統計数理(2018)
第 66 巻 第 2 号 225–233

特集「サービス科学の今」
[総合報告]

位置情報軌跡の統計的プライバシー保護

南 和宏†

(受付 2017 年 12 月 5 日；改訂 2018 年 3 月 27 日；採択 4 月 6 日)

要　　旨

スマートフォンの普及に伴い，我々の位置情報の取得が容易になり，多くのユーザーの移動履歴は，交通情報の提供，都市設計といった社会サービス，また商圏分析等の企業活動にも活用されている．一方，位置情報から個人の興味に関するプライバシーに関する情報が漏洩する危険性が懸念されている．位置情報の時系列データは，既存の匿名化手法の適用が困難な多次元データであり，個人の行動習慣，移動経路の制約等を反映した時空間の相関性を利用した統計的推論攻撃に対する防護策が必要となる．本記事では，位置情報軌跡を安全に分割する動的仮名更新手法，および，時空間の相関性による情報漏洩リスクを考慮した状態空間モデルに基づく匿名化データの安全性評価手法を紹介する．

キーワード：位置情報，匿名化，仮名化，マルコフ過程，状態空間モデル．

1. はじめに

近年，スマートフォンによる GPS 座標の位置情報の取得に加え，携帯電話の基地局，WiFi のアクセスポイント，IC カードによる電車の乗車履歴等，様々な種類の位置情報を入手することが可能になった．それら位置情報を統合することで多数のユーザーの長期間，広域での移動履歴が把握でき，首都圏でのリアルタイムの人口統計の提供（寺田 他, 2012），また商圏分析（清嶋, 2012）等の企業活動にも活用されている．さらには，運転操作データのような他の IoT データと位置情報を組み合わせることで，交通事故の発生する可能性の高い地点を表示するヒアリハット地図の作成（中野・豊田, 2015）に活用されている．

その一方，位置情報から，個人の習慣，興味，行動，交際範囲等，プライバシーに関する情報が明らかになる危険性がある．例えば，病院への定期的な訪問は重大な病気が推測され，カフェなど同一の場所での複数人の集まりは秘密の会議の開催を示唆するかもしれない．また位置情報は，ストーカーや空き巣のような犯罪に利用される可能性もある．よって位置情報の安全な 2 次利用には匿名化と呼ばれる個人の識別情報を取り除くデータ加工が不可欠である．

通常，匿名データを作成する際，氏名等の個人の識別子を削除するだけでは不十分である．なぜなら「年齢」，「性別」といった個人を断片的に識別する準識別子と呼ばれる属性情報が存在し，それらの組み合わせで個人の特定が可能になるからである．したがって，一般的には人々の属性に関する準識別子の情報を用いて k 未満のユーザーに絞り込むことを防ぐ k-匿名化処理（Sweeney, 2002）を行う．k-匿名化では元データを類似する k 個以上のレコードを含むグループに分割し，他のデータセットと照合されても特定のレコードの客体が再識別されることを防止するため，同一グループ内のレコードが同じ値を取るように一般化する．位置情報の

† 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

場合，情報の粒度を粗くすることが一般化処理に対応する．

しかし通常の k-匿名化の手法を位置情報軌跡に適用する場合，2 つの課題が存在する．1 つは，位置情報軌跡のような各時刻の位置情報を含む多次元データの場合，k-匿名化を実施するとデータの有用性が著しく劣化してしまう問題である．位置情報は個人のユニークな行動パターンを反映しており，長期的な位置情報軌跡を匿名化する場合，互いに類似する軌跡のグループを見つけることは困難である．そのような軌跡群をグループ化して一般化処理すると情報の損失が大きくなり，有益なデータ分析に堪えない．

2 つめは，位置情報軌跡のデータ間に時空間の相関性が存在し，匿名化した位置情報から統計的推論により元の軌跡情報が復元される問題である．位置情報軌跡には，人の移動に関する物理的制約が反映し，短時間での移動範囲は局所的であり，車，電車といった交通手段により移動経路は限定される．また長期的な移動軌跡には通勤，病院への通院といった個人の生活習慣を反映した特徴的なパターンが現れる．そのような移動パターンに関する外部知識を用いると一般化された匿名化データから元に位置情報が復元される危険性がある．

本記事では，この 2 つの課題を解決するための 2 つの手法を中心に位置情報の匿名化技術を紹介する．1 つは位置情報軌跡を複数のセグメントに分割する動的仮名交換手法（Tanjo et al., 2014）であり，ミックスゾーンと呼ばれる複数ユーザーの集積点でのランダムな仮名の再割当により移動先の不確定性を確保する手法である．もう 1 つは，状態空間モデルに基づく匿名化データの安全性評価手法である．ユーザーの移動パターンをマルコフチェーンでモデル化し，隠れマルコフモデルにおける内部状態の推定問題として匿名化データの安全性の評価を行う．

2. 攻撃者モデル

図 1 は位置情報データの流通形態を示す．位置情報サーバーは各ユーザーから定期的に時刻でタグ付けされた位置情報を受け取り，図 2 に示す表データに集計する．この表の各行はユーザーの位置情報を保持し，各列はある時刻の位置情報の値を示す．ただし，説明を簡略化するため，本記事では位置情報は座標値ではなく，座標値から変換された位置情報の領域に対応するグリッド ID を示す．位置情報サーバーは集計した表データから「氏名」等の識別子情報を削除する等，匿名化処理を施した匿名化データをデータ分析者に提供する．

匿名化技術における攻撃者は公開された匿名化データから標的とするユーザーの元データにおける位置情報軌跡を復元しようとする．この攻撃者は匿名化データ以外に標的とするユーザーの住所，勤務先，その他の目撃情報等の外部知識を利用する．このような攻撃者を想定すると単純な「氏名」等の識別子を仮名に置換する単純な匿名化処理では不十分である．例えば，図 3 は 3 人のモバイルユーザーの仮名化された軌跡を示す．もし攻撃者が「鈴木さん」の住所情

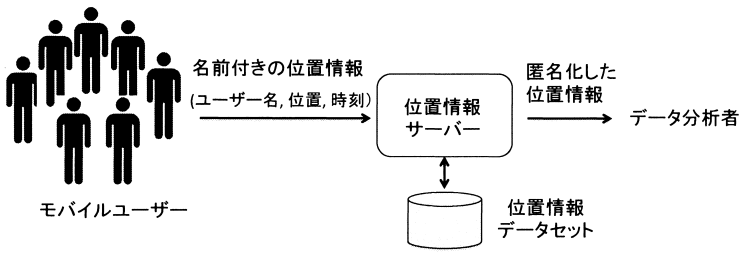


図 1．位置情報サービスの流通形態．データ分析者への位置情報の提供は一般公開の形態をとるため，元データから個人を識別する情報を除いた匿名化データを提供する．

2017年11月8日

氏名	8:00	8:30	9:00	9:30	10:00	10:30	11:00	
伊藤	1	5	4	8	12	15	9	
加藤	10	15	24	14	21	20	19	.....
鈴木	3	8	6	6	7	10	15	
高橋	23	24	19	11	9	4	5	

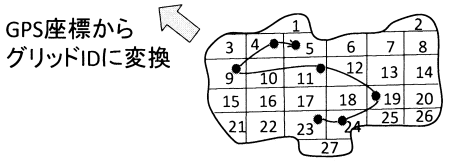


図 2. 位置情報軌跡の表データ.

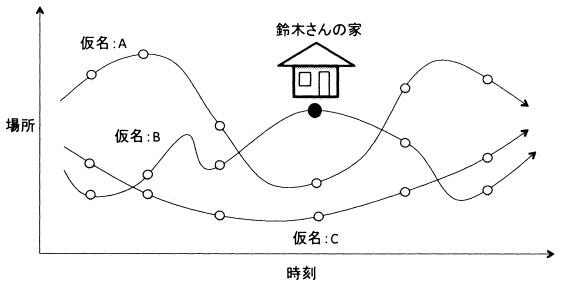


図 3. 仮名化された移動軌跡の概念図. 3 人のモバイルユーザーの実名を仮名 A，B，C に置き換える匿名化処理が行われている.

報を入手しており，仮名 B の軌跡がその場所を経由していることが分かると，仮名 B の軌跡は識別されてしまう.

実際これに類似する問題が 2012 年，JR 東日本が IC カード乗車券「Suica」の仮名化した乗車履歴を日立製作所に提供した時に起きている．Kikuchi ら（Kikuchi and Takahashi, 2015）は，駅の平均乗降数がジップの法則（Zipf's law）に従うと仮定してこの乗車履歴の安全性を評価し，個人が普段使う駅が 3 つ分かるだけで膨大な数の乗車履歴レコードから元の個人名が再識別されると分析している.

3. ミックスゾーンにおける動的仮名割当

2 章で示したように，個人の行動パターンが顕著に現れる位置情報軌跡の場合，その中の幾つかの点に過ぎない外部知識を用いて個人を識別することが可能性である．またいったん移動軌跡が識別されるとその軌跡全体の情報が開示されることとなり，位置情報軌跡の情報漏えいリスクは非常に高い.

本章では位置情報軌跡に紐付けられる仮名を動的に更新し長期間の軌跡データを複数の軌跡セグメントに分割することで情報漏えいリスクの局所化を実現する方式（Tanjo et al., 2014）を紹介する．この仮名の更新は複数のユーザーが同一の場所に存在する「ミックスゾーン」と呼ばれる場所で仮名交換の形式で実施され，ミックスゾーンを経由することで軌跡セグメント間の

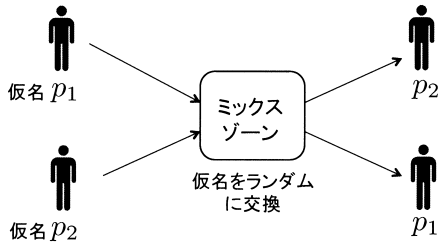


図 4. ミックスゾーンにおけるランダムな仮名再割当.

関連性を分断する.

3.1 ミックスゾーン

図 4 にミックスゾーンの概念を示す．各ユーザーの位置情報は実 ID を置き換えた仮名と紐付けられており，複数のユーザーが同時刻に集まるミックスゾーンで仮名をランダムに再割当が行われる．図 4 では仮名 p_1 と p_2 をもつ 2 人のユーザーがミックスゾーンに入り，そこで 2 つの仮名がランダムに再割当される．再割当の可能性は仮名が両者で交換される場合と交換されずに同じ仮名を持ち続ける場合の 2 通りである．ミックスゾーンを経由したとき，どちらの再割当が実施されたか匿名化データからは判別できないため，最初に仮名 p_1 のユーザーがミックスゾーン経由後，引き続き仮名 p_1 の経路をとる場合と p_2 の経路に移る場合の 2 つの可能性が共存する．このようにミックスゾーンを経由すると個人の移動軌跡の可能経路が分岐し，全体の位置情報軌跡の不確定性を増加させることができる.

3.2 安全性評価と排他的辺素パス問題

このミックスゾーンにおける仮名交換方式では，ユーザー u の時刻 t におけるプライバシー指標は到達可能な位置の数として定式化でき，到達可能な位置の数が多いほど高いプライバシーが保証できる．ただし一般のユーザーは自宅を始点として出発して最後はやはり終点である家に戻るという攻撃者が容易に知りうる拘束条件を持つため，全てのミックスゾーンの分岐経路が利用できるとは限らない．さらに，あるユーザーが特定の経路を使うと別のユーザーの始点から終点への経路が存在しなくなるという問題も生じる．つまりあるユーザーの代替経路を列挙する場合，他の全てのユーザーについても妥当な経路が存在することを保証する必要がある．もしデータ・セットに n 人の位置情報軌跡が含まれるとすると，この問題は図 5 に示すようなミックスゾーンをノードとするグラフ上で n 個の(始点，終点)の組が与えられたときに排他的辺素パスを列挙する問題に相当する．ユーザー u_2 が始点から終点に到達する経路は単独では $\langle 1 \to 3 \rangle$, $\langle 1 \to 2 \to 3 \rangle$ の 2 通りの順序でミックスゾーンを通過する経路が存在するが，後者の経路を選択するとユーザー u_1 が終点に向かう経路を分断してしまう．したがってこの例では排他的辺素パスは一組しか存在しない.

ユーザー数が入力として指定される場合，排他的辺素パス問題は NP 完全問題であることが知られている（Karp, 1975）．また攻撃者が始点，終点以外の中間の地点の情報(例えば，勤務先)を外部情報として持つ可能性もある．したがって仮名更新による安全性を評価するには排他的辺素パス問題をさらに一般化する必要がある．Tanjo et al., 2014 では一般化された排他的辺素パス問題を制約充足問題（Rossi et al., 2006)に変換し，仮名の変数の全ての異なる解の数を求めることで安全性の評価を行うシステムを開発した.

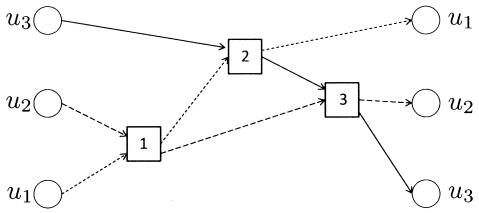


図 5. ミックスゾーンにおける排他的辺素パス問題．丸ノードは各ユーザーの始点，終点，四角のノードはミックスゾーンを表す．

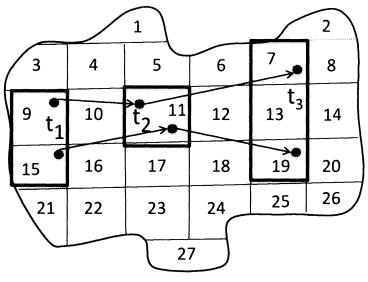
4. 状態空間モデルに基づく匿名化データの安全性評価

3 章で紹介した仮名更新による位置情報軌跡の分割は位置情報の識別リスクを局所化する手法である．しかしある位置情報が識別されると同じ軌跡セグメント内の位置情報は依然として漏洩してしまう．したがって分割して次元を削減した軌跡セグメント単位に対して k-匿名化を実施することが望ましい．ただし，位置情報軌跡には時空間の相関性が存在するので，通常の k-匿名化では不十分な場合が多い．本章では，統計的推論攻撃のリスクに対処するための状態空間モデルに基づく安全性評価の手法を紹介する．

4.1 k-匿名化の課題

k-匿名化は，標的ユーザーの軌跡の k 個未満への絞り込みを防ぐ実用的なプライバシー指標である．図 6 の例では，2 人のユーザーの位置情報を領域区分の ID で示しており，2 人の移動軌跡を同一にする 2-匿名化を実現するためには，粒度を粗くした太枠の区分に位置情報を変換する必要がある．

しかし位置情報の時系列データには，個人の行動習慣，移動経路の制約等を反映した時空間の相関性が存在し，k-匿名化した位置情報から元の軌跡の復元が可能な場合がある．例えば，図 7 は，2 人の車を運転するユーザーの移動軌跡を 2-匿名化した例である．太枠の区画に位置情報が一般化されているものの，破線で示す道路の経路情報が与えられれば，2 人のドライバーの位置を詳細に推測することは容易である．つまり k-匿名化は匿名化したデータの形式（シンタックス）のみを要件とするプライバシー指標であるため，軌跡データ間の相関性による情報漏えいのリスクが考慮できていない．



元データ

ユーザー	時刻 t_1	時刻 t_2	時刻 t_3
田中	{9}	{11}	{7}
鈴木	{15}	{11}	{19}

一般化処理

2-匿名化データ

仮名	時刻 t_1	時刻 t_2	時刻 t_3
A	{9, 15}	{11}	{7, 13, 19}
B	{9, 15}	{11}	{7, 13, 19}

図 6. 位置情報の 2-匿名化の例．位置情報はグリッドの集合として表現される．

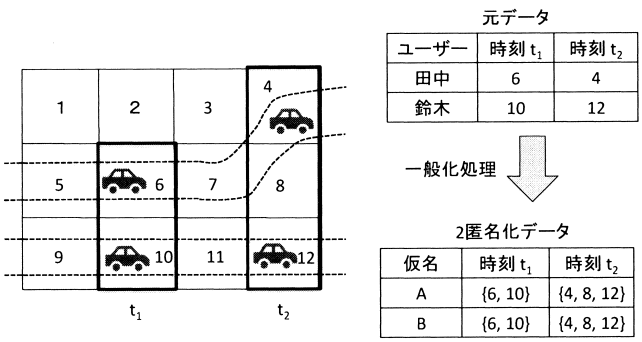


図 7. 道路情報を用いた位置情報の復元．破線は道路の形状を表す．

4.2 ユーザーの移動モデル

人々が移動する経路は，道路や建物の形状，通勤，通学等の生活習慣など様々な要因の影響を受ける．このような移動経路の特徴を統一的に捉えるには，時系列データの確率モデルであるマルコフ過程が適している．なぜなら人の移動を場所間の状態遷移と考える場合，移動経路に影響を与える要因は移動時の遷移確率に間接的に反映されるからである．例えば，歩行者が急に遠方の場所に移動できない事実や，侵入不可の建物による移動の制限は，それらの場所への遷移確率がないと記述すればよい．つまり移動に関する個別の要因をそれぞれ明示的に記述する必要がない．

このようなマルコフ行列は，各ユーザーの過去の移動軌跡を学習して作成することができる．ユーザー u の移動範囲が N 個の離散的な場所とすると，その状態遷移は $N \times N$ のマルコフ行列 P^u で記述される．マルコフ行列の各行 i が現在位置のグリッド i，各列 j が次の移動先のグリッド j に相当する．そしてグリッド i から j へ移動する確率はマルコフ行列の要素 $P^u_{i,j}$ で示される．

4.3 隠れマルコフモデルによる匿名化処理のモデル化

ユーザーの移動パターンをマルコフチェーンでモデル化し，匿名化技術の安全性評価を隠れマルコフモデルにおける観測情報から内部状態の推定問題として定式化する（Minami, 2014; Shokri et al., 2011）．図 8 のモデルの観測情報は匿名化データ，内部状態遷移は秘匿すべき元の位置情報に相当する．そして，匿名化処理は，内部状態から観測情報への確率的な変換を定義する記号出力行列として表現される．このモデル化の主な利点は，攻撃者が標的とする客体の移動パターンに関する知識を保持する場合，匿名化データの安全性を真の内部状態を推定する条件付き確率として定量的に評価できる点にある．

位置情報の匿名化手法は，4.1 章で述べた位置データの粒度を変える一般化処理以外にも幾つか存在する．例えば，図 8 において位置 l_1 は粒度が粗い l'_1 に一般化されているのに対し，位置 l_2 は省略を意味する空の文字($\perp$)に置換されており，また末端の l_T にはノイズが付加され，真の位置とは異なる l'_T に変換されている．

但し，図 8 は説明の簡略化のために一人のユーザー u_i の情報のみを表示しているが，実際のモデルは n 人のユーザーの情報を表現する必要がある．つまり内部状態は時刻 t における n 人の位置情報をもつベクトルとして表現し，それに応じて状態遷移を表すマルコフ行列，匿名化処理を記述する記号出力行列も拡張することになる．

例えば，図 9 は，2 人のユーザー u_1 と u_2 が 4 つのグリッド領域 $\{1, 2, 3, 4\}$ を移動する状況

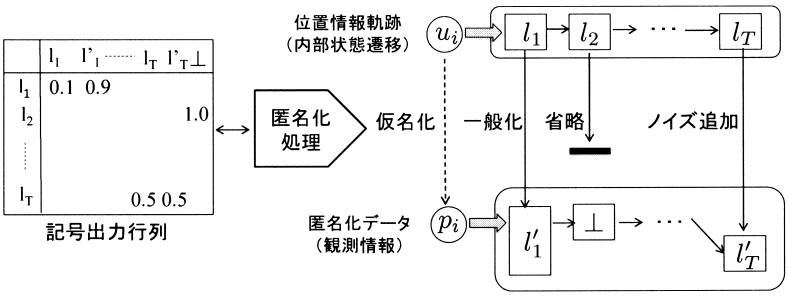


図 8. 隠れマルコフモデルによる匿名化技術のモデル化．元の位置情報軌跡は記号出力行列で定義された匿名化処理が適用され，匿名化データに変換される．省略された位置データは ⊥ で示す．

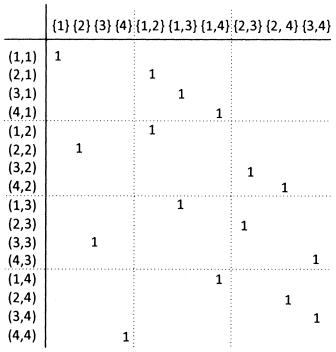


図 9. 2-匿名化関数を表現する記号出力行列．グリッド集合 $\mathcal{G} = \{1, 2, 3, 4\}$ とする．2 人のユーザーがそれぞれグリッド i, j に位置する状態を (i, j) と表す．2-匿名化関数は入力 (i, j) に対して決定的に $(\{i, j\}, \{i, j\})$ を出力する．2 人の匿名化された移動軌跡が同一なので，記号出力行列の各列の見出しは $(\{i, j\}, \{i, j\})$ を $\{i, j\}$ と簡潔化している．

で 2-匿名化を実施する場合の記号出力行列を示す．2 人のユーザーが時刻 t にそれぞれグリッド k, l に位置する場合，内部状態は (k, l) である．この 2-匿名化関数は単純な決定的関数であり，$f(k, l) = (\{k, l\}, \{k, l\})$ である．もし 2 人のユーザーが同じグリッド k に位置する場合，一般化加工は行われず出力は $f((k, k)) = (\{k\}, \{k\})$ となる．

図 9 は決定的で単純な匿名化アルゴリズムを記号出力行列として記述することが可能であることを示した．しかし，記号出力行列の形式で一般の匿名化アルゴリズムを記述するのは困難と予想され，隠れマルコフモデルの適用可能な匿名化アルゴリズムに一定の制限があることは明らかである．したがって，モデル化の適用範囲を明らかにし，匿名化アルゴリズムの汎用的な記述手法を確立することが今後の研究課題である．

5. まとめ

本記事では，一般に広く普及している k-匿名化手法を位置情報軌跡に適用する場合の 2 つの課題に対する解決策を提示した．1 つは，多次元データの匿名化における「次元の呪い」の問題 (Aggarwal, 2005) を回避するための位置情報軌跡の仮名更新による軌跡分割手法である．仮名

化された位置情報の安全性評価は排他的辺素パス問題の解列挙の問題に帰着され，制約充足問題ソルバーを利用した効率的な実施手段を実装した．

もう 1 つは位置情報の時空間の相関性を用いた推論攻撃に対する対策である．k-匿名化は標的となるユーザーと匿名化データの紐付け防止には一定の効果があるものの，移動に関する統計情報を用い，匿名化データから詳細の元データを復元する攻撃の危険性を見過ごしている．匿名化処理を隠れマルコフモデルでモデル化し，匿名化データの安全性を観測情報から内部状態への推定問題として定式化できることを示した．ただし，モデル化の対象に任意の匿名化アルゴリズムが含まれることになり，計算論的アルゴリズムと統計モデルの融合は今後の長期的課題となると予想される．

参 考 文 献

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality, *Proceedings of the 31st International Conference on Very Large Data Bases*, 901–909.

Karp, R. M. (1975). On the computational complexity of combinatorial problems, *Networks*, **5**, 45–68.

Kikuchi, H. and Takahashi, K. (2015). Zipf distribution model for quantifying risk of re-identification from trajectory data, *2015 13th Annual Conference on Privacy, Security and Trust* (*PST*), 14–21.

清嶋直樹 (2012). 電通 Draffic, http://itpro.nikkeibp.co.jp/article/JIREI/20121005/427881/.

Minami, K. (2014). Preventing denial-of-request inference attacks in location-sharing services, *2014 Seventh International Conference on Mobile Computing and Ubiquitous Networking*, 50–55.

中野美由紀, 豊田正史 (2015). ビッグデータがもたらす超情報社会—すべてを視る情報処理技術：基盤から応用まで．ビッグデータ時代を生きる, 情報処理, **56**(10), 958–961.

Rossi, F., van Beek, P. and Walsh, T. (2006). *Handbook of Constraint Programming*, Elsevier Science Inc., New York.

Shokri, R., Theodorakopoulos, G., Boudec, J. Y. L. and Hubaux, J. P. (2011). Quantifying location privacy, *2011 IEEE Symposium on Security and Privacy*, 247–262.

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

Tanjo, T., Minami, K., Mano, K. and Maruyama, H. (2014). Evaluating data utility of privacy-preserving pseudonymized location datasets, *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, **5**(3), 63–78.

寺田雅之, 永田智大, 小林基成 (2012). モバイル空間統計における人口推計技術（社会・産業の発展を支える「モバイル空間統計」：統計情報に基づく人口推計技術とその活用）, NTT DoCoMo テクニカル・ジャーナル, **20**(3), 11–16.

Statistical Privacy Protection of Location Trajectories

Kazuhiro Minami

Institute of the Statistical Mathemetics

Nowadays, trajectory location data, which is collected from peoples' smart phones, can be used for various analytic purposes, such as traffic monitoring, urban city planning. However, due to significant concern about location privacy, location data must be anonymized properly before making it available for secondary usage. Unfortunately, trajectory location data is inherently difficult to anonymize due to its high-dimensionality. Furthermore, we need to take additional measures to prevent inference attacks exploiting strong temporal and spatial correlations among data points. In this article, we present a technique of dynamically pseudonyms that devides a location trace into multiple segments and describe a state-space model to evaluate the safety of anonymized location data.

Key words: Location privacy, anonymization, Markov chain, state space model.