統計数理(2025)
第 73 巻 第 1 号 53–63

特集「空間統計モデリング：理論と応用」
[研究詳解]

ベイズ的モデル統合による空間予測

菅澤 翔之助†

(受付 2024 年 6 月 30 日；改訂 9 月 8 日；採択 9 月 9 日)

要　　旨

有限地点で観測された空間データからモデルを推定し，未観測地点を予測することは空間データ分析において重要なタスクの 1 つである．近年では，古典的な地球統計学や空間計量経済学的モデルから機械学習的アプローチまで様々なモデルが利用可能である．そのため，解析対象のデータに対して適切に分析手法を取捨選択することは分析上の重要な論点である．本稿では，予測モデルが複数個得られる状況においてベイズ的にモデルを統合した空間予測の方法論についてのレビューを行う．特に，近年提案されたベイズ的空間予測統合に関して，古典的な統合方法との違いや具体的な推定アルゴリズムに関する解説を行う．

キーワード：ベイズモデル平均，スタッキング，ガウス過程，ベイズ的予測統合．

1. はじめに

近年ではセンシング技術やデータ収集技術の発展に伴い，位置情報が付随した多様な空間データを用いた分析が可能になっている．空間データ分析の主要な目的の一つとして，未観測地点(観測データが得られていない地点)における値の予測が挙げられる．そのための方法として，クリギングと呼ばれる地球統計学的な手法 (e.g. Oliver and Webster, 1990)や空間自己回帰モデルなどの空間計量経済学的方法 (e.g. Anselin, 2022)，さらには地理的加重回帰 (e.g. Fotheringham et al., 2002)や階層ベイズ的アプローチ (e.g. Banerjee et al., 2014)などが知られている．近年では，勾配ブースティング木やランダムフォレスト，深層学習などの機械学習的アプローチも採用されるようになってきた (e.g. Du et al., 2020)．

候補として複数の予測モデルが考えられる場合，大きく 2 つの方向性が考えられる．1 つは何らかの規準に基づいて 1 つの最適なモデルを選択する方法である．代表的なアプローチとして交差検証法や情報量規準があるが，このような方法の弱点として，モデルを選択するステップにおける誤差や不確実性を評価することが難しい点が挙げられる．もう 1 つの方向性として，候補モデルの中から 1 つのモデルのみを選ぶ代わりに，各モデルごとに何らかのウエイトを与えて合成するアプローチがある．これは一般的にモデル平均と呼ばれており，モデル選択に基づく方法よりも予測精度が高くなる傾向があることが知られている．本稿では，空間予測において複数モデルが得られている状況でそれらを合成するアプローチに焦点を当て，特にベイズ的な枠組みで予測モデルを統合する方法論を解説する．標準的なアプローチとして，ベイズモデル平均やスタッキングを紹介し，さらにこれらの古典的なモデル統合手法を含む一般的な枠組みとしてベイズ的予測統合(Bayesian predictive synthesis)を紹介し，空間的なモデル重要度の異質性を考慮した新しい統合手法 (Cabel et al., 2022)を解説する．本稿は，従来のモデ

† 慶應義塾大学 経済学部：〒 108–8345 東京都港区三田 2–15–45

ル統合の手法から最近の統合手法についてある程度の詳細まで含めて解説することに重きを置いており，シミュレーションや実データにおける数値的な詳細に関しては Cabel et al.（2022）などの数値例を参照されたい．

2. ベイズ的モデル統合の古典的なアプローチ

2.1 ベイズモデル平均（Bayesian model averaging）

モデル統合の古典的かつ代表的な方法としてベイズモデル平均化（Bayesian model averaging, BMA）が知られている（cf. Hoeting et al., 1999）．これはモデルに対する事後確率による重み付き平均を考える方法で，空間データに限らず一般のモデルに対して使える方法である．$M_1, \ldots, M_K$ を候補モデルとし，$y = (y_1, \ldots, y_n)$ を観測データとする．さらに，モデル M_k のパラメータを θ_k（k ごとに次元が異なる可能性がある）とし，その事前分布を $p(\theta_k|M_k)$ とする．BMA は未観測データ $\tilde{y}$ に対する予測分布を

$$p(\tilde{y}|y) = \sum_{k=1}^{K} p(\tilde{y}|y, M_k)p(M_k|y)$$

で与える．ここで，$p(M_k|y)$ はモデル M_k の事後確率

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{k'=1}^{K} p(y|M_{k'})p(M_{k'})}$$

である．また，$p(M_k)$ はモデルの事前確率で，$p(y|M_k)$ はモデル M_k の周辺尤度（marginal likelihood）

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k$$

である．モデルの事前確率として一様事前分布 $p(M_k) = 1/K$ を考えた場合，モデル M_k の事後確率は周辺尤度に比例して決まることになる．すなわち，BMA は周辺尤度の大きさ（モデルのエビデンスの強さ）に応じて決まるウエイトによって予測分布を重み付き平均している方法であると解釈することができる．BMA は候補モデル $M_1, \ldots, M_K$ のどれかが真のデータ生成過程に一致している（どれかはわからない）状況では妥当な方法として知られているが，そうでない状況では漸近的に Kullback-Leibler（KL）ダイバージェンスが最も小さくなるモデルに対するウエイトが 1 になる（それ以外のモデルは 0 になってしまう）問題点が知られている．

空間データ分析の文脈では，複数の空間自己回帰モデルを BMA によって統合するアプローチが Debarsy and LeSage（2022）や LeSage and Parent（2007）で検討されている．具体例として，空間誤差モデルにおいて異なる説明変数行列を用いた複数のモデルを BMA で統合するケースを考えてみる．y を n 次元の被説明変数ベクトル，X を $n \times p$ の説明変数行列，W を $n \times n$ の空間重み行列として，空間誤差モデルは

$$y = X\beta + e, \quad e = \rho We + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

と表せる．このモデルは行列 $Q(\rho) = (I_n - \rho W)^\top (I_n - \rho W)$ を用いて $y \sim N(X\beta, \sigma^2 Q(\rho)^{-1})$ と表すことができる．BMA を求めるためには周辺尤度 $p(y)$ を求める必要があるが，それは

$$p(y) = \int \phi_n(y; X\beta, \sigma^2 Q(\rho))\pi(\beta|\sigma^2)\pi(\sigma^2)\pi(\rho)d\rho$$

で与えられる．ただし，$\pi(\beta|\sigma^2), \pi(\sigma^2), \pi(\rho)$ は β, σ^2, ρ の（条件付き）事前分布である．具体的には，β に対して g 事前分布 $\pi(\beta|\sigma^2) \sim N(0, \sigma^2\{gX^\top Q(\rho)X\}^{-1})$ および σ^2 に対して逆ガンマ事前分布 $\pi(\sigma^2) \sim \mathrm{IG}(\nu/2, \nu s^2/2)$ を用いる．すると周辺尤度は

$$p(y) = K\left(\frac{g}{1+g}\right)^{p/2}\int |Q(\rho)|^{1/2}\left[\nu s^2 + \frac{g}{1+g}\{y - X\hat{\beta}(\rho)\}Q(\rho)\{y - X\hat{\beta}(\rho)\}\right]^{-\frac{n+\nu-1}{2}}\pi(\rho)d\rho$$

となる．ただし，

$$K = \frac{\Gamma\left(\frac{n+\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}(\nu s^2)^{\nu/2}\pi^{-\frac{n-1}{2}}, \qquad \hat{\beta}(\rho) = \{X^\top Q(\rho)X\}^{-1}X^\top Q(\rho)y$$

である．$p(y)$ を求めるには $\pi(\rho)$ について積分をする必要があるが，一次元の有界区間 $(-1, 1)$ 上の積分のため比較的容易に数値積分を行うことができる．

2.2 スタッキング（Stacking）

スタッキング(Stacking)は複数モデルからの推定値を統合するアルゴリズムである（cf. Breiman, 1996）．$y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ をデータ y から y_i のみを除外したデータセットとし，$\hat{\mu}_k(y_{-i})$ をモデル M_k と y_{-i} に基づく未観測データ $\tilde{y}$ の leave-one-out（LOO）予測値とする．このとき，$\tilde{y}$ を $\hat{\mu}_k(y_{-i})$ の凸結合 $\sum_{k=1}^K w_k\hat{\mu}_k(y_{-i})$（ただし，$w_k$ は $\sum_{k=1}^K w_k = 1$ を満たす)によって求めることを考え，最適なウエイト $w = (w_1, \ldots, w_K)$ を

$$\sum_{i=1}^n\left\{y_i - \sum_{k=1}^K w_k\hat{\mu}_k(y_{-i})\right\}^2$$

を最小化するように決める．一般的に，二乗損失によって予測精度を比較した場合，BMA よりもスタッキングの方が予測精度が高くなることが知られている（Clarke, 2003）．

Clydec and Iversen（2013)は，スタッキングのアルゴリズムに対してベイズ的な解釈を与えている．$p_*(\tilde{y}|y)$ を真のデータ生成過程によって構成される予測分布(真の予測分布と呼ぶ)として，未観測データ $\tilde{y}$ を $\sum_{k=1}^K w_k\hat{\mu}_k(y)$ で予測することを考える．ここで，$\hat{\mu}_k(y)$ は予測分布の期待値

$$\int \tilde{y}p(\tilde{y}|\theta_k, M_k)p(\theta_k|y, M_k)d\theta_k d\tilde{y}$$

やパラメータの事後平均をプラグインした期待値

$$\int \tilde{y}p(\tilde{y}|\hat{\theta}_k, M_k)d\tilde{y}, \qquad \hat{\theta}_k = E[\theta_k|y, M_k]$$

などが考えられる．このとき，合成した推定量の期待二乗損失は

$$\int\left\{\tilde{y} - \sum_{k=1}^K w_k\hat{\mu}_k(y)\right\}^2 p_*(\tilde{y}|y)d\tilde{y} \tag{2.1}$$

と表せる．この量を最小にする w が最適な重みになるが，真の予測分布 $p_*(\tilde{y}|y)$ が未知のため上記の積分を評価することができない．その代わりに，スタッキングの考え方を使って積分の近似を構成することができる．最適なウエイトを LOO 二乗損失

$$\frac{1}{n}\sum_{i=1}^n\left\{y_i - \sum_{k=1}^K w_k\hat{\mu}_k(y_{-i})\right\}^2 \tag{2.2}$$

を最小化するように決める．このとき，LOO 二乗損失(2.2)が漸近的に期待二乗損失(2.1)に一致することが Le and Clarke (2017)によって示されている．

従来のスタッキングは点予測値を合成する手法であるが，Yao et al. (2018)では，スタッキングを用いた予測分布の統合手法が提案されている．

$$p(\tilde{y}|y, M_k) = \int p(\tilde{y}|\theta_k, M_k)p(\theta_k|y, M_k)d\theta_k$$

をモデル M_k に基づく予測分布として，凸結合した予測分布 $\sum_{k=1}^{K} w_k p(\tilde{y}|y, M_k)$ を考える．また，LOO 予測分布を

$$p(y_i|y_{-i}, M_k) = \int p(y_i|\theta_k, M_k)p(\theta_k|y_{-i}, M_k)d\theta_k$$

と定める．このとき，最適なウエイト $w_1, \ldots, w_K$ は以下の最適化問題によって求めることができる．

$$\max_{w} \frac{1}{n}\sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} w_k p(y_i|y_{-i}, M_k)\right\}, \quad \text{s.t.} \quad w_k \geq 0, \quad \sum_{k=1}^{K} w_k = 1.$$

この方法を実行する場合の計算上の問題点として，LOO 予測分布 $p(y_i|y_{-i}, M_k)$ を n 回計算する必要がある点が挙げられる．最も単純に実行する場合は，各 y_{-i} ごとに θ_k の事後分布を(マルコフ連鎖モンテカルロ法などで)計算する必要がある．Yao et al. (2018)では，この問題に対して重点サンプリングによる効率的な近似計算手法を与えている．全データによる事後分布 $p(\theta_k|y, M_k)$ を用いて，LOO 予測分布は

$$p(y_i|y_{-i}, M_k) = \int p(y_i|\theta_k, M_k)\frac{p(\theta_k|y_{-i}, M_k)}{p(\theta_k|y, M_k)}p(\theta_k|y, M_k)d\theta_k$$

と表すことができるため

$$\frac{p(\theta_k|y_{-i}, M_k)}{p(\theta_k|y, M_k)} \propto \frac{1}{p(y_i|\theta_k, M_k)} \equiv w(\theta_k)$$

を重点ウエイトとした重点サンプリングによって計算することができる．$w(\theta_k)$ は密度の逆数で与えられるため，数値計算が不安定になってしまう問題があるが，Yao et al. (2018)ではパレート平滑化重点サンプリング (e.g. Vehtari et al., 2024)を用いることで安定的な計算方法を提案している．

近年では，スタッキングの考え方を用いて階層空間回帰モデルの推定自体に利用する方法も開発されている．X を $n \times p$ の計画行列として，以下のような階層空間回帰モデルを考える．

$$y|z \sim N(X\beta + u, \delta^2\tau^2 I_n), \quad u \sim N(0, \tau^2 C(\phi)).$$

ここで，u が潜在的な空間効果を表す n 次元ベクトルであり，$C(\phi)$ はパラメータ ϕ によって定まる相関行列である．Zhang et al. (2023)は，パラメータ (δ^2, ϕ) が所与のもとで事後分布が解析的に計算できる点に注目し，スタッキングを用いた高速な事後分布の近似方法を与えている．具体的には，(δ^2, ϕ) に対する候補値の集合を L セット用意しておき，それぞれを 1 つのモデル M_l $(l = 1, \ldots, L)$ とみなすことで，スタッキングのアルゴリズムを適用して最適なウエイト $\hat{w}_1, \ldots, \hat{w}_L$ を計算している．その結果，例えば β に対する合成された事後分布(stacked posterior)は $\sum_{l=1}^{L} \hat{w}_l p(\beta|y, M_g)$ の形($p(\beta|y, M_g)$ はモデル M_g のもとでの β の周辺事後分布)で与えられる．

2.3 古典的なアプローチの限界

これまでに紹介した手法は，点予測や予測分布を定数のウエイトによる凸結合の形で統合する方針を採用している．しかし，空間データを用いた予測タスクにおいてはモデルの重要度が地点によって異なることが想定され，既存のモデル統合のアプローチでは精度の高い空間予測を実現することは難しいと考えられる．このようなモデルウエイトの空間異質性を考慮したモ

デル統合を実現するため，Cabel et al. (2022)では，ベイズ的予測統合と呼ばれる枠組みを用いた空間モデルの統合手法を導入している．

3. ベイズ的空間予測統合(Bayesian spatial predictive synthesis)

3.1 ベイズ的予測統合

s を緯度・経度などの位置情報，$y(s)$ を地点 s における確率変数とする．n 個の地点 $s_1,\dots,s_n$ におけるデータ $y(s_1),\dots,y(s_n)$ を用いて，任意の地点 s における未観測な確率変数 $y(s)$ を予測する問題を考える．このデータに対して J 個の統計モデルをそれぞれ並列に適用することで，J 個の予測分布 $h_j(\cdot)$ $(j=1,\dots,J)$ および観測地点において予測分布に従う確率変数 $f_j(s_i)$ を得ることができる．$h_j(\cdot)$ は様々な分布があり得るが，典型的な例としては予測平均と分散を持つ正規分布である．J 個の予測分布が与える情報集合を $\mathcal{H}(s)=\{h_1(f_1(s)),\dots,h_J(f_J(s))\}$ と定義する．ここで, $f_j(s)$ は地点 s における確率変数である．このとき，情報集合 $\mathcal{H}(s)$ からどのように $y(s)$ の予測分布を構成するかが問題となる．

この問いに対する理論的な答えとして，ベイズ的予測統合(Bayesian predictive synthesis, BPS)の枠組み（e.g. Genest and Schervish, 1985; West and Crosse, 1992; West, 1992)において一般的かつコヒーレントな分布形が与えられている．BPS によって与えられる事後分布は以下のような形になる．

$$\Pi_{\mathrm{BPS}}\big(y(s)|\Psi(s),\mathcal{H}(s)\big)=\int \alpha\big(y(s)|f(s),\Psi(s)\big)\prod_{j=1}^{J}h_j\big(f_j(s)\big)df_j(s). \tag{3.1}$$

ここで，$\alpha\big(y(s)|f(s),\Psi(s)\big)$ は統合関数，$\Psi(s)$ は地点ごとのパラメータ，$f(s)=(f_1(s),\dots,f_J(s))$ は潜在変数のベクトルである．統合関数は複数のモデルをどのように統合するかを決定する関数であるが，一般的な理論では具体的な形までは指定されておらず，問題に応じて適切な統合関数を与える必要がある．例えば，統合関数として

$$\alpha\big(y(s)|f(s),\Psi(s)\big)=\sum_{j=1}^{J}w_j\delta_{f_j(s)}(y(s))$$

を考える．ただし，$\delta_a(y)$ は $y=a$ 上の Dirac 測度である．このとき，$\Pi_{\mathrm{BPS}}\big(y(s)|\Psi(s),\mathcal{H}(s)\big)=\sum_{j=1}^{J}w_jh_j(y(s))$ となり，2.2 節で扱ったような J 個の密度関数の重み付き平均の形になる．したがって，BPS による統合分布の形(3.1)は，既存のモデル平均化を含んだ一般的な枠組みになっていることがわかる．空間予測の問題で BPS を使う場合は，どのような形の統合関数を与えるかが肝となるが，Cabel et al. (2022)では

$$\alpha(y(s)|f(s),\Psi(s))=\phi\left(y(s);\beta_0(s)+\sum_{j=1}^{J}\beta_j(s)f_j(s),\sigma^2\right) \tag{3.2}$$

という形の統合関数を提案している．ここで，$\Psi(s)=\{\beta_0(s),\beta_1(s),\dots,\beta_J(s),\sigma^2\}$ である．この形の統合関数の妥当性として，Cabel et al. (2022)では，確率変数 $f(s)_1,\dots,f_J(s)$ が与えられたもとでの $y(s)$ の予測問題を考えたときに，ある条件のもとで最良近似を与える形であることが示されている．統合関数(3.2)を利用した予測統合手法はベイズ的空間予測統合(Bayesian spatial predictive synthesis, BSPS)と呼ぶ．

統合関数(3.2)の重要な性質として，$f_j(s)$ に対する重み $\beta_j(s)$ が地点に依存している点が挙げられる．これは，場所によって各モデルの重要度が異なる可能性があること(モデル統合における空間異質性)を考慮した形になっている．また，J 個のモデルが統合されているのに加

えて地点に依存する切片項 $\beta_0(s)$ が付随している．これは，もし統合された項 $\sum_{j=1}^{J}\beta_j(s)f_j(s)$ が $y(s)$ の変動を捉えるのに十分なほど柔軟ではなかった場合に，空間トレンド $\beta_0(s)$ として柔軟さを補う役割を果たしている．さらに，ウエイトに対して $\sum_{j=1}^{J}\beta_j(s)=1$ などの制約は与えておらず，地点によっては負の値になることも許した定式化になっている．このような性質は，2 節で紹介した古典的な方法論とは大きく異なる特徴である．

3.2 BSPS の推定アルゴリズム

BSPS を実行するためには，統合関数のパラメータ $\Psi(s)$ を推定する必要がある．$\beta_j(s)$ は位置情報 s の関数であるが，ガウス過程事前分布を用いることで推定が可能となる．$i=1,\ldots,n$ に対して，$y_i=y(s_i)$, $\beta_{ji}=\beta_j(s_i)$, $f_{ji}=f_j(s_i)$ とする．以下では簡単のため，各モデルの予測分布として予測値と予測分散をそれぞれ平均・分散に持つ正規分布を想定する．すなわち，$f_{ji}\sim N(a_{ji},b_{ji})$ を仮定する．このとき，BSPS によって統合事後分布を求めることは，以下の潜在変数モデルを推定することと同値である．

$$y_i=\beta_{0i}+\sum_{j=1}^{J}\beta_{ji}f_{ji}+\epsilon_i,\quad (\beta_{j1},\ldots,\beta_{jn})\sim N(m_j1_n,\tau_j^2C(\phi_j)),\quad \epsilon_i\sim N(0,\sigma^2). \tag{3.3}$$

ここで，$(\beta_{j1},\ldots,\beta_{jn})$ に対する同時分布はガウス過程に由来するものであり，m_j はその平均である．全てのモデルの単純平均を事前平均として採用する方法（$m_0=0$ かつ $m_j=1/J$ と設定するアプローチ）が 1 つの自然な方法であるが，m_j 自体もデータから推定することも可能である．$C(\phi_j)$ はカーネル関数によって決まる分散共分散行列であり，例えば指数型のカーネルを使ったモデル化（$C(\phi)$ の (i,j) 成分が $\exp(-\|s_i-s_j\|/\phi)$ によって与えられるモデル）が考えられる．

未知パラメータに対して事前分布を与えることで，潜在変数 β_{ji} および f_{ji} の事後推論が可能となる．その際には，マルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo, MCMC）と呼ばれる計算アルゴリズムによって事後分布から乱数を生成する形で事後推測を行うことができる．今回のモデル (3.3) は潜在変数を用いた変化係数モデルと捉えることができるため，ギブスサンプラーによる MCMC によって乱数生成を行うことができる．ギブスサンプラーとは，パラメータ σ^2,τ_j^2,ϕ_j および潜在変数 β_{ji},f_{ji} それぞれの完全条件付き事後分布から繰り返し乱数を生成するアルゴリズムである．モデル (3.3) のもと，$\beta_j=(\beta_{j1},\ldots,\beta_{jn})$（地点 s_i におけるモデル j のウエイト）の完全条件付き分布は $N(A_j^\beta B_j^\beta,A_j^\beta)$

$$A_j^\beta=\left\{\frac{\Omega_j}{\sigma^2}+\frac{C(\phi_j)^{-1}}{\tau_j^2}\right\}^{-1},\quad B_j^\beta=\frac{1}{\sigma^2}f_j\circ\left(y-\beta_0-\sum_{k=1,k\neq j}^{J}f_k\circ\beta_k\right)+\frac{m_j}{\tau_j^2}C(\phi_j)^{-1}1_n$$

で与えられる．ここで，$\Omega_j=\mathrm{diag}(f_{j1}^2,\ldots,f_{jn}^2)$, $f_j=(f_{j1},\ldots,f_{jn})$, $\beta_j=(\beta_{j1},\ldots,\beta_{jn})$ であり，$\circ$ はアダマール積を表す．同様に，f_{ji} の完全条件付き分布は $N(A_{ji}^{(f)}B_{ji}^{(f)},A_{ji}^{(f)})$

$$A_{ji}^{(f)}=\left(\frac{\beta_{ji}^2}{\sigma^2}+\frac{1}{b_{ji}}\right)^{-1},\quad B_{ji}^{(f)}=\frac{\beta_{ji}}{\sigma^2}\left(y_i-\beta_{0i}-\sum_{k=1,k\neq j}^{J}\beta_{ki}f_{ki}\right)+\frac{a_{ji}}{b_{ji}}$$

で与えられる．σ^2,τ_j^2 については，事前分布として逆ガンマ分布を用いることで，完全条件付き分布も逆ガンマ分布となることが導出できる．ϕ_j の完全条件付き分布はよく知られた分布形にはならないが，ランダムウォーク型のメトロポリス・ヘイスティングスアルゴリズムを用いることでギブスサンプラーに組み込むことができる．このように，BSPS の MCMC アルゴリズムでは，f_{ji}（各モデルの予測値）の値に基づいて適切なウエイトを更新するステップ（β_{ji}

の生成ステップ)と β_{ji} (各モデルのウエイト)に基づいて予測値を更新するステップ (f_{ji} の生成ステップ)を繰り返し実行することで，モデルのウエイトや予測値に関する不確実性を事後分布を通して自然な形で得ることが可能となる．

MCMC によって事後サンプルが生成できると，任意の地点 s における統合予測を導出することができる．ガウス過程の定式化から，地点 s におけるモデルウエイト $\beta_j(s)$ の予測分布は $N(C_s(\phi_j)^\top C(\phi_j)^{-1}\beta_j, \{\tau_j^2 - \tau_j^2 C_s(\phi_j)^\top C(\phi_j)^{-1} C_s(\phi_j)\}^{-1})$ で与えられる(ただし，$C_s(\phi_j) = (C(\|s - s_1\|;\phi_j), \ldots, C(\|s - s_n\|;\phi_j))^\top$)ため，事後サンプルを用いて $\beta(s)$ の事後サンプルを生成することができる．さらに，地点 s における各モデルの予測分布が得られれば，BSPS のモデル式(3.2)を用いて $y(s)$ の乱数を生成することができる．この乱数の平均を計算することで点予測，分位点を計算することで区間予測を導出することができる．

3.3 最近傍ガウス過程を用いた高速実装

BSPS ではガウス過程を用いるため，β_j の完全条件付き分布から乱数を生成する(A_j^β を計算する)際には $n \times n$ 行列の逆行列を計算する必要があり，そのオペレーション回数は $O(n^3)$ である．したがって，地点数 n が大きい状況では計算コストが膨大なものになってしまい，現実的な計算時間で MCMC を実行することが困難になってしまう．このような問題を解決するための手法として，カーネル関数による重み付き平均，Karhunen-Loève 展開 (e.g. Banerjee et al., 2014, Section 12.3)や予測過程 (Banerjee et al., 2008)などの低ランク近似による方法が広く知られている．一般的に，低ランク近似による方法は，空間トレンドを過剰に平滑化してしまう(oversmoothing)問題が知られている (e.g. Datta et al., 2016)．近年では，低ランク近似によらないガウス過程の近似モデルが提案されているが，この節では特に Datta et al. (2016)で提案された最近傍ガウス過程(nearest-neighbor Gaussian process, NNGP)について解説する．Heaton et al. (2019)では，NNGP を含めた様々な高速化手法が紹介されており，手法間の包括的なレビューについてはこちらを参照されたい．

地点 $s_1, \ldots, s_n$ 上の確率変数をそれぞれ $x_i \equiv x(s_i)$ $(i = 1, \ldots, n)$ とする．通常の平均 0 のガウス過程では，$(x_1, \ldots, x_n) \sim N(0, C(\theta))$ となる．ここで，$C(\theta)$ はパラメータ θ を持つ $n \times n$ の共分散行列である．MCMC の各更新において，$C(\theta)$ の逆行列を扱う必要があり n が大きい状況では計算上のボトルネックとなる．Datta et al. (2016)では，n を固定したもとで，スパースな構造を持つ確率モデルによって元の多変量正規分布を近似する方法を与え，その後一般次元に対して確率分布を拡張することで確率過程として NNGP を構成している．一般に以下のような同時分布の分解公式が成り立つ．

$$p(x(s_1), \ldots, x(s_n)) = p(x(s_1)) \prod_{i=2}^{n} p(x(s_i)|x(s_1), \ldots, x(s_{i-1})).$$

添字 i が大きくなると，条件付ける変数の数が多くなるが，s_i から距離が離れている地点上の確率変数とは相関が小さく，そのような変数を無視したとしても大きな影響はないと考えられる．具体的に，$N(s_i)$ を $s_1, \ldots, s_{i-1}$ の中で地点 s_i から近い上位 m 地点を抽出した地点の集合として定義し，$x_{N(s_i)}$ を集合 $N(s_i)$ に含まれる各地点上の確率変数を並べた m 次元ベクトルとする．このとき，$p(x(s_i)|x(s_1), \ldots, x(s_{i-1}))$ を $p(x(s_i)|x_{N(s_i)})$ で置き換えることで同時分布の近似モデルを与えることができる．

$$\tilde{p}(x(s_1), \ldots, x(s_n)) = p(x(s_1)) \prod_{i=2}^{n} p(x(s_i)|x_{N(s_i)}).$$

$p(x(s_1), \ldots, x(s_n))$ が多変量正規分布の場合を考える．$C_{N(s_i)} \equiv \mathrm{Cov}(x_{N(s_i)})$ を $x_{N(s_i)}$ の分散

共分散行列とし，$C_{s_i,N(s_i)} \equiv \mathrm{Cov}(x(s_i), x_{N(s_i)})$ を x_i と $x_{N(s_i)}$ の共分散ベクトルとする．このとき

$$\begin{aligned}\tilde{p}(x(s_1),\ldots,x(s_n)) &= \phi(x(s_1);0,C(s_1,s_1))\prod_{i=2}^{n} p(x(s_i)|B_{s_i}x_{N(s_i)},F_{s_i}) \\ B_{s_i} &= C_{s_i,N(s_i)}C_{N(s_i)}^{-1}, \quad F_{s_i} = C_{s_i,s_i} - C_{s_i,N(s_i)}C_{N(s_i)}^{-1}C_{N(s_i),s_i}\end{aligned} \tag{3.4}$$

と表すことができる．上記の形で与えられる同時分布は多変量正規分布であり，元の分散共分散行列 $C(\theta)$ とは異なる共分散行列 $\tilde{C}(\theta)$ を持つ．また，$\tilde{C}(\theta)$ の重要な性質として，m が n に対して小さい場合，精度行列 $\tilde{C}(\theta)^{-1}$ がスパース行列(多くても $nm(m+1)/2$ 個の成分のみが非ゼロの行列)になる．一方で，低ランク近似の方法とは異なり，共分散行列 $\tilde{C}(\theta)$ は低ランク構造を持たない．これにより過剰な空間平滑化を防ぐことができる．近傍数 m の選択に関して，Datta et al. (2016) は n が大きいケースでも経験的に $m = 10, 15$ 程度で十分に正確な近似が得られると主張している．

この同時分布を元に，任意の地点集合 $(\tilde{s}_1,\ldots,\tilde{s}_k)$ における確率ベクトル $(x(\tilde{s}_1),\ldots,x(\tilde{s}_k))$ の同時分布を与えることができる．簡単のために，$(\tilde{s}_1,\ldots,\tilde{s}_k)$ と $(s_1,\ldots,s_n)$ は共通の地点を持たないと仮定する．このとき，$(x(\tilde{s}_1),\ldots,x(\tilde{s}_k))$ の同時分布は

$$p(x(\tilde{s}_1),\ldots,x(\tilde{s}_k)) = \int\prod_{l=1}^{k}\tilde{p}(x(\tilde{s}_l)|x(s_1),\ldots,x(s_n))\tilde{p}(x(s_1),\ldots,x(s_n))\prod_{i=1}^{n}dx(s_i)$$

で与えられる．ここで

$$\tilde{p}(x(\tilde{s}_l)|x(s_1),\ldots,x(s_n)) = \phi(x(\tilde{s}_l);B_{s_l}x_{N(x_l)},F_{s_l})$$

である．このように，任意の地点集合上で同時確率分布を定義することができ，確率過程として NNGP を構成することができる．

BSPS の推定を高速化するため，$\beta_j = (\beta_{j1},\ldots,\beta_{jn})$ の同時分布に対して以下のような NNGP を用いる．

$$p(\beta_{j1},\ldots,\beta_{jn}) = \prod_{i=1}^{n}\phi(\beta_{ji};B_{ji}\beta_j(N(s_i)),F_{ji}), \quad j=0,\ldots,J.$$

ただし

$$\begin{aligned}B_{ji} &= C_{s_i,N(s_i)}(\phi_j)C_{N(s_i)}(\phi_j)^{-1}, \\ F_{ji} &= \tau_j^2 - \tau_j^2 C_{s_i,N(s_i)}(\phi_j)C_{N(s_i)}(\phi_j)^{-1}C_{N(s_i),s_i}(\phi_j)\end{aligned}$$

であり，$N(s_i)$ は地点 s_i の m 近傍の地点の集合，$\beta_j(N_{s_i})$ は $s \in N(s_i)$ に対する $\beta_j(s)$ を並べた m 次元ベクトルを表す．また，$C_{s_i,N(s_i)}(\phi_j) = \mathrm{Cor}(\beta_j(s_i),\beta_j(N(s_i)))$ および $C_{N(s_i)}(\phi_j) = \mathrm{Cor}(\beta_j(N(s_i)),\beta_j(N(s_i)))$ である．このとき，$(\beta_{0i},\beta_{1i},\ldots,\beta_{Ji})$ の完全条件付き分布は $N(A_i^{(\beta)}B_i^{(\beta)}, A_i^{(\beta)})$

$$\begin{aligned}A_i^{(\beta)} &= \left\{\frac{f_i f_i^\top}{\sigma^2} + \mathrm{diag}(\gamma_{0i},\ldots,\gamma_{Ji})\right\}^{-1}, \quad \gamma_{ji} = \frac{1}{\tau_j F_{ji}} + \sum_{t;s_i\in N(s_t)}\frac{B_j(t;s_i)^2}{\tau_j F_{jt}}, \\ B_i^{(\beta)} &= \frac{f_i y_i}{\sigma^2} + (m_{0i},\ldots,m_{Ji})^\top, \\ m_{ji} &= \frac{B_{ji}^\top\beta_j(N(s_i))}{\tau_j F_{ji}} + \sum_{t;s_i\in N(s_t)}\frac{B_j(t;s_i)}{\tau_j F_{jt}}\left\{\beta_{jt} - \sum_{s\in N(s_t),s\neq s_i}B_j(t;s)\beta_j(s)\right\}\end{aligned}$$

で与えられる．ただし，$f_i = (1, f_{1i}, \ldots, f_{Ji})$ であり，$B_j(t;s)$ は係数ベクトル B_{jt} のうち $\beta_j(s)$ に対する係数を表す．f_{ji} の完全条件付き分布は 3.2 節と同様である．

3.4 他手法との比較

Cabel et al. (2022)では，数値実験やデータ解析例を通して BSPS と既存手法の予測精度を比較している．数値実験では，モデルの重要度が領域ごとに異なるシナリオにおいて予測精度の比較を行っており，(モデルの重要度の空間的異質性を考慮していない)既存手法に対する BSPS の優位性が数値的に示されている．また，データ解析においては，交差検証による精度比較を実施しており，その場面でも BSPS の予測精度が既存手法を上回ることが示されている．詳細については Cabel et al. (2022)を参照されたい．

4. まとめと議論

ベイズ的予測統合は空間予測の他にも時系列データ予測（e.g. McAlinn and West, 2019; Kobayashi et al., 2023)や因果推論（Sugasawa et al., 2023)などのタスクにおいて有効性が確認されている統合方法である．ベイズ的予測統合の実用上の限界点として，統合するモデルに対して何らかの不確実性が定量化されている必要がある点が挙げられる．機械学習的アプローチの中には点予測のみを与える手法も多く，それらの方法をベイズ的予測統合の枠組みで統合する場合は，ブートストラップ法などを使って予測の不確実性(例えば予測分散)を計算する必要があり，追加的な計算コストがかかってしまう点に注意が必要である．また，3 節の BSPS を実行するためには MCMC を用いる必要があり，(NNGP によって高速化をしているとはいえ)ある程度の計算時間がかかってしまうことが想定される．時系列予測の文脈では，時間ごとにモデルのウエイトが変化することを考慮しつつ，高速に逐次予測をするアルゴリズムなども開発されている（Bernaciak and Griffin, 2024)ため，空間データに対して同様なアプローチを開発していくことも今後の重要な課題になると考えられる．

参 考 文 献

Anselin, L. (2022). *Spatial Econometrics, Handbook of Spatial Analysis in the Social Sciences*, Edward Elgar Publishing, Glos, UK.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets, *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **70**, 825–848.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, CRC Press, Boca Raton, Florida, USA.

Bernaciak, D. and Griffin, J. E. (2024). A loss discounting framework for model averaging and selection in time series models, *International Journal of Forecasting*, **40**, 1721–1733.

Breiman, L. (1996). Stacked regressions, *Machine Learning*, **24**, 49–64.

Cabel, D., Sugasawa, S., Kato, M., Takanashi, K. and McAlinn, K. (2022). Bayesian spatial predictive synthesis, *arXiv preprint*, arXiv:2203.05197 (最終アクセス日 2024 年 10 月 23 日).

Clarke, B. (2003). Comparing Bayes model averaging and stacking when model approximation error cannot be ignored, *Journal of Machine Learning Research*, **4**, 683–712.

Clydec, M. and Iversen, E. S. (2013). Bayesian model averaging in the M-open framework, Bayesian Theory and Applications (eds. P. Damien, P. Dellaportas, N. G. Polson and D. A. Stephens), Oxford University Press, Oxford, UK.

Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian

process models for large geostatistical datasets, *Journal of the American Statistical Association*, **111**, 800–812.

Debarsy, N. and LeSage, J. P. (2022). Bayesian model averaging for spatial autoregressive models based on convex combinations of different types of connectivity matrices, *Journal of Business & Economic Statistics*, **40**, 547–558.

Du, P., Bai, X., Tan, K., Xue, Z., Samat, A., Xia, J., Li, E., Su, H. and Liu, W. (2020). Advances of four machine learning methods for spatial data handling: A review, *Journal of Geovisualization and Spatial Analysis*, **4**, 1–25.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2002). *Geographically Weighted Regression*: *The Analysis of Spatially Varying Relationships*, Wiley, New York.

Genest, C. and Schervish, M. J. (1985). Modelling expert judgements for Bayesian updating, *Annals of Statistics*, **13**, 1198–1212.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M. et al. (2019). A case study competition among methods for analyzing large spatial data, *Journal of Agricultural, Biological and Environmental Statistics*, **24**, 398–425.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with comments by M. Clyde, David Draper and EI George, and a rejoinder by the authors), *Statistical Science*, **14**, 382–417.

Kobayashi, G., Sugasawa, S., Kawakubo, Y., Han, D. and Choi, T. (2023). Predicting COVID-19 hospitalisation using a mixture of Bayesian predictive syntheses, *arXiv preprint*, arXiv:2308.06134 (最終アクセス日 2024 年 10 月 23 日).

Le, T. and Clarke, B. (2017). A Bayes interpretation of stacking for M-complete and M-open settings, *Bayesian Analysis*, **12**, 807–829.

LeSage, J. P. and Parent, O. (2007). Bayesian model averaging for spatial econometric models, *Geographical Analysis*, **39**, 241–267.

McAlinn, K. and West, M. (2019). Dynamic Bayesian predictive synthesis in time series forecasting, *Journal of Econometrics*, **210**, 155–169.

Oliver, M. A. and Webster, R. (1990). Kriging: A method of interpolation for geographical information systems, *International Journal of Geographical Information System*, **4**, 313–332.

Sugasawa, S., Takanashi, K., McAlinn, K. and Edoardo, A. (2023). Bayesian causal synthesis for meta-inference on heterogeneous treatment effects, *arXiv preprint*, arXiv:2304.07726 (最終アクセス日 2024 年 10 月 23 日).

Vehtari, A., Simpson, D., Gelman, A., Yao, Y. and Gabry, J. (2024). Pareto smoothed importance sampling, *Journal of Machine Learning Research*, **25**, 1–57.

West, M. (1992). Modelling agent forecast distributions, *Journal of the Royal Statistical Society* (*Series B*: *Methodological*), **54**, 553–567.

West, M. and Crosse, J. (1992). Modelling of probabilistic agent opinion, *Journal of the Royal Statistical Society* (*Series B*: *Methodological*), **54**, 285–299.

Yao, Y., Vehtari, A., Simpson, D. and Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion), *Bayesian Analysis*, **13**, 917–1003.

Zhang, L., Tang, W. and Banerjee, S. (2023). Exact Bayesian geostatistics using predictive stacking, *arXiv preprint*, arXiv:2304.12414 (最終アクセス日 2024 年 10 月 23 日).

Bayesian Model Synthesis for Spatial Prediction

Shonosuke Sugasawa

Faculty of Economics, Keio University

Estimating a model from spatial data observed at finite locations and predicting unobserved locations are the crucial tasks in spatial data analysis. Recently, various models ranging from classical geostatistics and spatial econometrics to machine learning approaches have become available. Therefore, selecting appropriate analysis methods for the data is a significant issue in spatial data analysis. This paper reviews the methodology of Bayesian model synthesis for spatial prediction in situations where multiple predictive models are obtained. Specifically, we discuss the differences between recently proposed Bayesian spatial predictive synthesis and classical synthesis methods, as well as provide explanations of concrete estimation algorithms.

Key words: Bayesian model averaging, stacking, Gaussian process, Bayesian predictive synthesis.