厚生労働科学研究費補助金（がん対策推進総合研究事業）
分担研究報告書
匿名化手法の検討・評価に関する研究

研究分担者　南　和宏　統計数理研究所教授

研究要旨: 匿名化処理の前処理としてランダムサンプリングを適用し、厳格な安全性規準である差分プライバシーを実現する手法を検討した。実証実験の結果、サンプリング率を適切に調整することにより、匿名データの安全性と有用性のバランスを保持することが可能であることが示された。

Ａ．研究目的

匿名化はがん登録情報に含まれる調査客体の識別を防止するために有効なプライバシー保護技術であるが、データの有用性を低下させる課題が存在するため、匿名化の強度の適切な調整が必要になる。本研究では、全国がん登録情報の利用者に対する代表的な安全管理装置を検討し、利用者の利用環境に応じて匿名化データの安全性強度を調整する手法を検討する。

Ｂ．研究方法

匿名データに厳格な差分プライバシーに基づく安全性を保証するため、匿名化処理の前処理として、ベルヌーイサンプリングを実施する手法を検討した。差分プライバシーは対象レコードが元データに含まれるかどうかの識別を確率的に困難にする概念であり、サンプリングの導入によりそのための識別不可能性を実現する。本研究では、代表的な匿名化アルゴリズムである Incognite を R 言語で実装し、サンプリングの前処理を行った匿名データの有用性に関する実証実験を実施した。具体的には、US センサスの Adult データセットとがん登録情報に対して検討手法を適用し、差分プライバシーのプライバシー予算 ε、匿名化のプライバシーパラメータ k、サンプリング率 β とデータの有用性指標である discernibility との関係性を網羅的に調べた。

Ｃ．研究結果

従来研究では、データの有用性を保持するために選択できるサンプリング率の上限の値を用いていたが、そのパラメータ設定では安全性の要件を緩めても匿名データの有用性が大幅に低減する逆説的な状況が生じた。この有用性低下の問題を解決するため、サンプリング率の値を下げたところ、同程度の安全性を維持しつつ、データの有用性が向上することが確認できた。さらにデータサイズが大きいがん登録情報の場合、低いプライバシー予算 ε を設定による強い安全性強度とデータの有用性保持が両立可能であることが示された。

E. 結論

ランダムサンプリングは匿名データの安全性強度を高める手法として有望であり、レコード数が多いがん登録情報の場合、低いサンプリング率による情報損失が相対的に少なく、安全性と有用性の両立の可能性を示すことができた。

Ｆ．健康危険情報

なし。

Ｇ．研究発表

1.　論文発表

特になし。

2.　学会発表

1) 杉山拓海, 南和宏. Empirical evaluation of anonymized data with ARX Anonymization Tool. 第 102 回コンピュータセキュリティ・第 52 回セキュリティ心理学とトラスト合同研究発表会 2023 年 7 月 25 日.

2) 南 和宏, 杉山拓海. 匿名化データに対する差分プライバシー適用の検討. 2023 年度統計関連学会連合大会 2023 年 9 月 5 日.

3) 南 和宏. ランダムサンプリングによる差分プライベートな度数表の検討. 研究集会「大規模データの公開におけるプライバシー保護の理論と応用」 2023 年 12 月 8 日.

4) Yutaka Abe, Kazuhiro Minami. A Case Study of Output Checking in Japan. UNECE Expert Meeting on Statistical Data Confidentiality 2023 2023 年 9 月.

5) Takumi Sugiyama, Kazuhiro Minami. Differentially Private Frequency Tables Based on Random Sampling. 2023 IEEE International Conference on BigData, 2023 年 12 月.