CSIS Discussion Paper No. 10

Accuracy of Areal Weighting Interpolation: Effects of Geometrical Properties of Zonal Systems

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 24, 1999

Accuracy of Areal Weighting Interpolation: Effects of Geometrical Properties of Zonal Systems

Keywords: areal interpolation, accuracy, geometrical properties
Short title: accuracy of areal weighting interpolation

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 Weighting Interpolation:
Effects of Geometrical Properties of Zonal Systems

Abstract

This paper analyzes the accuracy of count data transferred through the areal weighting interpolation with respect to the geometrical properties of zonal systems used for aggregating spatial data. A stochastic model is employed to measure the estimation error caused in data transfer between incompatible zonal systems. The relationship between estimation error and the geometrical properties of zones is approximately represented in analytical forms. The major results are as follows: 1) the perimeter of the target zone and the area and perimeter of the source zones are crucial to the accuracy of the areal weighting interpolation; 2) estimation error increases in proportion to the square root of the perimeter of the target zone; 3) concerning the lattice system, estimation error is proportional to the square root of the perimeter and the biquadratic root of the area of the cell, and inversely proportional to the biquadratic root of the number of cells; 4) the hexagonal lattice yields the most accurate estimates among all lattices.

1. Introduction

The areal weighting interpolation is a data transfer procedure between incompatible zonal systems. There are diverse zonal systems used for aggregating and reporting spatial data, say, census tracts, administrative districts, school districts, and so forth. Since they are often geographically incompatible, integration of spatial data requires data transfer between zonal systems. This process is called areal interpolation, and the areal weighting interpolation is one of the most popular interpolation methods in GIS (Markoff and Shapiro, 1973; Lam, 1983; Flowerdew and Green, 1991).

Assuming a uniform distribution of spatial objects, the areal weighting interpolation divides the count of spatial objects according to area in each zone, and sums up the counts in another incompatible zone. Obviously, obtained value is erroneous to some extent because the uniformity assumption does not hold in general (Goodchild *et al.*, 1993; Fisher and Langford, 1995). One method to improve the accuracy of estimates is to use source data based on smaller geographical zones. For instance, census data aggregated across census tracts are more desirable than those aggregated across higher levels of administrative districts, say, towns or counties. On the other hand, detailed data are generally expensive or hard to obtain, and they need much space for storage and high-performance processors for manipulation. Consequently, GIS users have to choose source data balancing the data handling cost and the expected accuracy of estimates.

Such data choice requires us to understand the relationship between the source data and the accuracy of estimates. Sadahiro (1999) proposes a method for analyzing the accuracy of count data estimated by the areal weighting interpolation, and investigates it in relation to the geometrical properties of zones, that is, their shape and size. The obtained results seem quite reasonable: source data consisting of small and convex zones give accurate estimates. In that paper, however, the accuracy measure is represented in an integral form, so that the computational cost reduces the accessibility to the analyzing method and results. For instance, it is not clear what properties are crucial to estimation accuracy. It is also difficult to answer simple questions such as "to what extent the estimation accuracy improves if the number of the source zones increases by twice?"

The motivation for the study described in this paper is to acquire more practical representations of the accuracy of the areal weighting interpolation. The accuracy measure is approximately represented in analytical forms on the basis of the stochastic areal weighting interpolation model developed by Sadahiro (1999). This allows us to understand what geometrical properties of zones are crucial to estimation accuracy, and how they affect the accuracy of the areal weighting interpolation.

2. Areal weighting interpolation model

This paper follows an approach taken by Sadahiro (1999): a stochastic model of the areal weighting interpolation. This section outlines the model and accuracy measure proposed in that paper.

Let S_0 be a region of area A_0 whose shape can cover a plane by its lattice, say, a triangle, a square, or a parallelogram. The region S_0 consists of K zones S_1, S_2, ..., S_K, which represent spatial units used for data aggregation such as census tracts (Figure 1). We call them *source zones* in this paper. The area and perimeter of S_i are denoted by A_i and L_i, respectively. This zonal system is referred to as the *source zonal system* Ω.

Figure 1 An example of the source zonal system Ω.

In the region S_0, N points (say, households) are independently distributed according to the uniform distribution. The location of point j is denoted by $\mathbf{y}_j$ This paper limits the case to the uniform distribution in order to focus on the effects of the geometrical properties of zones on estimation accuracy. We then assume that S_0 is surrounded by its copies, and the copies have the same zonal system and point distribution as those of S_0 (Figure 2). This assumption is called *periodic continuation* which often used in spatial statistics (Ripley 1981, Stoyan and Stoyan 1994).

Figure 2 Periodic continuation assumption.

Let us then consider a *target zone* T that represents a region in which the number of points needs to be estimated. The area and perimeter of T are denoted by A_T and L_T, respectively. We assume that T is fairly larger than the source zones, or to be exact, that T is large enough that the incircle of T is larger than the circumcircle of S_i's. This assumption seems reasonable in areal interpolation, because if T is not larger than S_i' s then it will lead to considerably imprecise interpolation.

The target zone T is dropped in such a way that it intersects S_0. If T does not completely lie in S_0, we replace the portion of T outside S_0 by its corresponding figure as shown in Figure 3. All possible positions of T appear randomly. The location of T is represented by an indicator function:

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

Figure 3 Transformation of T.

In the above stochastic model, the number of points in T and its estimator given by

the areal weighting interpolation are written as

$$M = \sum_j 1_T(\mathbf{y}_j) \tag{2}$$

and

$$\hat{M} = \sum_i \frac{\int_{\mathbf{x}\in S_i} 1_T(\mathbf{x})\mathrm{d}\mathbf{x}}{A_i} \sum_j 1_i(\mathbf{y}_j), \tag{3}$$

respectively, where

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

Estimation accuracy is measured by the mean square error (MSE) defined by

$$MSE[\Omega] = \mathrm{E}\left[\left(\hat{M} - M\right)^2\right]. \tag{5}$$

Substituting equations (2) and (3) into equation (5), we obtain

$$MSE[\Omega] = \frac{N}{A_0}\left\{A_T - \frac{1}{2\pi A_0}\sum_i \frac{1}{A_i}\int_{\mathbf{t}\in S_i}\int_{\mathbf{x}\in S_i} m(T;|\mathbf{x}-\mathbf{t}|)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}\right\} \tag{6}$$

(for details, see Sadahiro 1999), where $m(T; l)$ is the measure of the set of all figures congruent to T containing two points separated by a distance l.

3. Estimation error and geometrical properties of zones

Since equation (6) contains an integral term, calculation of $MSE[\Omega]$ requires numerical integration. Because of its computational cost, the accessibility to the measure is fairly reduced and thus it is not clear from equation (6) what geometrical properties of zones are crucial to estimation accuracy. We hence derive analytical representations of the measure using some approximations.

Instead of $MSE[\Omega]$, this paper adopts the root mean square error (RMSE) as the accuracy measure which is given by

$$\begin{aligned} RMSE[\Omega] &= \sqrt{\mathrm{E}\left[\left(\hat{M} - M\right)^2\right]} \\ &= \sqrt{\frac{N}{A_0}\left\{A_T - \frac{1}{2\pi A_0}\sum_i \frac{1}{A_i}\int_{\mathbf{t}\in S_i}\int_{\mathbf{x}\in S_i} m(T;|\mathbf{x}-\mathbf{t}|)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}\right\}}. \end{aligned} \tag{7}$$

To replace the integral term in equation (7), we employ an approximation

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

(Stoyan and Stoyan, 1994; Sadahiro, 1998). Substitution of equation (8) into equation (7) yields

$$RMSE[\Omega] \approx \sqrt{\frac{\lambda L_T}{\pi A_0}\sum_i A_i D_i}, \tag{9}$$

where λ is the density of points ($=N/A_0$) and D_j is the mean distance between two points

that are randomly distributed in S_j. The distance D_j is given in an analytical form if S_j has a simple shape. For instance, if S_j is a circle of radius r,

$$D_i = \frac{128}{45\pi} r . \tag{10}$$

For a rectangle of sides a, b, we have

$$D_i = \frac{1}{30a^2b^2}\left\{2\left(a^5+b^5\right) - 2\left(a^4 - 3a^2b^2 + b^4\right)\sqrt{a^2+b^2} + 5a^4 b \log\frac{\sqrt{a^2+b^2}+b}{a} + 5ab^4 \log\frac{\sqrt{a^2+b^2}+a}{b}\right\} \tag{11}$$

(Ghosh, 1951). If S_j has more complicated shape, an approximate expression obtained through an empirical study is useful:

$$D_i \approx 0.13 L_i \tag{12}$$

(Koshizuka, 1978). Substituting equation (12) into equation (9) we obtain

$$RMSE[\Omega] \approx \sqrt{\frac{0.13\lambda L_T}{\pi A_0} \sum_i A_i L_i} . \tag{13}$$

This equation indicates that the accuracy of the areal weighting interpolation depends on the perimeter of the target zone and the area and perimeter of the source zones. In other words, the effects of the geometrical properties of zones can be described by these parameters.

Having obtained analytical representations of the estimation error, we now discuss in detail how the geometrical properties of the source and target zones affect estimation accuracy.

3.1 *Geometrical properties of the target zone*

Let us first consider the effects of the target zone T. From equation (9) we notice that estimation error $RMSE[\Omega]$ increases in proportion to the square root of the perimeter of the target zone T. This implies that elongated forms of T yield less precise estimates than convex forms. Suppose, for instance, a rectangular target zone of v/h ratio γ ($\gamma \geq 1$). Then equation (9) becomes

$$RMSE[\Omega] \approx \sqrt{\frac{2\lambda}{\pi A_0}(1+\gamma)\sqrt{\frac{A_T}{\gamma}} \sum_i A_i D_i} . \tag{14}$$

Figure 4 shows the relationship between $RMSE[\Omega]$ and the v/h ratio of the rectangular target zone.

Figure 4 The relationship between $RMSE[\Omega]$ and the v/h ratio of the rectangular target zone. The area of the target zone is fixed. The value of $RMSE[\Omega]$ is standardized so that $RMSE[\Omega]=1.0$ for the square target zone.

Fixing the shape of the target zone, we obtain

$$RMSE[\Omega] \propto \sqrt[4]{A_T}\,. \tag{15}$$

This implies that *RMSE*[Ω] is proportional to the biquadratic root of the area of the target zone (Figure 5). Comparing Figures 4 and 5, we find that the size of the target zone is still more influential than its shape on estimation accuracy.

Figure 5 The relationship between *RMSE*[Ω] and the area of the target zone. The shape of the target zone is fixed. The value of *RMSE*[Ω] is standardized so that *RMSE*[Ω]=1.0 for A_T=1.0.

3.2 *Geometrical properties of the source zones*

We then turn to the effects of the source zones. As we mentioned earlier, they are fully described by equation (13). However, it is somewhat difficult to understand intuitively from the equation how the geometrical properties of the source zones affect *RMSE*[Ω]. We hence consider the lattice as the zonal system, though the lattice system is not so popular in geography. Consideration on the lattice would make the effects of the source zones more clear and easy to understand.

Let *S* be the fundamental cell of the lattice. The area and perimeter of *S* are denoted by A_S and L_S, respectively. Equations (13) then becomes

$$RMSE[\Omega] \approx \sqrt{\frac{0.13\lambda}{\pi} L_T L_S}\,. \tag{16}$$

Equation (16) indicates that the source and target zones are approximately symmetric with respect to the effects on *RMSE*[Ω]. Therefore, the relationships shown in Figures 4 and 5 equally hold for the source zones: estimation error is proportional to the square root of the perimeter of the cell; if the shape of the cell is fixed, the error increases in proportion to the biquadratic root of its area, that is,

$$RMSE[\Omega] \propto \sqrt[4]{A_S}\,. \tag{17}$$

The cell size is even more influential than its shape, which suggests that we only have to pay attention to the size of zones when we choose source data. These results are consistent with those obtained in Cockings *et al*. (1997) and Sadahiro (1999).

Equation (16) also shows that the lattice system of high convexity yields better estimates. Consequently, we can say that the hexagonal lattice is the most desirable among all lattices, which is also compatible with the result obtained in Sadahiro (1999).

We finally examine how the number of cells *K* affects the accuracy of areal interpolation. Fixing the shape of the cell, we have

$$RMSE[\Omega] \propto \frac{1}{\sqrt[4]{K}}\,. \tag{18}$$

This implies that estimation error decreases in proportion to the biquadratic root of the number of cells (Figure 6).

Figure 6 The relationship between $RMSE[\Omega]$ and the number of cells in the lattice system. The shape of the fundamental cell of the lattice is fixed. The value of $RMSE[\Omega]$ is standardized so that $RMSE[\Omega]=1.0$ for $K=1$.

4. Conclusions

In this paper we have analyzed the accuracy of count data estimated by the areal weighting interpolation. Using approximate expressions, we obtained analytical representations of estimation accuracy shown as equations (9), (13), and (16). Since computation of equations (13) and (16) requires only the area and perimeter of the source and target zones, $RMSE[\Omega]$ can easily be obtained in commercial GIS. From these equations we obtained the following results:

1) The perimeter of the target zone and the area and perimeter of the source zones are crucial to the accuracy of the areal weighting interpolation (equation (13)).

2) Estimation error increases in proportion to the square root of the perimeter of the target zone (equation (9)).

3) Concerning the lattice system, estimation error is proportional to the square root of the perimeter (equation(16)) and the biquadratic root of the area (equation(17)) of the cell. Moreover, it is in inverse proportion to the biquadratic root of the number of cells (equation (18)).

4) The hexagonal lattice yields the most accurate estimates among all lattices through the areal weighting interpolation (equation (16)).

References

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. 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.

Ghosh, B., 1951, Random Distances within a Rectangle and between Two Rectangles. *Bulletin of Calcutta Mathematical Society*, **43**, 17-24.

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.

Koshizuka, T., 1978, On the Random Distance within an Area. *Journal of the Operations Research Society of Japan*, **21**, 302-319 (in Japanese).

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

Markoff, J. and Shapiro, G. 1973, The Linkage of Data Describing Overlapping Geographical Units. *Historical Methods Newsletter*, **7**, 34-46.

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

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

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

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

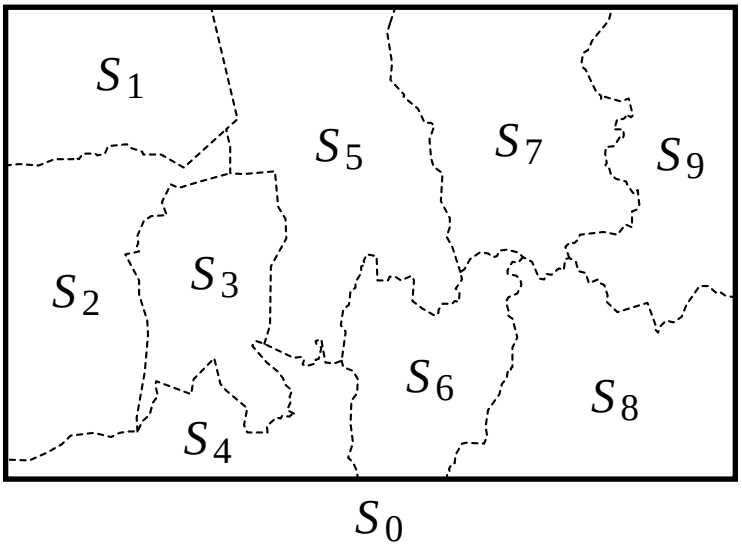

Figure 1

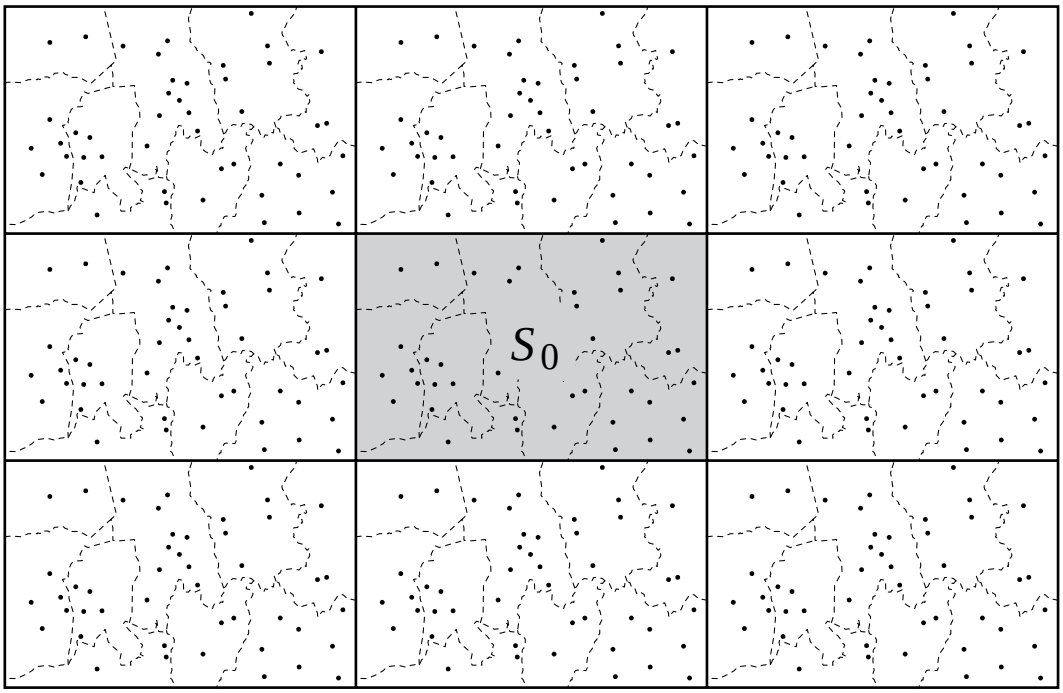

Figure 2

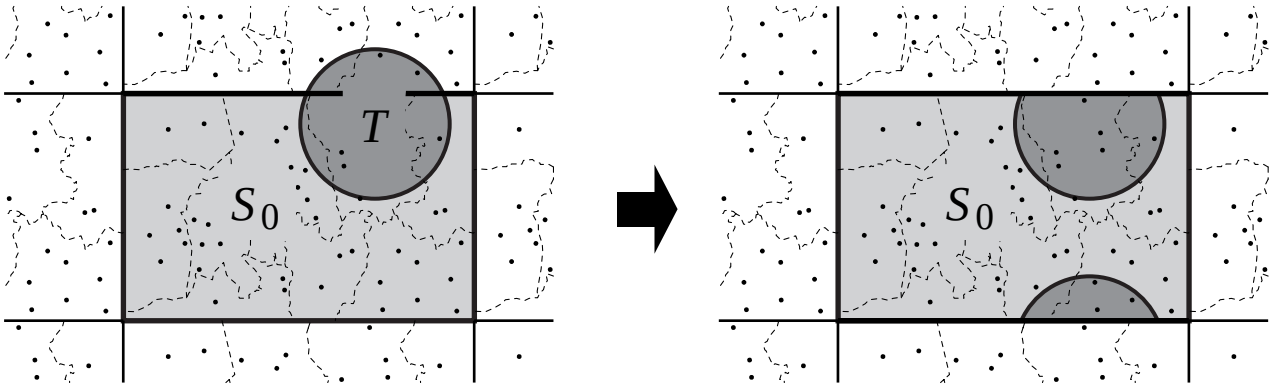

Figure 3

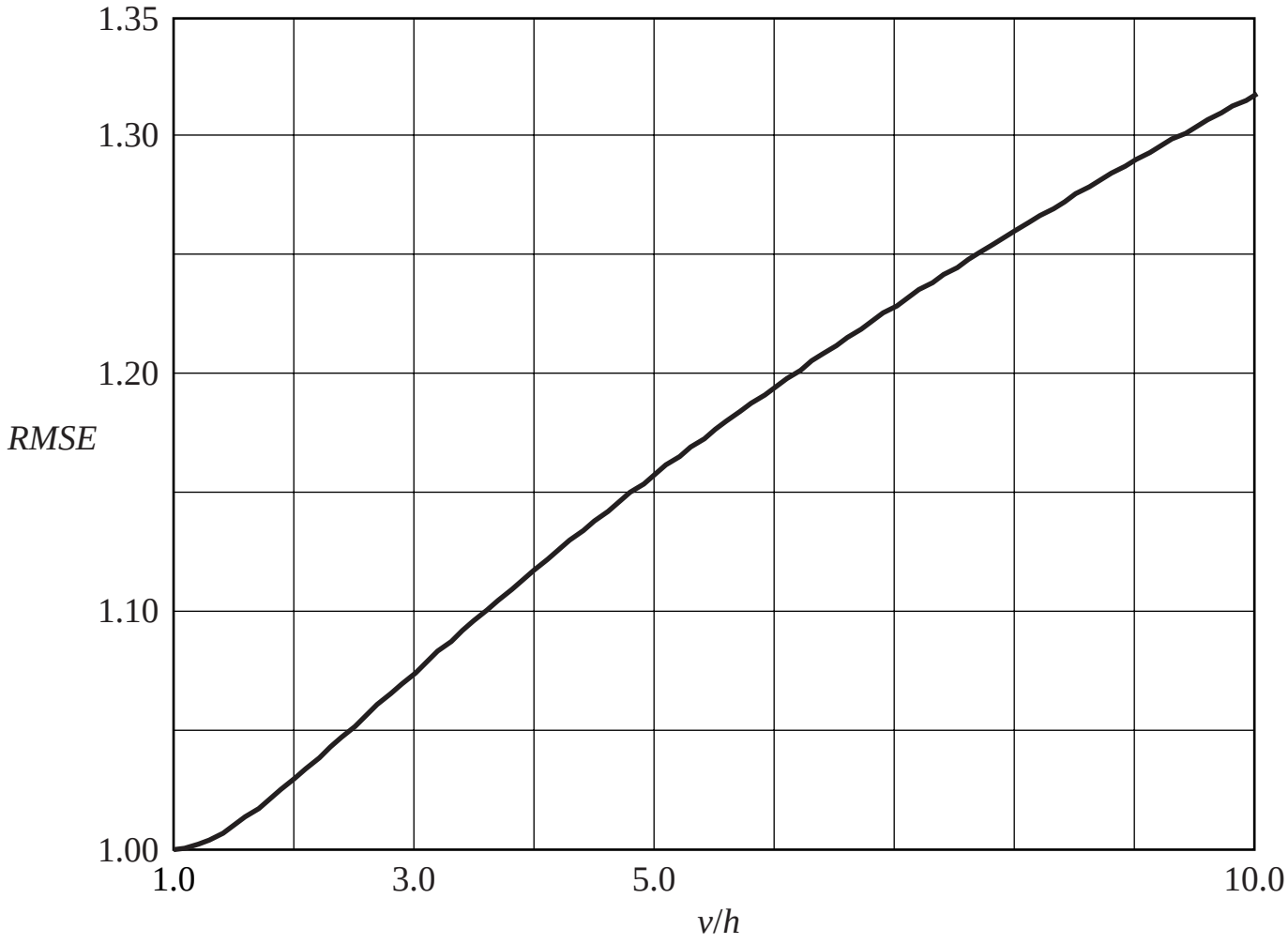

Figure 4

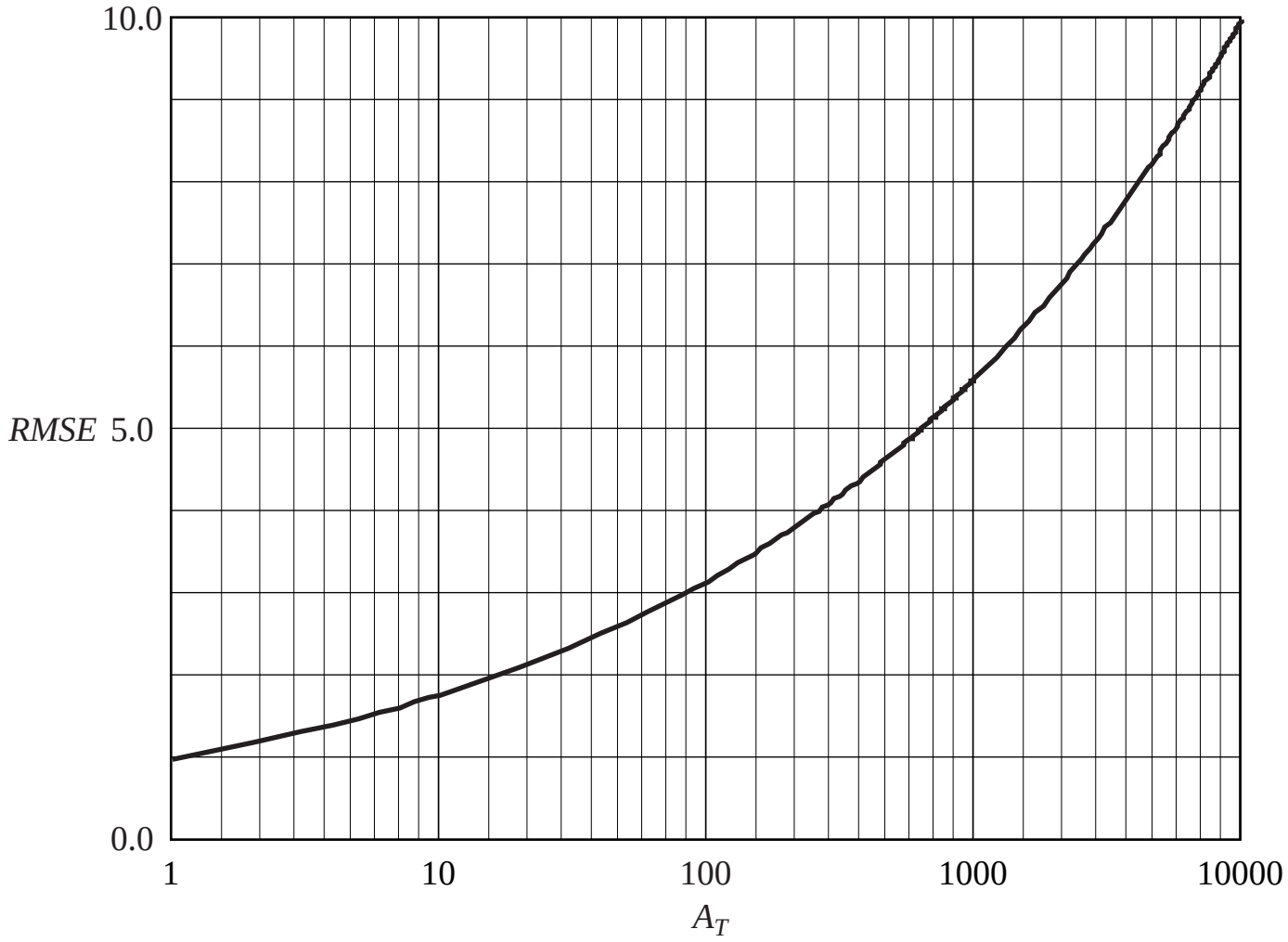

Figure 5

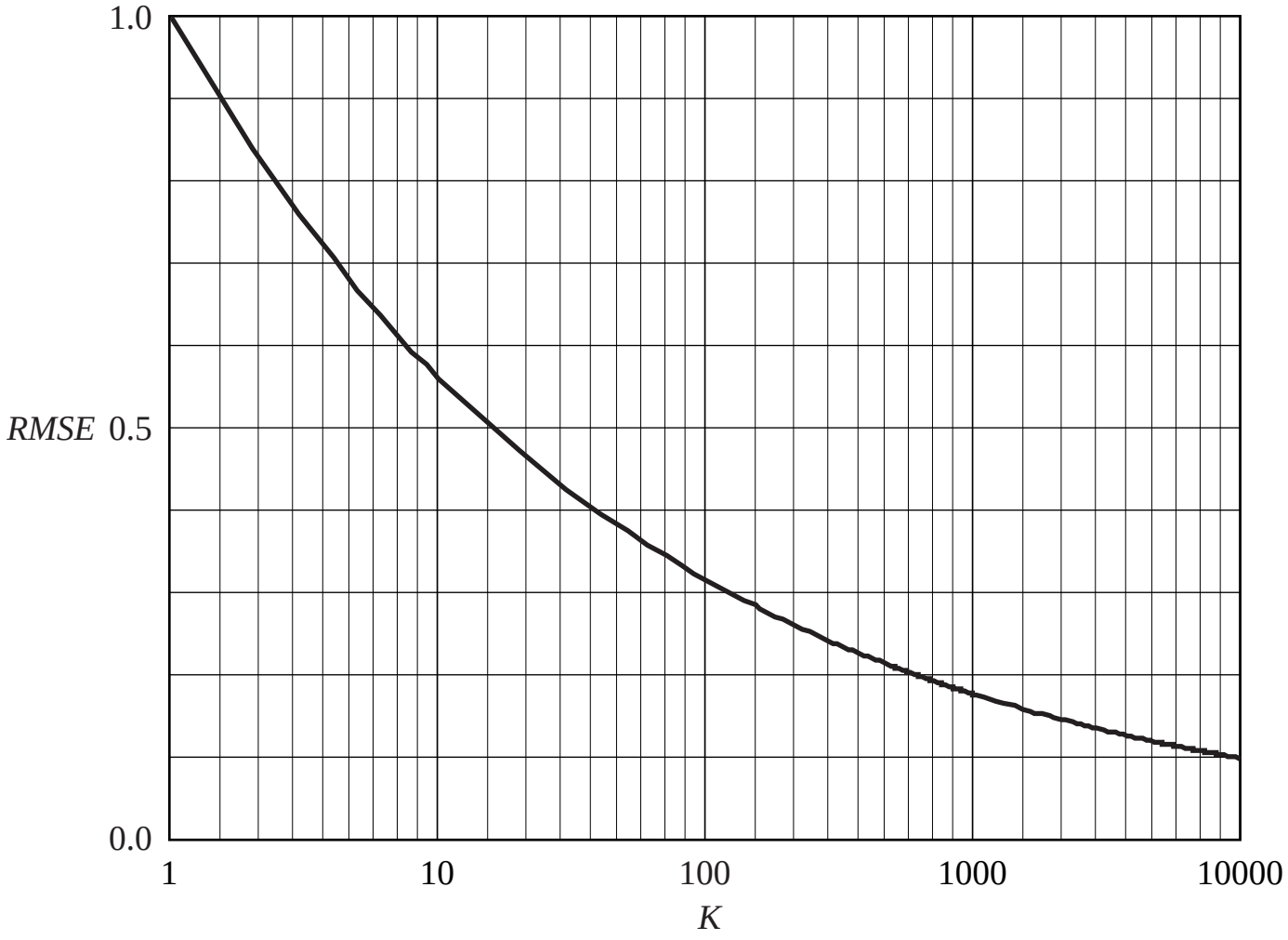

Figure 6