Discussion Paper No.335

公的統計における高度な攪乱的手法の適用可能性について
―アメリカ人口センサスにおける取り組みを中心に―

中央大学経済学部
伊藤　伸介

September　2020

INSTITUTE OF ECONOMIC RESEARCH
Chuo University
Tokyo, Japan

公的統計における高度な攪乱的手法の適用可能性について
—アメリカ人口センサスでの取り組みを中心に—*

中央大学経済学部
伊藤　伸介

1. はじめに—わが国の公的統計における二次利用の現状—

わが国の公的統計データにおいては、統計調査によって得られた調査票情報(個票データ)を集計することによって、オープンデータとして一般に公表されている統計表、および記入済みの調査個票に基づいて作成されたミクロデータの両面から、さまざまな形態で公的統計データの作成・提供が行われてきた。具体的には、統計法における条文の規定に基づき(第 33 条 調査票情報の提供、第 34 条 委託による統計の作成等、第 35 条 匿名データの作成、および第 36 条 匿名データの提供)、わが国における個票データの提供、および匿名データ(調査票情報に匿名化技法を適用することによって作成されたミクロデータ)の作成・提供、さらには、統計調査を対象にした有料のオーダーメイド集計サービスが行われている(伊藤(2018c))。

わが国における公的統計の統計表(集計結果表)は、政府統計の総合窓口である e-Stat でも公表されており、インターネット上で広範に利用することが可能である。また、総務省統計局で実施されている統計調査においては、API(Application Programming Interface)の機能や統計 GIS 機能を強化することによって、オープンデータとしての公的統計の利便性を一層高めている。それは、2016 年 12 月に施行された「官民データ活用推進基本法」によって、公的統計だけでなく、行政記録データや民間のビックデータを含むオープンデータの利活用の推進が図られていることと大いに関係している(伊藤(2016a))。

その一方で、欧米諸国において進展してきた「証拠に基づく政策立案(Evidence Based Policy Making= EBPM)」をわが国でも推進していくために、公的統計データだけでなく、民間のビックデータや行政記録データについても、その利用可能性が指摘された。2017 年に統計改革推進会議が設置され、EBPM を指向する形で、匿名データやオンサイト利用を含む公的統計のミクロデータの作成・提供のあり方が議論された。こうした状況の中で、わが国においては、2018 年 6 月 1 日に改正された統計法(「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「改正統計法」と呼称)が公布され、公的統計データの利用拡大のさらなる促進を図ろうとしている(伊藤(2018c))。

このように公的統計の二次利用が進展する中で、今後も、わが国では現在展開されていないオンデマンド集計の可能性も追究されるであろう。それによって、公的統計の統計表の公表のあり方にも影響を与えることも考えられる。

---

* 本稿の作成にあたり、寺田雅之氏(NTT ドコモ)から貴重なコメントをいただいた。記して感謝を申し上げたい。

ところで最近では、公的統計の分野において、秘匿の基準に従ってノイズをコントロールする方法として、主に情報工学の分野で展開されてきた差分プライバシー(differential privacy)の方法論の適用可能性が、海外の統計作成部局において注目されている。その理由として、とくに小地域レベルの統計表において、複数の統計表の組み合わせによって個体の特定化を行う差分攻撃(differencing attack)[1]のリスクに対する対応の必要性が指摘されている。そこで、2019 年の 10 月に、オランダ統計局において国際連合欧州経済委員会(The United Nations Economic Commission for Europe=UNECE)と欧州統計局(Eurostat)が共同で開催した「統計データの秘密保護に関するワークセッション(Work Session on Statistical Data Confidentiality、以下「秘密保護ワークセッション」と略称)」では、差分プライバシーに関する報告や議論が、統計作成部局の関係者から注目されてきた[2]。具体的には、差分攻撃のようなリスク回避の方法として、セルに含まれる結果数値にノイズを付与することによる集計表の作成および提供が、カナダ統計局やノルウェー統計局といった統計作成部局からも近年注目されている。例えば、Thomas(2019)は、カナダ統計局における個票データの二次利用の取り組みの１つとして、個票データによる分析結果の公表におけるリスク回避の観点から差分プライバシーの方法論の分析結果のリスク評価に対する適用可能性を議論している。Heldal et al.(2019)は、ノルウェー統計局で展開されているプログラム送付型のリモートエグゼキューションシステムである microdata.no において、集計表のノイズ付与における差分プライバシーの方法論を検討している。また、オーストラリア統計局(ABS)は、TableBuilder[3]におけるノイズ付与の方法論として、差分プライバシーの適用可能性を追究しており、これまでのノイズ付与の方法と差分プライバシーを用いたノイズ付与について実証的な比較研究を行っている(Bailie and Chien(2019))。

わが国の公的統計の実務の観点からは、差分プライバシーの方法論の有効性がこれまで議論されたことはなかったと言える。その意味では、公的統計データに対して差分プライバシーの方法論の可能性を議論することは、「公表可能な」統計表の安全性を追究する上でも有意義であると言える[4]。とくに、メッシュデータのような地域区分が詳細な統計データに対しても差分プライバシーの方法論の適用可能性を追究することは、小地域統計を含む統計表の公表可能性のさらなる拡大にも寄与しうる。

---

[1] 差分攻撃については、Fraser and Wooton(2005)を参照。

[2] 「秘密保護ワークセッション」は、2 年に 1 回、主として EU 域内で開催されており、ヨーロッパ諸国を中心に、統計作成部局の関係者、統計データの秘匿に関する専門家、さらには情報セキュリティの研究者が参加している。主として公的統計の統計表およびミクロデータを対象に、公的統計ミクロデータの提供状況、法制度的および統計技術的な匿名化措置の動向、匿名化に関する方法論の最新動向、匿名化されたミクロデータに対する秘匿性と有用性の評価方法等について、活発な議論と関係者による意見交換が行われている。

[3] オーストラリア統計局が開発した TableBuilder の特徴については、伊藤・谷道・小島(2018)を参照されたい。

[4] 統計表の公表可能性に関する議論については、伊藤(2015)を参照。

そこで、本稿では、最初に公的統計の公表に関する差分プライバシーの取り組みの先進的な事例として、差分プライバシーの概念を整理した上で、アメリカセンサス局が行った2010年人口センサスを用いた差分プライバシーの適用に関する実証研究の概要を述べる。最後に、わが国における公的統計の分野における差分プライバシーの可能性について私見を述べることにしたい。

2. 差分プライバシーについて

諸外国では、公的統計が、統計表(集計結果表)およびミクロデータという形で利用可能になっている。そして、公的統計のミクロデータにおいては、さまざまな形態による提供が進められてきた。具体的には、①個票データ(deidentified data)の提供サービス、②匿名化ミクロデータ(個票データに匿名化処理が施されたデータ、anonymized microdata)の作成・提供、③オーダーメイドによる集計結果表の提供、④オンデマンドによる集計サービス(リモート集計、remote execution)が展開されている。これらのミクロデータの提供形態は、個別具体的には、各国によって異なる様相を呈しているものの、各国の法制度を踏まえた形で秘匿性のレベルに留意しつつ、利用者のニーズも考慮した上で、様々な形態でデータ提供が行われている(伊藤(2016b), 伊藤(2018b))。

匿名化ミクロデータを作成するためには、各種の匿名化技法が適用される。具体的には、サンプリング(sampling)、リコーディング(recoding)、トップ(ボトム)・コーディング(top (bottom) coding)、データの削除(suppression)といった非攪乱的手法(non-perturbative methods)、さらにはスワッピング(data swapping)のような攪乱的手法(perturbative methods)が適用されてきた[5]。例えば、わが国で現在提供されている国勢調査の匿名データにおいては、１％のサンプリング、都道府県と50万以上市区の地域区分の提供、世帯人員が多い世帯や夫婦の年齢差が大きな世帯の削除、年齢といった属性のリコーディングやトップ・コーディングが行われてきた。また、地域間のスワッピングも適用されている。さらには、匿名データを作成するための実証研究としては、リコーディングにおけるしきい値の設定可能性、ミクロアグリゲーション(micro-aggregation)、スワッピングやPRAM(Post RAndomisation Methods)といった攪乱的手法の有効性の検討が行われた[6]。

---

[5] 攪乱的手法と非攪乱的手法を含む各種の匿名化技法の方法的な特徴については、Domingo-Ferrer and Torra(2001), Duncan et al.(2011), Willenborg and de Waal(2001)、星野(2016)等を参照。また、公的統計ミクロデータの作成の実務の観点から見た攪乱的手法と非攪乱的手法の適用可能性については、伊藤・村田・高野(2014)、伊藤(2019)等を参照されたい。

[6] 匿名データを作成するための数量的な参考資料となることを指向した、公的統計の個票データを用いた実証研究の事例は、以下の通りである。国勢調査を用いたリコーディングにおけるしきい値の設定可能性については伊藤(2018a)、ミクロアグリゲーションの有効性に関する実証研究については伊藤・村田・高野(2014)、スワッピングの方法的な可能性については伊藤・星野(2014)や伊藤(2017)、PRAMの適用可能性と攪乱的手法の有効性についてはIto et al.(2018)をそれぞれ参照されたい[6]。

公的統計ミクロデータの作成・提供にあたっては、どういった適切な匿名化技法を適用すべきかについては、有用性や秘匿性の両方について定量的な評価(Yancey et al.(2002), Shlomo(2010)等)を行うことが必要である。こうした定量的な評価によって、ミクロデータに含まれる情報量の粒度に応じて、どのように匿名化技法を適応すべきかに関する判断材料として有益な情報を提示することが可能になる。

ところで、秘匿性の定量的な評価に着目する、コンピュータサイエンスの分野では、プライバシーに関する様々な基準(metrics)が存在しており、それは、図 1 のように類型化されている。図 1 によれば、測定結果に基づくプライバシーの基準として、「不確実性」、「類似性/多様性」、「敵(adversary)が(個体の特定に)成功する確率」、「情報の利得/損失」、「識別不可能性」、「誤差」、「時間」、「精度/正確さ」の 7 つの基準に類型化することが可能である。「不確実性」については、主にエントロピーに基づく評価基準が含まれている。「類似性/多様性」に関しては、たとえば、k-匿名性(k-anonumity)といった個人情報の保護においてこれまでも議論されてきた評価基準が該当するだけでなく、回帰分析でモデルの説明力の指標として用いられている決定係数も含まれている。また、識別不可能性については、差分プライバシー(differential Privacy)が含まれる。

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(2015，1803 頁))[7]。

差分プライバシーは以下のように定式化される(寺田他(2015))。差分プライバシーでは、あるランダム化関数 Q': $\mathbf{D} \rightarrow \mathbf{R}$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))。

$$Pr[\boldsymbol{Q}'(D_1) \in \boldsymbol{S}] \leq e^{\varepsilon} \cdot Pr[\boldsymbol{Q}'(D_2) \in \boldsymbol{S}]$$

ここで、D1 と D2 (D1, D2 $\in$ $\mathbf{D}$)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 ($\subseteq$ $\mathbf{R}$) は $\mathbf{R}$ の任意の部分集合である。なお、ε(≧0)はランダム化関数 Q'の安全性を示すパラメータ(安全性指標)であり、ε が小さいほど Q'の安全性が高いことを示す。ε=0 のとき Q'は「完全な安全性 (perfect privacy)」を持ち(ただし、このとき Q'の出力からは何の有用な情報も得られない)、ε→∞のとき Q'の安全性は何も保証されない[8]。

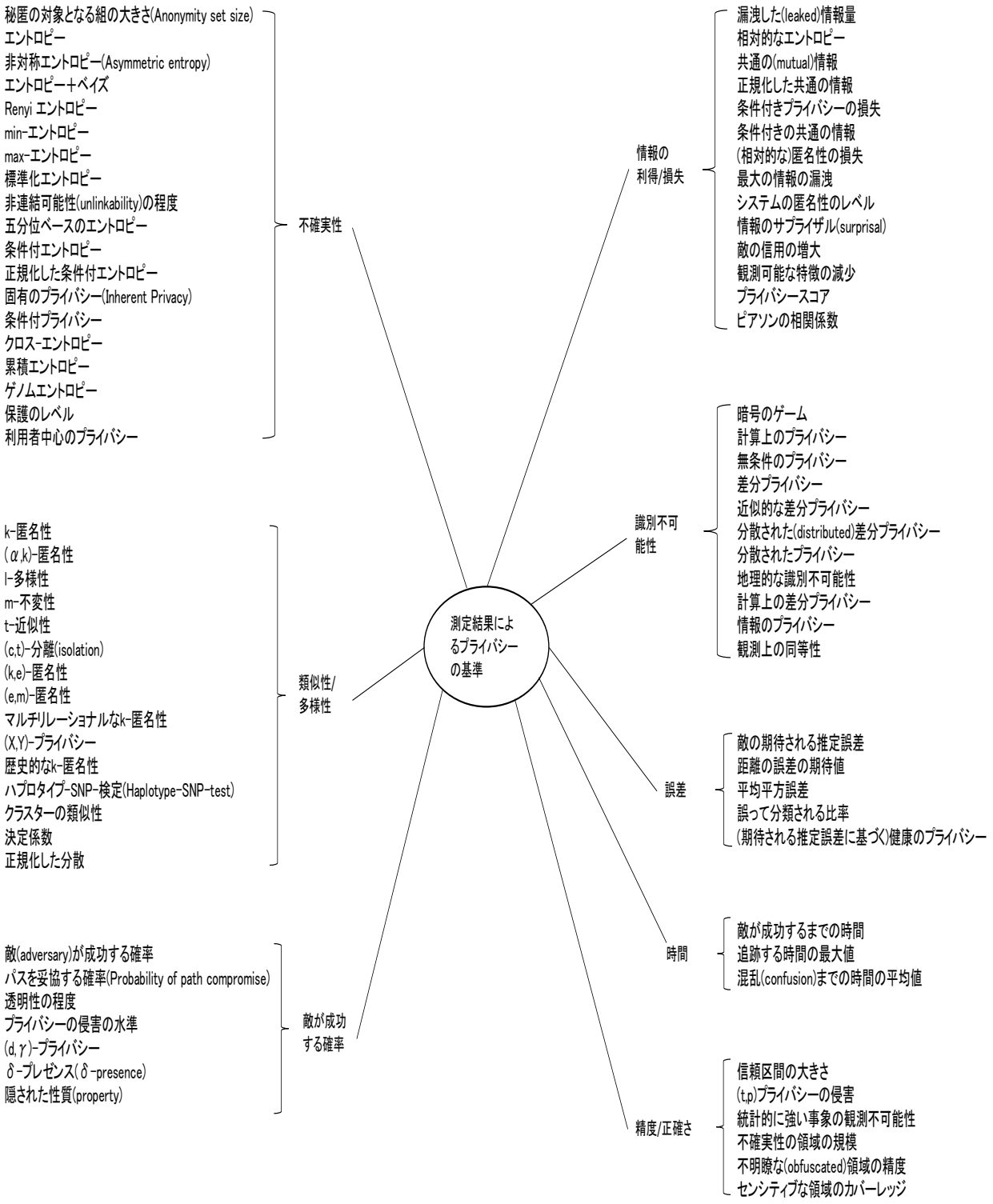
リコーディングといった非攪乱的な方法のみによる秘匿処理が困難になった場合、非攪乱的な方法だけでなくノイズやスワッピング等の伝統的な攪乱的手法を適用するによって、匿名化ミクロデータの作成が可能になる。攪乱的手法の適用の程度によっては、情報量損失が大きくなる可能性があることから、情報量損失の増大を回避しつつ、秘匿性を確

---

[7] 佐久間は、差分プライバシーのメカニズムについて数理的に説明するだけでなく、統計量の公開や、機械学習と差分プライバシーとの関連性を論じている。詳細については、佐久間(2016)の第 7 章～第 9 章を参照。

[8] εの安全性についての説明に関しては、寺田(2019)を参照されたい。

図１　プライバシーの評価基準に関する類型化



出所　Wagner and Eckhoff (2015, p.6)

保するための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。差分プライバシーは高度な攪乱的手法の１つということができ、現在、アメリカセンサス局で試みられている差分プライバシーに基づく合成データの作成方法の検討(Dajani et al.(2017))は、高度な攪乱的手法を用いて作成される高度な匿名化ミクロデータの可能性の１つであるとみなすことができる。

情報工学の分野では、フォーマルな(厳密な)プライバシー(formal privacy)の方法論が議論されてきた。差分プライバシーは formal なプライバシーの１つにすぎないが、差分プライバシーはプライバシーの定義として妥当性を有しており、実用性にも耐えうるものになってきた。そこで、次節では、アメリカセンサス局における差分プライバシーの取り組みについて述べることにしたい。

3. 公的統計の公表に関する差分プライバシーの取り組み―アメリカセンサス局の事例―

アメリカでは、公的統計の分野において差分プライバシーの実用化に向けた議論が進展している。アメリカセンサス局では、2020 年センサスの集計結果の公表およびミクロデータの作成に向けて、差分プライバシーを用いた公的統計作成の一大プロジェクトが展開されている。

Jon Abowd 氏は、集計表における database reconstruction attack を指摘している(Abowd(2018))。database reconstruction attack とは、個票データに含まれる個体情報を暴露するためにクエリを注意深く見なくても、少数のランダムなクエリを組み合わせることによって、クエリの元にある個票データに含まれる個体情報を暴露することができることである(Dinir and Nissim(2003))。この考えから、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考え方が出てきたと言える(Dwork(2006))。

アメリカセンサス局が、例えば 2010 年の人口センサスにおいて公表される統計表や一般公開型ミクロデータ(Public Use Microdata Sample=PUMS)の作成のために用いてきた方法は、情報の削除、トップ・ボトムコーディングといった非攪乱的手法による統計に含まれる情報の低減、およびノイズ付与、スワッピング、部分的な合成データの手法の適用を含む攪乱的手法である。これらの手法を適切に組み合わせることによって、公表可能な統計表や PUMS として提供可能なレベルにまで特定化のリスクを低減していることが知られている(Zayatz(2007), Lauger et al.(2014))。さらには、アメリカセンサス局の場合、人口センサスの統計表の元になる個票データにおいて、スワッピングが適用されてきた(Zayatz(2007))。しかしながら、個体が特定化されないように公表される統計表に含まれるセルの度数に対して削除を行うといったセル秘匿(cell suppression)等の秘匿処理を施したとしても、同一の統計調査で公表された他の統計表群を用いることによって個体情報が漏えいするリスクが存在することが、Abowd によって指摘されている(Abowd(2018))[9]。

[9] 南・阿部(2019)は、セル秘匿を行った集計表においても、セルの度数から個体情報が露見されるリスクが生じることを明らかにしたセル秘匿問題(cell suppression problem)の特徴を述べた上で、マッチング攻撃によって集計表に含まれるセンシティブなセルを推測で

フォーマルなプライバシーでは、プライバシーの数理的な定義から出発する。そして、このフォーマルなプライバシーの定義と一致する形で、個票データを用いてクエリを返すというメカニズムが想定される。この考え方を用いると、統計表は個票データに適用された一連のクエリとしてモデル化される。PUMS についても同様の考え方が適用され、個票データにサンプリングを用いる等、個票データに含まれる個体情報に対して「その場限り(ad hoc)」の秘匿措置を施す代わりに、フォーマルにプライベートな合成データ(formally private data)を作成するためのクエリが用いられる。このような考え方に立って、差分プライバシーの公的統計の適用可能性が追究されるようになった。

Abowd(2018)によれば、統計作成部局は、地域区分が細かい統計表を作成・公表していることから、ミクロデータが公開されてない場合であっても、このような地域区分な詳細な公表された統計表を組み合わせることによって、個体を特定するリスクが高まることが指摘されている。このような特定化のリスクを回避するためには、統計表に含まれる結果数値の精度を考慮しながら、ノイズを付与することによって、安全な統計表を公表すること求められる。その場合、集計表のセルの度数に対してランダムノイズを付与することによって、集計表を提供するオーストラリア統計局の TableBulder のようなオンデマンド型のシステムも存在する。TableBulder の場合、個票データの各レコードにランダムに割り振られた値である Record Key に基づいて、出力される集計表の中のセルの数値に対応するノイズが算出され、そのセルに対してノイズが含まれた出力結果が自動的に付与される(伊藤・谷道・小島(2018))[10]。それに対して、差分プライバシーの方法論を統計の実務レベルで全面的に導入しようとしているのがアメリカセンサス局であって、こうした動きは、ヨーロッパ諸国等の他国とは大きく異なる取り組みだと言える。

アメリカセンサス局における最初の差分プライバシーが適用された匿名化処理のシステムは、On the Map と呼ばれる居住地と勤務先の移動パターンを表す地理的なクエリ応答システムである(Dajani et al.(2017))である。On the Map で用いられる The Longitudinal Employer-Household Dynamics および Origin-Destination Employment Statistics は、合成データという形で設定される。こうした統計実務上の経験も踏まえながら、2020 年の人口センサスの実施に向けて、アメリカセンサス局は、2010 年の人口センサスの個票データを用いて検証を行っている。具体的には、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統計表を対象に、結果数値の精度を確保した上で安全性が保証された統計表を作成・公表可能にするために、差分プライバシーの実用性に関する検証を行っている。具体的には、全国レベル、州レベル、郡レベル、センサストラックレベルおよびセンサスブロックレベルで統計表を作成されていることから、それぞれにプライバシー損失予算(privacy-loss budget)ε を設定し、プライバシーの損失と精度のトレードオフの関係で最適な ε を決定する(Garfinkel et al. (2018))。

きるリスクを高めることを論じている。

[10] 例えば、イギリス国家統計局で現在開発が進められている Flexible Dissemination System においても、TableBuilder の方法論に基づいて、独自の cell key method の実用化が進められている。詳細については、Office for National Statistics(2020)を参照。

2010 年の人口センサスで検証したのは、PL94-171 と呼ばれる人種と地域に関する統計表と SF1 という性別と年齢別のサマリーファイルである。これらの統計表を作成するために用いた具体的な調査事項は、以下のとおりである(Abowd(2018))。

①センサスのトラクトレベルとセンサスブロックレベルのジオコード(15digit)
②性別(男性、女性)
③年齢(0 歳から 114 歳まで各歳年齢区分、および 115 歳以上)
④ヒスパニック系あるいはラテン系か(はい/いいえ)
⑤人種(以下の選択肢の組み合わせで分類区分が設定される)
(1)白人か(はい/いいえ)
(2)黒人かアフリカンアメリカ系か(はい/いいえ)
(3)アジア系か(はい/いいえ)
(4)アメリカンインディアン系あるいはアラスカネイティブ系か(はい/いいえ)
(5)ネイティブのハワイ系か他の太平洋の島々出身か(はい/いいえ)
(6)それ以外の人種か(はい/いいえ)

人種については、「(1)白人か」から「(6)それ以外の人種か」までの 6 つの選択肢に関するすべてを組み合わせることによって、63(=2^6-1)のユニークな人種が設定される。

アメリカセンサス局は、PL94-171 と SF1 を対象に、2010 年の人口センサスを用いて以下の手順で実験を行った(Abowd(2018), (Garfinkel et al. (2018))。最初に全国レベルで集計を行い、数理的に最適化されたプライバシー損失予算 ε に基づいてノイズを付与され、構造的ゼロ(structural zeros)を考慮した形で差分プライベートな統計表が作成される。その構築された全国レベル統計表に対して、個人単位のミクロデータ(3 億 3000 万レコード)が対応している。同様に、州のレベルについても、PL94-171 と SF1 に関して最も精度が高く、構造的ゼロを考慮したクエリとして州レベルの統計表が作成される。そして、数理的に最適化されたプライバシー損失予算 ε に基づいてノイズを付与された差分プライベートな統計表が作成されると、対応する個人単位のミクロデータに対して州レベルの地域区分が付与される。以下、同じような形で、郡レベル、センサストラクトレベル、センサストラックレベルで統計表が作成され、それに対応するレベルの地域区分がミクロデータに付与される。このように各地域レベルの統計表を作成した上で、それに対応するミクロデータとして、州区分、郡区分、センサストラクトの区分、センサストラックの区分のそれぞれの地域における区分が擬似的に付与されたミクロデータが新たに作成される。

アメリカセンサス局が差分プライバシーの適用可能性を検証するために用いているのは、経済学の生産可能性フロンティア(production probability frontier)の考え方に基づく、プライバシーの損失とデータの精度の最適の数値に関する視覚化である。具体的には、推定された最適な社会的便益に関する曲線(Estimated Marginal Social Benefit Curve)を引くことによって、最適な社会的便益(MSB=Marginal Social Benefit)と最適な社会的費用(MSC=Marginal Social Cost)が一致する点で ε を決定することが可能になる。なお、データの有用性の視点から結果数値の精度を重要視するか、あるいは秘匿性の観点

に立ってプライバシーの損失に重点を置くかいずれかによって、最適な社会的便益の曲線の位置が変わってくることに留意する必要がある[11]。

図２は、プライバシーの損失の指標(ε)とデータの精度の指標(I)に関する生産関数を示したものである。ε と I に関する限界変形率(marginal rate of transformation)、すなわち、プライバシー保護を考慮した上でデータの精度を上げる場合に要する限界費用(marginal Cost)が設定される。つぎに、個別主体のプライバシーの損失とデータの精度に関する効用の集計値で表す社会的厚生関数から導出された無差別曲線を描くことによって、生産関数と無差別曲線が接点を持つ最適な ε と I の組を求めることができる。このようにして、社会的限界便益と社会的限界費用が一致する ε と I が決定される。

このようなアメリカセンサス局による差分プライバシーの実用性については、ミネソタ人口センター(Minnesota Population Center)のスティーブン・ラグルス教授らが疑問を呈しており、アメリカセンサス局との間で論争になりつつある。ミネソタ人口センターは、IPUMS(=Integrated Public Use Microdata Sample)の作成・提供を広範に行ってきたことから[12]、ノイズが入った人口センサスのミクロデータに関しては、有用性が低いと考えていることが推察される(Ruggles et al.(2019))。

差分プライバシーの方法論の適用にあたって、人口センサスの作成者側と様々な立場の利用者側との間の対立をどのように解消しているかは、今後の大きな課題と言える。それに対して、アメリカセンサス局において、小地域レベルでの高次元クロス表に対して差分プライバシーを適用し、それをたたみ上げる形でより高次の地域レベルの統計表を作成・公表しようとするプロセスは、ミクロアグリゲーションの展開可能性の１つとみることができる(伊藤(2009))。したがって、アメリカセンサス局における差分プライバシーの方法論の適用の動向については、海外の統計作成部局における反応も含め、今後もその方向に注目していくべきではないかと考える。それは、わが国の公的統計データにおける差分プライバシーに基づく秘匿処理の基準の作成とそのような方法論の適用可能性を検討しようとする場合の参考資料になりうるからである。

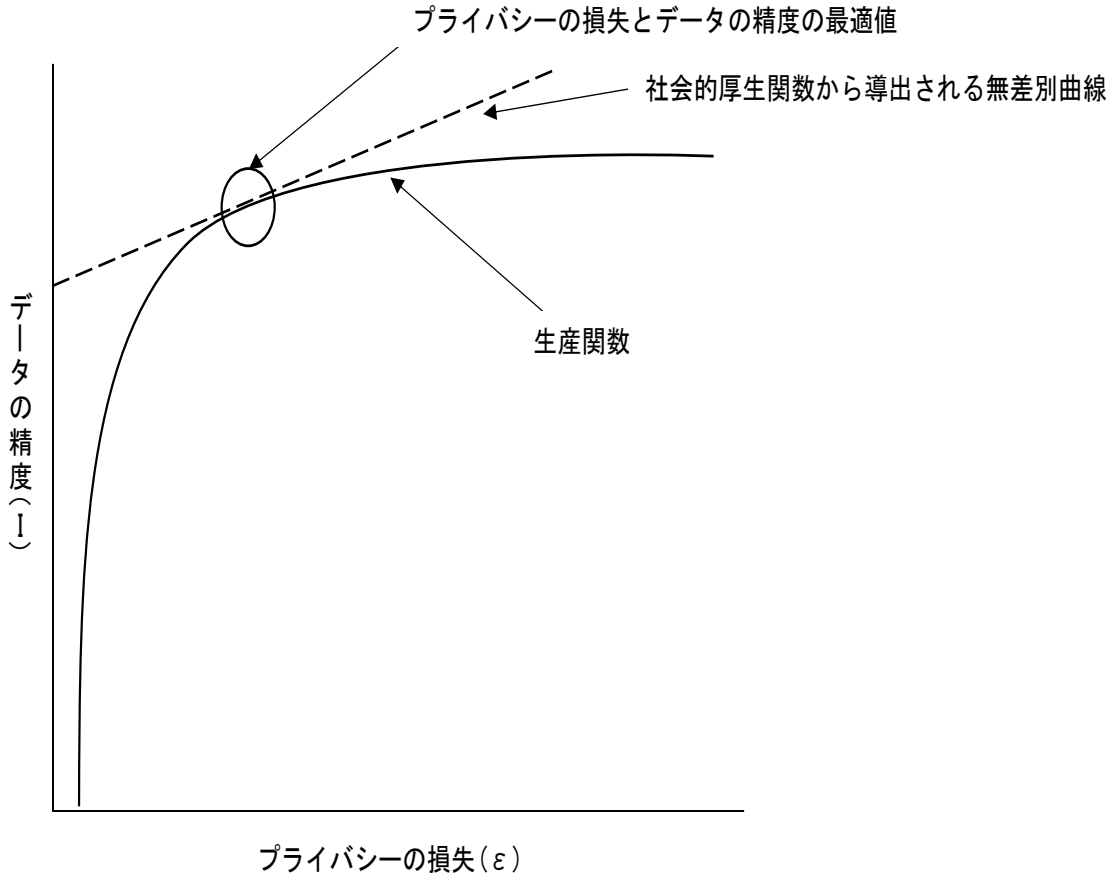
4. むすびにかえて

本稿では、最初にわが国の公的統計の二次利用の現状を概観し、アメリカセンサス局における人口センサスへの差分プライバシーの方法論の適用事例を紹介した。本稿でも明らかにしてきたように、Abowd が指摘した database reconstruction attack は、統計に含まれる個体情報を特定化するリスクを高めるという意味で、公的統計の公表の実務においても今後のさらなる議論の対象になりうると思われる。database reconstruction attack の可能性が出現して以降、その場限りの(ad hoc な)プライバシーの考え方は、個体情報の秘

[11] Abowd(2018)は、アメリカ人口センサスの公表にあたり、プライバシーの損失と精度のトレードオフに関するイメージ図を提示している。。

[12] IPUMS の詳細については、以下の URL を参照。
https://ipums.org/

図２　生産関数から見たプライバシーの損失とデータの精度との関係



注　Abowd and Schmutte (2019)の図１を修正した上で筆者が作成

密保護の観点で十分とは言えないという認識が海外の統計作成部局においても拡がりつつあるというのが現状である。そうした観点から見れば、アメリカセンサス局が、2020 年人口センサスの公表において、データ特性にしたがったプライバシーへの対応から汎用的な手法を適用するフォーマルなプライバシーへの転換を図っていることは、公的統計の作成・公表における今後のターニングポイントになる可能性がある。

差分プライバシーの方法論の可能性は、公的統計の分野においても、今後さらに追究されていると言えるが、本稿の最後に、その１つの試みとして、Ito and Terada(2019)を紹介することにしたい。Ito and Terada(2019)では、「秘密保護ワークセッション」において、差分プライバシーの方法論に基づいて、わが国の国勢調査のメッシュ統計に対して精度の高いラプラスノイズの適用に関する実験内容について研究報告を行っている。具体的には、Laplace メカニズムにより得られた結果数値は、①負の結果数値を多く含みうること、②非構造的なゼロも考慮することによって著しく増大すること、③個々のセルに対して適用されるノイズによって部分総計におけるより大きな誤差をもたらすことから、寺田

他(2015)によって提案された手法に基づいて、ε を変えた上で結果数値にラプラスノイズを適用した場合の結果が紹介された。本研究では、寺田他(2015)によって提案された非負 Wavelet 変換に基づく手法を 2010 年の国勢調査のメッシュ統計データに適用することによって、上記の 3 つの問題点が解決することが確認された。さらに、様々な ε を結果数値に用いた場合の結果を比較としても、単純な Laplace メカニズムを適用した場合、さらに出力に対して負値を 0 に補正した場合と比較して、元のメッシュデータにおける結果数値に対する情報量損失が小さいことが明らかになっている。

海外の統計作成部局からも、差分プライバシー等、情報工学の分野で議論されているリスク評価の方法や基準が近年注目されている。こうした状況を踏まえると、今後も公的統計の大規模なミクロデータを用いて、情報工学の分野で議論されているリスク評価の方法や基準の適用可能性に関する実証研究を模索するのは、有意義であると考えられる。

参考文献

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/jsm-presentation-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.
https://doi.org/10.1257/aer.20170627

Bailie, J., Chien, C. (2019) “ABS Perturbation Methodology Through the Lens of Differential Privacy”, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, The Hague, Netherlands, pp. 1–13.

Dajani, A. N., Lauger, A. D., Singer, P. E., Kifer, D., Reiter, J. P., Machanavajjhala, A., Garfinkel, S. L., Dahl, S. A., Graham, M., Karwa, V., Kim, H., Leclerc, P., Schmutte, I. M., Sexton, W. N., Thompson, K. J., Vilhuber, L., Abowd, J. M.,(2017) “The modernization of statistical disclosure limitation at the US Census Bureau”, Census Scientific Advisory Committee Meetings
https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf

Dinur, I., and Nissim, K. (2003) “Revealing information while preserving privacy”, in Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 202–210. ACM, 2003.

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

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

Dwork, C.(2006) *Differential privacy.* ICALP.
Fraser, B., and Wooton, J. (2005) "A Proposed Method for Confidentialising Tabular Output to Protect against Differencing", Paper Presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, Switzerland, pp. 1–6.
Garfinkel, S., Abowd, J. M., Martindale, C. (2018) "Understanding Database Reconstruction Attack in Public Data: These attacks on statistical databases are no longer a theoretical danger", *acmqueue*, Vol,16, issue 5, pp.1-26.
Heldal, J., Johansen, S., Risnes, Ø.(2019) "Instant Access to Microdata – microdata.no", Paper presented at New Techniques and Technologies for Statistics 2019, Brussels.
星野伸明 (2016)「エビデンスに基づいた匿名化」『日本統計学会誌』, 第 46 巻第 1 号, pp.1-42.
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・村田磨理子・高野正博(2014)「ミ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―」,『統計研究彙報』第 71 号, pp.83-124.
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』第 107 号, pp.1-16.
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, pp.457-478.
伊藤伸介(2016a)「わが国における政府統計のデータシェアリングの現状と課題」『情報管理』, Vol.58, No.11, 836～843 頁
伊藤伸介(2016b)「諸外国における公的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(2018a)「国勢調査における匿名化ミクロデータの作成可能性」『経済志林』, 第 85 巻第 2 号, pp.241-277.
伊藤伸介(2018b) 「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(2018c)「公的統計ミ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, pp.13-18.
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・谷道正太郎・小島健一(2018)「オーストラリアにおける公的統計の二次的利用について―オンデマンド集計システム TableBuilder を中心に―」,『経済学論纂(中央大学)』第 58 巻第 2 号, pp.187-208.
Ito, S., Yoshitake, T., Kikuchi, R., Akutsu, F.(2018) "Comparative Study of the Effectiveness of Perturbative Methods for Creating Official Microdata in Japan" Josep Domingo-Ferrer, J. and Montes, F. (eds.) *Privacy in Statistical Databases: UNESCO*

Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018, Proceedings (Lecture Notes in Computer Science), Springer,pp.200-214.
Ito, S. and Terada, M. (2019) "The potential of anonymization method for creating detailed geographical data in Japan", Paper Presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, The Hague, Netherlands, 2019, pp. 1–14.
Lauger A., Wisniewski, B., McKenna, L. (2014) "Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research", *Research Report Series (Disclosure Avoidance #2014-02)*, U.S. Census Bureau, pp.1-13.
南和宏・阿部穂日 (2019)「表データの最適セル秘匿処理に対するマッチング攻撃とその実証的評価」『コンピュータセキュリティシンポジウム 2019 予稿集』, 1–8.
Office for National Statistics (2020) Flexible dissemination system for Census 2021.
https://consultations.ons.gov.uk/census/initial-view-on-the-2021-census-output-design/supporting_documents/FINAL_Consultation_FDS_4.3ratio.pdf
Ruggles, S., Fitch, C., Magnuson, D., Schroeder, J. (2019) "Differential Privacy and Census Data: Implications for Social and Economic Research", *AEA Papers and Proceedings 2019*, 109, pp.403-408.
https://doi.org/10.1257/pandp.20191107
佐久間淳(2016)『データ解析におけるプライバシー保護』講談社
Shlomo, N. (2010) "Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility", *The Journal of Privacy and Confidentiality*, Vol.2, No.1, pp.73-91.
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』, Vol.56, No.9, pp.1801-1816.
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, pp.58-63.
Thomas, S.(2019) "Successes and Challenges in Increasing Accessibility at Statistics Canada" Paper presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, The Hague, Netherlands, pp.1-9.
Xiao, X., Wang, G., Gehrke, J. and Jefferson, T.(2011) "Differential Privacy via Wavelet Transforms", *IEEE Trans. Knowledge and Data Engineering*, 23(8), pp.1200–1214, IEEE.
Willenborg, L. and de Waal, T.(2001) *Elements of Statistical Disclosure Control*, Springer, New York.
Wagner, I., Eckhoff, D. (2015) "Technical Privacy Metrics: a Systematic Survey"
https://arxiv.org/pdf/1512.00327.pdf
Yancey, W. E., Winkler, W. E., Creecy, R. H.(2002) "Disclosure Risk Assessment in Perturbative Microdata Protection", Research Report Series(Statistics #2002-01), Statistical Research Division U.S. Bureau of the Census.
Zayatz, L. (2007) "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update", *Journal of Official Statistics*, Vol.23, No.2, pp.253-265.