

2014年度修士論文

人の流れデータを用いた
人々の人口統計学的属性の推定

Estimation of Human Demographic Attributes
from Person Flow Data

西村隆宏
Nishimura, Takahiro

東京大学大学院 新領域創成科学研究科
社会文化環境学専攻

目次

1 序論	7
1.1 研究の背景	7
1.2 人の流れに関するデータと既往研究	8
1.2.1 パーソントリップ調査	8
1.2.2 GPS データ	14
1.3 研究の目的	16
2 人口統計学的属性の推定方法	18
2.1 本研究で使用了データセット	18
2.1.1 人の流れに関するデータセット	18
2.1.2 場所特徴に関するデータセット	22
2.2 手法の概要	25
2.3 本研究に使用了パターン認識手法	28
2.3.1 Neural Network	29
2.3.2 Random Forest	30
2.3.3 Support Vector Machine	32
2.3.4 カーネル法	33
2.3.5 主成分分析	33
2.3.6 クラスタリング	35
2.3.7 転移学習	36
2.3.8 本研究で使用了パターン認識手法のまとめ	37
2.4 行動に関する特徴量抽出	39
2.5 滞在場所特徴に関する特徴量抽出	41
3 パーソントリップ調査を用いた人口統計学的属性の推定結果	44
3.1 特徴量の抽出	46
3.2 各属性推定手法の推定結果と精度	48
3.2.1 行動特徴のみ利用し属性推定した時の結果と精度	50
3.2.2 滞在場所特徴を考慮し属性推定した時の結果と精度	53
3.3 パーソントリップ調査を元に構築した属性推定モデルのまとめ	61
4 GPS データを用いた人口統計学的属性の推定結果	64
4.1 特徴量の抽出	65
4.2 各属性推定手法の推定結果と精度	65
4.2.1 行動特徴のみ利用し属性推定した時の結果と精度	66
4.2.2 滞在場所特徴を考慮し属性推定した時の結果と精度	70
4.2.3 パーソントリップ調査から構築したモデルを GPS データに適用した時の結果と精度	76
4.3 GPS データを用いた属性推定モデルの構築のまとめ	77

4.3.1	本研究で構築した属性推定モデルの構築のまとめ . . .	78
-------	------------------------------	----

5	結論	79
5.1	本研究の成果	79
5.2	本研究の課題と展望	80
5.3	性別推定モデルにおける主成分と元の特徴量の関係	89

表 目 次

1	人の行動に関するデータ	17
2	性別の属性コード	18
3	年齢の属性コード	19
4	職業コード	19
5	GPS データに付与されている被験者属性の対応表	20
6	真偽表	27
7	各学習手法の設定しなければならないパラメータ	28
8	各パターン認識手法のまとめ	37
9	推定モデル構築手法の特徴	38
10	ヒュベニの公式に用いる値	40
11	各クラスタの意味とクラスタサイズ	43
12	性別と年代のクロス集計表	44
13	年代と職業のクロス集計表	45
14	性別と職業のクロス集計表	45
15	使用データと東京都五歳階級別人口の相関検定の結果	46
16	パーソントリップ調査から抽出した基本統計量	47
17	与えられた属性をそのまま使用し属性推定モデルを構築した結果	48
18	性別推定の結果	49
19	年代推定の結果	49
20	職業推定の結果	49
21	行動特徴のみ利用した時の性別推定モデルの結果	51
22	行動特徴のみ利用した時の年代推定モデルの結果	51
23	行動特徴のみ利用した時の職業推定モデルの結果	51
24	各推定モデルにおいて変数重要度が高かった上位5つの変数	52
25	滞在場所を考慮した時の性別推定モデルの結果	54
26	性別推定モデルにおける変数重要度	55
27	滞在場所を考慮した時の年代推定モデルの結果	57
28	年代推定モデルにおいて変数重要度の高かった変数と変数重要度	57
29	滞在場所を考慮した時の職業推定モデルの結果	60
30	職業推定モデルにおいて変数重要度の高かった変数と変数重要度	61
31	各属性において最も性能の良かった推定モデル	62
32	変数重要度の高い変数と変数重要度	63
33	性別と年代のクロス集計表	64
34	年代と職業のクロス集計表	64
35	職業と性別のクロス集計表	65
36	行動特徴のみ利用した時の性別推定モデルの結果	67
37	性別推定モデルにおける変数重要度の高い特徴量と変数重要度	67
38	行動特徴のみ利用した時の年代推定モデルの結果	69

39	年代推定モデルにおける変数重要度の高い特徴量と変数重要度	69
40	行動特徴のみ利用した時の職業推定モデルの結果	70
41	滞在場所を考慮した時の性別推定モデルの結果	71
42	性別推定モデルにおける変数重要度	72
43	滞在場所を考慮した時の年代推定モデルの結果	74
44	滞在場所を考慮した時の職業推定モデルの結果	76
45	パーソントリップ調査を元に構築した属性推定モデルを GPS データに適用した結果	77
46	GPS データから構築された属性推定モデルのうち、最もよい 識別性能だった手法	78
47	行動履歴情報別の最も良い属性推定モデル	79
48	人の行動に関するデータ	80
49	滞留時間で集計した時の場所に関する特徴量の基本統計量	89
50	滞在回数で集計した時の場所に関する特徴量の基本統計量	89
51	性別と職業のクロス集計表	90
52	年代と職業のクロス集計表	90
53	年代と性別のクロス集計表	90
54	抽出した自宅勤務地の平均に関する基本統計量	91
55	抽出した平日の時間帯別行動距離の平均に関する基本統計量	92
56	抽出した休日の時間帯別行動距離の分散に関する基本統計量	93
57	抽出した自宅勤務地の分散に関する基本統計量	94
58	抽出した平日の時間帯別行動距離の分散に関する基本統計量	95
59	抽出した休日の時間帯別行動距離の分散に関する基本統計量	96
60	第2主成分と、元の特徴量との関係	96
61	第3主成分と、元の特徴量との関係	97
62	第1主成分と、元の特徴量との関係	97
63	第37主成分と、元の特徴量との関係	97
64	第9主成分と、元の特徴量との関係	97

目 次

1	今までに実施されたパーソントリップ調査の都市と回数	9
2	パーソントリップ調査を利用した, コミュニティバスの検討 (東京都圏交通計画協議会より引用)	10
3	パーソントリップ調査におけるトリップ例 (国土交通省, PT 調査とは?から引用)	10
4	播磨地域全体の人の動き (姫路市, 播磨都市圏パーソントリップ調査より引用)	11
5	播磨地域の事業所数の4次メッシュ集計	11
6	PT 調査の確率的時空間配分データの作成方法 (東京大学空間情報科学研究センター人の流れプロジェクトよ り引用)	13
7	日本国内の利用可能な PT 調査データセット (空間配分版) (東京大学空間情報科学研究センター人の流れプロジェクトよ り引用)	13
8	日本国外の利用可能な PT 調査データセット (空間配分 版) (東京大学空間情報科学研究センター人の流れプロジェクトよ り引用)	14
9	観光庁が実施した広域観光分析の対象エリア (観光庁 GPS データを用いた観光行動分析より引用)	15
10	リアルタイム混雑状況 (Google マップ)	16
11	VICS の概要 (一般財団法人 道路交通情報通信システムセンターより引用)	16
12	使用した GPS データの男女割合	21
13	使用した GPS データの年齢分布	21
14	使用した GPS データの職業割合	22
15	使用した事業所・企業統計調査のデータ	23
16	世田谷区における生産年齢人口	24
17	株式会社マイクロベースが作成した 500m メッシュにおける生 産年齢人口	24
18	性別・5 歳階級別人口データの一部	25
19	人の行動に着目した特徴抽出	26
20	本研究で検討した, 推定モデル構築パターン	26
21	データの使用方法	27
22	モデル構築段階で試した分類手法の一覧	28
23	Neural Network の構造	29
24	パーセプトロンの構造	30

25	Random Forest の構造	31
26	SVM の原理	32
27	ソフトマージン SVM の二次元解釈	33
28	主成分分析の例	34
29	主成分の方向	34
30	転移学習の概要	36
31	転移学習の種類	37
32	行動に関する特徴抽出手順	41
33	場所特徴の集約, クラスタリング	42
34	場所特徴の抽出方法	42
35	場所特徴データの構築結果	43
36	抽出した変数の相関	48
37	職業ラベルの属性集約	50
38	行動距離の違いによる男性判定率の変化	52
39	行動距離の違いによる主婦判定率の変化	53
40	行動距離の違いによる 20 代判定率の変化	53
41	クラスタ番号 6 の地域	55
42	クラスタ 6 の滞在時間の違いによる男性判定率の変化	56
43	帰宅時間の違いによる 30 代判定率の変化	58
44	退社時間の違いによる 30 代判定率の変化	58
45	帰宅時間の違いによる 10 代判定率の変化	59
46	退社時間の違いによるサラリーマン判定率の変化	61
47	パーソントリップ調査と GPS データで取得できるデータの違い	65
48	朝 7 時の行動傾向の違いによる男性判定率の違い	68
49	行動距離や勤務地訪問回数の違いによる男性判定率の変化	68
50	PC2 のとる値の違いによる, 男性判定率の違い	72
51	本稿で取り上げたテーマに関するロードマップ	81
52	今後の展望と参考手法	81

1 序論

1.1 研究の背景

近年、災害発生時の緊急対応や、交通状況の詳細な把握、マーケティング活動の支援資料、アドテクノロジーに代表されるデジタルアドバイジングの観点から、非集計の位置情報に対する産学官の関心は非常に高くなっている。特に「人の流れ」に関する非集計位置情報は利活用方法に関する活発な議論がなされている。これは位置情報等通知機能等の義務化により、総務省が全ての携帯電話にGPSチップを搭載することが義務付けているからである。[1]

特に近年はスマートフォンの普及により内蔵されたジャイロ等の様々なセンサからユーザーの行動情報を取得できる上に、スマートフォンのアプリケーション開発も容易であることから、アプリケーション開発によりユーザーの行動履歴を収集することが非常に容易になりつつある。大量に収集された行動情報を利用することで、様々な集客施設の時間帯別の来訪者の行動特徴がわかる。広域・俯瞰的な分析を行うことで、都市の活動を把握することができ、具体的には感染症キャリアの行動分析と対応立案の資料として有用である。

他にも、携帯電話のGPSデータをメッシュ集計した「混雑統計」(ゼンリンデータコム)は、企業のマーケティングデータとして利用されている[2]。株式会社ナイトレイは位置情報付きSNSデータを用いて、ユーザーから発信された緯度経度間の経路を補間して、擬似人流データとしてライセンスフリーで公開している[3]。観光庁では、GPSによる位置情報を活用し、観光客の行動、動態がわからない既存の調査票による調査を代替する政策を行っている。[4] また、学術分野では、長尾ほか[5]が行った、GPSログを用いた旅行者の周遊型観光の実態把握の分析や、山本ほか[6]はGPSログを分析し、新宿御苑における利用者の行動パターンについて分析を行った。[6] このように人の行動分析の関心は産学官全てにおいて高まっていることがわかる。

しかし、位置情報発信機能を持つ機器からは緯度経度情報や角速度情報しか取得できない。よって混雑度や推定居住地、推定勤務地を元にした分析や考察が限界である。そのため詳細な分析を行うには様々な属性情報を付与する必要がある。とくに人口統計学的属性と呼ばれる、性別、年齢、職業の3属性は様々な分析において、基本となる切り口であるため特に重要である。しかし、ユーザーの属性付きの人の行動履歴データはデータの性質上、非常に機密性が高い一方、しかし使用に対する高いニーズがあるため、様々な人が使用できるようになるには様々な困難を乗り越えなくてはならない。

この困難を克服する手段として、属性をユーザーから直接取得するのではなく、行動履歴を元にして推定する手段がある。しかし推定モデルを構築するには大量の属性付き行動履歴データが必要になるため、推定モデルを構築することは難しい。松尾ほか[7]はセンサの情報を使用し屋内における人の行動特徴から人々の属性を推定したが、センサで測位する情報が細かいため広域

におけるモデル適用の考察が難しい。

大規模、広域の人の流れの調査で最も利活用が進んでいるのがパーソントリップ調査である。この調査は調査票配布形式の調査方法であり、トリップと呼ばれる行動単位を用いて人々の行動を統計情報として整理している。このデータは人の行動履歴と属性情報が紐付いている非常に有用なデータである。しかし、パーソントリップ調査で取得できる人の行動情報は起終点のおおまかなエリアのみである。そのため、広域分析には適するが、一人ひとり細かく観察する分析には適さない。このように人の行動に着目した分析は社会的関心が非常に高い分析だが、現状では乗り越えなくてはならない課題が非常に多い。

1.2 人の流れに関するデータと既往研究

データ保存機器の容量単価の下落や様々な機器にセンサがつけられたことから、様々なデータが日々保存、分析されている。それは通常用いられていた構造化データだけでなく、非構造化データもある上、紙ベースのデータも存在する。本章では人の流れに関するデータセットについて、既往研究を交えながら紹介する。

1.2.1 パーソントリップ調査

一定の調査対象地域内において人の動きを知るパーソントリップ調査は交通に関する実態調査としては最も大きな調査の一つである。この調査は昭和42年に広島都市圏で大規模に実施されて以来、日本全国で数多く実施されている。図1はこれまでに全国で行われたパーソントリップ調査実施都市と回数である。

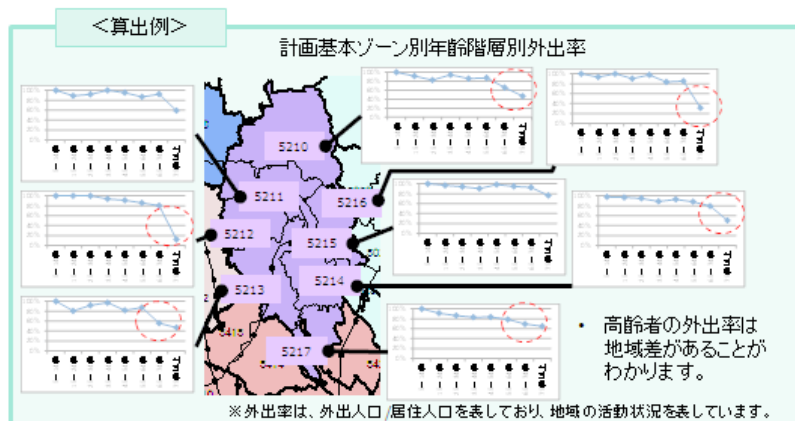


図 2: パーソントリップ調査を利用した、コミュニティバスの検討
(東京都市圏交通計画協議会より引用)

パーソントリップ調査はトリップという概念を用いて、人々の行動を構造化して捉える。トリップとはある目的をもって起点から終点へ移動する単方向移動を表す概念である。[9] 出勤のような一つのトリップを一つの目的の達成に起因するトリップをリンクトリップと呼び、交通手段による移動単位をアンリンクトリップと呼ぶ。よって一つのリンクトリップに1つ以上のアンリンクトリップが含まれることになる。これにより、交通量調査のようなある地点の定点観測による流動変化ではなく、人に着目して移動の変化を分析することができる。

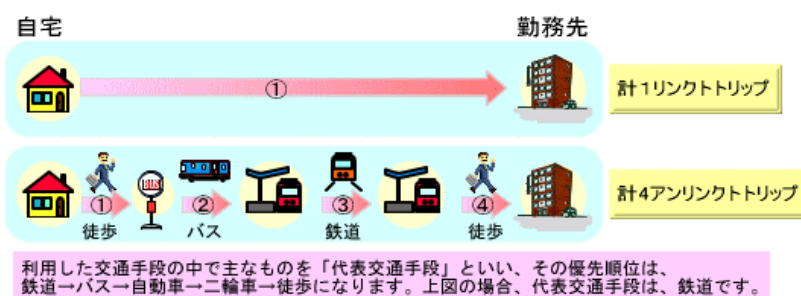


図 3: パーソントリップ調査におけるトリップ例
(国土交通省, PT 調査とは?から引用)

次にパーソントリップ調査の活用を公共団体での例と学術研究の例に分けて取り上げる。姫路市は平成 18 年に行われた、播磨都市圏パーソントリップ調査をインターネット上で公開している。[10] 図 4 は播磨地域全体のトリップ発集中量を集計した図である。図 4 の通り、播磨地域内に居住地を持つ人のほ

とんどが播磨地域内で移動している。また、神戸、阪神、淡路地域といった近隣の大都市圏への移動が多いことがわかる。



図 4: 播磨地域全体の人の動き
(姫路市, 播磨都市圏パーソントリップ調査より引用)

播磨地域内に焦点を当てると、図 4 の通り姫路市, 加古川市, 明石市の瀬戸内海に面した都市を中心に人の動きが多くなっている。これは臨海部に播磨地区で最も輸送量の多い山陽本線が走っている上に、図 5 の通り事業所数が近隣より多いことから、産業的に活発な地域であるためである。

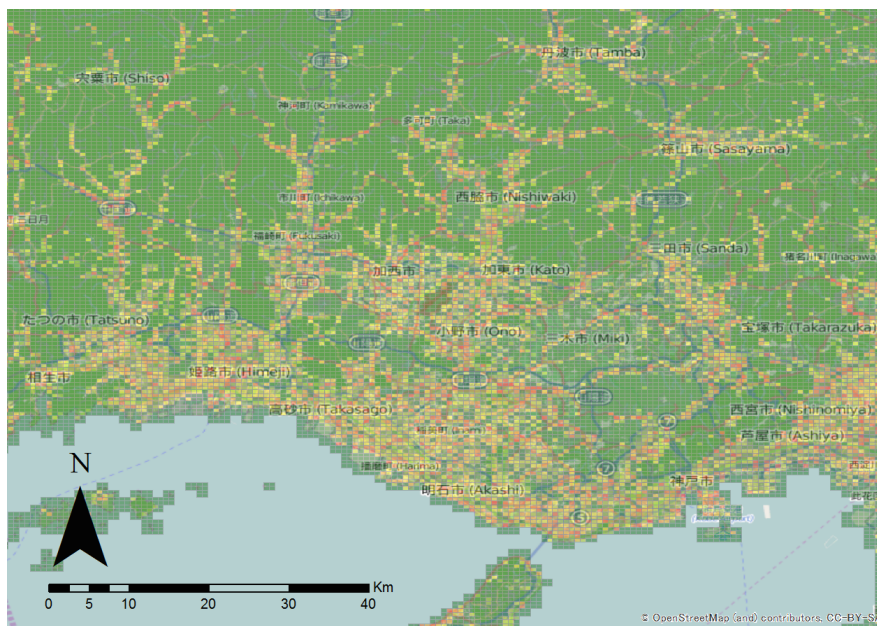


図 5: 播磨地域の事業所数の 4 次メッシュ集計

大佛, 島田 [11] はパーソントリップ調査を用いて地震被害の想定を行った。パーソントリップ調査は平日の1日分のみ行われる調査のため, 休日の人の行動解析には利用できないが, 「ある人の休日におけるトリップは, 他の人の平日におけるトリップで代替できる」という仮説のもと, 行動パターンを分類し, 平休日別駅乗降客数データを用いて, 各人の職業属性を元に休日トリップを構築した。これにより構築された休日トリップと平日トリップを用いて東京都世田谷区の建物倒壊による被災者数を推定した。その結果, 世田谷区のような住宅が多く立地する地域においては休日のほうが被害が大きくなる可能性がある一方, 平日の昼間は災害時要援護者の割合が高くなり, 救助活動については平日のほうがリスクが高いことがわかった。

齊藤ほか [12] はパーソントリップ調査のデータ粒度を細かくするために, 入口来街頻度と回遊パターンをI-射影モデリングにより逆問題へ定式化し, 混合制約の元での循環反復比率調整法により同時推定を行った。これにより, 来街地ベースサンプリングに基づいた, パーソントリップ調査からOD分布交通量を推計する課題へと貢献した。

このように, パーソントリップ調査を用いた都市の分析は俯瞰的分析には有用なものの, 詳細な分析を行うには前処理として統計的な推定を行う必要がある。東京大学空間情報科学研究センターはパーソントリップ調査を用いた詳細な人の流れデータセットを提供している。このデータは国土交通大臣や各都市圏の交通計画協議会, 国際協力機構の許諾を元に, 図6の手順で作成されている。図6の通り, アンリンクトトリップベースで起終点のゾーン代表点の時空間位置をジオコーディングし, 最短経路を元に経路探索を行い, 1分ごとの位置を各ネットワークの詳細データを元に内装を行うことで作成している。本データセットは国内だけでなく, マニラやハノイ, ダッカといった国外データも利用可能であり, 東京大学空間情報科学研究センターに共同研究申請を行う事により全24箇所の地域のデータが利用可能になる。(図7, 図8)

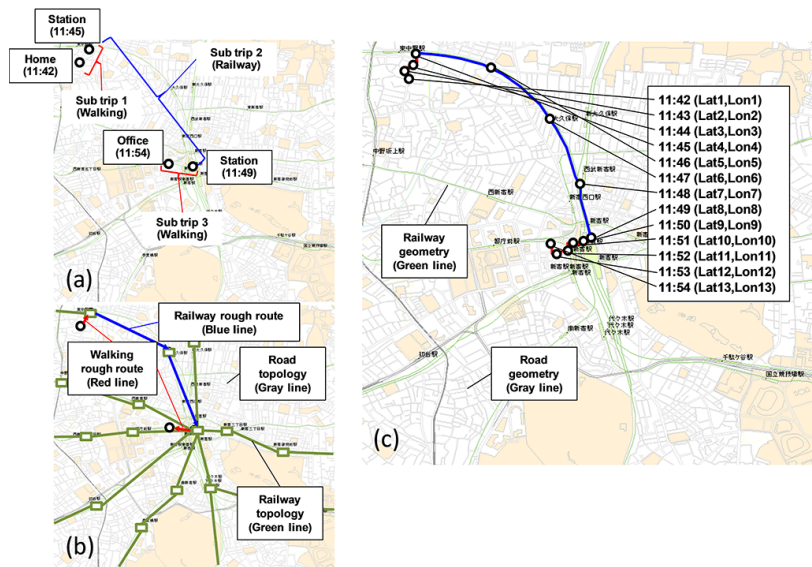


図 6: PT 調査の確率的時空間配分データの作成方法
 (東京大学空間情報科学研究センター人の流れプロジェクトより引用)

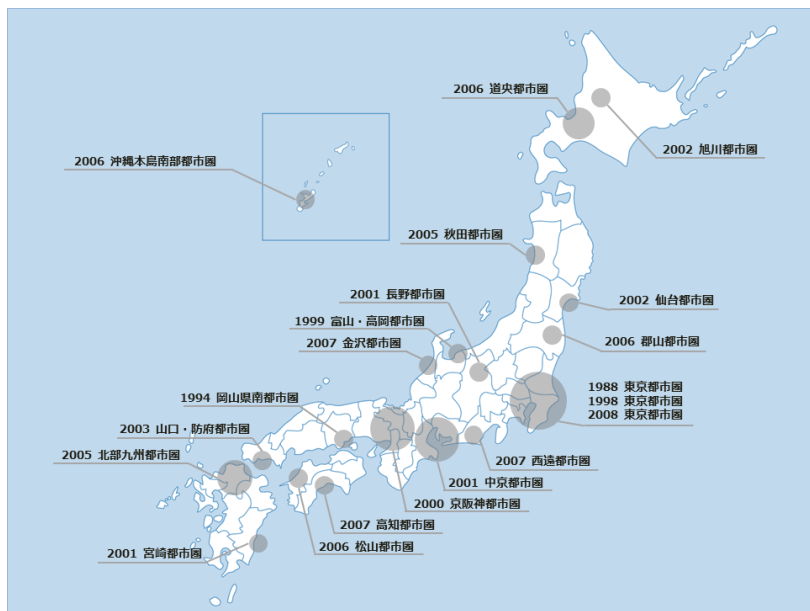


図 7: 日本国内の利用可能な PT 調査データセット (空間配分版)
 (東京大学空間情報科学研究センター人の流れプロジェクトより引用)

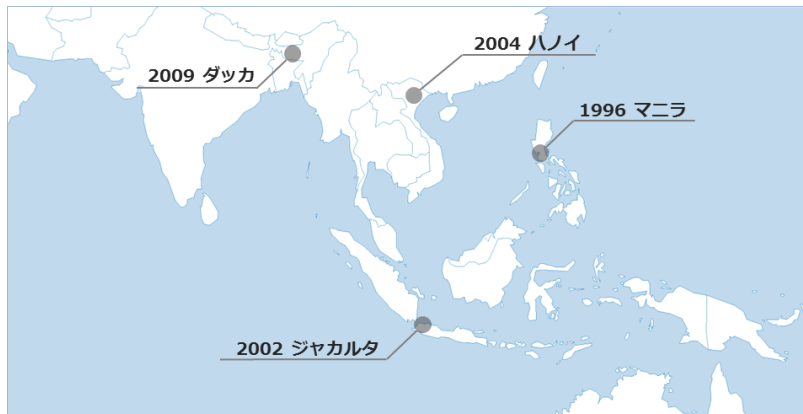


図 8: 日本国外の利用可能な PT 調査データセット (空間配分版)
 (東京大学空間情報科学研究センター人の流れプロジェクトより引用)

1.2.2 GPS データ

GPS(Global Positioning System) とはアメリカ合衆国運用による衛星測位システムであり, 準同期軌道にある複数の GPS 衛星からの信号を受信機で受信することにより受信者の現在位置を把握することができる. 現在位置の位置は緯度, 経度, 高さで一意に示せるため, 現在位置を知るためには最低三個の衛星から電波を受ける必要がある. 4 つ以上の衛星から電波を受けることが可能ならさらに正確な位置を知ることができる. しかし, 都市部では高層ビル等の施設の影響により, 電波の乱反射が起き不正確な位置を示すことがある. しかし, 2018 年に運用開始予定の準天頂衛星システムにより測位精度が向上するとされている [13]. これは少なくとも 1 台を日本のほぼ真上に準天頂衛星を配置することで受信者は地上障害物の影響を受けにくい電波の受信が可能になる.

金杉ほか [14] は利用者のオプトインを前提に, 個人が自分の情報を利活用していくことを想定し, 利用者負担が少なく, 大規模に利用できる可能性を持つ携帯電話の通信記録 (Call Detail Records: CDR) を対象に, 個人行動分析への適用可能性を探った. CDR は利用者負担が小さいが空間解像度が粗いため, 空間解像度が高いがバッテリー消費が大きい GPS データも同時に取得し精度を検証した.

GPS データを用いた分析は Ashbrook ほか [15] や Zheng ほか [16] が行った被験者のよく訪れる滞在地点の抽出や Liao ほか [17] や Jiang ほか [18], Ye ほか [19] が行った行動パターンの推定, Schussler ほか [20] が行った徒歩や電車といった交通モードの推定, 行動予測などがある. これらは教師なし学習手法を用い, 探索的に滞在地点を抽出している. 他にも, 予め対象エリアを選定し, エリア内に入ったユーザーを滞在とみなす方法をとっている分析もある.

観光庁と株式会社野村総合研究所は株式会社ゼンリンデータコム社提供の混雑統計を用いて観光客の行動分析を行った。[4] 混雑統計とは株式会社 NTT ドコモが提供するオート GPS サービスを利用している利用許諾取得済みユーザーの位置情報を秘匿処理を行い集計データに加工したデータである。この混雑統計を元に、観光地に滞在した人を業務目的と観光目的に分類し、観光目的ユーザーの行動傾向を明らかにした。しかし、図9の全国8つの対象エリアにおける観光行動分析は、性別、年代といった属性情報が付与されておらず、対象観光エリアに含まれる地方自治体からはこれらの属性を考慮した分析も行いたい、という声があがっている。[4]

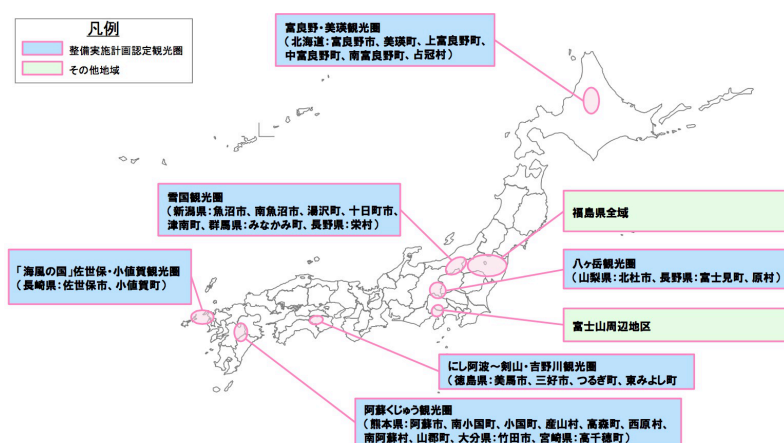


図 9: 観光庁が実施した広域観光分析の対象エリア (観光庁 GPS データを用いた観光行動分析より引用)

産業界での大規模 GPS データ活用事例の最たるものが Google 社提供の Google マップである。図 10 は Google マップで交通状況レイヤを表示した時の画面である。赤いほど低速、つまり渋滞を起こしている道路であり、緑は渋滞していない道路を表している。これは Google 社が提供しているスマートフォン用 OS の Android から発信される匿名化された GPS データをリアルタイムに収集し、単位時間あたりの移動距離を元に渋滞の判定を行っている。[21] [22] また過去に収集、分析した情報を元に曜日と時間別の渋滞状況も提供している。このサービスにより、道路交通情報通信システムセンター提供の VICS と呼ばれる全国約 4 万箇所の定点観測ビーコンで取得する道路交通情報 [23] より精度が高く交通状況を把握、予測することができる。

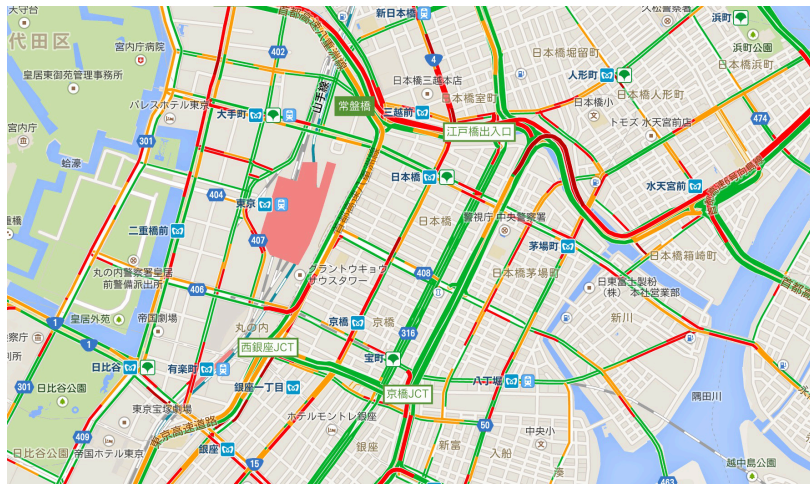


図 10: リアルタイム混雑状況(Google マップ)



図 11: VICs の概要
(一般財団法人 道路交通情報通信システムセンターより引用)

1.3 研究の目的

本研究の目的はユーザーの移動履歴を入力としたときに、ユーザーの人口統計学的属性を出力とするモデルを構築することである。このモデルを構築することで、ユーザーの移動履歴の取得容易性を保ったまま、属性情報の取得困難性を解決する。ユーザーの移動履歴情報は上述の通り、パーソントリップ調査と GPS データの二種類ある。表 1 のようにパーソントリップ調査はサン

ブル数が膨大な一方で、位置に関する集計単位がゾーンと粗い上に取得期間が平日1日のため、ユーザーの行動パターンを考慮した分析が行えない。一方でGPSデータはサンプル数が少ないが位置情報が緯度経度で収集でき、複数日取得できるためユーザーの行動パターンを考慮した分析が行える。本研究では特徴の異なる2つの人の流れデータセットに対し人々の人口統計学的属性の識別モデルを構築し、データの特徴による識別能力の違いについて考察する。

表 1: 人の行動に関するデータ

データ名	人数	ラベル	位置情報の粒度
PT 調査	膨大 (約 60 万人)	あり	粗い (ゾーン単位)
GPS データ	少数 (数百人程度)	あり (アンケートによる)	細かい (緯度経度)

2 人口統計学的属性の推定方法

本研究はユーザーの移動履歴を元に人口統計学的属性を推定するモデルの構築を試みた。ユーザーの移動履歴はパーソントリップ調査と GPS データを使用し、それぞれのデータを元に属性推定モデルを構築した。パーソントリップ調査と比較し、GPS データは空間解像度が細かいため、本研究の構成は、以下のようなになる。

1. パーソントリップ調査による属性推定モデルの構築
 - 行動特徴のみ使用して属性推定モデルを構築
 - 滞在特徴考慮済み属性推定モデルの構築
2. GPS データによる属性推定モデルの構築
 - 行動特徴を使用した属性推定モデルの構築
 - 滞在場所特徴考慮済み属性推定モデルの構築
 - パーソントリップ調査を元に構築した属性推定モデルの GPS データへの適用

2.1 本研究で使用したデータセット

2.1.1 人の流れに関するデータセット

パーソントリップ調査

本研究で用いた GPS データは東京大学空間情報科学研究センター提供の【空間配分版】2008 年東京都市圏 人の流れデータセットである。本データは東京都市圏交通計画協議会提供の 2008 年東京都市圏 PT データ [24] をもとに、独自加工した人の流れデータである。[25] 空間配分版は従来ゾーン代表点で再現していた起終点の位置を住宅地図データ等を用いてゾーン内部に確率的に再配置したデータである。

このデータに付与されている人口統計学的属性は性別、年代、職業の三種類でありそれぞれ表 2、表 3、表 4 に示す。

表 2: 性別の属性コード

コード	内容
1	男性
2	女性
9	不明

表 3: 年齢の属性コード

コード	内容	コード	内容
0	0 歳以上 5 歳未満	9	45 歳以上 50 歳未満
1	5 歳以上 10 歳未満	10	50 歳以上 55 歳未満
2	10 歳以上 15 歳未満	11	55 歳以上 60 歳未満
3	15 歳以上 20 歳未満	12	60 歳以上 65 歳未満
4	20 歳以上 25 歳未満	13	65 歳以上 70 歳未満
5	25 歳以上 30 歳未満	14	70 歳以上 75 歳未満
6	30 歳以上 35 歳未満	15	75 歳以上 80 歳未満
7	35 歳以上 40 歳未満	16	80 歳以上 85 歳未満
8	40 歳以上 45 歳未満	17	85 歳以上

表 4: 職業コード

コード	内容	コード	内容
1	農林水産業従事者	10	その他職業
2	生産工程・労務作業	11	園児・小学生・中学生
3	販売従事者	12	高校生
4	サービス職業従事者	13	大学生・短大生・各種専門学校生
5	運輸・通信従事者	14	主婦・主夫
6	保安職業従事者	15	無職
7	事務従事者	16	その他
8	専門的・技術的職業従事者	99	不明
9	管理的職業従事者		

GPS データ

本研究で用いた GPS データは金杉ほかの研究 [14] の実験において取得した GPS データである。このデータは移動履歴調査への参加に関する個人情報の取扱と、共同研究者の KDDI 株式会社が通信時位置情報 (CDR) を KDDI 研究所及び東京大学へ提供することへの同意を予め被験者から取得した。また被験者は以下の条件をすべて満たす人を選定した。なお、詳しくは金杉ほか [14] を参照されたい。

- 関東一都六県在住で 18 歳以上 (高校生は除く) である。

- au のスマートフォン端末を所有している
- パソコンを所有していて、インターネットに接続できること
- 1 習慣の外出頻度が週平均 1 日以上ある
- 通話, メール, Web サイト閲覧のいずれかが 1 日 1 回以上ある
- 調査期間中に海外旅行へ行く予定がない
- 本調査以外で携帯電話にアプリをインストールする調査に参加していない・調査期間中に参加する予定がない
- 通信時位置情報の取得に同意できる

本調査を始めるに先立ち、まず被験者の年齢、性別などの属性と、外出頻度や携帯電話の利用状況などのライフスタイルについて、Web アンケート形式での調査を実施した。また、調査実施期間は 2011 年 11 月 28 日から 12 月 22 日までの 25 日間で、被験者数は 184 名である。表 5 は被験者の各属性のラベルと意味の対応表であり、図 12, 図 13, 図 14 はそれぞれ、性別の割合、年代の分布、職業の割合である。なお、図 13 の赤点線は被験者の平均年齢である。

表 5: GPS データに付与されている被験者属性の対応表

コード	性別	職業
1	男性	農林漁業従事者
2	女性	生産工程・労務作業
3		販売従事者
4		サービス職業従事者
5		運輸・通信従事者
6		保安職業従事者
7		事務従事者
8		技術的・専門的職業従事者
9		管理的職業従事者
10		その他職業
11		大学生・短大生・各種専門学生
12		主婦・主夫（職業従事者を除く）
13		無職
14		その他

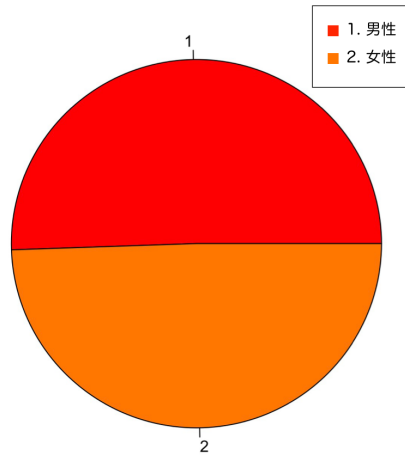


図 12: 使用した GPS データの男女割合

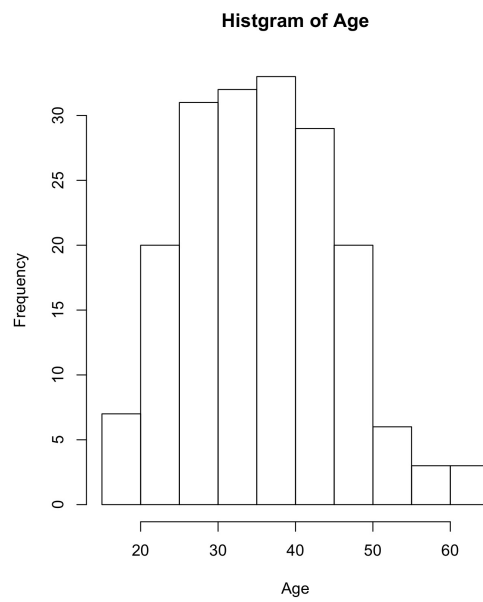


図 13: 使用した GPS データの年齢分布

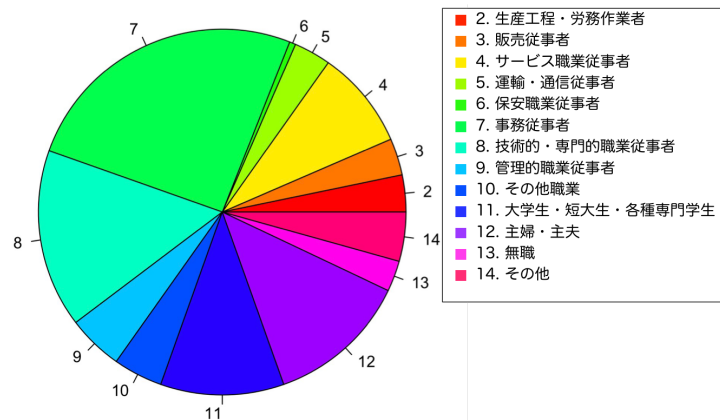


図 14: 使用した GPS データの職業割合

2.1.2 場所特徴に関するデータセット

事業所・企業統計調査

事業所・企業統計調査とは事業所及び企業の産業、従業者規模等の基本的構造を全国及び地域別に明らかにすると共に、各種標本調査実施のための母集団情報となる事業所及び企業の名簿を整備することを目的として行われる事業所及び企業についての国の最も基本的な統計調査である。[26] 本研究では公益財団法人 統計情報研究開発センター提供の事業所・企業統計調査の4次メッシュ集計版を利用した。[27] このデータは4次メッシュで、資本金規模、小業種別、大業種別のデータがある。この内、大業種別のデータを利用した。大業種の分類は以下の通りである。また図 15 は使用したデータのサンプルである。

- 鉱業
- 建設業
- 製造業
- 電器

- 情報サービス業
- 運輸業
- 小売・卸売業
- 金融業
- 不動産業
- 学術団体
- 飲食業
- 娯楽業
- 教育
- 医療
- 複合サービス業
- その他のサービス業

メッシュコード	鉱業	建設業	製造業	電気工事関連業	情報通信業	運輸業	小売業	金融業	不動産業	学術団体	飲食業	娯楽業	教育業	医療団体	複合サービス業	その他サービス業
533946014	1	39	97	0	183	39	325	28	129	262	47	43	5	5	0	116
533946121	0	72	125	0	148	44	376	56	105	168	70	32	5	10	0	86
533946002	0	12	14	2	48	12	220	29	127	87	579	57	2	9	0	55
533946214	0	55	73	0	115	27	409	65	157	125	110	34	6	7	0	68
533946011	0	20	23	1	127	19	334	23	139	187	162	60	16	4	0	74
533946312	0	67	92	0	105	9	422	31	127	100	83	43	8	4	0	77
533946001	2	25	76	1	92	64	253	43	155	145	105	60	10	5	0	70
533946112	0	44	86	0	101	29	304	57	153	138	64	29	5	4	0	84
533945264	0	63	40	0	170	5	230	16	156	206	93	51	13	9	0	44
533946013	0	21	13	0	82	15	274	23	165	136	183	70	15	10	0	84

図 15: 使用した事業所・企業統計調査のデータ

性別・5歳階級別人口500mメッシュデータ

総務省統計局が行っている国勢調査は全国の地域別人口や産業別就業者数などの統計を作成するための調査である。国勢調査は客観的なデータに基づく公正な行政を行うために実施され地方交付税の算出計算や衆議院選挙における小選挙区の確定等に用いられており、他にも公的統計の作成・推計のための情報基盤としての役割を担っている。[28] 5歳階級別人口は図16のように小地域別に集計されて総務省統計局 e-Stat において提供されている。[29] しかし、500mメッシュでは男女別の総人口のみ提供されておらず、メッシュベースで計算を行うのは難しい。株式会社マイクロベース [30] は国勢調査と住宅地図を用い、通常小地域に紐付けられている性別・5歳階級別人口情報を住宅地図に含まれる建物の容積情報を利用し500mメッシュに再配置した。図

18 は本研究で使用した, 500m メッシュにおける性別・5 歳階級別人口データの一部である.

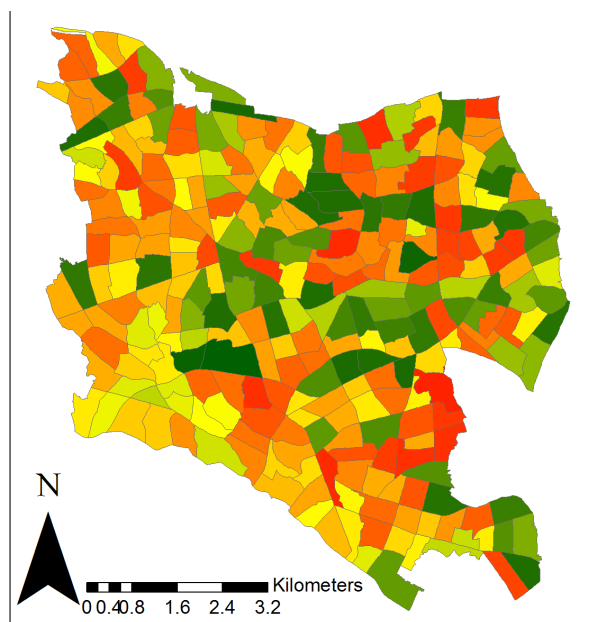


図 16: 世田谷区における生産年齢人口

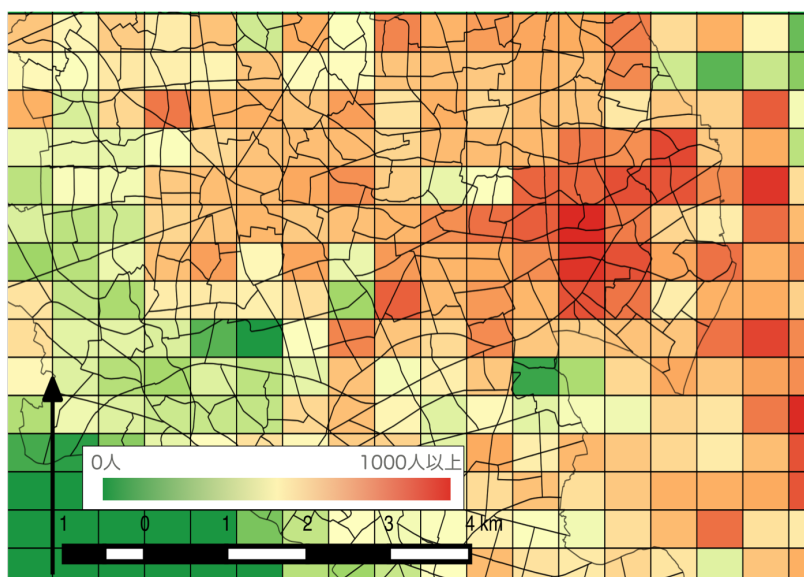


図 17: 株式会社マイクロベースが作成した 500m メッシュにおける生産年齢人口

mesh_code	m0	m5	m10	m15	m20	m25	m30	m35
533936814	416	209	141	138	197	379	569	711
533935082	173	118	83	68	250	515	546	580
533936743	383	198	115	84	94	200	530	717
533945244	90	65	52	96	265	565	530	447
533945774	104	72	60	102	388	522	513	403
533925824	223	151	112	141	230	418	508	504
533945872	135	91	101	127	351	511	507	439
533935084	161	110	91	90	247	483	491	537

図 18: 性別・5 歳階級別人口データの一部

2.2 手法の概要

前述のとおり人の流れデータセットはパーソントリップ調査を元に構築された詳細な人の流れデータである。人の流れデータセットを用いて人口統計学的属性推定モデルを構築し、GPS データへ適用した。人の流れデータセットはパーソントリップ調査を元に構築されたデータであるため、データの期間は平日 1 日のみである。よって、図 19 の人の移動に着目した特徴の内、1 日の移動履歴に関する特徴と、立ち寄り先に関する特徴の 2 つが人の流れデータセットより抽出できる。抽出した特徴を入力、推定する人口統計学的属性を出力とすると、モデルの構築は機械学習でいうクラス分類問題に帰着される。本研究で推定する人口統計学的属性は性別、年代、職業の 3 つであるため、構築されるモデルは 3 つとなる。クラス分類問題を解く手法は非常に多岐に渡るが、本研究では Neural Network, Support Vector Machine, Random Forest の 3 つを試し、構築されたモデルの分類性能が最も高い手法を採用した。モデルの分類性能は図 21 の通り、モデルを構築するための学習データとモデルを検証するために検証データにランダムに分け、学習データで構築されたモデルを元にして検証データへ適用し、検証データへ適用した時の正解率とした。また、これらの手法は事前に設定する必要のあるハイパーパラメータがある。最適なパラメータは解析的に求められないため、グリッドサーチとよばれる最適パラメータの決定手法を用いた。また、使用した特徴量である、一日の移動履歴に関する特徴、立ち寄り先に関する特徴は非常に膨大な数であり、特徴量間が独立とはいえない。そのため主成分分析を行い特徴空間を直交化した場合の識別性能についても考察した。よって図 20 の通り、機械学習の手法の違い、内部計算のパラメータの違い、前処理の違いによる、計 6 パターンを試行したほか、使用するデータの違い、滞在地特徴の抽出方法の違い、集計方法の違いにより、人口統計学的属性ごとに 51 パターンの推定モデルを構築した。

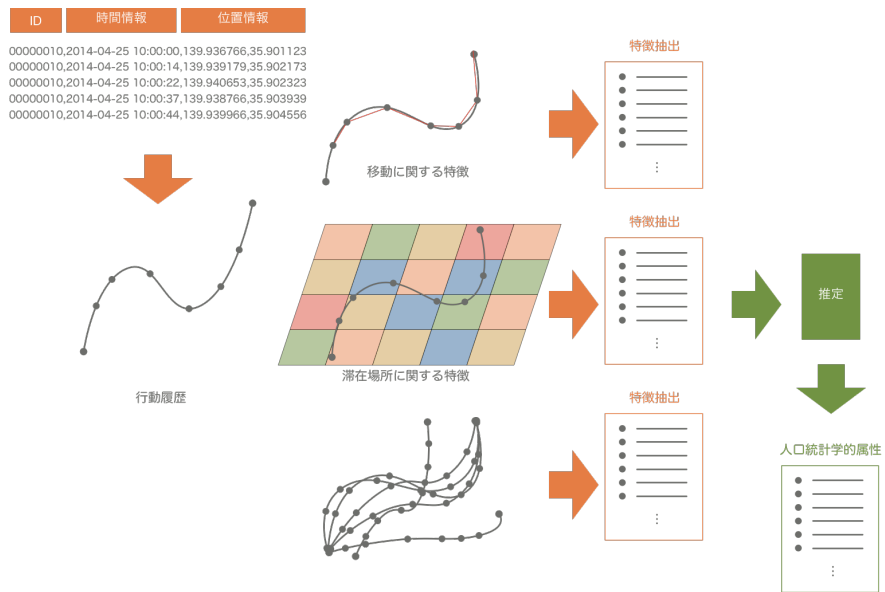


図 19: 人の行動に着目した特徴抽出

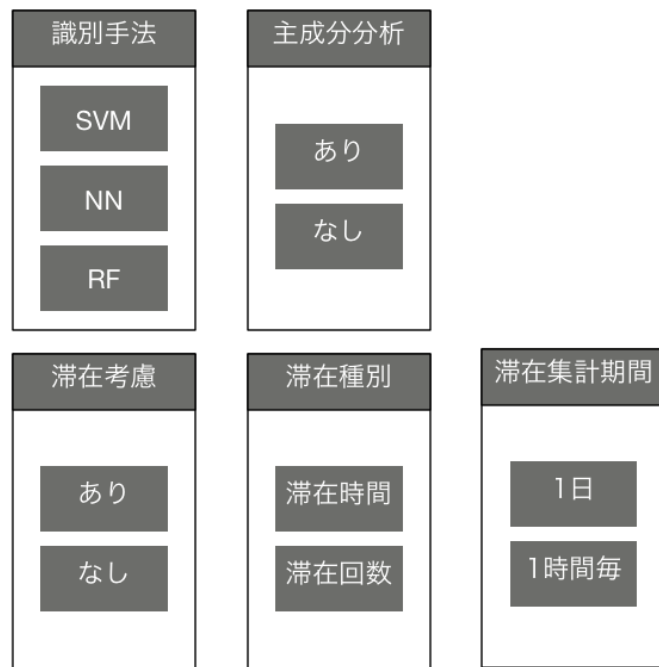


図 20: 本研究で検討した、推定モデル構築パターン

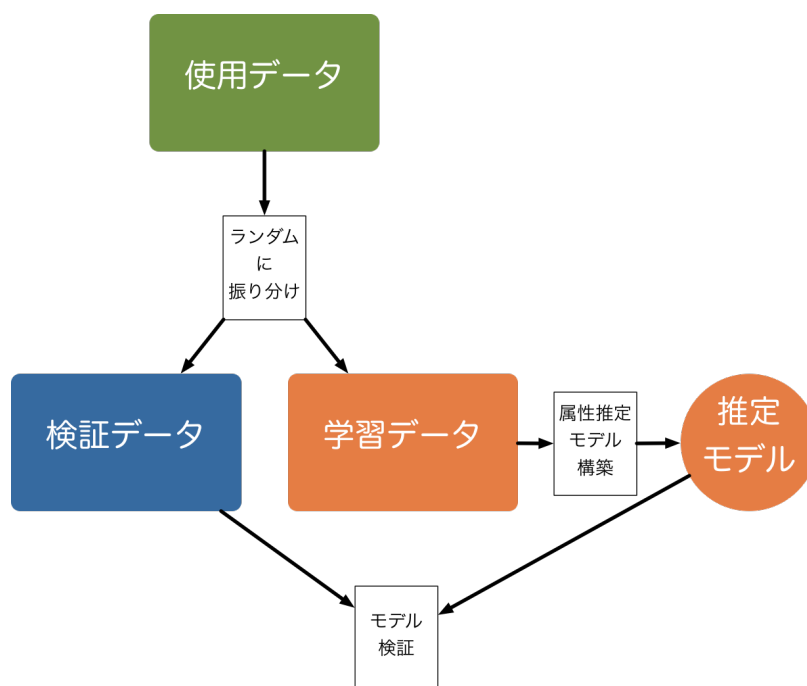


図 21: データの使用方法

構築した推定モデルを検証用データに適用し、すでに性別、年代、職業の3つの真値と比較し、Accuracy, Recall, Precision, F-measureを計算する。このうち、RecallとPrecisionの調和平均であるF-measureを属性推定モデルの性能とする。表6のように真値と予測値の真偽を表にまとめると、Accuracy, Recall, Precision, F-measureの計算方法は以下の通りである。式より、Accuracyは検証データ適用時の全体における正解率であり、Recallは真値が1であるときに正解した割合、Precisionは予測値が1であるときに正解した割合を意味する。

表 6: 真偽表

	答えが正	答えが負
予測が正	True Positive(TP)	False Positive(FP)
予測が負	False Negative(FN)	True Negative(TN)

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\begin{aligned}
 Precision &= \frac{TP}{TP + FP} \\
 Recall &= \frac{TP}{TP + FN} \\
 F - measure &= \frac{2Recall \times Precision}{Recall + Precision}
 \end{aligned}$$

2.3 本研究に使用したパターン認識手法

本論文は多岐にわたる手法を試しデータ構築, モデル構築を行っている. ここでは各手法について簡単な概要を示した上, 詳細を知りたい場合の参考文献を示している. 図 22 は本論文の流れにそって利用した手法の一覧である. なお, 本論文に用いた手法のみの説明であり, 例えば Random Forest Regression [31] や Recurrent Neural Network [32] のような類似手法については説明しない.

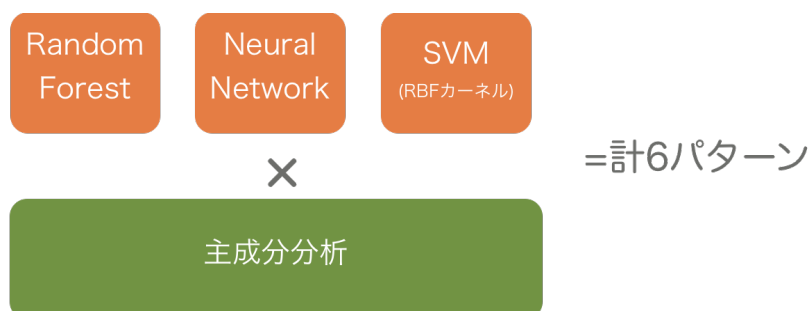


図 22: モデル構築段階で試した分類手法の一覧

本論文で用いた手法は事前に設定しなければならないパラメータが多く解析的に求めることが難しい. よってグリッドサーチという, パラメータの定義域を一定間隔で分割し, 全ての格子点において構築したモデルの性能を計測し, 最も性能が高かった時のパラメータを採用する手法を用いた. [33] 表 7 は本研究で用いた手法の事前に設定するパラメータの一覧である.

表 7: 各学習手法の設定しなければならないパラメータ

手法名	パラメータ
SVM	マージンパラメータ, カーネルパラメータ
RandomForests	生成する決定木の数, 決定木に使用する特徴量数
ニューラルネット	中間層の数, 結合荷重の初期値

2.3.1 Neural Network

Neural Network とは脳機能の特性をコンピュータ上で表現することを目指したものである。[34] 源流は生体脳のモデル化といった神経科学の分野だが、現在では乖離が激しくなり、人工 Neural Network と呼ばれる。[35] 人工的に計算機上で構築されたニューロン（ノード）は他のニューロンと人工シナプスにより結合している。Neural Network とは計算によってシナプスの強度を逐次的に変更する過程でモデル全体が学習能力を持つモデルである。図 23 は Neural Network の概要図である。

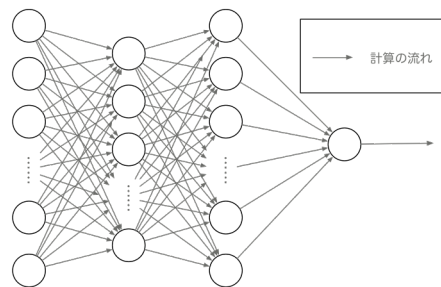


図 23: Neural Network の構造

本研究で使用した Neural Network はフィードフォワード型 Neural Network と呼ばれ、パーセプトロンを多層に積み上げた非常に一般的なモデルである。パーセプトロンとは視覚と脳の関係を模した非常にシンプルなアルゴリズムであり、単純ながらも現在でもよく用いられる非常に強力なアルゴリズムである [36]。図 24 はパーセプトロンの模式図であり、 w を重み、 x を入力、 H を閾値により出力が 1,0 のどちらかをとる関数であるとすると、

$$H\left(\sum_{i=1}^N w_i x_i - h\right) \quad (1)$$

と表せる。

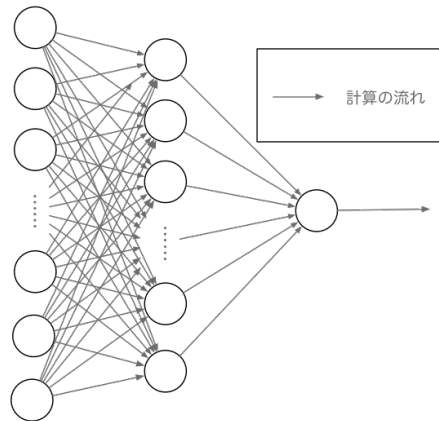


図 24: パーセプトロンの構造

しかしパーセプトロンは式 1 の通り, 線形分離のみ可能である. 線形分離とは直線の分類のことであり, 単純なパーセプトロンでは曲線による分類 (非線形分離) は行えない. しかし, パーセプトロンを多層につなぐ事で非線形分離が可能になる. これがフィードフォワード型 Neural Network である. フィードフォワード型 Neural Network を解くには誤差逆伝播法と言う手法を使う. 誤差逆伝播法は以下の手順で行われ, これにより, 適切な重みへと更新されていく. 数学的な処理手順は David ほか [37] を参照されたい.

1. 初期解を設定する.
2. 各出力ニューロン (ノード) に対し, 誤差を計算する.
3. 期待される出力値と, 実際の出力の差を計算する.
4. 上で計算された差を小さくするように, 重みを変更する.
5. 重みを変更後, 前段のノードに対し同じ計算を行う

しかし中間部分のノード数を適切に決める必要がある. この決定はノード数を変化させながら分類性能を確認し, 最も分類性能が高いノード数に設定した.

2.3.2 Random Forest

Random Forest はアンサンブル学習の一つである [38]. アンサンブル学習とは複数の弱学習器と呼ばれる, 精度は高くないが計算が早いモデルから導き出される複数の結果を用いて, 精度を向上させる手法である. Random Forest の流れは以下の手順で行われる.

1. 入力データから N 個のブートストラップサンプル B_1, B_2, \dots, B_N を作成する.
2. 各ブートストラップサンプル B_i を用いて決定木 T_i を作成する.
3. T_i を作成したデータ以外を用いて T_i の性能推定を行う.
4. T_1, T_2, \dots, T_N の性能推定を元に, 多数決をとって分類モデルとする.

Random Forest では作成する決定木の数, 決定木に使用する特徴量の数の 2 つのパラメータを事前に設定する必要がある. このパラメータはグリッドサーチにより決定した.

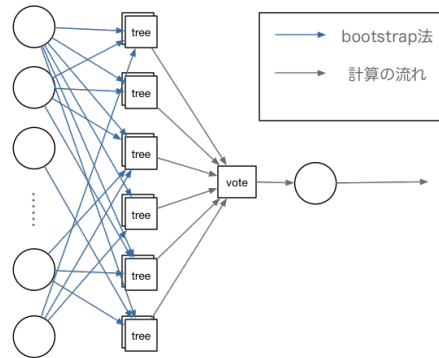


図 25: Random Forest の構造

また, Random Forest は内部構造として多数の決定木を作成するため, 多数の決定木の分割指標を元に変数の重要度を計算することができる. ある変数 X_m の重要度 $Imp(X_m)$ は決定木の数 N_T , ノード t に割り当てられた割合 N_t/N を $p(t)$, ノード t の分割点を $s_t = (X_m < c)$, $v(s_t)$ を分割点 s_t で使われた変数, ノード t の分割の際, 二分木に左に配置された割合を $p_L = N_{tL}/N_t$, 右に配置された割合を $p_R = N_{tR}/N_t$ とすると, 以下のように計算できる. [39]

$$Imp(X_m) = \frac{1}{N_T} \sum_T \sum_{t \in T: v(s_t) = X_m} p(t) \Delta i(s_t, t)$$

$$\Delta i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

ここで $i(t)$ は乱雑さの評価指標である. 本研究では決定木を CART [40] により構築したため, 評価指標はジニ係数である. 領域 R_t 内においてクラス $k \in 1, \dots, K$ に割り当てられているデータ点の割合を p_{tk} で定義すると, ジニ係数 $i(t)$ は以下のように計算できる.

$$i(t) = \sum_{k=1}^K p_{tk}(1 - p_{tk})$$

2.3.3 Support Vector Machine

Support Vector Machine は 1990 年代に Vapnik 等が考案した, Optimal Separating Hyperplane を起源とした線形識別の手法である. [41] この手法はカーネルトリックによる非線形への写像により, 非線形分離を可能とし, 様々な手法の中でも特に認識性能の優れた学習モデルの一つである. [42] Support Vector Machine(SVM) は図で示すと図 24 と同様の構造を示す. 仮に図 26 のように対象データの分類が線形分離可能だとしても, 2 つを誤りなく分ける式は一意に決まらない. しかし SVM ではなるべくデータから離れた部分で分離するようにマージンと呼ばれる概念で測り, マージンを最大とする直線の式を求める. [43]

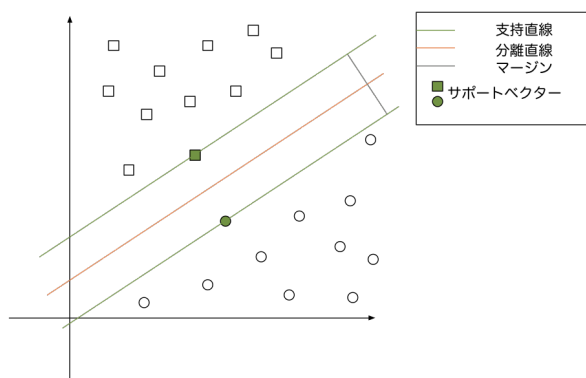


図 26: SVM の原理

本手法の便利な点は拡張が容易である所である. 実データは線形分離できないものが多いため, 上記手法を使うことが難しい. しかし SVM ではソフトマージンという, 線形分離できない事を許容しつつも誤分類したデータからの距離を最小にするように制約式を変更する. はソフトマージン SVM を解く問題の定式化である.

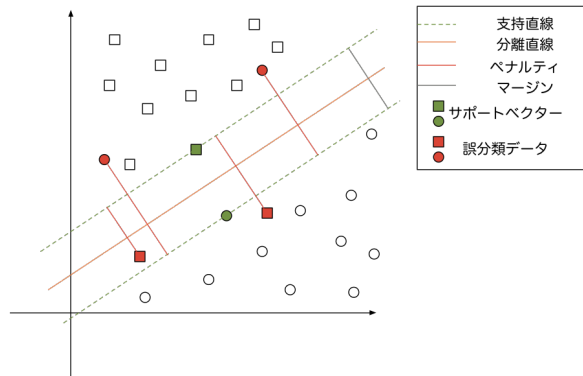


図 27: ソフトマージン SVM の二次元解釈

2.3.4 カーネル法

カーネル法とは2つのデータの間のある種の類似度を表すカーネル関数を通じてデータにアクセスするような学習モデルである [44] [43]. カーネル法を利用するには n 個のデータ x_1, x_2, \dots, x_n が与えられた時に α をモデルパラメータ, g をある関数とするとモデル $f(\cdot)$ が,

$$f(\cdot) = g\left(\sum_{i=1}^n \alpha_i k(\cdot, x_i)\right) \quad (2)$$

と表せる時に利用できる. また, $k(x_i, x_j)$ は内積として書ける必要があり,

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (3)$$

のように定義されている必要がある. [43]

カーネル法の利点は式 2, 式 3 のように, 全ての内積計算はカーネル関数を経由して行われることである. [43] よって, $\phi(x)$ を陽に計算せずに $\phi(x_i)\phi(x_j)$ を計算することが可能である. そのため特徴ベクトルの次元が非常に高い場合においても内積計算のみならば高速に計算することができる. この方法を使うことで入力データを高次元の特徴空間へ写像し, 特徴空間上での SVM, つまり線形分離を行うことで, 入力データから見れば非線形の分離曲面を得る事ができる. 詳細は赤穂の論文 [43] を参照されたい. 本論文では, 田中らで用いられたガウシアンカーネル $k(x_i, x_j) = (x_i x_j + 1)^d$ を用いて分析を行った. [45]

2.3.5 主成分分析

主成分分析とは少数の総合的指標を用いて変数間の関係や特徴を把握するための統計的な手法である [46]. たとえば図 28 のように身長と体重の2つの

変数が与えられたとする。この時身長が大きければ体重も大きくなるのが直感的にわかる。この傾向を「身体の大きさ」という一変数で示すとき、身長と体重という二変数の情報量をなるべく保ったまま行われるのが妥当である。[47] [48] つまり、得られる情報量を最大にする方向、データの分散が最も大きい方向に主成分を設定することで実現できる。

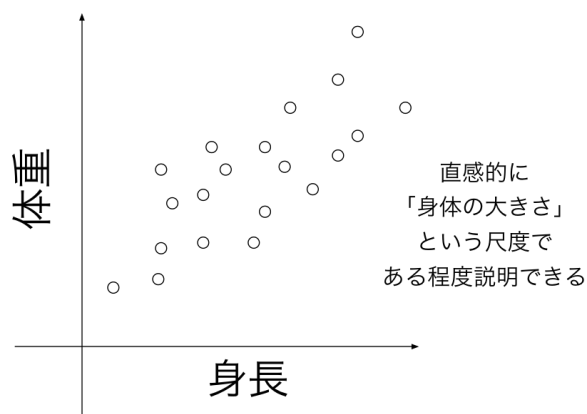


図 28: 主成分分析の例

よって図 29 で示すと新たな情報量は OB で表わせ、期待情報量を最大にする z は P 個のデータ点 $x_p (p = 1, 2, \dots, P)$ と重み $w_p (p = 1, 2, \dots, P)$ で示すと、

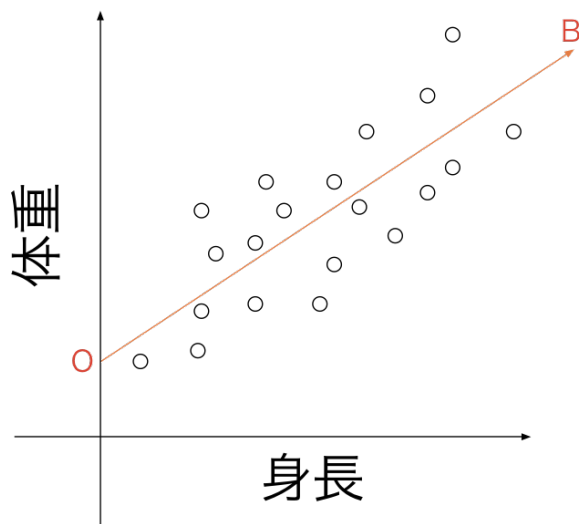


図 29: 主成分の方向

$$z = \sum_{p=1}^P w_p x_p \quad (4)$$

で表現できる。ただし、

$$\sum_{p=1}^P w_p^2 = 1 \quad (5)$$

とする。

主成分分析は、各変数の関係を主成分を用いて特徴を把握する方法なので、何個の主成分を用いれば元データの特徴を表せるのか知ることが重要である。この指標として寄与率、累積寄与率がある。寄与率とは主成分全体の分散の和のうち、その主成分が占める分散の割合であり、累積寄与率は求めたい第 m 主成分までの寄与率の和である。また、各主成分間は直交化する。本研究では抽出された全ての主成分数を採用し、変数間の線形性を除くために主成分分析を用いた。

2.3.6 クラスタリング

クラスタリングとは統計、パターン認識、データベース、データマイニング、ファジィ、そして人工知能などの分野で研究されている [49] データを分布に応じて分類する教師なし学習の一つの手法である [50]。クラスタリングには大きく分けて階層型分類と分割最適化手法に 2 分され、本研究では分割最適化手法を用いた。分割最適化手法は kmeans が最も有名である。kmeans とはクラスタ数を指定すると、データの特徴空間上で超球型のクラスタに分割する手法である。[51] この方法は以下のアルゴリズムでクラスタリングする。

1. データ $x_i \in X$ ($i = 1 \dots n$) にランダムにクラスタ番号 C_j ($j = 1 \dots K$) を付与する
2. クラスタ中心 V_j を計算する
3. 全ての x と V の距離を求め、 x_i のクラスタ番号を最も近い中心の番号に更新する。
4. 更新がされなくなったら終了

しかし、kmeans は実行時にクラスタ数を指定しなければならないほか、クラスタリング結果がクラスタ中心点の初期値に激しく依存するという問題がある。Pelleg ほか [52]、石岡ほか [53] は kmeans によるクラスタリング結果の評価をベイズ情報量規準 (BIC) を用いて定量化し、BIC をもとに最適クラスタ

数を決定する xmeans を提案した. この手法は kmeans で指定しなければならないクラスタ数を設定する必要がないため, クラスタ数が未知のデータに対して非常に有効な手法である. BIC は x をデータ点, C をクラスタ, n をサンプル数, μ を平均特徴量, V を分散共分散行列とすると 6 式のように表せる.

$$BIC = -2\log L(\hat{\theta}_i; x_i \in C_i) + 2p\log n_i \quad (6)$$

$$(\hat{\theta}_i = [\hat{\mu}_i, \hat{V}_i])$$

2.3.7 転移学習

転移学習とは, 図 30 のように求めたい問題を効率的に解くために, 別の類似した問題のデータや学習結果を再利用する方法である [54]. 1995 年ごろから機械学習の一分野として認識され始めたが, これまでに様々な呼び方がされている. Pan による転移学習のサーベイでは乱雑なこれらの呼び方を体系的にまとめている [55]. これによると, 転移学習は求めたい問題に使用するデータへのラベルの有無, 別の類似した問題に使用するデータのラベルの有無により, 図 31 の通り 4 パターンに分類される. この 4 パターンのうち, 最も標準的なのが両方のデータにラベルが付与されている場合である帰納転移学習 (Inductive Transfer Learning) であり, 他のパターンと違い両データの条件付き確率をやデータの特徴空間の定義域の一致仮定を必要としない [54]. 本研究では使用するパーソントリップ調査と GPS データの両方共ラベルが付与されているため, 帰納転移学習が適用できる. 転移学習の詳細は Pan らのサーベイを参照されたい. ① 先に簡単な

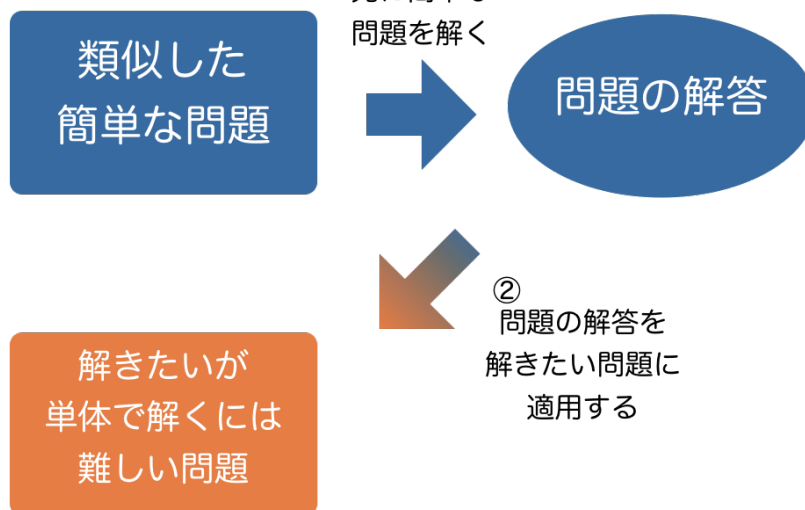


図 30: 転移学習の概要

		Does Target have Labels?	
		Yes	No
Does Source have Labels?	Yes	Inductive Transfer Learning	Transductive Transfer Learning
	No	Self-Taught Learning	Unsupervised Transfer Learning

図 31: 転移学習の種類

2.3.8 本研究で使したパターン認識手法のまとめ

表 8 は本研究で使したパターン認識手法のまとめである。本研究では大きく 7 つの手法を組み合わせで推定モデル構築を行った。推定モデル構築には Neural Network, Random Forest, Support Vector Machine の 3 つで Support Vector Machine の内部計算にカーネル法を使用した。またモデル構築前の処理において変数を直交変換する主成分分析の適用有無でパターン分けし、直交変換による推定モデルの識別性能について考察した。推定モデルに使用する特徴に滞在場所の特性を考慮したが、その時に予めメッシュの特徴を元に自動でクラスタ数を決定する xmeans を用いた。最後にパーソントリップ調査で構築した推定モデルの GPS データへの適用可能性を検討するため、転移学習を使用した。

表 8: 各パターン認識手法のまとめ

手法名	用途
Neural Network	推定モデル構築
Random Forest	推定モデル構築
Support Vector Machine	推定モデル構築
カーネル法	SVM の非線形分離
主成分分析	変数を独立にする
xmeans	クラスタリング
転移学習	PT によるモデルを GPS に転用する時の評価

また、表9は推定モデル構築に使用した3手法のまとめである。これらのモデルは単純な識別手法であるk-近傍法やロジスティック回帰、決定木と異なり、モデルが複雑になりやすい。そのため、識別性能は高くなるが、人間の可読性が落ちるという欠点がある。しかし、本研究は人口統計学的属性の推定可能性と識別性能に着目しているため、可読性より識別性能を優先した。本研究で使用した3手法はそれぞれ特徴がある。Neural Networkは単純なパーセプトロンを多層に積み上げたモデルであり中間層の数、初期値の設定によるパラメータ設定により、モデル全体が学習能力をもつ手法である。パーセプトロンの単純性と誤差逆伝播法により、荷重の収束が非常に早い一方で局所解で収束してしまうこともある。他にも入力層と中間層の接続関係、中間層と出力層の接続関係、入力層のノード配置といった設定があるが、これは検討数が爆発的に増加してしまい、探索的に最適な設計を行う事が難しい。

Random Forestはデータをブートストラップサンプリングにより多数の決定木を構築し、それらの多数決をもって出力値とする方法である。これは内部的に構築される決定木が非常に高速かつ、可読性の高いモデルであり、決定木の特徴を引き継ぐ側面がある。よって、分割指標による変数重要度の計算が行え、構築されたモデルの考察が行いやすい。しかし、サンプル数が少ない場合、サンプリングも少なくなるため、構築した多数の決定木による多数決が有効に作用しにくくなりモデルが不安定になりやすい。また、決定木の特徴を引き継ぐため、識別平面は変数に対し垂直に作成される。そのため変数間に線形性がある変数間付近に識別超平面が引かれる場合、非常に複雑な識別境界を描き過学習に陥りやすい。

Support Vector Machineはマージン最大化制約による識別平面の決定を、ラグランジュの未定乗数法により求める手法である。Support Vector Machine自体は線形分離のみ可能だが、カーネル法により使用するデータの特徴空間を別の空間に写像し、写像後の空間上で線形分離を行うことで結果的に非線形分離が行われる。この手法は非常に強力であるが、カーネルに使用する関数の選択により識別性能に大きな差がみられる。また、関数の設計は非常に柔軟であるため、データ集合に対し適切に設計しなければならない。

表 9: 推定モデル構築手法の特徴

手法名	チューニング	モデルの複雑性	大域解への収束	備考
Neural Network	難しい	非常に複雑	難しい	パラメータチューニングが非常に難しい
Random Forest	容易	複雑	中程度	変数間に線形関係があると過学習しやすい
Support Vector Machine	難しい	非常に複雑	必ず収束する	サンプル数 ≪ パラメータ数だと過学習しやすい

2.4 行動に関する特徴量抽出

行動に関する特徴抽出は時間, 距離, 回数について抽出した. 以下に抽出した特徴量をまとめる.

- 時間に関する特徴
 - 自宅出発時間 (時)
 - 自宅到着時間 (時)
 - 勤務地出発時間 (時)
 - 勤務地到着時間 (時)
- 距離に関する特徴
 - 1日に移動した距離 (m)
 - 24時間別移動距離 (m)
 - 自宅から勤務地までの距離 (m)
- 回数に関する特徴
 - 自宅訪問回数 (回/日)
 - 勤務地訪問回数 (回/日)

行動に関する特徴のうち, 時間に関する特徴はユーザーが何度も自宅や勤務地を訪問した場合, 複数の発着時間が生まれてしまう. そのため被験者は深夜に自宅に必ず帰ると仮定して以下のルールのもと, 発着時間を抽出した.

- 自宅出発時間: 一日の内, 最も早い自宅出発判定
- 自宅到着時間: 一日の内, 最も遅い自宅到着判定
- 勤務地出発時間: 一日の内, 最も遅い勤務地出発時間
- 勤務地到着時間: 一日の内, 最も早い勤務地到着時間

また, 距離に関する特徴の抽出には, 二点間の直線距離を使用した. しかし単純な三平方の定理では地球のような楕円体で近似できるような物体上の距離を求めると誤差が大きくなり生じてしまう. よって楕円体上の二点間距離を算出できるヒュベニの公式を用いて距離計算を行った. ヒュベニの公式は以下のとおりである. [56] また, 本研究で使用したデータの測地系は WGS84 であるため, 長半径, 短半径の値は上記測地系に準拠した値 ($a = 6,378,137.000, b = 6,356,752.314245$) [57] を用いた.

$$d = \sqrt{(d_y M)^2 + (d_x N \cos \mu_y)^2} \quad (7)$$

表 10: ヒュベニの公式に用いる値

変数	意味
d_x, d_y	経度差, 緯度差
μ_y	2点間の緯度の平均値
$M = a(1 - e^2)/W^3$	子午線曲率半径
$N = a/W$	卯酉線曲率半径
$W = \sqrt{1 - e^2 \sin^2 \mu_y}$	
$e = \sqrt{(a^2 - b^2)/a^2}$	第一離心率
a, b	長半径, 短半径

パーソントリップ調査を元に人口統計学的属性推定モデルを構築する時は、パーソントリップ調査の調査日が平日の1日であるためデータからそのまま特徴量を抽出した。しかしGPSデータの場合はデータ取得期間が複数日である。よってGPSデータの場合はデータ取得日を平日、休日に分け、上記特徴量の平均と分散をそれぞれ求め特徴量とした。よって行動に関する特徴量数はパーソントリップ調査を学習元データとした場合31個、GPSデータを学習元データとした時 $31 \times 2 \times 2 = 124$ 個となる。図32は特徴量抽出方法のフローチャートである。

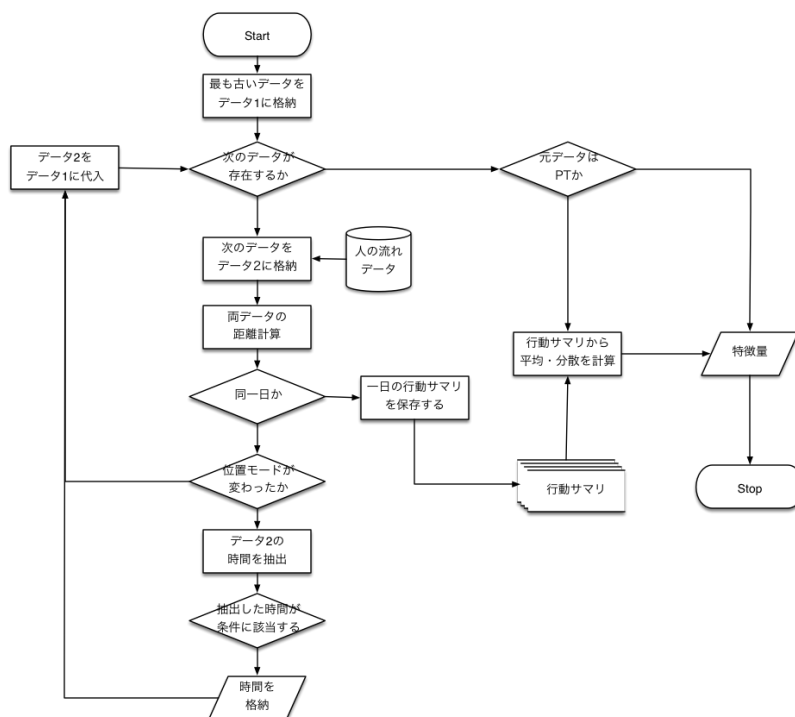


図 32: 行動に関する特徴抽出手順

2.5 滞在場所特徴に関する特徴量抽出

本研究で用いた場所特徴は全て地域メッシュで集計されたデータで、メッシュコードに対し複数の値が紐付けられている。しかしこのままの状態で使用した場合紐付けられた特徴量数×時間帯個の特徴量数になり推定モデルが非常に複雑になってしまう。そのため、地域メッシュを数〜数十グループへクラスタリングし、メッシュコードに対し一つのクラスタ番号が付与した。クラスタ数を n とすると、滞在場所の特徴を考慮するとき、滞在時間の考慮有無により 2 パターンに分類される。よって、本研究では滞在時間の 2 パターン試し、最も識別精度が高くなるパターンを採用した。図 34 は場所特徴データをモデルに組み込む方法の概要である。なおクラスタリングにはクラスタ数を自動で決定する手法の一つである、`xmeans` [53] を使用した。

mesh code	特徴1	特徴2	特徴3 … 特徴m
533935021	3	4	12 1
533936021	2	2	3 0
533935022	4	0	4 … 42
533934311	3	3	7 … 3
544035523	1	0	5 1
544020101	3	6	1 1

mesh code	クラスタ
533935021	1
533936021	1
533935022	2
533934311	4
544035523	4
544020101	7



図 33: 場所特徴の集約, クラスタリング

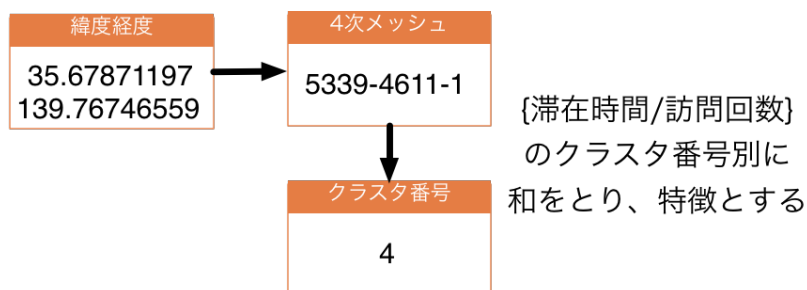


図 34: 場所特徴の抽出方法

本研究では 500m メッシュで集計された、業種別事業所数と男女別 5 歳階級別人口を元にクラスタリングを行った。図 35 は xmeans によるクラスタリング結果である。図より、各クラスタは表 11 のような意味付けができる。

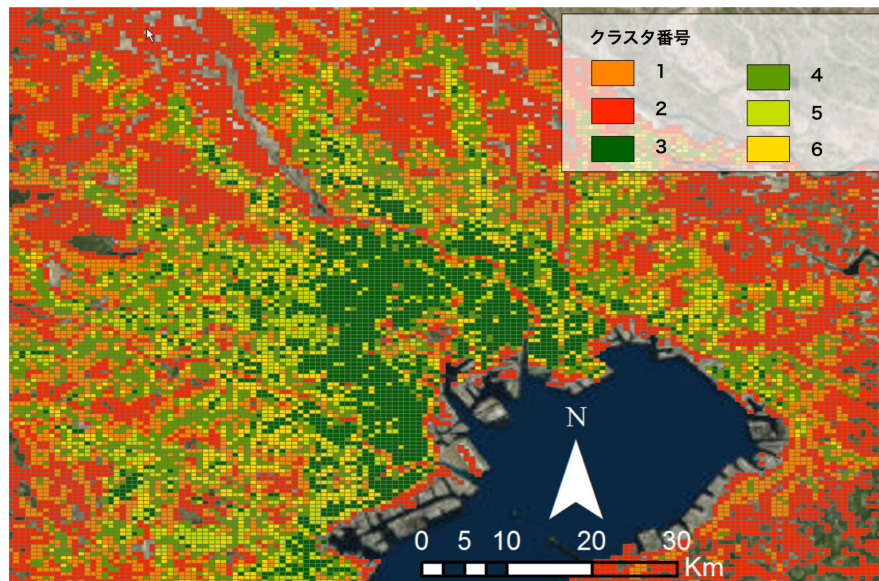


図 35: 場所特徴データの構築結果

表 11: 各クラスターの意味とクラスターサイズ

クラスター番号	クラスターサイズ	意味
1	5183	住宅地域
2	18320	居住者, 事業所共に少ない地域
3	1909	高密度集積地区
4	3310	都心近郊駅付近小規模店舗密集地域
5	1334	幹線道路隣接住宅地域
6	1047	幹線道路隣接商業地域

3 パーソントリップ調査を用いた人口統計学的属性の推定結果

使用したデータは2.1.1で取り上げた2008年東京都市圏パーソントリップ調査【空間配分版】である。表12, 表13, 表14はそれぞれ性別×年代, 年代×職業, 性別×職業のクロス集計表である。

表 12: 性別と年代のクロス集計表

年代 \ 性別	1	2	sum
2	715	709	1,424
3	744	685	1,429
4	664	655	1,319
5	756	910	1,666
6	991	1,388	2,379
7	1,401	1,791	3,192
8	1,686	2,190	3,876
9	1,542	2,026	3,568
10	1,351	1,537	2,888
11	1,140	1,323	2,463
12	1,284	1,511	2,795
13	1,309	1,512	2,821
14	1,127	1,387	2,514
15	964	1,142	2,106
16	729	980	1,709
17	530	670	1,200
18	274	664	938
sum	17,207	21,080	38,287

表 13: 年代と職業のクロス集計表

職業 \ 年代	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	sum
1	0	0	0	4	6	5	23	7	9	3	29	39	64	60	62	36	26	373
2	0	0	5	22	42	62	65	56	71	53	101	94	92	37	31	15	5	751
3	0	0	3	83	156	153	227	175	158	133	148	151	81	79	37	23	9	1,616
4	0	0	13	152	365	443	431	371	282	278	289	340	218	101	47	25	12	3,367
5	0	0	0	23	68	115	102	93	74	76	99	66	36	17	6	2	0	777
6	0	0	0	6	30	20	15	10	10	17	20	24	16	7	4	0	0	179
7	0	0	1	225	548	767	927	862	671	453	421	259	119	49	18	6	3	5,329
8	0	0	4	239	683	890	987	892	653	487	444	371	211	139	73	23	8	6,104
9	0	0	0	10	34	84	192	339	373	367	416	308	158	97	79	44	25	2,526
10	0	0	5	14	37	56	62	81	66	64	70	90	74	44	27	13	13	716
11	1,424	1,429	194	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3,047
12	0	0	740	2	1	0	0	0	0	0	0	1	0	1	0	0	0	745
13	0	0	327	772	102	33	15	9	4	2	4	3	2	3	0	0	0	1,276
14	0	0	1	14	129	390	621	518	378	374	486	482	534	412	323	196	78	4,936
15	0	0	19	79	131	128	132	103	94	110	195	486	774	922	882	718	701	5,474
16	0	0	2	1	2	5	6	3	1	3	1	1	4	1	2	1	4	37
91	0	0	3	12	30	27	46	32	29	27	46	50	61	39	28	12	5	447
99	0	0	2	8	15	14	25	17	15	16	26	56	70	98	90	86	49	587
sum	1,424	1,429	1,319	1,666	2,379	3,192	3,876	3,568	2,888	2,463	2,795	2,821	2,514	2,106	1,709	1,200	938	38,287

表 14: 性別と職業のクロス集計表

職業 \ 性別	1	2	sum
1	146	227	373
2	522	229	751
3	928	688	1,616
4	1,651	1,716	3,367
5	592	185	777
6	153	26	179
7	1,553	3,776	5,329
8	3,545	2,559	6,104
9	1,998	528	2,526
10	399	317	716
11	1,566	1,481	3,047
12	368	377	745
13	691	585	1,276
14	109	4,827	4,936
15	2,540	2,934	5,474
16	14	23	37
91	217	230	447
99	215	372	587
sum	17,207	21,080	38,287

また、東京都男女年齢 (5 歳階級別人口) の値と表 12 の値を相関検定した。表 15 は相関検定の結果である。この結果より、本データは母集団と非常に近い割合でサンプリングされていることがわかる。

表 15: 使用データと東京都五歳階級別人口の相関検定の結果

	correlation	p-value
男性	0.894	3.015×10^{-6}
女性	0.946	3.009×10^{-8}

3.1 特徴量の抽出

特徴量は 2.4 を利用した。表 16 は抽出した特徴量の基本統計量である。

表 16 を見ると、非常に標準偏差が大きく、一日の行動が人それぞれであるということがわかる。特に、一日の行動距離を示す `moveDist`、一時間あたりの行動を示す `m0` `m23` を見ると 500km 近く行動している人も見受けられる。これは新幹線や飛行機の高速移動手段を利用して移動したためと考えられる。

図 36 は各変数の相関をヒートマップにしたものである。相関が正に強い時赤く、負に強い時に緑に表示している。図を見ると、自宅や勤務地に関する特徴量間では比較的負の相関が強く、行動距離に関する特徴量では時間帯によって正の相関が見られる事がわかる。特に深夜帯はほとんどの人が行動をしていないため行動距離が非常に小さい時間が連続する人々が多く特に強い正の相関が見られる。

表 16: パーソントリップ調査から抽出した基本統計量

Statistic	N	Mean	St. Dev.	Min	Max
homeDpt	37,705	6.687	4.546	0	23
homeArv	37,705	12.894	7.902	0	23
WorkDpt	37,705	0.762	3.071	0	23
WorkArv	37,705	0.864	3.569	0	23
HomeCnt	37,705	0.856	0.619	0	6
WorkCnt	37,705	0.088	0.365	0	6
moveDist	37,705	3,786.386	3,809.443	0.000	144,654,500
m0	37,705	33.946	1,192.468	0.000	72,443.950
m1	37,705	5.267	394.334	0.000	43,985.150
m2	37,705	0.279	49.021	0.000	9,458.905
m3	37,705	4.970	185.583	0.000	19,001.630
m4	37,705	11.828	293.683	0.000	15,201.410
m5	37,705	46.544	610.450	0.000	19,828.130
m6	37,705	181.632	1,232.008	0.000	27,959.540
m7	37,705	1,071.625	3,025.487	0.000	53,848.300
m8	37,705	1,818.357	3,747.707	0.000	88,148.570
m9	37,705	893.881	2,978.627	0.000	122,349.100
m10	37,705	561.218	3,080.983	0.000	144,327.400
m11	37,705	518.480	3,056.841	0.000	169,634.300
m12	37,705	507.433	3,054.504	0.000	169,641.300
m13	37,705	542.912	3,224.532	0.000	169,641.900
m14	37,705	561.265	3,381.552	0.000	185,320.300
m15	37,705	660.842	3,578.909	0.000	185,316.200
m16	37,705	765.088	3,752.687	0.000	185,310.000
m17	37,705	1,004.538	3,944.625	0.000	185,304.300
m18	37,705	1,220.081	4,166.514	0.000	185,302.800
m19	37,705	885.139	3,807.522	0.000	145,171.700
m20	37,705	733.123	3,653.823	0.000	145,166.200
m21	37,705	600.671	3,604.124	0.000	145,170.100
m22	37,705	446.412	3,470.602	0.000	145,171.700
m23	37,705	267.576	3,234.801	0.000	140,649.500

	hDpt	hArv	wDpt	wArv	hCnt	wCnt	mDist	m0	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20	m21	m22	m23
hDpt	1	0.39	-0.6	-0.6	0.49	-0.6	-0.1	0.07	0.05	0.03	0.01	0.01	0	-0	-0	-0	-0	-0	-0	0.01	0.01	0.02	0.03	0.03	0.04	0.05	0.05	0.05	0.06	0.06	0.06
hArv	0.39	1	-0.6	-0.7	0.25	-0.6	0.16	-0.2	-0.1	-0.1	-0	-0	-0	-0.1	-0.1	-0.2	-0.2	-0.2	-0.2	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3
wDpt	-0.6	-0.6	1	0.81	-0.5	0.8	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
wArv	-0.6	-0.7	0.81	1	-0.5	0.75	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
hCnt	0.49	0.25	-0.5	-0.5	1	-0.5	-0.1	0.01	0.01	0.02	0.03	0.03	0.03	0.01	-0	0	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0	0	0.01	0.01	0.01
wCnt	-0.6	-0.6	0.8	0.75	-0.5	1	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
mDist	-0.1	0.16	-0	-0	-0.1	-0	1	0.05	0.02	0.01	0.01	0.01	0.01	0.06	0.09	0.1	0.1	0.11	0.11	0.12	0.12	0.12	0.13	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.1
m0	0.07	-0.2	-0	-0	0.01	-0	0.05	1	0.76	0.44	0.19	0.19	0.19	0.1	0.1	0.3	0.43	0.44	0.45	0.46	0.48	0.51	0.54	0.54	0.54	0.55	0.57	0.59	0.62	0.62	0.72
m1	0.05	-0.1	-0	-0	0.01	-0	0.02	0.76	1	0.71	0.35	0.35	0.34	0.16	0.03	0.11	0.17	0.21	0.24	0.24	0.26	0.29	0.28	0.29	0.28	0.29	0.3	0.31	0.34	0.33	0.39
m2	0.03	-0.1	-0	-0	0.02	-0	0.01	0.44	0.71	1	0.78	0.78	0.77	0.35	0.05	0.05	0.07	0.1	0.12	0.12	0.13	0.19	0.18	0.19	0.19	0.2	0.2	0.2	0.22	0.21	0.24
m3	0.01	-0	-0	-0	0.03	-0	0.01	0.19	0.35	0.78	1	1	0.99	0.46	0.05	0.03	0.02	0.01	0.01	0.01	0.02	0.1	0.1	0.1	0.1	0.12	0.12	0.12	0.12	0.1	0.11
m4	0.01	-0	-0	-0	0.03	-0	0.01	0.19	0.35	0.78	1	1	0.99	0.46	0.05	0.03	0.02	0.01	0.01	0.01	0.02	0.1	0.1	0.1	0.1	0.12	0.12	0.12	0.12	0.1	0.11
m5	0	-0	-0	-0	0.03	-0	0.01	0.19	0.34	0.77	0.99	0.99	1	0.56	0.12	0.07	0.04	0.04	0.03	0.03	0.04	0.12	0.12	0.12	0.14	0.14	0.14	0.14	0.13	0.11	0.12
m6	-0	-0.1	-0	-0	0.01	-0	0.06	0.1	0.16	0.35	0.46	0.46	0.56	1	0.67	0.38	0.23	0.21	0.2	0.19	0.19	0.22	0.21	0.22	0.22	0.22	0.23	0.22	0.18	0.17	0.18
m7	-0	-0.1	-0	-0	-0	-0	0.09	0.1	0.03	0.05	0.05	0.05	0.12	0.67	1	0.69	0.44	0.39	0.37	0.36	0.36	0.35	0.33	0.34	0.34	0.34	0.34	0.31	0.3	0.32	
m8	-0	-0.2	-0	-0	0	-0	0.1	0.3	0.11	0.05	0.03	0.03	0.07	0.38	0.69	1	0.85	0.77	0.72	0.7	0.69	0.67	0.64	0.65	0.65	0.65	0.66	0.65	0.64	0.63	0.61
m9	-0	-0.2	-0	-0	0.01	-0	0.1	0.43	0.17	0.07	0.02	0.02	0.04	0.23	0.44	0.85	1	0.97	0.91	0.88	0.86	0.84	0.8	0.81	0.82	0.81	0.82	0.82	0.81	0.81	0.79
m10	-0	-0.2	-0	-0	0.01	-0	0.11	0.44	0.21	0.1	0.01	0.01	0.04	0.21	0.39	0.77	0.97	1	0.96	0.93	0.91	0.89	0.85	0.86	0.86	0.86	0.85	0.84	0.84	0.8	
m11	0	-0.2	-0	-0	0.01	-0	0.11	0.45	0.24	0.12	0.01	0.01	0.03	0.2	0.37	0.72	0.91	0.96	1	0.99	0.97	0.95	0.9	0.91	0.89	0.88	0.85	0.83	0.83	0.82	0.79
m12	0.01	-0.3	-0	-0	0.01	-0	0.11	0.46	0.24	0.12	0.01	0.01	0.03	0.19	0.36	0.7	0.88	0.93	0.99	1	0.99	0.97	0.92	0.93	0.9	0.89	0.86	0.84	0.84	0.83	0.79
m13	0.01	-0.3	-0	-0	0.01	-0	0.12	0.48	0.26	0.13	0.02	0.02	0.04	0.19	0.36	0.69	0.86	0.91	0.97	0.99	1	0.99	0.94	0.94	0.91	0.9	0.87	0.86	0.85	0.84	0.81
m14	0.02	-0.3	-0	-0	0.01	-0	0.12	0.51	0.29	0.19	0.1	0.1	0.12	0.22	0.35	0.67	0.84	0.89	0.95	0.97	0.99	1	0.97	0.97	0.92	0.91	0.89	0.87	0.86	0.86	0.82
m15	0.03	-0.3	-0	-0	0.01	-0	0.12	0.54	0.28	0.18	0.1	0.1	0.12	0.21	0.33	0.64	0.8	0.85	0.9	0.92	0.94	0.97	1	0.98	0.9	0.89	0.87	0.85	0.85	0.84	0.81
m16	0.03	-0.3	-0	-0	0.01	-0	0.13	0.54	0.29	0.19	0.1	0.1	0.12	0.22	0.34	0.65	0.81	0.86	0.91	0.93	0.94	0.97	0.98	1	0.95	0.94	0.92	0.9	0.89	0.89	0.85
m17	0.04	-0.3	-0	-0	0	-0	0.12	0.54	0.28	0.19	0.1	0.1	0.12	0.22	0.34	0.65	0.82	0.86	0.89	0.9	0.91	0.92	0.9	0.95	1	0.99	0.97	0.96	0.95	0.94	0.9
m18	0.05	-0.3	-0	-0	0	-0	0.12	0.55	0.29	0.2	0.12	0.12	0.14	0.22	0.34	0.65	0.81	0.86	0.88	0.89	0.9	0.91	0.89	0.94	0.99	1	0.98	0.97	0.96	0.95	0.91
m19	0.05	-0.3	-0	-0	0	-0	0.12	0.57	0.3	0.2	0.12	0.12	0.14	0.23	0.34	0.66	0.82	0.86	0.85	0.86	0.87	0.89	0.87	0.92	0.97	0.98	1	0.99	0.98	0.97	0.94
m20	0.05	-0.3	-0	-0	0	-0	0.12	0.59	0.31	0.2	0.12	0.12	0.14	0.22	0.34	0.65	0.82	0.85	0.83	0.84	0.86	0.87	0.85	0.9	0.96	0.97	0.99	1	0.99	0.98	0.95
m21	0.06	-0.3	-0	-0	0.01	-0	0.11	0.62	0.34	0.22	0.12	0.12	0.13	0.18	0.31	0.64	0.81	0.84	0.83	0.84	0.85	0.86	0.85	0.89	0.95	0.96	0.98	0.99	1	0.99	0.96
m22	0.06	-0.3	-0	-0	0.01	-0	0.11	0.62	0.33	0.21	0.1	0.1	0.11	0.17	0.3	0.63	0.81	0.84	0.82	0.83	0.84	0.86	0.84	0.89	0.94	0.95	0.97	0.98	0.99	1	0.97
m23	0.06	-0.3	-0	-0	0.01	-0	0.1	0.72	0.39	0.24	0.11	0.11	0.12	0.18	0.32	0.61	0.79	0.8	0.79	0.81	0.82	0.81	0.85	0.9	0.91	0.94	0.95	0.96	0.97	1	0.97

図 36: 抽出した変数の相関

3.2 各属性推定手法の推定結果と精度

まず与えられた属性をそのまま使用し人口統計学的属性推定モデルを構築した。使用した方法はSVMでパラメータチューニングは行っていない。表 17 はその結果である。

表 17: 与えられた属性をそのまま使用し属性推定モデルを構築した結果

	gender	age	work type
accuracy	0.628	0.143	0.326
Recall	0.626	0.124	0.102
Precision	0.619	0.116	0.142
f-measure	0.618	0.067	0.104

表 17 より全ての人口統計学的属性において精度が非常に低い結果となった。理由として、一つの属性に含まれるクラスが多いため、非常に識別が難しくなっている事が原因になっていると考えられる。よって以降はクラスを集約

し、属性推定モデルを構築する。クラス集約はエラーテーブルを元に行った。表 18、表 19、表 20 は各属性のエラーテーブルである。

表 18: 性別推定の結果

推定値 \ 真値	1	2
1	2,082	1,289
2	2,021	3,505

表 19: 年代推定の結果

推定値 \ 真値	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2	249	212	58	7	14	22	57	55	39	37	62	51	53	35	22	16	9
3	51	69	30	13	14	15	25	18	24	26	17	23	11	9	6	0	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	2	2	9	9	5	0	4	1	0	1	0	0	0	0
7	1	4	3	21	29	42	41	21	28	8	12	11	10	4	1	1	1
8	70	90	252	293	463	620	691	657	498	381	419	328	241	168	98	49	20
9	5	11	8	8	12	29	29	26	12	20	19	19	13	4	8	3	0
10	0	0	0	0	0	1	1	0	1	0	0	2	0	0	0	0	0
11	0	0	0	2	0	4	3	2	1	3	3	6	1	0	1	0	0
12	0	0	1	1	2	1	2	3	4	1	0	1	2	2	0	0	0
13	3	3	1	2	15	20	19	14	13	16	22	16	10	20	11	3	1
14	3	4	14	42	49	87	116	117	108	93	104	145	173	153	146	88	27
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

表 20: 職業推定の結果

推定値 \ 真値	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	91	99
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
4	3	6	8	30	11	2	12	35	12	2	4	0	6	7	9	0	2	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	15	73	174	329	83	14	951	869	346	51	127	91	102	54	36	2	41	4
8	6	24	93	178	51	11	204	405	157	36	24	30	147	61	86	0	12	8
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	9	35	41	107	24	8	102	90	39	24	629	75	11	51	45	1	17	3
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	19	23	65	186	18	10	144	215	77	49	43	8	52	848	591	4	15	45
15	2	2	7	6	2	0	10	12	6	5	1	2	4	33	33	0	0	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

表 19、表 20 より、年代においては隣り合った年代において誤識別がされており、職業に関しては学生、主婦、サラリーマン、その他という大分類内の誤識別が多いことがわかる。よって年代のクラスを 5 歳階級から 10 歳階級に、職業については上記大分類を使用し属性推定モデルを構築する。図 37 は職業のクラス集約結果である。集約後のクロス集計表は付録を参照されたい。



図 37: 職業ラベルの属性集約

3.2.1 行動特徴のみ利用し属性推定した時の結果と精度

図 37 の通りにクラス集約し、再度推定モデルを構築した。表 21, 表 22, 表 23 は手法別の推定結果である。

表 21: 行動特徴のみ利用した時の性別推定モデルの結果

Method	Use PCA	Accuracy	Recall	Precision	F-measure
NN	yes	0.63	0.63	0.62	0.63
NN	no	0.61	0.59	0.53	0.46
RandomForest	yes	0.61	0.6	0.62	0.6
RandomForest	no	0.64	0.62	0.59	0.59
SVM	yes	0.62	0.62	0.61	0.6
SVM	no	0.59	0.5	0.5	0.37

表 22: 行動特徴のみ利用した時の年代推定モデルの結果

Method	Use PCA	Accuracy	Recall	Precision	F-measure
NN	yes	0.27	0.26	0.27	0.23
NN	no	0.6024	0.402	0.49	0.34
RandomForest	yes	0.26	0.27	0.26	0.26
RandomForest	no	0.65	0.51	0.45	0.46
SVM	yes	0.25	0.18	0.2	0.14
SVM	no	0.59	0.5	9.25	0.19

表 23: 行動特徴のみ利用した時の職業推定モデルの結果

Method	Use PCA	Accuracy	Recall	Precision	F-measure
NN	yes	0.69	0.54	0.5	0.49
NN	no	0.58	0.53	0.52	0.5
RandomForest	yes	0.69	0.56	0.49	0.51
RandomForest	no	0.65	0.58	0.52	0.53
SVM	yes	0.65	0.53	0.41	0.41
SVM	no	0.52	0.5	0.25	0.18

表 21, 表 22, 表 23 より, ほとんどの場合に置いて主成分分析を行っていない Random Forest が高い識別能力を示している. また, Recall と Precision

の値から、過学習も抑えられており、汎化されていると考えられる。理由として、図 36 の通り変数間の相関が非常に小さいため、Random Forest で生成される識別平面の複雑性が適切に保たれているためであると考えられる。表 24 はもっとも識別能力が高かった主成分分析を前処理で行ってない Random Forest の変数重要度の上位 5 つである。

表 24: 各推定モデルにおいて変数重要度が高かった上位 5 つの変数

	Gender's		Age's		Work type's	
変数名	$Imp(X_m)$	変数名	$Imp(X_m)$	変数名	$Imp(X_m)$	
1	moveDist	471.947	moveDist	628.908	homeArv	701.4229
2	m8	272.0467	homeArv	441.7423	moveDist	548.735
3	homeArv	263.669	m8	440.705	homeDpt	538.1984
4	m7	261.111	homeDpt	303.863	m8	457.144
5	homeDpt	243.2105	m18	299.6197	m15	290.0398

図 24 より、どの人口統計学的属性においても行動距離、8 時の移動距離、自宅の到着時間は性別、年齢、職業の違いに起因することがわかる。図 38 は行動距離の違いによる男性判定率の変化である。図より、1 日の行動距離が膨大な人や逆にほとんど行動していない人は男性の確率が高く、自宅付近の回遊行動を行う人は女性である確率が高い。また行動距離の違いによる主婦・主夫判定率の変化を表した図 39 より、自宅付近の回遊行動を行う女性は主夫や主夫であることがほとんどである。さらに、図 40 より 1 日の行動距離が 10km 程度の人々は 20 代であることが多い事が見て取れる。このようにパーソナリティ調査を用いて構築した属性推定モデルは、人口統計学的属性別の直感的な行動傾向の違いに則した精度の高いモデルであることがわかる。

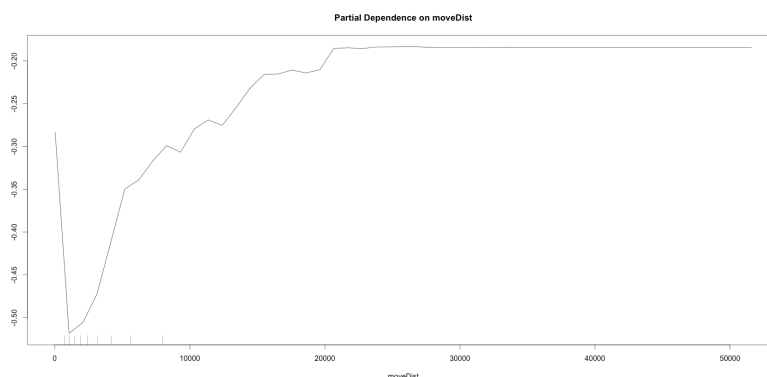


図 38: 行動距離の違いによる男性判定率の変化

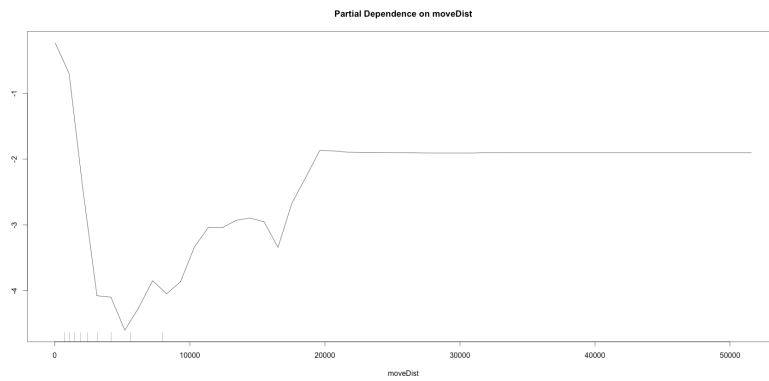


図 39: 行動距離の違いによる主婦判定率の変化

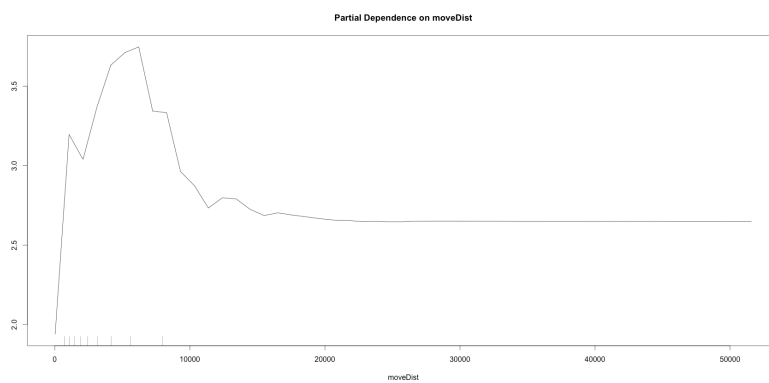


図 40: 行動距離の違いによる 20 代判定率の変化

3.2.2 滞在場所特徴を考慮し属性推定した時の結果と精度

次に滞在場所の特徴を考慮し属性推定した。表 25, 表 27, 表 29 は手法, 滞在場所特徴考慮パターン毎の推定結果である。

性別推定モデルの構築結果

手法別の性別推定モデルの構築結果をみると、行動特徴のみ使用して属性推定モデルを構築した時と同様に全体的に Random Forest を使用すると良い結果が得られる事がわかった。Neural Network を使用した時は収束しない場合もあったが、Random Forest に次ぐ識別性能を持っていることがわかる。一方で SVM については他の 2 手法と比較して識別性能は低い。

主成分分析を行った時とそうでない時では、行動特徴のみを利用して性別推定モデルを構築した時と比較し、主成分分析が有効に作用していることが考えられるが、抽出された特徴量をそのまま使用した時と比較し識別性能が大幅に向上したとはいえない。これは抽出された特徴量間の相関がそもそも低く、独立に近いことが考えられる。

表 25: 滞在場所を考慮した時の性別推定モデルの結果

Method	Use PCA	Consider stay time	Accuracy	Recall	Precision	F-measure
NN	Yes	Yes	収束せず			
NN	Yes	No	0.59	0.58	0.58	0.57
NN	No	Yes	収束せず			
NN	No	No	0.59	0.59	0.59	0.59
RandomForest	Yes	Yes	0.65	0.65	0.64	0.62
RandomForest	Yes	No	0.64	0.65	0.63	0.62
RandomForest	No	Yes	0.64	0.63	0.62	0.62
RandomForest	No	No	0.64	0.63	0.62	0.61
SVM	Yes	Yes	64	0.63	0.63	0.62
SVM	Yes	No	0.64	0.64	0.63	0.62
SVM	No	Yes	0.57	0.78	0.5	0.36
SVM	No	No	0.57	0.56	0.5	0.36

Random Forest は他の 2 手法と違い、変数重要度の計算が行えモデルの説明がし易い。また、表 25 より主成分分析なしの Random Forest は前処理で主成分分析を行った時の Random Forest と同程度の識別性能であることから、滞在時間による集計かつ前処理に主成分分析を使用していない Random Forest について、さらに掘り下げる。表 26 はこのモデルの変数重要度を計算し、上位 5 つを示した表である。

表 26: 性別推定モデルにおける変数重要度

	Variable	変数重要度
1	moveDist	3194.921
2	spVal	2941.177
3	hwDist	2695.569
4	m7	2050.485
5	c6	1696.148

これより行動特徴のみを利用して性別推定モデルを構築した時と同様に行動距離の違いに男女差が大きいことがわかる。また、滞留開始時間のばらつきや自宅勤務地の距離も男女差がある事がわかる。さらにクラスタ6の滞在時間の差に男女差があることがわかる。クラスタ6は図41より高速道路付近の商業地域であることがわかる。図42はクラスタ6の滞在時間の違いによる男性判定率の変化であるが、滞留が2000秒以上だと、高い割合で男性と判定されることが多い事がわかる。

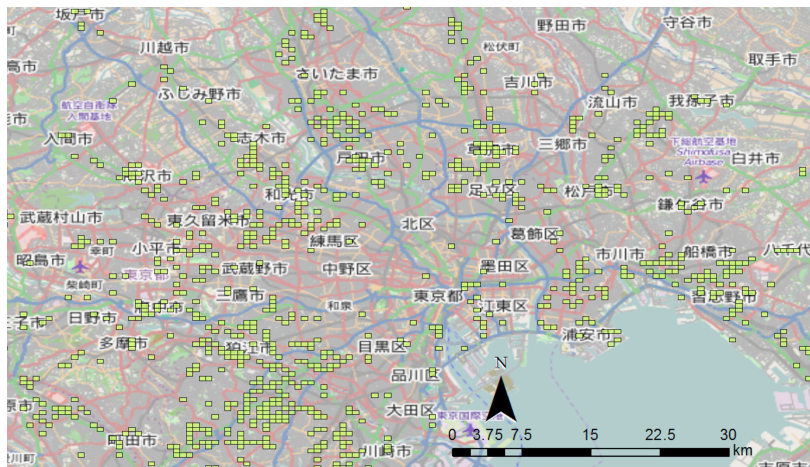


図 41: クラスタ番号6の地域

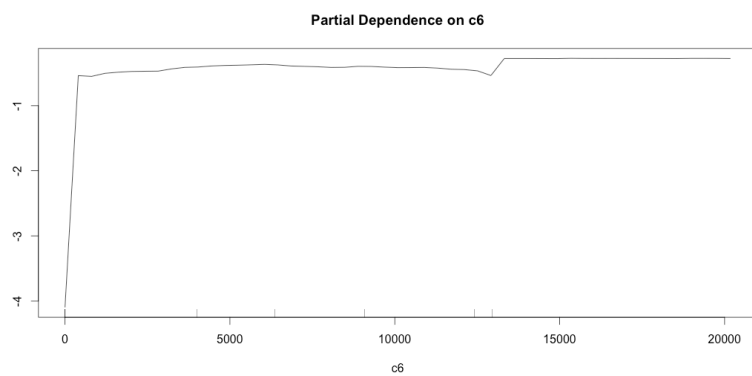


図 42: クラス 6 の滞在時間の違いによる男性判定率の変化

年代推定モデルの構築結果

滞在場所を考慮した年代推定モデルの構築は表 27 より、同一手法であっても前処理の方法が異なる事によって識別性能が大幅に変わる事がわかった。また、年代推定モデルの構築の場合も小さい差ではあるが性別推定モデルの構築時と同様に主成分分析を使用していなく、滞在時間で集計した Random Forest が最も高い識別性能を示している。このモデルの変数重要度を計算し、上位 5 つの変数と変数重要度を表 28 に示す。

表 27: 滞在場所を考慮した時の年代推定モデルの結果

Method	Use PCA	Consider stay time	Accuracy	Recall	Precision	F-measure
NN	Yes	Yes	収束せず			
NN	Yes	No	0.59	0.47	0.48	0.41
NN	No	Yes	収束せず			
NN	No	No	0.59	0.57	0.49	0.47
RandomForest	Yes	Yes	0.19	0.18	0.19	0.16
RandomForest	Yes	No	0.19	0.19	0.19	0.16
RandomForest	No	Yes	0.67	0.49	0.49	0.48
RandomForest	No	No	0.66	0.49	0.47	0.47
SVM	Yes	Yes	0.18	0.17	0.17	0.13
SVM	Yes	No	0.65	0.47	0.51	0.48
SVM	No	Yes	0.54	0.13	0.25	0.18
SVM	No	No	0.54	0.3	0.25	0.18

表 28: 年代推定モデルにおいて変数重要度の高かった変数と変数重要度

	特徴量	変数重要度
1	moveDist	4452.572
2	hwDist	4066.363
3	homeArv	3787.804
4	WorkDpt	3697.906
5	spVal	3626.717

表 28 より、行動距離や自宅勤務地の距離が年代の推定に重要な要素であることがわかる。また、性別推定モデルと違い、帰宅時間や退社時間も年代を推定するための重要な要素である。滞在に関する特徴では滞在開始時間の分散が識別において重要である。滞在場所の特性は変数重要度の上位には入らな

いものの、全変数のうち、中程度の重要度を示していた。また、変数重要度が高かった変数のとる値により、30代判定率の変化を計算した。図43、図44はそれぞれ帰宅時間、退社時間の違いによる30代判定率の変化である。これらの図より、帰宅時間や退社時間が早過ぎる、または遅すぎるときに30代と判定される割合が高いことがわかる。一方で10代に関しては図45より、帰宅時間が昼間から夕方である人々が10代と判定されることがわかる。このように検証した属性推定モデルはある程度の識別性能を持つ上に直感的にも納得できる結果となった。

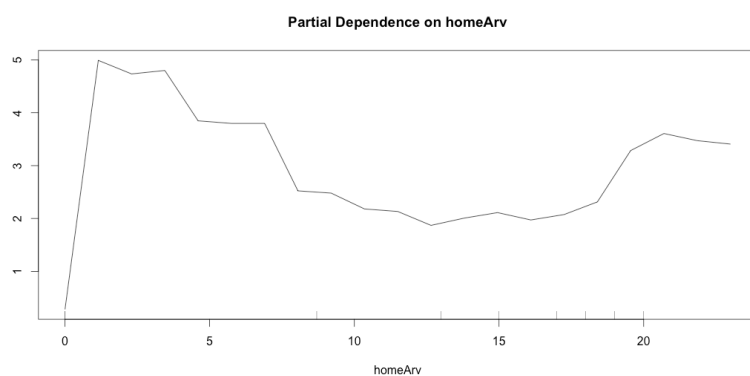


図 43: 帰宅時間の違いによる 30 代判定率の変化

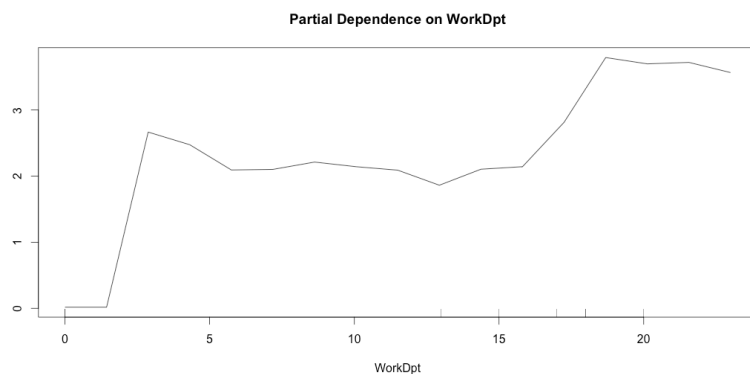


図 44: 退社時間の違いによる 30 代判定率の変化

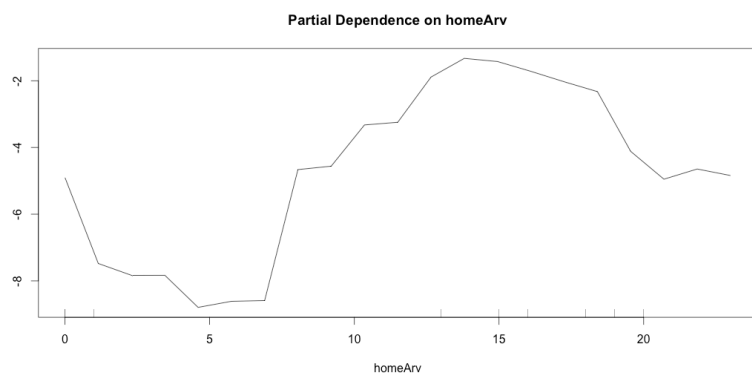


図 45: 帰宅時間の違いによる 10 代判定率の変化

職業推定モデルの構築結果

職業推定モデルについては他の属性と同じく Random Forest を用いたモデルが比較的高い識別性能をもつ一方、最も識別性能が高かったモデルは主成分分析を使用しなく、滞在場所特徴の抽出も滞在メッシュが属しているクラス別の滞在回数の場合の Neural Network という結果になった。これは Neural Network の特徴に起因すると考えられる。Random Forest は多数の決定木を構築し、構築した決定木の多数決でモデルを構築する手法であるため、決定木の特徴量のベクトルに対し垂直分割を繰り返すという特徴がある。そのため最適な識別超平面が複数の特徴ベクトルの線形和で表される場合、ジグザグな識別平面を構築するため、モデルの複雑性が向上してしまい、推定がうまくいかない場合がある。また SVM については特徴空間をカーネルトリックによる写像を行うため、特徴空間においては非線形の識別超平面であるものの、写像後の特徴空間においては線形の識別超平面が描かれる。そして SVM はラグランジュの未定乗数法により解を求めるため、ソフトマージンの許容コストを変化させても細かなチューニングをすることは難しい。一方で Neural Network は過学習に陥りやすいものの、元となるパーセプトロンが非常に単純であるため、過学習間近での細かな調整が可能である。今回の職業推定モデルの構築時には適切な Neural Network の構造をなしていたと考えられ、反復的に計算する誤差逆伝播法の収束が大域解付近で行われたことで良い識別性能が得られたと考えられる。

表 29: 滞在場所を考慮した時の職業推定モデルの結果

Method	Use PCA	Consider stay time	Accuracy	Recall	Precision	F-measure
NN	Yes	Yes	収束せず			
NN	Yes	No	0.51	0.42	0.4	0.38
NN	No	Yes	収束せず			
NN	No	No	0.57	0.57	0.67	0.61
RandomForest	Yes	Yes	0.41	0.37	0.2	0.19
RandomForest	Yes	No	0.41	0.39	0.2	0.19
RandomForest	No	Yes	0.66	0.6	0.59	0.59
RandomForest	No	No	0.65	0.6	0.59	0.58
SVM	Yes	Yes	0.38	0.18	0.16	0.14
SVM	Yes	No	0.62	0.59	0.57	0.58
SVM	No	Yes	0.44	0.11	0.25	0.15
SVM	No	No	0.44	0.3	0.25	0.15

識別性能は Neural Network に劣るものの、他の属性の場合と同様に主成分

分析を行わず、滞在時間を滞在場所の特徴指標に用いた Random Forest も良い識別性能をもっていることがわかる。あくまで参考ではあるが、このモデルの変数重要度を表 30 に示す。

表 30: 職業推定モデルにおいて変数重要度の高かった変数と変数重要度

	特徴量	変数重要度
1	spVal	6148.948
2	WorkDpt	5648.507
3	homeArv	5588.521
4	moveDist	4204.337
5	homeDpt	4137.308

表 30 より職業推定モデルにおいては滞在開始時間の分散が最も職業によって異なる事がわかった。また、退社時間や自宅の発着時間といった、時間に関する特徴量が変数重要度の高い結果となった。表 46 は退社時間の違いによるサラリーマン判定率の変化であるが、夕方 17 時にピークを迎え、20 時までに退社する人のサラリーマン判定率が高い事がわかる。試したモデルの中で 2 番目に高い識別性能だった Random Forest を使用したモデルの結果であるため参考ではあるが、構築されたモデルに関して、属性推定モデルは直感的に理解できる内容も含んでいることがわかる。

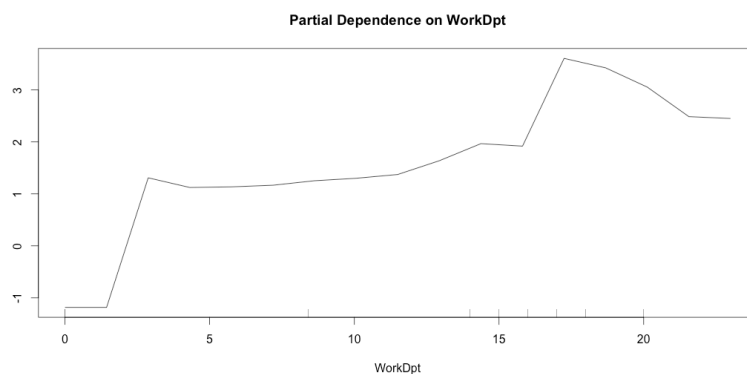


図 46: 退社時間の違いによるサラリーマン判定率の変化

3.3 パーソントリップ調査を元に構築した属性推定モデルのまとめ

表 25, 表 27, 表 29 より、年代、職業の属性推定に関しては場所の特徴を考慮したほうが良いモデルが構築されていることがわかる。一方で性別の属性推定

モデルの構築に関しては本研究で用いた場所の特徴データでは精度が向上しにくい事がわかった。さらに、行動特徴のみ考慮し属性推定モデルを構築した時と同様に主成分分析を前処理で行わない時の Random Forest が比較的安定して良いモデルを構築出来ている。また、滞在場所特徴を考慮する場合はメッシュに紐付いたクラスタ番号別に1日の総滞留時間を使用した場合に良い識別結果が得られることがわかる。これは滞留回数で集計した場合、滞在の特徴が弱まり、むしろノイズとして属性推定モデルに組み込まれてしまうためと考えられる。表 49, 表 50 は滞在場所に関する抽出した特徴量の基本統計量であるが、集計単位を滞留時間ではなく滞留回数にするような、各特徴量の値をなるべく小さくする方向に集計するほど人々に付与された属性別の特徴が消えてしまうと考えられる。また、各属性において Random Forest を使用したモデルの識別性能が非常に高い事から、識別性能の高かった Random Forest のモデルの変数重要度を計算したところ、直感的ではあるが、10代は帰宅が早く、30代は帰宅が遅いというような現実との乖離が少ない結果となった。

パーソントリップ調査を用い、使用した特徴量、手法の違いにより、各属性につき18個のモデルを構築した。表 31 は18個のモデルのうち、最も良い識別性能を持つモデルである。次節のパーソントリップ調査により構築された属性推定モデルをGPSデータに適用する時は表 31 の最良モデルを各属性に対し適用する。

表 31: 各属性において最も性能の良かった推定モデル

Attributes	Method	Use PCA	Use StayPoint/type	F-measure
性別	Random Forest	no	yes/time	0.61
年代	Random Forest	no	yes/time	0.48
職業	NN	no	yes/count	0.61

また、行動特徴のみ使用して属性推定モデルを構築したときと同様に、ほとんどの場合に置いて Random Forest を使用したモデルが最も良いか二番目、三番目に良いモデルとなっているため、各属性推定モデルにおける変数重要度を計算した。表 32 は変数重要度の上位5つのまとめである。

表 32: 変数重要度の高い変数と変数重要度

	Feature	Gender's $Imp(X_m)$	Feature	Age's $Imp(X_m)$	Feature	Work type's $Imp(X_m)$
1	moveDist	3194.921	moveDist	4452.572	spVal	6148.948
2	spVal	2941.177	hwDist	4066.363	WorkDpt	5648.507
3	hwDist	2695.569	homeArv	3787.804	homeArv	5588.521
4	m7	2050.485	WorkDpt	3697.906	moveDist	4204.337
5	c6	1696.148	spVal	3626.717	homeDpt	4137.308

4 GPSデータを用いた人口統計学的属性の推定結果

本研究で用いたGPSデータはKDDI研究所提供のGPSデータである。総被験者数184名のうち、滞留行動が確認できた158名のデータを使用して属性推定モデルを構築した。表33, 表34, 表35はそれぞれ性別×年代, 年代×職業, 職業×性別のクロス集計表である。なお, パーソントリップ調査でモデル構築を行った時と同様, 年代に関しては実年齢から, 10歳刻みに, 職業に関してはサラリーマン, 主婦, 学生, その他の4区分へ集約している。よって数字と属性の対応に関しては図37と同様である。

表 33: 性別と年代のクロス集計表

年代 \ 性別	1	2
4	3	8
5	11	14
6	10	15
7	17	17
8	13	9
9	12	9
10	4	7
11	2	2
12	3	1
13	1	0

表 34: 年代と職業のクロス集計表

年代 \ 職業	1	2	3	4
4	1	10	0	0
5	15	6	3	1
6	20	0	5	0
7	28	0	4	2
8	20	0	2	0
9	16	0	2	3
10	7	0	3	1
11	2	0	1	1
12	2	0	1	1
13	1	0	0	0

表 35: 職業と性別のクロス集計表

職業 \ 性別	1	2
1	65	47
2	7	9
3	1	20
4	3	6

4.1 特徴量の抽出

計測期間が平日の1日のみのパーソントリップ調査と異なり、GPS データは図 34 のうち、平日、休日の行動の違いや平日内の行動パターンの違い、休日内の行動パターンの違いといった、複数日にまたがった行動特徴の抽出を行う事ができる。よって、図 47 のようにパーソントリップ調査で抽出した自宅や勤務地の発着時間や訪問回数、行動距離といった特徴を平休日別の行動パターンを考慮して細分化することができる。よって、平休日別の平均特徴は取得した各日の特徴量の平均値、平休日別の行動パターンに関しては、取得した各日の特徴量の分散を使用した。そのため特徴量の数はパーソントリップ調査と比較すると4倍増えた。各特徴量の基本統計量は付録を参照されたい。

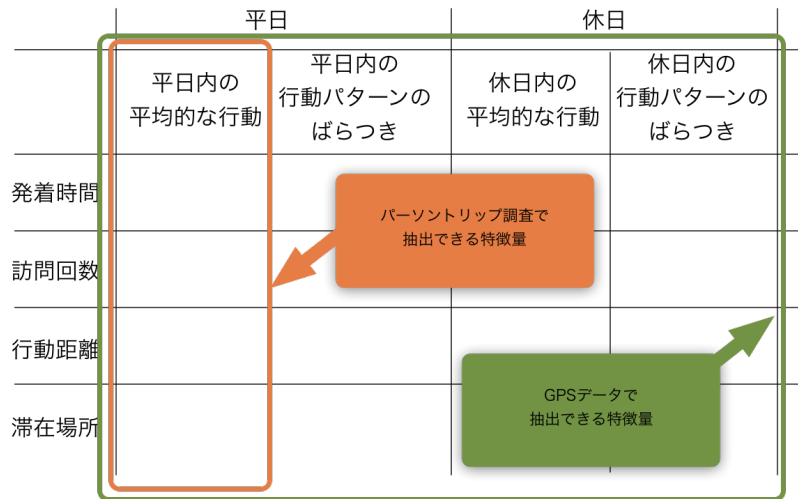


図 47: パーソントリップ調査と GPS データで取得できるデータの違い

4.2 各属性推定手法の推定結果と精度

パーソントリップ調査を用いて属性を推定した時と同様に行動特徴のみ使用した時、滞在場所の特徴を考慮した時の2パターンにおいてモデルを構築

していく。また、前節において構築したモデルを GPS に適用した時の精度についても考察する。

4.2.1 行動特徴のみ利用し属性推定した時の結果と精度

表 36, 表 38, 表 40 はそれぞれ性別, 年代, 職業について人口統計学的属性推定モデルを構築した時の結果である。パーソントリップ調査を用いた属性推定モデルの時と異なり全ての属性において Random Forest が良い手法とはいえ、各属性によって最も良い推定モデルで使われている手法が異なる。また、性別推定モデルに関してはパーソントリップ調査と比較し非常に少ないサンプル数で属性推定モデルを構築したにも関わらず、モデルの性能が良い事がわかる。

性別推定モデルの構築結果

表 36: 行動特徴のみ利用した時の性別推定モデルの結果

Method	Use PCA	Accuracy	Recall	Precision	F-measure
NN	yes	0.65	0.68	0.68	0.66
NN	no	0.5	0.52	0.52	0.5
RandomForest	yes	0.72	0.73	0.73	0.72
RandomForest	no	0.75	0.75	0.75	0.75
SVM	yes	0.65	0.66	0.66	0.65
SVM	no	0.68	0.7	0.7	0.69

これは性別推定モデルで最も性能が高かった主成分分析を前処理で行わない時の Random Forest の変数重要度 (表 37) の通り, 時間帯別の行動パターンは男女差が大きい事がわかり, 特に朝 7 時と夕方 17 時の行動パターンに男女差が大きい事がわかる. 図 48 より変数重要度の高い特徴量の値と, とる値による男性と推定される確率の関係をプロットすると休日の朝 7 時の行動パターンが画一的でなく, 行動距離の平均が大きいほど, 男性判定率が向上する. よって休日朝 7 時に自宅付近を習慣的に行動している人々は女性であることが高いと考えられる. また, 図 49 より, 平日の行動距離が多く, 勤務地の出入りが激しい人は男性である確率が高くなる. これはサラリーマンかつ, 営業担当者が頻繁に顧客企業へ訪問し, 会社へ帰る行動であると考えられる. そして, このような行動をする営業担当者には男性の従事者が多い事が考えられる.

表 37: 性別推定モデルにおける変数重要度の高い特徴量と変数重要度

	特徴量	変数重要度
1	avgM7Hday	3.158344
2	avgMoveDistWday	2.948406
3	avgM17Wday	2.477721
4	avgWorkCntWday	2.403706
5	varM7Hday	2.397615

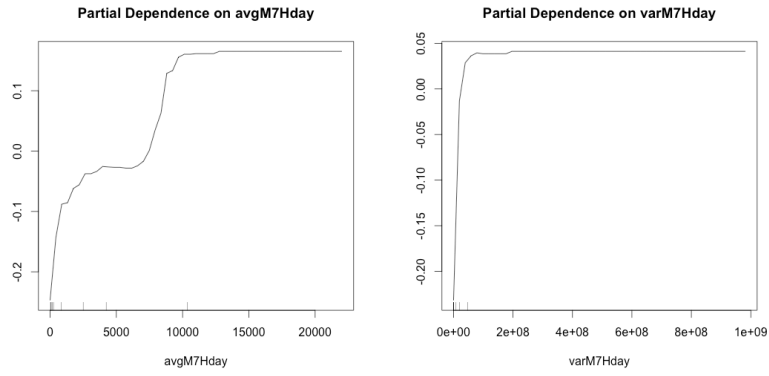


図 48: 朝 7 時の行動傾向の違いによる男性判定率の違い

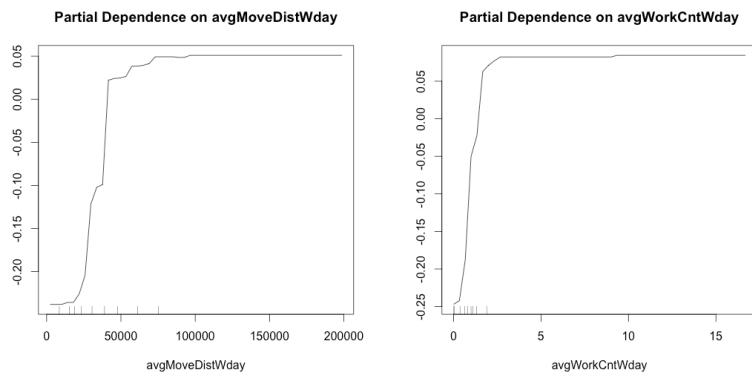


図 49: 行動距離や勤務地訪問回数の違いによる男性判定率の変化

年代推定モデルの構築結果

表 38 のとおり、年代推定に関しては非常に悪い結果となった。理由はパーソントリップ調査を用いた推定モデルの構築の時に精度向上のために行った属性集約を行っても、サンプル数の少ない GPS データにとってはまだまだクラス集約が必要であることが考えられる。職業の推定は主成分分析なしの Neural Network が非常に高い識別性能であることがわかる。これは Precision が 1 であることから、モデルの適合率が非常に高いためであると考えられる。しかし、データ数が少ない事による過学習も考えられる。

表 38: 行動特徴のみ利用した時の年代推定モデルの結果

Method	Use PCA	Accuracy	Recall	Precision	F-measure
NN	yes	0.31	0.21	0.29	0.24
NN	no	0.18	0.13	0.38	0.19
RandomForest	yes	0.19	0.07	0.13	0.07
RandomForest	no	0.25	0.1	0.14	0.11
SVM	yes	0.18	0.018	0.1	0.03
SVM	no	0.22	0.12	0.12	0.07

表 39: 年代推定モデルにおける変数重要度の高い特徴量と変数重要度

	特徴量	変数重要度
1	varSPHourWday	1.401767
2	varWorkDptHday	1.290073
3	varHomeCntWday	1.288909
4	avgM21Hday	1.153275
5	varM21Hday	1.150339

職業推定モデルの構築結果

表 40: 行動特徴のみ利用した時の職業推定モデルの結果

Method	Use PCA	Accuracy	Recall	Precision	F-measure
NN	yes	0.69	0.6	0.63	0.57
NN	no	0.78	0.78	1	0.87
RandomForest	yes	0.81	0.45	0.38	0.39
RandomForest	no	0.81	0.34	0.5	0.41
SVM	yes	0.78	0.19	25	0.21
SVM	no	0.87	0.43	0.49	0.46

職業推定モデルに関しては年代推定モデルの時と同様に Neural Network が他手法と比べて比較的 F-measure が高い結果になった。特に主成分分析を行わない Neural Network については Precision が 1 と非常に高い。しかし、サンプル数が少ないこともあり、検証データの数が非常に少なく、一人のユーザーの推定属性の正誤が F-measure に大きな影響を及ぼす。そのため、ランダムサンプリングされた学習データと検証データの組によって識別性能が大きく変わってしまう事が考えられる。しかし、パーソントリップ調査と比べサンプルサイズが非常に小さいにも関わらず、ある程度の識別性能を持っていることがわかる。

4.2.2 滞在場所特徴を考慮し属性推定した時の結果と精度

パーソントリップ調査から属性モデルを構築した時と同様に滞在場所の特徴を考慮して属性推定モデルを構築した。GPS データはパーソントリップ調査と違いデータ取得期間が 1 日でなく約 1ヶ月なので、行動パターンをモデルに組み込むために、滞在場所の特徴抽出は滞在時間の考慮有無の他に時間帯別と日別の 2 パターンを考慮し、計 4 パターンの試行を行った。表 41, 表 43, 表 44 はそれぞれ、性別、年代、職業の属性推定モデル構築結果である。

性別推定モデルの構築結果

表 41: 滞在場所を考慮した時の性別推定モデルの結果

Method	Use PCA	Consider stay time	Aggregate period	Accuracy	Recall	Precision	F-measure
NN	Yes	Yes	Hour	0.47	0.5	0.48	0.46
NN	Yes	Yes	Day	0.78	0.83	0.8	0.78
NN	Yes	No	Hour	0.72	9.72	0.72	0.72
NN	Yes	No	Day	0.53	0.52	0.53	0.52
NN	No	Yes	Hour	0.53	0.27	0.47	0.37
NN	No	Yes	Day	0.5	0.52	0.52	0.5
NN	No	No	Hour	0.59	0.59	1	0.74
NN	No	No	Day	NA	NA	NA	NA
Random Forest	Yes	Yes	Hour	0.69	0.68	0.7	0.68
Random Forest	Yes	Yes	Day	0.62	0.7	0.66	0.61
Random Forest	Yes	No	Hour	0.72	0.73	0.72	0.72
Random Forest	Yes	No	Day	0.59	0.59	0.59	0.59
Random Forest	No	Yes	Hour	0.53	0.54	0.54	0.53
Random Forest	No	Yes	Day	0.72	0.72	0.7	0.7
Random Forest	No	No	Hour	0.5	0.5	0.51	0.5
Random Forest	No	No	Day	0.69	0.69	0.67	0.68
SVM	Yes	Yes	Hour	0.34	0.18	0.5	0.26
SVM	Yes	Yes	Day	0.44	0.22	0.5	0.3
SVM	Yes	No	Hour	0.53	0.26	0.5	0.35
SVM	Yes	No	Day	0.5	0.26	0.47	0.33
SVM	No	Yes	Hour	0.56	0.28	0.5	0.36
SVM	No	Yes	Day	0.59	0.58	0.58	0.58
SVM	No	No	Hour	0.4	0.2	0.5	0.28
SVM	No	No	Day	0.62	0.61	0.62	0.62

この結果より, Random Forest が非常に高い値を示した. 特に訪問地回数を 1 時間単位で集計し, 主成分分析で前処理を行うとき, 識別性能が最も高い結果になった. この識別性能はサンプルサイズが非常に大きいパーソントリップ調査から構築した性別推定モデルよりも良い性能である. また, Neural Network については収束しなかったケースも見られたため, 収束させるよう

な初期解の調整を別途行う必要があると考えられる。表 42 は識別性能が最も高かったモデルの変数重要度である。

表 42: 性別推定モデルにおける変数重要度

主成分名	変数重要度	主成分の特徴
PC2	3.001211	都心近郊駅付近の訪問特徴
PC3	2.369962	行動距離と都心訪問数
PC1	1.337233	住宅地における深夜行動特徴
PC37	1.238659	ベッドタウン地域の早朝行動特徴
PC9	1.17786	勤務地訪問回数と平日昼間の移動特徴

表 42 より寄与率の高い第二主成分や第三主成分が性別の識別にとって重要な一方で、第 37 主成分といった寄与率の低い主成分も識別に重要な変数であると考えられる。

これより、吉祥寺や柏、船橋などの都心近郊駅付近への訪問特徴や、行動距離の差、ベッドタウン地域の行動特徴が性別に対し有効に作用していると考えられる。主成分と元の特徴量の関係については付録を参照されたい。また、図 50 は第二主成分のとり値による男性判定率の違いをプロットしたグラフである。表 42 より、都心近郊駅付近の訪問数が多いほど、男性判定率が向上すると考えられる。

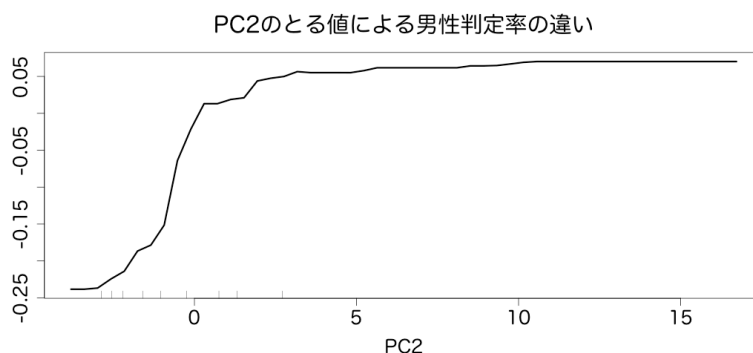


図 50: PC2 のとり値の違いによる、男性判定率の違い

年代推定モデルの構築結果

表 43 は滞在場所を考慮した時の年代推定モデルを構築した結果である。表より、ほとんどの場合において、推定が難しかった事がわかる。これは年代推定モデルではサンプル数が少ないが、パラメータ数、クラス数がサンプル数に比較して多い。モデルを構築しているデータは全体データ集合からランダム

に半数のデータを使用し構築しているため、約 80 人の行動特徴を 224 のパラメータに変換し 8 クラスの年代推定モデルを構築している。そのため、サンプル数 \ll パラメータ数であり、SVM で内部的に行われるカーネルトリックにより、写像空間での過学習が行われてしまう。また、Random Forest ではブートストラップサンプリングを行うため、約 1/3 のデータは使用されない。そのため、モデル構築に用いられるサンプル数が更に少なくなり、識別性能が低い結果となった。一方で Neural Network については他の 2 手法と比較すると単純なモデルである。そのため、サンプル数が少ない場合でも誤差逆伝播法による推定モデルの逐次更新が進み一定の識別性能を保ったと考えられる。

表 43: 滞在場所を考慮した時の年代推定モデルの結果

Method	Use PCA	Consider stay time	Aggregate period	Accuracy	Recall	Precision	F-measure
NN	Yes	Yes	Hour	0.16	0.16	0.5	0.24
NN	Yes	Yes	Day	0.09	0.12	0.18	0.15
NN	Yes	No	Hour	0.22	0.19	0.36	0.24
NN	Yes	No	Day	0.09	0.07	0.25	0.12
NN	No	Yes	Hour	0.06	0.06	1	0.11
NN	No	Yes	Day	0.19	0.27	0.42	0.29
NN	No	No	Hour	0.03	0.03	1	0.06
NN	No	No	Day	NA	NA	NA	NA
Random Forest	Yes	Yes	Hour	0.06	0.009	0.05	0.01
Random Forest	Yes	Yes	Day	0.21	0.04	0.12	0.06
Random Forest	Yes	No	Hour	0.12	0.02	0.08	0.03
Random Forest	Yes	No	Day	0.31	0.17	0.16	0.15
Random Forest	No	Yes	Hour	0.22	0.06	0.11	0.07
Random Forest	No	Yes	Day	0.21	0.06	0.08	0.07
Random Forest	No	No	Hour	0.19	0.04	0.13	0.07
Random Forest	No	No	Day	0.09	0.01	0.1	0.03
SVM	Yes	Yes	Hour	0.12	0.012	0.1	0.022
SVM	Yes	Yes	Day	0.19	0.02	0.1	0.03
SVM	Yes	No	Hour	0.19	0.02	0.1	0.03
SVM	Yes	No	Day	0.19	0.02	0.1	0.03
SVM	No	Yes	Hour	0.28	0.02	0.1	0.04
SVM	No	Yes	Day	0.37	0.2	0.17	0.16
SVM	No	No	Hour	0.16	0.01	0.1	0.03
SVM	No	No	Day	0.25	0.04	0.11	0.06

職業推定モデルの構築結果

表 44 は滞在場所を考慮した時の職業推定モデルの構築結果である。職業推定モデルにおいても年代推定モデルと同様に Neural Network が最も高い識別能力を持っていた。理由は年代推定モデルの時と同様に、SVM はサンプル数 ≪ パラメータ数であるため、モデルの構築が難しく、Random Forest では

ブートストラップサンプリングにより構築される各決定木に含まれるサンプルサイズがさらに小さくなる事である。また, Neural Network の中でも前処理に主成分分析を行わない方が推定モデルの識別性能が高かった。これは主成分分析を行うことで変数 (主成分) の重要度に差が生まれるため, 各変数に乘じられる荷重の更新量が変数によって異なる事が考えられる。パーソントリップ調査を元に構築した主成分分析適用後の Neural Network では多量のサンプル数があったため, 変数に乘じられる荷重の更新量に差があった場合でも全体的に荷重の更新がうまく行われていたが, GPS データではサンプル数が少ないため, 寄与率が低い主成分に関してはほとんど荷重の更新がされなかったと考えられる。

表 44: 滞在場所を考慮した時の職業推定モデルの結果

Method	Use PCA	Consider stay time	Aggregate period	Accuracy	Recall	Precision	F-measure
NN	Yes	Yes	Hour	0.6	0.44	0.55	0.48
NN	Yes	Yes	Day	0.59	0.31	0.35	0.32
NN	Yes	No	Hour	0.68	0.46	0.26	0.31
NN	Yes	No	Day	0.65	0.53	0.41	0.46
NN	No	Yes	Hour	0.81	0.81	1	0.89
NN	No	Yes	Day	0.75	0.75	1	0.85
NN	No	No	Hour	0.75	0.75	1	0.85
NN	No	No	Day	NA	NA	NA	NA
Random Forest	Yes	Yes	Hour	0.66	0.16	0.25	0.2
Random Forest	Yes	Yes	Day	0.71	0.17	0.25	0.21
Random Forest	Yes	No	Hour	0.81	0.2	0.25	0.22
Random Forest	Yes	No	Day	0.71	0.34	0.31	0.3
Random Forest	No	Yes	Hour	0.84	0.34	0.49	0.39
Random Forest	No	Yes	Day	0.84	0.72	0.47	0.54
Random Forest	No	No	Hour	0.81	0.38	0.34	0.37
Random Forest	No	No	Day	0.72	0.31	0.35	0.33
SVM	Yes	Yes	Hour	0.66	0.16	0.25	0.2
SVM	Yes	Yes	Day	0.72	0.18	0.25	0.21
SVM	Yes	No	Hour	0.81	0.2	0.25	0.22
SVM	Yes	No	Day	0.71	0.18	0.25	0.21
SVM	No	Yes	Hour	0.81	0.2	0.25	0.22
SVM	No	Yes	Day	0.78	0.44	0.3	0.3
SVM	No	No	Hour	0.75	0.18	0.25	0.21
SVM	No	No	Day	0.68	0.42	0.28	0.26

4.2.3 パーソントリップ調査から構築したモデルを GPS データに適用した時の結果と精度

3 節で最も良かった分類手法を用いて, GPS データへの適用を行った. 表 45 は適用後の結果である.

表 45: パーソントリップ調査を元に構築した属性推定モデルを GPS データに適用した結果

accuracy	Recall	Precision	f-measure
0.532	0.475	0.250	0.176
0.587	0.299	0.115	0.250
0.545	0.075	0.704	0.207

パーソントリップ調査から構築されたモデルを GPS データに適用したため、パーソントリップ調査からは取得できない平休日別の行動パターンは適用できなかった。そのため、識別性能は非常に低い結果となった。しかし、Accuracy を見ると年代推定モデルに関しては GPS データから構築した年代推定モデルのうち最も識別性能が良いモデルとほぼ同一の識別性能であり、モデルの転用がうまく行われている事がわかる。一方で、性別推定モデルに関しては GPS データ単体で構築したモデルのほうが識別性能が良いため、両データを組み合わせた年代推定モデルを構築する必要があると考えられる。

4.3 GPS データを用いた属性推定モデルの構築のまとめ

GPS データはパーソントリップ調査と異なりサンプルサイズが非常に小さく、各属性に付与されているクラスもサンプルサイズと比較すると大きいため、全体的に推定モデルが安定しない結果になった。しかし、性別推定モデルに関してはパーソントリップ調査から構築したモデルを凌駕する識別性能であり、サンプルサイズが小さいにも関わらず有用な推定モデルが構築できたと考えられる。年代推定モデルに関してはパーソントリップ調査で構築したモデルを GPS データに適用した場合でも、識別性能が変わらないことから、転移学習が有効に働いていると考えられる。しかし、今回はデータを組み合わせて属性推定モデルを構築していないため、両データを組み合わせて年代推定モデルを構築することでさらに良い識別性能のモデルを作成することができると考えられる。職業の推定モデルに関しては Neural Network を用いた場合、識別性能が高いモデルを構築することができた。しかし、サンプルサイズが小さいため、過学習の可能性を捨て去ることは難しい。パーソントリップ調査を元に構築した年代推定モデルは Recall が低い結果となった。よって、Recall の改善を行うことで、識別性能が非常に高くなると考えられる。

表 46: GPS データから構築された属性推定モデルのうち、最もよい識別性能だった手法

Attributes	Method	Use PCA	Consider Stay	Aggregate Period	F-measure
	Random				
性別	Forest	Yes	No	Hour	0.8
年代	NN	No	Yes	Day	0.29
職業	NN	No	Yes	Hour	0.89

4.3.1 本研究で構築した属性推定モデルの構築のまとめ

GPS データはパーソントリップ調査と異なりサンプルサイズが非常に小さく、各属性に付与されているクラスもサンプルサイズと比較すると大きいいため、全体的に推定モデルが安定しない結果になった。しかし、性別推定モデルに関してはパーソントリップ調査から構築したモデルを凌駕する識別性能であり、サンプルサイズが小さいにも関わらず有用な推定モデルが構築できたと考えられる。年代推定モデルに関してはパーソントリップ調査で構築したモデルを GPS データに適用した場合でも、識別性能が変わらないことから、転移学習が有効に働いていると考えられる。しかし、今回はデータを組み合わせる属性推定モデルを構築していないため、両データを組み合わせる年代推定モデルを構築することでさらに良い識別性能のモデルを作成することができると考えられる。職業の推定モデルに関しては Neural Network を用いた場合、識別性能が高いモデルを構築することができた。しかし、サンプルサイズが小さいため、過学習の可能性を捨て去ることは難しい。パーソントリップ調査を元に構築した年代推定モデルは Recall が低い結果となった。よって、Recall の改善を行うことで、識別性能が非常に高くなると考えられる。また、表 47 は本研究で構築した全ての属性推定モデルのうち、最も識別性能が高いモデルである。表より、主成分分析を前処理で使わずに属性推定モデルを構築するとほとんどの場合において良い属性推定モデルが構築できることがわかる。また、滞留地点の特徴を考慮した推定モデルを構築する時には滞留時間を使用すると良い場合が多いが、全ての場合において滞留時間を特徴量に用いるのが有効ではないため、推定モデルを構築するときには両方の検討が必要である。

パーソントリップ調査を元に構築した属性推定モデルは平日 1 日のみの取得であるため、抽出できる特徴量が少ないがサンプルサイズが大きいいため、ある程度の識別性能をもった推定モデルを構築できた。しかし、GPS データから取得できる特徴量はパーソントリップ調査から取得できる特徴量の約 4 倍なため、性別・職業推定モデルに関してはサンプルサイズが少ないにも関わ

らず、パーソントリップ調査から構築したモデルより良好な識別性能のモデルが構築できた。年代推定モデルに関しては良好な識別性能をもった推定モデルを構築することができなかったが、パーソントリップ調査を元に構築した推定モデルを GPS データに適用した結果、GPS データ単体から構築したモデルと同程度の識別性能であることがわかった。そのためパーソントリップ調査と GPS データの 2 つのデータを組み合わせて属性推定モデルを構築することで、2 つのデータの特徴を組み合わせた良好なモデルが構築できると考えられる。

表 47: 行動履歴情報別の最も良い属性推定モデル

Attribute	Period of Acquire Data	Method	Use PCA	Aggregate dimension/type	F-measure
性別	1 日	RandomForest	No	time	0.61
性別	複数日	RandomForest	Yes	count/hour	0.8
年代	1 日	RandomForest	No	time	0.48
年代	複数日	NN	No	time/day	0.29
職業	1 日	NN	No	count	0.61
職業	複数日	NN	No	time/hour	0.89

5 結論

5.1 本研究の成果

本研究は人の行動に着目して、人の行動履歴を元にその人の性別、年代、職業といった人口統計学的属性の推定モデルの構築を試みた。モデルの構築のために本研究ではパーソントリップ調査と GPS データといった 2 つの異なるデータを用いた他、推定モデル構築手法の違いで 6 パターン、滞在場所の特徴を考慮するために最大 4 パターンで計 48 個のモデルを構築した。さらにパーソントリップ調査で構築した最も識別性能の良いモデルを GPS データに適用し検証したため、計 51 個の検証を行った。

その結果、性別、職業の 2 つのモデルに関しては識別性能の高い属性推定モデルが構築でき、属性の違いが行動距離や特定の時間の行動パターンに影響を与えている事がわかった。一方で年代推定モデルに関しては本研究で属性推定モデルを構築するために用いた特徴量や場所の特徴データではうまく分類できない事がわかった。

また、パーソントリップ調査から構築したモデルを GPS に適用した結果、年代推定モデルに関しては類似した異なるデータで属性推定モデルを構築し

た場合でも識別が行える事がわかった。

5.2 本研究の課題と展望

本研究ではパーソントリップ調査でモデルを構築し構築したモデルの評価を行い、GPS データでモデルを構築し、構築したモデルの評価を行った。また、パーソントリップ調査で構築したモデルを GPS データへ適用する事も行った。

図 48 の通り、パーソントリップ調査はサンプルが多い一方で抽出できる特徴量は少なく、GPS データはサンプルが少ないが抽出できる特徴量が多い。また、本研究で使用しなかった GPS データは属性情報が付与されていないものの、サンプル数が約 60 万と非常に大きいデータである。よって、本研究で提示した課題のうち、本稿で行ったのは図 51 の通りであり、今回見つかった GPS データのサンプル数が少ない事による推定モデルの安定性が低い問題を解決することが必要になる。

この課題の解決のために、パーソントリップ調査で構築したモデルを本研究で用いた GPS データに適用し、モデルを更新する。さらに図 52 のように GPS データから抽出できる特徴量をパーソントリップ調査で構築したモデルから出力される値と GPS データからのみ抽出される特徴量に分解し、属性付き GPS データ単体でモデルを再構築する。そして再構築したモデルをラベルなし GPS データに適用しながら、モデルを更新していく。これら操作を行うことにより、パーソントリップ調査による頑健だが簡易なモデルの値を初期値としてラベルがある少量の GPS データで属性推定モデルを構築し、モデルの更新が取束したらラベルなしの GPS データを使用してさらに頑健なモデルを構築できると考えられる。

表 48: 人の行動に関するデータ

データ名	人数	ラベル
PT 調査	約 60 万人	あり
GPS データ	159 人	あり
ラベルなし GPS データ	約 60 万人	なし



図 51: 本稿で取り上げたテーマに関するロードマップ

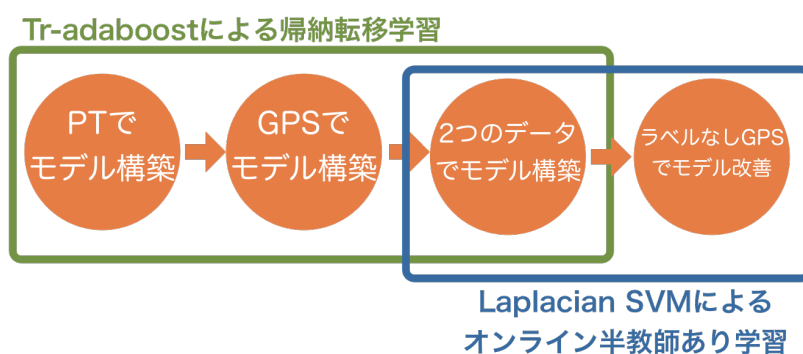


図 52: 今後の展望と参考手法

また、場所の特徴モデルについても検討する事がある。本研究では場所の特徴を属性推定モデルに含めるために、事前に場所に付与されている様々な特徴を元にクラスタリングし、その結果を用いて場所特徴考慮済みモデルを構築した。しかし、事前にクラスタリングを行っているため、もともと任意の場所に付与されている特徴量数を m とすると、 m 次元のデータを 1 次元に圧縮していることになる。また、クラスタリングについても簡易かつ恣意的なクラスタリング結果になりにくい $xmeans$ を用いているが、クラスタ数を自動決定するアルゴリズムは他にも複数ある上に、 $xmeans$ の元である $kmeans$ はデータを超球状にクラスタリングする手法である。よって、クラスタ形状に関する仮定を強く置いている上に隣接した場所の特徴を考慮していない。また事業所・企業統計調査と国勢調査の情報のみしか使用していない。この課題の解決として、クラスタリング手法のさらなる検討や、場所情報の使い分けによる属性推定モデルの精度の違い、Convolutional Neural Network を用いた特徴抽出を行う事が考えられる。

最後に、本研究では東京都市圏を対象に行なった。今後は別都市を対象に人口統計学的属性推定モデルの構築を行い、都市間の変数重要性の差を考察すると、都市の違いを人の行動を元に表せると考えられる。

参考文献

- [1] 総務省. 緊急通報の機能.
http://www.soumu.go.jp/menu_seisaku/ictseisaku/net_anzen/hijyo/tuho.html.
- [2] ゼンリンデータコム. 混雑統計.
<http://www.zenrin-datacom.net/business/campaign/s01.html>.
- [3] ナイトレイ株式会社. ナイトレイ GIS メッシュデータ.
<http://tool.nightley.jp/gis-demo/> 2013.
- [4] 観光庁. GPS を利用した観光行動の調査分析.
<https://www.mlit.go.jp/kankocho/shisaku/kankochi/gps.html>.
- [5] 光悦長尾, 秀憲川村, 雅人山本, 東大内. GPS ログからの周遊型観光行動情報の抽出 (<特集> 「ネットワークデータマイニング」「センサデータマイニング」). 電子情報通信学会技術研究報告. AI, 人工知能と知識処理, Vol. 105, No. 224, pp. 23-28, 2005-07.
- [6] 山本泰裕, 伊藤弘, 小野良平, 下村彰男. GPS を用いた新宿御苑における利用者の行動パターンに関する研究. ランドスケープ研究, Vol. 69, No. 5, pp. 601-604, 2006.
- [7] 松尾豊, 岡崎直観, 中村嘉志, 西村拓一, 橋田浩一, 中島秀之. 位置履歴からのユーザ属性の推定. 情報処理学会論文誌 (投稿中), 2006.
- [8] 東京都都市圏交通計画協議会. パーソントリップ調査検討事例.
<http://www.tokyo-pt.jp/person/03.html#04> 2014.
- [9] 国土交通省. PT 調査とは.
<http://www.mlit.go.jp/crd/tosiko.pt.html>.
- [10] 姫路市. 播磨都市圏パーソントリップ調査 (調査概要).
http://www.city.himeji.lg.jp/s70/2212533/_16711/_16715.html 2006.
- [11] 大佛俊泰, 島田廉. 平日と休日における都市内滞留者の時空間分布推定と地震被害想定への応用. 日本建築学会計画系論文集, Vol. 74, No. 635, pp. 145-152, 2009.
- [12] 斎藤参郎, 梶井昌邦, 中嶋貴昭. 都心商業空間における商業施設への消費者来街者数と回遊パタンの同時推定逆問題について. 地域学研究, Vol. 30, No. 1, pp. 213-229, 2000.

- [13] 準天頂衛星システムサービス株式会社. 準天頂衛星システムサービス.
<http://www.qzs.jp/index.html>.
- [14] 金杉洋, 黒川茂莉, 村松茂樹, 関本義秀. 携帯電話基地局通信情報の行動分析への適用可能性把握. 2012.
- [15] Daniel Ashbrook and Thad Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, Vol. 7, No. 5, pp. 275–286, 2003.
- [16] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from GPS trajectories. pp. 791–800. ACM, 2009.
- [17] Lin Liao, Dieter Fox, and Henry Kautz. Extracting places and activities from gps traces using hierarchical conditional random fields. *The International Journal of Robotics Research*, Vol. 26, No. 1, pp. 119–134, 2007.
- [18] Shan Jiang, Joseph Ferreira Jr, and Marta C. Gonzalez. Discovering urban spatial-temporal structure from human activity patterns. pp. 95–102. ACM, 2012.
- [19] Yang Ye, Yu Zheng, Yukun Chen, Jianhua Feng, and Xing Xie. Mining individual life pattern based on location history. pp. 1–10. IEEE, 2009.
- [20] Nadine Schssler, Kay W. Axhausen, Kay W. Axhausen, and Kay W. Axhausen. *Identifying trips and activities and their characteristics from GPS raw data without further information*. ETH, Eidgenössische Technische Hochschule Zrich, IVT, 2008.
- [21] Google Official Blog. The bright side of sitting in traffic: Crowdsourcing road congestion data.
<http://googleblog.blogspot.jp/2009/08/bright-side-of-sitting-in-traffic.html> 2009.
- [22] howstaffworks. How does Google Maps predict traffic?
<http://electronics.howstuffworks.com/how-does-google-maps-predict-traffic.htm> 2013.
- [23] 一般財団法人道路交通情報通信システムセンター. VICS の仕組み.
<http://vics.or.jp/structure/index.html>.
- [24] 東京都市圏交通計画協議会. パーソントリップ調査とは.
<http://www.tokyo-pt.jp/index.html>.

- [25] 動線解析プラットフォーム WEBAPI 仕様書. 東京大学人の流れプロジェクト.
<http://pflow.csis.u-tokyo.ac.jp/wp-content/uploads/webapi.pdf> 2014.
- [26] 総務省統計局. 事業所・企業統計調査の概要と沿革.
<http://www.stat.go.jp/data/jigyou/gaiyou/> 1996.
- [27] 公益財団法人統計情報研究開発センター. 事業所・企業統計調査の提供.
<http://www.sinfonica.or.jp/datalist/index.html> 2011.
- [28] 総務省統計局. 国勢調査の役割.
<http://kokusei2015.stat.go.jp/about/role.htm> 2015.
- [29] 総務省統計局. e-stat 政府統計の総合窓口.
<http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do>.
- [30] 株式会社マイクロベース. data shop.
<http://microgeodata.com/shop/> 2013.
- [31] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R news*, Vol. 2, No. 3, pp. 18–22, 2002.
- [32] Ken-ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural networks*, Vol. 6, No. 6, pp. 801–806, 1993.
- [33] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, and others. *A practical guide to support vector classification*. 2003.
- [34] 静岡理工科大学菅沼研究室. ニューラルネットワーク.
http://www.sist.ac.jp/~suganuma/kougi/other_lecture/SE/net/net.htm.
- [35] C.M. ビショップ. パターン認識と機械学習 上. 2008.
- [36] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, Vol. 65, No. 6, p. 386, 1958.
- [37] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 1988.

- [38] Leo Breiman. Random forests. *Machine learning*, Vol. 45, No. 1, pp. 5–32, 2001.
- [39] Gilles Louppe, Louis Wehenkel, Antonio Suter, and Pierre Geurts. Understanding variable importances in forests of randomized trees. pp. 431–439, 2013.
- [40] Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. *Classification and regression trees*. CRC press, 1984.
- [41] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, Vol. 9, No. 3, pp. 293–300, 1999.
- [42] 赤穂昭太郎, 津田宏治. サポートベクターマシン. *数理科学*, No. 444, pp. 52–58, 2000.
- [43] 赤穂昭太郎. カーネルマシン. *信学技報*, NC-2003-34, 2003.
- [44] Mingqing Hu, Yiqiang Chen, and JT-Y. Kwok. Building sparse multiple-kernel SVM classifiers. *Neural Networks, IEEE Transactions on*, Vol. 20, No. 5, pp. 827–839, 2009.
- [45] 田中成典, 中村健二, 加藤諒, 寺口敏生. マイクロブログの投稿時間に着目したユーザの職業推定に関する研究. *情報処理学会論文誌. データベース*, Vol. 6, No. 5, pp. 71–84, 2013.
- [46] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, Vol. 38, No. 8, pp. 904–909, 2006.
- [47] 加納学. 主成分分析. *京都大学大学院工学研究科化学工学専攻プロセスシステム工学研究室* 2002年, 2002.
- [48] 古川俊之, 田中博. 主成分分析. *臨床検査*, Vol. 24, No. 5, pp. 577–583, 1980.
- [49] 神島敏弘. データマイニング分野のクラスタリング手法 (1). *人工知能学会誌*, Vol. 18, No. 1, 2003.
- [50] 神島敏弘. データマイニング分野のクラスタリング手法 (2). *人工知能学会誌*, Vol. 18, No. 2, 2003.
- [51] John A. Hartigan and Manchek A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Applied statistics*, pp. 100–108, 1979.

- [52] Dan Pelleg, Andrew W. Moore, and others. X-means: Extending k-means with efficient estimation of the number of clusters. pp. 727–734, 2000.
- [53] Tsunenori Ishioka. An expansion of x-means for automatically determining the optimal number of clusters. pp. 91–96, 2005.
- [54] 神嶌敏弘. 転移学習. 人工知能学会誌, Vol. 25, No. 4, pp. 572–580, 2010.
- [55] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 22, No. 10, pp. 1345–1359, 2010.
- [56] 日本の山岳データベース. 二地点の緯度・経度からその距離を計算する.
<http://yamadarake.jp/trdi/report000001.html> 2012.
- [57] 国土地理院測地部. 日本の測地系.
<http://vldb.gsi.go.jp/sokuchi/> 2014.

謝辞

本修士論文を書き上げるにあたり、多くの方から、ご指導、ご協力を頂きました。心より感謝申し上げます。

柴崎亮介先生には、お忙しい中研究の相談に乗っていただきました。二度の国際学会を始め、数回の国内学会への参加、情報銀行コンソーシアム、G 空間 EXPO, g-spase と研究に関する一連の環境を提供していただきました。先生との研究打ち合わせは気を引き締めながらも楽しくお話でき、厳しい言葉も頂いた事は、今でも忘れられず私自身の非常に良い経験になりました。心より感謝いたします。

副指導教員としてご指導いただいた貞広先生には研究手法の点をご指導いただきました。私の手違いにより、副指導教員のお願いで小さな問題を起こしたにも関わらず親身に相談に乗っていただいた他、本研究の今後について話しながらも研究手法を深化させる方向について指導いただき、普段の研究室生活とは異なる刺激を頂きました。有難う御座いました。

柴崎研究室特任助教の秋山氏には研究の指導だけでなく、充実した修士生活を送ることができたきっかけを提供していただきました。右も左もわからない私に共同研究を行っていた株式会社ナイトレイを紹介していただき、ナイトレイでデータ解析やシステム開発の業務を学べたきっかけは秋山さんでした。この経験がきっかけで、私を支えてくれる大勢の人と知りあうことが出来ました。修士2年に入ってから秋山さんの無茶振りが増えましたが、技術的経験や企業との交渉といった学生では珍しい経験が積めて非常に楽しかったです。新卒入社先では学会への参加が推奨されている事もあるため、もしかしたら学会でお会いするかもしれません。学会以外でも、今後もお付き合いできると幸いです。

研究員の金杉さんには技術的な点でお世話になりました。中途半端な知識しかなかった私にプログラミングの真髄を見せていただき、研究のモチベーションの源泉を提供していただきました。上山さんには mobmap の使用方法という、研究成果の可視化システムを提供していただきました。研究における信頼性検証において mobmap は手軽にわかりやすく利用でき様々な人に研究を説明することが出来ました。

MGD チームの仙石さんには毎週のゼミでの的確な指摘を頂いた他、様々な機会を頂きました。仙石さんの指摘のお陰で研究の土台を整備でき順調に研究を進める事が出来ました。小川さんにはゼミ内の勉強会の他、プライベートでもお世話になりました。分析に関する勉強会のおかげで分析手法の基礎が身についた他、コーディングの知識も身につきました。また研究室内で趣味の話やスポーツの話を通じ、生活の視野が広まりました。

同期の水野君、若生君、木村君、Shams 君、河地さんには公私共にお世話になりました。研究室のイベントの運営、送別会と取りまとめが大変でしたが、楽しく運営できたのは皆様のおかげだと思っております。水野君、若生君は

就職先の業界が似ているためもしかしたら会うかもしれません。その時は是非昔話に花を咲かせて楽しく話せると幸いです。木村君とは普段いる棟が異なっていたため、直接会うことは少なかったですが、Facebookでのアクティビティを拝見しいつも刺激をもらっていました。会社の同期に木村君と共通の友人がいるので、今度はその友人も交えて雑談できると嬉しいです。Shams君には英語面で大変お世話になりました。駒場へ移った際、全く英語が出来なかった私ですが、Shams君と話すお陰で拙いものの日常会話ができるようになり、英語への苦手意識がなくなりました。有難うございます。河地さんには研究のみならず普段の生活でもお世話になりました。普段の雑談でのやりとりは毎回楽しかったです。就職後は全国への転勤可能性があるということですが、またどこかのホールやアリーナで会えることを楽しみにしています。

社会文化環境学専攻の皆様には様々な知識を頂きました。建築、哲学、水環境、都市工学、社会学等の様々なバックグラウンドの方と同じ部屋で雑談したり、研究の話をするのは今まで受けたことのない刺激でした。専攻柄、進路は本当に多様にわたりますが皆様のご活躍をお祈りいたします。

株式会社ナイトレイの皆様には技術指導をしていただきました。入学時から業務を通じて様々な手法を身につけさせて頂いたり、様々な場を提供していただき、学生ではなかなか体験できない事をさせて頂きました。今後とも何卒宜しくお願い致します。

株式会社マイクロベースの皆様には貴重なデータを頂きました。総務省統計局の公開データを活用して非常に便利なデータを提供して頂いた裏ではCTOの桑田さんを始め様々な方のご尽力があったと思います。提供していただいたデータにより、研究に深みができました。有難う御座いました。

就職先の同期の皆様には研究手法の数理的解釈について、多くの着想を頂きました。普段の雑談が機械学習、Deep Learning、ミドルウェア、言語、サービスである集団は振り返っても皆様だけでした。これからも切磋琢磨しあう関係でいられると幸いです。今後共何卒宜しくお願い致します。

ここに名前を上げた方以外にも研究内外を問わず多くの方々のお陰で充実した二年間を送ることができ、本修士論文を書き上げる事ができました。最後にこれまでの学生生活を支えてくれた家族に感謝の意を表して、謝辞とさせていただきます。

付録

RandomForest でモデルを構築した時の各モデルの変数重要度の一覧

パーソントリップ調査を元に構築したモデルの各統計量

表 49: 滞留時間で集計した時の場所に関する特徴量の基本統計量

Statistic	N	Mean	St. Dev.	Min	Max
c1	248,817	948.335	2,947.880	0	43,542
c2	248,817	1,072.088	2,791.433	0	41,994
c3	248,817	1,524.849	3,668.787	0	42,462
c4	248,817	1,438.619	3,524.099	0	50,022
c5	248,817	981.118	2,983.723	0	50,382
c6	248,817	341.675	1,767.713	0	34,902

表 50: 滞在回数で集計した時の場所に関する特徴量の基本統計量

Statistic	N	Mean	St. Dev.	Min	Max
c1	130,867	0.106	0.435	0	7
c2	130,867	0.289	0.841	0	13
c3	130,867	0.431	0.911	0	9
c4	130,867	0.434	0.954	0	11
c5	130,867	0.292	0.735	0	12
c6	130,867	0.106	0.445	0	16

GPS データから抽出した特徴量の基本統計量

5.3 性別推定モデルにおける主成分と元の特徴量の関係

表 51: 性別と職業のクロス集計表

職業 \ 性別	1	2
1	12,611	11,327
2	2,893	2,721
3	118	5,387
4	3,380	4,032

表 52: 年代と職業のクロス集計表

年代 \ 職業	1	2	3	4
1	0	3,173	0	0
2	889	2,251	17	143
3	4,983	145	570	391
4	6,420	24	1,254	405
5	4,764	7	831	327
6	4,221	8	1,097	970
7	1,891	6	1,049	2,231
8	655	0	605	2,091
9	115	0	82	854

表 53: 年代と性別のクロス集計表

年代 \ 性別	1	2
1	1,609	1,564
2	1,572	1,728
3	2,608	3,481
4	3,473	4,630
5	2,767	3,162
6	2,864	3,432
7	2,350	2,827
8	1,443	1,908
9	316	735

表 54: 抽出した自宅勤務地の平均に関する基本統計量

Statistic	N	Mean	St. Dev.	Min	Max
avgUnAquired					
MinuitesWday	160	388.017	455.976	11	3,865
avgUnAquired					
MinuitesHday	160	384.374	1,498.644	0	17,735.390
avgHomeDptWday	159	78,346.920	987,201.600	0	12,448,193
avgHomeDptHday	158	186.725	1,598.215	0	16,811.560
avgWorkDptWday	157	15.411	3.789	0	21.842
avgWorkDptHday	146	14.308	4.101	0	22.750
avgHomeArvWday	109	14.384	4.195	3.222	22
avgHomeArvHday	93	12.965	5.862	0	51.840
avgWorkArvWday	105	12.306	2.729	7.750	19.667
avgWorkArvHday	105	6.170	5.759	0	21
avgHomeCntWday	159	1.777	0.916	0	6.859
avgHomeCntHday	159	1.390	0.845	0	4.053
avgWorkCntWday	160	0.785	1.558	0	16.684
avgWorkCntHday	160	11,581.800	22,851.950	0	110,379.100
avgMoveDistWday	160	39,654.140	36,079.770	0	256,372.500
avgMoveDistHday	160	29,250.460	47,716.310	0	327,538.500

表 55: 抽出した平日の時間帯別行動距離の平均に関する基本統計量

Statistic	N	Mean	St. Dev.	Min	Max
avgM0Wday	158	14,559,945	133,417,969	0	1,452,133,704
avgM1Wday	160	4,572,757	42,085,997	0	458,942,954
avgM2Wday	160	175,590	2,180,465	0	27,580,056
avgM3Wday	160	167.599	718.149	0	8,804.610
avgM4Wday	160	213.168	844.802	0	7,244.857
avgM5Wday	160	393.123	1,590.423	0	14,266.260
avgM6Wday	160	703.469	2,036.847	0	15,142.210
avgM7Wday	160	2,157.087	9,484.401	0	115,030.600
avgM8Wday	160	2,103.776	5,222.499	0	41,333.220
avgM9Wday	160	3,015.494	12,010.670	0	144,947
avgM10Wday	160	2,064.334	4,109.436	0	35,631.150
avgM11Wday	160	2,294.540	6,096.570	0	59,312.880
avgM12Wday	160	2,222.109	5,451.170	0	62,733.110
avgM13Wday	160	546,271.600	6,880,956	0	87,040,254
avgM14Wday	160	12,845.320	129,119	0	1,634,971
avgM15Wday	160	4,877.018	21,387.850	0	263,968.600
avgM16Wday	160	3,559.104	11,698.060	0	142,153.700
avgM17Wday	160	584,313.300	6,852,534	0	86,468,239
avgM18Wday	160	2,021.534	2,872.464	0	24,005.630
avgM19Wday	160	463,444.600	5,830,072	0	73,747,632
avgM20Wday	160	48,693.350	597,374.800	0	7,557,668
avgM21Wday	160	1,404.586	2,364.053	0	12,630.330
avgM22Wday	160	1,223.882	2,532.513	0	12,165
avgM23Wday	160	414,827.700	5,233,834	0	66,204,400

表 56: 抽出した休日の時間別行動距離の分散に関する基本統計量

Statistic	N	Mean	St. Dev.	Min	Max
avgM0Hday	160	365.397	719.014	0	4,319.360
avgM1Hday	160	652,958.500	5,231,603	0	56,352,163
avgM2Hday	160	35,837.580	451,678.800	0	5,713,464
avgM3Hday	160	164.787	650.563	0	5,866.441
avgM4Hday	160	261.391	1,077.869	0	11,412.390
avgM5Hday	160	507.959	2,045.165	0	21,722.980
avgM6Hday	160	1,457.144	3,222.460	0	19,667.620
avgM7Hday	159	3,758.565	6,057.176	0	35,609.580
avgM8Hday	160	4,126.742	6,964.835	0	55,175.190
avgM9Hday	158	3,775.064	16,062.420	0	183,311.700
avgM10Hday	160	418,095.900	3,925,713	0	44,690,532
avgM11Hday	160	901,006	10,745,919	0	135,734,675
avgM12Hday	160	38,421.290	463,176.800	0	5,860,477
avgM13Hday	160	35,836.480	422,282.100	0	5,341,842
avgM14Hday	160	632,240.600	7,334,023	0	92,435,638
avgM15Hday	160	199,834,205	2,502,467,319	0	31,654,904,711
avgM16Hday	160	431,236.500	5,420,125	0	68,562,306
avgM17Hday	158	366,544	3,333,479	0	36,197,569
avgM18Hday	160	256,805.700	3,205,911	0	40,555,241
avgM19Hday	160	461,476.300	5,798,467	2.243	73,348,482
avgM20Hday	160	110,605.700	1,374,669	0	17,390,232
avgM21Hday	160	57,989.280	709,896.700	0	8,981,335
avgM22Hday	160	1,560.694	3,383.910	0	33,871.850
avgM23Hday	160	462,243.400	3,521,826	1.144	42,340,570

表 57: 抽出した自宅勤務地の分散に関する基本統計量

Statistic	N	Mean	St. Dev.	Min	Max
varUnAquired					
MinuitesWday	160	498,216.600	3,541,877	9	32,336,866
varUnAquired					
MinuitesHday	160	1,729,282	15,473,877	0	181,176,848
varHomeDptWday	157	587.875	6,718.573	0	84,071.720
varHomeDptHday	156	2,119,143	22,491,064	0	275,890,913
varWorkDptWday	156	38,676,561	483,057,580	0	6,033,388,238
varWorkDptHday	145	53,558,895	644,933,836	0	7,766,031,838
varHomeArvWday	109	612,015.300	4,497,619	0	33,866,288
varHomeArvHday	92	46,185,660	442,531,476	0	4,244,659,092
varWorkArvWday	106	77,874,946	801,771,302	0	8,254,740,813
varWorkArvHday	104	13,068,541	133,273,449	0	1,359,127,839
varHomeCntWday	158	28,100.120	353,196.900	0	4,439,617
varHomeCntHday	158	15,340.580	192,391.100	0	2,418,347
varWorkCntWday	158	6,511.220	81,720.690	0	1,027,222
varWorkCntHday	158	597,506,011	3,505,104,209	0	36,021,822,843
varMoveDistWday	158	3,489,165,423	15,719,299,256	322.157	129,981,124,258
varMoveDistHday	158	4,944,029,444	29,386,502,536	0	325,093,239,898

表 58: 抽出した平日の時間別行動距離の分散に関する基本統計量

Statistic	N	Mean	St. Dev.	Min	Max
varM0Wday	158	13,567,443	51,774,869	0	439,485,427
varM1Wday	158	3,498,169	26,421,795	0	316,861,463
varM2Wday	156	5,278,036	47,267,802	0	569,877,350
varM3Wday	156	428,822.500	2,854,985	0	32,387,930
varM4Wday	156	725,255.900	5,443,934	0	64,110,022
varM5Wday	156	7,347,991	56,473,069	0	645,782,544
varM6Wday	156	9,483,202	55,795,115	0	655,856,870
varM7Wday	156	263,409,752	2,899,463,935	0	36,200,328,433
varM8Wday	156	117,661,918	798,971,967	0	7,844,662,223
varM9Wday	156	708,138,244	7,869,248,389	0	98,178,458,813
varM10Wday	156	74,034,968	523,266,687	0	6,314,794,953
varM11Wday	156	68,410,135	629,886,857	0	7,866,574,245
varM12Wday	156	23,784,097	62,444,279	0	415,168,741
varM13Wday	156	28,139,004	77,341,488	0	539,083,758
varM14Wday	156	34,394,731	91,699,890	0	581,157,420
varM15Wday	156	2,176,399,534	26,711,220,228	0	333,658,422,316
varM16Wday	156	622,917,042	7,365,300,400	0	92,020,237,145
varM17Wday	156	63,613,357	463,529,790	0	5,747,694,078
varM18Wday	156	33,639,009	158,963,255	0	1,907,259,329
varM19Wday	156	613,205,150	7,357,127,150	0	91,911,339,086
varM20Wday	156	20,131,751	53,510,258	0	436,693,208
varM21Wday	156	20,978,326	62,058,735	0	471,074,546
varM22Wday	156	16,441,569	62,549,008	0	624,383,178
varM23Wday	156	15,263,457	44,836,976	0	345,756,961

表 59: 抽出した休日の時間別行動距離の分散に関する基本統計量

Statistic	N	Mean	St. Dev.	Min	Max
varM0Hday	156	3,551,214	12,744,838	0	95,309,753
varM1Hday	156	749,319.900	4,667,973	0	38,171,338
varM2Hday	156	390,787.300	1,908,658	0	17,340,487
varM3Hday	156	760,724.700	5,402,338	0	56,111,029
varM4Hday	156	1,574,130	12,068,943	0	143,717,712
varM5Hday	156	4,319,334	19,060,944	0	163,488,619
varM6Hday	156	33,125,605	270,056,875	0	3,347,116,343
varM7Hday	156	38,587,030	141,796,978	0	1,354,427,213
varM8Hday	156	299,254,610	2,647,919,470	0	31,625,940,578
varM9Hday	156	806,930,102	9,695,649,494	0	121,125,196,753
varM10Hday	156	31,698,384	80,639,611	0	764,919,916
varM11Hday	156	342,703,005	3,540,205,900	0	43,922,713,108
varM12Hday	156	37,006,172	141,547,359	0	1,551,824,655
varM13Hday	156	26,148,812	83,313,382	0	968,462,233
varM14Hday	156	20,714,869	50,451,941	0	457,012,889
varM15Hday	156	20,006,903	45,058,891	0	357,706,387
varM16Hday	156	17,440,499	29,261,359	0	207,298,326
varM17Hday	156	44,886,090	192,285,471	0	2,282,385,914
varM18Hday	156	55,757,253	197,518,274	0	1,992,357,350
varM19Hday	156	350,153,701	3,513,011,809	85.543	43,804,354,539
varM20Hday	156	49,739,676	307,441,977	54.478	3,815,648,687
varM21Hday	156	41,422,410	240,426,359	3.514	2,979,575,359
varM22Hday	156	21,003,603	53,318,205	0	386,403,794
varM23Hda	101	15,860,540	36,842,398	20.927	239,450,015

表 60: 第二主成分と, 元の特徴量との関係

変数名	固有ベクトル
avgC4.3Visit	0.1942792
varC4.3Visit	0.1747607
avgC4.19Visit	0.1660698
varC4.19Visit	0.1612222
varC4.16Visit	0.1466180

表 61: 第 3 主成分と, 元の特徴量との関係
変数名 固有ベクトル

変数名	固有ベクトル
avgMoveDistWday	0.1382669
avgC3.3Visit	0.1153758
avgC5.21Visit	0.1117290
avgMoveDistHday	0.1078960
avgC3.18Visit	0.1050005

表 62: 第 1 主成分と, 元の特徴量との関係
変数名 固有ベクトル

変数名	固有ベクトル
avgC1.3Visit	0.1878917
varC1.3Visit	0.1707204
avgC1.22Visit	0.1685699
varC1.22Visit	0.1634020
varC1.23Visit	0.1572304

表 63: 第 37 主成分と, 元の特徴量との関係
変数名 固有ベクトル

変数名	固有ベクトル
varC2.10Visit	0.1933161
avgC2.6Visit	0.1866215
avgC2.10Visit	0.1855387
varC2.6Visit	0.1804307
avgC1.19Visit	0.1689028

表 64: 第 9 主成分と, 元の特徴量との関係
変数名 固有ベクトル

変数名	固有ベクトル
avgM15Hday	0.01164394
varWorkCntWday	0.01175727
varWorkCntHday	0.01178161
avgM1Hday	0.01180227
varC5.0Visit	0.01610301