製表技術参考資料 10

ISSN 1348－7345

ミクロアグリゲーションに関する研究動向

及び

匿名化技法としてのミクロアグリゲーションの有効性に関する研究

－全国消費実態調査を例に－

N S T A C

Working Paper No.10

平成 20 年 9 月

独立行政法人 統計センター

製表技術参考資料は、独立行政法人 統計センターの職員がその業務に関連して行った製表技術に関する研究の結果を紹介するためのものである。

ただし、本資料の内容は執筆者の個人的見解を示すものであり、機関の見解を示すものではない。

目　次

序　文

現在、統計センターでは、平成21年度の新統計法の全面施行を受け、各府省より委託された場合の匿名データ提供事務における準備として、個別データの秘匿性を保持するための匿名化技法の開発及びその有効性の検証に着手する必要があると考え、匿名化技法の研究を進めている。

個別データに対する匿名化技法については、トップ・コーディング、ボトム・コーディング、スワッピング等、多くの手法が存在しており、諸外国の統計作成部局では、様々な匿名化技法が個別データに適用されていることが知られている。本研究でミクロアグリゲーションに着目した理由の１つとして、ミクロアグリゲーションが集計結果表の秘匿処理の考え方に着想を得ていることがあげられる。集計結果表の作成においては、セルに度数１又は２が存在した場合、個体が特定される可能性を回避するため、度数をＸに置き換える等の秘匿処理を行うことがある。この集計表の秘匿処理に関する実務上の経験に基づいて、政府統計の個別データに対する秘匿処理の方法を具体的に追究する場合、ミクロアグリゲーションは、個別データに対する秘匿処理の有力な方法の１つとして検討に値するのではないかと考え、この研究を行うこととした。

本報告書では、まず、主としてヨーロッパ諸国で展開されている匿名化技法に関する研究の１つである「ミクロアグリゲーション(microaggregation)」に関する研究動向を紹介(「Ⅰ　ミクロアグリゲーションに関する研究動向」）し、次に全国消費実態調査の個別データを用いて、ミクロアグリゲーションの有効性の検証結果(「Ⅱ　匿名化技法としてのミクロアグリゲーションの有効性に関する研究　―全国消費実態調査を例に）について述べている。なお、全国消費実態調査のデータを用いた理由は、主として量的属性に対する秘匿処理の方法として位置付けられているミクロアグリゲーションの有効性を検証するには、数多くの量的属性を調査項目として持つデータが適していると考えたためである。

匿名データの提供を考える場合、諸外国の事例を見ると、個別データの有用性を考慮しながらも秘匿性を確保することが求められている。我が国においても同様の視点が必要と考えられるが、今回の検証においては、主に個別データのミクロアグリゲーションについて有用性の視点から述べているものである。

ミクロアグリゲーションに関する研究動向

伊藤　伸介*

要　旨

近年、ヨーロッパ諸国を中心に「ミクロアグリゲーション(microaggregation)」による個別データに準じたレベルのデータの作成が注目されている。ミクロアグリゲーションとは、ミクロデータ(個別データ)を閾値ｋ個のレコードをそなえた同質的なレコード群にグループ化した上で、そのレコードにおける個々の属性値を平均値等の代表値に置き換えることであり、政府統計のミクロデータに対する匿名化技法の１つに位置付けられている。本稿では、ミクロアグリゲーションにおける研究動向を概観し、その基本的な特徴を考察した。

ミクロアグリゲーションについては、主として、①個別データに含まれる属性の性質、及び②レコードをグループ化する場合の基準となるレコード数の設定方法の観点から、類型化を図ることが可能である。量的属性のミクロアグリゲーションに関しては、単一軸法、第１主成分法、Ｚスコア総計法、個別ランキング法といった手法が展開されている。また、階層区分法といった探索的なミクロアグリゲーションに関する研究が進められている。さらに、質的属性のミクロアグリゲーションについても、スネーク法等の手法が存在することが知られている。

本稿では、ミクロアグリゲーションにおける先行研究を踏まえながらも、それとは異なる視点からミクロアグリゲーションの方法を追究している。それは、ミクロアグリゲーションの中に個別データに含まれるすべての属性群を集計事項の対象とした「超高次元クロス集計表」を位置付けたことにある。本稿では、次のように議論を展開している。第１に、個別データから作成された超高次元クロス集計表は、個別データに含まれる属性群のおのおのについて同一の属性値を有するレコード群(同質属性値レコード群)として捉えられることから、そこから個別に準じたレベルのデータを編成することが可能である。第２に、超高次元クロス集計表に含まれるセルの度数は、同質属性値レコード群内のレコード数と対応関係にある。よって、超高次元クロス集計表に基づいて、セルの中に度数１又は２が存在しないように集計表を作成した場合、そこから同質属性値レコード群内の最低３以上のレコード数を含むミクロアグリゲーション済みのデータ(ミクロアグリゲートデータ)を作成することができる。このような議論を踏まえて、本稿では、質的属性においては超高次元クロス集計表をもとに同質属性値レコード群を編成し、量的属性については、同質属性値レコード群内の属性値を平均値で置き換えることによって、ミクロアグリゲートデータを作成することを提案している。

* 統計センター情報技術部研究主幹非常勤職員(明海大学経済学部専任講師)

ミクロアグリゲーションに関する研究動向

伊藤伸介

１　はじめに

欧米諸国においては、1960 年代以降、政府統計のミクロデータの提供が広範に展開されてきた。諸外国では、一般公開型ミクロデータ(Public use microdata file)の公開、オンサイト施設の利用、個別契約方式によるミクロデータの提供、オーダーメイド集計等の様々な方法によって、政府統計ミクロデータの利用促進がはかられてきた。また、諸外国の統計作成部局では、政府統計のミクロデータに対して、様々な匿名化技法が適用されるだけでなく[1]、個体情報の保護に関する法的制度的措置が整備されてきた[2]。それによって、個体情報の秘匿性を確保するだけでなく、個別データの有用性を保持する形で、政府統計ミクロデータの提供が進められてきた。

ところで、近年、ヨーロッパ諸国を中心に「ミクロアグリゲーション(microaggregation)」による個別データに準じたレベルのデータの作成が注目されている。ミクロアグリゲーションは、政府統計ミクロデータに対する匿名化技法の１つと位置付けられている(Willenborg

---

1 諸外国の統計作成部局は、政府統計のミクロデータを提供するために、いろいろな匿名化技法を用いている。例えば、Federal Committee on Statistical Methodology(2005)によれば、アメリカセンサス局等の政府当局が、一般公開型ミクロデータを提供するために採用する基本的な匿名化技法は、次の４つとされている(Federal Committee on Statistical Methodology(2005, p.24))。
①標本データによるミクロデータの作成
②明示的な識別子(名前、住所等)の削除
③詳細な地域情報の制限
④属性群における分類区分数の限定
また、個体が特定される危険が高い属性(例えば所得等)については、上記の４つの方法だけでなく、次のような匿名化技法を追加的に導入することが考えられている(Federal Committee on Statistical Methodology (2005, p.25))。
①トップコーディング、ボトムコーディング
②分類区分の再符号化(recoding)(あるいは丸め込み(rounding))
③乱数(ノイズ)の導入
④データ・スワッピング(ある地域において人口社会的属性によって個体が特定される危険の高い世帯を別の地域の世帯と入れ替えること)あるいはランク・スワッピング(スイッチング(switching)とも呼ぶ)
⑤ランダムに選別されたレコードにおける変数値の削除(blank)と補定(impute)
⑥複数のレコード群における属性値の集計値(平均値)による置き換え(ブラーリング(blurring))

2 例えば、アメリカでは、連邦統計方法委員会(Federal Committee on Statistical Methodology)の秘匿・データアクセス委員会(the Confidentiality and Data Access Committee)が、ミクロデータにおける個体情報の漏洩の可能性を確認するために、1999 年に「データの公開における潜在的な露見(disclosure)の可能性についてのチェックリスト(Checklist on Disclosure Potential of Proposed Data Releases)」を開発している。このチェックリストは、現在、アメリカセンサス局、アメリカ労働統計局、アメリカ国立保健統計センター(National Center for Health Statistics)といった多くの統計作成機関でミクロデータの提供において開示リスクを検証するための基準として採用されている((Federal Committee on Statistical Methodology(2005, p.24))。また、アメリカセンサス局では、開示評価委員会(Disclosure Review Board)が設置されており、匿名化措置に関するチェックリスト等を用いて、センサス局で作成される政府統計ミクロデータの提供可能性を検討している(Zayatz(2007, p.255))。

and de Waal(2001, pp.30-31))[3]。ミクロアグリゲーションの研究は少なくとも 1980 年代に遡ることができる。Strudler *et al.*(1986)は、アメリカ内国歳入庁(Internal Revenue Service)が提供する所得税申告書(tax return)のミクロデータ(Tax Model)に対して、レコードのグループ化とグループ内における属性値の平均値への置き換え(ブラーリング(blurring)と呼ばれている）による秘匿処理を提唱し、その方法の有効性を検証している。また、Wolf(1988)は、アメリカセンサス局によって作成された事業所データに関する縦断的研究開発ファイル(Longitudinal Research Development file)に対する匿名化技法として、Spruill(1983)の研究に基づきミクロアグリゲーションの手法を追究している。Eurostat では、Strudler *et al.*(1986)の研究に着想を得て、90 年代初頭よりミクロアグリゲーションの調査研究を進めてきた(Defays(1997, pp.223-224))。そして、ヨーロッパの企業におけるイノベーションの活動状況を調査した Community Innovation Survey(1994)においては、匿名化技法の１つとしてミクロアグリゲーションが適用されている（Thorogood(1999))[4]。Eurostat が資金提供している CENEX プロジェクト(Hundepool(2006))では、ミクロデータの匿名化を行うソフトウェアμ-ARGUS が開発されており、μ-ARGUS を用いてミクロアグリゲーションを行うことが可能になっている[5]。イタリアにおいては、イタリア統計局が企業のミクロデータを対象にしたミクロアグリゲーションの研究を進めており(Pagliuca and Seri(1998)等)、System of Enterprises Accounts Annual Survey を用いた企業データの一般公開型ファイル(Public Use File)の作成を試みている(Pagliuca and Seri(1999, p.304))。さらに、イタリア統計局は、

---

[3] 近年、国際連合欧州経済委員会(United Nations Economic Commission for Europe(UNECE))は、諸外国の統計作成機関における匿名化措置の状況を把握するために、東欧諸国や旧ソ連諸国を対象に統計データの秘匿措置の現状について調査を行っている(Felsö *et al.*(2001))。付表(本稿 30 頁)は、対象となった統計作成機関の秘匿措置に関する調査結果を示したものであって、人口・社会統計と経済統計のそれぞれについて、秘匿措置の現状が把握される。付表を見ると、人口・社会統計と経済統計のいずれのミクロデータについても、匿名化措置として、データ項目の削除、分類区分の再符号化、標本抽出と並んでミクロアグリゲーションが用いられていることがわかる。

また、付図(本稿 31 頁)は、アメリカ、カナダ、ドイツ、オランダ等の 13 カ国の統計機関を対象に、人口センサス、人口・社会統計、経済統計のミクロデータに対して秘匿措置の現状に関する調査結果を示したものである（Felsö *et al.*(2001))。付図においては、標本抽出、識別子の削除、地域区分の制限、属性群における分類区分の制限が、匿名化技法として主に適用されていることがわかる。また、Federal Committee on Statistical Methodology(2005)では、ミクロアグリゲーションが、ブラーリングの一形態として位置付けられているが(Federal Committee on Statistical Methodology(2005, p.91))、ブラーリングを匿名化の方法として採用している統計機関が存在していることも、付図から明らかになっている。

[4] Thorogood(1999, pp.31-32)によれば、Community Innovation Survey(CIS)については、Eurostat やその傘下にある国家統計機関に所属していない外部の研究者に対してデータを提供することが指向されており、そのための匿名化措置として、CIS のデータにミクロアグリゲーションを適用することが定められた。しかしながら、実際の提供においては、個別企業の識別の禁止等に関する契約を結んだ上で、承認された(bona fide)研究者のみが、ミクロアグリゲーション済みの CIS データを提供されている。

[5] μ-Argus やτ-Argus を中心とした秘匿処理のソフトウェアの開発に関する研究動向については、Hundepool(2006)、瀧(2003，345～347 頁)等を参照されたい。

MASQ(Single Axis Microaggregation for Quantitative variables)と呼ばれるソフトウェアを開発することによって、企業ミクロデータへの適用可能性を模索している(Pagliuca and Seri(1999))。アメリカでも、アメリカセンサス局が、秘匿の観点からミクロアグリゲーションに基づいて作成されたデータの再識別の可能性に関する実験を試みている(Winkler(2002))。最近でも、統計データの秘匿措置に関するワークセッション(Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, 2007)の中で、ミクロアグリゲーションの研究成果が報告されている[6]。

このように、諸外国では、匿名化技法としてのミクロアグリゲーションの研究が展開されていることから、本報告では、ミクロアグリゲーションにおける研究動向を概観し、その基本的な特徴を明らかする。

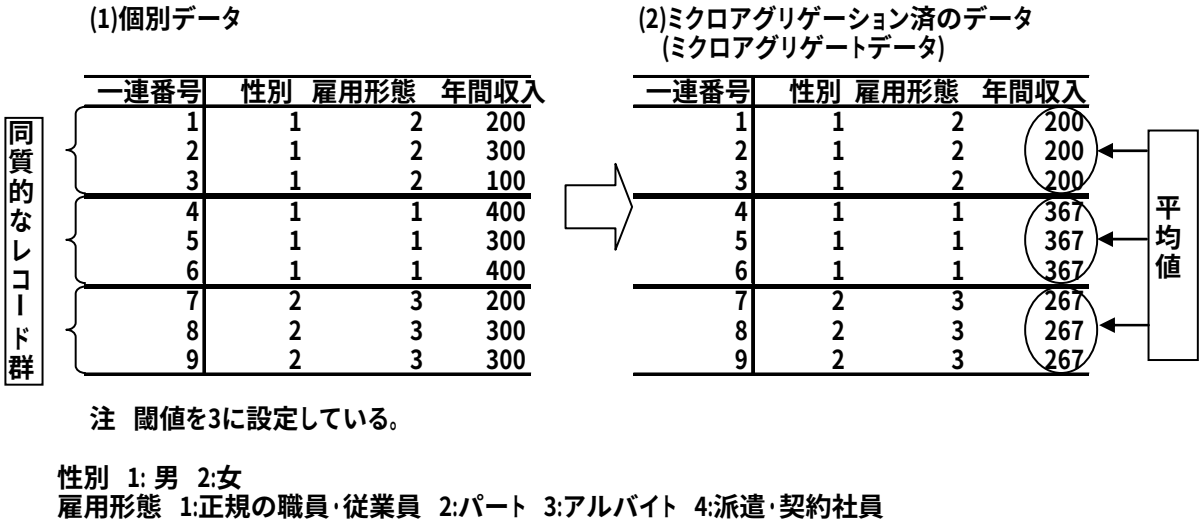
２　ミクロアグリゲーションの基本的特徴

ミクロアグリゲーションとは、ミクロデータ(個別データ)を k 個(k は閾値(threshold))のレコードを有する同質的なレコード群にグループ化した上で、そのレコードにおける個々の属性値を平均値等の代表値に置き換えることである(Domingo-Ferrer and Mateo-Sanz (2002, p.190))。例えば、属性群として性別、雇用形態と年間収入のみを持つ個別データを考えることにし、閾値を３に設定したとする(図１)。このデータ上にある属性群にミクロアグリゲーションを適用するということは、性別、雇用形態と年間収入の属性値のおのおのについて同質的であるとみなされるレコードを少なくとも３レコードずつグループ化し、各グループ内のレコードが持つ属性値を平均値等の代表値に変換することを意味している。図１では、最初に、性別と雇用形態に関して同一の属性値が選ばれるようにグループ化することによって同質的なレコード群が編成され、次に、各グループ内で年間収入を平均値に置き換えることによって、ミクロアグリゲーション済のデータ(以下「ミクロアグリゲートデータ(micro-aggregated data)と呼ぶ」)が作成されることが示されている。

ところで、ミクロアグリゲーションの方法については、主として次の２つの観点から整理することが可能だと考えられる。第１の観点は、個別データに設定される属性の性質に関する区分である。個別データに含まれる属性群は、年間収入や消費支出等といった数値項目を表す量的属性、及び性別や学歴といった分類項目を示す質的属性に大別される。先行研究に

[6] ミクロアグリゲーションに関する研究報告については、例えば Domingo-Ferrer, Sebé and Solanas (2007) 等を参照されたい。

図１　ミクロアグリゲーションのイメージ



注　閾値を3に設定している。

性別　1: 男　2:女
雇用形態　1:正規の職員･従業員　2:パート　3:アルバイト　4:派遣･契約社員

よれば、ミクロアグリゲーションの手法は、主として量的属性を対象とした匿名化技法として方法的に位置付けられてきたが(Domingo-Ferrer and Torra (2001))、近年、質的属性に関するミクロアグリゲーションについても実証的な研究が進められている(Torra(2004, p.224))。第２の観点は、レコードをグループ化する場合の基準となるレコード数の設定方法についてである。閾値に基づきながらも、グループ化の基準となるレコード数を固定的に設定した場合と(Defays(1997))、探索的な(heuristic)方法でグループ内のレコード数を定める場合とでは(Domingo-Ferrer and Mateo-Sanz (2002, p.192))、ミクロアグリゲーションの適用の仕方が大きく異なると考えられる。本節では、これらの点に着目し、ミクロアグリゲーションの概要を述べる。

(1)量的属性に関するミクロアグリゲーション

量的属性に関するミクロアグリゲーションについては、グループ化の基準となるレコード数を固定的に設定した場合に、量的属性値に対する加工の仕方によって、単一軸法(single axis method)、第１主成分法(first principal component method)、Ｚスコア総計法(sum of Z-scores method)、個別ランキング法(individual ranking method)等の手法が存在する(Anwar(1993))。以下で、各方法の概要を述べる。

①単一軸法

単一軸法では、ある特定の量的属性(ソートキー)に着目し、属性値を昇順又は降順にソートし、ソートされたレコード群を一定のレコード数ごとにグループ化した上で、グループ内

のレコードが有するそれぞれの量的属性値を平均値等の代表値に変換する。図２では、雇用者数、総売上高と店舗の数の３つの属性を含むレコード群を想定している。最初に、雇用者数に基づいてレコード群のソートが行われる。次に、グループ化の基準となるレコード数(図２ではレコード数を３に設定)にしたがってレコード群のグループ分けを行った後に、各グループ内のレコードに含まれる属性値が平均値に置き換えられる[7]。

②第１主成分法

単一軸法では、ある特定の属性に着目してレコード群のソートが行われるために、どの属性をソートキーとして選択するかによって、レコードの並び順が大きく変わる可能性がある。そこで、レコードが持つ属性群から統計指標を新たに作成し、その統計指標に基づいてソートを行うことが考えられる。これについては、主に２つの方法が存在するが、その１つが第１主成分法である。第１主成分法は、ミクロアグリゲーションに主成分分析を適用した方法である。図２では、基準となるレコード数を３に設定した場合に、雇用者数、総売上高、店舗の数の３つの属性値を標準化し、第１主成分のスコアを計算した上で、レコード群のソート、及びレコードのグループ化が行われている。

③Z スコア総計法

単一の統計指標によるソートの第２の方法が、Z スコア総計法である。Z スコア総計法は、各レコードにおける属性値群を標準化し、標準化された値の総計値(Z スコア総計値)に基づいてレコード群をソートし、レコードのグループ化を行う手法である。図２では、雇用者数、総売上高と店舗の数の属性値から算出された Z スコア総計値によって、レコード群がソートされている。

④個別ランキング法

個別ランキング法は、先述した単一軸法、第１主成分法と Z スコア総計法とは大きく異なる特徴を有している。単一軸法、第１主成分法、及び Z スコア総計法においては、ある単一の属性あるいは統計指標をソートキーとしてレコード群のソートが行われる。それに対して個別ランキング法は、量的属性のおのおのについて個別にソート化とグループ化を行う方法である。図３は、図２と同様に、雇用者数、総売上高と店舗の数を例に、個別ランキング法の概要を示したものである。最初に雇用者数をソートキーにしてレコード群をソートし、次に基準となるレコード数にしたがってレコードがグループ化され、レコードが有する属性値

[7] 我が国の政府統計の個別データの多くは、レコードが都道府県、市区町村といった地域順に並べられている。このような地域属性をソートキーとみなして、ミクロアグリゲーションを行うことも考えられる。

図２　単一の量的属性におけるミクロアグリゲーション

グループ化

一連番号	雇用者数	総売上高	店舗の数
1	12	1000	2
2	21	1500	6
3	39	2000	5
4	40	3000	3
5	42	1000	4
6	47	2000	10
7	53	1500	11
8	58	1500	10
9	60	3000	14

平均値

一連番号	雇用者数	総売上高	店舗の数
1	24	1500	4
2	24	1500	4
3	24	1500	4
4	43	2000	6
5	43	2000	6
6	43	2000	6
7	57	2000	12
8	57	2000	12
9	57	2000	12

②第1主成分法の適用

一連番号	雇用者数	総売上高	店舗の数	第1主成分のスコア
1	12	1000	2	-2.4516
2	21	1500	6	-1.1941
3	39	2000	5	-0.322
4	40	3000	3	0.0285
5	42	1000	4	-0.9596
6	47	2000	10	0.7402
7	53	1500	11	0.8237
8	58	1500	10	0.874
9	60	3000	14	2.4611

適用後

一連番号	雇用者数	総売上高	店舗の数	グループの番号
1	25	1167	4	1
2	25	1167	4	1
3	42	2333	6	2
4	42	2333	6	2
5	25	1167	4	1
6	42	2333	6	2
7	57	2000	12	3
8	57	2000	12	3
9	57	2000	12	3

③Zスコア総計法の適用

一連番号	雇用者数	総売上高	店舗の数	雇用者数のZ値	総売上高のZ値	店舗の数のZ値	Zスコア総計値
1	12	1000	2	-1.82093	-1.11111	-1.25939	-4.19413
2	21	1500	6	-1.26233	-0.44444	-0.29475	-2.00143
3	39	2000	5	-0.14485	0.22222	-0.58359	-0.45854
4	40	3000	3	-0.08277	1.55556	-1.10182	0.45455
5	42	1000	4	0.04138	-1.11111	-0.77707	-1.8468
6	47	2000	10	0.35177	0.22222	0.66989	1.24388
7	53	1500	11	0.72423	-0.44444	0.91105	1.19084
8	58	1500	10	1.03462	-0.44444	0.66989	1.26006
9	60	3000	14	1.15877	1.55556	1.63453	4.34884

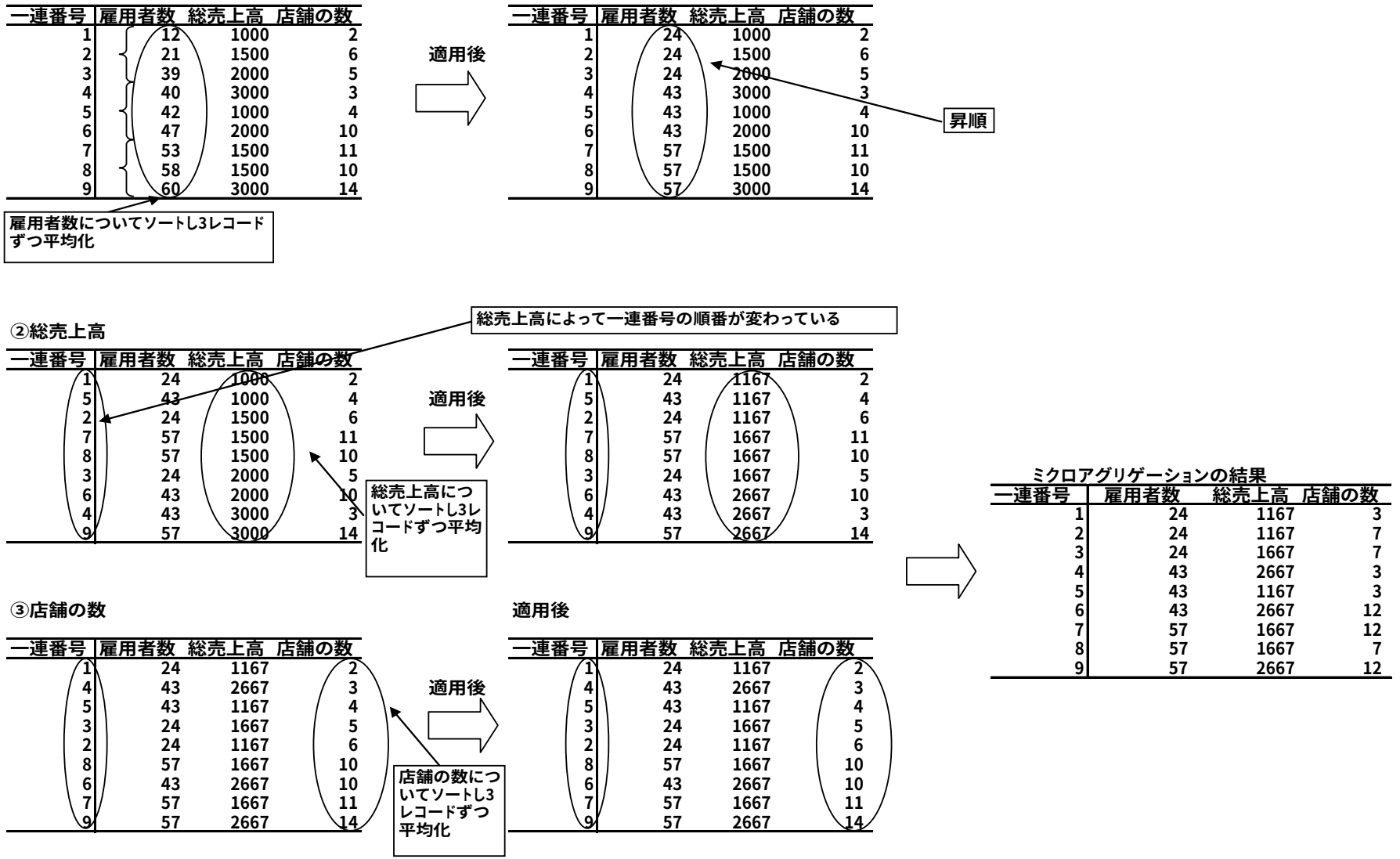
適用後

一連番号	雇用者数	総売上高	店舗の数	グループの番号
1	25	1167	4	1
2	25	1167	4	1
3	44	2167	6	2
4	44	2167	6	2
5	25	1167	4	1
6	55	2167	11	3
7	44	2167	6	2
8	55	2167	11	3
9	55	2167	11	3

注　ミクロアグリゲーション後の属性群も調査項目としての性質を継承すると考えられる。本図においては、雇用者数、総売上高、店舗の数が属性群として含まれているが、これらの属性群はすべて調査項目として整数値をとるものと仮定している。したがって、ミクロアグリゲーション後に小数点以下を四捨五入した場合、個々の属性値の合計は、原データにおける属性値の合計と必ずしも一致しないことに留意されたい。

出所　Tzavidis and Panaretos(2001, pp.11-19)より作成

図３　複数の量的属性群におけるミクロアグリゲーション―個別ランキング法の適用



出所　Tzavidis and Panaretos(2001, pp.26-30)より作成

が平均値に置き換えられる。総売上高、店舗の数についても同様に、レコード群のソート、及びレコードのグループ化を行った上で、それぞれの属性値が各グループ内の平均値に変換される。なお、Eurostat の Community Innovation Survey(1994)では、量的属性において個別ランキング法を採用していることが知られている(Thorogood(1999, p.31))。

他方、量的属性のミクロアグリゲーションにおいて、グループ化の基準となるレコード数を固定するのではなく、最初に閾値を決めた上で、個別データの分布特性に即した形でグループのレコード数を探索的に設定する手法が存在する。その１つが、Ward の階層区分法(hierarchical clustering method)をミクロアグリゲーションに適用することである(Domingo-Ferrer and Mateo-Sanz(2002, p.192))。階層区分法では、レコード群における同質性を最大にするようにグループ化が行われる。図４は、閾値を３に設定した場合のレコードのグループ化に関するイメージを示したものである。図４においてグループ内のレコード数を３に固定してレコード群をグループ分けした場合、グループ内のレコードの属性値が同質的になるようにレコードのグループ化がなされているとは言いがたい。そこで、階層区分法においては、図４に見られるように、閾値の基準を満たしながら、各グループ内にできるだけ同質的な属性値群が含まれるようにレコードのグループ化が行われる（k 分割（k-partition））。

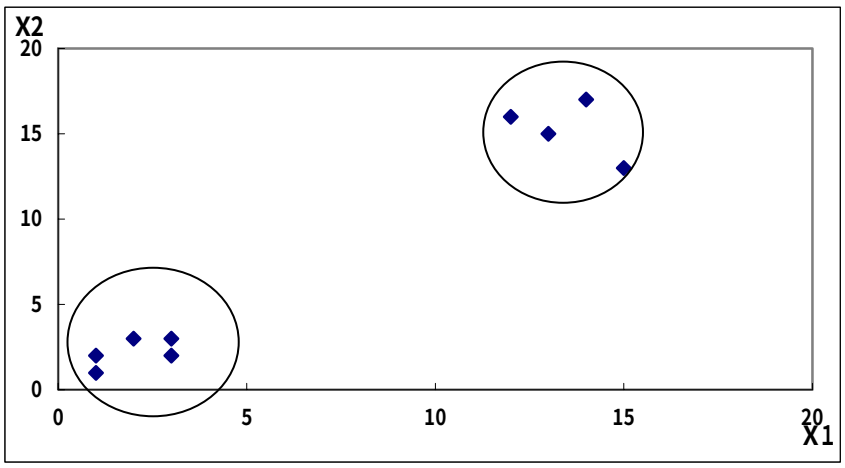
Domingo-Ferrer and Mateo-Sanz(2002, p.190)によれば、探索的なミクロアグリゲーションはつぎのように説明されている。n 個のレコードが p 個の属性を有しているとする。そのとき、p 個の変数(p は連続変数)をそなえた n 個のデータベクトルからなるミクロデータセットを想定することができる。このデータベクトルは、一般に $\mathbf{X}' = (X_1, \cdots, X_p)$（$X_i$ は変数)と表されている。n 個のデータベクトルが n_i 個のデータベクトルから成る g 個のグループに分割される場合（$n_i \geq k$ 及び $n = \sum_{i=1}^{g} n_i$）、i 番目のグループにおける j 番目のデータベクトルを $\mathbf{x}_{ij}$、i 番目のグループにおけるデータベクトルの平均値を $\bar{\mathbf{x}}_i$、n 個のデータベクトルにおける平均値を $\mathbf{x}$ と表す。

探索的なミクロアグリゲーションにおいては、グループ内平方和(within-groups sum of squares=SSE)を最小にするための閾値 k が探索的に求められる。グループ内平方和は、次の(1)式で与えられる。

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)' (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) \cdots (1)$$

このグループ内平方和が小さいほど、グループ内の同質性が高いと考えられる。次に、グループ間平方和(between-groups sum of squares=SSA)は、

図４　探索的な(heuristic)閾値の設定によるレコードのグループ化のイメージ



出所　Domingo-Ferrer and Mateo-Sanz(2002, p.191)より筆者が作成

$$SSA=\sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i-\bar{\mathbf{x}})'(\bar{\mathbf{x}}_i-\bar{\mathbf{x}})\cdots(2)$$

で表される。さらに、総平方和(total sum of squares=SST)は、グループ内平方和とグループ間平方和の合計、すなわち SST=SSA+SSE であり、

$$SST=\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}})'(\mathbf{x}_{ij}-\bar{\mathbf{x}})\cdots(3)$$

である。情報量の損失の程度を計測するために、グループ内平方和と総平方和の比、すなわち、次の尺度 L が定式化されている。

$$L=\frac{SSE}{SST}\cdots(4)$$

この尺度 L は 0 から１の間の数値をとるが、L が小さいほど、グループ内の同質性は高くなると考えられることから、L が最小になるような閾値 k が選択される。

次に、Domingo-Ferrer and Mateo-Sanz (2002)は、Ward の階層区分法に関して k-ward と呼ばれるアルゴリズムを提示している(Domingo-Ferrer and Mateo-Sanz(2002, p.193))。それは、以下のとおりである。

①データセットに含まれる最初の k 個のレコードがグループ化され、最後の k 個のレコードがもう１つのグループとして編成される。それ以外の中間に位置するレコード群が、単一のグループ(single-element group)を構成する。

②データセット内のすべてのレコードが、k 以上のレコードを含むグループに含まれるような操作が実行される。

③２k 以上のレコードを含むグループについては、①と②のアルゴリズムが繰り返される。

(2)質的属性に関するミクロアグリゲーション

近年、質的属性のミクロアグリゲーションについても研究が進められている。Torra(2004)によれば、質的属性のミクロアグリゲーションにおいても、閾値にそってレコード群のグループ化が行われるが、グループ内の属性値群は、平均値ではなく、主としてメディアンやモードに置き換えられている(Torra(2004))。質的属性に関するレコード群のソートについても、量的属性とは異なる方法が用いられている。

ソートについては、例えば次のような方法が提案されている。

①スネーク法(snake method)(Defays and Anwar(1998, pp.454-455))

スネーク法は、主に順序変数のソートに対して用いられる手法であり、質的属性に対する個別ランキング法の適用と考えられる[8]。スネーク法では、レコードに含まれる質的属性群を関連性の強い質的属性ごとに区分した上で(Thorogood(1999, p.31)、それらの属性値についてできるだけ同質的になるようにソートが行われる。また、属性値はメディアンといった代表値に置き換えられる。

図５は、２つの順序変数 X_1 と X_2 を用いてスネーク法のイメージを図示したものである。X_1 と X_2 は、それぞれ５つの分類項目に区分されているとする。図５では、(１，１)…(１，５)、(２，５)…(２，１)、…といった順序でソートを行った上で、３ずつグループ化され、属性値がメディアンに置き換えられる。

②エントロピーによる計測(Defays and Anwar(1998, p.457))

グループ化における同質性の尺度として、次の(5)式に基づいてエントロピーが計算される。

$$H = \left(-\sum_{i=1}^{L} p_i ldp_i \right) \Big/ ldL \quad \cdots(5)$$

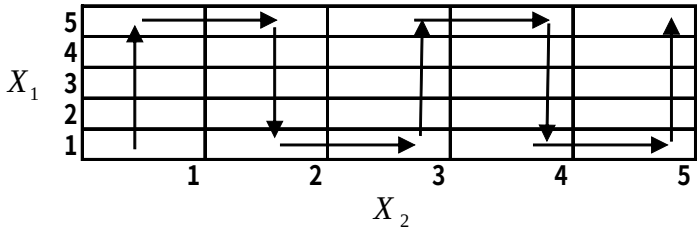
p_i…ある属性指標における i 番目の分類項目における頻度(出現確率)

ld…底を２とする対数

L…属性群における分類項目の数

[8] Community Innovation Survey(1994)では、順序変数に対してスネーク法が用いられている(Thorogood(1999, p.31))

図５　スネーク法のイメージ



スネーク法の適用

一連番号	X_1	X_2
1	1	1
2	1	2
3	2	4
4	2	5
5	2	4
6	3	3
7	4	3
8	4	2
9	5	5

適用後

一連番号	X_1	X_2
1	1	1
2	1	2
4	2	5
3	2	4
5	2	4
6	3	3
7	4	3
8	4	2
9	5	5

属性値のメディアンへの置き換え

一連番号	X_1	X_2
1	1	2
2	1	2
3	2	4
4	1	2
5	2	4
6	2	4
7	4	3
8	4	3
9	4	3

出所 Defays and Anwar(1998, pp.454-456)をもとに筆者が作成

各属性値におけるエントロピーを計測した上で、エントロピーの値に基づいてソートが行われる。

３　ミクロアグリゲーションにおける評価の基準

ミクロデータの秘匿措置においては、個別データに含まれる個体情報の保護とミクロデータの有用性の両面からその適用可能性が追究されてきた。そこで、匿名化技法としてミクロアグリゲーションが適用される場合においても、①ミクロアグリゲートデータの秘匿の程度、及び②ミクロアグリゲートデータの有用性の両面から、ミクロアグリゲーションを評価するための基準が追究される。

(1)ミクロアグリゲートデータにおける秘匿性

ミクロアグリゲーションは、政府統計の集計表で適用されている秘匿の方法にその着想を得ている(Defays (1999, p.223))。Federal Committee on Statistical Methodology(2005, p.24)によれば、集計表に含まれるセルの中の度数が１又は２である場合、そのセルは、個体を特定するリスクの高いセンシティブな(sensitive)度数であるとみなされる。そのために、集計表に度数１又は２となるセルが存在する場合には、集計表における秘匿の観点から、通常、該当するセルの度数をＸに置き換える欠測化(suppression)等の秘匿措置がとられてきた。

他方、集計表における秘匿の基準をミクロアグリゲーションの手法に適用した場合、次のように考えることができる。すなわち、ミクロアグリゲーションによって編成されたグループ内のレコードの数が１又は２である場合、個体情報が特定されるリスクが極めて高くなるが、０かあるいは少なくとも３レコードあればそのリスクは低下したと考えることが可能である[9]。なお、先行研究によれば、レコード群のグループ化の基準となる閾値は３～10 の間で設定されている。

(2)ミクロアグリゲートデータにおける有用性

ミクロデータの有用性については、秘匿処理が施されていない個別データ(以下「原データ」と呼ぶ)と秘匿処理済データ(protected data)の間のデータ構造の近似性を計測することにより評価される。そこで、秘匿処理済データの原データに対する情報量損失(information loss)(Mateo-Sanz, Domingo-Ferrer and Sebé(2005, pp.182-184))が考案されてきた。情報量損失は、秘匿処理済データが原データと比べてどの程度情報を失っているかを算出した指標である。William Winkler によれば、情報量損失の基準は、秘匿処理済データが「分析上有効であること(analytically valid)」、「分析上興味深いこと(analytically interesting)」だと考えられている(Mateo-Sanz, Domingo-Ferrer and Sebé(2005, p.182))。「分析上有効である」とは、原データと秘匿処理済データにおいて、①レコードに含まれる属性群に関する平均と共分散、②集計表に関する周辺分布、③少なくとも１つの分布上の特性が近似的とみなされることである。また、「分析上興味深い」とは、レコード群において分析上有効な属性群が複数個データセットに含まれていることである。分析上有効な属性の数については任意に定めることが可能であるが、Mateo-Sanz, Domingo-Ferrer and Sebé(2005)では、分析上有効な属性の数が６に設定されている。

秘匿処理済データの原データに対する情報量損失を算出するために、次の統計指標を用いて原データと秘匿処理済データとの間のデータ構造を比較することが提唱されている(Domingo-Ferrer and Torra(2001, p.104))。

①共分散行列

②相関係数行列

---

[9] 本稿では、ミクロアグリゲートデータにおける開示リスクの評価方法については考察の対象としていない。これについては、今後の課題だと考えている。また、開示リスクの評価方法については数多くの研究が存在するが、我が国の政府統計の個別データを用いた開示リスクの計測については、例えば、Takemura(2002)、佐井（1998）、Hoshino(2001)等を参照されたい。

③属性値と主成分分析から得られたそれぞれ因子との間の相関係数行列

④属性値のおのおのと第１主成分(それ以外の主成分)とのコモナリティ(commonality)(各属性が第１主成分（あるいはそれ以外の主成分によって）説明される比率)

⑤因子スコア係数行列(factor score coefficient matrix)

また、情報量損失の大きさについては、次のような尺度を用いて評価が行われる。

①平均平方誤差(Mean square error)

②平均絶対誤差(Mean absolute error)

③平均変量(Mean variation)

このような情報量損失の考え方は、ミクロアグリゲーションの有効性の検証においても適用可能であって、ミクロアグリゲートデータの原データからの情報量損失を計算し、その損失量が最小となるデータが最も望ましいミクロアグリゲートデータであるとみなされる。

4　匿名化技法としてのミクロアグリゲーションの適用可能性

Domingo-Ferrer and Torra (2001)によれば、ミクロアグリゲーションの手法は、主として量的属性を対象とした匿名化技法として方法的に位置付けられている(Domingo-Ferrer and Torra(2001, p.93))。しかし、政府統計の個別データには多くの質的属性が含まれていることから、ミクロアグリゲーションが秘匿処理の方法として適用されるためには、質的属性に関するミクロアグリゲーションの手法が具体的に追究される必要がある。その意味で、Torra(2004)が提唱するように、質的な属性値を平均値ではなくメディアンやモードといった代表値で置き換えることは、質的属性に対する匿名化技法の１つとして考慮に値すると思われる。

一方、ミクロデータの有用性の観点から見れば、ミクロアグリゲーションによって質的属性値がメディアンのような代表値で与えられた場合、このようなミクロアグリゲートデータにおける分布特性には、原データの分布と比較して、少なからず歪みが生じることも考えられる。それは、原データに含まれる情報量が、このミクロアグリゲートデータにおいて大きく失われる可能性があることを示唆している。

他方、質的属性のミクロアグリゲーションについては、対象となる質的属性群において属性値が同一であるレコードに着目し、同一の質的属性値を持つレコードをグループ化することが考えられる。グループ内のレコード群における質的属性値はすべて同一であるから、それらの属性値はグループの代表値に置き換えられたとみなすことができる。ゆえに、質的属性値に関するレコードのグループ化も「広義の」ミクロアグリゲーションの中に位置付けることが可能

である。

質的属性値に関するレコードのグループ化について具体的な例で見ていくことにする。図６では、属性群として性別(1：男、2：女)、雇用形態(1：正規の職員・従業員、2：パート、3：アルバイト、4：派遣・契約社員)、及び週間就業時間(1：35 時間未満、2：36～48 時間、3：49 時間～59 時間、4：60 時間以上)の３つの質的属性、及び量的属性として年間収入を有する個別データが想定されている。このとき、性別、雇用形態と週間就業時間の質的属性値にしたがって、この個別データに含まれるレコードをグループ化したとする。各グループは３つの質的属性値のいずれについても同一の属性値を持つレコードから構成されている。グループ化の対象となる属性群のおのおのについて同一の属性値を有するレコード群を、本稿では同質属性値レコード群と呼ぶことにする。図６で、１と３の一連番号が付与されているレコードはいずれも、性別は男(１)、雇用形態は正規の職員・従業員(１)、週間就業時間は 60 時間以上(４)という属性値を含む同質属性値レコード群の構成要素となっている。

ところで、属性群として性別、雇用形態、及び週間就業時間を含む個別データを用いて、これらの質的属性を集計事項としたクロス集計表を作成することが可能であるが、このクロス集計表におけるセルの度数と同質属性値レコード群内のレコード数は一致している。すなわち、性別が１、雇用形態が１、週間就業時間が４と付与されている同質属性値レコード群内のレコード数は２であるが、それは、性別、雇用形態と週間就業時間に関するクロス集計表において、属性値が男、正規の職員・従業員で週間就業時間が 60 時間以上に該当するセルの度数２と合致する。さらに、このクロス集計表を集計事項の分類項目の組合せとして表示すると、組合せのそれぞれに対して総数（N）と年間収入の総計が対応することがわかる。クロス集計表において質的属性値が男、正規の職員・従業員で 60 時間以上である場合、それは、分類項目の組合せとして性別１、雇用形態１、週間就業時間４に対応するだけでなく、その組合せの総数２及び年間収入の総計 360 万円という集計値が付与される。さらに、年間収入の合計を組合せ総数で割ることによって、性別１、雇用形態１と週間就業時間４という分類項目の組合せとそれに対応する年間収入の平均値 180 万円が導き出される。これらの数値群は、質的属性における分類項目の組、及び量的属性に関する平均値から構成されており、それは集計値として位置付けられる。しかし、この数値群を質的属性値群と量的属性値を含むレコードとして擬似的に捉えることも可能なように思われる。これらのレコードのおのおのについて該当する総数だけレコードを「複製」することによって、ミクロアグリゲートデータが編成される。

図６では、３つの質的属性群と１つの量的属性のみを含む仮想的な個別データを用いて議論

図６　個別データとミクロアグリゲートデータとの関係

(1)個別データ(属性群として性別，雇用形態，週間就業時間と年間収入のみが配列されていると想定)

一連番号	性別	雇用形態	週間就業時間	年間収入(千円)
1	1	1	4	1500
2	1	1	2	2300
3	1	1	4	2100
4	1	3	1	1500
5	1	3	2	2700
6	1	3	3	1800
7	2	2	3	3600
8	2	4	4	4000
9	2	2	3	2800

性別　1: 男　2:女
雇用形態　1:正規の職員･従業員　2:パート　3:アルバイト　4:派遣･契約社員
週間就業時間　1:35時間未満　2:35～48時間　3:49～59時間　4:60時間以上

(2)性別，雇用形態と週間就業時間に関する同質属性値レコード群

一連番号	性別	雇用形態	週間就業時間	年間収入(千円)
2	1	1	2	2300
1	1	1	4	1500
3	1	1	4	2100
4	1	3	1	1500
5	1	3	2	2700
6	1	3	3	1800
7	2	2	3	3600
9	2	2	3	2800
8	2	4	4	4000

対応関係

(3)性別，雇用形態，週間就業時間別クロス集計表

性別	男				女				計
雇用形態	正規の職員･従業員	パート	アルバイト	派遣社員	正規の職員･従業員	パート	アルバイト	派遣社員	
週間就業時間									
35時間未満	0	0	1	0	0	0	0	0	1
35～48時間	1	0	1	0	0	0	0	0	2
49～59時間	0	0	1	0	0	2	0	0	3
60時間以上	2	0	0	0	0	0	0	1	3
計	3	0	3	0	0	2	0	1	9

性別	雇用形態	週間就業時間	総数(N)	年間収入の総計
1	1	2	1	2300
1	1	4	2	3600
1	3	1	1	1500
1	3	2	1	2700
1	3	3	1	1800
2	2	3	2	6400
2	4	4	1	4000

(4)ミクロアグリゲートデータ

ミクロアグリゲーション後の一連番号	性別	雇用形態	週間就業時間	年間収入
1	1	1	2	2300
2	1	1	4	1800
3	1	1	4	1800
4	1	3	1	1500
5	1	3	2	2700
6	1	3	3	1800
7	2	2	3	3200
8	2	2	3	3200
9	2	4	4	4000

がなされているが、政府統計の個別データの場合においても、このような議論を拡張して展開することが可能だと考えられる。それは、政府統計の個別データが持つすべての属性群を集計事項とした多重クロス集計表を作成し、その集計表からミクロアグリゲートデータを作成することを意味している。本稿では、個別データが有するすべての属性群を集計事項の対象とした上で作成されるｎ次元の多重クロス集計表を「超高次元クロス集計表」と呼ぶことにする。図７で示されるように、超高次元クロス集計表では、あらゆる属性群の組合せが集計事項として設定可能だと考えられる。また、超高次元クロス集計表において、属性群における分類区分の設定を変えることによって、そこから新たに集計表を作成することもできる。このような超高次元クロス集計表から個別データに準じたレベルのデータを作成することは、統計データの２次的利用における新たな可能性を提示するように思われる[10]。それは、超高次元クロス集計に基づいて作成された個別データに準じたレベルのデータは、集計値の形態ではあっても、個別データと同様の属性群をそなえているとみなされるからである。

超高次元クロス集計の考え方については、これまでの先行研究にも見て取ることができる。例えば、松田(1999,124～125 頁)は、「できるだけ詳細な n 次元の多重(元)集計表」に基づいた「多重分類集計表」の作成と保管、さらには多重分類集計表から編成される「セミ・マクロ・データ」による利用可能性を議論している。また、寺崎(2000) は、集計表をリスト形式で捉え直すことによって、集計表の新たな利用のあり方を提唱している。

一方、総理府統計局(現 (独)統計センター)では、集計結果表の作成のために、一時期、セルレコード方式(タリー(Tally)方式)と呼ばれる集計方法によって製表業務が行われていたことが知られている。セルレコード方式とは、「統計表のイメージをコンピュータの内部メモリーに展開せずに、各セルごとにサマリーを作成する」方式(安野(1981,69 頁))である。図８に見られるように、セルレコード方式では、個々の集計表を作成するのに必要なすべての質的属性群の属性値とそれに対応する量的属性群(レコードの個数も含む)の集計値(集計表の１セルに対応)が１つのセットとして設定されている[11]。このセルレコード方式も超高次元クロス集計の発想に類似しているように見える。他方、本稿で議論している超高次元クロス集計がこれまでの先行研究と異なるのは、超高次元クロス集計を匿名化技法としてのミクロアグリゲーションの

[10] 我が国では、集計計画に基づいて、集計結果表(報告書に「掲載される」結果表、及び「非掲載」の結果表)が公表されている。これらの集計結果表(「結果原表」)においては、表章可能な集計事項の数に限りがあることから、統計データの 2 次的利用を行うにあたっては制約があると考えられる。それに対して、結果原表ではなく超高次元クロス集計表であれば、統計データの２次的利用の新たな展開を模索することも可能である。

[11] 当時の総理府統計局では、コンピュータの容量の制約に対して，業務の生産性の向上を目指して、機能別集計システムからセルレコード方式の集計システムが開発されている。例えば、安野(1981，63～76 頁)では、セルレコード方式による昭和 52 年、54 年の就業構造基本調査の集計方法が詳細に示されている。

図７　超高次元クロス集計のイメージ

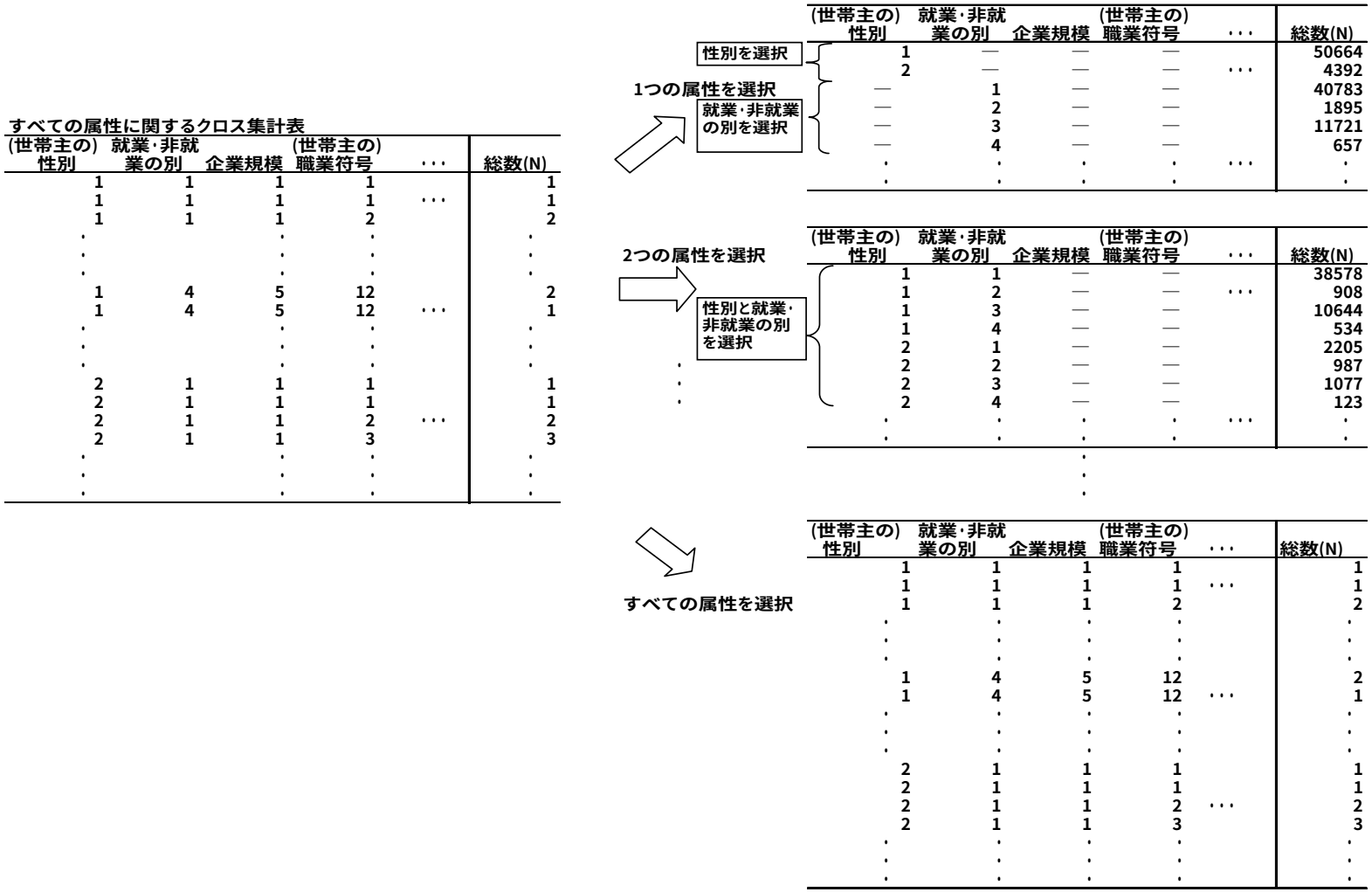


図８　セルレコードの形式

注

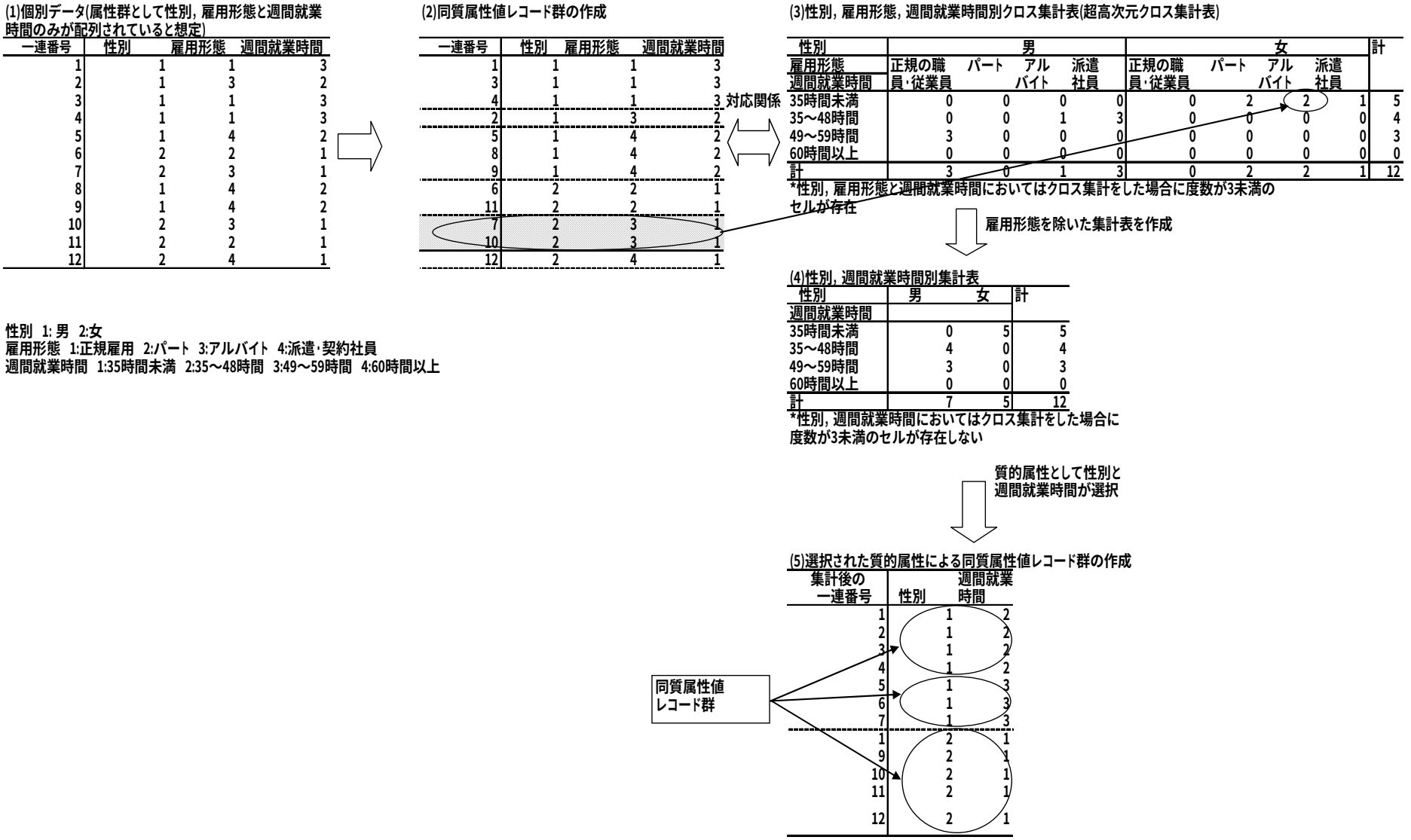
・調査名…調査アイデント
・表番号…結果表番号
・区分…1 つの表において、世帯数、世帯人員などのように異なった集計値を求める場合の識別符号
・集計地域…地域別に集計する場合の地域符号
・欄外項目…欄外項目の分類コード。項目間は 1 行あける。この 1 行は「ブランク」か「―」である。「―」は大分類，中分類などの関係がある項目を表す。
・表側項目…表側項目の分類コード。欄外項目と同じ形式。
・表頭項目…表頭項目の分類コード。欄外項目と同じ形式。
・表側連番…結果表上の表側行番号。
・表頭連番…結果表上の表頭セル番号。
・加工情報…平均値を算出する場合の表章桁数などをセットする。
・集計値…集計値は 1～３セルのいずれかである。集計値１は集計の対象となった個別データのカウントとして使われる。
・集計値１のみ：個別データのカウントのみにより結果をもとめる場合
・集計値１と２のみ：室数などの集計数をもとめる場合、または、推計乗率により集計する場合の推計値。
・集計値 1～３：平均値を算出する場合，集計値２は分母，集計値３は分子の値。

出所　安野（1981,70～71 頁）

観点から捉えていることである(Bethlehem *et al.*(1990),Höhne(2003))。ミクロアグリゲーションにおいて超高次元クロス集計を方法的に位置付けるということは、次のことを意味している。ミクロアグリゲーションでは、ミクロデータ(個別データ)が閾値 k のレコード群にグループ分けされ、グループ内のレコードにおける個々の属性値が平均値等の代表値に置き換えられる。先述したように、このグループについては同質属性値レコード群として把握することが可能であるが、対象となる属性群について編成された同質属性値レコード群内のレコード数は、同じ属性群を集計事項として作成された超高次元クロス集計表におけるセルの度数と対応している。よって、同質属性値レコード群内のレコード数を決めることは、超高次元クロス集計表に含まれるセルの度数に関する閾値を設定することを意味している。閾値ｋを設定した場合、超高次元クロス集計表の集計事項となる属性群から、属性の組合せを適当に選択することによって、超高次元クロス集計表に含まれるすべてのセルが０以外でかつｋ未満の数にならないように集計表を作成することができる。この集計表から同質属性値レコード群を編成することによって、ミクロアグリゲートデータを作成することが可能になる。

例えば、図９は、質的属性に関するミクロアグリゲートデータの作成の概略図を示したものである。図９では、閾値が３に設定されている。それは、超高次元クロス集計表の集計事項と

図９　質的属性に関するレコードのグループ化の概略図



なる属性群から、属性の組合せを適当に選択することによって、度数１又は２のセルが存在しないように集計表を新たに作成することを意味する。図９においては、性別、雇用形態と週間就業時間を用いてクロス集計を行うことによって、集計表の中に度数２のセルが出現しているが、そこから性別と雇用形態のみを集計事項として設定すれば、度数２のセルは集計表に存在しなくなる。このことから、質的属性として性別と雇用形態が選択されれば、同質属性値レコード群内のレコード数は少なくとも３以上になることがわかる。

その一方で、質的属性における分類区分を統合することによって、度数１または２のセルを含まない集計表を作成することも考えられる。図 10 は、図９と同じ個別データを使用し、雇用形態を従来の４区分から２区分(1：正規雇用者(正規の職員・従業員)、2：非正規雇用者(パート、アルバイト、派遣・契約社員))、同様に、週間就業時間を４区分から２区分（1：35 時間未満、 2：35 時間以上）に統合した場合における質的属性のグループ化の概略図 である。雇用形態と週間就業時間の分類区分を統合することによって、３つの属性群のすべてを集計事項に設定しても、作成された集計表に度数１又は２のセルが存在しないようにすることが可能になる。

個別データに量的属性と質的属性が含まれる場合のミクロアグリゲートデータの作成に関する概略図が、図 11 で示されている。属性群として性別、雇用形態、週間就業時間、及び年間収入を有する個別データが想定されている。最初に性別、雇用形態と週間就業時間の質的属性に着目し、質的属性に関する同質属性値レコード群が設定されている。同質属性値レコード群のおのおのについて、世帯総数と年間収入の合計が算出されている。次に、閾値が３に設定されていることから、同質属性値レコード群内にレコード数１又は２が存在しないように、質的属性として性別と週間就業時間のみが選択される。それによって、図 11 では、各同質属性値レコード群内における世帯総数が３以上になっていることがわかる。さらに、同質属性値レコード群における年間収入の総計をその世帯総数で割ると、年間収入の平均値が求められる。この平均値によって同質属性値レコード群内のレコードの属性値が置き換えられることによ

図 10　質的属性に関するレコードのグループ化の概略図―分類区分の統合済

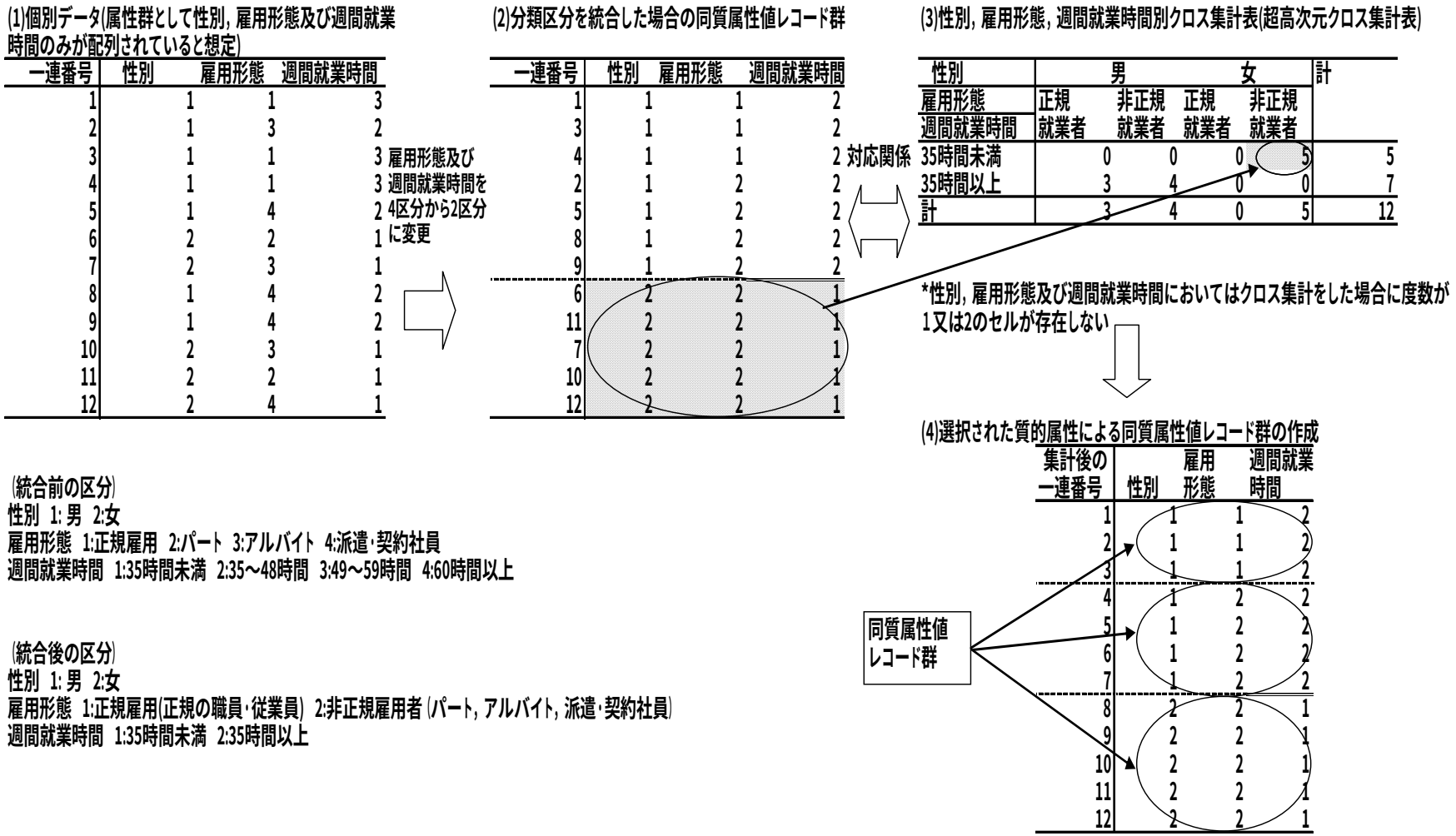


図 11　ミクロアグリゲートデータの作成に関する概略図

(1)個別データ(属性群として性別, 雇用形態, 週間就業時間及び年間収入のみが配列された個別データを想定)

一連番号	性別	雇用形態	週間就業時間	年間収入(千円)
1	1	3	2	2300
2	2	4	2	1500
3	1	3	2	2100
4	2	4	2	1500
5	1	3	2	2700
6	2	4	2	1800
7	1	1	4	3600
8	2	2	1	2800
9	1	1	4	4000
10	2	2	1	3200
11	1	1	4	4000
12	2	3	1	4000

性別　1: 男　2:女
雇用形態　1:正規の職員･従業員　2:パート
3:アルバイト　4:派遣･契約社員
週間就業時間　1:35時間未満　2:35～48時間
3:49～59時間　4:60時間以上

(2)質的属性に関する同質属性値レコード群

一連番号	性別	雇用形態	週間就業時間	年間収入(千円)
7	1	1	4	3600
9	1	1	4	4000
11	1	1	4	4000
1	1	3	2	2300
3	1	3	2	2100
5	1	3	2	2700
8	2	2	1	2800
10	2	2	1	3200
12	2	3	1	4000
2	2	4	2	1500
4	2	4	2	1500
6	2	4	2	1800

(3)超高次元クロス集計表

性別	雇用形態	週間就業時間	世帯総数(N)	年間収入の総計
1	1	4	3	11600
1	3	2	3	7100
2	2	1	2	6000
2	3	1	1	4000
2	4	2	3	4800

質的属性として
性別と週間就業時間を選択

(4)性別と週間就業時間に関するクロス集計表(超高次元クロス集計表)

性別	週間就業時間	世帯総数(N)	年間収入の総計
1	4	3	11600
1	2	3	7100
2	1	3	10000
2	2	3	4800

(5)ミクロアグリゲートデータ

ミクロアグリゲーション後の一連番号	性別	週間就業時間	年間収入(千円)
1	1	2	2367
2	1	2	2367
3	1	2	2367
4	1	4	3867
5	1	4	3867
6	1	4	3867
7	2	1	3333
8	2	1	3333
9	2	1	3333
10	2	2	1600
11	2	2	1600
12	2	2	1600

って、ミクロアグリゲートデータが編成される[12]。

5 結びにかえて

本稿は、諸外国で匿名化技法として近年注目されているミクロアグリゲーションの研究動向とミクロアグリゲーションの方法的な特徴を明らかにした。本稿は、ミクロアグリゲーションに関する試論的な考察に過ぎない。したがって、政府統計の個別データを用いて、ミクロアグリゲーションの方法的な有効性を検証することによって、ミクロアグリゲーションの展開可能性を模索する必要があろう。我が国の政府統計の個別データによるミクロアグリゲーションの有効性の検証については、稿を改めて述べることにしたい。

参考文献

Anwar M. N.(1993) “Micoaggregation: The Small Aggregates Method”, *Eurostat Internal Report*

Bethlehem, J. G., Keller, W. J. and Pannekoek, J.(1990) Disclosure Control of Microdata, *Journal of the American Statistical Association*, Vol.85, No.409, pp.38-45.

Defays, D.(1997) “Protecting Micro-Data By Micro-Aggregation:The Experience in Eurostat”, *QÜESTIIÓ*, vol.21, 1 i 2, pp.221-231

http://www.idescat.net/sort/questiio/questiiopdf/21.1.10.Defays.pdf

Defays, D. and Anwar, M.N.(1998) “Masking Microdata Using Micro-Aggregation”, *Journal of Official Statistics*, Vol.14, No.4, pp.449-461.

Domingo-Ferrer, J. and Torra, V.(2001) ”Disclosure Control Methods and Information Loss for Microdata”, Doyle *et al.*(eds.)(2001)*Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, Elsevier Science, Amsterdam, pp.91-110.

Domingo-Ferrer, J. and Mateo-Sanz, J. M.(2002) ”Practical Data-oriented Microaggregation for Statistical Disclosure Control”, *IEEE Transactions on Knowledge and Data Engineering,* vol.14, no.1, pp.189-201.

[12] 図 11 は、量的属性と質的属性が個別データに設定されている場合のミクロアグリゲーションの模式図を表したものに過ぎない。図 11 では、量的属性が年間収入のみとなっており、複数の量的属性がレコードに設定されている場合には、単一軸法、個別ランキング法等の量的属性に関するアグリゲーションの手法が、レコードに含まれる属性の性質にしたがって適用される。その場合、質的属性群についてのみ同質属性値レコード群を編成し、同質属性値レコード群内に件数１又は２が存在しない質的属性の組合せを選び出した上で、同質属性値レコード群内のレコードに含まれる量的属性群にミクロアグリゲーションの手法を適用することが考えられる。また、図 11 では、同質属性値レコード群ごとの世帯総数と年間収入の総計から平均値を求めた上で、そこから同質属性値レコード群内に含まれるレコードの量的属性値が平均値に置き換えられているが、これは、ミクロアグリゲーションの手法として単一軸法と同様の方法を適用していると考えることができる。

Domingo-Ferrer, J., Sebé, F. and Solanas, A.(2007)"Microaggregation Heuristics for P-Sensitive K-anonymity", Work Session on Statistical Data Confidentiality (Manchester, United Kingdom, 17-19 December 2007),

http://www.unece.org/stats/documents/2007.12.confidentiality.htm

Federal Committee on Statistical Methodology (2005) *Statistical Policy Working Paper 22(Second version): Report on Statistical Disclosure Limitation Methodology*. Federal Committee on Statistical Methodology, U.S. Office of Management and Budget, Washington, D.C.

Felsö, F., Theeuwes, J., Wagner, G. G.(2001)"Disclosure Limitation Methods in Use: Results of a Survey", Doyle *et al.*(eds.)(2001)*Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, Elsevier Science, Amsterdam, pp.17-42.

Hundepool, A.(2006) "The ARGUS Software in CENEX", Domingo-Ferrer, J. and Franconi, L.(Eds.) *Privacy in Statistical Databases: CENEX-SDC Project International Conference, PSD 2006 Rome, Italy, December 13-15, 2006 : proceedings*, Springer, Berlin, pp.334-346.

Hoshino, N.(2001) "Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment", *Journal of Official Statistics*, Vol.17, No.4, pp.499-520.

Höhne, J.(2003) "SAFE- A Method for Statistical Disclosure Limitation of Microdata", Joint ECE/Eurostat Work Session on Statistical Data Confidentiality (Luxenbourg, 7-9 April 2003)

http://unece.org/stats/documents/2003/04/confidentiality/wp.37.s.e.pdf

木村英典(1980)「機能別集計システムから新システムへ」『統計局研究彙報』第 34 号，37～64 頁

Mateo-Sanz, J. M. and Domingo-Ferrer, J.(1998) "A Comparative Study of Microaggregation Methods", *QÜESTIIÓ*, vol.22, 3, pp.511-526.

Mateo-Sanz, J. M., Domingo-Ferrer, J., Sebé, F.(2005) "Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata" *Data Mining and Knowledge Discovery,* vol.11, pp.181-193.

松田芳郎(1999)『ミクロ統計データの描く社会経済像』日本評論社

森博美(2007)「ミクロ統計とマクロ統計」『統計』2007 年 3 月号，2～7 頁

Pagliuca, D. and Seri, G.(1998) "The Release of Business Microdata: A Software Prototype for Microaggregation"

http://europa.eu.int/en/comm/eurostat/research/conferences/ntts-98/papers/cp/058c.pdf

Pagliuca, D. and Seri, G.(1999)"Masking Business Microdata with MASQ"

http://europa.eu.int/en/comm/eurostat/research/conferences/etk-99/papers/pagliuca-seri.pdf

佐井至道(1998)「個票データにおける個体数とセル数との関係」『応用統計学』Vol.27 No.3，127～145 頁

Spruill, N(1983) “The Confidentiality and Analytic Usefulness of Masked Business Microdata” in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 602-607.

Strudler, M., Oh, H. L. and Scheuren, F.(‘1986) “Protection of Taxpayer Confidentiality with Respect to the Tax Model” in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 375-381.

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

竹村彰通(2003)「個票開示問題の研究の現状と課題」『統計数理』第 51 巻第 2 号，241～260 頁

瀧敦弘(2003)「集計表におけるセル秘匿問題とその研究動向」『統計数理』第 51 巻第 2 号，337～350 頁

寺崎康博(2000)「リスト形式による集計表とパターン化変数」松田芳郎・伴金美・美添泰人(編著)『講座ミクロ統計分析② ミクロ統計の集計解析と技法』日本評論社，111～122 頁

Thorogood D.(1999) “Protecting the Confidentiality of Eurostat Statistical Outputs”, *Netherlands Official Statistics*, Volume 14, Spring, pp.30-33.

Torra, V.(2004) “Microaggregation for Categorical Variables: A Model Based Approach”, Domingo-Ferrer, J. and Torra, V.(eds) *Privacy in Statistical Databases CASC Project Final Conference PSD 2004 Barcelona Catalonia, Spain, June9-11, 2004 Proceedings*, Springer, pp.162-174.

Tzavidis, N. and Panaretos, J. (2001) *Aspects of Estimation Procedures at Eurostat with Some Emphasis in the Over-space Harmonisation*, Athens, Greece, Department of Statistics, Athens University of Economics

http://stat-athens.aueb.gr/~jpan/diatrives/Tzavidis/Index.html

United Nations Economic Commission for Europe(2007) *Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice*

http://www.unece.org/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

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

Winkler, W. E. (2002) “Single Ranking Micro-aggregation and Re-identification,” Statistical Research Division report RR 2002/08

http://www.census.gov/srd/www/byyear.html

Wolf, M. K.(1988) “Microaggregation and Disclosure Avoidance for Economics Establishment Data” *American Statistical Association 1988 Proceedings of the Business and Economics Statistics Section.*

Alexandria, Va.: American Statistical Association.

安野勝吾(1981) 「統計局における汎用統計集計システムの考察」『統計局研究彙報』第 37 号，53～101 頁

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

付表　ミクロデータの公開のために利用される開示制限法(disclosure limitation techniques)

人口統計のミクロデータ

	ミクロアグリゲーション	データ項目の削除	センシティブなレコードの削除	データスワッピング	分類区分の再符号化	トップコーディングないしはボトムコーディング	標本抽出	地域区分ないしは人口集団における閾値の設定
ブルガリア		○			○	○		
チェコ共和国			○					
エストニア	○							
ハンガリー		○						
ラトビア		○						
リトアニア	○		○			○	○	
ポーランド					○			○
ルーマニア	○	○			○		○	○
スロバキア	○	○			○	○	○	○
スロベニア		○			○		○	
旧ユーゴスラビアマケドニア共和国								
ユーゴスラビア								
アゼルバイジャン								
ベラルーシ								
キルギスタン	○				○			
ロシア								
トルクメニスタン								
計								

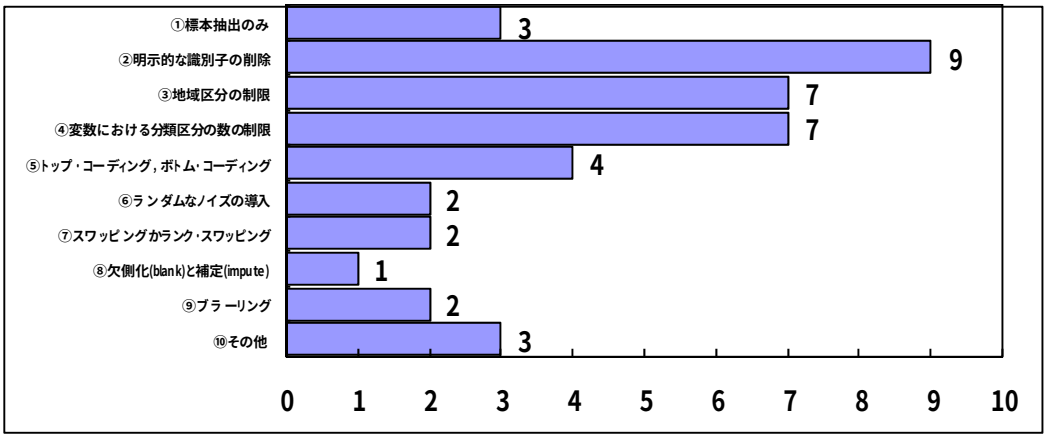
経済統計のミクロデータ

	ミクロアグリゲーション	データ項目の削除	センシティブなレコードの削除	データスワッピング	分類区分の再符号化	トップコーディングないしはボトムコーディング	標本抽出	地域区分ないしは人口集団における閾値の設定
ブルガリア		○			○	○		
チェコ共和国								
エストニア								
ハンガリー		○						
ラトビア		○						○
リトアニア	○	○	○	○	○	○	○	○
ポーランド	○				○			○
ルーマニア	○	○					○	
スロバキア	○				○	○	○	○
スロベニア								
旧ユーゴスラビアマケドニア共和国								
ユーゴスラビア	○							
アゼルバイジャン								
ベラルーシ								
キルギスタン	○			○	○			
ロシア								
トルクメニスタン								
計								

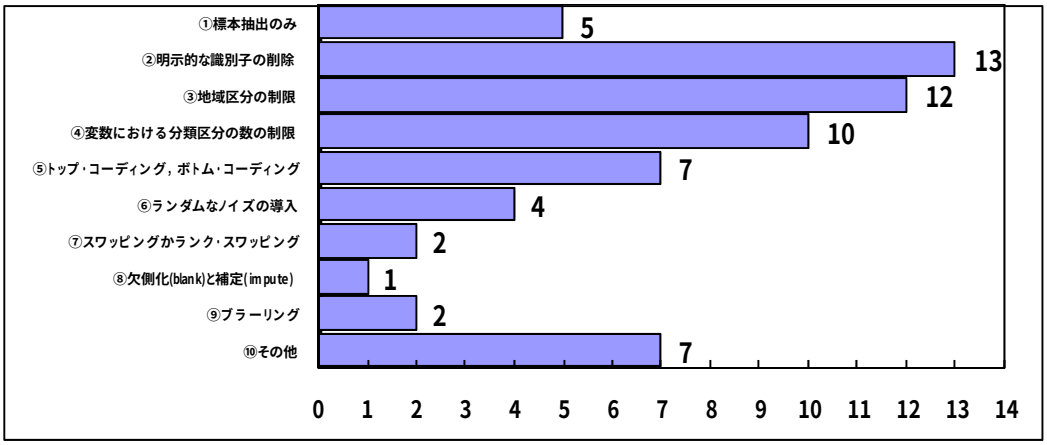
出所　Felsö *et al.*(2001, pp.22-23)より作成

付図 ミクロデータにたいする諸外国の匿名化措置の現状

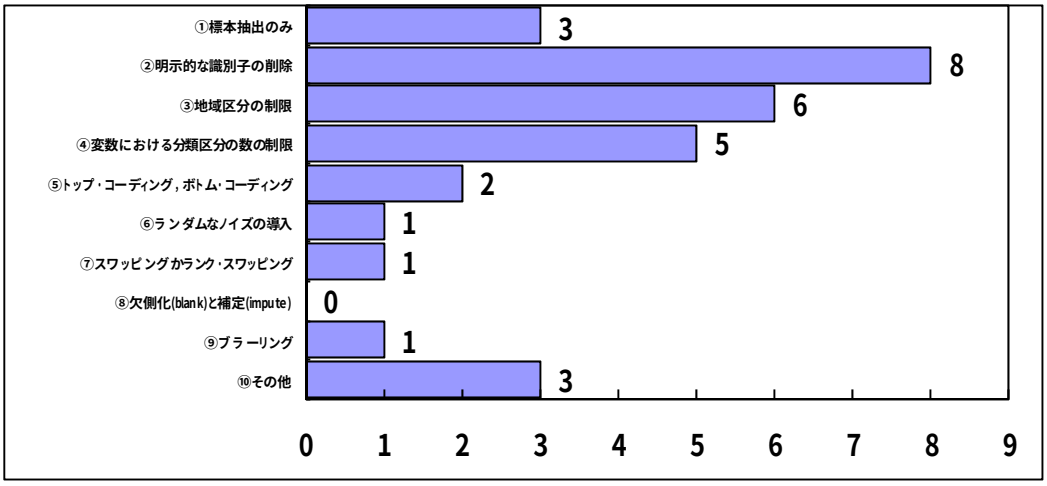
(1)人口センサスのミクロデータ



(2)センサス以外の人口統計のミクロデータ



(3)経済統計のミクロデータ



注　本図において調査の対象となっている国は、次のとおりである(Felsö *et al.*(2001, p.26)。
カナダ、チェコ共和国、デンマーク、エストニア、ドイツ、ハンガリー、イタリア、リトアニア、オランダ、ニュージーランド、ノルウェー、スウェーデン、アメリカ。

出所　Felsö *et al.*(2001, pp.33-34)

匿名化技法としてのミクロアグリゲーションの有効性に関する研究
―全国消費実態調査を例に―

伊藤　伸介*，磯部　祥子**，秋山　裕美**

要　旨

諸外国では、政府統計のミクロデータを提供するために、個体情報の保護に関する法的制度的措置が整備されてきただけでなく、トップ・コーディング、スワッピング等、個体情報の漏洩を回避することを指向した匿名化技法の開発が進められている。本稿では、諸外国で匿名化技法の１つとして展開されているミクロアグリゲーション(microaggregation)に着目し、『平成 16 年全国消費実態調査』の秘匿処理を施していない個別データ(「原データ」)を用いて、①ミクロアグリゲーションによる個別データに準じたレベルのデータ(ミクロアグリゲートデータ)の作成、及び②ミクロアグリゲートデータの原データに対する近似性の検証を試みた。

本研究では、第１に、対象となるすべての質的属性について同一な属性値を有するレコード群（「同質属性値レコード群」）の編成を行った。そして、秘匿の観点から閾値を３に設定した上で、同質属性値レコード群内にレコード数１又は２を含まない質的属性の組合せを検討した。第２に、同質属性値レコード群内においてレコードをグループ化し、各グループにおける量的属性値を平均値に置き換えることによって、ミクロアグリゲートデータを作成した。また、量的属性のおのおのに対して個別にソートを行う個別ランキング法を中心に、量的属性のミクロアグリゲーションを行った。第３に、作成されたミクロアグリゲートデータの原データに対する近似の程度を明らかにするために、ミクロアグリゲートデータの分布特性を原データのそれと比較した。さらに、原データとミクロアグリゲートデータのそれぞれの相関係数行列を計算し、その平均平方誤差を算出することによって、ミクロアグリゲートデータにおける情報量損失を計測した。

本研究においては、同質属性値レコード群を編成するために選出される質的属性の組合せが、全部で 255 パターン存在するが、その中で原データにおける同質属性値レコード群内にレコード数が 1 又は 2 でない組合せは、18 パターン（7.1%）となった。また、本研究では、個別ランキング法を用いた場合、原データに対してより近似的なミクロアグリゲートデータの作成が可能になることがわかった。

* 統計センター情報技術部研究主幹非常勤職員（明海大学経済学部専任講師）
** 統計センター情報技術部研究主幹（E-mail: research@nstac.go.jp）

匿名化技法としてのミクロアグリゲーションの有効性に関する研究
―全国消費実態調査を例に―

伊藤 伸介，磯部 祥子，秋山 裕美

１　はじめに

欧米諸国においては、1960 年代以降、政府統計のミクロデータが一般に提供されてきた。諸外国では、トップ・コーディング、スワッピング等、個体情報の漏洩を回避することを指向した匿名化の技術的な手法が開発されるだけでなく、個体情報を保護するための法的制度的措置が整備されてきた。それによって、ミクロデータの利用が可能になり、ミクロレベルの実証的な社会経済研究が広範に進められてきた。

他方、我が国では、政府統計の二次的利用の観点から、政府統計ミクロデータの提供に対する関心が高まっている。我が国においても、諸外国の場合と同様に、個別データの秘匿性を確保した上で、個別データの有用性を考慮することが求められる。個別データの秘匿措置については、諸外国において数多くの研究蓄積が存在する。こうした点を踏まえて、我が国の個別データに適用可能な秘匿処理の方法が具体的に検討されることが必要かと思われる。

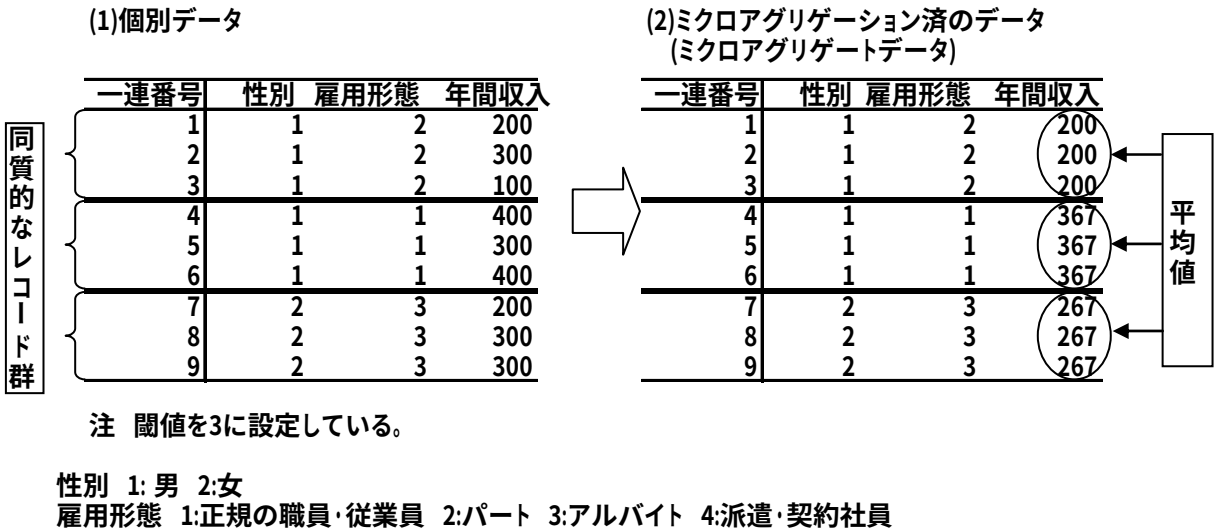
ところで、伊藤（2008）は、近年主にヨーロッパ諸国で調査研究が進められている「ミクロアグリゲーション（microaggregation）」に着目し、ミクロアグリゲーションの研究動向を洞察している。ミクロアグリゲーションは、個別データにおける匿名化技法の１つとして位置付けられており（Willenborg and de Waal（2001，pp.30-31））、Eurostat では、企業を調査対象とした Community Innovation Survey において、ミクロアグリゲーションが適用されている（Thorogood（1999））。また、イタリア統計局では、System of Enterprises Accounts Annual Survey を用いた企業データの一般公開型ファイル（Public Use File）の作成が進められている（Pagliuca and Seri（1999，p.304））。その一方で、我が国ではミクロアグリゲーションに関する実証的な研究がこれまで行われていなかったことから、我が国におけるミクロアグリゲーションの方法的な可能性を具体的に検討することは意義があると考えられる。

よって、本稿は、我が国におけるミクロアグリゲーションの適用可能性を追究するために、『全国消費実態調査』の個別データを用いて、個別データに準じたレベルのデータの作成を試み、ミクロアグリゲーションの有効性を検証する。

2　ミクロアグリゲーションの特徴

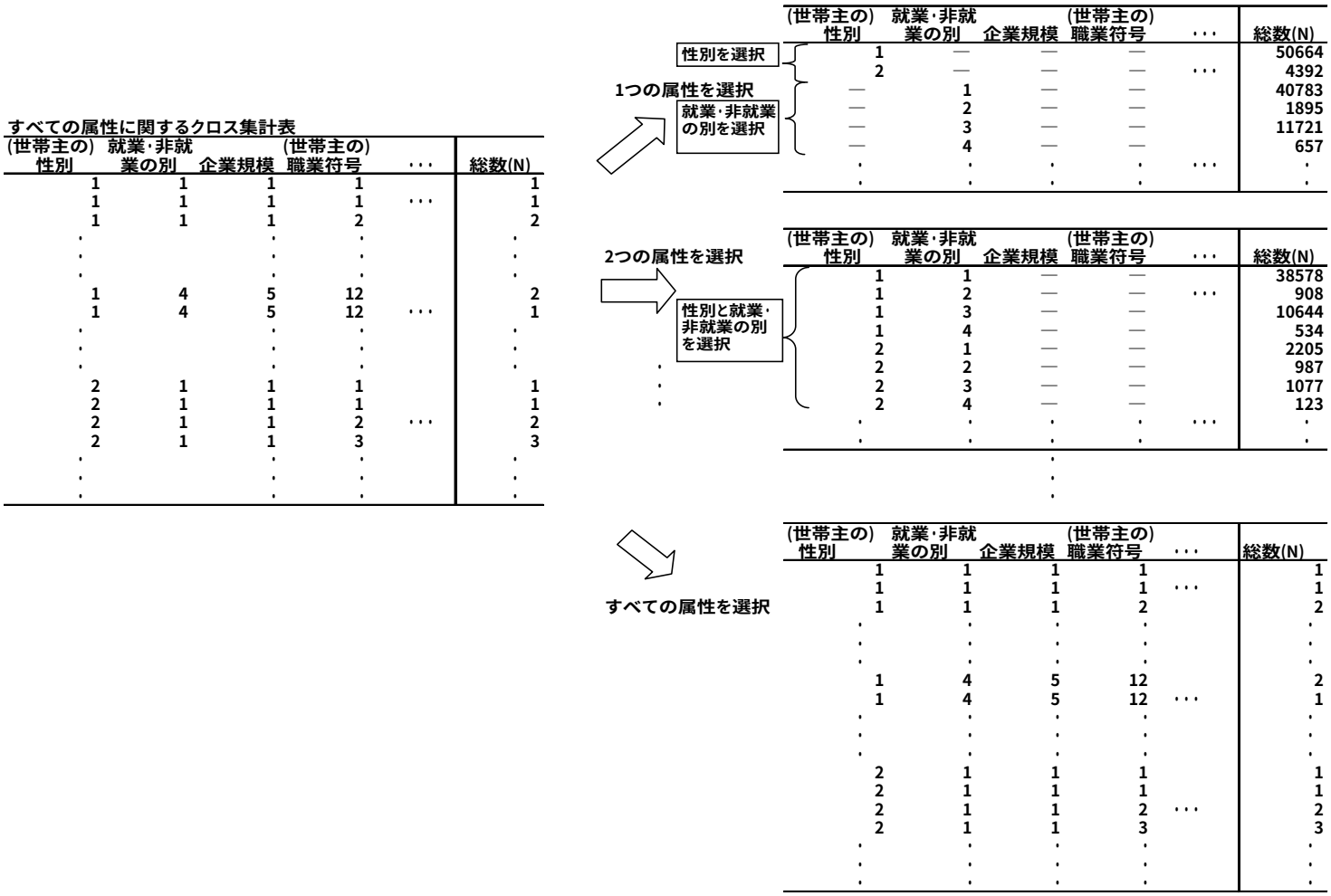
一般に、統計調査の個別データは、複数の調査項目（属性）と調査項目の回答値（属性値）から成り立っている。ミクロアグリゲーションとは、ミクロデータ（個別データ）を k 個（k は閾値）のレコードを持つ同質的なレコード群にグループ化した上で、そのレコードにおける個々の属性値を平均値等の代表値に置き換えることである（Domingo-Ferrer and Mateo-Sanz（2002,　p.190））。例えば、属性群として性別、雇用形態と年間収入のみを持つ個別データを考えることにし、閾値を３に設定したとする（図１）。このデータ上にある属性群にミクロアグリゲーションを適用するということは、性別、雇用形態と年間収入の属性値のおのおのについて同質的であるとみなされるレコード群を少なくとも３レコードずつグループ化し、各グループ内のレコードが有する属性値を平均値等の代表値に変換することを意味している。図１では、性別と雇用形態における属性値が同一になるように３レコードずつグループ化がなされており、各グループ内で年間収入が平均値に置き換えられていることを示している。

図１　ミクロアグリゲーションの例



ところで、個別データに設定される属性群は、年間所得や消費支出といった数値項目を表す量的属性、及び性別や配偶関係といった分類項目を示す質的属性に大きく区分される。そのために、量的属性と質的属性のそれぞれの特性に応じた形で匿名化技法の展開がはかられてきた。例えば、Domingo-Ferrer and Torra （2001）によれば、ミクロアグリゲーションの手法は、主として量的属性を対象とした匿名化技法として方法的に位置付けられている。しかし、近年では、質的属性に関するミクロアグリゲーションの手法についても実証的な研究が行われてい

図２　超高次元クロス集計のイメージ



出所　伊藤（2008，20 頁）

る。例えば、Torra（2004）は、質的な属性値については、それを平均値ではなくメディアンやモードで置き換えることを提唱している[1]。しかし、個別データの有用性の観点から見れば、ミクロアグリゲーションによって、質的属性値がメディアンのような代表値で与えられた場合、原データが持つ情報量と比較して、情報量が大きく損なわれる可能性がある。そこで、本研究では、質的属性に関しては、対象となるすべての属性について同一の属性値を有するレコード群（以下、「同質属性値レコード群」と呼ぶ）に着目し、特定の閾値に基づいた同質属性値レコード群の編成によって、質的属性のミクロアグリゲーションを行った（伊藤（2008，16～17頁））。また、量的属性については、同質属性値レコード群内のレコードを適当な大きさのグループに分割した上で、グループ内のレコードが有する属性値のおのおのを平均値に置き換えることにした。

さらに、本研究では、質的属性のミクロアグリゲーションを実行する上で、「超高次元クロス集計」を行っている。超高次元クロス集計とは、調査されているすべての属性群を対象にクロス集計を行うことである（伊藤（2008，19 頁））。超高次元クロス集計表においては、あらゆる属性群の組合せが集計事項として設定可能であると考えられる（図２）。ゆえに、超高次元クロス集計表に基づいて個体情報の秘匿の要件に適合する質的属性の組合せを探索することが可能である[2]。次に、質的属性と量的属性に関するミクロアグリゲーションの概要を述べる。

(1) 質的属性のミクロアグリゲーションの概要

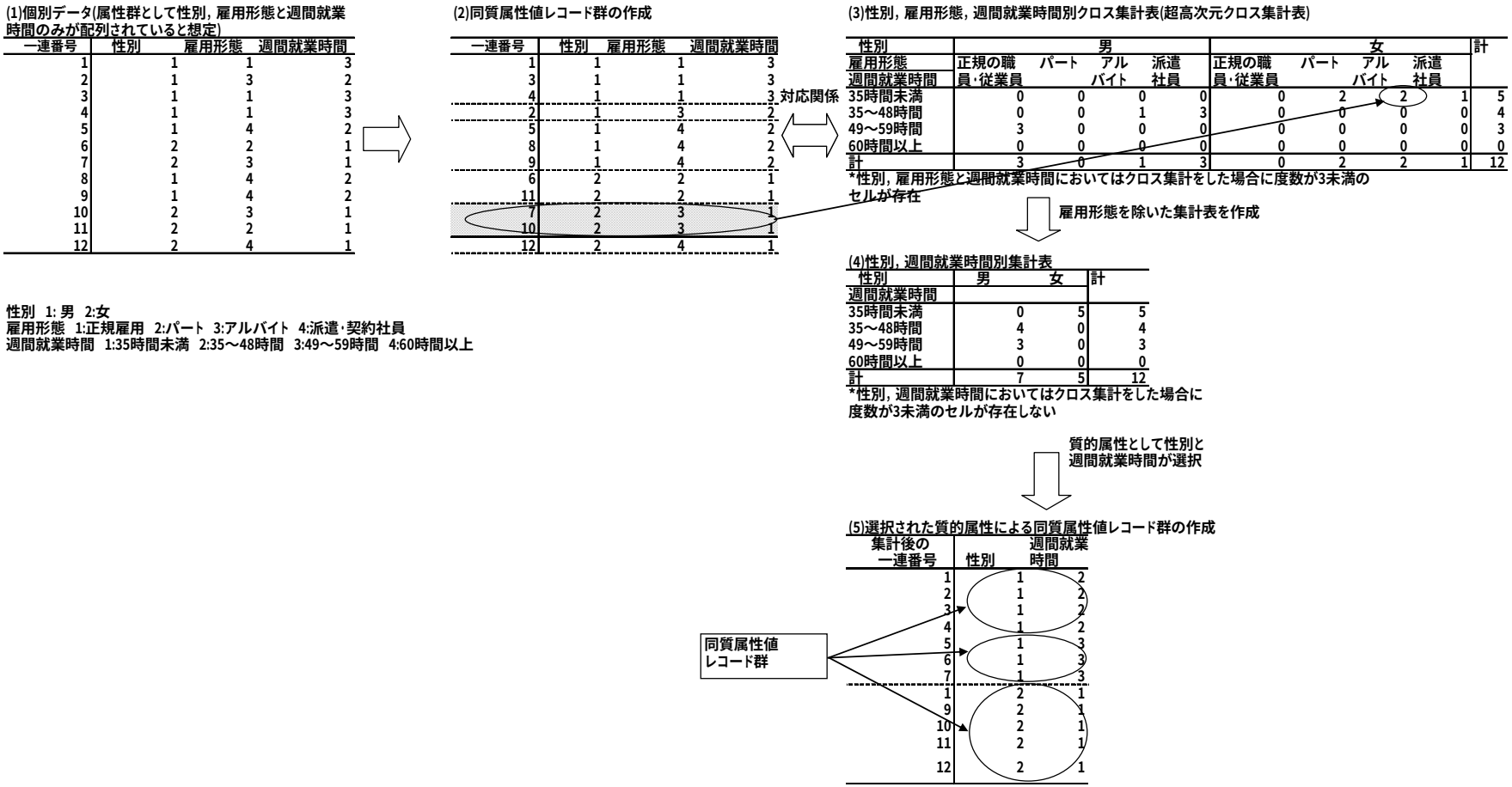
本研究で用いられる質的属性のミクロアグリゲーションの方法は、図３に示すとおりである。図３では、属性群として性別（１：男、２：女）、雇用形態（１：正規の職員・従業員、２：パート、３：アルバイト、４：派遣社員）、及び週間就業時間（１：35 時間未満、２：36～48 時間、３：49 時間～59 時間、４：60 時間以上）の３つの質的属性のみを有する個別データを想定している。この個別データにおけるレコードを性別、雇用形態及び週間就業時間に関する属性値にしたがってグループ化することによって、質的属性についての同質属性値レコード群を編成している。例えば、一連番号７と 10 のレコードにはいずれも、性別は２（女）、雇用形態は３（アルバイト）、週間就業時間は１（35 時間未満）という属性値を含んでいる。

これらの同質属性値レコード群内のレコード数は、性別、雇用形態及び週間就業時間別のク

[1] Community Innovation Survey においては、ミクロアグリゲーションによって質的な属性値をメディアンに置き換える方法が採用されている(Thorogood(1999))。

[2] 同質属性値レコード群内の属性値は同一であることから、属性値を代表値で置換していると捉えることによって、本研究で行われる質的属性のグループ化も「広義の」ミクロアグリゲーションとして位置づけることが可能である(伊藤(2008，16～17 頁))。

図３　質的属性に関するレコードのグループ化の概略図



出所　伊藤（2008，22 頁）

ロス集計表におけるセルの度数と一致している。すなわち、性別が２、雇用形態が３、週間就業時間が１である同質属性値レコード群内のレコード数は２であるが、それは、クロス集計表において、分類項目が「女」、「アルバイト」、及び「35 時間未満」に該当するセルの度数３と合致する。このことは、個別データが持つすべての属性群を対象にして作成された場合の同質属性値レコード群内のレコード数が、それらの属性群を集計事項として表章されたクロス集計表（すなわち「超高次元クロス集計表」）に含まれるセルの度数と対応関係にあることを示している。

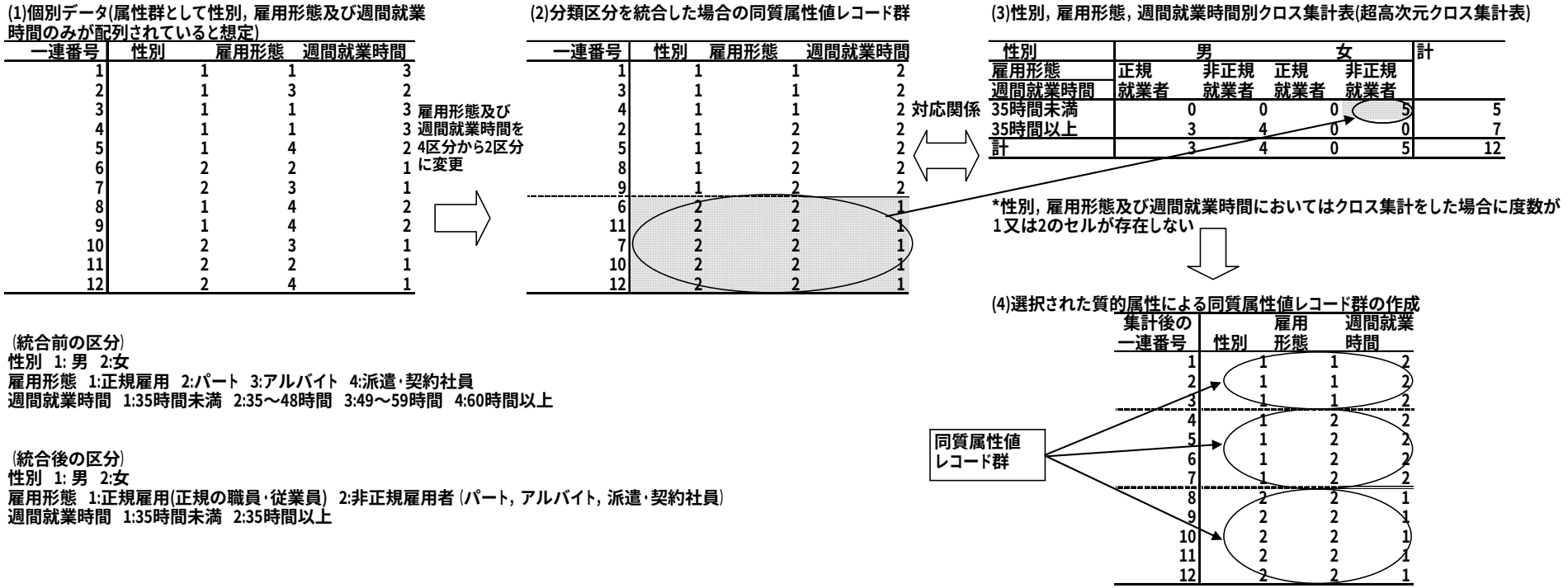
ところで、同質属性値レコード群の作成の基準となる閾値については、適当な値を任意に定めることが可能であるが（先行研究では３～10 で設定)、図３では、例として閾値を３に設定している。このことは、超高次元クロス集計表の集計事項となる属性群から、属性の組合せを適切に選択することによって、度数１又は２のセルが存在しないように集計表を作成することを意味している。例えば、図３では、性別、雇用形態、及び週間就業時間を用いてクロス集計を行うことによって、集計表の中に度数２のセルが出現しているが、そこから性別と週間就業時間のみを集計事項として再集計すれば、度数２のセルは集計表に存在しなくなる。このことから、質的属性として性別と週間就業時間を選択すれれば、同質属性値レコード群内のレコード数が３以上になることがわかる。

他方、質的属性における分類区分を統合することによって、度数１又は２のセルを含まない集計表を作成することも考えられる。図４は、図３と同様の個別データにおいて、雇用形態を従来の４区分から２区分（1：正規雇用者、2：非正規雇用者)、同様に、週間就業時間を４区分から２区分（1：35 時間未満、 2：35 時間以上）に統合した場合における質的属性のグループ化の概略図である。雇用形態と週間就業時間の分類区分を統合することによって、３つの属性群のすべてを集計事項に設定しても、作成された集計表に度数１又は２のセルが存在しないようにすることが可能になる。すなわち、分類区分の統合によって、質的属性の組合せに関する選択の可能性が広がっていると言える。このことは、質的属性の組合せの検討においては、属性群に設定されている分類区分だけでなく、区分を再統合した場合についても、質的属性の組合せを検証する必要があることを示唆している。

(2) 量的属性のミクロアグリゲーションの概要

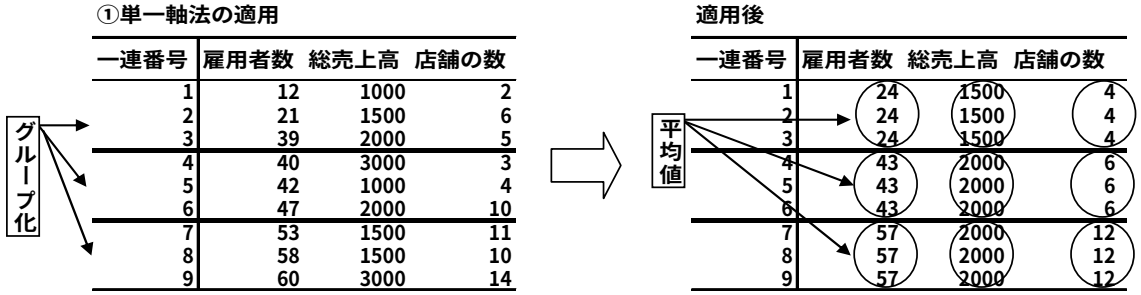
本研究では、最初に質的属性の組合せを検討した上で量的属性のミクロアグリゲーションを実行している。そのために、本研究では、同質属性値レコード群内のレコードを特定のレコー

図４　質的属性に関するレコードのグループ化の概略図　―　分類区分の統合済



出所　伊藤（2008，24 頁）

図５　単一の量的属性におけるミクロアグリゲーション



②第1主成分法の適用

一連番号	雇用者数	総売上高	店舗の数	第1主成分のスコア
1	12	1000	2	-2.4516
2	21	1500	6	-1.1941
3	39	2000	5	-0.322
4	40	3000	3	0.0285
5	42	1000	4	-0.9596
6	47	2000	10	0.7402
7	53	1500	11	0.8237
8	58	1500	10	0.874
9	60	3000	14	2.4611

適用後

一連番号	雇用者数	総売上高	店舗の数	グループの番号
1	25	1167	4	1
2	25	1167	4	1
3	42	2333	6	2
4	42	2333	6	2
5	25	1167	4	1
6	42	2333	6	2
7	57	2000	12	3
8	57	2000	12	3
9	57	2000	12	3

③Zスコア総計法の適用

一連番号	雇用者数	総売上高	店舗の数	雇用者数のZ値	総売上高のZ値	店舗の数のZ値	Zスコア総計値
1	12	1000	2	-1.82093	-1.11111	-1.25939	-4.19413
2	21	1500	6	-1.26233	-0.44444	-0.29475	-2.00143
3	39	2000	5	-0.14485	0.22222	-0.58359	-0.45854
4	40	3000	3	-0.08277	1.55556	-1.10182	0.45455
5	42	1000	4	0.04138	-1.11111	-0.77707	-1.8468
6	47	2000	10	0.35177	0.22222	0.66989	1.24388
7	53	1500	11	0.72423	-0.44444	0.91105	1.19084
8	58	1500	10	1.03462	-0.44444	0.66989	1.26006
9	60	3000	14	1.15877	1.55556	1.63453	4.34884

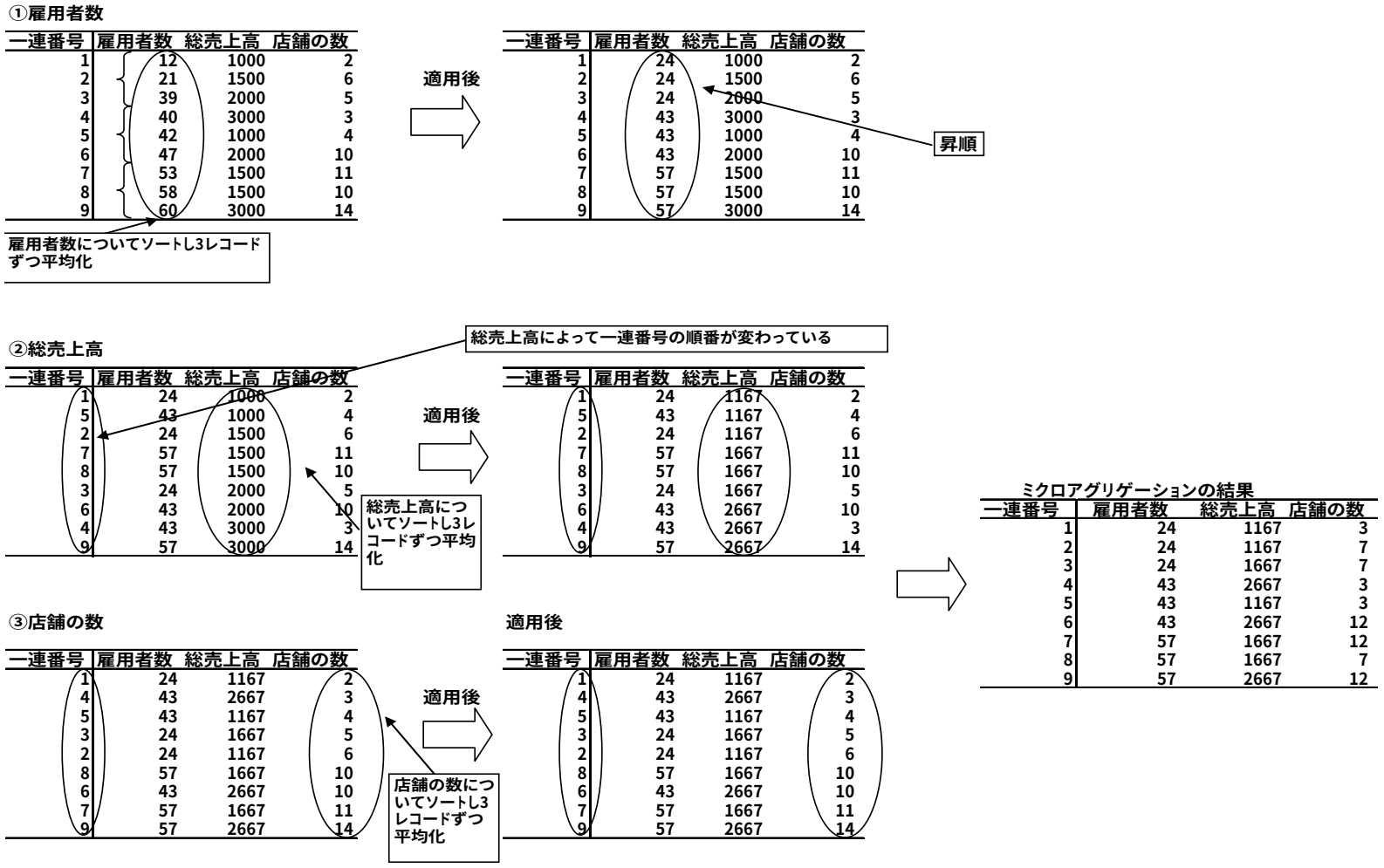
適用後

一連番号	雇用者数	総売上高	店舗の数	グループの番号
1	25	1167	4	1
2	25	1167	4	1
3	44	2167	6	2
4	44	2167	6	2
5	25	1167	4	1
6	55	2167	11	3
7	44	2167	6	2
8	55	2167	11	3
9	55	2167	11	3

*k=3で固定し,各グループにあるレコード群のすべての変数値を平均化

出所　伊藤（2008，９頁）

図６　複数の量的属性群におけるミクロアグリゲーション　―　個別ランキング法の適用



出所　伊藤（2008，10 頁）

ド数ごとにグループ化し、量的属性値のおのおのを平均値に置き換えている。また、本研究では、レコードのグループ化の基準となるレコード数を３に設定している。

量的属性のミクロアグリゲーションに関しては、単一軸法（single axis method）、第１主成分法（first principal component method）、Ｚスコア総計法（sum of Z-scores method）、個別ランキング法（individual ranking method）といった手法が存在する（伊藤（2008, ７～11頁））。最初に、単一軸法とは、ある特定の量的属性（ソートキー）についてソートし、ソートされたレコード群において特定のレコード数（例えば３）ごとにレコードのグループ化を行った上で、それぞれの量的属性値を代表値に置き換える手法である。図５を見ると、単一軸法を適用した場合、雇用者数をソートキーとしてレコードをグループ化することによって、各グループ内の雇用者数、総売上高と店舗の数がそれぞれ平均値に置き換えられていることがわかる。次に、第１主成分法及びＺスコア総計法は、単一の統計指標に基づいてソート及びグループ化を行う方法である。第１主成分法は、ミクロアグリゲーションの手法に主成分分析を適用している。図５では、雇用者数、総売上高及び店舗の数の３つの属性値を標準化した上で第１主成分のスコアを計算し、その数値にしたがってレコード群のソート、及び、レコードのグループ化を行っている。また、Ｚスコア総計法は、各レコードにおける属性値のおのおのを標準化し、それらを総計した値（Ｚスコア総計値）に基づいてソートし、レコードのグループ化を行う手法である。図５では、雇用者数、総売上高及び店舗の数の属性値から算出されたＺスコア総計値によって、レコード群のソートを行っている。

他方、個別ランキング法に関しては、先述の３つのミクロアグリゲーションとは、その特徴が大きく異なる。個別ランキング法は、量的属性のおのおのについて個別にソート及びグループ化を行う方法である。図６は、図５と同様に、雇用者数、総売上高及び店舗の数を例にして、個別ランキング法の概要を示したものである。最初に、雇用者数をソートキーにしてレコードのグループ化を行い、レコードが有する属性値を平均値に置き換えている。さらに、総売上高、店舗の数の順でソートを行い、おのおのの属性値を平均値に置き換えている[3]

３　ミクロアグリゲーションにおける評価基準

匿名化技法としてミクロアグリゲーションが適用される場合には、個別データにおける秘匿情報の保護とミクロデータの有用性の両面から、ミクロアグリゲーションの評価基準が具体的

[3] Eurostat の Community Innovation Survey(1994)では、量的属性において個別ランキング法が用いられている(Thorogood(1999, p.31))。

に検討されることが求められる。そこで本研究は、ミクロデータにおける秘匿性と有用性に関して次のような観点から評価を試みている(伊藤(2008，14～16 頁))。

(1)ミクロアグリゲーションにおける秘匿性

ミクロアグリゲーションは、政府統計の集計表で適用されている秘匿の方法にその着想を得ている（Defays （1999，p.223)）。集計表のセルに含まれる度数が３未満である場合、そのセルは、個体を特定化するリスクの高いセンシティブな（sensitive）度数であるとみなされる（Federal Committee on Statistical Methodology（2005，p.24)）。そして、集計表に度数１又は２となるセルが含まれる場合には、集計表における秘匿の観点から、通常、度数を X に置き換える欠測化（suppression）等の秘匿措置がとられてきた。この考え方を本研究で用いているミクロアグリゲーションの手法に適用すると、次のようになる。すなわち、ミクロアグリゲーションによって編成された同質属性値レコード群内のレコードの数が１又は２である場合、個体情報が特定されるリスクが非常に高くなるが、少なくとも３レコード存在すれば、秘匿性を高めることができると考えられる。このことから、本研究はミクロアグリゲーションにおける閾値を３に設定している。

(2)ミクロアグリゲーションにおける有効性

ミクロデータの有用性に関する評価基準としては、秘匿処理が施されていない個別データと秘匿処理済データ（protected data）においてデータ構造の近似性の程度を検証することが考えられる。ミクロアグリゲーション済のデータ（以下、ミクロアグリゲートデータ（micro-aggregated data）と呼ぶ） のデータ構造についても、秘匿処理が施されていない個別データのそれと比較・検証を行うことによって、ミクロアグリゲートデータの有用性が評価できる。そのための統計指標として、次のような指標が提案されている（Mateo-Sanz，Domingo-Ferrer and Sebé（2005，pp.182-184)）。

①平均、分散等の基本統計量

②分布上の特性

③情報量損失（information loss）

情報量損失では、秘匿処理済データが、秘匿処理が施されていない個別データと比べてどの程度情報量を失っているかを算出するために、それぞれのデータから求められた統計指標における差異の程度が評価される。本研究では、秘匿処理が施されていない個別データとミクロア

グリゲートデータの両方についてそれぞれ相関係数行列を算出し、相関係数行列における平均平方誤差（mean square error）を計算することによって、ミクロアグリゲートデータに関する情報量損失の程度を計測している。そして、情報量損失が最小となるデータを最も望ましいミクロアグリゲートデータであるとみなしている。

4 『全国消費実態調査』の個別データによるミクロアグリゲーションの検証

本節では、『全国消費実態調査（以下、『全消』と呼ぶ)』の個別データに基づいて、我が国におけるミクロアグリゲーションの方法的な可能性を検討する。先述したように、本研究の特徴は、属性群を量的属性と質的属性に類別した上で、ミクロアグリゲーションの手法を個別に適用していることである。そのために、本研究は、次の２つの手順から成っている。

研究１ 質的属性の組合せの検討

研究２ 量的属性のミクロアグリゲーションと有効性の検証

本研究では、最初に、超高次元クロス集計に基づいて、質的属性の組合せパターンを検討する（研究１）。次に、量的属性を対象にミクロアグリゲーションを行うことによって、ミクロアグリゲートデータの作成を試みた上で、『全消』を用いたミクロアグリゲーションの有効性の検証を行う（研究２）。なお、本研究で使用したデータは、平成 16 年の『全消』の個別データ（以下、「原データ」と呼ぶ）であり、用途分類に関する属性群をそなえた二人以上の世帯約 55,000 レコードである[4]。『全消』は、消費支出などの約 300 の量的属性を調査していることから、主に量的属性のミクロアグリゲーションに関する有効性の検証に適したデータであると考えることができる。さらに、本研究では、目的外使用申請及び報告書における調査項目の使用頻度に着目し、使用回数の多い調査項目を本研究で使用する属性群として選定している。次に、本研究の概要を述べる。

(1) 質的属性の組合せに関する検討

研究１では、ミクロアグリゲートデータを作成するための第１段階として、『全消』の個別データを用いた質的属性のミクロアグリゲーションを行った。本研究では、①世帯人員区分、

[4] 本研究では、二人以上の世帯に関する個別データを用いた研究に先立ち、ミクロアグリゲートデータの作成方法を検討するための予備的研究として、『全消』の単身世帯の個別データ(標本数は約 5,000)を用いたミクロアグリゲーションの実験を試みた。しかし、対象となった質的属性について編成された同質属性値レコード群において、レコード数１又は２を含まないような質的属性の組合せパターンは少なく、ミクロアグリゲートデータに含まれる質的属性の組合せのパターン数が限られることがわかった。よって、単身世帯の個別データについては、利用可能なミクロアグリゲートデータを具体的に作成するまでには至らなかった。

②就業人員区分、③住居の建て方、④住居の所有関係、⑤世帯主の性別、⑥世帯主の就業・非就業の別、⑦企業規模、⑧職業符号の８つの質的属性を分析の対象として選んでいる。

本研究では、最初に、研究の対象となるすべての質的属性群について、原データの属性における分類区分（以下、「原区分」と呼ぶ）にしたがって超高次元クロス集計表を作成した。次に、この超高次元クロス集計表に基づいて、クロス集計表の中のセルに度数１又は２を含まない質的属性の組合せの探索を行い（以下、「実験１」と呼ぶ）、これらの結果から、同質属性値レコード群内のレコード数１又は２の有無を判別するための質的属性の組合せリストを作成した。このリストを用いて、ミクロアグリゲートデータ上に設定可能な質的属性群を選別することが可能になる。例えば、図７は、性別及び就業・非就業の別という２つの質的属性を対象に組合せリストの作成手順を図示したもので、次の２つの手順からなっている。

ア　原データから、性別と就業・非就業の別に関するクロス集計表を作成する。図７では、㋐性別、㋑就業・非就業の別、㋒性別と就業・非就業の別の３つの質的属性の組合せがクロス集計の対象である。

イ　このクロス集計表に基づいて、質的属性の組合せリストを作成する。

質的属性の組合せリストは、質的属性の組合せのパターンごとに同質属性値レコード群内におけるレコード数１又は２の有無に関する判定結果を表示したもので、リスト上にレコード数１又は２の有無欄が無と表示されている質的属性の組合せパターンについてのみ、ミクロアグリゲートデータの作成が可能であると判断できる。なお、実験１の質的属性の組合せリストは、別添１を参照されたい。

実験１の結果から、原区分データを使用した場合、同質属性値レコード群内におけるレコード数が１又は２でない質的属性の組合せが、全 255 パターン中 18 パターン（全体の 7.1%）であった。また、質的属性の組合せの数は、最大で３になることがわかった。このうち、属性数が最大となる質的属性の組合せは、①性別２区分×就業・非就業４区分×企業規模５区分、及び②性別２区分×就業・非就業４区分×職業符号 12 区分の２パターンであった。

本分析の結果は、『全消』の原区分データを使用した場合には、質的属性の組合せを探索する上で大きな制約があることを示唆している。この問題点を解決する方法として、質的属性の分類区分を統合することによって、ミクロアグリゲートデータに設定可能な質的属性を増やすことが考えられる。そこで、本研究では、原データにおける各属性の分類区分を結果表の集計事項の中で最も少ない区分に統合したデータ（以下、「統合区分データ」と呼ぶ）を用いて、質的属性の組合せを再度検討した（以下、「実験２」と呼ぶ）。表１は、世帯人員区

図７　質的属性の組合せリスト作成

①データ

都道府県番号	市区町村番号	調査単位区符号	性別	就業・非就業の別
01	101	11	1	2
01	101	12	1	2
01	101	13	1	3
01	101	14	1	3
01	101	15	1	3
01	101	16	1	4
01	101	17	1	4
01	101	18	1	4
01	101	19	2	1
01	102	11	2	2
01	102	12	2	2
01	102	13	2	2
01	102	14	2	3
01	102	15	2	3
01	102	16	2	3
01	102	17	2	4
01	102	18	2	4
01	102	19	2	4

②質的属性別度数分布表

㋐性別

性別	
1	2
男	女
8	10

㋑就業・非就業の別

就業・非就業の別			
1	2	3	4
就業	うちパート	非就業	うち仕事を探している
1	5	6	6

㋒性別、就業・非就業の別

			就業・非就業の別			
			1	2	3	4
			就業	うちパート	非就業	うち仕事を探している
性別	1	男	0	2	3	3
	2	女	1	3	3	3

③質的属性の組合せリスト

	性別	就業・非就業の別	レコード数1又は2の有無
㋐	＊		無
㋑		＊	有
㋒	＊	＊	有

⇨ ミクロアグリゲートデータ作成可能

分に関して原区分と統合区分の相違を示したものである。なお、別添２は、８つの質的属性について原区分と統合区分を比較した一覧表である。また、実験２における質的属性の組合せリストを、別添３で示している。

表１　世帯人員区分における原区分と統合区分の相違

①原区分

符号	項目名	件数	構成比
2	２人	19,643	36%
3	３人	13,696	25%
4	４人	12,860	23%
5	５人	5,967	11%
6	６人	1,919	3%
7	７人	739	1%
8	８人	185	0%
9	９人	42	0%
10	10人以上	5	0%
総計		55,056	100%

②統合区分

符号	項目名	件数	構成比
2	２人	19,643	36%
3	３人	13,696	25%
4	４人	12,860	23%
5	５人	5,967	11%
6	６人以上	2,890	5%
総計		55,056	100%

統合区分データにおいては、同質属性値レコード群内において、レコード数１又は２が存在しない質的属性の組合せが、全 255 パターン中 77 パターン（全体の 30.2%）存在し、その組合せの数は、最大で５になることがわかった。また、属性数が最大となる組合せは、性別２区分×就業・非就業４区分×企業規模４区分×住居の所有関係２区分×職業符号４区分の１パターンであった。なお、表２は、原区分データと統合区分データによる質的属性の組合せの結果をまとめたものである。

表２　質的属性の組合せパターンに関する検証結果の要約

各属性の区分		実験１ 原区分	実験２ 統合区分
レコード数 １又は２がない	パターン数 全255パターン中	18パターン	77パターン
	パターン率	7.1%	30.2%
	属性数	1～3属性	1～5属性
	最多クロス数	96クロス	128クロス

(2) 量的属性のミクロアグリゲーションと有効性の検証

研究２では、『全消』の個別データを用いて、量的属性のミクロアグリゲーションを行った。本研究では、研究１で作成した質的属性の組合せリストの中から、質的属性群として性別２区分、就業・非就業４区分、及び企業規模５区分を選択した上で編成したデータ（以下、「質的属性選択済データ」と呼ぶ）について、同質属性値レコード群の中で３レコードずつグループ化した上で、量的属性値を平均値に置き換えた。また、本研究では、①年間収入、②消

費支出、③貯蓄現在高、④負債現在高、及び、⑤年齢（世帯主）の５つの量的属性を研究の対象として選んでいる。

次に、量的属性におけるミクロアグリゲーションの手順について述べる。

研究２では、質的属性選択済データを用いて、量的属性群に対して次の２種類のミクロアグリゲーションの方法を適用した。第１のミクロアグリゲーションの方法は、質的属性選択済データの最初の配列順にしたがって３レコードずつグループ化を行い、量的属性値のおのおのを平均値に置き換えている（以下、「ソートなし」と呼ぶ）。図８は、質的属性として性別、就業・非就業の別と企業規模、量的属性として年間収入と消費支出をそれぞれ有する原データに対して、ソートなしによるミクロアグリゲーションを適用した例である。図８では、最初に、同質属性値レコード群内で３レコードずつグループ化を行った。次に、年間収入と消費支出について乗率（母集団復元乗率）の重みをつけた加重平均値に置き換え、乗率についてはその平均値に置き換えることによって、ミクロアグリゲートデータを作成した[5]。なお、量的属性のミクロアグリゲーションにおいて、対象となる同質属性値レコード群内のレコードの総数が３で割り切れない場合には、そのレコード群内の最後のグループにおけるレコード数が４ないしは５になるように設定した。

図８　量的属性のミクロアグリゲーション―ソートなし

原データ

都道府県番号	市区町村番号	調査単位区符号	性別	就業・非就業の別	企業規模	年間収入	消費支出	乗率
08	109	13	1	1	2	2000	1000	100
21	101	11	1	1	1	2000	1000	100
44	104	16	1	1	2	2000	1000	100
04	106	17	1	2	3	2000	1000	900
15	104	11	1	2	4	3000	2000	800
18	105	17	1	2	4	4000	3000	700
30	106	13	1	2	5	5000	4000	600
34	105	19	1	2	2	6000	5000	500
20	105	15	1	3	4	1429	1184	354
22	108	14	1	3	5	5144	3643	184
26	107	15	1	3	4	6559	1010	678
28	106	19	1	3	4	7631	1824	920
41	109	18	1	3	1	8004	7437	163
43	101	18	1	3	2	9052	7542	101
：	：	：	：	：	：	：	：	：

ミクロアグリゲートデータ

	性別	就業・非就業の別	年間収入	消費支出	乗率
㋐	1	1	2000	1000	100
	1	1	2000	1000	100
	1	1	2000	1000	100
㋑	1	2	3714	2714	700
	1	2	3714	2714	700
	1	2	3714	2714	700
	1	2	3714	2714	700
	1	2	3714	2714	700
㋒	1	3	4851	1459	405
	1	3	4851	1459	405
	1	3	4851	1459	405
	1	3	7804	3085	395
	1	3	7804	3085	395
	1	3	7804	3085	395
	：	：	：	：	：

第２のミクロアグリゲーションの方法は、個別ランキング法の適用で、質的属性選択済データにおける量的属性のおのおのについてソートを行った上で、ミクロアグリゲートデータ

5 本研究では、ミクロアグリゲートデータによる母集団への復元を指向していることから、乗率を用いてミクロアグリゲーションを行っている。

を作成する方法である（以下、「個別ランキング法」と呼ぶ）。ソートなしと同様のデータを用いて行った個別ランキング法によるミクロアグリゲーションは、次のとおりである（図９）。最初に、原データについて、年間収入をキーとして昇順で並べ替えた上で（以下、「年間収入ソート済データ」と呼ぶ）、同質属性値レコード群内を３レコードずつグループ化し、年間収入を乗率の重みをつけた加重平均値に置き換えた。次に、乗率の平均値を年間収入用乗率として付与することによって、「年間収入ミクロアグリゲート済データ」を作成した。さらに、年間収入ミクロアグリゲート済データについて、消費支出をキーとして昇順で並べ替えた上で（以下、「消費支出ソート済データ」と呼ぶ）、消費支出を乗率の重みをつけた加重平均値に置き換えた。最後に、乗率の平均値を消費支出用乗率として付与し、ミクロアグリゲートデータを作成した[6]。

図９　量的属性のミクロアグリゲーション―個別ランキング法

年間収入ソート済データ

都道府県番号	市区町村番号	調査単位区符号	性別	就業・非就業の別	企業規模	年間収入	消費支出	乗率	
08	109	13	1	1	2	2000	1000	100	㋐
21	101	11	1	1	1	2000	1000	100	
44	104	16	1	1	2	2000	1000	100	
04	106	17	1	2	3	2000	1000	900	㋑
15	104	11	1	2	4	3000	2000	800	
18	105	17	1	2	4	4000	3000	700	
30	106	13	1	2	5	5000	4000	600	
34	105	19	1	2	2	6000	5000	500	
20	105	15	1	3	4	1429	1184	354	㋒
22	108	14	1	3	5	5144	3643	184	
26	107	15	1	3	4	6559	1010	678	
28	106	19	1	3	4	7631	1824	920	
41	109	18	1	3	1	8004	7437	163	
43	101	18	1	3	2	9052	7542	101	
：	：	：	：	：	：	：	：	：	

[6] 本研究では、乗率を使用していることから、ソートの対象となる量的属性については、その属性値に乗率の重みをつけた値にそってソートが行われている。そのために、ソートによるレコードの並び順が、乗率を適用しなかった場合の並び順と異なることが考えられる。また、『全消』の個別データにおいては、年間収入や消費支出といった総計値を表す量的属性は、その内訳を表す属性群の合計に一致するように設定されている（「加法性」）。このような加法性は、ソートなしのミクロアグリゲーションについてはそのまま保持されている。しかし、個別ランキング法では、量的属性のおのおのについてソートとグループ内の平均値への置き換えを行っているため、『全消』の個別データに設定されていた加法性が保持できない場合がある。

年間収入ミクロアグリゲート済データ

	都道府県番号	市区町村番号	調査単位区符号	性別	就業・非就業の別	年間収入	消費支出	年間収入乗率	乗率
㋐	08	109	13	1	1	2000	1000	100	100
	21	101	11	1	1	2000	1000	100	100
	44	104	16	1	1	2000	1000	100	100
㋑	04	106	17	1	2	3714	1000	700	900
	15	104	11	1	2	3714	2000	700	800
	18	105	17	1	2	3714	3000	700	700
	30	106	13	1	2	3714	4000	700	600
	34	105	19	1	2	3714	5000	700	500
㋒	20	105	15	1	3	4851	1184	405	354
	22	108	14	1	3	4851	3643	405	184
	26	107	15	1	3	4851	1010	405	678
	28	106	19	1	3	7804	1824	395	920
	41	109	18	1	3	7804	7437	395	163
	43	101	18	1	3	7804	7542	395	101
	：	：	：	：	：	：	：	：	

消費支出ソート済データ

都道府県番号	市区町村番号	調査単位区符号	性別	就業・非就業の別	年間収入	消費支出	年間収入乗率	乗率	
08	109	13	1	1	2000	1000	100	100	㋐
21	101	11	1	1	2000	1000	100	100	
44	104	16	1	1	2000	1000	100	100	
04	106	17	1	2	3714	1000	700	900	㋑
15	104	11	1	2	3714	2000	700	800	
18	105	17	1	2	3714	3000	700	700	
30	106	13	1	2	3714	4000	700	600	
34	105	19	1	2	3714	5000	700	500	
26	107	15	1	3	4851	1010	405	678	㋒
20	105	15	1	3	4851	1184	405	354	
28	106	19	1	3	7804	1824	395	920	
22	108	14	1	3	4851	3643	405	184	
41	109	18	1	3	7804	7437	395	163	
43	101	18	1	3	7804	7542	395	101	
：	：	：	：	：	：	：	：	：	

ミクロアグリゲートデータ

	性別	就業・非就業の別	年間収入	消費支出	年間収入乗率	消費支出乗率
㋐	1	1	2000	1000	100	100
	1	1	2000	1000	100	100
	1	1	2000	1000	100	100
㋑	1	2	3714	2714	700	700
	1	2	3714	2714	700	700
	1	2	3714	2714	700	700
	1	2	3714	2714	700	700
	1	2	3714	2714	700	700
㋒	1	3	4851	1425	405	651
	1	3	4851	1425	405	651
	1	3	4851	1425	405	651
	1	3	7804	5902	395	149
	1	3	7804	5902	395	149
	1	3	7804	5902	395	149
	：	：	：	：	：	：

ソートなしと個別ランキング法という２つの方法を用いて作成した２種類のミクロアグリゲートデータについては、それぞれの分布特性を原データの分布と比較し、量的属性のミクロアグリゲーションの有効性を検証した。

最初に、表３は、原データ、ソートなし、及び個別ランキング法の３種類のデータについて、５つの量的属性（年間収入、消費支出、貯蓄現在高、負債現在高、年齢）の平均値を比較したものである。当然ではあるが、ミクロアグリゲートデータの平均値については、ソートなしと個別ランキング法のいずれも原データの値に等しくなっている。

表３　原データ、ソートなしと個別ランキング法における量的属性の平均値

	年間収入	消費支出	貯蓄現在高	負債現在高	年齢
①原データ	692.47	320063.2267	1454.3004	542.5833	53.6547
②実験１　ソートなし	692.47	320063.2267	1454.3004	542.5833	53.6547
③実験２　個別ランキング法	692.47	320063.2267	1454.3004	542.5833	53.6547

また、表４は、データの散らばりの程度を比較するため、３種類のデータについて、量的属性の標準偏差を比較したものである。標準偏差については、個別ランキング法の方がソートなしよりも原データの値に近いことがわかる。

表４　原データ、ソートなしと個別ランキング法における量的属性の標準偏差

	年間収入	消費支出	貯蓄現在高	負債現在高	年齢
①原データ	466.5494	199778.3556	1993.8483	1246.4929	13.9158
②実験１　ソートなし	316.37	125780.5462	1247.6351	784.5321	11.2773
③実験２　個別ランキング法	464.9756	199246.4	1987.215	1242.348	13.9148

次に、図 10 及び図 11 はそれぞれ、３種類のデータにおける年齢 10 歳階級別世帯数分布別及び年間収入 10 区分階級別のヒストグラムである。図 10 と図 11 から、ソートなしにおける分布が原データの分布と大きく異なるのに対して、個別ランキング法における分布は原データの分布と比べてその形状が非常に似通っていることがわかる。さらに原データからの情報量損失の指標として、分布特性の相対係数行列（表５）を求めた上で、これらの相関係数行列から得られる平均平方誤差の値を算出している。平均平方誤差については、ソートなしが 0.0105557、個別ランキング法が 0.0000037 となることから、個別ランキング法の場合、ソートなしと比較して平均平方誤差の値が相対的に小さくなることがわかる。以上の結果から、個別ランキング法によって作成したミクロアグリゲートデータは、ソートなしによるデータ

図 10　原データ、ソートなし、個別ランキング法の年齢 10 歳階級別世帯数分布

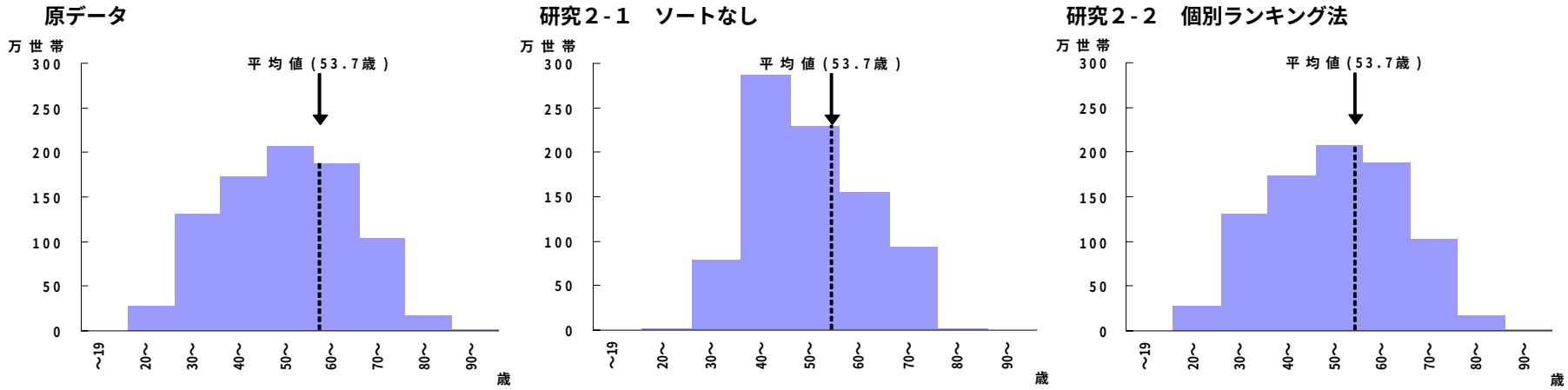


図 11　原データ、ソートなし、個別ランキング法の年間収入 10 区分階級別世帯数分布

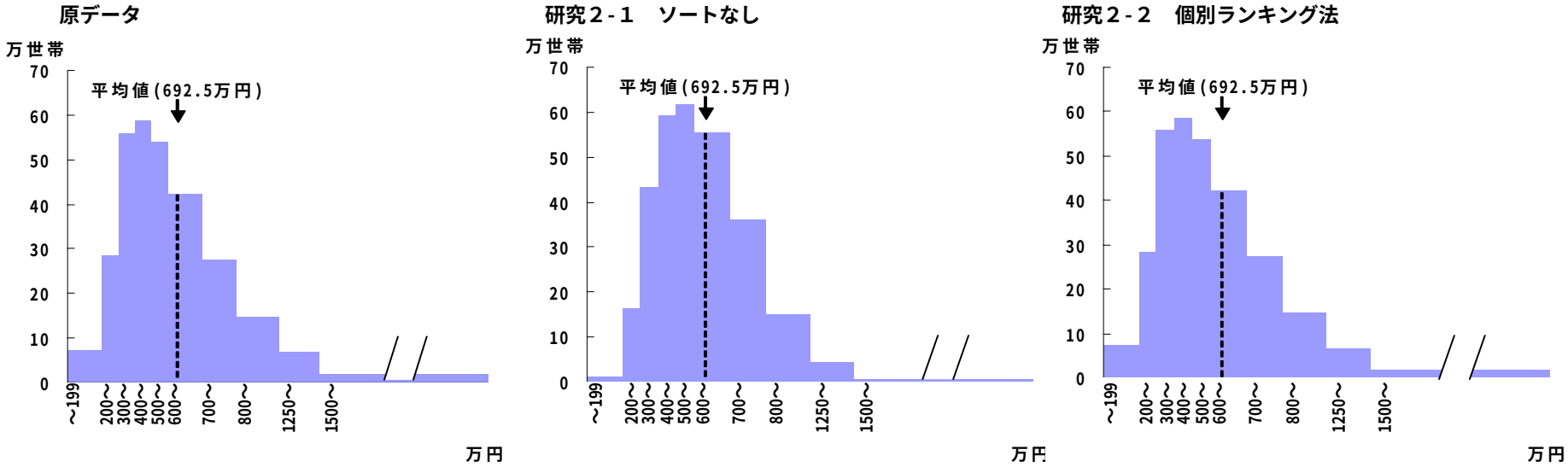


表５　原データ、ソートなし、個別ランキング法における量的属性間の相関係数行列

原データ

	年間収入	消費支出	貯蓄現在高	負債現在高	年齢
年間収入	1	0.4147353	0.3282868	0.2867635	-0.0461703
消費支出	0.4147353	1	0.2342448	0.0697177	-0.0279208
貯蓄現在高	0.3282868	0.2342448	1	-0.0388443	0.2754914
負債現在高	0.2867635	0.0697177	-0.0388443	1	-0.1864024
年齢	-0.0461703	-0.0279208	0.2754914	-0.1864024	1

ソートなし

	年間収入	消費支出	貯蓄現在高	負債現在高	年齢
年間収入	1	0.5027258	0.2543242	0.3694126	-0.2599332
消費支出	0.5027258	1	0.2235620	0.1495779	-0.1514459
貯蓄現在高	0.2543242	0.2235620	1	-0.0768227	0.3364917
負債現在高	0.3694126	0.1495779	-0.0768227	1	-0.3000032
年齢	-0.2599332	-0.1514459	0.3364917	-0.3000032	1

個別ランキング法

	年間収入	消費支出	貯蓄現在高	負債現在高	年齢
年間収入	1	0.4180874	0.3251327	0.2837790	-0.0467374
消費支出	0.4180874	1	0.2346149	0.0712573	-0.0287049
貯蓄現在高	0.3251327	0.2346149	1	-0.0396477	0.2762609
負債現在高	0.2837790	0.0712573	-0.0396477	1	-0.1880405
年齢	-0.0467374	-0.0287049	0.2762609	-0.1880405	1

よりも原データに近似的であり、個別ランキング法のデータが相対的に情報量損失の少ないミクロアグリゲートデータであると結論付けることができる。

5　結びにかえて

本稿では、平成16年の『全消』の個別データを用いて、我が国におけるミクロアグリゲーションの有効性の検証を行った。質的属性のミクロアグリゲーションを実行するために、対象となる質的属性群について超高次元クロス集計表を作成した上で、クロス集計表に含まれるセルに度数１又は２が存在しない質的属性の組合せを検討した。その結果、質的属性の組合せとして、原区分については最大３属性、統合区分について最大５属性となった。しかし、今回用いた『全消』の標本数が55,000程度（二人以上の世帯）であったことから、組合せ可能な質的属性の数を増やす上では限界があることもわかった。また、量的属性のミクロアグリゲーションの手法については、ソートなしと個別ランキング法を比較した結果、個別ランキング法のほうがソートなしよりも原データに近似したミクロアグリゲートデータを作成できることが明らかになった。

本研究は、我が国の政府統計の個別データを用いたミクロアグリゲーションを検証する最初の試みであることから、本研究で用いたミクロアグリゲーションの方法については、試行的な側面があることは否めないと思われる。よって、個別データに適用可能なミクロアグリゲーシ

ョンの方法に関しては、さらなる検討が必要ではないかと考えている。

本稿では、平成 16 年の『全消』を用いてミクロアグリゲーションの有効性を検証しているが、『全消』では標本数の制約もあることから、①地域区分を考慮した質的属性の組合せの検討、②質的属性における秘匿（質的属性値の再区分化等）に関する検証等を行うことができなかった。これについては、『住宅・土地統計調査』のような標本数の大きな調査の個別データを用いて、より精密な検証を行うことを考えている。なお、本研究では、ミクロアグリゲーションの有効性の観点からミクロアグリゲーションの手法について比較・検討を行ったが、秘匿の観点からミクロアグリゲートデータにおける開示リスクの計測を行っていない。今後は、ミクロアグリゲーショトデータに関する個体の特定の可能性を定量的に測るために、ミクロアグリゲーションにおける開示リスクの評価方法を具体的に検討することを考えている。また、ミクロアグリゲーションについては、本稿で検証したミクロアグリゲーションの手法以外にも、様々な手法が存在している（伊藤 (2008 , 6～ 14 頁) 。これらの手法が適用可能であるかについても、個別データを用いて具体的に検討を行う必要があると思われる。さらに、本研究で用いたミクロアグリゲーションの方法が、『全消』以外の調査（例えば、事業所・企業を対象とする『事業所・企業統計調査 』） にも一般的な方法として、適用可能であるかについても検証することを考えている。これらの研究課題について、ミクロアグリゲーションの有効性と秘匿性の両面から、我が国におけるミクロアグリゲーションの方法的な可能性をこれからも追究していきたい。

参考文献

Defays, D.(1997) “Protecting Micro-Data By Micro-Aggregation:The Experience in Eurostat”, *QÜESTIIÓ*, vol.21, 1 i 2, pp.221-231

http://www.idescat.net/sort/questiio/questiiopdf/21.1.10.Defays.pdf

Defays, D. and Anwar, M.N.(1998) “Masking Microdata Using Micro-Aggregation”, *Journal of Official Statistics*, Vol.14, No.4, pp.449-461.

Domingo-Ferrer, J. and Torra, V.(2001) ”Disclosure Control Methods and Information Loss for Microdata”, Doyle *et al.*(eds.)(2001)*Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, Elsevier Science, Amsterdam, pp.91-110.

Domingo-Ferrer, J. and Mateo-Sanz, J. M.(2002) ”Practical Data-oriented Microaggregation for Statistical

Disclosure Control", *IEEE Transactions on Knowledge and Data Engineering,* vol.14, no.1, pp.189-201.

Federal Committee on Statistical Methodology (2005) *Statistical Policy Working Paper 22(Second version): Report on Statistical Disclosure Limitation Methodology*. Federal Committee on Statistical Methodology, U.S. Office of Management and Budget, Washington, D.C.

伊藤伸介 (2008) 「ミクロアグリゲーションに関する研究動向」『製表技術参考資料』 No.10 , 3 ～31頁

Mateo-Sanz, J. M., Domingo-Ferrer, J., Sebé, F.(2005) "Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata" *Data Mining and Knowledge Discovery,* vol.11, pp.181-193.

Pagliuca, D. and Seri, G.(1998) "The Release of Business Microdata: A Software Prototype for Microaggregation"
http://europa.eu.int/en/comm/eurostat/research/conferences/ntts-98/papers/cp/058c.pdf

Pagliuca, D. and Seri, G.(1999)"Masking Business Microdata with MASQ"
http://europa.eu.int/en/comm/eurostat/research/conferences/etk-99/papers/pagliuca-seri.pdf

竹村彰通 (2003) 「個票開示問題の研究の現状と課題」『統計数理』第51巻第２号, 241 ～ 260頁

Thorogood D.(1999) "Protecting the Confidentiality of Eurostat Statistical Outputs", *Netherlands Official Statistics*, Volume 14, Spring, pp.30-33.

Torra, V.(2004) "Microaggregation for Categorical Variables: A Model Based Approach", Domingo-Ferrer, J. and Torra, V.(eds) *Privacy in Statistical Databases CASC Project Final Conference PSD 2004 Barcelona Catalonia, Spain, June9-11, 2004 Proceedings*, Springer, pp.162-174.

Tzavidis, N. and Panaretos, J. (2001) *Aspects of Estimation Procedures at Eurostat with Some Emphasis in the Over-space Harmonisation*, Athens, Greece, Department of Statistics, Athens University of Economics
http://stat-athens.aueb.gr/~jpan/diatrives/Tzavidis/Index.html

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

別添1
実験1 原区分の質的属性の組合せリスト

＊は質的属性として選択

世帯人員区分	就業人員区分	住居の所有関係(現住居)	住居の建て方(現住居)	就業･非就業の別(世帯主)	企業規模(世帯主)	職業符号(世帯主)	性別(世帯主)	レコード数1又は2の有無
*								無
	*							有
		*						無
			*					無
				*				無
					*			無
						*		無
							*	無
*	*							有
*		*						有
*			*					有
*				*				有
*					*			有
*						*		有
*							*	有
	*	*						有
	*		*					有
	*			*				有
	*				*			有
	*					*		有
	*						*	有
		*	*					有
		*		*				無
		*			*			無
		*				*		有
		*					*	無
			*	*				有
			*		*			有
			*			*		有
			*				*	無
				*	*			無
				*		*		無
				*			*	無
					*	*		有
					*		*	無
						*	*	無
*	*	*						有
*	*		*					有
*	*			*				有
*	*				*			有
*	*					*		有
*	*						*	有
*		*	*					有
*		*		*				有
*		*			*			有
*		*				*		有
*		*					*	有
*			*	*				有
*			*		*			有
*			*			*		有
*			*				*	有
*				*	*			有
*				*		*		有
*				*			*	有
*					*	*		有
*					*		*	有
*						*	*	有
	*	*	*					有
	*	*		*				有
	*	*			*			有
	*	*				*		有

＊ は質的属性として選択

世帯人員区分	就業人員区分	住居の所有関係(現住居)	住居の建て方(現住居)	就業･非就業の別(世帯主)	企業規模(世帯主)	職業符号(世帯主)	性別(世帯主)	レコード数1又は2の有無
	＊	＊					＊	有
	＊		＊	＊				有
	＊		＊		＊			有
	＊		＊			＊		有
	＊		＊				＊	有
	＊			＊	＊			有
	＊			＊		＊		有
	＊			＊			＊	有
	＊				＊	＊		有
	＊				＊		＊	有
	＊					＊	＊	有
		＊	＊	＊				有
		＊	＊		＊			有
		＊	＊			＊		有
		＊	＊				＊	有
		＊		＊	＊			有
		＊		＊		＊		有
		＊		＊			＊	有
		＊			＊	＊		有
		＊			＊		＊	有
		＊				＊	＊	有
			＊	＊	＊			有
			＊	＊		＊		有
			＊	＊			＊	有
			＊		＊	＊		有
			＊		＊		＊	有
			＊			＊	＊	有
				＊	＊	＊		有
				＊	＊		＊	無
				＊		＊	＊	無
					＊	＊	＊	有
＊	＊	＊	＊					有
＊	＊	＊		＊				有
＊	＊	＊			＊			有
＊	＊	＊				＊		有
＊	＊	＊					＊	有
＊	＊		＊	＊				有
＊	＊		＊		＊			有
＊	＊		＊			＊		有
＊	＊		＊				＊	有
＊	＊			＊	＊			有
＊	＊			＊		＊		有
＊	＊			＊			＊	有
＊	＊				＊	＊		有
＊	＊				＊		＊	有
＊	＊					＊	＊	有
＊		＊	＊	＊				有
＊		＊	＊		＊			有
＊		＊	＊			＊		有
＊		＊	＊				＊	有
＊		＊		＊	＊			有
＊		＊		＊		＊		有
＊		＊		＊			＊	有
＊		＊			＊	＊		有
＊		＊			＊		＊	有
＊		＊				＊	＊	有
＊			＊	＊	＊			有
＊			＊	＊		＊		有
＊			＊	＊			＊	有
＊			＊		＊	＊		有
＊			＊		＊		＊	有
＊			＊			＊	＊	有
＊				＊	＊	＊		有
＊				＊	＊		＊	有
＊				＊		＊	＊	有

＊ は質的属性として選択

世帯人員区分	就業人員区分	住居の所有関係(現住居)	住居の建て方(現住居)	就業･非就業の別(世帯主)	企業規模(世帯主)	職業符号(世帯主)	性別(世帯主)	レコード数1又は2の有無
*					*	*	*	有
	*	*	*	*				有
	*	*	*		*			有
	*	*	*			*		有
	*	*	*				*	有
	*	*		*	*			有
	*	*		*		*		有
	*	*		*			*	有
	*	*			*	*		有
	*	*			*		*	有
	*	*				*	*	有
	*		*	*	*			有
	*		*	*		*		有
	*		*	*			*	有
	*		*		*	*		有
	*		*		*		*	有
	*		*			*	*	有
	*			*	*	*		有
	*			*	*		*	有
	*			*		*	*	有
	*				*	*	*	有
		*	*	*	*			有
		*	*	*		*		有
		*	*	*			*	有
		*	*		*	*		有
		*	*		*		*	有
		*	*			*	*	有
		*		*	*	*		有
		*		*	*		*	有
		*		*		*	*	有
		*			*	*	*	有
			*	*	*	*		有
			*	*	*		*	有
			*	*		*	*	有
			*		*	*	*	有
				*	*	*	*	有
*	*	*	*	*				有
*	*	*	*		*			有
*	*	*	*			*		有
*	*	*	*				*	有
*	*	*		*	*			有
*	*	*		*		*		有
*	*	*		*			*	有
*	*	*			*	*		有
*	*	*			*		*	有
*	*	*				*	*	有
*	*		*	*	*			有
*	*		*	*		*		有
*	*		*	*			*	有
*	*		*		*	*		有
*	*		*		*		*	有
*	*		*			*	*	有
*	*			*	*	*		有
*	*			*	*		*	有
*	*			*		*	*	有
*	*				*	*	*	有
*		*	*	*	*			有
*		*	*	*		*		有
*		*	*	*			*	有
*		*	*		*	*		有
*		*	*		*		*	有
*		*	*			*	*	有
*		*		*	*	*		有
*		*		*	*		*	有
*		*		*		*	*	有

＊ は質的属性として選択

世帯人員区分	就業人員区分	住居の所有関係(現住居)	住居の建て方(現住居)	就業･非就業の別(世帯主)	企業規模(世帯主)	職業符号(世帯主)	性別(世帯主)	レコード数1又は2の有無
*		*			*	*	*	有
*			*	*	*	*		有
*			*	*	*		*	有
*			*	*		*	*	有
*			*		*	*	*	有
*				*	*	*	*	有
	*	*	*	*	*			有
	*	*	*	*		*		有
	*	*	*	*			*	有
	*	*	*		*	*		有
	*	*	*		*		*	有
	*	*	*			*	*	有
	*	*		*	*	*		有
	*	*		*	*		*	有
	*	*		*		*	*	有
	*	*			*	*	*	有
	*		*	*	*	*		有
	*		*	*	*		*	有
	*		*	*		*	*	有
	*		*		*	*	*	有
	*				*	*	*	有
		*	*	*	*	*		有
		*	*	*	*		*	有
		*	*	*		*	*	有
		*	*		*	*	*	有
		*		*	*	*	*	有
			*	*	*	*	*	有
*	*	*	*	*	*			有
*	*	*	*	*		*		有
*	*	*	*	*			*	有
*	*	*	*		*	*		有
*	*	*	*		*		*	有
*	*	*	*			*	*	有
*	*	*		*	*	*		有
*	*	*		*	*		*	有
*	*	*		*		*	*	有
*	*	*			*	*	*	有
*	*		*	*	*	*		有
*	*		*	*	*		*	有
*	*		*	*		*	*	有
*	*		*		*	*	*	有
*	*			*	*	*	*	有
*			*	*	*	*	*	有
	*	*	*	*	*	*		有
	*	*	*	*	*		*	有
	*	*	*	*		*	*	有
	*	*	*		*	*	*	有
	*	*		*	*	*	*	有
	*		*	*	*	*	*	有
		*	*	*	*	*	*	有
*	*	*	*	*	*	*		有
*	*	*	*	*	*		*	有
*	*	*	*	*		*	*	有
*	*	*	*		*	*	*	有
*	*	*		*	*	*	*	有
*	*		*	*	*	*	*	有
*		*	*	*	*	*	*	有
	*	*	*	*	*	*	*	有
*	*	*	*	*	*	*	*	有

別添2
原区分及び統合区分

●原区分

(1)世帯人員区分

符号	項目名	件数	構成比
2	2人	19,643	36%
3	3人	13,696	25%
4	4人	12,860	23%
5	5人	5,967	11%
6	6人	1,919	3%
7	7人	739	1%
8	8人	185	0%
9	9人	42	0%
10	10人以上	5	0%
総計		55,056	100%

(2)就業人員区分

符号	項目名	件数	構成比
0	0人	8,566	16%
1	1人	19,951	36%
2	2人	19,319	35%
3	3人	5,257	10%
4	4人	1,688	3%
5	5人	229	0%
6	6人	41	0%
7	7人	4	0%
8	8人	1	0%
総計		55,056	100%

(3)住居の所有関係（現住居）

符号	項目名	件数	構成比
1	持ち家（世帯員名義）	43,416	79%
2	持ち家（その他名義）	1,357	2%
3	民営賃貸住宅（設備専用）	5,494	10%
4	民営賃貸住宅（設備共用）	83	0%
5	県市区町村営賃貸住宅	2,285	4%
6	都市再生機構・公社等賃貸住宅	730	1%
7	社宅・公務員住宅（借上げ含む）	1,571	3%
8	借間	120	0%
総計		55,056	100%

(4)住居の建て方（現住居）

符号	項目名	件数	構成比
1	一戸建	43,003	78%
2	長屋建	950	2%
3	共同住宅（1・2階建）	2,482	5%
4	共同住宅（3～5階建）	5,080	9%
5	共同住宅（6～10階建）	2,107	4%
6	共同住宅（11階建以上）	1,337	2%
7	その他	97	0%
総計		55,056	100%

(5)就業・非就業の別[世帯主]

符号	項目名	件数	構成比
1	就業	40,783	74%
2	うちパート	1,895	3%
3	非就業	11,721	21%
4	うち仕事を探している	657	1%
総計		55,056	100%

(6)企業規模[世帯主]

符号	項目名	件数	構成比
1	1～4人	9,812	18%
2	5～29人	7,428	13%
3	30～499人	10,501	19%
4	500～999人	2,033	4%
5	1000人以上	6,063	11%
△	（非就業、又は官公）	19,219	35%
総計		55,056	100%

●統合区分

符号	個別符号	項目名	件数	構成比
1	2	2人	19,643	36%
2	3	3人	13,696	25%
3	4	4人	12,860	23%
4	5	5人	5,967	11%
5	6－	6人以上	2,890	5%
	総計		55,056	100%

符号	個別符号	項目名	件数	構成比
1	0	無し	8,566	16%
2	1	1人	19,951	36%
3	2	2人	19,319	35%
4	3	3人	5,257	10%
5	4－	4人以上	1,963	4%
	総計		55,056	100%

符号	個別符号	項目名	件数	構成比
1	1－2	持ち家	43,416	79%
2	3－8	借家・借間	11,640	21%
	総計		55,056	100%

個別データに同じ

符号	個別符号	項目名	件数	構成比
1	1－2	就業	42,678	78%
2	3－4	非就業	12,378	22%
	総計		55,056	100%

符号	個別符号	項目名	件数	構成比
1	1－2	1～29人	17,240	31%
2	3	30～499人	10,501	19%
3	4	500～999人	2,033	4%
4	5	1000人以上	6,063	11%
△	△	（非就業、又は官公）	19,219	35%
	総計		55,056	100%

●原区分

(7)職業符号[世帯主]

符号	項目名	件数	構成比
1	常用労務作業者	11,701	21%
2	臨時及び日々雇労務作業者	312	1%
3	民間職員	13,838	25%
4	官公職員1	1,165	2%
5	官公職員2	5,028	9%
6	商人及び職人	5,469	10%
7	個人経営者	575	1%
8	農林漁業従業者	2,126	4%
9	法人経営者	1,642	3%
10	自由業者	714	1%
11	その他	108	0%
12	無職	12,378	22%
総計		55,056	100%

(8)性別[世帯主]

符号	項目名	件数	構成比
1	男	50,664	92%
2	女	4,392	8%
総計		55,056	100%

●統合区分

符号	個別符号	項目名	件数	構成比
1	1－2	労務作業者	12,013	22%
2	3	民間職員	13,838	25%
3	4－5	官公職員	6,193	11%
4	6－12	勤労者以外	23,012	42%
	総計		55,056	100%

符号変換なし

別添3
実験2 統合区分の質的属性の組合せリスト

＊は質的属性として選択

世帯人員区分	就業人員区分	住居の所有関係(現住居)	住居の建て方(現住居)	就業・非就業の別(世帯主)	企業規模(世帯主)	職業符号(世帯主)	性別(世帯主)	レコード数1又は2の有無
*								無
	*							無
		*						無
			*					無
				*				無
					*			無
						*		無
							*	無
*	*							無
*		*						無
*			*					無
*				*				無
*					*			無
*						*		無
*							*	無
	*	*						無
	*		*					無
	*			*				無
	*				*			無
	*					*		無
	*						*	無
		*	*					無
		*		*				無
		*			*			無
		*				*		無
		*					*	無
			*	*				無
			*		*			有
			*			*		有
			*				*	無
				*	*			無
				*		*		無
				*			*	無
					*	*		無
					*		*	無
						*	*	無
*	*	*						無
*	*		*					有
*	*			*				無
*	*				*			無
*	*					*		無
*	*						*	無
*		*	*					有
*		*		*				無
*		*			*			無
*		*				*		無
*		*					*	無
*			*	*				有
*			*		*			有
*			*			*		有
*			*				*	有
*				*	*			無
*				*		*		無
*				*			*	無
*					*	*		有
*					*		*	有
*						*	*	無
	*	*	*					有
	*	*		*				無
	*	*			*			無
	*	*				*		有

＊は質的属性として選択

世帯人員区分	就業人員区分	住居の所有関係(現住居)	住居の建て方(現住居)	就業・非就業の別(世帯主)	企業規模(世帯主)	職業符号(世帯主)	性別(世帯主)	レコード数1又は2の有無
	*	*					*	無
	*		*	*				有
	*		*		*			有
	*		*			*		有
	*		*				*	有
	*			*	*			無
	*			*		*		無
	*			*			*	有
	*				*	*		有
	*				*		*	無
	*					*	*	無
		*	*	*				有
		*	*		*			有
		*	*			*		有
		*					*	無
		*		*	*			無
		*		*		*		無
		*		*			*	無
		*			*	*		無
		*			*		*	無
		*				*	*	無
			*	*	*			有
			*	*		*		有
			*	*			*	有
			*		*	*		有
			*		*		*	有
			*			*	*	有
				*	*	*		無
				*	*		*	無
				*		*	*	無
					*	*	*	無
*	*	*	*					有
*	*	*		*				有
*	*	*			*			有
*	*	*				*		有
*	*	*					*	有
*	*		*	*				有
*	*		*		*			有
*	*		*			*		有
*	*		*				*	有
*	*			*	*			無
*	*			*		*		無
*	*			*				有
*	*				*	*		有
*	*				*		*	有
*	*					*	*	有
*		*	*	*				有
*		*	*		*			有
*		*	*			*		有
*		*	*				*	有
*		*		*	*			無
*		*		*		*		無
*		*		*			*	無
*		*			*	*		有
*		*			*		*	有
*		*				*	*	有
*			*	*	*			有
*			*	*		*		有
*			*	*			*	有
*			*		*	*		有
*			*		*		*	有
*			*			*	*	有
*				*	*	*		有
*				*	*		*	有
*				*		*	*	無

＊は質的属性として選択

世帯人員区分	就業人員区分	住居の所有関係(現住居)	住居の建て方(現住居)	就業・非就業の別(世帯主)	企業規模(世帯主)	職業符号(世帯主)	性別(世帯主)	レコード数1又は2の有無
＊					＊	＊	＊	有
	＊	＊	＊	＊				有
	＊	＊	＊		＊			有
	＊	＊	＊			＊		有
	＊	＊	＊				＊	有
	＊	＊		＊	＊			有
	＊	＊		＊		＊		有
	＊	＊		＊			＊	有
	＊	＊			＊	＊		有
	＊	＊			＊		＊	有
	＊	＊				＊	＊	有
	＊		＊	＊	＊			有
	＊		＊	＊		＊		有
	＊		＊	＊			＊	有
	＊		＊		＊	＊		有
	＊		＊		＊		＊	有
	＊		＊			＊	＊	有
	＊			＊	＊	＊		有
	＊			＊	＊		＊	有
	＊			＊		＊	＊	有
	＊				＊	＊	＊	有
		＊	＊	＊	＊			有
		＊	＊	＊		＊		有
		＊	＊	＊			＊	有
		＊	＊		＊	＊		有
		＊	＊		＊		＊	有
		＊	＊			＊	＊	有
		＊		＊	＊	＊		無
		＊		＊	＊		＊	無
		＊		＊		＊	＊	無
		＊			＊	＊	＊	無
			＊	＊	＊	＊		有
			＊	＊	＊		＊	有
			＊	＊		＊	＊	有
			＊			＊	＊	有
				＊	＊	＊	＊	無
＊	＊	＊	＊	＊				有
＊	＊	＊	＊		＊			有
＊	＊	＊	＊			＊		有
＊	＊	＊	＊				＊	有
＊	＊	＊		＊	＊			有
＊	＊	＊		＊		＊		有
＊	＊	＊		＊			＊	有
＊	＊	＊			＊	＊		有
＊	＊	＊			＊		＊	有
＊	＊	＊				＊	＊	有
＊	＊		＊	＊	＊			有
＊	＊		＊	＊		＊		有
＊	＊		＊	＊			＊	有
＊	＊		＊		＊	＊		有
＊	＊		＊		＊		＊	有
＊	＊		＊			＊	＊	有
＊	＊		＊	＊	＊	＊		有
＊	＊			＊	＊		＊	有
＊	＊			＊		＊	＊	有
＊	＊			＊			＊	有
＊		＊	＊	＊	＊			有
＊		＊	＊	＊		＊		有
＊		＊	＊	＊			＊	有
＊		＊	＊		＊	＊		有
＊		＊	＊		＊		＊	有
＊		＊	＊			＊	＊	有
＊		＊		＊	＊	＊		有
＊		＊		＊	＊		＊	有
＊		＊		＊		＊	＊	有

＊は質的属性として選択

世帯人員区分	就業人員区分	住居の所有関係(現住居)	住居の建て方(現住居)	就業・非就業の別(世帯主)	企業規模(世帯主)	職業符号(世帯主)	性別(世帯主)	レコード数1又は2の有無
＊		＊			＊	＊	＊	有
＊			＊	＊	＊	＊		有
＊			＊	＊	＊		＊	有
＊			＊	＊		＊	＊	有
＊			＊		＊	＊	＊	有
＊				＊	＊	＊	＊	有
	＊	＊	＊	＊	＊			有
	＊	＊	＊	＊		＊		有
	＊	＊	＊	＊			＊	有
	＊	＊	＊		＊	＊		有
	＊	＊	＊		＊		＊	有
	＊	＊	＊			＊	＊	有
	＊	＊		＊	＊	＊		有
	＊	＊		＊	＊		＊	有
	＊	＊		＊		＊	＊	有
	＊	＊			＊	＊	＊	有
	＊		＊	＊	＊	＊		有
	＊		＊	＊	＊		＊	有
	＊		＊	＊		＊	＊	有
	＊		＊		＊	＊	＊	有
	＊			＊	＊	＊	＊	有
		＊	＊	＊	＊	＊		有
		＊	＊	＊	＊		＊	有
		＊	＊	＊		＊	＊	有
		＊	＊		＊	＊	＊	有
		＊		＊	＊	＊	＊	無
			＊	＊	＊	＊	＊	有
＊	＊	＊	＊	＊	＊			有
＊	＊	＊	＊	＊		＊		有
＊	＊	＊	＊	＊			＊	有
＊	＊	＊	＊		＊	＊		有
＊	＊	＊	＊		＊		＊	有
＊	＊	＊	＊			＊	＊	有
＊	＊	＊		＊	＊	＊		有
＊	＊	＊		＊	＊		＊	有
＊	＊	＊		＊		＊	＊	有
＊	＊	＊			＊	＊	＊	有
＊	＊		＊	＊	＊	＊		有
＊	＊		＊	＊	＊		＊	有
＊	＊		＊	＊		＊	＊	有
＊	＊		＊		＊	＊	＊	有
＊	＊			＊	＊	＊	＊	有
＊		＊	＊	＊	＊	＊		有
＊		＊	＊	＊	＊		＊	有
＊		＊	＊	＊		＊	＊	有
＊		＊	＊		＊	＊	＊	有
＊		＊		＊	＊	＊	＊	有
＊			＊	＊	＊	＊	＊	有
	＊	＊	＊	＊	＊	＊		有
	＊	＊	＊	＊	＊		＊	有
	＊	＊	＊	＊		＊	＊	有
	＊	＊	＊		＊	＊	＊	有
	＊	＊		＊	＊	＊	＊	有
	＊			＊	＊	＊	＊	有
		＊	＊	＊	＊	＊	＊	有
＊	＊	＊	＊	＊	＊	＊		有
＊	＊	＊	＊	＊	＊		＊	有
＊	＊	＊	＊	＊		＊	＊	有
＊	＊	＊	＊		＊	＊	＊	有
＊	＊	＊		＊	＊	＊	＊	有
＊	＊		＊	＊	＊	＊	＊	有
＊		＊	＊	＊	＊	＊	＊	有
	＊	＊	＊	＊	＊	＊	＊	有
＊	＊	＊	＊	＊	＊	＊	＊	有

製 表 技 術 参 考 資 料 10
平成 20 年 9 月発行

編集・発行　　独立行政法人　統計センター

〒162－8668
東京都新宿区若松町 19－1
電　話　代表　03 (5273) 1200

情報技術部研究主幹　　TEL : 03－5273－1286
E-mail : research@nstac.go.jp