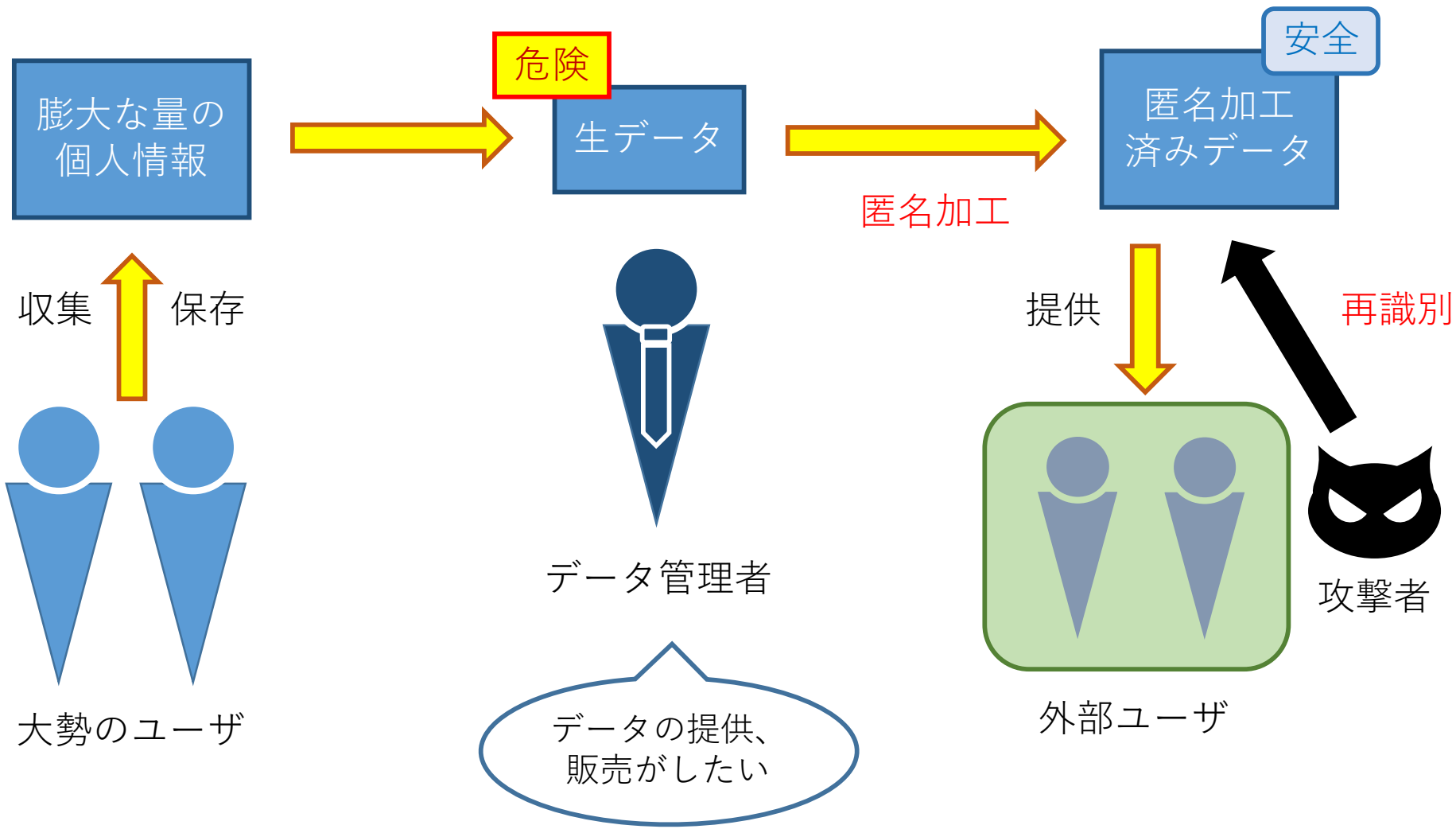
先端メディアサイエンス専攻
2019年度　修士論文発表会

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

二谷太郎

菊池研究室

匿名加工・再識別とは？


膨大な量の
個人情報
危険
生データ
匿名加工
安全
匿名加工
済みデータ
収集
保存
提供
再識別
データ管理者
攻撃者
大勢のユーザ
外部ユーザ
データの提供、
販売がしたい

加工するデータの例

- 乗降履歴（SuicaやPasmoなど）
- 購買履歴データ(Amazonや楽天市場など)
- 医療カルテ
- 軌跡情報(人流データ，GPSやWi-fiの履歴など)

↑これに焦点を置く！

先行研究

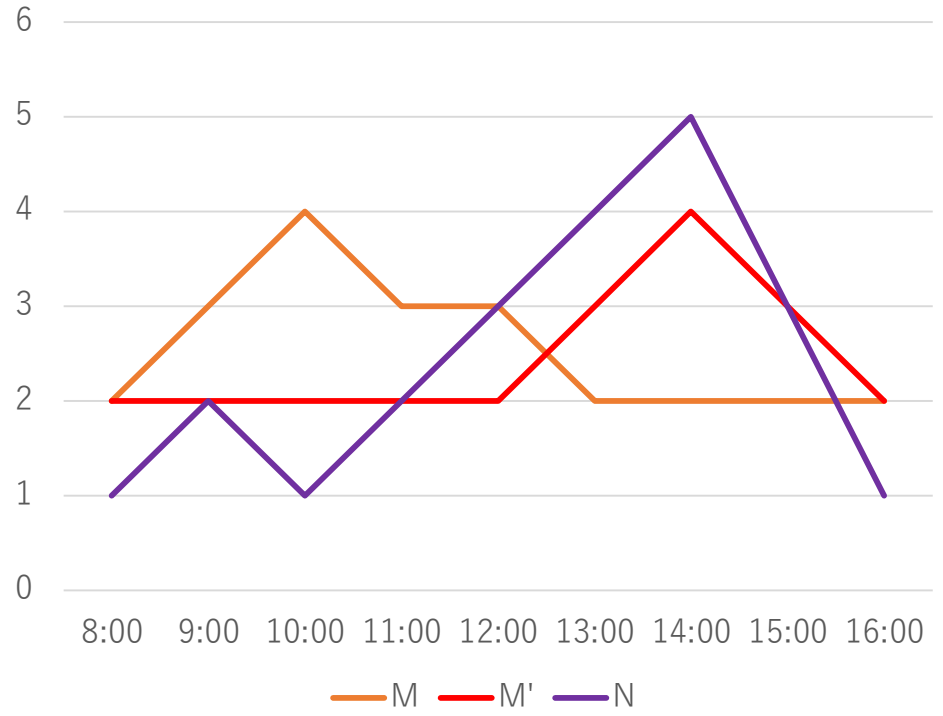
- 正木らによる，時空間におけるクラスタリングを用いた軌跡情報のk-匿名化法[1]
  - 緯度・経度，時間の3次元でクラスタリングすることにより人通りの少ない領域の軌跡情報もk-匿名性を充足させている.
  - k-平均法によりクラスタリングして，ミクロアグリケーションすることによりk匿名性を満たしている．（以後，k-平均法とk-匿名性のkの混乱を防ぐためにk-平均法をc-平均法と呼称する）
  - クラスタリングの距離の導出方法はユークリッド距離.

[1]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.

ユークリッド距離を用いた時の問題点

- MとM'は同ユーザの別日であり，同じルートを異なる時間で通っている，Nは別のユーザ．

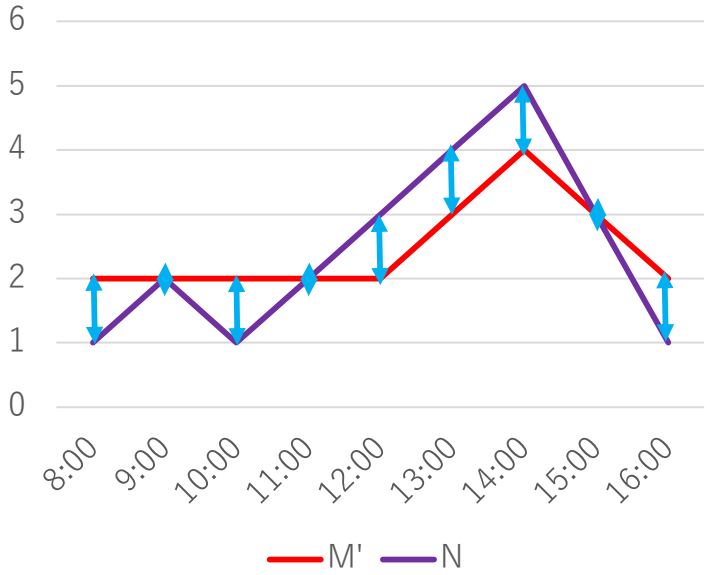
	時刻								
	8	9	10	11	12	13	14	15	16
M	2	3	4	3	3	2	2	2	2
M'	2	2	2	2	2	3	4	3	2
N	1	2	1	2	3	4	5	3	1



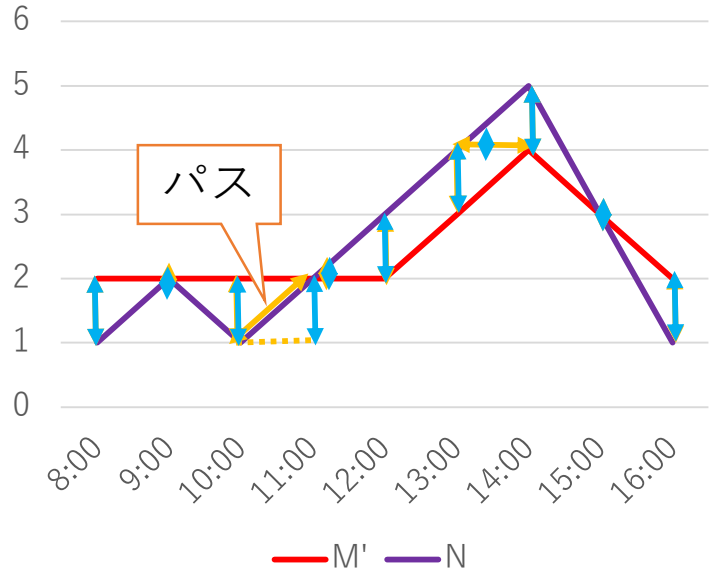
M-M'のユークリッド距離の総和：9，　M'-Nのユークリッド距離の総和：6

本研究の新規性

- **研究目的**
  - 前述のユークリッド距離の問題をDTW距離により解決する.
  - 動的時間伸縮法（Dynamic Time Warping）
    - 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方法.

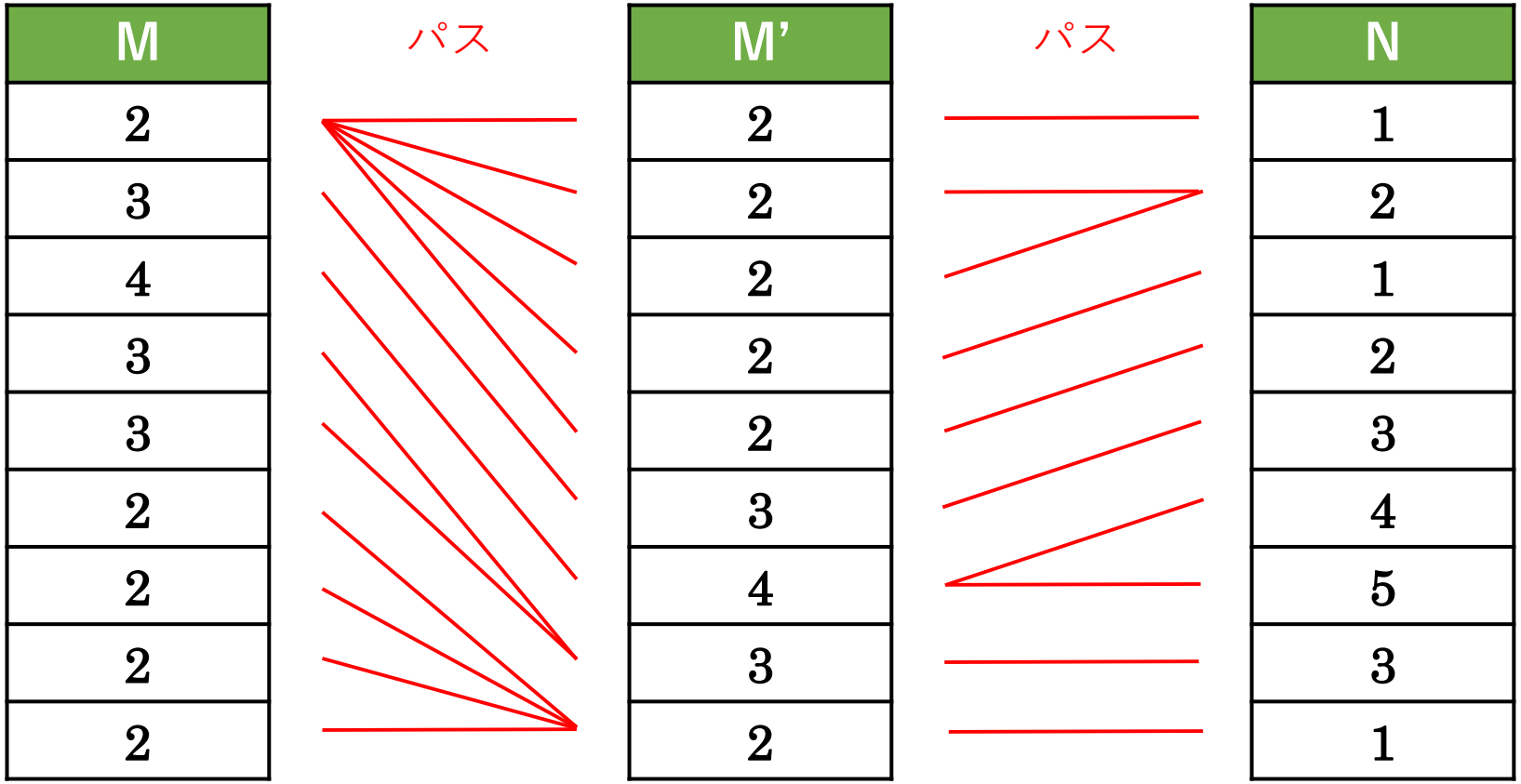


ユークリッド



DTW

例におけるDTW距離のパス



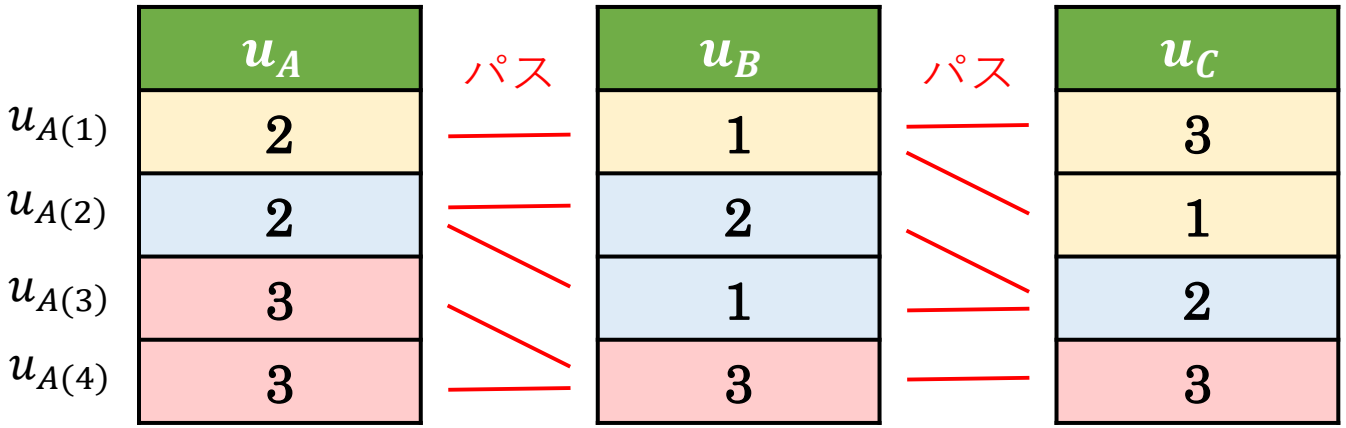
加工手法

1. 既存研究を基にした簡易手法
2. DTWを用いた提案手法

- 実装方法は以下の通り
  1. 簡易手法にはユークリッド，提案手法にはDTWを用いて緯度経度の2次元からなる100×100の距離行列を作成する.
  2. 距離行列を元に，c-平均法または群平均法によるクラスタリングを行う.
  3. k人以上のユーザを含んでいないクラスタを削除.
  4. クラスタ毎に，ユーザのデータにそれぞれの加工を施す.

説明用データ

- 100ユーザの中から，ユーザA，ユーザB，ユーザCの3人のみがクラスタリングによってある同一クラスタに分類されたとする．
- 各ユーザ間の距離は以下の通り．

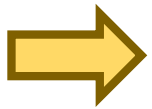


	ユークリッド	DTW
A-B	3	2
B-C	4	3
A-C	3	2

簡易手法

- クラスタ内のユーザの同時刻における緯度経度を分類されている全ユーザの平均値に置き換える.

	u_A	u_B	u_C
$u_{A(1)}$	2	1	3
$u_{A(2)}$	2	2	1
$u_{A(3)}$	3	1	2
$u_{A(4)}$	3	3	3

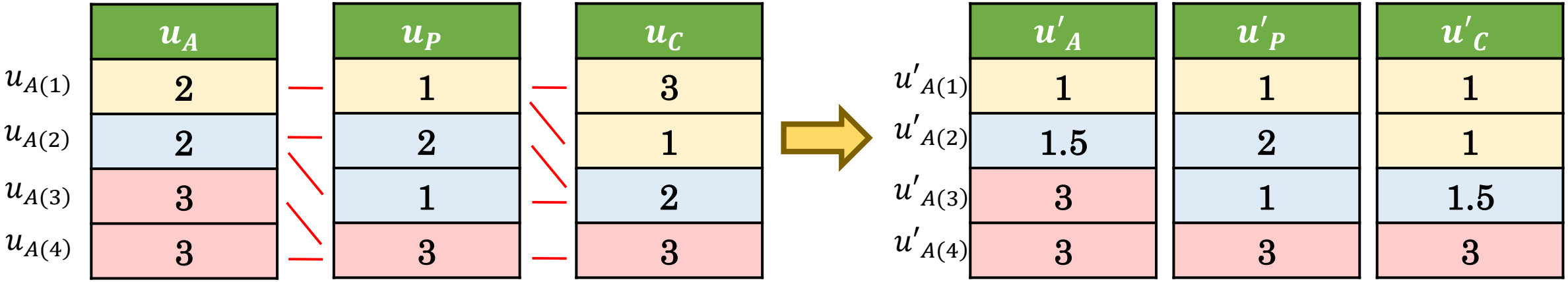


	u'_A	u'_B	u'_C
$u'_{A(1)}$	2	2	2
$u'_{A(2)}$	1.67	1.67	1.67
$u'_{A(3)}$	2	2	2
$u'_{A(4)}$	3	3	3

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

DTWを用いた提案加工手法

- クラスタ毎に完全にランダムな1ユーザにピンを立て(ピンが刺さったユーザには加工を施さない)，ピンが立っているユーザの対応パスの数値に置き換える，複数パスに対応している場合は，対応パスの平均値をとる．例ではユーザBにピンが刺さったとする．



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

提案手法の適応後のユーザ間のDTW距離

u'_A	u'_P	u'_C
1	1	1
1.5	2	1
3	1	1.5
3	3	3

	加工前	加工後
A-P	2	1
P-C	3	1
A-C	2	0
総和	7	2

使用するデータ

- 疑似人流データ（一部抜粋）[2]
- 簡易的に全ユーザ6432人のうち，100ユーザをランダムに取り出して実験した，5分おきで一人当たり288レコード.

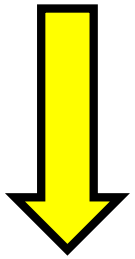
	A	B	C	D	E	F	G	H	I
1	10015	male	2013/12/22 0:00	35.66963	139.767	home	departure	STAY	8
2	10015	male	2013/12/22 9:15	35.66994	139.7665			MOVE	
3	10015	male	2013/12/22 9:20	35.67025	139.7661			MOVE	
4	10015	male	2013/12/22 9:25	35.67055	139.7658			MOVE	
5	10015	male	2013/12/22 9:30	35.67087	139.7656			MOVE	
6	10015	male	2013/12/22 9:35	35.67118	139.7654			MOVE	
7	10015	male	2013/12/22 9:40	35.67149	139.7654			MOVE	
8	10015	male	2013/12/22 9:45	35.67183	139.7657			MOVE	
9	10015	male	2013/12/22 9:50	35.67215	139.766			MOVE	
10	10015	male	2013/12/22 9:55	35.67248	139.7664			MOVE	
11	10015	male	2013/12/22 10:00	35.6728	139.7667			MOVE	
12	10015	male	2013/12/22 10:05	35.67313	139.7671			MOVE	

[2]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.

データの正規化

元データ（正規化前）

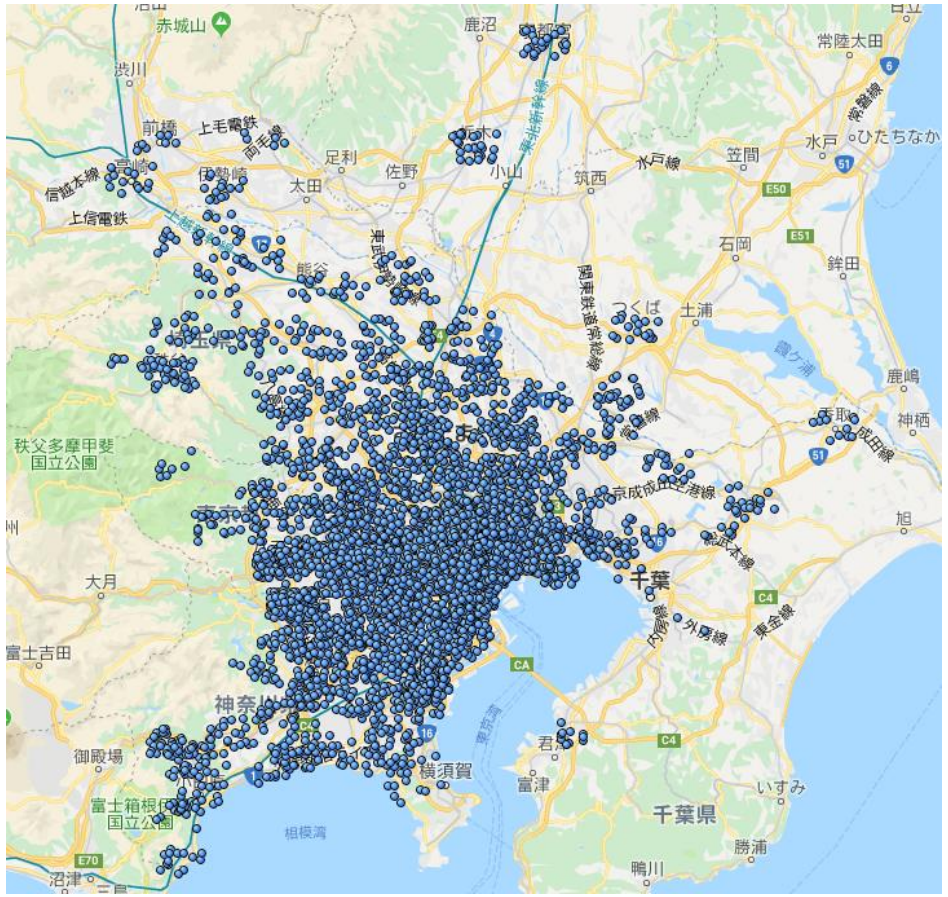
ID	時刻	緯度	経度	滞在情報
3	21:30	35.6679	139.4389	MOVE
3	21:35	35.6657	139.4382	MOVE
3	21:40	35.6660	139.4382	MOVE
3	21:45	35.6653	139.4370	STAY



正規化後

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

2013-07-01の散布図[3]



[3] Mobmap, https://shiba.iis.u-tokyo.ac.jp/member/ueyama/mm/.

6432ユーザ

合成データの作成

- 問題点に挙げられたような元データとは別の日を想定した合成データを作成する.

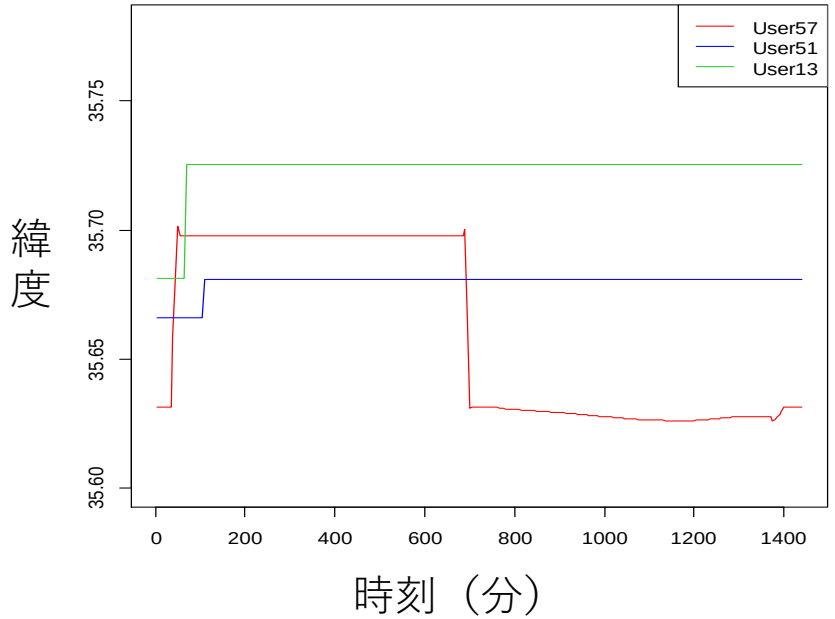
1. 元データを参照して，最も長い時間滞在している場所をユーザごとの家や職場の場所と推定する.

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.

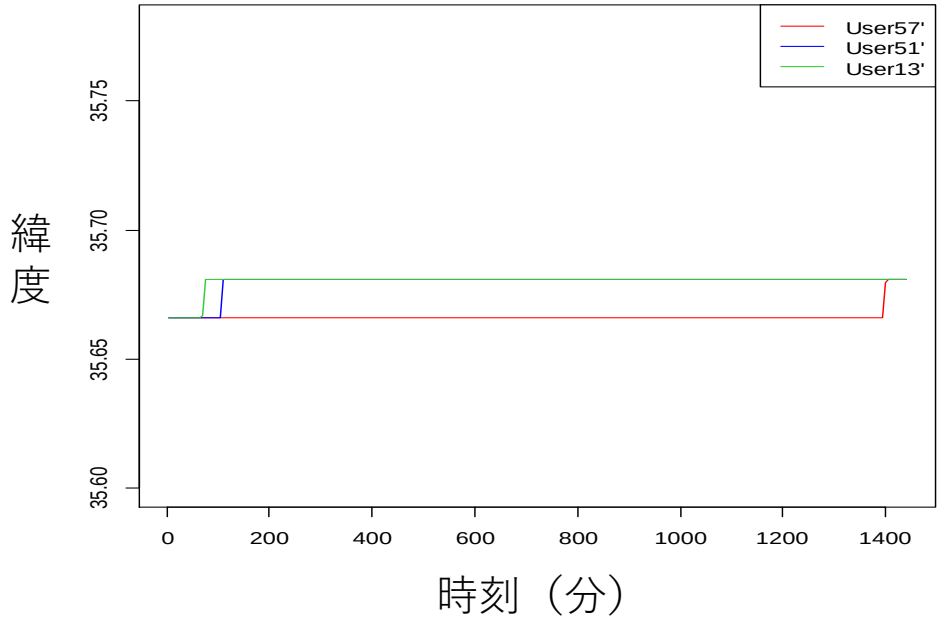
3. 家や職場と推定した以外の時間帯において，乱数（緯度経度において0.03から-0.03）を付加する.

提案手法における加工前後の同一クラスタ内のユーザの緯度の推移

- クラスタ数は40のとき
- ピンが刺さっているユーザは51
- 加工後の各ユーザ間のDTW距離が0になっている.



加工前



加工後

有用性

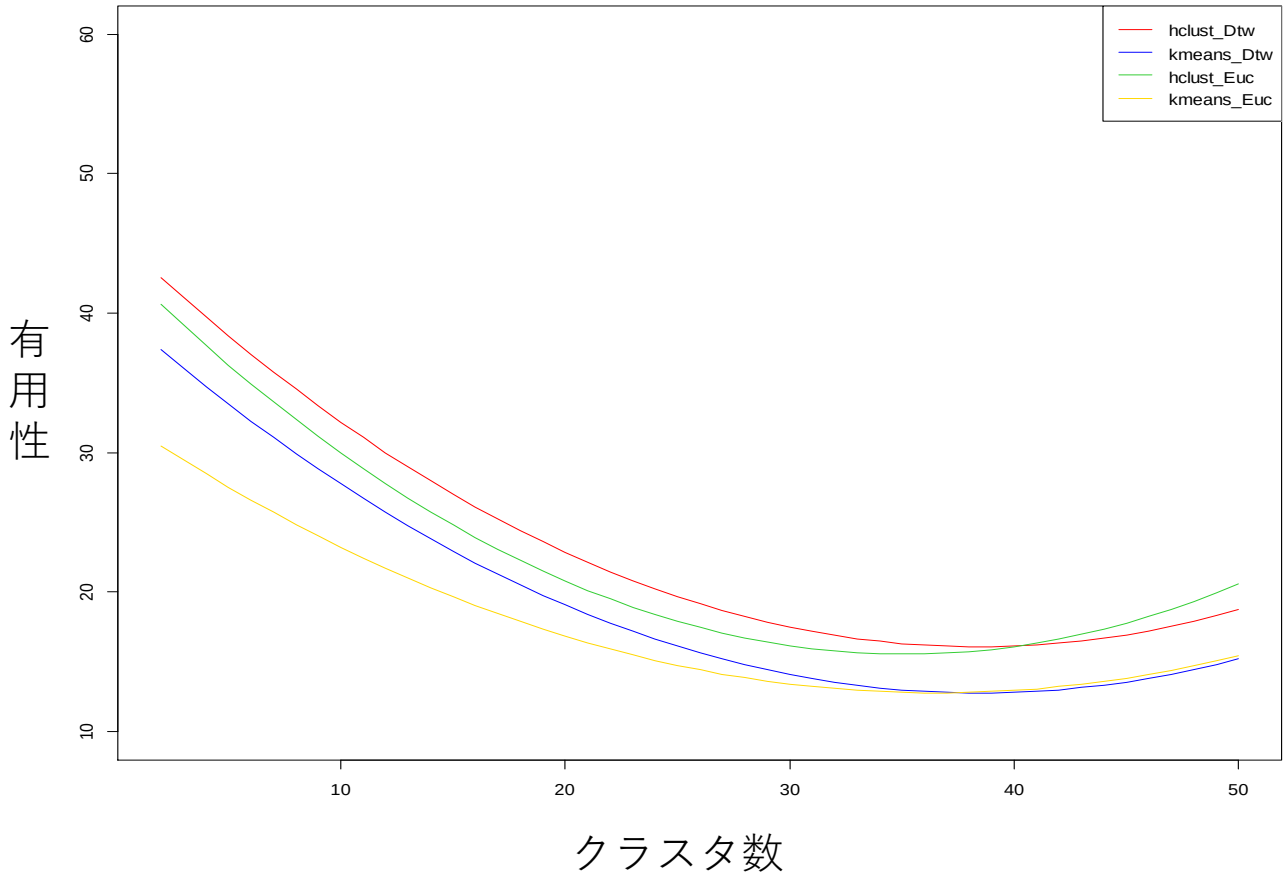
- 本実験ではクラスタの数cは2から50までの範囲で入力を行い，kの値は2で入力している.
- 元データ又は合成データと，加工後データとの同一ユーザの距離誤差を有用性とする.（数値が低い方が有用性が高い）
- 簡易手法の有用性は全ユーザのユークリッド距離の平均絶対誤差.
- 提案手法の有用性は全ユーザのDTW距離の平均絶対誤差.

元データにおける簡易手法と提案手法の有用性の比較

- hclust_Dtwは群平均法×提案手法
- kmeans_Dtwはc-平均法×提案手法
- hclust_Eucは群平均法×簡易手法
- kmeans_Eucはc-平均法×簡易手法

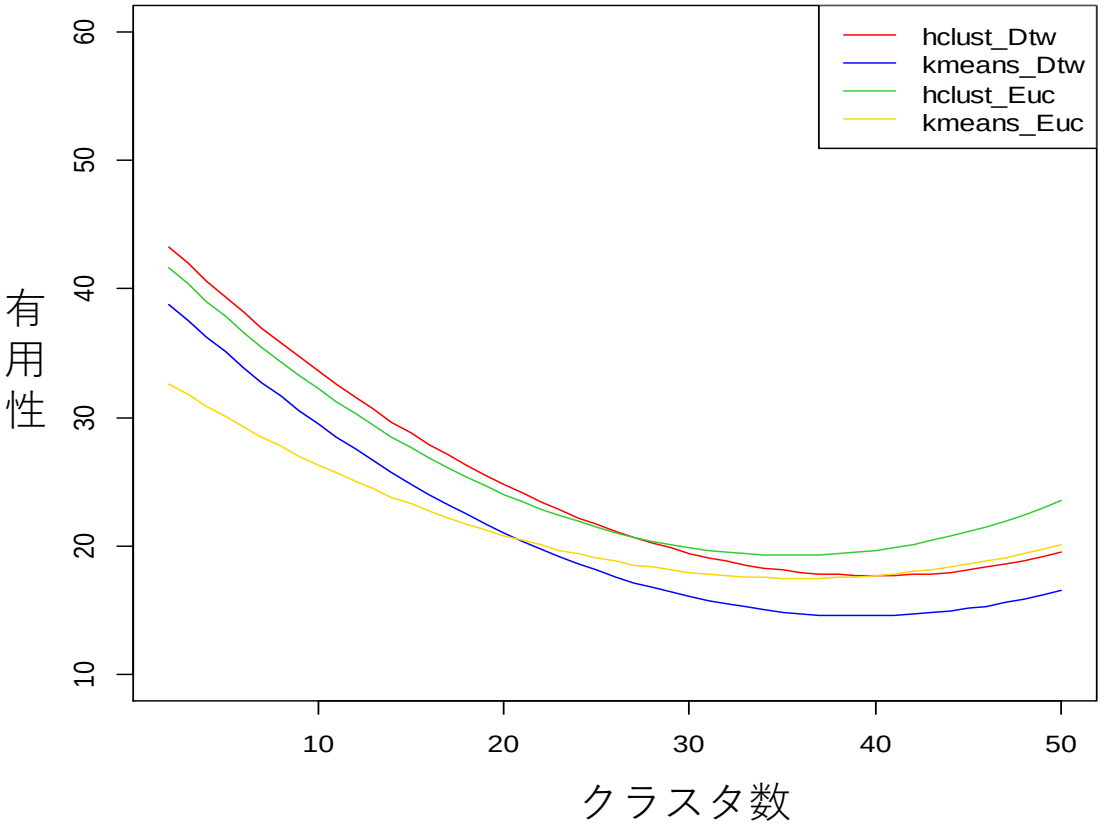
- kmeans_Dtwの平均絶対誤差の最低値：12.0（c=40）
- kmeans_Eucの平均絶対誤差の最低値：12.4（c=39）

- 提案手法(c=40)と簡易手法(c=39)の同一ユーザごとの有用性の勝敗を見たとき，提案手法が簡易手法に優っている割合：47%



合成データにおける簡易手法と提案手法の有用性の比較

- kmeans_Dtwの平均絶対誤差の最低値：13.4（c=40）
- kmeans_Eucの平均絶対誤差の最低値：17.5（c=39）
- 提案手法(c=40)と簡易手法(c=39)の同一ユーザごとの有用性の勝敗を見たとき，提案手法が簡易手法に優っている割合：58%



まとめ

- 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の平均絶対誤差の最小値を見ると，簡易的な手法の有用性に3.2%程ではあるが，優っている結果となった.

- 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，距離の平均絶対誤差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せた.

質疑応答用スライド

DTWの導出方法

1				
2				
1				
	1	**2**	**2**	**1**

- 行と列の差と[i][j-1],[i-1][j],[i-1][j-1]の中で最小のものを足す.

DTWの導出方法

1	1			
2	1			
1	0	1	2	2
	1	2	2	1

- 行と列の差と[i][j-1],[i-1][j],[i-1][j-1]の中で最小のものを足す.

DTWの導出方法

DTW距離

1	1	1	1	0
2	1	0	0	1
1	0	1	2	2
	1	2	2	1

- 行と列の差と[i][j-1],[i-1][j],[i-1][j-1]の中で最小のものを足す.

DTWの導出方法

パス
（最短経路）

1	1	1	1	0
2	1	0	0	1
1	0	1	2	2
	1	2	2	1

- 行と列の差と[i][j-1],[i-1][j],[i-1][j-1]の中で最小のものを足す.

本研究は2次元なのでこのように行っている

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

距離行列

	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

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	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

DTWによる距離行列の一部

提案手法の適応後のユーザ間の距離

u'_A	u'_P	u'_C
1	1	1
1.5	2	1
3	1	1.5
3	3	3

u'_A-u'_P

3	3	2.5	1	1
1	1	1	2.5	3.5
2	1	0.5	1.5	2.5
1	0	0.5	2.5	4.5
	1	1.5	3	3

u'_P-u'_C

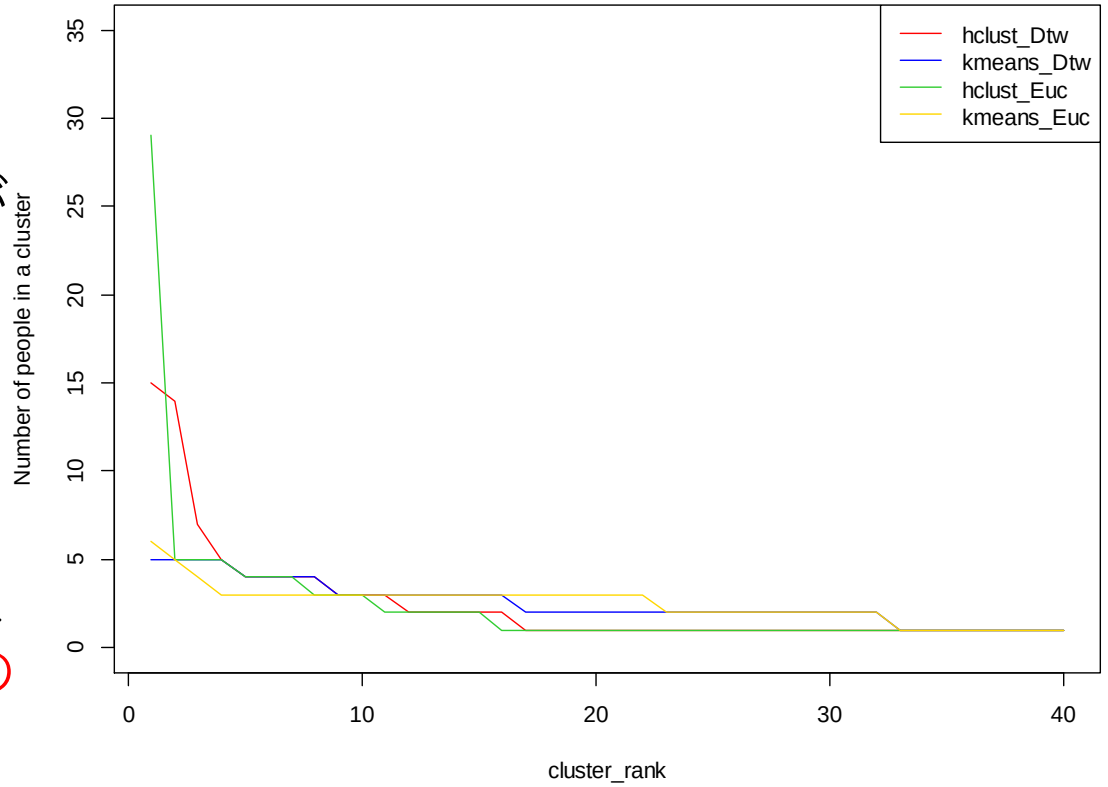
3	2.5	1.5	2.5	1
1.5	0.5	0.5	1	2.5
1	0	1	1	3
1	0	1	1	3
	1	2	1	3

u'_A-u'_C

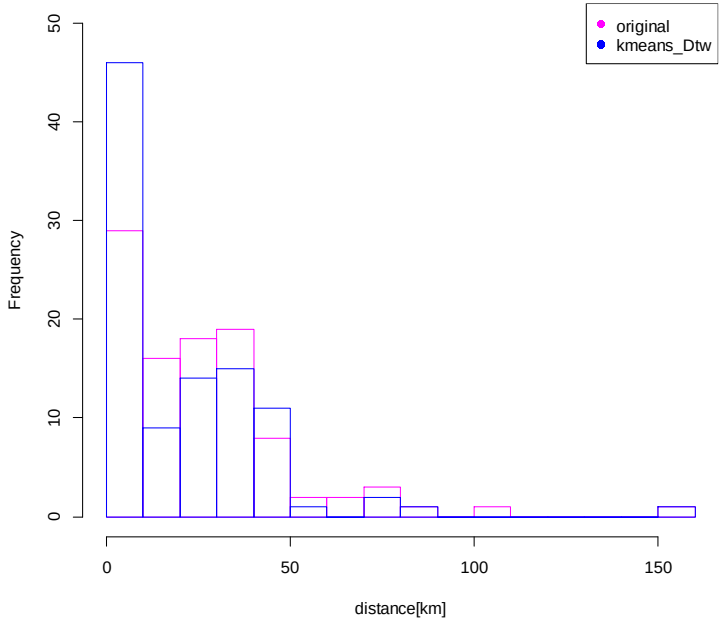
3	2.5	1.5	0	0
1.5	0.5	0	1.5	3
1	0	0.5	2.5	4.5
1	0	0.5	2.5	4.5
	1	1.5	3	3

1つのクラスタが持つユーザ数の推移（c=40）

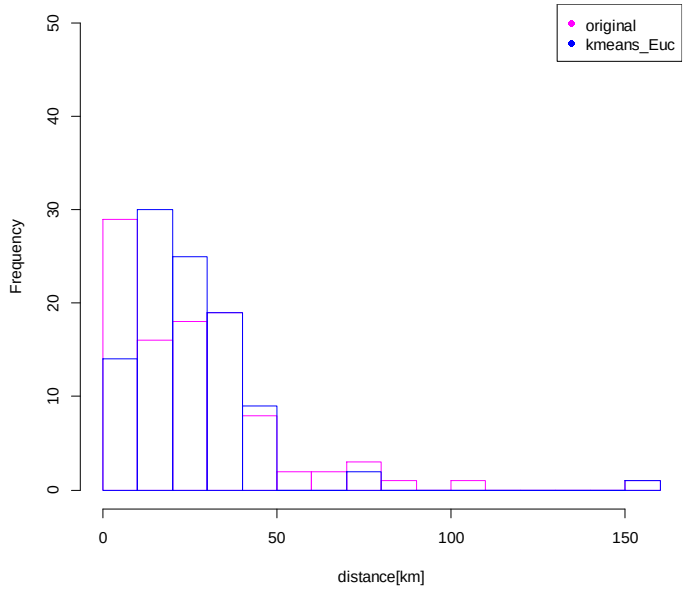
- 群平均法がc-平均化法と比べてクラスタ内の人数の最大数が大きい.
- 群平均法の方がクラスタ内のユーザ数が1となり削除されてしまうクラスタの数が多い.
- c-平均法は2人以上のユーザを所持しているクラスタが多く，最も大きいクラスタでも5や6であり，クラスタの大きさの差はそこまでない.



移動量分布

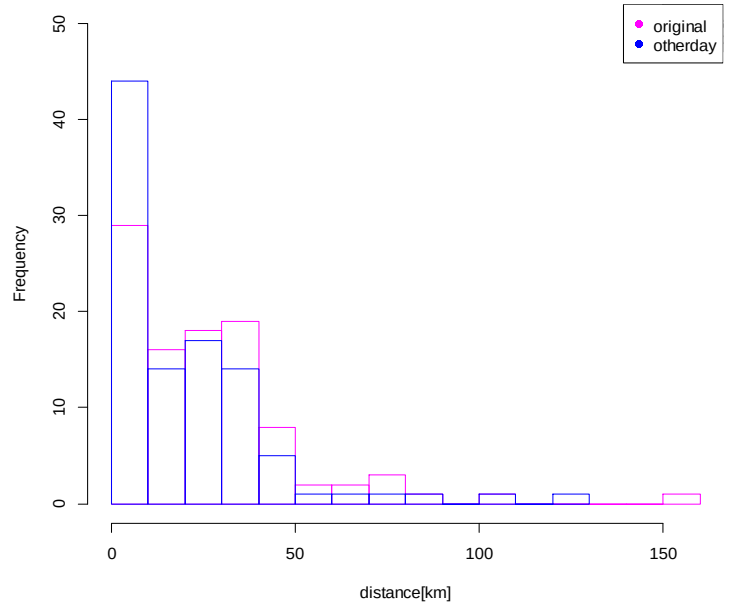


元データと提案手法の移動量分布

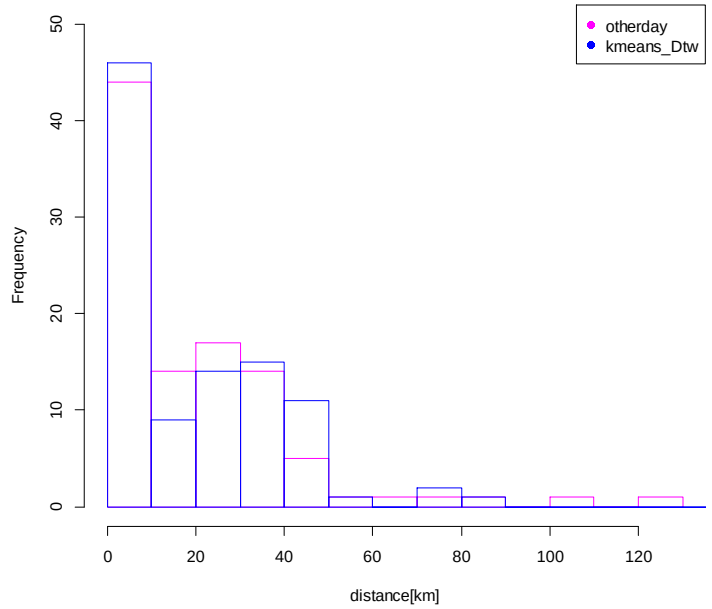


元データと簡易手法の移動量分布

移動量分布



元データと合成データの移動量分布



合成データと提案手法の移動量分布

Q&A

- 削除したユーザの有用性は？
  ＞削除されていないユーザーの最大の距離誤差を代入
- 攻撃者のモデルは？
  ＞元データを持っていて、DTW再識別を用いてくる
- 今回の加工のユースケースは？
  ＞持っていないユーザのデータに対しても有用性がある
- 加工の施行回数による変化は？
  ＞5回行った結果，合成データにおける提案手法の平均絶対誤差の平均（ｃ＝40）は，12.0から11.8となり大きなブレはないと思われる