匿名化技法としてのミクロアグリゲーションについて

伊 藤 伸 介

要　　旨

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 (microaggregation) の方法的な特徴を考察し，わが国の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有効性を検証した。最初に，本稿は，個別データに含まれる属性群を量的属性と質的属性に類別し，質的属性においては「超高次元クロス集計表」をもとに，対象となるすべての質的属性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」) の編成を行い，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，ミクロアグリゲーション済のデータ (「ミクロアグリゲートデータ」) が作成できることを提案した。つぎに，本稿では，『平成 16 年全国消費実態調査』の個別データを用いて，質的属性の組合せの検討，個別ランキング法等を用いた量的属性のミクロアグリゲーションによるミクロアグリゲートデータの作成，およびミクロアグリゲートデータと個別データにおける近似性の検証を試みた。本研究では，同質属性値レコード群内にレコード数 1 または 2 が存在しない組合せが，全 255 パターン中 18 パターン (7.1％) となった。また，個別ランキング法を用いた場合，個別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。

1　はじめに

わが国では，統計法の改正に伴い，政府統計のミクロデータの提供に対する関心が一層高まっている。これまでの旧統計法においては，政府統計ミクロデータの提供は，「統計目的外使用」という限定された形で行われてきたことから[1)]，ミクロデータの利用は一部の研究者に限られ

1)　旧統計法 (統計法 (昭和 22 年法律第 18 号)) の第 15 条では，「統計上の目的」がつぎのように明記されている。

「第 15 条　何人も，指定統計を作成するために集められた調査票を，統計上の目的以外に使用して

てきた。新統計法では，匿名データとしての作成・提供に関する条項が明記されており[2)]，これまで以上に，政府統計ミクロデータの利用の促進がはかられることから，わが国においてもミクロレベルの実証的な社会経済研究が，大きく進展することが期待できる。一方，政府統計ミクロデータの提供においては，ミクロデータの有用性を踏まえながらも，個体情報の保護を指向した形で個別データの秘匿性を十分に確保する必要がある。そのため，個別データに対する秘匿処理の方法を具体的に検討することが求められる。

ミクロデータを提供している欧米諸国では，個体情報の保護に関する法的制度的措置がとられていることが知られている[3)]。例えば，アメリカでは，1976 年に成立した現行の合衆国法典 (the U.S. Code) の第 13 編第 9 条に基づき，特定の事業所や個人に関する個体情報を識別することが可能なデータの提供が禁じられている (石田 (2000, 30 頁)，森 (2005，4〜5 頁)，Zayatz (2007, p.253))。また，2002 年には，秘密情報保護・統計効率化法 (Confidential Information Protection and Statistical Efficiency Act of 2002) が制定され，統計目的のために個人や企業から収集された秘密情報の保護が明記されている (森 (2005, 3〜4 頁)，Zayatz (2007, p. 253))。つぎに，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 (Confidentiality and Data Access Committee) において，ミクロデータの提供によって個体情報が露見される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可能性についてのチェックリスト

はならない。

2 前項の規定は，総務大臣の承認を得て使用の目的を公示したものについては，これを適用しない。」

ここで，「統計上の目的」とは，「第 7 条第 1 項 (総務大臣による「指定統計調査の承認および実施」) で承認を受けた調査により当該指定統計を作成するという目的」(坂本 (1991, 52 頁)) であるから，第 15 条の第 1 項は，「統計調査の企画の際に計画した集計表」を作成する以外には，調査票を使用することはできないことを意味している。そのために，第 15 条第 2 項において，「統計法第 14 条で規定されている秘密の保護」が担保され，「調査票の使用が公益性を有する」という条件のもとで，総務大臣が承認した場合に限り，統計目的外使用が認められているが，「公益性を有する」研究とは，「原則として政府からの委託研究であることや，少なくとも科学研究費補助金を受けているなど公に公益性があると認められている研究」を示唆している (松井 (2005, 13 頁))。よって，ミクロデータの提供は，「指定統計調査を実施する府省が研究等の目的のために，学者，研究者などに依頼する」場合のみに事実上限定されていた (井出 (2004, 39 頁))。

2) 新統計法 (統計法 (平成 19 年法律第 53 号)) では，政府統計の二次利用に関する規定として，オーダーメイド集計に関する条項 (第 34 条) および匿名データの作成・提供に関する条項 (第 35 条，第 36 条) が条文化されている。新統計法は，匿名データを「一般の利用に供することを目的として調査票情報を特定の個人又は法人その他の団体の識別ができないように加工したもの」(新統計法第 2 条第 11 項) と規定しており，「調査票情報」についても「統計調査によって集められた情報のうち，文書，図画又は電磁的記録によって記録されているもの」と明記している。なお，新統計法の特徴については，例えば森 (2007b) を参照されたい。

3) 欧米諸国におけるミクロデータの提供状況および個体情報の保護に関する法的制度的措置の詳細については，松田・濱砂・森編 (2000)，森 (2005) 等を参照されたい。

(Checklist on Disclosure Potential of Proposed Data Releases)」が作成されている[4]。さらに，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，主としてチェックリストを用いて，センサス局で作成される政府統計ミクロデータの提供に関する審査を行っている (石田 (2000, 35〜36 頁) , Zayatz (2007, p. 255))。

他方，諸外国の統計作成部局は，政府統計の個別データに対して様々な匿名化技法を用いている。Federal Committee on Statistical Methodology (2005, p. 24) は，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 (Public Use Microdata File) を提供するために採用する基本的な匿名化技法として，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という 4 つの方法を指摘している。さらに，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，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ノイズの導入，④データ・スワッピングあるいはランク・スワッピング (スイッチング (switching) とも呼ぶ)，⑤変数値の削除 (blank) と補定 (impute)，⑥ブラーリング (blurring) といった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))[5]。

ところで，近年，ヨーロッパ諸国を中心に，政府統計ミクロデータに対する匿名化技法とし

---

4) チェックリストが作成された契機としては，ミクロデータの秘匿に関して，①ミクロデータにおける個体情報の開示リスク (disclosure risk) についての尺度が明確ではないこと，②ミクロデータに適用される秘匿処理の妥当性についての基準が存在しないことが指摘されている (Federal Committee on Statistical Methodology (2005, p. 24))。なお，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関する定性的な基準として採用されている ((Federal Committee on Statistical Methodology (2005, p. 24))。

5) 近年，国際連合欧州経済委員会 (United Nations Economic Commission for Europe) 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 (Felso *et al.* (2001, pp. 20−24))。その調査結果によれば，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明らかにされている。また，Felso *et al.* (2001) では，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関する秘匿措置の現状が調査されているが，調査結果から，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っている (Felso *et al.* (2001, pp.31-34))。さらに，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作成機関が存在していることが調査結果から明らかになっている。

て，「ミクロアグリゲーション (microaggregation)」に関する研究が進められている (Willenborg and de Waal (2001, pp. 30−31))。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，ブラーリング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))[6]。イタリア統計局は，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，ドイツ連邦統計局でも，企業のパネルデータに対する匿名化技法の 1 つとして，ミクロアグリゲーションに関する研究が行われている (Brandt *et al.* (2008))。

その一方で，わが国ではミクロアグリゲーションについての実証的な研究がこれまで行われていなかったことから，諸外国における先行研究を踏まえて，わ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。

本稿では，つぎの 2 つの研究課題を扱うことにする。第 1 に，ミクロアグリゲーションにおける研究動向を概観することによって，その方法的な特徴を洞察する。第 2 に，わ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追究するために，政府統計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。

---

6) 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。

2　ミクロアグリゲーションの方法的特徴

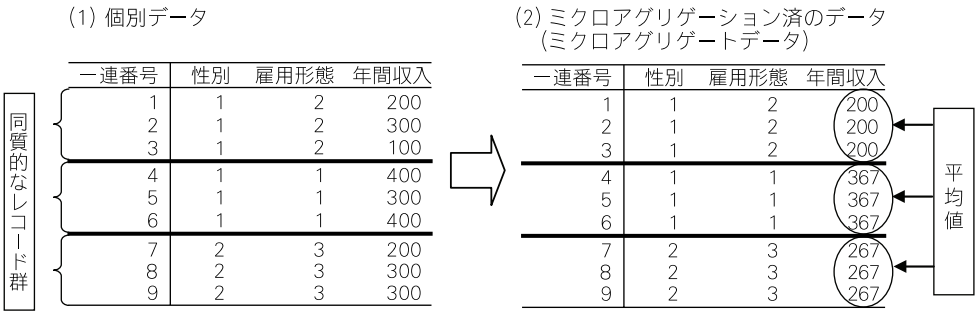
一般に，統計調査の個別データは，複数の調査項目 (属性) と調査項目の回答値 (属性値) から成り立っている。ミクロアグリゲーションとは，ミクロデータ (個別データ) を k 個 (k は閾値 (threshold)) のレコードを有する同質的なレコード群にグループ化した上で，そのレコードにおける個々の属性値を平均値等の代表値に置き換えることである (Domingo-Ferrer and Mateo-Sanz (2002, p. 190))。例えば，属性群として性別，雇用形態と年間収入のみを持つ個別データを想定し，閾値を 3 に設定したとする (図 1)。このデータ上にある属性群にミクロアグリ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，つぎに，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)」と呼称) 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，主としてつぎの 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，およ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ことから，先行研究では，属性値の特性に応じてミクロアグリゲーションの手法が個別に追究されている。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 (Defays (1997)) と，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 (Domingo-Ferrer and Mateo-Sanz (2002, p. 192)) とでは，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，主として属性の性質に関する区分をもとにして，ミクロアグリゲーションの基本的特徴を明らかにする。

⑴　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の設定方法と，量的属性値に対する処理の仕方に着目することによって，主として単一軸法 (single axis method)，第 1 主成分法 (first principal component method)，Z スコア総計法

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

(sum of Z-scores method)，個別ランキング法 (individual ranking method)，階層区分法 (hierarchical clustering method) に類別することが可能である (Anwar (1993)，Domingo-Ferrer and Mateo-Sanz (2002))。以下で，ミクロアグリゲーションの各手法の概要について述べる。

① 単一軸法

単一軸法では，ソートキーとなる特定の量的属性に着目し，その属性値を昇順または降順にソートし，ソートされたレコード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。つぎに，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

7)　わ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

③ Z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

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，およびレコードのグループ化が行われている。

③ Z 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，および 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，つぎ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，およ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))。

⑤ 階層区分法

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がグループ化されているとは言いがたい。

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

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

雇用者数についてソートし3レコードずつ平均化

適用後

一連番号	雇用者数	総売上高	店舗の数
1	24	1000	2
2	24	1500	6
3	24	2000	5
4	43	3000	3
5	43	1000	4
6	43	2000	10
7	57	1500	11
8	57	1500	10
9	57	3000	14

昇順

②総売上高

総売上高によって一連番号の順番が変わっている

一連番号	雇用者数	総売上高	店舗の数
1	24	1000	2
5	43	1000	4
2	24	1500	6
7	57	1500	11
8	57	1500	10
3	24	2000	5
6	43	2000	10
4	43	3000	3
9	57	3000	14

総売上高についてソートし3レコードずつ平均化

適用後

一連番号	雇用者数	総売上高	店舗の数
1	24	1167	2
5	43	1167	4
2	24	1167	6
7	57	1667	11
8	57	1667	10
3	24	1667	5
6	43	2667	10
4	43	2667	3
9	57	2667	14

③店舗の数

一連番号	雇用者数	総売上高	店舗の数
1	24	1167	2
4	43	2667	3
5	43	1167	4
3	24	1667	5
2	24	1167	6
8	57	1667	10
6	43	2667	10
7	57	1667	11
9	57	2667	14

店舗の数についてソートし3レコードずつ平均化

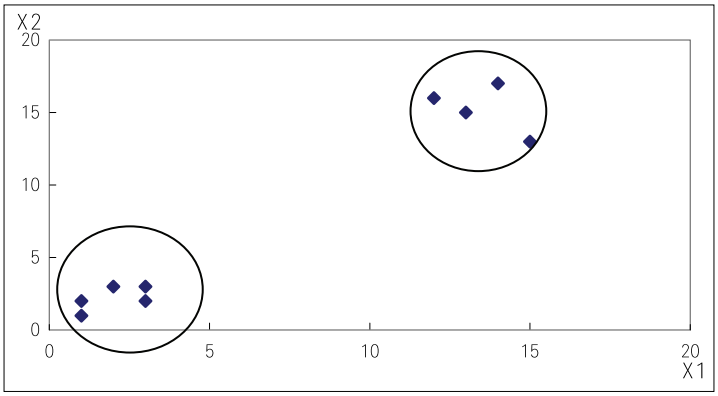
適用後

一連番号	雇用者数	総売上高	店舗の数
1	24	1167	2
4	43	2667	3
5	43	1167	4
3	24	1667	5
2	24	1167	6
8	57	1667	10
6	43	2667	10
7	57	1667	11
9	57	2667	14

ミクロアグリゲーションの結果

一連番号	雇用者数	総売上高	店舗の数
1	24	1167	3
2	24	1167	7
3	24	1667	7
4	43	2667	3
5	43	1167	3
6	43	2667	12
7	57	1667	12
8	57	1667	7
9	57	2667	12

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

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 $\mathbf{X}'=\left(X_1, \cdots, X_p\right)$ (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 ($n_i \geq k$ および $n=\sum_{i=1}^{g} n_i$)，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。

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

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。つぎに，グループ間平方和 (between-groups sum of squares＝SSA) は，

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

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，

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

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。

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

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。

つぎに，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。

3) 2k 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。

⑵ 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。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，メディアンやモードといった代表値に置き換えられている (Torra (2004, p. 165), Defays and Anwar (1998, pp. 456−457))。また，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，例えばつぎ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。

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

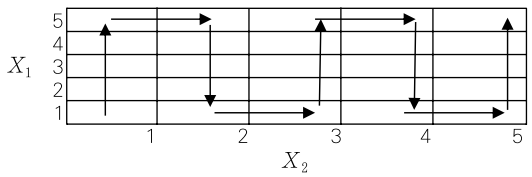
p_i …ある属性指標における i 番目の分類項目における頻度 (出現確率)

ld…底を2とする対数

L…属性群における分類項目の数

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。

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

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))

⑶ 匿名化技法としてのミクロアグリゲーションの展開可能性

Domingo-Ferrer and Torra (2001a) 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 (2001a, 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，個別データに対してミクロアグリゲーションを適用する上で，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，個別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，個別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なか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)，および週間就業時間 (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，およ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。

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 / 週間就業時間	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	
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

ところで，属性群として性別，雇用形態，およ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，同質属性値レコード群のなかで，性別 1，雇用形態 1，週間就業時間 4 という属性値を有するレコードがそれに該当するだけでなく，分類項目の組合せの総数 2 および年間収入の総計 360 万円という集計値がレコードに付与されている。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，およ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制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しているが，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[9)]。なぜなら，超高次元クロス集計に基

9) わが国では，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，および「非掲載」

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 独立行政法人統計センター) 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[10]。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，つぎ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) において同一の属性値を有するレコード群が閾値に基づいてグループ分けされた上で，グループ内のレコードに含まれ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設定し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の閾値を定めることは，

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。

10) 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

すべての属性に関するクロス集計表

(世帯主の)性別	就業・非就業の別	企業規模	(世帯主の)職業符号	…	総数 (N)
1	1	1	1		1
1	1	1	1	…	1
1	1	1	2		2
⋮		⋮	⋮		⋮
⋮		⋮	⋮		⋮
⋮		⋮	⋮		⋮
1	4	5	12		2
1	4	5	12	…	1
⋮		⋮	⋮		⋮
⋮		⋮	⋮		⋮
⋮		⋮	⋮		⋮
2	1	1	1		1
2	1	1	1		1
2	1	1	2	…	2
2	1	1	3		3
⋮		⋮	⋮		⋮
⋮		⋮	⋮		⋮
⋮		⋮	⋮		⋮

性別を選択

1つの属性を選択

就業・非就業の別を選択

(世帯主の)性別	就業・非就業の別	企業規模	(世帯主の)職業符号	…	総数 (N)
1	—	—	—		50664
2	—	—	—	…	4392
—	1	—	—		40783
—	2	—	—		1895
—	3	—	—		11721
—	4	—	—		657
⋮	⋮	⋮	⋮	…	⋮
⋮	⋮	⋮	⋮		⋮

2つの属性を選択

性別と就業・非就業の別を選択

(世帯主の)性別	就業・非就業の別	企業規模	(世帯主の)職業符号	…	総数 (N)
1	1	—	—		38578
1	2	—	—	…	908
1	3	—	—		10644
1	4	—	—		534
2	1	—	—		2205
2	2	—	—		987
2	3	—	—		1077
2	4	—	—		123
⋮	⋮	⋮	⋮	…	⋮
⋮	⋮	⋮	⋮		⋮

すべての属性を選択

(世帯主の)性別	就業・非就業の別	企業規模	(世帯主の)職業符号	…	総数 (N)
1	1	1	1		1
1	1	1	1	…	1
1	1	1	2		2
⋮	⋮	⋮	⋮		⋮
⋮	⋮	⋮	⋮		⋮
⋮	⋮	⋮	⋮		⋮
1	4	5	12		2
1	4	5	12	…	1
⋮	⋮	⋮	⋮		⋮
⋮	⋮	⋮	⋮		⋮
⋮	⋮	⋮	⋮		⋮
2	1	1	1		1
2	1	1	1		1
2	1	1	2	…	2
2	1	1	3		3
⋮	⋮	⋮	⋮		⋮
⋮	⋮	⋮	⋮		⋮
⋮	⋮	⋮	⋮		⋮

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決定することを意味している。閾値をｋとすると，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かｋ以上の数値になるようにクロス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。

図９は，個別データに量的属性と質的属性が含まれる場合の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。属性群として性別，雇用形態，週間就業時間，および年間収入を有する個別データが想定されている。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なる属性群から，属性の組合せを選び出すことによって，度数１または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。最初に，図９においては性別，雇用形態と週間就業時間の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または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９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

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)

図9　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

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における年間収入の属性値が置き換えられることによって，ミクロアグリゲートデータが編成される[11)]。

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処理においては，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，および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 (1997, 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なかの度数が 1 または 2 である場合，そのセルは，個体情報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 または 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，つぎのように考えることができる。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 または 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[12)]。なお，

---

11) 図 9 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 9 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 または 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。

12) 本研究では，ミクロアグリゲーションにおける秘匿性の定量的な評価方法については考察していない。これについては別稿の課題にしたい。秘匿性の定量的な評価を行うために開示リスクの評価方法

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。

⑵ 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」と呼称) 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って評価される。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，ミクロデータの有用性の基準に関しては，秘匿処理済データが「分析上有効であること (analytically valid)」，および「分析上興味深いこと (analytically interesting)」が考えられている (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，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) においては，属性数が 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 (2001a, p. 104))。

① 共分散行列
② 相関係数行列
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，つぎのような尺度を用いて評価が行われる。

を追究した研究は数多く存在するが，様々な匿名化技法が適用されたミクロデータに対して開示リスクを定量的に評価した研究については，例えば，Domingo-Ferrer and Torra (2001b) を参照。さらに，わ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。

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 『全国消費実態調査』によるミクロアグリゲーションの有効性の検証

前節までは，先行研究に基づきミクロアグリゲーションの研究動向を洞察することによって，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。本節では，ミクロアグリゲーションの手法を政府統計の個別データに適用することによって，ミクロアグリゲーションの方法的な有効性を探る。本研究では，個別データに含まれる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適用可能性を追究している。そのために，本研究は，つぎの 2 つの研究から成っている。第 1 の研究では，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 (以下「研究 1」)。第 2 の研究では，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，『全消』のミクロアグリゲートデータの作成を試み，ミクロアグリゲーションの有効性の検証を行う (以下「研究 2」)。

本研究では，『平成 16 年全国消費実態調査 (以下，『全消』と略称)』の原データを用いて，ミクロアグリゲーションの有効性を検討する。本研究で使用する『全消』の原データは，二人以上の世帯に関する約 55,000 レコードを有しているが[13]，消費支出などの約 300 の量的属性群が含まれることから，それは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，目的外使用申請およ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。つぎに，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』の原デー

13) 本研究では，『全消』の原データの中で単身世帯のレコード (標本数は約 5000) を分析の対象から除いている。

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，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，研究の対象となるすべての質的属性群について，『全消』 の原データにおける属性の分類区分にしたがって超高次元クロス集計表を作成した。つぎに，この超高次元クロス集計表に基づいて，クロス集計表のなかのセルに度数 1 または 2 を含まない質的属性の組合せの探索を行い，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 または 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 10 は，性別および就業・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示したもので，つぎの 2 つの手順からなっている。

1) 性別と就業・非就業の別に関するクロス集計を行う。図 10 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。

2) 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 または 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 または 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。

研究 1 の結果から，『全消』 の原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 または 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％) であることがわかる。また，質的属性の組合せの数は，最大で 3 になることが明らかになった。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，および 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。

⑵ 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』 の原データを用いて，量的属性のミクロアグリゲーションを行う。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なかから，質的属性群として性別 2 区分，就業・非就業 4 区分，および企業規模 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」と呼称) について，同質属性値レコード群のなかで 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

図 10　質的属性の組合せリスト作成の概略図

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	2	1
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

支出，③貯蓄現在高，④負債現在高，および，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。

つぎに，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る方法である (以下，「ソートなし」と呼称)。図 11 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 11 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い，つぎに，年間収入と消費支出について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が作成されている[14)]。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ている。

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あり，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を作成する方法である (以下，「個別ランキング法」と呼称)。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の手順は，つぎのとおりである (図 12)。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，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，グループ内のレコードに含まれる年間収入を平均値に置き換えた。つぎに，消費支出をキーとして昇順で並べ替え，レコードをグループ分けし，グループ内のレコードが有する消費支出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[15)]。

---

14) 本研究では，ミクロアグリゲーションの手法の相違が原データに対する情報量損失に及ぼす影響を把握することに焦点を当てていることから，本稿では，ミクロアグリゲーションにおけるレコードのソート化および属性値の平均値への置き換えにおいて，母集団復元乗率が適用されていないことに留意されたい。なお，伊藤・磯部・秋山 (2008) では，『全消』の個別データによるミクロアグリゲーションにおいて母集団復元乗率を適用した場合の研究成果が示されている。

15) 『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。

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することによって，量的属性のミクロアグリゲーションの有効性が検証される。

最初に，表1は，『全消』の原データとソートなしあるいは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

図11 『全消』における量的属性のミクロアグリゲーション―ソートなし

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		性別	就業・非就業の別	年間収入	消費支出
08	109	13	1	1	2	2000	1000	㋐	1	1	2000	1000
21	101	11	1	1	1	2000	1000		1	1	2000	1000
44	104	16	1	1	2	2000	1000		1	1	2000	1000
04	106	17	1	2	3	2000	1000	㋑	1	2	4000	3000
15	104	11	1	2	4	3000	2000		1	2	4000	3000
18	105	17	1	2	4	4000	3000		1	2	4000	3000
30	106	13	1	2	5	5000	4000		1	2	4000	3000
34	105	19	1	2	2	6000	5000		1	2	4000	3000
20	105	15	1	3	4	1429	1184	㋒	1	3	4377	1946
22	108	14	1	3	5	5144	3643		1	3	4377	1946
26	107	15	1	3	4	6559	1010		1	3	4377	1946
28	106	19	1	3	4	7631	1824		1	3	8229	3085
41	109	18	1	3	1	8004	7437		1	3	8229	3085
43	101	18	1	3	2	9052	7542		1	3	8229	3085
：	：	：	：	：	：	：	：		：	：	：	：

図12 『全消』における量的属性のミクロアグリゲーション―個別ランキング法

年間収入のソート

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	
08	109	13	1	1	2	2000	1000	㋐
21	101	11	1	1	1	2000	1000	
44	104	16	1	1	2	2000	1000	
04	106	17	1	2	3	2000	1000	㋑
15	104	11	1	2	4	3000	2000	
18	105	17	1	2	4	4000	3000	
30	106	13	1	2	5	5000	4000	
34	105	19	1	2	2	6000	5000	
20	105	15	1	3	4	1429	1184	㋒
22	108	14	1	3	5	5144	3643	
26	107	15	1	3	4	6559	1010	
28	106	19	1	3	4	7631	1824	
41	109	18	1	3	1	8004	7437	
43	101	18	1	3	2	9052	7542	
：	：	：	：	：	：	：	：	

図 12 のつづき

年間収入のミクロアグリゲーション

	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
㋐	08	109	13	1	1	2000	1000
	21	101	11	1	1	2000	1000
	44	104	16	1	1	2000	1000
㋑	04	106	17	1	2	4000	1000
	15	104	11	1	2	4000	2000
	18	105	17	1	2	4000	3000
	30	106	13	1	2	4000	4000
	34	105	19	1	2	4000	5000
㋒	20	105	15	1	3	4377	1184
	22	108	14	1	3	4377	3643
	26	107	15	1	3	4377	1010
	28	106	19	1	3	8229	1824
	41	109	18	1	3	8229	7437
	43	101	18	1	3	8229	7542
	：	：	：	：	：	：	：

消費支出のソート

	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
㋐	08	109	13	1	1	2000	1000
	21	101	11	1	1	2000	1000
	44	104	16	1	1	2000	1000
㋑	04	106	17	1	2	4000	1000
	15	104	11	1	2	4000	2000
	18	105	17	1	2	4000	3000
	30	106	13	1	2	4000	4000
	34	105	19	1	2	4000	5000
㋒	26	107	15	1	3	4377	1010
	20	105	15	1	3	4377	1184
	28	106	19	1	3	8229	1824
	22	108	14	1	3	4377	3643
	41	109	18	1	3	8229	7437
	43	101	18	1	3	8229	7542
	：	：	：	：	：	：	：

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
㋐	1	1	2000	1000
	1	1	2000	1000
	1	1	2000	1000
㋑	1	2	4000	3000
	1	2	4000	3000
	1	2	4000	3000
	1	2	4000	3000
	1	2	4000	3000
㋒	1	3	4377	1339
	1	3	4377	1339
	1	3	8229	1339
	1	3	4377	6207
	1	3	8229	6207
	1	3	8229	6207
	：	：	：	：

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『全消』の原データの値に等しくなっている。

また，表 2 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。

つぎに，図 13 および図 14 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および年間収入 10 区分階級別のヒストグラムである。図 13 と図 14 から，ソートなしにおける分布の形状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『全消』の原データのそれと非常に似ていることがわかる。さらに『全消』の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 3) 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 0.01068515，個別ランキング法が 0.00000210 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。

表 1　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の平均値

	年間収入 (万円)	消費支出 (万円)	貯蓄現在高 (万円)	負債現在高 (万円)	年齢 (歳)
○原データ	682.18	31.63	1422.97	511.76	53.73
①ソートなし	682.18	31.63	1422.97	511.76	53.73
②個別ランキング法	682.18	31.63	1422.97	511.76	53.73

表 2　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の標準偏差

	年間収入 (万円)	消費支出 (万円)	貯蓄現在高 (万円)	負債現在高 (万円)	年齢 (歳)
○原データ	446.57	20.20	1950.14	1179.8	13.82
①ソートなし	302.67	12.65	1209.11	738.33	11.14
②個別ランキング法	445.50	20.11	1944.42	1175.47	13.82

図 13　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

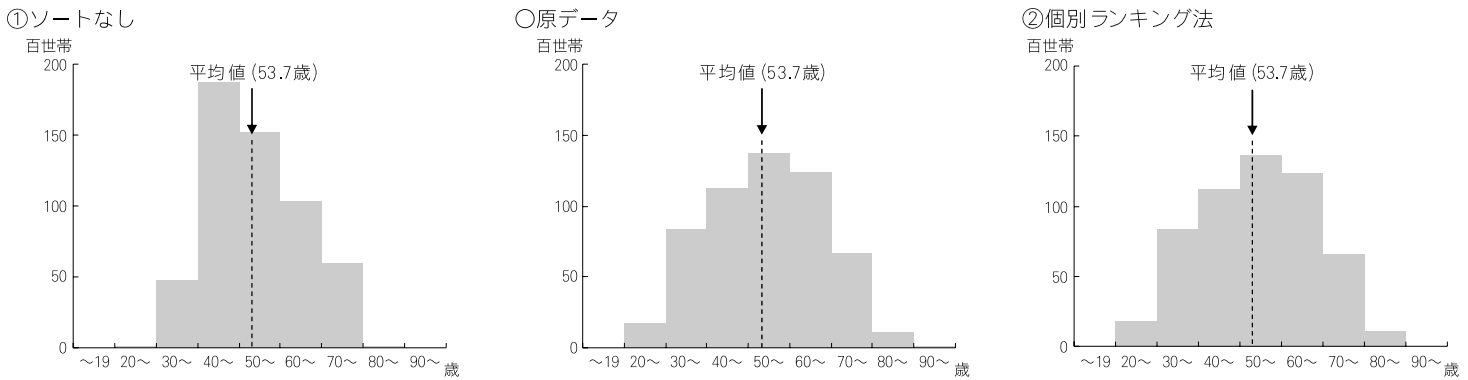


図 14　原データ，ソートなし，個別ランキング法の年間収入 10 区分階級別世帯数分布

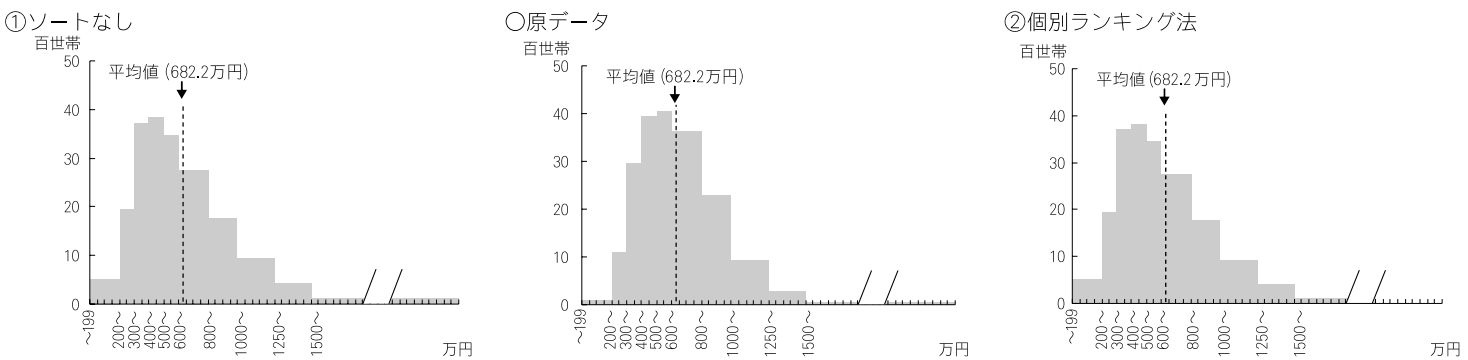


表 3　原データ，ソートなし，個別ランキング法における量的属性間の相関係数行列

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	1.00000				
消費支出	0.41474	1.00000			
貯蓄現在高	0.32829	0.23424	1.00000		
負債現在高	0.28676	0.06972	−0.03884	1.00000	
年　齢	−0.04617	−0.02792	0.27549	−0.18640	1.00000

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	1.00000				
消費支出	0.50139	1.00000			
貯蓄現在高	0.25780	0.22368	1.00000		
負債現在高	0.37106	0.15001	−0.07261	1.00000	
年　齢	−0.26453	−0.15286	0.33671	−0.29834	1.00000

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	1.00000				
消費支出	0.41793	1.00000			
貯蓄現在高	0.32606	0.23339	1.00000		
負債現在高	0.28531	0.07067	−0.03911	1.00000	
年　齢	−0.04716	−0.02826	0.27608	−0.18719	1.00000

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その基本的な特徴を考察するだけでなく，わが国の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有効性を検証した。本稿では，最初にミクロアグリゲーションのなか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方法的に位置付けることによって，ミクロアグリゲーションの方法論理の析出を試みた。ミクロアグリゲーションの方法的特徴は，つぎのように要約される。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同質属性値レコード群として捉えられる。このような同質属性値レコード群の編成に基づいて，個別に準じたレベルのデータを作成することが可能になる。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。ゆえに，超高次元クロス集計表をもとに，集計表のセルの閾値を k に設定した上でさらに集計を行った

場合，そこから同質属性値レコード群内に 0 かあるいは少なくとも k 個のレコード数を含むミクロアグリゲートデータを作成することができる。こうした論点を踏まえて，本稿では，個別データに含まれる属性群を質的属性と量的属性に類別した上で，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等の代表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できることを提案した。

つぎに，本稿は，ミクロアグリゲーションの方法の有効性を実証的に明らかにするために，『全消』の原データを用いて，ミクロアグリゲートデータの作成およびミクロアグリゲートデータの『全消』の原データに対する近似性の検証を行った。本研究では，第 1 に，同質属性値レコード群の編成を行い，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 または 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。そして，第 3 に，作成されたミクロアグリゲートデータと『全消』の原データにおける近似の程度を把握するために，ミクロアグリゲートデータの原データに対する情報量損失を計測し，個別ランキング法がソートなしと比較してより近似的なミクロアグリゲートデータであることを明らかにした。

わが国では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有効性を検証した研究がこれまで存在しなかったことから，本研究におけるミクロアグリゲーションの方法については，試論的な側面があることは否めない。しかしながら，わが国において政府統計ミクロデータの提供に関する議論が本格的に進められつつある状況において，個別データを用いて，匿名化技法の 1 つであ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追究したことの意義は小さくないと考えられる。その一方で，ミクロアグリゲートデータの有用性の観点からミクロアグリゲートデータと個別データの近似性を検証したことは，匿名化技法としてミクロアグリゲーションを適用した場合の個別データに対するバイアスを計測する試みだと捉えることもできる。このような秘匿処理によって生じるバイアスの取り扱いは，ミクロデータを用いて実証的なミクロ分析を行う上で，重要な論点となり得る。

政府統計のミクロデータの提供によって，ミクロデータに対する匿名化技法の適用可能性に関する議論が今後展開されることが考えられる。その場合，ミクロアグリゲーションだけでなく，トップ・コーディングやリコーディング等の様々な匿名化技法を対象に，匿名化技法にお

ける有効性や秘匿性を定量的に評価するための指標を具体的に検討することが必要である。本研究では，情報量損失に関する指標を用いてミクロアグリゲーションの有効性の検証を行ったが，秘匿性の観点に立ってミクロアグリゲーションにおける秘匿性の評価方法を追究することも検討すべき課題だと思われる。ミクロデータの有用性と秘匿性の両面において評価を行うことを可能にするフレームワークの構築が求められる。

謝 辞

本稿は，2008 年度統計関連学会連合大会 (2008 年 9 月 7 日～10 日，於 慶應義塾大学) において，筆者が独立行政法人統計センターで研究した成果の一部として発表した研究報告，および磯部祥子氏 (統計センター)，秋山裕美氏 (統計センター) との共同報告に基づいている。本稿の作成にあたっては，松井博氏 (元統計センター研究センター長，現統計センター顧問)，土井満喜氏 (元統計センター研究センター長，現統計センター情報技術部長)，山内晶仁氏 (統計センター情報技術部次長)，小林良行氏 (統計センター研究主幹) より様々なご助言を賜った。また，平成 19 年度データエディティング研究会 (第 2 回) (2008 年 3 月 4 日，於 独立行政法人統計センター) において本稿の内容に関する報告を行った際に，西郷浩教授 (早稲田大学)，稲葉由之教授 (当時総務省統計研修所，現慶應義塾大学)，樋田勉准教授 (群馬大学) より有益なコメントをいただいた。さらに，『全国消費実態調査』の個別データを用いたミクロアグリゲーションの有効性の検証については，磯部祥子氏と秋山裕美氏にミクロアグリゲートデータの作成およびミクロアグリゲートデータと個別データにおける近似性の検証をしていただいた。ここに記して感謝の意を表したい。なお，本稿における誤りは，すべて筆者に帰するものである。

参 考 文 献

Anwar M. N. (1993) “Microaggregation: 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.

Brandt, M., Lenz, R., Rosemann, M. (2008) “Anonymisation of Panel Enterprise Microdata－Survey of a German Project”, Domingo-Ferrer, J. and Yücel, S. (eds.) *Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference, PSD 2008, Istanbul, Turkey, September 24–26, 2008 Proceedings*, Springer, Berlin, pp. 139–151.

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. (2001a) “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 Torra, V. (2001b) “A Quantitative Comparison of Disclosure Control Methods for Microdata”, Doyle *et al.* (eds.) (2001) *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, Elsevier Science, Amsterdam, pp. 111−133.

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.

Domingo-Ferrer, J., Sebe, 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.

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

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.

井出　満 (2004)「日本におけるミクロデータ提供の現状」『ミクロデータとその利用』法政大学日本統計研究所『研究所報』No. 32，39 頁〜42 頁

石田　晃 (2000)「アメリカ」松田芳郎・濱砂敬郎・森博美編『講座ミクロ統計分析① 統計調査制度とミクロ統計の開示』日本評論社，24〜47 頁

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

伊藤伸介・磯部祥子・秋山裕美 (2008)「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 ―」,『製表技術参考資料』 No. 10, 33～66 頁

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., Sebe, 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

松田芳郎・濱砂敬郎・森　博美編 (2000)『講座ミクロ統計分析① 統計調査制度とミクロ統計の開示』 日本評論社

松井　博 (2005)「政府統計ミクロデータの利用」『ESTRELA』 2005 年 8 月号, 10～17 頁

森　博美 (2005)「諸外国におけるミクロデータ関連法規の整備状況とデータ提供の現状」法政大学日本統計研究所『オケージョナル・ペーパー』 No. 13

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

森　博美 (2007b)「新統計法の成立とわが国政府統計の課題」『計画行政』 第 30 巻第 4 号, 3～10 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

坂本佶三 (1991)『我が国の統計制度』 全国統計協会連合会

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)『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, June 9−11, 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.

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.

伊 藤 伸 介

Summary

On Microaggregation as Disclosure Avoidance Methods

Microaggregation has gained attention as a disclosure avoidance method especially in European countries in recent years. The aim of this paper is to examine its characteristics as a methodology, and judge the effectiveness of microaggregation for individual data from Japanese official statistics. First, this paper suggests how to create micro-aggregated data that closely resemble individual data based on multi-dimensional tabulation. The method of microaggregation this paper proposes is to compose records which have common values for all kinds of qualitative attributes based on multi-dimensional tabulation, to replace each value of quantitative attributes of records which have common values for the other qualitative attributes by a measure of central tendency (ex. average value etc.) within those records. Second, this paper attempts to create micro-aggregated data based on individual data from the 'National Survey of Family Income and Expenditure' (original data), and verify the degree of similarity between micro-aggregated data and original data. The findings of this empirical research are that the percentage of combinations which allow to create micro-aggregated data to ensure confidentiality of all kinds of combinations is 7.1 percent, and micro-aggregated data created by the individual ranking method is much closer to the original data than that by micro-aggregated data created microaggregation without sorting.