製表技術参考資料 22

ISSN 1348－7329

家計調査ミクロデータを用いた攪乱的手法の有効性に関する研究

NSTAC

Working Paper No.22

平成 25 年 5 月

独立行政法人 統計センター

製表技術参考資料は、独立行政法人 統計センターの職員がその業務に関連して行った製表技術に関する研究の結果を紹介するためのものである。

ただし、本資料に示された見解は、執筆者の個人的見解である。

目　次

家計調査ミクロデータを用いた攪乱的手法の有効性に関する研究

伊藤伸介*・村田磨理子**

要　旨

わが国では、統計法の改正以降匿名データ(政府統計ミクロデータ)の提供が進められているが、わが国の匿名データの作成においては、これまでトップ(ボトム)・コーディングやリコーディングといった非攪乱的な匿名化技法が採用されてきた。一方、諸外国では、2000年のアメリカ人口センサスのPUMS(Public Use Microdata Sample)におけるノイズの付加やスワッピングの適用、1998～1999年のオーストラリア家計調査(Household Expenditure Survey)のCURFs(Confidentialised Unit Record Files)における所得項目への攪乱的な秘匿処理等、政府統計ミクロデータを作成するための匿名化技法の1つとして、攪乱的手法(perturbation)が用いられてきた。

わが国において匿名データの作成・提供のさらなる展開を図るためには、トップ(ボトム)・コーディングやリコーディング等の非攪乱的な匿名化技法だけでなく、ノイズの付加、ミクロアグリゲーションといった攪乱的手法についても、その適用可能性を追究することが求められる。そのためには、ミクロデータを用いて様々な攪乱的手法の有効性に関する実証研究を行う必要があるが、わが国ではそうした研究は数少ない。そこで、本稿では、ミクロデータに対して各種の匿名化技法を適用した秘匿処理済データを対象に、有用性と秘匿性の定量的な評価を行うことによって、攪乱的手法の有効性の検証を行っている。

本研究は、平成21年家計調査の個別データに含まれる量的属性に攪乱的手法を適用した秘匿処理済データに対して、有用性と秘匿性の比較分析を試みた。攪乱的手法については、先行研究で議論してきたミクロアグリゲーションだけでなく、ノイズの付加やカテゴリー化等の匿名化技法を用いている。さらに、本研究では、ミクロアグリゲーションとカテゴリー化といった匿名化技法の併用についても検証の対象に含めている。その一方で、本研究においては、相関係数の平均平方誤差等を用いて情報量損失の計測を行うだけでなく、距離計測型リンケージ(distance-based record linkage)によって、秘匿処理を施していない個別データ(原データ)と秘匿処理済データの間で真のリンクと判定されるレコードの比率を算出することによって、有用性と秘匿性の定量的な評価を行った。

本分析の結果によれば、ノイズの付加の程度が高くなるにつれて、原データに対する秘匿処理済データの情報量損失が傾向的に大きくなることが実証的に確認された。また、ミクロアグリゲーションとノイズの付加の併用等の攪乱的手法の組み合わせによっては、真のリンクの比率が相対的に小さくなることがわかった。

* (独)統計センター統計情報・技術部統計技術研究課非常勤研究員(明海大学経済学部准教授)
** (公財)統計情報研究開発センター(シンフォニカ)主任研究員(元統計センター非常勤研究員)

家計調査ミクロデータを用いた攪乱的手法の有効性に関する研究

伊藤伸介・村田磨理子

1. はじめに

政府統計ミクロデータの提供においては，個体情報の保護に関する法的制度的な措置がとられるだけでなく，ミクロデータに対して様々な匿名化技法が適用される。ミクロデータに適用可能な匿名化技法は，非攪乱的な(non-perturbative)手法と攪乱的な(perturbative)手法に大別される(Willenborg and de Waal(2001))。非攪乱的な手法については、リコーディング(global recoding, local recoding)、データの削除(record suppression, attribute suppression)、トップ(ボトム)・コーディング等が含まれる。また、攪乱的な手法には、ノイズ(加法ノイズ(additive noise)，乗法ノイズ(multiplicative noise))、スワッピング(data swapping)[1]、ラウンディング(丸め)(rounding)、ミクロアグリゲーション(micro aggregation)[2]、PRAM(Post RAndomisation Method)[3]等が存在する(Domingo-Ferrer and Torra(2001a), Willenborg and de Waal(2001), Duncan *et al.*(2011))。

わが国では、統計法の改正以降、匿名データ(政府統計ミクロデータ)の提供が進められているが、就業構造基本調査や全国消費実態調査等、わが国で現在提供されている匿名データは、これまでトップ(ボトム)・コーディング、リコーディング、データの削除といった非攪乱的手法をもとに作成されてきた。その一方で、諸外国の統計作成部局は、政府統計ミクロデータを作成するための匿名化技法の 1 つとして、攪乱的手法(perturbation)を用いていることが知られている。例えば、アメリカセンサス局は、2000 年のアメリカ人口センサスの PUMS(Public Use Microdata Sample)において、加法ノイズ、スワッピングおよびラウンディングを採用している(Zayatz(2007))[4]。また、1998～1999 年のオーストラリア家計調査(Household Expenditure

---

[1] スワッピング(data swapping)とは、ミクロデータに含まれるレコード同士で属性値を入れ替えることである(Willenborg and Waal(2001, p.126))。なお、わが国におけるスワッピングの実証研究の事例としては、例えば、Takemura(2002)による実証研究がある。

[2] ミクロアグリゲーションとは、ミクロデータ(個票データ)を k 個(k は閾値(threshold))のレコードを有する同質的なレコード群にグループ化した上で、そのレコードにおける個々の属性値を平均値等の代表値に置き換えることである(伊藤(2008))。ミクロアグリゲーションの方法的な特徴については、伊藤(2008)を参照されたい。

[3] PRAM とは、ミクロデータにおける属性値に対して、事前に設定されたマルコフ連鎖遷移行列に基づいて攪乱を行うことである。なお、PRAM の概要については、例えば藤野・垂水(2003)を参照されたい。

[4] アメリカセンサス局は、2000 年人口センサスの PUMS の作成において、識別される危険の高い世帯(ex. 世帯人員が 10 人以上の世帯)について世帯員の年齢にノイズを導入している。具体的には、原データにおける年齢の属性値を削除し、ある特定の年齢階層における年齢分布から乱数によって発生させた年齢の属性値をレコードに新たに付与し、特定の年齢階層における結果表の分布が変わらないように、ノイズを付与している。具体的には、非常に粗い地域区分において特定の人口社会的属性群に基づいて一意性を有する世帯のレコードは、露見リスクが非常に高いと考えられることから、別の地域における他の世帯との入れ替えが行われている(Zayatz(2007,

Survey)の CURFs(Confidentialised Unit Record Files)において、所得項目への攪乱的な秘匿処理が行われている(Australian Bureau of Statistics(2007))。さらに、イギリスでは、2001 年の人口センサスの SARs において、PRAM が適用されている(De Kort and Wathan(2009))。なお、イギリスでは、人口センサスの個別データの作成において、レコードスワッピングが適用されていることが知られている(Shlomo(2007))[5]。

わが国において匿名データの作成・提供のさらなる展開を図る上では、トップ(ボトム)・コーディングやリコーディング等の匿名化技法だけでなく、攪乱的手法についても、ミクロデータを用いてその適用可能性を追究することは意義があると考える。それによって、わが国の政府統計ミクロデータに対して実用的な匿名化技法の範囲が拡大すると思われる。

一方、わが国では、これまで攪乱的手法を含む匿名化技法についての有効性に関する実証研究は数少なかった。そうした中で、統計センター統計技術研究課(旧統計センター研究主幹室)では、これまで諸外国で匿名化技法として近年注目されているミクロアグリゲーションを中心に、ミクロデータに適用可能な匿名化技法の有効性に関する実証研究を行ってきた(伊藤他(2008, 2009, 2010))。具体的には、個別データに様々な匿名化技法を適用したミクロデータ(以下「秘匿処理済データ」と呼ぶ。)の秘匿性と有用性に関する定量的な評価を行うことによって、ミクロデータにおける匿名化技法の適用可能性を検証した。その一方で、加法ノイズといったミクロアグリゲーション以外の攪乱的手法についても、ミクロデータを用いてその有効性に関する実証研究を行う必要があるが、わが国ではそうした事例は数少ない。

そこで、本稿では、攪乱的手法の適用可能性を検討する試みの１つとして、ノイズを中心に攪乱的手法が適用された秘匿処理済データに対して有用性と秘匿性の定量的な評価を行うことによって、攪乱的手法の有効性の検証を試みることにしたい。

2. ミクロデータにおける攪乱的手法の適用について―加法ノイズを例に―

ミクロデータに対する匿名化技法としての攪乱的手法に関する議論は、少なくとも 1970 年代に遡ることができ、加法ノイズ(Federal Committee on Statistical Methodology (1978))やスワッ

p.257)。

[5] アメリカでも、2000 年人口センサスの集計表における秘匿処理として、人口センサスの個別データにスワッピングを適用していることが知られている。スワッピングは、2000 年人口センサスにおける short form と long form の 2 種類の調査票情報に適用される。さらには、American Community Survey にもスワッピングが使用されている。具体的には、異なる地域に居住する世帯の組に対して、地域間におけるスワッピングが適用されているが、スワッピングの対象となる世帯の組については、最低限の人口社会的な属性に基づいた対応付けが行われている。なお、スワッピングされた個別データから PUMS および集計表が作成されている(Zayatz(2007, p.257)。

ピング(Dalenius and Reiss(1978))の可能性が議論されてきた。1980 年代には、ミクロアグリゲーション(ブラーリング)の方法的な有効性に関する研究が行われた(Strudler *et al.*(1986))。さらに、PRAM については、1990 年代後半に、Gouweleeuw 等が PRAM の理論的な特徴とその適用事例を紹介している(Gouweleeuw *et al.*(1998))。本節では、攪乱的手法の特徴を明らかにするために、加法ノイズに焦点を当てて述べることにしたい。

秘匿処理が施されていない個別データ(以下「原データ」と呼ぶ。)に、攪乱的な匿名化技法を適用することによって作成された秘匿処理済データは、つぎの(1)式のように行列で表示することが可能である(Duncan and Peason(1991), Domingo-Ferrer and Torra (2001a),Duncan *et al.*(2011) 等)。

$$\mathbf{V}' = \mathbf{AVB} + \mathbf{C} \quad \cdots (1)$$

ここで、

V・・・原データの行列

V′・・・秘匿処理済データの行列

A・・・レコードの変換に伴う秘匿処理(に関する行列)

B・・・変数値の変換に伴う秘匿処理(に関する行列)

C・・・攪乱的手法による変数値の置換(に関する行列)

データの削除(ex. record suppression)のような方法は、(1)式では、行列**A**における秘匿処理に該当すると考えられる。また、トップ(ボトム)・コーディングやリコーディングといった非攪乱的手法による変数値の変換は、行列**B**に含まれる。そして、ノイズ等の攪乱的手法の適用は、(1)式における行列**C**の分布構造に影響を与える。

つぎに、加法ノイズを例に行列**C**の分布構造を考えてみたい(Kim(1986), Domingo-Ferrer and Torra (2001a),Duncan *et al.*(2011) 等))[6]。加法ノイズでは、ランダムなノイズを発生させた上で、原データの属性値にノイズを付加することによって、秘匿処理が施される。したがって，加法のノイズはつぎのように定式化される(Domingo-Ferrer and Torra (2001a, p.94), Duncan *et al.*(2011, pp.112-114))。

原データにおいてｎ個の個体レコードがそれぞれｐ個の属性を持つ場合､原データの行列**V**は、ｎ×ｐ 個の属性値から構成される。おのおのの個体レコードが独立同一分布(independently and identically distributed)に従うとすると、

[6] 匿名化技法として加法ノイズを適用した場合の分布特性については、Kim(1986)や Kim and Winkler(1995)が詳しい。

$\mathbf{V} \sim (\boldsymbol{\mu}, \sum)$・・・(2)

と書くことができる。ここで、$\boldsymbol{\mu}$は、属性値群における平均値のベクトル、$\sum$は属性値群における分散共分散行列である。原データにノイズが適用されると、秘匿処理済データの行列$\mathbf{V}'$は、原データの行列$\mathbf{V}$とn×pのノイズの行列$\boldsymbol{\varepsilon}$の合計で示され、つぎの(3)式となる。

$\mathbf{V}' = \mathbf{V} + \boldsymbol{\varepsilon}$・・・(3)

ここで、ノイズの行列$\boldsymbol{\varepsilon}$は、

$\boldsymbol{\varepsilon} \sim (\mathbf{0}, c\sum)$(cはパラメータ)・・・(4)

である。各属性にランダムにノイズを発生させることから、各レコードに付与されるノイズの間に相関関係はないと考える。この場合，秘匿処理済データと原データとの間に、以下のような関係があることがわかっている。

$E(\mathbf{V}') = E(\mathbf{V}) + E(\boldsymbol{\varepsilon}) = E(\mathbf{V})$・・・(5)

$Var(\mathbf{V}') = Var(\mathbf{V}) + Var(\boldsymbol{\varepsilon}) = (1+c)Var(\mathbf{V})$・・・(6)

(5)式と(6)式から明らかなように、加法ノイズを適用した場合、秘匿処理済データにおける属性値のベクトルの期待値は、原データにおいて対応する属性値のベクトルの期待値と一致するが、(6)式を見ると、原データにおける属性値のベクトルの分散は、秘匿処理済データにおけるそれとは一致しないことから[7]、分散に関しては秘匿処理済データには原データに対するバイアスが生じる[8]。このことは、秘匿処理済データにおける相関係数や回帰係数においても原データからのバイアスが発生することを意味する(Matloff(1986))。このことから、攪乱的手法として加法ノイズを用いる場合には、秘匿処理済データの分布に生じるバイアスを考慮した上で、有用性の観点からノイズのパラメータcを設定する必要があると思われる。

3. ミクロデータの秘匿性と有用性に関する定量的な評価方法

ミクロデータに対する匿名化技法の適用可能性を検証するために、諸外国では、秘匿処理済データを用いた有用性と秘匿性の定量的な評価に関する実証研究が行われてきた。(Domigo-Ferrer

[7] (6)式を変形すると、

$Var(\mathbf{V}) = \frac{Var(\mathbf{V}')}{1+c}$・・・(6)'

となる。このことから、パラメータcが既知の場合、秘匿処理済データにおける分散を原データの分散に置き換えることは可能である。

[8] 秘匿処理済データの利用者にとっては(6)'式におけるパラメータcは未知であることから、秘匿処理済データの利用者が(6)'式のような置換によって、原データにおける属性値のベクトルの分散を算出するのは困難だと思われる。

and Torra(2001b), Yancey *et al.*(2002), Karr *et al.*(2006), Shlomo(2010), 伊藤他(2010)等)。

ミクロデータの有用性の評価方法については、様々な手法が提案されている。主な手法は、次の３つである。第１は、平均や分散等の基本統計量やクロス集計表における度数を直接比較することである(Domigo-Ferrer and Torra(2001b)。また、原データと秘匿処理済データの両方で集計表を作成した上で、セルごとの度数の差の絶対値に関する平均値(平均絶対距離(average absolute distance))を求めることも考えられる(Shlomo *et al.*(2010))。第２は、属性間の関連性の比較という観点から、セルに含まれる度数の変化率(伊藤他(2010))やクラーメルの V といった関連性の指標を計測し(Shlomo(2010))、原データと秘匿処理済データの近似性の比較を行うことである。第３は、秘匿処理済データの原データに対する情報量損失(information loss)に関する指標を計測することである。情報量損失については、量的属性と質的属性では異なる指標が提案されている。量的属性に関しては、Domigo-Ferrer 等が、属性値、相関係数行列、分散共分散行列等をもとに、平均平方誤差（Mean square error)、平均絶対誤差（Mean abs. error)、平均変化率（Mean variation）を用いて情報量損失を計測することを提唱している(Domigo-Ferrer and Torra(2001a))。それに対して、De Waal 等は、質的属性に関してもエントロピーに基づいて情報量損失(information loss)を計測する方法を議論している(De Waal and Willenborg(1999))。なお、それ以外の方法としては、有用性の評価方法として、回帰分析を行い、決定係数の比較や回帰係数の信頼区間を比較する方法(Karr *et al.*(2006))、さらには、傾向スコア、クラスター分析、経験分布関数等を用いて有用性を定量的に評価する方法が考案されている(Woo *et al.*(2009, pp.113-115))。

その一方で、秘匿性の評価に関しても、様々な手法が提案されてきた。第１は、外部情報とミクロデータのマッチングである。これについては、例えばドイツで事実上の匿名性を検証するために行われてきた、ミクロセンサスの個票データと研究者情報とのマッチングに関する研究を指摘することができる(Müller *et al.*(1995, p.135))。第２の方法は、ミクロデータにおいて母集団一意に関する指標を計測することである。例えば、イギリスでは、想定される様々なシナリオに基づいてキー変数を設定した上で、母集団一意の計測が行われている。具体的には，1991 年人口センサスの SARs の作成に関する露見リスクの研究において、キー変数を用いた母集団一意の検証が行われているだけでなく(Marsh *et al.*(1991))，2001 年の SARs では、母集団一意となるレコードの比率が，露見リスクに関する主要な指標として用いられた(Gross *et al.*(2004))。第３の方法は、原データと秘匿処理済データとのレコードリンケージを行うことである(Duncan *et al.*(2011), 伊藤(2010, 7～8 頁))。原データと秘匿処理済データとのレコードリンケージの方法に

ついては，確定的リンケージ(deterministic record linkage)，距離計測型リンケージ(distance-based record linkage)，および確率的リンケージ(probabilistic record linkage)が考えられる。確定的リンケージは、対応付けを行うためのキーとなる属性群（以下「リンクキー変数」と呼称）を用いて、原データと秘匿処理済データに含まれるそれぞれのレコード同士が１対１で照合するかどうかを判定する方法である(伊藤(2010，8～9 頁))。また、距離計測型リンケージでは、原データと秘匿処理済データにおけるレコード同士の距離を計算し、その距離の大きさに基づいて、2 つのデータが対応付け可能かを判定する(伊藤(2010，9～10 頁))。それに対して、確率的リンケージとは、原データと秘匿処理済データの全てのレコードの組み合わせ（ペア）を考え、各ペアがリンクされる集合又はリンクされない集合のどちらに属するかを、属性値の一致基準及び確率値にしたがって分類する方法である(伊藤他(2010，33～38 頁))。これら３つの方法に加えて、露見リスクの指標として「特殊な一意(Special Unique)」を探索する特殊な一意の分析(Special Uniques Analysis)(Elliot *et al.*(2002))が提案されている。特殊な一意の分析については、外観識別性の高い属性をもとに次元の低いクロス集計表を作成した上で、度数１のセル数の減少率を計測すること(Shlomo *et al.*(2010))も分析方法の１つとして考えられる。

さらに、近年では、様々な匿名化技法を適用した各種の秘匿処理済データにおける有用性と秘匿性の比較分析が行われている。有用性と秘匿性の相対比較を行う主な方法としては、①総合指標の計算、②R-Uマップの作成がある。総合指標を用いた有用性と秘匿性の相対比較については、例えば、Domingo-Ferrer 等の研究が知られている(Domigo-Ferrer and Torra(2001b))。Domingo-Ferrer 等は、主としてミクロアグリゲーション、加法ノイズ、スワッピング等の匿名化技法を用いて作成した秘匿処理済データを対象に、相関係数の平均平方誤差等を用いて情報量損失の計測を行うだけでなく、レコードリンケージ(record linkage)によって、原データと秘匿処理済データの間で真のリンクとなる比率を算定している。このような有用性と秘匿性に関する指標に基づいて総合指標を計算し、その総合指標をもとに匿名化技法の有効性に関する相対評価を行っている。具体的には、(7)式に基づいて、情報量損失と 3 種類のリスクに関する指標を用いてスコアを計算している。 なお、スコアの計算においては、情報量損失とリスクに関する指標に対してそれぞれ 0.5 のウェイトが付与されている。

$$Score = 0.5(IL) + 0.125(DLD) + 0.125(PLD) + 0.25(ID) \quad \cdots (7)$$

ここで

IL:情報量損失に関する指標[9]

DLD:原データと秘匿処理済データにおける距離計測型リンケージによる真のリンクの比率

PLD:原データと秘匿処理済データにおける確率的リンケージによる真のリンクの比率

ID:原データと秘匿処理済データにおける区間設定を考慮した場合の真のリンクの比率[10]

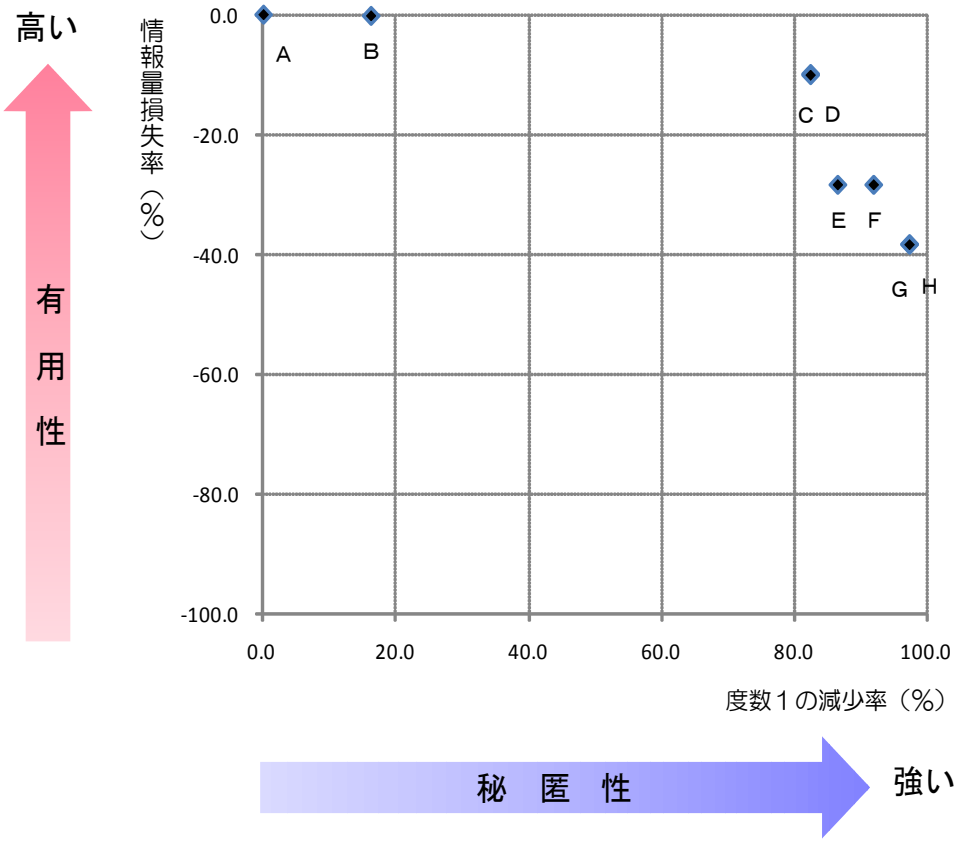
一方、Duncan 等は、R-U マップ(R-U confidentiality map)の作成を提唱している(Duncan *et al.* (2001))。R-U マップでは、有用性と秘匿性に関するそれぞれ指標を同時に図示することによって、各種匿名化技法の相対評価を視覚的に行うことを指向している。

ところで、伊藤他(2010)は、全国消費実態調査の個別データを用いて有用性と秘匿性の指標を計測し、それらの指標をもとに R-U マップの作成を試みた(図 1)。図 1 に示される R-U マップの縦軸は情報量損失率(%)を、横軸は度数 1 の減少率(%)を、それぞれ表している。また、〔A〕は原データであるが、〔B〕から〔H〕は、年齢ないしは職業に対してリコーディングおよびトップコーディングが適用された秘匿処理済データの組み合わせを示している。匿名化技法の適用によって、原データと比べて有用性が相対的に低くなり、秘匿性は相対的に高くなることが考えられる。このような有用性と秘匿性の指標を R-U マップで表示することによって、有用性と秘匿性がトレードオフの関係にあることが視覚的に把握できる。さらに、R-U マップを用いることによって、匿名化技法を変更する場合（例えば、世帯主の年齢に対してトップコーディングを適用する場合の閾値の変更、年齢全体ではなく年齢の一部を対象にしたリコーディングの適用、世帯主の職業における統合区分の変更等）、R-U マップにおける位置を確認した上で、有用性と秘匿性の相対比較を行うことが可能になる。このような比較分析は、様々な匿名化技法の有効性を検討する上で参考になると思われる。

[9] Domigo-Ferrer and Torra(2001b)は、情報量損失に関する指標として原データと秘匿処理済データにおける相関係数行列に関する平均変化率を含む 5 つの指標の平均値を用いている。

[10] 原データと秘匿処理済データにおけるレコード上の属性値が完全に一致しない場合でも、２つの属性値における近似の程度を判定することによって秘匿性を評価する手法が考えられる。その場合、具体的には、秘匿処理済データにおいてレコードの属性値を中心とした一定の区間(interval）を設定し、原データにおいて対応するレコードの属性値が、設定した区間の範囲内に存在するかどうかを確認する(Domingo-Ferrer and Torra(2001b、p.116)、伊藤他(2010))。

図１　ミクロデータにおける有用性と秘匿性の総合的な評価方法　R-U マップの例



	年齢		職業
	リコーディング（５歳階級）	トップコーディング（８５歳以上）	リコーディング（４区分）
[A]			
[B]		○	
[C]			○
[D]		○	○
[E]	○		
[F]	○	○	
[G]	○		○
[H]	○	○	○

出所　伊藤他(2010，26 頁)

4. ミクロデータにおける攪乱的手法の有効性に関する研究―家計調査ミクロデータを用いて―[11]

本節では、政府統計の個別データに対して攪乱的な手法を適用した場合のミクロデータの有用性と秘匿性の比較分析に関する実験成果を紹介することにしたい。具体的には、本実験では、ミクロアグリゲーション、加法ノイズ等を用いて作成した様々な秘匿処理済データを用いて、有用性と秘匿性の評価を試みる。

4-1 家計調査ミクロデータにおける有用性と秘匿性の評価

本研究で使用したデータは、家計調査の個別データである(平成 21 年(2009 年)1 月、勤労者世

[11] 本節は、Ito and Murata(2011)に加筆・修正を行ったものである。

帯 4,220 世帯)。ミクロデータに適用した匿名化技法は、ミクロアグリゲーション、加法ノイズ、量的属性のカテゴリー化である。さらに、ミクロアグリゲーションとカテゴリー化の併用といった 2 種類の匿名化技法に関する実験も行っている。

攪乱的手法が適用される量的属性は、勤め先収入、消費支出、年間収入、貯蓄現在高、負債現在高および延べ床面積である。攪乱的手法を適用する上では、基本的には、質的属性を用いてレコードの層化が行われる。具体的には、住居の所有関係を 5 区分(①持ち家(一戸建)、②持ち家(共同住宅、長屋・その他)、③民営の賃貸住宅・借間、④公営の賃貸住宅・都市再生機構・公社等の賃貸住宅、⑤給与住宅)にリコーディングした上で、層内のレコードに含まれる量的属性に対して攪乱的手法を適用する。

本実験における攪乱的手法の概要は次のとおりである。

ミクロアグリゲーションでは、個別ランキング法と Z スコア総計法の 2 種類の方法を用いて、実験を行った。Z スコア総計法については、住居の所有関係を用いてレコードの層化を行った上で、層内の量的属性値に Z スコア総計法を適用した場合とレコードの層化を行わずにレコード全体で Z スコア総計法を適用した場合の 2 つの実験を行った。さらに、Z スコア総計法を適用した上でノイズを付加する方法も用いている。具体的には、最初に、レコード層化を行わずに Z スコア法を適用し、つぎに、3 レコードずつのグループに分けた上でグループ内の量的属性値を平均値に置き換え、各グループ内で平均値から標準偏差の p 倍(p は 0.1、0.5 及び 1 のいずれかの値をとる)のノイズを控除したレコード、平均値に標準偏差の p 倍を加算したレコードと平均値の 3 つのレコードを作成した。加法ノイズに関しては、各量的属性において、平均が 0、標準偏差が原データの標準偏差の p 倍の正規分布に従うノイズを属性値に付加する（共分散は考慮していない)。パラメータ p の値については、0.01 から 0.5 までの値を設定した。なお、原データの数値が 0 になっている場合にはノイズは付与されない。ノイズを付与した結果、延べ床面積の数値が 20 未満になった場合には、その値は 20 と設定されている。同様に、他の属性についても値が 0 未満である場合、属性値は 0 と設定される。カテゴリー化については、十分位と二十分位の 2 種類について実験を行った。カテゴリー化の対象となる量的属性の値はそのカテゴリー内に含まれる属性値の平均値に置き換えられた。

一方、2 種類の匿名化技法の併用については、①個別ランキング法と Z スコア総計法[12]、②ミ

[12] Z スコア総計法は、各レコードにおける属性値群を標準化し、標準化された値の総計値に基づいてレコード群をソートし、レコードのグループ化を行う手法である。また、個別ランキング法は、量的属性のおのおのについて個別にソート化とグループ化を行う方法である(伊藤(2008，8～10 頁))。

クロアグリゲーションとカテゴリー化[13]、および③加法ノイズとカテゴリー化を試みた。①の場合、個別ランキング法とZスコア総計法の併用については、勤め先収入及び消費支出に対しては個別ランキング法を適用し、それ以外の年間収入、貯蓄現在高、負債現在高および延べ床面積についてはZスコア総計法を用いている。また、②については、勤め先収入及び消費支出に対してはミクロアグリゲーション（個別ランキング法等）を適用し、それ以外の属性についてはカテゴリー化（十分位 or 五分位）を用いる。さらに、③の場合、勤め先収入及び消費支出に対しては加法ノイズを適用し(p=0.50 等)、それ以外の量的属性についてはカテゴリー化（十分位）を行っている。

つぎに、有用性の評価方法については、原データと秘匿処理済データにおける相関係数行列をもとに、平均平方誤差、平均絶対誤差及び平均変化率をそれぞれ算出し、原データからの情報量損失を計測した。個別ランキング法について情報量損失を計算する場合、乗率については、各変数の秘匿処理後の乗率の単純平均を使った。その一方で、個別ランキング法とZスコア総計法の併用、及び個別ランキング法とカテゴリー化の併用の場合は、原データにおける乗率を用いて指標を計算した。

他方、秘匿性の評価方法に関しては、距離計測型リンケージによって真のリンクになるかどうかに関する評価を行っている。真のリンクとなる条件は、秘匿処理済データと原データをマッチングして、1対1に照合され、かつ、同一世帯番号となった場合に限定されている。

使用したリンクキー変数は､秘匿処理の対象となった6つの量的属性(勤め先収入等)に加えて、住居の所有関係(レコードの層化に使用)、世帯人員数、就業人員数と世帯主年齢である。なお、世帯主年齢については、原データでは各歳、秘匿処理済データでは5歳階級となっている。追加した3つのリンクキー変数と6つの量的属性を用いて、質的属性の層内で標準化ユークリッド距離を計測している。なお、世帯主年齢 5 歳階級については、階級中央値を使った。ただし、85歳以上は 92 歳に設定されている。

表1は、様々な攪乱的手法を用いて作成した秘匿処理済データにおける有用性の評価の結果を示したものである。表1では、標準化された属性値、相関係数行列及び標準化されていない分散共分散行列をもとに計算した平均平方誤差、絶対平方誤差と平均変化率が示されている。ミクロアグリゲーションについては、個別ランキング法のほうが、Zスコア総計法と比較して、原デー

[13] 本実験においてカテゴリー化を併用する場合には、カテゴリー内の平均値で置き換えている場合(例えば十分位では､「カテゴリー化(十分位)」と表示している)とそれぞれのカテゴリーに該当するランク(例えば十分位では、「カテゴリー化(十分位ランク)」と表示している)に置き換えている場合があることに留意されたい。

表１　攪乱的手法を適用した場合のミクロデータの有用性の結果

(1)標準化された属性値

有用性評価の指標		属性値（標準化済み）		
		平均平方誤差	平均絶対誤差	平均変化率
ミクロアグリゲーション	個別ランキング法	0.015208	0.013736	0.115500
	Ｚスコア総計法（層別）	0.574780	0.449410	25.538900
	Ｚスコア総計法	0.545510	0.441670	19.074600
	Ｚスコア総計法にノイズ付加（p=0.10）	0.546350	0.441450	17.738600
	Ｚスコア総計法にノイズ付加（p=0.50）	0.586940	0.445880	12.262200
	Ｚスコア総計法にノイズ付加（p=1）	0.689530	0.467770	11.271200
	個別ランキング２変数＋Ｚスコア総計法４変数	0.380570	0.304640	11.053000
加法ノイズ	p=0.01	0.000081	0.006097	0.314180
	p=0.02	0.000317	0.012608	0.628740
	p=0.04	0.001253	0.025447	1.346200
	p=0.05	0.001951	0.031796	1.660500
	p=0.06	0.002801	0.038148	1.972830
	p=0.08	0.004960	0.050808	2.691180
	p=0.10	0.007716	0.063400	3.313560
	p=0.12	0.011049	0.075867	3.993990
	p=0.14	0.014959	0.088274	4.640940
	p=0.16	0.019439	0.100660	5.315120
	p=0.18	0.024445	0.112900	5.949450
	p=0.20	0.029969	0.125050	6.614200
	p=0.25	0.045946	0.154910	8.181260
	p=0.30	0.064728	0.184010	9.700290
	p=0.35	0.086041	0.212250	11.232600
	p=0.40	0.109510	0.239570	12.666400
	p=0.45	0.134730	0.265840	14.015500
	p=0.50	0.161460	0.291060	15.295400
カテゴリー化	十分位	0.180430	0.160940	4.993160
	二十分位	0.119560	0.100180	1.849110
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（十分位）４変数		0.097787	0.100490	3.800780
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（十分位ランク）４変数		0.354600	0.329840	9.323420
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（五分位）４変数		0.164320	0.166560	2.369610
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（五分位ランク）４変数		0.378470	0.347060	4.151530
ミクロアグリゲーション（Ｚスコア）２変数＋カテゴリー化（十分位）４変数		0.183480	0.195100	6.276770
ミクロアグリゲーション（Ｚスコア）２変数＋カテゴリー化（五分位）４変数		0.250010	0.261170	4.845600
加法ノイズ（p=0.10）２変数＋カテゴリー化（十分位）４変数		0.092820	0.120790	4.286690
加法ノイズ（p=0.16）２変数＋カテゴリー化（十分位）４変数		0.097553	0.135580	4.583980
加法ノイズ（p=0.30）２変数＋カテゴリー化（十分位）４変数		0.115790	0.168600	5.259470
加法ノイズ（p=0.50）２変数＋カテゴリー化（十分位）４変数		0.154090	0.210320	6.159300

(2)相関係数行列

有用性評価の指標		相関係数行列		
		平均平方誤差	平均絶対誤差	平均変化率
ミクロアグリゲーション	個別ランキング法	0.000039	0.004124	0.020757
	Ｚスコア総計法（層別）	0.024383	0.125300	0.735740
	Ｚスコア総計法	0.025357	0.127050	0.736120
	Ｚスコア総計法にノイズ付加（p=0.10）	0.025538	0.127540	0.740470
	Ｚスコア総計法にノイズ付加（p=0.50）	0.030064	0.138940	0.840750
	Ｚスコア総計法にノイズ付加（p=1）	0.043124	0.165850	1.068240
	個別ランキング２変数＋Ｚスコア総計法４変数	0.013403	0.075929	0.543650
加法ノイズ	p=0.01	0.000000	0.000139	0.000708
	p=0.02	0.000000	0.000268	0.001343
	p=0.04	0.000001	0.000570	0.002699
	p=0.05	0.000002	0.000754	0.003517
	p=0.06	0.000003	0.000968	0.004446
	p=0.08	0.000006	0.001505	0.006690
	p=0.10	0.000012	0.002156	0.009386
	p=0.12	0.000021	0.002915	0.012499
	p=0.14	0.000034	0.003804	0.016183
	p=0.16	0.000053	0.004792	0.020264
	p=0.18	0.000078	0.005865	0.024646
	p=0.20	0.000110	0.007031	0.029471
	p=0.25	0.000233	0.010346	0.043048
	p=0.30	0.000432	0.014175	0.058772
	p=0.35	0.000723	0.018362	0.075714
	p=0.40	0.001120	0.022841	0.093847
	p=0.45	0.001631	0.027516	0.112470
	p=0.50	0.002264	0.032366	0.131800
カテゴリー化	十分位	0.002139	0.026404	0.107590
	二十分位	0.001198	0.020152	0.079517
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（十分位）４変数		0.000078	0.006402	0.039800
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（十分位ランク）４変数		0.002255	0.028588	0.210830
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（五分位）４変数		0.000201	0.009164	0.047305
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（五分位ランク）４変数		0.002394	0.029997	0.203430
ミクロアグリゲーション（Ｚスコア）２変数＋カテゴリー化（十分位）４変数		0.007535	0.033531	0.124640
ミクロアグリゲーション（Ｚスコア）２変数＋カテゴリー化（五分位）４変数		0.007471	0.034306	0.123610
加法ノイズ（p=0.10）２変数＋カテゴリー化（十分位）４変数		0.000078	0.006082	0.034988
加法ノイズ（p=0.16）２変数＋カテゴリー化（十分位）４変数		0.000088	0.006478	0.035931
加法ノイズ（p=0.30）２変数＋カテゴリー化（十分位）４変数		0.000186	0.008544	0.040967
加法ノイズ（p=0.50）２変数＋カテゴリー化（十分位）４変数		0.000690	0.016080	0.066631

(3)標準化されていない分散共分散行列

有用性評価の指標		分散共分散行列（標準化なし）		
		平均平方誤差	平均絶対誤差	平均変化率
ミクロアグリゲーション	個別ランキング法	7.82E+16	102,971,480	0.044867
	Ｚスコア総計法（層別）	2.67E+19	1,755,520,500	0.429140
	Ｚスコア総計法	2.75E+19	1,794,394,628	0.435970
	Ｚスコア総計法にノイズ付加（p=0.10）	2.72E+19	1,787,571,119	0.441070
	Ｚスコア総計法にノイズ付加（p=0.50）	2.05E+19	1,624,594,498	0.568050
	Ｚスコア総計法にノイズ付加（p=1）	9.60E+18	1,126,809,226	0.959780
	個別ランキング２変数＋Ｚスコア総計法４変	4.24E+16	67,428,551	0.389410
加法ノイズ	p=0.01	6.61E+12	883,375	0.000844
	p=0.02	4.66E+13	2,401,143	0.001640
	p=0.04	6.21E+14	8,950,577	0.002912
	p=0.05	1.52E+15	13,779,644	0.003744
	p=0.06	3.19E+15	19,644,230	0.004592
	p=0.08	1.03E+16	34,485,210	0.006410
	p=0.10	2.57E+16	53,461,541	0.008426
	p=0.12	5.39E+16	76,582,043	0.010644
	p=0.14	1.01E+17	103,838,133	0.013339
	p=0.16	1.74E+17	135,216,898	0.016093
	p=0.18	2.79E+17	170,460,478	0.019044
	p=0.20	4.25E+17	209,431,065	0.022421
	p=0.25	1.03E+18	323,023,618	0.031615
	p=0.30	2.09E+18	456,308,541	0.041808
	p=0.35	3.79E+18	610,868,187	0.053063
	p=0.40	6.30E+18	783,964,151	0.065697
	p=0.45	9.77E+18	972,204,138	0.078840
	p=0.50	1.45E+19	1,179,187,222	0.093120
カテゴリー化	十分位	2.03E+19	1,369,960,268	0.108720
	二十分位	1.08E+19	1,022,775,624	0.080608
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（十分位）４変数		4.24E+16	65,469,730	0.086390
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（十分位ランク）４変数		4.30E+16	77,659,871	0.853080
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（五分位）４変数		4.24E+16	66,413,399	0.150070
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（五分位ランク）４変数		4.30E+16	77,687,600	0.855820
ミクロアグリゲーション（Ｚスコア）２変数＋カテゴリー化（十分位）４変数		1.83E+19	1,615,766,577	0.127840
ミクロアグリゲーション（Ｚスコア）２変数＋カテゴリー化（五分位）４変数		1.83E+19	1,616,576,162	0.185850
加法ノイズ（p=0.10）２変数＋カテゴリー化（十分位）４変数		2.73E+16	58,478,350	0.083285
加法ノイズ（p=0.16）２変数＋カテゴリー化（十分位）４変数		1.81E+17	143,572,152	0.085098
加法ノイズ（p=0.30）２変数＋カテゴリー化（十分位）４変数		2.12E+18	468,371,877	0.091801
加法ノイズ（p=0.50）２変数＋カテゴリー化（十分位）４変数		1.46E+19	1,194,425,193	0.106660

タに近似することがわかる。また、ノイズ付加の場合、パラメータ p の値が大きくなるにしたがって、情報量損失が大きくなることが確認できる。カテゴリー化に関しては、二十分位における情報量損失が十分位におけるそれと比較してより小さくなることが明らかになっている。一方、相関係数行列の平均平方誤差を見ると、カテゴリー化(十分位)とノイズ付加(p=0.50)の値はほぼ等しい。このことは、有用性に関する評価指標を用いることによって、様々な匿名化技法における相対比較が可能なことを意味している。さらに、２種類の匿名化技法を併用した場合の標準化された属性値や相関係数行列に関する有用性の評価指標を見ると、年間収入等の属性については、カテゴリー化を適用した場合、Z スコア総計法と比較して、情報量損失が小さくなっていることがわかる。

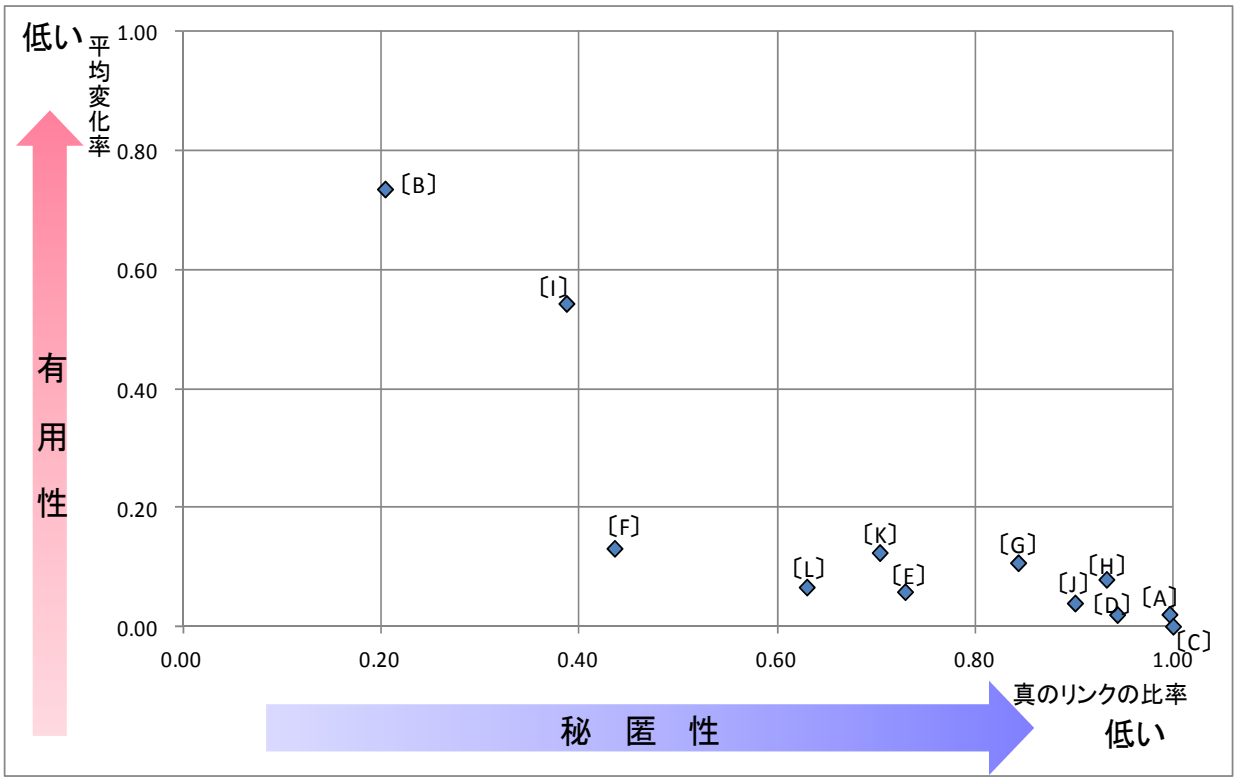
表２は、秘匿処理済データにおける秘匿性の評価の結果を示したものである。表２では、真のリンクと判定されたレコード数及び比率、1 対 1 の誤リンク、n 対 m(n 対 1、1 対 n を含む)のリンクが示されている。表２を見ると、Z スコア総計法の真のリンクの比率が最も小さく、個別ランキング法と Z スコア総計法の併用、ノイズ付加(p=0.50)における比率がそれに続いていることがわかる。また、加法ノイズについては、パラメータ p の値が大きくなるにつれて、真のリンクの比率が小さくなるだけでなく、1 対 1 の誤リンクや n 対 m のリンクの数が増大していることがわかる。一方、カテゴリー化における真のリンクの比率が非常に高いことは興味深い結果であると考える。

つぎに、図 2 は、表 1 と表 2 で示された有用性と秘匿性に関する評価結果の一部の数値をもとに R-U マップを図示したものである。本図においては、有用性については平均変化率を、秘匿性については真のリンクの比率をそれぞれ用いている。図 2 を見ると、ノイズやミクロアグリゲーションといった様々な攪乱的手法を適用した秘匿処理済データにおいて、有用性と秘匿性の間のトレードオフの関係が確認できる。図２では、加法ノイズ(p=0.01)(〔C〕)の場合、有用性は最も高いが、秘匿性は非常に低くなっている。それに対して、Z スコア総計法(〔B〕)では、有用性が相対的に低くなっているが、真のリンクの比率が最も低く、秘匿性が相対的に高いことがわかる。こうした有用性あるいは秘匿性の評価結果に対して、許容可能な閾値を設定することができれば、R-U マップのような指標をもとに、有用性と秘匿性の両面から望ましい匿名化技法を選択することが可能になるだろう。

表２　攪乱的手法を適用した場合のミクロデータの秘匿性の結果

		真のリンク		１対１の誤リンク	ｎ対ｍ
ミクロアグリゲーション	個別ランキング法	4,203	100%	0	17
	Ｚスコア総計法（層別）	827	20%	645	2,748
	Ｚスコア総計法	860	20%	611	2,749
	Ｚスコア総計法にノイズ付加（p=0.10）	853	20%	614	2,753
	Ｚスコア総計法にノイズ付加（p=0.50）	726	17%	720	2,774
	Ｚスコア総計法にノイズ付加（p=1）	566	13%	863	2,791
	個別ランキング２変数＋Ｚスコア総計法４変数	1,633	39%	435	2,152
加法ノイズ	p=0.01	4,218	100%	0	2
	p=0.02	4,216	100%	0	4
	p=0.04	4,216	100%	0	4
	p=0.05	4,214	100%	0	6
	p=0.06	4,208	100%	0	12
	p=0.08	4,194	99%	2	24
	p=0.10	4,165	99%	1	54
	p=0.12	4,131	98%	4	85
	p=0.14	4,080	97%	7	133
	p=0.16	3,980	94%	15	225
	p=0.18	3,863	92%	24	333
	p=0.20	3,748	89%	39	433
	p=0.25	3,419	81%	123	678
	p=0.30	3,076	73%	199	945
	p=0.35	2,702	64%	299	1,219
	p=0.40	2,361	56%	380	1,479
	p=0.45	2,077	49%	473	1,670
	p=0.50	1,838	44%	556	1,826
カテゴリー化	十分位	3,558	84%	6	656
	二十分位	3,934	93%	1	285
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（十分位）４変数		3,800	90%	2	418
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（十分位ランク）４変数		2,325	55%	139	1,756
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（五分位）４変数		3,098	73%	15	1,107
ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（五分位ランク）４変数		2,094	50%	157	1,969
ミクロアグリゲーション（Ｚスコア）２変数＋カテゴリー化（十分位）４変数		2,968	70%	68	1,184
ミクロアグリゲーション（Ｚスコア）２変数＋カテゴリー化（五分位）４変数		2,317	55%	119	1,784
加法ノイズ（p=0.10）２変数＋カテゴリー化（十分位）４変数		3,780	90%	4	436
加法ノイズ（p=0.16）２変数＋カテゴリー化（十分位）４変数		3,695	88%	6	519
加法ノイズ（p=0.30）２変数＋カテゴリー化（十分位）４変数		3,302	78%	56	862
加法ノイズ（p=0.50）２変数＋カテゴリー化（十分位）４変数		2,657	63%	206	1,357

図２ 家計調査の秘匿処理済データにおける有用性と秘匿性に関する R-U マップ



注１　本稿の表１と表２に基づいて作成した。なお、有用性については相関係数行列の平均変化率を用いている。

注２ 本図における匿名化技法の一覧

ミクロアグリゲーション	〔A〕個別ランキング法
	〔B〕Ｚスコア総計法
加法ノイズ	〔C〕p=0.01
	〔D〕p=0.16
	〔E〕p=0.30
	〔F〕p=0.50
カテゴリー化	〔G〕十分位
	〔H〕二十分位
〔I〕個別ランキング２変数＋Ｚスコア総計法４変数	
〔J〕ミクロアグリゲーション（個別ランキング）２変数＋カテゴリー化（十分位）４変数	
〔K〕ミクロアグリゲーション（Ｚスコア）２変数＋カテゴリー化（十分位）４変数	
〔L〕加法ノイズ（p=0.50）２変数＋カテゴリー化（十分位）４変数	

4-2 匿名化されたパネルデータの有用性と秘匿性の検証

本研究においては、秘匿処理が施されたパネルデータ(以下「秘匿処理済パネルデータ」と呼ぶ。)を試行的に作成し、その有用性と秘匿性を評価する。具体的には、本研究では、2 か月間の同一個体のレコードをリンクしたデータ(以下「2 か月パネル」と呼ぶ。)に、ミクロアグリゲーション、ノイズ付加等を適用した秘匿処理済パネルデータに対して、有用性と秘匿性の定量的な評価を行った。本研究で使用したデータは、平成 21 年家計調査の 1 月と 2 月の勤労者世帯(3427 世帯)に関する 2 か月パネルである。秘匿処理済パネルデータの作成のために、加法ノイズだけでなく、ミクロアグリゲーションとカテゴリー化といった 2 種類の匿名化技法の併用が行われる。ノイズ付加におけるパラメータ p の値については、0.16、0.30 と 0.50 の 3 つの数値が設定されている。また、匿名化技法の併用については、(1)個別ランキング法＋Z スコア総計法、(2)ミクロアグリゲーション(個別ランキング法等)＋カテゴリー化(十分位の平均値、十分位のランク)、(3)ノイズ付加＋カテゴリー化(十分位の平均値)の 3 つの方法が適用されている。これらの匿名化技法の併用については、前節で行った 1 時点の家計調査のミクロデータを用いた場合に採用した方法と基本的には変わらない。すなわち、匿名化技法を併用する場合には、勤め先収入及び消費支出とその他の量的属性については異なる手法を適用している[14]。

また、マッチングに使用した属性は、以下の 8 属性である。

1)市町村符号
2)単位区符号
3)世帯番号
4)一連世帯番号
5)抽出区分
6)住居の所有関係
7)延べ床面積
8)敷地面積

つぎに、有用性の評価については、標準化された属性値、相関係数行列と標準化されていない

[14] 匿名化パネルデータの試行的な作成においては、Brandt *et al.*(2008)も参考にした。

分散共分散行列についてそれぞれ平均平方誤差、平均絶対誤差及び平均変化率を計測した。また、秘匿性の評価に関しては距離計測型リンケージを用いている。

表３は、それぞれ加法ノイズおよびミクロアグリゲーション等の匿名化技法を併用した場合の有用性の結果を示したものである。全般的には、2 か月パネルにおける情報量損失の値の動きは、1 か月分のデータにおける数値のそれと比較して大きな違いは見られない。すなわち、2 か月パネルに攪乱的手法を適用した場合でも、ノイズ付加の場合、(ノイズとカテゴリー化の併用した場合でも)パラメータ p の値が小さいほど、情報量損失が小さくなるだけでなく、ミクロアグリゲーションを適用した場合、個別ランキング法を用いたほうが、有用性が高くなることがわかった。

表４は、秘匿性の評価の結果を示している。表６では、真のリンクと判定されたレコード数及び比率、1 対 1 の誤リンク、n 対 m(n 対 1、1 対 n を含む)のリンクが示されている。表６の結果から、2 か月パネルの場合、ノイズ付加(p=0.50)および個別ランキング法と Z スコア総計法の併用における真のリンクの比率が相対的に低いことが確認できる。また、ノイズ付加(p=0.50)の場合、1 か月分のデータと比較して、秘匿性に関する指標がそれほど大きく変わっていない。本実験結果は、秘匿処理済パネルデータの作成において、ノイズの適用可能性を示したものと考えることができる。

表３　攪乱的手法を適用した場合のミクロデータの有用性の結果―2 か月パネル

標準化された属性値

有用性評価の指標		属性値（標準化済み）		
		平均平方誤差	平均絶対誤差	平均変化率
ミクロアグリゲーション	個別ランキング＋Ｚスコア総計法	0.274280	0.224590	1.032780
加法ノイズ	p=0.16	0.020453	0.106010	0.806420
	p=0.30	0.068291	0.193820	1.476560
	p=0.50	0.170500	0.306630	2.321630
ミクロアグリゲーション（個別ランキング）＋カテゴリー化（十分位の平均値）		0.077829	0.080914	0.245470
ミクロアグリゲーション（個別ランキング）＋カテゴリー化（十分位のランク）		0.269020	0.252280	1.424610
ミクロアグリゲーション（Ｚスコア）＋カテゴリー化（十分位の平均値）		0.243720	0.238750	1.393770
加法ノイズ（p=0.10）＋カテゴリー化（十分位）		0.070396	0.110480	0.563300
加法ノイズ（p=0.16）＋カテゴリー化（十分位）		0.077517	0.132730	0.757410
加法ノイズ（p=0.30）＋カテゴリー化（十分位）		0.104870	0.182370	1.187850
加法ノイズ（p=0.50）＋カテゴリー化（十分位）		0.162300	0.245120	1.728290

相関係数行列

有用性評価の指標		相関係数行列		
		平均平方誤差	平均絶対誤差	平均変化率
ミクロアグリゲーション	個別ランキング＋Ｚスコア総計法	0.000414	0.008391	0.017424
加法ノイズ	p=0.16	0.000050	0.003735	0.008036
	p=0.30	0.000519	0.012502	0.027826
	p=0.50	0.003027	0.030517	0.068709
ミクロアグリゲーション（個別ランキング）＋カテゴリー化（十分位の平均値）		0.000049	0.003185	0.006776
ミクロアグリゲーション（個別ランキング）＋カテゴリー化（十分位のランク）		0.000394	0.006893	0.012356
ミクロアグリゲーション（Ｚスコア）＋カテゴリー化（十分位の平均値）		0.008996	0.047776	0.144770
加法ノイズ（p=0.10）＋カテゴリー化（十分位）		0.000063	0.003698	0.008031
加法ノイズ（p=0.16）＋カテゴリー化（十分位）		0.000106	0.005533	0.012568
加法ノイズ（p=0.30）＋カテゴリー化（十分位）		0.000473	0.012390	0.029149
加法ノイズ（p=0.50）＋カテゴリー化（十分位）		0.002237	0.026515	0.062659

表３続き

標準化されていない分散共分散行列

有用性評価の指標		分散共分散行列（標準化なし）		
		平均平方誤差	平均絶対誤差	平均変化率
ミクロアグリゲーション	個別ランキング＋Ｚスコア総計法	1.44E+17	131,679,773	0.313510
加法ノイズ	p=0.16	1.81E+17	155,738,524	0.023081
	p=0.30	1.87E+18	483,360,681	0.049110
	p=0.50	1.23E+19	1,242,591,211	0.096394
ミクロアグリゲーション（個別ランキング）＋カテゴリー化（十分位の平均値）		1.44E+17	129,796,443	0.079728
ミクロアグリゲーション（個別ランキング）＋カテゴリー化（十分位のランク）		1.44E+17	142,485,287	0.719340
ミクロアグリゲーション（Ｚスコア）＋カテゴリー化（十分位の平均値）		1.88E+19	1,981,683,979	0.160500
加法ノイズ（p=0.10）＋カテゴリー化（十分位）		5.23E+16	77,892,972	0.082296
加法ノイズ（p=0.16）＋カテゴリー化（十分位）		2.53E+17	175,921,875	0.086714
加法ノイズ（p=0.30）＋カテゴリー化（十分位）		2.38E+18	546,509,387	0.100330
加法ノイズ（p=0.50）＋カテゴリー化（十分位）		1.51E+19	1,388,062,063	0.127680

表４　攪乱的手法を適用した場合のミクロデータの秘匿性の結果―2か月パネル

		真のリンク		１対１の誤リンク	ｎ対ｍ
ミクロアグリゲーション	個別ランキング＋Ｚスコア総計法	1,886	55%	191	1,350
加法ノイズ	p=0.16	3,370	98%	0	57
	p=0.30	2,838	83%	58	531
	p=0.50	1,844	54%	299	1,284
ミクロアグリゲーション（個別ランキング）＋カテゴリー化（十分位の平均値）		3,197	93%	0	230
ミクロアグリゲーション（個別ランキング）＋カテゴリー化（十分位のランク）		2,439	71%	47	941
ミクロアグリゲーション（Ｚスコア）＋カテゴリー化（十分位の平均値）		2,173	63%	95	1,159
加法ノイズ（p=0.10）＋カテゴリー化（十分位）		3,211	94%	0	216
加法ノイズ（p=0.16）＋カテゴリー化（十分位）		3,176	93%	2	249
加法ノイズ（p=0.30）＋カテゴリー化（十分位）		2,896	85%	15	516
加法ノイズ（p=0.50）＋カテゴリー化（十分位）		2,288	67%	119	1,020

５．おわりに

本稿では、試行的にミクロデータにおける攪乱的手法の適用可能性に関する検証を行った。本分析の結果によれば、ノイズの付加の程度が高くなるにつれて、原データに対する秘匿処理済データの情報量損失が傾向的に大きくなることが実証的に確認された。また、ミクロアグリゲーションや加法ノイズ等を用いた攪乱的手法の組み合わせによっては、真のリンクの比率が相対的に小さくなることがわかった。このことから、攪乱的手法を適用した場合に、R-U マップを用いて有用性と秘匿性に関する相対評価を行うことできることから、わが国のミクロデータにおいて攪乱的手法の有効性の検証が可能なことが確認できた。

一方、本研究では、匿名化技法を適用したパネルデータの有用性と秘匿性の検証も行った。本研究の結果から、2 か月パネルに匿名化技法を適用した場合、匿名化技法の適用の仕方によっては、1 か月分の秘匿処理済データと秘匿性の程度が変わらない秘匿処理済パネルデータの作成も可能になることがわかった。これらの実験結果については、将来的に家計調査の匿名データの作成を検討する上で、参考資料になりうると考えている。

参考文献

Australian Bureau of Statistics (2007) *Technical Manual Household Expenditure Survey, Australia: Confidentialised Unit Record Files, Australia, 1998–99 (Third Edition - incl. Fiscal Incidence Study)*

Brandt, M., Lenz, R., Rosemann, M.(2008) “Anonymisation of Panel Enterprise Microdata-Survey of German Project”, Domingo-Ferrer, J. and Saygin, Y.(eds.) *Privacy in Statistical Databases UNESCO Chair in Data Privacy International Conference, PSD 2008 Istanbul, Turkey, September 2008, Proceedings*, pp.139-151.

Dalenius, T and Reiss, S. P. (1978) “Data-Swapping: A Technique for Disclosure Control (Extended Abstract)”, in Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, D.C., pp.191-194.

De Kort, S., and Wathan, J.(2009) “Guide to Imputation and Perturbation in the Samples of Anonymised Records”.

http://www.ccsr.ac.uk/sars/resources/imputation.doc.

De Waal, T. and Willenborg, L. (1999) “Information Loss through Global Recoding and Local Suppression”, *Netherlands Official Statistics (special issue on SDC),* Vol.14, pp.17-20.

Domingo-Ferrer, J. and Torra, V. (2001a) “Disclosure Control Methods and Information Loss for Microdata”, Doyle *et al.*(eds.) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier Science, Amsterdam, pp. 91-110.

Domingo-Ferrer, J. and Torra, V. (2001b) “A Quantitative Comparison of Disclosure Control Methods for Microdata”, Doyle *et al.*(eds.) *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, Elsevier Science, Amsterdam, pp.111-133.

Duncan, G. T., and Pearson, R. W. (1991) “Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future”, *Statistical Science*, Vol.6, pp.219-239.

Duncan, G. T., Elliot, M., Salazar-González, J.(2011) *Statistical Confidentiality*, Springer, New York.

Elliot, M. J., Manning, A. M., Ford, R. W.(2002) “A Computational Algorithm for Handling The Special Uniques Problem”, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No.5, pp.493-509.

Federal Committee on Statistical Methodology (1978) *Statistical Policy Working Paper 2: Report on Statistical Disclosure and Disclosure-Avoidance Techniques*, U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, Washington, D.C.

Gross, B., Guiblin, P., Merrett, K.(2004)“Implementing the Post Randomisation Method to The Individual

Sample of Anonymised Records (SAR) from The 2001 Census".

http://www.ccsr.ac.uk/////sars/2001/2001/pram.pdf.

藤野友和・垂水共之(2003)「PRAM の理論とその実用上の諸問題」『統計数理』第 51 巻第 2 号,321～335 頁

Gouweleeuw, J. M., Kooiman, P., Willenborg, L.C.R.J., de Wolf, P. P. (1998) "Post Randomization for Statistical Disclosure Control: Theory and Implementation", *Journal of Official Statistics*, Vol.14, No.4, pp.463-478.

星野伸明(2010)「公的統計ミクロデータ提供制度の課題」『日本統計学会誌』第 40 巻, 第 1 号, 23 ～45 頁

伊藤伸介(2008)「ミクロアグリゲーションに関する研究動向」,『製表技術参考資料』No.10, 3～32 頁

伊藤伸介・磯部祥子・秋山裕美(2008)「匿名化技法としてのミクロアグリゲーションの有効性に関する研究―全国消費実態調査を例に―」,『製表技術参考資料』No.10, 33～66 頁

伊藤伸介・磯部祥子・秋山裕美(2009)「秘匿性の評価方法に関する実証研究―全国消費実態調査のミクロアグリゲートデータを用いて―」,『製表技術参考資料』No.11, 1～35 頁

伊藤伸介(2010)「ミクロデータにおける秘匿性の評価方法に関する一考察」, 明海大学『経済学論集』第 22 巻第 2 号, 1～17 頁

伊藤伸介・高野正博・秋山裕美・後藤武彦(2010)「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』No.14, 1～40 頁

Ito, S. and Murata, M.(2011) "Quantitative Methods to Assess Data Confidentiality and Data Utility for Microdata in Japan", Paper presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Tarragona, Spain, pp.1-10.

http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/20_Japan.pdf.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., Sanil, A. P.(2006)"A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality", *The American Statistician*, Vol. 60, No.3, pp.1-9.

Kim, J. J.(1986) "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation", in Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 303-308.

Kim, J. J. and Winkler, W. E. (1995) "Masking Microdata Files", in Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 114-119.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., Walford, N. (1991)"The Case for Sample of Anonymized Records from the 1991 Census", *Journal of the Royal Statistical Society,* Series A, Vol. 154, No.2, pp.305-340.

Matloff, N. E.(1986) “Another Look at the Use of Noise Addition for Database Security”, in Proceedings of IEEE Symposium on Security and Privacy, pp.173-180.

Müller, W., Blien, U., Wirth, H.(1995) “Identification Risks of Micro Data: Evidence from Experimental Studies”, *Sociological Methods and Research*, Vol.24, No.2, pp.131-157.

Shlomo, N.(2007) “Statistical Disclosure Control Methods for Census Frequency Tables”, *S3RI Methodology Working Papers M07/04*, pp.1-40.

http://eprints.soton.ac.uk/44610/1/44610-01.pdf.

Shlomo, N.(2010) “Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility”, *The Journal of Privacy and Confidentiality*, Vol.2, No.1, pp.73-91.

Shlomo, N., Tudor, C., Groom, P. (2010) “Data Swapping for Protecting Census Tables”, Domingo-Ferrer, J. and Magkos, E.(eds) *Privacy in Statistical Databases UNESCO Chair in Data Privacy International Conference, PSD 2010 Corfu, Greece, September, 2010 Proceedings,* Springer, pp.41-51.

Strudler, M., Oh, H. L. and Scheuren, F.(1986) “Protection of Taxpayer Confidentiality with Respect to the Tax Model” in Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 375-381.

Takemura, A. (2002) “Local Recoding and Record Swapping by Maximum Weight Matching for Disclosure Control of Microdata Sets”, *Journal of Official Statistics*, Vol.18, No.2, pp.275-289.

Willenborg, L. and de Waal, T.(2001) *Elements of Statistical Disclosure Control*, Springer, New York.

Woo, M., Reiter, J. P., Oganian, A., Karr, A. F.(2009) “Global Measures of Data Utility for Microdata Masked for Disclosure Limitation”, *The Journal of Privacy and Confidentiality*, Vol.1, No.1,pp.111-124.

Yancey, W. E., Winkler, W. E., Creecy, R. H.(2002) “Disclosure Risk Assessment in Perturbative Microdata Protection”, *Research Report Series(Statistics #2002-01)*, Statistical Research Division U.S. Bureau of the Census.

http://www.census.gov/srd/papers/pdf/rrs2002-01.pdf.

Zayatz, L. (2007) “Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update”, *Journal of Official Statistics*, Vol.23, No.2, pp.253-265.

製 表 技 術 参 考 資 料 22

平成 25 年 5 月発行

編集・発行　　独立行政法人　統計センター

〒162－8668

東京都新宿区若松町 19－1

電　話　代表　03（ 5273 ）1200

統計情報・技術部 統計技術研究課　　TEL ： 03－5273－1368

E-mail ： research@nstac.go.jp