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

平成　２９　年　　４　月　１１　日現在

機関番号：　１３３０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３８０２６９

研究課題名（和文）エビデンスに基づいた匿名化の実証

研究課題名（英文）A Positive Study on Evidence Based Anonymization

研究代表者
星野　伸明（Hoshino, Nobuaki）

金沢大学・経済学経営学系・教授

研究者番号：００３１３６２７

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：個体情報が保護されていることの法的定義は、個体識別が出来ない状態である。しかしこの定義は技術的に曖昧なので、個票データは過剰に匿名化されたりする。このような社会的損失を正すには、個体識別可能性を明確化しなければならない。本研究は個体識別ができない状態を統計的推定の対象として明確化することを目的とし、細部まで詰めた理論を構築した。また本研究はこの理論を実証するため、既公開の匿名データが個体識別可能性の情報を持っていることに着目し、平成15年の住宅・土地統計調査匿名データの個体識別可能性を計測した。

研究成果の概要（英文）：The information protection of an individual is legally defined as unidentifiability. However, this definition is technically vague, and it often leads to the overanonymization of microdata. To mitigate this social loss, we need a clear technical definition of identifiability. Our study aims to clarify this unidentifiability as an object of statistical estimation, and has constructed the detailed theory of this idea. Also our study notes that the exting Anonymized Data have the information of unidentifiability of the data, and thus has measured the possibility of identification on these data of Housing and Land Survey of 2003.

研究分野：　経済統計学

キーワード：　個票開示リスク　離散分布論

１．研究開始当初の背景

公的統計などのミクロデータを分析利用に供する場合、調査客体の情報は保護しなければならない。ここでデータの変換による情報保護手法を「匿名化」と呼ぶ。大幅な匿名化により情報保護が破れる危険（開示リスク）は低くなるが、データの分析価値（有用性）は下がる。このようなトレードオフがあるので、匿名化水準の最適化が問題になる。いわゆるビッグデータ利用等における情報学の研究（プライバシー保護データマイニング）も、同じ問題を抱えている。

この最適化については、開示リスクと有用性を引数とする効用関数の最大化とみなす Duncan et al. (2001, Technical Report)の議論が有名である。このような考え方は、開示リスクや有用性を個別に研究する上で有用である。しかし現実の匿名化が許容範囲か否かは、実務家の価値判断の問題とされる。結果として実務家は途方に暮れ、判断を何らかの権威に投げているのが現状であろう。

現行法の下での匿名化の実務的判断について、客観化の余地が残っていると研究代表者は考える。統計法やいわゆる個人情報保護法では、個体情報が保護されていることを、個体識別が不可能な状態と定義する。これを制約とする匿名化の最適化は、個体識別が不可能な範囲で有用性を最大化することになる。このように考えれば、所与の開示リスクの下で個体識別が可能か否かの判定が、匿名化において決定的となる。個体識別の可能性による保護状態の区分は、U.S. Privacy Act など外国法でも採用されている。従って個体識別可能性の判定は、普遍的な問題である。

研究代表者はこの判定を価値判断ではなく、統計的推定問題として定式化する。星野(2013)では、個体識別の難易度δが閾値αを下回れば個体識別が可能と考える。難易度の測度については後述するとして、αは観測から推定可能である。すなわち個体識別がδに依存した確率で観測されるとして、過去の事例における個体識別の有無から母数αの尤度が書ける。αの最尤推定値は、これまで個体識別が発生していなければ、過去の事例で最小の難易度となる。

具体的な個体識別の難易度測度が問題になるが、開示リスクの研究蓄積が利用できる。Marsh et al. (1991, JRSS, A)によれば、(識別が起きる|識別を試みる)という事象は、(a)識別試行用の情報と公開情報が誤記等でずれていなく、(b)公開情報に個体が含まれ、(c)その個体は母集団一意であり、(d)母集団一意と確証できる、という条件が同時に満たされることである。これら条件の事象をそれぞれaからdと書けば、Pr(識別が起きる|識別を試みる)=Pr(a)Pr(b|a)Pr(c|a,b)Pr(d|a,b,c) ということになる。星野(2013)は、個体識別が不可能という状態が Pr(識別が起きる|識別を試みる)=0 と同値なことに着目する。だとすれば個体識別が不可能な状態の判定は、Pr(a)Pr(b|a)Pr(c|a,b)Pr(d|a,b,c) がゼロか否かの判断となる。そして Pr(a,b,c)は一般にゼロにならないので、結局 Pr(d|a,b,c)がゼロか、つまり所与のデータについて母集団一意が確証できる可能性の評価が、多くの問題で決定的となる。

Pr(d|a,b,c)の評価について、社会に流通する個体情報の網羅的調査が必要と Elliot et al. (2010, in LNCS 6344)は言う。しかしそのような調査は実施困難であろう。故に星野(2013)は必要な情報が不明瞭なことを前提として、Pr(d|a,b,c)が正か推定する。母集団一意の確証(d|a,b,c)は、正確に記録されている公開母集団一意(a,b,c)が増えれば容易になるはずだ。故に Pr(a,b,c)が閾値を超えれば、Pr(d|a,b,c)が正と考える。

結局細かい議論は省略するが、-Pr(a,b,c)を実質的にδとみなせる。このようなδを数値評価すれば、閾値αが共通する過去の事例が匿名化水準の統計的証拠となる。δの評価では母集団一意数の推定が必要となるが、研究代表者の手法(Hoshino (2001, JOS))が国際的に見ても最良だろう。αについては、社会に流通する個体情報などの要因に依存するはずだが、これらの要因は同制度の事例ではほぼ等しいと考えられる。故に本研究では、匿名データに着目する。

既公開の匿名データからαを推定すれば、これまで述べた理論が匿名化水準の決定方式として実証される。また匿名データ自体について、匿名化を緩和できる範囲の根拠が得られる。

匿名データ改善要望の多くは匿名化の緩和であり、本研究は匿名データの改善と直結する。例えば統計委員会諮問第 13 号の答申は、「匿名化措置を課す情報及びその程度が異なる複数の匿名データの作成の可能性について検討する必要」を指摘する。また諮問第 37, 44 号の答申でも同等の指摘がなされている。この問題については研究計画・方法の項で説明するが、匿名性の許容範囲（α）が分からないと明確な結論は出ない。

さらにαと個体情報などの要因の関係をモデル化すれば、匿名データ以外のデータについても、本研究によって個体識別可能性の根拠が得られる。またαの最尤推定は保守的だが、このようなモデル化によって、より精密にαを推定できる可能性がある。

２．研究の目的

利用可能な匿名データを全て用い、個体識別可能性を判定する測度δの閾値αの最尤推定値を求める。これにより匿名データにおける匿名化水準の許容範囲を明らかにする。この結果から、匿名データについて匿名化の緩和が可能か、特に複数匿名データの作成が

可能か検討する。また閾値αと社会に流通する個体情報等の要因との関係についてモデル化を試みる。

３．研究の方法

匿名データについて匿名化水準の許容範囲を求めるため、本研究は実績がある手段を選択した。まず匿名データの利用について、既に一部の利用手続きを先行して進めている。また残りの匿名データ利用の内諾も得た。それから匿名化水準の測度を求めるプログラムは、研究代表者が過去の研究で用いた物を修正して使う。平成 26 年度に総務省所管の匿名データを処理し、平成 27 年度以降は厚労省所管分を処理した後、複数匿名データ作成の問題等を考察し論文にまとめる。生成した統計は公開する。効率的に目的を達成するための工夫としては、高速な計算機の導入が重要である。また情報学を含む広範囲での情報収集、交換を行う。

４．研究成果

研究開始当初のアイデアは星野(2013)の理論モデルに基づいているが、細部が曖昧であった。まず測度δに基づいて個体識別の可能性を判定する必然性についてアドホックな面があった。それから測度が依存する母集団一意数がキー変数の選択によって大きく変化するので、キー変数の選択方針をかなり具体化する必要があった。その他、個体識別行為の成功とその観測が区別されていなかったなど、モデルの定式化に不十分な点があった。

これらの未検討であった点について、本研究期間を通して理論的に適切な整理をすることが出来た。結果として、現実の匿名データ審査体制の改善を理論的根拠に基づいて考えることが出来るようになり、具体的な改善点も浮かんだ。

それらの結果は査読付き論文として公刊した。実はキー変数の選択理論は「未解決問題」(Fung et al., 2010) と言われていて、満足できる理論展開は短くならない。従ってこの報告書で十分な説明は不可能だが、これを理論的にも実務的にも解決したと考える。このように本研究全体で扱っている問題は難しかったので、結果の論文が異例の長さになった。査読者にも論文が長すぎると指摘されたが、編集長の判断で公刊の機会を頂くことが出来た。

上記論文の執筆過程での考察内容は実際の匿名データ作成にも影響を与えており、本研究の成果は国民生活へ直接的に還元されている。研究代表者は総務省のオーダーメード集計・匿名データ作成検討会議の構成員を長らく務めており、この間、匿名データ作成の技術的問題について理論的なアドバイスを続けている。本研究の問題設定はこの場の課題から生まれており、成果はこの会議でのアドバイス内容に反映させている。

また論文公刊にはいたらなかったが、母集団一意数推定に用いる統計モデルについての研究においても進展があった。

Hoshino (2009, JOS)で提案している CCP 分布族について、本研究期間中に二次モメントを明らかにした。CCP 分布族は多項分布と異なる二次モメントを持つように多項分布を一般化してある。興味深いことに、CCP 分布族の分散共分散行列は多項分布のそれの定数倍になる。故に本分布族の相関係数行列は全て同じで多項分布のそれになることが導かれる。それから CCP 分布族の例として擬似多項分布が挙げられるが、特殊ケースの第二種の擬似二項分布について、上述の結果を利用して分散の新しい表現が得られる。この表現はセル平均の母数と分散の母数が積で分離しているので扱いやすい。これらの結果は研究代表者が独自に提案した離散多変数分布族に関する理解を大きく深めた。

その他これまでの研究で、コルチンモデルという正の整数の確率分割族について周辺分布を求めるには階乗モメントを反転すればよいことは分かっていたが、この反転公式の初出文献が不明であった。本研究期間中にこれを突き止めたと考えている。また、この確率分割族の重要な例であるピットマン分布について、その特殊ケースであるユーエンス分布と比較しながら、周辺分布の数値計算を体系的に重ねることが出来た。数値計算の結果は、導出した周辺確率が正しいことを示唆している。ピットマン分布の第一周辺分布は母集団一意数の分布として用いられるので、上記の結果は個票開示リスク評価理論の精緻化に寄与した。

このように理論研究は着実に進展したが、匿名データを用いた実証研究は、当初の目的を部分的に達成するにとどまった。結果の公刊に至ったのは平成 15 年の住宅・土地統計調査の匿名データのリスク評価のみである。その他、全国消費実態調査、社会生活基本調査、就業構造基本調査、住宅・土地統計調査の匿名データについては、提供を受けた全年度のリスク評価を比較可能な形でまとめた。当初、世帯単位の調査と個人単位の調査の比較方法を分かっていなかったが、本研究期間中に一意個人数の比較による解決へ至った。上記４調査を比較した結論として、匿名データの都道府県コードに施されている匿名化は緩和が可能なことが示唆される。この議論については英文査読誌に投稿したが、棄却されたため内容を再検討中である。労働力調査や国民生活基礎調査については、検討不足のまま研究期間の終了を迎えた。結果として閾値αの検討が不十分なので、複数匿名データ発行可能性の問題について妥当な指針を得るまで至らなかった。

本研究期間中は想定外の事由で多忙を極め、時間のかかる実証研究について憾みを残す結果となった。既に入手したデータを無駄にするつもりはなく、早い段階で分析して結果をまとめる予定である。

<引用文献>

星野伸明「エビデンスに基づいた匿名化」，金沢大学経済学・経営学系ディスカッションペーパー，No. 21，2013.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　2 件）

(1) 星野伸明「エビデンスに基づいた匿名化」，日本統計学会誌，Vol. 46，pp. 1-42，2016，査読あり

(2) 星野伸明「有用な匿名化データ---経験からの学習」，情報処理学会論文誌，Vol. 55，pp. 1956-1963，2014，査読あり

〔学会発表〕（計　6 件）

(1) Nobuaki Hoshino，“On the Marginals of a random partitioning distribution”，International Conference of Statistical Distributions and Applications 2016 (ICOSDA 2016)，Niagara Falls (Canada)，2016-10-15

(2) 星野伸明「多項分布の一般化について」，統計関連学会連合大会，岡山大学（岡山市），2015-09-07

(3) Nobuaki Hoshino，“Evidence Based Anonymization”，The 60th World Statistical Congress (ISI2015)，Rio de Janeiro (Brazil). 2015-07-27

(4) Nobuaki Hoshino，“Applying the Quasi-Multinomial Distribution”，The 24th International Workshop on Matrices and Statistics，Haikou (China)，2015-05-28

(5) 星野伸明「情報保護の統計モデル」，第 17 回情報論的学習理論ワークショップ (IBIS2014)，名古屋大学（名古屋市），2014-11-19

(6) 星野伸明「住宅・土地統計調査匿名データの開示リスク」，統計関連学会連合大会，東京大学（東京都），2014-09-15.

〔図書〕（計　0　件）

〔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://stat.w3.kanazawa-u.ac.jp/owner/papers.html

６．研究組織
(1)研究代表者
星野　伸明（HOSHINO，Nobuaki）
金沢大学・経済学経営学系・教授
研究者番号：00313627