様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １６ 日現在

機関番号：６２６０３
研究種目：若手研究(B)
研究期間：2017～2019
課題番号：１７Ｋ１２９７４

研究課題名（和文）時空間データの大規模化・多様化に向けた固有ベクトル空間フィルタリングの高度化

研究課題名（英文）Extension of eigenvectir spatial filtering approaches for large and diverse spatiotemporal datasets

研究代表者
村上　大輔（Murakami, Daisuke）

統計数理研究所・データ科学研究系・助教

研究者番号：２０７３８２４９

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究では，申請者提案の空間回帰法であるRE-ESFを拡張することで、既存の空間統計手法では困難であった大規模かつ多様なデータの（時）空間回帰モデリングを高速・柔軟に行うためのアプローチを開発した。そのために、まずはpre-conditioningを駆使してRE-ESEの高速化と省メモリ化を行い大規模データに適用可能とした。次に、同手法を、時空間データ、階層性を持つデータ、非ガウスデータなどに適用可能とするための拡張を行った。さらに、幅広い実データへの応用を通して提案手法の有用性を確認した。最後に、以上で開発した各手法を統計ソフトウェアRのパッケージspmoranに実装して一般公開した。

研究成果の学術的意義や社会的意義
近年急増する大規模な地理空間データを柔軟に解析するための空間回帰法を幅広く開発した。空間回帰法は空間疫学、空間計量経済学、計量地理学といった関連分野の高度化を、大規模データの解析手法の高度化の観点から後押ししうるものである。開発した各手法は既に統計ソフトウェアRのパッケージ化して一般公開済みであり、既に幅広い関連研究者に利用されている（例えば2019年度は7684回ダウンロードされた）。以上に加え、本研究で提案した空間回帰法は計算時間とメモリ消費が極めて小さく学術的にも新規的である。

研究成果の概要（英文）：This study develops fast and flexible spatial (and spatiotemporal) regression approaches for large and diverse geo-spatial datasets. This development is done by extending the random effects eigenvector spatial filtering (RE-ESF) approach, which is a spatial regression approach. First, we improve computational efficiency of RE-ESF by incorporating a pre-conditioning algorithm. Then, the memory consumption is drastically reduced for modeling very large samples through parallelization. After that, the developed fast approach is extended for spatio-temporal data, hierarchical data, non-Gaussian data, and other data by introducing latent variables capturing data properties. Usefulness of the proposed approach is verified by applying it to a wide variety of spatial and spatiotemporal data modeling. Finally, all the developed methods are implemented an R package spmoran to make them available for public.

研究分野：空間統計

キーワード：空間回帰

１．研究開始当初の背景

センサ技術の発達に伴う空間データの多様化・大規模化は著しく、それに応じた空間解析手法の高度化が求められている。特に、地理学、生態学、空間疫学を含む幅広い分野で利活用されている空間回帰法の高度化は喫緊の課題である。空間データの基本特性である空間的従属性や空間的異質性を考慮した解析の理論や方法は、空間統計学（spatial statistics）を中心に発展を遂げてきた。しかしながら、従来の空間回帰法の計算量は $O(n^3)$（標本数 n の 3 乗に比例して増加）であり非常に大きい。また、空間統計学では定常なデータが分析対象とされてきた経緯があり。分布・精度・構造の異なる幅広い時空間データのための回帰法の理論・方法は未だ発展途上である。以上ように、空間統計学の手法（特に空間回帰法）は必ずしも大規模化・多様化した最近の時空間データの分析に適合している言い難い。

空間統計学の比較的新しい一手法として、空間的な近接性を記述する行列の固有ベクトルを用いてデータの空間変動をモデル化しようという空間回帰法 eigenvector spatial filtering（ESF）が存在する。ESF の有用性は数多くの比較研究で指摘されてきてきた。しかしながら、ESF もまた計算量が巨大であり大規模データには適用できない。加えて ESF にはモデルの高度化に伴って不安定化するという傾向があり、多様なデータのための拡張の観点からも ESF には課題が残されている。

２．研究の目的

本研究では、申請者提案の空間回帰法 random effects eigenvector spatial filtering（RE-ESF）を発展的に応用することで、既存手法では困難であった大規模（例えば標本数 ＞10 万）かつ多様（な時空間データの解析を高速・柔軟に行うことのできる新たな方法を幅広く開発する。そのために、まずは(i)大規模データへの応用のための RE-ESF の高速化を行う。次に、(ii)同計算効率を保ちながら、多様な空間解析・時空間解析に応用できるように RE-ESF を幅広く拡張する。最後に(iii)開発した方法論を統計解析ソフト R のパッケージとして整理して一般公開する。

３．研究の方法

前節の目的(i)、(ii)、(iii)を達成するための方法について順に説明する。

(i)　大規模データモデリングのための RE-ESF の高速化

空間統計学ならびに応用統計学の関連文献を幅広くレビューした上で、回帰の既存の高速化手法を分野横断的に整理する。その上で、計算時間短縮とメモリ消費削減の両方を達成するような、RE-ESF の推定アルゴリズムを開発する。なお、一般に時間計算量が $O(n)$（標本数 n に比例して計算量が増加）またはそれ以下であることが、大規模空間データモデリングに求められる要件であることから、これを満たすアルゴリズムの開発を前提とする。

開発したアルゴリズムの性能は、既存の高速手法との比較を通して検証する。

(ii)　多様なデータのモデリングのための RE-ESF の拡張

時空間データ、階層性を持つデータ（例：マンション価格データは戸別属性と棟別属性からなる階層性を有するデータ）、非ガウスデータを含む幅広い（時）空間データをモデリングするための RE-ESF の拡張を行う。(i)で開発した高速推定アルゴリズムが適用可能な形での拡張を進めることで、大規模かつ多様な（時）空間データが柔軟に解析可能となるように RE-ESF を高度化していく。拡張した各手法の有用性を検証するために、それらを幅広い実データの分析に応用する。

(iii)　統計ソフトウェア R のパッケージとしての成果物の公開

上記(i)、(ii)で開発する各手法を統計ソフトウェア R のパッケージ spmoran 上に実装・公開する。ユーザーからのフィードバックも踏まえながら同パッケージの改良を進めていく。また、同パッケージの使い方や分析例をまとめたマニュアルを作成・公開することでユーザーを増やしていくことを目指す。

４．研究成果

(i)　大規模データモデリングのための RE-ESF の高速化

文献レビューの結果を踏まえ、pre-conditioning と呼ばれるアプローチを応用して空間回帰法 RE-ESF を高速化することとした。同アプローチは、RE-ESF の推定に必要となる尤度最大化を、データを直接的に用いて行うのではなく、データを内積に変換した後に行うものである。内積の次元は標本数に依存しないという特性を活かし、例えば標本数が数十万の大規模データであっても高速に RE-ESF モデルが推定できるようにした。

計算効率を確認するために、ここでは RE-ESF を用いた Spatially varying coefficients（SVC; 場所毎に変化する回帰係数）の推定に着目して、提案手法（RE-ESF）と従来手法（地理的加重加重）の計算時間を比較した。結果を要約した図 1 より、従来手法の計算時間は標本数が増えるに伴って指数的に増加するのに対して、提案手法の計算時間の増加は極めて小さいことが確認できる。以上より、提案手法の計算効率の良さを確認した。

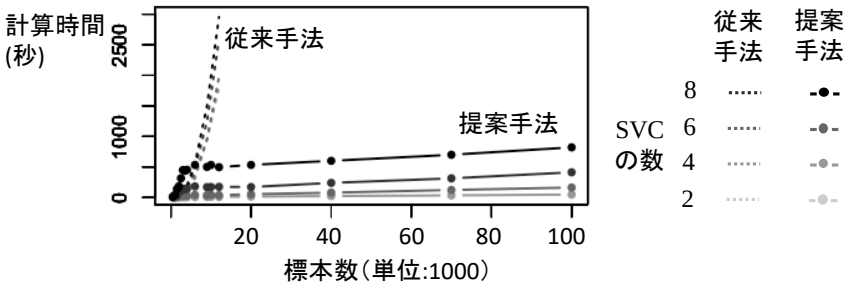


図 1：提案手法（RE-ESF）と従来手法（地理的加重回帰）との計算時間の比較。推定する SVC の数が大きくなるにつれて計算負荷は大きくなる。

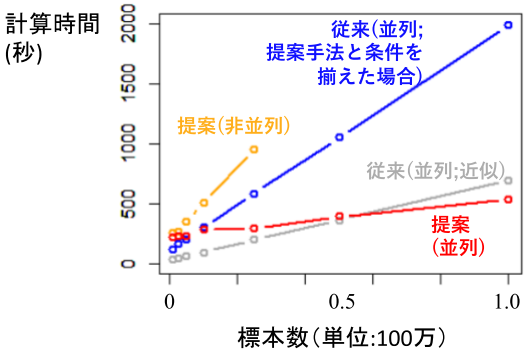


図 2 ：提案手法と従来手法の pre-conditioning 手法との計算時間の比較

以上で提案手法の計算効率を確認した。一方で、提案手法は標本数の増加に応じて消費メモリが増加していくという特性を有する。従って、例えば標本数数百万のデータには適用できない。衛星観測データや携帯 GPS データといった数十から数千万の標本からなるデータも増えてきていることを踏まえると、消費メモリーはできる限り削減する必要がある。この点を踏まえ、本研究では、一度にデータを処理するのではなく、提案手法で用いた pre-conditioning を並列化して、部分ごとに処理するというアルゴリズムに拡張することで、理論上のメモリ上限の存在しないアルゴリズムに拡張した。

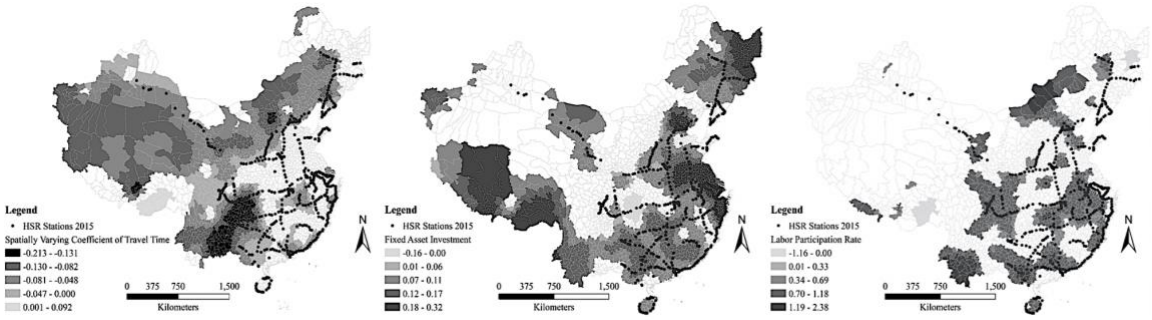
すでに類似の高速推定アルゴリズムが存在したため、同既存手法を提案手法と比較した。結果を要約した図 2 内の提案（並列）が提案手法、従来（並列）が従来手法である。この図より、既存の高速推定手法に比べ、標本数の増加に伴う計算時間の増加が小さく、極めて計算効率が良いこと、ならびに 100 万標本のデータでも 500 秒程度でモデルが推定できることを確認した。なお、1000 万標本のデータでも 70 分ほどでモデルが推定できることも確認している。

以上で提案した高速な空間回帰アルゴリズムは、大規模な地理空間データを柔軟に解析する上で有用と言えよう。

(ii)　多様なデータのモデリングのための RE-ESF の拡張

上記(i)で開発した高速推定アルゴリズムが適用可能な形で拡張していくことで、多様な時空間データを取り扱うことのできるように RE-ESF を拡張した。具体的には次のモデル開発を行った：(ii-1)時間効果と空間効果を捉える潜在変数を導入することで時空間データがモデリング可能となるように RE-ESF を拡張した；(ii-2)階層性を持つデータや小地域データの解析が可能となるように、グループ効果を捉える潜在変数を導入した；(ii-3)非ガウス分布に従う現象、特に異常気象などの極端現象がモデリング可能となるように、無条件分位点回帰モデルと組み合わせた RE-ESF の拡張を行った。

次に、 (ii-1)、(ii-2)、(ii-3)で開発した各手法の有用性を検証するために、幅広い実データに応用した。まず、(ii-1)時空間データに対する有用性の検証を行うために、中国の地域経済指数（郡別；2008～2015 年）の要因分析に拡張手法を用いた。結果の一部を図 3 に示す。図 3 はアクセシビリティ、固定資産投資、労働投入が経済成長に及ぼした効果（郡別）の推定結果をまとめた図でありであり、統計的に有意な効果が見られた群のみを着色している。なお、図中の黒点は高速鉄道路線の駅を表す。図 3 より、都市の集積する東部では固定資産投資や労働投入といった郡毎の経済変数が強い影響を及ぼしていること、並びにアクセシビリティは交通網の整備が遅れている西部で強い影響力を持つことを確認した。以上の結果は直感に整合する。提案手法から直感に整合する時空間モデリングの結果が得られることを確認した。



アクセシビリティの効果　　固定資産投資の効果　　労働投入の効果

図３：提案手法の地域経済指数（郡別；2008～2015 年）に対する応用結果

次に、(ii-2)階層性を持つデータに対する有用性を検証するために、ボストン住宅価格データ（非集計）を用いた地区毎（集計単位）の価格指数の評価に提案手法を応用した。結果を図４に要約した。図中の y は地区毎の住宅価格の集計値、pred が地区毎の価格指数の評価結果を表す。つまり pred は観測値 y からノイズを除去して推定された価格指数である。図４より、価格指数は概ね観測値と類似する点や、特に価格の高い（黄色い）地区を下方修正しているなど、一定の違いも確認できる。図４の xgroup は地区毎のグループ効果（各地区は独立と仮定して推定）、sf_residuals は地区毎の空間相関成分の推定結果である。この図より、xgroup が局所的な高価格街区の特徴を捉える成分、sf_residuals が価格の滑らかな空間パターンを捉える成分となっており、効果が適切に分離されたことが確認できた。以上より、提案手法は階層性を有するデータの解析にも役立つことを確認した。

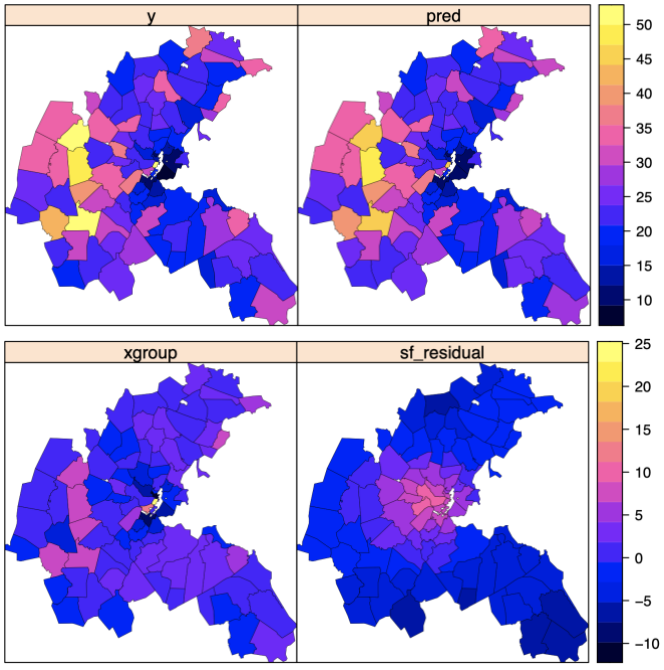


図４：提案手法の地区別の価格指数推定への応用結果

最後に(iii)極値事象の解析への応用として水害リスクの評価に提案手法を拡張した。結果の一部を図５に示す。同図は水害リスクが住宅価格に及ぼした影響を価格帯毎（横軸）に推定した結果であり、赤が従来手法（S-CQR）、黒が提案手法（SF-UQR）である。また黒の実線が影響の推定結果、灰色の領域がその 95%信頼区間である。図５より、従来手法とは異なり、提案手法の推定結果は、低価格帯において水害リスクが正の影響を持つという傾向、つまり「水害リスクの高いところに好んで居住する」という傾向が推定された。この結果は水害リスクの過小評価に起因するものと考えられる。この結果に基づけば、低価格帯に属する街区の水害リスクの対策が喫緊の課題と言えよう。

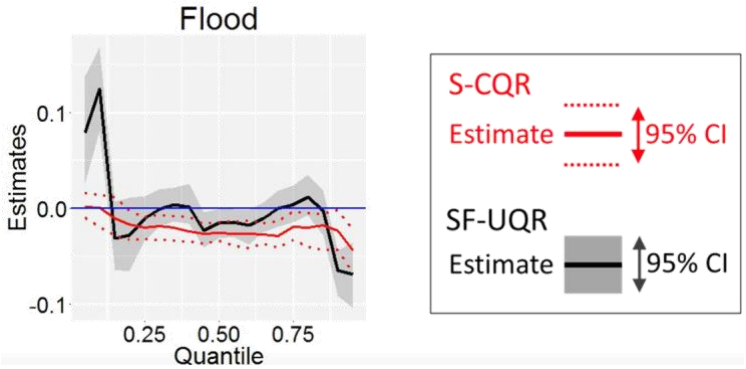


図５：水害リスク（Flood）が住宅価格に及ぼした影響の推定結果。赤が従来手法（S-CQR）で黒が提案手法（SF-UQR）である。

以上の分析を通して、本研究で開発・高度化してきた RE-ESF が幅広い（時）空間データの解析に役立つことを確認した。

(iii) 統計ソフトウェア R のパッケージとしての成果物の公開

以上の(i)、(ii)で開発した推定アルゴリズムや空間回帰法 RE-ESF は全て R パッケージ spmoran（https://cran.r-project.org/web/packages/spmoran/index.html）に実装済みであり、誰でも利用可能な状態とした（図６）。同パッケージのダウンロード数は 2017 年度に 2,783 回、2018 年度に 6,111 回、2019 年度に 7,684 回であり、着実に利用者が増えてきていることを確認している。なお、同パッケージの使い方や分析例をまとめたマニュアルも図６のページ上で公開済みである。

spmoran: Moran Eigenvector-Based Scalable Spatial Additive Mixed Modeling

Functions for estimating Moran eigenvector-based spatial additive mixed models, and other spatial regression models. For details see Murakami (2020) <arXiv:1703.04467>.

Version:	0.2.0-1
Imports:	sp, fields, vegan, Matrix, doParallel, foreach, ggplot2, spdep, rARPACK, RColorBrewer, splines, methods
Suggests:	R.rsp, rgdal
Published:	2020-05-31
Author:	Daisuke Murakami
Maintainer:	Daisuke Murakami <dmuraka at ism.ac.jp>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
NeedsCompilation:	no
In views:	Spatial
CRAN checks:	spmoran results

Downloads:

Reference manual:	spmoran.pdf
Vignettes:	Spatial regression using the spmoran package: Boston housing price data examples spmoran: An R package for Moran eigenvector-based scalable spatial additive mixed modeling
Package source:	spmoran_0.2.0-1.tar.gz

図６：spmoran パッケージ（https://cran.r-project.org/web/packages/spmoran/index.html）の公開ページ。最も標準的な R パッケージのレポジトリである CRAN 上で公開している。

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　3件／うち国際共著　2件／うちオープンアクセス　0件）

１．著者名 Murakami Daisuke、Lu Binbin、Harris Paul、Brunsdon Chris、Charlton Martin、Nakaya Tomoki、Griffith Daniel A.	４．巻 109
２．論文標題 The Importance of Scale in Spatially Varying Coefficient Modeling	５．発行年 2019年
３．雑誌名 Annals of the American Association of Geographers	６．最初と最後の頁 50～70
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/24694452.2018.1462691	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Murakami D.、Seya H.	４．巻 印刷中
２．論文標題 Spatially filtered unconditional quantile regression: Application to a hedonic analysis	５．発行年 2019年
３．雑誌名 Environmetrics	６．最初と最後の頁 e2556～e2556
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/env.2556	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Murakami Daisuke、Griffith Daniel A.	４．巻 51
２．論文標題 Eigenvector Spatial Filtering for Large Data Sets: Fixed and Random Effects Approaches	５．発行年 2019年
３．雑誌名 Geographical Analysis	６．最初と最後の頁 23～49
掲載論文のDOI（デジタルオブジェクト識別子） 10.1111/gean.12156	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Murakami Daisuke、Yoshida Takahiro、Seya Hajime、Griffith Daniel A.、Yamagata Yoshiki	４．巻 19
２．論文標題 A Moran coefficient-based mixed effects approach to investigate spatially varying relationships	５．発行年 2017年
３．雑誌名 Spatial Statistics	６．最初と最後の頁 68～89
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.spasta.2016.12.001	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Murakami Daisuke、Griffith Daniel A.	4．巻 30
2．論文標題 Spatially varying coefficient modeling for large datasets: Eliminating N from spatial regressions	5．発行年 2019年
3．雑誌名 Spatial Statistics	6．最初と最後の頁 39～64
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.spasta.2019.02.003	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Yu Danlin、Murakami Daisuke、Zhang Yaojun、Wu Xiwei、Li Ding、Wang Xiaoxi、Li Guangdong	4．巻 133
2．論文標題 Investigating high-speed rail construction's support to county level regional development in China: An eigenvector based spatial filtering panel data analysis	5．発行年 2020年
3．雑誌名 Transportation Research Part B: Methodological	6．最初と最後の頁 21～37
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.trb.2019.12.006	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　7件）

1．発表者名 Murakami D.、Seya H.、Griffith Daniel A.
2．発表標題 Low rank spatial econometric models
3．学会等名 XII World Conference of the Spatial Econometrics Association（国際学会）
4．発表年 2018年

1．発表者名 Murakami, D., Yamagata, Y., Daniel, A. Griffith
2．発表標題 Spatially varying coefficient model for large dataset: a rank reduction approach
3．学会等名 11th World Conference of the Spatial Econometrics Association（国際学会）
4．発表年 2017年

１．発表者名
Murakami, D., Yamagata, Y.

２．発表標題
Heatwave risk estimation with spatial BigData: A case study in Tokyo

３．学会等名
Spatial Statistics 2017: One World: One Health.（国際学会）

４．発表年
2017年

１．発表者名
Murakami, D., Yoshida, T., Daniel A. Griffith

２．発表標題
A Moran coefficient-based mixed effect approach to investigate spatially varying relationships

３．学会等名
2017 IASC-NZSA Joint Conference. Auckland（国際学会）

４．発表年
2017年

１．発表者名
村上大輔, Paul Harris, Binbin Lu, 中谷友樹

２．発表標題
The importance of scale in spatially varying coefficient modeling

３．学会等名
地理情報システム学会第26回研究発表大会

４．発表年
2017年

１．発表者名
Tsutsumi, M., Murakami, D.

２．発表標題
Parsimonious Modeling in Spatial Statistics and Spatial Econometrics

３．学会等名
2017年度 統計関連学会連合大会

４．発表年
2017年

１．発表者名
Murakami, D., Nakaya, T., Tsutsumida, N., Yoshida, T.

２．発表標題
Spatially varying coefficient modeling for large data: A case study of residential land price in Tokyo

３．学会等名
International Conference on Spatial Analysis and Modeling（国際学会）

４．発表年
2018年

１．発表者名
Murakami, D.

２．発表標題
Spatial regression modelling for large dataset: A precompression approach

３．学会等名
Workshop on High Dimensional and Bayesian Inference toward Quantifying Real World Uncertainties（国際学会）

４．発表年
2019年

１．発表者名
Murakami, D., Griffith, D.A.

２．発表標題
A precompression approach for fast spatial mixed effects modeling

３．学会等名
Spatial Statistics 2019（国際学会）

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
spmoran（オープンソースの統計ソフトウェアRのパッケージ）
https://cran.r-project.org/web/packages/spmoran/index.html

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考