Vol.2015-GN-95 No.3
Vol.2015-SPT-13 No.3
2015/5/14

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

小栗 秀暢[†1]　曽根原 登[†2]　松井 くにお[†3]　Mohammad Rasool Sarrafi Aghdam[†4]

Big Data 分析におけるパーソナルデータの処理は安全性を高めるために膨大な計算量を必要とし，重要な課題である．k-匿名化技術は個人情報の属性値の出現数が k 個(k>1)以上になるよう書き換え，個人識別性を減少させる手段として利用される．データ提供者間で互いに情報を公開できない場合，保持する情報を k-匿名化処理し，かつ互いに接続可能な属性値に加工することは難しい．そこで本稿では国勢調査及び 1635 サービスの登録者 434 万人のデータに対して一律の属性値の区分処理を行い，k-匿名レベルの推移調査を行った．その結果，累乗近似を用いた予測値と k-匿名レベルの推移に強い相関があることを明らかにした．これにより属性の区分数による匿名レベルの予測が可能となるため，個人情報を開示せずに匿名化処理が可能かの予測が行えるようになる．本稿では，実データを用いて予測値の正確性と計算回数の減少率を評価し，k-匿名化処理の効率化に寄与できることを確認した．

The prediction of limit number of classification that satisfies k-anonymity, And suggestion of efficient k-anonymizing process

Hidenobu Oguri[†1]　Noboru Sonehara[†2]　Kunio Matsui[†3]　Mohammad Rasool Sarrafi Aghdam[†4]

Privacy is important issue in Big data analysis. It requires an enormous amount of calculation in order to decrease the risk level of the data. A model that is widely used to protect privacy is k-anonymity, which can be generally defined as a clustering method in which any record in a dataset is indistinguishable from at least k (k>1) other records in the same dataset. If the two data providers need to exchange their data, and not be able to distribute detail information each other, It is difficult to recode these attributes that satisfy k-anonymization and be able to connect each other. We researched the transition investigation of k-value of 4.3 million users on 1635 services that has set same classification processes. As a result, we verified that the transition of the k-value has strong correlation with predicted value obtained by using the power approximation. By its predicted model, we can predict the number of the division of the attribute that satisfies the k-anonymity, and we can connect plural anonymized data by same taxonomy without disclosing personal information. In this paper, we evaluated the reduction rate of the calculation times and accuracy of the predicted value on the real data.

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ようになってきた．

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する範囲が狭く，有意な結果を出すことが難しい．そのため，他の事業者の保持するデータなどと結合し，分析する仕組みとして”Linked Open Data”[1]のような公開可能なデータを相互に結合して分析する仕組みが提案されている．

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[2]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[3]等への活用が期待できる．

だが，匿名化データを流通させる際の問題点は多く指摘されており，その多くは計算リソースの問題と情報の質の問題に分類できる．

計算リソースの問題とは，匿名状態の生成と検定処理に関する問題である．匿名化処理は属性値同士の組み合わせで行われるため，情報に含まれる選択肢数が増加すると組み合わせ爆発が発生し，計算リソースが大量に必要となる．

情報の質の問題とは，特に「次元の呪い」[4]と呼ばれる，属性の組み合わせが多くなると，組み合わせた属性値の出現数が劇的に減少し，結果的に情報量の少ない情報が生成される問題である．そのため [5][6]のように Utility を維持した匿名化処理が提案されているが，これは匿名化処理の根本に関わる問題であり，完全な解決方法は存在しない．そのため，匿名情報は情報の概要分析に用いて傾向を調査し，詳細な情報を利用する際には別途情報の利用契約を定める方式が合理的である．

†1 総合研究大学院大学 複合化学研究科 情報学専攻/ニフティ株式会社
The Graduate University for Advanced Studies, School of Multidisciplinary, Informatics Department, Tokyo, Japan. NIFTY Corporation
†2 国立情報学研究所
National Institute of Informatics
†3 ニフティ株式会社 NIFTY Corporation
†4 総合研究大学院大学 複合化学研究科 情報学専攻
The Graduate University for Advanced Studies, School of Multidisciplinary, Informatics Department, Tokyo, Japan.

Vol.2015-GN-95 No.3
Vol.2015-SPT-13 No.3
2015/5/14

匿名情報の利用方法や分析対象が明確に定まっている場合や，他の統計情報などとの情報の接続が必要な場合，情報の詳細性だけでなく，分析対象情報の保持と情報結合性を重視した処理が必要となる．

だが匿名情報の結合性は，元情報を類推させる手段と成りえるため，多くの情報は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行った後にデータを提供する．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，分析目的が失われてしまう場合がある．

例えば 図 1 のようにデータ提供者(DP)が匿名化データ P'を公開していたとしても，元情報の区分方法が異なることによってデータ利用者(DU)の分析対象 Q との接続が不可能となる．DU の数が増加すると，複数の DU と DP の間における共通の区分が必要になるが，元情報を全体に公開せずに匿名化を行うことは非常に困難である．

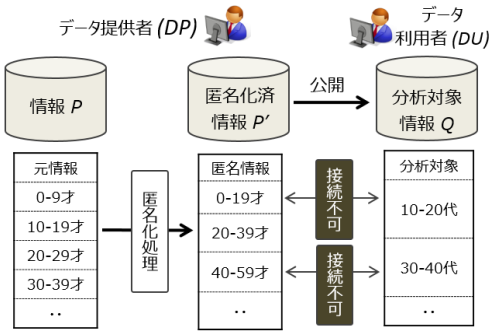


図 1　接続できない匿名化データの例

加えて，情報の変更によるプライバシー漏洩の問題も存在する．例えば複数の DP が保持する情報を匿名化して交換や共有する際に，「この属性値を匿名化できない」という現象自体が，情報の背景を類推させる材料となりうる．

もし，ある 2 者の持つ秘匿情報について，公開されている匿名化済情報と一般化のための階層から，両者ともに匿名化でき，かつ接続性を保持する最適な属性区分を推定することができれば，有益な形に成型された匿名化済情報を低コストで授受することが可能となり，より安全で有用性の高い情報同士の接続と分析が促進されるだろう．

そこで本稿では，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について 1635 個の実サービスにおける利用者数分布を用いて分析し，累乗近似型の予測式を提案する．これにより少ない検証回数から得られた k 匿名レベルの実績値と区分数によって 0.99 以上の相関係数での k 匿名レベルの予測が可能となった．

本論文の構成は次の通りである．第 2 章にて k-匿名化における情報加工の方式について述べる．第 3 章及び 4 章にて k-匿名レベルの減少度についての数学的特徴を分析し，累乗近似による予測式を提案. 第 5 章にてその評価を行い，第 6 章にて検証結果と課題点をまとめる．

2. 情報の匿名化に関する過去研究

まず，匿名化とは，個人情報やプライバシー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[2].

k-匿名化を実現するための手法として，Datafly 方式[2]や μ-Argus 方式 [7][8][9]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．

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．一般的に，アクセスログや購買ログなどのトランザクション型情報は，各属性値の出現数が頻繁に変更するため，再計算のコストとの兼ね合いで Local Recoding を利用し，住民台帳やサービス登録情報など，ある程度固定化されているマスター型の情報は，各属性の出現数を統計化したうえで Global Recoding を利用する．本稿ではマスター型に成型したデータを扱うため 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．

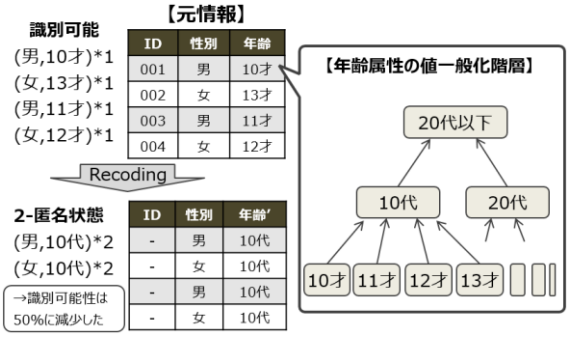


図 2　2-匿名化処理の書き換え例

例えば，図 2 において元情報の(男,10 才)という準識別子を組み合わせた属性値の出現数は 1 である．そのため ID を消去したとしても，属性の組み合わせによる再識別が可能であり，k-匿名状態の条件(k>1)を満たしていない．

そこで，匿名化処理として図 2 のような値一般化階層 *VGH(Value Generalization Hierarchies)*や，属性一般化階層 *DGH(Domain Generalization Hierarchies)* [10]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に書き換えることで，2-匿名レベルを達成することができる．また，数値情報などを含む属性値を簡素化させるために分割表(マージナル)と呼ばれる対照表を用いて値の書き換えを行う方式も提案されている[11].

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．

出現数の検証処理は作成される属性同士の掛け合わせの数だけ必要となる．一般化階層型の k-匿名化を実施した場合の理論的な計算回数は，2 以上の属性値を持つ属性同士の区分数を乗じた数値の合計になる．そのため，属性数の多い個人情報などでは容易に NP 困難な状況が発生する．

一般的に，企業が持つ個人情報において，属性数が数十以上に区分されていることは珍しくない．例えば，4 章の実験で用いるポータルサービスにおける会員データの属性種類は全 47 種類あり，その組み合わせ計算量を試算すると，最大で 1.51e+38 個の組み合わせとなる．

そこで，このような匿名化処理に伴う組み合わせ爆発状態を回避するためのアルゴリズムが多く提案されている．

Incognito[12]は，属性の抽象化候補を探索する中で，匿名化処理ができない属性が判明した場合，その属性を含む組み合わせをその後の計算から排除し，不必要な計算量を減少させている．[図 3]

この手法は情報同士の組み合わせ回数が限界になる点を予測することができないため，計算量の事前見積もりが難しいという問題がある．

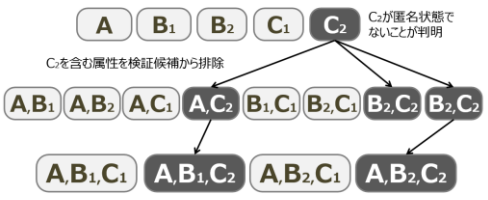


図 3　Incognito 方式による検証量削減方式

また OLA[13] (Optimal Lattice Anonymization)では，情報量に着目して最適な匿名化可能な属性の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”Globally Optimal Dataset”として利用する．[図 4]

これらの方式は，ボトムアップやトップダウンなどの順に匿名化処理可能な候補の探索を行い，条件を満たした場合にその後の処理を省略する．そのため処理削減効果は属性値の出現数の特徴に依存し，作業量を事前に予測することが難しい．k-匿名化処理が可能な限界点を予測し，適切な位置から処理を開始することで処理効率は更に高まる．

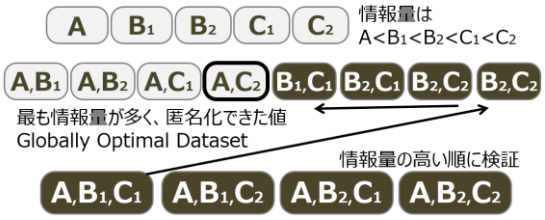


図 4　OLA 方式における最適値の検証順番

また，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，同一の基準で比較・分析する目的には利用できない．

そこで我々は複数の個人情報に対して，同一の値一般化階層を用いて匿名化処理を行い，k-匿名レベルの減少傾向を計測し，匿名レベルの減少予測モデルを提案する．

3. k-匿名レベル予測近似式の検討

我々は以前の研究[14]にて 1635 個のサービスに対し，同一の値一般化階層を用いて属性の書き換え処理を行い，サービスの規模ごとの k-匿名レベル(属性値区分における最小値)の推移を検証し，図 5 の結果を得た．

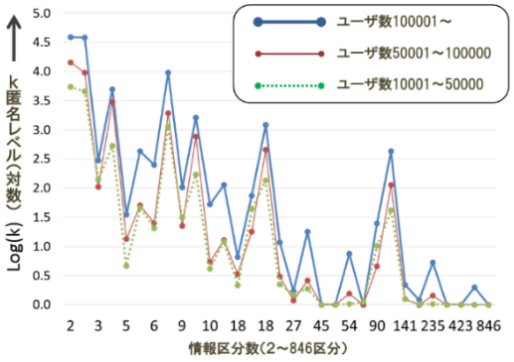


図 5　属性の区分数と k 値の推移(対数) [14]

属性の区分数が増加するたびに k-匿名レベルは減少するが，その減少値は元情報の規模に比して推移するのではなく，10 万人超のサービス群と 1 万人規模のサービス群を比較しても k-匿名レベルの減少率，減少傾向は大きく変化しなかった．本稿では，この結果を用いて k-匿名レベルの減少量を予測するモデルを検討する．今回，我々が着目したのは，k-匿名レベルの減少量が一定ではなく，特定の属性を掛け合わせた場合に，繰り返して高い，又は低い値が出るという特徴である，

基本的に k-匿名化処理は，属性をそのまま扱うのではなく，データの精度を維持するために，値一般化階層を何種

Vol.2015-GN-95 No.3
Vol.2015-SPT-13 No.3
2015/5/14

類か作成し，その範囲の中で属性同士を組み合わせて存在数を計算する．そのため，ある値一般化階層を用いて抽象度を上げた結果は，以前の数学的な特徴を引き継ぐのではないかという仮説を立てた．

図 6　一般化階層の例書き換えパターンの例

例えば図 6 における属性 A と属性 A'は同じ情報から作成されるため，その次の属性区分として A"を作成した場合，k-匿名レベルの減少傾向を引き継ぐ可能性がある．

そこで我々は k-匿名レベルは，ユーザが識別状態になる値(k=1)を最小値として減少していく，負の係数を持つ n 次方程式と近似する可能性を考えた．そこで実データに対して複数の近似式を当てはめ，最も相関係数の高いものについて精度の検証を行った．また，その近似式によって，k-匿名化可能な属性分類数を予測し，評価を行った．

4. 実験：k-匿名性予測近似式の比較と検証

実験は，ある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 7]. 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．

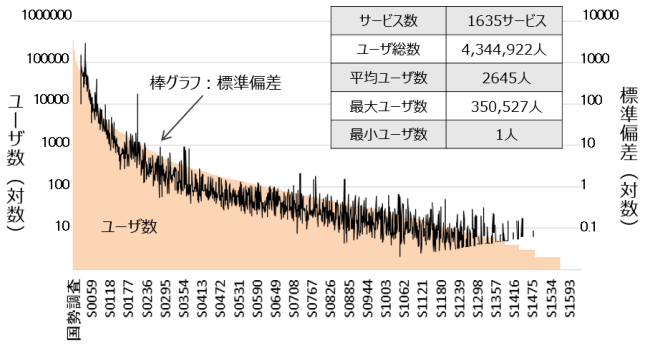


図 7 対象となったサービス群の顧客数と標準偏差

図 7 のサービス群は，分布が偏らずに存在するように，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-匿名レベルが低くなる可能性が高いものも多く存在している．

我々の過去研究[14]では，同データを用いた実験によって，標準偏差と k-匿名レベルの相関関係は 0.74 と，大きな影響を持つことが判明しており，標準偏差が異なる群を交えて検証することによってより予測式に一般性が出ると考えられる．

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．このパターンにおいても，正規分布による区分等を行わず，より一般的に利用される情報区分にするため，年代区分は定性的な意味で抽象化し，地域に関しては人口区分等を用いずに地理的な関係性でのみ区分を行った．

表 1　顧客群に対する値一般化階層

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								

各サービスのユーザ情報を表 1 の一般化階層によって書き換えていき，各区分数における k-匿名レベルの推移を求めた. その匿名化結果に対して, 一般的な表計算ソフト MS-Excel を用いて直線近似，多項式近似，指数近似，累乗近似を作成し，評価した．表 2 図 8 は国勢調査のデータを用いた試行結果である．

表 2 国勢調査の k 値から求めた近似式パターン

種類	近似式	相関係数
直線近似	y = -35.28x + 15457	0.258
多項式近似	y = -0.0008x^3+1.1143x ^2-375.57x+26233	0.435
指数近似	y = 2350.4e ^ -0.009x	0.442
累乗近似	y = 114659x^-1.414	0.997

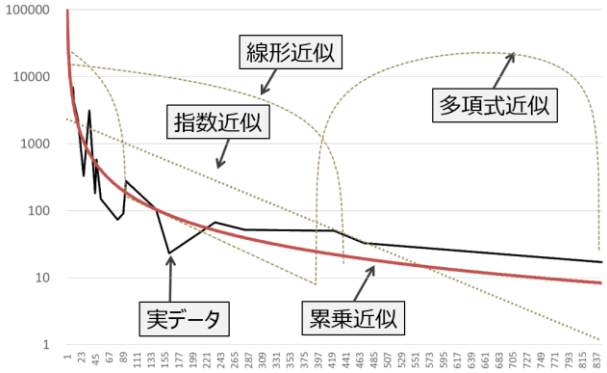


図 8 国勢調査の k 値から求めた近似式の比較

表 2 によると，以前我々が[14]にて検討した直線近似式が最も実データとの乖離が大きく(相関:0.258)，累乗近似式が最も相関が高い (相関:0.997)ことが判明した ．

k-匿名レベルは，各属性値の種類数を互いに乗じるたびに減少し，かつ最低値が 1 で確定していることから，負の乗数を持つ累乗近似と数学的性質が最も近い．[数式 1]

Vol.2015-GN-95 No.3
Vol.2015-SPT-13 No.3
2015/5/14

$$\alpha = \bar{y} - \beta\bar{x}$$
$$\beta = \ln\{\sum(x-\bar{x})(y-\bar{y})/\sum(x-\bar{x})^2\}$$
$$y = \alpha x^{\beta} \quad (+1)$$

数式 1　既知の x , y から累乗近似を求める数式

数式 1 は，既知の x：属性区分数(ex. [男,女]=2 区分)と y：k-匿名レベルの推移の平均値から累乗近似式を導きだすための計算方法となる，数式 1 では通常の累乗近似式の最後の部分に 1 を加えている．これは負の β 値を持つ累乗近似値は最終的に 0 に向けて収束するのに対し，k-匿名性は k=1 に収束していくためである．これによって精度が高くなり誤差計算も容易となる．

また，予測値を利用する中で重要な要素としては，ある区分を実行した結果，あらかじめ定められた k 値(k>1 等)が達成できるかどうかを予測することである．累乗近似式から目標 k-匿名レベルを達成できる限界区分予測値を算出する式は数式 2 の通りである．

$$k = \alpha x^{\beta}+1$$
$$(k-1) = \alpha x^{\beta}$$
$$x^{\beta} = (k-1)/\alpha$$
$$\beta\log(x) = \log((k-1)/\alpha)$$
$$\log(x) = 1/\beta\log(k-1) - 1/\beta\log(\alpha)$$
$$x = (k-1)^{1/\beta}/\alpha^{1/\beta}$$

k：目標とする k 値　α & β :累乗近似から取得
x：k 値を満たす選択肢の区分数

数式 2　累乗近似式から目標 k 値を取得する数式

これによって，匿名化するべき情報における x(選択肢の区分数)と y(その区分数における k 匿名レベル)の平均値を算出できれば，相関係数の高い近似式を作成することが可能となる．

だが，このような k 値を多量に算出する計算は非常に負荷が高い．一般的に k-匿名化処理は組み合わせ数が多くなるにつれて計算回数が増加していくため，なるべく組み合わせ数の少ない時点で近似式を作成した方が効率が良い．そこで，累乗近似式を作成するのに都合の良い情報区分について検証した．

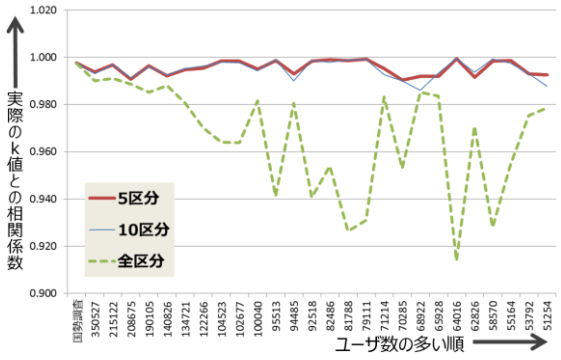


図 9　作成される近似式の相関係数と区分数の関係

図 9 は，5 万人以上のサービスにおいて，5 区分まで/10 区分まで/全区分を利用 の 3 パターンによって作成された累乗近似式が，元の k-匿名レベルの推移に対してどのような相関係数となるかの実験結果である．

この結果によると，5 区分と 10 区分で作成される累乗近似式に殆ど違いは存在しないが，全区分を用いて作成されたものは相関係数のばらつきが大きいことが解る．つまり，累乗近似式は元となるデータが大きい場合に正確になるのではなく，ある程度少ないサンプルによって作成されたものの方が正確な予測が可能となる．

また，少ないユーザ数でも同様の現象が発生するかを確認するため，ユーザ数に合わせて階級を作成し，小規模のサービス群でも同様の事象が発生するかを検証した．

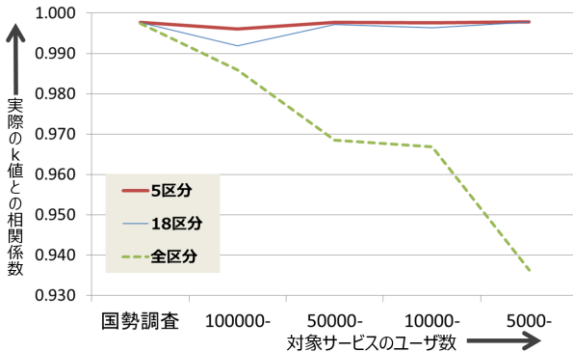


図 10　作成される近似式の相関係数と区分数の関係

図 10 は国勢調査/10 万人超/5-10 万人/1-5 万人/5 千-1 万人のサービス群の平均 k 値を取得して累乗近似式を作成し，相関係数を求めたものである．これによると，5 区分で作成された近似式が最も相関係数が高いという状況が，ユーザ数の規模に関係なく発生していることがわかる．

図 10 を実データの推移と予測式の対比によって検証したのが図 11 である．多くの群における k-匿名レベルは区分が細かくなるにつれて k=1 に近づいていくが，区分数に比して k 匿名レベルが多い地点がいくつか発生する．多い区分で作成された近似式は，そのような外れ値を含めた近似式になるため，結果的に相関係数が下がる要因となると考えられる．

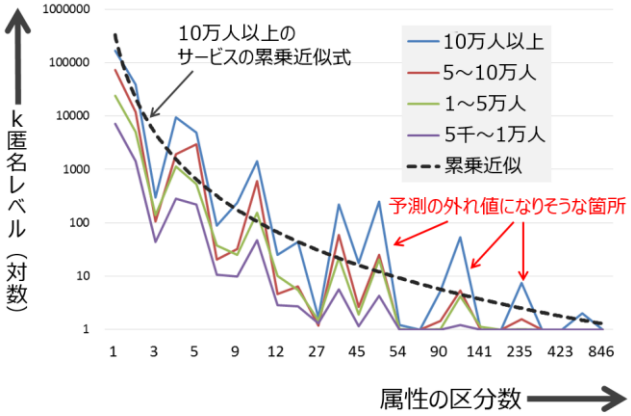


図 11　各サービス群の k 値の推移と累乗近似値の比較

この結果を受け，累乗近似式を用いて k-匿名レベルの

Vol.2015-GN-95 No.3
Vol.2015-SPT-13 No.3
2015/5/14

予測を行うために最も適した区分数はどのレベルであるかを検証した．図 11 は全ての区分数で累乗近似式を作成し，それぞれ実際の k 匿名レベルと予測値との相関係数を求めたものである．小規模サービスから 10 万人規模のサービス，及び国勢調査のほぼ全て群において，作成された累乗近似による予測値は 0.90 以上の相関係数となった．

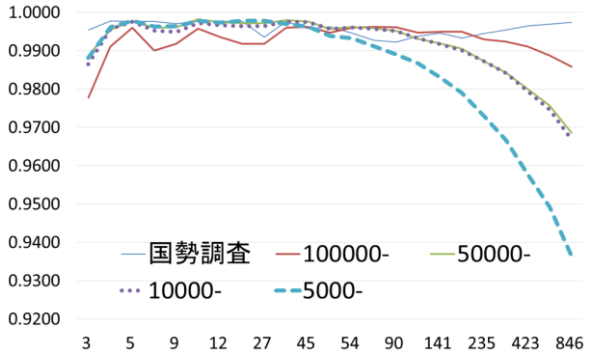


図 12 既知の x として与えられた区分数と，作成された累乗近似式の相関係数の推移

図 12 の結果より，累乗近似値を作るための区分数のサンプル(既知の x の値)が多くなるにつれて，その相関係数が減少するという傾向が判明した．例えば 846 区分全てを用いて作成された近似式の相関係数は，当初の区分数を用いて作成された近似式の相関係数と比べ，悪化している場合が多い．各区分における最高値と最低値とまとめたのが表 3 である．

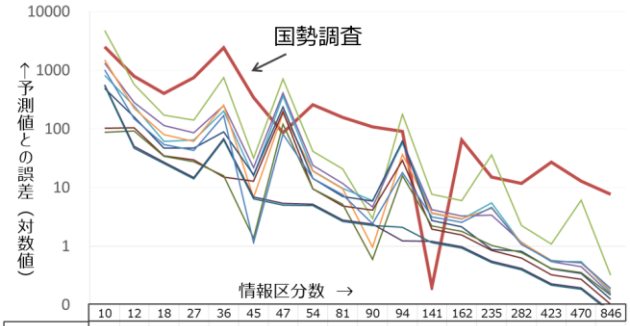
表 3　人数区分ごと相関係数の最高値と最低値

サービス群	最高値		最低値	
	区分数	相関係数	区分数	相関係数
国勢調査	4区分	0.998	90区分	0.992
10万人以上	45区分	0.996	3区分	0.978
5万-10万人	10区分	0.998	846区分	0.969
1万-5万人	36区分	0.998	846区分	0.967
5千-1万人	5区分	0.998	846区分	0.936

つまり，多くの区分による匿名化処理を行わなくとも．抽象度の高い 4～45 区分程度を抜き出して匿名化処理を実施し，そこから得られた減少数を累乗近似式にあてはめて予測値を作成することで，相関係数の高い近似式を作ることができると考える．

図 13 は国勢調査及び上位 10 サービスにおける予測値と実際の k 匿名レベルの誤差を絶対値化し，グラフ化したものである．また，国勢調査とあるサービスにおける予測値の比較結果を表 4 にて示す．いずれの値に関しても情報区分数が増加するに従い，誤差が少なくなる傾向がある．

このような近似式によって，ある程度の k-匿名レベルの予測が可能になることで，複雑な属性同士の掛け合わせ計算を全て行う前に，その値一般化階層における抽象化レベルの高い層を用いた試行によって k-匿名レベルを計測し，その数値から得られる近似式によって，k-匿名化可能で最も詳細な値一般化階層を予測することができる．



	10	12	18	27	36	45	47	54	81	90	94	141	162	235	282	423	470	846
国勢調査	2494	785	404	754	2419	345	87	258	157	108	90	0	64	15	12	27	13	8
A(350527)	4761	575	170	141	748	32	712	41	21	3	179	8	6	36	2	1	6	0
B(215122)	1331	281	115	87	251	22	408	24	12	5	61	4	3	3	1	1	0	0
C(208675)	820	245	61	65	202	16	361	19	10	6	57	4	3	5	1	1	1	0
D(190105)	1486	224	80	61	254	7	183	19	10	1	37	4	3	4	1	1	1	0
E(140826)	491	161	47	47	89	16	237	14	7	6	63	3	2	1	1	0	0	0
F(134721)	102	104	35	30	15	13	197	9	5	4	29	2	2	1	1	0	0	0
G(122266)	88	92	34	27	16	1	121	10	5	1	16	2	2	1	1	0	0	0
I(104523)	562	50	27	15	68	7	5	5	3	2	1	1	1	1	0	0	0	0
J(102677)	547	47	26	14	65	6	5	5	3	2	2	1	1	1	0	0	0	0
K(100040)	1011	148	54	43	170	1	81	14	8	2	18	3	3	5	1	1	1	0

A～K＝実験施行したサービス ，()内は人数

図 13 国勢調査及び上位 10 サービスの誤差推移

表 4 国勢調査及びサンプルの元データと予測値

	国勢調査		S0001(最大サービス)	
ユーザ数	127,794 (1000分の1)		350,527	
標準偏差	25186.7		50812.7	
α 値	114659.4		50316.4	
β 値	−1.414		−1.776	
相関係数	0.997		0.990	
区分数	元データ	予測値	元データ	予測値
2	54,556	43,028	95,033	14,697
3	22,782	24,253	595	7,155
4	22,392	16,148	25,504	4,294
5	15,960	11,778	19,226	2,889
6	9,858	9,102	156	2,091
9	4,977	5,131	669	1,018
10	6,915	4,421	5,606	845
12	4,201	3,416	36	611
18	2,330	1,926	128	298
27	332	1,086	5	146
36	3,142	723	836	88
45	183	528	91	59
47	583	496	767	55
54	150	408	2	43
81	73	230	1	22
90	91	199	21	18
94	277	187	196	17
141	106	106	1	9
162	23	87	1	7
235	67	52	40	4
282	52	40	1	3
423	50	23	1	2
470	33	20	8	2
846	17	9	1	1

5. 予測値の評価

本方式による k-匿名レベルの予測式には大きく分類して 2 種類の利用方法が存在する．1 つは k 値がある程度正確に判明することから，既存の匿名化手法に対して予測値を提供して匿名レベルの検証回数を減少させる目的で利用すること．これは予測値が正確である程，効果的である．

もう一つは個人情報の授受を行う場合に，互いの個人情報を開示せずに情報を授受する際に，両者の最大公約数となる情報区分数を予測する等の利用方法である．この場合，正確すぎる予測値は不特定多数に情報の識別可能性が高い群を類推させるリスクが増すため，予測値はあくまでも予測であり，失敗がある程度許容される．

そこで本方式による予測値に関する正確性を調査した．

Vol.2015-GN-95 No.3
Vol.2015-SPT-13 No.3
2015/5/14

図 14 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，点線は残処理数である．折れ線グラフはその予測によって計算回数が削減された率である．

ユーザ数上位6	A	B	C	D	E	F	平均
予測失敗数	2	1	4	4	2	2	2.5
残処理総数	5	7	9	9	10	13	8.8
失敗率	40%	14%	44%	44%	20%	20%	30%
処理減少率	73%	82%	88%	88%	91%	96%	86%

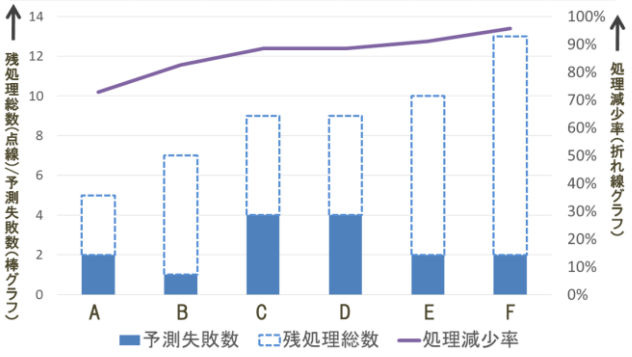


図 14　処理削減効果と匿名化予測失敗数

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．だが計算回数は大きく削減することができ，最高で 96%の処理数の削減を達成している．

この評価における「予測失敗数」とは，限界処理区分数を予測した値より，詳細な匿名化可能な値が存在した数を指している．予測値よりも匿名化可能な群が存在したため，厳密な処理として利用する価値は薄い．

そこで前研究[14]で得た知見によって，k-匿名レベルを高く維持する要因として国勢調査との相関の高い群を用いるという手法を活用して再度評価を行った．

「地域」情報は国勢調査との相関が高く，全体的に k-匿名レベルが高く維持される点に着目し，地域と性別のみを抽出して予測値を作成した．図 15 はその結果である．

ユーザ数上位6	A	B	C	D	E	F	平均
予測失敗数	0	0	2	1	1	1	0.8
残処理総数	1	3	5	1	6	9	4.1
失敗率	0%	0%	40%	100%	17%	11%	28%
処理減少率	27%	56%	73%	27%	78%	88%	58%

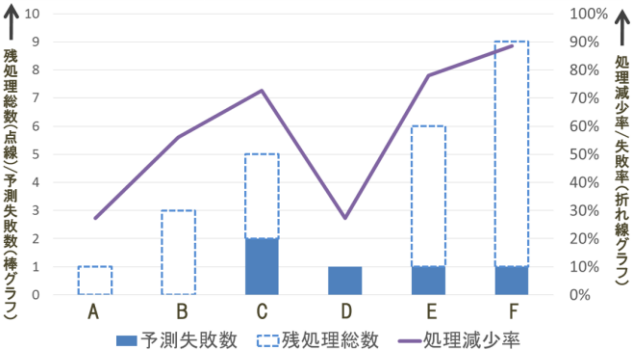


図 15　処理削減効果と予測失敗数(精度向上)

この場合，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た．だが，計算回数は多くなるために通常の予測値で作成した場合よりも計算回数の減少効果は少なくなることが判明した．

6. まとめと課題点

実データを用いた実験で判明した知見は以下の通り．

(1) 近似式を複数生成し試行した結果，累乗近似式は多くの実データと 0.99 以上の強い相関係数を持つ予測式となることが判明した．また，その近似式は抽象度の高い 4～45 区分程度での k 値によって生成された予測式の方が精度は高かった．

(2) 累乗近似式によって作成された予測値を用いて計算回数を取得する場合，予測の失敗は一定の確率で発生するため，匿名処理結果の厳密性が求められる状況での利用は難しい．予測失敗発生率は平均で 30%．その場合の処理削減効果は 86%以上を達成できた．

(3) **予測の精度を高めるため，より k 値の高い群を用いた予測値を用いた場合，匿名化予測の失敗率は減少した(30%→28%)が，処理減少効果は薄いものとなった(86%→58%)．処理削減効果と予測の正確性は両立しないため，用途に応じて使い分けるべき．**

本方式を用いた予測値はある程度の失敗を許容できる場合に用いる方が効果的である．例えば，大量の情報を同時に匿名化を行う際の共通の区分辞書を統一する際や，プラットフォームを介在して他の情報と接続できる情報区分を調査する場合などに用いることが想定される．

図 16 は本方式を利用するシステムの案である．情報を匿名化して公開しているデータ提供者 DP と，その保持情報と接続を行うデータ利用者 DU が互いの持つ情報を公開せずに最適な区分数を出力するために用いることができる．

属性 A は属性 D と接続可能だが，匿名化処理された属性 A'は属性 D と接続ができない．そこで属性 D と接続可能な書き換え候補を一般化階層 G から選択し，その区分による匿名化処理が可能かを予測する．

予測が正確であれば属性 A を，属性 D と接続可能な A'' に書き換えることが可能であるため無駄な匿名化処理計算と，属性の不一致による再試行を削減でき，また，匿名化できない情報の提供を最低限にできることから情報が類推されるリスクも減少できる．

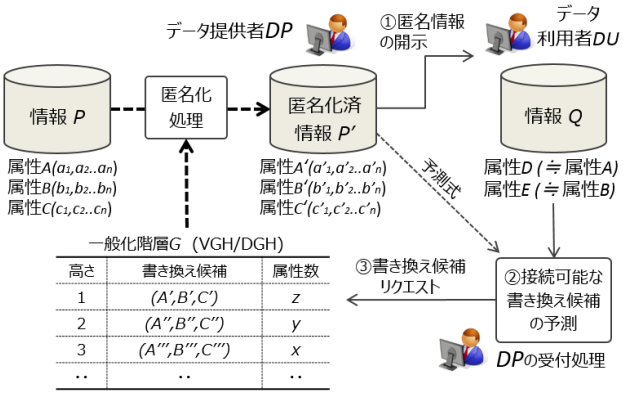


図 16　予測値を用いた匿名化システム案

表 5　図 16 における DP,DU の持つ情報種類

	データ提供者DP	データ利用者DU
公開情報	・匿名情報P' (A',B',C') ・P'を作成した一般化階層 G ・P'とDによるk-匿名レベルの予測値	
保持情報	元情報P (A,B,C)	情報Q (D,E)
目的	・Q(D,E)と接続可能な粒度で 匿名化処理されたP''(A'',B'',C'')の生成	

図 16 における処理は

① 匿名情報 *A'*の開示

② 接続可能な書き換え候補の予測

③ 書き換え候補 *A''*のリクエスト

を繰り返し行い接続可能な情報を生成する．だが，予測値が不正確な場合，この試行が数回繰り返されてしまう．元情報 $A(a_1...a_n)$と公開匿名化情報 $A'(a'_1... a'_n)$，生成される情報 $A''(a''_n... a''_n)$は，抽象化粒度が異なり[$a_n \in a''_n \in a'_n$]となる．そのため a''_nが複数回生成されることで a_nの情報が類推される可能性が高まる．

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[15] (Query Auditing)として問題提起されており，k-匿名レベルの予測手法によってリスクを軽減できる可能性がある．

現状の課題点は以下の通りである．

(1) 1 社における情報区分であるため，定式化に向けた，多くのサンプルと一般化階層による評価が必要．

(2) 現実的な情報流通を想定し，複数情報の組み合わせにより元情報が推定されるリスクの定式化が必要．

(3) 予測精度を高く保つ属性の区分方法を見つけ出すための手法の検討が必要．

今後，ビッグデータの流通が促進されると，匿名化情報の管理と分析が課題となってくる．その際に予測値や情報区分などを統一しておき，同一基準で検索や調査が可能な仕組みが求められてくるだろう．

本予測式は，そのような同一基準を作成するために必要な情報を事前に収集する際に利用可能である．本方式を発展させ，元情報を推定される・漏洩するリスクを抑えた形で,情報を簡便に結合･分析する手段が提供されることは，より多くの研究者・分析者にとって有益となるだろう．

参考文献

1) 大向一輝, オープンデータ活用:1. オープンデータと Linked Open Data, 情報処理 54(12), 1204-1210, 2013-11-15

2) L.Sweeney, k-anonymity: a model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, pp.557-570, 2002

3) 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),ケミカルエンジニヤリング 58(3), 188-192, 2013-03

4) 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

5) 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

6) 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.

7) 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

8) 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

9) Anco J. Hundepool, Leon C. R. J. Willenborg, Statistics,m-and t-ARGUS: Software for Statistical Disclosure Control,Record Linkage Techniques,1997

10) 村本俊祐，上土井陽子, and 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.

11) 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.

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) El Emam K, Dankar FK, Issa R, Jonker E, Amyot D, Cogo E, Corriveau 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

14) 小栗 秀暢,曽根原 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

15) 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