# 機械学習を用いた
# 社会生活基本調査・調査票 A の在宅・不在判定

荒尾　俊介[1]・円山　琢也[2]

[1]学生会員　熊本大学大学院自然科学教育部土木建築学専攻（〒860-8555 熊本県熊本市中央区黒髪2-39-1）
E-mail:217d8351@st.kumamoto-u.ac.jp

[2]正会員　熊本大学教授　大学院先端科学研究部 （〒860-8555 熊本県熊本市中央区黒髪 2-39-1）
E-mail: takumaru@kumamoto-u.ac.jp (Corresponding Author)

我が国における代表的な時間利用調査である社会生活基本調査の調査票 A は，活動場所の情報を収集していないため，交通分析への活用がしにくい課題がある．本研究は，在宅・不在が把握可能な調査票 B を教師データとし，活動場所が不明な調査票 A データの在宅・不在を機械学習で判定する手法を提案する．具体的には，決定木やランダムフォレスト等の複数の機械学習手法を用いて推定結果や特徴量を比較し，在宅・判定に適した学習法の選定を行う．在宅不在判定の特徴量としては，各時間帯における行動内容が最も大きな影響を及ぼしており，次いで一緒にいる人や行動時間帯等が影響していることが分かった．また，この手法を利用して，全国都市交通特性調査のトリップ記入漏れの分析等を行った．

***Key Words:*** *time-use survey, out-of-home, machine lerning*

## 1. はじめに

### (1) 研究の背景と目的

社会生活基本調査は，我が国における代表的な時間利用調査であり，国民の生活時間や生活行動の実態を把握することを目的として，5 年周期で実施されている．調査の結果は，ワーク・ライフ・バランスの推進や男女共同参画社会の形成といった各種行政施策のための重要な資料となると共に，その他にも生活時間の国際比較や学術研究等，幅広い分野で利用されている[注1)]．

社会生活基本調査には調査票 A と調査票 B (以下，単純に調査票 A，調査票 B と記述) の 2 種類の調査票があり，調査票 A は 1976 年，調査票 B は 2001 年から調査を開始している．そのうち調査票 B では活動場所の情報を収集しており，交通分析で重要な指標となる在宅・不在の回答値やトリップ数の推定値を得ることがで可能であることから，最近ではそれらの特徴を生かし交通分野での活用が行われている．

実際の交通分析への活用例として，筆者ら [1)]は全国都市交通特性調査 (全国 PT 調査) の精度検証を行っている．自記回答式の PT 調査は調査対象者の記憶に依存するため，トリップの記入漏れが発生しやすいことが古くから指摘されており [2)]，調査票 B を用いて時間帯別個人不在率・非外出率，およびトリップ原単位の指標を算出し，同時期の PT 調査と比較することで，PT 調査のトリップ記入漏れの実態を推測している．

一方で，調査票 A は調査票 B に比べ豊富なサンプルがあり (調査票 A = 約 19 万人，調査票 B = 約 1 万人)，全国の生活実態を網羅的に把握しているものの，データに活動場所の情報が含まれていないため，前述のような交通分析への活用がしにくいといった課題がある．

調査票 A の在宅・不在を推定した例として深堀・円山 [3)]があり，その時間帯の行動，その時間前後の行動，一緒にいた人の情報から在宅・不在を推定している．しかし，1) 在宅・不在判定に関連する特徴量が個人属性や行動時間帯など他にも考えられること，2) 判定不可の行動はすべて在宅としている等の課題が残されていた．

そこで本研究では活動場所が不明な調査票 A の在宅・不在を教師あり学習を用いて分類する．具体的には，活動場所の情報がある調査票 B を教師データとし，多くの特徴量を考慮した機械学習モデルを構築して調査票 A の在宅・不在を予測する．

調査票 A の在宅・不在推定から以下のことが期待できる．まず，外出状況が正確に把握できる時間利用調査か

ら，不在率や不在時間の指標が算出可能になる点がある．これらの指標を算出することで，既存研究で行われてきPT 調査の精度検証[1),3)]，時間帯別不在率[4)]や不在時間[5)]の経年推移等を明らかにすることができる．また，世帯年収や住居の詳細な種類といった PT 調査では入手していない項目による分析が可能になる．

以上を踏まえて本研究の目的を以下とする．

1) 調査票 B の行動場所データから，教師あり学習を用いて調査票 A の在宅・不在を分類するモデルを構築し，分類に重要な特徴量を明らかにする．
2) 調査票 A から，個人不在率や不在時間等を算出し，同時期の全国 PT と比較する．また，トリップ記入漏れの影響が多い属性に着目し経年での傾向を明らかにする．
3) 全国 PT には含まれない項目を利用した分析を行い，新たな知見獲得を試みる．

### (2) 本研究の構成

これ以降，本論文の構成として，2.では使用データのほか，機械学習の手法や推定に用いる説明変数について述べる．3.では，調査票 A の在宅・不在推定モデルを構築してその評価を行い，4.で調査票 A を用いて全国 PT との比較や，同調査に含まれない調査項目の分析を行う．最後に 5.で本研究の成果をまとめる．

## 2. データと分析手法

### (1) 使用データの概要

本研究では，在宅・不在の真値を含む調査票 B の 2016 年を教師データとして使用し，調査票 A の 1996，2001，2006，2011，2016 年次における在宅・不在を 15 分単位で推定する．**表-1** に 2016 年の調査票 B と調査票 A の概要を示す．表中の在宅割合をみると平日は 67.1%，休日は 74.1%で，不在よりも在宅の割合が高い．なお，調査票 A は 1996 年以前も実施されているが，「一緒にいた人」の情報を把握しておらず，正確な推定が困難と考えられるため対象としていない．

### (2) 個人・世帯不在率，不在時間の定義

既存研究[4)]と同様に世帯構成員全員が不在の時間帯を世帯不在の状態と定義する (**図-1**)．また，ある時間帯において，対象個人のうち不在の状態にある個人の割合を個人不在率，対象世帯のうち世帯不在の状態にある世帯の割合を世帯不在率とする．同様に対象個人のうち不在であった時間を不在時間とし，世帯構成員全員が不在である時間を世帯不在時間と定義する．なお，社会生活基本調査は 24 時間の行動を 15 分単位で集計しているため，不在時間も 15 分単位で算出している点には留意が必要である．

**表-1** 調査票Bと調査票Aの概要 (2016年)

| | | 調査票B (教師データ) | 調査票A (予測データ) |
|---|---|---|---|
| 平日 | サンプル (人) | 6,884 | 132,592 |
| | 活動場所数* (個) | 660,864 | 12,728,832 |
| | 在宅割合** (%) | 67.1 | - |
| 休日 | サンプル (人) | 11,475 | 218,152 |
| | 活動場所数* (個) | 1,101,600 | 20,942,592 |
| | 在宅割合** (%) | 74.1 | - |

*活動場所数＝サンプル (人) ×24 (時間) × 4 (15分単位)
**在宅割合= 在宅数 / 活動場所数×100 (%)

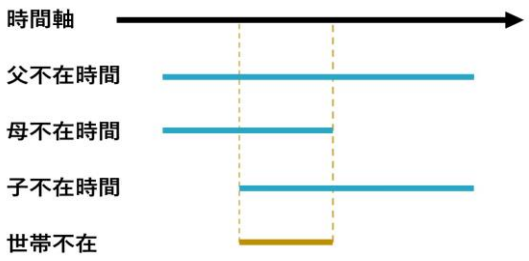



**図-1** 世帯不在率のイメージ図[4)]

### (3) 予測に用いる手法

調査票 A の在宅・不在を分類する手法として，教師あり学習を用いる．教師あり学習は，あらかじめ用意した入出力の組み合わせを訓練データとして与え，ここから入出力の関係を学習することで，新しく与えた入力に対して出力を予測する機械学習法である[6)]．本研究では，調査票 B の行動場所のデータから，その時間の行動や一緒にいた人，個人属性等の特徴量と在宅・不在の関係をモデルに学習させ，場所の情報が未知である調査票 A データの在宅・不在を分類する．教師あり学習には様々な種類がある[注 2)]が，分類結果の解釈がしやすいことや計算負荷が低いモデルであること，予測に用いる説明変数が多いこと等を踏まえ，決定木 (Decision Tree) および RF (Random Forest) を学習法として選定した．なお，本研究では計算コストの問題から，平日と休日それぞれでモデルを構築し，活動場所の分類を行う．

#### a) 決定木 (Decision Tree)

決定木とは，データベース上から重要な知識を抽出し木構造によるルールの組み合わせ表現するデータマイニング手法の一つであり，データ分類に用いられる[7)]．決定木は目的変数に強く関連している説明変数や注目したいサブグループを発見することを目的とした分析手法であり，得られた分類結果の意味を解釈したい場合，魅力的なモデルである．本研究では決定木の分析に CART

**表-2** 混同行列

| | | 予測 | |
|---|---|---|---|
| | | 正 (不在) | 負 (在宅) |
| 正解 | 正 (不在) | TP (true positive) | FN (false negative) |
| | 負 (在宅) | FP (false positive) | TN (true negative) |

(Classfication and Regression Trees) アルゴリズムを利用し，木の剪定には一般的な手法である Cost-complexity Pruning を用いた [8].

b) RF (Random Forest)

RF は，複数のモデルを融合させて 1 つの学習モデルを生成するアンサンブル学習の一つであり，ランダムに抽出した一部の特徴量とサンプルから相関の低い多様な決定木 (弱学習器) を作成し，分類問題の場合は構築した決定木の多数決をとって識別を行う手法である．RF は多様な決定木を用いることで，過学習を抑制し，説明変数が多い場合や説明変数間に相関が認められるような場合にも有効に機能する [9]．本研究では，RF のハイパーパラメータとして決定木の数，決定木の深さ，決定木に用いる特徴量を最適化した．

### (4) 評価指標

在宅・不在の分類をどの程度当てられたかを，正解率 (Accuracy)，F 値 (F-Value)，$\kappa$ (Kappa) 係数，AUC (Area under the curve) の 4 項目を用いて評価する．

正解率は予測と正解が一致したデータの割合を表し，F 値は適合率 (Precision) と再現率 (recall) の調和平均により算出される．正解率，適合率，再現率，F 値は，**表-2** に示す混同行列を用いて，式(1)～(4)のように算出される．なお，式(2)，(3)はそれぞれ正例を不在としているが，本研究では両方のクラスの分類精度を確かめるため，正例を在宅とした場合の F 値も算出する．

$$正解率 = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$適合率 = \frac{TP}{TP + FP} \tag{2}$$

$$再現率 = \frac{TP}{TP + FN} \tag{3}$$

$$F\,値 = \frac{2 \times 適合率 \times 再現率}{(適合率 + 再現率)} \tag{4}$$

$\kappa$係数は真値と予測値の一致度が偶然の一致からどのくらい逸脱するかを表す指標であり，式(5)で示される．式中において$P_0$は混合行列の真値と予測値が一致した確率，$P_e$は真値と予測値が偶然一致した確率であり，それぞれ**表-2** の混同行列から算出することができる．$\kappa$係数は 0 から 1 の値をとり，1 に近いほど予測精度が高いことを示す．

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \tag{5}$$

AUC はデータが正例に属する確率の順序をどの程度当てられたかを定量化する指標であり，ROC 曲線の下部の面積により定義される [10]．AUC の値が 1 に近づくほど確率の大きさによって正例と負例を区別できていると解釈でき，反対に AUC の値が 0.5 に近づくほどランダムな分類であることを意味する．

また，モデル性能の評価には$k$分割交差検証 (k-fold cross-validation) を用いた．$k$分割交差検証は，非復元抽出を用いて訓練データをランダムに$k$個分割した後，その

**表-3** 在宅・不在予測に用いる説明変数と概要

| | 説明変数 | 概要 | 種類数 |
|---|---|---|---|
| カテゴリ値 | 行動区分 | その時間帯にしていた行動 | 20 |
| | 一緒にいた人 | その時間帯に一緒にいた人 | 5 |
| | 行動時間帯 | 行動している時間帯 | 24 |
| | 行動分 | 行動している分 (15分単位) | 4 |
| | 曜日 | 平日は月～金の5種類を，休日は土日2種類を使用 | 7 |
| | 仕事の形態 | 仕事の雇用・勤務形態によって分類 | 5 |
| ダミー変数 | 性別 | 男性を1，それ以外を0とする | |
| | 昼食ダミー | 時間帯が12時かつ，行動区分が食事の場合なら1，それ以外を0とする | |
| | 前時間移動ダミー | 前時間の行動が「移動，通勤通学」なら1，それ以外を0とする | |
| | 休みの日ダミー | 対象日の状況を表すダミー変数．全6変数のうち当てはまるものすべてに1，それ以外を0とする | - |
| | 旅行・行楽ダミー | | |
| | 療養ダミー | | |
| | 行事・冠婚葬祭ダミー | | |
| | 出張・研修ダミー | | |
| | その他の日ダミー | | |
| 連続値 | 年齢 | 対象者の年齢 | |
| | 10歳以上の世帯員数 | 対象者世帯の10歳以上人数 | |
| | 10歳未満の世帯員数 | 対象者世帯の10歳未満人数 | |

うち$k-1$個をモデルの訓練に，1 個を性能の評価として使用し，この手順を$k$回繰り返し評価指標の平均値を算出することで，モデル性能を評価する．一般的に用いられる$k$の値は 10 であるが [11]，本研究ではデータセット数が大きく，計算コストを削減するために$k=5$とした．

### (5) 予測に用いる説明変数

**表-3** に在宅・不在の予測に用いる説明変数と概要を示す．名義尺度として項目を区別するカテゴリ変数には，行動区分 (20 種類) や仕事の形態 (5 種類) 等を用いた．カテゴリ変数の詳細な項目については本稿末尾の**付録**に示す．ダミー変数には性別をはじめ，対象日の状況等を用いた．対象日の状況は複数回答が可能であり，例えば仕事が無い休みの日で地域の行事等に参加していた場合[注 3]は，「休みの日ダミー」「行事・冠婚葬祭ダミー」が 1 となり，それ以外は 0 となる．また，仕事や学校のみの日の場合は「その他の日」が 1 となり，それ以外は 0 となる．連続値としては対象者の年齢や，世帯属性を用いた．

## 3. 在宅・不在推定の分析結果

### (1) 調査票 B の在宅・不在推定

**表-4** に，在宅・不在の分類の評価値をモデル別に示す．すべての評価指標において十分な精度があるといえる．2つのモデルを比較すると RF のほうが分類精度が高く，在宅・不在をうまく予測できていることがわかる．また，平休日でみると休日において分類精度が低くなっており，F 値の結果から不在の予測が低いことがわかる． 別途休日の不正解データを行動別に確認したところ，「睡眠」や「食事」の不在行動を在宅と予測した数が多く，外食や外泊といった行動をうまく分類できなかったことが考えられる．

**表-4** 分類モデルの評価値

| | | 正解率 | $\kappa$係数 | F値 (不在) | F値 (在宅) | AUC |
|---|---|---|---|---|---|---|
| 平日 | 決定木 | 0.956 | 0.900 | 0.933 | 0.968 | 0.962 |
| | RF | 0.988 | 0.973 | 0.982 | 0.991 | 0.999 |
| 休日 | 決定木 | 0.936 | 0.828 | 0.871 | 0.909 | 0.915 |
| | RF | 0.977 | 0.940 | 0.955 | 0.985 | 0.995 |

ここからは，在宅・不在の分類に使用される重要な特徴量やルールについて述べる．**図-2,3** に平休日における在宅・不在の決定木を示す．決定木の終端ノードのグラフはピンク色が在宅，水色が不在の割合を示し，各ノードに表示される説明変数のみが，実際に分類のルールとして使用される．平日では分類を行う特徴量として，行動区分，一緒にいた人，旅行・行楽ダミーがルールとして用いられている．休日では，平日の特徴量に加えて，行動時間帯，年齢，仕事の形態，が分類のルールとして用いられている．

**表-5** に RF から推定した特徴量の重要度の上位 6 項目を示す．ここでいう重要度とは RF 内のすべての決定木から算出されたジニ不純度の平均的な減少量のことであり，値が大きいほどその特徴量が目的変数の分類に影響を与えることを意味する．なお重要度は，特徴量全体から見た相対的な指標であるため，すべての特徴量の重要度の和は 1 となる．

表より，平休日ともに行動区分が分類において最も重要な変数であり，全体のおよそ半分を占めていることがわかる．また前述の深堀・円山 [3]の推定では使用されて

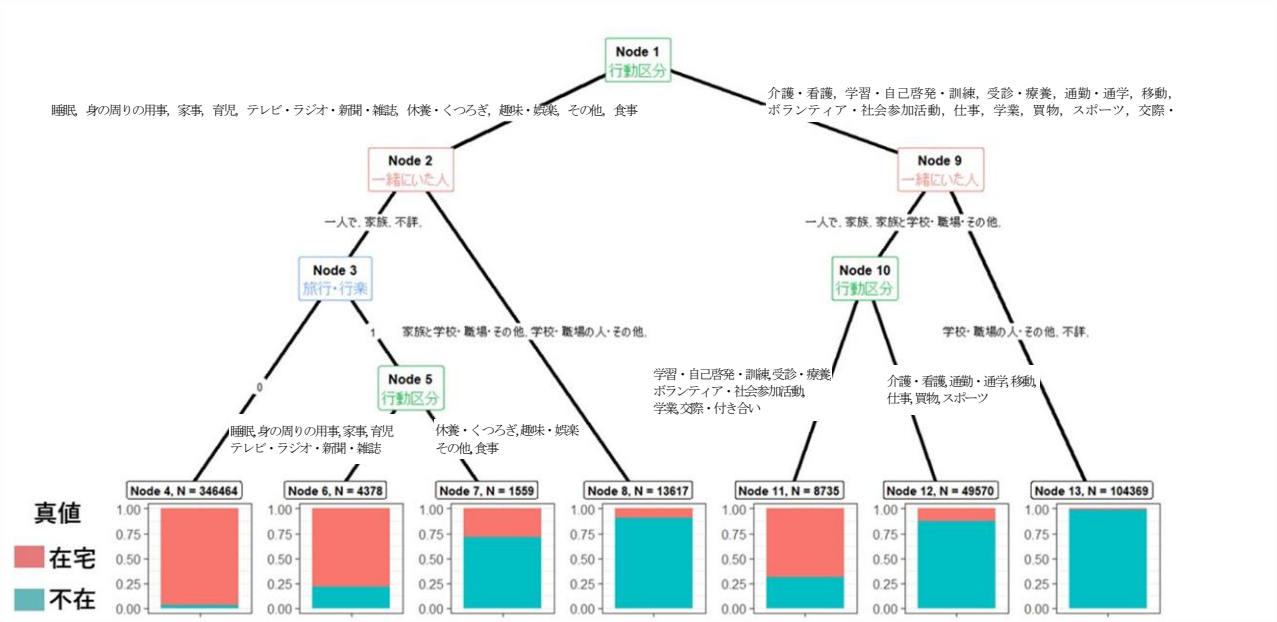



**図-2** 在宅・不在の決定木 (平日)

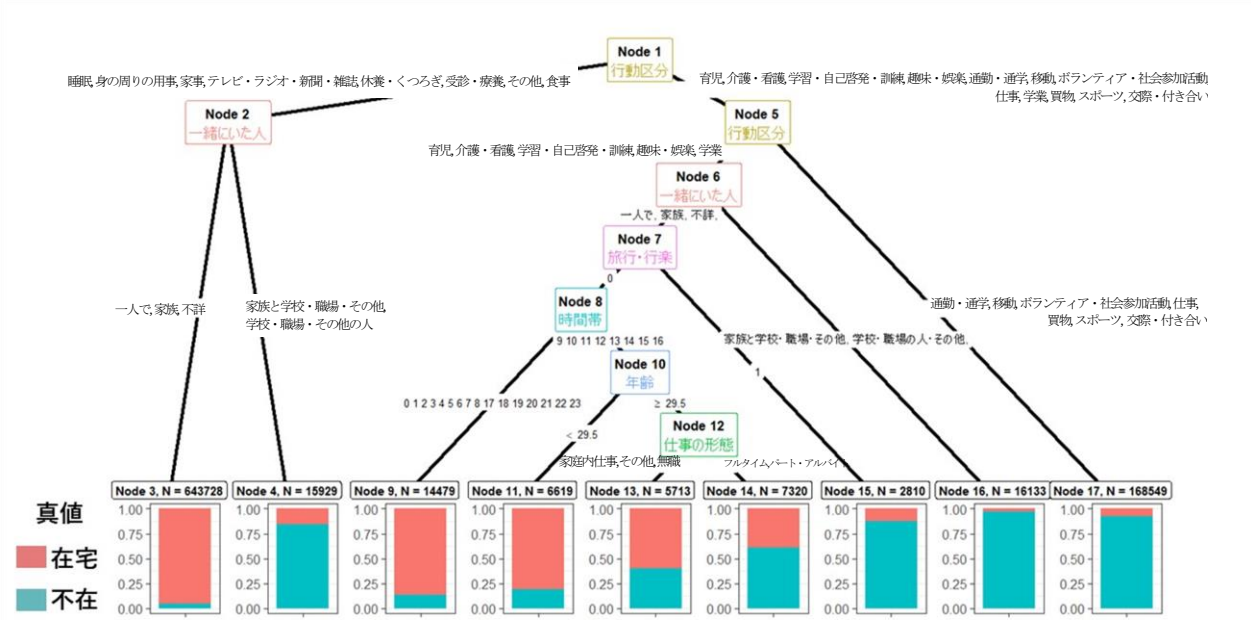



**図-3**　在宅・不在の決定木(休日)

いなかった，行動時間帯や年齢といった変数も上位に位置している．

### (2)　調査票 A の在宅・不在推定

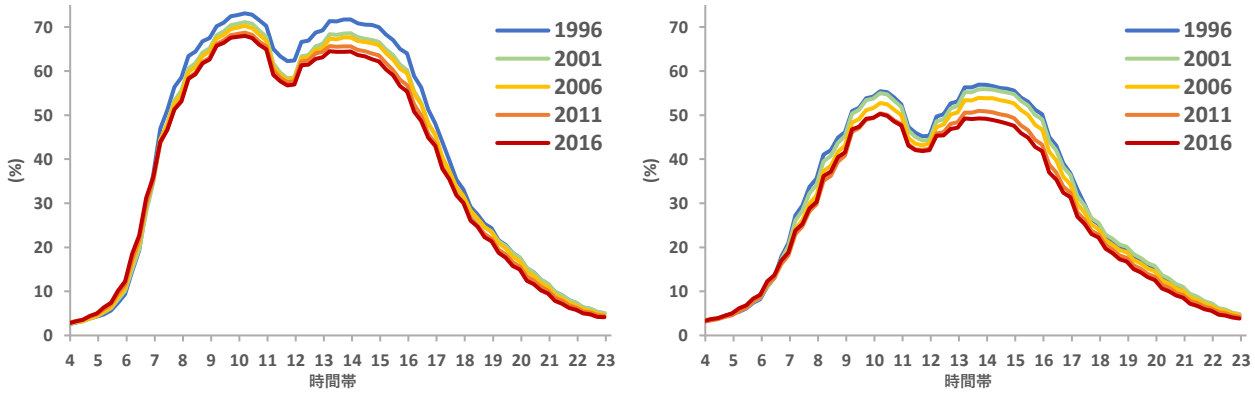
ここからは，3.(1)で作成したモデルを用いて複数年の調査票 A の行動場所の分類を行った．在宅・不在の予測には精度の高かった RF を用いた．

**図-4** は平休日別の調査票 A における個人不在率の経年推移を示している．平日では，明らかに減少の傾向が見受けられる．全ての年度で 11 時に最大値を取っており，1996 から順に 73.10%，71.10%，70.25%，68.72%，67.98%である．1996～2016 年での差を見ると，最大で 8.65% (16 時台) となり午後の不在率が特に減少していることが分かった．休日においても，明らかな減少傾向が見受けられ，1996～ 2011 年は 14 時台に，2016 年は 11 時台に最大値を取っており，1996 年から順に 56.92%，55.89%，53.93%，50.97%，50.28%である．1996～2016 年の差は 16 時台で最大となっており，その差は 8.37%であった．以上をまとめると，個人不在率は年々減少傾向にあり，特に午後 (夕方) の不在率が低下している．

**表-5**　重要度の上位6項目

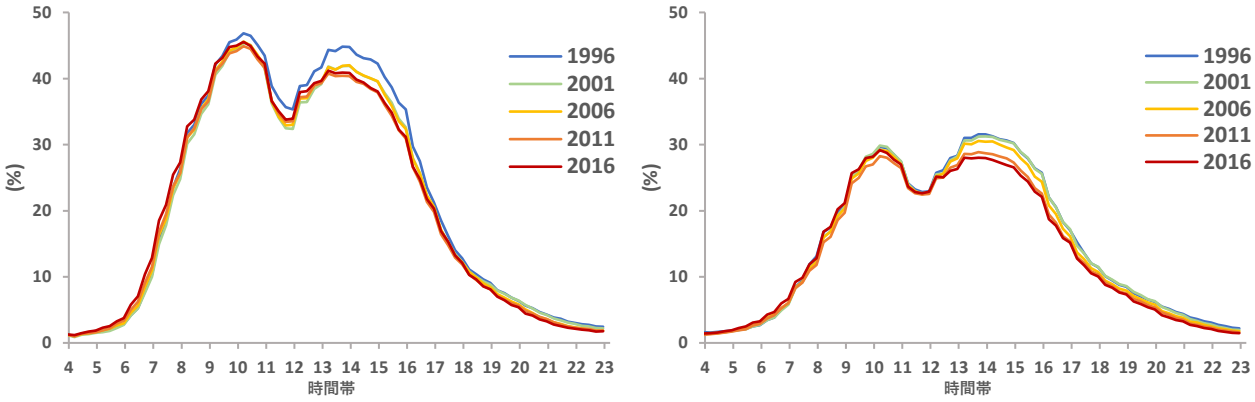
| 平日 | 値 | 休日 | 値 |
|---|---|---|---|
| 行動区分 | 0.493 | 行動区分 | 0.467 |
| 一緒にいた人 | 0.204 | 一緒にいた人 | 0.170 |
| 行動時間帯 | 0.106 | 年齢 | 0.098 |
| 年齢 | 0.066 | 行動時間帯 | 0.092 |
| 前行動移動ダミー | 0.028 | 前行動移動ダミー | 0.041 |
| 仕事の形態 | 0.025 | 10歳以上の世帯員数 | 0.034 |

**図-5** は平休日別の調査票 A における世帯不在率の経年推移を示している．平日では，1996～2001 年にかけて低下したものの，そこから 2016 年までには大きな変化は見られなかった．全年 11 時に最大値を取り，1996 年か




(a)　平日　　(b)　休日

**図-4**　調査票Aにおける個人不在率の経年推移




(a) 平日　　(b) 休日

**図-5** 調査票Aにおける世帯不在率の経年推移

ら順に，46.84%，45.37%，45.59%，44.89%，45.54%である．一方休日では，1996～2011 年は 14 時台に，2016 年は 11 時台に最大値を取っており，それぞれ順に 31.56%，31.24%，30.55%，28.87%，29.16%であった．1996~2016 年の差は最大 3.73%ポイント (15 時台) であり，午前中の世帯不在率には大きな変化はなかったものの，午後の時間帯に明らかな減少傾向が見られる結果となった．

## 4. 調査票 A を利用した分析

本章では，日本の代表的な交通調査である全国 PT と，在宅・不在を推定した調査票 A を用いて，両調査の不在率や不在時間を算出し比較を行う．また，全国 PT で入手していない世帯年収や住居の種類に関する分析も行う．

### (1) PT 調査との比較

**図-6** は調査票 A (2016) と全国 PT (2015) の個人不在率を示している．平休日共に調査票 A の方が明らかに高い値を示している．詳しくみると平日では 9～11 時，14～16 時で両調査の差が 10%を超えており最大で 11.15%ポイント (15:30) であった．休日では，8～18 時まで両調査の差

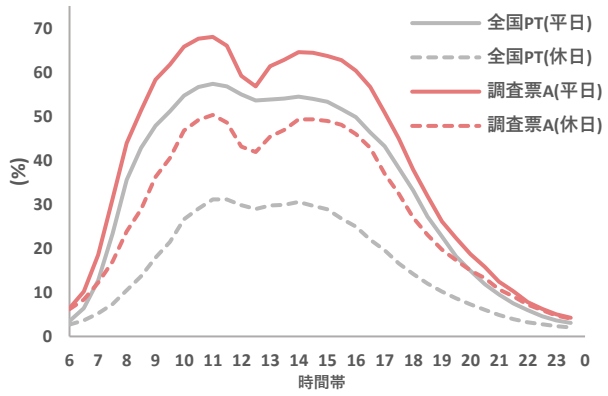
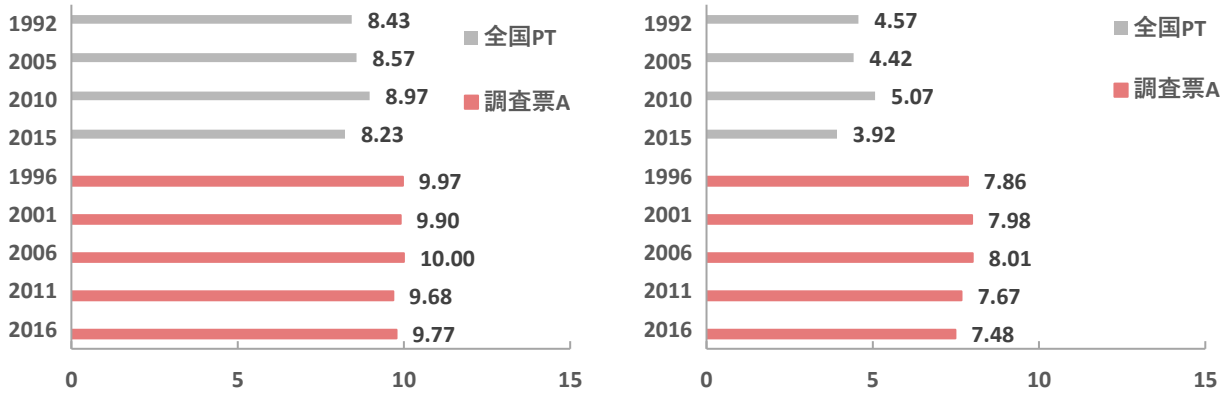



**図-6** 調査票Aと全国PT調査の個人不在率の比較

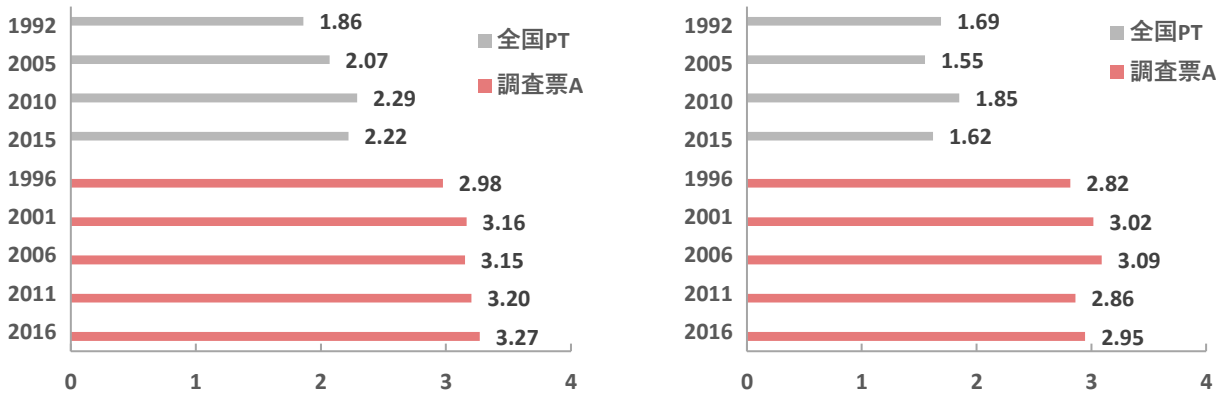
が 10%を超えており，最大で 21.28%ポイント (15:30) であった．このことから，既存研究[1]と同様に全国 PT のトリップ記入漏れ等の影響が示唆される．また特に休日においてその影響が大きいことが分かる．

次にトリップ抜け落ちの影響が大きい属性に着目する．わが国では全国 PT 調査の年齢別トリップ数の推移から若者のトリップ数の減少や高齢者のトリップの増加が強調されている．既存研究[1]ではその変化がトリップ記入




(a) 平日　　(b) 休日

**図-7** 20歳代における不在時間の経年推移 (時間 / 日)




(a) 平日　　(b) 休日

**図-8** 70歳以上における不在時間の経年推移(時間/日)

漏れの影響である可能性を示唆しており，ここで改めて検証を行うことは重要であると考える．また，本研究では不在率よりも比較が行いやすい不在時間に着目して分析を行う．比較を行うデータとして，1992，2005，2010，2015 の全国 PT 調査の結果を用い，この 4 回すべてで調査が行われた 41 都市のみを分析対象とした．

**図-7** に調査票 A と全国 PT 調査における 20 歳代の不在時間の経年推移を示す．まず**調査票 A について着目すると**休日のみ若干の減少傾向が見られる．2006 年に最大の 8.01 時間をとり，2016 年では 7.48 時間と 0.53 時間減少している．次に全国 PT では，平休日共に 2010 年に一時的な増加は見られるものの，全体としては減少傾向にある．同様に**図-8** は調査票 A と全国 PT 調査における 70 歳以上の不在時間の経年推移を示しており，両調査とも平日において増加傾向が見られる．調査票 A では 1996 年から順に 2.98，3.16，3.15，3.20，3.27 時間であり，20 年間で 0.29 時間増加している．一方全国 PT においても 1992 年から順に 1.86，2.07，2.29，2.22 と 23 年間で 0.36 時間増加している．　**表-6** は 20 歳代および 70 歳以上における調査票 A と全国 PT の不在時間の差であり，調査年の違いが 1 年であった場合を示している．調査票 A の方が不在時間が長く，特に 20 代の休日において差が大きいことが分かる．また，2005/2006～2016/2015 の 10 年間の差に着目すると 20 歳代では差が増加し 70 歳以上では減少している．このことから，トリップ抜け落ちが 20 歳代では増え，70 歳以上では減少している可能性があり，前述した既存研究[9]の考察を支持する結果となった．

**表-6**　調査票Aと全国PTの不在時間の差

| 調査票A(年次) | | 2006 | 2011 | 2016 | 10年間の差* |
|---|---|---|---|---|---|
| 全国PT(年次) | | 2005 | 2010 | 2015 | |
| 平日 | 20-29歳 | 1.43 | 0.90 | 1.82 | 0.38 |
| | 70-歳 | 1.08 | 0.88 | 0.76 | -0.32 |
| 休日 | 20-29歳 | 3.65 | 2.93 | 4.07 | 0.41 |
| | 70-歳 | 1.54 | 1.17 | 1.20 | -0.34 |

*(2016/2015年調査の差)−(2006/2005年調査の差) より算出
注) 調査年次差が1年以内の調査について，「調査票A-全国PT調査」の年齢別値(単位：時間/日)を計算．

### (2) 世帯年収・住居の種類による分析

**図-9** は世帯年収別の世帯不在時間の経年推移を示している．まず平日に着目すると年収 300 万円以下の世帯の不在率は減少傾向にある．1996 から順に，4.78，4.22，3.93，3.89，3.70 時間であり，1996 年からの 20 年間で 1.08 時間減少している．一方で世帯年収が 600～899 万円の世帯と 900 万円以上の世帯では，一転して世帯不在率が増加傾向にある．600～899 万円の世帯では 1996 年から順に，4.04，3.95，4.29，4.26，4.52 時間であり，1996 年から 20 年で 0.48 時間増加している．同様に 900 万円以上の世帯では，4.04，4.09，4.33．4.23，4.70 となっており 20 年間で 0.66 時間増加している．休日では，300 万円以下の世帯において平日と同様に不在時間の低下がみられる．1996 年から順に，3.84，3.55，3.25，3.19，3.04 であり 20 年間で 0.80 時間減少している．しかし，他の世帯年収クラスにおいて，不在時間の変化が 30 分を超えるような傾向は見られなかった．以上をまとめると，平休日共に 300 万円以下の世帯不在時間は減少傾向であり，500 万円以上の世帯では平日のみ増加傾向であることが分かった．

次に，住居の区分について分析を行う．ここでは，住居の区分が「持ち家」「民間の賃貸住宅」「公営の賃貸住宅の場合のみに絞った分析を行った．**図−10** に住居別世帯不在時間の経年推移を示す．まず，住居別にみると民・公営住宅が持ち家よりも世帯不在時間が長いことが分かる．2016 年において 2 つの住居の不在時間の差は平日では 2.53，休日では 1.83 時間であった．この最も大きい理由として，世帯人数の影響が考えられる．**表-7** は住居別の平均世帯人数の推移を示している．表より民・公営住宅では平均世帯人数が少なく，世帯不在の状態になりやすい単身世帯が多いことが考えられる．次に，住居ごとの経年推移をみる．持ち家世帯の不在時間は平休日共に，大きな傾向は見られない．一方，民・公営住宅世

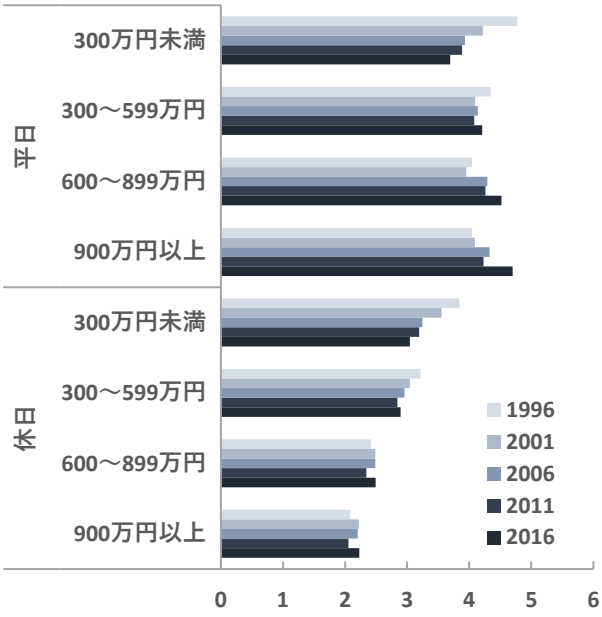



**図-9**　世帯年収別世帯不在時間の経年推移

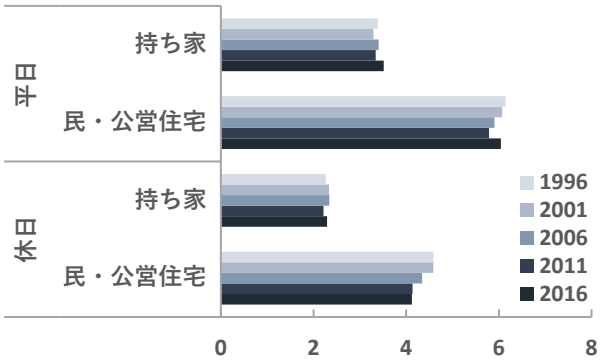



**図-10**　住居別世帯不在時間の経年推移

**表-7**　住居別平均世帯人数の推移(人/世帯)

| | 1996 | 2001 | 2006 | 2011 | 2016 |
|---|---|---|---|---|---|
| 持ち家 | 2.99 | 2.85 | 2.68 | 2.59 | 2.46 |
| 民・公営住宅 | 1.96 | 1.86 | 1.86 | 1.83 | 1.76 |

**表-8**　住居別の高齢者を含んだ世帯割合の推移(%)

| | 1996 | 2001 | 2006 | 2011 | 2016 |
|---|---|---|---|---|---|
| 持ち家 | 56.1 | 61.7 | 62.6 | 66.9 | 67.3 |
| 民・公営住宅 | 16.9 | 22.7 | 26.1 | 30.3 | 33.8 |

帯では，減少傾向が見られており，特に休日で減少幅が大きい．休日について詳しくみると 1996 年から順に．4.59，4.59，4.35，4.14，4.12 時間となっており，20 年で 0.47 時間減少している．この要因として，民・公営住宅に住む高齢者が急激に増えたことがあげられる．**表-8** は住居別の高齢者(60 歳以上)を含んだ世帯割合の推移を示している．民・公営住宅では 20 年間で 16.9%増加している．高齢者の不在時間の短さは**図-8** や先行研究 [4]より確認でき，その高齢者が増加したために世帯不在の状態が減少したと考えられる．一方で持ち家に関しては高齢者を含む世帯が増加傾向にあるものの，平均世帯人数の急激な減少等の世帯不在時間が長くなる要因と打ち消しあったため，経年推移の変化が生じなかった可能性がある．

## 5.　結論

本研究は，社会生活基本調査の調査票 B の行動場所データを利用し，機械学習を用いて調査票 A データの在宅・不在の分類を行った．また，その結果を用いて PT 調査との比較や PT 調査に含まれない項目についての分析を行った．具体的な成果を以下に述べる．

1) 社会生活基本調査の調査票 B データを用いて，調査票 A の在宅・不在を推定する方法を構築した．
2) 作成した決定木や RF の結果から，行動場所決定に影響を与える特徴量としては行動区分が最も大きく，次いで一緒にいた人，年齢，行動時間帯等が在宅・不在の分類に重要であることを示した．
3) 個人不在率は，平休日ともに減少傾向であり，特に午後(夕方)の不在率の低下幅が大きい．
4) 全国 PT 調査と個人不在率の差を比較した場合，平日では最大 11.15%ポイント，休日では 21.28%ポイントあり全国 PT 調査のトリップ記入漏れの影響を示唆した．
5) 全国 PT 調査と不在時間を比較した結果，20 歳代の若者世代でトリップ記入漏れが増加している可能性を示唆した．
6) 世帯年収別の分析から，世帯年収が 300 万円以下の世帯不在時間は減少しており，年収 500 万円以上の世帯は世帯不在時間が増加傾向にあることを示した．
7) 住居別の分析から，持ち家よりも民・公営住宅の世帯不在時間が長いことが明らかとなり，近年は高齢者の影響により減少傾向にあることを示した．

本研究は，教師あり学習を用いて調査票 A の在宅・不在を分類した．この手法の課題を以下に述べる．まず，調査票 A と調査票 B で行動区分が十分に対応できていない場合がある．教師データである調査票 B は行動の詳細が把握可能で 96 区分(2016 年の場合)に分類されており，本研究では既存研究 [1]の分類ルールを用いて調査票 A の 20 区分に対応させている．例えば，調査票 B の場合「送迎」の行動区分は 20 区分に対応させると「移動」としている．しかし，調査票 A では通勤・通学者の送迎の場合は「家事」の区分に分類される[注 4]．このことを踏まえると，調査票 A「家事」の場合の不在行動を，在宅側にに推定している可能性が考えられる．

今後の展望として，第一に調査票Aの在宅・不在推定モデルの改良があげられる．本研究では 2 種類の教師あり学習を用いたが，XGBoost (勾配ブースティング) 等の発展的なモデルを適用することでより精度の高い予測が期待できる．第二に，推定した調査票Aを利用した分析がある．具体的には，都市・地域別の外出状況の比較や世帯年収・住居別等の特性を踏まえた新たな世帯不在時間モデルの更新があげられる．

**謝辞**：本研究は JSPS 科研費 19K21997 の支援を受けた成果の一部です．また総務省統計局から社会生活基本調査について，国土交通省から全国 PT 調査について調査票情報の提供を受け独自集計した成果を含みます．深く謝意を表します．

## 付録　カテゴリ変数の項目

**表-7**　カテゴリ変数の項目詳細

| 項目名 (種類数) | 詳細 |
|---|---|
| 行動区分 (20) | 睡眠，身の回りの用事，家事，育児，介護・看護，テレビ・ラジオ・新聞・雑誌，休養・くつろぎ，学習・自己啓発・訓練 (学業以外)，趣味・娯楽，受診・療養，その他，食事，通勤・通学，移動，ボランティア活動・社会参加活動，仕事，学業，買い物，スポーツ，交際・付き合い |
| 一緒にいた人 (5) | 一人で，家族と一緒，学校・職場・その他の人と一緒，家族と学校・職場・その他の人と一緒，不詳 |
| 行動時間帯 (24) | 0時～23時までの24種類 |
| 行動分 (4) | 0，15，30，45 |
| 曜日 (7) | 月，火，水，木，金，土，日 |
| 仕事の形態 (5) | フルタイム，家庭内での仕事，パート・アルバイト，無職，その他 |

**NOTES**

注1) 総務省統計局: 平成 28 年社会生活基本調査, 調査からわかること.
https://www.stat.go.jp/data/shakai/2016/wakaru/index.html

注2) 総務省統計局：高等学校における「情報II」のためのデータサイエンス・データ解析入門，第 3 章.
https://www.stat.go.jp/teacher/dl/pdf/c4learn/materials/fourth/dai3.pdf

注3) 総務省統計局: 令和 3 年社会生活基本調査, 生活時間についての回答例 (調査票 A).
https://www.stat.go.jp/data/shakai/2021/pdf/kaitoua.pdf

注4) 総務省統計局：令和 3 年度社会生活基本調査，生活時間についての記入ポイント
https://www.stat.go.jp/data/shakai/2021/pdf/pointa.pdf

**REFERENCES**

1) 荒尾俊介，円山琢也：社会生活基本調査を利用した全国 PT 調査の精度検証，第 65 回土木計画学研究発表会 (春大会), 2022. [Arao, S. and Maruyama, T.: Validation of nationwide person trip survey using survey on time use and leisure activities, *Proceedings of the 65rd Infrastructure Planning Conference*, 2022.]

2) 北村隆一: 交通行動調査の展開, in: 北村隆一, 森川高行, (編), 交通行動の分析とモデリング, 技報堂出版, pp.53-68, 2002. [Kitamura, R. Advances in travel behavior surveys, in Kitamura, R. and Morikawa, T. (eds.) *Modeling Travel Behavior*, Gihodo-shuppan, pp. 53-68, 2002.]

3) 深堀達也, 円山琢也: 社会生活基本調査による個人・世帯不在率の経年変化: 交通調査のトリップ記入漏れ分析への示唆，土木学会論文集 D3,Vol.78, No.3, pp.93-104, 2022. [Fukahori, T. and Maruyama, T.: Inter-temporal changes in household- and individual-based out-of-home rates inferred from surveys on time use and leisure activities: implications for trip-misreporting analysis, *Journal of Japan Society of Civil Engineers, Ser. D3*, Vol.78, No.3, pp.93-104, 2022.]

4) 高橋瑠衣, 川野倫輝, 佐藤嘉洋, 円山琢也: PT 調査に基づく世帯単位の時間帯別不在率の経年比較分析, 土木学会論文集 D3,Vol.74, No.4, pp. 387-397, 2018. [Takahashi, R., Kawano, T., Sato, Y., and Maruyama, T.: Temporal analysis of household with every member out-of-home using person trip surveys, *Journal of Japan Society of Civil Engineers, Ser. D3*, Vol.74, No. 4, pp. 387-397, 2018.]

5) Kikuchi,K. and Maruyama, T.: Spatiotemporal change in duration of households with every member out-of-home: A case in Kumamoto, Japan, *International Journal of Urban Sciences*. in press. 2022.

6) 酒井幸市：改訂版 ディジタル画像処理の基礎と応用 -基本概念から顔画像認識まで-，CQ 出版，pp.36-38, 2007.

7) 渡邉萌, 佐藤嘉洋, 円山琢也: 集団離散選択モデルと決定木を利用した益城町仮設住宅入居世帯の住まいの意向分析, 土木学会論文集 D3 (土木計画学) , vol.74, No.5, pp. I_201-I_208, 2018. [Watanabe, H., Sato, Y. and Maruyama, T.: Residential preferences of households in mashiki temporary housing using group-based discrete choice model and decision tree, *Journal of Japan Society of Civil Engineers, Ser. D3*, Vol. 74, No. 5, pp. I_201-I_208, 2018.]

8) Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. : Classification and Regression Trees, CRC Press, 1984.

9) 花岡創: 森林遺伝育種のデータ解析方法（実践編 9）ランダムフォレスト, 森林遺伝育種, Vol.11, No.3, pp147-151, 2022. [Hanaoka, S: Shinrin iden ikushu no deta kaiseki houhou jissenhen 9 randomforest (in Japanese), *Forest Genetics and Tree Breeding*, Vol.11, No.3, pp147-151, 2022.]

10) R サポーターズ: パーフェクト R, データ分析, 技術評論社, pp. 304-311, 2017.

11) Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI'95)*,Vol.2 , pp. 1137-1143,1995.

**(Received July 1, 2022)**
**(Accepted November 1, 2022)**

# DETECTING IN-HOME AND OUT-OF-HOME OF THE DATA FOR SURVEY ON TIME USE AND LEISURE ACTIVITIES (QUESTIONNAIRE A) USING MACHINE LEARNING

Shunsuke ARAO and Takuya MARUYAMA

.