Vol.2020-DPS-182 No.60
Vol.2020-CSEC-88 No.60
2020/3/13

背景知識に基づいて非等長の移動履歴から個人を特定するベイズモデルの提案と評価

松本 瞬[1,a)]　大岡 拓斗[1]　市野 将嗣[1]　緑川 耀一[2]　吉井 英樹[2]　吉浦 裕[1,b)]

概要：近年，個人の移動履歴が産業界や公共分野で有効活用されているが，移動履歴にはプライベートな情報が推定される危険性があるため，これを保護する必要がある．そこで，プライバシーを保護するうえで必要となる移動履歴の機微の度合いを明確化するために，移動履歴からの個人特定のリスクの明確化を行うことにした．先行研究の手法では，不特定多数者の移動履歴を背景知識として活用することを考慮していない，時間情報を有効活用できていない等の問題が存在した．本稿では，これらの問題点を解決した個人特定の手法を提案した．また，手法の有用性を明らかにするために，不特定多数の Twitter アカウント，および 53 人の被験者の Wi-Fi 移動履歴・Twitter アカウントを用いた評価を行った．

キーワード：プライバシー，個人情報，移動履歴，SNS

1. はじめに

近年，スマートフォン等から得られる個人の位置情報を時系列で記録した移動履歴が，商品や飲食店の広告の最適化，交通システムの改善，災害時の避難計画などの様々な分野で有効利用されている．一方，移動履歴には個人の自宅や通勤通学先などのプライベートな情報が推定される危険性があるため，プライバシーを保護しながら活用することが重要である．プライバシーを保護しながら移動履歴の活用を行うための技術・制度を検討するうえで，まず最初に移動履歴の機微の度合いを明確化するべきである．特に，直接的にプライバシーに関わることが考えられる，移動履歴からの個人特定に関するリスクを明確化する必要がある．

先行研究として，マルコフ連鎖モデルに基づいた個人特定の手法が提案されている [1,2]．これらの手法は，**図 1** のように，持ち主が未知である仮名化移動履歴と持ち主が既知である実名移動履歴の 2 つの集合が与えられたときに，仮名化移動履歴と実名移動履歴の紐づけを行うことで，仮名化前の人物を明らかにする手法となっている．この手法の個人特定の流れは以下の通りである．

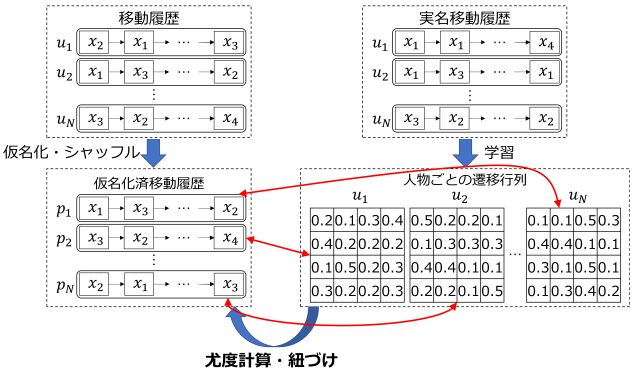

図 1　マルコフ連鎖モデルに基づいた個人特定の手法

(i) 実名移動履歴から，人物ごとの遷移行列を作成する．ここで，遷移行列とは当該人物が 2 地点間を移動する確率を各要素として持つ行列のことである．

(ii) 作成した遷移行列から各仮名化移動履歴の尤度を求める．これを遷移行列と仮名化移動履歴の全ての組合せについて行い，算出された尤度を類似度として用いることで，実名移動履歴と仮名化移動履歴の紐づけを行う．

しかし，これらの手法では

- 不特定多数者の移動履歴を背景知識として活用することを想定していない
- 全ての仮名化移動履歴が等長である（同じ個数の位置

[1] 電気通信大学大学院情報理工学研究科
Graduate School of Informatics and Engineering, The University of Electro-Communications
[2] ソフトバンク株式会社
〒 105-7317 東京都港区東新橋 1-9-1 東京汐留ビルディング, SoftBank Corp.,Tokyo Shiodome Bldg., 1-9-1, Higashi-shimbashi, Minato-ku, Tokyo 105-7303,Japan.
[a)] s.matsumoto@uec.ac.jp
[b)] yoshiura@uec.ac.jp

情報を含む）ことが前提になっている
- 移動履歴の取得対象地域に関するスケーラビリティが乏しい
- 時間に関する情報を有効活用できない

という問題点が存在する．そこで，本稿ではこれらの問題点を解決できる新たな手法を提案し，その有用性を明らかにする．

2. 関連研究

2.1 移動履歴からの個人特定の概要

Srivatsa らは，人物間の交流関係に基づいて，持ち主が同一人物である移動履歴と SNS アカウントの組合せを明らかにする手法を示した [3]．Srivatsa らは，移動履歴における人物同士の接触状況と，SNS における交友関係に類似性があることに注目した．この手法では，人物同士の接触状況・SNS における交友関係それぞれをグラフ化したうえで，グラフ理論に基づいた紐づけを行う．Ma らは，オリジナルの実名移動履歴とこれに対してノイズを加え仮名化を行った移動履歴が与えられたときに，仮名化前の人物を明らかにする手法を示した [4]．この手法では，移動履歴に加わるノイズの確率分布を仮定したうえで，2 つの移動履歴の類似度計算を行い，最尤推定によって仮名化前の人物を明らかにする．Riederer らは，異なる媒体から得られる 2 種類の移動履歴について，持ち主が同一人物である移動履歴の組合せを明らかにする手法を示した [5]．この手法では，各時刻・場所における訪問回数がポアソン分布に従うと仮定し，類似度の定式化を行った．この式を用いて全通りの移動履歴の組合せについて類似度を算出し，持ち主が同一人物である移動履歴の組合せを明らかにする．Gambs らは，実名移動履歴と仮名化移動履歴が与えられたときに，各々の移動履歴から人物ごとの遷移行列を作成し，これを用いて 2 つの移動履歴の紐づけを行う手法を提案した [6]．Shokri らは，位置情報に欠損があるような実名移動履歴から遷移行列を作成し，これを用いて仮名化移動履歴の仮名化前の人物を明らかにする手法を提案した [1]．Murakami は，位置情報の欠損が多数存在し，かつ位置情報数が少ない実名移動履歴を用いて仮名化前の人物を明らかにする手法を提案した [2]．また，Murakami はさらに位置情報数が少ない実名移動履歴から仮名化移動履歴の持ち主を特定する手法を示した [7]．この手法では，行列分解を用いることで，少ない実名移動履歴からより正確な人物ごとの訪問頻度の分布を作成している．その後，仮名化移動履歴の訪問頻度分布との JS ダイバージェンスを計算し，比較することで，仮名化前の人物を明らかにする．

2.2 類似研究

本節では，本研究の手法と類似したマルコフ連鎖モデルに基づく個人特定の手法に関する研究を示す．

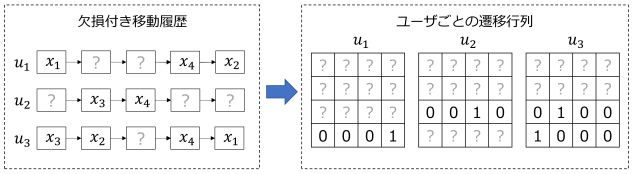

図 2 欠損を含んだ移動履歴からの遷移行列作成の具体例

Gambs らは，実名移動履歴・仮名化移動履歴それぞれから作成された遷移行列同士を比較することで個人特定を行う手法を示した [6]．この手法では，実名移動履歴と仮名化移動履歴それぞれについて，人物ごとの遷移行列を作成する．その後，複数の遷移行列間類似度指標を用いて類似度を計算し，遷移行列同士の紐づけを行う．これによって，仮名化移動履歴の持ち主を明らかにする．なおこの手法では，移動履歴の取得対象地域に含まれる地点について事前にクラスタリングを行い，似たような属性（例えば，家・職場など）を持つ地点が同じクラスタに属するようにしたうえで，その地点クラスタ間の移動に関する遷移行列を作成している．

Shokri らは，遷移行列を作成する際に用いる実名移動履歴に位置情報の欠損があることを考慮し，この欠損を補完しながら遷移行列を作成する手法を提案した [1]．この手法では，まず図 1 のように実名移動履歴から遷移行列を作成する．実名移動履歴は，位置情報サービスや追尾を通して得る必要があるが，これらの手段では完全な移動履歴を入手するのは難しく，位置情報の欠損が発生してしまう．位置情報の欠損が発生した場合，**図 2** のように，入手した移動履歴そのものから遷移行列を作成することができなくなる．そこで，この手法ではギブスサンプリングを用いることで，欠損していない位置情報を元に欠損した位置情報の補完を行いながら遷移行列の作成を行う．これにより，Gambs らの手法では実現できない位置情報の欠損の考慮が可能になる．

Murakami は，遷移行列を作成する際に用いる実名移動履歴について
- 位置情報の欠損がある
- 入手できる位置情報数が非常に少ない

ことを想定した際の遷移行列の作成方法を提案した [2]．この手法は EM アルゴリズムに基づいており，テンソル分解を合わせて用いることで，欠損部分の考慮及び人物間に共通した特徴の活用を実現している．これにより，Shokri らの手法よりも強力な補完が実現でき，上記 2 つの性質を持つ移動履歴からでも遷移行列を作成することができる．

2.3 類似研究の問題点

2.2 節で述べた手法には以下の問題点が存在する．

(P1) 不特定多数者の移動履歴を背景知識として活用できない

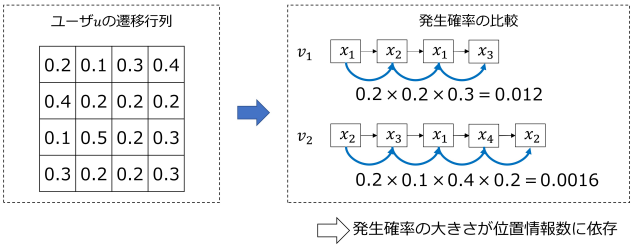

図 3 発生確率の比較の例

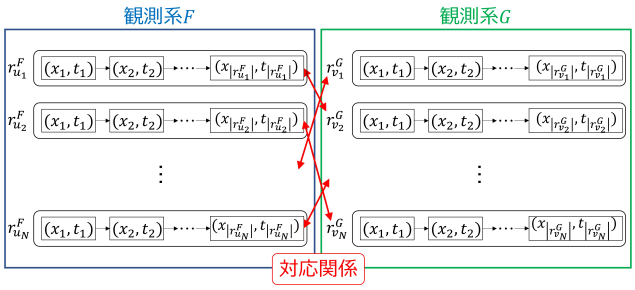

図 4 2 つの観測系の移動履歴

近年は，不特定多数の個人情報を匿名化状態，あるいは非匿名化状態で大量に入手できる．これらの情報を背景知識として用いることで，従来の精度を超えた精度の照合が可能になり得る．移動履歴に関しても SNS を通して大量に入手可能である．また，位置情報サービスを提供している企業や携帯会社，Wi-Fi 基地局設置企業であれば，より正確な移動履歴を大量に入手可能である．これらのデータを用いれば，従来の精度を超えた照合が可能となり得る．しかし，2.2 節で述べた手法では，このような不特定多数の移動履歴データの活用を想定した手法になっていない．

(P2) 全ての仮名化移動履歴が等長でなければならない
Shokri ら，Murakami の手法では，特定対象となる仮名化移動履歴の位置情報数が全人物について同一である必要がある．これは，**図 3** のように，遷移行列から各人物の移動履歴の発生確率を算出し比較する際に，位置情報数が同一でないと発生確率の比較を正しく行えないためである．現実において入手できる移動履歴がこのような性質を満たすことは考えにくい．

(P3) 対象地域に関するスケーラビリティが乏しい
2.2 節で述べた手法が実験で使用している移動履歴の取得対象地域は，サンフランシスコ，東京，北京等の 1 都市のみである場合が多い．仮に，取得対象地域を広くした場合，遷移行列のサイズが大きくなってしまう．その結果，次元の呪いによって正しく学習を行うことが困難になる．特に，2.2 節で述べた手法では人物ごとに遷移行列を作成しているため，使用できる位置情報が少なく，この問題が顕著に表れる．そのため，これらの手法は取得対象地域の広さに関して制約があると言える．

(P4) 時間に関する情報を有効活用できない
2.2 節で述べた手法では，移動履歴に含まれる情報のうち遷移のみに着目しているので，時間に関する情報については場所の前後関係しか利用できていない．そのため，時間に関する情報を十分に活用できているとは言えない．例えば，ある実名移動履歴と仮名化移動履歴が同一時刻に同一の場所を共有していた場合，その 2 つの移動履歴の持ち主は同一である可能性が高い．しかし，遷移のみに着目している場合，同一時刻・同一場所という情報は利用できない．仮に時間情報を扱おうとしたとき，遷移行列に対して時刻に関する次元を追加し，テンソルで表現することが考えられる．しかし，次元を拡張すると次元の呪いによって正しく学習できなくなってしまう．前項と同様に，この問題は人物ごとに作成するとき顕著に表れる．

3. 提案方式

3.1 記号の表記

図 4 のように，ある地理的環境 E において 2 つの観測系 F, G が存在し，それぞれの観測系で同一の N 人の移動履歴が観測される．観測系 F で観測される移動履歴の持ち主の集合を $\{u_1, \ldots, u_N\} \in U$ とし，観測系 G で観測される移動履歴の持ち主の集合を $\{v_1, \ldots, v_N\} \in V$ とする．本方式での個人特定とは，持ち主が同一人物である組合せ (u_i, v_j) が未知であるときに，観測された移動履歴を元にこの組合せを推定し，2 つの観測系における人物間の対応関係 $\sigma : U \longrightarrow V$ を明らかにすることである．

移動履歴について，詳細に定義を行う．移動履歴は，複数の位置情報から構成されている．位置情報は観測地点・時刻からなるタプル (x, t) で表現する．ここで，観測地点は緯度・経度で表現する．移動履歴は人物ごとに存在し，観測系 F における人物 u の移動履歴を r_u^F と表記する．すなわち，観測系 F における人物 u の移動履歴は $r_u^F = \{(x_1, t_1), (x_2, t_2), \ldots\}$ と表現される．また，移動履歴に含まれる位置情報の件数は $\left|r_u^F\right|$ のような形で表記し，この値は人物 u ごとに異なる．さらに，観測系 F で観測された全人物分の移動履歴の集合を $R^F = \left\{r_u^F | \forall u \in U\right\}$ と表記する．

3.2 基本アイデア

本方式は，「観測系 F における人物 u と観測系 G における人物 v が同一人物であれば，人物 u の移動履歴と人物 v の移動履歴を結合してできる移動履歴は地理的環境 E において自然な移動履歴になる」という仮定に基づいて構築

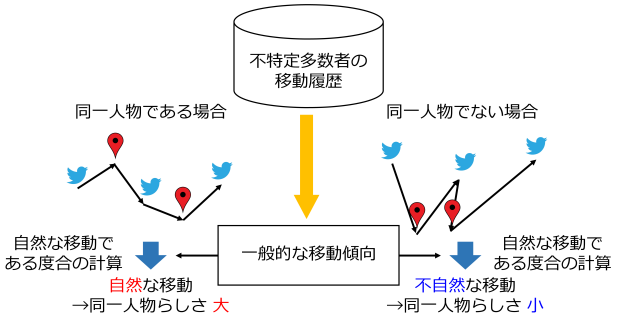

図 5 提案方式の基本アイデア

する．学習ステップでは，地理的環境 E における一般的な移動傾向を学習する．このステップでの目的は，人物ごとの移動傾向を学習することではなく，一般的な移動傾向を学習することであるので，不特定多数者の移動履歴を学習データとして使用できる．そのため，2.3 節における (P1) の問題を解決できる．類似度計算ステップでは，**図 5** のように，学習ステップで得られた一般的な移動傾向を用いて，結合移動履歴がどれだけ自然な移動履歴であるかを計算する．この値が大きければ大きいほど，同一人物らしさが高いとみなす．これを人物の組合せ全通りについて計算し比較することで，2 つの観測系における人物の対応関係を明らかにする．

3.3 理論

1. 観測系 F で観測された移動履歴の集合 R^F
2. 観測系 G で観測された移動履歴の集合 R^G
3. 一般的な移動モデル $\boldsymbol{\theta}$

の 3 つが与えられたもとで，u と v が同一人物である確率 $P\left(u \equiv v | R^F, R^G, \boldsymbol{\theta}\right)$ を定式化する．ここで，$u \equiv v$ は観測系 F における人物 u と観測系 G における人物 v が同一人物である事象を示している．本方式では，以下を仮定して理論を展開する．

(A1) モデル $\boldsymbol{\theta}$ と同一人物らしさの独立性

$$P(\boldsymbol{\theta}, u \equiv v) = P(\boldsymbol{\theta}) \cdot P(u \equiv v) \tag{1}$$

モデル $\boldsymbol{\theta}$ のみが与えられ，対象人物の移動履歴が与えられていない場合，同一人物らしさはモデル $\boldsymbol{\theta}$ には依存しない．

(A2) 事前情報がない場合の同一人物らしさに関する一様性

$$P(u \equiv v) = \text{const.} \qquad (\forall (u, v) \in U \times V) \tag{2}$$

事前情報が与えられない場合，任意の人物の組合せについて同一人物らしさは等しくなる．

(A3) 移動履歴の発生確率の条件付き独立性

$$\begin{aligned} &P(r_w^H, R^* | \boldsymbol{\theta}, u \equiv v) \\ &= P(r_w^H | \boldsymbol{\theta}) \cdot P(R^* | \boldsymbol{\theta}, u \equiv v) \quad (w \neq u, v) \end{aligned} \tag{3}$$

式 (3) における R^* は任意の観測系で観測された移動履歴の集合を示している．本仮定では，任意の移動履歴 r_w^H について，持ち主 w が事前情報として与えられる u, v と無関係である場合，当該移動履歴の発生確率を独立して計算してもよいものとしている．

(A4) 持ち主が同一の移動履歴の発生確率

$$P(r_u^F, r_v^G | \boldsymbol{\theta}, u \equiv v) = P(r_{(u,v)}^M | \boldsymbol{\theta}) \tag{4}$$

式 (4) における $r_{(u,v)}^M$ は，移動履歴 r_u^F と移動履歴 r_v^G を結合してできる移動履歴を意味している．この結合移動履歴 $r_{(u,v)}^M$ は，2 つの移動履歴 r_u^F, r_v^G に含まれる位置情報を 1 つにまとめ，時刻についてソートすることで作られる移動履歴であり，$r_{(u,v)}^M$ の位置情報数 $|r_{(u,v)}^M|$ は $|r_u^F| + |r_v^G|$ になる．本仮定では，$u \equiv v$ を満たすとき，r_u^F と r_v^G の同時確率は u, v の移動履歴を結合して得られる移動履歴の発生確率と等しくなるとしている．

上記の 4 つの仮定を用いることで，$P\left(u \equiv v | R^F, R^G, \boldsymbol{\theta}\right)$ は以下のように変形できる．

$$\begin{aligned} &P\left(u \equiv v | R^F, R^G, \boldsymbol{\theta}\right) \\ &= \frac{P(R^F, R^G, \boldsymbol{\theta} | u \equiv v) \cdot P(u \equiv v)}{\sum_{(u', v') \in U \times V} P(R^F, R^G, \boldsymbol{\theta} | u' \equiv v') \cdot P(u' \equiv v')} \\ &= \frac{P(R^F, R^G | \boldsymbol{\theta}, u \equiv v) \cdot P(\boldsymbol{\theta}) \cdot P(u \equiv v)}{\sum_{(u', v') \in U \times V} P(R^F, R^G | \boldsymbol{\theta}, u' \equiv v') \cdot P(\boldsymbol{\theta}) \cdot P(u' \equiv v')} \\ &\propto P(R^F, R^G | \boldsymbol{\theta}, u \equiv v) \\ &= P(r_{(u,v)}^M | \boldsymbol{\theta}) \cdot \prod_{u' \in U, u' \neq u} P(r_{u'}^F | \boldsymbol{\theta}) \cdot \prod_{v' \in V, v' \neq v} P(r_{v'}^G | \boldsymbol{\theta}) \\ &\propto \frac{P(r_{(u,v)}^M | \boldsymbol{\theta})}{P(r_u^F | \boldsymbol{\theta}) \cdot P(r_v^G | \boldsymbol{\theta})} \end{aligned} \tag{5}$$

1 行目から 2 行目への変形にはベイズの定理を用いた．また，2 行目から 3 行目への変形には (A1) の仮定に基づいて得られる式 (6) の変形を用いた．

$$\begin{aligned} &P(R^F, R^G, \boldsymbol{\theta} | u \equiv v) \\ &= \frac{P(R^F, R^G | \boldsymbol{\theta}, u \equiv v) \cdot P(\boldsymbol{\theta}, u \equiv v)}{P(u \equiv v)} \\ &= \frac{P(R^F, R^G | \boldsymbol{\theta}, u \equiv v) \cdot P(\boldsymbol{\theta}) \cdot P(u \equiv v)}{P(u \equiv v)} \\ &= P(R^F, R^G | \boldsymbol{\theta}, u \equiv v) \cdot P(\boldsymbol{\theta}) \end{aligned} \tag{6}$$

3 行目から 4 行目への変形では，(A2) の仮定を用いたうえで，定数項の除去を行った．4 行目から 5 行目への変形では，(A3)，(A4) の仮定に基づいて得られる式 (7) の変形を用いた．

$$\begin{aligned} &P(R^F, R^G | \boldsymbol{\theta}, u \equiv v) \\ &= P(r_u^F, r_v^G | \boldsymbol{\theta}, u \equiv v) \cdot \prod_{u' \in U, u' \neq u} P(r_{u'}^F | \boldsymbol{\theta}) \cdot \prod_{v' \in V, v' \neq v} P(r_{v'}^G | \boldsymbol{\theta}) \end{aligned}$$

$$= P(r^M_{(u,v)}|\boldsymbol{\theta}) \cdot \prod_{u' \in U, u' \neq u} P(r^F_{u'}|\boldsymbol{\theta}) \cdot \prod_{v' \in V, v' \neq v} P(r^G_{v'}|\boldsymbol{\theta}) \quad (7)$$

6 行目の式は，5 行目の式を $\prod_{u' \in U} P(r^F_{u'}) \cdot \prod_{v' \in V} P(r^G_{v'}|\boldsymbol{\theta})$ で割ることで得られる．$\prod_{u' \in U} P(r^F_{u'}) \cdot \prod_{v' \in V} P(r^G_{v'}|\boldsymbol{\theta})$ は $R^F, R^G, \boldsymbol{\theta}$ のみから決まる値であり，u, v に依存しないため，5 行目と 6 行目の式の間で比例の関係が保たれる．

以上により得られる式 (5) の値を，u と v の類似度 $L(u,v)$ として用いる．

$$L(u,v) = \frac{P(r^M_{(u,v)}|\boldsymbol{\theta})}{P(r^F_u|\boldsymbol{\theta}) \cdot P(r^G_v|\boldsymbol{\theta})} \quad (8)$$

類似度 $L(u,v)$ の大きさは，移動履歴の位置情報数に依存しない．仮に移動履歴の位置情報数が大きくなった場合，式 (8) の分子の値は小さくなるが，同時に分母の値も小さくなり，位置情報数の影響が相殺される．これによって，移動履歴間で位置情報数が異なる場合でも，比較が可能となる．よって，2.3 節における (P2) は解決する．

3.4 処理の流れ

3.3 節で定式化を行った類似度 $L(u,v)$ を用いて，2 つの観測系における人物の対応関係 σ を明らかにするための手順を以下に示す．

(i) 不特定多数者の移動履歴を用いて，一般的な移動モデル $\boldsymbol{\theta}$ を作成する．

(ii) 任意の人物の組み合わせ $(u,v) \in U \times V$ について，式 (8) から類似度 $L(u,v)$ を計算する．

(iii) (ii) で求めた類似度 $L(u,v)$ を用いて $u \in U$ と $v \in V$ の紐づけを行うことで，人物の対応関係 σ を決定する．ここで，σ の決定方法として以下の 2 つが考えられる．

(a) 人物別の最尤推定
各 u について最も類似度の大きい v を選択し，紐づけることで人物の対応関係 σ を決定する．この方法の場合，σ は全単射にならない可能性がある．

$$\arg\max_{v \in V} L(u,v) \quad (9)$$

(b) 全体を考慮した最尤推定
式 (10) を満たす人物の対応関係 σ を求める．これは，全ての人物の組み合わせ $(u,v) \in U \times V$ の類似度を考慮し，総合的に判断することで σ を決定することを意味している．この方法の場合，σ は全単射になる．

$$\arg\max_{\sigma} \prod_{u \in U} L(u, \sigma(u)) \quad (10)$$

4. 手法の具体化

4.1 一般的な移動モデル $\boldsymbol{\theta}$

本節では，3.4 節の各ステップの具体化を行う．一般的な移動モデル $\boldsymbol{\theta}$ として，2.2 節で挙げた手法が用いていた遷移行列を使用することも可能であるが，本手法では，2 つの位置情報間の時間差・距離の関係を用いる．

図 6 モデル $\boldsymbol{\theta}$ の具体例

本手法の移動モデルの作成方法は**図 6** のようになっている．まず，移動履歴の各位置情報間の時間差・距離を抽出し，列挙する．その後，あらかじめ定めた時間差区分・距離区分に該当する移動の発生回数をカウントし，行列を作成する．最後に各要素を要素の総和で割ることで正規化を行い，確率の行列を作成する．こうして得られた行列を $\boldsymbol{\theta}$ とする．なお 3.2 節で述べたように，ここで使用する学習用移動履歴は，人物ごとの移動傾向を学習するためのものではない．そのため，学習用移動履歴は仮名化済みのものであっても良い．

次に，任意の移動履歴 r^H_w に対する $P\left(r^H_w|\boldsymbol{\theta}\right)$ と類似度 $L(u,v)$ の計算方法を具体化する．本手法では，3.3 節で述べた 4 つの仮定に加えて，以下の 2 つを仮定する．

(A5) 移動履歴の発生確率に関するマルコフ性

$$P\left((x_k,t_k)\,|\,(x_{k-1},t_{k-1}),\ldots,(x_1,t_1),\boldsymbol{\theta}\right) = P\left((x_k,t_k)\,|\,(x_{k-1},t_{k-1}),\boldsymbol{\theta}\right) \quad (11)$$

(A6) 1 つの位置情報の発生確率に関する一様性

$$P\left((x,t)\,|\boldsymbol{\theta}\right) = \text{const.} \qquad (\forall x,t) \quad (12)$$

どんな地点・時刻でも，存在確率は一定であることを仮定している．本手法におけるモデル $\boldsymbol{\theta}$ には絶対的な観測地点・時刻に関する情報は含まれていないため，ある 1 つの位置情報の発生確率を計算するうえで，モデル $\boldsymbol{\theta}$ は無情報に等しい．よって，このような仮定を行った．

上記 2 つの仮定を用いることで，類似度 $L(u,v)$ は以下のように変形できる．

$$\begin{aligned} L(u,v) &= \frac{P(r^M_{(u,v)}|\boldsymbol{\theta})}{P(r^F_u|\boldsymbol{\theta}) \cdot P(r^G_v|\boldsymbol{\theta})} \\ &\propto \frac{\prod_{k=2}^{|r^M_{(u,v)}|} P\left((x_k,t_k)\,|\,(x_{k-1},t_{k-1}),\boldsymbol{\theta}\right)}{\prod_{k=2}^{|r^F_u|} P\left((x_k,t_k)\,|\,(x_{k-1},t_{k-1}),\boldsymbol{\theta}\right) \cdot \prod_{k=2}^{|r^G_v|} P\left((x_k,t_k)\,|\,(x_{k-1},t_{k-1}),\boldsymbol{\theta}\right)} \end{aligned} \quad (13)$$

2 行目から 3 行目への変形では，(A5)，(A6) の仮定に基づいて得られる式 (14) の変形を行ったうえで，定数項の除去

を行った．

$$
\begin{aligned}
&P\left(r_w^H|\boldsymbol{\theta}\right)\\
&= P\left((x_k,t_k),(x_{k-1},t_{k-1}),\ldots,(x_1,t_1)|\boldsymbol{\theta}\right)\\
&= \prod_{k=2}^{|r_w^H|} P\left((x_k,t_k)|(x_{k-1},t_{k-1}),\ldots,(x_1,t_1),\boldsymbol{\theta}\right)\cdot P\left((x_1,t_1)|\boldsymbol{\theta}\right)\\
&= \prod_{k=2}^{|r_w^H|} P\left((x_k,t_k)|(x_{k-1},t_{k-1}),\boldsymbol{\theta}\right)\cdot P\left((x_1,t_1)|\boldsymbol{\theta}\right)\\
&= \prod_{k=2}^{|r_w^H|} P\left((x_k,t_k)|(x_{k-1},t_{k-1}),\boldsymbol{\theta}\right)\cdot \mathrm{const.} \qquad (14)
\end{aligned}
$$

なお，$P\left((x_k,t_k)|(x_{k-1},t_{k-1}),\boldsymbol{\theta}\right)$ には，一般的な移動モデル $\boldsymbol{\theta}$ で時間差 $t_k - t_{k-1}$・距離 $|x_k - x_{k-1}|$ に対応する要素の値を用いる．本手法における類似度 $L(u,v)$ には，式 (8) の代わりに式 (13) を使用する．

4.2 類似研究の問題点の解決

3.2 節で述べたように，提案方式は類似研究での遷移行列に対応する移動モデル $\boldsymbol{\theta}$ を人物ごとに作成しない．本方式での移動モデル $\boldsymbol{\theta}$ は，一般的な人間の移動傾向を表現することが目的であるため，SNS 等から得られる不特定多数者の移動履歴を用いて作成する．よって 2.3 節における (P1) は解決する．これにより，作成する移動モデル $\boldsymbol{\theta}$ がより正確になることが期待される．

また，3.3 節で述べたように，本方式では移動履歴間で位置情報の件数が異なっている場合でも，類似度 $L(u,v)$ が比較可能なものとなっている．よって (P2) も解決する．

一般的な移動モデル $\boldsymbol{\theta}$ では，位置情報に含まれる観測地点そのものを用いるのではなく，2 つの位置情報間の距離を用いる．そのため，モデルとして遷移行列を用いる場合と比較すると，対象地域の拡大を行ってもモデルのサイズが大きくなりにくいため，(P3) も軽減できている．また，時間差という情報をモデル $\boldsymbol{\theta}$ に組み込んでいるため，(P4) も一部対応できている．

5. データセット

5.1 観測系 F の移動履歴

電気通信大学の学生 24 人と一般の被験者 29 人，合わせて 53 人の被験者のスマートフォンの MAC アドレスを元に，Wi-Fi アクセスポイントへのプローブ要求から移動履歴を作成し[*1]，これを観測系 F における移動履歴とした．移動履歴の取得対象期間は 2017 年 1 月 25 日から 2017 年 4 月 24 日である．観測系 F の移動履歴に含まれる位置情報数の統計値は**表 1** のようになっている．本稿では，この観測系 F の移動履歴を類似研究における仮名化移動履歴として扱う．

[*1] 被験者からは個別に明確な事前同意を得たうえで，移動履歴の取得を行った．また，本研究は電気通信大学の倫理委員会の審査を経て実施している．

表 1 観測系 F の移動履歴に含まれる位置情報数の統計値

	平均	中央値	標準偏差
1 人あたりの位置情報数	21361.36	13068	24862.88
1 日あたり 1 人あたりの位置情報数	244.90	161.04	260.72

表 2 観測系 G の移動履歴に含まれる位置情報数の統計値

	平均	中央値	標準偏差
1 人あたりの位置情報数	161.17	99	168.24
1 日あたり 1 人あたりの位置情報数	1.97	1.64	1.16

表 3 学習用移動履歴に含まれる位置情報数の統計値

	平均	中央値	標準偏差
1 人あたりの位置情報数	126.15	47	405.37
1 日あたり 1 人あたりの位置情報数	2.06	1.50	7.23

5.2 観測系 G の移動履歴

まず，観測系 F と同一の被験者 53 人の Twitter [8] アカウントの投稿文を収集した．投稿文の収集はアカウントごとに行い，2017 年 4 月を起点に過去に遡ったものを使用する．こうして得られた投稿文のうち，地名を含むものだけを取り出し，その地名を緯度・経度に変換することで，移動履歴と同様のデータ形式になるようにし，これを観測系 G の移動履歴とした．なお，この変換処理には GeoNLP [9] を用いた．観測系 G の移動履歴に含まれる位置情報数の統計値は**表 2** のようになっている．本稿では，この観測系 G の移動履歴を類似研究における実名移動履歴として扱う．

5.3 一般的な移動モデル θ の学習用移動履歴

まず，5.2 節で取得した被験者 53 人分の Twitter アカウントとは関係のない，30 万人分の Twitter アカウントから投稿文を取得した．その後，5.2 節と同様の処理を行うことで，30 万人分の疑似的な移動履歴を作成し，これを一般的な移動モデル $\boldsymbol{\theta}$ の学習用移動履歴とした．この移動履歴に含まれる位置情報数の統計値は**表 3** のようになっている．

6. 実装

3.4 節で述べた処理の流れに従い，Python でプログラムの実装を行った．

6.1 一般的な移動モデル θ の作成

5.3 節で述べた学習用移動履歴を用いることで，図 6 のように時間差・距離に関する確率の行列を作成する．なお，距離については 2km 単位・上限 500km，時間差については 30 分単位・上限 24 時間とした．

表 4 評価結果

	R=10	R=100	R=1000	R=10000	R=100000
(a) 人物別の最尤推定	11.44 (21.6%)	12.22 (23.1%)	13.72 (25.9%)	15.50 (29.2%)	15.12 (28.5%)
(b) 全体を考慮した最尤推定	27.28 (51.5%)	31.80 (60.0%)	33.56 (63.3%)	34.36 (64.8%)	34.46 (65.0%)

6.2 人物の対応関係 σ の導出

式 (9) や式 (10) によって人物の対応関係 σ を求める際，移動履歴の位置情報数によっては計算途中でアンダーフローが発生することが考えられる．そのため，対数を取った値を用いることで，人物の対応関係 σ を求める．具体的な方法を以下に示す．

(a) 人物別の最尤推定
各 u について最も類似度の大きい v を選択し，紐づけることで人物の対応関係 σ を決定する．

$$\arg\max_{v \in V} \log L(u, v) \tag{15}$$

(b) 全体を考慮した最尤推定
式 (16) を満たす人物の対応関係 σ を求める．

$$\arg\max_{\sigma} \sum_{u \in U} \log L(u, \sigma(u)) \tag{16}$$

これは，線形和の割当問題を解くことに相当する．そのため，線形和の割当問題の解法として良く知られている Hungarian アルゴリズム [10] を用いて，人物の対応関係 σ を求めた．

7. 評価結果

7.1 評価方法

まず，5.3 節で述べた一般的な移動モデル $\boldsymbol{\theta}$ の学習用移動履歴 30 万人分の中から，R 人分の移動履歴をランダムにサンプリングし，これを用いて一般的な移動モデル $\boldsymbol{\theta}$ を作成する．その後，5.1 節・5.2 節で述べた移動履歴を用いて被験者の対応関係 σ を求めた．対応関係 σ に含まれる被験者の組合せ 53 組の中で，正しい組合せが何組存在していたかをカウントし，この組数を元に評価を行った．

7.2 評価結果

結果を**表 4** に示す．なお，R 人分の移動履歴のサンプリングにはランダム性が存在するため，実験は R ごとに 50 回ずつ行い，これらの平均を取った値を実験結果とした．この実験は CPU：Intel Xeon Gold 6252，メモリ：128GB の計算機を使って行った．$R = 100000$ のとき，モデル $\boldsymbol{\theta}$ の学習に要した時間は 364 秒，類似度の計算に要した時間は 596 秒であった．

表 4 から，(a)，(b) 共に R の値が大きくなるにつれて精度が向上する傾向にあることがわかる．これは，モデル $\boldsymbol{\theta}$ の学習に用いる移動履歴の数が増えるにつれて，モデル $\boldsymbol{\theta}$ がより一般的なものとなり，類似度計算の精度が向上したことによるものだと思われる．ただ，R の値が大きくなるにつれ精度の向上の度合いが小さくなっている．特に (a) においては，$R = 100000$ のとき，精度がやや悪化している．これは，R の値がある程度大きくなると，モデルの一般性が十分に確保でき，追加で学習用のデータを加えてもモデルの一般性が大きく変化しないためであると思われる．また，(a) と (b) を比較すると，全ての R について (b) の方が精度が良いことがわかる．3.4 節で述べたように，(b) の方式では被験者の組合せ全通り分の類似度を考慮し，総合的に判断することで対応関係 σ の推定を行う．(a) の方式のように，着目した被験者 u に関する類似度のみを用いるのではなく，他の被験者の類似度も加味して対応関係 σ を推定している点が，良い精度の要因になっていると思われる．

8. まとめ

本稿では，2 つの系列においてそれぞれ複数人分の移動履歴が存在したとき，系列間の移動履歴の持ち主の対応関係を推定する手法を提案した．類似研究では，不特定多数者の移動履歴を背景知識として活用することを想定していない・仮名化移動履歴が等長でなければならないという問題点が存在した．提案手法では，類似度計算を行う際に使用する移動モデルを，対応関係の推定対象となる人物の移動傾向を表現するためのモデルではなく，一般的な移動傾向を表現するためのモデルとして扱うことで，不特定多数者の移動履歴の活用を可能にした．また，類似度計算式の導出をベイズの定理及びいくつかの仮定に基づいて行うことで，人物間で移動履歴の位置情報数が異なる場合でも，類似度の比較が可能となった．さらに類似研究では，移動履歴の取得対象地域に関するスケーラビリティが乏しい・時間に関する情報を有効活用できないという問題点が存在したが，本手法ではモデルの形式を類似研究で使用されていた遷移行列から，2 つの位置情報間の距離及び時間差に関する確率の行列に変更することで，これらの問題点を部分的に解決した．

実験では，推定対象となるデータセットとして，電気通信大学の学生 24 人と一般の被験者 29 人，合わせて 53 人の被験者の Wi-Fi 移動履歴及び Twitter アカウントの投稿文を使用した．また，一般的な移動モデルの学習用データとして，30 万人分の Twitter アカウントの投稿文を用意した．用意した 30 万人分の Twitter アカウントのうち，10 万人分の Twitter アカウントの投稿文をランダムサンプリングし，これらを元にモデルを生成した後，提案手法によって持ち主の対応関係の推定を行ったところ，平均して 34.46 人分 (65.0%) の対応関係の推定に成功した．対応関係の推定がある程度の精度で行えており，類似研究の問題

点を解決できていることから，本提案手法は類似研究の手法と比べてより実用的な手法であると言える．

参考文献

[1] Shokri, R., Theodorakopoulos, G., Le Boudec, J.-Y. and Hubaux, J.-P.: Quantifying location privacy, *2011 IEEE symposium on security and privacy*, IEEE, pp. 247–262 (2011).
[2] Murakami, T.: Expectation-maximization tensor factorization for practical location privacy attacks, *Proceedings on Privacy Enhancing Technologies*, Vol. 2017, No. 4, pp. 138–155 (2017).
[3] Srivatsa, M. and Hicks, M.: Deanonymizing mobility traces: Using social network as a side-channel, *Proceedings of the 2012 ACM conference on Computer and communications security*, ACM, pp. 628–637 (2012).
[4] Ma, C. Y., Yau, D. K., Yip, N. K. and Rao, N. S.: Privacy vulnerability of published anonymous mobility traces, *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, pp. 185–196 (2010).
[5] Riederer, C., Kim, Y., Chaintreau, A., Korula, N. and Lattanzi, S.: Linking users across domains with location data: Theory and validation, *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, pp. 707–719 (2016).
[6] Gambs, S., Killijian, M.-O. and del Prado Cortez, M. N.: De-anonymization attack on geolocated data, *Journal of Computer and System Sciences*, Vol. 80, No. 8, pp. 1597–1614 (2014).
[7] Murakami, T.: A succinct model for re-identification of mobility traces based on small training data, *2018 International Symposium on Information Theory and Its Applications (ISITA)*, IEEE, pp. 164–168 (2018).
[8] Twitter, Inc.: Twitter, Twitter, Inc. (online), available from ⟨https://twitter.com/⟩ (accessed 2020-02-19).
[9] 国立情報学研究所：GeoNLP - 文章を自動的に地図化する地名情報処理システム，国立情報学研究所（オンライン），入手先 ⟨https://geonlp.ex.nii.ac.jp/⟩（参照 2020-02-19）．
[10] Kuhn, H. W.: The Hungarian method for the assignment problem, *Naval research logistics quarterly*, Vol. 2, No. 1-2, pp. 83–97 (1955).