集合匿名化データの多変量解析評価

千田浩司† 木村映善‡ 五十嵐大† 濱田浩気† 菊池亮† 石原謙‡

†日本電信電話（株）NTT セキュアプラットフォーム研究所
180-8585 東京都武蔵野市緑町 3-9-11
‡愛媛大学医学部附属病院医療情報部
791-0295 愛媛県東温市志津川

あらまし 個人データの開示リスクを抑える手法の一つとして集合匿名化が知られているが，集合匿名化を施したデータの有用性については現状あまり明らかではない．特に多変量解析のように多変数からなるデータを扱う場合は，集合匿名化によってデータの有用性が著しく低下する「次元の呪い」の問題が指摘されている．本稿では次元の呪いの対策として，多変数からなる個人データを変数の少ない部分集合の組に分割して各々に集合匿名化を施す手法に着目し，多変量解析を例に本手法の有効性を考察する．また実際の医療情報に集合匿名化を施したデータを用いて，代表的な多変量解析である重回帰分析を行い，本手法の効果を実験的に検証する．

Evaluation of Multivariate Analysis on Group-Based Anonymized Data

Koji Chida† Eizen Kimura‡ Dai Ikarashi† Koki Hamada†
Ryo Kikuchi† Ken Ishihara‡

†NTT Secure Platform Laboratories
3-9-11 Midori-cho, Musashino-city, Tokyo 180-8585, JAPAN
‡Dept. Medical Informatics of Medical School of Ehime Univ.
Situkawa, Toon-city, Ehime 791-0295, JAPAN

Abstract A group-based anonymization such as k-anonymization method is known as a disclosure control method for personal data, however, its practical effort has been determined yet. In particular, the dimensionality curse problem makes the anonymization of multivariable data difficult while the multivariable data is also extremely useful for elaborate analyses. In this paper, we focus on an approach that selects several subset tables from the multivariable data and provides individually anonymized subset tables. We experimentally verify usability of the divided anonymized tables for multiple regression analysis using medical information.

1 はじめに

ICT の発達に伴い，組織や個人に関するデータの管理や利用の形態が年々変化してきている．特に最近では各種データを容易に収集し利活用できる環境が整備されつつあり，新たな価値創造による社会や産業の発展が期待できる．しかしながら，個人に関するデータ (個人データ) を利活用する際は，プライバシ保護に十分配慮しなくてはならない．例えば 1990 年の米国国勢調査の回答データは性別，生年月日，及び 5 桁の ZIP コードだけで全体の 87% が一意の値と

なっており，特定個人の回答データが高い確率で一意に識別されることが指摘されている [1].

前掲した個人データ利活用における問題への有望な対策として，**集合匿名化**技術が知られている．集合匿名化とは，開示リスクを考慮して個人データをグループ化する処理を指し [2]，個人データの集合から特定個人のデータを k 個未満に絞り込めるかどうかを匿名性の指標とする k-匿名性 [3] 及びその派生指標に基づく加工技術の総称である．しかし集合匿名化は**次元の呪い**と呼ばれる，性別や年齢といった個人データの変数の増加に伴いデータの有用性が低下する問題が指摘されている [4](3 節で詳述)．収集された個人データは一般に変数が多いほど様々な分析が可能となることから，大規模な個人データの利活用においては特に次元の呪いの問題の解決が強く望まれるだろう．

本稿では，多変数からなる個人データを変数の少ない部分集合の組に分割して各々を集合匿名化することで次元の呪いの回避を試み，多変数からなるデータを扱う多変量解析を例に本手法の有効性を考察する．具体的には，ある個人データから複数の部分データを抽出してそれぞれ集合匿名化するモデルの匿名性や処理効率，そして多変量解析に対する部分データの適用可能性について考察する．また実際の医療情報に集合匿名化を適用し，その集合匿名化データを用いて代表的な多変量解析である重回帰分析を行うことで本手法の効果を実験的に検証する．

2 準備

本節では本題に入る準備として，用語の定義と関連研究の紹介を行う．

2.1 個人データとテーブル

個人データは以下に分類される変数の組からなるものとする．

- **正識別子**: 個人を一意に識別できる変数，または変数の組．氏名，住所の組み合わせは無視できない確率で正識別子となる [5].
- **準識別子**: 間接的に個人を識別できる変数．性別や年齢は間接的に個人の識別に利用できる [6].
- **センシティブ変数**: 正識別子，準識別子以外で，個人のプライバシに関するもの等，他人にむやみに知られたくない変数．
- **非センシティブ変数**: 上記以外の変数．

表 1 に例示するように，先頭行に各変数名が記載され，他の各行 (レコード) には先頭行の変数名に従った個人データが記載される表形式のデータをテーブルと呼ぶ．

2.2 リスク指標：k-匿名性，Pk-匿名性

個人データを開示するリスクは，個人データが正識別子と対応付く身元開示リスク，及び個人データから特定個人のセンシティブ変数の値が知られる属性開示リスクとされる [7]．本稿では，身元開示リスクや属性開示リスクを抑えるために個人データを加工する処理を総称して匿名化と呼ぶ．個人データを利活用する際は，匿名化によって個人のプライバシがどの程度保護されるか (リスク指標)，また元のデータと比較して匿名化されたデータがどれだけ利活用に資するか (有用性指標) が重要な指標となる．

身元開示リスクを考慮したリスク指標として前述の k-匿名性が挙げられる．表 2 は $k = 2$ として表 1 のテーブルが k-匿名性を満たすように匿名化 (k-匿名化) した例である (一部の変数は記載を省略している)．同一の準識別子の値の組が常に k 個以上存在するように値の一般化や削除を行う．表 2 のテーブルからは，性別=‘男’, 年齢=‘24’ の特定個人のレコードが 2 行目または 5 行目のどちらか識別できないことが分かる．

k-匿名化は基本的に，個人データの一般化や削除により k-匿名性を満たすテーブルを作成するが，ノイズ付加やデータ置換といった撹乱処理によって匿名化する手法も研究されている．例えば撹乱処理に加え，撹乱されたデータから集計値を精度よく復元する再構築処理からなる撹乱・再構築法 [8, 9] が提案されている．さらに撹乱・再構築法については，k-匿名性と等価であ

表 1: テーブル (診療報酬要因分析に使用するデータの一例)

患者番号 (正識別子)	性別 (準識別子)	年齢 (準識別子)	入院日数 (非センシティブ変数)	感染症有無 (センシティブ変数)	手術有無 (センシティブ変数)	出来高点数 (非センシティブ変数)
0000001	男	24	2	有	無	385
0000002	女	30	30	有	有	17200
0000003	女	27	5	無	無	400
0000004	男	20	16	無	有	1200
0000005	女	34	4	無	無	539
0000006	女	29	52	有	有	237615

※ 変数の分類は一例であり，実際には所定の基準に従って適切に分類されることが望ましい．

表 2: 2-匿名テーブルの例

性別 (準識別子)	年齢 (準識別子)	入院日数 (非センシティブ変数)	感染症有無 (センシティブ変数)
男	[20..24]	2	有
女	[30..34]	30	有
女	[25..29]	5	無
男	[20..24]	16	無
女	[30..34]	4	無
女	[25..29]	52	有

※ [x..y] の表記は x 以上 y 以下を意味する．

る Pk-匿名性 [10] と呼ばれるリスク指標，及び Pk-匿名性を満たすテーブルを作成する Pk-匿名化アルゴリズムが提案されている [11]．Pk-匿名性は特定個人のレコードを $1/k$ を超える確率で識別できるかどうかをリスク指標とする，身元開示リスクを考慮した k-匿名性の派生指標である．またその他の派生指標として，属性開示リスクを考慮した ℓ-多様性や t-近似性等が提案されている．詳細は例えば [12] を参照されたい．

3 集合匿名化の課題と対策

k-匿名化は準識別子が増えるほど個人データの一般化や削除の度合いが増し，次元の呪いと呼ばれるデータの有用性が低下する傾向がある．これは準識別子の増加に伴い準識別子の値の組み合わせが指数的に増加し，準識別子の値の組が k 個以上同一となるレコードを作成するために一般化や削除の度合いを強めざるを得ないためである．Pk-匿名化も準識別子の増加に伴い個人データの撹乱度合いが増し，k-匿名化と同様の傾向が想定される．

一方，準識別子の基準が日本においては合意された見解がなく [13]，安全サイドに立てば多くの変数は準識別子とせざるを得ない．また個人データの再利用性の観点からも各変数を準識別子とみなすことが望ましい．すなわち，ある個人データから特定の (非) センシティブ変数の値を重複利用すると，図 1 に例示するように，テーブル T_1, T_2 はそれぞれ 2-匿名性を満たすものの，入院日数をキーに 2-匿名性を満たさないテーブル T_3 が作成可能となり身元開示リスクが高まる結果となる．

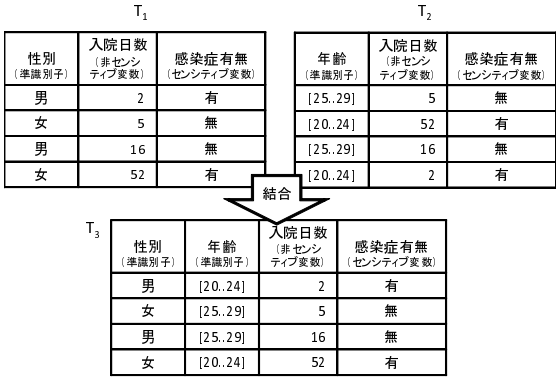

図 1: (非) センシティブ変数の重複利用問題

次元の呪いを回避する単純な方法は準識別子を減らすことだが，先に述べたとおりプライバシ保護の観点からは，多くの変数は準識別子とみなすべきであり，多変数からなるデータを扱う場合は特に匿名性と有用性の相反する命題を抱えることになる．そこで本稿では，特定の変数

の組からなる部分テーブルを元のテーブルからいくつか抽出し，各々の部分テーブルに集合匿名化を施すアプローチを考える．すなわち直感的には図1における結合と逆向きの操作を行う．これにより変数の少ない各々の部分テーブルについては一般化や撹乱等の度合いの軽減が期待できるが，図1で挙げたような結合による匿名性低下の問題や，部分テーブルの組の有用性については検討が必要であり，次節で考察する．

4 考察

4.1 部分テーブルの組の集合匿名化

部分テーブルの組の開示リスクについては，以下の指標が提案されている．

定義 4.1 (k-組み合わせ匿名性 [14])**.** r 個のテーブル $M = \{M_1, \ldots, M_r\}$ について，M に含まれる各準識別子の定義域の直積を D とするとき，任意の $t \in D$ が以下の少なくとも一方を満たすとき，M は k-組み合わせ匿名性を満たすという．

1. M と矛盾しない任意のテーブル T について，t は T に含まれない．
2. t が k 個以上含まれる，M と矛盾しないテーブル T が存在する．

すなわち k-組み合わせ匿名性は，M から k-匿名性を満たさないテーブルを構成できるかどうかをリスク指標としている．

以下に k-組み合わせ匿名性を満たす単純な方法を与える．そのために先ず**多重 k-匿名性**を定義する．

定義 4.2 (多重 k-匿名性)**.** r 個のテーブル $M = \{M_1, \ldots, M_r\}$ について，以下の条件を満たすとき，M は多重 k-匿名性を満たすという．

1. $M_1, \ldots, M_r$ は k-匿名性を満たす．
2. $M_1, \ldots, M_r$ は互いに同一の (非) センシティブ変数を持たない．

すると k-組み合わせ匿名性との関係について以下が成り立つ．

命題 4.1. M は多重 k-匿名性を満たすならば k-組み合わせ匿名性を満たす．

Proof. 定義 4.1 の $t \in D$ について，t を含むレコード Rc_1 を M から矛盾なく得られるとする．すると定義 4.2 の条件 1 より，Rc_1 と同一の変数からなる，t を含む $k'(\geq k)$ 個のレコード $Rc_1, \ldots, Rc_{k'}$ を M から得ることができる．そして $Rc_1, \ldots, Rc_{k'}$ から構成されるテーブル M_0 の (非) センシティブ変数は条件 2 より互いに異なるため，M_0 の任意のレコードは M と矛盾しない． □

ある部分テーブルの組 $M = \{M_1, \ldots, M_r\}$ について，M の全ての変数を準識別子とみなして $M_1, \ldots, M_r$ を k-匿名化すれば，M は多重 k-匿名性を満たし，命題 4.1 より M は k-組み合わせ匿名性を満たすことがいえる．

一方，撹乱・再構築法については，筆者らは先行して多重 k-匿名性と同様の考え方に基づく**多重 Pk-匿名性**を定義している [15]．簡単にいえば，M の全ての変数を準識別子とみなして $M_1, \ldots, M_r$ を Pk-匿名化すれば，M は多重 Pk-匿名性を満たし，M に含まれる特定個人のレコードを $1/k$ を超える確率で識別することはできない．

4.2 部分テーブルの組の有用性

複数の変数からなるデータについて変数の関連性を統計的に扱う手法は多変量解析と呼ばれ，回帰分析や主成分分析，因子分析等が知られる．特に代表的な多変量解析の一つとして重回帰分析がある．重回帰分析は，目的変数と呼ばれる変数と，説明変数と呼ばれる変数との関係を式 (回帰式) で表して変数間の関係を分析する回帰分析において，説明変数が複数からなる分析モデルである．

回帰式の導出には最小二乗法がよく用いられるが，通常レコード単位で変数の値を入力する．例えばオープンソースの統計解析ソフトウェア R[16] では，最小二乗法による回帰分析を行う関数 'lsfit' が提供されているが，変数を分けて処理する方法は自明でない．しかし重回帰分析

には変数を入れ替えながら適切な回帰式を導出する方法も知られており，その中でも**変数増加法**は少ない変数の組から分析を行い，適切な回帰式が導出できるまで変数を増加させていく．すなわち，必ずしもレコード単位で変数の値を入力する必要は無く，一部の変数の組から分析を行うことに適している分析モデルと考えられる．R では線形モデルによる回帰分析を行う関数 'lm' のオプションとして変数増加法が提供されている．

このように，一部の変数の組を入力して試行的に分析を行うような多変量解析においては，多重 k-匿名化や多重 Pk-匿名化と相性がよく，適合性が高いと考えられる．

5 実験評価

実際の医療情報を多重 k-匿名化し，重回帰分析の適用可能性を実験的に検証する．

5.1 対象データ

愛媛大学医学部附属病院が医療統計情報プラットフォーム研究会 [17](CISA: Platform for Clinical Information Statistical Analysis) に提出しているデータ (以降 CISA データと呼ぶ) を用いて実験を行う．CISA は研究会に参加している大学附属病院からレセプト・DPC データ [18] に基づいた独自形式の医療保健情報を収集・蓄積しており，日本では最大級規模の医療保健情報データベースを構築している．

実験対象とする CISA データは，2007 年 1 月から 2011 年 3 月分について，性別，生年 (1926〜2011)，入院日数 (0〜最大 1551)，感染症有無，手術有無，出来高点数 (6 桁の数値) の 6 変数からなるテーブルとし，これら全ての変数を準識別子とみなして評価する．入院日数が 1551 日 (4 年 3 か月) を超える異常値を含むレコードは削除し，最終的にレコード数は 29,669 となった．

5.2 実験方法

CISA データに対して多重 k-匿名化を施したデータを重回帰分析して得た結果を本来の CISA データから得られる結果と比較する．変数が少ない (実際には変数の値の組み合わせの数が小さい) テーブルほど良い結果が得られれば，4.2 節で例示した変数増加法等への適用により，多重集合匿名化が有効と考えられる．

重回帰分析の計算は R(version 2.14.1) の lm 関数を用い，出来高点数を目的変数，それ以外を説明変数と設定する．lm 関数から出力される値のうち，回帰式の係数 (回帰係数) の推定値，標準誤差，t 値，p 値，及びその有意レベル (最低 0 から最高 4 までの 5 段階) について，元の CISA データから得られる値を真の値とみなし，CISA データに多重 k-匿名化及び多重 Pk-匿名化をそれぞれ施したデータから得られる値 (測定値) の誤差または誤差率を求める．ここで

$$\text{誤差} = \{\text{測定値}\} - \{\text{真の値}\}$$
$$\text{誤差率} = \frac{\{\text{測定値}\} - \{\text{真の値}\}}{\{\text{真の値}\}}$$

と定める．

分析にあたり以下の前提を置く．

- 性別は男性を 1, 女性を 0，感染症有無及び手術有無は「有」ならば 1,「無」ならば 0 とそれぞれ数値化する．
- 以下の一般化を行う (一般化後の CISA データを一般化 CISA データと呼ぶ)．
  - 生年 : 5 年刻み ('1928' を '1925' とする等，自身を超えない最大の 5 の倍数に変換)
  - 入院日数 : 上位 2 桁以下を 0 に変換 (例えば '22', '123' はそれぞれ '20', '100' に変換)
  - 出来高点数 : 上位 3 桁以下を 0 に変換
- k-匿名化は，一般化 CISA データについてこれ以上の一般化は分析に支障をきたすと判断し，同一のレコードが k 個未満のものを削除する方法を採用する．
- 一般化 CISA データを多重集合匿名化する対象の部分テーブルは以下の 5 段階の変数の組み合わせからなるデータとする．

Lv.1 性別・出来高点数：740 通り

Lv.2 性別・感染症有無・手術有無・出来高点数：2,960 通り

Lv.3 性別・入院日数・出来高点数：14,800 通り

Lv.4 性別・生年・感染症有無・手術有無・出来高点数：65,120 通り

Lv.5 性別・生年・入院日数・感染症有無・手術有無・出来高点数：1,302,400 通り

一般化 CISA データの変数の値について，生年は 22 通り (1905〜2010)，入院日数は 20 通り (最小 0，最大 400)，出来高点数は 370 通り (最小 0，最大 980000) となった．すると一般化 CISA データの変数の値の組み合わせ数は，レコード数 29,669 よりもかなり大きい 1,302,400 通りとなり，このままではデータの削除数や撹乱の度合いが大きくなることが予想される．

5.3 実験結果

5.3.1 一般化データの誤差

一般化 CISA データから得られる重回帰分析の計算結果の誤差を図 2 に示す．

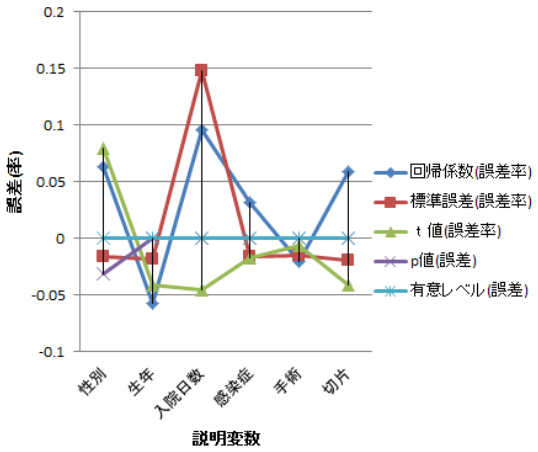

図 2: 一般化 CISA データの誤差

有意レベルに誤差は無く，回帰係数の誤差率も多少のばらつきはあるものの-5.8〜9.6%程度にとどまっている．

5.3.2 k-匿名化によるレコード数の変化

5.2 節で定めた k-匿名化の方法に従い，$k = 2, 3, 5, 10, 100$ について一般化 CISA データのレコード削除を行う．テーブル Lv.1〜5 の k-匿名化データのレコード数を図 3 に示す．

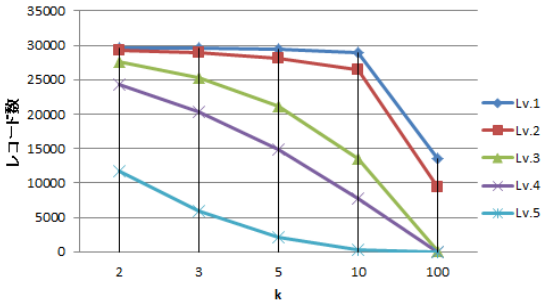

図 3: k-匿名化によるレコード数の変化

変数の値の組み合わせ数が最も多い Lv.5 では，$k = 2$ であっても半分以上のレコードが (非無作為に) 削除されてしまい，誤差の影響が懸念される．また $k = 100$ とした場合は，変数の値の組み合わせ数が最も少ない Lv.1 であっても，前記同様に半分以上のレコードが (非無作為に) 削除されてしまう．特に Lv.3 以上ではレコードが全く残らず分析不可能な状態となっている．

5.3.3 多重 k-匿名化データの誤差

多重 k-匿名化データ (Lv.1〜Lv.5) から得られる重回帰分析の計算結果を図 4 に示す．ここでは $k = 2$ として各変数の回帰係数，標準誤差，t 値，p 値，及び有意レベルの誤差の絶対値の平均を表している．

変数の値の組み合わせ数が比較的少ない Lv.1 や Lv.2 のデータは誤差が小さいことが分かる．Lv.3 以上では明らかに誤差が無視できないレベルになっているが，Lv.3 の結果の詳細をみると (表 3)，誤差が大きい回帰係数及び t 値は，性別における高い誤差の影響を受けていることが分かる．しかし性別は元々 p 値が大きく，また p 値が小さい入院日数については誤差が非常に小さいことから，全体でみれば分析結果への影響は少ないと考えられる．

表 3: Lv.3 の多重 k-匿名化データから得られる重回帰分析の結果

	回帰係数	標準誤差	t 値	p 値	有意レベル
性別 (真の値)	397.433	675.407	0.588	0.556	0
性別 (測定値)	-822.919	588.174	-1.399	0.162	0
性別 (誤差 (率))	-3.07	-0.129	-3.379	-0.394	0
入院日数 (真の値)	312.510	6.424	48.650	<2e-16	4
入院日数 (測定値)	343.244	6.811	50.393	<2e-16	4
入院日数 (誤差 (率))	0.098	0.06	0.036	0	0
切片 (真の値)	22204.667	552.106	40.218	<2e-16	4
切片 (測定値)	20982.233	491.123	42.723	<2e-16	4
切片 (誤差 (率))	-0.055	-0.11	0.062	0	0

括弧内の数値は p 値・有意レベルは誤差，それ以外は誤差率

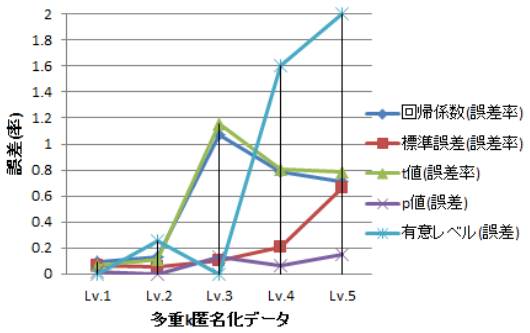

図 4: 多重 k-匿名化データの誤差

次に，Lv.2 の多重 k-匿名化データを対象とし，k の値を変化させた場合の誤差率の推移を図 5 に示す．図 4 同様，各変数の回帰係数，標準誤差，t 値，p 値，及び有意レベルの誤差率の絶対値の平均を表している．

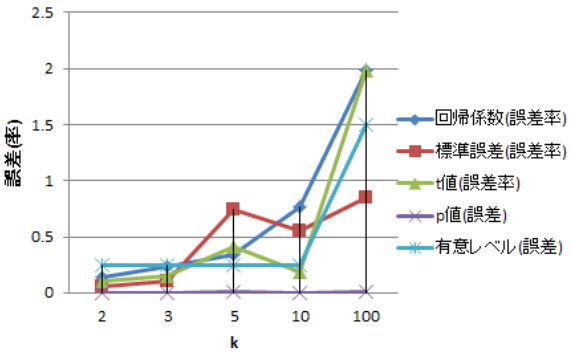

図 5: Lv.2 の多重 k-匿名化データの誤差 (k の値を変化)

$k = 5$ のとき誤差に若干特異な上昇がみられるが，全体的には k が大きくなるに従い誤差が上昇していることが分かる．

6 おわりに

個人データの安全な利活用に資する集合匿名化について，多変量解析のように多変数データを扱う場合は，次元の呪いと呼ばれるデータの有用性低下の問題が指摘されていることを踏まえ，複数のデータに分割して集合匿名化を行う多重集合匿名化に着目し，匿名性の考察や実験による有効性検証を行った．多重集合匿名化データは重回帰分析と適合性が良く，実際の医療情報に多重集合匿名化データを用いた重回帰分析を行いある条件下では誤差が小さいことを実験的に確認した．ただし今回は単純に匿名化前後の分析結果の誤差を評価したのみであり，今後はより高度な統計的手法により匿名化データの有用性を評価検討していく予定である．また，図 3 で示したように，変数が少ないテーブルであっても k の値が大きければ有用性の低下が問題となり得る．筆者らは [19] において，Pk-匿名化が k の値の増加に対して影響を受けにくいことを実験的に確認している．本稿では多重 k-匿名化の実験評価を与えたが，多重 Pk-匿名化との比較は今後の検討課題である．さらに，集合匿名化は ℓ-多様性や t-近似性等，様々なリスク指標に基づく手法が存在していることから，他のリスク指標に基づく多重集合匿名化についても匿名性の考察や実験評価を行うことを今後の課題としたい．

参考文献

[1] Sweeney, L.: Uniqueness of Simple Demographics in the U.S. Population, LIDAP-WP4, Carnegie Mellon University, Laboratory for International Data Privacy, 2000.

[2] 情報大航海プロジェクト パーソナル情報検討チーム: パーソナル情報の利用ガイドライン (案) <利用の在り方に関する提言> 平成 22 年 3 月, http://www.meti.go.jp/policy/it_policy/daikoukai/igvp/index/h22_report/sub/05.pdf.

[3] Sweeney, L.: k-anonymity: A Model for Protecting Privacy, *Int'l Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol.10, No.5, pp.557–570, World Scientific Publishing, 2002.

[4] Aggarwal, C.: On k-anonymity and the Curse of Dimentionality, *Proc. VLDB 2005*, pp.901–909, ACM, 2005.

[5] 独立行政法人 統計センター: 統計データ開示抑制に関する用語集 改訂版 (対訳) 2005 年 8 月, http://www.nstac.go.jp/services/pdf/skk-yogosyu2.pdf, 参照 Jul. 6, 2012.

[6] 竹村彰通: 個票開示問題の研究の現状と課題, Vol.51, No.2, pp.241–260, 統計数理, 2006.

[7] Lambert, D.: Measures of Disclosure Risk and Harm, *Journal of Official Statistics*, Vol.9, No.2, pp.313–331, Statistics Sweden, 1993.

[8] Agrawal, R. and Srikant, S.: Privacy-Preserving Data Mining, *Proc. SIGMOD 2000*, pp.439–450, ACM, 2000.

[9] Agrawal, R., Srikant, R., and Thomas, D.: Privacy Preserving OLAP, *Proc. SIGMOD 2005*, pp.251–262, ACM, 2005.

[10] 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, CSS2009 論文集, pp.763–768, 情報処理学会, 2009.

[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, SCIS2010 予稿集 (CD-ROM), 電子情報通信学会 (2010).

[12] Aggarwal, C. and Yu, P.: Privacy-Preserving Data Mining: Models and Algorithms, Springer-Verlag (2008).

[13] 木村映善: k-匿名性を利用した医療保健情報の利用可能性についての考察 — 国内外の医療情報利用に関する事例から —, 信学技法, Vol.111, No.484, SITE2011-50, pp.223–228, 2012.

[14] Kifer, D. and Gehrke, J.: Injecting Utility into Anonymized Datasets, *Proc. SIGMOD 2006*, pp.217–228, ACM, 2006.

[15] 五十嵐大, 千田浩司, 濱田浩気, 菊池亮: 秘匿計算とランダム化によるハイブリッド匿名化システム, SCIS2012 予稿集 (CD-ROM), 電子情報通信学会, 2012.

[16] The R Project for Statistical Computing, http://www.r-project.org.

[17] 医療統計情報プラットフォーム研究会, Platform for Clinical Information Statistical Analysis(オンライン), http://www.cisa.jp/index.html.

[18] 医学通信社編集部, DPC 点数早見表 2011 年 4 月増補版, 医学通信社, 2011.

[19] 千田浩司, 木村映善, 他: 撹乱手法を用いたプライバシー保護医療情報分析の実験評価, 医療情報学連合大会論文集, pp. 689–692, 2011.