製表技術参考資料 11

ISSN 1348－7345

秘匿性の評価方法に関する実証研究

－全国消費実態調査のミクロアグリゲートデータを用いて－

NSTAC

Working Paper No.11

平成 21 年 6 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

秘匿性の評価方法に関する実証研究
―全国消費実態調査のミクロアグリゲートデータを用いて―

伊藤　伸介 *，磯部　祥子 **，秋山　裕美 ***

要　旨

我が国では、統計法(平成 19 年法律第 53 号)の全面施行によって、政府統計のミクロデータ(匿名データ)の提供が本格的に進められている。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だけでなく個体情報の秘匿性について考慮することが求められる。諸外国では、各種の匿名化技法によって作成されるミクロデータの秘匿性に関して定量的な評価を行った研究が数多く存在する。一方、我が国においてミクロデータの提供を今後展開していく上では、調査単位の違いや匿名化技法の相違による秘匿性の比較という観点から、ミクロデータにおける秘匿性の評価方法を具体的に検討することが考えられる。本稿の目的は、秘匿性の定量的な評価方法に焦点を当て、ミクロアグリゲーションの手法によって作成したデータ(ミクロアグリゲートデータ)を用いて、様々なミクロアグリゲートデータに対する秘匿性の評価方法の適用可能性を追究することである。

ミクロデータにおける秘匿性に関する評価については、主として次の 2 つの方法が考えられる。第 1 の方法は、秘匿処理を施していない個別データ(以下「原データ」という。)における調査単位レベルの秘匿性を定量的に把握するために、標本の原データの中から母集団においても一意(unique)となるレコード数を計測することである。第 2 の方法は、様々な匿名化技法を個別データに適用した場合の秘匿性の程度を比較可能にするために、リンケージの手法を用いて、原データと秘匿処理を施した個別データの対応付けを行うことである。本稿では、後者のリンケージによる秘匿性の定量的な評価に着目し、完全照合型リンケージと距離計測型リンケージの 2 種類の秘匿性の評価方法の適用可能性を考察した。そして、平成 16 年全国消費実態調査（以下「全消」という。）の個別データから作成したミクロアグリゲートデータ(ソートなし、個別ランキング法、第 1 主成分法、Z スコア総計法)を用いて、完全照合型リンケージと距離計測型リンケージによる秘匿性の比較分析を行った。

本分析結果によれば、4 種類のミクロアグリゲートデータのいずれについても、距離計測型リンケージにおける真のリンクの比率のほうが、完全照合型リンケージにおける比率よりも大きくなっていることが明らかになった。また、距離計測型リンケージの場合、真のリンクとなるレコードについては、レコード間の距離の長さとは関係なく、真のリンクと判定される可能性があることがわかった。

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

秘匿性の評価方法に関する実証研究
―全国消費実態調査のミクロアグリゲートデータを用いて―

伊藤 伸介，磯部 祥子，秋山 裕美

１　本研究の背景と目的

我が国では、統計法（平成19年法律第53号）の全面施行によって、政府統計のミクロデータ（匿名データ）の提供が本格的に進められている。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の視点だけでなく、ミクロデータに含まれる個体情報の秘匿性(confidentiality)の観点についても考慮する必要がある。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。

諸外国では、ミクロデータの提供において、個体情報の保護に関する法的及び制度的な措置を整備していることが知られている[1]。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を禁じている（石田(2000, 31 頁)、森(2005, 4 頁), Zayatz(2007b, p.253)）。さらに、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(2007b, p.253)）。

また、アメリカでは、センサス局等の統計作成部局において開示評価委員会(Disclosure Review Board)が設置されている。例えば、センサス局の開示評価委員会では、センサス局が提供の対象としているすべてのミクロデータと集計表について、開示（disclosure）を回避するための方針を定め、開示の回避に関する手続きについて審議を行っている。そして、センサス局の開示評価委員会は、ミクロデータの提供による個体情報の開示の可能性を検証するために、「データの潜在的な開示可能性についてのチェックリスト(Checklist on Disclosure Potential of Data)」をミクロデータの提供に関する定性的な基準として採用している(石田(2000, 35～36 頁)、Zayatz(2007b, p.255))。

ところで、チェックリストについては、『統計的開示制限の方法論に関する報告書(Report on Statistical Disclosure Limitation Methodology, 1994、 以下『報告書』という)。』にまとめら

[1] 欧米諸国におけるミクロデータの提供状況、さらには個体情報の保護に関する法的及び制度的な措置の詳細については、松田･濱砂･森（2000）や森（2005）等を参照。

れた勧告が、作成の契機となっている [2]。それは、統計作成部局に対して推奨される実践についての勧告であり、全部で 12 の勧告から構成されている [3]。例えば、勧告２では、「開示制限が施されたデータの作成に関しては、部局レベルでの検討プロセスを中央集権化（centralize）すること」が明記されており、ミクロデータの作成・提供に関する手続きを統計作成部局において一元化する必要性が明らかにされている。また、勧告 12「ミクロデータファイルから直接的な識別子を削除し、他の識別される情報を制限すること」は、ミクロデータにおける基本的な匿名化技法に関する勧告だと言える。これらの 12 の勧告群に加えて、2005 年に刊行された『報告書』の第２版には、勧告 13「統計作成部局は開示リスク（disclosure risk）[4]の評価に関する情報を共有する必要があること」が新たに追加されている。このことは、近年、アメリカの統計作成部局において、ミクロデータの提供に伴う個体情報の開示の可能性を定量的に評価する必要性が高まっていることを示唆している [5]。

さらに、チェックリストにおいて、開示評価委員会がミクロデータと外部情報とのマッチングの可能性を考察していることが記載されており、外部情報とのマッチングによって生じるリスクの要因が列挙されている。「開示評価委員会は、マッチングの潜在的な可能性を回避するためにどのような措置がとられるべきかについて事前に厳密に定めることはできない。しかし、外部のデータベースとのマッチングの可能性が生じた場合には、開示評価委員会は、ミクロデ

---

[2] チェックリストを作成することになった理由としては、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安全性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))。

[3] 『報告書』によれば、統計作成部局に対して推奨される実践は、次のとおりである(Federal Committee on Statistical Methodology(1994, pp.74-78))。

「1. 結果表とミクロデータに関する全般的な勧告
勧告 1　回答者とデータの利用者から助言を求めること
勧告 2　開示制御が施されたデータの作成に関しては、部局レベルでの検討プロセスを中央集権化 (centralize) すること
勧告 3　政府全体にわたってソフトウェアと方法論を共有すること
勧告 4　データセットの重複に対しては、部局間の協力を求めること
勧告 5　(匿名化に関して)一貫した実務を行うこと
2. 度数データ(frequency count data)の集計表
勧告 6　匿名化の手法を比較・検証するための研究を行う必要があること
3. 数量データ(magnitude data)の集計表
勧告 7　劣加法的な(subadditive)開示のルールのみを用いること
勧告 8　p%ルール及び pq 曖昧性(pq-ambiguity)ルールを推奨すること
勧告 9　欠測化(suppression)に関するパラメータを明らかにしないこと
勧告 10　セル欠測化(cell suppression)を行うかあるいは集計結果表の行ないしは列を畳みあげること
勧告 11　集計データの審査(auditing)を行う必要があること
4. ミクロデータ
勧告 12　ミクロデータファイルから直接的な識別子を削除し、他の識別される情報を制限すること。」

[4] 開示リスクとは、ミクロデータの提供によって回答者が特定される危険性（竹村（2003，241 頁））だと考えられている。

[5] 『報告書』には、ミクロデータと集計データの開示制限に関する将来的な研究課題とそれらの課題の優先度が示されているが（Federal Committee on Statistical Methodology(1994, pp.79-88)）、『報告書』によれば、ミクロデータにおける開示リスクの定義を定め、リスクを評価することに関する優先順位が最も高くなっている。

ータファイルの公開に関するリスクを決定するいくつかの要因を考察している。その要因とは、(1)マッチングの目的のために利用可能な変数の数、(2)マッチングを行う上で必要なデータ源、(3)データの経年、(4)外部データの取得可能性、信頼性と完全性、(5)データのセンシティビティ（sensitivity)ないしは一意性(uniqueness)[6]である。」(Zayatz(2007a, pp.7-8))。

チェックリストのような定性的な資料において、ミクロデータにおける一意性といった開示リスクの定量的な評価基準に関連する概念が、ミクロデータの提供に関するリスク要因の１つとして指摘されていることは、非常に興味深い。

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。例えば、Zayatz(2007b, pp.256-257)によれば、アメリカセンサス局では、2000 年人口センサスの一般公開型ミクロデータ(Public Use Microdata Sample)を提供するために、(1)識別子（名前、住所等）の削除、(2)地域区分の制限、(3)丸め込み（ラウンディング）、(4)ノイズの導入、(5)質的変数における閾値の設定と質的属性値の再符号化（リコーディング）、(6)トップ・コーディング、ボトム・コーディング、(7)データ・スワッピングを主な匿名化技法として採用している。これらの匿名化技法の適用に対しては、開示リスクの定量的な評価を併せて考慮している可能性が指摘できる。例えば、地域区分についてはHawala(2001)による開示リスク評価の定量的な研究が存在する。また、センサス局は、「特殊な一意（special uniques）[7]」(Elliot(2001, p.84))という概念に基づいて、データ・スワッピングによる匿名化を行っている(Zayatz(2007b, p.257))。

このような先行事例を見ると、少なくともアメリカでは、統計作成部局がミクロデータに対して秘匿処理を行う上で、開示リスクの定量的な評価の必要性に対する認識を持っているように思われる [8]。

諸外国では、政府統計のミクロデータを用いて開示リスクの定量的な評価を行った先行研究が存在する(Bethlehem *et al.*(1990), Marsh *et al.*(1991), Müller *et al.*(1995), Hawala(2001), Dale and Elliot(2001), Kim and Jeong(2007)等)。これらの先行研究は、ミクロデータの提供における開示リスクの視点から、秘匿性の定量的な評価を行っている。一方、我が国においてミクロデータの提供を今後展開していく上では、諸外国における開示リスクの評価方法に関す

---

6 Federal Committee on Statistical Methodology(1978, pp.25)によれば、一意性(uniqueness)とは、ある個体が、ミクロデータで利用可能な情報を用いることによって、母集団に含まれる他のすべての個体と区別できる状況を表している。

7 Elliot(2001)によれば、「疫学的に特異であるために、本質的に（intrinsically）まれな属性群の組み合わせを有する」レコードは、母集団において一意となるレコードとして特定化される可能性が非常に高くなると考えられている。それは、「特殊な一意問題（special uniques problem）」と呼ばれている（Elliot(2001, p.84)）。

8 例えば、アメリカセンサス局に関しては、Hawala(2003)、Steel(2004)等を参照。

る研究事例を参考にしながらも、調査単位の違いや匿名化技法の相違による秘匿性の程度の比較という観点から、ミクロデータにおける秘匿性の定量的な評価方法を具体的に検討しておくことも意義があるものと考えられる。それによって、少なくとも研究段階においては、ミクロデータにおける調査単位の特性を踏まえ、各種の匿名化技法を適用することによって作成したミクロデータの秘匿性の程度を比較・検討することが可能になると思われる。

ところで、伊藤・磯部・秋山(2008)は、匿名化技法の１つであるミクロアグリゲーションに焦点を当て、ミクロデータの有用性の観点から、ミクロアグリゲーションの有効性の検証を行っている。しかし、ミクロアグリゲーションによって作成したデータ（以下「ミクロアグリゲートデータ」という。)の秘匿性については、定量的な評価を行っていない。そのためには、秘匿性の定量的な評価方法を考案した上で、ミクロアグリゲートデータに適用可能な秘匿性の評価方法を探究する必要があると思われる。

そこで、本稿では、ミクロデータの秘匿性に関する定量的な評価方法に焦点を当て、我が国の政府統計の個別データから作成した様々なミクロアグリゲートデータを用いて、ミクロアグリゲートデータに対する秘匿性の評価方法の適用可能性を追究する。

２　個体識別とは

諸外国では、政府統計ミクロデータに関する秘匿性の把握にあたって、ミクロデータに含まれる個体情報の開示リスクの評価方法が主に議論されている。先行研究によれば、開示は、主として、個体識別開示（identification disclosure）と予測開示（prediction disclosure）に類別することができる（Duncan and Lambert(1989), Skinner(1992), 佐井（2000）, 竹村（2003））。個体識別開示は、ミクロデータに含まれるレコードから目標となる個体情報の特定化を行うことによって、個体に関するセンシティブな情報が露見されることを表すのに対して、予測開示は、ミクロデータの提供によって、個体のセンシティブな属性について狭い範囲で予測可能になることである(Skinner(1992, p.23))。本節では、個体識別による開示に焦点を当て、個体識別の概要について簡単に述べることにしたい[9]。

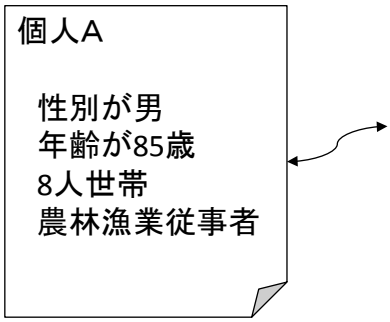
個体識別は、次のように考えることができる（Bethlehem *et al.*(1990), Marsh *et al.*(1991), Müller *et al.*(1995)）。侵入者（intruder）が、識別の対象となる特定の個体について把握している情報（事前情報（a priori knowledge））を含むファイル（識別ファイル）とミクロデータ

[9] 竹村(2003)は、個別データの開示問題に関するサーベイを行っているが、竹村(2003)においても、個体識別による開示に焦点を絞って、開示リスクの評価について議論している。

ファイルを持っていたと想定する。そのとき、①識別ファイルに含まれるレコードとミクロデータファイルに含まれるレコードにおいて、キー変数（key variable）を通じて１対１のマッチングがなされ、②対応関係にあるレコードが特定の個体のものであることが確認されるとき、個体識別が成立する。なお、キー変数とは、事前情報としてミクロデータ以外から取得可能であり、個体の識別が可能な変数である(佐井(2000，229 頁))。

母集団に関する外部情報(例えば電話帳等)を持つ侵入者は、ミクロデータを取得した場合に、外部情報とミクロデータのマッチングを行うことによって、個体識別を行うことが想定される。図１は、個体識別による開示のイメージを表した概略図である。侵入者は性別が男、年齢が 85 歳、世帯人員区分が８人世帯で職業が農林漁業従事者という個人 A の属性値に関する情報を有しているとする。侵入者はミクロデータを取得することによって、性別、年齢、世帯人員区分及び職業のキー変数を用いて、ミクロデータに含まれる属性群と個人 A について持っている外部情報とのマッチングを行い、ミクロデータの中から個人Ａに関する情報を特定しようとする。図１は、個人Ａに関する情報に該当するレコードが、一連番号が 00003 と付与されたレコードであることを示しているが、もし侵入者がそのレコードが個人Ａのレコードであることを確認できた場合、個人Ａとなるレコードが特定され、侵入者はそれまで持っていなかった年間収入等の情報を新たに入手することが可能になる。

図１　個体識別のイメージ例



一連番号	（世帯主）性別	（世帯主）年齢	世帯人員区分	（世帯主）職業	年間収入（万円）	・・・
00001	2	25	1	9	200	・・・
00002	1	36	4	1	800	・・・
00003	**1**	**85**	**8**	**7**	**6000**	**・・・**
00004	2	45	3	5	400	・・・
00005	1	70	2	5	300	・・・
・	・	・	・	・	・	
・	・	・	・	・	・	
・	・	・	・	・	・	

性別　1: 男　2:女
職業　1:専門的･技術的職業従事者　2: 管理的職業従事者　3:事務従事者
4:販売従事者　5: サービス職業従事者　6:保安職業従事者　7: 農林漁業従事者
8: 運輸･通信従事者　9: 生産工程･労務作業者

ところで、個体識別については、侵入者が入手可能なキー変数としてどのような属性を想定するかが重要な意味を持つと思われる。Elliot and Dale(1999, pp.8-9)によれば、キー変数の選択に関する明確な基準は存在しないが、キー変数についての類型化は可能であって、キー変数

の対象となる情報源を、(1)一般に利用可能な情報、(2)個人的な(非公式の)知識及び(3)企業や政府といった組織が保有する(organizational) データベースに類別している。しかし、選択されるキー変数は、侵入者による個体情報の識別に関していかなる戦略(strategy)が想定されるか[10]、さらには、侵入者にとってどのような属性がキー変数として入手可能かによって、異なっている。よって、開示リスクの定量的な評価を行う場合には、事前に個体識別に関する戦略やキー変数の入手可能性を検討した上で、キー変数の選定を行うことが必要かと思われる。

３　秘匿性の評価方法の概要

諸外国では、主として個体識別開示のリスクに関する尺度に基づいて、秘匿性を定量的に評価する研究がこれまで展開されてきた。『報告書』によれば、「開示リスクに関する尺度がなければ、ミクロデータファイルの開示制御に関する意思決定は、慣例及び個人的見解(judgment calls)に基づかなければならない」ことから、「ミクロデータにおいては、確率に基づいた開示の定義に向けた研究の優先度が高くなければならない」(Federal Committee on Statistical Methodology(1994, p.80))。そこで、『報告書』では、先行研究において提唱されている開示リスクの尺度として次の４点が指摘されている。

「・侵入者が突き止めようとしている回答者のレコードが、ミクロデータファイルと何らかのマッチング可能なファイル(matchable file)のいずれにも存在している確率

・マッチングの対象となる変数(matching variables)がミクロデータファイルとマッチング可能なファイル上に同一の形態で記録されている確率

・侵入者が突き止めようとしている回答者のレコードが、マッチング可能な変数群(matchable variables)に関して母集団の中で一意である確率

・侵入者が一意となった回答者のレコードが誰のレコードなのかを正しく特定した確実性の程度」(Federal Committee on Statistical Methodology(1994, p.65))[11]

---

[10] Müller *et al.*(1995, p.135)は、侵入者における識別の戦略として、(1)直接検索(directed search)と(2)釣り検索(the fishing strategy)に類別している。直接検索は、識別ファイルを用いて、ミクロデータファイルに含まれるある特定の個体のレコードを突きとめようとする戦略である。それに対して、釣り検索では、侵入者はミクロデータファイルの中で関心があるレコードに焦点を絞り、それらのレコードを識別するために、識別ファイルの中で対応付け可能なレコード群を探り出そうとする戦略である。また、直接検索においては、特定の対象となる個体のレコードがミクロデータの中に含まれるという情報(参入情報(participation knowledge))を侵入者が持っている場合が想定されている。なお、参入情報がある場合の直接検索がもっとも開示リスクが高いシナリオだと考えられている(Müller *et al.*(1995, p.139))。

[11] Marsh *et al.*(1991)では、定量的な開示リスクの評価が試みられている。
ある個体が正確に識別される確率は①式であたえられる。

P(識別)=P(識別|試み)P(試み)・・・①

そして P(識別|試み)は、次の②式で計算される。

これらの４つの尺度は、外部情報とミクロデータのマッチングによる個体識別開示のリスクに関連した尺度であると言うことができよう。

ところで、図２は、秘匿性の評価のイメージを図示したものであるが、侵入者がミクロデータを取得し、外部情報とのマッチングによって個体識別を試みる場合、ミクロデータの秘匿性に関する定量的な評価に関しては、２つの論点が考えられる。第１の論点は、外部情報の取得可能性及び外部情報とミクロデータのマッチング[12]に関する検討を行うことである。第２の論点は、様々な匿名化技法をミクロデータに適用した場合の秘匿性の程度を相対的に比較することである。本研究は、ミクロデータにおける秘匿性の程度を比較可能にするための定量的な評価方法を考察することを指向しており、個体識別開示のリスクを評価することを目指していない。そのために、本研究では、外部情報の取得可能性及び外部情報とミクロデータのマッチングについては考察の対象にしていない[13]。そこで、本研究では、第２の論点である匿名化技法の適用によるミクロデータの秘匿性の程度を比較・検討することに焦点を当てることにしたい。

P(識別|試み)=P(a)P(b|a)P(c|a,b)P(d|a,b,c)・・・②

ここで

P(a)：侵入者が有する事前情報と目標となる個体についてのミクロデータのいずれにおいてもキー変数が同一に記録されている確率

P(b|a)：条件(a)を満たすときに、目標となる個体がミクロデータの中に存在する確率

P(c|a,b)：条件(a)、(b)を満たすときに、目標となる個体におけるキー変数の値の組み合わせが一意である確率(母集団一意である確率)

P(d|a,b,c)：条件(a)、(b)と(c)を満たすときに、ミクロデータの利用者が、キー変数の値の組み合わせが母集団において一意であることを確証する確率

P(試み)の計測は困難であることから、Marsh *et al.*(1991)では、P(識別|試み)の計測を行っている。P(a)は格付けの誤り、未入力、分類区分の不一致等に起因している。また、P(b|a)は標本の抽出率に対応している。さらに、P(c|a,b)については、母集団においてキー変数の値の組み合わせが一意であるレコード数の比率が求められる。Marsh *et al.*(1991)では、P(c|a,b)を計測するために、イタリアのセンサスの個票データ(約 350 万)から１万レコードをランダムに抽出し、これらのレコードの各々が、同じ個票データから抽出された他のレコードと対応付けされる割合を求めている。また、P(d|a,b,c)については、非常に多くのキー変数に関する事前情報がない限り、その値は 0 に近くなると考えられている。

Marsh *et al.*(1991)によれば、実験の結果により P(a)＝0.6、P(b|a)=0.02、 P(c|a,b)=0.02、P(d|a,b,c)＝0.001と設定されている。よって、個体が正確に識別される確率は、P(識別)=0.6×0.02×0.02×0.001=2.4×10^{-7} となる。

[12] 外部情報とミクロデータのマッチングの実験については、例えば、Müller *et al.*(1995)による先行研究がある。

[13] Elliot(2001)は、外部情報とミクロデータのマッチングについては、侵入者が正確な対応付けを行うための手続きを把握する上で非常に有益であることを論じている。しかし、Elliot(2001)は、開示リスクを評価するために外部情報とミクロデータにおけるマッチング手法を利用する場合、2 つの重大な欠点があることを指摘している。第１の欠点は、特定の外部情報を用いてミクロデータとのマッチングを行った場合、その結果がミクロデータに対する開示リスクを評価するための十分な尺度を与えるということが確証されないことである。第 2 の欠点は、マッチングの実験を行う場合多くの時間を要することである(Elliot(2001, p.80))。

図２　秘匿性の評価のイメージ

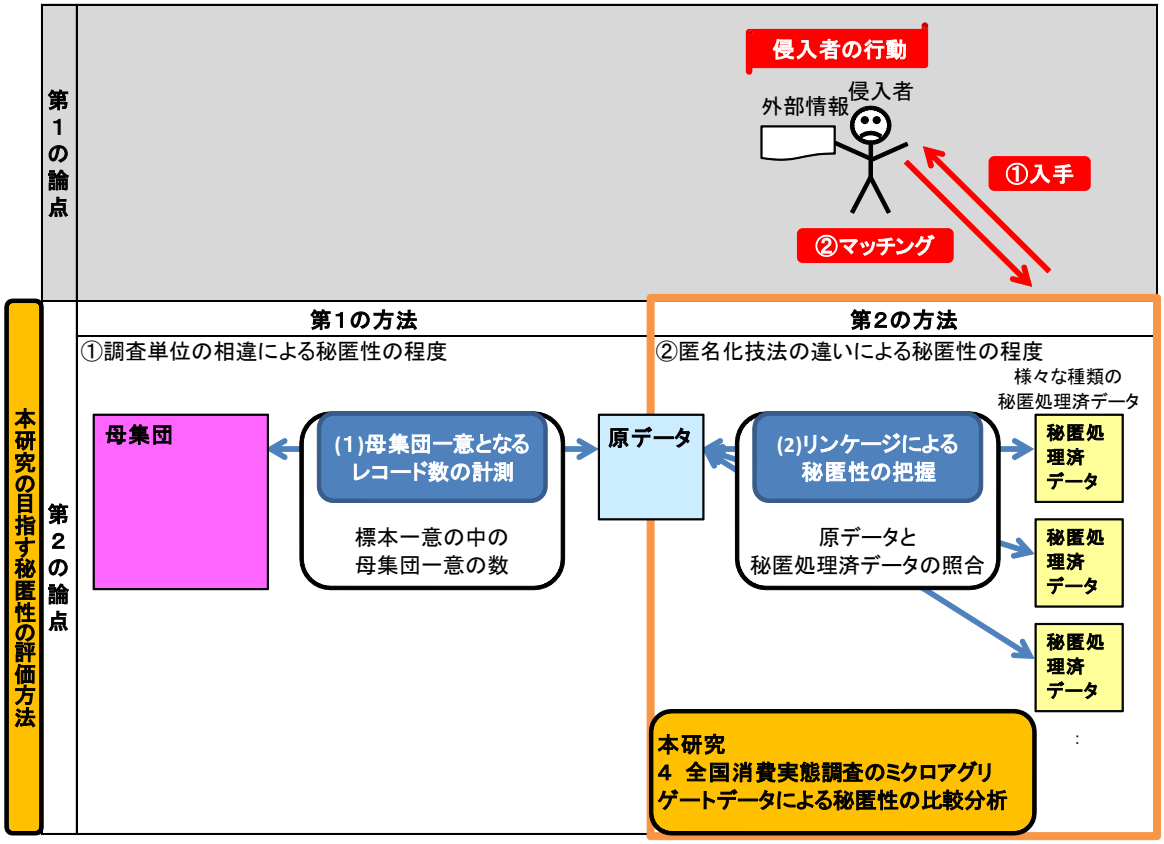


図２に示されるように、原データに対して、匿名化技法を適用することによって作成したミクロデータ（以下「秘匿処理済データ」という。）が想定される。一般に、秘匿処理済データの作成においては、原データにおける調査単位の特性によって、適用される匿名化技法が異なると考えられる。よって、本研究では、秘匿処理済データの秘匿性の程度を定量的に把握するために、①調査単位の相違による秘匿性の程度及び②匿名化技法の違いによる秘匿性の程度の２つの視点に着目した。それらの視点に基づき、本研究では、次の２つの秘匿性の評価方法を議論することにしたい。第１の方法は、標本の原データにおける調査単位レベルの秘匿性を定量的に把握するために、標本の原データの中から母集団においても一意となるレコード数を計測することである。第２の方法は、様々な匿名化技法間の秘匿性の程度を比較可能にするために、レコードリンケージ(record linkage)の手法を用いて、原データと秘匿処理済データとの対応付けを行うことである。

(1) 母集団一意となるレコード数の計測

ミクロデータには、「非常に目につきやすいレコード(high visibility record)」が含まれてい

る可能性がある(Federal Committee on Statistical Methodology(1994, p.62))。そのため、回答者の属性の組み合わせによっては、母集団の中に一意となる回答者が存在することも考えられる。それは、母集団一意(population unique)と呼ばれている。

母集団一意は、標本の中に一意となる回答者が存在する標本一意(sample unique)とは異なる概念である。図３は、母集団一意と標本一意の考え方を図示したものである。図３では、キー変数である性別、世帯人員区分と職業を用いて、標本と母集団のそれぞれについてクロス集計表が作成されている。標本に関する集計表を見ると、８人世帯に属し、生産工程・労務作業者として従事している男性ないしは女性については、セルの度数が１となることがわかる。しかし、同様の属性値の組み合わせを母集団についても確認すると、８人世帯に属し、生産工程・労務作業者として従事している女性に関するセルの度数のみが１となっている。このことは、これらの属性値群を持つレコードが、標本一意であり、かつ母集団一意でもあることを表している。それに対して、８人世帯に属し、生産工程・労務作業者として従事している男性については、標本一意ではあるが、母集団一意ではないことが明らかになっている。

図３　母集団一意と標本一意の考え方

（標本）

		性別							
		男				女			
		世帯人員区分							
		2人	…	8人	…	2人	…	8人	…
職業	専門的・技術的職業従事者	27		5		31		4	
	:	21		7		47		5	
	農林漁業従事者	48		9		20		9	
	運輸・通信従事者	48		4		15		3	
	生産工程・労務作業者	14		1		48		1	

標本一意

（母集団）

		性別							
		男				女			
		世帯人員区分							
		2人	…	8人	…	2人	…	8人	…
職業	専門的・技術的職業従事者	56		15		83		47	
	:	68		39		59		37	
	農林漁業従事者	57		45		77		33	
	運輸・通信従事者	83		39		67		48	
	生産工程・労務作業者	54		3		83		1	

母集団一意

標本一意かつ
母集団一意

母集団一意となるレコードが標本の原データに存在する場合、本研究では、それは、調査単位による秘匿性の相違を定量的に評価するための指標として捉えられる。そのため、母集団の中で一意となるレコード数を定量的に把握することが求められる。母集団一意の計測については、次の２つの方法が考えられる。第１に、母集団一意となるレコード数を数え上げることである。それは、具体的には、母集団のデータを用いて、選択されたキー変数に関するクロス集計を行い、度数が１となるセルを算出することであり、我が国では、これまで「寸法指標(size index)」の枠組みの中で議論されてきた(竹村(2003))。もし、センサスの原データ等が使用できるのであれば、母集団一意となるレコード数を数え上げることは可能かと思われる。第２に、標本一意となるレコード数から母集団一意となるレコード数を推定することである。この方法については、ポアソン・ガンマモデル(Poisson-gamma model)(Bethlehem *et al.*(1990))や対数線形モデル(Fienberg and Makov(1998))といったモデル等を用いて母集団一意となるレコード数を推測することが考えられる [14]。

Elliot(2001, pp.81-82)によれば、母集団一意となるレコード数を計測するために、母集団の一意性(population uniqueness)と共通一意(union uniques)という概念が用いられていることが知られている [15]。母集団の一意性とは、母集団全体のレコード数の中に占める母集団一意となるレコード数の比率である。また、共通一意とは、標本一意かつ母集団一意であることを表す概念であり、共通一意であるレコード数の標本一意となるレコード数に対する比率は、UUSU比率と呼ばれている。UUSU比率は、標本一意となるレコード群の中で母集団一意と判定されるレコードがどの程度存在するかを評価する指標である。これらの２つの指標のいずれも、一意性の概念を用いた秘匿性の定量的な評価方法だということができる。

(2) リンケージによる秘匿性の把握

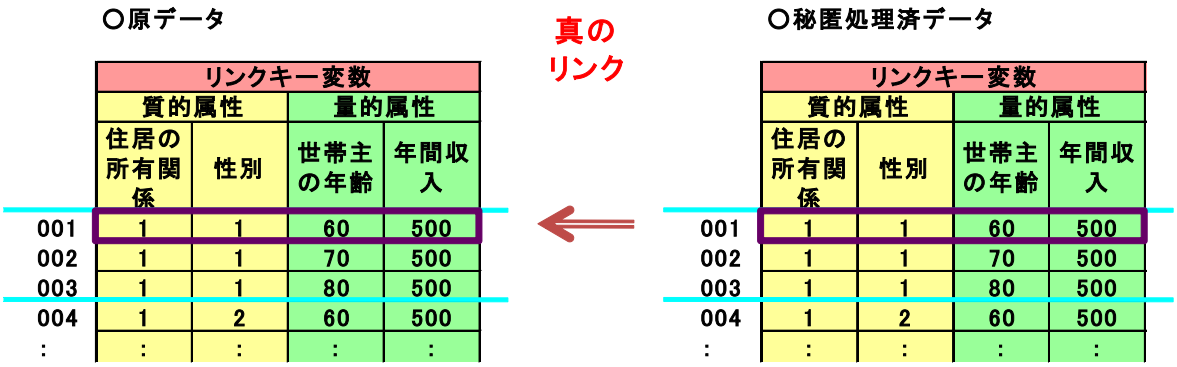
秘匿処理済データは、原データに様々な匿名化技法を適用することによって作成される。このような各種の匿名化技法を用いて作成する秘匿処理済データについて、その秘匿性を定量的に評価することが考えられる。本研究では、原データと秘匿処理済データの間でレコードリンケージ(Domingo-Ferrer and Torra(2001))を行うことによって、秘匿性について定量的に評価する方法を議論する。具体的には、本研究では、次の２つのリンケージの方法に着目した(Herzog *et al.*(2007), Domingo-Ferrer and Torra(2001), Winglee *et al.*(2002))。

14 モデルを用いた母集団一意の推測については、佐井(2000)等を参照されたい。

15 Elliot(2001, pp.81-82)は、母集団の一意性と UUSU 比率のいずれについても、それらの指標を計測するための要件として、母集団のデータにアクセス可能であることを指摘している。

第１の方法は、対応付けを行うためのキーとなる属性群（以下「リンクキー変数」という。）を用いて、完全照合によるリンケージを行うことである(以下「完全照合型リンケージ(deterministic record linkage)」という。)。図４は、原データと秘匿処理済データにおける完全照合型リンケージのイメージを図示したものである。完全照合型リンケージでは、秘匿処理済データのレコードとその元になる原データのレコードとの間で、リンクキー変数を用いて照合可能かどうかを判定する。例えば、図４では、リンクキー変数として、質的属性については住居の所有関係と性別、量的属性については世帯主の年齢と年間収入が想定されているが、これらの４つのリンクキー変数に関する属性値の各々が１対１で照合するかどうかを秘匿処理済データに含まれる各レコードについて検証する。そして、秘匿処理済データに含まれる特定のレコードが、その元になるレコードに１対１で照合可能な場合には(以下「真のリンク」[16]という。)、秘匿処理済データのレコードから原データにおいてその元になるレコードを対応付けることが可能になる。

図４　完全照合型リンケージのイメージ



第２の方法は、リンクキー変数を用いて、原データのレコードと秘匿処理済データのレコード間の距離を計測することである（以下「距離計測型リンケージ(distance-based record linkage)」という。)。図５は、原データと秘匿処理済データにおける距離計測型リンケージのイメージを図示したものである。距離計測型リンケージでは、秘匿処理済データに含まれる特定のレコードについて、原データの各レコードとの距離を測り、その距離が最も短いレコードが、原データのその元になるレコードに対応付けられるどうかを判定する。図５は、完全照合

16 本研究で議論しているリンケージの手法は、原データと秘匿処理済データの照合によって、秘匿処理済データに含まれる特定のレコードが、原データにおけるその元のレコードと対応付け可能かどうかを指向したものであり、個体識別に関する開示リスクの評価を目指していない。ゆえに、秘匿処理済データにおいて「真のリンク」となるレコードが見つかったとしても、そのことがミクロデータに含まれる個体情報を直接特定することにはならないことに留意されたい。

型リンケージと同様に４つのリンクキー変数を用いて、秘匿処理済データにおける特定のレコードと原データに含まれる各レコードの間の距離を計測した結果を示したものである。なお、図５では、距離の計測方法として、ユークリッド距離[17]を使用している。図５を見ると、質的属性である住居の所有関係と性別の属性値について層化を行った上で、秘匿処理済データにおいて一連番号 001 を持つレコードと原データにおいて同一の層内に含まれる各レコードとの間で、世帯主との年齢及び年間収入に関する距離が計測されている[18]。一連番号 001 のレコードとの距離が最短であるレコードが、原データにおける元のレコードに一致した場合に限り、真のリンクとみなすことができる。図５は、真のリンクが成立する例を示している[19]。

図５　距離計測型リンケージのイメージ

○原データ

	リンクキー変数			
	質的属性		量的属性	
	住居の所有関係	性別	世帯主の年齢	年間収入
001	1	1	60	500
002	1	1	70	500
003	1	1	80	500
004	1	2	60	500
:	:	:	:	:

真のリンク

ユークリッド距離
0.34
0.52
1.21

○秘匿処理済データ

	リンクキー変数			
	質的属性		量的属性	
	住居の所有関係	性別	世帯主の年齢	年間収入
001	1	1	61	500
002	1	1	70	500
003	1	1	80	500
004	1	2	60	500
:	:	:	:	:

４　全国消費実態調査のミクロアグリゲートデータによる秘匿性の比較分析

前節では、母集団一意となるレコード数の計測及びリンケージによる秘匿性の把握という２つの視点に基づき、秘匿性の評価方法についての整理を行った。本節では、リンケージによる秘匿性の評価方法に着目し、我が国の政府統計の原データから作成したミクロアグリゲートデータを用いて完全照合型リンケージ及び距離計測型リンケージを行うことによって、秘匿性の定量的な評価を試みることにしたい。

---

17 本研究で使用するユークリッド距離は、次のように表記される。

$d_{iI} = \sqrt{\sum_j (x_{ij} - X_{Ij})^2}$

ここで

d_{iI}：原データにおける i 番目のレコードと秘匿処理済データにおける I 番目のレコードに関するユークリッド距離

x_{ij}：原データにおいて i 番目のレコードにおける j 番目の属性

X_{Ij}：秘匿処理済データにおいて I 番目のレコードにおける j 番目の属性

18 図５では、ユークリッド距離を計測するために、世帯主との年齢及び年間収入を標準化した上で、距離の計測を行っている。

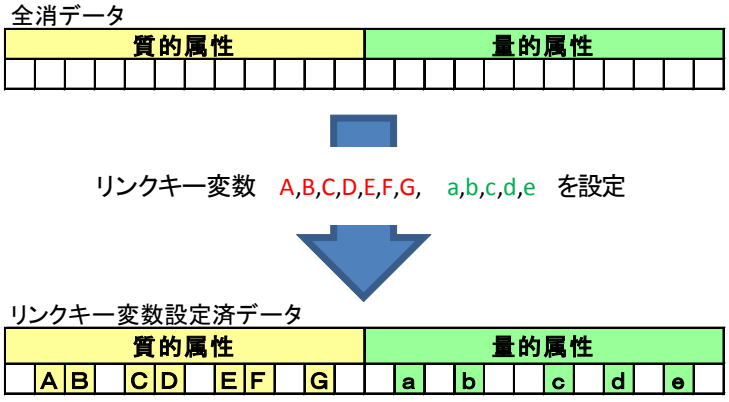
19 本研究は、確率的なレコードリンケージ(probabilistic record linkage)については考察の対象としていない。確率的なレコードリンケージの概要については、例えば Herzog *et al.*(2007, pp.83-91)を参照されたい。

（1）ミクロアグリゲートデータの作成方法

本研究で使用するデータは、全消の原データ(二人以上の世帯、55,056 世帯)より作成したミクロアグリゲートデータである。本研究では、伊藤・磯部・秋山（2008）で用いた方法に基づき、以下の手順でミクロアグリゲートデータの作成を行っている。最初に、全消の原データに含まれる質的属性群を対象に、ミクロアグリゲートデータに設定可能な質的属性の組合せリストを作る。次に、量的属性については、レコードのソート、レコードのグループ化及びグループ内のレコードが有する属性値の平均値への置き換えというミクロアグリゲーションの手法を適用することによって、ミクロアグリゲートデータを作成する。

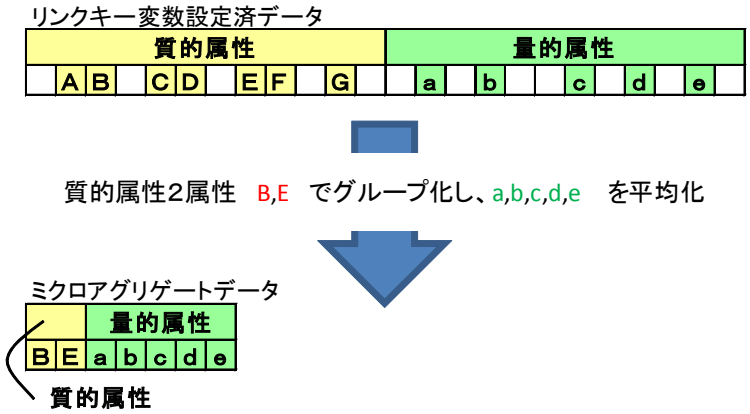
ところで、本研究は、秘匿性の評価方法の検討を目的としたミクロアグリゲートデータを作成することを指向している。そこで、全消の原データに含まれる属性群の中から、リンクキー変数となる属性群をあらかじめ選び出す必要がある。例えば、図６は、リンクキー変数の選定に関する模式図である。図６では、全消の原データに含まれる質的属性及び量的属性の中から、Ａ～Ｇ７つの質的属性、ａ～ｅの５つの量的属性がリンクキー変数として選ばれたことを示している。なお、本研究では、調査実施者の意見を踏まえた上で、全消の全調査項目の中から本研究のために必要だと思われる属性が、リンクキー変数の対象となっている。本研究では、全消の原データの中から、「住宅の所有関係」などの質的属性７属性と、「世帯主の年齢」など量的属性５属性をリンクキー変数として選定した。

図６　リンクキー変数の設定（イメージ）



次に、リンクキー変数を含む全消の原データのすべての質的属性群を用いて超高次元クロス集計を行った上で、セルに度数１又は２のない組合せに関する「質的属性の組合せリスト」を作成した(伊藤・磯部・秋山(2008，46 頁))。そして、質的属性の組合せリストから、質的属性について全てがリンクキー変数である組合せを選択した。図７では、Ａ～Ｇの７つの質的属性から属性Ｂと属性Ｅをリンクキー変数として選び出したことを表している。

図７　リンケージに使用したミクロアグリゲートデータ（イメージ）



最後に、選択された質的属性に量的属性群を追加的に設定し、量的属性に対してミクロアグリゲーションを実行することによって、ミクロアグリゲートデータを作成した。本研究では、量的属性に関するミクロアグリゲーションとして、「ソートなし」と「個別ランキング法」[20]に加え、「第１主成分法」及び「Zスコア総計法」[21]を用いることによって（伊藤（2008, 8 頁））、様々なミクロアグリゲーションの手法を適用した場合の秘匿性の程度について検証を試みている。図７は、リンクキー変数である量的属性a～eについてミクロアグリゲーションを行うことによって、質的属性ＢとＥにミクロアグリゲーション済の量的属性５属性が、ミクロアグリゲートデータとして追加されたことを示している。

以上の手順に従って、本研究では、全消の原データにおける属性群の中からリンケージに用いるリンクキー変数を選択し、ミクロアグリゲートデータを作成した。ミクロアグリゲートデータに含まれるリンクキー変数は、「住宅の所有関係」等の質的属性２属性及び「世帯主の年齢」等の量的属性５属性である。

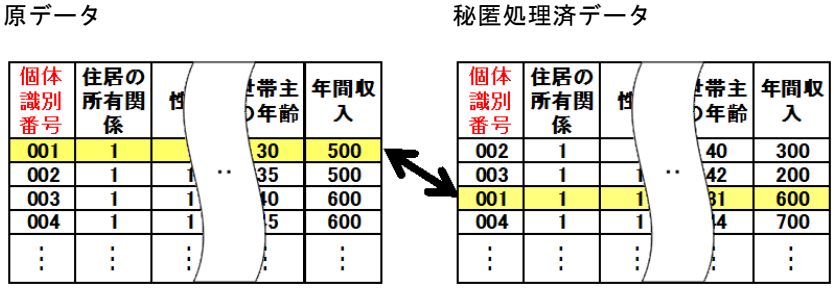
20 「ソートなし」と「個別ランキング法」の詳細については、伊藤・磯部・秋山（2008，48～51 頁）を参照。
21 「第１主成分法」と「Z スコア総計法」によるミクロアグリゲートデータの作成手順については、本稿の付録「「第１主成分法」と「Z スコア総計法」によるミクロアグリゲーションについて」を参照されたい。

（2）リンケージによる処理の手順

本研究では、完全照合型リンケージと距離計測型リンケージを行うために、次のような処理を行っている。

最初に、リンケージを行うための前処理として、原データと秘匿処理済データの各レコードに個体識別番号[22]を付与している。一般に、ミクロデータを提供する場合、ミクロデータの各レコードには原データとの対応付けが可能な個体識別番号は与えられていない。しかし、本研究では、秘匿処理済データを作成する際に、リンケージのための個体識別番号を事前に設定した上で、原データと秘匿処理済データのリンケージを行った。それによって、秘匿処理済データの各レコードが、原データにおけるどのレコードに対応しているかを確認することが可能となっている（図８）。

図８　個体識別番号の付与されたデータ（イメージ）



次に、完全照合型リンケージと距離計測型リンケージの各々について処理の手順を見ていくことにする。

完全照合型リンケージは、原データと秘匿処理済データにおいて同一の個体識別番号を持つレコード同士でリンクキー変数による照合を行うことによって、その一致の程度を検証する。そのために、完全照合型リンケージでは、次の手順でリンケージの操作を行っている。

第１に、リンクキー変数の中の質的属性を用いてレコードの層化を行う。図９は、秘匿処理済データにおける質的属性による層化のイメージを図示したものである。図９では、質的属性である住居の所有関係と性別のいずれについても、属性値が同一であるレコードを層化していることが示されている。

22 通常、個体識別番号として用いられているのは、都道府県番号、市区町村番号、調査区番号、一連番号等である。

図９　秘匿処理済データにおける質的属性による層化のイメージ

	リンクキー変数							
個体識別番号	質的属性		量的属性		質的属性		量的属性	
	住居の所有関係	性別	世帯主の年齢	年間収入	就業・非就業	…	消費支出	…
001	1	1	60	500				
002	1	1	70	500				
003	1	1	80	500				
004	1	2	60	500	:	:	:	:
:	:	:	:	:	:	:	:	:

第２に、量的属性を用いて、原データと秘匿処理済データにおいて個体識別番号が同一であるレコード間の照合を行う。このとき、秘匿処理済データが持つ量的属性値の桁数が原データにおける元のレコードのそれと異なることが予想されることから、本研究では、秘匿処理済データにおける量的属性値は、原データの量的属性と同じ桁数にそろえた上で、区間比較を行っている。

第３に、原データのレコードと秘匿処理済データにおいて個体識別番号が同一であるレコードについて、量的属性群が１対１で照合する場合を真のリンクとみなし、真のリンクとなるレコード数を数え上げる。図 10 と図 11 はそれぞれ、完全照合型リンケージにおける真のリンクと真でないリンクの事例を示したものである。図 10 は、真のリンクが成立した場合の例である。図 10 において個体識別番号が 001 であるレコードを見ると、原データと秘匿処理済データのいずれについても、量的属性値が一致しているだけでなく、原データにおいて個体識別番号 001 のレコードと同じ量的属性値を持つレコードが、他に存在しないことがわかる。それに対して、真でないリンクの場合は、少なくとも次の２つのケースが考えられる。第１のケースは、原データと秘匿処理済データで個体識別番号が同一であるレコード同士でリンケージを行った結果、量的属性値が照合しなかった場合である（図 11 のケース１）。第２のケースは、個体識別番号及び量的属性は一致するが、同一の量的属性値を持つレコードが、原データか秘匿処理済データのいずれかに複数存在する場合である（図 11 のケース２）。

図 10　完全照合型リンケージにおける真のリンクの例

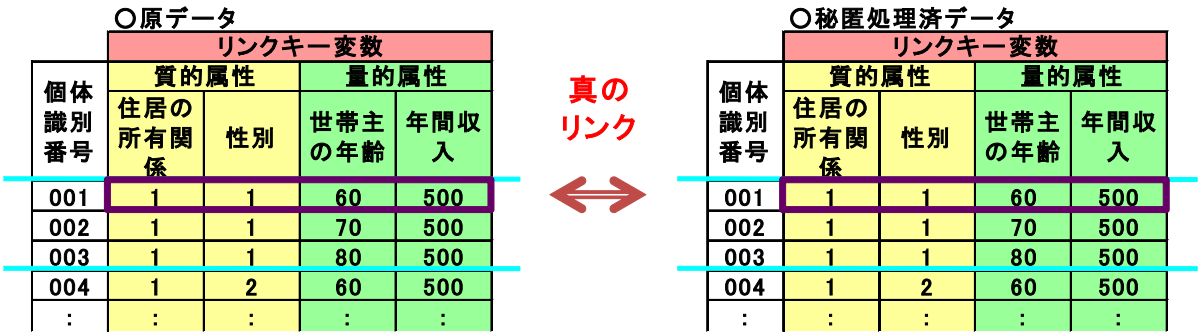
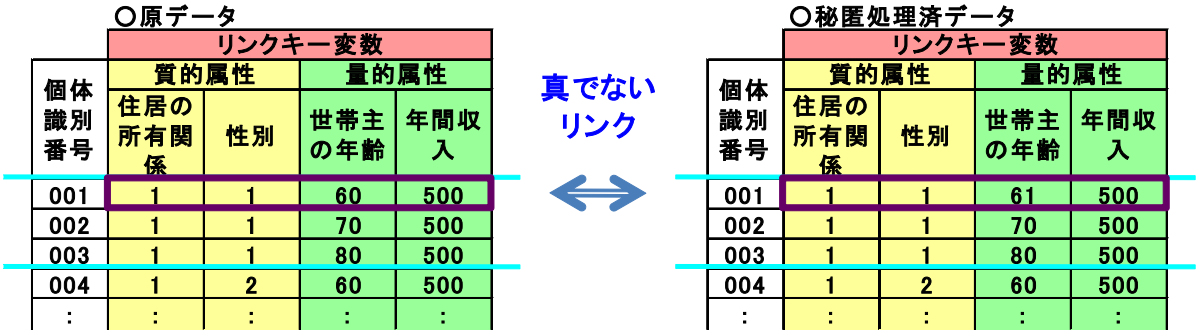


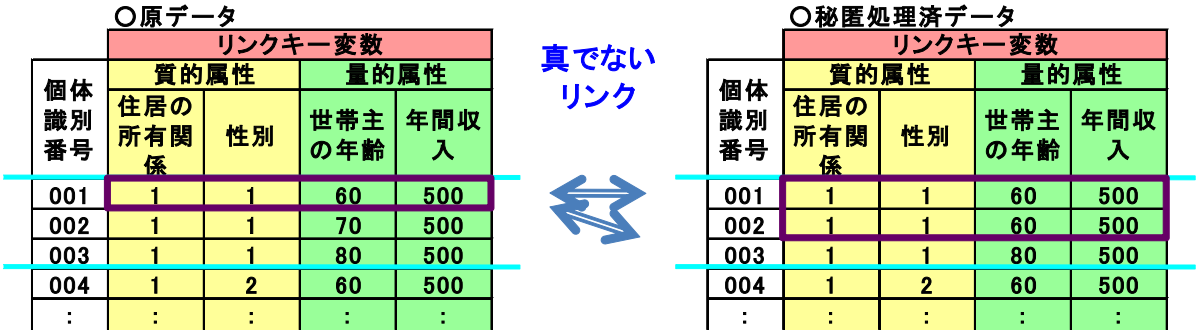
図 11　完全照合型リンケージにおいて真でないリンクの例

ケース１　個体識別番号は一致するが、量的属性値が照合しない場合



ケース２　個体識別番号及び量的属性は一致するが、同じ量的属性値を持つレコードが、原データか秘匿処理済データのいずれかに存在する場合

1) 秘匿処理済データに同一の属性値を有するレコードが複数存在する場合



2)原データに同一の属性値を有するレコードが複数存在する場合

○原データ

個体識別番号	リンクキー変数			
	質的属性		量的属性	
	住居の所有関係	性別	世帯主の年齢	年間収入
001	1	1	60	500
002	1	1	60	500
003	1	1	80	500
004	1	2	60	500
:	:	:	:	:

真でないリンク

○秘匿処理済データ

個体識別番号	リンクキー変数			
	質的属性		量的属性	
	住居の所有関係	性別	世帯主の年齢	年間収入
001	1	1	60	500
002	1	1	70	500
003	1	1	80	500
004	1	2	60	500
:	:	:	:	:

次に、距離計測型リンケージでは、原データと秘匿処理済データにおける距離の計測によって、秘匿処理済データにおけるレコードが、原データにおいてその元になるレコードに対応付けられるどうかを検証することを指向している。具体的には、秘匿処理済データの対象となるレコードと原データの各レコードとの距離を測定し、原データにおいてその距離が最短となるレコードが、秘匿処理済データにおいて距離の計測の対象にしていたレコードと同じ個体識別番号を持っていた場合に限り、真のリンクが成立したとみなす。また、本研究では、原データと秘匿処理済データにおけるレコード間の距離を計測するために、ユークリッド距離を使用している。

距離計測型リンケージにおける処理の手順は、次のとおりである。

第１に、完全照合型リンケージと同様に、距離計測型リンケージにおいても、リンクキー変数の中の質的属性を用いてレコードの層化を行う。

第２に、層内の原データと秘匿処理済データの量的属性値の各々を標準化した上で、秘匿処理済データにおける特定のレコードと原データの各レコードとの間のユークリッド距離を計測する。

第３に、秘匿処理済データにおける特定のレコードと、原データにおいて同一の個体識別番号を持つレコードとの間の距離が、原データにおける各レコードとの距離の中で最も短いことが確認され、原データの中に同じ距離を持つレコードが他に存在しない場合を真のリンクとみなし、真のリンクとなるレコード数を算出する。図 12 は、距離計測型リンケージにおける真のリンクの例であり、秘匿処理済データにおいて個体識別番号 001 をそなえたレコードと原データの各レコードとの距離が計測されている。原データにおいて距離が 0.34 となるレコードの個体識別番号は、001 と一致していることから、真のリンクが成立している。このことは、秘匿処理済データにおいて個体識別番号 001 に該当するレコードが、原データにおいて同一の

個体識別番号を持つレコードと対応付けされ得ることを意味している。また、図 13 は真のリンクと判定されない事例を示している。ケース１は、秘匿処理済データの特定のレコードからの距離が原データに含まれるレコードの中で最短ではあるものの、個体識別番号が一致しない場合である。ケース２は、秘匿処理済データの特定のレコードとの距離が原データの中で最短であるレコードが、原データにおいて同じ個体識別番号を有するレコード以外にも複数存在する場合である。

図 12　距離計測型リンケージにおける真のリンクの例

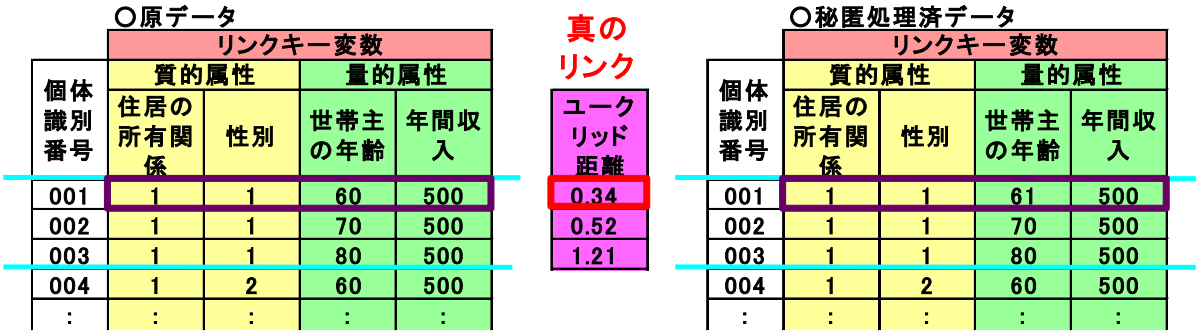
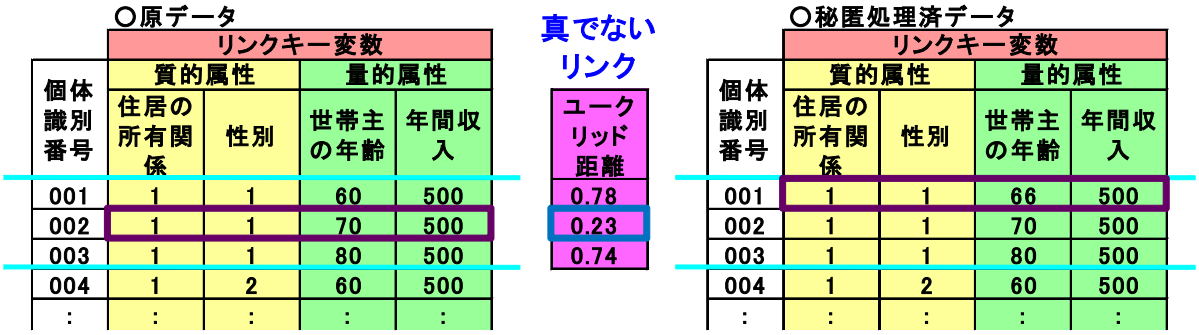
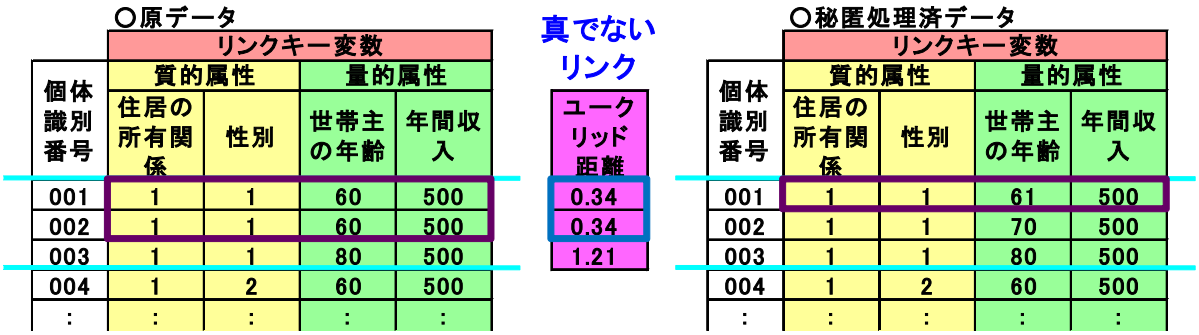


図 13　距離計測型リンケージにおいて真でないリンクの例

ケース１　秘匿処理済データのある特定のレコードに対して、原データの中で最も距離が短いレコードが、個体識別番号については一致しない場合



ケース２　秘匿処理済データの特定のレコードとの距離が原データの中で最短であるレコードが、原データにおいて同じ個体識別番号を有するレコード以外にも複数存在する場合



(3) 完全照合型リンケージと距離計測型リンケージに関する実証研究の結果

表１は、ソートなし、個別ランキング法、第１主成分法、Zスコア総計法の４手法によって作成したデータについて、完全照合型リンケージ及び距離計測型リンケージのそれぞれについて真のリンクとなったレコードの比率を示したものである。表１を見ると、いずれの手法についても、距離計測型リンケージにおける真のリンクの比率のほうが、完全照合型リンケージにおけるそれよりも高くなっていることが注目される。特に、個別ランキング法については、完全照合型リンケージにおいて真のリンクとなる比率が 58.90%であるのに対して、距離計測型リンケージでは、その比率が 98.58%となっている [23]。これは、２種類のリンケージにおける考え方の相違に起因していると思われる。完全照合型リンケージで真のリンクになるということは、原データにおけるレコード群が有する属性値をミクロアグリゲーションによって平均値に置き換えた結果、秘匿処理済データのあるレコードの属性値が、その元となる原データのレコードの属性値と完全に一致していることを表している。そのため、ほぼ同一の属性値群を持つレコード群内で属性値の平均値への置き換えが行われた場合に限り、完全照合型リンケージにおける真のリンクが成立する。よって、ソートなし、第１主成分法及びZスコア総計法のミク

23 リンケージによって真のリンクと判定されたレコードについては、そのレコードがある特定のリンクキー変数に関する分布においてどの分類区分に属するかによって、秘匿性の程度が異なることが考えられる。付図３は、世帯主の年齢に着目し、完全照合型リンケージと距離計測型リンケージにおける真のリンクの分布を全消の原データにおける年齢の分布特性と比較したものである。完全照合型リンケージ及び距離計測型リンケージについては、世帯主の年齢の各分類区分に含まれるレコード数を原データの総数で除することによって、真のリンクの分布を求めている。それによって、原データの分布特性と比較して、真のリンクの分布がどの程度異なるかを把握することができる。付図３を見ると、個別ランキング法では、特に距離計測型リンケージにおける真のリンクの分布が、原データの分布に非常に近似していることが確認できる。すなわち、原データでは、19 歳未満と 90 歳以上の比率が非常に低く、50～59 歳の年齢階層の比率が最も高い数値を示していることから、正規分布の形状を成していると言えるが、個別ランキング法において真のリンクとなったレコードの年齢分布を見ても同様の分布特性が見出される。なお、外れ値(特異値)として位置付けられるレコードの秘匿性に関する評価は、真のリンクの分布特性と関連しているように思われる。外れ値を評価するための秘匿性の評価方法については今後の検討課題にしたいと考えている。

ロアグリゲートデータにおいては同じ属性値を持つレコードが３ずつ存在するために、完全照合型リンケージにおいて真のリンクと判定される比率は 0%となっている。他方、距離計測型リンケージは、秘匿処理済データの特定のレコードに対して、その元になるレコードが原データの中で最短距離に位置するレコードかどうかを判定する方法である。個別ランキング法のように属性値が最も近いレコード同士で平均化の処理を行った場合、秘匿処理済データの特定のレコードと原データにおいて個体識別番号が同一である元のレコードとの間の距離が、原データにおいて個体識別番号が異なる他のレコードとの距離よりも小さくなる可能性が高くなる。こうした点を考慮すると、距離計測型リンケージの場合、個別ランキング法において真のリンクとなる比率は、完全照合型リンケージにおける比率よりも非常に高くなっていることは明らかだと言えよう。

表１　ソートなし、個別ランキング法、第１主成分法、Ｚスコア総計法において、完全照合型及び距離計測型リンケージで真のリンクと判定されたレコードの比率（%）

	ソートなし	個別ランキング法
① 完全照合型	0.00	58.90
② 距離計測型	0.21	98.58

	第1主成分法	Zスコア総計法
① 完全照合型	0.00	0.00
② 距離計測型	0.34	0.43

さらに、距離計測型リンケージによる結果の更なる特徴を探るために、距離計測型リンケージで真のリンクと判定されたレコードにおける平均距離、最短距離及び最長距離を調べたのが表２である。表２を見ると、個別ランキング法では、他の３手法と比較して平均距離や最短距離が非常に小さいことがわかる。このことは、距離計測型リンケージでは、真のリンクと判定される場合、リンケージの対象となるレコード間の距離の長さとは関係なく、真のリンクとなる可能性があることを意味している。

表２　ソートなし、個別ランキング法、第１主成分法、Ｚスコア総計法において真のリンクと判定されたレコードの平均距離、最短距離及び最長距離

	ソートなし	個別ランキング法
平均距離	0.53309	0.00987
最短距離	0.11870	0.00013
最長距離	2.30591	6.45367

	第1主成分法	Zスコア総計法
平均距離	0.77500	0.78671
最短距離	0.14178	0.08675
最長距離	9.35608	9.73848

ところで、表１を見ると、完全照合型リンケージと距離計測型リンケージのいずれについても、個別ランキング法で作成したミクロアグリゲートデータにおける真のリンクの比率が、ソートなし、第１主成分法及びZスコア総計法の３手法で作成したデータのそれと比較して著しく高いことが注目される。このことは、個別ランキング法では、秘匿処理済データの各レコードについて、その元となる原データのレコードに対応付けられる可能性が相対的に高くなっていることを示している。一方、本分析結果は、ミクロアグリゲートデータと原データの比較を行った場合、個別ランキング法で作成したミクロアグリゲートデータの分布特性が、他の３つの手法で作成したミクロアグリゲートデータと比較して、原データの分布と非常に近似していると言うこともできる[24]。個別ランキング法によるミクロアグリゲーションでは、他の３手法のように特定のソートキーに従ってレコードのソートを行うのではなく、属性の各々についてソートを行った上で、レコードのグループ化とグループ内の属性値の平均値への置き換えを行っている。よって、本分析結果は、ミクロアグリゲーションの秘匿性の観点から見ても、個別ランキング法によるミクロアグリゲートデータの分布特性が、他の手法で作成したミクロアグ

[24] 本研究では、ミクロアグリゲーションの有効性を検証するために、ソートなし、個別ランキング法、第１主成分法及びＺスコア総計法の４つの手法を用いて作成したミクロアグリゲートデータについて、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ている。付表１は、原データ及び４手法のミクロアグリゲートデータのそれぞれについて、年齢の平均値及び標準偏差を比較したものである。４手法のミクロアグリゲートデータのいずれも、平均値は原データにおける平均値と等しくなっている。また標準偏差を見ると、個別ランキング法の値が、他の３手法と比較して、原データの値に近くなっている。次の付図４は、４種類のデータにおける年齢 10 歳階級別世帯数分布のヒストグラムである。個別ランキング法のミクロアグリゲートデータにおける年齢の分布特性が原データのそれと近似しているが、他の３手法については、原データの分布と異なるように思われる。さらに、付表２は、原データからの情報量損失の指標として、分布特性の相関係数行列から得られる平均平方誤差を算出しているが、個別ランキング法の平均平方誤差が最も小さく、次いで、ソートなし、第１主成分法、Ｚスコア総計法の順に平均平方誤差が大きくなっていることがわかる。

リゲートデータとは異なる特徴を持っていることを示している[25]。

５　おわりに

本稿では、秘匿性の定量的な評価に関する一手法を提案し、それを全消のミクロアグリゲートデータに適用することによって、秘匿性の評価方法の有効性を検証した。本研究では、完全照合型リンケージ及び距離計測型リンケージという２種類のレコードリンケージ手法を用いて、ミクロアグリゲーションの様々な手法を用いて作成した秘匿処理済データのレコードと原データにおけるレコードとの対応付けを行い、秘匿処理済データにおける秘匿性の程度を検討した。

本研究で提唱したレコードリンケージによる定量的な評価方法は、様々な匿名化技法を用いて作成された秘匿処理済データの秘匿性の程度の比較を指向したものである。こうした視点に立って秘匿性の定量的な評価を試みた研究は、我が国ではこれまで存在しないと考えられることから、本研究の方法的な意義は小さくないと考える。しかし、本研究における秘匿性の評価方法は、試行的な域を出ないことから、秘匿性の定量的な評価方法については、更なる検討が必要ではないかと考えている。具体的には、本研究で提案した距離計測型リンケージについて、ユークリッド距離による計測だけでなく、マハラノビス距離といった他の距離計測法を用いた場合の距離計測型リンケージの結果を比較・検証することが考えられる。また、リンケージ以外の秘匿性の評価方法についても検討を進めることによって、様々な匿名化技法に対する秘匿性の評価方法の適用可能性をより一層追究する必要があると考えている。さらに、本研究では、ミクロアグリゲートデータを対象に秘匿性の定量的な評価を試みたが、今後は、リサンプリングやリコーディング等の各種の匿名化技法を用いて作成した秘匿処理済データに関して、秘匿性の程度の比較を行う予定である。これらについては、今後の検討課題としたい。

参考文献

Bethlehem, J. M., Keller, W. J., Pannekoek, J.(1990) “Disclosure Control of Microdata”, *Journal of American Statistical Association*, Vol. 85, pp.38-45.

Dale, A. and Elliot, M. (2001) “Proposal for 2001 Samples of Anonymized Records: An Assessment of Disclosure Risk”, *Journal of the Royal Statistical Society*, Series A, Vol.164, No.3, pp.427-447.

[25] 本分析の結果をもとにして､匿名化技法として原データに個別ランキング法によるミクロアグリゲーションを適用する場合には、秘匿性の強度をより高めるためにさらなる匿名化技法を追加的に適用することが必要かと思われる。

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

Duncan, G. and Lambert, D.(1989) “The Risk of Disclosure for Microdata” *Journal of Business and Economic Statistics*, Vol.7, pp.207-217.

Elliot, M., Dale, A. (1999) “Scenarios of Attack: the Data Intruder’s Perspective on Statistical Disclosure Risk”, *Netherlands Official Statistics*, Vol.14, pp.6-10.

Elliot, M. (2001) “Disclosure Risk Assessment”, Doyle *et al.*(eds.)*Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, Elsevier Science, Amsterdam, pp.75-90.

Federal Committee on Statistical Methodology(1978)*Statistical Policy Working Paper 2: Report on Statistical Disclosure and Disclosure-Avoidance Techniques*, U.S. Dept. of Commerce, Office of Federal Statistical Policy and Standards, Washington, D.C.

Federal Committee on Statistical Methodology (1994)*Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, U.S. Office of Management and Budget, Office of Information and Regulatory Affairs, Washington, D.C.

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

Fienberg, S. E., Makov, E. U.,(1998)“Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data”, *Journal of Official Statistics*, Vol. 14, No. 4, 1998, pp. 385-397

濱砂敬郎監訳 伊藤伸介訳(1997)「政府の統計調査から得られるミクロデータの秘匿性および索覗」,『ドイツにおけるミクロ統計データの匿名化の条件と開示状況』, 平成 8～10 年度文部省科学研究費補助金 : 重点領域『統計情報活用のニューフロンティア』A02 ミクロデータ利用の社会制度上の問題―資料 : Nr.4, 法政大学日本統計研究所, 16～37 頁

Hawala, S. (2001) “Enhancing the "100,000 rule" on the Variation of the Per Cent of Uniques in a Microdata Sample and the Geographic Area Size Identified on the File”, Paper Presented at Proceedings of the Annual Meeting of the American Statistical Association.

http://www.amstat.org/sections/SRMS/Proceedings/y2001/Proceed/00211.pdf

Hawala, S.(2003) “Microdata Disclosure Protection Research and Experiences at the US Census

Bureau", Paper presented at the Workshop on Microdata, Stockholm Sweden
http://www.census.gov/srd/sdc/microdataprotection.pdf

Herzog, T. N., Scheuren, F. J., Winkler, W. E.(2007) *Data Quality and Record Linkage Techniques*, Springler, New York.

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

Kim, J. J. and Jeong, D. M.(2007) "The Application of the Concept of Uniqueness for Creating Public Use Microdata Files", Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Manchester, United Kingdom, 17-19 December 2007).
http://www.unece.org/stats/documents/2007.12.confidentiality.htm

Longhurst, J. and Vickers, P.(2007) "Microdata Risk Assessment in an NSI Context", Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Manchester, United Kingdom, 17-19 December 2007).
http://www.unece.org/stats/documents/2007.12.confidentiality.htm

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., Walford, N. (1991)"The Case for Sample of Anonymized Records from the 1991 Census", *Journal of the Royal Statistical Society,* Series A, Vol. 154, No.2, pp.305-340.

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

Müller, W., Blien, U., Wirth, H.(1995) "Identification Risks of Micro Data: Evidence from Experimental Studies", *Sociological Methods and Research*, Vol.24, No.2, pp.131-157.

Paass, G.(1988) "Disclosure Risk and Disclosure Avoidance for Microdata", *Journal of Business and Economic Statistics,* Vol.6, No.4, pp.487-500.

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

佐井至道(2000)「予測個体数の期待値に基づく個票データのリスク評価」『統計数理』第 48 巻第 1 号，229～251 頁

Skinner, C. J. (1992) "On Identification Disclosure and Prediction Disclosure for Microdata", *Statistica*

Neerlandica, Vol.46, No.1, pp.21-32.

Spruill, N. (1983) “The Confidentiality and Analytic Usefulness of Masked Business Microdata”, Paper Presented at Proceedings of the Annual Meeting of the American Statistical Association.

http://www.amstat.org/sections/SRMS/Proceedings/papers/1983_114.pdf

Steel, P. M.(2004) “Disclosure Risk Assessment for Microdata” Presented at Proceedings of the Annual Meeting of the American Statistical Association.

http://www.census.gov/srd/sdc/Steel.Disclosure%20Risk%20Assessment%20for%20Microdata.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

Torra, V., Abowd, J. M., Domingo-Ferrer, J.(2006) “Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment”, 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*, Springler, Berlin, pp.233-242.

Winglee, M., Valliant, R., Clark, J., Lim, Y., Weber, M., Strudler, M. (2002) “Assessing Disclosure Protection for the SOI Public Use File”, Paper Presented at Proceedings of the Annual Meeting of the American Statistical Association.

http://www.amstat.org/sections/SRMS/Proceedings/

Zayatz, L. (1991) “Estimation of the Percent of Unique Population Elements on a Microdata File Using the Sample”, Statistical Research Division Report Series, Census/SRD/RR-91/08, Bureau of the Census, Statistical Research Division, Washington, D.C.

http://www.census.gov/srd/papers/pdf/rr91-08.pdf

Zayatz, L. (2002) “SDC in the 2000 U.S. Decennial Census”, Domingo-Ferrer, J.(ed.) *Inference Control in Statistical Databases：From Theory to Practice*, Springler, Berlin, pp.183-202.

Zayatz, L. (2007a) Supporting Document Checklist on Disclosure Potential of Data Version 1.3: Census Bureau Standard Disclosure Review, U.S. Census Bureau.

http://www.census.gov/srd/sdc/S14-1_v1.3_Checklist.doc

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

付録　「第１主成分法」と「Z スコア総計法」によるミクロアグリゲーションについて

(1) 第１主成分法

ソートなし法や個別ランキング法と同様に、すべての質的属性群を用いて超高次元クロス集計を行い、セルに度数１又は２のない組合せを対象とする質的属性の組合せリストを作成した上で(伊藤・磯部・秋山(2008，46 頁))、質的属性の組合せリストを用いて質的属性選択済データを作成した(質的属性選択済データの作成方法については、伊藤・磯部・秋山(2008，45～48 頁)を参照)。次に、質的属性選択済データに含まれる量的属性値の全てについて標準化した上で主成分分析を適用し、算出した第１主成分をキーとして同質属性値レコード群を昇順に並べ替えた。最後に、この並び替えたレコードを３レコードずつグループ化し、レコードに含まれる量的属性値の各々をグループ内の平均値に置き換えた（付図１）。

付図１　第１主成分法における量的属性のミクロアグリゲーションのイメージ

○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（万円）	消費支出（円）
08	109	13	1	1	2	200	100000
21	101	11	1	1	1	200	100000
44	104	16	1	1	2	200	100000
04	106	17	1	2	3	200	100000
15	104	11	1	2	4	300	200000
18	105	17	1	2	4	400	300000
30	106	13	1	2	5	500	400000
34	105	19	1	2	2	600	500000
20	105	15	1	3	4	142	118400
22	108	14	1	3	5	514	364300
26	107	15	1	3	4	655	101000
28	106	19	1	3	4	763	182400
41	109	18	1	3	1	800	743700
43	101	18	1	3	2	905	754200
:	:	:	:	:	:	:	:

⇨

第１主成分スコア
-1.32
-1.32
-1.32
-1.32
-0.73
-0.13
0.47
1.07
-1.43
-0.03
0.40
0.53
2.40
2.73
:

⇨

○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（万円）	消費支出（円）
1	1	200	100000
1	1	200	100000
1	1	200	100000
1	2	400	300000
1	2	400	300000
1	2	400	300000
1	2	400	300000
1	2	400	300000
1	3	437	194567
1	3	437	194567
1	3	437	194567
1	3	823	560100
1	3	823	560100
1	3	823	560100
:	:	:	:

(2) Z スコア総計法

他のミクロアグリゲーションの手法と同様に、最初に、質的属性選択済データを作成した。次に、質的属性選択済データに含まれる量的属性値の全てについて標準化した上で、それらの総計値（Z スコア総計値）を算出した。次に、算出した Z スコアをキーとして同質属性値レコード群を昇順に並べ替えた。そして、この並び替えたレコードを３レコードずつグループ化し、それぞれの量的属性値をグループ内の平均値に置き換えた（付図２）。

付図２　Ｚスコア総計法における量的属性のミクロアグリゲーションのイメージ

○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（万円）	消費支出（円）
08	109	13	1	1	2	200	100000
21	101	11	1	1	1	200	100000
44	104	16	1	1	2	200	100000
04	106	17	1	2	3	200	100000
15	104	11	1	2	4	300	200000
18	105	17	1	2	4	400	300000
30	106	13	1	2	5	500	400000
34	105	19	1	2	2	600	500000
20	105	15	1	3	4	142	118400
22	108	14	1	3	5	514	364300
26	107	15	1	3	4	655	101000
28	106	19	1	3	4	763	182400
41	109	18	1	3	1	800	743700
43	101	18	1	3	2	905	754200
:	:	:	:	:	:	:	:

Ｚスコア年間収入	Ｚスコア消費支出	Ｚスコア総計値
-1.03	-0.84	-1.87
-1.03	-0.84	-1.87
-1.03	-0.84	-1.87
-1.03	-0.84	-1.87
-0.63	-0.40	-1.03
-0.22	0.04	-0.18
0.18	0.49	0.66
0.58	0.93	1.51
-1.26	-0.76	-2.02
0.80	-0.84	-0.04
0.23	0.33	0.56
1.23	-0.48	0.76
1.38	2.01	3.39
1.80	2.06	3.86
:	:	:

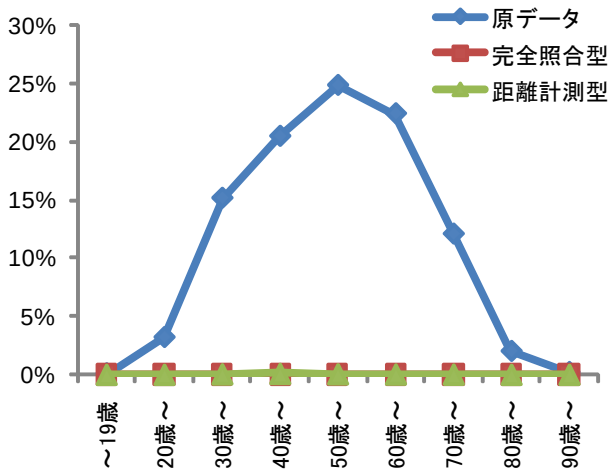
○ミクロアグリゲートデータ

性別	就業・非就業の別	年間収入（万円）	消費支出（円）
1	1	200	100000
1	1	200	100000
1	1	200	100000
1	2	400	300000
1	2	400	300000
1	2	400	300000
1	2	400	300000
1	2	400	300000
1	3	437	194567
1	3	437	194567
1	3	437	194567
1	3	823	560100
1	3	823	560100
1	3	823	560100
:	:	:	:

付図３　ソートなし、個別ランキング法、第１主成分法、Ｚスコア総計法における真のリンクの分布

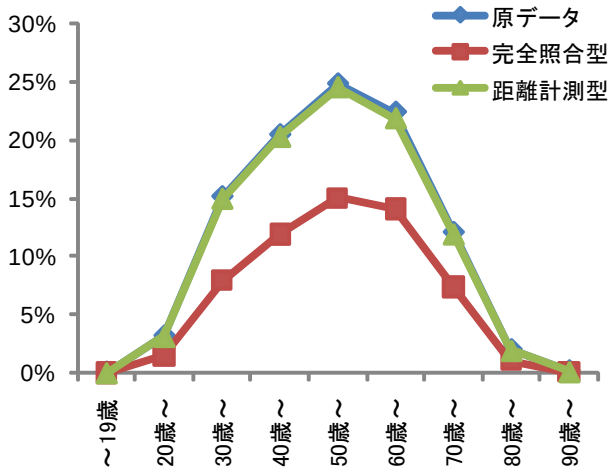
ソートなし

	原データ	完全照合型	距離計測型
～19歳	0.0%	0.0%	0.0%
20歳～	3.1%	0.0%	0.0%
30歳～	15.1%	0.0%	0.0%
40歳～	20.4%	0.0%	0.1%
50歳～	24.8%	0.0%	0.0%
60歳～	22.4%	0.0%	0.0%
70歳～	12.0%	0.0%	0.0%
80歳～	1.9%	0.0%	0.0%
90歳～	0.1%	0.0%	0.0%
総計	100.0%	0.0%	0.2%



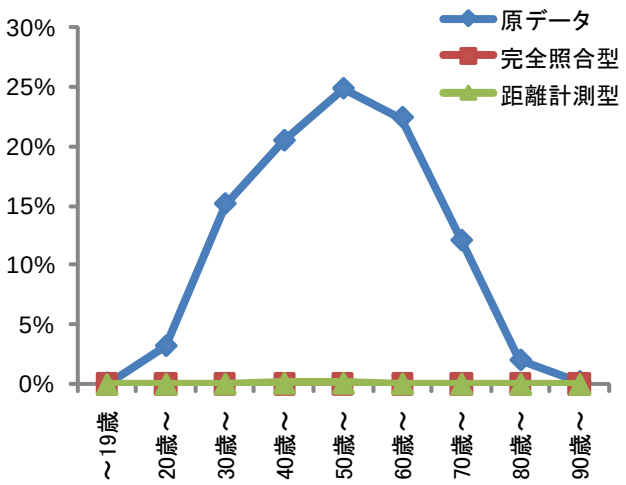
個別ランキング法

	原データ	完全照合型	距離計測型
～19歳	0.0%	0.0%	0.0%
20歳～	3.1%	1.4%	3.1%
30歳～	15.1%	7.9%	14.9%
40歳～	20.4%	11.9%	20.2%
50歳～	24.8%	15.0%	24.5%
60歳～	22.4%	14.1%	21.8%
70歳～	12.0%	7.4%	11.9%
80歳～	1.9%	1.1%	1.9%
90歳～	0.1%	0.0%	0.1%
総計	100.0%	58.9%	98.6%



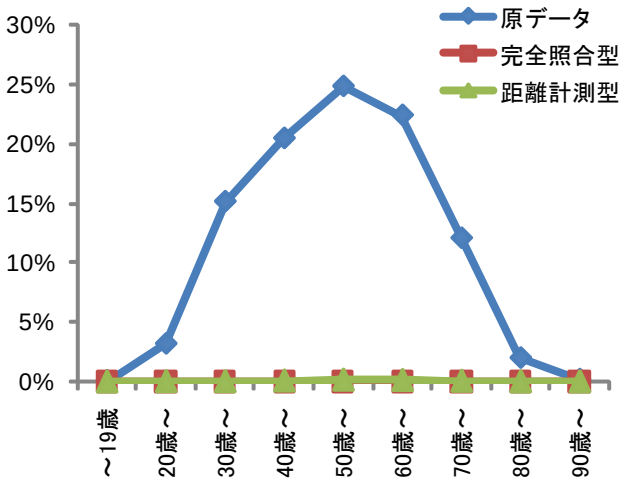
第１主成分法

	原データ	完全照合型	距離計測型
～19歳	0.0%	0.0%	0.0%
20歳～	3.1%	0.0%	0.0%
30歳～	15.1%	0.0%	0.0%
40歳～	20.4%	0.0%	0.1%
50歳～	24.8%	0.0%	0.1%
60歳～	22.4%	0.0%	0.1%
70歳～	12.0%	0.0%	0.0%
80歳～	1.9%	0.0%	0.0%
90歳～	0.1%	0.0%	0.0%
総計	100.0%	0.0%	0.3%



Ｚスコア総計法

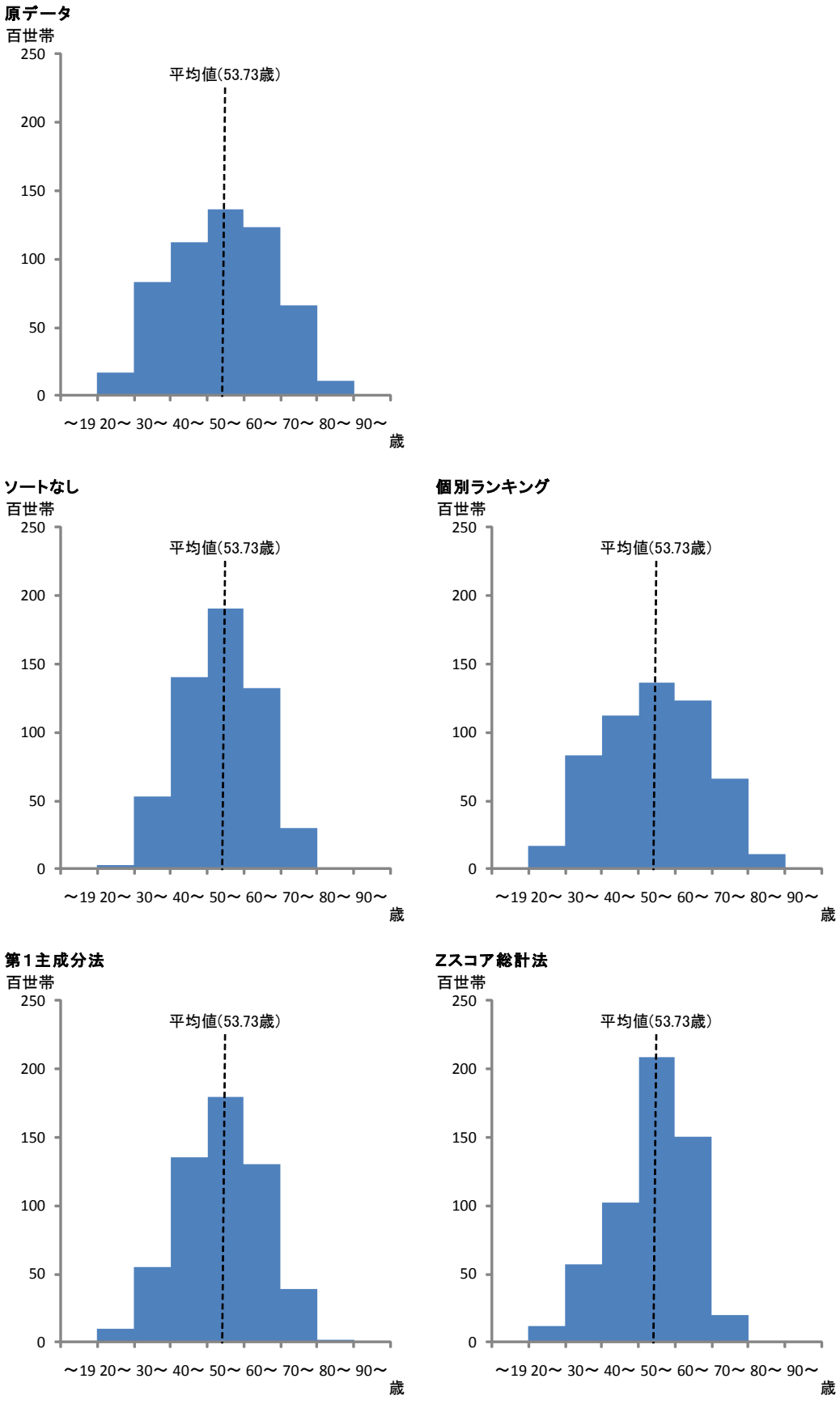
	原データ	完全照合型	距離計測型
～19歳	0.0%	0.0%	0.0%
20歳～	3.1%	0.0%	0.0%
30歳～	15.1%	0.0%	0.0%
40歳～	20.4%	0.0%	0.1%
50歳～	24.8%	0.0%	0.1%
60歳～	22.4%	0.0%	0.1%
70歳～	12.0%	0.0%	0.1%
80歳～	1.9%	0.0%	0.0%
90歳～	0.1%	0.0%	0.0%
総計	100.0%	0.0%	0.4%



付表１　原データ、ソートなし、個別ランキング法、第１主成分法、Z スコア総計法における量的属性の平均値及び標準偏差（世帯主の年齢）

	平均	標準偏差
原データ	53.73	13.82
ソートなし	53.73	10.26
個別ランキング法	53.73	13.82
第1主成分法	53.73	11.21
Zスコア総計法	53.73	10.63

付図４　原データ、ソートなし、個別ランキング法、第１主成分法、Ｚスコア総計法における世帯主の年齢階級別世帯数分布



付表２　原データ、ソートなし、個別ランキング法、第１主成分法、Z スコア総計法における平均平方誤差

	平均平方誤差
ソートなし	0.00414113
個別ランキング法	0.00000065
第1主成分法	0.00773992
Zスコア総計法	0.06876444

製 表 技 術 参 考 資 料 11

平成21年6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

掲載論文を引用する場合は、事前に下記まで連絡してください

情報技術部研究主幹　　TEL ： 03－5273－1368

E-mail ： research@nstac.go.jp