空間統計モデルに基づく面補間法の提案

村上大輔・堤盛人

An areal interpolation technique based on spatial statistical model

Daisuke MURAKAMI and Morito TSUTSUMI

Abstract. Differences in spatial units among spatial data often complicate analysis. Spatial unit convergence, called areal interpolation, is often applied to address this problem. Of the many areal interpolation methods that have been proposed, a few consider spatial dependence which is the general property of spatial data. Here, a new areal interpolation method that considers spatial dependence is presented based on a spatial statistical model called Kriging.

Keyword：面補間(Areal interpolation)、クリギング(Kriging)、空間従属性(Spatial dependence)、
体積保存則 (Pycnophylactic property)、Moore-Penrose の一般化逆行列(Moore-Penrose Inverse)

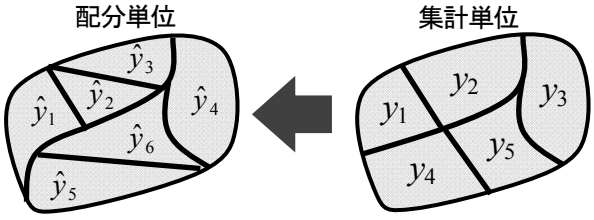
1. はじめに

社会経済データの多くは空間的な単位毎（以後、空間単位）に集計されて提供されているが、空間単位には行政界やメッシュなど様々なものが考えられるため、必要に応じ空間的単位の変換が必要となる。そのような変換は面補間（Areal interpolation）と呼ばれる。本稿では、面補間**前**の空間単位を『集計単位』、面補間**後**の集計単位を『配分単位』と呼び議論を進める。

面補間で考慮すべき性質として、まず、面補間前後で変数の総量が一定でなくてはならないという「体積保存則」が挙げられる。例えば、特定の集計単位内の予測人口を足しあわせた値は、同集計単位内の実際の人口と一致しなくてはならない。加えて、空間的に近接したデータ間に類似した傾向が見られるという空間従属性」が挙げられる。しかしながら、両性質を同時に考慮した面補間法は非常に限定的であり（例えば、Yoo and Kyriakidis (2006)、村上・堤(2009)）、これらとは異なる新たな手法の提案は、より高精度の面補間の実現とさらなる手法の開発可能性をもたらすという点で、非常に大きな意義を持つと考えられる。

そこで本研究では、空間統計学の空間内挿手法である Kriging（例えば Cressie (1993)）をベースとし、空間従属性と体積保存則を同時に考慮した新たな面補間法を提案する。

村上：〒305-8573 茨城県つくば市天王台 1-1-1
筑波大学大学院　システム情報工学研究科
E-mail: muraka51@sk.tsukuba.ac.jp



$\hat{y}_i$ ：データの予測値、　y_i ：データの実測値

図１　面補間の概要

2. 空間統計モデルの概要

空間統計学の代表的手法であるKrigingは、ユークリッド空間上の地点 $\mathbf{s}_i \in \boldsymbol{R}^2$ $(i = 1,\ldots, n)$ で観測されたデータから同空間上の任意地点 $\mathrm{s}_o \in \boldsymbol{R}^2$ のデータを予測することのできる手法であり、次式を基本式とする。

$$y_o = \mathbf{x}_o'\boldsymbol{\beta} + \varepsilon_o \qquad \varepsilon_o \sim N(0, \mathbf{C}(\mathbf{d}_{ij})) \qquad (1)$$

ここで、y_o、$\mathbf{x}_o'\boldsymbol{\beta}$及び$\varepsilon_o$は、予測地点のデータ、トレンド、及び、トレンドで説明されない局所的変動をそれぞれ表す。ただし、y_o及びε_oはスカラー、$\mathbf{x}_o$はp×1の説明変数ベクトル、$\boldsymbol{\beta}$はp×1のパラメータベクトルとする。$\mathbf{C}(\mathbf{d}_{ij})$は空間従属性を表現する共分散関数と呼ばれる距離の関数であり、次式で表される指数型共分散関数は代表的なものの一つである。

$$\mathbf{C}(d_{ij}) = \begin{cases} \sigma^2 \exp\left[-\left(\dfrac{d_{ij}}{w}\right)\right] & (s_i \neq s_j) \\ \tau^2 + \sigma^2 & ohterwise, \end{cases} \qquad (2)$$

ここで、σ^2 はpartial-sill、τ^2 はnugget、w はrange

と呼ばれるパラメータである。

同一空間上で観測された任意地点のデータを用いることで、(1)式の予測量は次式となる。

$$\hat{y}_o = \mathbf{x}'_o\hat{\boldsymbol{\beta}} + \mathbf{c}'_o\mathbf{C}^{-1}(\mathbf{y} - \mathbf{X}'\hat{\boldsymbol{\beta}}) \qquad (3)$$
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}$$

ただし、**y** 及び **X** は、観測地点についての被説明変数ベクトル及び説明変数行列を表し、それぞれ n×1 及び n×p のサイズを持つ。また、**C** は観測地点間の共分散行列、$\mathbf{c}_o$ は観測地点と予測地点との間の共分散ベクトルをそれぞれ表し、サイズはそれぞれ n×n 及び n×1 である。それらの各要素は共分散関数に基づいて与えられる。

Kriging のパラメータ推定手法として、Cressie (1985)により提案された、重み付き最小二乗法(WLS: Weighted Least Squared Method)及び一般化最小二乗法(GLS: Generalized Least Squared Method)を用いる手法（以下、WLS-GLS 法と略記）が広く用いられており、本研究でもこの手法を適用する。

3. Kriging に基づく新たな面補間法の提案

3.1.面補間法提案のための手法の前提

通常、面補間では計数データ（例：老年人口）の補間を前提とする。しかしながら、計数データ間相互の相関関係は、距離だけでなく、元となる個体の規模（計数データが老年人口の場合：総人口）にも強く依存するため、計数データのままでは空間従属性を捉えきれないことも考えられる。そこで本研究では、計数データを、その元となる個体の規模で割った変数、即ち、割合を表す変数（例：高齢化率）を対象とした空間単位変換のための手法を、広義の面補間法として提案する。従って、計数データへの適用に際しては、密度を表す変数に変換した後に適用することを前提とする。以降では、特に、人口あたりの個体数を表す変数を前提に議論を進める。

面補間法の構築に際し、線形回帰モデルに基づく面補間法を提案した Flowerdew and Green (1992)に倣い、集計単位及び配分単位の重ね合わせにより生成される空間単位を『細分単位』（図 2）と呼び定義する。以降では集計単位、配分単位及び細分単位の各領域を添え字 i, j, k によりそれぞれ表す。各空間単位の総数は、それぞれ I, J, K である。また、Flowerdew and Green (1992)では、図 2 の黒矢印で示すように集計単位のデータを細分単位に配分した後、同図の白矢印で示すようにそれを配分単位へと集計することで面補間を実行しており、本研究でも、その手順を前提とする。

集計単位と細分単位との間には(4)式、配分単位と細分単位との間には(5)式が、体積保存則を満足するための条件式としてそれぞれ成立しなければならない。

$$y_i = \sum_k \frac{n_{ik}}{n_i} y_k \quad (4), \qquad y_j = \sum_k \frac{n_{jk}}{n_j} y_k \quad (5)$$

ここで、n_i は集計単位 i の人口、n_j は配分単位 j の人口を表す。また、n_{ik}、n_{jk} は、それぞれ、n_i のうち細分単位 k で観測された人口を表す。

細分単位から配分単位への集計は、(8)式より予測誤差を伴わず行うことができるため、以降では、集計単位から細分単位へのデータの配分（面補間）のための手法に議論を絞る。

3.2.Kriging に基づく面補間のためのモデル構築

本研究では、Flowerdew and Green (1992)の手法を Kriging の基本モデルに援用することで面補間のための新たなモデルを構築する。なお、彼らの手法は、繰り返し計算による細分単位の予測値算出を前提としており、本研究で構築する手法も、同様の計算手順を踏むものとする。

Flowerdew and Green (1992)によるモデル構築の考え方を参考に、体積保存則満足のための制約式(4)を Kriging の基本式(1)式に考慮すると、面補間の基本式として次式が導かれる。

$$\mathbf{y}_k - \mathbf{N}\mathbf{y}_k + \bar{\mathbf{z}}_k = \mathbf{X}_k\boldsymbol{\beta} + \boldsymbol{\varepsilon}_k \quad \boldsymbol{\varepsilon}_k \sim N(0, \mathbf{C}(d_{ij})) \quad (6)$$

ここで **N** は n_{ik}/n_{iik} を要素とする K×K の行列であり、その k 行 k' 列の要素は、細分単位 k の人口のうち、細分単位 k' を含む集計単位 i' 内で観測された人口の割合 $n_{i'k}$ である。また、$\bar{\mathbf{z}}_k$ は各細分単位が含まれる集計単位の被説明変数を要素とする K×1 のベクトルである。なお、Flowerdew and Green (1992)のモデルは、(6)式の共分散が全て 0 に置き換えられた特殊形とみなすことができる。

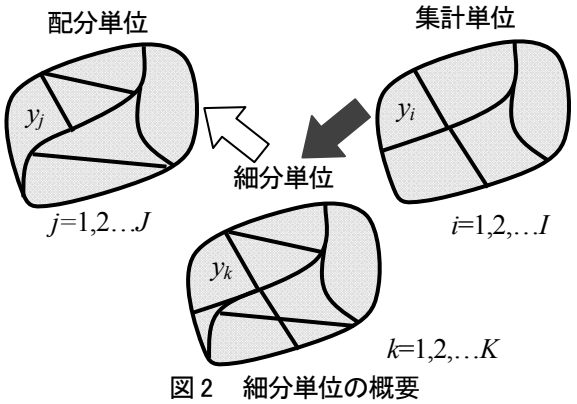


図 2　細分単位の概要

通常の Kriging の予測式(3)を援用することで、(6)式の予測式は(7)式となる。ここで、行列(**I**– **N**)は、その対角線上の要素が同一集計単位に含まれる細分単位毎の正方行列で与えられたブロック対角行列となり、体積保存則を考慮した結果として各正方行列式が 0 となるため、行列(**I**– **N**) の行列式も 0 となって通常の逆行列が存在しない。その結果、未知変数である y_k が一意に求まらないこととなる。

$$(\mathbf{I} - \mathbf{N})\mathbf{y}_k = \mathbf{X}_k\boldsymbol{\beta} + \mathbf{C}'_k\mathbf{C}_k^{-1}\{\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}\} - \bar{\mathbf{z}}_k \quad (7)$$

そこで、行列(**I**– **N**) の Moore-Penrose の一般化逆行列 (**I**– **N**) $^+$ を用いて次の方程式を導き、これを解いて y_k の解を得ることとする。

$$\mathbf{y}_k = (\mathbf{I} - \mathbf{N})^+[\mathbf{X}_k\boldsymbol{\beta} + \mathbf{C}'_k\mathbf{C}_k^{-1}\{\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}\} - \bar{\mathbf{z}}_k] \quad (8)$$

実際の適用に当たっては、$\boldsymbol{\beta}$ も未知であり、未知パラメータ$\boldsymbol{\beta}$ と未知変数である $\mathbf{y}_k$ を同時に求める必要がある。そこで、ここでは、(8)式による左辺の予測値を右辺の初期値として代入することで、繰り返し計算を行うこととする。しかしながら、(8)式は「データが既知である場合、それは、誤差を持たない唯一の実現値である」という空間統計学の仮定に基づいて構築されたため、(8)式により更新される $\mathbf{y}_k$ の値は、更新前の値と必ず一致する。しかしながら、$\mathbf{y}_k$ は実際には誤差を持つため、それを明示的に考慮した上での(8)式の修正が必要である。

空間統計学では、実現値を持つ地点の値が唯一であることを Nugget 効果と呼ばれる効果により表現する。この効果は、トレンド成分及び局所的変動成分により説明されないデータの分散の大きさを表す。この効果を置くことで、例え、トレンド及び局所的変動で説明されない成分が実測値に含まれていたとしても、それは、Nugget 効果により説明される成分とされる。従って、実測値は Nugget 効果まで考慮した尤もらしい値とみなされ、予測値は実測値から変化しない。一方で、実測値が未知である地点には、Nugget 効果は仮定されず、予測値はトレンド及び局所的変動のみにより説明される。

そこで、提案手法でも、Nugget 効果を表す共分散関数(2)式のパラメータ τ^2 を、実現値が未知である細分単位について排除することとする。それにより、(8)式は(9)式に修正される。

$$\mathbf{y}_k = (\mathbf{I} - \mathbf{N})^+[\mathbf{X}_k\boldsymbol{\beta} + \mathbf{C}'_{ok}\mathbf{C}_k^{-1}\{\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}\} - \bar{\mathbf{z}}_k] \quad (9)$$

$\mathbf{C}_{ok}$ は $\mathbf{C}_k$ と同様に共分散行列であるが、実測値が未知である細分単位の共分散は、Nugget 効果の大きさを表すパラメータ τ^2 を 0 とした共分散関数に基づき算出される点で $\mathbf{C}_k$ と異なる。

(9)式では、期待二乗誤差の最小化という基準の下、予測値を算出するが（後述）、同式において、実測値が未知である細分単位にのみ仮定された τ^2 が 0 となることは、その予測値に誤差が存在しないことを意味するため、τ^2 が 0 となった時点で期待二乗誤差は 0 とみなされ、計算は収束する。しかしながら、不完全な情報に基づく完全な予測は一般に不可能であり、一定の予測誤差が生じると考えるのが自然であろう。そこで、τ^2 は正という制約を与えることとする。これにより、予測データに対する予測誤差の存在が前提となる。また、τ^2 が 0 となることに起因した計算の収束は起こらなくなる。

3.3. 提案手法による面補間の手順

予測値の算出は以下の手順に基づき行う。

[1] 細分単位のデータ $\mathbf{y}_k$ の初期値を $\bar{\mathbf{z}}_k$ で与える。
[2] WLS-GLS 法等よりパラメータを推定する。
[3] 得られたパラメータを(9)式に代入することで $\mathbf{y}_k$ を更新する。
[4] [2]及び[3]を収束するまで繰り返す。

ここで、パラメータ推定（[2]）では、更新されたデータを完全データと見なし、予測地点に対しても Nugget 効果を仮定するのに対し、予測（[3]）では、通常の Kriging と同様に予測地点のみに対する Nugget 効果を排除して予測を行う。そのような方法により、パラメータ推定（[2]）では、期待二乗誤差を最小化するパラメータが推定される。一方で、予測（[3]）では、（[2]）で得られたパラメータから算出される予測値 $\hat{\mathbf{y}}_k$ を、実測値が未知である地点についてのみその誤差が 0 となるように動かすため、期待二乗誤差はパラメータ推定時よりも小さくなる。

従って、上記の繰り返し計算では予測値算出及びパラメータ推定のそれぞれで期待二乗誤差が小さくなるため、計算は必ず収束する。また、通常の Kriging では期待二乗誤差の最小化により予測値が算出されるが、提案手法もまた、同様の基準に基づいて予測値を算出する。

4. 実証による提案手法の有用性の考察

4.1.実証の概要

北関東 3 県（茨城県，栃木県，群馬県）のデータを用いた実証により、提案手法の有効性を検証する。本

実証では、2007年度の116市町村区分を集計単位、1995年度の203市町村区分を配分単位とした面補間を、①通常の線形回帰モデル[LM１]、②体積保存則を満足する線形回帰モデル、即ち(6)式の $\boldsymbol{\varepsilon}_k$ に $N(0,\sigma^2\mathbf{I})$ を仮定したモデル[LM２]、及び③提案手法[提案]によりそれぞれ行い比較する。

面補間の対象は国勢調査において得られた2007年度における大学・大学院卒業者比率［＝（大学又は大学院卒業を最終学歴とする人の数）／（25歳以上人口）］とし、対数変換をしたものを被説明変数として用いる。説明変数は東京駅までの距離（東京距離）[km]，最寄りの県庁所在地の主要駅までの距離（県庁距離）[km]、平均年齢[歳]、第三次産業就業者比率（三次比率）とする。

面補間の予測精度の検証には、1995年度の市町村区分毎に集計された2007年度の大学・大学院卒業者比率の実測値が（予測変数の真値として）必要である。そこで、2007年度における町丁目別大学・大学院卒業者比率データを、(5)式を用いて集計することで、それを作成した。

4.2.分析結果

表１はRMSE（*Root Mean Square Error*：平均二乗平方根誤差）の算出結果である。[LM２]と[LM１]との差は体積保存則の考慮による効果、[提案]と[LM２]との差は空間従属性の考慮による効果の大きさをそれぞれ表し、前者だけでなく後者についても一定の差が存在することから、空間従属性の考慮が面補間においても重要であることが示唆される。また、予測を行った全139市町村のうちの76%にあたる105市町村では、[提案]による予測精度が線形制約モデルによるそれを上回っていることからもそのことが確認できる。

最後に、(10)式で定義される誤差率を、予測した空間単位における[提案]及び[LM２]の両予測結果について計算し、その差をプロットした（図4）。

$$(\text{誤差率}) = 100 \times \sqrt{\{(\hat{y}_k - y_k)/y_k\}^2} \qquad (10)$$

図5から、[提案]による予測精度が向上した地点が対象地域全域で見られ、向上の程度は茨城県南部において特に著しいことがわかる。

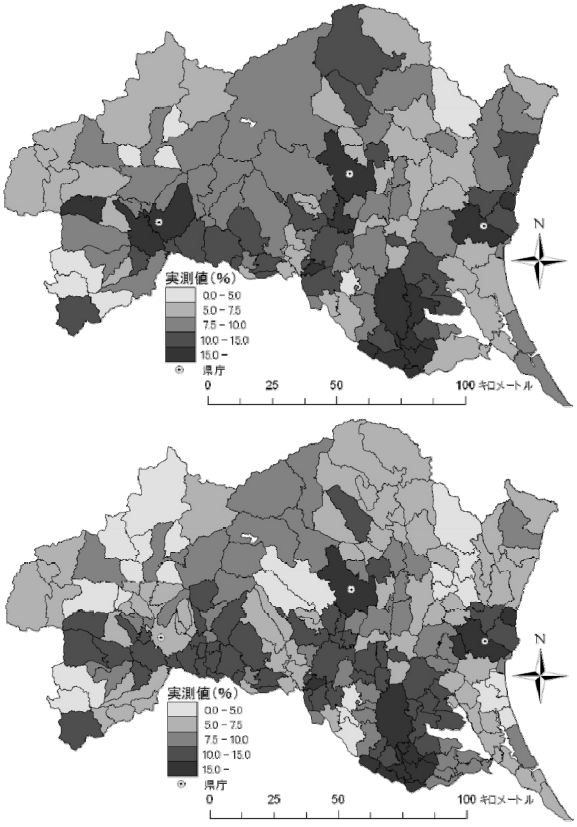


図3　2007年度区分（上）及び1995年度区分(下)での2007年度大学･大学院卒業者比率実測値

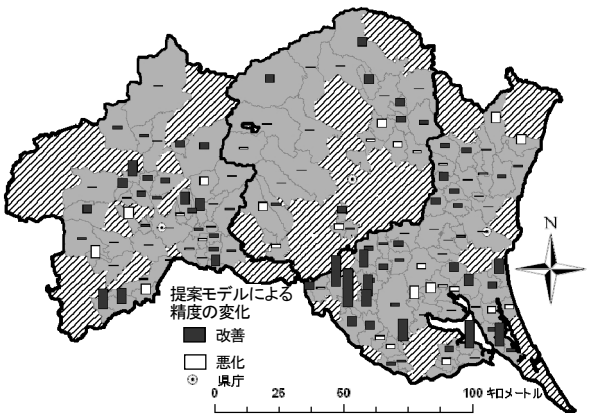


図4　[LM2]の誤差率 － [提案]の誤差率

表１　RMSEの算出結果

	LM１	LM２	提案
RMSE	2.87	1.98	1.82

参考文献

Cressie, N. (1985), Fitting variogram models by weighted least squares, *Mathematical Geology*, pp. 563-586.

Cressie, N. (1993), *Statistics for Spatial Data*, Revised Edition,John Wiley & Sons.

Flowerdew, R. and Green, M. (1992), Developments in areal interpolation methods and GIS,*Annals of Regional Science*, 26, pp.67–78.

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

Yoo, E. H, Kyriakidis, P. C. (2006), Area-to-point Kriging with inequality-type data. Journal of Geographical systems, 8(4), pp. 357-390.

貞広幸雄　(2000)：「空間集計データにおける面補間法の推定精度評価」，『都市計画』,225, pp.75–81.

村上大輔　・堤盛人(2009), 市町村合併による統計データの集計単位変更に対する方策の提案-空間計量経済モデルを用いた分析への対処法, 第 23 回応用地域学会発表論文 (Available: http://www.shiratori.riec.tohoku.ac.jp/~takita/ARSC2009/Paper/ARSC2009_03.pdf, accessed on 9 April 2010)