人口分布に着目した履歴匿名化手法の検討

疋田 敏朗[1]　山口 利恵[1]

概要：IoT 機器の位置情報の履歴から活用可能な移動履歴データを生成することを目的に，近年の ICT 技術の発展に応じて，スマートフォンのアプリケーションや IoT デバイスの移動履歴を利用して個人が特定されにくい個人情報保護法への対応を行った移動履歴データの生成交通統計データを生成することを検討することが必要であることを示す
そのうえで従来の履歴ごとにエリアを拡大しながら匿名化を行う方法ではなく，四分木を利用してグリッドごとの人口分布のデータを生成したうえで．エリア内の人口が一定以上になるようにエリア範囲を調節することで，人口の多いエリアでは小エリアに分割，人口の少ないエリアでは大エリアに分割することでよりアプリケーションで利用し易い匿名化手法を提案する．
さらに提案した方式に関して，東京大学空間情報科学研究センターの「人の流れプロジェクト」の「2008 年東京圏人の流れデータセット（空間配分版）」のデータを利用して実験を行い，提案手法を採用することで棄却率とメッシュサイズのバランスを考慮した加工が可能であることを示し，従来手法と異なり 30-40%程度の履歴は非対称な履歴を生成することを示す．
提案方式は確実な匿名加工方式ではないものの人口に着目することで従来手法に比べ，より解析ニーズに近いデータを生成できることを示す．

キーワード：CSS2016，PWS，人口分布，移動履歴，匿名加工

Anonymization methods using grid population

Toshiro HIKITA[1]　Rie Shigetomi Yamaguchi[1]

Abstract: Growing mobile networks and widely spread of Global Positing System (GPS) devices enables to collect large scale location and trajectory data. Using trajectory data to succeed, privacy and data characteristics are essential. Most anonymization methods are losing characteristic for application service, like adding noise to trajectories. Or it is hard to find anonymization pair of trajectories.
In this paper, firstly we present adaptive quadtree grid population calculation to determine grid size of trajectories.
Our anonymization method dynamically adjust cluster size to maintain trajectory data characteristics, small mesh to the dense populated area, large mesh to sparse populated area, base on quadtree geospatial data structure.
Proposed method is satisfied that adaptive size scaling and more efficient to maintain characteristic. Our experiment suggest that proposed method correctly anonymize Tokyo Urban flow data of 1.3M Trajectories.

Keywords: CSS2016, PWS, Population distribution, Trajetoriy, Anonymization

1. はじめに

近年，GPS を始めとする測位デバイスが携帯機器の機能の一部として搭載されるようになり，さらに携帯網の発展により携帯機器が測位した位置情報をセンター上のサーバに送信し，蓄積することが現実的になっている．さらにスマートフォンとそのアプリケーションや各種 IoT デバイスの普及によって，一般のユーザから大量に位置情報を収

[1] 東京大学情報理工学系研究科
The University of Tokyo

集・蓄積することができるようになっており，蓄積した位置情報を活用することで従来は困難であった新たなサービスが次々と生まれるようになっている．

また鉄道乗車券の IC カードが進展し，これらの乗車券の利用履歴を用いることで IoT デバイスの移動履歴と同様に利用者の移動履歴を得ることも難しくなくなった．

これらの各種 IoT デバイスや IC カードから得られる移動履歴情報を活用すれば移動履歴の推定ならびに予測を高精度に行うことができると考えられる．

しかしながら，IoT デバイスや IC カードから発生する移動履歴情報はその内容に位置情報やその移動履歴など，個人を特定しうる情報やセンシティブな情報が含まれることからそのまま利活用することはできない．

すでに，鉄道乗車に関する IC カードの移動履歴が事業者間で提供されたり [1]，個人の移動履歴に関するプライバシ保護に関する懸念が示されるなどの事例も発生している．それらの懸念に対応するためには履歴を匿名化することによって対応を行うことが考えられる．すでに携帯事業者では位置情報の履歴を匿名化してプライバシ保護を行いつつ統計を行う [2] という事業が始まっている．

交通履歴データに関しても同様に移動履歴はプライバシを保護し，個人を特定しないまま移動履歴を利用するための変換手法が望まれている．我々は都市開発においては最も必要と考えられる移動者の出発地と目的地の履歴に着目して匿名化手法を検討 [3] した．しかしながら既存の提案手法は出発地と目的地のグリッドサイズを同一とみなすという欠点があった．

本論文では，まず 2 章でまず位置情報と移動履歴に関する匿名化に関して紹介を行う、その上で従来の移動履歴情報の匿名化に関する研究はデータ提供者側のプライバシ保護に着目しており，移動履歴に対してノイズの形でダミーの移動履歴を加えるなどの加工を施すなど，データ利用の観点が存在しないという問題があることを示す．さらに加工方式の移動履歴の匿名化手法について説明をしたうえで，都心間での移動では問題がなくとも，移動履歴の多くが該当する郊外から都心部への移動に関して効果的ではないことを示す．

次に 3 章ではグリッドサイズの分割手法として Quadtree(四分木) の説明を行ったうえで，実データによる Quadtree グリッドの分布について論じる．

そのうえで提案手法として実際のデータを元に人口分布からグリッドサイズを規定し，移動履歴データの匿名化を行う手法の説明を行う．

さらに 4 章では実際の移動履歴のデータを元に Quadtree による人口分布に合わせたグリッドサイズを利用した匿名化を行い，従来手法よりも効果的に匿名化できていることを示す．

2. 従来の研究

本章では一般的な履歴の匿名化の説明と，従来の匿名化手法に関して論じる

2.1 履歴データの匿名化

まず一般的な履歴データの匿名化について説明を行う．

個人を直接的かつ一意的に識別する属性，たとえば氏名[*1]，個人番号[*2] などを示し，これを **個体識別属性** と呼ぶ

個人を一意に識別できないとしても複数の属性を組み合わせると個人を一意的に識別できるものもある．たとえば性別，生年月日，住所などが該当する．これらの属性を**疑似識別属性** (Quasi Identifier，以下 **QID**) と呼ぶ．

あるデータ T から個人が特定できないようなデータ T' を生成する変換作業を匿名化と呼ぶ．匿名化の手法としては $k-$ 匿名化 [4] が有名である．$k-$ 匿名化は概念で同一の擬似識別属性に対して，最低でも $n \geq k$ のデータが存在するように，擬似識別属性を曖昧化する．例えば氏名情報のみを削除し，会員番号のみを利用する方法は一般的には仮名化と呼ばれる．仮名化を行っても個体識別属性が残っていると一意に識別できるため，仮名化は厳密な意味での匿名化ではない [5]．また，$k-$ 匿名化の情報では，匿名化として不十分として，データの種別を定量的に計る手法 l-diversity [6] や データの全体の割合傾向を計る手法 t-clossness といった手法も提案されている [7].

2.2 個人情報保護法と個人の特定

2016 年 8 月現在施行されている我が国の個人情報保護法においては，個人情報とは個人が特定できるような情報のほかに，『(他の情報と容易に照合することができ、それにより特定の個人を識別することができることとなるものを含む。)』という形で他の情報と照合することで個人が特定できる情報もまた個人情報であるとされている．ここで識別とはそれが誰だかわからないが特有の 1 名に分離できるということであり，特定とはそれが固有の 1 名を示すこととされる，

ここであるデータ T が存在した場合にそのリストの全項目を QID として $k-$ 匿名化を行ったデータ T' については，1 つの情報について少なくとも 2 つ以上の列が該当することから一意に特定ができないことが知られている．すなわちデータ T の全項目を QID としてに $k-$ 匿名化による変換を行うことができればその情報について個人を特定することはできないとすることができる．

改正個人情報保護法による匿名加工に関しては，本稿執

[*1] 厳密には氏名だけでは同姓同名の個人が複数存在する可能性があるが，社会通念では個体識別属性とみなされている

[*2] 各個人に一意に割り当てられている番号，例えば日本でいえばマイナンバー，米国でいえばソーシャルセキュリティーナンバー．

筆時点で決着を見ていない．特定できてはいけないという記述と経産省のガイドラインにある識別子の削除，準識別子の k-匿名化を行うことで，履歴情報に関しては識別できることは問題ないという見解が存在しており，こちらに関しては今後の議論を待つこととしたいが，2.3 節に示すように移動履歴や位置情報に関しては LBS の進展に伴い非常に容易に照合可能であることから匿名加工にあたっては今後配慮が必要になるものと考えられる．

2.3 移動履歴の匿名化のために必要な要件

昨今，Swarm や Facebook などの SNS への投稿は位置情報を付加することが可能であったり，地点情報を付加してチェックインすることができるため，別の手段で収集された履歴情報や位置情報とデータ処理位行う履歴情報を照合することで個人を特定することも難しくはなくなりつつある．そのため個々の履歴情報に関して，他の情報を用いた場合でも個人を一意特定しうる状態ではないことが必要ということになる．

本研究では上記の事情を鑑みた上で移動履歴の長さが個人特定性にどのような影響をあたえるのかを検討することとする．移動履歴から個人が特定されるケースを列挙すると以下のようなケースが考えられる．

(1) **その情報自体が個人を特定できる情報を含む場合**
(2) **位置情報自体が自宅などを指し示す場合**
(3) **位置情報と時刻の組み合わせにより個人が特定される場合**
(4) **履歴が固有のために個人が特定される場合**

今回はこのうち 4 について検討をすることとする．これは組み合わせ情報が固有であるために個人が特定できることである．履歴情報は同じ ID の履歴をリンクさせることで個人が特定可能であると指摘をされている．

そこで実際の移動履歴を元にどの程度の履歴長であれば，個人が特定をされるのかについて実際のデータを元に検討することとする．

2.4 従来の移動履歴の匿名化の研究について

次に匿名化の位置情報への拡張について述べる．位置情報について $k-$ 匿名化を行った例 [8] は 2003 年に Gruteser らによって報告されている．この例では地点をグリッドごとに区切り，それぞれの地点情報をもとに $k-$ 匿名化が行われている．Gkoulalas-Divanis らによるまとめ [9] によれば，$k-$ 匿名化の手法は一般的に今あるデータを中心とした区切り方と地形情報を活用したグリッドベースの区切り方の 2 種類に分けることができると主張している．

$k-$ 匿名化の他の匿名化手法としてはノイズを混入するという手法が挙げられる．文献 [10] [11] では実際の位置情報の他に複数のダミーの位置情報を挿入させることでデータ自体の匿名性を担保する手法について記述されている．またダミーデータの混入手法についてはより高度な手法が提案されている，Niu ら提案 [12] によればダミーデータの配置場所を統計的に検討することで，ダミーユーザの現実的な配置が可能になり，より強固な配置が可能になるとされる．

移動履歴に関してもダミーデータを加えて匿名化するという手法が提案 [13] されている．この手法はランダムにダミーデータを加えた移動履歴情報を生成することで，リアルユーザのデータを秘匿化する．しかしながらダミーを利用する方法では受領した位置情報にダミー情報がかなりの確率で紛れ込むため位置情報の利用者側から見るとデータが使いにくいという問題が発生する．例えば実際の情報に 4 倍のダミーデータを混入した場合，位置情報を的中させることが出来る確率を 20%近くに低下させることができるが，利用者から見ると 1/5 でしか正確なデータが存在しないということになる．これは特にビッグデータ処理を前提とした場合にデータ自体の信頼性がなくなること言うことを意味しているため，データの利用目的によってはこの手法は使えない．

また移動履歴をグリッド化して加工を行うことにより，$k-$ 匿名化する方法はいくつか提案されている山口 [14] の手法では単一のグリッドで $k-$ 匿名化を実施するという手法が提案されており，著者ら [3] は可変グリッドを利用した単体移動履歴の匿名化を提案している．

従来の加工方式による匿名化は出発地と目的地を同一条件で加工を行うものであった．この手法では [渋谷] → [新宿] という都心間での移動に関してはうまく動作をするが，移動履歴の頻度に差がある場合は匿名化の効率が落ちることが明らかになっている．郊外の国分寺市で 4km 四方というのはそれほど大きなグリッドではないが，新宿-渋谷感が全て含まれる 4km のグリッドは匿名化データの利用に対して非常に大きな制約を与えることになる．

図 1 匿名化を行いたいエリアは異なる

そのため，図 1 のように地域によって異なるエリアサイズの匿名化ができるような手法が必要とされている．

3. 人口比に応じたジオコード手法の提案

本章では，ジオコーディングとしての四分木による領域分割について説明を行ったうえで，実際のデータを用いた四分木グリッドの人口分布について説明を行う．更にこの人口分布データを用いた匿名化手法の提案を行う．

3.1 領域分割手法と四分木 (Quadtree)

本節では地理空間情報で利用される領域分割手法を紹介しながら，本稿で利用する四分木に関する説明を行う．

位置座標は X,Y または緯度経度で表される二次元空間の座標であるがこの位置座標を計算機で扱いやすい情報に変換する必要がある．このような変換はジオコーディングと呼ばれており，階層化や符号化の方法によって複数の手法が使われている．

階層化手法で有名な符号化方法が Geohash[15] であり位置情報を符号化するジオコーディングの一種であり，階層構造を持ちつつも位置座標を空間分割する機能を持つ．

また米軍で採用されている MRGS と呼ばれる手法もあり，この手法は UTM 手法として国土地理院のコーディングでも採用されている．MRGS は全球をメルカトル図法で分割しており，単なる XY の分割よりも制度が高い特徴があるが XY を分離して保有するため，階層化する際の計算量が増えるという難点がある．

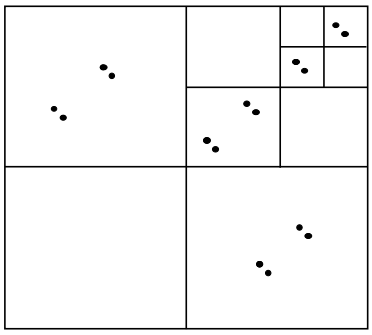

図 2 四分木 (Quadtree)

Quadtree(四分木) は二分木のような木構造による格納手法であり，2 分割する二分木と異なり，図 2 のように枝を 0,1,2,3 の四分割して格納する．木構造の四分木は非平衡木であるが位置情報を取り扱う場合には先頭からエンコードができることにメリットが有る．

今回はこの四分木の手法を位置情報の符号化に用いて整理したものを利用し，この符号化を Quadtree と記すことにする

3.2 実データを用いた四分木グリッドの人口分布

実際の移動履歴データを利用して，Quadtree の手法を適用した場合にグリッドあたりの分布がどのようになるのかを検討することにする．

3.2.1 使用データについて

匿名化実験のデータとしては人為的に作成したデータと実データの二種類が考えられる．完全に人為的に生成したデータはアルゴリズムの確認ならびに匿名化のベンチマークとしては利用可能であるが，現実的なデータとして利用するにはデータの特性が異なるため難しい．一方で実データの利用には本論文で議論したように ID と位置情報が個人を特定する情報になるため，自ら同意を取得してデータを収集するか，またはユーザの同意を得て第三者提供を受ける必要がある．現時点ではこのような実データの利用には困難が伴うと言える．

今回提案手法の実験を実施するためには現実的な移動履歴を持ち，人数が多く，なおかつユーザからの同意または個人情報保護上問題がないデータを利用する必要がある．そこで上記の条件を満たすデータとして，東京大学空間情報科学研究センターの「人の流れプロジェクト」[16] の「2008 年東京都市圏　人の流れデータセット（空間配分版)」のデータを利用することとした．

今回の実験データは東京都市圏交通計画協議会が収集したパーソントリップ調査によるデータを元にしている．実験データとしては関しては元のパーソントリップ調査のデータを用いて，住所詳細を記載していないものをベースとし，以下に示す空間配分を行った空間配分版とした．空間配分とはゾーンごとにまとめられた地点情報について，個々人の位置情報をゾーン範囲内の建物の分布に合わせて詳細位置に確率的に再配分し，現実のデータに近づける処理のことである．

表 1 に，人の流れプロジェクトのデータにおける位置履歴情報定義を利用した位置履歴情報の例を引用する．このデータ定義については実際のものだが，データ自体に関しては定義に合わせて著者が作成したダミーデータとなっている．

この例では 20-25 歳の学生でかつ女性であるユーザ 12345 は，東京大学構内から徒歩で本郷三丁目駅に移動し，本郷三丁目駅から新宿駅まで移動をした後に，新宿駅から高尾山口駅まで鉄道で移動し，最後に高尾山山頂まで徒歩で移動をした，この移動の目的はレジャーであったことがわかる．

今回利用したデータセットには本データセットには特定日付の 586882 ユーザの 1316100 の移動履歴が含まれている．今回はこの移動履歴を利用することにした．

3.2.2 Quadtree による人口分布の計算

まずはメッシュごとの人口計算を行う．今回はトリップの発着地点を双方用いることにする．すなわち発着地点ごとに人口を+1 しているので 1 名は 2 箇所に換算される．同一地点を往復した場合は 4 地点

表 1 人の流れプロジェクトのデータ例 (データはダミーデータ)

ID	番号	サブ	日時	経度	緯度	性別	年齢	職業	目的	手段
12345	1	1	2014/12/10 10:05	139.7619	35.7143	2	4	13	99	1
12345	1	1	2014/12/10 10:20	139.7605	35.7075	2	4	13	99	1
12345	1	2	2014/12/10 10:20	139.7605	35.7075	2	4	13	99	12
12345	1	2	2014/12/10 10:40	139.7001	35.6909	2	4	13	99	12
12345	1	2	2014/12/10 12:00	139.2696	35.6321	2	4	13	99	12
12345	1	3	2014/12/10 12:00	139.2696	35.6321	2	4	13	99	1
12345	1	3	2014/12/10 13:30	139.2436	35.6251	2	4	13	99	1

に対して+1 される．この Quadtree メッシュを 2 文字ごと (1/16 ごと) に計算を行う形で人口を算出した [313200312132223031] であれば，[313200312132223031], [3132003121322230] [31320031213222], [313200312132], [3132003121], [31320031] のエリアごとの人口を算出している

ここから抜き出したデータをトリップごとに整理を行う．たとえば「[本郷三丁目, 高尾山口]」を元に整理する．この際にこれ以下であれば領域を拡大し，これ以上であれば領域サイズをそのままにするスレショルド値の決定が必要である．この値を仮に 250 としてこの先の議論をすすめる．

まずは [本郷三丁目] エリアのデータを使って整理する．人口が多い箇所の場合は 18 文字の Quadtree 内の人口が 400 を超えることも稀ではない．そのため本郷地区については眺めの文字例すなわち，狭めのエリアで検討が可能である．

一方，高尾山口のようなエリアには人口がそれほどいない，そのため 250 というスレショルド値を満たすためには領域を拡大する必要がある．先ほど Quadtree メッシュごとの人口を作成したので，quadtree エリアの人口をプログラムに問い合わせることで適切な領域がわかる．

本郷エリアと高尾エリアのサイズが決まったところで，この経路ペアをシステムに登録する．システムでは発着エリアとも同エリア・同サイズであった場合に経路が重複しているとみなす．すなわち履歴において個人特定性があるとは，その履歴が全体の中で単独で存在することであり，同じ履歴が複数存在すれば個人特定性はないこととなる．

今回は移動履歴を表 2 のように管理する．この例の場合，トリップ数は 5 であり，経路数は [東京] → [新宿]，[新宿] → [渋谷]，[代々木] → [渋谷] の 3，棄却経路は [代々木] → [渋谷] の 1 となる．この棄却経路の履歴が個人特定性のある履歴ということになる．

また棄却率=個人特定率を [棄却経路数/総トリップ数] と定義し，この例の場合は 1/5 = 20% となる

このときに [代々木] → [渋谷] は区間としては [新宿] → [渋谷]　に内包されるが，今回はこのような区間の内包は計上せず別件として数えることとした．

実際にデータを適用して計算した分布が図 3 である．実

表 2 移動履歴と一意性

ID	乗車駅	降車駅
1	東京	新宿
2	新宿	渋谷
3	新宿	渋谷
4	東京	新宿
5	代々木	渋谷

データでは Quadtree の長さ 8 で最大 22 万のグリッドが存在したがプロットは 2500 で人口が 2500 で打ち切っている．

実際の計算では最大 5000 まで変動させたので，それ以上の領域も利用している

表 3 Quadtree の人口分布

QT 長	エリア数	倍率
8	10	-
10	50	5.0
12	491	9.82
14	5866	11.9
16	48543	8.28
18	247389	5.10

表 3 に Quadtree の長さと関東圏のデータで分割されたデータ数を示す．このひょうにおける分割数はデータがないところに関してはカウントされないのでデータが存在しているグリッドがいくつあるのか？ということを示している．

2 文字シフトでの理論上のグリッド増加量は 4x4 で 16 倍であるから，データの偏りによりデータの増加量が制限されていることが見て取れる．

3.3 人口比に応じた移動履歴の加工手法

本節では前説までの結果を踏まえて，人口比を利用したエンコード手法の提案を行う．図 4 に今回の提案手法を説明する．

(1) 全データからメッシュごとの人口を計算する
(2) 人口スレショルド値の決定
(3) 履歴を取り出し，発着点の人口メッシュを検索する

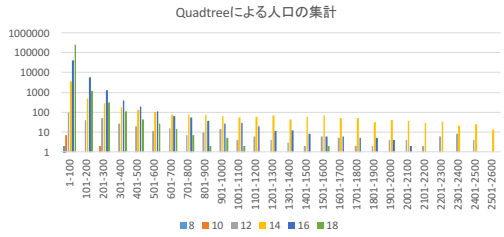

図 3 Quadtree を用いたグリッドの人口分布

(4) 発着点のグリッドサイズを調整して匿名加工を終了する

まず 1 のように，全データからメッシュごとの人口を計算する．今回は一日分のデータをすべて利用した．

次に 2 の人口スレショルド値を決める．今回は図 3 を参考に 100,250,500,1000,2500,5000 というスレショルド値を用意した．

ここからデータ処理に入る．3 のように履歴から発着点のみを抜き出す．その上で発地点，着地点ごとにスレショルド値よりも多い人口が存在する Quadtree メッシュを 4 発着地点のメッシュとして利用する．このようにして履歴データを加工していく．

図 4 Quadtree による人口分布適用法の概要

注意が必要なのは本提案手法で加工された情報の匿名性は保証されないということである．すなわち繁華街と広大な領域との移動であったとしてもその履歴が単独でしか存在しない一意な履歴であればこれは特定可能ということになる．

従来の手法では $k-$ 匿名化を利用していたので加工後のデータは必ず k よりも多いことが保証されていたが本件手法ではその保証はない．そのため加工処理後に必要に応じてスクリーニングを行う必要がある．実際の利用時に関してはこのような特定が可能な履歴をどのようにするかの検討が必要であり，必要に応じて履歴を削除することが求められる．

ただ，本稿執筆時に経済産業省から出されているガイドラインでは必ずしも履歴に関して，一意な履歴を削除しなければならないという立場のようであり，一意な履歴に関する取り扱いに関しては今後の検討が必要であると考えられる．

4. 実移動データを用いた移動履歴データと個人特定性の検討

本章では 3 章で示した提案手法をデータに適用した際にどのような加工結果が得られるのかを検討する．

まず提案手法を実際のデータに適用した場合に提案手法では人口スレショルド値の変化に対してどのような結果を出すのかを検討する．さらに実際に利用可能な移動履歴の加工情報とは何かについての議論を行う．

4.1 提案手法の実データへの適用結果

本説では提案手法を実データに適用した際の結果について説明をする．

実験データは東京圏の人の流れプロジェクトデータであり，トリップ数は述べたとおりに約 58 万人，1316100 トリップとなっている．このデータに対して，人口スレショルドを 100,250,500,1000,2500,5000 と適用し，提案手法による加工の効果を見ることとした．

表 4 提案手法の実験結果

エリア人口	棄却経路数	経路数	総 Trip	棄却率 (%)
100	761510	921918	1316100	57.9
250	417266	556062	1316100	31.7
500	282201	412932	1316100	21.4
1000	170561	293122	1316100	13.0
2500	24202	76701	1316100	1.83
5000	3698	22904	1316100	0.28

本件実験のプログラムは MacOS X 上の PC に python で実装し，実験を行った．

表 4 と図 5 に提案手法での実験結果を示す．メッシュのエリア人口を増加させると経路数が減少することがわかる．最低限の人口を増やすことで経路の多様性は減少する．その代わりに棄却される経路数，すなわち一意な経路 ($k = 1$ の経路数) も減少する．小さなメッシュではメッシュ間の接続に特徴があり，識別されるまたは特定の可能性を残る経路が生成されてしまうことがわかった．エリア人口を増加させると棄却される経路は減少していき，エリア人口を 1000 人にすると棄却率は 13%程度，2500 人で 2%弱でありこれくらいが実際に利用する際には適正な値になると思われる．エリア人口を増加させていくと経路の総数が減っ

てしまい，各ユーザのらしさを観測することができなくなる．大きなエリアで問題ないアプリケーションであれば構わないが，ユーザー属性に関して出来るだけの情報を収集したい場合には大きなメッシュによる経路数の現象に関してはどのように取り扱うかの検討が必要であろう．

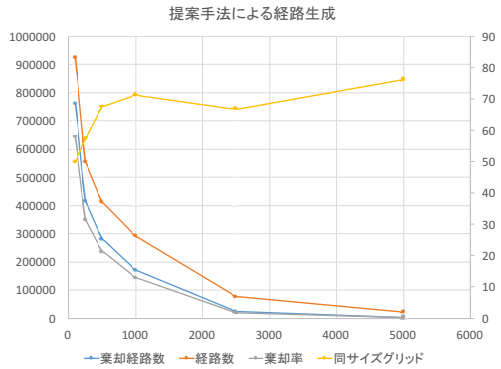

図 5　提案手法の実験結果

本件提案手法は原理的に捨てる経路がないので，$k-$ 匿名化とは異なり，総トリップ数は一定となる．また同一サイズのトリップ数を比較したところ，50-70%のトリップは同一サイズとなっていたこれにより 40-30%程度のトリップに関しては従来手法と異なり，非対称なメッシュサイズのトリップを生成することができていた．

4.2　生成した経路の真の棄却率または利用可能性について

本説では生成した経路の利用可能性について検討を行う．前節で説明したように棄却率についてはエリア人口を増加させることで，メッシュサイズを拡大させることができるので $k = 1$ となってしまう率を減らすことはできる．しかしながらメッシュを拡大させてしまうと発着地点が同一メッシュになってしまう可能性も増える．

発着が同一メッシュになってしまった場合は移動があったことは確認できるが，移動距離を算出したりすることはできない．この同一メッシュになる確率を図 6 に示す．実験結果からはスレショルドが 100 程度の時には 5%程度であった同一メッシュが 5000 人規模まで拡大すると 40%程度に拡大するということがわかった．

棄却率と同一メッシュで利用が制限される状況を合わせて考えると一番厳しい棄却率 (真の棄却率) を算出することができる．同一メッシュ発着の $k = 1$ 履歴は存在するので 2 つの数値をそのまま計算することはできず，実験結果から拾い上げることが必要である．

結果を図 6 に示す．真の棄却率は単調増加または単調減少は取らない，エリア人口 1000 人程度で最小値を取ることがわかった．この結果から実用においては 1 グリッドあたり 1000 人程度を目安にするのが良いと考えられる．

ただし，同一メッシュ内の発着は使えない情報かどうかはアプリケーションに依存する．渋谷区内の施設において渋谷区民が来訪している情報が使えないどころか，そのような近隣住民の情報は収集していることから考えると真の棄却率に関してはかなり厳しい仮定で計算していることになっている．

4.3　エリア人口と平均メッシュサイズについて

エリア人口と発着点のメッシュサイズに関しても検討を行った．今回の Quadtree では標準的なメッシュサイズは最小で 18 であり，8 が最大になる．8 では関東圏を 8 分割，10 では 50 分割になり，18 では 247389 分割である．

それぞれのエリア人口の時の平均サイズを図 6 に示す．平均サイズは 100 人で 15.5 から 5000 人で 12.0 である．5000 人にしてもサイズはそれほど減少していない事がわかる．

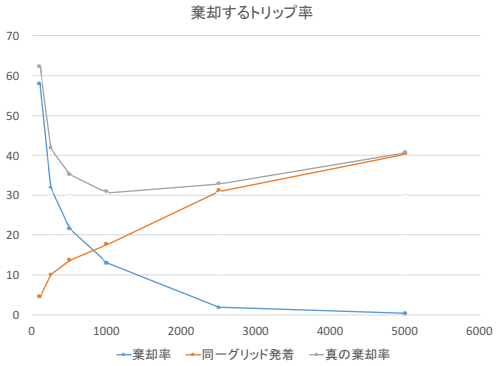

図 6　提案手法で生成した経路の棄却率

5.　おわりに

本論文では IoT 機器の位置情報の履歴から活用可能な移動履歴データを生成することを目的に，近年の ICT 技術の発展に応じて，スマートフォンのアプリケーションや IoT デバイスの移動履歴を利用して個人が特定されにくい個人情報保護法への対応を行った移動履歴データの生成交通統計データを生成することを検討することが必要であることを示した．

そのうえで従来の履歴ごとにエリアを拡大しながら匿名化を行う方法ではなく，四分木を利用してグリッドごとの人口分布のデータを生成したうえで．エリア内の人口が一定以上になるようにエリア範囲を調節することで，人口の多いエリアでは小エリアに分割，人口の少ないエリアでは大エリアに分割することでよりアプリケーションで利用し易い匿名化手法を提案した．

さらに提案した方式に関して，東京大学空間情報科学研究センターの「人の流れプロジェクト」の「2008 年東京圏人の流れデータセット（空間配分版）」のデータを利用して実験を行い，提案手法を採用することで棄却率とメッ

シュサイズのバランスを考慮した加工が可能であることを示し，従来手法と異なり 30-40%程度の履歴は非対称な離席を生成できることを示した．

提案方式は確実な匿名加工方式ではないものの人口に着目することで従来手法に比べ，より解析ニーズに近いデータを生成できる．

今後は本論文の成果を活かしてより実用的かつ高効率な移動履歴の匿名化に関する研究を進めていきたい．

謝辞 なお，本研究は科研費 (16K12548) の助成を受けたものである．また東京大学空間情報科学研究センターの「人の流れプロジェクト」との共同研究であり，データの整備並びに提供を行っていただいた空間情報科学研究センター各位に感謝する．

参考文献

[1] ”Suica に関するデータの社外への提供についての有識者会議”：“Suica に関するデータの社外への提供について” (2014).

[2] 寺田雅之：“モバイル空間統計：携帯電話ネットワークを活用した人口推計技術とその応用 (ビッグデータ特別セッション)”, pp. 63–66 (2014).

[3] 疋田敏朗, 山口利恵：“階層化符号表現を利用した移動履歴の匿名化手法”, マルチメディア、分散、協調とモバイル (DICOMO2015) シンポジウム 2015 情報処理学会 (2015).

[4] L. Sweeney: “k-anonymity: a model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, **10**, 5, pp. 557–570 (2002).

[5] 板倉陽一郎 伊藤孝一 菊池浩明 高木浩光 高橋克巳 中川裕志 疋田敏朗廣田啓一 山口利恵：“「完全な匿名化」幻想を超えて”, 暗号と情報セキュリティシンポジウム 2014 電子情報通信学会 (2014).

[6] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam: “l-diversity: Privacy beyond k-anonymity”, ACM Transactions on Knowledge Discovery from Data (TKDD), **1**, 1, p. 3 (2007).

[7] D. Rebollo-Monedero, J. Forné and J. Domingo-Ferrer: “From t-closeness to PRAM and noise addition via information theory”, Privacy in Statistical DatabasesSpringer, pp. 100–112 (2008).

[8] M. Gruteser and D. Grunwald: “Anonymous usage of location-based services through spatial and temporal cloaking”, Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, MobiSys ’03, New York, NY, USA, ACM, pp. 31–42 (2003).

[9] A. Gkoulalas-Divanis, P. Kalnis and V. S. Verykios: “Providing k-anonymity in location based services”, SIGKDD Explor. Newsl., **12**, 1, pp. 3–10 (2010).

[10] H. Lu, C. S. Jensen and M. L. Yiu: “Pad: privacy-area aware, dummy-based location privacy in mobile services”, Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile AccessACM, pp. 16–23 (2008).

[11] H. Kido, Y. Yanagisawa and T. Satoh: “An anonymous communication technique using dummies for location-based services”, Pervasive Services, 2005. ICPS’05. Proceedings. International Conference onIEEE, pp. 88–97 (2005).

[12] B. Niu, Q. Li, X. Zhu, G. Cao and H. Li: “Achieving k-anonymity in privacy-aware location-based services”, Proc. IEEE INFOCOM (2014).

[13] P. Shankar, V. Ganapathy and L. Iftode: “Privately querying location-based services with sybilquery”, Proceedings of the 11th international conference on Ubiquitous computingACM, pp. 31–40 (2009).

[14] R. S. Yamaguchi, K. Hirota, K. Hamada, K. Takahashi, K. Matsuzaki, J. Sakuma and Y. Shirai: “Applicability of existing anonymization methods to large location history data in urban travel”, Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference onIEEE, pp. 997–1004 (2012).

[15] G. Niemeyer: “Geohash”, http://geohash.org/.

[16] Y. Sekimoto, R. Shibasaki, H. Kanasugi, T. Usui and Y. Shimazaki: “Pflow: Reconstructing people flow recycling large-scale social survey data”, IEEE Pervasive Computing, **10**, 4, pp. 0027–35 (2011).