CSIS Discussion Paper No. 6

Accuracy of Count Data Estimated by the Point-in-Polygon Method

Yukio Sadahiro

JANUARY, 1999

Center for Spatial Information Science and Department of Urban Engineering
University of Tokyo
7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan

Accuracy of Count Data Estimated by the Point-in-Polygon Method

Abstract

This paper analyzed the accuracy of count data estimated by the point-in-polygon method. A point-in-polygon interpolation model was proposed which was based on a stochastic distribution of points and the target zone, in order to represent a variety of situations. The accuracy of estimates was numerically investigated in relation to the size of the target zone and the distribution of points, and the optimal location of representative points was discussed. The major findings obtained in this paper were as follows: 1) though the relative accuracy of estimates generally increases monotonously with the size of the target zone, the monotoneity is often disturbed by the periodicity in the spatial configuration of source zones and the point distribution; 2) the point-in-polygon and the areal weighting interpolation methods have the same accuracy of estimates when points are concentrated in less than 12-15% area around the representative point in source zones; 3) the point-in-polygon method is not so robust against the locational gap between points and the representative point; 4) the optimal location of representative points is given by the spatial median of points.

1. INTRODUCTION

Socioeconomic data based on the point location are often provided in an aggregated form. Census data, for instance, are aggregated across census tracts to open to the public, though they are originally collected as point data. Other data are aggregated on a variety of zonal systems such as regular lattices, municipal districts, school districts, and postal zones. It often happens that a zonal system is incompatible with the region in which an analysis is to be performed. In such a case it is necessary to estimate the count data in the study region. This type of data estimation is called areal interpolation, or to be exact, data transfer from source zones to a target zone. Since areal interpolation is frequently required in geographical analysis, a variety of mathematical methods have been proposed in geography and other related fields (Wright 1936; Markoff and Shapiro 1973; Tobler 1979; Goodchild and Lam 1980; Lam 1983; Rhind 1991).

Areal interpolation methods can be classified into two groups: intelligent methods which use supplementary data, say, remotely sensed data, and simple methods which do not use such data. An advantage of intelligent methods is accuracy of estimates. Using suitable supplementary data, they usually provide more accurate estimates of spatial data than simple methods (Flowerdew 1988; Langford *et al*. 1991; Fisher and Langford 1995). We should note, however, that intelligent methods are not always applicable. Remotely sensed data are often unavailable or expensive, and the computational cost is problematic if a great quantity of data has to be processed. Thus simple methods such as the point-in-polygon method (Burrough and McDonnell 1998), areal weighting interpolation method (Markoff and Shapiro 1973), and pycnophylactic interpolation method (Tobler 1979), are still widely used in geography and GIS (Okabe and Sadahiro 1997; Sadahiro 1999b).

Among simple methods, the point-in-polygon method is distinguished for its processing speed (Flowerdew and Green 1991; Okabe and Sadahiro 1997). This method transfers the count data of a source zone to a target zone if the representative point of a source zone is included in the target zone (see section 2 for detailed description). The point-in-polygon method is based on the point-location algorithm (Preparata and Shamos 1985), which runs very fast (linear time). Consequently, the point-in-polygon method is suitable for massive spatial data and in fact it is used for processing 1.65 million census tracts data by the Japanese Bureau of Statistics (Sinfonica 1994). The importance of processing speed cannot be underestimated in geographical analysis, because in recent years a number of huge spatial databases have become available to geographers. On the other hand, the point-in-polygon method is said to yield less accurate estimates than other methods. The method implicitly assumes that all the points are located on representative points, and gives the true count in a target zone if the assumption holds. However, this assumption is apparently too strong. Points are usually dispersed in spatial zones, which causes inaccuracy in estimation of count data. Estimation error is relatively small and often negligible if a target zone is even larger than source zones. Otherwise, estimation error is crucial to the quality of spatial analysis and thus care should be taken in handling the estimated data.

There are two alternatives to improve estimation accuracy when supplementary data are not available: to use source data consisting of smaller zones, or to adopt another simple interpolation method, say, the areal weighting interpolation method. This choice requires us to understand the nature of estimation accuracy, that is, how accurately areal interpolation methods estimate count data. Concerning the point-in-polygon method, unfortunately, there are few systematic studies analyzing the accuracy of estimates. Though it is known that accuracy depends on various factors such as the size of source zones and the location of representative points, their effects have not been quantitatively evaluated in the literature.

This paper analyzes the accuracy of count data estimated by the point-in-polygon method in order to investigate how it is determined by various factors. We hope that the study helps geographers to choose source data and interpolation methods in their analysis. In section 2, we briefly describe the point-in-polygon method. We also outline the areal weighting interpolation method since it is discussed later in comparison with the point-in-polygon method. In section 3, we propose a stochastic model representing the point-in-polygon interpolation. Using the model, we numerically examine the accuracy of estimates in section 4 in the case where source zones are a square lattice. In section 5, the optimal location of representative points is discussed. We finally summarize the conclusions in Section 6.

2. POINT-IN-POLYGON AND AREAL WEIGHTING INTERPOLATION METHODS

Suppose a zonal system S, a region Z_0 of area A_0 consisting of K zones Z_1, Z_2, ..., Z_K (Figure 1a). The region Z_0 may represent a ward whereas zone Z_i is its lower level spatial unit, say, a census tract. We call the zones Z_1, Z_2, ..., Z_K *source zones* and denote the area of Z_i as A_i (i=1, ..., K). Spatial objects which can be regarded as points, say, individual people or households, are distributed in Z_0 (Figure 1b). We refer to these spatial objects as *points*, and denote the location of point j as $\mathbf{y}_j$. The number of points is counted in each source zone, and the data are allocated to the *representative points*. The number of points and the location of the representative point in zone Z_i are denoted by n_i and $\mathbf{p}_i$, respectively (Figure 1c). The locational data of points are then eliminated to preserve the confidentiality of the subjects.

FIG. 1. A zonal system and count data. a) Source zones, b) distributed points, c) representative points and the number of points.

We next consider a *target zone T* of area *B* in which we want to know the number of points. Since the location of individual points is not known, we have to estimate the count using an areal interpolation method. In the point-in-polygon method, the estimate of count is given by summing up the n_is whose representative points are included in the target zone *T* (Figure 2a). In the areal weighting interpolation method, the count in each source zone is divided according to area and the estimate is given by summing up the assigned values (Figure 2b).

FIG. 2. Areal interpolation methods. a) The point-in-polygon method, b) the areal weighting interpolation method. The gray-shaded area indicates the target zone. The numbers in parentheses indicate the count data n_is.

The two methods are mathematically represented as follows. Let us denote the location of *T* using a binary function defined by

$$C(\mathbf{x}) = \begin{cases} 1 & if\ \mathbf{x} \in T \\ 0 & otherwise \end{cases}. \tag{1}$$

The true number of points in *T* is then given by

$$M = \sum_j C\left(\mathbf{y}_j\right). \tag{2}$$

The number of points in zone Z_i is written as

$$n_i = \sum_j U_i\left(\mathbf{y}_j\right), \tag{3}$$

where $U_i(\mathbf{x})$ is defined by

$$U_i(\mathbf{x}) = \begin{cases} 1 & if\ \mathbf{x} \in Z_i \\ 0 & otherwise \end{cases}. \tag{4}$$

The point-in-polygon method estimates the number of points in $T \cap Z_j$ as n_i if and only if *T* contains $\mathbf{p}_i$. Otherwise it assigns zero to $T \cap Z_j$. Consequently, the estimate of *M* given by the point-in-polygon method is written as

$$\begin{aligned} \hat{M} &= \sum_i C(\mathbf{p}_i) n_i \\ &= \sum_i C(\mathbf{p}_i) \sum_j U_i\left(\mathbf{y}_j\right) \end{aligned}. \tag{5}$$

On the other hand, the areal weighting interpolation method estimates the count in *T* to be

$$\hat{M} = \sum_i \frac{\int_{\mathbf{x}\in Z_i} C(\mathbf{x})\mathrm{d}\mathbf{x}}{A_i} n_i \,. \tag{6}$$

3. POINT-IN-POLYGON INTERPOLATION MODEL

Before developing a model, we briefly discuss existing approaches to estimation accuracy of areal interpolations. The accuracy of areal interpolations has been studied in various ways, and the approaches taken in the literature can be classified into two categories: empirical-based studies and Monte Carlo simulations. The former usually considers a particular geographic situation, that is, actual zonal systems and a point distribution, in order to evaluate the reliability of estimates and compare the accuracy between interpolation methods (Flowerdew 1988; Flowerdew and Green 1991; Langford *et al*. 1991; Goodchild *et al*. 1993). Though this approach is advantageous in the reality, it is an open question whether the obtained results are generally applicable (Fisher and Langford 1995). The Monte Carlo method, on the other hand, enables us to consider a wide variety of situations and thus to obtain general results. Fisher and Langford (1995) and Cockings *et al*. (1997) adopted this approach to discuss several interpolation methods in terms of estimation accuracy, and successfully obtained full distributions of estimation errors. Monte Carlo simulations, however, are computationally expensive, because areal interpolations involve the polygon overlay operation which requires rather complicated algorithms in GIS. The computational cost is problematic especially when a number of spatial relationships between source and target zones are to be realized.

Considering the discussion above, we follow the approach taken by Sadahiro (1999b), that is, evaluation of estimation accuracy on the basis of a stochastic model representing a class of situations. This assures us not only the generality in analysis but also far less computational cost than Monte Carlo simulations.

In the model, a set of source zones with representative points are given whereas points and the target zone are stochastically located on the source zones. To be explicit, we consider a situation that N points are independently and identically distributed in the region Z_0 according to a probability density function $f(\mathbf{x})$, and the target zone T is dropped in such a way that it intersects Z_0. For convenience of computation, we assume that Z_0 has such a shape that can cover a plane by its lattice, and that Z_0 is surrounded by its copies having the same zonal system and point distribution as those of Z_0 (Figure 3). This assumption is called the *periodic continuation* which is often used in spatial statistics (Ripley 1981; Stoyan and Stoyan 1995; Sadahiro 1999b). If T does not completely lie in Z_0, we replace the portion of T outside Z_0 by its corresponding figure (Figure 4). The location of T follows the probabilistic distribution such that all possible shapes and positions of *T* appear randomly.

FIG. 3. Periodic continuation assumption.

FIG. 4. Transformation of *T*.

One might think that the periodic continuation assumption is too strong and unrealistic. The assumption, however, is not essential for the analysis since it is introduced mainly in order to avoid computational difficulties arising from boundary effects. Actually the assumption is not necessary if the region Z_0 is larger enough than the target zone *T*. There are also several ways to solve boundary problems, and they usually give almost the same results (Sadahiro 1999a).

In the above setting the accuracy of count data is discussed. The source zonal system with representative points, the number of points, and the shape and size of the target zone are given whereas the location of points and the target zone is probabilistically determined. This implies that we are considering a class of situations that share the given conditions.

From equations (2) and (5) we have the estimation error of *M*:

$$\begin{aligned}\varepsilon &= M - \hat{M} \\ &= \sum_j C(\mathbf{y}_j) - \sum_i C(\mathbf{p}_i) \sum_j U_i(\mathbf{y}_j)^{\cdot}\end{aligned} \tag{7}$$

To evaluate the error, we use the mean square error (MSE) of ε defined by

$$MSE[S] = \mathrm{E}\left[\varepsilon^2\right]. \tag{8}$$

After several steps of calculation (see Appendix 1 for details) we obtain

$$\begin{aligned}MSE[S] = N(N-1)&\int_{\mathbf{t}\in Z_0}\int_{\mathbf{x}\in Z_0} f(\mathbf{x}) f(\mathbf{t}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ &+2N\left(\frac{B}{A_0}\right) \\ &-2N(N-1)\sum_i \int_{\mathbf{x}\in Z_i} f(\mathbf{x})\mathrm{d}\mathbf{x} \int_{\mathbf{x}\in Z_0} f(\mathbf{x}) \Pr\left[\mathbf{p}_i \cup \mathbf{x} \in T\right]\mathrm{d}\mathbf{x} \\ &-2N\sum_i \int_{\mathbf{x}\in Z_i} f(\mathbf{x}) \Pr\left[\mathbf{p}_i \cup \mathbf{x} \in T\right]\mathrm{d}\mathbf{x} \\ &+N(N-1)\sum_i \sum_{i'} \Pr\left[\mathbf{p}_i \cup \mathbf{p}_{i'} \in T\right] \int_{\mathbf{x}\in Z_i} f(\mathbf{x})\mathrm{d}\mathbf{x} \int_{\mathbf{x}\in Z_{i'}} f(\mathbf{x})\mathrm{d}\mathbf{x}\end{aligned} \quad . \tag{9}$$

In the above equation the probability Pr[**x**∪**t**∈ *T*] is not explicitly given. It can be computed as follows. Let *m*(*T*; *l*) be the measure of the set of all figures congruent to *T* containing two points separated by a distance *l* (Santaló 1976). Then Pr[**x**∪**t**∈ *T*] is given by

$$\Pr[\mathbf{x} \cup \mathbf{t} \in T] = \sum_{\mathbf{u} \in \Omega(\mathbf{t})} \frac{m(T; |\mathbf{x} - \mathbf{u}|)}{2\pi A_0}, \tag{10}$$

where $\Omega(\mathbf{t})$ is a set of points corresponding to $\mathbf{t}$ in the surrounding copies of Z_0 (Figure 5). The measure $m(T;\, l)$ is represented by an explicit form if T has a simple shape (Santaló 1976; Sadahiro 1999a, 1999b). For instance, if T is a circle of radius r,

$$m(T;l) = \begin{cases} 4\pi r^2 \arccos\left(\dfrac{l}{2r}\right) - \pi l\sqrt{4r^2 - l^2} & (l \le 2r), \\ 0 & (l > 2r). \end{cases}. \tag{11}$$

For complicated shapes of T, we can use an equation

$$m(T;l) = \frac{B^2}{l} g_T(l), \tag{12}$$

where $g_T(l)$ is the probability density function of the distance between two points randomly distributed in T. Since the function $g_T(l)$ is numerically computable, equation (9) can be evaluated for any T.

FIG. 5. A point set $\Omega(\mathbf{t})$. The gray-shaded area indicates the region Z_0.

4. NUMERICAL EXAMINATIONS

Having obtained a computable representation of MSE, we are now ready to analyze numerically the accuracy of estimates using the point-in-polygon interpolation model. We wish to investigate how various factors, especially the size of the target zone and the distribution of points, affect the accuracy of estimated count data.

As seen in the previous section, the model is applicable for evaluating estimation accuracy in a variety of situations. However, due to limitations of space, we focus on a few typical cases where source zones are a square lattice. Though a more realistic zonal system such as census tracts is also available, we choose a square lattice because of the following reasons. First, congruity of source zones permits us to ignore the effect of diversity among source zones on estimation accuracy, and consequently enables us to concentrate on other specific factors. Second, computational cost is reduced if source zones are regularly arranged. This brings high tractability to the numerical examination. Third, a square lattice is a good approximation of a zonal system whose zones are similar in shape and size. Existing studies such as Okabe and Sadahiro (1997) and Sadahiro (1999b) suggest that a slight difference in shape of source zones does not greatly influence the accuracy of estimates.

As mentioned above, we focus on how the size of the target zone and the distribution of points affect the accuracy of estimated count data. To this end, we suppose a square lattice whose every cell has its representative point on the centroid of the cell and has the same form of *f*(**x**) (Figure 6). We then consider the limit where the region Z_0 expands infinitely, keeping the size of cells and the density of points at *A* and μ=*N*/A_0, respectively.

FIG. 6. Representative points and *f*(**x**) on a square lattice.

To make the MSE comparable among different sizes of *T*, we divide it by B^2 for standardization. Hence we have

$$\lim_{Z_0\to\infty}\frac{MSE[S]}{B^2}=\lim_{Z_0\to\infty}\begin{bmatrix}\frac{1}{2\pi AB^2}\int_{\mathbf{t}\in U}\int_{\mathbf{x}\in Z_0}\{A_0 f(\mathbf{x})\}\{A_0 f(\mathbf{t})\}m(T;|\mathbf{x}-\mathbf{t}|)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}\\ -\frac{1}{\pi B^2}\int_{\mathbf{x}\in Z_0}\{A_0 f(\mathbf{x})\}m(T;|\mathbf{p}-\mathbf{x}|)\mathrm{d}\mathbf{x}\\ +\frac{A^2}{2\pi AB^2}\sum_i m(T;|\mathbf{p}-\mathbf{p}_i|)\end{bmatrix}\mu^2 \qquad , \quad (13)$$

$$+\left[\frac{2}{B}-\frac{1}{\pi AB^2}\int_{\mathbf{x}\in U}\{A_0 f(\mathbf{x})\}m(T;|\mathbf{p}-\mathbf{x}|)\mathrm{d}\mathbf{x}\right]\mu$$

where *U* and **p** indicate a unit cell and a representative point, respectively.

Equation (13) indicates that the mean square error of estimates is represented by a linear combination of μ and μ^2. Hence we rewrite the equation as

$$\lim_{Z_0\to\infty}\frac{MSE[S]}{B^2}=Q_2\mu^2+Q_1\mu, \qquad (14)$$

and employ Q_1 and Q_2 instead of the MSE to evaluate the accuracy of estimates, in order to discuss the effect of μ separately from other factors. Note that both Q_1 and Q_2 are independent of μ since they are determined by *U*, *T*, **p**, and *f*(**x**).

Given a certain value of μ, the MSE is determined by Q_1 and Q_2. If points are sparsely distributed so that μ has a small value, both Q_1 and Q_2 are influential on estimation accuracy. However, if points are densely distributed in Z_0, we may neglect Q_1 in evaluating the MSE since it is mainly governed by Q_2, the leading coefficient in equation (14).

In the following we discuss the point-in-polygon method in comparison with the areal weighting interpolation method because it is also widely used in geography. The standardized MSE of this interpolation method is given by

$$
\begin{aligned}
\lim_{Z_0\to\infty}\frac{MSE[S]}{B^2} &= \lim_{Z_0\to\infty}\left[\begin{array}{l} 1+\dfrac{1}{2\pi AB^2}\displaystyle\int_{\mathbf{t}\in U}\int_{\mathbf{x}\in Z_0}\{A_0 f(\mathbf{x})\}\{A_0 f(\mathbf{t})\}m(T;|\mathbf{x}-\mathbf{t}|)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ -\dfrac{1}{\pi AB^2}\displaystyle\int_{\mathbf{t}\in U}\int_{\mathbf{x}\in Z_0}\{A_0 f(\mathbf{x})\}m(T;|\mathbf{x}-\mathbf{t}|)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \end{array}\right]\mu^2 \\
&\quad+\left[\begin{array}{l} \dfrac{1}{B}+\dfrac{1}{2\pi A^2B^2}\displaystyle\int_{\mathbf{t}\in U}\int_{\mathbf{x}\in U}m(T;|\mathbf{x}-\mathbf{t}|)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ -\dfrac{1}{\pi A^2B^2}\displaystyle\int_{\mathbf{t}\in U}\int_{\mathbf{x}\in U}\{A_0 f(\mathbf{x})\}m(T;|\mathbf{x}-\mathbf{t}|)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \end{array}\right]\mu \\
&= Q_2\mu^2+Q_1\mu
\end{aligned}
\tag{15}
$$

(for details, see Sadahiro 1999b). Similar to the point-in-polygon method, the accuracy of the areal weighting interpolation is measured by Q_1 and Q_2.

4.1 Size of the target zone

We first analyze the relationship between the size of the target zone and the accuracy of estimates. Intuitively, it is expected that the relative accuracy increases as T becomes large (Langford *et al.* 1991). We test this hypothesis on a square lattice using the interpolation models.

For the target zone T, we adopt circles whose radii range from 0.1 to 5.0. The area of square cells is set to 1. Three forms of $f(\mathbf{x})$ are investigated: a uniform distribution (Figure 7a), a quadrangular pyramid distribution (Figure 7b), and a quadrangular prism distribution (Figure 7c). The latter two are adopted as approximations of points located on the representative point, the point distribution which the point-in-polygon method implicitly assumes.

FIG. 7. Forms of $f(\mathbf{x})$ in a cell. a) a uniform distribution, b) a quadrangular pyramid distribution, c) a quadrangular prism distribution.

The results are shown in Figures 8 and 9. As we have expected, the relative accuracy of estimates improves as T becomes larger. This tendency is found in both interpolation methods for any form of $f(\mathbf{x})$. We should note, however, that Q_2 changes periodically with the size of T, whereas Q_1 decreases monotonously with as increase of T. This is mainly because of the periodicity lying in both $f(\mathbf{x})$ and the lattice. This implies that the MSE may change drastically with a slight enlargement or reduction of T if points are regularly distributed like a population distribution in new towns.

FIG. 8. Estimation accuracy of the point-in-polygon method and the size of the target zone

T.

FIG. 9. Estimation accuracy of the areal weighting interpolation method and the size of the target zone *T*. The value of Q_2 is constant at zero for the uniform distribution of *f*(**x**).

4.2 Distribution of points - the degree of concentration

We next investigate the relationship between the distribution of points and the accuracy of estimates. The point-in-polygon method implicitly assumes that points are located exactly on representative points as mentioned earlier. The areal weighting interpolation method, on the other hand, assumes that points are uniformly distributed in zones. We expect that an interpolation method yields accurate estimates if a point distribution is close to its underlying assumption. To confirm this we analyze how the concentration of points affects the accuracy of the point-in-polygon method.

For the target zone *T* we try three sets of circles whose radii range from 0.62-1.12, 1.62-2.12, and 3.62-4.12 since the MSE changes periodically with the size of *T* as seen in the previous subsection. We calculate the averages of Q_1 and Q_2 for each set of *T*. The cell size is again set to 1. The form of *f*(**x**) changes from the uniform distribution to highly concentrated distribution as shown in Figure 10. The width of a quadrangular prism representing *f*(**x**) is denoted by *w*.

FIG. 10. Form of *f*(**x**) in a cell.

The results are shown in Figures 11a, 11b, and 11c. We notice that the figures are very similar though the vertical scale is different. This indicates that the relationship between the concentration of points and estimation accuracy is independent of the size of the target zone. Concerning the point-in-polygon method, Q_1 and Q_2 decreases as points concentrate around the representative point. This supports the hypothesis mentioned earlier. In contrast to this, the areal weighting interpolation method is not fully consistent with the hypothesis: as *w* decreases, Q_2 increases but Q_1 decreases. This implies that if the point density μ is low enough (say, μ=1) estimation accuracy improves as the point distribution becomes dissimilar to the underlying distribution of the areal weighting interpolation method. Though such a low density of points is not frequently observed in practice, it should be noted that estimation error does not always decrease when a point distribution becomes close to the underlying assumption of the interpolation method.

We now consider the degree of concentration of points where the two interpolation methods give the same accuracy. Interestingly, the two methods are equivalent in estimation accuracy when *w*=0.35-0.38 for both Q_1 and Q_2 for any size of *T*. From this we can say that the point-in-polygon method is more suitable than the areal weighting interpolation method when points are concentrated in less than 12-15% area around the representative point. If points are supposed to be more dispersed, the areal weighting interpolation method would be appropriate.

FIG. 11. Estimation accuracy and the concentration of points. Radius of T is a) 0.62-1.12, b) 1.62-2.12, c) 3.62-4.12.

4.3 Distribution of points - relative location to the representative point

As assumed in this section, the representative point is often located on the centroid of a source zone, especially in digital spatial data. The location of points, however, does not usually agree with representative points: points are sometimes dispersed in a zone, clustered apart from a representative point, or even located along the edge of a zone. Such a locational gap between points and the representative point is a source of estimation error in the point-in-polygon interpolation. Therefore, we finally analyze how the location of points in cells affects estimation accuracy.

To this end, we employ three quadrangular prism distributions as $f(\mathbf{x})$ whose widths are 0.1, 0.3, and 0.5. Points move gradually from the centroid of cells to the corner as shown in Figure 12. The distance between the center of prism and the representative point located on the centroid of a cell is denoted by d. For the target zone T we try the same set of circles as those used in the previous subsection, and calculate the averages of Q_1 and Q_2. The area of cells is set to 1.

FIG. 12. Location of points in a cell.

As seen in Figure 13, the obtained results do not much differ according to the size of the target zone. In general, estimation error increases consistently with d. It is interesting that this tendency is observed not only with the point-in-polygon method but with the areal weighting interpolation method (Q_1). The effect of locational gap is more serious in the point-in-polygon method, especially when points are highly concentrated. The point-in-polygon appears not so robust against the locational gap between points and the representative point.

When points are located distantly from the representative point, it often happens that the areal weighting interpolation method gives better estimates even if points are tightly clustered. The critical value of d where the two methods are equivalent in estimation accuracy depends on w, the degree of concentration of points. Figure 13 depicts that concentration of points increases the critical value of d.

FIG. 13. Estimation accuracy and the location of points. Radius of *T* is a) 0.62-1.12, b) 1.62-2.12, c) 3.62-4.12.

5. OPTIMAL LOCATION OF REPRESENTATIVE POINTS

In the previous section we analyzed the effect of the locational gap between points and the representative point. It was found that the point-in-polygon method is sensitive to the locational gap, and that the estimation error increases with the distance. This implies that if providers of spatial data such as national mapping agencies put the representative point on an appropriate position it would improve estimation accuracy of the point-in-polygon interpolation performed by users of the data. So where should the representative point be located in a zone? This section briefly discusses this subject.

We consider the optimal location of the representative point individually for each zone, and thus we focus on the representative point of zone Z_k. From the viewpoint of areal interpolation, it is desirable that the representative point is located so as to minimize estimation error originating in Z_k. This problem is formulated as follows.

Let M_k be the number of points in the intersection of Z_k and T. It is mathematically represented as

$$M_k = \sum_j C(\mathbf{y}_j) U_k(\mathbf{y}_j). \tag{16}$$

The estimate of M_k given by the point-in-polygon interpolation is

$$\hat{M}_k = C(\mathbf{p}_k) \sum_j U_k(\mathbf{y}_j). \tag{17}$$

The mean square error of M_k is calculated in a similar way as that of M (for details, see Appendix 2):

$$\begin{aligned} MSE_k[S] = {} & N(N-1)\int_{\mathbf{t}\in Z_k}\int_{\mathbf{x}\in Z_k} f(\mathbf{x}) f(\mathbf{t}) \Pr[\mathbf{x}\cup\mathbf{t}\in T]\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ & +2N\left(\frac{B}{A_0}\right)\int_{\mathbf{x}\in Z_k} f(\mathbf{x})\mathrm{d}\mathbf{x} \\ & -2N(N-1)\int_{\mathbf{x}\in Z_k} f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in Z_k} f(\mathbf{x})\Pr[\mathbf{p}_k\cup\mathbf{x}\in T]\mathrm{d}\mathbf{x} \\ & -2N\int_{\mathbf{x}\in Z_k} f(\mathbf{x})\Pr[\mathbf{p}_k\cup\mathbf{x}\in T]\mathrm{d}\mathbf{x} \\ & +N(N-1)\left(\frac{B}{A_0}\right)\left\{\int_{\mathbf{x}\in Z_i} f(\mathbf{x})\mathrm{d}\mathbf{x}\right\}^2 \end{aligned}. \tag{18}$$

The optimal location of $\mathbf{p}_k$ is obtained by solving

$$\min_{\mathbf{p}_k} MSE_k[S]. \tag{19}$$

Substitution of equation (18) yields

$$\min_{\mathbf{p}_k} MSE_k[S] \Leftrightarrow \max_{\mathbf{p}_k} \int_{\mathbf{x}\in Z_k} f(\mathbf{x})\Pr[\mathbf{p}_k \cup \mathbf{x} \in T]\mathrm{d}\mathbf{x}. \tag{20}$$

Assuming that the region Z_0 is even larger than T, we obtain

$$\max_{\mathbf{p}_k} \int_{\mathbf{x}\in Z_k} f(\mathbf{x}) m\left(T; |\mathbf{p}_k - \mathbf{x}|\right)\mathrm{d}\mathbf{x}. \tag{21}$$

The above problem is a kind of continuous location problems studied in operations research (Plastria 1995). If $f(\mathbf{x})$ and $m(T; l)$ are given in analytical and simple forms, the problem may be solvable by a computational procedure (Love 1972; Drezner and Wesolowsky 1980; Aly and Marucheck 1982; Love *et al.*, 1988). Otherwise we have to try numerous locations in Z_k to find the optimal location of $\mathbf{p}_k$.

Equation (21) implies that the optimal location of $\mathbf{p}_k$ depends not only on $f(\mathbf{x})$ and Z_k but also on the target zone T. In practice, however, the target zone is unknown when the location of $\mathbf{p}_k$ is determined. We hence consider the circle which is the most typical shape of the target zone.

Let T be a circle of radius r whose diameter $2r$ is larger than the maximum length of Z_k. The measure $m(T; l)$ is then given by

$$m(T;l) = 4\pi r^2 \arccos\left(\frac{l}{2r}\right) - \pi l\sqrt{4r^2 - l^2}\,. \tag{22}$$

Applying a Taylor series expansion around l=0 to equation (22), we obtain

$$\begin{aligned} m(T;l) &\approx m(T;0) + \frac{\mathrm{d}}{\mathrm{d}l}m(T;0)l + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}l^2}m(T;0)l^2 \\ &= 2\pi^2 r^2 - 4\pi r l \end{aligned}. \tag{23}$$

The above equation is a good approximation of $m(T; l)$ especially for small l (Figure 14). Substitution of equation (23) into equation (21) yields

$$\min_{\mathbf{p}_k} \int_{\mathbf{x}\in Z_k} f(\mathbf{x})|\mathbf{p}_k - \mathbf{x}|\mathrm{d}\mathbf{x}. \tag{24}$$

This indicates that the location of $\mathbf{p}_k$ is the spatial median of $f(\mathbf{x})$ in Z_k (Brown 1983; Small 1990; Stoyan and Stoyan 1994). As we will illustrate later, the spatial median does not always agree with the centroid of $f(\mathbf{x})$ on which the representative point is often located.

FIG. 14. The measure $m(T; l)$ of a circle of radius 3.

The above discussion holds for the shapes that allow the linear approximation shown in equation (23). Convex shapes similar to the circle, say, the square, regular triangle, regular hexagon, and rectangles meet this requirement (see Figures 15 and 16). Considering that the target zone usually has a simple convex shape, we may say that the spatial median is the optimal location of representative points in relation to estimation accuracy of the point-in-polygon interpolation.

FIG. 15. The measure $m(T; l)$ of a regular triangle and regular hexagon.

FIG. 16. The measure $m(T; l)$ of rectangles having a variety of horizontal to vertical ratios.

We finally show some examples of the optimal location of representative points. We compute both the spatial median and the centroid of $f(\mathbf{x})$ in square zones varying the distribution of points $f(\mathbf{x})$. Every zone is divided into 10 * 10 cells each of which has a constant value of $f(\mathbf{x})$ as shown in Figure 17.

As mentioned earlier, the spatial median does not always agree with the centroid. The spatial median is more strongly drawn by points, and it tends to be nearer the maximum of $f(\mathbf{x})$. Since the locational gap between points and the representative point is crucial in the point-in-polygon interpolation, estimation accuracy appears to be fairly improved if data providers locate the representative point on the spatial median rather than the centroid of $f(\mathbf{x})$.

FIG. 17. The optimal location of the representative point (spatial median) in a square zone. The white squares and black circles represent the spatial median and the centroid of $f(\mathbf{x})$, respectively. The value of $f(\mathbf{x})$ is zero in the white cells. The lightest gray cells indicate $f(\mathbf{x})$=1.0, and the values of $f(\mathbf{x})$ in the other gray cells are shown individually.

6. CONCLUSION

In this paper we have analyzed the accuracy of count data estimated by the point-in-polygon method. We first proposed a point-in-polygon interpolation model which is based on a stochastic distribution of points and the target zone. This allows us to replace diverse situations by their representative model. We then applied the model to the case of a square lattice and numerically investigated the accuracy of estimates. The major results are summarized as follows:

1) The relative accuracy of estimates generally increases with the size of the target zone. However, the periodicity in the spatial configuration of source zones and $f(\mathbf{x})$ disturbs monotoneity of the relationship, thus an enlargement of the target zone may reduce estimation accuracy.
2) The point-in-polygon method gives good estimates when points are concentrated around the representative point. The areal weighting interpolation method, on the other hand, fits uniform distributions of points. These methods are equivalent in terms of estimation accuracy when points are concentrated in less than 12-15%

area around the representative point.

3) The point-in-polygon method is not so robust against the locational gap between points and the representative point. Hence it often yields worse estimates than the areal weighting interpolation method even if points are strongly concentrated.

The third result implies that the location of representative points is crucial for the point-in-polygon interpolation. Thus we discussed the optimal location of representative points and found that in most cases it is given by the spatial median of points. This warns us as data producers that the centroid of points often used as the representative point is not the optimal location in relation to the point-in-polygon interpolation.

In the empirical study we considered only the square lattice due to limitations of space. We should emphasize, however, that the proposed method is applicable to any zonal system as mentioned earlier. Zonal systems which are less artificial than regular lattices, say, census tracts and municipal districts, should be analyzed in future researches.

The point-in-polygon interpolation model we proposed considers data transfer from multiple source zones to a single target zone. Data transfer between zonal systems still remains to be analyzed where estimation occurs in multiple target zones. This is an important subject in geography which is related to the modifiable areal unit problem (Openshaw 1984; Flowerdew and Green 1991; Fisher and Langford 1995; Okabe and Sadahiro 1997). We will examine an extension of our study to this case in order to analyze the accuracy of estimation between zonal systems.

LITERATURE CITED

Aly, A. A. and A. S. Marucheck (1982). "Generalized Weber Problem with Rectangular Regions." *Journal of the Operational Research Society* **33**, 983-989.

Brown, B. M. (1983). "Statistical uses of the Spatial Median." *Journal of the Royal Statistical Society, Series B* **45**, 25-30.

Burrough, P. A. and R. A. McDonnell (1998). *Principles of Geographical Information Systems*. New York: Oxford University Press.

Cockings, S., P. F. Fisher, and M. Langford (1997). "Parameterization and Visualization of the Errors in Areal Interpolation." *Geographical Analysis* **29**, 314-328.

Drezner, Z. and G. O. Wesolowsky (1980). "Optimal Location of a Demand Facility Relative to Area Demand." *Naval Research Logistics Quarterly* **27**, 199-206.

Fisher, P. F. and M. Langford (1995). "Modelling the Errors in Areal Interpolation between Zonal Systems by Monte Carlo Simulation." *Environment and Planning A* **27**, 211-224.

Flowerdew, R. (1988). "Statistical Methods for Areal Interpolation: Predicting Count Data from a Binary Variable." *Research Report* **15**, Northern Regional Research Laboratory.

Flowerdew, R. and M. Green (1991). "Data Integration: Statistical Methods for Transferring Data between Zonal Systems." In *Handling Geographical Information: Methodology and Potential Applications*, edited by I. Masser and M. Blakemore, pp.38-54, New York: Longman.

Goodchild, M. F. and N. N-S. Lam (1980). "Areal Interpolation: a Variant of the Traditional Spatial Problem." *Geo-processing* **1**, 297-312.

Goodchild, M. F., L. Anselin and U. Deichmann (1993). "A Framework for the Areal Interpolation of Socioeconomic Data." *Environment and Planning A* **25**, 383-397.

Lam, N. N-S. (1983). "Spatial Interpolation Methods: a Review." *American Cartographer* **10**, 129-149.

Langford, M., D. J. Maguire, and D. J. Unwin (1991). "The Areal Interpolation Problem: Estimating Population Using Remote Sensing in a GIS Framework." In *Handling Geographical Information: Methodology and Potential Applications*, edited by I. Masser and M. Blakemore, pp.55-77, New York: Longman.

Love, R. F. (1972). "A Computational Procedure for Optimally Locating a Facility with Respect to Several Rectangular Regions." *Journal of Regional Science* **12**, 233-242.

Love, R. F., J. G. Morris and G. O. Wesolowsky (1988). *Facilities Location: Models & Methods*. Amsterdam: North-Holland.

Markoff, J. and G. Shapiro (1973). "The Linkage of Data Describing Overlapping

Geographical Units." *Historical Methods Newsletter* **7**, 34-46.

Okabe, A. and Y. Sadahiro (1997). "Variation in Count Data Transferred from a Set of Irregular Zones to a Set of Regular Zones through the Point-in-polygon Method." *International Journal of Geographical Information Science* **11**, 93-106.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Norwich: Geobooks.

Plastria, F. (1995). "Continuous Location Problems." In *Facility Location: A Survey of Applications and Methods*, edited by Z. Drezner, pp.225-262, New York: Springer-Verlag.

Preparata, F. P. and M. I. Shamos (1985). *Computational Geometry - An Introduction* -. New York: Springer-Verlag.

Rhind, D. W. (1991). "Counting the People: the Role of GIS." In *Geographical Information Systems, Volume 2: Principles and Applications*, edited by D. J. Maguire, M. F. Goodchild, and D. W. Rhind, pp.127-137, New York: Longman.

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

Sadahiro, Y. (1999a). "Statistical Methods for Analyzing the Distribution of Spatial Objects in Relation to a Surface." *Geographical Systems*, to appear.

Sadahiro, Y. (1999b). "Accuracy of Count Data Transferred through the Areal Weighting Interpolation Method." *International Journal of Geographical Information Science*, to appear.

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

Sinfonica (1994). *Study on Mesh Data*. Report.

Small, C. G. (1990). "A Survey of Multidimensional Medians." *International Statistical Review* **58**, 263-277.

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

Tobler, W. R. (1979). "Smooth Pycnophylactic Interpolation for Geographical Regions." *Journal of the American Statistical Association* **74**, 519-530.

Wright, J. K. (1936). "A Method of Mapping Densities of Population with Cape Cod as an Example." *Geographical Review* **26**, 103-110.

APPENDIX 1

The mean square error of ε which is denoted by *MSE*[*S*] is calculated as follows.

$$\begin{aligned} MSE[S] &= \mathrm{E}\left[\varepsilon^2\right] \\ &= \mathrm{E}\left[M^2\right] - 2\mathrm{E}\left[M\hat{M}\right] + \mathrm{E}\left[\hat{M}^2\right]. \end{aligned} \qquad \text{(A 1)}$$

The first term of equation (A 2) becomes

$$\begin{aligned} \mathrm{E}\left[M^2\right] &= \mathrm{E}\left[\left\{\sum_j C\left(\mathbf{y}_j\right)\right\}^2\right] \\ &= \mathrm{E}\left[\sum_j \sum_{j'} C\left(\mathbf{y}_j\right) C\left(\mathbf{y}_{j'}\right)\right] \\ &= 2\sum_{j \neq j'} \mathrm{E}\left[C\left(\mathbf{y}_j\right) C\left(\mathbf{y}_{j'}\right)\right] + \sum_{j'} \mathrm{E}\left[C\left(\mathbf{y}_j\right)\right] \end{aligned} \quad . \qquad \text{(A 2)}$$

Substituting

$$\begin{aligned} \mathrm{E}\left[C\left(\mathbf{y}_j\right) C\left(\mathbf{y}_{j'}\right)\right] &= \Pr\left[\left(\mathbf{y}_j \in T\right) \cap \left(\mathbf{y}_{j'} \in T\right)\right] \\ &= \int_{\mathbf{t} \in Z_0} \int_{\mathbf{x} \in Z_0} \Pr\left[\left(\mathbf{y}_j \in \mathrm{d}\mathbf{x}\right) \cap (\mathrm{d}\mathbf{x} \in T) \cap \left(\mathbf{y}_{j'} \in \mathrm{d}\mathbf{t}\right) \cap (\mathrm{d}\mathbf{t} \in T)\right] \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ &= \int_{\mathbf{t} \in Z_0} \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) f(\mathbf{t}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \end{aligned}$$

(A 3)

and

$$\begin{aligned} \mathrm{E}\left[C\left(\mathbf{y}_j\right)\right] &= \Pr\left[\mathbf{y}_j \in T\right] \\ &= \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) \Pr[\mathbf{x} \in T] \mathrm{d}\mathbf{x}, \\ &= \frac{B}{A_0} \end{aligned} \qquad \text{(A 4)}$$

we have

$$\mathrm{E}\left[M^2\right] = N(N-1)\int_{\mathbf{t} \in Z_0} \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) f(\mathbf{t}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} + N\left(\frac{B}{A_0}\right). \qquad \text{(A 5)}$$

The second term of equation (A 1) is rewritten as

$$\begin{aligned} \mathrm{E}\left[M\hat{M}\right] &= E\left[\left\{\sum_j C\left(\mathbf{y}_j\right)\right\}\left\{\sum_i C(\mathbf{p}_i) \sum_j U_i\left(\mathbf{y}_j\right)\right\}\right] \\ &= E\left[\sum_i \sum_j \sum_{j'} C(\mathbf{p}_i) U_i\left(\mathbf{y}_j\right) C\left(\mathbf{y}_{j'}\right)\right] \\ &= 2\sum_i \sum_{j \neq j'} E\left[U_i\left(\mathbf{y}_j\right)\right] E\left[C(\mathbf{p}_i) C\left(\mathbf{y}_{j'}\right)\right] + \sum_i \sum_j E\left[U_i\left(\mathbf{y}_j\right) C(\mathbf{p}_i) C\left(\mathbf{y}_j\right)\right] \end{aligned} \quad . \qquad \text{(A 6)}$$

Substitution of

$$
\begin{aligned}
E\left[U_i\left(\mathbf{y}_j\right)\right]E\left[C(\mathbf{p}_i)C\left(\mathbf{y}_{j'}\right)\right] &= \int_{\mathbf{x}\in Z_i} f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in Z_0} \Pr\left[\left(\mathbf{p}_i \in T\right)\cap\left(\mathbf{y}_{j'} \in \mathrm{d}\mathbf{x}\right)\cap(\mathrm{d}\mathbf{x}\in T)\right]\mathrm{d}\mathbf{x} \\
&= \int_{\mathbf{x}\in Z_i} f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in Z_0} f(\mathbf{x})\Pr\left[\mathbf{p}_i \cup \mathbf{x}\in T\right]\mathrm{d}\mathbf{x}
\end{aligned}
\tag{A 7}
$$

and

$$
\begin{aligned}
E\left[U_i\left(\mathbf{y}_j\right)C(\mathbf{p}_i)C\left(\mathbf{y}_j\right)\right] &= \int_{\mathbf{x}\in Z_i} \Pr\left[\left(\mathbf{y}_j \in \mathrm{d}\mathbf{x}\right)\cap\left(\mathbf{p}_i \in T\right)\cap(\mathrm{d}\mathbf{x}\in T)\right]\mathrm{d}\mathbf{x} \\
&= \int_{\mathbf{x}\in Z_i} \Pr\left[\mathbf{y}_j \in \mathrm{d}\mathbf{x}\right]\Pr\left[\mathbf{p}_i \cup \mathrm{d}\mathbf{x}\in T\right]\mathrm{d}\mathbf{x} \\
&= \int_{\mathbf{x}\in Z_i} f(\mathbf{x})\Pr\left[\mathbf{p}_i \cup \mathbf{x}\in T\right]\mathrm{d}\mathbf{x}
\end{aligned}
\tag{A 8}
$$

yields

$$
\begin{aligned}
\mathrm{E}\left[M\hat{M}\right] &= N(N-1)\sum_i \int_{\mathbf{x}\in Z_i} f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in Z_0} f(\mathbf{x})\Pr\left[\mathbf{p}_i \cup \mathbf{x}\in T\right]\mathrm{d}\mathbf{x} \\
&\quad + N\sum_i \int_{\mathbf{x}\in Z_i} f(\mathbf{x})\Pr\left[\mathbf{p}_i \cup \mathbf{x}\in T\right]\mathrm{d}\mathbf{x}
\end{aligned}
\quad . \tag{A 9}
$$

The third term of equation (A 1) is given by

$$
\begin{aligned}
\mathrm{E}\left[\hat{M}^2\right] &= E\left[\left\{\sum_i C(\mathbf{p}_i)\sum_j U_i\left(\mathbf{y}_j\right)\right\}^2\right] \\
&= E\left[\sum_i\sum_{i'} C(\mathbf{p}_i)C(\mathbf{p}_{i'})\sum_j U_i\left(\mathbf{y}_j\right)\sum_{j'} U_{i'}\left(\mathbf{y}_{j'}\right)\right] \\
&= \sum_i\sum_{i'} E\left[C(\mathbf{p}_i)C(\mathbf{p}_{i'})\right]\sum_j\sum_{j'} E\left[U_i\left(\mathbf{y}_j\right)U_{i'}\left(\mathbf{y}_{j'}\right)\right]
\end{aligned}
\quad . \tag{A 10}
$$

This equation becomes

$$
\begin{aligned}
\mathrm{E}\left[\hat{M}^2\right] &= 2\sum_i\sum_{i'} E\left[C(\mathbf{p}_i)C(\mathbf{p}_{i'})\right]\sum_{j\neq j'} E\left[U_i\left(\mathbf{y}_j\right)\right]E\left[U_{i'}\left(\mathbf{y}_{j'}\right)\right] \\
&\quad + \sum_i\sum_{i'} E\left[C(\mathbf{p}_i)C(\mathbf{p}_{i'})\right]\sum_j E\left[U_i\left(\mathbf{y}_j\right)U_{i'}\left(\mathbf{y}_j\right)\right] \\
&= N(N-1)\sum_i\sum_{i'} \Pr\left[\mathbf{p}_i \cup \mathbf{p}_{i'} \in T\right]\int_{\mathbf{x}\in Z_i} f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in Z_{i'}} f(\mathbf{x})\mathrm{d}\mathbf{x} \\
&\quad + \sum_i E\left[C(\mathbf{p}_i)\right]\sum_j E\left[U_i\left(\mathbf{y}_j\right)\right] \\
&= N(N-1)\sum_i\sum_{i'} \Pr\left[\mathbf{p}_i \cup \mathbf{p}_{i'} \in T\right]\int_{\mathbf{x}\in Z_i} f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in Z_{i'}} f(\mathbf{x})\mathrm{d}\mathbf{x} \\
&\quad + N\left(\frac{B}{A_0}\right)
\end{aligned}
\quad . \tag{A 11}
$$

Using equations (A 5), (A 9), and (A 11), we obtain

$$\begin{aligned} MSE[S] = N(N-1)&\int_{\mathbf{t}\in Z_0}\int_{\mathbf{x}\in Z_0} f(\mathbf{x})f(\mathbf{t})\Pr[\mathbf{x}\cup\mathbf{t}\in T]\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ &+2N\left(\frac{B}{A_0}\right) \\ &-2N(N-1)\sum_i\int_{\mathbf{x}\in Z_i} f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in Z_0} f(\mathbf{x})\Pr\left[\mathbf{p}_i\cup\mathbf{x}\in T\right]\mathrm{d}\mathbf{x} \\ &-2N\sum_i\int_{\mathbf{x}\in Z_i} f(\mathbf{x})\Pr\left[\mathbf{p}_i\cup\mathbf{x}\in T\right]\mathrm{d}\mathbf{x} \\ &+N(N-1)\sum_i\sum_{i'}\Pr\left[\mathbf{p}_i\cup\mathbf{p}_{i'}\in T\right]\int_{\mathbf{x}\in Z_i} f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in Z_{i'}} f(\mathbf{x})\mathrm{d}\mathbf{x} \end{aligned} \quad . \qquad \text{(A 12)}$$

APPENDIX 2

The mean square error of M_k is calculated as follows.

$$MSE_k[S] = E\left[M_k^2\right] - 2E\left[M_k\hat{M}_k\right] + E\left[\hat{M}_k^2\right] \qquad \text{(A 13)}$$

The first term of equation (A 13) becomes

$$\begin{aligned} E\left[M_k^2\right] &= E\left[\left\{\sum_j C(\mathbf{y}_j)U_k(\mathbf{y}_j)\right\}^2\right] \\ &= 2\sum_{j\neq j'} E\left[C(\mathbf{y}_j)U_k(\mathbf{y}_j)C(\mathbf{y}_{j'})U_k(\mathbf{y}_{j'})\right] + \sum_j E\left[C(\mathbf{y}_j)U_k(\mathbf{y}_j)\right] \end{aligned} \quad . \qquad \text{(A 14)}$$

Using

$$\begin{aligned} &E\left[C(\mathbf{y}_j)U_k(\mathbf{y}_j)C(\mathbf{y}_{j'})U_k(\mathbf{y}_{j'})\right] \\ &= \int_{\mathbf{t}\in Z_k}\int_{\mathbf{x}\in Z_k}\Pr\left[\left(\mathbf{y}_j\in\mathrm{d}\mathbf{x}\right)\cap(\mathrm{d}\mathbf{x}\in T)\cap\left(\mathbf{y}_{j'}\in\mathrm{d}\mathbf{t}\right)\cap(\mathrm{d}\mathbf{t}\in T)\right]\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ &= \int_{\mathbf{t}\in Z_k}\int_{\mathbf{x}\in Z_k} f(\mathbf{x})f(\mathbf{t})\Pr[\mathbf{x}\cup\mathbf{t}\in T]\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \end{aligned} \qquad \text{(A 15)}$$

and

$$\begin{aligned} E\left[C(\mathbf{y}_j)U_k(\mathbf{y}_j)\right] &= \int_{\mathbf{x}\in Z_k}\Pr\left[\left(\mathbf{y}_j\in\mathrm{d}\mathbf{x}\right)\cap(\mathrm{d}\mathbf{x}\in T)\right]\mathrm{d}\mathbf{x} \\ &= \int_{\mathbf{x}\in Z_k} f(\mathbf{x})\Pr[\mathbf{x}\in T]\mathrm{d}\mathbf{x} \\ &= \frac{B}{A_0}\int_{\mathbf{x}\in Z_k} f(\mathbf{x})\mathrm{d}\mathbf{x} \end{aligned} \quad , \qquad \text{(A 16)}$$

we have

$$\begin{aligned} E\left[M_k^2\right] = N(N-1)&\int_{\mathbf{t}\in Z_k}\int_{\mathbf{x}\in Z_k} f(\mathbf{x})f(\mathbf{t})\Pr[\mathbf{x}\cup\mathbf{t}\in T]\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ &+N\left(\frac{B}{A_0}\right)\int_{\mathbf{x}\in Z_k} f(\mathbf{x})\mathrm{d}\mathbf{x} \end{aligned} \quad . \qquad \text{(A 17)}$$

The second term of equation (A 13) is rewritten as

$$\begin{aligned} E\left[M_k\hat{M}_k\right] &= E\left[C(\mathbf{p}_k)\sum_j\sum_{j'}C(\mathbf{y}_j)U_k(\mathbf{y}_j)U_k(\mathbf{y}_{j'})\right] \\ &= 2E\left[C(\mathbf{p}_k)\sum_{j\neq j'}C(\mathbf{y}_j)U_k(\mathbf{y}_j)U_k(\mathbf{y}_{j'})\right]+E\left[C(\mathbf{p}_k)\sum_j C(\mathbf{y}_j)U_k(\mathbf{y}_j)\right] \quad . \\ &= 2\sum_{j\neq j'}E\left[C(\mathbf{p}_k)C(\mathbf{y}_j)U_k(\mathbf{y}_j)\right]E\left[U_k(\mathbf{y}_{j'})\right]+\sum_j E\left[C(\mathbf{p}_k)C(\mathbf{y}_j)U_k(\mathbf{y}_j)\right] \end{aligned} \tag{A 18}$$

Substituting equation (A 8), we obtain

$$\begin{aligned} E\left[M_k\hat{M}_k\right] &= N(N-1)\int_{\mathbf{x}\in Z_k}f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in Z_k}f(\mathbf{x})\Pr\left[\mathbf{p}_k\cup\mathbf{x}\in T\right]\mathrm{d}\mathbf{x} \\ &\quad +N\int_{\mathbf{x}\in Z_k}f(\mathbf{x})\Pr\left[\mathbf{p}_k\cup\mathbf{x}\in T\right]\mathrm{d}\mathbf{x} \end{aligned} \quad . \tag{A 19}$$

The third term of equation (A 13) becomes

$$\begin{aligned} E\left[\hat{M}_k^2\right] &= E\left[\left\{C(\mathbf{p}_k)\sum_j U_k(\mathbf{y}_j)\right\}^2\right] \\ &= 2E\left[C(\mathbf{p}_k)\sum_{j\neq j'}U_k(\mathbf{y}_j)U_k(\mathbf{y}_{j'})\right]+E\left[C(\mathbf{p}_k)\sum_j U_k(\mathbf{y}_j)\right] \\ &= 2E\left[C(\mathbf{p}_k)\right]\sum_{j\neq j'}E\left[U_k(\mathbf{y}_j)\right]E\left[U_k(\mathbf{y}_{j'})\right]+E\left[C(\mathbf{p}_k)\right]\sum_j E\left[U_k(\mathbf{y}_j)\right] \\ &= N(N-1)\Pr\left[\mathbf{p}_i\in T\right]\left\{\int_{\mathbf{x}\in Z_i}f(\mathbf{x})\mathrm{d}\mathbf{x}\right\}^2+N\Pr\left[\mathbf{p}_i\in T\right]\int_{\mathbf{x}\in Z_i}f(\mathbf{x})\mathrm{d}\mathbf{x} \quad . \\ &= N(N-1)\left(\frac{B}{A_0}\right)\left\{\int_{\mathbf{x}\in Z_i}f(\mathbf{x})\mathrm{d}\mathbf{x}\right\}^2+N\left(\frac{B}{A_0}\right)\int_{\mathbf{x}\in Z_i}f(\mathbf{x})\mathrm{d}\mathbf{x} \end{aligned} \tag{A 20}$$

Using equations (A 17), (A 19), and (A 20) we obtain

$$\begin{aligned} MSE_k[S] &= N(N-1)\int_{\mathbf{t}\in Z_k}\int_{\mathbf{x}\in Z_k}f(\mathbf{x})f(\mathbf{t})\Pr[\mathbf{x}\cup\mathbf{t}\in T]\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ &\quad +2N\left(\frac{B}{A_0}\right)\int_{\mathbf{x}\in Z_k}f(\mathbf{x})\mathrm{d}\mathbf{x} \\ &\quad -2N(N-1)\int_{\mathbf{x}\in Z_k}f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in Z_k}f(\mathbf{x})\Pr\left[\mathbf{p}_k\cup\mathbf{x}\in T\right]\mathrm{d}\mathbf{x} \\ &\quad -2N\int_{\mathbf{x}\in Z_k}f(\mathbf{x})\Pr\left[\mathbf{p}_k\cup\mathbf{x}\in T\right]\mathrm{d}\mathbf{x} \\ &\quad +N(N-1)\left(\frac{B}{A_0}\right)\left\{\int_{\mathbf{x}\in Z_i}f(\mathbf{x})\mathrm{d}\mathbf{x}\right\}^2 \end{aligned} \quad . \tag{A 21}$$

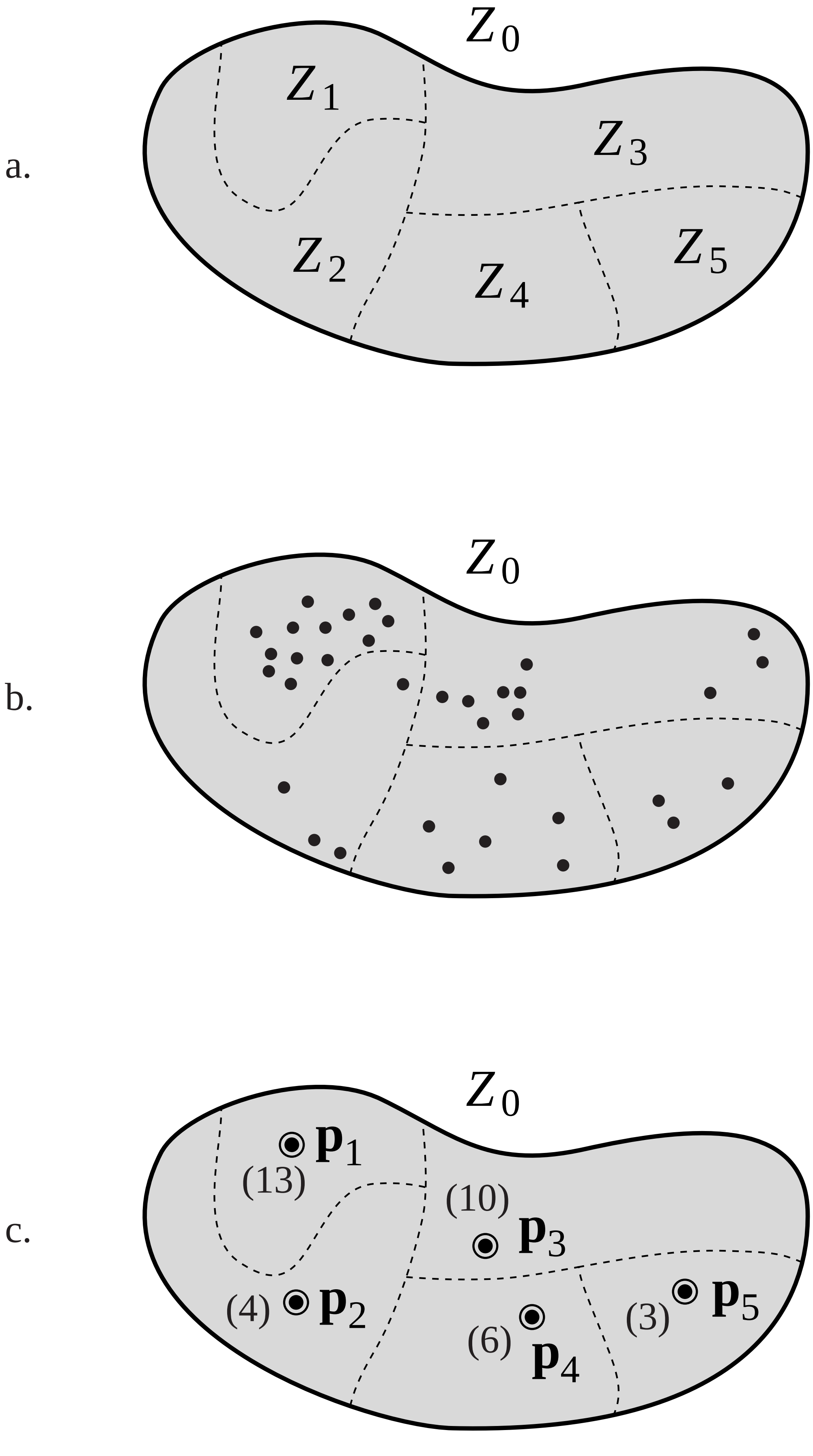
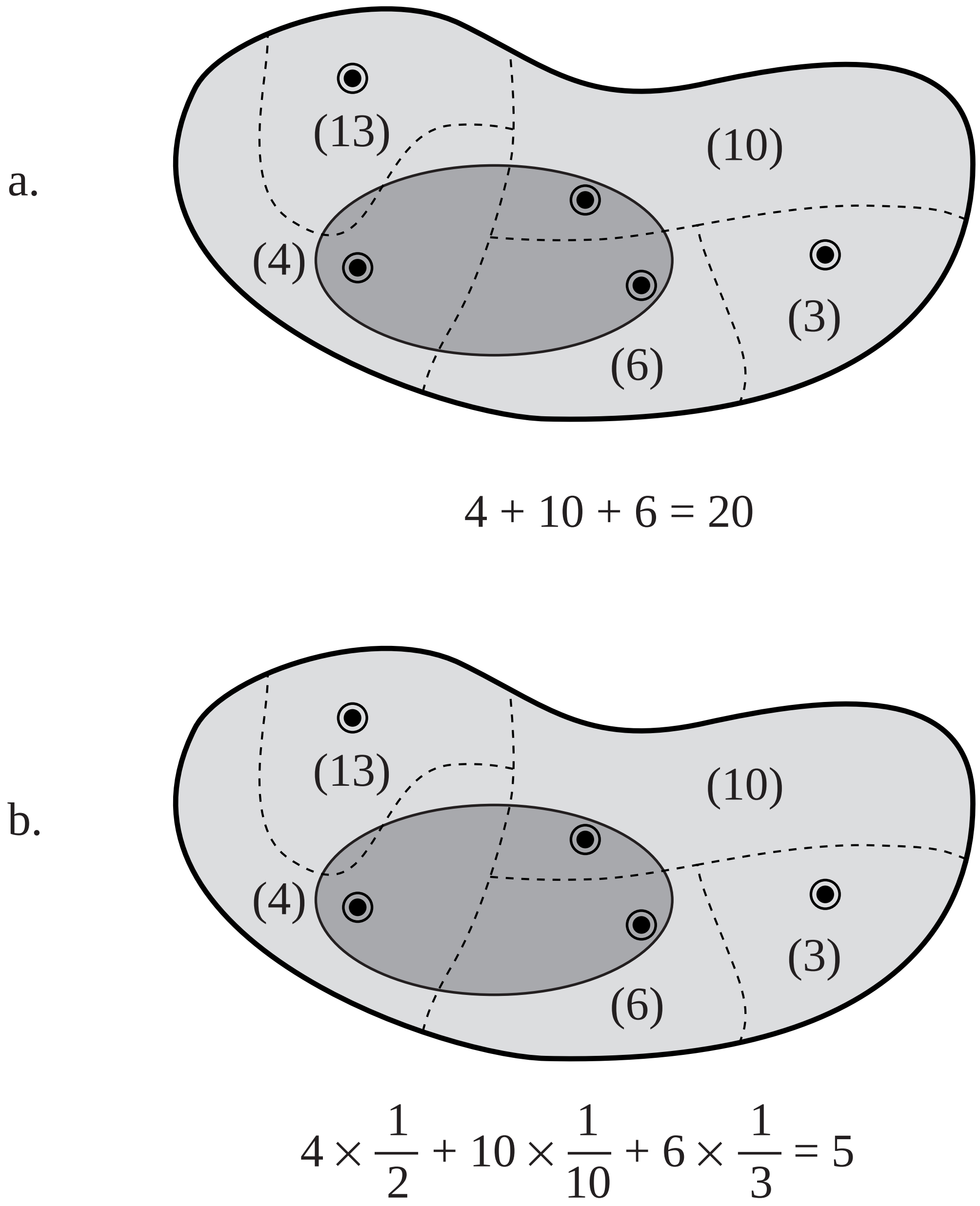

a.
Z_0
Z_1
Z_3
Z_2
Z_4
Z_5
b.
Z_0
c.
Z_0
$\mathbf{p}_1$
(13)
(10)
$\mathbf{p}_3$
(4)
$\mathbf{p}_2$
(6)
$\mathbf{p}_4$
(3)
$\mathbf{p}_5$

Figure 1



$$4 + 10 + 6 = 20$$

$$4 \times \frac{1}{2} + 10 \times \frac{1}{10} + 6 \times \frac{1}{3} = 5$$

Figure 2

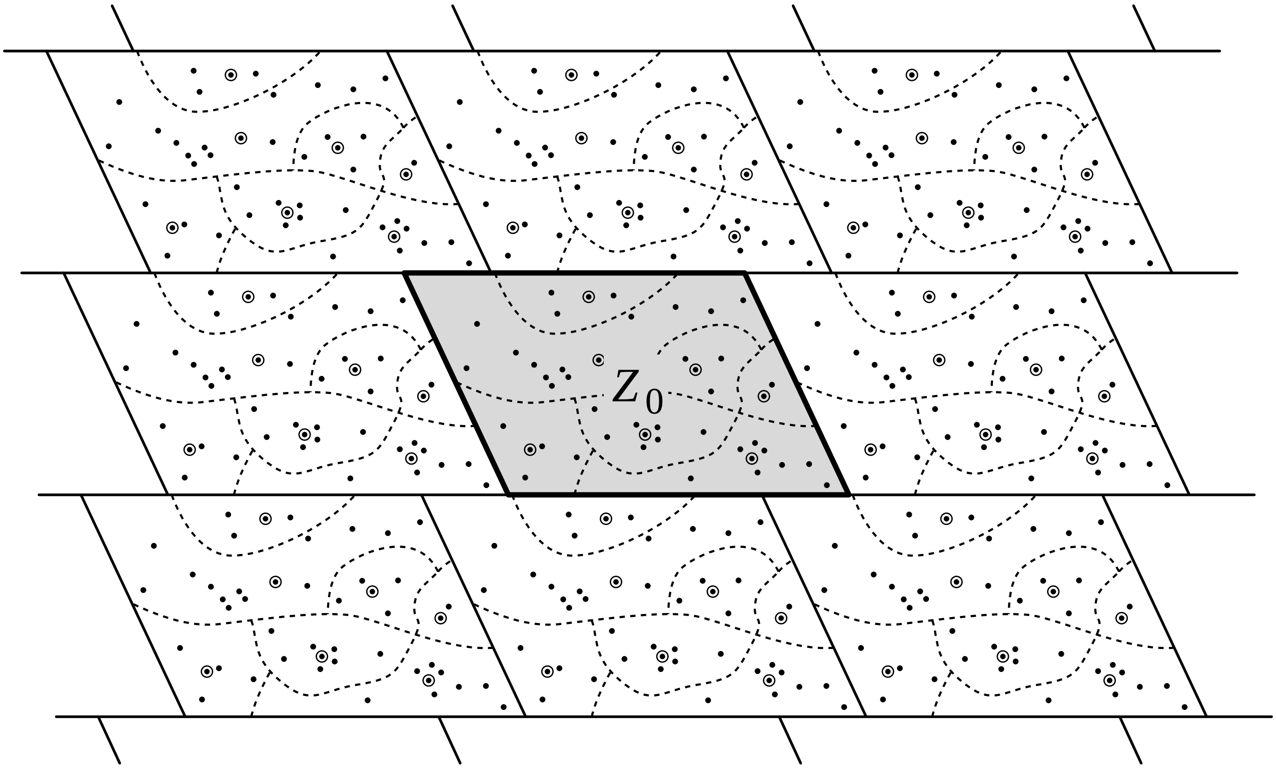


Figure 3

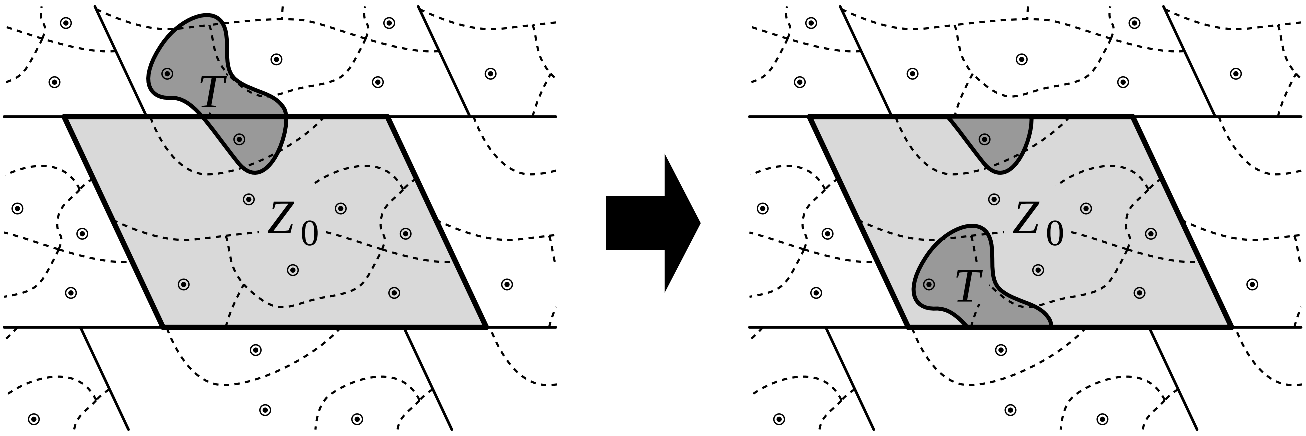


Figure 4

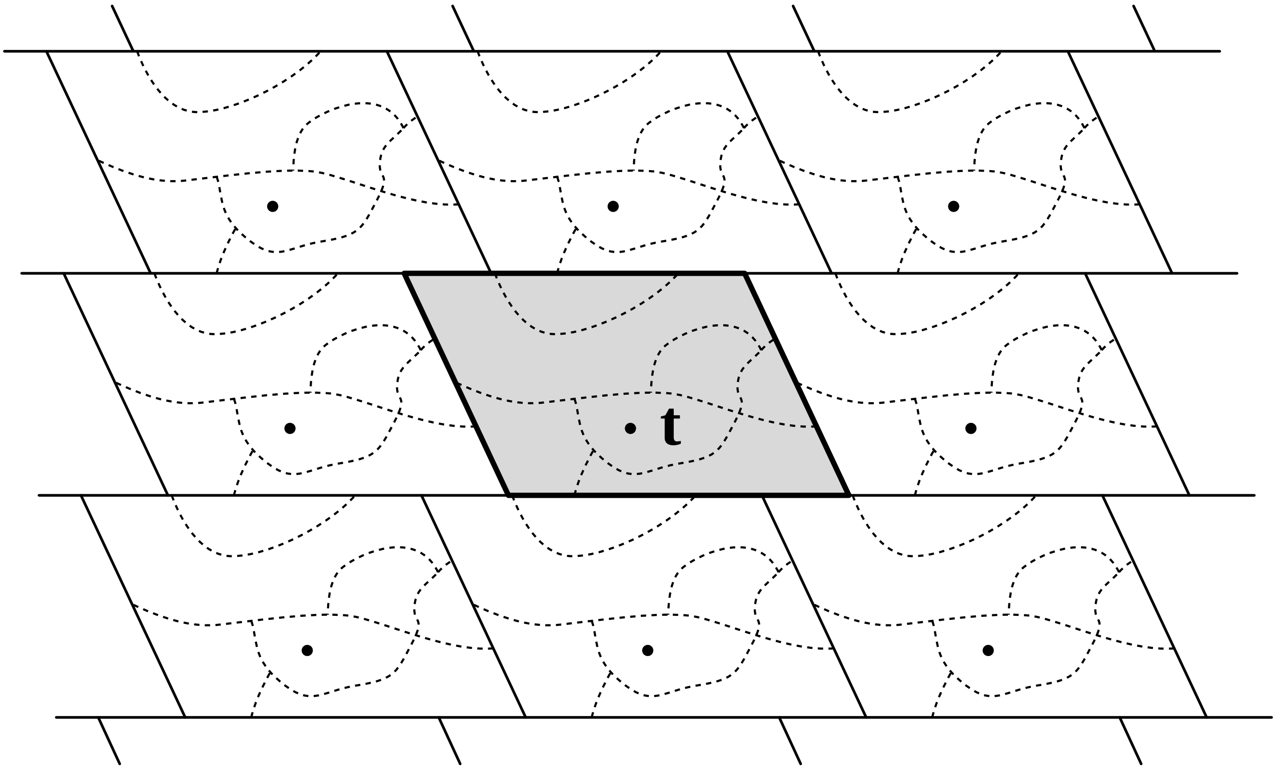


Figure 5

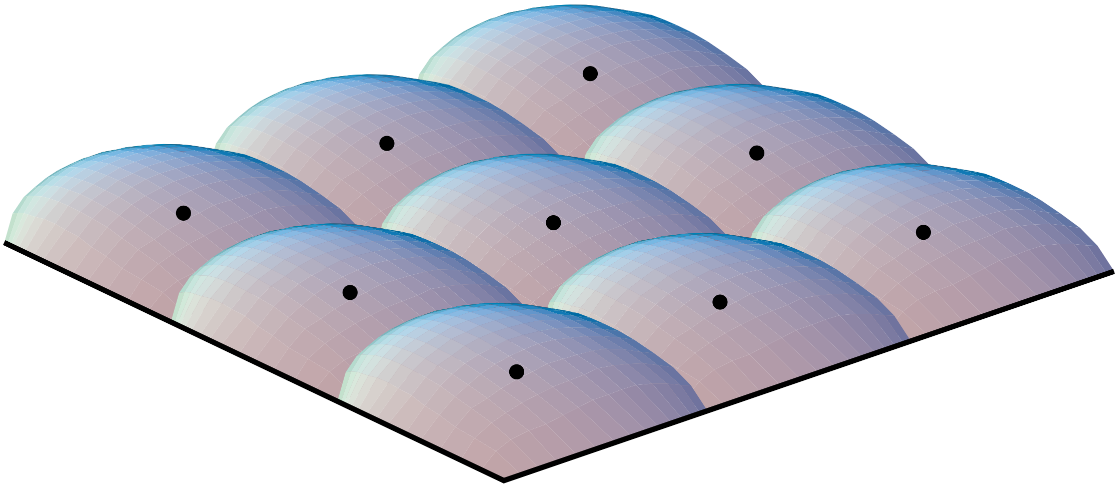


Figure 6

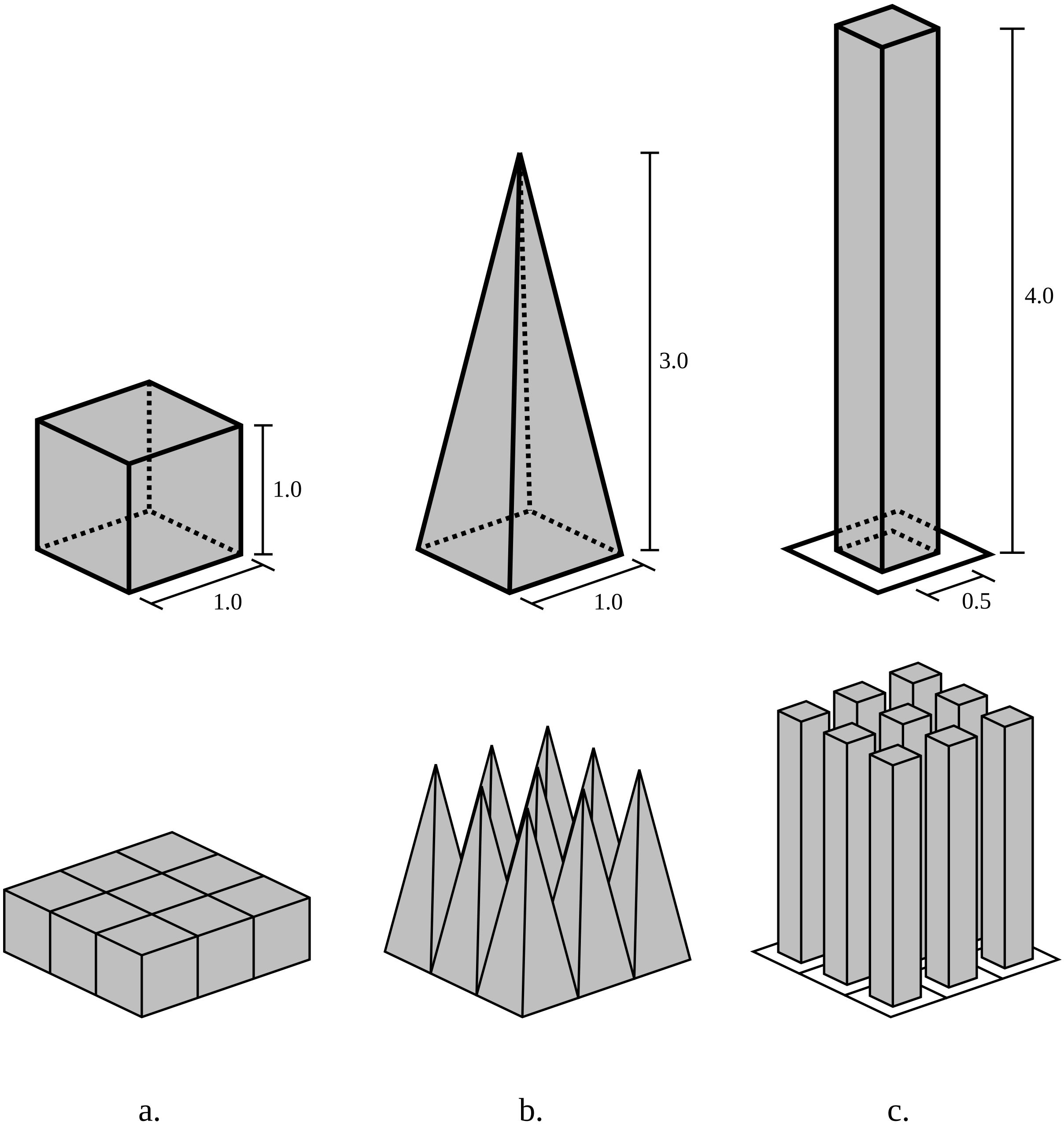


Figure 7

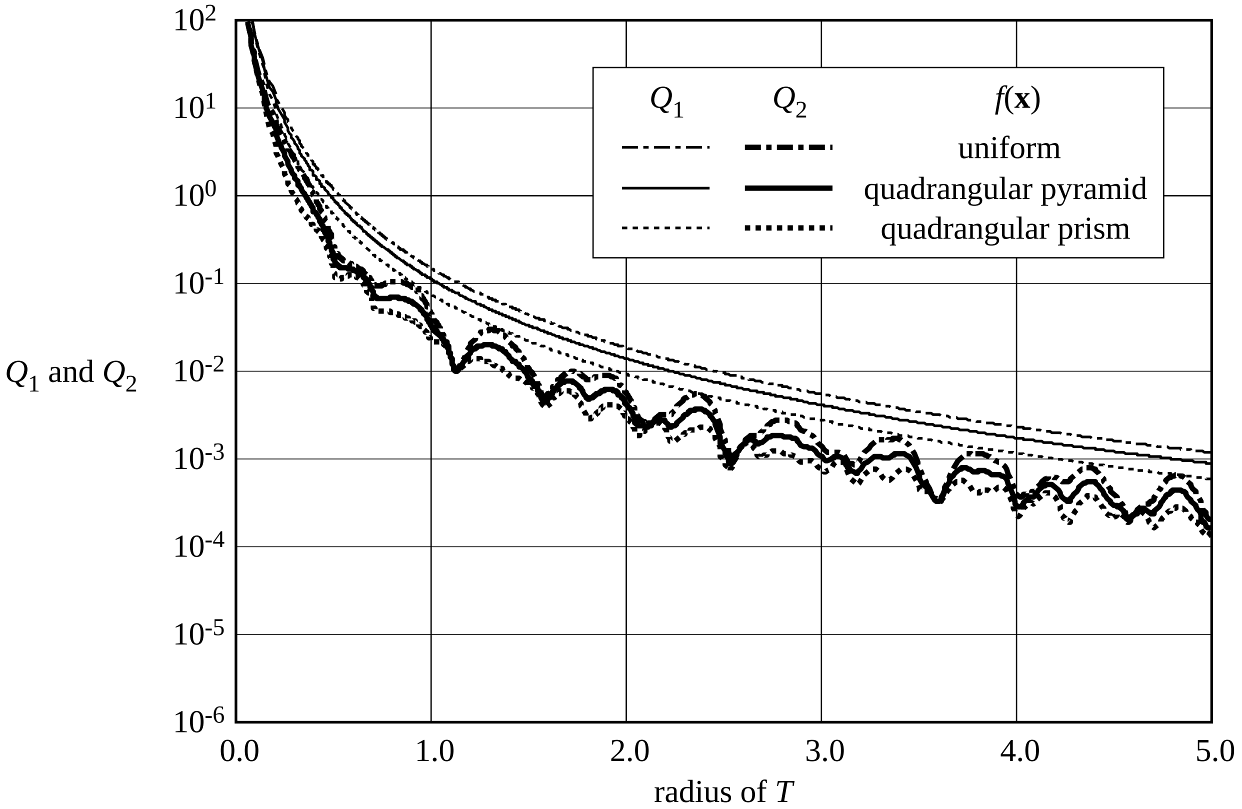
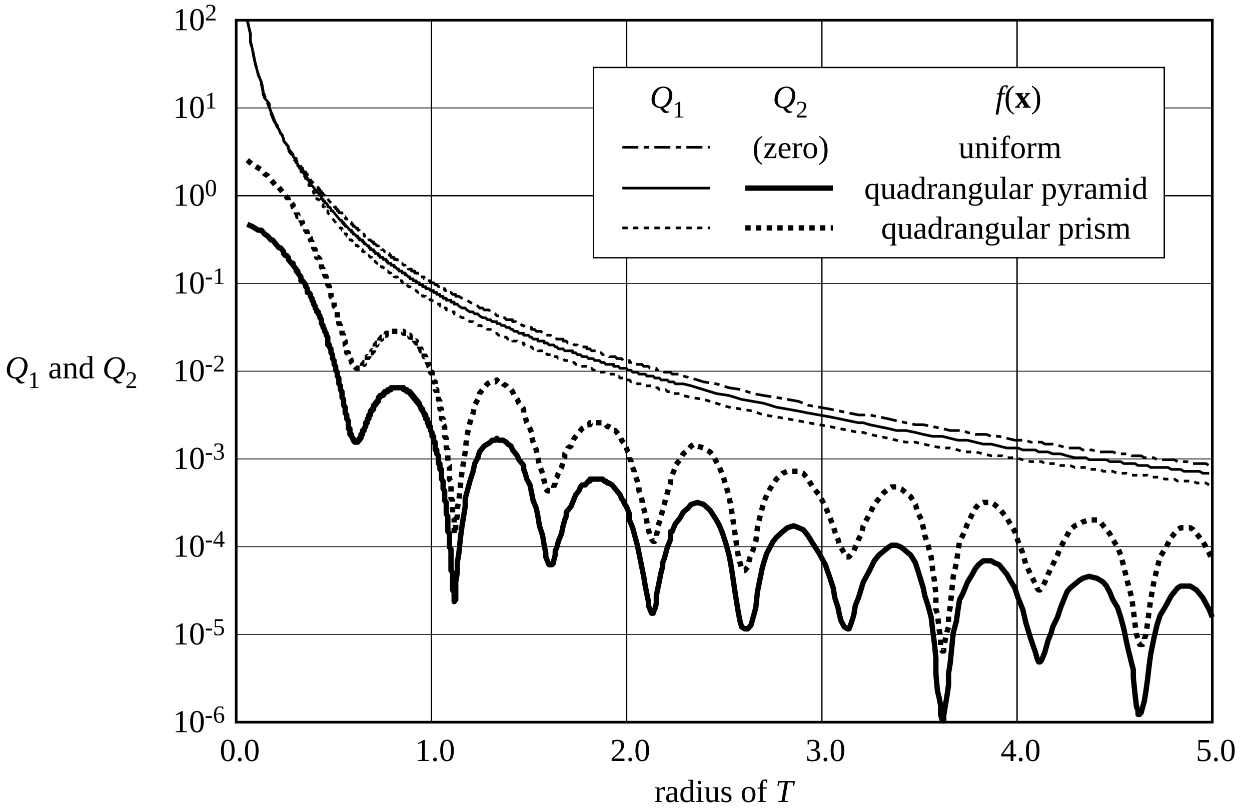

10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
Q_1 and Q_2
0.0
1.0
2.0
3.0
4.0
5.0
radius of T
Q_1
Q_2
$f(\mathbf{x})$
uniform
quadrangular pyramid
quadrangular prism

Figure 8



10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
Q_1 and Q_2
0.0
1.0
2.0
3.0
4.0
5.0
radius of T
Q_1
Q_2
$f(\mathbf{x})$
(zero)
uniform
quadrangular pyramid
quadrangular prism

Figure 9

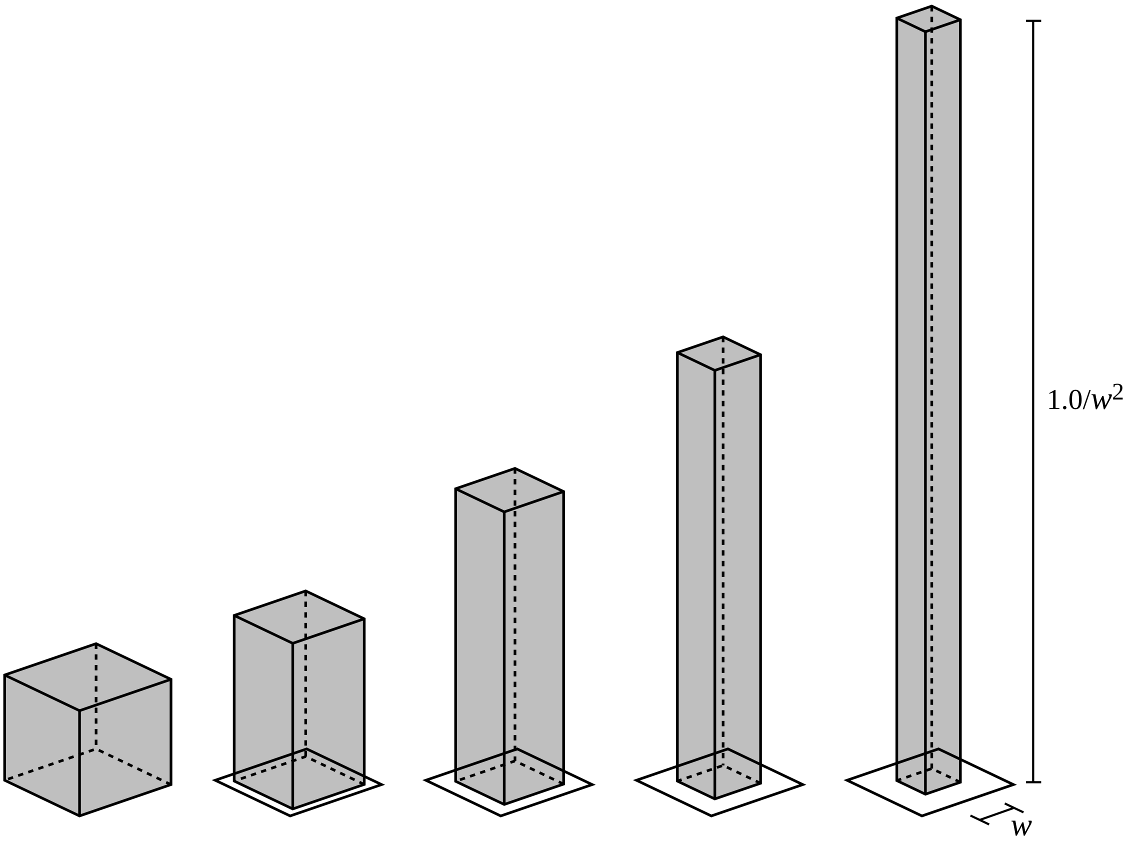


Figure 10

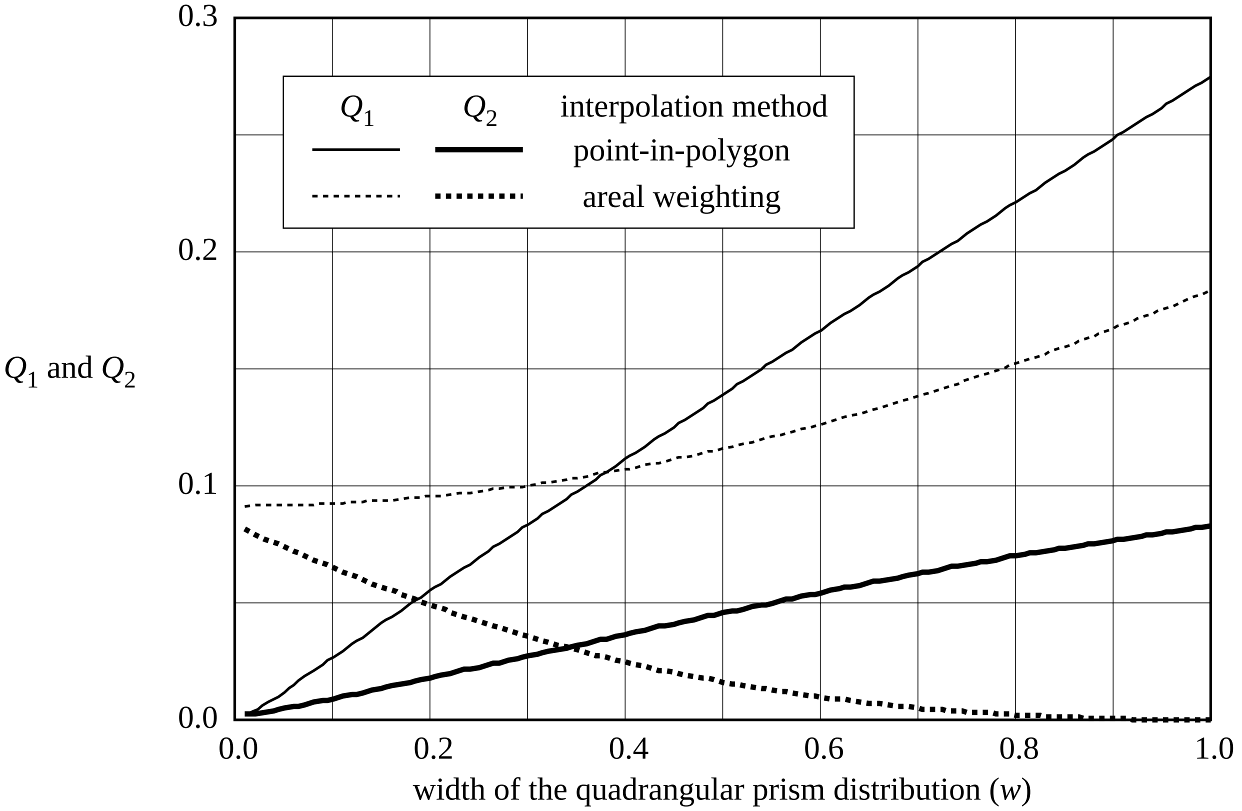

0.3
0.2
0.1
0.0
Q_1 and Q_2
0.0
0.2
0.4
0.6
0.8
1.0
width of the quadrangular prism distribution (w)
Q_1
Q_2
interpolation method
point-in-polygon
areal weighting

Figure 11a

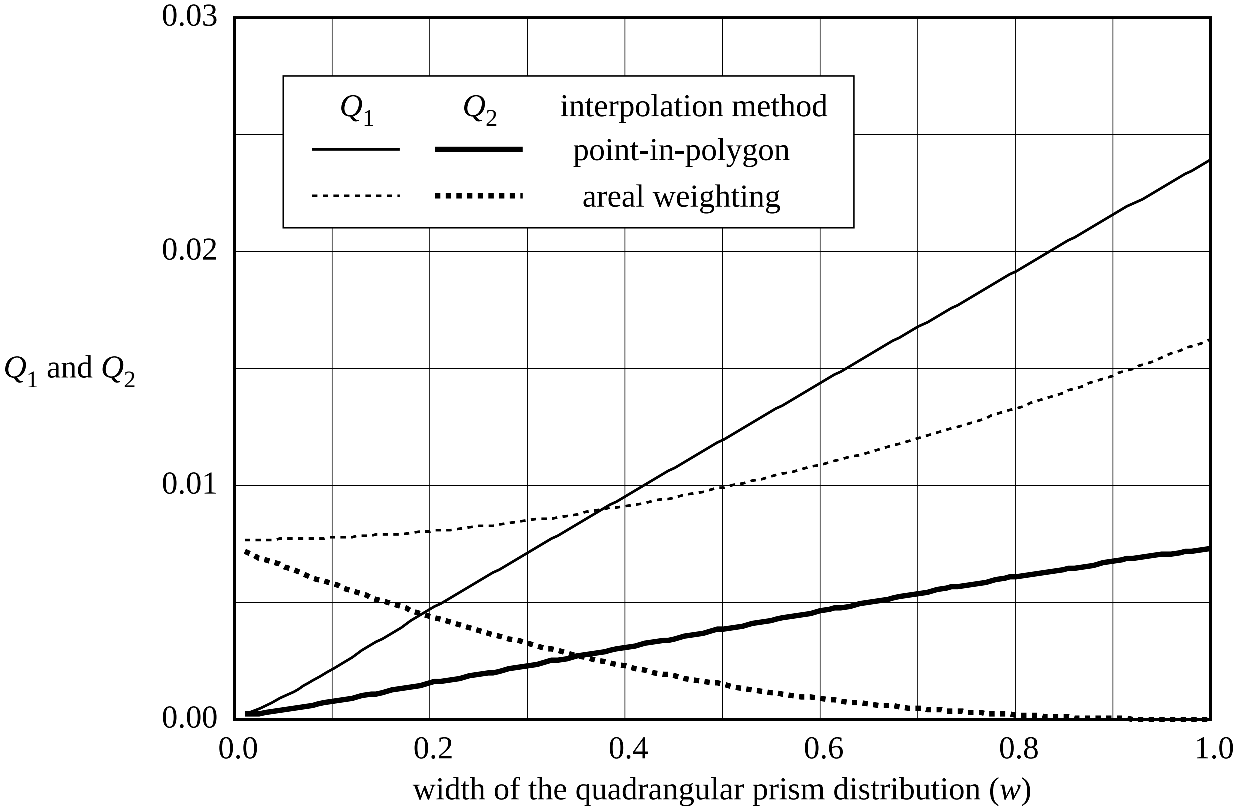
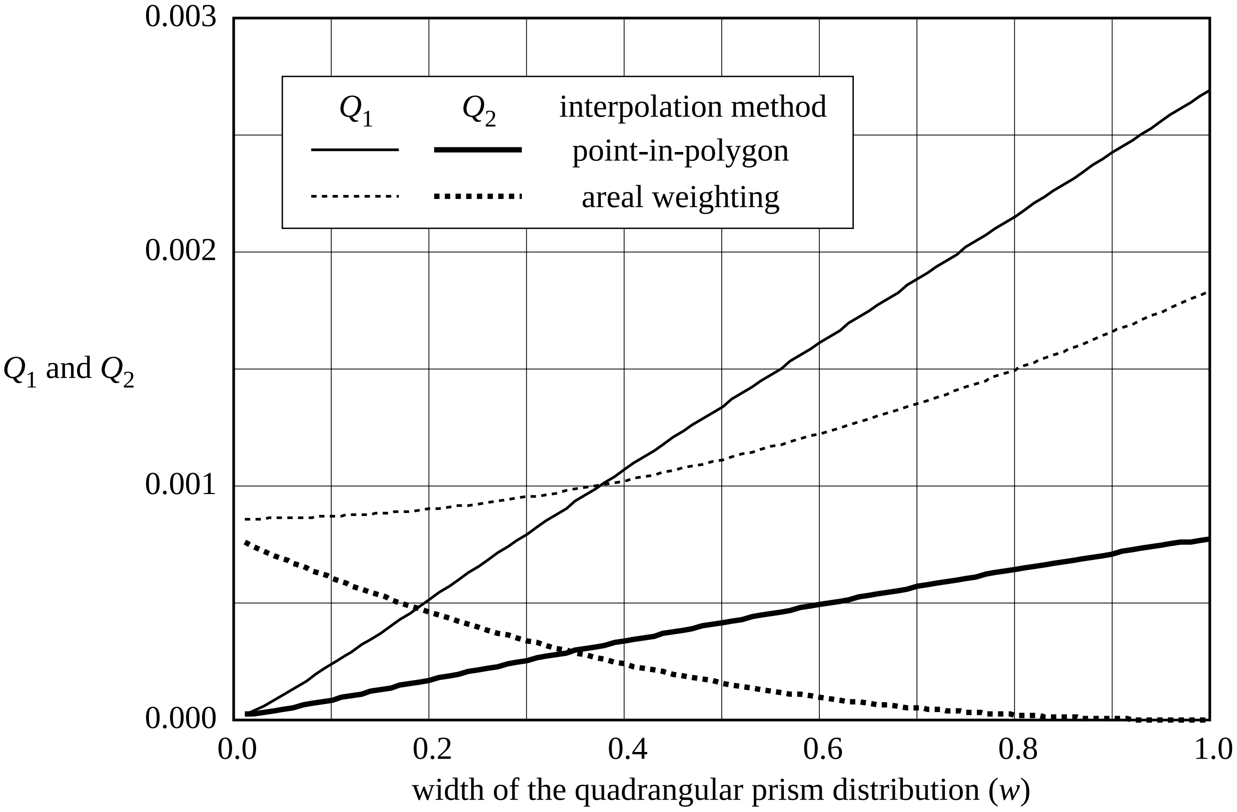

0.03
0.02
0.01
0.00
Q_1 and Q_2
0.0
0.2
0.4
0.6
0.8
1.0
width of the quadrangular prism distribution (w)
Q_1
Q_2
interpolation method
point-in-polygon
areal weighting

Figure 11b



0.003
0.002
0.001
0.000
Q_1 and Q_2
0.0
0.2
0.4
0.6
0.8
1.0
width of the quadrangular prism distribution (w)
Q_1
Q_2
interpolation method
point-in-polygon
areal weighting

Figure 11c

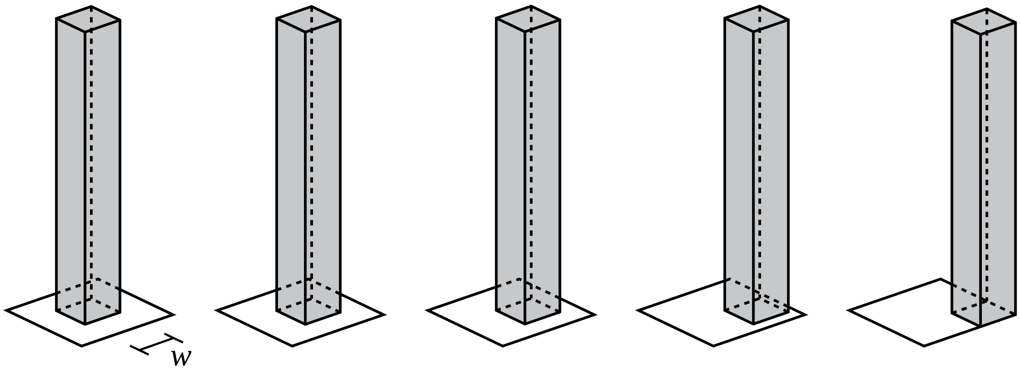


Figure 12

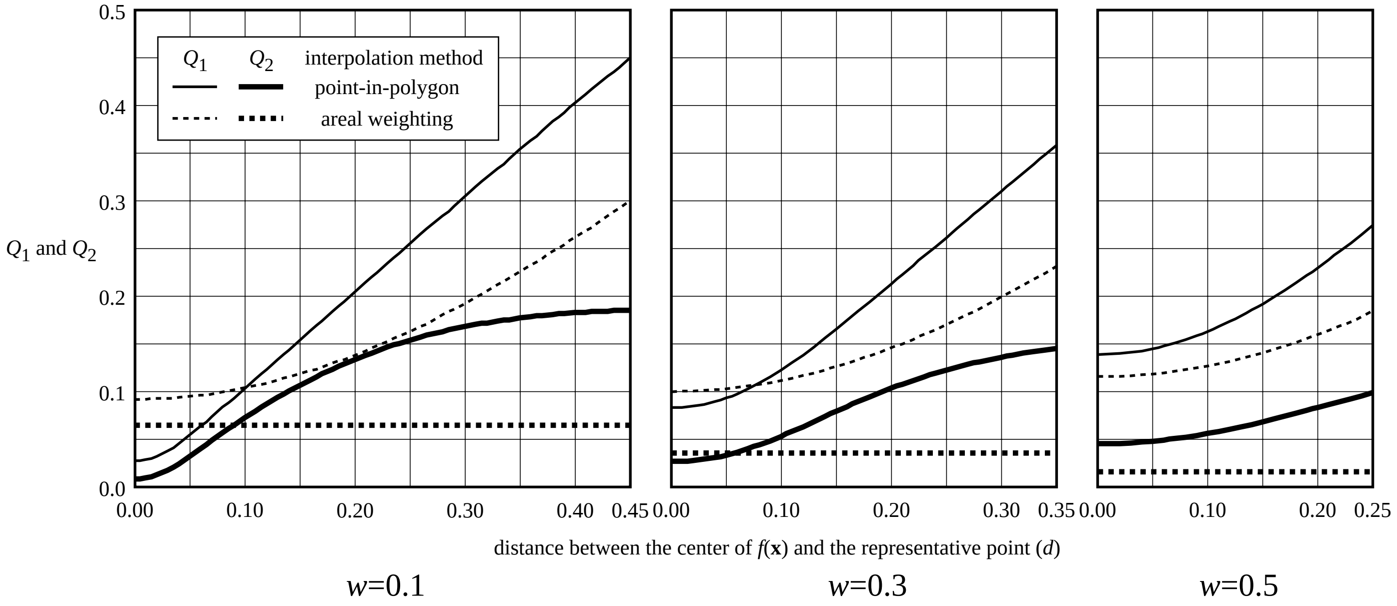

0.5
0.4
0.3
0.2
0.1
0.0
Q_1 and Q_2
Q_1
Q_2
interpolation method
point-in-polygon
areal weighting
0.00
0.10
0.20
0.30
0.40
0.45
0.00
0.10
0.20
0.30
0.35
0.00
0.10
0.20
0.25
distance between the center of $f(\mathbf{x})$ and the representative point (d)
w=0.1
w=0.3
w=0.5

Figure 13a

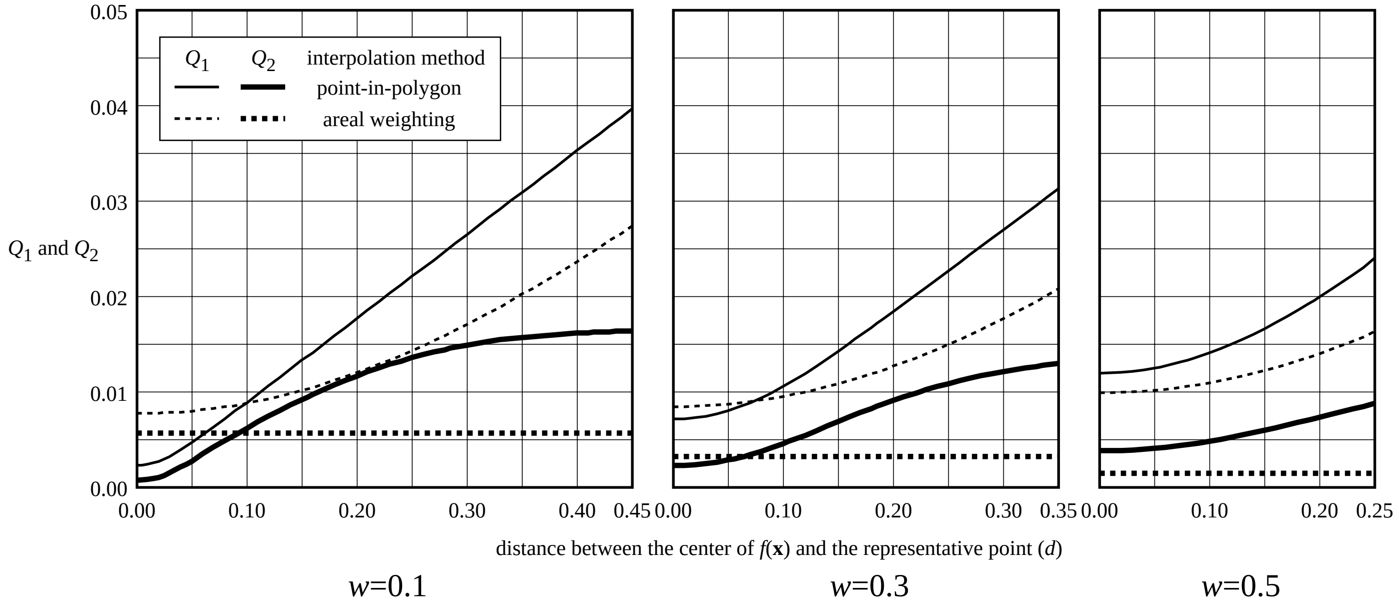

0.05
0.04
0.03
0.02
0.01
0.00
Q_1 and Q_2
Q_1
Q_2
interpolation method
point-in-polygon
areal weighting
0.00
0.10
0.20
0.30
0.40
0.45
0.00
0.10
0.20
0.30
0.35
0.00
0.10
0.20
0.25
distance between the center of $f(\mathbf{x})$ and the representative point (d)
w=0.1
w=0.3
w=0.5

Figure 13b

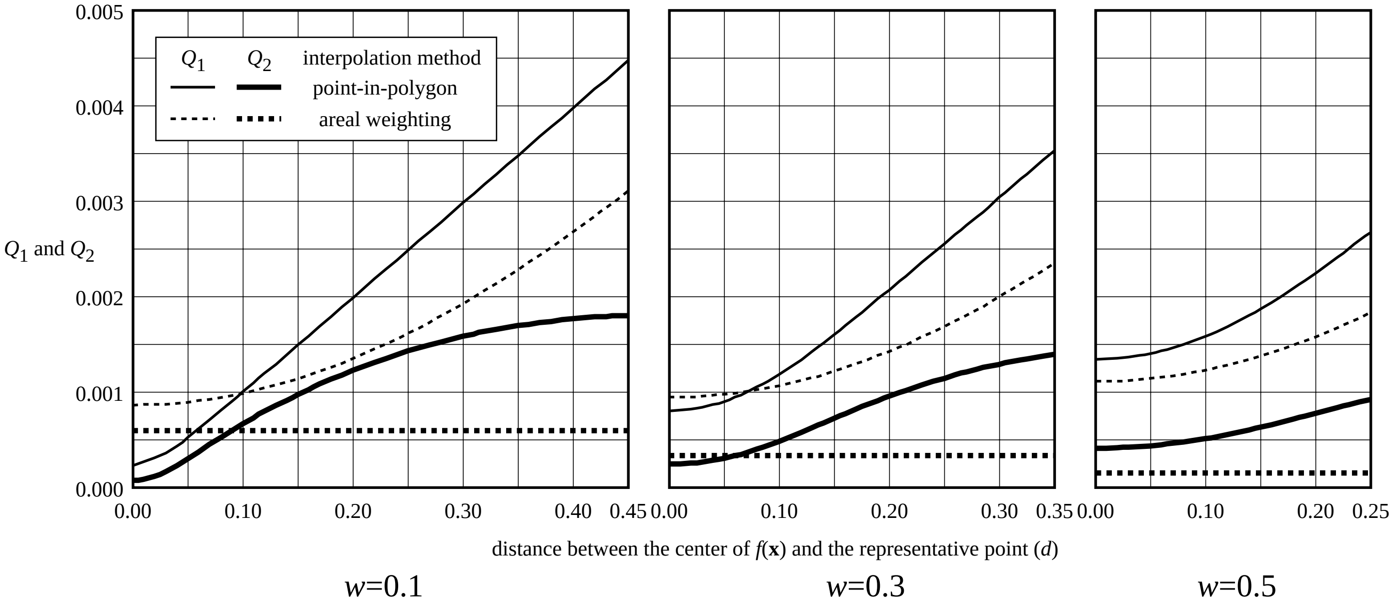

0.005
0.004
0.003
0.002
0.001
0.000
Q_1 and Q_2
Q_1
Q_2
interpolation method
point-in-polygon
areal weighting
0.00
0.10
0.20
0.30
0.40
0.45
0.00
0.10
0.20
0.30
0.35
0.00
0.10
0.20
0.25
distance between the center of $f(\mathbf{x})$ and the representative point (d)
w=0.1
w=0.3
w=0.5

Figure 13c

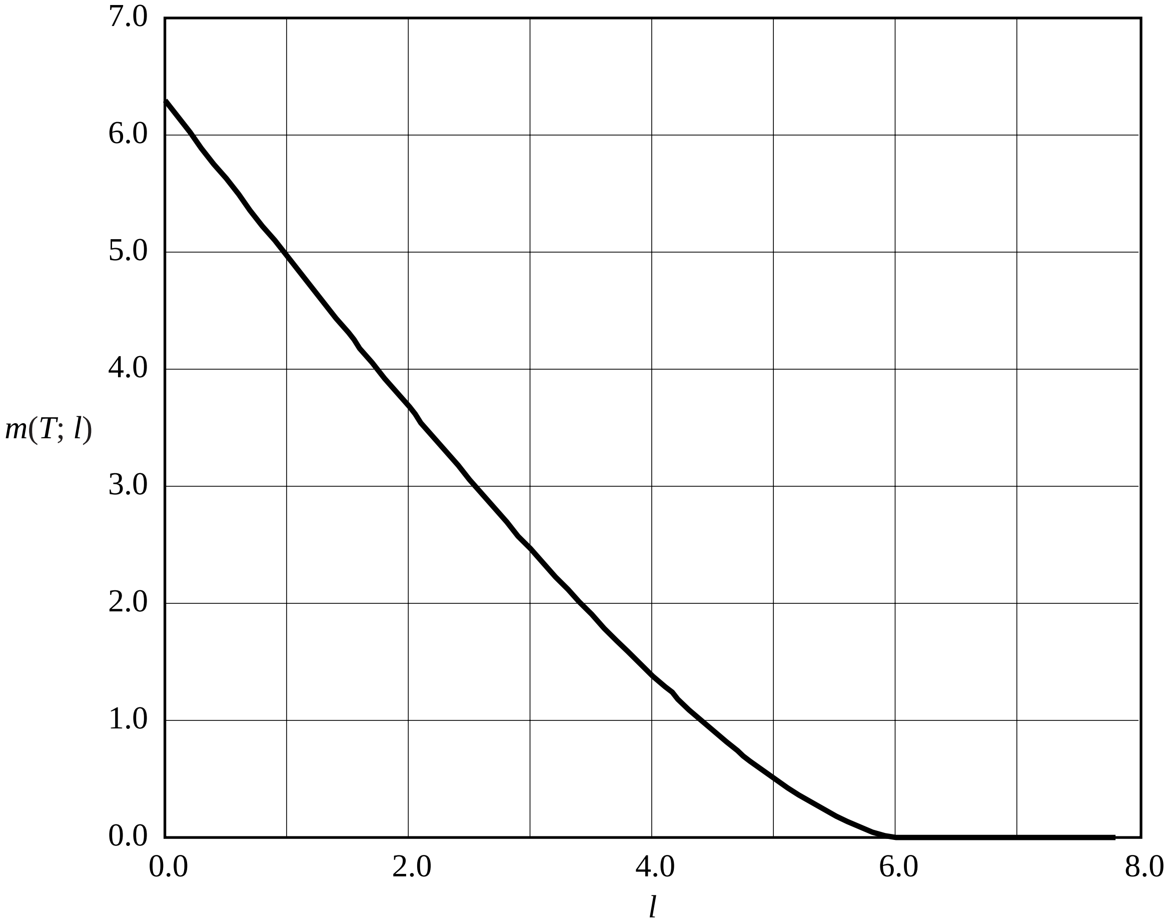
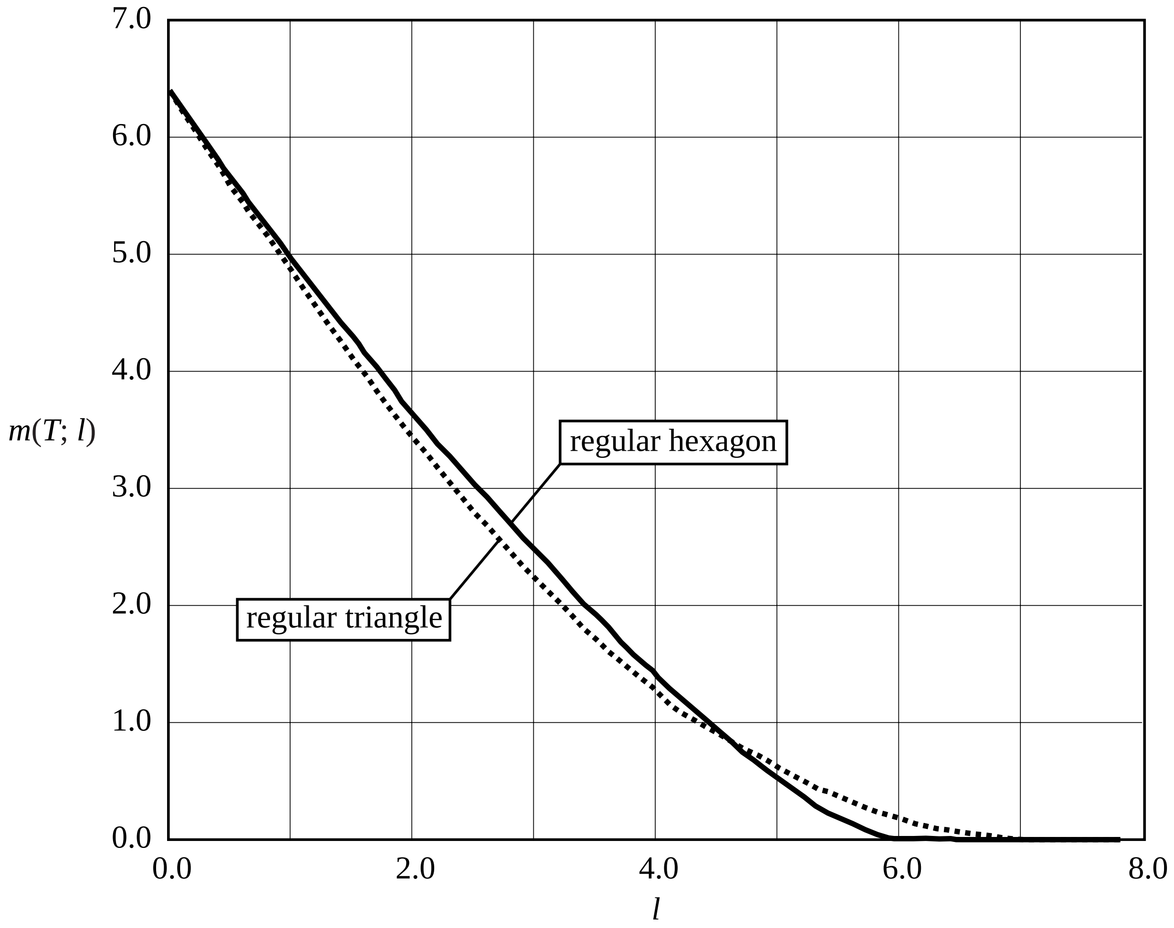


Figure 14



7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
$m(T; l)$
regular hexagon
regular triangle
0.0
2.0
4.0
6.0
8.0
l

Figure 15

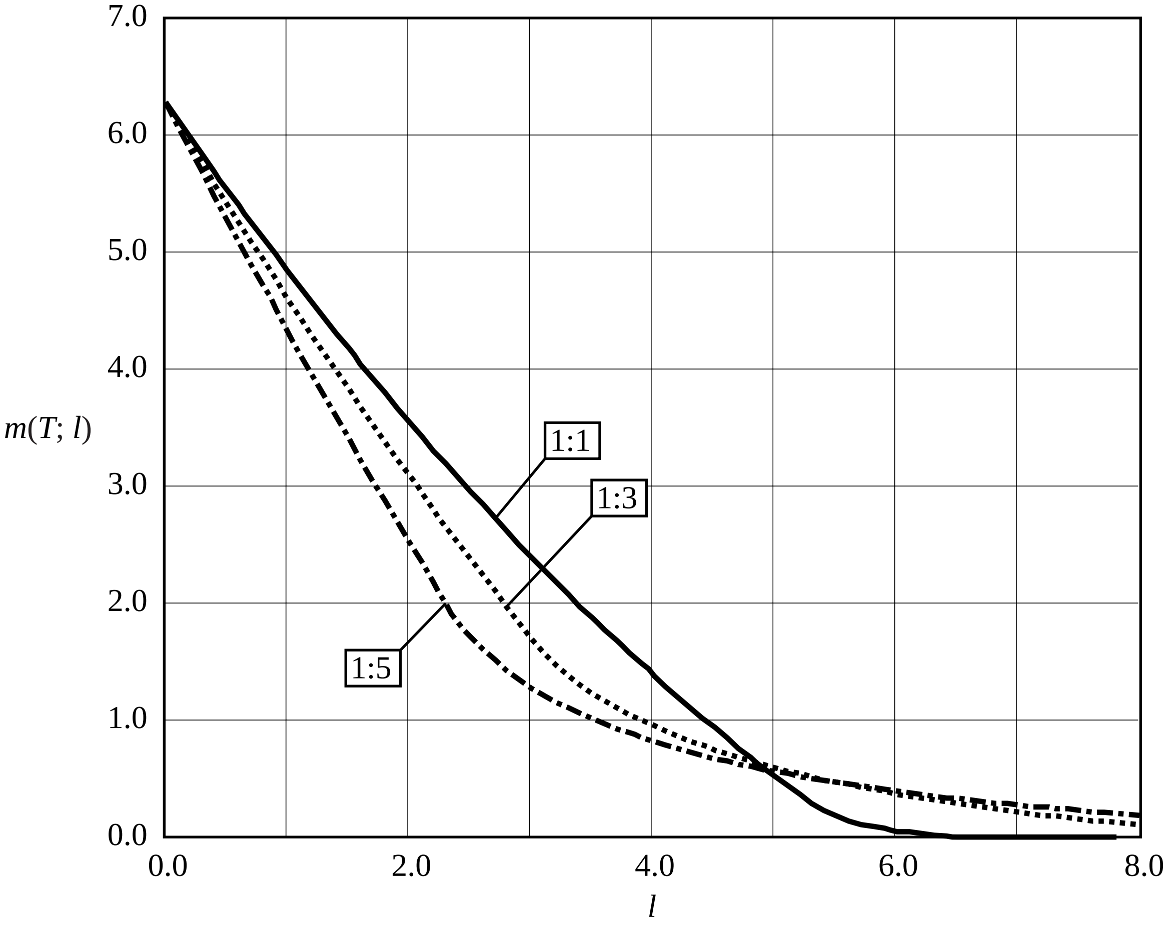


Figure 16

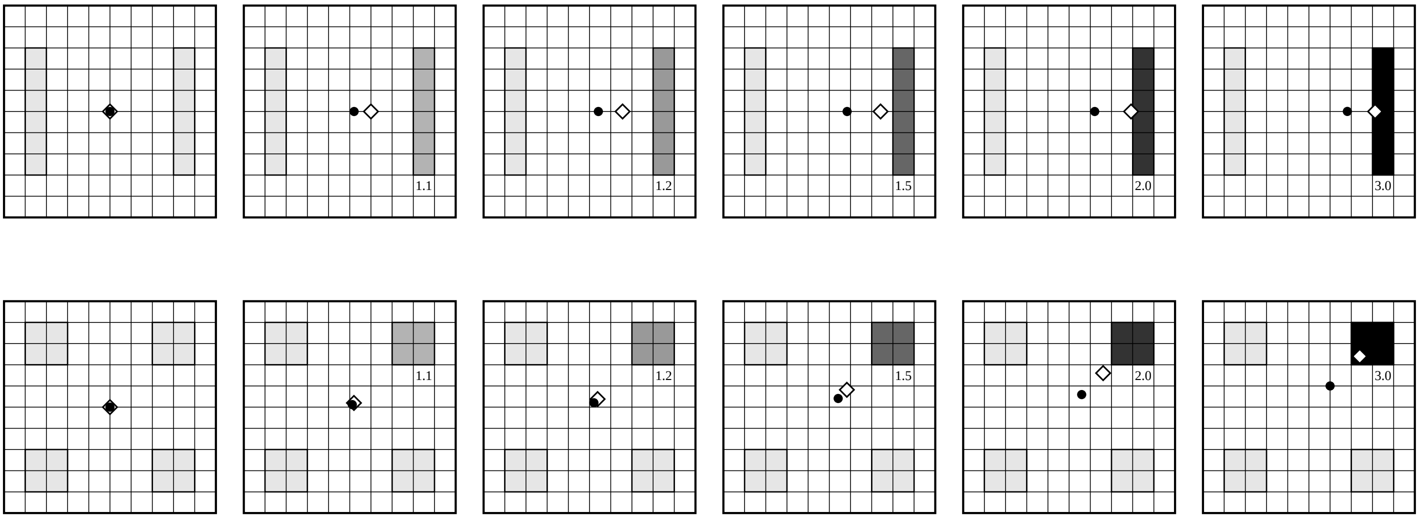


Figure 17