移動軌跡ストリームデータに対して遅延を許容することで情報損失を低減する k-匿名化手法

浜田　　凪† 　若林　真一†† 　上土井 陽子††

† 広島市立大学情報科学部　〒 731–3194 広島市安佐南区大塚東三丁目 4 番 1 号
†† 広島市立大学大学院情報科学研究科　〒 731–3194 広島市安佐南区大塚東三丁目 4 番 1 号
E-mail: †n_hamada@lcs.info.hiroshima-cu.ac.jp, ††{wakaba,yoko}@hiroshima-cu.ac.jp

あらまし　本稿では移動軌跡ストリームデータに対するリアルタイム k-匿名化手法を提案する．本稿で考察の対象とする移動軌跡ストリームデータは, 多数のユーザが所有するスマートフォン等から一定時間間隔で収集されるユーザの位置情報のストリームデータである．時々刻々と入力される移動軌跡ストリームデータに対してリアルタイムに k-匿名化することにより, ユーザのプライバシ保護を保証した匿名化ストリームデータがデータ解析者にリアルタイムで提供される．本稿では従来手法を改良し, データが収集される時刻からデータ解析者に匿名化されたデータが渡されるまでに一定の遅延時間を許容することで, より情報損失の少ない k-匿名化データが生成可能となる匿名化手法を提案する．

キーワード　移動軌跡, ストリームデータ, k-匿名化, プライバシ保護, リアルタイム処理

1 はじめに

近年, スマートフォンなどに搭載されている GPS によって, 位置情報を取得し活用するアプリケーションが増加している．例えば, カーナビは現在の位置情報を取得し, 目的地までの経路を提供する．また, 人の移動データを解析することで, 新たに有用なアプリケーションの開発につながる．しかし, 信頼できないサービス提供者や解析者に情報をそのまま渡してしまうと, 位置情報に付随した個人の嗜好や趣味などのプライバシ性の高い情報の漏洩につながる可能性がある．そのため, 位置情報は個人が特定できないように匿名化する必要がある．一方, 提供する位置情報を匿名化しすぎると, その分だけ情報損失となり, 情報としての価値が低い状態になり, 有用な解析やサービス提供を行うことができない．つまり, 匿名化と情報損失はトレードオフの関係にある．

データセットの準識別子を, 個人が特定される確率が $\frac{1}{k}$ 以下になるようにデータ値を変形することを k-匿名化 ($k-anonymization$) [6] という．準識別子とは, 名前や住所など, それ単体では個人を識別することはできないが, 複数組み合わせることで個人の識別が可能になる属性のことである．k 人以上の準識別子を同じ値に変形することで, 個人の特定を防ぐことができる．準識別子に対して, それ単体で個人を特定できる属性のことを識別子という．さらに, 他人には知られたくない属性をセンシティブ属性という．この属性は本人しか知りえない趣味や持病などが該当し, 解析などに使われるため匿名化されることはない．

位置情報の k-匿名化は距離が近い人同士で領域を構成し, 領域を位置情報とすることで k-匿名化を実現している．Aris らの論文 [3] では様々な位置情報の k-匿名化手法が紹介されている．

移動軌跡に対するリアルタイム k-匿名化の手法として,CMOA [7] が提案されている．移動軌跡の k-匿名化は最初から最後まで同じ k 人以上のメンバで匿名化をしなければならないため, 時間経過で領域の面積が大きくなり, 情報損失が増加するという問題点がある．リアルタイムで処理することを考えると, 最初の時刻だけで匿名化の領域のメンバを決定するため, 面積の増加は顕著に現れる．CMOA では動的再構成によって k-匿名性を維持しつつ, 領域内のメンバを変更することで, この問題に対処している．

また, 千葉らの研究 [8] ではリアルタイム処理ではないものの, 位置情報とタイムスタンプを修正することで移動軌跡の匿名性を高めることに成功している．

本論文では従来手法 CMOA を拡張し, 一定の遅延時間を許容することで動的再構成による領域のメンバ変更を正確に行い, より情報損失が少ないデータを生成することを目的とする．

本論文の構成は次の通りである．2 節で問題を定義し, 従来手法で用いられた評価指標について説明する．3 節で提案手法を詳細に説明する．4 節で実験方法と考察について述べる．5 節で本論文の結論をまとめる．

2 準　　備

2.1 移動軌跡

移動軌跡とは位置情報の時系列データのことである．本論文ではユーザの現在の位置情報として測位データ (id, x, y, t) を測位し, 時刻ごとに匿名化を行う．ここで id はユーザの移動軌跡を識別するための識別子であり,(x, y) はユーザの存在する座標, t は測位時刻を表す．また, 測位データにはセンシティブ情報を含むことが多いが, 本論文では省略する．測位データは t の全順序シーケンスである移動軌跡として蓄積される．移動軌跡は次のように定義される．

$$\tau = \{(id, x_i, y_i, t_i) | t_i < t_{i+1}, i = 0, 1, 2, \ldots, s\}$$

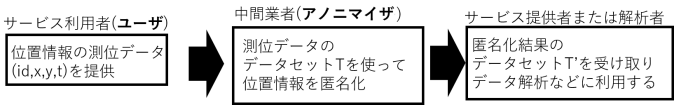

図 1　匿名化のモデル

表 1　2つの時刻においてユーザの存在する場所

ユーザ	時刻 t_i	時刻 t_{i+1}
a	駅	カフェ
b	駅	ー
c	駅	ー
d	ー	カフェ
e	ー	カフェ

ここで i は時系列の i 番目を示しており,t_s は現在の時刻を意味する.

図 1 に匿名化のモデルを示す. はじめにサービス利用者であるユーザは匿名化を行う中間業者 (アノニマイザ) に自身の位置情報とそれに付随する情報を提供する. アノニマイザはユーザから提供された測位データのデータセット T の匿名化を行い, 匿名化結果 T' をサービス提供者に渡す. この時, 第三者であるサービス提供者は信頼できないと想定する. 信頼できないとは, 何らかの方法で攻撃者に匿名化後のデータセット T' が知られてしまう可能性があることを示す. サービス提供者がアノニマイザとして匿名化を行うケースもあるが, その場合は信頼できるとして, 本論文では対象としない.

2.2 攻　撃　者

悪意を持って個人情報を漏洩しようとする攻撃者はユーザの移動軌跡の内, いくつかの位置情報を知っていると仮定する. これは, 攻撃者が特定の時刻にユーザを実際に観測したという事実に基づいており, ユーザの自宅や勤務地を知っている場合や, 偶然同じ場所に居合わせた場合が想定される. 最悪のケースとして, 攻撃者はユーザの移動軌跡の全ての位置情報を知っているという場合も考えられる. この仮定により, 攻撃者は複数の位置を使って個人の特定が可能になる. 個人が特定されると, 移動軌跡に付随したセンシティブ情報の漏洩につながってしまう. 例として, 攻撃者は2つの時刻に駅とカフェにいるユーザ A を観測したとする. このとき, 2つの時刻における匿名化の領域を構成するメンバがそれぞれ $(id = a, b, c)$ と $(id = a, d, e)$ であった場合, 2つの時刻に駅とカフェにいたユーザは a しかいないため, 攻撃者はユーザ A$= a$ だということが分かり, 個人が特定されてしまう. 表 1 に示す通り, ユーザ a は準識別子である位置情報 [時刻]t_i と t_{i+1} の組み合わせにおいて, 特有の値を持っていることがわかる. これを避けるため, 移動軌跡の匿名化はシーケンスの最初から最後まで同じ k 人以上のメンバで匿名性を維持しなければならない.

2.3 移動軌跡の匿名化

移動軌跡のデータセット T の匿名化したデータセットを T' とすると, 移動軌跡 τ を匿名化した匿名移動軌跡 τ' は以下のように定義される.

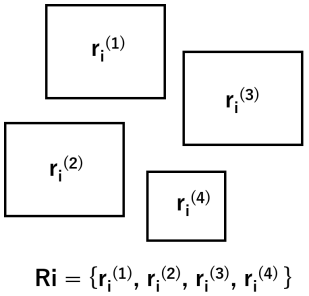

図 2　時刻 t_i における匿名移動軌跡の領域 R_i

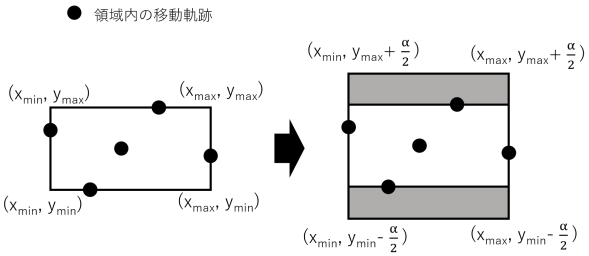

図 3　矩形領域から正方形領域への拡張

$$\tau' = \{(id', r_i, t_i) | t_i < t_{i+1}, i = 0, 1, 2, \ldots, s\}$$

ここで id' とは匿名化のためにランダムに割り振られた識別子であり, 識別子 id と対応付けられている. r_i は i 番目の時刻に, ユーザの厳密な位置座標 (x_i, y_i) を含めて, 面積を広げた領域である. 領域を構成する移動軌跡は, 全て領域内に存在する. よって, 領域内の移動軌跡は全て同じ値 r となり, 特定の個人と匿名移動軌跡との対応付けが困難になる.

i 番目の時刻 t_i における匿名移動軌跡の集合を

$$R_i = \{r_i^{(num)} | num = 1, 2, 3, \ldots, n\}$$

と定義する (図 3).

領域 r の面積を $S(r)$, 領域 r 内に存在する移動軌跡の数を $N(r)$ とする.

ある匿名移動軌跡 τ' が存在する全区間 $[t_0, t_s]$ において, 常に同じ位置情報 r_i を持った匿名移動軌跡が自身を含めて k 個以上存在するようにする. これを移動軌跡 τ の k-匿名化と呼ぶ.

2.4 領域の面積

領域の面積を定義する. 本論文では正方形の領域を想定する.

まずはじめに, 最小包含矩形を作成する. 匿名移動軌跡を構成するデータ点の中から x 軸方向 y 軸方向の最大値を x_{max}, y_{max} とし, 最小値を x_{min}, y_{min} とする. この時, 最小包含矩形は

$$(x_{min}, y_{min}), (x_{max}, y_{min}), (x_{max}, y_{max}), (x_{min}, y_{max})$$

の 4 点で構成される長方形となる. しかしこの場合, ユーザが各軸に一直線上に存在するときに面積が 0 になり, 不都合である. これを避けるため, 長方形の領域を正方形に変形する.

長方形の縦の辺 (y 軸) より横の辺 (x 軸) が長い場合, つまり $(x_{max} - x_{min}) > (y_{max} - y_{min})$ である場合は y 方向に領域を拡大する. α を両辺の差としたとき, 正方形領域は次の 4 点に

よって定義できる.

$(x_{min}, y_{min} - \frac{\alpha}{2}), (x_{max}, y_{min} - \frac{\alpha}{2}),$
$(x_{max}, y_{max} + \frac{\alpha}{2}), (x_{min}, y_{max} + \frac{\alpha}{2})$

また, 領域の面積は $(x_{max} - x_{min})^2$ となる. 縦辺が横辺より長い場合も正方形の面積を同様に定義する.

2.5 リアルタイム k-匿名化手法

従来手法である CMOA (Continuous Moving Objects Anonymization) は時々刻々と変化する位置情報に対してリアルタイムで移動軌跡を匿名化する手法である. CMOA は初期匿名化, 差分匿名化, 動的再構成の 3 つで構成されている. それぞれの動作について説明する.

1つ目の構成要素である初期匿名化では $i = 0$ 番目の時刻の匿名化, つまり最初の時刻の匿名化を行う. 移動軌跡が存在する各領域の面積が閾値 θ 以下になるようにユーザをクラスタリングする. ただし, 領域内に存在するユーザは k 人以上になるようにする.

2つ目の構成要素である差分匿名化では $i = 1$ 番目以降の時刻の匿名化を行う. 新たに取得した位置情報に対して前回匿名化した領域と同じメンバ構成で匿名化を行う.

3つ目の構成要素である動的再構成では, 差分匿名化で条件を満たした領域について, 分割や合成を行い領域を構成するメンバを変更する. ただし, メンバ変更は $\delta = \frac{N(r)}{S(r)}$ の値を計算し, 変更したときに値が大きくなる場合のみメンバ変更を行う.

さらに, 以下の 3 つの指標で CMOA によって生成された匿名移動軌跡のデータセットを評価した.

2.5.1 解像度指標 (RM)

匿名移動軌跡の情報損失の度合いからデータセットの有用性を評価する指標として, 解像度指標 ($ResolutionMetric$, RM) が用いられる. ピンポイントな位置情報である座標を曖昧な位置情報である領域に置き換えたとき, 面積が広がった分だけ情報損失となる. RM が高いほど情報損失は少なく, 有用なデータセットだといえる. あるユーザ id の存在する座標の解像度を 1 としたとき,i 番目の時刻における領域 r_i の解像度は $rm(id') = \frac{1}{S(r_i)}$ となる. ただし, 分割処理によって分割され, 領域内のメンバ数が k 人以下になった移動軌跡は結果として出力されないため, 解像度は 0 となる. 全ユーザの解像度の総和がその時刻の解像度指標になる.

$$RM[i] = \sum_{id'} rm(id')$$

2.5.2 最大継続時間 (MD)

移動軌跡がどれだけの追跡可能性を持っているかを評価する指標として, 最大継続時間 ($MaximumDuration$, MD) が用いられる. 長時間に渡って追跡できるほど移動軌跡の有用性が高いといえる. CMOA では分割や合成によって追跡可能性が失われる代わりに k-匿名性を実現している. あるユーザ id が同じ id' で移動し続けた時間を $Dur(id')$ と表すとき, 最大継続時間は

$$MD(id) = maxDur(id')$$

と定義する.

2.5.3 識別性指標 (DM)

匿名移動軌跡の識別の困難さからデータセットのプライバシ保護度合いを測る指標として, 識別性指標 ($DiscernibilityMetric$, DM) [2] が用いられる. DM が高いほどプライバシ保護度合が高いデータセットだといえる. 領域内に存在する匿名移動軌跡の数の 2 乗の総和がその時刻における識別性指標となる. ただし, 分割処理によって切り落とされた移動軌跡は識別不可能性が最大として扱う.

$$DM[i] = \sum_{num=1}^{n} N(r_i)^2 + |sup[i]||T|$$

ここで $|sup[i]|$ は i 番目の時刻に切り落とされている移動軌跡の数を表し,$|T|$ は匿名化前のデータセットの数を表す.

3 提案手法

従来手法 CMOA では動的再構成において, 領域を分割する際, 距離が近いメンバの移動軌跡同士で分割をする. そこで, 遅延を許容し匿名化結果の出力を 1 時刻だけ遅らせることで, 次の時刻の位置情報を考慮することができ, より高度な分割処理ができるのではないかと考えた. 合成処理に関しても, 合成する領域を選択する際に, 次の時刻の領域の位置情報を考慮することができる. これらの改良によって解像度指標の改善や最大継続時間の最大値や平均値が増加する可能性が考えられる. また, 従来手法では動的再構成を行う条件に δ の値を使用したが, 領域内に存在する移動軌跡の数で評価される識別性指標は k-匿名性を満たしていれば十分であると考えられるため, 提案手法では δ は使用しない.

3.1 問題の定式化

以上を踏まえて, 本論文の問題の定式化を行う.

複数のユーザの位置情報を定期的に取得し, 前回までの位置情報と対応付けを行い, そのユーザの移動軌跡とする. 位置情報を取得した時刻に匿名化を行い, 結果をリアルタイムで出力する.

本論文では現在の時刻 t_s を定め, 移動軌跡が存在する全区間 $[t_0, t_s]$ の全ての時刻について, 時刻ごとにリアルタイム処理することを想定して移動軌跡の k-匿名化を行う.

入力 移動軌跡 τ のデータセット $T[t_0, t_s]$
出力 匿名移動軌跡 τ' のデータセット $T'[t_0, t_s]$
目的関数 最大継続時間 MD を最大にする.
制約条件 1 全ての匿名移動軌跡が k-匿名性を満たす.
制約条件 2 領域の面積を閾値 θ 以下に抑える.

匿名化を行う時間間隔はデータセット作成時に適当に定める. 例えば, 時間間隔を 5 分とした場合では, 時刻 t_0 の測定時刻は 0 分, 時刻 t_1 の測定時刻は 5 分,t_2 の測定時刻は 10 分となる.

匿名化結果を出力する際は動的再構成のため, 一定時刻出力を遅らせなければならない. 遅延を d と表し, 遅延を許容しない場合は $d = 0$, $d > 0$ であれば,d 時刻の遅延を許容する. つま

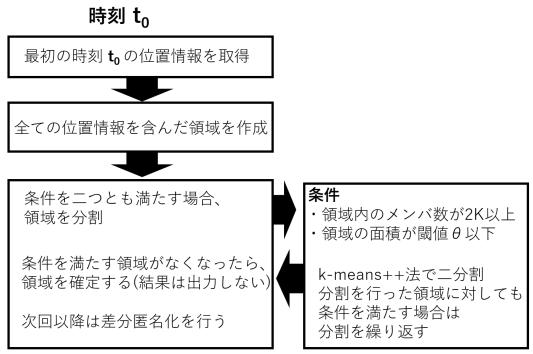

図 4　提案手法：初期匿名化の流れ

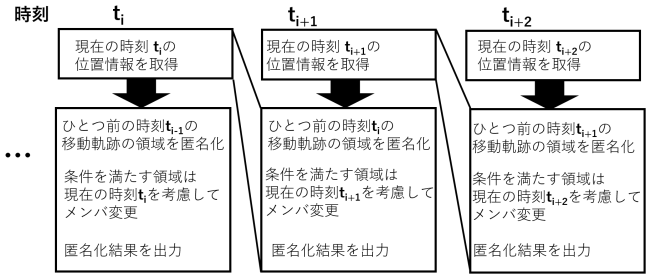

図 5　提案手法：差分匿名化の流れ

り, 時刻 t_i までの移動軌跡の匿名化の結果を出力する時刻は時刻 t_{i+d} となる．本論文では $d = 1$, すなわち出力に 1 時刻の遅延を許容するものとする．　3つの評価指標に関しては, それぞれトレードオフの関係にあるため, どれかを改善することでどれかが犠牲になってしまう．そこで本論文では最大継続時間の最大化を目的関数とし, 解像度指標は可能な限り改善を試みる．

3.2　遅延時間を許容する k-匿名化手法

3.2.1　初期匿名化

初期匿名化では, 測位データを取得したらはじめに全てのユーザの位置情報を含んだ領域を作成する．作成した領域は識別性指標が最大となり, 当然ながら k-匿名性も満たされている．しかし, 解像度指標は最低となり, 情報損失度合が最大であるため, 位置情報としての価値は乏しい．そのため, 領域の面積の閾値 θ を設け, 各領域の面積が θ 以下になるまで分割する必要がある．

領域の分割は k-means++法 [1] を用いる．これは非階層クラスタ分析の最も有名な手法である k-means 法 [4] を改良したものである．k-means++の k はクラスタ数を表しており,k-匿名化の k とは異なる．クラスタ数を 2 に設定することで領域を 2 分割することが可能になる．2 分割を再帰的に繰り返し行うことで, 初期匿名化のクラスタリングを行う．

まず, はじめに作成した領域を k-means++法によって 2 分割する．その後, 2 分割された領域が

- 領域を構成するメンバ数が $2k$ 以上
- 領域の面積が閾値 θ より大きい

この 2 つの条件を満たしている場合, 再度領域の分割を行う．以降, 分割された領域に対しても上記の条件を適用し, 条件を満たさなくなった領域はその領域で確定とする．k-means++法の分割は分割された領域に含まれるメンバ数に関して均等でないため, 全ての領域確定後にメンバ数 k 未満の領域が発生する場合がある．その場合は, 各領域の中心を計算し, 最も近い領域と合成する．

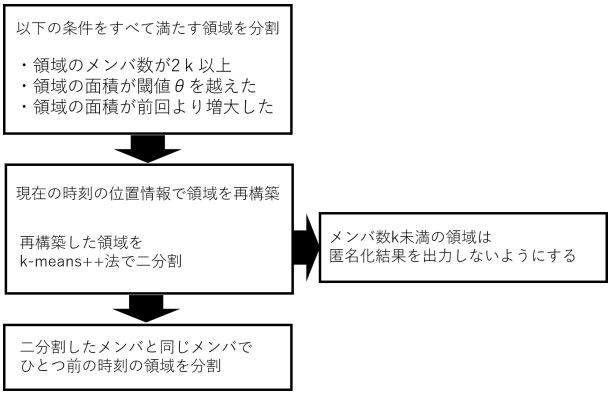

図 6　提案手法：分割処理の流れ

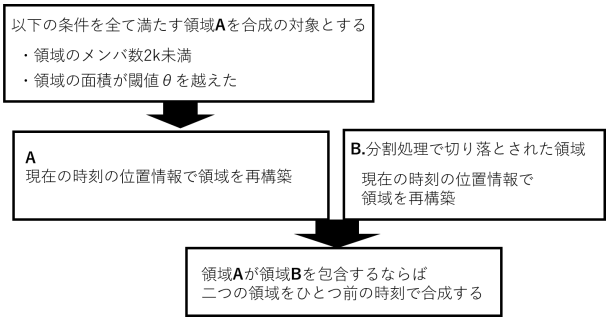

図 7　提案手法：合成処理の流れ

その後, ユーザに識別子 id' をランダムで振り分ける．

提案手法では $d = 1$ の遅延を許容しているため, 初期匿名化では匿名化結果を出力しない．提案手法の初期匿名化の流れを図 4 に示す．

3.2.2　差分匿名化

差分匿名化では, 測位データを取得したらはじめに id と id' の対応付けを行う．その後, 前回の位置情報を元に領域の面積を再計算し, 匿名化結果として出力する (前回が初期匿名化の場合はそのまま結果を出力する)．ただし, 条件を満たす領域については分割処理あるいは合成処理に基づく領域の動的再構成を行い, 領域のメンバを変更する．提案手法の差分匿名化の流れを図 5 に示す．

3.2.3　分 割 処 理

- 領域を構成するメンバ数が $2k$ 以上
- 領域の面積が前回より増加した
- 領域の面積が閾値 θ より大きい

以上の条件を満たした領域を分割処理の対象とする．現在の時刻に測位した増分位置情報で領域を構成する．構成した領域に対して,k-means++法 (クラスタ数 2) で 2 分割を行う．2 分割した結果のメンバで分割処理の対象となった元の領域を分割する．

2 分割した際に, メンバ数が k 未満の領域が発生する場合がある．そういった領域は切り落としを行い, 匿名化の結果として出力しない．提案手法の分割処理の流れを図 6 に示す．

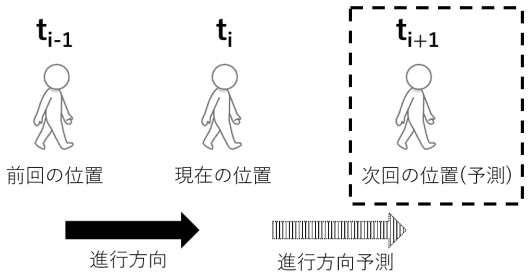

図 8　進行方向予測の例

3.2.4　合 成 処 理

以下で定義される領域 A と B に対し, 領域 B の重心が領域 A に包含される場合, この 2 つの領域を合成する.

A　領域を構成するメンバ数が $2k$ 未満かつ領域の面積が閾値 θ を超えた領域

B　分割処理によって切り落とされた領域

メンバ数 k-未満の切り落とされた領域が合成されるため, 合成処理を行った結果をそのまま出力すると移動軌跡の k-匿名性への違反となる. よって合成処理を行った場合は領域の識別子 id' をランダムに付け直す必要がある. 過去の移動軌跡との関連性はなくなるが, 集団の移動の流れだけは残る. 提案手法の合成処理の流れを図 7 に示す.

3.3　遅延の無い分割手法

提案手法の拡張として, 分割処理の際, ユーザの進行方向を予測することにより分割するメンバを決定する手法を新たに提案する. この手法では, 過去の移動軌跡に基づき予測を行うため遅延を発生させずに分割を行うことが可能である. 遅延時間は発生しないため $d = 0$ である.

進行方向予測は数時刻前の連続した位置情報から速度と方向の変化を計算し行う. 例えば, あるユーザの前回の位置が現在の位置と比べてマイナス方向であった場合, 次の時刻の位置は現在の位置と比べてプラス方向である可能性が高い (図 8). 変化が近いメンバ同士で領域を分割すれば, 領域を構成するメンバは同じ方向に進むと予測できる. しかし, あくまで予測であるため, 実際に予測通りに動くとは限らない. さらに, 時間間隔が開きすぎると有用性がなくなるという問題点もある. この拡張手法の評価は今後の課題とする.

4　実験と考察

従来手法と提案手法の比較実験を行うことで提案手法を評価する.

4.1　実 験 環 境

従来手法と提案手法のプログラムを C 言語で実装し, 比較を行った. 使用した PC は CPU:Intel Core i7(2.60GHz), 主記憶:8GB, OS:CentOS Linux7 である.

4.2　評価と考察

匿名化に用いるデータセットは移動体シミュレータである Siafu [5] を用いて 2000 人の移動軌跡を作成した. モデルはドイツの都市 Leimen であり, 移動軌跡が動く範囲は $1.148km \times 1.390km$ の範囲に設定した. そして, 午前 6 時から午後 8 時まで 5 分間隔で 14 時間, 計 169 時刻分の計測を行った. 提案手法は 1 時刻だけ遅延を許容するため, 実際に出力されるデータは 168 時刻分である. 比較のため従来手法の出力結果も提案手法に合わせて調整した. ユーザの自宅と勤務地はランダムに決められ, ユーザは計測開始時刻に自宅で目覚め, 勤務地へ向かう. その後, しばらく勤務した後, 帰宅する. このとき, ユーザは決められたルートは通らず, 寄り道する可能性がある. また, 移動手段として車を使うユーザも存在し, 徒歩で行動するよりも早く目的地に着く.

(k, θ)=$(5, 900)$, $(5, 1600)$, $(5, 2500)$, $(10, 1600)$, $(20, 1600)$ の 5 パターンで実験を行った.

各評価指標ごとに実験結果をまとめ, 考察をする.

表 2　解像度指標の平均値 ($k = 5$)

	解像度指標		
	θ =900	θ =1600	θ =2500
既存手法	3.854	3.726	3.510
提案手法	**4.057**	**4.067**	**3.699**

表 3　解像度指標の平均値 ($\theta = 1600$)

	解像度指標		
	k=5	k=10	k=20
既存手法	3.726	**2.238**	1.071
提案手法	**4.067**	2.216	**1.587**

4.2.1　解像度指標

ユーザのピンポイントな位置情報を $1m$ 四方の範囲とする. ユーザ 1 人の解像度の最大値が 1 であるため, 解像度指標の最大値は 2000 である. 図 9 に解像度指標の結果を示す.

従来手法と提案手法に大きな差は発生しなかった. 解像度指標が改善しなかった理由としては, 動的再構成のときに現在の時刻 t_i の位置情報で時刻 t_{i-1} の領域のメンバ変更を行うため, 時刻 t_{i-1} の領域に関してはむしろ面積が広がってしまったことが原因と考えられる. しかし, 表 2 と表 3 に示すように, 解像度指標の全時刻の平均値を見ると提案手法は従来手法と比較して改善していることがわかる.

4.2.2　最大継続時間

図 10 に最大継続時間の平均値の比率を示す.

全てのパターンにおいて, 最大継続時間は提案手法の方が上回っていた. k の値を増やすことで, その差はより広がった. これは,k-匿名性の k を増やすことで, ユーザの移動軌跡にばらつきが生まれ, 領域の面積の拡大が早まり, 頻繁に分割が行われたためだと考えられる. 従来手法に対して, 提案手法は分割を行う際に 1 時刻だけ追跡可能性を保証しているため, その分だけ, 最大継続時間の平均値が大きくなっていると考えられる.

4.2.3　識別性指標

識別性指標の最大値はデータセット数の 2 乗であるため,

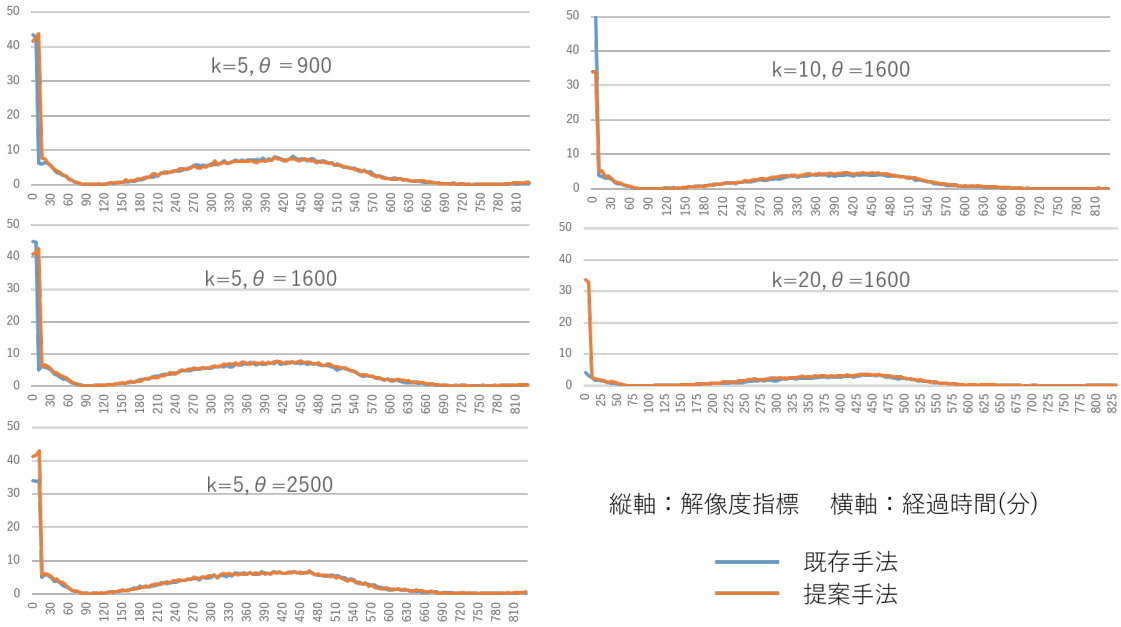

図 9　解像度指標

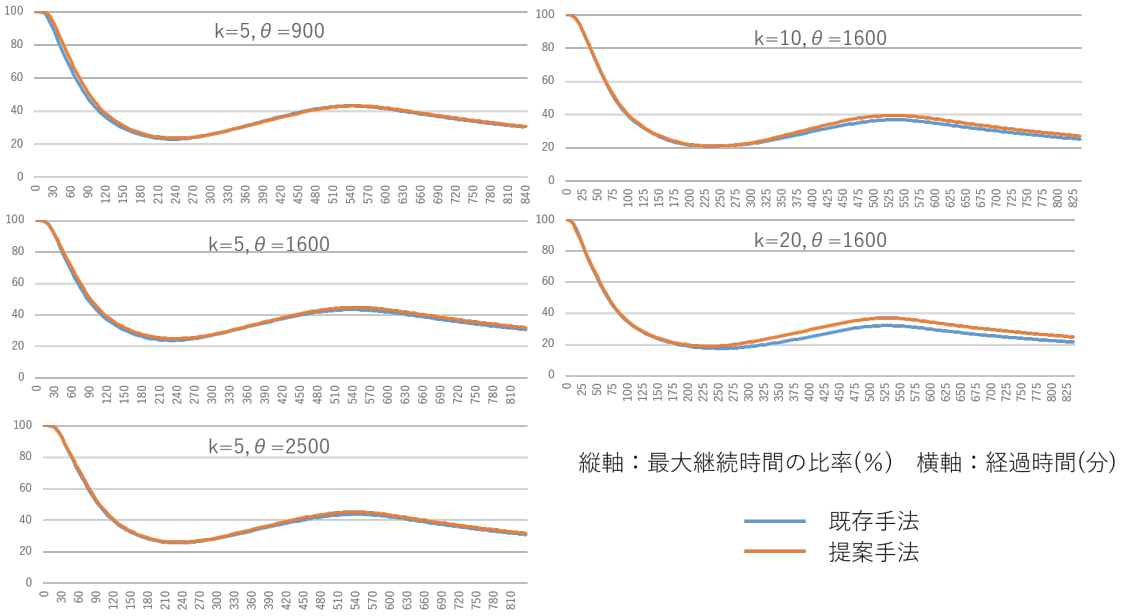

図 10　最大継続時間の平均値 (比率)

40×10^5 である．図 11 に識別性指標の結果を示す．

従来手法と提案手法に大きな差は発生しなかった．

識別性指標の結果は規則性がなく，値はランダムに上下していることがわかる．用いたデータセットでは中央の時間帯にユーザの移動が少ないため，識別性指標の変動が少なくなることがわかる．

また，図 12 に示すように，分割処理によって切り落とされた移動軌跡の数と識別性指標には相関があることがわかる．相関係数の値は 0.961 であり，その相関は非常に強い．つまり，切り落とされた移動軌跡が多ければ多いほど，識別性指標は高くなる．

表 4 と表 5 に従来手法と提案手法の識別性指標の全時刻の平均値を示す．識別性指標の平均値は従来手法の方が比較的高い値を持っている．これには 2 つの理由が考えられる．

1 つ目は切り落とした移動軌跡の数である．表 6 に全時刻で切り落とされた移動軌跡の数の平均値を示す．提案手法の方が切り落とされた移動軌跡の数が少ないことがわかる．

2 つ目は，従来手法は動的再構成に δ の値を使っている点である．3 つの評価指標はそれぞれトレードオフの関係にあるため，どれか 1 つの指標を重要視することで他の指標が犠牲になる．そこで従来手法では δ を導入し，これを最大化することで，解像度指標と識別性指標の両立を図った．これに対して本研究では前述した通り，識別性指標は k-匿名性を満たしていれば十分であると考え，識別性指標の向上は考慮しなかった．

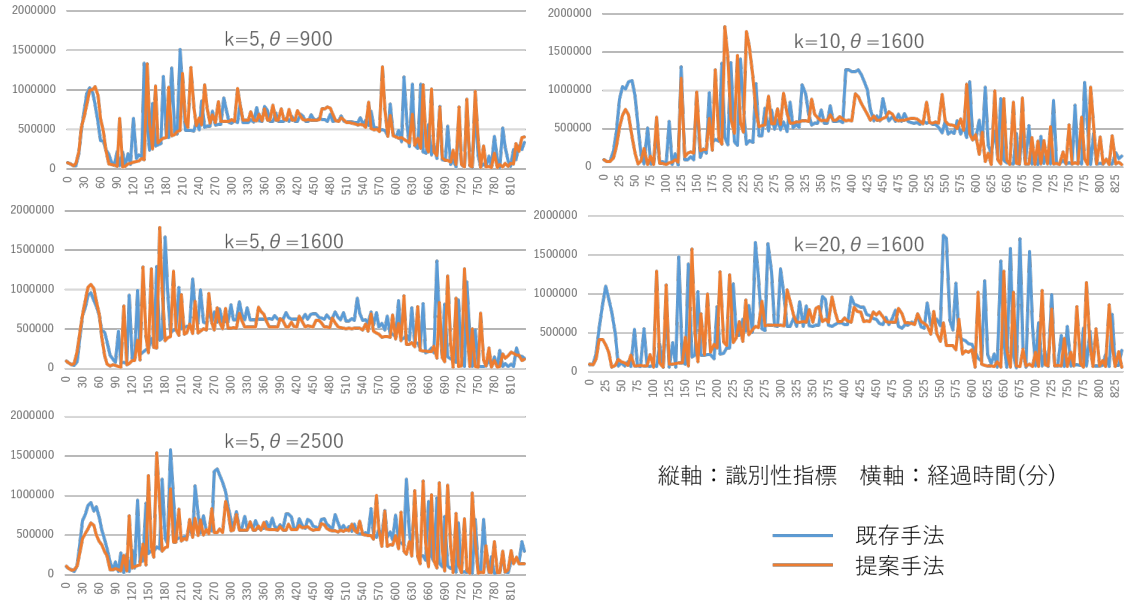

図 11　識別性指標

表 4　識別性指標の平均値 ($k = 5$)

	識別性指標(×10⁵)		
	θ =900	θ =1600	θ =2500
既存手法	4.865	**5.011**	**4.997**
提案手法	**4.941**	4.595	4.564

表 5　識別性指標の平均値 ($\theta = 1600$)

	識別性指標(×10⁵)		
	k=5	k=10	k=20
既存手法	**5.011**	**4.894**	**5.327**
提案手法	4.595	4.801	4.544

4.3　今後の課題

今後の課題として, プログラムの改良が挙げられる. 本研究の提案匿名化手法のプログラムでは, 初期匿名化において $t = 0$ の時刻で領域を作成したため, 初期分割において移動ベクトルは考慮しなかった. よって, 初期分割の際に動的再構成と同様に, 現在の時刻の位置情報を考慮することで, 計測開始時刻から数時刻先までの追跡可能性を伸ばすことができ, 最大継続時間の平均値が向上すると考えられる.

加えて, 遅延をさらに増やし, 移動ベクトルを分割に利用することで, 情報損失をより低減できるのかを考察する. 今回の研究では 1 時刻前の移動軌跡に動的再構成を行い, 匿名化結果を出力していたが, この遅延を 2, 3 と増やしていく. 最終的に遅延が全時刻分になると, リアルタイム匿名化ではなくなるため, 静的に匿名化した時と同じ最適な結果が得られると考えられる. よって, 遅延を増やすことにより, 評価指標がどれだけ改善され, 情報損失が低減するのかを, 遅延を許容する時間ごとに求めることで, 有益な解析結果が得られる移動軌跡ストリームデータに対するリアルタイム k-匿名化を調べることができると考えられる.

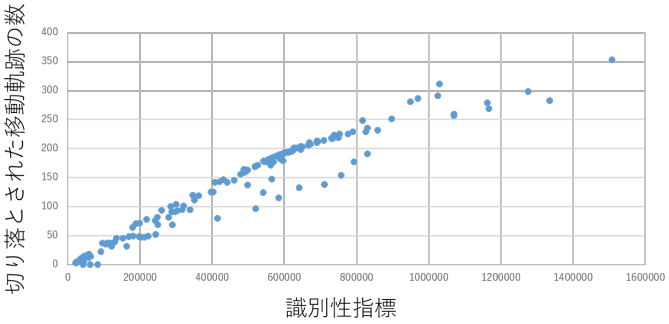

図 12　切り落としと識別性指標の関係 ($k = 5, \theta = 900$)

表 6　切り落とされた移動軌跡の数の平均値

	$k = 5$			$k = 10$	$k = 20$
	$\theta = 900$	$\theta = 1600$	$\theta = 2500$	$\theta = 1600$	
既存手法	143.2	152.4	138.8	79.9	41.2
提案手法	141.0	135.3	134.6	76.0	34.6

5　ま　と　め

本論文では移動軌跡をリアルタイムに k-匿名化する手法 CMOA を改良し, 移動ベクトルを利用することで情報損失を低減する手法を新たに提案し, 実験によりその有効性を確認した. 新たに提案した手法では移動軌跡の追跡可能性が向上し, 最大継続時間が改善した. また, 僅かながら解像度指標の改善も見られたが, 代わりに識別性指標は低下した. 識別性指標の低下

は情報損失度合いに影響しないため, 総合的に見て情報損失は低減したといえる.

移動軌跡をリアルタイムで匿名化する手法はユーザが一定の時間間隔で移動軌跡のサービスを受ける場合などに有用であると考えられる. しかし匿名性を満たすために, 結果として出力されるデータセットは元データと比べて情報損失が非常に大きくなる可能性がある. 初期匿名化において, 各領域を構成する移動軌跡をうまく選択すれば, 非リアルタイムで匿名化した場合と同程度の情報損失となる可能性がある. 従って, リアルタイム匿名化の初期段階で, いかにして同じ方向に移動するユーザの移動軌跡の組み合わせを選択するかが今後の課題である.

文　　献

[1] D.Arthur and S.Vassilvitskii, " k-means++: the advantages of careful seeding", Proc.18th Annual ACM-SIAM Symposium on Discrete Algorithms, pp.1027-1035, 2007.

[2] R.Bayardo and R.Agrawal, "Data privacy through optimal k-anonymity", Proc.ICDE2005, 2005.

[3] Aris Gkoulalas-Divanis, Panos Kalnis and Vassilios S. Verykios, "Providing K–Anonymity in location based services", SIGKDD Explorations, Volume 12, Issue 1, pp.3-10, 2010.

[4] J.MacQueen, "Some methods for classification and analysis of multivariate observations", Proc.Fifth Berkeley Symp. on Math.Statist. and Prob., Vol.1(Univ.of Calif.Press, 1967), pp.281-297, 1967.

[5] M.Martin and P.Nurmi, "A generic large scale simulator for ubiquitous computing", Proc.2006 Third Annual International Conference on Mobile and Ubiquitous Systems Networking Services, pp.1-3, 2006.

[6] P.Samarati and L.Sweeney, "Protecting Privacy when Disclosing Information: *k*-Anonymity and Its Enforcement through Generalization and Suppression", Harvard Data Privacy Lab., 1998.

[7] 高橋 翼, 宮川 伸也, 伊東 直子, "移動軌跡ストリームに対するリアルタイム k 匿名化手法の提案", DEIM Forum 2011, C5-1, 2011.

[8] 千葉 智樹, 清 雄一, 田原 康之, 大須賀 明彦, "位置情報とタイムスタンプの有用性を調整可能な移動軌跡匿名化手法", 電気学会論文誌 C(電子・情報・システム部門誌), Vol.140, No.8, pp.956-963, 2020.