様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ２ 日現在

機関番号：１４５０１
研究種目：挑戦的萌芽研究
研究期間：2013～2014
課題番号：２５６３０２１７

研究課題名（和文）匿名性を担保した交通行動データの流通促進のための理論

研究課題名（英文）Theory for promoting distribution of transport behavioural data with privacy protected

研究代表者
井料　隆雅（IRYO, TAKAMASA）

神戸大学・工学（系）研究科（研究院）・教授

研究者番号：１０３６２７５８

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：匿名性の担保とデータの価値の維持という2つの相反する目的を満たしつつ，公衆から交通行動データを収集し集計するシステムを開発するために必要な理論的基礎を構築した．「交通行動データの収集」と「交通行動データの活用」の2つを記述する抽象化されたモデルを提案し，これらを組み合わせて，取得されるデータの匿名性と有用性のトレードオフを評価した．あわせて情報科学分野での既存技術を調査しそれを応用した．交通運用への活用においては，集計による匿名化を行ってもデータの価値は大きく下がらないこと，前駆行動に関するデータがむしろ重要であることを示し，あわせて，交通計画における匿名性を担保した活用方法を提案した．

研究成果の概要（英文）：This study constructed a theoretical framework that is necessary to build a system collecting and processing transport behavioural data from public with privacy protected and value of the data preserved. Abstract models describing collection processes of transport behavioural data and its utilisation were built. Combining these two models, trade-offs between anonymously and value of the data obtained from public were evaluated. In addition, existing techniques in the information science field has been reviewed to apply them to the theory developed in this study. We showed that, in the context of the transport system management, anonymising privacy by aggregating individual data and cutting tracked trajectories does not deteriorate its value very much, while data indicating travellers' near-future action is more important. Further, we proposed an approach to utilising data for planning purpose with privacy protected.

研究分野：交通工学

キーワード：個人情報　匿名化　交通システム　ビッグデータ

１．研究開始当初の背景

本研究を開始した 2013 年はビッグデータとよばれる巨大なデータ群への注目が一般社会においても一気に集まった年であった．交通工学の分野においては，ビッグデータという言葉が流行するよりもかなり前から，交通サービスやそれに関連するサービスの提供に付随して自動的かつ継続的に交通行動データを収集する方法に注目が集まっていた．古典的には交通流の状態を断面ごとに観測する車両感知器（あるいは車両検知器）が活用されていたが，近年では交通系 IC カードや ETC(Electric Toll Collection; 自動料金収受システム)の利用履歴のような追跡型のデータの活用方法の研究もすでに着手されていた．

ビッグデータの有用性に着目が集まった 2013 年であったが，この年は，ビッグデータ活用におけるプライバシー問題もクローズアップされた年であった．社会的に注目されたのは大手鉄道会社の交通系 IC カードの使用履歴の他社への販売に関する問題である．データを収集した主体が，データを活用する際にデータの解析を第三者に依頼することは当然ありうることである．さらに，データより学術的な価値を見出そうというのであれば，任意の第三者がデータできることはなんらかの形で担保しなくてはならない．これらのことを考えれば，上述のような，プライバシーを理由とするデータの流通に関する障害は，データの利用価値を大きく毀損するものにもなりかねない．このようなプライバシー保護の問題は，最近では技術的制約よりも大きい障害となりつつあると考えても差し支えないであろう．

プライバシーを担保したままデータを流通させるには，データに含まれる個人情報を，データが第三者に渡される前に消去すること（以降ではこれを「匿名化」と名づける）が求められる．このことは，しばしば，単にデータから氏名や住所などの個人情報そのものを消す操作と間違えられる．しかしそのような操作は匿名化という意味ではまったく不十分である．個人情報を知ろうとする主体（攻撃者）は，当該データだけでなく，他のデータを組み合わせて攻撃を仕掛けることが可能だからである．著名な例として，当時の公開データだけを用いて米国マサチューセッツ知事の病歴を暴けることを示した研究が知られる[1]．このようなプライバシーに対する攻撃への対策の研究は，情報科学の分野で近年盛んに行われている[2]．ビッグデータ的手法で取得された交通行動データにおいても，それらの分野で開発された手法を応用することはある程度は有効であろう．

交通データの匿名化を議論する際には，交通データに含まれるプライバシーがどのようなものであるかというに加えて，交通データに含まれる「価値（有用性）」はどのようなものであるかを並べて考察しなくてはならない．匿名化の操作は必ずデータの質の劣化を伴う．もし匿名化だけが目的であれば，データの質がどう劣化しようとかまわないわけで，全く原型をとどめないような集計操作（たとえば，交通 IC カードの利用履歴データを，全路線の 1 日の総乗降数の合計に変換する）をしてしまえば目的は達成されたことになる．もちろんこれではデータの価値を十分に活かしたことにならない．匿名化の方法は常にデータが持つ価値とのトレードオフで評価されなくてはならない．このことは，交通行動データの匿名化手法の評価をする場合には，

○　その手法がどれだけプライバシーへの攻撃に耐えるか
○　その手法がどれだけデータの価値を保存できる（損ねない）か

のトレードオフを考慮しなくてはならないことを示唆する．

上述の議論で特にポイントとなるのは「プライバシー」と「データの価値」をどのように定めるかである．前者は既存研究による考え方を援用すればよいが，後者については，特に交通工学への応用というコンテクストにおいては多大な注意を要する．交通工学への応用というコンテクストのもとでプライバシーを含むデータの匿名化技術を開発するのであれば，その価値は，

○　交通システムの計画や運用にどれだけ役に立つか

で測られるべきである．この問題は交通工学の分野に特有の問題であり，交通工学特有の状況を考慮しながら解決しなくてはならない．このことが，交通行動データの匿名化技術の開発の際に，情報科学における一連の研究との差異を示す重要なポイントである．

２．研究の目的

本研究では，匿名性の担保とデータの価値の維持という 2 つの相反する目的のトレードオフを考慮し，これら 2 つの目的を満たしつつ公衆から交通行動データを収集し集計するシステムを開発するために必要な理論的基礎を構築することを目的とする．本研究では具体的なシステムそのものではなく，その構築のための理論的基礎を示すことを目的としている．これを実現するために，「交通行動データの収集」と「交通行動データの活用」の 2 つについて，できるだけ抽象化を行ったモデルを提案する．これらを組み合わせて，取得されるデータの匿名性とその有用性のトレードオフを定量的に評価する．匿名化技術については情報科学分野での既存技術の調査も行う．上記を総合して研究目的を達成させる．

３．研究の方法

研究の方法は以下からなる．

(1) 交通行動データの収集のモデル化
交通行動データの収集がどのように行われ，それがどのような情報をデータ収集者に提供するかをモデル化する．既存の技術的社会的制約にとらわれずに，理論的にどこまでどのようなデータが収集できるかを考察しながら，抽象的なモデルを構築する．

(2) 交通行動データの活用のモデル化
交通行動データの活用方法をモデル化する．一般に複雑な交通システム全体を仔細にわたってモデル化することは煩雑であり，本研究の目的を考慮すれば意味がない．よってここでは，交通システムの一般的な特徴を考慮しつつ単純かつ抽象的なモデルを構築することを目指す．交通計画への活用と交通運用（交通制御）への活用の 2 つそれぞれについてモデルを構築する．

(3) データの匿名性の有用性のトレードオフの評価
(1)(2)の知見を組み合わせてデータの匿名性と有用性のトレードオフを評価する．

(4) データの匿名化技術の調査
データの匿名化技術に関する既存研究等を調査する．このとき，そもそも匿名化とはどのようなことを成した状態と考えるべきなのかもあわせて調べる．調査は情報科学分野を対象に行う．

(5) 匿名性を担保した交通行動データの活用に対する理論的基盤の提案
(1)～(4)を組み合わせることにより，表題の研究目的を達成する．

４．研究成果

(1) 交通行動データの収集のモデル化
交通行動データの収集の抽象モデルとして，物理的に収集可能な交通行動データを最大限まであつめる装置を仮想的に考え，それに Extreme Life Logger (ELL)という名をつける．いま，交通行動を行う主体は人（個人）であり，ELL はこの個人にかかわる一切のデータを収集できるとする．一方で，ELL はあくまでも「装置」であり，人の心の中までをのぞく事はできないとする．このことを考慮すれば，ELL が収集できる情報は，

1. その人が過去にとった交通行動に関する情報
2. その人が将来どのような交通行動を取るかを示唆する行動に関する情報

の 2 つに分類することができる（両方に分類されることもある）．なお，ここでいう交通行動とは，実際に（目的を持った）移動を行う行動そのものを指すとしている．
実際の交通行動を直接観測して得られるデータ（たとえば GPS による位置情報データ）はすべて 1.に分類される．これらはいわゆる「顕示選好」を反映するものであり，当然ながら，必ず，過去の交通行動を反映したものとなる．
2.は，ELL が，観測対象である個人の将来の具体的な交通行動と直接関係する行動を観測することにより得られる．ここで「直接関係する」は，たとえば，過去の習慣から予測する，というようなあいまいな取得方法ではなく，「具体的に将来の交通行動に直結するロジックが存在する」というほどの強い関係を意味する．最も典型的なのは「予約」である．また，バス停に立っていることを観測すれば，そのバス停から出発するバスのいずれかに乗る，という関係，遠隔地での会議の約束を取り付けたメールから，その場所へ向かう交通手段のいずれかを使用する，という関係も考えられる．本研究ではとくに 2.に属する観測対象者の行動を「前駆行動」と名づけ，これまでの交通行動データでは一般的であった「過去の」交通行動を追跡するデータと質的に異なるものと捉える．

(2) 交通行動データ活用のモデル化
本研究では交通行動データの活用を交通計画と交通運用の 2 つの状況に分けて考える．いずれの状況においても，交通工学において交通行動データを活用する究極的な目標を「ある交通機関の将来の利用者数をできるだけ正確に予測する」ことにおく．
本研究では，交通行動データの交通計画への活用は，すべて「交通行動モデルのパラメータ推定」を通して行われると考える．たとえば，経路選択に関する過去の顕示選好を分析して経路選択モデルを構築し交通システムの将来計画に役立てる，あるいは，OD（起終点）交通量を需要モデルのパラメータとみなして観測する，という方法である．
理想的には，適切なデータとモデルを用いることにより，上記のようなモデルは季節変動や将来成長などによる時系列による変動も含めて完全に正確な予測を立てられると考えることもできるだろう．しかし，そうであったとしても，その予測は期待値に対するもの（あるいは，分散などを含めた確率分布に関する情報）にとどまり，確率的なゆらぎまで含めた正確に予測することは困難と考えられるだろう．これは，たとえば，交通量がポアソン分布に従うことまでわかり，かつ，その母数（期待値）までは予測できるにしても，具体的な値まではわからない，ということに対応する．本研究では，交通計画への適用においては，そのような確率的なゆらぎによる部分まで予測する必要はないと考える．
交通行動データの交通運用への活用は，特定の交通機関の近い将来（たとえば，1 時間後とか，明日など）における利用者数を，上記で示した「ゆらぎ」の部分まで含めて予測することと本研究では考える．このことの意味を，道路における交通信号を例に説明する．ある交差点における交通信号のパラメータ

は，もっとも初歩的には，その交差点の各枝にサイクルごとにやってくる「平均的な」交通量を基に決定した値に固定する．これは本研究でいうところの交通計画に相当する．一方，本研究でいうところの交通運用は，SCOOT のような適応型信号制御システムのように，各サイクルにおいて「実際に」やってくる車両の台数を予測し，その情報を活かして信号パラメータを動的に制御することに相当する．このような，近い将来の利用者数を，ゆらぎの部分まで精密に予測することは，特に，今後広く普及することが期待される PMV(Personal Mobility Vehicle)のような少量輸送サービスの効率的な運用に寄与するであろう．

交通行動データの交通運用への活用モデルの数理的な概要は以下の通りである：

1. ある容量を持つ交通機関を考える．
2. 容量は直近に調整可能である．大きい容量を設定すればするほど費用がかかる．
3. 運用者は直近の将来の利用者数を確率分布の形で予測し，それを用いて利益最大になるように容量を決定する．
4. 交通行動データを用いることにより，上記の確率分布の分散を小さく出来る．

(3) データの匿名性の有用性のトレードオフの評価

交通計画におけるデータの匿名性と有用性のトレードオフについては基本的には問題とならない．これは，交通計画においては，データの活用は「交通行動モデルのパラメータ推定」を通じて行われるとするからである．このパラメータ推定が一種の集計操作となることが期待できる．

パラメータ推定を安全に行うには，プライバシーを含んだデータそのものを流通させるのではなく，データを収集し１次的に保持する主体が，第 3 者からの要求（クエリ）に答えて集計し，その結果だけを返せばよい．ただし，このような手続きによるプライバシー漏洩を防ぐには，第 3 者が発行したクエリが，特定の個人のプライバシーを取得することを意図していないことをなんらかの形で確認しなくてはならない．これについては，(4)の成果を踏まえた上で(5)にて示す．

交通運用においては，(3)で示したモデルをベースに定量的な評価を行った．評価は 2 つの異なる設定において行った．1 つめ（問題１）は単一の交通機関が，過去の個人の移動履歴の追跡データを用いた際に，運用レベルでの近い将来（たとえば，明日）の利用者数をどれだけ正確に予測できて，どれだけ容量を調整でき，結果としてどれだけ利益を高められるかという問題である[3]．もう１つめ（問題２）は，複数の交通機関がネットワークを構成し，利用者はそれらを組み合わせて移動するときに，ある時間帯における各交通機関の利用台数（断面交通量）の情報を用いて，その直後の時間帯の各交通機関の利用台数の予測精度を向上しようという問題である[4]．前者の問題は，「プライバシー情報を含む追跡型ではあるが，過去の情報であるものが，交通運用においてどれだけの価値を持つか」を評価する．後者の問題は，「プライバシー情報を含まない前駆行動に関する情報がどれだけの価値を持つか」を評価する．

問題１の解析の結果は，端的に言うと，「個々人の過去の交通行動を追跡したデータは，交通運用の効率を高めることには大きくは寄与しない」であった．理由は２つある．ひとつは，個々人の過去の交通行動のパターンの解析からは，習慣的行動は予測できても，非日常的な行動を予測することは原理上不可能だからである．もうひとつは，交通機関を利用する人の多くは低頻度利用者であり，習慣的に利用する高頻度利用者が占める割合は決して大きくないからである．後者の特徴は，より具体的に言えば，利用頻度と利用回数の関係がジップ則に従うと表現できる．定量的に見ると，過去の追跡データから 90%の行動が予測できたとしても（既存文献[5]では 88%程度可能であるという実績がある），予測誤差の改善は，追跡データを使わないときに比べて 5%～20%程度にとどまる結果となった．図１にその結果をグラフで図示する．横軸の q は追跡データによる予測率を 1 から引いたものである（具体的な定義は当該文献を参照のこと）．s は利用頻度を示すジップ則のべき乗のパラメータであり，大きいほど低頻度利用者がより多い分布となる．なお，都市高速道路の実データによる解析では s=1.48 程度であった．

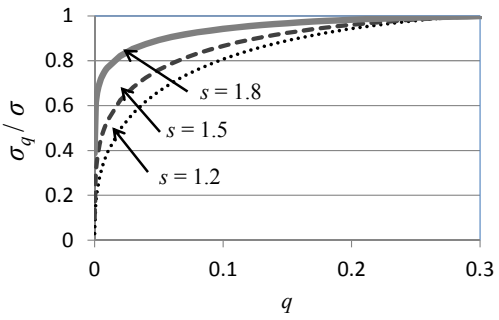

図１　行動予測精度（横軸，1-q）と利用者数予測誤差（縦軸）の関係[3]

問題２の解析の結果は，「断面交通量のような集計データであっても，それが前駆行動に属するものであれば，直後の交通量推定の精度向上に寄与する」であった．このことを説明するネットワークの例を図２に示す．この例では，ある時間帯にリンク１に相当する交通機関を利用した人は，次の時間帯にリンク２または３に相当する交通機関を使用する．過去の利用履歴より，リンク１から２または３のいずれへ移動するかの割合（分岐率）の期待値は正確に分かっているとすれば，

リンク１の断面交通量を観測した結果を用いて，リンク２と３の次の時間帯の断面交通量の予測精度を向上させることが可能である．この向上の程度は分岐率に左右されるため，ネットワーク構造と需要パターンに依存する．スケールフリーネットワークを数値的に生成して検証したところ，2 割程度のリンクについて，予測精度が 20%以上向上することがわかった．この値は，問題１で追跡データを用いたときの値と同等であるが，問題２では追跡データのようなプライバシーを含むデータを使用しないことに注意したい．

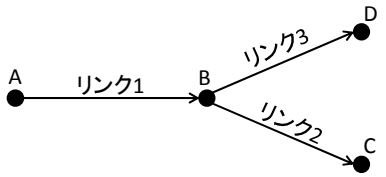

図２　問題２の例題のネットワーク構造[4]

(4) データの匿名化技術の調査

データの匿名化技術については情報科学において多くの既存研究がある．ここでは本研究に関連したものについてのみ簡単に述べる．

第１章で述べたように，データの匿名化を行うためには単に氏名等の個人情報をデータから削除するだけでは十分ではないと考えるのが既存研究における一貫した方針である．データの匿名性が議論される場合は，暗号の強度を議論するときと同様に，悪意をもった攻撃者の存在を仮定することが既存研究では一般的である．そのような攻撃者は匿名化を行おうとする当該のデータセットを，個人情報を含む別のデータセットと積極的にリンクさせることにより，当該のデータセットの個人情報を復元しようとする．このような悪意を持った攻撃者やリンク可能な個人情報データの存在の仮定は，匿名化という意味ではかなり安全側に機能することが考えられ，場合によっては過剰な匿名化を招く可能性もあるが，当面の社会的状況において，匿名化技術のベンチマークとしては有効に機能するだろう．

他のデータとのマッチングによる攻撃が成功する可能性を減らすために有効な操作には集計は集計処理である．特に *k*-匿名性とよばれる手法[1][2]はよく知られている．*k*-匿名性では，匿名化の対象となるデータを *k* 人ずつ分に集計することにより，そのデータと特定の個人が関連付けられることを抑止する．たとえば OD 交通量データであれば，ゾーンや時間帯を細かくすると，OD 交通量が 1 人（台）になってしまうところが出てしまうことがある．そのような場合に，周辺のゾーンや時間帯を結合させて OD 交通量が *k* 以上になるようにすれば *k*-匿名性を担保できる．この操作はETC-OD 交通量を時間帯別に分解する際にも有効である[6][7]．従来，時間帯別 OD は固定された時間帯で集計することが一般的であったが，平均的に *k* 人（台）ごとに集計できるように時間帯を指定することにより，時間分解能と *k*-匿名性を両立させることが可能である．

集計操作以外によく用いられる手法としてはランダマイゼーションがある．これは，データに人為的なノイズを加えて個人情報を隠蔽しようとする考え方である．この考え方を定量化する概念に差分プライバシーと呼ばれるものがある[8]．これは，データベース D とデータベース D’(D と 1 レコードのみ異なる)をクエリに対する答えから区別させないようにする概念である．

(5) 匿名性を担保した交通行動データの活用に対する理論的基盤の提案

匿名性を担保した交通行動データの活用に対する理論的基盤を，交通計画における活用と，交通運用における活用のそれぞれにわけて示す．なお，本研究における交通計画と交通運用の定義は(2)に示している．

交通計画における活用については，(3)の前半で示したように，「プライバシーを含んだデータそのものを流通させるのではなく，データを収集し１次的に保持する主体が，第 3 者からの要求（クエリ）に答えて集計し，その結果だけを返す」方法が有効である．第 3 者が発行したクエリが，特定の個人のプライバシーを取得することを確認するには，(4)で示したランダマイゼーションの考え方を適用できる．すなわち，元のデータにノイズを付加し，それを元にクエリで要求された集計操作を行ってその値を返す操作である．この際の匿名性の定量化に差分プライバシーの考え方を利用する．

過去の行動を追跡するデータは人々の交通行動の法則を発見する際には相当の有用性があるだろうが，この「法則」とはあくまでも集団としての法則であることに注意したい．(3)の後半でも示したように，個々人の行動法則を，習慣的なものを超えて発見することは現状では困難である．集団の法則を示すデータは一種の集計データであり，そのデータだけが流通するようにすれば，匿名性が担保されたまま，追跡型データの価値を活かせるであろう．

交通運用においては，(3)の後半で示したように，個々人の過去の追跡データの有用性は決して高くない一方で，前駆行動を捉えるデータは集計されていても一定の有効性があることがわかった．これらの結果は，有用性とプライバシー保護のトレードオフがよいとはいえない過去の追跡データをそのまま流通させることはあまり得策ではないことを示唆する．そのようなことにこだわるよりも，むしろ，追跡を切断し，*k*-匿名性を担保するような集計処理を行ったデータの流通を促進させることが優先されるべきである．より促進すべきは前駆行動に関するデータの流通である．このようなデータは集計的

であっても，交通運用の効率化には大きく寄与することが期待できる．このようなデータの有用性については他の研究でも示されている[9]. 前駆行動に関するデータは本質的にリアルタイム性の高いものであり，流通にも迅速性が求められ，そのための技術開発が今後求められることになるだろう．

＜参考文献＞

[1] Sweeney, L., *k*-anonymity: A Model for Protecting Privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), pp. 557-570, 2002.
[2] 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.
[3] 発表論文（学会発表②）
[4] 発表論文（学会発表①）
[5] Lu, X., Wetter, E., Bharti, N., Tatem, A. J., and Bengtsson, L., Approaching the Limit of Predictability in Human Mobility Scientific Reports, 3, 2013.
[6] 発表論文（雑誌論文①）
[7] 発表論文（学会発表③）
[8] Dwork, C., Differential Privacy, in: Automata, languages and programming, Springer, Berlin Heidelberg, pp. 1-12, 2006.
[9] 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，土木計画学研究発表会・講演集, 47, CD-ROM, 2013.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1　件）

① 上田大樹，井料隆雅，朝倉康夫，長期 ETC 統計データによる異なるランプ間ＯＤ交通量と旅行時間の相関分析，交通工学，査読有，　49(3)，2014，pp. 43-52

〔学会発表〕（計　3　件）

① 井料隆雅，日下部貴彦，原祐輔，限定的な利用者行動追跡データに基づく利用者数の短期間予測問題，第 12 回 ITS シンポジウム，2014.12.4，東北大学（宮城県）

② 井料隆雅，原祐輔，日下部貴彦，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，第 49 回土木計画学研究発表会，2014.6.7，東北工業大学（宮城県）

③ 小篠耕平，井料隆雅，朝倉康夫，粒子フィルタを利用した都市高速道路における潜在的ランプ間 OD 交通量の推定，情報処理学会第 76 回全国大会,2014.3.13，東京電機大学（東京都）

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕なし

６．研究組織
(1)研究代表者
井料　隆雅　（IRYO, Takamasa）
神戸大学・大学院工学研究科・教授
研究者番号： 10362758

(2)研究分担者
原　祐輔　（HARA, Yusuke）
東北大学・大学院情報科学研究科・助教
研究者番号： 50647683

日下部　貴彦（KUSAKABE, Takahiko）
東京工業大学・大学院理工学研究科・助教
研究者番号：80604610

(3)連携研究者
なし