Vol.2014-GN-92 No.3
Vol.2014-SPT-9 No.3
2014/5/15

確率的な安全性の指標に基づく パーソナルデータ匿名化システムの構築と評価

廣田 啓一[1]　正木 彰伍[1]　齋藤 恆和[1]　菊池 亮[1]　五十嵐 大[1]　冨士 仁[1]

概要：近年購買データ等のパーソナルデータの活用ニーズが高まる中，匿名化技術による安全な利活用の方法が検討されている．しかし，匿名化したデータの安全性と有用性の間にはトレードオフの関係があり，特に従来の匿名化技術において，安全性を高く加工すると情報としての有用性が低下することが知られている．我々は，このような匿名化の課題を解決するアプローチとして，匿名化の安全性を示す代表的な指標である k-匿名性と等価な安全性を持つ，確率的 k-匿名性に基づく匿名化システムを実装し，疑似データを用いた精度評価を行った．その結果, 同じパラメータを与えた場合でも，データ中の値の分布特性によって，出力データの精度にばらつきが生じることが明らかになった．本稿では確率的 k-匿名化において十分な精度を得るための条件について考察する．

Implementation and Evaluation of a Personal Data Anonymization System Based on Probabilistic Security Measure

Hirota Keiichi[1]　Masaki Shougo[1]　Saitou Tsunekazu[1]　Kikuchi Ryou[1]　Ikarashi Dai[1]
Fuji Hitoshi[1]

Abstract: Anonymization of personal data is expected to be one of the solution to use personal data in a secure manner. However, there is a trade-off between the safety and the usability of anonymized data, and it is known that when the data is processed to be highly safe, the usability of anonymized data decreases siginificantly using conventional anonymization techniques. As approach to solve this problem, we implemented probabilistic k-anonymity in our personal data anonymization system, and made an evaluation using published data and synthetic data. From the evaluation result, we found that the usability of anonymized data differ each even if the parameter given to the system is the same, and it depends on the distribution of the vale in data. In this report, we consider the condition to achieve high usability using probabilistic k-anonymization.

1. はじめに

近年，購買データや移動データ，視聴履歴データなどのパーソナルデータの活用ニーズが高まっている．サービス事業者に蓄積されたパーソナルデータを二次利用するニーズは従来からあり，レコメンデーションやパーソナライゼーションなどユーザにより良いサービスを提供するための利活用と合わせて，マーケティングによる新たなサービスの開発や商品の販売戦略，出店計画の策定といった，事業者視点での利活用への期待も高い．その一方で，個人に紐づくパーソナルデータの利活用にはプライバシーの問題があり，二次利用，特に第三者への提供はプライバシー侵害の懸念が強い．そのため，こうしたパーソナルデータを適切に匿名化することにより，本来とは異なる目的での二次利用や，第三者提供を可能とする動きが世界的に活発化している．日本においても個人情報保護法などの法制度の改正を含めた検討が急速に進んでいる．

ただし，どのように匿名化すれば適切に匿名化したと言えるかの明確な合意はまだ確立されていない．氏名や住所といった，特定の個人の識別が可能な情報を削除するだけ

[1] 日本電信電話株式会社 NTT セキュアプラットフォーム研究所
Nippon Telegraph and Telephone Corporation,
NTT Secure Platform Laboratories

の，いわゆる単純な匿名化では，外部の知識と照らし合わせた時に特定の個人が識別される再識別の問題が起こり得ることが指摘されており，実際に匿名化したデータから個人が識別された事例の報告も少なくない．こうした再識別の問題に対し，外部知識との照合を行っても個人の識別が起きるリスクが少ないと評価できる，安全性の指標および高度な匿名化の技術が必要とされている．

安全性の指標として代表的なものに，k-匿名性 [1][2] がある．k-匿名性は，同じ属性の値の組み合わせを持つレコードが k 個以上存在するように元データを加工することで，データの属性を知る攻撃者であっても “ある人のレコードを k 個未満に絞り込むことができない” 状態にし，個人の識別が起きるリスクを低減するものである．k-匿名性の考え方は広く受け入れられており，k-匿名性を持つように加工する k-匿名化の手法や k-匿名性を拡張した安全性の指標が数多く提案されている．

これに対し，近年，k-匿名性の考え方を確率的な指標に拡張し，一定の確率で元データにノイズを加えることで，“ある人のデータを $1/k$ 以上の確率で当てることができない” 状態にし，個人の識別が起きるリスクを低減する，確率的 k-匿名性 [3][4] が提案された．実際に確率的 k-匿名性を満たす確率的 k-匿名化の方法としては，カテゴリ属性に対する維持置換撹乱 [5]，数値属性に対するラプラスノイズ加算 [6] および有界ノイズ加算 [7]，データ分布に依存した撹乱処理を行う方法 [8] などが提案されている．

確率的 k-匿名化は，撹乱の処理が元データの値の分布によらず一定であるという特徴を持つ．そのため，確率的な安全性を保つために，データの性質によっては必要以上のノイズを加える場合があり，結果として加工後のデータの有用性を下げてしまう場合があることが考えられる．しかしながら，これまでの研究において，元データの値の分布に着目した撹乱処理の程度や，結果として得られる匿名化データの安全性と有用性のトレードオフについての明確な評価はなく，課題となっていた．

そこで，我々は確率的 k-匿名性に基づく匿名化システムを実装し，公開データおよび疑似データを用いた精度評価を行った．その結果，同じパラメータを入力として与えた場合でも，データ中の値の分布特性によって，匿名化したデータの精度にばらつきが生じることが明らかになった．

本稿では，評価実験の結果をもとに，確率的 k-匿名化において十分な精度を得るための条件について考察する．

2. 確率的 k-匿名化

確率的 k-匿名化は，データの値を確率的に維持あるいは置換することによりデータを撹乱する**維持置換撹乱**と，撹乱したデータから元データの統計量を推定する**再構築**からなる．

以下では，準備として処理の対象とするデータの概要を示した上で，基本的な確率的 k-匿名化における維持置換撹乱と再構築の処理の概要を説明する．

2.1 対象データ

確率的 k-匿名化では処理対象としてテーブルとクロス集計の二種類のデータを取り扱う．

テーブルとは個人に関する情報の集合である．各個人の情報はテーブル上では 1 行に表現され，これをレコードと呼ぶ．各レコードはあらかじめ決まった情報の項目に対する値から成り立っており，この項目を属性といい，属性として取りうる値の範囲を属性の値域という．**表 1** にテーブルの例を示す．“性別” 及び “年代” が属性であり，各行 “女性，10 代” 等がレコードである．

表 1 テーブルの例

ID	年代	性別
1	10 代	女性
2	40 代	男性
3	20 代	男性
4	30 代	女性
⋮	⋮	⋮

クロス集計とは，テーブルの複数の属性に着目し，その全ての属性の値が等しいレコードの数を表すものである．このクロス集計は，アンケート集計で使われるほかに数多くのデータマイニング手法の中で使用される基本的かつ重要な統計量である．一般的にクロス集計は行ベクトルとして扱われる．その大きさは各属性の組み合わせの総数，つまり各属性の値域の積となり，ベクトルの要素の和はレコードの総数と一致する．**表 2** は表 1 のテーブルのクロス集計の例である．ベクトルの第 2 要素である “{10 代，女性 }” はテーブルの中にある 10 代かつ女性である人数の総和であり，右辺のベクトルの第 2 要素はその属性の値を組を持つレコードが 45 件あることを示している．

表 2 クロス集計の例

({10 代，男性 }, {10 代，女性 },
{20 代，男性 }, {20 代，女性 }, . . .)
= (61, 45, 73, 50, . . .)

2.2 維持置換撹乱

確率的 k-匿名化の基本的な撹乱方法として，レコード毎に属性の値を一定の確率で維持し，それ以外の場合は取りうる値の中でランダムな値に置き換える維持置換撹乱がある [3]．この確率を，維持確率 ρ と呼ぶ．

維持確率 ρ を適切に設定することで，維持置換撹乱されたデータは Pk-匿名性を満たすようになる．

本稿は基本的な確率的 k-匿名化について評価を行ったものである．これらの拡張手法については別途評価を行う．

匿名性の指標として与えられた k の値に基づいて Pk-匿名性を満たすために，維持確率 ρ を含むパラメータ間で，以下の公式が成立する必要がある [3].

$$k = 1 + (|\mathcal{R}| - 1) \prod_{a \in \mathcal{A}} \left[\frac{1 - \rho_a}{1 + (M_a - 1)\rho_a} \right]^2 \qquad (1)$$

ここで，ρ_a は属性 a の維持確率，$|\mathcal{R}|$ はテーブルのレコード数，$|\mathcal{A}|$ は属性数，M_a は属性 a の値域を表し，$\prod$ はクロネッカー積である．維持確率 ρ_a は $0 \leq \rho_a \leq 1$ の範囲で，個々の属性毎に異なる値をとることもできるが，本稿では複数の属性で共通の値を用いる．

この維持置換撹乱により，ある属性の値は属性の持つ値域の範囲でランダムに置き換えられる．ある属性 a の値 v_a が，任意の値 v'_a に遷移する確率を，サイズ $M_a \times M_a$ の行列 $A^{(a)}_{v_a, v'_a}$ を用いて，

$$A^{(a)}_{v_a, v'_a} = \begin{cases} \rho_a + \frac{1-\rho_a}{M_a} & \text{for} \quad v_a = v'_a \\ \frac{1-\rho_a}{M_a} & \text{for} \quad v_a \neq v'_a \end{cases} \qquad (2)$$

と表すことができる．この行列 $A = \prod_{a \in \mathcal{A}} A^{(a)}$ を遷移確率行列と呼ぶ.

2.3 再構築

撹乱したデータから元データの統計量を推定する再構築の方法として，反復ベイズ推定法が知られている [9]．これは，撹乱した後のテーブルのクロス集計に対し，遷移確率行列 A を用いてベイズ推定を繰り返し行うことで，元のテーブルのクロス集計を推定するものである．

反復ベイズ推定法のアルゴリズムを Algorithm 1 に示す.

Algorithm 1 ベイズ推定法による再構築アルゴリズム

Input: A, $\vec{y}$
Output: $\vec{x}_i$
$\vec{x}_0 := \vec{y}$
$i := 0$
while $|\vec{x}_{i+1} - \vec{x}_i|_{L_1} \leq \epsilon$ **do**
  $\vec{x}_{i+1} := \vec{x}_i * (A(\vec{y}/\vec{x}_i A)^t)^t$
  $i := i + 1$
end while

ここで，アルゴリズム中の i は反復回数を表し，ベクトル $\vec{v}_1$, $\vec{v}_2$ に対して $\vec{v}_1 * \vec{v}_2$ と $\vec{v}_1/\vec{v}_2$ はベクトルの成分ごとの積と商を表す．$|\vec{x}_{i+1} - \vec{x}_i|_{L_1}$ は $\vec{x}_{i+1}, \vec{x}_i$ の L_1 距離と呼ばれ，各要素の差の絶対値の総和をレコード数で割ったものであり, ϵ はあらかじめ定める収束半径である. 本アルゴリズムは撹乱テーブルのクロス集計 $\vec{y}$ と遷移確率行列 A を入力とし，推定した元テーブルのクロス集計 $\vec{x}_i$ を出力とする.

3. パーソナルデータ匿名化システムの構築

前述の維持置換撹乱と再構築による確率的 k-匿名化の処理を実装した，パーソナルデータ匿名化システムを構築した.

なお，本システムにおいては属性の値の一般化とレコードの削除による従来の k-匿名化の処理も合わせて実装し，匿名化の対象とするデータの性質や用途によって匿名化の処理を使い分けできるものとした．本稿では k-匿名化の処理概要は説明を割愛する.

3.1 データの入力と属性の選択

匿名化システムは，テーブルを入力として読み込む．入力の際にはテーブルの構造や各属性の名称やカテゴリ属性，数値属性といった種類をメタデータとして与え，任意の形式のデータを読み込めるように設計した.

操作者は画面上で，テーブル内の匿名化処理の対象とする属性を選択し，匿名化の指標 k の値を与えて，Pk-匿名化処理を実行することができる.

図 1 に実行中のシステム画面の例を示す.

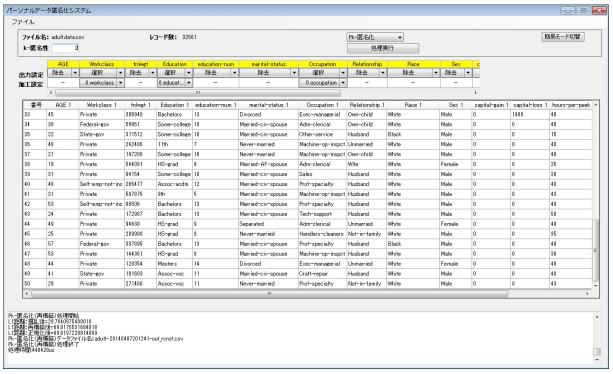

図 1 システム画面の例

3.2 Pk-匿名化処理

匿名化処理の対象とする属性を選択して，匿名化の指標 k の値を与え，処理を実行すると，システムは，パラメータから維持確率を算出し，撹乱処理を行って撹乱したデータを生成する．次に，撹乱したデータからクロス集計を生成し，遷移確率行列に基づいて再構築処理を行った後に，再構築したクロス集計をテーブル形式のデータに変換して出力する．処理の違いにより，Pk-匿名化処理では 2 種類のテーブルが生成される．便宜上，前者を撹乱データ，後者を再構築データと呼ぶ.

3.3 表示ツール

Pk-匿名化では，維持置換撹乱処理により元データの値を書き換えた上で，再構築処理により一旦クロス集計の形

式に変換されるため，その出力はレコードの順序が元のデータと異なり，出力されたデータを参照するだけでは匿名化処理によるデータの変化が分かりにくい．

そのため，匿名化前後でのデータの変化を可視化するための表示ツールとして，ヒストグラム表示とクロス集計表示を実装した．

3.3.1 ヒストグラム表示

個々の属性毎に処理を行うため，値の分布をヒストグラムにより表示する．**図 2** は後述する公開データにおける属性 age の値の分布を表示したものである．

左側は元データにおける値の分布を表し，右側は再構築データにおける値の分布である．

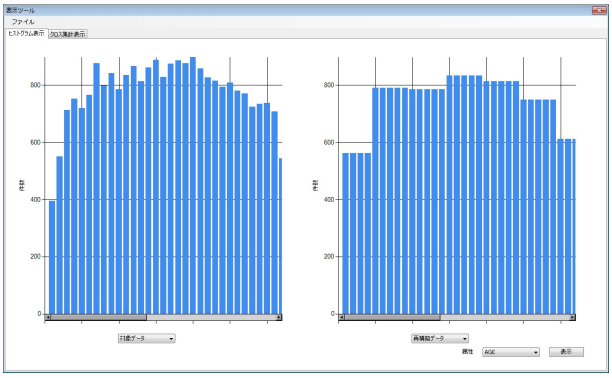

図 2 ヒストグラム表示の例

3.3.2 クロス集計表示

再構築処理では，クロス集計の形式に変換されるため，個々の属性の値の分布が正しくても，複数の属性の値の組み合わせとしては正しくない分布が推定される場合が考えられる．そのため，複数の属性の関係の変化を見るため，任意の 2 属性のクロス集計結果を表示するものとした．

図 3 は後述する公開データにおける属性 workclass と属性 education のクロス集計の表示例を示す．ヒストグラム表示と同様に，左側は元データにおけるクロス集計を表し，右側は再構築データにおけるクロス集計である．レコード数に対する集計値の割合に応じて色の濃淡をつけることにより，クロス集計値の変化が容易にわかるようにしている．

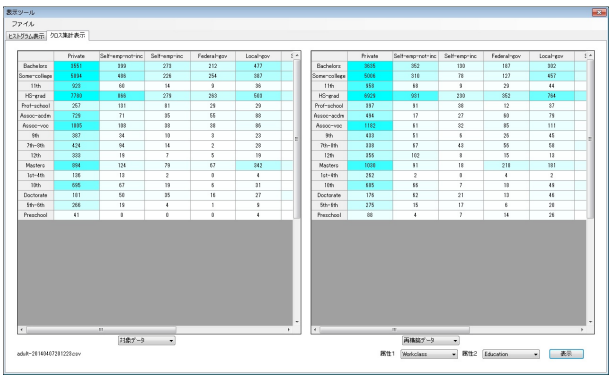

図 3 クロス集計表示の例

4. 評価実験

構築したパーソナルデータ匿名化システムを用いて，確率的 k-匿名化により生成されるデータの精度に関する評価実験を行った．

同じ匿名性指標 k を与えた場合でも，維持確率 ρ の値が高い程，維持置換撹乱における撹乱処理の程度が少ないため，再構築したデータの誤差が小さく，より有用性が高くなると考えられる．評価にあたっては，維持確率 ρ の値と，元のクロス集計との $L1$ 距離を評価の軸とし，匿名性指標 k の値を変えてこれらの評価値の変化を確認するものとした．

なお，クロス集計間の誤差を $L1$ 距離を使って，次のように算出した．これを本稿では仮に $L1$ 精度と呼ぶ．

$$precision_{L1}(x,y) = 1 - \frac{\sum_{i=1}^{n} |x_i - y_i|}{2|\mathcal{R}|} \qquad (3)$$

$L1$ 精度はクロス集計の各要素の差の絶対値の総和をレコード数の 2 倍で割ったものであり，直感的にはクロス集計の結果が総レコード数の何 % 合致するかを表す．$L1$ 精度が大きいほど，元のデータのクロス集計と近い結果が得られていることになり，データとしての有用性が高いと考えられる．

4.1 公開データを用いた評価実験

匿名化システムにおいて実装した Pk-匿名化の有用性を確認するために，公開データを用いた精度評価を行った．

4.1.1 対象データ

対象データとして，UCI Machine Learning Repository [10] の Adult Data Set を用いた．Adult Data Set は，1994 年のアメリカの国勢調査データベースから抽出された 15 の属性からなる 48,842 人分の公開データセットである．本実験では，Adult Data Set の訓練データ 32,561 件から，表 3 に示す 9 属性を対象としてテーブルを構成した．

表 3 テーブルの構成

属性	属性の値域数	備考
age	15	5 歳刻みに丸め
work class	9	損失値含む
education	16	
martial status	7	
occupation	15	損失値含む
relationship	6	
race	5	
sex	2	
native-country	42	損失値含む

なお，9 属性の内 ‘age’ については本来数値型の属性だが，Pk-匿名化の処理を前に，元の値域である 17 – 90 歳の

Adult data Set において損失値は?で示される．

値から 5 歳刻み（20 歳以下，21 – 25 歳，...，66 – 90 歳）の 15 種類のカテゴリ属性に丸め処理をしている．

4.1.2 実験方法

Adult Data Set のレコード数がたかだか 32,561 件であることから，テーブルの 9 属性全てを処理対象とした場合，維持確率がきわめて低くなり，十分な精度が得られないことは明らかである．そこで，9 属性の中から以下の**表 4** に示す 3 属性ずつの組み合わせ $t1$〜$t4$ を対象として，Pk-匿名化処理を実施するものとした．

組み合わせの $t1$ から $t3$ は 9 属性を単純に 3 属性ずつ分割したものである．表 4 における総値域数は，属性 1〜3 の値域の数から定まる属性の値の組み合わせの数の最大値を表す．$t4$ については，総値域数が $t3$ と同じになるように属性 1 のみ変えた組み合わせとし，組み合わせを構成する属性の違いによる精度の差異を見るものとした．

表 4 属性の組み合わせパターン

	属性 1	属性 2	属性 3	総値域数
t1	race	sex	native-country	420
t2	occupation	relationship	marital status	630
t3	age	work class	education	2160
t4	occupation	work class	education	2160

4.1.3 実験結果と考察

各組み合わせに対して，k の値を $k = 2, 5, 10$ と変えて Pk-匿名化処理を行った際の維持確率と，撹乱後のデータの $L1$ 精度，再構築後のデータの $L1$ 精度を**表 5** に示す．

表 5 組み合わせと k の値による実験結果

	維持確率			L1 精度（撹乱後）			L1 精度（再構築後）		
	k=2	k=5	k=10	k=2	k=5	k=10	k=2	k=5	k=10
t1	0.350	0.280	0.240	30.9	25.9	23.5	91.1	88.9	88.4
t2	0.350	0.287	0.252	36.5	33.4	31.8	79.8	77.6	73.8
t3	0.264	0.213	0.182	38.7	35.7	34.0	73.6	74.1	72.3
t4	0.264	0.213	0.182	30.1	27.4	25.9	71.0	68.0	65.3

表 4，表 5 から明らかなように，総値域数が少ない組み合わせほど維持確率は高くなる．しかし，再構築後のデータの L1 精度に関しては，必ずしも維持確率が高い方が良い結果とはならなかった．

また，同じ総値域数となる組み合わせの場合でも，維持確率は等しくなるが，属性の組み合わせによって $L1$ 精度に差がつく結果となった．

公開データを用いた実験で明らかになったように，レコード数と値域の数から同じ維持確率の値が算出される場合でも，対象とする属性の組み合わせによって，再構築したデータの $L1$ 精度に 5 ポイント以上の差異が見られる結果となった．

ヒストグラム表示により age の分布を確認したところ，全ての値域にある程度万遍なく分布しているのに対し，

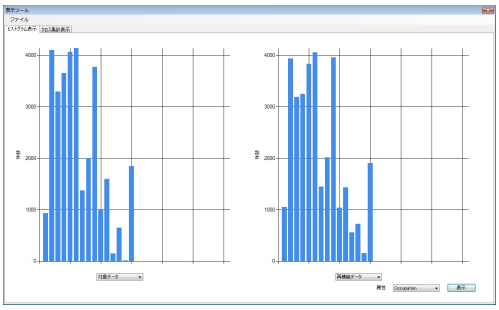

図 4 属性 occupation のヒストグラム表示

occupation の分布は属性の値によって分布が大きく異なり，頻度の高いものと低いものが散在していることが確認された．

4.2 疑似データを用いた評価実験

公開データを用いた実験から，匿名化データの $L1$ 精度は単純に維持確率からは定まらず，データの分布に依存するところが大きいと考えられる．

そこで，どのような分布の場合に影響が生じるか，また確率的な匿名化処理で十分な精度が得られるかを明らかにするために，疑似データを用いて，詳細な精度評価を行うものとした．

4.2.1 対象データ

疑似データを用いた実験の目的は，属性の値の分布の違いからくる精度の差を明らかにすることにある．そこで，対象データとしては血液型と生月の 2 属性のみで構成し，各属性の値域における値の分布を作為的に決定して，疑似データを作るものとした．

血液型は，A 型，B 型，O 型，AB 型の 4 つの値を値域とすることから，分布パターンを**表 6** のように構成した．

表 6 血液型の分布パターン構成

	分布パターン
A	A 型のみ 80% とし，他の値を均等とする一部突出した分布
B	A 型，B 型を 40%，C 型，D 型を 10% とする分布
C	A 型から順に 40%，30%，20%，10% とする傾斜状の分布
D	AB 型のみ 10% とし，他の値を均等とする一部陥没した分布
E	各血液型 25% とする均等分布

一方，生月については 1 月〜12 月までの 12 の値を値域として持つことから，分布のパターンを**表 7** のように構成した．

4.2.2 実験方法

血液型のパターン A〜E，生月のパターン 1〜5 を組み合わせた 25 通りの分布構造について，それぞれ 1,000 件，10,000 件の疑似データを生成した．これらの疑似データを入力として，指標 k の値を $k = 2$，$k = 5$，$k = 10$ と変えて，Pk-匿名化処理を実行し，撹乱後のデータの $L1$ 精度，再構築後のデータの $L1$ 精度をそれぞれ算出した．

撹乱処理は確率的な処理であり，その都度生成される乱数により処理結果が変わるため，評価に際しては 5 回ずつの試行を行って平均値を求めている．

4.2.3 実験結果と考察

まず組み合わせパターン毎に 1,000 件の疑似データを使った実験結果を**表 8** に示す．

25 通りの組み合わせの内，A-1 は一部突出した値域に値が集中するパターン，E-5 はもっとも平均的に値が分散して分布するパターンと考えられる．撹乱後のデータの $L1$ 精度を見てみると，A-1 から E-5 にかけて，すなわち突出したパターンから平均的なデータにかけて $L1$ 精度が高くなる傾向が見られる．これは維持置換撹乱によって属性の値がランダムに置き換えられることにより，撹乱した後のデータがより平均的な分布に置き換わり，安全性が高められることに由来する．なお，維持確率 ρ の値は，$k = 2$ から順に 0.393，0.291，0.236 と下がっている．

逆に，再構築後のデータの $L1$ 精度を見てみると，今度は E-5 から A-1 にかけて，$L1$ 精度が高くなる傾向が見られる．すなわち，突出したパターンにおいては，再構築処理によって本来の分布がある程度正しく推定されたためと考えられる．

次に，組み合わせパターン毎に 10,000 件の疑似データを使った実験結果を**表 9** に示す．

撹乱後のデータの $L1$ 精度については，1,000 件の場合と同様に，突出したパターンから平均的なデータにかけて $L1$ 精度が高くなる傾向が見られる．一方，再構築後のデータの $L1$ 精度については全てのパターンで 93〜96% 前後の高い結果が得られ，データの分布パターンの違いによる差異は見られなかった．これは十分なレコード数がある状態で高い維持確率によって維持置換撹乱処理が行われ，いずれのパターンにおいても，再構築処理によって本来の分布がある程度正しく推定されたためと考えられる．なお，維持確率は順に 0.561，0.463，0.400 であった．

実験結果に基づく再構築後のデータの $L1$ 精度のグラフを**図 5** に示す．既に述べたように，1,000 件の疑似データの場合は，匿名性の指標 k の下に同じ維持確率を算出しても，データの値の分布パターンによって再構築データの $L1$

表 7 生月の分布パターン構成

	分布パターン
1	1 月を 45%、2 月〜12 月を各 5% とする一部突出した分布
2	1〜6 月で 80%，7〜12 月で 20% を占める分布
3	1 月 15.4%〜12 月 1.3% とする傾斜状の分布
4	1〜11 月を 9%，12 月を 1% とする一部陥没した分布
5	各月約 8.3% とする均等分布

表 8 疑似データ 1,000 件での実験結果

	L1 精度（撹乱後）			L1 精度（再構築後）		
	k=2	k=5	k=10	k=2	k=5	k=10
A-1	66.24	59.48	57.34	79.54	73.86	74.02
A-2	64.06	59.74	57.64	78.52	75.62	77.76
A-3	64.06	58.68	55.58	81.84	74.58	71.24
A-4	65.46	60.46	57.30	80.68	74.26	73.06
A-5	66.24	60.58	58.08	75.26	72.02	71.70
B-1	72.84	69.38	67.84	74.86	66.94	65.28
B-2	75.04	70.24	68.26	73.24	67.00	64.24
B-3	73.74	70.94	68.94	70.86	66.12	65.18
B-4	78.40	75.94	74.48	69.22	64.94	62.60
B-5	80.74	76.76	75.68	70.00	62.46	61.80
C-1	75.14	71.58	70.26	75.48	69.58	66.52
C-2	78.52	75.98	74.36	68.18	62.90	60.92
C-3	79.48	77.24	75.22	68.28	63.32	58.20
C-4	82.86	80.04	77.98	69.88	60.82	56.34
C-5	85.44	82.16	82.82	68.68	60.94	62.58
D-1	75.04	71.60	70.20	73.48	69.96	64.00
D-2	78.98	72.74	71.20	70.14	65.06	63.30
D-3	79.40	77.22	74.88	67.64	61.20	57.10
D-4	83.84	81.92	81.36	66.60	61.54	59.16
D-5	86.30	83.62	83.66	69.46	62.12	61.24
E-1	75.26	70.98	69.12	76.38	65.12	74.52
E-2	81.40	78.04	76.28	68.90	63.64	56.62
E-3	82.46	80.42	76.84	71.60	67.34	59.78
E-4	88.76	87.40	86.96	72.14	61.26	55.70
E-5	89.88	88.46	88.36	64.12	58.88	59.58
平均	78.97	75.72	74.28	70.13	64.30	61.69

表 9 疑似データ 10,000 件での実験結果

	L1 精度（撹乱後）			L1 精度（再構築後）		
	k=2	k=5	k=10	k=2	k=5	k=10
A-1	75.89	70.38	66.55	95.85	95.01	93.56
A-2	73.42	68.36	65.01	94.35	93.26	92.90
A-3	73.45	67.97	64.33	94.99	93.83	93.53
A-4	74.21	69.09	65.45	94.63	92.69	92.87
A-5	75.94	70.30	67.23	95.02	93.85	92.22
B-1	82.75	79.74	77.52	94.61	92.87	92.46
B-2	80.89	77.18	74.51	94.06	91.62	90.66
B-3	82.38	78.96	76.70	93.94	93.14	91.29
B-4	84.87	81.92	79.98	93.25	93.64	91.22
B-5	86.76	83.68	82.23	93.49	91.10	90.04
C-1	83.25	79.66	77.57	94.29	92.58	91.90
C-2	83.98	80.47	78.33	93.78	92.31	90.85
C-3	85.70	82.93	80.94	92.89	91.58	90.44
C-4	88.91	87.12	85.75	93.46	92.70	88.29
C-5	90.85	89.22	87.49	93.02	92.04	90.46
D-1	83.12	79.81	77.59	93.91	93.34	91.85
D-2	83.57	80.17	77.99	94.25	92.51	91.05
D-3	86.36	83.28	81.57	93.87	92.04	90.74
D-4	90.54	88.68	87.78	93.10	92.16	89.27
D-5	93.04	91.76	91.27	93.96	91.87	91.43
E-1	84.01	80.41	78.13	94.09	92.60	91.21
E-2	87.08	84.18	82.33	94.46	92.76	91.38
E-3	89.61	87.19	85.39	93.91	92.49	90.40
E-4	95.16	94.81	94.52	93.41	91.67	89.97
E-5	96.87	96.56	96.42	93.15	90.74	88.30
平均	85.47	82.54	80.65	93.64	92.37	90.62

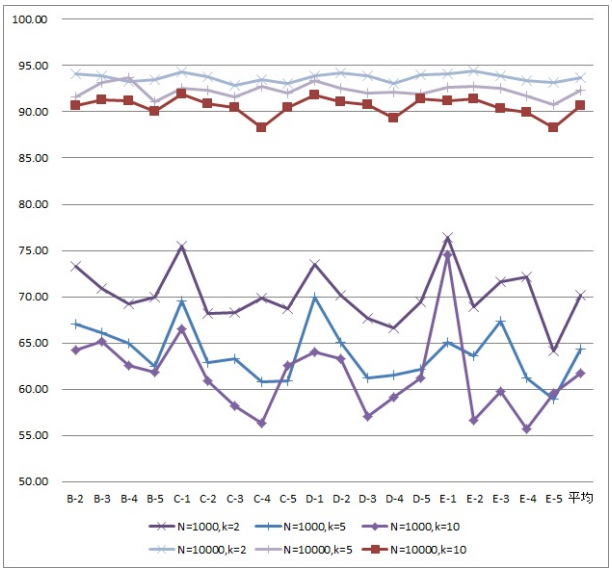

図 5 実験結果に基づく $L1$ 精度のグラフ

精度は大きく異なる結果となった．本実験においては，突出したパターンの分布ほど高い精度を示す傾向がみられた．

一方，10,000 件の疑似データの場合にはこうした差異は見られず，全般的に高い精度が得られる結果となった．

5. 考察と課題

公開データ，疑似データのそれぞれを用いた評価実験の結果，パラメータとして同じ匿名性の指標 k を与えた場合であっても，出力される再構築データの $L1$ 精度にばらつきが生じることが明らかになった．

以下では，確率的 k-匿名化において十分な精度を得るための条件について考察する．

まず維持確率 ρ を求める公式 (1) より，高い維持確率を得るためには，レコード数，属性の数，各属性の値域の数の 3 つが重要であることは自明である．

この内，属性の数については，与えられたテーブルの全ての属性を匿名化の対象とするのではなく，必要とする属性のみを匿名化するオーダーメイド匿名化を行うことで，少ない属性の数で高い維持確率を得ることが期待できる．Pk-匿名化は，同じデータから異なる属性の組を取り出して，繰り返し匿名化を行った場合でも，Pk-匿名性が保たれる性質を持つ．そのため，最小限の属性の組を対象として匿名化することが望ましい．

一方，各属性の値域の数については対象とするデータに依存するため，これを可変とすることは難しい．たとえば外れ値を持つレコードを削除する，あるいは属性の値を一般化するといった，従来の k-匿名化と同じアプローチにより値域を減らす方法が考えられるが，削除を行った場合には繰り返し匿名化する際の安全性が問題となるため，望ましい方法ではない．

もっとも単純には，対象データとして十分な数のレコードを確保することが望ましい．属性の数，各属性の値域の数と満たすべき匿名性の指標 k の値から，十分に高い維持確率を得るためのレコード数が逆算的に求められるため，精度を基準としてレコード数を定めることも考えられる．しかし，匿名化の対象とするテーブルは各個人の情報を 1 レコードとして扱うものであり，レコード数を増やすことは必ずしも容易ではない．

一方，本稿の実験結果から，維持確率が同じであっても，対象とするデータの属性の値の分布によって，精度にばらつきが生じることが示された．本稿で示した確率的な k-匿名化は，レコード毎に属性の値を一定の確率で維持または撹乱するものであり，各属性の中での値の分布を考慮したものとはなっていない．データ中の値の分布を考慮した撹乱方法を採用することが望ましいと考えられるが，分布に依存した情報や手法を用いることによって，再識別の可能性が増加することが懸念されるため，安全性の低下が起きないことの理論的な検討が必要である．そのため分布を考慮した撹乱方法の採用は今後の課題である．

6. おわりに

本稿では，確率的 k-匿名性に基づくパーソナルデータ匿名化システムの概要を示し，実装したシステムにより行った，公開データを用いた精度評価，疑似データを用いた精度評価の結果をそれぞれ示した．その結果, 同じパラメータを与えた場合でも，データ中の属性の値の分布によって，出力データの精度にばらつきが生じることが明らかになった．本稿では確率的 k-匿名化において十分な精度を得るための条件について考察を行ったが，その他にも幾つかの検討課題が存在する．

一点目は確率的 k-匿名化の精度向上である．本稿では，基本的な Pk-匿名化を実装の対象として，評価実験を行ったが，精度向上に向けたアプローチとして，数値属性に対する有界ラプラスノイズ加算やデータ分布に依存した撹乱処理を行う確率的 k-匿名化などの拡張手法が既に提案されており，これらの手法を採用することによる精度向上が期待される．また，考察において述べたように，データの分布に依存した撹乱方法を採用することによる精度向上の検討も行いたい．

二点目は，データ分析における精度の確認である．本稿では，データの有用性を示す指標として，元データのクロス集計と出力データのクロス集計の間の $L1$ 距離を軸として，精度評価を行った．実際には出力データは相関分析やクラスタリングといったデータマイニングの手法を適用した場合に，どれだけ元のデータと同じような分析結果が得られるか，あるいは有用な結果が得られるかが，匿名化データの有用性の観点からは重要となる．

三点目は，実データに適用した際の実用性の確認である．Pk-匿名化は意図的に誤差を入れて撹乱することにより，

安全性を高める手法である．マーケティングなどの元々のデータの誤差も含めて，ある程度の傾向が求められればいいデータ分析の領域があると考えている．今後，実サービスにおいてパーソナルデータの匿名化による安全な分析を検討する事業者などと議論を行って，どの程度の誤差であれば許容範囲として実用的なデータと言えるかを，実例を基に検討していきたい．

これらの検討課題に対しては，実装した匿名化システムを用いた実証実験などを行うことにより，明らかにしていく．

参考文献

[1] Pierangela Samarati & Latanya Sweeney, “Generalizing Data to Provide Anonymity when Disclosing Information”, PODS, 1998.
[2] Latanya Sweeney, “k-Anonymity:A Model for Protecting Privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002.
[3] 五十嵐大, 千田浩司, 高橋克己, “多値属性に適用可能な効率的プライバシー保護クロス集計”, CSS, 2008.
[4] 五十嵐大, 千田浩司, 高橋克巳, “k-匿名性の確率的指標への拡張とその適用例”, CSS, 2009.
[5] Rakesh Agrawal, Ramakrishnan Srikant & Dilys Thomas, “Privacy Preserving OLAP”, SIGMOD, 2005.
[6] 五十嵐大, 千田浩司, 高橋克巳, “数値属性における, k-匿名性を満たすランダム化手法”, CSS, 2011.
[7] 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-匿名性を保証するプライバシー保護クロス集計”, CSS, 2012.
[8] 菊池亮, 五十嵐大, 千田浩司, 濱田浩気, “データ分布依存処理によって高い有用性を実現する確率的 k-匿名化”, SCIS, 2013.
[9] Rakesh Agrawal & Ramakrishnan Srikant, “Privacy-Preserving Data Mining”, SIGMOD, 2000.
[10] UCI repository of machine learning databases, 1998, 入手先 ⟨http://archive.ics. uci.edu/ml/datasets/Adult⟩.
[11] 永井彰, 五十嵐大, 濱田浩気, 松林達史, “クロネッカー積を含む行列積演算の最適化による効率的なプライバシ保護データ公開技術”, SCIS, 2010.
[12] 千田浩司, 菊池亮, 濱田浩気, 五十嵐大, 廣田啓一, 冨士仁, 高橋克己, “動的集合匿名化データの有用性評価と開示リスクに関する一考察”, CSS, 2013.
[13] 齋藤恆和, 五十嵐大, 菊池亮, 濱田浩気, 正木彰伍, “撹乱再構築法における再構築アルゴリズムの高速化手法の提案”, to appear in SCIS, 2014.