匿名化された公的統計ミクロデータの作成における攪乱的手法の有効性の評価†

伊　藤　伸　介*
星　野　なおみ**
阿久津　文　香***
菊　池　　亮****

1．はじめに

近年，「統計改革」が注目されており，EBPM（=Evidence Based Policy Making, 客観的な事実に基づいて政策立案を行うこと）をわが国でも展開するために，公的統計のデータだけでなく，民間のビッグデータや行政記録データの利用可能性が指摘されている．2017年５月に刊行された『統計改革推進会議最終取りまとめ』では，EBPMの推進体制の構築のために，「統計等データ」[1)]の整備・改善が必要であることが明記された．こうした「統計改革」という形で，公的統計データの作成・提供への注目が高まっているだけでなく，2018年５月に国会で成立した「統計法及び独立行政法人統計センター法の一部を改正する法律（以下「改正統計法」と呼称）」との関連で，いわゆるpublic use fileを含む様々なタイプの公的統計ミクロデータの作成と提供の可能性も議論されている．

わが国では，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

†　本稿は，伊藤・星野・阿久津・菊池（2017）を加筆・修正したものである．共著者である星野なおみ氏（(独）統計センター），阿久津文香氏（総務省統計局），菊池亮氏（NTTセキュアプラットフォーム研究所）の了解をいただくだけでなく，（独）統計センターと総務省統計局より取りまとめの許可をいただいた上で，本稿を作成した．関係各位にお礼を申し上げたい．

1）「統計等データ」とは，統計，統計ミクロデータおよび統計的な利活用を行うために用いられる行政記録情報であって，それらのデータの利用・解釈を行うために必要な関連情報（メタデータ）を含む．

（調査票 A，調査票 B），労働力調査，国勢調査と国民生活基礎調査の７つの統計調査の匿名データが作成されている．例えば，国勢調査の匿名データの特徴としては，①提供年次は，平成12年と平成17年の２回分であること，②地域区分は，都道府県と人口50万以上市区であること，③データ量は母集団の１％であり，世帯単位で抽出されていること，④１種類のみの匿名データが作成・提供されていること，⑤リコーディング，トップコーディング，レコード削除といった非攪乱的手法（non-perturbative methods）だけでなく，スワッピングといった攪乱的手法（perturbative methods，パータベーション）が適用されていることが指摘できる．

公的統計の匿名データの作成に関する実務を勘案した場合，小地域分析のための匿名データに対するニーズが高いことなどを踏まえ，スペシャルユニーク（特殊な一意，special uniques，ex. 全国レベルの結果表のセルで度数１になるケース）といった，特定化のリスクが相対的に高いレコードを対象にした秘匿処理の方法が模索されている．こうしたことから，公的統計の匿名化ミクロデータ（匿名化技法が適用されたミクロデータ）における攪乱的手法の適用についても，さらなる検討の余地がある．

これまで，匿名化ミクロデータの作成に関する基礎研究として，例えば，全国消費実態調査，家計調査，さらには国勢調査の個票データを用いて，ミクロアグリゲーション，ノイズ，スワッピングといった攪乱的手法の適用可能性が追究されてきた（伊藤他(2014)，伊藤・星野(2014)，伊藤(2017)）．さらには，統計実務の観点から許容可能な閾値を設定した場合に，リコーディングやトップコーディングにおける区分統合の可能性が検討されてきた（伊藤(2018)）．これらの匿名化手法を適用することによって，様々なタイプの匿名化ミクロデータが作成される．そして，攪乱的手法が適用された匿名化ミクロデータについては，秘匿性と有用性の両面から定量的な比較・検討を行うことによって，利用者のニーズに応じた匿名化ミクロデータを作成することが可能になる．

本稿では，国勢調査を例に，匿名化された公的統計ミクロデータの作成の可能性を追究する．具体的には，本研究においては，スワッピング（data swapping）と PRAM（Post Randomization Methods）という２つの攪乱的手法に焦点を当てた上で，国勢調査の個票データを対象に，攪乱的手法の適用可能性を探ることにしたい．

２．公的統計ミクロデータにおけるスワッピング技法の有効性の評価

スワッピングという匿名化技法は，ミクロデータに含まれるレコードあるいは属性の組み合わせ同士で属性値群を入れ替える手法である（Willenborg and Waal（2001, p. 126））．スワッピングの可能性に関する議論については少なくとも1970年代に遡ることができる（Dalenius and Reiss (1978)）．図１は，スワッピングのイメージを示したものである．図１では，地域が異なるレコー

図1　スワッピングのイメージ

原データ

番号	地域	性別	雇用形態	週間就業時間
1	1	1	2	1
2	1	2	1	2
3	1	1	1	4
4	1	1	3	1
5	1	1	2	3
6	1	1	3	2
7	2	2	1	2
8	2	1	1	4
9	2	2	2	3

入れ替え

匿名化ミクロデータ

番号	地域	性別	雇用形態	週間就業時間
1	1	1	2	1
2	2	2	1	2
3	1	1	1	4
4	1	1	3	1
5	1	1	2	3
6	1	1	3	2
7	1	2	1	2
8	1	1	1	4
9	2	2	2	3

性別　1：男　2：女
地域　1：三大都市圏　2: それ以外
雇用形態　1：正規の職員・従業員　2：パート・アルバイト　3：派遣・契約社員
週間就業時間　1：35時間未満　2：35～48時間　3：49～59時間　4：60時間以上
原データ：秘匿処理を施していない個票データ
匿名化ミクロデータ：原データに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

出所）伊藤・星野（2014）

ド間の入れ替えを行っている．このことから，スワッピング後の匿名化ミクロデータにおいて作成された性別，雇用形態，週間就業時間別のクロス表は，スワッピングを行う前の原データ（秘匿処理が適用される前の個票データ）におけるクロス表の数値と変わらないことが確認できる（伊藤・星野(2014)）．

スワッピングは，特殊な一意に該当する露見リスクの高いレコードを対象に，特定のスワッピング率にしたがって適用される．スワッピングを適用する上で，基本的な属性間の関係性は変わらないことが要件だと言える．スワッピングは，露見リスクが相対的に高いレコードに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）と，スワッピングの対象となるレコードを無作為に選んだ上で，スワッピングを適用するランダム・スワッピング（random data swapping）に類別することが可能である（Shlomo *et al.* (2010)）．

諸外国におけるスワッピングの適用例として，つぎの事例が指摘できる．アメリカセンサス局は，2000年人口センサスのPUMS（Public Use Microdata Samples）やAmerican Community Surveyにおいて，スワッピングを適用している．スペシャルユニークの対象となるレコードを探索した上で，別のレコードに置き換えるという処理を行っている．具体的には，非常に粗い地域区分であっても，特定の人口社会的属性の組み合わせで一意となる世帯のレコードについては，露見リスクが非常に高いと考えられることから，別の地域における他の世帯との入れ替えが行われている（Zayatz（2007, p. 257））．また，イギリスでも，人口センサスの集計結果表を作成する前の個票データにおいて，レコードスワッピングが適用されている（Shlomo(2007)）．なお，アメリカにおいても，2000年人口センサスの集計表に対して秘匿処理を行うために，人口センサスの個

票データにスワッピングを適用していることが知られている．

スワッピング技法を適用する場合，２つの地域間でレコードの入れ替えを行うことが想定されるが，スワッピングにおける秘匿性および有用性を勘案した場合，n 地域間でスワッピングを行うことも考えられる．近年では，アメリカセンサス局において，n- サイクル スワッピング（n-cycle swapping）と呼ばれる複数地域間のスワッピングの考え方も提唱されている（Depersio *et al.* (2012)）．そこで，本研究では，複数地域間のスワッピングの可能性を追究している．

本研究では，平成22年国勢調査の調査票情報（個票データ）を使用する．本研究においては，A 県を対象に，平成12年，17年のすでに提供されている国勢調査の匿名データにおける秘匿性のレベルを超えない形で，様々なリコーディング（区分統合）を適用してテストデータを作成した．本研究ではテストデータを作成するためのキー変数の選出にあたって，母集団一意（population unique）の比率が計測された．また，地域を含む変数の原区分において，匿名データと同じ区分統合を行った場合の母集団一意の減少率も参照しながら，キー変数の区分が設定された．こうした実証的な研究を踏まえ，A 県における地域５区分を対象に，「年齢５歳階級90歳以上トップコーディング＆産業16区分＆職業７区分」の中で最も母集団一意が低いキー変数の組み合わせが選出された．なお，本研究における地域５区分とは，（１）地域 A（人口は50万人以上の地域），（２）地域 B（人口は50万人以上の地域），（３）地域 C（人口は20万人以上の地域），（４）地域 D（人口は20万人以上の地域），および（５）その他（地域 A，地域 B，地域 C，地域 D を除いた A 県における地域）である．なお，地域 A，地域 B，地域 C と地域 D の順に，人口規模は小さくなっている．

本研究において使用するキー変数および変数の分類区分はつぎのとおりである．

①建物の建て方（３区分）
②住居の種類（５区分）
③性別（２区分）
④配偶者の有無（４区分）
⑤国籍（２区分）
⑥労働力状態（７区分）
⑦従業上の地位（５区分）
⑧年齢（19区分）
⑨産業大分類（16区分）
⑩職業大分類（７区分）

本研究では，地域に関して，人口50人以上，あるいは人口20万人以上の人口規模に該当する，地域 A，地域 B，地域 C と地域 D の 4 地域が選ばれた．

本研究におけるスワッピングの方法は，伊藤・星野(2014)を参考にしながら，4地域（地域A，地域B，地域C，地域D）を対象として，以下の手順でスワッピングを行った．

① スワッピングを適用するにあたって，年齢，産業と職業についてグループ化を行う．本研究では，異なる年齢層，産業区分や職業区分の入れ替えに伴う情報量損失を小さくすることを指向していることから，具体的には，年齢については10歳区分，産業については3区分（第1次産業，第2次産業，第3次産業），職業については2区分（ブルーカラーとホワイトカラー）でグループ化を行った．なお，ホワイトカラーについては，専門的・技術的職業従事者，管理的職業従事者と事務従事者が該当するものとし，ブルーカラーについてはそれら以外の分類区分が該当するとする．

② スワッピングの対象レコードの中で優先順位が高いレコードを探索する．それは，つぎのように行われる．10個のキー変数から選ばれた任意の3変数のクロス表においてスペシャルユニークに該当する回数が多いレコードを，スワッピングの対象とする．具体的には，ある特定のレコードがスペシャルユニークに該当した回数をレコードごとに計測し，スペシャルユニークの回数を点数で表す（例えば，ある特定のレコードが10個の3変数の組み合わせにおけるクロス表でスペシャルユニークに該当したのであれば，10点とする等）．点数が高いほど，スペシャルユニークの中でもリスクが高いレコードと考えることができることから，スワッピングの優先順位が高くなる．なお，年齢，産業と職業の少なくともいずれかが含まれるキー変数の組み合わせの優先度が高くなっている．

③ ②で求めたスペシャルユニークに該当する点数に基づいて，地域ごとに，スペシャルユニークの点数が高いレコードから順番に，一定のスワッピング率にしたがってスワッピングを実行する（ターゲット・スワッピング）．本研究では，スワッピング率を0.1％とした．また，スワッピングにおいては，地域A，地域B，地域C，および地域Dの順番で行った．具体的には以下のとおりである．最初に，地域Aのスペシャルユニークを地域Bから探して入れ替える（ドナーファイルにおいて1対1で対応するレコードと入れ替えを行う）．入れ替えられたレコードをスワッピングのドナーのレコードとはしない．それを繰り返して，p％（本研究ではp=0.1％）のスワッピング率に達するまでスワッピングを行う．

④ 年齢，産業と職業で層化されたレコード群を対象に，ドナーファイルから層内で同一の属性値の組を有するレコードを探索した上で入れ替えを行う．層内において1対1で対応するレコードが存在しない場合には，ドナーファイルにおいて同一の属性値の組を有するn個のレコードの中から，ランダムに選定した上で入れ替えを行う．さらに，ドナーファイルにおいて同一の属性値の組を有するレコードがない場合には，分類区分数の逆数をウェイトとした上で，距離の計算を行い，年齢，産業と職業のグループ内で距離が一番近いレコードと入れ替える．

つぎに，本研究における距離の計算方法は，以下のとおりである（伊藤・星野(2014)）[2]．

（1）10個のキー変数について，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属性値が一致するかどうか検討する．具体的には，本研究では，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，分類区分の区分数で割る（不詳については除外している）．なお，年齢，産業，職業については，ドナーの層に含まれる区分数で除する．

（2）上記の10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，以下のように示される．

$$
\begin{aligned}
\text{距離} = {} & 〔\text{建物の建て方の分類区分数の逆数}〕 \times 〔\text{建物の建て方のスコア}〕 \\
& + 〔\text{住居の種類の分類区分数の逆数}〕 \times 〔\text{住居の種類のスコア}〕 \\
& + 〔\text{性別のスコア}〕 \\
& \cdots \\
& + 〔\text{職業の分類区分数の逆数}〕 \times 〔\text{職業のスコア}〕 \qquad (1)
\end{aligned}
$$

なお，同じ距離のレコードが複数ある場合には，ランダムに1つのレコードを選択している．

他方，本研究では，スワッピングの有効性に関する定量的な評価に関して，つぎの2つの方法を行った．

第1に，スワッピングが適用されたミクロデータの秘匿性に関しては，3変数の組み合わせのそれぞれにおいて探索されたスペシャルユニーク（3変数のクロス表における度数1）の中で，どの程度スワッピングが適用されているかについて検証を行った．それによって，どの変数に対してスワッピングが効果的に適用されているのかを確認した．

第2に，個票データと攪乱的手法が適用されたデータ（以下，「攪乱済ミクロデータ」と呼称．スワッピングが適用された場合には，「スワッピング済データ（スワッピングが適用されたデータ）」と呼称する）との間の絶対距離の平均値（average absolute distance）を計測することによって，情報量損失の計測を行った．

第2の情報量損失に基づく有用性の検証であるが，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，クラメールの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（伊藤・星野(2014)）が考えられる．また，伊藤(2017)においては，エントロピーに基づいて情報量損失の指標を作成した上で，秘匿の観点から「許容可能」な分類区分の組み合わせについて情報量損失の計測が行われた．そこで，本研究において

2） スワッピングの対象となるレコードのドナーファイルとの入れ替えに関する数理的な定式化については，伊藤・星野（2014）を参照．

も，攪乱済ミクロデータを行った場合の距離を計測することによって，情報量損失の計算を行った．具体的には，（2）式のように，原データと攪乱済ミクロ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計測した（Shlomo *et al.* (2010)）．

$$IL = \frac{\sum_{c} \left| T^{p}(c) - T^{o}(c) \right|}{n_T} \qquad (2)$$

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）：攪乱済ミクロデータをもとに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

表1-1～表1-4はそれぞれ，地域A～地域Dにおけるスワッピングの結果を示したものである．地域A，地域B，地域Cと地域Dにおけるスペシャルユニークとなるレコードとスワッピングされたレコード数との関係はそれぞれ，表2-1～表2-4で示されている．例えば，表1-1を見ると，地域Aにおいて，スワッピング対象となった278レコードの中で，地域Bのドナーファイルと置き換えられたのは195レコードであって，残りの83レコードについては地域Cと地域Dから置き換えられている．また，表1-2を見ると，地域Bでは，スワッピングの対象となった257レコードの中で，地域Cと入れ替えられたのは，129レコードのみであって，地域Dからは96レコードが入れ替えられている．さらに，地域Aのファイルですでにスワッピングのために用いられたレコードを除いたレコード群をドナーファイルとした場合，32レコードがスワッピングに用いられたことが確認できる．すなわち，複数の地域で入れ替えを行うことによって，1つの地域のみをドナーファイルとするよりも，よりスワッピングの対象レコードに近いドナーファイルのレコードと入れ替えを行うことが可能になっている．このことは，例えば，地域Aと地域Bの2地域間で入れ替えるよりも，複数の地域間で入れ替えを行ったほうがより効果的なスワッピングを行うことができることを示唆している．

つぎに，表2-1～表2-4は，120パターンの3変数の組み合わせにおいて，スペシャルユニークとなっているレコードの中のどの程度がスワッピングの対象となったかを示している．それによれば，組み合わせによって，スペシャルユニークの対象となっているレコードの中でスワッピングが施されたレコードの比率が異なることがわかる．例えば，地域Aにおいて，住居の建て方，従業上の地位，産業の3変数については，スペシャルユニークである18レコードの中の16レコードにスワッピングが施されており，スワッピングされたレコードの比率は88.9％となっている．一方，年齢，産業と職業の3変数については，スペシャルユニークである198レコードの中の66レコードにスワッピングが施されており，その比率は33.3％にすぎない．これについては，年齢，産業と職業の3変数でスペシャルユニークに該当したレコードは，他の変数の組み合わせではスペシャ

表1-1　地域Aにおけるスワッピングの結果

地域A	レコード数	10変数での母集団一意		スワッピング対象数			
		レコード数	比率		うち地域Bから	うち地域Cから	うち地域Dから
	277,665	18,191	6.55%	278	195	40	43

表1-2　地域Bにおけるスワッピングの結果

地域B	レコード数	10変数での母集団一意		スワッピング対象数			
		レコード数	比率		うち地域Cから	うち地域Dから	地域Aのスワッピングでドナーとして入れ替え
	257,451	17,708	6.88%	257	129	96	32

表1-3　地域Cにおけるスワッピングの結果

地域C	レコード数	10変数での母集団一意		スワッピング対象数			
		レコード数	比率		うち地域Dから	地域Aのスワッピングでドナーとして入れ替え	地域Bのスワッピングでドナーとして入れ替え
	85,640	8,768	10.24%	86	64	6	16

表1-4　地域Dにおけるスワッピングの結果

地域D	レコード数	10変数での母集団一意		スワッピング対象数				
		レコード数	比率		うち地域Aから	地域Aのスワッピングでドナーとして入れ替え	地域Bのスワッピングでドナーとして入れ替え	地域Cのスワッピングでドナーとして入れ替え
	76,442	10,011	13.10%	76	50	9	13	4

ルユニークに該当していなかった場合が少なくないことから，スワッピングの対象レコードの中で優先順位が相対的に低くなっていることが推察される．このように，スペシャルユニークに対するスワッピングの効果にはばらつきがあり，その効果は限定的であることが明らかになった．

表3は，①年齢5歳区分で90歳以上トップコーディング，産業（16区分）と職業（7区分）の3変数，②年齢10歳区分で90歳以上トップコーディング，産業（16区分）と職業（7区分）の3変数，および③年齢5歳区分で90歳以上トップコーディング，産業（3区分）と職業（2区分）でクロス表を作成した上で，地域A～地域Dを対象に，スワッピング済データにおける情報量損失を計測したものである．産業と職業をそれぞれ3区分，2区分と粗くした③のケースにおける情報量損失が最も低くなっていることがわかる．一方で，スワッピング率が0.1%であることから，情報量損失にそれほど大きな違いは見られないことが確認できる．

表2-1　スペシャルユニークとスワッピングされたレコード数との関係，地域 A

建て方 3区分	住居種類 5区分	性別	配偶者	国籍 2区分	労働力 7区分	従業上 5区分	5歳 90トップ 19区分	産業 16区分	職業 7区分	母集団一意数 (a)	スワッピングされた数 (b)	(b)/(a)
*	*	*								0	0	—
*	*		*							0	0	—
*	*			*						1	0	0
*	*				*					2	0	0
*	*					*				1	0	0
*	*						*			10	4	0.40000
*	*							*		5	1	0.20000
*	*								*	2	0	0
*		*	*							0	0	—
*		*		*						0	0	—
*		*			*					0	0	—
*		*				*				1	1	1.00000
*		*					*			5	4	0.80000
*		*						*		3	2	0.66667
*		*							*	1	1	1.00000
*			*	*						0	0	—
*			*		*					4	3	0.75000
*			*			*				3	2	0.66667
*			*				*			11	6	0.54545
*			*					*		9	7	0.77778
*			*						*	3	3	1.00000
*				*	*					2	1	0.50000
*				*		*				4	2	0.50000
*				*			*			10	6	0.60000
*				*				*		5	3	0.60000
*				*					*	3	1	0.33333
*					*	*				2	1	0.50000
*					*		*			67	41	0.61194
*					*			*		13	9	0.69231
*					*				*	6	4	0.66667
*						*	*			30	21	0.70000
*						*		*		18	16	0.88889
*						*			*	3	3	1.00000
*						*		*		18	16	0.88889
*						*			*	3	3	1.00000
*								*	*	29	20	0.68966
	*	*	*							0	0	—
	*	*		*						0	0	—
	*	*			*					0	0	—
	*	*				*				0	0	—
	*	*					*			6	5	0.83333
	*	*						*		3	2	0.66667
	*	*							*	1	1	1.00000
	*		*	*						0	0	—
	*		*		*					5	4	0.80000
	*		*			*				4	2	0.50000
	*		*				*			20	8	0.40000
	*		*					*		14	10	0.71429
	*		*						*	3	3	1.00000
	*			*	*					2	1	0.50000
	*			*		*				3	1	0.33333
	*			*			*			13	7	0.53846
	*			*				*		5	4	0.80000
	*			*					*	3	2	0.66667
	*				*	*				3	2	0.66667
	*				*		*			74	43	0.58108
	*				*			*		17	11	0.64706
	*				*				*	6	5	0.83333
	*					*	*			41	25	0.60976
	*					*		*		25	17	0.68000

表2-1続き

建て方 3区分	住居 種類 5区分	性別	配偶 者	国籍 2区分	労働力 7区分	従業上 5区分	5歳 90トップ 19区分	産業 16区分	職業 7区分	母集団 一意数 (a)	スワッピング された 数 (b)	(b)/(a)
	*					*			*	4	2	0.50000
	*						*	*		118	51	0.43220
	*						*		*	40	28	0.70000
	*							*	*	37	28	0.75676
		*	*	*						0	0	—
		*	*		*					1	1	1.00000
		*	*			*				0	0	—
		*	*				*			5	3	0.60000
		*	*					*		0	0	—
		*	*						*	0	0	—
		*		*	*					0	0	—
		*		*		*				1	1	1.00000
		*		*			*			4	3	0.75000
		*		*				*		1	1	1.00000
		*		*					*	0	0	—
		*			*	*				0	0	—
		*			*		*			19	13	0.68421
		*			*			*		1	1	1.00000
		*			*				*	0	0	—
		*				*	*			9	7	0.77778
		*				*		*		3	3	1.00000
		*				*			*	0	0	—
		*					*	*		29	25	0.86207
		*					*		*	8	5	0.62500
		*						*	*	6	3	0.50000
			*	*	*					3	1	0.33333
			*	*		*				2	1	0.50000
			*	*			*			5	3	0.60000
			*	*				*		8	3	0.37500
			*	*					*	2	1	0.50000
			*		*	*				1	1	1.00000
			*		*		*			42	24	0.57143
			*		*			*		3	2	0.66667
			*		*				*	2	1	0.50000
			*			*	*			22	14	0.63636
			*			*		*		12	8	0.66667
			*			*			*	1	0	0
			*				*	*		65	44	0.67692
			*				*		*	29	21	0.72414
			*					*	*	21	11	0.52381
				*	*	*				0	0	—
				*	*		*			17	10	0.58824
				*	*			*		8	3	0.37500
				*	*				*	2	0	0
				*		*	*			16	9	0.56250
				*		*		*		6	5	0.83333
				*		*			*	3	1	0.33333
				*			*	*		32	20	0.62500
				*			*		*	17	11	0.64706
				*				*	*	13	7	0.53846
					*	*	*			13	10	0.76923
					*	*		*		8	6	0.75000
					*	*			*	2	1	0.50000
					*		*	*		68	36	0.52941
					*		*		*	19	16	0.84211
					*			*	*	31	12	0.38710
						*	*	*		110	60	0.54545
						*	*		*	34	24	0.70588
						*		*	*	36	16	0.44444
							*	*	*	198	66	0.33333

表2-2　スペシャルユニークとスワッピングされたレコード数との関係，地域 B

建て方 3区分	住居種類 5区分	性別	配偶者	国籍 2区分	労働力 7区分	従業上 5区分	5歳 90トップ 19区分	産業 16区分	職業 7区分	母集団一意数 (a)	スワッピングされた数 (b)	(b)/(a)
*	*	*								0	0	—
*	*		*							0	0	—
*	*			*						1	0	0
*	*				*					2	0	0
*	*					*				2	2	1
*	*						*			12	5	0.41667
*	*							*		9	2	0.22222
*	*								*	4	1	0.25000
*		*	*							0	0	—
*		*		*						0	0	—
*		*			*					1	1	1
*		*				*				0	0	—
*		*					*			4	4	1
*		*						*		2	2	1
*		*							*	1	1	1
*			*	*						1	0	0
*			*		*					3	1	0.33333
*			*			*				3	3	1
*			*				*			16	10	0.62500
*			*					*		14	9	0.64286
*			*						*	2	1	0.50000
*				*	*					3	2	0.66667
*				*		*				1	1	1
*				*			*			8	3	0.37500
*				*				*		7	4	0.57143
*				*					*	3	2	0.66667
*					*	*				3	2	0.66667
*					*		*			61	33	0.54098
*					*			*		9	5	0.55556
*					*				*	3	3	1
*						*	*			25	20	0.80000
*						*		*		29	26	0.89655
*						*			*	6	6	1
*						*		*		29	26	0.89655
*						*			*	6	6	1
*								*	*	36	27	0.75000
	*	*	*							0	0	—
	*	*		*						0	0	—
	*	*			*					1	1	1
	*	*				*				0	0	—
	*	*					*			8	6	0.75000
	*	*						*		2	2	1
	*	*							*	1	1	1
	*		*	*						1	0	0
	*		*		*					4	1	0.25000
	*		*			*				5	4	0.80000
	*		*				*			23	15	0.65217
	*		*					*		12	8	0.66667
	*		*						*	2	1	0.50000
	*			*	*					2	1	0.50000
	*			*		*				1	1	1
	*			*			*			11	7	0.63636
	*			*				*		11	5	0.45455
	*			*					*	5	3	0.60000
	*				*	*				4	2	0.50000
	*				*		*			87	42	0.48276
	*				*			*		10	6	0.60000
	*				*				*	2	2	1
	*					*	*			38	28	0.73684
	*					*		*		27	22	0.81481

表2-2続き

建て方3区分	住居種類5区分	性別	配偶者	国籍2区分	労働力7区分	従業上5区分	5歳90トップ19区分	産業16区分	職業7区分	母集団一意数(a)	スワッピングされた数(b)	(b)/(a)
	*					*			*	8	7	0.87500
	*						*	*		116	48	0.41379
	*						*		*	42	33	0.78571
	*							*	*	46	26	0.56522
		*	*	*						0	0	—
		*	*		*					0	0	—
		*	*			*				0	0	—
		*	*				*			7	6	0.85714
		*	*					*		0	0	—
		*	*						*	0	0	—
		*		*	*					0	0	—
		*		*		*				0	0	—
		*		*			*			5	5	1
		*		*				*		0	0	—
		*		*					*	0	0	—
		*			*	*				0	0	—
		*			*		*			13	11	0.84615
		*			*			*		1	1	1
		*			*				*	0	0	—
		*				*	*			9	7	0.77778
		*				*		*		3	3	1
		*				*			*	0	0	—
		*					*	*		23	19	0.82609
		*					*		*	5	4	0.80000
		*						*	*	6	5	0.83333
			*	*	*					1	1	1
			*	*		*				1	1	1
			*	*			*			9	7	0.77778
			*	*				*		8	4	0.50000
			*	*					*	3	1	0.33333
			*		*	*				0	0	—
			*		*		*			54	26	0.48148
			*		*			*		2	1	0.50000
			*		*				*	0	0	—
			*			*	*			20	14	0.70000
			*			*		*		9	7	0.77778
			*			*			*	2	1	0.50000
			*				*	*		67	36	0.53731
			*				*		*	24	20	0.83333
			*					*	*	19	11	0.57895
				*	*	*				2	0	0
				*	*		*			17	7	0.41176
				*	*			*		6	2	0.33333
				*	*				*	1	1	1
				*		*	*			8	7	0.87500
				*		*		*		6	3	0.50000
				*		*			*	3	2	0.66667
				*			*	*		40	21	0.52500
				*			*		*	9	9	1
				*				*	*	21	7	0.33333
					*	*	*			12	9	0.75000
					*	*		*		8	4	0.50000
					*	*			*	0	0	—
					*		*	*		74	32	0.43243
					*		*		*	15	10	0.66667
					*			*	*	26	13	0.50000
						*	*	*		107	57	0.53271
						*	*		*	29	17	0.58621
						*		*	*	40	19	0.47500
							*	*	*	187	56	0.29947

表2-3　スペシャルユニークとスワッピングされたレコード数との関係，地域 C

建て方 3区分	住居種類 5区分	性別	配偶者	国籍 2区分	労働力 7区分	従業上 5区分	5歳 90トップ 19区分	産業 16区分	職業 7区分	母集団一意数 (a)	スワッピングされた数 (b)	(b)/(a)
*	*	*								1	0	0
*	*		*							0	0	—
*	*			*						0	0	—
*	*				*					10	3	0.300
*	*					*				5	1	0.200
*	*						*			16	6	0.375
*	*							*		8	1	0.125
*	*								*	3	0	0
*		*	*							1	1	1.000
*		*		*						0	0	—
*		*			*					3	3	1.000
*		*				*				1	1	1.000
*		*					*			4	4	1.000
*		*						*		8	6	0.750
*		*							*	1	0	0
*			*	*						1	0	0
*			*		*					11	9	0.818
*			*			*				5	3	0.600
*			*				*			22	12	0.545
*			*					*		16	4	0.250
*			*						*	9	3	0.333
*				*	*					3	2	0.667
*				*		*				2	0	0
*				*			*			11	5	0.455
*				*				*		10	5	0.500
*				*					*	4	0	0
*					*	*				4	3	0.750
*					*		*			59	21	0.356
*					*			*		12	6	0.500
*					*				*	6	3	0.500
*						*	*			38	22	0.579
*						*		*		27	18	0.667
*						*			*	7	5	0.714
*						*		*		27	18	0.667
*						*			*	7	5	0.714
*								*	*	38	16	0.421
	*	*	*							1	1	1.000
	*	*		*						0	0	—
	*	*			*					2	2	1.000
	*	*				*				3	1	0.333
	*	*					*			8	6	0.750
	*	*						*		6	3	0.500
	*	*							*	1	0	0
	*		*	*						0	0	—
	*		*		*					15	8	0.533
	*		*			*				5	3	0.600
	*		*				*			27	14	0.519
	*		*					*		22	3	0.136
	*		*						*	8	3	0.375
	*			*	*					7	2	0.286
	*			*		*				0	0	—
	*			*			*			15	6	0.400
	*			*				*		10	4	0.400
	*			*					*	4	0	0
	*				*	*				2	1	0.500
	*				*		*			82	29	0.354
	*				*			*		18	5	0.278
	*				*				*	6	2	0.333
	*					*	*			47	22	0.468
	*					*		*		36	15	0.417

表2-3続き

建て方 3区分	住居 種類 5区分	性別	配偶 者	国籍 2区分	労働力 7区分	従業上 5区分	5歳 90トップ 19区分	産業 16区分	職業 7区分	母集団 一意数 (a)	スワッピング された 数 (b)	(b)/(a)
	*					*			*	12	6	0.500
	*						*	*		183	55	0.301
	*						*		*	58	28	0.483
	*							*	*	59	19	0.322
		*	*	*						0	0	—
		*	*		*					1	1	1.000
		*	*			*				0	0	—
		*	*				*			8	3	0.375
		*	*					*		1	0	0
		*	*						*	0	0	—
		*		*	*					0	0	—
		*		*		*				1	0	0
		*		*			*			3	2	0.667
		*		*				*		7	6	0.857
		*		*					*	0	0	—
		*			*	*				0	0	—
		*			*		*			13	4	0.308
		*			*			*		6	1	0.167
		*			*				*	0	0	—
		*				*	*			12	10	0.833
		*				*		*		7	3	0.429
		*				*			*	1	0	0
		*					*	*		43	22	0.512
		*					*		*	12	9	0.750
		*						*	*	19	8	0.421
			*	*	*					2	1	0.500
			*	*		*				0	0	—
			*	*			*			12	2	0.167
			*	*				*		12	6	0.500
			*	*					*	2	0	0
			*		*	*				2	2	1.000
			*		*		*			56	13	0.232
			*		*			*		12	3	0.250
			*		*				*	3	1	0.333
			*			*	*			35	15	0.429
			*			*		*		20	7	0.350
			*			*			*	3	2	0.667
			*				*	*		95	28	0.295
			*				*		*	30	10	0.333
			*					*	*	43	5	0.116
				*	*	*				0	0	—
				*	*		*			19	6	0.316
				*	*			*		9	3	0.333
				*	*				*	4	0	0
				*		*	*			13	6	0.462
				*		*		*		13	7	0.538
				*		*			*	2	0	0
				*			*	*		52	22	0.423
				*			*		*	17	8	0.471
				*				*	*	17	10	0.588
					*	*	*			17	8	0.471
					*	*		*		18	6	0.333
					*	*			*	5	1	0.200
					*		*	*		124	34	0.274
					*		*		*	34	18	0.529
					*			*	*	45	7	0.156
						*	*	*		155	37	0.239
						*	*		*	57	21	0.368
						*		*	*	51	10	0.196
							*	*	*	231	38	0.165

表2-4　スペシャルユニークとスワッピングされたレコード数との関係，地域 D

建て方 3区分	住居 種類 5区分	性別	配偶 者	国籍 2区分	労働力 7区分	従業上 5区分	5歳 90トップ 19区分	産業 16区分	職業 7区分	母集団 一意数 (a)	スワッピング された 数 (b)	(b)/(a)
*	*	*								0	0	—
*	*		*							1	0	0
*	*			*						1	0	0
*	*				*					6	2	0.333
*	*					*				8	1	0.125
*	*						*			24	7	0.292
*	*							*		9	0	0
*	*								*	3	0	0
*		*	*							0	0	—
*		*		*						0	0	—
*		*			*					1	1	1.000
*		*				*				1	1	1.000
*		*					*			4	3	0.750
*		*						*		7	4	0.571
*		*							*	3	1	0.333
*			*	*						0	0	—
*			*		*					12	5	0.417
*			*			*				3	3	1.000
*			*				*			17	10	0.588
*			*					*		14	2	0.143
*			*						*	3	1	0.333
*				*	*					4	1	0.250
*				*		*				2	0	0
*				*			*			10	6	0.600
*				*				*		8	4	0.500
*				*					*	1	1	1.000
*					*	*				5	2	0.400
*					*		*			61	30	0.492
*					*			*		16	7	0.438
*					*				*	9	3	0.333
*						*	*			36	22	0.611
*						*		*		28	16	0.571
*						*			*	10	7	0.700
*						*		*		28	16	0.571
*						*			*	10	7	0.700
*								*	*	41	9	0.220
	*	*	*							0	0	—
	*	*		*						0	0	—
	*	*			*					1	1	1.000
	*	*				*				2	2	1.000
	*	*					*			9	6	0.667
	*	*						*		8	4	0.500
	*	*							*	2	1	0.500
	*		*	*						0	0	—
	*		*		*					13	5	0.385
	*		*			*				6	3	0.500
	*		*				*			28	12	0.429
	*		*					*		16	3	0.188
	*		*						*	2	1	0.500
	*			*	*					9	1	0.111
	*			*		*				3	0	0
	*			*			*			10	7	0.700
	*			*				*		18	4	0.222
	*			*					*	7	1	0.143
	*				*	*				4	3	0.750
	*				*		*			83	34	0.410
	*				*			*		17	7	0.412
	*				*				*	9	3	0.333
	*					*	*			40	22	0.550
	*					*		*		36	18	0.500

表2-4続き

建て方 3区分	住居 種類 5区分	性別	配偶 者	国籍 2区分	労働力 7区分	従業上 5区分	5歳 90トップ 19区分	産業 16区分	職業 7区分	母集団 一意数 (a)	スワッピング された 数 (b)	(b)/(a)
	*					*			*	14	6	0.429
	*						*	*		159	49	0.308
	*						*		*	47	22	0.468
	*							*	*	45	9	0.200
		*	*	*						0	0	—
		*	*		*					1	1	1.000
		*	*			*				0	0	—
		*	*				*			5	4	0.800
		*	*					*		0	0	—
		*	*						*	0	0	—
		*		*	*					1	0	0
		*		*		*				0	0	—
		*		*			*			2	2	1.000
		*		*				*		3	2	0.667
		*		*					*	0	0	—
		*			*	*				0	0	—
		*			*		*			16	9	0.563
		*			*			*		2	1	0.500
		*			*				*	0	0	—
		*				*	*			11	3	0.273
		*				*		*		5	3	0.600
		*				*			*	1	0	0
		*					*	*		40	27	0.675
		*					*		*	12	11	0.917
		*						*	*	8	3	0.375
			*	*	*					4	1	0.250
			*	*		*				1	0	0
			*	*			*			9	4	0.444
			*	*				*		15	3	0.200
			*	*					*	2	0	0
			*		*	*				1	1	1.000
			*		*		*			57	17	0.298
			*		*			*		9	2	0.222
			*		*				*	1	1	1.000
			*			*	*			28	11	0.393
			*			*		*		14	9	0.643
			*			*			*	3	1	0.333
			*				*	*		94	29	0.309
			*				*		*	34	17	0.500
			*					*	*	52	10	0.192
				*	*	*				2	1	0.500
				*	*		*			31	10	0.323
				*	*			*		8	4	0.500
				*	*				*	4	1	0.250
				*		*	*			17	6	0.353
				*		*		*		15	7	0.467
				*		*			*	4	0	0
				*			*	*		60	22	0.367
				*			*		*	21	6	0.286
				*				*	*	18	8	0.444
					*	*	*			19	9	0.474
					*	*		*		19	6	0.316
					*	*			*	2	0	0
					*		*	*		119	29	0.244
					*		*		*	41	19	0.463
					*			*	*	40	6	0.150
						*	*	*		124	35	0.282
						*	*		*	53	19	0.358
						*		*	*	48	8	0.167
							*	*	*	216	42	0.194

表 3　スワッピングにおける情報量損失の結果

① 年齢 5 歳区分で90歳以上トップコーディング，産業（16区分），職業（ 7 区分）

	地域 A	地域 B	地域 C	地域 D
差	248	400	260	222
差 / セル数	0.12	0.19	0.12	0.10

注）3 変数のクロス表のセル数は2,147

② 年齢10歳区分で90歳以上トップコーディング，産業（16区分），職業（ 7 区分）

	地域 A	地域 B	地域 C	地域 D
差	222	338	228	208
差 / セル数	0.20	0.30	0.20	0.18

注）3 変数のクロス表のセル数は1,130

③ 年齢 5 歳区分で90歳以上トップコーディング，産業（ 3 区分），職業（ 2 区分）

	地域 A	地域 B	地域 C	地域 D
差	10	22	12	8
差 / セル数	0.08	0.17	0.09	0.06

注）3 変数のクロス表のセル数は133

3．公的統計ミクロデータに対する PRAM の適用可能性の検証

カテゴリカルデータに対する攪乱的手法には，スワッピング技法の他にも Post Randomization Method（PRAM）と呼ばれる，ノイズ付加等の確率的処理を施してプライバシーを保護する方法が知られている（Kooiman（1998））．PRAM はオランダの公的統計（de Wolf *et al.*（1998), de Wolf and van Gelder（2004））や，データベース（Agrawal and Srikant（2000), Agrawal *et al.*（2005））の分野においていくつかの適用可能性に関する研究がなされてきている．

PRAM は「攪乱」と「再構築」と呼ばれる 2 つのステップから構成される．PRAM における「攪乱」とは，個票データの各セルの値をあらかじめ決められた確率に基づいて遷移させるステップであり，再構築とは，攪乱された個票データから原データが持つ分布を推定するステップである．

PRAM における「攪乱」をより詳細に説明するため，幾つかの定義を導入する．個票データの各属性を V とし，その取り得る属性値を $\{v_i\}_{0 \leq i < M}$ とする．また，ある属性 V に対する遷移確率行列 A を

$$\boldsymbol{A} := \begin{pmatrix} a_{0,\,0} & \cdots & a_{0,\,M-1} \\ \vdots & \ddots & \vdots \\ a_{M-1,\,0} & \cdots & a_{M-1,\,M-1} \end{pmatrix}$$

と定義する．ここで，$a_{i,j}$ は v_i が v_j に遷移する確率であり，匿名化処理を Δ としたとき $a_{i,j} = \Pr[v_j = \Delta(v_i)]$ と書ける．

PRAM においては，「攪乱」によって，各属性値が遷移確率行列 $\boldsymbol{A}$ にしたがって遷移する．例えば，性別という属性で，属性値が“男性”・“女性”の 2 つのみであり，遷移確率行列が

$$\boldsymbol{A} := \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}$$

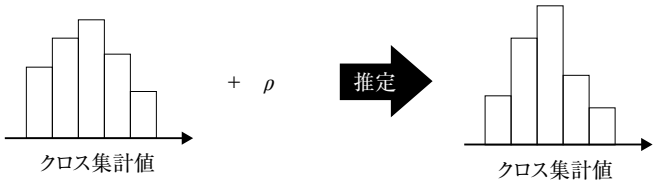
である場合，元々属性値が“男性”であったセルは，75％の確率でそのまま維持され，25％の確率で“女性”に遷移される．

PRAM における「攪乱」の効果は，直感的には以下のような例から得られる．攻撃者は，個票データの中から“男性かつ地域A”に住む，ある人物のデータを特定しよう試みているとする．このとき，「攪乱」によって“男性かつ地域A”が“女性かつ地域A”に遷移した場合，値が変わっているため攻撃者にとってその人物を特定することが難しくなっている．さらに，もし“男性かつ地域A”が遷移せずにそのまま残ったとしても，攻撃者から見れば“男性かつ地域A”が遷移せずにそのまま残っているのか，もしくは“女性かつ地域B”という人物のデータがたまたま“男性かつ地域A”に遷移したのか，どちらなのかを判別することは難しく，結果として個票データの中からある人物を特定することは難しくなっている．このようなプライバシー保護の効果を定量化した結果として，PRAM は，高々確率 *1/k* でしか個人を特定できないことを保証する Pk− 匿名性（Ikarashi *et al.*（2014））や，ϵ −差分プライバシー（Dwork（2006））を満たせることが知られている（Lin *et al.*（2012），Ikarashi *et al.*（2014））．

PRAM のもう 1 つのステップである「再構築」とは，元の個票データの統計的な分布を推定し，精度を向上させる処理である．一般に PRAM では，「攪乱」に用いた遷移確率行列は，攪乱済ミクロデータと共に公開される．これらの情報を用いると，原データのクロス集計値などを推定することができる．例えば，先ほどの“男性”・“女性”のみの例を考えると，攪乱済ミクロデータの男性と女性の人数の比率は，原データの男性と女性の人数の比率に比べ 1：1 に近づいている．そのため，例えば攪乱済ミクロデータで男性60人，女性40人であったならば，原データでは男性70人，女性30人のように，より大きな差があったと考えられる．このような推定の方法が「再構築」である．そのイメージを図 2 に示す．

再構築を行う具体的なアルゴリズムとしては，クロス集計値を再構築する逐次ベイズ法（Kooiman *et al.*（1998），Agrawal（2000））がよく知られている．また，近年では再構築によって得られた

図 2　再構築のイメージ



結果が原データとどの程度離れているかといった精度を保証できるような手法も研究されている（長谷川(2016)）.

本研究では，国勢調査の個票データに基づく匿名化ミクロデータにおいて，年齢，産業と職業のそれぞれに対して PRAM による攪乱を行い，その結果について考察する.

すでに述べた通り，PRAM では，ある個人の属性値 v_i が異なる属性値 v_j に変化し得ることに加えて，別の個人の属性値 v_k がたまたま v_i に変化し得ることも，プライバシーを保護する上で重要な役割を果たしている．そのため PRAM では，ある属性値 v_i は，攪乱によってその属性の取り得る値 $\{v_i\}_{0\leq i<Mi}$ の全てに遷移し得るのが一般的である．言い換えると，遷移確率行列の各要素について $a_{i,j}\neq 0$ である.

しかし，利用者の観点からは，そのような PRAM での攪乱が適切でない場合も考えられる．例えば産業といった属性を攪乱するにあたり，第一次産業である農業と，第三次産業である卸売・小売業といった異なる性質を持つものの間で値が遷移すると，攪乱後の分析手法によっては元の個票データを用いた場合と著しく異なる場合もある．また，そもそも第一次 / 第二次 / 第三次産業ごとに分析するといった用途では，産業区分を超えた攪乱ではなく，同一の産業区分内での攪乱に留めた方が，より分析精度が高くなることが期待できる.

そこで今回の実験では，属性全体を遷移する場合に加えて，属性値の幾つかをグループ化し，そのグループ内で遷移させることも行った．例えば産業の場合，16区分での攪乱に加え，図 3 のように，粗い 3 区分（第一次，第二次，第三次）にデータを分け，そのグループごとに攪乱を行った.

同様に職業は 7 区分と 2 区分の 2 パターン，年齢は 5 歳刻みと10歳刻みの 2 パターンで攪乱を行った.

攪乱方法には維持置換攪乱を用いた．これは，一定の確率 ρ で属性値を維持し，確率1−ρ で遷移候補の中からランダムな値に遷移するような攪乱であり，遷移確率行列の各要素は

$$a_{i,j}=\begin{cases}\rho+\frac{1-\rho}{|v|} & \text{if}\quad i=j\\ \frac{1-\rho}{|v|} & otherwise\end{cases}$$

で与えられる．ρ は0.95，0.9，0.85および0.8の 4 パターンを用いた．

本実験では地域 A，地域 B，地域 C と地域 D に対して PRAM を行い，精度の測定（差および差 / セル数）に関しては 3 回の実験の平均から求めた．これらを表4-1～表4-4に示している．維持確率が大きいほど精度が高いことが確認できる．

結果を見るとスワッピングに比べ精度が低いが，これはスワッピングが全レコードのうち0.1%をスワッピングしていることに対し，PRAM の実験では，維持確率が最も高い0.95であっても，全レコードのうち約15%のレコードのいずれかのセルがランダムな値に置き換わっているためと考えられる．スワッピングと PRAM の効果を正確に比較するためには，両手法の秘匿性・有用性を定量化する必要があるが，パータベーションの方法が大きく異なるため，定量化は簡単ではなく，今後の課題である．

また，本実験では再構築を行っていない．これは，既存の再構築においては「グループに分けて攪乱する」ことを想定していないことに由来している．詳細な説明は避けるが，既存の手法をそのまま用いた場合，再構築によって産業の区分が第一次産業から第二次産業に移るといったことが起こり得るだけでなく，大量の記憶領域を必要として実行が難しいなどといったことが発生している．そのため，再構築には既存と異なる新たな手法の考案が必要となる．

図 3　PRAM のイメージ

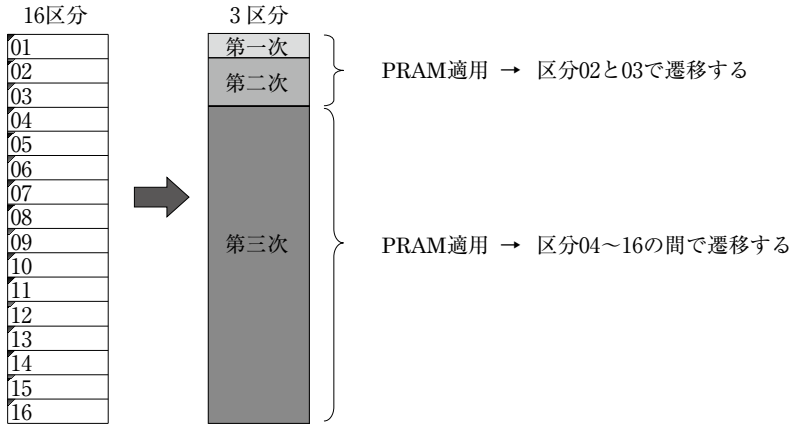


表4-1　PRAM における情報量損失の結果，地域 A

① 年齢 5 歳区分で90歳以上トップコーディング，産業（16区分），職業（ 7 区分）

維持確率	0.95	0.9	0.85	0.8
差	1808	2985	4019	4868
差 / セル数	0.84	1.39	1.87	2.27

② 年齢10歳区分で90歳以上トップコーディング，産業（16区分），職業（ 7 区分）

維持確率	0.95	0.9	0.85	0.8
差	1343	2453	3421	4263
差 / セル数	1.19	2.17	3.03	3.77

③ 年齢 5 歳区分で90歳以上トップコーディング，産業（ 3 区分），職業（ 2 区分）

維持確率	0.95	0.9	0.85	0.8
差	205	361	491	564
差 / セル数	1.539	2.712	3.694	4.241

表4-2　PRAM における情報量損失の結果，地域 B

① 年齢 5 歳区分で90歳以上トップコーディング，産業（16区分），職業（ 7 区分）

維持確率	0.95	0.9	0.85	0.8
差	1757	2811	3899	4859
差 / セル数	0.82	1.31	1.82	2.26

② 年齢10歳区分で90歳以上トップコーディング，産業（16区分），職業（ 7 区分）

維持確率	0.95	0.9	0.85	0.8
差	1357	2297	3297	4280
差 / セル数	1.20	2.03	2.92	3.79

③ 年齢 5 歳区分で90歳以上トップコーディング，産業（ 3 区分），職業（ 2 区分）

維持確率	0.95	0.9	0.85	0.8
差	216	339	542	544
差 / セル数	1.62	2.55	4.08	4.09

表4-3　PRAM における情報量損失の結果，地域 C

① 年齢 5 歳区分で90歳以上トップコーディング，産業（16区分），職業（ 7 区分）

維持確率	0.95	0.9	0.85	0.8
差	1028	1717	2261	2760
差 / セル数	0.48	0.80	1.05	1.29

② 年齢10歳区分で90歳以上トップコーディング，産業（16区分），職業（７区分）

維持確率	0.95	0.9	0.85	0.8
差	783	1373	1853	2351
差 / セル数	0.69	1.21	1.64	2.08

③ 年齢５歳区分で90歳以上トップコーディング，産業（３区分），職業（２区分）

維持確率	0.95	0.9	0.85	0.8
差	153	205	250	335
差 / セル数	1.15	1.54	1.88	2.52

表4-4 PRAM における情報量損失の結果，地域 D

① 年齢５歳区分で90歳以上トップコーディング，産業（16区分），職業（７区分）

維持確率	0.95	0.9	0.85	0.8
差	1115	1874	2485	3057
差 / セル数	0.52	0.87	1.16	1.42

② 年齢10歳区分で90歳以上トップコーディング，産業（16区分），職業（７区分）

維持確率	0.95	0.9	0.85	0.8
差	829	1482	2032	2599
差 / セル数	0.73	1.31	1.80	2.30

③ 年齢５歳区分で90歳以上トップコーディング，産業（３区分），職業（２区分）

維持確率	0.95	0.9	0.85	0.8
差	149	223	324	405
差 / セル数	1.12	1.67	2.44	3.05

４．むすびにかえて

本稿では，公的統計ミクロデータに対する攪乱的手法の適用可能性を追究するために，国勢調査を対象に，攪乱的手法の１つであるスワッピング技法と PRAM に焦点を当て，攪乱的手法が適用された匿名化ミクロデータの可能性を追究した．

本研究の結果を踏まえると，複数の地域間におけるスワッピングの場合，属性値がより近いレコードとの入れ替えを行うことが可能なことが明らかになった．また，年齢，産業と職業における情報量損失を小さくするために，年齢，産業と職業を層別した上で，スワッピング技法を適用することが可能なことが確認された．一方，PRAM についても年齢，産業と職業を層別した上で

の攪乱が可能であることが確認されたが，本研究では，再構築は適用されていないことから，今後さらなる研究・実験が必要になると思われる．

スワッピングや PRAM 等の攪乱的手法に関しては，「改正統計法」を踏まえた統計法制度の動向を見ながら，攪乱的手法が用いられた場合のミクロデータの特性を考慮した上で，統計実務における攪乱的手法の適用可能性を模索していく必要がある．また，近年の統計調査環境に伴い，公的統計データに不詳を含んだレコードが存在するが，そうした不詳を含むミクロデータに対する攪乱的手法の適用のあり方については，これからの課題になり得るものと考える．また，攪乱済ミクロデータの秘匿性と有用性を比較・評価するための定量的な評価方法についても，さらなる研究を進めていく必要があると思われる．これらについては今後の研究課題としたい．

参 考 文 献

伊藤伸介・村田磨理子・高野正博（2014）「ミクロデータにおける匿名化技法の有効性の検証―全国消費実態調査と家計調査を例に―」，『統計研究彙報』第71号，83-124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（2017）「国勢調査ミクロデータにおける匿名化の誤差の評価方法に関する一考察」，『経済学論纂（中央大学）』第57巻第 3 ・ 4 合併号，189-209頁

伊藤伸介・星野なおみ・阿久津文香・菊池亮（2017）「国勢調査の匿名化ミクロデータの作成方法に関する新たな取り組み」『製表技術参考資料』No. 37，1 -27頁

伊藤伸介（2018）「国勢調査における匿名化ミクロデータの作成可能性」『経済志林』，第85巻第 2 号，241-277頁

長谷川聡・正木彰伍・濱田浩気・菊池亮（2016）「確率的 k- 匿名化における再構築の正確度に関する理論的解析」『暗号と情報セキュリティシンポジウム』

Agrawal, R., R. Srikant (2000) *Privacy-Preserving Data Mining*, SIGMOD 2000.

Dalenius, T. and S. P. Reiss (1978) "Data-Swapping: A Technique for Disclosure Control (Extended Abstract)", in Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, D.C., pp. 191-194.

De Waal, T. and L. Willenborg (1999) "Information Loss Through Global Recoding and Local Suppression", *Netherlands Official Statistics* (*special issue on SDC*), Vol. 14, pp. 17-20.

De Wolf, P. P. and I. van Gelder (2004) An empirical evaluation of PRAM, Discussion paper 04012, Statistics Netherlands.

Depersio, M., M. Lemons, K. A. Ramanayake, J. Tsay, L. Zayatz (2012) "n-Cycle Swapping for the American Community Survey", J. Domingo-Ferrer and I. Tinnirello (eds.) *Privacy in Statistical Databases UNESCO Chair in Data Privacy International Conference, PSD 2012 Palermo, Italy, September, 2012 Proceedings, Springer*, pp. 143-164.

Domingo-Ferrer, J. and V. Torra (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.

Dwork, C. (2006) *Differential privacy*. ICALP.

Ikarashi, D., R. Kikuchi, K. Chida and K. Takahashi（2014）"k-anonymous Microdata Release via Post Randomisation Method", IWSEC 2014.

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

Lin, B. R., Y. Wang, and S. Rane（2012）*A framework for privacy preserving statistical analysis on distributed databases.* WIFS, 2012.

Shlomo, N.（2007）"Statistical Disclosure Control Methods for Census Frequency Tables", *S3RI Methodology Working Papers* M07/04, pp. 1-40.

Shlomo, N., C. Tudor, and P. Groom（2010）"Data swapping for protecting census tables". J. Domingo-Ferrer and E. *Magkos*（eds.）*Privacy in Statistical Databases UNESCO Chair in Data Privacy International Conference, PSD 2010 Corfu, Greece, September, 2010 Proceedings*, pp. 41-51. New York: Springer.

Zayatz, L.（2007）"Disclosure Avoidance Practices and Research at the U.S.Census Bureau: An Update", *Journal of Official Statistics*, Vol. 23, No. 2, pp. 253-265.

Willenborg, L. and T. de Waal（2001）*Elements of Statistical Disclosure Control* , Springer, New York.

（*中央大学経済学部教授　博士（経済学））

（**（独）統計センター 情報ソリューション課システム運用担当係長）

（***総務省統計局総務課国際第一係主査）

（****（株）NTT セキュアプラットフォーム研究所研究員　博士（工学））