公的統計ミクロデータ提供制度の課題

星野　伸明*

Issues on Disseminating Japanese Official Microdata

Nobuaki Hoshino*

概 要

公的統計のミクロデータ利用を促進する上で大きな課題を二点挙げると、一点目は利用目的制限の緩和、二点目は利用可能な統計調査の拡大である。本稿ではこれらの課題の解決策を検討する。前者については、一般目的汎用ファイルの提供が望ましい。これを実現するには、絶対的な匿名性を持ち有用なデータの作成を目標とするべきである。後者については、事業所・企業調査など匿名化が難しいデータも提供するのが望ましい。困難な匿名化について先進的事例を調査したところ、原データの統計的性質を一部保存するデータの作成が目標となっている。そのようなデータを確率的に生成するのが模造 (synthesis) という概念であり、一般目的汎用ファイルの作成でも有用である。従って本稿では、模造手法も整理して紹介する。

Among the most important issues on promoting the secondary analysis of Japanese official microdata, the author would like to focus upon two expansions: One is to expand purposes allowed to use microdata, and the other is to expand the variety of statistics allowed to be used. The former should be attained by producing public use files that are absolutely anonymous, while the utility of data is preserved. The latter requires to anonymize easily-identifiable statistics. These two goals seem to entail the same methodology known as synthesis. Terefore the present article reviews researches and practices on disseminating synthetic data.

Key Words and Phrases: Statistical Disclosure Control, Synthetic Data

1　はじめに

統計法（平成 19 年法律第 53 号）の第 1 条には「公的統計が国民にとって合理的な意思決定を行うための基盤となる重要な情報」と書かれている。このような理念を実現する方策の一つは、調査票情報を二次利用した分析の促進である。

*金沢大学経済学部, 〒 920-1192, 石川県金沢市角間町, E-mail: hoshino@kenroku.kanazawa-u.ac.jp

二次利用は旧統計法の下でも目的外使用として可能であったが、森 (2004) が説明するようにその運用は極めて厳しいものであった。現行統計法では一般の学術研究や高等教育について、「匿名データ」が利用可能である。匿名データは統計法第 2 条第 12 項で規定され、「一般の利用に供することを目的として調査票情報を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む。）ができないように加工したもの」である。また学術研究目的でいわゆる「オーダーメード集計」が認められているが、集計も個体識別ができないようような加工の一種である。なお高度な公益性を有する研究等を行う者に限り、原調査票情報の「目的外利用」（本稿では 33 条申請による利用を便宜上このように呼ぶ。）も認められる。

このように二次利用制度は整備されたが、運用が始まったばかりという事もあり、課題がないわけではない。例えば 2010 年 4 月の時点で匿名データが利用可能なのは、総務省の 4 調査（全国消費実態調査、社会生活基本調査、住宅・土地統計調査、就業構造基本調査）にとどまっている。今後は利用可能な統計調査の種類を増やすべきである。また二次利用は目的を限り認められるが、利用可能な者も増やすべきではないだろうか。統計法第 3 条 3 項には「公的統計は、広く国民が容易に入手し、効果的に利用できるものとして提供されなければならない」と書かれている。

自由に二次利用が出来ない理由は、調査客体の情報の保護である。調査客体が他者に知られたくない属性（変数）を「センシティブ」と言う。センシティブ変数の暴露—「開示 (disclosure)」を制限しなければ調査への協力を損ない、正確な統計の作成は望めない。なお本稿では暴露のニュアンスが無い「公開」と開示を区別する。

開示制限は、利用者の制限と情報量の管理を組み合わせて行う。利用者の制限とは、統計利用目的の審査や守秘義務契約、倫理規定等により、調査客体にとって不利益となる行為を直接規制する。データ情報量の管理とは、開示範囲が限定されるようにデータを変換する。このようなデータ変換を「匿名化」と呼ぶ。匿名化は、利用者制限が有効でない場合（契約違反や事故によるデータ流出）でも効果的である。

匿名化による開示制限は、「統計的開示制限 (Statistical Disclosure Limitation, SDL)」又は「統計的開示管理 (Statistical Disclosure Control, SDC)」と呼ばれる。SDC の現時点で最良の概説の一つとして、Skinner (2009) を挙げておく。また Willenborg and de Waal (1996, 2001) は基本文献である。

以上のように考えると、統計当局が提供するデータは匿名化が弱いほど利用者が制限され、匿名化が強いほど利用対象者が増える。これが二次利用制度設計の基本原理である。故に利用可能な者を増やすなら、より強い匿名化を施したデータが必要となる。また利用可能な統計調査を増やす際に個体識別が容易なものも含むならば、強力な匿名化が必要である。

強力な匿名化は、真の粗いデータか偽の詳細なデータを意味する。そしておおよそ統計的分析においては、偽の詳細なデータの方が価値が高い。従って SDC では「論理的に偽だが統計的に真のデータ」、すなわち原データの統計的性質を保存するデータの作成が目的となる。ところが日本の統計作成者とユーザ双方にとって、論理的に偽だが統計的に真のデータはなじみがない考え方のようである。故に本稿では、論理的には偽のデータによる匿名化の考え方を解説する。結論と

して公的統計の二次利用制度を充実させるため、論理的には偽のデータを確率的に生成する手法が不可欠である。

なおオーダーメード集計は、集計者が介在する事により、利用者を制限しデータ情報量も制限する制度と言える。このような二次利用形態は目的によっては十分だが、データの探索的解析等は不可能になる。公的統計の二次利用における制限を緩和するという大方針からは、オーダーメード集計は次善の策にすぎない。以下では議論の焦点を、ミクロデータ提供制度に絞る。

本論文の構成は以下の通りである。まず 1.1 節でデータの匿名化に関する概念を定義する。これをうけて 2 章の主題は、ミクロデータの利用促進における論点整理である。具体的には 2.1 節で、誰もがミクロデータを自由に使えるようにする手段が考察される。この理想の実現には、匿名化の技術的理解を背景にした法解釈が必要である。2.2 節では、多くのミクロデータを使えるようにする手段が考察される。匿名化が難しい統計調査としては、企業・事業所データや地理情報を含むデータ、パネルデータ、連結データ等が知られている。この節ではこれらのミクロデータ提供について、先進的事例を報告する。やはり匿名化が難しい統計調査については、論理的に偽だが統計的に真のデータを提供するのが現実解と思われる。そしてそのようなデータは、確率的に生成するのが主流である。1.1 節で正確に定義されるが、論理的には偽のデータを確率的に生成する手法を「模造」と呼ぶ。3 章では模造について様々なアプローチを紹介する。最後に 4 章で残された論点にふれる。

1.1 若干の準備

以下の議論を厳密にするため、本節では匿名化の基本概念を定義しておく。なお本稿の定義は、必ずしも一般的ではない。

まずはミクロデータの匿名化を形式的に整理しておこう。一般に p 個の変数からなる n 個体の（原）ミクロデータは、$n \times p$ 行列 $\boldsymbol{X}$ で表される。公開されるミクロデータは $m \times q$ 行列 $\boldsymbol{Y}$ で表す。通常は $m < n, q \leq p$ である。匿名化 μ とは $\boldsymbol{Y} = \mu(\boldsymbol{X})$ と書ける変換の事である。変換を行列で表現する場合も有る (Duncan and Pearson, 1991)。すなわち行列匿名化 $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ は

$$\mu(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{X}\boldsymbol{B} + \boldsymbol{C} \tag{1}$$

と表現される。ここで $\boldsymbol{A}$ はレコードを変換する操作、$\boldsymbol{B}$ はフィールドを変換する操作、$\boldsymbol{C}$ は移動操作である。なお Cox (1994) が議論しているように、同一の μ でも行列匿名化として複数の表現がありえる。いずれにせよ変換と言っても無限にあり、μ の良さの概念が必要となる。匿名化の良さは $\boldsymbol{Y}$ の「開示リスク」と「有用性」で測られる。

開示リスクとは技術的にセンシティブ変数が開示出来る可能性であり、実際に開示が起きる危険性と区別される。実際の危険性は、非統計的な開示制限の影響も受ける。例えば目的外利用であれば、匿名データの利用よりも審査等の非統計的開示制限が厳しいので、データの開示リスクが高くても実際の危険性は押さえられる。

$\boldsymbol{Y}$ のある行（レコード）が特定個体と識別されて起きる開示を、「識別開示」と呼ぶ。それから識別がない開示もありえる。例えば特定地域居住者の年収は必ず 300 万円とデータから分かるとしよう。この場合、特定地域に住んでいる A さんのレコードが識別出来ないとしても、年収は開示される。既知の個体属性を「キー変数」と呼ぶが、一般にキー変数所与でセンシティブ変数の狭い区間推定が可能な場合を「推測開示」と考える。

匿名データの定義では推測開示は問題とされず、識別（開示）の回避が匿名化の目的となっている。匿名化に対する有力な攻撃は既知の個体の識別なので、識別開示を一義的に管理するのは必ずしも不当ではない（現在提供されている匿名データでは推測開示もある程度制限されている）。識別開示を重視する理由を他にも挙げておこう。しばしばデータ利用の目的は、キー変数所与でセンシティブ変数を統計分析する事である。故に推測開示を制限するなら、データ分析の精度が制限される。これは後で説明するデータの有用性が低いという事である。従って、識別開示リスクを十分低くし推測開示は出来るだけ制限しないのは自然である。推測開示リスクの管理は難しくない事もあり、以下でリスクとは識別開示リスクを指す。

（識別）開示リスクの測度は各種提案されているが、基本的なアイデアは 2 つしかない。一つ目は $\boldsymbol{X}$ と $\boldsymbol{Y}$ が似ていれば開示リスクが高いとする。二つ目は $\boldsymbol{Y}$ の中でレコードが他と大きく異なるほど、またそのようなレコードが多いほど開示リスクが高いとする。従って m, q 所与で開示リスクの低い $\boldsymbol{Y}$ を作ろうとすれば、全レコードが $\boldsymbol{X}$ と関係ない同じ値を持てば良い。もちろんこのようなデータは分析の意味が無いので、有用性概念が必要となる。

匿名化の有用性測度も、基本的なアイデアは 2 つである。一つ目は開示リスクの裏返しで、$\boldsymbol{X}$ と $\boldsymbol{Y}$ が似ていれば有用性が高いとする。二つ目は、特定の分析手法を前提として、$\boldsymbol{X}$ の分析結果と $\boldsymbol{Y}$ の分析結果が似ていれば有用性が高いとする。

つまり一般に大きくデータを変換すれば、開示リスクは下がるが有用性も下がる。このようなトレードオフ下で、意思決定問題として最適な匿名化を達成するというパラダイムは Duncan et al. (2001) が明文化し、広く受け入れられている。

具体的な匿名化としては、「嘘」をつかない方法が基本である。例えば 78 歳というデータを 65 歳以上という区間で表示する方法は「トップコーディング」と呼ばれる。区間表示の考え方は、分割表による表章と同じである。区間表示により、ある個体の公開値は他と大きく異ならないようになるので、開示リスクは低下するとみなされる。「サブサンプリング」という手法も良く使われ、これは一部レコードの公開をしない。公開をしないという事は、全変数について区間 $(-\infty, \infty)$ と表示する事と同じである。

匿名化の「嘘」を定式化しよう。$\boldsymbol{Y}$ の i 行 j 列の要素 y_{ij} が真であるとは、$\boldsymbol{Y}$ の第 i レコードに対応する個体が母集団に存在し、その個体の第 j 属性の観測値が y_{ij} と矛盾しない事を言う。例えば公開データのある個体が原データ $\boldsymbol{X}$ に含まれていなくても、実在の個体の別の調査で得た観測値が表示されているならば真である。それから属性の観測値と表現値が同じでなくても真かもしれない。例えば年齢の観測値「78 歳」は、表現値「65 歳以上」と矛盾しない。

$\boldsymbol{Y}$ が $\boldsymbol{X}$ 所与で正の確率で偽の要素を持つ匿名化を「攪乱的 (perturbative)」と呼ぶ。基本的な

攪乱的手法は、「加法的ノイズ」と「スワッピング」である。まず加法的ノイズは、データに誤差を加える。その最も単純な場合、(1) において $\boldsymbol{AXB} = \boldsymbol{X}$ かつ $\boldsymbol{C}$ の各要素は平均 0 の正規乱数である。もちろん誤差は正規分布に従う必要はなく、例えば正のデータでは他の分布が望ましいだろう。故に Bowden and Sim (1992) では、誤差分布は経験分布をシフトさせたものである。次にスワッピングは、異なるレコード間で一部の変数の値を交換する事を言う。これも行列匿名化として表現可能である。スワッピングについては、例えば Gomatam et al. (2005) を参照すると良い。Gomatam 等は、スワッピングの開示リスクと有用性を複数の測度で評価した。結果として、おおまかなトレードオフ関係が見られる。

匿名化が「模造的 (synthetic)」とは、$\boldsymbol{Y}$ が $\boldsymbol{X}$ 所与でランダムな要素を持つ場合を言う。具体的には、（データの有用性を落とさないように）統計的性質を保つ分布から標本を抽出して $\boldsymbol{Y}$ の行とする。模造的匿名化によって生成されるデータを「模造 (synthetic) データ」と呼ぶ。

上で例に挙げた加法的ノイズの最も単純な場合は、攪乱的かつ模造的である。しかし $\boldsymbol{X}$ が条件を満たしたら必ずスワッピングを施すような匿名化は、攪乱的だが模造的とは言わない。本稿の定義では、模造と攪乱に包含関係は無い事に注意すべきである。例えば真の値を含む区間の長さをランダムに決めて表示するような匿名化は、模造的だが攪乱的ではない。

それから模造概念については「完全」と「部分」が区別される。すなわち $\boldsymbol{Y}$ の全ての要素が $\boldsymbol{X}$ 所与でランダムな匿名化を「完全模造的 (fully synthetic)」と呼び、一部の要素がランダムな場合を「部分模造的 (partially synthetic)」と呼ぶ。

2　ミクロデータ利用範囲の拡大にむけて

公的統計のミクロデータ利用を促進するため、本章では二方向の可能性を考察する。2.1 節では利用者の制限を緩和する方策を検討する。2.2 節では、利用可能な統計調査を増やす場合の課題を探る。結果として両方向ともに、攪乱による匿名化は避けられない。

2.1　日本版一般目的汎用ファイルの可能性について

一般に学術研究目的で、審査、契約を経て提供されるミクロデータを「科学目的汎用ファイル (Scientific Use File, SUF)」と呼ぶ。匿名データは SUF の一種である。他方、利用資格に制限がないデータを「一般目的汎用ファイル (Public Use File, PUF)」と呼ぶ。ただし米国 Federal Committee on Statistical Methodology (FCSM, 2005, p.5) によれば、National Center for Education Statistics 等の提供する “PUF” は、匿名化を破らないという誓約が利用に必須であった。現在日本では公的統計の PUF は提供されていない。しかし例えば米国の国勢調査や、労働力調査にあたる Current Population Survey 等は PUF が提供されている。

PUF は利用のコストが低い点で望ましい。例えば匿名データの利用コストは、実費程度の手数料と契約書類作成等の時間である。金銭的コストは、例えば学会が一括契約するなど支払制度に

改善の余地があると考えるが、大きな障害ではないかもしれない。しかし利用の審査をする側もされる側も、高度な専門性を有する人材を拘束する時間コストは無視できない。また研究の成算がみこめない段階でもデータ利用コストがかかるなら、挑戦的な研究の意欲をそぐだろう。

そもそも Fienberg (2005) は、データが公共財なら使用制限は非合理的と指摘する。データの使用を審査する「門番」には、データ使用を促進するインセンティブが働かない。故に匿名度の高いデータを自由に利用させる方が、匿名度の低いデータを制限的に利用させるより良いと Fienberg は主張する。「門番」にインセンティブを与えれば良いのかもしれないが、本稿ではその可能性は考察しない。

日本版 PUF は、「レプリカデータ」という名称で必要性を指摘されてきた。例えば日本学術会議 (2005, p.7) の報告では、「リサンプリングして匿名化し、さらにスワッピングしたデータ」をレプリカデータと呼び、研究者間で自由な配布を可能にすることが提案されている。松田 (2008) は「単に匿名化だけでなくスワッピング等の処理を施して、実際のデータに基づいているが個別の回答者とのリンケージ不可能なデータ」をレプリカデータと呼び、「大学院生にも自由に使えるデータにしたかった」と述べている。これらの用例ではデータの利用目的が研究教育と想定されているが、自由に使えるデータという主旨は PUF と同じである。

レプリカデータに関する検討状況は、総務省政策統括官（統計基準担当）(2008, p.10) に記載されている。曰くレプリカデータの作成・提供については、「匿名データの一形態と考えるか、全くの疑似データと考えるかによりその取り扱いが異なるので、今後さらに定義を明確化した上でその作成・提供について検討する必要がある」。

このように日本版 PUF に関する議論は未熟である。そのためか美添 (2008) が指摘するように、一般の社会人、高校生でも利用出来る PUF は、統計法に成文化されていない。しかし PUF は、広く国民が容易に入手出来るという統計法の理想に近い。故に日本版 PUF の可能性を検討してみよう。

まず PUF を匿名データの一形態と考える場合、提供条件は統計法第 36 条で定まる。本条によれば、研究教育目的以外での匿名データ利用は、総務省令で定めない限り認められない。PUF は利用目的を問わないので、総務省令で利用目的を広く定めなければ提供出来ない事になる。この場合、個体識別を試みないという条件で利用可能とするのが一案である。しかし統計法第 38 条は匿名データの利用について、手数料の納付を定めている。故に PUF は匿名データでない方が望ましい。匿名データ (SUF) でない、匿名のデータ (PUF) はありえるだろうか。

SUF と PUF の違いから考察しよう。法律的にはドイツの例が参考になる。Brandt et al. (2008, p.140) によれば、SUF は「事実上の匿名性」が保証されているファイルの事である。それに対し、PUF には「絶対的な匿名性」が要求される。事実上の匿名性とは、著しく時間と経費、労力をかけない限りファイルに含まれる個体を識別出来ない状態を言う。また絶対的な匿名性は、確実に個体識別の可能性が除かれた状態である。なお事実上の匿名性は、計算量的な困難性にも依存する。計算機科学の発展により、現在は計算が困難でも将来は問題ない場合もあるかもしれない。しかし時間が経過すれば、過去のデータの識別は難しくなり意味も薄れる。

濱砂 (2000) によれば、1987 年以前のドイツでは、研究用ミクロデータに絶対的な匿名性が要求された。その結果、研究者の要求する水準の情報を持つデータの作成は困難だったという。従って法改正をし、事実上の匿名性を持つデータの提供が開始された。このように、強い匿名化を施したデータが常に万人を満足させるわけではない。しかし PUF は、SUF ほど詳細でなくても良いはずだ。そして 1987 年に比べれば匿名化技術は進歩しており、現在でも絶対的な匿名性の「実効性がきわめて乏しい」とは限らない。濱砂 (2000) が紹介している当時の匿名化手法（ドイツ連邦統計局『標準的な匿名化措置』）は、初等的である。

Felsö et al. (2001) によれば、当時のドイツを含むヨーロッパ及び北米の匿名化実務は、米国のワーキングペーパー (FCSM, 1978) の影響を受けている。この論文はタイトルの “disclosure avoidance” を “disclosure limitation” に変えて刷新 (FCSM, 1994) され、更に第二版 (FCSM, 2005) が発行されている。タイトルが変わったのは、開示の回避ではなく制限が妥当な目標と理解されたからである。これらの改訂は、匿名化手法研究の進展を反映している。なお FCSM (2005) は、米国における匿名化実務の基本参考書である。

さて、日本版 SUF と PUF の匿名化水準がドイツの例にならうとして、制度的矛盾が生じないか検討しよう。

まず SUF としての匿名データが、事実上の匿名性しか持たないとする。冒頭で引用した匿名データの定義では、個体が識別できない事が要件であった。この「識別できない」の整合的解釈は、二通りあるように思われる。一つ目の解釈では、事実上の匿名性も識別できない事にはかわりないと考える。二つ目の解釈は、匿名データは事実上の匿名性しか持っていないが、審査、契約をしているので、全体として絶対的な匿名性が達成されていると考える。いずれにしても、問題はないように思える。

次に PUF についてだが、絶対的な匿名性は事実上の匿名性より強い。故に絶対的匿名性を持つデータは、匿名データの条件を満たす。従って PUF が調査票情報を加工したものなら、匿名データの一種となる。匿名データでない PUF を作るとしたら、調査票情報を加工したものと言えないデータとするしかない。そのようなデータは作れるだろうか。

一つの可能性は、匿名データを更に匿名化したデータは、調査票情報を加工したものでないと解釈する事であろう。まず統計法第 2 条 11 項によると、調査票情報とは「統計調査によって集められた情報のうち、文書、図画又は電磁的記録（電子的方式、磁気的方式その他人の知覚によっては認識することができない方式で作られた記録をいう。）に記録されているものをいう。」匿名データはこれを加工して作られるが、この加工は病的な例を除いて非可逆変換である。すなわち匿名データから原調査票情報を正確に再生出来ない。故に匿名データは、調査票情報と等価ではない。従って匿名データを更に匿名化したデータは、調査票情報でないものを加工したものである。

匿名データを更に匿名化して PUF を得る事の利点は大きい。まず PUF が匿名データにない情報を持つ可能性を捨てられるので、開示リスク評価が単純になる。また利用者にとっても、PUF に含まれる個体情報が匿名データにも基本的に含まれる事は望ましい。何故なら PUF を利用して試行的研究をし、成算を持って SUF の利用に移る場合、標本の違いによる結果の違いを限定出来

る。実務的にも、匿名データ作成時点で開示リスク要因が確認される（特に基幹統計の場合は統計委員会への諮問が有る）ため、そこで達成される事実上の匿名性から絶対的匿名性を目指す方が容易であろう。

しかし調査票情報を加工したものを更に加工したものも匿名データと考えれば、上記の整理は成立しない。その場合でも PUF の可能性を指摘しておこう。使われる匿名化技法が「加工」でなければ、結果のデータは匿名データではない。

3 章でレビューする模造的匿名化では、調査票情報からデータの分布モデルを推定し、その分布から確率標本を抽出する。この抽出された標本が公開されるミクロデータとなる。モデルの推定は統計量の算出に依存し、それは調査票情報の加工かもしれない。しかしモデルからの確率抽出は、もはや調査票情報の加工とは呼べないのではないだろうか。

模造は加工ではないとしよう。この場合、模造が一部でも入り絶対的な匿名性を持つファイルを、匿名データではない PUF として扱えば良い。模造が入ったものは全て匿名データではないとすると、次節で議論するように将来的に匿名データが作れなくなる恐れがあるので、そのような解釈は避けるべきである。

模造を適切に使えば、匿名データでない PUF が作成可能と思われる。ところが統計作成者とユーザ双方が、模造の前提である攪乱を避けているように見える。攪乱に抵抗が有るのは日本だけではない。例えば東欧や CIS の諸国は匿名化の経験が浅い時点で、スワッピングや加法的ノイズのような攪乱手法をほとんど使っていない。Felsö et al. (2001) によると、Eurostat の匿名化ガイドラインが出来たのは 1996 年である。その後 United Nations Economic Commission for Europe (UNECE) が 1998 年に東欧と CIS の国々で匿名化の実態を調査している。東欧では匿名化を支える法的基盤は在ったものの、匿名化の実装に困難を抱えている状態であった。CIS では法的基盤さえ無かった。2000-2001 年の UNECE の再調査では、全ての国が匿名化の重要性を認識していた。しかし匿名化について主要な関心は行政上の困難であって、匿名化の理論的側面への関心は弱かった。従って東欧と CIS 諸国は攪乱を避けて、場当たり的に匿名化していたという。

データが供給されても使われなければ意味がないので、ユーザが持つ攪乱への抵抗感を検討しなければならない。ユーザは攪乱されたデータが有用か、すなわち分析結果が原データのそれと近いか疑っているようだ。

実際、攪乱によるデータの歪みは例がある。2000 年の米国国勢調査の PUF は、65 歳以上の男女比が異常な値を示す場合があるという注意書き (Data Note 12, Census Bureau) が 2008 年 10 月になって加わった。これは攪乱的匿名化の副作用である（手法の詳細は非公開）。なお 2000-2006 年の American Community Survey の PUF も同様の問題を抱えていると、2009 年 4 月に警告が出ている。また Alexander et al. (2010) は、2004-2009 年の Current Population Survey も同様の問題が有ると指摘する。

このような事がありえるので、模造データが正当と保証されなければ、利用は広がらないだろう。究極的には、データ作成者の信用が問われる事になる。統計当局は最も信頼されているデータ提供主体なので、公的統計として慎重に模造データを発行するのが良いように思う。統計調査

の結果の公表は統計法で義務づけられているので、この「結果」もしくは公表しなければならない「政令で定める事項」として PUF を位置づけるのが良い。

他にデータの正当性を保証する方法として、検証サーバも研究されている（Reiter et al., 2009）。検証サーバは公開（模造）データと原データを保持し、ユーザは分析プログラムを入力する。出力は、原データと模造データでの分析結果の違いの大小である。

PUF はデータ「製品」として、作成・提供を独立行政法人や研究者が行う事も考えられる。ただし必要な匿名性を持たない製品を PUF として流通させてはならず、その匿名性審査は不可欠である。そして匿名性審査を通過したデータは、当局公認データアーカイブに蓄積すれば良い。製品の有用性については（あまり）審査せず、ユーザの選択にまかせて良いかもしれない。模造データは特定の分析手法を前提に作られる事が多く、いかなる目的にもふさわしいとは限らない。従って多様な分析に対応するには多様なデータがあった方が良く、データ供給のハードルは低い方が望ましい。なお PUF が統計でなくてもその作成を統計的研究とみなせば、調査票情報の二次利用の要件（統計法第 32 条 1 項）を満たす。この場合 PUF は、匿名データではなく原データから生成しても良いはずだ。しかし多様な PUF を結合して個体識別を試みる可能性を考えると、PUF は匿名データを匿名化して作成するものとした方が無難である。

いずれにせよ、模造データ作成は高度なノウハウを必要とする。経験を蓄積するには、作成したデータに関するユーザからのフィードバックを制度化しておく事も重要だ。その上で良くできた模造データ提供の実績を重ねれば、ユーザに受容されるだろう。

ここまで日本版 PUF の可能性を二通り指摘した。一つ目は、匿名データを更に匿名化したデータを PUF とする。二つ目は、模造データを PUF とする。前者の方が単純で望ましいと思うが、前者を採用するとしても模造を避けられるとは限らない。1987 年以前のドイツのように、模造を使わずに絶対的匿名性を達成しようとすれば、有用性の低いデータとなるだろう。有用かつ開示リスクの低いデータを作る上で、模造は重要である。

2.2 匿名化が困難な統計調査について

匿名データとして最初に利用可能となった総務省の 4 調査は、過去にミクロデータが一橋大学で試行的に提供されていた。試行的提供の成果については松田他 (2000)、運用実態については山口 (2008) を参照せよ。

山口は、試行的に提供された統計調査の選定基準も説明している。要約すると (a) 事業所・企業調査は匿名化が困難とされているので、世帯調査に限定する。(b) 国勢調査は全数調査という事もあり除外する。(c) 労働力調査のような月次調査は、「提供する場合にある程度の期間の時系列データを提供しないと利用価値が低くなることや、その間に調査項目などの変更があることなど、事務量の問題」があるので除外する。

結果として、試行的提供の第一期は 5 年毎に周期的に実施する大規模標本調査である 3 調査を対象に選定し、住宅・土地統計調査は第二期から対象とした。なお試行的提供の利用条件は匿名データのそれと比べて厳しく、全体的な開示制限は慎重である。

上で挙げた (a) から (c) の 3 点を見ると、匿名化が困難な統計調査の提供を避けた事が分かる。匿名化の困難が明記されているのは (a) のみだが、(b) で挙げられている全数調査は（調査対象の不確実性が消えるので）匿名化が比較的難しい。(c) の月次調査についても、余り認知されていないが匿名化は困難を抱えている。星野 (2009) は労働力調査について、標本数の変動効果を除いた開示リスクを月次で評価し、これが全て同等か検討した。結果として、最も危険な月と安全な月は開示リスクの構造が違うかもしれないという事であった。だとすれば既に匿名データが提供されている統計調査でも、新しい時期のデータを提供する場合、再度匿名性の審査が必要かもしれない。それから労働力調査のローテーション構造に起因して、開示リスクの自己相関が検出された。このような開示リスクの時間依存性の評価は十分研究されていない。なお労働力調査のローテーション構造については、加納 (2003) を参照のこと。

利用可能な統計調査を増やすなら、困難な匿名化に挑む必要がある。それから現在利用可能な匿名データは「調査実施から五年以上経過したもの（平成 21 年 2 月 13 日付内閣府統計委員会第 2 回匿名データ部会議事概要）」だが、時期的に新しいデータも必要だろう。新しいデータの匿名化も困難な課題である。以下では困難を克服し、オーダーメード集計という次善の策ではなく、ミクロデータを提供する方策を探ろう。

一般に匿名化が難しい種類のデータとしては

- 企業・事業所（ビジネス）データ
- 地理情報を含むデータ
- パネルデータ
- マッチングによる連結データ

が知られている。まずはこれらのミクロデータ提供について、先進的事例を見てみよう。

ビジネスデータ　大きな企業や事業所は全数調査される事も多く、その場合調査された事が既知となる。また外れ値として目立つので、ビジネスデータの匿名化は難しい。しかしミクロデータの提供実績が無いわけではない。Jabine (1993) によれば、米国で当時ビジネスミクロデータを発行していたのは、Statistics of Income Division, Internal Revenue Service (IRS) のみである。IRS は Spruill (1982, 1983) の匿名化研究を支援しており、その成果を活かし発行に踏み切ったという。ドイツでは Lenz (2008) によると、2002-2005 年にかけてビジネスミクロデータの事実上の匿名化を研究するプロジェクトが進められた。結論としては、模造（特殊な情報削減手法と攪乱的手法）を用いれば、クロスセクションのミクロデータについて、事実上の匿名化は可能という事である。他にビジネスミクロデータの匿名化の例として、外れ値の縮小と回帰モデルによる模造 (Franconi and Stander, 2002) を挙げておく。

地理情報を含むデータ　次に地理情報についてだが、これが個体識別について強力なキー変数となる事は良く知られている。例えば総務省政策統括官（統計基準担当）(2009) による「匿名データの作成・提供に係るガイドライン」（以下、ガイドライン）の匿名データチェックリスト（案）でも、地理的情報は特記事項である。また Franconi and Stander (2002) の重要な教訓は、匿名化における地理情報保護の有効性であった。再集計にしか地理情報を必要としないユーザーについては、オーダーメード集計で済ませれば良い。しかし小地域推定等で、詳細な地理情報が必要な場合もある。

Gutmann et al. (2008) は、社会科学研究者向けに書かれた、地理情報の匿名化に関する啓蒙的論文である。この論文に基づいて地理情報の匿名化手法を大まかに分類すると、攪乱（地理情報を不正確に変換)、点の集計、となる。つまり地理情報だからといって、匿名化手法の基本的アイデアは特別なものではない。ただし地理情報ならではの工夫はありえる。例えば Leitner and Curtis (2006) は、点を集計する広さを変えて地図を作り、人間が差をあまり認知できない範囲を実験で評価した。また地理的ブロックの中心や交差点など、地図的解釈が可能な場所に集計値を表示せよと主張している。なお視覚化は情報を削減するので、匿名化としてはたらく。しかしその理論的整理はなされていない。他に Young et al. (2009) は、近い地域に似た世帯が住む傾向を利用したスワッピングを提案している。Reiter (2009) は、模造データには詳細な地理情報を含められると主張する。実際に米国 Census Bureau は 2006 年 2 月から “OnTheMap” というウェブアプリケーションを一般公開しており、オンラインで地理的な労働力の属性をオーダーメード集計出来る。一般にオンライン集計は、範囲をずらした集計結果の差分を見る事で個体情報が分かる可能性がある。しかしこの場合集計するデータが模造なので、集計の履歴を神経質に管理しなくても構わない。

パネルデータ　ドイツでは Federal Statistical Office のリサーチデータセンター主導で、2006 年からビジネスパネルデータの事実上の匿名化プロジェクトが進行中である。Brandt et al. (2008) は、その成果の一部を報告している。パネルデータの匿名化手法は特別なものが知られているわけではないが、個体識別について時系列データが多重のキー変数となる事が重要である。Brandt 等は開示リスクを評価するため、市販のデータベースと匿名化パネルデータで類似する個体を探している。このような照合で、通常はキー変数ベクトルの距離を類似性とする。しかしパネルという構造を活かせば、キー変数の時系列変動の（順位）相関を類似度として利用可能である。ただ、結果としては通常の照合の方が強力であったという。その他の詳細については Lenz (2008) を見よ。

マッチングによる連結データ　パネルデータは、複数ファイルを連結して作る場合もある。登録情報（レジスター）等を連結すれば、単純な調査統計よりも有用なデータを作成可能である。しかしキー変数が増大するので、Zayatz (2007) は、おそらく模造を入れない限り米国 Census Bureau の PUF としては連結データを提供出来ないと述べている。Abowd and Woodcock (2001) は連結パネルデータの開示制限を議論した初期の論文だが、模造が採用されている。

米国 Census Bureau、IRS、Social Security Administration は、Survey of Income and Program Participation (SIPP) のパネルデータについて、詳細な収入と社会保障のデータを連結して提供する事を決定した。この連結パネルデータは非常にセンシティブな情報を含むので、強度の匿名化が必要である。結果として 600 以上の変数のうち、4 変数以外は全て模造する事になった。Abowd et al. (2006) は、この模造データの有用性を多角的に分析した詳細な報告書である。全てではないが、各種推定量の信頼区間は模造データと原データでおおよそ重なっていたそうだ。

上で例に挙げた SIPP 及び OnTheMap は、米国 Census Bureau の Longitudinal Employer-Household Dynamics (LEHD) 計画の成果物である。LEHD 計画は、労働者及び雇用者の縦断的な全国名簿の作成を目的とする。Wu and Abowd (2007) に基づき、計画の概要を紹介しよう。主要刊行物は Quarterly Workforce Indicators (QWI) であり、2003 年から “noise infusion” という簡易的模造手法を使用して一般公開されている。OnTheMap は、Census Bureau が公開する最初の本格的な（部分）模造データである。模造の目標は、Data Quality Act (2001) の一般的ガイドラインと Census Bureau (2006) の有用性、客観性、統一性に関わる基準に置かれている。なお模造データは、Census Bureau のリサーチデータセンターシステムの下で作られた。北米のリサーチデータセンターシステムについては、神林 (2008) が詳しい。SIPP 及び QWI は、コーネル大学のサーバ “Virtual Research Data Center (VirtualRDC)” で研究者に提供されている。SIPP のベータ版は 2007 年 5 月からユーザに提供され、フィードバックが集められている。模造データでは、分析における有用性が非常に重要と考えられているからだ。

ここまで主に米国の例を引用したが、個体識別が容易で匿名化が難しいデータを提供する場合、模造が積極的に使われている。3 章でより詳しく考察するが、模造手法により開示リスクは非常に低下するとみなされている。PUF に模造が使われたのも、開示リスクの大幅な低下が必要だからであった。有用性を保存しつつ匿名性を著しく増加させる場合、模造は現実解と思われる。

2010 年 4 月の時点で利用可能な匿名データには、全く攪乱は使われていない。そもそも試行的提供の時点で、山口 (2008) によれば、スワッピング及び誤差の付加は「我が国ではあまり望まれない方法」なので活用は考えられなかった。しかし攪乱が避けられないケースも、有ると思われる。ガイドラインにはスワッピング及び誤差の付加が匿名化手法の例として挙げられており、今後は使用の可能性が無いわけではない。

3　模造的匿名化について

日本で二次利用を促進する上で、PUF や匿名化の難しい統計調査のミクロデータ提供が課題となる事をこれまで述べた。これらの課題を実質的に解決するには、攪乱的匿名化が必要である。攪乱は特に危険なレコードを決定論的に選択して適用する場合もある。しかし Singh (2009) が指摘するように、決定論的な匿名化はバイアスを生じるかもしれない。従って確率的な匿名化、すなわち模造が正当化される。本章では模造に関するサーベイ結果を示す。

誤差の挿入による匿名化は自然な模造である。しかし模造の研究史上で画期的だったのは、確率モデルからの標本を公開するという発想である。計算機科学の分野で、このような発想の最も初期の例は Liew et al. (1985) と言われている。彼らはセンシティブな情報を、確率分布からの標本とする事を提案した。公的統計の文脈では、Rubin (1993) が「多重補定 (multiple imputation)」による完全模造データの公開を提案したのが最初である。

Rubin は完全模造により開示リスクは無くなると主張した。しかし Fienberg et al. (1998) が指摘するように、開示リスクは残る。例えば他から外れた属性値は、攪乱されていても照合により元の値がわかるかもしれない。模造データの開示リスク評価については後に再論する。

Little (1993) は、一部変数の模造を考察した。部分模造アプローチでは、キー変数とセンシティブ変数のいずれを模造するかで考え方が分かれる。1.1 節の議論を言い換えれば、キー変数の匿名化は識別の可能性を低減し、センシティブ変数の匿名化は推測開示を制限する。Little は開示リスクと有用性のトレードオフを考えて、キー変数を模造する方が望ましいと述べている。この主張は、匿名データが識別開示を管理している事と整合的である。

現在では模造の方法は多様である。それらは必ずしも排他的ではないが、以下のように 4 分類して紹介する。

- 多重補定
- ブートストラップ
- 十分統計量の保存
- 実験計画

これらに共通するのは、経験分布を基本に作られた母集団より標本を抽出して公開するという考えである。このようにすれば、原データのサイズ n より大きいサイズ m の公開データを自然に生成可能である。そのようなデータは、例えば教育用途で有益である。もし母集団作成に補助情報を用いるなら、$m > n$ とする意味は十分有る。また危険な一意の個体を保護するために、同様のキー属性を持つ架空の個体を公開データに加え、一意でなくす匿名化がありえる。この場合 $m > n$ とする意味がある。

多重補定　模造データの作成において主流の考え方は多重補定である。新しいサーベイ論文が書かれているので、Reiter (2009) を参照すると良い。なお補定は実務に定着しており、FCSM (2005, p.21) によれば既に 1990 年 Census of Population and Housing の summary tape files で空白にして補定するという模造の萌芽が見られる。

Raghunathan et al. (2003) によれば、多重補定の基本的アイデアは母集団の観測されていない個体を欠測とみなすという事である。そして欠測を補定して母集団を作り、そこから単純無作為抽出した標本を公開する。補定時に標本設計を考慮に入れる必要はあるかもしれないが、公開デー

タは単純無作為標本なので分析時に乗率は必要無い。なお匿名化における乗率の取り扱いは、十分研究されていない。例えばFienberg (2009) の広範なサーベイを見よ。

多重補定では一ファイルのみ公開するのではなく、同一の分布からの標本セットが複数公開される。このようにすれば、ユーザが比較的簡単に分析の分散を評価出来る。

多重補定による模造データ作成は、通常ベイズ的に理解されている。その場合は原データ所与での事後分布からの標本が公開データとして定義される。このように考えれば、レジスターやセンサス等の補助情報が事前分布として扱える。

連結データの匿名化で例示したSIPPは多重補定アプローチで作られており、Reiter (2009) によればAbowd et al. (2006) は多重補定による模造の最も包括的な検証結果である。PUF作成に多重補定を用いる詳細については、Reiter and Raghunathan (2007) を参照せよ。

ブートストラップ　データ分析では、真の多変量分布関数が推定できれば良い。従ってFienberg (1994) は、真の分布関数が推定出来る形態であればデータは模造でも良いはずだと主張した。そして経験分布もしくは平滑化した経験分布からの標本を模造データとする事が提案された。経験分布から復元抽出するなら、（単純な）ブートストラップである。

サブサンプリングをガイドラインでは「リサンプリング」と呼び、匿名データの作成に用いている。これは経験分布から非復元抽出する方法に他ならない。経験分布からの復元抽出は、安田 (2010) によれば、2.2節で言及した一橋大学の試行的提供においてユーザに不評であった。しかし著者は復元抽出を推奨する。復元によるデータの有用性低下はわずかだが、開示リスクの低下は無視出来ないように思われる。例えばサイズ n の原データで一意なあるレコードが、サイズ m の公開データに入る可能性は復元抽出の場合

$$1-\left(\frac{n-1}{n}\right)^m=\frac{m}{n}-\frac{m(m-1)}{2}\frac{1}{n^2}+O(n^{-3})$$

であり、非復元抽出の場合

$$1-\frac{n-m}{n}=\frac{m}{n}$$

である。つまり復元による保護効果は、おおよそ $-m(m-1)/(2n^2)$ となる。現在提供されている匿名データでは（住宅・土地統計調査を除き）$m/n=0.8$ であり、この場合の保護効果はかなり粗く見積もって-0.32もある。本来一意で危険なレコードが、標本に複数回入った場合の保護効果も大きい。

さて、次に経験分布の変換を考察する。有用性を損なわないように変換した経験分布から標本を抽出して公開するのは、匿名化の有力な形態である。Fienberg et al. (1998) は、原データとユーザが興味のあるデータは違うと指摘する。原データは、誤記やエディティング等の誤差を含んでいる。そしてユーザがこれらの誤差を除いたデータに興味があるなら、誤差を推定して除いた分布から標本を抽出すれば良い事になる。

それからGottschalk (2004)が指摘するように、多次元の真の分布を推定するには標本は余りにも疎であろう。故にベイズ的に事前分布を入れて情報を補い、事後分布から抽出して公開データとするのが一案である。これはベイズ的に理解された（多重）補定と同じ事になる。疎なデータの平滑化は開示リスク評価でも重要と認識されており、例えばIchim (2008)は局所尤度や罰則付き尤度の利用を考察している。またDi Consiglio and Polettini (2008)は疎な分割表について、前回の国勢調査を補助情報とした。

十分統計量の保存　有用性の測度は、適当な統計モデル分析を前提にする場合が多い。Little (1993)が主張するようにそのモデルの十分統計量だけを公開すれば、有用性が低下しない匿名化となる。しかし例えば教育用データセットは、ミクロデータとして与えないと教育にならない。故に原データの十分統計量の値を保存するミクロデータセット作成は、検討に値する。具体的には、十分統計量所与の条件付き分布（モデル）から標本を抽出すれば良い。

おそらく最も単純な例は、モデルが多変量正規分布に従う場合であろう (Burridge, 2003)。この場合の十分統計量は平均ベクトルと分散共分散行列になり、これらを保存する乱数の発生は容易である。そしてこのような匿名化なら、重回帰分析の結果が変わらない。これはそもそもLittle (1993)が例示した事である。

なおBurridgeが指摘するように、十分統計量所与の条件付き分布から生成する公開ファイルは一つである。多重補定のように、複数のファイルを提供しなくて良い。

他の例を挙げよう。対数線形モデルの十分統計量は周辺度数なので、Fienberg et al. (1998)は周辺度数を保存するような標本抽出を提案した。ただしFienberg (1998)が指摘するように、周辺度数が所与でセル度数の上限と下限が計算出来る事は、開示リスク評価で注意しなければならない。Fienberg等によれば、周辺度数の保存は多くの統計当局の慣行と整合的なので重要である。すなわち全数調査の結果を利用して事後層別（レイキング）する際、周辺度数は固定される。

スワッピングは低次元の周辺度数を保存する匿名化と言える。例えば世帯Aの居住地域と世帯Bの居住地域をスワップしたとして、両地域の世帯数は変わらない。故にランダムにスワッピングを繰り返せば、周辺度数を保存する模造データが生成出来るかもしれない。Takemura and Hara (2007)は、そのような事が可能な条件を議論している。

実験計画　Karr et al. (2006)のDiscussionにおいて、実験計画と匿名化の類似性を査読者が指摘したと書かれている。実験計画では、目的のデータ分析が所与で効率的な標本抽出を計画する。匿名化におけるデータの有用性を実験計画の最適性に置き換えれば、広範な成果が利用可能となる。実験計画では様々な最適性が提案されており、例えば線形モデルの推測にとって最適な計画についてはChaloner (1984)及びその参考文献を見よ。

Dandekar et al. (2002)はLatin Hypercube Sampling (LHS)により、一変数の周辺分布を崩さないように模造標本を抽出した。LHSは直交配列による実験計画とみなす事が可能である。Muralidhar and Sarathy (2003)は、原データと公開データの分布が近いほど有用性が高いと考え

る。LHS はそのような有用性を持つ模造手法である。なお Hoshino and Takemura (2000) が提案した強度 s の拡張 LHS を用いれば、s 変数の周辺分布を崩さないで模造標本を抽出可能である。Woodcock and Benedetto (2009) も、実験計画的ではないが周辺分布を保存するような模造を提案している。

このように様々なアイデアが提案されているが、模造は統計的な多変量モデリングに帰着する。ただし模造モデリングは通常と違い、オッカムの剃刀は適用されない。故に複雑なモデルで母集団を忠実に再現すれば良いかもしれない。しかし模造手法を（ある程度）公表し、データが分析に使えるかユーザが個別に判断するなら、複雑なモデルはふさわしくない場合もある。Woodcock and Benedetto (2009) が指摘するように、ノンパラメトリックな Classification and Regression Trees (CART) による模造 (Reiter, 2005) や機械学習系のモデルで生成した模造データは、分析に使えるかユーザが判断しにくい。単純で解釈が容易なモデルが望ましい事情は、違う意味で変わらない。

模造モデリングが通常と違う重大な点は、開示リスクの存在である。だからこそ Fienberg 等は、標本の再現ではなく母集団の再現を模造の目的としたのであった。模造データの開示リスクは、どのように考えたら良いだろうか。

撹乱されたデータの開示リスクについて、Hoshino (2009) の議論はおおよそ以下の通りである。まず攪乱がない場合、レコードが母集団で同じ属性を持つ個体群中で識別される可能性が開示リスクの主要な測度であり、これを第一の危険とする。この評価は母集団セル度数（特殊ケースが母集団一意）の推測として研究が蓄積されている。例えば Hoshino (2001) を参照のこと。攪乱がある場合の違いは、偽の要素の存在である。故に偽の要素を真の要素へ正確に戻せる可能性を、第二の危険と考える。このように考えれば、撹乱されたデータの開示リスクは、第一の危険と第二の危険の積となる。

この結論はベイズ的に正当化する事が出来る。Skinner (2008) は、攻撃者の所有するファイルと公表データセットの照合による個体識別が成功する可能性を考察した。あるレコードの照合が成功している事後確率を評価すれば、そのレコードと同じ属性を持つ母集団の個体数の逆数に比例する。Skinner の議論では攻撃者の所有するファイルが誤差を含む可能性を考慮に入れ、誤差が無い確率に母集団個体数の逆数を掛けて事後確率の近似値を得ている。ここで「誤差が無い確率」を「偽の要素を真の要素へ正確に戻せる確率」と置き換えれば、Hoshino (2009) の結論と同じである。

4　終わりに

匿名化されたデータの有用性を保つ上で、模造は有望なアプローチである。しかし模造データであっても、全ての分析ニーズを満たすわけではない。匿名化されたデータでは不十分な分析しか出来なければ、目的外利用を申請する事になる。二次利用分析を促進する上では、目的外利用の運用改善も別問題として必要であろう。匿名化されたデータの有用性を高くする意義は、相対的に低いコストで済むユーザが増える事である。

さて、Kinney et al. (2009) は 2008 年 5 月に行われた “Data Confidentiality: The Next Five Years” と称するワークショップの要約である。これを読むと、米国の研究者及び実務家の最近の問題意識が分かる。また面白い論点もあったので紹介しておこう。

最も重要性が高いのは SDC 技術の開発ではなく、SDC の健全な決定に資する方法論とツールの提供であると、Kinney 等は述べている。SDC は開示リスクと有用性のトレードオフ下での意思決定問題と了解されている。しかしこのアイデアは、実装の段階まで進んでいない。開示リスクと有用性の測度が十分発展していないのが理由と Kinney 等は述べている（私見では測度の精緻化は重要ではない）。そして開示リスクと有用性はデータの提供者と使用者に関わる概念だが、統計当局もデータ公開の利害関係者である。これを明示的に意思決定問題に組み入れるには、開示リスクと有用性のトレードオフという枠組を崩す必要が有るかもしれない。

計算機科学者と統計家の協力が必要という Kinney 等の指摘も正しい。計算機科学分野での匿名化研究は、Privacy Preserving Data Mining (PPDM) という名前で近年急速に発展している (Aggarwal and Yu, 2008)。SDC は統計当局にとって望ましい状態を達成するのが目的になっているが、PPDM ではこの前提が無く議論の自由度が高い。制約が緩い状況 (PPDM) での考察は制約がきつい状況 (SDC) でのブレークスルーを生む可能性が有り、重要である。

ところで少なくとも米国ではプライバシーは個人についての概念であり、法人の権利ではない。個人にも法人にも保証される権利が、「守秘性 (confidentiality)」である。Anderson and Seltzer (2009) は、米国において法人（企業）の守秘性概念が確立していく過程を、19 世紀後半から追っている。徴税等何らかの効率性の観点から、統計は常に目的外使用の圧力にさらされている。そして政治的に大きなテーマ、例えば戦争、独占禁止等のための目的外使用は、米国において実例があった。これは決して過去の事ではなく、2001 年の 9/11 テロ後においても、「愛国」目的の統計使用を許すため、議会が立法に動いた。今後も重大な事態が発生すれば、目的外使用の圧力に抗しきれないかもしれないと Anderson and Seltzer は言う。

上記の論文では、米国統計使節団長の Rice が統計基準部長時代に目的外使用の圧力と戦った事も述べられている。旧統計法の策定に強い影響を与えた Rice がこのような経験を持っていた事は、目的外使用が厳しく制限されていた事と関わりがあろう。現行統計法では二次利用の促進に舵を切ったが、守秘性が重要である事も社会に理解されなければならない。

日本のビジネスデータを例にとろう。上場企業は有価証券報告書の提出義務が有り、積極的に会社情報の公開を求められる。投資家の効率的な保護の為、上場企業のデータに守秘性を認めるべきではないと主張されるかもしれない。しかし個体識別が前提で提出する有価証券報告書は粉飾しても、守秘性の下の統計調査では真の値を報告するかもしれない。実際に米国では、ある企業の真の市場シェアを統計調査から算出させない事が過去の争点であった。センシティブな項目は、守秘性が前提で統計を作成するべきである。それから非上場企業についても、公的統計の情報は民間の企業情報データベースに（少なくとも一部は）含まれている。ここで買える「公知」の情報は、開示制限しなくて良いと主張されるかもしれない。しかし公知のセンシティブ変数は、守秘性の下で作成される統計と違う可能性がある。守秘性の下でのデータの質は、買えないのである。

データアクセスに関する社会的理解を醸成する事についても、米国の経験は参照されるべきである。Duncan et al. (1993) によれば、1989 年に米国 Committee on National Statistics 主導で守秘性とデータアクセスに関するパネルが招集された。パネルは広い範囲の専門家で構成され、守秘性を損なわないでデータアクセスを改善するための勧告作成を目的とした。重要なのは、全ての利害関係者—データの提供者と使用者及び作成者の三者が勧告の対象という事である。そして勧告内容が出来るだけ三者に理解されるように工夫され、関係する政策や実務の多くの例が提供された。このパネルの報告書はペーパーバックで出版されている (Panel on Confidentiality and Data Access, 1993)。日本でも特にデータ提供者への働きかけは工夫の余地が有るのではないだろうか。

最後にデータ作成者にむけて、当面の指針を提案する。Zayatz (2007) によると米国国勢調査の PUF では、地域以外の一部変数の値が同じ（近隣）世帯間で、地域をスワップしている。同様に地理的変数のスワッピングで経験を積むと良い。理由を説明しよう。攪乱は高度なノウハウを要求されるので、単純な手法から始めるべきである。単純な攪乱手法と言えば、スワッピングか（ミクロアグリゲーションを含む）誤差の付加となる。しかし誤差の付加は、似た個体を探すタイプの攻撃に対して脆弱な事が知られている（Brand, 2002) 。故にスワッピングを用いた方が良い。そして攪乱は、地理的変数のように識別について強力なキー変数に適用すべきである。

Jabine (1993) によれば、米国である官庁が外部者と契約を結び、PUF の識別を試してもらったそうだ。この試みは匿名化の弱点の除去につながり、有益であったという。日本でもそのような試みがあっても良いのではないか。

謝辞

渋谷政昭教授、竹村彰通教授及び匿名の査読者から、本論文の草稿について有益なコメントを頂いた。記して感謝の意を表したい。なお本研究は、科研費及び統計数理研究所の共同利用研究経費の助成を受けたものである。

参考文献

[1] Abowd, J., Stinson, M. and Benedetto, G. (2006). Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project, *Technical Report*, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program. Available at `http://www.census.gov/sipp/SSAfinal.pdf`

[2] Abowd, J.M. and Woodcock, S.D. (2001). Disclosure Limitation in Longitudinal Linked Data, in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* (Doyle et al. eds.), 215–278, Elsevier, New York.

[3] Aggarwal, C.C. and Yu, P.S. (2008). *Privacy-Preserving Data Mining: Models and Algorithms*, Springer, New York.

[4] Alexander, J.T., Davern, M. and Stevenson, B. (2010). Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications, *Working Paper No. 15703*, U.S. National Bureau of Economic Research, Massachusetts.

[5] Anderson, M.J. and Seltzer, W. (2009). Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues, *Journal of Privacy and Confidentiality*, **1**, 7–52.

[6] Bowden, R.J. and Sim, A.B. (1992). The Privacy Bootstrap. *Journal of Business and Economic Statistics*, **10**, 337–345.

[7] Brand, R. (2002). Microdata Protection through Noise Addition, in *Inference Control in Statistical Databases: From Theory to Practice, LNCS 2316* (Domingo-Ferrer ed.), 97–116, Springer, Berlin.

[8] Brandt, M., Lenz, R. and Rosemann, M. (2008). Anonymisation of Panel Enterprise Microdata – Survey of a German Project, in *Privacy in Statistical Databases, LNCS 5262* (Domingo-Ferrer et al. eds.), 139–151, Springer, Berlin.

[9] Burridge, J. (2003). Information Preserving Statistical Obfuscation, *Statistics and Computing*, **13**, 321–327.

[10] Census Bureau (2006). Census Bureau Section 515 Information Quality Guidelines, U.S. Census Bureau, Methocology and Standards Council. Available at `http://www.census.gov/qwuality`

[11] Chaloner, K. (1984). Optimal Bayesian Experimental Design for Linear Models, *Annals of Statistics*, **12**, 283–300.

[12] Cox, L.H. (1994). Matrix Masking Methods for Disclosure Limitation in Microdata, *Survey Methodology*, **20**, 165–169.

[13] Dandekar, R.A., Cohen, M. and Kirkendall, N. (2002). Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique, in *Inference Control in Statistical Databases: From Theory to Practice* (Domingo-Ferrer ed.), 117–125, Springer, Berlin.

[14] Di Consiglio, L. and Polettini, S. (2008) Use of Auxiliary Information in Risk Estimation, in *Privacy in Statistical Databases, LNCS 5262* (Domingo-Ferrer et al. eds.), 213–226, Springer, Berlin.

[15] Duncan, G.T., de Wolf, V.A., Jabine, T.B and Straf, M.L. (1993). Report of the Panel on Confidentiality and Data Access. *Journal of Official Statistics*, **9**, 271–274.

[16] Duncan, G.T., Keller-McNulty, S.A. and Stokes, S.L. (2001). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. *Technical Report*, U.S. National Institute of Statistical Sciences.

[17] Duncan, G.T. and Pearson, R.W. (1991). Enhancing Access to Microdata while Protecting Confidentiality: Prospects for the Future, *Statistical Science*, **6**, 219–239.

[18] Federal Committee on Statistical Methodology (1978). Report on Statistical Disclosure and Disclosure-Avoidance Techniques, *Statistical Policy Working Paper No. 2*, U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, Washington, D.C.

[19] Federal Committee on Statistical Methodology (1994). Report on Statistical Disclosure Limitation Methodology, *Statistical Policy Working Paper No. 22*, U.S. Office of Management and Budget, Statistical Policy Office, Washington, D.C.

[20] Federal Committee on Statistical Methodology (2005). Report on Statistical Disclosure Limitation Methodology, *Statistical Policy Working Paper No. 22 (Second version)*, U.S. Office of Management and Budget, Office of Information and Regulatory Affairs, Washington, D.C.

[21] Fienberg, S.E. (1994). A Radical Proposal for the Provision of Micro-data Samples and the Preservation of Confidentiality, *Technical Report No. 611*, Carnegie Mellon University, Pittsburgh.

[22] Fienberg, S.E. (1998). Fréchet and Bonferroni Bounds for Multi-way Tables of Counts with Applications to Disclosure Limitation. Statistical Data Protection (SDP' 98) Proceedings, IOS Press, Luxembourg.

[23] Fienberg, S.E. (2005). Confidentiality and Disclosure Limitation, in *Encyclopedia of Social Measurement* (Kempf-Leonard ed.), Vol. 1, 463–469, Elsevier, New York.

[24] Fienberg, S.E. (2009). The Relevance of Irrelevance of Weights for Confidentiality and Statistical Analyses, *Journal of Privacy and Confidentiality*, **1**, 183–195.

[25] Fienberg, S.E., Makov, U.E. and Steele, R.J. (1998). Disclosure Limitation Using Perturbation and Related Methods for Categorical Data. *Journal of Official Statistics*, **14**, 485–502.

[26] Felsö, F., Theeuwes, J. and Wagner, G.G. (2001). Disclosure Limitation Methods in Use: Results of a Survey, in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* (Doyle et al. eds.), 17–42, Elsevier, New York.

[27] Franconi, L. and Stander, J. (2002). A Model Based Method for Disclosure Limitation of Business Microdata, *Journal of the Royal Statistical Society*, D, **51**, 1–11.

[28] Gomatam, S., Karr, A.F. and Sanil, A.P. (2005). Data Swapping as a Decision Problem, *Journal of Official Statistics*, **21**, 635–655.

[29] Gottschalk, S. (2004). Microdata Disclosure by Resampling – Empirical Findings for Business Survey Data, *Allgemeines Statistisches Archiv*, **88**, 279–302.

[30] Gutmann, M.P., Witkowski, K., Colyer, C., O'Rourke, J.M. and McNally, J. (2008). Providing Spatial Data for Secondary Analysis: Issues and Current Practices Relating to Confidentiality, *Population Research and Policy Review*, **27**, 639–665.

[31] 濱砂敬郎 (2000).「事実上の匿名性の原則」,『講座ミクロ統計分析 (1) 統計調査制度とミクロ統計の開示』(松田他編), 196–224, 日本評論社.

[32] Hoshino, N. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment. *Journal of Official Statistics*, **17**, 499–520.

[33] 星野伸明 (2009).「労働力調査個票開示リスクの時間安定性について」, 統計関連学会連合大会講演.

[34] Hoshino, N. (2009). The Quasi-multinomial Distribution as a Tool for Disclosure Risk Assessment, *Journal of Official Statistics*, **25**, 269–291.

[35] Hoshino, N. and Takemura, A. (2000). On Reduction of Finite Sample Variance by Extended Latin Hypercube Sampling, *Bernoulli*, **6**, 1035–1050.

[36] Ichim, D. (2008). Extensions of the Re-identification Risk Measures Based on Log-Linear Models, in *Privacy in Statistical Databases, LNCS 5262* (Domingo-Ferrer et al. eds.), 203–212, Springer, Berlin.

[37] Jabine, T.B. (1993). Statistical Disclosure Limitation Practices of United States Statistical Agencies, *Journal of Official Statistics*, **9**, 427–454.

[38] 神林龍 (2008)「北米における政府統計個票公開の現状に関する調査報告—米国労働局，米国 Census Bureau およびカナダ統計局のオンサイトリサーチを中心に—」, 経済研究, **59**, 164–186.

[39] 加納悟 (2003).「労働力調査とローテーション・サンプリング」, 統計数理, **51**, 199–222.

[40] Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P. and Sanil, A.P. (2006). A Framework for Evaluation the Utility of Data Altered to Protect Confidentiality, *The American Statistician*, **60**, 224–232.

[41] Kinney, S.K., Karr, A.F. and Gonzalez, Jr., J.F. (2009). Data Confidentiality: The Next Five Years Summary and Guide to Papers, *Journal of Privacy and Confidentiality*, **1**, 125–134.

[42] Leitner, M. and Curtis, A. (2006). A First Step towards a Framework for Presenting the Location of Confidential Point Data on Maps — Results of an Empirical Perceptual Study, *International Journal of Geographical Information Science*, **20**, 813–822.

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

[44] Liew, C.K., Choi, U.J. and Liew, C.J. (1985). A Data Distortion by Probability Distribution, *ACM Transactions on Database Systems*, **10**, 395–411.

[45] Little, R.J.A. (1993). Statistical Analysis of Masked Data, *Journal of Official Statistics*, **9**, 407–426.

[46] 松田芳郎 (2008). 「日本におけるミクロ政府統計活用の新しい夜明け」, 統計, **59**, 2–9, 日本統計協会.

[47] 松田芳郎・濱砂敬郎・森博美 (2000). 『講座ミクロ統計分析 (1) 統計調査制度とミクロ統計の開示』, 日本評論社.

[48] 森博美 (2004). 「政府統計ミクロデータの提供とわが国統計制度の今日的課題」, 経済志林 , **72**, 33–65, 法政大学経済学会.

[49] Muralidhar, K. and Sarathy, R. (2003). A Theoretical Basis for Perturbation Methods. *Statistics and Computing*, **13**, 329–335.

[50] 日本学術会議 (2005). 「政府統計・世論調査等の一次データ（含む個票データ）の体系的保存と活用・公開方策について」, 学術基盤情報常置委員会報告（平成 17 年 9 月 15 日）.

[51] Panel on Confidentiality and Data Access (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, National Academies Press, Washington, DC.

[52] Raghunathan, T.E., Reiter, J.P. and Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics*, **19**, 1–16.

[53] Reiter, J.P. (2005). Using CART to generate partially synthetic, public use microdata, *Journal of Official Statistics*, **21**, 441–462.

[54] Reiter, J.P. (2009). Multiple Imputation for Disclosure Limitation: Future Research Challenges, *Journal of Privacy and Confidentiality*, **1**, 223–233.

[55] Reiter, J.P., Oganian, A. and Karr, A.F. (2009). Verification Servers: Enabling Analysts to Assess the Quality of Inferences from Public Use Data. *Computational Statistics and Data Analysis*, **53**, 1475–1482.

[56] Reiter, J.P. and Raghunathan, T.E. (2007). The Multiple Adaptations of Multiple Imputation, *Journal of the American Statistical Association*, **102**, 1462–1471.

[57] Rubin, D.B. (1993). Discussion: Statistical Disclosure Limitation, *Journal of Official Statistics*, **9**, 462–468.

[58] Singh, A.C. (2009). Maintaining Analytic Utility while Protecting Confidentiality of Survey and Nonsurvey Data, *Journal of Privacy and Confidentiality*, **1**, 155–182.

[59] Skinner, C. (2008). Assessing Disclosure Risk for Record Linkage, in *Privacy in Statistical Databases, LNCS 5262* (Domingo-Ferrer et al. eds.), 166–176, Springer, Berlin.

[60] Skinner, C. (2009). Statistical Disclosure Control for Survey Data, in *Sample Surveys: Design, Methods and Applications, Handbook of Statistics 29A* (Pfeffermann and Rao eds.), 381–396, Elsevier, Amsterdam.

[61] 総務省政策統括官（統計基準担当）(2008).「統計データの二次利用促進に関する研究会報告書」

[62] 総務省政策統括官（統計基準担当）(2009).「匿名データの作成・提供に係るガイドライン（平成 21 年 9 月 29 日改正版）」

[63] Spruill, N. (1982). Measures of Confidentiality, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 260–265.

[64] Spruill, N. (1983). The Confidentiality and Analytic Usefulness of Masked Business Microdata, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 602–607.

[65] Takemura, A. and Hara, H. (2007). Conditions for Swappability of Records in a Microdata Set when Some Marginals Are Fixed, *Computational Statistics*, **22**, 173–185.

[66] Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice* , Lecture Notes in Statistics 111, Springer, New York.

[67] Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control.* Lecture Notes in Statistics 155, Springer, New York.

[68] Woodcock, S.D. and Benedetto, G. (2009). Distribution-Preserving Statistical Disclosure Limitation, *Computational Statistics and Data Analysis*, **53**, 4228–4242.

[69] Wu, J. and Abowd, J. (2007). Synthetic Data for Administrative Record Applications at LEHD, *PN-2007-05*, Joint Statistical Meetings Invited Session 82 Presentation. Available at `http://lehd.did.census.gov/led/library/presentations/Wu-Abowd-20070831.pdf`

[70] 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**, 139–152.

[71] 安田聖 (2010). 私信.

[72] 美添泰人 (2008).「新統計法と統計情報の利用促進—国民の共有財産としての統計」, 統計関連学会連合大会基調講演.

[73] Young, C., Martin, D. and Skinner, C. (2009). Geographically Intelligent Disclosure Control for Flexible Aggregation of Census Data, *International Journal of Geographical Information Science*, **23**, 457–482.

[74] Zayatz, L. (2007). Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update, *Journal of Official Statistics*, **23**, 253–265.