

令和 3 年 6 月 1 日現在

機関番号：12601

研究種目：若手研究

研究期間：2018～2020

課題番号：18K12757

研究課題名（和文）グループデータ解析の安定化のための統計的方法論

研究課題名（英文）Statistical Methodology for Stabilizing Grouped Data Analysis

研究代表者

菅澤 翔之助（Sugasawa, Shonosuke）

東京大学・空間情報科学研究センター・准教授

研究者番号：50782380

交付決定額（研究期間全体）：（直接経費） 3,200,000円

研究成果の概要（和文）：個人や地域などの属性に基づいてグループ分けされる形式のデータはグループデータと呼ばれ様々な科学分野において頻繁に扱われるデータである。本研究ではこのようなデータに対する既存の統計分析手法が抱える問題点、特にモデルの柔軟性の制約、大規模データにおける計算コスト、外れ値が存在するもとの頑健性の問題などに注目し、効果的な解決策となる手法を開発した。また開発した手法の理論的性質も明らかにした。

研究成果の学術的意義や社会的意義

本研究結果によってグループデータの分析において障壁となっていた応用上の問題点に対して、いくつかの効果的な解決策を提示することができた。提案した手法の多くは実装したコードをオープンソースとして公開しており、グループデータを扱う関連分野の研究者や実務家が提案した方法を利用することで、これまでよりも効果的な統計分析を実行することが可能になると考えられる。

研究成果の概要（英文）：Data in the form of groupings based on attributes such as individuals and regions are called group data, and are frequently used in various scientific fields. In this study, we focused on the problems of the existing statistical analysis methods for such data, especially the limitation of model flexibility, computational cost for large-scale data, and robustness in the presence of outliers, and developed several effective methods to solve them. We also revealed some theoretical properties of the developed methods.

研究分野：経済統計学

キーワード：統計モデリング ロバスト統計 空間統計

科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

1. 研究開始当初の背景

個人や地域などの属性に基づいてグループ分けされる形式のデータはグループデータ (grouped data) またはクラスターデータ (clustered data) と呼ばれ、社会科学をはじめ、医学や薬学などの様々な科学分野において頻繁に扱われるデータである。例えば、ある個体を複数時間観測したパネルデータは、データを個体ごとにグループ分けすることができるため、グループデータの一つと言える。また、地域別の住民の経済状況を把握するために任意抽出された住民の所得データは、地域ごとにグループ化されているとみなすことができる。このようなグループデータでは、異なるグループ間に構造の異質性が存在し、異質性を無視した統計分析は、推定結果にバイアスが生じる危険性や、推定精度に致命的な悪影響を及ぼす可能性がある。

このような異質性を考慮する代表的なモデルのクラスとして階層モデルがあり、グループデータに対する効果的な統計分析手法として中心的な役割を果たしている。しかし、階層モデルにおいては、パラメトリックなモデルの仮定の制約の強さの問題、大規模データにおける計算コストの問題、外れ値が存在するもとの推定の頑健性の問題など、応用上重要であるが未解決な問題が多数存在する。一方で、既存の階層モデルが既にある程度の複雑な構造を保持しているため、前述のような問題点を解決する方法の提案は容易なものではなく、革新的かつ緻密なアイデアが必要とされる。

2. 研究の目的

前述のような階層モデルが持つ問題点を解決することで、グループデータから「正しい知見」を構成するための方法論の発展に貢献することを目的とする。具体的には、これらの問題の数理的な性質を明らかにし、効果的なモデルや有効な推定アルゴリズムの開発を行う。さらに、グループデータ解析に関わる多くの研究者・実務家が、本研究課題で開発した手法を容易に利用できるよう実装したコードをオープンソースとして公開する。

3. 研究の方法

本研究は、グループデータに対する統計分析手法の開発が研究対象であり、新たな方法論の開発と数理的な論証、新たに開発した方法論の実装と発信(関連する統計家が開発した手法を利用できるよう、開発した手法を実装したプログラムを作成し、公開する)、シミュレーション実験による既存手法との比較、実際のデータを用いた実証的な評価などの過程を経て、課題解決のための方法論の開発・評価を行う。

4. 研究成果

(1) 潜在有限混合モデルを用いた条件付き分布のモデリング

グループ毎に異なる条件付き分布(回帰モデルを一般化した概念)を推定することにより、グループデータに対する様々な情報を引き出すことができる。そのために、グループ毎の条件付き分布が全グループに共通の条件付き分布の有限混合で表現される新しいモデルの枠組みを提案した。各グループの特徴を決める混合割合に対してはディリクレ分布を用いた確率的なモデルと、クラスタリングと組み合わせた手法を提案した。さらに、それらのモデルを推定するための効率的な最適化アルゴリズムを与え、理論的な性質を明らかにした。

(2) 外れ値に対して頑健な線形混合効果モデルの推定手法の開発

階層モデルに含まれる代表的なモデルとして線形混合効果モデルがある。このモデルは誤差と変量効果に正規分布を仮定するのが標準的であり、データに外れ値が含まれている場合、正確にパラメータを推定することが困難になる。このような問題に対して、robust divergence の考えを導入して頑健な推定手法を提案し、最適化のための効率的なアルゴリズムを開発した。さらに、このような robust divergence を使った変量効果推定の考えを小地域推定の問題に適用し、他の地域と比べて特異な地域が存在する場合でも安定的に推計値を計算することが可能な手法を開発した。

(3) ホットスポットを考慮したカウントデータの階層モデリング

地域ごとに得られるローカルな情報から局所的な母数(犯罪リスクなど)を評価する場合、地域別母数に確率的な構造(事前分布)を仮定し、ベイズ法によって推定するのが一般的である。しか

し、ホットスポットのような他の地域と比べて特異的な地域の母数に対しては、典型的に用いられる事前分布の仮定が妥当でなく、誤った解析評価を与えてしまう可能性が知られている。そこでカウントデータに対する global-local shrinkage prior と呼ばれる事前分布のクラスを導入し、前述の問題を解決する方法論を提案した。そのために好ましい理論的な性質を定義し、それを満たすために事前分布が持つべき十分条件を導出し、その十分条件をもとに構成的に事前分布を開発した。また推定のための効率的なマルコフ連鎖モンテカルロ法のアルゴリズムも導出した。

(4) 空間クラスタリングを取り入れた空間回帰分析手法の開発

空間データに対して空間的な異質性を考慮した代表的な回帰分析手法として地理的加重回帰がある。この方法は数値的な不安定性や大規模データにおける計算コストの問題点が知られている。このような課題を解決するために、空間クラスタリングと回帰分析を同時に実行する革新的な方法論を開発した。提案モデルはクラスタリングと得られたクラスター毎の回帰分析を交互に実行することで推定することが可能であり、大規模データにおいても高速にかつ安定的に推定できることを数値的に示した。

(5) 外れ値に対して頑健なベイズ推測を実現するための新しい誤差分布の開発

データに外れ値が含まれている状況でベイズ分析を行う場合、 t 分布のような裾の厚い分布を誤差分布として用いることで頑健にパラメータ推定を実行するアプローチが標準的である。一方で、ベイズ分析における頑健性を担保するには、 t 分布よりも裾が厚い分布が必要であることが知られているが、現状提案されている誤差分布は事後分布を計算するのが容易ではない。そこで、簡易なマルコフ連鎖モンテカルロ法で事後分布の計算が実行できる誤差分布のクラスを新たに開発した。この分布により階層モデルのような複雑なモデルに対しても現実的な計算コストで頑健なベイズ分析が実行できるようになった。

(6) その他の成果

関連する成果として、外れ値が存在するもとの有限混合モデルの頑健な推定手法、robust divergence を用いた頑健なベイズ分析手法、空間的に異質な所得分布の推定手法の開発などを実施した。

5. 主な発表論文等

〔雑誌論文〕 計20件（うち査読付論文 20件 / うち国際共著 4件 / うちオープンアクセス 3件）

1. 著者名 Hashimoto Shintaro, Sugasawa Shonosuke	4. 巻 22
2. 論文標題 Robust Bayesian Regression with Synthetic Posterior Distributions	5. 発行年 2020年
3. 雑誌名 Entropy	6. 最初と最後の頁 661 ~ 661
掲載論文のDOI (デジタルオブジェクト識別子) 10.3390/e22060661	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 -
1. 著者名 Sugasawa Shonosuke, Noma Hisashi	4. 巻 77
2. 論文標題 Efficient screening of predictive biomarkers for individual treatment selection	5. 発行年 2020年
3. 雑誌名 Biometrics	6. 最初と最後の頁 249 ~ 257
掲載論文のDOI (デジタルオブジェクト識別子) 10.1111/biom.13279	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 Kobayashi Genya, Sugasawa Shonosuke, Tamae Hiromasa, Ozu Takayuki	4. 巻 14
2. 論文標題 Predicting intervention effect for COVID-19 in Japan: state space modeling approach	5. 発行年 2020年
3. 雑誌名 BioScience Trends	6. 最初と最後の頁 174 ~ 181
掲載論文のDOI (デジタルオブジェクト識別子) 10.5582/bst.2020.03133	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 -
1. 著者名 Ito Tsubasa, Sugasawa Shonosuke	4. 巻 153
2. 論文標題 Improved confidence regions in meta-analysis of diagnostic test accuracy	5. 発行年 2021年
3. 雑誌名 Computational Statistics & Data Analysis	6. 最初と最後の頁 107068 ~ 107068
掲載論文のDOI (デジタルオブジェクト識別子) 10.1016/j.csda.2020.107068	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 Sugasawa Shonosuke	4. 巻 NA
2. 論文標題 Grouped Heterogeneous Mixture Modeling for Clustered Data	5. 発行年 2020年
3. 雑誌名 Journal of the American Statistical Association	6. 最初と最後の頁 1~12
掲載論文のDOI (デジタルオブジェクト識別子) 10.1080/01621459.2020.1777136	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 Sugasawa Shonosuke, Kim Jae Kwang	4. 巻 NA
2. 論文標題 An Approximate Bayesian Approach to Model-assisted Survey Estimation with Many Auxiliary Variables	5. 発行年 2020年
3. 雑誌名 Statistica Sinica	6. 最初と最後の頁 NA
掲載論文のDOI (デジタルオブジェクト識別子) 10.5705/ss.202019.0239	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 Saegusa Takumi, Sugasawa Shonosuke, Lahiri Partha	4. 巻 NA
2. 論文標題 Parametric Bootstrap Confidence Intervals for the Multivariate Fay?Herriot Model	5. 発行年 2020年
3. 雑誌名 Journal of Survey Statistics and Methodology	6. 最初と最後の頁 NA
掲載論文のDOI (デジタルオブジェクト識別子) 10.1093/jssam/smaa038	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 Kobayashi Genya, Yamauchi Yuta, Kakamu Kazuhiko, Kawakubo Yuki, Sugasawa Shonosuke	4. 巻 NA
2. 論文標題 Bayesian Approach to Lorenz Curve Using Time Series Grouped Data	5. 発行年 2021年
3. 雑誌名 Journal of Business & Economic Statistics	6. 最初と最後の頁 1~16
掲載論文のDOI (デジタルオブジェクト識別子) 10.1080/07350015.2021.1883438	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 Hamura Yasuyuki, Irie Kaoru, Sugasawa Shonosuke	4. 巻 NA
2. 論文標題 On Global-Local Shrinkage Priors for Count Data	5. 発行年 2021年
3. 雑誌名 Bayesian Analysis	6. 最初と最後の頁 NA
掲載論文のDOI (デジタルオブジェクト識別子) 10.1214/21-BA1263	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 -

1. 著者名 Sugasawa Shonosuke, Noma Hisashi	4. 巻 38
2. 論文標題 Estimating individual treatment effects by gradient boosting trees	5. 発行年 2019年
3. 雑誌名 Statistics in Medicine	6. 最初と最後の頁 5146 ~ 5159
掲載論文のDOI (デジタルオブジェクト識別子) 10.1002/sim.8357	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 Sugasawa S	4. 巻 107
2. 論文標題 Robust empirical Bayes small area estimation with density power divergence	5. 発行年 2020年
3. 雑誌名 Biometrika	6. 最初と最後の頁 467 ~ 480
掲載論文のDOI (デジタルオブジェクト識別子) 10.1093/biomet/asz075	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 Sugasawa Shonosuke, Kobayashi Genya, Kawakubo Yuki	4. 巻 145
2. 論文標題 Estimation and inference for area-wise spatial income distributions from grouped data	5. 発行年 2020年
3. 雑誌名 Computational Statistics & Data Analysis	6. 最初と最後の頁 106904 ~ 106904
掲載論文のDOI (デジタルオブジェクト識別子) 10.1016/j.csda.2019.106904	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 Sugasawa Shonosuke, Noma Hisashi	4. 巻 22
2. 論文標題 A unified method for improved inference in random effects meta-analysis	5. 発行年 2021年
3. 雑誌名 Biostatistics	6. 最初と最後の頁 114 ~ 130
掲載論文のDOI (デジタルオブジェクト識別子) 10.1093/biostatistics/kxz020	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 Sugasawa Shonosuke	4. 巻 3
2. 論文標題 Small area estimation of general parameters: Bayesian transformed spatial prediction approach	5. 発行年 2019年
3. 雑誌名 Japanese Journal of Statistics and Data Science	6. 最初と最後の頁 167 ~ 181
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/s42081-019-00067-7	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 Sugasawa Shonosuke, Kubokawa Tatsuya	4. 巻 3
2. 論文標題 Small area estimation with mixed models: a review	5. 発行年 2020年
3. 雑誌名 Japanese Journal of Statistics and Data Science	6. 最初と最後の頁 693 ~ 720
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/s42081-020-00076-x	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 Sugasawa Shonosuke, Kawakubo Yuki, Ogasawara Kota	4. 巻 90
2. 論文標題 Small area estimation with spatially varying natural exponential families	5. 発行年 2020年
3. 雑誌名 Journal of Statistical Computation and Simulation	6. 最初と最後の頁 1039 ~ 1056
掲載論文のDOI (デジタルオブジェクト識別子) 10.1080/00949655.2020.1714048	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 Sugasawa Shonosuke, Kawakubo Yuki, Datta Gauri Sankar	4. 巻 173
2. 論文標題 Observed best selective prediction in small area estimation	5. 発行年 2019年
3. 雑誌名 Journal of Multivariate Analysis	6. 最初と最後の頁 383 ~ 392
掲載論文のDOI (デジタルオブジェクト識別子) 10.1016/j.jmva.2019.04.002	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 Sugasawa Shonosuke, Kobayashi Genya, Kawakubo Yuki	4. 巻 29
2. 論文標題 Latent mixture modeling for clustered data	5. 発行年 2018年
3. 雑誌名 Statistics and Computing	6. 最初と最後の頁 537 ~ 548
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/s11222-018-9821-7	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 Sugasawa Shonosuke, Kubokawa Tatsuya, Rao J. N. K.	4. 巻 46
2. 論文標題 Hierarchical Bayes small area estimation with an unknown link function	5. 発行年 2018年
3. 雑誌名 Scandinavian Journal of Statistics	6. 最初と最後の頁 885 ~ 897
掲載論文のDOI (デジタルオブジェクト識別子) 10.1111/sjos.12376	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 Sugasawa Shonosuke, Kubokawa Tatsuya	4. 巻 46
2. 論文標題 Adaptively transformed mixed model prediction of general finite population parameters	5. 発行年 2019年
3. 雑誌名 Scandinavian Journal of Statistics	6. 最初と最後の頁 1025 ~ 1046
掲載論文のDOI (デジタルオブジェクト識別子) 10.1111/sjos.12380	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 -

〔学会発表〕 計18件（うち招待講演 1件 / うち国際学会 9件）

1. 発表者名 菅澤翔之助
2. 発表標題 Spatially Clustered Regression
3. 学会等名 科研費シンポジウム "Recent Progress in Spatial and/or Spatio-temporal Data Analysis"
4. 発表年 2020年

1. 発表者名 菅澤翔之助
2. 発表標題 境界値過剰データに対する時空間モデリング_~ 海洋資源調査データへの適用 ~
3. 学会等名 CSIS DAYS 2020
4. 発表年 2020年

1. 発表者名 菅澤翔之助
2. 発表標題 Spatio-temporal modeling of zero-inflated count data: an application to marine resource monitoring
3. 学会等名 Webinar of Bayesian Econometrics 2020 (国際学会)
4. 発表年 2020年

1. 発表者名 菅澤翔之助
2. 発表標題 異質な集団が混在するデータに対する統計モデリング
3. 学会等名 第5回統計・機械学習若手シンポジウム (招待講演)
4. 発表年 2020年

1. 発表者名 Sugasawa, S.
2. 発表標題 On Global-local Shrinkage Priors for Count Data
3. 学会等名 Young Bayesian Workshop (国際学会)
4. 発表年 2019年

1. 発表者名 Sugasawa, S.
2. 発表標題 Robust Bayesian regression with shrinkage priors
3. 学会等名 The 4th Eastern Asia Meeting on Bayesian Statistics (国際学会)
4. 発表年 2019年

1. 発表者名 Sugasawa, S.
2. 発表標題 Robust Estimation and Joint Selection in Linear Mixed Models
3. 学会等名 62nd ISI World Statistics Congress 2019 (国際学会)
4. 発表年 2019年

1. 発表者名 菅澤翔之助
2. 発表標題 On Global-local shrinkage priors for count data
3. 学会等名 統計関連学会連合大会
4. 発表年 2019年

1. 発表者名 Sugasawa, S.
2. 発表標題 Estimation of spatially varying income distributions from grouped data
3. 学会等名 Workshop on Bayesian Modelling of Income and Wealth (国際学会)
4. 発表年 2019年

1. 発表者名 菅澤翔之助
2. 発表標題 無視できない欠測データに対するセミパラメトリックなベイズ推定
3. 学会等名 ベイズ計量経済学研究集会
4. 発表年 2019年

1. 発表者名 菅澤翔之助
2. 発表標題 ロバストな統計手法による 地域効果の推定法
3. 学会等名 研究会「公的データの利用とプライバシー保護の理論」
4. 発表年 2019年

1. 発表者名 Sugasawa, S.
2. 発表標題 Optimal discovery procedure for screening covariates associated with individual treatment effects
3. 学会等名 Workshop on the Frontiers in Data Science (国際学会)
4. 発表年 2020年

1. 発表者名 菅澤翔之助
2. 発表標題 Robust Bayesian Regression with Synthetic Posterior
3. 学会等名 科研費シンポジウム「多様な高次元モデルにおける理論と方法論」
4. 発表年 2020年

1. 発表者名 Sugasawa, S.
2. 発表標題 Estimation of area-wise spatial income distributions from grouped data
3. 学会等名 Computational Financial Econometrics (CFE) 2019 (国際学会)
4. 発表年 2019年

1. 発表者名 Shonosuke Sugasawa
2. 発表標題 Hierarchical Bayes Small Area Estimation with Unknown Link Function
3. 学会等名 Small Area Estimation 2018 (国際学会)
4. 発表年 2018年

1. 発表者名 Shonosuke Sugasawa
2. 発表標題 Small Area Estimation with Unknown Link Function
3. 学会等名 International Society of Bayesian Analysis 2018 World Meeting (国際学会)
4. 発表年 2018年

1. 発表者名 菅澤翔之助
2. 発表標題 clustered dataに対するグループ化有限混合モデル
3. 学会等名 統計関連学会連合大会
4. 発表年 2018年

1. 発表者名 菅澤翔之助
2. 発表標題 空間的確率構造を用いた地域別所得分布の同時推定
3. 学会等名 春季統計学会
4. 発表年 2019年

〔図書〕 計0件

〔産業財産権〕

〔その他〕

-

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

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
オーストラリア	オーストラリア国立大学			
米国	メリーランド大学	アイオワ州立大学	ジョージア大学	
カナダ	カールトン大学			