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

伊藤 聡志[1]　原田 玲央[1]　菊池 浩明[1]

概要：鉄道の乗降履歴には，市場調査の観点などで有益な情報が含まれている．しかしながら，様々な再識別リスクがあるために，その匿名加工は自明ではない．そこで，我々は 31 人の被験者から取得した小規模な乗降履歴データを用いてコンテスト形式の実験を行い，匿名加工手法を検討する．また，交通 IC カードから取得できる履歴データは交通用途のものだけではなく，物販やチャージ用途のものも含まれる．本稿ではそれらに対するリスク評価を行う．

キーワード：匿名加工，乗降履歴データ

Study on Identification Risk and Anonymization of Trajectory and Purchase History Data

Satoshi Ito[1]　Reo Harada[1]　Hiroaki Kikuchi[1]

1. はじめに

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，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．評価指標はデータの安全性や有用性を評価する指標であり，匿名加工は顧客データのような個人情報データから個人が特定されないように，データを加工することである．例えば，購買データを想定した評価指標・匿名加工に [1][2] がある．しかし，実際に顧客から同意を取りデータ収集するのは困難であった．加えて，乗降のみ，または，物販の購買のみのデータはあるが，その 2 つを照合する困難さを調べるには，それらの相関が不明であった．

そこで，本研究の目的を実際に顧客から収集したデータを用いた匿名加工の方法や，識別リスクと，複数のデータセットを突合させることによるリスクを評価することとする．31 人の交通 IC カードから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・匿名加工手法をコンテスト形式の実験を行い，検討する．

[1] 明治大学
Meiji University,Nakano,Tokyo,4-21-1

本研究では，交通 IC カードから取得できる履歴データが交通用途のものだけではなく，物販やチャージ用途のものも含まれることに注目した．乗降履歴と購買履歴を組み合わせることで，ユーザの識別に関するリスク (エントロピー) がどれくらい増大するかを明らかにする．本稿では，各用途について分析を行い，エントロピーや用途間の関係等を用いて，履歴データの識別リスクを定量化する．

本稿では，2 節で取得したデータの分析を行い，3 節でデータの識別リスクを分析し，4,5 節でデータの匿名加工を検討する．6 節で匿名加工データを評価する際に行ったコンテストについて述べる．

2. 交通系 IC カード履歴データの取得

2.1 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 人から同意を取り，交通 IC カードから，顧客データ M と利用履歴データ 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 [3] 」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 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 2 に取得した本データの概要を示す．顧客データ M(マ

表 1　取得できる履歴データの例

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 2　作成したデータの概要

	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	ℓ 584 件	顧客 ID	2 桁数値
			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　顧客データ M の例

顧客 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

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 3 に顧客データの例を示す．表 4 に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 1 の履歴をデータ化したものが表 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．

2.2　基本統計

第 4 節で示すユースケースへの適用可能性を明らかにするために，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 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，4,5,6 月の使用料金が多くなっている．また，図 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 5 に示す．表 6 に利用回数上位 5 位の駅名と回数を示す．表 7 に顧客属性 [性別, 学部学年 (教員を 0 とする)] のクロス集計表を示す．

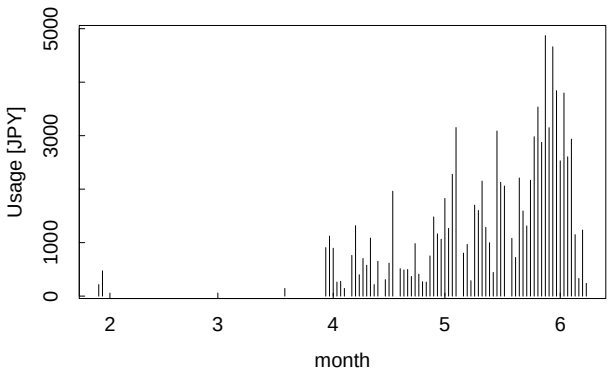

図 1　月日ごとの使用料金の変化

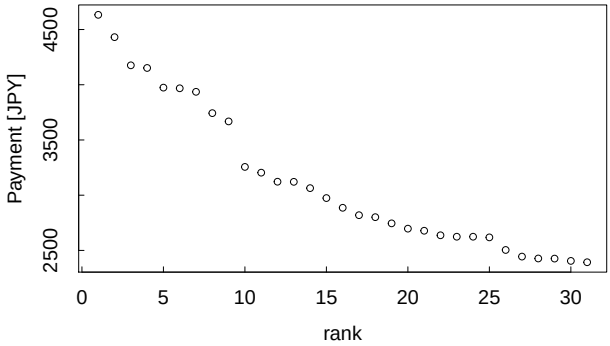

図 2　全ユーザの総使用料金

我々は，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 8 に取得した履歴の用途別の割合を示し，図 3 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 4 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，例えば，7 件の履歴の内，1 件のみ他と共通の時の距離 $1-\frac{1}{7+7-1}=\frac{12}{13}$ に相当している．従って，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．

2.3　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 5 ,6 ,7 に「交通」,「物販」,「チャージ」用途の履歴の

表 4　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

表 5　総使用料金の統計量 (円)

平均	最大	最小
3133.871	4633	2393

表 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

表 7　顧客属性のクロス集計表

性別 \ 学年	0	1	2	3	4	計
男	1	6	5	4	10	26
女	0	2	0	2	1	5

表 8　用途別の内訳

用途	交通	物販	チャージ	バスチャージ	共通	計
レコード数	364	100	84	2	34	584
割合	62.3%	17.1%	14.4%	0.3%	5.8%	100.0%

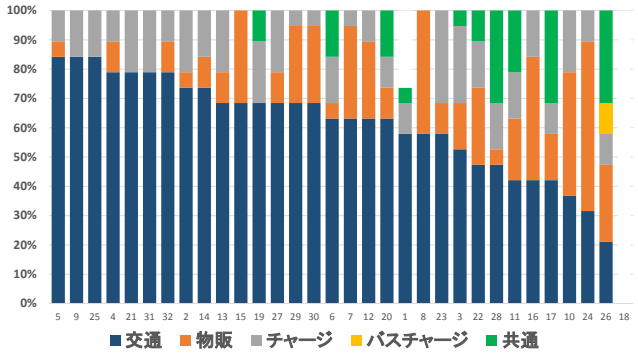

図 3　ユーザ別の用途の内訳

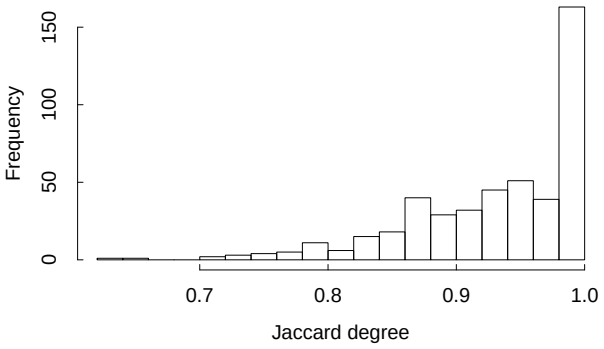

図 4　ユーザ間の駅利用についての Jaccard 距離の分布

出現頻度を示し，表 9 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．

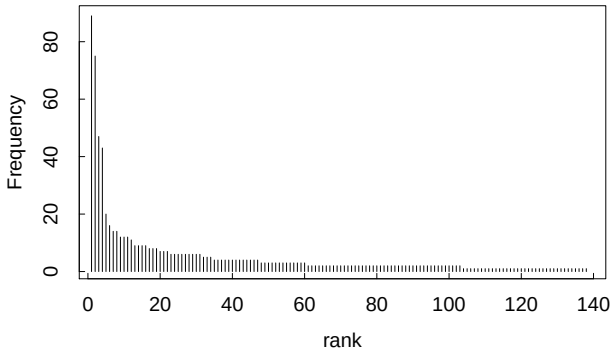

図 5　利用駅の出現頻度

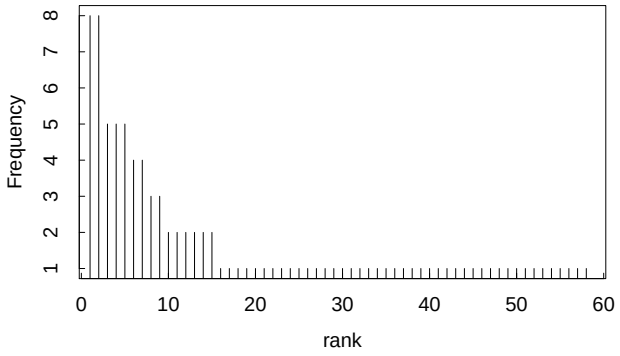

図 6　物販料金の出現頻度

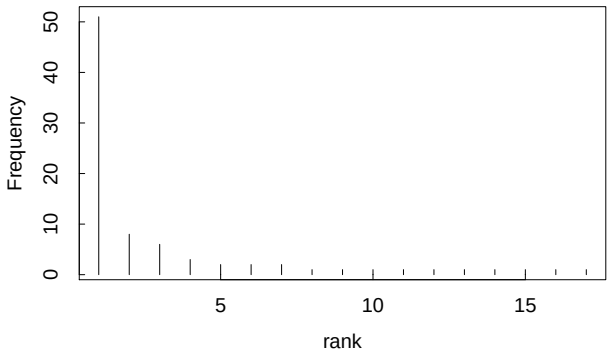

図 7　チャージ料金の出現頻度

表 9　特異なデータの割合

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

3. 識別リスク分析

3.1 リスクの考え方

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 10 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, s_2, s_3 は駅名である．例えば u_1 は s_1 を 2 回，s_2 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 $H(U)$ は，$P(U = u_i)$ をデータ E_S 中で u_i の履歴の生起確率，n をユーザー数としたとき，

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

で与えられる．E_S の場合，全 19 履歴のうち，u_1 のものは 3 回であるため，$P(U = u_1) = 3/19$,$P(U = u_2) = 8/19$,$P(U = u_3) = 8/19$ である．よって，$H(U) = 1.47$[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 $H(U|S)$ を考える．$P(S = s_i)$ を E_S 中で駅 s_i の履歴の生起確率，m を駅の種類数，

$$H(U|S = s_i) = -\sum_{j=1}^{n} P(U = u_j|S = s_i) \log_2 P(U = u_j|S = s_i)$$

を s_i 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，

$$H(U|S) = \sum_{i=1}^{m} P(S = s_i) H(U|S = s_i)$$

で与えられる．E_S の場合，

$$\begin{aligned} H(U|S) &= \sum_{i=1}^{3} P(S = s_i) H(U|S = s_i) \\ &= \frac{10}{19} 1.52 + \frac{5}{19} 0.72 \\ &= 0.99 \end{aligned}$$

である．

最後に，相互情報量 $I(U;S)$ 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，

$$I(U;S) = H(U) - H(U|S)$$

で与えられる．E_S の場合，$I(U;S) = 0.48$ である．

$H(U)$,$H(U|S)$,$I(U;S)$ 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，$H(U) = 1.47$ 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 $1/2^{H(U)} = 0.36$ 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 が分かると一意に u_2 であることが特定されるが，s_2 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 $H(U|S) = 0.99$ になり，その平均確率は $1/2^{H(U|S)} = 0.5$ である．このとき 1 つの駅利用履歴から得た情報量は $I(U;S) = 0.48$bit であるため，

$$H(U) = 1.47 < 1.92 = 4I(U;S)$$

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．

表 10　ユーザごとの利用駅集計のデータ例 E_S

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

表 11　用途別のエントロピー等の値

	交通 (S)	物販 (B)	チャージ (C)	交通・物販 (S, B)
$H(U)$	4.900	4.338	4.736	4.412
$H(U\|x)$	1.814	0.948	3.256	0.182
$I(U;x)$	3.085	3.389	1.479	4.230
$P(U\|x)$	0.284	0.518	0.105	0.881
n_x	31	25	29	31
m_x	138	58	17	8004

3.2 多様なエントロピー

取得した履歴データの，用途「交通」,「物販」,「チャージ」についてまとめた集計表を順に D_S,D_B,D_C とする．ユーザ (U) の数は 31 人，利用駅 (S) の種類は 138 種，物販料金 (B) の種類は 58 種，チャージ (C) 料金の種類は 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 11 に用途別のエントロピー等の値を示す．x は特定の用途を示す．$x =$「交通」用途の場合，$H(U) = 4.900$, $I(U;S) = 3.085$ 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 $1/2^{H(U)} = 0.033$ 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 $1/2^{H(U)-I(U;S)} = 1/2^{H(U;S)} = 0.284$ まで上がる．n_x はユニークユーザ数であり，交通 IC カードを用途 x で利用しているユーザの数である．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，$P(U|x)$ は用途 x の履歴を取得した場合に，データ中のあるユーザが特定される平均確率である．

3.3 用途の相関

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 8 に交通利用回数とチャージ料金の関係を示し，図 9 に交通利用料金と

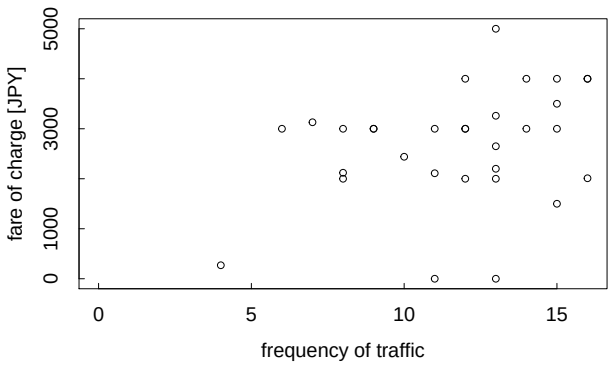

図 8　交通利用回数とチャージ料金の散布図

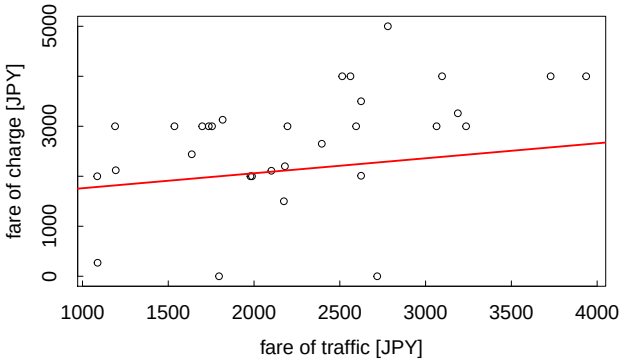

図 9　交通利用料金とチャージ料金の散布図

チャージ料金の関係を示す．相関係数は順に 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 12 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 と E_B の u_1, u_2, u_3 は同じユーザである．3.1 節と同様の手順で計算すると，$H(U) = 0.98$，$H(U|B) = 0.57$，$I(U;B) = 0.41$ 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 13 に示す．$E_{S,B}$ の場合，例として，

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

と表すことができ，$H(U) = 1.19$，$H(U|S,B) = 0.46$，$I(U;S,B) = 0.73$ が与えられる．E_S,E_B,$E_{S,B}$ の各値を表 14 に示す．$I(U;x)$ の行より，

$$I(U;S) + I(U;B) = 0.89 > 0.73 = I(U;S,B)$$

である．このことから，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 11 に示す．$I(U;S)$,$I(U,B)$ 等の結果と比較すると，

$$I(U;S) + I(U;B) = 6.474 > 4.230 = I(U;S,B)$$

となり，例と同様のことが言える．$m = 8004$ は交通 (138 種類) と物販 (58 種類) の組み合わせの数である．

4. 乗降履歴データのユースケースと評価指標

4.1 ユースケースと有用性評価指標

本データのユースケースを表 15 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 [4] 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．

想定したユースケースに対応する評価指標を検討する．[2] 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 (A_1)
(2) 駅利用回数の順位 (上位のみ)(A_2)
(3) 顧客属性 (性別，学年) のクロス集計の人数 (A_3)

また，これらの評価を行う有用性指標を順に A_1, A_2, A_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．$T_s(X_i)$ は T につい

表 12　ユーザごとの物販料金の例 E_B

ユーザ \ 物販料金	b_1	b_2	計	$P(U=u_i)$
u_1	2	0	2	2/7
u_2	1	3	4	4/7
u_3	0	1	1	1/7
$H(U\|B=b_i)$	0.92	0.31		
$P(B=b_i)$	3/7	4/7		

表 13　E_S と E_B から 1 履歴ずつ取得した場合の集計表 $E_{S,B}$

	s_1,b_1	s_1,b_2	s_2,b_1	s_2,b_2	s_3,b_1	s_3,b_2	計	$P(U=u_i)$
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
$H(U\|S=s_i,B=b_j)$	1	0.81	0	0	0	0		
$P(S=s_i,B=b_j)$	8/46	16/46	2/48	4/46	4/46	12/46		

表 14　E_S,E_B,$E_{S,B}$ の各値

$\backslash x$	s	b	s,b
$H(U)$	1.47	0.98	1.19
$H(U\|x)$	0.99	0.57	0.46
$I(U;x)$	0.48	0.41	0.73
$P(U\|x)$	0.50	0.67	0.73

表 15　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

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 駅の集合とする．$rank(T,s)$ 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，m_{sex} 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．

$$A_1(M,T,M^*,T^*) = \frac{\sum_{i=1}^{g} |T_s(X_i) - T_s^*(X_i)|}{g}$$

$$A_2(M,T,M^*,T^*)$$

$$= N - |s \in S_N(rank(T,s) = rank(T^*,s))|$$

$$A_3(M,T,M^*,T^*)$$

$$= \frac{\sum_{i=1}^{m_{sex}} \sum_{i=1}^{m_{grade}} |Cross_{sex,grade}(i,j) - Cross^*_{sex,grade}(i,j)|}{m_{sex} m_{grade}}$$

4.2　多様な評価指標

本節では，4.1 節で述べた例をもとに，実際に匿名加工を検討するために，いくつかの評価指標を定義した．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 や R 言語で行った．

安全性指標の概要を表 16 に示す．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 S_2, 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 S_3，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 S_5 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．

同様に，有用性指標の概要を表 16 に示す．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 U_1，駅の登場回数の平均絶対誤差を用いる指標 U_6，用途ごとの合計金額の平均絶対誤差を用いる指標 U_{10} がある．しかし，用意した有用性指標の多くが「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しまった．

5.　乗降履歴データの匿名加工

顧客データや乗降履歴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
(2) 特殊な駅 (利用回数が少ない，極端に離れた場所にある) を利用している

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．本節では，有用性を保ちつつ，これらの特殊な値を持たないデータを作成する手順を，簡易データを用いて説明する．表 17 に簡易顧客データ M_1，表 18 に簡易乗降履歴データ T_1 を示す．この場合，3 つのグループ (1:性別=男，学年=1)，(2:性別=男，学年=2)，(3:性別=女，学年=4) ができる．

まず，4.1 節で定義した有用性指標を損なわない加工手法を考える．A_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，A_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，A_2 を損なうこともなく，顧客データは加工しないため A_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 18 の乗降履歴データ T_1 の利用駅をグループ内シャッフルした結果 T_1^* を表 19 に示す．

次に個人が特定されない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 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．表 20 に加工した顧客データ M_1^* 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 2 と同じ顧客属性 (性別=男，学年=2) に加工する．

特殊な駅を利用している顧客も特定されやすい．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 A_1 と A_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，A_2 は損なわない．表 21 に加工された乗降履歴データ T_1^{**} 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．

以上の様にして，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_1^*,T_1^{**} を作成した．表 22 に T_1,T_1^*,T_1^{**},M_1,M_1^* についての A_1, A_2, A_3 の値の変化を示す．M_1 を加工したことによって A_1, A_3 が上がってしまったが，A_2 の値は変

表 16　指標スクリプト

指標	種類	概要
U_1	有用性	日付ごとの平均料金の差の平均絶対誤差
U_6	有用性	駅の登場回数の平均絶対誤差
U_{10}	有用性	用途ごとの合計金額の平均絶対誤差
S_2	安全性	顧客ごとの料金合計値が近いものを再識別
S_3	安全性	顧客ごとの [交通, 物販, チャージ, バスチャージ, 共通] の使用比率から再識別
S_5	安全性	顧客ごとの各交通レコードにおけるセルの一致率から再識別

表 17　簡易顧客データ M_1

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

表 18　簡易乗降履歴データ T_1

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

表 19　加工された乗降履歴データ T_1^*

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

表 20　加工された顧客データ M_1^*

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

化しておらず，有用性を保てている．

6. プチ PWSCUP

6.1 プチ PWSCUP

第 5 節では，乗降履歴データを匿名加工した一例を紹介した．本節では，実際に交通 IC カードの履歴を加工し，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. 本システムを使って,

表 21　加工された乗降履歴データ T_1^{**}

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

表 22　評価値の変化

	M_1, T_1	M_1, T_1^*	M_1^*, T_1^*	M_1^*, T_1^{**}
A_1	0	0	2.67	2.67
A_2	0	0	0	0
A_3	0	0	0.2	0.2

プチ PWSCUP として, 交通 IC カードの履歴に対する匿名加工手法の検討を行う. なお，評価指標には 4.2 節で定義したものを用いた.

6.2 評価プラットフォーム

匿名加工した交通 IC カードの履歴データを評価するために，評価プラットフォームを実装した．評価プラットフォームのシステム構成を図 10 に示す.

参加者は, 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, 各参加者が提出し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 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 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.

6.3 結果

プチ PWSCUP は 2016 年 8 月に開催し，明治大学菊池研究室の参加者 4 人から計 47 個のデータが提出された．図 13 に提出された匿名加工データの分布を示す．縦軸が

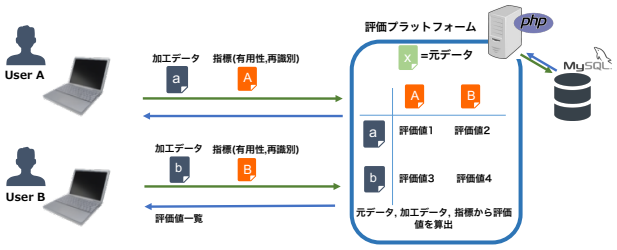

図 10　システム構成

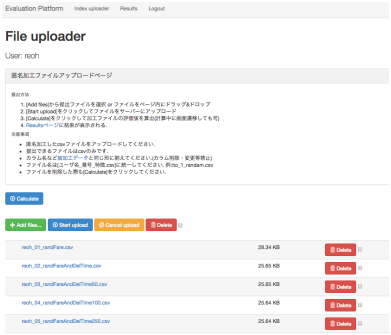

図 11　アップロード画面

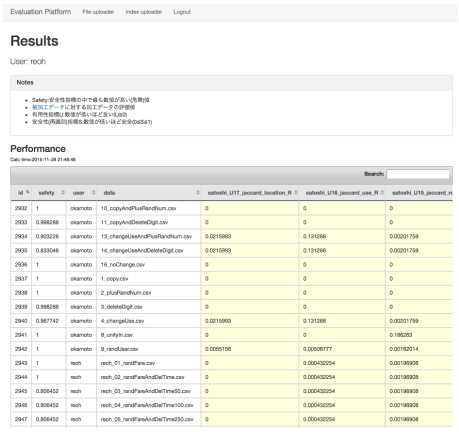

図 12　評価結果画面

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 位のデータは山岡匿名化 [1] 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．

7. おわりに

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

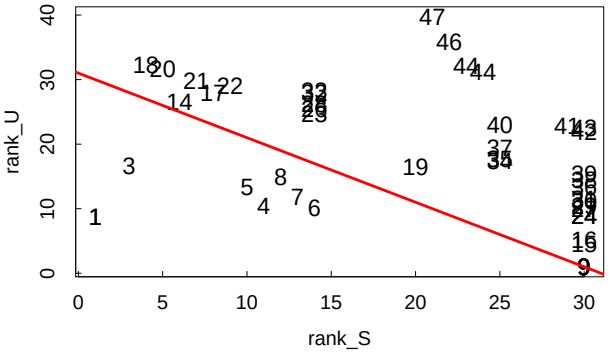

図 13　プチ PWSCUP の結果

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が判明した．また，「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．本稿では説明のため，作成した履歴データではなく簡易データに対して加工を行ったが，実際のデータにもこれらの手法は適用できる．本研究は非常に小規模なデータによるものであり，想定したユースケースも現実性に欠ける．

本稿では安全性をデータが特殊な値 (顧客，利用駅) を持つ場合を想定し，それらを持つ場合は危険なデータ，持たない場合は安全なデータと定義している．しかし本来，データの安全性評価は様々な再識別手法を想定する必要があるため，非常に困難である．データの評価指標についての議論は PWSCUP[1][2] を通して行われている．

より大規模なデータの取得・分析や，それを用いたより具体的なユースケースの想定とそれに対応する評価指標や加工手法の提案・実装，識別リスク分析を今後の課題とする．

謝辞　データを提供していただいた明治大学菊池研究室の皆さま，プチ PWSCUP に協力していただいた菊池研究室匿名加工班の皆さまに感謝いたします．

参考文献

[1] 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.
[2] 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.
[3] IC カードリーダー by マネーフォワード (https://play.google.com/store/apps/details?id=com.moneyforward.nfcreader&hl=ja)
[4] 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.)