個人情報の識別リスク評価に基づいた匿名化に関する研究

明治大学大学院先端数理科学研究科

２０２１年度

博士学位請求論文

個人情報の識別リスク評価に基づいた匿名化に関する研究

A Study on Anonymization of Personal Data based on Re-identification Risk

学位請求者　先端メディアサイエンス専攻
伊藤　聡志

あらまし

機械学習や AI 技術の発展により，ビックデータの利活用が企業・医療機関・金融機関など多様な場面で盛んになっている．移動履歴や購買履歴などのビッグデータを分析することによって我々は様々な知見を得ることができ，それを利活用することによって大きな利益を生むことができる．しかし，ビッグデータは有用であるのと同時に危険なものでもあり，例えば，2006 年には，Sweeny らによって ZIP コード，生年月日，性別の 3 つの属性の組み合わせから米国居住者の 87%を一意に識別できることが報告された．また，2019 年には Rocher らによってデータからランダムにサンプリングされた 0.5%のデータからでも，個人の一意性を誤差平均 0.028 で求められることが示され，大きな反響をよんだ．そのため，企業や組織は収集したビッグデータを利活用する際，そのデータから個人が識別されてしまうリスクを評価し，そのリスクを低減するために匿名化を行う必要がある．

匿名化とは，PII (personal identifiable information) から個人が特定されないようにデータの一部の削除やランダムなノイズを追加（摂動化）することであり，匿名化されたデータから特定の個人を識別する攻撃を再識別と呼ぶ．健康情報の匿名化の標準化文書 ISO/TS 25237 では，匿名化（anonymization）が「データ管理者が単独，または他者と協力して，データ対象者を直接的または間接的に識別できないように，個人データを不可逆的に変更する手順」と定義されている．国内においては，匿名化技術は 2017 年の個人情報保護法改正で導入された「匿名加工情報」に関わっている．匿名加工情報は,「特定の個人を識別することができないように個人情報を加工し，当該個人情報を復元できないようにした情報」であり，病歴などの要配慮個人情報を第三者に提供する際には，データに含まれる本人の同意をあらかじめとる（オプトイン）か，個人情報の第三者提供とならないようにデータを匿名加工情報とすることが必要となった．この匿名加工情報を作成するために，現在顧客のビッグデータを有する多くの企業や組織から注目を集めている．

一般に，データを匿名化すると個人が識別されるリスクが低くなるため安全性は上がるが，データ中の値を加工するため有用性は下がる．しかし，匿名化による安全性や有用性の変化度合いには，匿名化に用いる加工手法，データの安全性や有用性を評価する指標，加工の対象となるデータ，などの様々な要因が影響するため評価が困難であり，これまで明らかになっていなかった．本研究の目的は，このようなデータに対する匿名化の影響を明らかにすることである．本論文ではこの目的を達成するために，(1) 既存の攻撃者モデルの問題点，(2) 履歴データの識別リスクの問題点，(3)k-anonymity の問題点，(4) 実験データへの依存性，という 4 つの課題を設定し，これらを解決する．

目 次

第1章　序論

1.1　研究背景

ビッグデータの有用性と危険性

機械学習や AI 技術の発展により，ビックデータの利活用が企業・医療機関・金融機関など多様な場面で盛んになっている．移動履歴や購買履歴データ等の個人情報ビッグデータは非常に有用であり，利活用することによって大きな利益を生むことができる．例えば，野田らは厚生労働省と総務省の許可を得て人口動態統計死亡票を目的外利用して，茨城県に住む 92,277 人の住民健診データを分析することにより，検査項目と死亡との関係を相対リスク (relative risk) などを用いて明らかにした [13]．また，日本人の健康寿命や生活習慣病に影響を与える要因を明らかにする目的で，国が全国で実施した循環器疾患基礎調査 [14]，および，国民健康・栄養調査の参加者を対象に追跡調査した NIPPON DATA[15] 等の大規模コホート研究が数多く行われていた．川南 [16] らは，喫煙習慣によるがん，肺がん死亡へ影響を分析し，非喫煙者に対する，毎日喫煙する集団の肺がん死亡の相対リスクが男性で 6.67 倍，女性で 3.67 倍であることを明らかにした．また，Chen らは購買履歴データの顧客を直近購買日 (R : Recency)，購買頻度 (F : Frequency)，購買金額 (M : Monetary) の 3 軸でグルーピングする手法である RFM 分析を行い，各グループの特徴を特定している [77]．

このように，ビックデータを分析することによって我々は様々な知見を得ることができるが，それと同時に非常に危険なものでもある．例えば，2006 年に Sweeny らは ZIP コード，生年月日，性別の 3 つの属性の組み合わせから米国居住者の 87%を一意に識別できることを示している [2]．また，2019 年には Rocher らが，データからランダムにサンプリングされた 0.5%のデータからでも，個人の一意性を誤差平均 0.028 で求められることを示して大きな反響をよんだ [1]．そのため，企業や組織は収集したビッグデータを利活用する際，そのデータから個人が識別されてしまうリスクを評価し，そのリスクを低減するために匿名化を行う必要がある．

匿名化技術

匿名化とは，個人情報から個人が特定されないように PII (personal identifiable information) を加工（値の削除や摂動化等）することであり，匿名化されたデータから特定の個人を識別する攻撃を再識別と呼ぶ．健康情報の匿名化の標準化文書 ISO/TS 25237 [62] では，匿名化（anonymization）が「データ管理者が単独または他者と協力して，データ対象者を直接的または間接的に識別できないように，個人データを不可逆的に変更する手順 (“process by which personal data is irreversibly altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party”)」と定義されている．また，匿名化

に関する用語とその分類を標準化した ISO/TS 20889 [43] では，匿名化（de-identification）が「データの識別属性集合と主な部分の間の関連性を除去する手順 (“a process that removes the association between a set of identifying attributes and the data principal”)」 と定義されている．

匿名化技術は日本国内でも注目されており，2017 年の個人情報保護法 [38][44] 改正で導入された「匿名加工情報（Anonymously Processed Information）」に大きくかかわっている．近年のプライバシ意識の高まりを受けて，個人情報保護法では，利用目的を明確にしないで個人についてのデータを取得することを禁じており，特に検査結果や病歴などは要配慮個人情報[1]に分類され，特別な措置が必要となった[2]．匿名加工情報は,「特定の個人を識別することができないように個人情報を加工し，当該個人情報を復元できないようにした情報」であり，要配慮個人情報を第三者に提供する際には，データに含まれる本人の同意をあらかじめとる（オプトイン）か，個人情報の第三者提供とならないようにデータを匿名加工情報とすることが必要となった．個人情報保護法では，匿名加工情報を作成する際は法律施行規則 19 条 1-5 号までの表 1.1 の要件を満たして適切に加工する必要があるとされており，匿名加工情報に対する識別行為が法 38 条[3]で禁じられている．

この匿名加工情報を作成するために匿名化技術は必須であるため，現在多くの企業や組織から注目を集めており，金子らが収集した匿名加工情報取扱事業者公表データ [17] によると，これまで KDDI 株式会社やファイザー株式会社をはじめとした約 500 の企業/組織が匿名加工情報を利活用していたことが明らかになっている．しかし，加工方法の多様性や加工の度合いの自由度の影響で，実際に利活用している企業は 25%に過ぎないという報告 [18][19] がある．なお，同データは梶間らによって「匿名加工情報目録 [86]」という Web ページとしてまとめられており，誰でも自由にデータの閲覧や検索をすることができる．また，匿名化技術を発展させるために，2015 年から “PWS Cup” [45][58][6] という匿名化と再識別の技術を競うコンテストが開催されており，多くの学術機関や様々な企業が参加している．さらに，2018 年の金沢らの調査 [87] によると，匿名化技術の研究開発は年々盛んになっており，世界各国合計の年間論文発表本数が 2006 年から 2016 年にかけて 383 件から 2,157 件に大きく増加していることが報告されている．これらのことより，匿名化技術は国内外で大きく注目されていることが言える．

1.2 本研究の目的

一般的にデータを匿名化すると，個人が識別されるリスクが低くなるため安全性は上がるが，データ中の値を加工するため有用性は下がる．匿名化をする際には，この安全性と有用性の変化に注意する必要がある．例として，データを匿名化して値を全て削除した場合を考える．このデータから個人

[1]本人の人種、信条、社会的身分、病歴、犯罪の経歴、犯罪により害を被った事実その他本人に対する不当な差別、偏見その他の不利益が生じないようにその取扱いに特に配慮を要するものとして政令で定める記述等が含まれる個人情報

[2]ただし，個人情報の保護に関する法律 [38] では学術研究，行政機関の保有する個人情報の保護に関する法律 [39] では相当な理由がある，または業務遂行に必要な限度，あるいは学術研究での利用や提供が可能となっている．

[3]匿名加工情報取扱事業者は，匿名加工情報を取り扱うに当たっては，当該匿名加工情報の作成に用いられた個人情報に係る本人を識別するために，当該個人情報から削除された記述等若しくは個人識別符号若しくは第三十六条第一項，行政機関の保有する個人情報の保護に関する法律（平成十五年法律第五十八号）第四十四条の十第一項（同条第二項において準用する場合を含む.）若しくは独立行政法人等の保有する個人情報の保護に関する法律第四十四条の十第一項（同条第二項において準用する場合を含む.）の規定により行われた加工の方法に関する情報を取得し，又は当該匿名加工情報を他の情報と照合してはならない．

表 1.1: 法律施行規則 19 条 1-5 号 [38]

1	個人情報に含まれる特定の個人を識別することができる記述等の全部又は一部を削除すること（当該全部又は一部の記述等を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む.）.
2	個人情報に含まれる個人識別符号の全部を削除すること（当該個人識別符号を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む.）.
3	個人情報と当該個人情報に措置を講じて得られる情報とを連結する符号（現に個人情報取扱事業者において取り扱う情報を相互に連結する符号に限る.）を削除すること（当該符号を復元することのできる規則性を有しない方法により当該個人情報と当該個人情報に措置を講じて得られる情報を連結することができない符号に置き換えることを含む.）.
4	特異な記述等を削除すること（当該特異な記述等を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む.）.
5	前各号に掲げる措置のほか、個人情報に含まれる記述等と当該個人情報を含む個人情報データベース等を構成する他の個人情報に含まれる記述等との差異その他の当該個人情報データベース等の性質を勘案し、その結果を踏まえて適切な措置を講ずること.

が識別されることは無いため安全性は非常に高いが，全ての値が失われているので全く何の役にも立たず有用性は非常に低い．一方で，データ中の値がほとんど加工されていないデータは，元データの統計量などの特性が失われないため有用性は非常に高いが，簡単に個人が識別されてしまうので安全性は非常に低い．このように，データの安全性と有用性の間には必ずトレードオフの関係があり，どちらかを上げればもう片方が下がる．本研究の目的は，このような**データに対する匿名化の影響を明らかにすること**である．

1.3 匿名化の既存研究

匿名化に関する研究には，大きく 2 つの流れがある．様々な加工手法の研究と，それらの評価に関する評価である．

1.3.1 加工手法の研究

データを匿名化する手法の研究は盛んに行われており，以下のような様々なアルゴリズムが提案されている．

- レコード/セル削除：データ中の特異な値を持つレコード/セルを削除する手法
- 一般化：データ中の値をより一般的な区分や範囲に加工する手法（例：2021 年 4 月 3 日 →2021 年 4 月，静岡県 → 東海地方，など）
- 仮名化：データ中の識別子（単体で個人を一意に識別するもの）を仮 ID 等の仮名に置換する手法

- 摂動化：データ中の値にランダムノイズを加えたり丸め込みを行ったりすることにより，個人識別を防ぐ手法
- サンプリング：データの一部のみをサンプリングして公開する手法
- スワッピング：データ中の値を個人間で入れ替えることにより，個人識別を防ぐ手法
- トップ/ボトムコーディング：ある閾値に基づいて，データ中の閾値以下/以上の値をすべて閾値に置換する手法
- ミクロアグリゲーション：データ中の個人をクラスタリング等でグループ分けし，グループ内のデータの値を平均値等の代表値で置換する手法
- 疑似データ生成：元のデータと統計的に似た特性を持つ疑似データを生成する手法
- ダミーレコード追加：元のデータにダミーのレコードや架空の個人を追加することにより，個人識別をかく乱する手法

Aggarwal[64]，Hundepool[46]，Elliot[47]，ICO[63]，Torra[48] といった匿名化手法の網羅的なサーベイがなされている．他の匿名化手法として，三本らはデータの低ランク行列への分解に注目し，シーケンスデータに対するその手法の性能を評価している [85]．Domingo-Ferrer らは，ミクロデータを保護するために，様々な SDC(statistical disclosure control) 手法の比較をしている [76]．また，南らは複数データの集計表の差分から機密情報を得る攻撃（差分攻撃）に注目し，それを防ぐためのセル秘匿アルゴリズムを提案している [89][90]．さらに，本郷らは Laplace 分布に従うノイズをデータに加える手法（Laplace メカニズム）の 3 つの問題点（非負制約の逸脱，部分和精度の劣化，計算量の増大）を改善する手法を提案している [91]．杉山らは，匿名化手法の一つである疑似データ合成に注目し，既存手法の問題点である「マクロな特性は保存されるが，ミクロな特性は失ってしまう」という点を改善した疑似データ合成手法を提案している [92]．また，福嶋らはデータ中の 1 人のユーザに複数の仮名を割り当てて分割する多重仮名化手法を提案している [97]．山岡は，提供するデータを異なるパターンを用いたサンプリングによって加工することによって，第三者に提供したデータが提供先から漏洩した場合に提供元を特定できる仕組みを提案した [99]．

1.3.2 評価指標の研究

一般的に匿名化されたデータの評価は，統計量等の分析結果が加工によって元データからどのくらい変化してしまったか（有用性），加工されたデータからどれだけ多くの個人が再識別されるか（安全性）の 2 点で評価される．特に安全性指標が多く提案されており，例えば Koot らは，確率分布間の距離を測定する手法である Kullback–Leibler 距離を用いて一意性の近似値を求め，データの匿名性を定量的に評価する方法を提案している [65]．ここではそれらの安全性指標の研究を，差分プライバシ，攻撃者モデル，k-anonymity の 3 種類に分けて述べる．

差分プライバシ

Dwork によって提案された差分プライバシ [74] は，ランダムメカニズムで加工されたデータの分析結果から変化を識別できないことを保証する指標である．Ashwin らは，通勤データのような疎なデータにも適用できる (ϵ, δ)-確率的差分プライバシを提案し，それを満たすような合成データ生成アルゴリズムを提案した [108]．Ios らは，差分プライバシを満たすリレーショナルデータベースシステムである PrivateSQL を提案している [107]．Barak らは，プライバシ，正確性，一貫性の全てを保証する集計表作成手法として，差分プライバシ，フーリエ変換，線形計画法を用いた方法を提案している [109].

攻撃者モデル

匿名化されたデータの安全性を評価するためには，そのデータから個人を識別しようとする攻撃者の想定をする必要がある．攻撃者はデータについての何かしらの背景知識を持っており，それを手がかりにして個人の識別を試みることが考えられるが，この背景知識によって攻撃者の危険度は大きく変わる．

Domingo は匿名化されたデータを攻撃する攻撃者として，最悪のケースを想定した元のデータを丸ごと背景知識として有している最大知識攻撃者モデルを提案した [26]．小栗らは，管理者権限等を有しており匿名化手法を再現できる最高能力攻撃者モデルを提案し，最強の攻撃者想定として最大知識・最高能力攻撃者モデルを想定してリスク評価を行った [94]．また，濱田らは「匿名化アルゴリズムを知っていることでどれくらい再識別が容易になるのか？」という問題を実験的に解決した [95]．Li らは，攻撃者がデータを分析することによって背景知識を増やすことができることに注目し，δ-privacy という安全性指標を提案した [69]．また，早稲田らは攻撃者の持つ背景知識の質や量に基づいて，仮名化されたデータの識別耐性や漏洩量を評価している [88]．さらに，正木は個人の軌跡情報（線）の一部のみ（点）を背景知識として持つ攻撃者モデルを提案し，そのモデルに基づいた背景知識の精度評価やデータの匿名性評価をしている [93]．El Emam は現実世界で起こりうる「データが攻撃を受ける確率」や「情報漏洩が発生する確率」を想定して，“Intentional Attack”, “Unintentional Attack”, “Data Breach”, “Open Data Attack” の 4 つを提案している [75].

k-anonymity

安全性評価指標の代表的な研究として，Sweeney と Samaratiy によって提案された k-匿名性（k-anonymity）がある [49][2] ．k-anonymity は匿名化されたデータの安全性を評価する指標であり，データ中のいかなる個人のデータも少なくとも k 人の個人が同じ値を持っており区別がつかないことを保証するものである．これを満たすようにデータを加工することを k-匿名化と呼ぶ．例として，5 人のデータを k-匿名化するケースを図 1.1 に示す．図中左側に示すデータは元データであり，これをそのまま公開・提供してしまうと個人が一意に識別されてしまうため，匿名化をする必要がある．この例では，名前属性の仮名化と年齢属性と郵便番号属性の一般化を行うことによって，k-匿名化を行っている．図中右上のデータは 5 人全員が同じデータ（21-55,10***）に加工されており区別がつかない

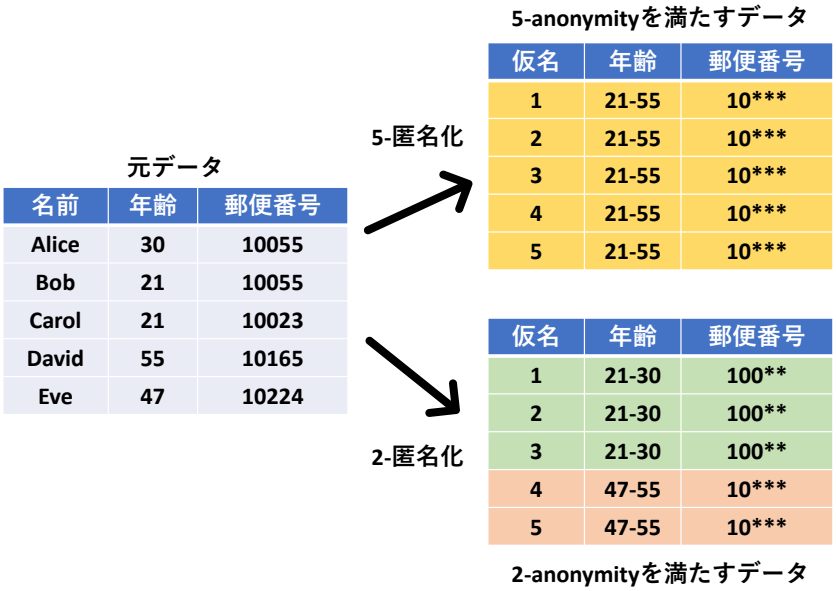


図 1.1: k-匿名化の例

ため，5-anonymity を満たしている（5-匿名化）．一方，図中右下のデータは 2 人 (仮名 1,2,3) と 3 人 (仮名 4,5) がそれぞれ同じデータに加工されており，少なくとも 2 人の区別がつかない状態になっているため，2-anonymity を満たしている（2-匿名化）．

この k-匿名化は多くの研究者によって議論されている．Meyeson と Williams によって，セル削除のみを用いた最適な k-匿名化を見つけることは NP 困難である（k-匿名化問題）ことが証明されており [50]，山田らによって，一定の条件下であれば k-匿名化問題は定数近似が可能であることが示されている [96]．Basu らは k-anonymity を満たすようなデータ公開をする場合のリスクを実験的に求めている [53]．前田ら [35][98] は，様々な性質を持つデータセットをセル削除による k–匿名化 [36] で加工することにより，加工で失われる有用性（機械学習の結果）の度合いがデータセットの性質によって異なることを示している．また，K-optimize [51] や Incognito [52] や Mondrian [41] では，ヒューリスティックな手法を用いて任意の k に対する k-匿名化の加工コストの評価を行っている．

1.4 既存研究の問題点

私は 1.3.2 節で述べた攻撃者モデルと k-anonymity に注目する．これらの既存研究には以下のような問題点がある．

課題 1：既存の攻撃者モデル

攻撃者がどのような背景知識を利用できるかは不明である．しかも，攻撃者はデータ中の最も危険（保護されていない）な個人を再識別してくることが想定される．例えば，ある個人のレコードにとても珍しい病気に罹患した履歴が残っていた場合，その個人は簡単に識別されてしまうだろう．とこ

ろが，情報漏洩のリスクを減らすためにはとても多くの属性の候補があり，攻撃される前にどの属性を優先的に加工すべきかを決めることは困難である．

そこで，データ中の加工すべき危険な属性を決めるために，Domingo-Ferrer らは元データ全てを背景知識として持つ最大知識攻撃者モデルを提案していた [26]．この攻撃者モデルでは，攻撃者が元データと加工データの両方を全て知っていることを想定しており，これは背景知識を持つ攻撃者として最悪のケースである．しかし，この仮定は強すぎることがしばしば指摘されている．例えば，匿名化の技術を競うコンテスト PWS CUP[45] でこのモデルを用いたところ，データの全ての属性を背景知識として使えるこの攻撃者はあまりにも強力であり，データの安全性を過度に低く見積もってしまうことが分かってきた．

一方，El Emam の攻撃者モデル [75] は,「平均的な人は 150 人の友人がいる」といった Dunbar 数等の経験則に基づいて，この 150 人の中に攻撃者がいる確率などを用いてリスクを評価しており，楽観的すぎる見積りである．現実の攻撃者は，最大知識攻撃者モデルほど強くなく，Dunbar 数モデルほど弱くもない．

課題 2：履歴データの識別リスク

商品の購買履歴や位置情報のように，仮名のもとに結びつけられた時系列データから，即時に個人が識別されることはないと考えられている．例えば，JR 東日本は 2013 年に交通系 IC カード Suica[78] の利用履歴データを個人が特定できる情報ではないとみなし，そのデータを他社に提供したことで批判を受けた [111]．それゆえ，既存の匿名化手法のほとんど（k-匿名性，l-diversity, t-closeness など）は，時刻情報の無い静的なデータを仮定している．しかしながら，数年間という長期間にわたる医療履歴からは，過去の病歴から容易に個人が特定できるだろう．また，位置情報に至っては，1 日の詳細な位置記録から自宅と勤務先が推測可能である．例えば，鉄道の乗降履歴に基づいて，高々3 駅分の履歴から 98%の個人が識別されることが知られている [101]．

しかしながら，このような動的に変化するデータから個人が識別されるリスクについては，限られた研究しか行われていない．例えば，Xiao らによる m-invariance[54] や柚木らの連続匿名化 [100] があるのみである．その理由には，前述の「履歴から個人は特定できない」という誤解に加えて，動的に変化するイベントを正確に定式化することの技術的な困難性があると考える．また，履歴データではデータの値だけでなく，データの量も個人を識別するための手掛かりとなり得る．例えば 1 年間の病歴データの場合，普通の風邪に一回罹ったという病歴を持つ個人は識別されにくいが，何十回も風邪に罹っている個人は珍しいため識別されやすい．こういった複雑さも，履歴データの識別リスクの研究が少ない理由の一つであると考えられる．

課題 3：k-anonymity の問題点

k-anonymity を満たすための加工（k-匿名化）には様々な手法があり，削除や一般化といった手法が組み合わされて用いられる．しかし，有用性と安全性の間にはトレードオフの関係があるので，最適な加工手法を見つけるのは困難である．

元データ

名前	年齢	郵便番号
Alice	30	10055
Bob	25	10055
Carol	21	10023
David	55	10165
Eve	47	10224

加工データ 1
(2-conceamlment)

仮名	年齢	郵便番号
1	25-30	10055
2	21-30	100**
3	21-25	100**
4	47-55	10***
5	47-55	10***

加工データ 2
(2-anonymity)

仮名	年齢	郵便番号
1	21-30	100**
2	21-30	100**
3	21-30	100**
4	47-55	10***
5	47-55	10***

図 1.2: k-anonymity によって生じる過度な加工の例

k-anonymity は広く用いられているが課題も多く，改善指標が数多く提案されている．k-anonymity だけでは属性推定のリスクを評価できないことを指摘し，Machanavajjhala らは l-diversity[3] を，Truta らは p-sensitive[4] という指標を提案した．Monreale らは c-safety という軌跡データの匿名化のためのフレームワークを提案している [66]．このフレームワークに基づいて，Basu らは k-匿名性を満たすデータ公開をする際のリスクを実験的に示した [67]．Xiao らは，データの再公開をすることによって生じるプライバシリスクに注目し，m-invariance という新しい一般化手法を提案している [54]．m-invariance を用いて柚木らは動的に変化するデータの連続匿名化の安全性指標を提案している [100]．この手法は，値の一般化と他のデータセットの顧客に似たダミー顧客の追加を組み合わせた手法である．Domingo-Ferrer らは k-anonymity を拡張した t-closeness[102] と ε-differential privacy が強く関係していることを示した [83]．Stokes は k-anonymity を緩めた (k,l)-anonymity [84] を提案し，k-anonymity をより一般的にした n-confusion を提案している [68]．

Tamir らは，k-anonymity の厳格さから生じる過度な加工を解決するために，2 部グラフと完全マッチングに注目して新たな指標 k-concealment[5] を提案した．例として，図 1.2 に示す 3 つのデータを考えよう．図中左の 5 人の個人についてのデータを，仮名化と一般化を用いて加工したものが加工データ 1,2 である．加工データ 2 では 5 人の仮名が 3 人と 2 人のグループに分けられ，それぞれのグループ内の区別がつかないように加工されているため 2-anonymity を満たしている．一方，加工データ 1 は加工データ 2 のいくつかの値を元に戻しているものであるが，このデータも「少なくとも 2 人の区別がつかない状態」になっている．これらの元に戻しても問題無い値は，k-匿名化によって生じる過度な加工といえるだろう．また，k-匿名化されたデータ内では，個人によって識別されるリスクに差が生じる．例えば，加工データ 2 で 3 人のグループに加工されている Alice, Bob, Carol は，2 人のグループに加工されている David, Eve よりも安全性が高い．こういった個人による識別リスクの不公平さも，k-匿名化の問題点といえるだろう．

課題 4：実験データへの依存性

K-optimize [51] や Incognito [52] や Mondrian [41] などの多くの研究では，k 匿名化の k などのパラメータを様々な値に変化させることにより，実験的に加工コストを求めている．例えば，Basu らは k-anonymity を満たすようなデータ公開をする場合のリスクを実験的に求めている [53]．実際に，加工コストや最適な k を求めることは，対象となるデータや想定するユースケースシナリオ等に大きく依存するので，非常に困難である．

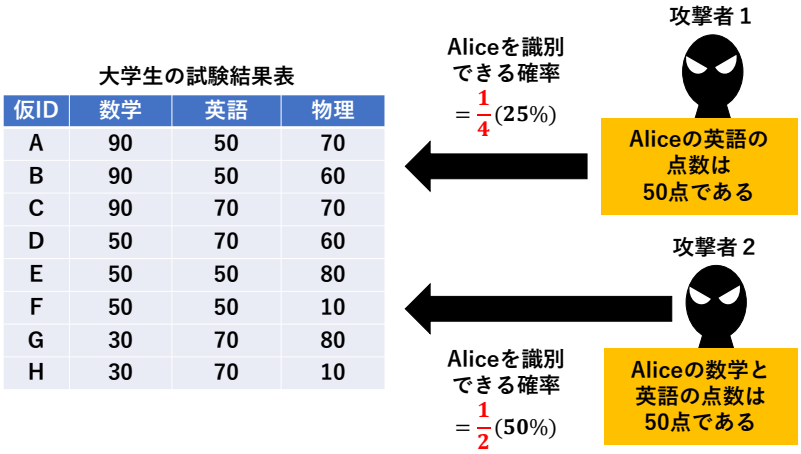


図 1.3: 攻撃者の部分的な背景知識の例

1.5 問題点を解決する手法と新規性

本研究では，前述した問題点を以下のように解決する．

(A) 部分知識を有する攻撃者の新たな数理モデル

攻撃者モデルの問題（課題 1）に対して，私は，部分的な背景知識を有する攻撃者モデルを提案する．最大知識攻撃者モデルほど強い仮定をせず，Dunbar 数モデルほど楽観的に考えなくても，部分的な知識を持つ攻撃者によるリスクを考えることが可能である．例として，図 1.3 に示す攻撃者が大学生の試験結果表からある個人を識別しようとするケースを考えよう．試験結果表では 8 人の学生が仮 ID に仮名化されているが，試験の点数は加工されていないものとする．このとき，試験結果表から Alice という個人を識別しようとする攻撃者 1,2 を考える．攻撃者 1 は背景知識として Alice の英語の点数を知っており，攻撃者 2 は Alice の数学と英語の点数を知っている．攻撃者 1 は背景知識を使って Alice の候補を 4 人 (A,B,E,F) に絞り込めるのに対し，より多くの背景知識を持つ攻撃者 2 は 2 人 (E,F) までに絞り込むことができる．このように，背景知識の量や質によって攻撃者の危険度（= データの安全性）は大きく変化する．

この例が，私の研究を動機づける根拠である．そして，このような部分知識を仮定したモデルによる安全性の評価に，私の研究の一つ目の新規性がある．本研究と先行研究の比較を図 1.4 に示す．比較対象として，前述した 2 つの研究（Domingo-Ferrer らの最大知識攻撃者モデル [26]，El Emam の攻撃者想定 [75]）と比較して，本手法の特徴を述べる．既存研究では様々な攻撃者の背景知識を想定しているが，本手法で想定する背景知識は「データ中のある 1 つの値」のみである．本モデルでは，この部分知識を有する攻撃者の平均識別確率を用いて，データセットの安全性を評価する．さらに，El Emam の研究と比較すると，本手法で提案する攻撃者モデルは普遍的であり，最大知識攻撃者モデルと同じように全てのデータセットに対して適用可能である．提案する攻撃者想定はデータの中のある属性から，レコード数等の統計量に基づく確率で背景知識を得ることを想定しているため，既存研究のような最悪のケース以外の多くの場合にも適用可能なものである．

	本研究 部分知識モデル	既存手法	
		最大知識 攻撃者モデル[26]	El Emamの 攻撃者想定[75]
攻撃者の 背景知識	データ中のある1つの 値のみを確率的に得る	元のデータそのものを 背景知識として持つ	情報漏洩の統計値や経験則に 基づいた背景知識を得る
安全性の 数理モデル化	○ 攻撃者の平均識別確率 を与える数理モデルを 提案する	✖ 実験的な安全性評価は行っ ているが，数理モデル化は していない	○ データが攻撃を受ける確率な どの値から安全性を求めるモ デルを提案している
モデルの 普遍性	○ 全てのデータに対して 適用できる	○ 全てのデータに対して適用 できる	✖ 計算に必要な客観的な値が ケースによって異なる

図 1.4: 本研究と先行研究の比較（攻撃者想定と安全性推定）

平均識別確率の欠点の一つは，評価結果の計算に時間がかかることである．提案モデルではデータセットの全レコードを調査するため，処理時間はデータのレコード数や属性数に比例し，ビッグデータのリスクを評価しようとすると膨大な時間がかかる．例えば，38,087 レコード，6 属性，400 人分の購買履歴データのリスク評価を行うために，27.5 秒かかることが私の調査によって明らかになっている（3.4.4 節参照）．この問題を解決するために我々は，平均モデル，低コストモデル，サンプリングモデルという 3 つの近似モデルを提案する．これらのモデルは近似精度と計算コストの間にトレードオフの関係があり，ユースケースに応じて使い分ける必要がある．

また，提案モデルの性能評価をするために，我々は幅広いデータを用いて実際にリスクの評価をする実験を行う．この実験では， 特徴の異なる 4 つのデータセット（購買履歴データ，患者の入院データ，世帯収入データ，ローン借り入れデータ）を用いて，データから個人が識別される確率を求めて安全性を評価する．

(B) 履歴データの数理モデル化による加工コスト推定

私は履歴データからの識別リスクの問題（課題 2）を解決するために,「x レコードのデータが与えられた時，その中のユニークな値の種類数 y は，全 ℓ 種類のうちいくらか？」という問題 (商品種類数問題) を考え，それを解決する履歴データの新たな数理モデルを提案する．提案モデルでは，レコードの値が一様分布で独立して発生するという仮定の下，次の 2 つの定理を示している．(a)x レコードのデータが与えられた時に，全 ℓ 種から選ばれるユニーク数 y の条件付確率の分布 (定理 4.2.1)．(b)x レコードのデータに含まれるユニーク数 y の期待値 (定理 4.2.2)．本モデルを用いることにより，データの基本統計量や加工パラメータといった加工前に手に入る値から加工コストの見積をできるようになるため，実際に様々な k の値を用いて k-匿名化を行わなくとも，最適な k の値を推定することができる．

(C) 新たな k-concealment 化手法の提案

Tamir らは図 1.2 の例で示した「k-匿名化の際にできてしまう $k+1$ 人以上のグループ」を k-匿名化の問題点であると指摘し，これを改善するために k-concealment という新たな指標を提案している [5]．この指標では，データの準識別子を完全に等しくする同値類（k この要素は全て同じ）の代わりに，元データと加工データの間のマッチの候補が k 個以上であれば，k-anonymity と同等の安全性が保障される．

しかし，Tamir らの [5] では，他の先行研究と同じように静的なレコードを仮定していたことと，「全ての個人が等しく k 人と区別がつかない」状態である完全 k-concealment が得られていない．そこで本研究では，以下の 2 つの加工手法を提案することによってこれらの問題（課題 3）の解決を試みる．

- 手法 1：顧客やレコードの削除・追加をすることなく履歴データを「最低でも k 人の区別がつかない」状態にするために，仮名の一般化とレコード間 k-concealment を用いた履歴データの k-concealment 化手法を提案する．
- 手法 2：レコードの削除・追加をせずにデータを「全ての個人が等しく k 人と区別がつかない」状態に加工するために，巡回セールスマン問題の解法やクラスタリングを用いた完全 k-concealment 化手法を提案する．

(D) 様々なデータに対する実験的評価

対象とするデータに依存する問題（課題 4）は，データサイエンスにおいては不可避な命題であろう．そこで，データに特有のふるまいによって想定外の評価をしてしまう危険性を最小化するため，私の研究では出来るだけ多くの分野のオープンデータを評価して対処する．そのために，表 1.2 に示す多様なデータについて分析する．本研究ではこれらのデータのうち，世帯収入データ，購買履歴データ，交通 IC カードデータ，健康診断データ，レセプトデータに注目し，これらに対する安全性評価や有用性評価を行う．

まず，データを次のように静的，動的に区別する．世帯支出データのように個人数とレコード数が等しいデータを静的なデータと呼ぶ．静的なデータから個人を識別する単純な手法として，ユークリッド距離のような距離尺度を用いて個人間の類似度を計算し，距離が近い 2 人を同一個人と識別する攻撃が考えられる．本稿では，静的なデータを匿名化することによって，ユークリッド距離攻撃をどれだけ防ぐことができるかを実験的に評価する．

購買履歴データは個人数 $\leq$ レコード数のデータであり，1 レコードにはある個人がある商品を買った履歴が記録されている．1 個人が複数のレコード（履歴）を持っているため，人によって購買商品の特徴（集合）が大きく異なることが考えられる．よって，このようなデータを履歴データと呼ぶ．履歴データに対して，Jaccard 係数等の集合間の類似度を測る尺度を用いて個人間の距離を測り，似ている 2 人を同一個人と識別する攻撃が考えられる．本稿では，履歴データに対する Jaccard 係数攻撃を防ぐためには，どのような匿名化をしたらよいのかということを実験的に評価する．

交通 IC カードデータは，購買履歴や乗降履歴といった複数の異なる用途のデータが含まれている珍しいデータである．こういった複雑なデータの安全性は，前述したユークリッド距離や Jaccard 係

表 1.2: 本研究で用いるデータセット

ID	内容	個人数	レコード数 (行)	属性数 (列)	扱う章	データの種類
1	購買履歴	400	38,087	7	3,4,5,9	オープンデータ
2	健康診断	198,740	964,636	49	6	匿名加工情報
3	交通 IC カード	31	584	10	7	個人データ (同意取得)
4	世帯支出	8,333	8,333	25	8	合成データ
5	糖尿病患者	71,518	101,766	50	3	オープンデータ
6	世帯収入	32,561	32,561	16	3,10	オープンデータ
7	ローン借入	42,538	42,538	145	3	オープンデータ
8	疑似人流	6,432	901,465	9	9,10	合成データ
9	傷病レセプト	288,568	39,363,878	15	6	匿名加工情報
10	医薬品レセプト	279,199	31,465,504	21	6	匿名加工情報

表 1.3: 本研究で実験的に評価するデータセット

章	用いるデータ	実験目的	実験内容	評価対象
5	購買履歴 (履歴データ)	履歴データから個人が識別されるリスクを評価する	Jaccard 係数を用いた攻撃を想定しそれを匿名化によってどれだけ防げるかを実験的に評価する	安全性 (Jaccard 係数を用いた再識別攻撃)
6	健康診断 (静的) 傷病/医薬品レセプト (動的)	データから得られる有益な情報が匿名化によってどれだけ変化するかを評価する	疾病に対する相対リスクや機械学習の精度が k-匿名化によってどれだけ変化するかを実験的に評価する	有用性 (F 値や相対リスクによる評価)
7	交通 IC カード (履歴データ)	複数の用途からなる複雑なデータから個人が識別されるリスクを評価する	エントロピーを用いてデータ中の用途ごとの危険度を実験的に評価し用途間に相関があるのかも調査する	安全性 (エントロピーによる評価)
8	世帯支出 (静的データ)	静的なデータから個人が識別されるリスクを評価する	ユークリッド距離を用いた攻撃手法を想定し，それを匿名化によってどれだけ防げるかを実験的に評価する	安全性 (ユークリッド距離による再識別攻撃)

数などでは評価が難しい．しかし，エントロピーを用いることによって，データのどの用途がどれくらい危険なのか，用途間の相関はあるのか，といった評価ができる．本稿では，複数用途からなる複雑なデータの安全性を正しく評価するために，交通 IC カードデータをエントロピーによって実験的に評価する．

健康診断データや傷病/医薬品レセプトデータは，分析することによって非常に有益な情報を得られる代表的なデータであり，多くの研究で用いられている．これらのデータから得られる有益な情報の例として，特定の病気に罹患する相対リスクや機械学習による疾病罹患予測などが挙げられる．本稿では，健診データとレセプトデータから得られる有用な情報（相対リスクや機械学習の結果など）が，k-匿名化によってどれだけ変化するのかを実験的に評価する．

これらのデータの種類と適した評価指標の関係を，表 1.3 にまとめる．

1.6 本研究の貢献

本研究の主な貢献は，以下の 3 つである．

1. 匿名化による安全性・有用性変化の理論的評価：
   (1) 新たな攻撃者モデルを提案することにより，データ中のどの属性が危険であるかを平均識別確率によって評価した．また，平均識別確率を近似する 3 つのモデルを提案し，購買履歴等の 4 つのデータに対して安全性の理論的な評価を行った．その結果，購買履歴データの時刻属性の値 1 つのみから，平均 32%の確率で個人が識別されることなどを明らかにした．(3 章)
   (2) 履歴データの値が一様に生起する仮定の下，履歴データ中に登場する項目の種類数の確率分布とその期待値を与える数理モデルを提案した．また，提案モデルを応用することにより，履歴データを k-匿名化するために必要なダミーレコード数の期待値を，元データの統計量やパラメータ k などから求めることができることを示し，匿名化の加工コストを理論的に評価した．その結果，k-匿名化するデータに対する最適な k の値を加工前に求めた．(4 章)

2. 匿名化による安全性・有用性変化の実験的評価：
   (1) 購買履歴データの個人が購買商品特徴から識別されるリスクを想定し，そのリスクへの耐性を高めるためにかかる加工コストを実験的に評価した．その結果，購買履歴データを 50 個のクラスタに分割して k-匿名化をするためには，約 18 万のダミーレコードの追加が必要であることなどを明らかにした．(5 章)
   (2) 健康診断データと傷病/医薬品レセプトデータを匿名化することによって，データの安全性と有用性がどのように変化するかを実験的に評価した．その結果，病歴/処方歴を k-匿名化することによって，識別される人数の割合が平均 2.9%まで減少することや，高血圧に対する相対リスクが相対誤差で 0.073 しか変化しないことなどを明らかにした．(6 章)
   (3) 乗降履歴や購買履歴などの複数用途の履歴が含まれている交通 IC カードデータから，個人が識別されるリスクをエントロピーを用いて評価した．その結果，個人を識別できる確率が 1 つの乗降履歴によって 3.3%から 28.4%まで上がることなどを明らかにした．(7 章)
   (4) 世帯支出データの個人がレコード間のユークリッド距離から識別されるリスクを想定し，そのリスクへの耐性を高めるための加工手法を検討した．その結果，単純なノイズ付加では再識別を防げないことや，k-匿名化によって再識別率を 17%まで下げられることなどを明らかにした．(8 章)

3. 新たな匿名化手法の提案：
   (1)「仮名の一般化」と「レコード間 k-concealment」を用いることによって，個人によってレコード数の異なる履歴データを k-concealment を満たすように加工する手法を提案した．提案手法を用いることにより，顧客やレコードを追加/削除することなく，履歴データを「最低でも k 人の区別がつかない」状態にすることができる．(9 章)
   (2) データを「全ての個人が等しく k 人と区別がつかない」状態に加工する完全 k-concealment 化に注目し，コストの低い加工をするために巡回セールスマン問題の近似解法を応用したアルゴリズムを提案した．提案手法によって，k-匿名化によって生じる過度な加工や，データ中の

個人識別リスクの不公平さを解消することができる.（10 章）

1.7　個人情報の取扱に対する配慮について

本研究では様々なデータを用いて実験を行っているが，これらのデータが個人情報保護法に抵触しないことを確認している．本研究で扱うデータの種類を，表 1.2 の「データの種類」列に示す．データ 1,5,6,7 は，いずれもインターネット上で公開されているオープンデータであるため，個人情報には該当しない．また，データ 2,9,10 はいずれもあるヘルスケア企業によって作成された匿名加工情報であるため，個人情報には該当しない．さらに，データ 4,8 は実データをもとに合成された疑似データであるため，これも個人情報には該当しない．データ 3 は私が独自で収集した実データであるため，これは個人情報に該当するが，収集対象者 31 人全員から研究利用への同意を取っている．

6 章では健康診断データと疾病や生活習慣との相関を明らかにして疾病予防，生活改善，健康施策づくりに有益な知見を得ることを目的に，匿名加工情報（法 [38] 第 2 条 9 項）を用いている．同法，関連する法令，ガイドラインなどを遵守して，適切な安全管理措置を施して研究を遂行している．本章で発表する研究結果には，特定の個人を識別可能な情報が含まれず，健康診断被験者のプライバシへ及ぼす影響がないことを，事前に (2020 年 7 月 30 日) ヘルスケア企業に相談，確認済みである．また，厚労省ガイドライン [27] 第 12 の 2「研究の成果の公表にあたっての留意点」に抵触している該当項目はないことを確認している．

1.8　本論文の構成

本論文の構成を図 1.5 に示す．2 章では，記号等の定義や先行研究について述べる．3,4 章では，貢献 1 にかかわる安全性と有用性の理論的評価について述べる．5,6,7,8 章では，貢献 2 にかかわる安全性と有用性の実験的評価について述べる．9,10 章では，貢献 3 にかかわる匿名化手法の提案について述べる．11 章では，本論文の結論を述べる．

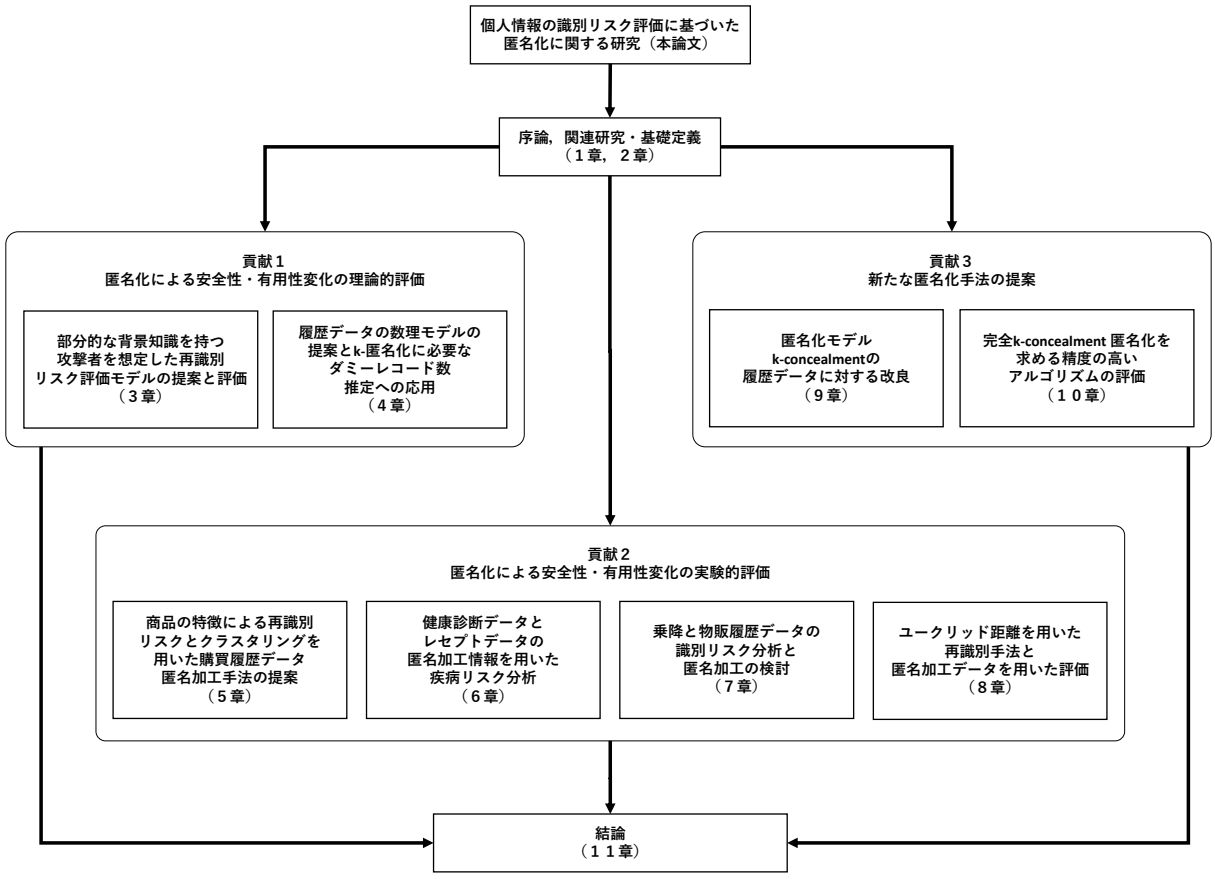

個人情報の識別リスク評価に基づいた
匿名化に関する研究（本論文）
序論，関連研究・基礎定義
（1章，2章）
貢献1
匿名化による安全性・有用性変化の理論的評価
部分的な背景知識を持つ
攻撃者を想定した再識別
リスク評価モデルの提案と評価
（3章）
履歴データの数理モデルの
提案とk-匿名化に必要な
ダミーレコード数
推定への応用
（4章）
貢献3
新たな匿名化手法の提案
匿名化モデル
k-concealmentの
履歴データに対する改良
（9章）
完全k-concealment 匿名化を
求める精度の高い
アルゴリズムの評価
（10章）
貢献2
匿名化による安全性・有用性変化の実験的評価
商品の特徴による再識別
リスクとクラスタリングを
用いた購買履歴データ
匿名加工手法の提案
（5章）
健康診断データと
レセプトデータの
匿名加工情報を用いた
疾病リスク分析
（6章）
乗降と物販履歴データの
識別リスク分析と
匿名加工の検討
（7章）
ユークリッド距離を用いた
再識別手法と
匿名加工データを用いた評価
（8章）
結論
（11章）

図 1.5: 本論文の構成

第2章　従来研究

2.1　従来研究

本研究と大きく関わる 4 つの従来研究（k-anonymity[2], k-concealment[5], 最大知識攻撃者モデル [26]，El Emam の攻撃者モデル [75]）について述べる．また，3 章で扱うデータのプライバシリスク評価の先行研究として，高崎による分析結果 [110] についても述べ，6 章で扱う医療情報分析の先行研究として，コホート研究を紹介する．

2.1.1　k-anonymity

k-anonymity[2] は Sweeny によって 2006 年に提案された匿名性指標であり，あるデータのすべてのレコードが少なくとも $k-1$ 個の他のレコードと区別できないことを保証するものである[1]．説明のためのデータ M_1, M_2, M_3 を表 2.1,2.2,2.3 に示す．M_1 は 5 人分の年齢と郵便番号のデータであり，M_2, M_3 は M_1 を一般化によって匿名化したデータである．M_1 はすべてのレコードが異なっているので 1-anonymity しか満たしていないが，M_2, M_3 はそれぞれ 5-anonymity, 2-anonymity を満たしている．

また，データを k-anonymity を満たすように加工する手法（k-匿名化）についての研究も盛んにされており，ここではその一例として Mondrian アルゴリズム [41] を紹介する．2006 年に Kristen らによって提案された Mondrian アルゴリズムは，複数の属性を持つレコードを k-匿名化するための手法である．Mondrian アルゴリズムでは，まずデータ中の複数の属性の中から加工の基準を定め，その基準によってデータを複数のグループ（レコード集合）に分割する．さらに，それらのグループがさらに分割可能である場合，分割処理を再帰的に繰り返すことによって，データをより細かいグループに分割していく．全グループの大きさが k を下回らないように分割を再帰的に繰り返したのち，それ

表 2.1: Original Data M_1

name	age	zipcode
Alice	30	10055
Bob	21	10055
Carol	21	10023
David	55	10165
Eve	47	10224

表 2.2: Anonymized Data M_2

id	age	zipcode
1	21-55	10***
2	21-55	10***
3	21-55	10***
4	21-55	10***
5	21-55	10***

[1] [2],3.1 節の定義 3 を参照されたし．

表 2.3: Anonymized Data M_3

id	age	zipcode
1	21-30	100**
2	21-30	100**
3	21-30	100**
4	47-55	10***
5	47-55	10***

表 2.4: Anonymized Data M_4（1-anonymity, 2-concealment）

id	age	zipcode
1	21-30	10055
2	21-30	100**
3	21	100**
4	47-55	10***
5	47-55	10***

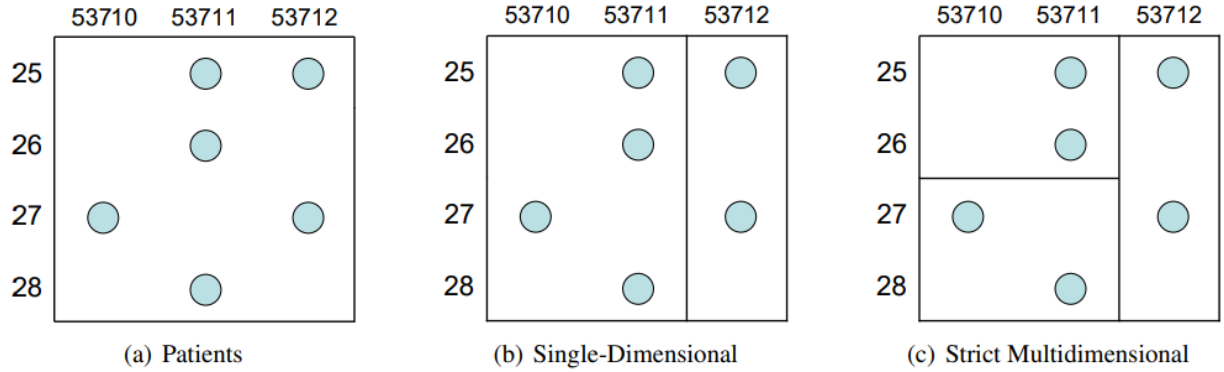


図 2.1: Kristen らが [41] で示した Mondrian アルゴリズムの概要図（[41], Fig.4）

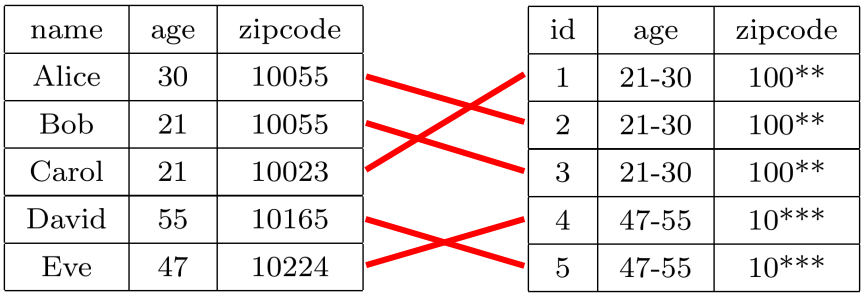
らのグループ内のレコードの値を平均値や中央値のような代表値に置き換えることによって，データを k-匿名化する．

Kristen らが [41] で示した Mondrian アルゴリズムの概要図[2]を，図 2.1 に示す．図中の (a) は，縦軸が age，横軸が zipcode である散布図として，元データの 6 人の個人をプロットしたものであり，このデータを 2-匿名化するケースを考える．まず，2 つの属性から zipcode を基準として選び，データを 2 つのグループに分割したものが (b) であり，この図では 4 人と 2 人のグループに分割されている．(b) の 4 人のグループはさらに分割することが可能であるため，age 属性を基準として 2 つのグループに分割する．その結果，(c) で示されたようにデータが 3 つのグループに分割されるため，各グループ内で個人の区別がつかないように加工をすれば，データを 2-匿名化することができる．例えば，2-匿名化された加工データ M_3 は，M_1 の age 属性と zipcode 属性を加工の基準として，データが 2 つのグループ（id=1,2,3 と id=4,5）に分割され，各グループ内の個人が一般化によって区別がつかないように加工されている．

2.1.2 k-concealment

k-anonymity はシンプルでわかりやすい指標であり，広く用いられている．しかしながら課題も多く，Tamir らは k-anonymity を満たすようにデータを加工する際に過度な加工が生じることを弱点と

[2][41], p.4 の Figure 4 より転載.



name	age	zipcode
Alice	30	10055
Bob	21	10055
Carol	21	10023
David	55	10165
Eve	47	10224

id	age	zipcode
1	21-30	100**
2	21-30	100**
3	21-30	100**
4	47-55	10***
5	47-55	10***

図 2.2: M_1 と M_3 の間の完全マッチングの一例

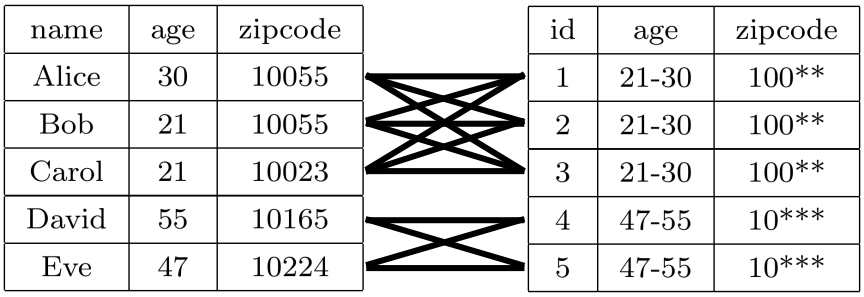
して指摘した [5]．例えば 2-anonymity を満たすように M_1 を加工する場合，M_1 は 5 レコードのデータであるため，どうしても M_3 のように 3 レコードの区別がつかないグループ（レコード 1,2,3）を作らなくてはならない.「少なくとも 2 レコードの区別がつかない」という状態さえ満たせば 2-anonymity を満たしているといえるため，3 レコードの区別がつかないグループは過度な加工であるといえる．

k-concealment[5] は Tamir らによって提案された匿名性指標であり，k-anonymity を改善・拡張したものである．k-concealment は元データと加工データの関係が 2 部グラフ [7] に置き換えられることに注目した指標であり，元データの全レコードが少なくとも k 種類の完全マッチングの辺（match）を持つことを保証する．なお，グラフ (V, E) において，V の分割 V_1, V_2 に対して，全ての辺 $e \in E$ が V_1 と V_2 に属すとき，その時に限り，2 部グラフ (bipartite graph) という．本稿では k-concealment を満たすようにデータを加工することを k-concealment 化と呼ぶ．

定義 2.1.1 (k-concealment) 表 M とその一般化を M'，E を M と M' の間の辺とし，(M, M', E) を *2* 部グラフとする．全ての $d \in M$ と $d' \in M'$ について，$(d, d') \in E$ であり，(d, d') を含むある完全マッチングが存在するとき，d は *match* を持つという．全ての $d \in M$ について，少なくとも k 個の異なる *match* を持つとき，M' を M の *k-concealment* と呼ぶ．

例 2.1.1 M_1 と M_3 の間の完全マッチングの一例を図 *2.2* に示す．これらの辺はすべて *match* である．また，M_1 と M_3 の間のレコード（行）関係を図 *2.3* の *2* 部グラフ (M_1, M_3, E) で表す．元データ M_1 の各レコードと加工後候補として当てはまるレコードの間に辺が張られている．例えば，元データ M_1 の *Alice* のレコードの加工後候補として，加工後データ M_3 では *id=1,2,3* のレコードが当てはまるため，*Alice* のレコードからは *3* 本の辺が張られている．図 *2.3* のすべての辺も完全マッチングの一部になりうるため *match* である．元データ M_1 の各レコードは加工後データ M_3 との間に少なくとも *2* 本の *match* を持つため，M_3 は *2-concealment* を満たしている.（元データと加工後データ間の完全マッチングを攻撃者の再識別パターンと考えるとよい）

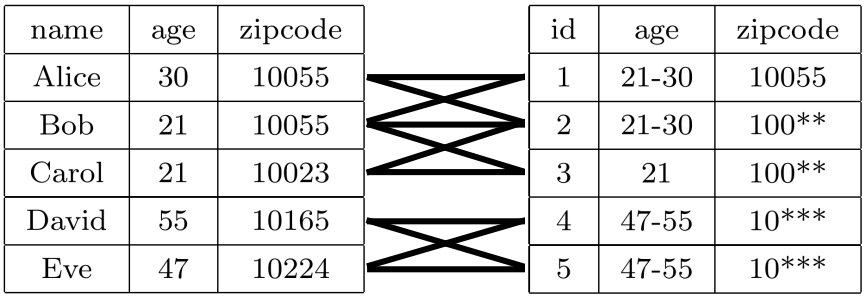
新たな加工後データとして表 2.4 の M_4 を考える．M_4 は M_3 の一部（1,3 行目）を加工前に戻したデータであり，k-anonymity の観点でこのデータを評価すると，1-anonymity しか満たしていないが，k-concealment の観点では 2-concealment を満たしている．M_1 と M_4 の関係を示す 2 部グラフを図 2.4 に示す．これらの辺はすべて match であり，データの一部を元に戻したことによって M_3 より辺の総数は減っているが，各レコードが少なくとも 2 本の match を持っている．そのため，M_4 も「最低でも 2 人の区別がつかない」という状態を満たしている．



name	age	zipcode
Alice	30	10055
Bob	21	10055
Carol	21	10023
David	55	10165
Eve	47	10224

id	age	zipcode
1	21-30	100**
2	21-30	100**
3	21-30	100**
4	47-55	10***
5	47-55	10***

図 2.3: M_1 と M_3 の関係を示す 2 部グラフ (M_1, M_3, E)



name	age	zipcode
Alice	30	10055
Bob	21	10055
Carol	21	10023
David	55	10165
Eve	47	10224

id	age	zipcode
1	21-30	10055
2	21-30	100**
3	21	100**
4	47-55	10***
5	47-55	10***

図 2.4: M_1 と M_4 の関係を示す 2 部グラフ

2-anonymity と 2-concealment を満たす M_3 と，1-anonymity と 2-concealment を満たす M_4 を比較してみよう．どちらも「最低でも 2 人の区別がつかない」という状態を満たす一方，M_4 は M_3 の一部を加工前に戻したデータであるため，M_3 よりも高い有用性を持っている．例えば加工されたセル数でデータの有用性を評価すると，M_3 の有用性は 10，M_4 は 8 であり，M_4 の方が元データに近い（値が小さいほど有用性が高い）．このように，k-concealment の観点でデータを評価・加工することにより，k-anonymity の場合と同等の安全性（最低でも k 人の区別がつかない）で，より高い有用性を持つデータを作ることができる．

2.1.3 最大知識攻撃者モデル

Domingo-Ferrer らは [26] の 3 章で，最大知識攻撃者モデル（maximum-knowledge attacker model）を定義している．最大知識攻撃者は，元のデータセットと匿名化されたデータセットの両方を背景知識として持っていると定める[3]．そのため，最大知識攻撃者は考えられ得る限り最も有効な再識別攻撃を匿名化データに対して行うことができる．このモデルは匿名化データに対する最悪の攻撃者想定であるため，最大知識攻撃者に対して安全である匿名化データは，他のいかなる攻撃者に対しても安全であるといえる．

一般的に，攻撃者が匿名化データを攻撃する動機や，攻撃者が個人の再識別を成功したときに得られる利益として，その個人についての未知の情報が挙げられる．例えば，ある攻撃者が仮名化された M_1 の age 属性のみから Alice の再識別を成功させたとき，Alice の zipcode 情報を追加で得ること

[3]Domingo-Ferrer らの [112] の定義 3 では，“An attack of this class is one in which the intruder knows the entire original data set (plaintext) and the entire corresponding anonymized data set (ciphertext), his objective being to recreate the correct linkage between the original and the anonymized records.” と定義されている．

ができる．しかし，最大知識攻撃者は元のデータセットの全ての値を知っているため，匿名化されたデータを攻撃する動機が無い．Domingo-Ferrer らは，最大知識攻撃者の動機は重要ではなく，あくまで匿名化データに対する攻撃の最悪のケースを想定することが目的であるため，最大知識攻撃者の関心は情報開示ではなく再識別のみである，と述べている[4]．強いて言うならば，最大知識攻撃者の目的は個人についての情報ではなく，匿名化データの評判を落とすこと，とも考えられる．

2.1.4 El Emam の攻撃者モデル

El Emam らは [75] の 2.3 節で，匿名化されたデータに対する 4 種類の再識別脅威 $T1, \ldots, T4$ を想定している．これらの再識別脅威では，いずれも現実世界で起こりうるケースが想定されている．

再識別脅威 $T1$ は，故意による再識別の試みである．例えば，データを管理している組織の不良職員が，そのデータから金銭的な利益を得ようと企むように，データにアクセスできる攻撃者がデータを悪用しようとするケースが考えられる．この場合，再識別脅威の評価に必要な要因は,「攻撃者の動機と能力」と「リスク低減コントロール」の 2 点であり，攻撃者が再識別を試みる確率を Pr(attempt)，再識別が試みられた時の再識別確率を Pr(re-id|attempt) とすると，$T1$ は Pr(re-id,attempt) = Pr(attempt)$\cdot$ Pr(re-id|attempt) と求めることができる．El Emam らは，Pr(attempt) は組織内の不良職員の割合から求められると述べている．

再識別脅威 $T2$ は，故意でない再識別の試みである．例えば，データセットを取り扱う人が，データを分析している時に年齢や郵便番号などの情報から，自分の知人や親戚がそのデータ内にいることに気づくケースが考えられる．この場合，データセット内に分析者の知り合いがいる確率を Pr(acquaintance)，データセット内に知人がいるときの再識別確率を Pr(re-id|acquaintance) とすると，$T2$ は Pr(re-id,acquaintance) = Pr(acquaintance) $\cdot$ Pr(re-id|acquaintance) と求めることができる．El Emam らは，Pr(acquaintance) は人の認知能力の限界数であるダンバー数から求めることができると述べている．

再識別脅威 $T3$ は，データ漏洩である．例えば，大学の職員がモバイル機器の紛失や盗難などによってデータセットを紛失して，そこから個人が識別されてしまうケースが考えられる．この場合，データセットが侵害される確率を Pr(breach)，侵害したデータセットから個人が再識別される確率を Pr(re-id|breach) とすると，$T3$ は Pr(re-id,breach) = Pr(breach) $\cdot$ Pr(re-id|breach) と求めることができる．El Emam らは，Pr(breach) は企業等の組織のデータ漏洩報告から得ることができ，この値は時間とともに変化するであろうと述べている．

再識別脅威 $T4$ は，公開データである．例えば，データが一般に公開されたときに，そのデータセットについての背景知識を持つ攻撃者によって再識別が試みられるケースが考えられる．この場合，$T4$ はデータセットだけに基づく Pr(re-id) と求められる．

攻撃者の振る舞いを表す $T1$ の Pr(attempt)，$T2$ の Pr(acquaintance)，$T3$ の Pr(breach) といった値は，組織によって異なるだけでなく時間によって変化する動的な値である．例えば，組織の職員の待遇によって Pr(attempt) の値が変化することや，組織のセキュリティ対策度合いによって Pr(breach)

[4][26] の 3 章「MAXIMUM-KNOWLEDGE ATTACKER MODEL」を参照されたし．

の値が変化することなどが予想できる．そのため，これらの値を入手するのは困難でありデータのリスクを評価するのは難しい．

2.1.5　高崎によるプライバシリスク評価

高崎は [110] の 4 章で，スマートフォンを介して収集したライフログデータのプライバシリスク評価を行っている．高崎らは，自らが開発したライフログ収集スマートフォンアプリケーション「マカロン」をダウンロードした被験者 1 万人のうち，アプリケーションを毎日起動した 463 名の被験者に対してアンケート調査を行った．そのアンケート結果をロジスティック回帰によって分析することにより，個人属性（性別，年齢，など），保護制度（透明性，変更可能性，など），懸念の質（漏洩，不正利用，など）といった説明要因が，懸念の増加，懸念の減少，懸念の消失といった被説明要因にどのように影響するかを調査している．

利用者のプライバシ懸念の増加に対して分析を行った結果，情報の取り扱いについて事前に規定しておくことによって利用者の懸念増加を抑制できることや，サービス利用前に発生したサービス自体への不安は払拭が難しいことや，インターネットを利用したサービス全般への本来的な懸念は特定のサービスへの懸念増加にはつながらないことが明らかになった．

また，利用者のプライバシ懸念の減少に対して分析を行った結果，いずれの説明要因にも有意と認められる変数はなく，サービス利用を通じて懸念の増加を食い止めることはできても，減少させることは難しいことが明らかになった．さらに，利用者のプライバシ懸念の消失に対して分析を行った結果，サービス提供事業者への信頼感を持つ利用者に対しては，特に対処をしなくても懸念が消える可能性が高いことや，パーソナルデータ提供後にその情報を変更/削除する制度があっても，利用者の懸念は消えないことが明らかになった．

このように，高崎はアプリケーション利用者から収集したアンケート結果を分析することによって，データやアプリケーションのどういった点が利用者のプライバシ懸念に影響を与えているのかを実験的に明らかにした．しかしながら，この調査結果は個人差が大きく，プライバシ侵害に対して必ずしも合理的に考えていないため，利用者のプライバシ指向度によって，具体的にデータのどの部分が危険であるのか，またどの部分を優先的に加工したらいいのか，という評価はできないことが示された．

2.1.6　医療情報分析の先行研究

Arafa ら [31] は，Japan Public Health Center (JPHC) 研究プロジェクトの多目的コホートを用いて，89,000 名の成人男女のデータを調査し，尿路結石の既往症の有無による心疾患のリスクを明らかにした．Cox 比例ハザードモデルにより定量化した尿路結石によるハザード比は，1.04 であり統計的な有意水準には達しないことが分かった．また，Islami ら [32] は，2011 年から 2015 年に米国で罹患したがん患者のデータ (the United States Cancer Statistics (USCS) Public Use database) を調べ，全米の州における肥満によるがん罹患のリスクの違いを調査した．BMI が 5 単位増加することによる相対リスク RR は，1.31 (胃がん)，1.59 (肝臓がん)，1.10 (乳がん）であり，南部や中西部の州では特に高いリスクを検出している．Saint-Maurice ら [33] は，4840 名の平均 56 歳の成人男性の活動を

加速度計で 7 日間測定し，10 年間に渡る追跡コホート調査により，平均歩数が死亡率に及ぼす影響を定量化している．Cox 比例ハザードモデルにより，4000 歩を基準 (=1) とした時の 8000 歩，12000 歩の被験者の死亡率が単調に下がり，ハザード比でそれぞれ 0.49 と 0.35 になることを示した．Chen ら [34] は，1999 から 10 年間の NHANES (National Health and Nutrition Examination Survey) データを用いて，3 万人の米国成人の栄養補助食品 (dietary supplement) の利用有無による，心疾患とがんの死亡率の変化を調査している．適量なサプリメントの取得は死亡率を下げるが，カルシウムを過剰に取るとがんによる死亡率が調整済みリスク比で，1.62 に増加することを明らかにしている．

このように，コホート研究においては死亡を目的変数とした Cox 比例ハザードモデルによる分析が主流である．罹患を目的変数としてロジスティック回帰分析を用いている研究 [32] は少ない．

第3章　部分的な背景知識を持つ攻撃者を想定した再識別リスク評価モデルの提案と評価

3.1　導入

データから個人が識別されるリスクは，そのデータからの個人識別を試みる攻撃者に大きく依存する．攻撃者はデータについての何らかの背景知識を持っていることが考えられるが，背景知識の量や質によってその危険度は大きく変化する．しかし，既存の攻撃者モデルである最大知識攻撃者モデル [26] は元データ全てを背景知識として持つ強すぎるものである．また，El Emam の Dunbar 数モデル [75] は知り合いの中に攻撃者がいることを想定している楽観的すぎるものである．

本章では，新たな攻撃者想定として，データ中の 1 つの値のみを背景知識として持つ攻撃者モデルを提案する．この攻撃者が個人の識別を成功させる確率の平均値（平均識別確率）を用いてデータの安全性を評価する．また，再識別リスク評価にかかる時間短縮のために，平均識別確率を近似する数理モデルを提案し，データの安全性を理論的に評価する．

3.2　基礎定義

3.2.1　データモデル

本研究では，レコード（行）と属性（列）によって構成され，個人を表す識別子を持つ履歴データ（トランザクションデータ）を研究する．記号等を以下のように定義する．

定義 3.2.1 T をトランザクションデータとする．T には n 人の個人による m 個の履歴を表すレコードが含まれている．D_A を T の属性 A の集合とする．R_a と U_a をそれぞれ，$a \in D_A$ を含むレコードの集合と属性 A に値 a を持つ個人の集合とする．T'' を T を仮名化（識別子属性を仮名に置き換える）したデータとする．

例 3.2.1 表 *3.1* に *3* 人の個人 $(1,2,3)$ の *3* 日間 (2010/12/1 – 2010/12/3) のトランザクションデータ T_{Example} を示す．例えば，個人 *2* は (2010/12/1) にパンを *3* つ買っていることがデータからわかる．この場合，$m = 10, n = 3$ であり，A が **date** のとき D_A は $\{(2010/12/1), (2010/12/2), (2010/12/3)\}$ であり，$|D_A| = 3$ である．また $a = (2010/12/1)$ のとき，$R_a = \{1,2,3,4\}$ であり $U_a = \{1,2\}$ である．表 *3.2* に加工データ例 T''_{Example} を示す．T''_{Example} は仮名化されており，T の **user ID** 属性の値（*1,2,3*）が仮名（*A,B,C*）に置き換えられている．

表 3.1: データ例 T_{Example}

ID	user ID	date	time	goods	price	quantity
1	1	2010/12/1	8:45	Bread	1.45	2
2	1	2010/12/1	8:45	Book	3.75	1
3	1	2010/12/1	20:10	Tea	0.85	2
4	2	2010/12/1	10:03	Bread	1.45	3
5	1	2010/12/2	15:07	Tea	0.85	3
6	3	2010/12/2	11:57	Bread	1.45	4
7	3	2010/12/2	11:57	Juice	1.25	4
8	3	2010/12/3	15:54	Book	3.75	1
9	3	2010/12/3	15:54	Tea	0.85	10
10	3	2010/12/3	15:54	Juice	1.45	10

表 3.2: 加工データ例 T''_{Example}

ID	pseudo ID	date	time	goods	price	quantity
1	A	2010/12/1	8:45	Bread	1.45	2
2	A	2010/12/1	8:45	Book	3.75	1
3	A	2010/12/1	20:10	Tea	0.85	2
4	B	2010/12/1	10:03	Bread	1.45	3
5	A	2010/12/2	15:07	Tea	0.85	3
6	C	2010/12/2	11:57	Bread	1.45	4
7	C	2010/12/2	11:57	Juice	1.25	4
8	C	2010/12/3	15:54	Book	3.75	1
9	C	2010/12/3	15:54	Tea	0.85	10
10	C	2010/12/3	15:54	Juice	1.45	10

3.2.2 攻撃者モデル

本研究では，攻撃者がトランザクションデータ T に属するユーザの属性 A についての背景知識 a を偶然得ることを想定する．攻撃者が背景知識 a を得る確率は，データ中のレコードが a を含む頻度に比例すると仮定しよう．例えば，攻撃者がスーパーマーケットの前で買い物をしている顧客を目撃したとき,「13 時に買い物をした顧客」のような頻繁に起こる出来事は,「深夜 2 時に買い物をした顧客」のような珍しい出来事よりも，背景知識として得やすいと考えられる．この場合，属性 A は「購買時刻」であり,「13 時」や「2 時」といった値が a である．

定義 3.2.2 攻撃者が背景知識 a を得る確率 $Pr(a)$ は，T 内の a の頻度に比例する．すなわち，$Pr(a) = |R_a|/m$ となる．

仮名化データ T'' を与えられた攻撃者は，背景知識として a を含む T のレコードにアクセスできる

とき，対応する T'' の仮名の真のユーザの候補として U_a を得る．従って，再識別を表す事象 idf が生起するリスクを，a の条件付確率として次のように定める．

定義 3.2.3 攻撃者に背景知識 a が知られたとき，個人を識別 idf する条件付確率を識別確率 $Pr(\mathrm{idf}|a) = 1/|U_a|$ とする．

定義 3.2.2,3.2.3 より，攻撃者が背景知識 a を得ることと，攻撃者が背景知識 a から個人を識別することの同時確率 $Pr(\mathrm{idf}, a)$ は，

$$Pr(\mathrm{idf}, a) = Pr(a)Pr(\mathrm{idf}|a) = \frac{|R_a|}{m}\frac{1}{|U_a|}$$

である．また，$\alpha_a = |R_a|/|U_a|$ とおくと，$Pr(\mathrm{idf}, a) = \frac{\alpha_a}{m}$ とも表せる．ここで，α_a は a についてのユーザ当たりの平均レコード数 [レコード/人] を意味しており，本論文の解析に重要な役割を果たす．そこで，これを次のように定義する．

定義 3.2.4 α_a を，属性 A についての背景知識 a についての平均レコード数とする．α_A を，属性 A についての α_a の平均値とする．すなわち，$\alpha_A = \frac{1}{|D_A|}\sum_{a\in D_A}\alpha_a$ である．

例 3.2.2 T_{Example} の **date** 属性についての a, $|R_a|$, $Pr(a)$, $|U_a|$, $Pr(\mathrm{idf}|a)$,$Pr(\mathrm{idf}, a)$ を表 *3.3* に示す．この場合，D_A は $\{2010/12/1, 2010/12/2, 2010/12/3\}$ である．攻撃者が背景知識 $a = 2010/12/3$ を得る確率は，$R_a = \{8, 9, 10\}, m = 10$ であるため $Pr(a) = 3/10$ である．その背景知識からユーザ u を識別できる確率 $Pr(\mathrm{idf}|a)$ は，$U_a = \{3\}$ なので，$Pr(\mathrm{idf}|a) = 1/1$ となる．この場合，攻撃者が背景知識 a によって顧客 u を識別できる確率は

$$Pr(\mathrm{idf}, a) = Pr(a)Pr(\mathrm{idf}|a) = 0.3 \cdot 1 = 0.3$$

である．または，$\alpha_a = 3/1 = 3$ であるので，

$$Pr(\mathrm{idf}, a) = \frac{\alpha_a}{m} = \frac{3}{10} = 0.3$$

とも計算できる．

私の提案する手法では，データがどのような手法で加工されているかに関わらず，攻撃者がターゲットの情報を得る確率とリスクの大きいレコードの数を用いて，リスクがどれだけ下がるかをモデル化する．

定義 3.2.5 (平均識別確率) 攻撃者がある属性 A の背景知識 a を用いて個人の識別に成功する確率の平均値を平均識別確率 $Pr(\mathrm{idf}, A)$ とする．

平均識別確率は

$$Pr(\mathrm{idf}, A) = \sum_{a\in D_A} Pr(a)Pr(\mathrm{idf}|a) = \sum_{a\in D_A} Pr(\mathrm{idf}, a)$$

と計算できる．

表 3.3: T_{Example} の **date** 属性についての識別確率

a	$\lvert R_a \rvert$	$Pr(a)$	$\lvert U_a \rvert$	$Pr(\text{idf}\lvert a)$	$Pr(\text{idf}, a)$	α_a
2010/12/1	4	0.4	2	0.5	0.2	2
2010/12/2	3	0.3	2	0.5	0.15	1.5
2010/12/3	3	0.3	1	1	0.3	3
Sum	10	1.0			0.65	

例 3.2.3 T_{Example} の場合，*"2010/12/1"* を含む *4* レコードの識別確率は $Pr(\text{idf}\lvert(2010/12/1)) = 1/2$ であり，*"2010/12/2"* を含む *3* レコードの識別確率は $Pr(\text{idf}\lvert 2010/12/2) = 1/2$ であり，*"2010/12/3"* を含む *3* レコードの識別確率は $Pr(\text{idf}\lvert 2010/12/3) = 1$ である．よって，A の平均識別確率は $Pr(\text{idf}, A) = 13/20$ である．

定義 3.2.4,3.2.5 より，平均識別確率は

$$Pr(\text{idf}, A) = \sum_{a \in D_A} Pr(\text{idf}, a) = \sum_{a \in D_A} \frac{\alpha_a}{m}$$

と計算することができる．

例 3.2.4 T_{Example} と $A =$ **date** が与えられた時，平均識別確率は

$$Pr(\text{idf}, A) = \sum_{a \in D_A} \frac{\alpha_a}{m} = \frac{2 + 1.5 + 3}{10} = 0.65$$

と計算できる．つまり，T_{Example} の **date** 属性からある背景知識を得た攻撃者は，平均 *65%*の確率で個人を識別することができる．

定義 3.2.6 平均識別確率 $Pr(\text{idf}, A)$ を属性 A のリスク評価値とみなす．この評価値を求めるために必要なデータのレコード数や計算時間を計算コストとする．

3.3 履歴データの属性の安全性

平均識別確率を求めるためには，定義 3.2.5 より，履歴 T の属性 A に出現するすべての a について，α_a を求める必要がある．しかしながら，ビッグデータに対してすべての α_a を計算するのは困難である．例えば，3.4.4 節で後述する計算コストの評価実験では，$n = 400, m = 38,087$, 属性数 6 の購買履歴データの全属性の平均識別確率を計算するために 27.5 秒かかることが明らかになっており，さらに大きなデータの計算にはより長い時間が必要であると予想できる．そのため，これ高速に近似する方法を検討する．平均識別確率を計算するモデルとして，本章では (1) 平均モデル，(2) 最小コストモデル, (3) サンプリングモデルの 3 つを提案する．

3.3.1 厳密解

匿名化データ T'' の再識別のリスクは，攻撃者に与えられる背景知識の属性 A に依存して決まる．そこで，A を与えられた時の再識別リスク $R(A)$ を，属性 A の平均識別確率と定める．すなわち，$R(A) = Pr(\mathrm{idf}, A)$ とする．$R(A)$ の厳密解を求めるためには，履歴 T の属性 A に出現するすべての a について α_a を求める必要があるため，この場合の計算コストは m である．

3.3.2 平均モデル

平均モデルは，属性 A のリスクを α_a の平均値 α_A を用いて求めるモデルである．以下のように定義を行う．

定義 3.3.1 属性 A のリスク $R_{mean}(A)$ を，

$$R_{mean}(A) = \frac{\alpha_A |D_A|}{m}$$

で与えるモデルを平均モデルと呼ぶ．

平均モデルで計算されたリスク評価値は，次のように厳密解と一致する．平均モデルによる属性 A のリスク評価値が $R_{mean}(A) = Pr(\mathrm{idf}, A)$ であることを示す．定義 3.3.1 より，$R_{mean}(A)$ は $R_{mean}(A) = \alpha_A |D_A|/m$，定義 3.2.4 より，$\alpha_A$ は $\alpha_A = \frac{1}{|D_A|}\sum_{a \in D_A} \alpha_a$ である．よって，以下のように $R_{mean}(A)$ の式を得る．

$$R_{mean}(A) = \frac{\alpha_A |D_A|}{m} = \frac{|D_A|}{m} \sum_{a \in D_A} \frac{\alpha_a}{|D_A|} = \sum_{a \in D_A} \frac{\alpha_a}{m} = \sum_{a \in D_A} Pr(a) Pr(\mathrm{idf}|a) = Pr(\mathrm{idf}, A)$$

例 3.3.1 $T = T_{\mathrm{example}}$, $A =$ **date** の場合，$Pr(\mathrm{idf}, A) = \frac{\alpha_A |D_A|}{m} = \frac{\frac{13}{6} \cdot 3}{10} = 0.65$ である．

このモデルでは α_a を求める際に，履歴 T の属性 A に出現するすべての a について α_a を計算する必要があるため，この場合の計算コストは m である．

3.3.3 最小コストモデル

3.4.2 節で後述する観測より，顧客ごとの平均レコード数 (α_a) が多くのデータで 1 に近いことを発見した．そこで，全ての顧客の α_a を 1 であると仮定することにより，α_a を計算するコストを 0 にする手法を以下のように提案する．

定義 3.3.2 属性 A のリスク評価値を $R_{cost}(A) = \frac{|D_A|}{m}$ とするモデルを最小コストモデルと呼ぶ．

例 3.3.2 $T = T_{\mathrm{example}}$, $A =$ **date** のとき，$R_{cost}(A) = \frac{|D_A|}{m} = \frac{3}{10} = 0.3$ である．

定理 3.3.1 最小コストモデルの誤差率は $|1 - \frac{1}{\alpha_A}|$ である．

(証明) $R_{cost}(A)$ の厳密解に対する誤差率 err は,

$$err = \frac{|R_{cost}(A) - Pr(\mathrm{idf}, A)|}{Pr(\mathrm{idf}, A)} = \frac{|\frac{|D_A|}{m} - \frac{\alpha_A |D_A|}{m}|}{\frac{\alpha_A |D_A|}{m}} = |\frac{1}{\alpha_A} - 1|$$

と与えられ，定理 3.3.1 を得る． (Q.E.D)

このモデルで用いる T のレコード数 m と A の種類数 $|D_A|$ は本研究では与えられているため，リスク計算のコストは 0 である．

3.3.4 サンプリングモデル

サンプリングモデルは，D_A からランダムに選んだ複数個の要素についての α_a を求め，これの平均を属性 A の平均レコード数 α_A の近似値とみなして属性 A のある値を知る攻撃者のリスクを求めるモデルである．レコードの一様サンプリングではないサンプリングの手順を説明する．まず，属性 A に含まれる $|D_A|$ 種類の値 $a_{(1)}, \ldots, a_{(|D_A|)}$ から，ランダムに複数個の値 $D'_A = \{a'_{(1)}, \ldots, a'_{(|D'_A|)}\}$ をサンプリングする．次に，データ T のレコードのうち，属性 A に値 $a' \in D'_A$ を持つレコードを全て抽出する．最後に，抽出したデータの全ての値 $a' \in D'_A$ について $\alpha_{a'}$ を求め，それらの平均値を求める．例えば，T_{Example} の **date** 属性のうち “2010/12/1” がランダムに選ばれた場合，T_{Example} からこれを満たすレコード（この場合 4 レコード）をすべて抽出し，これから $\alpha_{a'}$ や平均値を求める．以下のように定義を行う．

定義 3.3.3 D_A からランダムに $|D'_A|$ 個の要素をサンプリングした集合を $D'_A = \{a_1, \ldots, a_{|D'_A|}\}$ とし，$\alpha_{a'} = \frac{1}{|D'_A|} \sum_{i=1}^{|D'_A|} \alpha_{a_i}$ とする．属性 A のリスク評価値を $R_{sample}(A) = \frac{\alpha_{a'} |D_A|}{m}$ で与えるモデルをサンプリングモデルとする．

例 3.3.3 $T = T_{\mathrm{example}}$, $A =$ **date**, $|D'_A| = 2$, $D'_A = \{2010/12/1, 2010/12/3\}$ のとき，$\alpha_{a_1} = 2$, $\alpha_{a_2} = 3$ となるので，$R_{sample}(A) = \frac{\alpha_{a'} |D_A|}{m} = \frac{2.5 \cdot 3}{10} = 0.75$ と計算できる．

$|D'_A|$ 個のサンプルの標準偏差を $\sigma_{|D'_A|}$ とする．

定理 3.3.2 *90%*信頼区間を仮定したとき，サンプリングモデルの誤差率の最大値は $\frac{\sigma_{|D'_A|} m}{\sqrt{||D'_A||} |D_A| \alpha_A}$ である．

(証明) リスク評価値の厳密解は $Pr(\mathrm{idf}, A) = \alpha_A\ |D_A|/m$ であるため，90%信頼区間を仮定すると絶対誤差は $|R_{sample}(A) - Pr(\mathrm{idf}, A)| < Var[Pr(\mathrm{idf}, A)] = \sigma_{|D'_A|}/\sqrt{|D'_A|}$ となる．よって，$R_{sample}(A)$ と厳密解の相対誤差率は

$$= \frac{|R_{sample}(A) - Pr(\mathrm{idf}, A)|}{Pr(\mathrm{idf}, A)} < \frac{\frac{1}{\sqrt{|D'_A|}} \sigma_{|D'_A|}}{\frac{\alpha_A |D_A|}{m}} = \frac{\sigma_{|D'_A|} m}{\sqrt{||D'_A||} |D_A| \alpha_A}$$

となるので，定理 3.3.2 を得る． (Q.E.D)

表 3.4: 提案近似モデルの誤差率とコスト

Model	Risk	Error Rate	Cost
Exact Solution	$R(A)$	0	m
Mean	$R_{mean}(A)$	0	m
Low Cost	$R_{cost}(A)$	$\frac{1}{\alpha_A} - 1$	0
Sampling	$R_{sample}(A)$	定理 3.3.2	$\|D'_A\|m/\|D_A\|$

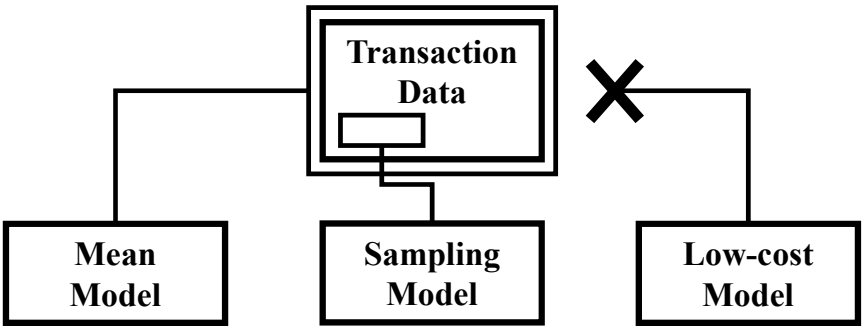


図 3.1: 提案する 3 つの近似モデルのイメージ図

このモデルでの $\alpha_{a'}$ の計算コストは，D'_A の要素が $|D_A|$ から確率 $1/|D_A|$ で一様に選ばれるならば，$|D'_A|m/|D_A|$ である．

表 3.4 に各モデルの概要をまとめる．厳密解と平均モデルの計算コストは最大（m）であるが，誤差はゼロである．一方，最小コストモデルのコストはゼロであるが，誤差は大きくなる．サンプリングモデルの計算コストと誤差はサンプリングサイズ $|D'_A|$ に依存し，$|D'_A|$ が大きくなるほど誤差は小さくなり，計算コストは大きくなる．図 3.1 に各近似モデルのイメージを示す．リスク評価値の近似をするために，平均モデル（Mean Model）はデータの全てを利用し，サンプリングモデル（Sampling Model）はデータの一部を利用し，最小コストモデル（Low-cost Model）はデータを全く利用しない．

3.4 評価実験

3.4.1 実験目的

前節で提案したモデルを用いて，実際のデータに対するリスク評価実験を行う．実験のために，以下に示す，UCI Machine Learning Repository[61] より公開されている 3 つのデータセットと Lending Club [80] より公開されている 1 つのデータを用いる．

1. T_1: Online Retail Dataset, イギリスのオンラインショップの 1 年間の購買履歴データ [57]
2. T_2: Diabetes Dataset, 10 年間の糖尿病患者入院データ [81]
3. T_3: Adult Dataset, 国勢調査によって収集された世帯収入データ [12]
4. T_4: LOAN DATA, 2007 年から 2011 年の間のローン借り入れデータ．[82]

表 3.5: T_1, T_2, T_3, T_4 の詳細と各属性のリスク評価値

T	m	n	#Attribute	A	Description	α_A	$\lvert D_A \rvert$	$Pr(\mathrm{idf}, A)$
T_1	38,087	400	7	**time**	Purchase time (hh:mm)	22.23	551	0.322
				date	Purchase date (yyyy/mm/dd)	24.42	290	0.186
				goods	ID of purchased goods (number and character)	1.32	2,781	0.097
				price	Price of purchased goods (Pound sterling)	2.49	184	0.012
				quantity	Quantity of purchased goods (number)	3.15	97	0.008
T_2	101,766	71,518	50	**days**	Days in hospital (number)	1.05	14	$1.45 \cdot 10^{-4}$
				age	Age of patient (number)	1.35	10	$1.33 \cdot 10^{-4}$
				ethnicity	Ethnicity of patient (character)	1.31	6	$7.73 \cdot 10^{-5}$
				gender	Gender of patient (character)	1.28	3	$3.78 \cdot 10^{-5}$
T_3	32,561	32,561	16	**age**	Age of user (number)	1	73	$2.24 \cdot 10^{-3}$
				occupation	Occupation of user (character)	1	15	$4.61 \cdot 10^{-4}$
				martial	Marital status of user (character)	1	7	$2.15 \cdot 10^{-4}$
				ethnicity	Ethnicity of user (character)	1	5	$1.54 \cdot 10^{-4}$
T_4	42,538	42,538	145	**employment**	Employment of customers (character)	1	30,661	0.721
				income	Annual income of customers (number)	1	5,597	0.132
				amount	Amount of loan (number)	1	898	0.021
				grade	Grade of customers (character)	1	8	0.000

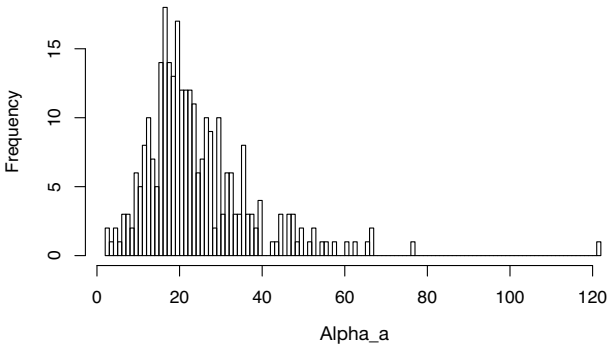


図 3.2: $T = T_1, A =$ **date** のときの α_a の分布

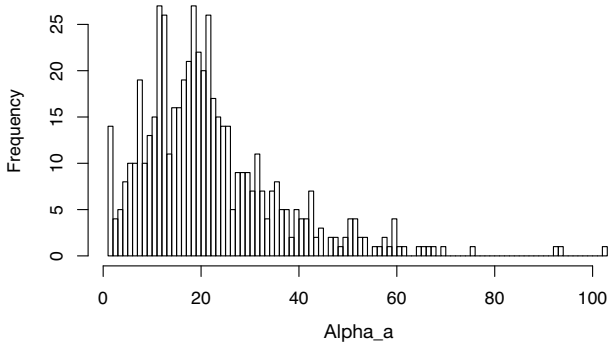


図 3.3: $T = T_1, A =$ **time** のときの α_a の分布

各データの m, n を表 3.5 に示す．T_1, T_2 は $m > n$ のデータであり，T_3, T_4 は $m = n$ のデータである．T_3 と T_4 は厳密にはトランザクションデータでは無いが，本実験ではトランザクションデータとして扱う．

3.4.2 データセットの分析

表 3.5 に T_1 の各属性の概要を示す．このデータは 7 属性から成るデータであるが，本研究ではユーザ ID・伝票 ID を除いた 5 属性 (**date**, **time**, **goods**, **price**, **quantity**) を A の候補として用いる．各属性の α_a の分布を図 3.2–3.6 に示し，各属性についての α_A と $|D_A|$ を表 3.5 に示す．

date,**time** 属性はユーザごとの平均レコード数 α_a が大きく，100 レコードを超える値もあり，例えば (2011/8/28) には 1 人のユーザが 122 レコードの購買をしている（a = 2011/8/28, $\alpha_{2011/8/28} = 122$）．本研究ではこういった値は，背景知識として得やすく，これを得た攻撃者が個人を識別しやすいので，大変危険であると評価される．一方，**goods**, **price**, **quantity** 属性では α_a が小さく，多くの値 a について $\alpha_a = 1$ である．

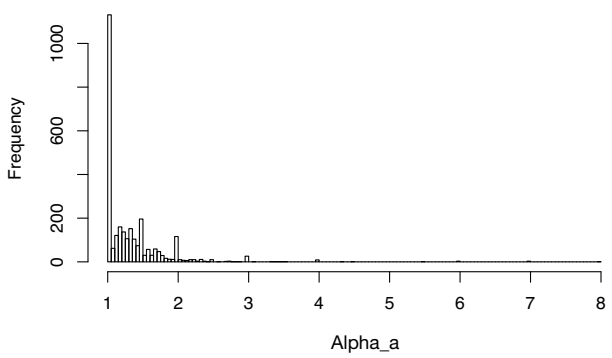


図 3.4: $T = T_1, A =$ **goods** のときの α_a の分布

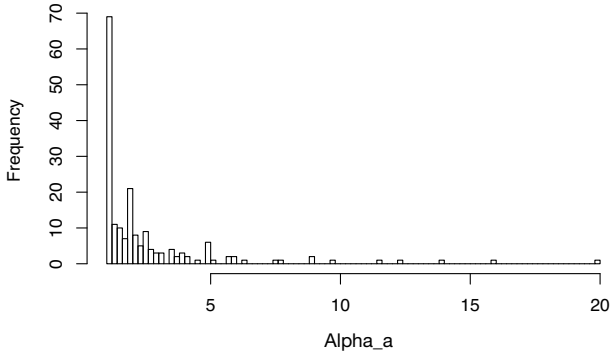


図 3.5: $T = T_1, A =$ **price** のときの α_a の分布

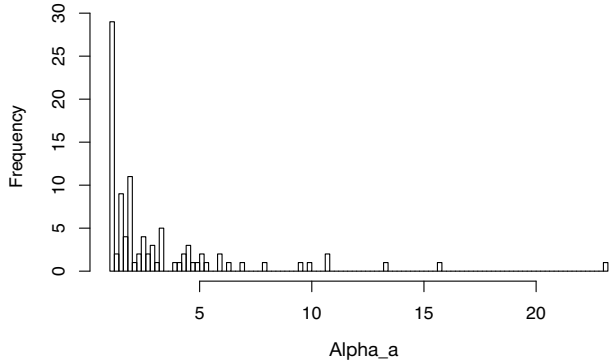


図 3.6: $T = T_1, A =$ **quantity** のときの α_a の分布

a 軸を $|U_a|$，y 軸を $|R_a|$ とした，T_1 の **date**, **price** 属性についての散布図を図 3.7, 3.8 に示す．赤直線は $|U_a| = \alpha_A |R_a|$（平均モデル）を示し，緑直線は $|U_a| = |R_a|$（最小コストモデル）を示す．

表 3.5 に T_2 , T_3, T_4 の各属性の概要を示す．T_2 は 50 属性から成るデータであるが，本研究ではそのうち，攻撃者が背景知識として得ることが想定される 4 属性 (**ethnicity**, **gender**, **age**, **time**) に注目する．T_3 は 17 属性から成るデータであるが，本研究ではそのうち，攻撃者が背景知識として得ることが想定される 4 属性 (**age**, **martial**, **occupation**, **ethnicity**) に注目する．同様に，表 3.5 に各データの各属性の α_A と $|D_A|$ も示す．図 3.9, 3.10 に T_2 の **age**, **date** 属性の α_a の分布を示す．

$T = T_3$ のとき，$m = n$, $|R_a| = |U_a|$ であるので，任意の A の任意の a で $\alpha_A = 1$ である．

3.4.3 リスク評価結果

前述した各属性について，リスク評価値の厳密解を求めた．表 3.5 に T_1, T_2, T_3, T_4 の各属性の $R(A)$ を示す．例えば，T_1 の **date** 属性のリスク評価値は 0.186 であり，T_1 で最も危険と評価された属性は **time** 属性（評価値 0.322）であった．つまり，T_1 を加工する際は **time** 属性を，値の丸め込み（例えば 8:45 から 8:00 に加工する）や摂動化（例えば 8:45 から 8:42 に加工する）等の手法で加工するべきである．匿名化では **time** 属性や **date** 属性などの属性が削除されることが多いので，この結果は理にかなっていると言える．

T_2 と T_3 の場合，$|D_A|$ と α_A が小さいのでリスク評価値もかなり小さくなっており，T_2 では **days** 属性，T_3 では **age** 属性が最も危険な属性と評価されている．T_2, T_3, T_4 では，平均レコード数 α_A が

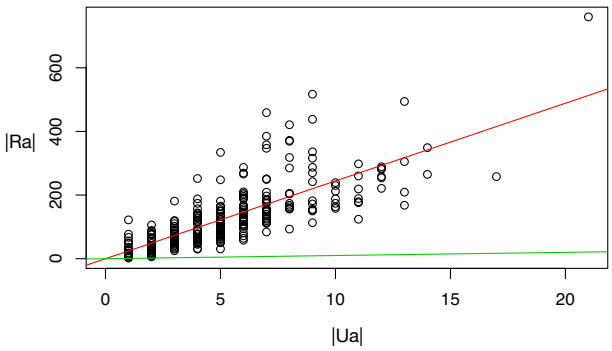


図 3.7: $T = T_1, A =$ **date** のときの $|R_a|$ と $|U_a|$ の散布図

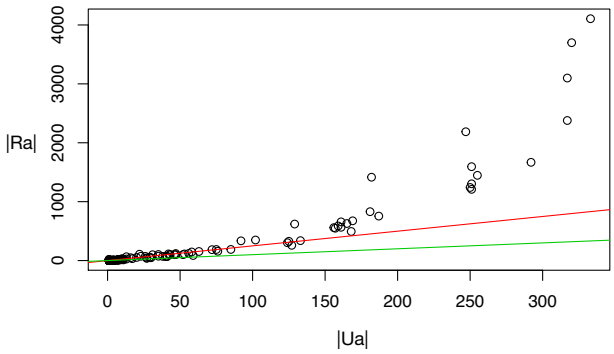


図 3.8: $T = T_1, A =$ **price** のときの $|R_a|$ と $|U_a|$ の散布図

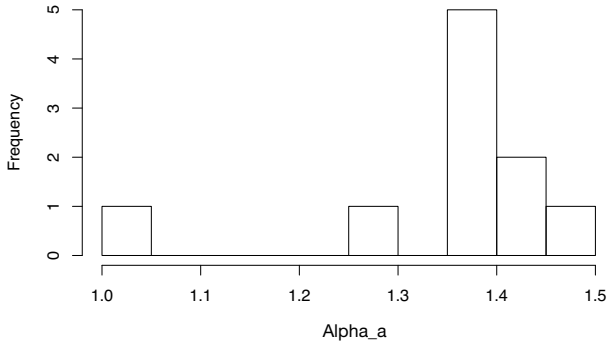


図 3.9: $T = T_2, A =$ **age** のときの α_a の分布

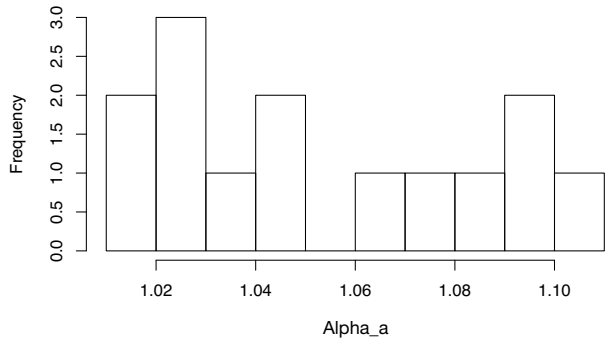


図 3.10: $T = T_2, A =$ **days** のときの α_a の分布

1 に近いことと，平均識別確率 $Pr(\mathrm{idf}, A)$ の順位が値の種類数 $|D_A|$ の順位と等しいことに注意せよ．T_1 のように大きな α_A を持つデータのリスクを評価する際には気を付ける必要があるが，提案モデルは任意のデータに適用することができる．

3.4.4 提案モデルの精度と計算コスト

T_1, T_2, T_3, T_4 の各属性の危険度を平均モデル，最小コストモデル，サンプリングモデルによって求める．表 3.6 に各モデルの評価値を示す．平均モデルによる評価値 $R_{mean}(A)$ は表 3.5 の $R(A)$ と一致する．サンプリングモデルによる評価値 $R_{sample}(A)$ は，$|D'_A| = 10$ のときの $90\%(\mu \pm \sigma)$ の信頼区間を示している．表中の*印がついている値は，そのデータで最も危険であると評価された属性のリスクである．例えば T_1 について，平均モデル（=厳密解）では **time** 属性が最も危険であると評価されているのに対し，最小コストモデルでは **goods** 属性が最も危険であると評価されている．サンプリングモデルにおいては，信頼区間の半順序関係における極大値となる属性は **time** であった．

図 3.11 に $T = T_1$, $A =$ **date** のときの，各モデルの計算コストと誤差の散布図を示す．横軸は計算コスト（レコード数）の対数であり，縦軸は厳密解 $Pr(\mathrm{idf}, A)$ との絶対誤差である．図中の赤い点が各モデルの結果を表している．灰色の点は D_A の要素数のリスク評価結果を示しており，それらの重心をサンプリングモデルの代表の点としている．サンプリングモデルはこれらの $|D_A|$ 個の点から $|D'_A|$ 個をランダムに選んでリスク評価をすることに注意せよ．表 3.7 に各モデルのコストと誤差の値

表 3.6: 各モデルによるリスク近似結果

T	A	$R_{mean}(A)$	$R_{cost}(A)$	$R_{sample}(A)(\lvert D'_A\rvert = 10)$
T_1	**time**	*0.3217	0.0145	*[0.1411, 0.5998]
	date	0.1860	0.0076	[0.1267, 0.2786]
	goods	0.0965	*0.0730	[0.0718, 0.0982]
	price	0.0121	0.0048	[0.0036, 0.0132]
	quantity	0.0080	0.0025	[0.0017, 0.0152]
T_2	**days**	*1.45E-04	*1.38E-04	*[1.46E-04, 1.52E-04]
	age	1.33E-04	9.83E-05	[1.21E-04 , 1.42E-04]
	ethnicity	7.73E-05	5.90E-05	[6.92E-05, 8.31E-05]
	gender	3.78E-05	2.95E-05	[3.08E-05 , 4.30E-05]
T_3	**age**	*2.24E-03	*2.24E-03	*[2.24E-03, 2.24E-03]
	occupation	4.61E-04	4.61E-04	[4.61E-04, 4.61E-04]
	martial	2.15E-04	2.15E-04	[2.15E-04, 2.15E-04]
	ethnicity	1.54E-04	1.54E-04	[1.54E-04, 1.54E-04]
T_4	**employment**	*0.7208	*0.7208	*[0.7208, 0.7208]
	income	0.1316	0.1316	[0.1316, 0.1316]
	amount	0.0211	0.0211	[0.0211, 0.0211]
	grade	0.0002	0.0002	[0.0002, 0.0002]

を示す.

T_1 の 6 つの属性を 3 つのモデル（平均モデル，サンプリングモデル，最小コストモデル）で評価する際にかかる計算時間を，表 3.9 に示す．計算時間の測定には R 言語 [8] を用いており，実験環境は OS：Windows 10, CPU：Intel Core i5-8400, メモリ：16GB である．なお，サンプリングモデルの計算時間は，サイズ 10 のランダムサンプリングを 10 回したときの平均値を示している．T_1 の場合，全属性の平均識別確率の厳密解を求めるために 27.49 秒かかるのに対し，サンプリングモデルでは平均 0.58 秒，最小モデルでは 0.12 秒で求めることができている．T_1 は個人数 400，レコード数 38,087，属性数 6（ID 属性を除く）のデータであるが，これらの数が大きくなるほどこの計算時間は増える．例として，T_1 のレコード数のみが 2 倍になったデータのリスク評価にかかる時間を表 3.10 に示す．表 3.9 の結果と比較すると，計算にかかる時間が約 2 倍になっていることがわかる．購買履歴データのような履歴データは時間経過によって個人数やレコード数が増えることが予想されるので，そういったデータのリスク評価の際にはサンプリングモデルや最小コストモデルが有用である．

T_1 の **date** 属性の D_A から 50 種類の値 a を 1,000 回ランダムサンプリングしたときの α_A の分布を図 3.12 に示す．また，サンプリング種類数毎の α_A の平均と標準偏差を表 3.8 に示す．これらの図表からわかるように，属性 A からランダムな値 a を選び，それについての α_a を求めることで，α_A を近似することができる．サンプリングのサイズに応じて，急速に真値に収束していることがわかる．これにより，本章では $|D'_A| = 10$ でリスク評価を行った．

表 3.7: 各モデルの計算コスト (レコード数) と誤差率

Model	Cost	Error
Mean	38087	0
Sample	131.3	0.073
# Records	0	0.178

表 3.8: $T = T_1$, $A =$ **date** のときのサンプリング種類数毎の α_A の平均と標準偏差（試行回数 1,000 回）

#Sample	mean	σ
1	24.03	13.33
50	24.47	1.75
100	24.39	1.09
150	24.41	0.78
200	24.41	0.53
250	24.42	0.33
$\lvert D_A \rvert$	24.42	0

表 3.9: 3 つのモデル（平均モデル=厳密解，サンプリングモデル，最小コストモデル）で T_1 の各属性を評価した際の計算時間

属性名	平均モデル（厳密解）[s]	サンプリングモデル [s]	最小コストモデル [s]
伝票 ID	19.14	0.13	0.01
購買日	0.22	0.02	0
購買時刻	0.47	0.03	0.02
購買商品	1.95	0.03	0.03
価格	5.57	0.34	0.04
数量	0.14	0.03	0.02
合計	27.49	0.58	0.12

3.4.5 攻撃者が個人を識別する現実的な方法

このビッグデータ社会の中では，善人悪人関係なしに誰もが大量の情報を手に入れることができるので，攻撃者が持つ背景知識を想定するのは難しくなっている．しかし，以下のような背景知識を想定する客観的な手法が存在すると私は考える．(1) 公開されている統計情報に基づいて，攻撃者に与えられる背景知識の属性を想定する．(2) ある基準に基づいて，異なる背景知識を持つ攻撃者をクラス分けする．(3) これまでに発生したサイバーインシデントに基づいて，漏洩した属性の量を調査して，そのデータの資産価値を想定する．これらの手法の中に，今後の研究が必要である．

3.5 まとめ

本章では，履歴データのある属性から部分的な背景知識を得る新たな攻撃者モデルを想定した．また，その攻撃者の平均識別確率を用いてデータのリスク評価を行うモデルを提案した．さらに，平均

表 3.10: 3 つのモデル（平均モデル=厳密解，サンプリングモデル，最小コストモデル）でレコード数が 2 倍の T_1 の各属性を評価した際の計算時間

属性名	平均モデル（厳密解）[s]	サンプリングモデル [s]	最小コストモデル [s]
伝票 ID	38.05	0.28	0.04
購買日	0.45	0.03	0.03
購買時刻	0.77	0.05	0.04
購買商品	3.68	0.05	0.03
価格	5.57	11.02	0.07
数量	0.30	0.05	0.04
合計	54.27	1.14	0.25

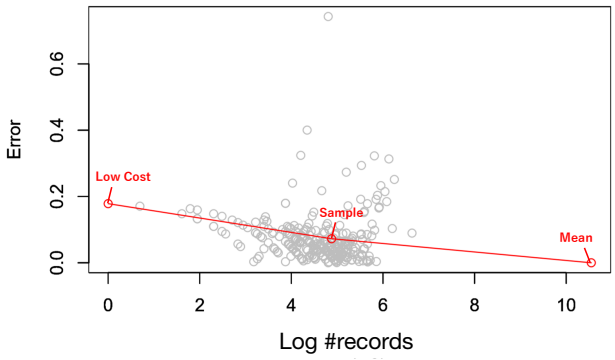


図 3.11: $T = T_1$, $A = \mathbf{date}$ のときの各モデルの計算に必要なレコード数と誤差率の散布図

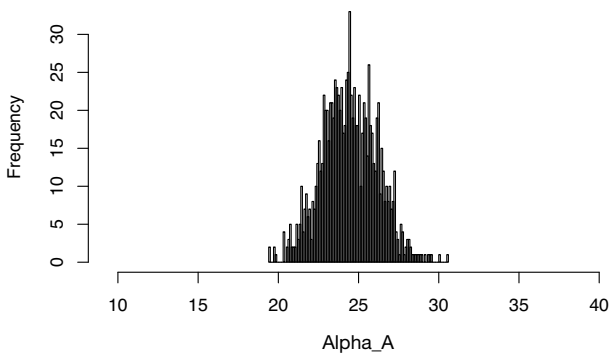


図 3.12: $T = T_1$, $A = \mathbf{date}$, $|D'_A| = 50$ のときの α_A の分布

識別確率を近似する 3 つのモデルを提案し，それらを用いて購買履歴データ，入院記録データ，世帯収入データの 4 つの実際のデータのリスク評価を行い，どの属性が危険であるのかを評価した．匿名加工をする際にこのリスク評価モデルを用いることによって，どの属性を加工・削除するか等の加工指針を立てることに活用できる．

第4章　履歴データの数理モデルの提案とk-匿名化に必要なダミーレコード数推定への応用

4.1　導入

これまで，購買履歴や位置情報のような履歴データから個人が識別されるリスクは，理論的には求められていなかった．その理由には,「履歴から個人は特定できない」という誤解 [111] に加えて，動的に変化するイベントを正確に定式化することの技術的な困難性等が考えられる．

私は，履歴データをダミーレコード追加によって k-匿名化する際の加工コストは，(1) 各グループの大きさ，(2) 各グループが持つ値の種類数，(3) 各個人が持つ値の種類数，の 3 つの値で求められることを発見した．ここでいう「グループ」とは，k-匿名化の際に区別がつかないように加工される個人の集まりのことであり,「個人/グループが持つ値」とは，購買商品などのデータ値のことである．ところが，これらの値は実際にデータを加工するまで得ることができないため，これらの値をそれぞれ期待値や平均値で近似することを考える．しかし，そのためには “全 ℓ 種類のデータ x レコードが与えられたとき，その中のユニークな値の種類数 y はいくらか？” という問題に取り組まなければならない．この問題を，本章では**商品種類数問題**と呼ぶ．商品種類数問題はクーポンコレクター問題 [55] に似ている．

> （クーポンコレクター問題）シリアルの箱に入っているクーポンが全て等しい確率で独立して発生するとき，全てのクーポンを集めるためにはシリアルを何箱買えばよいか？

この問題は，箱数の期待値が $E[X] = \sum_{i=1}^{\ell} \frac{\ell}{\ell-i+1} = \ell \sum_{i=1}^{\ell} \frac{1}{i}$ であり，ハーモニック数 $H(\ell) = \sum_{i=1}^{\ell} \frac{1}{i} = \ln \ell + O(1)$ を用いて，$E[X] = \ell \ln \ell + O(\ell)$ と解くことができる．しかし，商品種類数問題とクーポンコレクター問題の目的は異なるため，残念ながらこの解を商品種類数問題に適用することはできない．クーポンコレクター問題と同じようにシリアルとクーポンを用いると，商品種類数問題は以下のように例えることができる．

> （商品種類数問題）x 個のシリアルの箱を買ったとき，集まるクーポンの種類数 y は，全 ℓ 種類のうちいくらか？

表 4.1 に，商品種類数問題とクーポンコレクター問題の比較を示す．

本章では，商品種類数問題を解決するために，新たな履歴データモデルを提案する．提案モデルでは，レコードの値が一様分布で独立して発生するという仮定の下，次の 2 つの定理を示す．(a)x レコードのデータが与えられた時に，全 ℓ 種から選ばれる一意な種類の数 y の条件付確率 $Pr(y|x)$ の分布．(b)y の期待値 $E[y|x,\ell]$．本モデルを用いることにより，実際に様々な k の値を用いて k-匿名化を行わなくとも，最適な k の値を推定することができる．

表 4.1: 商品種類数問題とクーポンコレクター問題の比較結果

	クーポンコレクター問題	商品種類数問題
仮定	一様分布 $1/\ell$	一様分布 $1/\ell$
目的	ℓ 個のクーポンを集めるために 買う必要があるシリアルの箱の数	全 ℓ 種類のシリアルの箱を x 個買ったとき， 集まるクーポンの種類数 y はいくらか？
期待値	$\ell \ln \ell + O(\ell)$	$(-\ell)(1-\frac{1}{\ell})^x + \ell$

(a) T_{ex}

User IDs	Goods
Alice	Apple
Bob	Apple
Bob	Book
Carol	Book

(b) T'_{ex}

Pseudonym	Goods	
1	Apple	
1	Book	*
2	Apple	
2	Book	
3	Book	
3	Apple	*

(c)

User IDs	$I(u_i)$
Alice	$\{Apple\}$
Bob	$\{Apple, Book\}$
Carol	$\{Book\}$

(d)

Pseudonym	$I(u_i)$
1	$\{Apple, Book\}$
2	$\{Apple, Book\}$
3	$\{Apple, Book\}$

図 4.1: ダミーレコードを追加する手順

4.2 データモデル

4.2.1 基礎定義

本研究では，レコード（行）と属性（列）によって構成される履歴データを考える．記号等を以下のように定義する．

定義 4.2.1 履歴データを T とする．T はレコード数 m，顧客数 n のデータであり，個人を直接識別する顧客 ID 属性と，ℓ 種類の値をとる属性からなるものとする．T の顧客集合を $U = \{u_1, \ldots, u_n\}$ とし，U が属性でとる値の集合を $I(U) = \{g_1, \ldots, g_\ell\}$ とする．また，顧客 u_i が属性でとる値の集合を $I(u_i)$ とする．

例 4.2.1 T の例として，3 人の顧客の購買商品のデータ T_{ex} を図 *4.1* の表 *(a)* に示す．この場合，$U = \{Alice, Bob, Carol\}$ であるので $n = 3$，$m = 4$ であり，$I(U) = \{Apple, Book\}$ であるので $\ell = 2$ である．例えば，$Alice$ の購買商品集合は $I(Alice) = \{Apple\}$ である．図中の表 *(b), (c), (d)* については後述する．

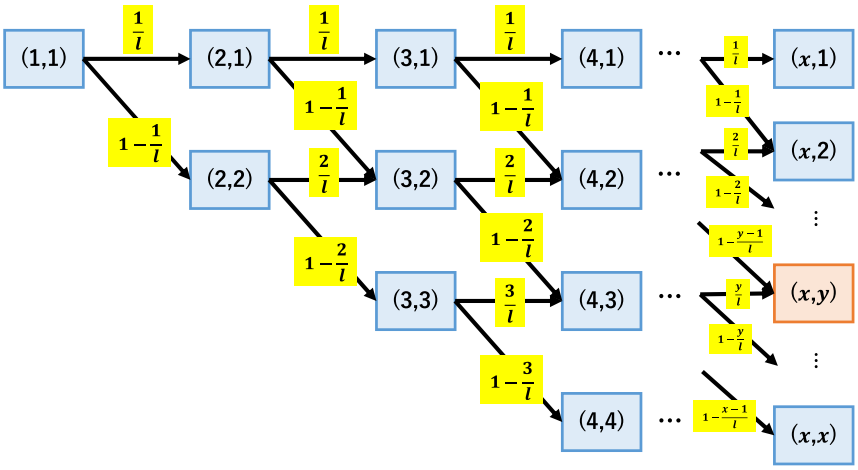


図 4.2: 状態 (x, y) の遷移図

仮定 4.2.1 （$1/\ell$ 仮定）T 中の ℓ 種の値は一様な確率 $1/\ell$ で生起する．

定義 4.2.2 X を $1, \ldots, m$ の値をとるレコード数についての確率変数，Y を $1, \ldots, \ell$ の値をとる種類数の確率変数とする．ℓ 種類の値をとる属性のデータが x レコードあるときに，値が y 種類ある状態を (x, y) とし，その状態になる条件付確率を $Pr(Y = y|X = x)$ とする．また，履歴データが X レコードあるときの属性 *(*全 ℓ 種類*)* の種類数 Y の期待値を $E[y|x, \ell]$ とする．

本研究では，購入という時系列イベントを，一様な確率で生起するレコード群とみなす．つまり，データの状態は新しい購買履歴が生まれるたびに更新される．

(x, y) になるパターンは，$(x-1, y)$ から確率 y/ℓ で値が重複する場合と，$(x-1, y-1)$ から確率 $1-(y-1)/\ell$ で新しい値が生起する場合の 2 パターンあるため，$Pr(y|x)$ は

$$Pr(y|x) = (1 - \frac{y-1}{\ell})Pr(y-1|x-1) + \frac{y}{\ell}Pr(y|x-1) \tag{4.1}$$

という漸化式で表せる．

例 4.2.2 (x, y) の状態遷移図を図 *4.2* に示す．例えば購買履歴データの場合，$(4, 2)$ はデータが *4* レコードあるときに，商品が全 ℓ 種類中 *2* 種類生起している状態である．図より，$Pr(Y = 2|X = 4) = ((1/\ell)(1/\ell)(1-1/\ell)) \cdot ((1/\ell)(1-1/\ell)(2/\ell)) \cdot ((1-1/\ell)(2/\ell)(2/\ell)) = 7(1-1/\ell)/\ell^2$ と計算できる．

4.2.2 履歴データモデル

本節では，$1/\ell$ 仮定のもと，履歴データ中に登場する項目の種類数 Y の確率分布 $Pr(y|x)$ とその期待値 $E[y|x, \ell]$ を与える数理モデルを提案する．

定理 4.2.1 ℓ 種類の値をとる x レコードが与えられた時に，y 種類の値がある状態 (x, y) が生じる確率 $Pr(y|x)$ は

$$Pr(y|x) = \prod_{j=0}^{y-1}(1 - \frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_y=x-y} (\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_y}, \tag{4.2}$$

である．ここで，$m_1, \ldots, m_y$ は総和が $x - y$（$x \geq y \geq 1$）になる正の整数である．

(証明) 式 (4.2) を数学的帰納法で証明する．$x = 1$ のとき，$x \geq y \geq 1$ であるため $y = 1$ であり，$Pr(1|1) = (1 - \frac{0}{\ell}) \cdot (\frac{1}{\ell})^0 \cdots (\frac{y}{\ell})^0 = 1$ となるため式 (4.2) は成り立つ．

$x = x' - 1$ $(x' \geq 2)$ のとき，任意の $y(1 \leq y \leq x' - 1)$ で式 (4.2) が成り立つと仮定する．つまり，$Pr(y|x' - 1) = \prod_{j=0}^{y-1}(1 - \frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_y=(x'-1)-y}(\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_y}$ とする．このとき，$Pr(y - 1|x' - 1) = \prod_{j=0}^{y-2}(1 - \frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_{y-1}=(x'-1)-(y-1)}(\frac{1}{\ell})^{m_1} \cdots (\frac{y-1}{\ell})^{m_{y-1}}$ となる．

これらの式を式 (4.1) に代入すると，

$$\begin{aligned}
Pr(y|x') &= (1 - \frac{y-1}{\ell})Pr(y - 1|x' - 1) + (\frac{y}{\ell})Pr(y|x' - 1) \\
&= \prod_{j=0}^{y-1}(1 - \frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_{y-1}=x'-y} (\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_{y-1}} + \prod_{j=0}^{y-1}(1 - \frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_y=x'-y} (\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_y+1} \\
&= \prod_{j=0}^{y-1}(1 - \frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_y=x'-y} (\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_y}
\end{aligned}$$

となるため，式 (4.2) を得る．$x = x'$ のときも式 (4.2) は成り立つ．よって，任意の $x(x \geq 1)$ について式 (4.2) は成り立つ． (Q.E.D)

この定理を適用することにより，状態遷移図からは求めるのが困難である状態 (x, y) の確率分布を実用的に求めることができるようになる．

定理 4.2.2 $1/\ell$ 仮定の下で，一様に分布する ℓ 種の値をとる x レコードのデータが持つ値の種類数 Y の期待値は $E[y|x, \ell] = (-\ell)(1 - \frac{1}{\ell})^x + \ell$ である．

(証明) $1/\ell$ 仮定の下で，x レコードのデータで全 ℓ 種の値が少なくとも 1 回は生起する確率は $1 - (1 - 1/\ell)^x$ である．期待値の線形性から，ℓ 種の値の数の期待値は $E[y|x, \ell] = \ell(1 - (1 - 1/\ell)^x) = (-\ell)(1 - \frac{1}{\ell})^x + \ell$ となる． (Q.E.D)

4.2.3 提案モデルの分析

本節では，提案したモデルによって与えられる $Pr(y|x)$ と $E[y|x, \ell]$ を検証する．レコード数 $x(= 10, \ldots, 100)$ を与えたときの y の確率分布 $(\ell = 100)$ を図 4.3 に示す．例えば，青いグラフは $Pr(Y = y|X = 50)$ を示しており，$y = 40$ のとき最大値 0.168 をとる．つまり，提案モデルに従う $x = 50$ レコードの履歴データは，全 100 種類中 $y = 40$ 種類の値を持つ確率が最も高いことを意味している．また，種類数 y を与えたときの x の確率分布 $(\ell = 100)$ を図 4.4 に示す．例えば，青いグラフは $Pr(Y = 25|X = x)$ を示しており，$x = 28$ のとき最大値 (0.250) をとる．つまり，提案モデルに従う履歴データが 100 種類中 25 種類の項目を持つ確率は，レコード数が 28 であるとき最も高いことを意

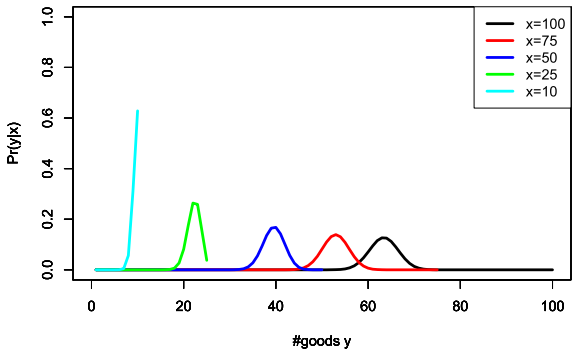


図 4.3: $\ell = 100$ のときの y と $Pr(Y|X)$ の関係 ($x = 10, 25, \ldots, 100$)

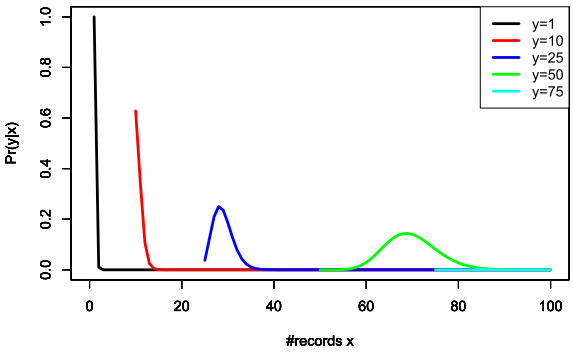


図 4.4: $\ell = 100$ のときの x と $Pr(Y|X)$ の関係 ($y = 1, 10, \ldots, 75$)

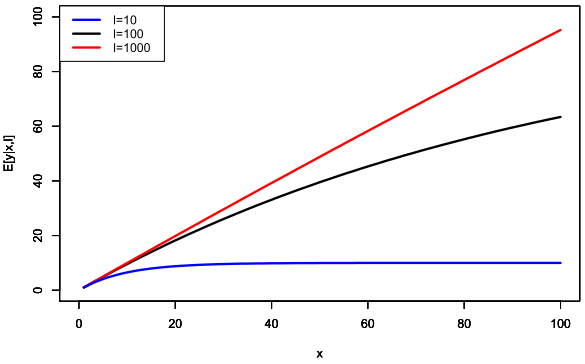


図 4.5: x についての $E[y|x, \ell]$ の分布

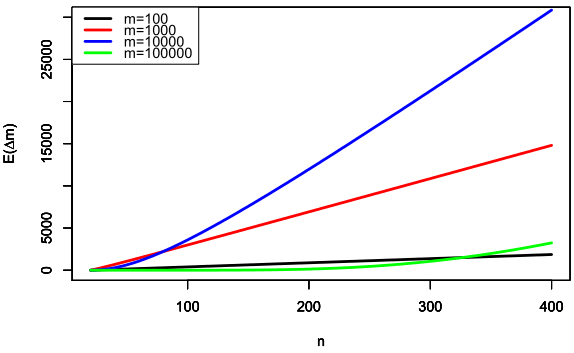


図 4.6: n の変化に伴う $E(\Delta m)$

味している．次に，ℓ を固定したときの $E[y|x, \ell]$ と x の関係を図 4.5 に示す．これは履歴データが x レコードあるときに，そのデータが持つ種類数 y(ℓ 種類中) の期待値の変化を示している．例えば黒いグラフは $E[y|x, 100]$ のグラフであるが，これから $\ell = 100$ のとき，100 レコードの履歴データが持つ項目種類数の期待値が 63.40 であることがわかる．

4.3 k-匿名化に必要な加工コスト

提案した履歴データモデルの応用として，本節では履歴データの k-匿名化に必要なダミーレコード数を見積もる．ダミーレコード数は有用性評価指標として用いられることもあり，匿名加工データにとっては重要な値の一つである．記号等の定義を行い，ダミーレコード数の厳密解について述べ，ダミーレコード数の期待値を提案履歴データモデルを用いて求める．

4.3.1 基礎定義

定義 4.3.1 T を加工したデータを T' とする．T' は仮名化された T に疑似レコードを追加して，複数の仮名が同じ履歴属性の値を持つように加工されたデータである．T' の識別子属性は仮名化されており，また，ダミーレコードの追加によって k-匿名化されている．追加された疑似レコード数を Δm とし，同じ履歴属性の値を持つ顧客のグループ数を c とする．i 番目のグループを $U_i = \{u_1^i, \ldots, u_{s_i}^i\}$ とし，$s_i = |U_i|$ をグループ U_i の大きさ *(*仮名数*)* とする．このとき，$U = U_1 \cup \cdots \cup U_c$ となる．$I(u)$ を顧客 u が持つ値の集合とし，$I(U_i) = \bigcup_{u \in U_i} I(u)$ とする．

例 4.3.1 T' の例として，T_{ex} を加工した T'_{ex} を図 *4.1* の表 *(b)* に示す．T'_{ex} では *Alice,Bob,Carol* はそれぞれ *1,2,3* に仮名化されており，購買商品が等しいグループの数が *1* つになるように，*2* つのダミーレコード（*印がついているもの）が追加されている *($c = 1, \Delta m = 2$)*．この場合のグループは $U_1 = \{1, 2, 3\}$ のみであり，$s_1 = 3$ である．

同図の表 *(c),(d)* に，T_{ex} と T'_{ex} の各顧客/仮名の購買商品リストを示す．この場合，ダミーレコードの追加によって *Alice, Bob, Carol* の *3* 人の購買商品集合が等しくなるように加工されているため，$I(1) = I(2) = I(3) = I(Alice) \cup I(Bob) \cup I(Carol) = \{Apple, Book\}$ となっている．表 *(d)* で示されたように *3* 人の顧客（仮名）の区別がつかないので，T'_{ex} は *3-anonymity* を満たしている．

4.3.2 ダミーレコード数の厳密解

命題 4.3.1 グループ数を c とすると，ダミーレコード数 Δm は

$$\Delta m = \sum_{i=1}^{c} s_i |I(U_i)| - \sum_{i=1}^{n} |I(u_i)|$$

である．

(証明) ダミーレコード数 Δm は $\Delta m = \sum_{i=1}^{c} \sum_{j=1}^{s_i} (|I(U_i)| - |I(u_j^i)|)$ と求められるため，この式を解くと $\Delta m = \sum_{i=1}^{c} \sum_{j=1}^{s_i} (|I(U_i)| - |I(u_j^i)|) = \sum_{i=1}^{c} \left(s_i |I(U_i)| - \sum_{j=1}^{s_i} |I(u_j^i)| \right) = \sum_{i=1}^{c} s_i |I(U_i)| - \sum_{i=1}^{n} |I(u_i)|$ となり，与式が得られる． (Q.E.D)

例 4.3.2 T_{ex} を $c = 1$ で加工する（T'_{ex} を作成する）場合を考える．このときグループは $U_1 = \{1, 2, 3\}$ のみであり，$s_1 = 3$ である．*3* 人の顧客の購買商品集合は $I(Alice) = \{Apple\}$, $I(Bob) = \{Apple, Bob\}$, $I(Carol) = \{Book\}$ であり，グループ全体の購買商品集合は $I(U_1) = \{Apple, Book\}$ である．この場合，必要なダミーレコードの数は $\Delta m = s_1 |I(U_1)| - |I(u_1)| - |I(u_2)| - |I(u_3)| = 3 \cdot 2 - 1 - 2 - 1 = 2$ と計算できる．

Δm は各グループの大きさ s_i，各グループの履歴属性の値の種類数 $|I(U_i)|$，各顧客の履歴属性の値の種類数 $|I(u_i)|$ から求めることができるが，s_i, $|I(U_i)|$,$|I(u_i)|$ の値はパラメータ c を決めて実際にデータを加工するまで手に入らないため，Δm も加工後まで求めることはできない．しかし最適な c を決めるためには，様々な c で加工コストとして Δm を求める必要がある．

4.3.3 ダミーレコード数の期待値

データを加工する前に Δm のおおまかな値を知ることができれば，c を決める際の指標になる．s_i, $|I(U_i)|$, $|I(u_i)|$ の値を，提案した履歴データモデルを用いることによって加工前に手に入る値に置き換えることにより，Δm の期待値を計算する．本章では以下を仮定する．

仮定 4.3.1 *n 人の顧客の c 個のグループの大きさは全て等しく n/c である (n/c 仮定).*

仮定 4.3.2 *n 人の顧客が持つ計 m 個のレコード数は全て等しく m/n である (m/n 仮定).*

上記の 2 つの仮定より，c 個のグループ全てが m/c 個のレコードを持つことが示される．これら 3 つの仮定 $(1/\ell, n/c, m/n)$ の下では，Δm の期待値は以下のように求められる．

定理 4.3.1 *3つの仮定 $(1/\ell, n/c, m/n)$* より，顧客数 n，レコード数 m であり，全 ℓ 種類の履歴属性を持つ履歴データを，c 個のグループに加工するために必要な疑似レコード数の期待値 $E(\Delta m)$ は，$E(\Delta m) = n\ell\left((1-\frac{1}{\ell})^{m/n} - (1-\frac{1}{\ell})^{m/c}\right)$ である．

(証明) Δm は $\Delta m = \sum_{i=1}^{c} s_i|I(U_i)| - \sum_{i=1}^{n}|I(u_i)|$ と計算できる．この式の s_i, $|I(U_i)|$, $|I(u_i)|$ の値を，それぞれ n/c, $E[y|m/c,\ell]$, $E[y|m/n,\ell]$ で置き換えることにより，以下の式を得られる．

$$
\begin{aligned}
\Delta m &= \sum_{i=1}^{c} s_i|I(U_i)| - \sum_{i=1}^{n}|I(u_i)| \\
E(\Delta m) &= \sum_{i=1}^{c}\frac{n}{c}E[y|\frac{m}{c},\ell] - \sum_{i=1}^{n}E[y|\frac{m}{n},\ell] \\
&= n\{(-\ell)(1-\frac{1}{\ell})^{m/c}+\ell\} - n\{(-\ell)(1-\frac{1}{\ell})^{m/n}+\ell\} \\
&= n\ell\{(1-\frac{1}{\ell})^{m/n} - (1-\frac{1}{\ell})^{m/c}\}
\end{aligned}
$$

よって，与式が証明される． (Q.E.D)

この定理と k-匿名性との関係を考える．ダミーレコードを追加されて c 個のグループに分割された履歴データが k-匿名性を満たしている場合，各グループの顧客数は少なくとも k である．そして 3 つの仮定より，履歴データにおいては，全顧客のレコード数が等しく (m/n)，全グループのレコード数も等しい (m/c) ため，全グループの顧客数が等しい (n/c) ことがわかる．つまり，$k < n/c$ であることが言えるため，その時のダミーレコード数は以下のように求められる．

系 4.3.1 顧客数 n，レコード数 m であり，全 ℓ 種類の履歴属性を持つ履歴データを k-匿名化するために必要な疑似レコード数の期待値 $E'(\Delta m)$ は，$E'(\Delta m) \geq n\ell\left((1-\frac{1}{\ell})^{m/n} - (1-\frac{1}{\ell})^{km/n}\right)$ である．

4.3.4 評価実験

本節では，疑似レコード数の期待値 $E(\Delta m)$ の値がどのように変化するのかを実データを用いて分析する．

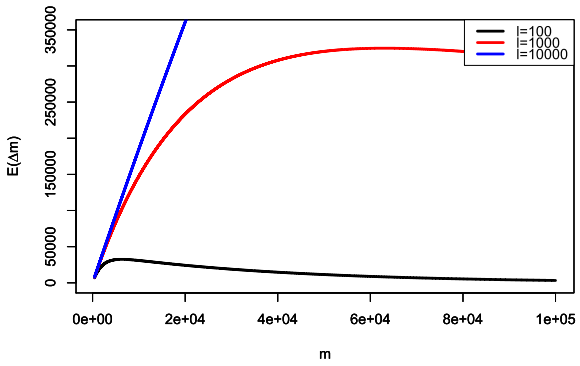


図 4.7: m の変化に伴う $E(\Delta m)$ の変化

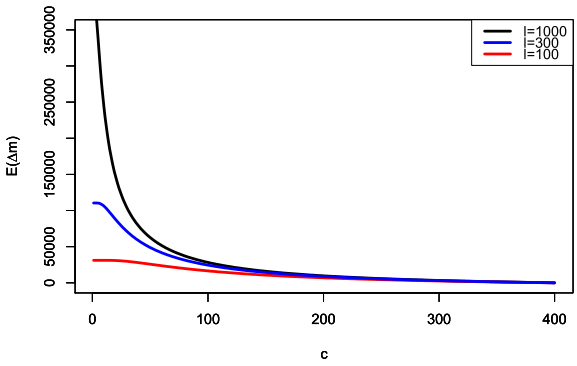


図 4.8: グループ数 c についての $E(\Delta m)$ の変化

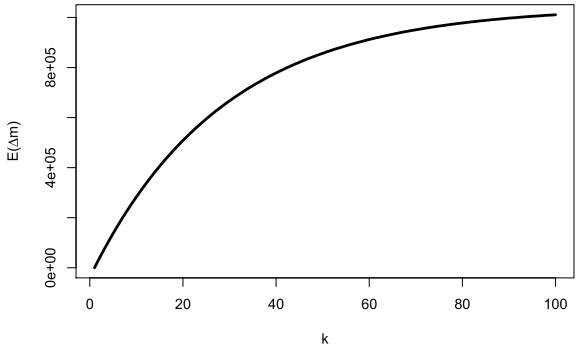


図 4.9: k についての $E'(\Delta m)$ の分布

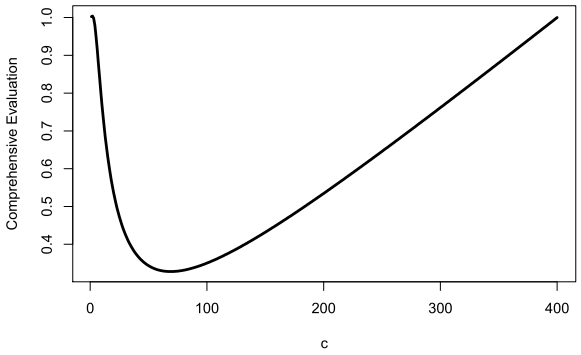


図 4.10: 総合評価値と c の関係

図 4.6 に，$\ell = 100, c = 20$ のときのレコード数 m と個人数 n の変化に伴う $E(\Delta m)$ を示す．例えば青いグラフは 10,000 レコードの履歴データ ($\ell = 100, c = 20$) についてのグラフであり，このデータの顧客数 n が 400 であるとき，追加するダミーレコード数の期待値 $E(\Delta m)$ が 30,850 であり，顧客数が増えるほど必要な疑似レコード数も増えることがわかる．

また，$n = 400, c = 20$, 全種類数 ℓ のときのレコード数 m の変化に伴う $E(\Delta m)$ を図 4.7 に示す．例えば赤いグラフは $\ell = 1,000, n = 400, c = 20$ のときのグラフであるが，$m = 63,037$ のときに極大値 (324,570 レコード) を取り，そこからは $E(\Delta m)$ は減少している．このことから，履歴データのレコード数が膨大になると，必要な疑似レコード数は減少することがわかる．

次に，グループ数 c と $E(\Delta m)$ の関係を図 4.8 に示す．この図より，グループ数が増えるほど必要な疑似レコード数は減少することがわかる．また，k-匿名化に必要な疑似レコード数の期待値 $E'(\Delta m)$ の最小値と k の関係を図 4.9 に示す．このとき，$n = 400, m = 38,000, \ell = 2,700$ である．k の値を大きくするほど $E'(\Delta m)$ の最小値も大きくなり，加工データの有用性が下がる．

表 4.2: Δm の結果と本研究で求められた $E(\Delta m)$ の比較

	Δm	$E(\Delta m)$	$E(\Delta m_2)$	$E(\Delta m_3)$	$E(\Delta m_4)$	Jaccard *Reid*	Random *Reid*
$k=2$	183,902	36,188	31,868	36,213	39,105	0.1729	0.1223
$k=3$	175,449	71,158	60,968	65,312	74,075	0.1726	0.1222
$k=4$	162,474	104,950	87,768	92,113	107,868	0.1723	0.1218
$k=8$	125,798	229,122	177,815	182,159	232,039	0.1681	0.1218

4.3.5 推定コストと実際のコストの比較

5 章 (研究 [56]) では，顧客 400 人，38,087 レコードの購買履歴データを 4-匿名化するために，約 160,000 のダミーレコードが必要になることを明らかにしている．表 4.2 に Δm の真値と推定コスト $E(\Delta m)$ の比較を示す．これらの結果が大きく異なる理由は，実際の Δm は全ての k でグループ数 c が 50 に固定されているため，3 つの仮定を満たしていないせいである．

5 章では，有用性と安全性の評価を総合的に行うために，ダミーレコード数と再識別率の値を用いた $\alpha E(\Delta m)+E(\mathit{Reid})$ という式を用いる．これは，PWS Cup 2016 で用いられた (Utility + Security)/2 という総合評価手法をもとにしたものであり，α は $E(\Delta m)$ を $0 \le \alpha E(\Delta m) \le 1$ の範囲に標準化するための係数である．図 4.10 に $n=400$, $m=38,000$, $\ell=2,700$, $\alpha=1/1,042,653$ のときの総合評価値と c の関係を示す．この場合，$c=69$ のとき総合評価値が最も小さくなっているので，本手法では $c=69$ で加工をするのが最適であり，$n/c=400/69=5.80$ なので k の最適な値は 5 であるといえる．

4.3.6 仮定の影響

本節では，本研究で置いた 2 つの仮定（$1/\ell$ 仮定，m/n 仮定）が推定結果に及ぼす影響を調査する．加工するデータがこれらの仮定を常に満たしているわけではない．図 4.11 に，Online Retail Data Set [57]（$n=400, \ell=2,781$）で生起している商品の頻度分布を示す.[1]最も頻度が高い商品は 1,000 回以上生起している一方で，多くの商品が 1 回のみしか生起しておらず，分布に非常に偏りがあることがわかる．図 4.12 にこのデータの顧客のレコード数分布を示す．図より，この分布にも似たような偏りがあることがわかる．このように残念ながら，Online Retail Data Set では $1/\ell$, m/n 仮定は満たされていなかった．

そこで本節では，これら 2 つの仮定 $(1/\ell,\ m/n)$ を提案モデルから取り除いたときの推定コストを以下のように調査する．

定義 4.3.2 p_j を ℓ 種類中 j 番目の値が生起する確率とする．b_i を n 人中 i 番目の顧客のレコード数とする．

$1/\ell$ 仮定を満たすとき，p_j は任意の j で $1/\ell$ となり，m/n 仮定を満たすとき，b_i は任意の i で m/n になる．

[1]ここで分析するデータは Online Retail Data Set そのものではなく，PWS Cup 用に作成された部分的なものであることに注意せよ．

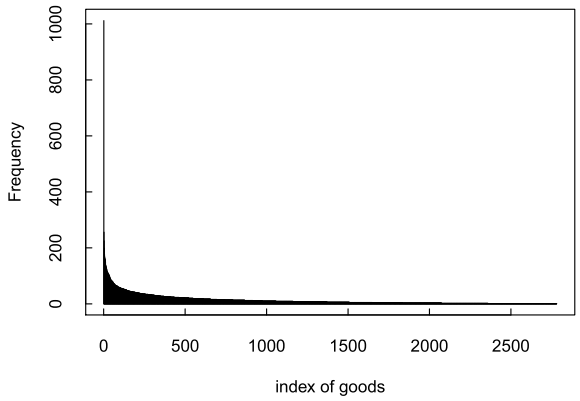


図 4.11: Online Retail Data Set で生起している商品の頻度分布

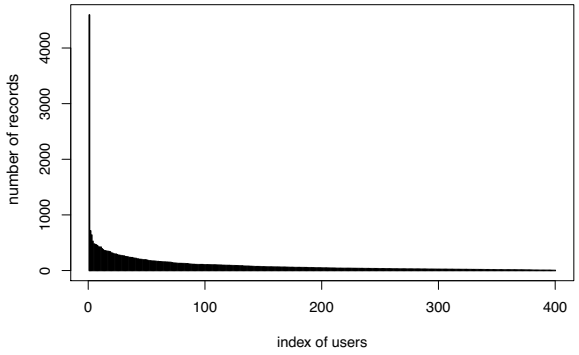


図 4.12: Online Retail Data Set の顧客のレコード数分布

系 4.3.2 n/c 仮定と m/n 仮定のもとで，データの加工に必要なダミーレコード数の期待値 $E(\Delta m_2)$ は $E(\Delta m_2) = n\sum_{j=1}^{\ell}\{(1-p_j)^{m/n} - (1-p_j)^{m/c}\}$ と計算できる．ここで，n は顧客数，m はレコード数，ℓ は値の種類数である．

(証明) Δm は $\Delta m = \sum_{i=1}^{c} s_i|I(U_i)| - \sum_{i=1}^{n}|I(u_i)|$ である．この式の s_i,$|I(U_i)|$,$|I(u_i)|$ の値をそれぞれ n/c,$E[y|m/c,\ell]$,$E[y|m/n,\ell]$ で置き換えると，与式を以下のように得ることができる．

$$\begin{aligned}
E(\Delta m_2) &= \sum_{i=1}^{c}\frac{n}{c}E[y|\frac{m}{c},\ell] - \sum_{i=1}^{n}E[y|\frac{m}{n},\ell] \\
&= n\sum_{j=1}^{\ell}\{1-(1-p_j)^{m/c}\} - \sum_{i=1}^{n}\sum_{j=1}^{\ell}\{1-(1-p_j)^{m/n}\} \\
&= n\sum_{j=1}^{\ell}\{1-(1-p_j)^{m/c}\} - n\sum_{j=1}^{\ell}\{1-(1-p_j)^{m/n}\} \\
&= n\sum_{j=1}^{\ell}\{(1-p_j)^{m/n} - (1-p_j)^{m/c}\}
\end{aligned}$$

(Q.E.D)

系 4.3.3 n/c 仮定の下で，データの加工に必要なダミーレコード数 Δm_3 の期待値 $E(\Delta m_3)$ は $E(\Delta m_3) = n\sum_{j=1}^{\ell}\{1-(1-p_j)^{m/c}\} - \sum_{i=1}^{n}\sum_{j=1}^{\ell}\{1-(1-p_j)^{b_i}\}$ と計算できる．ここで，n は顧客数，m はレコード数，ℓ は値の種類数である．

(証明) Δm は $\Delta m = \sum_{i=1}^{c} s_i|I(U_i)| - \sum_{i=1}^{n}|I(u_i)|$ である．この式の s_i,$|I(U_i)|$,$|I(u_i)|$ の値をそれぞれ

n/c, $E[y|m/c,\ell]$, $E[y|b_i,\ell]$ で置き換えると，与式を以下のように得ることができる．

$$\begin{aligned} E(\Delta m_3) &= \sum_{i=1}^{c} \frac{n}{c} E[y|\frac{m}{c},\ell] - \sum_{i=1}^{n} E[y|b_i,\ell] \\ &= nE[y|\frac{m}{c},\ell] - \sum_{i=1}^{n}\sum_{j=1}^{\ell}\{1-(1-p_j)^{b_i}\} \\ &= n\sum_{j=1}^{\ell}\{1-(1-p_j)^{m/c}\} - \sum_{i=1}^{n}\sum_{j=1}^{\ell}\{1-(1-p_j)^{b_i}\} \end{aligned}$$

(Q.E.D)

系 4.3.4 n/c 仮定と $1/l$ 仮定のもとで，データの加工に必要なダミーレコード数の期待値 $E(\Delta m_4)$ は $E(\Delta m_4) = \ell\sum_{i=1}^{n}\{(1-1/\ell)^{b_i} - (1-1/\ell)^{m/c}\}$ と計算できる．ここで，n は顧客数，m はレコード数，ℓ は値の種類数である．

(証明) 系 4.3.3 の $E(\Delta m_3)$ の式の p_j に $1/\ell$ を代入すると，与式を以下のように得ることができる．

$$\begin{aligned} E(\Delta m_4) &= n\sum_{j=1}^{\ell}\{1-(1-1/\ell)^{m/c}\} - \sum_{i=1}^{n}\sum_{j=1}^{\ell}\{1-(1-1/\ell)^{b_i}\} \\ &= \ell\sum_{i=1}^{n}\{(1-1/\ell)^{b_i} - (1-1/\ell)^{m/c}\} \end{aligned}$$

(Q.E.D)

Online Retail Dataset から全ての p_j（2,781 種類）と b_i（400 種類）を求めることにより，$E(\Delta m_2)$, $E(\Delta m_3)$, $E(\Delta m_4)$ を求めた．図 4.13 に 4 モデル ($E(\Delta m)$, $E(\Delta m_2)$, $E(\Delta m_3)$, $E(\Delta m_4)$) による，k-匿名化に必要なダミーレコード数の推定結果の比較を示す．図中の緑線と黒線はそれぞれ $E(\Delta m_4)$ と $E(\Delta m)$ を示しており，これらはほとんど重なっている．$E(\Delta m_4)$ は $E(\Delta m)$ から m/n 仮定のみを取り除いたものであるため，m/n 仮定の推定値への影響は殆ど無いといえる．また，赤線と青線はそれぞれ $E(\Delta m_2)$ と $E(\Delta m_3)$ を示しており，これら 2 本もほとんど重なっていて，任意の $k(1 \le k \le 200)$ で黒線 ($E(\Delta m)$) よりも下に位置している．$E(\Delta m_2)$ と $E(\Delta m_3)$ はどちらも $1/\ell$ 仮定を取り除いたものであるため，$1/\ell$ 仮定は推定値を増やしていることがわかる．表 4.2 には $k = 2,3,4,8$ のときの $E(\Delta m)$, $E(\Delta m_2), E(\Delta m_3), E(\Delta m_4)$ の推定結果を示している．

$E(\Delta m_2), E(\Delta m_3), E(\Delta m_4)$ を計算するためには，全ての p_j と b_i の値を求める必要があり，これらの値はデータを分析しないと得られない．よって，加工のパラメータ (k,c) の最適値を知りたいときは，これら 3 モデルよりも $E(\Delta m)$ の方が役に立つ．本節ではコストの推定をするために，m 以外のパラメータは検討していないことに注意せよ．

表 4.2 に示したように，推定コスト ($E(\Delta m)$, $E(\Delta m_2), E(\Delta m_3), E(\Delta m_4)$) の振る舞いと実際のコスト ($\Delta m$) の振る舞いは大きく異なり，$k$ の値が大きくなるほど，推定コストは増加して実際のコストは減少している．本実験結果より，推定結果と実際の結果の違いは置いている仮定によるものではないことが判明した．私は，これらの違いの原因は，比較対象（5 章）では加工コストを最小化するためにパラメータを調整しているためだと推測する．

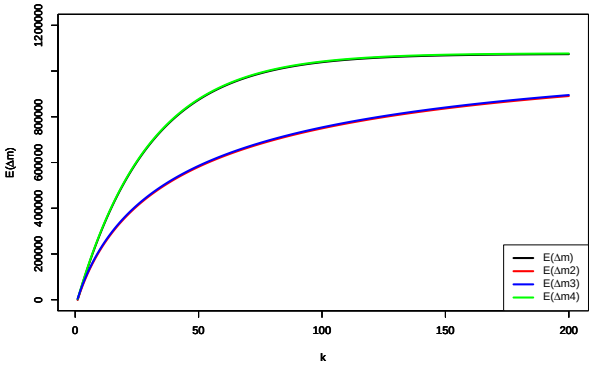


図 4.13: 4 モデルによるダミーレコード数の推定結果

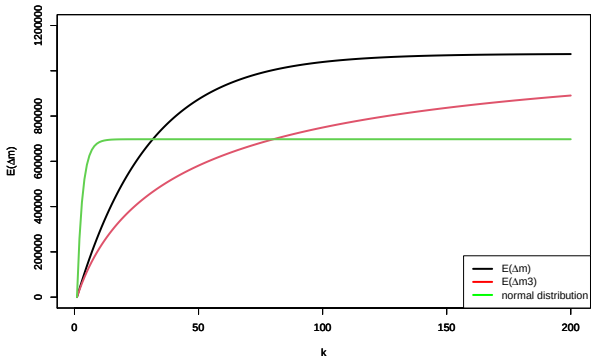


図 4.14: p_j を正規分布 $(\mu = m/\ell, \sigma = 2)$ で与えたときの推定値

また，p_j の値をパラメトリックな確率分布で与える場合を考える．図 4.14 に，$b_i = m/n$ と，正規分布 $(\mu = m/\ell, \sigma = 2)$ に従う p_j から求めた Δm の推定値を示す．図中の黒線は $E(\Delta m)$，赤線は $E(\Delta m_3)$ を示し，緑線は正規分布から求めた p_j に基づく推定値を示している．パラメトリックな確率分布から p_j を得た場合でも，推定値は似たような振る舞いをしており，k が増えるほど増加している．なお，b_i の値をパラメトリックな確率分布で与える実験も行ったが，いずれの結果も 3 つの仮定を置いたときのダミーレコード数の推定値 $(E(\Delta m))$ から大きな変化は見られなかった．

4.3.7　仮定と履歴の関係

本節では，本研究で用いた 3 つの仮定（$1/\ell$ 仮定，m/n 仮定，n/c 仮定）の効果を考える．履歴データ内の ℓ 種類の値は本来確率 $p_j (j = 1, \ldots, \ell)$ で生起しており，これはデータの更新等によって動的に変化する値であるため，そのふるまいをモデル化するのは難しい．仮定の 1 つである $1/\ell$ 仮定では，動的な値 p_j を全て平均値である $1/\ell$ に置き換えているため，本来動的なイベントである履歴データのふるまい（x レコードのデータが y 種類の値を持つ条件付き確率，x レコードのデータが持つ種類数 y の期待値）を，クーポンコレクター問題のようにパラメータ（$Pr(y|x), E[y|x, \ell]$）で表せるようになる．これにより，動的な履歴データを静的なデータのように扱えるようになっている．m/n 仮定も同様に，本来動的な値である各個人のレコード数 b_i を全て平均値 m/n に置き換えることにより，加工コスト（ダミーレコード数）を静的に推定できるようになる．最後に n/c 仮定では，n 人の個人がサイズが等しい c 個のグループに分割されることを仮定している．グループ数 c は加工のパラメータであるため，加工者が決めることのできる静的な値であり，データを n/c 仮定に従うように加工すれば加工コスト推定の精度が高くなる．n/c 仮定に従うように加工されたデータでは，全ての個人の識別リスクは等しく c/n となる．

4.4 まとめ

本章では，履歴データ中のある属性の値が全て等確率で生起する仮定のもと，履歴データ中に生起する項目の種類数の確率分布とその期待値を与える数理モデルを提案した．提案モデルによって,「x レコードの履歴データが持つ値 (全 ℓ 種類) の種類数 y の期待値」を求めることができる．また，提案モデルを応用することにより，履歴データを k-匿名化するために必要な疑似レコード数の期待値を，元データの統計量やパラメータ k から求めた．提案モデルを応用することにより，疑似レコード数以外のデータ評価値も加工前に求めることができることが期待できる．

第5章　商品の特徴による再識別リスクとクラスタリングを用いた購買履歴データ匿名加工手法の提案

5.1　導入

本章では，購買履歴データに対する匿名化の影響を実験的に評価する．個人の購買商品特徴と Jaccard 距離による再識別リスクを想定し，この手法の耐性を上げるためにダミーレコード追加による k-匿名化を行う．ダミーレコードを考え無しに追加してしまうと大量のダミーレコードが必要になってしまい，データの有用性を大きく損なう．従って，有用性と安全性のバランスをとるために，データセット内の似たような購買履歴を持つ顧客を慎重にクラスタリングし，少量のダミーレコードの追加で済むような手法を考える必要がある．

しかし，k-means のような既存のクラスタリング手法では，図 5.1 のイメージ図に示すような，以下の問題点が生じてしまう．

(1)「**単一の巨大なクラスタの独占問題**」顧客のほとんどを占める巨大なクラスタが発生してしまう．例えば PWS Cup 2016 では，38,087 レコードと 2,781 商品を含む購買履歴データが加工対象であったが，似たような購買履歴を持つ顧客が多かったため，巨大なクラスタが生まれてしまった．図 5.1 のオレンジ色のクラスタのような巨大なクラスタの顧客の区別をつかなくするためには，大量のダミーレコードが必要となる．

(2)「**大量の極小クラスタ問題**」1 人の顧客しかいない小さいクラスタが大量に発生してしまう．図 5.1 の青色のクラスタのように，1 人の顧客しかいないクラスタは攻撃者によって簡単に再識別されてしまう．

トランザクションデータでは，多くのレコードが似たような特徴を持っている．つまり，データ中の多様性が無いため，ダミーレコードのせいでデータが歪んでしまったり有用性が失われてしまう．そのため，クラスタサイズのバランスがとられるように改良されたクラスタリング手法を開発する必要がある．

そこで，前述した問題を解決するために以下の方法を試みる．(1) 購買商品に対して Term (*good*) Frequency–Inverse Document (*individual*) Frequency (TF-IDF) に基づいた重みをつけて顧客間の距離を測り，クラスタリングを試みる．TF-IDF の重みによって，珍しい商品には一般的なものより大きな重みが課されるため，小さいクラスタが発生しにくくなる．(2) クラスタサイズの最小値に制限をかける新たなクラスタリング手法を提案する．個人が識別できないように閾値を満たすまで，各クラスタは大きくなり続ける．この匿名化手法の有用性を，私はダミーレコード数を用いて評価する．

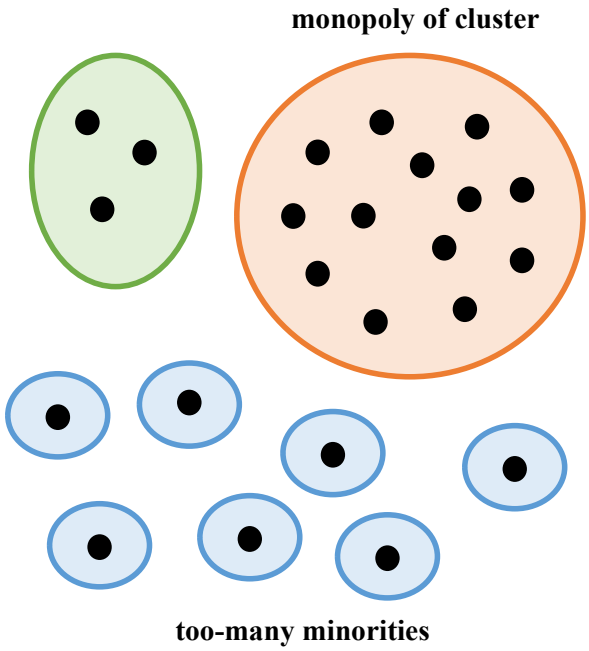


図 5.1: 顧客とクラスタのイメージ図

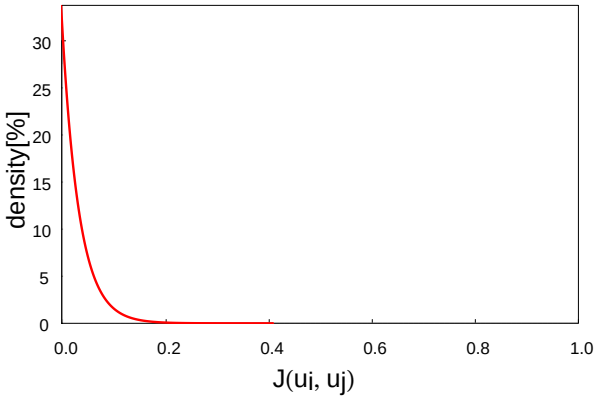


図 5.2: PWSCUP 2016 で用いられたデータの Jaccard 係数の分布

5.2 購買商品特徴と再識別リスク

5.2.1 購買履歴データ

本章では，Online Retail Data Set [57] を用いて研究を行う．このデータは，UCI Machine Leaning Repository [61] より公開されている，イギリスのオンラインショップの 1 年間の購買履歴データであり，PWS Cup 2016–2018 で加工対象のデータとして用いられている．本章では，データセットの記号等を以下のように定義する．

定義 5.2.1 n, m, ℓ をそれぞれ，データセットに含まれる顧客数，データセットのレコード数，データセットに含まれる商品の種類数と定義する．$U = \{u_1, \ldots, u_n\}$ をデータセットに含まれる n 人の顧客の集合とし，$U' = \{u'_1, \ldots, u'_n\}$ を加工されたデータに含まれる顧客の集合とする．$I(U) = \{g_1, \ldots, g_\ell\}$ をデータセットに含まれる商品の集合とする．$I(U)$ の部分集合 $I(u_i)$ を，顧客 u_i によって購入された商品の集合とする．b を *1* 人の顧客が *1* 年間に購入する商品種類数の平均値とする．$J(u_i, u_j)$ を顧客 u_i と u_j の間の *Jaccard* 係数（$J(u_i, u_j) = |I(u_i) \cap I(u_j)|/|I(u_i) \cup I(u_j)|$）とする．

本章で用いる，PWS Cup 2016 向けにカスタムされた Online Retail Dataset の概要，サンプルデータ，基本統計量を，それぞれ表 5.1, 5.2, 5.3 に示す．Online Retail Dataset は 7 属性 38,087 レコード ($m = 38,087$) のデータであり，400 人の顧客 ($n = 400$) と 2,781 種類の商品 ($\ell = 2,781$) を含んでおり，顧客の平均購買種類数は 65 種類 ($b = 65$) であった．また，図 5.2 に，2 顧客間の Jaccard 係数の分布を示す．2 顧客間の Jaccard 係数の平均値は 0.03 であり，最大値は 0.41 であった．これは，データ中で購買商品が最も似ている顧客ペアでも高々41%しか類似していないことを意味しており，顧客

表 5.1: PWSCUP 2016 で用いられたデータの概要

Attribute	Detail
User ID	顧客の ID (5 桁の番号)
Receipt ID	伝票の ID (6 桁の番号)
Date	購買年月日 (yyyy/mm/dd)
Time	購買時間 (hh:mm)
Goods	購買された商品の ID (文字列)
Price	購買された商品の価格 (ポンド)
Number	購買された商品の数量 (自然数)

表 5.2: PWSCUP 2016 で用いられたデータの例

User ID	Receipt ID	Date	Time	Goods	Price	Number
12583	536370	2010/12/1	8:45	22728	3.75	24
12583	536370	2010/12/1	8:45	22727	3.75	24
12431	536389	2010/12/1	10:03	22941	8.5	6
12431	536389	2010/12/1	10:03	21622	4.95	8
12431	536389	2010/12/1	10:03	21791	1.25	12
12838	536415	2010/12/1	11:57	22952	0.55	10
12567	537065	2010/12/5	11:57	22837	4.65	8
12567	537065	2010/12/5	11:57	22846	16.95	1
12748	537429	2010/12/6	15:54	84970S	0.85	12
12748	537429	2010/12/6	15:54	22549	1.45	8

の購買商品集合はかなり多様であるといえる．データのうち特に重要なのは，User ID, Receipt ID, Goods 属性である．

5.2.2 Jaccard 係数を用いた再識別リスク

匿名化されたデータを再識別する攻撃者にとって，顧客間の Jaccard 係数は大きな手掛かりとなる．なぜならば，識別ターゲット顧客の商品集合を偶然得た攻撃者がいた場合，全ての候補者との間の Jaccard 係数を求めることによって，ターゲット顧客のレコードを簡単に識別できることが予測されるからである．

攻撃者によって顧客が識別されるリスクを下げるために，データを加工して攻撃者が個人のデータを特定できないようにする必要がある．例えば，PWSCup2016 参加者はノイズ付加，レコード削除，ダミーレコード追加などによって匿名化をしていた．この加工について，記号等を以下のように定義する．

定義 5.2.2 Δm を，匿名化によって変化したデータのレコード数と定義する．加工されたデータのレコード数を m' とすると，$m' = m + \Delta m$ と表せる．

表 5.3: PWSCUP 2016 で用いられたデータの基本統計量

	記号	値
顧客の数	n	400
レコード（トランザクション）の数	m	38,087
伝票の数		1,763
商品の種類数	ℓ	2,781
商品の値段 (£)		0.04 – 4161
商品の種類数		1 – 74215
購買年月日		2010/12/1 – 2011/12/9
平均購買商品種類数	b	65
Jaccard 係数の平均値	μ	0.03

購買履歴データは，長期間にわたって 1 人の顧客が複数のレコードを持つ履歴データである．履歴データは，1 顧客が 1 レコードしか持たない静的なデータに比べて，長期間のデータが蓄積している分，個人が再識別されるリスクがかなり高まることが予測される．例えば，顧客の 1 年分の購買商品集合を使えば個人を簡単に再識別でき，3 章で述べた私の研究 [70] では，ターゲット顧客が購入した商品を一つでも背景知識として持っていれば，10%ほどの確率で個人を再識別できることが明らかになっている．

攻撃者の振る舞いをモデル化するために，顧客の購買商品集合の特徴を用いた再識別手法を提案し，詳細をアルゴリズム 1 に示す．この手法では最悪のケースとして，加工前のトランザクションデータを全て見ることができる非常に強力な攻撃者を想定する．この攻撃者は加工データを与えられたとき，Jaccard 係数を用いてターゲット顧客と最も近い個人を加工データから探し出し，その個人をターゲット顧客と同一人物であると識別をする．アルゴリズム 1 の計算量は $O(n^2)$ である．攻撃者についての定義を以下のように行う．

定義 5.2.3 攻撃者は，顧客 u_i の購買商品集合 $I(u_i)$ を背景知識として有しており，購買商品集合の特徴を用いた再識別手法（アルゴリズム *1*）を用いて個人の再識別を試みる．

3 章で提案した「データ中のある属性の 1 つの値のみを背景知識として持つ攻撃者」とは異なり，本章で想定する攻撃者が背景知識として有しているのは購買商品の集合であるため，購買した商品の種類は分かるが個数までは分からないということに注意せよ．購買商品の個数もまた，個人を再識別するために非常に有益な情報である．

5.2.3 PWS Cup 2016 における再識別リスク

提案した再識別手法（アルゴリズム 1）の評価を，PWS Cup 2016 に提出された加工データ $D_1, \ldots, D_{10}$ を用いて評価する．$D_1, \ldots, D_{10}$ は，大会の上位 10 チームによって匿名化されたデータであり，そのうち D_7 は私が作成したデータである．自分が匿名化したデータは識別の答えがわかっており，これ

Algorithm 1 Jaccard 係数を用いた再識別手法

Input: M, T, M', T'

Step 1.

M, T を加工前データ，M', T' を加工後データとし，$I(u_i), I(u'_i)$ $(i = 1, \ldots, n)$ を T の顧客 u_i と T' の顧客 u'_i が購買した商品の集合とする．

Step 2.

$i^*_j = \underset{i \in \{1,\ldots,n\}}{\arg\max} J(I(u'_j), I(u_i))$ $(j = 1, \ldots, n')$ を，顧客 u_i に最も近い T' 中の顧客インデックスとする．

Output: $Q = (i^*_1, i^*_2, \ldots, i^*_n)$

表 5.4: 購買商品の特徴を用いた再識別の結果 (PWS Cup 2016)

Data	最大再識別率 (a)	提案再識別手法 (b)
D_1	0.2225	*0.2225
D_2	0.2375	*0.2375
D_3	0.2550	*0.2550
D_4	0.2750	*0.2750
D_5	0.3025	*0.3025
D_6	0.3175	*0.3175
D_8	0.3725	0.2750
D_9	0.3850	*0.3850
D_{10}	0.5500	*0.5500

を再識別して評価結果としてしまうと公平でないため，本章では D_7 を除いた 9 個のデータを評価に用いる．表 5.4 にこれら 9 個のデータの評価結果を示す．(a) 列には，各データに対する大会参加者による再識別結果のうち，もっとも有効であった再識別率を示しており，(b) 列には提案再識別手法による再識別率を示している．(b) 列で*印がついた赤字の箇所は，提案再識別手法が他の参加者の再識別結果よりも優秀であったことを意味している．最も安全な加工データである D_1 でさえ，本提案再識別手法によって 22.25%の顧客が再識別されている．再識別率について，以下のように定義する．

定義 5.2.4 *Reid* を，データ中の個人数に対する再識別された個人の割合とする．

5.3 提案匿名化手法

5.3.1 Jaccard 係数を用いた再識別リスクへの対策

前述した Jaccard 係数を用いた再識別への耐性を高める匿名化手法を考える．私はこの問題を解決するために，複数顧客間の購買商品の統一のための以下の 3 つの手法を考える．(1) 既存レコードを変更する手法 $(m' = m)$．(2) 既存レコードを削除する手法 $(m' < m)$．(3) 疑似レコードを追加する手法 $(m' > m)$．手法 1,2 は元のデータ値を変更/削除する手法であるため，データの正確性を失ってしまうが，手法 3 は元のデータが失われることはない．表 5.5 にこれら 3 手法の長所と短所を示す．

表 5.5: 3 つの匿名化手法の長所と短所

手法	長所	短所
変更	レコードごとに独立に加工できる	k-匿名化が困難
削除	嘘の情報が含まれない	データ量が減ってしまう
追加	元の情報が全て残る	嘘の情報が含まれてしまう

(a) T

U	Receipt ID	Goods
u_1	100	g_1
u_1	300	g_2
u_2	500	g_1
u_2	500	g_3
u_2	600	g_5
u_3	...	...

(d) T' ＊: Dummy Record

U	Receipt ID	Goods	
u'_1	100	g_1	
u'_1	300	g_2	
u'_1	300	g_3	＊
u'_1	300	g_4	＊
u'_1	300	g_5	＊
u'_2	...	...	

(b)

	U	$I(U)$
1	u_1	$\{g_1, g_2\}$
2	u_2	$\{g_1, g_3, g_5\}$
3	u_3	$\{g_4, g_5\}$

(c)

	U'	$I(U')$
1	u'_1	$\{g_1, g_2, g_3, g_4, g_5\}$
2	u'_2	$\{g_1, g_2, g_3, g_4, g_5\}$
3	u'_3	$\{g_1, g_2, g_3, g_4, g_5\}$

図 5.3: トランザクションデータへのダミーレコード追加手法

本研究では上記 3 手法のうち，元の情報が全て残るダミーレコード追加手法に注目する．私の提案手法の説明のために，図 5.3 に加工例を示す．図中のパネル (a) は 3 属性（顧客 ID, 伝票 ID, 購買商品）と m レコードを持つ元トランザクションデータ T である．はじめに，T をもとに各顧客の購買商品をパネル (b) のようにまとめる．次に，パネル (d) のように顧客 u_1, u_2, u_3 の購買履歴にダミーレコードを追加し，購買商品集合を等しくなるように加工する．この場合，パネル (c) に示すように各顧客の購買商品が等しくなっている（$I(u'_1) = I(u'_2) = I(u'_3) = I(u_1) \cup I(u_2) \cup I(u_3) = \{g_1, g_2, g_3, g_4, g_5\}$）ため，Jaccard 係数を用いても個人を再識別することはできない．

図 5.4 に，Online Retail Dataset へのダミーレコードを追加する方法の例を示す．図中の表 (a) は 2 顧客と 5 レコードを含む元データ例であり，表 (b) は 2 顧客と 5 レコードを含む加工データ例である．攻撃者は $I(u_1) = \{A, B, C\}$ と $I(u_2) = \{A, B\}$ を背景知識として有しており，Jaccard 係数を用いた再識別を試みてくるため，ダミーレコードを追加してこれを防ぐ．この例では，$I(u_1)$ と $I(u_2)$ が同じ集合 $\{A, B, C\}$ になるように u_2 のダミーレコード（表 (b) の第 6 レコード）を追加している．このダミーレコードは u_2 の最後のレコードと 4 属性（顧客 ID，伝票 ID，日時，時刻）で等しい値を持っており，商品/価格属性では追加すべき商品についての値を持ち，数量属性には 1 が記録されている．本研究では，購買商品のみを背景知識として持っている攻撃者を想定しているため，他の属性の値は再識別リスクに影響しない．

加工に必要なダミーレコードの数 Δm とデータの有用性にはトレードオフの関係がある．そのため，

Algorithm 2 ダミーレコード追加アルゴリズム

Input: $M, T, X_c = \{x_1, x_2, \ldots, x_c\}$
u を顧客，x をクラスタとする．

1. 各クラスタ x の顧客 u のトランザクションに，商品 $I(x) - I(u)$ が含まれるダミーレコードを追加する．ダミーレコードが追加された加工データを T' とする．
2. クラスタ $x \in X_c$ 内の各顧客が購買した商品集合を $I(x) = \bigcup_{u \in x} I(u)$ に統一し，それを反映した顧客データ M' を作成する．

Output: T', M'

もし全ての顧客を統一化しようとするとダミーレコードの数は膨大になり，データは役立たずになってしまう．そのため，顧客を購買商品の特徴によって小さなクラスタに分類することによって，加工に必要なダミーレコード数を小さくしてやる必要がある．

顧客をクラスタリングする最も簡単な方法は，まず代表となる c 人の顧客をランダムに選び，その他の顧客を距離が近い各代表者のクラスタに割り当てていく k-means ベースのトップダウンのクラスタリング方法である．ここで，c をクラスタの数と定義し，それらのクラスタ集合を $X_c = \{x_1, \ldots, x_c\}$ とし，それらのサイズを $s_i = |x_i|$ と表記する．クラスタ x_i は顧客集合 U の分割（partition），すなわち，$\bigcup_{i=1}^{c} x_i = U, x_i \cap x_j = \varnothing$ である．クラスタ数 c と k-匿名化のパラメータ k を混同しないように注意せよ．アルゴリズム 2 に各顧客にダミーレコードを追加する手法を示す．

5.3.2 顧客間の TF-IDF 距離

アルゴリズム 2 は，トランザクションデータを加工するにはあまりにもシンプルな手法である．一般的に，多くの商品を含む購買履歴データはいわゆるロングテールな分布をすることが知られており，少数の顧客が殆どのレコードを占めているため，単純なクラスタリング手法では多くのダミーレコードが必要になってしまう．例として，図 5.8 に単純な手法（Jaccard 係数を用いた k-means 手法）でクラスタリングしたときのクラスタサイズの分布を示す．この手法で PWS Cup 2016 のトランザクションデータをクラスタリングした場合，最も大きなクラスタには全 400 人中 211 人の顧客が含まれてしまい，33 のクラスタには 1 人の顧客しか含まれていなかった．この結果より，クラスタサイズには大きな偏りがあることがわかった．

クラスタの一極化を防ぐために，新たなクラスタリングの手法を提案する．既存手法では 2 顧客間の距離を Jaccard 係数で測っていたのに対し，提案手法では商品集合の TF-IDF を用いて距離を測る．すなわち，各顧客の商品購買行列（TF）に顧客の購買商品の逆数（IDF）を合わせたもの，つまり，商品行列に重みを付けたものを用いる．以上の TF-IDF の重みを用いた提案クラスタリング手法の詳細をアルゴリズム 3 に示す．

例として，図 5.5 に提案アルゴリズム手法で 4 人の顧客 $U = \{u_1, u_2, u_3, u_4\}$ を 2 つのクラスタ $X_c = \{x_1, x_2\}$ にクラスタリングする場合を示す．表 (a) は各顧客の購買商品リストであり，これを 2 値の行列に書き換えると表 (b) のようになる．この 2 値行列を，表 (c) に示す TF-IDF による重みづけされた行列に修正する．例えば，顧客 u_1 の商品 g_1 は TF = 1/2，IDF = 1 であるため，特徴量は

(a) Original data

User ID	Receipt ID	Date	Time	Goods	Price	Number
u_1	1	2010/12/1	8:00	A	5	10
u_1	1	2010/12/1	8:00	B	10	20
u_1	2	2010/12/2	12:00	C	100	30
u_2	3	2010/12/1	15:00	A	5	5
u_2	4	2010/12/5	20:00	B	10	15

(b) Processed data　　∗ : Dummy Record

User ID	Receipt ID	Date	Time	Goods	Price	Number	
u'_1	1	2010/12/1	8:00	A	5	10	
u'_1	1	2010/12/1	8:00	B	10	20	
u'_1	2	2010/12/2	12:00	**C**	**100**	30	
u'_2	3	2010/12/1	15:00	A	5	5	
u'_2	**4**	**2010/12/5**	**20:00**	B	10	15	
u'_2	**4**	**2010/12/5**	**20:00**	**C**	**100**	1	∗

図 5.4: Online Retail Dataset へのダミーレコード追加手法

Algorithm 3 TF-IDF を用いて購買商品に重みをつける手法

Input: $u_i \in U, I(u_i), c$

Step 1. $\boldsymbol{v}_i = (f_{i1}, f_{i2}, \ldots, f_{i\ell})$ を顧客 u_i の ℓ 次元の特徴ベクトルとする．ここで，

$$f_{ij} = \begin{cases} 1 & \text{if} \quad g_j \in I(u_i) \\ 0 & \text{otherwise.} \end{cases}$$

である．

Step 2. $D_j = \{u_i \in U | g_j \in I(u_i)\}$ を，商品 g_j を購入した顧客の集合とする．$f'_{ij} = (f_{ij} / \sum_{k=1}^{\ell} f_{ik})(\log \frac{n}{|D_j|} + 1)$ を，TF-IDF を用いた f_{ij} への重みとし，$\boldsymbol{v}'_i = (f'_{i1}, f'_{i2}, \ldots, f'_{i\ell})$ を顧客 u_i の特徴ベクトルとする．

Step 3. 集合 U の顧客を，k-means 手法と特徴ベクトル $\boldsymbol{v}'$ 間のコサイン類似度を用いて分類する．

Output: $X_c = \{x_1, x_2, \ldots, x_c\}$

0.5 である．この行列をもとに顧客間のコサイン類似度を求めると，表 (d) のように $x_1 = \{u_1, u_2\}$ と $x_2 = \{u_3, u_4\}$ の 2 クラスタに分類される．TF-IDF の値が似ているため，クラスタサイズの分布に偏りが少なくなることに注意せよ．

5.3.3　手法 1 : k-means クラスタリングを用いた匿名化手法

手法 1 は，各商品に TF-IDF の重みをつけ，顧客間のコサイン類似度に基づく k-means クラスタリングを行い，同一クラスタ内の顧客の購買商品集合が等しくなるようにダミーレコードを追加する手法である．図 5.6 に，手法 1 を用いて $c = 50$ のクラスタリングをしたときのクラスタサイズ分布を示す．図 5.8 の分布と比較すると，TF-IDF の重みのおかげでクラスタサイズの偏りは小さくなっている．ここで，x_{max} と x_{min} をそれぞれ最も大きいクラスタと最も小さいクラスタと定義すると，$|x_{max}|$ は 32 であり，$|x_{min}|$ は 1 であった．このように，クラスタサイズ分布の偏りはまだ残っており，大きなクラスタに属する顧客の加工のために大量のダミーレコードが必要になるため有用性が大きく損な

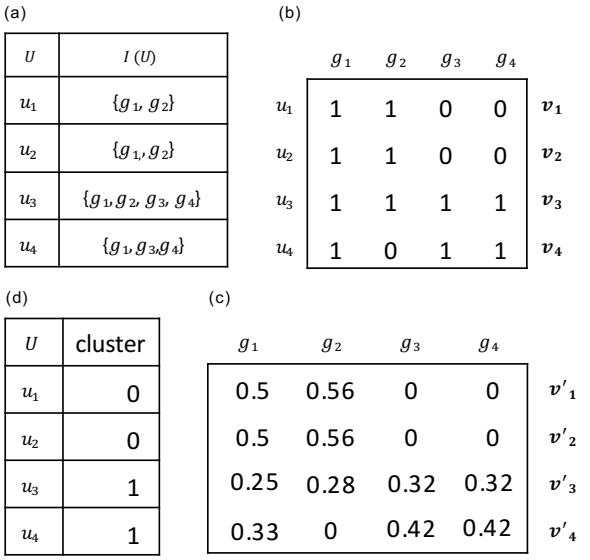


図 5.5: TF-IDF を用いたクラスタリングの一例

表 5.6: 単純な手法と提案手法の比較

項目	既存手法 (単純な k-means)	提案手法 1 (TF-IDF)	提案手法 2 (クラスタ調整)
アイデア	k-means	TF-IDF	クラスタの最小サイズに 制限を設ける (s_{min})
単一の巨大なクラスタ問題	問題あり	問題なし	問題なし
クラスタサイズの偏り	偏りあり	偏りあり	偏りなし
最大クラスタのサイズ $\lvert x_{\max}\rvert$	211	32	16
要素 1 のクラスタの数	33	0	0

われる．

5.3.4 手法 2：クラスタサイズを調整した匿名化手法

クラスタサイズの偏りが大きい問題を解決するために，最小クラスタサイズに制限を設ける手法 2 を提案する．手法 2 では，最小クラスタサイズ上限値を s_{min}（k-匿名化のパラメータ k と同じ意味を持つ）とし，全てのクラスタの大きさがこれを下回らないように加工を行う．

アルゴリズム 4 に手法 2 の詳細を示す．この手法では，最大クラスタ x_{max} に属する顧客をクラスタサイズが s_{min} より少ないクラスタに移していく作業を，全クラスタの大きさが s_{min} 以上になるまで繰り返す．s_{min} の値域はクラスタ数 c に依存し，$\{2, 3, \ldots, \lfloor n/c \rfloor\}$ の範囲の値をとる．

図 5.7 に，手法 2 を用いて $c = 50, s_{min} = 5$ のクラスタリングをしたときのクラスタサイズ分布を示す．図 5.6 の結果と比較すると，最大クラスタのサイズは 32 から 16 まで下がり，全てのクラスタが 5 人以上の顧客を含んでいる．

まとめとして，単純な k-means によるクラスタリング手法と，提案手法 1,2 の比較を表 5.6 に示す．

Algorithm 4 手法 1 の調整アルゴリズム (手法 2)

Input: s_{min}, c, M, T

手法 1 を用いてクラスタリングを行う.

クラスタ集合を $X_c = \{x_1, x_2, \ldots, x_c\}$ とする.

for x **in** $\{x_i \in X_c \big| |x_i| < s_{min}\}$ **do**

最大クラスタ: $x_{max} \in X_c$

while $|x'| < s_{min}$ **do**

$$u_j = \underset{u_j \in x_{max}, u_i \in X_c}{\arg\max} J(I(u_i), I(u_j)), \quad x'_{max} \leftarrow x_{max} - \{u_j\}, \quad x' \leftarrow x \cup \{u_j\}$$

end while

end for

ダミーレコードを追加する

Output: M', T'

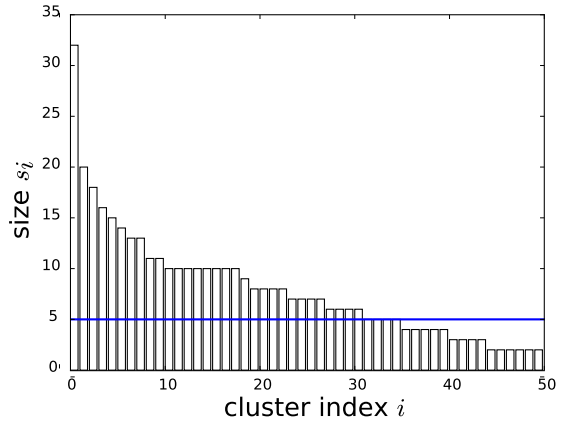


図 5.6: 手法 1 のクラスタサイズ分布 ($c = 50$)

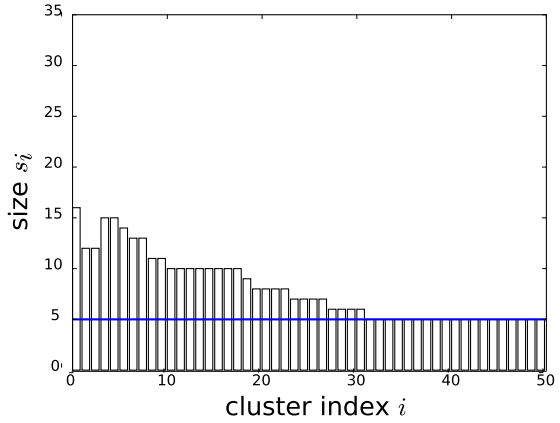


図 5.7: 手法 2 のクラスタサイズ分布 ($s_{min} = 5, c = 50$)

5.4 提案手法の評価

5.4.1 ダミーレコード数と有用性の関係

匿名化されたデータの有用性を，PWS Cup 2016[71, 72, 73] で用いられた 3 つの有用性指標（U1-cMAE, U2-cMAE, and U3-RFM）によって評価する．U1-cMAE と U2-cMAE は，元データと加工データの顧客の性別と国籍に着目して作られたクロス集計間の平均絶対誤差（MAE）で定める有用性を評価する指標である．PWS Cup 2016 で用いられたデータセットには 2 種類の性別と 36 種類の国籍が含まれているので，クロス集計には 72 個のセルが存在し，それらの値（平均購買価格など）は匿名化によって変化する．また，U3-RFM は匿名化されたデータの有用性を RFM（Recently, Frequency, Monetary）分析によって評価する指標であり，RFM の 3 軸によって顧客は 1,000 タイプに分けられ，各タイプの頻度（人数）が加工によってどれだけ変化したかを評価される．これらの指標の評価値には元データと加工データの誤差が用いられているので，匿名化データの有用性は有用性指標の評価値が小さいほど高くなる．

匿名化されたデータの有用性とダミーレコードの数 Δm には大きな相関がある．それゆえ，Δm と

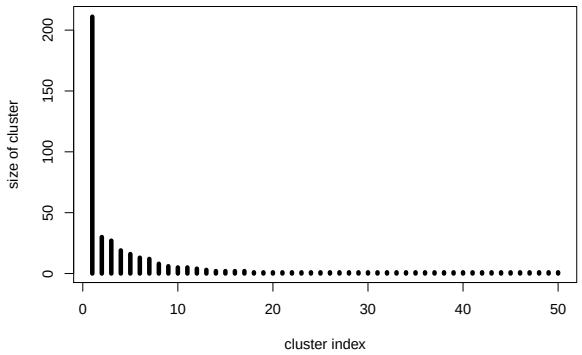


図 5.8: Jaccard 係数によるクラスタリングのクラスタサイズ分布

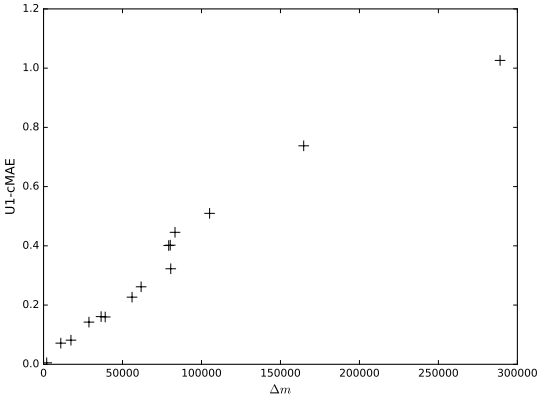


図 5.9: Δm と U1 指標の関係

表 5.7: Δm と有用性指標の相関関係

	Δm	$U1$-cMAE	$U2$-cMAE	$U3$-RFM	Jaccard 再識別	ランダム再識別	c
Δm	1.0000						
$U1$-cMAE	0.9798	1.0000					
$U2$-cMAE	0.9798	1.0000	1.0000				
$U3$-RFM	0.9547	0.9876	0.9876	1.0000			
Jaccard 再識別	-0.8586	-0.9327	-0.9327	-0.9494	1.0000		
ランダム再識別	-0.8489	-0.9247	-0.9247	-0.9432	0.9996	1.0000	
c	-0.8454	-0.9220	-0.9220	-0.9406	0.9994	0.9999	1.0000

有用性評価値の相関を示すことによって，Δm も有用性指標の一つになる．表 5.7 に，PWS Cup 2016 の有用性指標と Δm の関係を示す．Jaccard 再識別の行にはアルゴリズム 1 で示した Jaccard 係数を用いた再識別の結果（10 回試行の平均値）との相関を示し，ランダム再識別の行にはクラスタ内からランダムに個人を再識別した結果との相関を示す．Δm と 3 つの有用性指標の間には非常に強い負の相関があるため，Δm が増えるほどデータの有用性は下がることがわかる．クラスタ数 c を増やすとその分 Δm が減るため c と Δm の間の相関係数は –0.8454 であり，再識別率は増加する．

相関のある例として，図 5.9 に Δm と U1-cMAE 指標についての散布図を示す．ダミーレコード数が $0 < \Delta m \leq 300,000$ のとき，有用性評価値は $0.0 \leq U_1 \leq 1.02$ の値を取り，Δm が増えるほど有用性は下がっている．

5.4.2 Δm と Jaccard 係数の理論値

図 5.10 に，単純な k-means クラスタリング手法によって加工した際の Δm を示す．この手法では，25 クラスタの顧客を統一化するために 540,583 個のダミーレコードを追加する必要がある．図 5.11 に，提案手法 1,2 の Δm と c の比較結果を示す．黒実線 (Method 1) は手法 1 の Δm を示しており，赤実線 (Method 2) は手法 2 の Δm を示している．図 5.10,5.11 を比較すると，提案手法のダミーレコード数は単純な手法でクラスタリングするときよりも明らかに少なくなっていることがわかる．この実

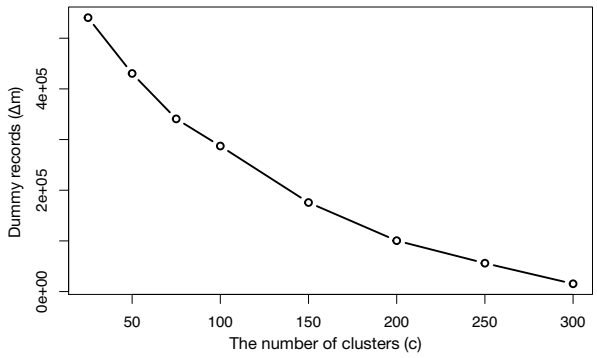


図 5.10: 単純な k-means 手法でクラスタリングしたときの Δm

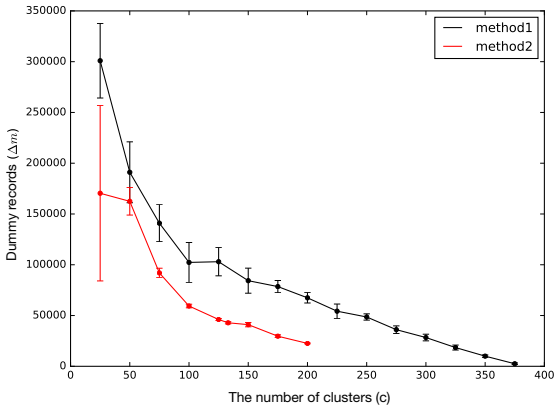


図 5.11: 手法 1 と 2 の Δm の比較

験では，400 人の顧客を閾値 $s_{min} = \lfloor \frac{n}{c} \rfloor$ でクラスタリングしている．提案手法 1,2 を比較すると，手法 2 は手法 1 よりも約 53%ダミーレコード数が少なくなっている．

ダミーレコード追加匿名化手法必要なに Δm の理論値の推定を試みる．m, n, ℓ, c が与えられた時の，Δm の期待値 $E(\Delta m)$ は，$E(\Delta m) = n\ell\{(1 - \frac{1}{\ell})^{\frac{m}{n}} - (1 - \frac{1}{\ell})^{\frac{m}{c}}\}$ で求められる．図 5.12 に $E(\Delta m)$ と c の分布を示す．

顧客 u_i と u_j の購買商品集合の類似度として，Jaccard 係数を以下のように定義する．

定義 5.4.1 $\mu = 1/\binom{n}{2} \sum_{i \neq j \in U} J(u_i, u_j)$ を T における *2* 顧客間の *Jaccard* 係数の平均値とする．ここで，$J(u_i, u_j) = |I(u_i) \cap I(u_j)|/|I(u_i) \cup I(u_j)|$ である．

データ統計量が与えられた時に，以下の方法で平均 Jaccard 係数を推定する．

命題 5.4.1 顧客が一年間に購買する商品の平均種類数 b と，*2* 顧客間の購買商品集合の平均重複数 h は，$h = |I(u_i) \cap I(u_j)|$ であり，$h = 2b\mu/1 + \mu$ である．

(証明) μ を

$$\mu = \frac{E(|I(u_i) \cap I(u_j)|)}{E(|I(u_i)|) + E(|I(u_j)|) - E(|I(u_i) \cap I(u_j)|)} = \frac{h}{2b - h}$$

と変形し，これを h について解くことにより，与式を得ることができる． (Q.E.D)

5.4.3 有用性と安全性

表 5.8 に Δm と s_{min} の関係を示す．Δm の値は，s_{min} が各 c の $\lfloor \frac{n}{c} \rfloor$ のときに最小化されている．前述したように，Jaccard 係数は小さな範囲に分布しており，標準偏差が 0.01 以下である．表中の各 c についての再識別の欄には，実際の再識別率の値が記録されている．

図 5.13 に，提案手法で加工されたデータの被再識別率を示す．各加工データにアルゴリズム 1 の Jaccard 再識別手法を適用した結果，各クラスタ内で少なくとも一人は購買の頻度が高い顧客を識別

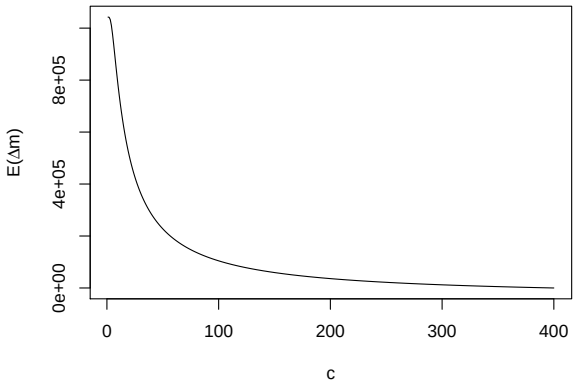


図 5.12: c と $E(\Delta m)$ の関係

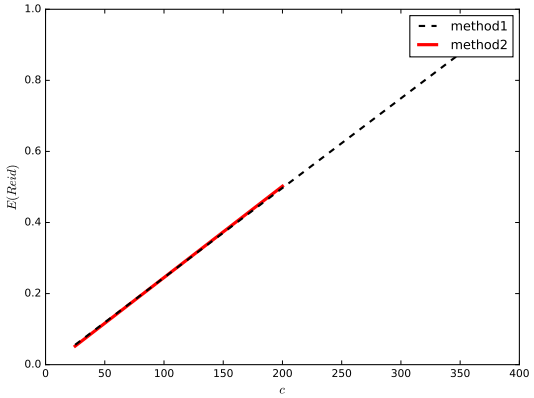


図 5.13: 手法 1 と手法 2 の再識別についての安全性の比較

表 5.8: s_{min} と Δm の関係

	$c = 50$			$c = 100$			$c = 125$		
	Δm	Jaccard 再識別	ランダム 再識別	Δm	Jaccard 再識別	ランダム 再識別	Δm	Jaccard 再識別	ランダム 再識別
手法 1	182,897	0.1728	0.1235	128568	0.3060	0.2488	97581	0.3692	0.3120
手法 2									
$s_{min} = 2$	183,902	0.1729	0.1223	99228	0.3061	0.2475	60492	0.3687	0.3105
$s_{min} = 3$	175,449	0.1726	0.1222	68357	0.3041	0.2480	46101	0.3667	0.3102
$s_{min} = 4$	162,474	0.1723	0.1218	59374	0.3044	0.2465			
$s_{min} = 8$	125,798	0.1681	0.1218						

することができた．提案手法 1,2 で匿名化されたデータから顧客が再識別される割合（再識別率）の期待値は，$E(Reid) = c/n$ という単純な結果になった．

5.5 まとめ

本章では，顧客の購買商品特徴によって個人が特定されるリスクがあることを示し，データにダミーレコードを追加することによって安全性を高める匿名化手法を提案した．提案手法では，TF-IDF に基づく重みを購買商品の特徴に与えることにより，顧客の集合をいくつかのクラスタに分割する．クラスタの最小サイズに制限をかけることによってダミーレコードの数を減らすことができることを示した．

第6章　健康診断データとレセプトデータの匿名加工情報を用いた疾病リスク分析

6.1　導入

本章では，健康診断データや傷病/医薬品レセプトデータに対する匿名化の影響を，実験的に評価する．実験に用いるデータとして，あるヘルスケア企業が取得した 10 年間の 20 万人分の健康診断データと 28 万人分のレセプトデータのから成る匿名加工情報を用いる．これらの匿名加工情報を未加工のデータとして扱い，加工を施した後に分析をして，匿名化が分析結果へ及ぼす影響を明らかにする．

このために，次の方法により匿名加工情報の分析を行う．

1. 健康診断データと，診断されたことのある傷病/処方されたことのある傷病/医薬品のレセプトデータをクロス集計して，疾病の相対リスクを分析する．

2. がんと脳卒中を対象とし，3 年以内の罹患と説明変数（健康診断結果）の関係をロジスティック回帰を用いて分析して，従来のコホート研究結果と比較する．

3. K 近傍法 (KNN)，RBF Support Vector Machine（SVM），Decision Tree (Tree)，Random Forest（RF）を予測アルゴリズムに使い，3 年以内の罹患を予測するモデルを 274 種類の傷病について作成する．

4. 健康診断データをいくつかの疑似識別子 (QI : Quasi Identifier) について k–匿名性 (k–anonymity) を満たす複数のアルゴリズムによる加工を行い，分析精度がどれだけ変化するのかを明らかにする．

上記の分析から，導かれた本章の主要な結論は以下の通りである．

1. 高血圧を危険因子としたときの循環器系の疾患の相対リスクが 1.75 であることを明らかにした．一度もレセプトに記録が無い（病院にかかっていない）特異な集団「健康集団」があるが，その健康診断結果はそれほど健康ではない．

2. 十分な睡眠をとる人が三年以内に脳卒中となる罹患リスクが，睡眠不足の人に比べて 0.787 倍になることや，加齢による脳卒中のリスクが [13] と整合した結果が得られたこと．

3. ランダムフォレストが最も予測精度が良く，274 種類の傷病の平均 F 値は 0.65 である．

4. 性別・年齢を QI として $k = 1,000$ までの追加の k–匿名化をした結果，$k = 1,000$ の時レコード数は約 10%減少するが，加工しても 274 種類の最大誤差は 0.007 であり，十分に精度良いモ

表 6.1: 先行研究との比較

	野田ら [13]	本分析
データ利用方法	人口動態統計死亡票の目的外使用	匿名加工情報
人数	92,277	68,629
説明変数数	12	37
傷病数	4	274
対象期間	1993–2001（9 年間）	2008–2016（9 年間）
被験者の年代	40 – 79	19 – 74
分析方法	Cox 回帰	ロジスティック回帰 機械学習等
目的変数	死亡	三年以内の罹患

デルが作れることを示した．病歴を QI とした k–匿名化においても，相対リスクの相対誤差が $k = 10$ で 0.073 であり，十分な精度を保持する．

5. 匿名加工情報と当該ヘルスケア企業より確認した予測健診データの OR の誤差は，平均 $2.5 \cdot 10^{-4}$ であった．

また，これらの分析結果と既存の研究の比較も行う．傷病と因子の関係を明らかにした野田ら [13] が行った約 10 万人を対象とするコホート研究と本研究の比較を，表 6.1 に示す．彼らは 10 万人のコホートについて 8 年間追跡調査を行い，住民健診の検査結果とその後の死亡の関係を男女別に Cox 比例ハザードモデルを用いて偏回帰係数を求める分析を行い，統計的に有意な因子とその相対危険度を明らかにした．一方，本研究は匿名加工情報を活用することで，従来の 4 種から 274 種の多くの疾病について分析することが可能になった．

6.2 健康診断データと傷病/医薬品レセプトデータ

6.2.1 概要

本章で分析する健康診断データの個人・レコード・属性数を表 6.2 に示し，各属性を表 6.3 に示す．第 3–17, 20–24 属性には連続値が，それ以外の属性には離散値が記録されている．第 1–27 属性は個人の身体情報を示し，第 28 属性以降は個人の問診 28 問 [20] への回答結果を示している．健康診断データには，2008 年から 2018 年までの 20 万人分のデータが記録されている．

第 25 属性の「健診ランク」は，bmi や中性脂肪等の 12 属性から個人のリスクを判定する指標の分布であり，A(非肥満) と B（肥満），1（リスクなし）–4（服薬投与）を組み合わせた 8 ランクに分類される．健康診断データにおける健診ランク（レコード）を表 6.4 に示す．13.2%のレコードが最も健康なランクである A1（非肥満・リスクなし）に，9.3%のレコードが最も不健康なランク B4（肥満・服薬投与）に該当した．また，この属性の情報を持たないレコードも多く，全体の 43.8%が “不明” となっていた．

表 6.2: 3 データの統計情報

データ名	健康診断データ	傷病レセプト	医薬品レセプト
個人数 n	198,740	288,568	279,199
レコード数 m	964,636	39,363,878	31,465,504
属性数 ρ	49	15	21
レセプト枚数	–	11,912,236	9,000,249
対象年	2008–2018	2012–2018	2012–2018

6.2.2 レセプトデータ

本レセプトデータには，各個人が診断された傷病の詳細が記録されている傷病レセプトデータと，各個人に処方された医薬品の詳細が記録されている医薬品レセプトデータの 2 種類がある．傷病/医薬品レセプトデータの統計量を表 6.2 に示す.

傷病レセプトデータの第 7–12 属性と医薬品レセプトデータの第 14–17 属性は傷病/医薬品分類コードである．傷病の分類コードには国際疾病分類第 10 版（ICD10）[21] が，医薬品の分類コードには解剖治療化学分類（ATC 分類）[22] が用いられており，これらの分類コードは大分類＞中分類＞小分類＞細分類とカテゴリ分けされている．例えば脳梗塞という病気は，循環器系の疾患（大分類コード：*I*）の中の脳梗塞カテゴリ（中分類コード：*I63*）の中の脳梗塞（細分類コード：*I639*）に分類される．

各レセプトデータは複数のレコードから成る．表 6.2 から，傷病レセプトでは平均 3.3 レコード/枚，医薬品レセプトでは平均 3.50 レコード/枚がある．しかし，図 6.1 に示すように傷病レセプトデータにおける各顧客ごとのレコード数分布（降順）は一様ではない．上位 9 名の個人のレコード数が飛びぬけて多く，歪んでいる．10 位の個人のレコード数が 4,015 であるのに対し，9 位の個人のレコード数は 321,828 であり，1 位の個人は 2,588,244 レコードも記録されている．

レセプトの枚数についても同様に歪んでおり，1 位の個人は 1 人で 855,147 枚のレセプトを処方されている．医薬品レセプトデータにおいても，上位 9 名の頻度とレセプト数が飛びぬけて多いことがいえる.[1].

6.2.3 傷病/医薬品と健康診断データ

3 つのデータ（健康診断データ，傷病レセプトデータ，医薬品レセプトデータ）を用いることにより，診断されたことのある傷病，処方されたことのある傷病，医薬品と健康診断データをクロス集計する．3 つのデータは，同一の個人情報取扱事業者により加工された単一の匿名加工情報である．法令に従った規則性を有しない方法で生成された，共通の仮 ID が振られている．図 6.2 に 3 データ間の包含関係をベン図で示す．3 データ全てに記録されている個人は 178,033 人であり，傷病/医薬品レ

[1]仮個人 id と仮レセプト id について分析を行った結果，1 枚のレセプトが 2 人の個人に対応するケースが存在した（傷病：8,275 枚，医薬品：6,504 枚）． 仮レセプト id 属性は仮名化されたものであるため，その際に重複が生じた可能性がある．

表 6.3: 健康診断データに記録されている情報

index	種類	属性名	欠損値数	一意な値の数	平均識別確率
1	離散/身体	仮個人 id	0	–	–
2	離散/身体	健診受診月	0	1	$1.31 \cdot 10^{-4}$
3	連続/身体	身長	1,048	5	$7.75 \cdot 10^{-4}$
4**	連続/身体	体重	1,060	19	$1.14 \cdot 10^{-3}$
5*	連続/身体	内臓脂肪面積	964,296	262	$3.15 \cdot 10^{-4}$
6	連続/身体	bmi	1,065	0	$3.60 \cdot 10^{-4}$
7**	連続/身体	腹囲 実測	76,519	28	$8.46 \cdot 10^{-4}$
8	連続/身体	収縮期血圧	154,021	0	$1.43 \cdot 10^{-4}$
9	連続/身体	拡張期血圧	154,023	0	$1.04 \cdot 10^{-4}$
10	連続/身体	中性脂肪	29,740	192	$1.38 \cdot 10^{-3}$
11	連続/身体	hdl コレステロール	29,765	173	$4.11 \cdot 10^{-4}$
12	連続/身体	ldl コレステロール	29,922	116	$4.00 \cdot 10^{-4}$
13	連続/身体	got ast	28,140	7	$2.01 \cdot 10^{-4}$
14**	連続/身体	gpt alt	28,141	5	$2.56 \cdot 10^{-4}$
15	連続/身体	γ gtp	28,161	88	$8.13 \cdot 10^{-4}$
16*	連続/身体	空腹時血糖	372,933	5	$2.81 \cdot 10^{-4}$
17	連続/身体	hba1c ngsp	111,921	15	$1.22 \cdot 10^{-4}$
18	離散/身体	尿糖	6,177	0	$1.21 \cdot 10^{-5}$
19	離散/身体	尿蛋白	5,301	0	$1.21 \cdot 10^{-5}$
20*	連続/身体	ヘマトクリット値	444,694	0	$3.72 \cdot 10^{-4}$
21**	連続/身体	血色素量	332,031	0	$1.54 \cdot 10^{-4}$
22	連続/身体	赤血球数	331,553	2	$3.74 \cdot 10^{-4}$
23*	連続/身体	クレアチニン	746,905	150	$3.83 \cdot 10^{-4}$
24*	連続/身体	尿酸	741,879	2	$1.25 \cdot 10^{-4}$
25*	離散/身体	健診ランク	422,239	0	$2.34 \cdot 10^{-5}$
26	離散/身体	メタボリックシンドローム判定	143,700	0	$1.18 \cdot 10^{-5}$
27	離散/身体	保健指導レベル	154,261	0	$1.40 \cdot 10^{-5}$
28	離散/問診	服薬 1 血圧	58,424	0	$1.18 \cdot 10^{-5}$
29	離散/問診	服薬 2 血糖	58,512	0	$1.16 \cdot 10^{-5}$
30	離散/問診	服薬 3 脂質	58,520	0	$1.13 \cdot 10^{-5}$
31	離散/問診	既住歴 1 脳血管	350,483	0	$1.20 \cdot 10^{-5}$
32	離散/問診	既住歴 2 心血管	350,393	0	$1.19 \cdot 10^{-5}$
33	離散/問診	既住歴 3 腎不全・人工透析	350,590	0	$1.14 \cdot 10^{-5}$
34	離散/問診	貧血	351,960	0	$1.18 \cdot 10^{-5}$
35	離散/問診	喫煙	40,513	0	$1.16 \cdot 10^{-5}$
36	離散/問診	体重変化 20 歳からの	356,876	0	$1.21 \cdot 10^{-5}$
37	離散/問診	運動習慣 30 分以上	205,592	0	$1.01 \cdot 10^{-5}$
38	離散/問診	歩行又は身体活動	205,783	0	$9.73 \cdot 10^{-6}$
39	離散/問診	歩行速度	357,278	0	$1.16 \cdot 10^{-5}$
40	離散/問診	体重変化 1 年間	371,610	0	$1.01 \cdot 10^{-5}$
41	離散/問診	食べ方 1 早食い等	357,880	0	$1.43 \cdot 10^{-5}$
42	離散/問診	食べ方 2 就寝前	205,902	0	$1.01 \cdot 10^{-5}$
43	離散/問診	食べ方 3 夜食・間食	220,089	0	$9.62 \cdot 10^{-6}$
44	離散/問診	食習慣	207,343	0	$1.06 \cdot 10^{-5}$
45	離散/問診	飲酒	271,688	0	$1.44 \cdot 10^{-5}$
46*	離散/問診	飲酒量	459,731	0	$1.52 \cdot 10^{-5}$
47	離散/問診	睡眠	357,548	0	$1.11 \cdot 10^{-5}$
48	離散/問診	生活習慣の改善	364,315	0	$1.54 \cdot 10^{-5}$
49	離散/問診	保健指導の希望	356,536	0	$1.13 \cdot 10^{-5}$

*：2.4 節のクレンジング手法 1 で削除を行った.

**：2.4 節のクレンジング手法 2 で削除を行った.

表 6.4: 健康診断データにおける健診ランクの分布

状態	健診ランク	レコード数	割合
非肥満	A1	127,550	0.132
	A2	87,487	0.091
	A3	45,155	0.047
	A4	66,744	0.069
肥満	B1	24,573	0.025
	B2	48,367	0.050
	B3	52,726	0.055
	B4	89,794	0.093
不明	不明	422,239	0.438

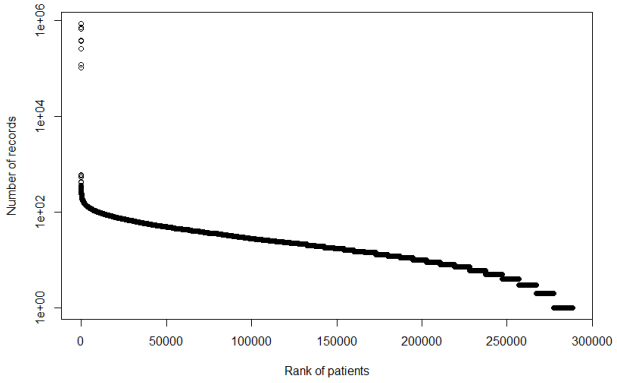


図 6.1: 傷病レセプトデータにおける各顧客ごとのレコード数分布

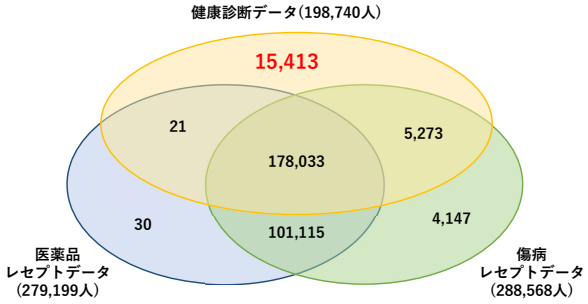


図 6.2: 3 データ間の包含関係

セプトデータにしか記録されていない個人も存在した．傷病/医薬品グループ間には個人の重複があり，個人は複数のグループに属することができる．

図 6.3 に傷病グループごとの健診ランクを示す．x 軸は傷病分類コード（大分類）を意味しており，“*He*” は 3.2.3 節で後述する健康集団，“*All*” は健康診断データ全体を示している[2]．傷病グループ O (妊娠, 分娩および産じょく）と傷病グループ P（周産期に発生した病態）の個人は A1(非肥満・リスクなし) の割合が飛びぬけて高く，どちらも 6 割を超えている（グループ O：0.606，グループ P：0.624）ため，健康な個人が多い．傷病グループ E（内分泌, 栄養および代謝疾患）や傷病グループ I（循環器系の疾患）は A4（非肥満・服薬投与）,B4（肥満・服薬投与）の割合が他グループより高い（グループ E：合計 0.420，グループ I：合計 0.470）ため，不健康な個人が多い．また，図 6.4 に傷病グループごとの拡張期/収縮期血圧の平均値を示す．この結果からも，平均血圧が低い健康なグループ（O, P）と，平均血圧が高い不健康なグループ（E, I）を観測できる．

[2]傷病グループ X に属する個人は健診ランクの情報を持たなかったため，省いている．

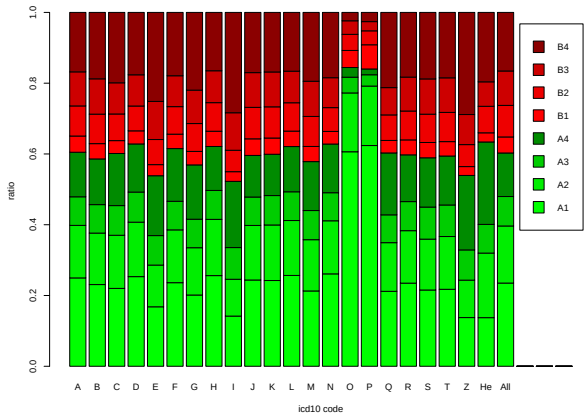


図 6.3: 傷病グループごとの健診ランク

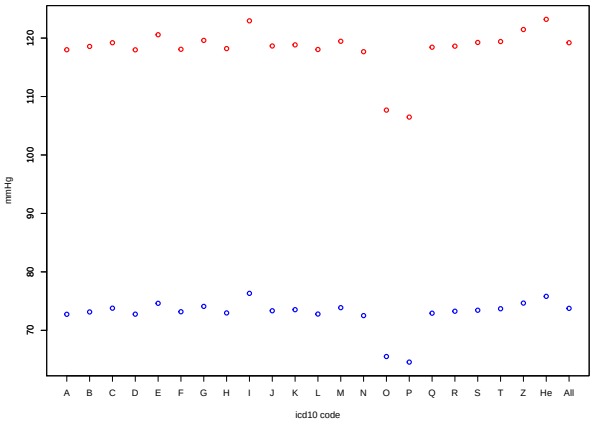


図 6.4: 傷病グループごとの拡張期/収縮期血圧の平均値

表 6.5: クレンジング後の健康診断データの統計量

	対象年	レコード数	ユーザ数 n	欠損値セル数	特徴量数 F
処理前	2008-2018	964,635	198,740	10,536,861	49
処理後	2008-2016	203,521	68,629	0	38

6.2.4 健康診断データのクレンジング

一般に，健康診断データには多くの欠損値が含まれている．本章のデータにおいても，全体の 23.8% のセルが情報を持たないセルである．そのため，分析の前にデータをクレンジングする必要がある．分析の障害になる欠損値を含むレコードや相関が高い冗長な属性，カテゴリカル変数には次の前処理を行う．

1. 欠損値レコードの多い 7 特徴量（列）を削除.
2. 多重共線性 [28] をなくすために，相関係数が 0.7 以上ある 2 変数の一方を削除（4 特徴量）.
3. 欠損値を含むレコード（行）の削除.
4. カテゴリカル変数をダミー変数に変更.

また，後述する 6.3.4 節では，分散 0 の属性を削除している．処理前後の健康診断データの統計量を表 6.5 に，データから削除したデータを表 6.3 の index 列に示す.

レセプトデータに含まれる ICD10 の中分類コード (1,490 種類) の傷病情報のうち，健診受診年 ≤ 傷病記録年 ≤ 健診受診年 + 2 の条件を満たすものを健康診断データに追加する．健康診断の結果と傷病の発病までには，一定の期間がかかると考えられる．そこで，生活習慣を改善し罹患を防止する期間を設け，発病までの期間を考慮するために 3 年の区間を定義した.

罹患情報追加後の ICD10 についての罹患者数分布の上位 100 件を図 6.5 に示す．1490 種類の傷病のうち 16 % (225 種類) はクレンジングによって記録が消え，レコード数が 0 であった．図中の赤線は zipf の法則 ($f(x)$ を x 番目に多い値の頻度，a, c を定数としたとき，$f(x) = \frac{a}{x^c}$ とするモデル) による近似値である.

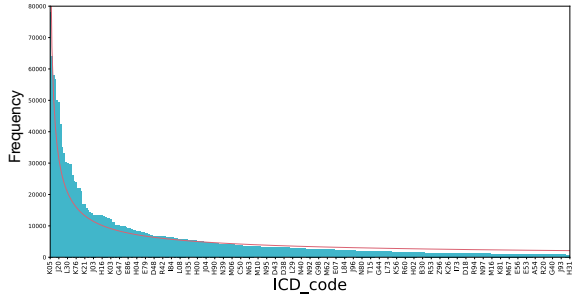


図 6.5: 疾病ごとの罹患者数

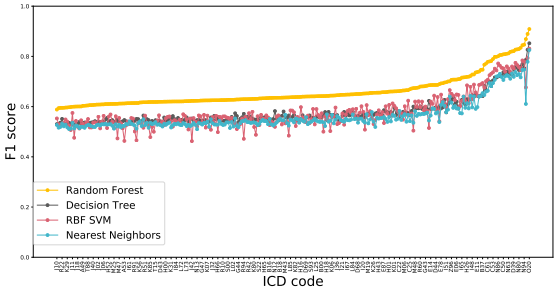


図 6.6: 疾病予測モデルの F 値

6.3 データ分析

6.3.1 概要

本章では，次の方法により匿名加工情報の分析を行う．

1. 健康診断データと，診断されたことのある傷病/処方されたことのある傷病/医薬品のレセプトデータをクロス集計して，疾病の相対リスクを分析する．
2. がんと脳卒中を対象とし，3 年以内の罹患と説明変数（健康診断結果）の関係をロジスティック回帰を用いて分析して，従来のコホート研究結果と比較する．
3. 4 種類の予測アルゴリズムを使い，274 種類の傷病について 3 年以内の罹患を予測するモデルを作成する．

6.3.2 傷病の相対リスク分析 (1)

分析手法

傷病/医薬品グループの相対リスク (Relative Risk) を求める．相対リスク [23] とは，ある危険因子（例えば「高血圧」）に曝露した場合，それに曝露しなかった場合に比べて何倍疾病に罹りやすくなるかを表す指標である．例として，表 6.6 の場合を考える．高血圧である個人が傷病 A に罹患する確率が 100/200 であるのに対し，高血圧でない個人の罹患率は 10/200 であるため，この場合の相対リスクは

$$RR_{\text{高血圧}} = \frac{Pr[A|\text{高血圧}]}{Pr[A|\text{正常域血圧}]} = \frac{(100/200)}{(10/200)} = 10$$

である．これは，高血圧の個人はそうでない個人の 10 倍傷病 A にかかりやすい，ということを意味している．

分析結果

高血圧を危険因子とした各傷病の相対リスクを，それぞれ表 6.7 に示す．ここで，相対リスクを求める際には罹患年や診断日などの時間情報は無視している．これらの表から，高血圧に対する相対リ

表 6.6: 高血圧の相対リスクに関する 2×2 分割表

	A に罹患している	A に罹患していない
高血圧	100	100
正常域血圧	10	190

表 6.7: 高血圧を危険因子とした各傷病の相対リスク $RR_{\text{高血圧}}$

分類コード	分類	相対リスク
I	循環器系の疾患	1.748
Z	健康状態に影響をおよぼす要因および保健サービスの利用	1.462
E	内分泌, 栄養および代謝疾患	1.305
G	神経系の疾患	1.136
C	新生物＜腫瘍＞血液及び造血器の疾患ならびに免疫機構の障害	1.104
T	損傷, 中毒およびその他の外因の影響	1.089
M	筋骨格系および結合組織の疾患	1.089
S	損傷, 中毒およびその他の外因の影響	1.059
Q	先天奇形, 変形および染色体異常	1.059
K	消化器系の疾患	1.000
R	症状, 徴候および異常臨床所見・異常検査所見で他に分類されないもの	0.993
B	感染症および寄生虫症	0.990
D	新生物＜腫瘍＞血液及び造血器の疾患ならびに免疫機構の障害	0.984
J	呼吸器系の疾患	0.973
N	尿路性器系の疾患	0.957
F	精神および行動の障害	0.951
H	眼および付属器の疾患，耳および乳様突起の疾患	0.943
L	皮膚および皮下組織の疾患	0.930
A	感染症および寄生虫症	0.904
O	妊娠, 分娩および産じょく	0.184
P	周産期に発生した病態	0.108

スクが高いグループ（傷病 I：1.748, 傷病 Z：1.462, 傷病 E：1.305）と低いグループ（傷病 O：0.184, 傷病 P：0.108）が観測できる．

考察と健康集団

本節では，健康診断データには登場するが，レセプトデータには登場しない個人に着目する．これらの個人を健康集団 He（傷病を診断されたことも，医薬品を処方されたこともない健康な集団）と呼ぶ．図 6.2 からわかるように，15,413 人の個人が健康集団に属している．図 6.3, 6.4 に健康集団 He についての分析結果を示す．

意外なことに，健康集団の診断結果はそれほど健康ではなく，むしろ健診ランクにおいては，図 6.3 からわかるように A4 や B4 の割合が高く，図 6.4 からわかるように血圧の平均値も他グループより高い（収縮期：1 位，拡張期：2 位）．健康診断データの他の属性についても分析した結果，健康集団

は診断結果はそれほど健康ではないわりに，問診結果は健康的（飲酒はしない，運動はしてる，等）であることが判明した．

6.3.3 傷病のロジスティック回帰分析 (2)

分析手法

がん（ICD10：C00-C99）と脳卒中（ICD10：I60-I69）を対象にして，3 年以内の罹患を目的変数，健康診断結果を説明変数として，ロジスティック回帰を用いて次のように分析する．

ある被験者 i の 3 年以内の傷病 y の罹患確率 p_{iy} を

$$p_{iy} = \frac{1}{1 + e^{-z_i}} \tag{6.1}$$

で表す．ここで，z_i は健康診断データから得られる 38 種類の説明変数 x_{log} と定数 α_{log}，各変数の係数 β_{log} について

$$z_i = \alpha_{log} + \beta_{log,1}x_{log,1} + \beta_{log,2}x_{log,2} + \cdots + \beta_{log,F}x_{log,F} \tag{6.2}$$

で定められる．

ある $x_{log,1}$ について，他の変数の影響を調整したオッズ比 (adjusted Odds Ratio) は，$OR = e^{\beta_{log,1}}$ で与えられる．罹患率が十分に小さい時，オッズ比と相対リスクが等しいことがよく知られており [24]，本章では説明変数 $x_{log,1}$ による罹患影響をオッズ比から確認する．

表 6.1 に野田らの実験と本分析の比較を示す．野田らの実験結果と比較する脳卒中とがんについては，母集団を先行研究と合わせるために健康診断データの 40 代以降のユーザーを抽出し，健康診断データの 38 特徴量，173,213 レコードを用いて分析を行う．また，他の傷病については母集団を全年代にするため 203,521 レコードを分析に使用する．分析には python の statsmodels ライブラリ [103] を用いる．

分析結果

表 6.8 に脳卒中，がん，インフルエンザについてのロジスティック回帰の結果を示す．estimate の正の値は罹患リスク増加，負の値は罹患リスク低下をそれぞれ表しており，*のついている値は統計的な有意差が確認できたものである．各 OR は，連続値の場合，値の増加による影響，2 値のカテゴリカル変数は 0 を基準に 1(質問に対して “はい” と答えた)，3 以上のカテゴリカル変数では最初の値をそれぞれ基準とした各値のオッズ比を表している（estimate が 0.000 の値は，estimate が極めて小さい値と基準値を区別するための表記である）．例えば，脳卒中で睡眠の OR=0.787 から，睡眠が十分に取れている人 (睡眠=1) は取れていない人 (睡眠=0) に比べて脳卒中の 3 年以内罹患リスクが 0.787 倍である．3 年以内の脳卒中罹患リスクには 22 因子，がん罹患には 33 因子，インフルエンザには 25 因が有意であった．

表 6.8 の相対リスク RR は，野田ら [13] の研究結果を表す．ただし，BMI は 19 未満をベースとした時の 19 以上 21 未満の相対リスク，尿蛋白は+以上を尿蛋白異常とした時の尿蛋白正常 (−, ±) に対する相対リスクである．

脳卒中の年齢，収縮期血圧では本分析の OR と野田らの RR から，同等の結果が得られていることがわかる．脳卒中とがんの両方で，既存研究と同様の結果が 5 項目から得られた．一方で，がんの喫煙による RR は既存研究が 1.51 に対して本分析では $OR = 0.920$ で不整合していた．この理由については考察で述べる．

また，先行研究には含まれなかった複数の問診結果（歩行又は身体活動，睡眠，食べ方１，飲酒など）で有意な差が見られた．十分な睡眠をとる生活習慣や 1 日 1 時間以上の歩行又は身体活動は，3 つ全ての疾病でリスクを下げる効果があった．

考察

本分析では，野田らの先行研究との比較を試みた．野田らの研究では Cox 比例ハザードモデルにより 5 年時生存関数を推定し，相対危険度 RR を求めているのに対し，本研究ではロジスティック回帰により，3 年以内の罹患のオッズ比 OR を求めている．手法が異なる理由の一つは，本分析で使用したデータは，主に健康で生存している被験者が中心であり，死亡者のデータが少なかったためである．代替として 3 年以内の罹患を目的変数としたが，罹患は死亡と異なり，一度罹患した後に再度罹患することもあるので，Cox 比例ハザードモデルで生存関数を算出するのが適切でないため，ロジスティック回帰を使用した．本分析により算出した OR と Cox 比例ハザードモデルによる RR は厳密には一致しないが，共に重要なリスク比の指標として知られており [40]，対象の疾患による死亡には罹患が前提条件になるため，両者は概ね同じ傾向を示すと考える．

喫煙に関しては先行研究との整合性が得られなかった．その原因として，喫煙についての変化が考えられる．日本人の喫煙率が先行研究の対象期間である 2001 年には 46%（男性）だったのに対し，その 15 年後の本分析の 2016 年には 30%（男性）であった [25]．健康診断で非喫煙と回答している人の中に元喫煙者が含まれていると予想される．この様な喫煙に関わる環境の変化の結果，喫煙による影響の整合性が保たれなかったと考える．従って，匿名加工情報により，従来のコホート研究と同等の分析結果が得られたと結論づける．

匿名加工後の性質について

本匿名加工情報は，当該ヘルスケア企業により，表 6.9 に従って法的に適切な加工（法 [38] 第 36 条 1 項，規則第 19 条）が行われており，この加工によって生データの持つ性質が劣化している可能性がある．例えば前田ら [98] は，様々な性質を持つデータセットをセル削除による k–匿名化 [36] で加工することにより，加工で失われる有用性の度合いがデータセットの性質によって異なることを示している．

そこで，表 6.9 の加工が罹患リスクに及ぼす影響を検討する．まず，削除と仮 ID への置換は罹患リスクに影響を与えない．次に，健康診断受診日を月単位に丸める加工は 3 年以内という条件を変える小さな可能性があるが，無視できる確率である．従って，影響があるのは表 6.8 の 12 種の連続量のトップ/ボトムコーディングのみである．そこで，いくつかの仮定を置いて加工前の健康診断データを予測し，当該ヘルスケア企業に確認した後に同様の分析を行った．

表 6.8: ロジスティック回帰結果

		脳卒中			がん			インフルエンザ	
	特徴量	estimate	OR	RR[13]	estimate	OR	RR[13]	estimate	OR
連続値	const	−3.643∗[1]	0.026		−1.041∗[2]	0.353		−0.494	0.610
	年齢 (歳)	0.024∗[1]	1.024	1.14	0.017∗[1]	1.017	1.09	−0.036∗[1]	0.964
	身長 (cm)	−0.002	0.998		0.003∗[4]	1.003		0.003∗[4]	1.003
	Body Mass Index (kg/m^2)	−0.004	0.996	1.00	−0.015∗[1]	0.985	0.86	0.009∗[4]	1.009
	収縮期血圧 (mmHg)	0.002	1.002	1.02	−0.002∗[4]	0.998	–	−0.001	0.999
	拡張期血圧 (mmHg)	0.003∗[4]	1.003		−0.002∗[4]	0.998		−0.005∗[1]	0.995
	中性脂肪 (mg/dl)	0.000∗[2]	1.000		0.000∗[3]	1.000		0.000	1.000
	hdl コレステロール (mg/dl)	0.000	1.000	‡	−0.001∗[3]	0.999	0.85	0.000	1.000
	ldl コレステロール (mg/dl)	0.002∗[1]	1.002		−0.002∗[1]	0.998		−0.001∗[2]	0.999
	got ast (IU/L)	−0.001	0.999		0.004∗[1]	1.004		0.003∗[1]	1.003
	γ gtp (IU/L)	0.000	1.000		0.001∗[1]	1.001		0.000	1.000
	hba1c(ngsp)	0.068∗[3]	1.070		0.048∗[2]	1.049		0.072∗[1]	1.074
	赤血球数 ($*10^4/\mu l$)	−0.001∗[1]	0.999		−0.001∗[1]	0.999		0.000	1.000
2 値の カテゴリカル	性別	−0.212∗[1]	0.809		−0.530∗[1]	0.588		0.012	1.012
	服薬 1 血圧	0.383∗[1]	1.466	1.56	0.224∗[1]	1.252	1.15	0.127∗[1]	1.136
	服薬 2 血糖	0.249∗[1]	1.283		0.165∗[1]	1.180		−0.116∗[4]	0.890
	服薬 3 脂質	0.255∗[1]	1.291		0.124∗[1]	1.132		0.061∗[4]	1.063
	既往歴 1 脳血管	1.84∗[1]	6.294		−0.117	0.889		0.078	1.081
	既往歴 2 心血管	0.204∗[2]	1.227		−0.025	0.976		−0.079	0.924
	既往歴 3 腎不全・人工透析	0.129	1.138		0.533∗[1]	1.704		0.543∗[1]	1.721
	貧血	0.174∗[1]	1.190		0.227∗[1]	1.255		0.108∗[1]	1.114
	喫煙	0.017	**1.017**	**1.27**	−0.084∗[1]	**0.920**	**1.51**	0.062∗[2]	1.064
	体重変化 20 歳からの	0.052	1.053		0.011	1.011		0.071∗[2]	1.074
	運動習慣 30 分以上	−0.022	0.978		−0.06∗[2]	0.941		−0.016	0.984
	歩行又は身体活動	−0.093∗[2]	0.911		−0.109∗[1]	0.897		−0.076∗[1]	0.927
	歩行速度	−0.034	0.966		−0.1∗[1]	0.905		−0.036∗[4]	0.965
	体重変化 1 年間	0.129∗[1]	1.137		0.096∗[1]	1.101		0.109∗[1]	1.116
	食べ方 2 就寝前	0.075∗[3]	1.078		−0.006	0.994		0.068∗[1]	1.070
	食べ方 3 夜食間食	0.014	1.014		0.093∗[1]	1.098		0.049∗[4]	1.050
	食習慣	−0.001	0.999		−0.098∗[1]	0.907		−0.067∗[2]	0.935
	睡眠	−0.239∗[1]	**0.787**		−0.118∗[1]	0.889		−0.178∗[1]	0.837
	保健指導の希望	0.075∗[3]	1.078		0.010	1.010		−0.011	0.989
3 値以上の カテゴリカル	メタボリックシンドローム判定								
	メタボ予備軍	0	1.000		0	1.000		0	1.000
	メタボ該当者	−0.096	0.908		−0.085∗[4]	0.918		−0.094∗[4]	0.911
	非メタボ	−0.075	0.928		−0.108∗[3]	0.898		−0.013	0.987
	食べ方 1（早食い等）								
	普通	0	1.000		0	1.000		0	1.000
	速い	0.095∗[1]	1.100		0.086∗[1]	1.090		0.025	1.026
	遅い	0.023	1.023		0.041	1.042		−0.033	0.968
	飲酒								
	時々	0	1.000		0	1.000		0	1.000
	ほとんど飲まない	0.083∗[3]	1.086		0.044∗[3]	1.045		0.018	1.018
	飲酒毎日	−0.032	0.969		−0.032	0.969		0.015	1.015
	保健指導レベル								
	情報提供	0	1.000		0	1.000		0	1.000
	動機付け支援	0.077	1.080		0.046	1.047		0.048	1.049
	対象外	−0.005	0.995		−0.074∗[4]	0.929		−0.124∗[3]	0.884
	積極的支援	0.072	1.075		0.036	1.037		−0.062	0.940
	尿糖								
	＋	0	1.000		0	1.000		0	1.000
	＋＋	−0.186	0.830		−0.039	0.961		−0.244	0.784
	＋＋＋	−0.304∗[4]	0.738		−0.099	0.906		−0.04	0.961
	−	−0.298∗[3]	0.742		−0.046	0.955		0.010	1.010
	±	−0.345	0.708		−0.058	0.943		0.089	1.093
	尿蛋白								
	＋	0	1.000	−	0	1.000	1.44	0	1.000
	＋＋	0.057	1.059		0.063	1.065		−0.232∗[4]	0.793
	＋＋＋	−0.241	0.786		−0.130	0.878		−0.178	0.837
	−	−0.123	0.884		−0.117∗[3]	0.890		−0.065	0.937
	±	0.031	1.031		0.043	1.043		0.014	1.014
	生活習慣の改善								
	改善予定 (１か月以内)	0	1.000		0	1.000		0	1.000
	改善するつもりである (概ね 6 か月以内)	−0.097∗[3]	0.907		−0.024	0.976		0.020	1.021
	改善するつもりはない	−0.181∗[1]	0.834		−0.125∗[1]	0.882		−0.071∗[3]	0.931
	既に改善に取り組んでいる (6 ヶ月以上)	−0.053	0.948		−0.002	0.998		−0.022	0.978
	既に改善に取り組んでいる (6 ヶ月未満)	−0.031	0.970		−0.005	0.995		−0.001	0.999

∗[1]：$P < 0.0001$，∗[2]：$P < 0.001$，∗[3]：$P < 0.01$，∗[4]：$P < 0.05$

–：有意な関連が全く示されなかったため表示せず [13].

‡：分析モデルに含めなかったため表示せず.

表 6.9: 当該ヘルスケア企業による匿名加工情報への加工手法

No	19 条規則	加工方法	該当（健康診断, レセプトにおける）
1	特定の個人を識別出来る記述等の全部または一部を削除	規則性のない方法で生成された仮 ID に置換	氏名（健康診断, レセプト）
		削除	住所（レセプト）
2	個人識別符号の全部を削除	削除	被保険者記号（健康診断, レセプト） 被保険者番号（健康診断, レセプト）
3	個人情報と他の情報を連結する符号を削除	規則性のない方法で生成された仮 ID に置換	レセプト ID（レセプト）
4	特異な記述等を削除	トップ・ボトムコーディング	健康診断データ (連続量)（健康診断）
5	他の個人情報との差異等の性質を勘案した措置	一般化	傷病名コードを ICD10 コードに変換（レセプト） 医薬品名コードを ATC コードに変換（レセプト） 診療行為コードをコード表用番号に変換（レセプト）
		日単位→月単位に変換	健康診断受診日（健康診断）
		削除	医療機関名称（レセプト）

表 6.10: 匿名加工情報と予測健診データとの OR 絶対誤差の統計量

病名	平均値	標準偏差	最大値	最小値
脳卒中	$2.5 \cdot 10^{-4}$	$4.2 \cdot 10^{-4}$	$1.9 \cdot 10^{-3}$	$5.6 \cdot 10^{-7}$
がん	$2.1 \cdot 10^{-4}$	$2.7 \cdot 10^{-4}$	$1.2 \cdot 10^{-3}$	$3.5 \cdot 10^{-7}$
インフルエンザ	$2.0 \cdot 10^{-4}$	$2.7 \cdot 10^{-4}$	$1.1 \cdot 10^{-3}$	$4.6 \cdot 10^{-7}$

表 6.10 に，匿名加工情報と予測された健診データを用いて算出した OR の絶対誤差の統計量を示す．ほとんどすべての説明変数について，両者の OR は一致しており，例えば，脳卒中についての絶対誤差は，最大で $1.9 \cdot 10^{-3}$，平均で $2.5 \cdot 10^{-4}$ である．この結果より，匿名加工情報による罹患リスクの変化は無視でき，匿名加工情報を用いても，ヘルスケア情報の分析結果として有用性が認められるレベルの品質の結果が得られると判断する．

6.3.4 疾病罹患予測モデル (3)

分析手法

本研究では，罹患者が 1,000 人以上の 274 種類の傷病を分析対象とする．各傷病を目的変数 y_{log}，健康診断データを説明変数 x_{log} として分析を行う．被験者 i が傷病 A04 に罹患するかを健康診断データから予測するモデルは，$y_{log,A04} = model_{A04}(x_{log,i1}, x_{log,i2}, \ldots, x_{log,i38})$ で表される．ここで，$model_{A04}$ は本分析で作成する機械学習モデルを表す．

健康診断データの有用性指標として，3 年以内の罹患予測モデルを傷病 274 種類作成する．学習時には罹患者数と同数の非罹患者レコードをランダムサンプリングして用いる．予測アルゴリズムには K 近傍法 (KNN)，RBF Support Vector Machine（SVM），Decision Tree (Tree)，Random Forest（RF）を使用する．また，本分析では高い精度を出すことが目的ではなく，匿名加工によって機械学習の精度がどれだけ変化するかの分析が目的であるため，各モデルのハイパーパラメータは表 6.11 のデフォルト値 [3] を使用する．

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`,
`sklearn.tree.DecisionTreeClassifier.html`,
`sklearn.neighbors.KNeighborsClassifier.html`,
`sklearn.svm.SVC.html`

表 6.11: 予測モデルのハイパーパラメータ

学習方法	パラメータ名	デフォルト値
k 近傍法	n_neighbors	5
	weights	uniform
	algorithm	auto
	leaf_size	30
	p	2
	metric	minkowski
	C	1
	kernel	rbf
SVC	degree	3
	gamma	scale
	criterion	gini
	splitter	best
決定木	min_samples_split	2
	min_samples_leaf	1
	min_weight_fraction_leaf，min_impurity_decrease min_impurity_split，ccp_alpha，min_weight_fraction_leaf	0
	n_estimaters	100
	criterion	gini
	min_samples_split	2
	min_samples_leaf	1
	max_features	auto
ランダムフォレスト	min_impurity_decrease, verbose, ccp_alpha	0
	bootstrap	TRUE
	oob_score, warm_start	FALSE

各モデルの評価は 5 分割交差検証によって行い，有用性は再現率と適合率の調和平均である F 値の平均を使用する．モデルは python の scikit-learn[104] を用いて実装する．

分析結果

図 6.6 に 274 種類の疾病の罹患を予測した 4 種類のモデルの F 値の分布を示す．表 6.12 に，学習手法ごとの各 274 種類の傷病予測モデルの F 値の統計量を示す．ランダムフォレストの平均 F 値が最も高く 66%であった．一方で，他のモデルの平均 F 値は 57%で大きな差はなかった．SVM では標準偏差が他のモデルに比べて大きく 0.07 で，疾病により大きく精度が変化する．

図 6.6 から傷病の種類により精度の偏りがある．表 6.13 に分析で使用した中分類を大分類に再集計した統計量を示す．各学習手法の列は，中分類の平均 F 値を表す．表 6.13 から，新生物，代謝疾患，尿路性器系の疾患，妊娠，健康状態に影響をおよぼす要因等は他に比べて精度が高く，F 値が 0.7 であることがわかる．

表 6.14 にランダムフォレストの F 値の上位 10 件の疾病を示す．10 件中 9 件の傷病が女性特有の

デフォルト値だけでの実験は一般的でないので，今後の実験で最適なパラメータを検討する．

表 6.12: 各学習手法精度（F 値）の統計量

	Mean	SD	Max	Min
RF	0.659	0.062	0.909	0.588
Tree	0.579	0.059	0.852	0.524
SVM	0.578	0.071	0.831	0.462
KNN	0.562	0.058	0.825	0.510

表 6.13: ICD10 大分類での平均精度

大分類	傷病名	中分類数	RF	Tree	SVM	KNN
A00-B99	感染症および寄生虫症	15	0.642	0.563	0.557	0.551
C00-D48	新生物＜腫瘍＞	24	0.700	0.617	0.625	0.603
D50-D89	血液障害等	5	0.666	0.578	0.580	0.564
E00-E90	内分泌, 栄養および代謝疾患	15	0.711	0.624	0.628	0.595
F00-F99	精神および行動の障害	4	0.631	0.551	0.549	0.533
G00-G99	神経系の疾患	7	0.636	0.554	0.550	0.533
H00-H59	眼および付属器の疾患	16	0.652	0.570	0.589	0.552
H60-H95	耳および乳様突起の疾患	9	0.630	0.549	0.544	0.536
I00-I99	循環器系の疾患	18	0.673	0.587	0.589	0.562
J00-J99	呼吸器系の疾患	23	0.624	0.550	0.547	0.535
K00-K93	消化器系の疾患	34	0.631	0.554	0.547	0.543
L00-L99	皮膚および皮下組織の疾患	20	0.638	0.558	0.563	0.544
M00-M99	筋骨格系および結合組織の疾患	24	0.645	0.565	0.562	0.550
N00-N99	尿路性器系の疾患	23	0.746	0.669	0.673	0.648
O00-O99	妊娠, 分娩および産じょく	1	0.909	0.852	0.831	0.825
Q00-Q99	先天奇形, 変形および染色体異常	1	0.656	0.569	0.564	0.554
R00-R99	異常検査所見で他に分類されないもの	24	0.634	0.559	0.541	0.537
S00-T98	損傷, 中毒およびその他の外因の影響	10	0.624	0.549	0.535	0.538
Z00-Z99	健康状態に影響をおよぼす要因等	1	0.707	0.627	0.651	0.616

疾患であり，ランダムフォレスト以外のモデルでも精度が 70%以上だった．日本の老衰を除いた 3 大死亡原因 [29] であるがん，心疾患，脳血管疾患 (脳卒中も含む) に該当する傷病の予測精度を表 6.15 に示す．脳梗塞は少なくとも 65%の精度で予測可能である．

考察

表 6.14 の一部の精度の高い疾病には女性特有（N97 女性不妊症，O20 妊娠早期の出血等）の傷病が多く，F 値の上位 10 件中 9 件の傷病であった．原因として，必ず罹患していない患者（男性）を簡単に分類できて，結果的に F 値が高くなったと考えられる[4]．例えば，女性データのみで作成した N97 に関するランダムフォレストによるモデルでは，F 値は 0.80 で全データを使用したモデルから

[4]N97 のモデルでは，6.3.4 節の手順から性別属性を削除するが，他の属性（年齢，bmi 等）から性別の推定は容易である

表 6.14: F 値上位 10 件の疾病

ICD10	傷病名	サンプル数	RF	Tree	SVM	KNN
O20	妊娠早期の出血	2,844	0.909	0.852	0.831	0.825
N97	女性不妊症	2,374	0.889	0.826	0.794	0.778
E10	1 型糖尿病	2,000	0.869	0.786	0.676	0.611
N94	月経周期の疼痛等	3,322	0.847	0.753	0.780	0.747
E28	卵巣機能障害	11,204	0.844	0.770	0.784	0.746
N95	閉経期障害等	6,564	0.835	0.760	0.745	0.717
N80	子宮内膜症	4,066	0.830	0.746	0.757	0.730
D25	子宮平滑筋腫	14,814	0.828	0.755	0.765	0.725
N76	膣及び外陰のその他の炎症	11,608	0.827	0.757	0.774	0.738

表 6.15: 日本 3 大死亡原因の罹患予測精度

ICD10	傷病名	サンプル数	RF	Tree	SVM	KNN
C18	結腸がん	20,470	0.604	0.531	0.538	0.524
I20	狭心症	13,178	0.652	0.570	0.580	0.543
I63	脳梗塞	8,806	0.648	0.565	0.587	0.545

約 0.10 劣化した．また，傷病によって精度が異なる原因には，T14(部位不明の損傷) など健康診断とあまり関係がない傷病があるためと考える．

6.4 k–匿名化と分析結果への影響

6.4.1 概要

健康診断データやレセプトから得られる病歴データを，代表的な匿名化手法である k–匿名性を満たすように加工することにより，3 章で提案した有用性指標がどれだけ変化するのかを明らかにする．k–匿名性は Sweeny によって提案された匿名性の指標 [2] であり，同じ QI を持つ個体の少なくとも k 人が同じ値を持つようにデータを加工すればこれを満たすことができる．k–匿名性を満たすには，レコード等の削除や値の一般化，及びそれの組み合わせが用いられる．各々の例として，本章では表 6.16 の 2 種類の方法（レコード削除加工，Mondrian アルゴリズム）を評価する．なお，Mondrian アルゴリズムについては 2 章を参照されたし．

なお，匿名加工情報に追加の加工を加えても，適法な匿名加工情報とみなせるので，この章では当該ヘルスケア企業による匿名加工情報を「ヘルスケア企業による匿名加工情報」，それらを k–匿名性を満たすように加工したものを「追加匿名加工情報」と区別して呼ぶ．

表 6.16: k–匿名化手法の詳細

	本論文（レコード削除）	Mondrian[41]
方法	該当する人数が k 人未満の QI（年齢，性別）を持つ個人を削除する	QI（年齢，性別）をもとに分割された各グループの QI の値を 中央値に置き換える処理を，k–匿名性を満たすまで繰り返す
実装	python による独自実装	Nithin による python スクリプト [42]

表 6.17: レコード削除によって k–匿名化された追加匿名加工情報の精度

k	レコード数	削除割合	RF	Tree	SVM	KNN
0	203,521	0.0000	0.659	0.579	0.578	0.562
3	203,521	0.0000	0.659	0.579	0.578	0.562
5	203,521	0.0000	0.659	0.579	0.578	0.562
10	203,521	0.0000	0.659	0.579	0.578	0.562
30	203,474	0.0002	0.659	0.579	0.578	0.562
50	203,311	0.0010	0.659	0.579	0.578	0.562
100	202,807	0.0035	0.659	0.579	0.577	0.562
500	196,719	0.0334	0.658	0.578	0.576	0.561
1,000	181,981	0.1058	0.656	0.576	0.571	0.558
最大誤差	–	–	0.003	0.003	**0.007**	0.004
平均誤差	–	–	0.001	0.001	0.001	0.001

6.4.2 QI=性別と年齢

分析手法

本分析では，年齢と性別のそれぞれの値を QI として，ヘルスケア企業による匿名加工情報の健康診断データを表 6.16 に示す 2 種類のアルゴリズムにより k–匿名化する．それぞれの加工アルゴリズムによって $k = 3, 5, 10, 30, 50, 100, 500, 1\,000$ で加工した時の追加匿名加工情報に対して，6.3.4 節と同様の分析を行いモデルの精度を比較する．

表 6.17 に，レコード削除による追加匿名加工情報の k の値による，レコード数と予測精度の変化を示す．各学習手法の値は，274 種類の傷病を予測した際の F 値の平均を表す．$k = 3$ から 1,000 の匿名化を行うと，最大で 10%のレコードが削除され，274 種類の傷病の $k = 0$ の基準データに対する平均 F 値の最大誤差は SVM の 0.007 であった．

また，Mondrian アルゴリズムによる追加匿名加工情報の k の値による予測精度の変化を表 6.18 に示す．$k = 3$ から 1,000 の匿名化を行うと，274 種類の平均 F 値の最大誤差は RF の 0.025 であり，レコード削除による追加匿名加工情報よりも誤差が大きくなった．

表 6.18: Mondrian アルゴリズムによって k–匿名化された追加匿名加工情報の精度

k	RF	Tree	SVM	KNN
0	0.659	0.579	0.578	0.562
3	0.634	0.562	0.567	0.553
5	0.635	0.562	0.567	0.553
10	0.634	0.562	0.567	0.553
30	0.635	0.562	0.567	0.553
50	0.634	0.562	0.567	0.553
100	0.634	0.562	0.567	0.553
500	0.635	0.562	0.567	0.553
1,000	0.634	0.562	0.567	0.553
最大誤差	**0.025**	0.017	0.011	0.009
平均誤差	0.025	0.017	0.011	0.009

考察

レコード削除による追加匿名加工情報では，QI を年齢と性別にした時に，$k = 1,000$ でレコードの 10%を削除したが，機械学習の精度は最大でも SVM による 0.007 の劣化であった．この原因の一つとして，年齢以外の要素が罹患予測に作用していたことが考えられる．例えば，ランダムフォレストの高血圧症 (I10) の罹患推定における，特徴量重要度 [30] は年齢が 0.06 であるのに対して，収縮期血圧が 0.117，BMI が 0.05 と同等もしくはそれ以上であった．従って，年齢の属性無しでも BMI などの属性が精度を補償していると考える．

図 6.7 に健康診断データの年齢頻度と，各年齢における高血圧症の罹患者数を示す．レコード数の少ない 10 代から 20 代や 65 歳以上では，罹患者数が極めて少ないことがわかる．従って，k–匿名化をしてそれらのレコードが削除されても精度が最大でも 0.007 の劣化であったことも，理由の一つと考える．

また，Mondrian アルゴリズムによる追加匿名加工情報の誤差は，最大，平均共にレコード削除による追加匿名加工情報よりも大きかった．Mondrian アルゴリズムではデータの QI の値をグループの中央値に書き換える処理を行うため，値を書き換える処理をしないレコード削除加工よりも誤差が大きくなったこと考えられる．k–匿名化には，他にも [36] などの様々な方式が知られているが，上記の理由から，おおむね同様の精度に収まることが予測される．

従って，これらの結果より，レコード削除や一般化によって k–匿名化を行って匿名加工情報を作成したとしても，機械学習の精度に対して重大な影響を与えるほどの低下をしないと本章では結論づける．

6.4.3 QI=病歴/処方歴

分析した傷病/医薬品の相対リスクを追加匿名加工情報の有用性とみなし，これらを評価する．

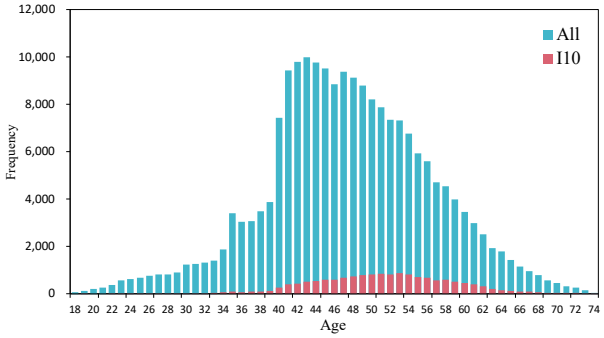


図 6.7: 健康診断患者全体（All）と高血圧症罹患者（I10）の年齢分布

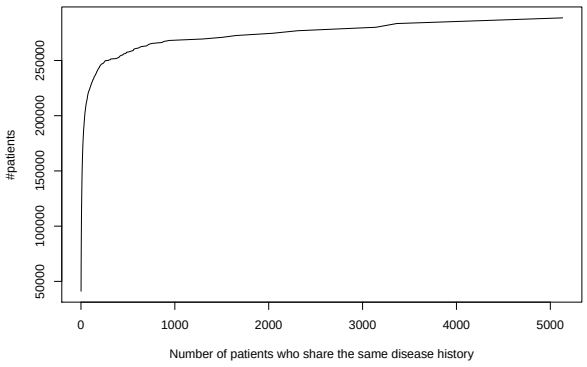


図 6.8: 各個人と同じ病歴を持つ個人数の累積分布

安全性：病歴/処方歴の一意性

傷病/医薬品レセプトデータから得られる病歴/処方歴の一意性に注目し，個人識別リスクを評価する．レセプトデータには 1 顧客についてのレセプトが複数枚分記録されている．それをまとめて各傷病/医薬品について 2 値のベクトル $\boldsymbol{dis} = (dis_1, \ldots, dis_\ell)$,

$$dis_i = \begin{cases} 1 & (i \text{ 番目の病歴/処方歴あり}) \\ 0 & (\text{なし}) \end{cases}$$

にし，これを各個人の病歴/処方歴ベクトルとする．

図 6.8 に，傷病レセプトデータの各個人と同じ病歴を持つ個人数の累積分布を示す．傷病レセプトデータの場合，最大で 5,131 人の個人が同じ病歴（傷病 K：消化器系の疾患のみに罹患したことがある）を持っており，一意な病歴を持っている個人は 41,099 人である．$n = 2.8 \cdot 10^5$, $\ell_{\text{病歴}} = 23$ のとき，一様ならば各病歴の発生確率は $p = 2^{-23}$ になるため，平均で $n \cdot p = 0.03$ 人の病歴が同じになる．しかし，本データでは平均 283 人の病歴が同じであるため，特定の病歴に著しく偏っていることがわかる．

加工による安全性/有用性の変化

病歴/処方歴は一意な値を持つ個人が多く，これらの人数を減らすために，データ削除による k–匿名化を検討する．

k–匿名化（$k = 1, \ldots, 10$）された病歴/処方歴からの識別率を図 6.9 に示す．ここでいう識別率は，元データを全て持っている最大知識攻撃者 [26] が k–匿名化された病歴から再識別するときの，(識別される人数の期待値）/（加工データに含まれる人数）とする．自分を含めて高々 k 人と同じ病歴/処方歴を持つ個人の数を n_k，病歴/処方歴に該当する個人数の最大値を n_{max} とすると，$\sum_{k=1}^{n_{max}} n_k/k$ で求めることができる．例えば $k = 1$ の病歴からは全体の 24.9%（71,864 人/288,568 人）の個人が識別されるが，$k = 10$ になるように該当人数が 10 人未満の病歴を持つ個人 132,736 人を削除すれば，識別される個人割合を 2.9%（4,563 人/155,832 人）まで減らすことができる．最大知識攻撃者は非常に

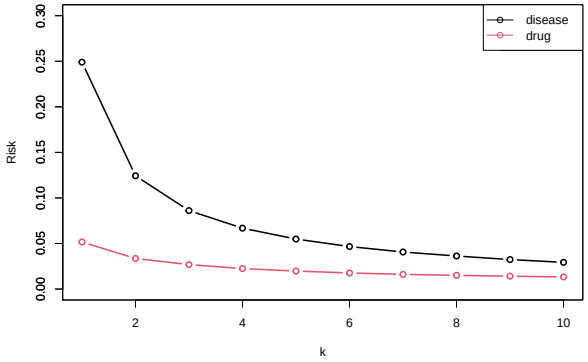


図 6.9: k–匿名化された病歴 (disease)/処方歴 (drug) の識別率

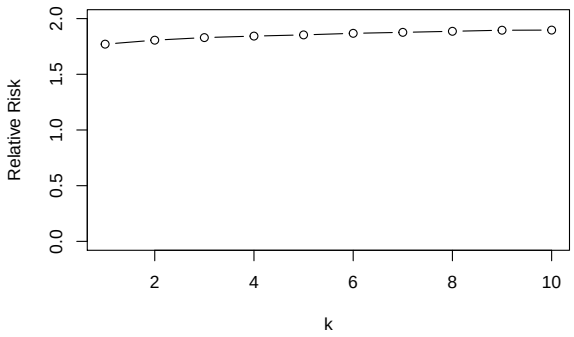


図 6.10: 病歴が k–匿名化されたときの相対リスクの変化

強い仮定であり，その仮定の下での識別率 2.9%は受容可能な範囲である [5] ．

次に，これらの追加匿名加工情報の有用性を評価する．病歴/処方歴は有用なデータであるため，それを評価する指標は数多く想定できるが，ここでは代表として，(1) 傷病/医薬品間の順位相関，(2) 高血圧を危険因子としたときの傷病 I の相対リスクの 2 点で有用性を評価する．

病歴/処方歴を k–匿名化した際に，各分類コード間の順位相関がどの程度変化するのかを，スピアマンの順位相関係数 ρ_{cor} で評価した結果を表 6.19 に示す．レコードを削除して k–匿名化をした場合でも，病歴・処方歴共に順位相関係 ρ_{cor} はあまり変化せず，10–匿名化をしても病歴の場合は 0.949，処方歴の場合は 0.996 までしか下がっておらず，データの有用性は失われていない．

また，高血圧を危険因子としたときの傷病 I の相対リスクが，病歴をレコード削除により k–匿名化した際にどのように変化するかを図 6.10 に示す．ヘルスケア企業による匿名加工情報の相対リスクが 1.77 であるのに対し，10–匿名化された追加匿名加工情報の相対リスクは 1.90 まで上がっている．これは相対誤差で $(1.90 - 1.77)/1.77 = 0.073$ であり，これは野田らの求めた，高血圧治療ありに対する脳卒中の相対リスクの 95%信頼区間の幅（0.71）より十分に小さいため，受容可能な精度である．これら 2 つの有用性評価結果より，追加匿名加工情報が有用であると結論付ける．

6.5 まとめ

本章では，あるヘルスケア企業が収集した 20 万人分の健康診断データと 28 万人分の傷病/医薬品レセプトデータを分析した．これらのデータはいずれも当該ヘルスケア企業によって適切に匿名加工されたものであるが，匿名加工情報を用いても，ヘルスケア情報の分析結果として有用性が認められるレベルの品質の結果が得られるかどうかを明らかにした．

相対リスクを用いて傷病/医薬品グループ間の違いを調査することによってデータの有用性を評価した．高血圧を危険因子としたときの傷病 I の相対リスクが 1.77 であることを明らかにした．飲酒をほとんどしない人が三年以内に脳卒中に罹患するリスクは，時々飲酒をする人に比べて 1.09 倍高くな

[5] 匿名加工コンテスト PWSCUP2016[37] では，攻撃者想定として最大知識攻撃者モデルが採用されているが，優勝チームの匿名加工データでさえ 22%の顧客が識別されている．

表 6.19: 病歴/処方歴が k–匿名化されたときの傷病/医薬品間の順位相関 ρ_{cor}

k	傷病レセプト	医薬品レセプト
1	1.000	1.000
2	0.996	0.999
3	0.989	0.998
4	0.982	0.998
5	0.976	0.998
6	0.969	0.997
7	0.962	0.997
8	0.958	0.997
9	0.953	0.997
10	0.949	0.996

ることや，十分な睡眠を取ることでリスクを 0.79 倍に下げるなどの新たな知見を得た健康診断データと傷病レセプトデータから 274 種類の傷病に 3 年以内に罹患するモデルをそれぞれ 4 種類の機械学習手法を用いて作成して評価した結果，ランダムフォレストが最も予測精度が良く，274 種類の傷病の平均 F 値は 0.65 であった．当該ヘルスケア企業による匿名加工情報の分析結果は，いくつかの仮定を置いて予測された加工前の健康診断データの結果と変わらず，予測健診データとの OR は 38 種の統計量で平均 $2.5 \cdot 10^{-4}$ の誤差しか生じないことを示した．

さらに，性別・年齢を QI として $k = 1,000$ までの k–匿名化を行い予測モデルの精度の変化を確認した．$k = 1,000$ の時レコード数は約 10%減少するが，加工しても十分に精度良いモデルが作れることを示した．病歴/処方歴を k–匿名化すると，識別される人数の割合は平均 2.9%まで減少するためデータの安全性を高めることができる一方で，相対リスクが相対誤差で 0.073 しか変化しないことを示した．

第7章　乗降と物販履歴データの識別リスク分析と匿名加工の検討

7.1　導入

本章では，私広く用いられている交通系IC カード「Suica[78]」のデータを用いた実験を行う．Suica の履歴データには電車の乗降履歴だけではなく，購買履歴やチャージ（入金）履歴等の他の用途（計5用途）の記録も残されており，用途の異なる履歴からなる珍しい形式のデータである．同一データ内に複数用途の履歴が記録されている複雑なデータのリスクの評価を行っている研究はこれまでない．そこで，情報エントロピーを用いて異なる用途間の相関関係をモデル化することにより，ある個人についての相互情報量やデータから個人が再識別されるリスクの定量化を行う．このリスクを合成データで測定し，購買履歴と移動履歴の相関関係を評価する．また，31 人の被験者から実際に IC カードデータを収集し，このデータのリスクを評価する実験を行い，安全性や有用性を評価する指標を提案してこのデータを評価する．

7.2　交通系 IC カード

本研究のために，明治大学総合数理学部に所属する学生 31 人の交通 IC カードから，顧客データ M と履歴データ T を作成した．交通 IC カードデータを収集する際には，対象者全員から研究利用への同意を得ている．なお，情報収集には Android のアプリケーション「IC カードリーダー by マネーフォワード [79] 」を使用した．一人あたりから収集できる履歴は最大 19 件である．表 7.1 にアプリケーションで取得できる乗降履歴データ T の例を示す．

表 7.2 に取得した本データの概要を示す．顧客データ M(マスターデータ) は 31 レコード 6 属性のデータであり，履歴データ T(トランザクションデータ) は 584 レコード 10 属性のデータである．表 7.3 に顧客データの例，表 7.4 に履歴データの例を各々示す．本来，交通 IC カードの利用履歴で得られる情報は「日付」,「利用内容」,「使用金額」の 3 属性のみであるが，本データでは「利用内容」属性を 6 属性に細分化している．例えば，表 7.1 の履歴をデータ化したものが表 7.4 であるが,「利用内容」属性を「乗車駅」,「降車駅」,「乗車路線」,「降車路線」,「用途」,「使用場所」の 6 属性に分けている.「用途」属性には IC カードの用途 (交通や物販等 5 種類) を示し,「使用場所」属性には IC カードを使用した場所 (券売機や自販機等 8 種類) を示している．

顧客データ M は IC カードから作成できないため，顧客本人から情報を取得し作成した．定期券の区間で乗り降りした履歴は取得できないため，顧客データ M に定期券の範囲を加えた．

表 7.1: 交通 IC カードデータの例

日付	詳細	料金 (円)
Oct. 30, 2016	in：上野 (JR 東日本) out：東京 (JR 東日本)	−194
Oct. 30, 2016	in：東京 (JR 東日本) out：上野 (JR 東日本)	−194
Oct. 8, 2016	券売機でのチャージ	2000
Oct. 1, 2016	自動販売機での購買	−150

7.3 再識別リスクの評価

7.3.1 エントロピーを用いた再識別リスク評価

データのリスク評価には様々な方法があるが，本節ではデータのエントロピー [bit/symbol] に注目してリスク評価を行う．表 7.5 のデータ例 E_S を用いて考え方を説明する．E_S は 3 人のユーザの計 19 回の駅利用履歴データについての集計表である．u_1, u_2, u_3 はユーザ，$s_1^{use}, s_2^{use}, s_3^{use}$ は駅名であり，例えば u_1 は s_1^{use} を 2 回，s_2^{use} を 1 回利用している．はじめに,「駅利用履歴が完全に不明である場合」のユーザのエントロピー $H(U_2)$ は，$P(U_2 = u_i)$ をデータ E_S 中で u_i の履歴の生起確率，n をユーザ数としたとき,

$$H(U_2) = -\sum_{i=1}^{n} P(U_2 = u_i) \log_2 P(U_2 = u_i)$$

で与えられる．この場合，全 19 履歴のうち，u_1 のものは 3 回であるため，$P(U_2 = u_1) = 3/19$, $P(U_2 = u_2) = 8/19$, $P(U_2 = u_3) = 8/19$ である．よって，$H(U_2) = 1.47$[bit/履歴] となる．

次に,「駅利用履歴が与えられた場合」のユーザの条件付きエントロピー $H(U_2|S)$ を考える．$P(S = s_i^{use})$ を E_S 中での駅 s_i^{use} の履歴の生起確率，ℓ を駅の種類数，$H(U_2|S = s_i^{use}) = -\sum_{j=1}^{n} P(U_2 = u_j|S = s_i^{use}) \log_2 P(U_2 = u_j|S = s_i^{use})$ を s_i^{use} の履歴が与えられた場合のユーザのエントロピーとしたとき,

$$H(U_2|S) = \sum_{i=1}^{\ell} P(S = s_i^{use}) H(U_2|S = s_i^{use})$$

で与えられる．E_S の場合，$H(U_2|S) = \frac{10}{19}1.52 + \frac{5}{19}0.72 = 0.99$ である．

最後に，相互情報量 $I(U_2; S)$ を求める．相互情報量とは 1 つの駅利用履歴から得られる情報量の期待値であり，$I(U_2; S) = H(U_2) - H(U_2|S)$ で与えられる．E_S の場合，$I(U_2; S) = 1.47 - 0.99 = 0.48$ である．

$H(U_2), H(U_2|S), I(U_2; S)$ の意味を考える．例えば駅利用履歴が完全に不明である場合，$H(U_2) = 1.47$ であり，u_1, u_2, u_3 の中のあるユーザを特定できる平均確率は $1/2^{H(U_2)} = 0.36$ である．しかし，1 つの駅利用履歴が判明した場合，例えば，s_3^{use} が分かると一意に u_2 であることが特定されるが，s_2^{use} ならば u_1 か u_3 らしいことしか分からない．平均すると $H(U_2|S) = 0.99$ になり，その平均確率は $1/2^{H(U_2|S)} = 0.5$ である．このとき 1 つの駅利用履歴から得た情報量は $I(U_2; S) = 0.48$_bit であるた

表 7.2: 交通 IC カードデータの詳細

Class	Quantity	Attribute	Detail
user data M	n 31	user ID	2 digit number
		sex	M/F
		grade	1 digit number
		address	place
		range of season ticket 1	place
		range of season ticket 2	place
history data T	m 584	user ID	2 digit number
		date	yyyy/mm/dd
		times	value
		entraining point	name of station
		alighting point	name of station
		entraining route	name of route
		alighting route	name of route
		usage	category
		location of use	category
		fare	value

表 7.3: 個人データ M の例

個人 ID	性別	学年	住所	定期券の範囲 1	定期券の範囲 2
1	M	1	千葉	NA	NA
2	F	3	東京	中野	新宿

め，$H(U_2) = 1.47 < 1.92 = 4I(U_2; S)$ より，4 つの駅利用履歴が判明した場合，u_1, u_2, u_3 の中の全てのユーザを特定できる確率はほぼ 1 になる．

7.3.2 交通 IC カードデータのエントロピー

ユーザの数は $n = 31$ 人，利用駅 (S) の種類は $\ell_S = 138$ 種，物販料金 (B) の種類は $\ell_B = 58$ 種，チャージ (C) 料金の種類は $\ell_C = 17$ 種である．なお，物販は簡単のため，料金の種類だけ商品の種類があると仮定する．

表 7.6 に用途別のエントロピー等の値を示し，X^{use} は特定の用途を示す．X^{use} =「交通」用途の場合，$H(U_2) = 4.900$，$I(U_2; S) = 3.085$ より，不明な値のある履歴からユーザが識別できる平均確率は $1/2^{H(U_2)} = 0.033$ である．1 つの履歴レコードには，交通，物販，チャージのどれかひとつしか記録されていない．従って，1 つの履歴が判明した場合には，その確率は $1/2^{H(U_2)-H(U_2|S)} = 0.284$ まで上がる．ユニークユーザ数 $n_{X^{use}}$ は，交通 IC カードを用途 X^{use} で利用しているユーザの数であ

表 7.4: 履歴データ T の例

User ID	Date	Times	Ent. point	Ali. point	Ent. route	Ali. route	Usage	Location	Fare
1	Oct 30 2016	2	Ueno	Tokyo	JR-EAST	JR-EAST	traffic	NA	-194
1	Oct 30 2016	1	Tokyo	Ueno	JR-EAST	JR-EAST	traffic	NA	-194
1	Oct 8 2016	1	NA	NA	NA	NA	deposit	ticket vending machine	2000
1	Oct 1 2016	1	NA	NA	NA	NA	purchase	vending machine	-150

表 7.5: 利用駅の集計表例 E_S

ユーザ \ 駅	s_1^{use}	s_2^{use}	s_3^{use}	合計	$P(U_2 = u_i)$
u_1	2	1	0	3	3/19
u_2	4	0	4	8	8/19
u_3	4	4	0	8	8/19
$H(U_2\|S = s_i^{use})$	1.52	0.72	0		
$P(S = s_i^{use})$	10/19	5/19	4/19		

る．例えば 31 人のユーザ中，交通 IC カードを「交通」用途で利用しているユーザは 31 人であるが,「物販」用途で利用しているのは 25 人である．また，$P(U_2|X^{use})$ は用途 X^{use} の履歴を取得した場合に，データ中のあるユーザが識別される平均確率である．

7.3.3 用途間の相関

本節では，交通 IC カードデータのリスク分析をするため，用途間の関係について分析を行う．図 7.1 に交通利用回数とチャージ料金の関係を示し，図 7.2 に交通利用料金とチャージ料金の関係を散布図で示す．相関係数は各々 0.469，0.315 であり，チャージ料金と交通利用回数・料金の間には弱い相関があることがわかる．よって，チャージ履歴の情報から交通履歴の情報を予測されるリスクがある．

交通履歴と物販履歴の関係を考える．表 7.7 にユーザごとの物販料金の例 E_B を示す．E_S の u_1, u_2, u_3

表 7.6: 各用途のエントロピー

	利用駅 (S)	購買 (B)	チャージ (C)	購買と乗降 (S, B)
$H(U_2)$	4.900	4.338	4.736	4.412
$H(U_2\|X^{use})$	1.814	0.948	3.256	0.182
$I(U_2; X^{use})$	3.085	3.389	1.479	4.230
$P(U_2\|X^{use})$	0.284	0.518	0.105	0.881
$n_{X^{use}}$	31	25	29	31
$\ell_{X^{use}}$	138	58	17	8004

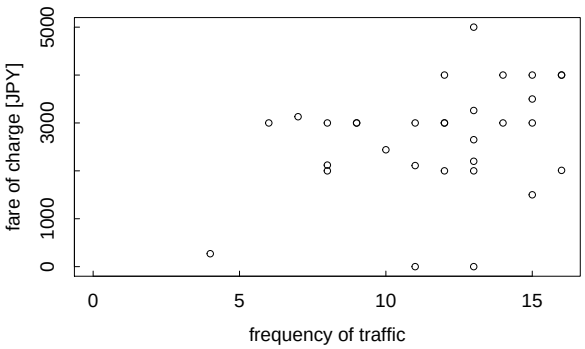


図 7.1: 交通利用回数とチャージ料金の散布図

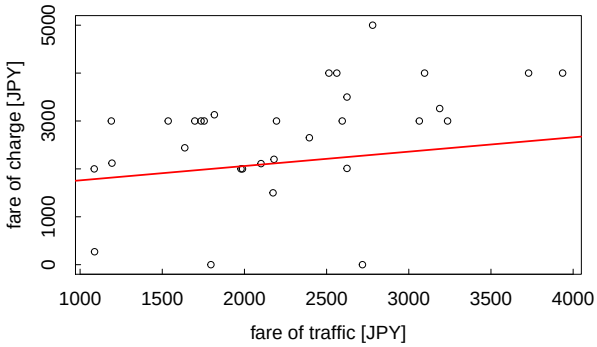


図 7.2: 交通利用料金とチャージ料金の散布図

表 7.7: 購買商品についての集計表例 E_B

ユーザ \ 料金	b_1^{use}	b_2^{use}	Sum	$P(U_2 = u_i)$
u_1	2	0	2	2/7
u_2	1	3	4	4/7
u_3	0	1	1	1/7
$H(U_2\|B = b_i^{use})$	0.92	0.31		
$P(B = b_i^{use})$	3/7	4/7		

と E_B の u_1, u_2, u_3 は同じユーザである．前述した例と同様の手順で計算すると，$H(U_2) = 0.98$, $H(U_2|B) = 0.57$，$I(U_2; B) = 0.41$ が与えられる．

また，E_S と E_B から 1 つずつ履歴を取得することを仮定した集計表 $E_{S,B}$ を表 7.8 に示す．この場合，例として，

$$P(u_1|s_1^{use}, b_1^{use}) = \frac{P(u_1|s_1^{use})P(u_1|b_1^{use})}{\sum_{i=1}^{n} P(u_i|s_1^{use})P(u_i|b_1^{use})} = \frac{4}{4+4} = \frac{1}{2}$$

と表すことができ，$H(U_2) = 1.19$, $H(U_2|S, B) = 0.46$, $I(U_2; S, B) = 0.73$ が与えられる．E_S, E_B, $E_{S,B}$ の各値を表 7.9 に示す．$I(U_2; X^{use})$ の行より，$I(U_2; S) + I(U_2; B) = 0.89 > 0.73 = I(U_2; S, B)$ であることから，交通と物販は独立ではないことがわかる．

交通 IC カードから取得した T の交通・物販用途を組み合わせた場合のエントロピー等の値を表 7.6 に示す．$I(U_2; S)$, $I(U_2; B)$ 等の結果と比較すると，$I(U_2; S) + I(U_2; B) = 6.474 > 4.230 = I(U_2; S, B)$ となり，例と同様のことが言える．$\ell = \ell_S \cdot \ell_B = 8004$ は交通 (138 種類) と物販 (58 種類) の組み合わせの数である．

7.4 評価

前述の例をもとに，適切な匿名加工を検討するために，いくつかの評価指標を定義する．加工データから料金や駅名などの属性と元データを比較し，仮名化された加工データの顧客を推測して算出された再識別率を安全性評価指標とする．また，特定の属性に注目し，平均絶対誤差を用いて元データ

表 7.8: E_S と E_B の履歴を組み合わせた集計表

	s_1^{use}, b_1^{use}	s_1^{use}, b_2^{use}	s_2^{use}, b_1^{use}	s_2^{use}, b_2^{use}	s_3^{use}, b_1^{use}	s_3^{use}, b_2^{use}	Sum	$P(U_2 = u_i)$
u_1	4	0	2	0	0	0	6	6/46
u_2	4	12	0	0	4	12	32	32/46
u_3	0	4	0	4	0	0	8	8/46
$H(U_2\|S = s_i^{use}, B = b_j^{use})$	1	0.81	0	0	0	0		
$P(S = s_i^{use}, B = b_j^{use})$	8/46	16/46	2/48	4/46	4/46	12/46		

表 7.9: E_S,E_B,$E_{S,B}$ の値

$\backslash X^{use}$	s^{use}	b^{use}	s^{use}, b^{use}
$H(U_2)$	1.47	0.98	1.19
$H(U_2\|X^{use})$	0.99	0.57	0.46
$I(U_2; X^{use})$	0.48	0.41	0.73
$P(U_2\|X^{use})$	0.50	0.67	0.73

表 7.10: 交通 IC カードデータの用途内訳

Usage	No. of records	Rate[%]
Traffic	364	62.3
Purchase	100	17.1
Deposit	84	14.4
Bus charge	2	0.3
Others	34	5.8
Sum	584	100.0

と加工データの差異を測定した値を有用性評価指標の評価値とする．評価指標の実装は python[105] と R 言語 [8] で行った．

7.4.1 交通 IC カードデータの分析

基本統計量

図 7.3 に月日ごとの使用料金の変化を示す．情報を収集したのが 2016 年 6 月であることと，収集できる履歴が直近 19 件までであることから，4,5,6 月の使用料金が多くなっている．また，図 7.4 にユーザごとの総使用料金を示す．直近 20 件の使用料金の最大値は 4,633 円であり，最小値は 2,393 円であった．

私は，交通 IC カードから取得できる履歴は交通のものだけではなく，用途が物販やチャージ等の履歴も含まれることに注目した．表 7.10 に取得した履歴の用途別の割合を示し，図 7.5 にユーザ別の用途の割合を示す．全体の 62.3%が交通用途の履歴であったが，交通 IC カードを交通より物販に多く使うユーザもおり，ユーザごとの多様性が高い．

図 7.6 に異なるユーザ間の利用駅についての類似度を表す Jaccard 距離の分布を示す．本データの全ての 2 組の平均 Jaccard 距離は 0.933 であり，本データのユーザは利用駅について，ほとんど似ていないことがわかる．

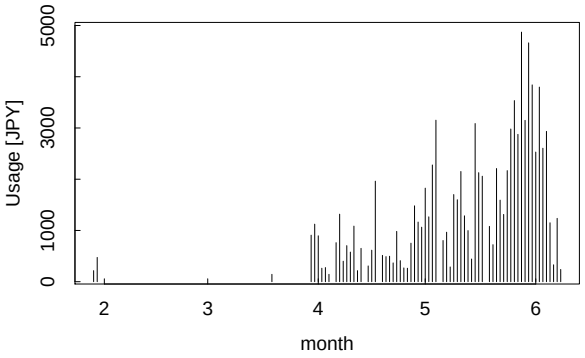


図 7.3: 月日ごとの使用料金の変化

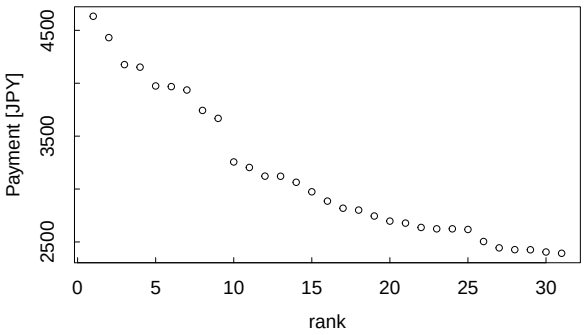


図 7.4: ユーザごとの総使用料金

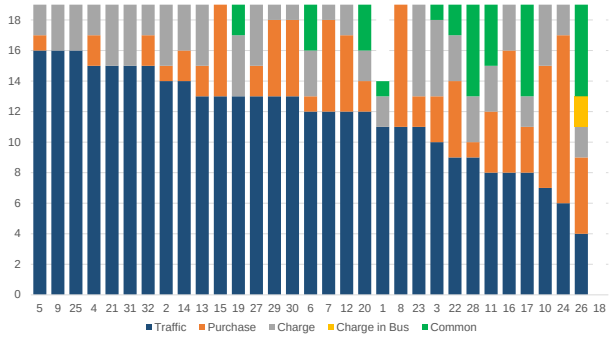


図 7.5: ユーザごとの用途内訳

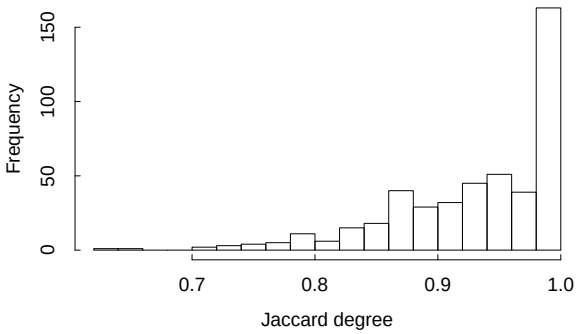


図 7.6: 異なるユーザ間の利用駅についての Jaccard 距離の分布

各値の出現頻度

取得した履歴データの用途は 5 種類あるが，そのうち「交通」,「物販」,「チャージ」用途の履歴が全体の約 94%を占めている．本節では，それらの 3 用途の出現頻度について分析を行う．

図 7.7, 7.8, 7.9 に「交通」,「物販」,「チャージ」用途の履歴の出現頻度を示し，表 7.11 に各用途ごとの特異なデータ（記録頻度が 2 回以下）の割合を示す．例えば「交通」用途の履歴は 364 レコードあり，その中に駅は 138 種類，727 回生起する．そのうち利用回数 1 回の駅が総利用回数の 4.8%（35 回）であり，利用回数 2 回以下の駅が 16.6%（78 回）であった．これらの特異なデータはユーザ特定の原因になりやすいため，匿名加工の際に削除などの対処をする必要がある．

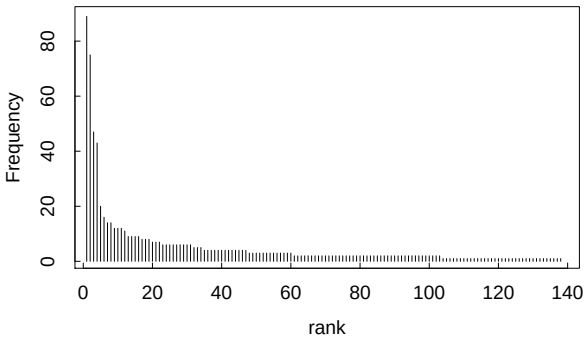


図 7.7: 交通用途の履歴の出現頻度

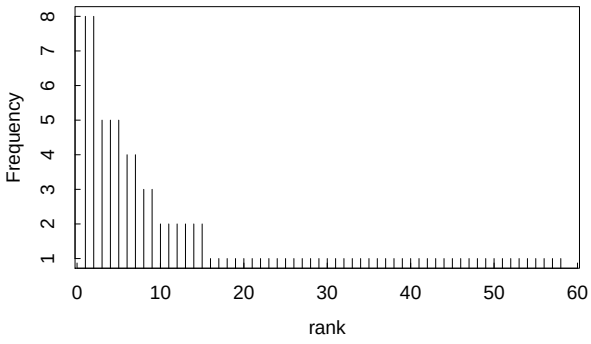


図 7.8: 物販用途の履歴の出現頻度

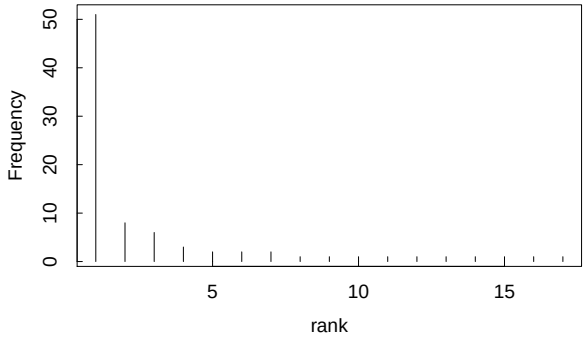


図 7.9: チャージ用途の履歴の出現頻度

表 7.11: 各用途ごとの特異なデータの割合

回数/用途	交通	購買	チャージ
1	4.8%	43.0%	11.9%
2	16.6%	55.0%	19.0%

7.4.2 匿名化手法

交通 IC カードデータを 2 つの手法で匿名化した．1 つ目の匿名化手法は，元データの SA(機微属性) の値にランダムなノイズを加える手法 (摂動化) である．例えば，表 7.4 に示す T を加工するときは，日付や料金などの属性の値にノイズを加えて別の値に変える．2 番目の手法は元データの SA の値を平均値で置き換える手法（ミクロアグリゲーション）である．例えば T を加工する場合，各レコードの料金の値 $(-194, -194, 2000, -150)$ を平均値 (365.5) で置き換える．

7.4.3 安全性指標

再識別リスクを評価する 6 つの安全性指標を用いて加工データの評価を行う．加工データから元データに対して，顧客ごとの料金合計値が近いものを識別する指標を $S2$, 顧客ごとに各用途 (交通, 物販，チャージ，バスチャージ，共通) のレコード数の比率が近いものを識別する指標を $S3$，乗降履歴のみに着目し, 各値の一致率から識別する指標を $S5$ とした．識別するにあたり，着目する属性に偏りが生じないように安全性指標を用意したが，複数の指標で同じ属性が用いられている場合もあることに注意せよ．再識別率を用いる安全性指標 $S2, S3$ は次のように定める．

安全性指標 $S2$

1. 入力：元データ T，加工データ T'，T のユーザ数 n，T' のユーザ数 n'
2. T の各ユーザ t_i の合計利用料金 $sum(t_i)$ を求め，全ユーザについての $sum(t_1), ..., sum(t_n)$ を求める．
3. T' の顧客 t'_1 の合計利用料金 $sum(t'_1)$ を求める．

4. $sum(t'_1)$ と $sum(t_1), ..., sum(t_n)$ を比較することにより，t'_1 と最も近いユーザ t_i を探す．もし候補が複数人いる場合，その中からランダムに選ぶ．

5. 手順 3,4 を $t'_2, ..., t'_{n'}$ についても行い，データについての再識別率を求める．

安全性指標 $S3$

1. 入力：元データ T，加工データ T'，T のユーザ数 n，T' のユーザ数 n'

2. T から各顧客 t_i の各用途のレコード数 $use(t_1), ..., use(t_n)$ を求める．本データの場合，$use(t_i)$ には 5 つの用途のレコード数を 5 次元のベクトルとして記録する．

3. T' の顧客 t'_1 の各用途のレコード数 $use(t'_1)$ を求める．

4. $use(t'_1)$ と $use(t_1), ..., use(t_n)$ を比較することにより，t'_1 と最も近いユーザ t_i を探す．もし候補が複数人いる場合，その中からランダムに選ぶ．

5. 手順 3,4 を $t'_2, ..., t'_{n'}$ についても行い，データについての再識別率を求める．

7.4.4 有用性指標

有用性を評価する 12 個の指標を用いて加工データの評価を行う．代表的なものとして，日付ごとの平均料金の平均絶対誤差を用いる指標 $U1$，駅の頻度の平均絶対誤差を用いる指標 $U6$，用途ごとの合計金額の平均絶対誤差を用いる指標 $U10$ がある．用意した有用性指標の多く (12 個中 5 個) が「料金」属性に注目するものであり，評価する際に用いる属性に偏りが生じていることに注意せよ．例として，有用性指標 $U1$ と $U10$ の説明を行う．

有用性指標 $U1$

1. 入力：元データ T，加工データ T'，T のユーザ数 n，T' のユーザ数 n'，T から T' へのマッピング p

2. T のユーザ $t_1, ..., t_n$ の料金合計 $sum(t_i)$ が記録された n 次元ベクトル $(sum(t_1), sum(t_2), ..., sum(t_n))$ を，p をもとに作成する．

3. ユーザ t' の料金合計ベクトル $sum(t')$ を作成する．$sum(t')$ は n 次元ベクトル $(sum(p(t'_1)), sum(p(t'_2)), ..., sum(p(t'_n)))$ である．

4. $sum(t)$ と $sum(t')$ の平均絶対誤差を計算する．もし $n \neq n'$ であるならば，小さいほうのベクトルにダミーを追加して次元数を揃える．例えば $n > n'$ の場合，$sum(t'_i)$ に $n - n'$ 個の 0 を追加してから $sum(t)$ と $sum(t')$ の間の誤差を求める．

有用性指標 $U10$

1. 入力：元データ T，加工データ T'
2. T の各用途の出現頻度を求めてベクトル $use(T)$ とする．この場合，$use(T)$ は用途数と等しい 5 次元のベクトルである．
3. T' の各用途の出現頻度を求めてベクトル $use(T')$ とする．
4. $use(T)$ と $use(T')$ の間の平均絶対誤差を求めて評価値とする．

7.4.5 評価実験

評価プラットフォーム

本節では，実際に交通 IC カードの履歴を加工し，前述した評価指標を用いて有用性と安全性について考察する．そこで, 加工データの評価を円滑に進めることを目的に, Linux 上に Web ベースの独自のプラットフォームを構築した．評価プラットフォームのシステム構成を図 7.10 に示す.

利用者は, 有用性と安全性を評価するスクリプト及び交通 IC カードの履歴を匿名加工したデータをアップロードする．評価プラットフォームは, 提出された評価指標や匿名加工データを集約し, 評価値をランキングとして出力する. 加工データに対する評価結果は SQL データベースに蓄積され, 評価の分析に活用する. 図 7.11 は, 評価プラットフォームのファイルのアップロード画面である．ドラッグ/ドロップで直感的な操作で利用することができる. 評価結果は, 図 7.12 のように, 各匿名加工データにおける評価値が出力される. また, 本プラットフォームは, 交通系 IC カード以外のデータに対しても, 評価を行えるように汎用的に設計している.

実験結果

評価実験では 47 個の匿名化データを評価した．図 7.13 に提出された匿名加工データの安全性/有用性の順位の散布図を示す．縦軸が有用性順位，横軸が安全性順位，図中の数字はデータの総合順位を示している．図中の線は元データ M, T と同じ評価になる境界線である．この線より下に位置するデータは元データより総合評価が高く，上に位置するデータは総合評価が元データより低い．この場合，1–8 位のデータは元データより総合評価が高く，9 位のデータは元データと同じ総合評価であり，それ以外のデータは元データより総合評価が低い．提出された 47 データ中 34 データが元データ以下と評価されている．上位 3 位のデータは id 属性のみをランダムに置き換えてかく乱する加工手法 [45] されたデータであり，4–8 位のデータは id 属性をスワップする手法によって加工されたものであった.

図 7.14 に加工データの安全性評価値分布を示し，図 7.15 に各データの最大値（最も有効だった指標の値）を示す．47 データ中 44 個の加工データでは 45%以上のユーザが再識別されており，購買履歴データと乗降履歴データの加工データから個人が再識別されるリスクは高いといえる.

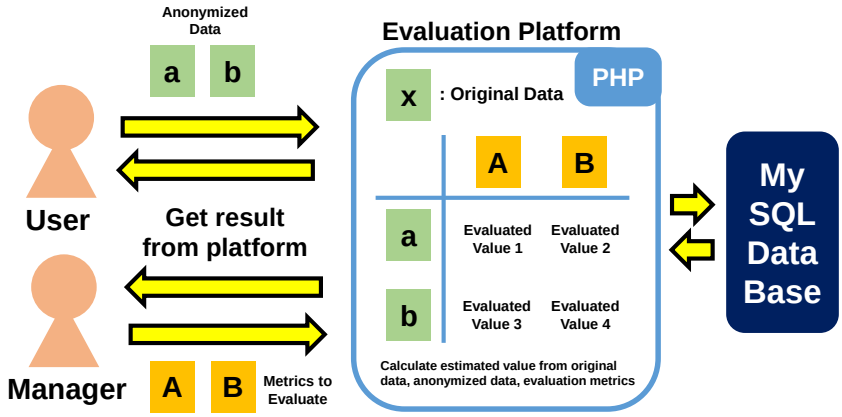


図 7.10: 評価システム概要

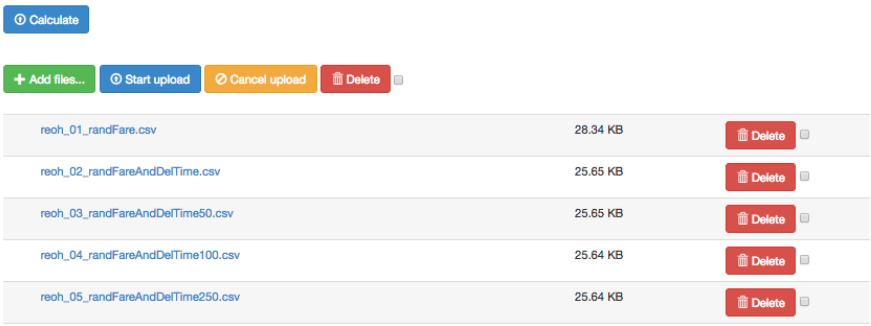


図 7.11: 評価プラットフォームのアップロード画面

図 7.16 に，利用料金属性を用いている指標 $S2$ と $U1$ の評価値の関係を示す．Data 1 は元データの値であり，$U1$ の評価値が 0（加工データと元データ間の誤差が 0）で $S2$ の評価値が 1（全てのユーザが正しく識別される）である．データが加工されて利用料金属性の値が変わるほど，有用性が失われるため $U1$ の評価値が上がり，安全性が上がるため $S2$ の評価値が下がっている．

7.5 まとめ

31 人の交通 IC カードから顧客データと履歴データを取得し，それらのデータのリスク評価をエントロピー等の値を用いて行った．その結果，用途「交通」,「物販」の履歴を 1 つ取得した場合，個人が特定されるリスクが大きく上がることや,「交通」と「チャージ」,「交通」と「物販」用途の間に相関があることが判明した．また，ユースケースを想定し，それに対応する評価指標と加工手法を考えた．

Performance

Calc time:2016-11-28 21:46:48

Search:

id	safety	user	data	satoshi_U17_jaccard_location_R	satoshi_U16_jaccard_use_R	satoshi_U15_jaccard_rc
2932	1	okamoto	10_copyAndPlusRandNum.csv	0	0	0
2933	0.998288	okamoto	11_copyAndDeleteDigit.csv	0	0	0
2934	0.903226	okamoto	13_changeUseAndPlusRandNum.csv	0.0215993	0.131266	0.00201759
2935	0.833046	okamoto	14_changeUseAndDeleteDigit.csv	0.0215993	0.131266	0.00201759
2936	1	okamoto	16_noChange.csv	0	0	0
2937	1	okamoto	1_copy.csv	0	0	0
2938	1	okamoto	2_plusRandNum.csv	0	0	0
2939	0.998288	okamoto	3_deleteDigit.csv	0	0	0
2940	0.967742	okamoto	4_changeUse.csv	0.0215993	0.131266	0.00201759
2941	1	okamoto	8_unityIn.csv	0	0	0.186263
2942	1	okamoto	9_randUser.csv	0.0055156	0.00506777	0.00162014
2943	1	reoh	reoh_01_randFare.csv	0	0.000432254	0.00196908
2944	1	reoh	reoh_02_randFareAndDelTime.csv	0	0.000432254	0.00196908
2945	0.806452	reoh	reoh_03_randFareAndDelTime50.csv	0	0.000432254	0.00196908
2946	0.806452	reoh	reoh_04_randFareAndDelTime100.csv	0	0.000432254	0.00196908
2947	0.806452	reoh	reoh_05_randFareAndDelTime250.csv	0	0.000432254	0.00196908

図 7.12: 評価プラットフォームの評価結果画面

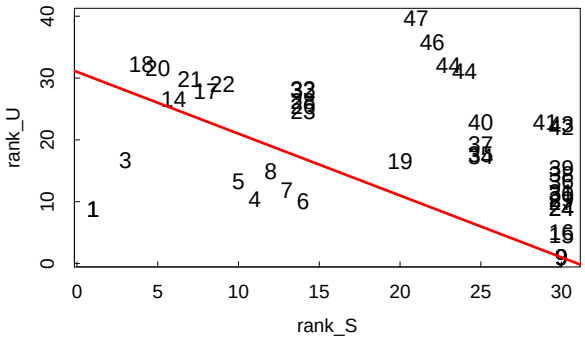


図 7.13: 評価実験の順位分布

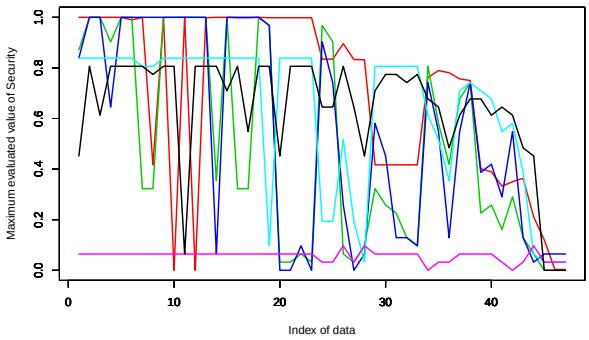


図 7.14: 安全性評価分布

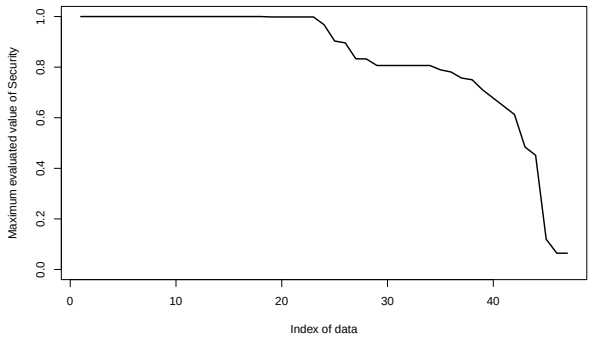


図 7.15: 安全性評価の最大値分布

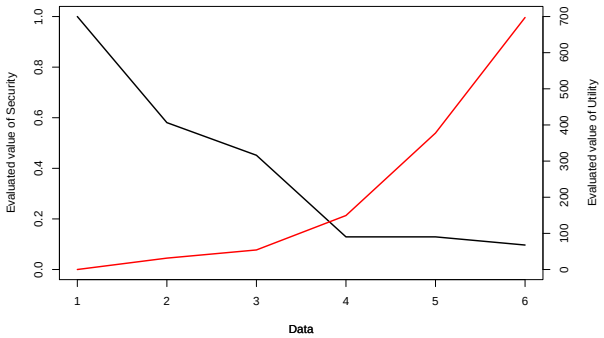


図 7.16: $S2$ と $U1$ 間の関係

第8章　ユークリッド距離を用いた再識別手法と世帯収入データの匿名化と評価

8.1　導入

本章では，レコード間のユークリッド距離による再識別リスクを想定し，その耐性を上げるための加工による効果を実験的に評価する．PWSCUP 2015[45] に提出された匿名化データを用いて匿名化の評価を行う．しかしながら，これらのデータの加工手法は不明であり，さらにいくつもの加工手法を組み合わせて作成されたことが予想される．そこで，まず様々な匿名化手法単体を用いた加工データを作成してその性能を評価し，大会に提出された匿名加工データを定量的に分析し，加工手法を推定する．匿名化手法の総合的性能（安全性と有用性）を評価するために，疑似ミクロデータ [60] を用いる．

8.2　有用性と安全性

8.2.1　疑似ミクロデータ

PWSCUP2015 では，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を加工の対象として用いている．疑似ミクロデータは，8,333 レコードと 25 属性から成る，2004 年の日本の年間世帯支出データである．第 1–13 属性には，家族の数や年齢といった離散値が記録されており，第 14–25 属性には食費や医療費といった連続値が記録されている．第 1–13 属性を準識別子（QI）として扱い，第 14–25 属性を機微属性（SA）として扱う．

8.2.2　有用性と安全性

PWSCUP2015 では，匿名化されたデータの性能評価のために，多くの有用性/安全性評価指標 [45] が作成されている．表 8.1 に各評価指標 (有用性: $U1$,...,$U6$, 安全性: $S1$,$S2$,$E1$,...,$E4$) の詳細を示す．これらのうち $E1$,...,$E4$ は既存再識別手法を用いた安全性評価指標であり，これらの詳細は次節にて説明する．

8.2.3　既存再識別手法

ユークリッド距離による再識別手法との比較対象として用いる，大会で用いられた 4 つの既存再識別手法の説明を行う．例として，元データ M と加工データ M'_B を表 8.2 と 8.3 に示す．M'_B は M に単純な加工を施したものであり，3 つの QI 属性と 2 つの SA 属性を持つ 4 レコードのデータである．複

表 8.1: 有用性/安全性指標の詳細

	Name	Detail	Target
$U1$	meanMAE	元データと加工データの SA 間の平均絶対誤差	SA
$U2$	crossMean	元データと加工データのクロス集計値間の平均絶対誤差	QI,SA
$U3$	crossCnt	元データと加工データのクロス集計数間の平均絶対誤差	QI,SA
$U4$	corMAE	元データと加工データの相関係数間の平均絶対誤差	SA
$U5$	IL	元データと加工データの各値間の平均絶対誤差	SA
$U6$	nrow	元データと加工データのレコード数の差	rows
$S1$	k-anony	k-匿名性指標の最小値	QI
$S2$	k-anonyMean	k-匿名性指標の平均値	QI
$E1$	identify-rand	QI が等しい個人からランダムに再識別する手法	QI
$E2$	identify-sa	QI が等しい個人のうち，第 15 属性目が近い個人を再識別する手法	QI,SA
$E3$	identify-sort	SA の値をソートすることによって再識別を行う手法	SA
$E4$	identify-sa21	第 21 属性のみの距離で個人を再識別する手法	SA

表 8.2: データ M

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	2	300	200
1	1	2	400	500

表 8.3: サンプル加工データ M'_B

QI1	QI2	QI3	SA1	SA2
2	1	1	110	90
2	1	1	220	390
1	1	2	280	210
1	1	2	390	520

数 QI/SA 属性の値を複数次元のベクトルとみなす．例えば，M の 1 レコード目は QI ベクトル (2,1,1) と SA ベクトル (100,100) からなる．

Identify-rand($E1$)

Identify-rand($E1$) 手法は，M'_B の攻撃対象と同じ QI ベクトルを持つ個人を M から探し，それらの候補の中からランダムに再識別を行う手法である．例えば，M'_B の第 1 レコードと同じ QI ベクトルを持つ個人は M の第 1,2 レコードであるため，これら 2 人からどちらかをランダムに識別結果とする．

Identify-sa($E2$)

Identify-sa($E2$) 手法は，M'_B の攻撃対象と同じ QI ベクトルを持つ個人を M から探し，それらの候補の中から特定の SA が最も近い個人を再識別結果とする手法である．例えば，M'_B の第 1 レコードと同じ QI ベクトルを持つ個人は M の第 1,2 レコードであり，そのうち SA1 が最も近い個人は第 1 レコードであるため，この個人を識別結果とする．

Identify-sort($E3$)

Identify-sort($E3$) 手法は，SA の和で M と M'_B のレコードを昇順にソートし，その順位で対応するレコードを再識別する手法である．例えば，SA の和 (SA1+SA2) のソート結果を用いる場合，M を昇順でソートすると第 1(200)，第 3(500)，第 2(600)，第 4 レコード (900) の順になり，M'_B を昇順でソートすると第 1(200)，第 3(490)，第 2(610)，第 4 レコード (910) の順になるため，この順で推定レコードとする．この場合，再識別は完全に成功しているため，再識別率は 1.0 となる．

Identify-sa21($E4$)

Identify-sa21($E4$) 手法は，レコードの QI は考慮せず，特定の SA の値だけで再識別を行う．例えば，M'_B の第 2 レコードの SA1 の値 (220) と最も近い値を持つのは M の第 2 レコードの 200 であるため，これを推定レコードとする．

8.3 ユークリッド距離を用いた再識別手法）

8.3.1 Identify-euc

本章では，M'_B の再識別したいレコードと同じ QI ベクトルを持つレコードを M から探し，それらの SA のユークリッド距離 $euc(a_{SA}, b_{SA}) = \sqrt{\sum_{i=1}^{m}(b_{SA,i} - a_{SA,i})^2}$ を用いて再識別を行う．例えば，M'_B の第 1 レコードの QI ベクトルは (2,1,1) であり，M のうち同じ QI ベクトルを持つ個人は第 1,2 レコードである．M のそれらのレコードの SA ベクトルを $a_{SA,1} = (100, 100)$ と $a_{SA,2} = (200, 400)$ とし，M'_B の第 1 レコードの SA ベクトルを $b_{SA,1} = (110, 90)$ とする．これらのユークリッド距離を測ると $euc(a_{SA,1}, b_{SA,1}) = 14.142 < 322.8 = euc(a_{SA,2}, b_{SA,1})$ となるため，M の第 1 レコードと M'_B の第 1 レコードを同一人物として再識別する．

8.3.2 $EUC1$ and $EUC2$

表 8.2,8.3 の例では，M と M'_B の QI 属性の値と完全に一致するように SA 属性のみを加工する場合を考えていた．しかし，もし QI 属性の値まで加工されていたら M'_B のレコードの QI 属性と M のレコードの QI 属性が一致しなくなってしまうため，ユークリッド距離による再識別手法が使えなくなってしまう．表 8.4 に，QI3 属性の値が全て 1 に加工されたデータ M'_D を示す．

M'_D の第 3,4 レコードを本手法で識別する場合，M に同じ QI ベクトル (1,1,1) を持つレコードが存在しないため，再識別ができなくなってしまう．この問題を解決するために，再識別手法に 2 つの手法 EUC1, EUC2 を用意する．表 8.5 にこれらの手法の詳細を示し，アルゴリズムと動作例を以下に示す．

表 8.4: 加工データ M'_D(or M'_E)

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	1	300	200
1	1	1	400	500

表 8.5: 再識別手法 $EUC1$ と $EUC2$ の詳細

$EUC1$	同じ QI ベクトルを持つレコードが存在しなかった場合，レコードの識別を諦める
$EUC2$	同じ QI ベクトルを持つレコードが存在しなかった場合，全レコードから個人の再識別を試みる

$EUC1$ のアルゴリズム

1. 入力：元データ M, 加工データ M'_B, M'_B のレコード数 m, 再識別に用いる QI 属性 qi, 再識別に用いる SA 属性 sa
2. qi の値を用いてインデックス ind とそれに対応するレコード番号を作成する．
3. M'_B の i 番目のレコードと，M のうち QI ベクトルが等しい全レコードの間の sa のユークリッド距離 $euc(a_{SA,j}, b_{SA,i})$ を計算し，最も近いレコードを M'_B の i 番目のレコードの再識別結果とする．
4. もし同じ QI ベクトルを持つレコードが M に存在しない場合，M'_B の i 番目のレコードの再識別結果を M の i 番目のレコードとする（再識別を諦める）．
5. ステップ 3,4 を M'_B の各レコードについて行い，それらの再識別結果を出力する．

$EUC1$ の例 1

M, M'_B の場合を考える．QI が第 1,2,3 列であるため，それらの属性から M'_B のインデックス ind を作成する．表 8.6 にこの場合の ind を示す．M'_B の第 1 レコード $b_{SA,1}$ を再識別する場合，M の第 1,2 レコード $a_{SA,1}, a_{SA,2}$ は同じ QI ベクトルの値 (2,1,1) を持っているので識別対象となり，$b_{SA,1}$ とのユークリッド距離 $euc(a_{SA,1}, b_{SA,1})$, $euc(a_{SA,2}, b_{SA,1})$ が計算される．その結果，$a_{SA,1}$ が $b_{SA,1}$ と最も距離が近いレコードとなり，識別結果として出力される．この手順を $b_{SA,2}, b_{SA,3}, b_{SA,4}$ に対しても行う．

$EUC1$ の例 2

M, M'_D の場合を考える．QI が第 1,2,3 列であるため，それらの属性から M'_D のインデックス ind を作成する．表 8.7 にこの場合の ind を示す．D の第 1 レコード $d_{SA,1}$ を再識別する場合，M の第 1,2 レコード $a_{SA,1}, a_{SA,2}$ は同じ QI ベクトルの値 (2,1,1) を持っているので識別対象となり，$b_{SA,1}$ とのユークリッド距離 $euc(a_{SA,1}, b_{SA,1})$, $euc(a_{SA,2}, b_{SA,1})$ が計算されて最も近いレコードが再識別結果が出力され，この手順が $b_{SA,2}, b_{SA,3}, b_{SA,4}$ に対しても行われる．しかし，$d_{SA,3}$ と $d_{SA,4}$ を再識別す

表 8.6: $EUC1$ の例 1 における ind

key	index
(2,1,1)	1,2
(1,1,2)	3,4

表 8.7: $EUC1$ の例 2 における ind

key	index
(2,1,1)	1,2
(1,1,1)	3,4

る場合，同じ QI ベクトル (1,1,1) を持つレコードが M には存在しないので，探索を諦めて $a_{SA,3}$ と $a_{SA,4}$ をそれぞれ識別結果として出力する．

$EUC2$ のアルゴリズム

1. $EUC1$ とステップ 1,2,3 は同じ動作をする．
2. M に同じ QI ベクトルの値を持つレコードが存在しない場合，$b_{SA,i}$ と M の全レコードとのユークリッド距離を求め，最も近いレコードを識別結果として出力する．
3. M'_B の全レコードについて上記の手順を繰り返し，識別結果を出力する．

$EUC2$ の例 1

M, M'_D の場合を考える．QI が第 1,2,3 列であるため，それらの属性から M'_D のインデックス ind を作成する．表 8.7 にこの場合の ind を示す．M'_D の第 1 レコード $d_{SA,1}$ を再識別する場合，M の第 1,2 レコード $a_{SA,1}, a_{SA,2}$ は同じ QI ベクトルの値 (2,1,1) を持っているので識別対象となり，$b_{SA,1}$ とのユークリッド距離 $euc(a_{SA,1}, b_{SA,1})$, $euc(a_{SA,2}, b_{SA,1})$ が計算されて最も近いレコードが再識別結果が出力され，この手順が $b_{SA,2}, b_{SA,3}, b_{SA,4}$ に対しても行われる．しかし，$d_{SA,3}$ と $d_{SA,4}$ を再識別する場合，同じ QI ベクトル (1,1,1) を持つレコードが M には存在しないので，M の全レコードとのユークリッド距離を求めて最も近いレコードを探し，それぞれを識別結果として出力する．

8.4 評価

この節では，以下の点に注目して $PWSCUP2015$ に提出された加工データを分析する．

- 単一の匿名化手法の影響
- $PWSCUP2015$ の評価に基づく，単一の匿名化手法の評価
- 再識別手法 (identify-euc) の評価

8.4.1 $PWSCUP2015$ の加工データ

$PWSCUP2015$ に提出された 12 個の加工データ $D_1, .., D_{12}$ を評価する．表 8.8 に $D_1, .., D_{12}$ の詳細を示す．これらのデータは，疑似ミクロデータをもとに 5 つの参加チーム $1, \cdots, 5$ によって作成されたものであり，チーム $1, \cdots, 5$ には総合順位上 3 チームも含まれている．

表 8.8: $D_1, ..., D_{12}$ の詳細

Name	Team	Rank
D_1, D_2	1	
D_3, D_4	2	2
D_5, D_6	3	
D_7, D_8, D_9	4	1
D_{10}, D_{11}, D_{12}	5	3

表 8.9: $D_a, .., D_h$ の詳細

Name	Method	Target
D_a	k-匿名化	QI
D_b	SA ノイズ付加	SA
D_c	山岡匿名化	ID
D_d	QI 統一化 1	QI
D_e	QI 統一化 2	QI
D_f	SA 平均化	SA
D_g	QI スワップ	SA
D_h	レコード削除	レコード

8.4.2 単一の加工手法による匿名化データ

匿名化データ $D_1, .., D_{12}$ は複数の加工手法を組み合わせたものであるため，単一の匿名化手法の影響は不明である．そこで，単一の加工手法によって作成された加工データを用いてこれらの影響を調査し，$D_1, ..., D_{12}$ に使われた加工手法を推定する．

私は単一の加工手法を用いて，8 つの加工データ $D_a, .., D_h$ を作成した．$D_a, .., D_h$ は，疑似ミクロデータからランダムにサンプリングされた 100 レコードのデータから作成されたものである．表 8.9 に $D_a, .., D_h$ の加工手法と加工対象を示す．各加工手法については後述する．これらのデータを用いて，$D_1, .., D_{12}$ がどの手法を組み合わせて作成されたデータであるかを推定する．

SA ノイズ付加

SA ノイズ付加手法は，元データの SA 属性にランダムノイズを付与することによって加工データを作成する手法であり，表 8.3 の M'_B はこの手法による加工の一例である．この手法は SA の値を加工するため，SA に関わる有用性指標 $U1, U2, U4, U5$ は悪化し，SA を用いる再識別手法 $E3, E4$ への耐性は高まる．

QI 統一化

QI 統一化手法は，QI の値を他の値に置き換えることによって加工データを作成する手法であり，表 8.4 の M'_D はこの手法の加工の一例である．この手法で加工を行うと，SA に関する有用性を落とすことなく安全性を高めることができるが，有用性指標で QI が用いられている場合は有用性が悪化してしまう．例えば，*PWSCUP2015* では $U2, U3$ の 2 指標が QI1,.., QI6 を用いている．そこで，有用性に用いられていない QI を加工した場合とそうでない場合で種類分けをして考える．

表 8.10: 加工データ例 F

Group	QI1	QI2	QI3	SA1	SA2
1	2	1	1	150	250
1	2	1	1	150	250
2	1	1	2	350	350
2	1	1	2	350	350

表 8.11: 加工データ例 G

Group	QI1	QI2	QI3	SA1	SA2
1	2	1	1	200	100
1	2	1	1	100	400
2	1	1	2	300	500
2	1	1	2	400	200

SA 平均化（ミクロアグリゲーション）

SA 平均化手法は，同じ QI ベクトルを持つレコード群の SA の値をそれらの平均値に置き換える手法である．表 8.10 にこの手法で M を加工したデータ M'_F を示す．この場合，QI によってレコードは 2 つのグループに分けられている．この手法で加工をすると有用性 $U4, U5$ が悪化し，安全性 $E3, E4$ が向上する．

QI スワップ

QI スワップ手法は，同じ QI を持つレコード間で SA をスワップする手法である．表 8.11 にこの手法で M を加工したデータ例 M'_G を示す．グループ 1（第 1,2 レコード）では SA1 の値がスワップされており，グループ 2（第 3,4 レコード）では SA2 の値がスワップされている．値のスワップはグループ内のみで行われるため，平均値や有用性 $U2, U3$ は変化しないが，相関係数を用いた有用性 $U4, U5$ は悪化し，安全性 $E2, E3, E4$ は高まる．

レコード削除

データ中のいくつかのレコードを削除する手法である．この手法で加工を行うと，有用性 $U1, U2, U3, U5, U6$ は悪化し，安全性 $E3, E4$ は高まる．しかし，*PWSCUP2015* ではこの手法は用いられなかったことに注意せよ．

k-匿名化，山岡匿名化

k-匿名化 [59] による加工で有用性 $U2, U3$ は悪化し，安全性 $S1, S2, E1, E2$ は悪化する．また，山岡匿名化 [58] はデータそのものは加工せず，レコードのインデックスを入れ替える加工手法である．この手法による加工は $U5$ 以外の有用性を落とすことなく，安全性 $E1, .., E4$ を高めることができる．

8.4.3 期待できる効果

一般的に，データを匿名化すると有用性が低下し安全性が高まる．表 8.12 に，前述した匿名化手法による影響の予測結果を示す．表には “positive”, “slightly”, “negative”, “-” の 4 つの値が示されている．有用性の欄の “negative” は大きく低下することを意味し，“significantly decrease” は僅かに低

表 8.12: 影響の予測

	k-匿名化	SA ノイズ付加	山岡匿名化	QI 統一化 1	QI 統一化 2	SA 平均化	QI スワップ	レコード削除
$U1$	-	slightly	-	-	-	-	-	negative
$U2$	negative	slightly	-	-	negative	-	-	negative
$U3$	negative	slightly	-	-	negative	-	-	negative
$U4$	-	slightly	-	-	-	negative	negative	negative
$U5$	-	slightly	negative	-	-	negative	negative	negative
$U6$	-	-	-	-	-	-	-	negative
$S1$	positive	-	-	-	-	-	-	-
$S2$	positive	-	-	positive	positive	-	-	-
$E1$	slightly	vulnerable	positive	slightly	slightly	vulnerable	vulnerable	vulnerable
$E2$	slightly	vulnerable	positive	slightly	slightly	vulnerable	slightly	vulnerable
$E3$	vulnerable	slightly	positive	vulnerable	vulnerable	positive	positive	vulnerable
$E4$	vulnerable	slightly	positive	vulnerable	vulnerable	positive	slightly	vulnerable
EUC	slightly	vulnerable	positive	slightly	slightly	vulnerable	positive	vulnerable

下することを意味し，“-” は変化しないことを意味する．安全性 $S1, S2$ の欄では，“positive” は向上することを意味し，“-” は変化しないことを意味する．再識別手法 $E1, \ldots, E4$ の欄では，“positive” は大きく耐性がつくことを意味し，“slightly” は僅かに耐性がつくことを意味し，“negative” は耐性が無いことを意味する．

8.4.4 評価結果

匿名化手法の予測と結果

表 8.13 に $D_a, ..., D_h$ の有用性/安全性評価値を示す．“original” の列には元データの評価結果を示している．指標 E4(identify.sa21) の値が全体的に低いように思えるが，これは疑似ミクロデータの 21 列目に 0 が多く，それらのレコードは正しく再識別することができないためである．このデータは 100 レコード中 76 レコードの 21 列目が 0 であるため，E4 の最大値は 0.24 となっている．

また，表 8.14 には $D_1, ..., D_{12}$ の有用性と安全性を示し，表 8.15 には $D_1, ..., D_{12}$ の評価結果と，それによる加工手法の予測を示す．匿名化手法は組み合わせると，それらの特徴を併せ持った加工になる．例えば，D_{10} は k-匿名化と SA 平均化を組み合わせて匿名化されたデータであるため，D_a と D_f の特徴を併せ持っている (表 8.12,8.15 に示す)．データ $M, M'_B, M'_C, M'_D, M'_F, M'_G$ と $D_a, ..., D_h$ は異なるが，記号が同じデータは同じ匿名化手法で加工されている．例えば，M'_B と D_b は両方 SA ノイズ付加で加工されたデータである．

8.4.5 EUC1 と既存手法との比較

再識別手法 EUC1 と既存手法 $E1, .., E4$ の比較を，元データ（疑似ミクロデータ）と加工データ $D_1, ..., D_{12}$ を用いて行う．表 8.16 に $E1, ..., E4$ と EUC1 の再識別率を示す．表中の赤い値は，その

表 8.13: $D_a, ..., D_h$ の有用性と安全性

	Original	D_a	D_b	D_c	D_d	D_e	D_f	D_g	D_h
$U1$	0	0	46.225	0	0	0	0	0	295.731
$U2$	0	38837.9	7808.7	0	0	15135.7	104.9	209.7	1094.4
$U3$	0	5.833	0	0	0	2	0	0	0.097
$U4$	0	0	0.02	0	0	0	0	0	0.049
$U5$	0	0	0.016	0.12	0	0	0	0	0
$U6$	0	0	0	0	0	0	0	0	10
$S1$	1	3	1	1	1	1	1	1	1
$S2$	1.031	7.692	1.031	1.031	1.053	1.053	1.031	1.031	1.034
$E1$	0	0.13	0.99	0	0.07	0.11	0.94	1	1
$E2$	0	0.17	1	0	1	1	1	1	1
$E3$	0	1	0.54	0	1	1	1	0.91	0.067
$E4$	0.24	0.24	0.22	0	0.24	0.24	0.24	0.24	0.089
$EUC1$	0	0.13	1	0	0.07	0.11	1	1	1
$EUC2$	0	0.17	1	0	1	1	1	1	1

加工データに対して最も有効な再識別手法であることを意味している．手法 EUC1 は，12 データ中 5 データに対して最も有効であり，他の手法よりも優れている．

8.4.6 再識別手法の処理性能評価

疑似ミクロデータには QI 属性が 13 あり，そのうちどれを identify.euc による再識別に用いるかによって計算時間と再識別率が変化する．再識別に用いる QI 属性の数を $|qi|$，SA 属性の数を $|sa|$ とおくと，$|qi|$ を増やせば計算量は少なくなるが，それに応じて QI の加工に弱くなり，再識別率が下がりやすい．図 8.1,...,8.4 に，100 レコード，25 属性のデータを用いた時の，$|qi|$ と $|sa|$ の変化に伴う計算時間と再識別成功率の変化を示す．図 8.1 より，計算時間は $|qi|$ について単調に減少しており，図 8.2 より，再識別率は $|qi|$ について単調に増加している (ただし，$|qi| = 5$ で飽和している)．図 8.3 より，計算時間は $|sa|$ について単調に増加しているが，小規模データでテストしているため誤差が大きい．また図 8.4 より，再識別率は $|sa|$ に依存しなかった．

8.5 まとめ

PWSCup2015 の加工データを用いて，再識別手法 identify.euc と既存手法の比較を行った．また，単独匿名加工手法で加工した小規模データを用いて *PWSCup2015* の匿名加工データの解析を行った．その結果，匿名化手法を組み合わせてデータを加工すると，用いた複数の手法の有用性/安全性指標への影響を併せ持つ加工データとなり，*PWSCup2015* 上位チームの匿名加工データはそれらをうま

表 8.14: $D_1, ..., D_{12}$ の有用性と安全性

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}
$U1$	0	0	0	0	0	0	0	0	0	0	0	0
$U2$	58340.87	0	31572.91	31400.95	0	0	4321.75	0	0	65093.42	52975.02	46100.64
$U3$	18.6	0	1.01	0.99	0	0	1.54	0	0	7.28	2.97	1.85
$U4$	0	0.01	0	0	0.07	0.07	0.03	0.09	0.09	0.15	0.11	0.11
$U5$	0	0.01	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.02	0.02	0.02
$U6$	0	0	0	0	0	0	0	0	0	0	0	0
$S1$	1	1	3	3	1	1	3	1	1	41	8	4
$S2$	2.66	1.88	4.91	4.86	36.07	36.07	36.07	13.71	13.68	106.83	42.3	31.09
$E1$	0.03	0.65	0.2	0.19	0	0	0	0	0	0.01	0.02	0.02
$E2$	0.82	0.65	0.24	0.24	0.02	0.02	0.02	0	0	0.01	0.02	0.02
$E3$	1	0	0.25	0.25	0	0	0.01	0	0	0	0	0
$E4$	0.19	0	0.05	0.05	0	0	0	0	0	0	0	0
$EUC1$	0.3	0.48	0.21	0.21	0.07	0.07	0.88	0	0	0	0.01	0.01

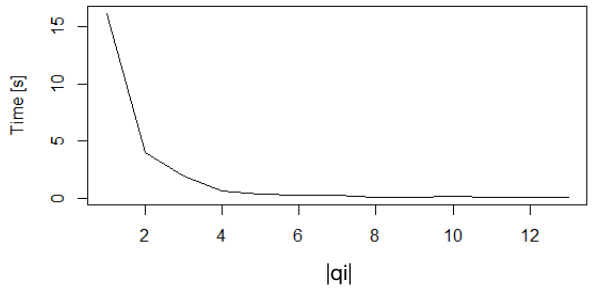


図 8.1: QI の属性数による計算時間の変化

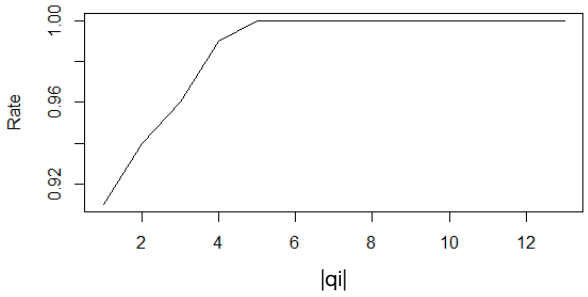


図 8.2: QI の属性数による再識別率の変化

く組み合わせて作成されていること，また，identify.euc(EUC1) は他の手法 (特に identify.sa) と比べて計算時間が大幅に多い割には再識別率にはあまり差はない.

表 8.15: $D_1, ..., D_{12}$ の評価と予測結果

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}
$U1$	-	-	-	-	-	-	-	-	-	-	-	-
$U2$	negative	-	negative	negative	-	-	negative	-	-	negative	negative	negative
$U3$	negative	-	slightly	slightly	-	-	slightly	-	-	negative	negative	slightly
$U4$	-	slightly	-	slightly	slightly	slightly	slightly	slightly	slightly	negative	negative	negative
$U5$	-	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly	slightly
$U6$	-	-	-	-	-	-	-	-	-	-	-	-
$S1$	-	-	slightly	slightly	-	-	slightly	-	-	positive	positive	slightly
$S2$	-	-	slightly	slightly	slightly	positive	positive	positive	positive	positive	positive	positive
$E1$	slightly	negative	negative	negative	positive	positive	positive	positive	positive	slightly	slightly	slightly
$E2$	negative	negative	negative	negative	slightly	slightly	slightly	positive	positive	slightly	slightly	slightly
$E3$	negative	positive	negative	negative	positive	positive	positive	positive	positive	positive	positive	positive
$E4$	negative	positive	slightly	slightly	positive	positive	positive	positive	positive	positive	positive	positive
$EUC1$	negative	negative	negative	negative	slightly	slightly	negative	positive	positive	positive	positive	positive
D_a	-	-	✓	✓	-	-	✓	-	-	✓	✓	✓
D_b	-	-	-	-	-	-	-	-	-	-	-	-
D_c	-	-	-	-	✓	✓	-	✓	✓	-	-	-
D_d	-	-	-	-	✓	✓	✓	-	-	-	-	-
D_e	✓	-	-	-	-	-	-	✓	✓	-	-	-
D_f	-	✓	-	-	-	-	-	-	-	✓	✓	✓
D_g	-	-	✓	✓	-	-	✓	✓	✓	-	-	-
D_h	-	-	-	-	-	-	-	-	-	-	-	-

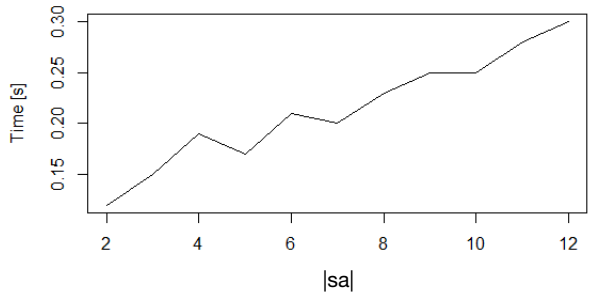


図 8.3: SA の属性数による計算時間の変化

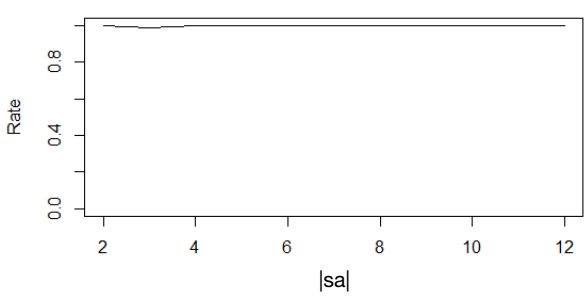


図 8.4: SA の属性数による再識別率の変化

表 8.16: EUC1 と既存手法との比較結果

加工データ	id-rand	id-sa	id-sort	id-sa21	EUC1
D_1	0.033	0.824	*1.000	0.186	0.301
D_2	0.649	*0.651	0.001	0.002	0.478
D_3	0.199	0.241	*0.248	0.051	0.207
D_4	0.189	0.240	*0.253	0.045	0.211
D_5	0.000	0.022	0.000	0.000	*0.074
D_6	0.000	0.022	0.000	0.000	*0.074
D_7	0.002	0.022	0.009	0.001	*0.876
D_8	0.000	0.000	0.000	0.000	*0.001
D_9	0.000	0.000	0.000	0.000	*0.002
D_{10}	0.006	*0.007	0.000	0.000	0.004
D_{11}	*0.018	0.016	0.000	0.000	0.008
D_{12}	*0.021	*0.021	0.000	0.000	0.008
average	0.093	0.172	0.126	0.024	*0.187
standard deviation	0.174	0.258	0.268	0.050	0.243
best score	2	3	3	0	5

第9章　匿名化モデル k-concealment の履歴データに対する改良

9.1　導入

Tamir らは k-匿名性 [2] の厳密さによって生じる過度な加工を指摘し，それを解決するために k-concealment という新しい指標を提案した [5]．k-concealment を満たすように加工すれば，k-匿名化されたデータと同等の安全性を持ちながら，より高い有用性を持つデータを作成することができる．しかしながら，[5] ではレコード数と個人数が等しい静的なデータしか想定されていない．

本章では，レコード数が個人数より多い履歴データにも適用できる k-concealment の改善指標を提案し，それを満たすための加工手法を提案する．提案加工手法は，仮名の一般化とレコード間 k-concealment を用いるものである．

9.1.1　履歴データの k-匿名化

定義 9.1.1 顧客数とレコード数が等しいデータを静的データ，レコード数が顧客数より多い（*1* 顧客が複数のレコードを持つ）データを履歴データとよぶ．

履歴データを k-匿名化する場合，k レコードではなく k 人の顧客の区別がつかなくなるようにする必要がある．また，履歴データでは 1 顧客ごとのレコード数が異なる場合が多いため，k-匿名化する場合にはレコードを削除または追加し，レコード数をそろえる必要がある．

履歴データの例として，購買履歴データ T_5 を表 9.1 に示す．T_5 は顧客 3 人の 4 日分の購買履歴であり，このデータを 3-匿名化する場合を考える．Alice は 3 レコード，Bob と Carol はそれぞれ 2 レコードを持っているため，この 3 人の区別がつかなくするためにはダミーレコードを追加し，すべての顧客の持つレコード数を 3 にそろえてやる必要がある．T_5 にダミーレコードを追加して 3-匿名化したデータ T_6 を表 9.2 に示す．id に*印がついているレコードがダミーレコードであり，この場合 2 レコードを追加してデータを一般化することによって 3 顧客の区別がつかなくなっている．また前述したように，このデータを完全 2-匿名化することはできない（1 人余ってしまうため）．

9.2　履歴データの k-concealment 化

9.2.1　アイデア

2 章で紹介した k-concealment は静的データを評価・加工する場合を想定したものであり，履歴データを評価・加工することは考えられていなかった．

表 9.1: 購買履歴データ T_5

name	date	goods
Alice	12/1	a
Alice	12/2	b
Alice	12/3	c
Bob	12/2	d
Bob	12/3	e
Carol	12/3	f
Carol	12/4	g

表 9.2: 3-匿名化された購買履歴データ T_6

id	date	goods
1	12/1	a
1	12/2-12/3	b,d,f
1	12/3-12/4	c,e,g
*2	12/1	a
2	12/2-12/3	b,d,f
2	12/3-12/4	c,e,g
*3	12/1	a
3	12/2-12/3	b,d,f
3	12/3-12/4	c,e,g

履歴データの k-匿名化については，2018 年に開催された匿名加工・再識別コンテスト PWSCUP-2018[6] にて議論されている．この大会では顧客 400 人の 1 年分の購買履歴データ（81,776 レコード）が加工対象として用いられており，上位チームの加工データにはレコードの削除・追加によって k-匿名化されたものが多かった．しかしながら，データが大規模であるほど顧客ごとのレコード数の違いも大きく，このデータ内で最もレコード数が多い顧客（4,289 レコード）と 2 番目に多い顧客（1,601 レコード）の区別をつかなくするには，2,688 レコードを削除する必要があった．この大会では，レコードを削除するとデータの有用性が大きく下がるようにルールが定められていたため，レコード数が多く k-匿名化にコストのかかる顧客を見捨てる（加工をしない）チームもあった．

しかしながら，**履歴データを「最低でも k 人の区別がつかない」状態にするためにはレコードの追加・削除や顧客の削除が必須なのだろうか？**本研究では，ここに k-concealment のアイデアを導入することにより，この問に否定的に回答する．キーとなる手法は，レコード数の異なる複数の顧客に対する「仮名の一般化」と「レコードの k-concealment 化」である．次節ではその手法・アルゴリズムを提案する．

9.2.2 仮名の一般化，レコードの k-concealment 化

加工データの仮名を一般化することにより，追加・削除するレコードの数を減らすことができる．例えば T_6 は 3-匿名化された T_5 であるが，これの仮名を表 9.3 の T_7 の 1 レコード目のように一般化することによって，ダミーレコードの数を 0 にすることができる．仮名 1,2,3 は実際には別のランダムな ID，例えば 4 とみなせばよい．仮名 4(1,2,3) は，Alice,Bob,Carol のどの顧客にも当てはまるようにすることで，加工レコードと真のレコードの関係を識別不能にする．

T_5 を 2-匿名化するためには顧客の削除が不可欠であるが，仮名の一般化だけではこの問題は解決できないため，さらにレコードの k-concealment 化を行う．表 9.4 の T_8 のように加工することにより，顧客と仮名が図 9.1 の関係を満たすように T_5 を 2-concealment 化することができる．T_8 は顧客 2 人の区別がつかないのに加え，2 レコードの区別がつかない状態にもなっており，図 9.1 の通り Alice

表 9.3: 仮名が一般化された購買履歴データ T_7

id	date	goods
1,2,3	**12/1**	**a**
1	12/2-12/3	b,d,f
1	12/3-12/4	c,e,g
2	12/2-12/3	b,d,f
2	12/3-12/4	c,e,g
3	12/2-12/3	b,d,f
3	12/3-12/4	c,e,g

表 9.4: 2-concealment 化された購買履歴データ T_8

id	date	goods
1,2	12/1-12/3	a,f
1	12/2-12/3	b,f
1	12/3-12/4	c,g
2	12/1-12/2	a,b,d
2	12/3	c,e
3	12/2-12/3	d,f
3	12/3-12/4	e,g

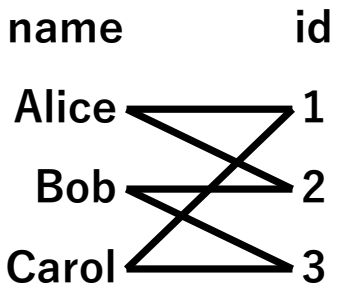


図 9.1: T_5 を 2-concealment した場合の 2 部グラフ

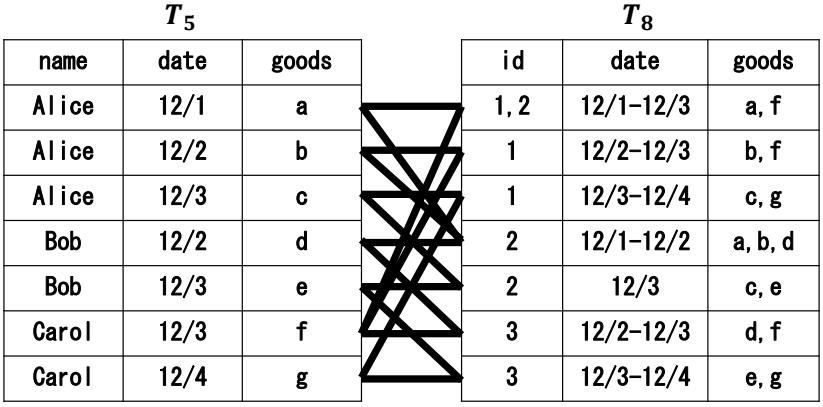


図 9.2: T_5 と T_8 間のレコードの対応を示す 2 部グラフ

と Carol が id1 に，Alice と Bob が id2 に，そして Bob と Carol が id3 に一般化されている．これは T_7 と比較すると加工の幅が小さくなっていることがわかる．また，T_5 と T_8 の間のレコードの対応を表す 2 部グラフを図 9.2 に示す.

9.2.3 基礎定義

履歴データに対する k-concealment 手法を提案するために，以下の定義を行う.

定義 9.2.1 履歴データを T，T のレコード数を m，顧客数を n，属性数を ρ とする．顧客の集合を $U = \{u_1, u_2, \ldots, u_n\}$ とし，T のうち顧客 u_i が持つ b_i 個のレコードの集合を $T_i = \{d_1^{(i)}, d_2^{(i)}, \ldots d_{b_i}^{(i)}\}$ とする．ここで，j 番目の要素は $d_j^{(i)} = (u_i, d_{j,2}^{(i)}, \ldots, d_{j,\rho}^{(i)})$ である.

$T = T_1 \cup T_2 \cup \ldots \cup T_n$ であり，$b_1 + \cdots + b_n = m$ である．また，T の 1 列目の属性は顧客の識別子であることを仮定している.

定義 9.2.2 仮名化されたTをT'とし，Uを仮名化した仮名の集合を$V = \{v_1, v_2, \ldots, v_{n'}\}$とする．ここで，仮名$v_i$が持つレコードの集合は$T'_i$とする．ここで，その$j$番目の要素は$d'^{(i)}_j = (v_i, d^{(i)}_{j,2}, \ldots, d^{(i)}_{j,\rho})$である．

1人の顧客に複数の仮名を振ることを許す．すなわち，$n \leq n'$となることがあることに注意せよ．

例 9.2.1 $T = T_5$の場合，$m = 7, n = 3$であり，U =*{Alice, Bob, Carol}*である．u_1 =*Alice*とすると$b_1 = 3$であり，T_1はT_5の*1,2,3*レコード$\{d^{(1)}_1$=*(Alice,12/1,A)*$, d^{(1)}_2, d^{(1)}_3\}$を要素とする集合である．この場合，Uは$V = \{1, 2, 3\}$に仮名化される．

定義 9.2.3 T_iとT_jについて定められる顧客u_iとu_j間の距離を$d_{i,j}$で表し，Uの顧客間の距離行列を$dist(T)$とする．$dist(T)$はn次正方行列である．レコード数の異なる顧客間の距離は，T_iとT_jの全レコードを一般化した場合を考えて求め，余ったレコードは削除したときのコストで定める．

例 9.2.2 例えば，T_5についての距離行列$dist(T)$は表*9.5*のようになる．*PWSCUP-2018*で濱田らは一般化されたデータと元データの距離を，一般化した集合の大きさに応じて，真の値となる期待値で定義している*[6]*.（ただし，表*9.5*はこの定義で求めたものではない）

定義 9.2.4 (顧客間対称 k-concealment) 顧客集合Uとその仮名集合Vとなる*k-concealment*な*2*部グラフ$G = (U, V, E)$において，$u_i \in U$と$v_j \in V$について$(u_i, v_j) \in E$ならば$(u_j, v_i) \in E$であるとき，Gは対称であるという．また，Gは$dist(T)$をもとにして距離を最小化されたものである．

例 9.2.3 図*9.1*と図*9.3*はどちらも*2-concealment*を満たす*2*部グラフであるが，図*9.1*は対称であり，図*9.3*は対称ではない．

定義 9.2.5 (レコード間 k-concealment) $G = (U, V, E)$を*k-concealment*を満たす対称な*2*部グラフ，TをUが有するレコード集合，T'をその仮名化とする．全ての$(u_i, v_j) \in E$について，$\{d^{(i)}_1, \ldots, d^{(i)}_{b_i}\} \subset T$と$\{d'^{(j)}_1, \ldots, d'^{(j)}_{b_j}\} \subset T'$が*2*部グラフとなる辺集合$E_{i,j} \subset E'$が存在し，$(d^{(i)}_x, d'^{(i)}_y)$ $(x \neq y) \in E_{i,j}$は存在しないとき，$G_k = (T, T', E')$をレコード*k-concealment*という．

例 9.2.4 図*9.3*は定義*9.2.5*の条件を満たしているため，レコード*2-concealment*である．

定義 9.2.6 レコード$(d^{(i)}_{x,1}, \ldots, d^{(i)}_{x,\rho})$と$(d^{(j)}_{y,1}, \ldots, d^{(j)}_{y,\rho})$の一般化とは，レコード$(v, g_2, \ldots, g_\rho)$とする．ここで，$w = 2, \ldots, \rho$について$g_w = g^{(i)}_{x,w} \cup g^{(j)}_{y,w}$である．また，仮名$v$は

$$v = \begin{cases} v^{(i)} & if\ b_i \geq b_j, \\ \{v^{(i)}, v^{(j)}\} & otherwise \end{cases}$$

とする．ここで，$v^{(i)}$はv_iに割り当てた一意でランダムな仮名である．

Algorithm 5 提案手法

Input: 履歴データ T, パラメータ k，顧客集合 U

Output: 加工データ T_{out}

Step 1.

T を仮名化して T' とする．T' の仮名集合を V とする．

Step 2.

$T_1, \ldots, T_n$ の距離を $n \times n$ の距離行列 $dist(T)$ で表す．

Step 3.

$dist(T)$ をもとにマッチングコストを最小化した，顧客間の対称かつ k-concealment な 2 部グラフ $G = (U, V, E)$（定義 9.2.4）を作る．

Step 4.

G をもとに，レコード間 k-concealment な T と T' の 2 部グラフ $G_k = (T, T', E')$（定義 9.2.5）を作る．

Step 5.

$d_w^{(j)} \in T'$ について（$w = 1, \ldots, m$），$(d, d'^{(j)}_w) \in E'$ となる全ての $d \in T$ の一般化を g_w とする（定義 9.2.6）．

Step 6.

$g_1, \ldots, g_m$ を結合し，T_{out} を作る．

9.2.4 提案アルゴリズム

3.1,3.2 節にて提案した，履歴データに対する k-concealment 化手法をアルゴリズム 5 に示す．入力する T は元データ，k は顧客に対する安全性のパラメータであり，出力される T_{out} は k-concealment を満たす加工された履歴データである．

例 9.2.5 例として，アルゴリズム *5* を用いて，T_5 を *2-concealment* 化する場合を考える．入力するものは $T = T_5, k = 2$ とする．

Step 1. T の顧客をレコード順にソートする．この場合 $U = \{u_1, u_2, u_3\}$ =$\{$*Alice,Bob,Carol*$\}$ であり，$b_1 = 3, b_2 = 2, b_3 = 2$ である．また，U は $V = \{v_1, v_2, v_3\} = \{1, 2, 3\}$ に仮名化するとし，仮名化された T を T' とする．

Step 2. T_1, T_2, T_3 の距離を測り，距離行列 $dist(T)$ を作る．この場合，$dist(T)$ は表 *9.5* のようになるとする．

Step 3. $dist(T)$ をもとに対称かつ *k-concealment* な *2* 部グラフ $G = (U, V, E)$ を作る．この場合，G は図 *9.3* のようになるとする．

Step 4. G をもとにレコードの *2* 部グラフ $G_k = (T, T', E')$ を作る．この場合 G_k は図 *9.4* のようになる．実線はマッチングの際に T' から張られた辺であり，点線は実線を反転することによって張られた辺である．赤実線はマッチングで余ってしまうレコード $(d'^{(1)}_3)$ から張られている辺であり，このようなレコードの仮名が一般化される．

Step 5,6. T' の各レコードを，G_k で辺がつながっているレコードと一般化し，T_{out} を作る．この場合，T_{out} は表 *9.6* のようになり，このデータは *2-concealment* を満たしている．

表 9.5: T_5 についての距離行列 $dist(T)$

name \id	1	2	3
Alice	0	1	5
Bob	1	0	4
Carol	5	4	0

表 9.6: G_k をもとに 2-concealment 化されたデータ T_{out}

id	date	goods
1	12/1-12/2	a,d
1	12/2-12/3	b,e
1,2	12/3	c,e
2	12/1-12/3	a,d,f
2	12/2-12/4	b,c,e,g
3	12/2-12/3	d,f
3	12/3-12/4	e,g

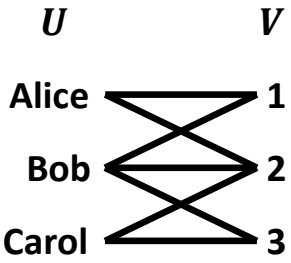


図 9.3: $dist(T)$ をもとに T_5 の顧客と仮名の対応を示した 2 部グラフ G

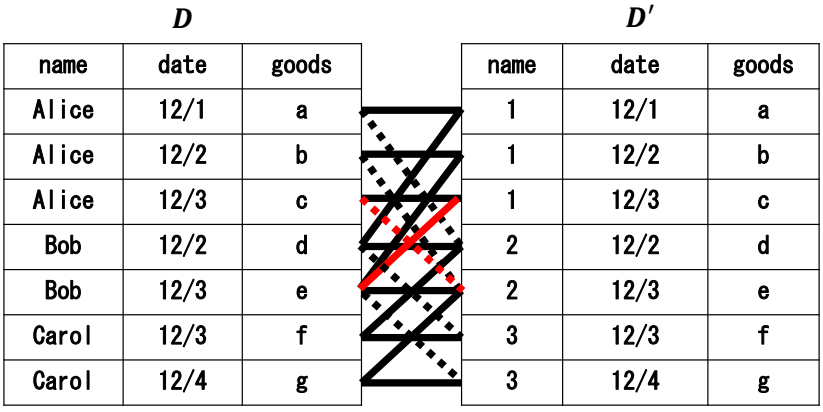


図 9.4: T_5 におけるレコード間 2 部グラフ G_k

9.3 実験

本章では履歴データを k-concealment 化することによって，データの特性がどう変化するかを実験によって評価する．しかし，この章で用いる k-concealment 化手法は 3 章で提案したものと厳密に同じではなく，レコードを補間することによる簡易的な手法であることに注意せよ．

9.3.1 レコード補間 k-concealment 化手法

履歴データの特徴は顧客ごとにレコード数が異なることであるが，レコードを補間してレコード数を揃えることによって静的データのように扱うことができ，容易に k-concealment 化をすることができる．T_5 を 2-concealment 化する場合を考える．このデータは顧客 3 人の 12/1〜12/4 の購買履歴データであるが，4 日分すべてのデータを持っている顧客はいない．そこで，各顧客の欠けているレコードを補間して，すべての顧客が 4 日分のデータ（4 レコード）を持つようにデータを変形する．

表 9.7: レコードを補間した購買履歴データ T_9

name	date	goods
Alice	12/1	a
Alice	12/2	b
Alice	12/3	c
Alice	12/4	*
Bob	12/1	*
Bob	12/2	d
Bob	12/3	e
Bob	12/4	*
Carol	12/1	*
Carol	12/2	*
Carol	12/3	f
Carol	12/4	g

表 9.8: 2-concealment を満たす購買履歴データ T_{10}

id	date	goods
1	12/1	a,*
1	12/2	b,*
1	12/3	c,f
1	12/4	g,*
2	12/1	a,*
2	12/2	b,d
2	12/3	c,e
2	12/4	*
3	12/1	*
3	12/2	d,*
3	12/3	e,f
3	12/4	g,*

T_5 のレコードを補間したデータ T_9 を表 9.7 に示す．例として，図 9.1 のような関係になる加工データを作る場合を考えると，Alice と Carol を一般化した id 1，Alice と Bob を一般化した id 2，Bob と Carol を一般化した id 3 を作成すれば 2-concealment を満たすことができる．T_9 を図 9.1 を満たすように一般化したデータ T_{10} を表 9.8 に示す．この手法は，顧客間の距離を測ることとレコード間の対応を決めることが容易である．

9.3.2 疑似人流データ

本章では，ナイトレイ社から公開されている疑似人流データ [11] を用いる．このデータから 10 人,50 人,100 人,500 人,1000 人の顧客をランダムに抽出したものを順に $T_{10}, T_{50}, T_{100}, T_{500}, T_{1000}$ とし，これらを実験に用いる．疑似人流データと 5 つの実験用データの統計量を表 9.9 に示す．

疑似人流データは 9 属性（顧客 ID，性別，日時，緯度，経度，位置カテゴリ 1，位置カテゴリ 2，状態，カテゴリ ID）のデータであるが，本章ではそのうち 4 属性（顧客 ID，日時，緯度，経度）だけを用いている．T_{100} をレコード補間したのち，各顧客間の緯度・経度のユークリッド距離の分布を図 9.5 に示す．

9.3.3 評価実験

k-anonymity と k-concealment を比較するために，いくつかの評価を行った．

k-anonymity 化されたデータと k-concealment 化されたデータの有用性と n の関係を図 9.6 に示す．ここで，有用性は元データと加工データの距離のユーザ平均値で定める誤差で評価する．青線が

表 9.9: 疑似人流データと実験データの統計量

データ	個人数 n	レコード数 m
疑似人流データ	6,432	901,465
T_{10}	10	1,402
T_{50}	50	6,608
T_{100}	100	13,503
T_{500}	500	68,699
T_{1000}	1,000	141,511

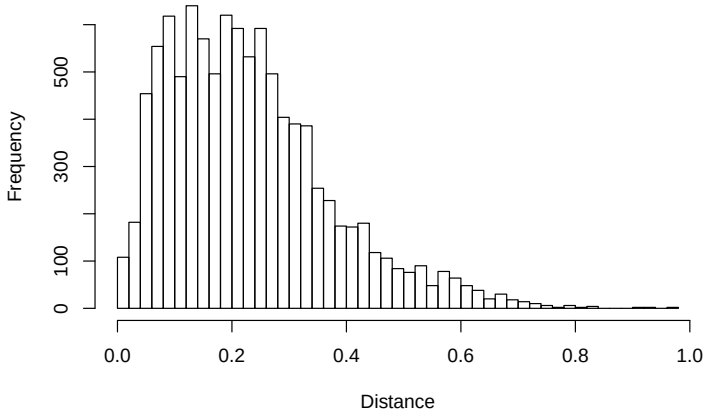


図 9.5: T_{100} の 2 顧客間の距離の分布

2-anonymity 化されたデータの，赤線が 2-concealment 化されたデータの各々の有用性を示している．この場合 n の値にかかわらず，2-anonymity 化されたデータの方が有用性が高い（=誤差が小さい）ことがわかる．

図 9.7 に $n = 100$ のときの誤差の分布を示す．図 9.6 と同じく，青線が 2-anonymity 化されたデータ，赤線が 2-concealment 化されたデータの分布を示している．2-concealment 化されたデータは誤差の分布が横に広がっており，誤差の平均値が大きい．

T_{100} を 2-anonymity 化したデータと 2-concealment 化されたデータを，実際の位置情報に投影して可視化した結果の一部をそれぞれ図 9.8,9.9 に示す．図中の黒点は元の座標情報を示しており，赤四角は一般化された範囲を示している．2-anonymity の方は独立した 2 点が一般化されているが，2-concealment の方は 2 つの四角（一般化された範囲）が 1 つの点を共有している場合がある．しかし，一般化後にできる四角形が大きいほど有用性は低くなり，図 9.9 の方が明らかに四角形の面積は大きい．

この実験では k-concealment の有用性は k-anonymity よりも低いという結果を得たが，マッチングの方法や 2 部グラフの作り方に改良の余地が残っている．

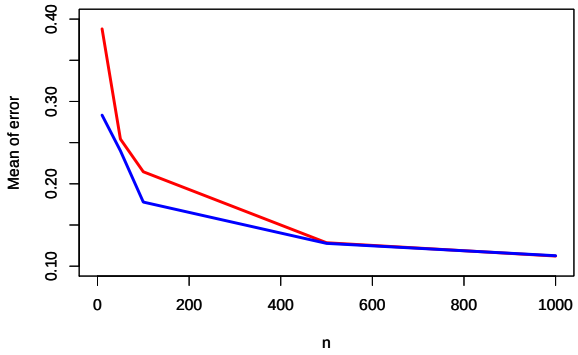


図 9.6: 2-anonymity 化されたデータと 2-concealment 化されたデータの顧客数 n についての有用性の平均誤差

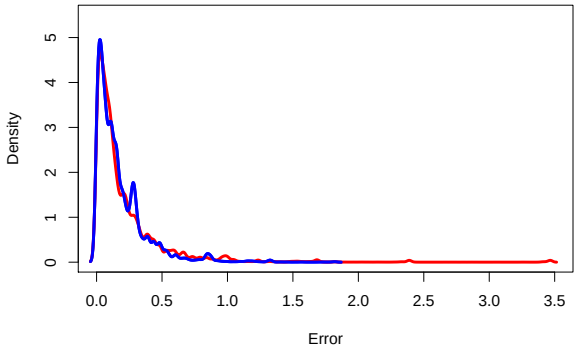


図 9.7: 2-anonymity 化されたデータと 2-concealment 化されたデータの誤差の分布

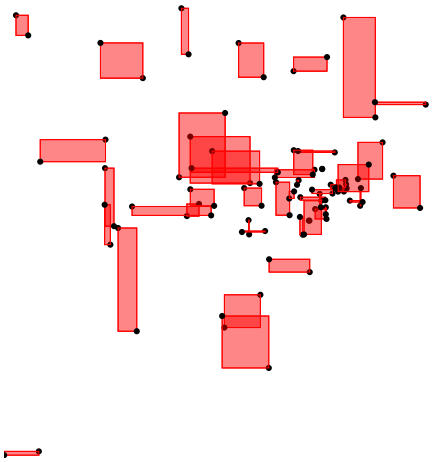


図 9.8: 2-anonymity 化された T_{100} の可視化

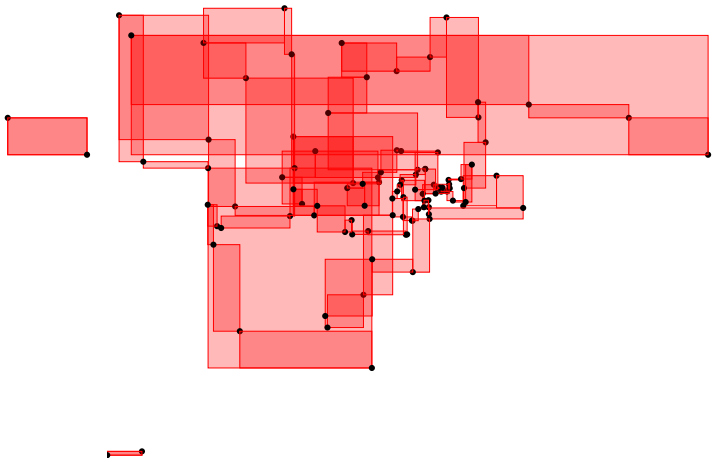


図 9.9: 2-concealment 化された T_{100} の可視化

9.4 まとめ

本章では**履歴データを「最低でも k 人の区別がつかない」状態にするためにはレコードの追加・削除や顧客の削除が必須なのだろうか？**という問題を解決するために,「仮名の一般化」と「レコード間の k-concealment」を用いた履歴データの k-concealment 化手法を提案した．提案手法を用いることにより，顧客やレコードの削除・追加をすることなく，履歴データを「最低でも k 人の区別がつかない」状態にすることが可能である．また，ナイトレイ社から公開されている疑似人流データを用いて，提案する手法の簡易版の実装をし，その手法の評価をした．

第10章　完全k-concealment 匿名化を求める精度の高いアルゴリズムの評価

10.1　導入

本章では，データを「全ての個人が等しく k 人と区別がつかない」状態に加工する完全 k-concealment 化手法を検討する．私は，完全 k-concealment 化にかかるコストを下げるために，巡回セールスマン問題の近似解法とクラスタリングを応用した手法を提案する．

10.2　基礎定義

10.2.1　データセット

本研究では，レコード（行）と属性（列）によって構成される個人データを考える．データ中の個人はレコードを一つのみ持ち，個人数とレコード数は常に等しくなる．記号等を以下のように定義する．

定義 10.2.1 （個人データ） 個人データを T とする．T のレコード数と個人数は n であり，ID 列を除く属性数は ρ である．T の個人集合を $U = \{u_1, \ldots, u_n\}$ とし，各個人は ρ 種類の属性にそれぞれ連続値または離散値の欠損値でない値を持つ．T の属性集合を $atr = \{a_1, \ldots, a_\rho\}$ とし，個人 u_i が属性 a_x に持つ値を $v_{i,x}$ とする．

例 10.2.1 T の例として，表 *10.1* に個人データ T_{ex} を示す．T_{ex} は *4* 人の個人 $U = \{u_1 = Alice, u_2 = Bob, u_3 = Carol, u_4 = David\}$ と *2* つの属性 $atr = \{a_1 = age, a_2 = sex\}$ を持つため，$n = 4, \rho = 2$ である．*4* 人の個人はそれぞれ *age* 属性の連続値と *sex* 属性の離散値を持ち，例えば $v_{1,1}$ は *Alice* が *age* 属性に持つ 10 を意味する．

表 10.1: 個人データの例 T_{ex}

ID	age	sex
Alice	10	F
Bob	20	M
Carol	40	M
David	50	F

表 10.2: T_{ex} の個人間の距離行列 $\mathrm{dist}(T_{\mathrm{ex}})$

	Alice	Bob	Carol	David
Alice	0	1.25	1.75	1.00
Bob	1.25	0	0.50	1.75
Carol	1.75	0.50	0	1.25
David	1.00	1.75	1.25	0

定義 10.2.2 （個人間の距離） T の距離行列を

$$dist(T) = \begin{pmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{pmatrix}$$

とする，ここで，d_{ij} は個人 u_i と個人 u_j の間の距離であり，$d_{ij} = \sum_{x=1}^{\rho} d_{ij}^{a_x}$ である．ここで，$d_{ij}^{a_x}$ は属性 a_x についての u_i と u_j の間の距離であり，a_x が連続値属性の場合，

$$d_{ij}^{a_x} = \frac{|v_{i,x} - v_{j,x}|}{max_{1 \leq i,j \leq n}(|v_{i,x} - v_{j,x}|)}$$

であり，a_x が離散値属性の場合，

$$d_{ij}^{a_x} = \begin{cases} 0 & if\ v_{i,x} = v_{j,x}, \\ 1 & otherwise, \end{cases}$$

である．

例 10.2.2 表 *10.1* の T_{ex} について，個人間の距離行列 $dist(T_{ex})$ を表 *10.2* に示す．$dist(T_{ex})$ は，各属性についての距離行列の和で求められる．

10.2.2 k-anonymity

一般的に，k-匿名化を行う際は，少なくとも k 個のレコードの値を同じに加工することによって，それらの個人の区別がつかないようにしている．複数個人のデータを等しく加工する手法として一般化やミクロアグリゲーションなどが挙げられるが，本章では一般化のみに注目し，以下のように加工を行う．また，k-匿名化の手法として，特定の属性の値の頻度が k 未満である特異な個人を削除する（行削除）ものがあるが，本章ではデータ中の個人は削除しないものとする．

定義 10.2.3 （一般化）（n_{ind} 人の個人 $u_1, \ldots, u_{n_{ind}}$ の見分けがつかないようにするとき，各個人が持つ）属性 $a_1, \ldots, a_\rho$ の値を加工する．a_x が連続値属性の場合，$v_{1,x}, \ldots, v_{n_{ind},x}$ の一般化を閉区間 $[min(v_{1,x}, \ldots, v_{n_{ind},x}), max(v_{1,x}, \ldots, v_{n_{ind},x})]$ とする．a_x が離散値属性の場合，$v_{1,x}, \ldots, v_{n_{ind},x}$ の一般化をそれらの値の和集合 $\{v_{1,x}, \ldots, v_{n_{ind},x}\}$ とする．

例 10.2.3 T_{ex} を一般化で加工した例として，表 *10.3, 10.4* に $T'_{ex\ 1}$ と $T'_{ex\ 2}$ を示す．$T'_{ex\ 1}$ は *Alice* と *Bob*，*Carol* と *David* のレコードが一般化によって等しくなり，区別がつかなくなっているため 2-匿名性を満たしている．また，$T'_{ex\ 2}$ は全てのレコードが一般化によって等しくなり，区別がつかなくなっているため 4-匿名性を満たしている．

定義 10.2.4 （データの加工コスト） 個人 u_i を加工したときのコスト $cost(u_i)$ は，一般化によって u_i と区別がつかなくなった全個人と u_i の距離の和であるとする．また，データ T の加工コスト $cost(T)$ は，全ての個人の加工コストの和であるとする．つまり，$cost(T) = \sum_{i=1}^{n} cost(u_i)$ である．

表 10.3: 加工データ例 $T'_{\text{ex 1}}$

仮名	age	sex
1	[10, 20]	{F, M}
2	[10, 20]	{F, M}
3	[40, 50]	{F, M}
4	[40, 50]	{F, M}

表 10.4: 加工データ例 $T'_{\text{ex 2}}$

仮名	age	sex
1	[10, 50]	{F, M}
2	[10, 50]	{F, M}
3	[10, 50]	{F, M}
4	[10, 50]	{F, M}

例 10.2.4 $T'_{ex\ 1}$ の加工コスト $cost(T'_{ex\ 1})$ について考える．*Alice* は一般化によって *Alice* と *Bob* の区別がつかない仮名 *1* に加工されているため，*Alice* の加工コスト $cost(Alice)$ は 0*(Alice* と *Alice* の距離*)*+1.25*(Alice* と *Bob* の距離*)*= 1.25 である．全ての個人の加工コストの和がデータの加工コストであるため，$cost(T'_{ex\ 1})$ は $1.25+1.25+1.25+1.25 = 5$ となる．また，$T'_{ex\ 2}$ の加工コスト $cost(T'_{ex\ 2})$ について考える．*Alice* は一般化によって全個人との区別がつかない仮名 *1* に加工されているため，*Alice* の加工コスト $cost(Alice)$ は $0+1.25+1.75+1 = 4$ であり，$cost(T'_{ex\ 2})$ は $4+3.5+3.5+4 = 15$ となる．

定義 10.2.5 （完全 k-匿名性）全ての個人が等しく $k-1$ 人の他の個人と区別がつかないことを保証する指標を，完全 k-匿名性とする．完全 k-匿名性を満たすようにデータを加工することを，完全 k-匿名化と呼ぶ.

例 10.2.5 $T'_{ex\ 1}$ は全ての個人が他の *1* 人の個人と区別がつかないデータであるため，完全 *2*-匿名化されたものである．また，$T'_{ex\ 2}$ は完全 *4*-匿名化されたものである．

10.2.3 k-concealment

k-anonymity を満たすデータと k-concealment を満たすデータは，どちらも「少なくとも k 人の区別がつかない」という条件を満たしているが，後者の方が加工コストが低くなる場合もある．本章では k-concealment を満たすようにデータを加工することを k-concealment 化と呼び，また，完全 k-匿名化と後述する完全 k-concealment 化について研究する．データを k-concealment 化するためには，まず加工の設計図として点（個人）と辺（個人間の対応関係）を持つ二部グラフを作成し，それに従って個人の情報を一般化すればよい.

定義 10.2.6 （完全 *k-concealment*）全ての点が等しく k 種類の辺の重複しない完全マッチングの辺（*match*）を持つデータを，完全 *k-concealment* と呼ぶ．完全 *k-concealment* を満たすようにデータを加工することを，完全 *k-concealment* 化と呼ぶ.

定義 10.2.7 （正解完全マッチング）同じ点同士を結ぶ辺を正解の辺とし，正解の辺のみで構成された完全マッチングを正解完全マッチングとする．完全 *k-concealment* 化を行うときに k 種類の完全マッチングを選ぶが，必ず一つの正解完全マッチングを含む.

例 10.2.6 図 *10.1, 10.2* に，T_{ex} の *4* 人の個人についての二部グラフ例 G_1, G_2 を示す．G_1 には *2* 種類の完全マッチング（赤実線，緑実線）があり，全ての個人が等しく *2* 本の *match* を持っている．こ

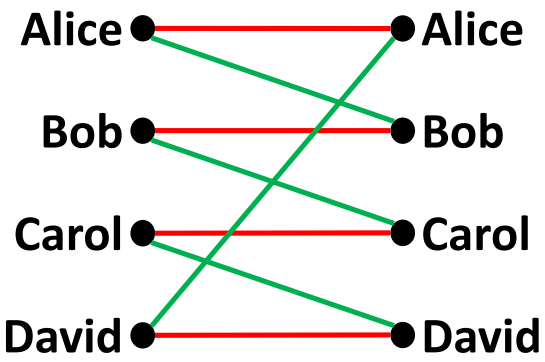


図 10.1: T_{ex} についての二部グラフ G_1

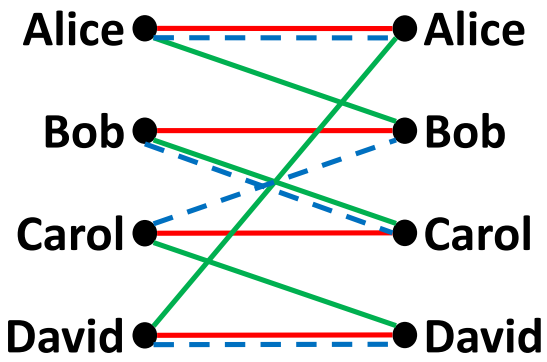


図 10.2: T_{ex} についての二部グラフ G_2

表 10.5: G_1 から加工した $T'_{\mathrm{ex}G_1}$

仮名	age	sex
1	[10, 50]	{F}
2	[10, 20]	{F, M}
3	[20, 40]	{M}
4	[40, 50]	{F, M}

表 10.6: G_2 から加工した $T'_{\mathrm{ex}G_2}$

仮名	age	sex
1	[10, 50]	{F}
2	[10, **40**]	{F, M}
3	[20, 40]	{M}
4	[40, 50]	{F, M}

の場合，赤実線の完全マッチングは正解完全マッチングである．G_2 は G_1 に *1* つの完全マッチング（青点線）が加えられたものであるが，既存の完全マッチングと一部の辺が重複しているので，*Bob* や *Carol* は *match* を *3* 本持っているが，*Alice* や *David* は *match* を *2* 本しかもっていない．

また，G_1 を基に加工した T'_{exG_1} と G_2 を基に加工した T'_{exG_2} を，表 *10.5, 10.6* に示す．例えば，G_1 の右側の *Alice* は左側の *Alice* と *David* との間に辺を持つので，*Alice* と *David* との区別がつかない仮名 *1* に一般化されている．T'_{exG_1} は T_{ex} を完全 *2-concealment* 化したものであり，T'_{exG_2} は T_{ex} を（不完全な）*2-concealment* 化したものである．二部グラフのみを見ると，G_2 は G_1 に辺を *1* 本足したものなので，その分加工の度合いが大きくなっている（仮名 *2* の *age* 属性）．

定義 10.2.8 （二部グラフの辺と加工コスト） *k-concealment* 化の加工コスト $cost(T)$ は，基になった二部グラフの辺の距離の総和である．

例 10.2.7 T'_{exG_1} の加工コスト $cost(T'_{exG_1})$ は，基にした G_1 の *8* 本の辺の距離の和であるため，$0+0+0+0+1.25+0.50+1.25+1=4$ である．また，G_2 は G_1 に左側の *Carol* から右側の *Bob* に向かう辺を *1* 本足したものなので，T'_{exG_2} の加工コスト $cost(T'_{exG_2})$ は $4+0.5=4.5$ である．

定義 10.2.9 （左右対称完全マッチング）正解の辺を含まない左右対称の完全マッチングを，左右対称完全マッチングと呼ぶ．

左右対称完全マッチング 1 種と正解完全マッチングを組み合わせた二部グラフをもとにすれば，データを完全 2-匿名化することができる．n が奇数のとき，左右対称完全マッチングは存在しない．

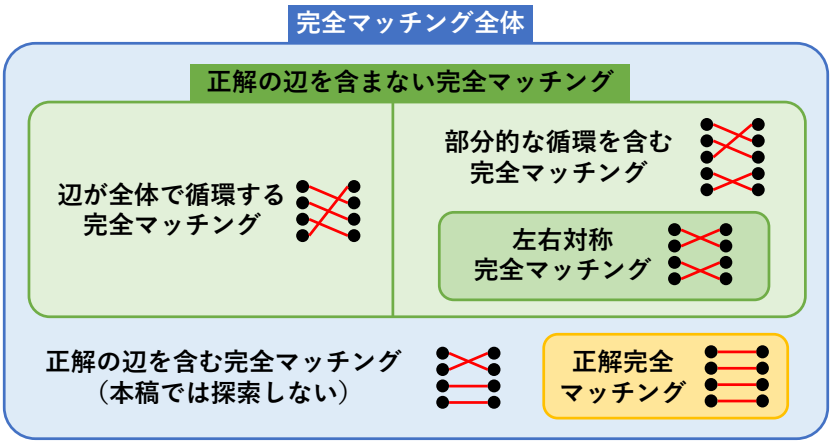


図 10.3: 完全マッチングのベン図

命題 10.2.1 （完全マッチングの種類数について） n 人の個人についての完全マッチングは全 $n!$ 通りあり，そのうち左右対称完全マッチングは $\prod_{i=1}^{n/2}(n-2i+1)$ 種類ある．また，$u_1 \to u_2 \to \cdots \to u_n (\to u_1)$ のように辺が全体で循環する完全マッチングは $(n-1)!$ 種類ある．

例 10.2.8 $n = 4$ の場合，完全マッチングは $4! = 24$ 通りあり，そのうち左右対称完全マッチングは $\prod_{i=1}^{4/2}(4-2i+1) = (4-2+1)(4-4+1) = 3$ 種類である．また，$n = 10$ の場合，完全マッチングは $10! = 3,628,880$ 通りあり，そのうち左右対称完全マッチングは $\prod_{i=1}^{10/2}(10-2i+1) = (10-2+1)(10-4+1)(10-6+1)(10-8+1)(10-10+1) = 945$ 種類である．

完全マッチングの分類（左右対称完全マッチングや，辺が全体で循環する完全マッチングなど）の関係を図 10.3 に示す．

定義 10.2.10 （完全マッチングの表記） n 人の個人についての完全マッチングを，置換 $P = (P[1], \ldots, P[n])$ と表記する．P には値 $\{1, \ldots, n\}$ が一度ずつ記録されており，$P[i] = j$ であるとき u_i から u_j に辺が張られている．

定義 10.2.11 *2* 種類の完全マッチング P_1, P_2 があり，$P_1[i] = P_2[i]$ であるとき，P_1 と P_2 は辺が重複している．P_1 と P_2 の辺が重複していることを $over(P_1, P_2)$ と表記する．

例 10.2.9 T_{ex} についての二部グラフ G_1（図 *10.1*）に注目する．この二部グラフには *2* 種類の完全マッチングが含まれており，$U = \{u_1 = Alice, u_2 = Bob, u_3 = Carol, u_4 = David\}$ とすると，赤実線のものは $P_1 = (1, 2, 3, 4)$，緑実線のものは $P_2 = (2, 3, 4, 1)$ と表記される．また，G_2 の青点線の完全マッチングは $P_3 = (1, 3, 2, 4)$ と表記できるが，$P_1[1] = P_3[1], P_1[4] = P_3[4], P_2[2] = P_3[2]$ であるため，この完全マッチングは前者 *2* つと辺が重複している．

定理 10.2.1 （完全 k-*concealment* 化の加工コストについて） 完全 k-*concealment* 化の加工コストは完全 k-匿名化の加工コスト以下である．

(証明) 完全 k-匿名化されたデータは完全 k-concealment であることを示す．完全 k-匿名なデータは k 人の個人 $u_1, \ldots, u_k$ が同じデータを持っており，区別がつかなくなっているため，これは二部グラフ上で $u_1, \ldots, u_k$ それぞれから $u_1, \ldots, u_k$ 全てに辺が張られている状態である．ここで，$P_1 = (1, 2, \ldots, k), P_2 = (2, \ldots, k, 1), \ldots, P_k = (k, 1, \ldots, k-1)$ という k 種類の完全マッチングを考える．これらの完全マッチングはいずれも辺が重複しておらず，全てを重ねると $u_1, \ldots, u_k$ それぞれから $u_1, \ldots, u_k$ 全てに辺が張られている二部グラフとなる．つまり，完全 k-匿名化されたデータの k 人の個人 $u_1, \ldots, u_k$ は，全て等しく k 本の match を持っているといえるため，完全 k-concealment も満たしている．よって，完全 k-匿名化されたデータの最小加工コストは完全 k-concealment 化されたデータの加工コストでもあるため，これが完全 k-concealment 化の最小加工コストを下回ることはない．(Q.E.D)

10.3 提案手法

加工コストが低い完全 k-concealment 化をするためには，辺が重複しないようにコストが低い完全マッチングを k 種類選んで二部グラフを作成し，それをもとにデータを加工すればよい．しかしながら，前述したように n 人の個人についての完全マッチングは全 $n!$ 通りあり，それらの加工コストを全て求めて最適解を見つけるのは計算量的に困難であるため，近似解を検討する．本章では，精度が高く加工コストが低い完全 k-concealment 化をするための近似アルゴリズムを提案する．

10.3.1 提案手法 1：貪欲法

貪欲法による提案手法のアルゴリズム 6 に示す．貪欲法はランダムな個人から最も近い個人へ辺を張って完全マッチング P を k 種類作成し，それらを記録したリスト L を出力するシンプルな手法である．定義 10.2.7 より，正解完全マッチングが必ず最初に選ばれていることを定義しているため，$k-1$ 回しか完全マッチングを作成していない点に注意せよ．また，この手法では張れる辺が無くなってしまう「詰み」が発生してしまう場合があることに注意せよ．

10.3.2 提案手法 2：くじ引き法

提案手法の一つであるくじ引き法をアルゴリズム 7 に示す．くじ引き法は完全マッチング P をランダムに t 回生成し，そのうちコストが最小のものを二部グラフに加えていく手法である．P が生成される際に，既存の完全マッチングと辺の重複が無いかを確認し，重複がある場合は再生成する．ループが進んでリスト L の要素が増えていくたびに，生成した P がチェックに通る確率が下がるため，k や t の値によって計算時間が増加していくことに注意せよ．また，完全マッチングの種類数は $n!$ であるため，個人数 n が増えるにつれてランダム生成によって低コストのマッチングを引く確率が下がり，加工コストも大きくなる．

Algorithm 6 提案手法 1：貪欲法

Input: $\mathrm{dist}(T), k$
  n：個人数
  d_{ij}：u_i と u_j の間の距離
  L：完全マッチングを記録するためのリスト
  $L[1] \leftarrow (1, 2, \ldots, n)$
  for i **in** $\{2, \ldots, k\}$ **do**
    P：n 次元の置換
    U_1, U_2：個人集合 $U_1 = U_2 = \{u_1, \ldots, u_n\}$
    while $|U_1| > 0$ **do**
      U_1 から，ランダムに u_x を選ぶ.
      $u_y = \arg\min_{u_y \in U_2} d_{xy}$
      $U_1 \leftarrow U_1 - \{u_x\}$, $U_2 \leftarrow U_2 - \{u_y\}$, $P[x] \leftarrow y$
    end while
    if $\forall_{j \in [1, i-1]} \overline{\mathrm{over}(P, L[j])}$ **then**
      $L[i] \leftarrow P$
    else
      i 番目のループをやり直す.
    end if
  end for
Output: L

10.3.3 提案手法 3：TSP 解法手法

巡回セールスマン問題の解を求める近似アルゴリズムを応用したものをアルゴリズム 8 に示す．TSP は，セールスマンが全ての都市を 1 回ずつ巡回する場合の最短経路を求める問題であり，NP 困難であることが知られている．私は，巡回経路を辺が循環する完全マッチングに置き換えられることに注目し，既存の TSP 解法を低コストの完全マッチングの探索に応用した．

例えば，T_{ex} についての二部グラフ G_1（図 10.1）に注目する．G_1 の緑実線の完全マッチングでは，辺が Alice → Bob → Carol → David → Alice と循環している．このような辺が循環する完全マッチングを都市の巡回経路に，個人間の距離を都市間の距離に置き換えることによって，TSP 解法を用いて低コストの完全マッチングを高速に探索することができる．しかし，この手法によって検索できるのは，完全マッチング全体（$n!$ 種類）のうち，辺が循環する完全マッチング $(n-1)!$ 種類のみであることに注意せよ．

本章では，R 言語 [8] の TSP パッケージ [9] を用いて TSP 解法手法を実装した．TSP の近似アルゴリズムには，TSP パッケージに含まれている表 10.7 の 9 種類の手法を用いる．k 回の探索 1 回ごとに，全ての手法を用いて完全マッチングを作成し，最もコストの低いものを採用している．また，コストの低い巡回経路を見つけたとき，その逆回りの経路も同様にコストの低い経路であることを利用し，計算時間短縮のために検索を一部で省いている．

Algorithm 7 提案手法 2：くじ引き法

Input: $\mathrm{dist}(T), k$
  n：個人数
  d_{ij}：u_i と u_j の間の距離
  L：完全マッチングを記録するためのリスト
  t：試行回数
  count：成功回数
  $L[1] \leftarrow (1, 2, \ldots, n)$
  for i **in** $\{2, \ldots, k\}$ **do**
    for count **in** $(1, \ldots, t)$ **do**
      ランダムに完全マッチング P を生成する．
      if $\forall_{j \in [1, i-1]} \overline{\mathrm{over}(P, L[j])}$ **then**
        $P_{\mathrm{count}} \leftarrow P$
      end if
    end for
    $P_{\min}$：$P_1, \ldots, P_t$ のうち，コストが最小であるもの
    $L[i] \leftarrow P_{\min}$
  end for
Output: L

10.3.4 提案手法 2,3+クラスタリング

提案手法 2 のくじ引き法には，n が増えるにしたがって加工コストが大きく増加してしまう問題点があり，提案手法 3 の TSP 解法手法には，完全マッチング全 $n!$ 種類のうち，辺が循環する完全マッチング $(n-1)!$ 種類だけしか探索できないという問題点があった．そこで私は，問題を小規模の部分問題に分解して効率よく解き，部分解を統合して大きな解を得るという分割統治法の考えに基づき，部分問題への分解にクラスタリングを用いてこれらの問題点の解決を試みる．

アルゴリズム 9 にクラスタリングを組み合わせる方法を示す．n 人の個人をクラスタリングによって c 個のデータに分割し，各データについて提案手法 2 や 3 で完全マッチングリスト $L_1, \ldots, L_c$ を出力し，最後にこれらを結合してリスト L を作成する．データを分割することによって，疑似的に完全マッチング種類数 $n!$ を減らすことができるので，提案手法 2 で低コストのマッチングをひける確率が高くなり，加工コストを下げることができる．また，クラスタリングと提案手法 3 を組み合わせることによって，TSP 解法単体では探索できない完全マッチングも作成できるようになる．本章では，類似度には前述した個人間の距離，クラスタリング手法にはウォード法 [10] を用いている．

10.4 評価実験

10.4.1 データセット

本章では，2 つのデータセット（疑似人流データ，世帯収入データ）を用いて提案手法 1,2,3 や提案手法 + クラスタリングの評価を行う．

疑似人流データ [11] はナイトレイ社より公開されている位置情報データである．このデータは都市圏周辺の数千人分の人流を，SNS の地域解析に基づいて疑似的にデータ化したものであり，緯度や

Algorithm 8 提案手法 3：TSP 解法手法

Input: $\mathrm{dist}(T), k$
  n：個人数
  L：完全マッチングを記録するためのリスト
  $L[1] \leftarrow (1, 2, \ldots, n)$
  for i **in** $\{2, \ldots, k\}$ **do**
    if i が偶数 **then**
      L の各要素の置換と辺の重複がないように，9 種類の TSP 解法で完全マッチング P を作成し，コストの最も低いものを P_i とする．
    else
      ひとつ前のループで作成された P_{i-1} の逆回りの完全マッチングを作成し，P_i とする．
    end if
    $L[i] \leftarrow P_i$
  end for
Output: L

Algorithm 9 提案手法 + クラスタリング

Input: $\mathrm{dist}(T), k$
  n：個人数
  L：完全マッチングを記録するためのリスト
  各クラスタの要素数が k を下回らないようにクラスタリングを行い，n 人の個人を c 個のデータ $T_1, \ldots, T_c$ に分ける．
  for i **in** $\{1, \ldots, c\}$ **do**
    $k, \mathrm{dist}(T_i)$ をアルゴリズム 7, または 8 に入力し，出力としてリスト L_i を得る．
  end for
  for j **in** $\{1, \ldots, k\}$ **do**
    $L_1[j], \ldots, L_c[j]$ を結合し，$L[j]$ に記録する．
  end for
Output: L

経度などの 9 つの属性が含まれている．疑似人流データには各個人の 5 分刻みの位置情報が記録されているため，レコード数 ≠ 個人数のデータであるが，0 時 0 分のレコードを用いることにより，定義 10.2.1 を満たすように処理している．また，本章ではランダムに抽出した個人 100 人と，緯度と経度の連続値属性 2 つのみのデータ $T_{\text{人流}}$ を用いる（$n = 100, \rho = 2$）．

世帯収入データには，UCI より公開されている Adult Data Set[12] を用いる．このデータは国勢調査によって作成された 32,561 レコードのデータであり，職種などの離散値属性 9 つと，年齢などの連続値属性 6 つを含んでいる．Adult Data Set はレコード数 m= 個人数 n のデータであり，ランダムに抽出した個人 1,000 人分の全ての属性を用い，これを $T_{\text{世帯}}$ とする（$n = 1,000, \rho = 15$）．

10.4.2 実験方法

実験 1：$T_{\text{人流}}$ を用いた手法比較

実験 1 では，提案手法 1,2,3 に加え，提案手法 2,3 にクラスタリングを加えた計 5 手法の完全 2-concealment 化手法と，貪欲法を用いた完全 2-匿名化手法の比較を行う．これら 6 種類の手法で加工

表 10.7: TSP 近似アルゴリズムの概要

ID	手法名	概要
1	identity	巡回経路を ID 順に出力する
2	random	巡回経路をランダムに出力する
		ある都市から始まる巡回経路に,
		まだ含まれていない都市を挿入して巡回経路を作っていく
3	nearest insertion	Nearest：経路上のある都市に最も近い都市を挿入する.
4	cheapest insertion	Cheapest：経路の増加具合が最も少ない都市を挿入する.
5	farthest insertion	Farthest：経路上のある都市に最も遠い都市を挿入する.
6	arbitrary insertion	Arbitrary：経路に含まれていない都市をランダムに挿入する.
7	nn	ランダムな都市から経路を始め,
		経路の終端の都市に最も近い都市を後ろに追加していく
8	repetitive nn	Repetitive：全ての都市をスタート地点として計算を行い,
		最も距離が短い経路を返す.
9	two opt	経路の一部の辺を交換 (順番を逆にする) して
		距離を短くしていく手法

された $T_{\text{人流}}$ の加工コストを求め，性能の評価を行う.

また，提案手法の加工度合いの可視化も試みる．$T_{\text{人流}}$ の 100 人の個人の位置情報（緯度と経度）を図 10.4 に示す．$T_{\text{人流}}$ のような位置情報を一般化で加工すると，後述する図 10.5 のように個人の位置情報が点から四角形の範囲になる．例えば，図 10.4 中の左下に位置する 2 人の個人の区別をつかないようにすると，この 2 点を対角の頂点とする四角形に一般化してやればよく，この四角形の面積が小さいほど有用性が高いといえる.

実験 2：$T_{\text{世帯}}$ を用いた手法比較

実験 2 では，提案手法 2,3 とクラスタリングを組み合わせた計 4 手法で $T_{\text{世帯}}$ を完全 k-concealment 化し，それらの性能の比較を行う．この実験では，$T_{\text{世帯}}$ を $k = 2, \ldots, 7$ で完全 k-concealment 化したときの加工コストを評価する.

10.4.3 実験結果

実験 1 の結果

6 つの加工データのコストを表 10.8 に示す．$T_{\text{人流}}$ を $k = 2$ で加工する場合，提案手法 2+クラスタリングが最もコストが低くなった．また，提案手法 1,2,3 単体はいずれも完全 2-匿名化に加工コストで劣っているが，クラスタリングを組み合わせることによって提案手法 2,3 は完全 2-匿名化より有用性の高いデータを生成した.

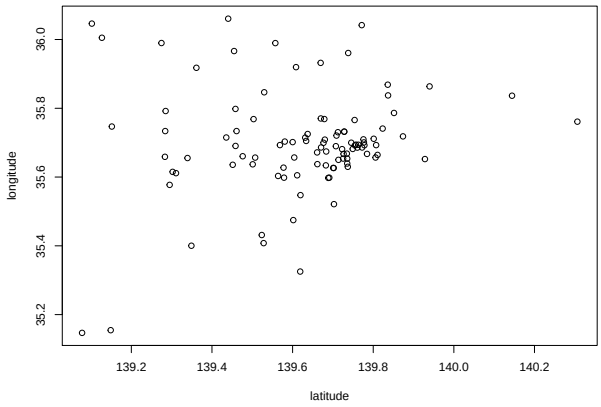


図 10.4: $T_{\text{人流}}$ の位置情報

表 10.8: $T_{\text{人流}}$ の加工コスト（$k = 2$）

加工手法	加工コスト
完全 2-匿名化（貪欲法）	0.2145
提案手法 1（貪欲法）	0.2328
提案手法 2（くじ引き法）	3.1490
提案手法 2+クラスタリング	**0.1493**
提案手法 3（TSP 解法手法）	0.3186
提案手法 3+クラスタリング	0.1640
LAP Solver[106] の応用	**0.1106**

まず，貪欲法での完全 2-匿名化と完全 2-concealment 化の比較を行う．図 10.5, 10.6 に，$T_{\text{人流}}$ を貪欲法にて完全 2-匿名化した結果と完全 2-concealment 化した結果を示す．どちらも個人 2 人の区別がつかないデータであるが，前者の加工コストは 0.2145，後者の加工コストは 0.2328 であり，貪欲法では完全 2-匿名化の方が有用性が高くなった．

次に，提案手法 2（くじ引き法，$t = 100,000$）のクラスタリングの有無で比較を行う．図 10.7, 10.8 に，$T_{\text{人流}}$ をクラスタリング無しで加工した結果とクラスタリング有りで加工した結果を示す．提案手法 2 のみでは一般化による四角形の面積が大きくなっている一方で，クラスタリングをした場合では距離が近い個人が一般化されていることがわかる．前者の加工コストは 3.149，後者の加工コストは 0.1493 であった．

最後に，提案手法 3（TSP 解法手法）のクラスタリングの有無で比較を行う．図 10.9, 10.10 に，$T_{\text{人流}}$ をクラスタリング無しで加工した結果とクラスタリング有りで加工した結果を示す．提案手法 3 のみでは，図 10.9 のようにデータ全体を循環する経路を探してしまうため，四角形の面積は大きくなってしまい，加工コストは 0.3186 である．一方，クラスタリングを組み合わせた場合は加工コストが 0.1640 まで下がっている．

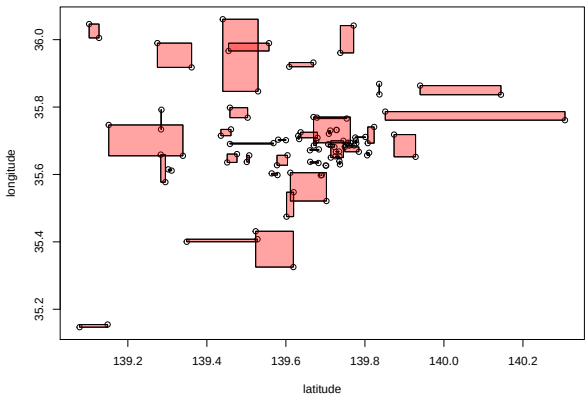


図　10.5:　完全 2-匿名化された $T_{\text{人流}}$（$\text{cost}(T_{\text{人流}}) = 0.2145$）

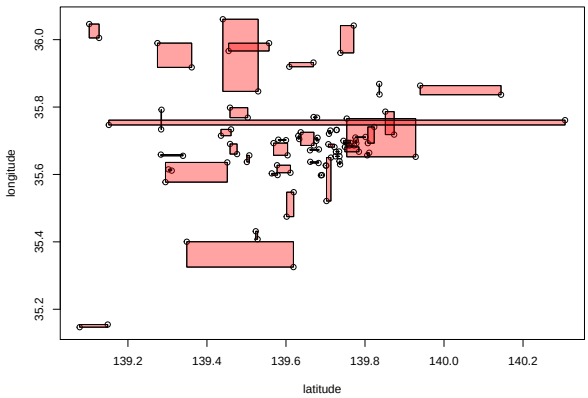


図　10.6:　提案手法 1 で加工された $T_{\text{人流}}$（$\text{cost}(T_{\text{人流}}) = 0.2328$）

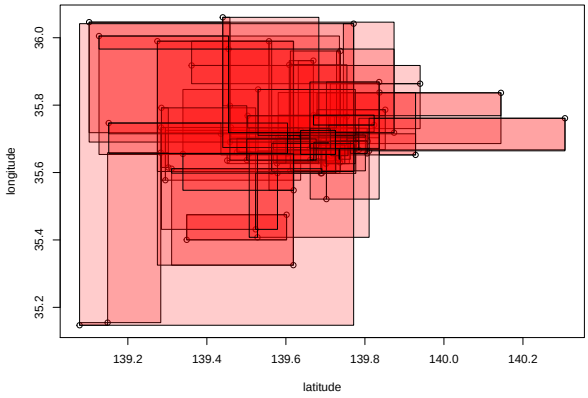


図　10.7:　提案手法 2 で加工された $T_{\text{人流}}$（$\text{cost}(T_{\text{人流}}) = 3.1490$）

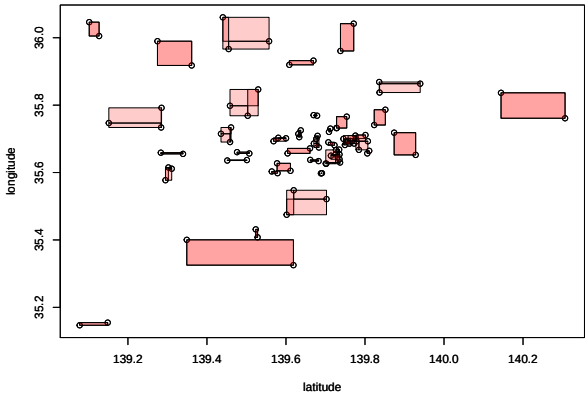


図 10.8: 提案手法 2+クラスタリングで加工された $T_{\text{人流}}$（$\text{cost}(T_{\text{人流}}) = 0.1493$）

実験 2 の結果

表 10.9 に，4 つの手法の加工コストを示す．まず，提案手法 2（$t = 10,000$）の性能評価を行う．例えば，提案手法 2 のみで完全 2-concealment 化を行うと加工コストは 5535.23 であるが，クラスタリングを組み合わせると加工コストが 1819.31 まで減少する．

次に，提案手法 3 の性能評価を行う．$T_{\text{世帯}}$ の場合，提案手法 3 は提案手法 2 よりも全ての k で加工コストが優れている．また，$T_{\text{人流}}$ の場合とは異なり，クラスタリングをしない方が有用性の高い加工ができており，例えば $k = 7$ のときはクラスタリングをすることによって加工コストが 1.07 倍ほどに増加している．

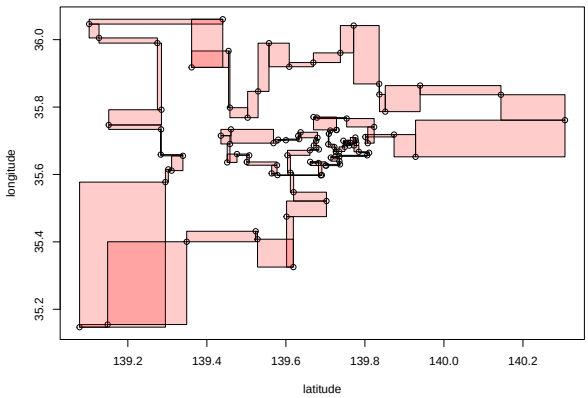


図 10.9: 提案手法 3 で加工された $T_{\text{人流}}$（$\text{cost}(T_{\text{人流}}) = 0.3186$）

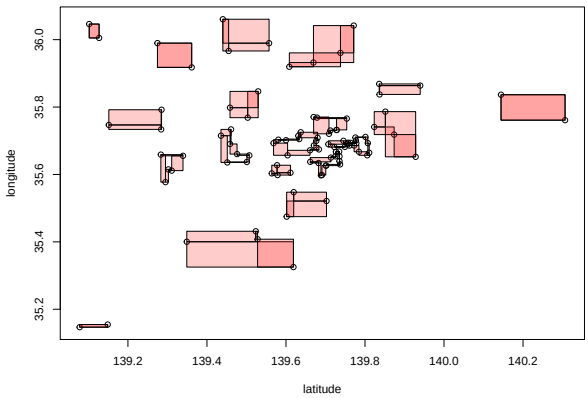


図 10.10: 提案手法 3+クラスタリングで加工された $T_{\text{人流}}$（$\text{cost}(T_{\text{人流}}) = 0.1640$）

表 10.9: $T_{\text{世帯}}$ の加工コスト（$k = 2, \ldots, 7$）

k	提案手法 2	+クラスタリング	提案手法 3	+クラスタリング	LAP Solver
2	5535.23	1819.31	1683.76	1718.13	1411.51
3	11066.13	4083.07	3367.52	3443.10	3122.19
4	16600.02	7413.35	5416.70	5656.75	4978.36
5	22153.13	10041.37	7465.88	7858.74	6962.03
6	27682.26	12741.39	9718.28	10364.58	9054.63
7	33205.50	15973.04	11970.68	12826.53	11230.81

10.4.4 考察

$T_{\text{人流}}$ を用いた実験 1 の結果では，クラスタリングを組み合わせたほうが提案手法 2,3 ともに加工コストが低くなっていたが，$T_{\text{世帯}}$ を用いた実験 2 の結果では，クラスタリングをしない方が提案手法 3 の加工コストが低くなった．この原因として，データによる個人間の距離の分布の違いが考えられる．図 10.11,10.12 に，各データの個人間距離の分布を示す．図からわかるように，$T_{\text{人流}}$ には近い距離に個人が多く分布しており，クラスタリングをすることによって局所的な最適解を求めやすくなると考える．一方，$T_{\text{世帯}}$ は属性数が多いためか，個人間の距離が大きく，このようなデータに対してはクラスタリングは効果が薄いと考えられる．

また，完全 k-concealment 化と普通の k-concealment 化の違いについて考える．普通の k-concealment 化で加工されたデータでは，区別のつかない最も小さいグループの大きさが k 人であり，$k+1$ 人以上のグループも存在する．しかし，このデータ中の個人が識別されるリスクには差があり，例えば k 人のグループに属している個人は，$2k$ 人のグループに属している個人よりも，識別されるリスクが 2 倍高いといえる．匿名化データ内の個人によって識別されるリスクに差があるのは，その当人達からすると公平ではないと感じるであろう．それに対して，完全 k-concealment 化された匿名化データでは，全ての個人が同じリスクを持つ（他の $k-1$ 人と区別がつかない）ようになるため，公平性

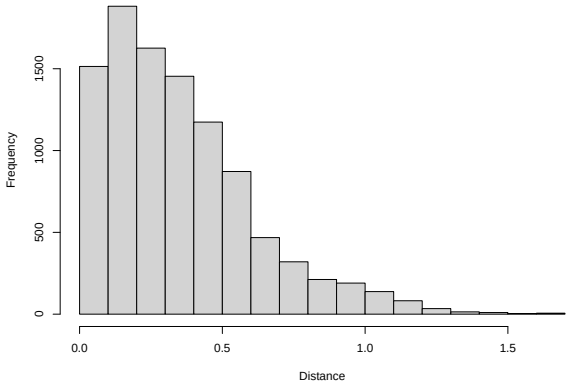


図 10.11: $T_{\text{人流}}$ における個人間距離の分布

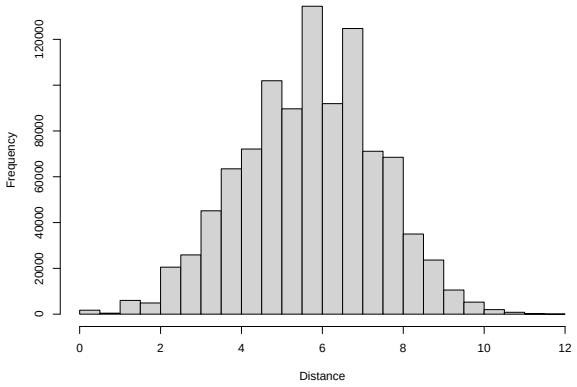


図 10.12: $T_{\text{世帯}}$ における個人間距離の分布

が高い．さらに，k-匿名化によって生じる過度な加工を完全に解消しているので，加工データの有用性も普通の k-concealment 化をされたデータよりも高くなりうる．これらのことより，私は普通の k-concealment 化よりも完全 k-concealment 化に着目している．

10.4.5 LAP Solver を用いた手法

線形和割り当て問題 (LSAP：linear sum assignment problem) のソルバーである python の LAP Solver ライブラリ [106] を用いて，コストが低い完全マッチングの探索を試みる．LSAP は,「グループ A の要素をグループ B の要素のどれに割り当てれば最も効率がよくなるか？」という問題であり，本研究では「元データの各個人を加工データのどの仮名に割り当てれば最もコストが低くなるか？」と言い換えることができる．図 10.13 に，LAP Solver を用いて完全 2-concealment 化された $T_{\text{人流}}$ を示す．この時の加工コストは 0.1106 であり，これは表 10.8 に示すように，図 10.5〜図 10.10 で示したどの加工データよりもコストが低い．また，$T_{\text{世帯}}$ を LAP Solver を応用して完全 2〜7concealment 化した際の加工コストを表 10.9 に示す．LAP Solver による加工データは，いずれの k でも提案手法より加工コストが低かった．

10.5 まとめ

本章では，レコード数と個人数が等しいデータを完全 k-concealment 化する研究を行った．完全 k-concealment 化を行うためには，データ中の個人を点とした二部グラフを加工の設計図に見立て，その中に辺の重複しない k 本の完全マッチングを作成する必要がある．そこで，私は辺の重複しない k 本の完全マッチングを作成する 3 つのアルゴリズム（1:貪欲法，2:くじ引き法，3:TSP 解法手法）を提案した．

また，100 人分の人流データと 1,000 人分の世帯収入データを用いて提案手法の性能評価を行ったところ，以下の結果が得られた．(1) 人流データを $k = 2$ で加工するとき，提案手法 2,3 とクラスタリングを組み合わせると，完全 2-匿名化されたデータよりも有用性の高い加工をすることができる．

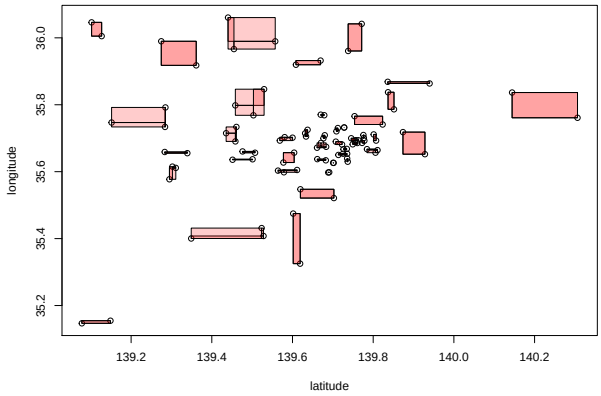


図 10.13: LAP Solver によって完全 2-concealment 化された $T_{\text{人流}}$（$\text{cost}(T_{\text{人流}}) = 0.1106$）

(2) 世帯収入データを完全 k-concealment 化するときは，提案手法 3 のみの加工が最もコストが低く，クラスタリングを組み合わせるとコストが 1.07 倍まで増加する．

本章で提案した完全 k-concealment 化手法は，レコード数と個人数が等しいならばどのようなデータに対しても適応可能であるが，なかでも近い距離に多くの個人が分布しているデータに対しては，クラスタリングを組み合わせると加工コストの向上が期待できる．

第11章　おわりに

本研究の目的は，データに対する匿名化の影響を明らかにすることである．研究目的達成のために，(課題 1) 既存の攻撃者モデルの問題点，(課題 2) 履歴データの識別リスクの問題点，(課題 3)k-anonymity の問題点，(課題 4) 実験データへの依存性，という 4 つの課題を設定し解決した．

各章の主な結論を述べる．1 章では，本研究の背景，研究目的，既存の研究における問題点と，それに対する本研究のアプローチと新規性を述べた．2 章では，匿名化技術における基礎定義と主要ないくつかの先行研究を解説した．

3 章では，課題 1 を解決するために，新たな攻撃者モデルを提案した．データ中のある一つの値を確率的に得る攻撃者モデルを提案し，その攻撃者の平均識別確率によってデータ中のどの属性が危険であるかを評価した．また，平均識別確率を近似する 3 つの数理モデル（平均モデル，最小コストモデル，サンプリングモデル）を提案し，これらを用いて 4 つのデータ（購買履歴データ，糖尿病患者データ，世帯収入データ，ローン借り入れデータ）に対して安全性の理論的な評価結果を示した．その結果，購買履歴データの時刻属性の値 1 つのみから平均 32%の確率で個人が識別されることや，データによっては提案モデルでも精度よくリスクを評価できることなどを明らかにした．

4 章では，課題 2 を解決するために，履歴データのふるまいを数理モデル化した．提案データモデルでは，履歴データの値が一様に生起する仮定の下，履歴データ中に登場する項目の種類数の確率分布とその期待値が与えられる．また，提案モデルを応用することにより，履歴データを k-匿名化するために必要なダミーレコード数の期待値を，元データの統計量やパラメータ k などから求めることができることを示した．さらに，匿名化の加工コストを理論的に評価した結果に基づき，従来は加工前に算出することが困難であった k-匿名化するデータに対する k の最適値を算出した．

5 章では，課題 4 を解決するために，購買履歴に対する匿名化の影響を実験的に評価した．購買履歴データの個人が購買商品特徴から識別されるリスクを想定し，そのリスクへの耐性を高めるためにかかる加工コストを実験的に評価した．その結果，購買履歴データを 50 個のクラスタに分割して k-匿名化をするためには，約 18 万のダミーレコードの追加が必要であることなどを明らかにした．

6 章では，課題 4 を解決するために，健康診断データと傷病/医薬品レセプトデータを匿名化することによって，データの安全性と有用性がどのように変化するかを実験的に評価した．その結果，病歴/処方歴を k-匿名化することによって，識別される人数の割合が平均 2.9%まで減少すること，高血圧に対する相対リスクが相対誤差で 0.073 しか変化しないことなどを明らかにした．

7 章では，課題 4 を解決するために，複数用途の含まれるデータから個人が識別されるリスクを実験的に評価した．乗降履歴や購買履歴などの複数用途の履歴が含まれている交通 IC カードデータから個人が識別されるリスクを，エントロピーを用いて定量化した．その結果，個人を識別できる確率が 1 つの乗降履歴によって 3.3%から 28.4%まで上がることや，購買履歴と乗降履歴を 1 つずつ知られた場合は，識別率が 88.1%まで上がることなどを明らかにした．

8 章では，課題 4 を解決するために，世帯収入に対する匿名化の影響を実験的に評価した．世帯支出データの個人がレコード間のユークリッド距離から識別されるリスクを想定し，そのリスクへの耐性を高めるための加工手法を検討した．その結果，ノイズ付加や値のスワップのような単純な摂動化では再識別を全く防げないことや，k-匿名化によって再識別率を 17%まで下げられることなどを明らかにした．

9 章では，課題 3 を解決するために，履歴データに対する k-concealment 化手法を提案した．Tamir らが提案した k-concealment 指標は，レコード数が個人数より多い履歴データは想定されていなかった．本章では，仮名の一般化とレコード間の k-concealment という新しい方式を提案し，個人によってレコード数の異なる履歴データを k-concealment を満たすように加工する手法を研究した．提案手法を用いることにより，顧客やレコードを追加/削除することなく，履歴データを最低でも k 人の区別がつかない状態にすることができる．

10 章では，課題 3 を解決するために，新たな完全 k-concealment 化アルゴリズムを提案した．データを全ての個人が等しく k 人と区別がつかない状態に加工する完全 k-concealment 化に着目し，コストの低い加工をするために巡回セールスマン問題の近似解法やクラスタリングを応用したアルゴリズムを提案した．提案手法によって，レコードの追加や削除をせずに，全ての個人の危険度が等しい公平な匿名化データを作成することができる．

これらの結果より，本研究を結論づける．本研究の貢献は以下の 3 つである．（**貢献１：匿名化による安全性・有用性変化の理論的評価**）私は，攻撃者や履歴データをモデル化することにより，データの安全性や有用性を理論的に評価し，本研究の課題 1,2 を解決した．（**貢献 2：匿名化による安全性・有用性変化の実験的評価**）多種多様なデータセットに対する識別リスクを想定し，それらを匿名化した際の影響を実験的に評価することにより，本研究の課題 4 を解決した．（**貢献 3：新たな匿名化手法の提案**）k-匿名性を改善した k-concealment 指標に注目し，これを満たすための新たな匿名化手法を提案することにより，本研究の課題 3 を解決した．

このように本研究では，4 つの課題を解決することによって，データに対する匿名化の影響を明らかにした．

業績

学術論文誌

1. Satoshi Ito, Hiroaki Kikuchi, “Estimation of cost of k-anonymity in the number of dummy records”, Journal of Ambient Intelligence and Humanized, Springer, 採択済み.

2. 伊藤聡志，池上和輝，菊池浩明, “健康診断データとレセプトデータの匿名加工情報を用いた疾病リスク分析”, 情報処理学会論文誌, 情報処理学会, 62 巻 9 号, pp.1560-1574, 2021 年 9 月.

3. Satoshi Ito, Reo Harada, Hiroaki Kikuchi, “De-identification for Transaction Data Secure against Re-identification Risk Based on Payment Records”, Journal of Information Processing, Volume 28, Pages 511-519, September 2020.

4. Ito,S., Kikuchi,H. and Nakagawa,H., “Attacker models with a variety of background knowledge to de-identified data”, Journal of Ambient Intelligence and Humanized Computing, Springer, July 2019.

国際会議投稿論文

1. Hiroaki Kikuchi, Atsuki Ono, Satoshi Ito, Masahiro Fujita, Tadakazu Yamanaka, “Web Crawler for an Anonymously Processed Information Database”, Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS 2021), Lecture Notes in Networks and Systems, vol 279, pp. 501-510, Springer, June 2021.

2. Masahiro Fujita, Yasuoki Iida, Mitsuhiro Hattori, Tadakazu Yamanaka, Nori Matsuda, Satoshi Ito, Hiroaki Kikuchi, “Proposal and Development of Anonymization Dictionary Using Public Information Disclosed by Anonymously Processed Information Handling Business Operators”, Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS 2021), Lecture Notes in Networks and Systems, vol 279, pp. 30-39, Springer, June 2021.

3. Satoshi Ito, Reo Harada and Hiroaki Kikuchi, “De-identification for Transaction Data Secure against Re-identification Risk Based on Payment Records”， The 16th International Conference on Modeling Decisions for Artificial Intelligence (MDAI-2019)，USB proceedings, Italy, pp. 158-169, September 2019.

4. Satoshi Ito, Hiroaki Kikuchi and Hiroshi Nakagawa, “Attacker Models with a Variety of Background Knowledge of Payment History”, The 15th International Conference on Modeling Decisions for Artificial Intelligence (MDAI-2018), USB proceedings, Spain, pp. 178-189, October 2018.

5. Satoshi Ito, Reo Harada and Hiroaki Kikuchi, “Risk of Re-identification from Payment Card Histories in Multiple Domains”, 2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA-2018), IEEE, Poland, pp. 934-941, May 2018.

6. Satoshi Ito and Hiroaki Kikuchi, “Risk of Re-identification Based on Euclidean distance in Anonymized Data PWSCUP2015”, Advances in Network-Based Information Systems (NBiS 2017), Lecture Notes on Data Engineering and Communications Technologies, vol 7. pp. 901-913, Springer, August 2017.

国内研究会投稿論文

1. 伊藤聡志，菊池浩明，“完全 k-concealment 匿名化を求める精度の高いアルゴリズムの評価”, Computer Security Symposium 2021 (CSS-2021), pp.1045-1052, オンライン開催, 2021 年 10 月.

2. 伊藤聡志，池上和輝，菊池浩明，“匿名加工情報の応用 (1): 健康診断データとレセプトデータの分析とプライバシーリスク評価”, Computer Security Symposium 2020 (CSS-2020), pp. 1222-1229, オンライン開催, 2020 年 10 月.

3. 池上和輝，伊藤聡志，菊池浩明，“匿名加工情報の応用 (2)：各種傷病を予測する健康診断モデル”, コンピュータセキュリティシンポジウム 2020 (CSS2020), pp. 1230-1237, 2020 年 10 月.

4. 伊藤聡志，菊池浩明，“履歴データの数理モデルの提案と k-匿名化に必要なダミーレコード数推定への応用”，第 88 回コンピュータセキュリティ研究発表会 （CSEC-88）, オンライン開催, 2020 年 3 月.

5. 金子侑紀, 小野敦樹, 伊藤聡志, 菊池浩明, 服部充洋, 飯田泰興, 藤田真浩, 山中忠和, “匿名加工情報取扱事情者を調査するクローラーシステムの開発”, 情報処理学会第 82 回全国大会, pp.3_447-3_448, 2020 年 3 月.

6. 伊藤聡志，菊池浩明，“履歴データに対する匿名化モデル k-concealment の改良手法の提案”, Computer Security Symposium 2019 (CSS-2019), pp.1477-1484, 2019 年 10 月.

7. 伊藤聡志，菊池浩明，“攻撃者の平均識別確率を用いた匿名加工情報の再識別リスク評価モデルの提案と評価”，第 84 回コンピュータセキュリティ研究発表会（CSEC-84）, 名古屋, 2019 年 3 月.

8. 小林祐貴, 中村幸輝, 伊藤聡志, 菊池浩明, “一般化匿名加工された購買履歴データの RFM 分析有用性評価”, 情報処理学会第 81 回全国大会, pp.3_425-3_426, 2019 年 3 月.

9. 伊藤聡志, 菊池浩明, 中川裕志, “背景知識の違いによる匿名加工データの攻撃者モデルの分類と評価”, 情報処理学会, コンピュータセキュリティシンポジウム 2017（CSS-2017）, pp. 136-142, 2017 年 10 月.

10. 伊藤聡志, 原田玲央, 菊池浩明, 乗降と物販履歴データの識別リスク分析と匿名加工の検討, 第 76 回コンピュータセキュリティ研究発表会 （CSEC-76）, pp. 1-8, 2017 年 3 月.

11. 原田玲央, 伊藤聡志, 菊池浩明, “商品の特徴による再識別リスクとクラスタリングを用いた購買履歴データ匿名加工手法の提案”, 電子情報通信学会, 暗号とセキュリティシンポジウム（SCIS-2017）, pp. 1-8, 2017 年 1 月.

12. 伊藤聡志, 菊池浩明, “ユークリッド距離を用いた再識別手法と PWSCup2015 の匿名加工データを用いた評価”, 情報処理学会, 第 73 回コンピュータセキュリティ研究発表会（CSEC-73）, pp. 1-8, 2016 年 5 月.

その他

1. 伊藤聡志, 原田玲央, 菊池浩明, “［招待講演］複数用途からなる交通 IC カードデータの再識別リスク分析 (from AINA 2018)”, 信学技報, vol. 119, no. 40, ISEC2019-9, pp.45-46, 2019 年 5 月.

2. 匿名加工・再識別コンテスト PWSCUP2018 総合 3 位

3. 匿名加工・再識別コンテスト PWSCUP2016 再識別賞

4. 2016 年度キャンドルスターセッション (CSSx2.0) 1 等星

謝辞

本稿は多くの方々のご指導・ご協力なくしては，完成しえないものである．指導教官である明治大学総合数理学部の菊池浩明教授からは，著者が学部 1 年生の頃から現在までの 9 年間，多大なるご指導を賜った．両親や祖父母は著者の学生生活を金銭的かつ精神的に，大いに支えてくれた．菊池研究室の同期や後輩たちは，著者の学生生活を楽しく，かけがえのないものにしてくれた．共同研究者である中川裕志先生，原田玲央君，池上和輝君には，それぞれ本稿の一部である攻撃者モデルの研究，購買履歴データの研究，健康診断データの研究を進めるにあたり，大いに協力していただいた．明治大学総合数理学部の斉藤裕樹教授と，統計数理研究所の南和宏教授には，本稿を改善するためのご助言をいただいた．明治大学総合数理学部の先生方や，静岡大学の西垣正勝教授や大木哲史先生，東京電機大学の稲村勝樹先生をはじめとする他大学の先生方，他研究室の学生の皆さんからは，合同発表会などを通じて，研究についての多くのご意見をいただいた．菊池研究室の社会人ドクターの新原功一氏，重本倫宏氏，仲小路博史氏，山口通智氏，ポスドクの黄緒平氏，馬瑞強氏は，著者に研究者としてあるべき姿を示してくれた．学部の 1 期生である著者にとって，静岡大学の藤田真浩氏をはじめとする他大学の先輩方の存在は非常にありがたく，大いに面倒を見ていただいた．研究会やコンテスト等の際には，企業や研究所の方々から，研究や進路についての様々な助言を賜った．著者の研究と学生生活を支えていただいた全ての方々に，心から感謝いたします．

参考文献

[1] Rocher L. et al., “Estimating the success of re identifications in incomplete datasets using generative models”, Nature Communications, 2019.

[2] L. Sweeny, “k-anonymity: a model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), pp.557–570. (2006)

[3] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-Diversity: Privacy beyond k-anonymity”, In International Conference on Data Engineering (ICDE), page 24. (2006)

[4] T. Truta, A. Campan, and P. Meyer, “Generating microdata with p-sensitive k-anonymity property”, In Secure Data Management (SDM), pages 124–141. (2007)

[5] Tamir Tassa, Arnon Mazza, Aristides Gionis, “k-Concealment: An Alternative Model of k-Type Anonymity”, TRANSACTIONS ON DATA PRIVACY 5, pp. 189–222. (2012)

[6] 濱田浩気，荒井ひろみ，小栗秀暢，菊池浩明，黒政敦史，中川裕志，西山賢志郎，波多野卓磨，村上隆夫，山岡裕司，山田明，渡辺知恵美,「PWS Cup 2018: 匿名加工再識別コンテストの設計 〜履歴データの一般化・再識別〜」，コンピュータセキュリティシンポジウム（CSS 2018），pp.935–940.（2018)

[7] 高橋磐郎，藤重悟,「離散数学」，岩波講座情報科学 17，pp.147–162.（1981)

[8] The R Project for Statistical Computing, https://www.r-project.org/, 2021 年 8 月 6 日参照.

[9] Michael Hahsler, Kurt Hornik, “Package TSP”, https://cran.r-project.org/web/packages/TSP/TSP.pdf, 2021 年 8 月 6 日参照.

[10] Ward, J. H., Jr., “Hierarchical Grouping to Optimize an Objective Function”, Journal of the American Statistical Association, 58, pp. 236–244, (1963).

[11] 疑似人流データ，ナイトレイ社，https://nightley.jp/archives/1954/，2021 年 8 月 6 日参照.

[12] Adult Data Set, UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/adult, 2021 年 8 月 6 日参照.

[13] 野田 博之，磯 博康，西連地 利己，入江 ふじこ，深澤 伸子，鳥山 佳則，大田 仁史，能勢 忠男，“住民健診（基本健康検査）の結果に基づいた脳卒中・虚血性心疾患・全循環器疾患・がん・総死亡の予測”，日本公衛誌，2006 年 53 巻 4 号，pp. 265–276，(2006).

[14] 厚生労働省，“循環器疾患基礎調査”，https://www.mhlw.go.jp/toukei/list/junkanki_chousa.html. (2020 年 11 月 15 日参照)

[15] NIPPON DATA (https://shiga-publichealth.jp/nippon-data/，2020.08.14 参照)

[16] 川南 勝彦，簑輪 眞澄，岡山 明、早川 岳人，上島 弘嗣，NIPPON DATA80 研究グループ，喫煙習慣の全死因，がん，肺がん死亡への影響に関する研究：NIPPON DATA80，日本衛生学雑誌 Jan;57(4):669–673，(2003).

[17] 金子 侑紀, 小野 敦樹, 伊藤 聡志, 菊池 浩明, 服部 充洋, 飯田 泰興, 藤田 真浩, 山中 忠和, “匿名加工情報取扱事情者を調査するクローラーシステムの開発”, 情報処理学会第 82 回全国大会, pp.3_447–3_448, (2020).

[18] 藤田 真浩, 飯田 泰興, 服部 充洋, 山中 忠和, 松田 規, 伊藤 聡志, 菊池 浩明, “匿名加工情報取扱事業者による公表情報を利用した匿名加工カタログの提案と実装”, コンピュータセキュリティシンポジウム（CSS 2020）, pp. 1214–1221, (2020).

[19] 株式会社三菱総合研究所，“匿名加工情報・個人情報の適正な利活用の在り方に関する動向調査”，https://www.ppc.go.jp/files/pdf/tokumeikakou_report.pdf. (2020 年 11 月 16 日参照)

[20] 特定保険組合連合会（けんぽれん），“平成 28 年度　特定検診の「問診回答」に関する調査”，https://www.kenporen.com/toukei_data/pdf/chosa_h30_08-2.pdf.（2020 年 7 月 31 日参照）

[21] World Health Organization (WHO)，“International Statistical Classification of Diseases and Related Health Problems 10th Revision”，https://icd.who.int/browse10/2016/en.（2020 年 7 月 31 日参照）

[22] World Health Organization (WHO)，“ATC/DDD Index 2020”，https://www.whocc.no/atc_ddd_index/.（2020 年 7 月 31 日参照）

[23] 日本疫学会，“疫学用語の基礎知識　相対危険”，https://jeaweb.jp/glossary/glossary017.html.（2020 年 7 月 31 日参照）

[24] 日本疫学会，“疫学用語の基礎知識　オッズ比”，https://jeaweb.jp/glossary/glossary019.html.（2020 年 11 月 16 日参照)

[25] 国立がん研究センター がん情報サービス，“がん登録・統計　喫煙率”，https://ganjoho.jp/reg_stat/statistics/stat/smoking.html.（2020 年 11 月 16 日参照)

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

[27] 厚生労働省，“レセプト情報・特定健診等情報の提供に関するガイドライン”，平成 23 年（平成 28 年改訂）.

[28] 松井秀俊，小泉和之，統計モデルと推測，講談社，p. 103, (2019).

[29] 厚生労働省：平成 29 年人口動態統計月報年計の概況 (https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/geppo/nengai17/index.html，2020 年 08 月 14 日参照)

[30] stackoverflow，“How are feature_importances in Random Forest Classifier determined?”, https://stackoverflow.com/questions/15810339/how-are-feature-importances-in-randomforestclassifier-determined?answertab=votes#tab-top. (2020 年 11 月 12 日参照)

[31] Ahmed Arafa, Ehab S Eshak, Hiroyasu Iso, Kokoro Shirai, Isao Muraki, Norie Sawada, Shoichiro Tsugane, for the JPHC Study Group, Urinary Stones and Risk of Coronary Heart Disease and Stroke: the Japan Public Health Center-Based Prospective Study, Journal of Atherosclerosis and Thrombosis, 2020, vol. 27, No. 11, pp. 1208-1215, 2020.

[32] Islami F, Goding Sauer A, Gapstur SM, Jemal A. Proportion of Cancer Cases Attributable to Excess Body Weight by US State, 2011-2015. JAMA Oncol. 2019; vol. 5, No. 3, pp. 384–392.

[33] Saint-Maurice PF, Troiano RP, Bassett DR Jr, Graubard BI, Carlson SA, Shiroma EJ, Fulton JE, Matthews CE. Association of Daily Step Count and Step Intensity With Mortality Among US Adults. JAMA. 2020 Mar 24, vol.323, No.12, pp. 1151-1160.

[34] Chen F, Du M, Blumberg JB, Ho Chui KK, Ruan M, Rogers G, Shan Z, Zeng L, Zhang FF. Association Among Dietary Supplement Use, Nutrient Intake, and Mortality Among U.S. Adults: A Cohort Study. Ann Intern Med. 2019 May 7, vol.170, No.9, pp. 604-613.

[35] W. Maeda, T. Shimizu, T. Fukuoka and I. Morikawa, ”Dataset Properties and Degradation of Machine Learning Accuracy with an Anonymized Training Dataset,” 2020 Eighth International Symposium on Computing and Networking Workshops (CANDARW), Naha, Japan, 2020, pp. 341-347.

[36] Y. Yamaoka and K. Itoh, “k-presence-secrecy: Practical privacy model as extension of k-anonymity,” IEICE TRANSACTIONS on Information and Systems, vol. 100, no. 4, pp. 730–740, 2017.

[37] 菊池浩明, 小栗秀暢, 野島良, 濱田浩気, 村上隆夫, 山岡裕司, 山口高康, 渡辺知恵美, “PWSCUP: 履歴データを安全に匿名加工せよ”, コンピュータセキュリティシンポジウム (CSS 2016), pp. 271–278, 2018.

[38] 個人情報の保護に関する法律（平成 15 年法律第 57 号，平成 27 年法律第 65 号，および，平成 28 年法律第 51 号，令和 2 年法律第 44 号により改正）

[39] 行政機関の保有する個人情報の保護に関する法律（平成 15 年法律第 58 号）

[40] 丹後俊郎，古川俊之，医学への統計学第 3 版，朝倉書店，p. 195 (2013).

[41] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, “Mondrian Multidimensional K-Anonymity”, 22nd International Conference on Data Engineering (ICDE’06), Atlanta, GA, USA, 2006, pp. 25-25.

[42] Nithin Prabhu (Nuclearstar), “k–anonymity”, `https://github.com/Nuclearstar/K-Anonymity` (2021 年 3 月 26 日参照)

[43] ISO, Privacy enhancing data de-identification terminology and classification of techniques ISO Technical Specification ISO/TS 20889 (2018).

[44] Personal Information Protection Commission Secretariat, Report by the Personal Information Protection Commission Secretariat: Anonymously Processed Information –Towords Balanced Promotion of Personal Data Utilization and Consumer Trust– (2017).

[45] Kikuchi H, Yamaguchi T, Hamada K, Yamaoka Y, Oguri H, Sakuma J, What is the best anonymization method? –a study from the data anonymization competition pwscup 2015. Data Privacy Management Security Assurance (DPM2016) LNCS 9963:230–237 (2016).

[46] Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt E, Spicer K, Wolf P, Statistical disclosure control. Wiley, New York (2012).

[47] Duncan G, Elliot M, Salazar J, Statistical confidentially. Springer, New York (2011).

[48] Torra V, Data privacy: Foundations, new developments and the big data challenge. Studies in Big Data 28, Springer (2017).

[49] Samarati P, Sweeney L, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Computer Science Laboratory, SRI International Technical Report SRI-CSL-98-04 (1998).

[50] Meyerson A, Williams R, On the complexity of optimal k-anonymity. ACM PODS, pp 223–228 (2004).

[51] Bayardo RJ, Agrawal R (2005) Data privacy through optimal k-anonymization. ICDE’05, pp 217–228 (2005).

[52] LeFevre K, DeWitt DJ, Ramakrishnan R, Incognito: efficient full-domain k-anonymity. SIGMOD’05, pp 49–60 (2005).

[53] Basu A, Monreale A, Trasarti R, Corena JC, Giannotti F, Pedreschi D, Kiyomoto S, Miyake Y, Yanagihara T, A risk model for privacy in trajectory data. Journal of Trust Management, pp 2–9 (2015).

[54] Xiao X, Tao Y, *m*-invariance: toward privacy preserving republication of dynamic datasets. Proc of SIGMOD'07, pp 689–700 (2007).

[55] Mitzenmacher M, Upfal E, Probability and conputing: Randomized algolithms and probabilistic analysis. Cambridge University Press pp 32–34, Section 2.4.1 (2005).

[56] Satoshi Ito, Reo Harada, Hiroaki Kikuchi, De-identification for Transaction Data Secure against Re-identification Risk Based on Payment Records, Journal of Information Processing, 2020, Volume 28, Pages 511-519 (2020).

[57] UCI Machine Learning Repository, Online Retail Data Set [online]. https://archive.ics.uci.edu/ml/datasets/online+retail, [Accessed 3 Dec 2020]

[58] H. Kikuchi, T. Yamaguchi, K. Hamada, Y. Yamaoka, H. Oguri and J. Sakuma, "Ice and Fire: Quantifying the Risk of Re-identification and Utility in Data Anonymization," 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), Crans-Montana, pp. 1035-1042, 2016.

[59] Latanya Sweeny, "*k*-anonymity", *Int. J. of Uncertainty, Fuzziness & Knowledge-Based System*, Vol. 10, pp. 571-588, 2002.

[60] H. Akiyama, K. Yamaguchi, S. Ito, N. Hoshino, T. Goto, "Usage and Development of Educational Pseudo Micro-data –Sampled from national survey of family income and expenditure in 2004 –", Technical Report of the National Statistics Center (NSTAC), 16, pp. 1-43, 2012.

[61] UCI Machine Learning Repository: A WEB Page, `http://archive.ics.uci.edu/ml/index.php` Accessed in 2018-12-17.

[62] ISO Technical Specification ISO/TS 25237: Health informatics – Pseudonymization (2008).

[63] Information Commissioner's Office (ICO): Anonymisation: managing data protection risk code of practice (2012).

[64] C.C. Aggarwal and P.S. Yu.: A General Survey of Privacy-Preserving Data Mining, Models and Algorithms, *Privacy-preserving Data Mining*, Springer, pp.11–52 (2008).

[65] Koot, M. R., Mandjes, M., van't Noordende, G., and de Laat, C.: Efficient probabilistic estimation of quasi-identifier uniqueness, In Proceedings of ICT OPEN 2011, 14–15, pp.119–126 (2011).

[66] A Monreale, R Trasarti, D Pedreschi, C Renso and V Bogorny: C-safety: a framework for the anonymization of semantic trajectories, Transactions on Data Privacy, Vol. 4(2), pp.73–101 (2011).

[67] A. Basu, A. Monreale, R. Trasarti, J. C. Corena, F. Giannotti, D. Pedreschi, S. Kiyomoto, Y. Miyake and T. Yanagihara: A risk model for privacy in trajectory data, Journal of Trust Management, pp.2–9 (2015).

[68] Klara Stokes, Vicence Torra: n-confusion: a generalization of k-anonymity, EDBT/ ICDT Workshops 2012, pp.211–215 (2012).

[69] Zhizhou Li, Ten H. Lai: δ-privacy: Bounding Privacy Leaks in Privacy, Preserving Data Mining, DPM/CBT 2017, LNCS 10436, pp. 124–142, Springer (2017).

[70] Ito, S., Kikuchi, H., and Nakagawa, H.: Attacker models with a variety of background knowledge to de-identified data, J Ambient Intell Human Comput, pp. 1–11 (2019).

[71] Ryo Nojima, et al: How to Handle Excessively Anonymized Datasets, Journal of Information Processing, 26, pp.477-485 (2018).

[72] H. Kikuchi, H. Oguri, R. Nojima, K. Hamada, T. Murakami, Y. Yamaoka, T. Yamaguchi, C. Watanabe: PWS CUP Competition: De-identify Transaction Data Securely, Computer Security Symposium, 2A1-2, pp. 271–278, in Japanese (2016).

[73] H. kikuchi: Data Anonymization and Quantifying Risk Competition (online), `https://project.inria.fr/FranceJapanICST/files/2017/05/HKikuchi\_\\presentation\_2017.pdf`, Accessed in 2020-3-26.

[74] C. Dwork, "Differential privacy", Proceedings of ICALP 2006, LNCS vol.4052, pp.1-12, 2006.

[75] Khaled El Emam, Luk Arbuckle, "Anonymizing Health Data Case Studies and Methods to Get You Started", *O'Reilly*, 2013.

[76] J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata", Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 111-133, 2001.

[77] Daqing Chen, Sai Liang Sain, and Kun Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197–208, 2012.

[78] EAST JAPAN RAILWAY COMPANY, `http://www.jreast.co.jp/e/`, June 24, 2017.

[79] Money Forward, `http://corp.moneyforward.com/`, June 24, 2017.

[80] Lending Club (2019b) [online]. https://www.lendingclub. com/, [Accessed 15 Apr 2019]

[81] UCI Machine Learning Repository, Diabetes 130-US hospitals for years 1999–2008 Data Set [online]. https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008, [Accessed 17 Dec 2018]

[82] Lending Club, LOAN DATA [online]. https://www.lendingclub.com/info/download-data.action, [Accessed 15 Apr 2019]

[83] Domingo-Ferrer J, Soria-Comas J, From t-closeness to differential privacy and vice versa in data anonymization. Journal Knowledge-Based Systems Volume 74 Issue 1:151—158 (2015).

[84] Stokes K, On computational anonymity. Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2012 pp 336—347 (2012).

[85] Mimoto T, Kiyomoto S, Hidano S, Basu A, Miyaji A, The possibility of matrix decomposition as anonymization and evaluation for time-sequence data. 2018 16th Annual Conference on Privacy, Security and Trust (PST), IEEE pp 139-–145 (2018).

[86] 梶間大地，伊藤聡志，菊池浩明，“国内の匿名加工情報を一覧する利活用サイト「匿名加工情報目録」の開発”，情報処理学会第 83 回全国大会，pp.3_415-3_416, 2021.

[87] 金沢 史明，岸野 徹，“特許出願からみた匿名化関連技術の技術動向–平成 29 年度特許出願技術動向調査より–”, コンピュータセキュリティシンポジウム 2018(CSS2018), pp.906–912, 2018.

[88] 早稲田篤志，野島良，盛合志帆，菊池浩明，“良い仮名化 悪い仮名化”, 暗号と情報セキュリティシンポジウム 2017(SCIS2017), pp.1–8, 2017.

[89] 南和宏，“集計表秘匿における差分攻撃の考察”， 暗号と情報セキュリティシンポジウム 2017(SCIS2017), pp.1–8, 2017.

[90] 南和宏，阿部穂日，“集計表セル秘匿問題の拡張によるデータ効用保持の有効性評価”, コンピュータセキュリティシンポジウム 2018(CSS2018), pp. 809–813, 2018.

[91] 本郷節之，大加瀬稔，手塚理貴，寺田雅之，稲垣潤，鈴木昭弘，“集計データへの差分プライバシ適用における特性の一考察 III”, 暗号と情報セキュリティシンポジウム 2019(SCIS2019), pp.1–8, 2019.

[92] 杉山歩未，香川椋平，川田涼平，“滞在・閉路性を表現した擬似経路データ生成法 経路データの滞在・閉路性がもたらすリスク評価を目指して”，コンピュータセキュリティシンポジウム 2020(CSS2020), pp. 1238–1244, 2020.

[93] 正木彰伍，“攻撃者のモデル化を用いた軌跡情報の匿名性評価法”, コンピュータセキュリティシンポジウム 2017(CSS2017), pp.143–150, 2017.

[94] 小栗秀暢，黒政敦史，“匿名加工情報の作成における攻撃者知識と安全性についての一考察”, コンピュータセキュリティシンポジウム 2017(CSS2017), pp.151–158, 2017.

[95] 濱田浩気，岡田莉奈，小栗秀暢，菊池浩明，中川裕志，野島良，波多野卓磨，正木彰伍，渡辺知恵美，“匿名化アルゴリズムの公開・非公開による再識別容易性の比較”, 暗号と情報セキュリティシンポジウム 2018(SCIS2018), pp.1–8, 2018.

[96] 山田古都子，大圖健史，石井将大，田中圭介，“大きい k に対する k-匿名化近似アルゴリズム”, 暗号と情報セキュリティシンポジウム 2018(SCIS2018), pp.1–8, 2018.

[97] 福嶋雄也，古坂浩輝，満保雅浩，“集合の分割に基づく仮名化手法の実装”, コンピュータセキュリティシンポジウム 2019(CSS2019), pp.1273–1276, 2019.

[98] 前田若菜，清水俊也，福岡尊，森川郁也，“匿名化によって機械学習の精度に影響を与えるデータの特徴の検討”, 暗号と情報セキュリティシンポジウム 2020(SCIS2020), pp.1–8, 2020.

[99] 山岡裕司，“データ流通における漏洩者を特定可能にするサンプリング方式の提案”, コンピュータセキュリティシンポジウム 2019(CSS2019), pp.275–280, 2019.

[100] 柚木塁，上土井陽子，若林真一，“動的データテーブルの連続的匿名化の安全性について”, 第 6 回データ工学と情報マネジメントに関するフォーラム（DEIM2014）, セッション E5(プライバシ)，pp.1–8, 2014.

[101] Hiroaki Kikuchi, Katsumi Takahashi, “Ziph Distribution Model for Quantifying Risk of Re-identification from Trajectory Data”, Journal of Information Processing, Vol. 24, No. 5, pp. 816–823, 2016.

[102] N. Li, T. Li and S. Venkatasubramanian, “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity”, 2007 IEEE 23rd International Conference on Data Engineering, pp. 106-115, 2007.

[103] statsmodels, `https://github.com/statsmodels/statsmodels` (2021 年 11 月 19 日参照)

[104] scikit-learn Machine Learning in Python, `https://scikit-learn.org/stable/` (2021 年 11 月 19 日参照)

[105] python, `https://www.python.org/` (2021 年 11 月 19 日参照)

[106] Christoph Heindl (cheind), “py-lapsolver”, `https://github.com/cheind/py-lapsolver` (2021 年 11 月 19 日参照)

[107] Ios Kotsogiannis, Yuchao Tao, Xi He, Maryam Fanaeepour, Ashwin Machanavajjhala, Michael Hay, Gerome Miklau, “PrivateSQL: a differentially private SQL query engine”, Proceedings of the VLDB Endowment, Volume 12, Issue 11, pp. 1371—1384, July 2019.

[108] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, Lars Vilhuber, “Privacy: Theory meets Practice on the Map”, 2008 IEEE 24th International Conference on Data Engineering, pp. 277-286, 2008.

[109] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, Kunal Talwar, “Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release”, PODS ’07: Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Pages 273—282, June 2007.

[110] 高崎晴夫，“プライバシーの経済学”，勁草書房，4 章，2018.

[111] JR 東日本，“Suica に関するデータの社外への提供について”, https://www.jreast.co.jp/information/aas/20151126_torimatome.pdf. (2021 年 11 月 29 日参照)

[112] J. Domingo-Ferrer and K. Muralidhar, “New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users.”, CoRR, 2015.