公的ミクロデータに対するk-匿名化加工の検討

南　和宏　データ科学研究系 教授

公的統計における匿名データ

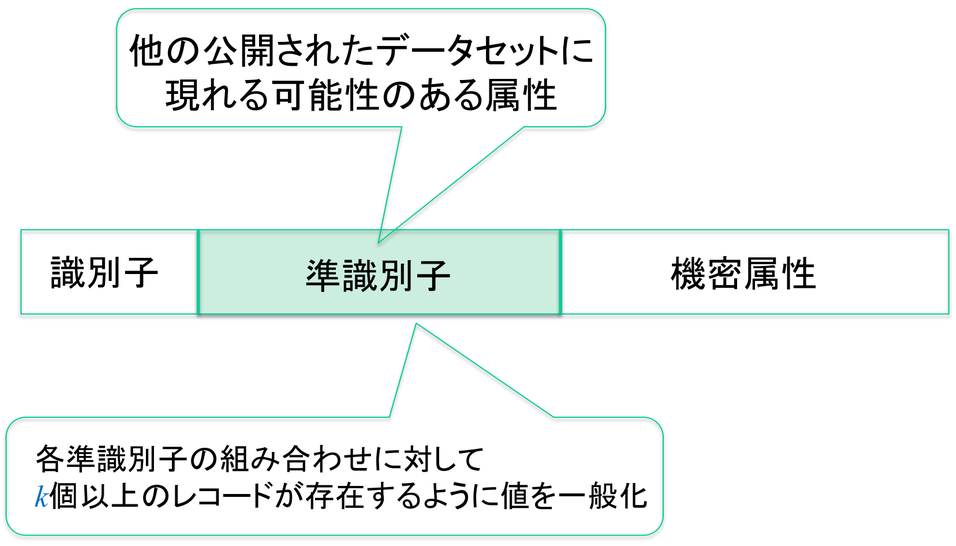
一般の利用に供することを目的として調査票情報を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む。）ができないように加工したものをいう。(統計法第2条第12項)

↓

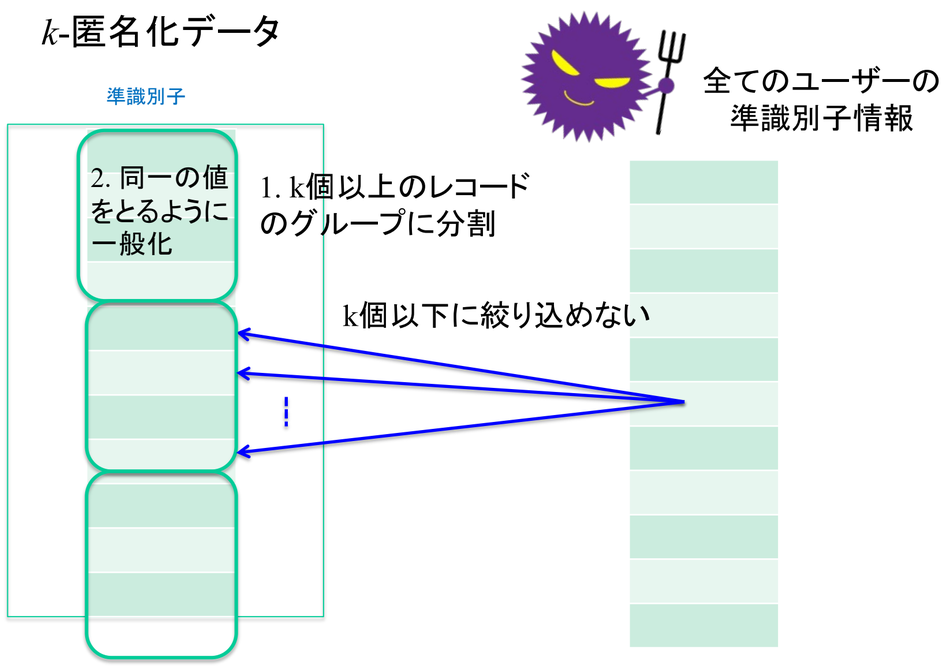
レコード識別の防止が要件

k-匿名化の概念

アイデア：必ず同一の準識別子の値を取るレコードをk個以上作成し、レコード識別をk個未満に絞り込ませない



k-匿名化によるレコード識別防止



k-匿名化の実施上の課題

- 安全性パラメータkをどう選ぶか？
  - kの値が大きければ安全になるが、情報損失は増大
- どの変数（属性）を準識別子とするか？
  - 攻撃者の外部知識を適切に想定する必要がある
- k-匿名化データ作成のためにどのような一般化処理（グループ化）を行うか？

↓

がん登録情報に適した匿名化手法を検討

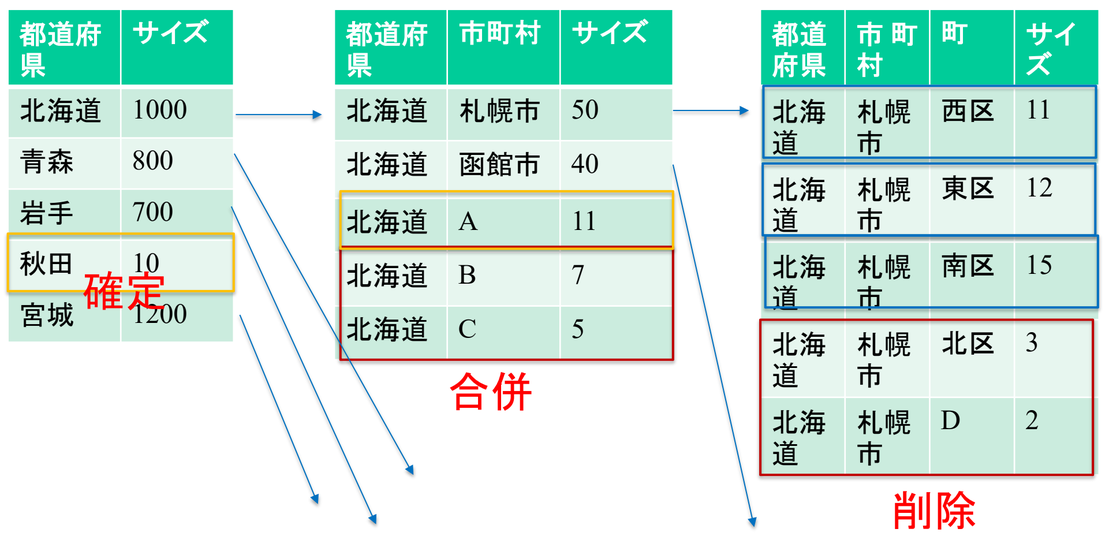
匿名データの作成

- R言語で準識別子属性のグループ化を行う関数の開発
  - 性別
  - 診断時年齢
  - 住所情報
- 安全のパラメータは $k=10$ に設定
  - 小地域による匿名化の可能性を試行
- 地域情報のトップダウンの自動分割により、柔軟なグループ化を目指す
  - 都道府県 → 市区町村→町丁目

一般化処理の手順

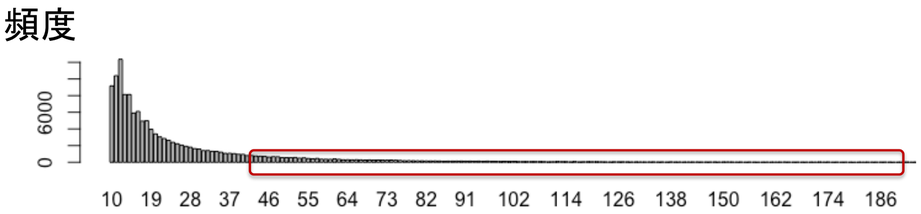
- 住所情報は一番一般化された「都道府県」レベルからスタートし、等価クラスのサイズはk未満にならない程度に市レベル、町レベルに分割していく
- 各レベルの情報の粒度には大きなギャップがあるので、それを埋めるための中間的なレベルとして、同レベルの住所情報のマージも行う
  - マージは同じ親をもつ同レベルの情報を無作為に選択 ➔ 今後改良の予定

トップダウン型分割アルゴリズムの概要



等価クラスのサイズの頻度分布

大部分は、10を少し超える値だが、かなり大きなグループも発生しており、都市部の町は丁目レベルの分割をすべきであった



等価クラスのサイズ

まとめと今後の課題

- 地域情報の柔軟な自動分割を目指す匿名化プログラムを実装し、がん登録情報に適用
- 地域によるレコード数のばらつきは大きく、k=10の場合、グループ化できないレコードが発生
  - この問題は、性別、年齢等の他の変数の組み合わせを考慮した場合により深刻な問題になる
- 地域情報のマージは有望な手法であるが、GPS位置情報等の参照による地理的な隣接性の考慮が必要
  - 現在はGPS位置座標に基づく地域分割を行うMondrianアルゴリズムを開発