交通行動データ活用とプライバシー保護のトレードオフ：理論モデルによる解析

井料　隆雅[1]・原　祐輔[2]・日下部　貴彦[3]

[1]正会員　神戸大学　大学院工学研究科（〒657-8501 神戸市灘区六甲台町1-1）
E-mail:iryo@kobe-u.ac.jp

[2]正会員　東北大学　未来科学技術共同研究センター
[3]正会員　東京工業大学　大学院理工学研究科

パッシブ型交通行動データの活用は交通システムの効率的な計画運用に有用と思われるものの，常にプライバシー保護の問題が制約条件として存在し，これと効率とのトレードオフが適正にとれているかどうかは重要な問題となりうる．本稿ではこのトレードオフがどのようになるかを知るために，交通システム運用の簡単なモデルと個々人の行動履歴を前提とした利用者行動予測モデルを考え，さらにジップの法則による利用頻度の分布を仮定して，利用者行動の予測によりどれだけ効率的な交通システムの運用が可能になるかを定量的に調べた．結果，利用者行動予測を相当程度高精度にしない限り運用効率をあまり改善できない場合があることがわかった．その原因は低頻度利用者の行動予測の困難性にあるようである．

Key Words : *travel behaviour data, privacy issue, zipf's law, passive data collection*

1. はじめに

GPS搭載の携帯電話，スマートフォン，交通サービスにおける電子決済など，パッシブ型の交通行動データを取得するために活用できそうなサービスは年を追うごとに一般化してきている．パッシブ型の交通行動データは取得コストがアンケート等の能動的な調査に比べて相対的に安いだけでなく，その網羅の範囲が時間軸方向にも空間方向にも大きい．これにより，アンケート調査のような古典的手法では事実上得られないような量と質のデータが取得できることが期待できるし，実際に取得されつつある．このようなデータの適切な活用が，交通システムのより効率的な計画と運用を実現させる可能性を秘めていることは疑いないであろう．

一方で，そのようなデータの活用によって，交通システムの計画や運用が「どれだけ」効率的になるかについての包括的な議論は少ない．もちろん，より大量でより良質な交通行動データを用いれば，そうでないときよりもより効率的な計画や運用が実現することは疑いない．取得コストが安いことを考えれば，そのような優れたデータを無制限に活用することが好ましいように見える．無視できない問題は，交通行動データ活用には常に交通主体個々人のプライバシーの問題が存在することである．パッシブ型の交通行動データには個々の交通主体がいつどこにいた，かという一般的に考えてセンシティブな情報が含まれる．個人識別符号を氏名などの個人を特定できる情報と結びつけないという匿名化処理（個人IDの匿名化）はプライバシー保護のための基本的な手段だが，この処理だけでは悪意を持って個人情報を暴露しようという攻撃に脆弱であることもよく知られている[1)]．プライバシーを強固に保護するのであれば，たとえば複数の交通主体の行動を集計したり，追跡を分断したり（たとえば，異なる日や異なるトリップで同一交通主体に異なるIDを割り振る）することが有効であろう．しかしこのような処理はデータの質を劣化させ，その結果として交通システムの「より効率的な計画や運用」に悪影響を及ぼすことになるかもしれないし，あるいは特段の影響を与えず，プライバシー保護を優先しても差し支えないのかもしれない．以上の考察は，プライバシー保護と交通システム計画運用の効率化のあいだには一定のトレードオフが存在しうることを示唆する．

本稿では，このプライバシー保護と交通システム運用の効率化のトレードオフがどのようになるかを簡単な交通システムと行動履歴データの活用のモデルを構築することにより定量的に分析する．具体的には，

○　ある交通システムの将来の利用者数予測に基づいて交通サービスの容量を調整し，交通システムの運用による利益または便益を最大化する問題

を「交通システム運用問題」と定式化する（第2章）．この問題において，行動履歴データを用いた予測精度の向上が交通システムの運用利益ないし便益をどれだけ向上させるかを，行動履歴データを用いて将来の行動を予測することに関する既存研究の知見を活用しつつ数理的に分析する．それによって行動履歴データの活用が交通システム運用をどれだけ効率化するかを調べる（第3章）．その結果を用いて，交通行動データ活用とプライバシー保護のトレードオフと，さらに，プライバシーを保護しつつ交通システムの効率的な計画や運用を実現するためにはどのようなアプローチが好ましいかを考察する（第4章）．

2. 交通システム運用問題

(1) 交通システム運用問題の定式化

本節では，単一種類のサービスを提供する交通システムを指定された容量の制約で提供する交通システムをできるだけ効率的に運用する問題（以降，「交通システム運用問題」と呼ぶ）を定式化する．いま，交通システムの運用者は期 t において交通サービスの容量 μ_t を確定的に決定するとする．容量を期ごとにどこまで変動させられるかは考える交通システムの特徴によって異なるであろう．比較的自由に変動させられるものとしては，例えば信号交差点（期＝サイクル，容量＝青時間×飽和交通流率，と考える）や共同利用型の自動車（配車の調整などを手段とする）などがあろう．そのほか臨時便を設定できる一般の交通システムでも一定程度の容量調整は可能と考えられよう．

交通システムの利用者数は事前には厳密にわからず，ある確率分布に従う確率変数 X_t で表されるとする．利用者数が容量よりも多い場合は，容量を超える利用者はその交通システムを利用できない．交通システムが1単位（1名分）利用されることによる運用者の利益は1であり，1単位用意するごとに $0<c<1$ の費用をこうむるとする．運用者は自身の期待効用

$$E(Y_t)=E(\min\{\mu_t,X_t\})-c\mu_t \tag{1}$$

を最大化するように容量を決定する．この最適化問題は

$$\underset{\mu_t\geq 0}{\mathrm{Max.}}\ E(\min\{\mu_t,X_t\})-c\mu_t \tag{2}$$

と定式化できる．これが交通システム運用問題の数式による定式化である．なお，利益を運用者に帰着する利益でなく社会全体の便益の合計と定義しなおせば，式(2)の目的関数は社会的便益と解釈しなおせる．以降では簡単のため目的関数の値を単に「運用利益」と記述する．

(2) 交通システム運用問題の最適解

式(2)の最適化問題を解いて最適な容量 μ_t^* とその際の運用利益を計算する．このために目的関数にある期待値を計算する．利用者数と容量の双方が連続値をとると考えれば，式(2)は

$$\underset{\mu_t\geq 0}{\mathrm{Max.}}\left\{\int_0^{\mu_t} xp(x)\mathrm{d}x+\mu_t\int_{\mu_t}^{\infty}p(x)\mathrm{d}x\right\}-c\mu_t \tag{3}$$

と書き直せる．ここで $p(x)$ は確率変数 X_t の値xに対する確率密度関数である．式(3)の目的関数の導関数は

$$\int_{\mu_t}^{\infty}p(x)\mathrm{d}x-c=(1-P(\mu_t))-c \tag{4}$$

となる．ただし

$$P(\mu_t)=\int_0^{\mu_t}p(x)\mathrm{d}x \tag{5}$$

である．式(4)は μ_t に対して単調減少であるため，式(4)が0になる μ_t が式(3)の最適化問題の解となる．すなわち

$$\mu_t^*=P^{-1}(1-c) \tag{6}$$

となる．式(6)を式(3)の目的関数に代入すれば，最適解における運用利益 $y_t^*(c;P)$ を

$$\begin{aligned} &y_t^*(c;P)\\ &=\left\{\int_0^{\mu_t^*}xp(x)\mathrm{d}x+\mu_t^*(1-P(P^{-1}(1-c)))\right\}-c\mu_t^*\\ &=\int_0^{\mu_t^*}xp(x)\mathrm{d}x=\mu_t^*(1-c)-\int_0^{\mu_t^*}P(x)\mathrm{d}x \end{aligned} \tag{7}$$

と得る（式(7)の式展開では，計算の便に応じて μ_t^* を式(6)の右辺で置換するか否かをわけている）．式(7)の最右辺の値は**図-1**のように累積分布関数のグラフ上の面積として図示できる．

図-1からも直感的にわかるように，$y_t^*(c;P)$ は X_t の標準偏差に応じてその値が小さくなる傾向がある．これを定量的に評価するために，$p(x)$ として期待値 μ，標準偏差 σ の正規分布を考えよう．このときの $y_t^*(c;P)$ を $y_t^*(c;N(\mu,\sigma^2))$ と記せば

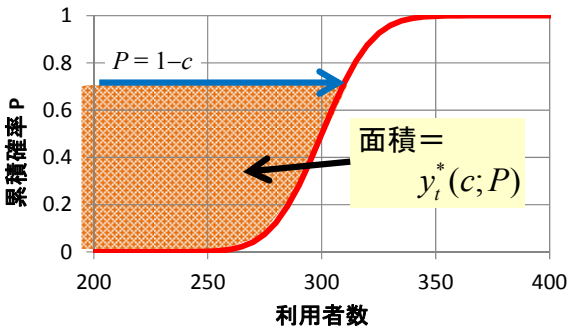

図-1 累積確率Pのグラフから運用利益 $y_t^*(c;P)$ を図解する方法（面積は横軸が0の部分まで含める）．

$$
\begin{aligned}
y_t^*(c;N(\mu,\sigma^2)) &= \int_{-\infty}^{\mu_t^*} \frac{x}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \mathrm{d}x \\
&= \mu \int_{-\infty}^{\mu_t^*} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \mathrm{d}x \\
&\quad + \int_{-\infty}^{\mu_t^*} \frac{x-\mu}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \mathrm{d}x \\
&= \mu(1-c) - \frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mu_t^*-\mu}{\sigma}\right)^2\right)
\end{aligned}
\tag{8}
$$

と計算できる．例えば$c=0.5$であれば$\mu_t^*=\mu$となるため式(8)は

$$
y_t^*(c;N(\mu,\sigma^2)) = 0.5\mu - 0.40\sigma \tag{9}
$$

となる．それ以外のcの場合，指数関数部分が1未満になるため第2項の係数は0.4よりも小さくなり，またσの変化により多少は変動するが，σに対して概ね線形に変化することには変わりない．

式(9)は，交通システムに対するX_tの標準偏差が小さければ小さいほど運用利益を大きくできることを示唆している．仮に，運用者が期tにおける利用者数を精度よく予測できれば，X_tの標準偏差は精度に応じて小さくなり，それに比例する分だけ運用利益を増加させることが可能である．次章ではX_tの標準偏差がとりうる値について分析する．

3. 行動履歴データを用いた利用者数の予測精度改善の可能性

本章ではX_tの標準偏差がどの程度の値になるかを利用者1人1人の行動を積み上げることによって計算する．この計算のためには各利用者がある期tにおいて交通システムを使用するか否かをどのように決定するかを記述するモデルを定式化しなくてはならない．この定式化を(1)節で行う．(2)節では(1)節の定式化に従って記述された各利用者の行動を積み上げたときにX_tの標準偏差が取る値を計算する．(3)節で，パッシブ型交通行動データから得られる過去の行動履歴を用いることによりX_tの標準偏差がどの程度改善できうるかを見積もる．

(1) 交通システム利用行動を記述する確率モデル

交通システムの利用行動は，当該の交通システムに対する人数nの潜在的利用者$i=1,\cdots,n$が，ある期tにおいて交通システムを利用する否かの二項選択行動として記述する．いずれを選択するかは確率で記述する．利用者iが期tに交通システムを利用する確率（利用確率）をp_t^iと書く．

本稿では「異なる利用者の行動には互いに相関がない」と仮定する．この仮定は，複数の利用者に共通して影響する説明変数が存在しないということに相当する．実際には曜日や天候など多くの利用者に同じ方向の影響を与える要因によって利用者間の行動に相関があるのが一般的であるためこのような独立性の仮定は強いものである．ただし，この仮定の非現実性は，たとえば，要因の有無により各期をカテゴリー分けし，カテゴリー内でのみこの仮定が成立する，などと考えることにより緩和が可能である．

p_t^iは利用者ごとに異なる値を取り得るため，その値の利用者内での分布形を事前に決めることができれば後の解析において便利である．このために，本稿では，ある一定の長期間（N個の期に相当する期間とする）で各利用者が交通システムを利用した回数はジップの法則に従うと仮定する．ジップの法則（Zipf's law, あるいはpower lawとも呼ばれる）は，都市のサイズ，文章中の単語の出現頻度など自然界のさまざまな現象が従うことが知られている法則である[2]．例えば単語の出現頻度であれば，ある英文テキストでの「単語の出現回数」を横軸にとり，各出現回数を記録した単語が何種類だったかを縦軸に取れば，その関係は両対数グラフ上で負の傾きの直線になることが知られている．これは，その英文テキストの中には「出現頻度が高い少数の単語」と「出現頻度が低い多数の単語」が混じっていることを意味する．同様に，交通システムの利用者についても，少数のヘビーユーザと多数の低頻度利用者が混在していると考えることは自然であろう．

ジップの法則はべき関数で記述される．本稿であつかう交通システムの利用回数に関する問題であれば，N個の期に相当する期間内における利用回数がξ回である利用者の全体に占める割合が

$$
f(\xi;s) = \frac{\xi^{-s}}{H_{N,s}} \tag{10}
$$

である，と記述できる．べき数sはこの分布の数学的性質（特に級数の収束性）に大きく影響するので，本稿では解析の便を考慮し特に$1<s<2$である場合のみを考える．なお，この仮定は実際に観測された各種の例[2]と大きくは乖離しない．$H_{N,s}$は一般化調和数で，

$$
H_{N,s} = \sum_{k=1}^{N} k^{-s} \tag{11}
$$

と定義される．この$H_{N,s}$は$s>1$のときにNを無限大にすれば一定の定数（具体的にはリーマン・ゼータ関数$\zeta(s)$）に収束する．Nが十分大きければN期にわたる

利用回数がξ回のときの利用確率をξ / Nと考えて差し支えないので，以降ではこれにより各利用者のp_t^iを定める．

交通システムの利用回数がジップの法則に従うという仮定を確かなものとするためには，実際の交通システムでの利用状況を確認することが望ましい．本稿では，このことを，匿名化されたETC統計データを集計して得られた阪神高速道路におけるETC通行車両の利用状況を例として実証的に確かめる．**図-2**は横軸に利用回数（軸上では日平均利用回数として表示している）を，縦軸に各利用回数を記録した車両数が総利用台数（期間中1回でも利用した車両の総数）にしめる割合をとり，その関係を両対数グラフとしてプロットしたものである．ここでいう「利用」は「阪神高速の任意のランプペアの利用」を意味している．集計期間は2009年7月から2010年6月までである．**図-2**はこの利用回数がジップの法則によく従うことを示している．グラフの左側の比較的低い利用頻度に相当する部分をべき関数で当てはめた結果を図中にあわせて示した（$s=1.482$）．

(2) 交通システム利用者数の標準偏差の計算

(1)節で定義した利用確率で各利用者が交通システムの利用の可否を決定する際の交通システム利用者数の期待値と分散を計算する．これらはそれぞれ

$$\mu = \sum_{i=1}^{n} p_t^i \tag{12}$$

$$\sigma^2 = \sum_{i=1}^{n} p_t^i (1 - p_t^i) \tag{13}$$

と書ける．これらより期待値については，

$$\mu = \sum_{\xi=1}^{N} \frac{\xi}{N} n f(\xi; s) = \frac{n H_{N,s-1}}{N H_{N,s}} \tag{14}$$

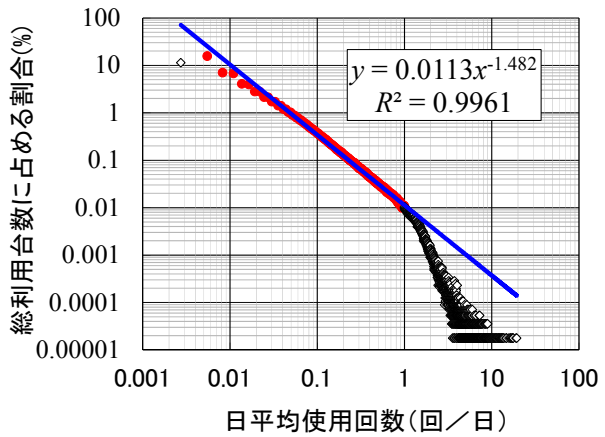

図-2 阪神高速道路における日平均利用回数の分布
回帰直線の計算には赤丸点（日平均使用回数が2番目に小さい値から1の間の点）のみ使用．

と計算できる．分散については，

$$\sum_{i=1}^{n} (p_t^i)^2 = \sum_{\xi=1}^{N} \left(\frac{\xi}{N}\right)^2 n f(\xi; s) = \frac{n H_{N,s-2}}{N^2 H_{N,s}} \tag{15}$$

と式(13)より

$$\sigma^2 = \mu \left(1 - \frac{H_{N,s-2}}{N H_{N,s-1}}\right) \tag{16}$$

と計算できる．分散が期待値に比べてどれだけ小さいかを知るには式(16)のカッコ内の値を知ればよい．これは，N が十分大きく，$r<1$ のときに

$$H_{N,r} = \int_1^N x^{-r} \mathrm{d}x \simeq \frac{N^{1-r}}{1-r} \tag{17}$$

と積分により近似的に計算して

$$1 - \frac{H_{N,s-2}}{N H_{N,s-1}} = 1 - \frac{\frac{N^{3-s}}{3-s}}{N \frac{N^{2-s}}{2-s}} = 1 - \frac{2-s}{3-s} = \frac{1}{3-s} \tag{18}$$

となるので，

$$\sigma^2 = \frac{\mu}{3-s} \tag{19}$$

とできる．$1<s<2$ であることを考えれば，式(19)の値は 0.5μ から μ の間になる．このことは，N が十分大きければ，σ^2 は μ の 0.5 倍より大きく 1 倍より小さいことを示す．

(3) q-予測モデルによる分散の改善の見積もり

σ^2 をより小さくするためには，予測対象の将来の期tにおける各利用者の行動を高い精度で予測し，それによって各利用者の p_t^i を0（使わないと予測）あるいは1（使うと予測）にできるだけ近づけることが求められる．利用者の行動履歴を活用してこのことを実現することを考えよう．「過去の行動履歴から将来の行動を予測する」ことがどこまで可能か，ということについては，Song et al. による，携帯電話の通話履歴より得られた3ヶ月にわたる行動データ（通話時の位置データと紐付けられているので，実質的には間欠的な位置追跡データとなっている）の履歴を用いた研究[3]がある．この研究では，実測されたデータをサンプルデータとみなし，そのようなデータが仮に十分長い期間取得されたときに正しく将来の行動を予測できる確率の理論的上限を，時系列データのエントロピーを求めることによって見積もっている．その値は93%であり，利用者による差はあまり見られない（最低でも80%と示されている）．2010年の震災後のハイチで行われた同様の研究では最頻値が80%程度で，最低でも50%程度の確率で正しく予測できうることが示されている[4]．一方で，Lu et al.はコートジボワールにお

ける50万人分の携帯電話の約4ヶ月分の通話履歴に対しマルコフチェーンをベースとした予測モデルを適用し，88%程度の確率で将来の行動を正しく予測することが実際に可能であることを示している[5]．これらの既存研究の知見を見る限りでは，一見，利用者の過去の行動履歴を活用することにより，その利用者の将来の行動をかなり正確に予測できるように見える．

しかしここで注意したいことは，この93%や88%という数字は「パッシブ型交通行動データを使えば，どのような利用者であっても，『いつ』交通システムを利用するかを正確に予測できる」ことを必ずしも意味していないことである．この問題は低利用頻度の利用者について深刻である．いま，$p_t^i = 0.01$ の利用者の行動を予測することを考えよう．この利用者に対して少なくとも「常に交通システムを使用しない」というナイーブな予測を立てておけば，それだけでその利用者の行動を99%の確率で正しく予測したことになる．もちろんこのような予測は σ^2 を小さくするためには何の役にも立たない．一方で，$p_t^i = 0.5$ の利用者に対してであれば，パッシブ型交通行動データを活用することによって，このようなナイーブな予測よりもよい予測を得る（すなわち，p_t^i を0.5よりも0または1に近い値にする）ことができよう．このことは，Song et al. において，周期性のような行動の順序を考慮しないときの予測確率の理論的上限は93%よりはかなり低い値になることが示されていることからも期待が持てることである．以上のことは，p_t^i が0ないし1に極端に近くない利用者については各利用者の過去の利用履歴の活用により予測精度を向上できうる一方で，そうでない利用者の予測精度は利用頻度によるナイーブな予測以上には向上しえないことを示唆する．

本稿では，すべての利用者について $\min\{p_t^i, 1-p_t^i\}$ を0.5以下の正数 q 以下にすることができる，各利用者の過去の利用履歴データに基づく仮想的な利用者行動予測モデル（以降では「q-予測モデル」と呼ぶ）の存在を仮定する．より具体的には，N期の期間における利用回数が ξ 回である利用者が将来の期tにおいて交通システムを使用する確率 $p_t(\xi)$ が，

$$\min\{p_t(\xi), 1-p_t(\xi)\} = \min\left\{\frac{\xi}{N}, q\right\} \tag{20}$$

を満たすように計算する予測モデルをq-予測モデルとする．式(20)は

1. ξ が Nq 以下の利用者と $N(1-q)$ 以上の利用者については，$p_t(\xi) = \xi / N$ （すなわち，期間中の利用頻度そのもの）とし，
2. それ以外の利用者については，$p_t(\xi) = q$ または $p_t(\xi) = 1-q$ を期tに応じて予測し設定する

ことを意味する．このうち1.は，「q-予測モデルはどの利用者についても，将来の利用確率を q ないし $1-q$ よりも精度よく（すなわち，これらが0ないし1に近い値になるように）設定できない．よって，過去の利用履歴が非常に少ないか多い利用者については，その利用頻度そのものを将来の利用確率とするナイーブな予測手法しか適用できない」ことを意味している．q-予測モデルの実装については本稿では考えないが，過去の利用履歴を相当長期間（最低でも$1/q$期分）同一個人について追跡することが要求されることは確かであり，プライバシー保護の観点からは負担の大きいモデルといえよう．

q-予測モデルを用いた際の分散 σ_q^2 を計算し，それが式(16)に比べてどれだけ小さくなるかを評価する．σ_q^2 は式(13)と式(20)により，まず

$$\sigma_q^2 = \sum_{\xi=1}^{N} \min\left\{q(1-q), \frac{\xi}{N}\left(1-\frac{\xi}{N}\right)\right\}\frac{n\xi^{-s}}{H_{N,s}} \tag{21}$$

と計算できる．右辺の min 演算子を展開すれば

$$\sigma_q^2 = \sigma_A^2 + \sigma_B^2 + \sigma_C^2 \tag{22}$$

ただし，

$$\sigma_A^2 = \sum_{\xi=1}^{N_1} \frac{\xi}{N}\left(1-\frac{\xi}{N}\right)\frac{n\xi^{-s}}{H_{N,s}} \tag{23}$$

$$\sigma_B^2 = q(1-q)\sum_{\xi=N_1+1}^{N-N_1} \frac{n\xi^{-s}}{H_{N,s}} \tag{24}$$

$$\sigma_C^2 = \sum_{\xi=N-N_1+1}^{N} \frac{\xi}{N}\left(1-\frac{\xi}{N}\right)\frac{n\xi^{-s}}{H_{N,s}} \tag{25}$$

$$N_1 = Nq \tag{26}$$

と計算できる．ここで，σ_C^2 は σ_A^2 に比べて σ_q^2 への寄与度が低いこと（特に q が 0 に近いとき）を考慮して，

$$\sigma_q^2 = \sigma_A^2 + \sigma_D^2 \tag{27}$$

ただし，

$$\sigma_D^2 = q(1-q)\sum_{\xi=N_1+1}^{N} \frac{n\xi^{-s}}{H_{N,s}} \tag{28}$$

と近似的に計算する．これはちょうど，式(21)の ξ が大きいときの min の項の部分を $q(1-q)$ で統一して計算していることに相当する．min 以外の部分はべき関数なので，ξ が大きい項の寄与度は小さくなることに注意．式(23), (28)を，式(14)を代入して書き直すと

$$\sigma_A^2 = \frac{\mu}{H_{N,s-1}}\left(H_{N_1,s-1} - \frac{H_{N_1,s-2}}{N}\right) \tag{29}$$

$$\sigma_D^2 = \frac{\mu N}{H_{N,s-1}} q(1-q)\left(H_{N,s} - H_{N_1,s}\right) \tag{30}$$

となる．$H_{N,s-1}$ および $H_{N,s-2}$ に式(17)を適用し

$$\sigma_A^2 = \mu q^{2-s}\left(1 - \frac{2-s}{3-s}q\right) \tag{31}$$

を得る．一方，式(30)は，和を積分に置き換えて

$$\sigma_D^2 = \mu(2-s)q(1-q)\frac{(q^{1-s}-1)}{s-1} \tag{32}$$

と近似的に計算できる．これらを足しあわせて

$$\sigma_q^2 = \mu\left\{q^{2-s}\left(1 - \frac{2-s}{3-s}q\right) + (2-s)q(1-q)\frac{(q^{1-s}-1)}{s-1}\right\} \tag{33}$$

となる．この式は s が 2 の近辺では

$$\sigma_q^2 \simeq \mu \tag{34}$$

と近似できる．また，s が 1 の近辺では

$$\sigma_q^2 \simeq \left\{\frac{2\log(q)(q-1)+2-q}{2}\right\} q \tag{35}$$

と近似できる（$\lim_{q\to 0} q\log q = 0$ に注意）．

式(34)，(35)の計算結果は q-予測モデルの導入による分散減少の効果は q の減少に対して緩慢であることを示唆する．特に s が 2 に近い場合には $\sigma_q^2 \simeq \mu$ であり，式(19)に $s=2$ を代入して得られる q-予測モデルを導入する前の分散と同じである．また，第 2 章で示したように，交通システム運用問題によれば予測値の精度向上の効果による運用利益の増加量は概ね予測値の標準偏差に線形に比例することがわかっている．このことは，q-予測モデルによる最終的な利益の増加量を計算するには式(33)からさらに平方根をとる必要があることを意味する．**図-3** に複数の s の値のときの標準偏差 σ_q を q-予測モデルを使用しなかったときの標準偏差 σ で割った値を，式(33)を用いて計算した上でグラフに示した．式(33)は q が小さいことを前提とした計算なので，**図-3** では $q \le 0.3$ の区間のみをグラフにしてある．

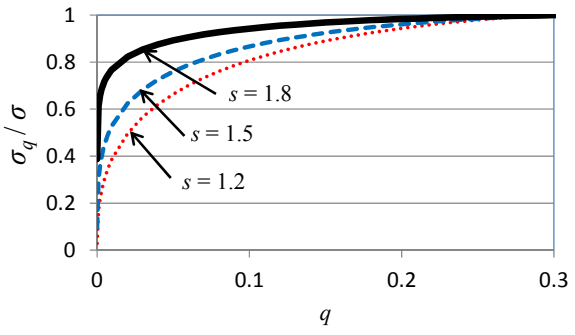

図-3　q-予測モデルによる予測誤差（予測利用者数の標準偏差）の改善の程度

4.　考察と今後の課題

(1)　データ活用とプライバシー保護のトレードオフ

3 章の結果は，データ活用により得られる利益の大小がどの程度になるかを定量的に示しており，これを用いれば，本稿で前提としている状況において，プライバシー保護の観点からは不利な個々人の行動履歴データの活用がどれだけ交通システムにメリットをもたらすかを考察できる．いま，**図-3** 示される計算結果と，σ は単に集計された利用者数の分散から算出できる（すなわちプライバシーの懸念が個人を追跡するよりかなり少ない）ことを考慮すれば，

○　s が 2 に近いか，あるいは q をあまり小さく（概ね 0.1 以下）できない場合には，利用履歴データを活用するメリットが少なく，プライバシー保護のメリットのほうが卓越する可能性が，そうでない可能性よりも相対的に大きい．

○　s が 1 に近く，かつ q をかなり（0.01 程度以下まで）小さくできる場合には，利用履歴データを活用するメリットが卓越する可能性が，そうでない可能性よりも相対的に大きい．

ことがいえよう．なお，式(19)および(35)でも示されるように，将来の利用者数の分散はその期待値に比例しているので，標準偏差は期待値の平方根に比例することになる．このことは，

○　利用者数が多い大規模な交通システムであれば，そもそも利用者数の予測の誤差による運用利益の損失が利益全体に比べて小さく，σ_q / σ が相当小さくなっても個々人の利用履歴データを活用するメリットが少ない．

○　利用者数が少ない小規模な交通システムであれば，利用者数の予測の誤差は全体の利用者数に対して相対的に大きくなる．よって，個々人の利用履歴データを活用して正確な予測を立てることは運用利益の改善に十分に貢献しうる．

ことを示唆する．以上のことは，プライバシー保護の問題を考慮したとても，個々人の利用履歴データを有効に活用したほうがよい場合は

○　利用者数が少ない小規模な交通システムである

○　s が 1 に近い（＝多頻度利用者が相対的に多い）

○　予測モデルの精度がよい（q を小さくできる）

であることを示唆するものである．

個々人の利用履歴データを有効に活用できる状況としては，以上の条件を満たす状況のほか，本稿で設定した仮定が適用できない状況がありうる．第 1 の可能性としてはジップの法則が適用できない状況が挙げられよう．これは，特に通勤時など利用者の移動の目的が偏在しい

ている時間帯や場所などで起きうるかもしれない．このような場合については別の解析が必要である．

第 2 の可能性としては利用者間の行動が互いに独立でないことである．例えば高速道路のランプにおける断面交通量に過分散がある（飯田・高山 [6]，井料ら [7]）ことを考えれば，このような可能性は十分ありうる．特に井料ら [7]は簡単な非集計的モデルを用いて過分散を利用者間の行動の相関として説明している．このような場合において予測精度を向上させるためには，利用履歴データを 1 人 1 人独立に分析するのではなく，全員分を同時に分析して異なる利用者間の行動の相関関係を調べることが必要となるだろう．もちろん，このことは利用者数が膨大な場合はかなり複雑な計算を要求することが予想されるため，解決すべき課題も多いと思われる．

(2) トレードオフの優れたデータ活用法の提案

本稿では交通行動データの活用法として「利用履歴から将来の行動を予測する」というものを想定し，そしてその活用法が必ずしも効果的なものとはいえないことを示した．このことは，逆に捉えれば「行動履歴からの予測アプローチ以外のアプローチを試す」ことが，より効果的な交通行動データの活用法であることを示唆しているともいえよう．

本稿の解析において予測精度の向上がよくなるとは限らなかった理由には，予測の際に相当な人数の低頻度利用者を相手にしなくてはならなかったことが挙げられる．低頻度利用者はその行動の予測性が低く，せいぜいランダムに動くとしか捉えられない．正確な予測には，彼らの動きを何とかして精度よく予測しなくてはならない．このために有効なアプローチのひとつとして，

○ 交通行動データから，ある利用者が当該の交通システムを使う前兆（前駆行動）を検出し，そのような利用者を潜在的利用者として抽出する．

方法があろう．これの最も簡単な例としては「駅周辺にいる人の数を数えて，近い将来の電車の利用者数を推定する」というものが挙げられよう．この例における前駆行動は「駅周辺にいる」ことである．より積極的なアプローチとしては，例えば石村らによる経路検索サービスの利用実績を用いた手法 [8]もある．これらの手法には利用者行動の情報を集計しても推定精度への影響がないものもあり，プライバシー保護とのトレードオフという意味でも優れているといえる．

上記以外にも，交通システムの運用よりも計画により適したアプローチとして

○ 多頻度利用者の利用履歴から交通行動モデルをつくり，それを低頻度利用者にもあてはめる

というものも考えられる．このアプローチはデータオリエンテッドというよりはむしろ伝統的なモデルベースに属するものであり，なおかつ日々の交通需要の変動など，運用への活用は苦手であることが予想される．平均的な利用者数の予測が問題となる計画への適用がより適していると思われる．このときはパッシブデータ以外のデータソース（Stated Preferene 調査など）との優位性の比較検証も必要であると思われる．

(3) 今後の課題

本稿で用いたモデルそのものの改善点をいくつか挙げる．本稿では利用者の選択肢は「使う」か「使わない」だけであり，交通サービスは単一で運用者は容量を増減させる以外の選択肢を持っていなかった．交通サービスについて解析する以上は，交通ネットワークを考慮した多様な選択肢下での分析が必要となろう．利用頻度分布の実証については，時間的空間的により細分化した集計単位で行い，ジップの法則がどこまで成立するかを検証する必要がある．これらの現行モデルの改善に加えて(2)節で議論した別のデータ活用アプローチを検証するためのモデル構築と解析も重要となろう．

謝辞：本研究で用いた ETC 統計データは阪神高速道路株式会社より提供をいただいたものである．本研究の一部分は科学研究費補助金（挑戦的萌芽：25630217）の支援によりなされた．この場を借りて感謝の意を表する．

参考文献

1) 竹之内隆夫：*k*-匿名化技術と実用化に向けた取り組み, 情報処理, Vol. 54, No. 11, pp. 1125-1129, 2013.
2) Newman, M. E. J: Power laws, Pareto distributions and Zipf's law, *Contemporary Physics*, Vol. 46, No. 5, pp. 323-351, 2005.
3) Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. : Limits of predictability in human mobility, *Science*, Vol. 327, No. 5968, pp. 1018–21, 2010.
4) Lu, X., Bengtsson, L., and Holme, P., Predictability of population displacement after the 2010 Haiti earthquake, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 109, No. 29, pp. 11576–115781, 2012
5) Lu, X., Wetter, E., Bharti, N., Tatem, A. J., and Bengtsson, L., Approaching the Limit of Predictability in Human Mobility. *Scientific Reports*, Vol. 3, 2013.
6) 飯田恭敬, 高山純一：高速道路における交通量変動特性の統計分析, 高速道路と自動車, Vol. 24, No. 12, pp. 22-32, 1981.
7) 井料隆雅, 岩谷愛理, 朝倉康夫：都市高速道路における時間帯別流入交通量の週変動分析, 第 27 回交通工学研究発表会論文報告集, pp. 173-176, 2007.
8) 石村怜美, 太田恒平, 富井規雄：経路検索サービスの実績データに基づく近未来の突発的移動需要の検出, 第 47 回土木計画学研究発表会, 2013.