高次元の公的統計データにおけるプライバシー保護をめぐって

中央大・経済　伊藤　伸介　(株)NTT ドコモ　寺田　雅之

諸外国では、学術研究を指向する特定の利用者に限定した個票データや匿名化ミクロデータの提供と、Public Use File の公開という２つの方向から、ミクロデータの作成・提供が行われている(伊藤(2018a))。わが国でも 2018 年 5 月に「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「改正統計法」と呼称)が成立したことから、公的統計のミクロデータの作成・提供に関する総務省令やガイドラインが、今後整備されると考えられる。改正統計法の第 33 条および第 36 条に明記されている「相当の公益性」に関しては、「客観的にみて合理的ないしはふさわしい」公益性の範囲について、利用目的に応じた形での個別具体的な検討が求められるが(伊藤(2018b))、秘匿性のレベルに留意しつつ、利用者のニーズも踏まえた上で、データ提供に関する様々な形態が模索されるものと考える。

一方、2017 年 5 月の改正個人情報保護法の全面施行に伴い、パーソナルデータを含む大規模データ(ビックデータ)の利活用が広範に進められている。個人の購買履歴や位置情報等に関する大規模データの第三者提供においては、何らかの匿名化の基準にしたがって、データに含まれる個体情報の特定性や識別性の低減を可能にするための匿名化措置が求められる。こうした匿名化措置の適用可能性を追究するためには、諸外国における大規模データの提供の状況や匿名化技術の動向を踏まえた上で、主要な匿名化技術に関する特徴およびそのメリットとデメリットを把握する必要がある。その意味では、公的統計におけるミクロデータ提供や匿名化技術の適用可能性を追究することは、ビックデータにおける匿名加工情報の作成においても参考になると思われる。

公的統計は、統計表とミクロデータという２つの形態によって作成・提供されているが、これらについては、「超高次元クロス集計表」という概念を介して関連付けることができ、統計表に含まれるセルの数とセルに含まれるレコード数の観点から、統計表とミクロデータが位置付けられる。変数における区分の統合やセルに含まれる度数の削除といった加工方法を用いることによって、公表可能な統計表が作成される。統計表に含まれるセル数が少なければ、レコード数の大小にかかわらず、統計表の形での提供が可能になる。セル数が増大するほど、セルに含まれるレコード数が小さくなり、度数 1 となるセルの数が増加することから、高次元の統計表は、数理的形式的にはミクロレベルのデータと同様のデータ構造を有すると考えられる。

こうした高次元の公的統計データに対して、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)攪乱的手法によって、Public Use File や匿名化ミクロデータの作成・提供が可能になる。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。例えば、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(synthetic data)の作成方法の検討は、その１つだと言える。こうした合成データの作成可能性の追究は、匿名化ミクロデータのさらなる展開の方向性を提示している。

本報告では、主要な匿名化技術に関する特徴およびそのメリットとデメリットを明らかにした上で、公的統計を含む大規模な高次元データへの適用可能性の検討を行うことによって、高次元の公的統計データにおけるプライバシー保護のあり方について議論していきたい。

参考文献

伊藤伸介(2018a)「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,77～101 頁

伊藤伸介(2018b)「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,13～18 頁