リサーチペーパー第 39 号　　Research Paper No.39

攪乱的方法を用いて作成する匿名データに関する基礎研究

Basic Research Related to the Production of Anonymized Data Using Perturbation

稲葉 由之
統計研修所客員教授
明星大学経済学部教授

INABA Yoshiyuki
SRTI Guest Professor
Professor, Faculty of Economics, MEISEI University

平成 29 年６月
June 2017

総務省統計研究研修所
Statistical Research and Training Institute (SRTI)
Ministry of Internal Affairs and Communications

受理日：平成 29 年 3 月 31 日

本ペーパーは，総務省統計研修所（現：総務省統計研究研修所）の客員教授が，その責任において行った統計研究の成果を取りまとめたものであり，その内容については，総務省統計局又は統計研究研修所の見解を表したものではない。

本研究では、統計法（平成 19 年法律第 53 号）第 32 条の規定に基づき、国勢調査に係る調査票情報を使用した。

攪乱的方法を用いて作成する匿名データに関する基礎研究

稲葉由之

概　要

匿名データとは，特定の個人や法人などの識別ができないように調査票情報を加工したものである．匿名データの作成方法は，偽のデータを含ませる方法と含ませない方法の 2 種類にわけることができる．偽のデータを含ませる方法を攪乱的方法と呼ぶ．本稿では，我が国において，攪乱的方法を用いて作成する匿名データに関する基本的な考え方を提案した．基本的な考え方とは，現在は「不詳」として表現されているデータに対して，スワッピングや異なるレコード間でのデータ交換を適用することである．これらの匿名データの作成方法は，米国やカナダにおいて匿名データ作成に使用されている方法や無回答へのインピュテーションに使用されている方法と同様である．

キーワード：匿名データ，攪乱的方法，国勢調査，インピュテーション，母集団一意

Basic Research Related to the Production of Anonymized Data Using Perturbation

INABA Yoshiyuki

Abstract

Anonymized data in questionnaire information refer to data that have been processed to prevent the identification of specific individuals or corporations. There are two methods of producing anonymized data: one that uses fake data, and one that uses authentic data. The former method is called perturbation. In this paper, we offer a fundamental approach for producing anonymized data using perturbation, in Japan. This approach is applied to data that have been currently encoded as “unidentified,” and involves the substitution of values through the application of “swapping” or exchanging of data between different records. The resulting methods for producing anonymized data are similar to methods used in both the US and Canada and methods for the imputation of nonresponses.

Keywords: anonymized data; perturbation; Population Census; imputation; population unique

1. はじめに

統計法第 2 条第 12 項に規定された匿名データとは，一般の利用に供することを目的として，調査票情報を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む）ができないように加工したものである．統計法第 32 条から第 38 条までの規定に基づく調査票情報の二次的利用として，匿名データを利用するほかに，オーダーメード集計や調査票情報の目的外利用がある．それぞれの二次的利用は，利用者の制限や調査票情報の管理の面などにおいて違いがある．

匿名データは，調査票情報を特定の個人又は法人その他の団体（以降，個体と呼ぶ）の識別ができないようにデータ変換するため，調査票情報よりも含まれる情報は粗くなる．匿名化におけるデータ変換の方法として，総務省では，2017 年現在，リサンプリングや識別情報の削除等，特異なレコードの削除，トップ/ボトムコーディング，リコーディング，スワッピングなどの方法を用いている．

匿名データの作成過程を匿名化と呼び，匿名化の基本概念を星野（2010）の定義を用いて説明する. 調査票情報を $n \times p$ 行列 $\boldsymbol{X}$, 匿名データを $m \times q$ 行列 $\boldsymbol{Y}$ として $(n \geq m,\ p \geq q)$ ，匿名化を行列 $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ で表すと，調査票情報と匿名データとの関係は,

$$\boldsymbol{Y} = \boldsymbol{AXB} + \boldsymbol{C} \qquad (1)$$

と表現される．このとき，$m \times n$ 行列 $\boldsymbol{A}$ はレコード操作を行う行列，$p \times q$ 行列 $\boldsymbol{B}$ は変数操作を行う行列，$m \times q$ 行列 $\boldsymbol{C}$ は移動操作を行う行列である．匿名化では，行列 $\boldsymbol{Y}$ の行が個体として識別されないように行列 $\boldsymbol{X}$ を行列 $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ によりデータ変換を施す．匿名化を実施する行列 $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ において，リサンプリングや特異なレコードの削除は行列 $\boldsymbol{A}$ ，識別情報の削除等やトップ/ボトムコーディング，リコーディングは行列 $\boldsymbol{B}$ ，ノイズ付加は行列 $\boldsymbol{C}$ によって実施する．

匿名化は，偽のデータを含ませる方法と含ませない方法の 2 種類にわけることができ，偽のデータを含ませる方法を攪乱的方法と呼ぶ．2017 年現在，総務省が用いている匿名化のうち，攪乱的方法に分類される方法はスワッピングのみであり，他の方法はすべて，非攪乱的方法である．非攪乱的方法による匿名化では，情報の粗いデータになるものの偽のデータにはならない．一方，攪乱的方法による匿名化では，偽のデータを含むことになるものの情報損失は少ないデータになることが多い．たとえば，62 歳を年齢 5 歳階級 60～64 歳に変換するリコーディングという非攪乱的方法は，調査票情報に比べて情報は粗くなるもののデータとしては真である．また，たとえば，異なる都道府県間で同じような世帯を交換するスワッピングという攪乱的方法は，地域情報は偽となるものの地域情報を除いた世帯データとしての情報損失は全くないことになる．

本稿では，攪乱的方法を用いて作成する匿名データの作成と評価に関する提案を行う．つぎの第 2 節において攪乱的方法を整理し，第 3 節において匿名データの開示リスクと国勢調査の調査票情報を利用した母集団一意に関する計算結果を示す．第 4 節では，フォローアップ調査と補定処理について我が国の状況を米国とカナダの状況とともに整理し，第 5 節では匿名化と補定処理の評価に係わる提案を行う．最後に第 6 節において匿名データ作成ならびに利用における課題に言及する．

2. 攪乱的方法

2.1 攪乱的方法の種類と性質

(1) スワッピング

スワッピングとは，異なるレコード間で値を交換して匿名データを作成することをいう．たとえば，異なる都道府県間の世帯を交換する，または同一地域内で世帯に含まれる世帯員を交換するといった方法である．世帯員を交換する場合，世帯に係るデータには偽を含むものの個人に係るデータとしては真となる．また，変数の値がほぼ同様な世帯間で世帯員を交換すれば，匿名化による調査票情報の情報損失はほとんどない．

(2) ノイズ付加

ノイズ付加とは，調査票情報の値にノイズを加えて匿名データを作成することをいう．たとえば，年間収入（万円）の値に正規分布 $N(0, 10)$ の乱数を加えて，調査票情報とは異なる値を匿名データとして代入する．ノイズ付加は前節（1）式における行列 $\boldsymbol{C}$ により実施する．乱数の散らばりを小さくすれば，匿名データと調査票情報との違いはほとんどない．

(3) 模造データ

模造データとは，経験分布に基づく母集団分布から個体を抽出して，その個体のもつ値を模造データとして代入することをいう．たとえば，10 年分の家計調査の調査票情報（120 回分の調査結果）に基づいて世帯属性別の年間収入の経験分布を作成する．同じ世帯属性の経験分布から抽出した値を一部分のレコードの年間収入の値として代入して匿名データとする．なお，確率分布を想定して抽出するのではなく，実際の調査票情報を格納したデータベースから値を抽出してもよい．このような方法は，異なる時点の調査間において世帯に係わるデータをスワッピングしたと解釈することもできる．

(4) その他

カテゴリカル変数におけるノイズ付加に該当する PRAM（Post Randomization Method）や合成データ生成などの方法がある．

2.2 諸外国における攪乱的方法の適用

(1) 匿名化における攪乱的方法の位置づけ

諸外国において，匿名データ作成に用いられている主な攪乱的方法は，スワッピングとノイズ付加である．攪乱的方法の適用について詳細は開示されていないため，攪乱的方法を適用している比率や詳細な手順はわからない．また，非攪乱的方法と合わせて匿名化を行っているため，攪乱的方法のみで匿名化を評価することは難しい．以下に，諸外国における攪乱的方法の適用について簡単に整理する．

(2) スワッピング

米国の ACS（American Community Survey）のミクロデータ提供では，地域間で世帯を入れ替えている．また，カナダの Public Use Microdata File において，世帯員の変数や世帯員自体，世帯の入れ替えを実施しているが詳細は公表されていない．米国やカナダでは，地域間での世帯の交換，世帯員の交換，世帯員の調査項目の交換などを行っている．

（3）ノイズ付加

米国やカナダにおいて，ノイズ付加は年齢や収入に対して適用されている．ただし，その散らばりの度合いや何%のデータに適用しているかなどについて，その詳細は公表されていない．

（4）その他の方法

PRAM（Post Randomization Method）や合成データ生成についても，その詳細は公表されていない．

3. 母集団一意

3.1 開示リスクと母集団一意，攪乱的方法

開示リスクの評価に関して，星野（2016）の提示した確率モデルを用いて，個体識別と母集団一意との関係を表す.

$$Pr(\text{識別成功}) = Pr(\text{識別成功}|\text{識別を試みる})Pr(\text{識別を試みる}) \quad (2)$$

$$Pr(\text{識別成功}|\text{識別を試みる}) = Pr(a)Pr(b|a)Pr(c|a,b)Pr(d|a,b,c) \quad (3)$$

a :公開ファイルのレコードと攻撃ファイルのレコードにおいてキー変数の値が同じならば両者のレコードの個体属性は同じである.
b ：公開ファイルに個体が含まれる.
c ：個体が母集団一意である.
d ：個体が母集団一意であると確証できる.

(2) 式のように，個体識別が成功するのは，識別を試みて，かつ識別が成功した状態を指す．(3) 式の確率モデルでは，個体が母集団一意であっても，母集団一意であることが確証されなければ個体は識別されたとは言わない．つまり， $Pr(d|a,b,c)$ が 0 であれば個体識別は不可能な状態になる．しかし，$Pr(d|a,b,c)$ 自体の評価は困難である．一方，$Pr(a,b,c)$ は調査票情報に基づいて評価することが可能である．星野（2016）は，$Pr(a,b,c)$ が適当な閾値を下回ることを $Pr(d|a,b,c)=0$ と同値として捉えて，住宅・土地統計調査により実際の開示リスクを評価した.

匿名化に攪乱的方法を用いた場合，(3) 式の確率モデルにおける $Pr(a)$ は 1 とは言えない状態となる．また，攪乱的方法を匿名化に利用して，かつ攪乱的方法を適用した調査項目や適用した比率を開示しないときには，攻撃者は $Pr(a)$ を評価することができない．このため，攻撃者による個体識別が実際に起こっているか否かは，攻撃者にはわからないことになる．この点が攪乱的方法を匿名化に用いる上での大きな利点である．匿名データ作成に攪乱的方法を用いて，その適用した調査項目や比率を明かさない場合，攻撃者による個体識別は不確実性をもつことになる.

3.2 米国における母集団一意に係わる考え方

(1) 母集団一意の計算に用いるキー変数

米国センサス局において提供するミクロデータは，地域区分を 10 万人以上とする基準を設けている．本節では，地域区分 10 万人以上の根拠例として，Hawala(2001) の内容を紹介し，米国センサス局の母集団一意に係わる考え方を示す.

Hawala(2001) による母集団一意に関する検討は，1990 年センサスの 3 種類のミクロデータのうち，都市圏 1%サンプルの Public-Use Microdata を使用した．これは，世帯データではなく個人データである．また，厳密には母集団一意ではなく，標本一意の計算であると考えられる．このため，説明では，母集団一意の比率ではなく一意となる比率と表現する.

キー変数に該当するカテゴリー作成の基となった調査項目は以下のとおりである．これらのカテゴリーの組合せにより 9 つのシナリオ（カテゴリー数 2.5×10^{18}（250 京）～324,000

（32 万 4 千））を作成した.

＜カテゴリー作成の基となった調査項目＞

年齢　90 カテゴリーまで（最大 90 カテゴリー）
人種　64 カテゴリー
性別　2 カテゴリー
Hispanic Origin　64 カテゴリー
祖先　143 カテゴリー
生誕地　167 カテゴリー
職業　443 カテゴリー
言語　74 カテゴリー

(2) 母集団一意の計算結果

9 つのシナリオ別一意の比率を表 1 に，グラフに表現したものを図 1 に示す．図 1 において一意の比率の人口規模による推移をみると，10 万人までの減少幅に比べて 10 万人から 50 万人までの範囲における減少幅は小さい．Hawala(2001) は，この減少率の状況を基づいて，地域区分 10 万人以上の根拠を支える一つの証拠であると論じている．つまり，Hawala(2001) は一意となる比率の減少率を根拠としており，シナリオによって大きく異なる一意となる比率自体を根拠としているわけではない．Hawala(2001) の行った評価は確率モデル（(3) 式）における $Pr(a,b,c)$ 自体を考慮していないため，本来は開示リスクの評価にはならない．

Hawala(2001) の示した計算は地域区分 10 万人以上の根拠とは言えないため，米国センサス局において別の考え方により 10 万人基準の根拠が示された可能性がある．ただし，Hawala(2001) の示した地域区分の規模に係わる扱いにより，米国センサス局では母集団一意であること自体を大きな問題にしていないということがわかる．

表1　シナリオ，地域区分の規模別 一意となる比率

	一意の比率(%) 人口(人)	シナリオ1 250京	シナリオ2 4.6京	シナリオ3 110兆	シナリオ4 4,500億	シナリオ5 390万	シナリオ6 450万	シナリオ7 65万	シナリオ8 43万	シナリオ9 32万
1	21,034	90.7	91.1	90.5	72.6	32.6	22.7	12.7	11.0	15.7
2	47,421	84.7	84.7	83.5	56.9	23.1	14.6	8.1	6.4	10.6
3	73,813	81.3	81.3	79.7	48.3	18.6	12.0	6.0	4.9	8.5
4	100,199	79.1	78.7	77.0	43.0	16.0	9.8	5.1	3.8	7.3
5	126,587	77.0	77.1	74.3	38.6	14.0	8.7	4.4	3.2	6.4
6	152,974	75.6	75.6	72.3	35.8	12.6	7.7	4.0	2.8	5.7
7	179,364	74.4	74.2	70.7	33.3	11.6	6.9	3.6	2.5	5.2
8	205,753	73.2	73.1	69.1	31.3	10.7	6.4	3.4	2.2	4.9
9	232,139	72.0	72.2	68.1	29.6	9.9	5.9	3.2	2.0	4.6
10	258,531	71.3	71.4	67.0	28.2	9.3	5.5	3.0	1.8	4.3
11	284,919	70.6	70.6	65.8	27.0	8.8	5.2	2.8	1.7	4.0
12	311,309	69.9	69.6	64.9	26.0	8.4	4.8	2.7	1.6	3.9
13	337,697	69.2	69.2	64.1	25.0	8.0	4.5	2.6	1.5	3.7
14	364,086	68.8	68.7	63.2	24.3	7.7	4.3	2.5	1.4	3.5
15	390,475	68.1	68.0	62.4	23.5	7.4	4.1	2.3	1.3	3.4
16	416,865	67.7	67.5	61.9	22.8	7.1	3.8	2.3	1.2	3.2
17	443,250	67.0	67.2	61.0	22.2	6.8	3.7	2.2	1.2	3.2
18	469,645	66.7	66.7	60.5	21.6	6.5	3.6	2.2	1.1	3.0
19	496,028	66.2	66.2	59.8	21.0	6.4	3.4	2.1	1.1	3.0
20	522,417	65.8	65.8	59.2	20.7	6.1	3.3	2.0	1.0	2.9

注:シナリオの下の値(250京, 4.6京, . . .)はカテゴリー数を表す.

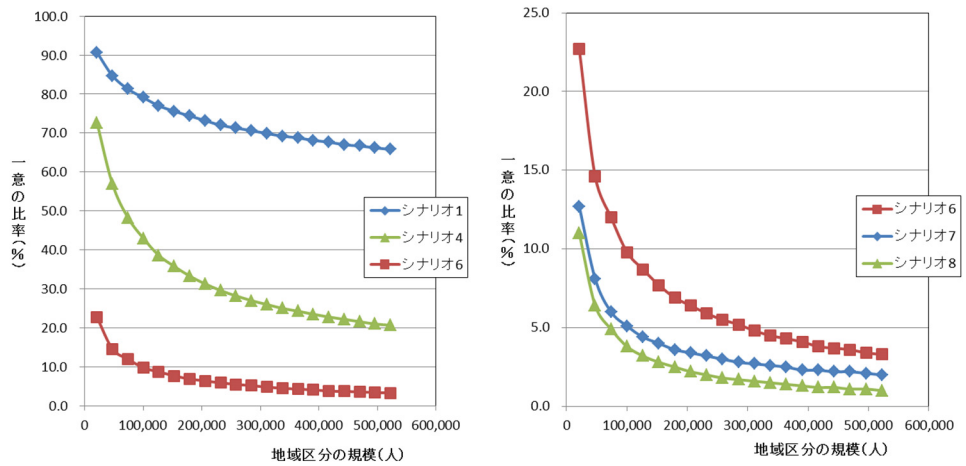

図1　シナリオ，地域区分の規模別 一意となる比率

3.3　2010 年国勢調査の調査票情報を用いた母集団一意

（1）母集団一意の計算に用いるキー変数

米国の事例を踏まえて, 2010 年国勢調査の調査票情報を用いた母集団一意の計算を統計研修所のミクロデータ利用室において行った．本節では，母集団一意の状況を簡単に示すとともに，$Pr(d|a,b,c)$ の評価を行う．具体的には，母集団一意となった個体を実際に観察することにより，母集団一意の個体が間違いなく母集団一意であると判断できるのか否かを確かめる．この手段は主観的な判断に基づいており，かつ内容が母集団一意に関するもののため証拠を詳細に示すことができない．ただし，攻撃者の用いる攻撃ファイルについて知る機会のない状況下では, この主観的判断は有効な情報になると考える. なお, 本節で示す内容は, 母集団一意に係わる内容であるため，変数の組合せや数値を詳細に示さない.

カテゴリー作成の基となった調査項目は以下のとおりである．これらのカテゴリーの組合せにより 3 つのシナリオ（カテゴリー数　1 万程度～約 100 万未満）を作成した. このうち, 外観から判別できる情報は，男女別，世帯の種類，住居の種類，住宅の建て方，住宅の床面積であり，それらに準ずる情報として地域区分，年齢 5 歳階級，家族類型を加えた．なお, 不詳はカテゴリーに含めずに除外して集計を行った.

＜カテゴリー作成の基となった調査項目＞

男女別　2 カテゴリー
世帯の種類　8 カテゴリー
住居の種類　8 カテゴリー
住宅の建て方　4 カテゴリー
住宅の床面積　14 カテゴリー
地域区分　47 カテゴリー
年齢 5 歳階級　23 カテゴリー
家族類型　16 カテゴリー

＜調査項目におけるカテゴリー＞

男女別　2 カテゴリー
男；　女

世帯の種類　8 カテゴリー
一般の世帯；　一人世帯（会社等の独身寮の入居者を含む）；
寮・寄宿舎の学生・生徒；　病院・療養所の入院者；　社会施設の入所者；
その他；　自衛隊営舎居住；　矯正施設入居者

住居の種類　8 カテゴリー
持ち家；　都道府県･市区町村営の賃貸住宅；　都市再生機構･公社等の賃貸住宅；
民営の賃貸住宅；　給与住宅(社宅・公務員住宅など)；　住宅に間借り；
会社等の独身寮･寄宿舎；　その他

住宅の建て方　4 カテゴリー
一戸建；　長屋建(テラスハウスを含む)；　共同住宅；　その他

住宅の床面積　14 カテゴリー

20 ㎡未満；　20～30 ㎡未満；　30～40 ㎡未満；　40～50 ㎡未満；
50～60 ㎡未満；　60～70 ㎡未満；　70～80 ㎡未満；　80～90 ㎡未満；
90～100 ㎡未満；　100～120 ㎡未満；　120～150 ㎡未満；　150～200 ㎡未満；
200～250 ㎡未満；　250 ㎡以上

地域区分　47 カテゴリー

47 都道府県

年齢 5 歳階級　23 カテゴリー

0～4 歳；　5～9 歳；　...　；100～104 歳；　105～109 歳；　110 歳以上

家族類型　16 カテゴリー

夫婦のみの世帯；　夫婦と子供から成る世帯；　男親と子供から成る世帯；
女親と子供から成る世帯；　夫婦と両親から成る世帯；
夫婦とひとり親から成る世帯；　夫婦，子供と両親から成る世帯；
夫婦，子供とひとり親から成る世帯；
夫婦と他の親族（親，子供を含まない）から成る世帯；
夫婦，子供と他の親族（親を含まない）から成る世帯；
夫婦，親と他の親族（子供を含まない）から成る世帯；
夫婦，子供，親と他の親族から成る世帯；　兄弟姉妹のみから成る世帯；
他に分類されない世帯；　非親族を含む世帯；　単独世帯

(2) 母集団一意の計算結果

3 つのシナリオ別のカテゴリー数や母集団一意となる比率を表 2 に示す．先に示したように，母集団一意に係わる内容であるため，変数の組合せや数値については詳細に示さない．ここでは，母集団一意を度数 1 または 2 の状況と定義した．それぞれのシナリオにおいて母集団一意となったのは 3%未満であった．

つぎに，個体が母集団一意であると確証できるか否かについて，該当するレコードを観察することにより主観的に判断した．3 つのシナリオにおいて母集団一意となるカテゴリーは同様な状況ではなく，シナリオによってカテゴリーの組合せは大きく異なっていた．また，該当するレコードを特異なレコードであると判断することもできなかった．それぞれのシナリオにおいて度数 0 のカテゴリーは全カテゴリーの 80%以上と大多数を占めている．一方，含まれるレコードが 1 万を超えるカテゴリーも比較的多く観察された．すなわち，カテゴリーに含まれる度数をみると，その度数には疎密があり，多くのカテゴリーが該当する「疎」の部分においてそのカテゴリーが母集団一意であると確証することは，攻撃者のもつ攻撃ファイルが国勢調査の調査票情報と同様に完全である場合を除いて，ほぼ不可能であると主観的に判断した．攻撃ファイルと照合した確定的な評価ではないものの，確率 $Pr(d|a,b,c)$ はほぼ 0，あるいはかなり小さい値であると仮定してよいと考えられる．ただし，攻撃ファイルに関する情報のない場合，$Pr(d|a,b,c)=0$ と確定できないため，識別成功の可能性がないと断言することはできない．このため，母集団一意となる確率 $Pr(a,b,c)$ の推定値が適当な閾値を下回ることで開示リスクを評価する方法（星野（2016）の提案した方法）は妥当

であると結論付けることができる.

表 2　平成 22 年国勢調査の調査票情報に基づくシナリオ別
カテゴリー数，度数 0 のカテゴリーの占める比率,
及び母集団一意となる比率

	シナリオ1	シナリオ2	シナリオ3
カテゴリー数	1万未満	30～40万	10～20万
度数0のカテゴリー	全カテゴリーの80%以上	全カテゴリーの80%以上	全カテゴリーの80%以上
母集団一意（度数1または2）	1%未満	3%未満	3%未満

注：母集団一意に係わる内容であるため，変数の組合せや数値を詳細には示さない

4. フォローアップ調査と補定処理の状況

4.1 米国における状況

（1）フォローアップ調査

米国においてセンサスの費用はセンサスを実施するたびに増加しており，1990 年センサスで 33 億ドル，2000 年センサスで 65 億ドル，2010 年センサスでは 147 億ドル（計画）に達した（Williams, J. D. (2011), U.S. census bureau (2009)）．2000 年センサスでは，調査員業務のうち約 62%が無回答フォローアップに費やされた．無回答フォローアップでは，郵便による調査地域での調査票未提出の世帯や空き家に対して，調査員の訪問による調査を行ったものである．無回答フォローアップによる調査票情報の補完を施しても全項目無回答になったのは，2000 年センサスでは約 580 万人（総人口の 2.1%），2010 年センサスでは約 570 万人（総人口の 1.9%）であった（U.S. census bureau (2009), Rothhaas *et al*. (2012)）．

（2）公開用ミクロデータの作成

米国センサス局では，エディティングを施していないファイル（Census Unedited File）をはじめに作成して，無回答や矛盾を含む回答に対するエディティングやインピュテーションを施してデータファイル（Census Edited File）を更新していく．この時点で，米国センサス局は個人の回答内容の開示を回避する技術を適用したデータファイルを作成した（U.S. census bureau (2009) p.273）．このような手順をとるため，米国では公開用ミクロデータの完成はサマリーファイル発表後すぐであり，2000 年センサスの場合，2003 年 4 月～7 月に最後のサマリーファイルである第 4 次サマリーファイルが発表されるのと同時期の 2003 年 4 月～6 月に 1%サンプルの公開用ミクロデータ，2003 年 8 月～9 月に 5%サンプルの公開用ミクロデータを完成させている．

（3）補定を行う調査票

補定の対象としては，全項目無回答となった場合，近隣の世帯の状況を代用する方法（Substitution）が適用された（U.S. census bureau (2009) p.309）．一部項目無回答については，以下の Assignment と Allocation という 2 種類の方法を採用した．補定率について一部の回答の状況を示した表 3 によると，2000 年センサスに比べて 2010 年センサスでは，同じ人の他の回答を用いて補定を行うこと（Assignment）が多くなっていることがわかる．たとえば，2000 年センサスにおける性別の補定は，Assignment 0.9% ，Allocation 1.1% であるのに対して，2010 年センサスにおける性別の補定は，Assignment 1.3% ，Allocation 0.3% と，他の人の回答を用いる補定よりも，同じ人の他の回答を用いる補定を多く用いるようになってきたことがわかる．ただし，続柄や年齢については，用いる補定方法の比率について大きな変化はない．

Assignment：無回答に対して，同じ人の他の回答を用いて補定を行った．例えば，「人種」の欄が無回答の場合，「ヒスパニック出身」の回答を利用した．2010 年センサスでは，2000 年センサスの結果や American Community Survey (ACS) の回答も利用した（Rothhaas *et al*. (2012) p.2）．

Allocation：無回答に対して，他の人の回答を用いて補定を行う．

Substitution：世帯員すべてが無回答である場合，近隣の世帯の状況を代用した（nearest neighbor hot deck と同様な方法）．

表 3　2000 年センサスと 2010 年センサスにおける
一部項目無回答に対する補定の種類と補定率

項　目		2010年センサス				2000年センサス			
		回答率 As Reported	補定率 Imputed	Assigned	Allocated	回答率 As Reported	補定率 Imputed	Assigned	Allocated
続柄	Relationship	97.9	2.1	0.5	1.7	97.4	2.6	0.4	2.2
性別	Sex	98.4	1.6	1.3	0.3	98.0	2.0	0.9	1.1
年齢	Age	95.0	5.1	1.5	3.6	94.9	5.1	1.5	3.6
ヒスパニック出身	Hispanic origin	95.5	4.5	1.7	2.8	95.6	4.4	0.2	4.2
人種	Race	95.9	4.1	1.7	2.8	96.0	4.0	0.0	3.9
住宅の所有関係	Tenure	96.5	3.5	n/a	3.5	94.5	5.5	0.7	4.8

Rothhaas *et al*. (2012) xi
Kevin (2003) p.13

(4) 補定方法の実際

2000 年センサスにおける補定方法について説明する. インピュテーションとは無回答などのデータを生成することであり，2000 年センサスでは約 580 万人（総人口の 2.1%）のすべての特性（all their 100 percent characteristics）がインピュテーションの対象となった．米国センサス局では，インピュテーションを count imputation と characteristics imputation とに分けている.

(a) count imputation

無回答フォローアップやデータ収集の終了時に，いくつかの住戸は居住しているのか空き家であるかの情報や世帯人員の情報を含んでいなかった．このとき，米国センサス局では，はじめに住戸の状況と世帯人員の値を割り付けた. count imputation は総人口に影響を与えるため，2000 年 12 月 31 日までに結果が必要とされ，実際には 2000 年 9 月中旬から作業をはじめて 10 月はじめに処理は完了した．米国センサス局では，近隣の住戸は類似しているものと仮定して, インピュテーションに nearest-neighbor hot deck imputation method を利用した. この方法では，一部項目無回答である世帯に対して，その世帯の住戸の状況と世帯人員の情報を割り付ける際に，地理的近接性（区域のブロック番号，通り名，住宅番号でソートされた順列）の情報を用いた．インピュテーションは，household size（住戸に居住しているが，その世帯人員はわからない場合），occupancy status（住居に居住しているか，あるいは空き家であるのかはわからない場合），housing unit status（「居住」,「空き家」,「削除」, のどれかを決めて, 住戸に居住している場合) に対して実施した. 2000 年センサスでは， count imputation によって 117 万人（総人口の 0.42%）が追加された.

(b) characteristics imputation

2000 年センサスにおける characteristics imputation のうち 全項目（性別，年齢，生年月日，続柄，ヒスパニック出身，人種）をインピュテーションの対象にしたのは，約 460 万人（総人口の 1.64%）であった．インピュテーションは，世帯全体に適用する場合と世帯内に適用する場合に分けられる．世帯全体に対するインピュテーションでは，nearest-neighbor hot deck imputation method を利用して世帯の複製を行った．このような世帯の代用によるインピュテーションによって，約 146 万世帯（居住者のいる住戸 1 億 550 万のうち 1.39%）約 344 万人のデータを作成した．世帯内におけるインピュテーションは，世帯内に少なくとも 1 人の data-defined person を含み，その他の世帯員の回答が欠測していたときに適用した．このインピュテーションでは，他の世帯員または他の世帯の回答における同様な情報に基づいて値を割り付けた．2000 年センサスでは，約 126 万世帯約 233 万人に対してインピュテーションを

行った.

＜インピュテーションの内訳＞
count imputation　約 117 万人
characteristics imputation　約 460 万人

世帯全体のインピュテーション　約 344 万人
（count imputation 約 117 万人を除外すると約 227 万人）
世帯員 のインピュテーション　約 233 万人

characteristics imputation　約 460 万人＝約 227 万人＋約 233 万人
count imputation　約 117 万人＋ characteristics imputation　約 460 万人＝ 約 577 万人
（総人口の約 2.1%）

補定の実施には，地理的情報と Dual System Estimation (DSE) の結果が用いられた．DSE の結果とは， Population sample (P サンプル) と Enumeration sample (E サンプル) の違いに関する情報である．P サンプルは抽出された地域内すべての住戸に対して面接調査を行って得られたデータであり，この調査はセンサスとは別に行われた．E サンプルは P サンプルを取得した地域と同じ地域におけるセンサスの結果である．センサスの結果（E サンプル）において無回答の箇所は P サンプルで調査されているため，両者を比較することによって，センサスで回答が得られなかった機構のモデル化を行うことができる．例えば，婚姻関係に無回答である場合はどのような状況にあるのかについては P サンプルと E サンプルの比較によって把握されている．

4.2　カナダにおける状況

（1）フォローアップ調査

カナダでは 2006 年センサスから，回答者負担を減らすために所得税レコードからのデータ使用の許可を求める質問がロングフォームに加えられた. また, 2006 年センサスにおいて，はじめてオンライン回答が導入され，2011 年センサスでは全体の 54.4%がオンラインによる回答であった（Statistics Canada (2012)）．カナダでは郵政省の協力のもと，郵便で返送された回答調査票は，データ処理センターへの配達の前に封筒のバーコードを登録された．これらの情報は日々Master Control System (MCS) へ送信され，無回答フォローアップの情報として利用された．無回答フォローアップでは，電話番号がある場合は初めに電話により調べて，電話で回答が得られなかった場合は訪問によって調査票を補完した．また，無回答フォローアップのほかに，Dwelling occupancy verification や Failed edit follow-up という手段で無回答部分の補完を実施した（Statistics Canada (2012)）.

（2）補定方法

2006 年センサスでは，2 種類の自動インピュテーションを適用した．一つは決定論的な当てはめ（deterministic imputation）であり，あらかじめ決められた規則に従って実施されるものである．もう一つはドナーによる当てはめ（minimum-change donor imputation）であり，ドナーに基づいて，出来る限り少ない項目の変更を施すものである．CANadian Census Edit and Imputation System (CANCEIS) は，2 種類のインピュテーションを自動的に行うシステムであ

る（Statistics Canada (2009)　p.14）．カナダ統計局の開発した補定用プログラム CANCEIS は世界各国に広く利用されている．各国センサスでの CANCEIS の利用（カッコ内は使用年）は，ペルー（2005），アルゼンチン（2010），ブラジル（2000），イスラエル（2008），スイス（2000），英国（2011）である．また，米国 2006 年試験センサスへの CANCEIS 適用に係わる検討（Chen(2007)）も実施されたことがある．

インピュテーションの評価に関連して，カバレッジ誤差（調査対象を捕捉できないことによる誤差）を測定するために以下の 3 つの研究を行った（Statistics Canada (2012)　pp.24-25）．

（a）Dwelling Classification Survey

実際には居住している住戸と空き家との分類間違いに関する研究であり，世帯人員の分布（インピュテーション後）を修正するために利用された．この研究は，はじめの発表としての人口算出に間に合わせるように実施された．

（b）Reverse Record Check

センサスで把握できなかった個人に関する研究であり，以前のセンサス結果などを利用して，実際にセンサスで把握できているか否かを確認した．このとき，移住した人や死亡した人は照合や聴取の段階で識別する．

（c）Census Overcoverage Study

2011 年と 2006 年のセンサスでは，二重カウントはセンサスデータベースにおけるレコード照合（名前，生年月日，性別を利用）によって検出された．この研究は Reverse Record Check と合わせて，カバレッジ誤差の推定に用いられた（Statistics Canada (2010)）．

カナダでは，無回答フォローアップ（Non-Response Follow-Up）を行っており，無回答を少なくする調査体制をとっている．また，カバレッジ誤差を測定するための研究などによりセンサスで把握できなかった世帯の状況を確認して，センサス結果の公表に役立てている．

4.3　米国やカナダと比較した我が国における状況

（1）フォローアップ調査や補定

米国やカナダでは，センサスにおける調査員の処理は何段階にも分かれており，無回答フォローアップを行って，無回答を減らすことに力を入れている．そして，無回答フォローアップでも捕捉できなかった世帯を補定の対象としている．補定の対象としては，一部項目無回答も含まれる．これに対して，我が国における無回答フォローアップは詳細な作業手順等を公表していない．また，一部項目無回答は基本的には「不詳」として表章されている．

（2）公開用ミクロデータの作成

米国では調査結果発表後すぐに公開用ミクロデータを作成している．集計処理と公開用ミクロデータ作成を同時期に実施しているということは，スワッピング等の攪乱的方法の適用も同時期に実施していることになる．また，個人の回答内容の開示を回避する技術を適用したデータファイルも同時に作成している．これに対して，我が国における公開用ミクロデータである匿名データの作成は，調査結果発表後数年経過した時点で行っている．調査票情報を基にして匿名化を行うための検討は，半年以上の期間をかけて実施している．

（3）不詳に係る問題

阿部（2013）は，国勢調査における近年の「不詳」の増加は分析上の大きな問題になっていることを指摘した．その指摘事項の一つとして，東京都杉並区における不詳率の上昇を挙げている．東京都杉並区では，年齢不詳率は平成 17 年国勢調査の 1.1%から平成 22 年国勢調査には 13.8%に達し，杉並区において年齢別人口の比較を行うことが難しいほどの急激な上

昇であった．

小池，山内（2014）は，国勢調査の調査事項別に潜在不詳割合（15 歳以上を対象とした調査事項や枝問に該当する調査項目の不詳を人口分布による按分で求めた不詳の割合）を都道府県別に示した．これによると，地域間の差や調査項目による不詳の割合の違いは大きく，地域分析を行う上での障害になりつつある．

（4）補定に係わる課題

不詳に係る問題を解決するためには，回答率を確保するための無回答フォローアップを行うことと，インピュテーションを行うことが挙げられる．米国におけるセンサス予算の増大でもわかるように，無回答フォローアップには大規模な予算が必要となる．また，米国やカナダと同様に，インピュテーションを行うためには，米国の DSE の結果のような無回答となる機構に関する情報と，それらの評価が必要である．これらの不足する情報や評価の経験を現時点ですぐに埋めることは困難であると予想される．

4.4　匿名データと補定との関連性

米国やカナダにおけるインピュテーション（4.1 と 4.2 を参照）は，基本的には，インピュテーションを行う世帯や世帯員に似た世帯や世帯員の情報を用いて値を代入する方法である．この方法には，似た世帯や世帯員そのものの値を利用する hot deck も含まれる．これらの方法は，目的は異なるものの，実施している内容としては匿名化における攪乱的方法のうちスワッピングと同様である．

5. 匿名化と補定処理の評価に係わる提案

5.1 匿名データへのインピュテーションの実施

匿名データを作成する際に,「不詳」の箇所にインピュテーションを行うことを提案する．これにより，匿名データは,「不詳」を含んだ調査票情報とは異なるものになる．また,「不詳」に対するインピュテーションの比率を公開しないものとする．そして，匿名データの作成に係る検討過程において，これらのインピュテーションの評価を行う．

5.2 統計調査の結果へのインピュテーション導入に関する提案

インピュテーションに関する数回の評価を行ったのち，調査票情報の匿名化は，統計調査の評価に係る過程と同時期に実施することを提案する．すなわち，補定処理前のデータに基づいて匿名データを作成し，匿名データと補定処理の評価を同時に行う．2017 年現在，匿名データの提供において，様々な状況を調べて評価を行う過程を設けている．母集団一意の確率の推定値を求める過程とインピュテーションを評価する過程について，米国で実施しているように，統計調査に係る過程に加えるならば，匿名化に関する評価は簡素化することができる．これにより，結果公表と合わせて匿名データの作成と評価が可能となる．

第 1 段階では，匿名データ作成において「不詳」へのインピュテーションの実施と評価を行い，第 2 段階において，統計調査結果へのインピュテーションの適用を行うことを提案する．今後,「不詳」の占める割合は大きくなり，何らかのインピュテーションを行わなければならない時期が来るかもしれない．本稿で提案した攪乱的方法を用いた匿名化は，そのような状況への検討作業も含んでいる．

6. 匿名データ作成ならびに利用における課題

匿名データの利用を促進し，分析結果を価値の高い成果とするためには，レプリカウエイトに係わる研究が必要になるだろう．レプリカウエイトとは，複数セットのウエイトを提供することにより，分析結果における標準誤差の評価を可能にするものである．このため，集計用乗率を 1 つのみ提供する調査票情報の目的外利用よりも，複数のレプリカウエイトを付与した匿名データの利用の方が，分析の結果判断が容易となる．レプリカウエイトの導入により，匿名データの利用価値は調査票情報の目的外利用よりも高くなる可能性がある．

参考文献

阿部隆（2013）国勢調査結果の「不詳数」に係わる諸問題，統計，**64**，51-54

小池司朗，山内昌和（2014）2010 年の国勢調査における「不詳」の発生状況：5 年前の居住地を中心に，人口問題研究，**70**，325-333.

星野伸明（2010）公的統計ミクロデータ提供制度の課題．日本統計学会誌，**40**，23-45

星野伸明（2016）エビデンスに基づいた匿名化．日本統計学会誌，**46**，1-42

Bor-Chung Chen (2007) CANCEIS Experiments of Edit and Imputation with 2006 Census Test Data, STUDY SERIES(Computing #2007-1).

Kevin J. Zajac (2003) Analysis of Imputation Rates for the 100 Percent Person and Housing Unit Data Items from Census 2000, Final Report, Census 2000 Evaluation B.1.a.

Hawala, S. (2001) Enhancing the "100,000 rule" on the variation of per cent of uniques in a microdata sample and the geographic area size identified on the file. *Proceedings of the annual meeting of the american statistical association*.

Statistics Canada (2009) 2006 Census Technical Report: Sampling and Weighting, Catalogue no. 92-568-X.

Statistics Canada (2010) 2006 Census Technical Report: Coverage, Catalogue no. 92-567-X.

Statistics Canada (2012) Overview of the Census, Census year 2011, Catalogue no. 98-302-XIE.

U.S. census bureau (2009) History: 2000 Census of Population and Housing (Volume 1, 2), PHC-R-V1, PHC-R-V2.

Rothhaas, Cynthia, Frederic Lestina, and Joan M. Hill (2012) 2010 Decennial Census: Item Nonresponse and Imputation Assessment Report, 2010 CENSUS PLANNING MEMORANDA SERIES, No.173.

Williams, J. D. (2011) The 2010 Decennial Census: Background and Issues, Congressional Research Service 7-5700.