滞在・閉路性を表現した擬似経路データ生成法 経路データの滞在・閉路性がもたらすリスク評価を目指して

杉山 歩未[1]　香川 椋平[1,a)]　川田 涼平[1]

概要：位置情報を利用したサービスや生活の変化検知など，パーソナルデータの利活用が期待されている．しかし，個人に密接に関わるデータであるため，プライバシーを尊重した取り扱いが求められ，その取り扱いについての検討が進められている．PWSCUP2019 においても，位置情報を想定し，匿名加工および再識別手法の検討が行われた．その結果，自宅位置が推測されることや滞在領域分布による再識別のリスクが検討された．しかしながら，使用された経路データから一部特徴が失われており，検討されていないリスクがあるのではないかと考えた．本稿では，PWSCUP で用いられたデータとその擬似データの元となったデータを比較し，表現されていない特徴量について調査を行った．その上で，その統計的な特徴を維持した擬似経路データ生成手法の提案・評価を行う．さらに，その特徴を含んだデータに対してのリスクの検討について述べる．

キーワード：経路データ，疑似データ，匿名加工

Generation Method for Trace Data with stay/cycle property Aiming for risk assessment by stay/cycle property of Trace Data

Ayumi Sugiyama[1]　Ryohei Kagawa[1,a)]　Ryohei Kawata[1]

Abstract: Services using location information and detection of changes in daily life, utilization of personal data is expected. However, since the data is closely related to the individual, the handling that respects privacy is demanded, and the examination about the handling is advanced. Also in PWSCUP2019, anonymization process and re-identification method were examined assuming position information. As a result, we estimated the home position and the risk of re-identification based on the distribution of stay areas. However, some features were lost from the route data used, and we suspect that there is a risk that it has not been examined. In this paper, we compared the data used in PWSCUP with the data that was the source of the pseudo data, and investigated the lost features. Then, we propose and evaluate a pseudo route data generation method that maintains the statistical characteristics. Furthermore, we discuss the examination of the risk to the data including the feature.

Keywords: Trace Data,Pseudo data,Anonymization

1. はじめに

パーソナルデータの利活用が期待されているが，個人に密接に関わるデータであるためその扱いには慎重さが求められ，研究に利用できるデータの種類やデータ数は限定的である．個人の生活と密接に結びついたパーソナルデータは，近年のビックデータ解析技術等の発展により保護だけでなく利活用が検討されている．しかし，これらの情報は個人の嗜好や個人情報を特定可能なセンシティブな情報が含まれる恐れがある．近年，データ取得・解析技術は急速に発展しており，様々なパーソナルデータが取得可能となっている．複数のデータを組み合わせることで多くの情

[1] セコム 株式会社 IS 研究所
Intelligent Systems Laboratory, SECOM CO.,LTD.
[a)] ryo-kagawa@secom.co.jp

報を推測できるため，どのようなデータがあれば個人特定の危険があるか，利活用に十分な分析が可能か等については未だ検討段階である．加えて，どのようなデータがセンシティブとみなされるかは法務的な基準はもちろんのこと，利用目的や提供先，提供者の感情によっても変わりうる．

個人情報保護のためにデータを匿名化して利用することも考えられるが，過度な匿名化により有用性が損なわれる場合もある．上記のように，保護と利活用の両立の検討にはデータの多様さや主観等の違いも考慮した検討が必要である．その検討のためには多くのデータが必要であるものの，センシティブ情報を含むことや組み合わせ時の影響推定の困難さから，利用できるデータは限られている．利用できるパーソナルデータが限定的であるため，有限なデータから疑似データ生成モデルを作成する手法が注目されている．例えば，データ保護と利活用を両立させる技術検討の場であるプライバシーワークショップ (PWS) が主催するコンテスト (PWSCUP2019[1]) においても疑似データが活用された．PWSCUP は匿名加工と再識別を競うコンテストであり，2015 年より毎年開催されており，PWSCUP2019 では，経路（位置）情報を利用したコンテストが開催された．本コンテストでは，匿名化されたデータの再識別を試みる攻撃者が元トレースの参考となる知識をもつ部分知識攻撃者モデルを採用している．そのため，チームごとに異なる識別対象となるデータ，部分知識となるデータという大量のデータが必要である．また，コンテストの性質上，チーム間で与えられたデータによって不平等が発生しないよう配慮する必要があり，これらの観点からも運営者が分析・生成の制御が容易な疑似データが有効であった．**図 1** に PWSCUP2019 で使用されたデータ生成過程の概要を示す．本コンテストでは図 1 に示すように，ＳＮＳから取得された位置情報データ等を基に生成された有限の疑似人流データ（ナイトレイ [2]）から疑似データ生成モデルを作成し，コンテストで利用する大量のデータを用意している．本コンテストでは，疑似データ生成モデルによって作られたデータを利用して，離散時間の位置情報集合であるトレースデータの匿名化と，匿名化されたトレースデータから ID 識別および加工前の元トレースの推定手法が競われた．PWSCUP2019 の疑似データ生成モデル (以下，既存手法) では，詳細は非公開であるが，マルコフモデルを使用し，ナイトレイの疑似人流データセット（以下，元データ）の時間ごとの人口分布や遷移行列が保存されていることが公開されている [3]．また，既存手法では擬似的なデータ提供者（以下，ユーザ）を作成し，ユーザがもつ属性に応じて経路を生成する．既存手法では，ユーザの特徴として自宅と時刻ごとの在宅確率しか設定されていない．しかし実世界では，自宅以外にも職場や頻繁に訪問する場所のような属性も経路に大きく影響すると考えられる．実際に，元データと既存手法の生成データを比較すると，一部の特陵

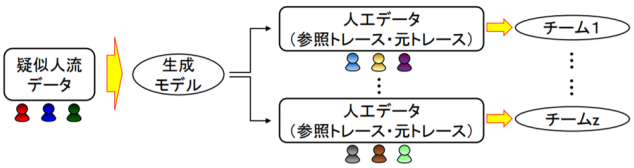

図 1　PWSCUP2019 で使用されたデータ [1]

において定性的に乖離がある事が確認された．そこで本研究では，既存手法において表現されなかった属性の調査および生成データ特性の乖離の原因を明らかにし，その属性が表現された自然な疑似経路データを生成する手法を提案し，その評価を行う．さらに PWSCUP2019 に用いられたデータセットにおいて表現されなかった属性に関してリスクの検討を行う．

2.　既存手法で表現されなかった属性の分析

2.1　既存手法の特徴

本検討では既存手法で生成されたデータと元データの比較を行うため，先に既存手法の特徴について簡単に説明する．詳細な説明についてはコンテスト主催者が公開している情報 [1] を参照されたい．既存手法は，疑似人流データのナイトレイのデータを基に，空間上の遷移を確率的に表現するマルコフモデルの遷移確率行列を構築する．PWSCUP2019 で使用されたナイトレイデータデータは，首都圏における 2013 年 7 月から 12 月の間の非連続な 6 日間分 (2013 年の 7/1, 7/7, 10/7, 10/13, 12/16, 12/22) にわたる東京近郊（首都圏）の人工的なトレースの公開データセットのデータである．カラムとしては，ユーザ ID，性別（推定値），日付・時刻（24 時間），緯度経度，滞在者カテゴリ（home，レストラン等の滞在目的），状態（滞在，移動）が含まれる．ユーザ ID は日別に割り当てられており，異なる日では同一の ID が存在する．ただし，この同一の ID が同一人物の位置上を参考にしたもので仮名化処理を行ったものなのか，異なる人物をベースに生成されたものかは公開情報からは不明であった．緯度は 1m，経度は 10m 単位の詳細なもので，時刻は 5 分毎のデータである．既存手法はこのデータを利用し疑似経路を生成する．生成される疑似経路データは**図 2** に示す都内の一部領域に限定されたものであり，緯度経度ではなく領域を 32 × 32 に分割して割り振った識別 id となっている．一つの領域 ID は約 300m 四方の広さであり，時間は 30 分離散のデータが生成される．なお，既存手法は限定した領域内外での人の移動は考えず，全ての移動がこの領域内に閉じるものとなっている．また，公開されているコンテストで利用されたデータは自宅等の容易にユーザの識別が可能な情報を隠すため，8 時から 18 時までのデータとなっている．

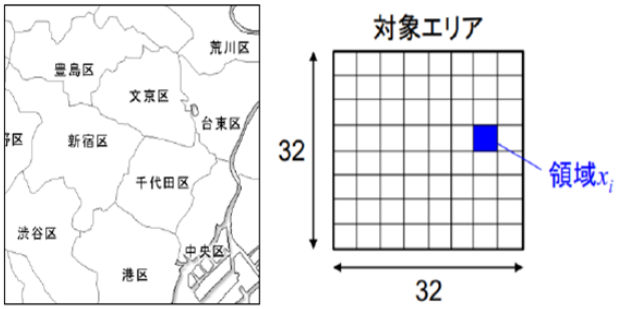

図 2　PWSCUP2019 におけるデータ領域

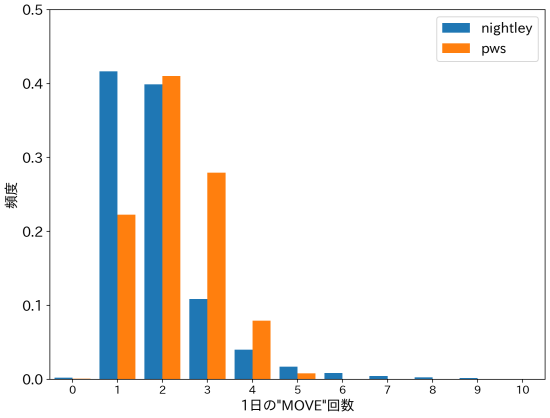

図 3　ナイトレイと PWSCUP2019 データセットの経路における移動回数のヒストグラム

2.2　ミクロな経路特徴の分析

既存手法が人口分布や遷移行列といったマクロな特徴を保持していることは明らかにされているため，既存手法と元データについて，ユーザごとの経路特徴というミクロな特徴に着目した分析を行った．両者で時間の範囲と離散間隔が異なるため，比較時には既存手法の範囲・間隔に合わせてナイトレイデータからランダムに抽出したデータを利用した．

はじめに，既存手法とナイトレイデータの経路データをユーザごとに確認したところ，両者の経路特性には大きく差があった．特に，既存手法の経路は元データに比べて特定箇所の滞在が少なく，移動し続けていた．1 人が 1 日に移動する回数は，平日 2.17 回，休日 1.68 回という報告 [4] もあり，不自然に感じられる．また，一般に経路はある場所から移動し，ある程度一貫した方向に進み，目的地に滞在することを繰り返しつつ，自宅等の拠点となる元の場所に戻る閉路性があると考えられる．しかし，既存手法の経路は元データに比べて移動の方向に一貫性がなく，開始点と終了地点のばらつきも大きく思われた．

次に，経路の差を定量的に評価するため，いくつかの評価指標で既存手法とナイトレイデータの比較を行った．図 3 に両データにおけるユーザごとの 8-18 時の経路中の移動回数のヒストグラムを示す．ここで，ヒストグラムによって分布形状を比較するために，合計が 1 になるように正規化を行っている．. また，既存手法のデータには滞在と移動の状態がないため，2 離散時間（1 時間）以上同一の領域 ID にいたものを滞在，そうでないものを移動とした．ナイトレイ側も 1 離散時間 (30 分) ごとにランダムに代表値として抽出したデータを利用している．図 3 の分布形状は大きく異なっており，特に既存手法は移動回数が多くなっている．この結果は前述した結果とも一致する．図 4 に両データにおけるユーザごとの 8-18 時の経路中の始点・終点間距離のヒストグラムを示す．距離は格子状に分割した領域におけるマンハッタン距離とし，ナイトレイデータも緯度経度から領域 ID に変換して比較した．図 4 の分布形状も差が大きく，同時間範囲の比較ながら，既存手法は始点・終点間距離が長い傾向にあり，閉路性が小さいことが分かる．ただし，ナイトレイのデータは 6 日間のうち半数は日曜日のデータであるため，ナイトレイのデータの始点・終点距離が短くなったのは，多くの人が休日であることも要因になったことも考えられる．

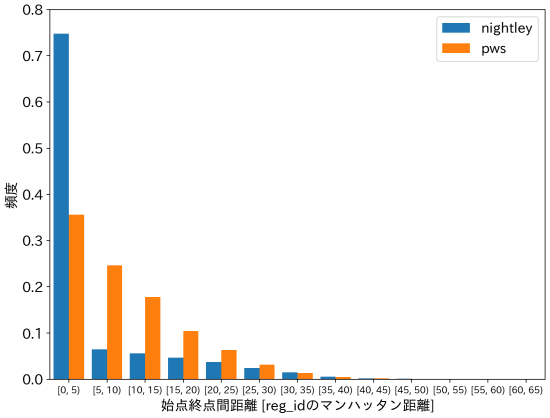

図 4　ナイトレイと既存手法 (pws) の 1 ユーザの 8-18 時の経路における始点・終点間の距離（領域 ID によるマンハッタン距離）

マクロ的な特性を維持しているにも関わらずミクロ的な特性は差異が大きい原因を検討した結果，マルコフモデルの特性によるものと，ユーザごとの時間的な行動パターンを保持していないためであると考えられた．マルコフモデルは次状態への遷移がこれまでの遷移に依存せず現在の状態のみで決定されるというマルコフ性から，都市計画等の分野で人流等の推定のため活用されてきた [5], [6]．しかし，都市計画分野においてマルコフモデルを使用する際の関心は，全体としての人流，あるいは位置履歴の一部の欠損を得ているデータから統計的に尤もらしく補完するという点である．このような補完技術の側面を利用して匿名加工に利用する研究も行われているが [7]，疑似データを 0 から生成する本研究とは目的が異なる．具体的には，欠損・補完したいデータが一部，あるいは人流を表現したい場合，ある場所 A から B に移動する確率が分かればよく，それが

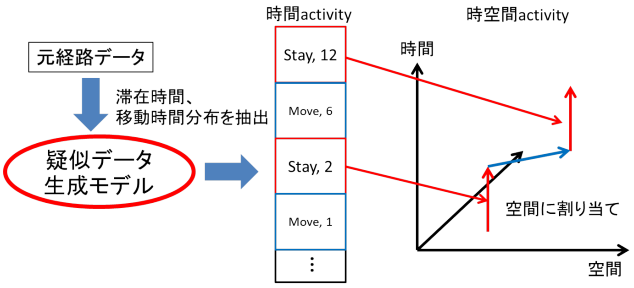

図 5 提案手法の概念図

誰かというのは重要性が小さい．しかし，パーソナルデータとしての経路を生成する場合にこれを適用すると，経路の目的地や方向性を無視して都度確率的な移動が行われてしまう．そのため，既存手法ではマクロな人流としての人口分布や遷移は維持されていたが，ミクロな個人としての経路特性が維持されたなかったと考えられる．

このような個人性の消失を避けるためには，ユーザの時間的な行動パターンを考慮する必要がある．都市計画の分野においても，マルコフモデルと対をなす形で，移動は行為の派生需要として発生するという視点のアクティビティベースドモデル (ABM)[5] が議論されている．ABM では，ユーザごとに行為を規定して，それに基づいて経路を生成する．行為とは，仕事や趣味のような意味をもったものや，目的達成のためのある場所での一定時間の滞在を表すこともある．例えば，[8] では，複数の公開されている統計情報や独自のアンケート情報，GPS 等によって取得された位置情報から，年齢，性別を入力として複数日の仕事や買い物等の行為パターンを生成している．しかし，このような統計情報を適用する地域に応じて取得することは困難であり，[8] においても人為的な調整が多く入っている．疑似データ生成においては，このような人為的な調整は極力小さく，必要な情報の種類も少ないことが望ましいと考えている．

3. 疑似データ生成手法の提案

3.1 提案手法

前項で述べた課題を踏まえ，ミクロな経路特性を維持した生成モデルを，極力少ないデータから抽出する手法を提案する．

図 5 に提案手法の概念図を示す．本手法は以下の 3 つのステップからなる．

(1) 時間の activity（滞在 or 移動）を元データから生成
(2) 滞在 activity の場所を元データから尤度に従って割り当て
(3) 滞在 activity から経路生成

最初のステップでは，出力データと同様の離散時間，領域分割に元データを整形し，各領域 ID と時刻 ID ごとに滞在時間と移動時間の分布を作成する．本研究では単にヒストグラムで分布を疑似的に表現した．次に，生成した分布を時間方向に重ね合わせた時間ごとの分布を作成する．そして，初期の滞在・移動状態を元データの割合に従って決定し，初期時刻における分布から該当の状態の滞在/移動時間を決定する．滞在/移動時間が決定した後，その時間経過後の時刻を次状態として，滞在/移動の状態を変化させて同様にその時間を決定する．これを 1 日の時間に達するまで繰り返すことで，図 5 の中央に示す時間の活動（activity）属性を得る．

次のステップでは，初期に作った時間方向に統合する前の時空間の元データ滞在分布から，決定した時刻と滞在時間を尤度に従って確率的に空間に割り当てる．より自然なデータに近づけるために，本研究でも既存手法と同様の家と各時刻の滞在確率のモデル化を行った．これにより，仕事的な滞在はオフィスエリアに多く割り当てられるといったマクロ的な特性の維持が期待できる．加えて，ナイトレイデータには自宅を示す home の状態が付与されているため，それを基に滞在確率分布と自宅領域分布を作成し，滞在の空間割り当て時にあらかじめ設定した自宅を確率的に選択するとした．ただし，PWSCUP と同様に自宅は図 2 の領域から選択している．都心のような場所では，郊外からの流入流出は無視できない要因であり，領域外の扱いについては追加検討が必要と思われる．home 情報の利用は人為的であるが，既存手法でも同様のモデル化を行っているため，ミクロな経路特性の維持を既存手法と提案手法間で比較する際には大きな影響はないと考えられる．さらに提案手法では，よく訪れる場所を表現するために，一度訪れた場所の再訪率を設定している．これは，ユーザの滞在場所には，職場や学校など，複数の日に訪れる場所があるのではないかとの考えから設けたものである．ただし，本研究で使用したナイトレイは連続した日付のデータがないため，ユーザが滞在場所に再訪する確率分布を得ることが出来なかったため．人為的に設定した．これは，連続した日の経路データを解析することで自動的に割り当てることができるのではないかと考えている．このステップにおける各領域における確率は領域全体で 1 となるように正規化する．

本研究では，限られた時間での空間的な移動距離には制約があると考え，元データからある領域における移動距離分布を抽出し，次の滞在場所はその距離範囲内から選択するとした．これはアクティビティベースと同様の移動は行為の派生であるという観点によるものである．疑似データ生成としては，移動のトラブルによって目的が変化するといった移動をベースにした手法も有効な可能性があるが，本研究の主眼からは外れるため今後の課題とした．

最後に，時空間上に割り当てが決定した滞在活動を埋めるように，経路を生成する．本研究では，単にマンハッタン距離を移動時間で等分割して，ランダムにマンハッタン

距離が小さくなるよう移動するとした．実世界では移動は一般に最短距離で行われ，道路や鉄道経路は制限されるため，経路の自由度は小さいと考えられる．また，近年では交通手段を事前に調べてから出かける人が増加しているとの報告 [9] がある．つまり，経路探索サービスで表示された経路をそのままユーザが通ることが考えられる．よって，移動経路の生成は経路探索サービスを用いることで代替できると思われ，移動を主体とした経路生成と組み合わせることでより元データに近い経路生成が期待できる．

提案手法の主な工夫点は 2 つあり，1 つは ABM における行為を，滞在の時間の場所としてとらえた点である．ABM で特に人為的な調整や大量のデータが必要となるのは，仕事等の行為の意味とそれに応じた滞在位置を割り当てる部分である．このような情報を生成モデルが保持できれば，より多様な属性を入力として入れることが期待できるが，情報取得コストとのトレードオフは利用目的によっても異なる．そこで本研究では，行為を滞在時間と場所としてとらえ，その情報を生成モデルが保持した場合に経路特性がどれだけ維持できるかを検討した．本報告では，検討の初期段階としてナイトレイに含まれる滞在と移動の状態を利用したが，滞在の時間と場所だけであれば，2 節で行ったように位置情報が変化しない状態を滞在として定義することでも取得できるため，必要な情報は少ない．

もう一つの工夫は，時間的な活動（滞在・移動）を先に抽出し，その活動に空間的な場所を割り当てるという時空間を分離した手法を採用した点である．1 日の経路（時間ごとの位置情報列）は，平面空間と時間の 3 次元空間上での状態遷移として扱うことができる．しかし，3 次元空間における遷移の組み合わせは膨大であり，限られた元データから遷移確率行列を計算することは困難である．マルコフモデルを使用した既存手法では，時間上の遷移に制約をもち，次時刻までの空間上の遷移を複数回繰り返すことで経路を得ており，組み合わせ爆発を抑えている．しかし，2 節で考察したように，実際の個人としての人の活動は目的を達成するための滞在があり，そのために移動が発生するという考えの方が適している．例えば，職場で 4 時間の滞在，昼食で 30 分の滞在，再び職場に戻って 4 時間の滞在という目的とその間の移動時間があり，それを満たすように移動経路も決定するという考えである．滞在や移動の時間的な制約を先に規定できれば，考慮すべき遷移の組み合わせを大きく減少でき，元データに必要なデータも時空間上の遷移でなく，時間上の滞在・移動の活動と少なくできる．

本手法では，元データから滞在特徴を抽出する際に，空間を無視して時間方向に射影することで，少量のデータでも有効に扱うことを可能とした．その後，ある時刻においてその滞在時間が空間上のどこで起こるかを元データの特徴に従って割り当てる．このように時空間を分離することで，与える属性としても住所や勤務先のような空間的な属性と，生活パターンのような時間的な属性を分離して与えることができる．

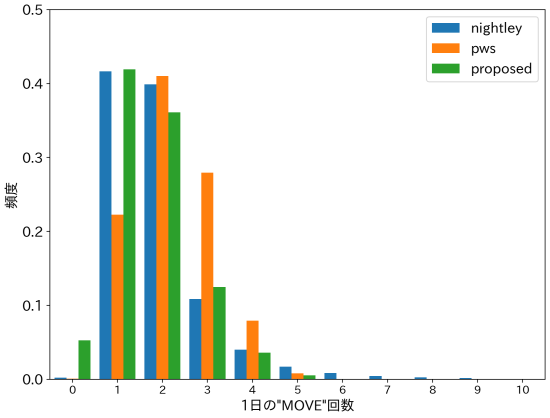

図 6 ナイトレイ，既存手法 (pws)，提案手法の 1 ユーザの 8-18 時の経路における移動回数のヒストグラム．

3.2 評価実験

提案した生成モデルの性能を評価するため，比較には 30000 日分のデータを生成して，提案手法の評価実験を行った．本実験では既存手法，ナイトレイデータとのミクロ経路特徴の比較を 2.2 節と同様に行った．また，ミクロな経路特性だけでなく，マクロな統計情報が維持されているかを評価するため，滞在場所の分布を比較した．今回はステップ 3 の経路生成について暫定的なものとしたため，正確な評価が可能な滞在時間のみで比較を行った．**図 6** にナイトレイ，既存手法，提案手法における個人ごとの 8-18 時の経路中の移動回数のヒストグラムを示す．図 6 から，提案手法の分布形状が既存手法に比べて元データのナイトレイに近いことが確認できる．元データよりも移動回数 0,1 の bin が多く，2,3 の bin が少ないが，分布の形状としては徐々に小さくなっていく同様の形状となっている．回数の差はデータ数が少ないことに起因する誤差の可能性もあるが，1 日のデータを自宅での滞在時間の長い 0 時から生成していることが原因の可能性も考えられる．しかし，本研究で使用したナイトレイは連続した日付のデータがないため，本研究では検証に至らなかった．次に，**図 7** に両データにおける個人ごとの 8-18 時の経路中の始点・終点間距離のヒストグラムを示す．この特徴に対しても，提案手法の方が分布形状も差が小さく，元のナイトレイデータの特徴を維持していることが分かる．

次に，マクロな特性維持の評価として，滞在場所の分布を比較した．**図 8** にナイトレイデータの，**図 9** に提案手法によって生成したデータの滞在領域分布を示す．両図を比べると，若干の差があるものの概ね同様の分布となっている．このことから，個人の特性に着目した経路生成手法においても，マクロな統計情報も維持できたことを示した．

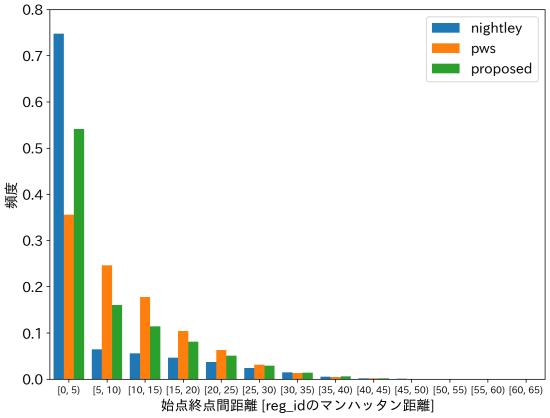

図 7 ナイトレイ，既存手法 (pws)，提案手法の 1 ユーザの 8-18 時の経路における始点・終点間の距離（領域 ID によるマンハッタン距離）

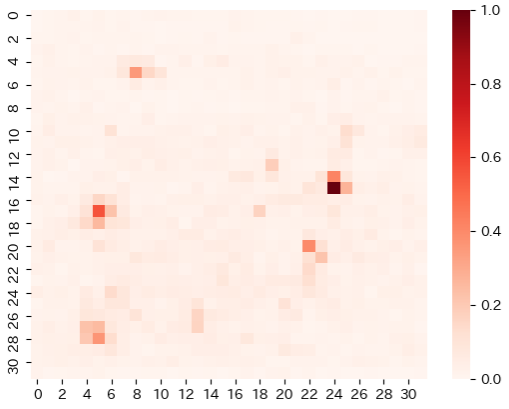

図 8 ナイトレイデータにおける，全時刻での滞在領域分布

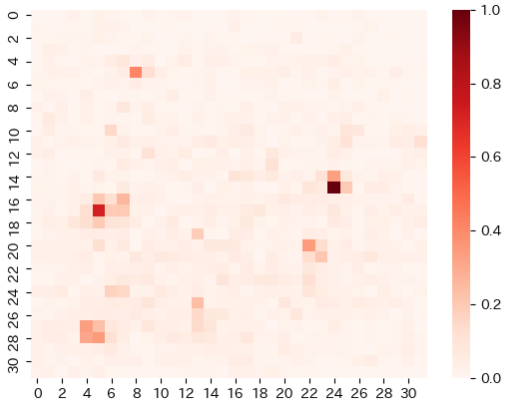

図 9 提案手法生成データにおける，全時刻での滞在領域分布

ミクロな特徴が維持されていてもマクロ的には異なる特徴となる例は多々あり，このような評価ではミクロマクロ両面での評価が必要であると考えている．

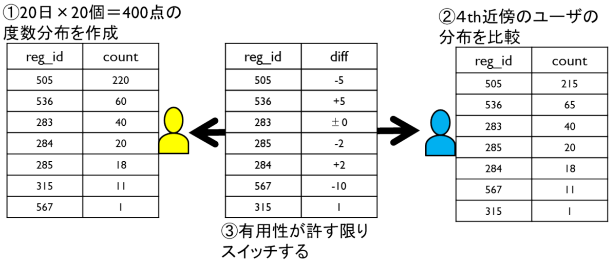

図 10 匿名加工 1 位のチームが用いた手法

4. 滞在・閉路性によるリスクの検討

本節では，これまでに検討した擬似経路データに保持されるユーザごとの滞在および閉路性を考慮した場合にはどのようなリスクがあるのかについて検討を行う．

4.1 PWSCUP2019 で上位チームによって使用された手法

まずは PWSCUP で検討された手法匿名加工 1 位のチームが用いた手法について説明する．まず，**図 10** のように各ユーザの滞在領域分布を計算する．この際の滞在領域分布は時間情報は考慮せず，8-18 時のデータをすべて含めて度数分布表を作成している．この度数分布表を他のユーザと誤認させるように加工を行う．これには，分布が 4 番目に近似しているユーザとの度数分布の入れ替えを行うように加工している．この手法では，時間の制約がないため，加工の際に有用性を節約できる利点がある．今回 PWSCUP では，マルコフモデルの遷移確率によって生成されたデータを用いている．前述した通り，ユーザごとの家以外にユーザの滞在などの固有の情報は含まれていないと考えられるため，全時間の滞在領域分布を他のユーザと誤認させるという手法が優れたものであった．続いて，再識別において ID 識別 1 位のチームが用いた手法について説明する．この手法も全時間における度数分布表を作成し，その分布にぼかしを加えた結果を特徴ベクトルとして，その最近傍のユーザを探索している．ぼかしの付加には，$c = ne^{-\lambda d}$（n：定数，λ：減衰定数，d：距離）を用いている．定数はこのチームが試して一番適した値を選択したとのことであった．結果的にラプラスフィルタに近いものになっているように見受けられた．この手法を用いて再識別を行う．この手法も，時間情報を考慮していないことや，ぼかしのフィルタの選択などが優れていたと考えている．

4.2 PWSCUP2019 で検討されたリスクとの違い

本稿にて我々が検討した滞在や閉路性が反映された場合には，PWSCUP2019 で検討されなかったリスクが存在するのではないかと考えている．既存手法で生成された経路データにおいても，8 時台のデータは高い確率で各ユー

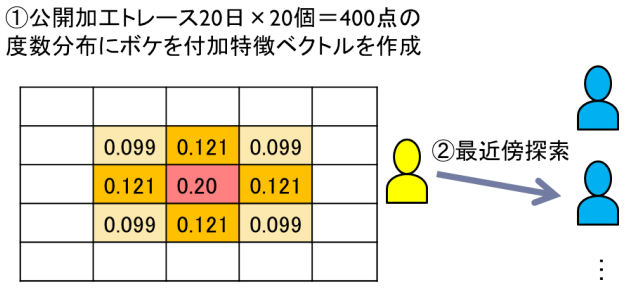

図 11 ID 識別 1 位のチームが用いた手法

ザに設定された自宅に滞在していることが示されており，PWSCUP2019 の参加チームにも，8 時台のデータへの加工が広く行われていた．同様に，自宅以外の場所においても高い確率で滞在する場所が存在すれば，滞在場所の情報は個人特定につながると考えられる．また，全時間の滞在領域分布を用いた場合には，午前中に滞在するユーザと午後に滞在するユーザは近似した分布になるが，滞在する時間に差が表れる恐れがある．また，多くのユーザは自宅に帰宅するために，その経路データには閉路性があると考えられる．実際にナイトレイのデータでは，多くのユーザが 8 時と 18 時のデータの距離が短いことから，自宅を知られないように加工するためには，夕方以降のデータも加工する必要がある．加えて，日常生活において外出の際の目的地と自宅の往復は，ほとんど同じ経路を通る可能性が高いと考えている．外出先での滞留と往復の経路が重複することで，様々な経路を通っていた PWSCUP のデータセットでの滞在領域よりも，滞在領域が密集する可能性があると考えている．これらの特徴がある経路データに対しては，全時間の滞在領域分布に加えて，時間ごとの滞在領域分布も考慮して加工する必要があると考えている．

5. さいごに

本研究では，パーソナルデータ利活用の検討に利用できるデータが限られているという課題を解決するため，経路（位置）データを題材に，属性情報を保持した疑似データ生成について検討した．最初の検討として，PWSCUP2019 で使用された疑似データ生成モデルを既存手法とし，生成データと元データの乖離と，その要因となる元データ属性を分析した．分析の結果，既存手法は時間ごとの人口分布のようなマクロな特性は保持していたが，滞在回数や経路の閉路性等，個人としてみたミクロな経路特性が保持されていないことを明らかにした．その要因として，使用しているマルコフモデルが空間上の遷移はよく表現していたものの，個人の 1 日の時間的なふるまいの属性（アクティビティ）を保持していないためであることを示唆した．そこで，より元データに近い自然な疑似経路を生成するための生成モデルを提案した．ここでは，経路を時空間上での遷移ととらえ，先に滞在と移動という時間上の遷移を決定し，それを空間に割り当てる手法を提案した．評価実験の結果，提案手法によって生成された疑似経路は既存手法よりもミクロな経路特性が元データに近いことを示し，マクロな滞在分布も維持していることを確認した．

また，経路データの持つ滞留や閉路性に着目し，時系列経路データにおけるリスクを検討した．長時間の時系列経路データは仮名化しか行われない場合には，非常に特異性が高いデータであり，PWSCUP2019 でも高い割合で ID 識別がなされる結果となっていた．さらに，滞在場所や，その場所に行くために通る経路が存在する場合には，ID 識別されるリスクがさらに高まると考えている．しかし，位置データや経路データであったとしても，センシティブではなく提供しても良いと感じるデータもあると考えている．そのため，ユーザが提供するデータを選択する手法や，住所や職場などがサービスの提供に必要がない場合には，そのセンシティブな属性を取り除く手法についても検討を行いたいと考えている．

参考文献

[1] PWS 2019 実行委員会：PWSCUP2019 個人データの匿名加工・再識別コンテスト (2020/3/31). https://www.iwsec.org/pws/2019/cup19.html.
[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 (2020/3/31). https://nightley.jp/archives/1954/.
[3] PWS Cup 実行委員会：PWS Cup 2019 データセットについて (2019).
[4] 国土交通省都市局都市計画課都市計画調査室：都市における人の動きとその変化～平成 27 年全国都市交通特性調査集計結果より～，技術報告 (2015).
[5] 羽藤英二，伊藤創太，伊藤篤志（BinN シリーズ）：ネットワーク行動学 -都市と移動- (2014(最終更新)). http://bin.t.u-tokyo.ac.jp/kaken/.
[6] 雄己大山：A Markovian route choice analysis for trajectory-based urban planning, 博士論文，東京大学 (2017).
[7] Murakami, T.: Expectation-maximization tensor factorization for practical location privacy attacks, *Proceedings on Privacy Enhancing Technologies* (2017).
[8] 日高健，大野宏司，志賀孝広：集計データの統合による都市内の移動行動データ生成，土木学会論文集 D3 (土木計画学) (2016).
[9] 小野 由樹子中人 美香：東京圏における駅を中心とした移動と消費に関する調査研究，techreport, JR EAST Technical Review (2008).