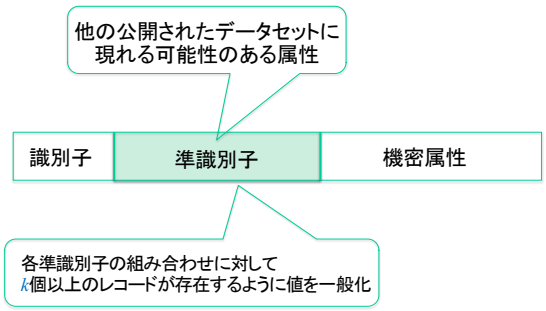
位置座標に基づく地域情報のk-匿名化手法の検討

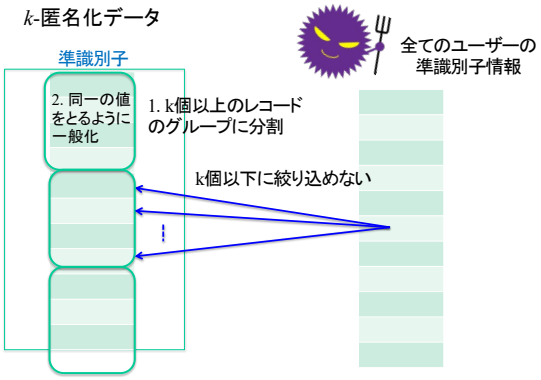
南　和宏　データ科学研究系 教授

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

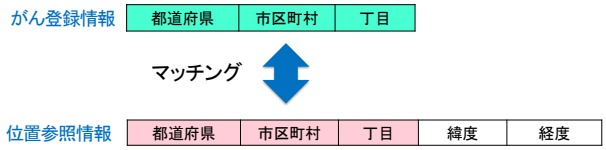
地域情報の一般化処理の問題点

地域間の人口密度の違いは大きく、地域レベルによる一般化では、等価クラスの粒度を均一にするのが困難

位置座標に基づく、地域領域の柔軟な分割

国土交通省位置参照情報とのリンケージ

都道府県、市区町村、丁目の値が同じレコードをMySQLで連結



k-匿名化データの作成と評価

R言語でGPS座標値に対するMondrianアルゴリズムをR言語で実装

地域情報のトップダウンの再帰的分割により、柔軟な地域グループの作成を目指す

安全性パラメータを $k=500,1000,2000$ に設定し、住所情報の等価クラスを作成

小地域による匿名化の可能性を試行

データの有用性をDiscernibility指標で評価

地域レベルの一般化による匿名化手法との比較

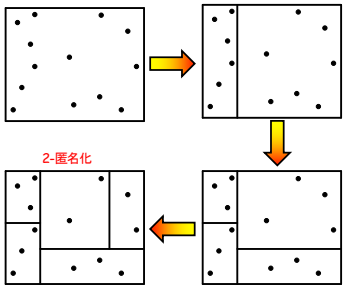
Mondrianアルゴリズム

- 今回の基準点は、領域内のGPS座標の中央値を使用
- 東西、南北の方向に交互に分割

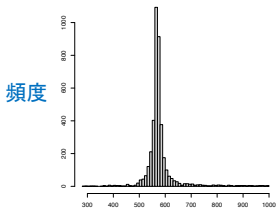
入力：領域L
　　領域上の点集合S
　　しきい値k
出力：部分領域の集合

1.領域Lを分割する次元（属性）を決める
2. その次元を軸に中間値で分割
3. 分割された2つの領域$L1$, $L2$について同様の処理を繰り返す。
4. 全ての分割がk個未満の部分領域を生成するとき終了

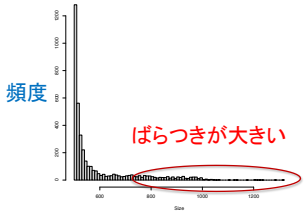
$$\forall i : k \leq |L_i| \leq 2k-1$$



等価クラスのサイズの頻度分布 (k=500) の比較



提案手法（Mondrian）



地域レベルに基づくk-匿名化

Discernibility指標によるデータ効用の比較