\{\frac{1}{a_0}\int_{\mathbf{x}\in S_0}f(\mathbf{x})d\mathbf{x}\right\}^2$$

APPENDIX 3. CALCULATION OF Z_{L1}, E[Z_{L1}], AND V[Z_{L1}]

Suppose a rectangle R of sides b, l ($b \le l$) and define the coordinate system whose axes are parallel to the sides of R and origin is located at the lower-left vertex of R. (Figure A 1). Let us denote the lower side of R as L.

Figure A 1. The rectangle R and the coordinate system.

We consider the integral of $f(\mathbf{x})$ over R given by

$$I_R(\mathbf{x})=\frac{1}{bl}\int_{\mathbf{x}\in R}f(\mathbf{x})d\mathbf{x} \quad \text{(A 14)}$$

and the integral of $f(\mathbf{x})$ on L,

$$I_L(\mathbf{x})=\frac{1}{l}\int_{\mathbf{x}\in L}f(\mathbf{x})d\mathbf{x}. \quad \text{(A 15)}$$

In the limit $b\to 0$, equation (A 14) reduces to equation (A 15) as below.

$$\lim_{b\to 0}I_R(\mathbf{x})=\lim_{b\to 0}\frac{1}{bl}\int_{\mathbf{x}\in R}f(\mathbf{x})d\mathbf{x}$$
$$=\frac{\partial}{\partial b}\frac{1}{l}\int_0^b\int_0^l f(\mathbf{x})dxdy \quad \text{(A 16)}$$
$$=I_L(\mathbf{x})$$

Applying equation (A 16) to equation (53), we obtain

$$
\begin{aligned}
Z_{L1} &= \lim_{b\to 0}\frac{1}{nbl}\sum_i \int_{\mathbf{x}\in S_i} f_1(\mathbf{x})d\mathbf{x} \\
&= \frac{1}{nl}\sum_i \int_{\mathbf{x}\in L_i} f_1(\mathbf{x})d\mathbf{x}
\end{aligned}
\quad . \tag{A 17}
$$

The expectation and the variance of Z_{L1} are calculated as follows.

$$
\begin{aligned}
\mathrm{E}\left[Z_{L1}\right] &= \lim_{b\to 0}\mathrm{E}\left[Z_{P1}\right] \\
&= \frac{1}{\pi a_0 + l_0 l}\left\{\pi \int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x} + l_0 l f_0\right\}
\end{aligned}
\tag{A 18}
$$

Similarly,

$$
\begin{aligned}
\mathrm{V}\left[Z_{L1}\right] &= \lim_{b\to 0}\mathrm{V}\left[Z_{P1}\right] \\
&= \frac{1}{2nl^2(\pi a_0 + l_0 l)}\int_{\mathbf{t}\in S_0}\int_{\mathbf{x}\in S_0}\left\{\lim_{b\to 0}\frac{m(R;\mathbf{x},\mathbf{t})}{b^2}\right\}\left\{f(\mathbf{x})f(\mathbf{t}) - 2f_0 f(\mathbf{x}) + f_0^2\right\}d\mathbf{x}d\mathbf{t} \cdot \text{(A} \\
&\quad -\frac{1}{n}\left\{\frac{\pi}{\pi a_0 + l_0 l}\right\}^2\left\{\int_{\mathbf{x}\in S_0} f(\mathbf{x})d\mathbf{x} - a_0 f_0\right\}^2
\end{aligned}
$$

19)

In the limit $b\to 0$, $\dfrac{m(R;\mathbf{x},\mathbf{t})}{b^2}$ reduces to

$$
\lim_{b\to 0}\frac{m(R;\mathbf{x},\mathbf{t})}{b^2} = \begin{cases} \lim_{b\to 0}\dfrac{\varphi(b)}{b^2} & (|\mathbf{x}-\mathbf{t}| \le l), \\ 0 & (l < |\mathbf{x}-\mathbf{t}|), \end{cases} \tag{A 20}
$$

where $\varphi(b)$ is given by

$$
\varphi(b) = 4l\sqrt{|\mathbf{x}-\mathbf{t}|^2 - b^2} - 4l|\mathbf{x}-\mathbf{t}| - 2b^2 + 4bl\arcsin\frac{b}{|\mathbf{x}-\mathbf{t}|}. \tag{A 21}
$$

Applying a Taylor series expansion to $\varphi(b)$, we obtain

$$
\begin{aligned}
\varphi(b) &= \varphi(0) + \varphi'(0)b + \frac{1}{2}\varphi''(0)b^2 + O\left(b^3\right) \\
&= \frac{2\left(l - |\mathbf{x}-\mathbf{t}|\right)}{|\mathbf{x}-\mathbf{t}|}b^2 + O\left(b^3\right)
\end{aligned}
\quad . \tag{A 22}
$$

Hence,

$$
\begin{aligned}
\lim_{b\to 0}\frac{\varphi(b)}{b^2} &= \lim_{b\to 0}\left\{\frac{2\left(l - |\mathbf{x}-\mathbf{t}|\right)}{|\mathbf{x}-\mathbf{t}|} + O(b)\right\} \\
&= \frac{2\left(l - |\mathbf{x}-\mathbf{t}|\right)}{|\mathbf{x}-\mathbf{t}|}
\end{aligned}
\quad . \tag{A 23}
$$

Using equations (A 19), (A 20) and (A 23), we obtain

$$\mathrm{V}\left[Z_{L1}\right]=\frac{1}{nl^{2}\left(\pi a_{0}+l_{0}l\right)}\int_{\mathbf{t}\in S_{0}}\int_{\mathbf{x}\in S_{0}}m'\left(\mathbf{x},\mathbf{t}\right)\left\{f\left(\mathbf{x}\right)f\left(\mathbf{t}\right)-2\,f_{0}f\left(\mathbf{x}\right)+\,f_{0}^{2}\right\}d\mathbf{x}d\mathbf{t}$$
$$-\frac{1}{n}\left\{\frac{\pi}{\pi a_{0}+l_{0}l}\right\}^{2}\left\{\int_{\mathbf{x}\in S_{0}}f\left(\mathbf{x}\right)d\mathbf{x}-a_{0}f_{0}\right\}^{2}\quad,\qquad\text{(A 24)}$$

where

$$m'\left(\mathbf{x},\mathbf{t}\right)=\begin{cases}\dfrac{l-\left|\mathbf{x}-\mathbf{t}\right|}{\left|\mathbf{x}-\mathbf{t}\right|} & \left(\left|\mathbf{x}-\mathbf{t}\right|\le l\right),\\ 0 & \left(l<\left|\mathbf{x}-\mathbf{t}\right|\right).\end{cases}\qquad\text{(A 25)}$$

FIGURES

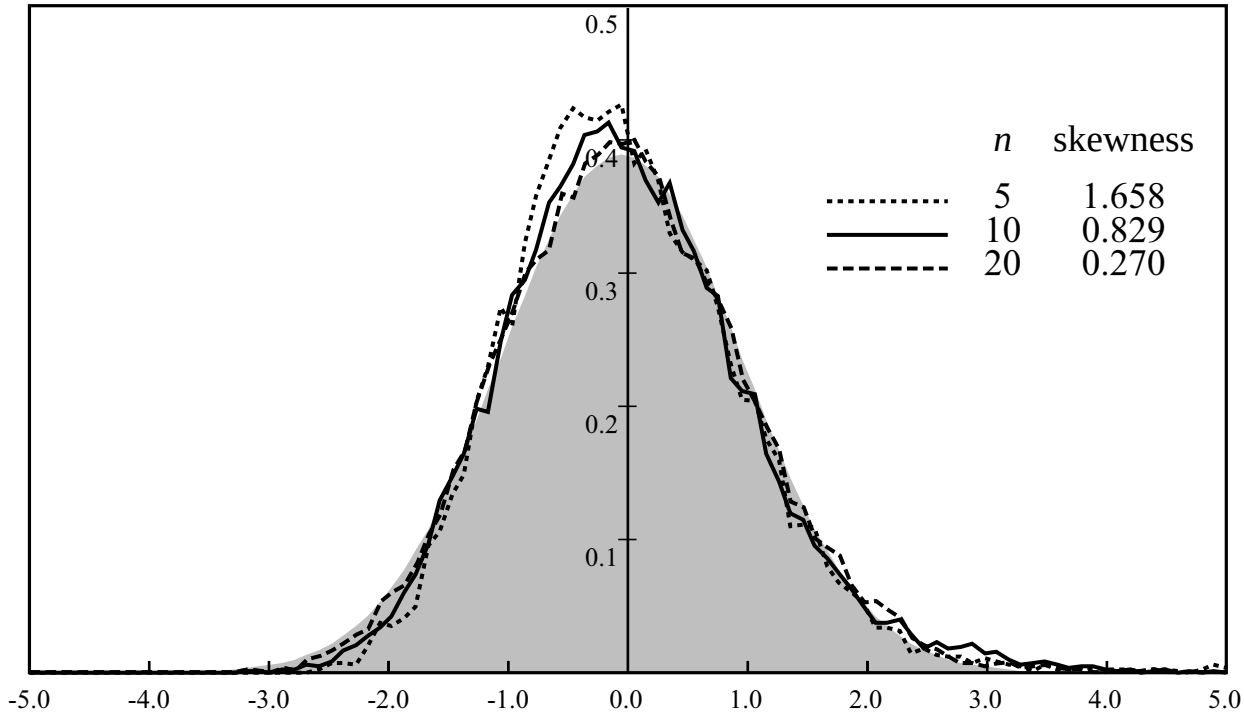
0.5
0.4
0.3
0.2
0.1
-5.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
5.0
n
skewness
5
1.658
10
0.829
20
0.270

Figure 11a

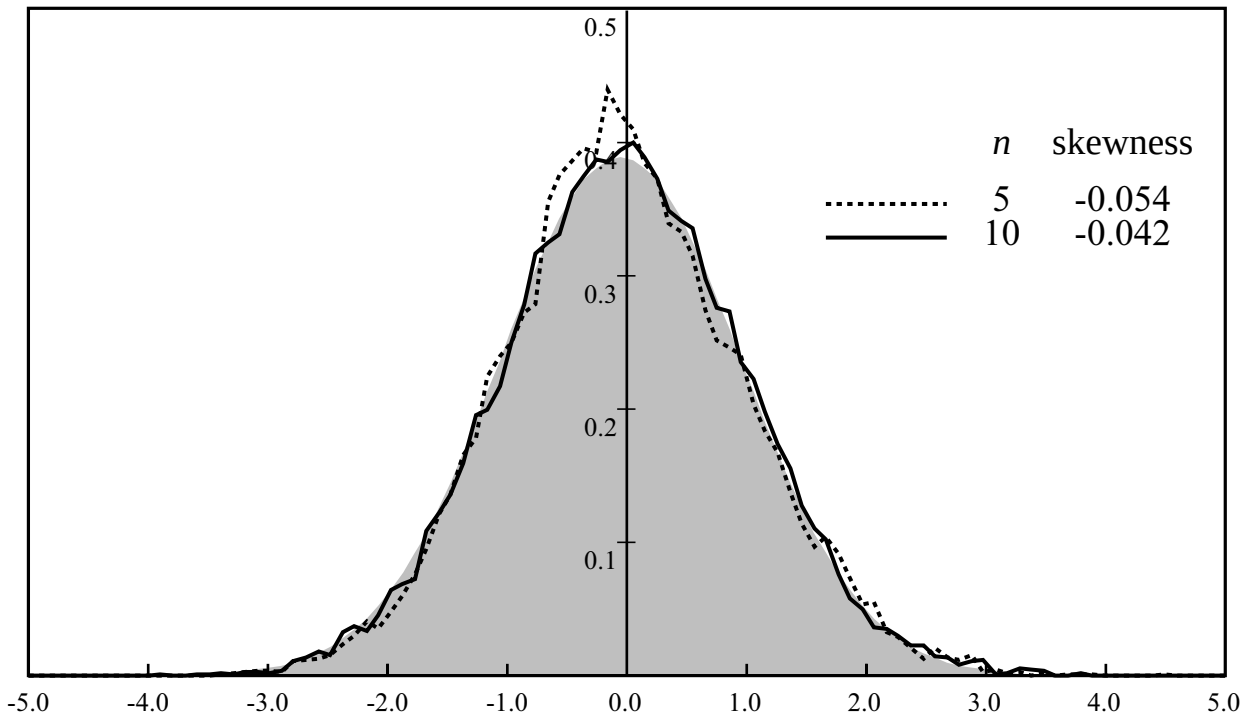
0.5
0.3
0.2
0.1
-5.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
5.0
n
skewness
5
-0.054
10
-0.042

Figure 11b

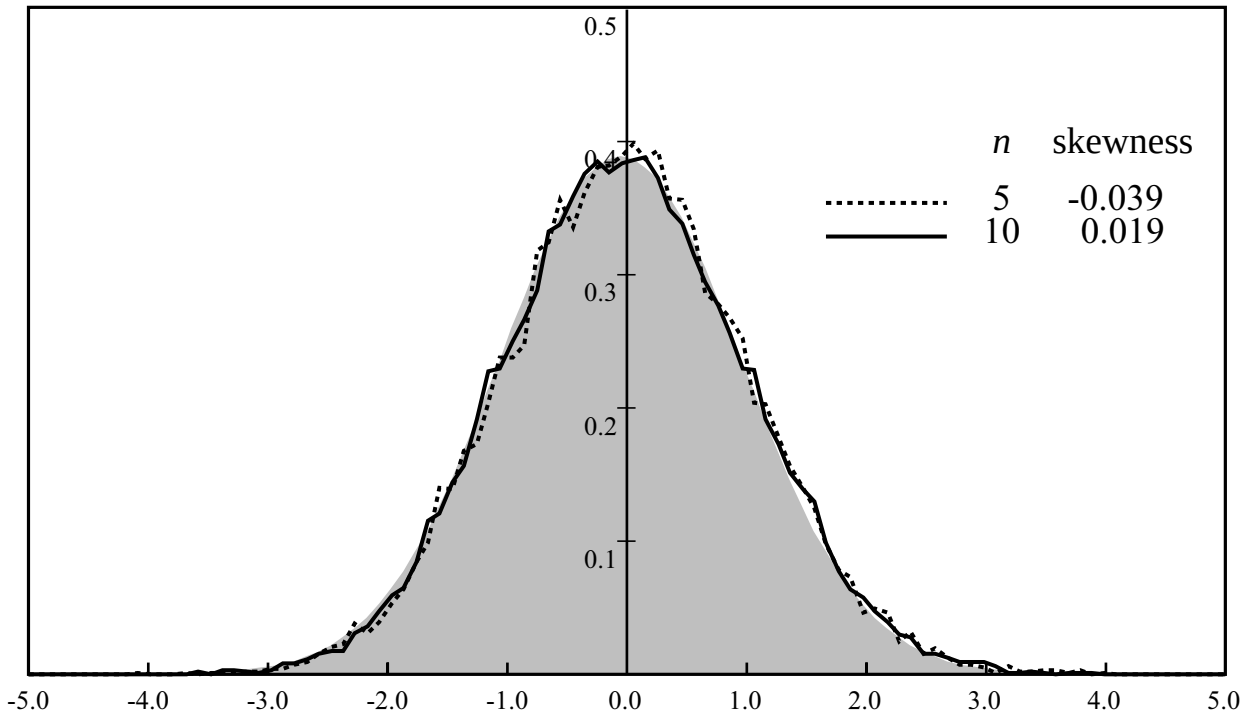
0.5
0.4
0.3
0.2
0.1
-5.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
5.0
n skewness
5 -0.039
10 0.019

Figure 11c

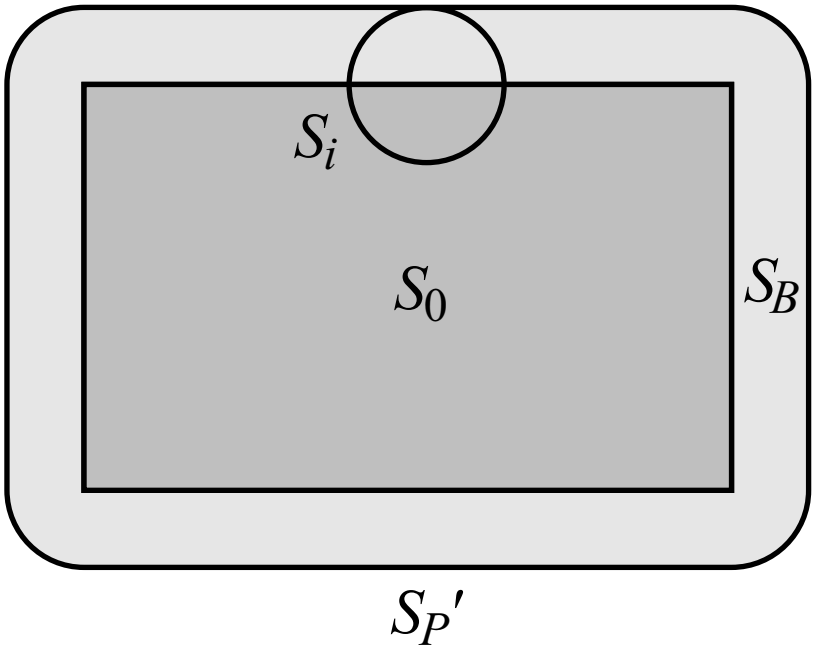

Figure 12

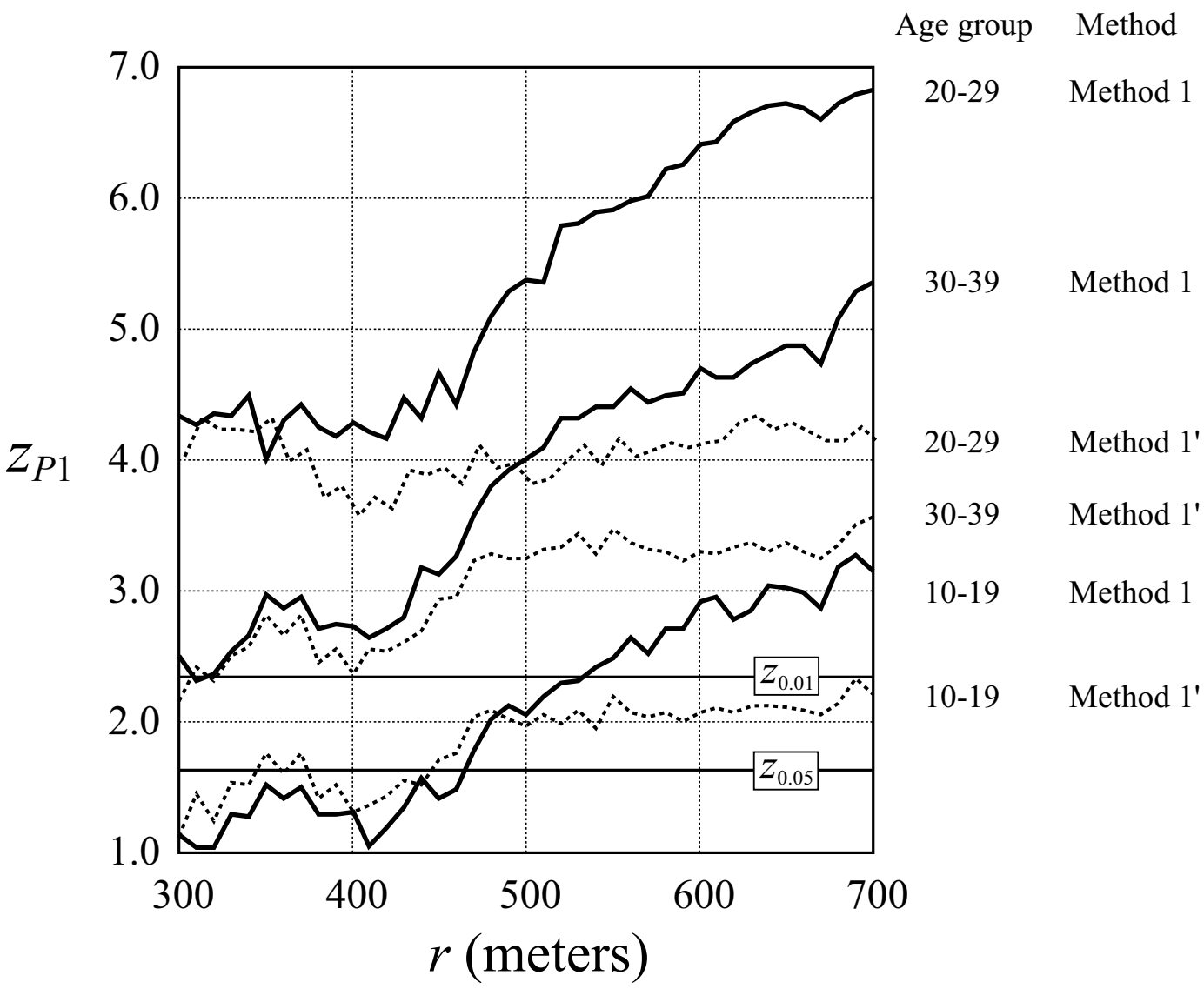
Age group
Method
20-29
Method 1
30-39
Method 1
20-29
Method 1'
30-39
Method 1'
10-19
Method 1
10-19
Method 1'
Z_{P1}
7.0
6.0
5.0
4.0
3.0
2.0
1.0
$z_{0.01}$
$z_{0.05}$
300
400
500
600
700
r (meters)

Figure 13

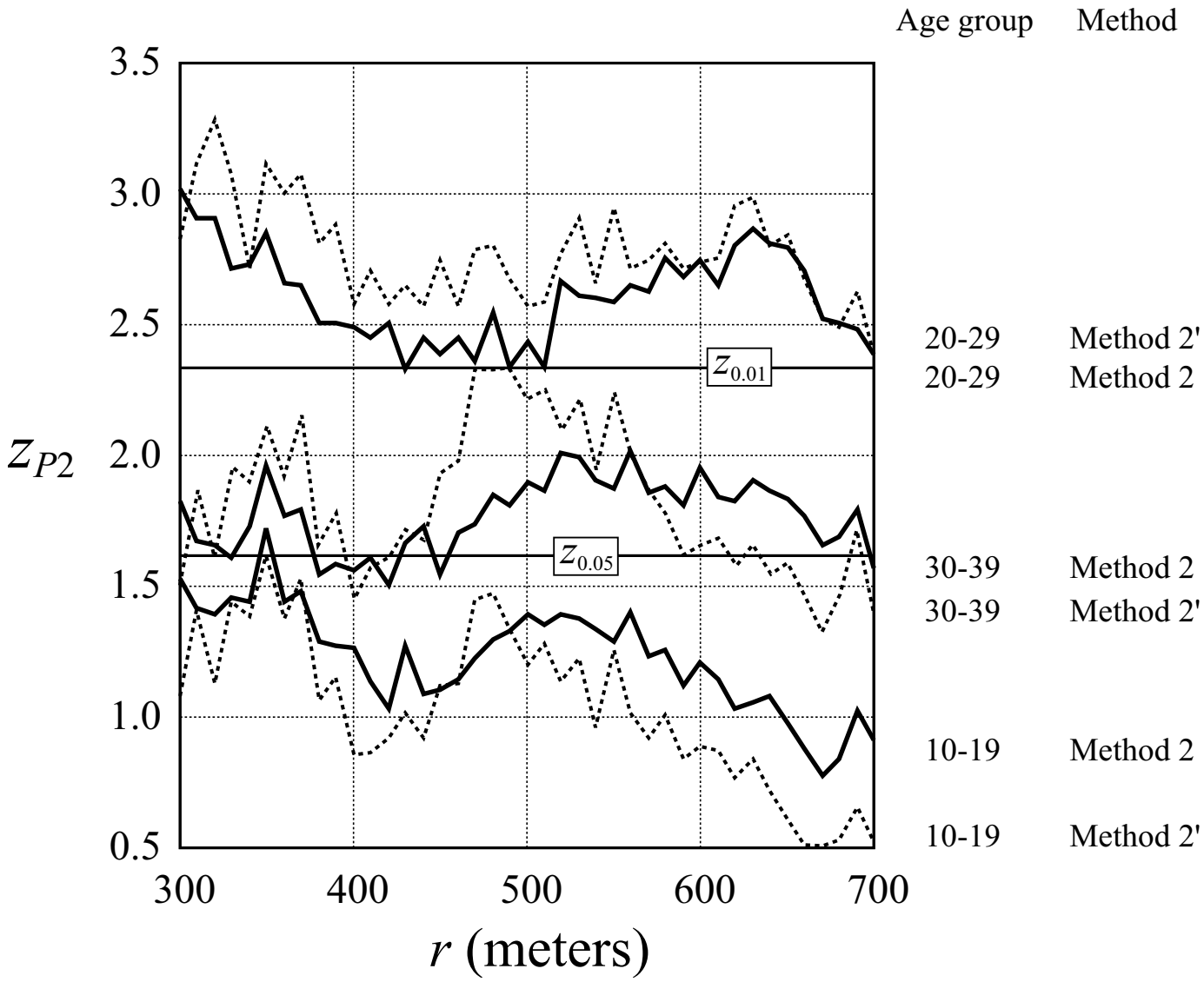
Age group
Method
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Z_{P2}
$z_{0.01}$
$z_{0.05}$
20-29 Method 2'
20-29 Method 2
30-39 Method 2
30-39 Method 2'
10-19 Method 2
10-19 Method 2'
300
400
500
600
700
r (meters)

Figure 14

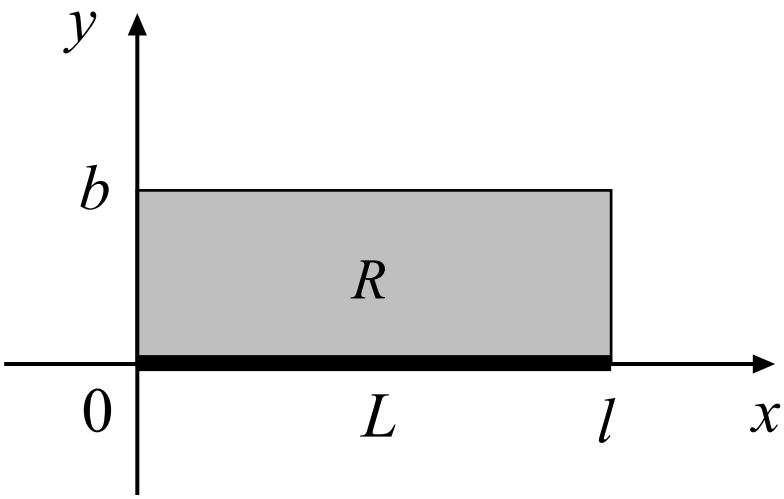

Figure A1