パーソントリップデータにおける匿名性に関する考察*

Experiments on Location Data Anonymization for Person Trip Survey*

松崎和賢**・廣田啓一***・高橋克巳****・白井康之*****
By Kazutaka MATSUZAKI**・Keiichi HIROTA***・Katsumi TAKAHASHI****・Yasuyuki SHIRAI*****

1. はじめに

近年のGPS端末の普及に伴い，ライフログとしての移動履歴データに基づく情報提供サービスや，蓄積されたデータを利用した道路・交通等の公共機関での利用，また，マーケティング分野への応用が期待されている．しかしながら，位置情報の精度が高まるほど，これらの時系列的な履歴から，個人が特定される危険性もまた高まっている．今後，移動履歴データをさまざまな用途において活用していくためには，これらのデータに対して個人が特定できないような匿名性を確保していくことは，2次的な利活用を進めるにあたって極めて重要な課題である．本稿では，平成10年首都圏パーソントリップデータを利用して，主に k-匿名性という考え方に基づいて検証ならびに実験を行った結果を記し，今後の移動履歴データの匿名化に関する指針を与える．

2. 移動履歴データの利用と匿名化

個人の位置に関する情報だけでなく，移動に関する情報を取得することが，近年極めて容易になってきている．例えば，携帯電話を使った位置情報の取得，カーナビなどによる経路情報の蓄積と利用，あるいは自動改札による入札・出札の記録といった形で，人の移動に関する履歴情報を情報技術システム上で蓄積することが可能な状況となっている[1)－3)]．移動履歴データは，単に移動の経路を示すものではなく，人々がその経路を辿って移動する理由までも含みうるものであるため，様々な観点から分析の対象となりうる．特に，個人の移動履歴に関するデータは，ビジネス上の分析においても重要なデータの一つとなる．例えば，特定の地域において，店舗に滞在する人や通過する人がどこから来て，どこへ行くのかがわかることのメリットは大きい．どの路線のどこの駅から来るのかということや，注目する店舗を素通りして競合店舗に行くという情報を有効に活用できれば，販促キャンペーンや広告配布を実施する際のエリアを検討する材料として事前に利用できる．

しかしながら，こうした個人の移動に関する履歴データにはプライバシーの問題があり，必ずしも十分な利活用が進んでいない現状がある．日々刻々の個人の正確な位置情報が取得され，かつ長時間にわたって記録される場合，その移動履歴データは個人にとって非常にセンシティブな情報であり，第三者への漏えいや流出が起こることにより，場合によってはプライバシーの侵害を引き起こす恐れがある．特に，個人の自宅・勤務地や知られたくない移動履歴を第三者により把握される危険性のある場合には，個人の移動履歴データの利活用に先立ち考慮が必要となる．

移動履歴データを様々な用途に利活用する場合，要件の一つとして，こうした移動履歴データから元の持ち主を特定されないことが挙げられる．匿名化は，そのためのアプローチの一つで，個人の移動履歴データから個人を特定可能な情報を削除し，データを加工することで，個人の特定を防ぐ手法である．もっとも単純な方法として，氏名や会員番号といった，直接的に個人を特定できるような情報を取り除くことによる匿名化が考えられるが，そうした情報を含まない場合でも，外部の情報と結合することにより個人が特定され，個人に関するセンシティブな情報が取得される状態が起こりえる．特に正確な位置情報を利用する場合，その場所にいた人物を指し示す，来店履歴や購買情報といった，外部の情報との結合により移動履歴データの持ち主が特定できてしまう場合が想定される．

そのため，安全な利活用を実現するためには，対象となる移動履歴データから「個人を特定されないこと」が一定の基準で保証できることが望ましい．本稿では，k-匿名化[4)]により，安全性が保証された移動履歴データの匿名化を行うことを考える．

k-匿名化とは，個人が特定されないようにプライバ

*キーワード：交通行動調査，交通行動分析
**非正員，博（情報理工学），株式会社三菱総合研究所 情報技術研究センター（東京都千代田区大手町二丁目3番6号，TEL:03-3277-0750，E-mail:kazutaka@mri.co.jp）
***非正員，博（情報学），日本電信電話株式会社 NTT情報流通プラットフォーム研究所
****非正員，博（情報理工学），日本電信電話株式会社 NTT情報流通プラットフォーム研究所
*****非正員，博（工学），株式会社三菱総合研究所（現・独立行政法人科学技術振興機構）

シーを考慮した匿名化データを作成する技術である．k-匿名化では，匿名化対象のデータに処理を施し，保護する対象となる属性情報の値の組み合わせが等しいレコードの件数を，匿名化したデータ中に少なくともｋ個含む状態を作る．特に，k-匿名化の手法としては，対象データに含まれる属性情報を，より粒度の粗い，あいまい化したデータに置き換える一般化（Global Recoding）と，k-匿名性を満たすことが困難な不適切な値やレコードを秘匿するための削除（Suppression）がよく用いられる．k-匿名化は，データマッチングによる個人特定リスクを回避できる基準として，広く認知されている．

本稿では，個人の移動履歴データについて，匿名性を満たしつつ，その利用範囲をビジネス上の分析などに広げることを検討する．本来の目的である移動履歴データの利活用の観点からは，ある程度匿名化した粒度の粗いデータであっても，有用であることが望ましい．移動履歴データの有用性に関する定量的な評価方法については，汎用性のある定式化は現状困難であるため，以下では事例として設定した例に対して匿名化をした後に，分析に耐えるデータが残るかどうかを定性的に検証する．例としては，移動履歴データを構成する位置情報を粒度の粗いメッシュに変換する場合と，移動履歴データから特定の位置情報（たとえば店舗など）を中心に前後の 3 点を取った場合のそれぞれについて検討する．

3. パーソントリップデータ

移動履歴データの匿名化を検討するにあたって，パーソントリップデータを対象データとして利用した．

パーソントリップデータ（以下 PT データと記す）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」[5] において研究目的での公開，利用がされている，大規模な移動履歴データである．国土交通省 東京都市圏交通計画協議会により首都圏在住・在勤者の日常的な移動傾向を把握するため，10 年ごとにアンケート調査により集計されている個人の移動経路情報を集積したパーソントリップ調査データを，東京大学空間情報科学研究センターにおいて，アンケート結果に基づく位置情報をクレンジングし，1 分単位で移動のポイントを内挿補間することで，個人の丸々1 日分の移動履歴をデータ化したものである．

PT データを構成する主要な属性とその概要を表 1に示す．PT データの詳細については，東京都市圏交通計画協議会の「東京都市圏パーソントリップデータ - PT データ利用の手引き」[6] に記載されている．

表 1 PTデータの構成

No	属性名	概要
1	パーソン ID	識別子
2	トリップ番号	目的を持った移動の単位
3	サブトリップ番号	トリップを細分した移動の単位
4	日時	データの日時． 日付は 1998/01/01 に集約
5	緯度	データの緯度
6	経度	データの経度
7	性別	男性・女性・不明
8	年齢	年代別に集計
9	住所コード	住所をコード化したもの
10	職業	職業をコード化したもの
11	移動の目的	通勤・通学など，目的をコード化したもの
12	交通手段	車・徒歩など，交通手段をコード化したもの

今回の検証では，1998 年に収集された 722,000 人分の PT データを対象に，15 分間隔で抽出したデータを分析対象として利用した．したがって，データの総件数は，722,000（人）×24（1 日: 1998/1/1 に集約）×4（15 分間隔）で，約 69,312,000 件である．

なお，PTデータはあらかじめ個人に関する属性情報などを丸めてあり，位置情報も厳密には正確なものではないため，基本的に個人を特定できる危険性は低い．本稿では，実データとしての規模と正確性から，移動履歴データ全般の匿名化を考える上でのテストデータとした．

4. メッシュを用いた移動履歴データの匿名化

緯度・経度により示される位置情報を一般化する手段としては，全体領域を一定の範囲で区切ったメッシュ[7] を用いる方法が一般的である．どのような属性情報を持つユーザがどの領域に分布しているかを見る場合などに用いられ，位置情報をそのまま使うのではなく，メッシュに変換することであいまい化して，個人の特定を防いでいる．

メッシュは通常1㎞四方や10㎞四方のものが用いられるが，さらに80㎞四方といったように，段階的にメッシュの大きさを大きくすることができる．また，逆にメッシュの中を細分化して，より細かなメッシュを定義することもできる．こうしたメッシュの大小を段階的に定義した情報を，一般化階層情報という．メッシュが大きいほど情報としての粒度が粗くなり，より多くの個別のレコードが含まれるため，個人の特定を防ぐことができると考えられる．

一方，移動履歴データの用途として，複数の位置情報を用いた移動シーケンスの分析が考えられる．たとえば，通勤時間帯における人の移動を分析する，あるいは

ある特定の時間帯に特定の地点にいる人の移動を分析するといった用途である．メッシュを用いる場合，移動シーケンスは元の点から点への流れではなく，メッシュからメッシュへの流れを表すものとなる．

こうした移動シーケンスをk-匿名化する場合，条件として同じメッシュからメッシュに移動するレコードが，少なくともk個あることが必要となる．領域を均等に分割したメッシュを用いると，人の流れが極端に少ない（最悪１人しか通らない）シーケンスが存在することで，全体としてメッシュの粒度を粗くする必要が生じる．

そこで，移動履歴データをメッシュによりk-匿名化する場合に，人の分布に適応したメッシュの一般化階層情報を構築することで，どのような匿名化が可能かを検証した．本稿では，これを適応型メッシュと呼ぶ．

PTデータから以下の手順にしたがって，注目する2点の位置情報を抽出し，適応型メッシュの構築と，移動シーケンスの匿名化と分析を行った．

1. ある時点でのユーザの分布（位置情報の分布）に対して，メッシュと一般化階層情報を定義
2. パーソントリップデータから，ある時間帯を起点とする，2地点の移動を示す位置情報を抽出して匿名化対象データを構築
3. 各点の位置情報を準識別子として匿名化を実施
4. k-匿名性を満たさないレコード数の評価

具体的には，次のような手順で実験を実行した．

まず，午前7時45分のPTデータの点データ（ユーザの位置情報）の分布に着目して，対象とする領域をメッシュに分割し，メッシュの一般化階層情報を構築した．構築の際には，メッシュ内での点データの分布状況を参照しながら，階層ごとに，同一階層内の各セルに同数の点データが属するように階層を設計し，二分木の形で構築した．結果として得られた適応型メッシュは，一般化階層情報として，0階層目として153,657個のメッシュのそれぞれに少なくとも1個の点が存在し，21階層目で全ての点が含まれるような1個のメッシュに一般化する，22階層の二分木からなる一般化階層情報を構築できた．適応型メッシュにおける点データ（位置情報）の分布状況を図 1に，構築した適応型メッシュの分割イメージを図 2に示す．

次に，対象とするPTデータ(全722,000レコード)から，午前8時の位置情報と午前9時の位置情報をそれぞれ切り出し，構築した適応型メッシュの定義にしたがって，各位置情報の緯度・経度をメッシュに変換し，表 2に示す構成をとるシーケンス化したPTデータを作成した．

さらに，シーケンス化したPTデータを対象に，適応型メッシュの一般化階層情報を使って， k-匿名化を行った．

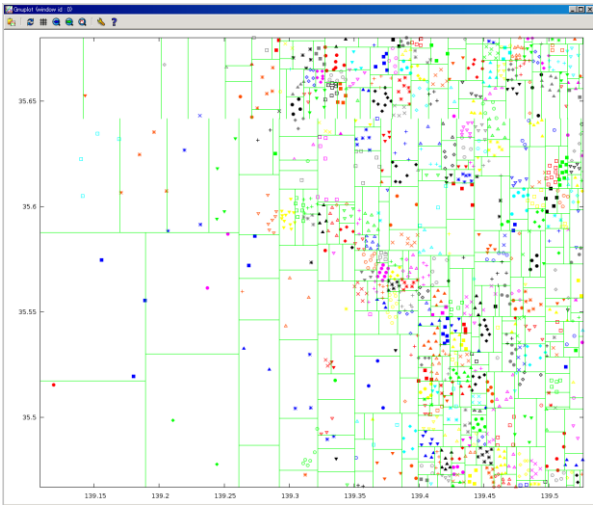

図 1　適応型メッシュにおける点データの分布状況

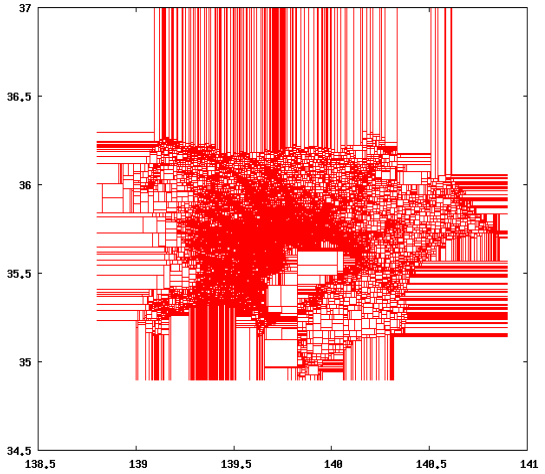

図 2　適応型メッシュの分割イメージ

表 2　シーケンス化したPTデータの構成

No	属性名	概要
1	パーソン ID	識別子
2	トリップ番号	目的を持った移動の単位
3	サブトリップ番号	トリップを細分した移動の単位
4	日付	1998/01/01 に集約
5	時刻１	午前 8 時
6	メッシュ１	メッシュに変換した緯度・経度
7	時刻２	午前 9 時
8	メッシュ２	メッシュに変換した緯度・経度
9	性別	男性・女性・不明
10	年齢	年代別に集計
11	住所コード	住所をコード化したもの
12	職業	職業をコード化したもの
13	移動の目的	通勤・通学など，目的をコード化したもの
14	交通手段	車・徒歩など，交通手段をコード化したもの

具体的には，時刻1，時刻2における位置情報（メッ

シュ1，メッシュ2）を段階的に一般化し，同じメッシュからメッシュに移動するシーケンスが2つ以上ある，すなわちk=2を満たす状態を探索した．また，一般化の過程で，k=2を満たさないシーケンスの数をカウントし，ユニークなレコードを削除することで，どの程度k=2を満たすために必要な，一般化の程度をさげることができるかを検証した．

表 3に検証結果を示す．2点の移動シーケンス722,000レコードを適応型メッシュにより匿名化した場合，22階層中19階層で，k=2が成り立つ，k-匿名化したデータを得られた．その粒度は，領域全体を8メッシュに分割した上で，各メッシュからメッシュへの移動シーケンスを表すことができる程度の粗いものとなった．

一方，匿名化の結果から，一般化してもk=2を満たさないような移動シーケンスの数をカウントし，ユニークなレコードを削除することによってどの程度一般化の程度をさげることができるかを評価したところ，全体の1%未満にあたる4,714レコードを削除することで，14階層（256メッシュ粒度）でのk-匿名化，10%程度の77,854レコードを削除することで，11階層（2,048メッシュ粒度）でのk-匿名化がそれぞれ可能であった．

表 3　一般化によるk-匿名化とメッシュ粒度

階層数	パターン数	ユニークレコード数	達成k値	メッシュ粒度
22	1	0	7,220,000	1
21	4	0	6,627	2
20	16	0	127	4
19	64	0	2	8
18	223	16	(2)	16
17	712	88	(2)	32
16	2,185	351	(2)	64
15	6,134	1,349	(2)	128
14	15,670	4,714	(2)	256
13	36,341	14,130	(2)	512
12	72,821	36,391	(2)	1,024
11	123,275	77,854	(2)	2,048
10	174,288	129,552	(2)	4,091
9	214,548	173,307	(2)	8,116
・・・	・・・	・・・	・・・	・・・
2	270,456	230,797	(2)	153,653
1	270,457	230,798	(2)	153,657

最終的に，1%未満のレコードを削除した256メッシュ粒度のk-匿名化データを対象として，移動履歴と個人属性（性別・年代・移動の目的）に関する簡単なデータマイニングを実行し，個人属性と移動シーケンスと間のルールを抽出した．抽出ルールとして，移動シーケンスと性別・年代・移動目的などの組み合わせを見ることで，エリアAからエリアBに移動する人は10代女性の通学目的が多い，エリアXからエリアYに移動する人は40代男性の通勤目的が多い，などの知見が得られた．

抽出したルールの例を図 3に示す．赤矢印は通勤目的の移動で，都心部に集中し，8時から9時の移動は流入する傾向が強い．青矢印は通学目的の移動で，分散する一方で，概ね一定の範囲内に集中し，地域単位での移動傾向が見られた．

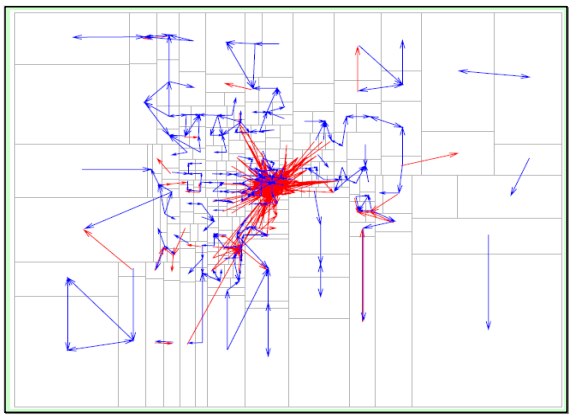

図 3　k-匿名化データからの抽出ルールの図示

本実験では，PTデータから抽出した2点の移動シーケンスを対象として，レコードの分布状況に応じて構築した適応型メッシュの一般化階層情報によるk-匿名化を試行した．結果としてレコード削除なしでは非常に粒度の粗い匿名化データとなるが，1%程度のレコード削除により一般化の粒度をさげることで，データマイニングを使って知見の抽出が見込める，十分に有用な粒度の匿名化データができることが確認できた．

5.　POI 情報を用いた移動履歴データの匿名化

マーケティングの分野では、位置情報に基づくプロモーション広告(Location-based Advertising: LBA)での利用が考えられる。LBAは、主にモバイル端末でのアプリケーションの操作や検索の際に、位置情報に基づいたプロモーション広告を含めて提示する仕組みである。パーソントリップデータは，位置情報の時系列データから構成されているが，POI情報等を付与することで特徴的な場所の移動シーケンス（たとえば，「A駅→デパートB→C駅」など）として再構成することができる．こうした再構成データは，LBAを利用する際の量的な予測などに役立つと考えられる．

検討方針として，ある地点に滞在・通過する人がどのような地点から来て，どのような地点に出て行くかに注

目する．具体的には，移動経路から3点（始点-中心点-終点）の移動シーケンスを抽出して匿名化した上でも有用性が残ることを検証する．

本稿では，位置情報をPOI（Point of Interest）情報に変換した上で，k-匿名性を満たす匿名化を行った．具体的には，移動履歴データから注目する3点を含む情報を取得し，匿名化と分析を以下の手順で行った．

1. 移動履歴データから注目地点(POI)を経由した人のデータを抽出
2. POIに対して，一般化階層を定義
3. 匿名化（始点，中心点，終点 の組合せをk-匿名化の対象として指定）を実施

まず，データの抽出については，データの件数が多く，主要な立ち寄り場所と考えられる銀座を分析地域とした．銀座の中で，特に多くの人が経由している地点をPOIとして定義しデータを作成した．図 4は銀座におけるPOIを通過した移動シーケンスの件数（シーケンス長，N=1）の上位を示す．

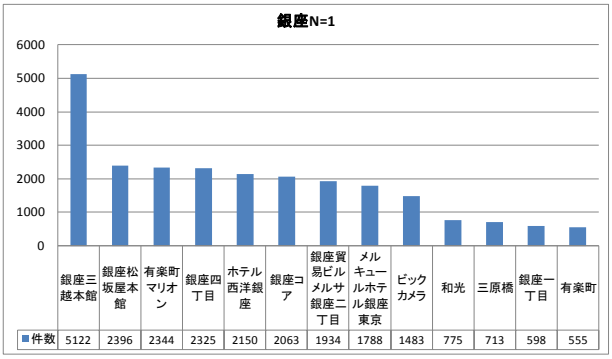

図 4　銀座N=1の件数

抽出した移動シーケンスの性質として，シーケンス長N=1の場合，k=1となる件数は0件である．これはPOIを選ぶ基準として，多くの人々が滞留する地点を選んだためであり，あらかじめサプレッションを行ったことと同様の影響があったと考えられる．N=2,3についてはk=1の件数がそれぞれ69件,331件だが，これらをサプレッションすることで，POI情報の一般化を行わずにk=2以上を満足させることができると考えられる．

次に，注目地点(POI)に対して図 5に示すような一般化階層の定義を行った．この階層定義に準じてk-匿名化を実施した．以下ではN=3，始点-中間点-終点という経路情報についての結果と分析を示す．

例えば，3点の中心(銀座三越本館)を軸に人の流れを解析する際には図 6のように匿名化することができる．この時，サプレッションを294件分行っている．これは，始点-中間点-終点の組み合わせでk<Kになるデータがあると，いずれかがすべて一段階一般化されるというグローバルレコーディングの特徴に対処するためである．

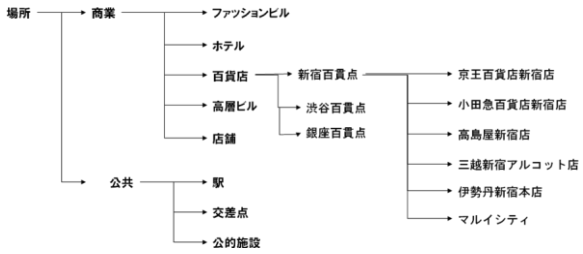

図 5　POIに対する一般化階層情報の定義例

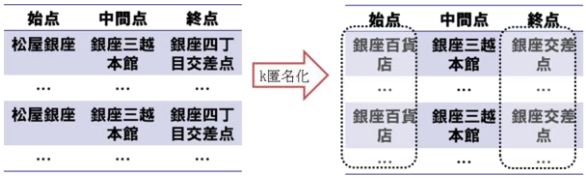

図 6　3点のk-匿名化例

図 6のデータは，銀座三越がLBAを配信する際に，重視するエリアの検討にも応用されうる．また，このデータにより，既存顧客の行動や既存顧客が滞在するエリアに加え，他店舗に行く人が多いエリアなどの情報も同様にして入手できることを示唆している．これらのデータはLBAを実施する際の定量的な事前検討データとして有用であると考えられる．また，LBAの効果を測る指標としても利用できる．これは，特定のエリアの特定の対象にターゲット広告を配布した結果，純粋に店舗に訪れるお客さんがどれくらい増えたか，滞在する人と通過するだけの人との割合の増減からコンバージェンスレートがどれくらい改善されたかを求めるなどの活用が考えられる．

6.　関連研究

移動履歴データの匿名化に関する研究は，日本だけでなく，欧米においても盛んである．特にEUでは，位置情報の匿名化とプライバシー保護に関する研究を具体的に行っているGeoPKDD（Geographic Privacy-aware Knowledge Discovery and Delivery），MODAP（Mobility, data mining, and privacy）というプロジェクトにおいて，位置情報サービスにおける匿名化（Private Queries），移動履歴データベースにおける匿名化（Trajectory Anonymization）を含む研究が進められている．

近年の研究として，移動シーケンスを構成する各位置情報を一般化するのではなく，複数の位置情報をグループとして平準化することで，k-匿名化を行う手法が幾つか提案されている．Abulら[8]は，同じ時間軸上で，一定の範囲δの中に，類似したk個の移動シーケンスを含

むような“不確かな”軌跡情報（匿名チューブ）を作ることにより，k-匿名化する手法を提案している．Nergizら[9]は，空間的に近いノード（位置情報）同士を関連付け，k個のノードを含むような時空間的な領域を作ることにより，k-匿名化する手法を提案している．

こうした複数の移動シーケンスのグループ化によるk-匿名化は，比較的アルゴリズムの実現が容易で，見た目上のk-匿名性が担保される一方で，データの正確性が位置情報や移動シーケンス間の距離の計算方法に依存し，データの性質によっては，元の移動履歴データから正確な情報が失われる場合が考えられる．匿名化データの有用性に関する定量的な評価方法については，汎用性のある定式化は困難なため，これらの匿名化手法との有用性や安全性の比較は今後の検討課題の一つである．

7. おわりに

本稿では，人の移動履歴の利用ニーズに応える上で必要となる，匿名性の検討を行った．移動履歴データとして，パーソントリップデータを用い，匿名化の方針としては，一般化と削除によるk-匿名化を採用した．一般化によるk-匿名化は，匿名化後のデータの抽象化レベルを均一にすることができ，分析上も扱いやすいデータを生成することが可能である．しかし，もとの移動履歴データに対して単純にk-匿名化を行うと，有用な情報はほとんど残らなくなってしまう．そのため，メッシュとPOIを用いたk-匿名化を検討した．

メッシュを用いた移動履歴データの匿名化では，レコードの分布状況に応じて構築した適応型メッシュの一般化階層情報を用いることで，1%程度のレコード削除でk-匿名化した移動履歴データを得ることができ，より細かい粒度での匿名化データの利用が可能であることを示した．

また，POI情報に着目した移動履歴データの匿名化では，注目する地点を中心とした人の流れを一定の匿名性を担保しつつ把握することができた．これは，LBAを行う際の基本データとして利用されうる．

今後，マーケティングの分野では，位置ターゲティングの様々な手法が検討されることが予想される．その際に，本稿で述べたような匿名性に関する検討を伴うことが期待される．

謝辞

本研究は、東京大学空間情報科学研究センターとの共同研究『個人情報の匿名化とその 2 次利用について（情報大航海プロジェクト）』による成果であり，研究用空間データをご提供頂いたことに謝意を表する．

参考文献

1）朝倉康夫，羽藤英二，大藤武彦，田名部淳：「PHSによる位置情報を用いた交通行動調査手法」，土木学会論文集 No.653/IV-48 pp.95-104，2000.

2）目黒浩一郎，佐藤賢：「行動分析調査ツールとしてのGPS携帯電話の可能性」，土木計画学研究講演集 Vol.34，2006.

3）目黒浩一郎：「土木計画におけるプローブデータの可能性」，第35回土木計画学研究発表会，2007.

4）Sweeney，L.：k-anonymity: a model for protecting privacy，International Journal on Uncertainty，Fuzziness and Knowledge-based Systems，10(5)，pp.557–570，2002.

5）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，http://pflow.csis.u-tokyo.ac.jp

6）東京都市圏交通計画協議会：「東京都市圏パーソントリップデータ － PTデータ利用の手引き」，http://www.tokyo-pt.jp/data/file/tebiki.pdf

7）環境省，基準地域メッシュ，http://www.biodic.go.jp/kiso/col_mesh.html

8）Abul，O.，Bonchi，F.，and Nanni，M：Never walk alone: uncertainty for anonymity in moving objects databases，ICDE2008，2008.

9）Nergiz，M. E.，Atzori，M.，and Saygin，Y.：Towards trajectory anonymization: a generalization-based approach，ACM SPRINGL'08，2008.