ミクロデータにおける匿名化技法の適用可能性の検証

―全国消費実態調査と家計調査を用いて―

伊藤伸介†、村田磨理子††、高野正博†††

Assessing Disclosure Limitation Methods for Japanese Official Microdata Based on Data from the ‘National Survey of Family Income and Expenditure’ and ‘Family Income and Expenditure Survey’

ITO, Shinsuke
MURATA, Mariko
TAKANO, Masahiro

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、匿名データで主として用いられているトップ(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等の攪乱的な手法についても、その有効性を実証的に明らかすることが求められるが、わが国ではそうした研究は数少ない。そこで、本稿では、全国消費実態調査と家計調査の個票データを用いて、ミクロデータに対する匿名化技法の適用可能性の検証を行った。

本研究では、様々な匿名化技法を用いて作成した秘匿処理済データの有用性と秘匿性の定量的な評価を行った上で、R-U マップを用いて、情報量損失と秘匿性の程度に関する比較分析を行っている。それによって、秘匿処理済データの有用性と秘匿性の相対的な評価が可能なことが明らかになった。

キーワード：政府統計ミクロデータ、匿名化技法、有用性、秘匿性、攪乱的手法

Further promoting the creation and use of Anonymized microdata from Japanese official statistics requires more extensive research on the potential of not only non-perturbative disclosure limitation methods such as top(bottom) coding and recoding, but also perturbative methods such as additive noise and microaggregation. However, only few studies that evaluate the various disclosure limitation methods exist in Japan. This paper assesses the potential of different disclosure limitation methods for Japanese official microdata based on anonymized data created from the ‘National Survey of Family Income and Expenditure’ and ‘Family Income and Expenditure Survey’.

This research quantitatively assesses data utility and data confidentiality for anonymized data created using disclosure limitation methods, and conducts a comparative analysis of information loss and degree of confidentiality using the R-U map. This allows a comparative evaluation of the effectiveness of various disclosure limitation methods for the creation of Anonymized microdata.

Key Words: Official Microdata, Disclosure Limitation Methods, Data Utility, Data Confidentiality, Perturbation

---

原稿受理日　平成 26 年 1 月 15 日

†独立行政法人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)
††元独立行政法人統計センター非常勤研究員((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)
†††総務省統計局統計調査部国勢統計課労働力人口統計室

１．はじめに

我が国では、統計法改正の全面施行に伴い、平成 21 年４月より、本格的な政府統計のミクロデータの提供が実施されている。我が国では，総務省統計局及び厚生労働省において 7 調査の匿名データが提供されているが、それらのほとんどについては、１つのタイプの匿名データのみが利用可能になっている。しかし、諸外国では、利用者のニーズに応じて、複数のタイプのミクロデータが提供されていることが少なくない。例えば、アメリカでは、1960 年代より人口センサスのミクロデータが公開されているが、2000 年人口センサスにおいては、地域区分や個人･世帯の社会経済的属性における分類区分の程度に応じて、標本抽出率が１％と５%の２種類の一般公開型ミクロデータ(Public Use Microdata Samples=PUMS)が利用可能になっている。また、イギリスの 2001 年人口センサスについては、データの構造や標本抽出率が異なる２種類の匿名化標本データ(Samples of Anonymised Records)が作成されているだけでなく、地域区分が詳細な小地域ミクロデータ(Small Area Microdata)が提供されている。さらに、カナダにおいては、2001 年人口センサスにおいて、個人ファイル(Individuals file)、家族ファイル(Families file)、及び世帯・住宅ファイル(Households and Housing file)の３種類のミクロデータファイルが提供されている。諸外国のこうした状況を踏まえると、我が国でも、将来的には利用者のニーズに対応した形で、複数のタイプの匿名データの作成・提供を本格的に進めることが考えられる。

一方，複数のタイプの匿名データの作成・提供を展開する場合，ミクロデータに対する匿名化技法の適用可能性を検討することが求められる。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)、PRAM(Post RAndomisation Method)[2]等が存在する(Domingo-Ferrer and Torra(2001a), Willenborg and de Waal(2001), Duncan *et al.*(2011))。

就業構造基本調査や全国消費実態調査等、我が国で現在提供されている匿名データのほとんどは、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 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))[3]。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 de Waal(2001, p.126))。なお、我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)や伊藤・星野(2013)による実証研究がある。

[2] 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。

[3] アメリカセンサス局は、2000 年人口センサスの PUMS を作成する上で、特定されるリスクの高い世帯(ex. 世帯人員が 10 人以上の世帯)を対象に世帯員の年齢にノイズを付与している。具体的には、個別データにおいて該当するレコードに含まれ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設定している。また、ノイズの導入においては、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な処理が施されている(Zayatz(2007, p.257))。

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))[4]。

こうした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は、匿名化されたミクロデータにおける有用性や秘匿性の定量的な評価(Yancey *et al.*(2002), Shlomo(2010)等)を行うことが必要だと思われる。こうした定量的な評価によって、ミクロデータの作成における判断材料として有益な数量情報を提示することが可能になる。

そこで、本稿は、これまでの先行研究を参考にしながら、ミクロデータにおける有用性と秘匿性の評価方法を検討する。次に、匿名化された政府統計ミクロデータを用いて、有用性と秘匿性の定量的な評価に関する実証研究を行うことによって，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。

２．ミクロデータにおける有用性の評価方法に関する検討

本節では、ミクロデータにおける有用性の定量的な評価方法について議論する。有用性の評価方法については、秘匿処理を施していない個別データ(以下、「原データ」という。)と、原データに様々な匿名化技法を適用することによって作成したミクロデータ(以下「秘匿処理済データ」という。)を対象に、基本統計量による比較や情報量損失(information loss)の評価、さらには、傾向スコアの計測、クラスター分析による検証、経験分布関数における差異の評価(Woo *et al.*(2009, pp.113-115))等、様々な方法が考案されてきた。一方、ミクロデータに含まれる属性の性質によって、適用可能な有用性の評価方法が異なると考えられる。本節では、量的属性と質的属性のそれぞれについて、有用性評価に関する主な方法を述べることにしたい。

(1) 統計指標を用いた有用性の評価

ミクロデータに含まれる量的属性に対して有用性の相対的な程度を評価する場合、原データと秘匿処理済データにおいてデータ構造の近似性を検証することが考えられる。具体的には、①平均、分散等の基本統計量、②分布上の特性、③情報量損失を比較・検証することが提案されている(Mateo-Sanz，Domingo-Ferrer and Sebé(2005，pp.182-184))。

情報量損失は、秘匿処理済データが原データと比べてどの程度情報量を失っているかを算出したものであって、原データと秘匿処理済データにおける統計指標の数値の差異によって評価される。具体的には、原データと秘匿処理済データに含まれる属性値の差や、分散共分散行列や相関係数行列に見られるデータ構造の変化によって、情報量損失の計測が行われている[5]。さらに、情報量損失の大きさについては、①平均平方誤差(mean square error)、②平

[4] アメリカにおいても、2000 年人口センサスの集計表に対して秘匿処理を行うために、人口センサスの個別データにスワッピングが適用されている。具体的には、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スワッピングが用いられる。さらには、American Community Survey にもスワッピングが使用されていることが知られている。スワッピングの対象となるレコードは、詳細な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であって、そのようなレコードについては、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。なお、スワッピングされた個別データから PUMS さらには集計表が作成されている(Zayatz(2007, p.257))。

[5] Domingo-Ferrer and Torra(2001a)は、次の統計指標を用いて原データに対する秘匿処理済データの情報量損失を計測することを提唱している(Domingo-Ferrer and Torra(2001a, p.104))。

均絶対誤差(mean absolute error)、③平均変化率(mean variation)といった尺度を用いて評価が行われる(Domingo-Ferrer and Torra(2001a, p.104))。表１は、k 個の属性、n 個のレコード数を持つ原データと秘匿処理済データを対象に、属性値の差、相関係数行列の差、及び分散共分散行列の差について、平均平方誤差、平均絶対誤差と平均変化率による情報量損失の算定式を表している。表１から明らかなように、平均変化率については、平均平方誤差や平均絶対誤差と異なり、各属性の単位の違いによる影響を受けないことが特徴である。平均平方誤差等の情報量損失の値が０に近いほど、原データと秘匿処理済データは近似しており、有用性が相対的に高いと判定することができる[6]。

(2) 質的属性値間の距離の計測

質的属性に匿名化技法を適用した場合の有用性の評価についても、様々な指標が考案されている。その１つが、質的属性値間の距離(distance for categorical variables)を定義し、その距離の近さの程度を計測することである(Domingo-Ferrer and Torra (2001a, pp.105-106))。

質的属性値間の距離を用いる場合、対象となる質的属性が順序尺度か名義尺度のいずれに該当するかによって、原データと秘匿処理済データにおける質的属性値間の距離の計測方法が異なる。順序尺度については、匿名化技法の適用による質的属性値の変化幅を属性値の分類区分の数で除した値が、質的属性値間の距離と定義される。その一方で、名義尺度に関しては、匿名化技法によって属性値が変化した場合の質的属性値間の距離は１、属性値が変化しなかった場合の距離は０、とそれぞれ定義される。

他方、順序尺度と名義尺度の両方の質的属性が含まれるデータも存在するが、その場合には、上記で定義した距離を結合した上で、相対的な属性の重要度を考慮し、それに応じた重みを付けた指標を作成することが考えられる(Takemura(1999, pp.4-5))。

このようにして算出した距離の値が小さいほど、原データと秘匿処理済データは近似したとみなされ、秘匿処理済データにおける有用性が高いと評価できる。

(3) 情報エントロピーを用いた有用性の評価

諸外国の先行研究によれば、質的属性に関する有用性の定量的な評価に関しては、Shannon が提唱する情報量（以下「シャノン情報量」という。）の概念に基づいた「情報エントロピー(entropy-based measures)」を用いて情報量損失を評価することが提案されている(Kooiman *et al.*(1998), Domingo Ferrer and Torra(2001a))。また、竹村(2003)は、個別データの持つ情報量を定量的に評価する上で、情報エントロピーを用いることの有効性を指摘している(竹村(2003, 250 頁))。そこで、本節では、質的属性の有用性の評価方法の１つとして、情報エ

---

①分散共分散行列
②相関係数行列
③属性と主成分分析から得られた主成分との間の相関係数行列
④属性の各々と第１主成分(それ以外の主成分)との共通性(commonality)(各属性が第１主成分(あるいはそれ以外の主成分)によって説明される比率)
⑤因子得点係数行列(factor score coefficient matrix)

[6] 分散共分散行列や相関係数行列などの複数の統計指標を組み合わせた評価式を定式化し、情報量損失の大きさを計測することも考えられる(Domingo-Ferrer and Torra(2001b, p.118))。

表１　平均平方誤差、平均絶対誤差と平均変化率による情報量損失の算定式

	平均平方誤差 (Mean square error)	平均絶対誤差 (Mean absolute error)	平均変化率 (Mean variation)
属性値の差	$\frac{\sum_{j=1}^{k}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{nk}$	$\frac{\sum_{j=1}^{k}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{nk}$	$\frac{\sum_{j=1}^{k}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{nk}$
相関係数行列の差	$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}(r_{ij}-r'_{ij})^2}{\frac{k(k-1)}{2}}$	$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{k(k-1)}{2}}$	$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{k(k-1)}{2}}$
分散共分散行列の差	$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{k(k+1)}{2}}$	$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{k(k+1)}{2}}$	$\frac{\sum_{j=1}^{k}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{k(k+1)}{2}}$

n：原データと秘匿処理済データにおけるレコードの総数
k：原データと秘匿処理済データに含まれる属性の数
x_{ij}：原データ上の i 番目のレコードにおける j 番目の属性の値
x'_{ij}：秘匿処理済データ上の i 番目のレコードにおける j 番目の属性の値
r_{ij}：原データにおける i 番目の属性と j 番目の属性に関する相関係数
r'_{ij}：秘匿処理済データにおける i 番目の属性と j 番目の属性に関する相関係数
v_{ij}：原データにおける i 番目の属性と j 番目の属性に関する分散ないしは共分散
v'_{ij}：秘匿処理済データにおける i 番目の属性と j 番目の属性に関する分散ないしは共分散
出所　Domingo-Ferrer and Torra(2001a, p.105)に基づいて作成

ントロピーに基づいた情報量損失の計測に焦点を当てることにしたい。

ある特定の状態が生じる確率を p とする。このとき、確率 p の対数を用いて、(1)式のようなシャノン情報量が定義される。

$$\text{シャノン情報量} = -\log p(0 \le p \le 1) \tag{1}$$

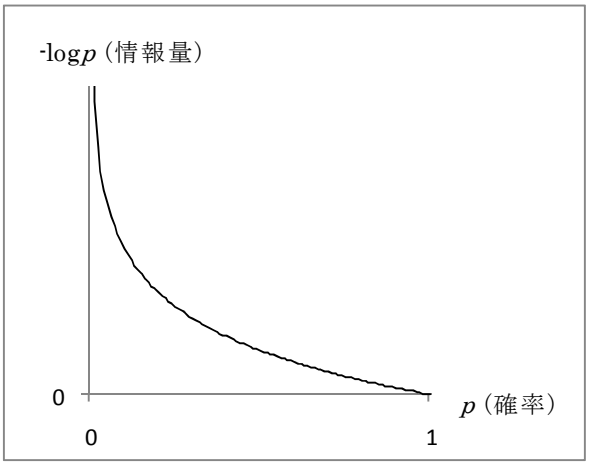
図１は、シャノン情報量をグラフで表示したものであって、横軸は確率 p を、縦軸はシャノン情報量$-\log p$を、それぞれ表している。図１を見ると、シャノン情報量は確率が０に近づくほど増加することが分かる。このことは、稀少な状態が生じたことを表す情報（確率の低い情報）ほど、シャノン情報量が大きくなることを示している。

次に、情報エントロピーは、シャノン情報量に確率 p を乗じ、その事象の数だけ総計したものと定義される。すなわち、情報エントロピーはシャノン情報量の期待値を表している。

$$\text{情報エントロピー} = -\sum_{i=1}^{n} p_i \log p_i \tag{2}$$

n：事象の数
p_i：i 番目の事象が起こる確率

図１　シャノン情報量と確率の関係



De Waal and Willenborg(1999)は、匿名化技法としてリコーディングを用いて作成した秘匿処理済データについて、情報エントロピーを用いて情報量損失を計測した。De Waal and Willenborg(1999)によれば、最初に、匿名化技法の適用によって属性値が変化する確率（以下「移行確率（transition probability)」という。）を用いて情報エントロピーが算出される。次に、情報エントロピーが計測された対象となるレコード数を乗じることによって、情報量損失が求められる。

移行確率については簡単な例を用いて説明することにしたい。図２の例は、職業区分「01. 国家公務員」(10 レコード）と「02. 地方公務員」(30 レコード）を統合して、新たに「03. 官公職員」（総数 40 レコード）という区分を作成したものである。このような分類区分の統合によって、「01. 国家公務員」から「03. 官公職員」に移行したレコード数は、40 レコードの中の 10 レコードとなる。したがって、属性値の移行確率は 10/40 と考えられる。同様に、「02. 地方公務員」から「03. 官公職員」に変化したレコード数は、40 レコードの中で 30 レコードであることから、移行確率は 30/40 と算出される。

次に、質的属性に対してリコーディングを適用することによって作成した２種類の秘匿処理済データを対象に、情報量損失を計測した上でデータ間の比較を行う。図３は、世帯人員の分類区分「５人」、「６人」、「７人以上」の分類区分を対象に、①「５人」と「６人以上」、②「５人以上」という２種類のリコーディングを適用したものである。

上記の２種類のリコーディングを対象に、匿名化技法の適用による属性値の移行を表す確率を算出する。最初に、①「５人」と「６人以上」について見ると、「６人以上」に統合された 400 レコードの中で、元の分類区分が「６人」及び「７人以上」であったレコード数は、それぞれ 300 レコードと 100 レコードとなっている。したがって、世帯人員に関する属性値が「６人」ないしは「７人以上」であるレコードが，新たな分類区分「６人以上」に統合された場合の移行確率は，それぞれ 300/400 と 100/400 と算出される。このとき、情報エントロピーは、

$$\text{情報エントロピー} = -\frac{300}{400}\log_2\frac{300}{400} - \frac{100}{400}\log_2\frac{100}{400} = 0.81128 \qquad (3)$$

と求められる。なお、情報エントロピーを計算する上で、対数の底については２が用いられている。次に、情報エントロピーに対してリコーディングの対象となるレコード数 400 を乗じることによって、情報量損失が算出される。例えば、分類区分「６人以上」に移行した場

図２　匿名化技法の適用による移行確率の考え方

区分	項目名	件数
01	国家公務員	10
02	地方公務員	30

リコーディング →

区分	項目名	件数
03	官公職員	40

・01→03 … 確率＝10/40
・02→03 … 確率＝30/40

図３　世帯人員の分類区分の中で、２区分あるいは３区分を統合した場合

区分	世帯人員	件数
01	５人	600
02	６人	300
03	７人以上	100

リコーディング →

区分	世帯人員	件数
01	５人	600
02	６人以上	400

区分	世帯人員	件数
01	５人	600
02	６人	300
03	７人以上	100

リコーディング →

区分	世帯人員	件数
01	５人以上	1000

合の情報量損失は、

$0.81128 \times 400 = 324.51$ (4)

である。

一方、分類区分「５人」における情報量損失は次のとおりである。「５人」に該当する 600 のレコードはリコーディング後も全て同じ区分に属するために、移行確率は 600/600 となる。したがって、情報エントロピーの値は０となることから、情報量損失も０となる。このことから、世帯人員３区分を「５人」及び「６人以上」にリコーディングした場合の情報量損失は、約 325 となる。同様に、②「５人以上」の場合における情報量損失は次のように算出される。最初に、世帯人員に関して「５人」、「６人」と「７人以上」に該当するレコードが「５人以上」の新しい分類区分に統合された場合の移行確率が、それぞれ 600/1000、300/1000、100/1000 であることから、情報エントロピーは 1.29546 と求められる。次に、情報量損失は、情報エントロピーにレコード数 1000 を乗じることによって、約 1295 と算出される((5)式)。

$$\text{情報量損失} = \left(-\frac{600}{1000}\log_2\frac{600}{1000} - \frac{300}{1000}\log_2\frac{300}{1000} - \frac{100}{1000}\log_2\frac{100}{1000} \right) \times 1000 = 1295.46 \quad (5)$$

図３の２つのケースについて情報量損失を比較すると、「５人以上」における情報量損失がより高い数値を示していることがわかる。このように、情報エントロピーを用いて情報量損失の程度を測ることによって、ミクロデータに含まれる質的属性についても、その有用性が定量的に評価できる。

次に、図４は世帯主の職業及び住居の所有関係という２つの質的属性に関するリコーディ

図４　２つの属性におけるリコーディング―世帯主の職業及び住居の所有関係

		住居の所有関係		
		持ち家（世帯員）	持ち家（その他）	借家
世帯主の職業	国家公務員	10	2	5
	地方公務員	70	18	25
	民間職員	180	20	50

リコーディング

		住居の所有関係	
		持ち家	借家
世帯主の職業	官公職員	100	30
	民間職員	200	50

ングを示したものである。図４を見ると、世帯主の職業については「国家公務員」と「地方公務員」の２つの区分が「官公職員」という新たな区分に統合されており、住居の所有関係については「持ち家（世帯員名義）」と「持ち家（その他名義）」の２つの区分が「持ち家」という区分に統合されている。次に、リコーディング後の世帯主の職業と住居の所有関係の組合せ（〔官公職員・持ち家〕、〔官公職員・借家〕、〔民間職員・持ち家〕、〔民間職員・借家〕）の４つのパターンについて、情報エントロピーが計算される。図５は、上記の４パターンを対象に情報エントロピーを算出したものである。例えば、〔官公職員・持ち家〕の組合せに対する情報エントロピーは、(6)式によって 1.25058 と求められる。

$$\text{情報エントロピー}=-\frac{10}{100}\log_2\frac{10}{100}-\frac{70}{100}\log_2\frac{70}{100}-\frac{2}{100}\log_2\frac{2}{100}-\frac{18}{100}\log_2\frac{18}{100}=1.25058 \qquad (6)$$

図６は、各区分に関する情報エントロピーに基づいて情報量損失を算出したものである。〔民間職員・借家〕の組合せについては、リコーディングによる区分の変更がないことから、情報エントロピーの値は０になっている。図６における４区分の情報量損失を総計することによって、世帯主の職業及び住居の所有関係の２つの属性をリコーディングした場合の情報量損失は約 238 と求められる。

３．ミクロデータにおける秘匿性の評価方法に関する検討

ミクロデータの有用性と秘匿性はトレードオフの関係にあると考えられる。例えば、原データと秘匿処理済データが近似的なデータ構造を有している場合、秘匿処理済データの有用性は相対的に高くなるのに対して、秘匿性の強度は低くなることが考えられる。一方、情報量損失といった有用性の評価方法は、原データと秘匿処理済データにおけるデータ構造の近似性を計測することを指向しているが、秘匿処理済データに含まれる特定のレコードが元のレコードと対応付け可能かどうかを検証することを目指していない。したがって、有用性を評価することとは別に、ミクロデータにおける秘匿性についても定量的な評価を行うことが求められる。

秘匿性の評価方法については様々な手法が議論されてきたが、主な手法の１つとして、外部情報とミクロデータのマッチングを指摘することができる。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))。また、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についてもこれまで多くの議

図５　リコーディング後における各分類区分の情報エントロピー

		住居の所有関係	
		持ち家	借家
世帯主の職業	官公職員	1.25058	0.65002
	民間職員	0.46900	0.00000

図６　リコーディング後における各分類区分の情報量損失

		住居の所有関係	
		持ち家	借家
世帯主の職業	官公職員	125.06	19.50
	民間職員	93.80	0.00

論が行われてきた。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))。その一方で、秘匿性の定量的な評価方法として、レコードリンケージとクロス集計表による評価方法も存在する。本節では、この２つの評価方法に焦点を絞って議論することにしたい。

(1) 確定的リンケージ及び距離計測型リンケージによる秘匿性の評価

量的属性群に匿名化技法を適用することによって作成した秘匿処理済データの秘匿性の強度を定量的に評価する方法として、レコードリンケージの手法が考えられる。レコードリンケージは、原データのレコードと秘匿処理済データのレコードとの間で対応付けが可能かどうか(以下では、対応付けられた場合を「真のリンク」という。）を判定することによって、秘匿性の強度を定量的に評価することを指向している。そして、真のリンクとなるレコードの割合は、様々な秘匿処理済データにおける秘匿性評価のための指標として用いられる。本節では、確定的リンケージ(deterministic record linkage)と距離計測型リンケージ(distance-based record linkage)について議論することにしたい。

最初に、確定的リンケージとは、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」という。）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))。確定的リンケージでは、原データのレコードと秘匿処理済データのレコードにおいてリンクキー変数の属性値がすべて一致した場合、そのレコードは真のリンクであると判定することができる（図

７)。真のリンクとなるレコードの割合が低いほど、秘匿性の強度は高いと考えられる。

一方、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。それは、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方法である(Domingo-Ferrer and Torra(2001b, p.116))。原データにおける属性値が区間の範囲内に存在した場合のレコード数と総レコード数との比率は秘匿性の評価指標と考えられ、その比率が低いほど、秘匿性の強度が高いと判定することができる。

区間の設定においては、順位統計量に基づいた区間(rank-based intervals)と標準偏差に基づいた区間(standard deviation-based intervals)を用いることが提案されている(Mateo-Sanz *et al.*(2004, pp.204-205))。順位統計量に基づいた区間の場合、最初に、原データと秘匿処理済データに含まれる各レコードについて、各々の属性値の全体における順位が付与される。次に、秘匿処理済データのレコード上にある特定の属性値の順位を基準として、レコード総数の p%(p は任意)の範囲内に、原データにおいて対応するレコードの属性値の順位が含まれるかどうかを確認することによって、秘匿性を評価することが可能になる（図８）。その一方で、標準偏差に基づいた区間を用いた場合には、秘匿処理済データのレコード上にある特定の属性値を中心とした p%の標準偏差の範囲に、原データにおいて対応するレコードの属性値が含まれるかについての検証を行うことによって、秘匿性の強度が確認できる（図９）。

他方、距離計測型リンケージは、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方法である(伊藤(2010，9～10 頁))。具体的には、最初に、秘匿処理済データのレコードから原データの各レコードへの距離を計測し（図 10)、次に、最も距離が短くなるレコードが、原データの元のレコードであり、かつ同じ距離となるレコードが他に存在しない場合に、そのレコードは真のリンクであると判定される（図 11)。

距離計測型リンケージを行う場合には、ユークリッド距離やマハラノビス距離といった距離が用いられる[7]。また、Torra *et al.*(2006, pp.235-236)は、ユークリッド距離に関して、①属性値の標準化(標準化ユークリッド距離)、及び②属性間の距離の標準化という２種類の標準化を提案している[8]。２種類の標準化の式については、それぞれ(7)式と(8)式で設定することが可能である。

① 属性値の標準化（標準化ユークリッド距離）

$$d_{il}^2 = \sum_j \left(\frac{x_{ij} - \bar{x}_j}{\sigma(x_j)} - \frac{X_{lj} - \bar{X}_j}{\sigma(X_j)} \right)^2 \qquad (7)$$

7 距離の計測方法については、ユークリッド距離やマハラノビス距離の他に、例えばカーネル距離による計測方法がある。例えば Torra *et al.*(2006)を参照。

8 秘匿処理済データに含まれるある特定のレコードの属性値が、原データにおいて対応する元のレコードの値と変わらない場合、その属性に関する標準偏差が原データと秘匿処理済データで異なるのであれば、それらの属性値における標準化ユークリッド距離は０にならない。一方、属性間の距離の標準化では、属性間の距離の平均値と標準偏差を用いることから、上記のような問題点は回避される。なお、伊藤・磯部・秋山(2009)では、標準化ユークリッド距離のみを用いて秘匿性の検証を行っている。

図 7　確定的リンケージのイメージ

原データ

一連番号	リンクキー変数			
	住居の所有関係	性別	年齢	収入
001	1	1	40	520
002	1	1	50	480
003	1	1	60	500

秘匿処理済データ

一連番号	リンクキー変数			
	住居の所有関係	性別	年齢	収入
001	1	1	40	520
002	1	1	50	450
003	1	1	60	530

一連番号001のレコードは真のリンクである

図 8　順位統計量に基づいた区間を用いた評価

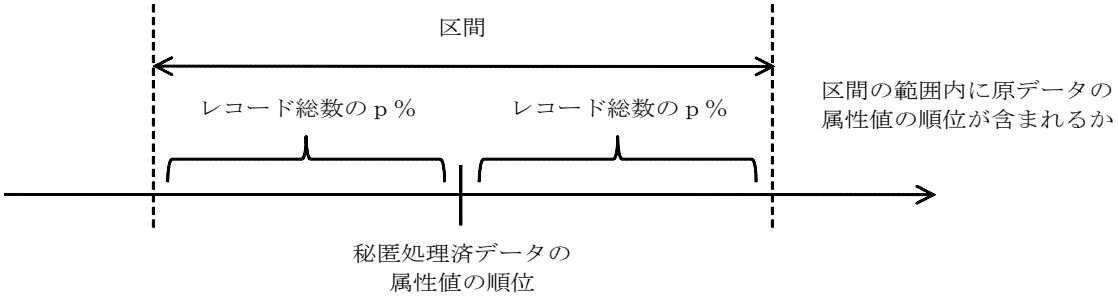


図 9　標準偏差に基づいた区間を用いた評価

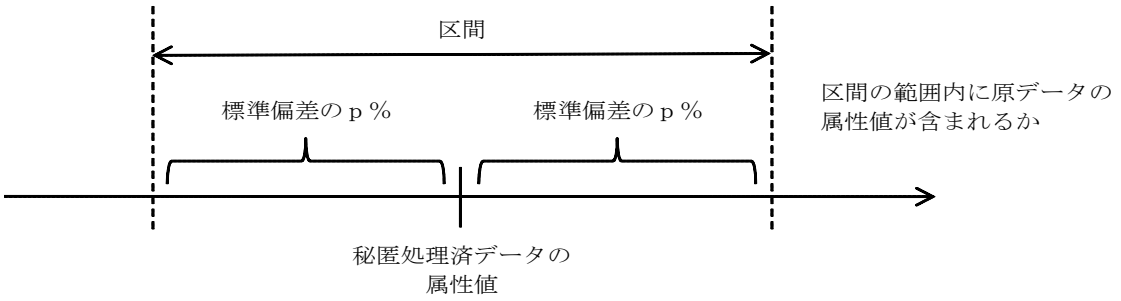


図 10　距離計測型リンケージのイメージ

原データ

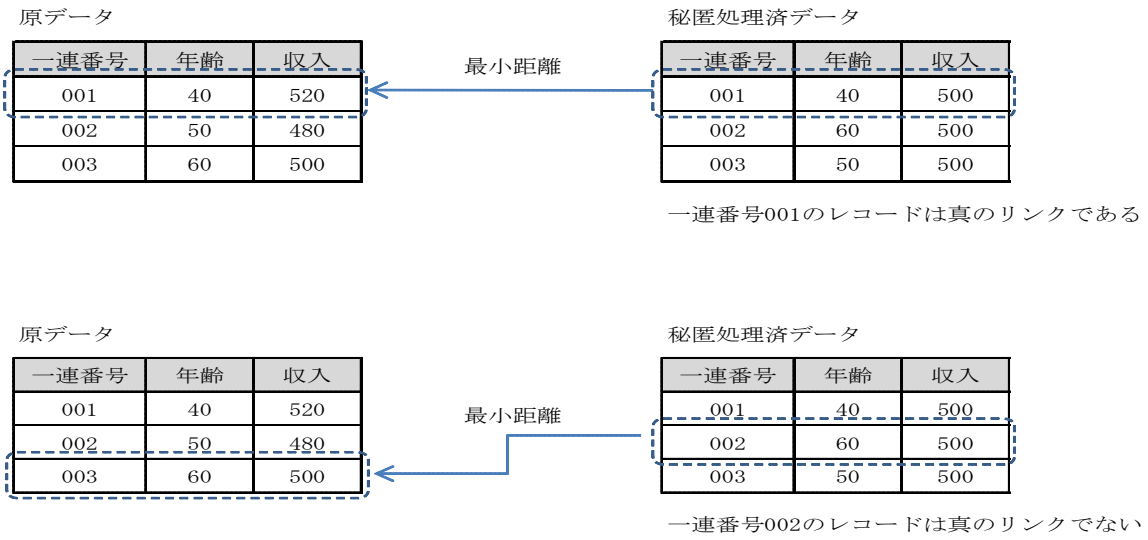
一連番号	年齢	収入
001	40	520
002	50	480
003	60	500

距離の計測

秘匿処理済データ

一連番号	年齢	収入
001	40	500
002	60	500
003	50	500

図 11　距離計測型リンケージにおける真のリンクのイメージ



② 属性間の距離の標準化

$$d_{iI}^2 = \sum_j \left(\frac{(x_{ij} - X_{Ij}) - (\overline{x_j - X_j})}{\sigma(x_j - X_j)} \right)^2 \tag{8}$$

d_{iI}：原データの i 番目のレコードと秘匿処理済データの I 番目のレコードとの間の距離
x_{ij}：原データの i 番目のレコードにおける j 番目の属性値
X_{Ij}：秘匿処理済データの I 番目のレコードにおける j 番目の属性値
$\bar{x}_j$ ：原データの j 番目の属性の平均値
$\bar{X}_j$ ：秘匿処理済データの j 番目の属性の平均値
$\sigma(x_j)$：原データの j 番目の属性の標準偏差
$\sigma(X_j)$：秘匿処理済データの j 番目の属性の標準偏差
$\overline{x_j - X_j}$ ：原データと秘匿処理済データにおける j 番目の属性間の距離に関する平均値
$\sigma(x_j - X_j)$：原データと秘匿処理済データにおける j 番目の属性間の距離に関する標準偏差

一方、マハラノビス距離は、属性値における散らばりの大きさと属性間の相関関係を考慮して算出される距離であり、次の(9)式で表される。

$$d^2 = (\mathbf{x}_i - \mathbf{X}_I)^T S^{-1} (\mathbf{x}_i - \mathbf{X}_I) \tag{9}$$

$\mathbf{x}_i$：原データにおける i 番目のレコードに含まれる属性値のベクトル
$\mathbf{X}_I$：秘匿処理済データにおける I 番目のレコードに含まれる属性値のベクトル
S^{-1}：分散共分散行列の逆行列

なお、分散共分散行列が、属性間に相関がない対角行列である場合、マハラノビス距離は標準化ユークリッド距離に一致する。さらに、分散共分散行列が単位行列の場合、マハラノビス距離はユークリッド距離に等しくなることが知られている。

(2) 確率的リンケージによる秘匿性の評価

確率的リンケージ(probabilistic record linkage)は、Fellegi and Sunter(1969)によって確立されたレコードリンケージの一手法であるが、ミクロデータの秘匿性に関する定量的な評価方法の１つとしても展開されてきた[9]。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。具体的には、原データにおけるレコードと秘匿処理済データに含まれるレコードの全てのペアを対象に、２つのレコード間における各属性値の一致の程度に関する情報に基づいて真のリンクであるレコードを判定する。そして、確定的リンケージや距離計測型リンケージと同様に、真のリンクであるレコードの比率が低いほど、秘匿性の強度は高いとみなすことができる。本節では、Torra and Domingo-Ferrer(2003)で示された事例を参照しながら、確率的リンケージによる秘匿性の評価方法の概要を説明する。

図 12 のように、年齢、職業、年収と貯蓄という４つの属性を持った原データ(A)と秘匿処理済データ(B)を考える。秘匿処理済データは、原データの年齢についてはトップコーディング、年収と貯蓄に関してはそれぞれ平均値に置き換えて作成したものである。

原データ(A)と秘匿処理済データ(B)のレコードの組 $(a,b)\in A\times B$ については、図 13 に示されるように 25 パターン(原データ５レコード×秘匿処理済データ５レコード)が作成される。また、原データと秘匿処理済データが同一のレコードに該当する場合には M(照合されたペア(Matched Pair))が、同一のレコードに該当しない場合には U(照合されないペア(Unmatched Pair))が、それぞれ図 13 の M/U 欄に表示されている。なお、M と U は、次の(10)式で与えられる。

$$M=\{(a,b);a=b,a\in A,b\in B\},\quad U=\{(a,b);a\neq b,a\in A,b\in B\} \tag{10}$$

また、図 13 の一致フラグ(γ) は、原データと秘匿処理済データの年齢、職業、年収、貯蓄の各々について、属性値が一致する場合には１を、一致しない場合には０を付与したものである。例えば、一致フラグを γ=(年齢, 職業, 年収, 貯蓄) とすると、一致フラグのパターン Γ は次の 16 通りで構成される((11)式)。

$$\Gamma=\{\gamma_1,\gamma_2,\cdots,\gamma_{16}\}=\{(1,1,1,1),(1,1,1,0),(1,1,0,1),(1,0,1,1),(0,1,1,1),(1,1,0,0),\cdots,(0,0,0,0)\} \tag{11}$$

図 14 では、図 13 に現れた一致フラグのパターン別の M と U のレコード数が算出されている。また、図 14 における条件付き確率 m と u は、レコードの組(a, b)が同一レコードである組(M) 及び同一レコードでない組(U)の中で、それぞれの一致フラグのパターンに該当する確率を表したものであって、次の(12)式で与えられる。

$$m=P(\gamma=\gamma(a,b)\,|\,(a,b)\in M),\ u=P(\gamma=\gamma(a,b)\,|\,(a,b)\in U) \tag{12}$$

さらに、Torra and Domingo-Ferrer(2003)は、レコードの照合、未照合を判別するための評価式として次の(13)式を定義している。

$$R_P(a,b)=R(\gamma(a,b))=\log(m/u) \tag{13}$$

(13)式を見ると、m の値が大きく u の値が小さいほど、R の値が大きくなることがわかる。よって、一致フラグ(γ)の各パターンに属するレコードについて、同一レコードでない組(U)よりも同一のレコードの組(M)の比率が大きい場合には、R は大きな値を示す。

---

[9] Fellegi and Sunter(1969)は、２種類のデータファイルに含まれるレコードの対応付け(link, possible-link, non-link)を行うために、新たな数理モデルを提案して理論的なフレームワークを与えた。その中で、確率に基づいたリンケージルール(optimal linkage rule)を定義し、対応付けを判定するための確率的リンケージの手法を確立した。また、Jaro(1989)は、フロリダ州タンパの 1985 年センサスに関するレコードリンケージを行う方法として、確率的リンケージの手法を導入した。

図 12　原データと秘匿処理済データの例―年齢、職業、年収と貯蓄

原データ(A)				秘匿処理済データ(B)			
年齢	職業	年収	貯蓄	年齢	職業	年収	貯蓄
30	官公職員	500	1000	30	官公職員	500	1000
30	官公職員	400	1000	30	官公職員	500	1000
25	民間職員	500	1000	25	民間職員	500	1000
88	民間職員	500	600	85	民間職員	500	1000
86	民間職員	600	1400	85	民間職員	500	1000

図 13　原データと秘匿処理済データにおける属性値のパターンと一致フラグ

原データ(A)				秘匿処理済データ(B)				M/U	一致フラグ(γ)			
年齢	職業	年収	貯蓄	年齢	職業	年収	貯蓄		年齢	職業	年収	貯蓄
30	官公職員	500	1000	30	官公職員	500	1000	M	1	1	1	1
30	官公職員	500	1000	30	官公職員	500	1000	U	1	1	1	1
30	官公職員	500	1000	25	民間職員	500	1000	U	0	0	1	1
30	官公職員	500	1000	85	民間職員	500	1000	U	0	0	1	1
30	官公職員	500	1000	85	民間職員	500	1000	U	0	0	1	1
30	官公職員	400	1000	30	官公職員	500	1000	U	1	1	0	1
30	官公職員	400	1000	30	官公職員	500	1000	M	1	1	0	1
30	官公職員	400	1000	25	民間職員	500	1000	U	0	0	0	1
30	官公職員	400	1000	85	民間職員	500	1000	U	0	0	0	1
30	官公職員	400	1000	85	民間職員	500	1000	U	0	0	0	1
25	民間職員	500	1000	30	官公職員	500	1000	U	0	0	1	1
25	民間職員	500	1000	30	官公職員	500	1000	U	0	0	1	1
25	民間職員	500	1000	25	民間職員	500	1000	M	1	1	1	1
25	民間職員	500	1000	85	民間職員	500	1000	U	0	1	1	1
25	民間職員	500	1000	85	民間職員	500	1000	U	0	1	1	1
88	民間職員	500	600	30	官公職員	500	1000	U	0	0	1	0
88	民間職員	500	600	30	官公職員	500	1000	U	0	0	1	0
88	民間職員	500	600	25	民間職員	500	1000	U	0	1	1	0
88	民間職員	500	600	85	民間職員	500	1000	M	0	1	1	0
88	民間職員	500	600	85	民間職員	500	1000	U	0	1	1	0
86	民間職員	600	1400	30	官公職員	500	1000	U	0	0	0	0
86	民間職員	600	1400	30	官公職員	500	1000	U	0	0	0	0
86	民間職員	600	1400	25	民間職員	500	1000	U	0	1	0	0
86	民間職員	600	1400	85	民間職員	500	1000	U	0	1	0	0
86	民間職員	600	1400	85	民間職員	500	1000	M	0	1	0	0

図 14　一致フラグ別のレコード数及び条件付き確率

一致フラグ（γ）				レコード数		条件付き確率		m/u	log(m/u)	インデックス番号
年齢	職業	年収	貯蓄	M	U	m	u			
1	1	1	1	2	1	2/5	1/20	8	2.08	1
1	1	0	1	1	1	1/5	1/20	4	1.39	2
0	1	1	0	1	2	1/5	2/20	2	0.69	3
0	1	0	0	1	2	1/5	2/20	2	0.69	4
0	1	1	1	0	2	0/5	2/20	0	−∞	5
0	0	1	1	0	5	0/5	5/20	0	−∞	6
0	0	1	0	0	2	0/5	2/20	0	−∞	7
0	0	0	1	0	3	0/5	3/20	0	−∞	8
0	0	0	0	0	2	0/5	2/20	0	−∞	9
合　計				5	20	1	1			

次に、確率的リンケージにおいては、以下の判別ルール(decision rule)にしたがって、レコード間の対応付けの有無が判定される。

1. $R_p(a,b) \geq ut$ ならば、レコードの組み合わせ(a,b)は対応付けされるペア(Linked Pair=LP)
2. $R_p(a,b) \leq lt$ ならば、レコードの組み合わせ(a,b)は対応付けされないペア(Non-Linked Pair=NP)
3. $lt < R_p(a,b) < ut$ ならば、レコードの組み合わせ(a,b)が LP と NP のいずれに該当するかが十分に判断できない(Clerical Pair=CP)

判別ルールにおいては、R の値を上限の閾値(upper threshold)である ut、さらには下限の閾値(lower threshold)である lt と比べることによって、レコード間の対応付けの有無が判定される。ut と lt は、それぞれ照合の誤りを許容する確率 μ、及び未照合の誤りを許容する確率 λ から導出される。先述のとおり、照合の誤りとは、同一レコードでない組に対して、照合されたと判定される誤りであって、未照合の誤りとは、同一レコードである組に対して照合されないと判定される誤りである。確率 μ と λ が設定されると、ut と lt は次の手順で導出される。

① m/u を降順に並べ替え、m/u が大きい値から順に指標 $\sigma(x)$ を設定する（x はインデックス番号)。

$$\frac{m^{\sigma(1)}}{u^{\sigma(1)}} > \frac{m^{\sigma(2)}}{u^{\sigma(2)}} > \frac{m^{\sigma(3)}}{u^{\sigma(3)}} > \cdots \tag{14}$$

② μ の下限と λ の上限については次の(15)式と(16)式を満たすように算出する。

$$\sum_{i=1}^{limit} u^{\sigma(i)} \leq \mu < \sum_{i=1}^{limit+1} u^{\sigma(i)} \tag{15}$$

$$\sum_{i=limit'}^{|\Gamma|} m^{\sigma(i)} \leq \lambda < \sum_{i=limit'-1}^{|\Gamma|} m^{\sigma(i)} \tag{16}$$

limit: μ の下限に対応するインデックス番号
limit':λ の上限に対応するインデックス番号
|Γ|：インデックス番号の総数
③　②で求めた limit、limit'をもとに、ut と lt を導出する((17)式と(18)式)。

$$ut = \log\left(m^{\sigma(limit)} / u^{\sigma(limit)}\right) \tag{17}$$

$$lt = \log\left(m^{\sigma(limit')} / u^{\sigma(limit')}\right) \tag{18}$$

例えば、μ=0.1、λ=0.2 に設定したとする。そのとき、(17)式と(18)式より、μ と λ の範囲は次のように与えられる((19)式と(20)式)。

$$\sum_{i=1}^{2} u^{\sigma(i)} \left(= 1/20 + 1/20 = 0.1\right) \le \mu < \sum_{i=1}^{3} u^{\sigma(i)} \left(= 1/20 + 1/20 + 2/20 = 0.2\right) \tag{19}$$

$$\sum_{i=4}^{9} m^{\sigma(i)} \left(= 0/5 + \cdots + 0/5 + 1/5 = 0.2\right) \le \lambda < \sum_{i=3}^{9} m^{\sigma(i)} \left(= 0/5 + \cdots + 0/5 + 1/5 + 1/5 = 0.4\right) \tag{20}$$

(19)式と(20)式より μ の下限に対応するインデックス番号は 2，λ の上限に対応するインデックス番号は 4 となる。したがって、μ=0.1、λ=0.2 における lt は 0.69、ut は 1.39 とそれぞれ算出される。これらの閾値と評価式(図 14 における log(m/u))の値から、上記の判別ルールを満たす M のレコード数は LP=3、NP=2、CP=0 となる。したがって、真のリンクと判定されるレコード数は 3 レコードで、その割合は 60.0%となっている。また確率 μ を 0.05 とした場合には、ut=2.08 と求められることから、判別ルールを満たす M のレコード数は LP=2、NP=2、CP=1 となる。このように、LP の値は閾値 ut に依存し、NP の値は閾値 lt に依存する。よって、もし照合誤りを許容する確率 μ を高く設定すれば、閾値 ut は低い値となり、LP と判定されるレコード数は増大する。その一方で、未照合誤りを許容する確率 λ を高く設定すれば、閾値 lt は高い値となり、NP と判定されるレコード数は増大する。

確率的リンケージの特徴は、照合の誤り(false match)と未照合の誤り(false non-match)を許容する上限確率を基準に真のリンクを判定することにある。照合の誤りとは、同一レコードでない組に対して、照合されたと判定される誤りである。一方、未照合の誤りとは、同一レコードである組に対して照合されないと判定される誤りである。確定的リンケージや距離計測型リンケージでは、属性値における一致の程度、さらには、属性間の距離の近さという視点から秘匿性を評価しているのに対して、確率的リンケージにおいては、確率的にどの程度まで照合の誤り、未照合の誤りが許容されるかという観点から秘匿性が評価される。さらに、確率的リンケージでは、量的属性と質的属性のいずれについても、秘匿性の相対的な評価を行うことが可能なことが特徴的だと言える。

(3) クロス集計表による秘匿性の評価

質的属性に匿名化技法を適用して作成した秘匿処理済データの秘匿性を評価する方法として、クロス集計表（分割表）を用いることが考えられる[10]。それは、データに含まれる複数の質的属性を対象に、クロス集計表における分布特性を比較することによって、秘匿性の強度を評価することを指向している。クロス集計表を用いることによって、原データと秘匿処

[10] Domingo-Ferrer and Torra(2001b)によれば、有用性の評価方法の１つとして、分割表による比較が提唱されているが、本研究は、その方法を参考にしながらも、秘匿性の強度を評価するための方法として、クロス集計表による秘匿性の相対比較を行うことにした。

理済データの間で度数が１となるセルの総数を比較し、度数１となるセル数の変化を確認することができる。このような度数１の変化を把握することは秘匿性の評価指標として適用可能だと思われる(Shlomo *et al.*(2010))。

図 15 は、原データと秘匿処理済データのそれぞれについて、世帯主の職業が民間職員である者を対象に、世帯主の年齢と住居の所有関係に関するクロス集計表を作成したものである。秘匿処理済データでは、世帯主の年齢が 80 歳以上の場合にトップコーディングが、住居の所有関係に対してリコーディングが、それぞれ適用されている。例えば、原データでは、住居の所有関係が持ち家で、世帯主の年齢が 80～84 歳に該当するセルの度数は１となっている。それに対して、秘匿処理済データの場合、トップコーディングの適用によって該当する年齢の分類区分が 80 歳以上に統合されているために、それに対応するセルの度数が２になっている。このことは、匿名化技法の適用によって秘匿性の強度が高まっていることを示している。このように、クロス集計表を用いた場合においても、質的属性値における秘匿性の強度を相対的に評価することが可能である。

４．全国消費実態調査による量的属性に関する有用性と秘匿性の定量的な評価―ミクロアグリゲーションを例に―

本節では、量的属性に匿名化技法を適用して作成した秘匿処理済データの有用性と秘匿性の評価について、平成 16 年全国消費実態調査（以下「全消」という。）の原データ（二人以上の世帯、55,056 世帯）を用いて実証研究を行う。本研究では、匿名化技法としてミクロアグリゲーションを適用しているが、ミクロアグリゲーションの手法に関しては、伊藤・磯部・秋山(2008)で展開された方法を採用している。それは、具体的には、最初にレコード群に含まれる質的属性を用いてレコードを層ごとに分け、層内のレコードについて特定のレコード数（あるいは特定の閾値）にしたがってグループ化を行い、グループ内の量的属性値を平均値等の代表値に置き換える方法である。本研究は、量的属性に対するミクロアグリゲーションとして次の２種類の手法を用いている。第１は、データの最初の配列順にしたがってグループ化を行い、グループ内の量的属性値を平均値に置き換える方法である（以下、「ソートなし」という）。

第２は、量的属性ごとにソート化とグループ化を行った上で、グループ内の量的属性値の各々を平均値に置き換える方法である（以下、「個別ランキング法」という）。なお、本研究では、ミクロアグリゲーションの対象となる属性群として、質的属性については性別と住居の所有関係、量的属性に関しては住居の延べ床面積、年間収入、貯蓄現在高、負債現在高と世帯主の年齢をそれぞれ用いている。

(1) 量的属性に関する有用性の評価結果

最初に、全消の原データを用いて行った有用性の評価に関する実証研究の結果を述べることにしたい。伊藤・磯部・秋山(2008)では、相関係数行列をもとに算出した平均平方誤差を情報量損失の指標として設定している。一方、本研究では、相関係数行列だけでなく、原データと秘匿処理済データの属性値や分散共分散行列を用いて情報量損失を計測した。

さらに、本研究では、情報量損失の指標として平均平方誤差、平均絶対誤差と平均変化率の３つの指標が用いられている。その理由としては、有用性を評価する場合、匿名化技法の適用による各レコードの属性値の変化とデータ構造全体の変化のどちらに比重を置くかによって、情報量損失の値が異なると考えられるからである。

図 15　クロス集計表による秘匿性の評価のイメージ

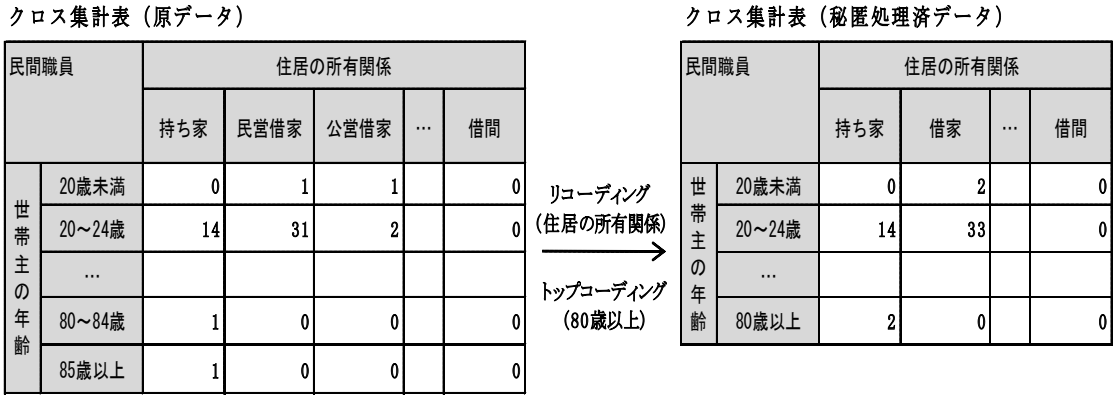


表２は、ソートなしと個別ランキング法の２つの手法について有用性を評価した結果を比較したものである。属性値による有用性の評価では、標準化した属性値が使用されている。また、分散共分散行列も、標準化した属性値を用いて算出されている。表２を見ると、平均平方誤差、平均絶対誤差、平均変化率の中では、個別ランキング法とソートなしのいずれの手法においても、平均変化率で計測した情報量損失の値が最も高いことが分かる。特に、ソートなしにおいては、原データと秘匿処理済データの属性値をもとに計測した平均変化率が、平均変化率の値の中では最大の値になっている。その理由としては、特にソートなしの手法においては、匿名化技法の適用によって属性値が大きく増加したレコードが見られたために、平均変化率が高い値を示したと考えられる。その一方で、本分析結果では、属性値や分散共分散行列を用いた場合に、情報量損失の値に違いが見られることがわかった。このことから、有用性を定量的に評価する際には、原データと秘匿処理済データにおけるレコード単位の属性値の差、あるいは、属性間の相関関係における差異という観点の違いを考慮した上で、評価指標を決定することが有効ではないかと考えられる。

(2) 量的属性に関する秘匿性の評価結果

次に、本実験で行った秘匿性の評価に関する実証研究の結果を紹介することにしたい。本研究は、秘匿性の評価手法として、確定的リンケージと距離計測型リンケージを試みている。本研究では、リンケージを行うためのリンクキー変数として、ミクロアグリゲーションで用いた住居の所有関係等の質的属性２属性と世帯主の年齢等の量的属性５属性を用いている。また、距離計測型リンケージで使用する距離として、①ユークリッド距離（属性値の標準化）、②ユークリッド距離（属性間の距離の標準化）、③マハラノビス距離[11]の３種類を用いて実験を行っている。さらに、本研究においては、秘匿処理済データに含まれるレコードの属性値を中心とした区間を設定した上で、原データにおいて対応するレコードの属性値が区間の範囲内に含まれる比率についても計測した。なお、使用する区間として、秘匿処理済データの各属性における標準偏差１％～10%(１％刻み)を設定した。

表３は、ミクロアグリゲーション(ソートなし及び個別ランキング法)で作成した秘匿処理済データについて、確定的リンケージ及び距離計測型リンケージを用いた場合の真のリンク

11 本研究では、マハラノビス距離を計測するために原データの分散共分散行列を用いて計算を行っている。その理由は、本研究における距離計測型リンケージでは、秘匿処理済データに含まれる特定のレコードから原データに含まれる各々のレコードとの間の距離を計測することが指向されているからである。

表２　秘匿処理済データにおける有用性の評価

有用性評価の指標		ソートなし	個別ランキング法
属性値	平均平方誤差	0.6432	0.0041
	平均絶対誤差	0.5284	0.0031
	平均変化率	4.0037	0.0062
相関係数行列	平均平方誤差	0.0041	0.0000
	平均絶対誤差	0.0574	0.0005
	平均変化率	0.4199	0.0036
分散共分散行列	平均平方誤差	0.0028	0.0000
	平均絶対誤差	0.0383	0.0004
	平均変化率	0.2799	0.0024

表３　真のリンクと判定されたレコード数（括弧内は全レコードに対する比率）

評価方法		ソートなし	個別ランキング法
確定的リンケージ		0（0.00%）	32,426（58.90%）
距離計測型リンケージ			
	ユークリッド距離（属性値の標準化）	116（0.21%）	54,272（98.58%）
	ユークリッド距離（属性間の距離の標準化）	141（0.26%）	54,276（98.58%）
	マハラノビス距離	141（0.26%）	54,276（98.58%）

と判定されたレコード数と全レコード(55,056 レコード)に対する真のリンクの比率を示している。真のリンクと判定されたレコード数が少ないほど、秘匿性の強度が高いと判定することができる。表３によると、ソートなしよりも個別ランキング法で作成したミクロアグリゲートデータのほうが、真のリンクと判定されるレコード数に関して、著しく高い値を示していることが分かる。このことは、本研究において個別ランキング法で作成したミクロアグリゲートデータの分布特性が、ソートなしの手法による分布特性と比較して、原データと非常に近似している可能性を示唆している(伊藤・磯部・秋山(2009, 23～24 頁))。また、表３について、評価方法の相違という観点から実験結果を見ていくと、次の３点を指摘することができる。

第１に、真のリンクと判定されたレコード数は、確定的リンケージよりも距離計測型リンケージで評価を行った場合のほうが高い値を示している。このことは、距離計測型リンケージで使用した距離のいずれについても同様の結果が得られている。その理由として、真のリンクの判定基準の相違に起因することが考えられる。確定的リンケージでは、原データと秘

匿処理済データにおけるレコードの属性値の完全な一致という基準によって、真のリンクの有無が判定される。一方、距離計測型リンケージでは、属性値が完全に一致しなくても、秘匿処理済データの特定のレコードが原データにおける元のレコードに対応付けられ、レコード間の距離が原データのレコードの中で最も短ければ、それは真のリンクと判定される。したがって、確定的リンケージについては距離計測型リンケージと比較して真のリンクの判断基準が厳しいと言うことができ、そのことが、表３での実験結果に表れている。

第２に、ユークリッド距離による距離計測型リンケージにおいて、属性値の標準化を行った場合と属性間の距離を標準化した場合では、真のリンクとなる比率にほとんど違いが見られなかった。特に個別ランキング法においては、その特徴が顕著に表れている。

第３に、距離計測型リンケージにおいて、ユークリッド距離とマハラノビス距離では結果の数値に大きな差異が存在しないことがわかる。その理由として、実証研究に使用した量的属性間の相関性が弱いことが考えられる。表４は、対象となる５つの量的属性に関する相関係数行列を示したものである。例えば、年間収入と貯蓄現在高の相関係数、及び年間収入と負債現在高の相関係数は、それぞれ 0.33 と 0.29 であって、相関関係が強いとは言いがたい。

表５は秘匿処理済データのレコード上にある属性群を対象に、属性値を中心とした区間（各属性の標準偏差１％～10％（１％刻み））を設定した上で、その区間の範囲内に原データにおける元のレコードの属性値が含まれる比率を示している[12]。この比率が高いほど、秘匿性の強度が低いと判定することができる。表５を見ると、ソートなしの場合、区間を標準偏差の10％に広く設定しても、原データにおいて該当するレコードの比率は 0.08％と低い値を示している。一方、個別ランキング法では、区間を標準偏差の１％と狭く設定しても、原データにおいて該当するレコードの比率は 93.64％と高い値を示しており、原データと秘匿処理済データが近似していることが分かる。標準偏差を１％～10％に設定して得られた 10 個の結果について平均値を取った場合においても、秘匿処理済データのレコードにおける属性値の各々に対して設定された区間内に、原データにおける元のレコードの属性値が含まれる比率は、ソートなしでは 0.03％、個別ランキング法では 97.58％という値を示した。このような区間を指標とした秘匿性の評価方法では、リンケージによる手法とは異なり、区間幅の設定を変えることによって、その区間内に含まれるレコード数がどの程度存在するのかを把握することが可能である。

次に、確率的リンケージを用いて秘匿性を評価した結果を紹介することにしたい。本研究では、平成 16 年全国消費実態調査の原データを用いている。本研究では、確定的リンケージや距離計測型リンケージと同様、匿名化技法としてミクロアグリゲーション(ソートなし、個別ランキング法)を適用している。

確率的リンケージで層別に評価を行う場合、閾値が各層で異なるために、確率μの変更による各層の閾値の導出が複雑になる。よって、本研究では、住居の所有関係が「持ち家(世帯主名義)」で性別が「男」であるレコード群(40,517 レコード)を対象に実験を行っている。また、本研究では、確率的リンケージにおける照合誤りの許容確率μを 0.0001(0.01%)、0.001(0.1%)、0.01(1%)の３種類に設定し、真のリンクとなるレコードとして対応付けされるペア(LP)となるレコードについての比率を算出した。

表６は、確率的リンケージで秘匿性を評価した結果を示している。表６においては、本研究の対象となったレコード群に対して、確定的リンケージと距離計測型リンケージを用いた

[12] 表 5 は、原データと秘匿処理済データに含まれる属性群を対象にしたとき、秘匿処理済データのレコードにおける属性値の各々に対して設定された区間内に、原データにおける元のレコードのすべての属性値が含まれた場合のレコード数を示している。

表４　原データに含まれる量的属性間の相関係数

	延べ床面積	年間収入	貯蓄現在高	負債現在高	世帯主の年齢
延べ床面積	1.00				
年間収入	0.23	1.00			
貯蓄現在高	0.19	0.33	1.00		
負債現在高	0.06	0.29	-0.04	1.00	
世帯主の年齢	0.18	-0.05	0.28	-0.19	1.00

表５　区間設定による秘匿性の評価―区間の範囲内に含まれるレコード数（括弧内は全レコードに対する比率）

区間幅	ソートなし	個別ランキング法
標準偏差１％	9（0.02%）	51,557（93.64%）
標準偏差２％	9（0.02%）	52,377（95.13%）
標準偏差３％	9（0.02%）	53,460（97.10%）
標準偏差４％	10（0.02%）	53,635（97.42%）
標準偏差５％	11（0.02%）	54,181（98.41%）
標準偏差６％	13（0.02%）	54,272（98.58%）
標準偏差７％	15（0.03%）	54,326（98.67%）
標準偏差８％	20（0.04%）	54,415（98.84%）
標準偏差９％	23（0.04%）	54,462（98.92%）
標準偏差 10%	42（0.08%）	54,535（99.05%）
平均	16（0.03%）	53,722（97.58%）

表６　確率的リンケージによる秘匿性の評価及び他の手法との比較―真のリンクに該当するレコード数及び真のリンクの比率

		ソートなし	個別ランキング法
確率的リンケージ（μ=0.0001）		66（0.16%）	39,320（97.05%）
確率的リンケージ（μ=0.001）		483（1.19%）	40,063（98.88%）
確率的リンケージ（μ=0.01）		665（1.64%）	40,370（99.64%）
参考	確定的リンケージ	0（0.00%）	28,877（71.27%）
	距離計測型リンケージ		
	①ユークリッド距離（属性値を標準化）	38（0.09%）	39,985（98.69%）
	②ユークリッド距離（属性間の距離を標準化）	27（0.07%）	39,987（98.69%）
	③マハラノビス距離	25（0.06%）	39,988（98.69%）
	区間設定による評価	10（0.02%）	40,232（99.30%）

注 住居の所有関係が「持ち家(世帯主名義)」で性別が「男」である 40,517 レコードを対象

場合に、真のリンクとなったレコード数、およびそれが全レコード数に占める比率を併せて載せている。また、表６では、秘匿処理済データにおいて属性値ごとに区間設定を行った場合の該当するレコード数とその比率も示されている。なお、表６の区間設定による評価に関しては標準偏差１％～10%(１％刻み)を用いて得られた結果の平均値が掲載されている。

表６を見ると、ソートなしと個別ランキング法のいずれについても、確率μの値が大きくなるにつれて、真のリンクとなるレコードの比率が上昇していることがわかる。ソートなしでは、μ=0.0001 の場合に真のリンクの比率は 0.16%であるが、μ=0.01 では、真のリンクの比率は 1.64%となっている。このように、確率的リンケージにおける真のリンクの比率が、確定的リンケージや距離計測型リンケージにおけるそれと比較して高い数値を示していることは、興味深い結果だと言えよう。

５．全国消費実態調査を用いた質的属性に関する有用性と秘匿性の比較分析

前節では、ミクロデータに含まれる量的属性を対象に、有用性と秘匿性の定量的な評価に関する分析結果を紹介した。本節では、質的属性に対して匿名化技法を適用することによって作成した秘匿処理済データの有用性と秘匿性について、全消の原データ（二人以上の世帯：55,056 世帯）を用いて定量的な評価を試みる。本研究では、質的属性として世帯主の年齢と住居の所有関係を使用する。そして、世帯主の年齢と住居の所有関係についてそれぞれ匿名化技法を適用することによって作成した秘匿処理済データを対象に、有用性と秘匿性の評価に関する実験を行った。

本研究においては、匿名化技法として、世帯主の年齢については、リコーディング（各歳区分を５歳階級区分に統合）とトップコーディング（85 歳以上の各歳区分を統合）を採用している。また、住居の所有関係に関しては、リコーディング(８区分を２区分に分類区分の統合）を行った。住居の所有関係におけるリコーディング後の分類区分(統合区分)は、結果表の集計事項で用いられる区分に基づいて、表７のように設定されている。また、表８は、それぞれ本研究で使用した匿名化技法の組合せを表している。表８に示されるとおり、世帯主の年齢と住居の所有関係について、匿名化技法の組み合わせのパターンは８パターンとなっている。

次に、本研究では、質的属性における有用性と秘匿性を定量的に評価するために、有用性については情報エントロピーを利用した情報量損失を、秘匿性についてはクロス集計表におけるセルの中で度数１となるセル数の減少率を、それぞれ算出した。度数１の減少率は、原データにおける度数１の数を基準にしたときに、匿名化技法の適用によって度数１が減少する比率を示したものである。なお、度数１の減少率は、次の(21)式によって算出されている。

$$\text{度数1の減少率}(\%)=\frac{\text{度数1のセル数〔原データ〕}-\text{度数1のセル数〔秘匿処理済データ〕}}{\text{度数1のセル数〔原データ〕}}\times 100 \qquad (21)$$

表９は世帯主の年齢と住居の所有関係に対して匿名化技法を適用した場合の有用性に関する実証研究の結果である。表９では、原データ（パターン〔A〕）を有用性評価の基準値（情報量損失＝０）として設定し、匿名化技法の適用による情報量損失の程度を計測した。また、本研究では、各属性の全ての区分を１つに統合したときに算出される情報量損失の値を仮想的な情報量損失の最大値と設定した。世帯主の年齢と住居の所有関係の分類区分をそれぞれ

表７　住居の所有関係の原区分と統合区分

住居の所有関係			
原区分		統合区分	
1	持ち家(世帯員名義)	1	持ち家
2	持ち家(その他名義)		
3	民営賃貸住宅（設備専用）	2	借家・借間
4	民営賃貸住宅（設備共用）		
5	県市区町村営賃貸住宅		
6	都市再生機構・公社等賃貸住宅		
7	社宅・公務員住宅		
8	借間		

表８　世帯主の年齢と住居の所有関係に対する匿名化技法の組合せ

世帯主の年齢		住居の所有関係
リコーディング	トップコーディング	リコーディング
各歳	なし	８区分
各歳	85歳以上	８区分
各歳	なし	２区分
各歳	85歳以上	２区分
５歳階級	なし	８区分
５歳階級	85歳以上	８区分
５歳階級	なし	２区分
５歳階級	85歳以上	２区分

表９　情報エントロピーを利用した情報量損失〔世帯主の年齢×住居の所有関係〕による有用性の評価

	世帯主の年齢		住居の所有関係	情報量損失	情報量損失率（%）
	リコーディング	トップコーディング	リコーディング		
〔A〕	各歳	なし	８区分	0	0.0
〔B〕	各歳	85歳以上	８区分	859	−0.2
〔C〕	各歳	なし	２区分	25,229	−6.7
〔D〕	各歳	85歳以上	２区分	26,114	−6.9
〔E〕	５歳階級	なし	８区分	126,421	−33.6
〔F〕	５歳階級	85歳以上	８区分	126,670	−33.7
〔G〕	５歳階級	なし	２区分	151,958	−40.4
〔H〕	５歳階級	85歳以上	２区分	152,220	−40.4

１つに統合した場合、仮想的な情報量損失の最大値は 445,017 となる。この情報量損失の最大値に対する情報量損失の比率(情報量損失率)[13]が表９に示されている。例えばパターン〔H〕のように、世帯主の年齢について、５歳階級のリコーディング及び 85 歳以上のトップコーディングを行い、住居の所有関係を２区分にリコーディングした場合は、情報量損失の値は 152,220 と算出されるが、情報量損失率は-40.4%となっている。

表９を見ると、世帯主の年齢に対して 85 歳以上のトップコーディングを原データに適用した場合には、それを適用しなかった場合と比較しても、情報量損失は大きく変わっていないことがわかる。例えば、パターン〔C〕と〔D〕を比較すると、情報量損失の違いはわずか 0.2%である。それに対して、世帯主の年齢にリコーディングを施すと、情報量損失率は、全般的に約 30%程度変化していることがわかる。このことは、各歳区分から５歳階級区分への区分統合が、秘匿処理済データの有用性に与える影響が大きいことを示唆している。さらに、パターン〔B〕と〔D〕における情報量損失の割合の変化に見られるように、住居の所有関係を８区分から２区分に分類区分を統合することによっても情報量損失の程度が大きいことが明らかになっている。

表 10 は世帯主の年齢と住居の所有関係に対して匿名化技法を適用したことによる秘匿性の相対的な評価結果を表したものである。表 10 における度数１の減少率は、パターン〔A〕(原データ)における度数１の数を基準とした場合に、匿名化技法の適用によって度数１が減少した割合を表している。よって、例えばパターン〔H〕のように、世帯主の年齢に対してリコーディング及びトップコーディングを適用し、さらに住居の所有関係に対してリコーディングを行った場合、度数１の減少率は 100.0%となっている。表 10 を見ると、世帯主の年齢あるいは住居の所有関係にリコーディングを施すことによって、度数１の減少率が大きくなっていることがわかる。

図 16 は、世帯主の年齢と住居の所有関係に匿名化技法を適用して作成した秘匿処理済データの有用性と秘匿性の関係を、R-U マップ(R-U confidentiality map)(Duncan *et al.*(2001))を参考に図示したものである。図 16 における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、図 16 でプロットされた〔A〕～〔H〕の点は、それぞれ表９及び表 10 における８パターンの匿名化技法の組合せ〔A〕～〔H〕の情報量損失率と度数１の減少率を図示したものである。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。図 16 のように有用性と秘匿性の関係を R-U マップにプロットすることによって、有用性と秘匿性が概ねトレードオフの関係にあることが視覚的に把握できる。

図 16 において注目すべき点は、パターン〔C〕及び〔D〕に関しては、情報量損失率が相対的に低いだけでなく、度数１の減少率が高い位置にプロットされていることである。すなわち、他の匿名化技法のパターンと比較して、有用性が高いだけでなく、秘匿性の強度も高いということができる。パターン〔C〕では、住居の所有関係に対してリコーディングが適用されている。また、パターン〔D〕については、住居の所有関係に対するリコーディングに加えて、世帯主の年齢に対するトップコーディングが施されている。このことから、秘匿性の強度が高くなった要因として、住居の所有関係に対してリコーディングを適用した場合、それが秘匿性の強度を高めることに対して大きな影響を及ぼすことが考えられる。一方、パターン〔C〕及び〔D〕については、原データと比較して情報量損失が非常に小さく、有用

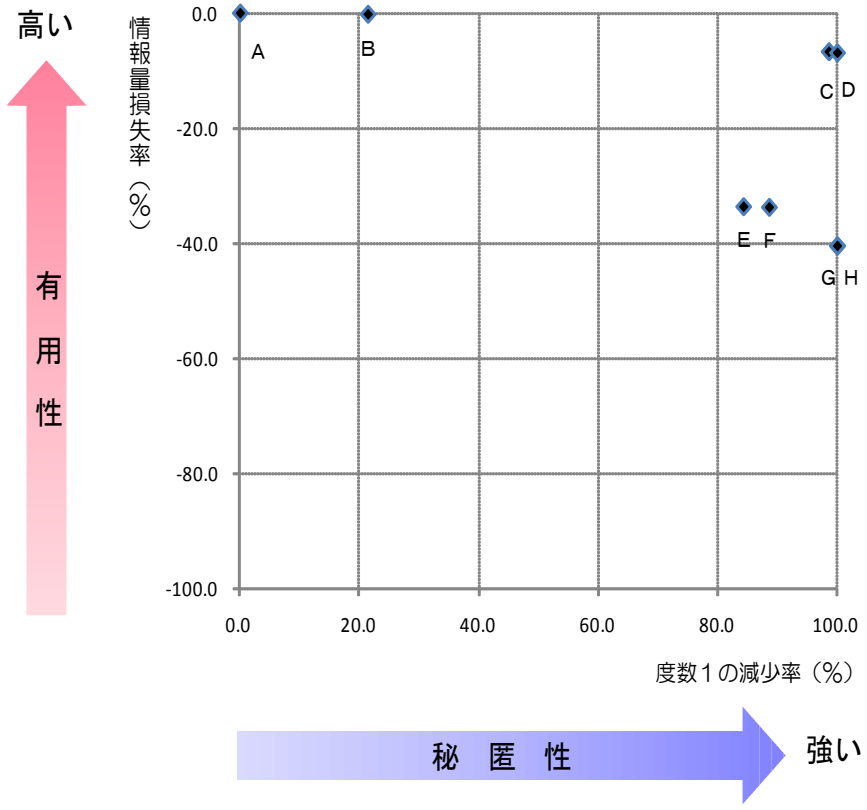
[13] 情報量損失率は、次式によって計算される。

$$\text{情報量損失率} = -\frac{\text{情報量損失}}{\text{情報量損失の最大値}} \times 100$$

表 10　クロス集計表〔世帯主の年齢×住居の所有関係〕による秘匿性の評価

	世帯主の年齢		住居の所有関係	度数１の減少率（%）
	リコーディング	トップコーディング	リコーディング	
〔A〕	各歳	なし	８区分	0.0
〔B〕	各歳	85歳以上	８区分	21.4
〔C〕	各歳	なし	２区分	98.6
〔D〕	各歳	85歳以上	２区分	100.0
〔E〕	５歳階級	なし	８区分	84.3
〔F〕	５歳階級	85歳以上	８区分	88.6
〔G〕	５歳階級	なし	２区分	100.0
〔H〕	５歳階級	85歳以上	２区分	100.0

図 16　世帯主の年齢と住居の所有関係に関する R-U マップ



	年齢		住居
	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（２区分）
〔A〕			
〔B〕		○	
〔C〕			○
〔D〕		○	○
〔E〕	○		
〔F〕	○	○	
〔G〕	○		○
〔H〕	○	○	○

注　本図におけるパターン〔G〕と〔H〕の各点については、R-U マップ上のほぼ同じ位置にプロットされていることに留意されたい。

性が保持されている。この要因として、世帯主の年齢に対してリコーディングを適用しないことによって有用性の相対的な低下に大きな影響を及ぼさないことが推測できる。

R-U マップを用いることは、匿名化技法の違いによって、秘匿処理済データの有用性と秘匿性がどのように変化するのかを捉えることが可能になることから、有用性と秘匿性の両面を考慮した匿名化技法を検討する上では有効ではないかと考える。さらに、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 マップにプロットされる位置を確認して有用性と秘匿性の変化を捉えることは、様々な匿名化技法を比較・検討する上で参考になると思われる。こうしたことから，R-U マップを用いて分析を行うことは、特定の匿名化技法の適用がミクロデータにおける有用性と秘匿性にどの程度大きな影響を与えるかが視覚的に捉えられる点で、有効な方法であると言うことができよう[14]。

6. 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―[15]

本節では、家計調査の個別データを例に、攪乱的な手法を適用した場合の匿名化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。

(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[16]、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[17]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

---

14 本研究では、クロス集計表のセルにおいて度数１だけでなく度数２も含めた場合の秘匿性の評価も試みている。それによれば、世帯主の年齢と住居の所有関係の組合せにおける度数１、２の減少率に関しては、度数１のみの減少率と比較しても、分布特性がほとんど変わらないことがわかった。したがって、度数１、２の減少率をもとに作成した R-U マップにおける分布も、度数１のみの減少率をもとに作成した R-U マップと基本的には変わっていない。

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

16 加法ノイズの数理的な特徴については、「補論 ミクロデータにおける加法ノイズの適用について」を参照。

17 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))。

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、次に、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)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、②ミクロアグリゲーションとカテゴリー化[18]、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。

次に、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 歳以上は階級値として 92 歳に設定されている。

表 11 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

[18] 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 11　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

果を示したものである。表 11 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 12 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 12 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 17 は、表 11 と表 12 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 17 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 17 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。

(2)匿名化されたパネルデータの有用性と秘匿性の検証

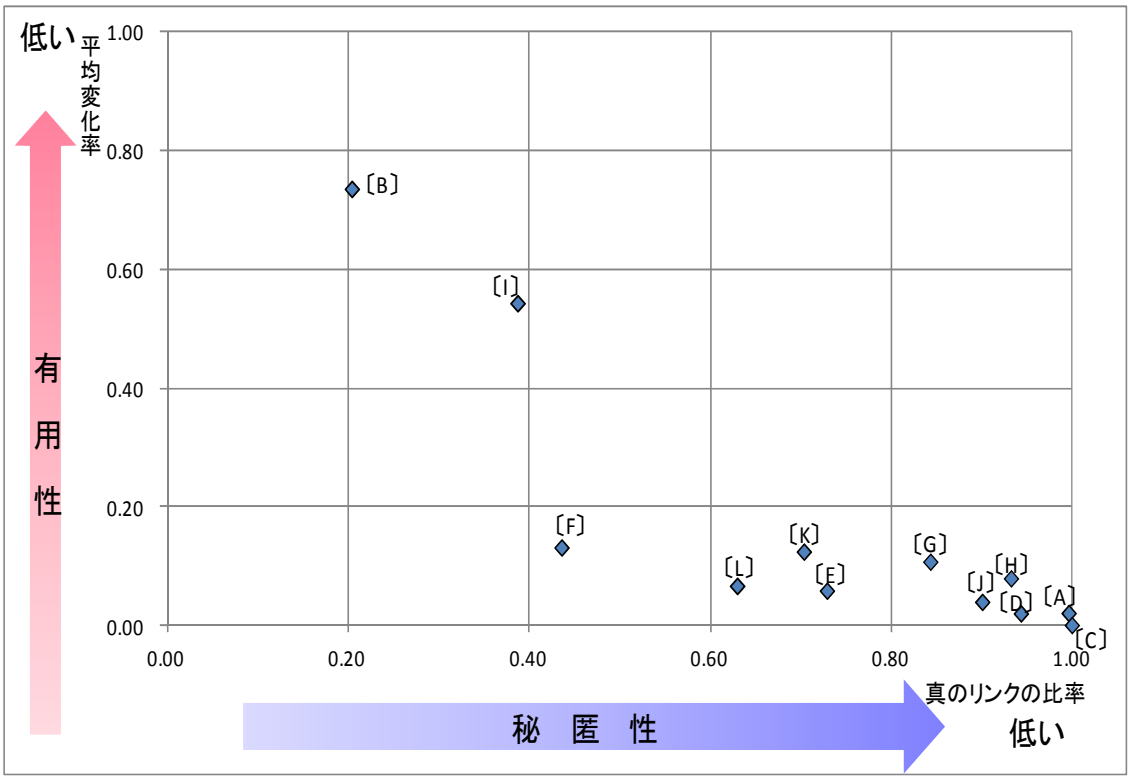
本研究においては、秘匿処理が施されたパネルデータ(以下「秘匿処理済パネルデータ」と呼ぶ。)を試行的に作成し、その有用性と秘匿性を評価する。具体的には、本研究では、2 か月間の同一個体のレコードをリンクしたデータ(以下「2 か月パネル」と呼ぶ。)に、ミクロアグリゲーション、ノイズ付加等を適用した秘匿処理済パネルデータに対して、有用性と秘匿性の定量的な評価を行った。本研究で使用したデータは、平成 21 年家計調査の 1 月と 2 月の勤労者世帯(3,427 世帯)に関する 2 か月パネルである。秘匿処理済パネルデータの作成のために、加法ノイズだけでなく、ミクロアグリゲーションとカテゴリー化といった 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

表 12　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 17 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 11 と表 12 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	

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[19]。

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

次に、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 13 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 14 は、秘匿性の評価の結果を示している。表 14 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 14 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。

７．おわりに

本稿では、量的属性と質的属性を対象に、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実験を行った。ミクロデータの有用性に関しては、量的属性だけでなく、質的属性についても情報量損失を計測することによって、その有用性を定量的に評価することが可能なことが明らかになった。また、ミクロデータの秘匿性については、量的属性においては、リンケージ技法による秘匿性の評価が展開可能であること、質的属性に対しては、クロス集計表を用いた秘匿性の相対的な評価が有効であることがわかった。さらに、本研究では、有用性と秘匿性の定量的な評価結果をもとに R-U マップを試行的に作成することによって、各種匿名化技法によるミクロデータの有用性と秘匿性を相対的に比較する上で、R-U マ

[19] 秘匿処理済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。

表 13　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
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＋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

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＋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 13 続き
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＋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 14　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

ップが有効な手法であることが確認できた。

また，本稿では、ノイズ等の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。さらに、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、我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家計調査のパネルデータについて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。

政府統計の匿名データの作成・提供に関するニーズは今後より一層高まることが考えられる。その意味では、本稿において、匿名化技法の適用可能性について定量的な評価方法を提示することは、ミクロデータの作成に関する実証研究という側面だけでなく、実務の面でも有益であると思われる。

補論　ミクロデータにおける加法ノイズの適用について

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。

原データに対して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(A1)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)。

$$\mathbf{V}' = \mathbf{AVB} + \mathbf{C} \qquad \text{(A1)}$$

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、(A1)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、(A1)式における行列 **C** の分布構造に影響を与える。

次に、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))[20]。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 $\mathbf{V}$ 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、

$$\mathbf{V} \sim (\boldsymbol{\mu}, \sum) \tag{A2}$$

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、次の(A3)式となる。

$$\mathbf{V}' = \mathbf{V} + \boldsymbol{\varepsilon} \tag{A3}$$

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

$$\boldsymbol{\varepsilon} \sim (\mathbf{0}, c\sum) \text{(c はパラメータ)} \tag{A4}$$

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。

$$E(\mathbf{V}') = E(\mathbf{V}) + E(\boldsymbol{\varepsilon}) = E(\mathbf{V}) \tag{A5}$$

$$Var(\mathbf{V}') = Var(\mathbf{V}) + Var(\boldsymbol{\varepsilon}) = (1+c)Var(\mathbf{V}) \tag{A6}$$

(A5)式と(A6)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、(A6)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[21]、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[22]。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。

---

[20] 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。

[21] (A6)式を変形すると、

$$Var(\mathbf{V}) = \frac{Var(\mathbf{V}')}{1+c} \tag{A6)'}$$

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。

[22] 秘匿処理済データの利用者にとっては(A6)'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(A6)'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。

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.

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.

Fellegi, I. P., and Sunter, A. B.(1969) "A Theory for Record Linkage," *Journal of the American Statistical Association*, Vol.64, No.328, pp.1183-1210.

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.

Herzog, T. N., Scheuren, F. J., Winkler, W. E.(2007) *Data Quality and Record Linkage Techniques*, Springer, New York.

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
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.
伊藤伸介・星野なおみ(2013)「匿名化技法としてのスワッピングの可能性について―国勢調査ミクロデータを用いた有用性と秘匿性の実証研究―」,『製表技術参考資料』No.24, 1～58 頁
Jaro, M. A.(1989) "Advances in Record-Linkage Methodology as Applied to the Matching 1985 Census of Tampa, Florida" *Journal of the American Statistical Association*, Vol. 84, No.406, pp.414–420.
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.
Kooiman, P., L. Willenborg and J. Gouweleeuw (1998) "PRAM: A Method for Disclosure Limitation of Microdata", *Research Paper*, No. 9705, Statistics Netherlands, Voorburg.
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.
Mateo-Sanz, J. M, Sebé, F, and Domingo-Ferrer, J.(2004) "Outlier Protection in Continuous Microdata Masking". Domingo-Ferrer, J. and Torra, V. (eds.) *Privacy in Statistical Databases: CASC Project Final Conference, PSD 2004 Barcelona, Catalonia, Spain, June 9-11, 2004 : Proceedings*, Springer, Berlin, pp.201-215.
Mateo-Sanz, J. M., Domingo-Ferrer, J. and Sebé, F.(2005) "Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata" *Data Mining and Knowledge Discovery*, vol.11, pp.181-193.
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.
Shannon, C. E.(1948) "A Mathematical Theory of Communication", *The Bell System Technical Journal*, Vol. 27, pp. 379-423.
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.(1999) “Local Recoding by Maximum Weight Matching for Disclosure Control of Microdata sets”, *ITME Discussion Paper*, No.11, Faculty of Economics, Univ. of Tokyo.
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.
竹村彰通(2003)「個票開示問題の研究の現状と課題」『統計数理』第 51 巻 第 2 号,241～260 頁
Torra, V. and Domingo-Ferrer, J. (2003) “Record Linkage Methods for Multidatabase Data Mining”, Torra, V. (ed.) *Information Fusion in Data Mining*, Springer, Berlin, pp.101-132.
Torra, V., Abowd, J. and Domingo-Ferrer, J (2006) “Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment”, Domingo-Ferrer, J. and Franconi, L.(eds.) *Privacy in Statistical Databases: CENEX-SDC Project International Conference, PSD 2006 Rome, Italy, December 13-15, 2006 : Proceedings*, Springer, Berlin, pp.233–242.
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.