パーソナルデータ流通における保護と利活用を均衡させる匿名化処理技術の研究

小栗 秀暢

博士(情報学)

総合研究大学院大学
複合科学研究科
情報学専攻

平成 28 年度
(2016)

本論文は総合研究大学院大学複合科学研究科情報学専攻に
博士(情報学)授与の要件として提出した博士論文である.

審査委員:

主査　曽根原　登　教授　国立情報学研究所／総合研究大学院大学
　　越前　功　　教授　国立情報学研究所／総合研究大学院大学
　　神門　典子　教授　国立情報学研究所／総合研究大学院大学
　　計　宇生　　教授　国立情報学研究所／総合研究大学院大学
　　小舘　亮之　教授　津田塾大学
　　山田　茂樹　教授　国立情報学研究所
　　(主査以外は 50 音順)

A study on anonymizing method to balance the protection and utilization for the personal data distribution

Hidenobu Oguri

DOCTOR of PHILOSOPHY

Department of Informatics
School of Multidisciplinary Sciences
SOKENDAI (The Graduate University for Advanced Studies)

2016

A dissertation submitted to the Department of Informatics,
School of Multidisciplinary Sciences
SOKENDAI (The Graduate University for Advanced Studies)
In partial fulfillment of the requirements for
The degree of Doctor of Philosophy

Advisory Committee:

Prof. Noboru Sonehara (Chair)	National Institute of Informatics / SOKENDAI (The Graduate University for Advanced Studies)
Prof. Isao Echizen	National Institute of Informatics / SOKENDAI (The Graduate University for Advanced Studies)
Prof. Noriko Kando	National Institute of Informatics / SOKENDAI (The Graduate University for Advanced Studies)
Prof. Yusheng Ji	National Institute of Informatics / SOKENDAI (The Graduate University for Advanced Studies)
Prof. Akihisa Kodate	Tsuda College
Prof. Shigeki Yamada	National Institute of Informatics

論文要旨

個人の属性や行動に関連する情報である「パーソナルデータ」を利用したサービスが増加している. パーソナルデータの流通と利活用の促進によって, 既存の産業へ新たな付加価値を与える効果や, 新サービス産業の創出などが期待されている. その一方で, コンピューティング能力の進歩により, パーソナルデータから個人が特定, または識別されるリスクが高まっている. そこで, パーソナルデータの安全性を高める匿名化処理技術が必要とされているが, 匿名化処理は, 処理コストが大きく, プライバシー保護とデータ利活用の両立が難しいという問題がある.

本論文は, パーソナルデータに対して, プライバシーの保護と利活用を両立する匿名化処理システムを提案するものである. 特に, 安全性基準を満たす属性値の組み合わせを効率的に探索するアルゴリズムを用いて, データ保持者とデータ利用者間のデータ授受におけるシステム全体の効率化を図り, パーソナルデータの利活用を可能とするプラットフォームの実現を目的とする.

パーソナルデータに対するリスクの指標として, 属性値に抽象化, 統合化, 削除等の処理を施すことで個人が再識別されるリスクを $1/k(k>1)$以下にする「k-匿名性」が知られている. 本論文では, 属性値の組み合わせ数から k-匿名性を予測する仕組みを通じて, 匿名化処理を効率化する手法について検討し, 匿名化処理されたパーソナルデータである「匿名化データ」の流通における, 以下の 3 つの課題を解決する.

1) 匿名化データの安全性の判定方法
2) パーソナルデータの匿名化処理リソースの軽減
3) 実社会に即した匿名化データの流通方法

第 1 章では, 本研究の背景と課題について述べ, 論文の構成について概観する.

第 2 章では, 本研究に関連する従来研究について述べる.

まず, パーソナルデータを匿名化処理して第三者に提供した場合の攻撃モデルに即した, 多様な安全性指標について概説する. 特に, 個人の識別可能性を低減させる k-匿名化処理に注目し, 主要なアルゴリズム, 及び技術的課題をまとめる. また, 匿名化処理を行うシステム全体に関する課題についても整理し, 従来技術の課題を分析する.

第 3 章では, 匿名化データの安全性基準について検討する. 企業・機関において匿名化処理を行う場合, プライバシーポリシーに基づき, データを匿名化する手法, その再識別方法, 及び, 安全性基準を達成するべく検討されたプロセス等は公開されない. そのため, 匿名化処理における, 技術と安全性のバランスが取れた基準値を探る先行事例が少ないという問題がある.

そこで，安全性基準を検討するための基礎データとして，実サービスの分布を反映した擬似パーソナルデータ群に対して，一律のクラスタ化処理を行った場合に，達成できる安全性基準の推移を検証した．その結果，単純なクラスタ化を行うだけでは，匿名化は達成できず，各データの分布の特徴に応じた匿名化処理を行わないと，データの安全性と有用性が損なわれることを実証した．

第 4 章では，この結果を受け，パーソナルデータの匿名化処理リソースの削減に向け，k-匿名性の減少予測式を用いて匿名化処理を効率化する手法を検討した．属性の出現数を統計的に利用するk-匿名化処理を行う場合，全体データの書き換え処理と，その安全性計測を連続的に行うことから，情報量の損失が最小の匿名化データを生成する処理は NP 困難であり，また，その予測も難しい．

そこで，課題解決のアプローチとして，クラスタ化による k-匿名性の減少傾向を予測し，安全性の基準を満たし，かつ最も詳細な属性の組み合わせを導く手法を検討した．具体的には，抽出条件によって分析に耐えうる十分なデータ量が出力される場合，出力データを基準化すると標準正規分布に近似する性質から，累乗近似型の予測式を提案した．

予測値を実際のk-匿名性と比較し，予測誤差を検証した結果，重相関係数 0.9 以上の値で実測値と近似したが，予測誤差が大きいことから，予測値を直接用いた効率化は困難であることが判明した．

第 5 章では，4 章にて提案した予測式を用いて，匿名化処理が達成可能な属性の組み合わせを予測し，その組み合わせから匿名化状態を検証するアルゴリズムを提案した．従来技術において，匿名化処理アルゴリズムは詳細な群から匿名化処理を試行するボトムアップ方式と，抽象的な群から匿名化処理を試行するトップダウン方式に分類される．提案するアルゴリズムは，予測地点における匿名性を検証した結果が，求める匿名性を満たす場合は，その地点からトップダウンの匿名化処理を選択し，満たさない場合はボトムアップ型の匿名化処理を選択することで，匿名化状態の検証回数を削減する．本方式を 4 章で用いたデータに対して適用し，他の匿名化アルゴリズムと比較し，処理削減効果が高いことを検証した．

第 6 章では，これらの研究成果を活用する枠組みとなる，実社会に即した匿名化データの流通方法について検討した．まず，4 章で検討した予測式を用いて匿名化処理の効率化が可能になる研究結果を受け，それを実装するべきパーソナルデータ流通プラットフォームを提案した．具体的には，匿名化データの利用者に対して，属性区分数と k-匿名性に関する予測式を提供することで，利用者が求める属性区分によって匿名化処理が達成できるかを，事前に検証出来る仕組みを考案した．これにより，元となるパーソナルデータの漏洩を抑制し，かつ匿名化データの授受に伴う検証作業を効率化できることを確認した．そこで，これらのプラットフォームを社会実装する上で必要な技術的課題を整理した．

第 7 章にて，本論文の成果をまとめる．

本研究の成果は，実サービスの分布を反映した疑似パーソナルデータ群を用いて，属性値によるクラスタリングを行った場合の k-匿名性の推移を検証し，相関係数の高い予測式を提案したことにある．加えて，その予測式を用いて，匿名化処理アルゴリズムを選択する手法を提案し，元情報の分布の特徴に依存せず，匿名化処理が効率化できることを実証した．これにより，元情報に含まれるデータの特徴を公開せずに，データ利用者の求める一般化階層を事前に評価するプラットフォームの社会実装を可能とした．

Abstract

Many enterprises and organizations collect information about individuals for various services. Among them, the information created from personal behavior, defined as "personal data", is increasing. On the other hand, by the progress of the computing power, the personal data contain risks of individual re-identification. Anonymization is a technology to prevent identification.

However, the load to make anonymized data that achieves minimum information loss is high, and it is difficult to achieve the balance between privacy protection and data utilization.

This thesis proposes the anonymization process to balance the privacy protection and utilization of personal data, which involve personal attributes and activities. Particularly, using an algorithm to search for the combination of attribute to meet a security standard effectively, aims to reduce the transaction cost between the data holder and data users. Moreover, we target the realization of a platform promoting the utilization of personal data.

As a barometer of a risk of personal data, “k-anonymity” is widely used. k-anonymity is an index of reducing the risk of individual re-identification less than 1/k (k>1) by abstraction and aggregation etc. In this thesis, we precisely studied a method to make the anonymization process effective through a system predicting k-anonymity by the number of combination of attributes. It also solves several problems in the distribution of anonymized data, which is anonymize-processed personal data, as well.

1) The method of evaluating the safety of anonymizing personal data.
2) The load reduction process in anonymization processing.
3) The distribution systems of anonymized data, which corresponds to the real society.

In Chapter 1, we describe the background and purpose of this research, and show the overview of this thesis.

In Chapter 2, we researched the related approaches about anonymization. At first, we researched the privacy regulations of various countries and regions, and we presented the survey of attack model of personal data and evaluations. After that, we focused on k-anonymous processing that reduces the risk of specific individuals, and surveyed the major algorithm technique

and their problems. In addition, we defined the problems containing the whole system of anonymization.

In Chapter 3, we suggested the security standard of anonymized personal data. Usually, enterprises do not release the detail process of anonymization based on the security policy. Hence, there are a few examples of making appropriate anonymizing standard.

In this situation, we researched the basic data to consider the safety standard. At first, we made the dummy personal data created by referencing the variance of actual service data. After that, we set a uniform clustering process of the data of the attribute values, and investigated the transition of the k-value.

In Chapter 4, we supposed the load reduction process in anonymization processing. To solve this problem, we supposed an efficient system using the predictive model of k-anonymization. It is generally acknowledged that the load required to realize optimal k-anonymization of microdata is too high and the process is NP-hard. Moreover, and it is difficult for us to predict whether the attributes abstracted from a dataset will satisfy k-anonymity.

We proposed a k-anonymity predictive model which uses a power approximation based on the property that most data approximates a normal distribution when extracted at a constant scale from large-scale data. Moreover, we investigate the accuracy of the predictive value and real k-value. The result of it, the proposed predictive model took out regression coefficient scores of more than 0.9. However, the model cannot be used when correct numerical values are required.

In Chapter 5, we therefore proposed an anonymizing algorithm that starts processing from a prediction spot and uses optimal anonymization algorithm. Generally, these methods suggest that it is possible to use a bottom-up or top-down order when processing data to achieve anonymity. Then we predict the point at which it is possible to achieve k-anonymity and the combination of attributes with the greatest utility, and begin the anonymity process with choosing optimal methods. We compared proposed algorithm experimentally with a general anonymizing algorithm.

In Chapter 6, we discussed the framework of anonymized data distribution systems, which corresponds to the real society. Therefore, we discussed the distribution platform of anonymized data to apply these studies. At first, a

data provider provides the predictive approximation between the number of the attribute and the k-anonymity to the data user. Thereby, the data user inspects the results that a generalization hierarchy can achieve k-anonymity, without requesting the detail of the personal data. By using this method, we can reduce the cost of negotiations in the process of data transactions.

In Chapter 7, we conclude the thesis with the summary of the results.

This thesis proposed a predictive model that meets high regression coefficient scores using the clustering process of the data by the attribute values. In addition, we proposed an anonymizing algorithm that starts processing from a prediction spot and chooses an optimal anonymization algorithm, and this algorithm achieves a stable and high efficiency, compared with other algorithms. Accordingly, we presented the methods achieved the safety and efficiency of the entire system according to the transaction of personal data.

目次

図目次

表目次

1 序論

1.1 研究の背景

現代社会では，様々な企業・機関がサービス提供を行うため，個人情報やパーソナルデータを多く収集している．

「個人情報」とは，個人情報の保護に関する法律（個人情報法保護法）によると，「生存する個人に関する情報であって，当該情報に含まれる氏名，生年月日その他の記述等により特定の個人を識別することができるもの」とされている．

それに対して「パーソナルデータ」は，「個人情報」よりもより広い意味を持ち，その情報が誰の情報かわからない場合でも，その個人に関するデータ全般を「パーソナルデータ」として扱うものとする[1]．

パーソナルデータには，位置情報，購買情報，インターネットログ，通信履歴などが含まれ，サービスを運営する上で必要なデータとして利用される．加えて近年では，そのデータを分析，または流通させることによって多方面で価値を生むと考えられている．

1.1.1 パーソナルデータの増加と利活用

パーソナルデータは，インターネットが発展する前までは個人の識別性・危険性がないデータとされ，長い間企業や機関が自由に取り扱っており，パーソナルデータの蓄積と利用は問題視されてこなかった．

しかし，パソコン・携帯電話・スマートフォン等の普及によって，パーソナルデータの重要性が高まっており，現在では，機器間で通信を行う M2M(Machine to Machine)も，その機器に触れた個人が存在する場合パーソナルデータになりうるものとされ，また，M2M よりも広い概念である IoT(Internet of Things)では，人と機械，ソフトウェア等，あらゆるモノがインターネットに接続するため，様々な領域において，パーソナルデータの対象は広がっている．

日本国内におけるデータ流通量は，平成 27 年情報通信白書によると，2014 年において 14.5 エクサバイト以上となっており，9 年間で 9.3 倍の量[2]となっている．同白書によると，POS データ，携帯電話等，過去から利用されているパーソナルデータの量は約 4.5 倍の量になっているのに対し，センサーデータは 2005 年比で約 12 倍，交通・渋滞情報データは約 9.2 倍と大きく拡大している．

それらの機器によって大量に生成され，記録されるパーソナルデータは，現代社会においては利便性の高いサービスを運営するために必要不可欠なものであると同時に，プライバシーとして守られるべき存在と認識されるようになった．

1.1.2　パーソナルデータ分析技術とプライバシー問題

パーソナルデータは様々な分野で重要視されているが，その半面，パーソナルデータを用いたプライバシー問題も多く発生している．その契機として，2002 年に発生した「ダブルクリック社訴訟」が存在する[3]．

これは，インターネット広告システムを運用していたダブルクリック社が，広告に利用している Cookie データを，実社会で使用されている名簿事業者の持つデータと照会することで，より精緻なターゲティング広告を実現しようとし，消費者団体から差し止めを要求された訴訟である[3]．

結果として，パーソナルデータの活用に対して，安全性基準，利用規則等（以降，規則等とする）が明確に定義され，その範囲内での活動が可能になり，新しい事業創出やサービス開発に発展した．特に，大量のパーソナルデータを活用し，かつ WEB サービス等と親和性の高い機械学習の研究は 2000 年台前半から活発となり，その後のビッグデータ分析などの研究につながっている．

機械学習や深層学習は，ある情報に対する結果データを学習させることで，その結果に至ったルール等をアルゴリズム化するものであり，WEB サービス，医療診断，金融分野等において，個人向けサービスの高度化やサービスの最適化等に活用される[4]．

パーソナルデータの授受が頻繁に発生する背景には，これらの研究分野が進展し，サービスやシステムに搭載されたことが関係する．その大きな事業分野として，インターネット広告分野がある．

インターネット上での個人の行動データを収集し，即時にオークション方式で広告取引と画面表示を行う，RTB(リアルタイム入札型広告配信)の技術は，多くのインターネットサービスで採用されている．

RTB における個人データのパターン解析は，機械学習や深層学習によって大きく効率化された．過去においては，個人の属性（性別，年齢等）に応じて広告の変更を行う「ターゲティング広告」方式が主流であったが，個人の行動（閲覧した URL，検索語履歴等）を分析して，最適な広告配信を行う「行動ターゲティング広告」方式に発展したことによって，個人への広告提供の効率が大きく向上した．

このようなパーソナルデータのインターネット広告分野への活用は，様々な問題点が指摘されている．例として，ある広告代理店にパーソナルデータを提供した場合に，そこから RTB を通じて他の広告代理店にパーソナルデータが共有されるといった，転々流通や第三者提供の問題が指摘されている．

特に，多国籍企業においては，自社サービスを利用するユーザのパーソナルデータを大量に取得し，各国のプライバシー法制度に抵触しない方式で結合，分析，共有を行い，行動ターゲティング広告に活用しているため，規則等が追いついていないという問題がある．

過去において発生したプライバシー問題は，主に国内問題であったのに対し，現在ではグローバル企業が世界中でサービスを行うことから，パーソナルデータの利用や管理について，国・地域ごとの対応が求められる．

例えば，仮想サーバ環境をネットワークを通じて遠隔から利用できるクラウドホスティングサービスは，世界中にサーバ設備を分散させてサービスを提供している．そのような仮想環境では，パーソナルデータの保存についても一箇所の設備に依存しておらず，分散保持している．このようなパーソナルデータの管理体制に対して，明確に一つの国・地域の規則等を適用することは難しい．

そのため，個人情報保護ポリシーが不明確な企業や，海外サービスなど自国民のデータに対する人権保護が行われない場所にパーソナルデータが流通することの危険性が認識され，多くの国・地域では，利用に対する制限が加えられている．

1.1.3　パーソナルデータ保護技術の国際動向

これら，新しい技術の進展によって，新しい脅威が生まれるため，常に最新技術に対応した規則等を新しく設定する必要が出てくる．

欧州連合（EU）ではインターネット広告などにおけるパーソナルデータを用いた過度な個人のプロファイリングを禁止する措置が強化された．EU の定義によると，プロファイリングとは「自然人に関する一定の個人的側面を評価すること，または，特に，当該自然人の職務上の成果，経済状況，位置，健康，個人的嗜好，信頼性若しくは行動を分析または予測することを意図した，あらゆる形式の自動個人データ処理」となっている[6]．

しかし，「行動を分析，または予測」という基準は曖昧であり，かつ，各個人において許容できるレベルは異なる．そのため，個別の行動ターゲティング広告がどこまで違法なプロファイリングにあたるのか，認定することは難しい．

そこで EU では機微情報に関わる処理を禁止し，かつプロファイリングされる対象を個人が選択できるよう，WEB サービスに自己情報コントロールの確保を義務付ける等の対策が行われている．

また，インターネットサービスだけでなく，パーソナルデータの提供についても制限を行っている．

EU と米国商務省が 2007 年に締結したセーフハーバー協定では，個別の企業が EU 諸国から転送された個人情報に対して十分な保護を行うことを保証する形で運用されていた[7]．しかし，2015 年に欧州裁判所によってセーフハーバー協定の無効判決が出さ

れたことから，同年に承認された EU データ保護規則[8,9]によって，欧州における国境をまたいだパーソナルデータの利用について，より厳格な管理が必要となった．

このような動きにあわせ，日本国内でも個人情報保護法が 2015 年に改正(個人情報の保護に関する法律及び行政手続における特定の個人を識別するための番号の利用等に関する法律の一部を改正する法律)され，パーソナルデータの利活用に対する定義が明確化された[10]．

日本では 2013 年に大手鉄道会社の持つ IC カードの利用履歴データを，個人識別性を残したまま販売しようとした事件が発生し，社会問題となったことから，個人情報を取り扱う事業者に対して，データの安全管理を強く求めている．

個人情報保護法の第 17 条において「個人情報取扱事業者は，偽りその他不正の手段により個人情報を取得してはならない」と定義されることで，企業から不正に流出した個人情報が販売，共有されることを禁じている．これによってパーソナルデータの転々流通へを防止し，データブローカー（名簿業者）への規制を強めた．

しかし，日本国内だけに限定しても，民間・行政機関・独立法人ごとに個人情報・パーソナルデータに関する定義が異なるという問題が指摘されている．2011 年の東日本大震災が発生した際に，国内に存在する 2000 の行政機関・独立法人毎にデータ提供ポリシーが異なる問題から，運営主体の異なる医療機関が持つ情報を相互に開示できず，被害地域に住む患者や老人を移設することができないという，所謂「2000 個問題」[11]である．

国内だけでも問題が山積する中，同様に米国・中国・OECD 諸国も独自のデータ保護政策を採用しており，全ての地域に対して適用可能であり，パーソナルデータの安全性と利用性を同時に向上させるシステムを作ることは技術的に困難な課題の一つである．

1.1.4 プライバシー保護データパブリッシング

このような，パーソナルデータから個人のプライバシー侵害が発生する要素を排除し，かつ，データを国・地域の定める規則等に合わせて加工することで，自由に流通可能な仕組みを整える研究分野として，プライバシー保護データパブリッシング（PPDP :Privacy Preserving Data Publishing）が必要とされている[12,13]．

これは，主に個人情報やパーソナルデータからセンシティブな要素を排除する，所謂，匿名化処理技術を適用した「匿名化データ」を流通させるものである．

類似した研究分野にオープンデータや Linked Open Data がある．オープンデータは，自由な流通，再利用，再配布が可能なデータである．そのため，個人のプライバシー侵害などが発生しないデータであることが求められる．Linked Open Data は，オープンデータの活用を促進するために，アクセス方法，データ形式，語彙などを統一し，結合や検索を容易にしたデータである[14,15]．

それに対して，プライバシー保護データパブリッシングは，個人情報やパーソナルデータに含まれる情報に対して，匿名化処理を施すことで個人が特定，識別されるリスクを定量的に低減し，かつデータ利用が可能な形に加工し，広く流通させることを目的とするため，パーソナルデータにおけるオープンデータの一部と考えることができる．

本研究では，パーソナルデータに対して匿名化処理を施し，個人が特定・識別されるリスクを定量的に低減し，かつデータ利用者の要求に即した形に加工されたデータを「匿名化データ」と定義する．

1.1.5　匿名化データの利活用

匿名化処理によって，パーソナルデータの安全性を定量的に高めることが可能となり，一定の基準を満たした場合に，個人情報に比べて簡易な手続きで第三者にデータを提供できる．

匿名化データのユースケースとしては，例えば，ある機関や組織が保有するパーソナルデータを分析するとき，単体ではデータ分析サンプルが不足している場合がある．その際には，類似したパーソナルデータを保持する機関から，分析対象を含む匿名化データを入手することで，比較分析や機械学習における教師用データとして使用することが可能となる．

特に，医療データや公共データの分野では，欧米や日本において安全性を高めたデータを研究機関等に提供する施策が進められている[16]．

匿名加工処理された情報を流通させるシステムは，欧米では政府機関や医療機関において定着しており，オランダの μ-Argus 方式 [17,18]や，カナダの CHEO における匿名加工情報の提供[19,20]などが知られている．

日本においては，レセプト情報・特定健診等情報データベース（NDB）から生成された匿名化データを，研究者に対して提供するなどの仕組みが整備されている[21,22]．

また，匿名化データを用いたレコメンドエンジンの開発[23]などの研究も進んでおり，今後も提供するデータの種類と形式に応じて活用方法が考案されている．

1.2　匿名化処理技術の課題

匿名化処理は，その扱うデータの種類，利用目的，達成するべき安全性指標と，その基準等の要因に対して柔軟に対応することが求められる．

安全性基準として，Sweeny が提案した k-匿名性 [24,25]により，個人が再識別されるリスクを 1/k(k>1 の整数)以下に加工することで，安全性を定量化することが可能となった．その後，ℓ-多様性[26]，t-近傍性[27]など，利用目的や攻撃モデルに対応した評価指標や，それを実現する匿名化処理アルゴリズムが提案されており，データから個人が特

定，または識別される可能性を定量的に低減させることで，悪用されるリスクを低減させることが可能となった．しかし，あらゆるデータの種類に対応できる匿名化処理や安全性指標を定義することは困難である．

そこでデータの目的に応じて有用性を最大化し，安全性を高める匿名化処理について，多くの研究がなされている．その一方で，匿名化処理は計算コストが高く，安全性基準を満たし，かつ情報量の損失が最小となる k-匿名化の実現は NP 困難であると知られている[28]．

また，匿名化処理を実施した場合には，属性の数が増加するごとにデータの有用性が低下する「次元の呪い」[29]や，過去に提供したデータとの差分を検証されることで個人の特定リスクが増加する「逐次公開」のリスク，個人情報を匿名化処理する際に，中間データ等，危険性の高いデータが含まれる可能性として「アウトソーシング」リスクなどの課題が挙げられており[30]，それぞれの分野で研究が進められているが，それらの基準を満たす処理を行うことによって，データの使用目的である有用性が損なわれる場合がある．

そのため，匿名化データの利活用には，データ保持者の安全性要求とデータ利用者の有用性要件の両方のニーズを満たす均衡点を探す，匿名化処理に関わる活動全体の効率化が必要である．

また，データ保持者の立場から考えると，匿名化データの安全性と有用性の指標を利用するためには，現在のサーバリソースにおいて達成可能な安全性の基準と，攻撃者が使用する再識別手法を比較し，求める安全性が達成できるかを具体的に検証することが必要である．

しかし，現実的には，匿名化処理は，企業や研究機関の内部で実施され，その強度や処理手法について，外部に情報が共有されることは少ない．また，攻撃者にとっても，そのデータの利用方法に応じた攻撃方法を適用するため，攻撃モデルについても定義することが難しい．パーソナルデータの再識別攻撃などの研究は，実際には違法な個人再識別につながる可能性も高いため，積極的に行われることは少ない．

そのため，匿名化処理アルゴリズムは，UCI(カリフォルニア大学アーバイン校)が公開する Adult Data Set[31]や独立行政法人 統計センターの擬似ミクロデータ[32,33]など，擬似データを用いて処理時間等の検証がされることが多い．しかし，実際の企業が保持するデータのような多様なユーザ分散パターンを持つデータでの検証が難しい．

また，有用性を担保できた場合においても，データ保持者とデータ利用者の折衝において，安全性要件を満たすかを予測することが出来ないという問題点がある．

安全性を満たしつつ，データの有用性を保ち，その折衝の中で発生する再匿名化処理コストを減少させることによって，社会全体における匿名化データの流通が活発化する方法について検討する．

1.2.1 オーダーメード型の匿名化処理の課題

匿名化データを第三者に提供するためには，匿名化処理技術だけでなく，データを保持する事業者と，データ利用者がデータの利用目的について討議し，その利用目的に沿った形に情報を整形し，データを授受するシステムが必要である．

しかし，現状において，事業者の求めるデータを提供する機関は少なく，公共機関や医療情報などに限られている．

独立行政法人 統計センター等の機関では，学術利用に限定して，個別利用者が必要とするデータ区分をリクエストしてデータを加工する「オーダーメード集計」[33]を提供することで，再識別可能性とデータ漏えいリスクを制御しつつ，個別のデータ利用ニーズに対応している．

オーダーメード集計では，データ利用者が，パーソナルデータの書き換え要求をまとめた統計作成仕様書を作成し，提供者と利用者間で折衝を繰り返してデータを提供する方式を採用している．統計作成仕様書には，加工する統計調査名，及び集計対象となる属性項目とその区分種類，属性区分数を記載し，データの再集計を依頼する．

図 1 はこの方式を参考に，オーダーメード型匿名化のシーケンス図を検討した例である．まず，データ提供者 DP はパーソナルデータ P に対して，k-匿名性における k 値 kp (kp>1 の整数)を満たす匿名化データ P'を作成し公開したとする．

しかし，匿名化データ P'はデータ利用者 DU が求める分析目的が達成できず，DU は新しい一般化階層 Gu を作成し，再匿名化処理を依頼した．DU は P についての知識は無く，また目的外利用を禁止する等の利用条件があるものとする．

DP は P に対して Gu を適用した結果が k-匿名性を満たすかが不明であることから匿名化処理を複数回試行する．その際に個人情報を含むデータベースに対して処理量の大きい匿名化処理を要求するため，作業負荷が大きい．

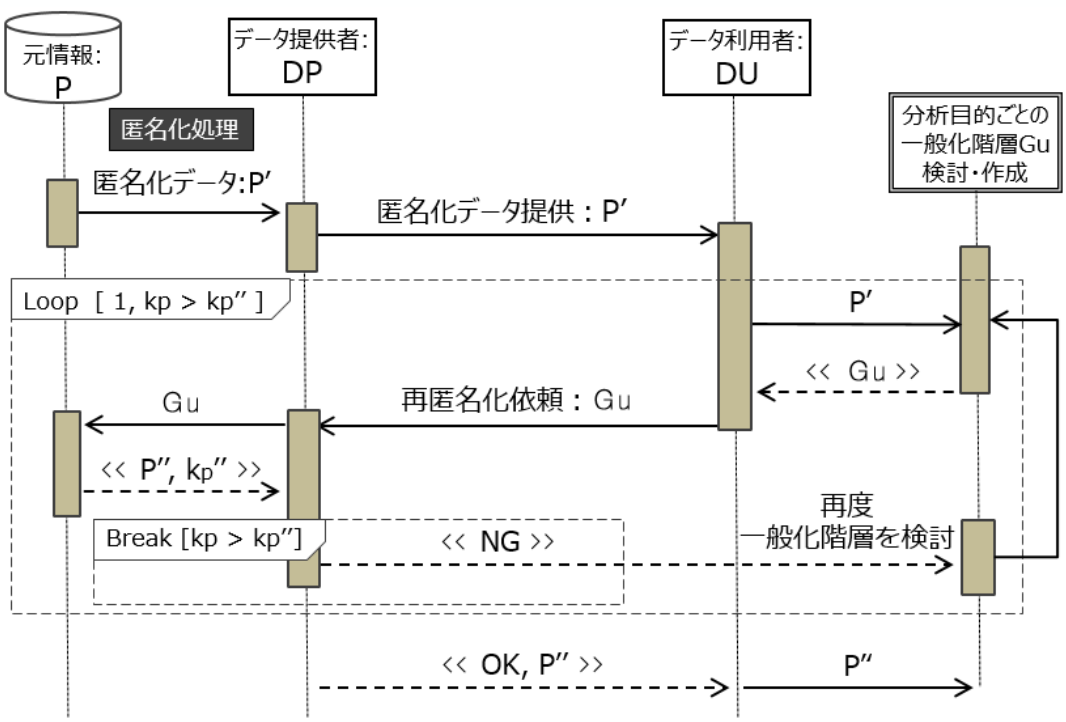


図 1 オーダーメード型匿名化処理のシーケンス図

一方 DU は P のユーザ分布を知り得ないため，新たに目的を達成するための，妥当な Gu を検討する指標が無く，双方の不一致問題が繰り返し発生する可能性がある．

また，DU が匿名化処理を要求する場合，データの利用目的を明確に設定していることが求められる．しかし，その場合，データ利用目的を達成する Gu を利用した場合の k-匿名性や，その他の安全性指標を満たすデータが出力されるかが不明である．DP は，元データにおけるユーザ属性の特徴を知られずに，DU への匿名化処理結果を伝達し，その結果についての折衝を行う必要がある．

これらの DP，DU 双方の作業負荷が匿名化データの流通を妨げている原因の一つと考える．

他の機関からの多様な要求に対応して匿名化データを作成するため，データ提供者 DP が求める k-匿名性を満たし，かつデータ利用者 DU のデータニーズに合致した匿名化データを，軽量に作成するアルゴリズムが必要とされている．

1.3 本研究の目的と貢献

本研究は，パーソナルデータに対して，プライバシーの保護と利活用を両立する匿名化処理システムを提案するものである．特に，安全性基準を満たす属性値の組み合わせを効率的に探索するアルゴリズムを用いて，データ保持者とデータ利用者間のデータ授

受におけるシステム全体の効率化を図り，パーソナルデータの利活用を可能とするプラットフォームの実現を目的とする．

その実現に向けた技術的課題として．

1） 匿名化データの安全性基準の策定

2） パーソナルデータの匿名化処理リソースの軽減

3） 実社会に即した匿名化データの流通方法

の 3 項目を抽出し，それぞれの解決に向けた検証と提案を行った．

本研究の成果は，実サービスの分布を反映した疑似パーソナルデータ群を用いて，属性値によるクラスタリングを行った場合の k-匿名性の推移を検証し，相関係数の高い予測式を提案したことにある．加えて，その予測式を用いて，匿名化処理アルゴリズムを選択する手法を提案し，元情報の分布の特徴に依存せず，匿名化処理が効率化できることを実証したことにある．これにより，元情報に含まれるデータの特徴を公開せずに，データ利用者の求める一般化階層を事前に評価するプラットフォームの社会実装を可能とした．

1.4　本論文の構成

本論文の構成は次の通りである．

第 2 章にて匿名化処理の従来研究を分析．第 3 章にて実データに即して作成した擬似パーソナルデータ群に対して一律に匿名化処理を行い，その安全性の減少傾向を検証する．第 4 章にて，匿名化処理を行った際の安全性と属性組み合わせ区分数との関係性について分析し，k 値の予測モデルを提案し，その予測モデルを擬似パーソナルデータ群にて検証．第 5 章にてその予測式を用いた処理削減アルゴリズムを提案し，既存アルゴリズムとの処理回数比較を行う．第 6 章にて，本研究を搭載するべきプラットフォームの要件について検討し，第 7 章にて研究内容を総括する．

2 従来研究の分析

2.1 本章の構成について

本章では，パーソナルデータに関する匿名化処理の従来研究について述べる．

まず 2.2 章にて用語の定義，2.3 章にて基本的な匿名化処理の考え方である k-匿名性について述べる．2.4 章ではパーソナルデータの攻撃方法から，安全性を高める指標の多様性について概説する．2.5 章にて，多様な安全性指標を適用するための，匿名化処理の手順と，データ処理単位，一般化階層，有用性指標の特徴について述べる．

2.6 章において，匿名化処理アルゴリズムについて説明し，それぞれの手法の特徴についてまとめる．その後，2.7 章からは，k-匿名性を含めた個人の識別性に関する日本，欧州，米国における安全性基準について述べ，2.8 章にてこれらの課題のまとめを行う．

2.2 用語定義

まず，本論文における用語を定義する．

パーソナルデータとは「属性(Attribute)」と「属性値(Value)」としてテーブルの形で表現される，ユーザに関する情報であり，図 2 にて示す通り，あるユーザのパーソナルデータをテーブルのレコードとして表現する．

パーソナルデータでは，「直接識別子(Explicit-Identifier)」として，氏名や識別番号等が設定され，他のレコードと区別して管理される．

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(QID :Quasi-Identifier)」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図 2 にて示すとおり，「センシティブ属性(SA :Sensitive Attribute)」と呼ぶ．

データ分析の分野では，準識別子は分析対象における「説明変数（Explanatory variable）」であり，センシティブ属性はその「目的変数（Target variable）」と考えることができる．また，属性と属性値を「特徴量」と呼称する場合もある．また，センシティブ属性には，個人のプライバシーに配慮した情報，という意味だけでなく，匿名化データ提供に伴う分析対象の属性であると考える．加えて，センシティブ属性は再識別を試みる攻撃者に秘匿されており，その値からは個人が特定されないものとする．

また，複数のレコードにおいて，同じ属性値を持つ群を「等価クラス(Equivalence Class)」，または「同値類」と呼び，各等価クラスにおけるクラスタの大きさを「等価クラスサイズ(Equivalence Class Size)」，等価クラスの存在する数を「等価クラス数

(Equivalence Class Number)」と定義する. 図 2 においては, 準識別子において同じレコードが存在するため, 等価クラスが存在するパーソナルデータである.

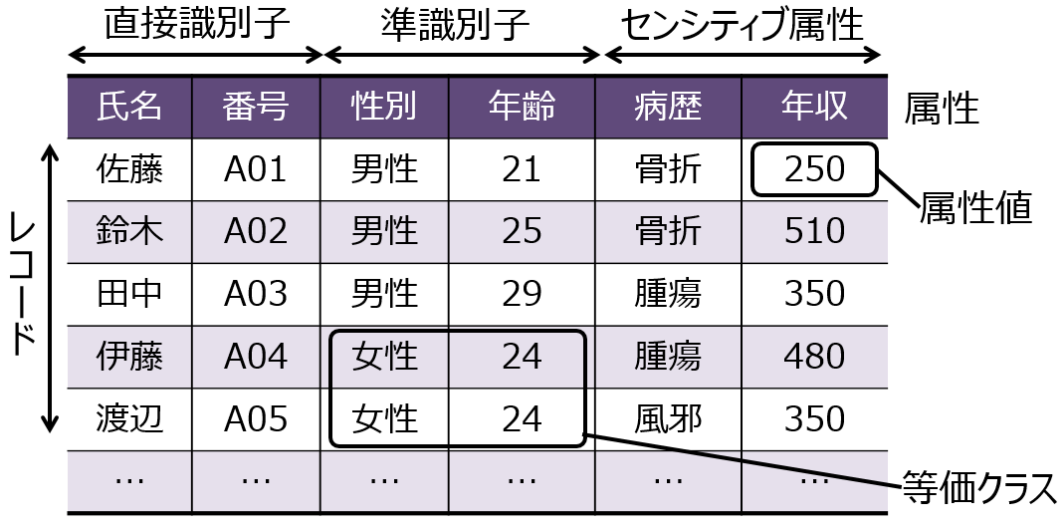


図 2 パーソナルデータの例

2.3 k-匿名性

ある個人情報やパーソナルデータを含むデータベースを公開する際, 個人が公開することを望まないセンシティブ属性の属性値が知られてしまった場合, プライバシー侵害が発生する. そこで, データベースから直接識別子を削除することで, 個人識別が出来なくなり, プライバシー侵害を防ぐことができる.

しかし, 直接識別子を削除した場合でも, 準識別子を複数組み合わせることで個人が識別され, センシティブ属性が知られてしまう場合がある. もし攻撃者があるユーザの準識別子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, センシティブ属性の値を知られてしまう.

そこで, 準識別子の出現数を計測, または, 属性値に抽象化, 削除等の処理を施すことで, 等価クラス数が少なくとも k 個以上(k>1)となる場合, そのデータは k-匿名性を満たす [24,25]. これにより, 個人が識別される可能性が 1/k まで低減される.

図 3 に直接識別子を削除し, k-匿名性を満たすデータの例を示す. 全ての準識別子を等価クラスに変更し, センシティブ属性は変更していない. これによって, 準識別子を説明変数, センシティブ属性を目的変数としたデータ分析が可能となる.

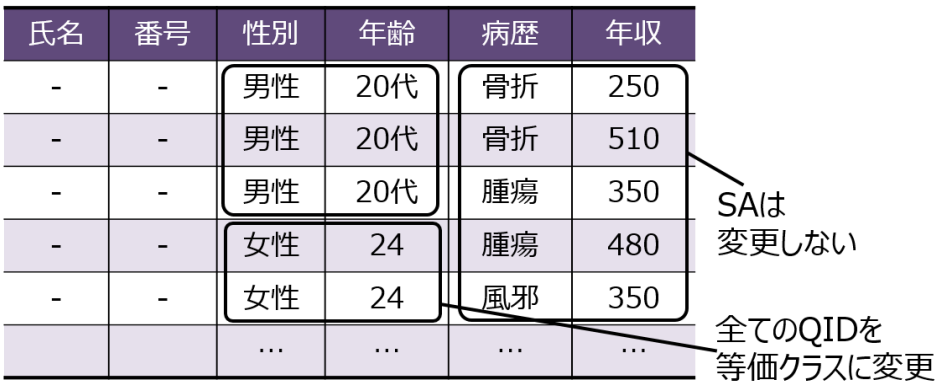


図 3 k-匿名性を満たすデータの例

しかし，センシティブ属性の処理方法については，使用する情報の種類や，国・地域による定義等によって変化する．次章より，センシティブ属性の扱いについて，値を変更せずに利用する米国方式と，値を変更する日本方式の扱いについて記す．

2.3.1 米国におけるセンシティブ属性の処理方法

米国では，1996 年に HIPAA 法（Health Insurance Portability and Accountability Act of 1996 ； 医療保険の携行性と責任に関する法律）が制定され，個人識別性が高く，公開されることが望ましくない属性について2つの方法が示されている．1つは匿名化技術などを用いた，確率的な再識別リスクの減少措置であり，これは他の国（EU や日本等）と同じである．米国における特徴的な手法として，消去するべき 18 個の識別子が指定されており，それ以外の情報をセンシティブ属性（この場合は統計における目的変数）として，そのまま利用することは違法ではない，と定義されている[34]．以下に参考として識別子となりうる 18 属性を示す．

○参考：HIPAA の定める，識別子となりうる 18 属性

1.氏名 2.住所 3.日付(登録日，誕生日等) 4.電話番号 5.FAX 番号
6.メールアドレス 7.社会保障番号 8.医療記録番号 9.健康保険番号
10.銀行口座番号 11.証明書，ライセンス番号 12.自動車等の免許番号
13.通信端末番号やシリアル番号 14.Web の URL 15.IP アドレス
16.生体認証データ 17.顔等が判別できる写真
18.その他，他者と区別するために作られた識別子全般

図 4 にセンシティブ属性を残す形の匿名化処理の例を示す．匿名化処理するべき元データには識別子となりうる 18 属性が含まれるため匿名化処理対象として準識別子とし

て匿名化処理を行う. しかし, 他の属性はセンシティブ属性としてそのまま利用し, データの変更を行わない.

元データ

QID		SA	
住所	年齢	購入	年収
New York	36	ジュース	500万
Boston	34	水	200万
Chicago	36	ジュース	450万
Tokyo	25	雑誌	800万
Osaka	29	マンガ	200万

SAを含むk-匿名化処理

QID		SA	
住所	年齢	購入	年収
USA	30代	ジュース	500万
USA	30代	水	200万
USA	30代	ジュース	450万
Japan	20代	雑誌	800万
Japan	20代	マンガ	200万

図 4 SA を残す匿名化処理例

一方, センシティブ属性をそのまま残すことによる弊害も大きく, 匿名化された状態から再識別されるリスクの大きな要因として, センシティブ属性の情報が漏洩した場合が想定される. センシティブ属性の属性値から個人を再識別される脅威については, パーソナルデータにおける攻撃モデルの項にて述べる.

2.3.2 センシティブ属性のを区分しない k-匿名化処理

それに対して, 日本における個人情報保護法では, 個人情報として定義されている 4 属性(氏名, 年齢, 住所, 性別)以外の情報についても, 個人を再識別することができる情報は基本的に全て準識別子として扱うため, 準識別子とセンシティブ属性の区別がされていない[35]. また, 個人情報の保護に関する法律の改正法[10]によって, 要配慮情報(例:人種, 信条, 病歴, 前科など)という新たな定義も行われており, 属性情報の提供が制限されている.

このように, 準識別子とセンシティブ属性を明確に区分する基準がない場合, 分析対象となるデータについても, 準識別子と同様に匿名化処理を行い, 匿名化データの提供を行う必要がある.

全ての属性を準識別子と設定して, 匿名化処理を行った匿名化処理例を図 5 に示す. 全ての属性を準識別子と設定して抽象化を行い k-匿名化を実現する. この場合, 米国の基準においてセンシティブ属性と認識される属性についても, 準識別子と同様に一般化階層を用いて匿名化処理を行う必要がある. その場合, 処理するべき属性が多いことから, より抽象的で, 多くの概念を含む語への書き換え処理が必要となる.

そのような過度な抽象化によって, 分析対象が失われてしまう場合, 多様な属性値を持つ属性に対して, かく乱, スワッピング, ノイズ付与などを施し, 確率的に元情報に戻すことができない処理を施すことで, 安全性を高める場合もある.

このように，国・地域の基準に応じて匿名化処理するべき属性の基準が異なる．そのため，国際的な情報流通や，基準が異なる組織同士における匿名化データの流通には，事前に安全性の基準のみならず，処理するべき属性と，匿名化処理を行った場合に失われる情報について同意する必要がある．

元データ

QID			
住所	年齢	購入	年収
東京	36	ジュース	500万
千葉	34	水	200万
埼玉	36	ジュース	450万
京都	25	雑誌	800万
大阪	29	マンガ	200万

SAを含まないk-匿名化処理

QID			
住所	年齢	購入	年収
関東	30代	飲料	200-500万
関東	30代	飲料	200-500万
関東	30代	飲料	200-500万
関西	20代	本	200-800万
関西	20代	本	200-800万

図 5 SA を含まない匿名化処理例

2.4 パーソナルデータに関する攻撃モデル

k-匿名性をはじめとするパーソナルデータの安全性に関する指標は，そのデータの種類や加工方法，提供先のリスク等に応じて異なるものが提案されている．

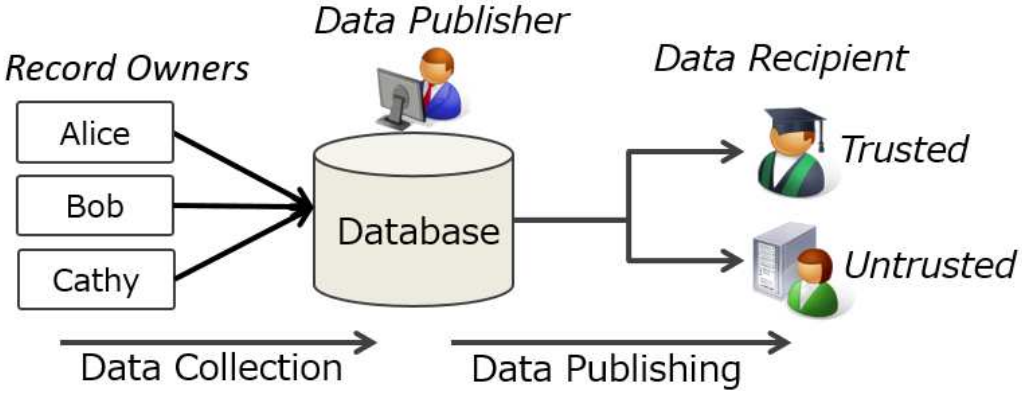


図 6 Fung らが考える攻撃モデルの範囲

図 6 に Fung らの攻撃モデルを示す [44]. Data publisher は，Record Owners からデータを収集し，Database に蓄積した後，Data Recipient に提供する．これを Data Publishing と定義する．この時，提供する相手は，契約等の締結と，Record Owners からの許諾を得ている Trusted な場合と，提供した Database を利用して Record Owners への攻撃を試みることが想定される Untrusted の場合がある．

このような，Data Publishing されたパーソナルデータに対して，Untrusted な Data Recipient から受ける攻撃モデルを以下の 4 種類と定義し，それぞれの攻撃モデルに対して有効な指標を整理している．

1.レコード結合(Record linkage)

2.属性結合(Attribute linkage)

3.テーブル結合(Table linkage)

4.確率的攻撃(Probabilistic Attack)

本章では，それらの攻撃モデルと，それぞれの防止のため，有効な匿名化処理アルゴリズム，指標について述べる．

2.4.1 レコード結合(Record linkage)

レコード結合(Record linkage)は，最も多く発生する攻撃モデルである．

パーソナルデータにおける識別子，準識別子を用いて，ユーザの一意絞込み(シングルアウト)が発生する．それによって個人が識別され，本来公開してはならない情報が漏洩する攻撃モデルである．

Sweeny は，[24]において，投票者リスト(Public voter list)に存在する名前と，医療履歴データベースに含まれる Zip コード，誕生日，性別を QID として結合することで，マサチューセッツ州知事の病気に関する情報を得ることができた．また，同様の手法を用いることで，米国国民の 87%を一意に識別することができることを報告した．

QID の組み合わせによって一意絞込みされたデータを基点として，他のパーソナルデータとの結合が可能となり，複数のデータが連結されることによって，本来知られてはいけない属性値が判明する．

そこで，パーソナルデータに含まれる QID を書き換え，より抽象的な概念に変更し，個人の再識別可能性 1/k まで減少させる k-匿名性(k-anonymity)が考案された[24]．

また，k-匿名性の概念を拡張し，レコード結合に含まれる他のリスクを定量化する指標も提案されている．これらの指標は，k-匿名性を満たした上での追加的な指標である．

特に，単一のパーソナルデータから複数の k-匿名化データを生成することによるリスクが多く指摘されており，Multi R k-anonymity[45]や(X,Y)-anonymity [46]が知られている．

Multi R k-anonymity[45]は，1つのパーソナルデータから複数の k-匿名化処理されたデータを生成する際に，他の k-匿名化データで利用されている一般化階層を用いずに，データの結合性と属性値が類推される可能性を弱める指標である．

(X,Y)-anonymity [46]は，属性値 X と，それに対応する属性値 Y について，少なくとも k 種類以上に区分されている，多様な属性が含まれていることを求める指標である．

また，k-匿名性は集合化や抽象化処理に着目し，ある個人に対する再識別確率という指標から生成されたため，データのかく乱やスワッピング処理等のデータ全体に対する確率調整処理を想定していない．そこで，データのかく乱，ノイズ追加，スワッピング処理等

を行った結果，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指標として，Pk-匿名性が提案されている[47,48].

匿名化処理の議論は，識別子が一意に存在するマスターデータ形式についてのみ適用されるのではなく，同一識別子が複数回出現するトランザクション型のデータに対しても適用される.

代表的なトランザクション型の匿名化指標として，k^m-privacy[49,50]，LKC-privacy[51]，(h,k,p)-privacy[45]，などが存在する.

k^m-privacy は，トランザクションデータにおける行動情報において，m個の情報が同一である等価クラスが最低でも k 個以上存在することを示す指標である. m=∞の時，全ての属性値について k 個以上存在することが保証される[49,50].

LKC-privacy は，位置情報について用いられる指標である. 攻撃者が知りえる位置情報の軌跡の長さの上限を L と設定し，L 以下の長さを持つレコードが k 個以上存在することを要求する指標である[51].

(h,k,p)-privacy は，攻撃者が p 個以下の外部情報を持つ場合を想定して，行動履歴の k-匿名性と属性値の多様性に関して，全ての値が h%以下であることを要求する指標である[52].

また，k-匿名性の確率的展開である Pk-匿名性のように，仮名化データであったとしても，かく乱やデータのスワッピングを行うことで，再識別を防ぐことが可能な仕組みが提案されている[53,54,55].

ユーザの ID を仮名に変更し，その仮名ユーザが同じ場所に存在したとき，ランダムで ID を交換することでかく乱を行い，ある個人が特定される可能性を減少させたことを示す指標として，(k,t)-privacy[56]が提案されている. これは時間 t における到達可能場所が k 個以上存在することを保証するという指標である.

k-匿名化処理やその関連指標は，属性の抽象化等によってレコード単位による再識別を防止するが，k-匿名性を満たし，Record Linkage を防いだ状態であっても，再識別や復元のリスクが発生する場合がある. 属性結合は，抽象化処理を行った場合の問題点を指摘するものである.

2.4.2 属性結合(Attribute linkage)

属性結合は，攻撃者が攻撃対象となる個人について詳細に情報を知らない場合でも，その所属している属性によって個人のプライバシーを侵害することができる攻撃である.

その種類としては，同種攻撃(homogeneity attack)と，背景知識攻撃(background knowledge attack)が知られている[26].

同種攻撃は，属性値を抽象化した場合でも，その属性に含まれる内容が単一である場合に，個人が知られたくない情報が推定されてしまう攻撃である.

背景知識攻撃は，準識別子とセンシティブ属性の組み合わせ，または属性値の出現数や分布の特性などから，センシティブ属性の値を知られてしまう攻撃である．

これらの攻撃を防止するため，属性値の抽象化と集合化が正しく行われているかを判定する指標が考案された．

匿名化処理における抽象化は，一般化階層などを用いて，より多くの意味を持つ語に変換することで，該当する属性値を増やす手法である．例えば(19 才,20 才)の属性値を(10 代,20 代)という，より抽象的な値に変換する処理を指す．

それに対して集合化は，属性値を抽象化した際に，複数の属性値が含まれる処理を伴う．例えば(10 歳,11 歳)の属性値を(10 歳又は 11 歳)という属性値にまとめることで，複数の意味を持つ値に変換される．抽象化と集合化は，同時に行われる場合があるが，複数の定義が 1 つのデータ内に混在することで元の属性が類推されるリスクが発生する場合がある．

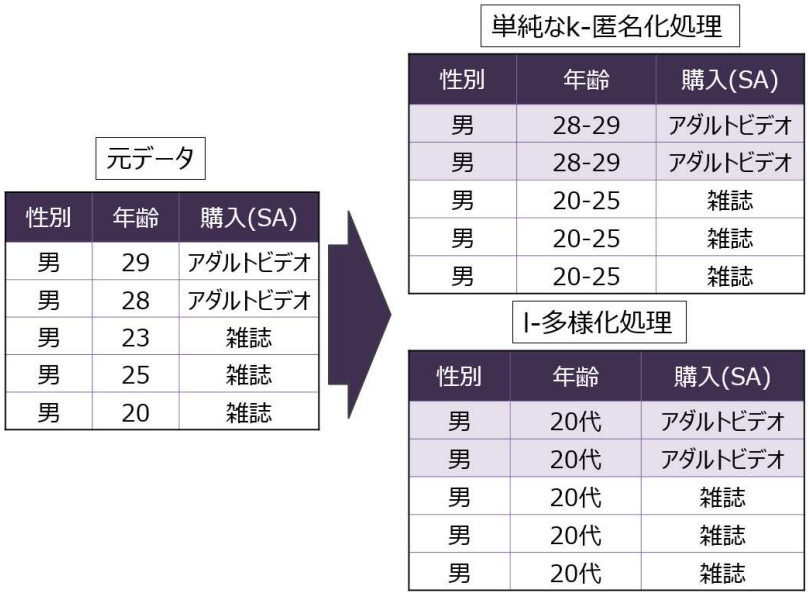


図 7 単純な k-匿名化処理の例(上)と，ℓ-多様化処理の例(下)

ℓ-多様性(ℓ-diversity)[26]は，情報を抽象化するだけでなく，集合化が正しく行われているかを計測する指標である．ある抽象化された等価クラスに含まれる属性の種類が，ℓ 個以上存在することを保証する．これによって，公開された属性からプライバシーの侵害が発生する確率を減少させることができる．

図 7 は，単純な k-匿名化処理を行った例と，同種攻撃を防止するため ℓ-多様化を実現した処理の比較例である．

単純に k-匿名化処理を行った場合，センシティブ属性も含めた全てのレコードが等価クラスとなっており，個人が識別されない．しかし，センシティブ属性に含まれる値が単一である場合，その結果データから該当する人物を特定することが容易である．図 7 の例では，

知人や家族がこのデータに含まれていることを知っている場合，その人物が該当する準識別子と同じ属性を持つ人物であるとき，センシティブ属性が判明する．

また，例えば，自分がその属性に含まれる人物である場合，自分と同じ属性の人物を調査すれば，その人物が同様のセンシティブ属性を持つことが類推できる．

そこで，準識別子の等価クラスに含まれるセンシティブ属性を多様化し，類推を防ぐ処理が必要となる．これがℓ-多様化処理である．このような脅威を検定する手法として[57]による評価手法の提案や，位置情報の安全性を増すための[58]の手法などが知られている．

しかし，ℓ-多様性を満たした場合でも，情報が類推される場合がある．図 8 に，ℓ-多様化処理を行った場合でも属性が推定されやすい状況を示す．ある属性値に対して，より抽象的な属性値が存在するとき，k-匿名性と ℓ-多様性を同時に満たした場合でも，抽象的な属性値に含まれる属性値の集合に，知人や近親者にとっては違いが認識しやすい概念が含まれている場合，プライバシー侵害が発生する可能性がある．

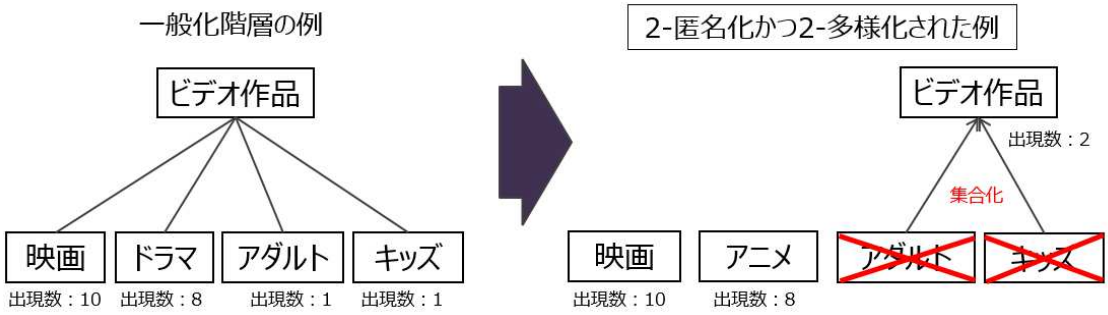


図 8 2-匿名化かつ 2-多様化が行われた例

このような対策として，t-近傍性(t-closeness)[27]が考案されている．これは，等価クラスにおけるセンシティブ属性の分布と，テーブル全体におけるセンシティブ属性の分布との差異を t 値以下に抑えるという処理である．

類似した指標として，(α, k)-匿名性[59]がある．こちらは，任意の等価クラスに含まれる任意のセンシティブ属性値が，α 以下の確率で存在していることを保証する指標である．

また，k-匿名性における確率的な展開方式として Pk-匿名性が提案されたように，ℓ-多様性を確率的に展開した Pℓ-多様性が提案されている[60]．

2.4.3 テーブル結合(Table linkage)

テーブル結合は，公開された匿名化データにおける個人のデータが，何らかの別の方法にて攻撃者に知られていた場合に，個人を再識別できる可能性が高まる攻撃方法である．

例えば，あるサービスプロバイダーが匿名化データ T を公開した場合を考える．

公開された匿名化データ T と，従業員に公開されている顧客台帳 E が存在し，E⊆T であることが期待できる場合，E に含まれる要素が T に含まれている確率を計算することができる．図 9 にその例を示す．

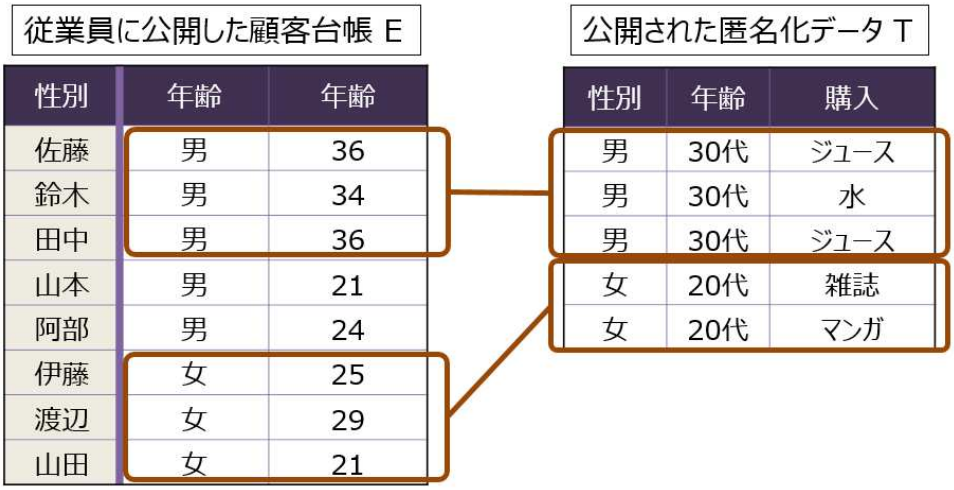


図 9 テーブル結合により個人の行動が類推される例

まず，E に含まれる等価クラスサイズと T に含まれる等価クラスサイズを比較することによって，E に含まれる人物が T に含まれる確率を計算できる．もし E の等価クラスと T の等価クラスが同数ならば，その人物が T に含まれる確率は 100%である．また，それによって，E に含まれる人物が，T に含まれるセンシティブ属性である確率を計算することができる．ある人物が T に存在する確率が 100%の時，その等価クラスに含まれるセンシティブ属性の出現確率は，その人物のセンシティブ属性である確率と同値となる．

δ-存在性(δ-Presence)[61]は，このような複数の公開テーブルによる個人再識別率をδ%以下に低減させたものである．

2.4.4 確率的攻撃(Probabilistic Attack)

確率的攻撃（Probabilistic Attack）は，パーソナルデータにおけるレコードや属性値を用いるのではなく，その公開されたデータの集計値に対して行う．過去に提供したデータについて，その後，時間を置いた後に再度提供したデータとの統計的差異を検証することで，変化した個人を識別する攻撃であり，差分プライバシーとも呼ばれる．

あるデータ保持者が，定期的に匿名化データをデータ利用者に提供する場合を想定する．データ保持者が n 回目に提供したデータと，n+1 回目に提供したデータは，ある属性値を保持する人数が 1 だけ異なるとする．そのとき，n 回目から n+1 回目のデータ提供時までに，その属性値の変化に該当する行動した人物を識別される可能性がある．

しかし，この差分に該当するユーザのプライバシーを守るため，変化した数値を前回と同じデータにしたり，またはノイズを増加させた場合，データ全体におけるユーザ構成比や属性同士の相関係数等が変化してしまい，分析目的が失われる場合がある．

そこで，一定の値の変化がないと提供しないポリシーを定める (c,t)-Isolation[62]，全ての準識別子の等価クラス内に m 個の情報が含まれていることを確認する，m-Invariance(m-不変性)[63]，等価クラスの出現数に応じてラプラス分布型のノイズを加えることでユーザの再識別を防ぐ ε-differential privacy(ε-差分プライバシー)[64]などが定義されている．

2.5 匿名化処理の手順

2.4 章で述べた安全性指標の多くは，Record Linkage を防止する k-匿名性の概念を拡張し，ユースケースに合わせた課題を解決，補完するための安全性指標である．そのため，匿名化処理においては，処理対象テーブルが k-匿名性を満たしたかを確認し，その上で必要となる他の安全性指標も検証するという手順が必要である．

そこで，k-匿名化を含む安全性指標を検証するための，匿名化処理の手順について検討する．

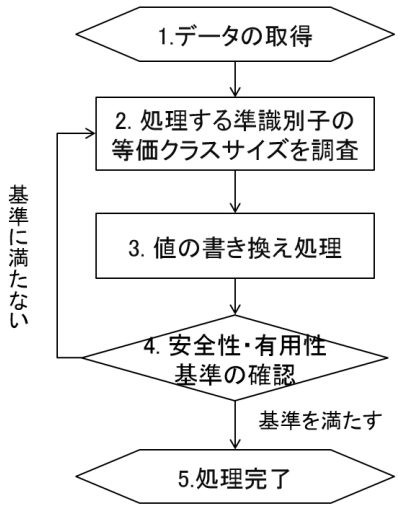


図 10 匿名化処理のフロー

匿名化処理は，準識別子における属性値の出現数を検証し，属性値において，個人識別のリスクが高い，または，プライバシー侵害が発生する場合，その値を書き換える(Recoding)ことで安全性を高める処理を行う．

その時，安全性を高める処理として，図 10 の流れで処理を行う．

まず 1. データの取得 を行い，2. 対象データに含まれる準識別子の等価クラスに含まれる属性値の出現数(等価クラスサイズ)を検証する．その後，3. 全ての等価クラス

において，安全性指標（k-匿名性，l-多様性など）を適用し，安全基準を満たしているかを検証する．もし，安全性指標を見てしていない値が存在した場合，再度，安全基準を満たすために必要な等価クラスのサイズを検証し，書き換え処理を繰り返す．全ての値について安全性基準を満たした場合，5．匿名化処理完了となる．

匿名化処理を行う場合，主に 2．等価クラスサイズの検証と，3．値の書き換え処理を繰り返し，基準を満たすまで処理を行うが，書き換えの手法はデータの特徴や使用目的に合わせて多岐にわたるため，評価指標や処理アルゴリズムが多く存在する．

2.5.1 等価クラスサイズの検証と書き換え処理

まず，k-匿名性を確認するため，全ての等価クラスのサイズを計測する必要があるが，その際に使用するデータ単位を決定する．

等価クラスの書き換え単位として，局所的な処理である Local Recoding と，属性値全体の統計情報から処理を行う Global Recoding の 2 種類が存在する．

Local Recoding はデータベースの部分集合に対して匿名化処理を行うため，単位あたりの処理量が少なく，分散処理に向く．また，各等価クラスサイズを自由に設定できるため，k-匿名性の予測は容易だが，等価クラスに含まれる属性値の制御が難しい．

Global Recoding はデータベース全体を用いて統計情報を作成するため，処理量は多いが，利用者が求める属性値について，統計情報を参照しながら調整できる利点がある．その一方で，求める属性値による処理の結果が，求める k-匿名性を実現するかについて，予測は難しい．その概念図について図 11，表 1 にて示す．Local Recoding は，等価クラスサイズを自由に設定できるため，k 値を予測することは可能だが，そこに含まれる属性値の種類について制御が難しい．そのため，分析目的に合致しない等価クラスが生成される可能性がある．逆に Global Recoding では，分析目的に合致しない等価クラスが作られることは少ないが，その k-匿名性を予測することは難しい．

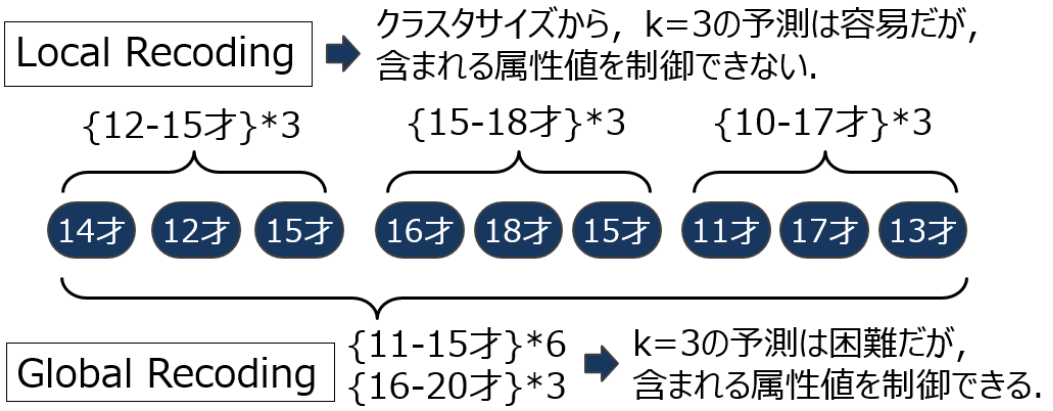


図 11 Recoding 手法の特徴

表 1 対象となったサービスと顧客群

	データ処理量	含まれる属性値の制御	k-匿名性の予測
Local Recoding	少ない	難しい	容易
Global Recoding	多い	容易	難しい

一般的に，位置情報や購買ログなどのトランザクション型データは，各属性値の出現数が頻繁に変更するため，局所的な単位で情報をクラスタ化する方が効率的である．そこで有用性と匿名性を維持しつつクラスタ化処理する [65][66]等の Local Recoding アルゴリズムを採用し，行動分析や機械学習に利用することが多い．

逆に，Global Recoding は，[67]等のアルゴリズムが存在し，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(ミクロデータ型)の情報に対して採用し，各属性値の出現数を統計化したうえで匿名化処理する．結果データは統計処理化や，回帰分析等に活用される．

2.5.2 属性値の書き換え処理の分析

等価クラスの数の処理単位を決定した後に行う，属性値の書き換え処理について述べる．

書き換え方式は，まずデータ全体への処理を行う場合と，個別の値に対する処理に区分される． データ全体への処理としては，元情報から一部を抜き出すサンプリング，レコードを追加する行追加，属性値を追加するノイズ付与，出現数の少ないものを排除する裾切り，値を入れ替えるスワップ処理などがある．

サンプリングや行追加，ノイズ付与，裾切り等の処理は処理するべきデータの初期の段階で行われ，全体の安全性を高める役割で用いられる．また，スワップ処理は値の合計や平均値が変化しないため，データの利用目的が明確な場合は有効である．

また，個別の値に対する処理としては，複数の値を組み合わせて新たな属性値を作るクラスタ化，平均値や最大値などの値に統一する代表値化，より抽象的な属性値に変化する抽象化，値の一部を削除して抽象化する部分削除等の方式がある．これらの処理は単体で行われるのではなく，複数組み合わせることで等価クラスサイズを制御する．

個別のデータに関する処理としては，最も距離が近い値のクラスタ化（Similarity based clustering）が用いられる．距離の指標にはマンハッタン距離やユークリッド距離等が用いられる．クラスタ化はデータの精度を集合化の限界点まで維持することが出来る点が優れているが，分析対象属性値が明確に決まっている場合には利用することが出来

ない. 例えば, 20 代のユーザを分析する際に, (29-30 才)というクラスタが生成された場合, 利用目的を達成しない, 等の問題点がある.

そこで, データの精度ではなく, データの利用目的に沿ったデータ書き換えを実現するために, 代表値化, 抽象化, 部分削除などの処理が用いられる[68]. 代表値化は, 平均値や最大値などの値に集約する方法であり, 数値属性に主に用いられる. 部分削除などの手法は, 属性値が正規化されている場合に有効だが, 不定形の場合には対応が難しい.

そのため, データ利用のニーズに応えるため, Datafly 方式[69]や μ-Argus 方式[17]などの匿名化処理アルゴリズムでは, データ全体の等価クラスの出現率を求め, 一般化階層によって情報を書き換える処理が使われる.

2.5.3 一般化階層を用いた匿名化処理について

匿名化処理において, 段階的に属性値を抽象化し, 安全性を高めるための手法として使われる, 一般化階層を用いた処理について述べる.

匿名化処理では, ある QID の属性値を一般化してより抽象的な値に変更し, その結果, 識別されるレコードが少なくとも k 個(k>1)以上になるように書き換える. しかし, 書き換えた結果が求めた安全性, または有用性を満たさない場合, 条件を満たさない属性値を, 更に抽象度の高い候補に書き換える. そのような抽象化処理を効率化するため, 一般化階層が用いられる.

あるパーソナルデータに性別と年齢属性が含まれており, 表 2 の一般化階層を用いて k-匿名化処理を行う場合, 図 12 のような Lattice Structure[70](格子構造)を作成し, 属性同士の全組み合わせを作成し, それぞれの属性組み合わせにおける k 値を検証する.

表 2 年齢属性の値一般化階層の例

性別	2区分	男				女			
年齢1	2区分	10代				20代			
年齢2	4区分	10-14才		15-19才		20-24才		25-29才	
年齢3	20区分	10才	・・	15才	・・	20才	・・	25才	・・

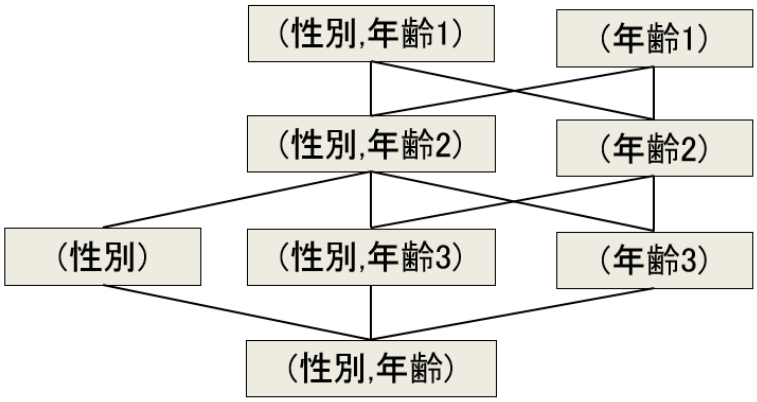


図 12 表 2 を用いた Lattice Structure の例

また，データの抽象化処理によって，分析対象の削除や過度な変更が発生し，データ利用目的が損なわれる場合がある．

例えば，一般化階層の n 区分では分析目的が達成できるが，k-匿名性は満たさない場合，一般化階層を，その分析目的だけが達成できる形に作り直す処理が必要となる．しかし，その新しい一般化階層を用いた場合でも，k-匿名性を満たすか，確認する手段が無いため，匿名化処理を繰り返し行い，有用性と安全性の条件を満たす一般化階層を探索する必要がある．

2.5.4 一般化階層の適用範囲

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検証しながら，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．

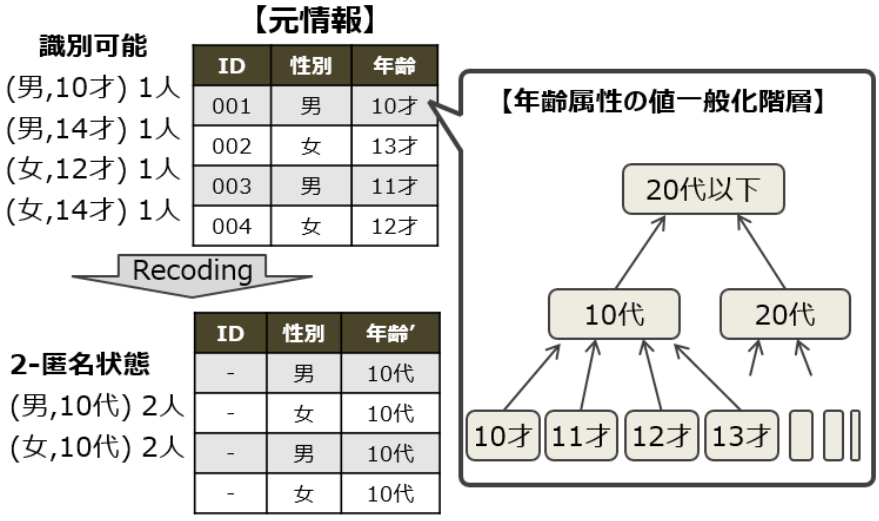


図 13 2-匿名化処理の書き換え例

例えば，図 13 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-匿名性の条件(k>1)を満たしていない．

そこで，匿名化処理として値一般化階層（VGH:Value Generalization Hierarchy）[69]や，属性一般化階層（DGH:Domain Generalization Hierarchy）[71]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．

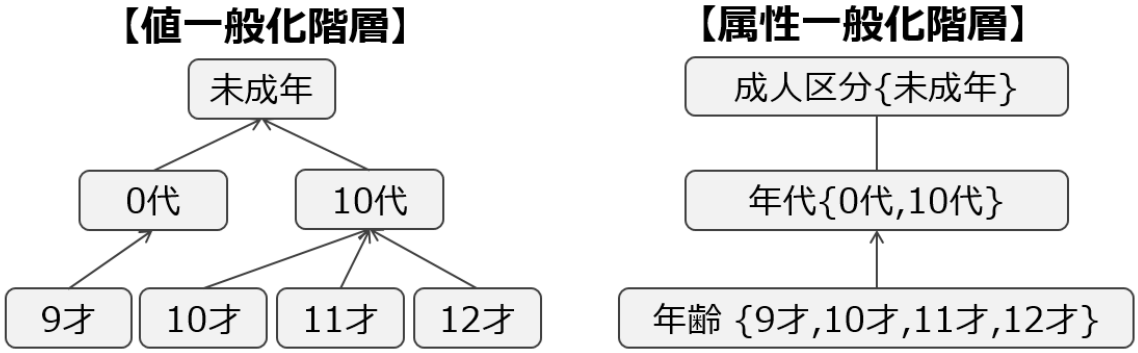


図 14 DGH と VGH の概念の違い

図 14 に，VGH と DGH の概念の違いを示す．VGH は階層が樹形状になっており，何か一つの属性値を抽象化した際に，他の属性値が抽象化されることは無い．そのため，データの情報量が失われないという利点がある．

DGH は，全ての属性が同じ抽象化レベルに属しているため，ある一つの属性値について抽象化が必要な場合，他の全ての属性値も同時に抽象化されるという違いがある．そのため，元データの分散の特徴によっては，情報量が失われる場合がある．

そこで，情報量が少なくなる場合に対して Datafly[69]等のアルゴリズムでは，値の削除(suppression)を行い，出現率が少ない値を排除することで対応する．

それに対して，μ-argus[17]方式や minDIS[71]等の VGH を用いるアルゴリズムにおいては，存在する値単位での抽象化を行うため，異なる階層が混在した結果が出力される．これによって，より詳細なクラスタを生成することを目的としており，DGH よりも情報量が維持される．しかし，元データの分散の特徴によっては，無意味なクラスタを生成する等の問題点が存在する．

また，匿名化処理に用いられる一般化階層を，出力データから自動的に生成し，VGH を自動的に生成する手法[72]も提案されている．

2.5.5 一般化階層を用いた場合の有用性指標

カテゴリ属性に対して，一般化階層を用いて抽象化した値に書き直した場合の有用性指標として，Datafly 方式における属性値の抽象化レベル評価のために考案された

Prec[69]や，元データからのデータ歪曲度を評価するために考案されたデータ歪曲度算出関数(DIS) [71]などが知られている．どちらの指標も，元情報を基準として，抽象化された値の，一般化階層における抽象化度を求める式であるため，一般化階層を用いない場合には適用できないことが難点である．

Precは，DGHを用いた場合の各属性値の変化を，平均の高さの変化数に変換したものである．(1)にPrecの定義を記す．元データRTと匿名加工データPTを比較する際の平均の高さの差を計測する．それぞれの属性値を $t \in PT=\{t_1,\ldots,t_N\}$，$t' \in RT=\{t'_1,\ldots,t'_N\}$ とし，準識別子 $QI_T=\{A_i,\ldots,A_j\} \subseteq \{A0,\ldots,A_{NA}\}$ とする．また，関数 h はその DGH における高さを返す関数とする．Prec においては，元データ PT に対して，匿名化処理後のデータ RT を作成する際に，一般化階層を用いずに匿名化処理が完了した場合 Prec(RT)＝0 となり，全ての高さが DGH の最高値（$\forall h=DGH_{Ai}$）まで抽象化された場合，Prec(RT)＝1 となる．

$$Prec(\mathrm{RT}) = 1 - \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N}\frac{h}{|DGH_{A_i}|}}{|PT|\cdot|N_A|} \quad \cdots(1)$$

DIS[71]は VGH の指標に対する各属性値の変化量を示す指標である．一般化階層全体の高さ DGH_{Ai} において，各 VGH_{Ai} の高さの平均を求めたものである．(2)に DIS の定義を記す．元データ RT と匿名加工データ PT を比較する際の歪み度を計測するため，それぞれの属性値を $t \in PT=\{t_1,\ldots,t_N\}$，$t' \in RT=\{t'_1,\ldots,t'_N\}$ とし，準識別子 $QI_T=\{A_i,\ldots,A_j\} \subseteq \{A_0,\ldots,A_M\}$ とする．また，関数 h はその VGH における高さを返す関数とする．こちらも一般化階層を用いなかった場合，DIS(RT)=0 であり，DGH_{Ai} の最高の高さまで抽象化された場合は DIS(RT)=1 となる．

$$DIS(\mathrm{RT}) = \frac{\sum_{A_i \in QI}\sum_{t_j \in PT}\frac{h\left(VGH_{A_i},tj(A_i)\right)-h\left(VGH_{A_i},tj'(A_i)\right)}{|DGH_{A_i}|}}{|PT|\cdot|QI|} \quad \cdots(2)$$

これらの指標は一般化階層が存在することが前提となっており，数値属性等との比較が出来ないため，カテゴリ属性のみに適用できる指標である．

数値属性の変化と比較し，全体の変化量を検証するためには，NCP(Normalized Certainty Penalty)[65]等の数値，カテゴリ属性の両方で利用できる指標を用いる．

NCP において，カテゴリ属性の場合は，匿名化処理前後における等価クラス数(Cardinality)の変化率で評価し，数値属性の場合，最大値から最小値を引いた値の変化率で評価する．(3)はカテゴリ属性に関する定義である．Card(A_j)に元データの等価クラス数，Card(u)に匿名化処理後の等価クラス数を用いて，その比率を求める．(4)は数値属性に関する定義である．属性値 A_i における最大値と最小値の差と，匿名化処理後の最大値と最小値の差の比率を求めている．

$$NCP_{A_i}(\mathrm{G}) = \frac{Card(u)}{Card(A_i)} \quad \cdots(3)$$

$$NCP_{A_i}(\mathrm{G}) = \frac{Max_{A_i}^{G} - Min_{A_i}^{G}}{Max_{A_i} - Min_{A_i}} \quad \cdots(4)$$

匿名化データの有用性指標 Prec, DIS, NCP の特徴について表 3 に示す.

Prec は DGH にのみ適用でき, DIS は VGH にも適用可能な点が優れている. NCP は, カテゴリ属性については等価クラス数の変化量, 数値属性については最大値と最小値の差の変化量について定義しているため, 数値とカテゴリ属性の両方で利用できるが, 一般化階層における値の変化量を評価できない.

表 3 評価指標とその適用可能範囲

評価指標	一般化階層の変化量評価	DGH カテゴリ属性	VGH カテゴリ属性	数値属性
Prec	○	○	×	×
DIS	○	○	○	×
NCP	×	○	○	○

また, これら全ての指標について, 値に対してノイズの付与やスワップ処理を行った場合に発生する, 9 才→10 才のような, 同じ階層同士の書き換えを行った場合に対応できず, 全ての匿名化処理結果を評価できる指標が求められている.

2.6 Global Recoding による匿名化処理アルゴリズム

一般化階層を用いて匿名化処理を行う手法として, Global Recoding について記述する.

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い, 匿名化処理の条件やデータの利用用途の条件を満たすまで, 処理を繰り返す.

出現数の検証処理は作成される属性同士の組み合わせの数だけ必要となる. Meyerson らの研究によると, 最適な k-匿名化を実現するには, 属性ごとに O(n log n) 回[28]の処理量が必要であり, NP 困難であるため属性数の多い個人情報では容易に計算困難となる.

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.

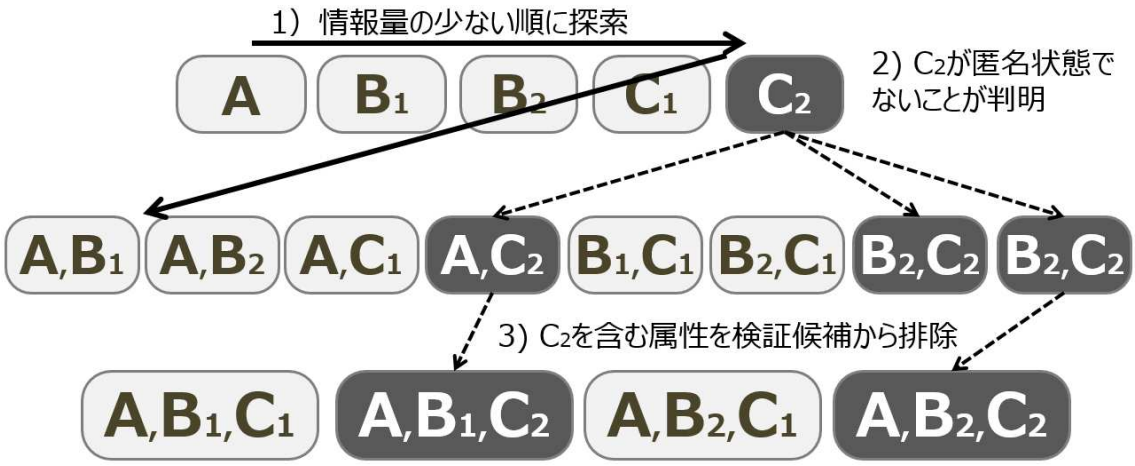


図 15 Incognito 方式による検証量削減方式の例

Incognito[67]は，トップダウン型で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，不必要な処理量を減少させている．図 15 にてその概念図を示す．まず，最も情報量が少ない 1 属性の群について，情報量の少ない順番に属性をソートし順番に探索する．その結果，ある属性は匿名化できないことが判明した場合，その属性を含む全ての属性組み合わせをその後の検証候補から排除する．これによって，探索するべき組み合わせ数が減少する．

トップダウン型の匿名化処理アルゴリズムは，情報を抽象的な群から処理するため，処理負荷の高い詳細な属性組み合わせ処理を省略する．そのため，検証処理量が少ない点で優れており，属性数，属性値の種類数が多い場合に，処理を効率化する目的に利用される．

Partition Algorithm[73]は，トップダウン型で属性値が大きいものから，一般化階層による細分化を進め， 細分化した属性値が k 値以下になるまで処理を繰り返すアルゴリズムである．これは，トップダウン型で Incognito と考え方が似ているが，最終的な結果データを VGH 型とすることを目的としており，また，結果データを最適化するために値の削除や変更を伴うため，結果データが Incognito とは異なる形で出力される．

逆に，ボトムアップ型の匿名化処理アルゴリズムは，最も情報が詳細な群から検証するため，情報量を多く維持したい場合に有益である．

Datafly[69]は，情報を詳細な群から一般化階層を適用して抽象化する，または値を削除していくことで，k-匿名性を達成するアルゴリズムである． しかし，Datafly は詳細な群から検証するためコストが高く，加えて複数の属性について，同一の階層で同時に複数の組み合わせで k-匿名性が達成される可能性があり，その中から最も有用性の高いものを，処理者が定性的に選択することを想定している．

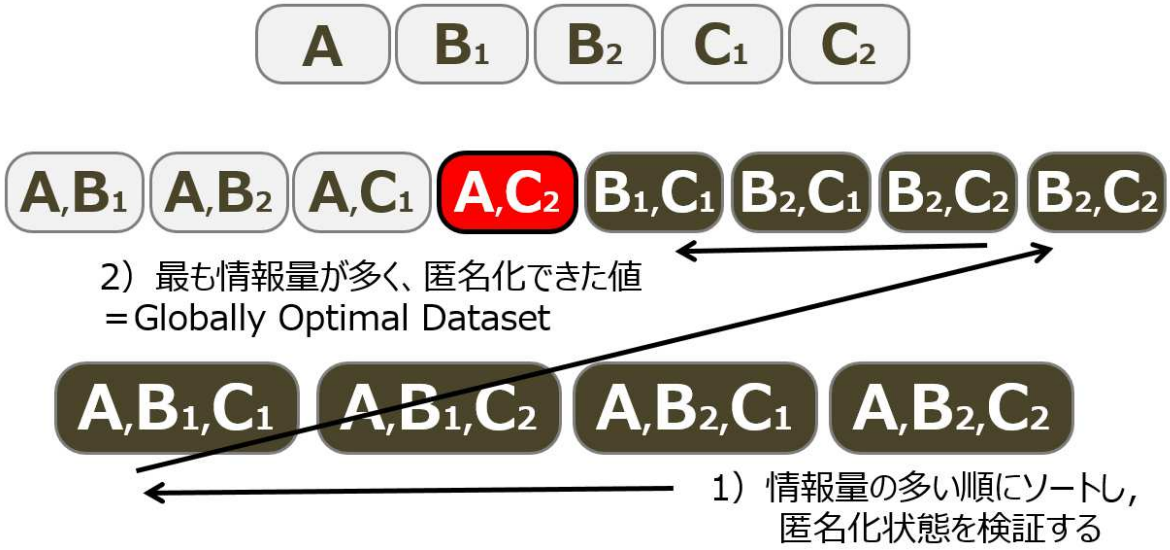


図 16 OLA 方式における GOD の探索方式の例

そこで OLA (Optimal Lattice Anonymization) [70]では，情報量に着目し，最も情報量の多い匿名化可能な属性の組み合わせを，ボトムアップ型で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探索する．図 16 にてその概念図を示す．まず，全ての属性値の組み合わせを作成し，最も情報量の多い順番にソートする．属性の組み合わせ数が多く，情報量が多い順番に匿名化状態を検証し，最も情報量が多く，匿名化できた属性組み合わせを”GOD（Globally Optimal Dataset）”として利用する

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組み合わせの探索を行い，条件を満たした場合にその後の処理を省略することで処理量を削減する．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．

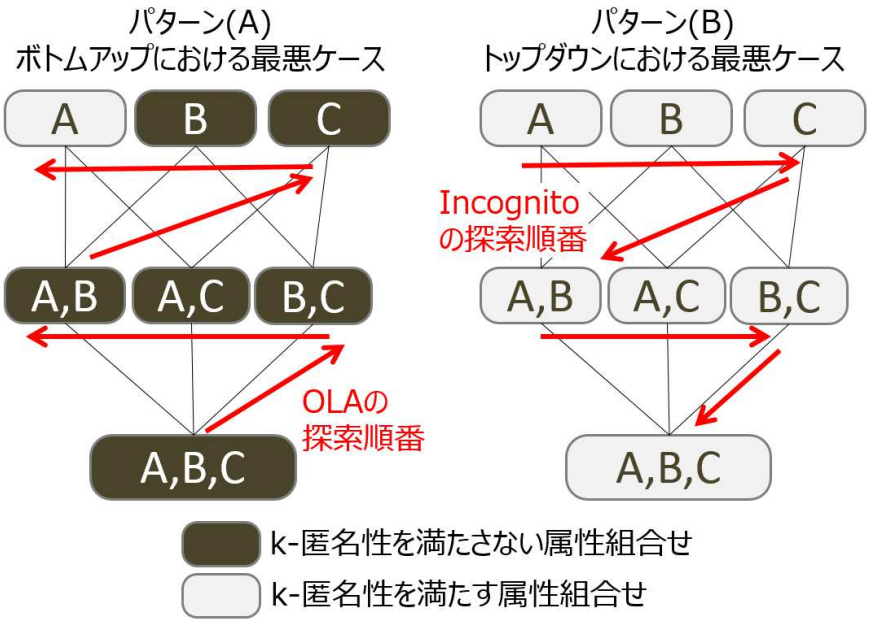


図 17 各アルゴリズムにおける最悪ケース

元データの属性値の分布の特徴によっては，属性の組み合わせの検証回数が全組み合わせの探索数と同じになる，最悪ケースになる場合がある．

例えば，図 17 のパターン(A)において，ボトムアップにおける最悪ケースの例を示す．求められた k-匿名性の条件によって k 値を大きくすることが求められる場合，最も抽象化された属性値によっても匿名化処理が達成できない場合がある．その場合，パターン(A)の形となり，その場合にボトムアップ型の OLA 方式を選択した場合，全組み合わせの探索が必要となり，最悪ケースとなる．

その逆に，求められた k-匿名性の条件によって k 値を小さくすることが求められた場合，最も詳細な群でも匿名化処理が可能な場合がある．その場合，パターン(B)の形となり，その場合にトップダウン型の Incognito を選択した場合，全組み合わせの探索が必要となり，最悪ケースとなる．

このような現象は，データの特徴や達成するべき k 値などによって変化するため，ボトムアップ型とトップダウン型のどちらが効率的に匿名化処理が可能かを事前に判断することは困難である．

2.6.1 有用性を維持した匿名化処理アルゴリズム

匿名化処理を行う際に，安全性のみを求めて処理を行うことは非常に少ない．実際の処理には，何らかの有用性要件が存在し，処理結果がそれを同時に満たすかを計測しつつ処理を行うことが求められる．

しかし，前述のとおり，最適な匿名化処理を求める場合における組み合わせ処理はNP困難であるため，有用性の次元を追加することによって，更に複雑さが増す．

そのため，有用性を何らかの指標に落とし込み，その指標と安全性の両要件を満たす組み合わせを探索するための手法が提案されている．

クラスタリング手法は，有用性を「情報のクラスタサイズの詳細さ」に求める手法である．図18にてUtility-Based Anonymization[65]の概念図を示す．属性値XとYの2軸を設定し，あるクラスタ同士で最も距離の近い値同士をLocal Recoding型でクラスタ化することで，有用性と安全性を満たす群を生成する．ある値と比較して最も値の差の総和が少ない値を探索してクラスタ化し，k-匿名性を満たすまでそれを繰り返す．この手法はトップダウン型でもボトムアップ型でも可能である．

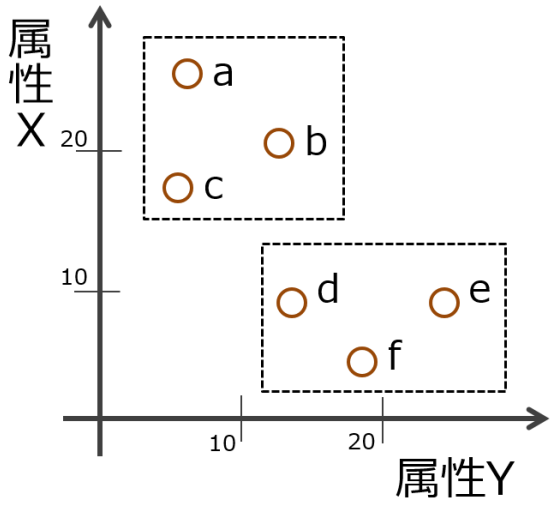


図 18 Utility-Based Anonymizationにおける2軸のクラスタ化の例

そこで，[74]では，有用性の軸が定まっていない属性値に対して，属性値の出現数から値の類似度を自動的に作成し，一般化階層を利用せずに有用性の高いクラスタ化を実現する手法を提案している．

多次元分割手法は，m個の属性を持つデータのk-匿名化処理を，m次元における空間分割の最適化命題と考える手法である．Mondrian[75]は，属性値XとYの2軸に対して，全てのクラスタがk値と同数になるようにトップダウン型で分割を繰り返していく．その概念図を図19にて示す．Raw Dataから開始し，全ての値がk-匿名性を満たすまで分割線を増加させる．

類似した研究として，多次元分割手法とクラスタリングを併用したk-匿名化処理方式[76]も提案されている．

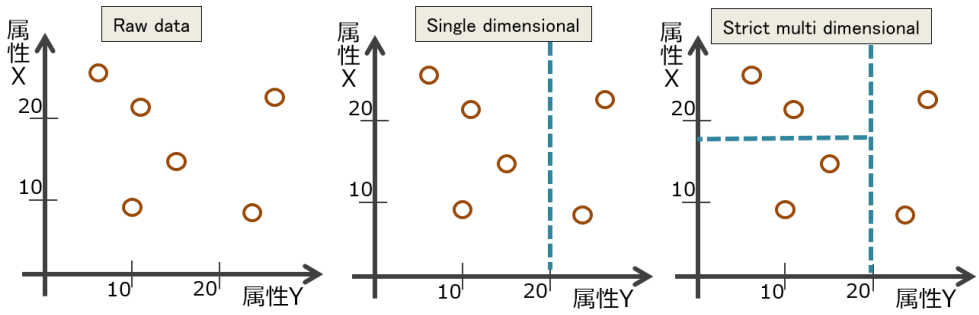


図 19 Mondorian における次元分割の例

クラスタリング手法と多次元分割手法は，両方とも詳細なクラスタを生成するための処理であるため，属性同士の組み合わせの制御が難しい．そのため，VGH のような属性同士の粒度が異なるクラスタが多く生成される場合があり，DGH のような同一の高さの階層に属性値を統一するなどの処理が不向きである．

マージナル[77]を用いた手法は，匿名化処理を行う際に一般化階層ではなく，別表(マージナル)によって変更するべき属性値を定義し，その定義に合わせた匿名化処理を行う手法である．例えば表 4 にてマージナルの例を示す．2 軸の抽象化範囲を定めた表を作成し，それぞれの範囲に含まれる値をカウントアップすることで等価クラスサイズを検証する方式である．最終的に全ての値が k 値以上になる場合，k-匿名性が成立する．

表 4 主な書き換え処理の例

Age/Income	{$0-$50k}	{$51k-$100k}	{$101k-$200k}
{20-24}	1	5	0
{25-29}	2	0	4
{30-34}	0	2	1

この手法は，マージナルを書き換えることで Age や Income の一般化レベルを変更することが出来るため，有用性を保った値に収束させることができる．これにより，一般化階層より詳細な条件を設定できるが，階層構造の設計が難しいため，一度設定した別表によって匿名化処理が達成できなかった場合の，探索型の匿名化処理には不向きである．

また，情報量を有用性の指標として定義し，VGH 型の一般化階層を用いて，最も情報量が維持され，かつ k-匿名性を満たすアルゴリズムが提案されている．

μ-Argus 方式[17]は，最も情報量の多いクラスタを生成するため，ある VGH に含まれる，各属性値の出現数を計測し，その中で出現数が k 値未満である属性値群を抑制(suppression)することで k-匿名性を達成する．本アルゴリズムにおける抑制処理は，値をより抽象化するだけでなく，値を消去する等の処理を含む．そのため，k 値未満の属性

値が多く存在する場合，多くの属性値が失われ，データの利用目的を達成できない可能性がある．

そこで，minDIS 方式[71]では，各属性値の出現数リスト（frequency list）から出現頻度が k 値未満の属性値をランダムで抽出し，最も歪み度（ここでは歪み度計測に 2.5.5 章における DIS を用いる）の近い値と集合化させ，出現数リストを更新する，という処理を繰り返す．これにより最も歪み度の少ない k-匿名化処理が達成できる．しかし，minDIS 方式は値をランダムで選ぶため，μ-Argus 方式と比較して，データの分散状態によっては，処理量が多くなる場合がある．加えて，実施する度にクラスタに含まれる属性値が異なるため，均一な結果が出力できないという問題がある．

これらの VGH 型の匿名化処理を行った場合，結果データにおける情報量は維持されるが，結果データに含まれる各クラスタの値を制御することが難しく，結果データが利用目的を達成できるかを事前に予測することができない．また，元となるパーソナルデータの分散の特徴によっては，結果データから個人の類推が可能になる点などが指摘されている．例えば，匿名化処理した結果が（男性，＊）であった場合，2値しか存在しない値を抑制したことによって，女性が k 人以下であることが類推される．または，連続値である結果データが削除された場合，例えば（16 才，＊，18 才）が出力された場合，削除された値は 17 才であることが類推される．このようなパターンによるプライバシー侵害が発生しないよう，μ-Argus 方式では，出力されたデータ群が適切であるか，データ保持者が確認することを求めている．

そこで，DGH と VGH の両方に対応し，一般化階層の探索型処理と有用性の両面を満たす処理として，属性値の SEM（検索エンジン広告）価格を用いた匿名化処理[78, 79]が提案されている．

これは，属性値がカテゴリ属性の場合にのみ適用できる手法で，各属性が独自の価値を設定できる場合に利用する．その例として，検索エンジン広告における取引価格を用いて，最も値段が高い語を含む属性の組み合わせが，最も有用性が高いと定義し，匿名化処理を達成するものである．

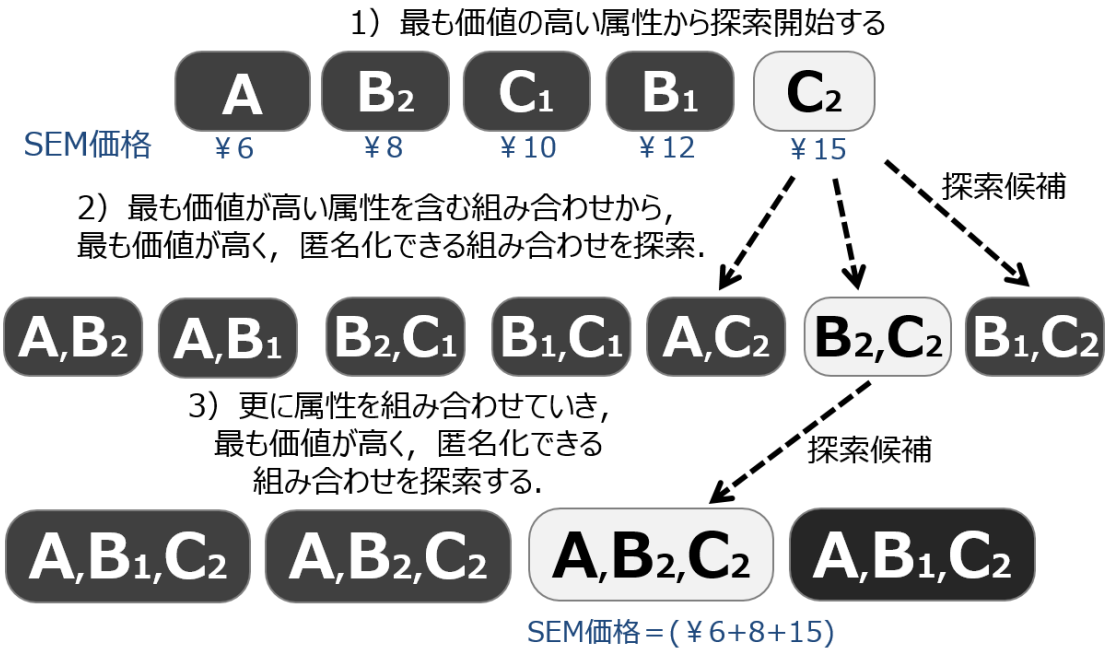


図 20 SEM 価格を用いた匿名化処理例

全ての属性値に価値がつくことで，属性*人数によって，その属性の持つ有用性が計測できる．そのため，その値の高い順に Lattice Structure を生成し，トップダウン型で匿名化処理を行っていく．図 20 にその概要を示す．まず，全ての属性値に SEM 価値が存在することが前提条件として必要である．そこで，各属性に含まれる属性値の SEM 価格の合計を出す．その中の最も価値が高い属性を検証し，それを含む属性組み合わせを探索候補とする．その時，他の組み合わせは検証しない．その中から匿名化状態を満たす組み合わせが発見できた場合，更にその値を含む探索候補を探していく．これを繰り返すことにより，総価格が最も高い属性組み合わせを効率的に探索できる．

しかし，語の価値によって有用性を計測する手法は，外部要因からの影響を受けやすく，金額の変動状況によっては，他のアルゴリズムと比して処理量が多くなる場合もあり，必ずしも効率的とは言えない．

そこで，より外部要因の影響が少ない指標として，検索エンジンにおけるリクエストクエリ数や，SNS における語数などを用いた匿名化処理を提案し [80]，災害時におけるパーソナルデータの匿名化に活用することを提案した．しかし，本質的な問題である，外的要因と処理量における問題点が改善されていない．

これら，有用性と安全性を両立させる匿名化処理は，探索範囲の次元が増加することから，匿名化処理を行った後に，安全性と有用性を検証し，条件を満たさない場合には，条件を変更して他の有益な組み合わせを探索する，Lattice Structure をベースとした探索型のクラスタ化処理が必要となる．

現状のアルゴリズムの多くが，データ保持者とデータ利用者が同一の組織・機関であり，有用性のチェックに際して，安全性を考慮していない．

社会実装を行うためには，データ利用者からの要求項目を有用性に変換し，かつ，安全性の高い仕組みによってデータの匿名化処理と授受が行われるべきである．

2.7　パーソナルデータの匿名化技術の動向

前章までは，匿名化処理技術に関する検討を行ったが，本章からは，匿名化データの流通プラットフォームの社会実装に向けた課題について検討する．そのため，匿名化データを流通させるためには，明確な安全性基準が必要である．

しかし，これらの匿名化データにおける定義と課題の設定は，各国や地域，自治体などによっても異なっており，多様な基準に対応できるシステムが必要とされている．

本章では，日本，欧州，米国における定義や手法を概説する．

2.7.1　日本における匿名化処理技術の動向

日本国内における匿名化処理の技術動向をまとめる．

日本における匿名化処理の定義として，内閣官房情報通信技術 IT 総合戦略室が設置した「パーソナルデータに関する検討会」は，個人識別性における「特定」と「識別」の概念を区分した[35]．

「識別特定情報（特定）」とは，表 5 に示すとおり「ある情報が誰の情報であるかが分かる」データである．具体的には，あるパーソナルデータや個人情報が攻撃者の手に渡ったときに，そのデータに含まれる氏名や住所，所属組織などから，その人物の社会的な状況が判明するデータである．

表 5 特定と識別の情報区分より

用語	用語の説明	データの種類
識別特定情報	個人が（識別されかつ）特定される状態の情報． それが誰か一人の情報であることがわかり，さらにその一人が誰であるかがわかる情報．	個人情報やパーソナルデータ
識別非特定情報	一人ひとりは識別されるが，個人が特定されない状態の情報． それが誰か一人の情報であることがわかるが，その一人が誰であるかまではわからない情報．	識別子が削除されたデータ（仮名化データ）．

非識別非特定情報	一人ひとりが識別されない(かつ個人が特定されない)状態の情報. それが誰の情報であるかがわからず, さらに, それが誰か一人の情報であることがわからない情報	匿名化データ.

「特定」は非常に強いプライバシー侵害を誘発する可能性がある. 例えば, パーソナルデータに記載されている属性値から, 直接的に個人への接触が可能になるため, 虚偽の連絡や脅迫行為, ストーカー行為などの犯罪行為に利用される可能性がある[36].

特定されたパーソナルデータは, プライバシー侵害を受ける対象個人に対する情報を多く知るものに悪用されやすい. そのため, 近親者や知人など, 対象者の属性値を良く知る利害関係者に, パーソナルデータから「特定」され, 情報が暴露されることは, 個人にとって不利益となる場合がある. また, 直接的な知人で無い場合でも, インターネット上においては, ソーシャルネットワーキングサービスなどを通じて知りえた情報から個人の情報を推測するなど,「特定」される情報の範囲も広がっている.

「特定」は, その人物の社会的な位置づけにあたる識別子, 例えば氏名や会員番号等の削除によって防止することが出来る場合がある. しかし, 識別子が無いデータであっても, 準識別子によって個人の特定が可能になる場合もある. このような, 準識別子による特定リスクを残すデータについて,「仮名化データ」と呼ぶ.

一方,「特定」と似た概念として, 社会的に誰であるかが判明しない場合でも, あるパーソナルデータ中において, その 1 名が存在することが判明する場合がある.

「識別非特定情報(識別)」は,「ある情報が誰か一人の情報であることが分かること」(ある情報が誰の情報であるかが分かるかは別にして, ある人の情報と別の人の情報を区別できること)と定義されている[35].

「識別」は, 特定よりも弱い概念と考えることができるが, 企業にとっては, ある個人を「特定」し, それぞれ属性値にターゲティングしながら対処することは, コストが高い作業である. 逆に,「識別」によって同じ属性値や同じ行動傾向を持つグループに対して, インターネット広告, メール配信, WEB サービスの提供, などの形態を用いて一斉に行うやり方が効率的である.

個人情報保護ポリシーが不明確な企業や, 海外サービスなど自国民のデータに対する人権保護が行われない場所にパーソナルデータを提供することによって, 不要なダイレクトメールの増加, 個人への過度なプロファイリングとサービス差別, マルウェアなどの標的に遭うなどの迷惑行為につながる場合もある.

概念としては, 識別 ⊆ 特定 であり, より広範な概念である, 個人の識別性を低減させることを, 本研究では匿名化処理, と呼び, 匿名化処理がなされたデータを匿名化データと定義する. しかし, これらの匿名化処理は万能ではなく, 汎用的に利用できる匿名化方法は存在しない[35], とされている.

また，これらのレポートを受け，匿名化データは 2015 年 9 月に成立した個人情報の保護に関する法律の改正法[10]によって，より明確に定義されることとなった．

個人情報保護委員会規則で定める基準に従い，個人情報を加工して特定の個人を識別することができないようにするとともに，当該個人情報を復元することができないようにしたもの(改正法三十六条)を，匿名加工情報という新た情報の類型とした．

「匿名加工情報」は，日本の法制度の中で認められた技術的手法と，提供ポリシーを満たした情報を指すため，匿名加工情報⊆匿名化データ　である．

本研究における匿名化データは，匿名加工情報と認定される際の条件のひとつである「特定の個人を識別することができないようにする」措置について，実施されたデータを指すものである．

匿名化データは，パーソナルデータに対して技術的な処理を行って再識別リスクを低減させ，かつ有用性を保持したデータであるのに対し，匿名加工情報は，その技術要件と，データの提供先や，情報公開の状況を，個人情報保護委員会などの機関が総合的に検討した上で認定されたデータである．

2.7.2　欧州における匿名化処理技術の動向

欧州における匿名化データ(Anonymous data)は，EU データ保護規則によって「**データ管理者からも他のいかなる者からもその者が識別することができない，自然人に関する全ての情報であり，その判断はケースバイケースで行う**」と定義されている[37]．

また，欧州連合におけるデータ保護のアドバイザリー機関である第 29 条作業部会は「匿名化技術に関する意見書」にて，以下のように定義している．

1. **個人を選別する(single out)ことは可能か**
2. **個人に関する記録と紐付けることは可能か**
3. **個人を推定することは可能か**

これらの基準に合わせて匿名化技術を選定することが必要とされるが，「完全な技術は存在しないため，ひとつの手法に依存しないこと」を求めている[37]．

また，匿名加工の手法に応じてもそれぞれにリスクを定義しており，最も詳細な区分がされており，それぞれのデータ提供先の定義などがされている．その概念図について図 21 にて示す．

Pseudonymized data は仮名化データとして解釈され，墨塗りなどの簡単な処理しか行っていないデータを指す．

Key-coded data は，匿名化処理を行っているのではなく，データをそのまま利用するが，個人の ID が適切に暗号化されており，元情報とのデータベース上での結合が不可能になった状態のデータである．これは，属性値を変更してはいけない医療用研究データや公共データの分析の際などに主に用いられている[38]．

本研究における匿名化データは個人を選別する(single out)することが出来ない情報であり，EU の定義における Anonymous data に最も近いものである．これは日本のおける非識別・非特定情報もその概念とほぼ同じものと言える．

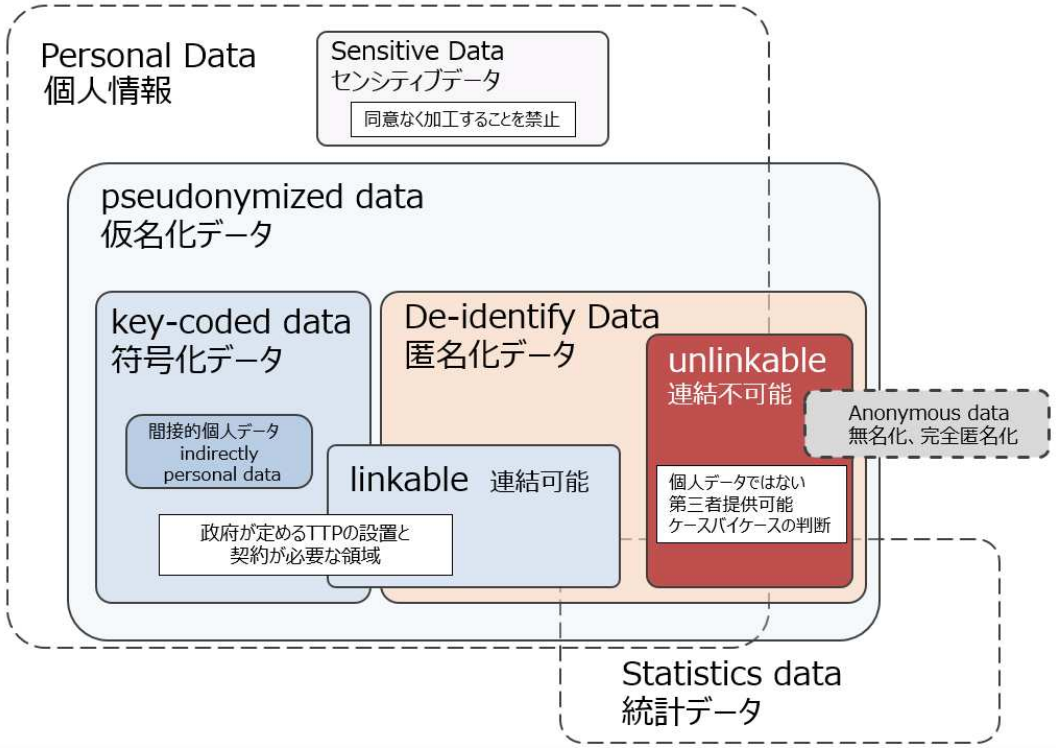


図 21 EU データ保護規則における匿名化データの種別

更に，個人を選別できないデータにおいて，Linkable と Unlinkable の区分がされている．Linkable とは，元データとの対照表（結合可能キー）をデータ作成者が保持していることを指している．一度外部に提供したデータについて，元データの対照表を保持している場合，元情報を保持する機関に所属している社員などが容易に元データに復元できることを示す．そのため，Unlinkable にすることで，復元されるリスクが大きく低減すると考えられている．

また，完全に Unlinkable にされることで Anonymous（無名）な状態にすることが理想とされるが，完全な匿名化処理は存在せず，また，完全に匿名化された情報は，統計情報や無価値なデータと限りなく近いもの，という区分がされている．

また，統計データを作成する際にも，その作成の途中段階で匿名化データが作成される場合もあり，それらの管理や統計の最終データの中に，個人への識別性が残るものが存在しないことも，同様に求められている．

2.7.3 米国における匿名化処理技術の動向

日本と EU が，匿名化データに対して明確な定義を行っているのに対し，米国における匿名化データ（De-identify data）の定義は曖昧にされていた．

また，2012 年に米国の連邦取引委員会(FTC: Federal Trade Commission)が発表した FTC3 要件によって，非識別化されたデータの悪用を契約で禁止することによってデータの保護を行った[39]．

しかし，この要件を設定した後においても，「合理的な措置」であると認定するための技術的基準や，「再識別」行為の範囲などを明確に定義することができず，EU 等との基準を統合することが難しかった．

そこで，2015 年に発行された「米国消費者プライバシー権法（ドラフト版）」において匿名化データの定義は「開示する情報について 1.合理的な根拠 2.再識別を抑制 3.契約と法的強制を持ち 4.公に告知したもの」とされ，EU 等の基準と同程度まで法的，技術的な用件が明確化された[40,41]．

米国ホワイトハウスが発行したレポート[43,44]には，「非識別化以上に再識別化が強力になっている，過度にデータ収集される現在の技術環境においては，パーソナルデータの収集と保持の管理に焦点を当てることは，重要ではあるものの，もはや個人のプライバシーを保護するために十分であるとはいえない」とされており，匿名化処理技術は，今後の再識別技術の進歩にあわせて常に進歩することが求められている．

2.7.4 匿名化データの安全性基準の課題

匿名化処理を行う際に適用される安全性基準は，国・地域や業種業態などの条件によって異なる．本章では，匿名化データの第三者提供に関する安全性基準の例として，カナダの小児医療研究所 CHEO における再識別リスクの基準について述べる．

CHEO における匿名化データ提供基準は，匿名化データを提供する際，提供先が信用できる組織（例として，公共団体や学術団体を挙げている）である場合，リスクが少ないと判断し，再識別確率は 0.33 まで許容される．これは k-匿名性で考えると 3-匿名性である．

逆に，公共利用を意図した，オープンデータに近いデータである場合は，それよりもリスクの閾値が高く設定され，0.05（20-匿名性）となる．これらの再識別リスク発生確率を定義することによって，各種の安全性指標との互換性が発生し，定量指標に展開が可能となる．

しかし，この場合の再識別リスクは，k-匿名化処理などの再識別率の減少処理以外にも，データへの攻撃者や関係者，関係機関などを対象に，その攻撃を行う動機と能力の有無とリスク低減コントロールの状態に応じて確率を算出し，最も高い確率を使用する．

例えば，一般的な医療機関にて情報漏洩が発生する確率は，政府発表の統計等から導き出した値として 0.27 と決まっており，それらのリスクを含めた，最も高い値をもってリスク発生確率と判断する．

リスク発生確率の閾値が明確化される事で，k-匿名化等の匿名化手法を選択することが容易になるが，この基準を全ての事業者や研究機関に適用することは難しい．

あらゆるパーソナルデータは，それぞれ独自の全体数，分布傾向や，隠すべき属性値を持つ．そのため，これらの安全性指標は，一律に設定するのではなく，データ提供者の置かれている状況と，利用者の要求に合わせて柔軟に設定される．

もし CHEO が採用している 0.33 という閾値について，妥当であると判断された場合，では，他の業種，例えばマーケティングデータの場合は，その n%が適正値なのか，などの定性的な比較基準も必要となり，事業者などでは妥当な安全性指標について主体的に決定することが難しい．

また，2.4 章にて述べたパーソナルデータ攻撃モデルに対する各種の攻撃方法が存在するため，状況に応じて必要な処理を行う必要がある．そのため，データ保持者が要件に対応した安全性指標を選択し，個別に評価を行う必要があり，これもデータ保持者の負担となっている．

2.7.5 匿名化処理プラットフォームの課題

匿名化データを流通させるシステムは，欧米では政府機関や医療機関において定着しており，オランダの μ-Argus [17]や，前章で述べたカナダの CHEO における匿名化データの提供[20,21]などが知られている．

日本においては，2010 年に経済産業省のプロジェクトとして「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」[30]が行われた．これは，前章にて述べた CHEO のシステムを日本の状況に合わせて運用したものである．

その結果として，永井らがデータ匿名化プラットフォームのアーキテクチャの提案[81]を行っている．その中で，匿名化処理は，一次事業者，委託第三者機関，第三者評価機関の 3 者が関係し，対象データが 1 種類の場合と複数種類の場合でそれぞれの役割が異なる点に着目し，マルチテナント型のアーキテクチャを実現した．

その課題を千田ら[82]がまとめており，匿名加工処理は，Datafly[69]やマージナル[77]等を用いた集合型匿名化アルゴリズムと，Pk-匿名性[47,48]などの撹乱・再構築型アルゴリズムでは手法が異なるため，それぞれ個別の評価指標が必要となり，一律に利用する事は難しいと報告している．

特に，集合型匿名化アルゴリズムでは，複数の属性値を抽象化することによって，次元の呪い[29]と呼ばれる有用性の低下が発生する．そのため，集合化や抽象化を行う匿名化手法と，確率撹乱手法を併用して再識別リスクを減少させる必要があるが，その際に結

果データの有用性と安全性を統合して評価する方式が定まっていないという問題点があった[82].

また, それ以外にも過去に提供したデータとの差分を検証されることで個人の特定リスクが増加する「逐次公開」のリスク, 個人情報を匿名化処理する際に, 中間データ等, 危険性の高いデータが含まれる可能性として「アウトソーシング」リスクなどの課題が挙げられており[30], それぞれの分野で研究が進められているが, 全てを同時に解決することは難しい.

また, 医療用のレセプト NDB(National Data base)を匿名化処理し, 契約した研究者にのみ提供する, レセプト 2 次利用実証事業[83]が開始している. その中で, 実データと k-匿名化処理された医療データによる研究結果の比較を行った結果, 集計単位等の変更を加えることによって情報損失を少なく抑えることが可能であることを報告しており, センシティブな属性を匿名化処理する手法が発表されている[84,85]. しかし, 医療用データは情報漏えい時の問題が大きく, 公開範囲や目的を設定することが難しいため, 継続して提供方法などの議論が続けられている.

2.8 従来研究のまとめと課題整理

本章では, 匿名化処理に関連する従来研究について分析した.

まず, 2.3 章と 2.4 章にて, k-匿名性をはじめとした, 安全性指標について述べた. そこで, 代表的なパーソナルデータに対する 4 種類の攻撃方法について概説した. それぞれの攻撃方法においても多様な課題が存在するため, 個別の安全性指標が提案されており, それら全ての指標を網羅することは困難である.

しかし, それらの指標には k-匿名性によって個人識別性を確率的に減少させた上で, 更に個別の用途に即した安全性を計測するという, k-匿名化処理後の追加処理として検証する指標が多く存在する. 例えば, レコード結合における Multi R k-anonymity や (X,Y)-anonymity, 属性結合における ℓ-多様性, t-近傍性及び(α, k)-匿名性, テーブル結合における δ-存在性, 確率的攻撃における(c,t)-Isolation や m-Invariance の指標は, k-匿名性の検証を行った上での追加的な安全性指標である.

そのため, これらの安全性指標を効率的に検証するには, ある属性に対して, 抽象化・集合化・削除処理等を行って k-匿名性を満たし, その処理結果が他の指標の基準も満たしているかについて, 追加検証する, という手順が必要となる. その追加検証の結果, k-匿名性を満たすが, 他の指標を満たさない場合も発生するため, その場合, より抽象的な値を探索する処理を繰り返す.

そのため, k-匿名性を満たす属性の組み合わせ探索を効率化することによって, 匿名化処理のプロセス全体の効率化が達成できる. そこで k-匿名性を達成するための処理手順について 2.5 章及び 2.6 章にて概説した.

まず，データ処理を行う際に用いる処理窓単位を設定することが求められる．出現した属性値を小規模の窓で処理する Local Recoding では，求める k-匿名性を実現することは容易だが，属性値に含まれる要素をコントロールすることが難しい．逆に，データ全体の出現率を用いて処理する Global Recoding では，属性値に含まれる要素をコントロールすることは容易だが，得られる k-匿名性を予測することが難しい．

また，値の書き換え処理を行う手法は多く存在するが，構造的に有用性を維持するための手法として一般化階層を用いて段階的に属性を抽象化する処理が行われる．その際に，局所的に情報を抽象化する VGH のみを用いて処理を行う場合と，同一階層毎の処理を行う DGH を用いて行う場合がある．VGH による局所的な抽象化処理を行う場合，情報量は多くなるが，等価クラスに含まれる属性値の制御が難しい．逆に DGH は等価クラス内に含まれる属性を制御できるが，その階層を用いた場合に実現される k-匿名性を予測することが難しい．

処理窓，及び一般化階層の適用において，局所的な処理と全体的な処理を組み合わせて匿名化処理が行われるが，高い有用性と k-匿名性の両立が共通の課題である．

そのような課題を解決するため，有用性と k-匿名性を両立するアルゴリズムも提案されている．しかし，有用性の基準として情報量，または情報損失量で計測するため，データ利用者にとって有用性が高い属性の組み合わせを適用し，かつ安全性基準を満たす処理の実現が困難である．利用目的や対象とするデータの特徴が多様であるため，アルゴリズムの特徴によっては非効率となる場合が発生する．

そのため，効率的な Lattice Structure の探索方法が求められているが，ボトムアップ型やトップダウン型の探索手法は，対象データ毎の分散や必要とされる k 値に依存し，処理の効率化が難しいという課題がある．

加えて，匿名化データの流通を実現するには，匿名化処理技術に関する課題だけでなく，国や地域によって定められる安全性基準に対応する必要がある．そこで，2.7 章では，匿名化データに関する，日本，欧州，米国の基準の違いについて述べた．しかし，それらの法律等では，パーソナルデータの安全性基準を明確に定めず，匿名化データの提供先の利用目的と契約内容に沿った処理を行うことが求められている．そこで，プラットフォームに関する課題を検討するため，カナダの CHEO における安全性基準について概説し，そのシステムをベースに行われた実証実験における課題を分析した．それによって，安全性と有用性に関する指標と基準が多様に存在しており，汎用的に用いる基準が存在しないことが問題提起されている．

従来研究から，匿名化データの流通に関する課題として，以下の 3 点を抽出した．

1） 匿名化データの安全性基準の策定

2） パーソナルデータの匿名化処理リソースの軽減

3） 実社会に即した匿名化データの流通方法

これらの課題の解決に向け，有用性の高い一般化階層を適用した際に k-匿名性が達成できるかについて，予測式を用いて導く手法によって効率化する処理を検討する．

従来研究では，データ利用者にとって有用性の高い属性組み合わせ処理を要求した場合，提供者側の規則等で求められている k-匿名性等の安全性基準を達成できるかを予測できないため，データ提供者は負荷の高い匿名化処理を繰り返し実施する．また，データ利用者においても一般化階層の改良が発生し，非効率的である．

そのため，匿名化データの授受を行うプラットフォームとして実装した際に，匿名化データの提供者と利用者における，安全性と有用性の要望の不一致問題が発生する．

それぞれのパーソナルデータは個別の分散や特徴を持つため，一律のアルゴリズムを適用するのではなく，データの性質に応じた予測値や，アルゴリズムを適用することによって，属性組み合わせの探索処理の軽減や，データ利用者による事前検証などに活用が可能である．

次章より，予測式を導くための基礎データとして，多様な分散を持つ実データに即した擬似データを用いて，k-匿名性の推移を検証し，予測式の提案につなげる．

3 k-匿名性減少特性の検討

3.1 はじめに

本章では，まず **1）匿名化データの安全性基準の策定** の課題に対応するため，基礎データとして，通常のパーソナルデータを扱う中で，最も基本的な安全性の概念である k-匿名性がどのように変化するかを検証する．

本来，k-匿名性における k 値とは，データ保持者のリスク許容度から設定されるが，本研究では，ある一般化階層を適用した際の等価クラスサイズの最小値と定義する．

匿名化処理における課題は，安全性の基準がそれぞれの保持するデータ量やその分布の特徴によって異なり，一律に適用できる基準値が存在しない，という点である．

また，この基準値を検討するためには，企業が持つパーソナルデータの分布状態に対して，どのような匿名化処理を施したことによって，どのレベルの安全性が実現できるのか，というサンプル評価データセットが大量に必要となる．

しかし，事業者から実データを集めることはプライバシーポリシー等で用途が設定されており，目的外の利用である匿名化処理の評価に用いることは現実的には難しい．また，それを公開することによって発生するパーソナルデータの侵害行為の可能性を排除できない．

本章では，実サービスに登録しているユーザの属性分布を参考に作成した擬似パーソナルデータに対して，一律に一般化階層を適用し，その k-匿名性の推移を計測，そこから導かれる k-匿名性の減少特性を検討することで，k-匿名化処理における安全性基準の達成可能性について検討する．

3.2 k-匿名性における k 値の定義

まず，k-匿名性で利用される k の値が，どのような数値的な特徴を持つかの定義を行う．ある個人情報全体のレコード数を S とし，それらの情報を k-匿名化処理した際の最小値を k とする．

まず，準識別子が 1 種類しか存在しない場合（k=S, k/S＝1)があるが，その場合は属性値に意味がないため，ここでは定義しない．それ以外の場合，選択肢の分類は 2 種類以上存在するため，S/2 が実質的な最大値となり，かつ S は1以上存在する．(1≦k＜S/2, k/S＜0.5）その中で，2 を最低ラインとして識別確率を 1/k まで低下させる処理が k-匿名化処理である．これにより，k-匿名性を維持する最小値は(5)で定義できる．

$$\left(\frac{S}{2}\right) > k \quad \therefore S \geq 2k+1 \quad \cdots(5)$$

3.3　実験概要

実験は，実サービスに登録しているユーザの属性分布を参考に作成した 1635 サービス，4,344,922 人分の擬似パーソナルデータを用いて行われた．図 22 にてサービス群の顧客数と，元データにおける標準偏差を示す．

また，本データを入手できない場合にも比較検証を可能とするため，2012 年の国勢調査を顧客データと同様に区分し，サービスデータの一つとして設定した．その際に，擬似パーソナルデータのスケールが 1～350527 人であることから，データスケールを合わせるため，全体の値を 1/1000 として処理した．これにより，本研究で実施した内容を検証することが可能となる．

本研究で用いられたデータによる，主要な k-匿名性の減少傾向は，巻末資料として掲載している．より詳細な資料について必要な場合は，筆者に問い合わせを頂きたい．

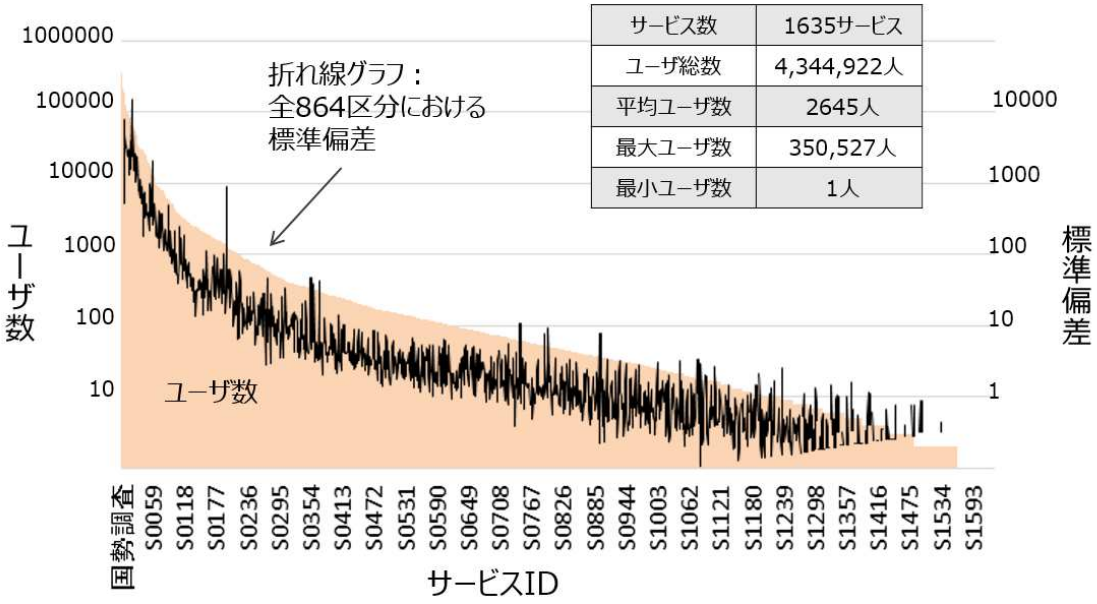


図 22 対象サービス ID の顧客数と標準偏差

対象となる属性は，性別・年齢・都道府県であり，それらの顧客の属性を一律に分類し，分類の方法によって k 値がどのように変化するかを検証した．

サービスの区分は，サービスの内容によって区分するのではなく，k 値の特性を確認するため人数による階級を作成した．個人情報の分類方法として，表 6 の一般化階層を作成し，それぞれの値での分類を行った．

これらのサービス群の顧客全体に対して，上記の区分を実施し，それぞれの組み合わせ(性別 2 区分×年代 3 区分等)も含めて実施した．

実施総パターンは　単体[1+3+3＝7 種], 2 種組み合わせ[(1*6+3*4+3*4)/2)=15 種], 3 種組み合わせ[1*3*3=9 種]の合計 31 種類の情報区分を行い, サービスのユーザ総数と区分数の関係性などを検証した. 表 7 にて各顧客区分の種類を示す.

表 6 調査対象サービスの人数による階級区分

登録人数	サービス数	人数	平均ユーザ数
100001人以上	10	1,669,482	166,948
50001～100000人	16	1,147,872	71,742
10001～50000人	36	870,965	24,193
5001～10000人	34	241,927	7,116
1001～5000人	124	266,613	2,150
1000人以下	1415	148,063	105
合計	1635	4,344,922	2,657

表 7 顧客群に対する区分方式の種類

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	西日本	東日本							
	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	岩手	・・・	鹿児島	沖縄			

3.4 区分数と k 値の関係性

まず, 各区分数と k 値の推移を検証した. 縦軸は平均 k 値, 横軸は区分数となっている. 区分数とは, その属性が持つ選択肢の区分数を合計したものである. 例えば[性別＝2 区分], [都道府県＝47 区分], [男女 2 区分*年代 9 区分*都道府県 47 区分の組合せ＝846 区分(グラフ内の最端値)]となる. 図 23 にてその推移を示す.

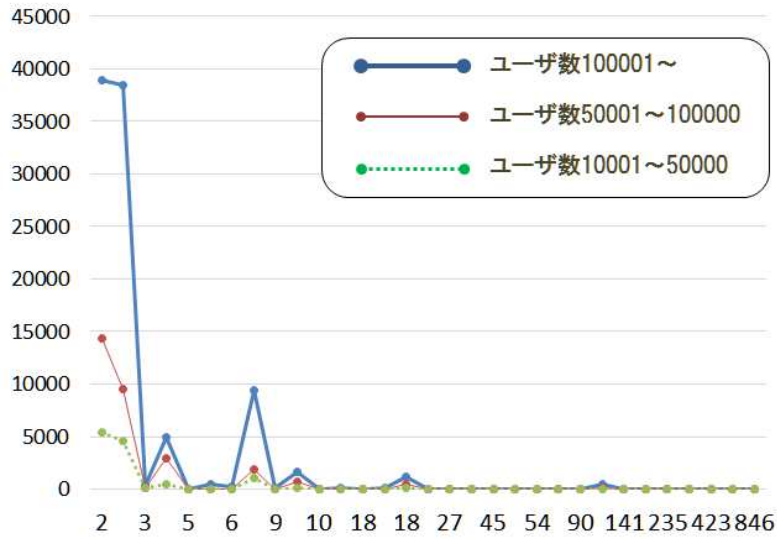


図 23 属性の区分数と k 値の推移

区分数が 2 までは非常に大きな値が出力されるが，5～9 区分になると極端に値が減少し，そこから同じような数値が続く．特徴が解り難いため，k 値という意味は薄れるが，対数表にしたものが図 24 となる．

これを見ると，値に対する大小はあるが，区分数が多くなるほど k 値が小さくなる傾向が見られるが，これは想定通りである．しかし，10 万人超のサービスと 1～5 万人のサービスでも，10 区分以下になると大きく値が変化しないことは想定外な結果である．

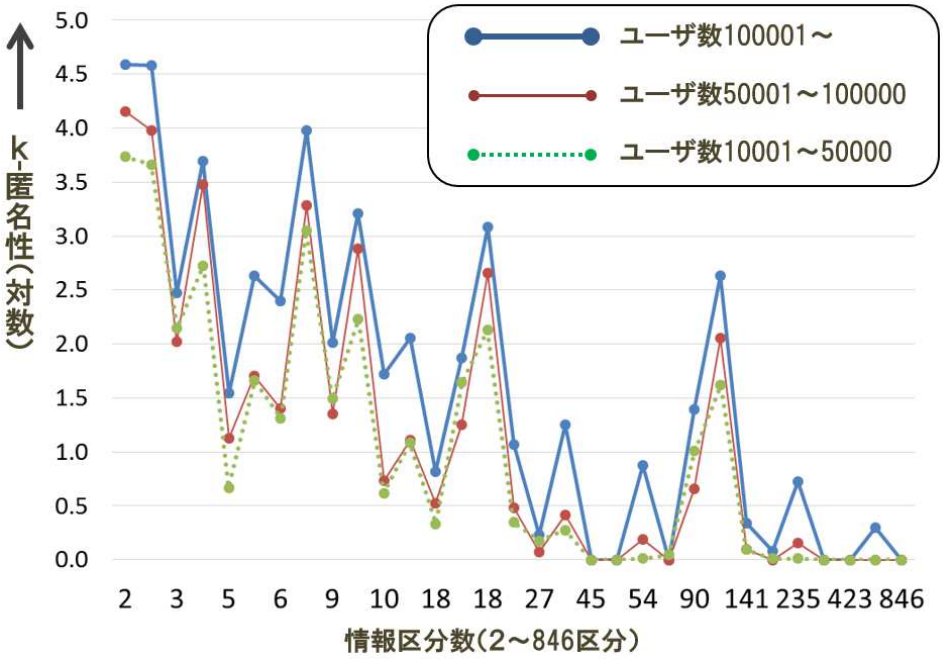


図 24 属性の区分数と k 値の推移

各増減の状況を見ると特定の区分の際に数値が増減している傾向が見えるため，数学的な区分の大小よりも，区分の方法による増減の差の方が大きいと考えられる．グラフの中で急激に上がっている箇所は，多くが“地域”の属性が含まれている箇所であることが判明した．

“地域”は 47 区分を行った場合でも万遍なく存在し，標準偏差が小さいことが判明している．図 25 は，標準偏差と属性区分数の比較を行った結果である．その場合，相関係数 0.74 と比較的高い数値が出た．標準偏差と k 値の増減には一定の関係性があると考えられる．

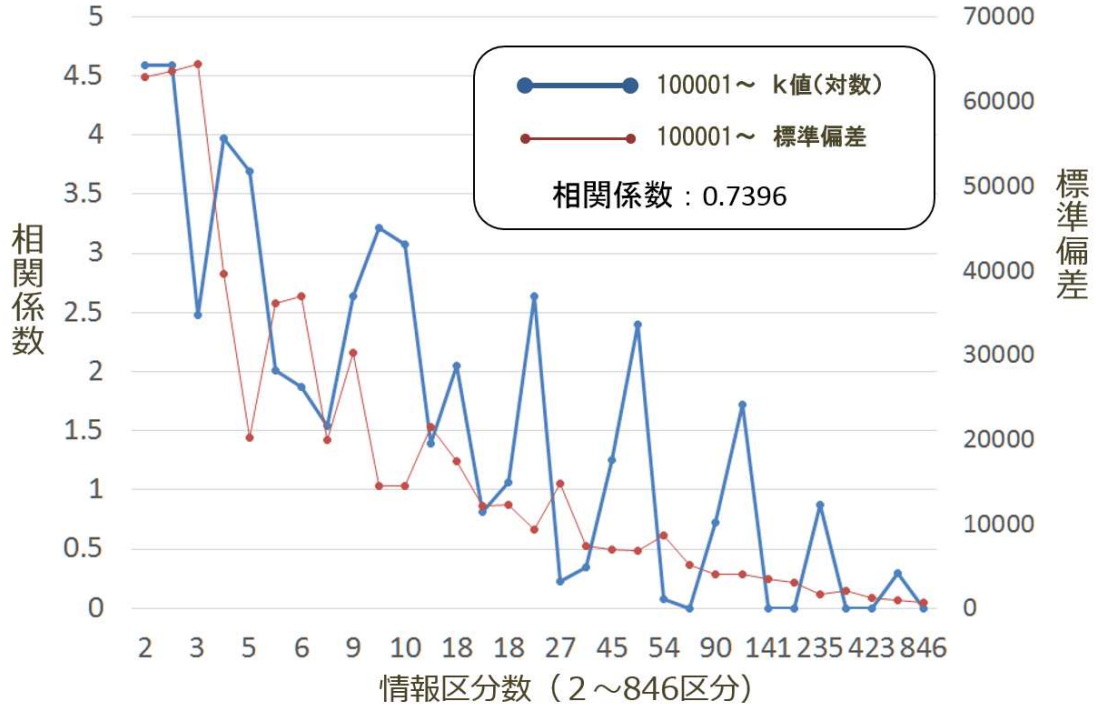


図 25 属性の区分数(対数)と標準偏差平均の比較

また，各サービス群を対象に，k値の減少値を計測し，近似直線を引いたところ，10 万人以上のサービスにおけるk値の平均減少値は 2938.9 であり，平均減少率は 1.76%であった．図 26 と表 8 にて線形近似式にて予測式を作成した場合の結果を示す．

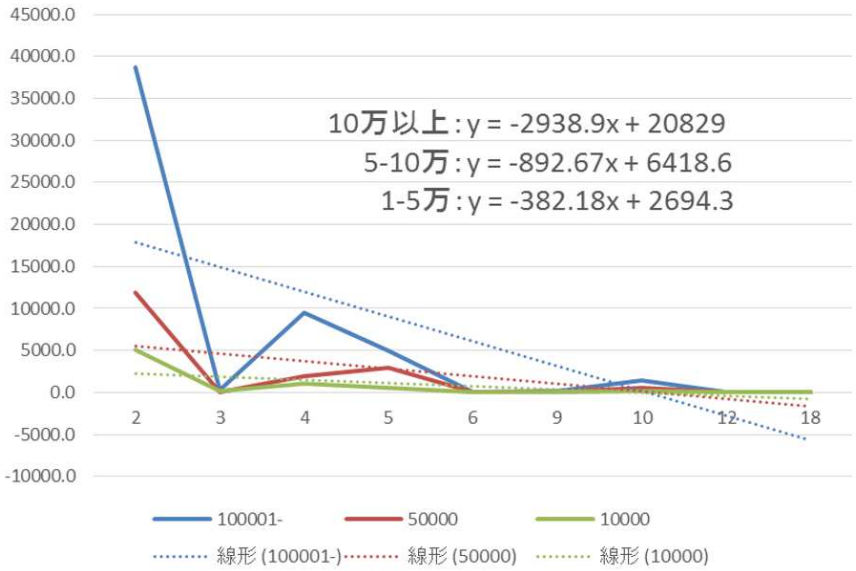


図 26 k 値(実数)の線形近似値の比較

表 8 k 値の平均減少数と平均減少率

サービス分類	サービス数	平均人数	平均減少k値	平均減少率
100001人以上	10	166,948	−2938.9	−1.76%
50001～100000人	16	71,742	−892.7	−1.24%
10001～50000人	36	24,193	−382.2	−1.58%

3.5 k 値と区分方式の関係性

前項の実験は，属性を区分する数に対しての関係性を確認したが，本項は属性値を一般化階層で抽象化した際に変化する値について検証した．

情報を抽象化することで k 値がどのように変化するかについて性別と年代についての属性値の組み合わせによる変化を確認した．

18 分類は，元となった情報と同じものであり，10 分類は，情報の両端値をすそ切りして得た分類である．

図 27 は，18 分類における各階級の平均 k 値と，k=1 となったサービス数の推移である．18 分類はデータの詳細さが一番高い群である．これは，サービスの要件定義を行う際にある程度不必要と考えられる階層も確認できるよう，保証のために入れられた区分が存在していると考える．

また，サービスの中には，ある年代が全く存在しない群も含まれるため，k 値が低くなるようになっている(例えば男性向けのサービスのため，女性は少ない等)．しかし，この 18 分類を行った場合の 10 万人超のユーザを持つサービス群での k 値は 6.6 と非常に小さな値となった．

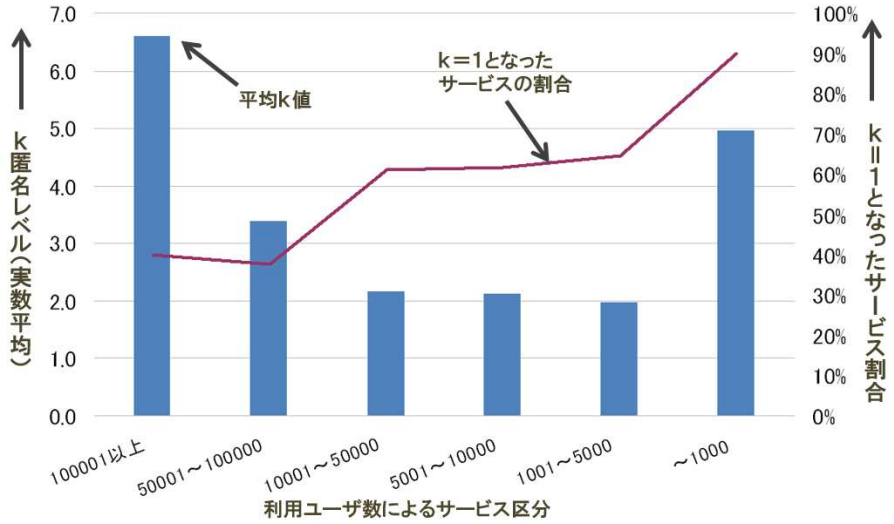


図 27 18 分類における k 値平均と，k=1 サービス率

また，10 万人超のサービスの場合でも 40%のサービスで k=1(匿名化できない)状態となった．1000 人以下のサービスにおいては 90%以上のサービスが k=1 となった．つまり通常のサービスで利用される顧客区分をそのまま用いた場合，k 値を高く維持することは難しいことが解る．

同様の分析を 6 分類/10 分類で行った結果が図 28 と図 29 となる．6 分類は年代情報をほぼ最大限に抽象化し，簡単な分析にしか利用できない形としたものであるため k 値は 18 分類と比較すると大きいが，10 区分よりも k の値は低い．k 値平均は分類数の多い 10 分類の方が，6 分類の 10 倍以上となる．安全性という面においても，情報の詳細さという面においても 10 分類は 6 分類よりも優れていると考えられる．

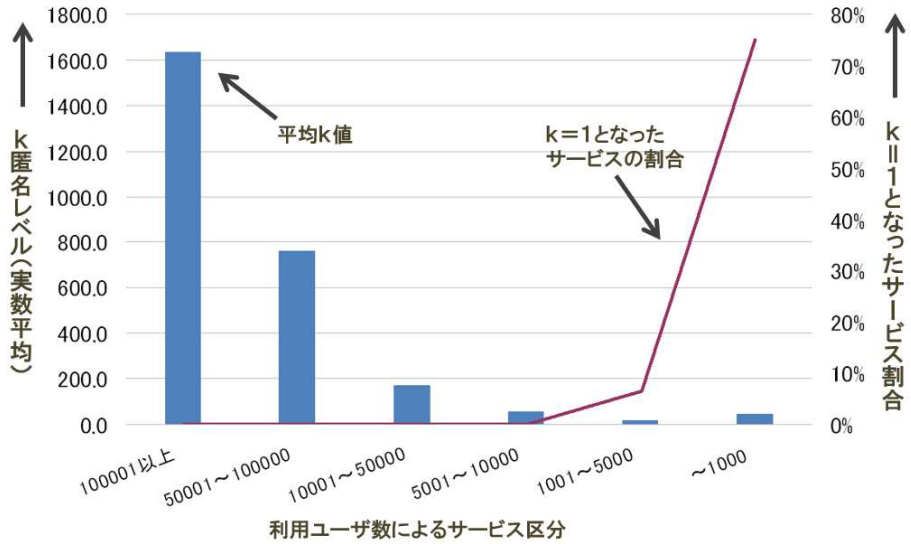


図 28 10 分類における k 値平均と，k=1 サービス率

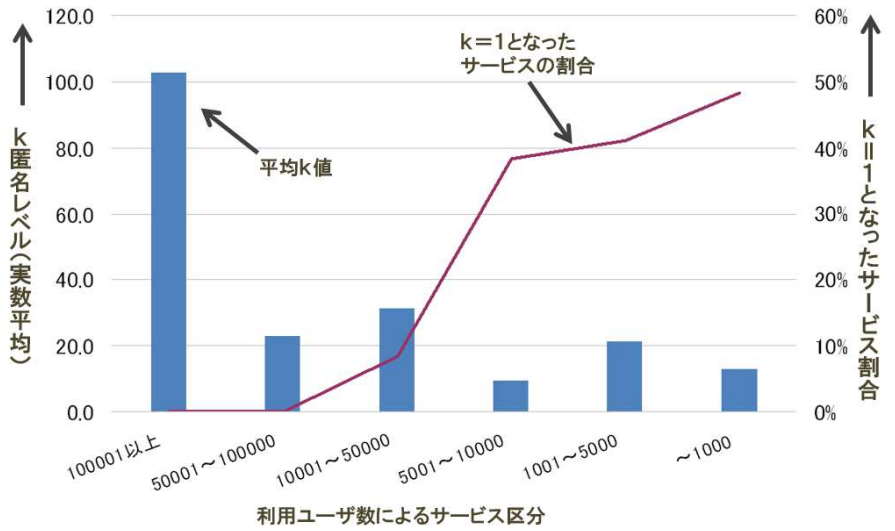


図 29 6 分類における k 値平均と，k=1 サービス率

10 分類の特徴として，1001-5000 ユーザのサービスまでの全てのサービスで k≧2 が維持されており，非常にユーザ数の少ないサービスにおいても匿名状態が維持できていることが解る．

通常の技術者であった場合，この情報は情報の詳細性も維持されている上に安全性も格段に高い．そのためにこの方針で匿名化処理を進めると考えられる．この 10 区分の問題点は，k 値を高めるために”**20 代以下**”という属性を使用している点にある．サービス登録者として少数である 10 代を排除するために行った措置のため，安全性の観点から作成した区分であり，利用性の減少については配慮されていない．

3.6　国勢調査の相関とk値の分布調査

今回対象とした情報群の中で，地域分布については日本の国勢調査(2010 年)との相関関係が比較的高いことが判明した．そこで，基準値として日本国内の人口分布を用いて，その値との相関係数を計測．相関関係とk値に関連性があるかについて検証した．

○国勢調査の相関係数とk値の検証対象

-地域 9 分類:(北海道,東北,関東,中部,近畿,中国,四国,九州,沖縄) 国勢調査との相関 0.934

-地域 47 分類:(北海道,青森県,秋田県. . . 鹿児島県,沖縄県)
国勢調査との相関 0.914

図 30 に対象の 2 種の k 値を国勢調査との相関係数の比較結果を示す．相関係数 0.93 までのサービス群は k 値が 200 以上と高い値で推移したのに対し，それ以下の相関係数になるとk値が多く1で推移する．これはある程度サンプルを多く入れ込んだ場合でも傾向は同じである．

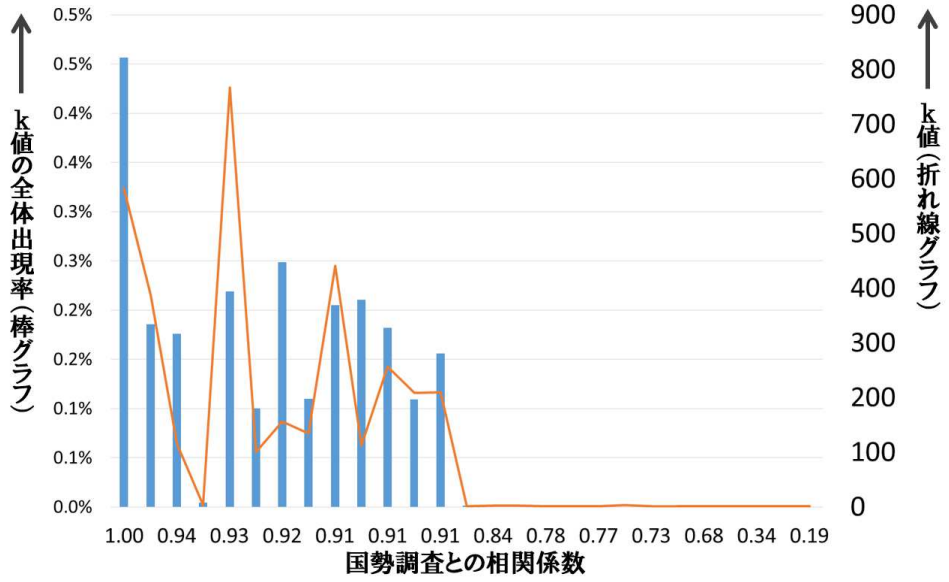


図 30 地域 47 分類のk値と国勢調査の相関係数:50001 人以上のサービス

また，図 31 にて 50001 人以上の大規模サービスでの数字ではなく，上位 100 サービスでの分布状態を検証した．相関係数が 0．9 以上のサービスに k 値が高いものが集中しており，相関係数が低いものについては殆どk値を維持することができなくなっている．

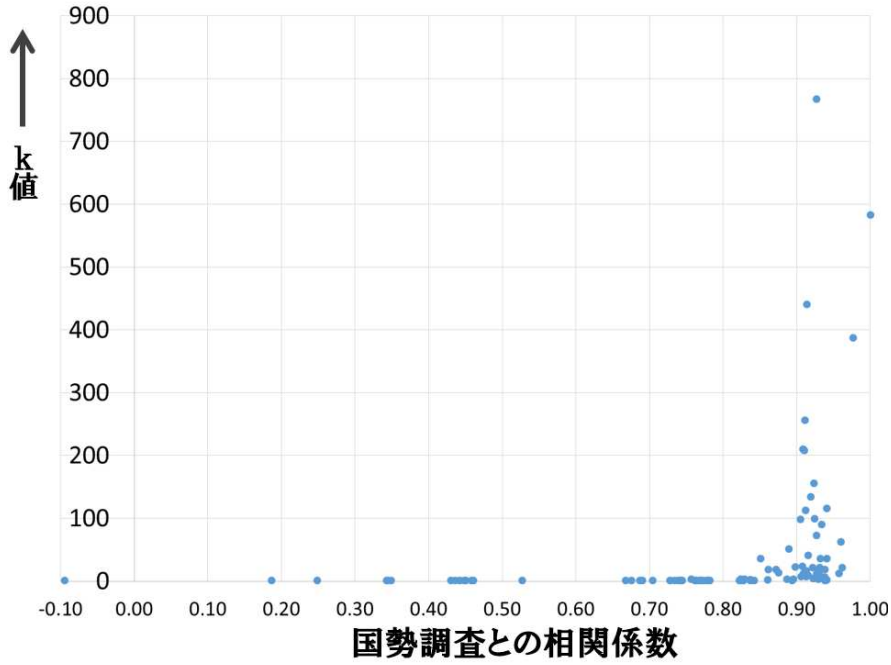


図 31 地域 47 分類の k 値と国勢調査との相関係数:上位 100 サービス

上記 2 つの結果は，元が同じ顧客群から生成されている情報であるため，似た数値となることはある程度想定できる．本情報の中には”東日本”のみで展開されているサービスなど国勢調査と反する結果を持つものが多く存在する．

匿名化するための区分を決定するアプローチとしては，国勢調査やオープンデータと情報の区分の方法を同一にして相関係数を取得することで，情報を初期の段階でスクリーニングすることが可能であると考えられる．

3.7 本章のまとめ

本章における実験で判明した知見をまとめる．

1) 本データに対して，年代/地域などの一律的な区分を用いた場合，次元の呪いによってk値は急速に低下する．10 万人超のサービスであっても 18 区分を行うと平均 k 値は 6.6 となった．
2) データを区分する中で，区分数が1増加するごとに減少する平均k値を求めると 10 万人以上のサービスの場合 2938.9 であり，全体の 1.76%であった．この定数を用いることで匿名化可能な区分数と元情報の規模を類推することを試みたが，実測値と予測値の相関係数が低く，予測式として不十分である．
3) 国勢調査との相関係数が高い群(相関係数 0.91 以上)はk-匿名化処理結果が高くなるが，相関係数が低い群は,一律の処理を行った場合に，多くがk＝1 となり，有為な匿名化処理が達成できない可能性が高くなる．

4) 出現率の低い情報のすそ切りなどの処理を行うことで k 値を高く維持することができる. しかし, 出現率だけに着目した分類では利用性が低くなる場合がある.

これらの結果より, 例えば 10 万人超の顧客を持つサービスであっても, 元データに近い形でデータ区分を行った場合, k-匿名化は達成できない場合が多く存在するが, データの分散状況に合わせた抽象化のための一般化階層を作成することで, k-匿名性を維持することは可能である.

また, その k-匿名性の減少傾向については, 属性値の区分数によって, ある程度定式化可能であると考える.

次章では, k-匿名性の減少傾向予測式について検討する.

4 k-匿名性の予測近似式

4.1 はじめに

前章の検証によって，匿名化処理を行う場合，データの分布状況に即した形で一般化階層を生成することで，有用性が高く，安全性を一定値以上に保つことが出来ることが確認できた．しかし，安全性と有用性の検証は処理コストが高く，かつ複数回発生する可能性がある．

k-匿名化処理は，属性情報同士を組み合わせた際の最小出現数(k 値)が，求める基準以上に存在するかを検証し，もし組み合わせた属性情報が k-匿名性を満たさない場合，属性値をより粗い粒度のデータに抽象化し，属性値の書き換えを行うことで，安全性を高める処理を行う．そのため，安全性が達成できない場合，最も抽象化されたデータまで全ての組み合わせにおける匿名性を検証する必要がある．

その一方で，データ保持者に対して，ある匿名化データの提供要求があった場合，その条件に沿って処理されたデータが，k-匿名化状態を満たすかも予測できないという問題もある．安全性を検証して提供した結果データが，利用者のデータ利用目的に合致しない場合は，再度属性値を変更するなど，処理条件を変更して匿名化処理を行う．このようなデータ提供者の負担を軽減する仕組みが必要とされている．

そこで本章ではこれらの問題を解決するため，k-匿名化状態を満たす情報を効率的に導くため，k-匿名性の予測式を検討し，予測式によって不要な安全性検定を排除する方法を検討し，3 章で作成した擬似パーソナルデータを用いて評価した．

4.2 予測式の検討

まず，k-匿名性の性質について検討する．

本研究は，公共データなどの大きなデータ群から，特定の条件に合致した群を抽出し，その情報に対して一般化階層を適用して匿名化処理を行う方式を想定する．

通常，複数の抽出条件によってデータの出現量を予測する場合は，多次元正規分布の同時密度関数によって出現率を予測する．しかし，その際には，全組み合わせによる分散と相関係数を導く必要があるため，各一般化階層における属性値の出現数検証を，Lattice Structure における組み合わせ回数だけ行い，同時密度を計算する必要がある．これは匿名化処理以上のコストがかかるため，実用的でない．

逆に，出力された結果から考えると，抽出条件によって，分析に耐えうる充分なデータ量が出力される場合，出力データを基準化すると中心極限定理によって

標準正規分布に近似できる．その分布の特徴から k 値を予測する方が効率的である．

出力されたデータが正規分布である場合，k 値は，データの全体数 P に対して，属性区分数 x 個のクラスタ化を行った場合のクラスタサイズの最小値として定義でき，その場合の k 値は平均値から最も遠い位置にある．

まず，各属性の区分によって生成されたクラスタの出現確率が，標準正規分布であった場合の k 値を検討する．

全体数 P に対して，属性区分数 x でクラスタ化した際の結果分布が常に標準正規分布であると設定する．そのときのクラスタサイズの最小値 $k_{(x)}$は，標準正規分布の最端値として存在し，平均値から最端値までの距離を $a\sigma_{(x)}$と設定すると，$k_{(x)}$は$P/x \ *\left(1-\Phi_{a(x)}\right)$と定義する．

全体数 P が変化せずに区分数が x+n に増加した場合のクラスタサイズの最小値 $k_{(x+n)}$は，$P/(x+n) \ *\left(1-\Phi_{a(x+n)}\right)$となる．図 32 にてその概念図を示す．

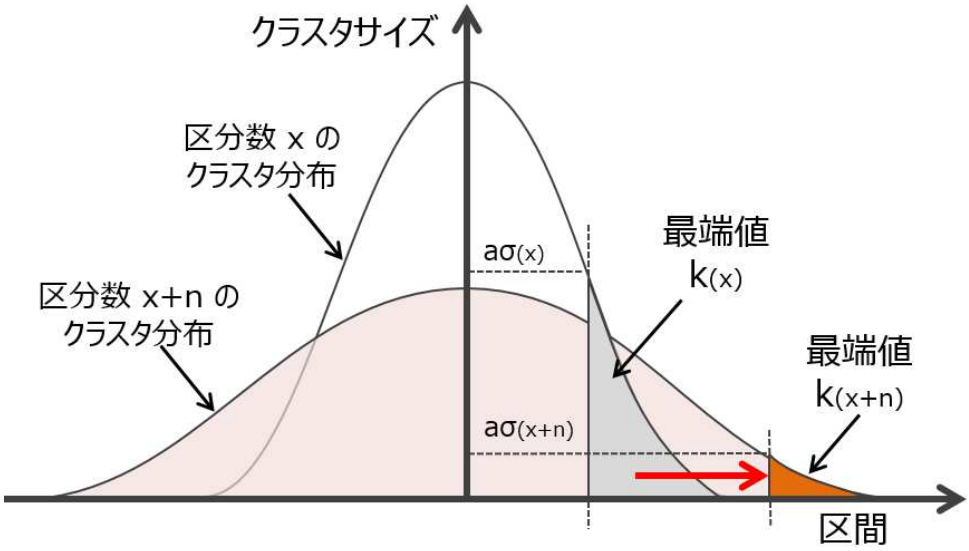


図 32 区分数の増加と最端値の変化

この $k_{(x)}$ と $k_{(x+n)}$の関係性を検証するため，区分数が均一に増加する正規分布によるクラスタサイズの最小値の推移について実験を行った．

図 33 は分散が 1 である正規分布で 10 万サンプルを生成し，最端を 4σ とした場合のクラスタサイズの最小値(k 値)と設定し，属性区分数を増加させ，各区分数で 50 回試行した際の平均 k 値の推移である．x 軸が属性区分数，y 軸が k 値である．これによって，平均 k 値は，$48770x^{-3.21}$ で $R^2=0.9612$ で近似することが判明した．

実際には，それぞれのデータの傾向によって異なる分散の特徴を持つため，実データにおける k 値は$\mathbf{P} * \mathbf{x^{-n}}$のような，べき乗型で漸減する傾向があると仮説を設定する．

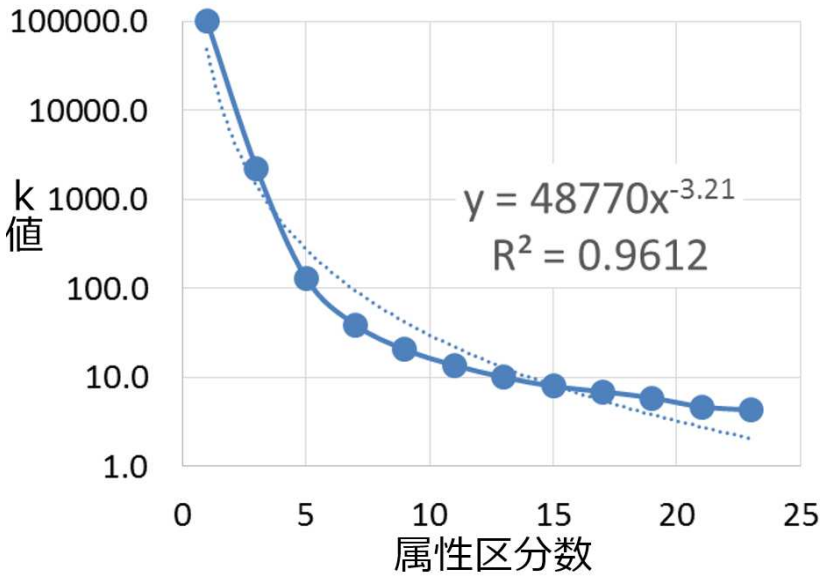


図 33 正規分布によるk値推移の実験結果

そこで，少量のサンプルを用いた匿名化処理の結果を用いて，属性の区分数からその後の k 値を導くため，累乗近似型の予測式を提案する．

一般的なデータの匿名化処理を想定した場合，匿名化処理結果が1つも存在しないことは稀であり，通常は少量の匿名化の結果を保持していると考えられる．

それらの属性区分数を既知の x，その区分数における最小クラスタのサイズ（k 値）を既知の y と設定し，累乗近似式によって，その後の区分数における k 値を予測する．累乗近似式は，既知の x，y によって最小二乗法における切片 α と傾き β を求め，(6)に代入する形で求める．

$$\alpha = \bar{y} - \beta\bar{x} \cdots (6a)$$

$$\beta = \ln\{ \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \} \cdots (6b)$$

$$k = \alpha x^{\beta} + 1 \cdots (6)$$

(6)では累乗近似式の結果に 1 を加えている．これは負の β 値を持つ累乗近似値は k=1 に収束するため，オフセットを 1 とした．

また，予測式の利用法としては，定められた k 値を満たす区分数 x を予測する場合もある．その場合は(6)に求める k 値を代入する形となり，(7)で表す．

$$x^{\beta} = (k-1)/\alpha \cdots (7a)$$

$$x = \frac{(k-1)^{1/\beta}}{\alpha^{1/\beta}} \cdots (7)$$

本予測式が成立する条件をまとめる.

1. 対象データが正規分布である.

2. 一般化階層の区分数と分散の増加量に規則性がある.

3. 既知の x,y として設定可能な, 属性区分数と k 値の実測値が存在する.

このため, 3 章で利用した, 国勢調査と擬似パーソナルデータを用いて, 区分数の異なる複数の一般化階層を適用し, k 値の推移を計測. その結果と予測式が出力する結果とを比較する実験を行った.

4.3 累乗近似式の比較と検証

実験は, 実サービスに登録しているユーザの属性分布を参考に作成した 1635 サービス, 4,344,922 人分の擬似パーソナルデータを用いて行われた.

また, 本データを入手できない場合にも比較検証を可能とするため, 2012 年の国勢調査を顧客データと同様に区分し, サービスデータの一つとして設定した. その際に, 擬似パーソナルデータのスケールが 1～350527 人であることから, データスケールを合わせるため, 全体の値を 1/1000 として処理した. これにより, 本研究で実施した内容を検証することが可能となる.

表 9 対象ユーザの階級ごとの状況

階級	登録人数	サービス数	人数	平均ユーザ数
1	100001人以上	10	1,669,482	166,948
2	50001～100000人	16	1,147,872	71,742
3	10001～50000人	36	870,965	24,193
4	5001～10000人	34	241,927	7,116
5	1001～5000人	124	266,613	2,150
6	1000人以下	1415	148,063	105
	合計	1635	4,344,922	2,657

通常, 企業の持つサービス会員データは多様な分布パターンがあるが, そのサービスがターゲットとしている顧客層を中心とした正規分布の性質を持つと考えられる. 本対象データには, 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まれている.

これらのサービス毎の会員データに対し, 登録人数によって階級を作成して表 9 に分類する. また, 本資料の追加資料として, 5 万人以上のサービスにおける k-匿名性の推移, 及び, 今回の検証で得られた結果の回帰係数等を記載してある.

国勢調査は, 分布が擬似パーソナルデータに比して均一に近いが, 過去における「団塊世代」などは出生数が多く, 現代に近づくと減少する傾向を持つ. 結果

として 10 年単位で出現数をクラスタリングすると，分散の小さい正規分布が成立する．属性区分数を大きくしても高い k-匿名性を維持する特徴を持つ．

性別(2 区分)，年代(3,5,9 区分)，地域(2,9,47 区分)の 3 属性を組み合わせて 2 区分から 846 区分まで設定し，各区分における k-匿名性を計測した結果を実測値として使用する．その実測値に対して，前章で検討した累乗近似式が近似することを，実データを用いて検証する．

まず，国勢調査を用いて，仮説設定した累乗近似式を作成し，値の一致率を検証する．

属性区分数を既知の x，一般化階層を適用して 5 区分まで組み合わせた場合の k 値を既知の y と設定し，累乗近似式を作成し，実測値と比較したものが図 34 である．重相関係数が 0.998 と非常に高いが，標準誤差が 1790.88 と大きな値になる．

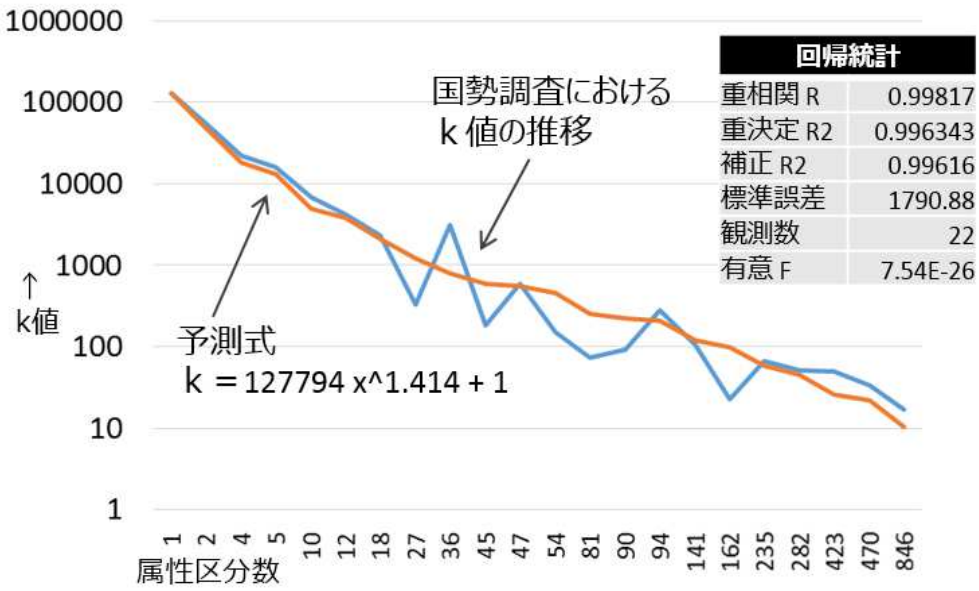


図 34 国勢調査の k-匿名性の実測値と予測値の比較

4.3.1 必要サンプル数の検証

k 値のサンプルをいくつか取得することによって，予測が可能であることが確認できたが，k 値を多量に算出する計算は非常に負荷が高い．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 35 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 値の推移に対してどのような相関係数となるかを評価した結果である．

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．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であるといえる．

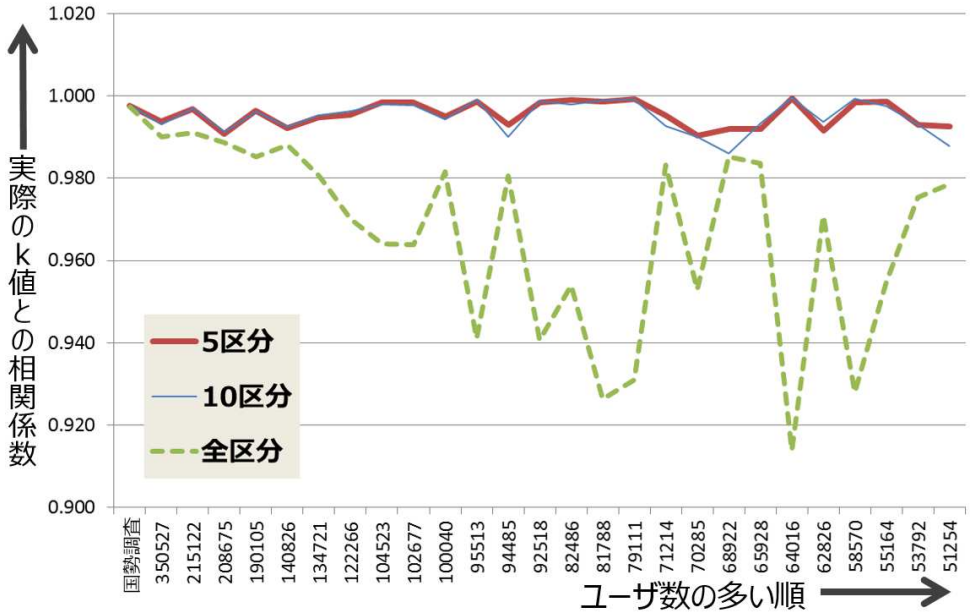


図 35 作成される近似式の相関係数と区分数の関係

この結果を受け，累乗近似式を用いて k 値の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．

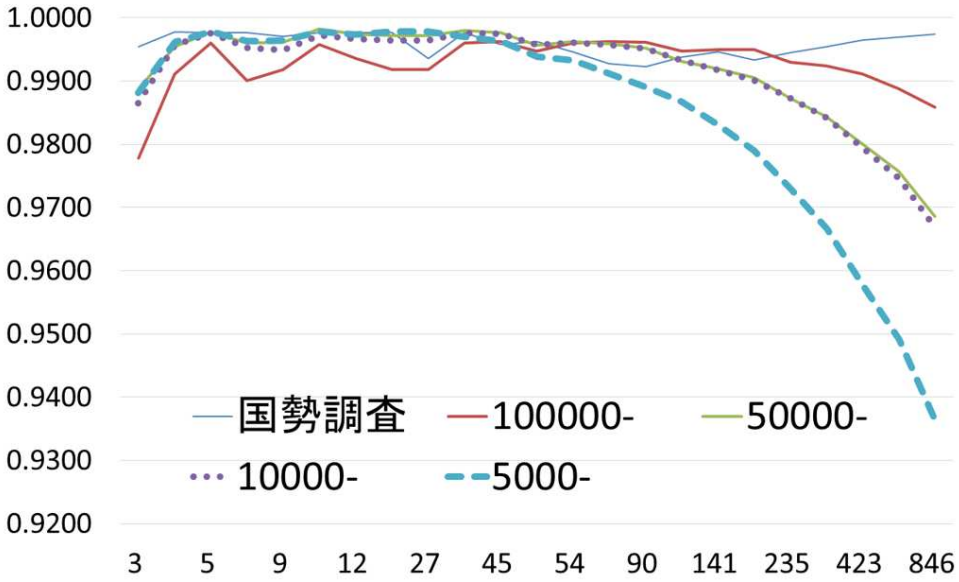


図 36 既知の x として使用した区分数と相関係数の推移

図 36 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，少ない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をまとめたのが表 10 である．

表 10 相関係数の最高値と最低値を出した区分数の検証

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，抽象度の高い 4～45 区分程度をサンプリングして匿名化処理を実施し，そこから得られた値によって累乗近似式を作成することで相関係数の高い予測式を作ることができると考える．

このような近似式によって，ある程度の k 値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 値をサンプリングし，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．

4.3.2 予測誤差の計測

ここまでの結果を受け，他のデータ群においても累乗近似式を適用した．表 11 は，対象データ群に対して，一律に一般化階層を適用し，9 区分までを既知の x，y として利用して作成した累乗近似式の回帰分析結果である．

また，所属人数が大きいサービスにおける，絶対誤差を計測したものが図 37 である．

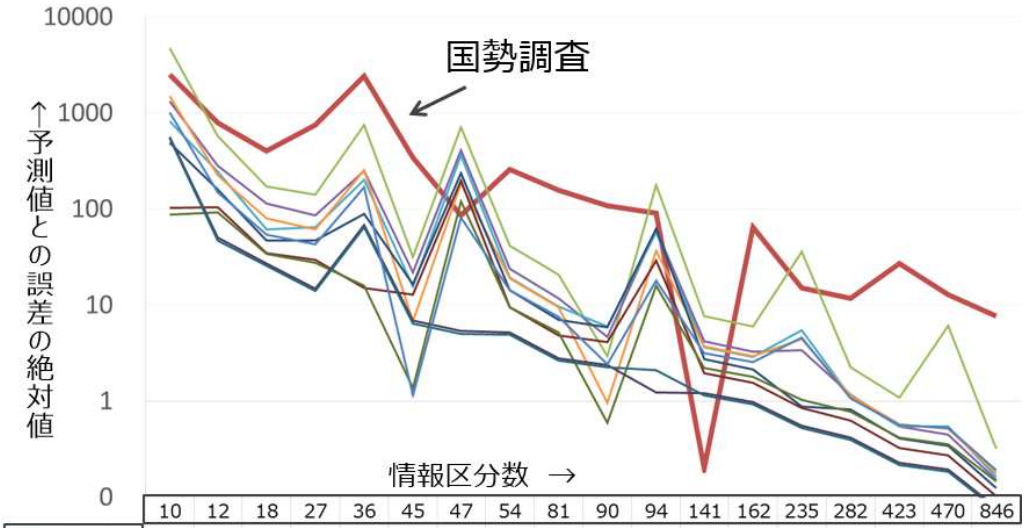
結果は，サービス人数規模ごとにばらつきはあるが，全ての群において，重相関係数が 0.99 を超え，自由度修正後の決定係数においても 0.97 以上の関係を持つ．また有意 F は 1.2E-16 以下と低い値となっており，回帰式として当てはまりが良いことを示している．

表 11 サービス人数毎の回帰分析結果(人数規模比較)

サービス人数規模	国勢調査	10万人～	5～10万人	1～5万人	5千～1万
対象サービス数	1	10	16	36	34
重相関R 平均	0.99705	0.99104	0.99143	0.99000	0.99047
重相関R 標準偏差	-	0.00344	0.00777	0.00870	0.00927
重決定R2 平均	0.99410	0.98217	0.98299	0.98018	0.98111
自由度修正決定係数 平均	0.99384	0.98139	0.98225	0.97932	0.98029
標準誤差 平均	2238.56	4267.16	1903.85	651.01	145.41
有意F 平均	3.8.E-27	4.1.E-20	1.7.E-17	2.5.E-16	7.3.E-17
平均誤差比率 平均	1.86	14.06	8.88	5.69	4.22

表 12 属性組み合わせ数による結果比較(全体平均)

属性組合せ	1属性	2属性	3属性
重相関R 平均	0.98995	0.98807	0.99058
重決定R2 平均	0.98008	0.97637	0.98132
標準誤差　平均	1300.61	1852.58	1069.61
有意F　平均	1.1E-15	3.0E-16	1.2E-16
平均誤差比率　平均	6.83	8.03	6.52



	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 37 国勢調査及び上位 10 サービスの誤差推移

また，累乗近似式を 3 属性以下の組み合わせで作成した場合について検証したのが表 12 である． 3 属性を組み合わせた k 値から作成された累乗近似式の重相関係数が最も高い． 1 乃至 2 属性によって得られた k 値から作成された近似式についても，重相関係数が 0.98 を超えるため，簡易的なものとして用いることが可能である．

これらの回帰分析結果の結果によると，重相関係数は高く出るが，標準誤差平均について，国勢調査で 2238.56 と非常に誤差が大きいことが判明した．しかし，標準誤差値は k 値が大きい場合における誤差も含めた平均値であるため，小さな k 値における誤差を表現するには不適当である．

そこで，k 匿名性が 1 に向けて収束していく性質から，各属性区分数の誤差を相対化して評価するため平均絶対比率(8)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 N で割り平均化したものである．

$$\frac{1}{N} * \Sigma \left(\max \left[\frac{a}{a'}, \frac{a'}{a} \right] \right) \quad \cdots \ (8)$$

これによって累乗近似型の予測式は，全体平均で 6.52 倍以内の値で予測が可能であることが判明した．このような誤差範囲の大きい予測式では，k-匿名性を満たす属性区分を正確に出力することは難しい．

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%

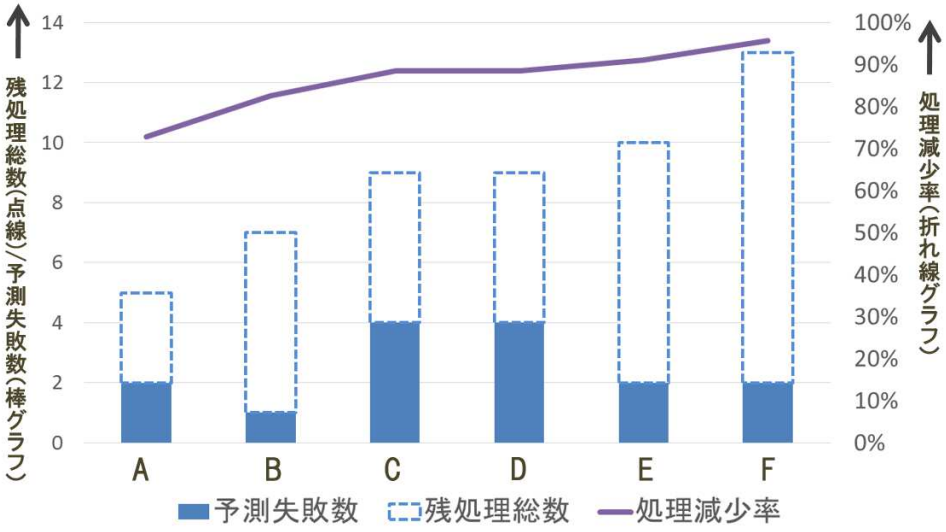


図 38 処理削減効果と匿名化予測失敗数

また，本予測式の評価軸として，処理の削減効果を考えることもできる．図 38 はサービス A～F における匿名化予測の失敗率と，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，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．また，全ての処理結果において 1～4 項目の予測失敗群を残している．

4.4　本章のまとめ

これらの実験結果により，本予測式は，重相関係数は 0.99 以上の高い数値が出るが，予測誤差が大きく，予測値によって匿名化処理を行った場合に，匿名化処理が達成できない場合が多く発生することが判明した，

しかし，重相関係数が高いことから，匿名化可能な限界までの大まかな傾向をつかむことが可能である．

そこで，累乗近似型の予測式を用いて，匿名化処理が実現できる属性区分数を予測し，最も予測値に近い属性組み合わせから匿名化処理を開始する．予測値の k-匿名性の検証結果によって匿名化アルゴリズムを選択することで，処理回数を削減するアルゴリズムを提案する．

5 累乗近似式を用いた匿名化処理選択方式

5.1 提案アルゴリズム概要

k 値が累乗近似型に減少する性質を持ち, その予測値と実測値が高い重相関係数を持つことが期待できる場合, 最も予測値との差が小さい属性組み合わせを選択し, その地点から匿名化処理を開始する方式を提案する. また, その開始地点における属性組み合わせがk-匿名性を満たすかを確認することによって, ボトムアップ型, またはトップダウン型の匿名化処理を選択し, 既存アルゴリズムの課題を解決する.

まず, Lattice Structure によって作成された属性の組み合わせについて, それぞれの属性区分数を検証する.

表 13 一般化階層の例

A		2区分	男			女		
B	B1	2区分	10代			20代		
	B2	4区分	10-14才		16-19才	20-24才		25-29才
C	C1	2区分	東日本			西日本		
	C2	6区分	北海道・東北	関東	中部	関西	中国・四国	九州・沖縄

k-匿名化処理を行う候補として, 表 13 として属性 A,B,C に対して, 一般化階層 A,B_1,B_2,C_1,C_2 を適用し, 最も詳細で k-匿名状態を満たす群を Lattice Structure の候補から探索する場合を想定する. また, アルゴリズム上で探索する際の優先度として C>B>A と設定する. これは, (A,B_1) と(A,C_1)は両方とも 4 区分であるため, 探索の際の検証順を決定するためである.

対象データに一般化階層を適用し, 属性の組み合わせとその区分数を記録する. 例えば, (A,B_1)=4 区分{(男,10 代), (女,10 代) , (男,20 代) , (女,20 代)}を構成し, Lattice Structure を作成すると図 39 の形で全候補が作成される. この属性の組み合わせについて, 区分数が少ないものから順番に, 既知の y として使用する k 値を一定数取得し, 記録する. また, 本組み合わせにおいては(B1,C1)が最も予測値に近い属性組み合わせであった場合を記載している.

既知の x, y によって作成された累乗近似式から, k 匿名性を満たす属性区分数を予測し, Lattice Structure 内で最も差分の絶対値が少ない属性組み合わせを選択する.

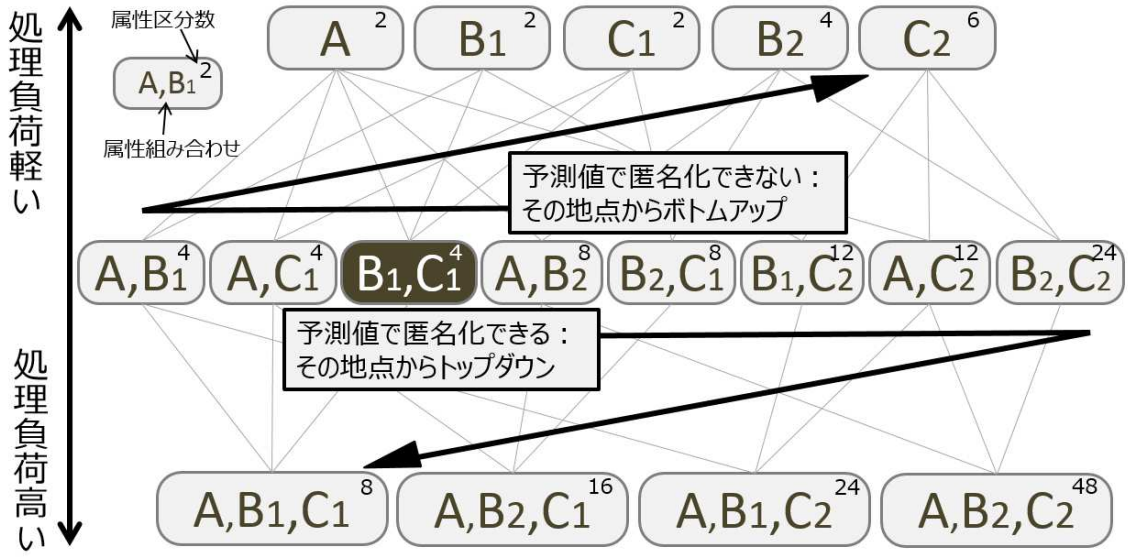


図 39 提案アルゴリズム概要

最も予測値に近い属性組み合わせで，k-匿名化処理が達成できるならば，そこで匿名化処理終了である．その上で，より詳細な組み合わせの存在を疑うならば，その地点を開始点としてトップダウン型の匿名化処理アルゴリズム(Incognito 等)を用いて匿名化可能な組み合わせを探索することも可能である．

逆に，指定した組み合わせでは匿名化処理が達成できない場合，その地点を開始点としてボトムアップ型の匿名化処理アルゴリズム(OLA 等)を用いることで検証回数を減少させることができる．

匿名化処理は組み合わせ数が増加するごとに処理数が増加するため，処理を行う範囲を限定することで，大きく処理回数を削減できる．

5.2 サンプルコード

サンプルコード ： 累乗近似予測値による匿名化処理選択方式(PAK)

Input: 匿名化すべき複数属性($D_1, D_2 \ldots D_N$)のテーブル T.

Power Approximation k-anonymity Method(PAK)::={

1. 処理開始:属性番号と属性値の種類を定義する.

Array[1]=D_1\{ $d1_1, d1_2 \ldots d1_n$\}... Array[N]= D_N\{$dn_1, dn_2 \ldots dn_m$\}

2. D_1～D_nの一般化階層に沿った組み合わせと属性区分数を記録.

3.累乗近似式の作成に必要なサンプル数z回の k-匿名性を計測する.

for (i_1 =1, i_1 <= z , i_1 ++) {

query1=Select count(*) as *Count*, Array[i_1] *as Att* from *T* group by Array[i_n] ； // *query1 の結果を table R に記録.*

for (i_2 = 1, i_2 <= *z*, i_2 ++) {

query2=Select min(*Count*) as *k_val*, count(*) as *k_cnt* from *R* where *Att* = Array[i_2] ; // *query2 の結果を table C に記録.*

} end for

} end for

4. 累乗近似値 PowApp_k を求める

query3 = update *C* set *k_calc* = LN(*k_val* – AVG*(k_val)*;

query4 = update *C* set *c_calc* = LN*(k_cnt*– AVG*(k_cnt))*;

query5=select EXP{sum(k_calc*c_calc)/sum(k_calcˆ2)} as A,

AVG(c_calc) – (***k*** * AVG(*k_calc*)) as B from *table C* ;

PowApp_k = (k-1)^(1/ B) / A^(1/ B)； // k 値の予測値

5. 全組み合わせ中から最も予測値に近い組み合わせを取得.

for (i_3 = 0, i_3 <=*N*, i_3 ++) {

If (abs(Count(Array[i_3])***Count(**Array[i_{3+m}])－**PowApp_k)) <** *min 値* **)**

{ Array[*x*]と Array[*y*]の組み合わせが最も予測値に近いと判明 }

} **end for**

6. 最も予測値に近い組み合わせにおける k-匿名性(min(*Count*))を計測.

7. 求める k-匿名性を満たさない場合, ボトムアップでより粗い情報を探索する. 満たす場合は, 処理を終了しても良いが, トップダウンのアルゴリズムを用いて, より詳細な情報を検証することもできる.

If (min(*Count*) < *k*){

Use Bottom Up Algorithm } **else** {

GODval1 = *x* , **GODval2 =** *y* , **GODk=** min(*Count*)

Use Top Down Algorithm } **}**

Output: 最も情報量が多く k-匿名性を満たす Globally Optimal Dataset を含む Table *C*[表 14]

図 40 は本アルゴリズムのフローチャートである.

1. 属性とその属性に適用する一般化階層について定義を行う. 表 14 を用いた場合, A→D_1{男,女}, B_1→D_2{10 代,20 代..}, B_2→D_3{10-14 才, 15-19 才..}と定義し, 一般化階層を数値で取得できるよう変換する.

2. 属性の全組み合わせを作成し, 属性区分数を求め, table C に記録する. (例: [D_1:2 区分] [D_1,D_2:4 区分] [D_1,D_3:8 区分]...) また, 表 14 の場合 B∈D_2,D_3 ,C∈D_4,D_5 であるため, [D_2,D_3] [D_4,D_5]を含む候補は除外する.

3. 精度の高い累乗近似式を取得するために必要な k-匿名性と属性区分数を既知の x, y と設定する.

4-5. 累乗近似式により予測値を作成し, 予測した属性区分数に最も近い属性組み合わせを求める. サンプルコードでは絶対値で差分の少ない値を求めているが, 予測値よりも大きい場合に排除する手法もある.

6-7. 予測による組み合わせによる k-匿名性を取得し, 求める k-匿名性を実現した場合は, Top Down Algorithm によって, より詳細な値を探索し, 実現しない場合は Bottom Up Algorithm によって, より抽象化された値を探索する.

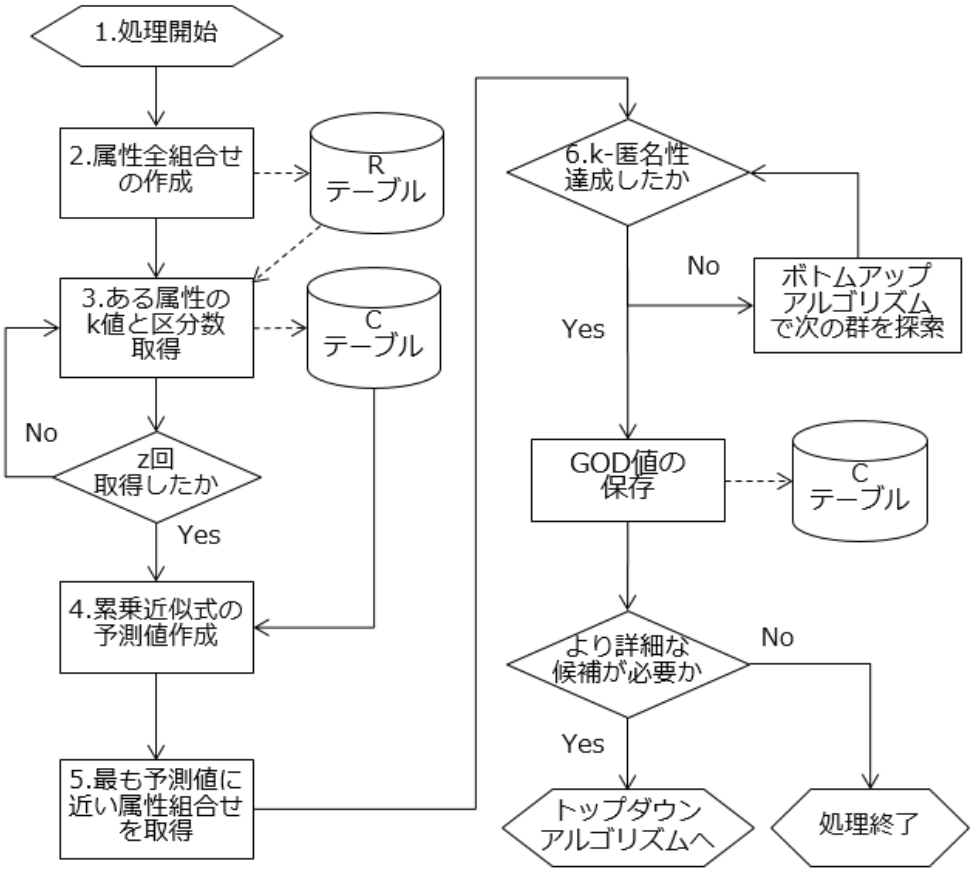


図 40 提案アルゴリズム(PAK)のフローチャート

表 14 table C のサンプル

No.	Att	k_val	k_cnt	k_calc	c_calc
	属性組合せ	k 値	属性区分数	k 平均値	属性数平均値
1	(D_1, D_2)	$k\_val(d1_n, d2_m)$	$Count(Array[D_1 * D_2])$	k_calc	C_calc
2	(D_1, D_3)	$k\_val(d1_n, d3_m)$	$Count(Array[D_1 * D_2])$	k_calc	C_calc
:	:	:	:	:	:
GOD	(D_x, D_y)	$k\_val(dx_n, dy_m)$	$Count(Array[D_x * D_y])$	-	-

ここで出力された table C[表 14]は，既知の x，y に加え，本アルゴリズムの処理過程で得られた新たな属性区分数と k 値が記録されていき，値がより正確となる．そのため，もし再度匿名化処理を行う場合は 1～3 までの手順を省略し，直接累乗近似式を作成することが出来る．

5.3 提案方式の評価

本方式(PAK:Power Approximation k-anonymity method)は，直接的な匿名化処理ではなく，データの特徴に合わせた匿名化処理を選択することで処理の削減を図る方式のため，従来アルゴリズムの単独処理と比較した処理削減率にて効果を計測する．

比較対象とするアルゴリズムを検討するため，必要条件を整理する．

本方式は，均一の一般化階層を用いるため，属性一般化階層（DGH:Domain Generalization Hierarchy）を用いた場合と同じく，結果データにおける階層の違いが発生しない．そのため，値一般化階層（VGH:Value Generalization Hierarchy）を用いるアルゴリズムを適用した場合，結果データが異なるため比較対象として利用できない．

また，結果データから値を削除，または別の値に変更（スワップ処理やノイズ付与処理等）する処理は，出力するデータ形式が異なるため同一の形で評価できない．そこで，2.6 章で検討した匿名化処理アルゴリズムを検証し，表 15 に条件を整理した．

その結果，条件に合致する Bottom Up 型アルゴリズムは，Datafly，及び OLA である．OLA は情報量の多い群から処理することから，Datafly と比して処理検証回数が少なくて済み，かつ，最も情報量の多い組み合わせ地点を決定することが可能であるため，OLA を選択する．

また Top Down 型アルゴリズムで条件に合致するのは，Incognito，及び SEM 価格を用いた匿名化処理である．SEM 価格を求める処理は，有用性を最大化することを目的とするため，処理回数の削減を考慮していない．そのため，Incognito を選択する．

表 15 比較対象候補のアルゴリズム

アルゴリズム	処理方式	一般化階層	削除・変更
Datafly[69]	Bottom Up	DGH	あり/なし両方可
OLA[70]	Bottom Up	DGH	なし
μ-Argus[17]	Bottom Up	VGH	あり
マージナル[77]	Bottom Up	使用しない	なし
Incognito[67]	Top Down	DGH	なし
Mondrian[75]	Top Down	VGH	なし
minDIS[71]	Top Down	VGH	なし
Partition Algorithm[73]	Top Down	VGH	あり
SEM 価格を用いた匿名化処理[79]	Top Down	DGH/VGH 両方可	なし
Utility-Based Anonymization[65]	Top Down	VGH	なし

OLA, Incognito, 及び PAK 方式は, それぞれ処理の目的が異なる. Incognito は k-匿名性を満たす全ての属性組み合わせを導くことを目的としているが, OLA は, 最も情報量が多く, かつ k-匿名性を満たす属性組み合わせ(GOD: Globally Optimal Dataset)を 1 つだけ求めることを目的としている. そこで本実験では目的を OLA に合わせ, GOD 値を求めるまでの属性組み合わせの検証回数によって比較する.

また, PAK 方式に関しては, 予測値から検証を開始した際に, 予測誤差によって, より詳細な群について匿名化処理が可能な属性組み合わせが存在する場合がある. そのため, 予測誤差発生後の修正として, ボトムアップ型の処理を行った場合, トップダウン型の匿名化処理検証を行った計算回数を加えることで, 条件を均一化して実験を行った.

処理削減率を検証する実験は, 4.3 章で用いた国勢調査, 及び擬似サービス群において人数の多い上位 6 サービス(S_1～S_6)に対して実施した. それぞれのデータ量, 及び予測式と実測値との相関係数を表 16 に示す. 総処理回数は, 匿名状態を検証するべき属性数の合計(総数:3185)である.

表 16 対象サービスの人数と予測式

	国勢調査	S1	S2	S3	S4	S5	S6
総人数	127794	350527	215122	208675	190105	140826	134721
予測式α値	114659	50316	40048	19664	16326	13618	6641
予測式β値	−1.414	−1.776	−1.852	−1.735	−1.695	−1.722	−1.642
予測式と実測値の相関係数	0.9974	0.9901	0.9911	0.9887	0.9853	0.9880	0.9805

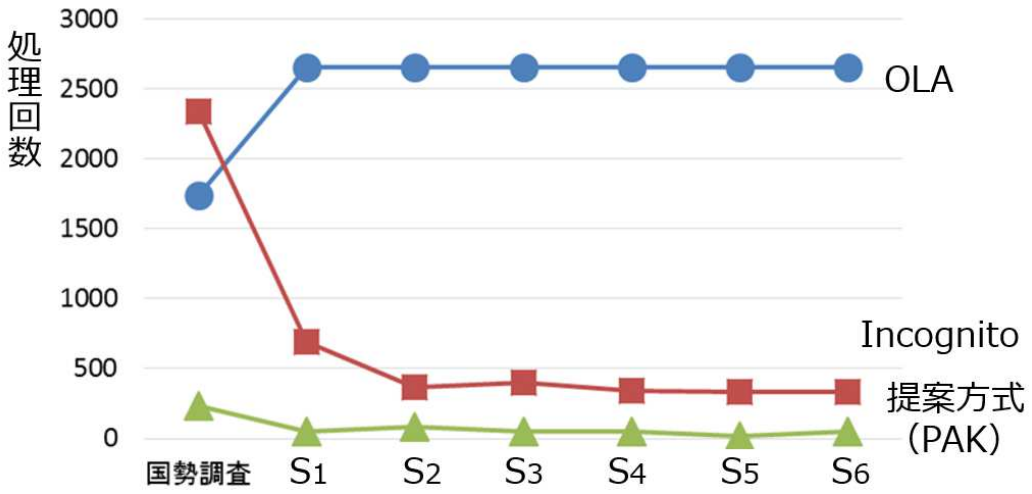


図 41 k≧50 における匿名化処理量比較

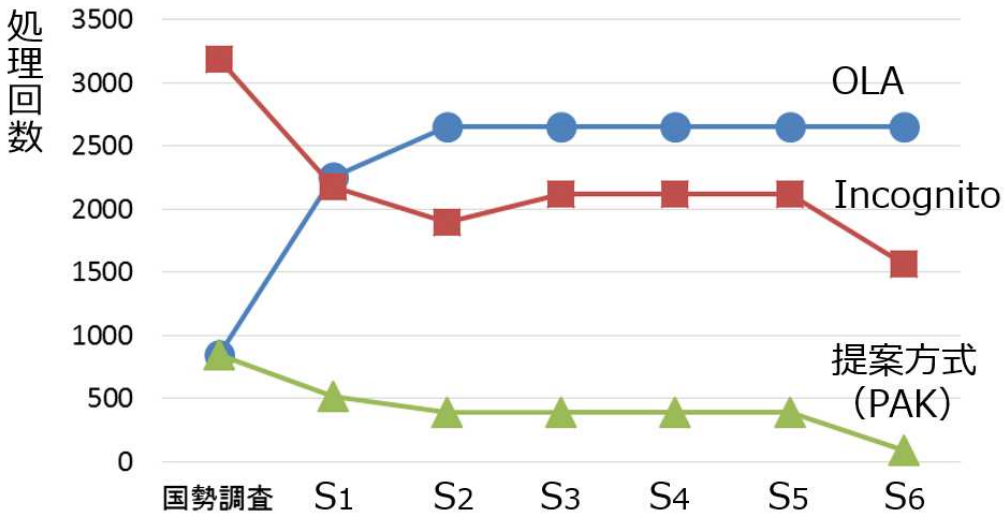


図 42 k≧2 における匿名化処理量比較

図 41 は，k≧50 における匿名化処理にかかった処理量の比較である．予測値に基づいた PAK 方式は OLA 方式と比較して平均で 3.5%，Incognito 方式と比較して平均で 12.5%の処理量で匿名化処理結果を出力した．

図 42 は，k≧2 における処理量比較である．k 値を小さく設定した場合，OLA 方式と比較して平均で 26.6%，Incognito 方式と比較して平均で 19.1%の処理量となった．特に，国勢調査において，OLA 方式と比較すると処理量が同じとなっている．これは最も詳細な群においても k≧2 が達成できたため，ボトムアップ方式における最良ケースで探索が終了したためである．PAK 方式も，最も詳細な組み合わせでの匿名化が可能であるとの予測ができたため，結果として処理回数が同じとなった状態を示している．図 43 にて，ボトムアップ方式における最良ケースの概念図を示す．

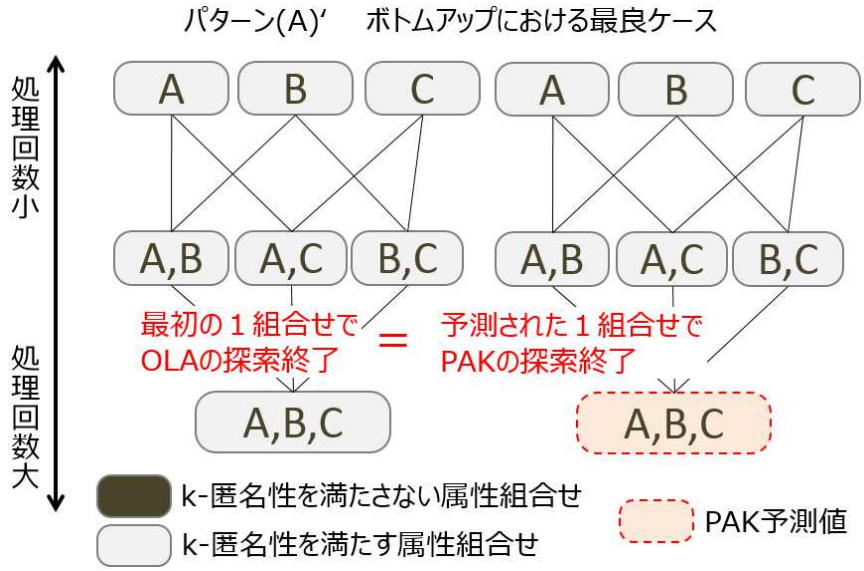


図 43 ボトムアップの最良ケースと PAK の比較

図 44 は，図 42 の実験にて OLA 方式と同一の値であった国勢調査を除き，人数の多い順から 50 サービスにおける PAK 方式の処理削減量と相関係数の関係性を検証した結果である．予測値と実測値の相関係数が低い群の方が，OLA 方式と比較した相対的な処理量が少ないことが解る．

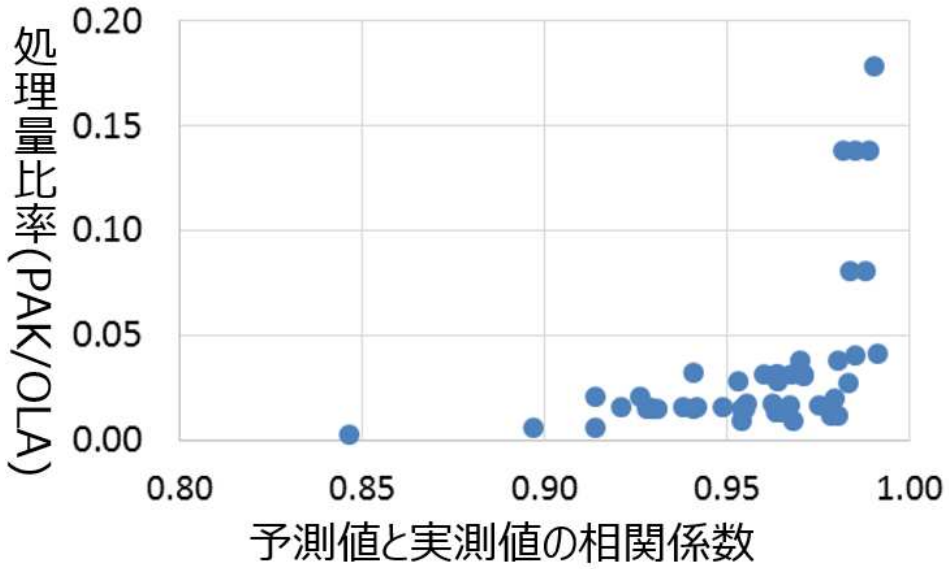


図 44 相関係数と処理量比率の関係性グラフ

この現象は，予測値と実測値の相関係数が低い群は，属性値の分散が大きく，k 値が低くなる傾向が強いために発生している．表 17 は，図 44 の結果を階級ごとの処理量平均に換算したものである．これによると，対象データの予測値と実測値の相関係数が 0.98 以上における OLA と比較した処理量平均が 8.31%であるのに対し，相関係数 0.94 以下の群は平均 1.37%で処理が終了している．

表 17 相関係数の階級ごとの OLA 処理量平均

相関係数	対象データ数	OLA 比処理量平均
0.98 以上	11	8.31%
0.96-0.98	19	2.18%
0.94-0.96	9	1.85%
0.94 以下	11	1.37%

ボトムアップ型の OLA 方式は，k 値が低い場合には，全探索を行うため，図 45 における最悪ケースであるパターン(A)に近い結果となる．それと比較して PAK は相関係数が低く，予測値による匿名化処理が実現出来ない場合でも，探索回数が少なくて済み，相対的に OLA と比べて探索効率が向上した．図 46 にて，PAK 方式とボトムアップアルゴリズムの探索回数の差を示す．

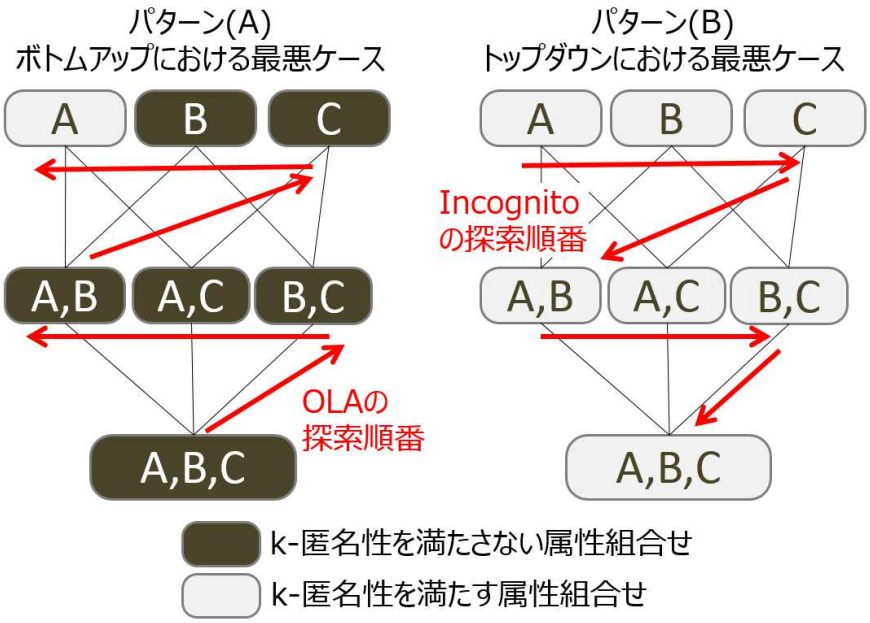


図 45 各アルゴリズムにおける最悪ケース（再掲）

同様の現象は，処理量比較にて，OLA が最も効率的に処理できる国勢調査の際に incognito の効率が最も下がる結果となっていることから，トップダウン型アルゴリズムに関しても同様の現象によって処理効率の変化が発生している．

その一方で，PAK 方式は予測した地点から処理を行っているため，他のアルゴリズムにおいて，探索が非効率になる場合においても処理削減効果が高く維持できたと言える．

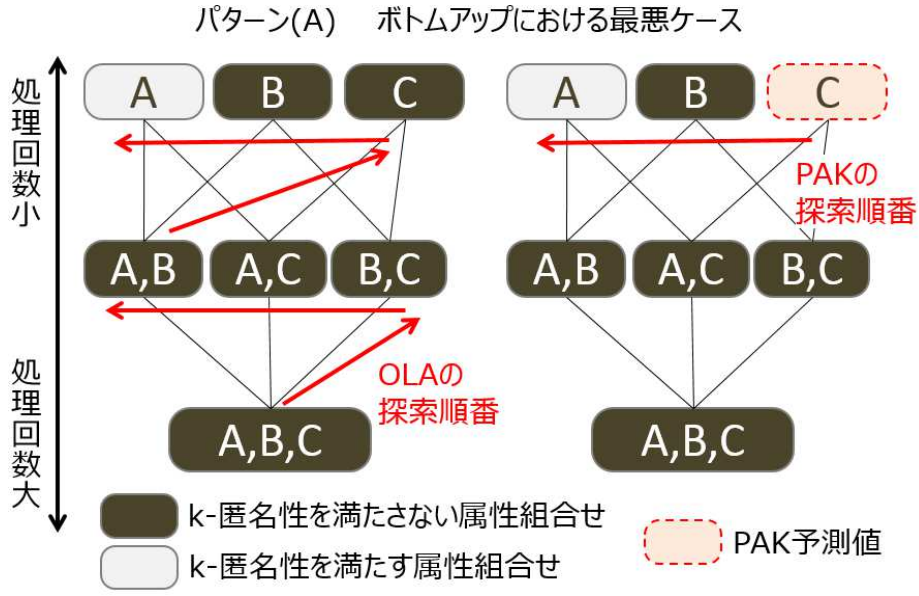


図 46 ボトムアップの最悪ケースと PAK の比較

5.4　本章のまとめ

累乗近似型の k-匿名性予測式について，国勢調査や擬似データを用いて検証したところ，5000 人以上の群において重相関係数が 0.99 以上と，高い値が得られた．しかし，全体平均で絶対誤差比率 6.52 であり，直接的な k 値予測に利用するのは難しい．

そこで，k 値を得られる属性組み合わせの情報を，匿名化処理の開始地点として利用し，その地点で匿名化処理が可能な場合は，より詳細な情報を探索するトップダウン方式を採用し，匿名化処理が失敗した場合は，より抽象的な情報を探索するボトムアップ方式を採用することで，探索作業を効率化する，PAK 方式の匿名化処理を提案した．

PAK 方式は k≧50 の時，平均で OLA 方式と比較して 3.5%，Incognito 方式と比較して 12.5%の処理量で匿名化処理結果を出力した．

PAK 方式は，対象データの分布が正規分布で，かつ各属性同士が独立している場合の予測に用いることを想定していたが，検証によって，特徴的な分散を持つデータに対しても，他のアルゴリズムと比べ，処理回数を削減できることが判明した．

6 匿名化データの流通プラットフォーム

本研究では，3 章にて匿名化処理における安全性の推移検証を行い，4 章，5 章では，安全性の予測式とそれを用いた匿名化処理の効率化アルゴリズムを提案し，検証した．

そこで，6.1 章では，**3）実社会に即した匿名化データの流通方法　の解決に向け，**匿名化処理プラットフォームの社会実装に向けての課題について検討する．具体的には，4 章で提案した予測式によって，データ利用者による安全性と有用性の事前検証を行うことで，データ作成者，データ利用者双方の負荷を軽減させる仕組みを提案する．

6.2 章では，6.1 章で検討した匿名化データの授受プラットフォームの社会実装に向けた課題を洗い出すため，プロトタイプを構築し，評価した．また，評価指標の検討のため，研究者が匿名化データを作成し，それを再識別する公開実験に対してプラットフォームを提供したことにより，社会実装に向けて必要となる安全性指標について検討した，

6.1　匿名化データの流通プラットフォームの検討

本章では，匿名化処理プラットフォームの課題である，匿名化データの提供者とその利用者における，安全性と有用性の要望の不一致問題を解決するための仕組みを検討する．これらの課題は，データ利用者が求める利用性を満たす匿名化データを生成するとき，データ提供者が求めるk-匿名性が実現するかを予測できない，という従来の匿名化処理アルゴリズムの問題に起因している．

本章では，データ提供者が 4 章で検討した予測式を用いることで，データ利用者の求める一般化階層でのk-匿名性が実現できるか，事前に予測するシステムについて検討する．

6.1.1　k-匿名性の予測式を共有するプラットフォームの提案

図 47，表 18 は本システムの概要である．あるパーソナルデータ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一般化階層について合意することを目的とする．

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．

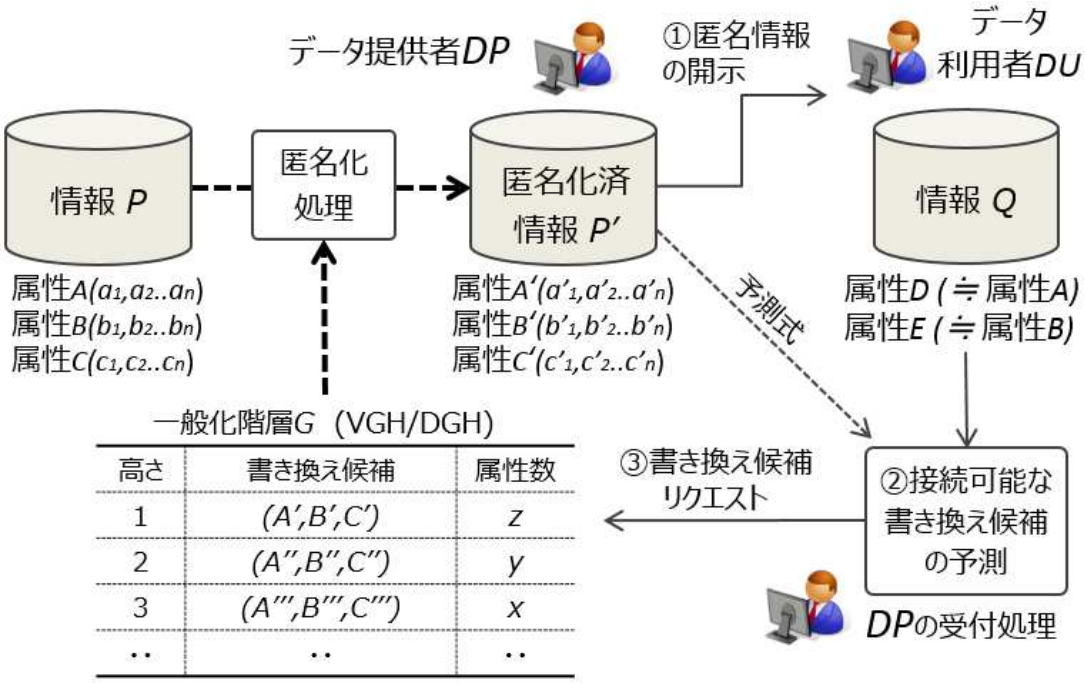


図 47 予測値を用いた匿名化システム案

表 18 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	

このようなデータ利用者からリクエストを受けた形で対応する匿名化処理は

① 匿名情報 P'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 P''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．しかし，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 $P(a_1 \ldots a_n)$と公開匿名化データ $P'(a'_1 \ldots a'_n)$，生成される情報 $P''(a''_n \ldots a''_n)$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[86,87] (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．

本問題は，多次元にわたる属性値ごとの減少傾向を，区分数という 1 次元に統合して x 軸と設定し，y 軸に k-匿名性の減少レベルを設定した領域の問題として定義できる．

図 48 にその関係性を示す．DU が x_u 区分によって情報の再処理要求を行った場合の DP の k 値(k>1 を満たす整数)を k_u とする．また DP が情報を提供できる安全基準 k_p を満たす情報区分数 x(x>1 を満たす整数)の最小値を x_n と定義したとき，DU の利用性は満たすが DP の安全性基準を満たさない，結合不能領域 $S_1=(k_p\ k_u)*(\ x_u\ x_n)$ を定義できる．この領域が最も DU にとって価値が高く，背景情報を類推される可能性が高い．

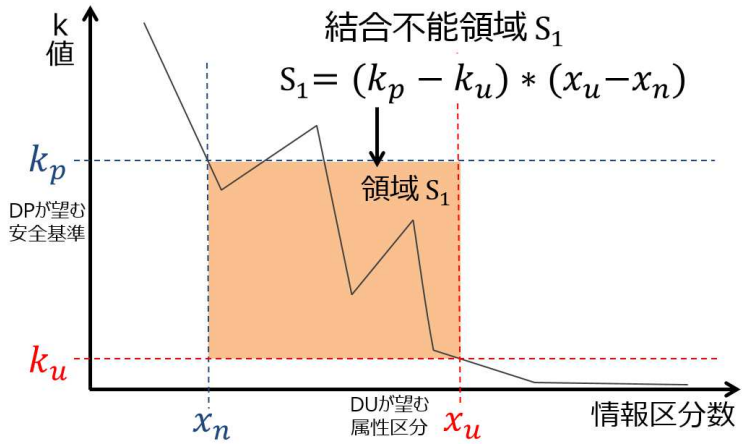


図 48 情報区分数と k 値の関係と結合不能領域

このとき，S_1 に対する再処理要求 $[x_u, x_{u-1}... x_n]$ への結果$[k_u,k_{u-1}..k_n]$の比較によってプライバシー情報の推定が可能となる．

図 49 は予測式を用いてデータ利用者 DU 側が匿名化を依頼できる区分を制限する方式の例である．

1)データ提供者 DP は匿名化データ P'を提供する際に，一般化階層 G_p, k-匿名レベル予測式 $f_{(DP)}$，安全基準 k_p を提供する．

2)データ利用者 DU は必要な情報区分を G_p から選択し，その区分数 x_u を求め $f_{(DP)}$に適用する．

3)$f_{(DP)}$により出力された予測値 k_u が安全基準 k_p より大きい場合，DP は G_p の x_u 区分による再処理要求を受領し，匿名化処理を行う．

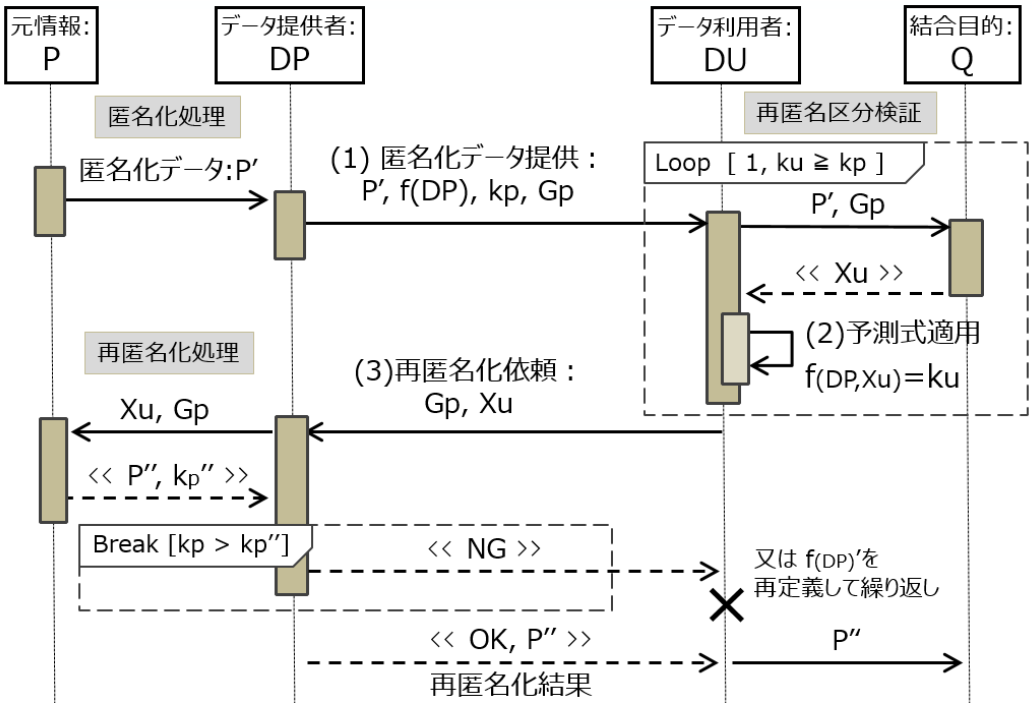


図 49 提案手法のシーケンス図

6.1.2 提案方式の評価方法

図 49 の提案手法を用いた場合の結合不能領域 S_2 の面積は $f_{(DP)}$の予測誤差によって図 50 として定義される. S_1 は x_p で分割され, 誤差領域 S2= (kp－ku)*(xp－xn) まで削減される. また, $S_2 = S_1 - T$ であるため, S2 のリスクが S1 より大きくなることはない.

予測式 $f_{(DP)}$が正確であれば個人情報 P は, k_pを満たす匿名化データ P"の作成が可能となる. 逆に予測誤差が多く発生した場合, 領域 S_2 が大きくなり, 予測式を用いる利点が減少する. もし予測式が大きく安全性を損なっている場合は, 予測式の傾きを変更することによって安全基準 k_p を変更することも可能である.

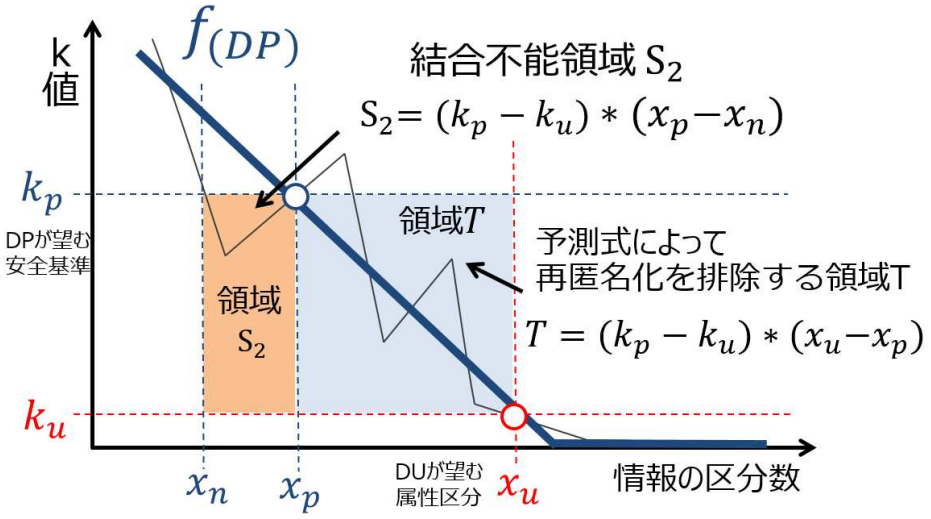


図 50 提案手法における領域の概念図

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．

しかし，本予測式による失敗を完全に排除することは出来ないため，図 50 で示した結合不能領域 S_2 は，予測誤差の範囲だけ広くなる．本方式はそれに対応するため，結合不能領域 S_2 は $f_{(DP)}$ の角度を修正した予測式 $f'_{(DP)}$ に変更することで削減できる．だが S_2 を最小にした場合，予測式によって情報提供リクエストを排除される非提供領域 N も増加する．図 51 にてその領域 N を示す．

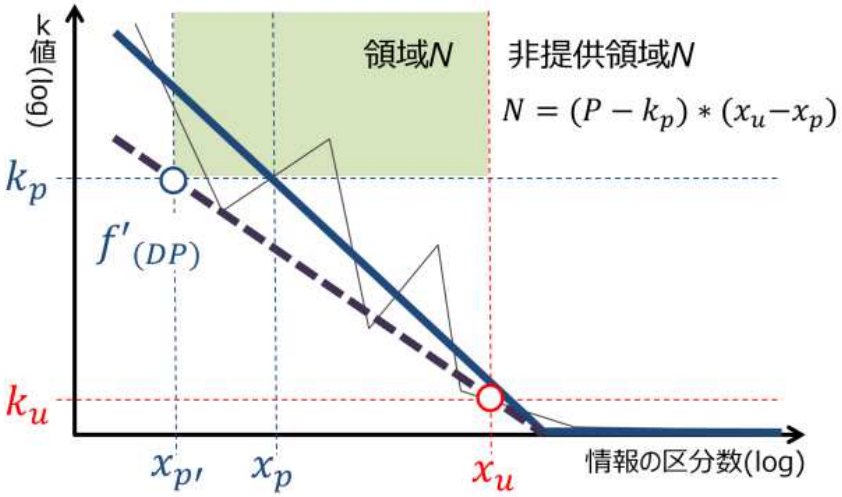


図 51 提案手法における結合不能領域

DP が安全性を増すために近似式の精度を厳しく設定した場合，DU が利用できる範囲が減少するというトレードオフの関係がここで成り立つ．そのため安全性だけでなくと利用性の面でも評価できる方式が望ましい．

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 52 にて，その区分方法を示す．

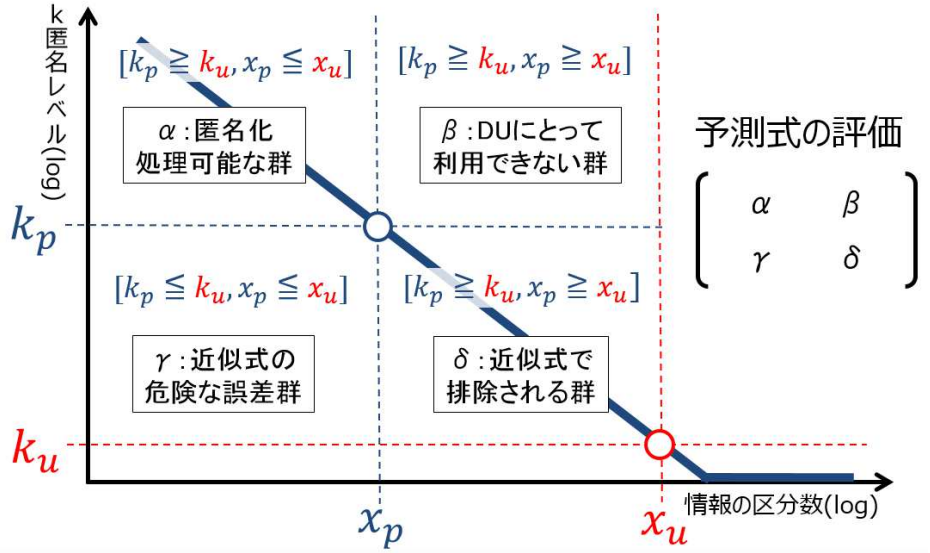


図 52 提案手法の評価方法の提案

図 52 における四象限はそれぞれ

α:正常な予測範囲

β:DU に対する提供機会の逸失

γ:情報推定リスク

δ: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可能である．本評価方式を用いて 3 章，4 章で利用した国勢調査，及び各階級による評価を行った結果が図 53 である．

図 53 によると，国勢調査群の結果は γ:結合不能領域が存在しない．しかし，その他の実サービス群では 5～10%程度の γ 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 α 領域を大きくすることも容易である．

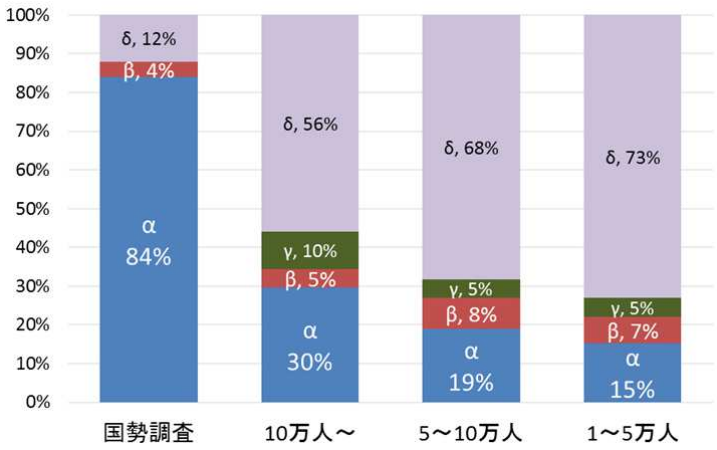


図　53　kp=50 における 4 象限評価結果

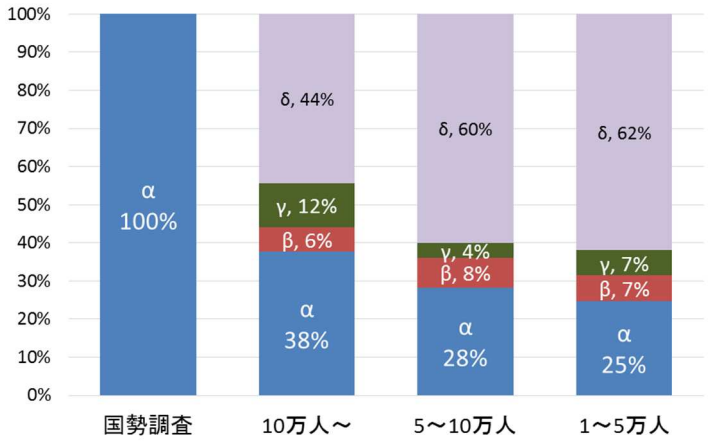


図　54　kp=10 における 4 象限評価結果

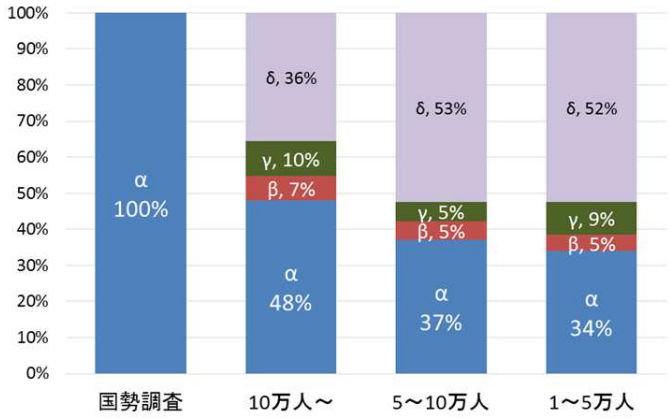


図　55　kp=2 における 4 象限評価結果

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 54，図 55 である．この場合，国勢調査の情報は全て提供可能になり，10 万人超のデータにおいては α 値が 30%→48%に向上しているが，結合不能領域 γ 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．

6.1.3 提案方式のまとめ

表 19 は，提案方式と従来方式を分類するものである．提案方式は，従来方式では提供しなかった一般化階層と予測式を公開する．匿名化処理回数の最大値は従来方式が($x_u - x_n$)であるのに対し，提案方式は x_u –($x_p + x_n$)となり，情報の予測値 x_p 分の処理回数が減少する．それによって結合不能領域は従来方式の S_1 よりも領域 T の面積分減少する．

また，従来方式が DU の信頼性などによって安全基準 k_p を変更すると，匿名化処理が再度必要になるのに対し，提案方式は自由な k_p を提供できる点が優れる．また，クエリ監査問題に対しても，提案方式は領域 S_2 へのアクセスが確認された時点で，予測式の傾きを変更し，領域 T を増加させることで安全性を変化させることができ，連続的なクエリ要求を排除できる利点がある．

表 19 従来方式と提案方式の比較

	従来方式	提案方式
公開情報	匿名化情報 P'	匿名化情報 P' 一般化階層 G_p 予測式 $f_{(DP)}$
匿名化処理回数	$x_u - x_n$	$x_u - (x_p + x_n)$
リスク対象範囲	S_1	$S_1 - T$
基準値の変更	不可能	$f_{(DP)}$ により予測可能
リスク軽減措置	オフライン監査 シミュレータブル監査	従来方式＋ 予測式の修正方式

本方式によって，個人情報を含むデータベースに対して，少ない負荷で匿名化処理を行うことが可能となる．また，匿名化処理の結果データが，属性区分数と k 値として蓄積されていくことで，オーダーメード型匿名化処理を行う際に，データ利用者の求める属性値によって，k-匿名化処理が可能であるかを予測する精度も向上する．

これによりデータ提供者とデータ利用者の双方の処理が効率化し，匿名化データの流通を促進することが期待できる．

6.2　パーソナルデータ流通プラットフォームの社会実装に向けた検討

本章では，前章まで検討した匿名化処理の安全性を向上させるプラットフォームを社会実装するために必要なシステム構成，及び，安全性指標について検討する．

個人情報の保護に関する法律に，2015 年 9 月に成立した同法の改正法[10]により，匿名加工情報という新たな情報の類型が定義された．匿名加工情報は，一定の条件の下で，本人の同意がなくても第三者に提供することが可能となる．匿名化処理されたパーソナルデータは提供の条件を満たすことで匿名加工情報と認定される場合がある．そこで，匿名化データをデータの評価者と共有し，評価，第三者提供までが簡便に実現できるプラットフォームが求められている．

前章まで検討していた匿名化データの流通方式を社会実装するためには，データの提供者，利用者の区分だけでなく，一般顧客への情報開示や，データ提供者の責任者による提供の可否を決定する仕組みが必要となる．

データ作成者が利用契約を結んだデータ利用者に対して，匿名化データ Y を提供する場合を考えたものが図 56 である．

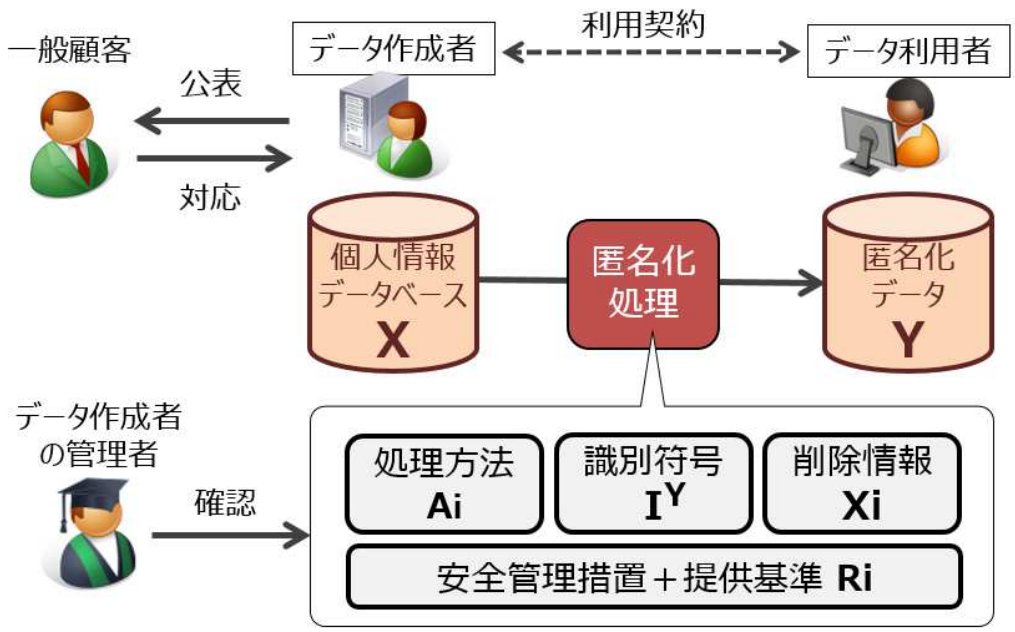


図 56 匿名化データの流通に関わるプレイヤーの関係図

まず個人情報データベース X に対して，匿名化アルゴリズム（情報の加工方法）A_i を用いて処理し，匿名化データ Y，個人識別符号 I^Y，加工に伴う中間データや削除された情報 X_i を出力した．また，それらの情報を，安全管理措置＋提供基準 R_i に沿って漏えいを防止する措置を講じた．この場合における提供基準とは，k-匿名性などの再識別リスクの定量基準や倫理基準などの定性基準を指す．

このとき，R_i，I^Y，X_i，A_i の基準を，データ提供者が定める基準に合致させる際，2 章で検討したような多様な指標に対応する必要があるが，多くのデータ利用目的に応じて共通で適用できる安全性指標が求められている．

また，データ提供者は，加工方法を類推されない範囲で一般顧客に対して情報提供を行う必要がある．例えば，企業 A が企業 B に対し，個人情報を匿名加工し，データベース形式で提供した，という情報が公開された場合を考える．

これを企業 A が公開した場合，顧客から企業 A へ問い合わせが来ることが予想される．例として「提供した情報は安全なのか」「安全性基準は何か」「匿名加工ソフトウェアは何か」等が挙げられる．

公開範囲を広くしたり，丁寧に質問に対応したりすると情報の加工方法の漏えいにつながる可能性がある．現在，多様な個人情報の匿名化アルゴリズムや評価指標が提案されており，それらの顧客からの質問，情報開示要求全てに対して，合理的な回答を行うことは難しい．

しかし「加工方法の詳細を伝えることは出来ない」という対応を行った場合，一般顧客の不信感の増大，オプトアウト希望ユーザの増加，風評被害など，データ作成者のリスクが大きく，自社データを提供するモチベーションが低下する．

6.2.1 社会実装に向けたプロトタイプシステムの検討

前章における課題を以下の 3 項目としてまとめる．

1) 匿名加工処理とその評価指標への柔軟な対応

2) 一般顧客に対する公表の方法とその根拠

3) データ利用者に対する匿名加工措置の保証

そこで，匿名加工情報を流通する上での課題を解決できる仕組みとして，データ作成者とデータ利用者の間をプラットフォーム事業者がつなぐ方式を検討し，検証用のプロトタイプを開発した．

プラットフォームは，主に以下の要素で構成されている．

1)サービス利用者管理・履歴保存システム

2)匿名加工システム

3)安全性評価システム

4)利用者管理/安全性基準 管理システム

まず前提として，図 57 のプロトタイプは，プライベートクラウド内に構築され，限られたユーザしかアクセス出来ない．かつ，その環境下でデータ作成者ごとに，独立したハード/データベース領域を生成し，処理を行う．

1)サービス利用者管理システムは，システムの根幹である．まず，利用者同士はネットワーク上のアクセス管理権限とセッション上の権限情報の二重の制限によって，他ユーザのデータベース環境を閲覧することは出来ない．

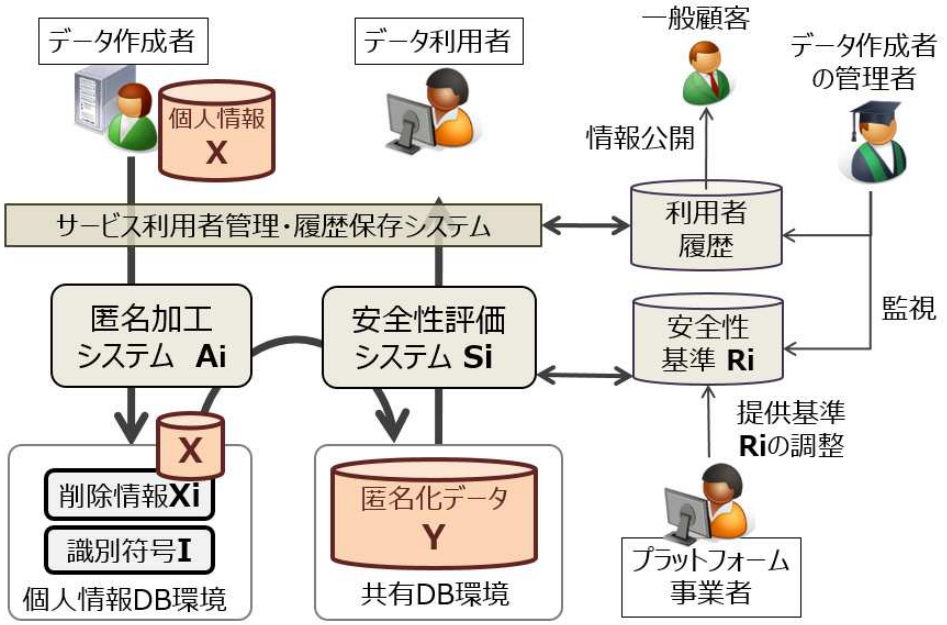


図 57 プロトタイプのシステム概念図

同環境下において，ユーザは 1 人に 1 つずつ別環境における個別データベース領域を設置し，識別リスクが評価されていない個人情報 X や，削除・中間ファイル X_i，個人識別符号 I^Xも含め，全てが個別の環境下で安全に管理される．

ユーザは，個人の権限に合致した範囲で **2)匿名加工システム**にデータを投入する．匿名加工システムは，データの種類に合わせて複数存在し，ユーザ自身が選択する．

作成された匿名加工情報は **3)安全性評価システム**を通過する際にデータの評価を行い，問題ないものだけがデータ利用者に対して公開される．また，公開する匿名加工情報のデータベースを共有にすることで，検索エンジンを搭載することが出来るため，公開情報を横断で検索し，そのまま複数情報ソースを比較検証することが可能である．

データ作成者が投入した R_i, I^Y, X_i, A_i 等のデータを **4) 利用者管理/安全性基準 管理システム**にて，正式な手順を踏んで匿名加工措置を行っていることを検証する．一般顧客に情報公開する場合は，このシステムと履歴データベースから適切なもののみを出力する．

本プロトタイプの問題点は大きく 2 点存在する．

1 つ目は，データ型式の多様性を受容するためにデータの型式とエラーチェック処理が多く必要となる点である．

データ作成者は，常に投入するべきデータのプロパティについて正確に認識している訳ではない．具体的には，マスター型かトランザクション型の選択，またはカテゴリ属性と

数値属性の区分，または言語情報や位置情報における辞書データ等を多く必要とし，システムが肥大化する．

もう 1 つは評価システムである．評価システムは既存のプラットフォーム実験[30]等で行われた定性評価と定量評価を含めた総合評価システムを参考に複数搭載し，総合的に評価した

しかし，全てのデータに一律に適応できる評価方式が存在せず，結局は審査する側の定性評価が主となった．そこで，プロトタイプでは，最終的に全てのデータを識別子がユニークに存在するマスターデータ型に変換し，k-匿名性を共通評価指標として，関係者によるデータ流通実験等に利用してきた．

結果として，データの多様性に関する課題は，匿名化処理プログラムを自由に設置，改良可能なプラットフォームとして構成することで，軽量化と運用性を高めることに成功した．また，評価システムの課題については，次章における匿名化データの評価プラットフォームによって検討を行った．

6.2.2 匿名化データの流通と評価を行うプラットフォームの検討

匿名化データの流通プラットフォームを実現するにあたり，複数の匿名化手法に対して共通で適用できる評価システムの実現は大きな課題である．菊池らは[88,89]において，匿名化データの安全性と有用性の評価を行う仕組みとして，公開実験に参加した研究者らが，共通のパーソナルデータに対して匿名化処理を行い，その結果データを他の研究者らが再識別する手法を提案している．

そこで，この公開実験に対し，匿名化データの流通と評価を行うプラットフォームを提供し，評価システムの検証を行った．

公開評価実験は，最大知識攻撃者モデル(maximum-knowledge attacker [90,91])を想定し，匿名化・再識別の参加者双方が元データセットを共有した状態で行われ，匿名加工と再識別の 2 つのフェイズが存在する．図 58 にてその流れを示す．

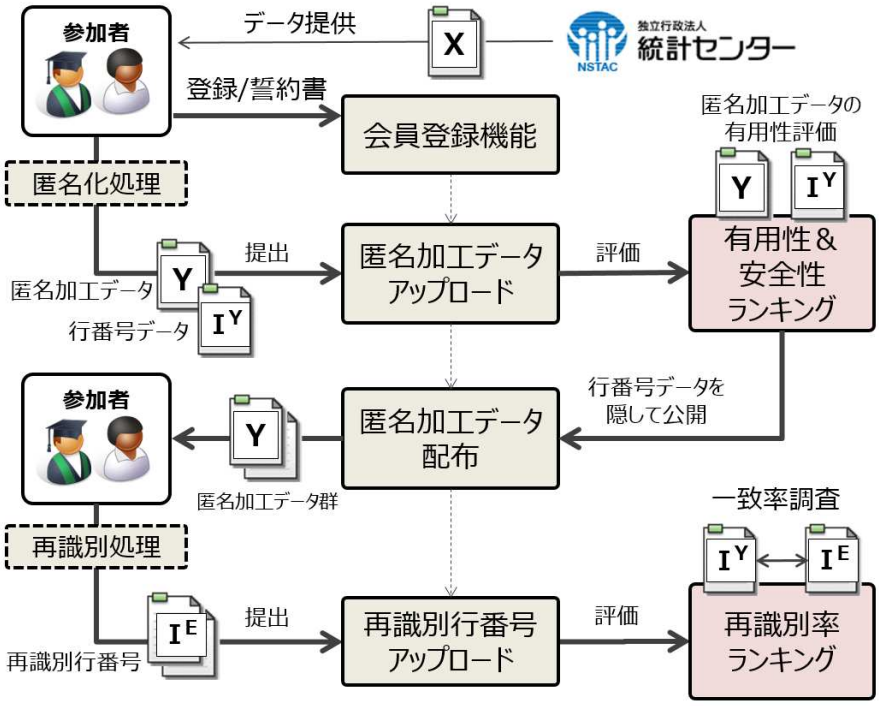


図 58 公開実験の流れ

本公開実験で使用したデータは，教育機関などの演習用として独立行政法人 統計センターが作成した疑似ミクロデータ[32]である．これを個人情報 X とした時，処理された匿名加工データを Y と表す．図 59 にてユースケースを示す．

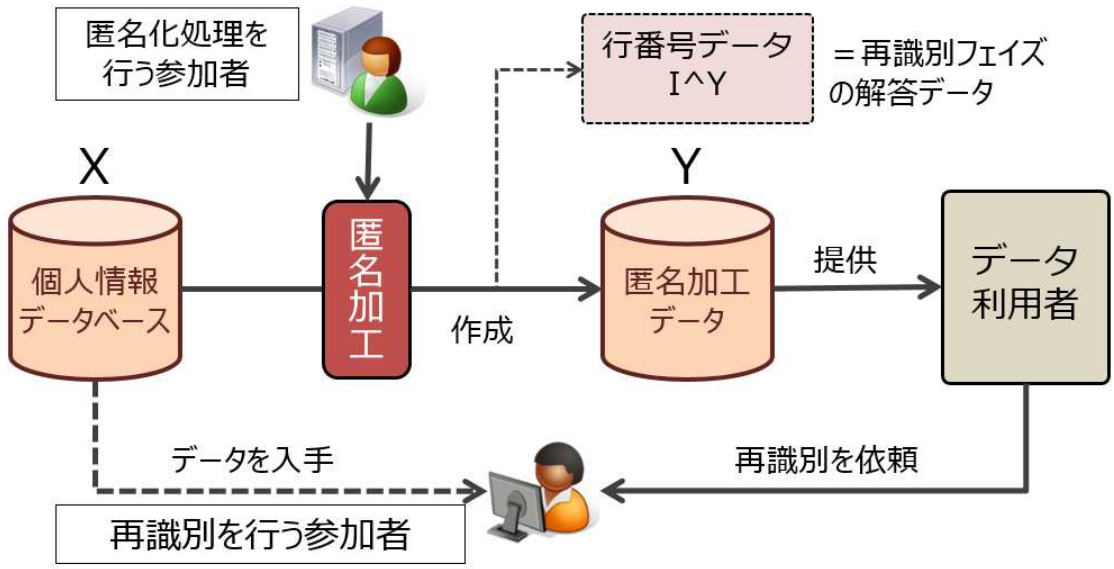


図 59 公開実験のユースケース

1）匿名加工フェイズ: データ提供者が個人情報 X の匿名加工データ Y を生成．連結可能キー I^Y を廃棄してデータ利用者に提供する．

2）再識別フェイズ:X を入手可能な人物が，Y の再識別を依頼されたため，X と Y を対照して再識別処理を行う.

本ユースケースは，ある機関や組織の中で作成された匿名化データを外部に提供する際に，同組織内の技術者によって，提供データによる再識別リスクを評価し，安全管理指標として活用できる.

6.2.3 公開実験による評価プラットフォームの検討

本公開実験では，匿名化処理は各参加者のローカル環境にて実施し，プラットフォームでは，その結果データの管理と評価を行う. そのため，本章では匿名化処理ではなく，評価システムの検討を行う.

匿名化データの流通と評価を行うプラットフォームの実現にあたり，ベースとなったプロトタイプから変更点が多く発生した. 代表的なものをまとめる.

1)個人単位ではなくチーム単位で処理を行う

2)匿名加工/再識別処理は参加者のローカル環境

3)使用データを擬似ミクロデータに固定

4)評価指標が 13 個あり，各値の順位で評価

本公開実験の概念図を図 60 にて示す. まず **1）チーム単位での参加システム**を開発した. 社会実装を行う際には，同様の組織・機関によるデータ共有の仕組みが求められる. そこで，互いに独立した処理であったシステムに対して，内部でチームとして結合するシステムを追加して解決した. 近年では，開発ツールや機械学習などの共同作業ツール等の要件でも同様の機能が求められており，ロール毎の権限設定については今後も改良が必要である.

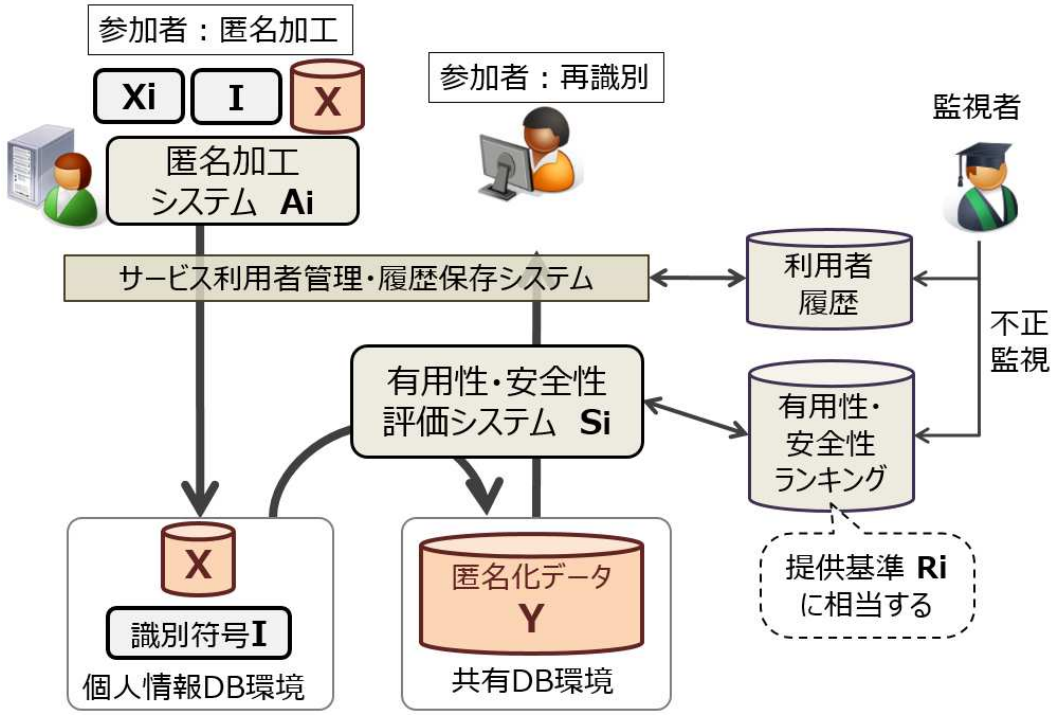


図 60 匿名化データの流通と評価を行うプラットフォーム概念図

2)匿名加工/再識別処理が参加者のローカル環境で行われることに関して，システム上の対応は軽微であった．

3)使用データが擬似ミクロデータに固定されたことは，システムの簡素化に貢献した．その反面，提出データの正規化チェックが必要となったため，システムを追加して対応した．

本公開実験で用いられた擬似ミクロデータは，QID と SA 属性が明確に区分されており，各属性値の表示バイト数も指定されていたことから，内部ロジックとして持っていたデータの正規チェックシステムの多くが不必要となったが，社会実装に向け，より多様なデータの正規化チェックシステムが求められる．

4) 評価指標が 13 個あり，各値の順位で評価 については，最も改修が多いものであった．プロトタイプにおいては，匿名化データは全て 1 つの評価指標（k-匿名性）によって評価される仕組みで設計されており，複数，それも 13 個の指標による評価は想定しておらず，評価システムは全て再設計された．

匿名化データの評価プラットフォームの簡易機能構成図を図 61 にて示す．本システムでは，評価システム，ファイル管理システム，共有 DB，個別 DB 領域等を 1 サーバ～全て独立した環境に展開することも可能である．

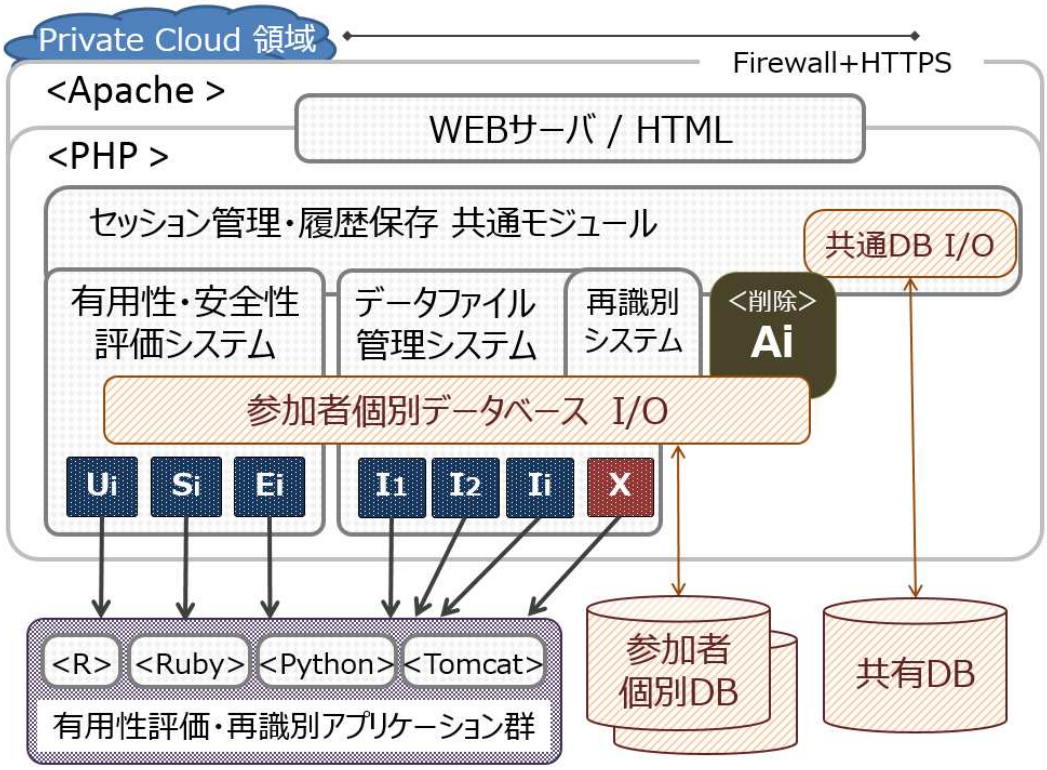


図 61 匿名化データの評価プラットフォームの機能構成図

また，本システムはクラウド内に設定されたプライベートクラウド領域における SaaS を志向して開発された．全体のセキュリティ要件として，Firewall やアクセス権限管理があるが，それらの要件は事業者ポリシー，サーバ構成，アクセスするユーザの役割等によって異なるため，本研究ではスコープ外とする．

通常の SaaS の場合，利用者数が増加するとサーバの CPU コア数やメモリ量を増加する形でのスケールを志向するが，本システムでは，各利用者による独立性を重視するため，個人環境のサーバ台数を増加する方向に設計されている．

クラウドサービスの発達により，近年ではハイスペックなサーバ性能を更に向上させるより，追加のサーバ台数を増加させる方が，コストパフォーマンスが良い場合がある．

各 WEB サーバ環境や DB 環境がセキュアな状態で独立している場合，結合可能キー(行番号データ I^Y)や中間・削除ファイル Xi 等の管理は各 WEB サーバ内で行われている．個人情報に類する情報を KVM やファイルサーバ等の複数事業者の共有ストレージに設置することはセキュリティ上好ましくないため，この構成とした．しかし WEB サーバ上に設置することもセキュリティ上問題がある．プロトタイプでは全てのデータは個人情報として DB に格納しており，消去や撹乱，ハッシュ化等の操作が可能であった．これらファイル管理方法は今後も検討が必要である．

また, 増加するデータベースとWEBサーバの制御も共有DB内で行われており, ユーザは個人のサーバ環境を意識することなくセキュアな匿名化処理が可能である.

処理サーバが分散される場合, 各アプリケーションは軽量で運用性が高い方が望ましい. 近年では企業向けHPの作成にはWordPressが多く用いられており(w3techs.com調べ2016年1月CMSシェア58.8%), 運用性の高いApache + PHP + Mysqlの, 所謂 LAMP 環境での開発需要が高まっている. 本プラットフォームも LAMP 環境で作成されており, インターネットサービスとの親和性が高い.

評価システムの実装にあたり, 必要とされるミドルウェアとして, PHP, R 言語, Ruby, Python, Javaと, 合計 5 種類のプログラム言語がインストールされている. 今後もデータ型式の多様性に伴い, 評価用プロトタイプアプリケーションが増加することが予想される. そのため, 今後は評価プログラムとプラットフォームを抽象化するための共通 API 層が必要になるだろう.

最終的なサーバスペックとインストールされたミドルウェア等の情報を表 20 に示す. 予備選はWEB＋DBを含めた1台での構成. 本戦はWEB1台＋DB1台の2台構成で運営された.

表 20 システム仕様

種別	詳細
サーバ	ニフティクラウドサーバ
システム	ニフティ匿名化処理プラットフォーム
CPU	Intel(R) Xeon(R) 3.00GHz
Memory	4 GB (負荷に応じて可変)
Apache	Apache/2.2.3
PHP	PHP 5.5.30
R 言語	R version 3.2.0
Java	Java(TM) SRE 1.8.0 45
Ruby	ruby 2.2.2p95
Python	Python 2.7.10
Mysql	mysql Ver 14.14 Distrib 5.6.15

6.2.4 匿名化データの評価指標に関する課題

公開実験を通じた, 匿名化データの評価指標の検証を通じて得られた, 社会実装に向けた課題を検討する. 本実験を通じて設定された評価指標として, Ui:有用性評価,

Si:安全性評価(k-匿名性), Ei:安全性評価(再識別率), 計 13 指標が表 21 として設定された. 各指標に対応した評価アプリケーションは, 実行委員が作成し, プラットフォーム側からの処理要求によって稼動する.

本指標は以下の 3 種類に分類できる.

1)匿名化データ Y だけで評価する指標

2)元情報 X と匿名化データ Y を比較した指標

3)元情報 X, 匿名化データ Y, 行番号 I^Y を利用する指標

本システムが社会実装された場合, 上記区分は大きな意味を持つ.

表 21 公開実験で利用した指標

No	指標	指標説明(使用言語)	必要情報
U1	meanMAE	SA 平均絶対誤差(R)	X,Y
U2	crossMean	クロス集計値の平均絶対誤差	X,Y
U3	crossCnt	クロス集計数の平均絶対誤差	X,Y
U4	corMAE	SA の相関係数の平均絶対誤	X,Y
U5	IL	データ各値の平均絶対誤差	X,Y,I^Y
U6	nrow	データのレコード数(システム)	Y
S1	k-anony	k-匿名性指標の最小値(R)	Y
S2	k-anonyMean	k-匿名性指標の平均値(R)	Y
E1	IdRand	QI からランダムな再識別率	X,Y
E2	IdSA	QI から SA15 列による再識別	X,Y
E3	Sort	SA 総和ソートによる再識別率	X,Y
E4	SA21	SA21 列について再識別率	X,Y
E5	AYA	山岡匿名化における検知率	X,Y, I^Y

匿名化データ Y を評価し, 流通させるだけの場合, 個別のデータ作成者と個人情報取扱契約を結ぶ必要はない. しかし, 2) 3)の指標が必須になる場合, プラットフォーム事業者は複数の個人情報取扱事業者との間で契約を交わす必要があり, 情報の管理と授受に関わる要件が大きく異なる.

本公開実験では, 数値の絶対評価を用いず, 最終的なスコアの決定に相対的な順位(Rank)を用いた. しかし, これらの評価システムを社会実装する際には, 他者との相対評価よりも絶対的な数値評価が必要となる. データを利用する目的に合わせて評価数値を設定し, 今後の評価システムに適用する必要がある.

また, 本システムは, 匿名化データの流通と評価を行うプラットフォームであるため, 内部に匿名加工アルゴリズムを持たず, 公開実験参加者によって作成された匿名化データを投入し, 評価するシステムである. しかし, 社会実装に向け, 匿名化処理の実行中に発生した不正要件を検知する仕組みが求められる.

そこで，匿名化データを提供する際に，あらかじめ定められた安全性基準を達成しているか，または，データに対して過度な改ざん等を行っていないか，を確認することを意図してシステム履歴機能を設計した．

しかし，公開実験では，各参加者の開発環境にて匿名加工処理を行われた後の，結果データのみを投入する仕組みであるため，投入された結果データのみから不正を検知することが困難であることが判明した．

その例として，ユーザによる山岡匿名化についての対応を紹介する．山岡匿名化とは，匿名加工データと，その再識別時の解答となる行番号データを作成する際に，行番号データだけを故意にランダム化して並び替える処理を指す．本処理によって，匿名化データの平均絶対誤差や k-匿名性を変化させずに，他の参加者からの再識別だけを防止することが出来る．

本問題は U5(IL 指標)にて検証できることが予想されていたが，登録されたデータを検証したところ，山岡匿名化を行っているデータに対して特徴的な傾向を見出すことが出来なかった．図 62 にその有用性指標の分布を示す．

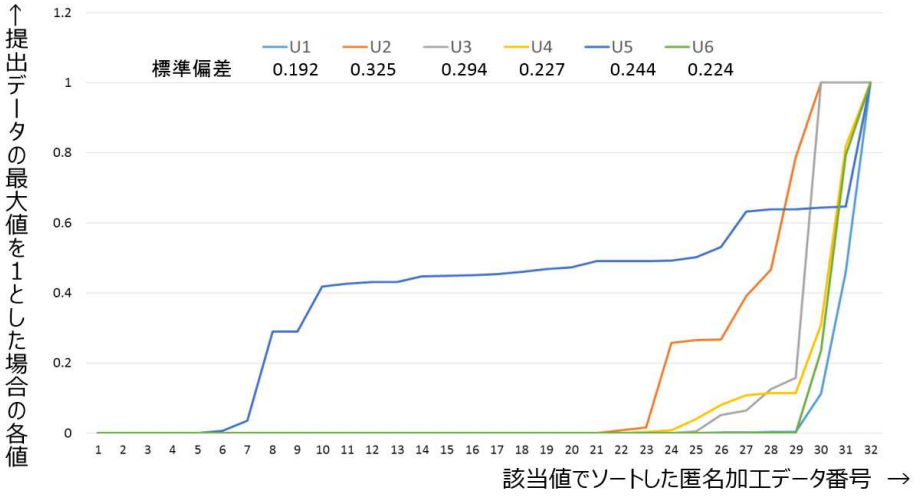


図 62 結果データの有用性指標分布

そこで，山岡匿名化に対する完全な検知が不可能と判断し，新たな指標 AYA を作成し，検出を試みた．結果として，本戦では IL 指標，AYA 指標ともに機能したが，今後も想定していないデータの不正処理が発生した場合の対応が継続的に必要となる．

本問題は，提出された匿名加工データについて，データ作成者が偽装を行うことによって，プラットフォーム側が検出することなく市場に流れる可能性を示唆している．

本来ならば，匿名加工データは個人情報保護委員会または認定個人情報保護団体が作成する規則等が定める基準に従い適切に加工されることが前提となっているが，もし指定の処理を採用しなかった場合，プラットフォーム側での検知を行う仕組みを必要とする．その際，プラットフォーム側に処理の手順等が示されていない場合，検知が非常に難

しい. 各業種業態に特徴的なデータが存在するように, その検知にも独自のアルゴリズムが必要になる.

6.3 本章のまとめ

本章では, 匿名化データを流通する上で必要となるプラットフォームの社会実装に向けた課題について検討した.

6.1 章では, 4 章で提案した累乗近似式をデータ利用者に提供することで, 元となるパーソナルデータの分布を公開せず, 予測式から匿名化処理のリクエストを行う方式について検討した. このシステムの実現によって, データ作成者側の処理負荷が減少する. しかし, 予測式の精度が高い場合は, データ作成者とデータ利用者の処理に問題が発生しないが, 精度が低い場合に発生する情報漏えいリスクが残ることが判明した.

6.2 章では, 匿名化処理プラットフォームを社会実装する上でのシステム要件と, 匿名化データの評価プラットフォームに発生する課題について検討した. プロトタイプ開発を通じて得られた社会実装に向けての課題として, 一般顧客への情報公開, 及び統一的な安全性指標について検討し, 他の研究者による再識別による評価方式の実装可能性について検討した. しかし, それらの指標を悪用するための手法も同時に判明しており, 社会実装に向け, 不正検知などの仕組みが必要とされていることを確認した.

7 匿名化処理に関する議論

本章では，各章における残課題や研究の展開について述べる．

7.1 第 3 章:k-匿名性減少特性に関する議論

第 3 章ではインターネットサービスに所属する顧客の分布を参考にしたデータに対して，同一の属性区分処理を行った後の k-匿名性を計測した．本実験で利用した一般化階層，及び属性値の選択に関する課題が存在する．

本データの元となったパーソナルデータは，本来 60 以上の属性情報を保持しており，本研究では，その中から，個人情報として法律等に明文化される 4 属性（氏名，性別，年齢，住所）から，氏名以外の属性を用いて分析を行った．これらの基本属性は，分布が想定しやすく，多くが正規分布に近似することが経験的に判明している．

しかし，それ以外にも多くのカテゴリ属性と数値属性が存在しており，現実的にはそのうちのどの属性を利用するかは決まっていない．

特に，商品名や金額など，多様な種類を持つ属性値に関しては，ポワソン分布など，他の分布に近似する属性値も多く存在する．分布の性質が異なる属性値を組み合わせた結果から発生する k-匿名性の変化の計測と，その性質の調査に関しては，今後の検討課題と位置づけている．

また，パーソナルデータに数値属性が含まれる場合，その処理方法は多様である．例えば，2.5.2 章では，値の書き換え手法について述べており，一般的に数値属性を処理する際には，サンプリングや行追加，ノイズ付与，裾切りなどの処理を行う．

しかし，年齢や所得などの数値による区分は，そこに意味的な区分が適用される場合もある．例えば，国政選挙，自動車免許，公営ギャンブルなどに関連する調査を行う場合は「18 歳以上」という具体的な年齢の区切りが必要となる．企業などが調査を行う場合，独自に作成している年齢区分によって実施する場合があることから，区分する基準は無数に存在しており，それら全ての概念を包括することは困難である．

本研究では年齢，性別，地域の 3 属性に対して，分析等に利用されている一般化階層を用いて検討を行ったが，更に多くの属性を交えた検証を行う必要がある．

そのため，社会に存在する年齢区分等の種類について多くのサンプルを収集し，最も利用されている区分手法を採用することを検討したが，公平な判断方法が見つからなかった．今後，ある属性について，世間においてより多く利用されている，または，最も優れていると判定できる一般化階層自体を検証，評価する手法について，検討が必要である．

7.2　第 4 章:k-匿名性の予測近似式についての議論

4 章にて検討した k-匿名性の予測式は，本来，企業の中において匿名化データを効率的に作成するために用いた手法である．そのため，正確性よりも概要を示すことを目的とし，属性の種類によらずに k-匿名性の予測を行えるよう，処理の軽量化を目指した．

実際には，選択した属性によって，k-匿名性の変化が大きい箇所と，小さい箇所が発生することから，属性ごとの k-匿名性変化の波を予測し，波の合成の結果として，詳細な k-匿名性を予測する手法も検討した．

しかし，本手法は，一般化階層によって区分される属性値の数値が，連続分布ではなく，一般化階層の結果によって発生する独立した結果の集合であることから，波を表す式を作成した際に，定式化できない箇所が出ることから断念した．

逆に，生成された波の性質に合わせて，一般化階層を調整することも可能だが，そのような行為は技術側の都合による条件の変更であることから，実際の調査ユースケースとしてはそぐわないものである．

そのため，本章では，一般化階層を固定した状態で，どこまで予測式の精度を高めることが可能であるか，という点に閉じたアプローチとなっている．

最も利用価値が高い予測式は，未知の一般化階層と既存の一般化階層を比較し，予測式による類似性の比較を行ったうえで処理を行う手法である．これは，3 章の課題である，一般化階層自体の評価と共に今後の検討課題と考えている．

7.3　第 5 章:累乗近似式を用いた匿名化処理アルゴリズムについての議論

第 5 章にて提案した匿名化処理アルゴリズムは，少量のサンプルによる計算によって，予測値を出力できることから，小規模の研究者，事業者による匿名化データ生成に係る負担を軽減することを目的として検討したものである．

本アルゴリズムでは，一度作成した予測近似式のみをデータベースに格納しておき，他のデータを出力する際にその近似式を再利用することで効率化が実現できる．

本近似式は，その精度が成功率として明確に定義できることから，今後，機械学習等のアプローチによって匿名化処理を行う場合，匿名化処理が可能かを判定する手法に展開が可能である．

その場合においては，アルゴリズムにおける「匿名化可能か否かの予測」を行う部分を，機械学習による結果に置き換えることで，本方式だけではない，他の方式のよる効率化処理を同時に利用することが可能になると考える．複数の手法の組み合わせによる効率化手法については，今後の検討課題である．

7.4　第 6 章:匿名化データの流通プラットフォームの議論

第 6 章で提案したプラットフォームの仕組みを確立させるためには多くの課題が存在する. 本研究では, 匿名化データ作成リクエストに伴う情報漏えいリスクの計測手法について検討したが, 実際には有用性の指標や利用している一般化階層の種類, 及び, 5 章における予測近似式の数式などの数値が存在する.

基本的に匿名化データは, その用途に応じて作成されることから, 特殊な一般化階層などを適用するため, 他のデータの接続が困難である. 他のデータとの接続, 比較が出来ないことによって, 含まれるデータの検索性や, 不正データの検知システムなど, 関連した業務にも影響が出る. 全ての匿名化データに対して, それぞれの一般化階層の特徴に応じた検索エンジンや不正検知システムなどの開発は困難である.

しかし, 本研究で提案している予測式は, ある性別, 年齢などの属性に対して, 一律の一般化階層を適用した際の結果を表示していることから, 他のデータと一律で比較可能な指標として利用できる. 事前に評価されたデータの近似式を比較することによって, そのデータの持つ属性の分布傾向だけでなく, 有用性や不正利用などの情報と紐付けることが期待できる.

例えば, 全く同じ傾向の予測式を持つデータが投入された場合, 過去のデータを再投入している, または, 他社データのコピーを利用している可能性がある. また, 予測式では k-匿名性が低いものしか実現できない, と予測されたにも関わらず, 他社に提供したデータの k-匿名性が非常に高く維持されていた場合, データの不正な改ざんが行われている可能性がある.

逆に, 類似性が高いデータ同士は, 同様の匿名化処理を行った際に詳細なデータに区分が可能である可能性が高いことから, 共同調査のマッチングを行うなど, 利用を促進するための仕組みにも展開が可能である.

しかし, このような分野の研究には多くの匿名化データのサンプルが必要であることから, プラットフォームの有用性を高め, 研究対象となるデータを増加させることが求められる. 今後のパーソナルデータ流通の状況に伴い, これらの課題について検討していきたい.

8 結論

本研究の内容をまとめる.

第 1 章にて, 研究の背景をまとめた. 現代においてパーソナルデータの価値は高く評価されているが, 同時にプライバシー侵害に対する問題も多く提起されている. 多くの国・地域ではパーソナルデータに対して規則等を強化しているが, 常に対応できるとは限らない. そのため, 状況に対応して安全性を高め, かつ利用者が求める有用性を保持したデータを流通させる, プライバシー保護データパブリッシングが必要とされている. その大きな技術的要素が匿名化処理技術である. 匿名化データを第三者提供する際には, 匿名化処理の要件を整理したオーダーメイド型匿名化処理が必要だが, 有用性と安全性を両立させる匿名化処理アルゴリズムが存在しないという課題がある.

第 2 章にて, 匿名化処理の従来研究を検証した.

まず, パーソナルデータの安全性の評価のため, 代表的な 4 つのパーソナルデータに対する攻撃モデルについて, 関連する安全性指標を検討し, 個人の識別性を低減させる安全性指標である k-匿名性とその発展指標について概説した. その結果, 匿名化データにおける安全性指標には, まず k-匿名性を満たした上で, 追加的に検証を行うべき指標が多く存在することを確認した. そのため, 効率的に k-匿名性を満たす属性の組み合わせを探索することで, 他の指標の検証効率も向上すると言える. そこで, k-匿名性を満たす属性組み合わせを効率的に探索する匿名化処理アルゴリズムについて概説し, 有用性を保つ処理方式について検討した.

その結果, 従来研究は元情報の持つ分散や特徴に応じて処理の効率が異なり, そのデータに対しての最悪ケースに至る方式を選択した場合, 安全性検証に係る検証回数が増加することを確認した.

これらの検討から, 現状の匿名化処理における問題点より, 以下の 3 点を抽出した.

1) 匿名化データの安全性基準の策定

2) パーソナルデータの匿名化処理リソースの軽減

3) 実社会に即した匿名化データの流通方法

そこで, 匿名化処理がどの地点で達成されるかを予測する手法を用いた, 属性組み合わせ探索処理の効率化の観点から詳しく検討した.

第 3 章では, 問題点 1) における, 安全性基準を検討するための基礎データとして, 擬似サービス群に対して, 一律の匿名化処理を行った場合に達成できる安全性を検証した.

その結果, 10 万人以上の登録ユーザがいる場合でも, 3 属性を組み合わせた結果, 平均の k 値は 6.6 と, 非常に低い値となった. 検証を行った結果, k 値が高く出るサービ

ス群は，国勢調査との相関が高く，また，k 値の減少傾向を元にした直線近似式は，精度が低く，値の予測に利用することが出来なかった．

第 4 章にて，具体的な k 値の減少傾向を，累乗近似式を用いて予測する手法について検討した．実データに即した疑似パーソナルデータを用いて検証したところ，累乗近似式による予測式は，他の分布の予測式よりも精度が高く，全ての群において平均 0.9 以上の重相関係数であった．そこで1635種類の擬似データを用いて検証したところ，累乗近似式は相関・回帰の数値は高いが，平均誤差量が大きいという問題点が解決できなかった．

そこで，問題点 2)として定義されている，匿名化処理コストの低減に活用する方式と合わせ，予測式を用いて匿名化処理コストを削減する方法を検討した．

第 5 章では，4 章にて提案した予測式を用いて，匿名化処理が達成可能な属性の組み合わせを予測し，その組み合わせから匿名化状態を検証するアルゴリズムを提案した．従来の匿名化処理アルゴリズムは，詳細な群から匿名化処理を試行するボトムアップ方式と，抽象的な群から匿名化処理を試行するトップダウン方式に分類される．提案するアルゴリズムは，予測地点における匿名性を検証した結果が，求める匿名性を満たす場合は，その地点からトップダウンの匿名化処理を選択し，満たさない場合はボトムアップ型の匿名化処理を選択することで，匿名化状態の検証回数を削減する．

累乗近似式による匿名化処理選択方式(PAK 方式)は，一般的なボトムアップアルゴリズムと比して 3.5%，トップダウンアルゴリズムと比して 12.5%の処理量で匿名化処理結果を出力することが出来た．これらの結果は，各パーソナルデータのユーザ分布に依存するが，実験により，本方式は匿名化処理が難しい，分散が偏っているパーソナルデータに対しても処理削減効果を発揮することを検証した．

第 6 章では，これらの技術を搭載するべき情報流通システムを検討するためのプラットフォームについて主に検討した．まず，提案する予測式を用いて匿名化処理の効率化が可能になる研究結果を受け，それを実装するべきプラットフォームを提案した．

具体的には，匿名化データの利用者に対して，属性区分数と k-匿名性に関する予測式を提供することで，利用者が求める属性区分によって匿名化処理が達成できるかを，事前に検証出来る仕組みを提案した．これにより，元となるパーソナルデータを漏洩せずにデータ授受における折衝を効率化できる．

また，これらのプラットフォームを社会実装する上で必要な技術的課題を，プロトタイプの作成と，匿名化データの流通と，その再識別によって安全性を評価するプラットフォームとして検討した．それにより，今まで計算上のリスク問題であった再識別可能性について，実際に再識別される実数値に変換することで統一的な指標として採用できることを確認した．

本研究の成果は，実サービスの分布を反映した疑似パーソナルデータ群を用いて，属性値によるクラスタリングを行った場合の k-匿名性の推移を検証し，相関係数の高い予測式を提案したことにある．加えて，その予測式を用いて，匿名化処理アルゴリズムを選択する手法を提案し，元情報の分布の特徴に依存せず，匿名化処理が効率化できることを実証したことにある．これにより，元情報に含まれるデータの特徴を公開せずに，データ利用者の求める一般化階層を事前に評価するプラットフォームの社会実装を可能とした．

謝辞

本論文の作成にあたり，大変多くの方の御協力を賜りました．

はじめに，本論文の主査，並びに主任指導教官でもありました，国立情報学研究所/総合研究大学院大学の曽根原 登教授におかれましては，朝晩を問わず，御指導，御鞭撻を賜りましたことに深く感謝致します．何よりも，先生より繰り返し頂きました，研究者としての正しい精神の持ち方，生き方に関する御言葉は，今後の人生の指針とさせて頂きます．謹んで御礼申し上げます．

また，プライバシーと匿名化処理に関する研究分野を深く進めて行く中で，多くの方から御指導を賜りました．副主任指導教官でもありました越前 功教授には，セキュリティに関する様々な知見から御指導を賜りました．同じ研究室に在籍しておりました Aghdam Rasool 氏には，匿名化処理の根幹に関する御討論，御協力を賜りましたことで，研究の視野を広く持つことが出来ました．また，神門 典子教授より賜りました御指導によりまして，本研究分野の研究方針を定めることが出来ました．心より御礼申し上げます．

また，本論文をまとめるにあたり，国立情報学研究所の山田 茂樹教授より分析不足の点につきまして御指摘，御指導賜りました．国立情報学研究所／総合研究大学院大学の計 宇生教授におかれましては，研究内容に本質に関する課題の明確化について御指摘，ご指導賜りました．津田塾大学の小舘 亮之教授におかれましては，大変お忙しい中，お時間を割いて頂き，論文の細かな箇所まで御指摘・御指導賜りました．皆様の御指導，御鞭撻により，本論文の質を高めることができました．重ねて御礼申し上げます．

また，本研究に取り組む機会を与えて頂きました，ニフティ株式会社の多くの方々に感謝を申し上げます．松井 くにお氏には，進学にあたっての御推薦，本研究への御指導，また公私共への御配慮など，長期間に亘り支えて頂けましたことに，深く感謝致します．加えまして，黒政 敦史氏には，夜遅くまで御指導賜り，論文の完成にも御協力頂きました．御二方と継続して匿名化処理の社会的な意義に関する御討論を賜りました経験が，研究を進める大きなモチベーションとなっております．また，御退職なされました須田 智紀氏には，社会とプライバシーに関する様々な示唆を賜りました．研究・開発を進める上で業務のサポートを頂きました木村 孝氏，阿部 妙子氏，加藤 奈美氏，廣川 謙氏，その他支えて頂いた多くのメンバーにも，御礼を申し上げます．

最後に，自分の長年の夢でもあった進学を認め，何度も失敗，挫折した状況にも関わらず，生活を支えてくれた，妻 直子に感謝致します．

参考文献

[1] 田中 裕, "パーソナルデータについて", 日本データ通信 No.196, pp.26-27, (2014).

[2] 総務省, "第 4 節(1) 我が国におけるビッグデータ流通量の推計", 平成 27 年版情報通信白書, (2016).

[3] 高木 浩光, "情報取得手段ごとに相当な同意確認基準の提案", 総務省 利用者支店を踏まえた ICT サービスに係る諸問題に関する研究会, (2012).

[4] 高井 正三, "ビッグデータとコグニティブ・コンピューティングが変える世界", 富山大学総合情報基盤センター広報, 13 pp.28-34, (2016).

[5] 株式会社 MM 総研, "国内クラウドサービス需要動向(2015 年版)", マルチクライアント・レポート, (2015).

[6] 石井夏生利, "個人情報保護法の現在と未来", 勁草書房, (2014).

[7] 経済産業省 商務情報政策局, "パーソナルデータに関する海外動向", 経済産業省, (2012).

[8] 個人情報の安心安全な管理に向けた社会制度・基盤の研究会, "個人情報の安心安全な管理に向けた社会制度・基盤の研究会 報告書", JIPDEC(一般財団法人 日本情報経済社会推進協会), (2011).

[9] 個人情報の安心安全な管理に向けた社会制度・基盤の研究会, "「個人データ保護規則」案 仮訳", JIPDEC(一般財団法人 日本情報経済社会推進協会), (2011).

[10] 個人情報の保護に関する法律及び行政手続における特定の個人を識別するための番号の利用等に関する法律の一部を改正する法律(平成 27 年法律第 65 号)

[11] 鈴木 正朝, "プライバシーの権利と個人情報保護法", マイナンバーシンポジウム資料, (2012).

[12] 南 和宏, "プライバシー保護データパブリッシング", 情報処理, 54(9), pp.938-946, (2013).

[13] Fard, A.M., "Privacy Preserving Web Query Log Publishing: A Survey on Anonymization Techniques", arXiv preprint arXiv:1211.2354, (2012).

[14] 武田 英明, "リンクするデータ(Linked Data)-広がり始めたデータのクラウド-：6.日本における Linked Data の現状と普及に向けた課題", 情報処理, 52(3), pp.326-333, (2011).

[15] 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), pp.1204-1210, (2013).

[16] 財団法人日本情報処理開発協会 パーソナル情報認証スキーム検討委員会, "パーソナル情報の利用のための調査研究", 平成22年度情報化推進に関する調査研究等補助事業報告書, (2011).

[17] Hundepool, A., Willenborg, L. and Statistics Netherlands, "m-and t-ARGUS: Software for Statistical Disclosure Control", Record linkage techniques--1997 Proceedings of an International Workshop and Exposition. March, pp.142-149, (1997).

[18] 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", ERATO 湊離散構造処理系プロジェクトセミナー, (2011).

[19] Emam, K. E., Arbuckle, L., "データ匿名化手法 –ヘルスデータ事例に学ぶ個人情報保護", 株式会社オライリー・ジャパン, (2015).

[20] El Emam, K. and Arbuckle, L., "Anonymizing health data: case studies and methods to get you started", O'Reilly Media, Inc., (2013).

[21] 厚生労働省 保険局 医療介護連携政策課 保険システム高度化推進室, "レセプト情報・特定健診等情報データベース(ＮＤＢ)について", 厚生労働省, (2014).

[22] 森川 博之, "ビッグデータの活用に関するアドホックグループの検討状況", 総務省 ビッグデータの活用に関するアドホックグループ, (2014).

[23] 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), pp.188-192, (2013).

[24] Sweeney, L., "k-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), pp.557-570, (2002).

[25] Sweeney, L., "Achieving K-anonymity Privacy Protection Using Generalization and Suppression", Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5), pp.571-588, (2002).

[26] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M., "l-diversity: Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), pp.3, (2007).

[27] Li, N., Li, T. and Venkatasubramanian, S., "t-closeness: Privacy beyond k-anonymity and l-diversity", Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on, pp.106-115, (2007).

[28] Meyerson, A. and Williams, R., "On the complexity of optimal k-anonymity", Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp.223-228, (2004).

[29] Aggarwal, C.C., "On k-anonymity and the curse of dimensionality", Proceedings of the 31st international conference on Very large data bases, pp.901-909, (2005).

[30] (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", 経済産業省 平成 23 年度次世代高信頼・省エネ型 IT 基盤技術開発・実証事業報告, (2012).

[31] Lichman, M., "UCI Machine Learning Repository", University of California, Irvine, School of Information and Computer Sciences, (2013).

[32] 秋山 裕美, 山口 幸三, 伊藤 伸介, 星野 なおみ, 後藤 武彦, "教育用擬似ミクロデータの開発とその利用 平成 16 年全国消費実態調査を例として", 統計センター製表技術参考資料, (2012).

[33] 亀本 信康, 齋藤 敦, “統計データの二次的利用に関する統計センターの取組状況", 2013 年度 統計関連学会連合大会, (2013).

[34] Health Insurance Portability and Accountability Act(HIPAA), "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule", U.S. Department of Health & Human Services, (2012).

[35] 「パーソナルデータに関する検討会」技術検討ワーキンググループ, "技術検討ワーキンググループ報告書", 内閣官房情報通信技術 IT 総合戦略室, (2013).

[36] 中川 裕志, "プライバシー保護入門:法制度と数理的基礎", 勁草書房, (2016).

[37] ARTICLE 29 DATA PROTECTION WORKING PARTY, "Opinion 05/2014 on Anonymisation Techniques", European Commission, 0829/14/EN WP216, (2014).

[38] EPAG, “White Paper on Key-Coded Data as a Data Protection Best Practice”, EUROPEAN PRIVACY ADVISARY GROUP, (2013).

[39] 松前恵環, "個人識別性/識別可能性といわゆる「FTC3 要件」", 堀部政男情報法研究会第 9 回シンポジウム, (2013).

[40] 佐藤 慶浩, “米国消費者プライバシー権法検討素案におけるパーソナルデータの定義”, URL: http://yoshihiro.cocolog-nifty.com/postit/2015/03/post-0de3.html, (2015).

[41] The White House, "Administration Discussion Draft: Consumer Privacy Bill of Rights Act", (2015).

[42] 藤井秀之, "パーソナルデータの匿名化に関する米欧の議論動向~最新公表レポートから", 情報通信総合研究所 Infocom Law Report 2014, (2014).

[43] President's Council of Advisors on Science and Technology(PCAST), "BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE", Executive Office of the President of the United States, (2014).

[44] Fung, B., Wang, K., Chen, R. and Yu, P.S., "Privacy-preserving data publishing: A survey of recent developments", ACM Computing Surveys (CSUR), 42(4), pp.14, (2010).

[45] Nergiz, M.E., Clifton, C. and Nergiz, A.E., "Multirelational k-anonymity", Knowledge and Data Engineering, IEEE Transactions on, 21(8), pp.1104-1117, (2009).

[46] Wang, K. and Fung, B., "Anonymizing sequential releases", Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.414-423, (2006).

[47] 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", コンピュータセキュリティシンポジウム 2009 (CSS2009) 論文集, 2009 pp.1-6, (2011).

[48] 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", コンピュータセキュリティシンポジウム 2011 論文集, 2011(3), pp.450-455, (2011).

[49] Poulis, G., Loukides, G., Gkoulalas-Divanis, A. and Skiadopoulos, S., "Anonymizing data with relational and transaction attributes",

Machine learning and knowledge discovery in databases, pp.353-369, (2013).

[50] Poulis, G., Skiadopoulos, S., Loukides, G. and Gkoulalas-Divanis, A., "Apriori-based Algorithms for Km-anonymizing Trajectory Data", Trans. Data Privacy, 7(2), pp.165-194, (2014).

[51] Mohammed, N., Fung, B., Hung, P.C. and Lee, C., "Anonymizing healthcare data: a case study on the blood transfusion service", Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.1285-1294, (2009).

[52] Xu, Y., Wang, K., Fu, A.W. and Yu, P.S., "Anonymizing transaction databases for publication", Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.767-775, (2008).

[53] Tanjo, T., Minami, K., Mano, K. and Maruyama, H., "Evaluating data utility of privacy-preserving pseudonymized location datasets.", JoWUA, 5(3), pp.63-78, (2014).

[54] 有住 なな, 丹生 智也, 南 和宏, 丸山 宏, "仮名化データの安全性検証アルゴリズムの考察", マルチメディア, 分散協調とモバイルシンポジウム 2014 論文集, 2014 pp.823-827, (2014).

[55] Tanjo, T., Minami, K., Mano, K. and Maruyama, H., "On Safety of Pseudonym-Based Location Data in the Context of Constraint Satisfation Problems", Information and Communication Technology, pp.511-520, (2014).

[56] Mano, K., Minami, K. and Maruyama, H., "Protecting location privacy with k-confusing paths based on dynamic pseudonyms", Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on, pp.285-290, (2013).

[57] 村本 俊祐, 上土井 陽子, 若林 真一, "プライバシ保護データ公開に向けた l-多様化適性の評価", 情報処理学会論文誌データベース(TOD), 4(2), pp.126-141, (2011).

[58] 山岡 裕司, 伊藤 孝一, 牛田 芽生恵, 津田 宏, "プライバシー保護データ開示における 2-多様性を満たす機微データ曖昧化法(情報セキュリティ基礎)", 電子情報通信学会論文誌. A, 基礎・境界, 97(3), pp.197-208, (2014).

[59] Wong, R.C., Li, J., Fu, A.W. and Wang, K., "(α, K)-anonymity: An Enhanced K-anonymity Model for Privacy Preserving Data Publishing", Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.754-759, (2006).

[60] 五十嵐 大, 千田 浩司, 高橋 克巳. "Pl-多様性:属性推定に対する再構築法のプライバシーの定量化", コンピュータセキュリティシンポジウム2010(CSS2010)論文集, pp.813–818, (2010).

[61] Nergiz, M.E., Atzori, M. and Clifton, C., "Hiding the presence of individuals from shared databases", Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pp.665-676, (2007).

[62] Chawla, S., Dwork, C., McSherry, F., Smith, A. and Wee, H., "Toward privacy in public databases", Theory of Cryptography, pp.363-385, (2005).

[63] Xiao, X. and Tao, Y., "M-invariance: towards privacy preserving re-publication of dynamic datasets", Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pp.689-700, (2007).

[64] Dwork, C., "Differential privacy", Automata, languages and programming, pp.1-12, (2006).

[65] Xu, J., Wang, W., Pei, J., Wang, X., Shi, B. and Fu, A.W., "Utility-based anonymization using local recoding", Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.785-790, (2006).

[66] Aghdam, M.R.S. and Sonehara, N., "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).

[67] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R., "Incognito: Efficient full-domain k-anonymity", Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pp.49-60, (2005).

[68] 伊藤 伸介, 村田 磨理子, 高野 正博, "ミクロデータにおける匿名化技法の適用可能性の検証 : 全国消費実態調査と家計調査を用いて", 統計研究彙報, (71), pp.83-123, (2014).

[69] Sweeney, L., "Guaranteeing anonymity when sharing medical data, the Datafly System.", Proceedings of the AMIA Annual Fall Symposium, pp.51, (1997).

[70] El Emam, K., Dankar, F.K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J., Walker, M., Chowdhury, S., Vaillancourt, R. and others, "A globally optimal k-anonymity method for the de-identification of health data", Journal of the American Medical Informatics Association, 16(5), pp.670-682, (2009).

[71] 村本 俊祐, 上土井 陽子, 若林 真一, "データを極小歪曲し k-匿名性を保持したデータに変換するプライバシー保護アルゴリズム", 日本データベース学会 letters, 6(1), pp.97-100, (2007).

[72] 原田 邦彦, 佐藤 嘉則, "一般化階層木の自動生成と情報エントロピーによる歪度評価を伴う k-匿名化手法", 研究報告コンピュータセキュリティ (CSEC), 2010(47), pp.1-7, (2010).

[73] He, Y. and Naughton, J.F., "Anonymization of set-valued data via top-down, local generalization", Proceedings of the VLDB Endowment, 2(1), pp.934-945, (2009).

[74] Aghdam, M.R.S. and Sonehara, N., "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).

[75] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R., "Mondrian multidimensional k-anonymity", Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on, pp.25-25, (2006).

[76] 新井 淳也, 鬼塚 真, 塩川 浩昭, "クラスタリングと空間分割の併用による効率的な k-匿名化", DBSJ Japanese journal=日本データベース学会和文論文誌, 13(1), pp.72-77, (2014).

[77] Aggarwal, C.C. and Philip, S.Y., "A general survey of privacy-preserving data mining models and algorithms", Springer, (2008).

[78] Oguri, H. and Sonehara, N., "A K-Anonymity Method Based on SEM (Search Engine Marketing) Price of Personal Information", Social Computing (SocialCom), 2013 International Conference on, pp.1011-1015, (2013).

[79] 小栗 秀暢, 曽根原 登, "RL-005 個人情報の SEM(検索エンジン広告)価格に基づいた k-匿名化手法の提案(L 分野:ネットワーク・セキュリティ,査読付き論文)", 情報科学技術フォーラム講演論文集, 13(4), pp.25-32, (2014).

[80] Oguri, H. and Sonehara, N., "A k-anonymity method based on search engine query statistics for disaster impact statements", Availability, Reliability and Security (ARES), 2014 Ninth International Conference on, pp.447-454, (2014).

[81] 永井 康彦,五十嵐 亮基,羅川 広行,松岡 健,藤田 麻里央,佐藤 祥太郎,美馬 正司, "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「匿名化クラウド」のアーキテクチャ提案", 研究報告情報セキュリティ心理学とトラスト(SPT),2011-SPT-1(26), pp.1-8, (2011).

[82] 千田 浩司,五十嵐 大,高橋 克巳,濱田 浩気,冨士 仁, "集合匿名化クラウドの課題と対策", 研究報告インターネットと運用技術(IOT), 2011-IOT-13(21), pp.1-6, (2011).

[83] 次世代高信頼・省エネ型ＩＴ基盤技術開発・実証事業, "レセプト情報等利活用に関する調査・検証 H23 年度事業報告書", 経済産業省, (2011).

[84] 高橋 翼, 側高 幸治, 豊田 由起, 竹之内 隆夫, 森 拓也, "大規模レセプトに対する匿名化システムの開発", 第 33 回医療情報学連合大会抄録集, (2013).

[85] 側高 幸治, 高橋 翼, 豊田 由起, 竹之内 隆夫, 森 拓也, "センシティブ情報からの個人特定を防ぐレセプト匿名化の検討", 第 33 回医療情報学連合大会抄録集, (2013).

[86] 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), pp.1135-1140, (2013).

[87] Nabar, S.U., Kenthapadi, K., Mishra, N. and Motwani, R., "A survey of query auditing techniques for data privacy", Privacy-Preserving Data Mining, pp.415-431, (2008).

[88] 菊池 浩明, 山口 高康, 濱田 浩気, 山岡 裕司, 小栗 秀暢, 佐久間 淳, "匿名加工・再識別コンテスト Ice & Fire の設計", コンピュータセキュリティシンポジウム 2015 論文集, 2015(3), pp.363-370, (2015).

[89] Kikuchi, H., Yamaguchi, T., Hamada, K., Yamaoka, Y., Oguri, H. and Sakuma, J., "Ice and Fire: Quantifying the Risk of Re-identification and Utility in Data Anonymization", The 30th IEEE International Conference on Advanced Information Networking and Applications (AINA 2016), (2016).

[90] Domingo-Ferrer, J., Ricci, S. and Soria-Comas, J., "Disclosure risk assessment via record linkage by a maximum-knowledge attacker", Privacy, Security and Trust (PST), 2015 13th Annual Conference on, pp.28-35, (2015).

[91] Martin, D.J., Kifer, D., Machanavajjhala, A., Gehrke, J. and Halpern, J.Y., "Worst-case background knowledge for privacy-preserving data publishing", Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on, pp.126-135, (2007).

研究成果

1. 査読付きジャーナル論文

[1]. 小栗 秀暢, 曽根原 登, 松井 くにお, Mohammad Rasool Sarrafi Aghdam, “累乗近似式を用いた k-匿名化処理の効率化”, 情報処理学会論文誌, pp.2034-2044, (Sep.2016). (査読付き, 研究論文)

2. 査読付き国際会議・国内会議論文

[1]. Oguri, H. and Sonehara, N., "A k-anonymity method based on search engine query statistics for disaster impact statements", Availability, Reliability and Security (ARES), 2014 Ninth International Conference on, pp.447-454, (Sep.2014). (査読付き, full paper)

[2]. Oguri, H. and Sonehara, N., "A K-Anonymity Method Based on SEM (Search Engine Marketing) Price of Personal Information", Social Computing (SocialCom), 2013 International Conference on, pp.1011-1015, (Sep.2013). (査読付き,poster)

[3]. 小栗 秀暢, 曽根原 登, “個人情報の SEM（検索エンジン広告）価格に基づいた k-匿名化手法の提案”, 情報科学技術フォーラム講演論文集, 13(4), pp.25-32, (aug.2014). (査読付き, 研究論文)

3. 学位論文に関連する特許

[1]. 小栗 秀暢, 特許 2014-202232, "減少係数算出装置, それを用いた匿名処理装置, 方法及びプログラム", (2014.9.30).

[2]. 小栗 秀暢, 特許 2015-125021, "判定方法, 判定装置及び判定プログラム", (2015.6.22)

4. 学位論文に関連する発表論文（査読なし）

[1]. 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", 研究報告コンピュータセキュリティ(CSEC), 2014(4), pp.1-8, (2014).

[2]. 小栗 秀暢, 曽根原 登, 松井 くにお, 黒政 敦史, Mohammad Rasool Sarrafi Aghdam, "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", コンピュータセキュリティシンポジウム 2015 論文集, 2015(3), pp.387-394, (2015).

追加資料

5 万人以上サービスにおける k 匿名性推移データ，及び予測式の回帰係数

予測式の回帰分析結果							
重相関 R	0.989	0.933	0.947	0.944	0.945	0.944	0.956
重決定 R2	0.979	0.871	0.897	0.892	0.893	0.892	0.915
補正 R2	0.978	0.865	0.892	0.887	0.888	0.887	0.911
有意 F	0.000	0.000	0.000	0.000	0.000	0.000	0.000
標準誤差	1841.1	7335.2	3280.5	3998.2	2970.9	2571.9	1726.3
観測数	24	24	24	24	24	24	24
実データ							
属性区分数	S0000	S0001	S0002	S0003	S0004	S0005	S0006
1	127794	350527	215122	208675	190105	140826	134721
2	54556	95033	48366	57003	42704	36823	28005
3	22782	595	1137	290	237	270	173
4	22392	25504	10982	15447	10395	9603	6118
5	15960	19226	5589	3568	5819	2255	1030
6	9858	156	385	111	63	83	31
9	4977	669	417	366	230	232	158
10	6915	5606	1895	1183	1817	751	255
12	4201	36	122	20	19	29	9
18	2330	128	76	70	43	48	24
27	332	5	4	1	1	1	1
36	3142	836	305	243	293	118	35
45	183	91	14	12	20	4	1
47	583	767	441	387	208	256	210
54	150	2	2	1	1	1	1
81	73	1	1	1	1	1	1
90	91	21	6	3	8	1	1
94	277	196	71	65	45	69	34
141	106	1	1	1	1	1	1
162	23	1	1	1	1	1	1
235	67	40	6	8	7	3	1
282	52	1	1	1	1	1	1
423	50	1	1	1	1	1	1
470	33	8	1	2	2	1	1
846	17	1	1	1	1	1	1

予測式の回帰分析結果							
重相関 R	0.959	0.954	0.955	0.940	0.935	0.966	0.892
重決定 R2	0.919	0.911	0.912	0.883	0.875	0.934	0.795
補正 R2	0.916	0.907	0.908	0.877	0.869	0.931	0.786
有意 F	0.000	0.000	0.000	0.000	0.000	0.000	0.000
標準誤差	1402.3	989.6	961.4	1833.7	810.4	1257.0	903.3
観測数	24	24	24	24	24	24	24
実データ							
属性区分数	S0007	S0008	S0009	S0010	S0011	S0012	S0013
1	122266	104523	102677	100040	95513	94485	92518
2	23443	15601	15313	25039	10648	23277	8839
3	55	60	60	110	61	69	382
4	4798	2548	2491	6639	33	4334	4
5	843	3396	3311	4143	3203	1643	4144
6	12	7	7	29	4	14	31
9	118	5	4	141	14	17	3
10	213	631	612	1216	483	445	711
12	3	2	2	6	1	3	1
18	18	2	1	28	1	3	1
27	1	1	1	1	1	1	1
36	35	79	75	198	3	52	1
45	15	1	1	19	1	1	1
47	134	2	2	100	1	1	1
54	1	1	1	1	1	1	1
81	1	1	1	1	1	1	1
90	6	1	1	5	1	1	1
94	21	2	1	25	1	1	1
141	1	1	1	1	1	1	1
162	1	1	1	1	1	1	1
235	1	1	1	7	1	1	1
282	1	1	1	1	1	1	1
423	1	1	1	1	1	1	1
470	1	1	1	2	1	1	1
846	1	1	1	1	1	1	1

予測式の回帰分析結果							
重相関 R	0.973	0.890	0.936	0.936	0.631	0.912	0.901
重決定 R2	0.947	0.792	0.877	0.877	0.399	0.832	0.811
補正 R2	0.944	0.782	0.871	0.871	0.371	0.825	0.803
有意 F	0.000	0.000	0.000	0.000	0.001	0.000	0.000
標準誤差	578.0	782.2	532.3	1292.8	1977.3	1706.0	1566.1
観測数	24	24	24	24	24	24	24
実データ							
属性区分数	S0014	S0015	S0016	S0017	S0018	S0019	S0020
1	82486	81788	79111	71214	70285	68922	65928
2	11945	7627	7080	17168	8167	18878	16240
3	46	48	73	87	229	96	236
4	1832	7	11	3591	65	5632	4279
5	1089	3444	2028	4392	9504	5873	6048
6	5	6	14	30	41	29	66
9	5	4	9	10	8	5	225
10	211	360	225	1272	1160	2095	1412
12	1	1	2	9	2	8	25
18	1	2	2	4	2	3	41
27	1	1	1	1	1	2	2
36	22	3	2	143	3	395	170
45	1	1	1	1	1	1	25
47	1	1	1	1	3	3	116
54	1	1	1	1	1	1	1
81	1	1	1	1	1	1	1
90	1	1	1	1	1	1	8
94	1	1	1	1	1	2	21
141	1	1	1	1	1	1	1
162	1	1	1	1	1	1	1
235	1	1	1	1	1	1	10
282	1	1	1	1	1	1	1
423	1	1	1	1	1	1	1
470	1	1	1	1	1	1	1
846	1	1	1	1	1	1	1

予測式の回帰分析結果							
重相関 R	0.949	0.957	0.949	0.958	0.950	0.957	
重決定 R2	0.900	0.916	0.901	0.917	0.903	0.915	
補正 R2	0.895	0.912	0.897	0.914	0.899	0.912	
有意 F	0.000	0.000	0.000	0.000	0.000	0.000	
標準誤差	339.9	824.7	420.3	531.6	886.1	853.3	
観測数	24	24	24	24	24	24	
実データ							
属性区分数	S0021	S0022	S0023	S0024	S0025	S0026	
1	64016	62826	58570	55164	53792	51254	
2	5100	13536	6367	8757	13406	13872	
3	44	137	36	32	73	50	
4	6	2741	10	1586	3381	3097	
5	1041	181	1173	1513	842	1497	
6	5	35	12	6	18	14	
9	8	110	11	3	75	9	
10	117	44	200	299	262	426	
12	4	9	2	1	2	2	
18	2	14	3	3	16	5	
27	1	2	1	1	1	1	
36	3	13	4	31	48	52	
45	1	1	1	1	3	1	
47	1	156	1	1	113	1	
54	1	1	1	1	1	1	
81	1	1	1	1	1	1	
90	1	1	1	1	1	1	
94	1	22	1	1	29	1	
141	1	1	1	1	1	1	
162	1	1	1	1	1	1	
235	1	1	1	1	1	1	
282	1	1	1	1	1	1	
423	1	1	1	1	1	1	
470	1	1	1	1	1	1	
846	1	1	1	1	1	1	