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
研究分担者　東　尚弘　国立がん研究センターがん対策研究所がん登録センターセンター長
研究分担者　柴田　亜希子　国立がん研究センターがん対策研究所がん登録センター全国がん登録室室長
研究分担者　祖父江友孝　大阪大学大学院医学系研究科環境医学分野　教授

研究要旨: 本研究では、がん登録情報を対象とする匿名化手法の確立を目指す。特に、外観識別性の高い地域情報の一般化処理に着目し、２種類の匿名化アルゴリズムを実装した。評価実験の結果、通常の地域レベル（都道府県、市区町村、町丁目）による一般化処理では人口密度が大きく異なる様々な地域を適切にグループ化することは困難であり、GPS 位置情報の基準点に基づく地域分割の有効性を示すことができた。

A．研究目的

全国がん登録の情報には、医療機関の受診者に関する機密情報が含まれており、がん登録情報を用いた調査研究を行う際に、匿名化データからの機密情報が外部に漏洩しないような安全性の担保が必要である。現在、匿名化データの代表的な安全性指標として、k-匿名性および、その派生指標が多く提案されているが、がん登録情報に対して具体的にどの手法を選択すべきかその要件は明らかでない。本研究では、匿名化手法の安全性の評価手法を確立し、その要件を満足する匿名化手法の確立を目指す。

B．研究方法

本年度は具体的な k-匿名化アルゴリズムを実装し、特に外観識別性の高い地域情報に着目し、一般化処理に関する実証評価を行った。年齢、性別、住所情報の３つを準識別子情報とする k-匿名化アルゴリズムを検討し、特に識別リスクの高い住所情報の一般化処理に着目して、２種類の k-匿名化アルゴリズムを考案、実装した。１つ目のアルゴリズムは、住所情報を地域レベル（都道府県、市区町村、町丁目）に基づき一般化処理を行う方式であり、必要に応じて同じレベルの地域情報の統合を行う。２つ目は、前処理として、がん登録情報の住所情報に国土交通省の公開する GPS 位置情報を紐付け、GPS 位置情報の基準点に基づき、広域レベルの地域情報の２分割を再帰的に繰り返すトップダウン型の匿名化アルゴリズムである。２つのアルゴリズムが生成する k-匿名化グループのレコード数のばらつきを比較した。

C．研究結果

一つめの地域レベルの一般化による匿名化アルゴリズムは、効率的に動作するものの地域の人口密度を反映した柔軟なグループ化が困難であり、都市部では安全性のパラメータである k 値を大きく上回るグループが多く生成され、また人口の少ない地域では k 値を満たすグループが作成されず、多くのレコードが削除される結果となった。この結果を踏まえ、地域情報に国土交通省の位置参照情報に含まれる町丁目レベルの GPS 座標を付与し、GPS 基準点に基づく動的な領域分割処理を行ったところ、匿名化グループのサイズの均一化について大幅な改善が実現できた。

D．考察

GPS 位置情報を参照することで、適切な粒度の地域情報の定義が可能となったが、一方市区町村の境界を超えた地域グループも現状定義されている。このような地域情報が分析者のニーズを満たすものであるかの検討は必要であり、匿名化処理された地域情報の可視化方法を検討する予定であ

る。

E. 結論

地域情報の匿名化処理において、地域の隣接性を考慮できる GPS 位置情報を用いた領域分割の手法は、匿名化データの粒度であるグループサイズの均一化に有効な手法であることが確認できた。

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.南和宏. 公的ミクロデータに対する k-匿名化加工の検討. 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」. 2021 年 12 月 10 日.

2.南和宏. 表データの最適秘匿処理に対するマッチング攻撃とその防御手法の検討. 2021 年度統計関連学会連合大会, 2021 年 9 月 7 日.