DEIM Forum 2014 E5-5

# Pk-匿名化手法の一改良法の検討

柿澤　美穂† 渡辺知恵美†† 古川　諒††† 高橋　翼†††

† お茶の水女子大学大学院人間文化創成科学研究科　〒 112–8610 東京都文京区大塚 2–1–1
†† 筑波大学システム情報工学研究科　〒 305–8573 茨城県つくば市天王台 1–1–1
††† NEC クラウドシステム研究所　〒 211–8666 神奈川県川崎市中原区下沼部 1753
E-mail: †kakizawa.miho@is.ocha.ac.jp, ††chiemi@cs.tsukuba.ac.jp,
†††r-furukawa@cb.jp.nec.com, †††t-takahashi@nk.jp.nec.com

**あらまし**　ランダム化を用いた匿名化手法の一つに Pk-匿名化 [1] [2] [3] がある．Pk-匿名化の既存手法は, k-匿名性を確率的指標に拡張した Pk-匿名性を保証するために, データ主体を 1/k 以上の確信度に絞り込めないように属性値にノイズを付与する．既存手法は, 元の属性値にラプラス分布に従ったノイズを付与することで Pk-匿名化を実現し, 所望の k の下で Pk-匿名性を満たすようにラプラス分布の分散を決定している．本稿では, より小さい分散で Pk-匿名化を実現するよう改良したアルゴリズムを提案する. 提案手法を既存手法と比較, 評価することで提案手法の優位性を示す.
**キーワード**　データベース, プライバシ保護, 匿名化

## 1. はじめに

近年, データベースサービスの普及に伴い, 個人情報等の機密情報をデータベースに格納する際にプライバシ保護が要求されている．特に, データベースに格納された機密情報を公開する際, データ公開者はデータベースのレコード所持者をデータ利用者に特定させずに公開したいと望む．そのような場合にレコード所持者を隠すため, データを匿名化する手法が研究されている．データ匿名化の一手法として, k-匿名化 [7] [5] がある．k-匿名化とは, 属性値の抽象化, 削除等を行うことによって, レコード所持者を k 人未満に絞れないようにする手法である．この k-匿名化を確率的指標に拡張した手法として, Pk-匿名化 [1] [2] [3] がある．Pk-匿名化は, 属性値の置換やノイズ付与といった確率的な操作を用いて, レコード所持者を 1/k 以上の確信度で絞り込めないようにする．既存研究では, 数値属性に対してラプラス分布に従うノイズを付与することで Pk-匿名化を実現する方法が提案されている．

既存の Pk-匿名化では, ラプラス分布の分散値を決定する際に, 属性値間の最大距離とレコード数を用いる．しかし, 属性値の分布によっては, この方法では過剰にノイズが付与され匿名化が過剰になされる場合がある．例えば, レコードの属性値の分布に特徴があり, いくつかのグループに明確に分類できるような場合を考える．つまり各グループ内のレコード間の類似度が高く, グループ間の距離が大きく離れている場合である．このようなデータに対して Pk-匿名化を適用する際, レコード全体に対してラプラス分布に従うノイズを適用すると, ラプラス分布の分散が大きくなり匿名化前に見られていたデータの特徴が失われてしまう場合がある．

そこで我々は, ラプラス分布に基づくノイズ適用による Pk-匿名化手法において, 匿名化前のデータの特徴をできるだけ維持するための改良法を提案する．基本的なアイデアとして, 匿名化前のデータの類似性を考慮していくつかのグループに分割し, 各グループの属性値集合に対して Pk-匿名化を適用する．上記のような例の場合, 各グループに Pk-匿名化を適用した際のラプラス分布に従うノイズの分散が, 既存手法に比べて小さく抑えられる．

本稿では, 第 2 章にて Pk-匿名化について例を用いて説明し, 第 3 章でラプラス分布に従うノイズを適用した Pk-匿名化手法が持つ問題点を指摘する．第 4 章で提案手法について述べ, 第 5 章でラプラス分布に従うノイズの分散を抑制するためのデータ分割法に関しての考察を行う．第 6 章で, 属性値の分布が二つのグループに明確に分けられる属性値集合を例に提案手法を適用し, 既存手法に比べてラプラス分布に従うノイズの分散が抑制されることを示す.

## 2. Pk-匿名化

Pk-匿名性は, k-匿名性の ”データの持ち主を k 人未満に絞り込めない ”という直感的な指標を確率的に拡張し, ”データの持ち主を 1/k 以上の確信度に絞り込むことができない ”ことを保証する．k-匿名性はランダム化を用いた匿名化手法に適用できなかったが, Pk-匿名性はランダム化に対応した手法であり, ランダム化の一手法である維持-置換攪乱 [8] において Pk-匿名性が実現可能であることが証明されている. [1]

例として, ある病院の患者の氏名, 住所の緯度, 経度, 病名が格納されているデータセットがあり, これを Pk-匿名化し, 年齢, 体重と病気の関係について第三者に分析を依頼することを想定する．攻撃者はある人物の病名の特定を目的とし, 匿名化後のデータを何らかの手段で入手したとする．この時攻撃者は, 背景知識として対象者の体重と年齢を知っており, このデータをもとに対象者のレコードを特定して, 病名を突き止めようとする．Pk-匿名化は, レコードの属性値に確率的なノイズを加算して攪乱させることによって, 対象者のデータ特定の確信度を 1/k 以

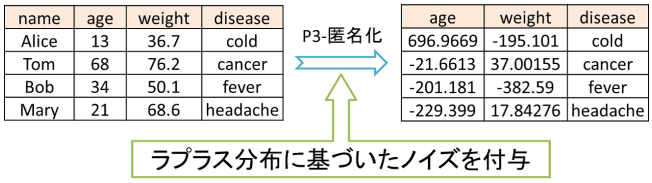


図 1: Pk-匿名化

下にする. 匿名化後のデータセットは図 1 の右に示すような値となり, 攻撃者が対象者の体重と年齢をあらかじめ知っていたとしても, それが匿名化後の値のどれと対応するのかを 1/k 以上の確信度で突き止めることができない.

さらに, この Pk-匿名性を保証する処理を Pk-匿名化と呼ぶ. 文献 [2] では, 数値属性において Pk-匿名化を実現する手法として, レコードに対しラプラス分布に従うノイズを付与する手法を提案している. ラプラス分布に従うノイズを付与すると, 匿名化後の属性値が確率的に存在するようになる. これより, ある属性値 $A$ を Pk-匿名化した値を $A'$ とすると,「$A$ を Pk-匿名化した値が $A'$ である」と確信する確率を $1/k$ 以下に抑えることができる. ここでは, k を下記の式で与えられる値として Pk-匿名化であると証明している.

$$k = 1 + (|R| - 1) \prod_{V:\text{属性}} exp(-2\frac{sup_{u,v \in V}||u - v||_1}{\sigma}) \quad (1)$$

$\sigma$ はラプラス分布の分散, $|R|$ はレコード数を表す. $sup_{u,v \in V}||u - v||_1$ は, 各属性値の絶対値の総和, すなわち属性値間の最大距離を表す. 所望の k の値の下で, $\sigma$ の値を式 (1) で決定する.

複数属性の場合は, まず各属性の値域を, 属性毎に [0,1] の範囲に正規化する. これは, ラプラス分布の分散の値域を属性間で揃え, 属性間でノイズの強度が異なるのを防ぐためである. 正規化した後, 式 (1) で $\sigma$ の値を算出し, $\sigma$ の値に基づいたランダムなノイズを各属性値に付与する. 最後に, 正規化の逆変換で値のスケールを元に戻す.

## 3. 既存手法の課題

### 3.1 予備実験による分析

Pk-匿名化はラプラス分布に従うノイズを付与することで実現される. そこで我々は, ノイズ付与による属性値の分布の広がり方を調べるため, 疑似データを用いて予備実験を行った.

初めに, あるデータに k-匿名化を施した場合と Pk-匿名化を施した場合の属性間の相関係数を比較する. 相関係数を用いて比較するのは, 匿名化後の値が元の値の特徴をどれだけ保持して匿名化を実現できているかを確かめるためである.

ここで, 2 属性間の相関係数が 0.5 と 1.0 である疑似データを準備する. k-匿名化後のデータは, 各データに Mondrian [9] をベースにした k-匿名化を施し, 匿名化後の値の範囲の中で一様分布に従って値を決めたものとする. 結果は以下の図のようになった. 相関係数 0.5 を持つデータを各匿名化処理した場合の相関係数の推移を表したものが図 2, 同様に相関係数 1.0 のデータの場合が図 3 である.

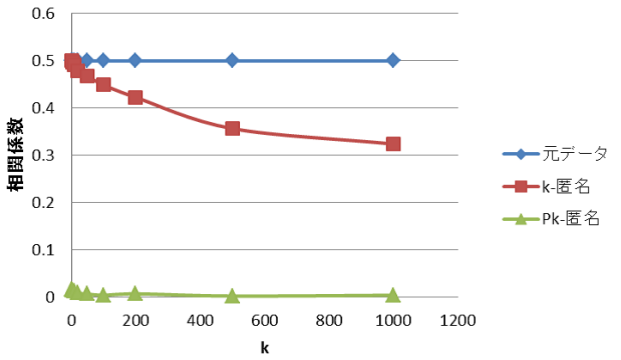


図 2: 相関係数 0.5 のデータを各匿名化処理した時の相関係数の推移

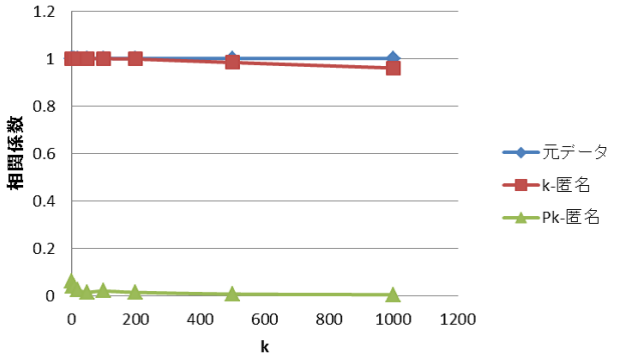


図 3: 相関係数 1.0 のデータを各匿名化処理した時の相関係数の推移

図 2,3 から読み取れるように, Pk-匿名化した場合の相関係数は, 元の相関係数から大きく離れてしまっている. k-匿名化した場合の相関係数は元の相関係数と大きく変わっていないことから, k-匿名化は元の属性値の特徴をある程度保持して匿名化できる手法であることが言える. それに比べて Pk-匿名化をした場合は, 元の相関係数から大きく離れることから, 匿名化後の値が元の属性値から大きく広がって分布する可能性が考えられる.

これを確認するため, 上記の疑似データを匿名化した場合の属性値の分布を可視化する. 各データの Pk-匿名化前の属性値の分布を示したものが図 4 と図 6, この疑似データを $k = 10$ として Pk-匿名化した結果の分布を示したものが図 5 と図 7 である.

匿名化後のデータはラプラス分布に従って分散していることが見て取れる. また, 匿名化後の値の分布は元の属性値の分布とは大きく異なっていることが分かる. 匿名化前後で値の分布が大きく異なるという結果から, ラプラス分布の分散が大きいために過剰なノイズが付与されている可能性が考えられる.

### 3.2 解析的考察

第 3.1 節で, 既存の Pk-匿名化の問題点を実験を用いて示した. そこで今度は, Pk-匿名化の問題点を数式で解析的に考察する. 既存手法で提案されているラプラス分布の分散決定手順では, まず各属性の値域が $[0, 1]$ になるように正規化を行った後, ノイズを付与し正規化した値を元のスケールに戻す. この過程

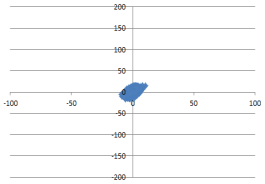

図 4: 相関係数 0.5 : 匿名化前

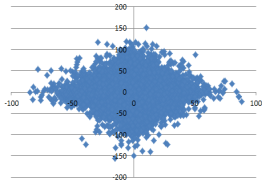

図 5: 相関係数 0.5 : 匿名化後

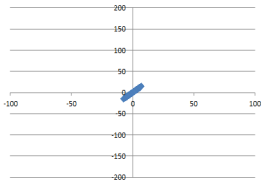

図 6: 相関係数 1.0 : 匿名化前

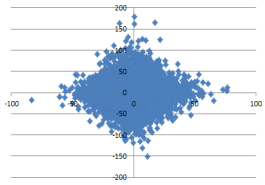

図 7: 相関係数 1.0 : 匿名化後

において, ノイズを付与する場合に Pk-匿名化を満たすためのラプラス分布の分散 $\sigma$ を求める.

まず, 第 2 章で説明した Pk-匿名化の k の算出式を再掲する.

$$k = 1 + (|R| - 1) \prod_{V:\text{属性}} exp(-2\frac{sup_{u,v\in V}||u - v||_1}{\sigma})$$

正規化後の属性値に対する $\sigma$ の値を $\sigma'$ とし, 属性数を $n$ とおく. 正規化を適用すると各属性の値域は $[0, 1]$ になるので, $sup_{u,v\in V}||u - v||_1$ は必ず 1 となる. したがって, 正規化した属性値における k の算出式は,

$$k = 1 + (|R| - 1)exp(-\frac{2n}{\sigma'}) \tag{2}$$

となり, この式から $\sigma'$ は下記の式で導出できる.

$$\sigma' = 2\frac{n}{log\frac{|R|-1}{k-1}}$$

また, 各属性の元の値域を $d_i$ とすると, 正規化された状態からスケールを元に戻すには $\sigma'$ に $d_i$ 倍することになる. したがって, 各属性に適用するラプラス分布に従うノイズの分散を $\sigma_i$ とすると,

$$\sigma_i = 2\frac{d_i n}{log\frac{|R|-1}{k-1}}$$

となる.

ここで例として $|R| = 14842$, $k = 101$, $n = 2$ を代入し, ノイズを加えた属性値がどの程度分散するか計算してみる.

$$\sigma_i = \frac{4}{5}d_i$$

この結果より, あるレコードにラプラス分布に従うノイズを加えた場合, データ全体の値域の半分以上の広さで分散することがわかる.

## 4. 提 案 手 法

既存の Pk-匿名化手法の問題点として, 匿名化後の属性値が大きく広がって分布し, データの有用性が維持できない可能性を示した. 過剰なノイズが付与されると, 元のデータの特徴が失われデータの有用性が低くなるため, できるだけ高い有用性を保持して Pk-匿名化するには, ラプラス分布の分散を小さくすることが望まれる.

そこで本研究では, ラプラス分布の分散を小さくし, ノイズの過剰付与を防ぐ手法を提案する.

まず, 第 2 章の既存の k の算出式 (1) を変形すると, レコード数と所望の k の値を固定した際, $\sigma$ の値は属性値間の最大距離 $sup_{u,v\in V}||u - v||_1$ に比例して大きくなることが分かる.

$$\sigma = 2\frac{sup_{u,v\in V}||u - v||_1}{log(|R| - 1) - log(k - 1)} \tag{3}$$

元の属性値がある程度均等に分布している場合には, 全ての属性値間の最大距離を用いて $\sigma$ の値を決定する. また, 例えば元の属性値が大きく二つに分類された分布を持つ場合も, 値が分布していない空間を含んでいるにも関わらず, 全ての属性値間の最大距離を用いて $\sigma$ の値を決定する.

図 8 のようにデータがいくつかのまとまりとなって分布している場合, 全ての属性値をまとめて Pk-匿名化を行うより, 纏まって分布しているグループ毎に Pk-匿名化する方が分散を小さくでき, かつ過度のノイズを付与することがないと考えられる. そこで我々は, 属性値の分布に関わらず属性値をいくつかのグループにあらかじめ分割し, グループ毎に Pk-匿名化を行うという改良法を提案する (図 8).

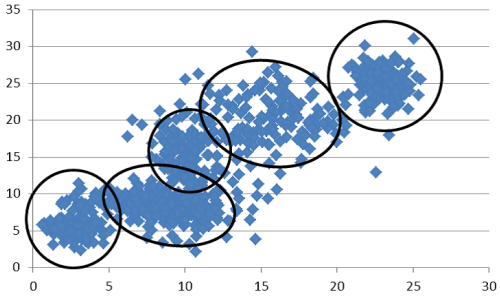


図 8: 提案手法 : 全体の属性値をグループに分割

Pk-匿名化は, 所望の k の値から匿名化後の値の存在を確率的に与えるため, ラプラス分布の分散が属性値毎に異なっても k の値は保たれる. したがって, 全体の属性値を分割して Pk-匿名化することが可能である. 全体の属性値をグループに分割してグループ毎に匿名化を行うことで, ラプラス分布に従うノイズが過度に付与されることを抑えられ, それに伴い匿名化後の値が元の属性値の特徴を維持することが可能になると期待できる.

我々の提案手法は, ある一つの属性値とその周辺に分布している属性値をまとめて一つのグループとし, そのグループ毎に $\sigma$ の値を算出してノイズを付与するという手法である. これにより, 過剰のノイズが付与されることを防ぐ.

## 5. ラプラス分布の分散と属性値のグループ分割の最適化

全体の属性値をあらかじめ分割してから Pk-匿名化を施すことによって, ラプラス分布の分散を小さくして過剰なノイズ付与を防ぐことができる. そこで, グループ分割方法の有用性, ラ

プラス分布の分散と属性値のグループ分割の最適化の方法を考える.

### 5.1 属性値のグループに対して Pk-匿名化を満たすラプラス分布の分散の考察

全体の属性値集合 $R$ （以下, 全体属性値集合と呼ぶ）に対して, 一部分の属性値集合 $C$ （以下, 部分属性値集合と呼ぶ）だけを対象に Pk-匿名化をすることを考える. この時 Pk-匿名化を満たすためのノイズのラプラス分布の $\sigma$ の値を求める.

データ全体にラプラス分布に従うノイズを付与する場合の $\sigma$ の値は, 第 3.2 節より

$$\sigma_i^R = 2\frac{d_i n}{log\frac{|R|-1}{k-1}}$$

である.

同様に部分属性値集合に対する $\sigma_i$ の値は,

$$\sigma_i^C = 2\frac{d_i \beta}{log\frac{|C|-1}{k-1}}$$

$$\text{ただし}, \beta = \Sigma_{j=1}^{n}\frac{c_j}{d_j}$$

となる. ここで $|C|$ は部分属性値集合の要素数, $d_i$ は全体属性値集合がとりうる値域, $c_i$ は i 番目（以下, i 次元と呼ぶ）の属性値の値域とする.

まず, $\sigma_i^R$ に対する $\sigma_i^C$ の比率を求める.

$$\frac{\sigma^C}{\sigma^R} = \frac{\Sigma_{j=1}^{n}\frac{c_j}{d_j}}{n}\frac{log\frac{|R|-1}{k-1}}{log\frac{|C|-1}{k-1}} \tag{4}$$

以降これをパラメタ比率と呼ぶことにする. 式 (4) で比率を求めた時点で次元に依存しなくなるため, 次元にかかわらず同じパラメタ比率となる.

### 5.2 全体属性値集合に対する割合を用いた指標の導入

この節で定義するそれぞれの指標では, 実際の値ではなく全体の属性値集合に対する割合で考える. 部分属性値集合 $C$ の要素数, 全体属性値集合の値域に対する $C$ の値域の比率をそれぞれ要素数比率, 幅比率と呼び, $frac_{card}(R,C)$, $frac_{width}(R,C)$ と表す. これらの定義は以下のとおりである.

$$frac_{card}(R,C) = \frac{|C|-1}{|R|-1}$$

$$frac_{width}(R,C) = \Sigma_{j=1}^{n}\frac{c_j}{d_j}$$

式 (4) に $frac_{card}(R,C)$ と $frac_{width}(R,C)$ を代入して関係式を求めると

$$\frac{\sigma^C}{\sigma^R} = \frac{frac_{width}(R,C)}{1+log_{\frac{|R-1|}{k-1}}frac_{card}(R,C)} \tag{5}$$

となる. 上式を読みとくと, $\sigma^R$ に対する $\sigma^C$ の比率が幅比率に比例する一方, 要素数比率に反比例するため, 要素数を小さくし過ぎるとパラメタ比率は逆に大きくなってしまい, 場合によっては 1 を超える可能性もある. しかしながら, 要素数比率は対数スケールで反比例するためその影響は要素数比率に対しては小さいと考えられる.

つまり, 全体属性値集合から部分属性値集合を抽出して Pk-匿名化を適用したとき, 全体属性値集合に対して小さな値域の部分属性値集合を選ぶことができれば, それによって要素数が減ったとしてもその影響は少なく, 部分属性値集合の値域分だけ $\sigma$ の値を抑えることができる.

## 6. 予備実験と評価

元の属性値をあらかじめ分割して Pk-匿名化を施すことで, より小さい $\sigma$ の値でのノイズ付与を実現することが期待できる. これを確認するため, サンプルデータを用いて全体の属性値を二つのグループにあらかじめ分割して Pk-匿名化を施す予備実験を行う. サンプルデータには, 東京と大阪の全国地方公共団体コードと郵便番号が格納されたレコードを用いる. 東京と大阪のレコードの分布は, 図 9 のようにはっきりと分かれている. これを東京グループと大阪グループとする.

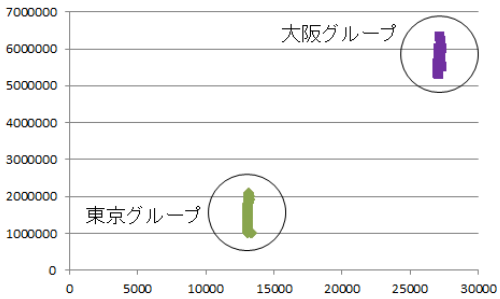


図 9: 東京グループと大阪グループの分布

東京と大阪をまとめたレコードを東京_大阪グループと呼ぶ. この東京_大阪グループに対して $\sigma$ の値を求めた場合と, 東京グループ, 大阪グループで別々に $\sigma$ の値を求めた場合との $\sigma$ の値を比較する. 図 10 が, $\sigma$ の値を比較したグラフである. 東京グループもしくは大阪グループだけで $\sigma$ の値を求めた場合に比べて, 東京_大阪グループで $\sigma$ の値を求めた場合の $\sigma$ の値の方が 6〜7 倍大きくなっていることが分かる.

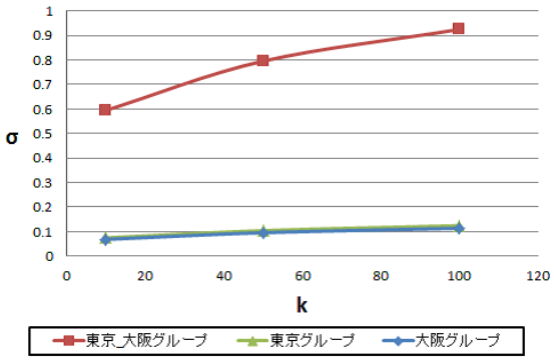


図 10: $\sigma$ の値の変化グラフ

この結果より, あらかじめレコードをグループに分類をしてからグループ毎に $\sigma$ の値を決定し, その $\sigma$ の値に従ってノイズを付与する方法が有用であると言える. また, $k = 20$ として東京_大阪グループで Pk-匿名化した場合と, 東京グループもし

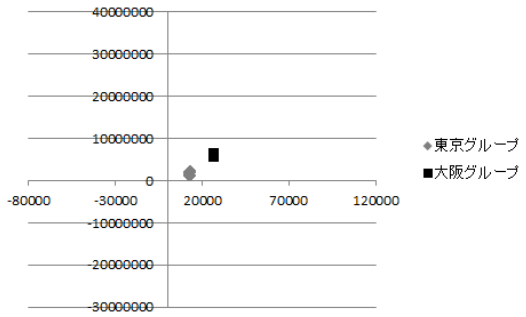


図 11: 匿名化前

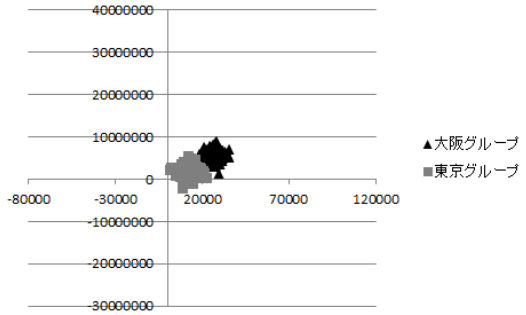


図 12: 分類してから匿名化した場合

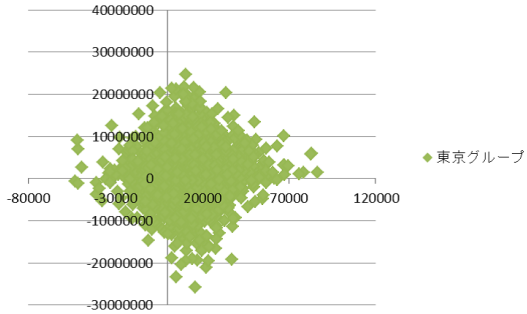


図 13: 東京_大阪グループ匿名化後（東京）

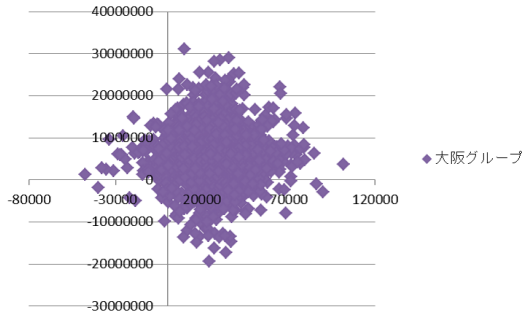


図 14: 東京_大阪グループ匿名化後（大阪）

くは大阪グループで Pk-匿名化した場合の値の分布を比較する (図 10〜14).

この結果を見ると, あらかじめ分類をせずに東京_大阪グループとして匿名化を行うと, 元の属性値から大きく離れて分散することが分かる (図 13, 図 14). しかし, あらかじめ分類をしてから東京グループと大阪グループに分けてそれぞれ Pk-匿名化した場合, 図 12 を拡大すると以下の図 15 のようになり, 各属性値はグループとして分布していた元の属性値の周辺に分散していることが分かる.

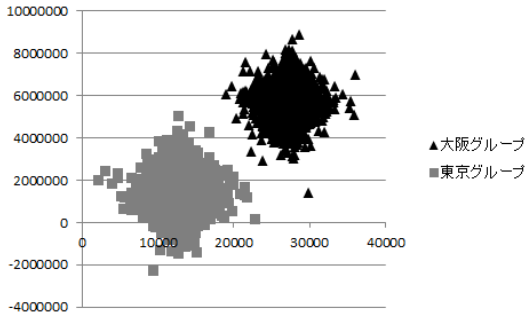


図 15: 分類してから匿名化した場合（図 12 の拡大）

この予備実験では属性値全体を単純に明確な二つのグループに分けたため, 第 5. 章で提案した比率を考慮して導出したグループの要素数で分割できれば, さらに $\sigma$ の値を抑えた Pk-匿名化を行うことができると考える. 今後, それらの比率を考慮したグループの要素数の最適化にも取り組んでいきたい.

## 7. ま と め

本稿では, ラプラス分布のパラメータを決める際に, 元の属性値をあらかじめいくつかのグループへ分類してからグループ毎に Pk-匿名化を施すという, 一改良法を示した. この手法により, 過剰にノイズを付与することなく Pk-匿名化を実現することが可能になると考える. 今後の課題として, さらに複雑な分類を持つデータや実際の機密情報を用いた検証実験や, k の値と $\sigma$ の値との関係の数式的定義の確立などが考えられる. また, 本稿で導出したデータ分割の条件の実験的な検証や, Randomization を用いた数値データ [6] に対する k-匿名化手法との比較もあげられる.

## 文 献

[1] 五十嵐 大, 千田 浩司, 高橋 克巳,”k-匿名化の確率的指標への拡張とその適用例”,CSS2009,2009

[2] 五十嵐 大, 千田 浩司, 高橋 克巳,”数値属性における k 匿名化を満たすランダム化手法”,CSS2011,2011

[3] 五十嵐 大, 千田 浩司, 高橋 克巳,”ランダム化データベース上の k-匿名性の一般的算出法”,CSS2011,2011

[4] 五十嵐 大, 長谷川 聡, 納 竜也, 菊池 亮, 千田 浩司, 数値属性に適用可能なランダム化により k 匿名化を保証するプライバシクロス集計,CSS2012

[5] 千田 浩司, 木村 映善, 五十嵐 大, 濱田 浩気, 菊池 亮, 石原 謙,”集合匿名化データの多変量解析評価”,CSS2012,2012

[6] J. Li, ”Preservation of proximity privacy in publishing numerical sensitive data”, SIGMOD2008

[7] L. Sweeney. ”k-anonymity: a model for protecting privacy”, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,10(5), pp.555-570, 2002.

[8] R. Agrawal and R. Srikant. Privacy-preserving data mining. Proc. of the 2000 ACM SIGMOD Intl. Conf. on Management of Data, 2000.

[9] Kristen LeFevre, David J. DeWitt, Raghu Ramakrishnan, ”Mondrian Multidimensional K-Anonymity”,ICDE,2006