統計数理研究所 75 周年記念事業（2019）
©2019 統計数理研究所

大規模空間データのための空間可変係数モデリング

Spatially varying coefficient modeling for large dataset

データ科学研究系 村上 大輔 （Daisuke Murakami）

1. 背景

センサ技術の発展に伴い，地理空間データが急速に大規模化してきている．例えば土地被覆や気象関連の情報は高解像度の衛星画像として提供されており，世帯構成や産業構成のような社会経済統計は街区のような空間詳細な単位毎に整備されつつある．それら標本数 N の大きな地理空間データを柔軟にモデル化する方法が研究・実務で求められている．

Gaussian process (GP) は，地理空間データの背後にある空間過程をモデル化するために幅広く用いられてきた．GP の共分散が距離減衰関数に従うことを仮定することで，地点間の空間的な従属関係を柔軟にモデル化することができる．一方で，GP を推定するためには $N \times N$ 次元の共分散行列の逆行列を評価する必要があり，標本数 N が数百万あるいはそれ以上になりうる昨今，その計算負荷は実用上の課題となってきている．そのような中，GP を高速に推定するための近似手法が近年活発に議論されてきた．

本節では，GP に基づいた空間モデルの高速化の一環といて取り組んでいる Spatially varying coefficient (SVC) モデルの高速化に関する研究を紹介する．

2. SVC モデルとその高速化

SVC モデルとは，各回帰係数の背後に GP を仮定することで回帰係数を場所毎に推定しようというモデルである．例えば都市部と郊外部の違いなどを柔軟に捉えることができるなど，SVC モデルは実用上便利である．しかしながら，回帰係数毎の GP を推定しようという同モデルの計算量は極めて大きいことが知られている．そこで本研究では，大規模データへの適用を見据え，SVC モデルの推定を次の手順で高速化した：(i) 各 GP を主成分のみを残して低ランク近似する；(ii) サイズが N に依存する行列・ベクトルを予め処理する（内積をとる）ことで，N に依存しない計算量で評価できるように事後確率を書き直す；(iii) 同事後確率を逐次的に最大化していくことで回帰係数の空間分布を決める各 GP(低ランク) を推定する．手順 (i) は上述の逆行列の計算量を削減するための近似，手順 (ii) と (iii) は各 GP を特徴づけるパラメータの推定を高速化するための処理である．

以上で高速化した SVC モデルの計算時間を，ベイズ SVC モデル（提案手法はこれを近似）および地理的加重回帰モデル（従来手法）と比較した．計算はすべて統計ソフトウェア R 上で行った．計算時間の比較結果は図 1 に示すとおりである．同図より，ベイズ SVC モデルは計算負荷が極めて大きく，実用上難があることを確認した．従来手法の計算時間もまた標本数 N が大きくなるにしたがって急激に増加しており，大規模データのモデリングの観点からは課題が残されているとの示唆を得た．

対照的に，提案手法の計算時間は N に関して線形にしか増加しておらず，例え SVC 数 K が 8 の（8 個の GP を同時推定する）場合であってもその増加は極めて遅い．例えば $N = 5,000$

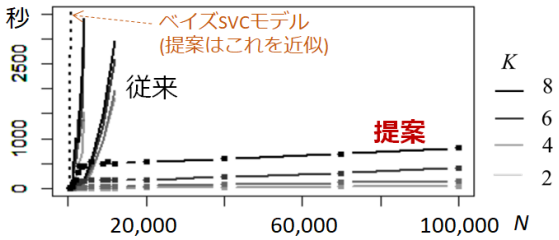

図 1. 計算時間の比較（N は標本数，K は SVC の数）.

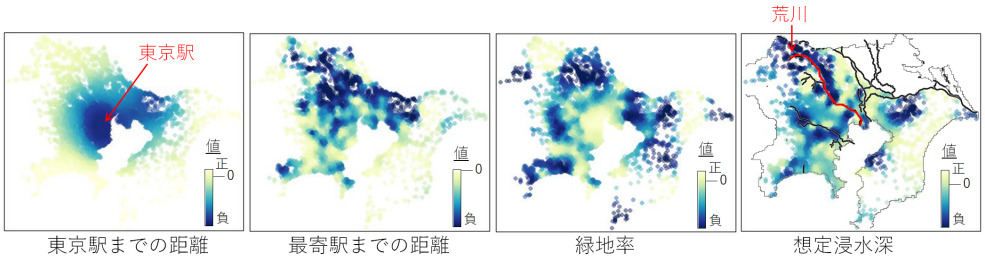

図 2. 推定された場所毎の回帰係数

の場合の提案モデルの推定時間は平均で 454 秒となり，従来モデルが適用困難な $N = 100,000$ の場合でも，その推定時間はわずか 836 秒（平均値）となった．提案手法を用いることで，計算時間が大幅に短縮されることを確認した．なお，多くの場合に提案手法の SVC 推定精度が従来手法を上回ることもまたモンテカルロシミュレーションにより確認している．

3. 住宅地価分析への応用

提案モデルを東京都市圏の住宅地公示地価 (2010) 年の分析に適用した．説明変数は東京駅までの距離，最寄駅までの距離，1km グリッド内の緑地率，想定浸水深である．各データは国土数値情報ダウンロードサービス（http://nlftp.mlit.go.jp/ksj/）で公開されている．

推定された回帰係数（SVC）を図 2 にプロットした．この図から東京駅までの距離は都心を中心とした広域的な影響パターンを持つことが確認された．対照的に，最寄り駅までの距離は郊外部での影響力が大きく，特に鉄道網が比較的疎な北部ではその影響力が強まっていることが確認できる．緑地率もまた郊外部で強い効果を持つが，その影響は負であり，緑地の多さは地価を低下させるという結果が得られた．これは緑地よりも都市施設の多い土地のほうが好まれる傾向があるためである可能性がある．なお，例外的に都心部と横浜市中心部では緑地率は正に有意となっており，それら地域では緑地が好まれているの示唆を得た．想定浸水深は荒川沿岸の地価をより強く低下させていると推定された．この結果は，荒川沿岸の地価が水害リスクが適切に反映された値付けになっており，水害リスクにより適応した都市パターンであることを意味するものである．以上の結果は直感に整合する．