2024年度公開研究会(於: 広島大学高等教育研究開発センター)

公的統計データの利活用における プライバシー保護のあり方について

2024年5月7日

中央大学経済学部

伊藤 伸介

1.はじめに

・欧米諸国では、様々なチャンネルを通じて、公的統計や行政記録情報を対象にしたミクロデータの作成・提供が行われてきた(伊藤(2018), (伊藤(2020a), 伊藤・木村・佐々木(2024)) 。

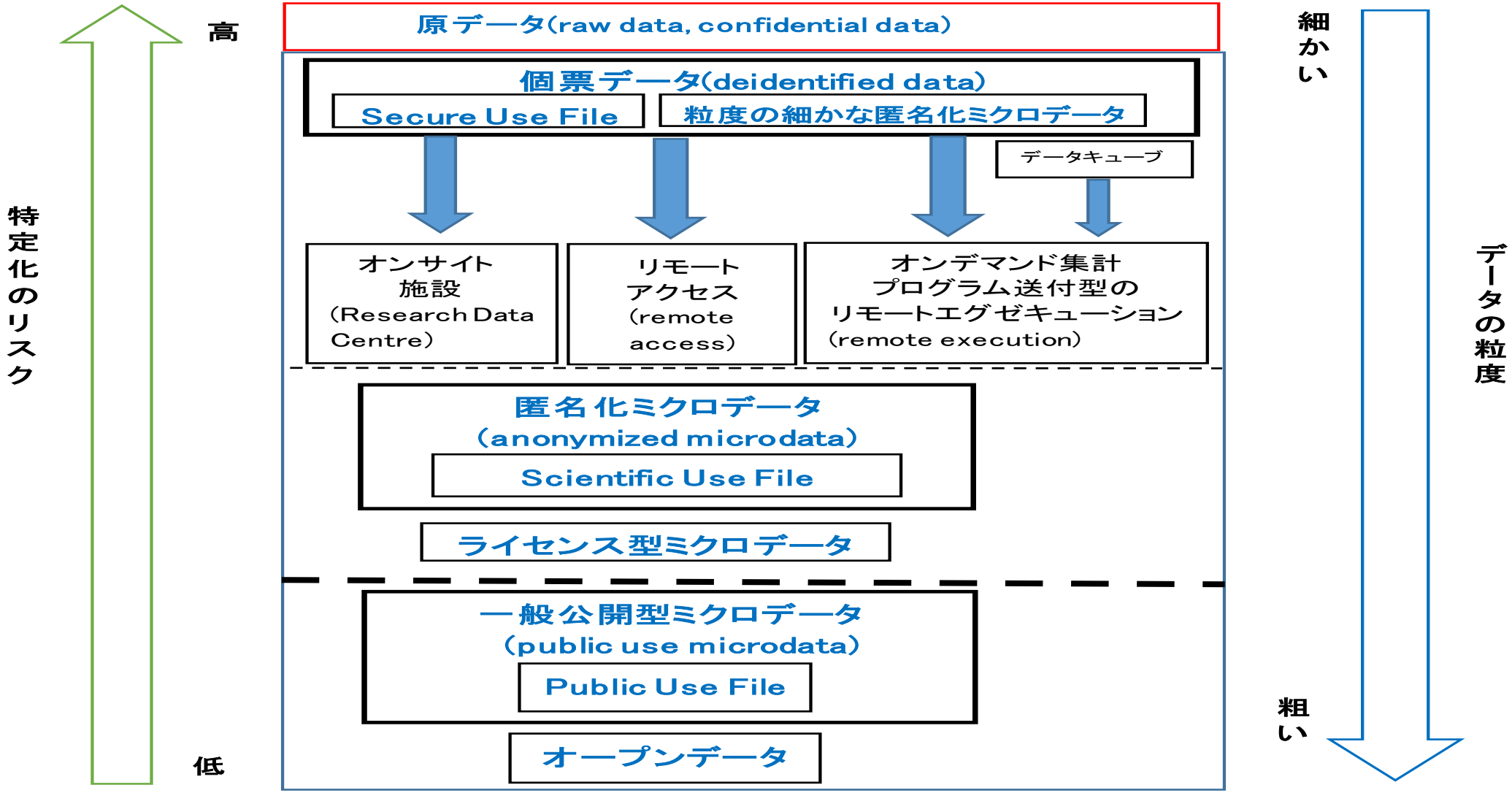
①原データ(raw data, confidential data)から直接的な識別子のみが削除された研究用の個票データ(非識別データ、deidentified data)へのアクセス

②個票データに匿名化技法が適用された学術研究用の匿名化された公的統計ミクロデータ(以下「匿名化ミクロデータ(anonymized microdata)」と呼称)の提供

③秘匿性の強度を高めることによって、誰でも入手可能なオープンデータとして位置づけられる一般公開型ミクロデータ(public use microdata)の作成・提供

・公的統計ミクロデータの利活用を推進するために、攪乱的手法の適用等、各国の統計法制度に基づき、様々な秘匿措置が追究されている(伊藤・寺田(2023))。また、わが国の匿名化された公的統計のミクロデータを対象に、秘匿性や有用性の評価方法についての実証研究もおこなわれてきた(伊藤・星野(2014), 伊藤(2017, 2019)等)。

図1　公的統計の提供形態に関する概略図



出所 伊藤(2020a)，図1を一部修正

公的統計ミクロデータに対する匿名化措置(伊藤(2022))

①個票データへのアクセス

(1)セキュアな環境の元での個票データの利用(**オンサイト施設**の利用、**リモートアクセス**)

(2)個票データの利用の手続きに係る法制度的措置

※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

②利用者のニーズに対する集計表(統計表)・分析結果の提供

(1)**オンデマンド型**の提供システム(リモート集計システム)の構築

(2)プログラム送付型の**リモートエグゼキューション**の適用

(3)集計表(統計表)に対する秘匿処理の適用

(4)集計表・分析結果のチェック(**アウトプットチェック**、output checking)

③学術研究用の匿名化ミクロデータの提供

(1)個票データ(元データ)に対する各種の**匿名化技法(非攪乱的手法、攪乱的手法)**の適用

(2)匿名化ミクロデータに対する**秘匿性と有用性の定量的な評価**

(3)匿名化ミクロデータの利用の手続きに係る法制度的措置

④一般公開用ミクロデータの作成・提供

(1)個票データ(元データ)に対する(攪乱的手法を含む)各種の匿名化技法の適用

(2)一般公開型ミクロデータに対する秘匿性と有用性の定量的な評価

公的統計ミクロデータに対する匿名化技法の適用(伊藤(2022))

①非攪乱的手法(nonperturbative methods)

・レコード削除(suppression)

・トップ(ボトム)・コーディング(top(bottom) coding)

・リコーディング(区分統合)(global recoding)

・サンプリング(あるいはリサンプリング)

②攪乱的手法(perturbative methods)

・ノイズ付与(noise addition)

・スワッピング(data swapping)

・ミクロアグリゲーション(microaggregation)

・PRAM(Post RAndomization Methods)

①統計作成部局は、各種の匿名化技法の適用を検討するにあたって、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するだけでなく、調査事項から把握される調査客体の外観識別性や個体情報の特定化を可能にする外部情報の入手可能性も勘案する。

②統計作成部局が匿名化ミ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する傾向にある。

③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、統計政策的な判断に基づき、適当な匿名化技法を選択する。

参考　秘匿性の定量的な評価方法について(伊藤(2019))

(1)外部情報とミクロデータのマッチング

・個体情報の特定化を行うために用いられると想定される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、個体識別リスクを検証

(2)母集団一意に関する指標の計測

・ミクロデータの中で母集団一意(population unique)に該当するレコード数を計測

(3)特殊な一意の分析(Special Uniques Analysis)

・母集団一意に該当するレコードの中でも、特異な形で存在するレコードは、「特殊な一意」に該当することから、特殊な一意に該当するレコード数に関して検証を行う。

(4)レコードリンケージ(対応付け)によるリスク評価

・元データのレコードと匿名化されたミクロデータのレコードとの間で対応付けが可能かどうかを判定することによってリスク評価を行う方法

(5)クロス集計表によるリスク評価

・原データと匿名化ミクロデータのそれぞれに含まれる質的属性を用いてクロス集計表を作成し、クロス表の中で度数が1となるセルの総数をそれぞれ比較し、度数1となるセル数の変化の程度を比較

参考　有用性の定量的な評価方法について(伊藤(2019))

(1)記述統計量やクロス表における分布特性の比較

・原データと匿名化ミクロデータの間で、平均、分散等の記述統計量やクロス表における分布特性を比較するだけでなく、原データと匿名化ミクロデータに含まれる属性値の差や、分散共分散行列や相関係数行列に見られる分布特性を比較・検証することが考えられる。

⇒属性値間の距離を定義し、その距離の近さを測ることも考えられる。

(2)情報量損失(information loss)に関する指標の評価

・情報量損失に関する指標を定義した上で、原データから匿名化ミクロデータを作成した場合の情報量の提言の程度を定量的に評価する。

※傾向スコア(propensity score)の計測、クラスター分析による検証、経験分布関数における差異の評価等を用いて有用性を定量的に評価する方法も考案されている(Woo et al.(2009))。

・ミクロデータの秘匿性と有用性はトレードオフの関係にあると言える。そのため、秘匿性と有用性の関係を定量的に明らかにし、秘匿性と有用性のバランスを勘案しながら、匿名化措置の適用可能性を模索することが求められる。例えば、**R-Uマップ(R-U Confidentiality Map)**に基づいて匿名化措置の有効性を比較・検討することが考えられる(わが国の公的統計ミクロデータを用いた定量的な研究事例としては、例えば伊藤・星野(2014)等)。

本報告の目的

・海外における公的統計に関する匿名化措置の現状を明らかにすることは、わが国における公的統計ミクロデータの作成・提供における今後の方向性を具体的に議論する上でも有益だと言うことができる。

・本報告では、イギリスやアメリカを中心に、海外における公的統計データの匿名化と利活用の動向について述べる。

・本報告においては、わが国における公的統計ミクロデータの利活用にあたって適用される匿名化措置の最近の動向を踏まえつつ、ミクロデータの利活用とプライバシー保護の今後のあり方についても論じる。

2. 欧米諸国における公的統計ミクロデータの作成・提供

・欧米諸国では、様々なチャンネルを通じて、公的統計ミクロデータの作成・提供が行われてきた。

①個票データ(非識別データ、deidentified data)へのアクセス

(1)Research Data Centre(RDC)におけるセキュアな環境(オンサイト施設)での利用

(2)リモートエグゼキューションにおける個票データの利用

・リモートアクセス(remote access)

・プログラム送付型リモートエグゼキューション

・オンデマンド集計(リモート集計)

②匿名化ミクロデータ(anonymized microdata)」の作成・提供

・学術研究用ファイル(Scientific Use File=**SUF**)の提供

⇒利用申請が必要であり、利用にあたってライセンス(誓約書)が必要

③一般公開型ミクロデータ(public use microdata)の作成・提供

・一般公開型ファイル(Public Use File=**PUF**)という形での提供

⇒利用目的に関する制約はないが、**教育目的やプログラム作成のための利用**が推奨されている。

利用登録が必要な場合もあり

※近年において一般公開型ミクロデータは、合成データ(synthetic data)の作成の方向と符合している。

合成データ・・・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，個体情報の秘匿性が確保された、ミクロレベルの擬似的なデータ(Templ(2017, p.157))。

表1　海外の主な統計作成部局におけるミクロデータの提供形態

データの種類/提供形態 各国の統計作成部局	個票データ(非識別データ)のオンサイト施設による提供	個票データ(非識別データ)の磁気媒体による提供	個票データ(非識別データ)のリモートアクセスによる提供	プログラム送付型のリモートエグゼキューションによる提供	オンデマンドシステムによる提供	匿名化ミクロデータの提供	一般公開型ミクロデータによる提供
Eurostat	○		○			○	○
イギリス国家統計局	○		○		○	○(UKDSから提供)	○
ドイツ連邦統計局	○		●(SUFとして提供)	○	○	○	○
フランスINSEE	○(ドイツの個票データ)		○			○	○
オランダ統計局			○	△	○	○	○
デンマーク統計局			○		○		
フィンランド統計局	△		○			○	
アメリカセンサス局	○		○		○		○
カナダ統計局	○		●	○	○		○
オーストラリア統計局			○		○	○	
ニュージーランド統計局	○					○	
ノルウェー統計局	○	○		○			○(NSDから提供)
総務省統計局	○	○	●			○	○

注1：UKDSとはイギリスエセックス大学のU.K. Data Service、NSDとはNorwegian Centre for Research Dataの略である。

○・・・運営されている。●・・・計画中である。△・・・運営中であるが、活動を休止している(2020年2月時点(一部を除く))。

注2 伊藤・木村・佐々木(2024)の表1、およびBrenzel　et al.(2023)、Bujnowska et al.(2023)、Cramswick（2019）に基づいて、一部加筆・修正(2024年1月時点)

(1)イギリスにおけるミクロデータの提供状況(伊藤(2020), 伊藤(2022), 伊藤・寺田(2023), 伊藤・木村・佐々木(2024))

イギリス国家統計局(ONS)におけるミクロデータ提供の現状

1) 個票データの提供サービス

・Secure Research Service(SRS)が管理・運営するオンサイト施設による個票データの提供

・SRSやUK Data Serviceのthe Secure Lab等が管理するリモートアクセスシステムによる個票データの提供

⇒公的統計の個票データだけでなく、行政記録情報、公的統計ミクロデータとのリンケージされた行政記録情報、行政記録情報同士でリンケージされたデータの提供が可能

→2018年の**Digital Economy Act 2017**の施行とADR UKの支援による行政記録情報のリンケージサービス

・ADR UK: ESRC(=Economic Social Research Council)による行政記録情報の利活用の推進に関するプロジェクトであり、ADRN(=Administrative Data Research Network)の後継プロジェクト

2) 匿名化ミクロデータの提供

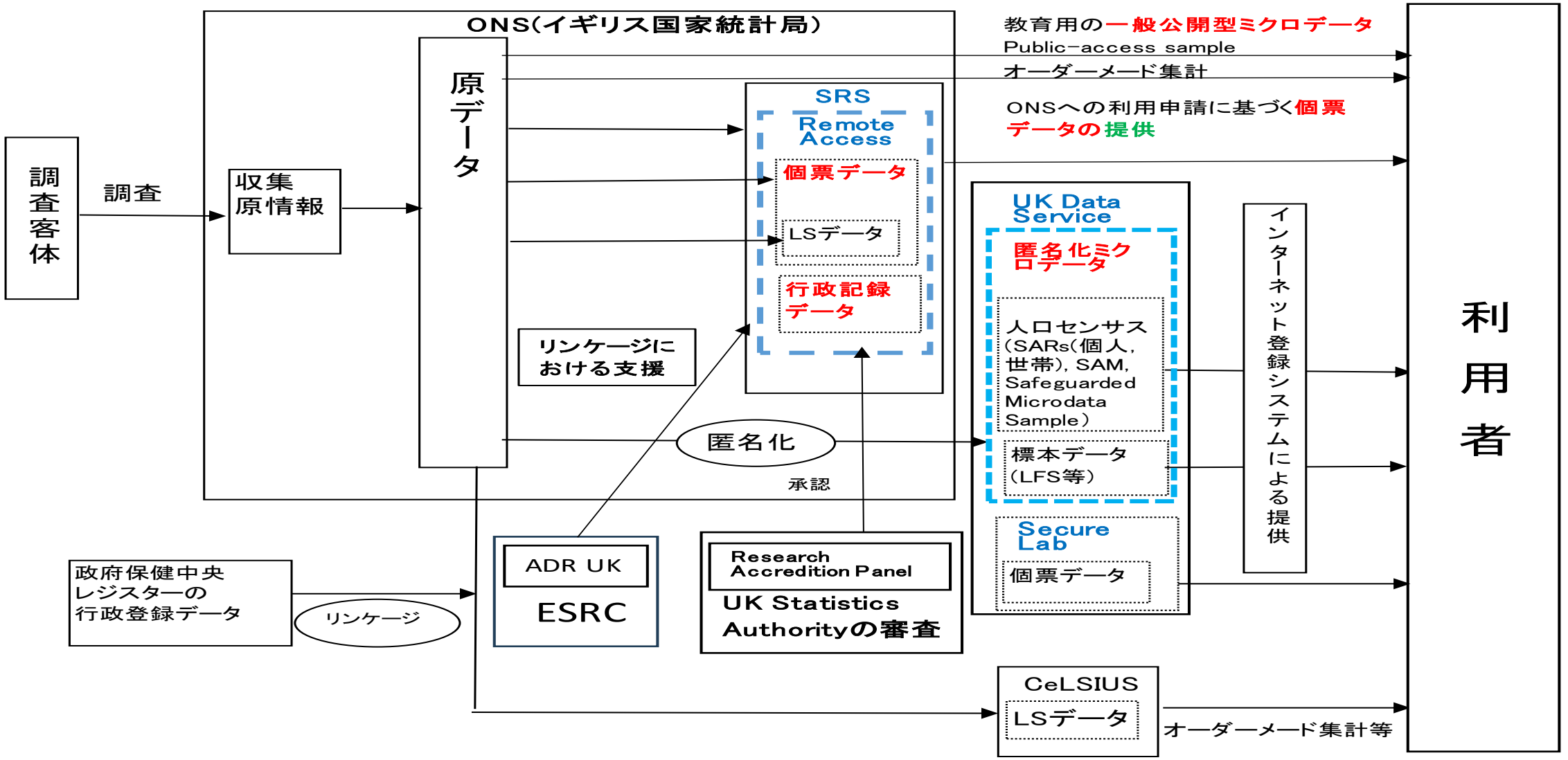
⇒U.K. Data Serviceにおけるライセンス(End User Licence)を必要とするタイプのミクロデータ(「ライセンス型ミクロデータ」)の提供

3) 一般公開型ミクロデータの公開

Ex. 人口センサスの教育用ミクロデータの作成・提供

※5つの安全モデル(Five Safes Model)によるデータ提供の枠組み(伊藤(2016b), 伊藤(2020b))

図2　イギリスにおけるミクロデータの提供形態に関する概略図



注 SARs: Samples of Anonymised Records,　SAM: Small Area Microdata,　CeLSIUS: Centre for Longitudinal Study Information and User Support

出所 伊藤・木村・佐々木(2024)，図1

参考　イギリス統計法(The Statistics and Registration Service Act 2007)(2016a)

第39条　個人情報の秘密保護

(中略)

(2)この条文における「個人情報(personal information)」とは、(法人を含む)特定の個人と関連付けられ、およびその個人を識別する情報を意味する。

(中略)

(4)条文(1)は以下の開示には適用されない。

(中略)

(i)承認された研究者(approved researcher)によって行われる場合

(5)条文(4)(i)の目的に関して、「承認された研究者」とは、統計研究の目的のために、統計委員会(Statistics Board)によって保有される個人情報を取得することについて委員会が容認した者を意味している。

(以下省略)

参考　Digital Economy Act 2017について(2020a)

・イギリスでは、2018にDigital Economy Actが施行

↓

研究目的のために、あらゆる統計データ、行政記録データ、政府によって(潜在的に)生み出されるビックデータの利用(リンケージされたデータの利用も含む)の可能性が指向されている(民間企業が所有するパーソナルデータについては対象外)

第5章　研究目的のための共有

第64条　研究目的のための情報の開示(disclosure)

第1項　公的な当局(public authority)によって保有される、公的当局と関連している情報は、実施されている研究の目的のために、他の個人に開示される。

(以下省略)

※Digital Economy Act(DEA)に基づいて行政記録情報を利用するにあたっては、認可された研究者(accredited researcher) になるための資格取得が求められる。

⇒ONSとしては、利用者が「承認された研究者」よりも「認可された研究者」の資格を取得することが通常の考え方となっている。

・行政記録情報の大部分については、DEAに基づいて学術研究のための利用が可能になっている。

5つの安全モデルについて(Five Safes model)(伊藤(2016b, 2018), 伊藤・木村・佐々木(2024))

＊以下の5つの基準から成るミクロデータの提供に関する実践的なモデル

①安全なプロジェクト(safe projects)

→妥当な統計目的のために個票データのアクセスが行われること。

②安全な利用者(safe people)

→研究者は個票データを適切に利用し、利用手続きに従うことについて、信頼されていること。

③安全なデータ(safe data)

→データそれ自体が、個体情報が露見しないデータであること。

④安全な施設(safe settings)

→個票データのアクセスに関する技術的な管理措置によって、容認されていないデータの移動が回避可能なこと。

⑤安全な分析結果(safe outputs)

→統計分析の結果に個体情報が露見されるような結果が含まれないこと。

※分析結果に関する最終成果物については、担当者がチェックを行った上で公表することが可能

↓

安全な利用(safe use)

5つの安全モデルの考え方によれば、ひとつ1つの基準を組み合わせて、「安全な統計(safe statistics)」となるレベルに達するまで安全性の強度を高めることによって、安全な統計の作成・提供が可能になる。

⇒統計作成部局が個票データを提供する上では、安全なプロジェクト、安全な利用者、安全な施設と安全な分析結果を重要視する。これらの4つの安全性を確保した上で、安全なデータについては、個票データは、個体に関する直接的な識別子の削除のみが適用された非識別データが、オンサイト施設やリモートアクセスのような安全な施設の中で可能になる。

攻撃者のシナリオに基づく公的統計の匿名化ミクロデータの作成・提供(伊藤・寺田(2023))

・攻撃者に関してどのようなシナリオを想定するかによって、また、どのような利用目的を想定して匿名化されたミクロデータを作成するかによって、匿名化の対象となるキー変数(あるいは準識別子)や適用される匿名化技法も異なる。

・ONSが想定している個体情報の特定に関するシナリオ

シナリオ1　公開されているデータセットを利用すること

シナリオ2　特異な属性の組み合わせを有している等、悪意がないような形で個体が偶発的に特定されること(偶発的な個体特定(spontaneous recognition))

偶発的な個体特定:珍しい属性の組み合わせを持つ個体が、データの利用者によって、偶発的に母集団の中で特定されること(Duncan et al.(2011))

③シナリオ3　個人に関する私的な情報を持っており、詮索好きの隣人(nosy neighbour)がいること

→<u>ONSが特に重視しているのは、シナリオ1とシナリオ2</u>

・ONSは、社会調査のミクロデータに関する露見制御のためのガイダンス―事例研究(GSS/GSR Disclosure Control Guidance for Microdata Produced from Social Surveys – Case Studies)」では、ミクロデータに対して匿名化技法を適用した事例が紹介されている。

⇒例えば、資産調査(The Wealth and Assets Survey)のライセンス型ミクロデータを作成するために、①世帯人数10人以上の世帯の削除、②80歳以上の年齢のトップコーディング、③家計資産に関する変数に対する秘匿処理が適用された。

例　シナリオ2—偶発的な個体特定

攻撃者(intruder)は、公開されている情報によって、特定のミクロデータの中にある個人や企業が存在することを偶発的に把握する可能性がある。

⇒回答者が特異な特性を有している、あるいは特定の個人が一般に知られていたり、ある企業が攻撃者によって認識されていたりする場合

→攻撃者が悪意を持って個体の特定化を行わなかったとしても、偶発的な個体特定が発生する可能性がある。

⇒意図しない形での攻撃者

※攻撃者が使用する変数の例

個人の場合：名前、年齢、性別、結婚状態、所得、職業、住所、民族グループ

企業の場合：産業、所在地

参考　イギリス国家統計局における世帯・人口系の匿名化ミクロデータ作成のステップ

1. 選択された変数の組み合わせに関する集計表を作成し、度数の小さなセルの分布状況を探索的に確認する。

2. データの中に一意あるいは特異な組み合わせがあった場合には、対象となる変数やレコードに対して匿名化措置を施す(変数やレコードの削除、リコーディング、**スワッピング**等)。

3. 秘匿処置済のミクロデータに対して、再度選択された変数の組み合わせに関する集計表を作成する。露見(disclosure)の可能性があるセルがなければ、次のステップに進むが、露見のリスクがある場合には、さらなる匿名化措置を施す。

4. 侵入者によるマッチングの検証(intruder testing)を実行する。数多くのマッチングの成功事例が確認される場合には、ステップ2からステップ4を繰り返す必要がある。

5. 要求されるライセンス(オープンライセンスか制限付きのライセンスか)に基づいて、データの公開(あるいは提供)を行う。

2021年人口センサスにおけるcell key methodとtargeted data swappingの実用化の追究

・イギリスの人口センサスでは、学術研究用の匿名化ミクロデータの作成にあたって、元データとなる個票データにランダム・スワッピング(random data swapping)や**ターゲット・スワッピング(targeted data swapping)**が採用されてきた。

⇒教育用の公開型ミクロデータの作成においても、**スワッピングを適用**したことが指摘される(伊藤(2020a))。

・ヨーロッパ諸国においては、オンデマンド集計システムに対する社会的な関心が高まっている。

⇒イギリスやドイツでは、オンデマンド集計システムの開発にあたって、オーストラリア統計局のTableBuilderを参考にして、集計表の各セルにランダムなノイズを付与するcell key methodの実用化に向けた技術的な検討がなされてきた。

・欧州委員会(European Commission)は、2021年人口センサスデータの作成・公表において適用すべき攪乱的手法として、ターゲット・スワッピングとcell key methodを推奨している。

図3　Cell Key Methodの適用のイメージ

ミクロデータの例

ID	性別	年齢グループ	学歴区分	n	record key
1	M	A	中卒	1	0.346
2	F	B	高卒	1	0.684
3	F	B	高卒	1	0.959
4	F	C	短大卒	1	0.629
5	M	B	短大卒	1	0.866
6	F	C	大卒	1	0.363
7	M	A	大卒	1	0.914
8	M	A	高卒	1	0.696
9	M	B	大卒	1	0.535
10	F	B	大卒	1	0.685
11	F	B	短大卒	1	0.034
12	M	B	高卒	1	0.337
13	F	B	中卒	1	0.112
14	F	A	短大卒	1	0.565
15	M	A	大卒	1	0.01

性別　M：男性　F：女性

年齢グループ　A：15～39歳　B：40～59歳　C：60歳以上

T：総計

性別と年齢グループによる集計

性別	年齢グループ	総数	cell key
M	A	4	0.97
M	B	3	0.74
M	B	0	0
F	C	1	0.57
F	B	5	0.47
F	C	2	0.99

Cell Keyに基づくノイズの選定

セルキー

セル度数	1	2	3	…	47	48	…	71	72	73	74	…	99
0		+1											
1			+1										
2													
3	-1										+1		
4+			-1		0	-1					+1		

集計表においてノイズが付与された総数の算出

性別	年齢グループ	総数	セルキー	ノイズ	ノイズつき総数
M	A	4	0.97	2	6
M	B	3	0.74	1	4
M	B	0	0	0	0
F	C	1	0.57	1	2
F	B	5	0.47	0	5
F	C	2	0.99	2	4

cell key methodの手順

①攪乱のルールを定める(ノイズ関数の設定)

②一様分布の乱数であるrecord keyを各レコードに追加する

③原データの集計を行う

④原データの集計と同時にrecord keyも集計することによって、cell keyの数値も算出する。

⑤ノイズ関数（ONSの場合には、look up table(p tableとも呼ばれる)）を用いて、ノイズの値を確定させ、元の集計値にノイズを付与する。

出所　伊藤他(2022)

・ONSが開発した「オンデマンド型公表システム(Flexible Dissemination System)」において、**cell key methodの適用可能性が追究されてきた**(Office for National Statistics(2017))

⇒データの有用性を重視する形での攪乱(“a light touch cell key perturbation”)が指摘されている。

・ONSは、2021年人口センサスに関して、「Create a custom dataset」というwebサイトで、オンデマンドによる多次元集計表の提供を開始した。

・集計項目の数やそれに含まれるカテゴリーの数に関する利用者の要求に応じて、多次元集計表が作成される。

・2021年人口センサスにおけるオンデマンド型公表システムには、作成された集計表から個体情報が露見されないような**自動化された露見チェック(automated disclosure check)**の仕組みが備わっており、チェック済みの集計表を利用者がダウンロードすることができる。

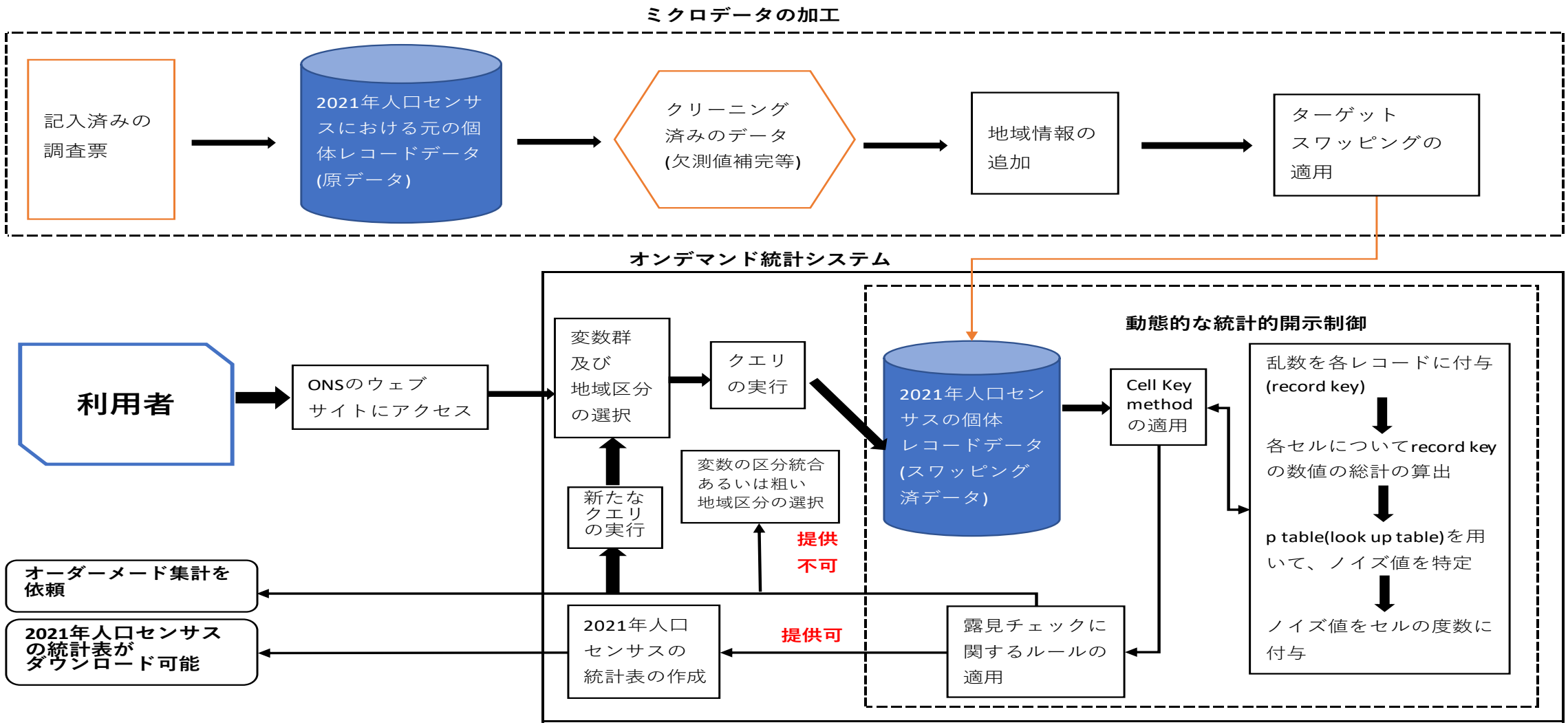
自動化された露見チェックのプロセスにおいては、集計された統計表において、以下の点が主たる露見チェックの対象となる(Blanchard(2019))。

①極端にスパース(sparce)な統計数値(当該度数は1でそれ以外のセルがゼロになるような数値)が存在するか

②統計表に含まれるセルの度数が個別の属性値あるいは集団の属性値と対応関係にあるために、個体の属性値が漏洩されるか(属性漏洩(attribute disclosure))、

③統計表の中に個体の特定につながる度数1となるセルが発見されるか(個体識別漏洩(identity disclosure))

図4 オンデマンド型公表システムのイメージ図



注　https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2017/Meeting-Geneva-Oct/Day3_1000_UK_UNECE_ONS_Flexible_dissemination_for_2021_Census.pdfをもとに筆者が作成。
出所　伊藤・寺田(2023)

・ONSが、統計表や元になるミクロデータに対して、攪乱的な秘匿処理を行うだけでなく、プライバシー保護に関わるパラメータについて公開可能な情報を公表する方向に変化してきたことは、重要な特徴だと言える。

・2011年の人口センサスまでは、統計表の公表や匿名化されたミクロデータの作成・提供において、スワッピング率のような個体情報の特定につながる可能性のある情報は公表されなかった。

・イギリスの場合、2021年センサスにおいて透明性を高める方向に転換し、これまで公開しなかった秘匿処理に関する情報を可能な範囲で公表する方向に進んでいる。

・具体的には、「オンデマンド集計システム」で用いられる自動化された露見チェックのパラメータは公表されることになっている。

⇒p-tableにおける record keyやターゲット・スワッピングにおけるスワッピング率は公表されていない。

イギリスにおける個票データ提供の現状(伊藤(2020a), 伊藤(2022), 伊藤・木村・佐々木(2024))

イギリスにおける個票データ提供の現状

・Secure Research Service(SRS)といったオンサイト施設による個票データの提供

・SRSやUK Data ServiceのThe Secure Lab等によるリモートアクセスによる個票データの提供

※2023年6月時点で、個票データにアクセスしている研究者の90%以上が、研究室からのリモートアクセスで、個票データを利用している。

→Assured Organisational Connectivity (AOC) agreementをONSとの間で締結した研究機関に所属する研究者は、所属する機関からSRSを通じてリモートで個票データにアクセスすることが可能(伊藤・木村・佐々木(2024))

・公的統計の個票データだけでなく、行政記録情報、公的統計ミクロデータとのリンケージされた行政記録情報、行政記録情報同士でリンケージされたデータの提供が可能

→2018年のDigital Economy Act 2017の施行とADR UKの支援による行政記録情報のリンケージサービス

ADR UK: ESRC(=Economic Social Research Council)による行政記録情報の利活用の推進に関するプロジェクトであり、ADRN(=Administrative Data Research Network)の後継プロジェクト

・承認された研究者(approved researcher)あるいは認可された研究者(accredited researcher)の資格を取得し、申請している研究プロジェクトが、UK Statistics Authority(UKSA)のResearch Accreditation Panel(RAP)(ONSに設定されていたMicrodata Release Panelの後継の審査委員会)による審査を通ると、個票データの利用が可能←審査の一元化

・125のデータセットが利用可能(2023年6月時点)

→教育データ、裁判データ、税務データが、利用申請が多い行政記録情報である。

・直接的な個体識別子を用いて**ONS内部でリンケージが行われた**公的統計の個票データに関して SRSにおいてアクセスすることが可能

参考　個票データの申請書類に必要な記載事項

①「認可された研究者」(あるいは「承認された研究者」)の資格を取得するための申請書類において必要な記載事項

- 申請者の所属機関
- 計量分析を行った経験
- **統計データ(個票データ)の利用経験**
  - **利用した統計データ**
  - **統計的手法**
  - **利用したソフトウェア**
  - **分析結果に対する秘匿措置**

・個票データを利用するためのトレーニングも義務付けられている。

②「認可されたプロジェクト」となるための申請書類に関する主な記載事項

- 研究プロジェクトの目的や概要
- 研究プロジェクトの期間
- モデル分析の手法
- 利用する個票データの名称、年次
- 使用するソフトウェア等

※申請対象となる研究が公共の利益(public good) をもたらすかどうかが検討されたうえで、容認された場合に個票データのアクセスが行われる。

ONSによるリモートアクセスの特徴

・分析の成果物に関するチェックについては、公的統計と行政記録情報のいずれも、提供者と利用者との相互の信頼関係に基づいて最適な分析結果の持ち出しを可能にする「原則対応型アプローチ(Principles Based Approach)」が採用されており、**マニュアル**で成果物のチェックが行われている。

→データのタイプや規模によって、アウトプットチェックの基準が異なる。世帯・人口系の統計調査に関する集計表のセルにおける閾値は10だが、事業所・企業系の調査の場合には、閾値は30となっている。

・SRSは機械学習モデルによる分析と成果物の持ち出しを想定していない。

→SRSにおいて、機械学習モデルにおいて実証研究を行う研究プロジェクトはない(2023年6月時点)

・個票データの利用申請に必要な変数について利用する個票データのタイプ、およびどういう目的で個票データを利用するかによるという点で、**ケースバイケース**である。サーベイデータについては、事前に必要な変数を提示する必要はない。人口センサスのような大規模なデータについては、利用する変数を事前に提示する必要がある。行政記録情報とのリンケージデータについても、事前に必要な変数を示すことが求められる。

⇒ウェブページで閲覧可能なメタデータを確認した上で、申請者は事前に必要な変数を選定する。

・個票データの利用申請に要する期間については、イギリスの場合、利用申請から最短で1ヶ月で提供可能であるが、データの特性によって、提供されるまでの期間は、ケースバイケースで異なる。

⇒認可された研究者(承認された研究者)の取得で早くて1週間、プロジェクトがUKSAのRAPで認められるまでに3週間で、最短で1か月

・個票データの利用が容認された場合、利用者は、平均して1年～1年6か月個票データを利用している。

ONSにおけるデータリンケージの動向について

・イギリス全体(UK)を統括する ADR UK の下で、イングランド、ウェールズ、スコットランドと北アイルランドのそれぞれにおいて、行政記録情報のリンケージされたデータの利用サービスが個別に展開されている(伊藤(2020a))。

・イングランドの場合、現在ではONSが行政記録情報間のリンケージおよびリンクされた行政記録情報の提供を一元的に担っている。

ADR UK ・・・ONS とのパートナー関係を結んでおり、資金提供を担う研究インフラ支援機関(funding agency)として、ONS におけるデータリンケージの仕組みを整備するための資金援助を行うだけでなく、行政記録情報を保有する行政機関との交渉も担っている。

・ADR UK とONS で役割分担がなされているだけでなく、ONS 内部でデータリンケージの手続きが完結していることが指摘できる。その意味では、行政機関との調整、データのリンケージ業務、研究者へのリンケージデータの提供は、一体的に展開されているため、データリンケージの業務にあたって、より効率的な手続きが進められていると言うことができる。

ONS が行っている非識別データの提供方法

① ONS でリンケージを専門的に行っている部署(SRS とは異なる)において、10～20 人程度でリンケージの作業を行う。リンケージの対象となる複数のミクロデータの中の各レコードから個人を直接識別することが可能な情報(氏名、住所、生年月日等)のみ切り離す。

② 直接識別子以外の情報は切り落とされた上で、仮名化を行った上で、リンケージ用ID が生成される。直接識別子のマッチングを行い、マッチングされたリンケージ用ID の対応表のみを残す。このとき、直接識別子は消去される。

③ ONS のデータリンケージ担当チームは、このリンケージ用ID の対応表に基づいて複数のミクロデータ間のリンケージを行う。

④ リンケージ後に、リンケージ用ID が削除されたリンケージデータがSRS に保管される。研究者は、SRS のオンサイト施設においてリンケージデータにアクセスするか、研究室からリモートアクセスすることによって、研究を進める。

※一度リンケージされたデータは、他の研究グループも申請することで、利用することが可能になっている。

(2)アメリカにおけるミクロデータの提供の現状(伊藤(2018), 伊藤他(2022), 伊藤・寺田・加藤(2024))

1. 一般公開型ミクロデータサンプル(Public Use Microdata Sample=PUMS)および一般公開型ファイル(Public Use File=PUF)の提供

＊HP上で無料でダウンロードすることが可能

Ex. 人口センサス(Population Census)や経常人口調査(Current Population Survey)

2. リサーチデータセンターによる個票データの利用サービス

・センサス局の経済研究センター(Center for Economic Study)を拠点として、全国33か所の大学等に連邦統計リサーチデータセンター(Federal Statistical Research Data Centers=FSRDCs、以下「リサーチデータセンター」)が設置され、FSRDCsにおいて、世帯・人口系だけでなく、事業所・企業系の公的統計の個票データだけでなく行政記録情報のアクセスが可能になっている。

←宣誓職員制度(Special Sworn Status)による個票データの利用

⇒SAP(Standard Application Process) Portalにおいて、16の統計作成部局における個票データの検索が可能(2023年5月時点)

3. リモートアクセスによる個票データの提供

・リサーチデータセンターでの個票データの利用経験がある研究者を対象に、リモートアクセスによる個票データの利用が可能な仕組みが整備されている(2023年5月時点)。

アメリカの人口センサスデータにおける提供の動向について

①一般公開型ミクロデータ作成の最新動向

・アメリカセンサス局(以下「センサス局」)は、1960年から2010年まで人口センサスのPUMSを作成・提供(McKenna(2019b), 伊藤他(2022))

センサスの調査年次	適用された匿名化技法
1960年	サンプリング(0.1%)、直接的な識別子の削除、人口規模が25万人以上の地域区分の設定
1970年	サンプリング(1%)、直接的な識別子の削除、人口規模が25万人以上の地域区分の設定
1980年	サンプリング(1%、5%)、直接的な識別子の削除、丸め(所得額)、トップコーディングの適用(所得額、年齢)、人口規模が10万人以上の地域区分の設定
1990年	サンプリング(1%、3%、5%)、直接的な識別子の削除、トップコーディング(全体の分布の上位0.5%かあるいは非負である分布の上位 3%)およびリコーディングの適用(居住地、勤務地、集合住宅 (group quarters)、所得、年齢等)、欠測値の補完、地域区分の閾値の設定
2000年	サンプリング(1%、5%)、 直接的な識別子の削除、トップ・(ボトム)コーディング とリコーディングの適用、 公表統計表の元になる個票データに対するスワッピングの適用、カテゴリカルな変数における分類区分のしきい値の設定(1万人以上)、ノイズの付与(世帯人員が多い世帯)、地域区分の閾値の設定
2010年	サンプリング、直接的な識別子の削除、トップ・ボトムコーディング、地域区分の閾値の設定、カテゴリカルな変数における分類区分のしきい値の設定、ノイズ付与、スワッピング、部分的な合成データの手法の適用

・2020年センサスについては、PUMSではなく、PUMF（＝Public Use Microdata File）を作成する予定。センサス局の担当者によれば、2024年にPUMFが提供予定とのことである。

・2020年センサスにおいては、サンプルデータではなく、全レコードが提供の対象となるため、PUMFが作成される予定。

・PUMFは、差分プライベートな(differential private)ミクロデータとして、作成・提供の予定である。

⇒ノイズを付与しているという意味で、2020年センサスのPUMFは、（センサス局の実務家の認識としては事実上は）**合成データ**であるとみなしている。

アメリカにおける公的統計のプライバシー保護をめぐる最近の動き

・センサス局は、プライバシーに関する懸念を2010年センサス以前から持っていた。それが、2001年にセンサス局内部にDSEP((=Data Stewardship Executive Policy Committee)が設置される契機となる(McKenna（2019a）)。

DSEP・・・センサス局がアメリカ国民及びアメリカ経済に関するデータを効率的かつ合理的に収集・利用することを保証することがDSEPの任務であり、データの収集過程全体にわたって回答者のプライバシーとデータの秘匿情報を保護することが求められる。

・2000年代～2010年代にセンサス局が一般に公開した「所得・プログラム参加調査(Survey of Income and Program Participation)」やAmerican Community SurveyのPublic Use Microdata File(PUF)と外部の公開データとのマッチングによる再識別に関する実験によって、PUFについて露見リスクがあることが明らかにされた(McKenna（2019a）)。

⇒2020年人口センサスについて、センサス局のDSEPは、2010年センサスまで適用されたスワッピングとは異なる新たな方法論の適用の方向性を模索

→差分プライバシー(differential privacy)の方法論の適用可能性(Jamin(2021）)

←データベース再構築攻撃(database reconstruction attack)への対策

アメリカセンサス局による2020年人口センサスにおける差分プライバシーの適用について

・Abowd(2018)による集計表における「データベース再構築攻撃(database reconstruction attack)」の指摘

データベース再構築攻撃・・・複数の集計結果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体情報を暴露することである(Dinir and Nissim(2003))。

→適切なパラメータ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**の考え方が出てきた(Dwork(2006))。

差分プライバシー:「『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(寺田他(2015,p.1803))

「差分プライバシーにおいて、あるランダム化関数Q’：D →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))。

$$Pr[Q'(D1)\in S]\leq e^{\varepsilon}\cdot Pr[Q'(D2)\in S]$$

ここで、D1とD2 (D1, D2∈D)は任意の隣接する(互いにたかだか1レコードしか異ならない) データベースであり、S (⊆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, p.147))。

→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。

・アメリカセンサス局では、統計実務上の経験も踏まえながら、2020年の人口センサスに向けて、2010年の人口センサスを用いた検証を行ってきた。

⇒全国レベルの性別、人種、年齢、世帯主との続き柄に関する様々な集計結果表を対象に、結果数値の精度を確保した上で公表することを可能にしつつ、安全性を確保するために、差分プライバシーの実用性に関する検証を行っている。

①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（privacy-loss budget）ε を管理しつつ、プライバシーの損失と精度のトレードオフの関係を勘案した上で、ε を決定する。

②プライバシー損失予算 ε を設定した上で、州レベル、郡レベル、センサストラクトレベル、センサスブロックレベルの各地域レベルにおいて、プライバシー損失予算を割り当てる(Garfinkel et al.(2018), Abowd et al.(2019))。

アメリカセンサス局における差分プライバシーの公的統計への適用について

・2019年10月に、センサス局は2010年センサスの回答結果を用いて差分プライバシーが適用された2010 Demonstration Data Products と呼ばれる人口センサスの統計データ(以下「センサスデータ」)を公表

⇒データの利用者に差分プライバシーを用いた新たな「露見回避システム(Disclosure Avoidance System=DAS)」が公的統計データの精度に与える影響を示すためであった。

→データの利用者は、新たに公表されたセンサスデータとすでに公表されている元のセンサスデータとの分布特性の近似性についてセンサス局に対してフィードバックを行うことができた。

・さらに、アメリカセンサス局は、2010年におけるセンサスデータを用いて、データの利用者が精度検証を可能にするための近似性を評価するための指標を新たに提案した。具体的には、平均絶対誤差(MAE)や平均2乗誤差の平方根(RMSE)が用いられている。

→この指標は、アメリカセンサス局によって実施された2020年のセンサスデータの精度の検証に用いられている。

・差分プライバシーの適用におけるパラメータεの設定にあたって、統計数値の秘匿性の観点だけでなく、データの利用者や利害関係者が要求する統計数値の精度も考慮した上で、DSEPでの議論等に基づいて、数度の修正がなされた。

・2021年6月に公開された最終版のプライバシー保護済ミクロデータファイル(Privacy-Protected Microdata Files=PPMFs)の作成においては、DSEPにおいて、DASにおける全体のプライバシー損失予算のパラメータがε=19.61と設定された(伊藤他(2022))。

⇒2021年8月16日に、2020年センサスの区画改定データ(PL94-171) がセンサス局で公表された。

・センサス局によれば、2030年センサスにおいても差分プライバシーの実現手法を適用することを計画している。

参考　差分プライバシーの方法論の有効性に関する実証研究例

Christ et al. (2022)：スワッピング技法と差分プライバシーの方法論の有効性に関する比較研究

使用するデータ：2010年センサスの公表されたブロックや群レベルのデータを用いて作成した合成データ⇒ブロックグループごとに性別、年齢、人種、ヒスパニックの有無、世帯の規模、住宅の所有形態(household tenure)に関する変数

(1)スワッピングの方法

①真の(true)ランダム・スワッピング・・・州に関する真の分布からスワッピングの対象レコードがランダムに選ばれて、レコードの入れ替えが行われる。

②擬似的な(pseudo)ランダム・スワッピング・・・州のセンサスデータの中で年齢と性別について一致するサブデータセット同士で入れ替えが行われる。人種については、州の分布からランダムな抽出がなされる。

③類似レコードスワッピング(similar swapping)・・・類似性のしきい値を設定した上で、人種、年齢と性別といった変数においてしきい値の基準を満たすように、入れ替えが行われる。具体的には、年齢に関しては年齢階級区分、人種については分類区分における近接するカテゴリーが類似性に関するしきい値の設定対象となっている。

(2)差分プライバシーの出現方法

・年齢階級区分、ヒスパニックの有無(2区分)、人種(63区分)、世帯規模(4区分)でクロスされるセルを対象に、センサス局が実施したTop Downアルゴリズムが本実験でも適用されている。

⇒クロスされた各セルの度数にノイズを付加した出力結果が得られる。

・州レベルの人口数が公表数値と正確に一致するだけでなく、負の値については0に変換するといった処理も考慮した上で、ノイズが付与されている。

有用性と秘匿性の定量的な評価について

(1)有用性の評価

スワッピングや差分プライバシーの実現手法によって攪乱が施されたデータの元データに対する有用性に関する評価指標として、対象となる変数のヒストグラムについて平均平方誤差(MSE)および「平滑化された」カルバック・ライブラー情報量(μ −smoothed Kullback−Leibler divergence、以下「KL情報量」と呼称)を用いている。

⇒スワッピング率が減少およびパラメータ ε が上昇するにしたがって、MSEおよびKL情報量が小さくなっており、攪乱が施されたデータが元データに対して相対的に近似する傾向にある。

(2)秘匿性の評価

・スワッピング済のデータについては、一般に公表されている外部情報としてのデータセットとのリンケージを行い、スワッピングされたレコードが特定化されるリスクに関する定量的な評価を行っている。

・差分プライバシー適用済みデータについては、クロス集計された各セルに対してノイズ付加を行うことから、スワッピングと同様にレコードリンケージによる攻撃で秘匿性の検証を試みようとしても、レコードに個体識別子が付与されていないため、リスク評価を行うことは困難になっている。

本実験の結果

スワッピングと差分プライバシーの比較研究の成果から、①パラメータ ε の値にもよるが、差分プライバシーの実現手法が適用されたデータにおいても、スワッピングを適用した場合と同等のデータの精度を担保することが可能なこと、②スワッピングにおいては、人種に着目した場合、多くの異なる人種である人々が居住する地域、とくにマイノリティが居住する地域においては、スワッピング率を上昇しても、個体識別リスクが増大傾向にあることが明らかになった。

(3)デンマークにおけるミクロデータの提供状況(伊藤(2020b), 伊藤(2024))

・デンマークの場合、リモートアクセスを通じてのみ、個票データ(非識別データ、deidentified data)の利用サービスが行われている。

・デンマーク統計局におけるミクロデータ提供の経緯

1980年代半ば:デンマークにおける学術研究のための公的統計のミクロデータの利用に対する関心の高まり

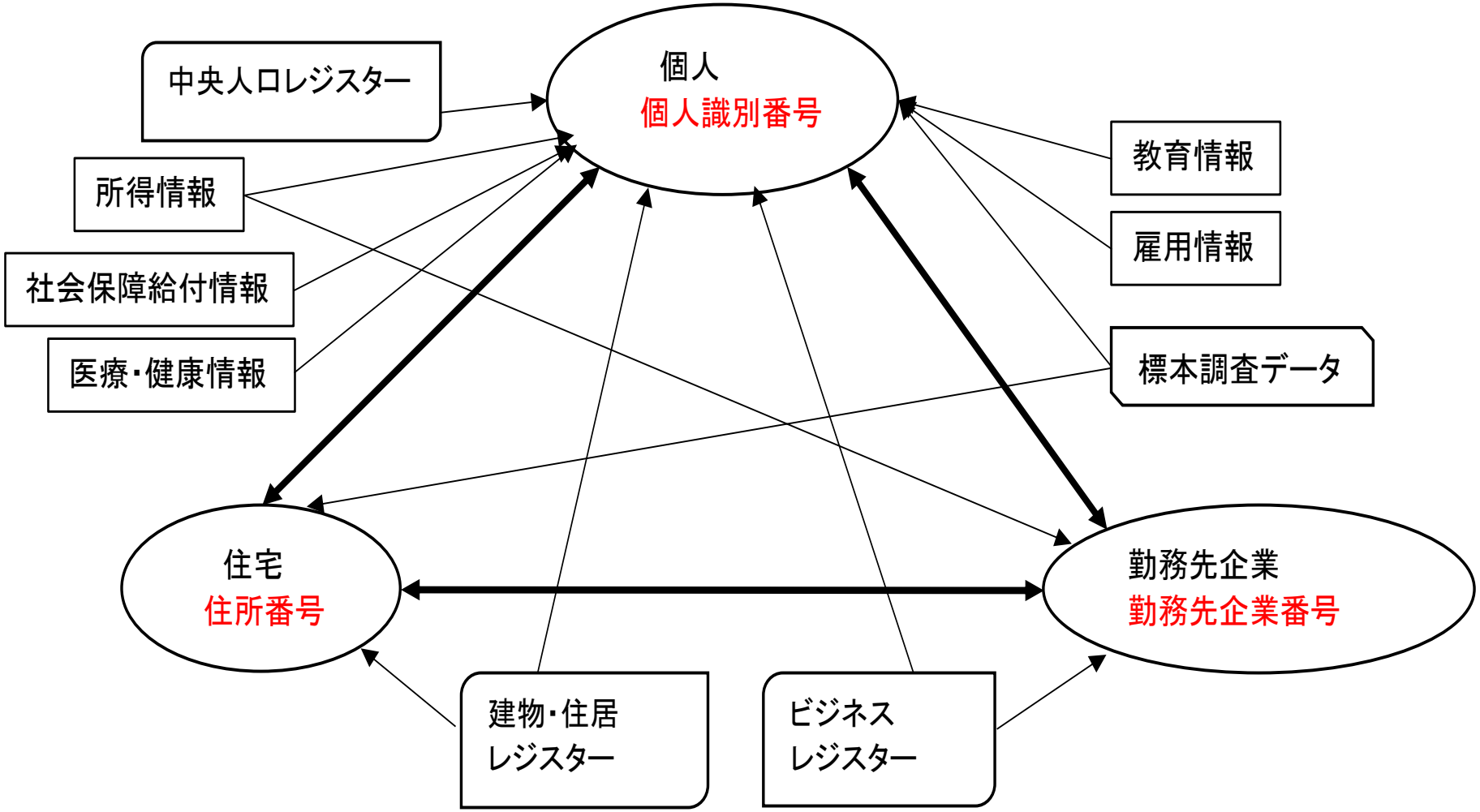
1988年:デンマーク統計局内にセキュアな環境であるオンサイト施設が設置

2001年:リモートアクセスサービスが開始

2008年:オンサイト施設は閉鎖され、リモートアクセスのサービスが拡充

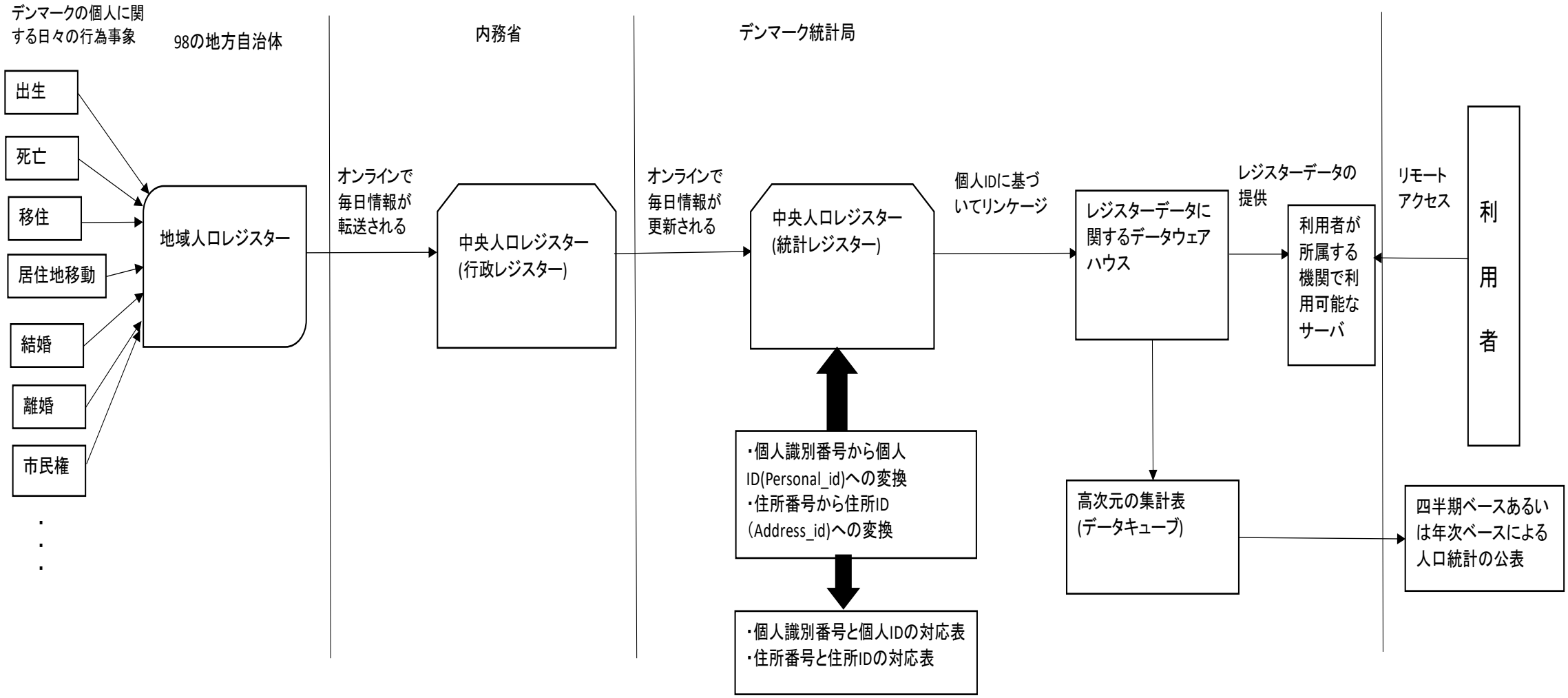
※人口社会情報(結婚と離婚、出生と死亡、移住等)、労働・所得情報(就業、職業訓練、賃金、家族所得等)、教育に関する情報(最終学歴、教育歴等)、企業経済情報(企業の雇用、購入・販売等)、医療健康情報(国民患者レジスター(National Register of Patients)、医薬品統計レジスター(Register of Medical Product Statistics)等)といった様々な情報を対象に、利用者はリンケージされたデータにリモートでアクセスすることが可能

図5 デンマークにおける統計情報を作成するためのシステム



出所　伊藤(2020b)、図2

図6 人口・社会統計に関するデンマークにおける統計作成のプロセス



出所　伊藤(2020b)、図3

デンマークにおけるリモートアクセスの特徴

・ミクロデータのアクセスに関しては、**認証された機関（authorized institution)**においてのみ個票データの提供が可能

→認証された機関の研究者及びその共同研究者であれば、認証された機関を通じて個票データにリモートでアクセスすることができる。

・「知る必要性(need to know)」原則に基づいて、利用者において必要な変数群を含むミクロデータのみアクセスすることが可能

→プロジェクトごとに別々のIDが振られており、そのIDにしたがって、デンマーク統計局がリンケージを行う。そのため、研究者がIDに直接アクセスすることはできない。

・2018年デンマークデータ保護法(Danish Act on Protection of Data 2018)、1985年行政機関法(Public Administration Act 1985)、デンマーク刑法(Danish Panel Code)に基づく法的な根拠に基づいて、ミクロデータが提供される。

⇒2018年デンマークデータ保護法に基づいて、統計目的で集められた情報を他の統計目的や研究目的のために加工することが可能になっている。

・2018年に、デンマーク統計局によって開発されたスキャンシステムに基づいて、研究者は分析結果のチェックを自動的に行うことが可能である。

→分析結果のチェックについては、一部の結果をランダムに選んだ上で、職員が事後的にマニュアルでチェックを行う。

・デンマーク統計局は、利用者の分析目的に応じて、個票データの提供を行っている。

→個票データの利用者については、研究者だけでなく、官庁や地方自治体の職員、さらには企業に所属する者も利用者の対象に含まれる。

※例えば、企業がマーケティングリサーチを行うために個票データを用いるようなケースも容認されている。

⇒社会に還元した研究、公共の利益に資する研究であることが求められる。商用目的のために個票データを利用することは容認されない。

・個票データの利用申請に要する期間については、デンマークの場合、利用申請から6～10日で提供可能となっている。

⇒利用者からのニーズに応える形

・デンマークにおいては、**データの提供者と研究者との間の信頼関係**を重視しており、それに対応する形でリモートアクセスが広範に展開されていることが見逃せない点だと言える。

4.わが国における公的統計のミクロデータの提供状況と今後の課題

・現行の統計法(平成十九年法律第五十三号)に基づき、わが国における公的統計データの提供については、データの秘匿性と利用者のニーズに合わせた形で様々なチャンネルが存在する。

(1)統計表(集計結果表)

⇒政府統計の総合窓口(e-Stat)による統計表の公表→オープンデータとして公開

(2)個票データ

⇒「調査票情報」(統計法第2条第11項)として提供

調査票情報・・・「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」ミクロデータ

(3)匿名化されたミクロデータ

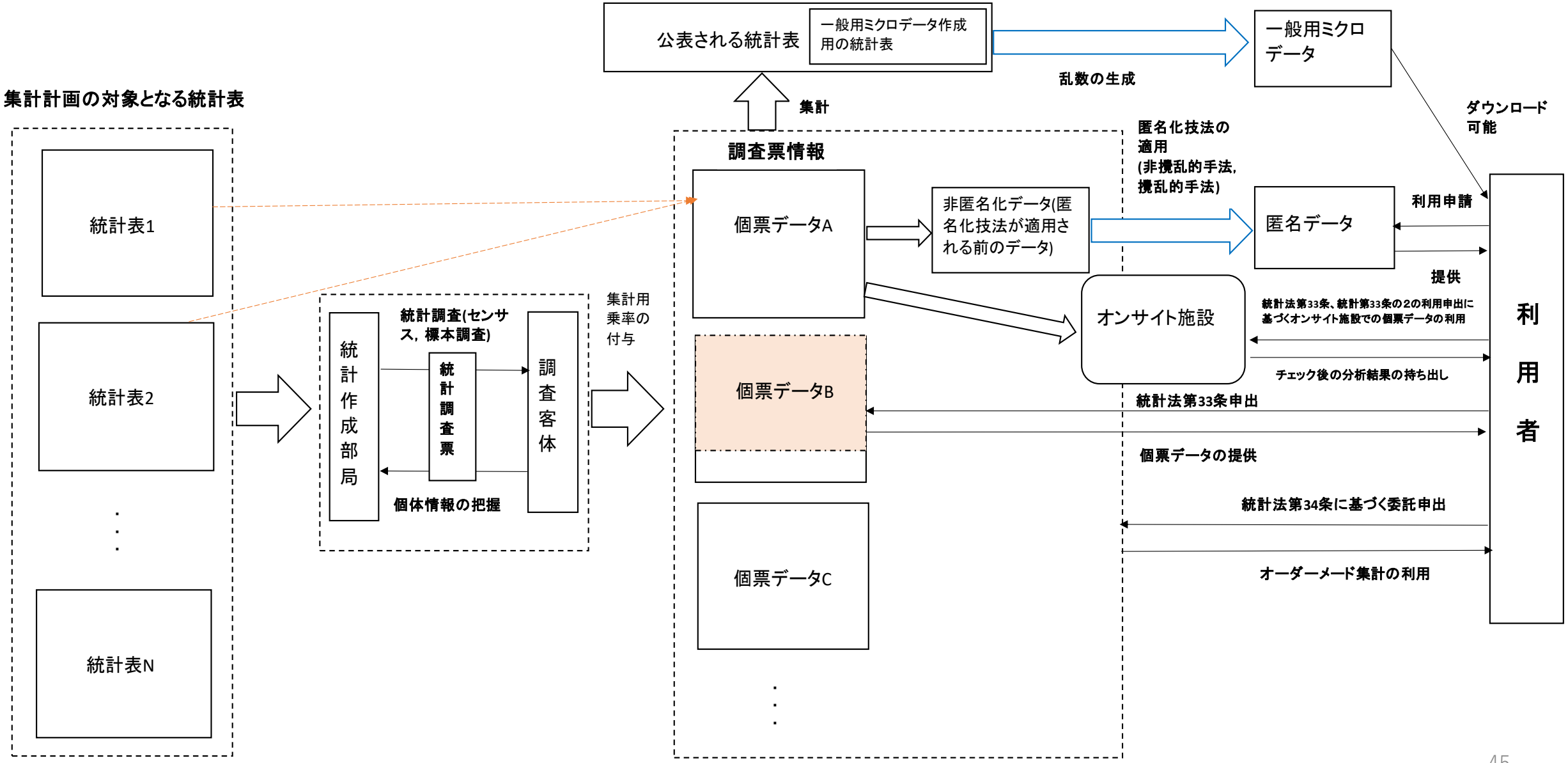
⇒「匿名データ」(統計法第2条第12項)として提供

「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」ミクロデータ

(4)オーダーメード集計

※わが国では、2種類の統計調査を対象に、一般用ミクロデータも公開されている。

図7　わが国の公的統計ミクロデータの提供に関する概略図



出所　伊藤(2022)，図1

(1)e-Statによる統計表の活用

e-Stat： 2008年から運用が開始された政府統計に関するポータルサイトであり、各府省が公表している統計データの検索だけでなく、統計表のダウンロードが可能

※e-Statが実用化される前は、紙媒体で刊行される報告書に掲載されている統計表から結果数値を入手していた。

←報告書に掲載されない統計表(非掲載統計表)も存在

対象となる分野

①国土・気象、②人口・世帯、③労働・賃金、④農林水産業、

⑤鉱工業、⑥商業・サービス業、⑦企業・家計・経済、

⑧住宅・土地・建設、⑨エネルギー・水、⑩運輸・観光、

⑪情報通信・科学技術、⑫教育・文化・スポーツ・生活、⑬行財政、

⑭司法・安全・環境、⑮社会保障・衛生、⑯国際、⑰その他

(2)個票データの利活用

わが国における公的統計の調査票情報(個票データ)については、現行の統計法の下で2つの提供形態がある。

(1)磁気媒体による調査票情報の提供(法第33条)

・実証分析を行う上で**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**

(2)オンサイト施設による調査票情報の提供(法第33条、法第33条の2)

・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を行う**事後審査方式**

⇒世帯・人口系および事業所・企業系の統計調査のミクロデータが利用可能

・特にオンサイト施設での利用においては、仮説検証的な研究だけでなく、探索的な実証研究を行うことも想定される。

調査票情報の提供状況　952件(2024年4月24日時点)

※調査票情報の利用にあたっては、統計法における安全管理措置(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。

(3)匿名データの利用

・2007年に改正された統計法(統計法(平成19年法律第53号))が公布されたことにより、匿名データの提供は、2009年4月から開始された。

⇒統計法36条に基づく提供

現在、国勢調査や労働力調査等、8つの統計調査の匿名データが作成・提供されている。

←調査票情報に対して、リサンプリング(サンプリング)、レコードの削除やリコーディング(区分の統合)といった匿名加工を施すことによって、匿名データが作成される。

・匿名データの利用目的

①学術研究、②教育、③国際比較統計利活用事業、④デジタル社会形成統計利活用事業

・匿名データも、調査票情報と同様に、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。

・匿名データの利用においては、手数料が発生する。

・匿名データの提供状況　　159件(2024年4月24日時点)

⇒全国消費実態調査、就業構造基本調査、社会生活基本調査の匿名データの申請数が比較的多くなっている。

(4)オーダーメード集計の利用

・オーダーメード集計・・・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

⇒31の統計調査を対象に、有料でオーダーメード集計が可能

→統計法第34条(委託による統計の作成等)

オーダーメード集計の提供状況　　118件(2024年4月24日時点)

※2019年における改正統計法の施行に伴い、オーダーメード集計によって作成した統計表は、2019年5月以降、ミクロデータ利用ポータルサイトmiripoで公表されている。

一般用ミクロデータについて

・わが国では、現在全国消費実態調査と就業構造基本調査の**一般用ミクロデータ**が公開されている。

・一般用ミクロデータは、公表された統計表に含まれるセルの数値からミクロレベルの値を生成させることによって作成されることから、web 上での取得が可能になっている。

⇒一般用ミクロデータは、利用目的の制限がなく誰でも入手可能であることから、それは、わが国における**一般公開型ミクロデータ**に該当すると言うことができる。

・現在公開されている全国消費実態調査の一般用ミクロデータの作成については、2011年8月から(独)統計センターで試行提供された**教育用擬似ミクロデータ(Synthetic Microdata)**の作成に関する手法が、その方法的な基礎になっている。

→具体的には、高次元の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、多変量の正規乱数の生成が行われている(山口他(2013))。

⇒この手法は、**合成データ**の方法論の適用例の1つとみなすこともできる。

・教育用擬似ミクロデータは、匿名化技法の1つである**ミクロアグリゲーション(microaggregation)**(伊藤(2009))が方法的に展開されたものと位置付けられうる。

・ミクロアグリゲーションの一形態としても位置づけられる**「超高次元クロス集計表」**に基づき作成された集計表のセルに含まれる数値群を質的属性値群と量的属性値を含む個票データに準じたレコード群(ミクロアグリゲートデータ(ミクロアグリゲーションが適用されたデータ))として擬制的に捉えることも可能なように思われる(伊藤(2009, pp.211-212))。

超高次元クロス集計表・・・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(伊藤(2009, p.211))

公的統計データをめぐる最近の動き

2016年12月 **「官民データ活用推進基本法」**の施行

→オープンデータ及び統計データの利活用の推進

2017年1月 **統計改革推進会議**の設置

2017年5月 『統計改革推進会議最終取りまとめ』の刊行

→EBPM（=Evidence Based Policy Making、証拠に基づく政策立案）を指向

・Public Use Fileの提供可能性

2017年9月 総務省政策統括官（統計基準担当）**「調査票情報等の利用、提供等に関する法制研究会」の開催**（2017年9月～2018年3月）（座長 廣松 毅（情報セキュリティ大学院大学情報セキュリティ研究科客員教授））

→調査票情報等の提供における法律面での課題について審議

1）調査票情報のオンサイト施設における提供サービスに関する法的な位置付け

2）匿名データの提供に関する法律上の整理

2018年3月 新たな**「公的統計の整備に関する基本的な計画」**（**「第Ⅲ期基本計画」**）

（平成30年3月6日閣議決定）

・「一般の人も利用できる匿名データ」の提供に関する法制面や技術面からの検討

・賃金構造基本統計調査の個人票を中心とした匿名データの提供の検討

2018年3月6日 「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案」が国会に提出

→統計法の一部改正

① 行政機関等の責務等の規定設置

② 事業所母集団データベースに記録されている情報の提供を受けることができる 調査の範囲等の拡大

③ 調査票情報の提供対象の拡大

④ 統計委員会の機能強化

「公的統計の整備に関する基本的な計画」』(第Ⅲ期基本計画)(2018年3月、2020年6月一部変更)においては、証拠に基づく政策立案(Evidence Based Policy Making=EBPM)の推進が強く指向されている。

・統計の利活用促進・環境改善

①調査票情報等の利用促進の必要性

⇒個票データ(調査票情報)を対象にした、セキュリティレベルの高いオンサイト施設による利用の拡充

②e-Statのさらなる改善による利便性の向上

⇒機械判読可能な形式でのデータ提供の拡大

第Ⅳ期「公的統計の整備に関する基本的な計画」(令和5年3月28日閣議決定)では、公的統計の作成・提供・利用の基盤整備が、基本計画として提示されている。

①e-Statのユーザーインターフェース等の改善、メタデータ整備の改善、データカタログ機能の追加

②調査票情報(個票データ)の利活用の柔軟化や提供手続等の**迅速化・簡素化**

⇒調査票情報の安全性を確保する方式として、研究室の端末からセキュリティを確保しつつ調査票情報を利活用できる**リモートアクセス**の実証実験への取り組み

「調査票情報等の提供についても、公共の利益に資する様々な分析ニーズ等に応えるため、デジタル技術も活用し、情報管理を徹底しつつ、迅速化・円滑化等に取り組む」(6頁)

わが国の公的統計ミクロデータの提供に関する課題(伊藤(2022))

わが国においては、**①個票データの提供サービス、②匿名化ミクロデータの作成・提供**および**③一般公開型ミクロデータ**が実現していると言うことはできる。

・**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**は現状では実現していない。

・匿名化ミクロデータの作成に関しては、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

⇒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可能性

・一般用ミ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。

⇒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の作成方法のさらなる検討

・リモートエグゼキューション(オンデマンド集計、プログラム送付型)は、現在もわが国では実現に至っていない。

・統計法の定義に基づけば、調査票情報は、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。

・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。

※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、このような事情が関係していると考えられる。

→海外で実用化され、人口センサスのミ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。

統計法における目的規定について

・統計法の目的に関する条文(法第1条)

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。

・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のではないか。

・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、公的統計の二次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。

規制改革推進会議における公的統計の二次利用をめぐる議論

・規制改革推進会議の第7回医療・介護・感染症対策ワーキング・グループ(以下「WG」と略称)(2023年3月6日開催)において、「公的統計の調査票情報の二次利用について」というテーマで、公的統計の調査票情報の提供に関する今後のあり方について議論がなされた。

・「規制改革推進に関する答申～転換期におけるイノベーション・成長の起点～」(以下「答申」と呼称)(2023年6月1日発表)においては、「公的統計の調査票情報の円滑な二次的利用の確保」を目指した今後の方針が提示された。

(1)調査票情報の提供に関する審査についてのマニュアルの作成、利用申出の様式の統一に関する方針

①利用する変数の選定(「塗り絵」作業)に関する手続きの簡略化

②「公益性」の審査において統計所管府省庁による「外形的な確認」のみの実施

⇒「具体的な研究内容に踏み込んで公益性の有無に関する判断を行うものではないことを明確化すること」

③論文等の成果物において、研究者等がプライバシー保護について適切に実施することが明確化←磁気媒体による調査票情報の提供

⇒「初回利用時の管理状況等の確認に基づく資格認定、誓約の徴取その他の手段」によって、プライバシー保護が適切に行われているかを確認

(2)統計所管府省庁が利用申出から調査票情報の提供までに要する期間を、令和5年度中に平均1か月以内、令和6年度中に平均1週間以内、かつ、遅くとも4週間に短縮

(3)リモートアクセスの実証実験の開始およびリモートアクセスによる調査票情報の提供の実現

(4)公的統計に関するメタデータの整備および公的統計データの機械可読化の推進

(5)調査票情報に含まれる回答者ごとの連結(「紐付け」)を可能にするための方策

わが国の公的統計ミクロデータの作成・提供に関する将来的な方向性について（伊藤(2022)）

海外におけるミクロデータの提供形態を参考にした場合、わが国における公的統計ミクロデータの作成・提供に関する将来展望を図る上での課題が考えられる。

①個票データのアクセスの利便性の観点からの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の追究

②利用者のニーズを踏まえた匿名化ミクロデータさらには一般公開型ミクロデータの作成の方向性の探究

③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の検討

①個票データのアクセスの利便性の観点からの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の追究

・リモートアクセスにおける論点は、①個人情報の保護の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。

⇒5つの安全モデルの中の**安全なプロジェクト**と**安全な利用者**をリモートアクセスに対応する形で位置付けた上で、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。

・Eurostatのような「認証された機関(authorized institution)」(伊藤(2018))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へのリモートアクセスの要件とすることも議論の対象になりうる。

・将来的には、研究室からの個票データのリモートアクセスだけでなく、デンマークのようなモバイルPCによる個票データのリモートアクセスの可能性も模索されてよいのではないか(伊藤(2020b))

・リモートエグゼキューションの場合、用いられる保管用データについて議論する必要がある。

(1)個票データ、(2)匿名化ミクロデータ、(3)データキューブ(hypercube、高次元の集計表)が想定

・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)の場合、統計法の第34条(オーダーメード集計)の運用面での拡大可能性

(2)の場合、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ミクロデータの作成が可能かどうかについて、法解釈の議論の必要性

(3)については、集計計画における高次元の集計表の設定可能性、オンデマンド集計システムへの展開の可能性の追究

・リモートエグゼキューションによる分析結果については、安全性を担保した上で、有用な結果数値をどのように提供するかということも重要な課題となる。

「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、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の必要性(伊藤(2020b))

②利用者のニーズを踏まえた匿名化ミクロデータさらには一般公開型ミクロデータの作成の方向性を追究

・統計作成部局は、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を採用する。

⇒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。

→近年の海外の動向として、一般公開型ミ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(2018))。

・秘匿性の強度を高めた教育用の匿名化ミクロデータの作成は、将来的な検討課題になりうる。

→わが国で、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場合、統計法における安全管理措置や罰則に関する法解釈や運用の変更についての措置が必要になるだろう。

⇒ミ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に関する検討が求められる。

・海外で展開されている合成データ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ミ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」(伊藤(2022, p.13))。

③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の検討

・行政記録情報の二次利用については、法制度な措置と秘密保護に関する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。

→公的統計のミクロ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

・現行統計法は、あくまで調査統計を対象にしていることから、行政記録情報や民間の個人情報が追加された公的統計ミクロデータは、法第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。

→行政記録情報を含む個人情報と調査票情報の連結可能性、リンケージデータの利活用を想定しようとすれば、統計法における調査票情報の定義を踏まえた上で、行政記録情報や民間の個人情報も調査票情報に(あるいは個票データという形で)包含できるかどうかが、行政記録情報の二次利用に向けた今後の検討の方向性の1つだと考える。

・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、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も考えられる。

・統計技術的な課題としては、公的統計ミクロデータと行政記録情報が連結された場合のメタデータの整備とその提供可能性も検討課題となりうる(公的統計ミクロデータにおいて提供可能な符号表(codebook)の標準化等)。

5. まとめ

・本報告では、海外における公的統計ミクロデータを対象にした匿名化措置の現状とミクロデータの利活用の動向を明らかにした。つぎに、このような海外の状況を踏まえた上で、わが国における公的統計ミクロデータの提供をめぐる課題と今後の方向性を提示した。

・公的統計における匿名化技法の有効性については、海外では攪乱的手法が積極的に適用される傾向にあるが、ヨーロッパとアメリカでは方向性が異なる。

→アメリカでは、最新の人口センサスに差分プライバシーの方法論が適用されたが、ヨーロッパでは、人口センサスにターゲット・スワッピングやcell key methodが推奨

・センサス局のように、差分プライバシーの方法論をPUFの作成・提供に援用する動きが見られる。一方、わが国では、統計表の作成、さらには公的統計の二次利用に関する展開の方向性を見据えた上で、差分プライバシーの適用可能性についての検討が必要になるだろう。

・海外では、公的統計や行政記録情報の個票データに対するリモートアクセスが広範に展開されている。

→デンマークやフランスのように提供者と利用者との信頼関係に依拠したリモートアクセスの仕組みを導入している国々があるが、Eurostatやオランダのように提供者と利用者の間で分析結果の安全性に関して責任の分担を指向している国々もある。

⇒ドイツのように個票データのリモートアクセスが容認されていないが、SUFのリモートアクセスを検討している国も存在している。

・わが国の個票データを利活用のさらなる促進を図る上で、公的統計の個票データや行政記録情報の利用申請に関する手続き(資格取得制度、審査機関の一元化等)、複数のミクロデータを対象にしたリンケージデータに対するリモートアクセスの方法、個票データに基づく分析結果のチェック方法等、海外の事例については参考になる点が少なくないように思われる。

・わが国では、リモートアクセスの実現に向けた議論が展開されているが、わが国のリモートアクセスの推進にあたっては、提供者側と利用者側の信頼関係についてどのような形で制度設計を行うかが重要だと考える。

⇒5つの安全モデル(特に安全な利用者と安全なプロジェクト)の適用可能性の追究

・Public Use Fileを含む匿名化されたミクロデータの作成可能性、合成データの生成可能性、高度な攪乱的手法の統計データへの適用可能性については、実証研究の成果に基づき、わが国においてさらなる検討を進めることが求められよう。

参考文献

伊藤伸介(2009)「匿名化技法としてのミクロア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,, pp.197-232.

伊藤伸介・星野なおみ(2014)「国勢調査ミクロデータを用いたスワッピングの有効性の検証」『統計学』107号, 1～16頁

伊藤伸介(2016a)「政府統計における個票データの提供と秘密保護について―イギリスを例に―」,『経済学論纂(中央大学)』第56巻第5・6合併号, pp.1-19.

伊藤伸介(2016b)「諸外国における政府統計ミ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, pp.233-249.

伊藤伸介(2017)「国勢調査ミクロデータにおける匿名化の誤差の評価方法に関する一考察」,『経済学論纂(中央大学)』第57巻第3・4合併号, pp.189-209.

伊藤伸介(2018)「公的統計ミ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, pp.77-101.

伊藤伸介(2019)「公的統計データにおける秘匿性と有用性の評価のあり方に関する一考察―スワッピングを中心に―」, 坂田幸繁編『公的統計情報―その利活用と展望』中央大学出版部, pp.39-62.

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, pp.1-16.

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, pp. 109-138.

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, pp.139-166.

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, pp.131-150.

伊藤伸介(2022)「ミクロデータの匿名化と統計情報の秘匿可能性について」『経済学論纂(中央大学)』第63巻1・2合併号, pp.1-23.

伊藤伸介・寺田雅之(2023)「海外における公的統計に関するプライバシー保護の現状―アメリカとイギリスの事例をもとに―」『統計研究彙報』第80号, pp.117-136.

参考文献

伊藤伸介・横溝秀始(2024)「わが国の公的統計における合成データの展開可能性に関する一考察―事業所・企業系の統計調査を例に―」『経済学論纂(中央大学)』第64巻3・4合併号, pp.147-164.

伊藤伸介・木村映善・佐々木香織(2024) 「イギリスにおける大規模データの二次利用の新たな展開―スコットランドの事例を中心に―」,『統計研究彙報』第81号, pp.89-108.

伊藤伸介・寺田雅之・加藤駿典(2024) 「公的統計に対する差分プライバシーの適用と有効性の評価に関する検討―国勢調査を例に―」,『統計研究彙報』第81号, pp.69-88.

伊藤伸介(2024)「海外におけるミクロデータの利活用推進に向けた取り組み状況―リモートアクセスを中心に―」(『経済学論纂(中央大学)』第65巻1号において掲載予定)

規制改革推進会議(2023)「規制改革推進に関する答申～転換期におけるイノベーション・成長の起点～」

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の適用」『情報処理学会論文誌』, 第 56巻第 9号, pp.1801-1816.

濱砂敬郎(1999)「ドイツ連邦統計法におけるミクロデータ規定の匿名化措置」, 法政大学日本統計研究所『研究所報』No. 25, pp.69-99.

山口幸三・伊藤伸介・秋山裕美(2013)「教育用擬似ミクロデータの作成―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.

Abowd, J. M. (2018). Staring-down the database reconstruction theorem, Joint Statistical Meetings, Vancouver, BC, Canada. https://www.census.gov/content/dam/Census/newsroom/press-kits/2018/jsm/jsmpresentation-database-reconstruction.pdf

Abowd, J, and Schmutte, I. M. (2019) "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices", *American Economic Review*, Vol.109, No.1, pp.171-202.

Blanchard S., "The Methodological Challenges of Protecting Outputs from a Flexible Dissemination System", *Survey Methodology Bulletin* 79, pp. 1-15.

Brenzel, H., Cramer, K., Güttgemanns, V., Mathes, M. (2023) "Remote Access for Scientific Use Files - a New Pathway for German Official Statistics Microdata Access", Paper presented at UNECE Expert Meeting on Statistical Data Confidentiality 2023, pp.1-6.

参考文献

Bujnowska, A., Espelage, F., Stocchi, M.(2023) “Remote access to European microdata”, Paper presented at UNECE Expert Meeting on Statistical Data Confidentiality 2023, pp.1-6.

Christ, M., Radway, S., Bellovin, S. M. (2022), “Differential Privacy and Swapping: Examining De-Identification’s Impact on Minority Representation and Privacy Preservation in the U.S. Census”,Paper Presented at Conference: 2022 IEEE Symposium on Security and Privacy, pp.457-472.

Cramswick, K. (2019) “Virtual data labs – A more flexible approach to access Statistics Canada microdata”, Paper presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, pp.1-13.

Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy, in *Proceedings of the twentysecond ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 202-210

Duncan, G. T., Elliot, M., Salazar-González, J-J.(2011) *Statistical Confidentiality*, Springer.

Dwork, C. (2006) “Differential privacy”, Proc. 33rd intl. conf. Automata, Languages and Programming, LNCS 4052, Springer, pp. 1-12.

Garfinkel, S., Abowd, J. M. and Martindale, C. (2018). Understanding database reconstruction attack in public data: these attacks on statistical databases are no longer a theoretical danger, *Acmqueue*, 16(5), 1-26.

Jamin, R. (2021), “Disclosure Avoidance for the 2020 Census: An Introduction”, U.S. Census Bureau.

McKenna, L. (2019a), “U.S. Census Bureau Reidentification Studies”, U.S. Census Bureau.

McKenna, L. (2019b), “Research and Methodology Directorate: Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Censuses of Population and Housing Public Use Microdata Samples”, U.S. Census Bureau.

Office for National Statistics (2017) “Development of flexible dissemination for 2021 Census”.

Templ, M. (2017). *Statistical Disclosure Control for Microdata: Methods and Applications in R*, Springer International Publishing

Woo, M., Reiter, J. P., Oganian, A., Karr, A. F.(2009) “Global Measures of Data Utility for Microdata Masked for Disclosure Limitation”, *The Journal of Privacy and Confidentiality*, Vol.1, No.1,pp.111-124.