乗降履歴データの匿名化に関する理論モデルと実データとの比較

崔 誠云† 疋田 敏朗† 山口 利恵†

† 東京大学大学院 情報理工学系研究科
113-8654 東京都文京区本郷 7-3-1
{song,hikita,yamaguchi}@yamagula.ic.i.u-tokyo.ac.jp

あらまし ビッグデータ活用において，ユーザーの移動履歴は都市開発において貴重なデータとなる．しかし，移動履歴データは個人情報に密接に関係するので，データ活用ではプライバシー保護のための加工する必要がある．既存研究では電車の乗降客数から乗降履歴を数学的にモデリングし，評価を行った．しかしこの手法では限られたデータに基づいてモデルを立て，乗降駅間の相関関係を考慮に入れず評価を行ったため，実環境のデータと乖離している．本稿では，電車の乗降履歴の実データを用いて匿名化を行い，理論結果との比較を行う．その後，理論研究モデルを設定するときの問題点を指摘する．そして，より正確なモデルを提案するために必要要件を述べる．

A Comparison of Anonymizing between Real Log of Transit Ridership and Theoretical Model

Seongun CHOI† Toshiro HIKITA† Rie Shigetomi YAMAGUCHI†

†Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo, Tokyo 113-8654
{song,hikita,yamaguchi}@yamagula.ic.i.u-tokyo.ac.jp

Abstract Big data of trajectory history becomes valuable in urban development. However, since the trajectory history is closely related to personal information, it should be processed for privacy protection. There was a study evaluating the process by generating theoretical model of train passenger log. However, since the model was made based on limited data and without considering correlation between stations, it was deviated to real data. In this paper, we compare anonymizing process between real log of transit ridership and theoretical model. Also, we point out problems of theoretical model and describe the requirement to propose an accurate model.

1 はじめに

近年，多様なビッグデータを活用しようとする動きが増えている．その中でも，人々の移動履歴データは，分析することによって日々どのような動きをするのかが分かり．混雑を解消するための交通サービスを提供するなど都市開発において役に立つ貴重なデータとなる．しかし，移動履歴データは個人情報に密接に関係するので，データを活用するためには，プライバシー保護ができるようデータ加工を行う必要がある．既存研究 [1] では電車の乗降客数から乗降履歴を数学的にモデリングし，データをどこまで公開しても安全で有効に活用できるかを評価した．しかし，この手法では:

- 限られたデータに基づいてモデルを立て，
- 乗降駅の選択を独立だと仮定した

うえ評価を行ったため，実環境のデータと乖離するという問題がある．人々から実データを収集するのは，ユーザーの同意や費用問題が生じるので，正しいモデルの設計は，それらの問題が回避できるというメリットがある．本稿では，菊池ら [1] の既存研究を紹介し，菊池らが用いたモデルについて説明する．そして，そのモデルの問題点を 2 つ述べ，「人の流れプロジェクト」[4] の平成 20 年東京圏のデータセットから電車の乗降履歴の実データを得て匿名化を行い，菊池モデルとの比較を行うことで問題点を示す．さらに，新たに log を使ったモデルを提案し，同じく実データを用いて比較を行う．その後，理論モデルを設計するとき考慮すべき点を指摘する．そして，より正確なモデルを提案するために必要要件を述べる．

本稿は, 下記の通りの手順で説明する．まず，2 章でプライバシーを保護するためデータ加工の必要性を述べ，データ加工手法を述べる．3 章で乗降履歴の既存研究について述べ，4 章で「人の流れプロジェクト」の東京圏での電車の乗降履歴データから菊池モデルを評価する．5 章で log を使ったモデル提案し，同様に電車利用履歴データから評価を行い，6 章でモデルを立てる際に考慮すべき要素について述べ，7 章でまとめる．

2　データ加工の必要性

この章では，プライバシーについて述べ，プライバシー保護のためデータ加工をする必要があることを述べる．その後，データ加工の有名な手法の一つである k 匿名化 [3] を紹介する．

2.1　プライバシー，匿名化

近年，様々なビッグデータをベースにそれらを解析し，その結果を第 3 者に提供するなど新たなビジネスが生まれている．プライバシーに関する意識が高まったいるなか，このようなビッグデータは敏感な個人情報が含んでおり，これを解析して得られた情報を第 3 者に提供するときには，個人が特定されるなどプライバシー問題を起こさないようにしなければならない．そのため，情報提供の際に，個人が特定されないよう，匿名化処理などのデータ加工を行い，安全に情報を提供する必要がある．

2.2　データ加工

データ加工の有名な手法として，k 匿名化という手法がある．k 匿名化とは，同じ属性と持ったユーザーが少なくとも k 人存在するようにデータを加工して，ユーザーが特定できなくする手法である．k 匿名化を満たすようにデータを加工するとき，k を満たさないデータを削除する方法や，それらのデータを曖昧化させる変形を行うなど多様な方法がある．しかし，データ自体に過剰な変形を加える方法を使うと，データとしての特性を失ってしまう．同様に，データを余りにもたくさん削除してしまうと，それもデータとしての特性を失ってしまう．データの特性を失ってしまうということは，データの価値を失うことと同じであるので，データの有効な活用において，これらを念頭し，元データを最大限維持できるような匿名化手法を考えないといけない．

3　乗降履歴の既存理論モデル

菊池ら [1] は電車の乗降履歴を数学的にモデリングし，匿名化に伴う有用性の損失を，削減されたレコードの割合で定めて，その評価方法を提案した．この章では，菊池らが提案した菊池モデルについて述べる．

3.1　既存研究の概要

菊池らは電車の乗降履歴を個人が特定されるリスクなしに何箇所の駅情報を公開しても安全かついて理論モデルを立て示した．菊池らの研究では，駅の利用頻度がジップ則に従うことに着眼し，ジップ則に従う菊池モデルを提案した．そして，乗降駅が独立に選ばれると仮定した上，理論実験を 2012 年の東京地域の JR 駅利用客数データ [2] から行った．その結果，利用履歴の

駅を 3 個以上公開するためには，$k = 2$ 匿名化を実現するためほぼ全データを削除すべきであることを示した．

3.2 理論モデル

菊池らは乗車数がジップ則に従うことに着眼し，駅の利用頻度がジップ則に従う菊池モデルを提案した．ジップ則は，様々な自然現象によく見られるもので，ジップ則によると，乗降履歴の場合，利用順位 x と利用客数 $f(x)$ はパラメータ a と c を使って次のような関係が成り立つ．

$$f(x) = \frac{a}{x^c} \tag{1}$$

ランクを各駅の利用客数の順位と定義する．高いランクは利用者数が多く，低いランクは利用者数が低い．ジップ則により得られたランクを 1 から総駅数の m までの利用客数で割ると，客がランク x の駅に存在する離散確率分布 $p(x)$ が得られる．

$$p(x) = \frac{f(x)}{\sum_{i=1}^{m} f(i)}$$

例えば，ランク 1 の駅がある人の乗車駅として選ばれる確率は $p(1)$ となる．データ加工前，収集されたデータのセット数 n が与えられたら，$n \cdot p(x)$ はランク x の駅の利用客数に相当する．各ユーザーは乗降履歴として多数の駅，レコードを持っているとする．ユーザーは公開するレコード数 s が多くなるほど特定しやすくなる．乗降履歴の公開レコード数 s が 1 のときはデータ加工が簡単で，k が 2 の匿名化を満たすように，$n \cdot p(x) < 2$ となるランク x のデータを削除すればよい．削除処理を行い残ったデータは，どの駅であっても少なくとも 2 人以上のユーザーが存在するようになり，$k = 2$ 匿名化を満たすことになる．ここで，そのデータが有効かどうかを判断するためには，そのデータがどの程度削除されたのかを知ることが重要になる．その理由で，匿名化率 ap を定義し，削除するデータの割合とする．$n \cdot p(x) < 2$ となる x の最小値を x^* とすると，次のように匿名化率が求まる．

$$ap = \sum_{i=x^*}^{m} p(i) \tag{2}$$

匿名化率 ap が高いほどデータとしての価値を失ってしまい，匿名化率 ap が 1 となると，全データを削除したことになる．連続した乗車駅の選択が独立事象であると仮定すると，s が 2 以上の場合に対しても匿名化率 ap の計算ができる．例えば，公開するレコード数 s が 2 のとき，収集されたデータのセット数 n と順位 x の確率 $p(x)$ と順位 y の確率 $p(y)$ の掛け算をランク 1 から総駅数 m まで行う，その結果をソートで順位を付け，その値が k/n 以下となる順位 z^* から先のデータを削除する．この計算は計算コストが高いため，菊池らは総和を積分で計算した．離散値である各駅の匿名化率 ap の総和を積分で計算すると，正しい計算結果と背離する．本稿ではより正確な比較を行うため，計算コストは考えないで正確に計算を行う．

3.3 厳密な計算結果と比較

表 1 は JR 東日本の各駅の 1 日平均乗車人員 [2] をランク 20 まで表したものである．菊池らはケーススタディで JR 東日本の乗車数のデータの中，ランク 100 までの乗車数を最小二乗法で当てはめて，パラメータ $a = 794132$，$c = 0.580$ のジップ則に従う菊池モデルを生成した．関東圏の駅数は $m = 2497$，関東地方の人口 $n = 42598300$ を用いて計算を行った．本稿では，4 章で $s = 2$ の結果を評価するので，$s = 2$ に対して計算比較を行う．厳密に計算した結果，$s = 2$ のときの匿名化率 ap は，菊池らの結果である 0.06556 より少ない 0.04514 となった．菊池らは，計算コストのため，総和を積分で計算した．ジップ則のグラフは単調減少するので，次のような不等式が常に成り立つ．

$$\sum_{i=1}^{m} f(i) \geq \int_{1}^{m} f(i)di \tag{3}$$

これにより，$N = \sum_{i=1}^{m} f(i)$ が実際より小さくなって，$p(x)$ は大きくなり，実際消さざるを得ない

表 1: (JR) ランク 20 までの利用客数

rank	駅名	1 日平均乗車人員
1	新宿	742,833
2	池袋	550,756
3	渋谷	412,009
4	東京	402,277
5	横浜	400,655
6	品川	329,679
7	新橋	250,682
8	大宮	240,143
9	秋葉原	234,187
10	高田馬場	201,765
11	北千住	198,624
12	川崎	188,193
13	上野	183,611
14	有楽町	164,929
15	立川	157,468
16	浜松町	153,104
17	田町	145,724
18	吉祥寺	138,483
19	大崎	138,311
20	蒲田	135,668

駅が，匿名化を満たしていることになってしまうケースが起こる．また，匿名化率 ap も上の同じ計算をするので，実際より小さく計算される．積分で計算を行うとこのように実際の結果と乖離してしまうので，この計算を総和で計算することで正確にした．本稿では正確な計算を行い，4 章で菊池モデルを評価する．

3.4 理論モデルの問題点

菊池モデルには 2 つの問題点がある．まず一つは，ランク 100 までのデータに基づいて生成したジップ則モデルにある．上位ランクに基づいて生成された菊池モデルは下位ランク，つまりテールを正しく反映しない．テールは一般には重要に扱われないが，k 匿名化は，テールが k 匿名化を満たすかどうかによって，データ削除を決めるため，ここのデータがきちんと反映されないと大きな問題が生じる．もう一つは，連続した乗車駅の選択が独立事象であると仮定した上，計算を行ったことである．実際には，乗降駅には強い相関関係がある．テール問題や相関関係については，4 章で実データを用いて検証する．

4 実データを用いた菊池モデル評価

本実験では，東京大学空間情報科学研究 センター「人の流れプロジェクト」[4] より,「平成 20 年 東京都市圏 人の流れデータセット」を用いて菊池モデルの評価を行い，問題点を述べる．

4.1 乗降履歴データを用いた評価

「平成 20 年 東京都市圏 人の流れデータセット」は，東京都市圏交通計画協議会による第 5 回東京都市圏パーソントリップ調査 [5] で得られたデータに，東京大学空間情報科学研究センターが独自に処理を加え，研究者を対象に公開しているものである．データセットから東京地域の電車利用ユーザーの乗車駅，降車駅を 1 セットに抽出し，各々のセットをユーザーの実データとして用いる．今回用いるデータは，$s = 2$ の履歴が 404005 セット，そのデータで出現された駅数は 1504 個である．この実データから $k = 2$ 匿名化を行う．つまり，同じ乗車駅と降車駅を持ったユーザーが 1 人のみであるとき，そのユーザーの履歴を削除する．$k = 2$ 匿名化を行った結果，匿名化率は 0.25279 となった．これは，データセットのうち，個人が特定できるユーザーのデータが全体の約 25 %を示していて，プライバシーのためにこれらのデータを削除しなければならないことを意味する．比較のために，理論研究と同様にランク 100 までの駅の利用客数からジップ則に従うモデルを生成させる．実データと生成した菊池モデルを図 1 と 2 に示す．図 1 は菊池モデルを生成するときに使ったランク 100 までの範囲の図であり，図 2 は全範囲の図である．菊池モデルのパラメータは，$a = 15559$，$c = 0.476$ である．図 1 によると，菊池モデルが実データを正しく反映してい

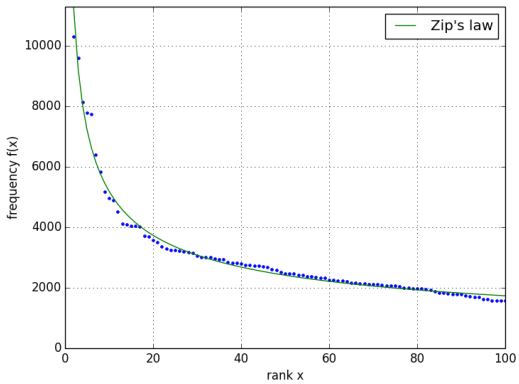

図 1: ランク 100 までの利用客数の実データと菊池モデル

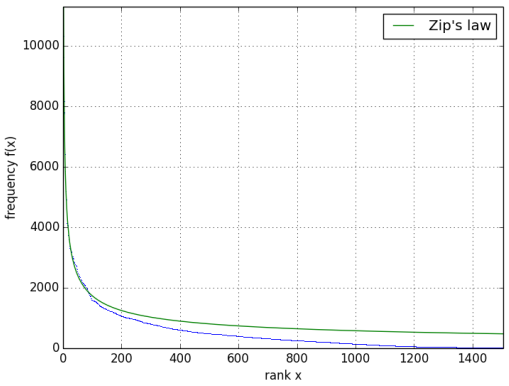

図 2: 全範囲での利用客数の実データと菊池モデル

るように見えるが，図 2 から見ると，テールが正しく反映されたいない．匿名化処理を行った結果，モデルによる匿名化率は 0.94173 となり，実際の結果と大きく外れた．

4.2 菊池モデルの問題点

このように実結果と外れたのは，3 章で述べた 2 つの原因が考えられる．まず 1 つに，ランク 100 までのデータから生成した菊池モデルは実世界を正しく反映しないことである．図 2 よりジップ則を適用した菊池モデルは，テールを正しく反映しないことを示した．さらに，ランク 1504 番目の駅までの実際のセット数は 404005 セットであることに対して，菊池モデルによるランク 1504 番目の駅までのセット数は，1351464 個になり，同様にテールが正しく反映されないことがこの違いからよくわかる．もう 1 つ，相関関係を考慮しなかったことについて，表 2 はランク 20 までの利用客を $k = 2$ 匿名化処理したときの匿名化率の比較である．利用客が少なくなることに連れて匿名化率は急激に増加するが，実際は乗降駅に相関関係があり，激しく変動しない．図 3 からも，菊池モデルは実データをと乖離していることがわかる．

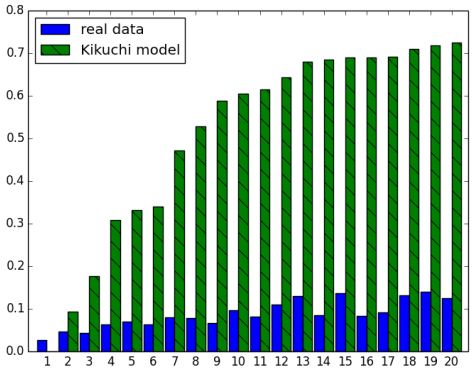

図 3: ランク 20 までの実データと菊池モデルの匿名化率比較

5 log モデル提案

4 章で，菊池モデルはテールが正しく反映されてないことと，相関関係を考慮しないことの問題点があった．本章では，log を使い，前者をより実データに近くしたモデルを提案する．

5.1 提案

この章の用語は 3 章で定義したものと同じものを使う．テールの問題を改善することで，匿名化率がどの程度現実に近くなるのかを確かめるため，次のように log を使ったモデルを提案

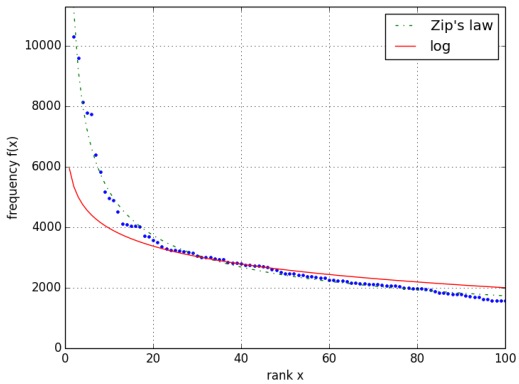

図 4: ランク 100 までの利用客数の実データと提案モデル

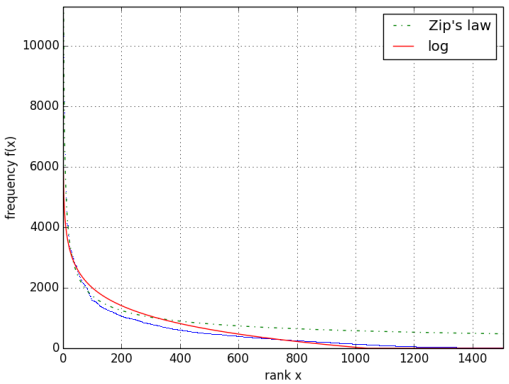

図 5: 全範囲での利用客数の実データと提案モデル

する．

$$f(x) = \begin{cases} a \log x + c & (f(x) > 0) \\ 0 & (otherwise) \end{cases}$$

提案モデルは，菊池モデルと同様に相関関係は考慮しないが，テールをより実データに近くしたモデルである．

5.2 乗降履歴データを用いた評価

提案モデルに対し，人の流れプロジェクトのデータを用いて，比較評価を行う．同じ条件での比較のため，菊池モデルと同様にランク 100 までの駅の利用者数からモデルを生成する．実データと生成したモデルが図 4 と 5 である．図 4 は提案モデルを生成するときに使ったランク 100 までの範囲での図であり，図 5 は全範囲での図である．図 4 によると，提案モデルは菊池モデルより性能が悪く見えるが，図 5 によると，菊池モデルよりテールのところで実データに近いモデルが得られたことがわかる．提案モデルのパラメータは，$a = -855.8$，$c = 5945.2$ である．提案モデルの匿名化率は，0.76293 となり，菊池モデルよりは改善されたが，実結果とはまだ大きく外れている．

5.3 提案モデルの問題点

提案モデルはテールは改善されたものの，高いランク，ヘッドを正しく反映しない問題が残っている．また，提案モデルは菊池モデルと同様に相関関係を考慮に入れてないので，実結果とは多く異なる結果になった．表 2 と図 6 からも，提案モデルは実データをと乖離していることがわかる．

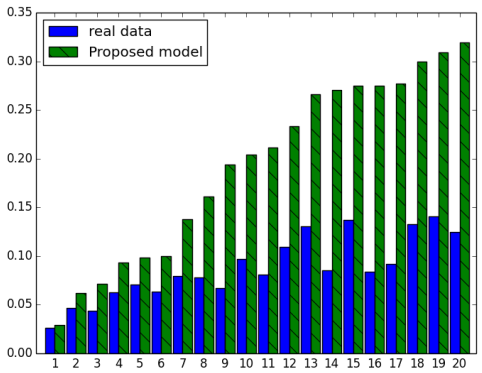

図 6: ランク 20 までの実データと提案モデルの匿名化率比較

6 考慮すべき要素

この章では，4 章と 5 章の比較結果をまとめ，モデルを設計するときの考慮要素を述べる．そ

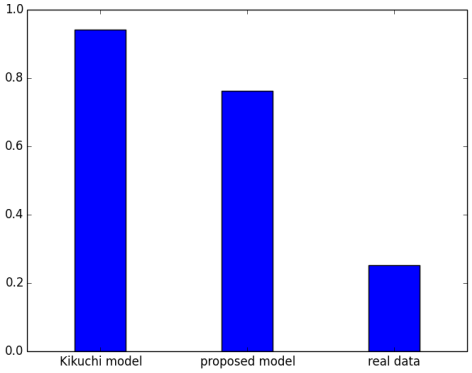

図 7: 各モデルと実データの匿名化率

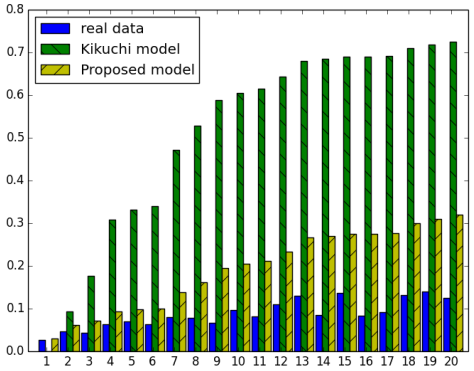

図 8: ランク 20 までの実データと各モデルの匿名化率比較

して，相関関係を考慮することの難しさを述べ，実データを用いることの重要性を述べる．

6.1 考慮要素

図 7 に各モデルと実データの匿名化率を示した．表 2 と図 8 に各モデルと実データのランク 20 までの匿名化率を示した．図 8 を見ると，菊池モデルも提案モデルも匿名化率が実結果より高くなっている．乗降駅に相関関係がないと仮定したら，ある駅の次の駅が図 5 の実データが描く曲線のような右下がり曲線に従う確率分布で駅が選ばれ，発駅の人数 n が少なくなると，$n \cdot p(x) < 2$ となる x のランク上がって，匿名化率が上がる．しかし，図 8 の実データの匿名化率を見ると，匿名化率が若干上がる傾向はあるが，菊池モデルが想定するほど上がっていない．その理由は，乗降駅に相関関係があるからである．その影響で，図 7 に見えるように，匿名化率が実結果より高くなっている．実世界を反映するためには，相関関係を考慮することが重要である．しかし，相関関係は都市の歴史や規模によって大幅に異なると予想される．実世界を反映するため，相関関係を求めることが課題となる．

6.2 実データ

菊池モデルと提案モデルは，実世界に近いモデルを生成したとしても，相関関係を考慮しない限り，限界があるため，実世界を正しく反映するモデルを生成することは難しい．しかし，実データを使うにも，調査方法によっては偏りが生じる可能性があることを念頭しなければならない．表 2 を表 1 と比較してみると，人の流れプロジェクトのデータセットにも偏りが生じていることがわかる．実データを使うなら，実データを集める際に偏りが生じないようにすることも重要であろう．

7 終わりに

本稿では，電車の乗降履歴の理論モデルの匿名化率の厳密な計算を行い，実データにより計算された匿名化率と比較を行い，理論モデルが実世界を正しく反映しないことを示した．そして，そういう結果となった理由を理論モデルの問題点を 2 つ挙げて示した．さらに，テールが正しく実世界を反映しない問題点を直した log を使ったモデルを提案し，菊池モデルより少し改善されたことを示した．しかし，提案モデルも相関関係を考慮しなかったため，実結果と大きく外れていることを示し，モデルを作るとき，相関関係を考慮することが重要であることを述べ，それを満たすモデルを作ることの難しさを述べた．

表 2: ランク 20 までの実データと各モデルの匿名化率比較

rank	駅名	客数	匿名化率		
			実データ	菊池モデル	提案モデル
1	新宿	7825	0.02633	0	0.02945
2	渋谷	5177	0.04655	0.09381	0.06166
3	横浜	4755	0.04374	0.17595	0.07175
4	池袋	4068	0.06293	0.30875	0.09316
5	東京	3945	0.07047	0.33221	0.09814
6	新橋	3906	0.06375	0.33986	0.10018
7	品川	3208	0.07918	0.47141	0.13759
8	川崎	2906	0.07777	0.52768	0.16120
9	大宮	2573	0.06724	0.58903	0.19418
10	田町	2485	0.09698	0.60485	0.20448
11	立川	2425	0.08124	0.61544	0.21182
12	浜松町	2271	0.10920	0.64354	0.23314
13	秋葉原	2067	0.13062	0.68043	0.26641
14	町田	2045	0.08509	0.68432	0.27030
15	飯田橋	2019	0.13720	0.68925	0.27523
16	戸塚	2019	0.08370	0.68925	0.27523
17	柏	2006	0.09222	0.69124	0.27722
18	有楽町	1896	0.13238	0.71072	0.29979
19	神田	1851	0.14046	0.71819	0.30940
20	上野	1809	0.12493	0.72583	0.31925

参考文献

[1] 菊池 浩明, 高橋 克巳,「乗降履歴データの安全な匿名化は可能か?」SCIS 2014

[2] JR 東日本, 各駅の乗車人員 (2012 年度), (http://www.jreast.co.jp/passenger/, 2013 年 10 月参照).

[3] L. Sweeney. k-anonymity: A model for protecting privacy. In International Journal on Uncertainty, Fuzziness and Knowledge Based Systems,10(5), pages 557-570, 2002

[4] 東京大学空間情報科学研究センター 人の流れプロジェクト, People Flow Project. (http://pflow.csis.u-tokyo.ac.jp/index-j.html)

[5] 東京都市圏交通計画協議会 パーソントリップ調査 (http://www.tokyo-pt.jp/person/index.html)