GPSやコンテクストデータを活用した行動目的の推定手法に関する研究

塚本　健太郎[1]・佐藤　仁美[2]・森川　高行[3]

[1]非会員　名古屋大学大学院環境学研究科都市環境学専攻（〒464-8603 愛知県名古屋市千種区不老町）
E-mail:tsukamoto.kentarou@b.mbox.nagoya-u.ac.jp
[2]正会員　未来社会創造機構　特任講師
E-mail:sato@trans.civil.nagoya-u.ac.jp
[3]正会員　未来社会創造機構　教授
E-mail:morikawa@nagoya-u.jp

従来行われているパーソントリップ調査では，莫大な費用がかかることや回答率の低下による調査の継続性や精度低下が問題となっている．そこで本研究では，被験者の補助入力を一切必要とせずに，スマートフォンのGPS情報などから行動目的を推定する手法の検討を目的とし，決定木とRandomForestによるモデル構築，分析を行った．その結果，行動目的の判別精度は7割を超えることが示された．さらに，調査実施時における適切な実施期間，被験者数の検討も行ったところ，適切な調査実施期間は1か月程度であることが示され，被験者数の増加に伴って判別精度は向上することが示された．

Key Words : *Trip Purpose,GPS, Decision Tree,Random Forest*

1.　背景

従来，交通行動分析には伝統的にパーソントリップ調査(以下PT調査)が用いられてきた．PT調査とは，10年に1度一定地域(大都市圏)において無作為に選ばれた個人を対象に，ある平日の1日の人の移動を調べ，「どのような人が」，「どのような目的で」，「どこからどこへ」，「いつ」，「どのような交通手段を使って」移動しているかを把握することを目的とした調査である．調査方法としては，個人に対し調査票を配布し，記入・送付してもらう場合と，WEBアンケートを利用する場合とがある．

しかし，近年アンケート調査にかかる莫大な費用や，回答率の低下による調査の継続性やデータの精度低下が問題となっている．

この問題の改善を期待されているのがGPS搭載の携帯電話である． GPS搭載の携帯電話を活用したシステムはアンケート調査と比較して，被験者の負荷が少なく長期間の調査も可能であるなどの多くのメリットがあるものの，移動の概念の説明や携帯電話の操作方法，PCでの入力方法を説明するために被験者は説明会への参加が必須となることや，長期にわたる調査では操作が面倒になってしまい，いい加減な入力が行われることによる，精度の低下がデメリットとして挙げられる．そこで，被験者の操作をなくし，鞄などに入れているだけで活動データを自動収集することが可能なモバイル端末も開発されているが，移動手段や行動目的まで完全自動で精度高く把握することはできていない．また，GPS情報を用いた行動目的の推定を行った研究は少ない．さらに，実際の調査実施時において，行動目的の判別に用いるための適切な期間，必要な被験者数を検討する研究は見られない．

そこで本研究では，主に2つの目的を設定した．まず1点目は被験者の補助入力を一切必要とせずに行動目的を推定する手法の検討である．スマートフォンから得られるGPS情報に加え，場所や曜日，時間などのコンテクスト情報を長期的に取得，補足し，決定木モデル，RandomForest(RF)の2つの手法により検討する．2点目は調査実施時における，行動目的を推定するために必要な情報量の検討である．実際の調査に用いられる場合，どれだけの期間に，どれだけの被験者を対象に調査を行えば調査結果を有効に活用できるかを検討する．

2.　研究の概要

(1)　使用データ概要

本研究で使用したデータの概要を表−1にまとめた．被験者数の合計は156名であるが，①と②に重複した19名の被験者がいるため，全被験者数は137名である．また，「その他」トリップの全体に占める割合は約9.5%であり，全体の1割に満たないこと，また「その他」トリップを分類することは困難である

と考えられるため，「その他」目的以外の5つの行動目的を推定するものとし，モデル構築データに「その他」トリップを含めない．ただし，実際の調査において予測するデータから「その他」トリップは除けないため，予測データには含むものとする．よって判別精度の最大値は約90%となる．

表-1　使用データ概要

項目	内容
調査期間	①2010/11/22~12/19 ②2008/9/24~10/30 ③2008/11/19~12/31 ④2011/12/5~2012/2/7
調査方法	プローブパーソン(PP)調査
調査地域	中京都市圏
被験者数	①50名, ②30名, ③50名, ④26名
トリップ数	①3846, ②2568, ③4678, ④3038
GPS取得間隔	数分から数秒まで様々
行動目的	出勤，帰宅，帰社 業務，自由目的，その他

(2)　自宅と勤務先の位置推定

本研究では出勤・帰宅の判別精度を上げるため，自宅と勤務先の位置を推定する．取得した GPS 情報から各トリップの出発地(O メッシュ)，目的地(D メッシュ)を取得し，1 日の最終トリップの D メッシュのうち，調査期間中の最頻出メッシュを被験者ごとに集計する．また 1 日の最初トリップの D メッシュのうち，調査期間中の最頻出メッシュを被験者ごとに集計する．これらをそれぞれその被験者の自宅，勤務先の存在するメッシュと定義する．

図-1 に GPS 取得期間を 1 週間から 5 週間まで増加させたときの自宅，勤務先の位置推定精度の変化を示す．自宅の推定精度は 3 週間から 80%を超える精度となり，4 週間目からは 82.4%の精度で安定し，勤務先の推定精度は 3 週間までは単調に増加し，3 週間目から 56.3%の精度で安定している．よって自宅の位置推定に 4 週間程度の GPS 取得期間があれば約 80%の精度，勤務先(登校先)の位置推定にも 3 週間程度の GPS 取得期間があれば 56.3%の精度となることがいえる．

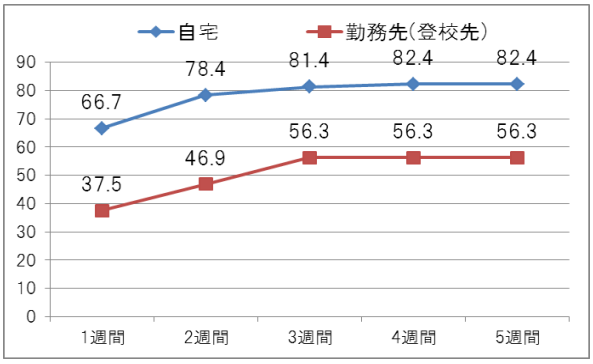

図-1　自宅と勤務先の推定精度の推移

(3)　本研究の流れ

本研究の分析の流れを図-2に示す．前節で説明したように自宅と勤務先の位置推定を行った後，それらの情報やGPS情報から行動目的の判別に有効であると思われる説明変数を作成する．本研究では33の説明変数を用いた．そしてクロスバリデーション(交差確認・検証法)を用いて，決定木モデルとRFにより判別精度，適切な実施期間，被験者数を検証する．

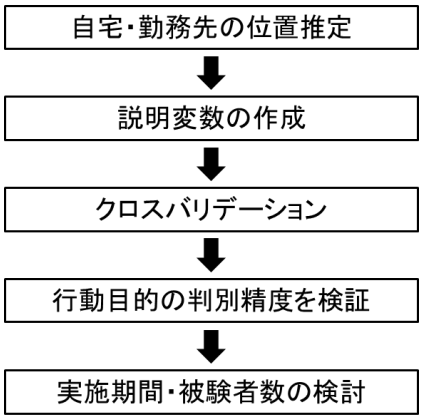

図-2　本研究の流れ

ここで，クロスバリデーションの概要を以下に示す．

1)すべてのデータをできる限り均等になるようにN 個のデータに分割.
2)N 個のうち，1 データを予測データ，その他すべてのデータを学習データとして，N 回検証を繰り返す.
3)N 回繰り返した検証精度の平均値を判別精度とする.

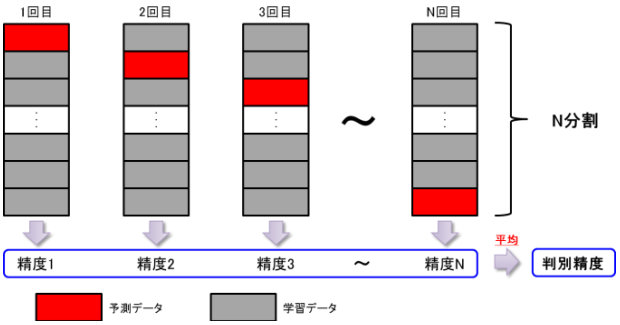

図-3　クロスバリデーション概要

本研究では個人の長期的データを用いるため，被験者ごとに，各グループのトリップ数の差が最も小さくなるように5つのグループに分割した．

以下の表-2に分割の詳細を示す．ただし，各行動目的のトリップ数は学習データにおけるものである．

表-2　データ分割の詳細

項目	グループ				
	1	2	3	4	5
被験者数	28	28	27	27	27
学習データトリップ数	9973	9944	10474	10415	10306
予測データトリップ数	3096	3046	2639	2657	2692
出勤	1550	1553	1673	1731	1761
帰宅	3197	3287	3275	3157	3360
帰社	491	576	640	674	631
業務	2538	2212	2506	2622	2258
自由目的	2197	2316	2380	2231	2296

(5)　使用手法概要

a)　決定木モデル概要

決定木(Decision tree)は，データの特徴量を用いた簡単な相関ルールで分岐を次々と作り，特徴空間を複数の矩形領域に分割し，そこに単純なモデルを当てはめることで判別を行うモデルである．決定木のモデルは木構造で表現され，木の根(root)からラベルに対応するターミナルノード(terminal node)への経路によって個々のデータを表現する．決定木では，作成した判別ルールを木構造で図示することが可能であり，これにより判別構造が高次元である場合でも木構造を見ることで判別の過程を見ることが容易にでき，その判別ルールの可読性の高さが決定木モデルの利点である．

本研究では，このような木構造を用いたモデルにRを使用し，代表的な決定木のアルゴリズムであるCARTを用いる．CARTでは以下の流れで判別木構造を構築する．(岡田,2011；金森,2009；金,2006；R-Tips)

1)木の構築：何らかの停止基準を満たすまで，あらかじめ定義したコストに基づいて特徴空間を2分割する操作を繰り返して木を構築する．

2)剪定(pruning)：構築された木の深さが深ければ深いほど，複雑なデータを扱うことができる一方，あまりに複雑にしすぎた場合，データに過学習してしまう．そこでモデルの複雑度を制御するために，木の深さなどのあらかじめ決めておいたパラメータに基づき木を剪定する．これを「枝刈り」ともいう．

木の構築に関して，ターミナルノードのコストQとして以下Gini係数を用いる．

$$Q = 1 - \sum_{i=1}^{c} p\left(\frac{i}{t}\right)^2 \qquad (1)$$

ここで，$p(\frac{i}{t})$はノードtでクラスiをとる確率である．Gini係数は不純度を表す指標であり，その減少幅が最も大きくなるように各分岐で分類を行う．

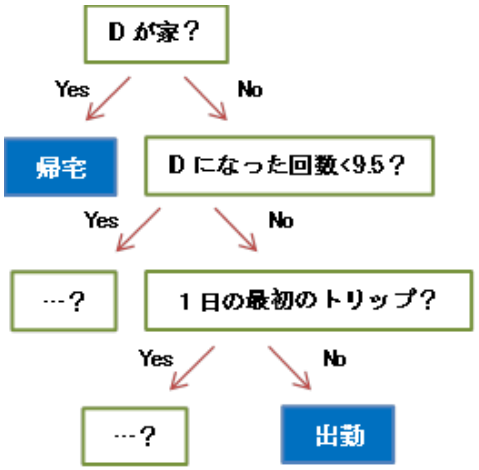

図-4　決定木モデル概要

b)　RandomForest概要

RFとは，決定木モデルの精度を高めるために開発された手法であり，機械学習アルゴリズムの一種として，分類，回帰及びクラスタリングに用いられる．以下にRFの概要を示す．(岡田,2011；金森,2009；金,2006；R-Tips)

1)与えられたデータからB組のランダムサンプルを生成(本研究ではB=500)．総説明変数がMのとき，分類木では$\sqrt{M}$の変数，回帰木では$M/3$の変数をそれぞれランダムサンプリングする．

2)各サンプルを用いてB本の未剪定かつ，最大の決定木または回帰木を生成．分岐ノードはランダムサンプリングされた変数の中の最善のものを用いる．

3)すべての結果を組み合わせ，一つの新たな分類器を構築．

RFは決定木をランダムにいくつも生成・統合し，より有効性の高い分類器を構築して分類を行うため，決定木モデルに比べて精度が高く，また過学習を行わないという利点がある．

本研究では，行動目的の判別精度はRFにより検証する．また，説明変数ごとのGini係数の値とターミナルノード決定に使用された総回数が算出されるため，説明変数の影響度がわかることも利点といえる．ただし，RFではデータに欠損値が存在するトリップは判別に用いられないため，欠損データが存在するとデータの母体数が少なくなってしまう．またランダムに決定木を生成するため，学習過程が不明瞭であるという欠点をもつ．

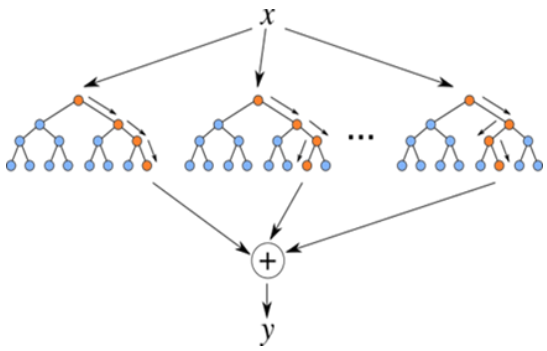

図-5　RF概要

3.　判別精度の検証

(1)　判別精度

決定木モデルとRFそれぞれにクロスバリデーションを行った結果を図-6に示す．

決定木モデルの精度の平均が63.0%，RFの精度の平均が68.3%と，両手法とも60%を超える精度となり，RFの精度が決定木モデルの精度を上回る結果となった．

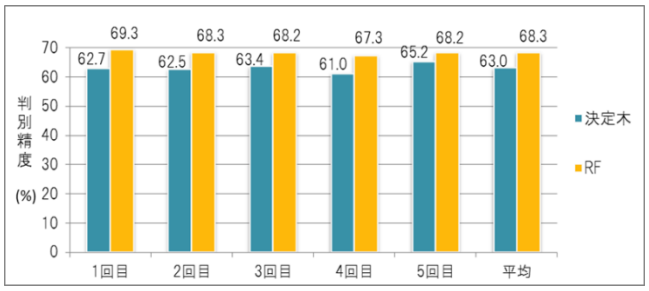

図-6　判別精度

(2)　行動目的別判別精度

次に，RF により行動目的別の判別精度の検証を行った．図-7 に，クロスバリデーションを行った結果を示す．

「帰宅」トリップは 91.4%と，非常に高い精度となった．これは自宅の判別の精度が高いことによるものと考えられる．一方，「出勤」トリップは「帰宅」トリップに次いで高い精度となったものの，76.3%と「帰宅」トリップに比べて精度が低い結果となった．この結果は勤務先の判別精度があまり上がらなかったことが原因であると考えられる．また，「帰宅」，「出勤」トリップが他に比べて高い精度となった理由として，平日であればほぼ毎日決まった時間帯に行われる可能性が高い行動であり，被験者ごとにパターン化されているため多くの説明変数が影響していると考えられ，多角的な判別を行うことが可能であるために精度が高くなったと考えられる．また，「業務」トリップは約 60%の精度，「自由目的」トリップは約 70%の精度となった．一方，「帰社」トリップに関しては，他 4 つの行動目的に対して比較的精度が上がらなかった．これは本研究で用いた「帰社」トリップの総数が少ないことがまず考えられる．また「帰宅」トリップや「自由目的」トリップなどのように全被験者に取得されているトリップではないため，総数が少ないことでクロスバリデーションを行うためのデータ分割時に「帰社」トリップの少ないグループができてしまい，学習データ数が少なくなることで，そのグループの精度が上がらずに平均精度も下がってしまうと考えられる．

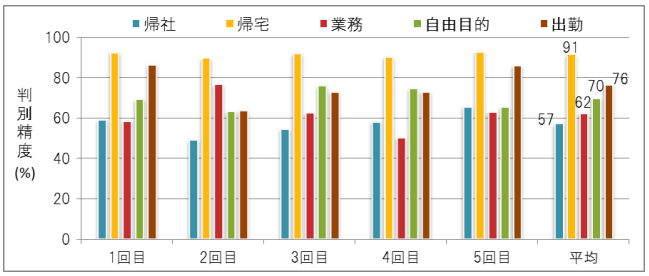

図-7　RF による行動目的別判別精度

(3)　決定木モデルによる判別過程の可視化

RF を用いることで，高い精度で判別が可能になるが，実際にどのような過程で判別が行われているかは不透明である．そこで，決定木モデルを用いて判別過程を可視化する．ただし，RF と決定木モデルによる分類器は異なるため，大まかな判別過程を可視化するために行うこととする．

以下の図-8 にクロスバリデーションの分割のうち，グループ 1 を例として決定木モデルを示す．なお，判別率は 62.7%(1942/3096)である．

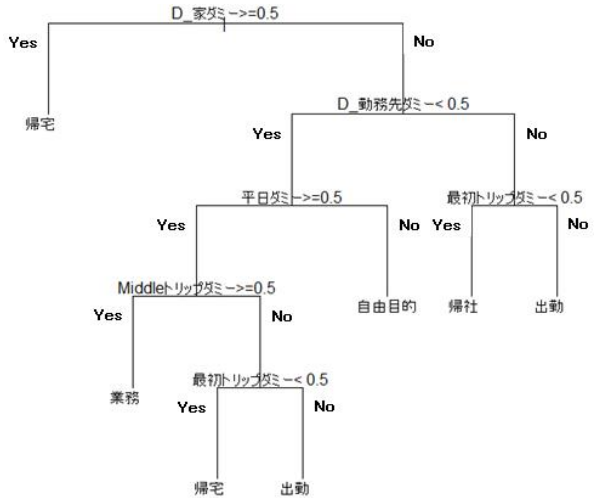

図-8　決定木モデル例（グループ 1）

図-8 の決定木モデルでは 5 変数のみを用いている．これらの 5 変数が存在していれば行動目的の判別は 60%を超えることがいえる．他のグループの決定木モデルにおいても 5 個から 8 個の有効な変数が存在していれば判別精度は 60%を超えることが分かった．

(4)　実施期間の検討

前節までにクロスバリデーションによる判別精度の検証を行ったが，ここで実際の調査施行時においてどれだけの期間調査を行えばどの程度の判別精度が得られるか検証し，行動目的の判別に適切な調査実施期間を検討する．

表-1 によれば，同一の被験者が実施した調査のうち，最も取得期間の長いデータは 13 週間であり，その被験者数は 19 名である．学習データとしてこの 13 週間分のデータを用い，1 週間から 13 週間分までデータ量を増加させ，また予測データをその他 118 名の取得開始から 4 週間目までのデータに統一した．最も取得期間の短いデータは 4 週間であるため，予測データを 4 週間分とした．そして週を増加させる順序をランダムに入れ替え，この操作を 10 回繰り返し，その平均を検証精度とした．

以下の図-9 に，決定木モデル，RF それぞれを用いて検証した GPS 取得期間の変化による判別精度の推移を示す．

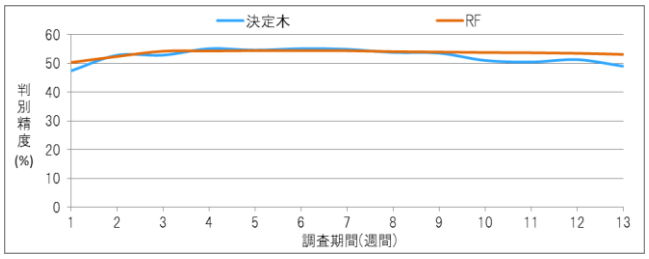

図-9　GPS 取得期間による判別精度の推移

図-9 からわかるように，2 週間を超えるあたりまで決定木の精度が RF の精度を下回り，実施期間が 4～9 週間程度までは，決定木と RF の精度にほとんど差がなくなることがわかる．また，9 週間を超えるあたりから再び決定木の精度が RF の精度を下回る結果となった．また決定木，RF ともに GPS 取得期間が 3～4 週間程度まで判別精度は向上し続け，4 週間を超えたあたりで判別精度は安定する．その後 7 週間程度までは，ほとんど精度に変化は見られない．一方，実施期間が 7 週間程度を超えたあたりから両手法とも精度が緩やかに下がり始めていることがわかる．

そこで，どの行動目的が精度低下の原因であるかを検証するため，図-10，図-11 に決定木モデルと RF それぞれの行動目的別の判別精度推移を示した．

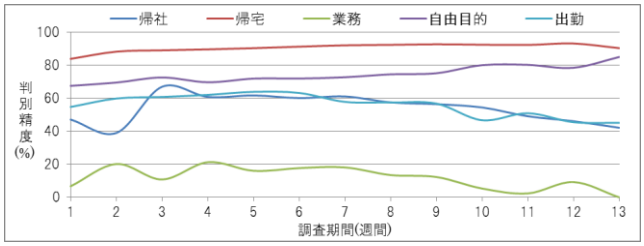

図-10　決定木による行動目的別の判別精度推移

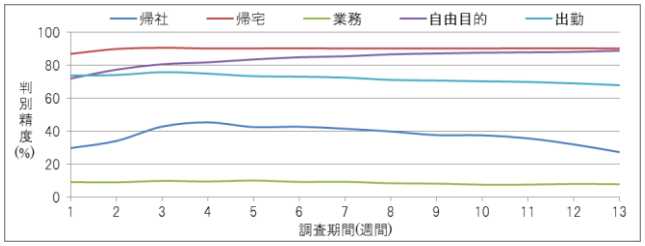

図-11　RF による行動目的別の判別精度推移

両グラフから，「帰宅」トリップは決定木モデルでは 12 週目まで緩やかに上昇し，13 週目で低下がみられる．RF では 3 週目あたりで精度は安定することがわかる．また，「帰社」，「業務」及び「出勤」トリップは両グラフとも 4～7 週間程度でピークを迎え，その後精度は低下している．

これらのことから，仕事に関与する「帰社」，「業務」，「出勤」が主に精度低下の原因であるといえる．これは実施期間が 4 週間程度となるまでにそれぞれ被験者の行動パターンは十分把握可能であり，それ以上データを増加させることによって，「帰社」や「業務」など，普段の行動パターンから外れやすいトリップによって特異なパターンが取得され，精度が低下してしまう可能性が考えられる．

以上のことから，実施期間が 7～8 週間を超えると判別精度は低下してしまうため，実施期間を増加させ続けることは適切とはいえず，実施期間が 4 週間程度あれば判別精度はピークを迎え，その調査の最も有意な結果が得られるといえる．また，自宅と勤務先の位置推定も 3～4 週間程度の取得期間があれば精度のピークを迎えるため，実施期間は 4 週間程度が適切であるといえる．

(5)　被験者数の検討

次に，被験者数を増加させたときの判別精度の変化を検証する．全被験者 137 名の取得開始から 1 週間のデータを学習データとし，10 名毎に 137 名まで増加させ，また予測データを全被験者 137 名の 2 週間目以降の 3 週間分のデータに統一した．GPS データの長期間の取得による被験者への負担が調査の問題点として挙げられていることを考慮し，できる限り被験者の負担を減らすため，調査期間を 1 週間と仮定し，学習データは 1 週間分とした．調査実施期間の検討と同様に，被験者を増加させる順序をランダムに入れ替え，この操作を 5 回繰り返し，その平均を検証精度とした．

以下の図-12 に決定木モデル，RF それぞれを用いて検証した被験者数の変化による判別精度の推移を示す．

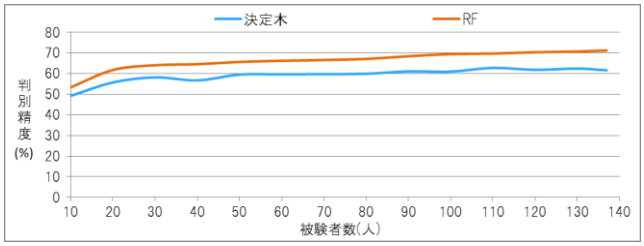

図-12　被験者数による判別精度の推移

図-12 を見ると，決定木，RF ともに被験者数が 2～30 名を超えたあたりで精度の上昇は緩やかになり，その後被験者数の増加に伴って精度は向上し続けていることがわかる．

決定木モデルでは，被験者数が 90 名程度で精度は 60%を超え，RF では 20 名程度で 60%，120 名程度で 70%を超えることがわかる．今回の検証における学習データは全被験者 137 名の 1 週間分のデータであることから，学習データの GPS 取得期間が 1 週間程度と短い場合でも，被験者数を増加させることで精度は 70%以上を期待できる．

さらに，本研究では被験者数が最大で 137 名であるが，図-12 を見ると，さらに被験者数を増加させることで判別精度は向上するものと考えられ，ある程度の被験者数で安定する可能性が高い．ただし，今回の検証に用いた学習データと予測データの被験者は同一であるため，学習データにおける被験者毎の行動パターンがその被験者の予測データに反映されやすいと考えられる．そのため，学習データと予測データの被験者が異なる場合において，判別精度は今回の検証結果を下回る可能性が高い．

5.　まとめと課題，展望

RF を用いて，GPS 情報から出勤，帰宅，帰社，業務，自由目的の 5 行動目的を約 75%判別可能であることが分かった．自宅の位置推定精度が 80%を超えたことで，帰宅目的のトリップを 90%を超える精度で判別可能であったが，勤務先の位置推定精度が高いとはいえないため，出勤，帰社，業務目的のトリップを高い精度で判別することはできなかった．

また，実際の調査施行時においては，約 4 週間程度の調査実施期間が最も適切であることがいえ，7 週間を超える調査期間は適切ではないと分かった．さらに，被験者数の増加に伴って判別精度は向上すること分かった．

今回の結果から，実際に調査時には 1 週間程度，200 名程度のデータを取得してマスターデータを作成し，予測する被験者は GPS 情報のみを用いて予測を行う形式が最適であると考えられる．

また，本研究の手法を用いたアプリ等を開発することで，従来の PT 調査における若者の回答率低下問題に対し，精度の補完を行うことが可能であると考えられる．このように従来の PT 調査と組み合わせ，補完し合うことでより精度の高い調査が行うことが可能であるといえよう．

また，出勤，帰社，業務の判別精度向上のためには勤務先の位置推定精度の向上が必要である．

さらに，本研究では被験者数の増加に伴って判別精度は向上したが，どの程度の被験者数で判別精度は最大となるかを検討する必要である．

謝辞：本研究はJSPS科研費25630215の助成を受けたものです.

参考文献

1) 朝倉康夫・羽藤英二・大藤武彦・田名部淳(2000)：PHS による位置情報を用いた交通行動調査手法，土木学会論文集，pp.95-104.
2) 大野夏海・関本義秀・中村敏和・Horanont Treerayut・柴崎亮介(2013)：東京都市圏における長期の GPS データを用いた移動経路の推定に関する研究，地理情報システム学会講演論文集，No.F-41.
3) 岡田昌史(2011)：R パッケージガイドブック，東京図書，pp.210-224.
4) 金森敬文・竹之内高志・村田昇(2009)：R で学ぶデータサイエンス　Vol.5.パターン認識，共立出版，pp.107-113.
5) 金明哲(2006)：フリーソフトによるデータ解析.マイニング　Vol.32.R と集団学習，pp199-205
6) 前司敏昭・堀口良太・赤羽弘和・小宮粋史(2005)：GPS 携帯端末による交通モード自動判定法の開発，第 4 回 ITS シンポジウム 2005.
7) 中京都市圏総合都市交通計画協議会(2002)：第 4 回中京都市圏パーソントリップ調査報告書 Vol.1 実態調査の企画と実施，国土交通省中部地方整備局・愛知県・岐阜県・三重県・名古屋市，pp.49-54.
8) 羽藤英二・小島英史・横田幸哉(2005)：BCALs を用いた行動文脈の推定，土木計画学研究発表会講演集，Vol.31.
9) Li.S，Peter.R.S.(2006)：A process for trip purpose imputation from Global Positioning System data，Transportation Research Part C36，pp.261-267.
10) McGowen,P.，McNally,M.(2007)：Evaluating the Potential to Predict Activity Type from GPS and GIS Data，pp.5-18.
11) Wendy B，Kees M.(2008)：Deriving and validating trip purposes and travel models for multi-day GPS-based travel survey: A large-scale application in the Netherlands，Transportation Research Part C: Emerging Technologies，Vol.17，Issue3，pp.285-297.
12) R-Tips：http://cse.naro.affrc.go.jp/takezawa/r-tips/r2.html
13) 株式会社トランスフィールド HP：

http://www.transfield.co.jp/company.html
14) Sideswipe HP： (2014.8.1 受付)
http://kazoo04.hatenablog.com/

STUDY ON ESTIMATION METHODS FOR TRIP PURPOSE USING GPS AND CONTEXT DATA

Kentaro TSUKAMOTO, Hitomi SATO and Takayuki MORIKAWA

Due to the features of significant cost and low respondent rate, traditional Personal Trip (PT) survey has the problem about low accuracy and disability of continuity. In order to avoid these disadvantages, GPS data collected by smart phone were utilized in this study to infer trip purpose data. Based on the method of Decision Tree and Random Forest, model for inferring trip purpose was developed and some analyses proceeded subsequently. It is found that the accuracy of trip purpose inference calculated by the model developed in this study is more than 70%. Furthermore, it can be concluded that 1 month continuous GPS data is minimum requirement to get the satisfied results. Also, the results show that the accuracy of trip purpose reference can be improved by increasing the number of respondents.