

# 移動実態に即した履歴情報のグループ化手法

疋田 敏朗<sup>1</sup> 山口 利恵<sup>1</sup>

**概要:** 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の普及によって, 一般のユーザから大量に位置情報を収集・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きている。しかしながら, 位置情報を活用した従来の地理空間情報処理の方法は位置座標を用いた二次元のグリッド符号化を利用していたため, 目的の街区が分割されてしまったり, ふたつの異なる街区を融合してしまうこと, 鉄道の乗車履歴の場合は各駅ごとに履歴を整理するため単独となる履歴が多数発生し, 破棄しなければならない履歴が多数発生するなどの課題があった。

本論文では従来の位置座標を用いた二次元のグリッド符号化ではなく, 移動実態を解析し, ヒューリスティックにより鉄道路線ごとに符号化する方法, そして統計的な手法と機械学習を用いて自動的にエリアを区切る方法の2つの方法を提案した。

さらに提案した縮約方式のうちヒューリスティックな方法に関して, 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一意な履歴数を4.9%から0.05%に大幅に減らせることを確認した。

## Adaptive Encoding of Commute Trajectory

Toshiro HIKITA<sup>1</sup> Rie Shigetomi Yamaguchi<sup>1</sup>

### 1. はじめに

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の普及によって, 一般のユーザから大量に位置情報を収集・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きている。

これらの位置履歴を用いたサービスにより, 全国各地の日々の渋滞情報を瞬時に把握したり, 都市内での人の動きを理解したり, 震災時の人々の移動経路の記録や自動車の通行可能道路を可視化することが可能となっている。

また, ユーザの履歴を蓄積することでよりユーザの好みにあった情報の提供や地点の推薦を行うシステムも提案されており, GPS や携帯網の普及により, 各種デバイスから得られる移動履歴情報により新サービスを生み出し, 経済を活性化させることが期待されている。しかしながら, 移動履歴は個人にとってセンシティブな情報であるため, プライバシーを保護し, 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。GPS などから得られる移動履歴情報の活用が期待されているが, 移動履歴にはプライバシーの考慮が必要なため, 個人を特定しないまま移動履歴を利活用するための変換手法が望まれている。

### 2. 履歴データの匿名化に関する従来の研究

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

<sup>1</sup> 東京大学 大学院 情報理工学系研究科  
University of Tokyo

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 [1] が有名である.  $k$ -匿名化は概念で同一の疑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, 疑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 [2]. 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 [3] や データの全体の割合傾向を計る手法  $t$ -closeness といった手法も提案されている [4].

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 [5] 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 [6] 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 [7] [8] 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 [9] 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 [10] 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  $k$ -匿名化する方法はいくつか提案されている山口 [11] 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 [12] 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.

一方で乗車のような履歴の匿名化については菊池ら [13] が数学的モデルにより, 鉄道駅の乗降客数データの分布から類推した移動履歴の匿名性に関する検討を行っている.

著者らは GeoHash のような階層化符号方式を活用した匿名化手法 [12] を提案している. 現在利用している匿名化手法は, グリッド手法である GeoHash をベースに, 出発地  $(X_o, Y_o)$  と目的地  $(X_d, Y_d)$  の四次元を一次元に縮約して移動履歴を計算している. しかしながら, この方法では位置座標の  $X, Y$  により区分けを行うので, 人の流動状況と匿名化の区分けが上手に対応せず綺麗なエリア分けができないという問題が存在した.

### 3. 地理空間情報の符号化処理の課題

本来, 匿名化なりデータの縮約は図 3 のように [元データ] の特性を解析・理解を行い, その [特性] に合わせた手法を用いて, [匿名化] を行うのが正しい方法である.



図 3 Processing Outline

しかしながら, 一般的には図 4 のように移動軽度を区切るグリッドコーディングまたはある地点の半径でまとめる空間コーディングと呼ばれる方法で区切ることが一般的な手法となってしまう.

グリッドコーディングは緯度経度という一般的な方法を  $X, Y$  座標の値だけを用いて計算できるため非常に幅広く利用されている. 地理空間情報の処理に関しては一般的な処理はグリッドコーディングであると言っても過言ではない.

しかしながらグリッドコーディングを実際に適用すると様々な問題に直面することになる. 図 1 に幾つかの例を示

\*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.



図 1 Problems of Grid Coding



図 2 Optimal Coding

す。例えば東京駅の東西が同じグリッドに収まっており、丸の内側と八重洲側という人の移動があまりなく性格が違う領域を区切りことができず同一視するしかなくなる。また、海（または隅田川の河口）を挟んで反対側の芝浦と月島が同一のグリッドの扱いになっているという問題がある。

さらにはグリッドの境界が実際の人流の境界と一致しな

い場合があり、霞が関や六本木のように領域が分割されてしまい求めている解析を行うことができないという事例が起こることもある。

本来は図2のように街や人の生活圏ごとに分割することが望ましい。

しかし、このような妥当な分割には一般的に困難が伴う。

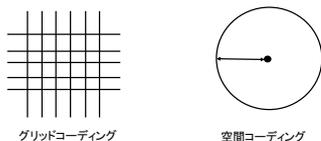


図 4 Grid or Area Coding

特に地理情報の場合ながら妥当な分割かということを含め理解して入力しておくことは大変困難である。

#### 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

本章では著者らが [14] において検討をした移動履歴と個人特定性の関係について概要を説明する。

##### 4.1 実験用の移動履歴データの生成

移動データのうち交通手段が鉄道であるものを抜き出した。

抜き出したデータを各ユーザごとに整理を行う。ここで駅数を履歴に関与する駅の数、履歴超を移動区間の数と定義する。[本郷三丁目],[高尾山口],[渋谷] の 3 駅が履歴にある場合は移動履歴における駅数は 3 であり、履歴長は 2 となる。

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。

移動履歴を表 1 のように管理する。この例の場合、総数は 5 であり、パターン数は [東京] → [新宿], [新宿] → [渋谷], [代々木] → [渋谷] の 3、一意な履歴は [代々木] → [渋谷] の 1 となる。この一意の履歴が個人特定性のある履歴ということになる。個人特定率を [一意数/総数] と定義し、この例の場合は  $1/5 = 20\%$  となる。[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。

表 1 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  |

#### 4.2 移動履歴データの検討 (東京地区)

そのうえで鉄道を利用した移動について 1 日のデータを履歴長とその数で整理すると表 2 に示す。

履歴の総数はユーザー数にして 188275 ユーザ、各ユーザごとの 1 日の履歴を記録しているため、2 区間の履歴が最も多く、83%の履歴が 2 区間の利用履歴となっている。2 区間の内訳を見てみると表??の様になる。146321 件 94%の履歴が自宅最寄り駅→出先最寄り駅→自宅最寄り駅という経路である。

表 2 移動履歴数と履歴長 (東京地区)

| 履歴長 | 履歴数    | 一意履歴数 | 個人特定率 |
|-----|--------|-------|-------|
| 1   | 8000   | 5679  | 71.0  |
| 2   | 154997 | 64837 | 41.8  |
| 3   | 17396  | 16766 | 96.4  |
| 4   | 6257   | 6185  | 98.8  |
| 5   | 1132   | 1132  | 100   |
| 6   | 370    | 370   | 100   |
| 7   | 79     | 79    | 100   |
| 8   | 37     | 37    | 100   |
| 9   | 5      | 5     | 100   |
| 10  | 1      | 1     | 100   |
| 11  | 1      | 1     | 100   |

この場合に一意である履歴は 95092 件であり全履歴の 50.5%が他の履歴と重複しない一意なものであることがわかる。また 2 区間の履歴を除くと個人特定率は高まる 3 区間以上に限定すれば特定率は 96%以上となる。

##### 4.3 移動履歴データの履歴と個人特定性について

大規模な鉄道の移動履歴を見る限り、過半数の履歴は乗降駅レベルで一意である、東京地区と大阪地区という 2 つの大都市圏のデータで調査した結果は同様の傾向を示した。ある程度鉄道網が発達した大都市圏 (実質的には東京圏と大阪圏、名古屋圏) においては同じような傾向を示すものと考えられる。

購買や移動に関してはユーザの志向を理解するためにもある程度の履歴長が履歴長が必要であると言われている。しかしながら 18 万人のデータでなおかつ履歴長が 2 であるデータであっても半数のデータは重複しない可能性が高く、履歴長が 3 を超えた場合にはほとんどの場合に履歴が一意であることをこの実験データは示している。

一意性を持たない履歴は全体の履歴の 80%以上である単なる往復履歴であり、移動履歴として価値が高いと思われる 3 区間以上の履歴に関しては一意性があることを実験結果は示している。

このように移動履歴の一意性が極めて高いことは問題である。そのため何らかの方法で履歴を縮約する形で一意性を下げる努力を行う必要がある。

## 5. 手法提案

今までの検討のように鉄道の移動履歴について何らかの方法で履歴を縮約するで一意性を下げることができれば匿名化なり履歴の有効利用が可能になると考えられる。

首都圏の場合はその移動の多くが鉄道を利用していることから、鉄道の路線を前提として、路線ごとに駅をグループ化していくことがより移動実態にあった符号化を可能にすると考えられる。例えば、小田急線沿線で考えると川崎市、町田市、相模原市と区分けして符号化するよりは、小田急線沿線を駅ごとに区切って符号化するほうが移動の実態にあった符号化ができると考えられる。

### 5.1 ヒューリスティックによる類型化

これらの分割を行う方法としてまず考えられるのが発見的的手法により分割を行うヒューリスティックな手法である。これは解析者が予め保有する事前知識をプログラムに入力することで類型化を可能にする。

図5に示すように鉄道の特性を活かした形で縮約することが考えられる。

前述したように各駅停車の駅の利用客は少なく、これらの駅を利用した履歴が小さくなるのが実験により確認されている。これらの各駅停車の駅をひとまとめにグループ化することで人の流動をそのままにより精度の高い匿名化が可能であるとかんがえられる。

例えば小田急線沿線の場合であれば新百合ヶ丘から先の柿生、鶴川、玉川学園前、町田、相模大野の各駅について一括して町田と類型化を行うことができる。

このような類型化についてはある程度の鉄道の知識は必要となるが、人為的な入力のほかに小田急線であれば「次の急行停車駅に類型化する」のようなある程度の規則を元に機械的に区分けをするという方法も考えられる。

### 5.2 機械学習による類型化

上記のようなヒューリスティックな手法は人間の気づきを必要とし、データ処理のコーディング負荷が大きいという問題があり、さらにデータが想定と異なる分布であった場合や解析者の思い込みが強く影響する、解析者がその土地について詳しく理解していないと適切な類型化ができないという課題がある。

特に鉄道移動履歴において日本全国において先の例のような類型化を行うには鉄道マニア同様の知識を必要とする上、日本以外の諸外国では適用することができない。鉄道以外に自動車や自転車での移動の類型化を行う際にはそれぞれ道路網に関する知識が必要となるなど、ヒューリスティック手法は実用的に利用するには高い課題があるのもまた事実である。

そこである程度多数の移動履歴が存在することを前提

に、統計的にエリアを区切ったり、利用駅を統計的に算出することで機械学習によりエリアを生成する方法を検討する。

まずは駅ごとの乗降者数でコアとなる駅を決定する。コア駅の前後の駅についてコア駅と同様の目的地の分布になる駅の集合を類型化していく。この方法を取ることで似たような傾向の駅を集積することができる。

現在提案中の方法では山手線のような円周方向への移動がある場合は問題がないが、小田急線のような放射状に移動している場合にはたぶん問題が発生するものと考えられる。なぜならば利用客は基本的に路線を放射状に移動する上に、急行停車駅から各駅停車に乗り換える形で乗客数が減衰する方向に作用してしまうからである。

たとえば新百合ヶ丘は百合ヶ丘までの駅の傾向のほうが多く、柿生以遠とは傾向が異なる可能性が高い。実際に新百合ヶ丘から先は多摩線に乗り入れる列車が存在することを考えると乗客の傾向は異なると考えられる。このような線区ごとの特性を事前知識なく、機械学習または統計的な処理のみによって処理することが必要であり、今後の検討が待たれるところである。

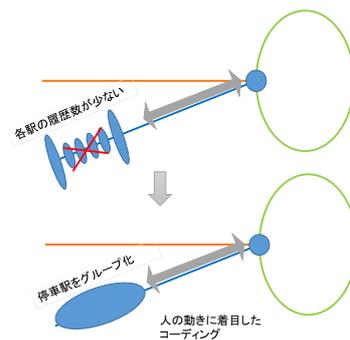


図5 Proposed Method

## 6. 実験

### 6.1 使用データについて

前章までに提案した手法と従来手法で匿名化に差があるかどうかを検討するために人の流れプロジェクトのデータを用いて実験を行うことにした

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っており、このようなデータの

流通がないことが匿名化に関して検証を難しくしている理由と考えられる。

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」[15]の「2008年東京都市圏 人の流れデータセット（空間配分版）」（以下、東京地区）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（東京地区）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3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ついて、東京地区データセットには特定日付の576806ユーザ分の移動履歴が含まれている。この移動履歴のうち、鉄道乗車券のデータを念頭におき、一日のうち一回でも鉄道を利用した188257人の履歴を用いることにした。

## 6.2 実験手法

まず今回の実験は、まず移動履歴を履歴長1の着発データとして抽出したうえで、目的の路線沿線を発着する履歴のみを抽出し、その上で該当線区内の履歴をまとめることで結果が得られるかどうかを確認することとした。

今回はJR東日本山手線と小田急小田原線の交互発着のデータを用いることとした。鉄道利用であれば移動経路は問わないこととした。具体的には渋谷～東京や下北沢～町田は抽出対象であり、渋谷～町田も対象とするが、新百合ヶ丘～大手町に関しては対象とはしなかった。

履歴長1の着発データの総数は404005トリップであり、このうち小田急線と山手線の中で発着していたデータは15808トリップとなる。

これらの発着データをもとに元の発着データで集計したもの（以下 Standard）と、区間ごとに集約したもの（以下 Summarized）の二種類について比較を行うこととした。Summarizedの区間については山手線は6区間に、小田急小田原線は快速急行停車駅を中心に9区間に分割をした。新宿駅に関しては同一液とし、山手線側の区間に包含することとした。

表4 山手線の分割

| 駅名   | 類型化した駅名 |
|------|---------|
| 大崎   | 品川      |
| 五反田  |         |
| 目黒   | 渋谷      |
| 恵比寿  |         |
| 渋谷   |         |
| 原宿   |         |
| 代々木  | 新宿      |
| 新宿   |         |
| 新大久保 |         |
| 高田馬場 |         |
| 目白   | 池袋      |
| 池袋   |         |
| 大塚   |         |
| 巣鴨   |         |
| 駒込   |         |
| 田端   | 上野      |
| 西日暮里 |         |
| 日暮里  |         |
| 鶯谷   |         |
| 上野   |         |
| 御徒町  |         |
| 秋葉原  | 東京      |
| 神田   |         |
| 東京   |         |
| 有楽町  |         |
| 新橋   |         |
| 浜松町  |         |
| 田町   | 品川      |
| 品川   |         |

また発着データについては発駅と着駅の入替えは考慮していないので新宿～東京と東京～新宿は別のデータであることに注意が必要である。

Standardは合計75駅の交互発着データであり、Summarizedは集約により15区間の交互発着データとなっている。

75駅間の交互発着データは理論的には $75 \times 74 = 5550$ 通りとなるが、Standardは合計で2602通りの発着となっている。15区間の交互発着データは210通りであるがSummarizedは205通りの発着データが存在する。

表 3 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  |

### 6.3 実験結果

これらのデータについて、発着データごとの利用者数と発着履歴の数をグラフ化したもの図6である。これはX軸が利用者数、Y軸が該当する履歴の数になる。K=2のK-匿名化を意識するのであれば、利用者数が1である履歴を対象外としなければならない。X軸、Y軸ともに対数表記であることに注意してほしい。Summarizeすることによって、明らかに利用者数が少ない履歴の数を減らすことができている。

表 5 提案手法の効果

| 履歴長              | 履歴数   | 一意履歴数 | 個人特定率 |
|------------------|-------|-------|-------|
| Standard (K=1)   | 15808 | 785   | 4.9%  |
| Summarized (K=1) | 15808 | 8     | 0.05% |
| Standard (K=2)   | 15808 | 1226  | 7.7%  |
| Summarized (K=2) | 15808 | 12    | 0.07% |

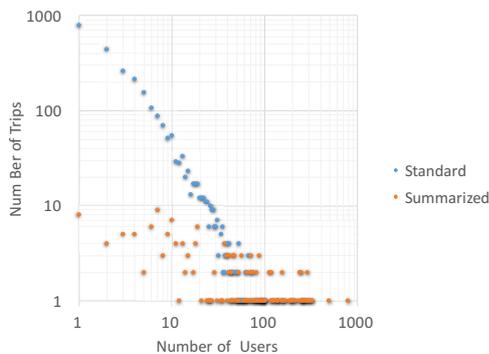


図 6 提案手法の効果

### 6.4 考察

本実験では、人の流れデータを利用して、履歴長1の区間データを生成することで提案手法の有効性を確認した。4章の実験との違いは、往復をそれぞれ1の履歴にするように履歴長を強制的に1にしていること、そのため重複が増えて一意性が低下している。

また路線を山手線と小田急線に限定したことで履歴の重複が増加しており、標準的な手法においても重複は全体の

7%となっている。これは都心部と放射路線の区間のみに限定了ことによって発生しているものであり、全体としてこのような傾向を示すというものではない。

そのような状況においても最頻値は利用者数が1であり、次に多い値は利用者数が2である履歴であることは明らかである。一方今回提案の縮約を行った場合は700を超えるような最頻値ではなくなり、他の利用者数のデータを同様に値になることが確認できたことは非常に良い発見である。

これは適切な区間で利用駅を縮約することで従来よりも遥かに効率的な匿名化が可能であることを示している。

発見的な手法で工数がかかるとい問題はあものものの提案しているような手法を用いることで従来よりも遥かに有用な匿名化が可能になることを示唆している。

## 7. まとめ

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の普及によって、一般のユーザから大量に位置情報を収集・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。しかしながら、位置情報を活用した従来の地理空間情報処理の方法は位置座標を用いた二次元のグリッド符号化を利用していたため、目的の街区が分割されてしまったり、ふたつの異なる街区を融合してしまうこと。鉄道の乗車履歴の場合は各駅ごとに履歴を整理するため単独となる履歴が多数発生し、破棄しなければならない履歴が多数発生するなどの課題があった。

本論文では従来の位置座標を用いた二次元のグリッド符号化ではなく、移動実態を解析し、ヒューリスティックにより鉄道路線ごとに符号化する方法、そして統計的な手法と機械学習を用いて自動的にエリアを区切る方法の2つの方法を提案した。

さらに提案した縮約方式のうちヒューリスティックな方法に関して、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一意な履歴数を4.9%から0.05%に大幅に減らせることを確認した。

ヒューリスティックな手法は、鉄道路線に対する知識ならびにプログラムの負荷が極めて大きいことが明らかになっている。今後は学習により区分を自動的に行う手法に関して検討を進めたい。

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] L. Sweeney: “k-anonymity: a model for protecting privacy”, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **10**, 5, pp. 557–570 (2002).
- [2] 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).
- [3] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkatasubramanian: “l-diversity: Privacy beyond k-anonymity”, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **1**, 1, p. 3 (2007).
- [4] D. Rebollo-Monedero, J. Forné and J. Domingo-Ferrer: “From t-closeness to PRAM and noise addition via information theory”, *Privacy in Statistical Databases* Springer, pp. 100–112 (2008).
- [5] 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).
- [6] 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).
- [7] 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 Access* ACM, pp. 16–23 (2008).
- [8] 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 on IEEE*, pp. 88–97 (2005).
- [9] B. Niu, Q. Li, X. Zhu, G. Cao and H. Li: “Achieving k-anonymity in privacy-aware location-based services”, *Proc. IEEE INFOCOM* (2014).
- [10] P. Shankar, V. Ganapathy and L. Iftode: “Privately querying location-based services with sybilquery”, *Proceedings of the 11th international conference on Ubiquitous computing* ACM, pp. 31–40 (2009).
- [11] 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 on IEEE*, pp. 997–1004 (2012).
- [12] 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).
- [13] 菊池浩明, 高橋克巳: “乗降履歴データの安全な匿名化は可能か?”, 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).
- [14] 疋田敏朗, 山口利恵: “大都市圏での交通機関の利用履歴による匿名加工条件の検討”, 暗号と情報セキュリティシンポジウム 2016 電子情報通信学会 (2016).
- [15] 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).