匿名データの利用改善に向けた
調査研究報告書

平成 29 年 2 月

一橋大学経済研究所

はじめに

統計の個票データ利用は今や実証研究に欠かせない。質の高い公的統計においては，調査対象者の匿名性を担保しつつ，これを十分に利用できる体制を整備すること，またそのための人材育成をすることは，社会の重要なインフラストラクチャーである。

日本において，平成 19 年の統計法改正まで，公的統計の個票データ利用は厳しく制限されてきた。しかし，公的統計は，その規模，回収率の高さ，偏りの少なさといった質の点で，他の調査に比肩できないものであり，その活用は国民の資産として重要である。こうした中で統計法改正を経て，平成 21 年 4 月の本格施行以降，個人が特定されないように匿名性を高めた上で，個票データを利用できる匿名データが開発され，研究利用に供されるようになった。同時に第三十三条第二号申請により，公的統計の調査票情報を利用するための手続きも整備された。匿名データはより幅広い研究者層が利用すべく，現在では『就業構造基本調査』，『社会生活基本調査』，『全国消費実態調査』，『国民生活基礎調査』，『労働力調査』，『国勢調査』，『住宅・土地統計調査』が提供されている。

こうして匿名データが開発されたのだが，提供開始後の利用実績を見ると，全般に公的統計の二次的利用は停滞している。匿名データに限定すれば，その理由は，厳格な手続きを経ているにもかかわらず，データがやや古いこと，都道府県等の地域情報が開示されていないこと，年齢が 5 歳階級のみでしか開示されていないこと，利用の手続きや時間のハードルが高いことであろう。匿名データにおけるこれらの情報非開示は，匿名性を高める工夫として行われているものである。しかし，匿名性を担保できる範囲で，可能な情報は開示されることが望ましい。さらに，学生や大学院生等が公的統計の利用を学べるような仕組みも作ることが統計利用の裾野を広げるだろう。

コンピュータの発達は統計利用の在り方を大きく変えてきた。1990 年代以前は，国が集計する統計報告書を参照することが，研究者を含めほとんどの人にとっての公的統計の主な利用方法であった。しかし，パーソナルコンピュータの普及と集計・統計解析用ソフトウェアの開発によって，個票データの利用による研究分析が，いまや研究の当たり前の標準になりつつある。個票データを利用することで，新たに統計調査を行うことなく，既存の集計統計とは別の角度から社会課題に迫れる場合が多い。このような統計利用は，研究者に限らず，政府，地方自治体や NPO 等にとっても重要である。また，個票データを用いた計量分析は政策効果の測定，Evidence Based Policy Making には不可欠である。日本の現状をとらえ，処方箋を考えるために，質の高いデータを，最新のものを含めて研究者や大学院生が活発に統計分析できる環境を整備することは，重要な社会のインフラストラクチャーである。しかし，公的統計の利用のハードルは依然として高いために，必ずしもこうした質の高い統計の利用が十分に行われているとは言えない。

本研究会の目的は，日本の公的統計の匿名データの利用が低い現状を鑑み，何が問題なのか，そしてどう具体的に改善が図れるかを検討することである。海外の事例や匿名化の方法

を参考にしつつ，調査客体の匿名性を担保した上で，質の高い大規模データとしてその利用の改善を行うことについて，具体的な方策と可能性を検討する。

お茶の水女子大学教授
永瀬伸子

目次

1. 本報告書の概要 1
2. 匿名データを含む二次的利用の現状と課題....2
2.1. 国内で提供されている公的統計の二次的利用データの位置づけ 2
2.1.1. 提供されているデータの種類....2
2.1.2. データ提供実績....4
2.2. 匿名データの提供開始時期....7
2.3. 現在行われている匿名化措置....8
2.3.1. 調査間で共通の措置....8
2.3.2. リサンプリング....8
2.3.3. 地域情報....9
2.3.4. 世帯削除の基準....10
2.3.5. リコーディング....11
2.3.6. トップコーディング・ボトムコーディング....12
2.4. 匿名データへの意見・要望等....15
2.4.1. 研究者からの意見・要望....15
2.4.2. 過去に行われた利用者へのアンケート結果....16
2.4.3. 提供側からの意見・要望....17
2.5. 今後の二次的利用の動向と匿名データのあるべき姿....18
2.5.1. リモートアクセス型オンサイト利用....18
2.5.2. パブリックユースファイル....19
2.5.3. 匿名データ....20
3. 海外の事例 21
3.1. UNECE による研究目的データの分類（抜粋）....21
3.2. アメリカセンサス局のパブリックユースファイル：
PUMS (Public Use Microdata Samples)....21
3.3. カナダ統計局のパブリックユースファイル：
PUMFs (Public Use Microdata Files)....23
3.4. ミネソタ大学のパブリックユースフィル：
IPUMS-I (Integrated Public Use Microdata Series, International)....25
4. 匿名データの利用改善に向けて....27
4.1. 匿名性の検証における本報告書の用語の定義....27
4.2. 匿名化の前提条件....28

4.3. 利用者が必要とする情報.....31
4.4. 識別情報として扱う項目.....33
4.5. 匿名データ作成のための実証実験.....36
4.5.1. 匿名化技法の検討.....39
5. 提言.....41
5.1. 研究の結果による提言.....41
5.2. ガイドライン改正案.....43
5.2.1. 新旧比較表.....43
5.2.2. ガイドライン別紙改正案.....48
6. まとめと今後の課題.....55
7. 集計結果詳細.....57
7.1. 共通の集計条件.....57
7.1.1. サンプル数.....57
7.1.2. 各変数 集計除外条件.....57
7.1.3. 年齢の集計条件.....57
7.1.4. 地域の集計条件.....57
7.2. 国勢調査調査票情報を用いた基本４情報の組合せ集計.....59
7.3. 基本４情報に他の準識別子を加えた母集団一意の推計.....63
7.3.1. 就業構造基本調査の標本データにおける職業情報の最小度数.....63
7.3.2. 年齢・地域情報・性別・職業の組合せにおける最小度数の推計結果.....65
7.4. 国勢調査調査票情報を用いた基本４情報及び職業による検証.....67
7.4.1. 国勢調査調査票情報による職業別度数.....67
7.4.2. 国勢調査 10%抽出詳細データを用いた性別・地域・年齢・職業中分類の組合せにおける標本一意.....70
7.4.3. 国勢調査調査票情報を用いた性別・地域・年齢・職業大分類の組合せにおける母集団一意.....72
「匿名データの利用改善に向けた研究会」について.....74
付録１ アメリカ統計局のパブリックユースファイルの貸与方法と匿名性のための工夫.....75
付録２ アメリカにおいてパブリックユースファイルが作られているその他の調査の事例.....90

1. 本報告書の概要

新統計法が全面施行された平成 21 年 4 月から約 8 年が経過し，新法において新たな制度として設けられた二次的利用が定着しつつあるが，新たな課題も指摘されている。「公的統計の整備に関する基本的な計画」（平成 26 年 3 月 25 日，閣議決定）において，オーダーメード集計では利用条件の緩和，匿名データでは提供する統計調査の種類の追加等の課題が指摘されている。このうち，世帯系統計調査で作成されている匿名データでは，地域情報等について，より詳細な情報を求めるユーザーの声が出ており，また統計委員会における審議においても同様の意見が出ている。これらを受け，統計委員会では，匿名データにおける地域情報の提供方法等の匿名化手法の検討を匿名データ部会に付議した。

「匿名データの利用改善に向けた研究会（以下研究会という）」は，「匿名データの作成・提供に係るガイドライン（以下ガイドラインという）」の改正に向け，利用者側からの意見等を反映させるとともに，技術的助言を得ることを目的として，平成 28 年 11 月から平成 29 年 2 月にかけて開催された。本報告書は，研究会の内容をまとめ，ガイドラインの改正案を示すものである。

第 2 章では，二次的利用の現状と意見・要望及び今後の動向について述べる。第 3 章では，海外のデータ提供事例について述べる。第 4 章では，匿名データの利用改善に必要な条件の明確化と，その実証分析を行う。第 5 章では，分析の結果に基づき提言を行い，ガイドラインの改正案を示し，第 6 章で今後の課題を示す。

2. 匿名データを含む二次的利用の現状と課題

この章では，匿名データの利用状況とその課題について述べる。次いで，匿名データにおける利用者側・提供側からの問題点を示し，匿名データに関わる二次的利用の今後の動向について述べる。

2.1.国内で提供されている公的統計の二次的利用データの位置づけ

公的統計の二次的利用は，匿名データに限られたものではない。研究者はさまざまな公的統計の利用方法を勘案した上で，匿名データあるいは他の二次的利用の方法を選んでいる。そこで，公的統計の二次的利用として提供されているデータの種類についてまず述べ，データの提供実績を示す。

2.1.1. 提供されているデータの種類

調査票情報

これは統計調査によって集められた情報（統計法第二条第 11 項）について，集められた属性・値が加工されていないデータの利用である。平成 19 年の統計法改正以後，その利用方法が明確となり，統計法第三十三条に基づいて提供される。ここでは，統計法第三十三条第二号による研究者への提供に限定する。

利用者	以下のいずれかの者 ➢ 公的機関が委託又は共同して調査研究を行う者 ➢ 公的機関が公募の方法により補助する調査研究を行う者 ➢ 行政機関等が政策の企画・立案，実施又は評価に有用であると認める統計の作成等を行う者
手数料・データの消去	手数料は不要 データの消去は必要
審査等	研究内容・提供を希望する変数・作成する結果物の概要等詳細な分析計画の提出が必要

オーダーメード集計

統計法第三十四条に基づき，作成する集計表のレイアウトを指定し，調査票情報から統計の作成を依頼するものである。調査票情報そのものの提供を受けなくとも，オーダーメード集計を担当部局に依頼することで，既存の統計表にない統計表を公的統計から得ることができる。ただし，変数の数には制限があり，結果の一部は秘匿される場合がある。

利用者	研究者等
手数料・データの消去	手数料が必要 （主な費用：集計にかかった時間 × 5,900 円）
審査等	平成 28 年から以下の通り緩和 ➢ 企業活動等の一環として行う研究であっても研究意義があり，成果等を公表すれば利用可能 ➢ 法人利用の場合, これまで必要だった法人代表者(社長等)の本人確認は不要

匿名データ

幅広い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データ（統計法第二条第 12 項)。統計法第三十六条に基づいて提供される。属性・値はグループ化，ノイズやスワッピング等の匿名化処理により調査票情報と異なる場合があり，また一部変数やレコードは削除されているが，学術的研究に必要な精度は確保されている。

利用者	学術研究又は高等教育を目的とする者
手数料・データの消去	どちらも必要 （主な費用：申出 1 件につき 1,850 円, 1 ファイルにつき 8,500 円）
審査等	研究概要・分析方法の概要等

一般用ミクロデータ

公開可能な集計表から合成されたデータ。主に教育目的での利用を想定しており，研究目的での利用は想定していない。現在は全国消費実態調査から作成した簡易版(変数が限定)が提供中で，将来は詳細版を提供予定。

利用者	主に大学生の演習等，利用の限定はない
手数料・データの消去	どちらも不要
審査等	なし

2.1.2. データ提供実績

統計法第三十三条第二号申請による調査票情報の提供件数

過去 5 年間における統計法第三十三条第二号申請のうち，統計法施行規則第九条第二号(調査研究)による調査票情報の提供件数は表 2-1 のとおりである。

統計法第三十三条第二号は，科研費等，公的機関が公募の方法により補助する調査研究を行う者への貸与である。ここでは比較のため，匿名データが提供されている調査について提供件数を示す。

最も提供件数が多い調査は国民生活基礎調査であり，年 15～20 件程度となっている。その他の調査の提供件数は年 3 ～ 4 件前後となっている。

提供件数の傾向としては，国民生活基礎調査，就業構造基本調査が比較的利用件数が多く，かつ増加傾向となっている。

表 2-1　統計法第三十三条第二号申請提供件数[1]

調査名	H23	H24	H25	H26	H27
国勢調査	1	1	2	3	4
社会生活基本調査	6	2	3	6	3
就業構造基本調査	4	4	7	9	12
住宅・土地統計調査	2	1	0	3	1
全国消費実態調査	11	6	2	10	3
労働力調査	1	1	5	6	5
国民生活基礎調査	11	14	20	15	23

[1] 注）匿名データを提供している調査に限る。

統計法第三十四条によるオーダーメード集計の利用件数

過去 5 年間における統計法第 34 条によるオーダーメード集計の利用件数は表 2-2 のとおりである。

最も利用件数が多い国勢調査では年 5 ～10 件程度となっている。その他の調査の利用件数は年 1 ～ 2 件前後となっている。

表 2-2　オーダーメード集計利用件数[2]

調査名	H23	H24	H25	H26	H27
国勢調査	2	8	5	9	7
社会生活基本調査	1	0	0	3	1
就業構造基本調査	0	1	2	6	3
住宅・土地統計調査	4	3	2	3	3
全国消費実態調査	1	1	0	0	1
労働力調査	0	3	0	0	0

2 注）匿名データを提供している調査に限る。

統計法第三十六条による匿名データ提供件数及び相談件数

総務省の匿名データは試行的提供を経て，平成 21 年から本格的に提供が開始された。以降では平成 23 年に「平成 18 年社会生活基本調査」の匿名データが追加され，労働力調査の提供も開始された。平成 25 年には「平成 12 年，17 年国勢調査」匿名データの提供が開始され，平成 27 年には「平成 13 年，18 年 社会生活基本調査 調査票 B（生活時間編）」匿名データの提供が開始された。また，厚生労働省「国民生活基礎調査」の匿名データは，平成 23 年に提供が開始された。

平成 21 年以降の提供件数は，表 2-3 に示すとおりである。また，利用者への提供及び利用の検討のための利用相談件数は表 2-4 のとおりである。

表 2-3　匿名データ提供件数

調査名	H21	H22	H23	H24	H25	H26	H27
全国消費実態調査	6	17	12	13	8	14	9
社会生活基本調査	10	9	16	11	15	6	10
就業構造基本調査	7	10	7	5	10	13	11
住宅・土地統計調査	0	6	1	1	3	2	2
労働力調査			0	0	2	2	5
国勢調査					1	4	1
国民生活基礎調査			2	5	8	4	10

表 2-4　匿名データ利用相談件数

調査名	H21	H22	H23	H24	H25	H26	H27
全国消費実態調査	20	92	115	95	82	143	71
社会生活基本調査	28	91	138	73	75	118	58
就業構造基本調査	24	71	82	53	53	151	83
住宅・土地統計調査	6	33	8	14	15	29	20
労働力調査			3	8	13	8	22
国勢調査					4	38	18

2.2.匿名データの提供開始時期

総務省統計局，厚生労働省が提供する匿名データの提供開始時期は，表 2-5 のとおりである。

この表の調査年次と提供開始日を見ると，提供開始時期は調査年から 5 年以上経過している場合が多い。なお，この表に掲載されていない調査年次の匿名データは，調査年から 5 年以上が経過しているにも関わらず，未だ公表されていない。現時点で 5 年以上経過している調査は表 2-6 のとおりである。

表 2-5　匿名データ提供開始時期

省庁	調査名	調査年次	調査区分	提供開始日
総務省	国勢調査	平成12年		平成25年12月27日
		平成17年		平成26年３月28日
	住宅・土地統計調査	平成５年、10年、15年		平成21年４月１日
	全国消費実態調査	平成元年、６年、11年、16年		平成21年４月１日
	就業構造基本調査	平成４年、９年、14年		平成21年４月１日
	社会生活基本調査	平成３年、８年、13年（調査票A）	生活行動編	平成21年４月１日
			生活時間編	
		平成18年（調査票A）	生活行動編	平成23年10月28日
			生活時間編	
		平成13、18年（調査票B）	生活時間編	平成27年7月31日
	労働力調査	平成元年１月～19年12月	基礎調査票	平成23年12月27日
		平成20年１月～20年12月	基礎調査票	平成24年７月20日
		平成21年１月～21年12月	基礎調査票	平成25年10月31日
		平成22年１月～22年12月	基礎調査票	平成26年11月28日
		平成23年１月～23年12月	基礎調査票	平成27年11月30日
		平成24年１月～24年12月	基礎調査票	平成28年11月30日
厚生労働省	国民生活基礎調査	平成10年		平成28年９月
		平成13年		平成24年５月
		平成16年		平成23年９月
		平成19年		平成26年３月
		平成22年		平成27年９月

表 2-6　調査年から５年以上経過した調査

省庁	調査名	調査年次
総務省	国勢調査	平成22年
	住宅・土地統計調査	平成20年
	全国消費実態調査	平成21年
	就業構造基本調査	平成19年
	社会生活基本調査	平成23年

2.3.現在行われている匿名化措置

匿名データの作成に用いられる調査票情報は個人情報に該当するが，統計法第五十二条のとおり，個人情報保護法ではなく統計法が適用される。個人情報保護法は適用されないが，同法の基本法部分を踏まえた厳格な措置は講じる必要がある。

統計法で定義される匿名データは，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。また，匿名データの作成にあたっては，匿名性及び有用性の確保について統計委員会で審議されている。

国勢調査，全国消費実態調査，社会生活基本調査，就業構造基本調査，住宅・土地統計調査，労働力調査及び国民生活基礎調査の直近に作成された匿名データにおける匿名化措置は以下のとおりである。

2.3.1. 調査間で共通の措置

いずれの調査でも，レコードの無作為な並べ替え及び区分の統合（リコーディング）が行われている。

レコードの無作為な並べ替えにより，データの並びがある情報と対応することを防ぐことができる。

リコーディングにより，複数のカテゴリーがより抽象度の高いカテゴリーに統合されることで，識別を防ぐことができる。

2.3.2. リサンプリング

世帯を単位としてまとめた上で，リサンプリングが行われている。リサンプリング率は表 2-7 のとおり，国勢調査では 1%，住宅・土地統計調査では 10%，国民生活基礎調査では 20%であり，その他の調査では 80%である。ただし，労働力調査は，沖縄県の抽出率が 20%であるため，全体の抽出率は 80%より若干低い。

抽出方法は，国勢調査，全国消費実態調査，住宅・土地統計調査及び国民生活基礎調査では確率比例抽出，その他の調査では単純無作為抽出が用いられている。

確率比例抽出の基準は，国勢調査では世帯の種類(一般世帯又は施設等)，住宅・土地統計調査では都道府県別，全国消費実態調査では 3 大都市圏か否かの地域 2 区分，国民生活基礎調査では国勢調査区（又は単位区）と世帯の二段階で抽出されている。

表 2-7　各匿名データにおけるリサンプリング率

調査名	リサンプリング率	リサンプリング方法
国勢調査	1%	確率比例(世帯 2 区分)
全国消費実態調査	80%	確率比例(地域 2 区分)
社会生活基本調査	80%	単純無作為
就業構造基本調査	80%	単純無作為
住宅・土地統計調査	10%	確率比例(都道府県)
労働力調査	80%(沖縄県は 20%)	単純無作為
国民生活基礎調査	20%	確率比例 (地域及び世帯の 2 段階)

2.3.3. 地域情報

表 2-8 のとおり，最も詳細なデータを含むのは国勢調査で，都道府県及び人口 50 万人以上の市区に居住する場合にはその市区名が含まれており，次いで住宅・土地統計調査では都道府県の情報を含む。全国消費実態調査，社会生活基本調査及び就業構造基本調査では 3 大都市圏か否かのみが含まれ，労働力調査及び国民生活基礎調査では地域情報は含まれない。

表 2-8　各匿名データにおける地域情報

調査名	地域情報
国勢調査	都道府県及び人口 50 万人以上の市区名
全国消費実態調査	2 区分(3 大都市圏か否か)
社会生活基本調査	2 区分(3 大都市圏か否か)
就業構造基本調査	2 区分(3 大都市圏か否か)
住宅・土地統計調査	都道府県
労働力調査	なし(全国)
国民生活基礎調査	なし(全国)

2.3.4. 世帯削除の基準

世帯人数が多い場合や，三つ子等の同一年齢の子供が 3 人以上いる等の，小数となる組合せが含まれる世帯は削除されている（表 2-9)。

社会生活基本調査の調査票Ｂについては，国勢調査（母集団情報）において，地域，親の年齢，子供の数及び住宅の所有の関係を組み合わせた結果，小数となる組合せに該当する世帯を削除している。

国勢調査においては，世帯人員が一定以上(地域により 7 ～ 9 人)の世帯，父子世帯(未婚，死別又は離別の男親と，その未婚の 20 歳未満の子供のみからなる世帯)，　年齢差の大きい夫婦のいる世帯(年齢差が 25 歳以上の夫婦のいる世帯)，年齢差の大きい又は小さい親子のいる世帯(年齢差が 55 歳以上の男親と子，45 歳以上の女親と子，14 歳以下の親と長子又は 19 歳以下の親と末子のいる世帯)，世帯主又は配偶者のいずれか一方，若しくは双方が外国人で子供の数が多い世帯(地域により 3 ～ 7 人)が削除されている。さらに，公表統計により母集団一意又は二意であることが判明しているレコードを含む世帯が削除されている。

表 2-9　各匿名データにおける削除基準

調査名	世帯人員	三つ子等の年齢基準	その他
国勢調査	7 ～ 9 人以上	なし	世帯類型， 夫婦年齢差， 親子年齢差， 公表統計における母集団一意又は二意等
全国消費実態調査	8 人以上	15 歳未満	
社会生活基本調査	8 人以上	10 歳未満 (平成 13 年以降)	小数となる組合せ
就業構造基本調査	8 人以上	15 歳未満	特定施設の 調査対象世帯
住宅・土地統計調査	8 人以上 (準世帯を除く)	15 歳未満	家計を主に支える者の年齢が 15 歳未満
労働力調査	8 人以上	15 歳未満の同一年齢階級(平成 14 年以降)	特定施設の調査対象（自衛官，受刑者等）
国民生活基礎調査	8 人以上	同一年齢階級に 4 人以上	世帯類型， 夫婦年齢差， 親子年齢差等

2.3.5. リコーディング

年齢のリコーディング

年齢は，どの調査においても基本的に 5 歳階級で，85 歳以上は一つの区分とされている（表 2-10)。ただし，住宅・土地統計調査の平成 5 年調査では，家計を主に支える者を除いた各世帯員の年齢は含まれておらず，代わりに年齢階級別の世帯員数が含まれる。

末子の年齢の区分は調査年次によって異なり，社会生活基本調査では平成 3 年では 2 区分(6 ～17 歳，18～26 歳)，平成 8 年及び 13 年では 7 区分(0 歳， 1 ～ 2 歳， 3 歳， 4 ～ 5 歳， 6 歳， 7 ～ 8 歳， 9 歳，10 歳以上)，平成 18 年では 6 区分(0 歳， 1 ～ 2 歳， 3 歳から 11 歳まで 3 歳階級，12 歳以上)である。一方，就業構造基本調査では平成 4 年では 18 区分(0 ～14 歳まで各歳 15 歳以上は在学中か否か)，平成 9 年と 14 年では 20 区分（ 0 ～17 歳まで各歳 18 歳以上は在学中か否か)である。

表 2-10　各匿名データにおける年齢のコーディング

調査名	年齢階級	上限	末子の年齢	子供の年齢
国勢調査	5 歳階級	85 歳		
全国消費実態調査	5 歳階級	85 歳	なし	15 歳未満は各歳
社会生活基本調査	5 歳階級	85 歳	２～７区分	10 歳未満は各歳 (平成 13 年以降)
就業構造基本調査	5 歳階級	85 歳	18～20 区分	15 歳未満は各歳
住宅・土地統計調査	5 歳階級	85 歳	なし	15 歳未満は各歳 (平成 10 年以降)
労働力調査	5 歳階級	85 歳	なし	なし
国民生活基礎調査	5 歳階級	90 歳	なし	なし

2.3.6. トップコーディング・ボトムコーディング

国勢調査，労働力調査，全国消費実態調査，住宅・土地統計調査及び国民生活基礎調査においてはトップコーディング・ボトムコーディングが行われている。

国勢調査における閾値は以下のとおりである（表 2-11)。

表 2-11 国勢調査におけるトップ（ボトム）コーディング[3]

項目	上限(トップコーディング)	下限(ボトムコーディング)
建物全体の階数（注）	6 階建以上，11 階建以上， 15 階建以上のいずれか	なし
世帯が住んでいる階（注）	3 階以上，6 階以上，11 階以上， 15 階以上のいずれか	なし
週間就業時間	90 時間以上	なし

労働力調査では，週間就業時間が 90 時間以上の場合，トップコーディングの対象となっている。

全国消費実態調査における閾値は以下のとおりである（表 2-12)。

表 2-12　全国消費実態調査におけるトップ（ボトム）コーディング

項目	二人以上・単身の別	上限(トップコーディング)	下限(ボトムコーディング)
延べ床面積	二人以上の世帯	200 平方メートル	30 平方メートル
	単身世帯	200 平方メートル	なし
うち業務用面積	二人以上の世帯	150 平方メートル	なし
	単身世帯	100 平方メートル	なし
敷地面積	二人以上の世帯	1,000 平方メートル	なし
	単身世帯	1,000 平方メートル	なし
年間収入	二人以上の世帯	2,500 万円	なし
	単身世帯	1,000 万円	なし
貯蓄現在高	二人以上の世帯	9,500 万円	なし
	単身世帯	5,500 万円	なし
負債現在高	二人以上の世帯	4,500 万円	なし
	単身世帯	1,500 万円	なし

3 （注）調査年次や地域によって，閾値は異なる。

住宅・土地統計調査における閾値は以下のとおりである（表 2-13)。

表 2-13　住宅・土地統計調査におけるトップ（ボトム）コーディング

項目	上限(トップコーディング)	下限(ボトムコーディング)
延べ面積	200 平方メートル	30 平方メートル
居住室数	10 室	なし
居住室の畳数	60 畳	9 畳
住宅の敷地面積 （一戸建・長屋建）	700 平方メートル	50 平方メートル
住宅の建築面積 （＝1 階の床面積） （一戸建・長屋建）	150 平方メートル	30 平方メートル
家賃・間代	9 万円	なし
建物の敷地面積 （長屋建・共同住宅）	2,000 平方メートル	100 平方メートル
建物の建築面積 （長屋建・共同住宅）	500 平方メートル	100 平方メートル
階数（一戸建・長屋建）	2 階	なし
階数（共同住宅）	11 階	2 階

国民生活基礎調査における閾値は以下のとおりである（表 2-14)。

表 2-14　国民生活基礎調査におけるトップ（ボトム）コーディング

項目	二人以上・単身の別	上限(トップコーディング)	下限(ボトムコーディング)
家計支出額 総額	二人以上世帯	100 万円	なし
	単独世帯	55 万円	なし
育児費用 総額		7 万円	なし
仕送り額		6 万円	なし
学業仕送り額		16 万円	なし
室数		10 室	なし
床面積		300 平方メートル	20 平方メートル
総所得	二人以上世帯	2,200 万円	なし
	単独世帯	1,100 万円	なし
拠出金合計(税金＋社会保険料)	二人以上世帯	490 万円	なし
	単独世帯	250 万円	なし
企業年金・個人年金等掛金	二人以上世帯	80 万円	なし
	単独世帯	40 万円	なし
貯蓄現在高	二人以上世帯	9,000 万円	なし
	単独世帯	6,300 万円	なし
貯蓄減少額	二人以上世帯	1,300 万円	なし
	単独世帯	800 万円	なし
借入金額	二人以上世帯	4,000 万円	なし
	単独世帯	2,400 万円	なし
支払額	二人以上世帯	18 万円	なし
	単独世帯	6 万円	なし
一週間の就業時間		80 時間	なし
就業期間		50 年	なし
普段の活動ができなかった日数		25 日	なし

2.4.匿名データへの意見・要望等

2.4.1. 研究者からの意見・要望

匿名データについて，相談件数に比べて実際の利用者数が少ない理由として以下の理由が想定される。

1. 申請からデータ入手までの手続きの手間がかかること

申請後にデータ入手まで２～３か月かかる。事前に研究・分析までを含めた申請しなければならないが，データを見ないとわからないことも少なくない。しかし，申請書類には分析の手法，公表予定や公表先等，事前に予想して記入しなければならず，容易には作成できない部分がある。

データの利用場所は，見取り図等を出す必要がある。独立した分析スペースの申請やインターネットにつながないコンピュータの申請が必要である等設備面の対応が必要となる。このように，調査票情報の申請と同様に，厳格な条件が課されることが利用のハードルを上げている。

2. 論文修正の際に再び申請が必要となること

研究者にとっては論文発表が重要であるが，匿名データは利用期間とその期間終了後のデータや中間生成物の消去があらかじめ決められている。しかし，論文投稿から掲載に至るまでに２年ほどかかることもあり得る。投稿過程でレフェリーから修正を提案された場合，再び上記の時間のかかる上記の申請過程を経ないとデータを入手し論文修正ができない。さらに，その際にデータを借りるにも時間と手間がかかり，論文の出版が遅れてしまう。投稿から審査に至る期間については，匿名性を高めたデータが提供され，データの保管が認められればこうした困難は少なくなる。

3. 情報の粗さ

厳密な申請過程を経るにもかかわらず，科研費等を通じた三十三条第二号申請と比べると，地域情報や年齢情報が粗く，分析内容が制限される。三十三条第二号申請はデータ読み込みの手間等から，匿名データよりは使い勝手が悪いものの，どうしても公的統計を使って分析したいとなれば地域情報等に制約のない三十三条第二号申請も検討されるだろう。

4. 利用経験の少なさ

匿名データを一度利用すれば，容易に入手できる民間調査データを利用する場合に比べて，データの偏りの少なさ，データ数の大きさ等，公的統計のメリットと魅力がわかる。しかし，公的統計利用の経験がないために，このような手間をかけてまで匿名データを借りる

ことの意義がわかりにくい。

日本の匿名データとパブリックユースファイル

日本の匿名データに行われている匿名化処理は，例えばアメリカの国勢調査のパブリックユースファイルに比べると，地域情報がほとんど開示されていない等，パブリックユースファイル以上に情報が非開示となっている場合も少なくない[4]。また，トップコーディング等については，パブリックユースファイルと同程度の匿名化が行われている。

しかし，アメリカのパブリックユースファイルはだれもが（海外にいる個人を含めて），簡単にインターネット上からダウンロードできる。このためそのアクセスはきわめて容易で，大勢によって利用されている。これに対して，日本の匿名データは，ほぼ同程度の匿名化の工夫が行われているが，本当にそれほど厳格に運用すべきかわからないほど厳格な運用が行われている。このために利用者は非常に少ない。

データの匿名化や厳密な利用手続きは，調査対象を特定させない工夫として非常に重要な手続きである。しかし，手続きの厳密さ等は，データに含まれるセンシティブ情報が開示される可能性にさらされる可能性の不利益と，統計利用を容易にすることで得られる利益の大きさとの比較の度合によって決まるべきである。つまり，すべての情報を厳格に管理するのではなく，（統計分析には影響を与えない程度に）スワッピング等の一定の加工等を行った上で容易にアクセス可能なミクロデータと，詳細がわかるがその利用手続きも利用場所も厳密に制御されるようなミクロデータというように，手続きに差を設けるべきではないか。

2.4.2. 過去に行われた利用者へのアンケート結果

独立行政法人統計センターが行ったアンケート調査から，利用者からの意見・要望の要約は以下のとおりである。

- 地域区分の情報が含まれないことで，学術誌に投稿してもレフェリーからのコメントに対応できず結果として掲載されなかった。（就業構造基本調査利用者）
- 比推定用乗率だけではなく，本来の線形推定用乗率や抽出率の提供を希望。（全国消費実態調査・社会生活基本調査利用者）
- 都道府県レベルによる地域区分の提供を希望。（就業構造基本調査利用者）

4 日本の国勢調査は 50 万人規模で地域情報が開示されているとはいえ，アメリカの国勢調査と比較すると，収入情報，住宅情報などがないという点ではアメリカの国勢調査とは比較しにくい。提供されている情報量でいえば，アメリカの国勢調査は日本の就業構造基本統計調査に近いものがある。その就業構造基本統計調査では 3 大都市圏以上には地域情報の開示がない。このように日本の匿名データはかなり情報の匿名性の管理には厳格である。

- 大都市とその他よりも詳細な地域区分の提供を希望。（全国消費実態調査・社会生活基本調査利用者）

2.4.3. 提供側からの意見・要望

独立行政法人統計センターのサテライト機関である，一橋大学・神戸大学の匿名データ提供業務の担当者に意見・要望のヒアリングを行った。意見・要望等は以下のとおりである。

匿名データ申請時の要望について

- 海外からの申出を簡略化してほしい。日本人の研究代表者，または海外の利用者が属する組織が責任を持てばよいのではないか。

匿名データ利用中の要望について

- 所属等変更届出書は所属だけ変わる場合にのみ使われるもので，利用場所の変更等所属以外が変わる場合は記載事項変更になるが，利用者にはご理解いただけない。所属等変更届を廃止し，記載事項変更届に統一してもよいのではないか。

利用延長時の要望について

- 利用期間延長依頼申出書は様式だけ存在しているが，利用期間の延長は研究期間や公表予定の変更が必要なため，この様式が使われることはありえない。様式を廃止したほうがよいのではないか。

その他

- 申出書やメールのやりとり等，申出者の個人情報はデータの提供手続きが終了次第削除しているが，再度の申込や提供実績のまとめ，匿名データ利用者による講演会の開催を検討する際に支障があり，安全に保管すればよいようにしてほしい。
- 匿名データ複製の際に，入力チェックに大変時間がかかるため，改善できないか。一橋大学・神戸大学どちらでも時間がかかるため，コンピュータ等の環境の問題ではない。
- 匿名データの直接受取に来られた方が，出張の証明を求められることがある。独自のデータ受渡書を渡しているが，統一の書式が必要ではないか。
- 郵送提供の料金は実費を後払いにできないか。仮申出の際に一括で見積もると，郵送する際に重量制限で苦労することがある。

2.5.今後の二次的利用の動向と匿名データのあるべき姿

今後の匿名データのあるべき姿を，調査票情報やパブリックユースファイル等，その他の二次的利用に関わる動向と共に整理する。

2.5.1. リモートアクセス型オンサイト利用

リモートアクセスによる調査票情報の利用と匿名データ

現在，リモートアクセス型オンサイト利用によって，調査票情報の利用拡大が推進されている。どのような形で進むかはまだ明確ではないが，リモートアクセス型オンサイトのメリット・デメリットによっては，調査票情報の利用者数が増減する可能性もある。調査票情報の変わりに匿名データを利用する，または調査票情報の利用が容易となったので匿名データを利用しないといったように，リモートアクセス型オンサイトの動向は匿名データの利用者数にも影響し得る。

リモートアクセス型オンサイト利用の概要：

メリット 利用者 1. 利用前審査の短縮・簡易化 2. 利用できる調査事項の増加（標準的な調査事項を利用可） 管理・提供者 1. 事前申請に係る審査事務の負担軽減 2. ファイルの保管について厳格に管理可能
デメリット 1. 利用可能な時間や場所が限定される 2. 集計結果や集計に利用したプログラムを持ち出すには審査が必要

＊ リモートアクセス型オンサイト利用は，上記デメリットから利用者にとっては使いづらいデータの提供方法である。リモートアクセス型オンサイト利用による調査票情報の利用は，匿名化すべきリスクの高い情報がある場合に限定すべきであろう（海外でも，オンサイトはデータ利用のハードルをあげることから，健康情報や税務情報等リスクの高い情報が含まれる場合に限定されている。ただし，フィンランド等オンサイトではなく，研究室や自宅から利用可能な例もある）。その場合は，匿名データの地域区分や年齢区分を詳細化し，匿名データを用いて詳細な研究ができるようにすべきである。

2.5.2. パブリックユースファイル

諸外国の世帯統計調査と比べて，日本の公的統計における二次的利用の申請手続きは複雑であり，簡単には使えない問題があることから，パブリックユースできるミクロデータの開発は急務である。現在日本では一般用ミクロデータと呼ばれているファイルが提供されているが，作成方法が限られることからデータの有用性について懸念する声がある。センシティブ情報が少ない公的統計調査については，個人の匿名性を高めた上で，研究者がより容易に使うことができる提供の条件を探る必要がある。

パブリックユースファイルに求められる役割

1. 簡単な手続きで入手できるが，公的統計の持つ分布精度は備えている統計であること
2. 調査対象については絶対に識別されないように匿名化されていること
3. 利用データを消去する必要がないように匿名化されていること
4. 日本全国を反映した個票データの利用によりエビデンスに基づく研究を活発にすること

パブリックユースファイルの利用・作成に関する問題

- データの信頼性

現在提供されている「一般用ミクロデータ」は日本のパブリックユースファイルといえるかもしれない。これは教育用として合成作成された変数であり，簡単に誰もが利用できる。しかし，これは合成された変数なので，教育には用いられても，研究・分析には堪えない。パブリックユースファイルであったとしても，個票データとして，公的統計の持つ精度を正しく反映したものでないと分析に堪えないため，信頼性のあるデータであることが必要である。

- 作成可能な方法が限られている

日本では，調査票情報を直接用いたデータにおいては，例えばトップコーディング等の情報匿名化を行っているとしても，アメリカセンサス局のパブリックユースファイルのような扱いはできないという見方がある。調査票を用いたデータは「匿名データ」という位置づけとなり，厳密な申請や管理の手続きが統計法から要請されるという見解がある。

2.5.3. 匿名データ

現在の匿名データに求められている役割

1. 幅広く研究者，学生等が利用できる。
2. 調査票情報そのものに比べて，匿名化すべきリスクの高低に応じて，利用申請等の手間がかからない。
3. 利用目的等，事前の申請による審査手続きを通じて，信頼できる利用者に提供されるデータである。
4. 調査票情報の情報とほぼ同様の精度を持って研究・分析に用いることができる。
5. 公益に資する分析に用いることができる。
6. 公開されている統計表に比べ，詳細な情報が取得できる。
7. 論文審査等限られた用途の場合，必要な部分に限りデータの再利用ができる。

今後の匿名データの方向性の例

1. 調査票情報の利用のためのステップ（大学院生・若手研究者用）
2. 社会人学生等，オンサイト利用ができない人が，夜間を含めて研究，分析をするための詳細なデータ
3. SPSS や STATA 等の統計解析ソフトウェアですぐ読み込めるデータ形式で，利用しやすいデータの提供
4. 調査年次による変数の差の調整等，扱いやすいデータを整えた上での提供

現在の匿名データの利用・作成に関する問題

1. 分析に詳細な区分が必要な場合や，分析対象のサンプルサイズが小さい場合
   例：市町村レベルの地域情報，外国人世帯，人数の多い世帯
2. 利用申請は短縮されるべきだが，地域や年齢の詳細化は必要

3. 海外の事例

匿名データの利用改善の参考とするため，海外において提供されている匿名データの分類をまず示す。次いで，アメリカ，カナダのパブリックユースファイルはどのようなものがあるか，またそのデータ匿名化の方法について示す。最後に，ミネソタ大学人口センターが中心となって運営されている，IPUMS-I という世界の個票データの提供サイトを事例として示す。

3.1.UNECE[5]による研究目的データの分類（抜粋）[6]

パブリックユースファイルとは

名前やアドレスの削除だけではなく，識別情報を粗くする等識別を不可能にし，公共の利用を目的として公開されたデータ。

インターネットを通じて提供される場合でも，守秘義務の誓約を必要とする等，リスク管理の措置を取ることは許容される。

ライセンス制ファイル

匿名化されたデータであるが，パブリックユースファイルとは異なり，事前に審査され承諾された研究者等に限定して提供されるデータ。他のデータとのリンク等の手法による潜在的に識別可能なデータを含むことができる。

日本の匿名データはライセンス制ファイルに該当する。

3.2.アメリカセンサス局のパブリックユースファイル：PUMS (Public Use Microdata Samples)

サービス概要

1. 利用目的は限定しない
2. インターネットからダウンロードして利用

5 国連欧州経済委員会（United Nations Economic Commission for Europe）

6 http://www1.unece.org/stat/platform/display/confid/V.+Methods+of+supporting+the+research+community （平成 29 年 1 月 31 日アクセス）

3. 複数の統計解析ソフト用のファイルあり
4. 国外からの使用可
5. 無料
6. 利用後の報告義務なし，公表義務なし
7. 利用規約への署名等は不要
8. 利用後の消去・返却は不要
9. ライセンス制ファイルの利用申請には，パブリックユースファイルでは用を成さない理由を添える必要あり

データ概要

1. データ品質よりも匿名性を優先し，いかなる手法を用いても個人特定に至らないレベルで匿名化している。
   （これに対してアクセスが制限される匿名データは，統計知識がなければ個人の特定はできないレベルで匿名化。）
2. 合成データ，ノイズの付加等による匿名化。スワッピングは限られたデータのみに適用。
3. 地域情報の匿名化を重視しており，一定の基準以下の母集団となった地域は自動検出し対応する。
4. 産業・職業分類は 3 桁符号。
5. １％抽出サンプルの場合，地域情報は 40 万人以上を地域単位とする。5%抽出サンプルでは，10 万人以上を地域単位とする。

アメリカセンサス局がパブリックユースファイル（PUMS）として公開している調査の例は以下のとおり[7]

- 2000 年センサス
- 1990 年センサス
- 1980 年センサス
- アメリカ社会調査（American Community Survey）
- 現在人口調査（Current Population Survey）―労働力調査にあたる
- 所得・事業参加調査（Survey of Income and Program Participation）

＊ 1970 年以前のセンサスについては国立公文書記録管理局が公開

上記のうち，2000 年センサスのパブリックユースファイルで行われている匿名化措置とデータ内容については付録 1 に示した。付録のとおり，40 万人以上の地域単位での開示，

7 https://www.census.gov/main/www/pums.html （平成 29 年 1 月 31 日アクセス）
https://dataferrett.census.gov/AboutDatasets.html （平成 29 年 1 月 31 日アクセス）

世帯のスワッピングや年齢の 1 歳程度の誤差の部分的導入のかく乱等一定の保護措置が行われた上であるが，カテゴリーデータでは，全米で 1 万人以上いれば開示するといった，産業，職業，収入等を含めて非常に詳細に内容が開示されている。また，他の調査についても付録 2 に説明を付した。

3.3.カナダ統計局のパブリックユースファイル：PUMFs（Public Use Microdata Files)

サービス概要

1. 研究・教育目的に限定
2. パスワード保護された FTP からダウンロードして利用
3. 複数の統計解析ソフト用のファイルあり
4. 国外からの使用可
5. 有料。ただし，協定を締結した大学の関係者は無料
6. 利用後の報告義務なし，公表義務なし
7. 適正な管理維持のため，利用規約への署名等が必要。個人への罰則あり
8. 利用後の消去・返却は不要
9. カナダは集中型統計機構のため，統計局が全ての調査を行っており，公開される PUMF も多い。PUMF を公開している調査の例は以下のとおり[8]
   - 国勢調査（1971 年以降）
   - 総合的社会調査（General Social Survey）
   - National Graduates Survey
   - Food Expenditure Survey
   - National Graduates Survey

10. ライセンス制ファイルの利用申請には，パブリックユースファイルでは用を成さない理由を添える必要あり

データ概要

1. データ品質よりも匿名性を優先し，いかなる手法を用いても個人特定に至らないレベルで匿名化している。
   （アクセスが制限される匿名データは，統計知識がなければ個人の特定はできないレ

8 http://www.statcan.gc.ca/pub/11-625-x/2010000/collection-eng.htm
（平成 29 年 1 月 31 日アクセス）

ベルで匿名化。）
2. 外れ値の除去，グループ化，ランダム丸め，トップコーディング等による匿名化。スワッピングは行わない。
3. 地域情報の匿名化を重視しており，一定の基準以下の母集団となった地域は自動検出し対応する。
4. 産業・職業分類は 2 桁符号。

PUMFs の公開基準について

カナダ統計法について

カナダの統計法では，第 17 条にて守秘義務が規定されており，カナダ統計局の職員以外による調査票情報の利用と，特定の個人，企業，組織を識別可能な情報の漏洩が禁止されている。カナダ統計局はこれに基づき，調査対象を識別できないデータを公開している。

PUMFs について

PUMFs は調査票情報に基づいて作成される。PUMFs は機密性が優先され，広く配布するために有用性を低くしている。公開に先立つ修正により，信頼性に違反するリスクが取り除かれていることが保証されており，分析結果の公表前にいかなる分析結果も吟味する必要がないため，”Public”とみなされる。一定の分析ニーズを満たす傾向情報は保持しているが，より詳細な分析を行いたい場合は，別の詳細データの利用が推奨される。

PUMFs の利用については，カナダ統計局が協定を締結した教育機関の関係者は無料で利用できるが，個人や企業向けには，年間 5,000 カナダドルの費用が請求される。

ミクロデータ提供ポリシー

カナダ統計局は，ミクロデータに関する方針として，以下の二点を満たす場合に一般利用のためのミクロデータの公開を許可すると規定した。

1. 公開により，調査データの分析価値が大幅に向上すること
2. 調査単位の特定を妨げるために妥当な措置が取られていることが確認できること

そこで，以下の基準により公開されるミクロデータの審査を行う組織「ミクロデータ提供委員会（Microdata Release Committee）」が設置された。

1. 名前，住所，識別番号等の明示的な識別子はすべてファイルから削除する必要がある。
2. 調査単位を特定すると合理的に予想される値，特性，固有の組合せは，抑制されなければならない。
3. 変数は，非常に小さな集団の性質を他の変数と組み合わせることで，それ自体が準識別子となって特定されないよう，吟味されなければならない。
4. 以下の場合には，２つ以上のバージョンのパブリックユースファイルが公開され得る：
   A） それぞれのバージョンは，相互に排他的な調査単位の集合に基づく
   B） Microdata Release Committee によるレビューにより，全公開データの全変数の合計が機密性を危険に晒すとはみなされない場合

これらの基準に基づき，PUMFs の匿名化は行われており，完全に個人を特定することができないと確認されたうえで公開されている。

参考資料

神林龍（2007），「北米における政府統計個票調査公開の現状に関する調査報告: アメリカ労働統計局，アメリカセンサス局およびカナダ統計局のオンサイトリサーチを中心に」，http://hermes-ir.lib.hit-u.ac.jp/rs/handle/10086/13576 （平成 29 年 1 月 31 日アクセス）

3.4.ミネソタ大学のパブリックユースフィル：IPUMS-I (Integrated Public Use Microdata Series, International)

サービス概要

1. ミネソタ大学人口センターが中心となって運営
2. 研究・教育目的に限定
3. パスワード保護されているホームページ上からダウンロード
4. 複数の統計解析ソフト用のファイルあり
5. 国外からの使用を認める
6. 利用料は無料
7. 利用後の消去・返却は不要
8. さまざまなデータ間で共通変数が作られている等の使い勝手の良いデータを提供

データ概要

1. 2017 年 1 月末において 85 ヶ国 301 の人口センサスミクロデータを公開
2. 国際比較・時点の比較が容易に可能なように，データ及びコードは標準化済
3. 一定の基準に加えて，データ提供元の要請にあわせて匿名化処理

提供データについて

データの匿名化については，データ提供者がすでに実施済みであったとしても重ねて行われる。地理上のエリアについての参照情報には，データ提供者の指示に従い制限をかける。

必要に応じて量的変数はグループ化が行われ，間接的な識別子は削除又はトップ／ボトムコード化されることがある。ある種の機密扱いの変数は，除外又は間接的な識別子として取り扱われることがある。また，より細かいレベルの地理単位を入れ替える世帯のスワッピングも適用されることがある。

提供されている標本は，主に 5 %から 10%の範囲である。また，完全なデータセットは提供されず，必要な変数や条件により抽出したデータが提供される。

ミクロデータ提供ポリシー

電子申請フォームは，申請者の善意ならびに提案される研究に対するミクロデータの適切さを確かめるような構成になっている。不正な申請には厳しい警告が行われ，申請者の身元および加入情報を検証するためのチェックが行われる。

譲渡・供与の禁止，学術用途への限定，商用利用の禁止，機密性に関する厳重な規則，データの安全保護，適切な引用，データにおけるエラーの通知等の使用制限事項に同意が必要となる。

参考資料

MCCAA, Robert, et al. IPUMS-International high precision population census microdata samples: balancing the privacy-quality tradeoff by means of restricted access extracts. In: International Conference on Privacy in Statistical Databases. Springer Berlin Heidelberg, 2006. p. 375-382.

4. 匿名データの利用改善に向けて

研究会における議論の中で，匿名化・利用者・リスクについて前提条件を置くべきこと，識別情報として扱う項目を限定すべきことが確認され，これらの条件に基づいて国勢調査調査票情報を用いた実証分析を行うとの結論が得られた。この章では，匿名化の条件を示す。また，利用者が必要とする情報の要望を示す。これらに基づき，識別情報を整理し，匿名データ作成のための実証実験を行う。

4.1.匿名性の検証における本報告書の用語の定義

外観識別情報

比較的容易に入手できる識別情報であり，外観からでも把握できるような基本的な属性を指す。例としては，都道府県，市町村等の地域情報や，世帯員数，世帯員の性別，住宅の大きさ等が挙げられる。

また，情報源が他者の場合を含め，日常生活において入手が可能である情報であれば外観識別情報に該当する。例としては，自宅で営業している世帯であればその産業・職業を知ることができ，子供の年齢は通学している学年で分かる。

外観識別情報では，性別等の質的属性だけでなく，年齢や住居の面積等の量的属性も，階級区分レベルでは識別が可能とする。

識別

匿名データにおける「識別」とは，利用者が既に知っている客体を対象とし，対象について特別の調査をすることなく入手が可能な情報（変数）の組合せが母集団で一意となる場合である。

準識別子

変数単体では客体を識別することはできないが，他の変数と組み合わせることにより，識別に用いられる変数を指す。

準識別子として扱う変数は，利用者や識別の動機に合わせて検討する。

例として，パブリックユースファイルでは，外観識別情報が準識別子に該当する。

基本 4 情報

個人情報の基本となる氏名，性別，住所，生年月日の 4 つに関する情報を指す。

個人の特定において最も重要となる基本的な情報の種類を表した概念であり，個人情報とその保護に関する制度や議論に用いられる。

自治体が作成・管理している住民基本台帳に記録されている個人についての基本的な情報で，住民票等で閲覧及び証明をすることができる。

なお，氏名は単一で客体を識別可能な場合があり，また分析には一般的に必要がないことから匿名データには含まれない。

母集団一意

国勢調査において，準識別子の組合せが一意となる場合を指す。

準識別子は基本 4 情報に加え，どの変数・組合せが該当するか調査や年次ごとに決定する。

母集団一意となる場合，提供においてはリコーディング等の非撹乱的手法により，母集団一意とならないようにする，またはノイズ付与等の撹乱的手法を適用するといった，匿名化処理を施す必要があるものとする。

4.2.匿名化の前提条件

研究会における議論の結果，匿名データ利用者，匿名データ提供申出における審査や誓約書提出の義務付け等から，匿名化の議論の際は以下の前提条件を置くべきことが確認された。ここでは，想定される違反や，利用者に鑑み，この前提条件の妥当性を示す。

前提条件

1. 匿名化については，完全な匿名化[9]は考慮しない。
2. 匿名データの利用者は，個人・団体等の識別を意識的に行わない。
3. 匿名データ漏洩のリスクは，調査票情報の取り扱いに準ずる。

9 完全な匿名化とは，いかなる手段・外部情報を用いても識別されない状態を指す。

想定される違反

匿名データの提供における違反について，対象者と原因及び内容として以下が想定される。

対象者	原因	内容
有資格者 (研究等を目的とする利用者)	故意	(1) データの目的外利用，譲渡，複製等
		(2) 客体の識別，情報の漏洩
	過失	(3) データの流出
		(4) 中間生成物の保存
		(5) 匿名データの紛失
	偶然 (違反ではない)	(6) 客体の識別
無資格者 (上記以外)	故意	(7) 成りすまし，虚偽の申請

(1)から(5)に関しては申請時に確認や誓約が求められる。(7)は申請時に身分証明書の提示が求められることから，違反することは困難である。なお，違反した場合は統計法第六十一条により罰則が定められている。

ただし，(6)の客体識別は偶然起こる可能性があるため，違反とはならない。

想定されている利用者

匿名データの提供は，統計法第三十六条で規定されており，学術・教育を目的とする利用者に限られている。なお，前提条件を満たす利用者への提供に限定するため，匿名データの利用手続きにおいては以下の措置が行われている。

1. 利用目的を，学術研究の発展や高等教育の発展に資するものに限定
2. 利用目的等を事前の申請により審査
3. 利用者には統計法により適正な管理の義務と守秘義務が課せられ，自己又は第三者の不正な利益を目的として不正使用した場合は，統計法により罰則が適用
4. 利用者は提供を受ける際に，身分証明書を提示し，個人・団体等を特定しようとする試みは行わないこと等を明記した誓約書を提出

匿名データは，個人・団体等が識別されないことを条件として作成されているが，上記の前提条件を満たす利用者に限定して提供しているため，「識別されない」という部分については，誰がどのように利用した場合においても客体を識別することができない「絶対的な匿

名性」ではなく，故意に識別を行わない利用者に対して，識別の可能性が十分に低く保たれている「事実上の匿名性」であるとしてよいものとする。

識別に用いられる項目の分類

匿名データは，統計法第二条第 12 項で，「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をいう。」と定義されている。

ここでは，「特定の個人又は法人その他の団体の識別」とは，識別子による特定を指し，「他の情報との照合による識別」は，準識別子の組合せによる特定を指すものとする。識別子と準識別子は，例としてデータが個人に関するものである場合，識別子は名前や住所等，個人を直接特定できる属性を指す。準識別子は性別，年齢や職業等，複数の準識別子を組み合わせて個人を特定し得る属性を指す。

母集団一意について

実務上の識別の基準には母集団一意が用いられている。しかし，準識別子の組合せが，母集団一意である場合でも，「特別の調査をすれば入手し得るかもしれないような情報は，基本的に『他の情報』に含めて考える必要はない。」（総務省政策統括官〔統計基準担当〕「逐条解説統計法」， 2009， P.60）ことから，現地調査等特別な調査を行わなければ対応情報が特定できない場合は，「他の情報との照合による識別」にはあたらないものとする。

区分の詳細化に伴い想定されるリスク

提供される匿名データは数年前の調査の結果であり，その当時個々の調査対象がどのような属性を有していたか知ることは，たとえ世帯の基本的な属性であっても難しい。例として年齢の区分を各歳にした場合，現在の正確な年齢を知っており，かつ調査日と誕生日についても考慮しなければ，正確に 1 歳単位で絞り込むことはできない。そのため実際には 3 歳程度の範囲でしか絞り込むことができない，あるいは対応関係に誤りが生じることになるため，識別につながるリスクは見かけ上よりも小さい。

多くの標準的な客体の場合には，識別の可能性は比較的低いものにとどまり，一部の客体についてだけ識別できたとしても，数年前の統計情報でもあり，商業目的での利用価値は低い。また，個人情報の流出の際に，特に問題となるような，住所，メールアドレス，氏名，クレジットカード番号，利用者 ID とパスワード等の商業目的にそのまま利用できるような情報は匿名データには含まれない。したがって，対象を識別しようとする試みが，商業目的で行われる可能性は低い。

これは，対応関係の特定を試みるのに必要なコストとしては，まず匿名データの利用申請にかかる時間及び手数料，データの入手後にはデータの読み込みや条件による候補の絞込

み等のデータ操作が必要であり，さらに識別には多くの情報が不可欠であることから，情報の収集・整理を行う必要がある。また，違反が発覚した場合には統計法によって罰則が定められており，また所属機関による罰則を受けるリスクもある。

このように匿名データより特定の客体に関する情報を入手するには，得られる情報に対してコストが高く，居住・所在地への訪問や SNS の閲覧等，他の手段も存在することから，匿名データを用いて故意に識別が試みられる可能性は低い。

これらのことから，匿名データの利用においては法律上及び実務上のどちらにおいても，利用者を制限することにより識別のリスクは低減されている。また，罰則も定められていることから各種の違反は想定する必要はなく，一般的な利用者が分析・集計中に偶然に客体を識別する場合に限定される。そのため，匿名データにおける「識別」とは，利用者がすでに知っている客体を対象とし，対象について特別の調査をすることなく入手が可能な準識別子の組合せが一意となる場合である。

この識別の定義によって定められたリスクを下回る範囲において，地域情報等の項目を詳細にして提供することにより，事実上の匿名性を満たしながらデータの有用性を高めることが可能である。

4.3.利用者が必要とする情報

匿名データの利用者として，ここでは研究者を想定する。研究者が必要とする情報は，例えば労働統計の場合，就業状態，家族構成，雇用形態，勤続年数，勤め先属性，配偶者をはじめとする家族の属性等多岐にわたる。家計消費系の調査や，住宅系の調査ではそれぞれ利用する項目が変わる。また，調査票にある情報をいかに組み合わせて興味深い学術上の疑問に答えるかは，それ自体が独創性を要する。そのため，利用者が必要とする情報を事前に一般的な形で提示することは極めて難しい。ここでは，現在の匿名データでは提供されていない情報の中で，利用者が必要とする情報に焦点を当てて，利用者が必要とする情報を記す。就業構造基本調査を用いて例示を行うが，あくまでも例示であり，これら以外の情報が必要とされることもあり得る点に留意されたい。

ターゲットとする分析

利用者が分析において被説明変数とする変数は，

- 個人の就業状態
- 子供の有無・子供の数
- 家計の消費水準
- 住居の形態

等がありうる。これらを通じて，高齢者の就業，非正社員としての就業，引退，離職，生活水準，借家と持ち家の選択行動等様々な家計行動の決定要因が分析される。

キーとなる属性

上で例示した被説明変数の決定要因となりうるのは,

- 産業
- 職業
- 地域
- 性別
- 年齢
- 家族構成

等である。

必要なサンプルサイズ

年齢や地域に依存して各個人・世帯の受ける政策が異なることがある。例えば細かな年齢に応じて年金の受給資格は変化し，市町村によって保育所の利用可能性は大きく異なる。研究者は，公的統計の個票データに年齢や地域によって異なる政策介入を年齢や地域情報を用いて結合し，政策変数を説明変数とし個票データに含まれる諸変数を被説明変数とする回帰分析を行い，政策効果を推定している。また，そのような推定に当たっては，これらの政策の効果が世帯の属性等によって異なる，政策効果の異質性にも十分な注意が払われている。これらの点を勘案すると大きなサンプルサイズが必要となることが多いが，公的統計はその要求を満たしている場合が多い。匿名データにおいても公的統計の大規模サンプルのメリットが十分に生かされることが必要である。

簡略化・詳細化

上述の通り，研究者が必要としており現在の匿名データが提供していない情報として各歳の年齢情報と地域情報が挙げられる。少なくとも各歳情報と都道府県情報は有意義な研究の遂行のため必要である。さらに，都道府県情報に加えて都市部か地方部かの情報が付与されていると望ましい。また，市町村情報も貴重な情報であり，都道府県内市町村スワッピング等を用いて匿名性を保ちつつ開示されることが望ましい。

匿名性を保つ観点から，個人単位で抽出する場合，他の世帯員のレコードとの照合は不能となり匿名性は保たれやすくなる。ただし，その場合でも，分析では家族変数が不可欠である。つまり，世帯員情報は状況に応じてまるめることが必要である。例えば夫の年齢，子供の年齢と人数は必要だが，子供が 4 人以上は末子のみの年齢とし，子供 4 人以上とする等。世帯主の親についても同居していれば世帯主の母親同居，世帯主の父親同居，といった同居世帯員ダミー及びその年齢や就業状態，労働時間，収入等をつけて提供するというのは一案である。

データ抽出の際の除外条件

特に除外できるような条件はない。匿名化にあたり必要な場合には，レコードを除外するのではなく，グループ化によって対応したほうがよい。例えば，高齢者をサンプルから除外するのではなく，年齢のトップコーディングを行うといった対応が望ましい。

利用頻度・重要性の低い項目(変数)

あくまでも例示であるが，就業構造基本調査の場合，従な収入，副業に関しては詳細まで使う研究者は少ない。これらの情報は統計法三十三条に基づく利用申請を行うと整理するのも一案である。

4.4.識別情報として扱う項目

前提とする条件

利用者が「信頼できる」研究者に限定されることから，先述のとおり，一般的な利用者が分析・集計中に偶然に客体を識別してしまう状況を考える。すなわち，利用者は故意に識別を行わず，特別な調査によって入手し得る情報は知らないことを前提とする。その場合，外観識別情報のように容易に入手できる変数についての考慮が必要となる。

匿名データの利用者

匿名データでは，4.2 のとおり下の条件を前提とすることができることから，信頼できる利用者とみなすことができる。

1. 利用者が限られている
2. 故意に識別が試みられない
3. 情報漏洩は発生しない

利用者の分類

準識別子として扱う項目を検討するために，利用者が統計調査以外から知り得る情報に関して以下の分類を行った。

① **特定個人の基本４情報を知ることができる** 想定される利用者攻撃者：不特定多数すべて
② **役所や一般的なサービス事業者が保有する情報を知ることができる** 想定される利用者攻撃者：事業者
③ **特定個人の見た目で分かる身体的特徴や，本人との日常の会話等から得られる情報を知ることができる** 想定される利用者攻撃者：知人・友人
④ **特定個人の見た目で分かる身体的特徴や世帯情報等を知ることができる** 想定される利用者攻撃者：隣人
⑤ **インターネット等を通じて特定個人の情報収集を試みる** 想定される利用者攻撃者：ジャーナリスト
⑥ **特定個人の詳細な世帯情報や行動パターンの情報収集を試みる** 想定される利用者攻撃者：ストーカー

①の基本情報は，個人情報の基本となる氏名，性別，住所，生年月日の４つに関する情報を指す。個人の特定について最も重要となる基本的な情報の種類を表した概念であり，個人情報とその保護に関する制度や議論に用いられる。また，自治体が作成・管理している住民基本台帳に記録されている個人についての基本的な情報で，不特定多数の人が住民票等で閲覧することができる。そのため，基本情報は最低限考慮すべき変数である。なお，氏名については匿名データには含まれない。

②については，事業者が保有するデータと照合することで識別され得るが，このようなケースは以下のガイドラインによって制限されていることから除外する。

> なお，法第 33 条に基づいて提供された調査票情報及び法第 36 条に基づいて提供された他の匿名データ及びその他の個体識別が可能となる可能性があるデータとのリンケージを行う場合には，提供を認めない。
>
> （匿名データの作成・提供に係るガイドライン　第８ 提供依頼申出に対する審査　総則）

③，④については，提供されるミクロデータに利用者の知人・友人あるいは隣人が含まれる可能性があるため，考慮すべきである。ただし，②の事業者が保有する客体と比べ，知人・友人や隣人は一般に少なく，偶然に識別してしまう可能性は低い。

最後に⑤，⑥は，故意に識別を試みており前提条件に反するため，本検討から除外する。

以上から，本調査では基本４情報及び知人・友人や隣人が知り得る情報を，準識別子として扱う項目の候補として検討する。

準識別子として扱う項目

就業構造基本調査を例に，準識別子として扱う項目について検討した。まず，基本４情報は下記の変数が該当する。

氏名：該当なし 性別：性別 住所：地域情報 生年月日：年齢

次に，基本４情報以外で知人・友人が知り得る情報を検討し，以下の変数が該当するとした。

- 世帯に関する事項
  - 基本属性
    - 一般・単身の別
- 個人に関する事項
  - 基本属性
    - 有業・無業の別
    - 仕事の主従
    - 就業状態
    - １年前との就業異動
    - 就業異動履歴
    - １年前の就業状況
    - 前職の有無
  - 有業者
    - 従業上の地位 － ８区分
    - 雇用形態
    - 経営組織
    - 産業 - 農林・非農林
    - 産業 - 中分類（旧）
    - 産業 - 中分類（新）
    - 職業 - 中分類
    - 従業者規模
    - 就業の規則性
    - 現職就業時期 - 元号
    - 現職就業時期 - 月
    - 継続就業期間 - 年
    - 副業の有無，従業上の地位

ただし，知人・友人は親密度等から知っている情報に個人差が生じるため，一概に決めることは難しい。一方で，準識別子が多いほど有用性の確保が困難となる問題もある。そこで，先述したように，知人・友人や隣人は一般に少なく，偶然に識別してしまう可能性は低いこ

とから，上記の変数のうち比較的多くの個人の情報を入手できる以下の変数のみを準識別子と定める。

- 有業・無業の別
- 仕事の主従
- 就業状態
- 従業上の地位
- 雇用形態
- 経営組織
- 産業
- 職業中分類
- 従業者規模

最後に，隣人が知り得る情報については，準識別子が多いほど有用性の確保が困難となる問題に加え，一般に知人よりもさらに小数であり，偶然に識別してしまう可能性は極めて低いことから，準識別子としないこととした。

4.5.匿名データ作成のための実証実験

前節で定めた年齢・地域情報・性別・就業に関する変数を準識別子として，各変数の区分を変えながら母集団一意性を確認または推計した。結果，基本 4 情報である年齢・地域情報・性別の組合せにおいては，地域や年齢をある程度詳細にしても母集団一意とはならず，最小度数が十分大きくなる場合も多く見られることを国勢調査結果から確認した。例えば，地域 3（区分：都道府県），年齢 3（区分：年齢各歳で 90 歳トップコーディング）としても，最小度数は 500 以上となる（7.2 国勢調査調査票情報を用いた基本 4 情報の組合せ集計）。

就業情報は国勢調査結果から得られないため確認ができないが，就業情報を組み合わせた場合の母集団一意性を推計できれば，対策の必要性の目安となる。本調査では，就業構造基本調査の標本データから推計を試みる方法を提案する。具体的には，年齢・地域情報・性別の組合せの最小度数をM，そして就業情報の最小度数が全体のp%と推計されたとき，年齢・地域情報・性別・就業情報の組合せの最小度数をM × p%と推計する。この値が一定数以上となれば母集団一意となる可能性は低いと判断し，そうでなければ区分の見直しや撹乱的な手法の適用といった対策を行う。7.3 節では，就業構造基本調査の標本データにおける各就業情報の最小度数と，年齢・地域情報・性別・職業の組合せにおける最小度数の推計値を示す。ただし，職業の情報については，各変数の個別の最小度数となっているが，実際には当該変数の組合せの最小度数を推計できることが望ましい。そのため，より詳細な就業構造基本調査のデータを用いた最小度数の推計が今後の課題である。

7.4 節では，国勢調査調査票情報及び 10%抽出詳細データを用いて，基本 4 情報及び職業による母集団一意・標本一意の検証結果を示す。

基本 4 情報による母集団一意の確認

平成 22 年国勢調査の結果を用いて，基本 4 情報に該当する準識別子である性別・地域・年齢の 3 つに関する変数を組み合わせた場合，各組合せに該当する客体数を確認した。国勢調査は悉皆調査であるため，性別・地域・年齢の組合せの母集団一意性を確認できる。客体数が少ない場合は，リコーディング等の非撹乱的手法により母集団一意を回避する，あるいはノイズ付与等の撹乱的手法を適用する方法が挙げられる。

検討する地域・年齢の区分

基本 4 情報に該当する性別・地域情報・年齢のうち，地域情報と年齢について下のような区分を設定し，各組合せで最小となる数を調査した。

地域情報

3 大都市圏か否か(現行基準)
地方
都道府県
市区名(40 万人以上)
市区名(10 万人以上)
市区町村名(5 万人以上)
市区町村名(3 万人以上)
市区町村名(1 万人以上)

年齢

5 歳階級	85 歳トップコーディング（現行基準）
各歳	85 歳トップコーディング
各歳	90 歳トップコーディング
各歳	95 歳トップコーディング
各歳	100 歳トップコーディング

集計結果

基本 4 情報に該当する準識別子である性別・地域・年齢の 3 つに関する変数を組み合わせた場合，地域情報と年齢の区分を以下の組合せとした場合に，母集団一意に該当する客体があった。

地域情報	年齢
市区町村名(1 万人以上)	各歳（85 歳トップコーディング）
市区町村名(3 万人以上)	各歳（95 歳トップコーディング）
市区町村名(5 万人以上)	各歳（100 歳トップコーディング）
市区町村名(10 万人以上)	各歳（100 歳トップコーディング）

以上から，これらの条件の場合には提供しない，あるいは撹乱的手法の適用を検討する必要がある。
詳細な結果は「7.2 国勢調査調査票情報を用いた基本 4 情報の組合せ集計」を参照。

その他の準識別子に関する検討

その他の準識別子の割合を用いた母集団一意の推計

基本 4 情報以外の準識別子については，悉皆データが無いため，母集団一意の確認ができない。逆に，利用者にとっても母集団一意かどうか知ることが困難ともいえる。本調査では基本 4 情報以外の準識別子である職業について，平成 24 年就業構造基本調査を用いて最小度数とその割合を求め，性別・地域・年齢の組合せの最小客体数に当該割合を掛けることで，性別・地域・年齢およびその他の準識別子の組合せにおいても十分な客体数が期待できるかどうかを確認した。

詳細な結果は「7.3 基本 4 情報に他の準識別子を加えた母集団一意の推計」を参照。

基本 4 情報及び職業による母集団一意の確認

基本 4 情報以外の準識別子については，最小となる変数が職業中分類であったことから，職業分類についてどの程度一意となる客体があるか確認した。職業中分類については国勢調査の 10%抽出(抽出詳細集計)を用いて，職業大分類は国勢調査を母集団として確認を行った。10%抽出の場合ではその標本では一意であっても，全数の場合では一意ではない可能性もあり，該当する客体の数が 2 以上であれば，母集団においても一意ではない。

その結果，どの区分の組合せでも，母集団一意に該当する客体がいくつか存在した。例えば，地域 1（区分：3 大都市圏か否か），年齢 1（区分：5 歳階級で 85 歳トップコーディング）としても，母集団一意となる準識別子の組合せが 9 組存在した。この場合は他の匿名化手法の適用し，公開できる区分を検討する必要がある。

詳細な結果は「7.4.2 国勢調査 10%抽出詳細データを用いた性別・地域・年齢・職業中分類の組合せにおける標本一意」及び「7.4.3 国勢調査調査票情報を用いた性別・地域・年齢・職業大分類の組合せにおける母集団一意」を参照。

サンプルサイズの小さい組合せ

個人に関する変数について，組み合わせた場合にサンプルサイズが小さくなる組合せを確認した。見つかった組合せとしては，下のような例が挙げられた。

- 飛び級の学生（18 歳未満の大学生）
- 海のない都道府県の漁業者（養殖等）
- 18 歳未満の自営業主，役員等
- 低年齢の父母，祖父母
- 高齢の子・孫
- 18 歳未満の就業者

4.5.1. 匿名化技法の検討

センシティブな情報に対する匿名化

匿名データは識別ができないように作成されているが，非常に低い確率で識別できてしまうことが起こりうる。そのため，収入等のセンシティブな変数は，トップコーディングや階級値の統合を検討する。以下では匿名データにおいて，センシティブな変数であるとされる収入がどのように扱われているかを例示する。

現在の匿名データにおける収入の扱い

平成 14 年就業構造基本調査では，調査票と同様で以下のとおりである。

個人所得	世帯収入階級
収入なし，50 万円未満	100 万円未満
50～99 万円	100～199 万円
100～149 万円	200～299 万円
149～199 万円	300～399 万円
200～249 万円	400～499 万円
250～299 万円	500～599 万円
300～399 万円	600～699 万円
400～499 万円	700～799 万円
500～599 万円	800～899 万円
600～699 万円	900～999 万円
700～799 万円	1,000～1,249 万円
800～899 万円	1,250～1,499 万円
900～999 万円	1,500 万円以上
1,000～1,499 万円	
1,500 万円以上	

＊平成 19 年調査以降の調査票では上限は 1,500~1,999 万円，2,000 万円以上となっている

参考情報：全国消費実態調査

- 二人以上の世帯は，2,500 万円以上トップコーディング（連続値）
- 単身世帯は，1,000 万円以上トップコーディング（連続値）

収入の分布

公的統計の結果表において，2,500 万以上の収入が表章されている，民間給与実態統計調査の収入分布は以下の図 1 のとおりである。

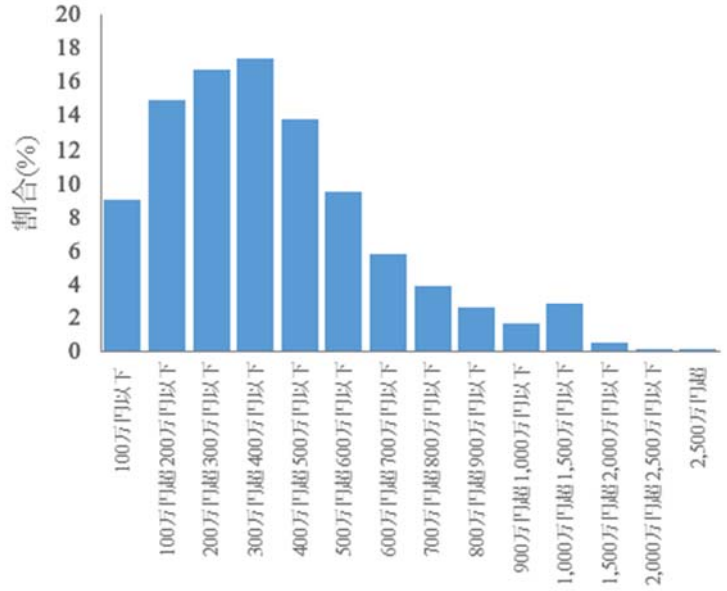

図 1　給与階級別給与所得者 構成比（平成 25 年民間給与実態統計調査から）

匿名データにおける収入は，基本的に階級区分となっており，また階級値および連続値の上限となる区分の割合は，アメリカ統計局で用いられている基準である 0.5%よりも多いことから，センシティブな情報にはすでに十分に配慮がなされている。

母集団からの抽出率

母集団からの抽出率が低い場合には，データの利用者は標本一意であっても，母集団において一意に該当するかを推測することは基本的に難しい。例として，抽出率が母集団の 10%であれば，標本一意であっても，母集団では約 10 レコードが該当すると予想される。ミネソタ大学の IPUMS-I では，抽出率を最大 10%としている。

そのため，有用性を損なわない範囲で抽出率を低くすることにより，母集団で一意であると確信することが難しくなることから，匿名性は強くなる。

撹乱的手法を用いる際の留意点

母集団一意となる客体が含まれる場合，撹乱的手法を用いることにより，母集団で一意であると確信することが難しくなり，識別ができなくなる。スワッピング等は海外では主要な匿名化手法であり，詳細なデータの提供をするためには有用である。しかし，適用する範囲によっては，有用性の低下も懸念されることから，利用者の意見を確認しながら，有用性を損なわない範囲で行うことが望ましい。

5. 提言

5.1.研究の結果による提言

申請や保管，教育用利用等の問題

匿名データは教育用の利用も一つの大きな目的となっているが，現在の管理方法では授業で使うことは難しい。他方，現在の「一般用ミクロデータ」ではデータの有用性が不十分である。このことから，教育用のデータをどのように整備・拡充すればよいか，検討が必要である。

現在は，調査票情報を用いて作成されたデータはすべて匿名データとして扱い，その厳密な利用手続きに従う必要があると解釈されている。しかし，匿名性の度合いによって管理や貸与手続きの厳しさを変えることが必要である。匿名化の度合を高めパブリックユースファイルとして利用できるデータを作ることが望ましいのではないか。

匿名データにおける識別の定義

匿名データにおける「識別」とは，利用者が既に知っている客体を対象とし，対象について特別の調査をすることなく入手が可能な情報（変数）の組合せが母集団で一意となる場合である。これまでは，外観識別情報を準識別子として扱っていたが，現在の匿名データは利用者や利用環境に限定をかけている。そのため，準識別子として扱う必要のある変数は外観識別情報よりも限定される。匿名データにおける準識別子としては少なくとも基本 4 情報が該当し，より厳格な措置として知人が有していると想定される情報を準識別子として扱う。その際に，すべての準識別子を同時に組み合わせて検討する必要はなく，妥当な条件で組み合わせればよいとする。

母集団一意性

標本調査の場合やリサンプリングによって，抽出後が母集団の 10%程度となっていれば，その中で標本一意であっても母集団一意ではない可能性が高い。このような場合，標本一意であっても母集団で一意でなければ問題はない。

年齢と地域情報の詳細化について，考慮したほうがよい基本 4 情報の組合せを，国勢調査による実験からみた結果によれば，母集団において地域情報として市区町村名の開示を人口 3 万人以上かつ年齢は各歳で 95 歳以上をトップコーディングとした場合，または人口 10 万人以上・5 万人以上かつ年齢各歳（100 歳以上トップコーディング）とした場合に，一意となる客体があることがわかった（7.2 国勢調査調査票情報を用いた基本 4 情報の組合せ集

計)。このことから，攪乱的手法を合わせて用いない場合は，少なくともこれよりもいずれかの情報を粗くする必要がある。

知人が有していると想定される基本 4 情報以外の情報のうち，区分が最も詳細になっているのは職業である。基本 4 情報に加え職業を組み合わせた場合には，職業大分類でも現行の基準でも一意となる客体がある（7.4.3 国勢調査調査票情報を用いた性別・地域・年齢・職業大分類の組合せにおける母集団一意）ため，撹乱的手法を用いることを検討したほうが良い。例えば，地域内最小人口が 5 万人以上を一つの基準とし，これ以下の地域については，スワッピング等攪乱的手法を用いる等して個人が特定されるリスクを減じつつ，統計の有用性を減じない工夫が匿名データに求められる。

撹乱的手法を用いない場合には，職業情報が大分類までしか含まれない調査であれば，地域情報は都道府県レベル，年齢は各歳であっても，一意となるのは特定の条件が加わった場合に限られる。このため，そのような条件を考慮すれば，調査ごとに他に準識別子として考慮する必要がある変数がない場合には，都道府県，各歳年齢は，きわめて高齢の区分でなければ開示可能である。

職業が中分類まで含まれる調査の場合では，幅広い条件で一意となる場合が出ることが予想されることから，地域情報または年齢を粗くすることや，攪乱的手法を用いることにより，識別の可能性を低く抑えることを検討すべきである。

5.2.ガイドライン改正案

「匿名データの作成・提供に係るガイドライン」別紙１～３について，以下を改正案とする。

5.2.1. 新旧比較表

別紙１			
章・節	旧	新	変更の種類
(4) 識別情報		そのため，どの情報が識別情報に該当するかについては，利用者や識別の動機に合わせて検討する。	追加
(6) 特定目的のための匿名データの作成について		匿名化処理を行うことで，匿名データに含まれる識別情報は概略化されるが，研究・教育の目的によっては，特定の識別情報が詳細に提供されることが望まれる場合がある。 この要望については，あらかじめ特定の識別情報を詳細化した匿名データの作成を検討すべきである。 ただし，ある識別情報を詳細化することは，特定の可能性を高めるものであるから，他の識別情報について更に匿名化処理を行い，特定の可能性を低く保つようにしなければならない。	章の追加

別紙２			
章・節	旧	新	変更の種類
(1) 匿名化処理の技法	対応関係を特定しにくくする匿名化処理の方法としては，下記のような方法がある。	調査単位（世帯や個人等）とミクロデータの対応関係を特定しにくくする匿名化処理の方法としては，下記のような方法がある。	変更
		非攪乱的手法	追加
	① 識別情報等の削除	① 識別情報の削除	変更
	② 識別情報のトップ・コーディング	② 準識別情報のトップ（ボトム）・コーディング	変更
	対応関係を特定できる可能性が高くなる特殊な属性を，まとめる方法である。	対応関係を特定できる可能性が高くなる特殊な属性をまとめる方法である。	変更
	例えば，100 歳以上の高齢者がいる世帯や世帯員が 10 人いる世帯の数は少ないので，対応関係を特定しやすくなるので，特に大きい値や小さい値を「○○以上」,「○○以下」というようにまとめる。	例えば，100 歳以上の高齢者がいる世帯や世帯員が 10 人いる世帯の数は少ないので，特定される可能性が高い場合，特に大きい値や小さい値を「○○以上」,「○○以下」というようにまとめる。	変更
	海外では，トップ・コーディングされるのが対象全体の 0.5%以上としている例などがある。		削除
	例えば，年齢を例にすると，22 歳ではなく, 21～25 歳とする方法である。	例えば，年齢が 22 歳ではなく, 21～25 歳とする方法である。	変更
	海外では，人口 10 万人未満の地域区分は提供しないなどの基準が設けられている例などがある。		削除
	ミクロデータの配列順を並べ替えることでランダムにし，対応関係を探り出すことができ	ミクロデータの配列順をランダムに並べ替えることで，対応関係を探り出す	変更

	ないようにする方法である	ことができないようにする方法である	
	別の概念からの匿名化処理の技法としては，ミクロデータから正確な対応関係を知ることができないようにする方法がある。	攪乱的手法 ミクロデータから正確な対応関係を知ることができないようにする方法である。	変更
(2) 匿名化処理の方法の決定	上記のような問題があるものの，実際に海外で行われている匿名化処理の方法をみるとかなり詳細なデータをそのまま提供しているのが普通である。		削除
	そのような現実的な判断を行うために，海外では権威ある委員会などが匿名化処理の方法を最終承認する方式をとっている。我が国においても同様の手続きを踏むべきであり，試行的提供では，統計局の「匿名標本データ作成・利用研究会」の承認を得ている。	匿名化処理の方法においては，統計法第三十五条第２項に基づき，統計委員会の承認を得ている。	変更

別紙３			
章・節	旧	新	変更の種類
1. 地理的情報について	(1) 地理的情報としては，地域内に最小でも人口 50 万人以上いなければならない。	(1) 地理的情報としては，母集団一意が発生しない最小の人口を含む地域区分とする。	変更
	(2) 直接的な地理的情報以外で, 地理的情報が明らかになる項目（例えば，サンプリング情報など）についても，上記(1)の最小人口 50 万人の基準に適合させなければならない。	(2) 直接的な地理的情報以外で，地理的情報が明らかになる項目（例えば，サンプリング情報など）についても，上記(1)の基準に適合させなければならない。	変更
	(1) 地域分析用として，人口 50 万人未満の地理的情報を提供するような匿名データを作成する場合には，他の識別情報などの匿名化の程度を高めなければならない。	(2) 地域分析用として，(1)の基準未満の地理的情報を提供するような匿名データを作成する場合には，他の識別情報などの匿名化の程度を高めなければならない。	変更
2. 個人・世帯の識別情報について	トップコーディングにおいては，母集団（個人又は世帯）全体の 0.5%を目安にすることが望ましい。	トップコーディングにおいては，準識別情報を組み合わせた母集団における個人又は世帯数等が 10 以上となることが望ましい。	変更
	(3) 小数の特定の集団を対象とする場合, トップコーディングの基準を３～５%にすることを考慮すべきである。		削除
	世帯単位のデータを提供する場合, 調査単位が特定されることがないよう，必要があれば，匿名化を考慮する必要がある。	世帯単位のデータを提供する場合，調査単位が特定されることがないよう，必要があれば，匿名化手法を考慮する必要がある。	変更
5. 外部ファイルとのマッチングの可能	(1) ミクロデータと外部の既存ファイルのデータを突き合わせることにより調査単位が		削除

性	識別されるような可能性があれば, それを回避するための措置をとらなければならない。		
	(2) 調査のための標本フレームが, 国勢調査の母集団情報以外の情報によって提供されている場合には, 調査データと標本フレームの元の情報とを一致させることが可能となるおそれがあるので, 事前に回避する措置をとらなければならない。		削除
		禁止事項として定められていることから，データ作成においては，考慮しない。	追加
6. その他の問題		提供時期は調査ごとに設定し，前回調査については速やかに提供されるようにすることが望ましい。	追加

5.2.2. ガイドライン別紙改正案

「匿名データの作成・提供に係るガイドライン」別紙１改正案

匿名化処理の考え方

(1) 匿名化処理とは

ミクロデータから世帯や個人の秘密の情報を知るということは，調査対象である調査単位（世帯や個人）とミクロデータの対応関係を特定し，特定されたミクロデータから調査単位の秘密に属する事項を知るということを意味する。どの調査事項が，秘密の情報に当たるかは一概には決めることができないし，時代とともに変化し，普遍的ではないと思われるので，匿名化処理とは，基本的には，調査単位とミクロデータの対応関係を特定されないようにするということである。

(2) 対応関係

提供するミクロデータには，氏名，住所などの直接的に世帯や個人が特定できる情報は付与されていないので，調査単位とミクロデータの対応関係は，性別や年齢などの属性（識別情報）が同じかどうかで判断することになる。

全国の全調査単位のミクロデータが提供されていて，かつ，全調査単位について識別情報が分かる場合，識別情報が一致する調査単位とミクロデータがそれぞれ一つしかない場合には同じ世帯や個人と判断でき，それぞれ複数ある場合はそのうちのいずれかと判断できる。実際のミクロデータの提供の場合，一部の調査単位のミクロデータが提供されていて，かつ，一部の調査単位の識別情報がわかるに過ぎず，このような状況では，対応関係を特定するのは現実的ではないと考えられる。

(3) 特定の可能性

特定の可能性を考えると，地域範囲が狭い場合には，調査対象が絞り込まれるので，識別情報を収集することが容易になり，ミクロデータの地域情報が詳細であれば，特定の可能性が高くなる。また，調査を受けていることが知られていると，その調査単位のミクロデータに必ず存在することが分かるため，対応関係を特定される可能性が高まる。しかし，調査対象のリストは厳格に管理されており，外部の者が調査を受けている調査単位を知る可能性は低く，調査時から数年が経過すれば外部の者が知ることは不可能と言える。

しかし，特殊なデータのときに，特定の可能性は高くなる。例えば，100 歳以上の高齢者がいる世帯や世帯員が 10 人いるというような世帯の数は少ないので，母集団のある個別の世帯に対応するデータ数が少なくなり，そのどれに当たるか決定するのが比較的容易になる。また，複数の属性の特殊な組合せも特定の可能性が高くなる。これに対し，標準的な対

象の場合には同じ条件のデータが多数出現することになるので，特定の可能性は比較的低いものにとどまる。

(4) 識別情報

調査対象である調査単位とミクロデータの対応関係を特定しようとするときに用いる識別情報とは，提供するミクロデータに含まれていて，かつ，統計調査以外からも知ることができる情報である。

個人又は世帯を対象とした統計の場合，比較的容易に入手できる識別情報としては，外観からでも把握できるような基本的な属性が考えられ，例えば，県，市町村などの地域情報や，世帯員数，世帯員の性別，住宅の大きさなどが挙げられる。このほか，自宅で営業している世帯であればその産業・職業を知ることができるし，子供の年齢は通学している学年で分かると思われる。ただし，これらの情報だけでは，一般には対応関係を特定することはできない。また，これらの情報の収集は比較的簡単ではあるが，多数の調査単位について情報を収集しようとすれば大きな作業量を必要とする。

そのため，どの情報が識別情報に該当するかについては，利用者や識別の動機に合わせて検討する。

実際の問題としては，時間が経つとともに識別情報を正確に知ることは難しくなる。提供されるミクロデータは数年前の調査の結果であり，そのときに個々の調査対象がどのような属性を有していたか知ることは，たとえ世帯の基本的な属性であっても難しい。既存のリストのようなものの場合も，そのリストとミクロデータの時点が一致していないと対応関係の特定には多くの誤りが生じることになる。

(5) 特定の試み

匿名化処理の方法を決めるときには，現実にどのような危険があるかについても考えておく必要がある。最近，個人情報の流出がよく問題となるが，そのような例では，住所（メールのアドレス等も含む。），氏名などが流出しており，それは，商業目的などにそのまま利用できる。しかし，統計情報の場合，住所，氏名が流出することはあり得ない。また，前述のとおり，特殊な対象の場合には特定の可能性が比較的高くなるが，多くの標準的な対象の場合には特定の可能性は比較的低いものにとどまる。一部の対象についてだけ特定できたとしても，商業目的での利用価値は少ないであろう。したがって，対象を特定しようとするような試みが，最近問題になっているような商業目的で行われる可能性は低いものと考えられる。そもそも，数年前の統計情報では利用する価値もないであろう。

しかし，もし対象を特定するような試みが実際に行われたら，それはミクロデータ提供の危険性，ひいては統計調査の危険性を指摘するものとして利用されてしまうであろう。ところが，絶対的な匿名性を担保しようとすると，ドイツでの経験のように提供できる情報が極めて限られてしまう。したがって，この問題は匿名化処理だけで対策を考えるべきものでは

なく，そのような試みを行うこと自体を制限しておくことが必要となる。このため，データを提供するときには，利用目的を限定し，データの管理を適正に行わせることを義務付けておかなくてはならない。

注：ドイツは，1980 年の連邦統計法で「絶対的な匿名化」条項によるミクロデータの提供を行ってきたが，多くの情報が失われることになり，科学研究の要求に応じられず，ほとんど利用されなかった。そのため，1987 年の連邦統計法ではミクロデータが莫大な時間や経費をかけない限り識別できないという「事実上の匿名性」の概念に法規定を改正している。

(6) 特定目的のための匿名データの作成について

匿名化処理を行うことで，匿名データに含まれる識別情報は概略化されるが，研究・教育の目的によっては，特定の識別情報が詳細に提供されることが望まれる場合がある。この要望については，あらかじめ特定の識別情報を詳細化した匿名データの作成を検討すべきである。

ただし，ある識別情報を詳細化することは，特定の可能性を高めるものであるから，他の識別情報について更に匿名化処理を行い，特定の可能性を低く保つようにしなければならない。

「匿名データの作成・提供に係るガイドライン」別紙２改正案

匿名化処理の技法

(1) 匿名化処理の技法

調査単位（世帯や個人等）とミクロデータの対応関係を特定しにくくする匿名化処理の方法としては，下記のような方法がある。

非攪乱的手法

① 識別情報の削除

対応関係を特定する危険性の高い識別情報である，世帯や居住地を直接的に特定できるような情報を削除する方法である。

② 準識別情報のトップ（ボトム）コーディング

対応関係を特定できる可能性が高くなる特殊な属性をまとめる方法である。例えば，100歳以上の高齢者がいる世帯や世帯員が 10 人いる世帯の数は少ないので，特定される可能性が高い場合，特に大きい値や小さい値を「○○以上」,「○○以下」というようにまとめる。

③ 識別情報のグルーピング

特定の値をグループ分けして階級区分に変更する方法である。例えば，年齢が 22 歳ではなく，21～25 歳とする方法である。また，市町村コードなどの地域情報の場合は，外部の者にも把握しやすい情報であること，対応関係を調べなくてはならないデータの範囲を限定できることなどから特に注意が必要となる。

④ リサンプリング

ミクロデータをすべて提供するのではなく，そこから抽出した一部のミクロデータだけを提供する方法である。この方法によれば，提供するミクロデータが少なくなるので，対応関係を特定できる可能性を低下させることができる。

また，特定できたとの主張に対し，特定できたと考えることが適当ではないと主張する方法でもある。

⑤ ミクロデータのソート

ミクロデータの配列順をランダムに並べ替えることで，対応関係を探り出すことができないようにする方法である。

攪乱的手法

ミクロデータから正確な対応関係を知ることができないようにする方法である。具体的には，ミクロデータを加工して正しくないものにしてしまう方法である。

① スワッピング

任意の２つの調査単位の間で，一部の調査事項の値を入れ替える方法である。

② 誤差の導入

ミクロデータの一部の調査事項（識別情報又は秘密の情報自体）に誤差を導入する方法である。

(2) 匿名化処理の方法の決定

匿名化処理は，論理的に可能性だけを考えると極めて厳しく行わなくてはならないことになるが，実際には，匿名化の必要性や利用面も考慮して現実的な判断の下で決定している。

匿名化処理の方法においては，統計法第三十五条第２項に基づき，統計委員会の承認を得ている。

「匿名データの作成・提供に係るガイドライン」別紙３改正案

匿名化処理の目安

1. 地理的情報について

(3) 地理的情報としては，母集団一意が発生しない最小の人口を含む地域区分とする。
(4) 直接的な地理的情報以外で，地理的情報が明らかになる項目（例えば，サンプリング情報など）についても，上記(1)の基準に適合させなければならない。
(5) 地域分析用として，(1)の基準未満の地理的情報を提供するような匿名データを作成する場合には，他の識別情報などの匿名化の程度を高めなければならない。
(6) 入手可能な外部情報により，ある特定の種類の施設であることが明らかになるようなことがないようにしなければならない。

2. 個人・世帯の識別情報について

(1) 氏名，住所など個人又は世帯を直接的に識別できる情報は削除されなければならない。
(2) 間接的に個人又は世帯を識別できる情報，例えば年齢，世帯人員，居住室数などの情報については，年齢の高い個人，世帯員数が多い世帯，居住室数の多い住宅など特定される可能性が高い場合，トップコーディング，グルーピングまたは削除を施す必要がある。トップコーディングにおいては，準識別情報を組み合わせた母集団における個人又は世帯数等が 10 以上となることが望ましい。
(3) トップコーディングするデータ項目については，その情報（平均値や中央値など）を明らかにすることが望ましい。
(4) 世帯単位のデータを提供する場合，調査単位が特定されることがないよう，必要があれば，匿名化手法を考慮する必要がある。

3. 誤差（ノイズ）

(1) ミクロデータに誤差を加えることによって，調査データと外部情報との対応関係を特定する可能性を低めることができる。他に適当な匿名化の技法がない場合には，研究・分析上の有用性を損なわない範囲で誤差を付加することを考慮すべきである。
(2) 誤差を加える方法としては，①乱数による誤差の付加（random noise），②調査単位間の調査報の交換（swapping），③ブランク（blank）への置換え又は補定（imputation）がある。

4. リサンプリング

ミクロデータを全て提供する場合は，その一部を提供する場合に比べて，調査単位の特定

の可能が高くなる。例えば，ある人が調査を受けたことがわかっている場合には，ミクロデータの中に必ずその人のデータがあるはずとの前提で探すことができる。したがって，必要に応じて，ミクロデータの全てではなく，一部のデータだけを提供することを考慮すべきである。

5. 外部ファイルとのマッチングの可能性

禁止事項として定められていることから，データ作成においては，考慮しない。

6. その他の問題

(1) データの一連番号，データの並び順によって，およその地域範囲が推測されるおそれがあるので，削除，付替え又は並べ替えをするべきである。
(2) サンプリングに関する情報によっては，地理的情報以外に特定の地域や集団であることが明らかになるおそれがあるので，そのような情報は削除すべきである。
(3) 秘密の情報のうち匿名化の必要性の高い調査項目については，その調査項目自体についてグルーピング，削除等の匿名化を施す必要がある。
(4) 時間の経過とともに，調査データを外部情報と照合することは困難になる。提供時期は調査時点から最低限２年間以上は離すべきである。ただし，提供時期は調査ごとに設定し，前回調査については速やかに提供されるようにすることが望ましい。

6. まとめと今後の課題

本報告書は，現在の匿名データが地域情報に欠け，また年齢も 5 歳階級であることが研究の幅を限定していることを，第一の課題として実証実験を行った。すなわち，「就業構造基本調査」で調査されている調査内容を想定した上で，「国勢調査」を用いて，地域情報や各歳年齢を詳細にした場合に，個人が特定されるリスクについて検証した。

現在の匿名データは，研究者による利用に限定し，厳密な手続きを経て提供を行っている。そのため，提供にやや時間はかかるものの，匿名データの利用者は限定されている。そのため，匿名性としては，絶対的な匿名性ではなく事実上の匿名性で良いものとし，基本 4 情報で母集団一意となるかどうかを検証した。この結果，地域内最小人口 5 万人以上で居住地を示し，年齢各歳でデータを提供しても，よほど高齢の場合を除き，母集団一意とはならないことが示された。さらに，提供するデータに攪乱的手法の導入等の匿名化措置を追加することで，より人口規模が少ない地域や職業中分類等を含めて，データの有用性を下げずに情報提供を行うことができよう。

この研究報告書を受けて，今後提供される匿名データにおいて地域情報が，あるいは，年齢各歳別情報が，匿名性を担保した上で提供されるようになれば，研究者がより容易に日本の代表的な公的統計にアクセスでき，かつ地域の雇用情勢を含めて分析することができることとなる。これは，匿名データ利用の拡充に大きなプラスであり，誠に喜ばしいことであり，現在の手続き等を前提とした上での匿名データの利用改善の方向である。

また，今後の課題として，パブリックユースファイルについても検討が行われることが望ましい。日本では現在はまだ検討されていないが，海外では一般的な雇用や家族状況の調査である世帯統計については，一定の匿名化措置をした上で，幅広く個票データ利用を可能とすることの価値が高いと考え，そのようなパブリックユースファイルが簡単な手続きで手に入るように提供されている。

海外のデータ利用の状況は大いに参考になる。そこで，本報告書は海外のパブリックユースファイル等についても章を設けて示した。例えば，隣接地域の世帯を交換して居住地域情報を攪乱するデータスワッピングであるが，狭い地域の分析をすればデータの精度に影響が出るが，都道府県別情報による分析とすれば，分析結果に影響を与えない。その上で個人の特定を難しくする。このように作成されたパブリックユースファイルは大学生，大学院生の統計の授業等でも広く使われている。例えば，アメリカでは公的統計である「Currrent Population Survey（労働力調査）」のパブリックユースファイルを例に，統計分析の授業が行われる。そのため，パブリックユースファイルのダウンロード数は多く，なじみも深い。

日本については，もっとも開放性の高い匿名データについてさえも，申請による利用が 1 統計あたりせいぜい年間 5 件程度から十数件程度である。日本の現状をもっともよく伝える公的統計の分析利用が低調であることは懸念である。きわめて精度の高い統計があるに

もかかわらず，コンピュータ技術の発展に見合った利用がされず，公的機関が行っている統計表作成以上の利用がされていないとすれば，それはきわめて残念なことである。

公的統計は，小規模のデータには到底望めないような精度で日本の現状を伝えられる大きな力を持っている。この統計を，現状分析に，また社会変化の把握に，現在のコンピュータ技術を生かして活用するためのインフラストラクチャーを作ることはきわめて重要である。

今回は，匿名データの現在の厳格な利用審査手続き等を踏まえた上で，地域情報や各歳年齢の開示が可能との実証実験結果を示した。地域情報や年齢各歳情報が増えれば，匿名データの活用の幅が広がり利用が増えると考えられる。

今後については，パブリックユースファイルも考えていくことが望ましい。これは匿名データとは別のものとなるだろう。パブリックユースファイルは，ウェブからダウンロードできる等，幅広い利用者が容易に入手できる体制を前提とする。そのため，教育利用にも向いている。研究利用としても，入手のしやすさが改善されることで，質の高い公的統計を用いた社会課題の分析が広がるものと考えられる。ただし，個人の識別に関する匿名化措置の度合いは匿名データよりも強いものと期待される。

7. 集計結果詳細

7.1.共通の集計条件

7.1.1. サンプル数

総数	128,057,352
うち一般世帯・15 歳以上	108,785,754

7.1.2. 各変数 集計除外条件

年齢	VVV	不詳（基本項目記入不備世帯）	972,269

7.1.3. 年齢の集計条件

区分	年齢階級	トップコーディング
年齢 1	５歳階級	85 歳
年齢 2	各歳	85 歳
年齢 3	各歳	90 歳
年齢 4	各歳	95 歳
年齢 5	各歳	100 歳

7.1.4. 地域の集計条件[10]

区分	地域情報	区分の数
地域 1	３大都市圏か	2
地域 2	地方	10
地域 3	都道府県	47
地域 4	市区町村（40 万人以上）	84
地域 5	市区町村（10 万人以上）	455
地域 6	市区町村（５万人以上）	816
地域 7	市区町村（３万人以上）	1,145
地域 8	市区町村（１万人以上）	1,730

10 区分には地域コードを用いた

3 大都市圏に該当する市区町村

国勢調査の大都市圏・都市圏 のうち，関東・中京・近畿大都市圏に含まれる市区町村。

市区町村の人口

国勢調査の人口基本集計の結果を用いた。

3 大都市圏に該当する市区町村および市区町村の人口は e-Stat（https://www.e-stat.go.jp/）から取得した。

都道府県と地方名

都道府県	地方	都道府県	地方
北海道	北海道	滋賀県	近畿
青森県	東北地方	京都府	近畿
岩手県	東北地方	大阪府	近畿
宮城県	東北地方	兵庫県	近畿
秋田県	東北地方	奈良県	近畿
山形県	東北地方	和歌山県	近畿
福島県	東北地方	鳥取県	中国
茨城県	北関東・甲信	島根県	中国
栃木県	北関東・甲信	岡山県	中国
群馬県	北関東・甲信	広島県	中国
埼玉県	南関東	山口県	中国
千葉県	南関東	徳島県	四国
東京都	南関東	香川県	四国
神奈川県	南関東	愛媛県	四国
新潟県	北陸	高知県	四国
富山県	北陸	福岡県	九州
石川県	北陸	佐賀県	九州
福井県	北陸	長崎県	九州
山梨県	北関東・甲信	熊本県	九州
長野県	北関東・甲信	大分県	九州
岐阜県	東海	宮崎県	九州
静岡県	東海	鹿児島県	九州
愛知県	東海	沖縄県	九州
三重県	東海		

7.2.国勢調査調査票情報を用いた基本 4 情報の組合せ集計

性別	地域	年齢		性別	地域	年齢	レコード数	割合
性別	－	－		男性	－	－	52,720,674	48.463%
－	地域 1	－		－	その他	－	52,841,633	48.574%
－	地域 2	－		－	四国	－	3,358,504	3.087%
－	地域 3	－		－	鳥取県	－	494,812	0.455%
－	地域 4	－		－	宮崎市	－	332,790	0.306%
－	地域 5	－		－	筑紫野市	－	82,247	0.076%
－	地域 6	－		－	田川市	－	41,612	0.038%
－	地域 7	－		－	時津町	－	23,730	0.022%
－	地域 8	－		－	1 万人未満(大分県)	－	1,945	0.002%
－	－	年齢 1		－	－	85 歳以上	2,970,141	2.730%
－	－	年齢 2		－	－	84 歳	664,132	0.610%
－	－	年齢 3		－	－	89 歳	276,393	0.254%
－	－	年齢 4		－	－	94 歳	87,641	0.081%
－	－	年齢 5		－	－	99 歳	13,019	0.012%
性別	地域 1	－		男性	その他	－	25,196,795	23.162%
性別	地域 2	－		男性	四国	－	1,583,175	1.455%
性別	地域 3	－		男性	鳥取県	－	234,421	0.215%
性別	地域 4	－		男性	宮崎市	－	154,358	0.142%
性別	地域 5	－		男性	筑紫野市	－	38,640	0.036%
性別	地域 6	－		男性	田川市	－	18,614	0.017%
性別	地域 7	－		男性	時津町	－	11,175	0.010%
性別	地域 8	－		男性	1 万人未満(大分県)	－	889	0.001%
性別	－	年齢 1		男性	－	85 歳以上	913,453	0.840%
性別	－	年齢 2		男性	－	84 歳	251,199	0.231%
性別	－	年齢 3		男性	－	89 歳	79,665	0.073%
性別	－	年齢 4		男性	－	94 歳	21,719	0.020%
性別	－	年齢 5		男性	－	99 歳	2,698	0.002%
－	地域 1	年齢 1		－	三大都市圏	85 歳以上	1,196,263	1.100%
－	地域 1	年齢 2		－	三大都市圏	84 歳	271,988	0.250%

性別	地域	年齢		性別	地域	年齢	レコード数	割合
−	地域 1	年齢 3		−	三大都市圏	89 歳	111,156	0.102%
−	地域 1	年齢 4		−	三大都市圏	94 歳	35,649	0.033%
−	地域 1	年齢 5		−	三大都市圏	99 歳	5,407	0.005%
−	地域 2	年齢 1		−	四国	85 歳以上	124,355	0.114%
−	地域 2	年齢 2		−	四国	19 歳	27,470	0.025%
−	地域 2	年齢 3		−	四国	89 歳	11,657	0.011%
−	地域 2	年齢 4		−	北海道	94 歳	3,451	0.003%
−	地域 2	年齢 5		−	北海道	99 歳	478	0.000%
−	地域 3	年齢 1		−	鳥取県	85 歳以上	20,173	0.019%
−	地域 3	年齢 2		−	鳥取県	19 歳	4,196	0.004%
−	地域 3	年齢 3		−	鳥取県	89 歳	1,888	0.002%
−	地域 3	年齢 4		−	鳥取県	94 歳	597	0.001%
−	地域 3	年齢 5		−	福井県	99 歳	90	0.000%
−	地域 4	年齢 1		−	柏市	85 歳以上	5,808	0.005%
−	地域 4	年齢 2		−	柏市	84 歳	1,375	0.001%
−	地域 4	年齢 3		−	柏市	89 歳	545	0.001%
−	地域 4	年齢 4		−	柏市	94 歳	177	0.000%
−	地域 4	年齢 5		−	江東区	99 歳	21	0.000%
−	地域 5	年齢 1		−	戸田市	85 歳以上	1,094	0.001%
−	地域 5	年齢 2		−	浦添市	84 歳	263	0.000%
−	地域 5	年齢 3		−	戸田市	89 歳	107	0.000%
−	地域 5	年齢 4		−	戸田市	94 歳	33	0.000%
−	地域 5	年齢 5		−	戸田市	99 歳	1	0.000%
−	地域 6	年齢 1		−	みよし市	85 歳以上	486	0.000%
−	地域 6	年齢 2		−	豊見城市	83 歳	120	0.000%
−	地域 6	年齢 3		−	みよし市	89 歳	46	0.000%
−	地域 6	年齢 4		−	長久手町	94 歳	10	0.000%
−	地域 6	年齢 5		−	12 の組合せ		1	0.000%
−	地域 7	年齢 1		−	松伏町	85 歳以上	418	0.000%
−	地域 7	年齢 2		−	対馬市	19 歳	74	0.000%
−	地域 7	年齢 3		−	松伏町	88 歳	39	0.000%
−	地域 7	年齢 4		−	松伏町	94 歳	7	0.000%

性別	地域	年齢		性別	地域	年齢	レコード数	割合
−	地域 7	年齢 5		−	46 の組合せ		1	0.000%
−	地域 8	年齢 1		−	1 万人未満 (富山県)	85 歳以上	33	0.000%
−	地域 8	年齢 2		−	1 万人未満 (大分県)	21 歳	1	0.000%
−	地域 8	年齢 3		−	2 の組合せ		1	0.000%
−	地域 8	年齢 4		−	3 の組合せ		1	0.000%
−	地域 8	年齢 5		−	266 の組合せ		1	0.000%
性別	地域 1	年齢 1		男性	三大都市圏	85 歳以上	368,975	0.339%
性別	地域 1	年齢 2		男性	三大都市圏	84 歳	104,633	0.096%
性別	地域 1	年齢 3		男性	三大都市圏	89 歳	31,901	0.029%
性別	地域 1	年齢 4		男性	三大都市圏	94 歳	8,766	0.008%
性別	地域 1	年齢 5		男性	三大都市圏	99 歳	1,123	0.001%
性別	地域 2	年齢 1		男性	四国	85 歳以上	38,859	0.036%
性別	地域 2	年齢 2		男性	四国	84 歳	10,463	0.010%
性別	地域 2	年齢 3		男性	四国	89 歳	3,561	0.003%
性別	地域 2	年齢 4		男性	四国	94 歳	931	0.001%
性別	地域 2	年齢 5		男性	北陸	99 歳	124	0.000%
性別	地域 3	年齢 1		男性	鳥取県	85 歳以上	5,866	0.005%
性別	地域 3	年齢 2		男性	鳥取県	84 歳	1,506	0.001%
性別	地域 3	年齢 3		男性	鳥取県	89 歳	521	0.000%
性別	地域 3	年齢 4		男性	鳥取県	94 歳	140	0.000%
性別	地域 3	年齢 5		男性	福井県	99 歳	15	0.000%
性別	地域 4	年齢 1		男性	枚方市	85 歳以上	1,847	0.002%
性別	地域 4	年齢 2		男性	川口市	84 歳	535	0.000%
性別	地域 4	年齢 3		男性	松戸市	89 歳	152	0.000%
性別	地域 4	年齢 4		男性	川口市	94 歳	39	0.000%
性別	地域 4	年齢 5		男性	川口市	100 歳以上	3	0.000%
性別	地域 5	年齢 1		男性	美浜区	85 歳以上	344	0.000%
性別	地域 5	年齢 2		男性	浦添市	84 歳	79	0.000%
性別	地域 5	年齢 3		男性	朝霞市	89 歳	30	0.000%
性別	地域 5	年齢 4		男性	城南区	94 歳	3	0.000%

性別	地域	年齢		性別	地域	年齢	レコード数	割合
性別	地域 5	年齢 5		126 の組合せ			1	0.000%
性別	地域 6	年齢 1		男性	みよし市	85 歳以上	151	0.000%
性別	地域 6	年齢 2		男性	糸満市	84 歳	36	0.000%
性別	地域 6	年齢 3		男性	みよし市	88 歳	10	0.000%
性別	地域 6	年齢 4		男性	白岡町	93 歳	2	0.000%
性別	地域 6	年齢 5		428 の組合せ			1	0.000%
性別	地域 7	年齢 1		男性	松伏町	85 歳以上	123	0.000%
性別	地域 7	年齢 2		男性	南風原町	84 歳	19	0.000%
性別	地域 7	年齢 3		男性	美原区	88 歳	8	0.000%
性別	地域 7	年齢 4		6 の組合せ			1	0.000%
性別	地域 7	年齢 5		797 の組合せ			1	0.000%
性別	地域 8	年齢 1		男性	1 万人未満 (東京都)	85 歳以上	13	0.000%
性別	地域 8	年齢 2		4 の組合せ			1	0.000%
性別	地域 8	年齢 3		8 の組合せ			1	0.000%
性別	地域 8	年齢 4		108 の組合せ			1	0.000%
性別	地域 8	年齢 5		1999 の組合せ			1	0.000%

7.3.基本 4 情報に他の準識別子を加えた母集団一意の推計

7.3.1. 就業構造基本調査の標本データにおける職業情報の最小度数

平成 24 年就業構造基本調査から，準識別子に該当する変数それぞれの区分とレコード数を確認した。

データは e-Stat から取得し，各変数とその区分はレコード数が少ない順に示している。

職業分類(中分類)	レコード数
採掘従事者	4,000
船舶・航空機運転従事者	24,500
家庭生活支援サービス職業従事者	29,800
鉄道運転従事者	43,100
林業従事者	51,400
管理的公務員	53,300
外勤事務従事者	89,800
音楽家，舞台芸術家	92,600
法務従事者	93,600
著述家，記者，編集者	132,100
宗教家	133,300
研究者	150,100
その他の輸送従事者	152,700
漁業従事者	161,800
事務用機器操作員	197,400
経営・金融・保険専門職業従事者	203,900
その他の管理的職業従事者	242,000
居住施設・ビル等管理人	330,100
包装従事者	330,800
機械検査従事者	345,100
美術家，デザイナー，写真家，　映像撮影者	350,600
運輸・郵便事務従事者	360,200
保健医療サービス職業従事者	400,400
定置・建設機械運転従事者	407,900
製品検査従事者	412,900

職業分類(中分類)	レコード数
販売類似職業従事者	475,700
その他のサービス職業従事者	566,300
生産関連事務従事者	578,300
電気工事従事者	582,900
営業・販売事務従事者	666,900
生産関連・生産類似作業従事者	723,900
生活衛生サービス職業従事者	894,700
その他の専門的職業従事者	916,000
その他の運搬・清掃・包装等従事者	996,700
社会福祉専門職業従事者	1,008,800
清掃従事者	1,110,700
機械整備・修理従事者	1,118,800
法人・団体役員	1,131,900
保安職業従事者	1,146,500
保安職業従事者	1,146,500
製品製造・加工処理従事者（金属製品）	1,313,100
管理的職業従事者	1,427,100
機械組立従事者	1,500,700
教員	1,560,200
介護サービス職業従事者	1,600,700
運搬従事者	1,663,500
自動車運転従事者	1,681,400
会計事務従事者	1,728,000
接客・給仕職業従事者	1,808,600
飲食物調理従事者	2,091,100
農業従事者	2,155,100
分類不能の職業	2,233,800
建設・土木作業従事者	2,268,300
輸送・機械運転従事者	2,309,600
農林漁業従事者	2,368,300
技術者	2,654,900
保健医療従事者	2,845,600
建設・採掘従事者	2,855,200

職業分類(中分類)	レコード数
営業職業従事者	3,507,200
製品製造・加工処理従事者（金属製品を除く）	3,733,100
運搬・清掃・包装等従事者	4,101,800
商品販売従事者	4,576,300
サービス職業従事者	7,721,700
販売従事者	8,559,200
一般事務従事者	8,788,000
生産工程従事者	9,147,400
専門的・技術的職業従事者	10,141,600
事務従事者	12,408,600

7.3.2. 年齢・地域情報・性別・職業の組合せにおける最小度数の推計結果

年齢・地域情報・性別の組合せの最小度数をM，就業情報の最小度数が全体のp%としたとき，年齢・地域情報・性別・就業情報の組合せの最小度数をM × p%と推計した。

性別	地域	年齢	各組合せのうち最小レコード数	職業中分類（採掘従事者）による推計最小度数	職業大分類（保安職業従事者）による推計最小度数
性別	地域 1	年齢 1	368,975	13	3,817
性別	地域 1	年齢 2	104,633	4	1,083
性別	地域 1	年齢 3	31,901	1	330
性別	地域 1	年齢 4	8,766	0	91
性別	地域 1	年齢 5	1,123	0	12
性別	地域 2	年齢 1	38,859	1	402
性別	地域 2	年齢 2	10,463	0	108
性別	地域 2	年齢 3	3,561	0	37
性別	地域 2	年齢 4	931	0	10
性別	地域 2	年齢 5	124	0	1
性別	地域 3	年齢 1	5,866	0	61
性別	地域 3	年齢 2	1,506	0	16
性別	地域 3	年齢 3	521	0	5
性別	地域 3	年齢 4	140	0	1

性別	地域	年齢	各組合せのうち最小レコード数	職業中分類（採掘従事者）による推計最小度数	職業大分類（保安職業従事者）による推計最小度数
性別	地域 3	年齢 5	15	0	0
性別	地域 4	年齢 1	1,847	0	19
性別	地域 4	年齢 2	535	0	6
性別	地域 4	年齢 3	152	0	2
性別	地域 4	年齢 4	39	0	0
性別	地域 4	年齢 5	3	0	0
性別	地域 5	年齢 1	344	0	4
性別	地域 5	年齢 2	79	0	1
性別	地域 5	年齢 3	30	0	0
性別	地域 5	年齢 4	3	0	0
性別	地域 5	年齢 5	1	0	0
性別	地域 6	年齢 1	151	0	2
性別	地域 6	年齢 2	36	0	0
性別	地域 6	年齢 3	10	0	0
性別	地域 6	年齢 4	2	0	0
性別	地域 6	年齢 5	1	0	0
性別	地域 7	年齢 1	123	0	1
性別	地域 7	年齢 2	19	0	0
性別	地域 7	年齢 3	8	0	0
性別	地域 7	年齢 4	1	0	0
性別	地域 7	年齢 5	1	0	0
性別	地域 8	年齢 1	13	0	0
性別	地域 8	年齢 2	1	0	0
性別	地域 8	年齢 3	1	0	0
性別	地域 8	年齢 4	1	0	0
性別	地域 8	年齢 5	1	0	0

7.4.国勢調査調査票情報を用いた基本４情報及び職業による検証

7.4.1. 国勢調査調査票情報による職業別度数

国勢調査調査票情報及び 10%抽出詳細データを用いて，職業別の度数を確認した。ただし，調査票情報には職業大分類までの情報しかないため，職業中分類は 10%抽出詳細のみを用いて確認した。

国勢調査調査票情報及び 10%抽出詳細データを用いた職業大分類別度数

職業大分類	全数	10%抽出詳細
保安職業従事者	979,051	101,131
管理的職業従事者	1,419,712	144,232
輸送・機械運転従事者	2,087,487	223,182
農林漁業従事者	2,323,559	391,721
建設・採掘従事者	2,671,355	291,108
分類不能の職業	3,387,733	276,080
運搬・清掃・包装等従事者	3,689,494	377,053
サービス職業従事者	6,828,683	700,924
販売従事者	7,996,040	748,029
生産工程従事者	8,453,309	893,105
専門的・技術的職業従事者	8,628,598	816,631
事務従事者	10,977,604	1,053,719
対象外	49,343,129	4,924,997
合計	108,785,754	10,941,912

国勢調査 10%抽出詳細データを用いた職業中分類別度数

職業中分類	10%抽出詳細
管理的公務員	8,918
法人・団体職員	107,131
その他の管理的職業従事者	28,183
研究者	9,822
技術者	190,078
保険医療従事者	252,161
社会福祉専門職業従事者	85,657
法務従事者	6,571
経営・金融・保険専門職業従事者	12,702

職業中分類	10%抽出詳細
教員	139,764
宗教家	13,719
著述家，記者，編集者	8,518
美術家，デザイナー，写真家，映像撮影者	23,198
音楽家，舞台芸術家	5,651
その他の専門的職業従事者	68,790
一般事務従事者	724,769
会計事務従事者	161,122
生産関連事務従事者	46,950
営業・販売事務従事者	49,837
外勤事務従事者	15,514
運輸・郵便事務従事者	36,912
事務用機器操作員	18,615
商品販売従事者	423,199
販売類似職業従事者	38,856
営業職業従事者	285,974
家庭生活支援サービス職業従事者	2,135
介護サービス職業従事者	140,269
保険医療サービス職業従事者	33,136
生活衛生サービス職業従事者	85,605
飲食物調理従事者	201,210
接客・給仕職業従事者	168,915
居住施設・ビル等管理人	23,773
その他のサービス職業従事者	45,881
保安職業従事者	101,131
農業従事者	345,767
林業従事者	11,886
漁業従事者	34,068
製品製造・加工処理従事者（金属製品）	129,955
製品製造・加工処理従事者（金属製品を除く）	369,459
機械組立従事者	150,599
機械整備・修理従事者	106,059
製品検査従事者	40,187
機械検査従事者	35,139

職業中分類	10%抽出詳細
生産関連・生産類似作業従事者	61,707
鉄道運転従事者	3,529
自動車運転従事者	159,801
船舶・航空機運転従事者	3,687
その他の輸送従事者	14,635
定置・建設機械運転従事者	41,530
建設・土木作業従事者	234,666
電気工事従事者	55,705
採掘従事者	737
運搬従事者	158,547
清掃従事者	98,190
包装従事者	30,331
その他の運搬・清掃・包装等従事者	89,985
分類不能の職業	276,080
対象外	4,924,997
合計	10,941,912

7.4.2. 国勢調査 10%抽出詳細データを用いた性別・地域・年齢・職業中分類の組合せにおける標本一意

国勢調査の 10%抽出詳細データを用いて，性別・地域・年齢・職業中分類の組合せで集計を行い，標本一意の発生数と割合を確認した。その結果，どの区分の組合せでも，標本一意に該当する客体がいくつか存在した。

性別	地域	年齢	職業中分類
−	−	−	−
性別	地域 1	年齢 1	職業中分類
性別	地域 1	年齢 2	職業中分類
性別	地域 1	年齢 3	職業中分類
性別	地域 1	年齢 4	職業中分類
性別	地域 1	年齢 5	職業中分類
性別	地域 2	年齢 1	職業中分類
性別	地域 2	年齢 2	職業中分類
性別	地域 2	年齢 3	職業中分類
性別	地域 2	年齢 4	職業中分類
性別	地域 2	年齢 5	職業中分類
性別	地域 3	年齢 1	職業中分類
性別	地域 3	年齢 2	職業中分類
性別	地域 3	年齢 3	職業中分類
性別	地域 3	年齢 4	職業中分類
性別	地域 3	年齢 5	職業中分類
性別	地域 4	年齢 1	職業中分類
性別	地域 4	年齢 2	職業中分類
性別	地域 4	年齢 3	職業中分類
性別	地域 4	年齢 4	職業中分類
性別	地域 4	年齢 5	職業中分類
性別	地域 5	年齢 1	職業中分類
性別	地域 5	年齢 2	職業中分類
性別	地域 5	年齢 3	職業中分類
性別	地域 5	年齢 4	職業中分類
性別	地域 5	年齢 5	職業中分類

性別	地域	年齢	職業中分類	標本内レコード数	割合
−	−	−	採掘従事者	737	0.007%
68 の組合せ				1	0.000%
545 の組合せ				1	0.000%
701 の組合せ				1	0.000%
835 の組合せ				1	0.000%
900 の組合せ				1	0.000%
705 の組合せ				1	0.000%
5,437 の組合せ				1	0.000%
6,159 の組合せ				1	0.000%
6,605 の組合せ				1	0.000%
6,691 の組合せ				1	0.000%
6,032 の組合せ				1	0.000%
40,454 の組合せ				1	0.000%
42,526 の組合せ				1	0.000%
43,368 の組合せ				1	0.000%
43,462 の組合せ				1	0.000%
13,551 の組合せ				1	0.000%
84,288 の組合せ				1	0.000%
86,687 の組合せ				1	0.000%
87,557 の組合せ				1	0.000%
87,737 の組合せ				1	0.000%
92,732 の組合せ				1	0.000%
496,698 の組合せ				1	0.000%
500,511 の組合せ				1	0.000%
501,587 の組合せ				1	0.000%
502,808 の組合せ				1	0.000%

性別	地域	年齢	職業中分類
性別	地域 6	年齢 1	職業中分類
性別	地域 6	年齢 2	職業中分類
性別	地域 6	年齢 3	職業中分類
性別	地域 6	年齢 4	職業中分類
性別	地域 6	年齢 5	職業中分類
性別	地域 7	年齢 1	職業中分類
性別	地域 7	年齢 2	職業中分類
性別	地域 7	年齢 3	職業中分類
性別	地域 7	年齢 4	職業中分類
性別	地域 7	年齢 5	職業中分類
性別	地域 8	年齢 1	職業中分類
性別	地域 8	年齢 2	職業中分類
性別	地域 8	年齢 3	職業中分類
性別	地域 8	年齢 4	職業中分類
性別	地域 8	年齢 5	職業中分類

性別	地域	年齢	職業中分類	標本内レコード数	割合
			162,307 の組合せ	1	0.000%
			784,647 の組合せ	1	0.000%
			789,022 の組合せ	1	0.000%
			790,650 の組合せ	1	0.000%
			792,703 の組合せ	1	0.000%
			218,411 の組合せ	1	0.000%
			1,007,169 の組合せ	1	0.000%
			1,011,900 の組合せ	1	0.000%
			1,013,943 の組合せ	1	0.000%
			1,016,769 の組合せ	1	0.000%
			320,313 の組合せ	1	0.000%
			1,378,330 の組合せ	1	0.000%
			1,383,764 の組合せ	1	0.000%
			1,386,641 の組合せ	1	0.000%
			1,390,828 の組合せ	1	0.000%

7.4.3. 国勢調査調査票情報を用いた性別・地域・年齢・職業大分類の組合せにおける母集団一意

国勢調査調査票情報を用いて，性別・地域・年齢・職業大分類の組合せで集計を行い，母集団一意の発生数と割合を確認した。その結果，どの区分の組合せでも，母集団一意に該当する客体がいくつか存在した。

性別	地域	年齢	職業大分類
－	－	－	－
性別	地域 1	年齢 1	職業大分類
性別	地域 1	年齢 2	職業大分類
性別	地域 1	年齢 3	職業大分類
性別	地域 1	年齢 4	職業大分類
性別	地域 1	年齢 5	職業大分類
性別	地域 2	年齢 1	職業大分類
性別	地域 2	年齢 2	職業大分類
性別	地域 2	年齢 3	職業大分類
性別	地域 2	年齢 4	職業大分類
性別	地域 2	年齢 5	職業大分類
性別	地域 3	年齢 1	職業大分類
性別	地域 3	年齢 2	職業大分類
性別	地域 3	年齢 3	職業大分類
性別	地域 3	年齢 4	職業大分類
性別	地域 3	年齢 5	職業大分類
性別	地域 4	年齢 1	職業大分類
性別	地域 4	年齢 2	職業大分類
性別	地域 4	年齢 3	職業大分類
性別	地域 4	年齢 4	職業大分類
性別	地域 4	年齢 5	職業大分類
性別	地域 5	年齢 1	職業大分類
性別	地域 5	年齢 2	職業大分類
性別	地域 5	年齢 3	職業大分類
性別	地域 5	年齢 4	職業大分類

性別	地域	年齢	職業大分類	母集団内レコード数	割合
－	－	－	保安職業従事者	979,051	0.900%
女性	その他	85 歳以上	保安職業従事者	1	0.000%
9 の組合せ				1	0.000%
18 の組合せ				1	0.000%
36 の組合せ				1	0.000%
62 の組合せ				1	0.000%
15 の組合せ				1	0.000%
133 の組合せ				1	0.000%
202 の組合せ				1	0.000%
273 の組合せ				1	0.000%
516 の組合せ				1	0.000%
140 の組合せ				1	0.000%
1,229 の組合せ				1	0.000%
1,619 の組合せ				1	0.000%
2,563 の組合せ				1	0.000%
3,345 の組合せ				1	0.000%
335 の組合せ				1	0.000%
3,607 の組合せ				1	0.000%
4,765 の組合せ				1	0.000%
6,578 の組合せ				1	0.000%
7,643 の組合せ				1	0.000%
4,159 の組合せ				1	0.000%
44,222 の組合せ				1	0.000%
56,465 の組合せ				1	0.000%
61,448 の組合せ				1	0.000%

性別	地域	年齢	職業大分類
性別	地域 5	年齢 5	職業大分類
性別	地域 6	年齢 1	職業大分類
性別	地域 6	年齢 2	職業大分類
性別	地域 6	年齢 3	職業大分類
性別	地域 6	年齢 4	職業大分類
性別	地域 6	年齢 5	職業大分類
性別	地域 7	年齢 1	職業大分類
性別	地域 7	年齢 2	職業大分類
性別	地域 7	年齢 3	職業大分類
性別	地域 7	年齢 4	職業大分類
性別	地域 7	年齢 5	職業大分類
性別	地域 8	年齢 1	職業大分類
性別	地域 8	年齢 2	職業大分類
性別	地域 8	年齢 3	職業大分類
性別	地域 8	年齢 4	職業大分類
性別	地域 8	年齢 5	職業大分類

性別	地域	年齢	職業大分類	母集団内レコード数	割合
			652,593 の組合せ	1	0.000%
			9,681 の組合せ	1	0.000%
			89,652 の組合せ	1	0.000%
			106,679 の組合せ	1	0.000%
			115,101 の組合せ	1	0.000%
			116,451 の組合せ	1	0.000%
			15,955 の組合せ	1	0.000%
			132,981 の組合せ	1	0.000%
			153,883 の組合せ	1	0.000%
			162,840 の組合せ	1	0.000%
			164,472 の組合せ	1	0.000%
			31,322 の組合せ	1	0.000%
			229,099 の組合せ	1	0.000%
			253,522 の組合せ	1	0.000%
			262,775 の組合せ	1	0.000%
			265,481 の組合せ	1	0.000%

「匿名データの利用改善に向けた研究会」について

目的

「匿名データの利用改善に向けた研究会」は「匿名データの作成・提供に係るガイドライン」の改正に向け，利用者側からの意見等を反映させるとともに，技術的助言を得るために開催する。

検討事項

（１）　匿名データの利用状況，課題の整理
（２）　匿名データの利用形態の在り方
（３）　匿名データにおける匿名性と有用性
（４）　匿名データにおける地域情報提供の在り方の整理
（５）　匿名データの利用改善に向けた提案

構成員

（委員）　（50 音順）

川口　大司　　国立大学法人東京大学大学院経済学研究科教授
千田　浩司　　NTT セキュアプラットフォーム研究所主任研究員
永瀬　伸子　　国立大学法人お茶の水女子大学基幹研究院教授
南　和宏　　大学共同利用機関法人情報・システム研究機構統計数理研究所准教授

（アドバイザー）（50 音順）

宇南山　卓　　国立大学法人一橋大学経済研究所准教授
岡室　博之　　国立大学法人一橋大学経済学研究科教授
神林　龍　　国立大学法人一橋大学経済研究所教授
小林　良行　　総務省統計研修所教授
山口　幸三　　総務省統計研修所教授

＜事務局＞

佐々木　健一　　総務省統計委員会担当室室長補佐
白川　清美　　国立大学法人一橋大学経済研究所准教授
阿部　穂日　　国立大学法人一橋大学経済研究所助教

※座長は必要があると認めるときは，関係者を研究会に出席させ，意見を聴くことができる。

付録 1　アメリカ統計局のパブリックユースファイルの貸与方法と匿名性のための工夫

以下は，アメリカ国勢調査のパブリックユースファイルの作成方法の概要を示した。英語を直接読めるよう，日本語で概要を示したうえで，原典である英文を入れている。

アメリカは国勢調査の Public Use Microdata Sample は，1980 年，1990 年，2000 年について，１％抽出と５％抽出サンプルを提供している。

以下は 2000 年について 2000 Census Population and Housing Technical Documentation をみたものである。

出典　https://www.census.gov/prod/cen2000/doc/pums.pdf （2017 年 1 月 15 日アクセス）

Ⅰ　アメリカ国勢調査 Public Use Microdata Sample 2000

貸与は，申請も，研究計画も必要がなく，ホームページから簡単にダウンロードできるという簡易な方法である。

利用のためのコードブックや注意等は，ホームページに掲載されている 724 ページの上記 Technical Document にある。

当局は個人が特定されないような十分な匿名化がされているとしている。その方法は，一つは近隣地域について，キー変数が同じ世帯の地域情報の入れ替えをしていること，世帯員が 10 名以上の場合には，匿名性を高めるために，世帯員の年齢をわずかにずらしていること，カテゴリー変数は全米で 1 万人以上いることを条件としてこれより対象者が少ない場合は合体していること，トップコーディングをしていること等である。産業分類，職業分類は 3 桁分類であり，また例えば賃金収入のトップコーディングが 175,000 ドルであるから 100 円で計算すれば，年収 1,750 万円以上でトップコーディング，しかしそれ以下は 1 ドル単位で開示している。また，世帯情報，続いて個人情報であるが，世帯主，その配偶者，他の世帯員情報という順で提供している。

その上で 1990 年のパブリックユースファイル作成時と比べて 2000 年になるとコンピュータ技術が一層発達したことから，匿名性を高めるために 1%抽出サンプルの地域情報は 40 万人以上という大きい地域単位にまとめたとする。そして１％抽出サンプルについては，1990 年データとほぼ連結できるような細かさで出しているが，一方，５％抽出サンプルの地域情報は 1990 年同様の 10 万人以上の地域単位にまとめているが，情報はやや粗くしたとある。かなり詳細な内容を出しているが，地域を広くすれば，個人が特定されないと考えているものと思われる。州を 40 万サンプルで分ける場合，例えばアーカンソー州は 4 地域となっていた。カリフォルニア州は人口が多いため，はるかに地域数は多くなっている。

入手できる情報は多岐にわたり，世帯については，概要であるが，例えば農業売り上げ，寝室数，家賃，水道電気代，家族収入，家族と続き柄，保険加入，世帯収入，世帯累計，台所の設備類，住宅ローン支払い，子供の年齢と人数，同居家族，固定資産税，建物の広さ，電話があるか，居住年数，空き家かどうか，家の価格，自動車があるかどうか，いつこの家に引っ越したか，何年に建築されたか，世帯ウェイト等。
個人については，英語が話せるか，年齢，人種，国籍，障害，1999 年の収入，学歴，ケアをしてくれる人としての祖父母の存在, 就業時間, 収入の種類, 産業, 自宅で使っている言語，婚姻状況，勤務先への交通手段，転居や移動の状況，1995 年に PUMA 単位でどこに住んでいたか，軍隊経験，PUMA 単位での勤務地，出勤時間，通勤時間，自動車を持っているか，1999 年に年間何週働いたか，1999 年の就業状況，何年から働き始めているか，個人ウェイト等である。
以下は国勢調査のパブリックユースファイルについての詳細である。

SUBJECT CONTENT

Public Use Microdata Sample (PUMS) files contain records representing 5-percent or 1-percent samples of the occupied and vacant housing units in the U.S. and the people in the occupied units. Group quarters people also are included. The file contains individual weights for each person and housing unit, which when applied to the individual records, expand the sample to the relevant total. Please see Chapters 6 and 7 - Data Dictionary for a complete list of the variables and recodes.

Some of the items included on the housing record are: acreage; agricultural sales; allocation flags for housing items; bedrooms; condominium fee; contract rent; cost of utilities; family income in 1999; family, subfamily, and relationship recodes; farm residence; fire, hazard, and flood insurance; fuels used; gross rent; heating fuel; household income in 1999; household type; housing unit weight; kitchen facilities; linguistic isolation; meals included in rent; mobile home costs; mortgage payment; mortgage status; plumbing facilities; presence and age of own children; presence of subfamilies in household; real estate taxes; rooms; selected monthly owner costs; size of building (units in structure); state code; telephone service; tenure; vacancy status; value (of housing unit); vehicles available; year householder moved into unit; and year structure built.

Some of the items included on the person record are: ability to speak English; age; allocation flags for population items; ancestry; citizenship; class of worker; disability status; earnings in 1999; educational attainment; grandparents as caregivers; Hispanic origin; hours worked; income in 1999 by type; industry; language spoken at home; marital status; means of transportation to work; migration Public Use Microdata Area (PUMA); migration state; mobility status; veteran period of service; years of military service; occupation; person's weight; personal care limitation; place of birth; place of work PUMA; place of work state; poverty status in 1999; race; relationship; school enrollment and type of school; time of departure for work; travel time to work; vehicle occupancy; weeks worked in 1999; work limitation status; work status in 1999; and year of entry.

地域

匿名性を確保するために，地域情報は，パブリックユースファイルでは州の中をいくつかにわけている。

匿名性のために，地域については大きくくくっている。1％抽出の州レベルのファイルでは，super PUMA は PUMA もしくは隣接 PUMA を合わせて 40 万人以上の地域にくくりなおしている。 5％抽出の州レベルデータでは，PUMA として地域を表彰している。PUMA は 10 万人以上の地域としてつくられている。PUMA は州をまたがっては設定されていない。州の中は一つか一つ以上の super PUMA または複数の PUMA で構成されている。大都市は複数の superPUMA や PUMA に分けられていることもある。地域情報は，1995 年 1 月の住居，2000 年の住居，勤務地の場所の 3 か所わかるようになっている。

GEOGRAPHIC CONTENT

The Public Use Microdata Sample (PUMS) files contain geographic units known as super-Public Use Microdata Areas (super-PUMAs) and Public Use Microdata Areas (PUMAs). To maintain the confidentiality of the PUMS data, minimum population thresholds are set for PUMAs and super-PUMAs. For the 1-percent state-level files, the super-PUMAs contain a minimum population of 400,000 and are composed of a PUMA or a group of contiguous PUMAs delineated on the 5-percent state-level PUMS files. Super-PUMAs are a new geographic entity for Census 2000. The 5-percent state-level files contain PUMAs, each having a minimum population of 100,000; the 5-percent files also will show corresponding super-PUMAs codes. Each state is separately identified and may be comprised of one or more super-PUMAs or PUMAs. Large metropolitan areas may be subdivided into super-PUMAs and PUMAs. PUMAs and super-PUMAs do not cross state lines. Super-PUMAs and PUMAs also are defined for place of residence on April 1, 1995 and place of work.

データの匿名化

データの匿名性を高めるため，１％抽出ファイルは詳細な内容がわかるかわりに，40 万人単でしか地理がわからない。５％抽出ファイルは 10 万人単位で地理がわかるが個人情報はより限定されている。

１％抽出ファイルでは，すべてのカテゴリーを示す。ただし，国全体で 8,000 人より少ない人種については示していない。

世帯員が 10 人以上の場合は，匿名性を高めるために，世帯員の年齢を少しずらしている。カテゴリー変数は，一定数以下の場合には合体している。分布がとても少ない場合にはトップコーディングをしている。

PROTECTING CONFIDENTIAL INFORMATION

All data released (in print or electronic media) by the Census Bureau are subject to strict confidentiality measures imposed by the legislation under which our data are collected: Title 13, U.S. Code. Responses to the questionnaire can be used only for statistical purposes, and Census Bureau employees are sworn to protect respondents' identities.

Because of the rapid advances in computer technology since 1990 and the increased accessibility of census data to the user community, the Census Bureau has had to adopt more stringent measures to protect the confidentiality of public use microdata through enhanced disclosure limitation techniques. At the same time, the Census Bureau recognizes the data user's need for characteristic detail and geographic specificity. Hence, there are two sets of files: one that provides a fuller range of detailed characteristics (the 1-percent files) and one that provides greater geographic detail but less characteristic detail (the 5-percent files).

Confidentiality is protected, in part, by the use of the following processes: data-swapping, topcoding of selected variables, geographic population thresholds, age perturbation for large households, and reduced detail on some categorical variables.

Data swapping is a method of disclosure limitation designed to protect confidentiality in tables of frequency data (the number or percent of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases. Swapping is applied to individual records and, therefore, also protects microdata.

Top-coding is a method of disclosure limitation in which all cases in or above a certain percentage of the distribution are placed into a single category.

Geographic population thresholds prohibit the disclosure of data for individuals or housing units for geographic units with population counts below a specified level.

Age perturbation, that is, modifying the age of household members, is required for large households (households containing ten people or more) due to concerns about confidentiality.

Detail for categorical variables is collapsed if the number of occurrences in each category does not meet a specified national minimum threshold.

データスワッピングの詳細

キー変数が同じ隣接する地域の世帯を入れ替える。

また，個人や世帯がわからないように，カテゴリーデータでは，全米で 1 万人以上いることを条件とする（例えば産業や職業分類や人種等と思われる）。また，地域変数は 10 万人以上（５％抽出），40 万人以上（１％抽出）の単位とする。トップコードをされたり，また実変数のかわりに平均値を入れたりしてある。

Data swapping. Data swapping is a method of disclosure limitation designed to protect confidentiality in data (the number or percentage of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases. A sample of households is selected and matched on a set of selected key variables with households in neighboring geographic areas that have similar characteristics. Because the swap often occurs within a neighboring area, there is usually no effect on the marginal totals for the area or for totals that include data from multiple areas. Data swapping procedures were first used in the 1990 census and were also used for Census 2000.

Since microdata records are the actual housing unit and person records, the Census Bureau takes further steps to prevent the identification of specific individuals, households, or housing units. The main disclosure avoidance method used is to limit the geographic detail shown in the files. A

Accuracy of the Microdata Sample Estimates 4-1

U.S. Census Bureau, Census 2000

minimum threshold of 10,000 for the national population was set for identification of groups within categorical variables in the state level PUMS files. A geographic area must have a minimum of 100,000 population to be fully identified in the 5 percent file, and 400,000 for the 1 percent sample file. Furthermore, certain variables are topcoded, or the actual values of the characteristics are replaced by a descriptive statistic, such as the mean.

トップコードの事例

賃金収入 1 ドルから 174,999 ドルまではそのまま示す。175,000 ドル以上はトップコード，自営業の場合 126,000 ドル以上はトップコード。

```
D INCWS          6                    244                  249
T Wage/Salary Income in 1999
V                                blank  . Not in universe (Under 15 years)
V                               000000  . No/none
R                       000001..174999  . $1 to $174,999
V                               175000  . Topcode
V                              175000+  . State mean of topcoded values

D INCWSA         1                    250                  250
T Wage/Salary Income in 1999 Allocation Flag
V                                     0  . Not allocated
V                                     1  . Allocated

D INCSE          6                    251                  256
T Self-Employment Income in 1999
V                                blank  . Not in universe (Under 15 years)
V                               -09999  . Loss of $9,999 or more
R                       -00001..-09998  . Loss of $1 to $9,998
V                               000000  . No/none
V                               000001  . $1 or break even
R                       000002..125999  . $2 to $125,999
V                               126000  . Topcode
V                              126000+  . State mean of topcoded values
```

平均値を入れる事例

年金収入 18,000 ドル以上でトップコード。トップコードについてはその階級の実額の州平均を入れているという意味か。

```
D INCSS          5                    265                  269
T Social Security Income in 1999
V                                blank  . Not in universe (Under 15 years)
V                                00000  . No/none
R                        00001..17999  . $1 to $17,999
V                                18000  . Topcode
V                               18000+  . State mean of topcoded values

D INCSSA         1                    270                  270
T Social Security Income in 1999 Allocation Flag
V                                     0  . Not allocated
V                                     1  . Allocated

D INCSSI         5                    271                  275
T Supplemental Security Income in 1999
V                                blank  . Not in universe (Under 15 years)
V                                00000  . No/none
R                        00001..13799  . $1 to $13,799
V                                13800  . Topcode
V                               13800+  . State mean of topcoded values
```

パブリックユースファイルは自分でさまざまなクロス集計をすることができる。データは，サンプル数の制限，地域情報の制限，匿名化措置をしたうえで，ほぼ Long Form のほとんどの内容がはいっているため，自分で全米の標本調査をしたように，ただしより正確でより大きい国勢調査データを使える。

USES OF MICRODATA FILES

Public use microdata files essentially allow "do-it-yourself" special tabulations. The Census 2000 files furnish nearly all of the detail recorded on long-form questionnaires in the census, subject to the limitations of sample size, geographic identification, and confidentiality protection. Users can construct a wide variety of tabulations interrelating any desired set of variables. They have almost the same freedom to manipulate the data that they would have if they had collected the data in their own sample survey, yet these files offer the precision of census data collection techniques and sample sizes larger than would be feasible in most independent sample surveys.

Microdata samples are useful to users who are doing research that does not require the identification of specific small geographic areas or detailed crosstabulations for small populations. Microdata users frequently study relationships among census variables not shown in existing census tabulations, or concentrate on the characteristics of specially defined populations.

2000 年の５％抽出サンプルは 1,400 万人，500 万の住宅の調査として提供されている。１％抽出では，280 万人，100 万の住宅の調査として提供されている。アメリカ全体の分析であれば，0.1%抽出の方が使いやすいかもしれない。0.1%抽出ファイルは作ってはいないが，特定変数を使うと，例えばサンプルを 1,000 に一つ等と小さくすることができる（subsampling）。

There are two independently drawn samples, designated "5 percent" and "1 percent," each featuring a different geographic scheme. Nationwide, the Census 2000 5-percent sample provides the user records for over 14 million people and over 5 million housing units. For the 1-percent sample, there are records for over 2.8 million people and over 1 million housing units. Since processing a smaller sample is less resource intensive, some users may want to produce extracts using the subsample numbers provided in the housing record. The sample design is discussed more thoroughly in Chapter 5. Sample Design and Estimation.

データは世帯情報，その中の世帯員情報が，まずは世帯主，その配偶者，その他の家族員，家族以外の者の順番で並ぶ形で提供されている。

欠損値がないように編集されており，欠損については数字をあてはめている。しかし，実データか，あてはめかはわかるようになっているので，実データのみでの集計は可能である。

提供されるデータの形式

ASCII フォーマット, ソフトウェア付きの DVD ではそれを使いクロス集計できる。また, 読み込みのための Dictionary ファイルも提供している。

SELECTION OF THE PUBLIC USE MICRODATA SAMPLES

A stratified systematic selection procedure with equal probability was used to select each of the public use microdata samples. The sampling universe was defined as all occupied housing units including all occupants, vacant housing units, and group quarters people in the census sample. The sample units were stratified during the selection process. The stratification was intended to improve the reliability of estimates derived from the public use microdata samples by defining strata, within which there is a high degree of homogeneity among the census sample households with respect to characteristics of major interest.

The occupied housing unit stratification was performed using a matrix containing 34,080 cells made by combining 71 race groups, 5 Hispanic origin groups, 3 family types, 2 tenure groups, 4 groups based on maximum age of household members, and the 4 long form sampling rates. In the case of occupied housing units the primary sampling units selected by the systematic selection process are housing units and all person records are extracted after the housing units are chosen. Therefore, the race and Hispanic origin correspond to the householder. The maximum age variable, in contrast, can come from any household member. For group quarters people, the race, Hispanic origin, and age will be those of the individual group quarters person. Table A contains a representation of the occupied housing unit stratification matrix.

The vacant housing unit stratification was performed within a matrix consisting of 12 cells made by combining the four long form sampling rates with three vacancy statuses. Table B contains a representation of the vacant housing unit stratification matrix.

The group quarters stratification used a matrix of 2,840 cells made by combining 71 race groups, five Hispanic Origin groups, four group quarters person age groups, and two types of group quarters. Table C contains a representation of the group quarters person stratification matrix.

産業コードの例（一部のみ）

DETAILED INDUSTRY CODE LIST

1997 NAICS and Census 2000 sorted by 1997 NAICS codes and subsequent OMB directives
(Census codes may not be in sequential order)

NAICS Based Census 2000 Category Title	Census 2000	1997 NAICS Equivalent
Agriculture, forestry, fishing and hunting, and mining:	001-056	11, 21
Agriculture, forestry, fishing and hunting:	001-036	11
Unused codes	001-016	
Crop production	017	111
Animal production	018	112
Forestry except logging	019	1131, 1132
Unused codes	020-026	
Logging	027	1133
Fishing, hunting, and trapping	028	114
Support activities for agriculture and forestry	029	115
Unused codes	030-036	
Mining:	037-056	21
Oil and gas extraction	037	211
Coal mining	038	2121
Metal ore mining	039	2122
Unused codes	040-046	
Nonmetallic mineral mining and quarrying	047	2123
Not specified type of mining	048	Part of 21
Support activities for mining	049	213
Unused codes	050-056	
Utilities census codes 057-076 moved to Transportation and Warehousing NAICS subsector 48-49		
Construction:	077-106	23
Construction	077	23
Unused codes	078-106	
Manufacturing:	107-406	31-33

職業コードの例（一部のみ）

OCCUPATION DETAILED CODE LIST

Decennial 2000 SOC and Census 2000 sorted by Census 2000 SOC equivalent

SOC Based Census 2000 Category Title	Census 2000	2000 SOC Equivalent
Management, professional and related occupations:	001-359	11-0000 through 29-0000
Management, business and financial operations occupations:	001-099	11-0000 and 13-0000
Management occupations:	001-049	11-0000
Chief executives	001	11-1011
General and operations managers	002	11-1021
Legislators	003	11-1031
Advertising and promotions managers	004	11-2011
Marketing and sales managers	005	11-2020
Public relations managers	006	11-2031
Unused codes	007-009	
Administrative services managers	010	11-3011
Computer and Information Systems managers	011	11-3021
Financial managers	012	11-3031
Human resources managers	013	11-3040
Industrial production managers	014	11-3051
Purchasing managers	015	11-3061
Transportation, storage, and distribution managers	016	11-3071
Unused codes	017-019	
Farm, ranch, and other agricultural managers	020	11-9011
Farmers and Ranchers	021	11-9012
Construction managers	022	11-9021
Education administrators	023	11-9030
Unused codes	024-029	
Engineering managers	030	11-9041
Food service managers	031	11-9051
Funeral directors	032	11-9061
Gaming managers	033	11-9071

移転コード

州単位。海外も。

Code	Description
050	Vermont
051	Virginia
052	Not used
053	Washington
054	West Virginia
055	Wisconsin
056	Wyoming
057-059	Not used
060	American Samoa (See code 555)
061-065	Not used
066	Guam (See code 555)
067	Johnston Atoll (See code 555)
068	Not used
069	Northern Marianas (See code 555)
070	Not used
071	Midway Islands (See code 555)
072	Puerto Rico (See code 555) *
073-075	Not used
076	Navassa Island
077	Not used
078	U.S. Virgin Islands (See code 555)
079	Wake Island (See code 555)
080	Not used
081	Baker Island
082-083	Not used
084	Howland Island
085	Not used
086	Jarvis Island
087-088	Not used
089	Kingman Reef
090-094	Not used
095	Palmyra Atoll
096	U.S. Island Area not specified (See code 555)
097-099	Not used
100	Albania (See code 555)
101	Andorra
102	Austria (See code 555)

トップコードをした場合のトップコード以上に入る人々についての州平均の開示

Appendix H.
Topcoded Variables and Corresponding State Means for Values at and Above the Topcode for the Housing Record and Person Record

Table 1. **Topcoded Variables for the 1-Percent PUMS Housing Record**

States	Elec	Gas	Water	Oil	Rent	Mrt1 amt	Mr2t amt	Insamt	Condo fee	Mhcost
Topcode:										
United States	4800	3000	2000	2100	1700	3000	1100	2500	720	10000
Corresponding state means for values at and above the topcode:										
Alabama	5600	4500	2900	3300	2100	4600	1400	3100	0	14900
Alaska	5700	3900	2600	2900	1900	3500	1300	2800	0	11900
Arizona	5600	4100	2700	2900	2200	4000	1800	3400	1100	12400
Arkansas	5700	4500	2900	3200	2200	4000	1400	3500	1300	14200
California	5700	4200	2700	4000	2100	4100	1800	3500	1100	13200
Colorado	5800	4200	2600	3300	2100	4100	1600	3400	1100	11900
Connecticut	5700	3900	2700	2900	2500	4400	1700	3500	1300	13000
Delaware	5700	4000	2800	2600	2400	4000	1800	2900	1200	13800
District of Columbia	6000	4200	2900	2200	2200	4100	1400	3400	910	0
Florida	5600	4300	2700	3800	2200	4200	1800	3300	1100	13400
Georgia	5600	4100	2900	3300	2200	4000	1700	3300	1000	13300
Hawaii	5800	4400	2600	0	2200	4000	1700	3300	1000	0
Idaho	5800	4500	3400	3200	2100	4100	1800	3000	780	10700
Illinois	5800	4100	2800	3500	2100	3900	1600	3200	1000	12800
Indiana	5700	4300	3000	2800	2200	3900	1700	3200	1000	14000
Iowa	5600	4300	2900	3700	2200	3800	1400	3200	1400	13800
Kansas	5600	4300	3000	5000	2300	3800	1500	3100	1200	11800
Kentucky	5700	4600	3000	3000	2100	3900	1700	3000	880	14100
Louisiana	5800	4200	2900	4200	2000	3700	2000	3400	1300	14600
Maine	5700	4100	2900	3000	2300	3200	1200	2900	0	15400
Maryland	5600	4200	2700	3200	2200	3800	1700	3300	830	13500
Massachusetts	5700	3700	2500	2900	2100	4000	1700	3300	1200	13300
Michigan	5800	4000	2800	2700	2100	4200	1700	3300	1100	12400
Minnesota	5700	4000	3100	3000	2100	3800	1500	3200	1100	12100
Mississippi	5800	4400	2900	3100	2300	3600	1400	3100	1200	13600
Missouri	5900	4100	2800	3600	2300	3800	1700	3300	1000	12100
Montana	5800	3900	3700	2600	2000	5100	1200	3600	0	12600
Nebraska	5700	4300	2800	2900	1900	4100	1400	3500	1200	12700
Nevada	5700	4000	2800	2600	2100	4200	1700	3600	1000	12700
New Hampshire	5500	3800	2700	2900	2000	3500	1400	3300	960	13100

年齢，通勤時間，収入等のトップコード

Table 2. **Topcoded Variables for the 1-Percent PUMS Person Record**

States	Age	Trvtime	Incws	Incse	Incint	Incss	Incssi	Incpa	Incret	Incoth
Topcode:										
United States	90	120	175000	126000	50000	18000	13800	12300	52000	37800
Corresponding state means for values at and above the topcode:										
Alabama	93	171	344000	255000	150000	26300	18000	20500	169000	64000
Alaska	94	175	323000	227000	169000	24800	23300	18000	150000	47200
Arizona	93	174	318000	249000	132000	24000	17100	21500	110000	65000
Arkansas	93	170	320000	266000	147000	25700	18400	20800	123000	65000
California	93	156	330000	263000	134000	23800	17800	19400	117000	63000
Colorado	92	170	333000	222000	113000	24600	18300	21100	102000	63000
Connecticut	93	156	342000	225000	131000	22900	18100	23400	105000	63000
Delaware	92	168	326000	188000	146000	23500	17900	16800	107000	52000
District of Columbia	93	169	343000	290000	203000	26100	16700	27000	86000	75000
Florida	93	170	324000	248000	134000	23900	17900	20500	102000	62000
Georgia	93	171	319000	243000	131000	25400	19700	21000	99000	61000
Hawaii	93	157	318000	195000	132000	24000	16200	17800	95000	59000
Idaho	92	172	297000	280000	180000	24000	17600	14100	109000	68000
Illinois	93	161	322000	266000	131000	23700	18500	23200	135000	61000
Indiana	93	169	324000	264000	143000	24200	19700	25800	176000	58000
Iowa	93	172	311000	224000	136000	27300	15800	20300	162000	58000
Kansas	93	169	319000	192000	118000	23400	17300	21000	169000	58000
Kentucky	93	169	332000	238000	124000	26200	19600	25800	205000	67000
Louisiana	93	172	319000	227000	120000	25000	20200	19000	143000	56000
Maine	93	163	298000	306000	160000	24600	20900	19800	132000	81000
Maryland	93	163	328000	232000	156000	23700	18500	18700	81000	62000
Massachusetts	93	165	310000	249000	111000	23700	16900	19900	117000	68000
Michigan	93	171	305000	261000	129000	24200	16200	19700	172000	61000
Minnesota	93	170	336000	223000	134000	23900	17000	23300	93000	62000
Mississippi	93	174	329000	232000	118000	23400	17600	30000	159000	62000
Missouri	93	171	312000	198000	121000	24600	18300	23800	157000	67000
Montana	92	170	229000	228000	146000	24700	18300	26100	146000	58000
Nebraska	92	162	340000	233000	94000	23600	16400	20100	115000	60000
Nevada	92	170	322000	310000	135000	23500	17500	22700	92000	61000
New Hampshire	93	159	294000	227000	117000	24500	17700	27100	97000	67000

サンプル数

Appendix I.
Total Unweighted and Weighted Population and Housing Counts

Control Counts for the 1-Percent PUMS Files

State	Total population unweighted	Total housing unweighted (includes pseudo-housing units)	Total population weighted	Total housing weighted
Alabama	44487	20782	4445562	1963448
Alaska	6422	2803	628493	261389
Arizona	51901	22990	5129713	2189281
Arkansas	26978	12471	2675687	1173605
California	338725	130341	33879320	12217313
Colorado	43135	19113	4301983	1808330
Connecticut	34118	14942	3406431	1386039
Delaware	7786	3676	783683	343123
District of Columbia	5770	3105	572781	274971
Florida	159704	76920	15985411	7303716
Georgia	81446	35158	8186026	3282330
Hawaii	12218	4962	1211064	460544
Idaho	13112	5595	1293454	527912
Illinois	123613	52080	12420669	4885588
Indiana	60669	27106	6076995	2531689
Iowa	29212	13367	2923524	1231936
Kansas	26767	12132	2687848	1131091
Kentucky	40217	18659	4041737	1750944
Louisiana	44538	19833	4470170	1847377
Maine	12877	6869	1274571	651752
Maryland	52764	22793	5297739	2146221
Massachusetts	63760	28436	6349715	2622935
Michigan	99184	44841	9934066	4233707
Minnesota	49780	22018	4920116	2065465
Mississippi	28446	12574	2845775	1162028
Missouri	56051	26044	5592082	2441456
Montana	9151	4375	902423	412894
Nebraska	17161	7736	1710928	722261
Nevada	20065	8612	1997348	827116
New Hampshire	12430	5826	1236607	547217
New Jersey	84117	35053	8415300	3310587

地域コードの例

アーカンソー州で５地域，カルフォルニア州はかなり多い。

		移動の Super PUMA	Super PUMA
05	Arkansas	05100	05100
		05200	05200
		05300	05300
		05400	05400
		05500	05500
06	California	06010	06010
		06020	06020
		06030	06030
		06040	06040
		06050	06050
		06069	06060-06072
		06080	06080
		06090	06090
		06100	06100
		06110	06110
		06120	06121-06122
		06130	06130
		06140	06140
		06150	06151-06153
		06160	06161-06163
		06170	06170
		06180	06180
		06190	06190
		06200	06201-06203
		06210	06210
		06220	06220
		06230	06230
		06309	06301-06411

付録２　アメリカにおいてパブリックユースファイルが作られているその他の調査の事例

カレント・ポピュレーション・サーベイ（Current Population Survey，労働力調査）

コンピュータ補助つきの面接調査（CAPI）およびコンピュータ補助つきの電話調査（CATI）でもって毎月１週間の状況を調査する労働力調査である。また，３月追加調査では，仕事経験，年収，移動を，また 10 月追加調査では学校教育を調査している。このほかに他の追加調査が他の官庁等により行われる。統計局および労働調査局が共同して実施している。２週間後に速報が，６週間語に書く法が出る。追加調査についての結果は３～６か月後に速報が，１年から 18 か月後に確報が出る。パブリックユースファイルはデータ収集ののち６か月から１年後に利用できるようになっている。

Purpose: To provide estimates of employment, unemployment, and other characteristics of the general labor force, of the population as a whole, and of various subgroups of the population. Monthly labor force data for the country are used by the Bureau of Labor Statistics (BLS) to determine the distribution of funds under the Job Training Partnership Act. These data are collected through combined computer-assisted personal interviewing (CAPI) and computer-assisted telephone interviewing (CATI). In addition to the labor force data, the CPS basic funding provides annual data on work experience, income, and migration from the March Annual Demographic Supplement and on school enrollment of the population from the October Supplement. Other supplements, some of which are sponsored by other agencies, are conducted biennially or intermittently.

Sponsoring agencies and legal authorities: The U.S. Census Bureau and the BLS jointly sponsor the survey under the authorities of Title 13, United States Code, Section 182, and Title 29, United States Code, Sections 1-9.

Periodicity: A continuing survey with interviewing conducted during one week of each month.

Release of results: The first release of monthly employment data by the BLS occurs approximately two weeks after completion of data collection. The final report, Employment and Earnings, is published by the BLS approximately six weeks after data collection. On a quarterly basis, earnings data for people in the labor force are published in the form of a press release, and characteristics of people not in the labor force are published in Employment and Earnings. Advance reports on supplement data are usually released approximately 3 to 6 months after data collection; final reports for supplements are typically released within one year to 18 months. Public use microdata files are made available within six months to one year after data collection.

Supplements: Additional information about specific supplements can be found on the Census Bureau website at http://www.census.gov/cps/methodology/techdocs.html