Discussion Paper No. 78

代表点法による空間集計地区変換からみた最適な代表点の位置について

貞広幸雄*

1998 年 9 月

*東京大学空間情報科学研究センター・工学部都市工学科

〒113　東京都文京区本郷 7-3-1

代表点法による空間集計地区変換からみた最適な代表点の位置について

摘要

本論文は，空間単位に基づいて集計された空間データの代表点について，その最適な位置を分析したものである．最も代表的な空間集計地区の変換手法である代表点法を取りあげ，変換時に生ずる誤差が最小となるような代表点の位置について考察を行っている．まずはじめに，分析の前提となる条件を整理し，現実性の高い仮定のもとで，推定値の平均二乗誤差を導出した．そして，点分布と連続分布の 2 種類の分布について，推定誤差が最小となる代表点の位置を求めた．その結果，いずれの場合についても最適な代表点の位置は分布の幾何中央点であることが明らかになった．さらに実証分析から，分布に偏りがある場合には明らかに幾何中央点が最適であり，分布が比較的均一な場合には，分布重心も比較的推定誤差を小さくする位置であることが分かった．計算費用の問題を考え合わせると，実際上はこれら 2 つの位置を適宜使い分けることが望ましいと言える．

キーワード：空間集計，代表点法，推定精度，幾何中央点

I はじめに

地理学において最もよく用いられる空間データの一つとして，空間単位に基づいた集計データがある．人口や世帯に関するデータは，データの機密保持のため，県，市町村，町丁目，1km メッシュなど，様々な空間単位に集計した後に公刊される．また，都市施設や動植物などについても，個々のデータを記述するとデータ量が膨大になる場合には，適当な空間単位ごとにデータを集計する．集計されたデータでは，個々の世帯や施設の位置は秘匿され，空間単位の境界を示すデータと，多くの場合，個々のデータの代表的な位置を表す代表点データの 2 つによって位置の概略が示される．

このように空間集計されたデータは，その集計単位が分析に適していればそのまま利用できるが，データと分析の集計単位が異なる場合にはデータの変換が必要になる．例えば，データが町丁目単位で与えられており，分析を最寄り駅に基づいた地区ごとに行うという場合，空間集計地区の変換が行われる．この操作は面補間（areal interpolation）と呼ばれ，その代表的な方法として代表点法がある（Goodchild and Lam 1980; Lam 1983; Okabe and Sadahiro 1997; Burrough and McDonnell 1998）．この方法では，元の集計データに新たな集計領域（以下，集計先地区）を重ね，その中に含まれる全ての代表点に割り当てられている数値を合計する．例えば図 1a のような元データに対して，図 1b のように集計先地区を重ねた場合，図 1c のような計算によって集計先地区内の数値が与えられる．この方法は面補間の方法の中で最も単純であるため，データ量が非常に多い場合に有効であり（Preparata and Shamos 1985），実際，国勢調査データのメッシュデータへの変換にはこの方法が用いられている（総務庁 1994）．

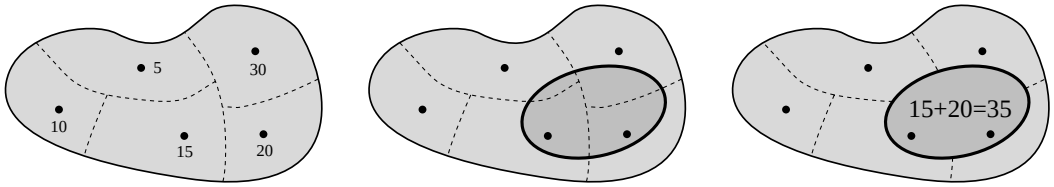


図 1　代表点法による空間集計地区変換

代表点法で推定されたデータは，通常，誤差を伴う．データ利用の観点からみると，推定誤差は小さいほど望ましく，データ提供者にとっても大きな誤差の生ずるような集計方法は好ましくない．推定誤差はいくつかの要因に基づいて発生し（Sadahiro, 1998b），それらはデータ利用段階において利用者の選択により改善可能なものと，データ作成段階において改善可能なものの 2 つに分類することができる．前者の例としては空間単位の大きさがあり，空間単位の小さなデータを選択することで推定精度を高くすることが可能である．後者の例としては代表点の位置があり，適切な位置を選択することで推定誤差を小さくすることができる．本論文ではこれら誤差要因のうち，特に代表点の位置に着目し，誤差最小化という観点から最適な代表点の位置について分析を行う．元データとしては，主に人口や世帯などを想定し, 点分布として表される場合と連続分布として表される場合の 2 つを考える．後者については，例えば大気中のある物質の濃度のように，真に連続な分布に対しても適用である．

以下，2 章では空間集計された点分布について，代表点法による推定誤差を導出する手法を提案し，それを用いて最適な代表点の位置を求める．3 章では，同様の手法を連続分布に対して適用する．4 章では，いくつかの代表的な連続分布について，実際に最適な代表点の位置を算出する．5 章では，結論と今後の課題をまとめる．

II 点分布における最適な代表点の位置

本論文では，元の空間集計地区の一つ（以下，集計元地区）S に着目し，そこに集計先地区 T が重ねられた場合の重複部分における推定誤差を最小化するという観点から， S の代表点 $\mathbf{z}$ の最適な位置を考える．推定誤差は，S の形状と大きさ，点分布，代表点の位置，及び，集計先地区 T の形状・面積・位置の 4 つの要因によって決定される．これらの要因のうち，代表点の位置決定に際しては前 2 者が既に与えられており，また推定誤差から最適な代表点の位置を定めるには，推定誤差の計算のために代表点の位置も与える必要がある．従って，これら 3 つの要因は全て所与のものとする．集計先地区 T の形状・面積・位置は，通常，データ作成段階では不確定である．しかし，推定誤差はこれらの要因に大きく依存し，未定のまま議論を進めることは困難である．そこで当面，集計先地区 T の形

状・面積は所与としておこう（面積を B とする）．また T の位置については，集計元地区 S とのあらゆる位置関係が起こりうるということを鑑み，空間集計地区 S と少なくとも一部分が重なるようにランダムに置かれるものとする．

以上の議論に基づき，ここでは次のような設定を考える．集計元地区 S には N 個の点 1, 2, ..., N が分布しており，それぞれの位置を $\mathbf{p}_1$, $\mathbf{p}_2$, ..., $\mathbf{p}_N$ と表す．$\mathbf{z}$ に位置する代表点には，点の個数 N が属性値として割り当てられている．そして，集計元地区 S に対して集計先地区 T が重ねられ，S と T の重複部分に含まれる（真の）点の個数を M とおく．M の値の推定は代表点法によって行われる．即ち，T が代表点 $\mathbf{z}$ を内包すれば N が，内包しなければ 0 が，それぞれ M の推定値として与えられる．

このような設定のもとで，代表点法の推定誤差，即ち，M の真の値と推定値を定式化しよう．いま，関数 $C(\mathbf{x})$ を

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

と定義する．すると，真の個数 M は

$$M = \sum_i C(\mathbf{p}_i) \tag{2}$$

と表される．一方，代表点法による M の推定値は

$$\hat{M} = NC(\mathbf{z}) \tag{3}$$

となる．従って，代表点法の推定誤差は式 2 と式 3 の差によって与える．

もちろん，代表点法による推定誤差は，集計先地区 T の位置によって異なる．そこでここでは，全ての T の位置に関する平均的な誤差の程度を表すために，推定値の平均二乗誤差を用いる．この値は具体的には以下の式で与えられる．

$$\begin{aligned} E\left[\varepsilon^2\right] &= E\left[\left(M - \hat{M}\right)^2\right] \\ &= E\left[M^2 - 2M\hat{M} + \hat{M}^2\right] \\ &= E\left[\left\{\sum_i C(\mathbf{p}_i)\right\}^2\right] - 2NE\left[\sum_i C(\mathbf{p}_i)C(\mathbf{z})\right] + N^2 E\left[\left\{C(\mathbf{z})\right\}^2\right] \end{aligned} \tag{4}$$

この式を展開すると以下のようになる．

$$\begin{aligned} E\left[\left\{\sum_i C(\mathbf{p}_i)\right\}^2\right] &= \sum_i \sum_j E\left[C(\mathbf{p}_i)C(\mathbf{p}_j)\right] \\ &= \sum_i \sum_j \Pr\left[\mathbf{p}_i \cup \mathbf{p}_j \in T\right] \end{aligned} \tag{5}$$

一方第 2 項は，

$$E\left[\sum_i C(\mathbf{p}_i)C(\mathbf{z})\right] = \sum_i E\left[C(\mathbf{p}_i)C(\mathbf{z})\right] = \sum_i \Pr\left[\mathbf{p}_i \cup \mathbf{z} \in T\right] \tag{6}$$

最後に第 3 項は,

$$E\left[\{C(\mathbf{z})\}^2\right] = E\left[C(\mathbf{z})\right] = \Pr\left[\mathbf{z} \in T\right] \tag{7}$$

となる．従って式 4 は,

$$E\left[\varepsilon^2\right] = \sum_i \sum_j \Pr\left[\mathbf{p}_i \cup \mathbf{p}_j \in T\right] - 2N\sum_i \Pr\left[\mathbf{p}_i \cup \mathbf{z} \in T\right] + N^2 \Pr\left[\mathbf{z} \in T\right] \tag{8}$$

となる．

式 8 は，集計先地区 T の位置がランダムに与えられる場合の平均的な誤差の程度を表している．T の位置の特定が困難であることを考えると，この値をできるだけ小さくするような代表点の位置が望ましいと言える．即ち,

$$\min_{\mathbf{z}} E\left[\varepsilon^2\right] \tag{9}$$

の解が，集計元地区 S の最適な代表点の位置であると考えて良いであろう．

式 8 において，第 1 項は所与であり，第 3 項は代表点 $\mathbf{z}$ の位置によらず一定である．従って第 2 項だけが制御可能な項であり，上の問題は

$$\max_{\mathbf{z}} \sum_i \Pr\left[\mathbf{p}_i \cup \mathbf{z} \in T\right] \tag{10}$$

と置き換えて良い．この式は即ち，各点と代表点が同時に T に内包される確率ができるだけ高くなるように代表点を選ぶと良い，ということを示している．

式 10 のうち，$\Pr[\mathbf{p}_i \cup \mathbf{z} \in T]$はさらに以下のように書き換えることができる．

$$\Pr\left[\mathbf{p}_i \cup \mathbf{z} \in T\right] = \frac{m\left(T; |\mathbf{p}_i - \mathbf{z}|\right)}{m'(T; Z)} \tag{11}$$

但し，$m(T;\ |\mathbf{p}_i\text{-}\mathbf{z}|)$は距離が$|\mathbf{p}_i\text{-}\mathbf{z}|$だけ離れた 2 点を内包する T と合同な図形の集合の測度，$m(T;\ Z)$は Z と少なくとも一部分重なる T と合同な図形の集合の測度である（Santaló 1976）．後者は代表点 $\mathbf{z}$ の位置とは無関係であるので，式 10 は

$$\max_{\mathbf{z}} \sum_i m\left(T; |\mathbf{p}_i - \mathbf{z}|\right) \tag{12}$$

となる．ここで $m(T; l)$は，T が簡単な図形の場合には明示的に与えられる（Santaló 1976; Sadahiro 1998a,

1999)．例えば，T が半径 r の円の場合には，

$$m(T;l)=\begin{cases}4\pi r^2\arccos\left(\dfrac{l}{2r}\right)-\pi l\sqrt{4r^2-l^2} & (l\le 2r),\\ 0 & (l>2r).\end{cases} \tag{13}$$

である．また，T が短辺 b，長辺 c の長方形の場合には，

$$m(T;l)=\begin{cases}2\pi bc-4(b+c)l+2l^2 & (l\le b),\\ 4c\sqrt{l^2-b^2}-4cl-2b^2+4bc\arcsin\dfrac{b}{l} & (b<l\le c),\\ 4c\sqrt{l^2-b^2}+4b\sqrt{l^2-c^2}-2\left(b^2+c^2+l^2\right)+4bc\left(\arcsin\dfrac{c}{l}-\arccos\dfrac{b}{l}\right) & \left(c<l\le\sqrt{b^2+c^2}\right),\\ 0 & \left(\sqrt{b^2+c^2}<l\right).\end{cases} \tag{14}$$

である．さらに複雑な図形の場合には，測度 $m(T;\ l)$は以下の式を用いて数値的に計算することができる．

$$m(T;l)=\frac{l}{B^2}g_T(l) \tag{15}$$

但し，$g_T(l)$は T 内にランダムに 2 点をとったときの点間距離 l の確率密度関数である．従って，最悪の場合でも式 12 は数値的に解くことが可能である．即ち，点分布と集計先地区 T が与えられれば，式 12 を用いて最適な代表点の位置を導出することが可能である．

以上の議論では，集計先地区 T の形状・面積は所与であると仮定してきた．そして式 12 は，最適な代表点の位置が T の形状・面積に依存するということを示している．しかし前述したように，T の形状や面積はデータ作成段階，即ち，代表点の位置を決定するときには不確定であり，それらを特定して式 12 を解くことは実際上あまり意味を持たない．そこで次に，集計先地区 T について多少緩やかな以下の仮定を考えてみよう．

1) 集計先地区 T は円形である．

2) 集計先地区 T は集計元地区 S と比べて十分大きい．具体的には，円の直径 $2r$ が S の絶対最大長（単位地区の 2 点間距離の最大値）よりも大きい．

円形は，集計先地区 T の形状としては最も典型的であり，この仮定が当てはまる場合は決して少なくはないであろう．また，2 番目の仮定も十分に妥当であると考えられる．なぜならば，T の大きさが S と比べてそれほど大きくない場合には，推定誤差が非常に大きくなる可能性が高く（Sadahiro 1998b），空間集計地区変換，特に，代表点法による変換はほとんど行われないためである．

これら 2 つの仮定が満たされる場合には，測度 $m(T;\ l)$は

$$m\left(T;|\mathbf{p}_i-\mathbf{z}|\right)=4\pi r^2\arccos\left(\frac{|\mathbf{p}_i-\mathbf{z}|}{2r}\right)-\pi|\mathbf{p}_i-\mathbf{z}|\sqrt{4r^2-|\mathbf{p}_i-\mathbf{z}|^2} \tag{16}$$

とおくことができる．ここで，

$$M(l)=4\pi r^2\arccos\left(\frac{l}{2r}\right)-\pi l\sqrt{4r^2-l^2} \tag{17}$$

とおき，l=0 のまわりで 2 次の項まで Taylor 展開すると，

$$\begin{aligned} M(l) &\approx M(0)+M'(0)l+\frac{1}{2}M''(0)l^2 \\ &= 2\pi^2r^2-4\pi rl \end{aligned} \tag{18}$$

となる．この近似は，l が小さいときには非常によい近似である（図 2 参照）．そこで，式 18 を式 12 に代入すると，

$$\begin{aligned} \max_{\mathbf{z}}\sum_i m\left(T;|\mathbf{p}_i-\mathbf{z}|\right) &\Leftrightarrow \max_{\mathbf{z}}\sum_i\left(2\pi^2r^2-4\pi r|\mathbf{p}_i-\mathbf{z}|\right) \\ &\Leftrightarrow \min_{\mathbf{z}}\sum_i|\mathbf{p}_i-\mathbf{z}| \end{aligned} \tag{19}$$

を得る．

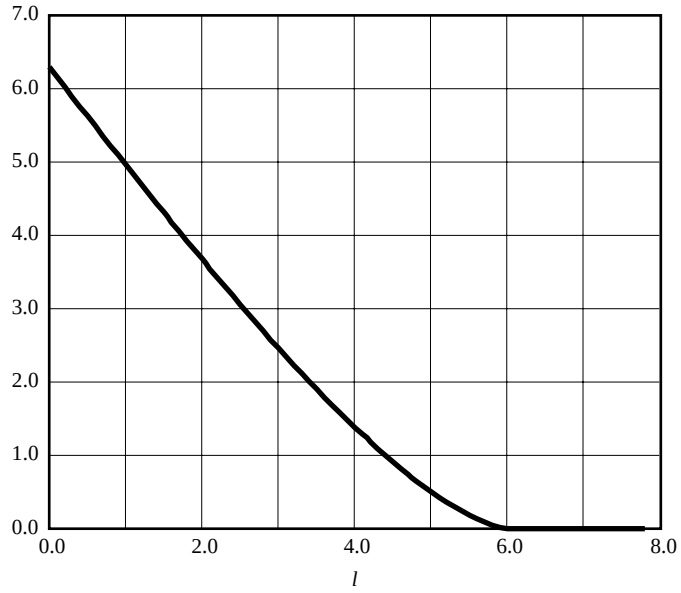


図 2　円に関する測度 $m(T; l)$

式 19 は，最適な代表点の位置が点 1, 2, ..., N の幾何中央点（geometrical median, spatial median）であることを意味している（Brown, 1983; Small, 1990）．点分布の幾何中央点の導出は，OR における Weber 問題と同等であり，具体的な方法は，Cooper (1968)，Francis and White (1974)，Love et al. (1988)，Plastria

(1995)などに詳しい．従って，既存の方法を用いることで最適な代表点を具体的に求めることが可能である．

ところで上の結果は，式 18 にあるように， l の 1 次式による $m(T;\ l)$の近似が妥当であれば，集計先地区 T の形状によらず成立する．実際に，例えば T が長方形の場合，縦横比がそれほど大きくなければこの近似は可能であり，また，T が正三角形や正六角形の場合でも同様である（図 3 参照）．一般的には，T が円形度の高い凸多角形の場合には，この近似は妥当であると言える（Sadahiro, 1998b）．もちろん，集計先地区 T の形状は必ずしも凸多角形とは限らない．しかし，円形に代表される凸多角形は，T の形状としては非常によく用いられる．従って，

1) 集計先地区 T は円形度の高い凸多角形である．
2) 集計先地区 T は集計元地区 S と比べて十分大きい．

という 2 つの仮定が成立すれば，最適な代表点の位置は点分布の幾何中央点であると言って良いであろう．

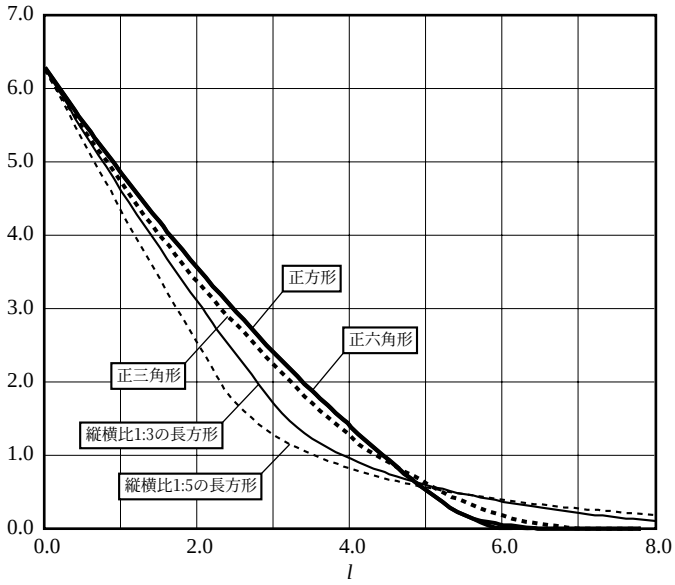


図 3　様々な図形に関する測度 $m(T;\ l)$

III 連続分布における最適な代表点の位置

本章では，ある集計元地区 S 内に連続分布，即ち，S 内のある位置に対して値が一意に定められる関数を想定し，代表点法による推定誤差の最小化という観点から，地区 S の最適な代表点 $\mathbf{z}$ の位置を考える．この場合，推定誤差は，S の形状と大きさ，連続分布，代表点の位置，及び，集計先地区 T の形状・面積・位置の 4 つの要因によって決定されるが，ここでも前章と同様に，前 3 者及び集計先

地区 T の形状・面積を所与とし， T の位置については，空間集計地区 S と少なくとも一部分が重なるようにランダムに置かれるものとする．

具体的には，以下のような設定を行う．集計元地区 S では，全ての位置 $\mathbf{x}$ について関数 $f(\mathbf{x})$が定義され，その S 内での積分値が $\mathbf{z}$ に位置する代表点に割り当てられている．そして，集計元地区 S に対して集計先地区 T が重ねられ，S と T の重複部分における $f(\mathbf{x})$ の積分値を Q とおく．Q の値の推定値としては，T が代表点 $\mathbf{z}$ を内包すれば $f(\mathbf{x})$ の S 内での積分値が，内包しなければ 0 が与えられる．

このような設定のもとでは，Q の真の値と推定値はそれぞれ以下のように表される．

$$\begin{aligned} Q &= \int_{\mathbf{x}\in S\cap T} f(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int_{\mathbf{x}\in S} C(\mathbf{x}) f(\mathbf{x})\mathrm{d}\mathbf{x} \end{aligned} \tag{20}$$

$$\hat{Q} = C(\mathbf{z})\int_{\mathbf{x}\in S} f(\mathbf{x})\mathrm{d}\mathbf{x} \tag{21}$$

前章と同様，ここでも推定値の平均二乗誤差を求めると，

$$\begin{aligned} E\left[\varepsilon^2\right] = &\int_{\mathbf{t}\in S}\int_{\mathbf{x}\in S} \Pr[\mathbf{x}\cup\mathbf{t}\in T] f(\mathbf{x}) f(\mathbf{t})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ &-2\int_{\mathbf{x}\in S} f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in S} \Pr[\mathbf{x}\cup\mathbf{z}\in T] f(\mathbf{x})\mathrm{d}\mathbf{x} \\ &+\Pr[\mathbf{z}\in T]\left\{\int_{\mathbf{x}\in S} f(\mathbf{x})\mathrm{d}\mathbf{x}\right\}^2 \end{aligned} \tag{22}$$

となる．この値を最小化するには，

$$\max_{\mathbf{z}} \int_{\mathbf{x}\in S} \Pr[\mathbf{x}\cup\mathbf{z}\in T] f(\mathbf{x})\mathrm{d}\mathbf{x} \tag{23}$$

を解けばよい．この問題はさらに，式 11 を代入すると，

$$\max_{\mathbf{z}} \int_{\mathbf{x}\in S} m\left(T;|\mathbf{x}-\mathbf{z}|\right) f(\mathbf{x})\mathrm{d}\mathbf{x} \tag{24}$$

と置き換えられ，式 12 を連続分布にそのまま拡張した問題になる．従って，集計先地区 T が単純な形状であれば解析的に，そうでなくとも数値的に最適な代表点の位置を導出することが可能である．

次に，先程と同様に，集計先地区 T が集計元地区 S と比べて十分大きい凸多角形である場合について検討しよう．この場合，式 16～18 を適用し，

$$\min_{\mathbf{z}} \int_{\mathbf{x}\in S} |\mathbf{x}-\mathbf{z}| f(\mathbf{x})\mathrm{d}\mathbf{x} \tag{25}$$

を得る．即ち，最適な代表点の位置は関数 $f(\mathbf{x})$の幾何中央点である．但し，具体的な幾何中央点の導出は，点分布の場合ほど容易ではない．$f(\mathbf{x})$が S 内の一部の円領域や長方形領域で一定値を持つ，といった特別な場合については，いくつか有効な解法アルゴリズムが示されているが（Love 1972; Drezner and Wesolowsky 1980; Aly and Marucheck 1982; Plastria 1995），それ以外の場合については悉皆法を用い

ることになる．

IV 実証分析例

本章では，前章の結果に基づき，いくつかの具体的な連続分布について最適な代表点の位置を導出し，それぞれの場合の推定値の平均二乗誤差を算出する．集計元地区 S は最も典型的な形状である正方形，集計先地区 T は S よりも十分大きな凸多角形を想定する．集計元地区 S は 10×10 の正方形格子網で分割し，各セルに関数 $f(\mathbf{x})$の値を一定値として与える．即ち，関数 $f(\mathbf{x})$の各セル内での値は一定であり，S 内では階段関数となる．具体的な $f(\mathbf{x})$の値は図 4 の通りである．

この場合，推定値の平均二乗誤差は，式 11 及び 22 より

$$\begin{aligned} E\left[\varepsilon^2\right] = & \int_{\mathbf{t}\in S}\int_{\mathbf{x}\in S}\frac{m\left(T;|\mathbf{x}-\mathbf{t}|\right)}{m'\left(T;Z\right)}f(\mathbf{x})f(\mathbf{t})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ & -2\int_{\mathbf{x}\in S}f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in S}\frac{m\left(T;|\mathbf{x}-\mathbf{z}|\right)}{m'\left(T;Z\right)}f(\mathbf{x})\mathrm{d}\mathbf{x} \\ & +\Pr[\mathbf{z}\in T]\left\{\int_{\mathbf{x}\in S}f(\mathbf{x})\mathrm{d}\mathbf{x}\right\}^2 \end{aligned} \tag{26}$$

と表される．ここで，

$$\Pr[\mathbf{z}\in T]=\frac{2\pi B}{m'\left(T;Z\right)} \tag{27}$$

及び

$$m'\left(T;Z\right)=2\pi(A+B)+L_S L_T \tag{28}$$

（A は S の面積，L_S 及び L_T はそれぞれ S と T の周長を表す）を用いると，式 26 は

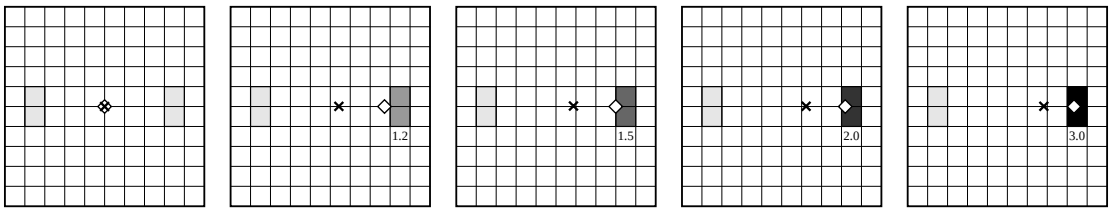
$$E\left[\varepsilon^2\right]=\frac{1}{2\pi(B+A)+L_S L_T}\begin{Bmatrix} \int_{\mathbf{t}\in S}\int_{\mathbf{x}\in S}m\left(T;|\mathbf{x}-\mathbf{t}|\right)f(\mathbf{x})f(\mathbf{t})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ -2\int_{\mathbf{x}\in S}f(\mathbf{x})\mathrm{d}\mathbf{x}\int_{\mathbf{x}\in S}m\left(T;|\mathbf{x}-\mathbf{z}|\right)f(\mathbf{x})\mathrm{d}\mathbf{x} \\ +2\pi B\left\{\int_{\mathbf{x}\in S}f(\mathbf{x})\mathrm{d}\mathbf{x}\right\}^2 \end{Bmatrix} \tag{29}$$

となる．但しここでは，関数 $f(\mathbf{x})$の絶対値による影響を排除するために，平均二乗誤差を基準化して用いている．即ち，

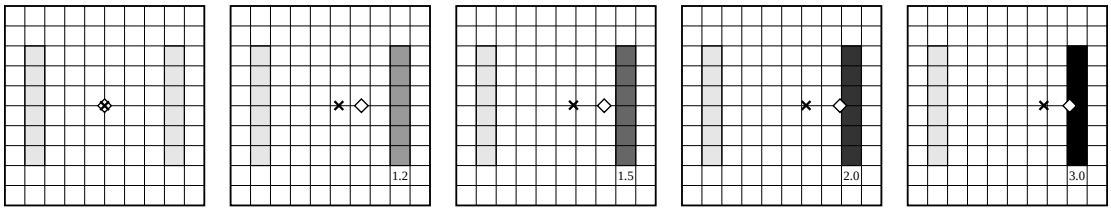
$$\frac{E\left[\varepsilon^2\right]}{\left\{\int_{\mathbf{x}\in S} f(\mathbf{x})\mathrm{d}\mathbf{x}\right\}^2} = \frac{1}{2\pi(B+A)+L_S L_T}\left\{\begin{array}{l} \dfrac{1}{\left\{\int_{\mathbf{x}\in S} f(\mathbf{x})\mathrm{d}\mathbf{x}\right\}^2}\int_{\mathbf{t}\in S}\int_{\mathbf{x}\in S} m\left(T;|\mathbf{x}-\mathbf{t}|\right) f(\mathbf{x}) f(\mathbf{t})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\ -2\dfrac{1}{\int_{\mathbf{x}\in S} f(\mathbf{x})\mathrm{d}\mathbf{x}}\int_{\mathbf{x}\in S} m\left(T;|\mathbf{x}-\mathbf{z}|\right) f(\mathbf{x})\mathrm{d}\mathbf{x} \\ +2\pi B \end{array}\right\} \quad (30)$$

である．なお，式 30 からも明らかな通り，平均二乗誤差の算出には S と T を具体的に与える必要がある．そこで，S としては一辺 1.0 の正方形，集計先地区 T は半径 2.0 の円をそれぞれ仮定した．

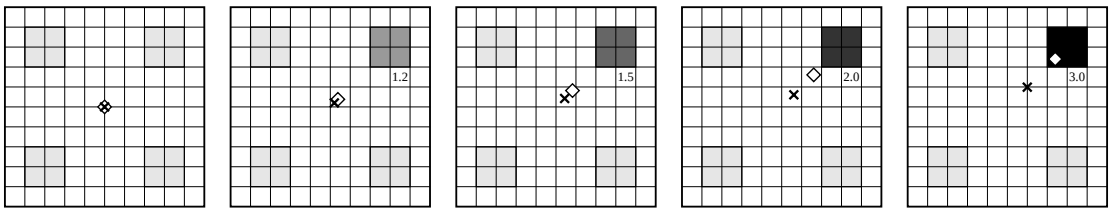
最適な代表点，即ち幾何中央点の導出には，悉皆法を用いている．またここでは比較のために，代表点が分布重心或いは地区重心（セルの中心）に位置している場合の平均二乗誤差も併せて示す．これらの位置はいずれも代表点として良く用いられており，比較によってそれぞれの妥当性を検討することができる．



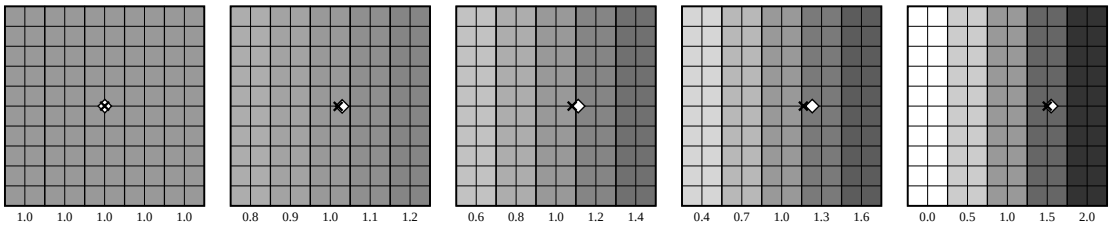
	座標	平均二乗誤差								
幾何中央点	(50, 50)	0.05896	(77, 50)	0.05257	(80, 50)	0.04329	(82, 50)	0.03290	(83, 50)	0.02221
分布重心	(50, 50)	0.05896	(54, 50)	0.05829	(58, 50)	0.05586	(62, 50)	0.05156	(68, 50)	0.04204
地区重心	(50, 50)	0.05896	(50, 50)	0.05961	(50, 50)	0.06163	(50, 50)	0.06597	(50, 50)	0.07427



	座標	平均二乗誤差								
幾何中央点	(50, 50)	0.05671	(66, 50)	0.05448	(74, 50)	0.04888	(79, 50)	0.04126	(81, 50)	0.03259
分布重心	(50, 50)	0.05671	(54, 50)	0.05614	(58, 50)	0.05411	(62, 50)	0.05050	(68, 50)	0.04290
地区重心	(50, 50)	0.05671	(50, 50)	0.05727	(50, 50)	0.05901	(50, 50)	0.06272	(50, 50)	0.06981



	座標	平均二乗誤差								
幾何中央点	(50, 50)	0.05854	(53, 53)	0.05836	(58, 58)	0.05737	(66, 66)	0.05376	(74, 74)	0.04312
分布重心	(50, 50)	0.05854	(52, 52)	0.05844	(54, 54)	0.05799	(56, 56)	0.05712	(60, 60)	0.05373
地区重心	(50, 50)	0.05854	(50, 50)	0.05885	(50, 50)	0.05982	(50, 50)	0.06228	(50, 50)	0.06836



	座標	平均二乗誤差								
幾何中央点	(50, 50)	0.04540	(56, 50)	0.04502	(61, 50)	0.04397	(65, 50)	0.04242	(72, 50)	0.03878
分布重心	(50, 50)	0.04540	(54, 50)	0.04512	(58, 50)	0.04422	(62, 50)	0.04279	(70, 50)	0.03903
地区重心	(50, 50)	0.04540	(50, 50)	0.04609	(50, 50)	0.04784	(50, 50)	0.05063	(50, 50)	0.05937

図 4　様々な連続分布に関する最適な代表点（幾何中央点）と分布重心の位置，及び，平均二乗誤差の値．関数 $f(\mathbf{x})$の値は，1) 白いセル内では 0，2) 数字の付された灰色のセルでは数字の値，3) 数字の付されていない灰色のセルでは 1.0，である．また，白い四角は幾何中央点，×印は分布重心をそれぞれ表す．

結果は図 4 に示すとおりである．この図から分かるとおり，幾何中央点と分布重心とは必ずしも一

致しない．一般的に言って，幾何中央点の方が分布重心よりも関数形により強く影響を受け，$f(\mathbf{x})$の最大値を与える領域により近くなる傾向がある．

3 つの位置における誤差の程度は，場合によって異なる．全体的に見ると，図 4 右列の場合のように連続分布が偏っているほど，3 つの平均二乗誤差の違いは顕著であり，幾何中央点の最適性が確認できる．図 4 右上の場合が最も端的な例であり，幾何中央点と地区重心の平均二乗誤差は 3 倍以上異なる．しかしながら，分布重心と幾何中央点の位置がそれほどずれていない場合には，平均二乗誤差も大きく違うわけではない．例えば図 4 下列のように，連続分布の偏りが小さな場合には，いずれの位置も良い精度をもたらす．但し地区重心は，多くの場合，他の二つの位置よりもかなり大きな誤差を生む可能性がある．地区重心は，分布とは無関係に決定される位置であり，この結果は当然であろう．

以上の分析結果は，次のようにまとめることができる．代表点法の推定誤差という観点から見ると，理論上は幾何中央点が最適な代表点の位置である．特に，分布に偏りがある場合には，その最適性は顕著である．しかし，分布がそれほど偏っていない場合には，分布重心による推定も比較的良い結果をもたらす．幾何中央点の導出がそれほど簡単ではない一方，分布重心の導出は極めて容易であることを考えると，極端ではない分布については，分布重心も実際的な選択肢として有用であろう．

V おわりに

本論文では，代表点法による空間集計地区変換という観点から，最適な代表点の位置について分析した．解析的な分析の結果，元データが点分布，連続分布のいずれの場合にも，推定誤差を最小化する代表点の位置は分布の幾何中央点で与えられることが分かった．さらに実証分析からは，分布に偏りがある場合には明らかに幾何中央点が最適であり，分布が比較的均一な場合には，分布重心も比較的推定誤差を小さくする位置であることが分かった．従って，計算費用の問題を考え合わせると，実際上はこれら 2 つの位置を適宜使い分けることが望ましい．もちろん，代表点の位置は様々な観点から決定されるものであり，幾何中央点あるいは分布重心が常に最適であるというわけではない．しかしながら，空間集計地区変換は極めて基本的な空間操作であり，その推定誤差を小さくすることでその後の多くの分析結果の精度を高めることができる．従って，本論文において得られた結果は，代表点の位置を決定する際，十分考慮に値すると言えよう．

最後に，今後の研究課題について簡単に整理しておこう．まずはじめに，望ましい代表点の位置に関する多様な視点からの分析がある．本論文では推定誤差という観点からこの問題を取りあげたが，代表点の位置が影響を及ぼす場面は他にも数多い．例えば，空間分析において地区間の平均距離を代表点間距離で代用したり，影響圏をボロノイダイアグラムで近似するような場合には，代表点の取り方が結果に大きく影響する．また，ドットマップ作成は点分布を代表点で近似するという操作と同等であるが，そのときの代表点の位置は見やすい地図の作成という観点から決定される．このように，

代表点の位置については様々な観点から検討することが必要であり，今後の研究課題としても重要である．

第二に，集計先地区が凸ではない，或いは，あまり大きくない場合の分析がある．集計先地区が十分大きな凸多角形であるという仮定の適用範囲は非常に広いが，もちろん，この仮定が成り立たない場合も有り得る．そのような場合について，最適な代表点の位置がどのように与えられるのか，この点についてはさらに研究を進めていきたい．

文献

総務庁 1994. 『地域メッシュ統計の概要』総務庁統計局.

Aly, A. A. and Marucheck, A. S. 1982. Generalized Weber problem with rectangular regions. *Journal of the Operational Research Society* **33**: 983-989.

Brown, B. M. 1983. Statistical uses of the spatial median. *Journal of the Royal Statistical Society, Series B* **45**: 25-30.

Burrough, P. A. and McDonnell, R. A. 1998. *Principles of geographical information systems*. New York: Oxford University Press.

Drezner, Z. and Wesolowsky, G. O. 1980. Optimal location of a demand facility relative to area demand. *Naval Research Logistics Quarterly* **27**: 199-206.

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.

Love, R. F. 1972. A computational procedure for optimally locating a facility with respect to several rectangular regions. *Journal of Regional Science* **12**: 233-242.

Love, R. F., Morris, J. G. and Wesolowsky, G. O. 1988. *Facilities location: models & methods*. Amsterdam: North-Holland.

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.

Plastria, F. 1995. Continuous location problems. In *Facility location: a survey of applications and methods*. ed. Z. Drezner. 225-262. New York: Springer-Verlag.

Preparata, F. P. and Shamos, M. I. 1985. *Computational geometry - an introduction* -. New York: Springer-Verlag.

Ripley, B. D. 1981. *Spatial statistics*. New York: John Wiley.

Sadahiro, Y. 1998a. Accuracy of count data transferred through the areal weighting interpolation method. *Discussion Paper Series* **76E**, Department of Urban Engineering, University of Tokyo.

Sadahiro, Y. 1998b. Accuracy of count data estimated by the point-in-polygon method. *Discussion Paper Series* **77E**, Department of Urban Engineering, University of Tokyo.

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.

Small, C. G. 1990. A survey of multidimensional medians. *International Statistical Review* **58**: 263-277.

ABSTRACT

The optimal location of the representative point was discussed in relation to the estimation accuracy of the point-in-polygon interpolation method. The point-in-polygon interpolation is an areal interpolation method that is most frequently used in geography and GIS. This method implicitly assumes that all the points are located on the representative points, and thus it yields estimation error if the assumption does not hold. Accuracy of estimated data heavily depends on the location of the representative point. Hence this paper analyzed the relationship between estimation accuracy and the location of the representative point, and discussed the optimal location of the representative point that minimizes the estimation error of the point-in-polygon method. The point distribution and the surface were investigated and it was found in both cases that the geometrical (spatial) median of the distribution is the optimal location of the representative point. An empirical study was performed on the basis of the analytical study. The results showed that the estimation accuracy given by the geometrical median is remarkably higher than that given by the centroid when the surface distribution has a strongly concentrated form. On the other hand, these two locations yield similar accuracy of estimation if the surface distribution is rather smooth. These results and the fact that the geometrical median cannot be easily computed suggest that in practice the geometrical median and the centroid should be chosen according to the surface distribution.

Keywords: spatial aggregation, point-in-polygon method, estimation accuracy, geometrical median

英文摘要和訳

本論文では，代表点法の推定精度という観点から最適な代表点の位置を論じている．代表点法は，地理学や GIS において最もよく用いられる面補間の方法である．この方法は，全ての点が代表点の場所に集まっているという仮定を暗に置いており，その仮定が成り立たない場合には推定誤差を生ずる．推定誤差は，代表点の位置に大きく影響される．そこでここでは，推定精度と代表点の位置の関係を分析し，代表点法における推定誤差を最小化する代表点の位置について論じている．分布としては点分布と連続分布の 2 つを取りあげ，いずれの場合にも分布の幾何中央点が代表点の最適な位置であるという結論を得た．実証分析の結果からは，分布が偏っている場合には幾何中央点に基づいた推定精度が際立って高く，他方，分布が比較的平滑な場合には，幾何中央点と分布重心の与える推定精度はそれほど異なっていないことが明らかになった. この結果及び幾何中央点の導出が比較的困難であることを考え合わせると，実用上は幾何中央点と分布重心を分布形に応じて適宜使い分けることが望ましいと言えよう．

図表

図 1　代表点法による空間集計地区変換

Data transfer between zonal systems through the point-in-polygon method.

図 2　円に関する測度 $m(T; l)$

The measure $m(T; l)$ for the circle.

図 3　様々な図形に関する測度 $m(T; l)$

The measure $m(T; l)$ for a variety of shapes.

図 4　様々な連続分布に関する最適な代表点（幾何中央点）と分布重心の位置，及び，平均二乗誤差の値．関数 $f(\mathbf{x})$の値は，1) 白いセル内では 0，2) 数字の付された灰色のセルでは数字の値，3) 数字の付されていない灰色のセルでは 1.0，である．また，白い四角は幾何中央点，×印は分布重心をそれぞれ表す．

The optimal location of the representative point (geometrical median), the centroid of the distribution, and the centroid of the cell. The mean square error of estimated values is also shown. The value of the function $f(\mathbf{x})$ is 1) zero in white cells, 2) 1.0 in the lightest gray cells, 3) the value indicated in the figure. The white square and the cross represent the geometrical median and the centroids.