

マイクロデータの匿名化と統計情報の秘匿可能性について

伊藤伸介

1. はじめに
2. わが国の公的統計マイクロデータ提供に関するこれまでの展開と現状
3. わが国の公的統計マイクロデータの作成・提供についての諸課題
4. わが国の統計情報の作成・提供システムに関する将来設計に向けて
5. おわりに

1. はじめに

諸外国では、各国の法制度に基づきながら、個人情報特定化されないように秘匿措置を施した上で、マイクロデータ¹⁾の提供が行われている。提供の際には、マイクロデータに対する秘匿性の程度（強度）と利用者の各種のニーズ（有用性）を考慮し、利用者の利用目的も勘案した上で、粒度（詳細さの程度）が異なる各種のマイクロデータが作成・提供される。それは、学術研究を指向する特定の利用者限定した、個票データ（原データ（raw data, confidential data）から直接的な識別子が削除されたデータ（非識別データ, deidentified data）や匿名化マイクロデータ（個票データに匿名化処理が施されたデータ（本稿では、わが国の統計法における「匿名データ」とは区別して使用する）（anonymized microdata）の作成・提供、および一般には利用目的の制限がなく誰でも入手可能な一般公開型マイクロデータ（個票データに対する匿名化の程度を高めることによって作成される一般に利用可能なマイクロデータ, public use microdata）の作成・公開に大別することができる（伊藤（2018c））。このような海外の状況を踏まえながら、わが国においても、公的統計情報²⁾の一形態であるマイクロデータの利活用の推進を図るための統計法制度が整備されることによって、公的統計（政府統計）のマイクロデータの提供が行われてきた。

本稿の目的は、わが国の公的統計マイクロデータの作成・提供に関して、その現状と課題を示し、課題の解決に向けて、マイクロデータの作成者の視点から論じることである。したがって、本

1) 本稿におけるマイクロデータとは、記入済みの調査個票を含む個別主体について把握される個人情報に基づいて作成された、複数の調査項目（属性、変数）と調査項目の回答値（属性値、変数値）から構成されるレコード群を表すものとする。

2) 本稿では、公的統計の公表された統計表、マイクロデータといった統計作成部局によって作成されている各種の公的統計に関する情報（データ）を「公的統計情報」と定義する。

稿では、統計利用者あるいはマイクロデータの利用者の視点からマイクロデータの特徴を明らかにすることを主として指向していない。統計利用論の観点に立ったマイクロデータの特徴は、吉田(2004)、森(2004)、森(2011)において論じられ、単位、標識、時間、場所という大量(観察)の4要素とマイクロデータの構造との関係が論述されている。また、坂田(2006)は、マイクロデータの利用者の側から、統計作成者側の社会の理論的認識に基づく統計表(系列)と統計利用者側の社会経済の理論に基づく統計解析のための系列(社会経済の実態に関する模像(理論モデル))との間のずれを個票データに関して指摘した上で、個票データの作成プロセスを念頭においた利用上の制約と利用可能性について考察している³⁾。

それに対して、本稿においては、様々なマイクロデータの利活用を想定する統計作成者側の実務の観点に立った場合に、マイクロデータの匿名化の可能性および統計情報⁴⁾の秘匿可能性を追究するために、国際的な動向も踏まえつつ、マイクロデータの作成・提供に関するわが国の課題とその将来的な方向性を議論する。本稿では、最初に、わが国の公的統計マイクロデータの提供に関するこれまでの展開と現状を概括する。つぎに、海外における公的統計マイクロデータの提供状況を参考にした上で、マイクロデータの作成者側の視点から見た場合のわが国の公的統計マイクロデータの作成・提供に関する課題を明らかにする。さらに、わが国におけるマイクロデータを含む統計情報の作成・提供に関する将来設計に向けた試論的な考察を行う。

2. わが国の公的統計マイクロデータ提供に関するこれまでの展開と現状

わが国における公的統計の個票データは、1947年に公布された統計法(昭和22年法律第18号)(以下「旧統計法」と呼称)の制度の下では、「統計目的外使用」という形で、「公益性」に合った学術研究を行う一部の限定された研究者に提供されてきた。2007年に、旧統計法から60年ぶりに改正された統計法(平成19年法律第53号)(以下「2007年統計法」と呼称)が公布され、公的統計の二次的利用に関する規定が明文化された。具体的には、第2条第11項において「調査票情報」と「匿名データ」が新たに条文化されるだけでなく、調査票情報の二次利用に関する条項(第32条)、調査票情報の作成・提供に関する条項(第33条)、オーダーメイド集計(委託に応じる形での利用者の分析目的に対応した集計表の作成)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)が新たに設けられた。2007年統計法におけるこれらの条文の規定およびマイクロデータの利用とデータの管理に関するガイドラインが整備されることによって、公的統計ミ

3) 岩井(2006)も参照されたい。

4) 本稿における「統計情報」は、マイクロデータから作成されたかあるいは高次元の集計表に基づいて編成された集計表(あるいは集計データ)だけでなく、集計表の元となるマイクロデータも含んでいる。

クロデータの利用促進が図られてきた。

匿名データの提供は、内閣府（当時）の統計委員会において匿名データの作成に関する審議を行った上で、2009年4月から開始されたが、当初は、総務省統計局所管の世帯・人口系の統計調査である、就業構造基本調査、社会生活基本調査、全国消費実態調査と住宅・土地統計調査の4調査の匿名データが提供された。これらの匿名データの特徴は、①対象となる統計調査がすべて標本調査であるだけでなく、5年に1度実施される周期調査であること、②複数のタイプではなく、1種類のみ匿名データが作成されたことである。その後、2011年9月に厚生労働省所管の国民生活基礎調査の匿名データの提供が開始された。国民生活基礎調査の場合、利用者の分析目的を踏まえた形で、2種類の匿名データが作成されたことが特徴的である。また、2011年12月から経常調査である労働力調査（基礎調査票）の匿名データが提供されている。さらに、2013年12月には、全数調査である国勢調査の匿名データの提供が開始された。このように、わが国では現在7つの世帯・人口系の統計調査を対象として、統計作成部局が作成・提供を行っている。

2007年統計法に基づき策定された「匿名データの作成・提供に関するガイドライン」（以下「ガイドライン」と略称）（2009年2月制定）によれば、匿名データの利用目的は、①学術研究、②教育、③国際比較統計活用事業の3つである⁵⁾。匿名データの利用にあたっては、わが国の統計法制度の下で、利用申請が必要であるだけでなく、利用する場所の指定や匿名データの適正管理措置が求められる。教育目的での利用に関しては、こうした利用申請や適正管理措置が求められないタイプのマイクロデータに対するニーズがわが国に存在する。そこで、個票データとは統計法制度的に異なるデータである集計表から擬似的なマイクロデータを作成する方法が追究され（山口他（2013））、2011年8月から（独）統計センターで教育用擬似マイクロデータが試行提供された。このデータは、匿名化技法の1つであるマイクロアグリゲーションの方法論（伊藤（2009））も援用する形で、高次元の集計表を作成し、集計表のセルに含まれる平均や標準偏差だけでなく、変数間の相関性も考慮した上で、多変量の正規乱数を生成することによって作成したものである。

教育用擬似マイクロデータは、「教育における利用を想定したもので」、「研究目的には適さないなど不十分な点もあるが、利用目的に制限なく利用することができる」データと位置付けられた（統計データの二次的利用促進に関する研究会（2012、31頁））。その後、第Ⅱ期「公的統計の整備に関する基本的な計画（平成26年3月25日閣議決定）」において、「広く一般に提供可能な「一般用マイクロ

5) 2018年統計法が成立した後に、「ガイドライン」は数回改正がなされたが、現行の「ガイドライン」（2021年8月31日改正）では、匿名データの利用目的は、①学術研究、②教育、③国際比較統計活用事業に加えて、④デジタル社会形成統計活用事業が設定されている。なお、「ガイドライン」によれば、「デジタル社会形成統計活用事業」とは、「デジタル社会形成基本法（令和3年法律第35号）第37条第2項第13号に規定する特定公共分野に関する統計の作成等であって、国民経済の健全な発展又は国民生活の向上に寄与すると認められる行為」（3-4頁）である。

データ（仮称）」の提供に向けた取組の推進が明記された（山口（2019））。そこで、この教育用擬似マイクロデータの作成方法に基づいて、平成21年全国消費実態調査を対象に、公表可能な統計表を作成し、その統計表に含まれるセルごとの平均や標準偏差を用いて、正規乱数を発生させて作成した一般用マイクロデータが、web上で入手可能となっている。この一般用マイクロデータは、利用目的の制限がなく誰でも入手可能であることを勘案すると、一般公開型マイクロデータに該当すると言えよう。

わが国におけるEBPM（=Evidence Based Policy Making、証拠に基づく政策立案）の推進の動きの中で、2017年に創設された統計改革推進会議（議長 菅義偉内閣官房長官（当時））では、公的統計のマイクロデータの作成・提供のあり方が議論された。『統計改革推進会議最終取りまとめ』（平成29年5月19日統計改革推進会議決定）において、「統計等データ」（「統計、統計マイクロデータ及び統計的な利活用を行うために用いられる行政記録情報」が含まれる）の整備・改善の必要性が指摘されたことから、統計法の改正に向けた動きが展開された。そして、2018年に「統計法及び独立行政法人統計センター法の一部を改正する法律」（以下「2018年統計法」と呼称）が公布され、2019年に本格施行された（伊藤（2018c））。

2018年統計法では、公的統計マイクロデータのさらなる利用推進を図ることを指向している。2007年統計法では、第33条第1項第2号に基づいて、高度な公益性⁶⁾を持つ学術研究を行う研究者に対して、調査票情報を提供することが認められていた。2018年統計法における改正では、「第33条の2」が新たに設けられた。この条文では、「相当の公益性を有する統計の作成等として総務省令で定めるものを行う者」が、調査票情報の提供の対象となることが明記された。また、第36条においても、「相当の公益性を有する」者が匿名データの提供の対象になることが新たに記されている。これは、統計作成者の側から見れば、統計実務の観点から、公的統計マイクロデータの作成・提供に関する総務省令やガイドラインに基づきながら、利用目的の「公益性」、および調査票情報と匿名データにおける利用者の範囲を運用のレベルで定めることの必要性を示唆している（伊藤（2018e）⁷⁾）。

2007年統計法で規定された第33条第1項第2号に基づく調査票情報の利用申出の基本的な考え方は、旧統計法の「統計目的外使用」における申請手続きを踏襲していると言える。それは、実

6) 金子（2006）は、公的統計を一般的な「公共財」として理解することは不十分であるとした上で、公的統計データの「公開性」としての特質を考察している。なお、本稿では、マイクロデータの利用者による分析結果が「公共の利益」に資することが、学術研究目的のためのマイクロデータの利用申請の要件であることを述べているが、それは、マイクロデータの提供者と利用者との間の信頼関係が成立するための前提条件であることを指摘しておきたい。

7) 「調査票情報」と「匿名データ」の定義の記載内容は、2018年統計法の改正前の統計法（平成19年法律第53号）と変わっていない。なお、2018年統計法の特徴については、山口（2019）も参照されたい。

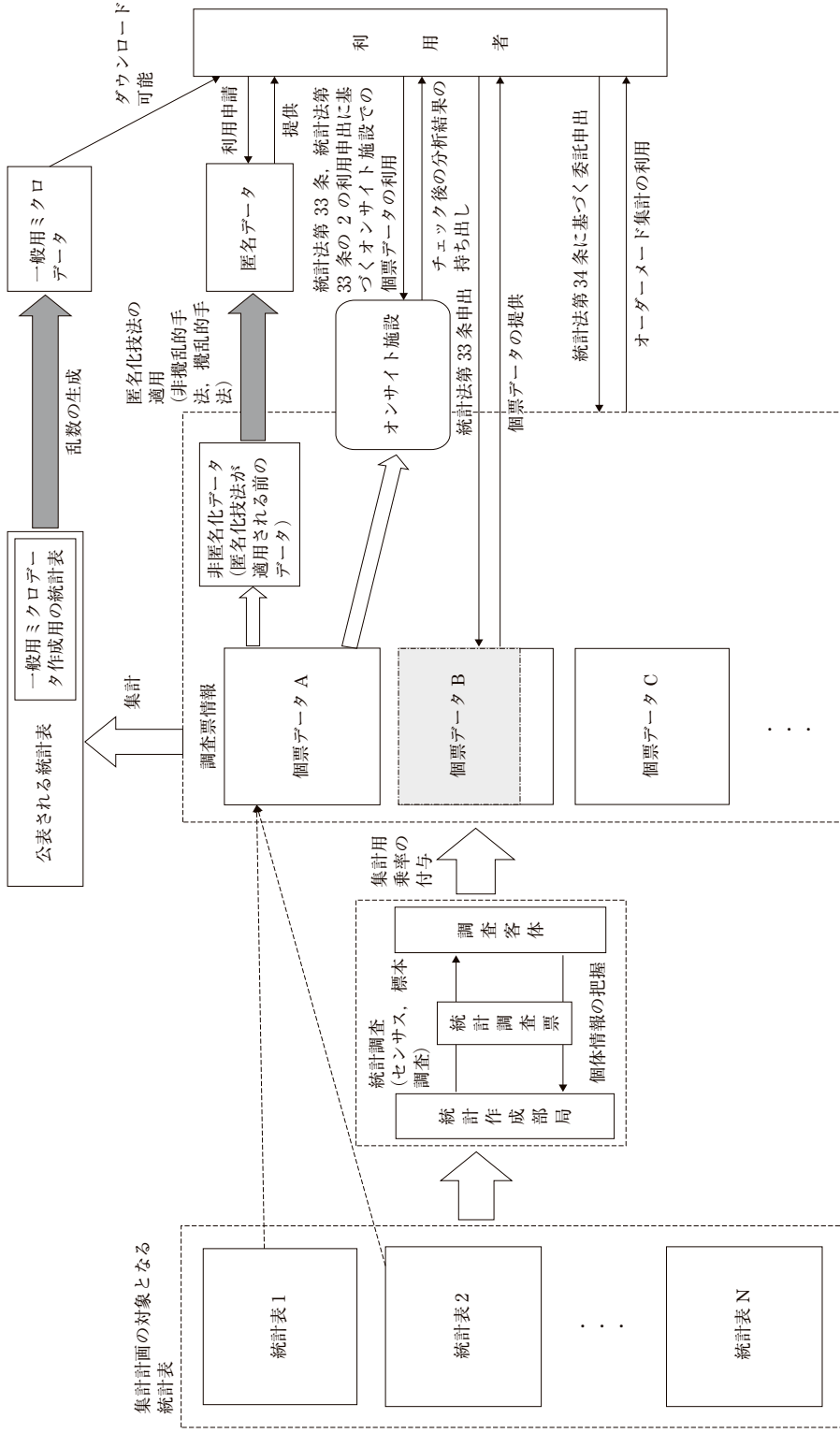
証分析を行う上で最低限必要な調査事項（変数）のみの利用を想定し、統計作成部局が調査票情報の利用目的との適合性の観点から、利用申出において実証分析で作成する集計表と分析に用いるモデルの内容を審査する「事前審査方式」である。これに対して、2018年統計法における公的統計の二次的利用制度の特徴は、「事前審査方式」に基づく調査票情報が入った磁気媒体による提供の仕組みを残しながらも、第33条第1項第2号および第33条の2において、オンサイト施設での調査票情報の利用を前提とした「事後審査方式」が採用されていることである。「事後審査方式」では、オンサイト施設からの分析結果の持ち出しにおいて、分析結果についての安全性を確認するために、分析結果の秘匿性に関するチェックが事後的に行われる。これによって、統計作成部局が利用申出において集計表と分析モデルに関する事前審査を行う必要がなくなり、公的統計マイクロデータを用いた探索的な実証研究が可能になっている。また、オンサイト施設での調査票情報の提供に関する業務は、第37条に基づき独立行政法人統計センターに委託されており、分析結果を持ち出すための安全性の基準に基づく分析結果のチェックも行われている。

3. わが国の公的統計マイクロデータの作成・提供についての諸課題

本節の目的は、海外におけるマイクロデータの作成状況も参考にしながら、わが国の公的統計マイクロデータの作成・提供に関する課題を提示することである。公的統計マイクロデータについては、各国の法制度や利用者のニーズによってデータの提供形態が異なる。個別具体的には、利用目的、利用対象、利用場所、および利用の仕方によって、①個票データの提供サービス（リモートアクセス（大学の研究室のような研究を行う場所で使用するPCから個票データが保管されているサーバにリモートでアクセスできる仕組み）も含む）、②リモートエグゼキューション（プログラムの送付やオンライン上で集計事項を選択する（オンデマンド集計を行う）ことによって分析結果や集計表がリモートで提供される仕組み）、③匿名化マイクロデータの作成・提供、④一般公開型マイクロデータが主に展開されている（伊藤（2018b））。

わが国では、第Ⅲ期「公的統計の整備に関する基本計画」（平成30年3月6日閣議決定、一部変更後令和2年6月2日閣議決定）においても、調査票情報等の利用促進の必要性が指摘されている。前節で論じたわが国における公的統計マイクロデータの作成・提供の現状は、図1の概略図で整理することができる。集計計画の対象となる統計表の元データとしての調査票情報（個票データ）は、磁気媒体による提供かオンサイト施設による調査票情報の提供という形で利用可能である。また、匿名化技法が適用される前の個票データ（非匿名化データ）に各種の匿名化技法を適用することで作成された匿名データの提供も行われている。統計法第34条に基づいて調査票情報から集計するオーダーメイド集計の仕組みも存在する。さらに、統計作成部局によってその作成の元になる統計表が公表された上で、乱数の生成等によって、一般用マイクロデータが作成・公開されて

図 1 わが国の公的統計マイクロデータの作成・提供に関する概略図



出所) 伊藤 (2018b, 98頁) 図 3 を一部修正

いる。

このように見ていくと、わが国においては、①個票データの提供サービス、③匿名化マイクロデータの作成・提供、および④一般公開型マイクロデータが実現していると言うことはできる。しかしながら、個票データの提供サービスについては、海外で展開されているリモートアクセスは実現していない。また、匿名化マイクロデータの作成に関しても、わが国では、一部の統計調査を除き、1種類の匿名データが作成・提供されているが、地域区分の詳細化、年齢の各歳区分での提供に加え、複数のタイプの匿名データの作成が、今後の課題として指摘される⁸⁾。一般用マイクロデータの精度は、公表統計表のセルに含まれる結果数値の粒度に依存するが、現行の作成方法では、生成される正規乱数の元になる数値がセルの中の平均や標準偏差に限定されることから、変数間の関連性を元の個票データにより近似的にするために、作成方法のさらなる検討が求められる。さらに、②リモートエグゼキューションは、現在もわが国では実現に至っていない。

2018年統計法でも、調査票情報と匿名データがそれぞれ、第2条第11項と第12項で定義されている。それによれば、調査票情報は、「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」の形で「記録されている」マイクロデータであるから、統計調査によって収集されない行政記録情報のようなデータは、調査票情報には含まれない。また、匿名データは、「一般の利用に供することを目的として、調査票情報を特定の個人又は法人その他の団体の識別」が「できないように加工した」マイクロデータであるが、調査票情報と匿名データのいずれも、2018年統計法上、安全管理措置（第42条第1項第2号）、秘密保護（法第43条第2項）、罰則（第61条第3号）の対象となっている。したがって、匿名データは、「一般の利用に供することを目的として」作成されているとは言えず、安全管理措置が必要な、学術研究目的等の利用目的に関する制限が課されたデータである。このように、統計作成部局においてデータの安全性が重視されていることか

8) 国勢調査は、現在1種類の匿名データのみが提供されているが、統計委員会による「諮問第44号の答申 国勢調査に係る匿名データの作成について」（2013年2月15日）によれば、今後の課題として、（1）トップコーディングを行う高齢者の年齢の検討、（2）複数の匿名データの作成の可能性に関する検討の2点が指摘されている。とくに（2）については、本答申の中で「地域区分を全国のみとすることや提供データを個人単位とすることなどにより、匿名化措置を緩和できる可能性も考えられることから、今後、利用者のニーズを踏まえて、匿名化措置の内容や組合せ、抽出単位の異なる複数の匿名データの作成の可能性について検討する必要がある」ことが明記されている。したがって、国勢調査において、一般公開型マイクロデータの作成の可能性も含め、複数の匿名データの提供可能性を検討することはわが国における将来的な課題ということができ、そのための具体的な検討が求められる。さらに、就業構造基本調査に関しても、「諮問第87号の答申 就業構造基本調査に係る匿名データの作成について」（2016年4月26日）では、今後の課題として、「年齢区分、産業・職業の分類区分、地域区分の細分化など利用者のニーズが高度化してきているとの指摘もある」ことから、「匿名性の確保に留意しながら、年齢あるいは地域の詳細な情報を含む匿名データの提供の在り方を引き続き検討する必要がある」ことが記載されている。

ら、法制度で規定され、安全性が担保された提供形態のみがわが国で採用されている。個票データのリモートアクセスやリモートエグゼキューションがわが国で実現していないのは、こうした事情が関係していると思われる。さらに、海外で実用化され、人口センサスのマイクロデータ等で適用されてきた各種の攪乱的な手法をわが国の匿名データの作成において積極的に採用するための動機付けも、統計作成部局においては弱かったと言える。そのため、匿名データの有用性を勘案した場合、国勢調査において採用されたスワッピング（data swapping）技法を除けば、攪乱的な手法は、現在提供されている匿名データには適用されなかった。

ところで、2018年統計法の第1条には、統計法の目的に関する条文が明記されている。それによれば、統計法の目的は、「公的統計が国民にとって合理的な意思決定を行うための基盤となる重要な情報であることにかんがみ、公的統計の作成及び提供に関し基本となる事項を定めることにより、公的統計の体系的かつ効率的な整備及びその有用性の確保を図り、もって国民経済の健全な発展及び国民生活の向上に寄与すること」である。2007年統計法からの改正をめぐる議論では、匿名データやオーダーメイド集計の要件が緩和されるのは、法第1条にある「公的統計の体系的かつ効率的な整備及びその有用性の確保」に資する場合となっている。また、法第33条の2の新設に関する議論では、オンサイト施設において「学術研究の発展に資する統計の作成等」を行う場合が想定されている⁹⁾。法第1条で明記されている「国民経済の健全な発展及び国民生活の向上に寄与する」ために「公的統計の体系的かつ効率的な整備及びその有用性の確保」を指向する法制度的・技術的措置の可能性がさらに議論されてよいと考える。また、調査票情報を用いて「学術研究の発展に資する統計の作成等」を行う上で、法第33条で規定されている現行の磁気媒体による提供やオンサイト利用以外にどのような形で安全性が担保できるかを議論することも、二次的利用のさらなる促進をもたらすことから、「国民経済の健全な発展及び国民生活の向上に寄与する」ことが期待できる。

こうした状況を踏まえ、わが国の公的統計マイクロデータの作成・提供に関する将来的な方向性について、以下の3つの課題を指摘したい。

第1の課題は、個票データのアクセスの利便性の観点から、わが国においても、リモートアクセスやリモートエグゼキューションシステムの構築可能性を追究することである。それは具体的には、①大学の研究室等からの個票データのリモートアクセスの可能性の検討、および②プログラム送付型のリモートエグゼキューションやオンデマンド集計システムの展開可能性の追究である。

前者については、イギリス、オランダ、デンマークといったヨーロッパの国々やオーストラリ

9) 総務省「調査票情報等の利用、提供等に関する法制研究会（第3回）」の「資料2-2 調査票情報の二次利用制度の改正案」を参照。

アにおいて、研究室からの個票データのリモートアクセスが実現していることから、こうした海外の事例を参考にした上で、わが国におけるリモートアクセスの可能性を追究することが求められる。中長期的には、デンマークのようなモバイル PC による個票データ（非識別データ、deidentified data）のリモートアクセスの可能性も検討されてよい（伊藤（2020b）¹⁰）。ただし、2018年統計法を踏まえるのであれば、法制度上は、リモートアクセスは高度の公益性を求める第33条第1項第2号に基づく運用に限定されるだろう。また、技術的には、利用者が個票データにリモートでアクセスする上での（指紋等の認証の可能性を含む）認証システムの構築やセキュリティの確保が検討課題になる。

リモートアクセスにおける重要な論点は、①個体に関する情報を保護するための法制度的技術的対応を前提にしつつも、統計作成部局と利用者との間に信頼関係があること、②個票データを用いた実証研究の成果物が公共の利益に該当することである。これらは個票データの提供の前提として求められるが、リモートアクセスでは、磁気媒体による個票データの提供で求められる事前審査やオンサイト施設という物理的な安全管理措置といった制度的・技術的な安全性が担保されない可能性がある。そこで、リモートアクセスでは、上記の2つの論点をどのように法制度あるいは運用面で具体化していくかを方法的に検討する必要がある。その意味で、主としてヨーロッパ諸国で展開され、マイクロデータ提供に関する統計実務の方法論的な基礎になっている「5つの安全モデル（Five Safes Model）」をリモートアクセスに対応する形で位置付けることが考えられる。5つの安全モデルは、安全なプロジェクト（safe projects）、安全な利用者（safe people）、安全なデータ（safe data）、安全な施設（safe settings）と安全な分析結果（safe outputs）の5つの基準から成る個票データのアクセスに関する実践的なモデルである（伊藤（2016））。この「5つの安全モデル」に基づいて、「安全な統計（safe statistics）」が作成される。安全な統計の作成・提供にあたって、5つの安全モデルのすべての要素を満たすことが求められるのではないが、マイクロデータの提供形態によって、統計実務としてどの要素を重視するかが異なってくる。例えば、わが国の匿名データの提供については、調査票情報に各種の匿名化技法を適用していること、および利用申請が必要であることを勘案すると、5つの安全モデルの中の安全な利用者と安全なデータを重視している。それに対して、オンサイト施設における調査票情報の利用は、5つの安全モデルの中の安全な施設と安全な分析結果に対応する形でわが国では実用化されている。

わが国でリモートアクセスを実現するにあたっては、利用申出者が安全なプロジェクトでかつ安全な利用者であるかどうかを審査し、リモートアクセスによるデータの提供の可否を判断するプロセスを新たに設けることが考えられる。そのために、統計作成部局の内部にリモートアクセスの可否に関する審査機関を設置することも検討に値すると思われる。さらに、Eurostatのよう

10) デンマークにおけるリモートアクセスシステムについては、伊藤（2020b）を参照。

な「認証された機関 (authorized institution)」(伊藤 (2018b)) に関する制度をわが国でも議論することによって、利用者が認証された機関に所属しているかどうかを個票データの提供の要件とすることも議論の対象になるだろう。

後者については、プログラム送付型のリモートエグゼキューションやオンデマンド集計システムで用いられる保管用データを議論する必要がある。海外の状況を勘案すると、①個票データ、②匿名化マイクロデータ、③データキューブ (hypercube, 高次元の集計表)¹¹⁾のいずれかが想定される。①から③に示したデータのタイプに応じて、統計法におけるリモートエグゼキューションやオンデマンド集計の位置付けも異なる。①個票データであれば、2018年統計法の第34条 (オーダーメード集計) が運用面で拡大可能かどうか論点になる。②匿名化マイクロデータの場合には、法第35条 (匿名データの作成) を適用することによって、オンデマンド集計用の匿名化マイクロデータの作成が可能かどうかについて、法解釈の議論が必要になるだろう。さらに、③データキューブ (高次元の集計表) に関しては、法第34条を適用した場合に、どこまで粒度の細かいデータキューブがオーダーメード集計によって作成可能か、作成された高次元の集計表が、オンデマンド集計システムに活用することが可能かを模索することが考えられる。

つぎに、リモートエグゼキューションによる分析結果について、安全性を担保した上で、有用な情報をどのように提供するかということが重要である。そのためには、先述の「5つの安全モデル」における安全な分析結果を保証するための法制度的および技術的な手続きが求められる。また、分析結果の安全性のチェックを行うために、ノルウェーのプログラム送付型のリモートエグゼキューションのように、「ルール対応型アプローチ (Rule-Based Approach)」を適用した上で機械的なチェックを行うか (伊藤 (2020a))、あるいは「原則対応型アプローチ (Principles-Based Approach)」に基づいて、人手によるチェックを行うかについての議論も必要になるであろう。

第2の課題は、利用者のニーズを踏まえた匿名化マイクロデータさらには一般公開型マイクロデータの作成の方向性を追究することである。統計作成部局は、データの特性を踏まえ、秘匿性と有用性の定量的な評価指標も参考情報にしながら、適当な匿名化技法を選択する。具体的には、個人や世帯といった個体識別による露見 (identity disclosure) のリスクの低減 (統計的露見制御 (statistical disclosure control = SDC)) や個体のセンシティブな属性値の露見 (attribute disclosure) の

11) データキューブは、ヨーロッパでは hypercube とも呼ばれている。欧州委員会 (European Commission) によって2010年に出された勧告、「人口・住宅センサスの統計データとメタデータの作成計画に関する勧告 (Regulation 2010/519-Programme of the statistical data and of the metadata for population and housing censuses provided for by Regulation 763/2008) によれば、「第2条 定義」の2において、データキューブはつぎのように定義されている。すなわち、「データキューブ (hypercube) は、多次元のクロス集計表を意味しており、データキューブに含まれる集計項目のカテゴリーがそれ以外のあらゆる集計項目のカテゴリーとクロス集計されることによって算出されるセルの度数が含まれている」。

制限を勘案した上で、統計作成部局の統計政策的な判断に基づき、適用可能な匿名化技法が検討される。諸外国では、匿名化技法を選定するにあたって、各種の匿名化マイクロデータの秘匿性と有用性について、定量的に評価するための様々な手法が追究され¹²⁾、定量的な評価結果に基づく秘匿性と有用性に関する閾値（しきい値、threshold）の設定方法についても検討が行われてきた¹³⁾。統計作成部局にとってのこれらの定量的な評価結果は、匿名化技法の採用にあたっての参考情報と位置付けることができる。また、秘匿性と有用性に関する閾値に関しても、統計実務の観点からは、統計作成部局は統計数理に基づいて決定するというよりも、むしろ統計政策的な判断基準を重要視している。

統計作成部局は、統計調査の調査設計、とりわけ調査票に含まれる調査事項を考慮し、調査事項から把握される調査客体の外観識別性や個人情報の特特定化を可能にする外部情報の入手可能性も勘案した上で、レコード削除、トップ（ボトム）・コーディング、リコーディング（区分統合）、サンプリングといった非攪乱的手法、さらには、ノイズ付与、スワッピング、マイクロアグリゲーション（microaggregation）、PRAM（Post RAndomization Methods）等の攪乱的手法を匿名化技法として選択している。統計作成部局が匿名化マイクロデータを作成するにあたっては、世帯・人口系の統計調査に対しては、非攪乱的手法と攪乱的手法のメリットとデメリットを勘案しつつも、非攪乱的手法を優先的に適用している。それは、データの削除やリコーディングのような非攪乱的手法は、統計表の作成における個体レコードの処理や集計事項の分類区分の設定と整合的だからである。ゆえに、非攪乱的手法のみでは匿名化マイクロデータの安全性を確保できない場合に、追加的な手法としてスワッピングといった攪乱的手法が採用されるのが一般的なやり方と言える。統計作成部局側としては、公的統計の統計表の作成プロセスにおいて攪乱的手法が用いられておらず、また、攪乱的手法の適用によって実証分析における分析結果の有効性を担保することができないことが想定されるのであれば、統計作成部局が攪乱的手法を採用することに消極的になることが考えられる。

このような形で各種の匿名化処理が施されたマイクロデータは、諸外国においては、匿名化の強度によって学術研究用ファイル（Scientific Use File=SUF）と一般公開型ファイル（Public Use File=PUF）に類別される。前者が匿名化マイクロデータ、後者が一般公開型マイクロデータにそれぞれ

12) 匿名化マイクロデータに対する有用性と秘匿性の定量的な評価に関する先行研究については、Domigo-Ferrer and Torra (2001), Karr et al. (2006), 伊藤他 (2014) 等を参照されたい。

13) イギリスでは、1991年の人口センサスの匿名化標本データ（SARs）の作成において、閾値ルール（thresholding rule）に基づいて、地域区分、個人・世帯属性の分類区分とサンプリング率の関係についての定式化がなされた（Marsh et al. (1994, p. 43), 伊藤 (2011)）。なお、伊藤 (2018a) では、国勢調査の個票データをもとに、「許容可能な母集団一意の比率」に関する閾値や「地域の人口規模の閾値」を算定した上で、秘匿性と有用性に関する定量的な評価を行っている。

れ位置付けられるが、前者が学術研究を主たる利用目的として限定するのに対して、後者は、欧米では高等教育機関の授業での活用やプログラムのコードの作成のために利用されることを想定している。アメリカやカナダにおいては、各種の匿名化措置が施されて作成されたアメリカ人口センサスの一般公開型マイクロ標本データ（Public Use Microdata Sample）やカナダ統計局が作成・提供している人口センサス等の一般公開型マイクロデータファイル（Public Use Microdata Files）については、学術研究目的よりも教育目的の利用が指向されている（伊藤（2018b））。それに対して、わが国における匿名データの利用は、主として学術研究目的と教育目的の両方を包含しているため、2018年統計法での定義を勘案すると、諸外国の統計作成部局が作成・提供する SUF と PUF のいずれかに符合しているとは言い難い。

そのため、わが国における公的統計の二次的利用のさらなる促進を図る上では、利用者のニーズを勘案した上で、匿名データの役割を明確に方向付ける必要がある。その1つの方向性としては、教育目的での匿名データの利用促進を今後追究していくことが指摘できる。ただし、匿名データは、2018年統計法における安全管理措置、秘密保護と罰則の対象となっているから、教育目的を一層指向する形で匿名データを提供するのであれば、2018年統計法における安全管理措置や罰則に関する法解釈やガイドラインにおける運用面の拡張についての議論が求められる。その上で、秘匿性の強度を高めた一般公開型マイクロデータに相当する教育用 PUF を匿名データの1つとして作成することも、将来的な検討課題になりうるだろう¹⁴⁾。そのためには、マイクロデータに対する攪乱的手法の有効性に関する実証研究の成果を参考にしながら、攪乱的手法の適用可能性を追究することが求められよう。

なお、海外で展開されている合成データ（synthetic data）の方法論をわが国で追究することも考えられる。合成データは、元になるデータから、その分布特性が近似するように属性値を新たに生成することによって作成され、個人情報秘匿性が確保された、マイクロレベルの擬似的なデータと位置付けられる（Templ（2017, p.157））。合成データの方法論は、欠測値を補正するために用いられる多重代入法（multiple imputation）の応用としても展開されてきた（Rubin（1993））。また、合成データの作成においては、回帰モデルといった統計モデルを用いたパラメトリックな手法の適用だけでなく、CART（=classification and regression tree）等を用いたノンパラメトリックな手法の開発も議論されてきた（Reiter（2005）等）。さらに、合成データは、その対象となるすべてのレコードに含まれておらず、欠測値となっている属性群に対して多重代入法を適用して、擬似的に属性値を補完する完全合成データ（fully synthetic dataset）（Rubin（1993））と一部のレコード群において該当するセンシティブな属性に対してのみ補完を行うことによって、擬似的な値が生成される部分合成データ（partially synthetic dataset）（Little（1993））に大別

14) 教育目的に特化した匿名データの作成に関する議論については、小林（2011）も参照されたい。

することができる（Drechsler（2011）、高部（2022））。

欧米諸国では、Eurostat、アメリカセンサス局といった統計作成部局が、欠測値補完やシミュレーションの手法を用いる形で合成データの方法論を適用することによって、PUFのような一般公開型マイクロデータを作成している事例が存在する（伊藤（2018b））。こうした海外の合成データの方法論の適用事例を踏まえて、わが国において一般用マイクロデータの作成のために合成データの方法論を追究することは、議論の対象となりうる。その場合、2018年統計法の下では、個票データではなく、公表されている（高次元の）統計表が合成データの方法論の適用対象となるデータであることが、一般用マイクロデータの必要不可欠な要件になる。そのため、わが国における一般用マイクロデータの作成においては、まずは高次元の統計表の公表可能性を検討した上で、元データの分布特性に近似的になることを指向する形で、合成データの方法論を探究することが求められる。その一方で、教育用あるいはプログラムコード作成のためのテストデータ用の匿名データの作成を目指して、秘匿性を確保するための匿名化技法の1つとして、個票データに含まれる全レコードあるいは一部のレコード群に合成データの方法論を適用し、PUFの形で提供することも考えられる。このように、合成データの方法論をわが国で検討することは、それを適用して作成されたデータをわが国の一般用マイクロデータのような匿名データの外に位置付けることができるか、あるいは2018年統計法ないしは統計法施行規則の下で、匿名データに適用される匿名化技法の1つとみなすことが可能かを議論する意味で、統計法制度上の課題にもなるだろう。

第3の課題は、学術目的のための行政記録情報のさらなる利活用の可能性を追究することである。海外では各国の法制度に基づき、プライバシーの保護に配慮する形で、公的統計マイクロデータだけではなく、行政記録情報の広範な利活用が展開されている。わが国では、2015年9月に施行された改正個人情報保護法における「匿名加工情報」という新たな概念の明記に伴い、行政機関等個人情報保護法においても、「非識別加工情報」が明文化された¹⁵⁾。また、例外的な措置という形で、学術研究のための行政記録情報の利用可能性も想定されている。

わが国でも、学術研究目的や政策目的での行政記録情報の利活用に対する関心が高まっていることから、行政記録情報の利用に関する法制度的な措置と秘密保護についての統計技術的な手法の両面から、その可能性を模索する必要がある。その中には、公的統計の個票データと行政記録情報のリンケージを行うことによって作成されるリンケージデータに対する統計法制度上の対応も含まれる。

15) 個人情報保護法における「匿名加工情報」とは、「特定の個人を識別することができないよう個人情報を加工して得られる個人に関する情報であって、当該個人情報を復元することができないようにしたもの」であり（法第9条）、第9条における個人情報の区分にしたがった匿名化措置が講じられる。なお、公的統計との関連から見た個人情報保護法における匿名化の特徴については、伊藤（2015）および伊藤（2021a）を参照。

2018年統計法は、あくまで調査統計を対象にした法律である。ゆえに、行政記録情報が連結された公的統計の個票データは、2018年統計法の第2条第11項における調査票情報の定義では想定されていないと言える。したがって、公的統計の調査票情報と行政記録情報が連結された個票データの利活用を現行の統計法制度の下で想定しようとするれば、2018年統計法における調査票情報の定義において、行政記録情報も調査票情報に包含できるかどうか、公的統計の二次利用に向けた今後の検討の対象となるように思われる。そのための1つの方向性としては、デンマーク等で適用されている「知る必要性 (need to know)」という概念 (伊藤 (2020b)) を参考にして、統計法の第33条の枠内で、高度な公益性に基づく統計的研究の観点から、公的統計の個票データの中に必要最低限の行政記録情報の属性群を連結可能にする仕組みを模索することではないだろうか。

また、中・長期的には、わが国において公的統計の調査票情報と行政記録情報が連結された個票データの利用がわが国で定着するためには、この連結された個票データを対象にした安全管理措置や個人情報漏えいに関する罰則規定といった個票データの秘密保護に関する法制度的措置についての議論が求められる。そのためには、行政記録情報も対象となる令和3年改正個人情報保護法の適用の範囲や2018年統計法と令和3年改正個人情報保護法との関係を整理する必要も出てくるだろう。

さらに、統計技術的な課題としては、公的統計マイクロデータと行政記録情報が連結される場合のメタデータの整備とその提供可能性が指摘できる。これについては、公的統計マイクロデータに含まれる互いに関連する調査項目群の統制語彙 (controlled vocabulary) の方法論についての整理、および統制語彙の整備に関する議論等、公的統計マイクロデータと行政記録情報の連結を念頭に置いた場合のメタデータの標準化についてさらなる検討が必要になるだろう¹⁶⁾。

4. わが国の統計情報の作成・提供システムに関する将来設計に向けて

本節では、中・長期的に見た公的統計の二次的利用のさらなる課題を追究するために、わが国における公的統計情報の作成・提供に関する将来設計のための試論的な考察を行う。

統計作成部局は、統計調査のプロセスにおいて、調査客体から収集された個体に関する情報を対象に、個人情報の秘密保護について十分考慮した上で、集計計画で設定された統計表を作成し、一般の利用者に対して公表している。公的統計の統計表では、統計表の集計事項として用い

16) 統制語彙とは、「情報検索において、索引語として利用する語を限定し、その意味範囲や使用方法を規定したもの」であって (日本図書館情報学会用語辞典編集委員会 (2020))、統制語彙によって「同義語の存在で検索漏れが生じないように」機械で一意的に判別することが可能になる (伊藤・前田 (2019)、伊藤 (2021b))。

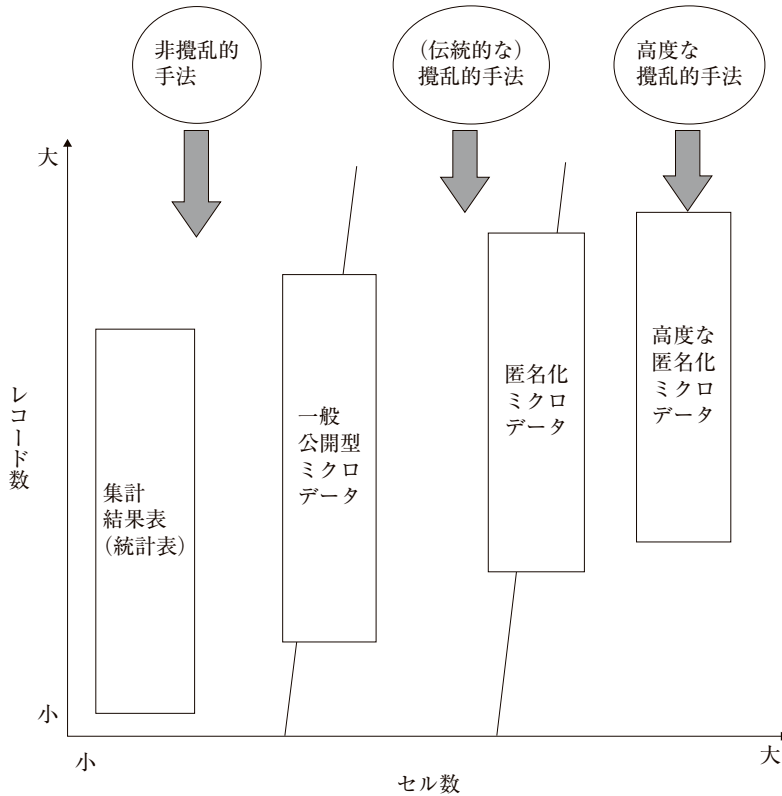
られる属性の数だけでなく、属性の種類や分類区分、セルに含まれる結果数値の情報特性（度数、平均値、比率、重み付きの度数等）の制約を踏まえた形で、統計の単位、標識（変数）、時間、場所の規定を受けた統計数値が獲得される。この統計数値から構成される統計表は、わが国の場合統計法の条文の下ではなく、統計作成部局の責任において個人情報の特特定化が困難なように秘匿処理を施した上で、公表されている（伊藤（2015, 475頁））。

ところで、社会経済の実態における動的な変化を把握するために、利用者が迅速に統計情報を取得することへのニーズが高まっている。このような動きに対して、海外の統計作成部局の中には、利用者のニーズに対してオンデマンドで統計表を作成し、公表するシステムの構築を展開している国々がある（例えば、イギリス国家統計局で進められている「オンデマンド型公表システム（Flexible Dissemination System）」の開発の試みやデンマーク統計局における公的統計の作成等（伊藤（2020b）、伊藤・寺田（2020））。わが国においてそのような方向を模索する場合、統計作成のプロセスにおいて集計計画に基づいて統計表が作成・公表されることを勘案すると、集計計画において、対象となる集計事項をすべて含む「超高次元クロス集計表」（伊藤（2009））を設定することが考えられる。この場合、匿名化技法としてのマイクロアグリゲーションの方法論を適用すると、 k 次元の「超高次元クロス集計表」は、マイクロアグリゲーションが適用されたデータ（以下「マイクロアグリゲートデータ」）と1対1の対応関係になりうる。このようなマイクロアグリゲートデータは、個票データと近似的なデータ構造を保持していることから、統計表とマイクロデータについては、「超高次元クロス集計表」の概念を介して関連付けることが可能である¹⁷⁾。

図2は、統計表に含まれるセル数あるいはマイクロデータにおけるレコード数の大小、および統計表やマイクロデータに適用される匿名化措置との関連から統計表、一般公開型マイクロデータと匿名化マイクロデータを位置付けたものである。統計作成部局は、提供するデータのタイプ（統計表あるいはマイクロデータ）とデータの粒度を勘案した上で、攪乱的手法を含む適切な匿名化技法を選択する。リコーディングやセル削除等の非攪乱的な方法を用いることによって、公表可能な統計表を作成することができる。統計表に含まれるセル数がレコード数と比較して十分に少なければ、統計表という形でのみの提供が可能になる。セル数が増大するほど、度数1となるセルが増加するため、マイクロレベルのデータに近似する。そして、匿名化マイクロデータや一般公開型マイクロ

17) 例えば、年齢や収入といった量的変数であっても、各歳年齢や一円単位の分類区分を想定することによって、質的変数として捉えることも可能である。したがって、例えばマイクロデータに含まれるすべての変数で高次元のクロス集計表を作成した場合に、それが度数1か0の値のみを含むセルから構成される集計表となるのであれば、当該高次元のクロス集計表とマイクロデータは1対1で対応可能となる。なお、ヨーロッパの公的統計において展開されているデータキューブは、秘匿の観点から個人情報特定されないように作成されるが、超高次元クロス集計表の一形態として、データキューブを位置づけることも方法論的には可能である。

図2 統計表のセル数、マイクロデータにおけるレコード数、および適用される匿名化措置から見た統計表やマイクロデータの位置付け



出所) 伊藤伸介・寺田雅之「詳細な地域データにおける秘匿処理の適用可能性について」(2019年度統計関連学会連合大会学会報告(滋賀大学, 2019年9月12日))の発表資料における図を一部修正

データの作成を指向する場合、リコーディングといった非攪乱的な方法のみによる秘匿処理が困難であれば、非攪乱的な方法だけでなくノイズやスワッピング等の(伝統的な)攪乱的手法が適用される。さらに、攪乱的手法の適用によって、情報量損失が大きくなる可能性があることから、情報量損失の増大を回避しつつ、秘匿性を確保するために、情報工学の分野で発展してきた「差分プライバシー(differential privacy)」のような高度な攪乱的手法の有効性を検討することも考えられる。

近年、安全な統計表の公表を保証することを指向した、差分プライバシーの統計実務への適用可能性が注目されている。差分プライバシーとは、「『ある個人のデータを含むデータベースに対する問い合わせ結果が、その個人のデータを含まないデータベースへの問い合わせ結果と区別できないなら、その問い合わせは安全である(個人に関するプライバシーを開示しない)』という考え方によりプライバシーを規定する」基準である(寺田他(2015, 1803頁))¹⁸⁾。アメリカセンサス局で

は、データベース再構築攻撃（database reconstruction attack）（伊藤他（2022）¹⁹）に備えるために、この差分プライバシーの方法論を用いた2020年人口センサスに関する統計表の作成・公表を行っている。そのために、差分プライバシーの適用可能性の検証（階層的な（「Topdown」方式による）ラプラスノイズの付与に関する実証研究²⁰）を行ってきただけでなく、差分プライバシーが適用された検証用のマイクロデータ（「プライバシー保護済マイクロデータファイル（Privacy-Protected Microdata Files=PPMFs）」）の作成・公開も進めてきた（伊藤他（2022））。

わが国の公的統計の実務において、差分プライバシーの方法論に関する議論はなされていない。しかしながら、アメリカセンサス局が、差分プライバシーの方法論を用いて、攪乱的手法を本格的に適用したことについては、わが国における公的統計情報の秘匿処理の将来的な方向性、さらには公開可能なマイクロデータの作成方法を議論する上での1つの参考事例として今後も注視すべきであろう。

ところで、中・長期的な視点に立てば、わが国において、個人に関するデータ（リンケージデータも含まれる）として、公的統計マイクロデータ、行政記録情報と民間の個人情報を位置付けた場合、これらの利活用を将来的にどのように展開するかについてグランドデザインを描くことが求

18) 寺田他（2015）によれば、差分プライバシーの定義は、数理的には以下のように定式化される。

「差分プライバシーにおいて、あるランダム化関数 $Q: D \rightarrow R$ が下式を満たすとき、 Q は ϵ -差分プライバシー（ ϵ -differential privacy）を満たすと定義される（Dwork（2006））。

$$\Pr[Q(D_1) \in S] \leq e^\epsilon \cdot \Pr[Q(D_2) \in S]$$

ここで、 D_1 と D_2 ($D_1, D_2 \in D$) は任意の隣接する（互いにたかだか1レコードしか異なる）データベースであり、 $S (\subseteq R)$ は R の任意の部分集合である。」

差分プライバシーの特徴は、「攻撃者がどのような背景知識を持ち、どのような攻撃手法を適用しようとも、 ϵ -差分プライバシーを満たすメカニズム Q は ϵ の値によって定まる安全性が保証される」ことである（伊藤・寺田（2020, 147頁））。そこで、公的統計への適用においては、適切な ϵ の値によって算出されるラプラスノイズ（平均0の両側指数分布（ラプラス分布）に従う乱数）を統計数値に付与した上で、数値の精度も考慮することが求められる。なお、 ϵ の安全性に関する説明については、例えば寺田（2019）を参照。

19) データベース再構築攻撃とは、複数の集計表を組み合わせることによって、元となる個票データに含まれる個人情報が見えることである（Dinur and Nissim（2003））。このデータベース再構築攻撃を避ける観点から、適切な ϵ を決めた上で、ノイズを入れてクエリを返すという差分プライバシーの方法論が展開されてきた（Dwork（2006））。

20) アメリカセンサス局における「Topdown」アルゴリズムの適用に関する検証方法に関する特徴はつぎの通りである。第1は、作成した統計表の公表によって消費されるプライバシー損失予算（許容されるプライバシー損失の総量、privacy loss budget） ϵ を管理した上で、プライバシーの損失と精度のトレードオフの関係を踏まえて、 ϵ の値を決めることである。第2は、プライバシー損失予算 ϵ を設定し、州、郡、センサストラクト（census tract）、およびセンサスブロック（census block）の各地域のレベルで、プライバシー損失予算の割り当てを行うことである（Garfinkel et al.（2019）、Abowd and Schmutte（2019）、伊藤他（2022））。

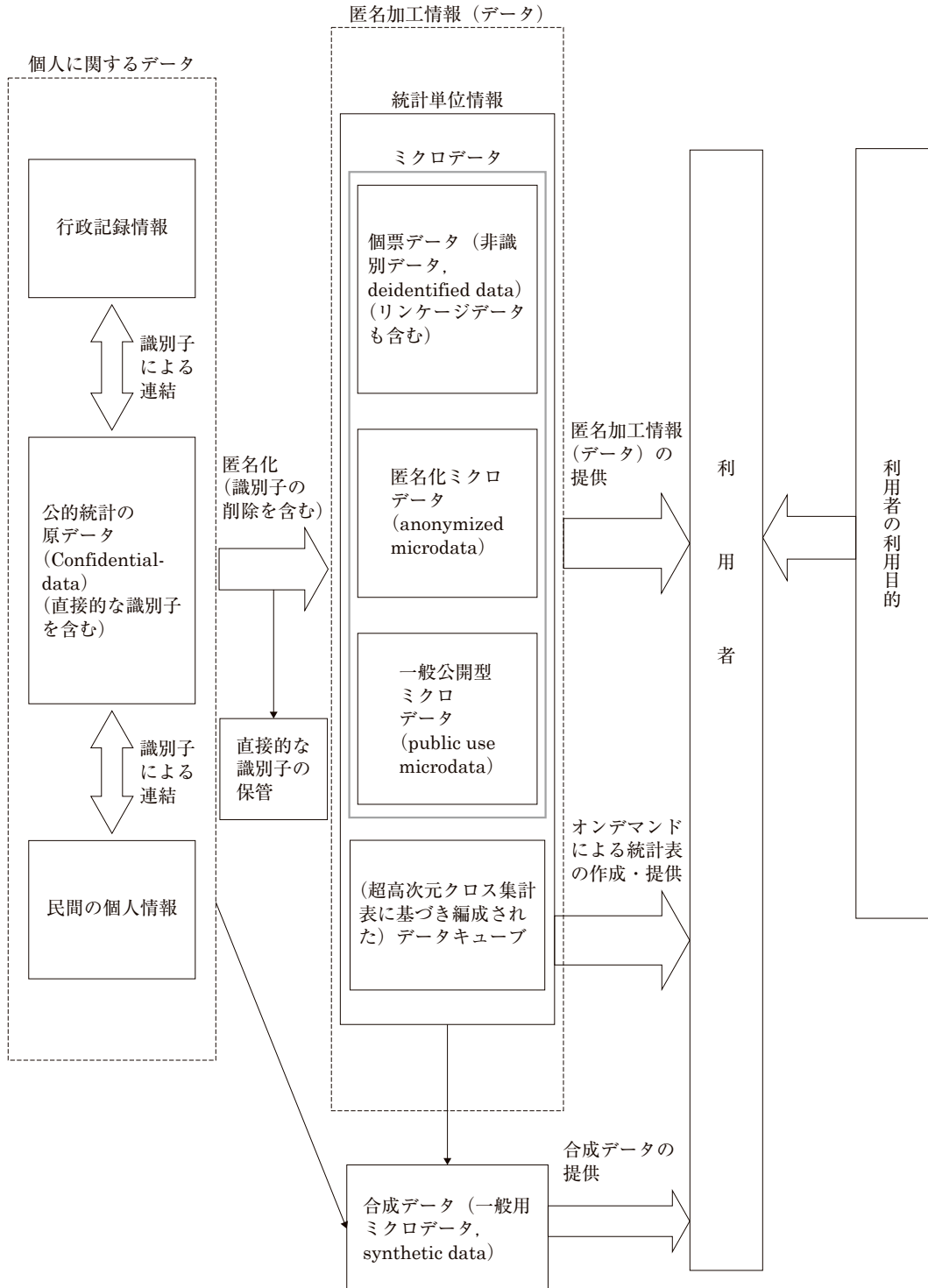
められる。それは、匿名化された個人に関するデータをどのような形で保管し、利用者のニーズに合わせて、どのような形態で提供するかについての方向性を模索することを示唆している。図3は、統計単位情報²¹⁾をキーとなる情報と捉えた上で、匿名化の観点から見た統計単位情報を位置付けたものである。統計作成部局が、公的統計の原データだけでなく、行政記録情報や民間の個人情報についても収集可能な状況を想定する。この場合、公的統計の原データだけでなく、行政記録情報や民間の個人情報に共通の直接的な識別子（名前や住所、ID等）が付与されることが前提となっており、そのための法制度的・統計技術的な対応は求められる。こうした個人に関するデータから直接的な識別子の削除（仮名化（pseudonymization）も該当する）を含む匿名化措置が施されたデータを、本稿では「匿名加工情報（データ）」と定義する²²⁾。そして、図3に示されるように、「社会経済的な集団の構成要素としての社会的個体に関する情報」というデータ特性を備え、「単なる個別のレコードの集合体ではなく、社会経済事象の集団特性が反映された統計単位情報」であるマイクロデータ（伊藤（2018b, 94頁））は、匿名加工情報（データ）の一形態と位置付けられる。利用者にとってマイクロデータという形でアクセスすることが可能な匿名加工情報（データ）は、利用者の利用目的を想定した上で、識別化および特定化の困難性に従って様々なデータの形態を備える。具体的には、公的統計の統計表の元になる個票データ、匿名化マイクロデータさらには一般公開型マイクロデータと情報量が異なる複数のタイプのマイクロデータが包含される。なお、個票データのあり方については今後検討の余地があると思われる。具体的には、匿名化技法を適用する前の非匿名化データの位置付け、さらには公表されない（できない）（高次元の）統計表と調査票情報との関係についての整理は、中・長期的な検討課題となる。

個人に関するデータのアーカイビングの観点から保管すべき統計単位情報の形態を考慮すれ

21) 大屋は、統計単位情報を「調査票に記入された個々の記録すなわち記入済みの調査個票」であり、「調査目的に適合するように設計された一連の統計表の様式に従って集計され、表示あるいは表章される」単位情報と定義している（大屋（1995, 77-78頁））。また、森（2004）は、「統計原単位情報」の概念を提示した上で、「大屋の統計単位情報を基礎に、個別識別情報を含めた調査個票の情報を統計原単位情報、統計に編成される個体情報を統計単位情報と区別」している（岩井（2006, 43頁））。本稿では、マイクロデータの匿名化の観点から、統計作成目的のために編成される「統計単位情報」に焦点を当てて、マイクロデータの匿名化との関係からその特徴を捉え直すことを目指している。そのため、本稿では、匿名加工情報（データ）の作成のために匿名化措置が施される前の元データという意味で、直接的な識別子を備えた「公的統計の原データ」を設定している。

22) 令和3年版改正個人情報保護法において、「匿名加工情報」の対象となる情報は、民間の個人情報と行政記録情報であって、公的統計の調査票情報はその対象に含まれない。本稿では、統計法上、世帯・人口系の統計調査を対象にした場合、非識別データとしての調査票情報には含まれない直接的な識別子を含む公的統計の原データが、民間の個人情報や行政記録情報と連結可能な状況を想定していることから、改正個人情報保護法における「匿名加工情報」と区別するために、公的統計の原データ、民間の個人情報と行政記録情報のすべてを包含する概念として「匿名加工情報（データ）」を設定している。

図3 統計単位情報と匿名加工情報（データ）の位置付け



出所) 伊藤 (2018b, 96頁), 図2を加筆・修正

ば、個票データ、匿名化マイクロデータさらには一般公開型マイクロデータだけが含まれるのではなく、想定される様々なタイプの集計表に対応するために、超高次元クロス集計表に基づいて編成されたデータキューブも、マイクロデータの形式とは異なるが、統計単位情報の派生的な形態として該当しうる（伊藤（2018b））。それに対して、行政記録情報や民間の個人情報から直接的な識別子を仮名化した場合、それは、図3における「匿名加工情報（データ）」に該当するものの、それだけでは統計分析を行うための統計単位情報としての特性を備えていないと考えられる。したがって、マイクロデータのデータ構造の観点から見れば、行政記録情報や民間の個人情報のような個人に関するデータを統計単位情報に転換するためには、統計目的に適合する形で、行政記録情報や民間の個人情報に含まれる個別主体の属性情報に関する欠測値や外れ値への対応が検討課題になるだろう²³⁾。

さらに、利用者にとっての統計表の迅速な提供へのニーズの高さを勘案すると、オンデマンドによる統計の作成・提供システムの可能性も追究する必要がある。先述のように、オンデマンドによる統計表の作成を可能にするための保管用データは、図3で示される統計単位情報の中の個票データ、匿名化マイクロデータあるいは超高次元クロス集計表に基づき編成されるデータキューブである。保管用データの安全管理の状況やオンデマンドで提供される統計表のセルに含まれる結果数値の秘匿処理の方法は、統計作成部局にとっての保管用データの選択に影響を及ぼす。このような統計表に関するオンデマンド集計システムの構築もわが国の公的統計情報の作成・提供に関する中・長期的な課題となる。

なお、図3で表示されている合成データは、個人に関するデータを元データとした上で合成データの方法論を適用した上で作成するデータと、データキューブを対象にして合成データの方法論が適用されたデータ（現行のわが国の統計法制度の下では「一般用マイクロデータ」もそれに位置付けられうる）の2つのタイプの合成データが想定されている。また、図3では、匿名化マイクロデータを作成するための匿名化技法の1つとして、合成データの方法論が適用される場合も示されている。わが国の公的統計情報における合成データの位置付けおよびその有用性の追究は、中・長期的に見た場合の検討課題になるだろう。

23) 森（2004）や森（2011）では、大量の四要素との関連で、「統計原単位情報」に含まれる個別識別情報の活用可能性を論じている。その場合、例えば、個別識別情報である住所が緯度・経度情報に変換することによって、位置情報が個体レコードに付与されうるが、住所は統計調査のために把握されることから、それが変換された位置情報は、基本的に静態的な情報となっている。それに対して、鉄道会社のICカードに記録される情報の中に住所等の識別情報が登録されている場合、ICカードに保存されている個体に関する識別情報は個人情報であるだけでなく、公共交通機関の乗降履歴との組み合わせによって、個人の居住地からの動態的な移動履歴情報として利用することが可能になる。こうした民間の個人情報を統計分析のために用いることも考えられるが、その場合、統計分析を可能にする統計単位情報としての集団特性をどのように考えるかが求められる（伊藤（2021a））。

5. おわりに

本稿では、最初にわが国の公的統計マイクロデータのこれまでの提供に関するこれまでの経緯と現状を述べた。つぎに、海外における公的統計マイクロデータの提供状況を踏まえた上で、わが国における公的統計のマイクロデータ提供の課題を明らかにした。さらには、試みにわが国の公的統計の作成・提供システムにおける将来設計に関する考察を行った。

公的統計情報だけでなく、行政記録情報や民間の個人情報など、様々なタイプの個人に関するデータの利活用の可能性が、欧米諸国で近年追究されている。行政記録情報や民間の個人情報は、それぞれ行政目的や営利目的で収集された情報である。しかしながら、海外では、公的統計の作成や学術研究目的のために、公的統計マイクロデータだけでなく、民間の個人情報や行政記録情報を統合する形で大規模データの利活用が展開されていることから、わが国でも、公的統計情報の整備における行政記録情報や民間の個人情報の活用可能性を模索する必要がある。その場合、統計目的で公的統計マイクロデータと行政記録情報や民間の個人情報がリンクされた個人に関するデータにおいて、社会経済事象を把握する上で求められる集団特性を備えているかを具体的に検討する必要がある。また、各種の個人に関するデータを対象にしたリンクの方法論についてもさらなる検討が求められる。各種の個人に関するデータがリンクされたデータに含まれる属性において、論理的に矛盾する数値（例えば、無業者における現職の就業期間のような属性値）が存在するかどうかをチェックすることも必要になるだろう。こうした個人に関するデータの利用促進を図るためには、データの有用性と安全性の両面から法制度的方策と統計技術的な措置を追究することが求められる。

北欧諸国におけるレジスター型の統計作成システムのように、情報技術の進展によって、中・長期的に、個々人の様々な社会経済的な情報がつながり、個体情報の経常的な更新が可能になるような公的統計の作成がわが国で模索されることも考えられる。そのようなシステムのわが国での実現可能性を具体的に検討しようとするれば、北欧諸国だけでなく、イギリスのような共通の個人識別番号を持たない国々を参考にしながら、個体識別番号のような個体に関する識別情報の設定可能性と個体識別番号を統計目的に対して利用可能にするための制度設計が必要になるであろう。

わが国において公的統計マイクロデータの作成・提供の可能性を追究する中で、公的統計情報だけでなく、民間の個人情報や行政記録情報を含む大規模データを対象にした統計情報のあり方が改めて問われている。その意味でも、様々な情報の中で公的統計情報と他の情報との類型化を行った上で、公的統計マイクロデータの作成・提供目的に適合させる形で、大規模データの活用可能性を模索することは、公的統計情報の整備および公的統計の二次的利用に関する将来展望に向けて課せられた課題である。

謝辞：本稿の旧稿において、関係各位より数多くの貴重なコメントをいただき、本稿を大きく改善することができた。記して謝意を申し上げたい。

参考文献

- 伊藤伸介（2009）「匿名化技法としてのマイクログリゲーションについて」熊本学園大学『経済論集』第15巻第3・4号合併号，197-232頁。
- 伊藤伸介（2011）「わが国におけるマイクロデータの新たな展開可能性について—イギリスにおける地域分析用マイクロデータを例に一」明海大学『経済学論集』第23巻第3号，36-54頁。
- 伊藤伸介（2015）「公的統計データの匿名化について—パーソナルデータの利活用における基盤整備との関連を中心に—」『中央大学経済研究所年報』第46号，457-478頁。
- 伊藤伸介（2016）「諸外国における公的統計マイクロデータの提供の現状とわが国の課題」『中央大学経済研究所年報』第48号，233-249頁。
- 伊藤伸介（2018a）「国勢調査における匿名化マイクロデータの作成可能性」『経済志林』第85巻第2号，241-277頁。
- 伊藤伸介（2018b）「公的統計マイクロデータの利活用における匿名化措置のあり方について」『日本統計学会誌』第47巻第2号，77-101頁。
- 伊藤伸介（2018c）「公的統計マイクロデータの利活用の動向とわが国における課題」『統計』2018年6月号，13-18頁。
- 伊藤伸介（2020a）「諸外国における公的統計と行政記録データの二次利用に関する展開方向」『経済学論纂（中央大学）』第61巻第2号，1-16頁。
- 伊藤伸介（2020b）「デンマークとオランダにおける医療健康データの二次利用について」『日本統計学会誌』第50巻第1号，109-138頁。
- 伊藤伸介（2021a）「大規模データの利活用のあり方と匿名加工に関する一考察」『経済学論纂（中央大学）』第61巻第3・4合併号，235-251頁。
- 伊藤伸介（2021b）「わが国における分野横断的なデータカタログの整備に関する現状と課題」『経済統計学会第64回全国研究大会報告集』，104-107頁。
- 伊藤伸介・村田磨理子・高野正博（2014）「マイクロデータにおける匿名化技法の有効性の検証—全国消費実態調査と家計調査を例に一」『統計研究彙報』第71号，83-124頁。
- 伊藤伸介・前田幸男（2019）「分野横断的なデータカタログの整備に向けて—現状と課題—」『ESTRELA』No. 308，8-14頁。
- 伊藤伸介・寺田雅之（2020）「詳細な地域データにおける秘匿処理の適用可能性について」『日本統計学会誌』第50巻第1号，139-166頁。
- 伊藤伸介・寺田雅之・赤塚裕人・北井宏昌（2022）「海外における公的統計に対する攪乱的手法の新たな取り組み—アメリカセンサス局による差分プライバシーの適用を中心に—」『統計研究彙報』第79号，131-150頁。
- 岩井浩（2006）「コメント：個票データと統計利用」『統計学』第90号，42-44頁。
- 大屋祐雪（1995）『統計情報論』九州大学出版会。
- 金子治平（2006）「統計の作成・公表・利用における公共性」『統計学』第90号，1-14頁。
- 小林良行（2011）「匿名データの教育目的利用に関する一考察」『統計学』第100号，100-105頁。
- 坂田幸繁（2006）「個票データと統計利用」『統計学』第90号，31-42頁。
- 高部勲（2022）「合成データの考え方に基づく公的統計疑似マイクロデータの作成方法の検討」『統計研究彙報』第79号，111-130頁。

- 寺田雅之 (2019) 「差分プライバシーとは何か」『システム／制御／情報』, Vol. 63. No. 2, 58-63頁.
- 寺田雅之・鈴木亮平・山口高康・本郷節之 (2015) 「大規模集計データへの差分プライバシーの適用」『情報処理学会論文誌』, Vol. 56, No. 9, 1801-1816頁.
- 統計改革推進会議 (2017) 『統計改革推進会議最終取りまとめ』.
- 統計データの二次的利用促進に関する研究会 (2012) 『統計データの二次的利用促進に関する研究会 平成23年度報告書』.
- 日本図書館情報学会用語辞典編集委員会 (2020) 『図書館情報学会用語辞典 第5版』丸善出版.
- 森博美 (2004) 「マイクロデータの利用特性と統計利用論」法政大学日本統計研究所『研究所報』 No. 32, 9-38頁.
- 森博美 (2011) 「統計の基本的要素との関連から見た時空間個体データベースの可能性について」『オケージョナル・ペーパー』 No. 24, 1-26頁.
- 山口幸三 (2019) 「改正された統計法と二次的利用の現状と課題」坂田幸繁編著『公的統計情報—その利用と展望』中央大学出版部, 3-21頁.
- 山口幸三・伊藤伸介・秋山裕美 (2013) 「教育用擬似マイクロデータの作成—平成16年全国消費実態調査を例として—」『統計学』104号, 1-15頁.
- 吉田忠 (2004) 「統計利用論からみたマイクロデータ」法政大学日本統計研究所『研究所報』 No. 32, 1-7頁.
- Abowd, J. and I. M. Schmutte (2019) “An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices”, *American Economic Review*, Vol.109, No.1, pp.171-202.
- Dinur, I. and K. Nissim (2003) “Revealing Information while Preserving Privacy” in Proceedings of Twenty-Second ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems (PODS '03), ACM, NEW York, NY, USA, DOI: 10.145/773153.773173.
- Domingo-Ferrer, J. and V. Torra (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.
- Drechsler, J. (2011) *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Springer.
- Dwork, C. (2006) Differential Privacy. ICALP.
- Garfinkel, S., J. M. Abowd and C. Martindale (2019) “Understanding Database Reconstruction Attack in Public Data”, *Communications of the ACM*, Vol. 62, No. 3, pp. 46-53.
- Karr, A. F., C. N. Kohnen, A. Oganian, J. P. Reiter, A. P. Sanil (2006) “A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality”, *The American Statistician*, Vol. 60, No. 3, pp.1-9.
- Little, R. J. A. (1993). “Statistical Analysis of Masked Data”, *Journal of Official Statistics* Vol. 9, pp. 407-426.
- Marsh, C., A. Dale and C. Skinner (1994) “Safe Data versus Safe Settings: Access to Microdata from the British Census”, *International Statistical Review*, Vol. 62, No. 1, pp. 35-53.
- Reiter, J. P. (2005) “Using CART to Generate Partially Synthetic, Public Use Microdata”, *Journal of Official Statistics* Vol. 21, pp. 441-462.
- Rubin, D. B. (1993) “Discussion: Statistical Disclosure Limitation”, *Journal of Official Statistics*, Vol. 9, pp. 462-468.
- Templ, M. (2017) *Statistical Disclosure Control for Microdata: Methods and Applications in R*, Springer International Publishing.