旅客乗降履歴の匿名化における情報削減量に関する検討

河村 悟[1,a)]　横山 大作[2,b)]　富田 美光[1,c)]　伊藤 正彦[2,d)]　豊田 正史[2,e)]　喜連川 優[3,2,f)]

概要：IC カード乗車券から得られる旅客乗降履歴は、需要予測、適切な案内等のサービス改善の観点で有用な情報である。その利用に当たっては、個人情報保護の観点からデータの匿名性が重要な課題となっている。首相官邸 IT 総合戦略本部に設けられた「パーソナルデータに関する検討会」においては、首都圏の乗降客統計に基づいて乗降履歴データの匿名化についてシミュレーションを実施し、短い履歴データでも特定の個人を識別可能であると指摘している。本稿では 1 日当たり 30 万件程度の無記名 PASMO の乗降履歴を解析して、匿名化レベルとそれに必要なデータ削減量の関係を検証した。

1. はじめに

地下鉄などの都市公共交通機関は住民の生活を支える重要インフラとしての役割を果たしている。しかしながら、2011 年 3 月に発生した東日本大震災時には、運転再開の際に渋谷駅などのターミナルが混乱し、さらに同じ年の秋に発生した台風ではやはり、運転再開時に同様の混乱が発生し、旅客の状況把握は困難を極めている。IC カード乗車券から得られる乗降履歴には、日常的な旅客流動、イベント、事故、災害等発生時の流動の変化などが分かる情報が含まれており、これを分析することにより、今後想定される首都直下地震発生時、2020 年の東京五輪等の大規模イベント時において、機動的な輸送サービス及び案内を提供するために有用な知見を抽出することが期待されている。これがひいては都市交通を重要インフラとして維持可能にすることにも通じる。

乗降履歴は、IC カード乗車券利用者の移動履歴を含んでおり、個人情報保護の観点から慎重な取り扱いが求められる。首相官邸高度情報通信ネットワーク社会推進戦略本部 (IT 総合戦略本部) に設けられた「パーソナルデータに関する検討会」[*1](以下「検討会」と言う) においては、首都圏の乗降客統計に基づく乗降履歴データの匿名化についてシミュレーションを実施し、氏名等の情報を除去し、ID を仮 ID に変換した短い履歴からでも特定の個人を再識別可能である (2-匿名性が担保されない) ことを指摘している [1]。

本研究では、首都圏で流通している IC カード PASMO の乗降履歴データを用い、少数回しか観測されない乗降駅の組み合わせを持つ履歴のレコードを削除する、という匿名化手法を用いた場合に必要なデータ削減量の関係を検証する。今回の実験では、2013 年 4 月中 1 週間分の東京メトロに関する無記名カードのデータを用い、k-匿名性が担保される乗降履歴の割合がどの程度あるかを検証し、検討会でのシミュレーションの結果と比較を行った。乗降客数の分布を元に算出した駅の利用確率を用い、独立に複数の駅を選択するという検討会のユーザモデルから推定される出現確率と、乗降履歴から集計された出現確率とは乖離しており、乗降履歴では利用される駅の組み合わせがより偏った分布になっていることが判明した。乗車駅と降車駅の間の遷移確率を検証した結果、遷移確率は 1 駅めの選択確率より偏った分布を持つことが確認され、駅の選択が独立であるという仮定が成り立っていないのではないかという結果が得られた。

さらに、各レコードに時間情報が付与されている事を考慮し、複数の利用シナリオを念頭に置いて記録する時間粒度を変化させた時の再識別可能性について検証を行った。キャパシティプランニングなどに利用できる、ピーク需要を考慮した旅客状況把握のためには 1 時間程度の時間粒度が必要であり、その状況下では 2 駅の情報を利用して 20%程度、5 駅以上の情報を利用すると 70%程度のレコードについて識別可能であるとの結果が得られた。また、災害・事故・イベント等の状況把握においては 15 分程度の

[1] 東京地下鉄株式会社 情報システム部
Tokyo Metro Co., Ltd.
[2] 東京大学 生産技術研究所
The University of Tokyo
[3] 国立情報学研究所
National Institute of Informatics
[a)] s.kawamura@tokyometro.jp
[b)] yokoyama@tkl.iis.u-tokyo.ac.jp
[c)] y.tomita@tokyometro.jp
[d)] imash@tkl.iis.u-tokyo.ac.jp
[e)] toyoda@tkl.iis.u-tokyo.ac.jp
[f)] kitsure@tkl.iis.u-tokyo.ac.jp
[*1] http://www.kantei.go.jp/jp/singi/it2/pd /index.html

時間粒度による解析が求められるが、この場合には 5 駅以上の情報を利用することで 90%弱のレコードが識別可能であることが判明した。

本稿の流れは以下のとおりである。まず 2 章で、IC カード PASMO の実情を説明する。次に 3 章で PASMO 乗降履歴データを使用した検証結果について説明する。4 章で関連研究を紹介する。最後に 5 章でまとめと今後の方向性について述べる。

2. IC カード PASMO

2.1 PASMO とは

PASMO[*2] は 2007 年 3 月に供用開始され、発行枚数は 2,400 万枚を超えている。さらに昨年 2013 年 3 月から全国の主要都市でも相互利用可能になり、8,000 万枚のカードがどこでも使用可能になった。

PASMO にはカードごとに ID 番号が付与され、自動改札機や自動券売機等で使用されると履歴データがカードに蓄積されるとともに、センターサーバーに定期的に情報がアップロードされている。この履歴に関するログ情報を一件明細データ (以下「利用ログ」と言う) と呼んでいる。

本稿では、この利用ログを用いて検証した。PASMO は記名カードと無記名カードの 2 種類からなるが、今回は無記名カードのみを用いた。東京メトロに保管されているデータを用い、東京メトロの管理の下で実験を行った。解析対象としては 2013 年 4 月の 1 週間分の利用ログを用い、カードの ID についてはランダムに変換し、もとに復元できないような状態で取り扱った。

2.2 PASMO ログからの乗降駅抽出

PASMO の利用ログには、乗車駅、乗車時刻、降車駅、降車時刻、利用状況 (改札機入出場、券売機での切符購入等)、利用金額、残額等の履歴情報が記録されている。今回解析の対象としたのは東京メトロの路線を一度でも利用したと考えられる乗降に関する履歴であり、直通運転などにより東京メトロ以外の駅を利用した履歴も含まれている。しかし、東京メトロ以外の区間に関しては輸送の一部を記録しているに過ぎず、需要予測等の解析対象とするには実態との乖離が大きいので、ここでは東京メトロ内の駅を利用した情報のみを解析対象とできるよう、以下のような前処理を行った。

東京メトロ以外の駅を利用した乗降に関しては、接続駅で打切り、他線から乗りついできた乗客を接続駅以遠の仮想駅を想定して集約することとした。例えば、メトロ内の s_1 から乗車してメトロ以外の駅 s_{other} で降車した履歴が存在するとき、s_1 から s_{other} までの経路探索を行い、メトロの路線から他社線へ接続する駅 s_m を求める。そして、s_m に対し s_m^{outer} という仮想的な駅を作成し、$[s_1, s_m^{outer}]$ という乗降の履歴が存在したとして解析を行う。経路探索においては、標準的な駅間所要時間、乗り換え所要時間、及び列車間隔時間をもとに、所要時間が最短となるような乗り換え経路を求めた。

[*2] http://www.pasmo.co.jp/

表 1 解析対象ログ

	1 日分	1 週間分
トリップ数	29.9 万	205 万
ユニーク旅客数	15.1 万	52.3 万
ユニーク駅数	184	184

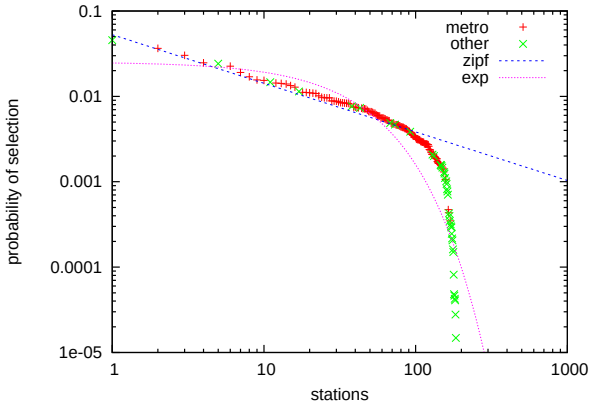

図 1 駅毎の乗降確率

2.3 解析対象ログ

2013 年 4 月 15 日 (月曜日)、及び 4 月 15 日から 21 日までの 1 週間の利用ログに関して、2.2 節の手法を用いて乗降駅の組 (以下、トリップとする) を抽出した。表 1 に抽出できた乗降履歴の情報を示す。

1 日分の乗降履歴について、駅毎の乗降客数を集計し、全乗降客数に対する割合を多い順に並べた結果を図 1 に示す。metro の系列は東京メトロ内の駅を、other はトリップ抽出過程で仮想的に生成された駅を示している。また、この分布が Zipf 則に従うと仮定して、最小 2 乗法で関数近似したものを Zipf の系列に示す。近似関数は

$$p(x) = 0.0528x^{-0.569}$$

となった。上位 120 駅程度までは Zipf 分布によく従っていると考えられる。なお、メトロの全駅数は 142 駅であり、おおむね 8 割以上の駅がこの分布で近似できていると考えられる。また、exponential な関数で近似した場合 $p(x) = 0.0253e^{-0.0277x}$ となる。結果を exp の系列に示しているが、Zipf の方が分布として近い。

3. 匿名化における情報削減量の解析

3.1 ユーザ毎のトリップ履歴解析

ユーザ毎に 1 日の間に何回利用があったのかを集計した結果を図 2 に示す。user の系列はユーザ数に占める割合

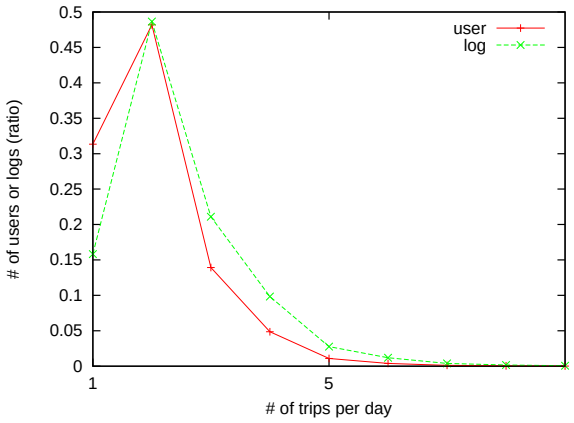

図 2　同一ユーザのトリップ数 (1 日)

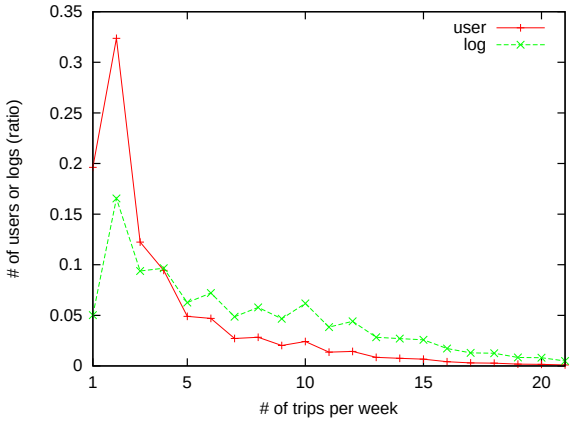

図 3　同一ユーザのトリップ数 (1 週間)

を、log の系列は乗降履歴内のレコード数に占める割合をそれぞれ示している。1 日の間には、2 回利用する人の割合が半数近くを占め、3 トリップ以上利用している人は急激に減少していることがわかる。また、1 週間分の集計を図 3 に示す。1 週間のログであっても、2 回しか鉄道を利用していない人の割合が高いことがわかる。ただし、1 日分と比較して 5 トリップ以上利用した割合も増えており、特にレコード数に対する割合が高くなっている。つまり、多数の駅を利用した履歴が残っているユーザに関するレコードが、無視できない割合で存在していることが想定され、長時間履歴を集めることで識別可能なレコードの量が増えていくことが予想される。

3.2　検討会モデルとの比較

検討会では、

- 各ユーザは Zipf 則に従って駅を選択する
- 各ユーザは複数の駅を独立に選択していく (前に利用した駅は次の駅の選択に影響を与えない)

というユーザモデルに従った検証を行っていた。本論文では、まずこの検討会モデルと乗降履歴との振る舞いの違いを検証する。

乗降履歴を利用した解析においては、旅客の移動経路を

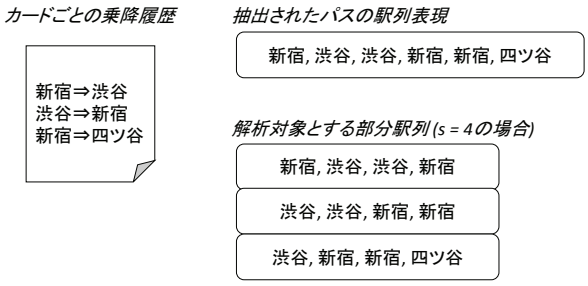

図 4　カードごとの乗降履歴から得られるパス

推定するなどの目的で、同一旅客の連続するトリップを取り出し、解析対象とすることが多いと考えられる。例として、図 4 のような乗降履歴があったとき、連続したトリップを取り出すことで新宿-渋谷-新宿-四ツ谷という移動経路 (以下「パス」と言う) をこの旅客が取っていたことがわかる。利用するパスが長くなるほど解析で得られる情報は増える一方、匿名性の担保は難しくなると考えられるため、このパス長と識別可能性との関係を明らかにすることが重要である。ここでは、パスの表現として、トリップから得られた利用駅の列を用いるものとする。図 4 の例で、乗降駅両方の情報を利用できると仮定した場合、このトリップは (新宿, 渋谷, 渋谷, 新宿, 新宿, 四ツ谷) という長さ 6 の駅列によって表現される。また、もし乗降履歴において降車駅の情報のみしか利用できない場合には、このトリップは (渋谷, 新宿, 四ツ谷) という長さ 3 の駅列で表現される。

本論文では、解析に利用するこの駅列の長さ (s とする) を変化させ、それぞれの s においてどの程度の乗降履歴に識別可能性があるのか、すなわち同一の利用パターンのない履歴がどの程度存在しているのかを求める。今、あるユーザが $[s_1, s_2, s_3, s_4]$ という 4 駅を利用したという履歴があったとする。ここで、$s = 2$ とした場合、このユーザは $[s_1, s_2], [s_2, s_3], [s_3, s_4]$ の 3 通りの駅列を持つと考え、それぞれについて同一の駅列を持つユーザ数を集計した。ただし、あるユーザの乗降履歴が s 以下の駅しか含んでいないとき、長さ s に到達するまでダミーの駅を利用したという履歴が後ろに付いているものと仮定して集計を行っている。例えば $[s'_1]$ という履歴を持つユーザについては、$[s'_1, dummy]$ という履歴であるとして集計を行った。

3.2.1　駅選択確率

検討会では、降車駅の情報のみを利用した場合を想定していたため、まず降車駅履歴のみに関する集計を行った。結果を図 5 に示す。$s = 1, 2, 3$ の系列がそれぞれの s で集計を行ったときの、パス毎の乗降履歴中の存在確率を示している。また、3.1 節で得られた Zipf 分布を Zipf(s = 1) の系列に示す。検討会モデルでは、$s = 2$ の時は Zipf 則に従った駅選択を独立に 2 回繰り返すという仮定を行っており、これをプロットしたのが Zipf(s = 2) である。同様に 3 回繰り返したものが Zipf(s = 3) となる。乗降履歴の集計

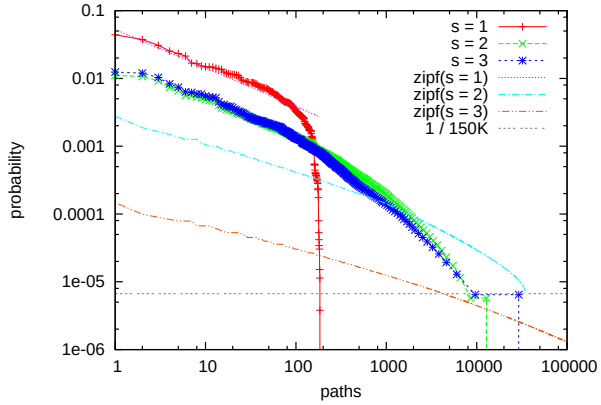

図 5　パス毎の存在確率

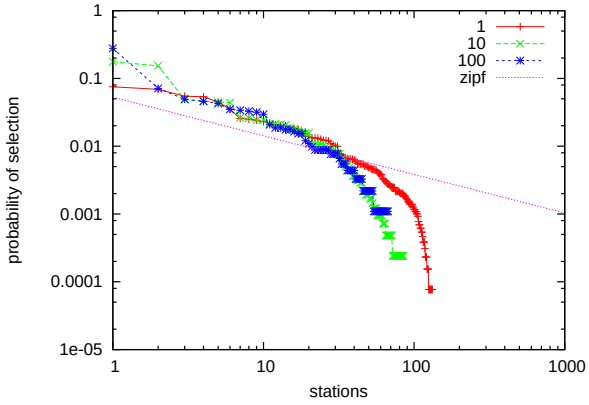

図 7　駅遷移確率 (1 日)

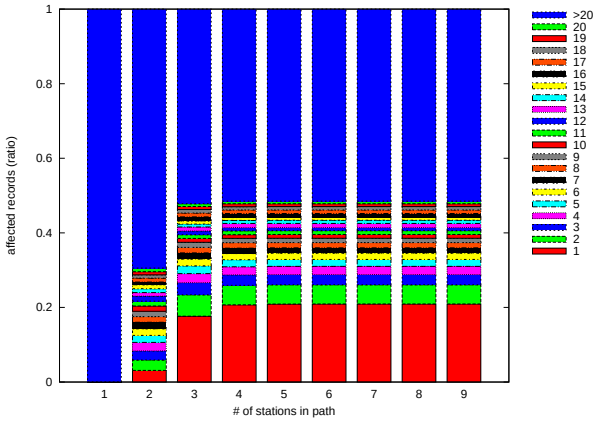

図 6　識別可能レベルごとのログ中の存在割合

結果と検討会モデルとを比較すると、$s = 1$ の場合はよく一致しているが、$s = 2$ より長い駅列長に関しては乗降履歴は利用の多い駅の組み合わせでは検討会モデルより高い確率を持ち、利用が少なくなるに従って急激に確率が低下するような分布を持つ。つまり、実際の乗降は少数の駅の組み合わせに偏って行われており、1 ユーザの履歴にしか現れないような識別可能なエントリが発生しにくい、ということがわかる。

なお、ユニークユーザ数をおよそ 15 万人と考え、識別可能エントリが発生するであろう確率、すなわち $1/150,000$ を図中の”1 / 150K”の系列に示している。検討会モデルでは $s = 2$ ではぎりぎり識別可能レコードが存在せず、$s = 3$ で識別可能になると予想されているが、乗降履歴では識別可能なレコードが $s = 2$ から存在している。

また、$s = 2$ から $s = 3$ に駅列長を延ばしても確率の分布があまり変化していないが、これはおそらく 3 駅以上利用するようなユーザの割合が少ないことに起因すると考えられる。図には示していないが、$s = 4$ 以上にした場合も分布はあまり変化しない。

3.2.2　識別可能レベルごとの履歴ログへの影響

前節では、パスごとにログ中の出現回数を集計した。ここで、出現回数が k 回 $(k = 1, 2, ...)$ のパスがログ中のどのトリップから現れたかを調べ、ログ内での全トリップ数に対する該当したトリップ数の割合を求めた。結果を図 6 に示す。横軸は $s = 1$ から 9 まで変化させたときの集計結果を表しており、それぞれの系列はパスの出現回数を示している。”>20”の系列は 21 回以上出現しているようなトリップの割合である。

$s = 1$ では全てのパスが 3 回以上出現しており、$s = 2$ で 1 回しか出現しないパスがログ中に占める割合は 3%程度である。また、s を増やしていった時、1 回しか出現しない割合はおよそ 20%程度で収束している。これは、2-匿名化を考えるとき、何らかの対処が必要となるログの割合が 20%程度であることを示している。また、20-匿名化を必要とするような場合には 50%程度のログが影響を受けることが見て取れる。

3.2.3　駅選択における遷移確率

検討会モデルと乗降履歴の解析結果はあまり一致していない。その原因には以下が考えられる。

- 駅の選択が「独立」であるという仮定が成り立っていない
- 2 駅以下の利用履歴しかないユーザが大半で、長い利用履歴を残すユーザが少ない
- 利用者の少ない駅は Zipf 分布から外れている

2 つめ以降の原因はログの集計期間を延ばしていったときに変化する可能性があるが、1 つめの原因はどのような場合でも影響が大きいと考えられる。あるユーザがある駅を利用した時、次に利用する駅は前の駅によってかなり制約を受けることが容易に想像でき、選択が独立であるという仮定は成り立たないことが予想される。これを検証するため、駅間の遷移確率を調査する。

ここでは、1 トリップにおける乗車駅と降車駅の遷移確率、つまりある駅から乗った旅客がどこで降りるかの確率分布を乗降履歴から求めた。乗車駅毎に降車駅分布を求めた結果を図 7 に示す。”1”、”10”、”100”の系列はそれぞれ、乗車旅客数が 1 番多い駅、10 番目、100 番目に多い駅の降車駅確率分布を表している。また、Zipf の系列は 3.1 節で

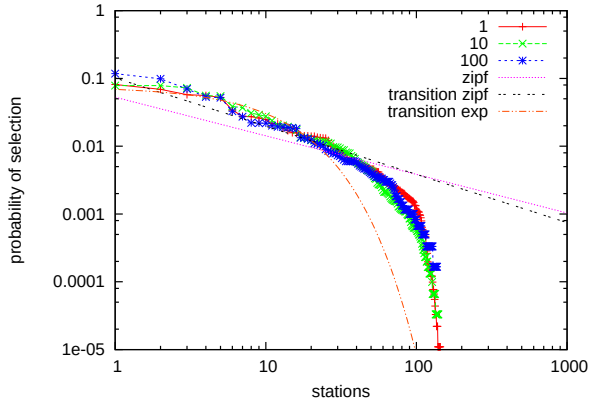

図 8 駅遷移確率 (1 週間)

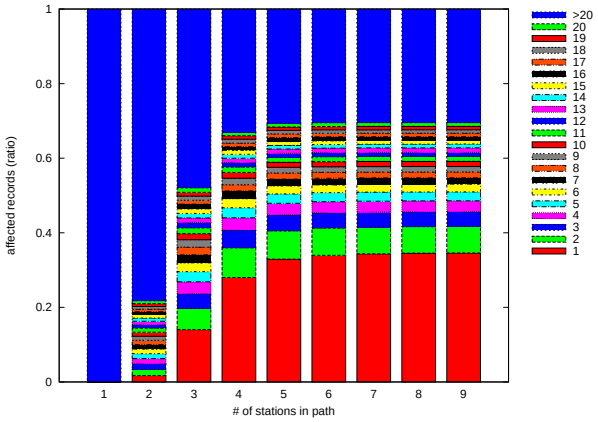

図 9 識別可能レベルごとのログ中の存在割合 (乗降駅双方利用)

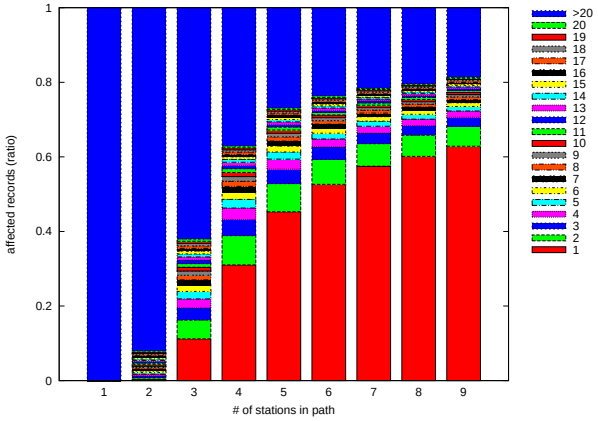

図 10 識別可能レベルごとのログ中の存在割合 (1 週間, 乗降駅双方利用)

得られたものである。どの乗車駅においても、降車駅分布が Zipf 則から外れ、より偏った分布を示していることが見て取れる。複数駅を選択するユーザモデルを構築するときには、このような遷移確率の偏りがあることを考慮する必要がある。

図 7 では、乗車客数が少なくなるにつれ Zipf からのずれが目立つ結果となっているが、これは該当するレコード数が少なくなっていくことに起因している可能性もある。1 週間の乗降履歴を利用して同じ集計を行った結果を図 8 に示す。1 週間の集計では、乗車客数の異なる駅での分布変化が小さくなっている。

また、“1” の系列に対して Zipf 及び exponential な分布による関数近似を行った結果をそれぞれ “transition zipf”, “transition exp” に示す。1 駅に関する乗降客数分布とは異なった分布を取っていると考えられるが、分布の形についてはより詳細に検討する必要がある。

3.3 乗降駅を両方利用する場合

検討会モデルでは降車駅のみを集計対象として議論を行っている。しかし、ログの解析利用においては、旅客フローの算出など乗車駅の情報も必要とする場面が多いと考えられるため、乗車駅の情報を削除しなかった場合について検討を行う。

ここでは、乗車駅と降車駅は区別せず、あるユーザが利用した駅の履歴として扱われると仮定して集計を行った。パスの出現回数ごとに乗降履歴中に占める割合を求めた結果を図 9 に示す。降車駅のみの場合 (図 6) と比較して、$s = 2, 3$ では識別可能なレコードが減っているが、それ以上の長いパスを利用するときには影響が大きくなり、35%程度のレコードが識別可能となることがわかる。

なお、乗客は交通機関を往復同経路で利用することが多いと考えられており、$s = 2$ で降車駅のみを解析した場合と、$s = 4$ で乗降駅両方を解析した場合は同様な傾向を示すのではないかとの予想もあったが、実際には乗降駅両方を解析した方が影響が大きく、往復同経路の利用にとどまらないランダム姓の高い利用が行われている事を示唆している。

3.4 集計期間を 1 週間に延ばした場合

図 2 に示したように、1 日の間に多数回鉄道を利用するような旅客は少なく、長いパスでの解析を行っても識別可能となるレコードの量が少なく保たれる 1 つの原因になっていると考えられる。そこで、より長期間の集計を行った場合を想定し、1 週間分の乗降履歴を利用して同様の解析を行った。結果を図 10 に示す。ここでは乗降駅の両方を利用している。

$s = 3$ よりパス長が短い領域での影響は 1 日分の解析と比較してより小さくなっており、乗降履歴の量が増えたことによる影響ではないかと考えられる。一方、それより長い領域では影響が大きくなっており、今回解析を行った最長パスの $s = 9$ の場合で 6 割以上が識別可能状態にあった。また、さらに長いパス長で解析を行うともう少し影響が増えそうな傾向が見て取れる。履歴の集計期間が長くなると匿名化が非常に困難である事がわかる。

3.5 時間情報を利用する場合

3.5.1 解析シナリオ

前節までは、乗降履歴の時間情報を利用せずに解析を行う場合の影響を検討してきた。災害時、イベント時に有用な輸送サービス、案内を行うためには利用者がいつ交通機関を利用したかの情報が不可欠である。そのため、乗降履歴内の駅利用時間の情報を残した状態で解析を行う場面も多いと想定される。

ここでは、駅情報に加えて時間情報を残した場合の識別可能性について検討を加える。前節まではパスの情報として利用駅のみを利用していたが、ここでは (利用駅, 利用時刻) のペアによってパスが構成されるとする。ただし、今回利用する PASMO の一件明細には降車時間のみが記録されているため、乗車駅については利用時刻は存在しないものとして扱った。例えば、s_1 から乗車、s_2 で降車、降車時刻 t_2 の履歴は、$[(s_1, None), (s_2, t_2)]$ というパスであると扱っている。なお、メトロ以外の駅で降車したトリップについては、s_2 がメトロ外の仮想的な駅となるため、最短経路探索で得られた予想乗車時間を利用して東京メトロ路線との接続駅を通過した時間を求め、仮想駅の降車時間とした。ただし、特にメトロ外の予想乗車時間は誤差が大きいと考えられるため、15 分単位で丸めた時刻を t_2 としている。

解析に必要な時刻情報の精度は目標ごとに異なることが予想される。必要以上の情報を残していると匿名性が損なわれる可能性が高くなるため、何らかの時間幅を設定して時間情報を丸めて乗降履歴を利用することが想定される。ここでは、いくつかの解析シナリオを想定し、時間の丸め幅と乗降履歴の識別性の関係を検証した。

3.5.2 1 時間単位での集計

駅施設の設計、避難計画策定などにおいては旅客の需要状況把握が必要であり、特に最大需要が発生した際のキャパシティを正確に理解する必要がある。地下鉄の場合にはラッシュ時と閑散時の需要の差が大きいため、実態を把握するには時間情報を利用することは必須と考えられる。

このような、ピークを考慮したキャパシティ解析においては、少なくとも 1 時間程度の時間スパンごとに旅客数等を集計することが求められる。このような利用想定のもと、乗降駅両方の情報を利用し、1 日分の乗降履歴について、1 時間単位で降車時間を丸めた結果を図 11 に示す。$s = 2$ の解析でも 20%程度、長いパス長では 70%以上のログについて識別可能であるとの結果が出た。匿名化を行う際、1 時間単位の時間丸めを施したとしても、相当量のレコードに影響が出ることが判明した。

3.5.3 15 分単位での集計

過去の災害時や事故時における旅客状況を理解して情報伝達計画、避難計画等を検討するために、その障害に巻き込まれた旅客、すなわちある時点にある地域で鉄道を利用している旅客の数を求めたい、という要求がある。この場合、少なくとも 15 分程度の時間幅で旅客数を求める必要があると考える。

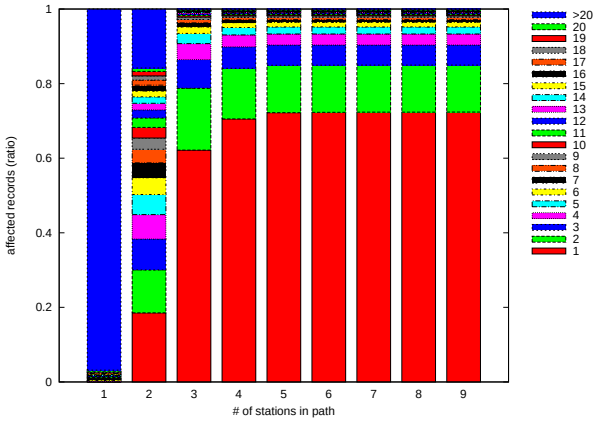

図 11 識別可能レベルごとのログ中の存在割合 (時刻情報 60 分単位)

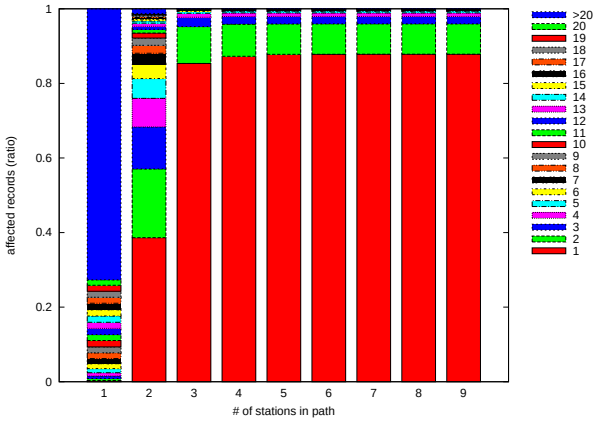

図 12 識別可能レベルごとのログ中の存在割合 (時刻情報 15 分単位)

また、スポーツ、花火、エンタテインメントなどの大規模イベントの際にも、旅客の円滑な誘導や人員配置に関する計画、効果的な情報伝達などを計画するために同様の旅客数把握を行うことが必要である。この場合においても、少なくとも 15 分程度の時間幅が求められる。

このような利用想定のもと、乗降駅両方の情報を利用し、1 日分の乗降履歴について、15 分単位で降車時間を丸めた結果を図 12 に示す。60 分単位の時間丸め操作よりさらに多くの乗降履歴が識別可能となり、90%弱のレコードに影響が出ることが見て取れる。

なお、東京メトロをはじめとして都市交通においては、ラッシュ時の最小運転時隔が 2 分以下の路線も多い。この点では 1 時間や 15 分という数字はやや大ざっぱであり、より精度の高い解析も可能であると考えられるが、災害時、イベント時等の応用を考えると、この程度の集約であっても十分有用な知見が得られる可能性はあると思われる。

4. 関連研究

著者等の知る限り、乗車券履歴と匿名化についての研究はこれまで存在していないが、応用可能性に着目した研究

は多く見られる。Pelletier ら [2] は IC カードシステムを 4 つのオブジェクトとして捉えたうえで、乗車券データの応用可能性について戦略的レベル、戦術的レベル、運用レベルの 3 つの場面を想定してそれぞれでの活用について触れている。彼らの視点から見ると、以下の研究は主に運用レベルの場面と言える。

筆者ら [3], [4] は、乗降履歴を用いることで旅客流動の時間変化を探索すべく 3 次元可視化を行った。これは路線図上に乗客の流れを表現したもので、時間とともに推移し色の濃淡で通常からの偏移を現している。無記名カードで ID 情報を用いずに履歴を可視化する試みである。大震災発災時、五輪イベント時等の乗客行動予測に有効であると考えている。定性的に可視化する試みにより匿名化の課題を克服し、活用の可能性を示した。一方、角田ら [5] は IC カードデータを活用する試みとして JR 東日本が発行している Suica カードの履歴から旅客の鉄道乗車所要時間を算出し、輸送障害時の利用者の影響を定量化する手法を提案している。

Sun ら [6] はシンガポール地下鉄 IC 乗車券カードのデータを用いて乗客の時空間分布と列車の軌跡を抽出する試みを行い、乗客の駅滞在時間と列車乗車時間分布を明らかにした。さらに列車の移動軌跡を描くことに成功した。しかしながらこれは単一路線のみの試みであり、ネットワーク全体に適用することを今後の課題としている。Liu ら [7] も中国シンセン地下鉄で同様な研究を行っている。また、Capla ら [8], [9], [10] はロンドンの地下鉄、バス等の公共交通で使用されている IC カード乗車券 (Oyster card) のデータとアンケート調査の結果を比較することにより、交通利用に関する意識と実際の利用状況の違いがあることを見出した。それによると 83 日間で 40 万ポンドの節約 (年間換算で 2 億ポンド) が可能であることを示した。この結果からは事業者には減収の懸念が生じるが、本来のあるべき運賃政策にフィードバックさせることにより公共交通促進につながることが期待される。しかしながら、この結果に対して乗客が如何に反応して行動するかは不明であるとしている。この点は匿名化することにより却ってその効果が把握しにくくなるという課題を指摘したい。さらに彼らは乗客の行動予測が可能であることに着目し、平日のラッシュ時において数時間後の混雑状況を予測し、旅客案内に応用することも提案しており、サービス改善の取組みとして注目される。

以上は都市鉄道における研究だが、公共交通の一端を担っているバスについてもやはり応用研究ではあるが行われている。Yuan ら [11] はバスのカード利用ログからカードホルダーの動きを抽出する手法を提案している。バスは鉄道のようにカード利用ログをリアルタイムに近い形で収集することが困難であり、その制約を乗り越える試みとして注目されるが、匿名化については残念ながら記述が見られない。

我々の研究はこれらの研究とは異なり、検討会での報告を受けて、匿名化によるデータ活用とデータ量削減の関係を検証した。これらは今後の様々な応用場面で、パーソナルデータを保護しつつ、鉄道業界での戦術、戦略立案を行うために効果的であると確信している。

5. おわりに

本研究では、IC カード PASMO の乗降履歴データを用いて匿名化レベルとデータ削減量の関係を検証した。検討会ではシミュレーションで検証したが、我々は乗降履歴を用いて検証した結果、検討会結果と一致する部分だけでなく差異があることも判明した。

今回は PASMO の履歴に関する匿名化レベルの検討を行った。今後は冒頭で述べたような震災発生時、五輪イベント等の乗客が集中する場面を想定して、どの程度データを削減しても、臨時列車運行、旅客案内と言ったサービス改善に支障しないかについて検証していきたい。この点に関しては、検討会との結果の違いをさらに詳しく分析する必要があると思われる。

本結果が今後のパーソナルデータの保護と有効活用の一助になれば幸いである。

参考文献

[1] パーソナルデータに関する検討会技術検討ワーキンググループ：技術検討ワーキンググループ報告書 2013/12/10, http://www.kantei.go.jp/jp/singi/it2/pd/dai5/siryou2-1.pdf (2013).

[2] Pelletier, M.-P., Trépanier, M. and Morency, C.: Smart Card Data Use in Public Transit: A Literature Review, *Transportation Research Part C: Emerging Technologies*, Vol. 19, No. 4, pp. 557 – 568 (2011).

[3] Itoh, M., Yokoyama, D., Toyoda, M., Tomita, Y., Kawamura, S. and Kitsuregawa, M.: Visualization of Passenger Flows on Metro, *Proc. VAST '13 (poster)* (2013).

[4] Yokoyama, D., Itoh, M., Toyoda, M., Tomita, Y., Kawamura, S. and Kitsuregawa, M.: A Framework for Large-Scale Train Trip Record Analysis and Its Application to Passengers' Flow Prediction after Train Accidents, *Proc. PAKDD '14*, pp. 533–544 (2014).

[5] 角田史記，加藤　学，大塚理恵子，助田浩子，大関一博：交通系 IC カードを利用した鉄道輸送障害時の影響を定量化する方法の研究，情報処理学会論文誌データベース (TOD)，Vol. 6, No. 3, pp. 187–196 (2013).

[6] Sun, L., Lee, D.-H., Erath, A. and Huang, X.: Using Smart Card Data to Extract Passenger's Spatio-temporal Density and Train's Trajectory of MRT System, *Proc. UrbComp'12*, pp. 142–148 (2012).

[7] Liu, L., Hou, A., Biderman, A., Ratti, C. and Chen, J.: Understanding Individual and Collective Mobility Patterns from Smart Card Records: A Case Study in Shenzhen, *Proc. ITSC '09*, pp. 1–6 (2009).

[8] Ceapa, I., Smith, C. and Capra, L.: Avoiding the Crowds: Understanding Tube Station Congestion Patterns from Trip Data, *Proc. UrbComp'12*, pp. 134–141 (2012).

[9] Lathia, N. and Capra, L.: How Smart is Your Smartcard?: Measuring Travel Behaviours, Perceptions, and Incentives, *Proc. UbiComp '11*, pp. 291–300 (2011).

[10] Lathia, N. and Capra, L.: Mining Mobility Data to Minimise Travellers' Spending on Public Transport, *Proc. KDD '11*, pp. 1181–1189 (2011).

[11] Yuan, N. J., Wang, Y., Zhang, F., Xie, X. and Sun, G.: Reconstructing Individual Mobility from Smart Card Transactions: A Space Alignment Approach, *Proc. ICDM '13*, pp. 877–886 (2013).