

国勢調査マイクロデータを用いた匿名化技法の有効性の検証*

明海大学経済学部 伊藤 伸介**
(独)統計センター 星野 なおみ

1. はじめに

アメリカ、イギリス、カナダ等の欧米諸国では、人口センサスのマイクロデータが作成・提供されていることが知られている。一方、わが国ではこれまで、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のみが提供されてきたが、全数調査である国勢調査についても、平成 25 年中に匿名データが提供される予定となっている。提供予定の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)提供される年次は平成 12 年と 17 年であること、(4)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、小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マイ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マイ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マイ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別データ)に秘匿処理を施すことによって匿名化マイ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、スワッピングやノイズ(加法ノイズ等)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マイ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、H17 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」と呼称)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。

匿名化マイクロデータの作成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。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、各歳区分を 5 歳区分にリコーディングを施した上で(19 区分)、85 歳以上の区

* 本稿の内容は個人的な見解を示すものであり、(独)統計センターの見解を表すものではないことに留意されたい。

** (独)統計センター非常勤研究員

分についてトップコーディングを行っている。(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¹。(3)(1)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。

具体的なスワッピングの手順は、以下のとおりである(伊藤・星野(2013))。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。標本一意となるレコード数の算出のために使用するキー変数は、つぎの11変数である。

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

第2に、スワッピングの対象となるレコードの中で、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、標本一意であ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、対象となる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、そのレコードにスワッピングを行うランダム・スワッピング(random data swapping)を適用した。

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。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、本実験においてはスワッピング率として1%、2%、3%、5%、10%、20%および30%を設定した上で、ターゲット・スワッピングとランダム・スワッピングについてはそれぞれ以下のような方法でスワッピングの対象レコードを選別した。

(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。

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とは異なる地域(以下「地域B」とする)を対象にドナーファイル(約5,000レコード)から探索する。

¹ 本稿では、紙面の制約上、サンプリング率が10%における匿名化マイクロデータを用いて行った実験の結果のみをもとに議論を進めることにしたい。

つぎに、本研究でスワッピングにおいて適用されたリンケージの方法はつぎのとおりである(Domingo-Ferrer and Torra (2001), Takemura(1999), 伊藤・星野(2013))。

(1)キー変数 11 変数について地域 A におけるスワッピングの対象レコードとドナーファイルの中のレコードの間で一致するかどうか検討する。具体的には、年齢以外の 10 変数については、それぞれスワッピングの対象レコードに含まれる属性値とドナーファイルの中のレコードにおける値が一致する場合には 0、それ以外には 1 というスコアを新たに設定した上で(ただし「住居の建て方」で「共同住宅」の場合を除く)、分類区分数で割る(不詳も 1 つの区分と考える)。一方、年齢については、ドナーファイルの中のレコードにおける属性値からスワッピングの対象レコードに含まれる属性値を引いた上で、19 で割った値をスコアとして新たに付与する。また、「住居の建て方」で「共同住宅」に該当する場合には、ドナーファイルの中のレコードにおける属性値からスワッピングの対象レコードに含まれる属性値を引いた上で、30 で割った値をスコアとして新たに付与する。

(2)上記の 11 変数に関して値を合計して、以下の計測式に基づいて距離を計測する。

距離＝〔世帯主との続き柄の分類区分数の逆数〕×〔世帯主との続き柄のスコア〕

＋〔男女の別の分類区分数の逆数〕×〔男女の別のスコア〕

＋〔年齢 5 歳区分のスコア〕

・・・

＋〔建て方の種類の分類区分数の逆数〕×〔建て方の種類のスコア〕

(共同住宅の場合には、〔建て方の種類のスコア〕のみ)

この距離の絶対値を計測し、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。

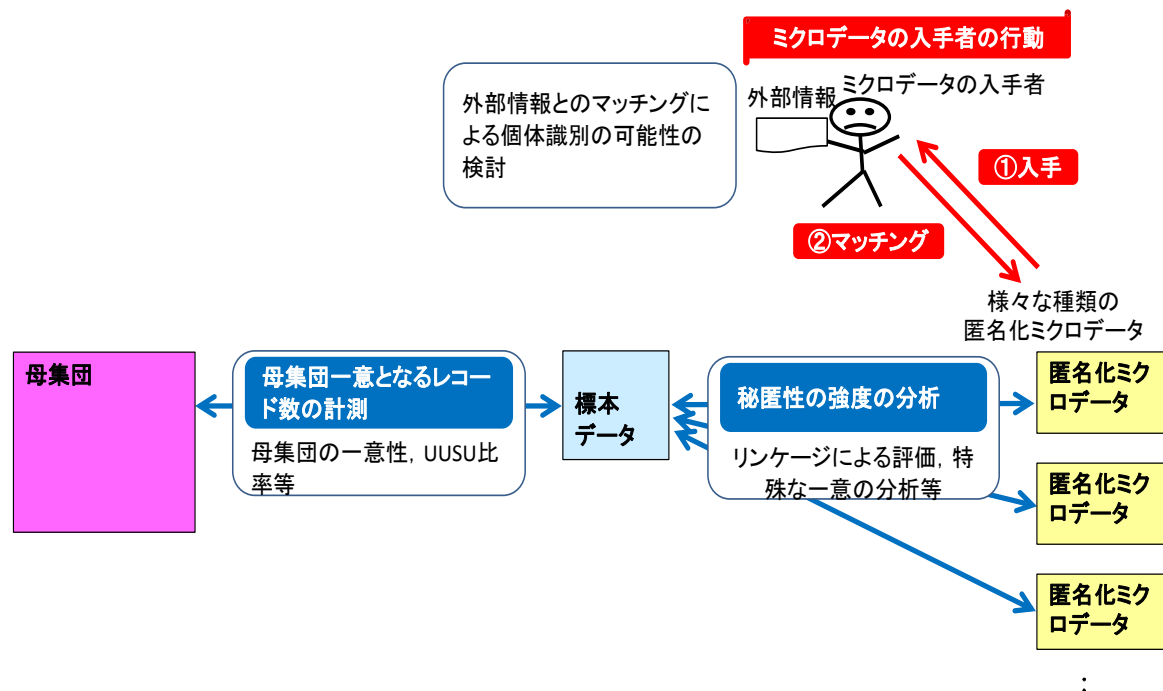
3. 国勢調査マイクロデータを用いた秘匿性の評価—マッチングの試み—

政府統計マイ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))。マイクロデータの入手者(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)を持っていたと想定する。そのとき、(1)識別ファイルに含まれるレコードとマイ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 のマッチングがなされ、(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るとき、個体識別が成立する。

マイクロデータの入手者(intruder)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。第 1 は、提供されるマイクロデータにおいて「母集団一意」となるレコード数を計測することである(Bethlehem *et al.*(1990), Marsh *et al.*(1991) 等)。第 2 は、外部情報の取得可能性および外部情報とマイクロデータのマッチングの実験による個体識別の可能性を検討することである(Müller *et al.*(1995))。

図 1 は、マイ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(population unique)の比率で表される「母集団の一意性(population uniqueness)」や標本一意に対する母集団一意かつ標本一意の比率である 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

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。

マイクロデータについては、秘匿処理を施していない原データと匿名化マイ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。他方で、このようにして作成された匿名化マイクロデータについては、マイクロデータの入手者の行動を想定し、識別の戦略²に基づいてマッチング率を計測することで、匿名化マイクロデータの秘匿性を定量的に評価することが可能である。

こうした先行研究を踏まえ、本節では、秘匿性の評価方法の1つとして、外部情報とマイ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、本研究では、国調以外の政府統計のマイクロデータを外部情報とみなした上で、匿名化マイ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、スワッピングが施された10%抽出の国調の匿名化マイクロデータに対して、H20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、国調のテスト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、例えば、ド

² Müller *et al.*(1995)で議論されているように、マイ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、マイ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とめようとする戦略である。それに対して、釣り検索とは、マイ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群を探り出そうとする戦略である。また、直接検索においては、対象となる個体のレコードがマイ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))をマイ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))。

ドイツにおける事実上の匿名性(factual anonymity)について実証的に明らかにするため、ドイツのマイクロセンサス(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。

本研究では、これらの先行研究を参考にしながら、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: 県市区町村番号、世帯人員、性別、年齢、配偶関係

ケース2: 県市区町村番号、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: 県市区町村番号、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、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、本研究では、国調の匿名化マイ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を行っている³。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、年次の調整がなされている。

表1～表6は、国調の匿名化マイクロデータと住調の個票データと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。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、国調の匿名化マイ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確認できる。例えば、表1～表3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マイ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2000～2400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が確認できる。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、国調の匿名化マイ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1～表6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。国調の匿名化マイ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、約4%と大きく減少している。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、国調の匿名化マイ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

³ 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。

表1 マッチングの結果：ケース1，ターゲット・スワッピング

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2 マッチングの結果：ケース2，ターゲット・スワッピング

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3 マッチングの結果：ケース3，ターゲット・スワッピング

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 4 マッチングの結果：ケース 1, ランダム・スワッピング

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 5 マッチングの結果：ケース 2, ランダム・スワッピング

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 6 マッチングの結果：ケース 3, ランダム・スワッピング

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%

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、国調の匿名化マイ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、本実験では、スワッピング率を上

げても、国調の匿名化マイ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、本実験では国調の標本一意に対する「真のマッチング」の比率が算出されている。本研究における真のマッチングとは、国調の匿名化マイ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、調査区番号も同一であることが確認されることである⁴。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であるレコードの比率は、約10%~30%となっている。本実験の結果から、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マイ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. 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、国調の匿名化マイ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(special uniques)」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定量的に評価する必要がある。そこで、本節では、10%抽出された国調の匿名化マイクロデータを対象に、その有用性と秘匿性の定量的な評価を行った⁵。

本研究では、Shlomo *et al.*(2010)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、(1)式で示されるような絶対距離の平均値(average absolute distance)で与えられる。

$$\text{有用性の評価指標}(DU) = \frac{\sum_c |T^P(c) - T^O(c)|}{n_T} \quad (1)$$

$T^O(c)$: 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

$T^P(c)$: 匿名化マイクロ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

(1)式では、原データと匿名化マイクロ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ている。

一方、本研究では、秘匿性の評価の尺度としてつぎの(2)式で与えられる指標を用いる。

⁴ 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、本実験では、統計センターの内部資料に基づいて、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番号が同一か否かについて検証作業を行った。

⁵ 有用性の評価を行うにあたって、本研究で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。世帯人員を例に挙げると、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、本研究の結果から確認されている。

$$\text{秘匿性の評価指標(DR)} = \frac{\sum_c I(T^O(c)=1, T^P(c)=1)}{\sum_c I(T^O(c)=1)} \quad (2)$$

$\sum_c I(T^O(c)=1)$: 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

$\sum_c (T^O(c)=1, T^P(c)=1)$: 匿名化マイクロ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、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。なお、R・U マップにおける有用性と秘匿性の評価指標に関してはそれぞれ、スワッピングで用いられるキー変数の中のある 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使用される。

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指標(DU)と秘匿性の指標(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。

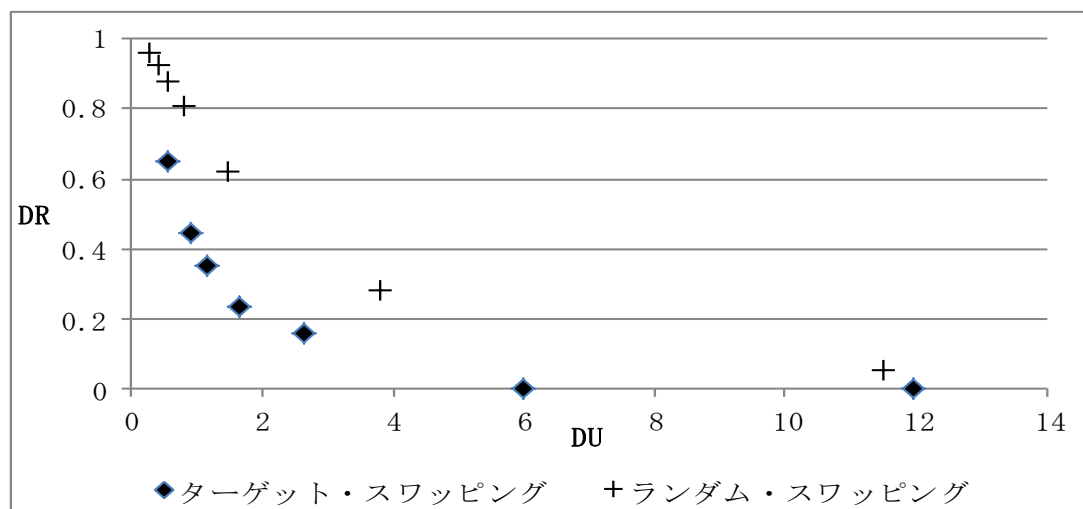
ところで、図 2 において、例えばスワッピング率が 5% の場合のターゲット・スワッピングに着目すると、その DR は、スワッピング率が 30% の場合を除けば、ランダム・スワッピングにおける DR よりも低くなっている。それに対して、スワッピング率 5% におけるターゲット・スワッピングの DU は、スワッピング率が 10% 未満の場合のランダム・スワッピングにおける DU よりも高くなっている。したがって、本図からは最適なスワッピングの方法やスワッピング率を選定することは困難だと思われる。このことは、有用性あるいは秘匿性に関する閾値に基づいて、ターゲット・スワッピングかランダム・スワッピングの選択、さらにはスワッピング率の設定を行う必要があることを示唆している。

5. むすびにかえて

本稿では、国勢調査のマイ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、国調の匿名化マイ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、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。

他方、本研究では、R・U マップを用いて、匿名化マイクロデータにおける有用性と秘匿性

図2 R-U マップの結果



の検証も行った。適切な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、望ましいスワッピングの方法およびスワッピング率を選択することが可能になることがわかった。

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.
- 伊藤伸介(2010)「マイ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—国勢調査マイ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』として刊行予定)
- 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.