明治大学総合数理学部

2016 年度

卒　業　研　究

乗降履歴データの
有用性評価指標と匿名加工

学位請求者　先端メディアサイエンス学科

伊藤聡志

目次

1. はじめに

1.1. 研究背景

2013 年 6 月，JR 東日本が Suica の乗降履歴データを加工したものを日立製作所に提供することが明らかになった．しかし，このデータ提供に対して多くの Suica 利用者から批判を受けた．詳細は[1]を参考されたい．

この事件をきっかけに，企業間での顧客データ売買の際のデータ加工が重要視されるようになり，2015 年 9 月には個人情報保護法が改正されて，匿名加工情報が定義された．同年 10 月には匿名加工技術の開発と安全性の評価手法の確立を目的として，匿名加工・再識別コンテスト(PWSCUP)[2]が開催された．匿名加工技術が確立されれば，Suica データのようなビッグデータを安全に利活用することが可能となる．

1.2. 匿名加工・再識別について

データを個人が特定されないように加工することを匿名加工といい，データから個人を特定することを再識別という．例えば，表 1.1 のような学生(名前を A,B,C,D とする)の試験結果データ X を想定する．ID は名前を仮名化したものである．

表 1.1　学生の試験結果データ X

ID	数学	物理	英語
1	100	20	20
2	80	30	60
3	40	60	80
4	20	70	90

X をそのまま公表してしまうとプライバシー保護の観点から様々な問題が生じる．例えば，A が「B は数学の試験で 100 点を取った」という情報を知っている場合，A は公表された X から B の物理と英語の点数を知ることができてしまう．つまり，「ID=1 は B のデータである」と再識別ができてしまう．

そこで，公表する前に X を匿名加工する必要が生じる．X を匿名加工した一例として，匿名

加工データ X'を表 1.2 に示す．なお，X'は X を「k-匿名化」という手法で匿名加工したものである．k-匿名化は，データの値を書き換えることにより個人が特定されるリスクを 1/k 以下に抑える手法である．

表 1.2　匿名加工データ X'

ID	数学	物理	英語
1	80~100	20~30	20~60
2	80~100	20~30	20~60
3	20~40	60~70	80~90
4	20~40	60~70	80~90

X'ならば公表しても，個人を特定するのは困難である．例えば，A が「B は数学の試験で 100 点を取った」という情報を知っている場合，A は公表された X'から「B のデータは ID=1 か ID=2 のどちらかである」と絞り込むことはできるものの，完全に特定することは困難であり，B の物理と英語の点数を知ることもできない．個人が特定されるリスクが 1/2 であるため，X'は 2-匿名化されたデータである．

しかし，匿名加工をする際に注意しなければならないのが，「データの有用性と安全性」である．データの有用性は「そのデータがどれだけ役に立つか」を表し，データの安全性は「そのデータがどれだけ安全であるか」を表す．例えば X と X'の例だと，元データである X は有用性が高いが，簡単に個人が特定されてしまうため，安全性は低い．一方，匿名加工データである X'は安全性が高いが，元データを加工しているため，有用性は低い．このように，匿名加工はデータの安全性を高めることができるが，代わりに有用性を下げてしまう．1.1 節で説明した PWSCUP では，データに対応した有用性評価指標と安全性評価指標が用意され，できるだけ有用性を下げずに安全性を高める匿名加工手法の研究が進められている．

また，企業は収集した顧客データやトランザクションデータを利活用する際，そのデータのユースケースに応じてリスク評価と匿名加工手法を考える必要がある．

1.3. 研究目的

本研究の目的は，実際に個人情報データを収集・作成し，そのデータを匿名加工することである．完全な実データを用いた匿名加工の例は多くない．例えば，購買データを想定した評価指標・匿名加工に[2][4]があるが，これらの例で用いられているデータの一部は擬似データであ

り，完全な実データではない.

1.4. 研究方法

本研究では完全な実データを用いた匿名加工を行う．31 人の交通 IC カードから乗降履歴データを取得し，そのデータのユースケースを想定して，それに対応する評価指標・匿名加工手法を検討する．本稿では「世帯データ」と「乗降履歴データ」の 2 種類のデータを用いる．これらのデータの関係を表 1.3 に示す.

表 1.3　本稿で用いるデータの関係

	世帯データ	乗降履歴データ
データの種類	静的データ	動的データ
データの大きさ	25 属性*8333 レコード	6 属性*31 レコード 10 属性*584 レコード
作成者	独立行政法人統計センター	明治大学菊池研究室
取り扱う章	2 章	3,4,5 章

2. 静的データに対する匿名加工と再識別

2.1. 静的データと動的データ

個人情報データには静的データと動的データの 2 種類がある．前章の X のようなデータを静的データといい，乗降履歴データや購買履歴データのような，時間属性が付加されたデータを動的データという．動的データは静的データと異なり，マスターデータとトランザクションデータの組になっており，時々刻々とデータが増加する．動的データ(マスターデータ M，トランザクションデータ T)の例として，学生の試験結果推移のデータを表 2.1，2.2 に示す．この例の場合，M は学生の名簿，T は学生の試験結果の推移となる．

表 2.1　マスターデータ M

ID	性別	学年	組
1	男	2	1
2	女	2	1
3	女	2	2
4	男	2	2

表 2.2　トランザクションデータ T

ID	年/月	数学	物理	英語
1	2017/1	90	90	40
2	2017/1	90	40	60
3	2017/1	80	70	10
4	2017/1	10	50	100
1	2016/7	100	20	20
2	2016/7	80	30	60
3	2016/7	40	60	80
4	2016/7	20	70	90

2.2. 静的データに対する匿名加工と再識別

本節では，静的データの匿名加工と再識別について，2015 年 10 月に長崎で開催された

PWSCUP2015[2]の結果や，私が 2016 年 5 月に鳥取で開催された CSEC 研究発表会で発表した内容を用いて説明を行う．

2.2.1. PWSCUP2015 について

PWSCUP2015 では，コンテスト参加者が匿名加工を行う対象として，独立行政法人統計センター[5]が作成した擬似ミクロデータが用いられた．このデータは 8333 レコード，25 属性のデータであり，平成 16 年世帯別年間支出額を表している．1～13 列目は世帯の人数，年齢などの離散的なデータを与えており，本研究ではこれを準識別子(QI)に分類する．一方，14～25 列目は食費や医療費などの支出額であり，連続値を持つ．本研究ではこれを属性値(SA)とする． QI，SA の厳密な定義は[2]を参考されたい．

匿名加工手法の評価には 6 つの有用性指標と 6 つの安全性指標が用いられた．表 2.3 に指標一覧とその対象を示す．これらの指標から出された評価値によって匿名加工データの評価が行われた．PWSCUP2015 の詳細や結果については[2][6]を参考されたい．

表 2.3 PWSCUP2015 のデータ評価指標一覧

		指標名	指標の内容	対象
有用性指標	U1	meanMAE	SA 平均絶対誤差	SA
	U2	crossMean	クロス集計値の平均絶対誤差	QI,SA
	U3	crossCnt	クロス集計数の平均絶対誤差	QI
	U4	corMAE	SA の相関係数の平均絶対誤差	SA
	U5	IL	匿名加工データの各値の平均絶対誤差	SA
	U6	nrow	匿名加工データのレコード数	行数
安全性指標	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 からランダムな再識別率	QI
	E2	identify.sa	QI から SA15 列目による再識別率	QI,SA
	E3	identify.sort	SA の総和でソートによる再識別率	SA
	E4	Identify.sa21	SA21 列について再識別率	SA

2.2.2. 再識別手法について

本章では提案手法との比較に，匿名加工・再識別コンテスト　PWSCup2015 で匿名加工データの安全性の評価に用いられた以下の 4 つの手法を用いる．ここで，元データを X，X を匿名加工(SA へのノイズ付加)したデータを B とする．説明のために表 2.4，2.5 にこれらの例を示す．3 つの QI 属性，2 つの SA 属性についての 4 つのレコード(行)から成っている．本編では QI 属性の値の組み合わせを，QI のベクトルと呼ぶ(SA も同様)．例えば，X の第 1 レコードの QI のベクトルは(2, 1, 1)であり，SA のベクトルは(100,100)である．

再識別率を，再識別手法によって求めた行番号と匿名加工データの行番号との一致率，すなわち，一致したレコード数／元データのレコード数と定義する．

表 2.4　サンプルオリジナルデータ X

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 2.5　匿名加工データ(ノイズ付加)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

(1) identify.rand(E1)

この手法では，B の再識別したいレコードと同じ QI のベクトルを持つレコードを X から探し，その中からランダムに再識別を行う．例えば，B の第 1 レコードと同じ QI のベクトル(2,1,1)を持つレコードは X の第 1，第 2 レコードであるため，それらの 2 レコードからランダムに 1 つを選ぶ．

(2) identify.sa(E2)

この手法では，B の再識別したいレコードと同じ QI のベクトルを持つレコードを X から探

し，その中から特定の SA が最も近いレコードを再識別する．例えば，B の第 1 レコードと同じ QI のベクトル(2,1,1)を持つ 2 レコードの中で SA1 の値が 110 に最も近いのは，100 と 200 の内第 1 レコードである．そこで，X の第 1 レコードを加工したと推定する．

(3) identify.sort(E3)

この手法では，SA の和で X と 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．表 1 の例では， SA の和(SA1+SA2)でソートした．X 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，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．

(4) identify.sa21(E4)

この手法ではレコードの QI は考慮せず，特定の SA の値だけで再識別を行う．例えば，B の第 2 レコードの SA1 の値(220)と最も近い値を持つのは X 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．

2.2.3. 提案手法 identify.euc

本提案 identify.euc では，B の再識別したいレコードと同じ QI のベクトルを持つレコードを X から探し，それらの SA のユークリッド距離 $\mathrm{D}(\boldsymbol{a},\boldsymbol{b}) = \sqrt{\sum_{i\in s}^{n}(b_i - a_i)^2}$ を用いて再識別を行う．例えば，B の第 1 レコード(SA のベクトルは$\mathbf{b_1}$ =(110, 90))と同じ QI のベクトル(2, 1, 1)を持つ X のレコードは第 1 レコード(SA のベクトルは$\mathbf{a_1}$ =(100, 100))と第 2 レコード(SA のベクトルは$\mathbf{a_2}$ =(200, 400))であり，ユークリッド距離は，$\mathrm{D}(\mathbf{a_1},\mathbf{b_1}) = 14.142 < 322.8 = \mathrm{D}(\mathbf{a_2},\mathbf{b_1})$ より，$\mathbf{b_1}$を加工前のレコード$\mathbf{a_1} = \mathbf{b_1}$と推定する．

B は SA のみ加工されたデータであるため，X と B の QI 属性はすべて適合する．しかし QI 属性を加工されている場合，QI のベクトルが適合しないことがある．例えば，表 2.6 の匿名加工サンプルデータ D を考えよ．D は X の 3 つの QI 属性の内，QI3 の値をすべて 1 に統一した匿名加工データである．

表 2.6　匿名加工データ(QI 統一)　D (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

D の 3,4 レコードを identify.euc で再識別する際，問題が生じる．(1, 1, 1)という QI のベクトルを持つレコードは A には存在しないためである．このように QI のベクトルが適合しない場合の動作を 2 種類用意し，それらを実装した identify.euc をそれぞれ EUC1，EUC2 とした．動作の詳細は表 4 に示す．EUC1, EUC2 のアルゴリズムを表 2.7 に示す.

表 2.7　EUC1 と EUC2

EUC1	QI のベクトルが適合しない場合，D の探索中のレコード$\mathbf{d_i}$の i を返す
EUC2	QI のベクトルが適合しない場合，そのレコードと，元データの全レコードの SA についてユークリッド距離を求め，最も近いものを探し，その ID を返す.

Algorithm: identify.euc (EUC1)

---

1. 入力: 元データ X，匿名加工データ B，B の長さ n，再識別に用いる QI のベクトル q，SA のベクトル s
2. key = q の QI，value = 対応するレコードでインデックス F を作成する.
3. B の第 i 番目のレコード$\mathbf{b_i}$と同じ QI を持つ全レコードを X から探し，それらの全てについてレコード間のユークリッド距離$D(\mathbf{a_j}, \mathbf{b_i})$を s の SA で求める．最もユークリッド距離の近い j レコードを i の同一レコードと推定する.
4. QI のベクトルが適合しない場合，B の探索中のレコード$\mathbf{b_i}$ の i を返す.
5. 3 または 4 を B のすべてのレコード i=1,…,n について行い，推定行番号 ID を返す.

例 1）X, B: 2.2.2 項のサンプルデータ X, B

q = (1, 2, 3)　s = (4, 5)

q = (1, 2, 3 }であるため，B の 1~3 列目を用いてインデックス F を作成する．この場合の F を表 2.8 に示す.

表 2.8　例 1 における F

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

B の第 1 レコード$\mathbf{b_1}$を再識別する場合，X で QI のベクトルが$\mathbf{b_1}$と同じ(2, 1, 1)であるのは$\mathbf{a_1}$と$\mathbf{a_2}$である．そのため，$\mathbf{a_1}$と$\mathbf{b_1}$，$\mathbf{a_2}$と$\mathbf{b_1}$間のユークリッド距離を求め，$\mathbf{b_1}$との距離が最小の X のレコードを$\mathbf{b_1}$の推定レコードとする．この工程を$\mathbf{b_2}$，$\mathbf{b_3}$，$\mathbf{b_4}$についても行い，推定行番号 ID を作成する．

例 2）X, D: 2.2.2，2.2.3 項のサンプルデータ X, D

q = (1, 2, 3)　s = (4, 5)

q = (1, 2, 3 }であるため，D の 1~3 列目を用いてインデックス F を作成する．この場合の F を表 2.9 に示す.

表 2.9　例 2 における F

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

D の第 1 レコード$\mathbf{d_1}$を再識別する場合，X で QI のベクトルが$\mathbf{d_1}$と同じ(2, 1, 1)であるのは$\mathbf{a_1}$と$\mathbf{a_2}$である．そのため，$\mathbf{a_1}$と$\mathbf{d_1}$，$\mathbf{a_2}$と$\mathbf{d_1}$間のユークリッド距離を求め，$\mathbf{d_1}$との距離が最小の X のレコードを$\mathbf{d_1}$の推定レコードとする．この工程を$\mathbf{d_2}$，$\mathbf{d_3}$，$\mathbf{d_4}$についても行い，推定行番号 ID を作成する．しかし，$\mathbf{d_3}$，$\mathbf{d_4}$の QI のベクトルは(1, 1, 1)であるが，X にはこの QI のベクトルを持つレコードは存在しない．そのため，$\mathbf{d_3}$は 3，$\mathbf{d_4}$は 4 を推定行番号として返す.

Algorithm: identify.euc (EUC2)

1~3．EUC1 と同一である.

4.　QI のベクトルが適合しない場合，そのレコードと，元データの全レコードの SA についてユークリッド距離を求め，最も近いものを探し，その ID を返す.

5．3 または 4 を D のすべてのレコード i=1,…,n について行い，推定行番号 ID を返す.

例 3）X, D: 2.2.2 項，2.2.3 項のサンプルデータ X, D

q = (1, 2, 3)　s = (4, 5)

q = (1, 2, 3)であるため，D の 1~3 列目を用いてインデックス F を作成する．この場合の F を表 2.10 に示す.

表 2.10　例 3 における F

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

D の第 1 レコード$\mathbf{d_1}$を再識別する場合，X で QI のベクトルが$\mathbf{d_1}$と同じ(2, 1, 1)であるのは$\mathbf{a_1}$と$\mathbf{a_2}$である．そのため，$\mathbf{a_1}$と$\mathbf{d_1}$，$\mathbf{a_2}$と$\mathbf{d_1}$間のユークリッド距離を求め，$\mathbf{d_1}$との距離が最小の X のレコードを$\mathbf{d_1}$の推定レコードとする．この工程を$\mathbf{d_2}$，$\mathbf{d_3}$，$\mathbf{d_4}$についても行い，推定行番号 ID を作成する．しかし，$\mathbf{d_3}$，$\mathbf{d_4}$の QI のベクトルは(1, 1, 1)であるが，X にはこの QI のベクトルを持つレコードは存在しない．そのため，$\mathbf{d_3}$と X の全レコード間，$\mathbf{d_4}$と X の全レコード間のユークリッド距離を求め，距離が最も近いレコードの行番号 ID を返す.

2.2.4.　identify.euc の評価

本項の目的は，PWSCup2015 の匿名加工データの解析である．本項では，次の 3 つについて評価する.

・単独匿名加工手法の効果

・PWSCup2015 を用いた評価

・提案手法(identify.euc)の性能評価

identify.euc を評価する際に用いるデータを，D_1,…,D_{12}とする．これらのデータは PWSCup2015 の本戦に参加した上位 3 チームを含む 5 チームから提出された，擬似ミクロデータを匿名加工したデータである．表 2.11 に示す.

表 2.11　PWSCup2016 のD_1,...,D_{12}の作成チーム

データ名	作成チーム	成績
D_1, D_2	T_A	
D_3, D_4	T_B	2 位
D_5, D_6	T_C	
D_7, D_8, D_9	T_D	1 位
D_{10}, D_{11}, D_{12}	T_E	3 位

D_1,...,D_{12}は複数の匿名加工手法を組み合わせて作成されたデータである．そのため，どの加工手法が安全性と有用性にどれだけ効果していたかが不明であった．そこで，これらの加工に用いられていた主な手法を，単独に適用したデータを用いて，各々の効果を調査する．

疑似ミクロデータからランダムにサンプリングした小規模データ(100 レコード，25 属性)をもとに，8 つの匿名加工データを単独手法によって作成した．これらのデータをD_A,...,D_Hとし，詳細を表 2.12 に示す．D_A,...,D_Hの有用性と安全性を調べることにより，D_1,...,D_{12}がどの手法を組み合わせて加工されたデータであるかを予測する．また，D_A,...,D_Hを作成する際に用いた 8 つの匿名加工手法について，2.2.2 項の X を用いて説明する．ただし，k-匿名化については[3]を，山岡匿名化については[2]を参考されたい．

表 2.12　単独加工データD_A,...,D_H

データ名	匿名加工手法	加工対象
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（対象外）	QI
D_E	QI 統一（対象内）	QI
D_F	SA 平均化	SA
D_G	QI 内スワップ	SA
D_H	レコード削除	レコード

(1) SA ノイズ付加

元データの SA にノイズを付加する手法(ランダマイズ，摂動化)である．2.2.2 項の B は X を

この手法で加工した匿名加工データである．この手法で加工を行うと，SA を対象とした有用性指標である U1, U2, U4, U5 が下がり，同様に SA を対象とした安全性指標である E3 と E4 が上がると考えられる．

(2) QI 統一

元データ X の QI のいくつかの属性をある値に統一する手法である．2.2.2 項の D は X の QI の内，QI3 を 1 に統一した匿名加工データである．この手法で加工を行うと，有用性を下げずに安全性を上げることが可能である．しかし QI の内，クロス集計の評価値 U2, U3 の対象である属性を統一してしまうと有用性が下がってしまう．例えば PWSCup2015 本戦の場合，U2, U3 で用いる QI の属性は 1 列目～6 列目であったため，1 列目～6 列目を統一してしまうと有用性が下がった．この手法の対象となる属性を加工に含めるか否かで，D と E の 2 種類を区別する．

(3) SA 平均化(ミクロアグリゲーション)

元データのレコードの内，QI のベクトルが同じ値のレコードの SA を属性ごとに平均値で置き換える手法である．X を SA 平均化によって匿名加工した結果 F を表 2.13 に示す．この手法で加工を行うと，U4, U5 が下がり，E3, E4 が上がる．この場合，QI のベクトルによって，4 つのレコードが{1, 2}, {3, 4}の 2 グループに分類されている．各グループの平均値を用いるため，全体の平均値 U1 には影響を与えない．

表 2.13　匿名加工データ(SA 平均化)F

グループ	QI1	QI2	QI3	SA1	SA2
(1)	2	1	1	**150***	**250***
	2	1	1	**150***	**250***
(2)	1	1	2	**350***	**350***
	1	1	2	**350***	**350***

(3) QI 内スワップ

元データのレコードの内，QI のベクトルが同じ値のレコードの SA を属性ごとにランダムにスワップする手法である．X を QI 内スワップによって匿名加工した結果 G を表 2.14 に示す．グループ(1)では SA1 を，(2)では SA2 を入れ替えている．スワップなので平均値は変わらず，

QI 内なのでクロス集計値 U2, U3 も変化しない．この手法で加工を行うと，相関係数等の U4, U5 が下がり，安全性 E2, E3, E4 が上がる．

表 2.14　匿名加工データ(QI 内スワップ)G

グループ	QI1	QI2	QI3	SA1	SA2
(1)	2	1	1	**200***	100
	2	1	1	**100***	400
(2)	1	1	2	300	**500***
	1	1	2	400	**200***

(4) レコード削除

元データ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 U1, U2, U3, U5, U6 が下がり，E3, E4 が上がる．(PWSCup2015 では該当データは提出されなかった)

(5) k-匿名化，山岡匿名化

これらの手法については参考文献[3],[2]を参考されたい．k-匿名化によって加工を行うと U2, U3 が下がり，S1, S2, E1, E2 が上がり，山岡匿名化で加工を行うと U5 が下がり，E1~E4 が上がる．

データを匿名加工すると，一般的に有用性が低くなり，安全性が高くなる．手法の特性をもとに，D_A,…,D_Hの作成に用いた 8 つの匿名加工手法が U1～U6, S1, S2, E1~E4, EUC1 の値をどのように変化させるのか定性的な効果の予測を表 2.15 に示す．有用性 U1～U6 の欄における「×」は「大きく損なう」，「∆」は「少し損なう」，「ー」は「変化しない」を意味し，安全性 S1, S2 における「○」は「高まる」，「×」は「変化しない」を意味し，E1~E4, EUC における「○」は「この手法には強い」，「∆」は「この手法には少し強い」，「×」は「この手法には弱い」を意味する．

D_A,…,D_Hの有用性・安全性を表 2.16 に示す．Original の列にはD_A,…,D_Hの元データの指標値を示す．　E4(identify.sa21)の値が全体的に低いように思えるが，これは擬似ミクロデータの 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，U4 と E3 についての散布図を図 2.1 に示す．

表 2.15　期待される効果

		匿名加工手法							
		k-匿名化	ノイズ付加	YA	QI統一（対象外）	QI統一（対象内）	ミクロ	QI内スワップ	レコード削除
有用性	U1	−	△	−	−	−	−	−	×
	U2	×	△	−	−	×	−	−	×
	U3	×	△	−	−	×	−	−	×
	U4	−	△	−	−	−	×	×	×
	U5	−	△	×	−	−	×	×	×
	U6	−	−	−	−	−	−	−	×
安全性	S1	○	×	×	×	×	×	×	×
	S2	○	×	×	○	○	×	×	×
	E1	△	×	○	△	△	×	×	×
	E2	△	×	○	△	△	×	△	×
	E3	×	△	○	×	×	○	○	×
	E4	×	△	○	×	×	○	△	×
	EUC	△	×	○	△	△	×	○	×

表 2.16　元データと$D_A,\ldots,D_H$の有用性と安全性

	Original	DA	DB	DC	DD	DE	DF	DG	DH
	元データ	k-匿名化	ノイズ付加	YA	QI統一（対象外）	QI統一（対象内）	ミクロ	QI内スワップ	レコード削除
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	0.020	0	0	0	0.000	0.000	0.049
U5	0	0	0.016	0.120	0	0	0.000	0.000	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	1	0.13	0.99	0	0.07	0.11	0.94	1	1
E2	1	0.17	1	0	1	1	1	1	1
E3	1	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	1	0.13	1	0	0.07	0.11	1	1	1
EUC2	1	0.17	1	0	1	1	1	1	1

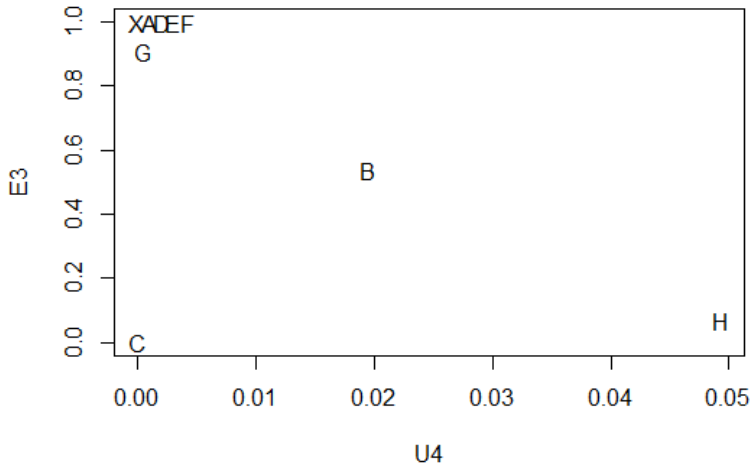

図 2.1　U4 と E3 についての散布図

表 2.17，2.18 にD_1,…,D_{12}の有用性と安全性についての表と，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．匿名加工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加工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 2.19 に示す)．サンプルデータ A, B, C, D, F, G とD_A,…,D_Hは異なる．しかし，記号が同じデータは同じ匿名加工手法で加工されている．例えば，2.2.2 項の 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．

表 2.17　D_1,…,D_{12}の有用性と安全性

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
U1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
U2	58340.87	0.00	31572.91	31400.95	0.00	0.00	4321.75	0.00	0.00	65093.42	52975.02	46100.64
U3	18.60	0.00	1.01	0.99	0.00	0.00	1.54	0.00	0.00	7.28	2.97	1.85
U4	0.00	0.01	0.00	0.00	0.07	0.07	0.03	0.09	0.09	0.15	0.11	0.11
U5	0.00	0.01	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.02	0.02	0.02
U6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S1	1.00	1.00	3.00	3.00	1.00	1.00	3.00	1.00	1.00	41.00	8.00	4.00
S2	2.66	1.88	4.91	4.86	36.07	36.07	36.07	13.71	13.68	106.83	42.30	31.09
E1	0.03	0.65	0.20	0.19	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.02
E2	0.82	0.65	0.24	0.24	0.02	0.02	0.02	0.00	0.00	0.01	0.02	0.02
E3	1.00	0.00	0.25	0.25	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
E4	0.19	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EUC1	0.30	0.48	0.21	0.21	0.07	0.07	0.88	0.00	0.00	0.00	0.01	0.01

表 2.18　$D_1,\ldots,D_{12}$の評価結果とそれによる加工手法の予測

		匿名加工データ											
		D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
有用性	U1	−	−	−	−	−	−	−	−	−	−	−	−
	U2	×	−	×	×	−	−	×	−	−	×	×	×
	U3	×	−	△	△	−	−	△	−	−	×	×	△
	U4	−	△	−	△	△	△	△	△	△	×	×	×
	U5	−	△	△	△	△	△	△	△	△	△	△	△
	U6	−	−	−	−	−	−	−	−	−	−	−	−
安全性	S1	×	×	△	△	×	×	△	×	×	○	○	△
	S2	×	×	△	△	△	○	○	○	○	○	○	○
	E1	△	×	×	×	◎	◎	○	◎	○	△	△	△
	E2	×	×	×	×	△	△	△	◎	○	△	△	△
	E3	×	○	×	×	○	○	○	○	○	○	○	○
	E4	×	○	△	△	○	○	○	○	◎	○	○	○
	EUC1	×	×	×	×	△	△	×	○	○	○	○	○
匿名加工手法	D_A	−	−	○	○	−	−	○	−	−	○	○	○
	D_B	−	−	−	−	−	−	−	−	−	−	−	−
	D_C	−	−	−	−	○	○	−	○	○	−	−	−
	D_D	−	−	−	−	○	○	○	−	−	−	−	−
	D_E	○	−	−	−	−	−	−	○	○	−	−	−
	D_F	−	○	−	−	−	−	−	−	−	○	○	○
	D_G	−	−	○	○	−	−	○	○	○	−	−	−
	D_H	−	−	−	−	−	−	−	−	−	−	−	−

表 2.19　D_A, D_FとD_{10}の特徴

	D_A	D_F	D_{10}
U1	−	−	−
U2	×	−	×
U3	×	−	×
U4	−	×	×
U5	−	×	△
U6	−	−	−
S1	○	×	○
S2	○	×	○
E1	△	×	△
E2	△	×	△
E3	×	○	○
E4	×	○	○
EUC1	△	×	○

EUC1 と既存の 4 つの手法との比較を行う．比較に用いるデータは，元データに擬似ミクロデータ(8333 レコード，25 属性)，匿名加工データに$D_1,\ldots,D_{12}$を用いる．

既存手法の再識別成功率を表 2.20 に示す．赤い数値(*が付いている数値)は，その匿名加工データに対して最も再識別成功率が高かった再識別手法を示している．EUC1 は 12 個中 5 個が最高値であり，既存手法よりも多い．

表 2.20　既存手法と EUC1 の再識別成功率

	既存方式				提案方式
匿名加工データ	Id-rand	Id-sa	Id-sort	Id-sa21	EUC1
D_1	0.0326	0.8238	*1.0000	0.1858	0.3010
D_2	0.6485	*0.6507	0.0012	0.0022	0.4780
D_3	0.1990	0.2412	*0.2482	0.0511	0.2070
D_4	0.1894	0.2401	*0.2526	0.0455	0.2110
D_5	0.0000	0.0223	0.0004	0.0002	*0.0743
D_6	0.0000	0.0223	0.0004	0.0002	*0.0743
D_7	0.0023	0.0223	0.0091	0.0014	*0.8762
D_8	0.0000	0.0000	0.0004	0.0002	*0.0011
D_9	0.0001	0.0002	0.0004	0.0000	*0.0024
D_{10}	0.0060	*0.0066	0.0001	0.0005	0.0043
D_{11}	*0.0180	0.0164	0.0001	0.0001	0.0080
D_{12}	*0.0214	*0.0214	0.0004	0.0001	0.0080
平均	0.0931	0.1723	0.1261	0.0240	*0.1871
標準偏差	0.1741	0.2578	0.2681	0.0499	0.2426
最適数	2	3	3	0	5

EUC1 の最高値が従来手法より多かった理由は，提案手法で再識別に用いる属性の数が既存手法より多いためである． 例えば，既存手法の identify.sa は特定の SA からレコードを再識別する手法であるが，その特定の SA に大きいノイズが加えられると，正しく再識別することができなくなる．対して EUC1 は再識別の際に全ての SA を用いるため，それらの内 1 つが大きく加工されても，他の SA から再識別をすることができる．

しかし，$D_1,\ldots,D_4$，$D_{10},\ldots,D_{12}$においては identify.sa に再識別率で劣っている．これは identify.sa が EUC2 と同様に，「QI が適合しない場合，元データの全レコードの SA と計算を行い，ID を返す」という仕組みが実装されているためと考えられる．本来ならば EUC2 と

identify.sa を比較するべきであるが，EUC2 は計算時間が identify.sa と比べてはるかに多く，再識別率を出すのが困難であるため断念した．また表 2.17 より，k-匿名化をされているデータでは EUC 1 よりも identify.sort の方が再識別率が高い．

また，もう一つの理由として，比較に用いた匿名加工データはコンテストに提出されたものであるため，安全性指標である 4 つの既存手法に対抗できるように作られたものが多い．そのため，提案手法が有利になった可能性も考えられる．

擬似ミクロデータには QI 属性が 13 あり，そのうちどれを提案手法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．

用いる QI 属性の数を|q|，SA 属性の数を|s|とおくと，|q|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 2.2，2.3，2.4，2.5 に，100 レコード，25 属性のデータを用いた時の，|q|と|s|の変化に伴う計算時間と再識別成功率の変化を示す．計算時間は|q|について単調に減少しており，再識別率は|q|について単調に増加している(ただし，|q|=5 で飽和している)．計算時間は|s|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，再識別率は|s|に依存しなかった．図 2.6 に|q|=1 のときのレコード数の増加に伴う計算時間の変化を示す． 計算時間はレコード数に対して増加している．なお，擬似ミクロデータの SA 属性にノイズを付加した匿名加工データでテストした結果，|q|=13 のとき計算時間は約 1 分，再識別率は約 17%であり，|q|=6 のとき計算時間は約 31 分，再識別率は約 20%であった．ただし，以下の図はすべて EUC1 についてのものである．EUC2 は EUC1 に比べて計算量がはるかに多いため，EUC2 では計算時間と再識別成功率の両方が EUC1 より増加すると考えられる．

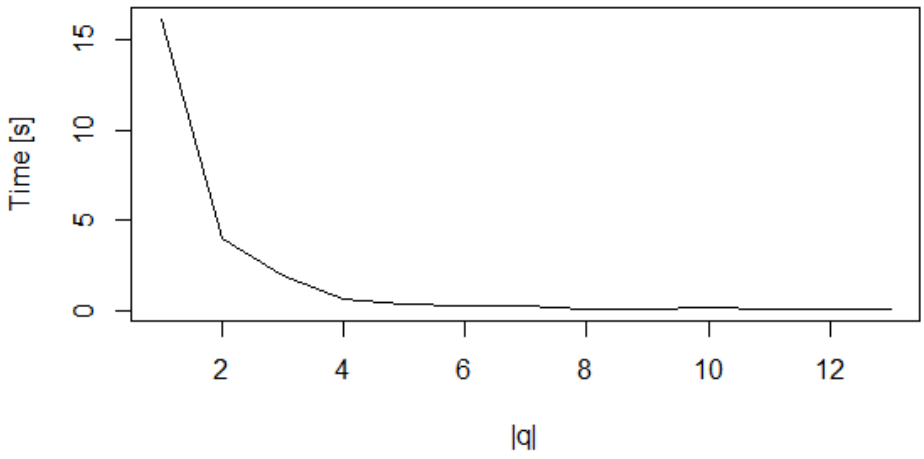

図 2.2　|q|についての計算時間

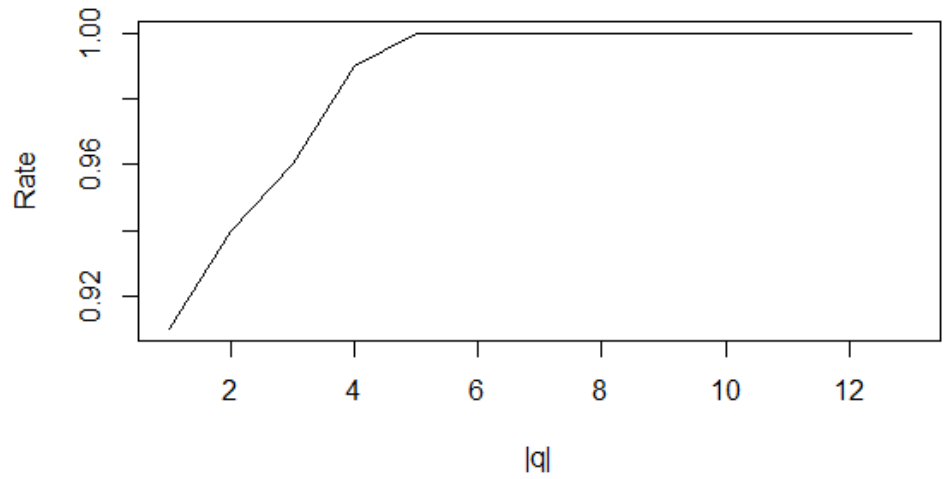

図 2.3　|q|についての再識別成功率

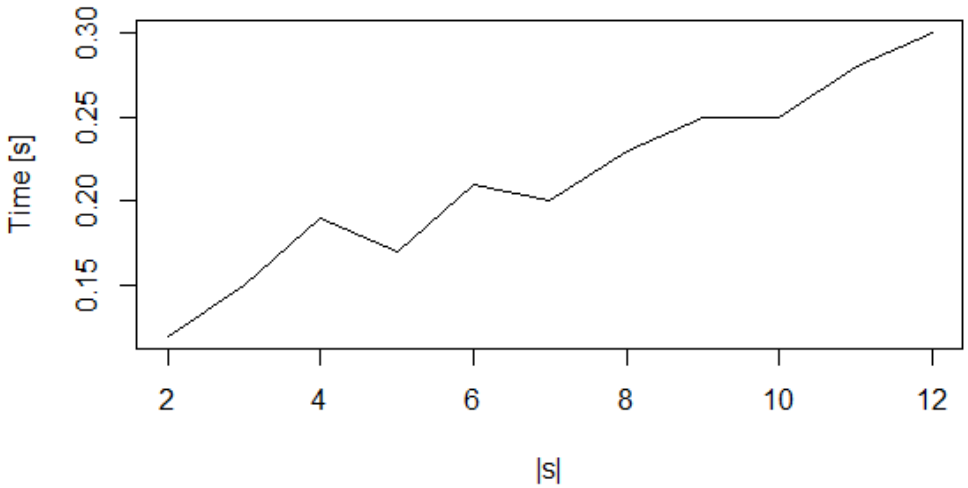

図 2.4　|s|についての計算時間

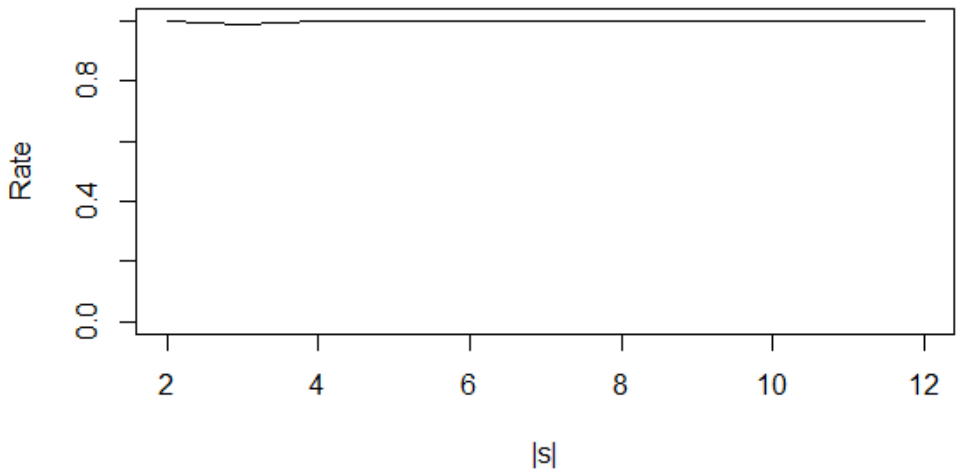

図 2.5　|s|についての再識別成功率

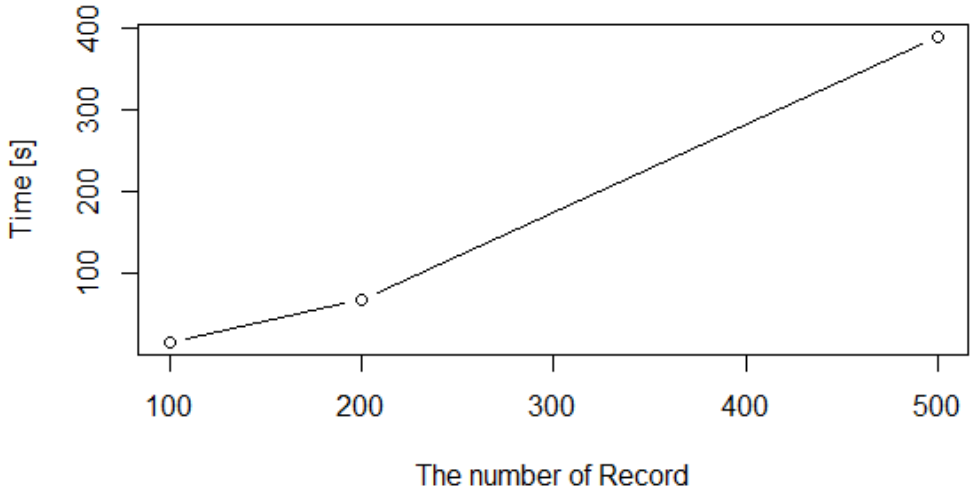

図 2.6　レコード数についての計算時間

PWSCup2015 の匿名加工データを用いて提案再識別手法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，匿名加工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匿名加工データとなり，PWSCup2015 上位チームの匿名加工データはそれらをうまく組み合わせて作成されていること，また，identify.euc (EUC1)は既存手法(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く，差をつけるためには更に計算時間を増やす必要があることが判明した．しかし，EUC2 のように再識別にかなりの時間を必要とする手法のパフォーマンス性能は悪い．

identify.euc のさらなる改善，新たな再識別手法の開発，それらを考慮した上での新たな匿名加工手法の開発を今後の課題とする．

3. 乗降履歴データの取得と分析

前章では静的データに対する匿名加工と再識別について論じたが，本章では動的データの匿名加工手法や評価指標を検討するために，それらのデータからどのような情報を得られるかを調査する．そのため，実際に乗降履歴データを収集し，それらのデータの分析を行う．

3.1. 乗降履歴データの取得

本研究のために，明治大学総合数理学部に所属する 31 人の交通 IC カードから顧客データ M と乗降履歴データ T を作成した．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』を使用した．アプリケーションの仕様上，一人あたりから収集できる履歴は最大 19 件である．表 3.1 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 3.1　取得できる乗降履歴の例

日付	利用内容	使用金額
2016/10/30	入 上野（JR 東北本線） 出 高田馬場（JR 山手線）	-194
2016/10/30	入 高田馬場（JR 山手線） 出 上野（JR 東北本線）	-194
2016/10/8	チャージ 券売機等	2000

表 3.2 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，乗降履歴データ 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 3.3 に顧客データの例を示す．表 3.4 に乗降履歴データの例を示す．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 3.1 の乗降履歴をデータ化したものが表 3.4 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 3.2　取得したデータの概要

	データ種別	データ件数	データ項目	項目
個人情報	顧客データ M	n 31 件	顧客 ID	2 桁数値
			性別	男女
			学年	1 桁数値
			住所	名称
			定期券範囲 1	名称
			定期券範囲 2	名称
	乗降履歴データ T	m 584 件	顧客 ID	2 桁数地
			日付	yyyy/mm/dd
			回数	数値
			乗車駅	名称
			降車駅	名称
			乗車路線	名称
			降車路線	名称
			用途	カテゴリ
			使用場所	カテゴリ
			料金	数値

表3.3　顧客データMの例

ID	性別	学年	住所	定期券範囲1	定期券範囲2
1	男	1	千葉県	NA	NA
2	女	3	東京都	中野	新宿

表3.4　乗降履歴データTの例

顧客ID	日付	回数	乗車駅	降車駅	乗車路線	降車路線	用途	使用場所	料金
1	2016/10/30	2	上野	高田馬場	JR東北本線	JR山手線	交通	NA	-194
1	2016/10/30	1	高田馬場	上野	JR山手線	JR東北本線	交通	NA	-194
1	2016/10/8	1	NA	NA	NA	NA	チャージ	券売機	2000

3.2. 乗降履歴データの分析

本データのユースケースへの適用可能性を明らかにするために，乗降履歴データの「使用料金」と「駅利用回数」に注目して分析を行う．図 3.1 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 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，4～6 月の使用料金が多くなっている．また，図 3.2 にユーザごとの総使用料金を示す．ユーザの総使用料金の統計量を表 3.5 に示す．

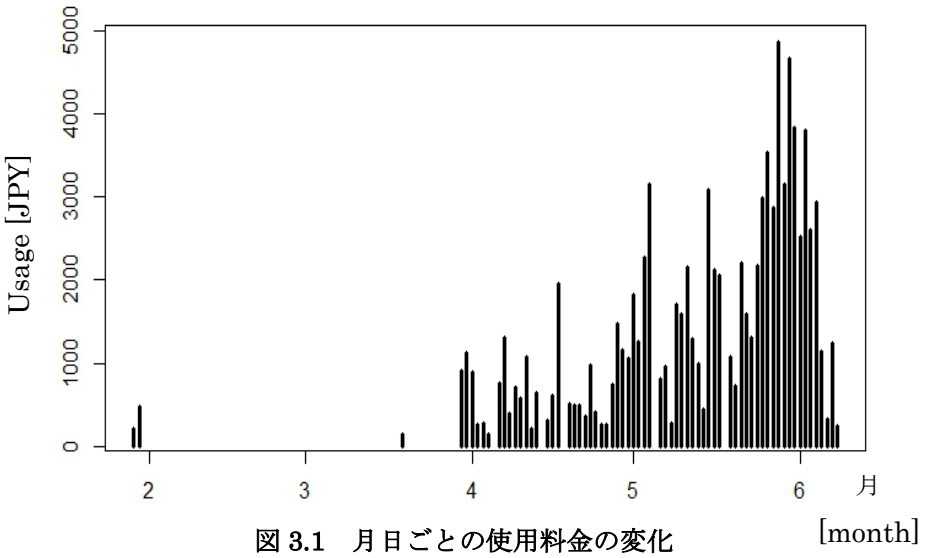

図 3.1　月日ごとの使用料金の変化

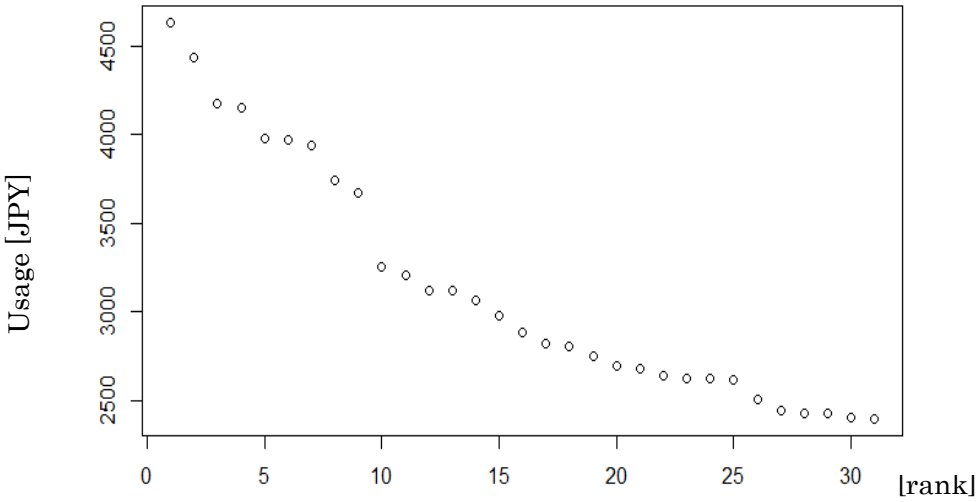

図 3.2　全ユーザの総使用料金

表3.5　総使用料金の統計量(円)

平均	3133.871
最大	4633
最小	2393

図 3.3 に駅ごとの利用回数を示す．被験者の所属する中野キャンパス周辺の中野駅や新宿駅の利用回数が非常に多く，その他の駅の利用回数と大きな差が見られた．表 3.6 に利用回数上位 5 位の駅名と回数を示す．図 3.4 に図 3.3 の結果を地図上にプロットした結果を示す．

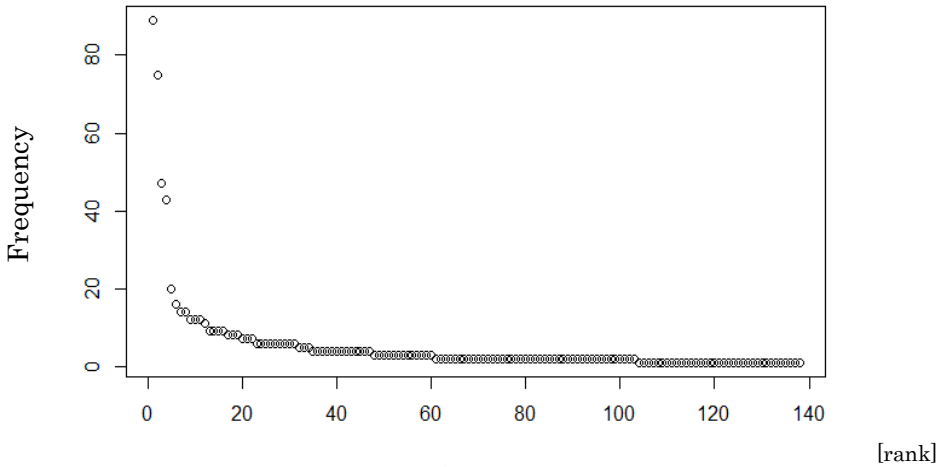

図 3.3　駅ごとの利用回数

表3.6　利用回数上位の駅名と回数

新宿	中野	渋谷	高田馬場	明大前
89	75	47	43	20

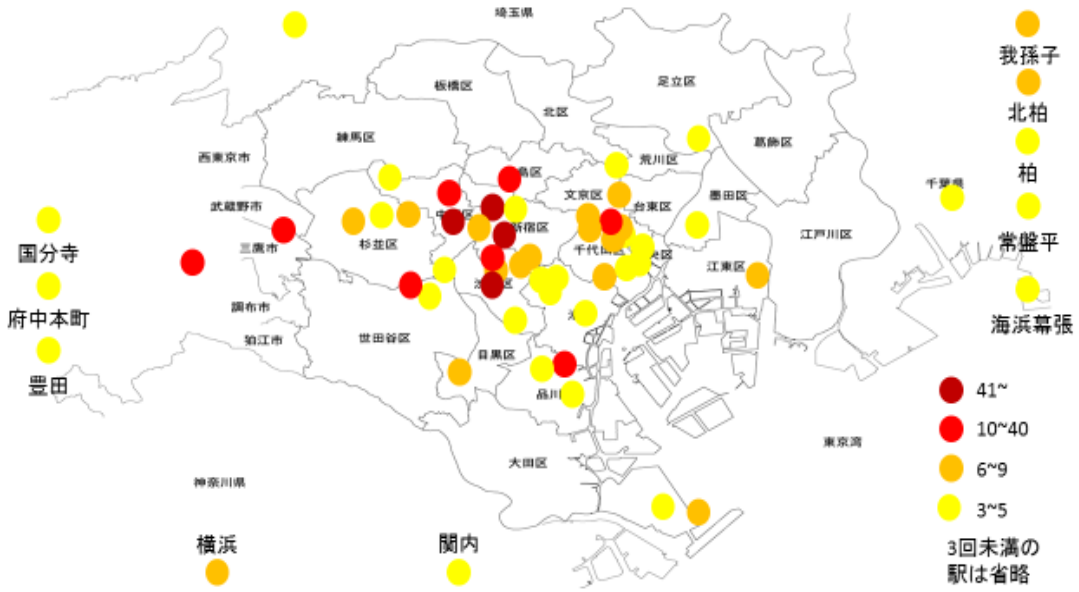

図 3.4　利用駅の分布

図 3.5 にユーザ間の利用駅についての類似度の度合いを表す Jaccard 距離を示す分布を示す．Jaccard 距離の定義式と計算例を以下に示す．

定義式　A, B: 集合　J(A, B): AB 間の Jaccard 距離

$$\mathrm{J(A,B)} = 1 - \frac{|A \cap B|}{|A \cup B|}$$

計算例　A=(新宿，中野，高田馬場)　B=(新宿，上野)

$$\mathrm{J(A,B)} = 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \frac{|(新宿)|}{|(新宿,\ 中野,\ 高田馬場,\ 上野)|} = 1 - \frac{1}{4} = 0.75$$

Jaccard 距離が 0 に近いほど集合は類似している．本データの平均 Jaccard 距離は 0.933 であった．このことより，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．表 3.7 に顧客属性(性別，学年)のクロス集計表を示す．

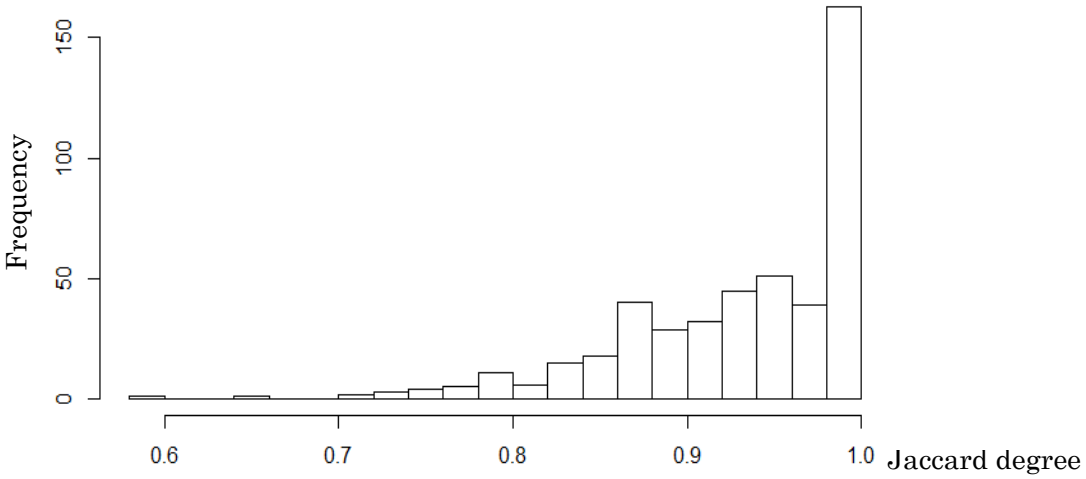

図 3.5　ユーザ間の駅利用についての Jaccard 距離の分布

表3.7　顧客属性のクロス集計表

性別/学年	0(教授)	1	2	3	4	計
男	1	6	5	4	10	26
女	0	2	0	2	1	5

4. 乗降履歴データのユースケース・評価指標・加工手法について

4.1. 乗降履歴データのユースケース

本データのユースケースを表 4.1 のように想定する．本ユースケースは，本データを用いて明治大学総合数理学部に所属する人に対して広告・勧誘を行う効果的な場所を選定することを想定している．なお，ユースケースの作成には経済産業省の匿名加工情報作成マニュアル[7]を参考にした．例えば，外部組織が明治大学総合数理学部に所属する 3・4 年生の男性に対して広告・勧誘を行う場合，顧客属性が「性別=男，学年=3or4」の全ユーザの駅利用回数を用いる．

表 4.1　想定するユースケース

匿名加工情報	顧客属性に応じた駅利用回数
業務サービス概要	明治大学総合数理学部に所属する人が利用している駅， またその回数を顧客属性(性別・学年)に応じて適した広告を配信する
提供する属性	M(顧客 ID，性別，学年) T(顧客 ID，乗車駅，降車駅，乗車路線，降車路線)
匿名加工情報利用目的	利用者に応じた最適な広告・勧誘を行うこと

4.2. 乗降履歴データの評価指標

4.2.1.　　有用性指標

前節で想定したユースケースに対応する有用性評価指標を検討する．[4]の匿名加工データの有用性評価の多くは「元データの特徴をどれだけ保持しているか」という観点で評価されている．そこで，本ユースケースでは，以下の特徴を保持できているかどうかで匿名加工データの有用性を評価する．

1　顧客属性(性別，学年)毎の駅利用回数(U'1)

2　駅利用回数の順位(上位のみ)(U'2)

3　顧客属性(性別，学年)のクロス集計の人数(U'3)

また，これらの評価を行う有用性指標を順に U'1, U'2, U'3 とする．各指標の式を以下に示す．M，T が元データを表し，M*，T*が加工されたデータを表す．$T_{station}(X_i)$は T についてのグループX_iの駅利用総回数を表す．g は T のグループ数を表す．S_Nを上位 N 駅の集合とし，N は変数である．rank(T,s)は駅 s の T における利用回数順位を表している． $Cross_{sex,grade}$は M の(性別，学年)属性についてのクロス集計値を表し，$num(\textit{属性名})$はその属性の種類数を表す．これらの有用性指標の値が 0 に近いほど，データ(T*,M*)の有用性は高い．

$$U'_1(M,T,M^*,T^*) = \frac{\sum_g^{i=1} |T_{station}(X_i) - T^*_{station}(X_i)|}{g}$$

$$U'_2(M,T,M^*,T^*) = N - |\{s \in S_N(rank(T,s) = rank(T^*,s))\}|$$

$$U'_3(M,T,M^*,T^*) = \frac{\sum_{num(sex)}^{i=1} \sum_{num(grade)}^{j=1} |Cross_{sex,grade}(i,j) - Cross^*_{sex,grade}(i,j)|}{num(sex) * num(grade)}$$

4.2.2. 安全性指標

本データをそのまま外部組織に提供してしまうと顧客個人が特定されてしまう場合がある．本ユースケースでは以下の場合が考えられる．

1 顧客属性(性別，学年)の組み合わせが特殊である(S'1)
2 特殊な駅(利用回数が 3 回未満，または東京都外にある)を利用している(S'2)

例えば本データの場合，表 7 より顧客属性が「性別=女，学年=4」であるユーザは 1 人しかいないため，個人が特定されてしまう．また，特殊な駅を利用している場合も個人が特定されやすい．例えば静岡駅を繰り返し利用している履歴があった場合，その履歴は住所が静岡県の顧客のものである可能性が高い．これらの評価を行う有用性指標を順に S'1, S'2,とし，各指標の評価値の定義を以下に示す．S'1, S'2 の値が大きいほど匿名加工データの安全性は低く，再識別されやすい．

$S'_1(M,T,x)$

$= M$の特殊なユーザ(同じ顧客属性の組み合わせを持つユーザが本人含めx人以下)の数

$S'_2(M,T,y) = T$の特殊な駅(利用回数がy回以下，または東京都外にある)の数

4.3. 乗降履歴データの加工手法

本節では，前節で定義した評価指標について，有用性と安全性がともに高いデータを作成する．加工対象として，表 4.2 に簡易顧客データ M，表 4.3 に簡易乗降履歴データ T を示す．これらのデータの安全性は，x=1，y=1 とすると$S'_1(M,T,1) = 1$，$S'_2(M,T,1) = 1$である．

表 4.2　簡易顧客データ M

顧客ID	性別	学年	group
1	男	1	A
2	男	1	A
3	男	2	B
4	男	2	B
5	女	4	C

表 4.3　簡易乗降履歴データT

顧客ID	乗車駅	降車駅	group
1	新宿	品川	A
1	品川	新宿	A
2	高田馬場	新宿	A
2	新宿	中野	A
3	中野	新宿	B
3	新宿	中野	B
4	高田馬場	品川	B
4	品川	熱海	B
5	中野	東京	C
5	東京	中野	C

まず，前章で定義した U'1, U'2, U'3 を損なわない加工手法を考える．U'1 は駅利用回数についての有用性指標であるため，なるべく駅利用回数を保持する必要がある．しかし，顧客属性

が同じユーザのグループ内で利用駅(乗車駅，降車駅)をシャッフルしても顧客属性ごとの駅利用回数は変化しないため，U'1 を損なうことはない．全体の駅利用回数も変化しないため，U'2 を損なうこともなく，顧客データは加工しないため U'3 も損なわない．この手法を「グループ内シャッフル」とする．表 4.3 の乗降履歴データ T の利用駅をグループ内シャッフルした結果 T*を表 4.4 に示す．*の部分が加工された箇所である．

表 4.4　グループ内シャッフルされた乗降履歴データT*

顧客ID	乗車駅	降車駅	group
1	**新宿***	**新宿***	A
1	**高田馬場***	**品川***	A
2	**品川***	**中野***	A
2	**新宿***	**新宿***	A
3	**品川***	**新宿***	B
3	**中野***	**品川***	B
4	**高田馬場***	**中野***	B
4	**新宿***	**熱海***	B
5	中野	東京	C
5	東京	中野	C

次に，S'1, S'2 の値を下げるようにデータを加工する．グループ内シャッフルのみだと顧客データは無加工であるため，表 3.7 より「性別=男，学年=0」と「性別=女，学年=4」の顧客が容易に特定されてしまう(S'1)．表 4.5 に加工した顧客データ M*を示す．この場合，顧客データの属性の組み合わせが特殊な顧客は(性別=女，学年=4)であるため，グループ B と同じ顧客属性(性別=男，学年=2)に加工する．

表 4.5　加工された顧客データM*

顧客ID	性別	学年	group
1	男	1	A
2	男	1	A
3	男	2	B
4	男	2	B
5	**男***	**2***	**B***

特殊な駅を利用している顧客も特定されやすい(S'2)．これらを解決するために，顧客属性の組み合わせが特殊な顧客を別のグループに移し，特殊な駅の利用履歴を全て利用回数 1 位の「新宿」に置き換える．これらの加工では顧客属性ごとの駅利用回数や人数が変わってしまうため U'1 と U'3 を損なってしまうが，駅利用順位は変わらない(1 位の利用回数がさらに増えるだけ)ので，U'2 は損なわない．表 4.6 に加工された乗降履歴データ T**を示す．この場合特殊な駅は「熱海」であるため，利用回数 1 位の「新宿」に置き換える．

表 4.6　加工された乗降履歴データT**

顧客ID	乗車駅	降車駅	group
1	新宿	新宿	A
1	高田馬場	品川	A
2	品川	中野	A
2	新宿	新宿	A
3	品川	新宿	B
3	新宿	品川	B
4	高田馬場	中野	B
4	新宿	新宿*	B
5	中野	東京	B
5	東京	中野	B

よって，想定したユースケースに対応する評価指標を作成し，それを満たし，かつ個人が特定されにくい匿名加工データ M*，T**を作成した．表 4.7 に T, T*, T**, M, M*についての U'1~U'3, S'1, S'2 の値の変化を示す．M を加工したことによって U'1 と U'3 が上がってしまったが，U'2 の値は変化しておらず，有用性を保てている．また S'1, S'2 の値も下がっており，安全性を高めている．

表 4.7　評価値の変化

	M,T	M,T*	M*,T*	M*,T**
U'1	0	0	2.67	2.67
U'2	0	0	0	0
U'3	0	0	0.2	0.2
S'1	1	1	0	0
S'2	1	1	1	0

5. プチ PWSCUP とその有用性指標について

5.1. プチ PWSCUP について

3 章で作成した乗降履歴データを用いた実験として，小規模な匿名加工・再識別コンテスト「プチ PWSCUP」を研究室内で開催した．運営と参加は私を含む明治大学総合数理学部に所属する 4 名で行い，私は主に「データの収集・作成」と「有用性評価指標の実装」を担当した．プチ PWSCUP は 2016 年 8 月に開催され，計 47 個のデータが提出された．本章では主に「プチ PWSCUP の有用性評価指標」について論ずる．プチ PWSCUP の評価プラットフォームについては[8]を参考されたい．匿名加工データの評価は PWSCUP2015 と同様，「(安全性順位+有用性順位)/2」で行っている．

5.2. プチ PWSCUP の有用性指標

プチ PWSCUP の有用性指標一覧と，それらの指標が注目している属性を表 5.1，5.2 に示す．これらの指標は R 言語と python で実装した．

表 5.1　プチ PWSCUP の有用性指標一覧

指標名	作成者	内容
reoh_U1_name.py	原田	id 毎の合計値の平均絶対誤差
reoh_U2_sma5.py	原田	5 日移動平均の平均絶対誤差
satoshi_U1_mae_fare.R	伊藤	日付ごとの平均料金の差の平均絶対誤差
satoshi_U2_mae_record.R	伊藤	日付ごとのレコード数の差の平均絶対誤差
satoshi_U3_mae_user.R	伊藤	ユーザごとのレコード数の差の平均絶対誤差
satoshi_U4_AYA.R	伊藤	レコード番号ごとの合計料金の平均絶対誤差
satoshi_U5_time.R	伊藤	回数のレコード数の差の平均絶対誤差
satoshi_U6_station1.R	伊藤	駅の登場回数の平均絶対誤差
satoshi_U7_station2.R	伊藤	路線の登場回数の平均絶対誤差　※
satoshi_U8_AYA2.R	伊藤	レコード番号ごとの ID の平均絶対誤差
satoshi_U9_use1.R	伊藤	用途と使用場所の組み合わせの平均絶対誤差
satoshi_U10_use2.R	伊藤	用途ごとの合計金額の平均絶対誤差

表 5.2　プチ PWSCUP の有用性指標が注目している属性

属性名/指標名	レコード	名前	顧客ID	日付	回数	乗車駅	降車駅	乗車路線	降車路線	用途	使用場所	料金
reoh_U1_name.py	-	-	○	-	-	-	-	-	-	-	-	○
reoh_U2_sma5.py	-	-	-	○	-	-	-	-	-	-	-	○
satoshi_U1_mae_fare.R	-	-	-	○	-	-	-	-	-	-	-	○
satoshi_U2_mae_record.R	-	-	-	○	-	-	-	-	-	-	-	-
satoshi_U3_mae_user.R	-	-	○	-	-	-	-	-	-	-	-	-
satoshi_U4_AYA.R	○	-	-	-	-	-	-	-	-	-	-	○
satoshi_U5_time.R	-	-	-	-	○	-	-	-	-	-	-	-
satoshi_U6_station1.R	-	-	-	-	-	○	○	-	-	-	-	-
satoshi_U7_station2.R	-	-	-	-	-	-	-	○	○	-	-	-
satoshi_U8_AYA2.R	○	-	○	-	-	-	-	-	-	-	-	-
satoshi_U9_use1.R	-	-	-	-	-	-	-	-	-	○	○	-
satoshi_U10_use2.R	-	-	-	-	-	-	-	-	-	○	-	○
計	2	0	3	3	1	1	1	1	1	2	1	5

これらの指標では主にクロス集計と平均絶対誤差を用いて評価値を出しており，表 5.2 はその計算にどの属性を用いているかを表している．例えば satoshi_U10_use2.R では乗降履歴データの「用途」「料金」属性を計算に用いている．以下に定義式と数値例を示す．

定義式　T：加工前データ　T'：加工後データ　$U_{satoshi\_10}(\mathrm{T},\mathrm{T}')$ ：T'の評価値

$Cross_{use,fare}(\mathrm{T},\mathrm{i})$: T についての用途ごとの合計料金のクロス集計の i 番目の要素

num(use)：用途属性の種類数(本データの場合 5 種類)

$$U_{satoshi\_10}(\mathrm{T},\mathrm{T}') = \frac{\sum_{i=1}^{num(use)} |Cross_{use,fare}(\mathrm{T},\mathrm{i}) - Cross_{use,fare}(\mathrm{T}',\mathrm{i})|}{num(use)}$$

数値例　T：表 5.3　T’：表 5.4　$Cross_{use,fare}$(T)：表 5.5　$Cross_{use,fare}$(T)：表 5.6

$$U_{satoshi\_10}(\mathrm{T},\mathrm{T}') = \frac{|(-700)-(450)|+|(-400)-(-300)|+|2000-750|+|0|+|0|}{5} = 500$$

表 5.3　簡易加工前データ T

顧客 ID	用途	料金
1	1	-200
1	2	-400
2	1	-500
2	3	2000

表 5.4　簡易加工後データ T’

顧客 ID	用途	料金
1	1	-300
1	2	-300
2	1	750
2	3	750

表 5.5　T についてのクロス集計

用途	1	2	3	4	5
合計料金	-700	-400	2000	0	0

表 5.6　T’についてのクロス集計

用途	1	2	3	4	5
合計料金	450	-300	750	0	0

また，「レコード」「名前」属性は作成した乗降履歴データには本来無いものだが，評価指標の仕様上加えている．「レコード」属性は履歴の番号を示した属性であり，「名前」属性は顧客の本名を示した属性である．また，匿名加工データは PWSCUP2015 と同様に，全有用性指標の平均順位によって有用性評価が求められる．詳細は[2]を参考されたい．

5.3. プチ PWSCUP の問題点・反省点

プチ PWSCUP を開催し，その結果を分析したところ，様々な問題点や反省点が見つかった．主な問題点・反省点は以下の 2 つである．

1　乗降履歴データの属性間で価値の差が生じてしまった

2　山岡匿名化の対策が不十分だった

問題点 1 について説明をする. 5.2 節の表より, 有用性指標が注目している属性に偏りがある. 例えば「料金」属性は 5 つの指標で用いられているのに対し,「乗車駅」属性は 1 つの有用性指標でしか用いられていない. つまり,「料金」属性を加工するとデータの有用性が下がりやすいが,「乗車駅」属性を加工してもデータの有用性は下がりにくい. 属性間の価値の差が生じてしまうと匿名加工データを正しく評価することができないため, 解決する必要がある.

また, 問題点 2 について説明をする. PWSCUP2015 では山岡匿名化という匿名加工手法が猛威を振るい, それによって加工された匿名加工データが上位を独占した. この結果を受けて翌年の PWSCUP2016 では足切り指標が導入され, 山岡匿名化をしていると思われる匿名加工データは提出できない仕組みができた. しかし, プチ PWSCUP ではこれらの仕組みが導入されていないため, PWSCUP2015 同様山岡匿名化が猛威を振るう結果となった. 図 5.1 にプチ PWSCUP に提出された匿名加工データの分布を示す.

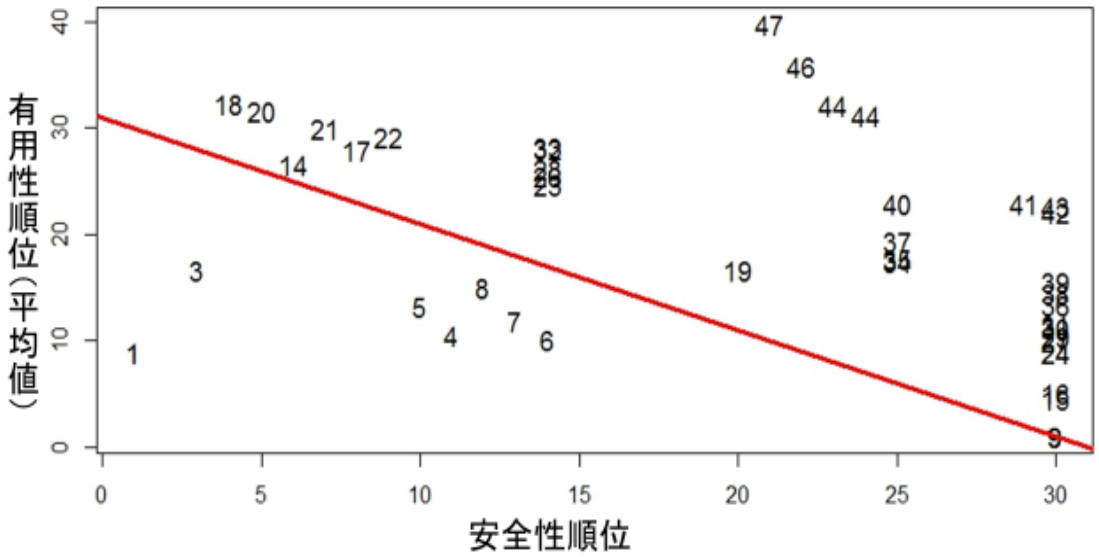

図 5.1　プチ PWSCUP のデータ分布

縦軸が有用性順位を示し, 横軸が安全性順位を示している. また, 図中の数字はデータの総合順位を示しており, 1 位(2 個重なっている)と 3 位のデータは山岡匿名化されたデータである.

図中の線は元データ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おり, これは問題点 1 の影響と考えられる.

5.4. 追加で実装した有用性指標とその結果

前節で説明した問題点 1, 2 を解決するために，新たな有用性指標を実装した．追加指標一覧と注目している属性を表 5.7，5.8，5.9 に示す．また，問題点 2 を解決するために，5.2 節で紹介した有用性指標のうち satoshi_U4_AYA.R と satoshi_U8_AYA2.R を足切り指標に切り替えた．

表 5.7　追加で実装した有用性指標

指標名	作成者	内容
U11_rank_station.R	伊藤	利用駅の順位が保存できているかどうか
U12_rank_route.R	伊藤	利用路線の順位が保存できているかどうか
U13_rank_use_location.R	伊藤	用途と使用場所の順位が保存できているかどうか
U14_jaccard_station.R	伊藤	利用駅についての類似度の平均絶対誤差
U15_jaccard_route.R	伊藤	利用路線についての類似度の平均絶対誤差
U16_jaccard_use.R	伊藤	用途についての類似度の平均絶対誤差
U17_jaccard_location.R	伊藤	使用場所についての類似度の平均絶対誤差
satoshi_U4_AYA.R	伊藤	山岡匿名化を足切りするための指標
satoshi_U8_AYA2.R	伊藤	山岡匿名化を足切りするための指標

表 5.8　追加した指標で注目している属性

属性名/指標名	レコード	名前	顧客ID	日付	回数	乗車駅	降車駅	乗車路線	降車路線	用途	使用場所	料金
U11_rank_station.R	-	-	-	-	-	○	○	-	-	-	-	-
U12_rank_route.R	-	-	-	-	-	-	-	○	○	-	-	-
U13_rank_use_location.R	-	-	-	-	-	-	-	-	-	○	○	-
U14_jaccard_station.R	-	-	-	-	-	○	○	-	-	-	-	-
U15_jaccard_route.R	-	-	-	-	-	-	-	○	○	-	-	-
U16_jaccard_use.R	-	-	-	-	-	-	-	-	-	○	-	-
U17_jaccard_location.R	-	-	-	-	-	-	-	-	-	-	○	-
計	0	0	0	0	0	2	2	2	2	2	2	0

表 5.9　有用性指標で注目している属性の数

属性名/指標名	レコード	名前	顧客ID	日付	回数	乗車駅	降車駅	乗車路線	降車路線	用途	使用場所	料金
既存有用性指標(U4, U8 除く)	0	0	2	3	1	1	1	1	1	2	1	4
追加有用性指標	0	0	0	0	0	2	2	2	2	2	2	0
計	0	0	2	3	1	3	3	3	3	4	3	4

有用性指標を追加することにより，問題点 1 を解決することができた．また，足切り指標を導入することによって山岡匿名化をしたデータをはじくことも可能となり，問題点 2 を解決することもできた．有用性指標を追加したプチ PWSCUP を「新プチ PWSCUP」，追加する前のものを「旧プチ PWSCUP」とし，新プチ PWSCUP のデータ分布を図 5.2 に示す．旧プチ PWSCUP は大半のデータ(47 個中 34 個)が元データより悪い評価であったが，新プチ PWSCUP ではその数が半分ほど(47 個中 23 個)になっている．しかし，元データより評価が悪いデータが少ないことが良いことであるのかは不明である．また，図中の 1～3 位のデータは山岡匿名化で加工されたデータであるため，足切り指標により実際には評価対象外となる．しかし，有用性評価指標が 17 個と非常に多くなってしまったため，改善する必要がある．表 5.10 に旧プチ PWSCUP と新プチ PWSCUP を比較した結果を示す．

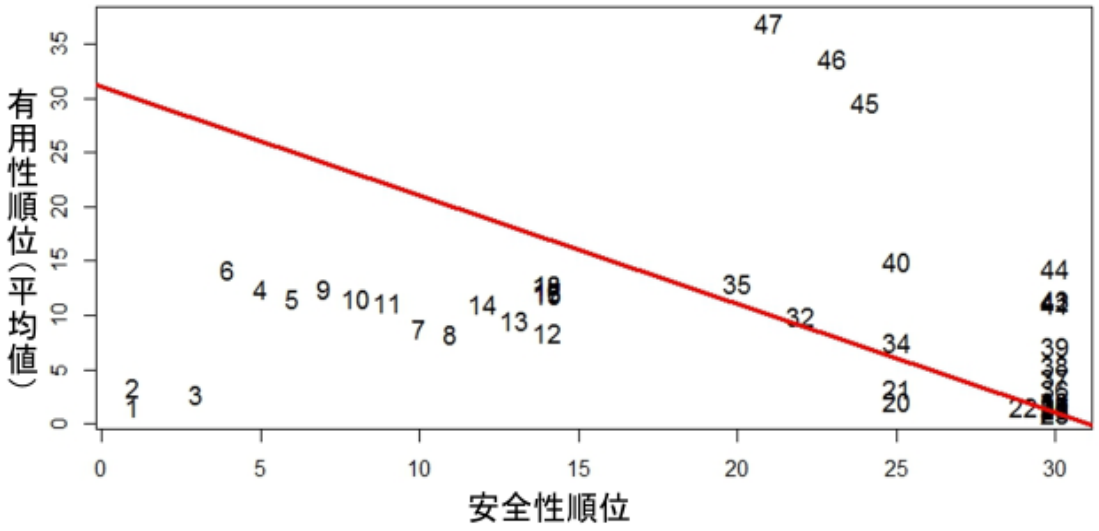

図 5.2　新プチ PWSCUP のデータ分布

表 5.10　旧プチ PWSCUP と新プチ PWSCUP の比較結果

	旧プチ PWSCUP	新プチ PWSCUP
有用性指標の数	12	17
安全性指標の数	6	6
足切り指標の数	0	2
元データより評価が悪いデータの数	34	23
乗降履歴データの属性間での価値の差	大	小
元データより評価が良い山岡匿名化されたデータの数	3	0

6. おわりに

本稿では，「静的データの匿名加工・再識別」，「乗降履歴データの作成・分析」，「動的データの有用性指標と匿名加工」，「乗降履歴データを用いたプチ PWSCUP」について論じた.

謝辞

本研究を進めるにあたり，多くのご指導をいただいた菊池浩明教授，様々な点で研究に協力してくれた菊池研究室匿名加工班の皆さま，identify.euc の評価を行うにあたり，匿名加工データとその行番号データを提供していただいた匿名加工・再識別コンテスト PWSCup2015 の参加チームの方々，動的データを作成するにあたり，乗降履歴を提供していただいた明治大学総合数理学部菊池研究室の方々に感謝いたします.

参考文献

[1] 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/press/2013/20130716.pdf, 2017 年 1 月参照.)

[2] 菊池浩明，山口高康，濱田浩気，山岡裕司，小栗秀暢，佐久間 淳，“匿名加工・再識別コンテスト Ice & Fire の設計”, CSS 2015, pp.363-370, 2015.

[3] 南和宏，“プライバシー保護データパブリッシング”，情報処理 Vol. 54，No. 9，pp. 938-946，2013.

[4] 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加工せよ”, CSS2016, pp.271-278, 2016. (https://pwscup.personal-data.biz, 2016 年 12 月参照.)

[5] 秋山他，“教育用擬似ミクロデータの開発とその利用～平成 16 年全国消費実態調査を例として～”，統計センター製表技術参考資料, 16, pp.1-43, 2012.

[6] 菊池浩明，山口高康，濱田浩気，山岡裕司，小栗秀暢，佐久間淳，“匿名加工コンテスト PWSCUP2015 の報告と匿名加工方法の評価”，SCIS，2016.

[7] 経済産業省，事業者が匿名加工情報の具体的な作成方法を検討するにあたっての参考資料 (「匿名加工情報作成マニュアル」)Ver1.0 (http://www.meti.go.jp/press/2016/08/20160808002/20160808002-1.pdf, 2016 年 12 月参照.)

[8] 原田玲央，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，明治大学総合数理学部 2016 年度卒業論文