

## k-匿名レベルと属性区分数の関数近似式とその評価方法の提案

小栗 秀暢<sup>†1†3</sup> 曾根原 登<sup>†2</sup> 松井 くにお<sup>†3</sup> 黒政 敦史<sup>†3</sup> Mohammad Rasool Sarrafi Aghdam<sup>†1</sup>

†1 総合研究大学院大学 複合科学研究科 情報学専攻

101-8430 東京都千代田区一ツ橋 2-1-2 oguri.hidenobu@nifty.co.jp

†2 国立情報学研究所 †3 ニフティ株式会社

**あらまし** データの結合性とk-匿名性を同時に満たす処理は、データ提供者側の個人再識別リスクと計算コストの増加を伴うため、データ利用を阻害する要因となっている。本論文は k-匿名化における個人識別性kと属性区分数の累乗関数による予測式によって、匿名化データの識別可能性の増加リスク領域と結合性を評価する方式を検討した。実データに対して評価を実施し、予測式とk値の設定によって識別可能性と結合性が制御できることを確認した。

### A proposal on data prediction and evaluation method, using the relationship between division number of dataset and k-anonymity

Hidenobu Oguri<sup>†1†3</sup> Noboru Sonehara<sup>†2</sup> Kunio Matsui<sup>†3</sup> Atsushi Kuromasa<sup>†3</sup> Mohammad Rasool Sarrafi Aghdam<sup>†1</sup>

†1 The Graduate University for Advanced Studies, School of Multidisciplinary, Informatics Department, Tokyo, Japan

2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 oguri.hidenobu@nifty.co.jp

†2 National Institute of Informatics †3 NIFTY Corporation

**Abstract** If the two data providers need to connect their datasets each other and satisfy their k-anonymous level, it causes increase of re-identification risk and computational cost. The method that can connect datasets securely is required widely. In this paper, we examined the method to evaluate the increase risk area and the binding of identifiability of anonymous data, by the relationship of personal identifiability and Prediction formula by a power function of the attribute number of segments. We also evaluated them by real data, it was confirmed that can be controlled binding and identifiability by setting the prediction expression and k-values.

## 1 はじめに

近年の個人情報保護意識の高まりによって、個人情報を保持する事業者は、情報の有効利用を促進する施策と、情報の漏洩や不正利用を防止する施策を両立させることが求められる。

企業が情報を活用して新規顧客獲得や事業拡大を達成するためには、自社が保持する情報だけでは分析対象が限定されてしまい、有為な

施策や改善提案に結びつけることは難しい。そのため、他の事業者の保持するデータなどと結合し、比較・分析する仕組みが検討されている。

情報内からセンシティブな要素を排除することでコンプライアンスリスクを軽減させ、データを他社に提供し、分析やマーケティング等に利用する手段として、個人情報の匿名化技術が有望視されている。特に k-匿名化処理[1]をはじめとする個人の特定性・識別性を低減させる手法は、他

のデータベースとの名寄せによる結合や公開情報同士の再結合と再識別化を防ぐ手段として効果的である。

高いレベルで  $k$ -匿名化が施された情報群は、再識別化による攻撃や、悪用の可能性が低くなるため、個人情報よりも簡単な手続きで利用可能となり、第三者への情報提供、ノウハウの共有やマーケティング分析、協調フィルタリングによるレコメンドエンジン[2]等への活用が期待できる。

だが、匿名化データの流通と結合による分析には、再識別化の危険性が常に存在する。そのため多くの匿名化データは再識別を防ぐために抽象化された概念に書き換えられ、適切なノイズを加えるなどの処理を施される。そのため、分析を行うべき対象データとの抽象化レベルの違いによって、定義や情報の粒度が変化し、結合性が失われることによって分析目的が達成できない場合がある。

そこで本稿では、データ同士の結合性を維持するため、個人情報を開示せずに結合可能な属性の区分数を予測式で算出した場合の匿名化データの結合性を評価する手法を提案する。また、その評価手法を共通のユーザ群から成り立つ実サービスの利用者数分布に適用し、識別可能性が増加する領域を明らかにした。

本論文の構成は次の通りである。第 2 章にて現状の匿名化データ結合の問題点を述べ、第 3 章にて匿名化処理に関する従来研究と比較する。第 4 章にて  $k$ -匿名レベルの予測式を説明し、第 5 章にて実データによる実験を行う。第 6 章にてその評価方法の提案と実験を行い、第 7 章で結果をまとめる。

## 2 匿名化データの結合に関する問題点

匿名化データの利用に伴い、図1のような一般化階層の相違によってデータの結合性が失われ、再識別可能性が高まり、匿名化処理コストが増大する問題が発生する場合がある。

図1は、異なる事業者間でマーケティング情報等を匿名化して共同で活用する状況を表したものである。例えば同じ商店街における顧客データ共有や、商品の販売代理店同士の販売状況の

比較などを想定すると解りやすい。

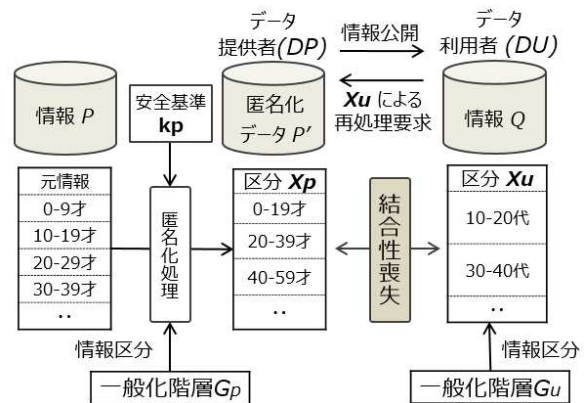


図1 結合できない匿名化データの例

データ提供者 DP は個人情報 P に対して安全基準  $k_p$  を満たす匿名化データ  $P'$  を公開した。だが、 $P'$  はデータ利用者 DU の持つ情報 Q との結合ができない。そのため、DU は一般化階層  $G_u$  から  $x_u$  を指定して DP に再度の匿名化処理を要求する。図 2 にてそのシーケンス図を示す。

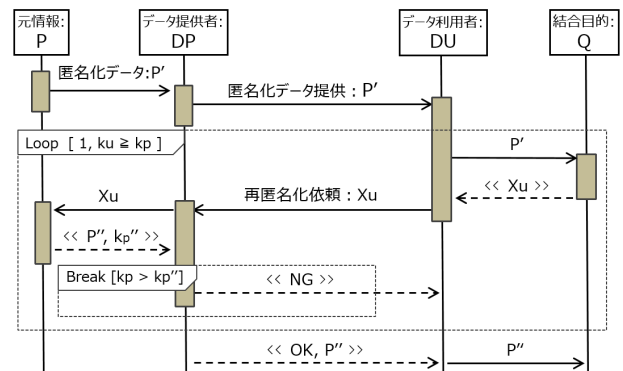


図2 結合できないデータ処理のシーケンス図

DP は安全基準  $k_p$  と一般化階層  $G_p$  から求められた属性区分  $x_p$  (=0-19 才, 20-39 才...) による匿名化データ  $P'$  を提供する。DU が保持する情報 Q は区分  $x_u$  (=10-20 代, 30-40 代...) で情報が統一されており、区分  $x_p$  の情報との結合性が失われている。DU は DP に対して  $x_u$  での再匿名化処理を要求するが、 $x_u$  区分の結果  $k_u$  が、DU の安全基準  $k_p$  を達成できるかは元情報 P に問い合わせないと判明しない。DP は求められた  $x_u$  に対する匿名化処理結果を DU に提供するが、このとき「匿名化できない」という情報が知られるだけでも背景情報が漏洩する可能性がある。

例えば、年齢属性:  $x_n$  に対する結果  $k_u$  = [99 才: 2 人] が出力されたのに対して、年齢\*性別属性:  $x_{nm}$

に対する結果が出力できないことが判明した場合、 $k_{nm} = [99 \text{ 才男:1 人}, 99 \text{ 才女:1 人}]$ が判明し、識別可能性が高まる。このような複数回のクエリ発行による情報漏洩リスクは、クエリ監査問題[3] (Query Auditing)として提起されている。

このような情報漏洩リスクは匿名化処理の結果予測によって発生を抑えることができる。だが匿名化処理の予測は一般的に難しく、全ての方式に当てはまるものは存在しない。本稿では、属性同士の組み合わせによって発生する $k$ -匿名レベルの減少度を近似式によって予測し、結合性を維持した匿名化情報を生成し、評価する。

### 3 データの匿名化に関する従来研究

まず、匿名化とは、個人情報やパーソナル情報などのパーソナル情報を加工して、他の情報との容易照合性を減少させる処理のことである。

パーソナル情報とは「属性」と「属性値」として表現されるユーザに関する情報であり、あるユーザのパーソナル情報をテーブルのレコードとして表現する。そして、単一の属性ではユーザを特定できないが、複数組み合わせるとユーザを特定できる可能性のある属性の組合せを準識別子 (quasi-identifier, QID) と呼ぶ。

また、ユーザを特定された状態で開示されることが望ましくない属性をセンシティブ属性 (sensitive attribute : SA) と呼ぶ。この時、もし攻撃者があるユーザの QID の属性値を知っていたとすると、そのユーザのレコードを特定できてしまい、SA の属性値を知られてしまう。これを防ぐために、QID の属性値を一般化して、より抽象的な値にする方法が知られている。そして、QID の属性値によって識別されるレコードが少なくとも  $k$  個 ( $k > 1$ ) 以上ある場合、そのテーブルは  $k$ -匿名性を満たすという[1]。

$k$ -匿名化を実現するための手法として、Datafly 方式[1]や  $\mu$ -Argus 方式 [5][6][7]などのアルゴリズムによって情報を書き換える(Recoding) 処理が主に使われており、公共データや医療データの配信システムとして利用されている[8]。情報の Recoding の方法は、大きく分けて、局所的な変更である Local Recoding と、属性値全体の統

計情報から変更を行う Global Recoding の二種類が存在する。本稿では Global Recoding について詳細を記述する。

Datafly 方式をはじめとする Global Recoding では、主に各属性値の出現数を計測しながら、匿名化条件を満たさない属性値を抽象度の高い候補に書き換えるという、一般化階層型の集合匿名化処理を行う。値の書き換えには値一般化階層 VGH、属性一般化階層 DGH [9]、マージナル [10]やマイクロアグリゲーション[11]などを利用し、出現数の少ない属性値を抽象度の高い属性値や代表値に書き換える処理を行う。

Global Recoding では値の書き換え前後に各属性値の出現数の検証を行う。書き換えの試行と検証は 1 度とは限らず、匿名化条件やデータの利用用途の条件を満たすまで、値一般化階層の修正と値の書き換え、出現数の検証を繰り返し行うことになる。このような匿名化処理に伴う計算回数を削減するため、incognito[12]方式などの処理削減アルゴリズムが多く提案されている。だが、情報量や  $k$ -匿名化レベルによる計算量の減少に関するロジックでは、単体の分析結果としての精度は向上するが、複数の情報同士の定義を統合し、同一の基準で結合性を高める目的には利用できない。

[13][14]などにおいて情報の精度の向上や情報有用性を保持しつつ匿名化するアルゴリズムも提案されているが、複数情報の結合容易性は検討されていない。

特定区分による Recoding に利用しやすい方式として OLA[15] (Optimal Lattice Anonymization) は、情報量に着目して最も情報量の大きい組み合わせを導き出す方式を提案している。出現した属性値の組み合わせを情報量の多い順番にソートし、最も情報量が多く、匿名化条件を満たした群を”GOD :Globally Optimal Dataset”として利用する。GOD 値は最も情報量が多い情報だが、データ利用者の求める情報とは限らないため、クエリ監査問題が発生する可能性がある。

また、複数回の情報提供時に発生するクエリ監査問題に対しては、提供情報のリスクを算出するオフライン監査方式[3]や、情報の提供先や時

期に問題がある場合、応答ノイズをランダムに加えるシミュレータブル監査方式[16]が提案されているが、これらは情報提供者側が情報を秘匿する手法であり、他の情報との結合性を高めることはできない。また、 $k$  値は 1 以上の整数値であることが求められるため、ノイズによってデータの整合性が取れない場合が出てくる。

そこで我々は、匿名化データの作成に利用した一般化階層の持つ情報区分数を用いた匿名化レベルの予測式を利用する方式を提案する。

#### 4 k-匿名レベル予測近似式

研究[18]では、マーケティング等の目的でユーザを収集した群において、 $k$ -匿名レベルの減少率は属性情報の区分数  $x$  と累乗近似型の関係性で表すことができることを確認した。

$k$ -匿名レベルは、ある情報を複数の属性値に区分し、クラスタを作成した際の最小出現数  $k$  の値である。そのため、Local Recoding に採用される一様分布では、 $k = P * x^{-1}$  で  $k$  値を表現することができ、Global Recoding で採用される正規分布の場合  $k = P * x^{-2}$  の形で表すことができる。

通常のマーケティングデータは標準正規分布に近い形で収集されるが、実際にはノイズが多く存在するために分布が変化する。そこで、少数の  $k$ -匿名結果をサンプリングし、最小二乗法における傾き  $\beta$  と切片  $\alpha$  を用いて累乗近似式  $k = \alpha x^\beta + 1$  を作成する方法が最も相関係数が高い予測式を構築することが出来た[図 3][18]。

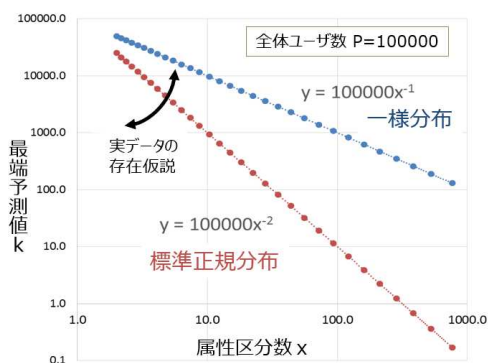


図 3 一様分布と標準正規分布の近似式

一般的に、 $k$  値の出現可能性は属性値の分散と適用する一般化階層に依存し、 $k$  値に規則性は無いと考えられる。

だが、通常のマーケティング分析等で利用する一般化階層は、利用データの出現クラスタに対して、ある程度適合した情報区分を採用していると期待できる。

そこで、何らかの目的に沿って顧客をターゲットングして収集された群に対して、その収集や管理に利用している顧客区分を一般化階層に変換して匿名化処理に適用した場合、ある程度、近似式等による定式化が可能という仮説を考える。

また、属性の区分数についても、区分の方法は 1 種類とは限らない。例えば「4 区分」を構成する群は 2 属性  $\times$  2 区分と 1 属性  $\times$  4 区分など、複数の可能性が存在する。それら区分数が同じ群を平均化してサンプリングしたならば、予測式を生成する群として普遍性が出ることを期待できる。

本予測式について、情報の結合性を重視した評価指標を検討する。

#### 5 k-匿名性予測式の比較と検証

実験は、ポータルサービスに属する 1635 個のサービスに対して、2013 年 10 月に 1 度以上課金決済を行った顧客群に対して行った。顧客群は、対象顧客の中から性別/年齢/都道府県の 3 つの要素を全て入力している顧客を抜き出した[図 6]。各サービスごとに登録している顧客の属性値を一律に分類した結果を求め、群における最小値 (=  $k$ -匿名レベル) を取得した。また、一般性を高めるため、国勢調査(2010 年)のデータ量を 1/1000 に変更し、他のデータと比較可能な量に変更して同様の処理を行った。以下、国勢調査群は全てこの 1/1000 の群を指すものとする。

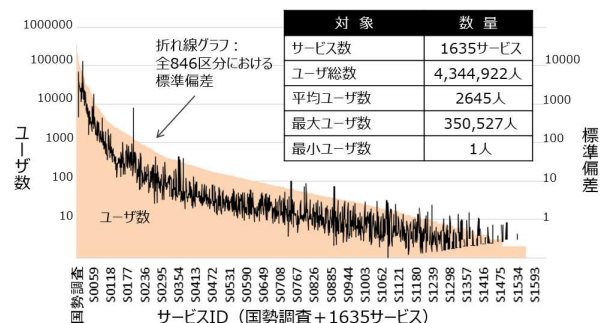


図 4 対象となったサービス群の顧客数と標準偏差

図 4 のサービス群は、特定の地域にしか提供しないもの、男性の利用が多いものなど、同一の

基準を用いた場合に k 値が低くなる可能性が高いものも多く含んでいる。

それに対して適用する匿名化パターンとなる値一般化階層は表 1 の基準で作成した。これらは n 社内で一般的に利用される情報区分に統一されている。

表 1 顧客群に適用する一般化階層

属性	区分数	分類1	分類2	分類3	分類4	分類5	分類6	分類7	分類8	分類9
性別	2区分	男性	女性							
年代	3区分	未成年		成人				老人		
	5区分	20代以下			30代	40代	50代	60代以上		
	9区分	0代	10代	20代	30代	40代	50代	60代	70代	80代以上
地域	2区分	東日本				西日本				
	9区分	北海道	東北	関東	中部	近畿	中国	四国	九州	沖縄
	47区分	北海道,青森...沖縄までの47都道府県								

まず, k 値の累乗近似式  $k = \alpha x^\beta + 1$  と比較して, 他の分布を評価し比較する。表 2 は, 分布の種類とその相関係数である。

表 2 分布の種類と実データとの相関比較

分布種類	$\alpha$ 値	$\beta$ 値	相関係数
一様分布	127794	-1.000	0.985
標準正規分布	127794	-2.000	0.979
実データ累乗近似	114659	-1.414	0.997

この結果を受け, n 社の実データに対しても同様の累乗近似式を作成し, 相関係数を求めた。

図 5 は, 累乗近似式, 標準正規分布, 累乗近似式の 3 種類について, 各ユーザー群の規模ごとの値を計測したものである。国勢調査は正規分布の形よりも一様分布の相関係数が高く出ており, ユーザが均一に分布していることがわかる。だが, 他の群においては, ほぼ相関係数は変化しておらず, ユーザの規模に関係なく累乗近似式が利用できることが解る。

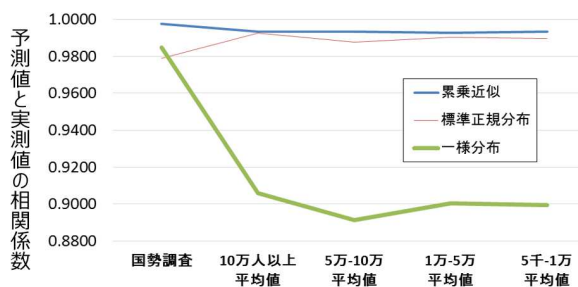


図 5 累乗近似式と他分布の相関係数比較

だが, 相関係数のみを利用した場合, 各値における誤差が大きく, 予測値として利用できない場合がある。それを回避するため, 平均絶対誤差  $\Sigma(y_i - \bar{y})/N$  を求め, 他の分布との比較を行

った[図 6]。この指標においては, 標準正規分布は 10 万人以上の群において正確性が高くなる。また, 一様分布は非常に誤差が大きく不正確であると言える。

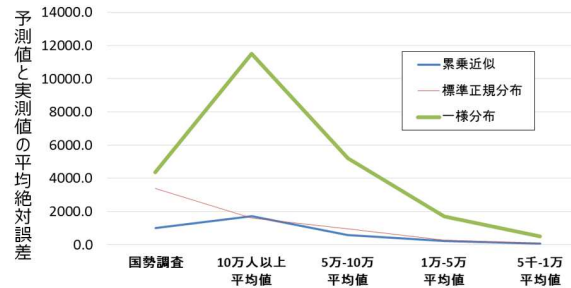


図 6 累乗近似式と他分布の平均絶対誤差比較

このような誤差の発生は, k 値が 1 に向けて収束していく性質と関係がある。平均絶対誤差で計測した場合, 値が 1 に向けて収束するため, 本来必要となる詳細な情報における誤差は相対的に影響が少なくなる。

本方式の利用者の立場に立つと, 大きな k 値の際の予測値の正確性は必要ではなく,  $k > 1$  近辺における区分数の予測の正確性が必要である。そのため,  $k > 1$  の近辺における予測値との誤差と, 区分数が少ない群における誤差は相対化して評価する必要がある。

そこで我々は平均絶対比率(1)を利用した。これは実測値 a と予測値 a' の比の大きい方を取得し, 試行数 P で割り, 平均化したものである。

$$\frac{1}{N} * \Sigma \left( \max \left[ \frac{a}{a'}, \frac{a'}{a} \right] \right) \quad (1)$$

本評価方法を用いて分布ごとの比較を行ったのが図 7 である。10 万人以上のサービスにおいて, 平均絶対誤差の値は標準正規分布が最も相関係数が高かったが, 誤差の比率で見ると累乗近似は全ての群において優れた予測値を出力している。

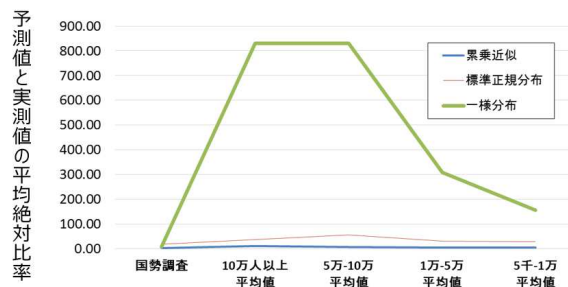


図 7 累乗近似式と他分布の平均絶対比率の比較

これによって累乗近似は他の分布と比べて精度が高いことが判明し、また区分数が少ない場合においても他の分布よりも優れた予測値を出すことが判明した。

だが本評価方法は、値の正確さの指標として利用できるが、結合性の評価としては利用できない。そこで精度の計測に対しては平均絶対比率を用いるが、これに加え結合性と安全性を評価できる指標について検討を行った。

## 6 k-匿名レベル予測式の評価方法提案

本予測式は縦軸にk-匿名レベルk、横軸に情報区分数xを設定し、区分数によってk-匿名性が減少する領域における、DPの識別可能性基準 $k_p$ とDUが望む属性区分 $x_u$ における問題として定義できる[図8]。

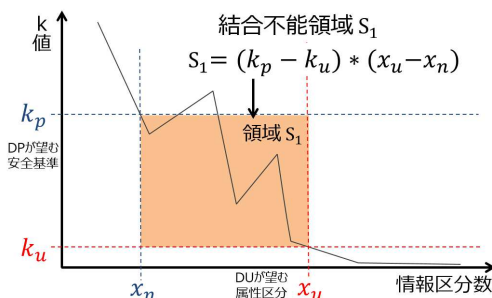


図8 予測式を用いない場合の結合不能領域

$x_u$ におけるk-匿名レベル $k_u$ と、 $k_p$ を満たす最も粗い情報区分 $x_n$ を定義したとき、クエリ監査問題が発生する結合不能領域 $S_1=(k_p-k_u)*(x_u-x_n)$ として定義する。この領域 $S_1$ が小さい程、DUの望む情報が少ないコストで提供されるため、情報の結合性は高いといえる。それに対して、図9はk匿名性の予測式 $f_{(DP)}$ を用いた場合の結合不能領域を定義したものである。

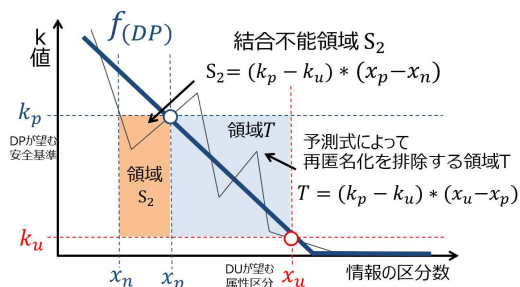


図9 予測式を用いた場合の結合不能領域

予測式 $f_{(DP)}$ における $k_p$ を満たす予測区分 $x_p$ を求めることによって、実際のデータを用いた匿名

化を行わなくとも予測式で排除可能な領域 $T=(k_p-k_u)*(x_u-x_p)$ を新たに設定することができる。これによって結合不能領域 $S_2=(k_p-k_u)*(x_p-x_n)$ まで減少する。

この領域 $S_2$ は、予測における誤差の範囲であり、予測式にて匿名化可能であると予測したが、実際には匿名化できない群である。誤差の発生によって識別可能性が低下する場合、クエリ監査問題を最小限にするため、 $f_{(DP)}$ の角度を修正した予測式 $f'_{(DP)}$ に変更することで領域 $S_2$ を減少させることができる。[図10]

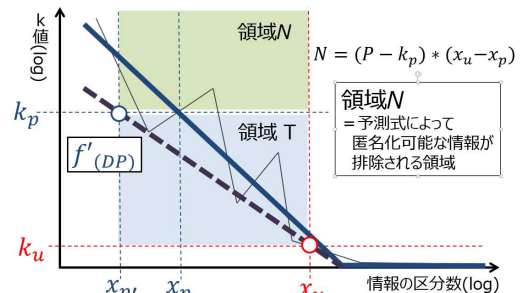


図10 予測式を修正した際に広がる非提供領域N

図11は実サービスA~Fにおける $k>1$ 状態の匿名化予測失敗率と計算量の削減効果のグラフとなる。棒グラフが予測の失敗数、点線は計算削減回数である。折れ線グラフはその予測によって全体の計算回数が削減された割合である。

多くの結果は完全な予測を出せず、1~4項目の予測失敗群を残している。そこで、この数式について安全性の高い群のみを抜き出し、より安全性の高い累乗近似式を作成して評価したのが図12である。

この場合、全ての群において匿名化処理の予測値の精度は向上し、予測失敗率は大きく減少し、識別可能性が高まる領域 $S_2$ が小さくなったことを示す。

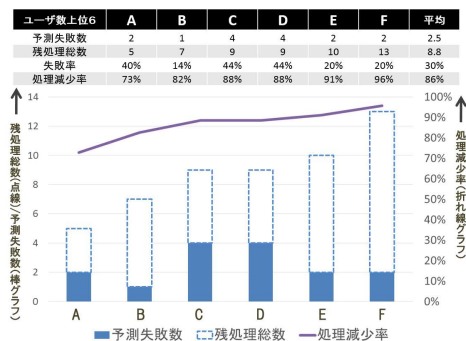


図11 処理削減効果と匿名化予測失敗数

だが  $S_2$  を減少させた場合、予測式によって匿名化不可能と判定され、情報提供リクエストを排除される非提供領域  $N$  が増加する[図 11]. ここに DP が識別可能性リスクを減少させるために近似式を厳しく設定した場合、DU の結合可能な情報範囲が減少するというトレードオフの関係が成り立つ. そのため識別可能性だけでなく結合性も評価できる手法が望ましい.

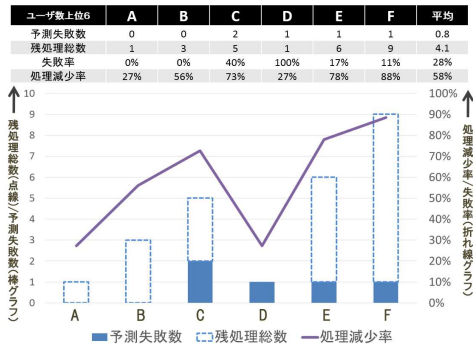


図 12 処理削減効果と予測失敗数(精度向上)

この問題を踏まえ、予測式に関する評価方法として、 $(k_p, x_p)$  を中心とした匿名処理数や、情報利用数にて評価する方法を提案する. [図 13]

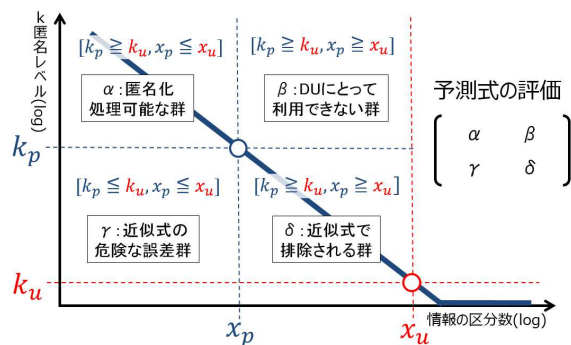


図 13 提案手法の評価方法の提案

図 13 における 4 象限はそれぞれ

- $\alpha$  : 正常な予測範囲
- $\beta$  : DU に対する提供機会の逸失
- $\gamma$  : 情報推定リスク
- $\delta$  :  $k_p$  の安全基準の妥当性

と定義することができ、利用目的に合わせて  $k_p$  と  $f_{(DP)}$  を設定することで制御が可能である.

本評価方式を用いて国勢調査、及び実データの評価を行った結果が図 14 である.

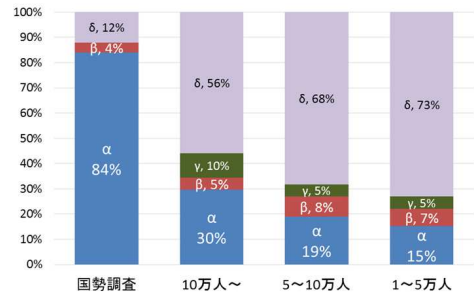


図 14  $k_p=50$  における 4 象限評価結果

図 14 によると、国勢調査群の結果は  $\gamma$  : 結合不能領域が存在しない. だが、その他の実サービス群では 5～10% 程度の  $\gamma$  領域が発生した. この領域の情報提供を制御することでクエリ監査問題を担保できる. また、結合性を高めるため  $k_p$  を低く変更して  $\alpha$  領域を大きくすることも容易である.

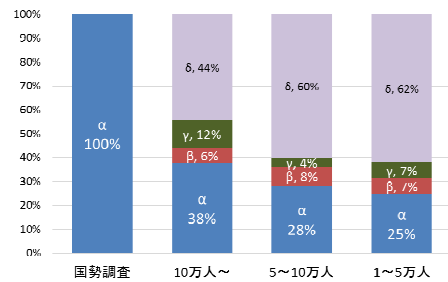


図 15  $k_p=10$  における 4 象限評価結果

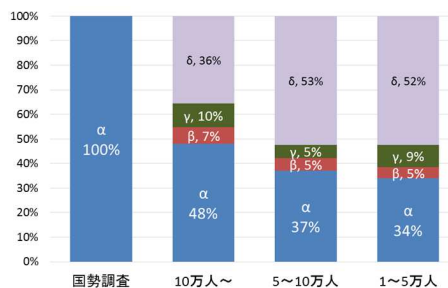


図 16  $k_p=2$  における 4 象限評価結果

上記データの提供基準  $k_p=10$ ,  $k_p=2$  に変化させた結果が図 15, 図 16 である. この場合、国勢調査の情報とは全て提供可能になり、10 万人超のデータにおいては  $\alpha$  値が 30%→48% に向上しているが、結合不能領域  $\gamma$  値は  $k_p=10$  の時に増加している. 誤差の発生による識別可能性低減リスクは、 $k$  値との関係なく増減することを示している. そのため  $k$  値の検討時に本評価によってリスクの少ない値を定めることも可能である.

このような評価を行うことによって、DP は  $k$  匿名レベルの設定基準と、利用する一般化階層の改

良,そして実際に利用者に提供できる情報について,目的に合わせてコントロールすることが可能となる.

## 7 まとめ

本稿における実験結果をまとめる.

(1) 累乗近似式の正確性を評価するためには,平均絶対誤差では  $k$  が小さい値における詳細部分の影響が反映されない.そのため,平均絶対比率を併用して評価することで正確な評価が可能となる.

(2) 情報の結合性を評価する手法として,  $(k_p, x_p)$  を中心とした 4 象限のデータ出現数で評価する手法を提案し,結合不能領域と非提供領域のトレードオフの関係は  $k_p$  の値だけでは制御できない.

(3) 実データを用いた評価を行ったところ,  $k$  値を減少させてデータ提供できる領域を高めたが,その際には結合不能領域は大きく変化しない.サンプル群を変更して数式自体を変化させることで結合不能領域を減少させることできた.

本予測式とその評価方法を用いて,匿名化データの流通における再識別可能性リスクを抑制しつつ,結合性を保つ一般化階層を選択することができる.

今後,個人の行動データの流通が促進されると,本提案のような,複数の DP に対して匿名化処理を行い,かつ属性値を共通化するニーズが高まると予想する.政府レベルの大規模な分析結果と小規模な個人情報を持する企業といった,データの規模や次元が異なる情報同士を,各種の区分方法で組み合わせる機会が増加することで,多様な属性を持つ情報同士を共通化する手法の必要性は高まるだろう.

## 参考文献

- 1) L.Sweeney, k-anonymity: a model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, pp.557-570, 2002
- 2) 本多 克宏,個人情報のクラスタリングによる匿名化と安心・安全な推薦システム (特集 安全社会における情報科学の役割),ケミカルエンジニアリング 58(3), 188-192, 2013-03
- 3) 荒井 ひろみ,佐久間 淳, プライバシーを守った IT サービスの提供技術:6.データベース問合せにおけるプライバシー保護モデル, 情報処理 54(11), 1135-1140, 2013-10-15

- 4) 荒井 ひろみ,津田 宏治,佐久間 淳, 類似検索におけるプライバシー保護のためのクエリ監査法, 情報処理学会 コンピュータセキュリティシンポジウム論文集 2013(4),pp.102-109,2013-10-14
- 5) Mitsubishi Research Institute, Inc. 情報技術研究センター 松崎和賢, データ匿名化の現状に関する一考察. 医療・統計分野を中心とした国内外の動向. 2011-7-8
- 6) 日本情報経済社会推進協会(JIPDEC), パーソナル情報の利用のための調査研究報告書, 2011-3
- 7) Anco J. Hundepool, Leon C. R. J. Willenborg, Statistics,m-and t-ARGUS: Software for Statistical Disclosure Control,Record Linkage Techniques,1997
- 8) 経済産業省 (株)日立コンサルティング, 「行動情報活用型クラウドサービス振興のためのデータ匿名化プラットフォーム技術開発事業」 事業報告書,2012-3
- 9) 村本俊祐, 上土井陽子, 若林真一, k-匿名性を利用したデータ一般化によるプライバシー保護. DEWS2007, 2007.
- 10) D. Kifer and J. Gehrke, "Injecting Utility into Anonymized Datasets", Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, pp. 217-228, 2006.
- 11) 独立行政法人統計センター, ミクロアグリゲーションに関する研究動向 及び 匿名化技法としてのミクロアグリゲーションの有効性に関する研究, 2008
- 12) Kristen LeFevre David J and DeWitt Raghu Ramakrishnan, Incognito: Efficient Full-Domain K-Anonymity, SIGMOD '05 Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pp.49-60, 2005
- 13) Mohammad Rasool Sarrafi Aghdam, Noboru Sonehara, EFFICIENT LOCAL RECODING ANONYMIZATION FOR DATASETS WITHOUT ATTRIBUTE HIERARCHICAL STRUCTURE, The Second International Conference on Cyber Security, Cyber Peacefare and Digital Forensic (CyberSec2013), pp.130-140, 2013
- 14) J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. Fu, "Utilit -Based Anonymization Using Local Recoding," Proc. Int. Conf. on Knowl. discovery and data mining (KDD), pp.785.790, 2006.
- 15) El Emam K, Dankar FK, Issa R, Jonker E, Amyot D, Cogo E, Coriveau JP, Walker M, Chowdhury S, Vaillancourt R, Roffey T and Bottomley J, A globally optimal k-anonymity method for the de-identification of health data, September-October 2009
- 16) S.Nabar, K.Kenthapadi, N.Mishra and R.Motwani, A Survey of Query Auditing Techniques for Data Privacy, Privacy-Preserving Data Mining Advances in Database Systems Volume 34, 2008, pp.415-431
- 17) 小栗秀暢,曾根原登, 実サービスのデータを用いた k-匿名状態の推移調査と,合理的な匿名状態評価指標の検討, 情報処理学会研究報告.CSEC, 2014-CSEC-64(4), 1-8, 2014-02-2
- 18) 小栗 秀暢,曾根原 登,松井 くに, Mohammad Rasool Sarrafi Aghdam,近似式を用いた k-匿名性の予測手法と匿名化処理の効率化提案,情報処理学会研究報告. SPT, セキュリティ心理学とトラスト 2015-SPT-13(3), 1-8, 2015-05-07