GPS 軌跡解析器の開発と長期観測データを用いた新たな個人属性の提案

原　祐輔 [1]

[1]正会員　東京大学大学院助教　工学系研究科社会基盤学専攻・JST さきがけ（〒 113-8656 東京都文京区本郷 7-3-1）
E-mail: hara@bin.t.u-tokyo.ac.jp

GPS データを用いた交通行動調査や解析が一般的となって久しいが，GPS データのみでは交通行動を人間が解釈することが難しく，付随的にアクティビティダイアリー調査などが行われてきた．しかし，これらの調査は被験者にとって著しく調査コストを高めている．一方で，GPS 軌跡データを入手した実務者・研究者にとっても，それらのデータ分析を行うコストは大きく，結果として紙ベース・Web ベースでの社会調査から GPS オリエンテッドな調査へと切り替わっていない．これらを問題と考え，本研究では誰もが GPS を簡単に解析可能な GPS 軌跡解析器:Catsudon を開発する．また，長期間の観測データに対して，この GPS 軌跡解析機を用いることで，非定常時のレア事象の行動予測に対して，これまで重要視されてきた年齢，性別，職業よりも，過去の行動の習慣性や癖を捉えた行動履歴の方が，より大きな情報量を有している可能性を示した．

Key Words : *human mobility, activity recognition, travel behavior analysis*

1. はじめに

GPS 移動軌跡データを用いた行動調査や GPS 移動軌跡データを用いた交通状態モニタリングは一般的なツールとなっており，これらをもとに詳細な人の交通行動の観測やリアルタイムな交通状態の観測が実現可能となっている．

このような人の行動データが詳細かつ長期的に蓄積・分析可能な時代になることで，人の行動の多様性と予測可能性に関する研究が生まれてきた (e.g. González et al.[1])．Eagle and Pentland[2] は人間の行動には，強い習慣性・定常性があることをデータから明らかにしており，Song et al.[3] によれば，長期的な個人データがあれば，平均的に 93%の確率でどこにいるか予測できると報告されている．一方で，行動のエントロピーにより，定常的な行動ではない場合においては，予測限界があることも同時に指摘されている．

計画論の観点から議論すれば，日常的な行動もさることながら，災害時等の発生頻度の低いレア事象時における人の行動を理解・予測したいという要望が存在する．また，長期的な計画を策定するためには，政策や社会変動等の変化に対する応答を知りたいという要望も存在するだろう．このような場合において，これまで重要視されてきたのは，サービス変数と社会経済属性（年齢，性別，職業等）である．特に，社会経済属性は制御変数ではないため，統計モデルの構築時にはある種の調整項の役割を実務的に担うことがしばしば存在する．しかし，社会経済属性が行動に与える影響・メカニズムの分析はなされることが少なく，また，働き方・暮らし方の多様化することによって，固定化された社会経済属性の分類と現代的な生活スタイルとの乖離が大きくなっている．

このような観点から，長期的な予測における政策や社会変動，異常時等の変化に対する応答を分析するためには，古典的な社会経済属性の分類を用いるよりは，各個人の行動履歴そのものを用いた各個人の活動の型や癖を用いた方が，人々の交通行動・活動の本質を捉えられるのではないかと考えた．そこで，本研究では従来型の「社会経済属性」と行動履歴をベースにした個人の「行動スタイル」のどちらがレア事象時の行動を予測可能か，という問題設定をする．

本研究の貢献は以下の 3 つである．

- GPS 点列データから，自動的に移動滞在判別，目的地属性付与，経路特定を行い，行動・活動データへと変換する GPS 軌跡解析器を開発した．これはオープンソースとして github 上で公開されている．
- 東京都市圏の約 600 名・3ヶ月の GPS 行動調査を実施し，新たな個人属性である行動スタイルを提案した．これは中期的に安定した個人の fingerprint になりえることを示した．
- 社会経済属性よりも，個人の行動履歴である行動スタイルの方が未観測レア事象を予測できることを実証的に示し，行動スタイルのもつ潜在的可能性を示唆した．

2. 関連研究

(1) 人の移動・活動パターンに関する既往研究

近年の ICT とスマートフォン等のデバイスの発展により，人々の移動・活動を時間的・空間的に詳細に記録可能な時代となっている．そのデータソースとして，CDR(Call Detail Record), GPS, WiFi, IC カード等が利用されており，都市内の人々の複雑な活動やシステムが理解・分析されている．

González et al.[1] は 100,000 人，6ヶ月間の CDR データを用いて，人の移動パターンを分析し，人々の移動パターンは時空間的な規則性をもっていることを示した．特に，個人の移動パターンはシンプルで再現性の高い分布へと分解できることを示し，人の移動パターン研究に大きな影響を与えた．Song et al.[3] は個々の人々の行動はどの程度予測可能か？という問題設定のもと，CDR データの分析を行い，人間行動には本質的に規則性が存在し，平均的に人の移動は 93%で予測可能であることを示した．また，訪問場所数や移動距離は人によって様々であるにもかかわらず．人口全体での予測可能性のばらつきは非常に小さいことを示した．一方で，人の行動エントロピーのために，どんなに正確なモデルを構築したとしても，予測性能への限界があることも同時に示している．Pappalardo et al.[4] は CDR データから，反復的な行動傾向にある returner と多くの異なる場所を訪問する傾向にある explorers という 2 つのクラスに大きく分けることができ，またそれらは感染症の伝播や社会的相互作用に異なる影響を与えていることを示した．

Eagle and Pentland[2] は人々の長期間の活動パターンに対して主成分分析 (PCA) を行い，各コンポーネントを *eigenbehaviors* と名付けた．特に，上位の eigenbehavior は個人の日常的な行動を表していることを示し，人間の行動には，強い習慣性・定常性があることを実証的に示した．Farrahi and Gatica-Perez[5] は Blei et al.[6] の確率的トピックモデルを用いて，日常的な行動パターンを発見する手法を提案している．Sun and Axhausen[7] はシンガポールにおける公共交通スマートカードデータに対して，その利用パターンに対してテンソル分解と潜在的意味解析を行うことで，利用者の時空間 OD の潜在的なパターンを明らかにしている．

空間側に着目した分析として，Reades et al.[8] は Eagle and Pentland[2] を参考に，モバイル通信の通信量をもとにして，都市を *eigenplace* に分割する手法を提案している．これは都市空間ごとに固有の訪問パターンが存在することを示している．Roth et al.[9] は人の流れが都市のサブセンター周辺に集約・組織化されており，都市の空間構造が入れ子型の複雑な階層構造になっていることを実証的に示している．

災害時の人の避難行動や当時の交通モニタリングに関する研究として，2011 年の東日本大震災時の石巻市の交通状態をモニタリングした研究に Hara and Kuwahara[10] が，2016 年の熊本地震における研究として Kawasaki et al.[11] が存在する．また，東日本大震災時の避難行動と平常時の行動パターンの関係性を分析した研究に関塚ら [12] がある．本研究はこれらの研究結果をもとに，より詳細な行動理解・行動分析を行うことをモチベーションとしている．

(2) GPS データ分析に関する既往研究

GPS データの分析，特に交通行動分析の文脈においてはデータの前処理として捉えられている要素の整理を行う．対象とするのは，マップマッチング，手段判別，滞在場所属性推定である．

マップマッチングは交通工学や交通行動分析においては，一般的な GPS 軌跡データに対する前処理であり，三谷 [13]，Miwa et al.[14], Hunter et al.[15] 等が存在する．基本的な考え方は各 GPS 測位点と各道路リンクとの空間的距離から実際の通行リンクを特定するものであり，それにデータ同化の考え方を導入してシステムモデルに移動を記述する拡張や交差点付近での停止行動を記述することで特定精度を高める拡張が提案されている．

GPS 軌跡から交通手段を判別する手法については，Zheng et al.[16] が各測位点間の距離，ストップレート，速度変化レートを用いて，自動車，バス，自転車，徒歩の 4 手段を判別するモデルを提案している．Shafique and Hato[17] は加速度データを用いた交通手段判別を提案している．

滞在場所属性推定として，Liao et al.[18] は教師ありデータをもとに，Conditional Random Field (CRF) の一種を拡張することで，GPS 軌跡から滞在場所や活動を判別するモデルを構築している．ラベルとしては Work, Sleep, Leisure, Visiting, Pickup, On/Off Car, Other の 7 種類であるが，高精度で判別可能なモデルを構築している．室内での WiFi を用いた場所特定の先駆的研究は Bahl and Padmanabhan[19] による RADAR と呼ばれるシステムであろう．これは室内の各地点での電波強度や距離から，最近傍法 (Nearest Neighborhood 法) を用いて位置特定をした．

マップマッチングや判別問題は，一般に教師データを作成した教師あり学習モデルを構築される例が多い．近年は Eagle and Pentland[2] のように，教師なしアプローチ，具体的には PCA や k-means などのクラスタリング，Latent Dirichlet Allocation(LDA) などの確率的トピックモデルを用いたアプローチも増加している．

3. GPS 軌跡解析器の開発

本章では，本研究において開発した GPS 軌跡解析器である GPS trajectory analyzer: Catsudon について記述する．この GPS 軌跡解析器 Catsudon はタイムスタンプ, 緯度, 経度のシークエンスデータのみから，移動・滞在判別，滞在場所特定，滞在施設属性特定，マップマッチングによる経路特定を行う解析ツールである．特に，個人の行動解析を行うことを目的として，GPS 移動軌跡データから，解釈可能性の高い活動・行動データへと変換することを目的とする．**図–1** は Catsudon の概要を示している．

想定する入力データは，同一個人の数週間以上の GPS 移動軌跡のシークエンスデータである．これは，自宅・職場判定などを行うためには数日以上のデータを必要とするためである，出力データとしては，活動・行動が把握可能なトリップ・アクティビティデータと，後述する日々の活動を集計した行動スタイルデータである．

この解析器で行う解析の多くは，データ分析の前処理として，土木計画学の専門家やエンジニアにとって既に用いられており，また各構成要素はより洗練された手法が提案されているものも多い．しかし，この解析器開発の動機はテクノロジーの民主化にある．スマートフォンなどの GPS 測位器をほぼ全ての個人が持つ時代を鑑みて，すべての個人が自身のデータを分析可能な状況とするテクノロジーの民主化は，現代的な課題である．そこで，GPS 移動軌跡データさえ入力することで，全自動で解釈可能性の高いデータへと変換するツールが必要であると感じ，この GPS 軌跡解析器を開発した．本解析器のコードは github にて公開 [20] している．

(1) 空間データの整備

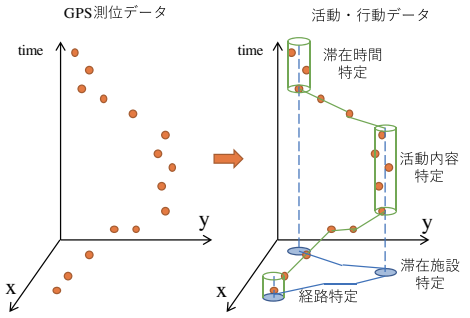
まず，空間データの整備として，OpenStreetMap をベースにネットワークデータを作成した．OpenStreetMap には，可視化すると接続しているように見えるが，リンク ID やノード ID 上ではリンク間の接続関係が記述されていない道路リンクが多く存在する．そこで，日本全国の道路リンクデータに対して，トポロジー情報を付与した．これらは日本全国で 10km×10km の二次メッシュ数で 4688，道路リンク数は 16, 447, 188 本に及ぶ．ここで，OpenStreetMap 上では，道路リンクは必ずしも 1 つの交差点間を道路リンクと定義していない点に注意されたい．

経路特定上，これらの全リンクから経路特定を行うのは計算上，非効率であるため，三次メッシュごとに道路リンクを格納し，それらの三次メッシュを格納した二次メッシュという階層構造をもった json ファイルとして，全道路リンクを整備・格納した．これにより，マップマッチング時の近隣リンク探索を高速化できる．

次に，施設データの整備として，位置情報 SNS である foursquare の施設データを用いる．foursquare は API が公開されており，位置座標を API に送ることで近隣施設情報を入手することができる．OpenStreetMap と同様，foursquare も利用者が施設登録をすることでデータベースが構築されているので，完全な施設情報を網羅しているわけではない．しかし，利用頻度の高い商業施設や飲食店などの多くは登録されているため，行動履歴から活動特定をするための施設データとして利用することが可能である．

timestamp	lat	lon
2014/4/1 7:45:12	38.220633	140.728537
2014/4/1 7:46:31	38.22076	140.72721
2014/4/1 7:47:54	38.22092	140.72833
2014/4/1 7:49:21	38.22094	140.72851

出発地	目的地	出発時刻	到着時刻	経路
自宅	A大学 B研究科	2014/4/1 7:46:31	2014/4/1 8:15:50	
A大学 B研究科	学食	2014/4/1 12:15:31	2014/1/1 12:21:40	
学食	A大学 B研究科	2014/4/1 12:53:46	2014/4/1 12:58:34	
A大学 B研究科	自宅	2014/4/1 20:16:49	2014/4/1 20:45:37	



図–1 GPS 軌跡解析器の概要

(2) 移動・滞在判別

最初に，GPS 移動軌跡シークエンスデータから移動・滞在判別を行い，トリップデータと滞在データに分割する．移動・滞在判別は，ルールベースにて実装した．

まず，第 1 段階目の処理として，各 GPS 点が「移動中」か「滞在」かのフラグをルールベースに設定する．連続する 2 つの GPS 点を用いて,「距離が 30m 以上かつ二地点間の速度が 3km/h 以上」または「距離が 500m 以上」のとき，その GPS 点を「移動中」と判定し，それ以外は「滞在中」と判定する．次に，挟み込み処理として，各観測点が「移動中」のフラグがあるが，前後の観測点が「滞在中」と判定されており，かつ前後の点のタイムスタンプ差が 300 秒以下の場合,「移動中」のフラグを「滞在中」へと変更する．同様に，観測点が「滞在中」のフラグがあるが，前後が「移動中」の観測点に挟まれた場合には「移動中」へと変更する．

次に，これらの移動中・滞在中とフラグを振られた GPS 点をもとに，トリップデータを生成する．連続する 2 つの「移動中」のタイムスタンプ差が 900 秒以内であれば，同一のトリップとしてまとめる．まとめた同

一トリップの所要時間が 5 分以上かつ移動距離が 500m 以上のとき，それをトリップとして判定し，トリップデータを生成する．

(3) 滞在場所特定

次に，期間中の全トリップデータの起点・終点の位置座標から，滞在場所を特定する．起点・終点の位置座標はトリップ発着地の候補である．GPS の性質より，同一施設が発着地のトリップであっても，それらの緯度経度は必ずしも一致しない．そのため，それらが同一施設であれば，まとめる必要がある．

そのためのアルゴリズムとして，Canopy アルゴリズム (McCallum et al.[21]) と k-means アルゴリズムを用いる．基本的な考え方として，同一の地点から出発・到着したトリップの起点座標，終点座標はその地点の位置座標周辺にばらついているはずである．しかし，現在のデータでは，トリップ数はわかっても，訪問施設数を事前に把握することはできない．そこで，Canopy アルゴリズムを用いて，クラスター数の特定を行う．Canopy アルゴリズムを用いた施設特定は以下の流れである．

(a) トリップの起点または終点の位置座標を施設候補集合とする．
(b) 施設候補集合の中から，ランダムにトリップの起点または終点の座標を 1 つ選ぶ
(c) その座標と他のすべてのトリップの起点または終点の直線距離を計算する．
(d) 2 点の距離が T_1 以内の起点または終点座標は同一の Canopy に含める．
(e) 2 点の距離が $T_2 (< T_1)$ 以下の起点または終点座標を施設候補集合から除外する．
(f) 施設候補集合の個数が 0 個になるまで，(b) から (e) の処理を繰り返す．

本研究では $T_1 = 500\text{m}$, $T_2 = 250\text{m}$ と設定した．Canopy アルゴリズムでは，同じデータが複数の Canopy に所属しうる．また，すべての Canopy 間の中心点は違いに T_2 より大きいという性質がある．そこで，Canopy アルゴリズムによって前処理を行い，そこから計算された Canopy 数を用いて，k-means アルゴリズムを行うことで，頑健性のある滞在場所特定を行うことができる．k-means アルゴリズムを行なったのちに，各トリップの起点および終点位置座標は属するクラスターの中心点の位置座標へと変更する．これにより，同一施設・座標の特定を行うことができる．

(4) 滞在施設属性特定

上記の滞在場所特定を行うことで，すべてのトリップの起点または終点の位置座標から，限られた数の滞在場所の位置座標へと特定することができた．次にこれらの位置座標を用いて，施設属性の特定を行う．

(a) 日本全国の通過リンク　(b) 仙台周辺の通過リンク

図–2 Catsudon による 1 年間の通行リンク特定の例

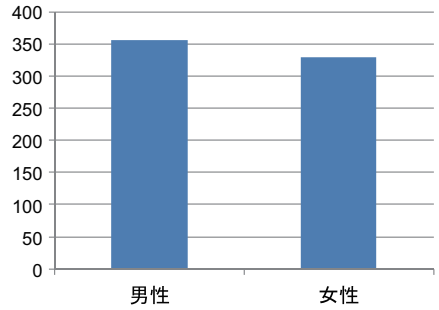
まず，自宅と職場については次のようなルールベースで特定を行う．滞在場所候補地点のうち，全期間の 23 時から 6 時までの時間帯で，最も滞在している滞在場所候補地点を自宅として属性を付与する．同様に，11 時から 17 時までの時間帯で，最も滞在している滞在場所候補地点を職場として属性を付与する．ただし，その場所の滞在頻度が自宅の 1/5 にも満たない場合は職場としての属性を付与しない．逆に，そのように職場の属性を付与しなかった場合，全時間帯にわたって，自宅の 1/5 以上の滞在頻度箇所が存在する場合は職場としての属性を付与する．

自宅と職場以外の滞在施設属性は次のようなステップで行う．Foursquare API を用いて滞在場所候補地点の緯度経度座標から，最近隣にある周辺に存在する施設名称と施設属性を取得する．ここで取得される施設属性は詳細な施設属性であるため，詳細な施設属性から次の 12 属性（飲食店，店舗・サービス，娯楽，旅行・交通施設，アウトドア・レクリエーション，イベント，大学，ナイトスポット，医療施設，住宅，教育，その他）へと変換する辞書を事前に準備し，施設の大属性を特定する．

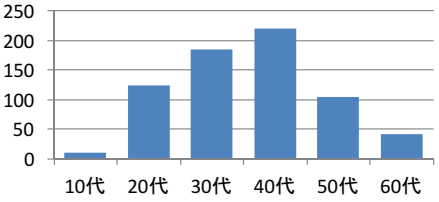
最近隣の施設を滞在場所として特定するという現在のアプローチは非常にナイーブであり，都市部のような店舗が密集している地域や建物内に複数の店舗が入っている場合では特定精度が低下する．トリップ時間帯や 1 時点前の滞在箇所の施設属性を用いた滞在施設特定への拡張が今後の課題である．

(5) マップマッチングによる経路特定

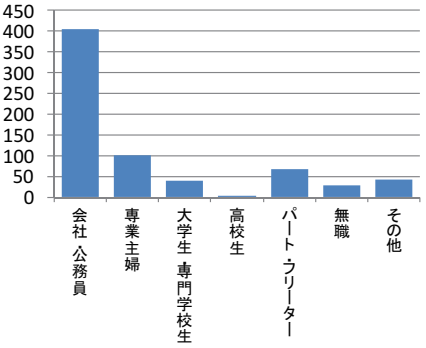
最後に，トリップの起点・終点および移動中の GPS 点を用いることで，OpenStreetMap から構築した道路リンクデータに対して，マップマッチングを行い，経路特定を行う．今回は，GPS 軌跡シークエンスが，数十秒から 1 分単位のインターバルであることを想定して，原・桑原 [22] のスパースな GPS 測位点に対するマップマッチングを利用する．結果は Google Earth の kml ファイルとして出力する．**図–??** は筆者の 1 年間の GPS 軌跡を描画したものであり，暖色になるにつれて，通過回数の多い道路リンクである．



図–3 調査対象群の男女比



図–4 調査対象群の年齢比



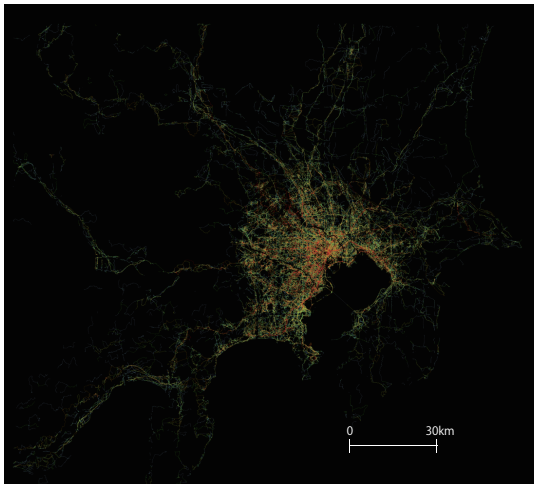
図–5 調査対象群の職業比

4. 首都圏での交通行動調査概要

次に，本研究で実施した交通行動調査の概要を説明する．本調査では，東京都市圏（東京都，神奈川県，埼玉県，千葉県）に自宅および職場が存在する個人を対象に，Moves[23] と呼ばれるスマートフォンアプリを用いた GPS 軌跡と活動場所の取得調査を 2016 年 12 月 1 日から 2017 年 2 月 28 日の三ヶ月間を対象期間として実施した．これらの調査対象者を選定・収集するにあたって，調査会社サーベイリサーチセンターを通して，男女比や年齢比，職業比率の構成に注意して，688 名の調査対象者の選定を行なった．

(1) データの基礎統計

図–2，**図–3**，**図–4** には，調査対象群の男女比，年齢比率，職業比率を表している．男女比については，ほぼ均等に調査を行えている．年齢比については，10 代や 60 代が少ないものの，一部の年齢層に偏ることなく調査を実施した．職業比率についても同様である．



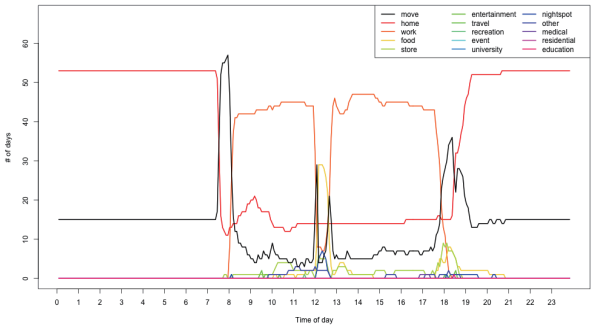
図–6 調査対象群の首都圏での通行リンク

図–5 には期間中の全個人の GPS 軌跡データを Cat-sudon にて解析し，マップマッチングした道路リンクを通行数に応じて色分けした首都圏を示している．688 名 ×90 日なので，約 60,000 人日分を重ねたデータである．

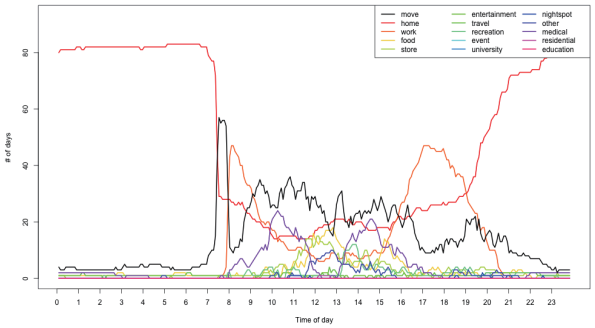
(2) 各個人の行動スタイルデータ

GPS 軌跡解析器での目的地分類と同様に，移動中，自宅，職場，飲食店，店舗・サービス，娯楽，旅行・交通施設，アウトドア・レクリエーション，イベント，大学，ナイトスポット，医療施設，住宅，教育，その他の 15 分類を各個人の移動・滞在施設に対して付与した．一つのタイムスロットを 5 分と仮定し，1 日を 288 のタイムスロットに分割した．このように，1 日の各タイムスロットにおいて，どの属性の施設に滞在していたかまたは移動中であったかを示すヒストグラムを，本研究では行動スタイルヒストグラムと呼ぶ．

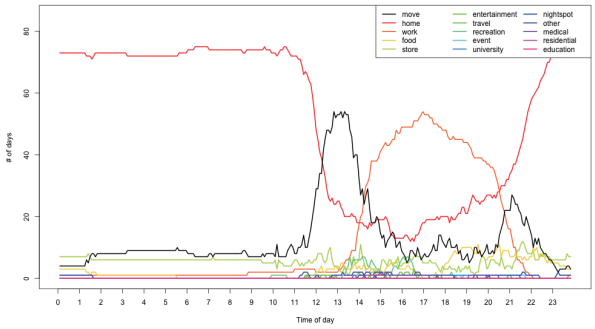
いくつか特徴的な個人の例を示す．**図–6** は伝統的な有職者の行動スタイルといえよう．毎日，午前 7 時半頃から 8 時過ぎにかけて通勤をし，昼の 12 時から 13 時の間にランチをとっている．18 時頃に帰宅を開始し，時折，店舗・サービスに寄ってから帰宅している．このモニターは 54 歳，会社・公務員の男性である．**図–7** も同様に，午前 7 時頃に自宅を出ることが多く，8 時には職場に到着している．一方で，モニター ID(0015) と比較すると日中の移動が多く，また 10 時から 12 時，13 時から 16 時に医療施設に滞在していることが多い．また，16 時から 20 時にかけて再度職場に戻っている．これらの行動から，このモニター ID(0005) は有職者の中でも営業職，特に医療関係の営業職であることが推察される．このモニターは 46 歳，会社・公務員の男性である．**図–8** も定期的に職場に向かっているモニターであるが，職場に滞在する時間は 14 時頃から 21 時頃である．この通勤パターンは伝統的な会社員・公務員とは異なるようにも思えるが，このモニターは 49 歳，会



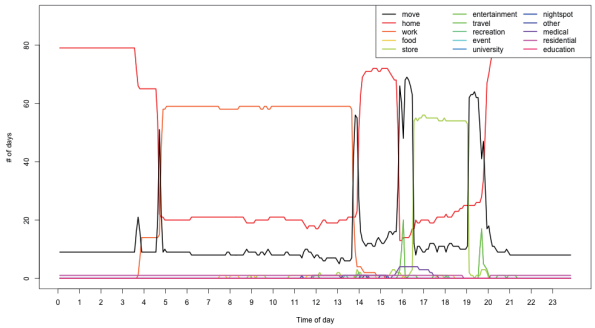
図–7 モニター ID(0015) の行動スタイルヒストグラム



図–8 モニター ID(0005) の行動スタイルヒストグラム



図–9 モニター ID(0360) の行動スタイルヒストグラム



図–10 モニター ID(0413) の行動スタイルヒストグラム

社・公務員の男性である．**図–9** は逆に，早朝 5 時から 14 時まで職場で仕事をしているモニターである．この通勤パターンも伝統的な会社員・公務員とは異なるようにも思えるが，このモニターは 50 歳，会社・公務員の男性である．

この 4 つの例が示すことは，どれも 50 歳前後の会社員・公務員の男性という，年齢・職業・性別の 3 つの社会属性がほぼ一致しているにもかかわらず，1 日の行動パターンが大きく異なることである．また，この結果はこれら 4 名に特殊な例ではない．一般に会社員・公務員といっても働き方は多様であり，また専業主婦という社会属性であっても，いつ買い物にいくのか，どこに買い物に行くのかといった活動パターンは個人によって大きく異なる．そのため，この結果は，都市内の活動の多様性やライフスタイルの多様性によって，年齢や職業，性別といった古典的な社会経済属性から交通行動や都市活動を説明するという既存のアプローチの一つの限界を示しているといえよう．

5. 行動スタイルの fingerprint 性

一方で，上記の例を見る限り，個人の行動スタイルは非常に安定している（同じ時間帯に通勤したり，同じ時間帯に同じ場所にいる）ように思われる．そこで，この行動スタイル自体を新たな個人属性として，定量化できないかについて，本章では検討する．以降の分析では，調査終了後に 30 日以上の行動が観測された 579 名を対象とする．

行動スタイルヒストグラムから離散的な多項分布である行動スタイル分布へと変換することを考えよう．行動スタイル分布は行動スタイルヒストグラムを正規化したものであり，タイムスロット t に行動 a をしている確率を $P(a_t)$ とする．これは期間中に観測された回数 n_{a_t} を用いて，以下のように計算される．

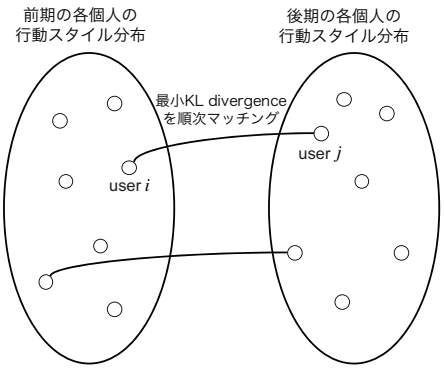
$$P(a_t) = \frac{n_{a_t}}{\sum_{a_t} n_{a_t}} \tag{1}$$

また，一度も観測されなかった行動の確率が 0 となることを防ぐため，この多項分布の事前分布にディリクレ分布を仮定する．このとき，上記の式は以下のように書き直すことができる．

$$P(a_t) = \frac{n_{a_t} + \alpha}{\sum_{a_t} (n_{a_t} + \alpha)} \tag{2}$$

本研究ではディリクレ分布のハイパーパラメータは $\alpha = 0.1$ と設定する．これは Dirichlet スムージングと解釈してもよい．

各個人の行動スタイル分布を安定性を確認するために，各個人の観測された日数を機械的に前期と後期に二等分する．各個人の前期と後期の行動ヒストグラムから，それぞれの行動スタイル分布を $P(a_t), Q(a_t)$ と定義する．安定性の確認のために，前期の行動スタイル分布 $P(a_t)$ と後期の行動スタイル分布 $Q(a_t)$ の差異を計量することを考える．2 つの確率分布間の差異は一般に，カルバック・ライブラー情報量 (KL divergence)



図–11 前期と後期の確率分布のマッチング手続き

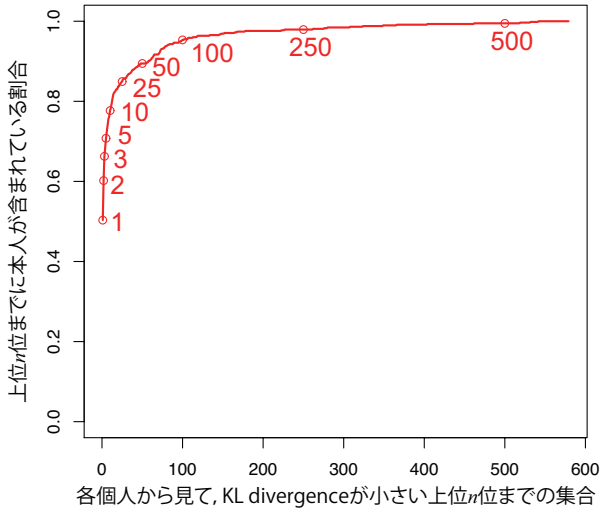
によって計量することが多い．この KL divergence は 2 つの確率分布が一致するときに 0 に，差異が大きくなるにつれて，値が大きくなる性質をもっており，以下の式で表される．

$$D_{KL}(P||Q) = \sum_{a_t} P(a_t) \log \frac{P(a_t)}{Q(a_t)} \tag{3}$$

次に，579 名の前期の行動スタイル分布と後期の行動スタイル分布の組み合わせペア 335,241 通りの KL divergence を計算する．そして，**図–10** に示すように，KL divergence の値が小さくなるペアから順次，前期と後期のユーザーを最近傍マッチングさせていくことで，579 ペアのマッチングを貪欲的に作成する．その結果，579 名中，348 名が前期と後期が本人と一致した．これは割合でいえば，0.601 であり，約 6 割のモニターは個人の行動履歴同士で本人確認が行えることを示している．

図–11 は別の見方を示している．各個人の前期から見たときに，後期の KL divergence が小さい順にランキングを 1 位から 579 位まで付与し，自分（正解）が何位に出てきたかを表したのが，**図–11** である．横軸は何位までに自分が出てきたかを示しており，縦軸はその順位までに本人が含まれているモニターの割合を示している．この結果では，上位 1 位が自分がである個人の数は 291 名，比率にして 0.503 であり，上位 10 位以内に自分が入っている個人の数は 449 名，比率にして 0.776 である．

どちらの結果も，複雑なモデルを用いることなく，頻度ベースの行動スタイルの確率分布が中期的（三ヶ月）に安定していることを示し，かつその行動スタイルそのものがその個人を正確に捉えていることを表している．つまり，行動スタイル分布（過去の行動履歴の集計的な分布）が個人の fingerprint としての役割を果たしていることを示している．これまで用いられてきた社会経済属性（性別，年齢，職業など）の代わりにこの行動スタイル分布自身を一つの個人属性として利用する可能性が示された．



図–12 行動スタイルによる本人適合率の高さ

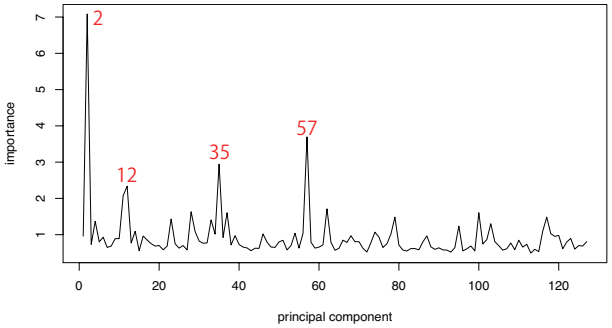
6. 行動スタイルのレア事象との関連性

最後に，本研究の問題設定であるレア事象の行動を予測・説明するのに，従来型の「社会経済属性」と行動履歴をベースにした「行動スタイル」のどちらの方が望ましいかについて，検討を行う．本研究ではレア事象として，今回は年末年始（2016 年 12 月 29 日から 2017 年 1 月 3 日まで）の間に長距離トリップを行うかどうかを対象とする．長距離トリップの定義は自宅または職場から 100km 以上離れた場所への訪問とする．

行動スタイルデータ構築にあたっては，実際の行動を含まないようにするために年末年始の 6 日間を除いた期間で行動スタイル確率分布を作成する．この行動スタイル分布は要素数として 4320 個存在するため，主成分分析 (PCA) で次元圧縮を行い，累積寄与率が 0.95 を超える上位 127 個の主成分を特徴量として用いる．社会経済属性データを用いる場合は，5 歳刻みの年齢ダミー，男性ダミー，職業ダミーの 17 個のダミー変数を特徴量として用いる．

予測モデルとして，今回は簡単のために Random Forest を利用する．検証にあたっては，標本群から 1 つの事例だけを抜き出してテストデータとし，残りを訓練データとする leave-one-out cross-validation (LOOCV) を用いて，2 つのモデルの比較を行なった．また，標本群は長距離トリップを行なった人とそうでない人の割合を 1:1 にしている．そのため，すべてのテストデータに対して，どちらかを予測した場合の精度は 0.5 となる．

分析の結果，社会経済属性を用いた場合の精度は 0.550，行動スタイルを用いた場合の精度は 0.652 であった．この結果，社会経済属性を用いて予測するのに比べて，行動スタイルを用いることで大きく予測精度が向上していることがわかる．この結果は，年末年始を含まない平常時の行動スタイルデータの中にも，年末年始



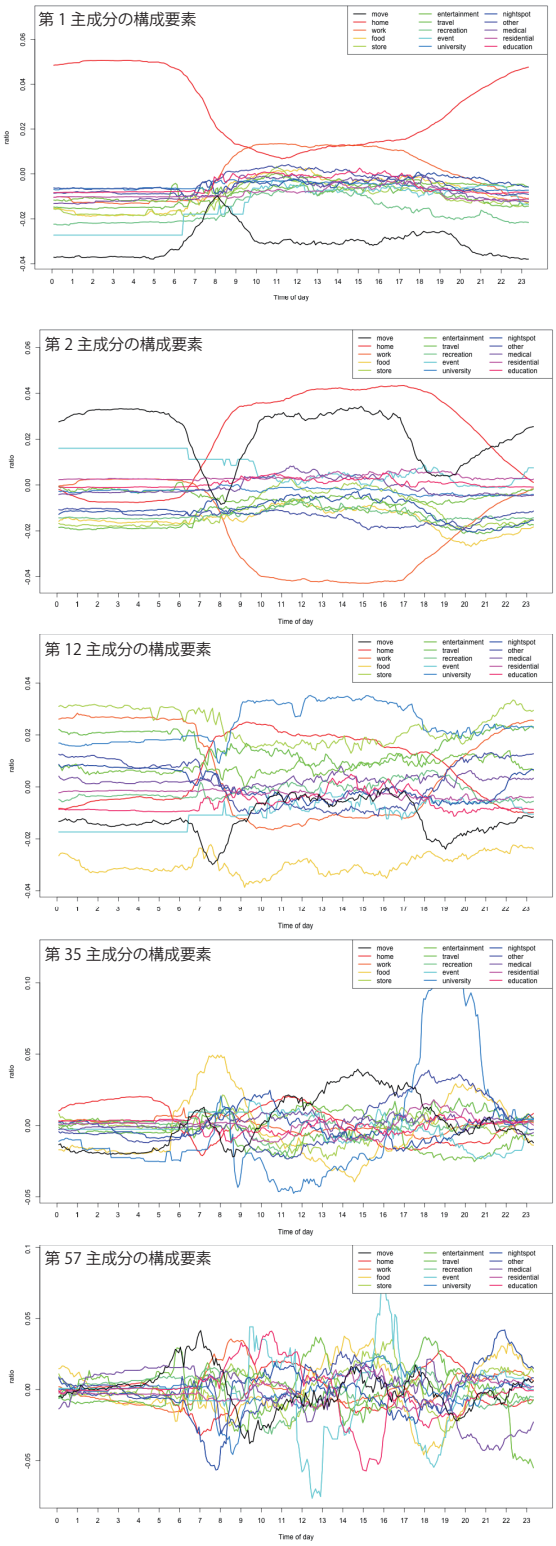
図–13 行動スタイルによる予測における重要度

にどのような行動を行うかを部分的に示すような活動・行動の癖が含まれていることを示唆している．つまり，平常時の行動スタイルデータは年末年始のようなレア事象時に対して，社会経済属性よりも大きな情報量をもつ可能性があることを示している．ただし，行動スタイルを用いたとしても，予測精度は 0.652 に過ぎず，高い精度で予測を当てられるわけではない．この点に関しては，レア事象時の活動・行動予測に，どのような平常時の行動を特徴量として用いるべきか，今後も研究を進めていく必要がある．

最後に，行動スタイルデータによる予測精度が社会経済属性に比べて高かった要因について，考察を行う．Random Forest では特徴量の重要度を出力することができるため，行動スタイルデータでの予測に用いた主成分 1 位から 127 位の重要度を示したのが**図–12** である．その結果，他と比べて大きな重要度を示した主成分 2, 12, 35, 57 がどのような主成分であったかを次に示す．参考のため，第 1 主成分は「夜間に自宅に，日中は職場にいて，8 時周辺に通勤を行う」という行動スタイルであり，これは我々の日常的な行動を考えても，現実的な結果である．ここで，各主成分は正規化されているため，縦軸は確率を表すわけではないが，図示したパターンは相対的な頻度として解釈できる．

年末年始の長距離トリップに影響を与えた行動スタイルの主成分として，第 2 主成分は「日中に自宅にいて，かつ日中のトリップが相対的に多い行動パターン」である．第 12 主成分は「日中に大学にいることが相対的に多い行動パターン」である．第 35 主成分は「夜間に大学にいることが相対的に多い行動パターン」であり，第 57 主成分は「夕方ごろにイベント場所にいること，夜間に外食またはナイトスポットにいることが多い行動パターン」である．これらの行動パターンから，簡易的に解釈を行うならば，第 12, 35 主成分からはいわゆる学生であり，年末年始に帰省を行なっていると考えられる．第 57 主成分は，子育て世代の行動パターンのように解釈ができ，年末年始の帰省や旅行を行いやすいと考えられる．

ただし，社会経済属性では年末年始の長距離トリップを予測できなかったことから，影響を与えている要素は必ずしも学生や子育て世代といった解釈性の高い理屈ではない可能性がある．そのため，行動スタイルがレア事象に対して有している情報量の分析については，より慎重な分析が必要である．



図–14 行動スタイルの重要な主成分の表示

7. おわりに

本研究では，各個人の長期的な行動履歴を新たな社会経済属性として利用することを目的に，GPS trajectory analyzer の開発・公開，首都圏での約 600 名の交通行動調査とその分析，活動履歴の fingerprint 性の示唆，年末年始の長距離トリップに対して，従来型の社会経済属性よりも，普段の活動の癖の方がより情報量を持っている可能性を示した．

今後の課題として，GPS 軌跡解析器に対する既往手法の実装と更なる精度向上，社会経済属性と行動スタイルの関係性，レア事象と行動スタイルの関係性のさらなる検証を必要とする．

謝辞： 本研究は JST さきがけ (JPMJPR15D6) の助成を受けたものです．

参考文献

1) González, M.C., Hidalgo, C.A., Barabási, A.L.: Understanding individual human mobility patterns, *Nature*, Vol.453(7196), pp.779-782, 2008.
2) Eagle, N., Pentland, A.S.: Eigenbehaviors: Identifying structure in routine, *Behavioral Ecology and Sociobiology*, Vol.63(7), pp.1057–1066, 2009.
3) Song, C., Qu, Z., Blumm, N., and Barabási, A.L.: Limits of predictability in human mobility, *Science*, Vol.327(5968), pp.1018–1021, 2010.
4) Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., Barabási, A. L.: Returners and explorers dichotomy in human mobility, *Nature communications*, Vol.6, 2015.
5) Farrahi, K., Gatica-Perez, D.: Discovering routines from large-scale human locations using probabilistic topic models, *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol.2(1), 3, 2011.
6) Blei, D. M., Ng, A. Y., Jordan, M. I.:Latent dirichlet allocation, *Journal of machine Learning research*, Vol.3(Jan), pp.993–1022, 2003.
7) Sun, L., Axhausen, K.W.: Understanding urban mobility patterns with a probabilistic tensor factorization framework, *Transportation Research Part B*, Vol.91, pp.511–524, 2016.
8) Reades, J., Calabrese, F., Ratti, C.: Eigenplaces: analysing cities using the space–time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, Vol.36(5), pp.824–836, 2009.
9) Roth, C., Kang, S. M., Batty, M., Barthélemy, M.: Structure of urban movements: polycentric activity and entangled hierarchical flows, *PloS one*, Vol.6(1), e15923, 2011.
10) Hara, Y., Kuwahara, M.: Traffic Monitoring immediately after a Major Natural Disaster as Revealed by Probe Data - a Case in Ishinomaki after the Great East Japan Earthquake, *Transportation Research Part A*, Vol.75, pp.1–15, 2015.
11) Kawasaki, Y., Kuwahara, M., Hara, Y., Mitani, T., Takenouchi, A., Iryo, T., Urata, J.: Investigation of Traffic and Evacuation Aspects at Kumamoto Earthquake and the Future Issues, *Journal of Disaster Research*, Vol.12, No.2, pp.272–286, 2017.
12) 関塚貴一, 原祐輔, 桑原雅夫, 足立龍太郎: 平常時の行動特性が震災時の避難行動に与える影響に関する研究, 土木計画学研究・講演集, Vol.50, CD-ROM, 2014.
13) 三谷卓摩: プローブパーソン型交通情報発信システムの適用可能性に関する研究, 愛媛大学博士論文, 2005.
14) Miwa, T., Kiuchi, D., Yamamoto, T., Morikawa, T.: Development of map matching algorithm for low frequency probe data, *Transportation Research Part C: Emerging Technologies*, Vol.22, pp.132–145, 2012.
15) Hunter, T., Abbeel, P., Bayen, A.: The path inference filter: model-based low-latency map matching of probe vehicle data, *IEEE Transactions on Intelligent Transportation Systems*, Vol.15(2), 507–529, 2014.
16) Zheng, Y., Li, Q., Chen, Y., Xie, X., Ma, W. Y.: Understanding mobility based on GPS data, *In Proceedings of the 10th international conference on Ubiquitous computing*, pp.312–321, 2008.
17) Shafique, A., Hato, E.: Use of acceleration data for transportation mode prediction, *Transportation*, Vol.42, No.1, pp.163–188, 2015.
18) Liao, L., Fox, D., Kautz, H.: Location-based activity recognition. *Advances in Neural Information Processing Systems (NIPS)*, Vol.18, 787, 2006.
19) Bahl, P., Padmanabhan, V. N.: RADAR: An inbuilding RF-based user location and tracking system, *In INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, Vol. 2, pp.775–784, 2000.
20) Yet Another GPS Trajectory Analyzer: Catsudon, https://github.com/harapon/catsudon
21) McCallum,A., Nigam, K., Ungar, L. H.: Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching, *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.169–178, 2000.
22) 原祐輔, 桑原雅夫: スパースなデータに対するマップマッチング手法と頻度変更型測位に関する研究，交通工学研究発表会論文集，Vol.33, 2013.
23) Moves, https://www.moves-app.com/

(2017. 4. 28 受付)

DEVELOPMENT OF GPS TRAJECTORY ANALYZER AND A PROPOSAL OF NEW INDIVIDUAL ATTRIBUTE USING LONG-TERM DATA

Yusuke HARA