製表技術参考資料 28

ISSN　1348－7345
ISSN－L 1348－7329

ミクロデータにおける匿名化の誤差の評価に関する研究

―国勢調査を例に―

及び

スワッピングの適用可能性に関する評価研究

―国勢調査ミクロデータを用いて―

NSTAC

Working Paper No.28

平成 27 年 3 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

ミクロデータにおける匿名化の誤差の評価に関する研究

―国勢調査を例に―

伊藤伸介*、星野なおみ**

要旨

我が国では、これまで、就業構造基本調査、全国消費実態調査といった標本調査の匿名データだけでなく、全数調査である国勢調査についても、平成 25 年度に平成 12 年と平成 17 年の国勢調査の匿名データが提供されてきた。国勢調査の匿名データの特徴としては、「匿名データの作成・提供に関するガイドライン」に沿った形で、(１)地域区分については都道府県と人口 50 万以上市区が利用可能なこと、(２)標本抽出率は１％でありかつ世帯単位でレコードが抽出されていること、(３)リコーディングやトップコーディング等の様々な匿名化技法が適用されていることが指摘される。

他方で、将来的には、地域分析用の匿名データ等、別のタイプの国勢調査の匿名データの要望が出てくる可能性があることから、ミクロデータにおける匿名化技法の有効性を検証することは有用であると考えられる。そこで、本稿では、国勢調査を例に、個票データに秘匿処理を適用することによって作成したミクロデータ(以下「匿名化ミクロデータ」と呼ぶ。)の秘匿性と有用性に関する実証研究を行うことによって、匿名化の誤差の検証を行った。

本研究では、平成 17 年国勢調査の個票データにおける特定の地域(以下「地域 A」と呼ぶ。)のレコードをもとに作成したテストデータ(約 100,000 レコード)を用いて、秘匿性の検証を行った。秘匿性の検証方法としては、第 1 に、リコーディング、トップコーディング、及びサンプリング(1%, 5%,10%)を行うことによって作成された匿名化ミクロデータを対象に、母集団一意(population unique)かつ標本一意(sample unique)の比率(UUSU 比率)を計測した。本研究においては、母集団一意(および標本一意)の計測のために使用するキー変数として、性別、年齢等の質的属性を用いている。第 2 に、外部情報と匿名化ミクロデータとのマッチングを試みた。具体的には、スワッピングが施された国勢調査の匿名化ミクロデータに対して、平成 20 年住宅・土地統計調査の地域 A に該当するレコードを含む個票データ(約 10,000 レコード)とのマッチングの実験を行った。なお、スワッピングについては、ターゲット・スワッピング(targeted data swapping)とランダム・スワッピング(random data swapping)の２つの手法を適用し、

* （独）統計センター統計情報・技術部統計技術研究課非常勤研究員（中央大学経済学部准教授）
** （独）統計センター統計情報・技術部統計技術研究課

異なるスワッピング率における秘匿性の強度を検証した。

その一方で、本研究では、主として、サンプリングとスワッピングにおける誤差を算定することにより、有用性を検証することを試みた。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は、クラメール V といった関連性の指標の算出や原データからの絶対距離の平均値(average absolute distance)の計測等を行うことが考えられるが、本研究においては、サンプリングにおける誤差にも着目し、キー変数以外の属性を対象にスワッピングが適用された匿名化ミクロデータ(以下「スワッピング済データ」と呼ぶ。)と原データとの分布の差を確認した。

さらに、スワッピング済データの秘匿性と有用性を比較・検討するために、秘匿性と有用性の評価指標に基づいて、R-U マップ(R-U Confidentiality Map)を作成し、スワッピング率及びスワッピングの方法を変えた場合の R-U マップの変化を確認した。

本分析の結果については、次の３点に要約される。第１に、外部情報とのマッチングの試みとして、国勢調査の匿名化ミクロデータと住宅・土地統計調査の個票データとのマッチングを行ったが、本研究においては、マッチング率は低いという興味深い結果が得られた。また、「特殊な一意(special uniques)」に該当するようなレコードに対して、追加的な匿名化手法としてスワッピングを適用した場合、秘匿性の強度が高まることが定量的に明らかになった。第２に、特定のサンプリング率においてスワッピング済データと原データとの差がサンプリングの誤差の範囲で収まるかどうかを確認することによって、サンプリングの誤差の観点からスワッピングの有効性を検証することが可能なことがわかった。第３に、R-U マップに基づいて、スワッピング済データにおける有用性と秘匿性の検証も行った結果、有用性あるいは秘匿性に関する閾値を設定することができれば、適切なスワッピング率およびスワッピングの方法を選択することが可能なことが実証的に確認された。

ミクロデータにおける匿名化の誤差の評価に関する研究

―国勢調査を例に―

伊藤伸介、星野なおみ

１．はじめに

我が国では、これまで、就業構造基本調査、全国消費実態調査といった標本調査の匿名データだけでなく、全数調査である国勢調査についても、平成 25 年度に平成 12 年と平成 17 年の国勢調査の匿名データが作成・提供されてきた。現在提供されている国勢調査の匿名データの特徴としては、「匿名データの作成・提供に関するガイドライン」に沿った形で、(１)地域区分については都道府県と人口 50 万以上市区が利用可能なこと、(２)標本抽出率は１％でありかつ世帯単位でレコードが抽出されていること、(３)リコーディングやトップコーディング等の様々な匿名化技法が適用されていることを指摘することができる。

他方で、将来的には、地域分析用の匿名データ等、別のタイプの国勢調査の匿名データの要望が出てくる可能性があり、その予備的な研究とし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ことは有用であると考えられる。そこで、本稿では、各種匿名化技法を用いて作成された国勢調査の匿名化ミクロデータを対象に、秘匿性と有用性に関する実証研究を行うことによって、匿名化技法の有効性の検証を行うことにしたい。

２．本研究における匿名化ミクロデータの作成方法

原データ(秘匿処理が施される前の個票データ)に秘匿処理を施すことによって匿名化ミクロデータを作成する場合、情報の削除(特異なレコードの削除も含む)や区分の再編(リコーディング、トップ((ボトム)コーディング)、サンプリングといった非攪乱的手法だけでなく、スワッピング、ノイズ(加法ノイズ等)、ミクロアグリゲーションといった攪乱的手法についても、その適用可能性を議論することが求められる。そこで、本研究では、国勢調査を例に、匿名化ミクロデータを試行的に作成した。本研究で使用するデータは、平成 17 年国勢調査(以下「国調」と略称)の個票データをもとに、特定の地域(以下「地域 A」と呼ぶ。)のレコードから作成したテストデータ(約 100,000 レコード)である。このテストデータには、個人単位で抽出した一般世帯の世帯主のレコードのみが含まれている。

匿名化ミクロデータの作成の手順は以下のとおりである。①国調のテストデータに

対して、労働力状態、従業上の地位と年齢についてはリコーディングを、世帯人員と年齢に関してはトップコーディングをそれぞれ適用した[1]。具体的には、労働力状態については 9 区分を 6 区分に、従業上の地位については 8 区分を 4 区分に、それぞれリコーディングを行い、世帯人員については 8 人以上の分類区分にトップコーディングを適用している。なお、年齢については、各歳年齢区分から 5 歳年齢区分へのリコーディングを行った上で、85 歳以上の区分についてはトップコーディングを施している。②リコーディングとトップコーディングを施したデータに対して、様々な標本抽出率(1%、5%、10%)によるサンプリングを行った[2]。③①と②の匿名化技法が適用されたデータに対して、複数のタイプのスワッピングを適用した。

スワッピングは、以下のように行われた(伊藤・星野(2014))。第１に、キー変数(key variable)を用いて標本一意(sample unique)を計測し、スワッピングの対象となるレコードを選出する。なお、標本一意となるレコード数を算出するために使用するキー変数は、表 1 に示される 12 変数である。

第 2 に、スワッピングの対象レコードの中で、優先度の高いレコードをスコアに基づいて探索する。具体的には、標本一意となるレコードを対象に、キー変数のすべての組み合わせでクロス集計を行い、ある特定のレコードが標本一意に該当した回数がスコアとして計測される。

第 3 に、対象レコードに対してスワッピングを適用する。具体的には, 算出されたスコアに基づいて特定化のリスクの高いレコードに焦点を絞ってスワッピングを行うターゲット・スワッピング(targeted data swapping)、および対象となるレコードの中からランダムにレコードを選んで、そのレコードに対してスワッピングを行うランダム・スワッピング(random data swapping)が行われた。

本研究では、地域 A のレコード(10%抽出の場合約 10,000 レコード)に対してスワッピングを適用する。スワッピングの方法は、以下のとおりである(伊藤・星野(2014))。最初に、標本一意に該当した回数が 1 回以上のレコードをスワッピングの候補となるレコードとして選出する。そして、スワッピング率として 1%、2%、3%、5%、10%、20%、30%を設定した上で、

①ターゲット・スワッピングの場合, 母集団一意かつ標本一意のレコードを含むスコアの高い上位 p%(p はスワッピング率)に該当するレコードをスワッピングの対象レコードとした。

②ランダム・スワッピングの場合, スワッピングの候補となるレコードからランダムに p%選別されたレコードをスワッピングの対象レコードとした。

---

[1] 労働力状態、従業上の地位に関する原データとテストデータにおける分類区分の比較について付表１を参照されたい。

[2] 本稿では、基本的にはサンプリング率が 10%における匿名化ミクロデータを用いて行った実験の結果に基づいて議論を進めることする。

表 1　本研究において標本一意の計測のために用いたキー変数

変数	区分数
世帯主との続き柄	13
男女の別	2
年齢 5 歳階級(トップコーディング済)	19
配偶関係	5
国籍	13
労働力状態(リコーディング済)	6
従業上の地位(リコーディング済)	4
産業大分類	19
職業大分類	10
住居の種類	9
建て方の種類	5
建物の階数[3] (建物の階数については共同住宅のみ)	30

スワッピングの対象レコードに対してその入れ替えの候補となるレコードについては、地域 A とは異なる地域(以下「地域 B」と呼ぶ。)を対象に、ドナーファイル(約 5,000 レコード)から探索する。

ところで、スワッピングの対象となるレコードは、「特殊な一意(special uniques)」[4]として出現する可能性がある。その場合、スワッピングの対象レコードとキー変数の値が完全に一致するレコードが、ドナーファイルにおいて見つかる可能性は低い。したがって、本研究においては、スワッピングの対象レコードに対して、ドナーファイルに含まれるレコードとの距離を計算した上で、ドナーファイルの中で最も距離が小さいレコードとスワッピングを行っている。具体的には、距離計測型リンケージ(Domingo-Ferrer and Torra (2001),Takemura(1999))の方法を援用し、以下の手順に従っている(Domingo-Ferrer and Torra (2001),Takemura(1999), 伊藤・星野(2013), 伊藤・星野(2014, 8～9 頁))。

最初に、i(i=1,...,m)および j(j=1,...,n)を、それぞれスワッピング対象レコードの番号およびドナーファイルのレコード番号とし(m と n は、それぞれスワッピング対象

[3] 建物の階数については、「建て方の種類」が共同住宅に該当するレコードのみが対象となる。

[4] 特殊な一意とは、「K 個のキー変数の集合において標本一意であるだけでなく、K の部分集合である k 個(のキー変数の集合)においても標本一意となること」である(Elliot and Manning(2004), 伊藤・星野(2014))。具体的には、「疫学的に特異であるために、本質的に(intrinsically)まれな属性群の組み合わせを有する」レコードは、標本一意の中で母集団一意に該当するレコードの中でも個人が特定化される可能性が特に高くなることから、特殊な一意に該当するレコードとみなされる(Elliot(2001), 伊藤・星野(2014))。

レコードの数およびドナーファイルのレコード数)、k(j=1,...,11)をキー変数の番号[5]とする。このとき、i 番目のレコードにおけるキー変数 k の分類区分の数値をCs_{ki}、j 番目のドナーファイルのレコードにおけるキー変数 k の分類区分の数値をCd_{kj}とすれば、キー変数 k に関する i と j の質的属性値間の距離(distance for categorical variables)を次の(1)式で定義することができる(Domingo-Ferrer and Torra, (2001, pp.105-106))。

$$Sd_{kij} = \left| Cs_{ki} - Cd_{kj} \right| \quad (1)$$

なお、年齢および住居の建て方の「共同住宅」以外の場合、$\left| Cs_{ki} - Cd_{kj} \right| > 0$であれば、$Sd_{kij} = 1$とする。

次に、質的属性値間の距離をスコア化するために、k 番目のキー変数における分類区分数C_kでSd_{kij}を除することによって、k 番目のキー変数におけるスコアである$Score_{kij}$が(2)式によって求められる。すなわち、

$$Score_{kij} = \frac{1}{C_k} \cdot Sd_{kij} \quad (2)$$

さらに、各キー変数のスコアの総計を算出することによって、全てのキー変数に関する i 番目と j 番目のレコード間の距離についての総合指標D_{ij}が導出される。

$$D_{ij} = \sum_k Score_{kij} \quad (3)$$

最後に、スワッピングの対象レコードとドナーファイルに含まれるおのおののレコードとの間で総合指標D_{ij}を計測し、ドナーファイルの中でD_{ij}が最も小さいレコードをスワッピング対象レコードと置き換える(Domingo-Ferrer and Torra(2001), Takemura(1999))[6]。

３．国勢調査ミクロデータを用いた秘匿性の評価―マッチングの試み―

本節では、前節で述べた国勢調査のテストデータを用いて秘匿性の検証を行う。具体的には、(1)サンプリングと(2)スワッピングにおける秘匿性の程度を検証することに焦点を当てる。

ところで、政府統計ミクロデータに関する秘匿性については、諸外国では主として

[5] マッチングの実験では、「建て方の種類」と「建物の階数」を組み合わせた変数を用いる。ゆえに、距離計測型リンケージで用いるキー変数は 11 である。

[6] 距離を計算した際に、ドナーファイルの中でもっとも距離が小さいレコードが複数存在する場合もある。その場合には、最小の距離を有する複数のレコードの中からランダムに 1 つのレコードを選んでいる。

個体識別(identification)に伴う露見リスク(disclosure risk)の評価として議論が展開されてきた。個体識別は, つぎのように考えることができる（Bethlehem *et al.*(1990), Marsh *et al.*(1991), Müller *et al.*(1995), 伊藤(2010), 伊藤・星野(2014)）。ミクロデータの入手者（侵入者, intruder）が、識別の対象となる特定の個体情報に関するファイル（識別ファイル, identification file）を持っていたとする。その場合に、(1)識別ファイルに含まれるレコードとミクロデータ上に存在するレコードにおいて、キー変数を通じて１対１のマッチングが行われ、(2)対応関係にあるレコードが特定の個体のものであることが確認された場合に、個体識別が成立したとみなされる。

ミクロデータの入手者が、外部情報とのマッチングによって個体識別を試みることを想定した場合、個体識別による露見リスクの定量的な評価に関しては、主に次の 2 つの方向からの研究が行われてきた(伊藤(2010))。第 1 は、提供されるミクロデータにおいて「母集団一意」に関連した指標を計測することである(Bethlehem *et al.*(1990), Marsh *et al.*(1991)等)。第 2 は、ミクロデータの入手者における外部情報の取得可能性を検討するだけでなく, 外部情報とミクロデータのマッチングの検証を行うことによって個体識別による露見リスクの可能性を追究することである(Müller *et al.*(1995))。

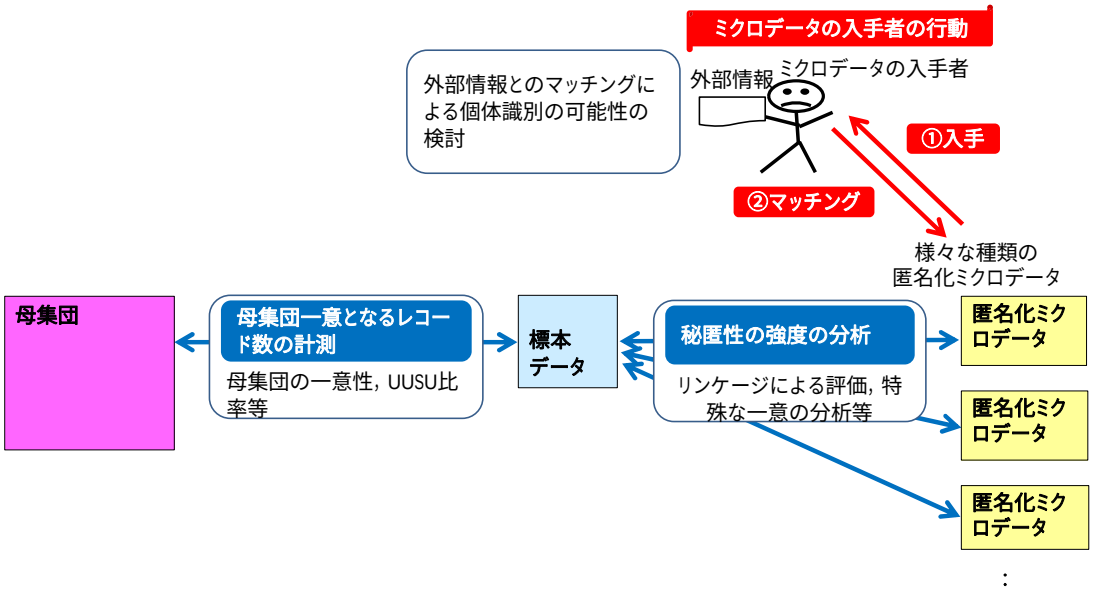
図１は、ミクロデータにおける秘匿性の評価に関する概略図(伊藤(2010))を示したものであるが、母集団一意の計測については、標本データから母集団一意(population unique)となるレコード数を計測することによって、母集団に対する母集団一意の比率で表される「母集団の一意性(population uniqueness)」や標本一意(sample unique=SU)に対する母集団一意かつ標本一意(union unique=UU)の比率である UUSU 比率(UUSU ratio)といった指標に基づいて、個体が識別される確率を計測することができる。その一方で、各種匿名化技法を適用することによって作成した匿名化ミクロデータについては、原データと匿名化ミクロデータとのリンケージや、特殊な一意の分析(special unique analysis)によって秘匿性の相対的な強度を定量的に評価することが考えられる。

他方、このようにして作成された匿名化ミクロデータについては、ミクロデータの入手者の行動を想定し、識別の戦略[7]に基づいて外部情報とのマッチングの程度を計測することによって、匿名化ミクロデータの秘匿性の定量的な評価が可能になる。

---

[7] Müller *et al.*(1995)で議論されているように、ミクロデータの入手者における識別の戦略としては、①直接検索(directed search)と②釣り検索(fishing strategy)がある。直接検索とは、識別ファイルを用いて, ミクロデータファイルに含まれるある特定の個体のレコードを突き止める戦略である。それに対して、釣り検索とは、ミクロデータファイルの中で関心があるレコードに焦点を絞り、それらのレコードを識別するために、識別ファイルの中で対応付け可能なレコードを突き止める戦略である。また、直接検索においては、対象となる個体のレコードがミクロデータファイルに存在するという情報(調査参加情報(participation knowledge))をミクロデータの入手者が持っている場合が考えられる。こうした調査参加情報がある場合の直接検索は、もっとも露見リスクが高いシナリオだと考えられている(Müller *et al.*(1995, p.139))。

図 1　ミクロデータにおける秘匿性の評価に関する概略図



出所　伊藤(2010,8 頁)の図 3 を一部修正した。

こうした先行研究を踏まえ、本研究では、最初に、母集団一意となるレコード数を計測するために、リコーディングおよびトップコーディングを行ったデータに対してサンプリング(1%, 5%,10%)を行った上で、母集団一意かつ標本一意の計測を行った。なお、母集団一意かつ標本一意の計測に用いたキー変数は、表 1 で示された 12 変数を用いている。

表 2 は、サンプリング率が 1%(レコード数は 1,000 レコード)、5%(レコード数は 5,000 レコード)と 10%(レコード数は 10,000 レコード)の場合における標本一意に該当するレコード数と母集団一意かつ標本一意に該当するレコード数の計測結果を表したものである。1%のサンプリング率の場合、母集団として位置付けられるテストデータ(10 万レコード)から 100 組のサンプルデータが抽出可能なことから、表 2 では、100 組のサンプルデータの平均値、標準偏差、最小値と最大値を示している。同様に、5%のサンプリング率の場合には 20 組のサンプルデータを対象に、10%のサンプリング率の場合には 10 組のサンプルデータについて、それぞれ平均値、標準偏差、最小値と最大値を算出している。なお、母集団一意に該当するレコード数は 14,568 レコードである。

1%のサンプリング率の場合の標本一意および母集団一意かつ標本一意の平均値は、それぞれ 573 と 146 である。したがって、標本一意に占める母集団一意かつ標本一意の比率(UUSU 比率)は、25.4%である。一方、5%のサンプリング率および 10%のサンプリング率における UUSU 比率は、それぞれ 36.7%と 44.6%となっている。したがって、サンプリング率が上がるにしたがって、UUSU 比率が上昇することが

表２　母集団一意かつ標本一意の計測結果

サンプリング率	一意の類型	平均	標準偏差	最小値	最大値
1%	標本一意	573	15.0726	540	621
(1,000レコード)	母集団一意かつ標本一意	146	10.6751	122	170
5%	標本一意	1,987	27.8662	1,936	2,039
(5,000レコード)	母集団一意かつ標本一意	728	21.9167	687	769
10%	標本一意	3,270	37.7506	3,206	3,317
(10,000レコード)	母集団一意かつ標本一意	1,457	31.2438	1,389	1,500

確認できる。

次に、秘匿性の評価方法として、外部情報とミクロデータのマッチングに焦点を当てることにしたい。具体的には、国調以外の政府統計のミクロデータを外部情報とみなした上で、匿名化ミクロデータと外部情報とのマッチングを試みた。本研究では、サンプリング率 10%で抽出され、スワッピングが施された国調の匿名化ミクロデータに対して、平成 20 年住宅・土地統計調査 (以下「住調」と略称)の地域 A に該当するレコードを含む個票データ(約 10,000 レコード)とのマッチングの実験を行った。住調については、国調の匿名化ミクロデータと調査区が重複するように、対象レコードが選定される。

ところで、外部情報とのマッチングに関する諸外国の先行研究については、ドイツにおける事実上の匿名性(factual anonymity)[8]について実証的に明らかにするため、ドイツのミクロセンサス(Microcensus)と研究者名鑑(Kürschners Deutscher Gelehrtenkalender 1987)を用いたマッチングに関する研究(Müller *et al.*(1995))や、人口センサスの 2001 年 SARs の作成に関する実証研究として行われた 1991 年の 2%個人 SAR と一般世帯調査(General Household Survey)とのマッチングの実験(Elliot and Dale(1998))がある。

これらの先行研究を参考にした上で、本研究では、国調と住調の両方の共通の調査事項から選ばれた以下のケース 1 からケース 3 におけるキー変数に基づいて、国調と住調のマッチングを行った。

ケース 1：市町村番号(５桁)、世帯人員、性別、年齢、配偶関係
ケース 2：市町村番号(５桁)、世帯人員、性別、年齢、配偶関係、住宅の建て方、住宅所有の関係
ケース 3：市町村番号(５桁)、世帯人員、性別、年齢、配偶関係、住宅の建て方、住

[8] 「事実上の匿名性」とは、「著しく大きな時間、経費および労力の支出によってしか個別データから回答者を突きとめることができない」ことであって、1987 年ドイツ連邦統計法には、政府統計ミクロデータが「事実上匿名」であれば、学術研究のためにミクロデータを提供してもよいことが明記されている。なお、「事実上の匿名性」の概念に基づくミクロデータの匿名化措置については，濱砂(2000)を参照されたい。

宅所有の関係、建物の階数

本研究では、国調と住調においてキー変数の区分を合わせた上で、マッチングを行う。また、国調の匿名化ミクロデータにおいて上記のキー変数で一意になったレコードを対象に、住調とのマッチングが行われている[9]。さらに、マッチングにおける国調と住調の調査年次の違いについては、国調のレコードにおいて年齢を加算することによって、年次の調整を行っている。

表３から表８は、国調の匿名化ミクロデータと住調の個票データのマッチングの結果を３つのマッチングのケースとスワッピングの種類別に示したものである[10]。ケース１からケース３にかけてキー変数の数が多くなるにつれて、国調の匿名化ミクロデータにおける標本一意の数が増大していることがわかる。例えば、表３～表５を見ると、ターゲット・スワッピングが施された国調の匿名化ミクロデータについて、ケース１のキー変数でマッチングした場合の標本一意の数は約 900 であるのに対して、ケース３によるマッチング場合の標本一意の数は約 2,000～2,400 となっている。この傾向は、ターゲット・スワッピングだけでなく、ランダム・スワッピングの場合にも当てはまることを確認することができる。その一方で、ターゲット・スワッピングとランダム・スワッピングのいずれにおいても、スワッピング率が高くなるにつれて、国調の匿名化ミクロデータにおける標本一意の数が小さくなっていることが興味深い。その理由としては、スワッピングによって、標本一意に該当するレコードが度数 2 以上のセルに該当するグループに移動したことが考えられる。

表３から表８では、1 対 1 でマッチングされたレコード数が示されている。国調の匿名化ミクロデータに含まれる標本一意のレコード数の中で、住調の個票データと 1 対 1 でマッチングされたレコードの比率は、ケース１の場合、ターゲット・スワッピングとランダム・スワッピングのいずれについても約 20%であるのに対して、ケース 3 においては、その比率が約 4%に大きく減少していることがわかる。この結果については、国調と住調における「建物の階数」に関する調査対象者が異なっているため、国調のデータに「建物の階数」の値がブランクになっているレコードが少なからず存在することが指摘される。こうしたことから、キー変数がケース 3 の場合に、国調の匿名化ミクロデータにおいて標本一意に該当するレコード数に占めるマッチングされたレコード数の比率は、ケース 1 やケース 2 と比較して小さくなっていることが考えられる。

他方、スワッピング率が上昇するにつれて、マッチングされたレコードの中でスワッピングされたレコードの比率が高くなっているものの、スワッピング率が 30%の

9 本研究で行ったマッチングについては、識別の戦略の１つである釣り検索を適用したものをみなすことができる。

10 サンプリング率が 1%と 5%の場合における国調の匿名化ミクロデータと住調の個票データとのマッチングの結果は、それぞれ付表２－１～付表２－３及び付表３－１～付表３－３を参照。なお、スワッピング率については、1%、2%と 3%のみが適用されている。

表３　マッチングの結果：ケース 1, ターゲット・スワッピング, 10%サンプリングデータ

スワッピング率	国調における標本一意の数	住調における標本一意の数	マッチング		真のマッチング		国調の標本一意に対してマッチングされたレコードの比率	国調の標本一意に対する真のマッチングの比率
			マッチングされたレコード数	スワッピングされたレコード数	マッチングされたレコード数	スワッピングされたレコード数		
1% (100rcd)	933	866	195	3	20	0	20.90%	2.14%
2% (200rcd)	934	866	195	5	20	0	20.88%	2.14%
3% (300rcd)	931	866	194	6	20	0	20.84%	2.15%
5% (500rcd)	930	866	194	11	20	1	20.86%	2.15%
10% (1000rcd)	930	866	191	21	19	1	20.54%	2.04%
20% (2000rcd)	909	866	185	39	18	2	20.35%	1.98%
30% (3000rcd)	898	866	180	55	17	3	20.04%	1.89%

表４ マッチングの結果：ケース 2, ターゲット・スワッピング, 10%サンプリングデータ

スワッピング率	国調における標本一意の数	住調における標本一意の数	マッチング		真のマッチング		国調の標本一意に対してマッチングされたレコードの比率	国調の標本一意に対する真のマッチングの比率
			マッチングされたレコード数	スワッピングされたレコード数	マッチングされたレコード数	スワッピングされたレコード数		
1% (100rcd)	1,989	1,750	341	4	51	1	17.14%	2.56%
2% (200rcd)	1,979	1,750	342	11	51	1	17.28%	2.58%
3% (300rcd)	1,967	1,750	341	16	50	1	17.34%	2.54%
5% (500rcd)	1,952	1,750	339	28	49	3	17.37%	2.51%
10% (1000rcd)	1,913	1,750	336	53	44	4	17.56%	2.30%
20% (2000rcd)	1,820	1,750	322	98	38	7	17.69%	2.09%
30% (3000rcd)	1,740	1,750	311	135	31	8	17.87%	1.78%

表５ マッチングの結果：ケース 3 , ターゲット・スワッピング, 10%サンプリングデータ

スワッピング率	国調における標本一意の数	住調における標本一意の数	マッチング		真のマッチング		国調の標本一意に対してマッチングされたレコードの比率	国調の標本一意に対する真のマッチングの比率
			マッチングされたレコード数	スワッピングされたレコード数	マッチングされたレコード数	スワッピングされたレコード数		
1% (100rcd)	2,392	2,388	114	2	34	0	4.77%	1.42%
2% (200rcd)	2,379	2,388	111	4	32	0	4.67%	1.35%
3% (300rcd)	2,365	2,388	110	8	30	0	4.65%	1.27%
5% (500rcd)	2,337	2,388	108	14	27	1	4.62%	1.16%
10% (1000rcd)	2,274	2,388	102	25	22	2	4.49%	0.97%
20% (2000rcd)	2,133	2,388	94	43	15	3	4.41%	0.70%
30% (3000rcd)	2,002	2,388	84	54	10	3	4.20%	0.50%

表6 マッチングの結果：ケース 1，ランダム・スワッピング，10%サンプリングデータ

スワッピング率	国調における標本一意の数	住調における標本一意の数	マッチング		真のマッチング		国調の標本一意に対してマッチングされたレコードの比率	国調の標本一意に対する真のマッチングの比率
			マッチングされたレコード数	スワッピングされたレコード数	マッチングされたレコード数	スワッピングされたレコード数		
1% (100rcd)	934	866	196	2	20	0	20.99%	2.14%
2% (200rcd)	933	866	196	4	20	0	21.01%	2.14%
3% (300rcd)	935	866	195	6	19	0	20.86%	2.03%
5% (500rcd)	934	866	194	9	20	0	20.77%	2.14%
10% (1000rcd)	930	866	195	21	19	1	20.97%	2.04%
20% (2000rcd)	920	866	192	40	18	3	20.87%	1.96%
30% (3000rcd)	897	866	184	57	17	3	20.51%	1.90%

表7 マッチングの結果：ケース 2，ランダム・スワッピング，10%サンプリングデータ

スワッピング率	国調における標本一意の数	住調における標本一意の数	マッチング		真のマッチング		国調の標本一意に対してマッチングされたレコードの比率	国調の標本一意に対する真のマッチングの比率
			マッチングされたレコード数	スワッピングされたレコード数	マッチングされたレコード数	スワッピングされたレコード数		
1% (100rcd)	2,004	1,750	343	4	52	0	17.12%	2.59%
2% (200rcd)	1,998	1,750	342	9	51	1	17.12%	2.55%
3% (300rcd)	1,993	1,750	343	15	51	1	17.21%	2.56%
5% (500rcd)	1,984	1,750	339	22	48	2	17.09%	2.42%
10% (1000rcd)	1,956	1,750	339	50	44	3	17.33%	2.25%
20% (2000rcd)	1,887	1,750	322	91	39	6	17.06%	2.07%
30% (3000rcd)	1,766	1,750	310	133	32	8	17.55%	1.81%

表8 マッチングの結果：ケース 3，ランダム・スワッピング，10%サンプリングデータ

スワッピング率	国調における標本一意の数	住調における標本一意の数	マッチング		真のマッチング		国調の標本一意に対してマッチングされたレコードの比率	国調の標本一意に対する真のマッチングの比率
			マッチングされたレコード数	スワッピングされたレコード数	マッチングされたレコード数	スワッピングされたレコード数		
1% (100rcd)	2,407	2,388	116	1	35	0	4.82%	1.45%
2% (200rcd)	2,400	2,388	114	3	34	0	4.75%	1.42%
3% (300rcd)	2,393	2,388	113	5	34	0	4.72%	1.42%
5% (500rcd)	2,385	2,388	112	10	32	1	4.70%	1.34%
10% (1000rcd)	2,348	2,388	108	21	27	1	4.60%	1.15%
20% (2000rcd)	2,259	2,388	98	39	19	2	4.34%	0.84%
30% (3000rcd)	2,048	2,388	84	49	13	4	4.10%	0.63%

場合でもその比率はケース 3 の場合で約 60%となっている。したがって、国調の匿名化ミクロデータの中でスワッピングされたレコードは、住調の個票データとマッチングされたレコードのすべてとは重複していないことがわかる。このことは、本研究では、スワッピング率を上げても、国調の匿名化ミクロデータのレコードの中で住調の個票データとマッチングされたレコードの一部に対してのみ、スワッピングが施されていることを意味する。

ところで、本研究では、国調の標本一意に対する「真のマッチング」の比率を算出している。本研究における真のマッチングとは、国調の匿名化ミクロデータと住調の個票データの間で 1 対 1 にマッチングされたレコードにおいて、調査区も同一であることが確認されることである。そこで、マッチングされたレコードが真のマッチングに該当するレコードかどうかを確認するために、国調と住調でマッチングされたレコードの組を対象に、それらのレコードの調査区が同一か否かについて検証作業を行った。標本一意に占める真のマッチングの比率を確認すると、ターゲット・スワッピングおよびランダム・スワッピングのいずれについても、その比率は約 1%～2%あり、非常に低くなっている。また、マッチングされたレコードの中で真のマッチングに該当するレコードの比率は、約 10%～30%となっている。このことは、国調と住調においてマッチングされたレコードの組については、同じ調査区であっても、それらが同一の世帯のレコードに該当する可能性が低いことを示している[11]。これらの結果を踏まえると、住調を外部情報として想定した場合、国調における匿名化ミクロデータの露見リスクは低いと言うことができよう。

４．国勢調査ミクロデータにおける有用性の検証

本節では、有用性の評価を行うため、サンプリングおよびスワッピングにおける誤差の検証を行った。サンプリングにおける誤差の検証については、テストデータからのサンプルデータにおいてサンプリング率を上げた場合に、サンプリングにおけるキー変数以外の属性について原データとの差を確認した。その一方で、スワッピングが適用された匿名化ミクロデータ(以下「スワッピング済データ」と呼ぶ。)における有用性については、クラメールの Ⅴ といった関連性の指標や原データからの絶対距離の平均値(average absolute distance)の計測等を用いることが考えられるが、本研究では、キー変数以外の属性を対象にスワッピング済データと原データとの分布の差を計測し、スワッピングにおける誤差の評価を試みた。

表９～表14は、それぞれ年齢と世帯人員の相対度数を例に、サンプリング率が１%、

[11] サンプリング率が 1%と 5%の場合においても、標本一意に占める真のマッチングの比率は、ターゲット・スワッピングおよびランダム・スワッピングのいずれについても、その比率は約 1%～2%あり、非常に低くなっている。なお、マッチングされたレコードの中で真のマッチングに該当するレコードの比率は、約 5%～20%となっている。

表９　年齢におけるサンプリングの誤差，サンプリング率 1%

	相対度数(%)	95%信頼区間最小値	95%信頼区間最大値	信頼区間外のサンプルの数
15～19歳	0.8	0.25	1.35	1
20～24歳	3.8	2.61	4.97	5
25～29歳	4.4	3.14	5.67	9
30～34歳	6.0	4.55	7.48	5
35～39歳	5.7	4.24	7.10	0
40～44歳	6.4	4.87	7.89	3
45～49歳	7.8	6.13	9.44	5
50～54歳	10.3	8.47	12.23	7
55～59歳	12.2	10.15	14.18	3
60～64歳	10.0	8.14	11.84	5
65～69歳	9.0	7.20	10.72	6
70～74歳	8.9	7.17	10.68	5
75～79歳	7.6	5.99	9.27	3
80～84歳	4.5	3.26	5.83	3
85歳以上	2.6	1.62	3.58	4

表 10　世帯人員におけるサンプリングの誤差，サンプリング率 1%

	相対度数(%)	95%信頼区間最小値	95%信頼区間最大値	信頼区間外のサンプルの数
1人	25.0	22.35	27.69	4
2人	25.3	22.64	28.00	2
3人	18.4	15.98	20.76	6
4人	15.4	13.20	17.65	2
5人	8.0	6.34	9.69	2
6人	4.8	3.48	6.12	3
7人	2.3	1.35	3.18	4
8人以上	0.8	0.24	1.33	4

表 11　年齢におけるサンプリングの誤差，サンプリング率 5%

	相対度数(%)	95%信頼区間最小値	95%信頼区間最大値	信頼区間外のサンプルの数
15～19歳	0.8	0.56	1.04	1
20～24歳	3.8	3.27	4.30	0
25～29歳	4.4	3.85	4.96	1
30～34歳	6.0	5.37	6.66	3
35～39歳	5.7	5.04	6.29	0
40～44歳	6.4	5.72	7.04	0
45～49歳	7.8	7.06	8.51	0
50～54歳	10.3	9.52	11.17	0
55～59歳	12.2	11.28	13.05	1
60～64歳	10.0	9.18	10.80	1
65～69歳	9.0	8.19	9.73	0
70～74歳	8.9	8.15	9.69	0
75～79歳	7.6	6.91	8.35	0
80～84歳	4.5	3.98	5.11	1
85歳以上	2.6	2.17	3.03	0

表 12　世帯人員におけるサンプリングの誤差，サンプリング率 5%

	相対度数(%)	95%信頼区間最小値	95%信頼区間最大値	信頼区間外のサンプルの数
1人	25.0	23.85	26.19	0
2人	25.3	24.14	26.49	0
3人	18.4	17.32	19.41	2
4人	15.4	14.45	16.40	1
5人	8.0	7.28	8.75	0
6人	4.8	4.22	5.38	3
7人	2.3	1.86	2.67	0
8人以上	0.8	0.55	1.02	1

表 13　年齢におけるサンプリングの誤差，サンプリング率 10%

	相対度数(%)	95%信頼区間最小値	95%信頼区間最大値	信頼区間外のサンプルの数
15～19歳	0.8	0.64	0.97	0
20～24歳	3.8	3.43	4.14	0
25～29歳	4.4	4.03	4.79	0
30～34歳	6.0	5.57	6.46	0
35～39歳	5.7	5.24	6.10	0
40～44歳	6.4	5.93	6.83	0
45～49歳	7.8	7.29	8.28	1
50～54歳	10.3	9.78	10.91	0
55～59歳	12.2	11.56	12.77	0
60～64歳	10.0	9.43	10.54	0
65～69歳	9.0	8.43	9.49	0
70～74歳	8.9	8.39	9.45	0
75～79歳	7.6	7.14	8.12	0
80～84歳	4.5	4.16	4.93	0
85歳以上	2.6	2.30	2.89	0

表 14　世帯人員におけるサンプリングの誤差，サンプリング率 10%

	相対度数(%)	95%信頼区間最小値	95%信頼区間最大値	信頼区間外のサンプルの数
1人	25.0	24.21	25.82	0
2人	25.3	24.51	26.13	0
3人	18.4	17.65	19.09	1
4人	15.4	14.76	16.10	0
5人	8.0	7.51	8.52	0
6人	4.8	4.40	5.20	0
7人	2.3	1.99	2.54	0
8人以上	0.8	0.62	0.95	0

５％と 10%におけるサンプリングの誤差を示したものである。相対度数は、母集団であるテストデータにおける相対度数を示している。なお、色づきの網掛けされた箇所は、95%信頼区間を設定した場合に、サンプルデータで計測した相対度数が、95%信頼区間から外れたサンプルの数を示している。例えば、サンプリング率が１％の年齢の相対度数を見ると、25～29 歳の年齢階級区分では、95%信頼区間の外にあるサンプルの数は 9 となっている。本分析結果では、一部の分類区分については、サンプリングの誤差が大きくなっているものの、全般的にはサンプリングの誤差は小さいことが確認できる。

一方、表 15 と表 16 はそれぞれ、世帯人員を例に、ターゲット・スワッピングとランダム・スワッピングを適用した場合の原データとスワッピング済データにおける分布の差を図示したものである。また、本分析では、サンプリング率が 10%のサンプルを用いている。本表においては、スワッピング率を上げるにしたがって、95%信頼区間におけるサンプリングの誤差よりスワッピングの誤差が上回っているかどうかが確認される。本分析結果によれば、例えば、サンプリング率が 10%の場合、ターゲット・スワッピングとランダム・スワッピングのおのおのを適用した場合のスワッピングの誤差は、スワッピング率が 30%の場合を除けば、サンプリングの誤差の範囲に収まっていることが確認されている。

５．スワッピングにおける有用性と秘匿性の評価

第３節では、外部情報とのマッチングの観点から、国調の匿名化ミクロデータにおける秘匿性を検証した。その一方で、特殊な一意に該当すると思われるレコードの秘匿性の程度については、十分に把握されているとは言えない。

このようなレコードレベルの秘匿性の程度を把握するためには、主として特殊な一意のレコードを対象に適用されたスワッピングの有効性を評価する必要がある。そこで、本節では、スワッピング済データを対象に、その有用性と秘匿性の評価を行った 。

本研究では、Shlomo *et al.*(2010)や伊藤・星野(2014)に基づいて，有用性と秘匿性の評価指標を作成した。有用性の評価指標は、(4)式で示されるような絶対距離の平均値で与えられる。

$$\text{有用性の評価指標}(DU)=\frac{\sum_{c}\left|T^{S}(c)-T^{O}(c)\right|}{n_{T}} \qquad (4)$$

$T^{O}(c)$：原データを用いて作成したクロス表におけるセルの度数

表 15　原データとスワッピング済データにおける分布の差：世帯人員，
ターゲット・スワッピング

	原データ	スワッピング率							95%信頼区間	95%信頼区間
		1%	2%	3%	5%	10%	20%	30%	最小値	最大値
1人	24.9	24.9	24.9	24.8	24.8	24.7	24.9	23.8	24.2	25.8
2人	25.7	25.7	25.7	25.6	25.6	25.7	25.4	25.3	24.5	26.1
3人	18.1	18.1	18.1	18.1	18.2	18.2	18.1	18.3	17.6	19.1
4人	15.5	15.5	15.6	15.7	15.6	15.5	15.6	15.9	14.8	16.1
5人	8.0	8.0	8.0	8.0	8.0	8.2	8.1	8.2	7.5	8.5
6人	4.7	4.7	4.7	4.7	4.7	4.7	4.7	5.1	4.4	5.2
7人	2.3	2.4	2.3	2.4	2.4	2.3	2.4	2.6	2.0	2.5
8人以上	0.8	0.8	0.7	0.7	0.7	0.8	0.8	0.8	0.6	1.0
合計	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0		

表 16　原データとスワッピング済データにおける分布の差：世帯人員，
ランダム・スワッピング

	原データ	スワッピング率							95%信頼区間	95%信頼区間
		1%	2%	3%	5%	10%	20%	30%	最小値	最大値
1人	24.9	24.8	24.8	24.8	24.7	24.8	24.7	23.7	24.2	25.8
2人	25.7	25.7	25.7	25.6	25.8	25.9	25.8	25.5	24.5	26.1
3人	18.1	18.2	18.2	18.1	18.1	18.2	18.3	18.4	17.6	19.1
4人	15.5	15.6	15.6	15.7	15.6	15.5	15.5	15.9	14.8	16.1
5人	8.0	8.0	8.0	8.0	8.0	7.9	7.9	8.1	7.5	8.5
6人	4.7	4.7	4.7	4.7	4.6	4.6	4.7	5.1	4.4	5.2
7人	2.3	2.3	2.3	2.4	2.3	2.4	2.4	2.5	2.0	2.5
8人以上	0.8	0.8	0.8	0.7	0.8	0.8	0.8	0.8	0.6	1.0
合計	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0		

$T^S(c)$：スワッピング済データを用いて作成したクロス表におけるセルの度数

n_T ：集計表におけるセルの数

(4)式では、原データとスワッピング済データの両方で集計値を作成した上で、セルごとの度数の差の絶対値に関する平均値が算出される。

一方、本研究では、秘匿性の評価の尺度として(5)式の指標を用いる。

$$\text{秘匿性の評価指標}\,(DR) = \frac{\sum_c I\left(T^O(c)=1, T^S(c)=1\right)}{\sum_c I\left(T^O(c)=1\right)} \qquad (5)$$

$\sum_c I\left(T^O(c)=1\right)$:原データにおけるクロス表の中で度数 1 であるセルの数

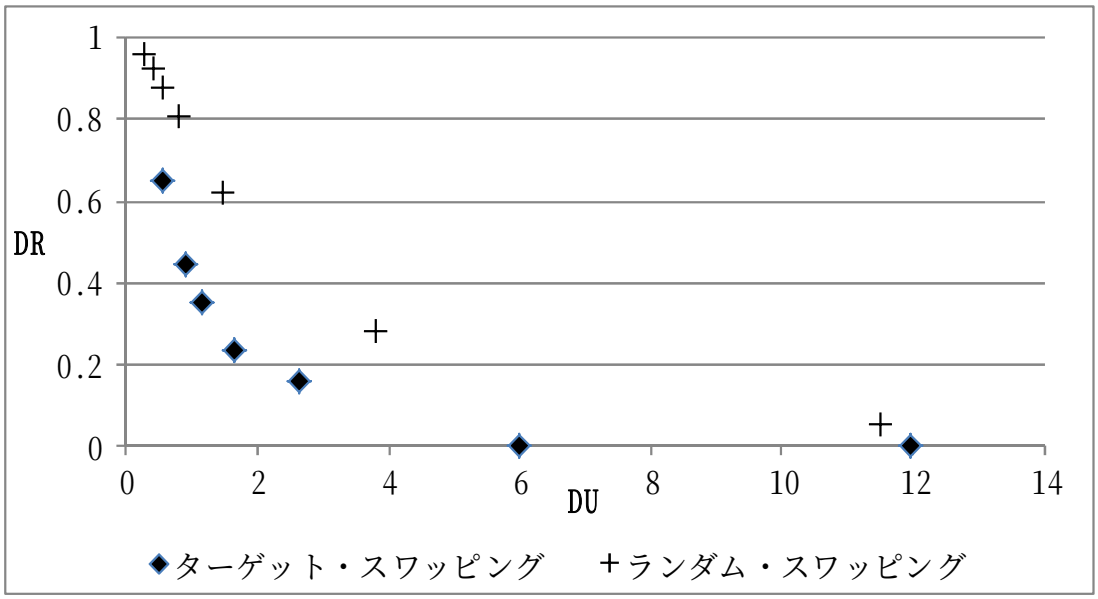
$\sum_c \left(T^O(c)=1, T^S(c)=1\right)$:スワッピング済みデータにおけるクロス表の中で度数 1 でありかつスワッピングされていないセルの数

これらの有用性と秘匿性の評価指標に基づいて、本研究では、キー変数の中のあらゆる 2 変数の組み合わせについて評価指標を計算するだけでなく、その指標に基づいて R-U マップ(R-U Confidentiality Map)を作成し、有用性と秘匿性の相対比較を試みた[12]。なお、R-U マップで使用する有用性と秘匿性の評価指標に関しては、これらの 2 変数の組み合わせにおいて算出された評価指標の平均値がそれぞれ用いられている。

図 2 は、R-U マップをもとにターゲット・スワッピングとランダム・スワッピングにおいて、スワッピング率を変えた場合の有用性の評価指標(DU)と秘匿性の評価指標(DR)の比較を行ったものである。図 2 を見ると、スワッピング率が上昇するほど、DU の数値が大きくなる傾向にあることが確認できる。また、ランダム・スワッピングと比較して、ターゲット・スワッピングにおける DU の数値が大きいことが確認できる。このことは、ターゲット・スワッピングにおける有用性がランダム・スワッピングにおけるそれと比較して低いことを示唆している。その一方で、DR は、スワッピング率が高いほど小さくなるだけでなく、ランダム・スワッピングにおける DR と比べて、ターゲット・スワッピングにおける DR が低くなる傾向にある。このことから、ターゲット・スワッピングにおける秘匿性の程度は、ランダム・スワッピングよりも高いことが確認できる。

12 キー変数の中のあらゆる 2 変数の組み合わせに基づく有用性の評価指標(DU)と秘匿性の評価指標(DR)の一覧表については参考表「有用性と秘匿性の評価結果の一覧表」を参照されたい。

図２　R-U マップの結果



次に、スワッピング率が 5%の場合のターゲット・スワッピングに着目すると、その DR は、スワッピング率が 30%の場合を除けば、ランダム・スワッピングにおける DR よりも低くなっている。それに対して、スワッピング率 5%におけるターゲット・スワッピングの DU は、スワッピング率が 10%未満の場合のランダム・スワッピングにおける DU よりも高くなっている。このことは、有用性あるいは秘匿性に関する閾値をどこに設定するかによって、ターゲット・スワッピングかランダム・スワッピングの選択、さらにはスワッピング率の設定が変わる可能性があることを示唆している。

６．おわりに

本稿では、国勢調査のミクロデータを用いて、各種匿名化技法の有効性の検証を行った。外部情報とのマッチングの試みとして、国調の匿名化ミクロデータと住調の個票データとのマッチングを行ったが、マッチングの精度は高くないことから、住調を外部情報とみなした上で、外部情報とのマッチングの可能性という観点から見た場合、露見のリスクは低いとみなすことができる。さらに、追加的な匿名化手法としてスワッピングを適用することによって、特殊な一意に該当するようなレコードが特定化されるリスクを低減することでき、秘匿性の強度を高めることが実証的に明らかになった。

また、本研究では、R-U マップをもとに、スワッピング済データにおける有用性と秘匿性の検証も行った。有用性あるいは秘匿性に関する閾値を設定することが

できれば、適切なスワッピング率およびスワッピングの方法を選択することが可能になることがわかった。

我が国でも、政府統計のミクロデータにおける匿名化手法に関する研究成果を生かしながら、統計実務における匿名化ミクロデータの作成可能性とミクロデータ利用のニーズへの対応の両面から、政府統計ミクロデータの作成・提供における将来展望を図っていくことが必要と考える。その意味では、本研究における秘匿性を評価するための外部情報とのマッチング手法や匿名化技法としてのスワッピングの手法は、将来的には、利用者のニーズを踏まえながら国勢調査の匿名データの作成方法を議論する上でも、有益な研究事例になると思われる。今後も、我が国において匿名化手法の有効性に関する研究および秘匿性と有用性の評価に関する実証研究を進めていくことが求められよう。

参考文献

Bethlehem, J. M., Keller, W. J., Pannekoek, J.(1990) “Disclosure Control of Microdata”, *Journal of American Statistical Association*, Vol. 85, pp.38-45.

Domingo-Ferrer, J. and Torra, V. (2001) “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.

Elliot, M. J. and Dale, A. (1998) *Disclosure Risk for Microdata*. Report to the European Union ESP/ 204 62/DG III.

Elliot, M.(2001) “Disclosure Risk Assessment”, Doyle *et al*.(eds.*) Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, Elsevier Science, Amsterdam, pp.75-90.

Elliot, M. J. and Manning, A.(2004) “The Methodology used for the 2001 SARs Special Uniques Analysis”, Paper Presented to An Open Meeting on the Samples of Anonymised Records form the 2001 Census, CCSR.

濱砂敬郎(2000)「事実上の匿名性の原則」松田芳郎・濱砂敬郎・森博美編『講座ミクロ統計分析① 統計調査制度とミクロ統計の開示』日本評論社, 109～128 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』Vol.22, No.2, 1～17 頁

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

伊藤伸介・星野なおみ(2014) 「国勢調査ミクロデータを用いたスワッピングの有効性の検証」『統計学』107 号，1～16 頁

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.

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

Shlomo, N., 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.

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.

付表１　原データとテストデータにおける符号表の比較―労働力状態と従業上の地位―

	原データ		テストデータ	
労働力状態	1	主に仕事	1	就業者
	2	家事などのほか仕事	2	完全失業者
	3	通学のかたわら仕事	3	家事
	4	仕事を休んでいた（休業者）	4	通学
	5	仕事を探していた（完全失業者）	5	その他
	6	家事	V	不詳（労働力状態が不詳，聞き取り調査世帯）
	7	通学		
	8	その他		
	V	不詳（労働力状態が不詳，聞き取り調査世帯）		
従業上の地位	1	雇用者（常雇）	1	雇用者（役員を含む）
	2	雇用者（臨時雇）	2	自営業主（家庭内職者を含む）
	3	雇用者（役員）	3	家族従業者
	4	自営業主（雇人のある業主）	V	不詳（従業上の地位不詳，聞き取り調査世帯）
	5	自営業主（雇人のない業主）		
	6	家族従業者		
	7	家庭内職者		
	V	不詳（従業上の地位不詳，聞き取り調査世帯）		

付表２－１　マッチングの結果：ケース 1, 1%サンプリングデータ

スワッピングの種類	スワッピング率	国調における標本一意の数	住調における標本一意の数	マッチング		真のマッチング		国調の標本一意に対してマッチングされたレコードの比率	国調の標本一意に対する真のマッチングの比率
				マッチングされたレコード数	スワッピングされたレコード数	マッチングされたレコード数	スワッピングされたレコード数		
ターゲットスワッピング	1%	383	866	51	1	0	0	13.4%	0.1%
	2%	383	866	51	1	0	0	13.4%	0.1%
	3%	383	866	51	2	0	0	13.4%	0.1%
ランダムスワッピング	1%	383	866	51	0	0	0	13.4%	0.1%
	2%	384	866	51	1	0	0	13.4%	0.1%
	3%	383	866	51	2	0	0	13.4%	0.1%

付表２－２ マッチングの結果：ケース 2, 1%サンプリングデータ

スワッピングの種類	スワッピング率	国調における標本一意の数	住調における標本一意の数	マッチング		真のマッチング		国調の標本一意に対してマッチングされたレコードの比率	国調の標本一意に対する真のマッチングの比率
				マッチングされたレコード数	スワッピングされたレコード数	マッチングされたレコード数	スワッピングされたレコード数		
ターゲットスワッピング	1%	540	1750	83	1	1	0	15.3%	0.1%
	2%	539	1750	83	3	1	0	15.3%	0.1%
	3%	538	1750	82	4	1	0	15.3%	0.1%
ランダムスワッピング	1%	540	1750	82	1	1	0	15.3%	0.1%
	2%	540	1750	82	2	1	0	15.3%	0.1%
	3%	540	1750	82	3	1	0	15.3%	0.1%

付表２－３ マッチングの結果：ケース３, 1%サンプリングデータ

スワッピングの種類	スワッピング率	国調における標本一意の数	住調における標本一意の数	マッチング		真のマッチング		国調の標本一意に対してマッチングされたレコードの比率	国調の標本一意に対する真のマッチングの比率
				マッチングされたレコード数	スワッピングされたレコード数	マッチングされたレコード数	スワッピングされたレコード数		
ターゲットスワッピング	1%	569	2388	23	0	5	0	4.0%	0.8%
	2%	568	2388	23	1	5	0	4.0%	0.8%
	3%	568	2388	23	2	5	0	4.0%	0.8%
ランダムスワッピング	1%	568	2388	23	2	5	0	4.0%	0.8%
	2%	569	2388	23	1	5	0	4.0%	0.9%
	3%	569	2388	23	1	5	0	4.0%	0.8%

付表３－１　マッチングの結果：ケース 1, 5%サンプリングデータ

スワッピングの種類	スワッピング率	国調における標本一意の数	住調における標本一意の数	マッチング		真のマッチング		国調の標本一意に対してマッチングされたレコードの比率	国調の標本一意に対する真のマッチングの比率
				マッチングされたレコード数	スワッピングされたレコード数	マッチングされたレコード数	スワッピングされたレコード数		
ターゲットスワッピング	1%	775	866	154	2	1	0	19.9%	0.1%
	2%	773	866	153	4	1	0	19.8%	0.1%
	3%	772	866	153	5	1	0	19.9%	0.1%
ランダムスワッピング	1%	775	866	154	2	1	0	19.9%	0.1%
	2%	776	866	154	3	1	0	19.8%	0.1%
	3%	774	866	154	4	1	0	19.8%	0.1%

付表３－２ マッチングの結果：ケース 2, 5%サンプリングデータ

スワッピングの種類	スワッピング率	国調における標本一意の数	住調における標本一意の数	マッチング		真のマッチング		国調の標本一意に対してマッチングされたレコードの比率	国調の標本一意に対する真のマッチングの比率
				マッチングされたレコード数	スワッピングされたレコード数	マッチングされたレコード数	スワッピングされたレコード数		
ターゲットスワッピング	1%	1429	1750	260	4	2	0	18.2%	0.2%
	2%	1422	1750	260	8	2	0	18.3%	0.2%
	3%	1419	1750	261	12	2	0	18.4%	0.2%
ランダムスワッピング	1%	1435	1750	260	3	3	0	18.1%	0.2%
	2%	1434	1750	260	7	3	0	18.1%	0.2%
	3%	1428	1750	258	10	2	0	18.1%	0.2%

付表３－３ マッチングの結果：ケース 3 , 5%サンプリングデータ

スワッピングの種類	スワッピング率	国調における標本一意の数	住調における標本一意の数	マッチング		真のマッチング		国調の標本一意に対してマッチングされたレコードの比率	国調の標本一意に対する真のマッチングの比率
				マッチングされたレコード数	スワッピングされたレコード数	マッチングされたレコード数	スワッピングされたレコード数		
ターゲットスワッピング	1%	1631	2388	82	1	20	0	5.0%	1.3%
	2%	1624	2388	81	3	20	0	5.0%	1.2%
	3%	1618	2388	80	5	19	0	5.0%	1.1%
ランダムスワッピング	1%	1637	2388	82	2	21	0	5.0%	1.3%
	2%	1634	2388	82	3	21	0	5.0%	1.3%
	3%	1630	2388	81	4	20	0	5.0%	1.2%

スワッピングの適用可能性に関する評価研究

―国勢調査ミクロデータを用いて―

伊藤伸介*、星野なおみ**

要旨

現在提供している国勢調査の匿名データの作成においては、外部情報とのマッチングによる特定化のリスクを低減するために、国勢調査における全国レベルの公表された結果表の中の度数 1 と 2 に該当するレコードの削除が行われている。しかしながら、このようなレコードの削除は、大変な労力を伴うため、匿名化措置として効率的な方法ではないように思われる。それに対して、結果表における度数 1 や 2 に該当するレコードを対象にスワッピングといった匿名化技法を適用することによって、結果表における特異なセルの削除を行う必要性がなくなることも考えられる。その一方で、国勢調査の匿名データの将来的な作成・提供の可能性を模索するにあたって、ミクロデータに対する匿名化技法の適用可能性を追究することは有用であることから、本稿では、スワッピングに焦点を当て、その適用可能性を追究した。

本研究では、平成 17 年の国勢調査の個票データを用いて、スワッピングの適用可能性を検証する。具体的には、ある都道府県における特定の市町村(以下「地域 A」)のレコードから作成したテストデータ(約 50,000 レコード)および同一の都道府県内の別の市町村(以下「地域 B」) から作成したテストデータ(約 10,000 レコード)である。本研究では、地域 A と地域 B のいずれにおいても、10%のリサンプリングを行った上で、各種のスワッピングの技法を適用した。また、各歳年齢区分の提供の可能性を追究するために。5 歳年齢区分だけでなく、各歳年齢区分の 2 種類の匿名化ミクロデータを作成し、検証を行った。つぎに、公表されている結果表の集計事項の中で外観識別性の高い属性を対象にしたクロス集計表を作成した上で、クロス集計表の一意に該当するレコードとスワッピングが施された匿名化ミクロデータ(以下「スワッピング済データ」と呼ぶ。)とのマッチングを行った。一方、「特殊な一意(special uniques)」に該当するレコードに対して適用したスワッピングの有効性を明らかにするために、各種のスワッピング済データにおける有用性と秘匿性の比較・検証を行った。

本分析結果によれば、スワッピング率が上がるにつれて、クロス集計表における度数 1 の

* （独）統計センター統計情報・技術部統計技術研究課非常勤研究員（中央大学経済学部准教授）
** （独）統計センター統計情報・技術部統計技術研究課

レコードの露見のリスクが低減することを確認することができた。さらに、年齢を各歳階級にした場合でもスワッピングを適用することによって、年齢 5 歳階級と同様に露見リスクを低減することができたことから、スワッピングの適用によって、匿名データの各歳年齢の提供も追究することが可能になると思われる。一方、R-U マップに基づく匿名化ミクロデータの秘匿性と有用性の検証結果からは、ランダム・スワッピングとターゲット・スワッピングにおける適切な割合を設定することによって、有用性を保持したまま、秘匿性をより高めるようなスワッピングの方法について検討することが可能なことがわかった。

スワッピングの適用可能性に関する評価研究

―国勢調査ミクロデータを用いて―

伊藤伸介、星野なおみ

1．本研究の目的

アメリカ、イギリス、カナダ等の欧米諸国の統計作成部局は、1960 年代以降、非攪乱的手法(non-perturbative methods)だけでなく、攪乱的手法(perturbation)を適用することによって、人口センサスのミクロデータを提供してきた(伊藤(2015))。それに対して、我が国でも、総務省統計局が、「匿名データの作成・提供に関するガイドライン」に沿った形で、平成 12 年と平成 17 年について国勢調査の匿名データを作成・提供している。現在、提供している国勢調査の匿名データにおいては、サンプリング、リコーディング、トップコーディング、ボトムコーディング、特異なレコードの削除といった非攪乱的手法を適用している。その一方で、国勢調査に関して公表されている結果表において度数１を含むセルの中で特異な分布を持つセルが散見される場合(例えば、詳細な地域における結果表で度数１が含まれる場合等)、そのセルからミクロデータに含まれる個体情報が露見される可能性がある。こうした可能性を考慮すると、レコード削除では不十分なことから、攪乱的手法としてスワッピングを適用している。さらに、外部情報とのマッチングよる特定化のリスクを低減するために、全国レベルの公表された結果表の中の度数１と２に該当するレコードの削除も行われている。

しかしながら、公表されている結果表に対して、このような度数１ないしは２に該当するレコードを削除するのは、大変な労力を伴うことから、統計実務の観点からは効率的な匿名化措置ではないようにも思われる。もし、これらの度数１や２に該当するレコードを対象にスワッピングのような攪乱的手法を適用することによって、公表された結果表の特異なセルに該当するレコードの削除を行う必要性がなくなるのであれば、統計リソースの有効活用の観点から、より効果的な匿名データの作成が可能になるかもしれない。

他方、将来的には、地域分析用の匿名データや各歳年齢区分の匿名データ等、別のタイプの国勢調査の匿名データの要望が出てくる可能性があることから、その予備的な研究とし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ことは有用であると考えられる。

こうした国勢調査の匿名データの将来的な作成・提供の可能性を追究するために、本稿では、ミクロデータにおける匿名化技法の１つであるスワッピングに焦点を当て、

スワッピングが適用された匿名化ミクロデータ(以下「スワッピング済データ」と呼ぶ。)を対象に、秘匿性と有用性の両面からスワッピングの有効性を探ることにしたい。

2．本研究におけるスワッピングの方法

本研究では、平成 17 年の国勢調査の個票データを用いて、スワッピングの適用可能性を検証する。本研究では、ある都道府県における特定の市町村(以下「地域 A」と呼ぶ。)のレコードから作成したテストデータ(約 50,000 レコード)および同一の都道府県内の別の市町村(以下「地域 B」と呼ぶ。) から作成したテストデータ(約 10,000 レコード)である。これらのテストデータには、 個人単位で抽出した一般世帯の世帯主のレコードのみが含まれている。

本研究のために、地域 A と地域 B のいずれにおいても、10%のリサンプリングを行った上で、各種のスワッピングの技法が適用されている。具体的なスワッピングの手順としては、以下のように行われる(伊藤・星野(2014), 伊藤・星野(2015))。

第１に、キー変数(key variable)を用いて標本一意(sample unique)を計測し、スワッピングの対象となるレコードを選び出す。使用するキー変数は、表１で示される 12 変数であり、年齢については 5 歳年齢階級と各歳年齢階級の 2 つのパターンでスワッピングが行われた。

第 2 に、標本一意となるレコードを対象に、キー変数のすべての組み合わせでクロス集計を行い、ある特定のレコードが標本一意に該当した回数をスコアとして計測する。そのスコアに基づいて、スワッピングの対象レコードの中で、優先度の高いレコードを探り出す。

第 3 に、対象となるレコードに対してスワッピングを施す。本研究では、ターゲット・スワッピング(targeted data swapping)、(2)ランダム・スワッピング(random data swapping)だけでなく、(3)ターゲット・スワッピングとランダム・スワッピングの併用も行った。

本研究では、標本一意に該当した回数が 1 回以上のレコードをスワッピングの候補となるレコードとして選出する。本研究においては、キー変数 12 変数のすべての組み合わせについてスコアを計算し、10 ファイルにおけるスコアの平均値を算出した。表２－１と表２－２はそれぞれ、5 歳年齢階級と各歳年齢階級におけるスワッピング済データの基本統計量を示したものである。本表から、地域 A(50,000 レコード)におけるサンプルデータ(5,000 レコード)の場合、スコアの平均値、中央値、最小値と最大値はそれぞれ、556、478、21 と 1,772 であるのに対して、地域 B(10,000 レコード)におけるサンプルデータ(1,000 レコード)の場合、スコアの平均値、中央値、最小値と最大値はそれぞれ、749、707、58 と 1,828 になっていることがわかる。一方、各歳年齢階級のスワッピング済データでは、地域 A の場合、スコアの平均値、中央値、最小値と最大

表１　本研究において標本一意の計測のために用いたキー変数

変数		区分数
男女の別		13
年齢	5 歳年齢階級	25
	各歳年齢階級	122
配偶関係		5
国籍		13
労働力状態		9
従業上の地位		8
産業大分類		19
職業大分類		10
住居の種類		9
建て方の種類		5
建物の階数 (建物の階数については共同住宅のみ)		30
世帯の住んでいる階 (建て方の種類を考慮。世帯の住んでいる階については共同住宅のみ)		30

表２－１　スワッピング済データの基本統計量―5 歳年齢階級

		平均値	中央値	最小値	最大値
10,000 レコード	サンプル1	727.71	665	56	1,794
	サンプル2	749.58	734	64	1,840
	サンプル3	727.44	690	64	1,754
	サンプル4	772.90	736	112	1,826
	サンプル5	726.97	651	32	1,840
	サンプル6	745.23	708	56	1,794
	サンプル7	764.56	722	32	1,832
	サンプル8	762.01	736	64	1,882
	サンプル9	766.50	726	48	1,850
	サンプル10	744.53	704	56	1,865
	平均値	749	707	58	1,828
50,000 レコード	サンプル1	556.53	468	16	1,740
	サンプル2	559.20	478	32	1,838
	サンプル3	560.53	495	24	1,767
	サンプル4	542.99	449	20	1,698
	サンプル5	547.60	464	16	1,718
	サンプル6	552.77	476	16	1,758
	サンプル7	568.77	497	20	1,724
	サンプル8	559.61	480	12	1,902
	サンプル9	566.16	494	32	1,801
	サンプル10	550.54	476	24	1,778
	平均値	556	478	21	1,772

表２－２　スワッピング済データの基本統計量―各歳年齢階級

		平均値	中央値	最小値	最大値
10,000レコード	サンプル1	848.34	816	128	1,834
	サンプル2	852.51	779	112	1,858
	サンプル3	852.06	848	128	1,824
	サンプル4	870.06	864	64	1,856
	サンプル5	832.47	768	64	1,840
	サンプル6	859.47	818	128	1,824
	サンプル7	863.57	848	128	1,858
	サンプル8	889.35	896	128	1,905
	サンプル9	857.23	832	128	1,886
	サンプル10	865.46	832	128	1,894
	平均値	859	830	114	1,858
50,000レコード	サンプル1	655.32	607	32	1,764
	サンプル2	653.72	608	32	1,862
	サンプル3	658.78	608	64	1,818
	サンプル4	642.11	584	48	1,812
	サンプル5	652.57	596	32	1,772
	サンプル6	653.56	597	48	1,782
	サンプル7	667.19	624	48	1,770
	サンプル8	658.05	608	24	1,915
	サンプル9	658.27	608	48	1,869
	サンプル10	645.72	576	48	1,814
	平均値	655	602	42	1,818

値はそれぞれ、655、602、42 と 1,818 であるが、地域 B の場合、スコアの平均値、中央値、最小値と最大値はそれぞれ、859、830、114、1,858 となっている。本分析結果からは、年齢階級の区分が細かくなることによってスコアの平均値が大きくなっているが、地域 B においては、5 歳年齢階級区分と各歳年齢区分でスコアの最大値がそれほど変わらないのが興味深い。

次に、本研究では、レコード数全体に対するスワッピング率として 1%、5%、10%を設定した上で、

(1)ターゲット・スワッピングについては、スコアの高い上位 p%(p はスワッピング率)に該当するレコードをスワッピングの対象レコードとした。

(2)ランダム・スワッピングに関しては、スワッピングの候補となるレコードから p%ランダムに選別されたレコードをスワッピングの対象レコードとした。

(3)ターゲット・スワッピングとランダム・スワッピングの併用の場合、スコアの高い上位 1/2p%にターゲット・スワッピングを適用した上で、残りのレコードから 1/2p%ランダムに選別されたレコードにランダム・スワッピングを適用した。

本研究では、 スワッピングの対象レコードに対してその入れ替えの候補となるレコードについては、地域 A や地域 B とは異なる地域(以下「地域 C」と呼ぶ。)を対象にドナーファイル(約 5,000 レコード)から探索する。

スワッピングの対象となるレコードは、「特殊な一意(special uniques)」[1]として出現する可能性が高い。このようなレコードとキー変数の値が完全に一致するレコードがドナーファイルで見つかる可能性は低いことから、本研究においても、伊藤・星野(2015)と同様に、スワッピングの対象レコードに対して、ドナーファイルに含まれるレコードとの距離を計測し、ドナーファイルの中で最も距離が小さいレコードとスワッピングを行った。具体的には、質的属性値間の距離(distance for categorical variables)を定義した上で(Domingo-Ferrer and Torra(2001,pp.105-106))、スワッピングの対象レコードとドナーファイルとの間の距離計測型リンケージを行い(Domingo-Ferrer and Torra(2001), Takemura(1999))、リンケージの結果からキー変数ごとのスコアを算定する。次に、各変数のスコアに基づいて、スワッピングの対象レコードとドナーファイルの距離を計測した上で、ドナーファイルの中でもっとも距離が小さいレコードと置き換える。なお、本研究における距離計測型リンケージは、伊藤・星野(2015)と同様の以下の手順で行われる。

①i(i=1,...,m)および j(j=1,...,n)を、それぞれスワッピング対象レコードの番号およびドナーファイルのレコード番号とし(m と n は、それぞれスワッピング対象レコードの数およびドナーファイルのレコード数)、k(j=1,...,11)をキー変数の番号[2]とする。このとき、i 番目のレコードにおけるキー変数 k の分類区分の数値をCs_{ki}、j 番目のドナーファイルのレコードにおけるキー変数 k の分類区分の数値をCd_{kj}に基づいて、キー変数 k に関する i と j の質的属性値間の距離を次の(１)式で計測する (Domingo-Ferrer and Torra, (2001a,pp.105-106))。

$$Sd_{kij} = \left| Cs_{ki} - Cd_{kj} \right| \tag{1}$$

なお、年齢および住居の建て方の「共同住宅」以外の場合、$\left| Cs_{ki} - Cd_{kj} \right| > 0$であれば、$Sd_{kij} = 1$とする。

②質的属性値間の距離をスコア化するために、k 番目のキー変数における分類区分数C_kでSd_{kij}を除することによって、k 番目のキー変数におけるスコアである$Score_{kij}$を算出する((２)式)。

$$Score_{kij} = \frac{1}{C_k} \cdot Sd_{kij} \tag{2}$$

③各キー変数のスコアの総計を求めることによって、i 番目と j 番目のレコード間の

[1] 特殊な一意の定義については、伊藤・星野(2015)の脚注４を参照されたい。

[2] マッチングの実験では、「建物の階数」と「世帯の住んでいる階」については、それぞれ「建て方の種類」と組み合わせた変数が用いられる。ゆえに、距離計測型リンケージで用いるキー変数は 11 となっている。

距離に関する総合指標 D_{ij} を導出する((３)式)。

$$D_{ij} = \sum_k Score_{kij} \quad (3)$$

④総合指標 D_{ij} に基づいて、スワッピングの対象レコードとドナーファイルとの間の距離計測型リンケージを行う。

表３は、スワッピングの対象レコードとドナーファイルに含まれるレコードとの距離に関する平均値、中央値、最大値と最小値を示したものである。ランダム・スワッピングの場合、距離の最小値がほぼ 0 となっていることから、スワッピングの対象レコードとキー変数の値がほぼ同一であるレコードと入れ替えられていることがわかる。それに対して、ターゲット・スワッピングの場合、ランダム・スワッピングと比較して、距離の最小値が大きいレコードとの入れ替えが行われている。このことは、効果的なスワッピングを行うためには、スワッピングの対象レコードとドナーファイルのレコードとの距離に基づいて、スワッピングの種類が選ばれなくてはならないことを示唆している。

3.マッチングによる秘匿性の評価研究

先述のように既存の結果表を外部情報の１つとみなした場合、結果表から匿名化ミクロデータに含まれる個体情報が特定化されるリスクが考えられる。そこで、本節では、秘匿性の評価方法の１つとして、外部情報と匿名化ミクロデータのマッチングに焦点を当てることにしたい。本研究では、国勢調査において既存の結果表を外部情報とみなした上で、スワッピング済データとのマッチングを行った。

一方、国勢調査においてすべての既存の結果表の中から度数 1 ないしは 2 に該当するセルを探索するのは大変な労力を伴うことから、本研究を行う上では効率的な作業とは言えない。そこで、結果表の中で外観識別性の高い属性のみを選び出し、その属性のすべての組み合わせに関するクロス表を作成し、そのクロス表を擬似的な結果表とみなした。外観識別性の高い属性は、以下の 6 変数である。

・男女の別
・住居の種類
・住居の建て方
・建物の階数
・世帯の住んでいる階
・延べ床面積(14 区分)

表３　スワッピングの対象レコードとドナーファイルに含まれるレコードとの距離

5 歳年齢階級

	ランダム											
	1%スワッピング				5%スワッピング				10%スワッピング			
	最大値	最小値	平均値	中央値	最大値	最小値	平均値	中央値	最大値	最小値	平均値	中央値
5000レコード	0.527	0.000	0.136	0.109	0.733	0.000	0.136	0.105	0.789	0.000	0.141	0.113
1000レコード	0.240	0.000	0.081	0.060	0.315	0.000	0.074	0.050	0.398	0.000	0.077	0.051

	ターゲット											
	1%スワッピング				5%スワッピング				10%スワッピング			
	最大値	最小値	平均値	中央値	最大値	最小値	平均値	中央値	最大値	最小値	平均値	中央値
5000レコード	0.882	0.122	0.433	0.421	0.896	0.039	0.337	0.315	0.896	0.003	0.274	0.250
1000レコード	0.579	0.155	0.335	0.328	0.619	0.033	0.235	0.213	0.619	0.010	0.195	0.177

	ターゲット&ランダム											
	1%スワッピング				5%スワッピング				10%スワッピング			
	最大値	最小値	平均値	中央値	最大値	最小値	平均値	中央値	最大値	最小値	平均値	中央値
5000レコード	0.869	0.004	0.300	0.260	0.970	0.000	0.271	0.227	0.979	0.000	0.248	0.210
1000レコード	0.559	0.004	0.231	0.200	0.607	0.000	0.180	0.159	0.626	0.000	0.153	0.139

各歳年齢階級

	ランダム											
	1%スワッピング				5%スワッピング				10%スワッピング			
	最大値	最小値	平均値	中央値	最大値	最小値	平均値	中央値	最大値	最小値	平均値	中央値
5000レコード	0.475	0.000	0.103	0.066	0.609	0.000	0.109	0.076	0.700	0.000	0.113	0.076
1000レコード	0.209	0.002	0.062	0.044	0.338	0.000	0.067	0.039	0.370	0.000	0.069	0.040

	ターゲット											
	1%スワッピング				5%スワッピング				10%スワッピング			
	最大値	最小値	平均値	中央値	最大値	最小値	平均値	中央値	最大値	最小値	平均値	中央値
5000レコード	0.930	0.113	0.451	0.436	0.964	0.045	0.373	0.342	0.964	0.023	0.325	0.292
1000レコード	0.561	0.136	0.320	0.309	0.621	0.021	0.239	0.218	0.621	0.008	0.198	0.181

	ターゲット&ランダム											
	1%スワッピング				5%スワッピング				10%スワッピング			
	最大値	最小値	平均値	中央値	最大値	最小値	平均値	中央値	最大値	最小値	平均値	中央値
5000レコード	0.857	0.001	0.287	0.254	0.953	0.000	0.256	0.213	0.964	0.000	0.238	0.200
1000レコード	0.552	0.002	0.209	0.164	0.604	0.000	0.167	0.148	0.621	0.000	0.148	0.135

スワッピングに使用されていない延べ床面積が、外観識別性の高い属性の１つに含まれていることに留意されたい。

次に、これらの 6 変数を用いて作成したクロス表において度数 1 に該当するレコードを対象にスワッピング済データとの間でマッチングを行った。これは、スワッピングを行うことによって、既存の結果表において度数１に該当すると思われるレコードが低減するかどうかの検証を目指している。なお、50,000 レコードのデータ及び 10,000 レコードのデータの中でクロス集計表の度数 1 に該当するレコードは、それぞれ 776 レコードと 260 レコードである。

表４と表５は、それぞれ 5 歳年齢階級と各歳年齢階級におけるスワッピング済データとクロス表において度数１に該当するレコードとのマッチングの結果を示したものである。表４を例にすると、地域 A(50,000 レコード)の 5 歳年齢階級において、10%のスワッピング率でランダム・スワッピングを適用した場合に、クロス表に含まれる度数１のレコードと１対１でマッチングし、マッチングされたレコードがスワッピング済である比率は 0.13%であるのに対して、1 対 1 でマッチングするが、該当するレコードがスワッピングされていない比率は 77.96%となっている。次に、クロス表に含まれる度数１のレコードと１対２でマッチングし、マッチングされたレコードの一部がスワッピング済である比率は 1.03%であるのに対して、1 対 2 でマッチングされ、該当するレコードのすべてがスワッピングされている比率は 0.13%となっている。さらに、度数１のレコードと１対ｎ(ｎは 3 以上)でマッチングする比率は 1.80%であり、マッチングしないレコードの比率は 18.94%となっている。

それに対して、ターゲット・スワッピングを適用した場合(5 歳年齢階級、10%のスワッピング率)、マッチングしないレコードの比率は 75.90%となっている。また、ターゲット・スワッピングとランダム・スワッピングの併用については、マッチングしないレコードの比率は 57.99%である。

一方、各歳年齢階級の場合(表５)、10%のスワッピング率でランダム・スワッピングを適用したときのマッチングしないレコードの比率は、13.66%であるのに対して、ターゲット・スワッピングにおいてマッチングしないレコードの比率は 75.64%となっている。さらに、ターゲット・スワッピングとランダム・スワッピングの併用については、マッチングしないレコードの比率は 56.06%である。

本分析結果を見ると、各歳年齢階級における秘匿性の程度は、5 歳年齢階級におけるそれと比較して、基本的には変わらないことが確認できる[3]。したがって、公表されて

[3] 本研究では、クロス表において母集団一意に該当するレコードと年齢区分が異なる場合のスワッピング済データ(ターゲット・スワッピング、スワッピング率は 10%)とのマッチングの結果についてさらなる検証を行った。具体的には、年齢 5 歳階級におけるスワッピング済データと年齢各歳階級におけるスワッピング済データを対象に、スワッピングの際にドナーファイルにおいて入れ替えの対象となったレコードを確認した。付表１は、5 歳年齢階級と各歳年齢階級においてドナーファイルのレコードが重複する比率を示したものである。それによると、50,000 レコードの場合と 10,000 レコードの場合においては、ドナーファイルのレコードが重複する比率は、それぞれ 66.92%と 46.52%であった。このことから、5 歳年齢

表４　スワッピング済データとクロス表におけるマッチングの結果、5 歳年齢階級

50,000 レコード

スワッピング率	ランダム			ターゲット			ターゲット＆ランダム		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
１：１ 該当するレコードはスワッピング済	0.00%	0.00%	0.13%	0.13%	0.00%	0.13%	0.00%	0.00%	0.00%
１：１ 該当するレコードはスワッピングされてない	96.78%	87.24%	77.96%	81.96%	43.81%	21.39%	88.14%	59.41%	39.18%
１：２ 該当するレコードの一部はスワッピング済	0.64%	0.90%	1.03%	0.52%	0.26%	0.13%	0.77%	0.39%	0.26%
１：２ 該当するレコードの全てがスワッピング済	0.00%	0.00%	0.13%	0.00%	0.13%	0.00%	0.00%	0.13%	0.00%
１：ｎ（nは３以上）	0.00%	0.90%	1.80%	0.77%	1.93%	2.45%	0.64%	1.93%	2.58%
マッチングしない	2.58%	10.95%	18.94%	16.62%	53.87%	75.90%	10.44%	38.14%	57.99%

10,000 レコード

スワッピング率	ランダム			ターゲット			ターゲット＆ランダム		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
１：１ 該当するレコードはスワッピング済	0.00%	0.77%	1.15%	0.00%	0.77%	1.92%	0.00%	0.38%	0.77%
１：１ 該当するレコードはスワッピングされてない	98.46%	91.92%	80.77%	83.46%	43.08%	25.00%	90.00%	59.23%	36.92%
１：２ 該当するレコードの一部はスワッピング済	1.15%	2.31%	3.85%	2.31%	1.92%	1.92%	1.54%	2.31%	2.69%
１：２ 該当するレコードの全てがスワッピング済	0.00%	0.00%	0.00%	0.00%	1.92%	1.15%	0.00%	0.38%	1.92%
１：ｎ（nは３以上）	0.00%	0.00%	0.77%	0.00%	2.31%	4.23%	0.00%	1.54%	2.31%
マッチングしない	0.38%	5.00%	13.46%	14.23%	50.00%	65.77%	8.46%	36.15%	55.38%

階級のスワッピング済データと各歳年齢階級のスワッピング済データにおけるスワッピングの対象になったレコードの多くについては、ドナーファイルにおいて入れ替えの対象になったレコードの大部分が重複していたことがわかった。ドナーファイルにおいて入れ替えの対象となるレコードが重複している理由としては、スワッピングで用いる距離計測型リンケージにおける計算式の特徴から、(２)式で算出される年齢のスコアは小さくなるため、そのスコアが総合指標 D に与える影響は小さいことが指摘できる。このことは、5 歳年齢階級と各歳年齢階級の年齢区分に関わりなく、年齢以外の変数によってドナーファイルで入れ替えの対象となるレコードが決まることを意味している。

なお、付表１からは、5 歳年齢階級のスワッピング済データと各歳年齢階級のスワッピング済データにおいて入れ替えの対象となったレコードが重複していない場合においても、スワッピング後の変数の大半に関して同じ値を持つレコードが存在することが確認できる。具体的には、10,000 レコードの場合、5 歳年齢階級と各歳年齢階級において入れ替えの対象となったレコードは重複していないものの、床面積を除く 5 変数が一致している割合は、8.07%となっている。

表５　スワッピング済データとクロス表におけるマッチングの結果、各歳年齢階級

50,000 レコード

スワッピング率	ランダム			ターゲット			ターゲット&ランダム		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
１：１ 該当するレコードはスワッピング済	0.00%	0.00%	0.26%	0.00%	0.00%	0.00%	0.00%	0.26%	0.00%
１：１ 該当するレコードはスワッピングされてない	97.68%	90.59%	83.63%	82.86%	44.20%	21.39%	88.66%	61.86%	41.24%
１：２ 該当するレコードの一部はスワッピング済	0.77%	0.90%	0.64%	0.39%	0.13%	0.26%	0.52%	0.52%	0.13%
１：２ 該当するレコードの全てがスワッピング済	0.00%	0.00%	0.13%	0.13%	0.00%	0.00%	0.00%	0.26%	0.13%
１：ｎ（nは３以上）	0.00%	1.16%	1.68%	0.90%	2.32%	2.71%	0.90%	1.68%	2.45%
マッチングしない	1.55%	7.35%	13.66%	15.72%	53.35%	75.64%	9.92%	35.44%	56.06%

10,000 レコード

スワッピング率	ランダム			ターゲット			ターゲット&ランダム		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
１：１ 該当するレコードはスワッピング済	0.00%	0.00%	0.00%	0.00%	0.38%	2.31%	0.00%	0.00%	0.77%
１：１ 該当するレコードはスワッピングされてない	98.85%	92.31%	88.46%	83.08%	42.31%	25.38%	91.92%	61.92%	39.62%
１：２ 該当するレコードの一部はスワッピング済	0.38%	1.54%	2.69%	1.15%	3.08%	2.69%	0.38%	2.31%	3.46%
１：２ 該当するレコードの全てがスワッピング済	0.00%	0.00%	0.00%	0.00%	1.15%	1.54%	0.00%	0.77%	1.54%
１：ｎ（nは３以上）	0.00%	0.38%	0.77%	0.77%	2.69%	3.46%	0.38%	1.15%	3.08%
マッチングしない	0.77%	5.77%	8.08%	15.00%	50.38%	64.62%	7.31%	33.85%	51.54%

いる結果表において特異なセルが出現した場合でも、それに該当するレコードに対してスワッピングを行うことによって、公表されている結果表を外部情報と考えた場合の露見リスクを低減することが可能になる。さらに、年齢の階級を各歳区分にした場合でも秘匿性の程度が大きく変わらなかったことから、本研究を踏まえると、国勢調査の匿名データに関しては、年齢を各歳区分で提供する可能性も考えられよう[4]。

4．スワッピングの有効性の評価に関する研究

前節では、外部情報とのマッチングの観点から、国調の匿名化ミクロデータにおける秘匿性を検証した。その一方で、特殊な一意に該当すると思われるレコードの秘匿性の程度を把握する上で、主として特殊な一意のレコードを対象に適用されたスワッピングの有効性を評価することも考えられる。そこで、本研究では、伊藤・星野(2015)と同様に、スワッピング済データにおける有用性と秘匿性の評価を行った。最初に、Shlomo *et al.*(2010)や伊藤・星野(2014, 2015)に基づいて、以下の(４)式で示される有用性の評価指標と(５)式で示される秘匿性の評価指標を算出した。

$$\text{有用性の評価指標}\ (DU)=\frac{\sum_c \left|T^S(c)-T^O(c)\right|}{n_T} \tag{4}$$

$T^O(c)$：原データを用いて作成したクロス表におけるセルの度数
$T^S(c)$：スワッピング済データを用いて作成したクロス表におけるセルの度数
n_T：集計表におけるセルの数

$$\text{秘匿性の評価指標}\ (DR)=\frac{\sum_c I\left(T^O(c)=1, T^S(c)=1\right)}{\sum_c I\left(T^O(c)=1\right)} \tag{5}$$

$\sum_c I\left(T^O(c)=1\right)$:原データにおけるクロス表の中で度数 1 であるセルの数
$\sum_c \left(T^O(c)=1, T^S(c)=1\right)$:スワッピング済みデータにおけるクロス表の中で度数 1 であり かつスワッピングされていないセルの数

[4] 本研究では、延べ床面積を入れなかった 5 変数においてもマッチングの実験を行った上で、比較・検証を行った。検証結果によれば、延べ床面積を含む６変数におけるマッチングの結果と大きな違いは見られなかった。これについては、次のように考えることができる。スワッピングに用いたキー変数のみの 5 変数の場合、スワッピング率を上げると、母集団一意となるレコードのほとんどがスワッピング済レコードになることから、スワッピングによる秘匿の効果は高くなる。一方で、スワッピングに用いたキー変数以外の変数(延べ床面積)を加えた 6 変数の場合でも、スワッピング率を上げることによって、キー変数 5 変数の場合と同様に、秘匿の効果を高めると考えることができる。このことは、キー変数の一部を選択した上で、キー変数以外の変数との組み合わせを考えた場合、それに関する結果表において母集団一意に該当するセルが存在した場合でも、スワッピング率の程度によっては、秘匿を高めることが可能なことを示している。

次に、これらの有用性と秘匿性の評価指標を用いて、R-U マップ(R-U Confidentiality Map)を作成し、有用性と秘匿性の相対評価を行った。なお、R-U マップで使用する有用性と秘匿性の評価指標に関しては、スワッピングデータとのマッチングのために作成された擬似的なクロス集計表で用いられた 6 変数の中のあらゆる 2 変数の組み合わせについて計算された評価指標の平均値がそれぞれ使用されている[5]。

図１は、R-U マップをもとにターゲット・スワッピング、ランダム・スワッピングおよびターゲット・スワッピングとランダム・スワッピングを併用した場合に、スワッピング率を変化に伴う有用性の評価指標(DU)と秘匿性の評価指標(DR)の比較結果を図示したものである。図１を見ると、スワッピング率の上昇に伴って、DU の数値が大きくなることがわかる。また、ランダム・スワッピングと比較して、ターゲット・スワッピングにおける DU の数値が大きいことから、ターゲット・スワッピングにおける有用性がランダム・スワッピングにおけるそれと比較して低いことが確認できる。それに対して、DR は、スワッピング率の上昇に伴って小さくなる傾向にある。さらに、ランダム・スワッピングにおける DR は、ターゲット・スワッピングにおける DR と比較して高くなることから、ターゲット・スワッピングにおける秘匿性の程度は、ランダム・スワッピングよりも高いことが確認できる。なお、ターゲット・スワッピングとランダム・スワッピングを併用した場合の R-U マップは、ターゲット・スワッピングとランダム・スワッピングの中間に位置していることから、全般的に見て、ターゲット・スワッピングのみを用いた場合よりも有用性は高く、秘匿性が低くなっていることがわかる。

5．むすびにかえて

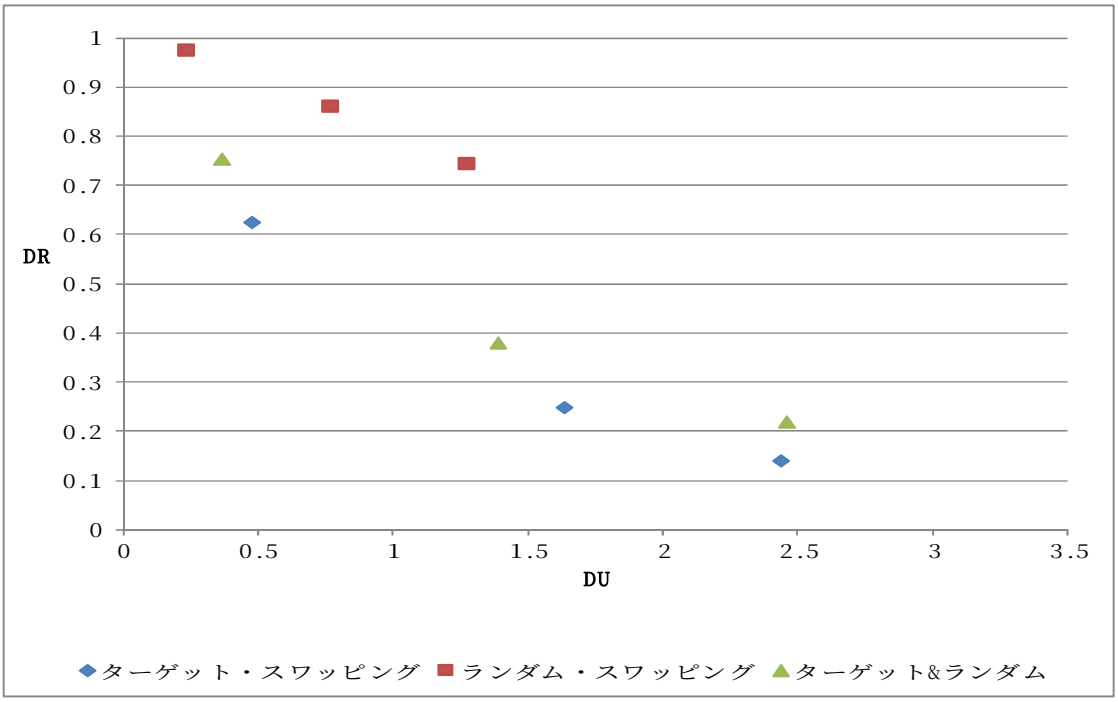
本稿では、国勢調査のミクロデータを用いて、スワッピングの適用可能性の検証を行った。国勢調査の既存の公表されている結果表が外部情報になりうることから、公表されている結果表の集計事項の中で外観識別性の高い属性を対象にしたクロス集計表を作成した上で、クロス集計表において度数１に該当するレコードと国調のスワッピング済データとのマッチングを行った。本研究では、スワッピング率が上がるにつれて、クロス集計表における度数 1 のレコードの露見のリスクが低減することを確認することができた。さらに、年齢については、各歳年齢階級にした場合でもスワッピングを適用した場合、5 歳年齢階級と同様に露見リスクを低減することができたことから、スワッピングの適用の仕方によっては、匿名データにおいて各歳提供の可能性も考えることができると思われる。

他方、本研究では、R-U マップをもとに、スワッピング済データにおける有用性と

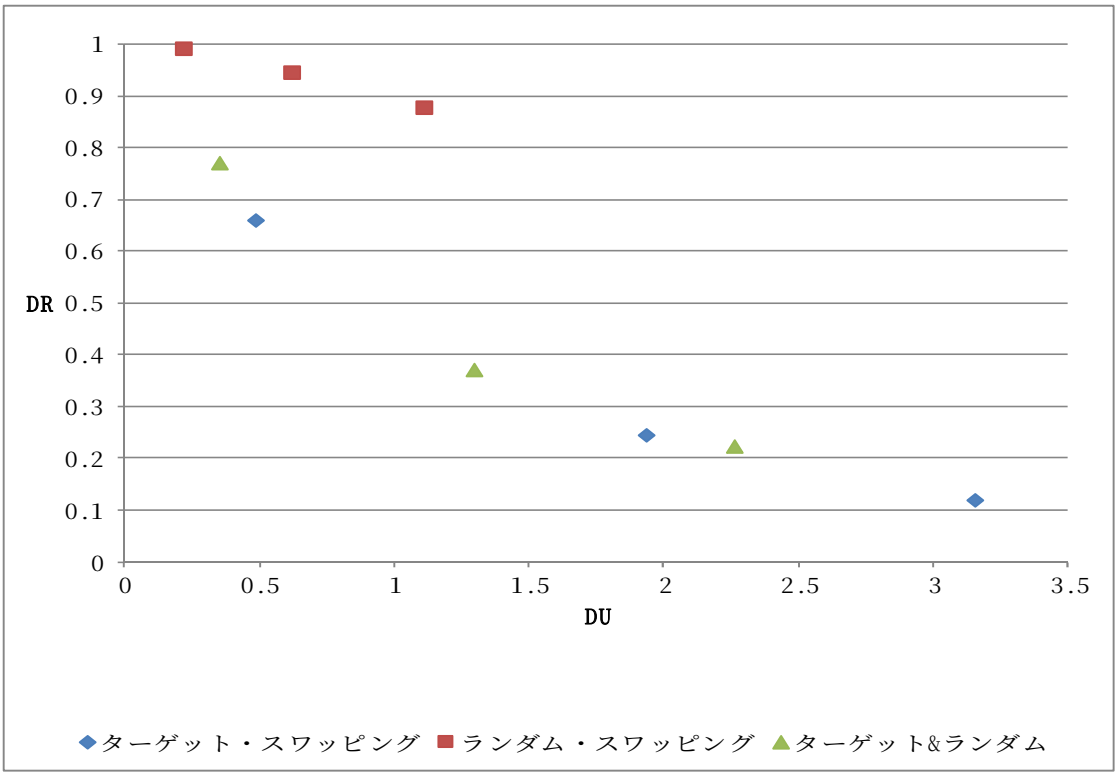
[5] 本分析における有用性の評価指標は、スワッピング済データに基づいて擬似的なクロス表を作成した場合の有用性を検証することを指向している。

図 1　R-U マップ、地域 A

5 歳年齢区分



各歳年齢区分



秘匿性の検証も行った。またターゲット・スワッピングとランダム・スワッピングを併用した場合の分析も行った。ランダム・スワッピングとターゲット・スワッピングにおける適切な割合を設定することによって、有用性を保持したまま、秘匿性をより高めるようなスワッピングの方法についても検討することが可能になると考えられる。

参考文献

Domingo-Ferrer, J. and Torra, V. (2001) “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.

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』Vol.22, No.2, 1〜17 頁

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

伊藤伸介・星野なおみ(2014)「国勢調査ミクロデータを用いたスワッピングの有効性の検証」『統計学』107 号，1〜16 頁

伊藤伸介(2015)「人口センサスにおけるミクロデータの作成状況について」，『統計』2015 年 1 月号，8〜13 頁

伊藤伸介・星野なおみ(2015)「ミクロデータにおける匿名化の誤差の評価に関する研究―国勢調査を例に―」(『製表技術参考資料』No.28 で刊行予定)

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.

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.

付表１　5 歳年齢階級と各歳年齢階級においてドナーファイルのレコードが重複する比率、ターゲット・スワッピング、スワッピング率 10%

	母集団一意	５歳年齢階級と各歳年齢階級でドナーファイルのレコードが重複する比率	５歳年齢階級と各歳年齢階級でドナーファイルのレコードが異なる比率			
			5変数同じ床面積同じ	5変数同じ床面積違う	5変数違う床面積同じ	5変数違う床面積違う
10,000レコード	260	66.92%	2.69%	5.38%	4.23%	20.77%
50,000レコード	776	46.52%	1.42%	3.99%	5.28%	42.78%

製 表 技 術 参 考 資 料 28

平成 27 年 3 月発行

編集・発行　　独立行政法人　統計センター

〒162－8668

東京都新宿区若松町 19－1

電　話　代表　03（ 5273 ）1200

統計情報・技術部統計技術研究課　　TEL ： 03－5273－1368

E-mail ： research@nstac.go.jp