経済統計学会　第57回(2013年度)全国研究大会(2013年9月13日～14日，於 静岡市産学交流センター)

ミクロデータにおける匿名化の誤差の検証―国勢調査を例に―

2013年9月13日

明海大学経済学部　伊藤 伸介
(独)統計センター　星野なおみ
(独)統計センター　後藤武彦

目次

1. 本報告の背景と目的
2. 本研究で使用する匿名化技法
3. 国勢調査ミクロデータを用いた秘匿性の検証
4. 国勢調査ミクロデータを用いた有用性の検証
5. おわりに

＊本報告の内容は、個人的見解を示すものであり、統計センターの見解を示すものではありません

1. 本報告の背景と目的

現在、総務省統計局で作成・提供されている匿名データ

住宅・土地統計調査
全国消費実態調査
就業構造基本調査
社会生活基本調査
労働力調査

標本調査の匿名データ

全数調査である国勢調査の匿名データの提供を望む声が多い

＊国勢調査の匿名データは、平成25年中に提供される予定

国勢調査の匿名データの主な特徴

●地域区分は、都道府県と人口50万以上市区
　→　細かい地域区分に対するニーズが高い

●データ量は母集団の1%、世帯単位で抽出
　→　1%あれば人口50万以上市区でも集計可能

●提供年次は、平成12年と平成17年
　→　2ヵ年あれば時系列での比較が可能

●項目は、世帯員の数、続柄、年齢、住宅の種類など

←「匿名データの作成・提供に関するガイドライン」に沿って匿名データが作成されている。

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、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。

本報告の目的

本報告では、各種匿名化技法を用いて作成された国勢調査のミクロデータ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試みることにしたい。

2.本研究で使用する匿名化技法

原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、例えば、以下のような匿名化技法を用いることが考えられる。

(1) 情報の削除(record suppression)

→特異なレコードの削除も含む

(2) 区分の再編(global recoding, top(bottom) coding)

→将来的にはlocal recodingの可能性あり

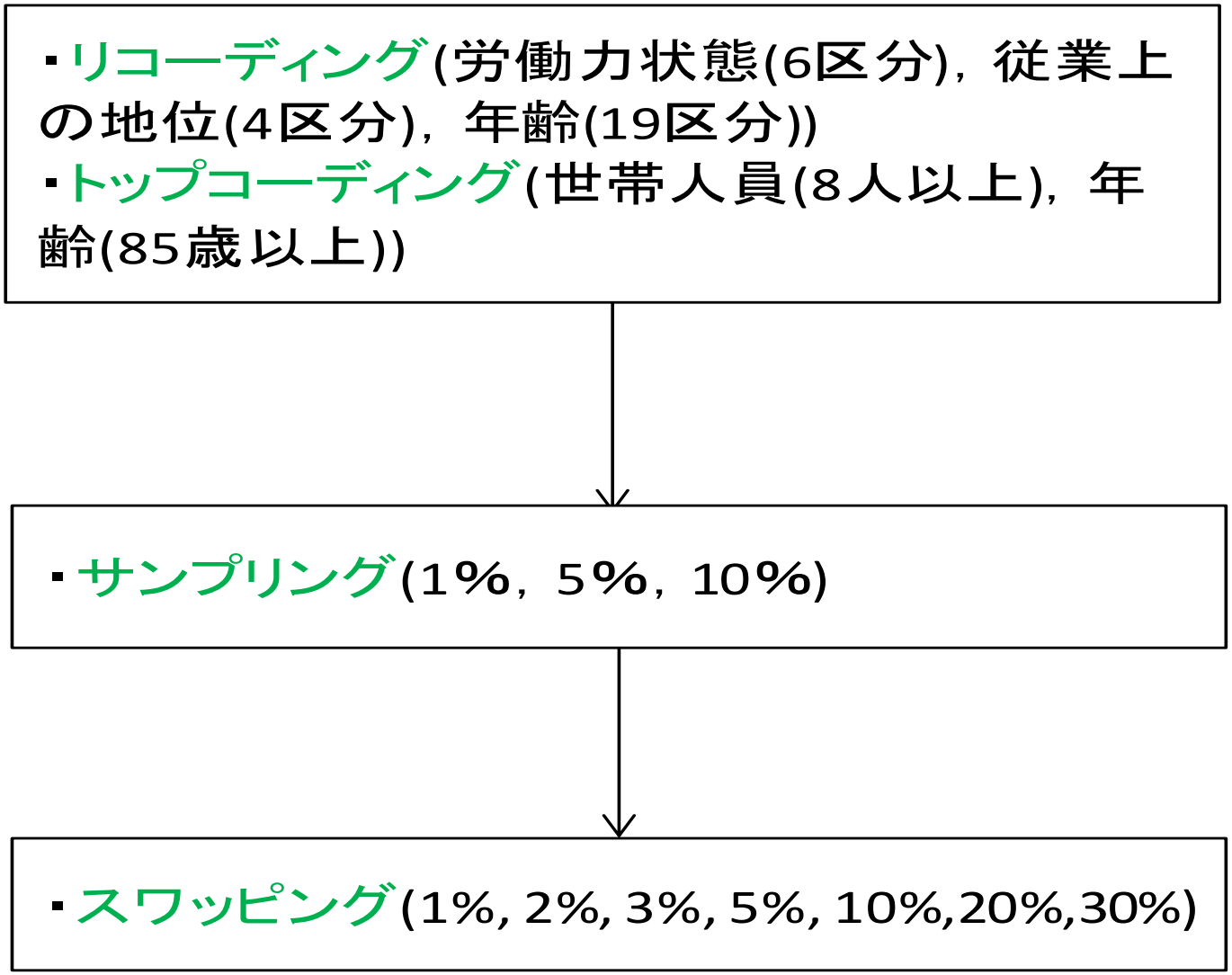
(3) サンプリング

→サンプリングの方法についても将来的な検討課題となりうる。

(4) 攪乱的手法(ex.スワッピング等(data swapping))

→ノイズ(加法ノイズ等)といった攪乱的手法についても適用可能性を議論する必要あり。

本研究における匿名化ミクロデータの作成手順



３．国勢調査ミクロデータを用いた秘匿性の検証

本実験では、国勢調査ミクロデータを用いて秘匿性の検証を行う。

使用するデータ: H17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

⇒個人単位で抽出した一般世帯の世帯主のみを対象

→本研究では、(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。

秘匿性と露見リスク

・政府統計ミクロデータに関する秘匿性
→諸外国では、主としてミクロデータに含まれる個体情報の露見リスク(disclosure risk)の評価として議論が展開

↓

本研究では、個体識別による露見リスクの定量的な評価方法を検討する。

参考　個体識別(identification)

(Bethlehem *et al.*(1990), Marsh *et al.*(1991), Müller *et al.*(1995))

＊識別の対象となる特定の個体について侵入者(intruder)が把握している情報(事前情報(a priori knowledge))を含むファイル(識別ファイル)とミクロデータファイルを想定

1)識別ファイルに含まれるレコードとミクロデータファイルに含まれるレコードがキー変数(key variable)を通じて1対1のマッチング(matching)がなされること

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こと

↓

識別が成立

参考　個体識別のイメージ(伊藤(2010))

侵入者

個人A

性別が男
年齢が85歳
8人世帯
農林漁業従事者

政府統計のミクロデータのイメージ

一連番号	（世帯主）性別	（世帯主）年齢	世帯人員区分	（世帯主）職業	年間収入（万円）	…
00001	2	25	1	9	200	…
00002	1	36	4	1	800	…
00003	**1**	**85**	**8**	**7**	**6000**	**…**
00004	2	45	3	5	400	…
00005	1	70	2	5	300	…
・	・	・	・	・	・	
・	・	・	・	・	・	
・	・	・	・	・	・	

性別　1: 男　2:女
職業　1:専門的・技術的職業従事者　2: 管理的職業従事者　3:事務従事者
4:販売従事者　5: サービス職業従事者　6:保安職業従事者　7: 農林漁業従事者
8: 運輸・通信従事者　9: 生産工程・労務作業者

侵入者が個人Aの属性値について情報を持っていた場合

侵入者がミクロデータを手に入れると、ミクロデータに含まれる属性群と個人Aに関して持っている情報に関して、**キー変数**(この場合は、性別、年齢、世帯人員区分、職業)によるマッチングを行うことによって、ミクロデータの中から個人Aに関する情報を特定しようとする。

個体識別による露見リスクの定量的な評価に関する先行研究(伊藤(2010))

(1)外部情報の取得可能性および外部情報とミ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(Müller *et al.*(1995))

(2)提供されるミ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(Bethlehem *et al.*(1990), Marsh *et al.*(1991)等)

秘匿性の検証について

(1)母集団一意(population unique)かつ標本一意(sample unique)の検証

・UUSU比率(母集団一意かつ標本一意(共通一意(union uniques))であるレコード数の標本一意となるレコード数に対する比率)による計測

→母集団一意かつ標本一意の対象レコードは追加的な匿名化技法の適用の対象となりうる。

(2)外部情報と匿名化ミクロデータとのマッチング

・外部情報の1つとして政府統計ミクロデータも想定することが可能である。

母集団一意と標本一意の考え方

キー変数：性別、世帯人員区分、職業

（標本）

		性別							
		男				女			
		世帯人員区分							
		2人	…	8人	…	2人	…	8人	…
職業	専門的・技術的職業従事者	27		5		31		4	
	：	21		7		47		5	
	農林漁業従事者	48		9		20		9	
	運輸・通信従事者	48		4		15		3	
	生産工程・労務作業者	14		1		48		1	

性別が男
8人世帯
生産工程・労務作業者

性別が女
8人世帯
生産工程・労務作業者

標本一意

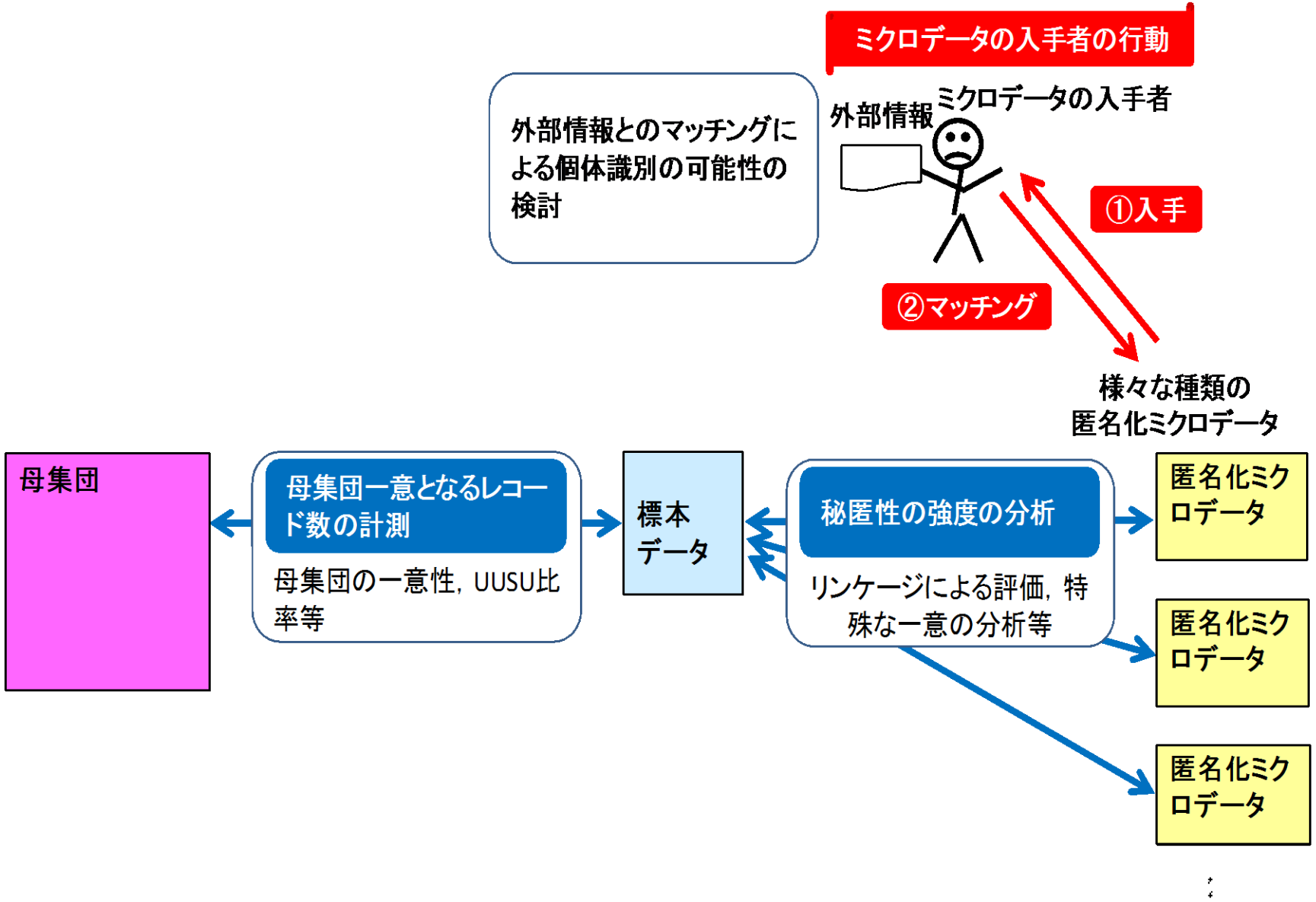
（母集団）

		性別							
		男				女			
		世帯人員区分							
		2人	…	8人	…	2人	…	8人	…
職業	専門的・技術的職業従事者	56		15		83		47	
	：	68		39		59		37	
	農林漁業従事者	57		45		77		33	
	運輸・通信従事者	83		39		67		48	
	生産工程・労務作業者	54		3		83		1	

母集団一意

標本一意かつ母集団一意

図　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))



侵入者における識別の戦略
(Müller *et al.*(1995))

1)直接検索(directed search)

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とめようとする戦略

2)釣り検索(fishing strategy)

ミクロデータのなか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

＊調査参加情報(participation knowledge)がある場合の直接検索戦略が最も開示リスクが高いシナリオだと考えられている。

(1)母集団一意かつ標本一意の計測

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を計測する。

母集団一意(標本一意)の計測のために使用するキー変数

以下のキー変数を用いて，母集団一意(標本一意)を計測する。

- 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)

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

母集団一意に該当するレコード数は14568レコード

(2)外部情報と匿名化ミクロデータとのマッチング

本研究では、国調のサンプリングデータ(10%)にスワッピングを施した上で、スワッピング済データに対して外部情報とのマッチングの実験を行う。

→国調のスワッピング済データとH20年住宅・土地統計調査の地域Aに該当するレコードを含んだ個票データ(約10,000レコード)とのマッチング

←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。

スワッピングの手順

(1) キー変数を用いて標本一意を計測し，スワッピングの対象となるレコードを探す(先述の11変数を使用)。

(2) 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のスコアをもとに探索する。

⇒標本一意の対象レコードについて，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**をレコードごとに計測し，スコアで表す。

←特殊な一意(special uniques)を考慮

(3) スワッピングの適用

1)ターゲット・スワッピング(targeted data swapping)

⇒対象となるレコードの中で，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。

2)ランダム・スワッピング(random data swapping)

⇒対象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。

参考　スワッピングの方法について

本実験では、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。

＊標本一意に該当した回数が1回以上のレコードがスワッピングの候補となるレコード群である。

(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。

＊スワッピング率pについては1,2,3,5,10,20,30%を適用した。

⇒本実験では，対象レコードに対して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)から作成したドナーファイル(約5,000レコード)から探索する。

参考　本研究におけるリンケージの方法(Domingo-Ferrer and Torra (2001),Takemura(1999))

(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。
(2)上記の11変数に関する値を合計して距離を計測する。
＊距離の計測式は以下のとおり
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。

外部情報とのマッチングに関する先行研究

・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((Müller *et al.*(1995)))←ドイツにおける事実上の匿名性(factual anonymity)に関する実証研究

・ 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))←人口センサスの2001年SARsの作成に関する実証研究

国調と住調のマッチングに使用したキー変数

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

＊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。

＊国調のデータにおいて、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う。←釣り検索の適用

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%

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%

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%

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%

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%

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%

4. 国勢調査ミクロデータにおける有用性の検証

スワッピング済データ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平均絶対距離の計測等を用いることが考えられるが、本研究ではサンプリング誤差に着目し、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。

＊本分析では、10%抽出のサンプリングデータを使用する。

参考　サンプリングの誤差(1%):年齢

	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

サンプリングの誤差(1%):世帯人員

	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	22.3	27.7	4
2人	25.3	22.6	28.0	2
3人	18.4	16.0	20.8	6
4人	15.4	13.2	17.7	2
5人	8.0	6.3	9.7	2
6人	4.8	3.5	6.1	3
7人	2.3	1.3	3.2	4
8人以上	0.8	0.2	1.3	4

参考　サンプリングの誤差(5%):年齢

	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

参考　サンプリングの誤差(5%):世帯人員

	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

参考　サンプリングの誤差(10%):年齢

	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

参考　サンプリングの誤差(10%):世帯人員

	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

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。

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 最小値	95%信頼区間 最大値
		1%	2%	3%	5%	10%	20%	30%		
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		

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 最小値	95%信頼区間 最大値
		1%	2%	3%	5%	10%	20%	30%		
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		

スワッピング率が30%の場合を除けば，世帯人員におけるスワッピングの誤差についてはサンプリングの誤差の範囲に収まっている。

参考　スワッピングにおける有用性と秘匿性の評価について

有用性と秘匿性の評価指標をもとに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。

⇒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、キー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用いられている。

有用性の評価指標

絶対距離の平均値*による評価

⇒個票データとスワッピング済データの両方で集計表を作成した上で，セルごとの度数の差の絶対値に関する平均値を求める(Shlomo *et al.*(2010))。

$$DU = \frac{\sum_c \left| T^P(c) - T^O(c) \right|}{n_T}$$

$T^O(c)$：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

$T^P(c)$：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

＊情報量損失(Information loss)に関する指標

秘匿性の評価指標(Shlomo *et al.*(2010))

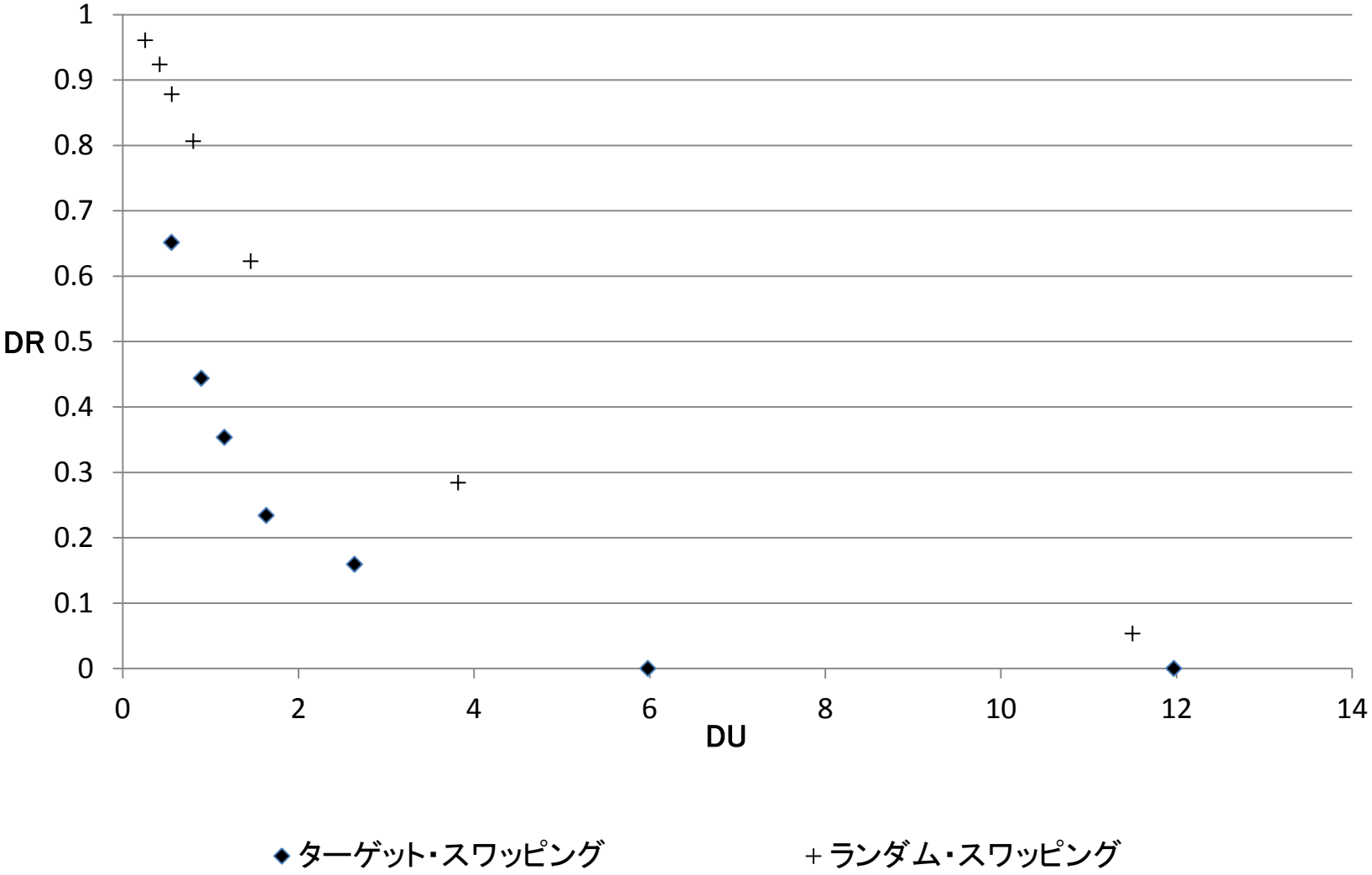
本研究では、リスク評価の尺度として以下の指標を用いる。

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

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

$\sum_c I\left(T^O(c)=1, T^P(c)=1\right)$ **秘匿処理済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**

R-Uマップの結果


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
DR
0
2
4
6
8
10
12
14
DU
ターゲット・スワッピング
ランダム・スワッピング

5. 終わりに

本報告では、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、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。

・本研究では、スワッピングの有効性を検証するために、スワッピングにおける誤差に着目した。本分析結果を見た限りでは、スワッピング済データと原データとの差は、概ねサンプリングの誤差の範囲であることが確認できた。

・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。

主要参考文献

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)「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』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.