様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年４月１２日現在

機関番号：３５３０１
研究種目：基盤研究（Ｂ）
研究期間：2010～2012
課題番号：２２３０００９７
研究課題名（和文）　調査データベース公有化における個人データ保護の統計理論

研究課題名（英文）　Statistical theory on data protection for public database

研究代表者
佐井　至道（SAI SHIDO）
岡山商科大学・経済学部・教授
研究者番号：30186910

研究成果の概要（和文）：母集団寸法指標の推定に用いられる超母集団モデルについて，その性質が新たに解明されるとともに，ノンパラメトリック法との融合が図られた。多重寸法指標についても研究が進められ，時系列データ，層化抽出されたデータから得られた個票データ，事後層化された個票データに対するリスク評価に適用され，その有効性も示された。また，PPDM，空間統計，医学統計，生物統計などの分野に，これまで蓄積された理論や手法が応用された。

研究成果の概要（英文）：Some properties of the superpopulation models, which are used for the estimation of the population size indecies, are investigated. The nonparametric estimation method associated with the Pitman model, which is one of the superpopulation models, are proposed. The multi size indecies are also proposed and applied to the risk assessment of the microdata set made from the sequential data, the stratified sample and the post-stratified sample. These theories and techniques are adopted to the other fields, for example, PPDM, the spatial statistics, the medical statistics and biostatistics.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	5,600,000	1,680,000	7,280,000
２０１１年度	4,700,000	1,410,000	6,110,000
２０１２年度	3,400,000	1,020,000	4,420,000
年度			
年度			
総　計	13,700,000	4,110,000	17,810,000

研究分野：総合領域
科研費の分科・細目：情報学・統計科学
キーワード：官庁統計，個票データ，リスク評価，寸法指標，秘匿措置，プライバシー保護，多重寸法指標，疑似個票データ

１．研究開始当初の背景

統計審議会の統計行政に対する新中・長期構想答申を受け，平成８年度から科学研究費補助金で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，本研究参加者によって様々な研究が行われてきた。個票データのリスク評価法，秘匿方法，有用性の研究で進捗があったが，特に個票データを公開する際のリスク評価の理論構築は，欧米諸国の研究を質，量とも凌いでいる。

これらの研究の成果も後押しとなり，四つの官庁統計についての匿名化データの試行的提供を経て，平成 19 年 5 月には統計法が全部改正され，平成 21 年 4 月から全面施行

されている．改正の柱の一つは「統計データの利用促進と秘密の保護」であり，その中には，(a)従来の目的外使用を継承した調査票情報の二次利用に関する規定に加えて，(b)調査票情報について適切な秘匿措置を施し，必要に応じてサブサンプリングを行った匿名データの提供に関する規定，(c)利用者の求めに応じて，調査票情報を基に集計結果を計算するオーダーメード集計に関する規定が含まれていた。

しかし，当時(b)において提供されている統計は四つの官庁統計にとどまっており，(c)においても国勢調査しか扱っていなかった。また，研究者から要望の多かった自由な配布は認められていなかった。

更に，プライバシーに配慮しなければならないデータは官庁統計以外の分野でも数多く存在するが，理論的蓄積はそれらの分野にはほとんど活かされていなかった。

２．研究の目的

本研究では次の四つを目的に掲げた。

(1) 個票データに対する秘匿方法とリスク評価方法の確立

個票データのリスク評価については，これまで基礎的な理論が確立されているが，複雑なサンプリング法で得られた個票データや，時系列的に得られた個票データセットの同時リスク評価についてはこれからの課題であり，研究を進める。個票データの秘匿措置については，これまで国内においては十分に研究されてきたとは言えない。ミクロアグリゲーションなど，ここ数年研究に進展のある方法を含め検討を行う。なお，秘匿措置はリスクを減少させるとともにデータが持っている情報量をも減少させるため，その両方を同時に評価することも必要となる。

上記二つの研究を統合し，官庁統計のサンプリング方法，時系列情報，調査項目の特性などを考慮に入れながら適切な秘匿措置を施すとともにリスク評価を行い，プライバシー漏洩の観点から見て安全で，しかも分析の面から見ても十分な情報量を持つ個票データを作成するための理論構築を第一の目的とする。

(2) 表形式データに対する秘匿方法とリスク評価方法の確立

本研究では，表形式データについても情報量を残しながら秘匿を行う手法の確立と，秘匿措置の自動化を第二の目的に掲げる。

(3) 調査票情報の縮約と個票データへの復元方法の確立

本研究の第三の目的として，調査票情報などの個票データから，個体を特定できる情報を含まず，しかも分析に必要な情報量を残す統計量の組や表形式データなどを作成し，その情報のみを用いて疑似個票データを復元する方法の確立を掲げる。その実現によって，個票データに対する分析方法の開発が容易になるとともに，分析に精通した人材育成の助けになる。

(4) 官庁統計以外の個票データおよび他分野のデータに対する応用

地方自治体，企業，NPO などでも多くの個票データを所有しているものの，適切な公開方法を見いだすことができずにいる場合も多い。また，インターネットなどのオンラインサービスにおける個人情報の扱いでは，従来の秘匿方法やリスク評価の枠では解決できない問題も数多くある。第四の目的として，これらの問題整理とこれまで蓄積された理論と手法を用いた解決策の提示を掲げる。

３．研究の方法

研究目的に掲げた四つの目的について，それぞれ小グループを配置し，各研究を並行して行う。

(1)，(2)の個票データと表形式データの秘匿方法とリスク評価方法に関する研究では，これまでの経験と理論的な蓄積があるため，個々の研究者それぞれで研究を進めるとともに，研究会や研究集会において，それらの成果を統合する。(1)については，佐井，渋谷，大和，稲葉，伊藤，竹村，星野，和合，大森が中心に研究を行い，(2)については，佐井，竹村，稲葉，瀧，大森，田村が中心に研究を行う。

(3)の調査票情報の縮約と疑似個票データの作成は全く新しい研究であるため，打ち合わせを目的とする研究会を開催し，綿密な計画を立てながら，段階的にその実現に近づけていく。この研究は，佐井，渋谷，星野，和合，大森，小林が中心に行う。

(4)の他分野のデータに対する応用では，個々の分野ごとに研究会を開催しながら，全員が参加する研究会や研究集会において成果の共有を図る。この研究は，佐井，竹村，星野，丸山，佐久間が中心に行う。

本研究の成果は，速やかに学会や各種シンポジウムで公表するとともに，インターネットなどでも情報発信を行う。

４．研究成果

(1) 個票データに対する秘匿方法とリスク評価方法の確立については，多くの成果が得られた。母集団寸法指標の推定に用いられる Ewens-Pitman sampling formula などの確率

分割モデルについては，個々の性質やモデル相互の関係が新たに解明され，確率論などの他分野との共同研究も行われた。また，母集団寸法指標の推定において，ノンパラメトリック法のペナルティー関数の代わりにピットマンモデルの確率関数を用いることによって，ノンパラメトリック法とパラメトリック法の融合が図られた。

多重寸法指標についても研究が進められ，ピットマンモデルの多変量への拡張が行われるとともに，ノンパラメトリック法についても多重指標への拡張が行われた。また，時系列データや層化抽出されたデータから得られた個票データのリスク評価，さらにはリスク評価の精度改善を目的として事後層化された個票データに対しても多重寸法指標が用いられ，その有効性も示された。

(2) 表形式データに対する秘匿方法とリスク評価方法の確立では，マルコフ基底を用いたスワッピングなどにおいて研究が引き続き行われた。また諸外国での動向についても紹介が行われ，国内のデータへの適用について検討が行われたが，大きな成果は得られていない。

(3) 調査票情報の縮約と個票データへの復元方法の確立でも前進があった。ミクロアグリゲーションなど本研究での成果を取り入れる形で，国内の官庁統計では初めて，独立行政法人統計センターにおいて疑似個票データの試行的提供が行われた。作成方法としては質的変数について高次元の集計表を作成し，多数の量的変数については対数正規乱数で置き換えるものである。

これとは若干異なる手法で一橋大学においても疑似個票データが作成された。作成途中に集計表を用いず，個票データに直接誤差を追加したりスワッピングを行ったりすることによって，個票データを安全にする方法についても研究が行われている。

(4) 官庁統計以外の個票データおよび他分野のデータに対する応用でも大きな進捗があった。空間統計データでは，重回帰分析などの結果に影響を与えない個票データの秘匿措置について提案がなされた。医学統計では，あるデータベースに重複して登録されているレコード数を推定する問題において，別のデータベースと寸法指標そのものや，寸法指標に当てはめるモデルのパラメータを比較する方法が提案された。またプライバシー保護データマイニング，生物統計，環境統計の問題についても，国内外のエキスパートとともに問題解決を図っているところである。

(5) 本研究の成果については国内外の学会，シンポジウム，研究集会などで随時報告した。特に，統計関連学会連合大会においては，2010 年度に 14 件，2011 年度に 17 件，2012 年度に 15 件の報告を行っている。

また，2010 年 10 月 29 日～30 日，2011 年 10 月 21 日～22 日，2012 年 10 月 26 日～27 日には，それぞれ統計数理研究所において研究集会「官庁統計データの公開における諸問題の研究と他分野への応用」を開催した。毎年度，40 名前後の参加者があり，この分野に関係する研究者や官庁統計の実務者と意見交換を行った。

５．主な発表論文等

〔雑誌論文〕（計８０件）

① Ogawa, M., Hara, H. and Takemura, A., Graver basis for an undirected graph and its application to testing the beta model of random graphs, Annals of the Institute of Statistical Mathematics, 査読有, VOL.65, 2013, 191-212. DOI:10.1007/s10463-012-0367-8

② 佐井至道，事後層化による個票データのリスク評価の改善,岡山商大論叢，査読無, 48 巻，1-25，2012.

③ 瀧敦弘，集計表におけるセル秘匿問題（２），広島大学経済学部 Discussion Paper，査読無，NO.2012-03，2012，1-7.

④ 伊藤伸介，出島敬久，若年の就業状況に与える家計の資産所得の影響,一橋大学経済研究所 Discussion Paper Series A，査読無，NO.571，2012，1-23.

⑤ 秋山裕美，山口幸三，伊藤伸介，星野なおみ,後藤武彦，教育用疑似ミクロデータの開発とその利用～平成 16 年全国消費実態調査を例として～，製表技術参考資料，査読無，NO.16，2012，1-43.

⑥ Ogawa, M. and Takemura, A., Markov bases for typical block effect models of two-way contingency tables, Journal of Multivariate Analysis, 査読有, VOL.112, 2012, 219-229. DOI:10.1016/j.jmva.2012.06.007

⑦ Takemura, A. and Hara, H., Markov chain Monte Carlo test of toric homogeneous Markov chains, Statistical Methodology, 査読有, VOL.9, 2012, 392-406. DOI:10.1016/j.stamet.2011.10.004

⑧ Nakajima, J., Kunihama, T., Omori, Y. and Fruhwirth-Schnatter, S., Generalized extreme value distribution with time-dependence using the AR and MA models in state space form, Computational Statistics and Data Analysis, 査読有, VOL.56, 2012,

3241-3259.
DOI:10.1016/j.csda.2011.04.017

⑨ Ishihara, T. and Omori, Y., Efficient Bayesian estimation of a multivariate stochastic volatility model with cross leverage and heavy-tailed errors, Computational Statistics and Data Analysis, 査読有, VOL.56, 2012, 3674-3689.
DOI:10.1016/j.csda.2010.07.015

⑩ Nakajima, J. and Omori, Y., Stochastic volatility model with leverage and asymmetrically heavy-tailed error using GH skew Student's t-distribution, Computational Statistics and Data Analysis, 査読有, VOL.56, 2012, 3690-3704.
DOI:10.1016/j.csda.2010.07.012

⑪ Hoshino, N., Random partitioning over a sparse contingency table, Annals of the Institute of Statistical Mathematics, 査読有, VOL.64, 2012, 457-474.
DOI:10.1007/s10463-011-0327-8

⑫ Kakamu, K., Polasek, W. and Wago, H., Production technology and agglomeration for Japanese prefectures during 1991-2000, Papers in Regional Science, 査読有, VOL.91, 2012, 29-41.
DOI:10.1111/j.1435-5957.2011.00360.x

⑬ Yamato, H., Asymptotic distribution of number of distinct observations among a sample from mixture of Dirichlet processes, Bulletin of Informatics and Cybernetics, 査読有, VOL.44, 2012, 41-47.

⑭ 佐井至道, 層化無作為標本から得られる個票データに対するリスク評価, 岡山商大論叢, 査読無, 47巻, 2011, 1-22.

⑮ 瀧敦弘, 集計表におけるセル秘匿問題（１）, 広島大学経済学部 Discussion Paper Series, 査読無, No.2011-02, 2011, 1-8.

⑯ 稲葉由之, 事業所・企業を対象とした統計調査データに関する二次利用の課題, 2010 年度統計技術研究会報告, 査読無, 2011, 1-9.

⑰ Ito, S., The employment status and involvement in society of Japanese youth: Based on microdata from the 'survey on time use and leisure activities', Meikai Economic Review, 査読有, Vol.23, 2011, 30-48.

⑱ 伊藤伸介, わが国におけるミクロデータの新たな展開可能性について―イギリスにおける地域分析用ミクロデータを例に―, 明海大学 経済学論集, 査読有, Vol.23, 2011, 36-54.

⑲ Ito, S. and Takano, M., A method to quantitatively assess confidentiality and potential usage of official microdata in Japan, The 58th World Statistics Congress of the International Statistical Institute, 査読無, 2011, 1-5.

⑳ Kamiya, H., Takemura, A. and Terao, H., Periodicity of non-central integral arrangements modulo positive integers, Annals of Combinatorics, 査読有, VOL.15, 2011, 449-464.
DOI:10.1007/s00026-011-0105-6

 Hara, H. and Takemura, A., A Markov basis for two-state toric homogeneous Markov chain model without initial parameters, Journal of the Japan Statistical Society, 査読有, VOL.41, 2011, 33-49.

 Kashimura, T., Sei, T., Takemura, A. and Tanaka, K., Properties of semi-elementary imsets as sums of elementary imsets, Journal of Algebraic Statistics, 査読有, VOL.2, 2011, 14-35.

 Yamato, H., Residual fractions of size-biased permutations of discrete prior associated with Gibbs partitions, Bulletin of Informatics and Cybernetics, 査読有, Vol.43, 2011, 41-52.

 Maruyama, Y. and George, E. I., Fully Bayes factors with a generalized g-prior, Annals of Statistics, 査読有, VOL.39, 2011, 2740-2765.
DOI:10.1214/11-AOS917

 Miyawaki, K., Omori, Y. and Hibiki, A., Panel data analysis of Japanese residential water demand using a discrete/continuous choice approach, Japanese Economic Review, 査読有, Vol.62, 2011, 365-386.
DOI:10.1111/j.1468-5876.2010.00532.x

 伊藤伸介, 高野正博, 秋山裕美, 後藤武彦, ミクロデータにおける有用性と秘匿性の定量的な評価に関する研究, 製表技術参考資料, 査読無, No.14, 2010, 1-40.

 Hara, H., Aoki, S. and Takemura, A., Minimal and minimal invariant Markov bases of decomposable models for contingency tables, Bernoulli, 査読有, Vol.16, 2010, 208-233.
DOI:10.3150/09-BEJ207

 Hara, H. and Takemura, A., A localization approach to improve iterative proportional scaling in Gaussian graphical models, Communications in Statistics Theory and

Methods, 査読有, Vol.39, 2010, 1643-1654.
DOI:10.1080/03610920802238662

- Omori, Y. and Miyawaki, K., Tobit model with covariate dependent thresholds, Computational Statistics and Data Analysis, 査読有, Vol.54, 2010, 2736-2752.
DOI:10.1016/j.csda.2009.02.005
- 星野伸明, 公的統計ミクロデータ提供制度の課題, 日本統計学会誌, 査読有, Vol.40, 2010, 23-45.

〔学会発表〕(計１８６件)

① 伊藤伸介, 出島敬久, 小林良行, 就業構造基本調査と賃金センサスを用いた賃金分布の比較とその応用, 2012 年度統計関連学会連合大会, 2012 年 9 月 12 日, 北海道大学（北海道）.
② 大和元, Ewens sampling formula の分割数のエッジワース展開, 2012 年度統計関連学会連合大会, 2012 年 9 月 11 日, 北海道大学（北海道）.
③ 佐井至道, 寸法指標のノンパラメトリック推定に対する種々の改善, 2012 年度統計関連学会連合大会, 2012 年 9 月 10 日, 北海道大学（北海道）.
④ 原尚幸, 赤坂拓哉, 竹村彰通, Toric homogeneous Markov chain モデルのマルコフ基底と格子基底, 2012 年度統計関連学会連合大会, 2012 年 9 月 10 日, 北海道大学（北海道）.
⑤ 大森裕浩, 渡部敏明, Realized Stochastic Volatility モデル-日次リターンと Realized Volatility の同時モデル化, 2012 年度統計関連学会連合大会, 2012 年 9 月 10 日, 北海道大学(北海道).
⑥ 星野伸明, エビデンスに基づいた匿名化, 2012 年度統計関連学会連合大会, 2012 年 9 月 10 日, 北海道大学（北海道）.
⑦ 渋谷政昭, Pitman 確率分割とミクロデータ公開リスク管理, 2012 年度統計関連学会連合大会, 2012 年 9 月 10 日, 北海道大学（北海道）.
⑧ Hoshino, N., Invitation to mathematical statistical disclosure control, imsAPRM2012, 2012 年 7 月 3 日, Tsukuba International Congress Center（茨城県）.
⑨ Sai, S., Nonparametric estimation for population size indices, imsAPRM2012, 2012 年 7 月 3 日, Tsukuba International Congress Center（茨城県）.
⑩ Sibuya, M., Random partition of number and multi-index, imsAPRM2012, 2012 年 7 月 3 日, Tsukuba International Congress Center（茨城県）.
⑪ Yamato, H., Random partitions of integers based on mixtures of Dirichlet processes, imsAPRM2012, 2012 年 7 月 3 日, Tsukuba International Congress Center（茨城県）.
⑫ Omori, Y., Realized stochastic volatility models and their applications using high frequency financial time series, imsAPRM2012, 2012 年 7 月 3 日, Tsukuba International Congress Center（茨城県）.
⑬ Kamishima, T., Akaho, S. and Sakuma, J., Fairness-aware learning through regularization approach, IEEE International Workshop on Privacy Aspects of Data Mining 2011, 2011 年 12 月 11 日, Vancouver, Canada.
⑭ 大和元, 混合ディリクレ過程からの標本に基づく確率分割, 2011 年度統計関連学会連合大会, 2011 年 9 月 7 日, 九州大学（福岡県）.
⑮ 伊藤伸介, 村田磨理子, 後藤武彦, ミクロデータにおける攪乱的手法の有効性の検証, 2011 年度統計関連学会連合大会, 2011 年 9 月 6 日, 九州大学（福岡県）.
⑯ 秋山裕美, 後藤武彦, 星野なおみ, 伊藤伸介, 山口幸三, 教育用ミクロデータの試行提供について, 2011 年度統計関連学会連合大会, 2011 年 9 月 6 日, 九州大学（福岡県）.
⑰ 原尚幸, 青木敏, 竹村彰通, Running Markov chain without Markov basis, 2011 年度統計関連学会連合大会, 2011 年 9 月 6 日, 九州大学（福岡県）.
⑱ 稲葉由之, 荒木万寿夫, 世帯定義の拡張に関する考察, 2011 年度統計関連学会連合大会, 2011 年 9 月 6 日, 九州大学（福岡県）.
⑲ 渋谷政昭, 種の多様性調査におけるサブサンプリング, 2011 年度統計関連学会連合大会, 2011 年 9 月 6 日, 九州大学（福岡県）.
⑳ 星野伸明, 自然数の確率分割における周辺分布, 2011 年度統計関連学会連合大会, 2011 年 9 月 7 日, 九州大学（福岡県）.

- 小林良行, 公的統計ミクロデータのオンサイト利用 － 一橋大学オンサイト利用施設の現状と課題, 2011 年度統計関連学会連合大会, 2011 年 9 月 5 日, 九州大学（福岡県）.
- 佐井至道, 事後層化による個票データのリスク評価の改善, 2011 年度統計関連学会連合大会, 2011 年 9 月 5 日, 九州大学（福岡県）.
- 渋谷政昭, Galton-Watson 確率木と Riordan 配列: Lagrange 分布族再論,

2010 年度統計関連学会連合大会，2010 年 9 月 8 日，早稲田大学（東京都）.

- 原尚幸，竹村彰通，マルコフ連鎖の homogeneity の正確検定について，2010 年度統計関連学会連合大会,2010 年 9 月 8 日，早稲田大学（東京都）.
- 大和元，混合ディリクレ過程からの標本による確率分割の分布，2010 年度統計関連学会連合大会，2010 年 9 月 8 日，早稲田大学（東京都）.
- 石原庸博，大森裕浩，交差レバレッジのある多変量確率的ボラティリティ変動モデルのベイズ推定，2010 年度統計関連学会連合大会，2010 年 9 月 8 日，早稲田大学（東京都）.
- 佐井至道，サンプリング法を考慮に入れた個票データのリスク評価，2010 年度統計関連学会連合大会，2010 年 9 月 6 日，早稲田大学（東京都）.
- 小林良行，集計データを用いた疑似個別データ作成について，2010 年度統計関連学会連合大会，2010 年 9 月 6 日，早稲田大学（東京都）.
- 伊藤伸介，高野正博，秋山裕美，後藤武彦，ミクロデータにおける有用性と秘匿性の定量的な評価の試み，2010 年度統計関連学会連合大会，2010 年 9 月 6 日，早稲田大学（東京都）.
- 山口幸三，秋山裕美，後藤武彦，伊藤伸介，教育用ミクロデータの作成方法について，2010 年度統計関連学会連合大会，2010 年 9 月 6 日，早稲田大学（東京都）.
- 星野伸明，模造個票データの必要性について， 2010 年度統計関連学会連合大会，2010 年 9 月 6 日，早稲田大学（東京都）.

〔図書〕（計５件）

① 竹村彰通，佐井至道，ミネルヴァ書房，ファイナンス・景気循環の計量分析（第 11 章 官庁統計データ匿名化の統計学），2011，291-310.
② 和合肇，各務和彦，ミネルヴァ書房，ファイナンス・景気循環の計量分析（第 6 章 ベイズ型パネル空間プロビット・モデルを用いた地域景気循環モデル），2011，135-168.

〔その他〕

ホームページ等
http://www.osu.ac.jp/~sai/index.html

６．研究組織

(1)研究代表者
佐井　至道（SAI　SHIDO）
岡山商科大学・経済学部・教授
研究者番号：30186910

(2)研究分担者
渋谷　政昭（SIBUYA　MASAAKI）
慶應義塾大学・理工学部・名誉教授
研究者番号：20146723
瀧　敦弘（TAKI　ATSUHIRO）
広島大学・大学院社会科学研究科・教授
研究者番号：40216809
稲葉　由之（INABA YOSHIYUKI）
慶應義塾大学・経済学部・教授
研究者番号：80312437
伊藤　伸介（ITO　SHINSUKE）
明海大学・経済学部・准教授
研究者番号：90363316
小林　良行（KOBAYASHI　YOSHIYUKI）
統計情報研究開発センター・研究開発本部・研究員
研究者番号：80553643
（H23→H24：研究協力者）

(3)連携研究者
竹村　彰通（TAKEMURA　AKIMICHI）
東京大学・情報理工学系研究科・教授
研究者番号：10171670
星野　伸明（HOSHINO　NOBUAKI）
金沢大学・人間社会研究域経済学経営学系・教授
研究者番号：00313627
和合　肇（WAGO　HAJIME）
京都産業大学・経済学部・教授
研究者番号：00091934
大森　裕浩（OMORI　YASUHIRO）
東京大学・大学院経済学研究科・教授
研究者番号：60251188
田村　義保（TAMURA　YOSHIYASU）
統計数理研究所・モデリング研究系・教授
研究者番号：60150033
丸山　祐造（MARUYAMA　YUZO）
東京大学・空間情報科学研究センター・准教授
研究者番号：30304728
佐久間　淳（SAKUMA　JUN）
筑波大学・システム情報工学研究科・准教授
研究者番号：90376963
大和　元（YAMATO　HAJIME）
鹿児島大学・理学部・名誉教授
研究者番号：90041227