CSIS Discussion Paper No. 9

Accuracy of Areal Interpolation:
A Comparison of Alternative Methods

Yukio Sadahiro

FEBRUARY, 1999

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

February 12, 1999

Accuracy of Areal Interpolation: A Comparison of Alternative Methods

Yukio Sadahiro

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

Phone: +81-3-3812-2111 (ext. 6273)
Fax: +81-3-5800-6965
E-mail: sada@ua.t.u-tokyo.ac.jp

Accuracy of Areal Interpolation: A Comparison of Alternative Methods

Abstract

This paper discusses the accuracy of spatial data estimated by areal interpolation. A stochastic model is proposed which represents areal interpolations in diverse geographic situations. The model is used to examine the relationship between estimation accuracy and the spatial distribution of estimation error from a theoretical viewpoint. The analysis shows that the uniformity in error distribution improves the accuracy of areal interpolation. Four areal interpolation methods are then assessed through numerical examinations. From this it is found that the accuracy of simple interpolation methods heavily depends on the appropriateness of their hypothetical distributions, whereas the accuracy of intelligent methods depends on the fitness of the range of supplementary data for that of true distribution.

Keywords: areal interpolation; accuracy; error distribution; stochastic model.

1. INTRODUCTION

Spatial data, especially socioeconomic data, are often provided in an aggregated form though they are originally disaggregated. Census data, for instance, are aggregated across census tracts in order to keep confidentiality of subjects. Landuse data are sometimes aggregated across administrative units or square lattice for the reduction of data volume. Since there are a wide variety of zonal systems used for data aggregation, it is a basic function of GIS to integrate multiple spatial databases having incompatible zonal systems into one database on a common zonal system.

Integration of spatial data requires a data transfer from one zonal system to another. This process is called areal interpolation, which is inevitably uncertain to some extent because it involves data estimation based on arbitrary assumptions on the distribution of spatial objects. Hence, in order to improve estimation accuracy, numerous areal interpolation methods have been proposed in the literature (Wright, 1936; Markoff and Shapiro, 1973; Tobler, 1979; Goodchild and Lam, 1980; Lam, 1983; Flowerdew and Green, 1991; Rhind, 1991; Goodchild *et al*., 1993; Burrough and McDonnell, 1998).

Along with the development of new methods, there has arisen a need for comparison of estimation accuracy between interpolation methods. To meet this demand, geographers have recently compared the accuracy of areal interpolation methods. Langford *et al*. (1991) investigated the accuracy of three areal interpolation methods. They estimated the population distribution of northern Leicestershire by regression models where landuse data were used as independent variables. Goodchild *et al*. (1993) proposed an interpolation method using ancillary data called "control zones" in which population density was reasonably assumed to be constant. They employed seven interpolation methods including their new method to estimate the population of counties in California and compared the accuracy of estimates.

Most of comparative studies including the above two papers have been performed on a single geographic situation. Hence it is questionable whether the obtained results have global applicability (Fisher and Langford 1995). To solve the problem, Fisher, Langford, and their colleagues have advocated the use of Monte Carlo simulation (Langford *et al*., 1993; Fisher and Langford, 1995, 1996; Cockings *et al*., 1997). Monte Carlo simulation allows us to test interpolation methods in a variety of geographic circumstances so that the results have wider applicability. Fisher and Langford (1995), for instance, discussed estimation accuracy of five interpolation methods considering diverse combinations of source and target zonal systems. They successfully obtained full distributions of estimation errors which were valuable information to assess the interpolation methods.

Unfortunately, Monte Carlo simulation is computationally expensive. The computational cost is a serious problem especially in its application to areal interpolation, because it involves the polygon overlay operation which requires rather complicated algorithms in GIS. Realization of a number of spatial relationships between source and target zones requires considerable calculation. Furthermore, estimation accuracy depends not only on the relationship between source and target zones but also on the distribution of spatial objects, say, population distribution. Analysis of its effect imposes further complexity on computation and consequently makes Monte Carlo simulation impractical.

Motivated from the above discussion, we employ a theory-based approach to compare estimation accuracy among areal interpolation methods. Our approach is based on a stochastic model that represents diverse geographic situations, and the model requires far less computational cost than Monte Carlo simulation. In addition to this advantage, it allows us to discuss theoretically the relationship between estimation accuracy and areal interpolation methods.

In the following section we outline the process of areal interpolation and propose a method for analyzing its accuracy. In Section 3 we discuss the relationship between the accuracy of areal interpolation and the spatial distribution of estimation errors from a theoretical point of view. In Section 4 we numerically examine the accuracy of four areal interpolation methods. Finally we summarize the conclusions obtained in this paper with discussion.

2. AREAL INTERPOLATION MODEL

Areal interpolation is a process of transferring data from one zonal system to another. Let us consider, for instance, a zonal system shown in Figure 1a. Point objects such as individual people or households are distributed over the zones, and the number of point objects is recorded in each zone. A target zone is overlaid on the zones (Figure 1b) and we want to know the number of points contained in the target zone which is directly unobservable. In such a case we employ areal interpolation to estimate the number of points from the source data.

Figure 1. Areal interpolation. a) A zonal system, b) an overlaid target zone. The number in parenthesis indicates the number of points in each zone.

Areal interpolation can be mathematically represented as follows. In this paper we focus on the estimation of count data in the union of a source zone S and a target zone T. Let $f(\mathbf{x})$ be the *density function* of point objects in S. In usual areal interpolation methods an estimator function of $f(\mathbf{x})$ is implicitly employed. For instance, the areal weighting method assumes a constant density in S, that is, it assumes $f(\mathbf{x})$ to be a uniform distribution. Pycnophylactic interpolation method uses another smooth continuous surface (Tobler, 1979). Hence we denote an estimator of *f*(**x**) as $\hat{f}(\mathbf{x})$ which we call *estimator function* (Figure 2). Note that the volume-preserving principle (Lam, 1983) restricts $\hat{f}(\mathbf{x})$ to satisfy

$$\int_{\mathbf{x}\in S}\hat{f}(\mathbf{x})\mathrm{d}\mathbf{x}=\int_{\mathbf{x}\in S}f(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{1}$$

For convenience we denote the volume shown above as *V* hereafter.

Figure 2. A density function *f*(**x**) and its estimator function $\hat{f}(\mathbf{x})$ in one dimensional case.

In the above setting, the number of points in *S*∩*T* and its estimator are given by

$$M=\int_{\mathbf{x}\in S}f(\mathbf{x})1(\mathbf{x})\mathrm{d}\mathbf{x} \tag{2}$$

and

$$\hat{M}=\int_{\mathbf{x}\in S}\hat{f}(\mathbf{x})1(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{3}$$

respectively, where 1(**x**) is the indicator function of *T* defined by

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

To evaluate the accuracy of areal interpolation, we consider a variety of spatial relationships between the source and target zones. More explicitly, we assume that the target zone *T* is dropped randomly in such a way that it intersects *S*, given the density function *f*(**x**), its estimator $\hat{f}(\mathbf{x})$, and the shape and size of *S* and *T* (similar approach was employed in Sadahiro, 1998a, 1999b). We call this stochastic process *areal interpolation model*. The accuracy of interpolation is measured by the mean square error given by

$$\mathrm{E}\left[\left(\hat{M}-M\right)^2\right]=E\left[\left\{\int_{\mathbf{x}\in S}\Delta f(\mathbf{x})1(\mathbf{x})\mathrm{d}\mathbf{x}\right\}^2\right], \tag{5}$$

where

$$\Delta f(\mathbf{x})=\hat{f}(\mathbf{x})-f(\mathbf{x}). \tag{6}$$

We refer the function Δ*f*(**x**) as *error distribution function* since it represents the spatial distribution of estimation errors. After a few steps of calculation (see Sadahiro 1998b for details), we obtain

$$\mathrm{E}\left[\left(\hat{M}-M\right)^2\right]=\frac{1}{m'\left(T;S\cap T\neq\varnothing\right)}\int_{\mathbf{t}\in S}\int_{\mathbf{x}\in S}m\left(T;\left|\mathbf{x}-\mathbf{t}\right|\right)\Delta f(\mathbf{x})\Delta f(\mathbf{t})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}, \tag{7}$$

where *m*(*T*; *S*∩*T*≠∅) and *m*(*T*; *l*) are the measure of the set of all figures congruent to *T* intersecting *S* and that containing two points separated by a distance *l*, respectively (Santaló, 1976). These measures are computable by at least numerical calculations (Sadahiro, 1999a), hence it is possible to compare areal interpolation methods in terms of estimation accuracy using equation (7).

The above equation, however, is somewhat intractable because the computation of *m*(*T*; *l*) often involves complicated geometrical calculations. To obtain a more operational representation of accuracy, we assume that the target zone *T* is large enough that the incircle of *T* is larger than the circumcircle of *S*. If both *S* and *T* have rounded shapes, the size of these zones can be almost equal. This assumption seems realistic in areal interpolation, because *S* and *T* are usually convex in a global scale or *T* is fairly larger than *S* (otherwise areal interpolation may yield considerably imprecise estimates).

If the above assumption holds, we can approximate *m*(*T*; *l*) by

$$m(T;l) \approx 2\pi A_T - 2L_T l, \tag{8}$$

where A_T and L_T are the area and perimeter of *T*, respectively (Stoyan and Stoyan, 1994; Sadahiro, 1998a). Substitution of equation (8) into equation (7) yields

$$\mathrm{E}\left[\left(\hat{M}-M\right)^2\right] = -\frac{2L_T}{m'\left(T;S\cap T\neq\varnothing\right)}\int_{\mathbf{t}\in S}\int_{\mathbf{x}\in S}|\mathbf{x}-\mathbf{t}|\Delta f(\mathbf{x})\Delta f(\mathbf{t})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}. \tag{9}$$

Since we are interested in comparing various areal interpolation methods, we keep the shape and size of *S* and *T* fixed hereafter. Then it is enough to consider

$$I_\varepsilon = -\int_{\mathbf{t}\in S}\int_{\mathbf{x}\in S}|\mathbf{x}-\mathbf{t}|\Delta f(\mathbf{x})\Delta f(\mathbf{t})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \tag{10}$$

for the purpose of assessing the accuracy of areal interpolation. We call this value *error index*, and use it as a measure of estimation accuracy in the following sections.

3. ACCURACY OF AREAL INTERPOLATION AND SPATIAL DISTRIBUTION OF ESTIMATION ERRORS

As seen in equation (10), the accuracy of areal interpolation depends on the error distribution represented by Δ*f*(**x**). To understand this relationship more deeply, we investigate how the form of Δ*f*(**x**) affects the error index I_ε taking some typical examples.

In the following discussion, we focus on a small subregion of *S*∩*T* denoted by *U* (Figure 3). Keeping Δ*f*(**x**) in *S*∩*T*\U fixed, we discuss how the form of Δ*f*(**x**) in *U* affects I_ε. The region *U* is divided into four congruent subregions (not necessarily squares) U_1, U_2, U_3, and U_4, in each of which Δ*f*(**x**) has a constant value (Figure 4).

Figure 3 The region *U* in *S*∩*T*.

Figure 4 An example of the error distribution Δ*f*(**x**) in *U*.

3.1 Degree of Estimation Errors

We start with the two cases shown in Figure 5. As shown in the figure, the absolute value of Δ*f*(**x**) in Case 1b is larger than that in Case 1a by *d* (*d*>0).

Figure 5 The error distribution function Δ*f*(**x**) in *U* (*a*>0, *b*<0).

We wish to examine in which case estimates are more accurate. Though the answer is intuitively obvious, it would be meaningful to discuss this question in a theoretical way.

Let us denote the error index of the Case *i* as $I_\varepsilon(i)$. The index I_ε(1a) is then written as

$$\begin{aligned} I_\varepsilon(1\mathrm{a}) &= -\int_{\mathbf{t}\in S}\int_{\mathbf{x}\in S}|\mathbf{x}-\mathbf{t}|\Delta f(\mathbf{x})\Delta f(\mathbf{t})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ &= -\int_{\mathbf{t}\in S\setminus U}\int_{\mathbf{x}\in S\setminus U}|\mathbf{x}-\mathbf{t}|\Delta f(\mathbf{x})\Delta f(\mathbf{t})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ &\quad -2a(K_3+K_4)-2b(K_1+K_2) \qquad , \\ &\quad -a^2(k_{33}+2k_{34}+k_{44})-b^2(k_{11}+2k_{12}+k_{22}) \\ &\quad -2ab(k_{13}+k_{14}+k_{23}+k_{24}) \end{aligned} \tag{11}$$

where

$$K_i = \int_{\mathbf{t}\in S\setminus U}\int_{\mathbf{x}\in U_i}|\mathbf{x}-\mathbf{t}|\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \tag{12}$$

and

$$k_{ij} = \int_{\mathbf{t}\in U_i}\int_{\mathbf{x}\in U_j}|\mathbf{x}-\mathbf{t}|\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}. \tag{13}$$

The index I_ε(1b) is also given by a similar equation. Then we have

$$\begin{aligned} I_\varepsilon(1\mathrm{b})-I_\varepsilon(1\mathrm{a}) &= 2dK_1+2dK_2-2dK_3-2dK_4 \\ &\quad +4d(b-a)(k_{11}+k_{12}) \\ &\quad -4d(b-a-d)(k_{13}+k_{14}) \qquad . \\ &\approx 4d\{-(a-b)(k_{11}+k_{12})+(a-b+d)(k_{13}+k_{14})\} \\ &= 4d\{(a-b)(k_{14}-k_{11})+d(k_{13}+k_{14})\} > 0 \end{aligned} \tag{14}$$

From equation (14) we obtain

$$I_\varepsilon(1\mathrm{b}) > I_\varepsilon(1\mathrm{a}). \tag{15}$$

This confirms that estimation accuracy increases as the absolute value of the error distribution function $\Delta f(\mathbf{x})$ decreases.

3.2 Uniformity in Error Distribution

We then proceed to the second example shown in Figure 6. Though the sum of $\Delta f(\mathbf{x})$ in *U* is equal in both cases, its spatial allocation is different: in Case 2a the error is uniformly distributed whereas in Case 2b the error distribution has spatial variation. Using this example we investigate whether the uniformity of $\Delta f(\mathbf{x})$ improves estimation accuracy.

Figure 6 The error distribution function $\Delta f(\mathbf{x})$ in *U*.

The difference between I_ε(2a) and I_ε(2b) is given by

$$\begin{aligned} I_\varepsilon(2\mathrm{b}) - I_\varepsilon(2\mathrm{a}) &= 2dK_1 + 2dK_2 - 2dK_3 - 2dK_4 \\ &\quad +4d^2\left(k_{14} - k_{11}\right) \qquad . \\ &\approx 4d^2\left(k_{14} - k_{11}\right) > 0 \end{aligned} \tag{16}$$

From equation (16) we have

$$I_\varepsilon(2\mathrm{b}) > I_\varepsilon(2\mathrm{a}). \tag{17}$$

Inequality (17) implies that the spatial variation of estimation error increases the error index I_ε. This leads to a conclusion that the uniformity of $\Delta f(\mathbf{x})$ improves the accuracy of areal interpolation.

3.3 Scale of Uniformity in Error Distribution

The analysis of Case 2 suggests that the uniformity is a desirable property of $\Delta f(\mathbf{x})$. The uniformity, however, can be obtained in a variety of scales, from global to local. Consequently, we next examine which uniformity in $\Delta f(\mathbf{x})$ is more desirable, global (Case 3a) or local (Case 3b).

Figure 7 The error distribution function $\Delta f(\mathbf{x})$ in U.

The index I_ε(3a) is written as

$$\begin{aligned} I_\varepsilon(3\mathrm{a}) &= -\int_{\mathbf{t}\in S}\int_{\mathbf{x}\in S} |\mathbf{x} - \mathbf{t}| \Delta f(\mathbf{x}) \Delta f(\mathbf{t}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ &= -\int_{\mathbf{t}\in S\setminus U}\int_{\mathbf{x}\in S\setminus U} |\mathbf{x} - \mathbf{t}| \Delta f(\mathbf{x}) \Delta f(\mathbf{t}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ &\quad -2a\left(K_2 + K_3\right) - 2b\left(K_1 + K_4\right) \qquad . \\ &\quad -a^2\left(k_{11} + 2k_{14} + k_{44}\right) - b^2\left(k_{22} + 2k_{23} + k_{33}\right) \\ &\quad -2ab\left(k_{12} + k_{13} + k_{24} + k_{34}\right) \end{aligned} \tag{18}$$

The index I_ε(3b) is also given by a similar equation. The difference of the indices is given by

$$\begin{aligned} I_\varepsilon(3\mathrm{b}) - I_\varepsilon(3\mathrm{a}) &= 2(a - b)\left(K_2 - K_4\right) \\ &\quad +2\left(k_{14} - k_{12}\right)\left(a^2 - 2ab + b^2\right). \\ &\approx 2\left(k_{14} - k_{12}\right)(a - b)^2 \end{aligned} \tag{19}$$

If $a \neq b$ then equation (19) is positive. Thus we have

$$I_\varepsilon(3\mathrm{b}) > I_\varepsilon(3\mathrm{a}). \tag{20}$$

Inequality (20) suggests that the global uniformity is more desirable than the local uniformity. Areal interpolation should be performed so that $\hat{f}(\mathbf{x})$ fits globally for $f(\mathbf{x})$.

From the above analyses, we obtained some desirable properties of areal interpolation methods. The first analysis showed that the estimation accuracy increases as the absolute value of the error distribution function $\Delta f(\mathbf{x})$ decreases. This is an intuitively understandable result because I_ε increases in proportion to the square of $\Delta f(\mathbf{x})$ as shown in equation (10). In the second analysis we found that the uniformity of $\Delta f(\mathbf{x})$ improves estimation accuracy. This implies that local but serious estimation error may considerably reduce estimation accuracy. The third analysis showed that the global uniformity is more important than the local uniformity in terms of estimation accuracy. These results can be summarized as shown in Figure 8.

Figure 8 The error distribution function $\Delta f(\mathbf{x})$ and estimation accuracy.

4. NUMERICAL EXAMINATIONS

In the previous section we have investigated the relationship between the accuracy of areal interpolation and spatial distribution of estimation errors. Since we employed a theory-based method in the analysis, the obtained results are rigid and deepen our understanding of areal interpolation. On the other hand, since no specific interpolation method was referred in the discussion, one might think that the results are abstract and difficult to apply to the choice of areal interpolation methods. Hence in this section we analyze the accuracy of areal interpolation in a more concrete way. To this end, we numerically evaluate four areal interpolation methods that are widely used in geography and GIS using the areal interpolation model proposed in Section 2.

We adopt a circle of radius 1.0 as the source zone *S*. This is because the circle is a good approximation of convex figures in terms of the accuracy of areal interpolation (Sadahiro 1999b). The representative point is located at the center of *S* whose locational vector is indicated by **z** (Figure 9). For the volume *V*, we try both $1/2\pi$ and 2π.

Figure 9 The source zone *S* and its representative point.

4.1 Areal Interpolation Methods

In numerical examinations we consider four methods of areal interpolation: 1) areal weighting, 2) point-in-polygon, 3) kernel, and 4) intelligent methods. The first three methods are often called *simple methods* in contrast to *intelligent methods* which use supplementary data in estimation process. We successively outline these methods with the representation of the estimator function $\hat{f}(\mathbf{x})$ (for derivations of $\hat{f}(\mathbf{x})$, see Sadahiro, 1998b).

(1) Areal weighting method

The areal weighting method is one of the most popular simple methods, which is frequently used when no information is available on the distribution of spatial objects. This method assumes that spatial objects are uniformly distributed in the source zone (Figure 10a). Consequently, the estimator function $\hat{f}(\mathbf{x})$ is written as

$$\hat{f}(\mathbf{x}) = \frac{\int_{\mathbf{x}\in S} f(\mathbf{x})\mathrm{d}\mathbf{x}}{\pi} = \frac{V}{\pi}. \tag{21}$$

(2) Point-in-polygon method

In GIS the number of spatial objects in a zone is often recorded as an attribute of the representative point of the zone. The point-in-polygon method sums up all the counts allocated to representative points that are included in the target zone (Okabe and Sadahiro, 1997; Sadahiro, 1998a). This implies that the method assumes all the spatial objects to be located exactly on the representative point (Figure 10b). Hence the estimator function $\hat{f}(\mathbf{x})$ is given by a delta function which is represented as the equation

$$\hat{f}(\mathbf{x}) = \begin{cases} \dfrac{V}{\pi r^2} & \mathbf{x} \in C(\mathbf{z}, r) \\ 0 & \mathbf{x} \notin C(\mathbf{z}, r) \end{cases} \tag{22}$$

in the limit $r \to 0$, where $C(\mathbf{z}, r)$ is the circle of radius r whose center is located at **z**.

(3) Kernel method

Kernel method is originally a statistical method used for nonparametric density estimation (Silverman, 1986; Scott, 1992). However, it is also applicable to areal interpolation when the locational data of representative points are available (Bracken and Martin, 1989, 1995; Bracken, 1993). Kernel method treats the representative point as a "high information point," that is, it assumes point objects to be clustered around the representative point. The probability density function called *kernel* is placed on the representative points to create a smooth surface.

There have been proposed various kernel functions in the literature. From these functions we choose the conical kernel for numerical examinations (Figure 10c). Mathematical representation of $\hat{f}(\mathbf{x})$ is somewhat complicated because of the volume-preserving principle. Suppose that the slope of the cone λ is given. The estimator function $\hat{f}(\mathbf{x})$ is then represented as follows:

if $\frac{1}{3}\pi\lambda < V$ then

$$\hat{f}(\mathbf{x}) = \frac{1}{\pi}\left(V + \frac{2}{3}\pi\lambda\right) - \lambda|\mathbf{x} - \mathbf{z}|, \tag{23}$$

otherwise,

$$\hat{f}(\mathbf{x}) = \begin{cases} \sqrt[3]{\dfrac{3\lambda^2 V}{\pi}} - \lambda|\mathbf{x}-\mathbf{z}| & if\ |\mathbf{x}-\mathbf{z}| \le \sqrt[3]{\dfrac{3V}{\pi\lambda}} \\ 0 & otherwise \end{cases}. \tag{24}$$

For the value of λ we try 1.0, 3.0 and 5.0.

(4) Intelligent method

As GIS and remotely-sensed satellite images become widely available, new areal interpolation methods called *intelligent methods* have been developed (Fisher and Langford, 1995, 1996; Langford *et al.*, 1991; Goodchild *et al.*, 1993). They use supplementary data such as satellite images or landuse data in areal interpolation to improve estimation accuracy.

Intelligent methods can be described mathematically by use of a supplementary function φ(**x**) which represents additional data. Let us consider, for instance, the dasymetric method which is often used for population estimation. Using landuse data this method divides a source zone into subregions in each of which population density is assumed to be constant. Typically, a source zone is divided into two subregions, that is, residential and non-residential areas. In this case the function φ(**x**) is written as a binary function

$$\varphi(\mathbf{x}) = \begin{cases} 1 & if\ \mathbf{x} \in R \\ 0 & otherwise \end{cases}, \tag{25}$$

where R indicates the residential area. The estimator function $\hat{f}(\mathbf{x})$ is then given by

$$\hat{f}(\mathbf{x}) = \frac{\varphi(\mathbf{x})}{\int_{\mathbf{x}\in S} \varphi(\mathbf{x})\mathrm{d}\mathbf{x}} V. \tag{26}$$

Since population density is highly correlated with landuse category, equation (25) may be represented as

$$\varphi(\mathbf{x}) = \begin{cases} 1 & if\ \ f(\mathbf{x}) \ge \alpha \\ 0 & otherwise \end{cases}, \tag{27}$$

where α is the lower limit of the population density of residential areas.

If more than two landuse categories are distinguishable, φ(**x**) is given by a stepwise function

$$\varphi(\mathbf{x}) = k_i \quad if\ \mathbf{x} \in L_i \tag{28}$$

(Figure 10d), where L_i indicates the type i landuse area. Substituting equation (28) into equation (26), we obtain the estimator function $\hat{f}(\mathbf{x})$.

Linear regression models can be regarded as the case where the φ(**x**) is a linear function of f(**x**) having a probabilistic error term ε, that is,

$$\varphi(\mathbf{x}) = \alpha + \beta f(\mathbf{x}) + \varepsilon. \tag{29}$$

Non-linear regression models can also be represented in a similar way.

In numerical examinations, we try three forms of the supplementary function φ(**x**) as follows.

Method I$_1$:
$$\varphi_1(\mathbf{x}) = \begin{cases} 1 & if\ \ f(\mathbf{x}) \geq 1 \\ 0 & otherwise \end{cases} \tag{30}$$

Method I$_2$:
$$\varphi_2(\mathbf{x}) = \begin{cases} 1 & if\ \ f(\mathbf{x}) \geq 1 \\ 2/3 & if\ 1 > f(\mathbf{x}) \geq 2/3 \\ 1/3 & if\ 2/3 > f(\mathbf{x}) \geq 1/3 \\ 0 & if\ 1/3 > f(\mathbf{x}) \end{cases} \tag{31}$$

Method I$_3$:
$$\varphi_3(\mathbf{x}) = \begin{cases} 3 & if\ \ f(\mathbf{x}) \geq 3 \\ 2 & if\ 3 > f(\mathbf{x}) \geq 2 \\ 1 & if\ 2 > f(\mathbf{x}) \geq 1 \\ 0 & if\ 1 > f(\mathbf{x}) \end{cases} \tag{32}$$

If φ(**x**)=0 for any **x**, we redefine the function as

$$\varphi(\mathbf{x}) = k. \tag{33}$$

In this case the intelligent method is equivalent to the areal weighting method.

Figure 10 Estimator functions of areal interpolation methods. a) Areal weighting, b) point-in-polygon, c) kernel, and d) intelligent methods.

4.2 Results

In order to assess the accuracy of the above interpolation methods, we employ four types of point distributions as *f*(**x**): 1) cylinder, 2) conical, 3) annular, and 4) linear distributions (Figure 11). The results of numerical examinations are successively described in the following.

Figure 11 Forms of the function *f*(**x**). a) Cylinder, b) conical, c) annular, and d) linear distributions.

(1) Cylinder distribution

Equation (10) and the results obtained in Section 3 indicate that areal interpolation yields better estimates if the estimator function is close to the true distribution. Consequently, for instance, the point-in-polygon and kernel methods are expected to give better results when point objects are clustered around the representative point. To confirm this we examine how the degree of concentration affects the accuracy of estimates using the cylinder distribution, a typical concentrated distribution of point objects. The radius of the cylinder indicated by *r* changes from 0 to 1 as shown in Figure 12. The height of the cylinder is denoted by *h*.

Figure 12 Cylinder distributions.

The relationship between the radius *r* and estimation accuracy is depicted in Figure 13. The results are quite consistent with our intuition. As the distribution becomes concentrated estimation accuracy of the point-in-polygon method linearly increases whereas that of the areal weighting method decreases. The kernel methods are in the middle of these methods. The accuracy of the intelligent methods is noteworthy: the error index I_ε is nearly always zero. This is because the stepwise functions assumed in the intelligent methods fit locally-uniform distributions.

Figure 13 Estimation accuracy on the cylinder distribution. *r*: the radius of the cylinder. *h*: the height of the cylinder. K_i: the kernel method of $\lambda=i$. I_i: the intelligent method of the function $\varphi_i(\mathbf{x})$. a) $V=1/2\pi$, b) $V=2\pi$.

(2) Conical distribution

We next discuss the conical distribution as another example of centralized distributions. The slope of the cone denoted by *m* increases gradually from 0 to 100 (Figure 14). The height of the cone is indicated by *h*.

Figure 14 Conical distributions.

The results are shown in Figure 15. Concerning the areal weighting and point-in-polygon methods the results are quite understandable as in the case of the cylinder distribution. On the other hand, it is counterintuitive that the intelligent methods do not always give the best result among the four interpolation methods. For instance, let us take a look at Figure 15a-2. When the slope *m* is between 1 and 2, the intelligent methods I_1 and I_3 are far less accurate than the areal weighting and kernel methods. The intelligent methods do not suit the gently-sloping distribution, though they work successfully for the distribution of a steep slope (see Figures 15a-1 and 15b-1). This can be partly understood by looking carefully at Figure 15b-2. When the cone height *h* is around 5.0, the intelligent method I_3 whose boundary values are 1.0, 2.0, and 3.0 (recall equation (32)) gives better estimates than I_1 and I_2 having boundaries of smaller values. For h=2.1, however, I_3 is the least accurate among the intelligent methods. From this we can say that estimation accuracy of the intelligent methods depends on the fitness of the range of the stepwise function $\varphi(\mathbf{x})$ for that of the density function $f(\mathbf{x})$. They yield poor estimates if the boundaries of $\varphi(\mathbf{x})$ do not agree with the range of $f(\mathbf{x})$.

Another counterintuitive result is also found in Figure 15b. The intelligent method I_2 is less accurate than I_1 for *h*>4.0, though I_2 has a finer stepwise function than I_1. This is also caused by the lack of agreement between the ranges of $\varphi(\mathbf{x})$ and $f(\mathbf{x})$. The fineness of the stepwise function sometimes makes areal interpolation less accurate, if the ranges of supplementary data and true distribution do not agree.

One might think it strange that the intelligent methods yield the best estimates for very large *h*, say, *h*=20.0. In such a case, however, $f(\mathbf{x})$ is equal to zero in almost everywhere so that the boundary setting of $\varphi(\mathbf{x})$ scarcely affects the estimation accuracy.

Figure 15 Estimation accuracy on the conical distribution. *m*: the slope of the cone. *h*: the height of the cone. K_i: the kernel method of $\lambda=i$. I_i: the intelligent method of the function $\varphi_i(\mathbf{x})$. a-1) $V=1/2\pi$, a-2) $V=1/2\pi$ (magnified), b-1) $V=2\pi$, b-2) $V=2\pi$ (magnified).

(3) Annular distribution

Contrary to the previous two distributions we next consider a disconcentrated distribution, that is, the annular distribution. Using this distribution we analyze the relationship between the degree of discentralization and the estimation accuracy of areal interpolation. The radius of the hole indicated by *r* changes from 0 to 1 as shown in Figure 16. The height of the annulus is denoted by *h*.

Figure 16 Annular distributions.

The results are shown in Figure 17. It depicts that the simple methods fail to give accurate estimates when point objects are distributed along the edge of the source zone. This is obviously because the annular distribution is quite different from the distribution assumed in the simple methods. Only the intelligent methods yield satisfactory results for the annular distribution. The results warn us that it is quite dangerous to assume blindly an arbitrary distribution in areal interpolation.

Figure 17 Estimation accuracy on the annular distribution. *r*: the radius of the hole. *h*: the height of the annulus. K_i: the kernel method of $\lambda=i$. I_i: the intelligent method of the function $\varphi_i(\mathbf{x})$. a) $V=1/2\pi$, b) $V=2\pi$.

(4) Linear distribution

We finally examine the linear distribution. This type of monotonous distribution, though not necessarily linear, is often found in global population distribution. The slope indicated by *m* changes from 0 to V/π as shown in Figure 18.

Figure 18 Linear distributions.

The results shown in Figure 19 are similar to those in Figure 17 in that the simple methods are generally inaccurate. Estimation accuracy of the simple methods monotonously reduces with an increase of *m*. On the other hand, the intelligent methods yield better estimates for any *m*. Among the intelligent methods, I_1 and I_2 are better than I_3 for h<2.6 (Figure 19b), whereas I_3 is better for larger *h*. This tendency was also found in the conical distribution case (Figure 15b), which is caused by the disagreement between the ranges of $\varphi(\mathbf{x})$ and $f(\mathbf{x})$ as discussed earlier.

Figure 19 Estimation accuracy on the linear distribution. *m*: the slope of the linear distribution. *h*: the height of the distribution. K_i: the kernel method of $\lambda=i$. I_i: the intelligent method of the function $\varphi_i(\mathbf{x})$. a) V=1/2π, b) V=2π.

5. CONCLUDING DISCUSSION

In this paper we have analyzed the accuracy of areal interpolation from both theoretical and empirical viewpoints. To keep generality in the analysis, we first developed a stochastic process which we call areal interpolation model that represents areal interpolation in diverse geographical situations. We then proposed a measure of estimation accuracy called error index on the basis of the model.

The results obtained from theoretical considerations in Section 3 can be summarized as follows: 1) estimation accuracy improves as the error at individual locations decreases; 2) estimation accuracy improves as the error distribution becomes smooth and uniform; 3) the global uniformity in the error distribution is more important than the local uniformity in terms of estimation accuracy. Though the first conclusion is not surprising, the second and third results are useful for the evaluation of areal interpolation methods. The spatial uniformity in the error distribution, or the global fitness of estimator function, is a desirable property of areal interpolation methods.

From numerical examinations we obtained the following results: 1) intelligent methods are generally more accurate than simple methods; 2) estimation accuracy of simple methods heavily depends on the appropriateness of their hypothetical distributions; 3) estimation accuracy of intelligent methods depends on the fitness of the range of supplementary data for that of true distribution; 4) if the ranges do not agree, intelligent methods may yield poorer estimates than simple methods.

The above results are quite suggestive for the choice of areal interpolation methods. However, they are clearly limited in several aspects. We finally discuss the limitations of this paper for further research.

First, the effects of the shape and size of the target and source zones are remained to be analyzed. We have fixed these factors throughout the analysis in order to focus on how the relationship between two spatial distributions, one is actual and the other is assumed in areal interpolation, affects estimation accuracy. Though the influence of this relationship is quite significant, geometrical properties of spatial units are also important. Concerning the areal weighting method, their effects on estimation accuracy were numerically investigated (Sadahiro, 1999b). However, the study was limited to one interpolation method, so accuracy of other methods should be analyzed in relation to geometrical properties of zones.

Second, we have discussed the accuracy of areal interpolation only for a pair of source and target zones. This allows us to examine the relationship between estimation accuracy and error distribution from a theoretical point of view. However, there is a clear demand for theoretical analysis in more realistic situations - a combination of multiple source and target zones. Though it is quite difficult to discuss estimation accuracy keeping generality in complicated geographical circumstances, the difficulties have to be overcome in future research.

Third, there still remain areal interpolation methods not examined in this paper. For instance, the Poisson regression method (Flowerdew and Green, 1989) and the pycnophylactic method (Tobler, 1979) should also be tested. Since they involve a convergence calculation in estimation process, the first step to assess their accuracy is to analyze the properties of the estimator function yielded from convergence calculation. Once the estimator function is given in an analytical form, their accuracy can be evaluated by the method proposed in this paper.

LITERATURE CITED

Bracken, I. and D. Martin (1989). "The Generation of Spatial Population Distributions from Census Centroid Data." *Environment and Planning A* **21**, 537-543.

Bracken, I. (1993). "An Extensive Surface Model Database for Population-Related Information." *Environment and Planning B* **20**, 13-27.

Bracken, I. and D. Martin (1995). "Linkage of the 1981 and 1991 UK Censuses Using Surface Modelling Concepts." *Environment and Planning A* **27**, 379-390.

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.

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.

Fisher, P. F. and M. Langford (1996). "Modeling Sensitivity to Accuracy in Classified Imagery: A Study of Areal Interpolation by Dasymetric Mapping." Professional Geographer **48**, 299-309.

Flowerdew, R. and M. Green (1989). "Statistical Methods for Inference between Incompatible Zonal Systems." In *Accuracy of Spatial Databases*, edited by M. Goodchild and S. Gopal, pp. 239-247, London: Taylor and Francis.

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.

Langford, M., P. Fisher, and D. Troughear (1993). "Comparative Accuracy Measurements of the Cross Areal Interpolation of Population." *Proceedings of EGIS*

'93, 663-674.

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.

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.

Sadahiro, Y. (1998a). "Accuracy Count Data Estimated by the Point-in-Polygon Method." *Discussion Paper Series* **77E**, Department of Urban Engineering, University of Tokyo.

Sadahiro, Y. (1998b). "Accuracy of Areal Interpolation for Surface." *Discussion Paper Series* **79**, Department of Urban Engineering, University of Tokyo (in Japanese).

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

Scott, D. W. (1992). *Multivariate Density Estimation*. New York: John Wiley.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

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.

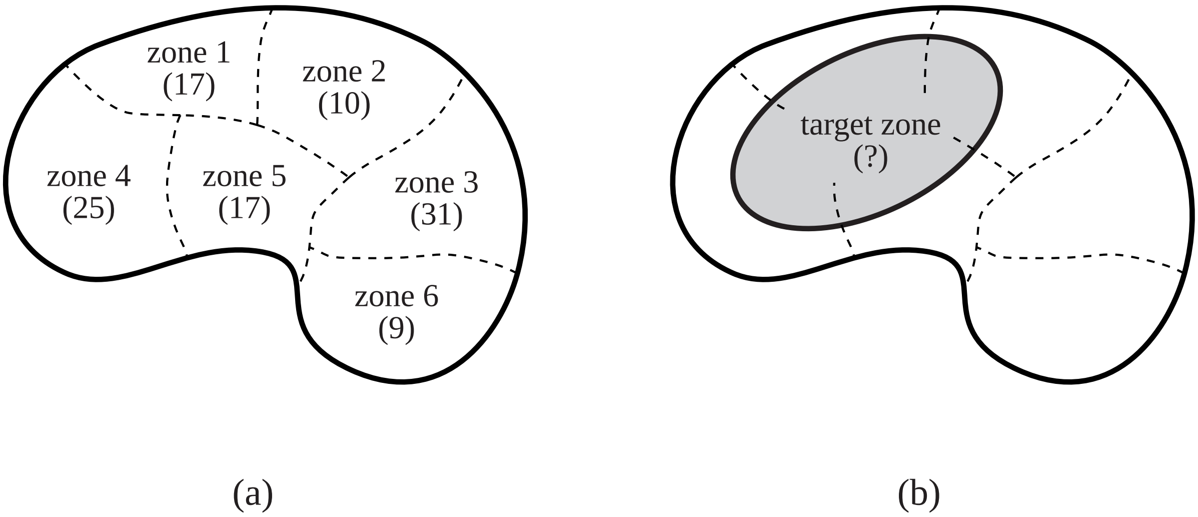

zone 1
(17)
zone 2
(10)
zone 4
(25)
zone 5
(17)
zone 3
(31)
zone 6
(9)
target zone
(?)
(a)
(b)

Figure 1

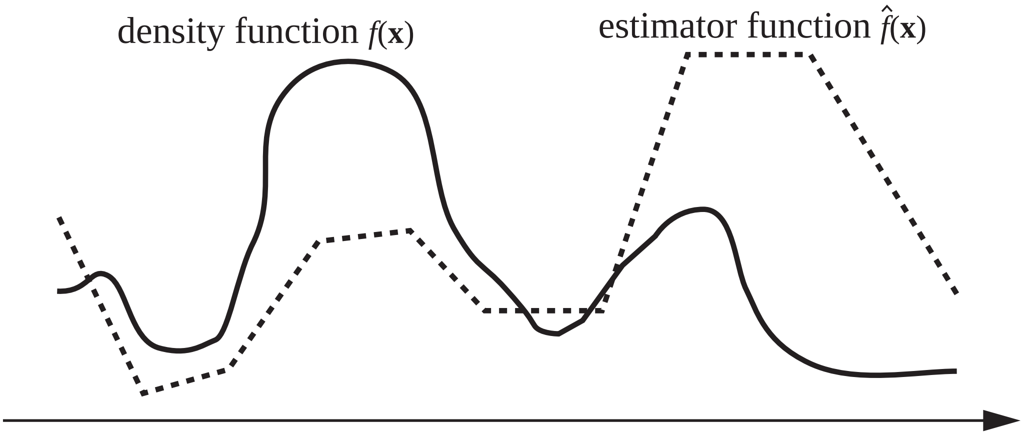

density function $f(\mathbf{x})$
estimator function $\hat{f}(\mathbf{x})$

Figure 2

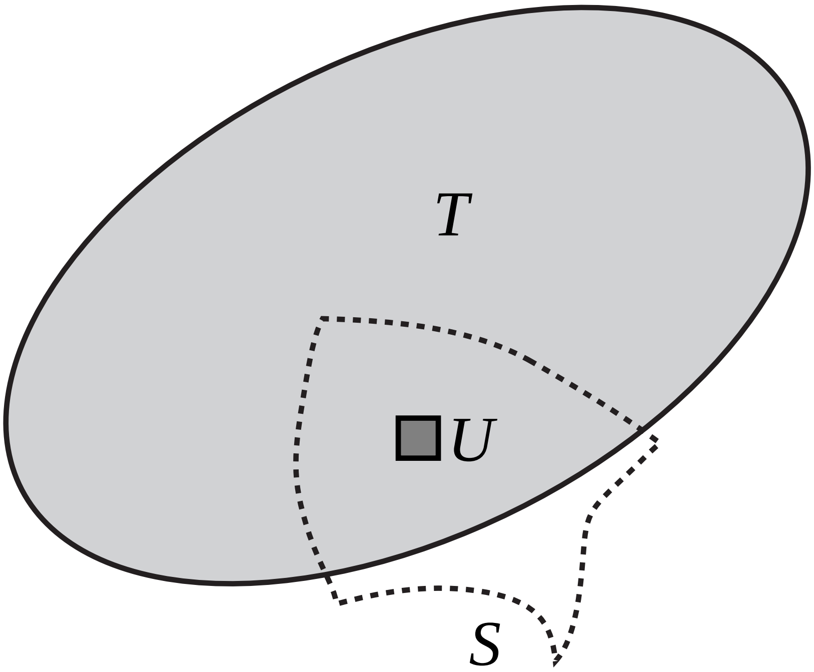


Figure 3

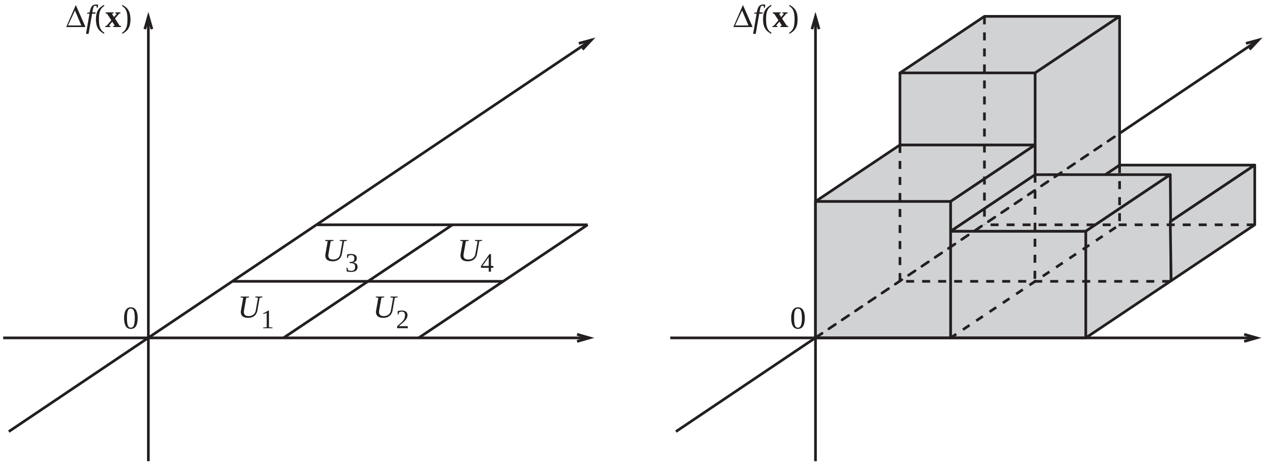
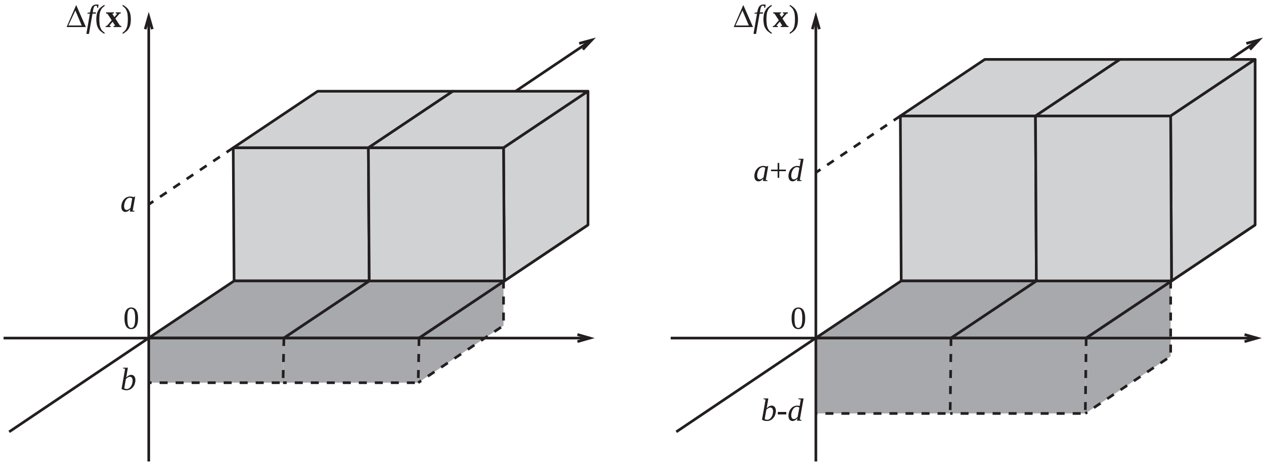


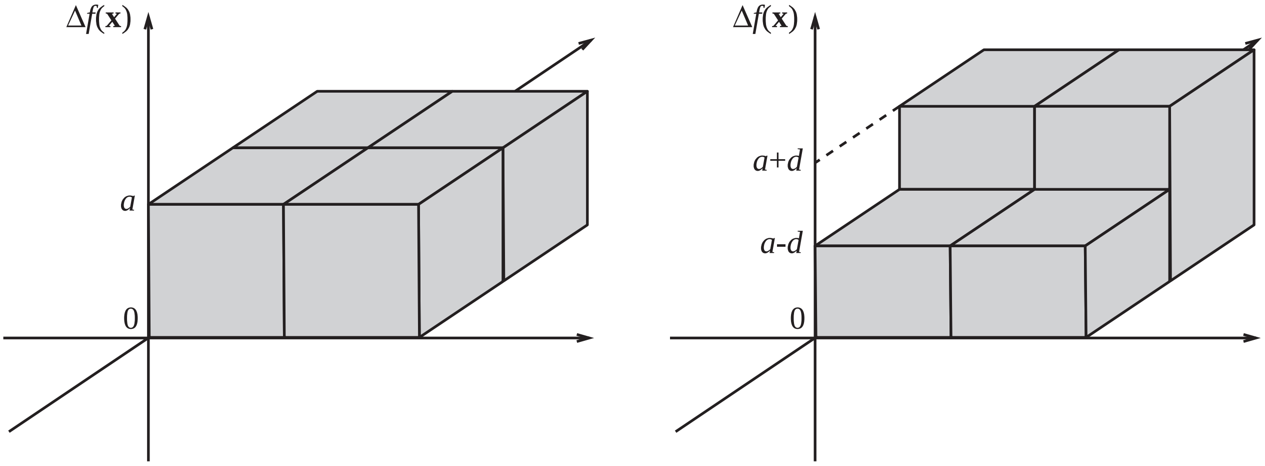
Figure 4



Case 1a

Case 1b

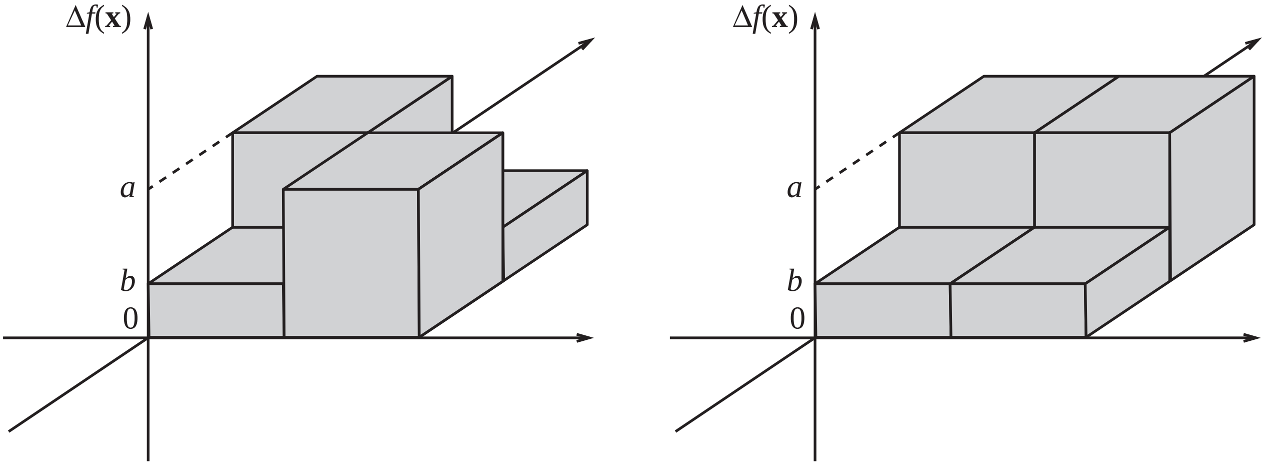
Figure 5



Case 2a

Case 2b

Figure 6



Case 3a

Case 3b

Figure 7

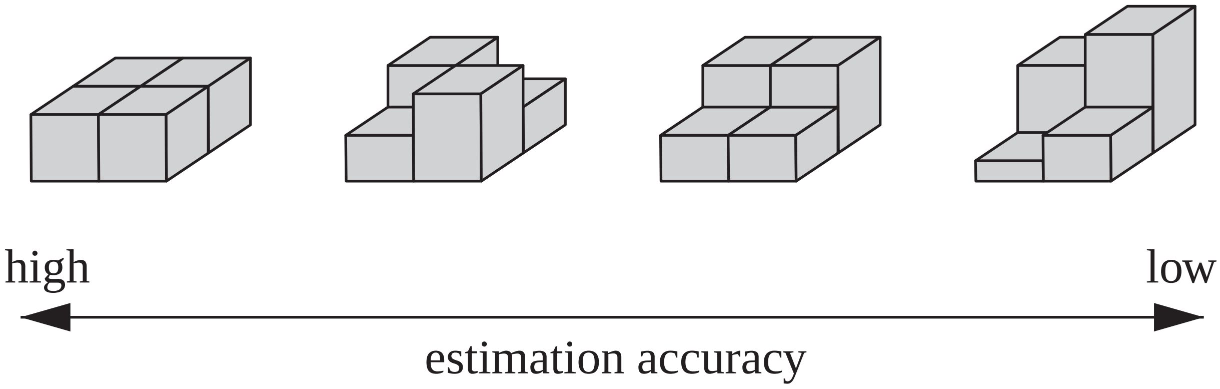


Figure 8

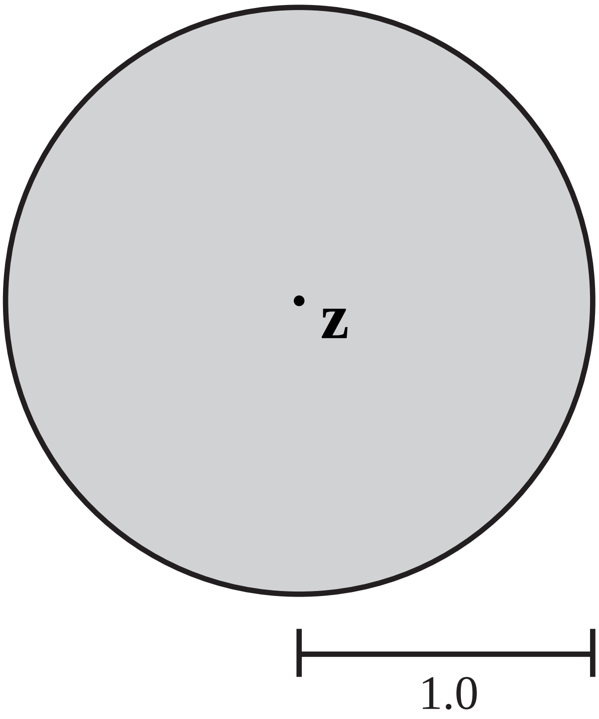


Figure 9

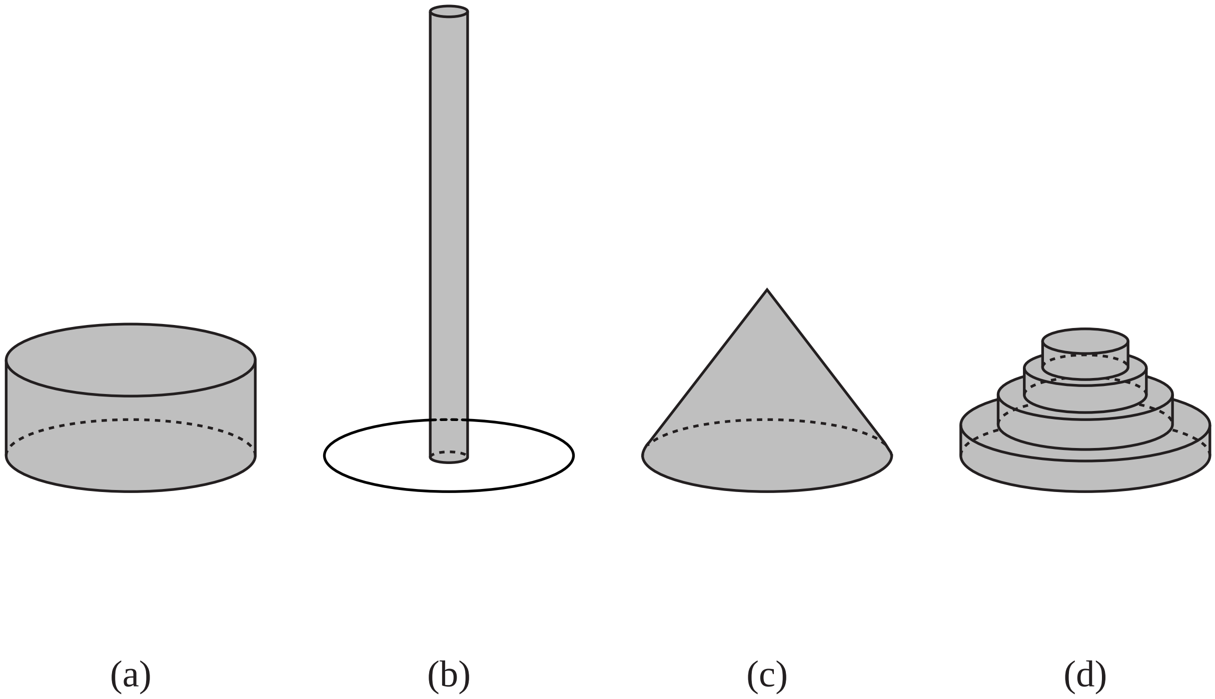


Figure 10

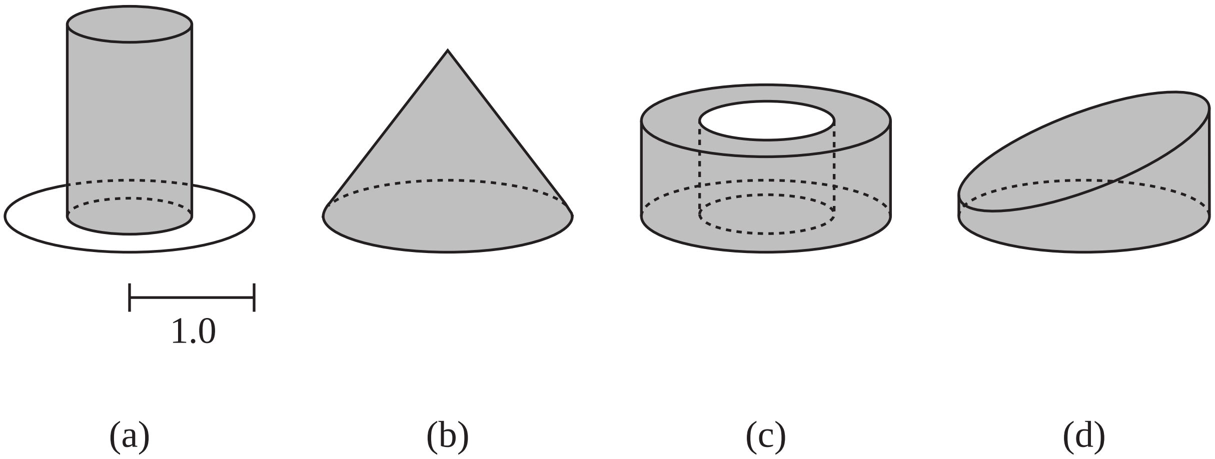


Figure 11

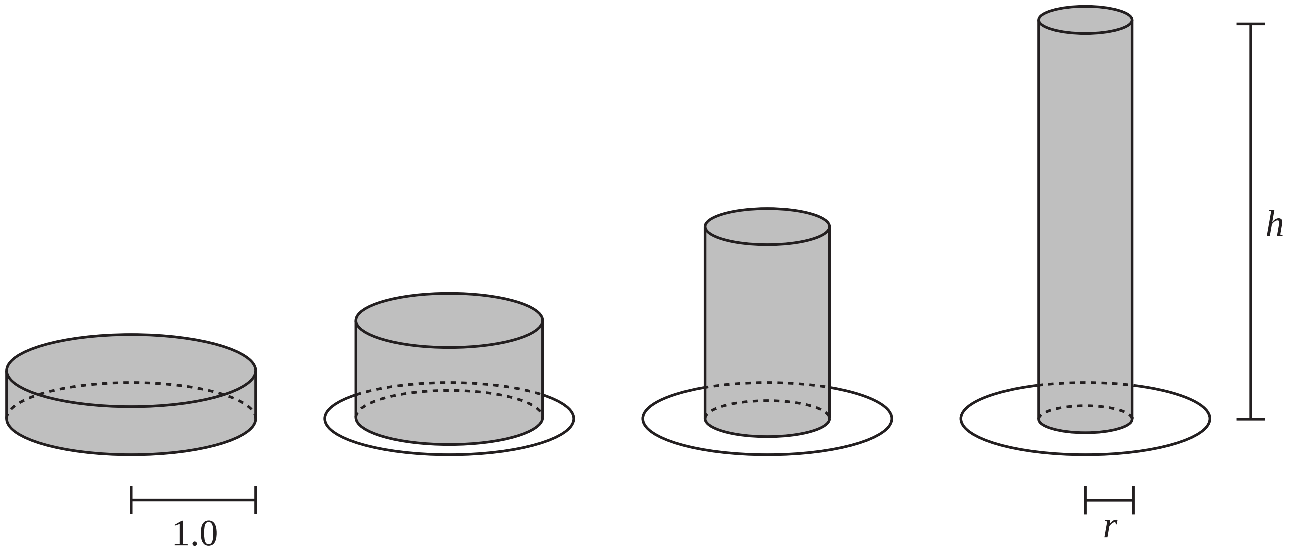


Figure 12

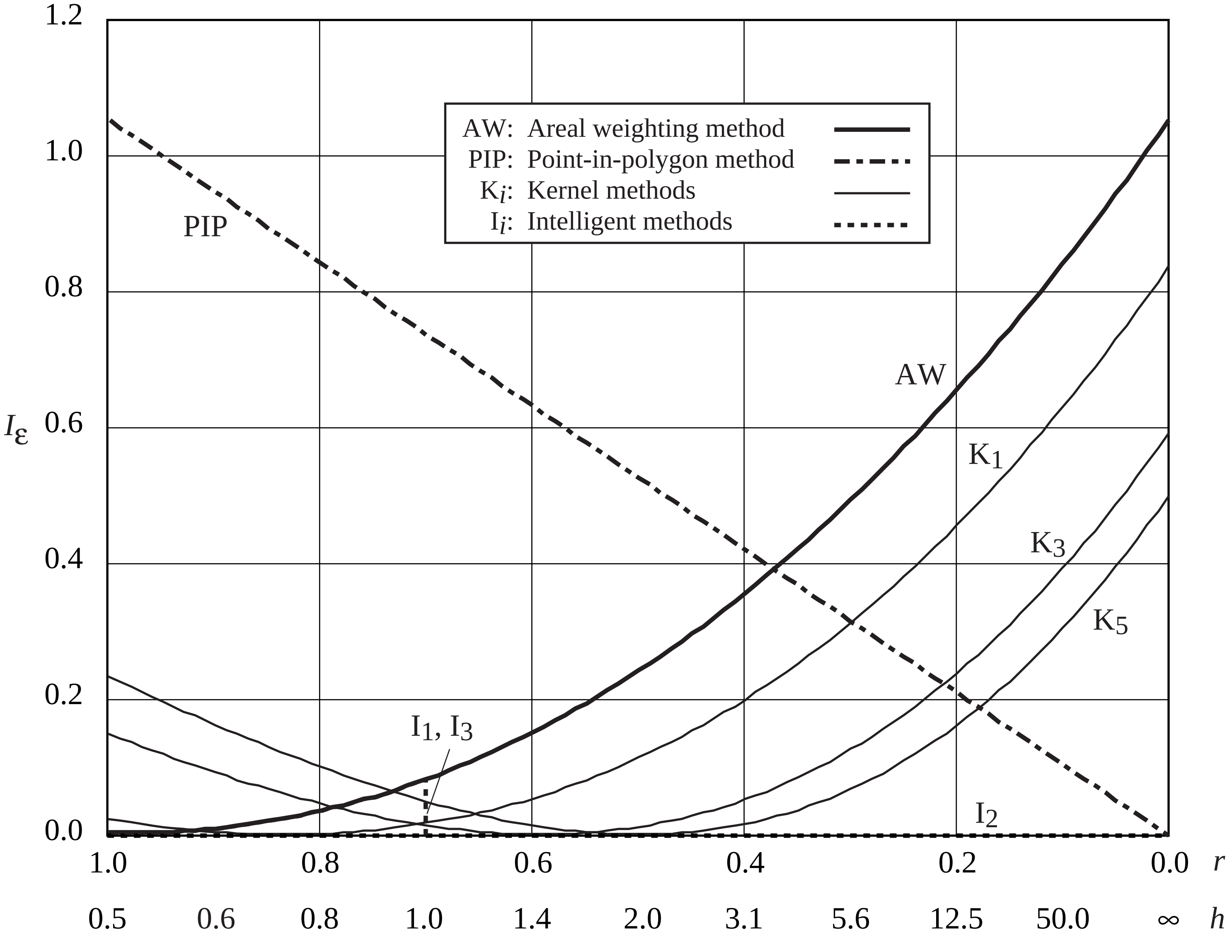

1.2
1.0
0.8
I_ε 0.6
0.4
0.2
0.0
AW: Areal weighting method
PIP: Point-in-polygon method
K_i: Kernel methods
I_i: Intelligent methods
PIP
AW
K_1
K_3
K_5
I_1, I_3
I_2
1.0
0.8
0.6
0.4
0.2
0.0
r
0.5
0.6
0.8
1.0
1.4
2.0
3.1
5.6
12.5
50.0
∞
h

Figure 13a

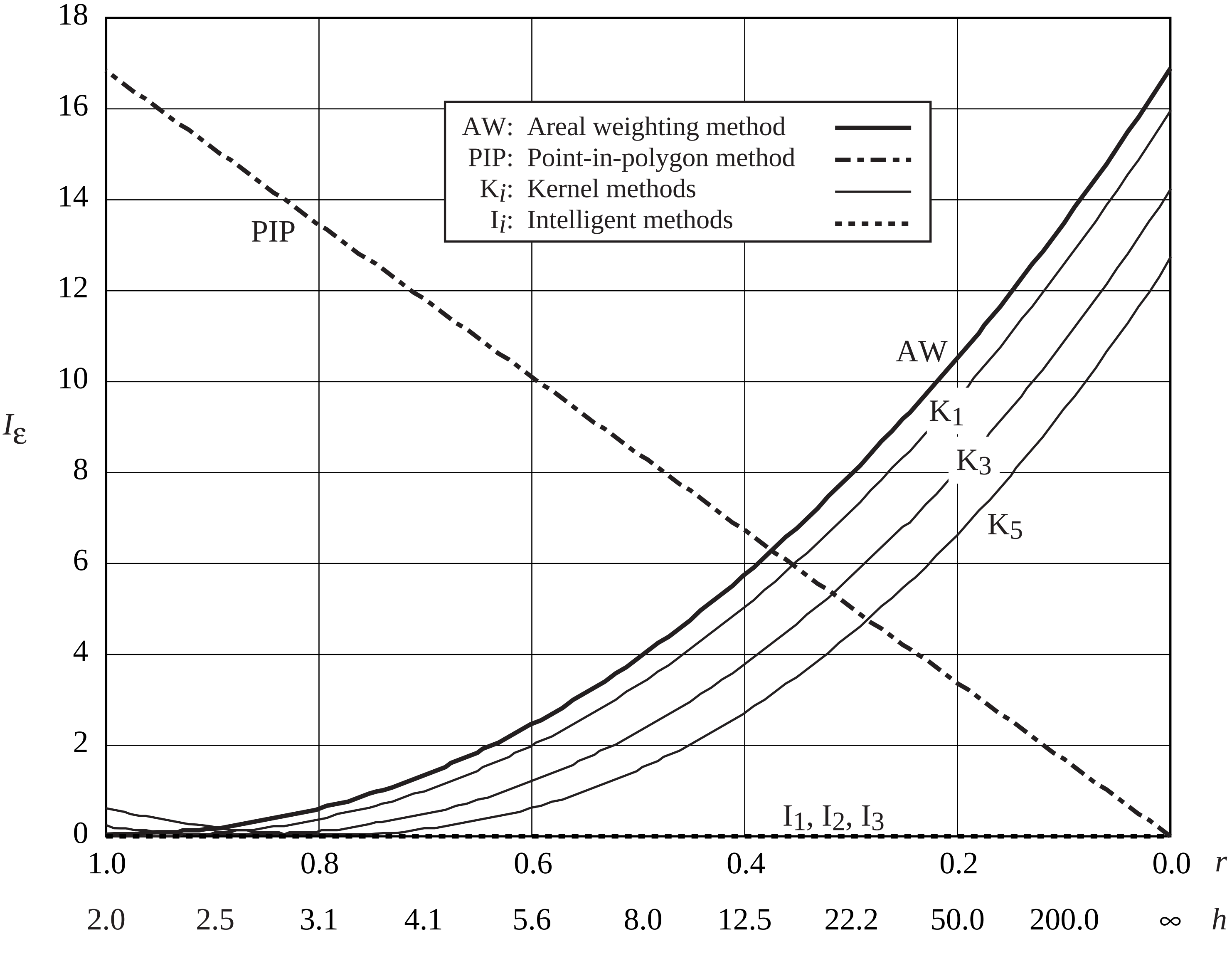

18
16
14
12
10
8
6
4
2
0
I_ε
AW: Areal weighting method
PIP: Point-in-polygon method
K_i: Kernel methods
I_i: Intelligent methods
PIP
AW
K_1
K_3
K_5
I_1, I_2, I_3
1.0
0.8
0.6
0.4
0.2
0.0
r
2.0
2.5
3.1
4.1
5.6
8.0
12.5
22.2
50.0
200.0
∞
h

Figure 13b

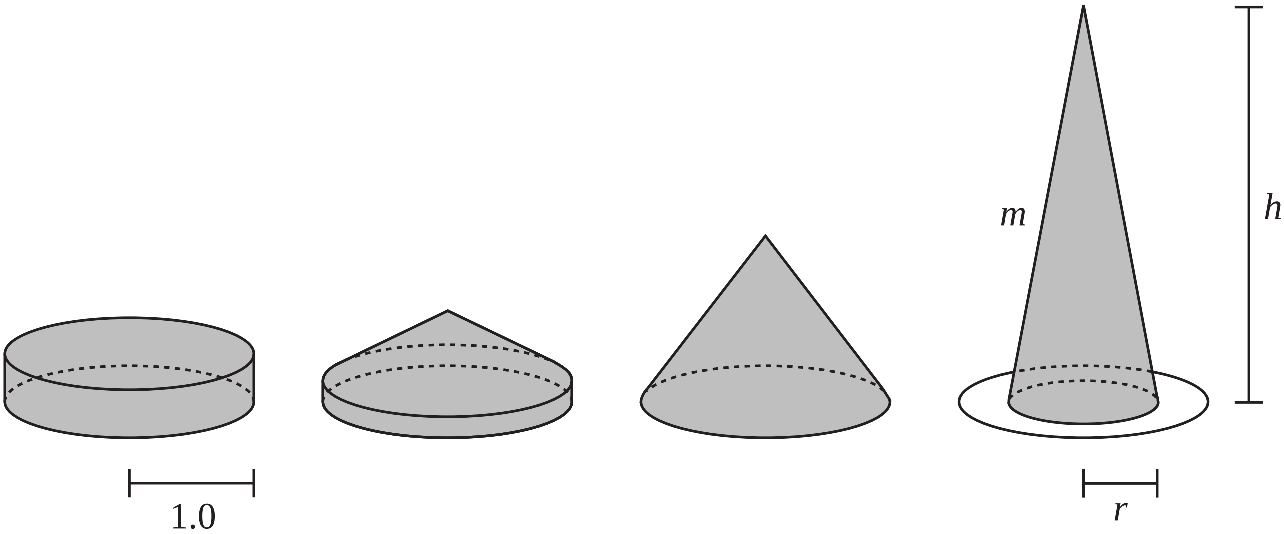


Figure 14

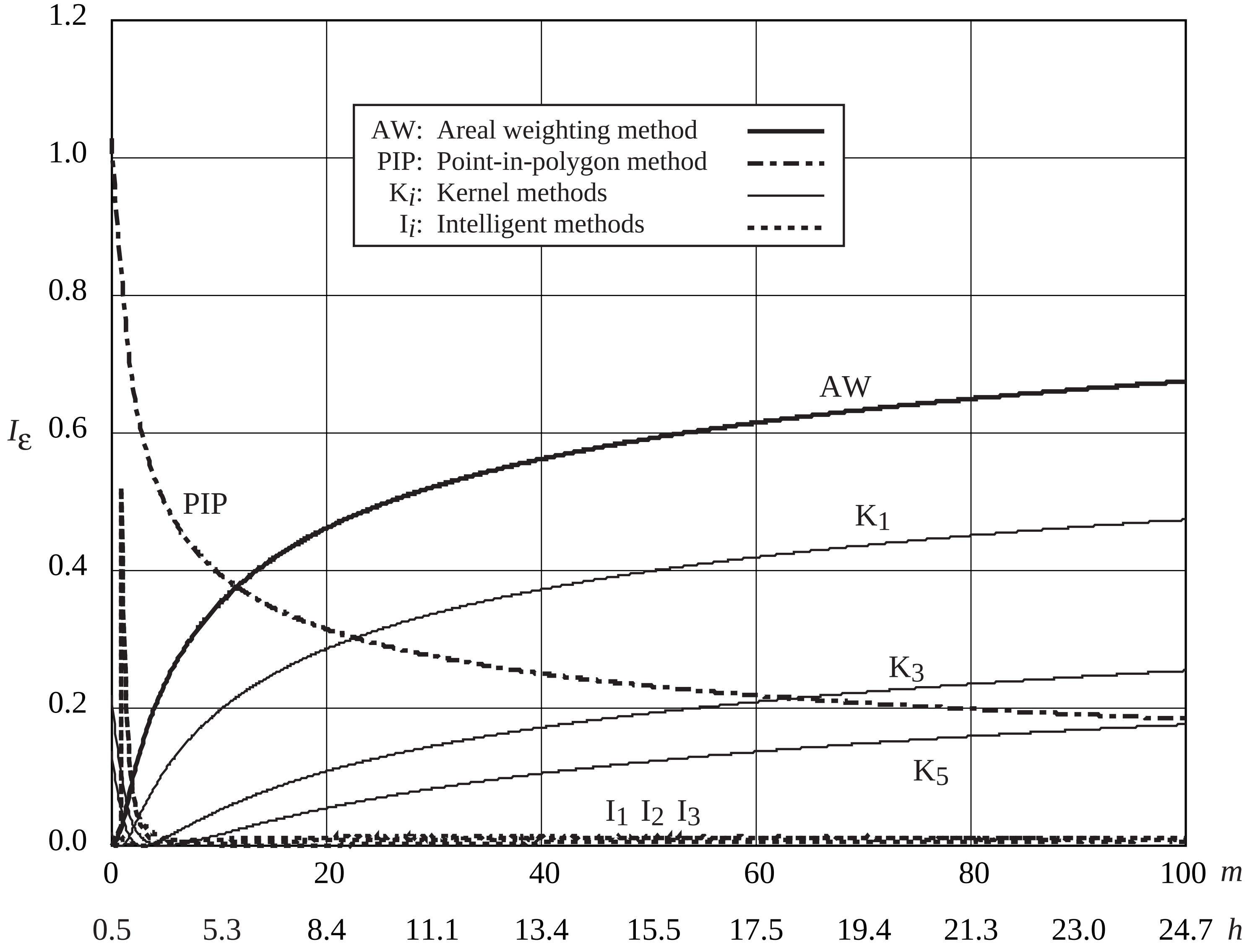

1.2
1.0
0.8
0.6
0.4
0.2
0.0
I_ε
AW: Areal weighting method
PIP: Point-in-polygon method
K_i: Kernel methods
I_i: Intelligent methods
AW
PIP
K_1
K_3
K_5
I_1 I_2 I_3
0
20
40
60
80
100
m
0.5
5.3
8.4
11.1
13.4
15.5
17.5
19.4
21.3
23.0
24.7
h

Figure 15a-1

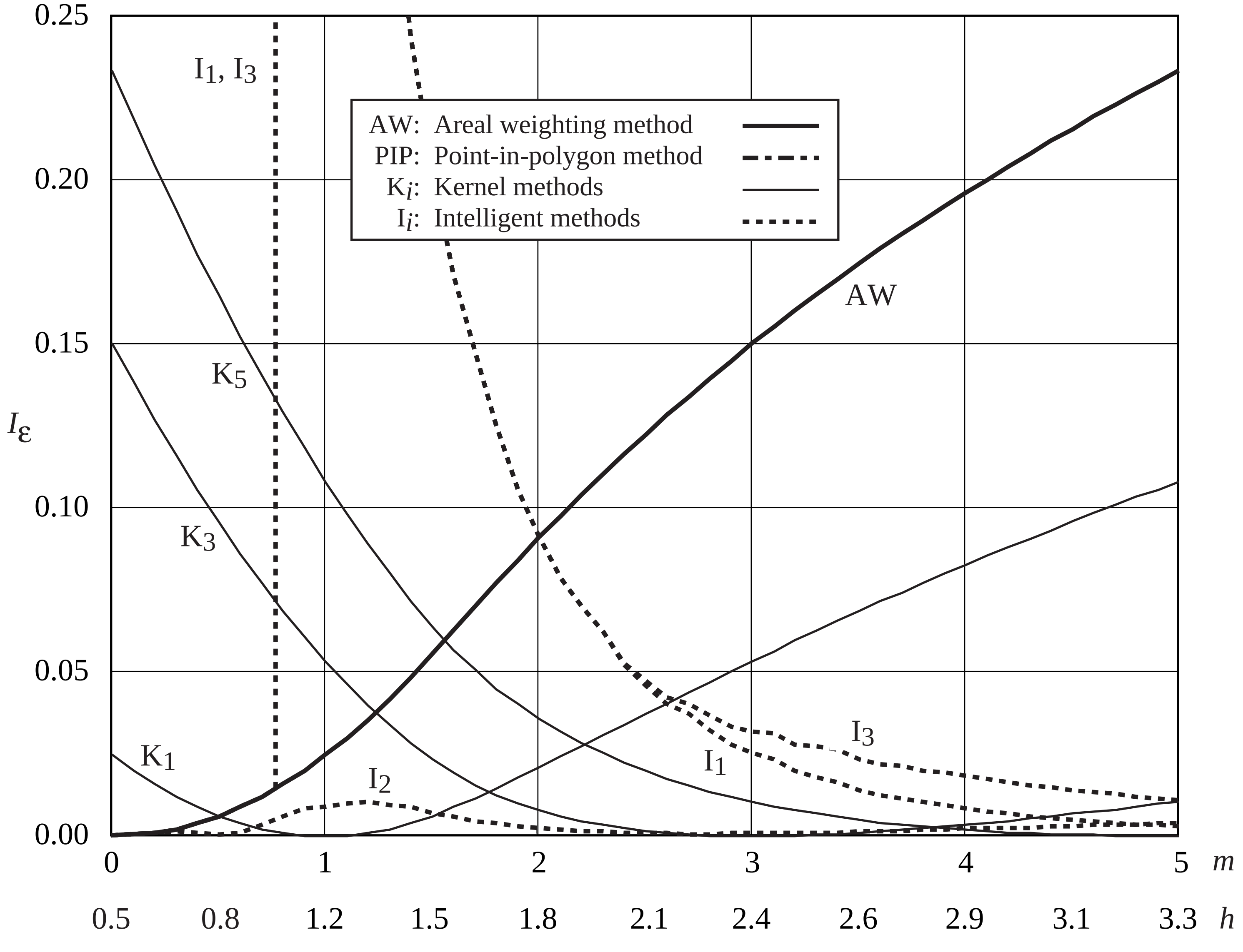

0.25
0.20
0.15
0.10
0.05
0.00
I_ε
AW: Areal weighting method
PIP: Point-in-polygon method
K_i: Kernel methods
I_i: Intelligent methods
I_1, I_3
K_5
K_3
K_1
I_2
I_1
I_3
AW
0
1
2
3
4
5
m
0.5
0.8
1.2
1.5
1.8
2.1
2.4
2.6
2.9
3.1
3.3
h

Figure 15a-2

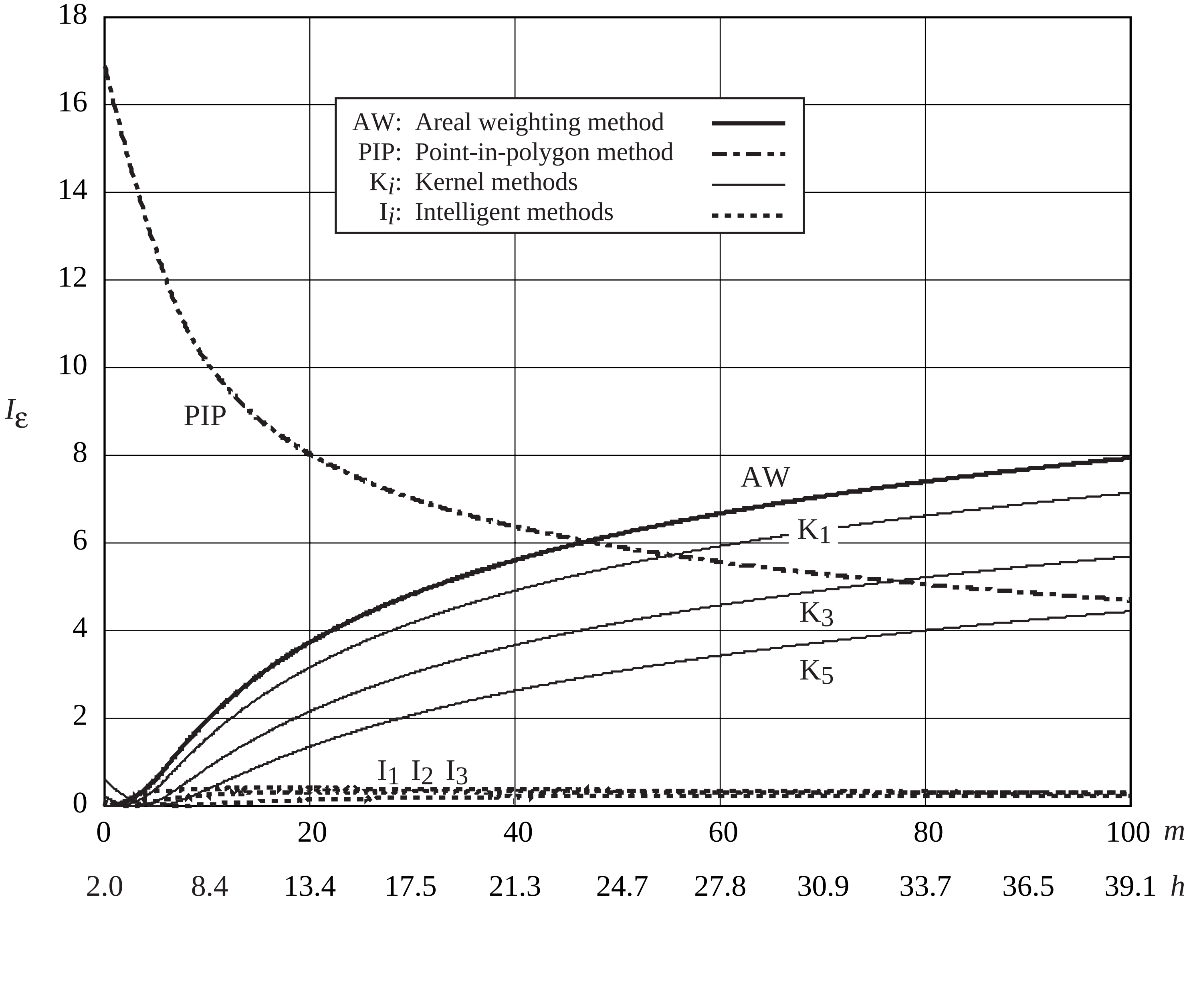

18
16
14
12
10
8
6
4
2
0
I_ε
AW: Areal weighting method
PIP: Point-in-polygon method
K_i: Kernel methods
I_i: Intelligent methods
PIP
AW
K_1
K_3
K_5
I_1 I_2 I_3
0
20
40
60
80
100
m
2.0
8.4
13.4
17.5
21.3
24.7
27.8
30.9
33.7
36.5
39.1
h

Figure 15b-1

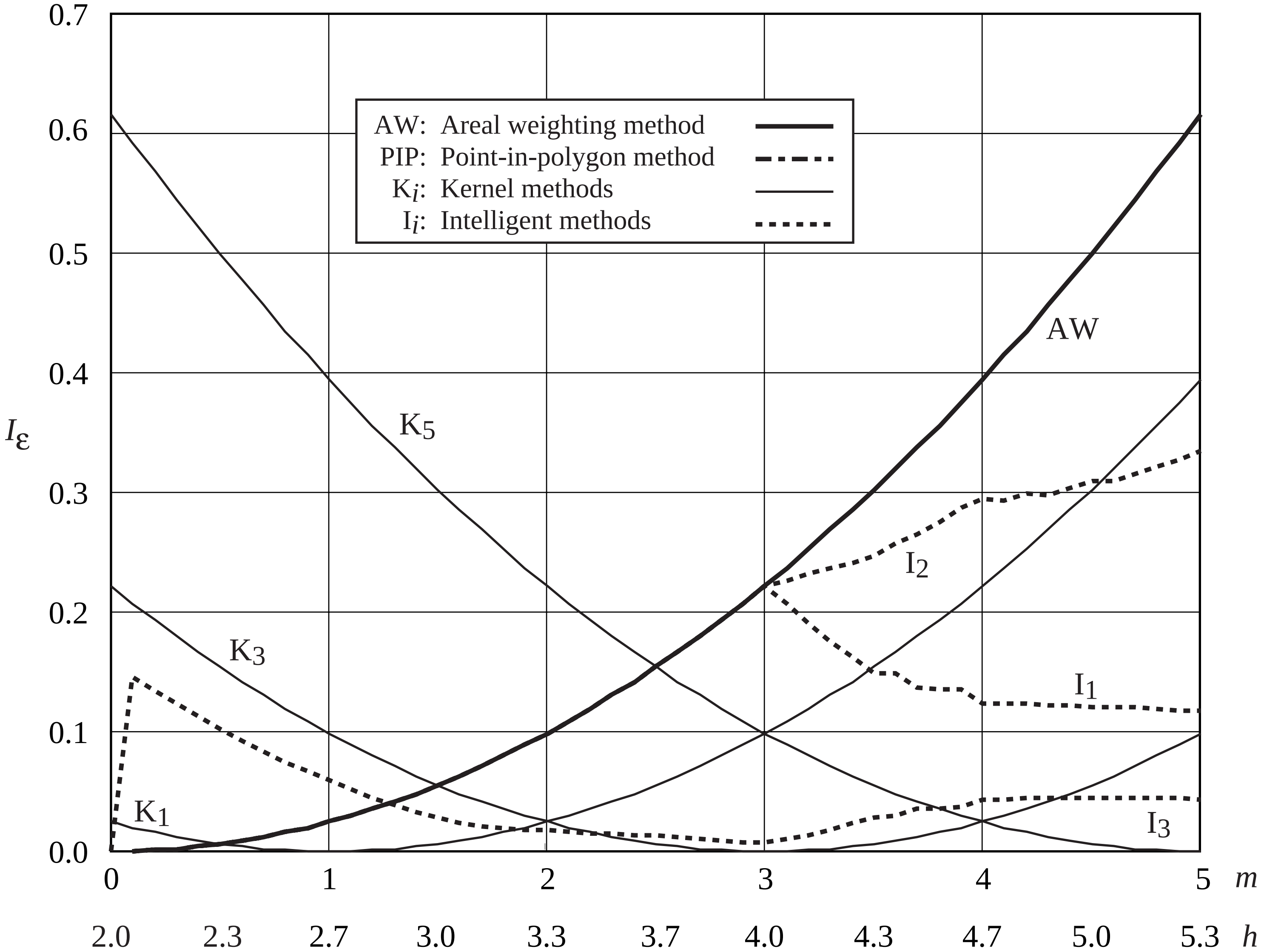

AW: Areal weighting method
PIP: Point-in-polygon method
K_i: Kernel methods
I_i: Intelligent methods
I_ε
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
AW
K_5
K_3
K_1
I_1
I_2
I_3
0
1
2
3
4
5
m
2.0
2.3
2.7
3.0
3.3
3.7
4.0
4.3
4.7
5.0
5.3
h

Figure 15b-2

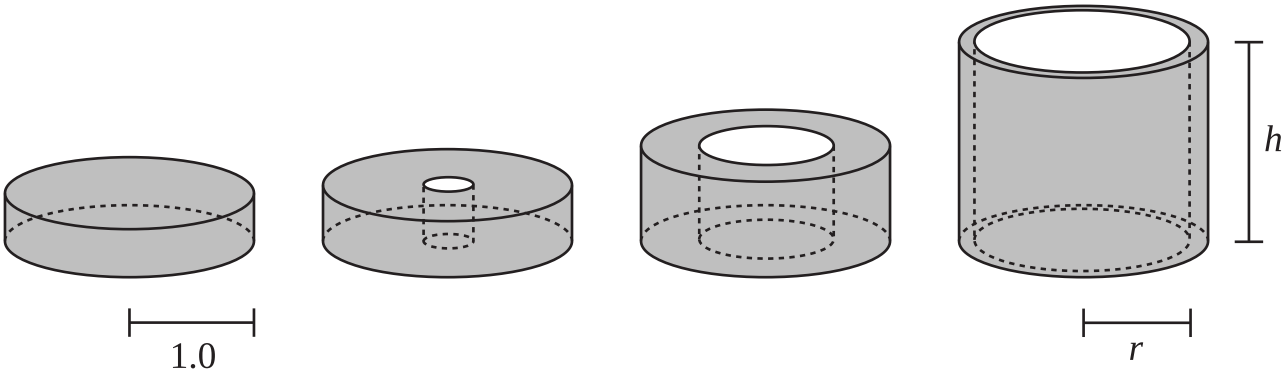


Figure 16

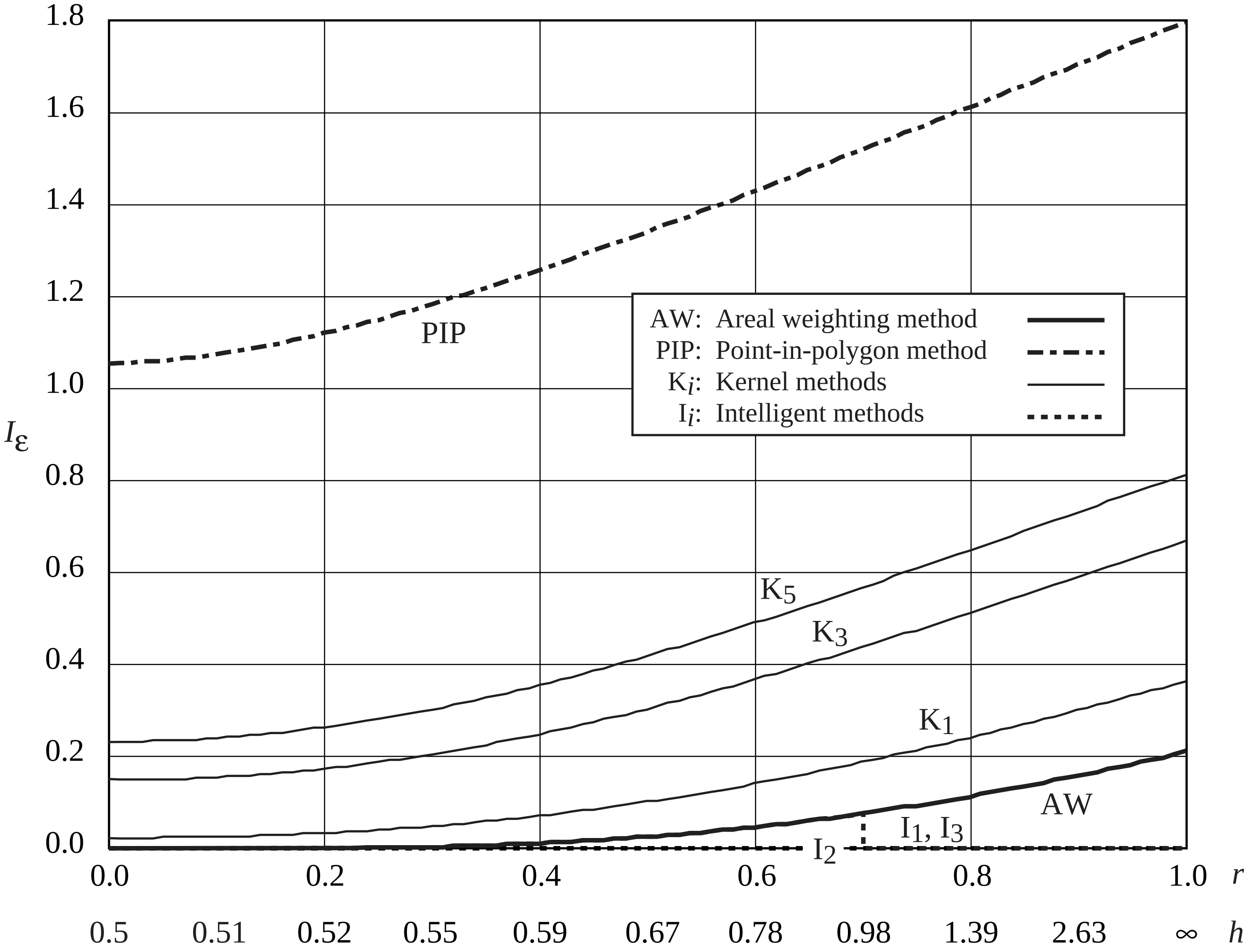
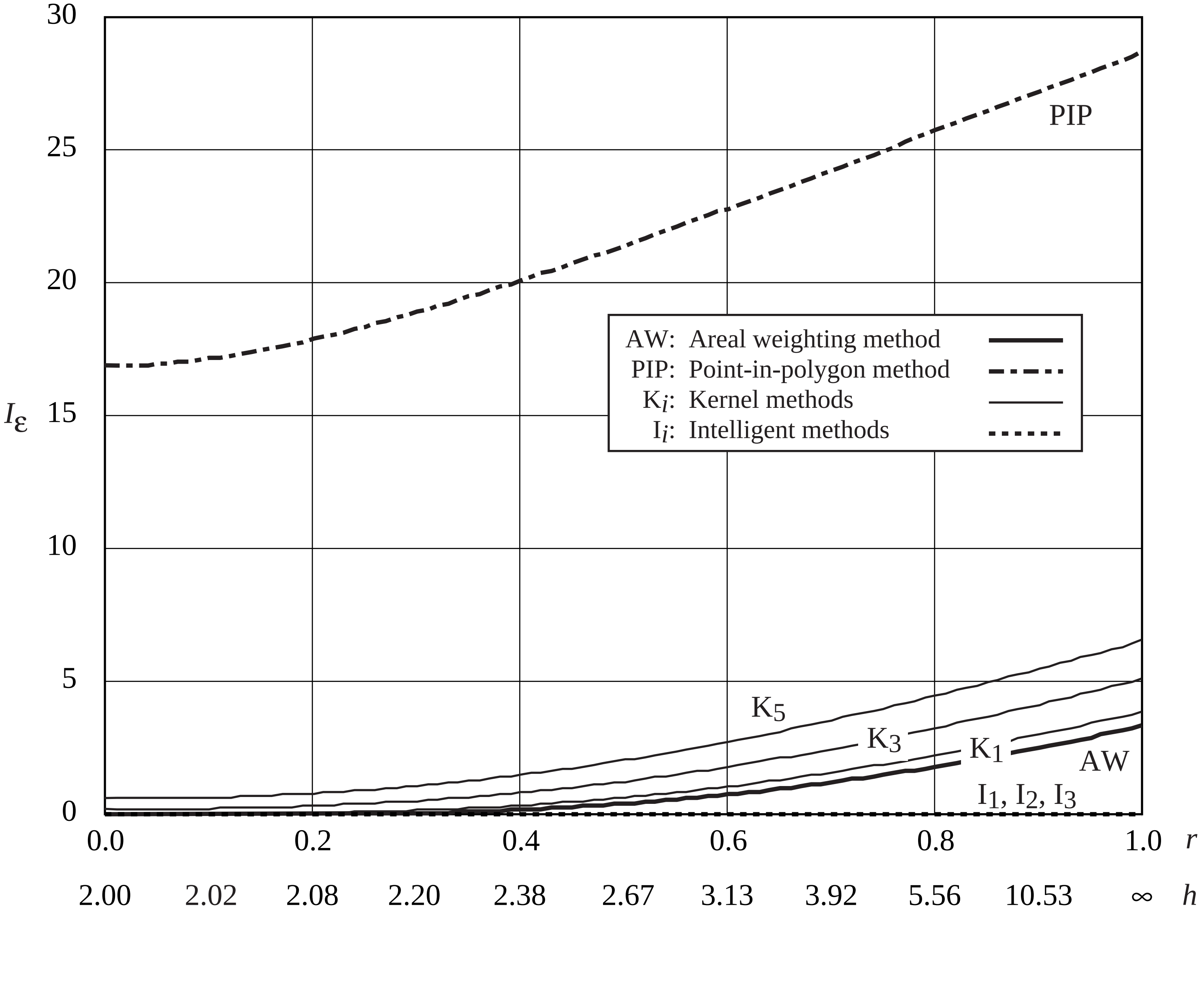

1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
I_ε
PIP
AW: Areal weighting method
PIP: Point-in-polygon method
K_i: Kernel methods
I_i: Intelligent methods
K_5
K_3
K_1
AW
I_1, I_3
I_2
0.0
0.2
0.4
0.6
0.8
1.0
r
0.5
0.51
0.52
0.55
0.59
0.67
0.78
0.98
1.39
2.63
∞
h

Figure 17a



30
25
20
15
10
5
0
I_ε
PIP
AW: Areal weighting method
PIP: Point-in-polygon method
K_i: Kernel methods
I_i: Intelligent methods
K_5
K_3
K_1
AW
I_1, I_2, I_3
0.0
0.2
0.4
0.6
0.8
1.0
r
2.00
2.02
2.08
2.20
2.38
2.67
3.13
3.92
5.56
10.53
∞
h

Figure 17b

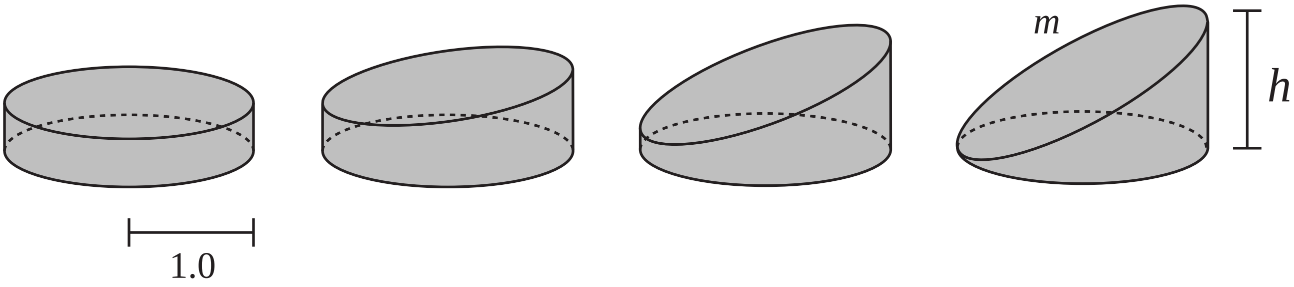


Figure 18

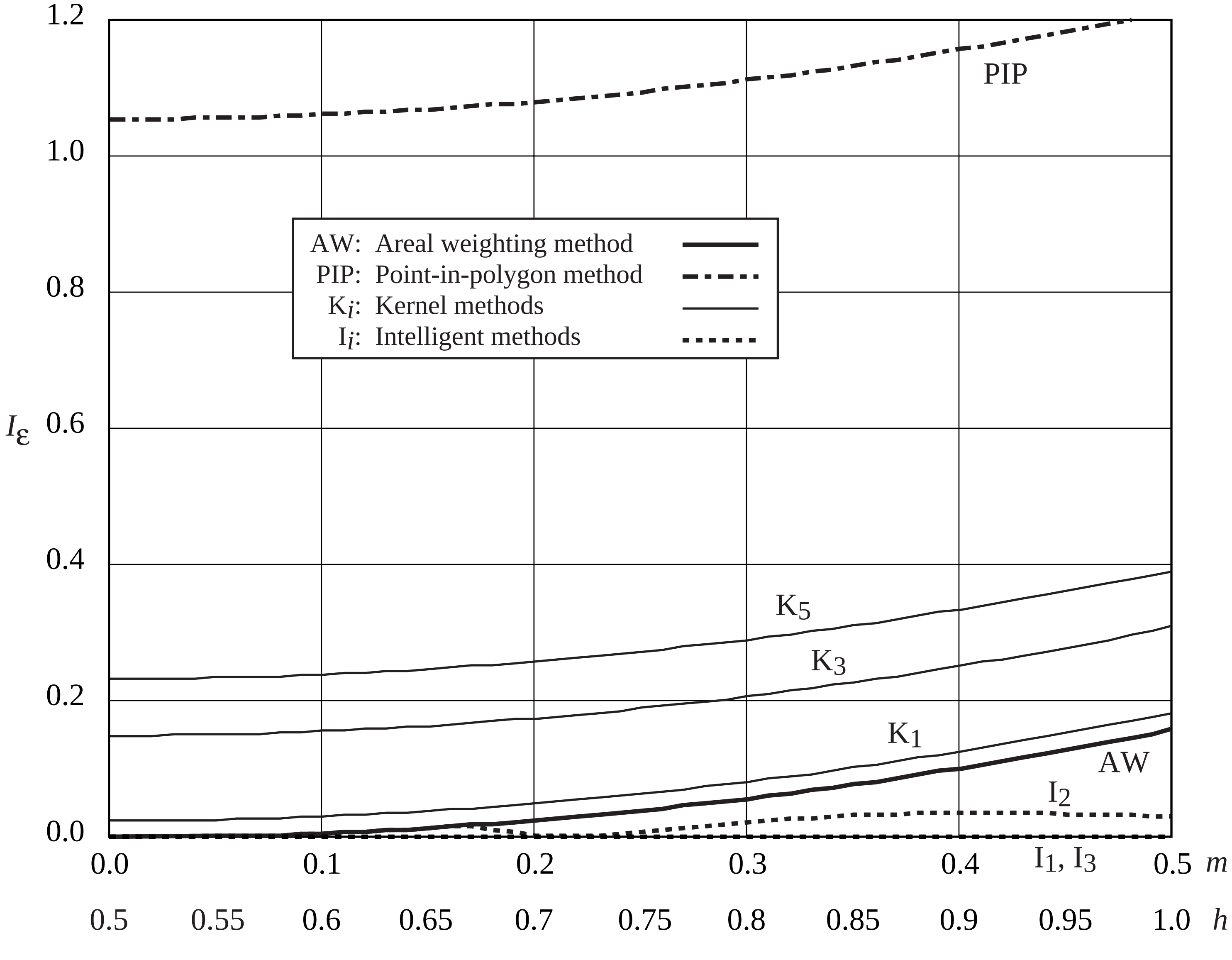

1.2
1.0
0.8
0.6
0.4
0.2
0.0
I_ε
PIP
AW: Areal weighting method
PIP: Point-in-polygon method
K_i: Kernel methods
I_i: Intelligent methods
K_5
K_3
K_1
AW
I_2
I_1, I_3
0.0 0.1 0.2 0.3 0.4 0.5 m
0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1.0 h

Figure 19a

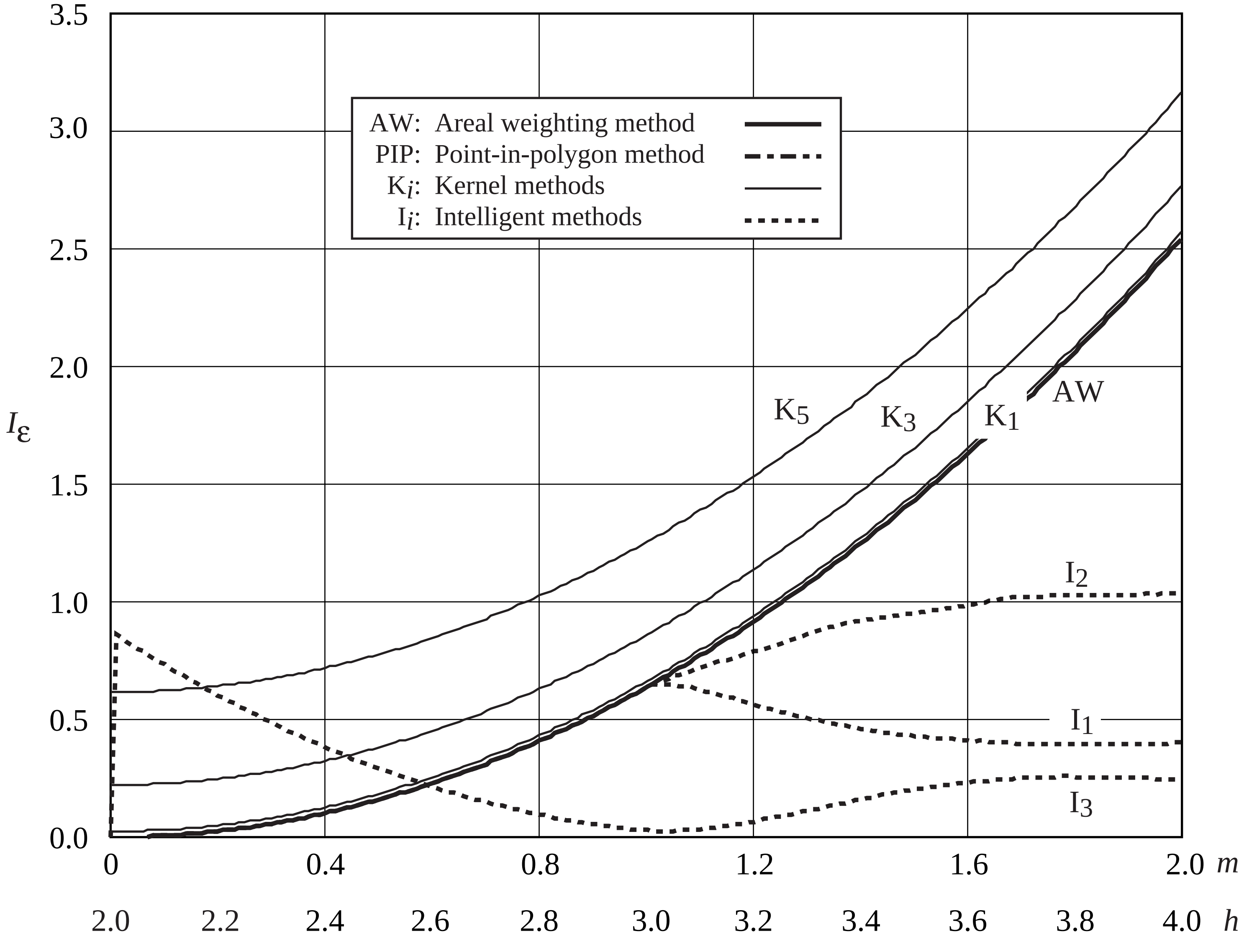

AW: Areal weighting method
PIP: Point-in-polygon method
K_i: Kernel methods
I_i: Intelligent methods
I_ε
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
0.4
0.8
1.2
1.6
2.0
m
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0
h
K_5
K_3
K_1
AW
I_2
I_1
I_3

Figure 19b