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　元 年　 ６ 月 １０ 日現在

機関番号：１２６１３
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００４２６

研究課題名（和文）EDINET等を活用した企業調査における匿名化技法の考察

研究課題名（英文）Anonymization methods in enterprise surveys using the EDINET

研究代表者
白川　清美（SHIRAKAWA, KIYOMI）

一橋大学・経済研究所・准教授

研究者番号：２０７５５０９５

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では、企業の特定に使用する統計量と度数別数値パターンを登録したデータベースを構築した。これにより、特定のための数値を基に検索が可能となったので、元データの値が特定できる場合とできない場合の閾値を見つけることが可能となった。さらに、企業調査である科学技術研究調査のミクロデータを使用し、最大値・最小値などの統計量や度数を変数ごとに作表した。また、最大値・最小値を秘匿するためのトップ・ボトムコーディングの境界値の分析を行い、どの様にそれぞれの境界値を決定すればよいかの回帰モデルを作成した。これらの検証により、研究者に有用な匿名化技法の明確化が可能となった。

研究成果の学術的意義や社会的意義
学術的意義は、二次利用推進を図るため、企業調査における匿名化技法の研究であり、攪乱手法による有用性を損なわない秘匿性向上の検証である。特に、基本統計量のデータベース作成とその機能の活用により、トップ・ボトムコーディング、リコーディング（区分統合）、センシティブルールの閾値の設定など、より多くのシミュレーションに基づいた匿名化を可能とすることである。
社会的意義は、①企業情報の利用を要望している研究者等の有用性が向上する。②数値パターンに基づいた基本統計量のデータベースを利用し、匿名化に有用な手法を活用することにより、データの匿名化や作成した合成データが研究者等の利便性を格段に向上させる、である。

研究成果の概要（英文）：In this study, we built a database that registered statistics and frequency patterns for identify companies.
Therefore, it became possible to search by statistics for identifying the original data, and to find a threshold when the value of the original data can be identified and when it cannot be identified. Furthermore, using micro data of Survey of Research and Development, statistics such as maximum value and minimum value, frequency by each variable were tabulated. In addition, we analyzed top and bottom coding boundary values to conceal maximum and minimum values. We created a regression model of how to determine each boundary value. These examinations have made it possible to clarify useful anonymization techniques for researchers.

研究分野：公的統計二次利用

キーワード：統計的開示抑制　トップ・ボトムコーディング　持ち出し審査　基本統計量　合成データ　オープンデータ　オンサイト施設

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

近年、「オープンデータ」「ビッグデータ」や「パブリックユースミクロデータ」の利用が活発化している。これらの背景を受け、公的統計個票データの利用を促進するため、2009 年 4 月に統計法が改正され、国勢調査、就業構造基本調査及び国民生活基礎調査、等の 7 調査の匿名データが提供されている。しかし、これらの匿名データは、世帯調査に限定され、それらの所管は国民生活基礎調査以外は総務省統計局のみである。＊1

このことから、現時点において、企業を対象とした匿名データは作成されていない。山口（2008）＊2 は、「企業調査を対象とした匿名データ作成は困難である。」と述べている。また、星野(2010)＊3 は、企業調査を、匿名化が困難である 4 種類のデータの１つに挙げている。ただし、これらの研究には「何故、匿名化が困難であるのか」等、詳細な分析結果の記述はない。しかしながら、この困難性は以下に示す様に、①エンティティベースの集計と②企業の公開情報が起因していると推測ができる。つまり、①企業調査における企業の産業分類格付けが、本来的には 1 企業・多企業産業分類となるところ、エンティティベースの集計であるため、1 企業・１企業産業分類となっている点、②企業の HP や有価証券報告書によって、詳細な経営状況を公表している点である。

そのため、①と②の照合により企業の特定が容易となり、またこれらが起因して、識別子（企業番号等企業を明確に特定できる項目）や準識別子（企業名、住所、電話番号等）の情報を削除しただけでは、匿名化が困難（資本金、従業者数、企業の産業分類等の項目から当該企業の特定が可能）であることを示している。一方、Lenz(2008)＊4 は、「ドイツにおける企業の匿名化は、特殊なデータ削除と攪乱によって可能となった。」と述べている。さらに、近年、フランスの The National Institute for Statistics and Economic Studies (INSEE) ＊5 では、企業情報をオープンデータとして提供している。このことから、企業の匿名データ作成は、企業の公開情報とは異なるユニットに変換する等、創意工夫を凝らした匿名化技法によって可能となる。

なお、日本における匿名データ作成では、前述の「企業調査を対象とした匿名データ作成は困難である。」の定説が研究者間で根付いているため、このテーマの研究が停滞しているのが実情である。そこで、本研究では、これまでの経済調査の実務経験と諸外国の事例に基づき、企業調査を対象とした匿名データ作成に取り組むこととする。

[参考文献]

*1 総務省統計局「匿名データの作成・提供及びオーダーメード集計」 http://www.stat.go.jp/info/tokumei/

*2 山口幸三 (2008). 「政府統計の個票利用と統計法改正–試行的提供の経験を踏まえて–」,経済研究, 59, pp.139‐152

*3 星野伸明（2010）.「公的統計ミクロデータ提供制度の課題」, 金沢大学経済学経営学系 Discussion Paper Series No.15, pp.1-23

*4 Lenz,R. (2008). Risk Assessment Methodology for Longitudinal Business Microdata, Wirt Sozialstat Archiv, 2, pp.241–257

*5 INSEE http://www.insee.fr/en/bases-de-donnees/fichiers-detail.asp

２．研究の目的

日本において、公的企業調査の匿名データは存在しない。それは、トレードオフ関係にある「有用性（情報量の損失の変動）」と「秘匿性（秘匿すべきカテゴリ）」の両立が困難なためである。ただし、秘匿すべき対象は特異性のある「大企業」に集中しており、全てのカテゴリではない。

本研究では、金融庁の EDINET(Electronic Disclosure for Investors' NETwork)データを活用し、公的企業調査における秘匿すべき対象の「特異性・有用性・秘匿性」を検証する。具体的には、「企業等の財務情報をセグメント別に分割することと、ミクロアグリゲーション（詳細集計）やリコーディング（区分統合）の手法により、特異性のある企業等の匿名化が可能であるか否か」を実証分析する。特に、秘匿性の検証における「秘匿ルールの曖昧な閾値の明確化」を図る。

３．研究の方法

(1)金融庁の EDINET(Electronic Disclosure for Investors' NETwork)データの活用

大企業を対象としている財務データを取得する。

(2)公的企業調査における秘匿すべき対象「特異性・有用性・秘匿性」の検証

EDINET データをセグメント別に分割することと、ミクロアグリゲーション（詳細集計）やリコーディング（区分統合）の手法により、特異性のある企業等の匿名化が可能であるか否かを実証分析する。

(3) 数値パターン別の基本統計量に基づいたデータベースの構築

基本統計量の閾値を設定するため、度数別に取り得る値のレンジや最頻値を求めることができるデータベースを構築する。

(4)秘匿性の検証における秘匿ルールの曖昧な閾値の明確化

作成したデータベースに基づき、統計量別の閾値を設定する。
(5)匿名化技法の検証
リサンプリング、リコーディング、トップ・ボトムコーディングなど、データベースに基づいた検証を行う。
(6)匿名化技法の適用
匿名化する変数別に、最適な技法を適用する。

４．研究成果

本研究では、匿名化技法に関する研究により、企業調査の二次利用推進と研究者の利便性の向上にあるため、経済センサスなどの全数調査ではなく、サンプリング調査のミクロデータを利用した匿名化技法の検証をした。

これまでの成果には、本研究以外の成果である白川・阿部*1 があり、これをベースに企業調査に有用な方法が導出できた。匿名化技法には、撹乱的手法と非撹乱的手法があるが、非撹乱的手法にいくつかの有用な方法がある。それは、リサンプリング率、トップ・ボトムコーディング、基本統計量に基づいた合成データ作成に加え、EDINET データの部分利用である。

具体的には、企業調査である科学技術研究調査のミクロデータを使用し、最大値・最小値など個々の値や集計後の度数の少ない区分にならないように作表した。また、最大値・最小値を秘匿するためのトップ・ボトムコーディングの境界値の分析を行い、どの様にそれぞれの境界値を決定すればよいかの回帰モデルを作成した。さらに、EDINET データでの置き換えによる匿名化も検証した。

今後も引き続き、企業調査の匿名化技法の研究を継続し、企業調査に基づいた合成データの作成と提供を行う予定である。さらに、合成データの作成に利用するために構築した基本統計量のデータベースも公開する予定である（*2、*3、*4、*5)。

[参考文献]
*1 白川清美, 阿部穂日「匿名データの利用改善に向けた調査研究報告書」, 統計委員会, 2017, pp1-90 http://www.soumu.go.jp/main_content/000482460.pdf
*2 Kiyomi Shirakawa, Yutaka Abe, Shinsuke Ito, “Creating an 'Academic Use File' Based on Descriptive Statistics: Synthetic Microdata from the Perspective of Distribution Type,” In: Josep Domingo-Ferrer and Mirjana Pejić-Bach eds., Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2016, Dubrovnik, Croatia, September 14 – 16, 2016, Proceedings, Springer, 2016, pp.149-162
*3 白川清美、大規模データベースに基づく秘匿すべきセルの抽出―MM形式によるスパース行列の効果的な活用―、2016 年度統計関連学会連合大会、2016、p.239
*4 阿部穂日・白川清美・千葉亮太、匿名データ作成のためのリコーディング-決定木による最適な境界値の分析-、2016 年度統計関連学会連合大会、2016、p. 238
*5 中松建・白川清美、数量表への汎用的な秘匿ルールの適用-「τ-ARGUS」の有効性の検証-、2016 年度統計関連学会連合大会、2016、p.174

５．主な発表論文等

〔雑誌論文〕（計　2 件）

① 白川清美、千田浩司、田中哲士、高橋慧、菊池亮、“公的統計の実証分析における秘密計算とその部分計算過程を公開することの安全性の検討”、経済研究 69 巻 2 号、2018、pp.145-152、査読有
② 田中哲士、阿部穂日、高橋慧、菊池亮、土井厚志、千田浩司、白川清美、“公的統計への秘密計算適用に向けたミクロデータの統計分析”、マルチメディア、分散、協調とモバイル（DICOMO 2017)論文集、2017、pp. 424-429、査読有

〔学会発表〕（計　12 件）

① 中松建、白川清美、変数の組み合わせパターンに基づく合成データの作成、平成 30 年度研究集会 「ミクロデータから見た我が国の社会・経済の実像」、2019
② Kiyomi Shirakawa, ’Optimal Boundary Value for Creating Anonymized Microdata: Empirical Analysis based on Economic Survey Data’, uRos 2018
③ Kiyomi Shirakawa, Koji Chida, Satoshi Takahashi, Satoshi Tanaka, Ryo Kikuchi, Dai Ikarashi, ‘Analysis Of Official Microdata Using Secure Statistical Computation System’, New Zealand Statistical Association and the International Association of Statistical Computing (Asian Regional Section) Joint Conference 2017
④ Kiyomi Shirakawa, ’A Proposal of a Simple and Secure Statistical Processing System using Secret Sharing’, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, 2017
⑤ 白川清美、経済統計分析への秘密計算技術適用に向けた一考察、第 61 回経済統計学会、2017
⑥ 白川清美、SDC に基づく線形回帰係数の安全性の検証、第 61 回経済統計学会、2017

⑦ 白川清美、秘密計算及び統計的開示制御を組み合わせたセキュアな統計処理システムの提案、2017 年度統計関連学会連合大会、2017
⑧ 白川清美、記述統計量に基づく秘匿すべき回帰モデルの検証、2017 年度統計関連学会連合大会、2017
⑨ Kiyomi Shirakawa, 'Challenges in improving the quality and amount of statistical utilization -New uses of official statistics in Japan', ISI 2017 MARRAKECH 61st World Statistics Congress, 2017
⑩ 白川清美、オンサイト利用における持出し審査：「τ-ARGUS 」の有効性の実証分析、政府統計の二次的利用研究会、2017
⑪ 白川清美、利用者指向に基づく匿名データの作成、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」、2016
⑫ 白川清美、阿部穂日、二次利用のための大規模データベースの作成－合成データの作成と秘匿セルの抽出－、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」、2016

〔その他〕
ホームページ等　http://shirakawa-kiyomi.strikingly.com/

６．研究組織

(1)研究分担者
研究分担者氏名：相良　直哉
ローマ字氏名：Sagara, Naoya
所属研究機関名：一橋大学
部局名：経済研究所
職名：助教
研究者番号（8 桁）：70433852