製表技術参考資料 14

ISSN 1348－7345

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

NSTAC

Working Paper No.14

平成 22 年 9 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

要旨……1
１．はじめに……3
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……4
(1) 統計指標を用いた有用性の評価……4
(2) 質的属性値間の距離の計測……5
(3) 情報エントロピーを用いた有用性の評価……6
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……10
(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……11
(2) 確率的リンケージによる秘匿性の評価……15
(3) クロス集計表による秘匿性の評価……16
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……17
(1) 量的属性に関する有用性の評価結果……17
(2) 量的属性に関する秘匿性の評価結果……18
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……22
(1) 質的属性に関する有用性と秘匿性の比較―世帯主の年齢と職業を例に―……24
(2) 質的属性に関する有用性と秘匿性の比較―世帯主の年齢と住居の所有関係の場合―……27
６．むすびにかえて……29
参考文献……30
付録　確率的リンケージによる秘匿性の評価方法について……33

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

伊藤　伸介*，高野　正博**，秋山　裕美**，後藤　武彦**

要　旨

現在、我が国では、４調査の政府統計(公的統計)のミクロデータ(匿名データ)が作成・提供されているが、これらについては、１つのタイプの匿名データのみが利用可能となっている。しかし、諸外国では、利用者のニーズに応じて、様々なタイプのミクロデータが提供されていることから、我が国でも、将来的には利用者のニーズに対応した形で、複数のタイプのミクロデータが作成されることが考えられる。その場合、ミクロデータにおける有用性や秘匿性の定量的な評価を行うことができれば、その評価結果は、ミクロデータの作成における数量的な参考情報として有意義だと思われる。本稿は、ミクロデータの有用性と秘匿性に関する評価方法を検討し、公的統計の個別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を試みたものである。

本稿は、秘匿処理を施していない個別データ(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(秘匿処理済データ)の有用性と秘匿性について、ミクロデータに含まれる量的属性と質的属性の両面からその評価方法を検討している。本研究では、有用性の評価方法として、量的属性については平均平方誤差や平均変化率等を用いることによって、原データに対する秘匿処理済データの情報量損失の計測を追究している。また、質的属性における有用性の評価に関しては、情報エントロピーに基づいて情報量損失を測ることに焦点を当てている。一方、秘匿性の評価方法に関しては、量的属性については、レコードリンケージの手法を用いて秘匿性の相対的な強度を比較することを指向している。それに対して、質的属性における秘匿性の評価においては、本研究では、原データと秘匿処理済データを対象にクロス集計表を作成し、秘匿性の観点から２つの集計表における分布特性を比較することを提案している。

次に、本稿は、平成 16 年全国消費実態調査の原データを利用して、有用性と秘匿性の定量的な評価に関する実証研究を行っている。原データに適用する匿名化技法として、本研究では、量的属性についてはミクロアグリゲーションを、質的属性に関してはトップコーディング及びリコーディングを、それぞれ適用した。本分析結果によれば、ミクロデータの有用性に関しては、量的属性と質的属性のいずれについても、情報量損失を計測することによって、その有用性が評価可能なことがわかった。また、ミクロデータの秘匿性に関しては、量的属性についてはリンケージ技法による秘匿性の評価方法が有効であること、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可能なことが明らかになった。

*　統計センター情報技術部研究主幹非常勤研究員（明海大学経済学部専任講師）
**　統計センター情報技術部研究主幹（E-mail: research@nstac.go.jp）

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

伊藤伸介，高野正博，秋山裕美，後藤武彦

１．はじめに

我が国では、統計法の全面施行に伴い、平成 21 年４月より、政府統計(公的統計)のミクロデータ(匿名データ)の提供が実施されている。総務省統計局では４調査の匿名データが提供されているが[1]、それらについては、１つのタイプの匿名データ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知られている。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 Sample)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を作成・提供することが考えられる。その場合、ミクロデータにおける有用性や秘匿性の定量的な評価を行うことができれば、その評価結果は、ミクロデータの作成における数量的な参考情報として有意義だと思われる。

これまで、伊藤・磯部・秋山(2008)は、匿名化技法の１つであるミクロアグリゲーション(microaggregation)に焦点を当て、ミクロデータの有用性の観点から、平均平方誤差等の情報量損失の評価を行うことによって、ミクロアグリゲーションの有効性の検証を行った。また、伊藤・磯部・秋山(2009)は、秘匿性の相対的な程度(秘匿性の強度)に関する定量的な評価方法を検討するために、ミクロアグリゲーションによって作成したデータ(以下「ミクロアグリゲートデータ」と呼ぶ。)を用いて、レコードリンケージ(record linkage)による秘匿性の評価方

[1] 総務省統計局では，現在，就業構造基本調査（平成４年，９年，14 年），全国消費実態調査（平成元年，６年，11 年，16 年），社会生活基本調査（平成３年，８年，13 年），住宅・土地統計調査（平成５年，10 年，15 年）の４調査が提供されている。

法を提唱した。しかし、ミクロデータに対する匿名化技法には、ミクロアグリゲーションだけでなく、トップコーディング、リコーディング(再符号化)等の様々な方法が存在しており、匿名化技法によっては、これまで検討してきた評価方法では有用性と秘匿性が十分に計測できない場合がある。さらに、一般に、ミクロデータには量的属性と質的属性が含まれていることから、属性の性質に即した形でミクロデータの有用性と秘匿性を評価することが求められる。

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、公的統計の個別データを用いて、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。

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り、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る[2]。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[3]。

(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))。

---

[2] Domingo-Ferrer and Torra(2001a)は、次の統計指標を用いて原データと秘匿処理済データとの間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))。
①共分散行列
②相関係数行列
③属性値と主成分分析から得られたそれぞれ因子との間の相関係数行列
④属性値の各々と第１主成分(それ以外の主成分)とのコモナリティ(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因子スコア係数行列(factor score coefficient matrix)

[3] 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

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。

シャノン情報量 ＝ $-\log p (0 \leq p \leq 1)$ ・・・（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。

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

n：事象の数
p_i：i 番目の事象が起こる確率

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

図１　シャノン情報量と確率の関係

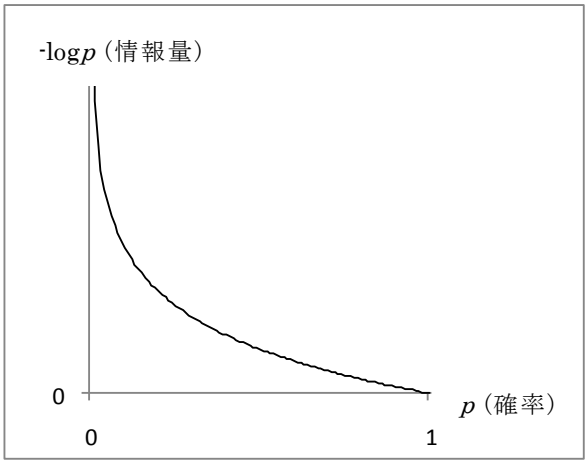


図２　匿名化技法の適用による移行確率の考え方

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

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、

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

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」に移行した場合の情報量損失は、

$0.81128\times400=324.51\cdots(4)$

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)。

$$\text{情報量損失}=\left(-\frac{600}{1000}\log_2\frac{600}{1000}-\frac{300}{1000}\log_2\frac{300}{1000}-\frac{100}{1000}\log_2\frac{100}{1000}\right)\times1000=1295.46\cdots(5)$$

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

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

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。

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

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

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

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。本節では、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ることができる。

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

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

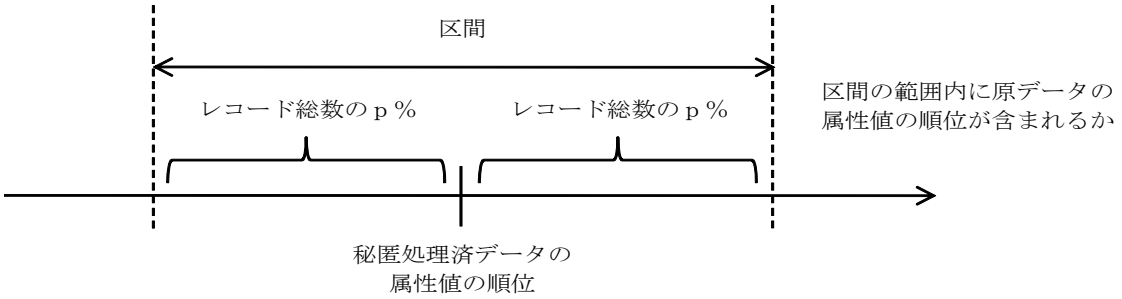


図９　標準偏差に基づいた区間を用いた評価

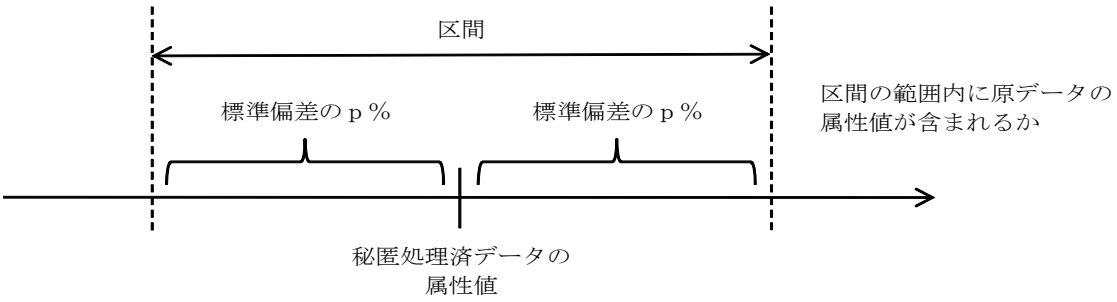


図 10　距離計測型リンケージのイメージ

原データ

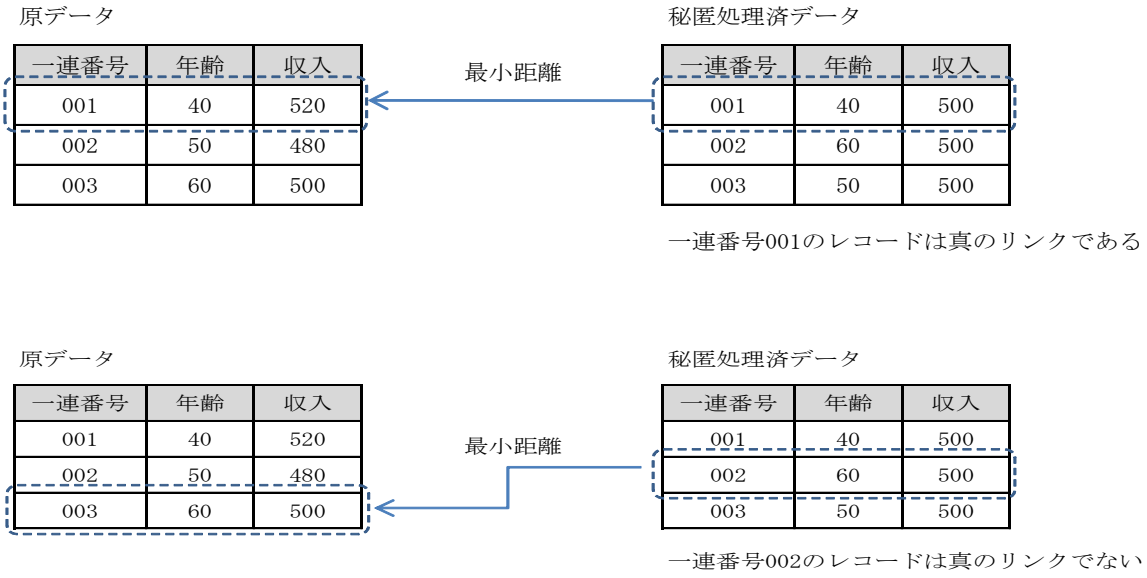
一連番号	年齢	収入
001	40	520
002	50	480
003	60	500

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



が用いられる[4]。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、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[5]。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）

$$d_{iI}^2 = \sum_j \left(\frac{x_{ij} - \bar{x}_j}{\sigma(x_j)} - \frac{X_{Ij} - \bar{X}_j}{\sigma(X_j)} \right)^2 \quad \cdots (7)$$

②属性間の距離の標準化

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

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

---

4 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。

5 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。

$\sigma(x_j)$：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。

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

$\mathbf{x}_i$：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[6]

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[7]。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。

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

6 確率的リンケージの詳細については、本稿の付録「確率的リンケージによる秘匿性の評価方法について」を参照されたい。

7 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。

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

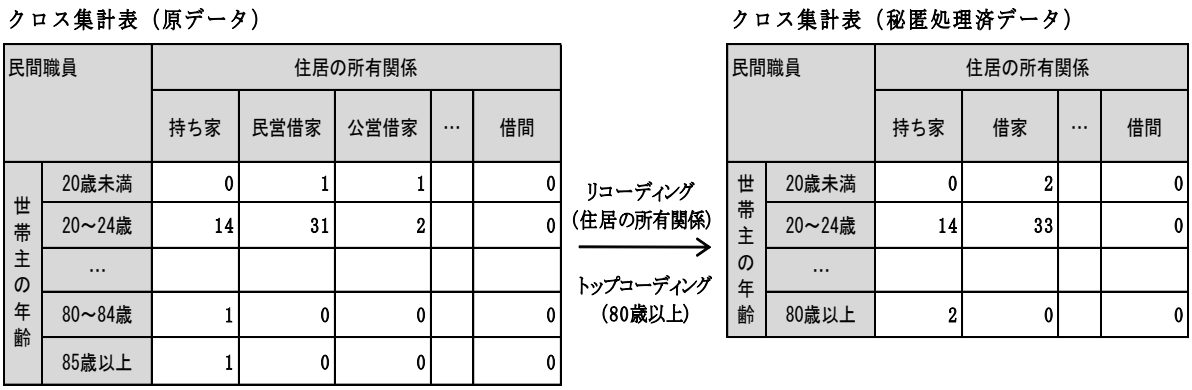
質的属性に匿名化技法を適用して作成した秘匿処理済データの秘匿性を評価する方法として、クロス集計表（分割表）を用いることが考えられる[8]。それは、データに含まれる複数の質的属性を対象に、クロス集計表における分布特性を比較することによって、秘匿性の強度を評価することを指向している。クロス集計表を用いることによって、原データと秘匿処理済データの間で度数が１となるセルの総数を比較し、度数１となるセル数の変化を確認することができる。このような度数１の変化を把握することは秘匿性の評価指標として適用可能だと思われる。

図 12 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。

---

[8] 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。

図 12　クロス集計表による秘匿性の評価のイメージ



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

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。また、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）。また、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。

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[9]の３種類を用いて実験を行っ

[9] 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。その理由としては、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

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

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%）

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。

次に、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

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%）

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[10]。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

[10] 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。

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。

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、世帯主の職業と住居の所有関係を使用する。そして、①世帯主の年齢と世帯主の職業、及び②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２種類の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、世帯主の職業及び住居の所有関係に関しては、リコーディング（世帯主の職業については 12 区分を４区分に、住居の所有関係に関しては８区分を２区分にそれぞれ分類区分の統合）を行った。世帯主の職業及び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６のように設定されている。また、表７と表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７と表８に示されているとおり、世帯主の年齢と世帯主の職業、及び世帯主の年齢と住居の所有関係のいずれについても、匿名化技法の組み合わせのパターンはそれぞれ８パターンである。

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

表６　世帯主の職業及び住居の所有関係の原区分と統合区分

世帯主の職業			
原区分		統合区分	
1	常用労務作業者	1	労務作業者
2	臨時及び日々雇労務作業者		
3	民間職員	2	民間職員
4	官公職員１	3	官公職員
5	官公職員２		
6	商人及び職人	4	勤労者以外
7	個人経営者		
8	農林漁業従業者		
9	法人経営者		
10	自由業者		
11	その他		
12	無職		

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７　世帯主の年齢と世帯主の職業に対する匿名化技法の組合せ

世帯主の年齢		世帯主の職業
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	12区分
各歳	85歳以上	12区分
各歳	なし	４区分
各歳	85歳以上	４区分
５歳階級	なし	12区分
５歳階級	85歳以上	12区分
５歳階級	なし	４区分
５歳階級	85歳以上	４区分

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

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(10)式によって算出されている。

$$\text{度数1の減少率}(\%) = \frac{\text{度数1のセル数〔原データ〕} - \text{度数1のセル数〔秘匿処理済データ〕}}{\text{度数1のセル数〔原データ〕}} \times 100 \cdots (10)$$

(1) 質的属性に関する有用性と秘匿性の比較―世帯主の年齢と職業を例に―

表９は世帯主の年齢と世帯主の職業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。世帯主の年齢と世帯主の職業の分類区分をそれぞれ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)[11]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、職業を４区分にリコーディングした場合は、情報量損失の値は 170,789 と算出されるが、情報量損失率は-38.4%となっている。

[11] 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×世帯主の職業〕による有用性の評価

	世帯主の年齢		世帯主の職業	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	12区分	0	0.0
〔B〕	各歳	85歳以上	12区分	861	−0.2
〔C〕	各歳	なし	４区分	43,988	−9.9
〔D〕	各歳	85歳以上	４区分	44,879	−10.1
〔E〕	５歳階級	なし	12区分	126,165	−28.4
〔F〕	５歳階級	85歳以上	12区分	126,421	−28.4
〔G〕	５歳階級	なし	４区分	170,526	−38.3
〔H〕	５歳階級	85歳以上	４区分	170,789	−38.4

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、世帯主の職業を 12 区分から４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。

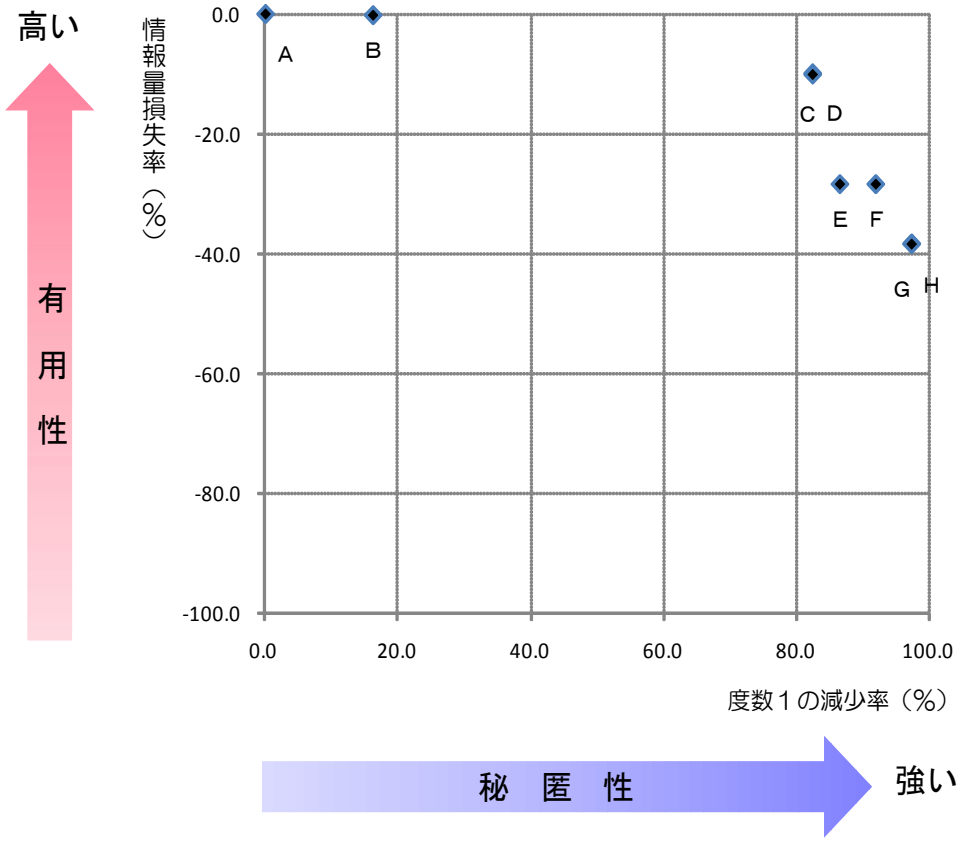
表 10 は世帯主の年齢と世帯主の職業に対して匿名化技法を適用したことによる秘匿性の相対的な評価結果を表したものである。表 10 における度数１の減少率は、パターン〔A〕（原データ）における度数１の数を基準とした場合に、匿名化技法の適用によって度数１が減少した割合を表している。よって、例えばパターン〔H〕のように、世帯主の年齢に対してリコーディング及びトップコーディングを適用し、さらに世帯主の職業に対してリコーディングを行った場合、度数１の減少率は 97.3%となっている。表 10 を見ると、世帯主の年齢あるいは職業にリコーディングを施すことによって、度数１の減少率が大きくなっていることがわかる。

図 13 は、世帯主の年齢と世帯主の職業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 13 における縦軸は情報量損失率(%)を、横軸は度数１の減少率

表 10　クロス集計表〔世帯主の年齢×世帯主の職業〕による秘匿性の評価

	世帯主の年齢		世帯主の職業	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	12区分	0.0
〔B〕	各歳	85歳以上	12区分	16.2
〔C〕	各歳	なし	４区分	82.4
〔D〕	各歳	85歳以上	４区分	82.4
〔E〕	５歳階級	なし	12区分	86.5
〔F〕	５歳階級	85歳以上	12区分	91.9
〔G〕	５歳階級	なし	４区分	97.3
〔H〕	５歳階級	85歳以上	４区分	97.3

図 13　世帯主の年齢と世帯主の職業に関する R-U マップ



	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〕	○	○	○

注　本図におけるパターン〔C〕と〔D〕及びパターン〔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。

(%)を、それぞれ表している。また、図 13 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。しかし、図 13 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、有用性と秘匿性がトレードオフの関係にあることが視覚的に把握できる。また、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。

(2) 質的属性に関する有用性と秘匿性の比較―世帯主の年齢と住居の所有関係の場合―

次に、世帯主の年齢及び住居の所有関係の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について、有用性と秘匿性の定量的な評価を行う。表 11 では、表９と同様に情報エントロピーを用いて、匿名化技法の適用による情報量損失の程度を比較している。また、表 12 は、表 10 と同様に、クロス集計表を用いて度数１の減少率を算出している。さらに、図 14 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に関する R-U マップを図示したものである。

図 14 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、有用性が保持されている。

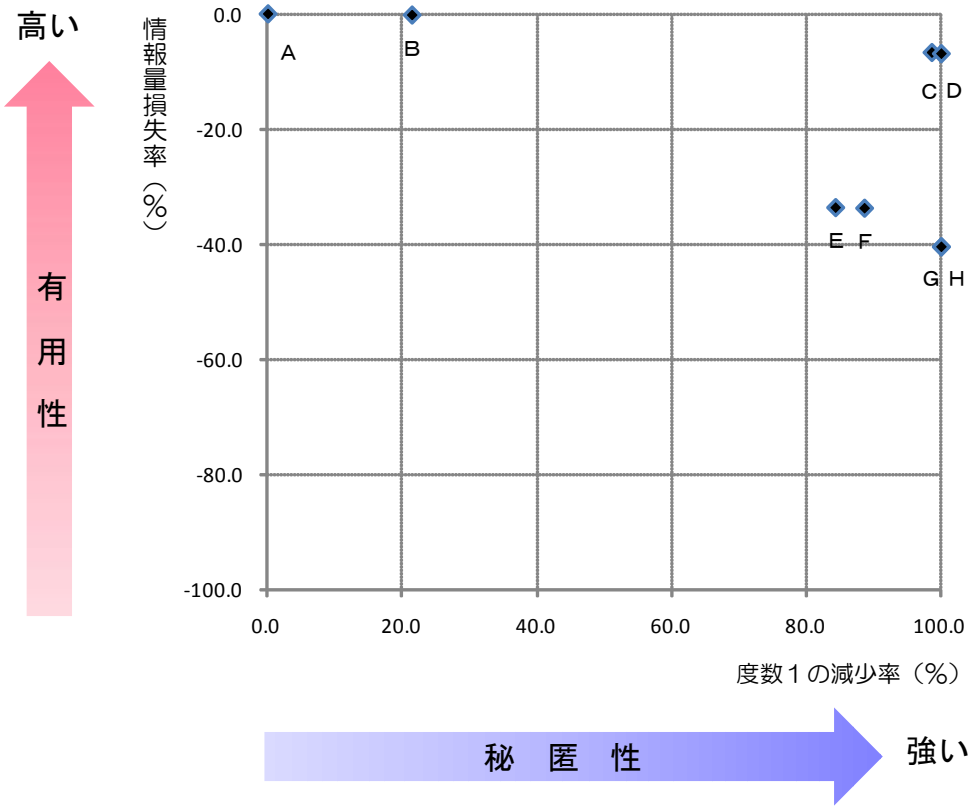
表 11　有用性の実証結果―情報エントロピーを利用した情報量損失〔世帯主の年齢×住居の所有関係〕

	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

表 12　秘匿性の実証結果―クロス集計表〔世帯主の年齢×住居の所有関係〕

	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 14　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。

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。このように、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[12]。

６．むすびにかえて

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

[12] 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(参考１～参考２を参照)。それによれば、世帯主の年齢と世帯主の職業、および世帯主の年齢と住居の所有関係のいずれの組合せにおいても、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(図 13、図 14)と基本的には変わっていない。

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 マップが有効な手法であることが確認できた。この R-U マップを用いることによって、本研究では考察の対象とされなかったノイズやスワッピング等の攪乱的手法(perturbation)についても、R-U マップを用いて有効性の検討を行ってみたいと考えている。一方、R-U マップだけでなく、様々な匿名化技法を適用した場合のミクロデータにおける有用性と秘匿性の相対比較を行うためのフレームワークとして、有用性と秘匿性に関する総合的な評価指標の構築に向けた展開を図ることも考えられる(Domingo-Ferrer and Torra(2001b))。他方、ミクロデータに含まれる地域属性に関しては、地域区分が個体情報の特定化において識別力の高い情報であることから、情報量損失やクロス集計表における分布特性の比較だけでなく、有用性と秘匿性に関する評価方法の検討をさらに追究する必要があると考えている。これらについては今後の課題としたい。

参考文献

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., Keller-McNulty, S. and Stokes, S. L.(2001) ”Disclosure Risk vs. Data Utility: the R-U Confidentiality Map” *Technical Report* 121, US National Institute of Statistical Sciences, Durham, North Carolina.

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.

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

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

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.

Kooiman, P., L. Willenborg and J. Gouweleeuw (1998) "PRAM: A Method for Disclosure Limitation of Microdata", *Research Paper*, No. 9705, Statistics Netherlands, Voorburg.

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.

Shannon, C. E.(1948) "A Mathematical Theory of Communication", *The Bell System Technical Journal*, Vol. 27, pp. 379-423.

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.

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.

Winglee, M., Valliant, R., Clark, J., Lim, Y., Weber, M., Strudler, M. (2002) “Assessing Disclosure Protection for the SOI Public Use File”, Paper Presented at Proceedings of the Annual Meeting of the American Statistical Association.

http://www.amstat.org/sections/SRMS/Proceedings/

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.

付録　確率的リンケージによる秘匿性の評価方法について

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。付図１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、付図２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、それぞれ付図１の M/U 欄に表示されている。なお、M と U は、次の(A1)式で与えられる。

$$M=\{(a,b);a=b,a\in A,b\in B\},\quad U=\{(a,b);a\neq b,a\in A,b\in B\}\cdots(A1)$$

また、付図２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((A2)式)。

$$\Gamma=\{\gamma_1,\gamma_2,\cdots,\gamma_{16}\}=\{(1,1,1,1),(1,1,1,0),(1,1,0,1),(1,0,1,1),(0,1,1,1),(1,1,0,0),\cdots,(0,0,0,0)\}\cdots(A2)$$

付図３では、付図２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、付図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(A3)式で与えられる。

$$m=P(\gamma=\gamma(a,b)|(a,b)\in M),\quad u=P(\gamma=\gamma(a,b)|(a,b)\in U)\cdots(A3)$$

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(A4)式を定義している。

$$R_p(a,b)=R(\gamma(a,b))=\log(m/u)\cdots(A4)$$

(A4)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。

付図１　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

付図２　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

付図３　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)。

$$\frac{m^{\sigma(1)}}{u^{\sigma(1)}} > \frac{m^{\sigma(2)}}{u^{\sigma(2)}} > \frac{m^{\sigma(3)}}{u^{\sigma(3)}} > \cdots \quad \cdots (A5)$$

②μの下限とλの上限については次の(A6)式と(A7)式を満たすように算出する。

$$\sum_{i=1}^{limit} u^{\sigma(i)} \leq \mu < \sum_{i=1}^{limit+1} u^{\sigma(i)} \quad \cdots (A6)$$

$$\sum_{i=limit'}^{|\Gamma|} m^{\sigma(i)} \leq \lambda < \sum_{i=limit'-1}^{|\Gamma|} m^{\sigma(i)} \quad \cdots (A7)$$

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
$|\Gamma|$: 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((A8)式と(A9)式)。

$$ut = \log\left(m^{\sigma(limit)} / u^{\sigma(limit)}\right) \cdots (A8)$$

$$lt = \log\left(m^{\sigma(limit')} / u^{\sigma(limit')}\right) \cdots (A9)$$

例えば、μ=0.1、λ=0.2に設定したとする。そのとき、(A8)式と(A9)式より、μとλの範囲は次のように与えられる((A10)式と(A11)式)。

$$\sum_{i=1}^{2} u^{\sigma(i)} \left(= 1/20 + 1/20 = 0.1\right) \leq \mu < \sum_{i=1}^{3} u^{\sigma(i)} \left(= 1/20 + 1/20 + 2/20 = 0.2\right) \cdots (A10)$$

$$\sum_{i=4}^{9} m^{\sigma(i)} \left(= 0/5 + \cdots + 0/5 + 1/5 = 0.2\right) \leq \lambda < \sum_{i=3}^{9} m^{\sigma(i)} \left(= 0/5 + \cdots + 0/5 + 1/5 + 1/5 = 0.4\right) \cdots (A11)$$

(A10)式と(A11)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(付図３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。

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。また、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が選ばれている。

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。

付表１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。付表１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、付表１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、付表１の区間設定による評価に関しては標準偏差１％～10％(１％刻み)が用いられているが、付表１には、10 個の結果の平均値が掲載されている。

付表１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。

付表１　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

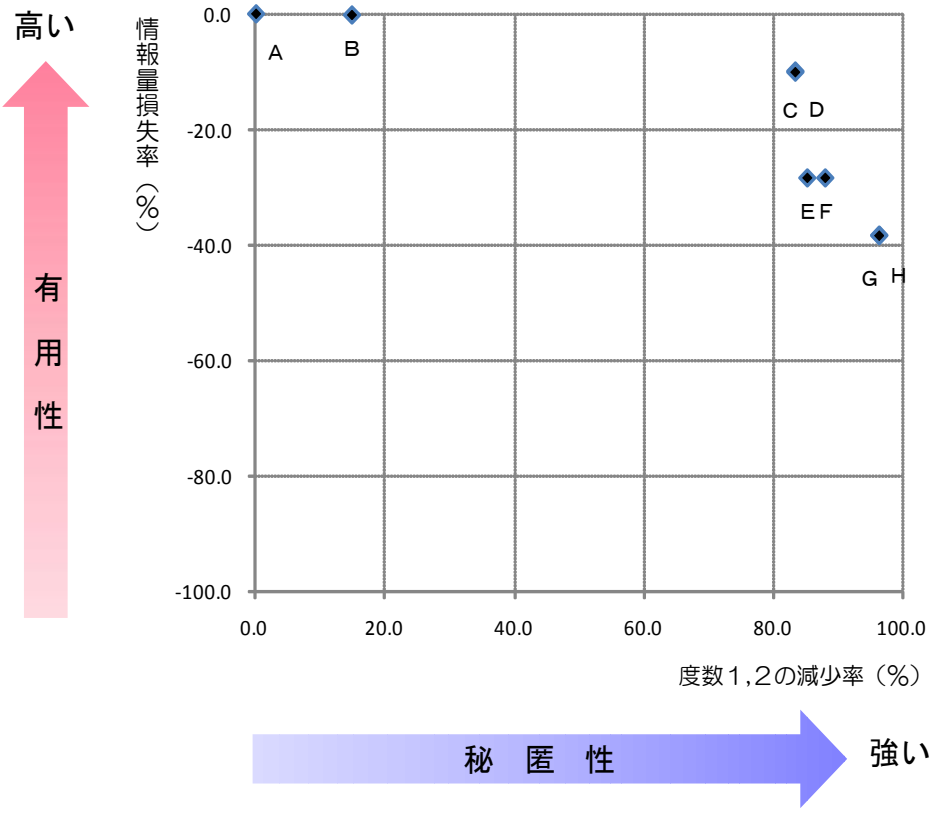
		ソートなし	個別ランキング法
確率的リンケージ（μ=0.0001）		66（0.16%）	39,320（97.05%）
確率的リンケージ（μ=0.001）		483（1.19%）	40,063（98.88%）
確率的リンケージ（μ=0.01）		665（1.64%）	40,370（99.64%）
参考	確定的リンケージ	0（0.00%）	28,877（71.27%）
	距離計測型リンケージ		
	①ユークリッド距離（属性値を標準化）	38（0.09%）	39,985（98.69%）
	②ユークリッド距離（属性間の距離を標準化）	27（0.07%）	39,987（98.69%）
	③マハラノビス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

参考１　質的属性に関する秘匿性の評価と R-U マップ―世帯主の年齢と職業―

クロス集計表による秘匿性の実証結果：度数１，２の減少率による評価

	世帯主の年齢		世帯主の職業	度数１，２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	12区分	0.0
〔B〕	各歳	85歳以上	12区分	14.8
〔C〕	各歳	なし	４区分	83.3
〔D〕	各歳	85歳以上	４区分	83.3
〔E〕	５歳階級	なし	12区分	85.2
〔F〕	５歳階級	85歳以上	12区分	88.0
〔G〕	５歳階級	なし	４区分	96.3
〔H〕	５歳階級	85歳以上	４区分	96.3

世帯主の年齢と世帯主の職業に関する R-U マップ



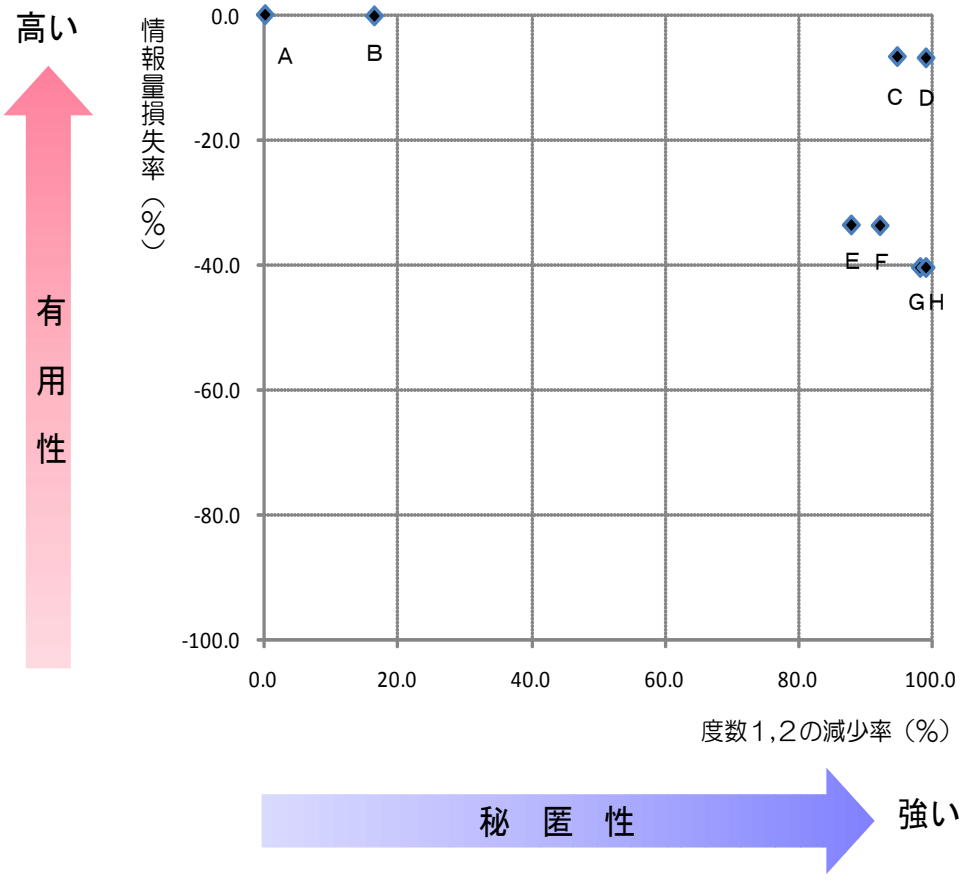
	年齢		職業
	リコーディング（５歳階級）	トップコーディング（85歳以上）	リコーディング（４区分）
[A]			
[B]		○	
[C]			○
[D]		○	○
[E]	○		
[F]	○	○	
[G]	○		○
[H]	○	○	○

参考２　質的属性に関する秘匿性の評価と R-U マップ―世帯主の年齢と住居の所有関係―

クロス集計表による秘匿性の実証結果：度数１，２の減少率による評価

	世帯主の年齢		住居の所有関係	度数１，２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	16.4
〔C〕	各歳	なし	２区分	94.8
〔D〕	各歳	85歳以上	２区分	99.1
〔E〕	５歳階級	なし	８区分	87.9
〔F〕	５歳階級	85歳以上	８区分	92.2
〔G〕	５歳階級	なし	２区分	98.3
〔H〕	５歳階級	85歳以上	２区分	99.1

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]	○	○	○

製 表 技 術 参 考 資 料 14

平成 22 年 9 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

掲載論文を引用する場合は、事前に下記まで連絡してください

情報技術部研究主幹　　TEL ： 03－5273－1368

E-mail ： research@nstac.go.jp