ミクロデータにおけるスワッピングの適用可能性の検証

明海大・経済　伊藤　伸介*　統計センター　星野　なおみ

諸外国では、ノイズの導入やスワッピングといった攪乱的手法(perturbation)の有効性に関する研究は、少なくとも 1970 年代に遡ることができ(Dalenius and Reiss(1978)等)、これまでも数多くの実証研究が行われてきた。一方、わが国における攪乱的手法に関する実証的な研究については、Takemura(2002)による人口動態調査死亡票の個票データを用いたスワッピングの研究、さらには伊藤・村田(2013)による家計調査の個票データを用いたミクロアグリゲーションや加法ノイズの有効性の研究等があるが、諸外国と比べると研究蓄積は非常に少ないと思われる。ミクロデータに対する攪乱的手法の適用可能性を検証することによって、匿名データの作成において実用的な匿名化技法の範囲が拡大することが期待されることから、わが国でも攪乱的手法についてはさらなる実証的な研究の必要性は高いと思われる。そこで、本報告では、攪乱的手法の 1 つであるスワッピングに焦点を当て、わが国の政府統計ミクロデータに対するスワッピングの適用可能性について検討を試みた。

本研究では、平成 17 年国勢調査の個票データの中から特定の地域(以下「地域 A」と呼ぶ。)を対象に、個人単位で抽出した約 100,000 レコードを用いて、スワッピングの実験を行った。本実験では、最初に、外観識別性等を考慮した 11 個のキー変数を用いて母集団一意を計測し、スワッピングの対象となるレコードを探す。つぎに、スワッピングの対象レコードの中で、相対的にリスク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、母集団一意の対象レコードについて、11 個の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母集団一意に該当した回数をスコアとして算出した上で、スワッピングの対象レコードの中でリスクが高い(優先順位が高い)レコードを選び出した。最後に、リスクの高いレコードに対してスワッピングを試行的に適用した。本実験では、スワッピング率を p%に設定した上で、(1)ターゲットスワッピングと(2)ランダムスワッピングの 2 種類のスワッピングを行った。また、本実験においては、対象レコードに対してスワッピング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(「地域 B」)のドナーファイル(約 50,000 レコード)から探索し、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を計測した上で、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た。

このようにしてスワッピングが施された個票データに対して、本研究では、有用性と秘匿性の定量的な評価を行った。有用性については絶対距離の平均値(average absolute distance)やクラーメルの V を、秘匿性に関してはクロス表に基づいた評価指標をそれぞれ算出するだけでなく、R-U マップによる有用性と秘匿性の比較・検討も行っている。本分析結果によれば、今回使用した国勢調査のミクロデータに関しては、ターゲットスワッピングのほうがより有効であることが明らかになった。

参考文献

Dalenius, T. and Reiss, S. P. (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.

伊藤伸介・村田磨理子(2013)「家計調査ミクロデータを用いた攪乱的手法の有効性に関する研究」,『製表技術参考資料』No.22, 1～26 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)

Takemura, A.(2002) "Local Recoding and Record Swapping by Maximum Weight Matching for Disclosure Control of Microdata Sets", *Journal of Official Statistics*, Vol.18, No.2, pp.275-289.

* (独)統計センター非常勤研究員

2013年度統計関連学会連合大会(於 大阪大学)

ミクロデータにおけるスワッピングの適用可能性の検証

2013年9月9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

目次

＊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

政府統計ミクロデータの作成・提供においては、様々な匿名化技法が適用されている(Domingo-Ferrer and Torra(2001a), Willenborg and de Waal(2001), Duncan *et al.*(2011))

(1)非攪乱的(non-perturbative)な手法

・リコーディング(global recoding, local recoding)
・データの削除(record suppression, attribute suppression)
・トップ(ボトム)・コーディング等

(2)攪乱的な(perturbative)な手法

・加法ノイズ(additive noise)
・乗法ノイズ(multiplicative noise)
・**スワッピング(data swapping)**(ランク・スワッピングを含む)
・ラウンディング(rounding)
・ミクロアグリゲーション(microaggregation)
・PRAM(Post RAndomisation Method)等

わが国では、攪乱的手法(perturbation)の有効性に関する実証的な研究は多くない。

→攪乱的手法に関する実証的な研究によって、その実用可能性を検討することができれば、匿名データ作成において実用的な匿名化技法の範囲が拡大することが期待できる。

ミクロアグリゲーションの適用可能性に関する実証研究(伊藤他(2008), 伊藤他(2009), 伊藤他(2010))

ノイズの付加の有効性に関する研究(Ito and Murata(2011), 伊藤・村田(2013))

⇒それ以外の攪乱的手法についても実証的な研究が必要である。

本報告の目的

攪乱的手法の1つであるスワッピングに焦点を当て、スワッピングの可能性について検討を試みる。

2. スワッピングの概要

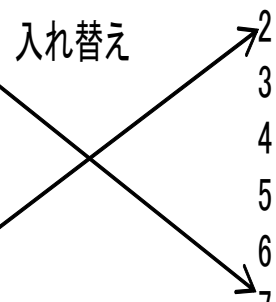
スワッピング(data swapping)とは・・・ミクロデータに含まれるレコード同士で属性値を入れ替えること (Willenborg and Waal(2001, p.126))

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))

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



秘匿処理済データ

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別データ
- 秘匿処理済データ : 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

地域のみ入れ替えを行っているので、上記の例の場合、スワッピング後の秘匿処理済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、スワッピング前の原データにおけるクロス表の数値と変わらない。

諸外国におけるスワッピングの適用事例

・アメリカでは、2000年人口センサスのPUMS(Public Use Microdata Samples)において，スワッピングを適用している。

⇒「特殊な一意(special uniques)」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において特定の人口社会的属性群に基づいて一意性を有する世帯のレコードは、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))

・イギリスでは、人口センサスの個別データの作成において、レコードスワッピングが適用されている (Shlomo(2007))。

⇒アメリカでも、2000年人口センサスの集計表における秘匿処理として、人口センサスの個別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。

３．ミクロデータにおける有用性と秘匿性の評価について

ミクロデータに対する匿名化技法の有効性

→様々な匿名化技法が適用されたミクロデータに対する有用性と秘匿性の定量的な評価(伊藤他(2010))

・スワッピングにおいても、スワッピングが適用された秘匿処理済データにおいて有用性と秘匿性の定量的な評価研究を行うことが求められる。

ミクロデータにおける主な有用性の評価方法(Domingo-Ferrer and Torra(2001a), Karr *et al.*(2006), Shlomo(2010), 伊藤(2012)等)

(1)基本統計量の計測⇒平均、分散等

(2)クロス集計表による比較

①セルに含まれる度数の変化の程度

②クロス集計表における関連性の指標　ex. クラメールのV等

(3)情報量損失(Information loss)の計測

1)量的属性の場合

①平均平方誤差（Mean square error）

②平均絶対誤差（Mean abs. error）

③平均変化率（Mean variation）

⇒属性値、相関係数行列、分散共分散行列等をもとに、情報量損失を計測

2)質的属性の場合

①エントロピーによる情報量損失の計測

＊回帰分析を用いた有用性の評価方法も提案(Woo *et al.*(2009), Karr *et al.*(2006))

ミクロデータにおける秘匿性の主な評価方法(Duncan *et al.*(2011),伊藤(2012))

(1)外部情報とミクロデータのマッチング

(2)ミクロデータにおける母集団一意の計測
⇒ファイルレベルのリスク評価法

・シナリオに基づいてキー変数を設定した上で、母集団一意を計測

⇒イギリスでは、2001年のSARsにおいて母集団一意となるレコードの比率が、露見リスクに関する主要な指標として用いられた(Gross *et al.*(2004))。

(3)特殊な一意の分析(Special Uniques Analysis)(Elliot *et al.*(2002))⇒レコードレベルのリスク評価法

(4)原データと秘匿処理済データとのレコードリンケージ

・確定的リンケージ (Deterministic record linkage)

・距離計測型リンケージ (Distance-based record linkage)

・確率的リンケージ (Probabilistic record linkage)

母集団一意の計測のイメージ

キー変数の例：性別、世帯人員区分、職業のクロス表

（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

キー変数：個別データ以外から取得可能であり、個体の識別が可能な変数(佐井(2000，229頁))

露見リスクの評価については、母集団一意性(Population Uniqueness)等によって評価

$$\text{母集団の一意性} = \frac{\text{母集団一意のレコード数}}{\text{母集団のレコード数}}$$

キー変数について

キー変数の選択に関する明確な基準は存在しない
(Elliot and Dale(1999, pp.8-9))。

キー変数の情報源

1)一般に利用可能な情報

2)個人的な(非公式の)知識

3)企業や政府といった組織が保有する(organizational)データベース

＊選択されるキー変数は、ミクロデータの入手者が個体情報の識別においていかなる戦略(strategy)を想定するか 、さらには、ミクロデータの入手者にとってどのような属性がキー変数として取得可能かによって、異なっている(伊藤(2010, 6頁))。

特殊な一意分析(Special Uniques Analysis)について(Elliot(2001))

特殊な一意問題(special uniques problem)

「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、母集団において一意となるレコードとして特定化される可能性が非常に高くなると考えられる」

↓

特殊な一意の分析(Special Unique Analysis)

⇒SUDA(Special Uniques Detection Algorithm)の開発(Elliot *et al.*(2002))

→イギリス国家統計局だけでなく、オーストラリア統計局、ニュージランド統計局等でも用いられている。

本研究における「特殊な一意」のイメージ

（クロス表1）

		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	
	:	:		:		:		:	
	農林漁業従事者	57		45		1		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	

同じレコードが母集団一意に該当

（クロス表2）

		性別							
		男				女			
		世帯人員区分							
		2人	…	8人	…	2人	…	8人	…
産業	農業	56		15		43		8	
	林業	68		39		18		5	
	漁業	57		35		20		1	
	:	:		:		:		:	
	公務	350		200		83		39	

クロス表1では、母集団一意のセルが2つ存在するが、その中で、性別が女、世帯人員区分が8人世帯、職業が生産工程・労務作業者の属性値を持つレコードは、クロス表2においても母集団一意のセルに該当している⇒母集団一意のレコードの中でも、リスクが相対的に高いレコード⇒特殊な一意となるレコードの可能性が高い。

４．スワッピングの有効性に関する検証の試み―国勢調査ミクロデータを用いて―

本実験では、国勢調査ミクロデータを用いて、スワッピングの有効性に関する検証を試みる。

使用するデータ: H17国勢調査の個別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 をもとに作成したサンプルデータ、約100,000レコード

⇒個人単位で抽出したデータを使用している。

←将来的には地域分析用のミクロデータに関するニーズが考えられることを考慮

本研究における分析方法

(1) キー変数を用いて母集団一意を計測し、スワッピングの対象となるレコードを探す。

(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をスコアをもとに探索する。

(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。

(1)キー変数を用いた母集団一意の計測

以下のキー変数を用いて、母集団一意を計測する。

- 世帯主との続き柄(13区分)
- 男女の別(2区分)
- 年齢5歳階級(25区分)
- 配偶関係(5区分)
- 国籍(13区分)
- 労働力状態(9区分)
- 従業上の地位(8区分)
- 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(4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)

キー変数の組み合わせによって、地域Aにおける母集団一意の数は変わる。

キー変数の組み合わせ	母集団一意の数	母集団一意の比率(%)
世帯主との続き柄, 性別, 年齢5歳階級(3変数)	33	0.03
世帯主との続き柄, 性別, 年齢5歳階級, 国籍, 従業上の地位(5変数)	617	0.59
世帯主との続き柄, 性別, 年齢5歳階級, 国籍, 建て方の種類(5変数)	1543	1.47
世帯主との続き柄, 性別, 年齢5歳階級, 国籍, 従業上の地位, 産業, 建て方の種類(7変数)	9852	9.42
世帯主との続き柄, 性別, 年齢5歳階級, 配偶関係, 国籍, 労働力状態, 産業, 従業上の地位, 建て方の種類(9変数)	17509	16.74
世帯主との続き柄, 性別, 年齢5歳階級, 配偶関係, 国籍, 労働力状態, 産業, 職業, 従業上の地位, 住居の種類, 建て方の種類(11変数)	32064	30.65

キー変数が11変数の場合の母集団一意となるレコードをスワッピングの対象となるレコードとする。

(2)スワッピングの優先度が高いレコードの探索

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方法

⇒母集団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**母集団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(例えば、10個のクロス表で母集団一意に該当したのであれば、10点とする等)。

→スコアが高いほどリスクが高いレコードと考えることができることから(特殊な一意に該当するレコードの可能性大)、スワッピングの優先順位が高くなる。

(3) スワッピングの試行的な適用

本実験では、地域Aのレコード(約100,000レコード(建て方の種類が空欄であるレコードを削除している))に対して試行的にスワッピングを適用する。

＊母集団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、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、8、10、15、20%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50,000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については、特殊な一意のような形で出てくる可能性が高い。

⇒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で見つかる可能性は低いと考えられる。

↓

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えばよいと考えられる。

⇒質的属性値間の距離（distance for categorical variables）を定義した上で（Domingo-Ferrer and Torra (2001a, 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実行する。

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

①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「建て方の種類」で「共同住宅」の場合を除く)、分類区分の区分数で割る(不詳も1つの区分と考える)。

②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、25で割った値をスコアとして新たに付与する。また、「建て方の種類」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。

スワッピングにおける有用性と秘匿性の評価について

(1)有用性の評価

・クラメールのVによる評価

・絶対距離の平均値(average absolute distance)による評価

(2)秘匿性の評価

・クロス表による評価

(3) R-Uマップによる有用性と秘匿性の検証

(1)スワッピングにおける有用性の評価

1)クラメールのVを用いた評価

有用性についてm×nのクロス表における関連性の尺度であるクラメールのVを用いて評価を行う(Shlomo *et al.*(2010))。

クラメールのVの公式

$$CV = \sqrt{\frac{\chi^2}{N \cdot Min(m-1, n-1)}}$$

なお

$$\chi^2 = \sum_i^n \sum_j^m \frac{\left(f_{ij} - F_{ij}\right)^2}{F_{ij}}$$

f_{ij}:i行j列のセルの観測値

F_{ij}:i行j列のセルの期待度数

クラメールのVを用いた有用性の評価指標 (Shlomo *et al*.(2010))

$$\text{有用性の評価指標} = \frac{CV(T^P) - CV(T^O)}{CV(T^O)} \times 100$$

$CV(T^O)$: 原データを用いて作成したクロス表におけるクラメールのV

$CV(T^P)$: 秘匿処理済データを用いて作成したクロス表におけるクラメールのV

＊有用性の指標が大きいほど、スワッピングの影響が大きい。

2)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

（2）秘匿性の評価(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

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

本研究では、最初に、キー変数の中から2変数を選んだ場合のすべての組み合わせについてクロス表を作成した上で、有用性の評価を試みた。

⇒一部の分析結果について紹介

1)年齢5歳階級×国籍

2)従業上の地位×産業

3)産業×職業

4)住居の種類×建て方の種類

5)世帯主との続き柄×労働力状態

→クラメールのVによる指標と絶対距離の平均値による有用性の評価の比較

有用性の評価指標(絶対距離の平均値)の試算結果

スワッピング率とスワッピングの種類	年齢×国籍	従業上の地位 × 産業	産業×職業	住居の種類 × 建て方の種類	世帯主との続き柄 × 労働力状態
1%(Targeted)	1.9323	2.1000	1.6636	1.0109	5.5043
1%(Random)	0.4369	0.9556	0.9636	0.5018	1.5043
2%(Targeted)	3.0831	3.8111	2.9091	1.8145	9.6752
2%(Random)	0.6954	1.6444	1.8273	0.9091	2.8889
3%(Targeted)	4.0615	5.3000	4.2636	2.5164	12.7863
3%(Random)	0.8677	2.1778	2.5545	1.3382	4.0855
4%(Targeted)	4.6523	6.7556	5.4273	3.2036	16.4103
4%(Random)	1.0954	2.7667	3.1273	1.7818	5.2137
5%(Targeted)	5.2862	7.8889	6.5818	3.8327	19.2650
5%(Random)	1.4092	3.4778	3.6727	2.1564	6.0855
8%(Targeted)	6.6954	11.9222	10.2818	5.7673	26.5812
8%(Random)	2.2523	5.2000	5.9818	3.3636	9.6581
10%(Targeted)	7.2677	14.5333	12.5364	6.9455	30.3077
10%(Random)	2.8738	6.4333	7.0273	4.1236	11.9316
15%(Targeted)	8.8615	20.4333	16.8091	9.1636	37.2821
15%(Random)	3.9692	9.6222	9.5091	6.4364	17.7094
20%(Targeted)	9.8708	30.2111	22.4636	11.4945	46.6325
20%(Random)	5.2677	17.0111	14.8545	9.1018	28.6496

有用性の評価指標(クラメールのV)の試算結果

スワッピング率とスワッピングの種類	年齢×国籍	従業上の地位 × 産業	産業×職業	住居の種類 × 建て方の種類	世帯主との続き柄 × 労働力状態
1%(Targeted)	9.4878	0.0521	0.6741	0.0102	0.2939
1%(Random)	0.0147	0.0178	0.0483	0.0289	0.0775
2%(Targeted)	9.4466	0.1228	1.4620	0.0070	0.5645
2%(Random)	0.0165	0.0309	0.1812	0.0380	0.1335
3%(Targeted)	9.4495	0.1649	1.7908	0.0027	0.7143
3%(Random)	0.0271	0.0316	0.5744	0.0579	0.1975
4%(Targeted)	9.4466	0.2271	2.7296	0.0128	0.8871
4%(Random)	0.0279	0.0471	0.7651	0.0757	0.2454
5%(Targeted)	15.3597	0.2292	3.3252	0.0185	1.0226
5%(Random)	0.0336	0.0663	0.8579	0.0740	0.2671
8%(Targeted)	15.3475	0.3105	3.9002	0.0249	1.3830
8%(Random)	0.0448	0.0593	1.3946	0.0472	0.3948
10%(Targeted)	15.3458	0.3727	4.1604	0.0212	1.5268
10%(Random)	0.0609	0.1268	1.3567	0.0221	0.5102
15%(Targeted)	4.3197	0.3195	4.2946	0.0987	1.5899
15%(Random)	0.0651	0.1868	1.9799	0.0812	0.7511
20%(Targeted)	4.3106	0.3415	4.5316	0.0468	1.9065
20%(Random)	0.0629	0.1786	2.2610	0.1158	1.1204

- スワッピング率を上げるにつれて、有用性の程度が小さくなることが確認される。

- ランダムスワッピングのほうが、ターゲットスワッピングと比較して，全般的に有用性が高い。

- 有用性の評価指標として、クラメールのVを用いた場合、スワッピング率を上げるにつれて、結果数値の動きが傾向的に示されない場合がある。

つぎに、本研究では、キー変数の中から3変数を選んだ場合のすべての組み合わせについてクロス表を作成した上で、有用性と秘匿性の評価を試みた。

⇒一部の分析結果について紹介

1)年齢5歳階級×性別×国籍

2)従業上の地位×産業×国籍

3)産業×職業×労働力状態

4)住居の種類×建て方の種類×配偶者の有無

5)年齢×世帯主との続き柄×労働力状態

有用性の評価指標(絶対距離の平均値)の試算結果

スワッピング率とスワッピングの種類	年齢×性別×国籍	従業上の地位×産業×国籍	産業×職業×労働力状態	住居の種類×建て方の種類×配偶者の有無	年齢×世帯主との続き柄 ×労働力状態
1%(Targeted)	0.9785	0.3607	0.2899	0.2111	0.2790
1%(Random)	0.2554	0.0966	0.1586	0.1280	0.1149
2%(Targeted)	1.5569	0.5957	0.5162	0.3803	0.4855
2%(Random)	0.3815	0.1718	0.2697	0.2123	0.2072
3%(Targeted)	2.0492	0.7932	0.7253	0.5200	0.6475
3%(Random)	0.4738	0.2333	0.3798	0.2837	0.2735
4%(Targeted)	2.3754	0.9470	0.9222	0.6400	0.8253
4%(Random)	0.5908	0.2991	0.4808	0.3655	0.3344
5%(Targeted)	2.6769	1.0821	1.0848	0.7557	0.9668
5%(Random)	0.7569	0.3829	0.5616	0.4437	0.3870
8%(Targeted)	3.3692	1.4906	1.6030	1.1200	1.3354
8%(Random)	1.1662	0.5966	0.9111	0.6708	0.5983
10%(Targeted)	3.7108	1.7248	1.8879	1.3465	1.5385
10%(Random)	1.4738	0.7376	1.0626	0.8098	0.7268
15%(Targeted)	4.5108	2.2333	2.4879	1.8308	1.9385
15%(Random)	2.2185	1.0889	1.4293	1.2597	1.0393
20%(Targeted)	5.1938	2.9521	3.1727	2.3225	2.5347
20%(Random)	3.3200	1.7162	2.0414	1.7858	1.5856

・3変数の場合でも、スワッピング率を上げるにつれて、有用性の程度が小さくなるだけでなく、ランダムスワッピングのほうが、ターゲットスワッピングと比較して、全般的に有用性が高いことが確認できる。

秘匿性の評価指標の試算結果

スワッピング率とスワッピングの種類	年齢×性別×国籍	従業上の地位×産業×国籍	産業×職業×労働力状態	住居の種類×建て方の種類×配偶者の有無	年齢×世帯主との続き柄 ×労働力状態
1%(Targeted)	0.2586	0.2500	0.6727	0.7021	0.6687
1%(Random)	0.9828	0.9853	1.0000	0.9149	0.9755
2%(Targeted)	0.1724	0.0735	0.4727	0.5745	0.5337
2%(Random)	0.9828	0.9853	1.0000	0.8511	0.9202
3%(Targeted)	0.1034	0.0294	0.2182	0.4468	0.4233
3%(Random)	0.9828	0.9706	0.9455	0.8085	0.9080
4%(Targeted)	0.1034	0.0147	0.1455	0.3191	0.3374
4%(Random)	0.9655	0.9559	0.9273	0.7872	0.8773
5%(Targeted)	0.0517	0.0147	0.0727	0.1915	0.2515
5%(Random)	0.9138	0.9265	0.9091	0.7660	0.8466
8%(Targeted)	0.0345	0.0147	0.0364	0.1064	0.1288
8%(Random)	0.7586	0.7794	0.7273	0.6170	0.7546
10%(Targeted)	0.0345	0.0147	0.0364	0.0851	0.0859
10%(Random)	0.6897	0.6765	0.6364	0.5957	0.7055
15%(Targeted)	0.0172	0.0147	0.0182	0.0638	0.0429
15%(Random)	0.5172	0.4559	0.4909	0.4043	0.5276
20%(Targeted)	0.0172	0.0147	0.0182	0.0638	0.0429
20%(Random)	0.3103	0.2647	0.3818	0.3404	0.3558

- スワッピング率を上げるにつれて、秘匿性の程度が相対的に大きくなることが確認される。

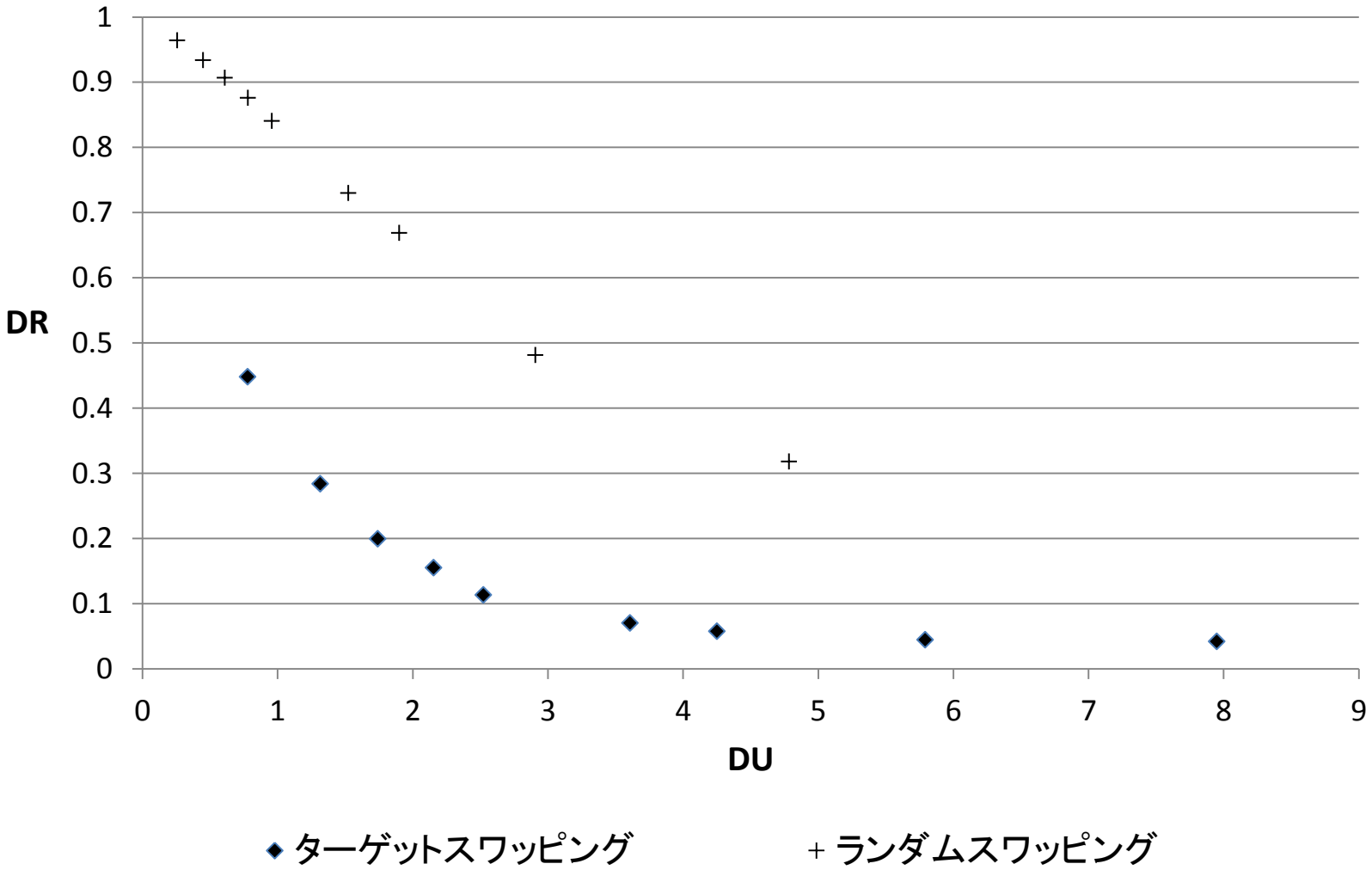
- ターゲットスワッピングのほうが、ランダムスワッピングと比較して、全般的に秘匿性が高くなることがわかる。

（3）R–Uマップによる有用性と秘匿性の検証

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3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。

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
1
2
3
4
5
6
7
8
9
DU
ターゲットスワッピング
ランダムスワッピング

- 2%のスワッピング率において、ターゲットスワッピングを適用した場合、あらゆるランダムスワッピングよりも秘匿性が高くなる。
- 2%のスワッピング率において、ターゲットスワッピングを適用すると、8%のスワッピング率でランダムスワッピングを行った場合よりも有用性が高い。

⇒本分析結果によれば、今回使用した国勢調査のミクロデータに関しては、ターゲットスワッピングのほうがより有効であることがわかる。

5. おわりに

本報告では、匿名化技法としてのスワッピングに焦点を当て、スワッピングの適用可能性について検証を試みた。

本研究では、匿名データ作成のための実用性の観点も踏まえ、「特殊な一意」となるレコードの探索方法、スワッピングを行うための質的属性におけるリンケージ技法、クロス表を用いた秘匿性と有用性の評価方法を展開することによって、スワッピングの有効性を議論することが可能になった。