明治大学大学院　先端数理科学研究科

２０１９年度

修士学位請求論文

軌跡情報の DTW 距離を保存する時空間の匿名加工手法の提案

学位請求者　先端メディアサイエンス専攻

二谷　太郎

目次

第 1 章　序論 1
1.1　研究背景 1
1.2　従来研究 1
1.3　研究目的 2
1.4　本稿の構成 3

第 2 章　事前準備 4
2.1　用いるデータセット 4
2.2　データセットの正規化 4
2.3　データモデルの合成 4

第 3 章　軌跡情報の DTW 距離を保存する匿名加工手法の提案 7
3.1　簡易手法の定義 7
3.2　提案手法 8

第 4 章　評価実験 11
4.1　有用性指標 11
4.2　簡易手法と提案手法の比較 11
4.3　加工前後のデータの差異 13
4.3.1　元データと加工データの緯度分布の差 13
4.3.2　元データと加工データの移動量分布の差 13
4.4　合成データにおける簡易手法と提案手法の比較 16
4.5　合成データにおける加工前後のデータの差異 17
4.5.1　合成データと加工データの緯度分布の差 17
4.5.2　合成データと加工データの移動量分布の差 18
4.6　考察 18

第 5 章　まとめ 21

参考文献 22
謝辞 24

第 1 章　序論

1.1　研究背景

昨今の技術革新により，購買履歴情報や乗降履歴等の個人情報を含むパーソナルデータが大量に蓄積されるようになっている．その中でも時刻，緯度，経度からなる位置情報の時系列データである軌跡情報は非常に多くの情報を含む需要の多いパーソナルデータであり，その情報を解析，利用することによってより旅行支援システムや，商業店舗の出店における効果的な立地案を求める手法など，様々な利益をもたらすことができる．しかし，パーソナルデータを利活用するにあたって，情報漏えいによる家や職場の特定など，プライバシー侵害のリスクが常に付きまとう．

例えば，2000 年に起こったマサチューセッツ州の医療記録が特定された事件が挙げられる．州に雇用されている 135,000 人分の医療データ収集し，氏名の削除をして研究者に提供したところ，郵便番号，生年月日，年齢の 3 属性の組み合わせで約 87%のユーザが一意に特定できてしまうことを Sweeney らが示している[1]．他にも，2013 年に JR 東日本が Suica の履歴データを販売しようとして顧客からの多数の苦情を受け取りやめた事件[2]や，2019 年にリクナビで内定辞退率のリストを販売し，厚生労働省から行政指導が行われた事件[3]など，多数の問題点があった．そのため，個人情報の管理には細心の注意を払わなければいけない．

2017 年 5 月に個人情報保護法[4]が改正され，新たに「匿名加工情報」が定義された．これにより，位置情報を含むパーソナルデータを非個人情報として企業間や組織間で取引することができるようになった．

1.2　従来研究

匿名加工を行うアプローチには様々なものがある[5][6][7]．[5]は 2012 年に Tamir らにより行われた研究であり，元データの各レコードが加工後データから見て k 個以上の対応を持たせるように加工することにより，従来の k-匿名化[1]と比べてデータの損失を減らす手法である．[6]は，2019 年に山添らによって行われた研究であり，高次元データに対して k-匿名化を行った際に有用性が著しく下がってしまう現象を，次元を削除するアルゴリズムを用いることにより解消した手法である．[7]は，2018 年に前田らによって行われた研究であり，匿名化データの受領者が加工者に対してどのような情報を求めるかの意向を送付し，その意向を元に匿名化を施すことにより有用性を保証する手法である．

さらに，位置情報を含む個人情報を匿名化するアプローチも多数存在する

[8][9][10][11][12]. [8][9][10]は 2016 年から 2017 年にかけて正木らによって行われた研究であり，従来は緯度経度の 2 次元でクラスタリングを行い，それに対して加工を施すのが主流であったがそこに 3 次元目として時刻を加えてクラスタリングし，加工することにより位置精度を保証する手法であった. [11]は 2017 年に疋田らによって行われた研究であり，東京大学が主催である MITHRA プロジェクトにより集めた膨大な位置履歴データから，どの程度の履歴データであれば個人特定性があるかの検討を行っている．[12]は 2018 年に河内らによって行われた研究であり，道路沿いに設置されている路側機の ID である RSU-ID の集合を地図として用い，車両から取得した GPS 位置情報を RSU-ID に変換したデータに対して k-匿名化を行うことにより，安全性を保障した.

1.3 研究目的

上記の手法では従来通りのユークリッド距離を用いて匿名加工を施して有効な結果を出してはいるが，次にあげるデータを適切に処理できない課題があった．表 1.1 にユークリッド距離を用いた際にうまくいかないデータの例を示す．あるユーザ M の時系列データと，M の別日の時系列データ M'を考える．別日のデータは，辿るルートは同じであるが辿る時間帯が異なる．さらに，ルートも時間帯も全く異なるユーザを N とする．そこで M-M'間のユークリッド距離の総和と，M-N 間のユークリッド距離の総和を比較する．M のxセル目を$M_{(x)}$，M'のxセル目を$M'_{(x)}$とすると，M-M'の距離の算出式は以下の通りである.

$$\text{M-M'間の距離} = \sum_{x=1}^{9} \left| M_{(x)} - M'_{(x)} \right|$$

M-M'の距離が 9 なのに対して，M'-N の距離は 6 となり，別日の同じユーザよりも他のユーザとの距離の方が近くなり M'は N と似ていると推定されてしまう.

この問題点に対して，本研究では DTW（Dynamic Time Warping）を用いる．DTW は 2 つの時系列データのパターンマッチにより 2 者間の距離を算出することができる[13]．そのため時間帯は違えど，似たパターンの同一ユーザに対しては本来の期待通り距離が小さくなることが考えられる． M-M'間の DTW 距離を$f(\max(x), \max(x))$と定義する．この際，$f(i,j)$は以下の通りに再起的に求める.

$$f(i,j) = \left| M_{(i)} - M'_{(j)} \right| + \min \begin{cases} f(i,j-1), \\ f(i-1,j), \\ f(i-1,j-1), \end{cases}$$

$$f(0,0) = 0, f(i,0) = f(0,j) = \infty.$$

なお，本研究においては 1 つのセルにおいて，緯度経度の 2 次元からなるデータを用い

るため，仮に M が緯度経度の 2 次元データであると仮定し，M のxセル目の緯度データを$M_{(x,lon)}$，経度データを$M_{(x,lat)}$としたとき，DTW の内部計算には 2 次元のユークリッド距離を以下のように用いる．

$$\left|M_{(i)} - M'_{(j)}\right| = \sqrt{(M_{(i,lon)} - M'_{(j,lon)})^2 + (M_{(i,lat)} - M'_{(j,lat)})^2}$$

実際にこの例で考えると，M-M'間の DTW 距離は 0，M'-N 間の DTW 距離は 4 となり，正しく M と M'が似ていると算出することができた．そこで本稿では表 1.1 のような例でも有用性を保証することを目的として，手法を提案する．また，距離誤差による有用性を用いて評価実験を試みる．

表 1.1: ユークリッド距離でうまくいかないデータの例

	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00
M	2	3	4	3	3	2	2	2	2
M'	2	2	2	2	2	3	4	3	2
N	1	2	1	2	3	4	5	3	1

1.4 本稿の構成

本稿は５章で構成される．

- 第１章：本稿の研究背景と先行研究と研究目的を述べた．
- 第２章：本稿で取り扱うデータと表記を定義する．
- 第３章：提案手法と簡易手法を提案する．
- 第４章：データの分析と提案手法の評価実験，考察を行う．
- 第５章：本稿のまとめを行う．

第２章　事前準備

本章では，取り扱うデータと表記の定義を行う．また，本章以降におけるシンボル表を表 2.1 に示す.

2.1　用いるデータセット

本研究では株式会社ナイトレイによって無料公開されている疑似人流データ[14]を用いる．

2013 年 7 月 1 日のデータを Web サービス Mobmap[15]を用いて実際の地図上にプロットした散布図を図 2.1 に，ナイトレイ疑似人流データ Dの一部を表 2.2 に，Dを正規化したデータD_1を表 2.3 に示す.

Dは，ユーザ ID，性別（推定値），時刻，緯度，経度，滞在者カテゴリ（大分類），滞在者カテゴリ（小分類），滞在情報の 9 つの属性からなる．本研究では，データの分析に必要ない属性を除き，ユーザ ID，時刻，緯度，経度を使用する．

2.2　データセットの正規化

Dは, 2013 年 7 月 1 日の一日のユーザの移動軌跡を 5 分おきに取得したものであり, ”STAY” セルから次の”MOVE”セルまでの時刻において緯度経度が同じ座標が入る. そこで本研究では，その間を補完するように，Dに正規化を施して全ユーザデータを完全な 5 分おきのデータに変換したデータをD_1とする．正規化後のデータD_1の概要を表 2.4 に示す．本研究では 6,432 ユーザの中から 100 ユーザをランダムに抜粋しており，その 100 ユーザからなるデータをD_2とする．以後，D_2を元データと呼称する．

2.3　データモデルの合成

ナイトレイのデータは日ごとに異なるユーザ ID が割り振られていた．そのため，本研究で扱う 2013 年 7 月 1 日とは異なる合成データD_3を，表 1.1 のような例を想定し，以下のように作成する．

1. D_2を元にして，最も長い時間滞在している場所をユーザごとの家や職場の場所と推定する．

2. ユーザごとに家や職場の時間をランダムに（5 時間から-５時間）引き伸ばして，その時間に応じてそれ以外の時間帯をずらす.

3. 家や職場と推定した以外の時間帯において，乱数（緯度経度において 0.03 から-0.03）を付加する.

表 2.1: 本稿におけるシンボル表

シンボル	意味
D	ナイトレイ疑似人流データ
D_1	Dに正規化を施したデータ
D_2	D_1を 100 ユーザサンプリングしたデータ
D_3	D_2を元にして作成した合成データ
c	クラスタリングする際のクラスタの数
k	k-匿名性を保証する k の値
D'_{kmens_Euc}	ユークリッドによる距離行列と k-平均法から匿名化したデータ
D'_{hclust_Euc}	ユークリッドによる距離行列と群平均法から匿名化したデータ
u_A	あるクラスタ内に分類されたあるユーザ
u_B	ユーザ A と同一のクラスタに分類されたユーザ
u_C	ユーザ A，ユーザ B と同一のクラスタに分類されたユーザ
u'	あるユーザに加工を施したデータ
$u_{(x)}$	あるユーザのxセル目の数値
u_p	ピンが刺さり，加工を施さないユーザ
D'_{kmens_Dtw}	DTW による距離行列と k-平均法から匿名化したデータ
D'_{hclust_Dtw}	DTW による距離行列と群平均法から匿名化したデータ

表 2.2: 疑似人流データD（一部）

ID	時刻	緯度	経度	滞在情報
3	21:30	35.6679	139.4389	MOVE
3	21:35	35.6657	139.4382	MOVE
3	21:40	35.6660	139.4382	MOVE
3	21:45	35.6653	139.4370	STAY

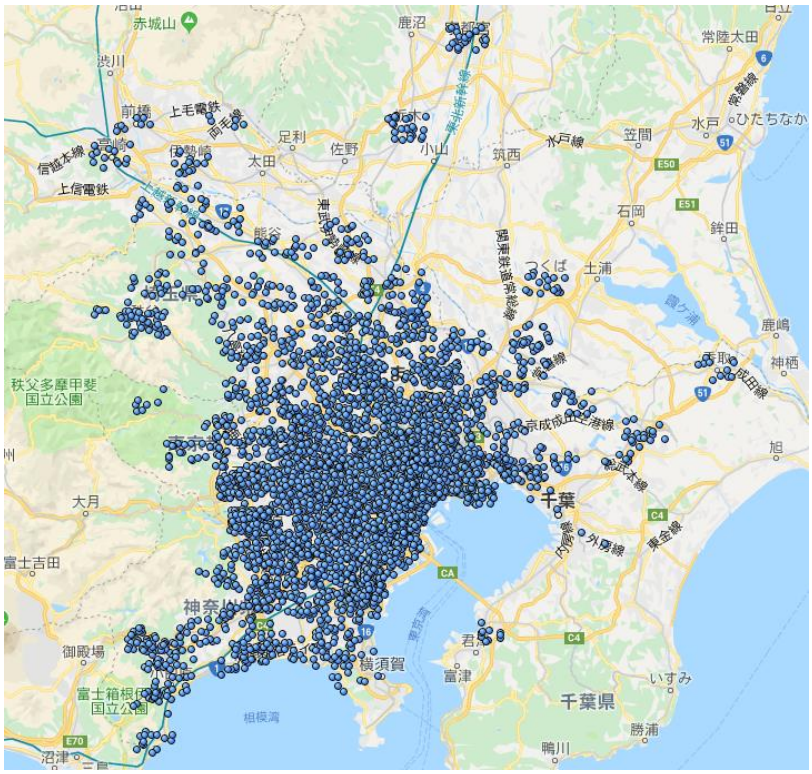

図 2.1: ナイトレイデータセットにおける 2013-07-01 のユーザ位置の散布図[15]

表 2.3: 正規化後データD_1

ID	時刻	緯度	経度
3	21:35	35.6657	139.4382
3	21:40	35.6660	139.4382
3	21:45	35.6653	139.4370
:	:	:	:
3	23:55	35.6653	139.4370

表 2.4: 正規化後データD_1の概要

総ユーザ数	6,432
1 ユーザ当たりのレコード数	288
総レコード数	1,852,416

第 3 章　軌跡情報の DTW 距離を保存する匿名加工手法の提案

本章では問題を解決するための匿名加工手法を提案する．まず，比較手法として[5]に基づいて，ユークリッド距離を距離行列にしてそれを用いてクラスタリングし，ミクロアグリゲーションすることにより匿名化を行う簡易手法を定義する．

3.1　簡易手法の定義

1. D_2を元に，緯度経度の 2 次元からユークリッド距離を用いて全ユーザ 100×100 の距離行列を作成する．ユークリッドによる距離行列の一部を表 3.1 に示す．

2. その距離行列を元に，非階層的クラスタリングである k-平均法によるクラスタリングと，階層的クラスタリングである群平均法の 2 種類のクラスタリングを行い，全ユーザをc個のクラスタに分割する．以後，k-匿名と k-平均法との k における混乱を防ぐために，k-平均法を c-平均法と呼称する．

3. k-匿名性を満たすために，k人以上のユーザを含んでいないクラスタを削除する．

4. 図 3.1 に簡易手法の適応前後の例を示す．ここで，ステップ 2 によってあるクラスタに 100 ユーザの中からユーザ A とユーザ B，ユーザ C の 3 人のみが分類されたと仮定し，ユーザ A をu_A，ユーザ B をu_B，ユーザ C をu_C，加工が施されたあるユーザをu'，あるユーザのxセル目の数値を$u_{(x)}$と定義する．時刻毎に同じクラスタに分類されているk人以上のユーザの緯度経度それぞれの平均値を求めて，同じクラスタに分類されているユーザの同時刻の緯度経度を平均値で置き換える．図 3.1 における加工後の各セルの数値は，

$$u'_{(x)} = \frac{u_{A(x)} + u_{B(x)} + u_{C(x)}}{3}$$

で表される．

5. c-平均法による匿名化されたデータをD'_{kmens_Euc}，群平均法による匿名化されたデータをD'_{hclust_Euc}とする．

3.2 提案手法

前節の手法では 1.3 節のような合成データに対して有用性が保証できない.

そこで，本稿では DTW を用いることにより問題の解決を計った．DTW はある一所にとどまっていることや，時間帯は異なるが道筋は同じなどの日常的にあり得るケースに対して 1.3 節のように距離が小さいと算出することができる．そこで，DTW の性質に則した加工手法を施すことによって合成データにおける有用性を保証することを試みる．しかし，DTW は 2 者間の距離しか測ることができないため，クラスタ内で収束させるような加工を施すことは難しい．そこで本研究では，以下のように提案する．

1. D_2を元に，緯度経度の 2 次元から DTW を用いて全ユーザ 100×100 の距離行列を作成する．DTW による距離行列の一部を表 3.2 に示す.

2. その距離行列を元に，非階層的クラスタリングである c-平均法によるクラスタリングと，階層的クラスタリングである群平均法の 2 種類のクラスタリングを行い，全ユーザをc個のクラスタに分割する.

3. k-匿名化に近しい安全性を確保するために，k人以上のユーザを含んでいないクラスタを削除した.

4. 同じクラスタに分類されているユーザ同士の DTW 距離を小さくするように加工をするために，クラスタごとに完全にランダムな 1 ユーザにピンを立てる．ピンが刺さっているユーザには加工を施さない.

5. 図 3.2 に提案手法の適応前後の例を示す．3.1 節のステップ 4 と同様に，ステップ 2 によってあるクラスタに 100 ユーザの中からユーザ A とユーザ B，ユーザ C の 3 人のみが分類されたと仮定する．新たに，そのユーザの中でピンが刺さっているユーザをu_pと定義する．図 3.2 においてはユーザ B にピンが刺さったと仮定し，説明を行う．この際の，u_A-u_P間の DTW 距離は 2，u_P-u_C間の DTW 距離 3，u_A-u_C間の DTW 距離は 2 であった．DTW 距離を求めるにあたって最短経路を結んだ時の経路はワーピングパスと呼称されている（以後パスと呼称する）．パスの選び方は 1.3 節の計算式の一部を用いて，

$$\min\begin{cases} f(i,j-1), \\ f(i-1,j), \\ f(i-1,j-1) \end{cases}$$

で示される．図 3.2 では加工前のユーザ間の線でつながっているセル同士が，パスがつながっていることを示している．このステップでは，同一クラスタ内の全ユーザ間の DTW 距離をパスの対応関係をできるだけ崩さずに小さくなるような加工を施す．パスがつながっているセル同士の数値を近づければ近づけるほど DTW 距離を小さくすることができるため，加工を施すユーザのセルの値を，ピンが刺さっているユーザのパスがつながっているセルの値と同じ数値に置き換える．図 3.2 では，$u_{A(1)}$とパスがつながっているのは$u_{P(1)}$であるため，$u'_{A(1)}$は，

$$u'_{A(1)} = u_{P(1)} = 1$$

となる．ところが，ピンが刺さっていないユーザの 1 つのセルに，ピンが刺さっているユーザの複数のセルからパスがつながることがあり，単純に同じ数字を置き換えることができない例がある．その際は，パスがつながっているセルの平均値で置き換える．図 3.2 では，$u_{A(2)}$とパスがつながっているのは$u_{P(2)}$と$u_{P(3)}$であるため，$u'_{A(2)}$は，

$$u'_{A(2)} = \frac{u_{p(2)} + u_{p(3)}}{2} = \frac{2 + 1}{2} = 1.5$$

となる．同一クラスタ内の加工が施されていないユーザにもピンが刺さ亭るユーザと同様の処理を行う．加工を施すことにより，u'_A-u'_P間の DTW 距離は 1，u'_P-u'_C間の DTW 距離 1，u'_A-u'_C間の DTW 距離は 0 となり，同一クラスタ内の全ユーザ間の DTW 距離が小さくなった．更に，加工後もユーザ間のパスの対応関係は加工前と同じであった．

6. c-平均法による匿名化されたデータをD'_{kmens_Dtw}，群平均法による匿名化されたデータをD'_{hclust_Dtw}とする．

加工前データ

u_A	u_B	u_C
2	1	3
2	2	1
3	1	2
3	3	3

加工後データ

u'_A	u'_B	u'_C
2	2	2
1.67	1.67	1.67
2	2	2
3	3	3

図 3.1: 簡易手法の適応前後の例

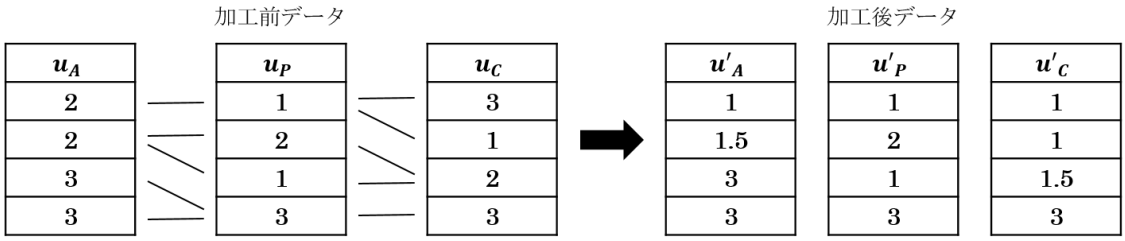

図 3.2: 提案手法の適応前後の例

表 3.1: ユークリッドによる距離行列の一部

	1	2	3	4	5	6	7
1	0	79.26362	43.81662	65.41121	69.38171	31.28206	183.3204
2	79.26362	0	65.30549	79.6104	116.8457	108.9196	115.0083
3	43.81662	65.30549	0	25.87817	60.14051	64.15221	156.7081
4	65.41121	79.6104	25.87817	0	53.72348	81.08877	161.1342
5	69.38171	116.8457	60.14051	53.72348	0	64.7698	207.4395
6	31.28206	108.9196	64.15221	81.08877	64.7698	0	210.0761
7	183.3204	115.0083	156.7081	161.1342	207.4395	210.0761	0

表 3.2: DTW による距離行列の一部

	1	2	3	4	5	6	7
1	0	79.26362	43.81662	65.41121	69.38171	31.28206	183.3204
2	79.26362	0	64.70825	79.6104	116.8457	104.5083	115.0083
3	43.81662	64.70825	0	25.87817	60.14051	64.15221	156.7081
4	65.41121	79.6104	25.87817	0	53.40195	81.08877	161.1342
5	69.38171	116.8457	60.14051	53.40195	0	64.74831	207.4395
6	31.28206	104.5083	64.15221	81.08877	64.74831	0	204.1189
7	183.3204	115.0083	156.7081	161.1342	207.4395	204.1189	0

第 4 章　評価実験

本章では，データの分析や提案手法の評価実験を行う．なお，本研究ではクラスタの数cは 2 から 50 までの範囲で入力を行い，kの値は 2 で入力している．

4.1　有用性指標

本稿においては，元データD_2または合成データD_3と加工後データとの距離誤差を有用性と定義する．その際，簡易手法はユークリッド距離により加工を行っているため，各ユーザのユークリッド距離の総和の平均を有用性とし，提案手法は DTW により加工を行っているため，各ユーザの DTW 距離の平均を有用性としている．

4.2　簡易手法と提案手法の比較

まず，7 月 1 日時点で簡易手法に劣らないかどうかを検証する．クラスタ数cについての各手法の有用性の関係を図 4.1 に示す. 有用性は距離誤差の平均なので低ければ低いほど有用性が高い．いずれもc=30 から 40 のあたりで最小値となっている．その中でもc=40 の時にD'_{kmeans_Dtw}が全体の最小値である 12.0 を記録し，次点で低い数値はc=39 の時のD'_{kmeans_Euc}の 12.4 であった. また, の時のD'_{kmeans_Dtw}(c=40) と, の時のD'_{kmeans_Euc}(c=39) の距離誤差をユーザごとに調べたところ，ユーザごとにD'_{kmeans_Dtw}がD'_{kmeans_Euc}よりも距離誤差が少なくなる割合は，47 %だった．僅かな差で割合では負けているものの，平均値で見ると簡易手法に劣らないどころか僅かな差ではあるが提案手法が 3.2%程よい有用性を示していた.

次に，1 つのクラスタが持つユーザ数の推移（c=40）を図 4.2 に示す．群平均法が c-平均化法と比べてクラスタ内の人数の最大数が大きく，更にクラスタ内のユーザ数が 1 となり削除されてしまうクラスタの数が多くなっている．それに対して，c-平均法は 2 人以上のユーザを所持しているクラスタが多く，最も大きいクラスタでも 5 や 6 であり，クラスタの大きさの差はそこまでない.

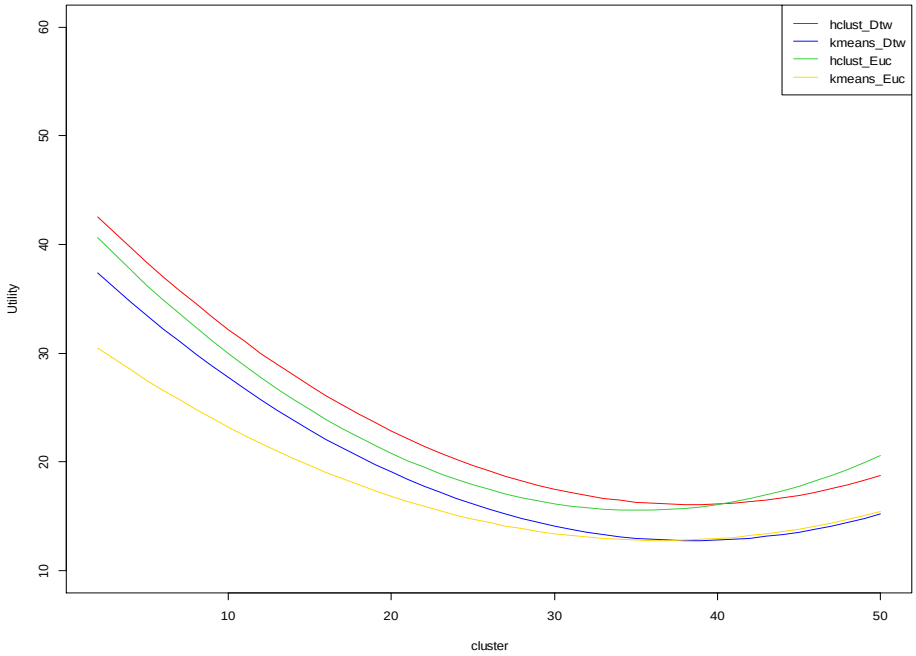

図 4.1: 有用性とクラスタ数の関係

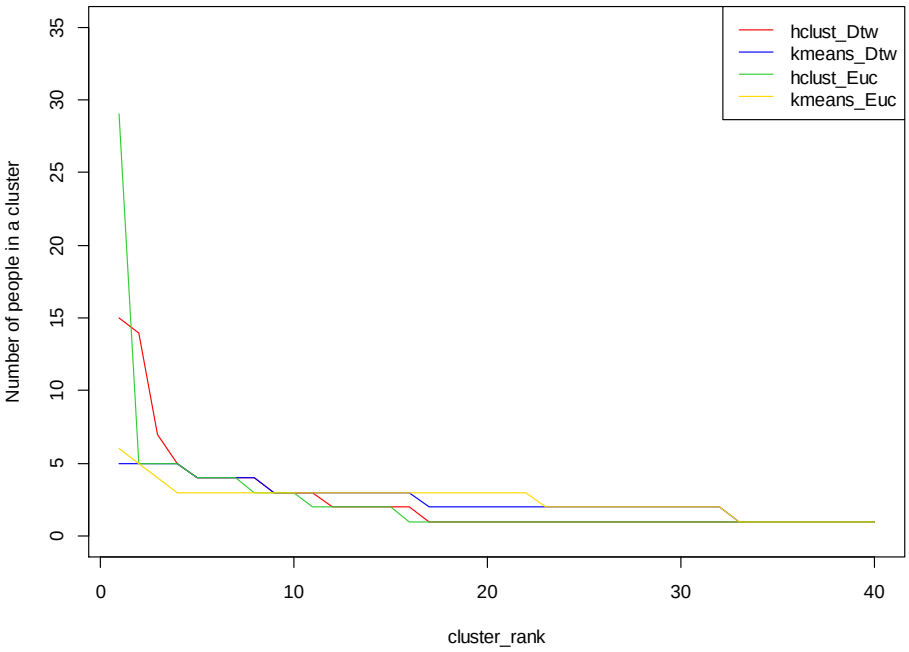

図 4.2: 1 クラスタが持つユーザの数

4.3　加工前後のデータの差異

4.3.1　元データと加工データの緯度分布の差

図 4.3 にD'_{kmeans_Dtw}（c=40）による加工後のデータと元データD_2との緯度の分布，図 4.4 にD'_{kmeans_Euc}（c=40）による加工後のデータと元データD_2との緯度の分布を示す．

中心部を緯度が 35.6 以上 35.75 以下の範囲と定義した場合，元データの中心部の割合は 64.8%，D'_{kmeans_Dtw}の中心部の割合は 68.2%，D'_{kmeans_Euc}の中心部の割合は 67.9%となった．どちらの手法もクラスタ内で加工を行うため，約 3.1%から 3.4%ほど中心に寄る結果になった．いずれの手法でも割合に大きな差は無かったが，DTW による加工後データの方が 0.3%ほど中心に寄っていた．

4.3.2　元データと加工データの移動量分布の差

図 4.5 にD'_{kmeans_Dtw}（c=40）による加工後のデータと元データD_2との移動量の分布，図 4.6 にD'_{kmeans_Euc}（c=40）による加工後のデータと元データD_2との移動量の分布を示す．

どちらの手法も右端に存在する移動量が多いユーザの割合は変わらなかった．移動量が 10km 以下のユーザを移動量が少ないユーザとすると，元データの移動量が少ない割合は 29%であり，DTW による加工後データでは割合が 17%増え，ユークリッド距離によるデータでは割合が 15%減っていた．

実際に DTW の加工法について，1 クラスタを抜き出して加工前のユーザの緯度のプロットを図 4.7 に，D'_{kmeans_Dtw}（c=40）による加工後のユーザの緯度のプロットを図 4.8 に示す．赤線と青線と緑線は同クラスタに分類されたユーザを示す．

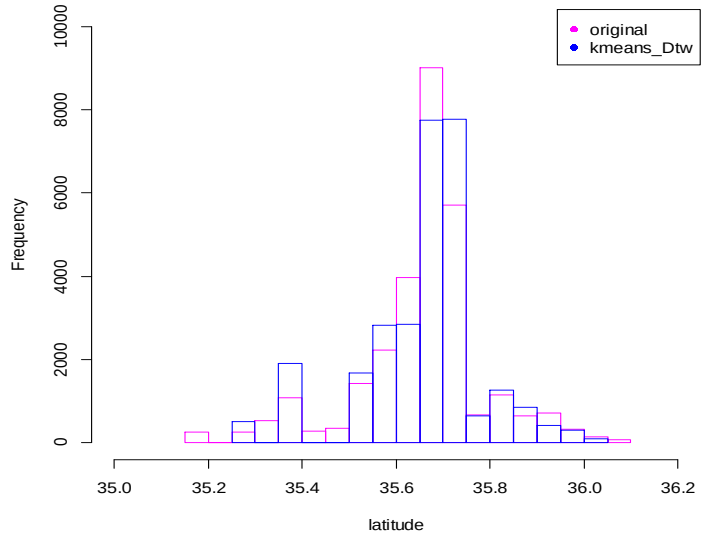

図 4.3: D'_{kmeans_Dtw}と元データD_2の緯度分布

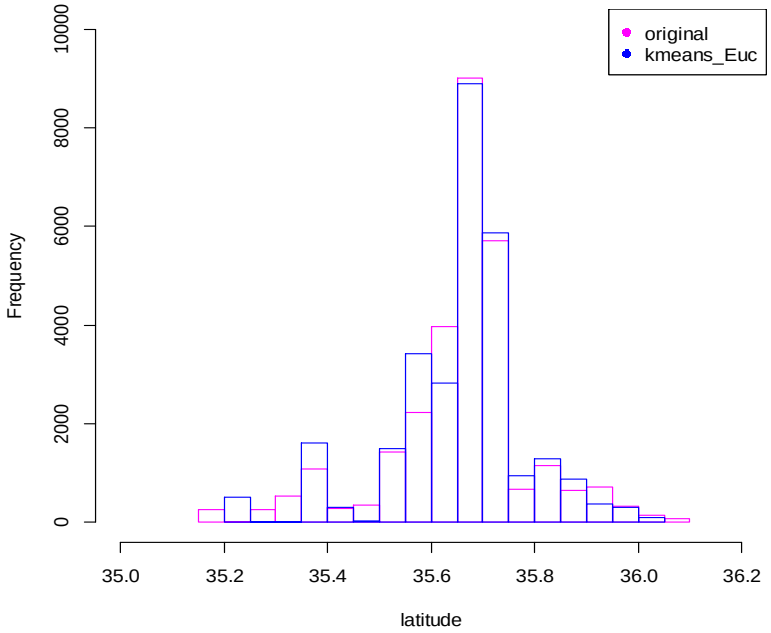

図 4.4: D'_{kmeans_Euc}と元データD_2の緯度分布

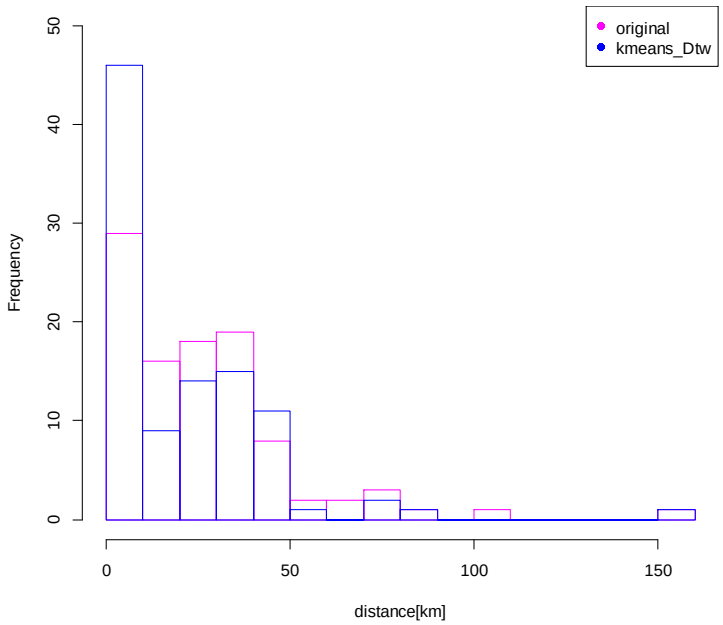

図 4.5: D'_{kmeans_Dtw}と元データD_2の移動量分布

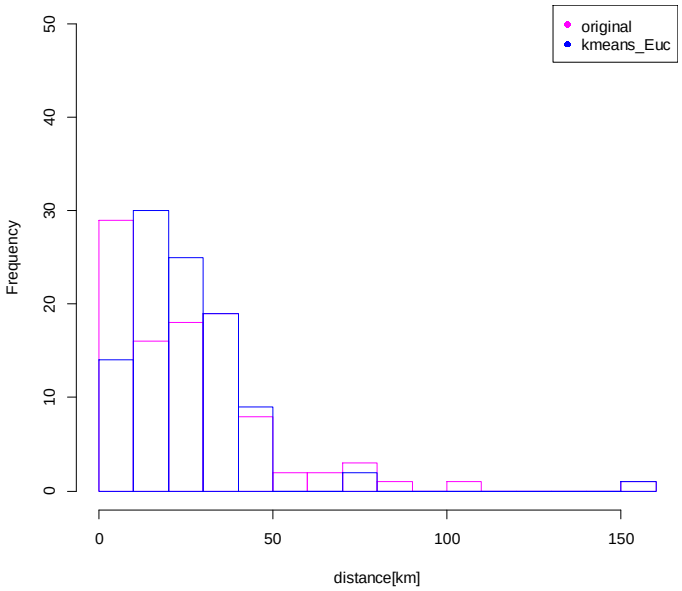

図 4.6: D'_{kmeans_Euc}と元データD_2の移動量分布

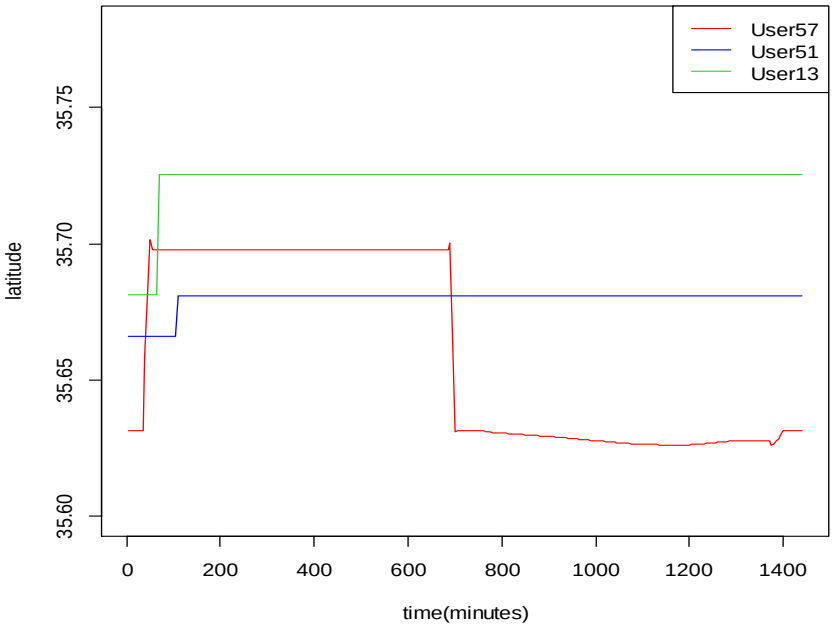

図 4.7: 元データD_2の緯度のプロット

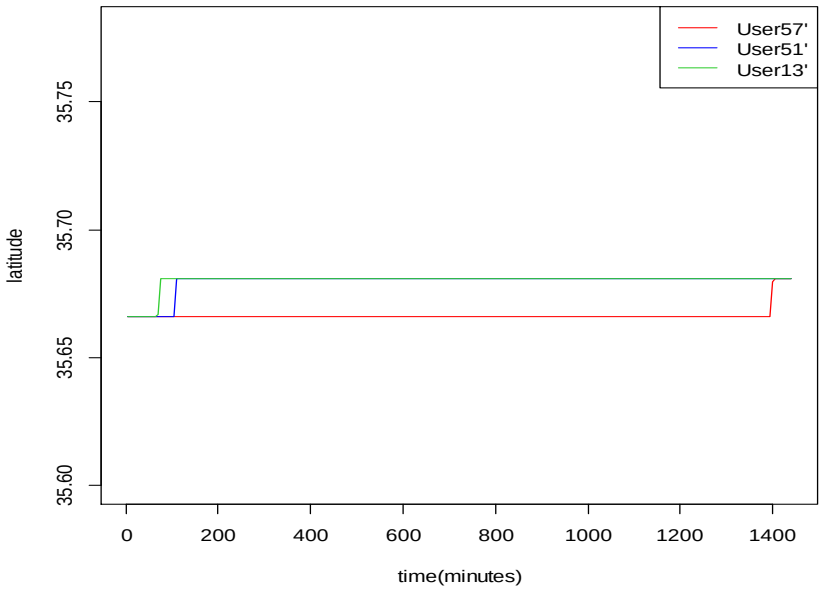

図 4.8: D'_{kmeans_Dtw}の緯度のプロット

4.4　合成データにおける簡易手法と提案手法の比較

次に，本手法が同一ユーザの日による挙動に対してロバストであるか検証する．合成データにおけるクラスタ数cについての各手法の有用性の関係を図 4.9 に示す．

いずれも図 4.1 と同様にc=30 から 40 のあたりで距離誤差が最小となっている．最小値についても 4.2 節同様に，D'_{kmeans_Dtw}（c=40）が全体の最小値である 13.4 を記録し，次に小さい手法はD'_{kmeans_Euc}（c=39）で 17.5 であり，平均値の最小値においては提案手法が 23.4%程優っていた． 4.2 節とは変わって DTW とユークリッド距離に顕著な差が見られた．ユーザごとにD'_{kmeans_Dtw}がD'_{kmeans_Euc}よりも距離誤差が少なくなる割合は，58 %にまで上がった．c-平均法，群平均法のそれぞれのクラスタリング手法においても DTW による加工方法がユークリッド距離による加工方法よりも距離誤差が小さくなった．

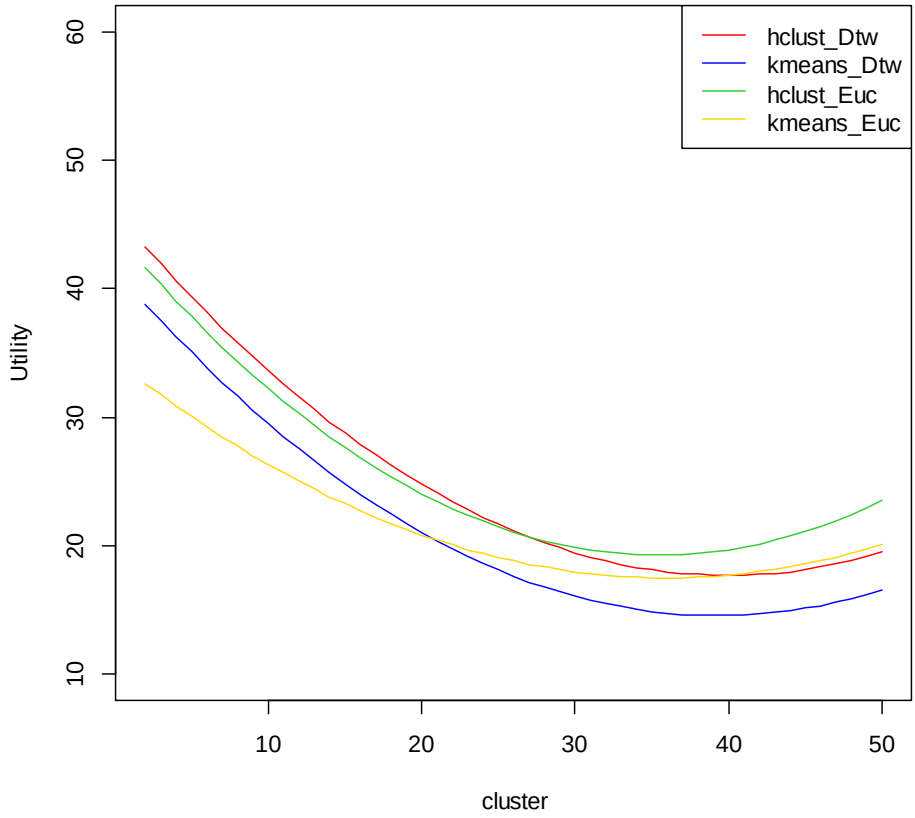

図 4.9: 合成データにおける有用性とクラスタ数の関係

4.5 合成データにおける加工前後のデータの差異

合成データにおいては DTW による加工方法とユークリッド距離による加工方法の有用性に大きな差があったため，D'_{kmeans_Dtw}のみに焦点を当ててデータの分析を行う．

4.5.1 合成データと加工データの緯度分布の差

図 4.10 に元データD_2と合成データD_3との緯度の分布を，図 4.11 にD'_{kmeans_Dtw}（c=40）による加工後のデータと合成データD_3との緯度の分布を示す．

元データの中心部の割合は 64.8%%で，合成データの中心部の割合は 66.8%へと変化した．元データで頻度が高かった座標に少し集まったと考えられる．

4.5.2　合成データと加工データの移動量分布の差

図 4.12 に元データD_2と合成データD_3との移動量の分布を，図 4.13 にD'_{kmeans_Dtw}(c=40)による加工後のデータと合成データD_3との移動量の分布を示す．

先述したように合成データは元データの滞在時間が長い時間帯を引き伸ばして作成しているため，元データと比べて移動量が少ない割合が 15%大きくなっている．図 4.13 に関しては，合成データも DTW による加工データも移動量が少なくなっているため，全体の分布が似た形になっている．

4.6　考察

4.2 節において，c-平均法と群平均法で大きく有用性に差が出たのは，群平均法は階層的クラスタリングであり，貪欲的に分類していくことから同クラスタ内のユーザ間の距離が開いてしまったためと考えられる．

4.3 節において，DTW とユークリッド距離で分布に差異があるのは，加工する方法に原因があると考えられる．ユークリッド距離による加工法は時刻毎の平均値をとるため，もともと動いていないユーザと動いているユーザとの加工になった場合，加工後は動いているユーザとなったため移動量が増えたと考えられる．それに対して DTW による加工法は，加工される側のユーザの複数セルとパスがつながることが多数あり，加工後にその複数のセルが複数時刻をまたいで同じ数値になったために移動量が下がったと考えられる．

また，図 4.7 においてピンが刺さっているのは青線で示されたユーザ 51 であり，赤線のユーザ 57 と緑線のユーザ 13 が青線との DTW 距離が 0 になるように加工されている．赤線に着目した場合，元のデータでは移動量があるのに対して加工後のデータは移動量が格段に減っている．このような現象により，図 4.5 のように移動量が小さいユーザが増えたと考えられる．

緯度分布や移動量分布に差があっても DTW によって距離を算出する際のパスのつなぎ方が合成データであっても元データとあまり変わらないため，有用性に大きな変化が出なかった．これにより，提案手法は滞在時間に差があっても行動パターンがほぼ一緒である同一ユーザに対して有用性を保証する手法であることが明らかになった．

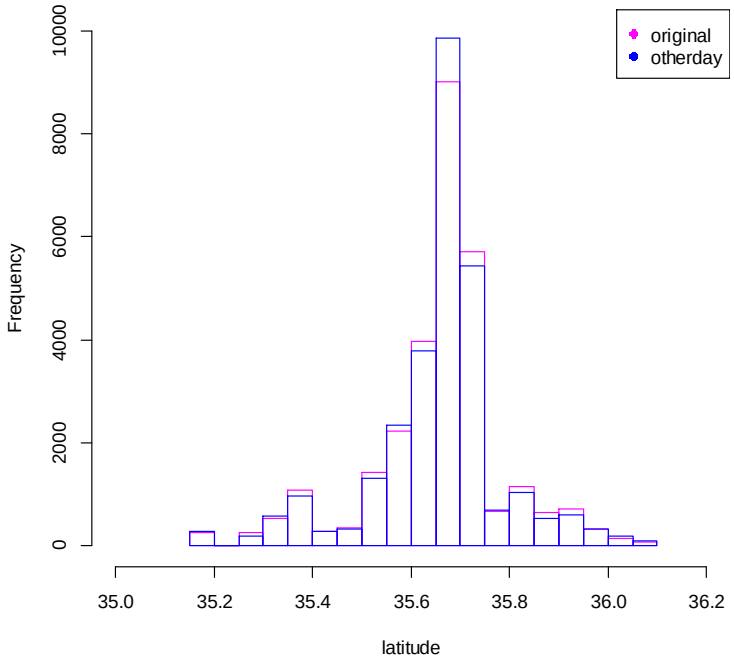

図 4.10: 元データD_2と合成データD_3の緯度分布

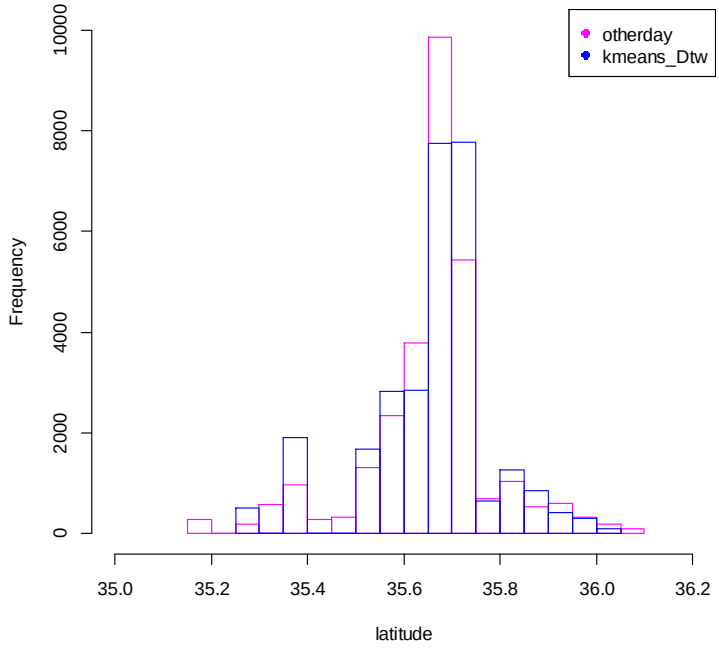

図 4.11: D'_{kmeans_Dtw}と合成データD_3の緯度分布

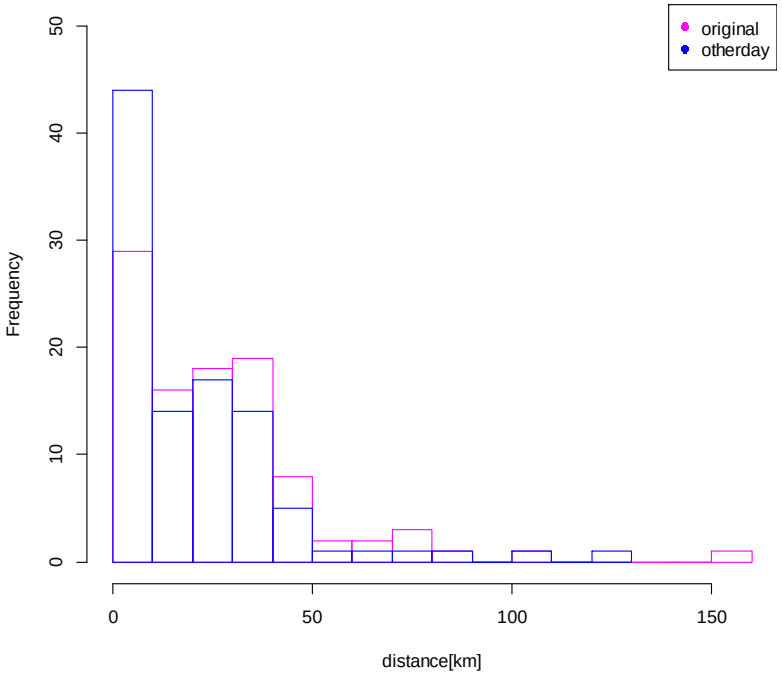

図 4.12: 元データD_2と合成データD_3の移動量分布

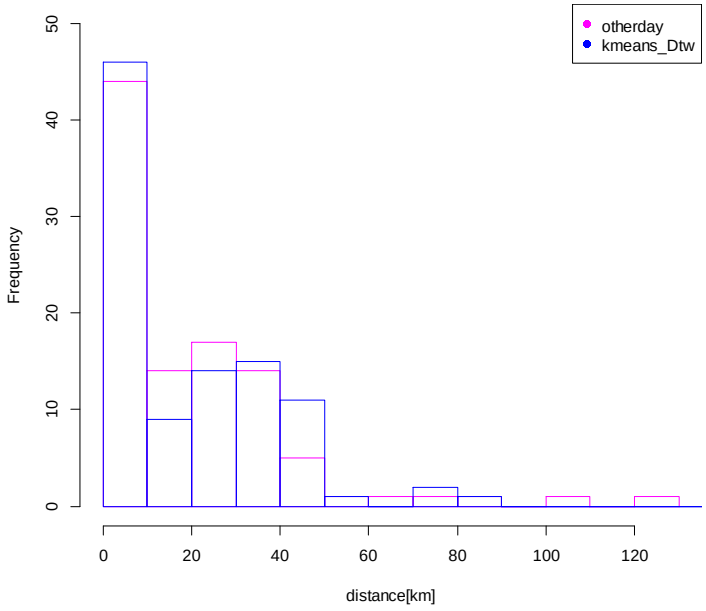

図 4.13: D'_{kmeans_Dtw}と合成データD_3の移動量分布

第５章　まとめ

本稿では，元データとは異なる日を想定した合成データセットにおいても有用性を保てるような加工手法を提案した.

評価実験の結果，元データにおける有用性については，距離誤差が小さいユーザの割合から見ると提案手法が 6%程劣っている結果であったが距離誤差の平均値の最小値を見ると，簡易的な手法の有用性に 3.2%程ではあるが，優っている結果を出すことができた．その上で，合成データにおいては，距離誤差が小さいユーザの割合から見ると提案手法が 16%優っており，距離誤差の平均の最小値においても提案手法の有用性が簡易手法と比べて 23.4%程高くなることを示すことができた.

以上により，今まで通り元データとの有用性に対しては従来手法とさほど変わらない有用性を持っていることを示すことができたうえに，1.3 節であったような事例に対しても有用性を保証することができた.

参考文献

[1] L. Sweeny, “k-anonymity: a model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557–570, 2006.

[2] 日経 xTECH,「Suica 履歴販売」は何を誤ったのか, https://tech.nikkeibp.co.jp/it/article/NC/20131010/510322/, refered in October 1, 2018.

[3] 時事ドットコム, リクナビの「内定辞退率」販売問題, https://www.jiji.com/jc/v7?id=201908rikunabi, refered in December 28, 2019.

[4] 個人情報保護委員会, 個人情報保護法等, https://www.ppc.go.jp/personalinfo/, refered in Junuary 30, 2019.

[5] Tamir Tassa, Arnon Mazza, Aristides Gionis, ” k-Concealment: An Alternative Model ofk-Type Anonymity”, Transaction on Data Privacy 5:1 (2012), pp.189-222.

[6] 山添貴哉, 面和成,「次元削減を利用した匿名化データに対する有用性と安全性の評価」, コンピュータセキュリティシンポジウム 2019(CSS2019), pp. 1277-1282, 2019.

[7] 前田若菜, 山岡裕司,「属性推定攻撃を抑止可能なプログラム送付型匿名化方式の提案」, 情報処理学会, コンピュータセキュリティシンポジウム 2018(CSS2018), pp. 913-919, 2018.

[8] 正木彰伍, 長谷川聡, 千田浩司,「時空間におけるクラスタリングを用いた軌跡情報の k-匿名化法」, 情報処理学会, コンピュータセキュリティシンポジウム 2016 (CSS2016), pp. 921-928, 2016.

[9] 正木彰伍,「時空間クラスタリングを用いた軌跡情報 k-匿名化法の位置精度向上に関する考察」, 電子情報通信学会, 暗号と情報セキュリティシンポジウム SCIS 2017 予稿集, 3A4-4, 2017.

[10] 正木彰伍,「攻撃者のモデル化を用いた軌跡情報の匿名性評価法」, 情報処理学会, コンピュータセキュリティシンポジウム 2017 (CSS2017), pp. 143-150, 2017.

[11] 疋田敏郎，山口理恵，「移動履歴からの個人特定とそのリスクについて」，情報処理学会，コンピュータセキュリティシンポジウム 2017 (CSS2017)，pp. 1180-1187，2017.

[12] 河内尚，鈴木貴之，吉野雅之，佐藤尚宣，兼平晃，「車両移動データ利活用時における k-匿名化技術の有用性評価」，電子情報通信学会，暗号と情報セキュリティシンポジウム SCIS 2018 予稿集，3C4-1，2018.

[13] 森駿文，菊池浩明，「歩容データの DTW 距離に基づく個人識別手法の提案と外乱に対する評価」, マルチメディア, 分散, 協調とモバイルシンポジウム(DICOMO 2018)，pp. 672-680，2018.

[14] 株式会社ナイトレイ，東京大学 CSIS との研究活動成果として SNS 解析データを元とした「疑似人流データ」を無料公開，http://nightley.jp/archives/1954，refered in Febrary 12，2019.

[15] Mobmap，https://shiba.iis.u-tokyo.ac.jp/member/ueyama/mm/，refered in July 30，2019.

謝辞

本稿は多くの方からのご指導，ご協力の上で完成したものとなっています．

指導教員である明治大学総合数理学部先端メディアサイエンス学科の菊池浩明教授からは，学部生 2 年から修士 2 年までの 5 年間に渡り，大変多くのご指導をいただけたと思っております．私は，自分の意思があまり強くない方なので研究テーマで右往左往したり，なかなか進捗が挙げられない時期が続いたりしてしまったのですが，それでも見守り，時には愛のある厳しい言葉で火をつけていただけたこと，大変感謝しております．

また，合同ゼミを通じて助言をいただいた，静岡大学創造科学技術大学院の西垣正勝教授，静岡大学情報学部情報科学科講師の大木哲史先生，東京電機大学理工学部理工学科情報システムデザイン学科の稲村勝樹先生に心より感謝いたします．同じセキュリティ系ではありながらも，専門外の発表を聞いたり，質問をしていただけたりしたのはとても刺激になりました．

明治大学大学院の半澤映拓氏と，明治大学大学院卒業生である森駿文氏には本研究を進めるうえで，とても大きなヒントをいただくことができました．この場をもって感謝の意を申し上げるとともに，今後ますますのご活躍をお祈りしております．

著者の大学生活，大学院生活を支えてくれた家族と友人，全ての方々に深く御礼申し上げます．