

CSIS Discussion Paper No. 5

**Accuracy of Count Data Transferred through
the Areal Weighting Interpolation 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 Transferred through the Areal Weighting Interpolation Method

Abstract

This paper analyzes the accuracy of count data transferred from a zonal system to an incompatible zone through the areal weighting interpolation method. To treat a variety of situations in a theoretical framework, stochastic models representing areal weighting interpolation are developed. The relationship between the accuracy of estimates and the size of source zones is analytically investigated by use of a proposed model. The results strongly suggested that smaller zones give better estimates on a wide range of zonal systems. The effect of the shape of source zones and the target zone on estimation accuracy is also numerically examined, and it was found to be significant as well as the effect of the size of source zones.

1. Introduction

Social, economic, and demographic data used in GIS are usually provided in an aggregated form based on a zonal system (Rhind 1991). In Japan, for instance, socioeconomic data are aggregated across census tracts and municipal districts. It often happens, however, that an analysis on the data is to be performed in a zone incompatible with the source zones. Let us consider, for instance, a process of store site selection. Given several alternative sites, a market analyst draws hypothetical market areas around the sites in which population count data are necessary for evaluating the alternatives. Unfortunately, market areas are typically represented as circles and incompatible with the census tracts in which demographic data are recorded. Thus the analyst has to estimate the population within the areas from the data aggregated across census tracts. This type of data estimation, that is, data transfer from the source zones to the target zone, is called areal interpolation (Lam 1983, Goodchild et al. 1993), and there have been proposed a lot of areal interpolation methods in the literature (Wright 1936, Markoff and Shapiro 1973, Tobler 1979, Goodchild and Lam 1980, Lam 1983, Flowerdew 1988, Flowerdew and Green 1991, Goodchild et al. 1993, Fisher and Langford 1995, Burrough and McDonnell 1998).

Data estimation through areal interpolation is inevitably inaccurate to some extent, which is undesirable in subsequent spatial analysis. There are at least two methods to improve the accuracy of estimates. One is to choose an areal interpolation method yielding more accurate estimates. Intelligent methods which use supplementary data are generally better than simple methods such as the point-in-polygon method (Okabe and Sadahiro 1997) and the areal weighting interpolation method (Fisher and Langford 1995). It seems desirable to choose intelligent methods if suitable supplementary data, say, remotely sensed data, are available. However, such data are not always available, and the computing cost is problematic especially when the number of zones is very large. Consequently, simple methods are still widely used in GIS.

Another method to improve estimation accuracy is to employ source data whose zones are sufficiently small. If source zones are even smaller than a target zone, we can obtain a highly accurate estimate. Population count data, for instance, are often available in several zonal systems such as census tracts, towns and villages, states, and so forth. It is possible to choose the data aggregated across census tracts which have the smallest zones among them, because they provide more accurate estimates than the others. In practice, however, handling of spatial data consisting of small zones causes some problems: cost of data acquisition, computation, and storage. High-resolution data are more costly than low-resolution data, and they consume much space and time in spatial operation including areal interpolation. Therefore, we try to choose the source data

balancing the cost of data handling and the expected accuracy of estimates.

Such data choice requires us to understand the relationship between the accuracy of estimates and a source zonal system. There are several studies investigating the accuracy of areal interpolation in diverse situations (Flowerdew and Green 1991, Langford et al. 1991, Goodchild et al. 1993). However, since they employed particular sets of spatial objects and zonal systems, it remains unknown whether the results have global applicability. One exception is an analysis based on the Monte Carlo simulation (Fisher and Langford 1995) where a variety of zonal systems were used. Even in this study the distribution of spatial objects is given and the form of zone boundaries is quite limited, thus the obtained results are not proved to be generally applicable.

The objective of this paper is to develop a theoretical framework for analyzing the accuracy of count data transferred from a zonal system to an incompatible zone, and to obtain some general results. We focus on the areal weighting interpolation method (Markoff and Shapiro 1973) because it has been widely used in GIS. Other areal interpolation methods are left for future research.

In Section 2, we propose areal weighting interpolation models which are theoretical basis for the analysis of estimation accuracy. Introducing stochastic models we discuss a set of situations by using their representative probabilistic functions. One of the models is used in Section 3, where the relationship between the accuracy of estimates and the size of source zones is analyzed. In Section 4 we perform a numerical examination of estimation accuracy on lattice systems. Finally, we summarize the conclusions in Section 5.

2. Areal weighting interpolation models

As mentioned in the previous section, estimation accuracy depends on the situation, that is, the source zonal system, the target zone, and the point distribution, which has impeded a general discussion of this subject. One solution to the problem is to perform a computed-assisted simulation of areal interpolation processes (Fisher and Langford 1995). This approach, however, has some limitations. First, because of its computational cost, the realized situations are quite limited. Areal interpolation of count data depends on the above three factors, and it is impossible to try numerous situations varying in all the factors. Second, a simulation-based approach is not directly connected with a theoretical analysis. Suppose, for instance, the relationship between the accuracy of estimates and the size of source zones. Performing computed-assisted simulations we can obtain numerical information on this relationship, say, a correlation coefficient. However, this result is not supported by a firm theory, and thus the applicability of the result still remains unknown.

To overcome these difficulties, we propose two stochastic models which we call the *areal weighting interpolation models* as a framework for analyzing estimation accuracy. We replace a variety of situations by their representative stochastic models, and discuss the models rather than individual situations. By this we can avoid computationally expensive simulations to consider numerous situations.

Among the factors affecting the accuracy of estimates we are most concerned with the source zonal system. Hence the source zonal system is given and fixed in models. The shape and size of the target zone are also given. In contrast to these factors, the point objects are assumed to follow a probability distribution, which is a representative of diverse distributions. The location of the target zone also follows a probability distribution. We should note that, however, the location of points and the target zone can be fixed by adopting a suitable probability distribution, if a specific situation has to be considered.

2.1 Basic model

Suppose a source zonal system S , a region Z_0 of area A_0 which consists of K zones Z_1, Z_2, \dots, Z_K (Figure 1). The area of Z_i is denoted by A_i . The region Z_0 represents, say, a county, and each zone corresponds to a census tract, a school district, or a postal zone by which count data are reported. A target zone T of area B in which count data need to be estimated, say, a hypothetical market area, is assumed to be randomly dropped in such a way that it intersects Z_0 (Figure 2). The location of T is represented by a binary function

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

In the region Z_0 , N points (say, population) are independently and identically distributed according to a probability density function $f(\mathbf{x})$. The location of point j is denoted by \mathbf{y}_j .

Figure 1 A source zonal system S . Gray-shaded area represents the region Z_0 .

Figure 2 Possible locations of the target zone T (the ellipses). Gray-shaded area represents the region Z_0 . Note that the target zone is not necessarily wholly contained in Z_0 .

Let us consider the areal weighting interpolation in the above setting. The number of points in T is given by

$$M = \sum_j C(\mathbf{y}_j). \quad (2)$$

This value is unknown to an analyst and has to be estimated through the areal weighting

interpolation as follows. Suppose a function $U_i(\mathbf{x})$ defined by

$$U_i(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in Z_i \\ 0 & \text{otherwise} \end{cases}. \quad (3)$$

The number of points in Z_j is

$$n_i = \sum_j U_i(\mathbf{y}_j). \quad (4)$$

The areal weighting interpolation assumes that the points are uniformly distributed in each zone. Therefore, given the number of points for every zone, we have an estimate of M ,

$$\begin{aligned} \hat{M} &= \sum_i \frac{\int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x}}{A_i} n_i \\ &= \sum_i \frac{\int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x}}{A_i} \sum_j U_i(\mathbf{y}_j) \end{aligned} \quad (5)$$

Estimation error of M is given by

$$\begin{aligned} \varepsilon &= M - \hat{M} \\ &= \sum_j C(\mathbf{y}_j) - \sum_i \frac{\int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x}}{A_i} \sum_j U_i(\mathbf{y}_j) \end{aligned} \quad (6)$$

In the areal weighting interpolation models, the estimation error ε changes stochastically because the point objects and the target zone are located according to probability distributions. To evaluate the error, we adopt the mean square error (MSE) of ε as a measure defined by

$$MSE[S] = E[\varepsilon^2]. \quad (7)$$

The calculation of $E[\varepsilon^2]$ is shown in Appendix A1, thus we show only the result.

$$\begin{aligned} MSE[S] &= N(N-1) \int_{\mathbf{t} \in Z_0} \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) f(\mathbf{t}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \\ &\quad + N \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) \Pr[\mathbf{x} \in T] d\mathbf{x} \\ &\quad - 2N(N-1) \sum_i \frac{1}{A_i} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \\ &\quad - 2N \sum_i \frac{1}{A_i} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \\ &\quad + N(N-1) \sum_i \sum_{i'} \frac{1}{A_i A_{i'}} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_{i'}} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} f(\mathbf{x}) d\mathbf{x} \\ &\quad + N \sum_i \frac{1}{A_i^2} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (8)$$

Equation (8) contains the probabilities $\Pr[\mathbf{x} \in T]$ and $\Pr[\mathbf{x} \cup \mathbf{t} \in T]$ which are not explicitly given. These values can be computed as follows.

Let $m(T; Z_0)$ be the measure of the set of all figures congruent to T intersecting Z_0 . The probabilities $\Pr[\mathbf{x} \in T]$ is written as

$$\Pr[\mathbf{x} \in T] = \frac{2\pi B}{m(T; Z_0)}. \quad (9)$$

If both T and Z_0 are convex, $m(T; Z_0)$ is given by

$$m(T; Z_0) = 2\pi(A_0 + B) + P_0 P_T, \quad (10)$$

where P_0 and P_T are the perimeters of Z_0 and T , respectively. Otherwise, computation of $m(T; Z_0)$ requires a numerical simulation or a spatial sampling.

The probability $\Pr[\mathbf{x} \cup \mathbf{t} \in T]$ is given by

$$\Pr[\mathbf{x} \cup \mathbf{t} \in T] = \frac{m(T; |\mathbf{x} - \mathbf{t}|)}{m(T; Z_0)}, \quad (11)$$

where $m(T; l)$ is the measure of the set of all figures congruent to T containing two points separated by a distance l . If T has a simple shape, $m(T; l)$ is represented by an explicit form (for details, see Santaló 1976, Sadahiro 1999). For instance, if T is a circle of radius r ,

$$m(T; l) = \begin{cases} 4\pi r^2 \arccos\left(\frac{l}{2r}\right) - \pi l \sqrt{4r^2 - l^2} & (l \leq 2r), \\ 0 & (l > 2r). \end{cases} \quad (12)$$

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

$$m(T; l) = \begin{cases} 2\pi bc - 4(b+c)l + 2l^2 & (l \leq b), \\ 4c\sqrt{l^2 - b^2} - 4cl - 2b^2 + 4bc \arcsin \frac{b}{l} & (b < l \leq c), \\ 4c\sqrt{l^2 - b^2} + 4b\sqrt{l^2 - c^2} - 2(b^2 + c^2 + l^2) & (c < l \leq \sqrt{b^2 + c^2}), \\ +4bc \left(\arcsin \frac{c}{l} - \arccos \frac{b}{l} \right) & \\ 0 & (\sqrt{b^2 + c^2} < l). \end{cases} \quad (13)$$

If T has more complicated shape, $m(T; l)$ is numerically computable by using

$$m(T; l) = \frac{B^2}{l} g_T(l), \quad (14)$$

where $g_T(l)$ is the probability density function of the distance between two points that are randomly distributed in T (derivation of equation (14) is shown in Appendix A2). The function $g_T(l)$ can be easily obtained in GIS: 1) overlay a square lattice on T , 2) calculate the point-to-point distance for all pair of grid points in T , 3) make a histogram of the distance.

2.2 Periodic continuation model

The basic model proposed in the previous subsection allows us to evaluate the estimation accuracy of areal weighting interpolation in diverse situations. Calculation of

equation (8), however, is sometimes computationally expensive because calculation of $\Pr[\mathbf{x} \in T]$ may require numerical simulations. To reduce the computational cost, we propose another areal weighting interpolation model which we call the *periodic continuation model*.

The model setting is almost the same as that of the basic model. We hence describe only the differences.

1) The region Z_0 has such a shape that can cover a plane by its lattice (Figure 3). Rectangles, parallelograms, and regular hexagons meet this requirement.

2) The region Z_0 is surrounded by its copies, and the copies have the same zonal system and point distribution as those of Z_0 (Figure 4). This assumption is often called *periodic continuation* (Ripley 1981, Stoyan and Stoyan 1995).

3) If T does not completely lie in Z_0 , we replace the portion of T outside Z_0 by its corresponding figure as shown in Figure 5. We then assume that all possible shapes and positions of T appear randomly.

Figure 3 Possible shapes of the region Z_0 (the upper row). It is confirmed from the lower row figures that those shapes satisfy the requirement 1).

Figure 4 The region Z_0 and its surrounding copies.

Figure 5 Transformation of T .

In the above setting we have

$$\Pr[\mathbf{x} \in T] = \frac{B}{A_0}. \quad (15)$$

Substituting equation (15) into equation (8), we obtain

$$\begin{aligned}
MSE[S] &= N(N-1) \int_{\mathbf{t} \in Z_0} \int_{\mathbf{x} \in Z_0} f(\mathbf{x})f(\mathbf{t}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x}d\mathbf{t} \\
&\quad + N \left(\frac{B}{A_0} \right) \\
&\quad - 2N(N-1) \sum_i \frac{1}{A_i} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x}d\mathbf{t} \\
&\quad - 2N \sum_i \frac{1}{A_i} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x}d\mathbf{t} \\
&\quad + N(N-1) \sum_i \sum_{i'} \frac{1}{A_i A_{i'}} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_{i'}} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x}d\mathbf{t} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} f(\mathbf{x}) d\mathbf{x} \\
&\quad + N \sum_i \frac{1}{A_i^2} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x}d\mathbf{t} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x}
\end{aligned} \tag{16}$$

The probability $\Pr[\mathbf{x} \cup \mathbf{t} \in T]$ is given by

$$\Pr[\mathbf{x} \cup \mathbf{t} \in T] = \sum_{\mathbf{u} \in \Omega(\mathbf{t})} \frac{m(T; |\mathbf{x} - \mathbf{u}|)}{2\pi A_0}, \tag{17}$$

where $\Omega(\mathbf{t})$ is a set of points corresponding to \mathbf{t} in the surrounding copies of Z_0 (Figure 6). Equation (16) then becomes

$$\begin{aligned}
MSE[S] &= \frac{N(N-1)}{2\pi A_0} \int_{\mathbf{t} \in Z_0} \int_{\mathbf{x} \in Z_0} f(\mathbf{x})f(\mathbf{t}) \sum_{\mathbf{u} \in \Omega(\mathbf{t})} m(T; |\mathbf{x} - \mathbf{u}|) d\mathbf{x}d\mathbf{t} \\
&\quad + N \left(\frac{B}{A_0} \right) \\
&\quad - \frac{N(N-1)}{\pi A_0} \sum_i \frac{1}{A_i} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) \sum_{\mathbf{u} \in \Omega(\mathbf{t})} m(T; |\mathbf{x} - \mathbf{u}|) d\mathbf{x}d\mathbf{t} \\
&\quad - \frac{N}{\pi A_0} \sum_i \frac{1}{A_i} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) \sum_{\mathbf{u} \in \Omega(\mathbf{t})} m(T; |\mathbf{x} - \mathbf{u}|) d\mathbf{x}d\mathbf{t} \\
&\quad + \frac{N(N-1)}{2\pi A_0} \sum_i \sum_{i'} \frac{1}{A_i A_{i'}} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_{i'}} \sum_{\mathbf{u} \in \Omega(\mathbf{t})} m(T; |\mathbf{x} - \mathbf{u}|) d\mathbf{x}d\mathbf{t} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} f(\mathbf{x}) d\mathbf{x} \\
&\quad + \frac{N}{2\pi A_0} \sum_i \frac{1}{A_i^2} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} \sum_{\mathbf{u} \in \Omega(\mathbf{t})} m(T; |\mathbf{x} - \mathbf{u}|) d\mathbf{x}d\mathbf{t} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x}
\end{aligned} \tag{18}$$

Note that equation (18) does not contain the term $m(T; Z_0)$ whose computation often requires a numerical simulation. The measure $m(T; l)$ is easily computable as mentioned earlier, and so is $MSE[S]$. The periodic continuation model is more tractable than the basic model.

Figure 6 Point located at \mathbf{t} and the set of its corresponding points $\Omega(\mathbf{t})$ in the surrounding

regions (Gray-shaded area represents the region Z_0).

The choice of the above two models depends on the case. Concerning the model setting, the basic model is more flexible than the periodic continuation model. The latter puts some additional assumptions, and there may be the case where they are not acceptable. If the assumptions are allowable, the periodic continuation model would be a better choice because it is computationally less expensive. Moreover, since equation (18) is more simple than equation (8), the periodic continuation model seems more suitable for an analytical investigation of estimation accuracy.

3. Accuracy of estimates and the size of source zones

Having defined the areal weighting interpolation models, we are now ready to analyze the accuracy of estimates. In this section we take an analytical approach to the relationship between the estimation accuracy and the size of source zones.

We consider the case where the points (say, population) are distributed in the region Z_0 according to the uniform distribution. This implies that in a statistical sense an even distribution of population is assumed in Z_0 . Formally, the assumption is represented as

$$f(\mathbf{x}) = \frac{1}{A_0}. \quad (19)$$

Indeed, this assumption scarcely holds in a strict sense. However, analyzing the case of uniformly distributed points would be helpful in considering a variety of point distributions, especially for those analogous to the uniform distribution.

From the two available models we choose the periodic continuation model because of its tractability. Substitution of equation (19) into equation (16) yields

$$\begin{aligned} MSE[S] &= N(N-1) \left(\frac{1}{A_0} \right)^2 \int_{\mathbf{t} \in Z_0} \int_{\mathbf{x} \in Z_0} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x}d\mathbf{t} \\ &\quad + N \left(\frac{B}{A_0} \right) \\ &\quad - 2N(N-1) \left(\frac{1}{A_0} \right)^2 \sum_i \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_0} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x}d\mathbf{t} \\ &\quad - N \frac{1}{A_0} \sum_i \frac{1}{A_i} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x}d\mathbf{t} \\ &\quad + N(N-1) \left(\frac{1}{A_0} \right)^2 \sum_i \sum_{i'} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_{i'}} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x}d\mathbf{t} \end{aligned} \quad (20)$$

Noticing that

$$\sum_i \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_0} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} = \int_{\mathbf{t} \in Z_0} \int_{\mathbf{x} \in Z_0} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t}, \quad (21)$$

and that

$$\sum_i \sum_{i'} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_{i'}} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} = \int_{\mathbf{t} \in Z_0} \int_{\mathbf{x} \in Z_0} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t}, \quad (22)$$

we obtain

$$\begin{aligned} MSE[S] &= N \left(\frac{B}{A_0} \right) - N \frac{1}{A_0} \sum_i \frac{1}{A_i} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \\ &= \frac{N}{A_0} \left\{ B - \frac{1}{2\pi A_0} \sum_i \frac{1}{A_i} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} \sum_{\mathbf{u} \in \Omega(\mathbf{t})} m(T; |\mathbf{x} - \mathbf{u}|) d\mathbf{x} d\mathbf{t} \right\}. \end{aligned} \quad (23)$$

Equation (23) indicates that the MSE does not depend on the spatial arrangement of zones in Z_0 . Three zonal systems shown in Figure 7 have the same MSE.

Figure 7 Zonal systems having the same MSE.

Using equation (23), we analyze the relationship between the accuracy of estimates and the size of source zones. To this end, we consider three typical zonal systems: the hierarchical zonal system, the zonal system consisting of sets of congruent figures, and the lattice system.

3.1 Hierarchical zonal systems

Socioeconomic data are often aggregated in a hierarchical zonal system. Population count data, for instance, are available at various levels of hierarchy, say, census tracts, towns and villages, and states. In a hierarchical zonal system, a higher level system is obtained by combining several zones of a lower level system (Figure 8).

We now have a question whether a lower level zonal system, that is, a system consisting of smaller zones, really yields more accurate estimates than a higher level system. In other words, we wish to examine whether the combination of zones lowers the estimation accuracy.

Figure 8 A hierarchical zonal system.

Suppose hierarchical zonal systems S_1 and S_2 , where S_1 is at the lower level of S_2 . We assume for simplicity that S_1 consists of two zones Z_1 and Z_2 , while S_1 consists of zone Z_0 . The MSE's are given by

$$\begin{aligned}
MSE[S_1] &= \frac{N}{A_0} \left\{ B - \frac{1}{2\pi A_0} \sum_i \frac{1}{A_i} \int_{t \in Z_i} \int_{x \in Z_i} \sum_{u \in \Omega(t)} m(T; |\mathbf{x} - \mathbf{u}|) d\mathbf{x} d\mathbf{t} \right\} \\
&= \frac{N}{A_0} \left\{ B - \sum_i \frac{1}{A_i} \int_{t \in Z_i} \int_{x \in Z_i} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \right\}
\end{aligned} \tag{24}$$

and

$$MSE[S_2] = \frac{N}{A_0} \left\{ B - \frac{1}{A_0} \int_{t \in Z_0} \int_{x \in Z_0} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \right\}, \tag{25}$$

respectively. From these equations we have

$$\begin{aligned}
\Delta MSE &= MSE[S_2] - MSE[S_1] \\
&= \frac{N}{A_0} \left\{ \begin{aligned} &\frac{1}{A_1} \int_{t \in Z_1} \int_{x \in Z_1} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \\ &+ \frac{1}{A_2} \int_{t \in Z_2} \int_{x \in Z_2} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \\ &- \frac{1}{A_0} \int_{t \in Z_0} \int_{x \in Z_0} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \end{aligned} \right\}.
\end{aligned} \tag{26}$$

After a few steps of calculation equation (26) becomes

$$\Delta MSE = \frac{N}{A_0^2 A_1 A_2} \mathbb{E} \left[\left\{ A_2 \int_{x \in Z_1} C(\mathbf{x}) d\mathbf{x} - A_1 \int_{x \in Z_2} C(\mathbf{x}) d\mathbf{x} \right\}^2 \right]. \tag{27}$$

and thus

$$\Delta MSE \geq 0. \tag{28}$$

This indicates that S_1 is better than S_2 with regard to the mean square error.

The above discussion can be easily extended to a hierarchical zonal system consisting of more than two zones, which leads to a conclusion that in any hierarchical system a lower level system always gives a smaller MSE than a higher level system. It is desirable to choose a lower level system in a hierarchical zonal system regardless of the shape and size of the source and target zones.

3.2 Zonal systems consisting of sets of congruent figures

We next consider the zonal system that consists of sets of congruent figures. Suppose a zonal system S , the region Z_0 consists of m types of figures (Figure 8a). We denote the covering ratio of type i figures as α_i ($\alpha_1 + \alpha_2 + \dots + \alpha_m = 1$). The MSE of this system is given by

$$MSE[S] = \frac{N}{A_0} \left\{ B - \frac{1}{2\pi} \sum_i \frac{\alpha_i}{A_i^2} \int_{t \in Z_i} \int_{x \in Z_i} \sum_{u \in \Omega(t)} m(T; |\mathbf{x} - \mathbf{u}|) d\mathbf{x} d\mathbf{t} \right\}. \tag{29}$$

We then suppose a zonal system S_i , the region Z_0 consists of only type i figures (Figures 8b and 8c). The mean square error $MSE[S_i]$ is given by

$$MSE[S_i] = \frac{N}{A_0} \left\{ B - \frac{1}{2\pi A_i^2} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} \sum_{\mathbf{u} \in \Omega(\mathbf{t})} m(T; |\mathbf{x} - \mathbf{u}|) d\mathbf{x} d\mathbf{t} \right\}. \quad (30)$$

From equations (29) and (30) we obtain

$$MSE[S] = \sum_i \alpha_i MSE[S_i]. \quad (31)$$

As seen in equation (31), $MSE[S]$ is given by the linear combination of $MSE[S_i]$'s weighted by the covering ratio (Figure 8). This implies that the MSE of a zonal system consisting of sets of congruent figures lies among MSE's of zonal systems consisting of all congruent figures. The following inequation holds for the zonal systems shown in Figure 9.

$$MSE[S_1] < MSE[S] < MSE[S_2] \quad (32)$$

Figure 9 A zonal system consisting of sets of congruent figures.

3.3 Lattice systems

Count data aggregated into a lattice form are now widely used in GIS. It sometimes happens that several lattice data are available differing in the cell size though not having a hierarchical structure (Figure 10). To treat such a case, we analyze the relationship between the accuracy of estimates and the cell size of a lattice.

Figure 10 Lattice systems differing in the cell size.

Suppose two lattice systems L_1 and L_2 whose cells are similar in shape but different in size. The unit cell and its area of L_i are denoted by Z_i and A_i , respectively ($A_1 < A_2$). We assume that Z_0 is still larger than T so that the following equation holds for any \mathbf{x} and \mathbf{t} in both Z_1 and Z_2 .

$$\sum_{\mathbf{u} \in \Omega(\mathbf{t})} m(T; |\mathbf{x} - \mathbf{u}|) = m(T; |\mathbf{x} - \mathbf{t}|). \quad (33)$$

The MSE of L_1 is given by

$$\begin{aligned} MSE[L_1] &= \frac{N}{A_0} \left\{ B - \frac{1}{2\pi A_1^2} \int_{\mathbf{t} \in Z_1} \int_{\mathbf{x} \in Z_1} m(T; |\mathbf{x} - \mathbf{t}|) d\mathbf{x} d\mathbf{t} \right\}, \\ &= \frac{N}{A_0} \left\{ B - \frac{1}{2\pi} \int_{l=0}^{\infty} g_{Z_1}(l) m(T; l) dl \right\} \end{aligned} \quad (34)$$

where $g_{Z_1}(l)$ is the probability density function of the distance between two points which are randomly distributed in Z_1 . Similarly,

$$MSE[L_2] = \frac{N}{A_0} \left\{ B - \frac{1}{2\pi} \int_{l=0}^{\infty} g_{Z_2}(l) m(T; l) dl \right\}. \quad (35)$$

Noticing that Z_1 and Z_2 are similar in shape, we have

$$g_{z_2}(l) = \sqrt{\frac{A_1}{A_2}} g_{z_1}\left(\sqrt{\frac{A_1}{A_2}} l\right). \quad (36)$$

Substitution of equation (36) into (35) yields

$$\begin{aligned} MSE[L_2] &= \frac{N}{A_0} \left\{ B - \frac{1}{2\pi} \sqrt{\frac{A_1}{A_2}} \int_{l=0}^{\infty} g_{z_1}\left(\sqrt{\frac{A_1}{A_2}} l\right) m(T; l) dl \right\} \\ &= \frac{N}{A_0} \left\{ B - \frac{1}{2\pi} \int_{l=0}^{\infty} g_{z_1}(l) m\left(T; \sqrt{\frac{A_2}{A_1}} l\right) dl \right\}. \end{aligned} \quad (37)$$

From equations (34) and (37) we have

$$\begin{aligned} \Delta MSE &= MSE[L_2] - MSE[L_1] \\ &= \frac{N}{2\pi A_0} \int_{l=0}^{\infty} g_{z_1}(l) \left\{ m(T; l) - m\left(T; \sqrt{\frac{A_2}{A_1}} l\right) \right\} dl. \end{aligned} \quad (38)$$

Since $\sqrt{A_2/A_1} > 1$ and $g_{z_1}(l) \geq 0$ for any l , ΔMSE is positive if $m(T; l)$ is a monotonic decreasing function of l . The condition is satisfied at least when the target zone T is convex. The reason for this is as follows. The measure $m(T; l)$ is equal to the measure of the set of all point pairs separated by a distance l in T . If T is convex, $m(T; l)$ is also equal to the measure of the set of all line segments of length l in T . The set of possible locations of a line segment of length $\sqrt{A_2/A_1} l$ includes the possible locations of a line segment of length l (recall $\sqrt{A_2/A_1} > 1$). Therefore, an inequation

$$m(T; l) \geq m\left(T; \sqrt{\frac{A_2}{A_1}} l\right) \quad (39)$$

holds for any l . Inequation (41) indicates that $m(T; l)$ is a monotonic decreasing function of l , and consequently ΔMSE is always positive. From this we can say that it is desirable to choose a lattice system consisting of smaller cells when T is convex.

Unfortunately, theoretical approach appears difficult in case of non-convex T . We thus experimentally computed the MSE for non-convex T such as U-shape, T-shape, and H-shape using numerical calculations, yet ΔMSE was always positive in all the cases. The MSE seems to increase monotonically with the cell size even if T is non-convex, though it still remains unproved.

In this section, we have analytically investigated the relationship between the size of source zones and estimation accuracy. The obtained results strongly suggest that the MSE decreases as source zones become smaller. From this we can say that smaller zones give better estimates on a wide range of zonal systems. Though this has been reported on an empirical basis (say, Fisher and Langford 1995, Cockings *et al.* 1997), it is meaningful to prove it on a theoretical basis because it warrants the global applicability.

4. Numerical examination of estimation accuracy on lattice systems

In this section we analyze the effect not only of the size but of the shape of source and target zones on estimation accuracy. To this end, we numerically examine the accuracy of count data estimated on lattice systems. We focus on the lattice system because it is widely used in GIS. It should be emphasized, however, that any zonal system can be numerically analyzed in a similar way.

We follow the approach taken in the Subsection 3.3, that is, we employ the periodic continuation model, assuming that the points are uniformly distributed and that the region Z_0 is still larger than the target zone T . The MSE of a lattice system L is given by

$$MSE[L] = \frac{NB}{A_0} \left\{ 1 - \frac{B}{2\pi} \int_{l=0}^{\infty} \frac{g_{Z_1}(l)g_T(l)}{l} dl \right\}. \quad (40)$$

Let μ be the expectation of the number of points contained in T , that is,

$$\mu = E[M]. \quad (41)$$

From equations (2) and (A 4) we obtain

$$\begin{aligned} \mu &= E \left[\sum_j C(\mathbf{y}_j) \right] \\ &= \frac{NB}{A_0} \end{aligned} \quad (42)$$

Substituting equation (42) into equation (40), we have

$$MSE[L] = \left\{ 1 - \frac{B}{2\pi} \int_{l=0}^{\infty} \frac{g_{Z_1}(l)g_T(l)}{l} dl \right\} \mu. \quad (43)$$

In this section we fix the values of N , A_0 , and B , in order to focus on the effect of the size and shape of source zones on estimation accuracy. This implies that μ in equation (43) is constant. The area of the target zone T is set to 1.

4.1 Square lattices

We first examine the square lattice, a zonal system consisting of a set of congruent squares. For the shape of the target zone T , we consider the circle, square, and six types of rectangles whose ratios of the vertical to the horizontal length (denoted by *v/h ratios* hereafter) are 2, 3, 4, 8, 16, and 32.

The relationship between $MSE[L]$ and the cell size is illustrated in Figure 11. As shown in the previous section, $MSE[L]$ monotonically increases with the cell size. It increases rapidly at smaller cells while slowly at larger cells.

Figure 11 The relationship between $MSE[L]$ and the cell size of the square lattice.

Numbers in circles indicate the v/h ratios of rectangular target zones.

Let us turn to the effect of the shape of T . Figure 11 shows little difference between $MSE[L]$ of the circle T and that of the square T . However, when T is a rectangle, $MSE[L]$ increases as T becomes elongated. The shape of T clearly affects the accuracy of estimates. Concerning the cell size, we can say that a lattice consisting of small cells is desirable if T has a oblong shape. Suppose, for instance, that we want to keep $MSE[L]$ at 0.5μ . The cell size needs to be smaller than 0.46 if $v/h=4$, while 0.77 if T is a square. When T is a rectangle of $v/h=16$, a lattice of cells smaller than 0.16 is necessary. This implies that, if we want to estimate the count data in a rectangular area of $v/h=16$, the amount of required source data is approximately as five times large as that of the data required for the estimation of a square area. Attention should be paid to the size of source zones when the target zone has a oblong shape, say, the buffer zone of a road.

4.2 Regular lattices

There are three possible regular lattices on a plane, that is, triangular, square, and hexagonal lattices. The strength and weakness of these lattices have been discussed in the literature (Burt, 1980; Star and Estes, 1990; Okabe and Sadahiro, 1997). We compare these lattices from the viewpoint of estimation accuracy. For the shape of T we try the circle, square, and four types of rectangles whose v/h ratios are 4, 8, 16, and 32.

Figure 12 illustrates the relationship between $MSE[L]$ and the cell size of the lattices. Since regular triangles and hexagons are convex, $MSE[L]$ of the lattices monotonically increases as the cells become larger. The relationship between $MSE[L]$ and the cell size on the triangular and hexagonal lattices looks similar to that on the square lattice. Especially the square and hexagonal lattices show quite similar results regardless of the shape of T , though the hexagonal lattice always gives a slightly smaller $MSE[L]$ than the square lattice.

Considering the result and the tractability of spatial database, we can say that the square lattice is practically the best choice among all possible regular lattices. This conclusion is compatible with that reported in Okabe and Sadahiro (1997).

Figure 12 Comparison of the regular lattices. Numbers in circles indicate the v/h ratios of rectangular target zones.

4.3 Rectangular lattices

We finally investigate the rectangular lattice. Though the rectangular lattice is not popular, the analysis will help us to understand the relationship between the accuracy of

estimates and the shape of source zones.

In the analysis we consider five types of lattices consisting of rectangles whose v/h ratios are 2, 3, 4, 8, and 16. For the shape of T , we adopt the circle, square, and six types of rectangles whose v/h ratios are 2, 3, 4, 8, 16, and 32.

Let us examine Figures 13a and 13b, which depict the relationship between $MSE[L]$ and the cell size on the rectangular lattices of $v/h=4$ and 16, respectively. Both figures are similar in general to Figures 11 and 12: the effect of the cell size on $MSE[L]$ decreases as cells becomes large; $MSE[L]$ increases as the shape of T becomes elongated from the circle and the square to the rectangles.

Figure 13 The relationship between $MSE[L]$ and the cell size of the rectangular lattice. The v/h ratios of the rectangular cells are (a) 4, (b) 16. Numbers in circles indicate the v/h ratios of rectangular target zones.

Figure 14 compares rectangular lattices varying the v/h ratio of the lattice cells. Interestingly, the effect of the cell shape is very similar to that of the shape of T : $MSE[L]$ increases as cells become elongated.

Figure 14 Comparison of rectangular lattices. The v/h ratios of the rectangular target zone are (a) 1 (square), (b) 4, (c) 16. Numbers in squares indicate the v/h ratios of rectangular cells.

From the above results, we notice that estimation accuracy is significantly influenced by not only the size but the shape of cells and the target zone. Estimates become inaccurate as the shape of cells and the target zone becomes elongated, which is compatible with the results reported in Cockings *et al.* (1997). Unfortunately, it seems difficult to give a theoretical explanation on this relationship, yet an intuitive discussion based on integral geometry is possible as follows.

Let us consider a lattice L of unit cell Z . We denote the area and perimeter of Z as A and P , respectively. A target zone T of area B and perimeter P_T is dropped randomly on the lattice L .

Suppose that T is now overlaid on L as shown in Figure 15. Estimation error in count data occurs only in the cells intersected by the boundary of T (dark gray cells in Figure 15). Hence it is supposed that estimation error increases with the number of cells intersected by the boundary of T , which is denoted by κ .

Figure 15 A locational relationship between the lattice L and the target zone T .

The expectation of κ is given in an analytical form as below (Santaló, 1976).

$$E[\kappa] = \frac{2\pi(A + B) + P_T P}{2\pi A} \quad (44)$$

Assuming A and B fixed, we find that $E[\kappa]$ increases monotonically with P and P_T . These perimeters are minimized when Z is a regular hexagon and T is a circle, which consequently gives the minimum of $E[\kappa]$. As Z and T become elongated, both P and P_T increase and so does $E[\kappa]$. Therefore, we surmise that an increase of P and P_T makes the estimate inaccurate.

Certainly the above discussion is somewhat vague. However, it helps us to understand intuitively the relationship between the accuracy of estimates and the shape of cells and the target zone.

5. Concluding discussion

In this paper we have analyzed the accuracy of count data transferred from a zonal system to an incompatible zone through the areal weighting interpolation method. We proposed areal weighting interpolation models on a stochastic basis as a framework for handling a variety of point distributions and locations of the target zone. Using an areal weighting interpolation model, we analyzed the relationship between the accuracy of estimates and the size of source zones where points followed the uniform distribution. The results suggest that smaller zones give better estimates on a wide range of zonal systems.

In the previous section we performed numerical examinations of estimation accuracy on lattice systems. We focused on the shape and size of lattice cells and the target zone, and found that not only the size but the shape of cells and the target zone significantly affects the accuracy of estimates. We discussed the effect of the shape in relation to the perimeter and suggested that it may be partly explained from the view of integral geometry.

Analytical and numerical investigations of estimation accuracy assumed that points were randomly distributed in the region, which may be unrealistic in some cases. We should emphasize, however, that this assumption is not essential as seen in the setting of the areal weighting interpolation models. The models are applicable when the points follow a nonuniform distribution.

The areal weighting interpolation models have wide applicability, yet they have some limitations. We finally discuss them for further research. First, the models assume that a target zone is randomly distributed. This assumption does not always hold in spatial analysis. Recall a process of store site selection mentioned in the introductory

section. The process includes repeated areal interpolations which estimate the population within the market area. However, the distribution of proposed sites is usually nonuniform, which does not satisfy the assumption. A model to treat nonuniformly distributed target zones should be developed. Second, the models consider only the areal weighting interpolation method. Though this method is simple and its algorithm runs so fast, it often gives inaccurate estimates as reported by Fisher and Langford (1995). If source data are not available which have a resolution high enough for areal interpolation, intelligent interpolation methods should be attempted. Areal interpolation methods other than the areal weighting interpolation need to be modeled and evaluated.

Appendix A1

The expectation of ε^2 is given by the following equation.

$$\mathbb{E}[\varepsilon^2] = \mathbb{E}[M^2] - 2\mathbb{E}[MM\hat{M}] + \mathbb{E}[\hat{M}^2] \quad (\text{A } 1)$$

We separately derive the three terms in the above equation.

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

Since the points are independently distributed, equation (A 2) becomes

$$\mathbb{E}[M^2] = 2\sum_{j \neq j'} \mathbb{E}[C(\mathbf{y}_j)]\mathbb{E}[C(\mathbf{y}_{j'})] + \sum_j \mathbb{E}[C(\mathbf{y}_j)]. \quad (\text{A } 3)$$

Using

$$\begin{aligned} \mathbb{E}[C(\mathbf{y}_i)] &= \Pr[\mathbf{y}_i \in T] \\ &= \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) \Pr[\mathbf{x} \in T] d\mathbf{x} \end{aligned} \quad (\text{A } 4)$$

we have

$$\begin{aligned} \mathbb{E}[M^2] &= N(N-1) \left\{ \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) \Pr[\mathbf{x} \in T] d\mathbf{x} \right\}^2 + N \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) \Pr[\mathbf{x} \in T] d\mathbf{x} \\ &= N(N-1) \int_{\mathbf{t} \in Z_0} \int_{\mathbf{x} \in Z_0} f(\mathbf{x})f(\mathbf{t}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x}d\mathbf{t} + N \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) \Pr[\mathbf{x} \in T] d\mathbf{x} \end{aligned} \quad (\text{A } 5)$$

Let us proceed to the second term of equation (A 1).

$$\begin{aligned} \mathbb{E}[MM\hat{M}] &= \mathbb{E}\left[\left\{\sum_j C(\mathbf{y}_j)\right\} \left\{\sum_i \frac{\int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x}}{A_i} \sum_j U_i(\mathbf{y}_j)\right\}\right] \\ &= \sum_i \sum_j \sum_{j'} \frac{1}{A_i} \mathbb{E}\left[U_i(\mathbf{y}_j)C(\mathbf{y}_{j'}) \int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x}\right] \\ &= 2\sum_i \sum_{j \neq j'} \frac{1}{A_i} \mathbb{E}[U_i(\mathbf{y}_j)] \mathbb{E}\left[\int_{\mathbf{x} \in Z_i} C(\mathbf{y}_{j'})C(\mathbf{x}) d\mathbf{x}\right] \\ &\quad + \sum_i \sum_j \frac{1}{A_i} \int_{\mathbf{x} \in Z_i} \mathbb{E}[U_i(\mathbf{y}_j)C(\mathbf{y}_j)C(\mathbf{x})] d\mathbf{x} \end{aligned} \quad (\text{A } 6)$$

Substitution of

$$\begin{aligned} \mathbb{E}[U_i(\mathbf{y}_j)] &= \Pr[\mathbf{y}_j \in Z_i] \\ &= \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (\text{A } 7)$$

and

$$\begin{aligned}
\mathbb{E}\left[\int_{\mathbf{x} \in Z_i} C(\mathbf{y}_{j'}) C(\mathbf{x}) d\mathbf{x}\right] &= \int_{\mathbf{x} \in Z_i} \mathbb{E}[C(\mathbf{y}_{j'}) C(\mathbf{x})] d\mathbf{x} \\
&= \int_{\mathbf{x} \in Z_i} \int_{\mathbf{t} \in Z_0} \Pr[(\mathbf{y}_{j'} \in d\mathbf{t}) \cap (d\mathbf{t} \in T) \cap (\mathbf{x} \in T)] d\mathbf{t} d\mathbf{x} \\
&= \int_{\mathbf{x} \in Z_i} \int_{\mathbf{t} \in Z_0} \Pr[\mathbf{y}_{j'} \in d\mathbf{t}] \Pr[d\mathbf{t} \cup \mathbf{x} \in T] d\mathbf{t} d\mathbf{x} \\
&= \int_{\mathbf{x} \in Z_i} \int_{\mathbf{t} \in Z_0} f(\mathbf{t}) \Pr[\mathbf{t} \cup \mathbf{x} \in T] d\mathbf{t} d\mathbf{x}
\end{aligned} \tag{A 8}$$

yields

$$\begin{aligned}
\mathbb{E}[MM\hat{M}] &= N(N-1) \sum_i \frac{1}{A_i} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_i} \int_{\mathbf{t} \in Z_0} f(\mathbf{t}) \Pr[\mathbf{t} \cup \mathbf{x} \in T] d\mathbf{t} d\mathbf{x} \\
&\quad + \sum_i \sum_j \frac{1}{A_i} \int_{\mathbf{x} \in Z_i} \mathbb{E}[U_i(\mathbf{y}_j) C(\mathbf{y}_j) C(\mathbf{x})] d\mathbf{x}
\end{aligned} \tag{A 9}$$

The second term of the above equation becomes

$$\begin{aligned}
\mathbb{E}[U_i(\mathbf{y}_j) C(\mathbf{y}_j) C(\mathbf{x})] &= \Pr[(\mathbf{y}_j \in Z_i \cap T) \cap (\mathbf{x} \in T)] \\
&= \int_{\mathbf{t} \in Z_i} \Pr[(\mathbf{y}_j \in d\mathbf{t}) \cap (d\mathbf{t} \in T) \cap (\mathbf{x} \in T)] d\mathbf{t} \\
&= \int_{\mathbf{t} \in Z_i} f(\mathbf{t}) \Pr[\mathbf{t} \cup \mathbf{x} \in T] d\mathbf{t}
\end{aligned} \tag{A 10}$$

Thus we have

$$\begin{aligned}
\mathbb{E}[MM\hat{M}] &= N(N-1) \sum_i \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_i} \int_{\mathbf{t} \in Z_0} f(\mathbf{t}) \Pr[\mathbf{t} \cup \mathbf{x} \in T] d\mathbf{t} d\mathbf{x} \\
&\quad + N \sum_i \frac{1}{A_i} \int_{\mathbf{x} \in Z_i} \int_{\mathbf{t} \in Z_i} f(\mathbf{t}) \Pr[\mathbf{t} \cup \mathbf{x} \in T] d\mathbf{t} d\mathbf{x}
\end{aligned} \tag{A 11}$$

The third term of equation (A 1) is written as

$$\begin{aligned}
\mathbb{E}[\hat{M}^2] &= \mathbb{E}\left[\left\{\sum_i \frac{\int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x}}{A_i} \sum_j U_i(\mathbf{y}_j)\right\}^2\right] \\
&= \sum_i \sum_{i'} \sum_j \sum_{j'} \mathbb{E}\left[\frac{\int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} C(\mathbf{x}) d\mathbf{x} U_i(\mathbf{y}_j) U_{i'}(\mathbf{y}_{j'})}{A_i A_{i'}}\right] \\
&= \sum_i \sum_{i'} \frac{1}{A_i A_{i'}} \mathbb{E}\left[\int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} C(\mathbf{x}) d\mathbf{x}\right] \sum_j \sum_{j'} \mathbb{E}[U_i(\mathbf{y}_j) U_{i'}(\mathbf{y}_{j'})] \\
&= 2 \sum_i \sum_{i'} \frac{1}{A_i A_{i'}} \mathbb{E}\left[\int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} C(\mathbf{x}) d\mathbf{x}\right] \sum_{j \neq j'} \mathbb{E}[U_i(\mathbf{y}_j)] \mathbb{E}[U_{i'}(\mathbf{y}_{j'})] \\
&\quad + \sum_i \sum_{i'} \frac{1}{A_i A_{i'}} \mathbb{E}\left[\int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} C(\mathbf{x}) d\mathbf{x}\right] \sum_j \mathbb{E}[U_i(\mathbf{y}_j) U_{i'}(\mathbf{y}_j)]
\end{aligned} \tag{A 12}$$

Substituting equation (A 7), we have

$$\begin{aligned}
\mathbb{E}[\hat{M}^2] &= N(N-1) \sum_i \sum_{i'} \frac{1}{A_i A_{i'}} \mathbb{E} \left[\int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} C(\mathbf{x}) d\mathbf{x} \right] \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} f(\mathbf{x}) d\mathbf{x} \\
&\quad + \sum_i \sum_{i'} \frac{1}{A_i A_{i'}} \mathbb{E} \left[\int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} C(\mathbf{x}) d\mathbf{x} \right] \sum_j \mathbb{E} [U_i(\mathbf{y}_j) U_{i'}(\mathbf{y}_j)] \\
&= N(N-1) \sum_i \sum_{i'} \frac{1}{A_i A_{i'}} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_{i'}} \mathbb{E} [C(\mathbf{x}) C(\mathbf{t})] d\mathbf{x} d\mathbf{t} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} f(\mathbf{x}) d\mathbf{x} \\
&\quad + \sum_i \frac{1}{A_i^2} \mathbb{E} \left[\int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_i} C(\mathbf{x}) d\mathbf{x} \right] \sum_j \mathbb{E} \left[\{U_i(\mathbf{y}_j)\}^2 \right] \\
&= N(N-1) \sum_i \sum_{i'} \frac{1}{A_i A_{i'}} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_{i'}} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} f(\mathbf{x}) d\mathbf{x} \\
&\quad + N \sum_i \frac{1}{A_i^2} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x}
\end{aligned} \tag{A 13}$$

Using equations (A 5), (A 11), and (A 13), we obtain

$$\begin{aligned}
\mathbb{E}[\mathcal{E}^2] &= N(N-1) \int_{\mathbf{t} \in Z_0} \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) f(\mathbf{t}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \\
&\quad + N \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) \Pr[\mathbf{x} \in T] d\mathbf{x} \\
&\quad - 2N(N-1) \sum_i \frac{1}{A_i} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_0} f(\mathbf{x}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \\
&\quad - 2N \sum_i \frac{1}{A_i} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \\
&\quad + N(N-1) \sum_i \sum_{i'} \frac{1}{A_i A_{i'}} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_{i'}} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x} \int_{\mathbf{x} \in Z_{i'}} f(\mathbf{x}) d\mathbf{x} \\
&\quad + N \sum_i \frac{1}{A_i^2} \int_{\mathbf{t} \in Z_i} \int_{\mathbf{x} \in Z_i} \Pr[\mathbf{x} \cup \mathbf{t} \in T] d\mathbf{x} d\mathbf{t} \int_{\mathbf{x} \in Z_i} f(\mathbf{x}) d\mathbf{x}
\end{aligned} \tag{A 14}$$

Appendix A2

Suppose a region T of area B and two points that are independently and randomly distributed in T . We denote the location of the points by the vectors \mathbf{x} and \mathbf{t} , and consider a function

$$D(\mathbf{x}, \mathbf{t}, l) = \begin{cases} 1 & \text{if } |\mathbf{x} - \mathbf{t}| \leq l \\ 0 & \text{otherwise} \end{cases}. \quad (\text{A } 15)$$

The probability distribution function of the distance between the two points is written as

$$G_T(l) = \frac{1}{B^2} \int_{\mathbf{t} \in T} \int_{\mathbf{x} \in T} D(\mathbf{x}, \mathbf{t}, l) d\mathbf{x} d\mathbf{t}. \quad (\text{A } 16)$$

The function $g_T(l)$, the probability density function of the distance between the points, is given by

$$\begin{aligned} g_T(l) &= \frac{d}{dl} G_T(l) \\ &= \lim_{dl \rightarrow 0} \left\{ \frac{G_T(l + dl) - G_T(l)}{dl} \right\} \\ &= \frac{1}{B^2} \lim_{dl \rightarrow 0} \int_{\mathbf{t} \in T} \int_{\mathbf{x} \in T} \frac{D(\mathbf{x}, \mathbf{t}, l + dl) - D(\mathbf{x}, \mathbf{t}, l)}{dl} d\mathbf{x} d\mathbf{t} \end{aligned}. \quad (\text{A } 17)$$

We then turn to the measure $m(T; l)$, which is equivalent to the measure of the set of all point pairs separated by a distance l in T . To derive $m(T; l)$, we first fix one of the points at \mathbf{x} and examine possible locations of the other.

Obviously, the location of the movable point is restricted on the arc Γ , which is given as the union of a circle of radius l centered at \mathbf{x} and the region T (Figure A1). Thus we have

$$\lim_{dl \rightarrow 0} \frac{\theta(\mathbf{x}, l)}{2\pi} \left\{ \pi(l + dl)^2 - \pi l^2 \right\} = \lim_{dl \rightarrow 0} \int_{\mathbf{t} \in T} \frac{D(\mathbf{x}, \mathbf{t}, l + dl) - D(\mathbf{x}, \mathbf{t}, l)}{dl} d\mathbf{t}, \quad (\text{A } 18)$$

where $\theta(\mathbf{x}, l)$ is the included angle of Γ . The left side of equation (A 18) is rewritten as

$$\begin{aligned} \lim_{dl \rightarrow 0} \frac{\theta(\mathbf{x}, l)}{2\pi} \left\{ \pi(l + dl)^2 - \pi l^2 \right\} &= \lim_{dl \rightarrow 0} \frac{\theta(\mathbf{x}, l)}{2} \left\{ 2ldl + dl^2 \right\} \\ &\approx l\theta(\mathbf{x}, l) \end{aligned}. \quad (\text{A } 19)$$

Substitution of equation (A 19) into equation (A 18) yields

$$\theta(\mathbf{x}, l) = \frac{1}{l} \lim_{dl \rightarrow 0} \int_{\mathbf{t} \in T} \frac{D(\mathbf{x}, \mathbf{t}, l + dl) - D(\mathbf{x}, \mathbf{t}, l)}{dl} d\mathbf{t}. \quad (\text{A } 20)$$

The measure $m(T; l)$ is given by

$$m(T; l) = \int_{\mathbf{x} \in T} \theta(\mathbf{x}, l) d\mathbf{x}. \quad (\text{A } 21)$$

Substitution of equation (A 20) into equation (A 21) yields

$$\begin{aligned}
m(T;l) &= \int_{\mathbf{x} \in T} \theta(\mathbf{x}, l) d\mathbf{x} \\
&= \int_{\mathbf{x} \in T} \frac{1}{l} \lim_{dl \rightarrow 0} \int_{\mathbf{t} \in T} \frac{D(\mathbf{x}, \mathbf{t}, l + dl) - D(\mathbf{x}, \mathbf{t}, l)}{dl} d\mathbf{t} d\mathbf{x}. \tag{A 22} \\
&= \frac{1}{l} \lim_{dl \rightarrow 0} \int_{\mathbf{x} \in T} \int_{\mathbf{t} \in T} \frac{D(\mathbf{x}, \mathbf{t}, l + dl) - D(\mathbf{x}, \mathbf{t}, l)}{dl} d\mathbf{t} d\mathbf{x}
\end{aligned}$$

From equations (A 17) and (A 22) we have

$$\begin{aligned}
m(T;l) &= \frac{1}{l} \lim_{dl \rightarrow 0} \int_{\mathbf{t} \in T} \int_{\mathbf{x} \in T} \frac{D(\mathbf{x}, \mathbf{t}, l + dl) - D(\mathbf{x}, \mathbf{t}, l)}{dl} d\mathbf{x} d\mathbf{t} \\
&= \frac{B^2}{l} g_T(l) . \tag{A 23}
\end{aligned}$$

Figure A1 The arc Γ defined by the union of a circle of radius l centered at \mathbf{x} and T .

References

- Burrough, P. A. and McDonnell, R. A., 1998, *Principles of Geographical Information Systems* (New York: Oxford University Press).
- Burt, P. J., 1980, Tree and Pyramid Structures for Coding Hexagonally Samples Binary Images. *Computer Graphics and Image Processing*, **14**, 271-280.
- Cockings, S., Fisher, P. F., and Langford, M., 1997, Parameterization and Visualization of the Errors in Areal Interpolation. *Geographical Analysis*, **29**, 314-328.
- Fisher, P. F. and Langford, M., 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 Green, M., 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 (New York: Longman), 38-54.
- Goodchild, M. F. and Lam, N. N-S., 1980, Areal Interpolation: a Variant of the Traditional Spatial Problem. *Geo-processing*, **1**, 297-312.
- Goodchild, M. F., Anselin, L., and Deichmann, U., 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., Maguire, D. J., and Unwin, D. J., 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 (New York: Longman), 55-77.
- Markoff, J. and Shapiro, G. 1973, The Linkage of Data Describing Overlapping Geographical Units. *Historical Methods Newsletter*, **7**, 34-46.
- Okabe, A. and Sadahiro, Y., 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.
- 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 (New York: Longman), 127-137.
- Ripley, B. D., 1981, *Spatial Statistics* (New York: John Wiley).

- Sadahiro, Y., 1999, Statistical Methods for Analyzing the Distribution of Spatial Objects in Relation to a Surface. *Geographical Systems*, to appear.
- Santaló, L. A., 1976, *Integral Geometry and Geometric Probability* (London: Addison-Wesley).
- Star, J. and Estes, J., 1990, *Geographic Information Systems - An Introduction -* (Englewood Cliffs: Prentice-Hall).
- Stoyan, D. and Stoyan, H., 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.

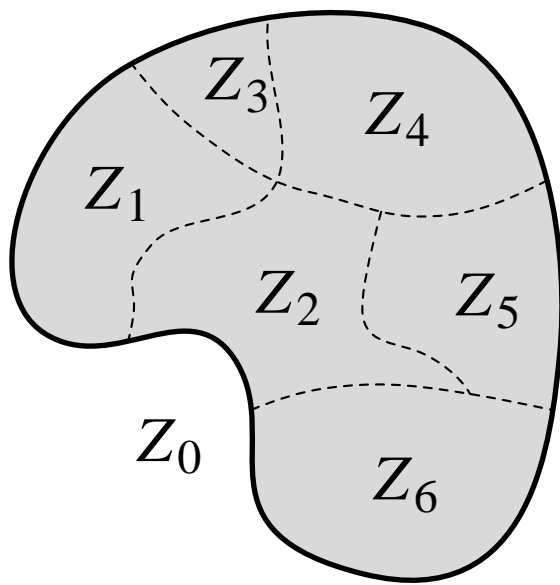


Figure 1

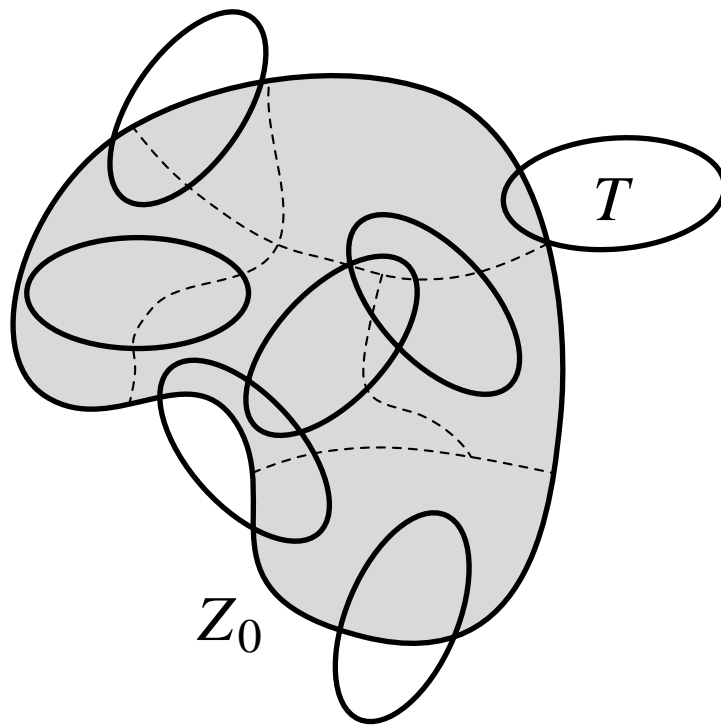


Figure 2

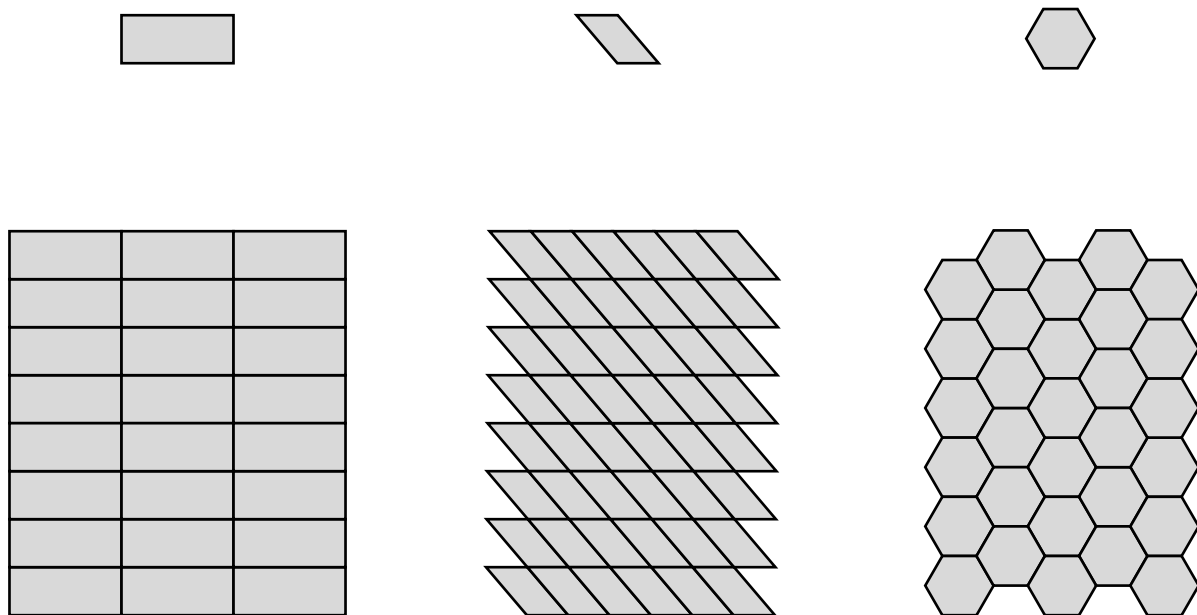


Figure 3

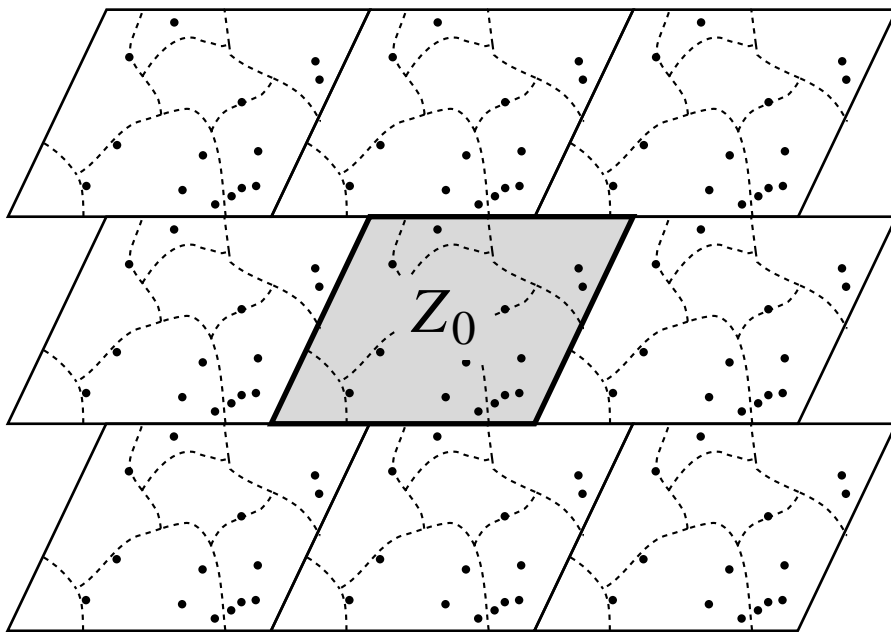


Figure 4

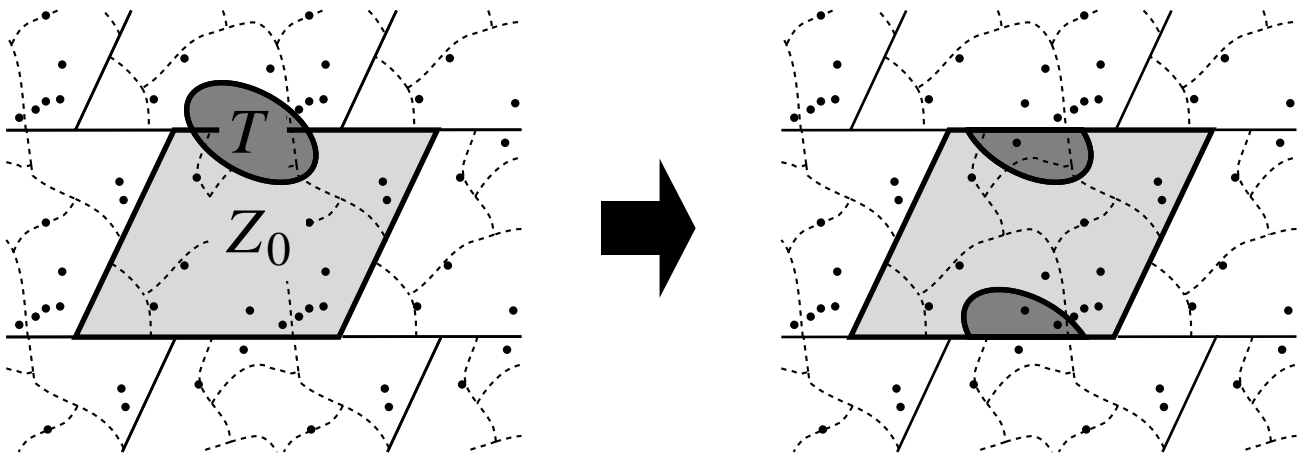


Figure 5

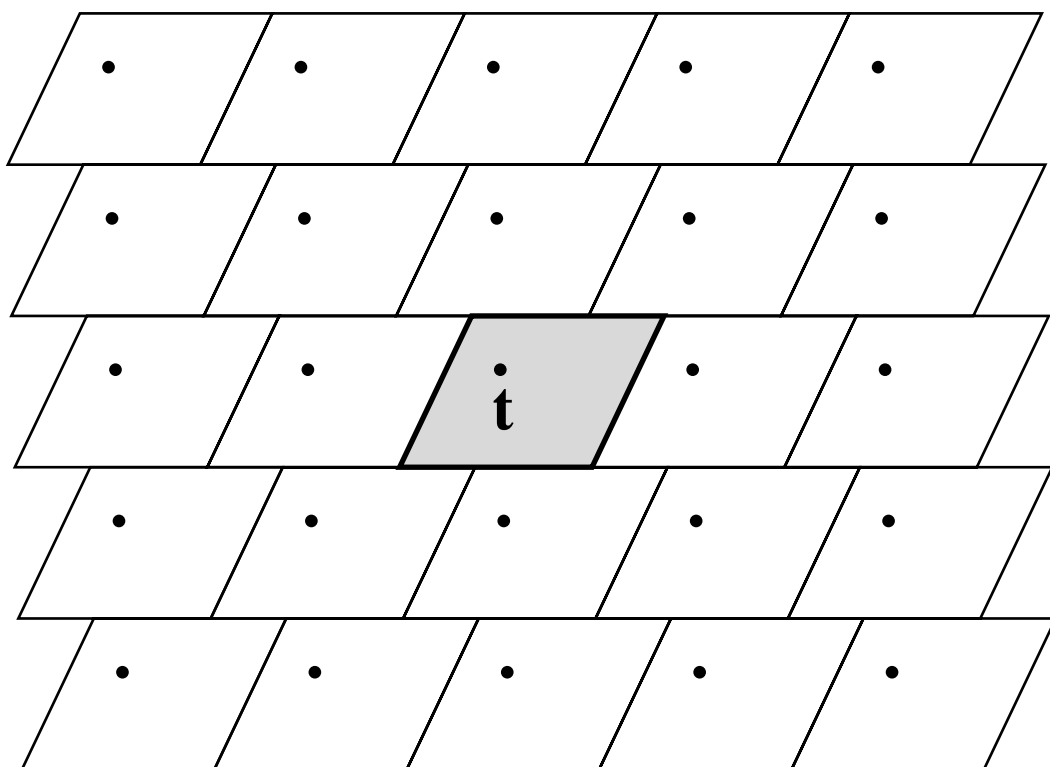


Figure 6

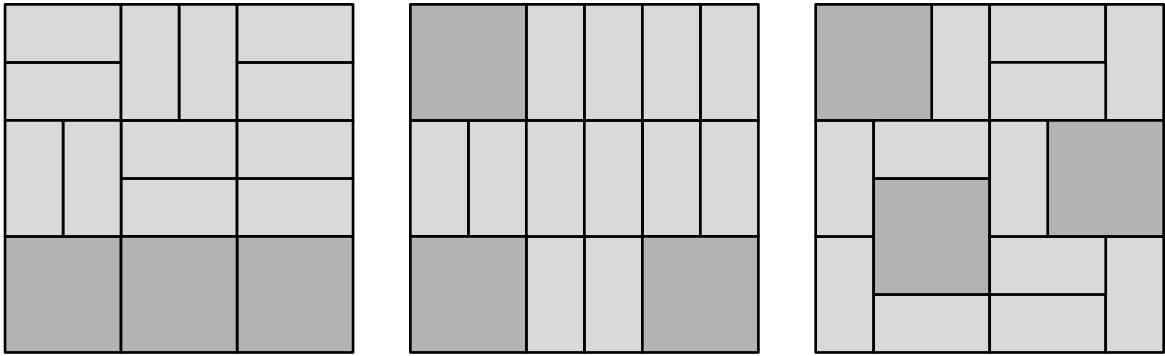
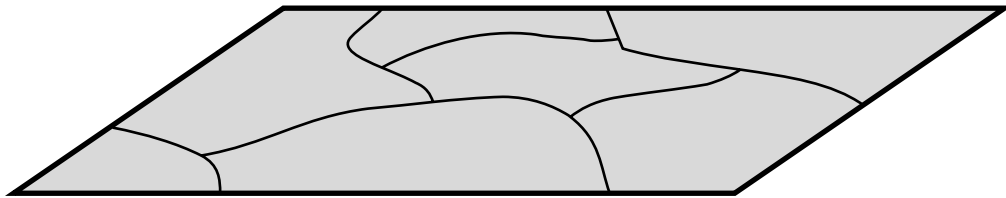


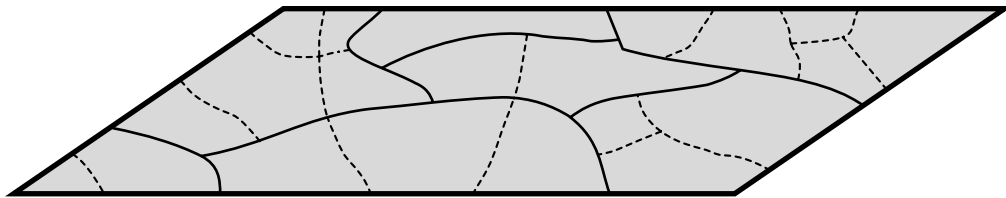
Figure 7



1st level
(the highest level)

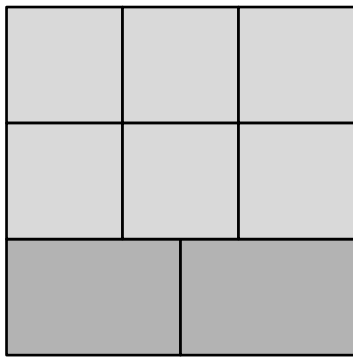


2nd level



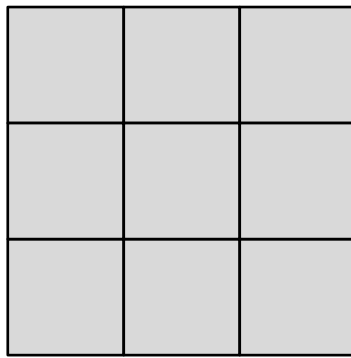
3rd level
(the lowest level)

Figure 8



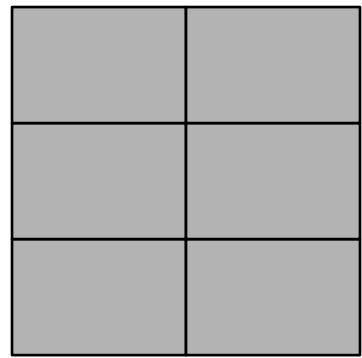
$$MSE[S] = \frac{2}{3}MSE[S_1] + \frac{1}{3}MSE[S_2]$$

(a)



$$MSE[S_1]$$

(b)



$$MSE[S_2]$$

(c)

Figure 9

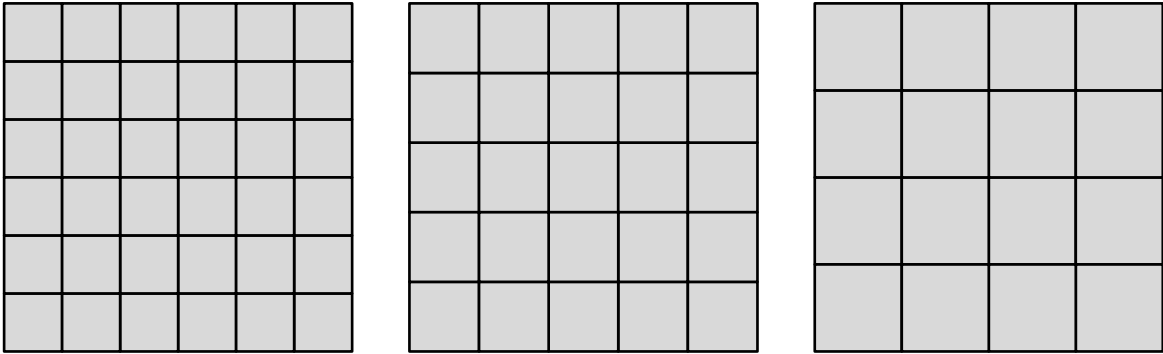


Figure 10

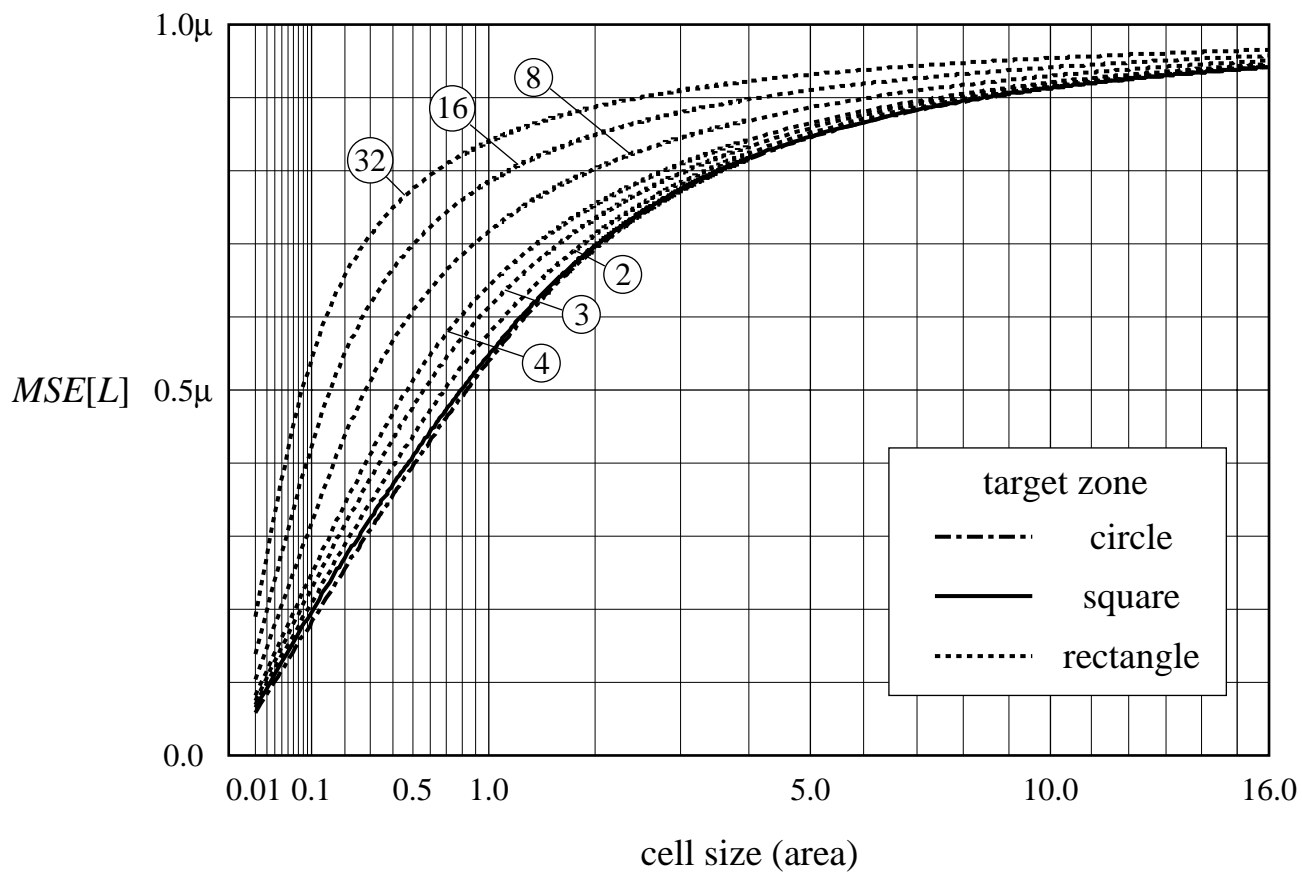


Figure 11

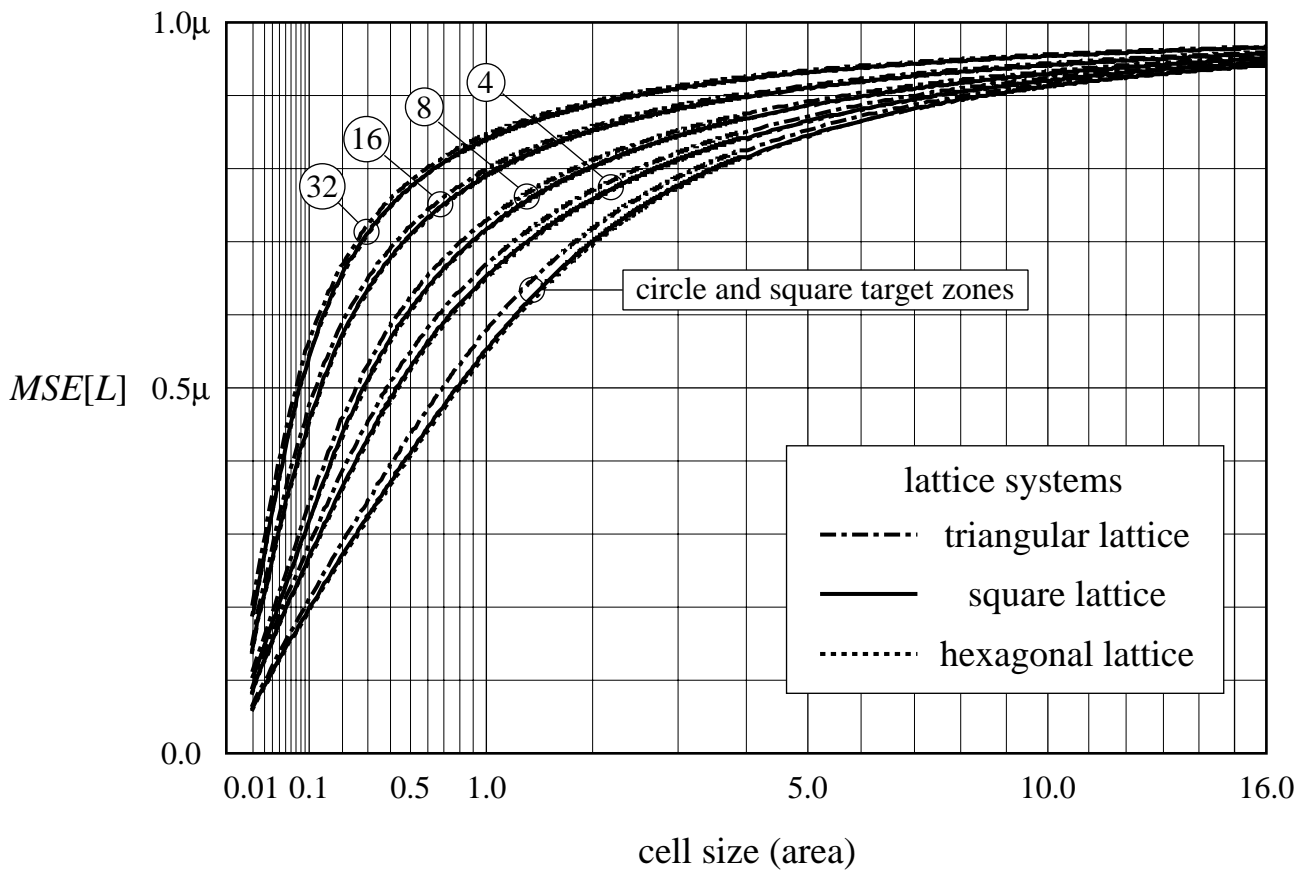


Figure 12

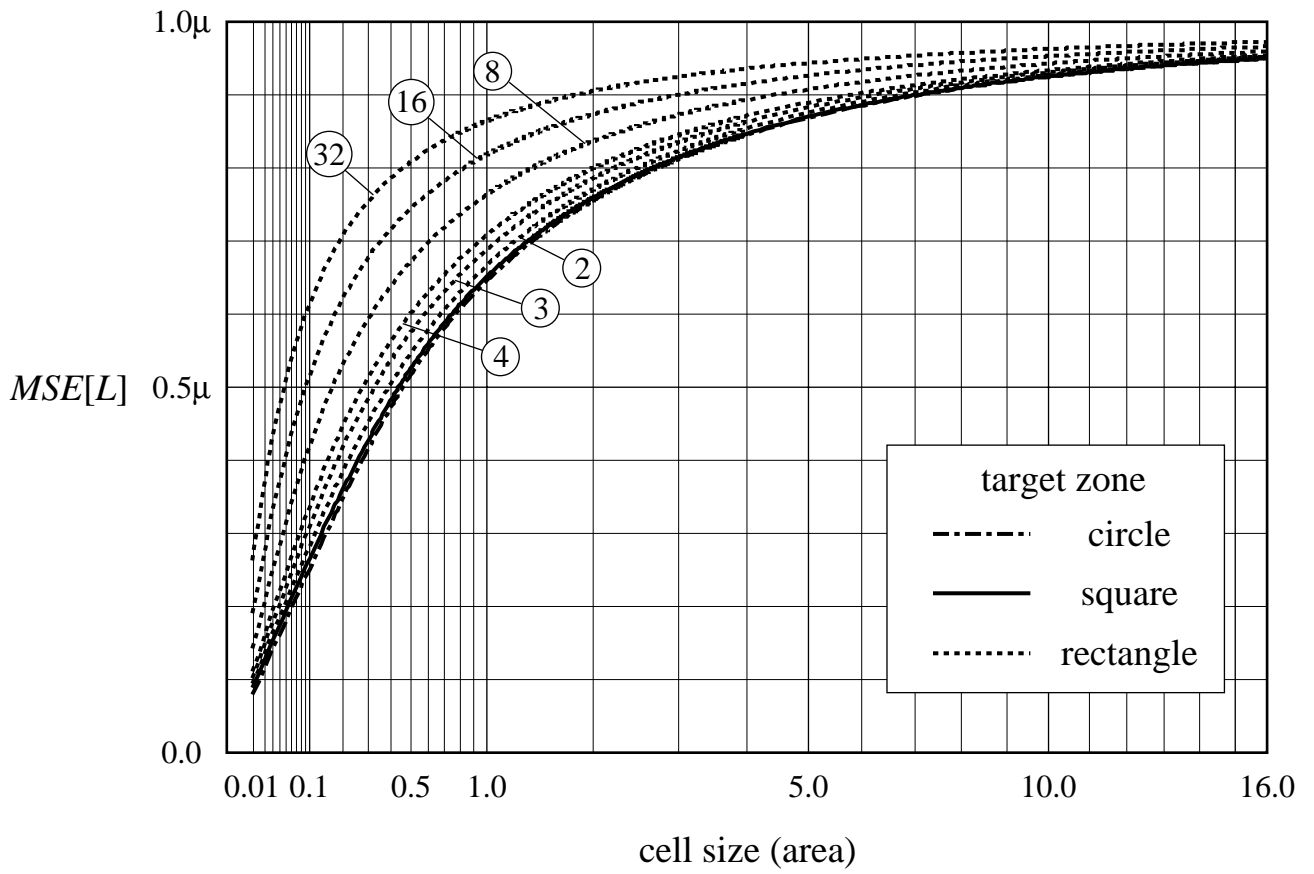


Figure 13a

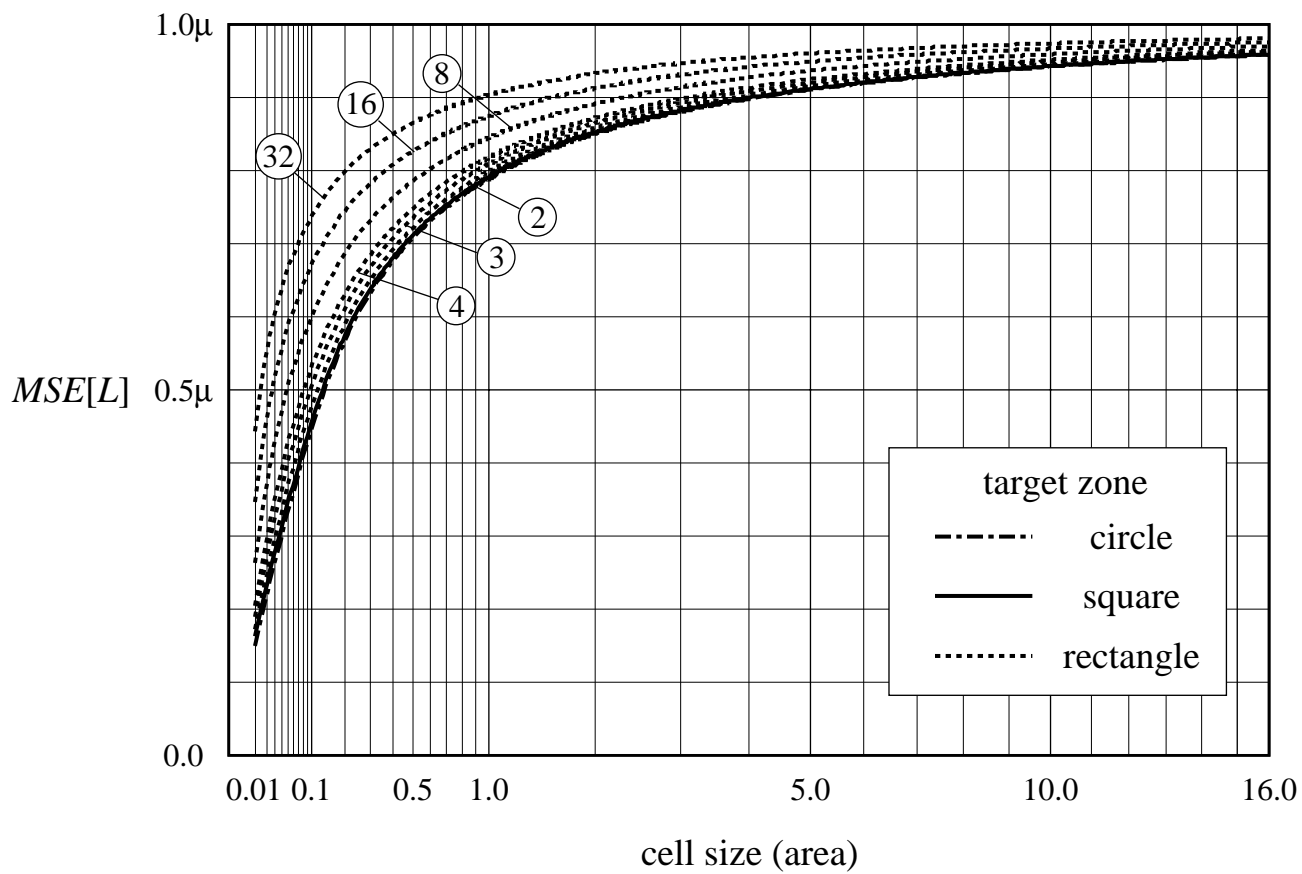


Figure 13b

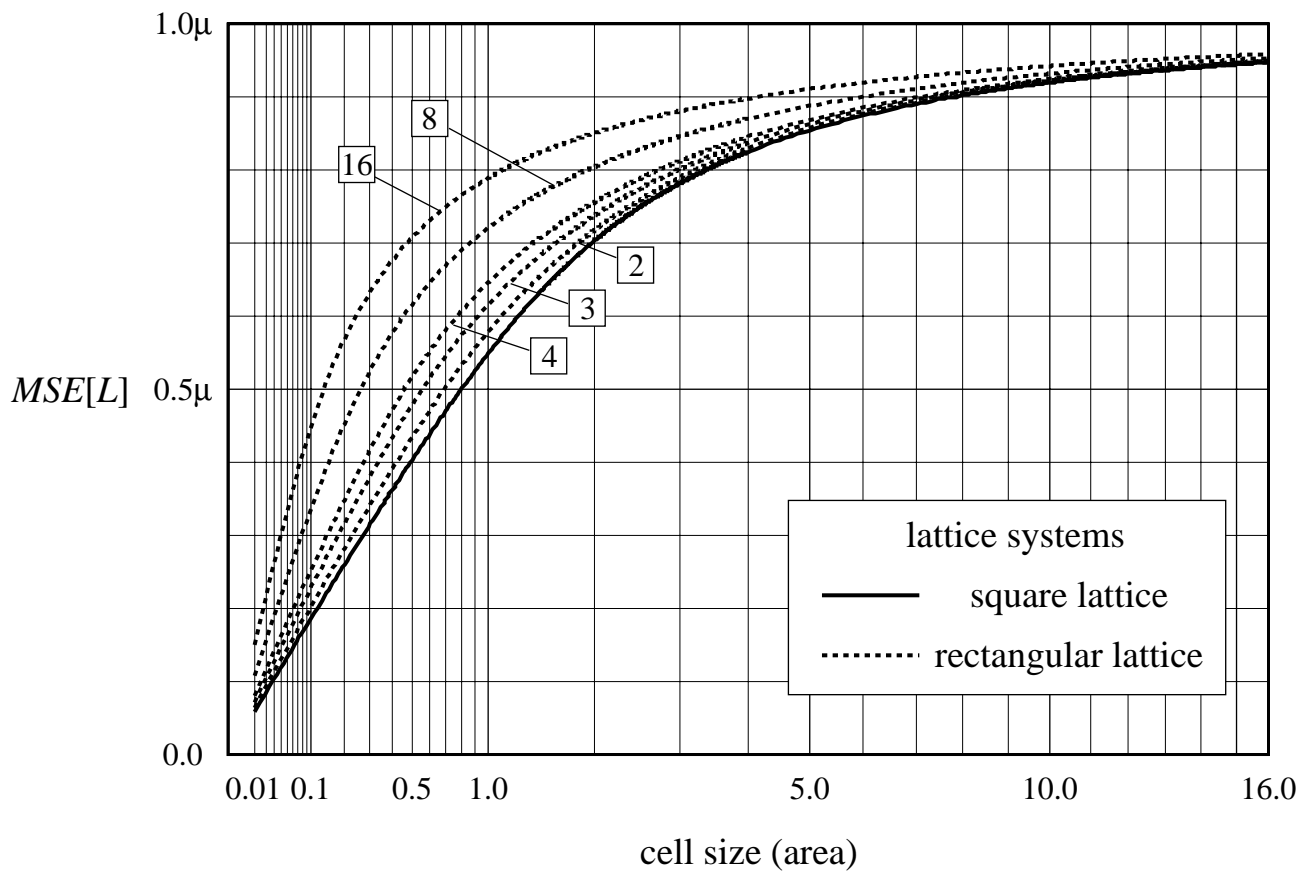


Figure 14a

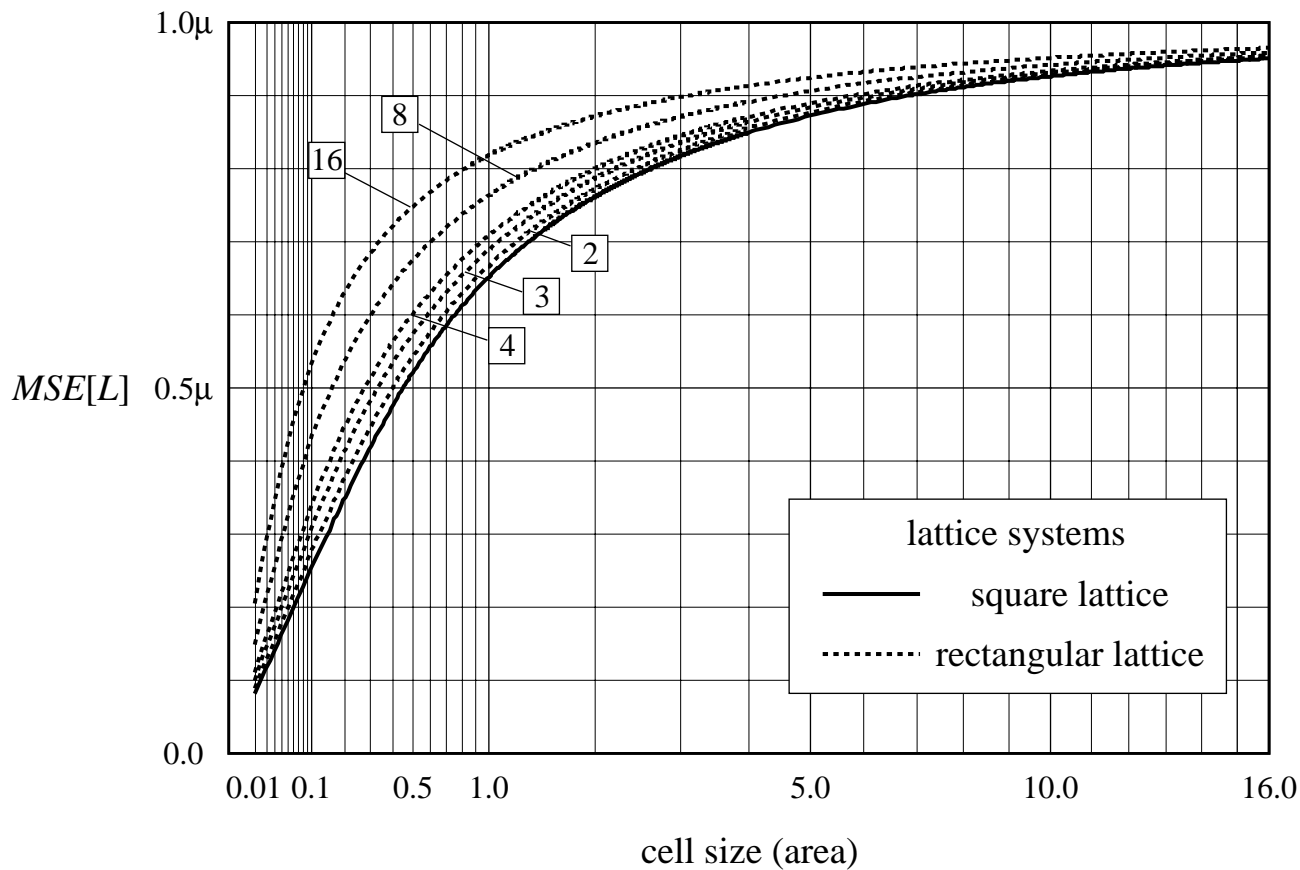


Figure 14b

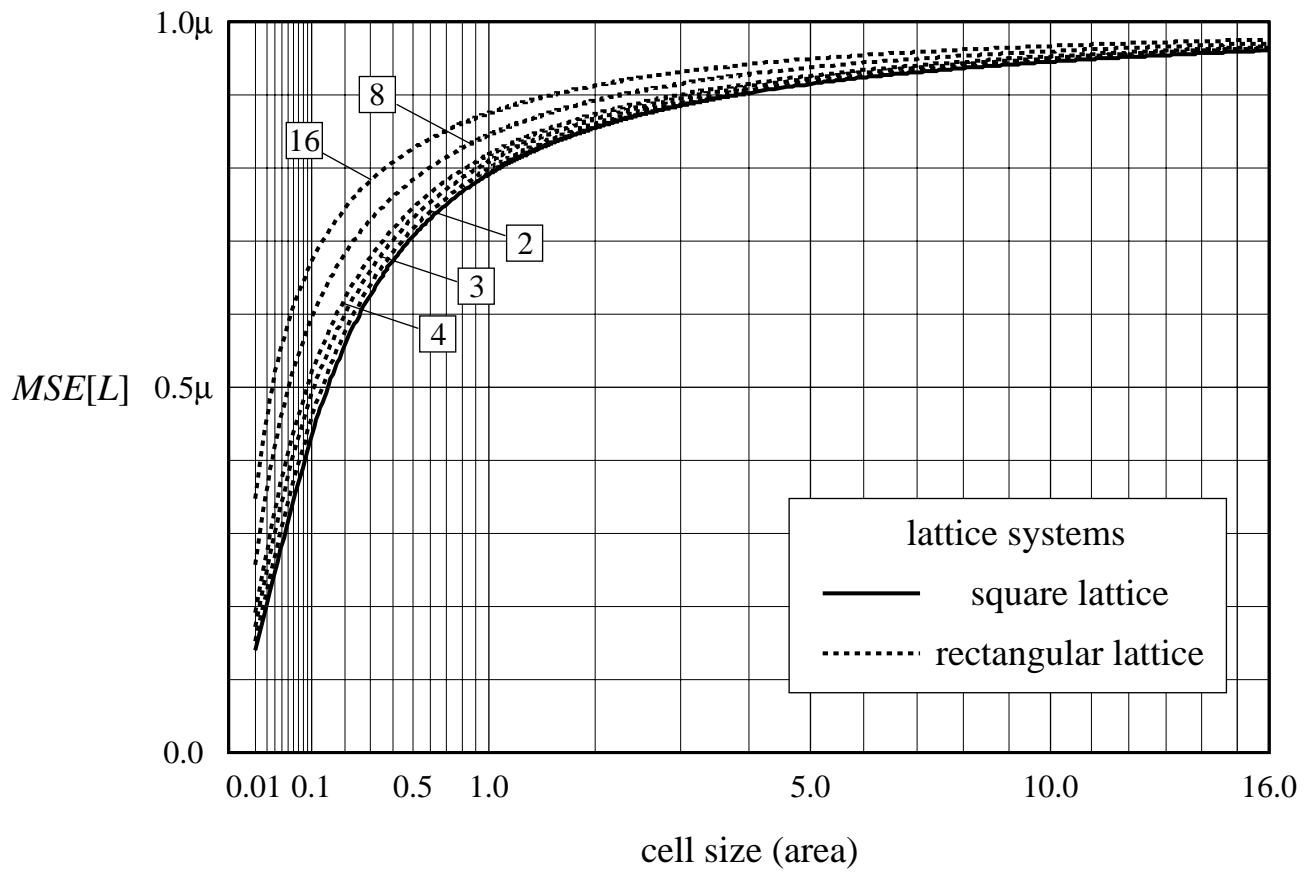


Figure 14c

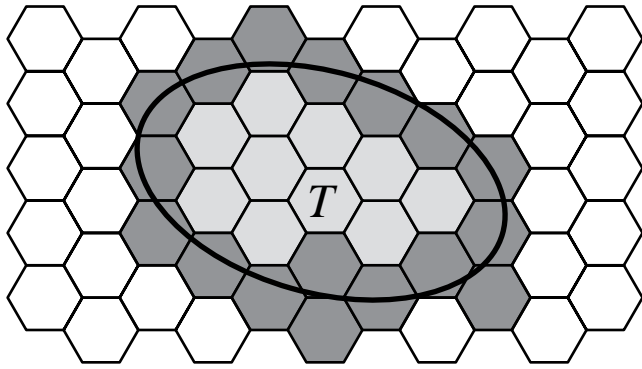


Figure 15

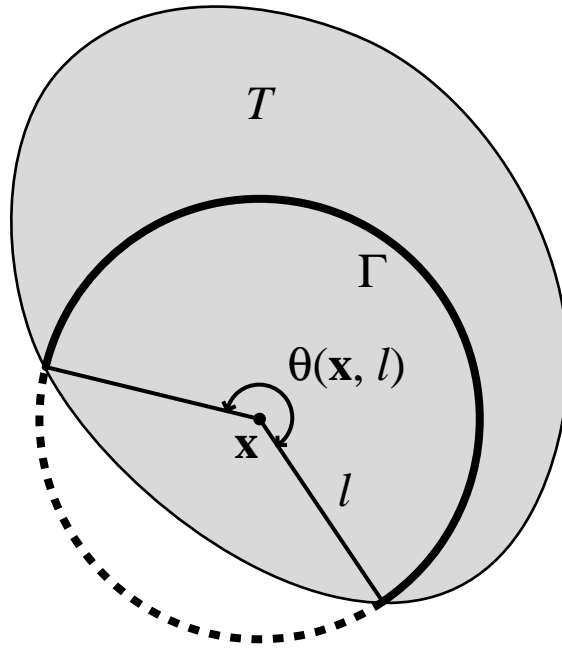


Figure A1