

平成 31 年 4 月 18 日現在

機関番号：35301

研究種目：基盤研究(B) (一般)

研究期間：2016～2018

課題番号：16H02791

研究課題名(和文) 公的大規模データの利用におけるプライバシー保護の理論と応用

研究課題名(英文) Theory of privacy protection and its application for public big data

研究代表者

佐井 至道 (Shido, Sai)

岡山商科大学・経済学部・教授

研究者番号：30186910

交付決定額(研究期間全体)：(直接経費) 12,700,000 円

研究成果の概要(和文)： 個票データに対するリスク評価に用いられるピットマンモデル、ユーエンスモデルなどについての重要な性質が解明された。疑似マイクロデータ作成のためのノイズの挿入やスワッピングなどについて新たな手法が提案されるとともに、ノイズが挿入されたデータについて母集団も考慮に入れたリスク評価方法が新たに提案された。また、PPDM、空間統計などの分野に、これまで蓄積された理論や手法が応用された。

研究成果の学術的意義や社会的意義

リスク評価に用いられる確率分割モデルの性質などについては、本研究グループでは長期にわたる研究の蓄積があり、本研究期間に絞ってみても、その量、質とも他国の研究を凌いでいる。また、キー変数にノイズを挿入した疑似マイクロデータのリスク評価において、データが標本調査で得られた場合でも、これまでは母集団が考慮されておらず、提案されたリスク評価方法は全く新しい手法と考えられる。

これらの研究成果は、データを安全に広く利用する上で大いに役立つものとする。

研究成果の概要(英文)： The important properties of the models those are used for the risk assessment for the microdata, for example, Pitman model, Ewens model, are found out. New techniques of the noise adding and data swapping to create the synthetic data are proposed, and new risk assessment methods for these data are also proposed. These theories and techniques are adopted to the other fields, for example, PPDM, the spatial statistics.

研究分野：数理統計学

キーワード：官庁統計 個票データ ミクロデータ リスク評価 秘匿措置 ビッグデータ

1．研究開始当初の背景

平成8年度に科学研究費補助金の重点領域研究として、官庁統計の個票データとしての公開に向けた研究が開始されて以来、本研究参加者が代表者を務めた研究を含め、様々な研究が行われてきた。研究開始当初までの本研究グループの研究において、個票データのリスク評価方法、秘匿方法の研究で大きな進捗があり、リスク評価の理論構築においては欧米諸国の研究を質、量とも凌いでいた。特に確率分割モデルについての基礎研究はかなり成熟しており、それを利用できる比較的単純な構造の個票データについては十分な精度でのリスク評価が可能となっていた。

企業や各種団体でもビッグデータが収集されており、その公開や利用も始まっていたが、リスク評価としては推測開示に関する議論にとどまるなど十分な議論が行われていないため、公開に否定的な意見が増えてきていた。また、他のデータセットとの併用による開示リスクの増加についても問題視されていた。

2．研究の目的

本研究では次に挙げる(1)、(2)を目的に掲げた。

(1) データに対する秘匿方法、リスク評価方法の理論の確立と充実

これまで培ってきた理論を基礎として、官庁統計データに代表される個票データの秘匿方法、リスク評価方法、さらには有用性の評価方法について研究を進め、個々の理論の完成と、それらの研究の統合、そしてその応用を目的とする。

(1-1) データの秘匿方法と有用性の研究

秘匿方法については国内における研究の上積みの少ない状況が続いていたが、ミクロアグリゲーションの研究などでここ数年新たな成果が得られており、従来型の方法とともに、秘匿方法の効果について理論的な成果を得ることを目的とする。

海外では攪乱やスワッピングなどの嘘をつく秘匿方法が利用されているが、国内では否定的な意見が多かったこともあり十分な議論が行われてこなかった。平成23年度から全国消費実態調査の疑似マイクロデータが国内では初めて試行的提供されているが、攪乱などを高次元集計表作成後でなく、元データに直接加える方法をとることによって、リスクを押さえながら統計量が元データにより近い疑似データを作成する可能性がある。既に攪乱やスワッピングについて基礎研究が始まっており、さらに研究を進めて、新たな疑似マイクロデータ作成のための理論構築を行うことを目的とする。

また国内でこれまで分析手法ごとに論じられてきた秘匿後のデータの有用性について、より一般的な指標の提案を目指し、最終的にはリスク指標との関係についての理論的な成果を得ることも目的とする。

(1-2) データのリスク評価方法の研究

個票データのリスク評価は、これまでの本研究参加者の貢献が最も大きい領域である。ここ数年の間にも、層化抽出されたデータや短期的な時系列データについてのリスク評価方法が提案され、推定に用いられるビットマンモデルなどの確率分割モデルの基礎研究でも大きな発展があった。本研究では、長期時系列データ、階層的な構造を持つデータ、多次元の表形式データなどについてのリスク評価方法についての理論の確立も目的とする。

また国内ではこれまで用いられなかったため、攪乱などの秘匿を施したデータについての安全性については十分な検討が行われてこなかった。本研究ではその解決も目的とする。

(2) 種々の分野のビッグデータの公開に関する研究

(2-1) 種々の分野のデータに対する秘匿とリスク評価の研究

秘匿方法とリスク評価方法の研究成果は蓄積されてきているが、その適用対象の多くは官庁統計であった。しかし他分野でも本研究に関連する問題が多数存在しており、現在は本研究参加者が個別に対応し研究を行っている。本研究参加者は、医学、空間統計、生物統計、PPDM（プライバシー保護データマイニング）で扱うデータやゲノムデータなどについても研究を行っているが、既存の理論が適用できない場合も多く、研究期間内に問題の類別を行い、それぞれの問題について解決を図ることを目的とする。

(2-2) 種々の形式のデータに対するリスク評価方法の確立

今後公開される予定のオープンデータの中には、細かい地理情報が多数を占めるデータやテキストデータなど、多様な形式のデータが含まれることが予想される。テキストデータではキー変数の定義や、ファイルからの抽出などが必要になり、これらと既存の理論との接続を図ることも目的とする。

3．研究の方法

本研究は下記の(1)と(2)における複数の研究を統合する形で行われた。それぞれの研究について少人数の研究者からなるグループを配置し、他の研究者がそれをサポートした。

(1) データに対する秘匿方法，リスク評価方法の理論の確立と充実

(1-1) データの秘匿方法と有用性の研究については，佐井，星野，伊藤，稲葉，瀧，丸山，竹村，小林が中心に取り組む。

個票データの秘匿方法としてはグローバルリコーディングなどが用いられてきたが，数年前から，伊藤，小林が中心となりマイクロアグリゲーションについての研究が進められており，その理論の成熟を図る。これまで国内では否定的な意見も多かった攪乱（誤差を加える方法）については，伊藤，小林によって疑似マイクロデータの作成に活かされるなど抵抗感が薄れてきており，現在も総務省統計研修所のプロジェクトメンバーとして稲葉が作成に取り組んでいる。佐井によって理論的な研究も始まっており，その理論の構築を進める。

秘匿措置の前後におけるデータの有用性とリスクの変化については，伊藤によって国内では初めて本格的に取り組まれており，秘匿方法，リスク評価方法，有用性の評価方法の関係について，より広い方法や指標について議論する。

(1-2) データのリスク評価方法の研究については，佐井，渋谷，星野，間野，大和，伊藤，竹村，小林が中心に取り組む。

寸法指標の推定によるリスク評価方法の研究では本研究参加者による 19 年余りの研究成果が蓄積されている。特に理論が成熟しつつある確率分割モデルによる推定について理論の完成を目指す。

攪乱されたデータでは識別は起きないが，事実上識別が可能なケースはありうる。その評価については佐井によって検討が始められており，基本的な設定の問題から順次取り組んでいく。

(2) 種々の分野のビッグデータの公開に関する研究

(2-1) 種々の分野のデータに対する秘匿とリスク評価の研究

本研究参加者は，官庁統計以外の個票データに関する研究にも着手している。その中には(1)の研究成果が応用できるものもあるが，データの構造や性質に違いがあるため，新たな検討も必要となる。平成 28 年度はその洗い出しを行い，29 年度以降はその解決を図る。

佐井は平成 23 年度から院内がん登録データの公開の研究に携わっている。この個票データには，各レコードに膨大な時系列データであるレセプトデータが付随している。この問題については，佐井，星野，竹村，田村が研究を行う。

丸山は地理情報などを扱う空間統計を専門としている。地理情報は，キー変数の中でも個体の特定に最も利用されやすい変数となる。一方で，秘匿された情報から特定の地点，あるいは地域全体の地価を推定したいとするニーズがあるため，秘匿と利用価値の両立を図らなければならない。この研究には丸山，稲葉，星野が携わる。

PPDM（プライバシー保護データマイニング）では，例えば，2 つのオンラインサービスが所持している個人情報を，互いに他のオンラインサービスに開示することなしに，その両方を用いなければ得られない値を計算することが要望されている。この研究には PPDM を専門としている佐久間とともに，星野，佐井，瀧が携わる。

間野はゲノムデータに精通している。ゲノムデータは個々の要素よりもサイズの多様性に興味があるのが特徴であり，高度にセンシティブなデータでもあるため，通常のデータと異なる扱いが必要になる。これらの研究は間野，渋谷，大和，佐井が行う。

(2-2) 種々の形式のデータに対するリスク評価方法の確立

今後公開される予定のオープンデータの中には，通常のデータ形式のものだけでなく，細かい地理情報が多数を占めるデータ，時系列情報を持つデータ，テキストデータなど，様々な形式のデータが含まれる。地理情報や短期的な時系列情報を持つデータについては，これまで培ってきた理論でどの程度の秘匿やリスク評価が可能かを見極める。この研究は，当初はテキストマイニングに精通している佐久間とともに，佐井，間野，丸山，竹村が担当することとする。

本研究では参加者が研究を分担して行うが，これらの研究は相互に深く関連している。そのため全員の共通認識の構築と相互の意見交換のために，毎年度 3 回程度の研究会を開催する。また，幅広い意見や知識の収集のために，官庁統計に携わる研究者や実務者にも参加を呼びかけて研究集会を 2 日間の日程で開催する。研究成果については，統計関連学会連合大会などの国内外の学会において公表するとともに，進捗状況については Web ページでも随時公開する。

4．研究成果

本研究の主目的は，データに対する秘匿方法・リスク評価方法の理論を確立し充実させることであり，データとしては個票データ，表形式データ，疑似マイクロデータが対象であった。

このうち個票データに対するリスク評価に用いられるピットマンモデル，ユーエンスモデル，ディリクレ過程など，確率分割やその関連領域における研究では，渋谷，間野，佃，大和などによってモデルの性質などについて多くの成果が得られた。（〔雑誌論文〕7, 8, 9, 10, 12, 13, 14, 19, 20）また，汎用的なリスク評価についての提案も星野によってなされた。（〔雑誌論文〕18）さらに，伊藤らによって，個票データの有用性とリスクを同時に考慮した評価手法について引き続き検討が進められている。（〔雑誌論文〕4, 5, 21）

疑似マイクロデータについては、統計センターにおいて本研究参加者の業績を取り入れて一般用マイクロデータとして提供が行われており、現在も改良が進められている。〔雑誌論文〕16) 伊藤はデータスワッピングやマイクロアグリゲーションを用いた秘匿について、官庁統計の実個票データを用いた詳細な検討を行った。〔雑誌論文〕3, 6, 15, 22, 23) 星野は疑似多項分布を用いた模造データの作成法を提案し、他の手法と比較して優れていることを示した。〔雑誌論文〕2) 佐井はキー変数とノイズ変数に分布を導入して、母集団も考慮に入れたリスク評価方法を考案した。両変数について一様分布や正規分布などの理論的分布を導入した方法についてはほぼ完成したものの、応用面での当てはまりの悪いデータが存在することから、局所的なキー変数の分布を利用したノンパラメトリック推定法を提案し、前述の方法よりも優れたケースのあることを示した。〔雑誌論文〕1, 11, 17) これらの研究は本研究でも最も進捗が大きいものとなったが、まだ残されている問題も多く今後の課題である。

本研究のもう一つの目的は、種々の分野のビッグデータの公開に関する研究を行うことであった。情報処理学会内の PWS (プライバシーワークショップ) とは初年度からそれぞれの研究集会へ相互に参加し合い情報共有と意見交換を行ってきた。2017 年度と 2018 年度の統計関連学会連合大会、PWS2017 (CSS2017) と PWS2018 (CSS2018) においては、それぞれ合同の企画セッションを設置し、互いに招待講演を行うなどして、それぞれの学会内での知識の習得にも務めた。

本研究の成果については、統計関連学会連合大会などの学会や国内外の各種シンポジウム、研究集会において報告を行うとともに、統計数理研究所で主催した研究集会や研究会でも報告し、討論や意見交換を行った。以下に、本研究が主催 (あるいは共催) した研究集会と研究会の概要を示す。

- | | | |
|-------------------------------------|----------------------------|----------------------|
| 研究集会「公的大規模データの利用におけるプライバシー保護の理論と応用」 | 日時: 2016 年 12 月 8 日, 9 日 | 場所: 統計数理研究所・セミナー室 1 |
| | 参加者: 41 名 | 報告者: 18 名 |
| 研究集会「公的大規模データの利用におけるプライバシー保護の理論と応用」 | 日時: 2017 年 12 月 14 日, 15 日 | 場所: 統計数理研究所・セミナー室 5 |
| | 参加者数: 35 名 | 報告者数: 18 名 |
| 研究集会「公的大規模データの利用におけるプライバシー保護の理論と応用」 | 日時: 2018 年 12 月 13 日, 14 日 | 場所: 統計数理研究所・セミナー室 2 |
| | 参加者: 25 名 | 報告者: 14 名 |
| 研究会「公的大規模データの利用におけるプライバシー保護の理論と応用」 | 日時: 2016 年 7 月 23 日 | 場所: 統計数理研究所サテライトオフィス |
| | 参加者: 7 名 | 報告者: 4 名 |
| 研究会「プライバシー保護の統計数理」 | 日時: 2017 年 2 月 13 日 | 場所: 岡山大学東京オフィス |
| | 参加者: 6 名 | 報告者: 3 名 |
| 研究会「公的大規模データの利用におけるプライバシー保護の理論と応用」 | 日時: 2017 年 7 月 20 日 | 場所: 岡山大学東京オフィス |
| | 参加者数: 10 名 | 報告者数: 5 名 |
| Kanazawa Workshop on SDC | 日時: 2018 年 1 月 21 日 | 場所: 金沢大学サテライトプラザ |
| | 参加者数: 10 名 | 報告者数: 5 名 |
| 研究会「公的大規模データの利用におけるプライバシー保護の理論と応用」 | 日時: 2018 年 7 月 26 日 | 場所: 岡山大学東京オフィス |
| | 参加者: 6 名 | 報告者: 5 名 |

5. 主な発表論文等

〔雑誌論文〕(計 43 件)

1. 佐井至道, 個票データのキー変数の型とリスクとの関係, 岡山商大論叢, 査読有, 54, 2018, 1-28.
2. Hu, J. and Hoshino, N., The quasi-multinomial synthesizer for categorical data, Privacy in Statistical Databases, PSD 2018, Lecture Notes in Computer Science, Domingo-Ferrer, J., Montes, F. (eds), 査読有, 11126, 2018, 75-91.
DOI: https://doi.org/10.1007/978-3-319-99771-1_6
3. 伊藤伸介, 国勢調査における匿名化マイクロデータの作成可能性, 経済志林, 査読無, 85, 2018, 241-277.
4. Ito, S., Yoshitake, T., Kikuchi, R. and Akutsu, F., Comparative study of the effectiveness of perturbative methods for creating official microdata in Japan, Privacy in Statistical Databases, PSD 2018, Lecture Notes in Computer Science, Domingo-Ferrer, J., Montes, F. (eds), 査読有, 11126, 2018, 200-214.
DOI: https://doi.org/10.1007/978-3-319-99771-1_14

5. 伊藤伸介, 星野なおみ, 阿久津文香, 菊池亮, 匿名化された公的統計マイクロデータの作成における攪乱の手法の有効性の評価, 経済学論纂(中央大学), 査読無, 59, 2018, 37-60.
6. 伊藤伸介, 星野なおみ, 阿久津文香, 菊池亮, 国勢調査の匿名化マイクロデータの作成方法に関する新たな取り組み, 製表技術参考資料, 査読無, No.37, 2018, 1-27.
7. Tsukuda, K. and Matsuura, S., High-dimensional testing for proportional covariance matrices, Journal of Multivariate Analysis, 171, 2019, 412-420.
DOI: <https://doi.org/10.1016/j.jmva.2019.01.011>
8. Tsukuda, K., On Poisson approximations for the Ewens sampling formula when the mutation parameter grows with the sample size, The Annals of Applied Probability, 査読有, 29, 2019. 1188-1232.
DOI: <https://doi.org/10.1214/18-AAP1433>
<https://projecteuclid.org/euclid.aoap/1548298939>
9. Tsukuda, K., Functional central limit theorems in $L^2(0,1)$ for logarithmic combinatorial assemblies, Bernoulli, 査読有, 24, 2018, 1033-1052.
DOI: <https://doi.org/10.3150/16-BEJ847>
10. Tsukuda, K. and Mano, S., A reversal phenomenon in estimation based on multiple samples from the Poisson-Dirichlet distribution, arXiv, 査読無, 1802.00578, 2018.
11. 佐井至道, 曖昧な母集団情報を考慮に入れたノイズを含む個票データのリスク評価, 岡山商大論叢, 査読無, 53, 1, 2017, 27-57.
12. Segers, J., Sibuya, M. and Tsukahara, H., The empirical beta copula, Journal of Multivariate Analysis, 査読有, 155, 1, 2017, 35-51.
DOI: <https://doi.org/10.1016/j.jmva.2016.11.010>
13. Mano, S., Partition structure and the A-hypergeometric distribution associated with the rational normal curve, Electronic Journal of Statistics, 査読有, 11, 2017, 4452-4487.
DOI: <https://doi.org/10.1214/17-EJS1361>
14. 間野修平, 分割の可乗測度と可換代数, 確率論シンポジウム, 査読無, 京都大学数理解析研究所講義録, 2030, 2017, 105-107.
<http://www.kurims.kyoto-u.ac.jp/~kyodo/kokyuroku/contents/2030.html>
15. 伊藤伸介, 国勢調査マイクロデータにおける匿名化の誤差の評価方法に関する一考察, 経済学論纂(中央大学), 査読無, 57, 2017, 189-209
16. 稲葉由之, 攪乱の方法を用いて作成する匿名データに関する基礎研究, 総務省統計研修所リサーチペーパー, 査読無, 39号, 2017, 1-17.
17. 佐井至道, ノイズが挿入された個票データのリスク評価, 岡山商大論叢, 査読無, 52, 1, 2016, 23-50.
18. 星野伸明, エビデンスに基づいた匿名化, 日本統計学会誌, 査読有, 46, 1, 2016, 1-42.
<http://www.terrapub.co.jp/journals/jjssj>
19. Mano, S., Extreme sizes in the Gibbs-type exchangeable random partitions, Annals of the Institute of Statistical Mathematics, 査読有, 69, 2016, 1-37.
DOI: <https://doi.org/10.1007/s10463-015-0530-0>
20. Griffiths, R. C., and Mano, S., The star-shaped Lambda-coalescent and Fleming-Viot process, Stochastic Models, 査読有, 32, 2016, 606-631.
DOI: <https://doi.org/10.1080/15326349.2016.1188404>
21. 伊藤伸介, 星野なおみ, 阿久津文香, 国勢調査における匿名化マイクロデータの有用性と秘匿性の定量的な評価, 製表技術参考資料, 査読無, No.32, 2016, 1-33.
22. Ito, S., Hoshino, N. and Akutsu, F., Potential of disclosure limitation methods for census microdata in Japan, Paper presented at Privacy in Statistical Databases 2016, 査読有, 2016, 2016, 1-14.
23. 伊藤伸介, 星野なおみ, 阿久津文香, 国勢調査マイクロデータに対する匿名化措置の可能性に関する研究, 製表技術参考資料, 査読無, No.34, 2016, 1-59.

以下省略

〔学会発表〕(計118件)

1. Sibuya, M., Gibbs base random partitions, Pioneering Workshop on Extreme Value and Distribution Theories in Honor of Professor Masaaki Sibuya(特別講演)(国際学会), 2018.
2. 佐井至道, 秘匿措置が施されたデータのリスク評価の試み, 研究集会「公的大規模データの利用におけるプライバシー保護の理論と応用」, 2018.

以下省略

〔図書〕(計1件)

- Mano, S., JSS Research Series in Statistics, SpringerBriefs in Statistics, Springer, Partitions, Hypergeometric Systems, and Dirichlet Processes in Statistics, 2018, 135.
ISBN: 978-4-431-55888-0

〔産業財産権〕

- 出願状況（計 0 件）
- 取得状況（計 0 件）

〔その他〕

ホームページ等

<http://sai.in.coocan.jp/research/index.html>

6. 研究組織

(1)研究分担者

研究分担者氏名：星野 伸明

所属研究機関名：金沢大学

職名：教授

ローマ字氏名：(HOSHINO, nobuaki)

部局名：経済学経営学系

研究者番号（8桁）：00313627

研究分担者氏名：渋谷 政昭

所属研究機関名：慶應義塾大学

職名：名誉教授

ローマ字氏名：(SIBUYA, masaaki)

部局名：理工学部（矢上）

研究者番号（8桁）：20146723

研究分担者氏名：間野 修平

所属研究機関名：統計数理研究所

職名：准教授

ローマ字氏名：(MANO, shuhei)

部局名：数理・推論研究系

研究者番号（8桁）：20372948

研究分担者氏名：伊藤 伸介

所属研究機関名：中央大学

職名：教授

ローマ字氏名：(ITO, shinsuke)

部局名：経済学部

研究者番号（8桁）：90363316

研究分担者氏名：稲葉 由之

所属研究機関名：明星大学

職名：教授

ローマ字氏名：(INABA, yoshiyuki)

部局名：経済学部

研究者番号（8桁）：80312437

研究分担者氏名：佃 康司

所属研究機関名：東京大学

職名：特任講師

ローマ字氏名：(TSUKUDA, koji)

部局名：大学院総合文化研究科

研究者番号（8桁）：30764972

研究分担者氏名：瀧 敦弘

所属研究機関名：広島大学

職名：教授

ローマ字氏名：(TAKI, atsuhiro)

部局名：社会科学研究科

研究者番号（8桁）：40216809

(2)研究協力者

研究協力者氏名：大和 元

研究協力者氏名：丸山 祐造

研究協力者氏名：竹村 彰通

研究協力者氏名：佐久間 淳

研究協力者氏名：田村 義保

研究協力者氏名：小林 良行

ローマ字氏名：(YAMATO, hajime)

ローマ字氏名：(MARUYAMA, yuzo)

ローマ字氏名：(TAKEMURA, akimichi)

ローマ字氏名：(SAKUMA, jun)

ローマ字氏名：(TAMURA, yoshiyasu)

ローマ字氏名：(KOBAYASHI, yoshiyuki)