確率的訪問POI分析:
時空間行動軌跡からのユーザモデリング

西田 京介[1,†1,a)] 戸田 浩之[1] 倉島 健[1] 内山 匡[1]

概要：GPS やネットワーク位置情報源（携帯基地局や Wi-Fi など）により得られるユーザの時空間行動軌跡から，そのユーザが訪問した場所（Point of Interest; POI）を推定する確率的訪問 POI 分析技術を提案する．提案技術は（1）時空間カーネルを用いた Mean-shift クラスタリングによる滞留点抽出法（2）ユーザの真の訪問 POI を潜在変数とした，滞留点の位置とその滞留時間に関する確率的生成モデル，から構成され，真の訪問 POI が未知の滞留データも学習に利用することで訪問 POI を高精度に推定できる．本技術が実現する訪問 POI を基にした個々のユーザの行動・嗜好の理解は，情報提供や生活支援などパーソナルアシスタントサービスの品質向上に貢献できる．本論文では，GPS/Wi-FI により得られた実データによる実験を行い，提案技術が従来手法に比べて滞留点の抽出と訪問 POI の推定を精度良く行えたことを示す．

キーワード：ユーザモデリング，位置情報，POI（Point of Interest），階層ベイズモデル，クラスタリング

Probabilistic Visited-POI Analysis:
User Modeling from Spatio-Temporal Trajectories

Kyosuke Nishida[1,†1,a)] Hiroyuki Toda[1] Takeshi Kurashima[1] Tadasu Uchiyama[1]

1. はじめに

近年では，GPS や Wi-Fi を搭載したスマートフォンやタブレット PC の普及と，ソーシャルメディアの発展により，ユーザの現在位置を容易に取得・共有できるようになった．このブレークスルーによって，時間やユーザ属性と結びついた質の高い位置情報がユーザに提供される様になった．たとえば，携帯端末から特定の場所にチェックインして自分の体験を友人に共有しポイントやバッジを取得する，というゲーミフィケーションをサービスに組み込んだ Foursquare は，位置情報に基づいたソーシャルゲームに留まらず，個々のユーザに対して興味を持つであろう場所を推薦可能な情報提供基盤として世界中で大きな発展を遂げている [18,19]．さらに，Google Now や NTT ドコモの i コンシェルなどのようにパーソナルアシスタントとしてユーザの行動を理解し、気象・交通・イベントなどの位置関連情報を個々人向けに提供・推薦するサービスの需要が高まっている．このような背景の中で我々は，個々のユーザに対してより品質・満足度の高い情報推薦を行うためには，明示的に記録・蓄積されにくいユーザの行動を深く理解することが重要と考えた．ここで，ユーザの行動理解のために重要な情報源でありながら，明示的な行動履歴

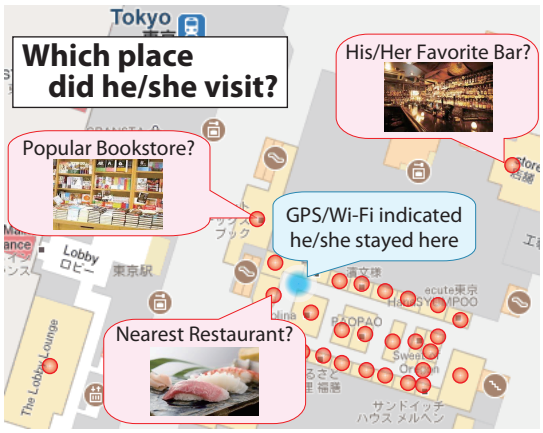

図 1 訪問 POI（Point of Interest）分析の概念図．
Fig. 1 A concept of visited-POI (Point of Interest) analysis.

[1] 日本電信電話株式会社 NTT サービスエボリューション研究所
NTT Service Evolution Laboratories, NTT Corporation
[†1] 現在，NTT レゾナント株式会社 サーチ事業部
Presently with NTT Resonant Inc., Search Division
[a)] k-nishi@nttr.co.jp

として残りにくい情報の 1 つが，ユーザが興味を持って訪問した場所（Point of Interest; POI）である．前述したように，GPS や Wi-Fi を搭載した機器によりユーザの位置情報については容易に自動収集出来るようになったが，その位置情報からユーザの行動を理解するためには，緯度・経度で表される位置情報に対して実際に訪問した POI の情報を意味づけする必要がある．そして，訪問 POI について多数の情報提供をユーザに求めるのは非常に負担が大きいため，訪問 POI を自動的に推定する技術が必要となる．ここで，訪問 POI の推定には，(a) 真の訪問 POI の位置と測位データに誤差が生じる (b) 真の訪問 POI の周辺に多くの POI が存在する，という 2 つの大きな問題がある（**図 1**）[15,33]．測位誤差については、理想的な環境では現状のスマートフォン搭載の GPS，Wi-Fi を用いて 10m 未満の誤差で位置を特定できるが [27]，高層ビルなどが存在する実環境では 100m 以上の大きな誤差が生じる場合がある．実際に，Foursquare におけるチェックイン場所と携帯端末の測位位置の差の中央値は 70m との報告がある [23]．また，測位誤差に比べて POI は密集して存在しているため，測位データの最近傍の POI と，真の訪問 POI が一致する事例は少ない．特に，ターミナルビルなどの大型商業施設においては，隣接する POI や，異なる階層にあって同じ緯度・経度を持つ複数個の POI が存在するため，測位誤差が今後 1m 未満まで改善したとしても訪問 POI の推定は容易ではない．

そこで本研究では，GPS やネットワーク位置情報源（携帯基地局や Wi-Fi など）により得られるユーザの時空間行動軌跡から，そのユーザが訪問した POI を精度良く推定する確率的訪問 POI 分析技術（Probabilistic Visited-POI analysis; PV-POI）を提案する．この提案にあたり，以下の 2 点の Research Questions の解決に取り組んだ．

RQ1. 測位点の軌跡から，ユーザが滞留した地点とその時間を高精度に抽出できるか？

RQ2. 訪問 POI の正解情報が与えられていないデータを含む滞留点の集合から，ユーザが真に訪問した POI を精度良く推定できるか？

RQ1 については従来も研究が行われてきたが [2,32]，距離と時間の両面で高精度な手法は確立されていないため，本研究では新たに時空間カーネルを用いた Mean-shift クラスタリングによる滞留点抽出法を提案する．RQ2 については，従来の訪問 POI 推定技術は教師有りのランキング学習 [15,23] を用いるため，高い推定精度を実現するには，個々のユーザに教師データとなる訪問 POI の情報を多く提供して貰う必要があった．これはユーザの負担が大きいため，本研究では，少数の教師データに加えて，教師データが与えられていない多数のデータも利用することで訪問 POI 推定の精度を高めることを目指し，滞留点の位置とその滞留時間に関する確率的生成モデルを提案する．この提案モデルにより，各滞留点がどの POI を訪問した結果生成されたものかを確率的に推定可能になる．ここで，教師無し／半教師有り学習による POI 推定は本研究が初めての取組であり，ユーザがスマートフォンやタブレットを持ち歩くだけで品質の高い情報推薦を受けられるパーソナルアシスタントサービスの実現に向けた貢献は大きいと考える．

本研究の貢献: 本研究の貢献を以下にまとめる．

1. 教師無し／半教師あり学習可能な確率的滞留点生成モデルの提案による訪問 POI 推定の精度向上
2. 時空間 Mean-shift クラスタリングの提案による滞留点抽出の精度向上
3. POI の密集度合に関する定量的知見

貢献 3 については，これまで定量的な知見が報告されていないため，本研究で併せて分析する．

本研究の構成: 2 章にて，関連研究を示し本研究の位置づけについて整理する．3 章にて，測位誤差や POI の密集度に関する予備分析の結果を示す．4 章にて，距離情報と時間情報を併せて考慮して滞留点抽出を行う時空間 Mean-shift クラスタリングについて説明する．5 章にて，ユーザの真の訪問 POI を潜在変数とした確率的滞留点生成モデルについて述べる．6 章にて評価実験の結果を示し，7 章にて結論を述べる．

2. 関連研究

本章では，訪問 POI 分析，滞留点抽出，測位誤差の分析について関連研究を示し，本研究の貢献点を明らかにする．

2.1 訪問 POI 推定

ユーザの時空間行動軌跡を意味的に解釈する試みとして様々な研究が行われてきた．活動状態（例えば，仕事中か否か）の推定 [6,9,16,17]，移動手段（例えば，徒歩，車，電車など）の推定 [30,31]，移動経路の抽出・予測 [2,3,13,32] などが代表的なタスクとして挙げられる．そして，これらのタスクの基礎的な技術として，ユーザが高頻度で訪れる場所を抽出する研究が行われてきた（2.2 節に詳細を示す）．しかし，これらの研究ではユーザが訪れた場所の位置情報（緯度・経度）が抽出されるのみで，実際に訪問した POI の具体的な名前の推定は行われていなかった．間接的なアプローチとしては，訪問地を含むエリアの意味的解釈を，該当エリアに含まれる POI の情報を基に行う試みが行われているが [25,26]，訪問 POI を具体的に明らかにすることは難しい課題とされてきた [33]．

近年になって，Foursquare などのチェックインサービスの流行により，訪問 POI の推定（チェックイン候補の適切な提示）を行う為の教師データが得られるようになったことで研究が大きく進展した．Lian と Xie は，1 都市 545 人のチェックイン履歴のデータセットを用いて，POI の人気度，チェックイン（測位）点と POI 間の距離，チェッ

クイン時間，ユーザの訪問履歴を素性としてランキング学習を行い，訪問 POI の推定にはユーザの訪問履歴と，チェックイン点と POI 間の距離が役立つことを明らかにした [15]．Shaw らは，Foursquare の大量のチェックイン情報を教師データとして，時間と場所の検索に対して適切な訪問 POI を推定する学習モデルを提案した．このモデルでは，各 POI に関するチェックイン場所・時間の確率値や Foursquare 特有の情報（メイヤー，友人の存在など）を素性としてランキング学習により訪問 POI を推定し，P@1 指標で 53.1% まで精度を高めた [23]．

ここで，上記した従来研究の両方が，ユーザの訪問 POI 履歴が重要な素性として報告している [15,23]．すなわち，訪問 POI 推定については個々のユーザの行動履歴に基づいたモデリングが重要となるが，教師有りランキング学習を用いる両手法において推定精度を高めるには，個々のユーザから訪問 POI の教師データを多数提供して貰わなければならない．本研究は，真の訪問 POI が未知のデータも併せて利用することで，教師有りのデータだけを用いる従来手法よりも訪問 POI 推定の精度を高めた．また，従来手法が利用していない POI のカテゴリと，カテゴリごとの平均滞留時間の違いを考慮することでデータのスパースさの問題を解決し，個人の少量の教師データから確率的に訪問 POI を推定可能にした（**貢献 1**）．

2.2 滞留点抽出

これまで，測位点集合からユーザが高頻度で訪れる位置や，長時間滞留した点を抽出するために，既存のクラスタリング手法が利用されてきた．例えば，Ashbrook ら [2,3] は k-means 法 [12] を，Adams ら [1] は DBSCAN [8] を用いている．また，Kurashima ら [14] は，写真に付与された位置情報の集合から，Mean-shift クラスタリング [5,10] によりランドマークを抽出している．

これらのクラスタリング手法は距離情報のみを考慮したものであるため，POI が密集した地域を異なる日時に繰り返し訪問する様な場合に，高い距離分解能を持って滞留点を抽出することができなかった．Kang ら [13] は，上記の問題意識から，一定範囲内に一定時間以上滞留している点を抽出する time-based clustering 手法を提案したが，1 つの長い滞留点がノイズにより分断されてしまう問題が残っている．なお，Zheng ら [25,32,33] も一連の研究において Kang らと同様の手法を用いている．本研究では Mean-shift を距離と時間情報を併せて考慮する手法に拡張することで，ノイズの含まれる測位点の軌跡からロバストに，かつ高い距離分解能を有して滞留点を抽出できるようにした（**貢献 2**）．

2.3 測位誤差

Zandbergen により，様々なデバイス・環境による測位誤差が報告されている．測位誤差の中央値について抜粋すると，携帯電話の Assisted GPS（A-GPS）による測位では，屋外で 5.0–8.5m，屋内で 6.2–9.8m [29]，iPhone 3G の A-GPS による測位では，理想的な屋外条件で 4.3–12.6m [27]，iPhone 3G の Wi-Fi（Skyfook Wireless 社の測位システム）によるアメリカ都市部のスターバックス店内における測位では 41.6–92.4m [28] であった．また，Paek らによる報告では，良条件の環境における測位において，GPS，Wi-Fi，携帯基地局（GSM）の精度はそれぞれ 10 m，40m，300m 以上としている [21]．また，NTT ドコモでは携帯電話搭載の GPS を利用した測位の誤差範囲をおおむね 50m 未満，基地局の情報を利用した測位では 300m 以上としている [20]．この様に，測位環境・測位デバイスに応じて測位誤差は大きく異なるが，実環境においては 10–100m 程度の測位誤差の存在を無視することはできない．実際に，様々な測位デバイスが用いられる Foursquare では，チェックイン場所と測位位置の差の中央値は 70m との報告がある [23]．そして，訪問 POI 推定のためには，上記した測位誤差に比べて POI が密集していることが問題となる [15,33] が，POI の密集度に関する定量的な知見はまだ得られていない．そこで，本研究では 3.2 節に Foursquare を例に分析を行った結果を示す（**貢献 3**）．

3. 予備分析

本章では，訪問 POI の推定において問題となる，GPS・Wi-Fi における測位誤差と，POI の密集度合について，ソーシャル位置情報サービス Foursquare を題材として分析した結果を示す．

3.1 GPS・Wi-Fi の測位誤差

図 2 と**表 1** に，宅内における Nexus 7 を用いた GPS と Wi-Fi の測位結果と，真の地点（地図上から緯度・経度を取得）の誤差について示す．Nexus 7 は，GLONASS や準天頂衛星の「みちびき」にも対応する最新の高感度 GPS チップ（Broadcom BCM4751）を搭載しており，現在普及しているスマートフォンやタブレットの中では精度良く測定可能な機器の一つである．本実験では，宅内の窓際で Wi-Fi/GPS 信号を常時捕捉可能な場所に端末を固定し，3 秒間隔で測位した．

まず，GPS による測位データは，測位点が真の地点から放射線状に広がって得られた．75% の測位点は真の地点から 13m 以内である一方で，最大で 110m の誤差があった．これは，衛星位置の時間変化が測位精度に影響しているためと考える．なお，2 点間の距離計算には，Vincenty の公式 [24] を用いた（本論文では特別な注釈の無い限り，本公式を用いて距離を計算する）．また，GPS の測位点に正規性は認められなかった（Shapiro-Wilk Multivariate Normality Test; n.s.）これは，測位点が i.i.d-サンプリング

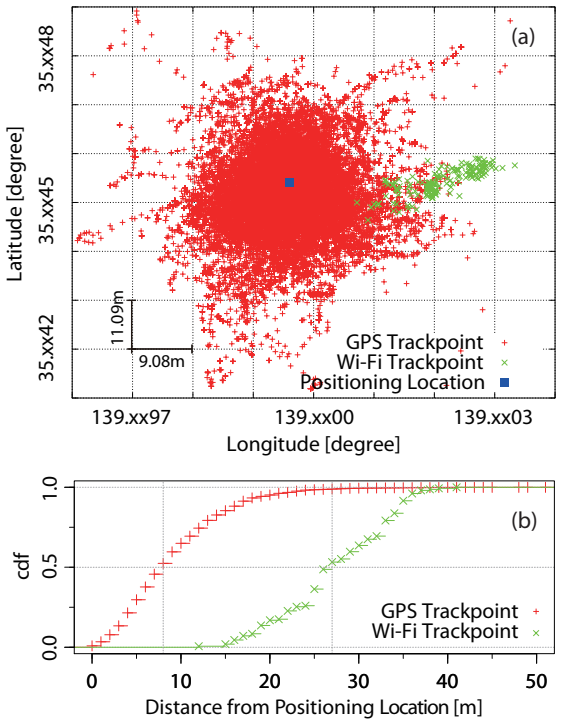

図 2 (a) Nexus 7 の GPS と Wi-Fi による同一場所に滞留時の測位結果．(b) 各測位点の真の地点からの距離の累積密度関数．

Fig. 2 Positioning results from Nexus 7 (GPS and Wi-Fi) while staying in same location. (b) Cumulative distribution of distance from positioning location to each trackpoint.

表 1 真の地点から各測位点までの距離の統計量．

Table 1 Statistics of distance from positioning location to each trackpoint.

Device	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
GPS	30555	0.0	5.0	8.0	9.309	13.0	110.0
Wi-Fi	154	12.0	23.25	27.0	27.54	33.0	41.0

ではないためと考える．次に Wi-Fi では，GPS に比べて測位点のばらつきが小さい代わりに，真の地点からの平均誤差は 27m と大きくなった．図 2 に示すように，測位環境によっては，GPS・Wi-Fi ともに測位点が真の地点から一定の方位にずれる場合がある．

3.2 Foursquare における POI の密集

本分析では，Foursquare Search Venues API を用いて，(a) 東京都のスターバックス（名前に “Starbucks” が含まれるもの）264 件 (b) 東京都のコンビニエンスストア（カテゴリが “Convenience Store” なもの）2000 件 (c) 東京都の任意の POI 2000 件，をランダムにサンプリングして取得し，さらに，各 POI について近接する POI を取得した．

図 3 に最近傍 POI 間距離の累積分布を，**表 2** に中央値，四分位点などの統計量を示す．ここで，POI 間の距離は Foursquare の Search Venue API の実行結果（1m 単位）をそのまま利用した．まず，東京都のスターバックス (a) は

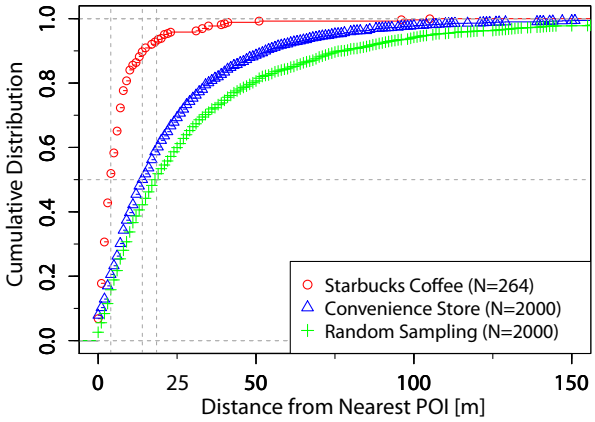

図 3 Foursquare における最近傍 POI 間距離の累積分布．東京都におけるスターバックスコーヒー 264 件，コンビニエンスストア 2000 件，ランダムサンプリング 2000 件について調査．

Fig. 3 Cumulative distribution of distance between each POI and its nearest POI in Foursquare. We investigated Starbucks Coffees, Convenience Stores, and POIs Sampled Randomly in Tokyo.

表 2 Foursquare における最近傍 POI 間距離の統計量．

Table 2 Statistics of nearest POI pair distances in Foursquare.

Group	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Starbucks	264	0.0	2.0	4.0	7.254	8.0	105.0
Convenience	2000	0.0	6.0	14.0	23.11	30.0	431.0
Random	2000	0.0	7.0	18.5	33.38	41.0	1380.0

大規模な商業施設に多く出店しているため，半数の店舗の近傍 4m 以内，75% の店舗の近傍 8m 以内に他の POI が存在している．また，東京都のコンビニエンスストア (b) や他の POI (c) でも近隣に POI が存在しており，コンビニエンスストアでは，半数の店舗の近傍 14m 以内，75% の店舗の 30m 以内に他の POI が存在している．なお，(c) のデータにおいて，POI が最も多く所属するカテゴリはコンビニエンスストアであった（180/2000 件）．

本調査結果より，GPS や Wi-Fi の測位点の最近傍が示す POI を訪問 POI と推定する方法では，GPS や Wi-Fi の測位誤差（10–100m 程度）に比べて多数の POI が密集して存在しているため，正しく推定が行えないことを定量的に示すことができた．

3.3 訪問 POI 推定に向けたその他の問題

上記した問題以外にも，訪問 POI の推定を行う際には，(1) POI データベースに誤った情報（緯度・経度など）が登録されている (2) POI データベースに訪問した POI が登録されていない，などの問題がある．特に Foursquare のようなソーシャルサービスではユーザが自由に POI を登録可能なため，正確な緯度・経度が登録されていなかったり，新しい POI に関する情報が未登録なケースがある．本研究は，真の訪問 POI の位置と測位データに誤差が生じ

る点と，真の訪問 POI の周辺に多くの POI が存在する点の影響を主に考慮したものであり，上記（1–2）の直接的な検討・解決は本研究のスコープ外とする．ただし，(1) については，5.4 節に示す訪問 POI に関する教師データを利用することにより，その影響を軽減することができる．

4. 滞留点抽出

本章では，RQ1 の解決に向けた滞留点抽出方法を提案する．まず用語の定義を行った後に，距離と時間情報を両方考慮する時空間 Mean-shift クラスタリングについて示す．

4.1 用語定義

本章で用いる用語の定義を以下の通り与える．

定義 1 (測位点). GPS やネットワーク位置情報源（携帯基地局や Wi-Fi など）から得られる測位点 $p_i = (x_i, t_i)$ は，位置（緯度，経度）$x_i = (\mathrm{lat}_i, \mathrm{lng}_i)$，タイムスタンプ t_i から構成される．

定義 2 (測位点集合). 測位点集合 $\mathcal{P} = \{p_1, p_2, ..., p_N\}$ は，タイムスタンプによって昇順に整列された測位点系列から構成される．また，$t^{\mathrm{bgn}} = t_1$, $t^{\mathrm{end}} = t_N$ と定義する．

定義 3 (δ-軌跡). δ-軌跡 $\mathcal{T} = \{p_j \mid j = 1, \ldots, N\}$ は，全ての j について $t_{j+1} - t_j \leq \delta$ を満たす測位点集合である．

定義 4 (クラスタ). クラスタ $C = (x^{\mathrm{cls}}, t^{\mathrm{cls}}, \mathcal{P})$ は，中心位置 $x^{\mathrm{cls}} = \sum_{j \in \mathcal{P}} x_j / |\mathcal{P}|$，中心時間 $t^{\mathrm{cls}} = \sum_{j \in \mathcal{P}} t_j / |\mathcal{P}|$，クラスタに所属する測位点集合 $\mathcal{P}$ から構成される．

定義 5 (滞留点). 滞留点 $(x, s, \mathcal{T})$ は，中心位置 $x = \sum_{j \in \mathcal{T}} x_j / |\mathcal{T}|$，滞留時間 $s = t^{\mathrm{end}} - t^{\mathrm{bgn}}$，$\lambda s$-軌跡 $\mathcal{T}$，から構成される．ただし，$s \geq \psi^{\mathrm{time}}$，かつ，$\mathcal{T}$ に含まれるすべての測位点の位置 x_j について $d(x, x_j) < \phi^{\mathrm{dist}}$ を満たす．ψ^{time} は最低滞留時間，ϕ^{dist} は距離カーネル幅を表す．λ は離脱許容係数（$0 < \lambda < 1$）であり，滞留点からの離脱について，滞留時間 s を λ 倍した時間以下は一時的なものとし，滞留点を分断しない．

4.2 時空間 Mean-shift クラスタリング

本研究では Mean-shift を距離と時間情報を併せて考慮することで，ノイズの含まれる測位点の軌跡からロバストに，かつ高い距離分解能を有する手法に拡張する．

以下に，提案手法である時空間 Mean-shift クラスタリングアルゴリズムの擬似コードを示す．提案手法は，測位点 $\mathcal{P}$ が与えられたとき，N 個の滞留点 $\{(x_n, s_n, \mathcal{T}_n)\}_{n=1}^{N}$ を出力する．

1: 各測位点をクラスタとして初期化する
2: 収束まで繰り返し:
3: 　各クラスタについて時空間カーネル内のクラスタを発見する
4: 　各クラスタについて中心位置・時間をミーンシフトする
5: 　同じ中心位置を持つクラスタを結合する
6: クラスタの滞留期間の隔たりが小さいものを連結する
7: 滞留期間に一定時間以上の離脱を含むクラスタを分割する
8: 最低滞留時間を満たさないクラスタを削除する
9: 残存するクラスタ集合を滞留点とする

まず，各測位点 $p_i = (x_i, t_i)$ を，1 データのみが所属するクラスタ $C_i = (x_i^{\mathrm{cls}}, t_i^{\mathrm{cls}}, \mathcal{P}_i)$ として初期化する（行 1）．

$$C_i \leftarrow (x_i, t_i, \{p_i\}) \tag{1}$$

次に，各クラスタ C_n について，時間・空間の両カーネル（カーネル幅をそれぞれ ϕ^{time}，ϕ^{dist} とする）内に含まれるクラスタのインデクス集合 I_n を取得し（行 3），

$$I_n \leftarrow \{j \mid d(x_n^{\mathrm{cls}}, x_j^{\mathrm{cls}}) < \phi^{\mathrm{dist}} \cap |t_n^{\mathrm{cls}} - t_j^{\mathrm{cls}}| < \phi^{\mathrm{time}}\} \tag{2}$$

クラスタ C_n の中心 $(x_n^{\mathrm{cls}}, t_n^{\mathrm{cls}})$ をインデクス集合 I_n に含まれるクラスタの加重平均に移動する（ミーンシフト，行 4）．

$$x_n^{\mathrm{cls}} \leftarrow (\textstyle\sum_{j \in I_n} x_j^{\mathrm{cls}} |\mathcal{P}_j|) / (\sum_{j \in I_n} |\mathcal{P}_j|) \tag{3}$$

$$t_j^{\mathrm{cls}} \leftarrow (\textstyle\sum_{j \in I_n} t_j^{\mathrm{cls}} |\mathcal{P}_j|) / (\sum_{j \in I_n} |\mathcal{P}_j|) \tag{4}$$

ミーンシフト後に，同じ中心位置を持つクラスタ C_n と C_m が存在する場合は C_n にマージし，C_m を削除する（行 5）．

$$C_n \leftarrow (x_n^{\mathrm{cls}}, \tfrac{t_n^{\mathrm{cls}} |\mathcal{P}_n| + t_m^{\mathrm{cls}} |\mathcal{P}_m|}{|\mathcal{P}_n| + |\mathcal{P}_m|}, \mathcal{P}_n \cup \mathcal{P}_m) \tag{5}$$

上記を繰り返し，ミーンシフトが収束（行 2〜行 5）した後に，滞留点の条件を満たすクラスタの抽出を開始する．まず，ミーンシフト収束後は，時間カーネルの効果により短い滞留時間を持つ多数のクラスタが存在しているため，下記の式 (6) を満たす全てのクラスタのペア (C_n, C_m) について式 (5) に従ってマージする（行 6）．

$$d(x_n, x_m) < \phi^{\mathrm{dist}} \text{ and } \mathrm{timegap}(C_n, C_m) < \phi^{\mathrm{time}} \tag{6}$$

ここで，timegap 関数は以下の通り定義される．

$$\begin{cases} 0 & \text{if } t_n^{\mathrm{bgn}} < t_m^{\mathrm{bgn}} < t_n^{\mathrm{end}} \\ 0 & \text{if } t_m^{\mathrm{bgn}} < t_n^{\mathrm{bgn}} < t_m^{\mathrm{end}} \\ \min(|t_n^{\mathrm{end}} - t_m^{\mathrm{bgn}}|, |t_m^{\mathrm{end}} - t_n^{\mathrm{bgn}}|) & \text{otherwise} \end{cases} \tag{7}$$

次に，クラスタ C_n の測位点集合 $\mathcal{P}_n = \{p_h \mid h = 1, \ldots, N_n\}$ について $t_{h+1} - t_h > \lambda s$ となる h が存在する場合は，クラスタ C_n を同じ中心位置 x_n^{cls} を持つ 2 つのクラスタ C_m と C_l に分割する（行 7）．

$$C_m \leftarrow (x_n^{\mathrm{cls}}, \tfrac{\sum_{j \in \mathcal{P}_m} t_j}{|\mathcal{P}_m|}, \mathcal{P}_m = \{p_1, \ldots, p_h\}) \tag{8}$$

$$C_l \leftarrow (x_n^{\mathrm{cls}}, \tfrac{\sum_{j \in \mathcal{P}_l} t_j}{|\mathcal{P}_l|}, \mathcal{P}_l = \{p_{h+1}, \ldots, p_{N_n}\}) \tag{9}$$

最後に，クラスタの滞留時間 $s_n = t_n^{\mathrm{end}} - t_n^{\mathrm{bgn}}$ が最低滞留

時間 ψ^{time} より小さいクラスタを削除する（行 8）．全ての処理（行 1〜行 8）を実行した後に残存するクラスタは定義 5 を満たすため，これを滞留点として抽出する（行 9）．

提案手法では時間カーネルを用いることで，異なる時間帯に同一地域を複数回訪れた際の測位点を区別してクラスタリングする様になるので，POI が密集した地域においても滞留点抽出の距離分解能を向上させることができる．

5. 確率的訪問 POI 分析モデル: PV-POI

本章では，RQ2 の解決に向けて我々が提案する確率的訪問 POI 分析（Probabilistic Visited-POI analysis; **PV-POI**）モデルについて説明する．

5.1 用語定義

本章で用いる用語の定義を以下の通り与える．

定義 6 (POI). POI（Point of Interest）k は，POI の中心位置（緯度，経度）$\mu_k = (\text{lat}_k, \text{lng}_k)$ と，POI が所属するカテゴリ $\text{cat}(k)$ から構成される．POI は必ず 1 つのカテゴリに所属するものとする．

定義 7 (訪問 POI). 訪問 POIz_n は，滞留点 (x_n, s_n) の近傍 Rm 以内に存在する POI とし，各滞留点に対して 1 つの POI が割り当てられるものとする．

また，滞留点については，前章の定義 5 に従う．

5.2 滞留点生成モデルの概要

提案モデルでは，個々のユーザの時空間行動軌跡から抽出した滞留点（位置 x_n，滞留時間 s_n）の集合 $\{(x_n, s_n)\}_{n=1}^N$ が，**図 4** に示す過程により生成されると仮定して，潜在変数である訪問 POI $\{z_n\}_{n=1}^N$ を推定する．**図 5** にグラフィカルモデルを示す．

1: **for each** category $c = 1, \ldots, C$:
2: Draw mean of stay time $\nu_c \sim \text{Normal}(\nu_{0c}, \tau_{0c})$.
3: Draw precision of stay time $\tau_c \sim \text{Gamma}(a_c, b_c)$.
4: **for each** stay point $n = 1, \ldots, N$:
5: Draw category proportion $\theta_n \sim \text{Dirichlet}(\boldsymbol{\beta})$.
6: **for each** category $c = 1, \ldots, C$:
7: Draw POI proportion $\pi_{nc} \sim \text{Dirichlet}(\boldsymbol{\alpha})$.
8: Draw visited-category $y_n \sim \text{Multinomial}(\theta_n)$.
9: Draw visited-POI $z_n \sim \text{Multinomial}(\pi_{ny_n})$.
10: Draw stay location $x_n \sim \text{Normal}(\mu_{z_n}, \sigma^2)$.
11: Draw stay time $s_n \sim \text{Log-Normal}(\nu_{y_n}, \tau_{y_n})$.

図 4 滞留点の生成過程．
Fig. 4 Generative process of stay points.

上記モデルの特徴は，少ない教師データから高精度に訪問 POI を推定するために，POI のカテゴリに関するユーザの嗜好・平均滞留時間を考慮したことにある．次節より，上記生成過程についてモデリングの観点毎に説明する．

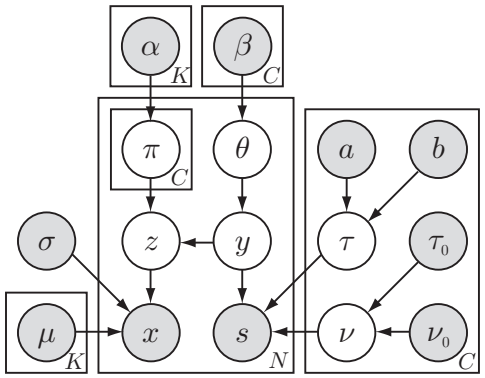

図 5 提案モデルのグラフィカルモデル．塗りつぶし円は観測変数，中抜き円は潜在変数，矢印は依存関係，方形は繰り返しを表す．
Fig. 5 Graphical model of proposed model. Filled circles and circles represent observed and latent variables. Directed edges indicate dependencies between variables. Rectangles represent repetitions.

5.2.1 ユーザ嗜好のモデリング

我々は，ユーザの POI 訪問プロセスを（1）最初に訪問するカテゴリを決定し（2）決定したカテゴリに所属する POI の中から訪問 POI を決定する，と仮定した．

この際，訪問カテゴリ y_n は，過去の行動履歴（ユーザ嗜好）に基づくカテゴリ訪問比率 θ_n をパラメータとして持つ多項分布に従って選択される（行 7）．

$$y_n \sim \text{Multinomial}(\theta_n) \tag{10}$$

そして，選択されたカテゴリにおける POI 訪問比率 π_{ny_n} をパラメータとして持つ多項分布に従って，訪問 POI z_n が決定される（行 8）．

$$z_n \sim \text{Multinomial}(\pi_{ny_n}) \tag{11}$$

ここで，各滞留点の地域特性差を考慮して，滞留点毎に固有のカテゴリ訪問比率 θ_n，POI 訪問比率 π_{ny_n} を持つとする．

5.2.2 滞留位置のモデリング

訪問 POI が $z_n = k$ のとき，滞留位置 x_n は，POI の位置 μ_k を中心とした正規分布に従って生成される（行 9）．

$$x_n \sim \text{Normal}(\mu_k, \sigma^2) = \mathcal{N}(x_n | \mu_k, \sigma^2) \tag{12}$$

ここで，μ_k はデータベースに登録された POI の位置，σ^2 は GPS や Wi-Fi の測位誤差から決定される既知のパラメータとする．

なお，μ_k については，訪問 POI が既知の情報として与えられた滞留点データが存在する場合に補正を行うことで，訪問 POI の推定精度を向上できる（5.4 節に詳細を示す）．

5.2.3 滞留時間のモデリング

訪問 POI が $z_n = k$ のとき，滞留時間 s_n は，z_n が所属するカテゴリ $y_n = c$ の平均滞留時間（対数スケール）ν_c と精度（分散の逆数）τ_c をパラメータとして持つ対数正規

分布に従って生成される（行 10）．

$$s_n \sim \text{Log-Normal}(\nu_c, \tau_c) = \mathcal{LN}(s_n|\nu_c, \tau_c) \tag{13}$$

5.2.4 事前知識のモデリング

提案モデルでは，カテゴリ訪問比率 θ_n，POI 訪問比率 π_{nc} について，ディリクレ事前分布を仮定する．

$$\theta_n \sim \text{Dirichlet}(\boldsymbol{\beta}) \tag{14}$$

$$\pi_{nc} \sim \text{Dirichlet}(\boldsymbol{\alpha}) \tag{15}$$

これらのハイパーパラメータ $\{\alpha_k\}_{k=1}^K$，$\{\beta_c\}_{c=1}^C$ をユーザ全体の傾向から設定することによって，（教師データが多く得られない）個々のユーザ嗜好のモデリングについて事前知識を与えることが出来る．本研究では，

$$\alpha_k = \alpha \cdot N_{\text{users}}(k) \tag{16}$$

$$\beta_c = \beta \left(\textstyle\sum_{k:c=\text{cat}(k)} N_{\text{users}}(k) \right) \tag{17}$$

として，Foursquare における各 POI へのチェックインユーザ数 $N_{\text{users}}(k)$ と，そのカテゴリ毎の合計を定数倍したものをそれぞれ α_k，β_c とした．ここで，α，β は定数パラメータとする．

また，カテゴリ c の平均滞留時間（対数スケール）について，正規事前分布を仮定する．

$$\nu_c \sim \text{Normal}(\nu_{0c}, \tau_{0c}) \tag{18}$$

ここで，ハイパーパラメータ ν_{0c}（平均），τ_{0c}（精度）をユーザ全体の滞留時間の傾向から設定することができる．さらに，カテゴリ c の平均滞留時間の精度について，ガンマ事前分布を仮定する．

$$\tau_c \sim \text{Gamma}(a_c, b_c) \tag{19}$$

a_c，b_c の値により，事前知識が学習に与える影響の度合いを制御することができる．

提案手法では，これらの事前知識を積極的に利用することで，教師データが少ない（あるいは，存在しない）新規ユーザに対する推薦の困難さ（コールドスタート問題 [22]）を軽減することができる．

5.3 ギブスサンプリングによる訪問 POI の推定

提案モデルでは，Collapsed ギブスサンプリング [11] を用いて未知のパラメータ（z_n, y_n, ν_c, τ_c）を推定する．以下に，推定手順を示す．

(1) 訪問 POI が既知であるすべての滞留点 n について，z_n に真の訪問 POIz_n^* をアサインする．

(2) 訪問 POI が既知でないすべての滞留点 n について：

(a) x_n からの距離が R 以内の POI 集合 Z_n（うち，カテゴリ c に所属するもの Z_{nc}）と，Z_n に含まれる POI が所属するカテゴリの集合 Y_n を取得する．

$$Z_n = \{k \,|\, ||x_n - \mu_k|| < R\} \tag{20}$$

$$Z_{nc} = \{k \,|\, k \in Z_n \cap c = \text{cat}(k)\} \tag{21}$$

$$Y_n = \{c \,|\, k \in Z_n \cap c = \text{cat}(k)\} \tag{22}$$

(b) $z_n \in Z_n$ をランダムにアサインし，y_n に $\text{cat}(z_n)$ をアサインする．

(3) 訪問 POI が既知でないすべての滞留点 n について，$z_n \in Z_n$, $y_n \in Y_n$ をサンプリングして更新する．

$$\begin{aligned} &p(z_n = k, y_n = c \mid \mathbf{x}, \mathbf{s}, \mathbf{z}_{-n}, \mathbf{y}_{-n}) \\ &\quad \propto \mathcal{N}(x_n|\mu_k, \sigma^2) \cdot \mathcal{LN}(s_n|\nu_c, \tau_c) \cdot \\ &\qquad \frac{N_{ck,-n} + \alpha_k}{\sum_{k \in Z_{nc}} (N_{ck,-n} + \alpha_k)} \cdot \frac{N_{c,-n} + \beta_c}{\sum_{c \in Y_n} (N_{c,-n} + \beta_c)} \end{aligned} \tag{23}$$

$N_{ck,-n}$，$N_{c,-n}$ はそれぞれ，k かつ c がアサインされた滞留点数，c がアサインされた滞留点数（ただし，z_n，y_n のアサイン値を除く）である．

(4) すべてのカテゴリ $c \in C$ について：

(a) $\nu_c|\tau_c$ をサンプリングして更新する．

$$\nu_c \sim \text{Normal}\left(\frac{\mu_{0c}\tau_{0c} + \tau \sum_n \ln(s_n)^{y_{nc}}}{\tau_{0c} + N_c\tau}, \tau_{0c} + N_c\tau \right) \tag{24}$$

(b) $\tau_c|\nu_c$ をサンプリングして更新する．

$$\tau_c \sim \text{Gamma}\left(a + \frac{N_c}{2}, b + \frac{\sum_n (\ln(s_n) - \nu_c)^{2y_{nc}}}{2} \right) \tag{25}$$

ここで，y_{nc} は，$y_n = c$ のとき 1，それ以外の時 0 となる変数である．

(5) 3, 4 を N_{iter} 回繰り返す．繰り返し終了後，訪問 POI が既知でないすべての滞留点 n について，z_n にアサインされた値の最頻値（あるいは上位 m 件）を訪問 POI の推定結果とする．■

繰り返し終了後，滞留点 n におけるカテゴリ訪問比率 $\hat{\theta}_n$，POI 訪問比率 $\hat{\pi}_{nc}$ の推定値は以下の通り得られる．

$$\hat{\theta}_n = \frac{N_c + \beta_c}{\sum_{c \in Y_n} (N_c + \beta_c)} \tag{26}$$

$$\hat{\pi}_{nc} = \frac{N_{ck} + \alpha_k}{\sum_{k \in Z_{nc}} (N_{ck} + \alpha_k)} \tag{27}$$

また，全体のカテゴリ訪問比率 $\hat{\theta}$ は以下の通り得られる．

$$\hat{\theta} = \frac{N_c + \beta_c}{\sum_{c \in C} (N_c + \beta_c)} \tag{28}$$

5.4 POI の位置補正

上記の提案モデルでは，各 POI の位置 μ_k は既知のパラメータとして扱ったが，POI のデータベースに登録された位置が，真の位置と異なっている場合が多く存在する．ま

た，測位環境によっては，測位点が POI の真の位置を中心とした分布にならない場合がある．

そこで，訪問 POI が既知の滞留点データが得られる場合は，μ_k の補正を行う．補正前の POI の位置を μ_k，真の訪問 POI が $z_n^* = k$ となる滞留点の集合を $X_k = \{n \mid z_n^* = k\}$ としたとき，補正後の POI の位置 μ'_k を

$$\mu'_k = \frac{\mu_k + \sum_{n \in X_k} x_n}{1 + |X_k|} \tag{29}$$

とする．この補正後に，5.3 節に示すギブスサンプリングを行って訪問 POI を推定する．

6. 評価実験

本章では，我々が提案する確率的訪問 POI 分析技術を構成する（1）時空間カーネルを用いた Mean-shift クラスタリングによる滞留点抽出法（2）滞留点の位置とその滞留時間に関する確率的生成モデル，について，人工データ・実データを用いて評価を行った結果を示す．

6.1 滞留点抽出

提案する時空間 Mean-Shift クラスタリング（**ST-MS**）について，Kang らの手法（time-based clustering; **Kang04**）[13] [*1]，Mean-Shift（**MS**）と比較し，提案手法の優位性を示す．提案手法のパラメータは $\phi^{\text{dist}} = 20$，$\phi^{\text{time}} = 600$，$\psi^{\text{time}} = 180$，$\lambda = 0.1$ と設定した．なお，Mean-shift は $\phi^{\text{time}} = \infty$ とした時空間 Mean-Shift に相当する．

6.1.1 擬似データにおける滞留点の抽出

まず，**図 6** に示す様に地点 $A \to B \to C$ を移動する（各地点で滞留した後，次の地点へ移動する）擬似的な時空間行動軌跡 $\{p_i = (x_i, t_i)\}$ において，3 つの滞留点を高精度に抽出できるかを評価した．

本実験では，時空間行動軌跡 $\{p_i\}$ を，移動距離 $d = 30$ と $d = 50$ の 2 条件で，各 100 軌跡ずつ生成した．ここで，真の位置系列を $\{x_i^*\}$ とするとき，各擬似測位点の位置 x_i を $x_i \sim \text{Normal}(x_i^*, 10^2)$ として生成した．また，擬似測位間隔は 10 秒（$(\forall i)\, t_{i+1} - t_i = 10$）とした．各地点の滞留時間 s_A, s_B, s_C については，各軌跡毎に $U(600, 5400)$ からそれぞれ独立にサンプリングして決定した．その他，移動速度 $v = 8/6$[m/s]（徒歩の速度を想定），移動角度 $\theta = 4\pi/9$[rad] と設定した．なお，$d = 30$ のとき，$A \to C$ 区間の距離は 10.4m，$d = 50$ のときは 17.3m である．

滞留点抽出結果の評価を行うため，抽出した滞留点集合 $\{(x_n, s_n)\}_{n=1}^{N^{\text{ext}}}$ と真の滞留点 (x_A, s_A)，(x_B, s_B)，$(x_C, s_C) \in \mathbf{X}^{\text{true}}$ の差について，滞留点の抽出個数，位置，滞留時間の 3 つの観点で評価を行った．まず，抽出個数については，述べ抽出個数 N^{ext} と異なり位置抽出個数 N^{uniq}

[*1] Zheng らの滞留点抽出手法 [25,32,33] は，Kang らの手法と同一であるとして良い．

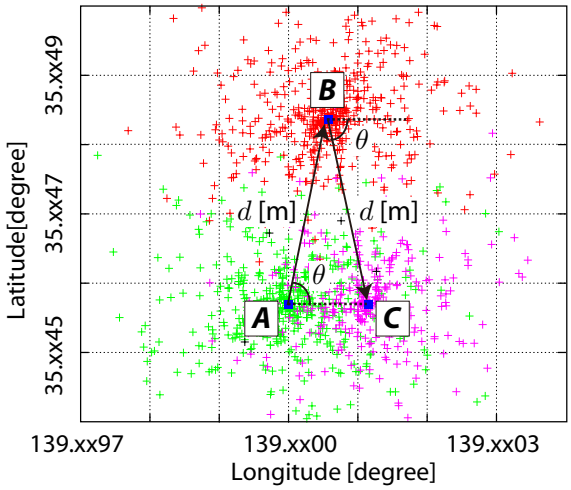

図 6 滞留のある時空間行動軌跡の擬似データ．地点 A→B→C と擬似的に移動する（真の位置系列を $\{x_i^*\}$ とする）とき，擬似測位点を $\mathcal{N}(x_i^*, 10^2)$ に従って生成．滞留点間距離 $d = 30$ or 50[m]，移動速度 $v = 8/6$[m/s]，移動角度 $\theta = 4\pi/9$[rad]，各地点の滞留時間 $s_{[A-C]} \sim U(600, 3600)$[s].

Fig. 6 Pseudo-trackpoint data $\{x_i \sim \mathcal{N}(x_i^*, 10^2)\}$. distance $d = 30$ or 50[m], velocity $v = 8/6$[m/s], angle $\theta = 4\pi/9$[rad], and stay time $s_{[A-C]} \sim U(600, 3600)$[s].

の 2 指標で評価した．次に，抽出位置については，抽出した点の付近に真の滞留点があるか（$E_{\text{pr}}^{\text{dist}}$），真の滞留点の付近で滞留点を抽出できたか（$E_{\text{re}}^{\text{dist}}$）の 2 つの指標を用いた．これらの誤差指標は，それぞれ precision と recall に相当する．

$$E_{\text{pr}}^{\text{dist}} = \frac{\sum_{n=1}^{N^{\text{ext}}} \min_m \{d(x_n, x_m)\}}{N^{\text{ext}}} \tag{30}$$

$$E_{\text{re}}^{\text{dist}} = \frac{\sum_{m \in \mathbf{X}^{\text{true}}} \min_n \{d(x_n, x_m)\}}{|\mathbf{X}^{\text{true}}|} \tag{31}$$

同様に，滞留時間についても以下の 2 指標を用いた．

$$E_{\text{pr}}^{\text{time}} = \frac{\sum_{n=1}^{N^{\text{ext}}} \min_m \{\, |s_n - s_m| \,\}}{N^{\text{ext}}} \tag{32}$$

$$E_{\text{re}}^{\text{time}} = \frac{\sum_{m \in \mathbf{X}^{\text{true}}} \min_n \{\, |s_n - s_m| \,\}}{|\mathbf{X}^{\text{true}}|} \tag{33}$$

表 3 に，$d = 30$ と $d = 50$ の場合の滞留点抽出結果について 100 試行の平均値を示す．まず，Kang らの手法（**Kang04**）は，$E_{\text{pr}}^{\text{dist}}$ の誤差が小さく滞留点の抽出位置については正確であるが，抽出個数 N^{ext} が真の滞留点に比べて非常に多く，滞留時間の誤差（$E_{\text{pr}}^{\text{time}}$，$E_{\text{re}}^{\text{time}}$）も大きい．これは，Kang らの手法では滞留点を連続する測位点と定義しているため，測位誤差の大きいデータが含まれると，1 つの長い滞留点が，複数個の滞留点に分割されて抽出してしまうためである．次に，Mean-shift（**MS**）は，$d = 50$ 条件では滞留点 A と C を区別できず（$N^{\text{uniq}} = 2$），$d = 30$ 条件では全ての滞留点を区別できていない（$N^{\text{uniq}} = 1$）．一方，提案手法である時空間 Mean-shift クラスタリング（**ST-MS**）は，時間カーネルを用いることで，$d = 30$ 条

表 3 滞留のある時空間行動軌跡の擬似データ（$d = 30$, $d = 50$）に対する各手法の抽出結果（100 試行の平均）．誤差 E に関して太字は他手法に比べ有意差有り（paired t-test, $p < .05$）．

Table 3 Results of extracting stay points for pseudo data. Bold-faced error values E reached statistical significance (paired t-test, $p < .05$).

$d = 30$						
Method	N^{ext}	N^{uniq}	$E^{\text{dist}}_{\text{re}}$	$E^{\text{dist}}_{\text{pr}}$	$E^{\text{time}}_{\text{re}}$	$E^{\text{time}}_{\text{pr}}$
Kang04	19.14	19.14	**0.678**	1.597	2755.6	3155.1
MS	1.00	1.00	13.98	9.615	5987.3	4864.6
ST-MS	3.00	3.00	0.760	**0.760**	**49.9**	**49.9**
$d = 50$						
Method	N^{ext}	N^{uniq}	$E^{\text{dist}}_{\text{re}}$	$E^{\text{dist}}_{\text{pr}}$	$E^{\text{time}}_{\text{re}}$	$E^{\text{time}}_{\text{pr}}$
Kang04	19.26	19.26	0.723	1.609	2815.8	3257.2
MS	3.00	2.00	5.999	4.113	934.2	922.9
ST-MS	3.00	3.00	**0.573**	**0.573**	**29.0**	**29.0**

表 4 時空間行動軌跡の実データに対する各手法の抽出結果．

Table 4 Results of extracting stay points for real GPS-logs.

Method	N^{ext}	N^{uniq}	$E^{\text{dist}}_{\text{re}}$	$E^{\text{dist}}_{\text{pr}}$	$E^{\text{time}}_{\text{re}}$	$E^{\text{time}}_{\text{pr}}$
Kang04	39	39	12.877	23.456	1016.8	1348.1
MS	26	20	15.444	19.913	948.5	999.6
ST-MS	27	26	12.108	15.060	568.9	847.5

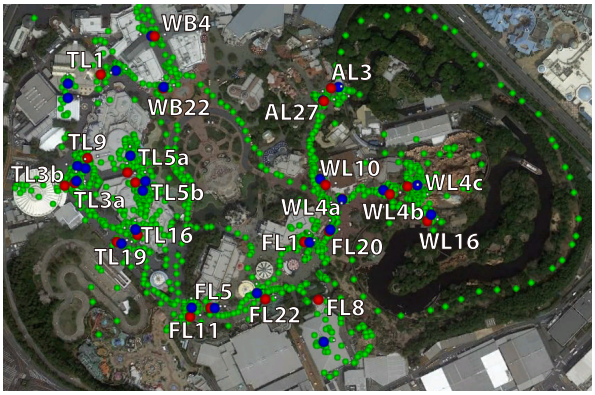

図 7 提案手法による実データからの滞留点抽出結果．緑:測位点，赤:真の滞留点（アトラクションの位置），青:抽出した滞留点．図中のラベル（例: FL11）はアトラクション名を表す．

Fig. 7 Results of extracting stay points with proposed method. Green, red, and blue points are trackpoints, true and extracted stay points. Labels are attraction names.

件でも地点 A での滞留と地点 C での滞留を区別できるようになり，全試行で 3 つの滞留点を正しく抽出できた．また，滞留時間の誤差 $E^{\text{time}}_{\text{pr}}$，$E^{\text{time}}_{\text{re}}$ も他手法に比べて有意に小さいことが確認できた（paired t-test; $p < .001$）．

6.1.2 実データからの滞留点の抽出

次に，実際に GPS ロガーから得られた時空間行動軌跡から滞留点を抽出する実験を行った．あるテーマパークを GPS ロガーを持って移動し，3 分以上滞留した地点（アトラクション）について，手動で滞留期間の記録を残した．滞留地点の正解データは Foursquare に登録されている POI の緯度・経度を利用し，GPS ロガーには V-990（Victory 社，MTK3329 チップ）を利用した．なお，測位環境の影響で GPS ログが記録されない期間の滞留点については正解データから除外した．

表 5 提案手法による実データからの滞留点抽出結果．表中の距離は真の滞留点と抽出された滞留点の位置の距離を表す．

Table 5 Results of extracting stay points with proposed method from real GPS logs. “Distance” means the distance between true and extracted stay points.

True Stay Point		Extracted Stay Point		Distance
FL11	10:29–10:33	SP1	10:29:41–10:33:41	12.958
FL5	10:34–10:38	SP2	10:33:51–10:38:41	6.070
TL5a	10:44–10:56	SP3	10:43:51–10:56:00	23.560
TL5b	10:57–11:17	SP4	10:57:19–11:17:00	15.463
TL16	11:23–11:48	SP5	11:21:30–11:36:26	8.154
		SP6	11:36:36–11:40:39	75.664
		SP5	11:40:49–11:48:58	8.154
TL3a	11:50–12:23	SP7	11:50:08–12:27:51	3.883
TL3b	12:23–12:44	SP8	12:28:01–12:45:26	32.616
TL9	13:05–13:15	SP9	13:05:44–13:11:54	14.621
TL19	13:22–13:30	SP10	13:27:00–13:31:54	6.974
FL1	13:38–14:13	SP11	13:39:06–14:13:36	7.538
TL20	14:14–14:23	SP12	14:13:46–14:24:18	2.721
AL3	14:29–14:35	SP13	14:28:30–14:35:41	10.426
AL27	14:47–14:51	SP14	14:46:43–14:51:15	1.410
WL10	14:54–14:56	SP15	14:52:15–14:57:25	10.989
WL4a	15:03–15:19	SP16	14:57:45–15:16:30	0.797
WL4b	15:19–15:49	SP17	15:16:40–15:45:00	10.603
WL4c	15:49–16:37	SP18	15:45:55–16:31:48	13.792
WL16	16:38–16:43	SP19	16:37:32–16:43:00	10.275
WB4	16:50–16:54	SP20	16:50:10–16:55:34	4.349
FL22	19:09–19:42	SP21	19:08:55–19:43:15	14.336
FL8	19:43–19:55	SP22	19:51:29–19:55:01	59.173
TL1	20:02–20:46	SP23	20:02:03–20:06:24	56.551
		SP24	20:06:34–20:14:31	21.793
		SP25	20:41:56–20:45:18	47.720
WB22	20:49–20:58	SP26	20:49:17–20:58:01	2.133

表 4 に，提案手法と従来手法の比較を示す．提案手法は，距離・時間の誤差指標 E の全てで最も良い結果であった．通常の Mean-Shift クラスタリングでは密集した地域における滞留点抽出の距離分解能が低いため，実際に滞留したアトラクションの種類数（23）に比べて，抽出できた異なり滞留点数（N^{uniq}）が少ないが，提案手法では時間カーネルの利用によりこの問題を解決できた．次に，**図 7** と**表 5** に提案手法による滞留点抽出結果の詳細を示す．提案手法によって，実際の滞留点を，POI が密集した地域においても漏れなく抽出できた．一部の滞留点については，Foursquare に登録された POI の緯度・経度と，実際に滞留する（アトラクションを待つ）地点とに差がある（例: FL8）ため，滞留点の正解データと抽出データの距離が大きくなった．また，アトラクション TL16 に対応する滞留点については，GPS の測位誤差の一時的な増大の影響で，複数の滞留点に分割されて抽出されていた．

6.2 訪問 POI 推定

本節では，提案手法である確率的訪問 POI 分析モデルについて，実データセットによる評価結果と，訪問 POI の推定結果を用いた行動分析の例を示す．

6.2.1 データセット

本実験では，被験者 1 人により 30 日間 Nexus 7（GPS・Wi-FI）による測位を行い，測位データから時空間 Mean-shift クラスタリングにより 5 分以上（$\psi^{\text{time}} = 300$）の滞留点を抽出した[*2]．そして，各滞留点の位置 x_n をクエリとして Foursquare API により近傍 $R = 100$m 以内の POI 集合 Z_n と，真の訪問 POIz_n^* を取得したものをデータセットとした．なお，カテゴリに所属しない POI はデータセットから除き，複数個のカテゴリに所属する POI は 1 つ目のカテゴリを所属カテゴリとした．測位は Android OS の location クラスを用いて GPS と Wi-Fi により常時行い，3 秒毎に location クラスが提供する精度指標が最も良い測位結果を記録した．また，自宅など未登録の POI についてはカスタムの POI を作成し，データセットに含めた．

表 6 に，本データセットに含まれる滞留点の訪問 POI について，上位 5 カテゴリに関する統計を示す．192 個の滞留点（#SP）と，62 個の訪問 POI （#POI）が含まれ，訪問回数 1 回の POI が 44 個，3 回以下の POI が 51 個を占めた（**表 7**）．また，各滞留点の近傍 100m 以内には平均 39.1 個，述べ 2,215 個の POI が存在した．

6.2.2 評価方法

データセットを k 分割し，分割後の 1 つのデータセットに含まれる滞留点に関する訪問 POI は既知とし，残りのテストセットに含まれる滞留点について訪問 POI の推定精度を評価した．訪問 POI が既知のデータの割合を決定する k の値は，20，10，5，4，3，2 と変更して，教師データが少量の場合でも正しく推定できるかを評価した．

比較手法として，最近傍モデル（POI 位置補正なし／あり）**Nearest** ／ **Nearest-C**，Lian らの教師有りランキング学習（POI 位置補正なし（元論文 [15] に相当）／あり）**LianX11** ／ **LianX11-C**，提案手法（POI 位置補正あり）**PV-POI** について評価した[*3]．なお，LianX11(-C) の実装には，RankLib [7] を利用し，ランキング学習器には元論文 [15] と同じく ListNet [4] を用いた．LianX11(-C) は教師有り学習のため，訪問 POI が未知のデータは学習に用いていない，Nearest，LianX11 以外の手法では，5.4 節に示す POI の位置補正を，訪問 POI が既知のデータセットを基に実施した．提案手法のパラメータは，$\sigma = 15$, $\alpha = 10^{-4}$, $\beta = 10^{-4}$, $(\forall c) a_c = 2$, $(\forall c)\, b_c = 5$, $(\forall c)\, \mu_{0c} = 8.0$, $(\forall c)\, \tau_{0c} = 0.3$ と設定した．すなわち，本実験ではカテゴリの平均滞留時間について全てのカテゴリで同じ事前知識を与えた．

[*2] 測位システムの都合により，日付を跨って連続した滞留点は，日付変更時（0AM）で分割されて抽出された．

[*3] Shaw らの手法 [23] の評価には Foursquare のチェックインデータが必要なため比較対象に含めていない．

表 6 データセット中の訪問 POI の上位 5 カテゴリに関する統計
Table 6 Statistics of Top 5 POI-categories in our dataset.

Category	#SP	#POI	mean(s_n)	sd(s_n)
Home (*custom)	66	2	18210.8	15271.7
Ramen / Noodle House	27	12	654.48	345.18
Office	22	2	29385.82	9403.42
Convenience Store	17	5	529.47	201.91
Japanese Restaurant	17	9	1430.18	1827.45
All	192	62	10622.90	14116.05

表 7 各 POI の繰り返し訪問回数に関する統計量．
Table 7 Statistics of number of visits for each POI.

N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
62	1	1	1	3.097	2	39

評価指標には，P@1（Precision@1; 各滞留点に対する推定結果の上位 1 位が真の訪問 POI に一致する割合），R@3（Recall@3; 各滞留点に対する推定結果の上位 3 個に真の訪問 POI が含まれる割合）を用いた．

$$\text{P@1} = \frac{\sum_{n=1}^{N} [\![\hat{z}_{n,1} = z_n^*]\!]}{N} \tag{34}$$

$$\text{R@3} = \frac{\sum_{n=1}^{N} \sum_{i=1}^{3} [\![\hat{z}_{n,i} = z_n^*]\!]}{N} \tag{35}$$

ここで，$\hat{z}_{n,i}$ は滞留点 n に関する POI 推定結果の上位 i 番目を意味する．また，z_n^* は真の訪問 POI とする．

6.2.3 評価結果

図 8 と**表 8** に，データセットの分割数 k を変えた際の，各手法の P@1，R@3（k 回の平均値）の変化について示す．

まず，滞留点に最も近い POI を出力する Nearest モデルでは，測位誤差の影響，POI データベースに登録された位置の誤り，高密集した POI の存在，などの問題により P@1 で 0.209，R@3 で 0.340 と低い精度であった．従来手法である LianX11 は，距離，時間，過去の訪問履歴を素性としてランキング学習を行うことで Nearest モデルに比べて高い推定精度を実現した．また，本実験により，訪問 POI が既知のデータから POI の位置補正を行う（Nearest-C，LianX11-C）ことで，補正を行わない場合に比べて精度が向上することが確認された．提案手法である PV-POI は，データセット（192 個の滞留点）の 25% 以下（$k \geq 4$）の滞留点について訪問 POI が既知のとき，P@1，R@3 の指標において LianX11-C よりも有意に良い結果を得た（paired t-test, $p < .05$）．分割数が少ないため $k \leq 3$ のとき有意差は認められなかったが，50% の訪問 POI が既知のとき，PV-POI は P@1 で 0.796，R@3 で 0.882 の精度を実現し，LianX11-C に比べてそれぞれ 11.9%，4.7% 精度が向上した．PV-POI は，LianX11-C に比べて，訪問 POI が既知でないデータも含めて学習を行うことと，POI のカテゴリ

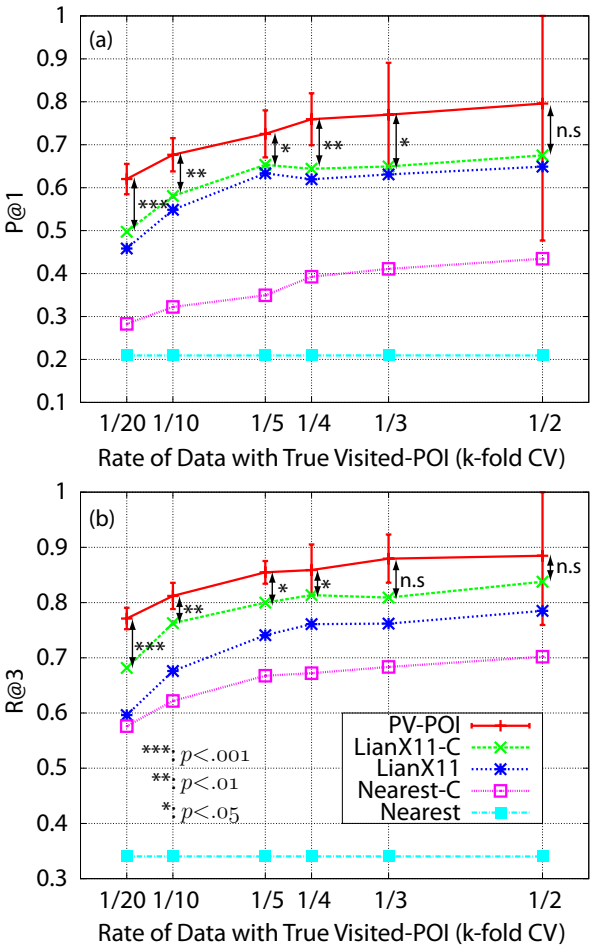

図 8 訪問 POI の推定結果．訪問 POI が既知のデータ数の割合 $(1/k)$ を変更して実験した．(a) P@1 (b) R@3．エラーバーは 95% 信頼区間を示す．

Fig. 8 Results of estimating visited-POIs. (a) P@1 (b) R@3 metrics. Error bars mean 95% confidence intervals.

に関するユーザの嗜好・平均滞留時間をモデリングすることで推定精度を高めることができた．

6.2.4 訪問 POI によるユーザモデリング例

訪問 POI を高精度に推定することでユーザの行動の理解に近づくことができる．例として，前節で示した訪問 POI の推定結果（$k = 10$）から求めた POI カテゴリの時間帯別訪問確率の推定値 $\hat{\theta}$ を**図 9** に示す．図に示す通り，少ない教師データ（20 件弱）から訪問 POI 推定を行った場合でも，このユーザの出勤時間帯（9 時半前），食事の時間帯（8 時半頃）・嗜好（外食，特に蕎麦・ラーメン），コンビニエンスストア訪問の習慣などを高精度に知ることができ，関連する最新情報の推薦・提供に役立てることができる．また，上記の分析結果はカテゴリ単位での分析のため，初めて訪れた地域で各 POI への訪問履歴が全く無い場合でも，ユーザへ情報推薦することができる利点を持つ．

7. おわりに

本研究では，GPS や Wi-Fi，携帯基地局などに基づく測位システムから得られるユーザの時空間行動軌跡から，そのユーザが訪問した POI （Point of Interest）を精度良く推定する確率的訪問 POI 分析技術を提案した．提案技術は（1） 時空間カーネルを用いた Mean-shift 法によるユーザの時空間行動軌跡からの滞留点抽出（2）滞留点の位置とその滞留時間に関する，ユーザの真の訪問 POI を潜在変数とした確率的生成モデル，から構成される．滞留点抽出に関して時間情報と距離情報を併せて考慮することで，滞留点の抽出精度を大きく向上することが出来た．また，POI のカテゴリに関してユーザの嗜好・平均滞留時間を考慮し，教師データが付与されていないデータも含めて学習を行うモデルを提案することで，少ない教師データから，教師あり学習を行う従来手法に比べて有意に高い推定精度を実現出来た．さらには，訪問 POI の推定結果を基にユー

表 8 提案手法 PV-POI と従来手法の訪問 POI 推定結果．(a) P@1 (b) R@3．太字は PV-POI と LianX11-C 間に有意差有り (paired t-test, $p < .05$)．

Table 8 Results of estimating visited-POIs. (a) P@1 (b) R@3 metrics. Bold-faced values mean statistical differences between PV-POI and LianX11-C ($p < .05$).

(a) P@1

Method	$k = 20$	10	5	4	3	2
PV-POI	**0.620**	**0.676**	**0.725**	**0.759**	**0.770**	0.796
LianX11-C	0.497	0.580	0.653	0.644	0.649	0.675
LianX11 [15]	0.458	0.549	0.634	0.620	0.631	0.649
Nearest-C	0.283	0.322	0.349	0.393	0.411	0.435
Nearest	0.209	0.209	0.209	0.209	0.209	0.209

(b) R@3

Method	$k = 20$	10	5	4	3	2
PV-POI	**0.771**	**0.812**	**0.854**	**0.859**	0.879	0.885
LianX11-C	0.682	0.763	0.800	0.813	0.809	0.838
LianX11 [15]	0.596	0.676	0.741	0.761	0.762	0.785
Nearest-C	0.576	0.622	0.668	0.672	0.683	0.702
Nearest	0.340	0.340	0.340	0.340	0.340	0.340

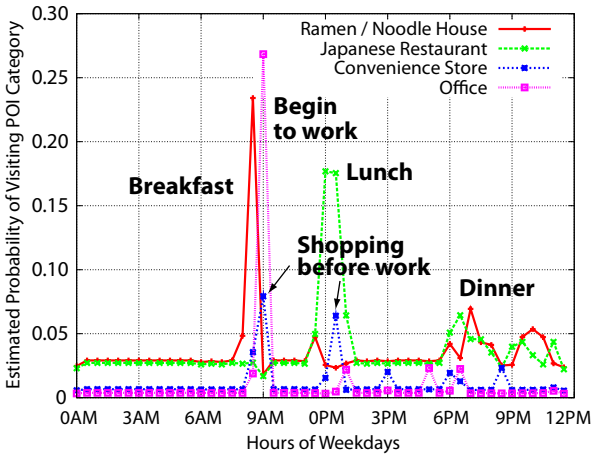

図 9 POI カテゴリの時間帯別訪問確率の推定値（30 分単位）．

Fig. 9 Estimated probability of visiting POI categories every 30 minutes.

ザの行動・嗜好を分析することで，情報提供や生活支援の品質向上に貢献できることについて実例を上げて示した．

提案手法は 1 ヶ月分の個人データ（192 個の滞留点）において，5% のデータの訪問 POI が既知（滞留点 10 個弱）のとき，未知のデータに対する正解率（P@1）が 62.0% であった．そして，50% のデータの訪問 POI が既知のとき，P@1 は 79.6% まで向上した．この結果は，従来手法 [15] に比べて 10% 以上の精度改善であった．また，提案手法は，ユーザ全体の傾向などから POI の人気度やカテゴリごとの滞留時間について事前知識を与えることができる点も従来手法に比べ優れている．事前知識を積極的に利用することで，ユーザが情報推薦を受け始める際のコールドスタート問題 [22] を軽減することができる．

今後は，ユーザが POI を訪問する時間帯や，ユーザの属性（性別，年代），POI 訪問時の天候，さらにはユーザの検索履歴などを滞留点生成モデルで考慮することで，さらに訪問 POI の推定精度を高めていきたい．また，POI の位置情報の誤り，未登録の POI に関する問題の解決に今後取り組みたい．

参考文献

[1] Adams, B., Phung, D. Q. and Venkatesh, S.: Extraction of social context and application to personal multimedia exploration, *ACM Multimedia*, pp. 987–996 (2006).

[2] Ashbrook, D. and Starner, T.: Learning Significant Locations and Predicting User Movement with GPS, *ISWC*, pp. 101–108 (2002).

[3] Ashbrook, D. and Starner, T.: Using GPS to learn significant locations and predict movement across multiple users, *Personal and Ubiquitous Computing*, Vol. 7, No. 5, pp. 275–286 (2003).

[4] Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F. and Li, H.: Learning to rank: from pairwise approach to listwise approach, *ICML*, pp. 129–136 (2007).

[5] Cheng, Y.: Mean Shift, Mode Seeking, and Clustering, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 17, No. 8, pp. 790–799 (1995).

[6] Cho, E., Myers, S. A. and Leskovec, J.: Friendship and mobility: user movement in location-based social networks, *KDD*, pp. 1082–1090 (2011).

[7] Dang, V.: RankLib v2.1, http://www.cs.umass.edu/~vdang/ranklib.html [Online] (2012).

[8] Ester, M., Kriegel, H.-P., Sander, J. and Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *KDD*, pp. 226–231 (1996).

[9] Farrahi, K. and Gatica-Perez, D.: Discovering routines from large-scale human locations using probabilistic topic models, *ACM TIST*, Vol. 2, No. 1, p. 3 (2011).

[10] Fukunaga, K. and Hostetler, L. D.: The estimation of the gradient of a density function, with applications in pattern recognition, *IEEE Trans. Information Theory*, Vol. 21, No. 1, pp. 32–40 (1975).

[11] Griffiths, T. L. and Steyvers, M.: Finding Scientific Topics, *PNAS*, Vol. 101, No. suppl. 1, pp. 5228–5235 (2004).

[12] Jain, A. K.: Data clustering: 50 years beyond K-means, *Pattern Recognition Letters*, Vol. 31, No. 8, pp. 651–666 (2010).

[13] Kang, J. H., Welbourne, W., Stewart, B. and Borriello, G.: Extracting places from traces of locations, *WMASH*, pp. 110–118 (2004).

[14] Kurashima, T., Iwata, T., Irie, G. and Fujimura, K.: Travel route recommendation using geotags in photo sharing sites, *CIKM*, pp. 579–588 (2010).

[15] Lian, D. and Xie, X.: Learning location naming from user check-in histories, *GIS*, pp. 112–121 (2011).

[16] Liao, L., Fox, D. and Kautz, H. A.: Location-Based Activity Recognition using Relational Markov Networks, *IJCAI*, pp. 773–778 (2005).

[17] Liao, L., Fox, D. and Kautz, H. A.: Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields, *I. J. Robotic Res.*, Vol. 26, No. 1, pp. 119–134 (2007).

[18] Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J. I. and Zimmerman, J.: I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application, *CHI*, pp. 2409–2418 (2011).

[19] Noulas, A., Scellato, S., Mascolo, C. and Pontil, M.: An Empirical Study of Geographic User Activity Patterns in Foursquare, *ICWSM* (2011).

[20] NTT ドコモ：測位方法，http://www.nttdocomo.co.jp/service/safety/search/usage/gps/ [Online] (2012).

[21] Paek, J., Kim, K.-H., Singh, J. P. and Govindan, R.: Energy-efficient positioning for smartphones using Cell-ID sequence matching, *MobiSys*, pp. 293–306 (2011).

[22] Schein, A. I., Popescul, A., Ungar, L. H. and Pennock, D. M.: Methods and metrics for cold-start recommendations, *SIGIR*, pp. 253–260 (2002).

[23] Shaw, B., Shea, J., Sinha, S. and Hogue, A.: Learning to rank for spatiotemporal search, *WSDM*, pp. 717–726 (2013).

[24] Vincenty, T.: Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations, *Survey Review*, Vol. 22, No. 176, pp. 88–93 (1975).

[25] Xiao, X., Zheng, Y., Luo, Q. and Xie, X.: Finding similar users using category-based location history, *GIS*, pp. 442–445 (2010).

[26] Yuan, J., Zheng, Y. and Xie, X.: Discovering regions of different functions in a city using human mobility and POIs, *KDD*, pp. 186–194 (2012).

[27] Zandbergen, P. A.: Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning, *T. GIS*, Vol. 13, No. s1, pp. 5–25 (2009).

[28] Zandbergen, P. A.: Comparison of WiFi positioning on two mobile devices, *J. Location Based Services*, Vol. 6, No. 1, pp. 35–50 (2012).

[29] Zandbergen, P. A. and Barbeau, S. J.: Positional Accuracy of Assisted GPS Data from High-Sensitivity GPS-enabled Mobile Phones, *J. Navigation*, Vol. 64, No. 3, pp. 381–399 (2011).

[30] Zheng, Y., Li, Q., Chen, Y., Xie, X. and Ma, W.-Y.: Understanding mobility based on GPS data, *UbiComp*, pp. 312–321 (2008).

[31] Zheng, Y., Liu, L., Wang, L. and Xie, X.: Learning transportation mode from raw gps data for geographic applications on the web, *WWW*, pp. 247–256 (2008).

[32] Zheng, Y., Zhang, L., Xie, X. and Ma, W.-Y.: Mining interesting locations and travel sequences from GPS trajectories, *WWW*, pp. 791–800 (2009).

[33] Zheng, Y. and Zhou, X.(eds.): *Computing with Spatial Trajectories*, Springer (2011).