

# 職業属性を考慮した人流データの居住地及び通勤通学地推定

## Identification of the Significant Places from Trajectory Data

### Considering Occupation

学籍番号 47186760  
氏 名 小林 稜介 (Kobayashi, Ryosuke)  
指導教員 柴崎 亮介 教授

#### 1. 背景と目的

近年、スマートフォンアプリの普及により、アプリ開発キット SDK を介して多くの事業者がユーザーの位置情報を取得できるようになった。また、高精度な測位を可能とする準天頂衛星システムの本格運用が開始され、人々の移動に関わる時空間データ活用への期待が高まっている。同データの活用は多岐にわたるが、中でもアドテクノロジーや交通計画において、大いに有益である人々の居住地及び通勤通学地の推定は最も重要な活用法の一つである。

既往研究 Isaacman et al. (2011) は、基地局通信履歴を用いて滞留時間を元に居住地・通勤地を推定している。しかし、ユーザーが通勤者以外である場合の考慮に欠ける。また、Sari Aslam et al. (2018) は交通利用履歴データを用いて、居住地・通勤地を推定している。長期データから定期的な行動を元に通勤者を抽出しているが、長期データの利用にはデータ取得から推定までに大きなラグが生じ、デジタル広告などの即時性を求める場面では損失となりうる。

そこで本研究では、1 日という短期間の移動履歴を元に、職業属性を考慮した居住地及び通勤通学地の推定手法を構築することを目的とする。

#### 2. データセット

##### 2.1 データセット一覧

用いるデータセットは以下の通りである。

- 1) 空間配分版 パーソントリップ調査 (東京大学 空間情報科学研究センター, 2008)
- 2) 人流データ (株式会社 Agoop, 2016)

##### 2.2 PT データと人流データの概要

空間配分版パーソントリップ (PT) データは、東京都市圏交通計画協議会によって行われた 10 年に 1 度の大規模な PT 調査の結果であり、人々のデモグラフィック属性やトリップ目的が付随する 1 日間の移動履歴データである。一方、人流データは、株式会社 Agoop が提供するスマートフォンアプリから、ユーザーの同意を得て収集したポイント型流動人口データである。表 1 に両データの概要を示す。

表-1 両データの概要

項目	PT データ	人流データ
対象エリア	関東地方	関東地方
対象期間	1 日間(平日)	1 ヶ月間(平日)
総ユーザー数	587,434	(延べ)2,272,810
職業属性ラベル	あり	なし

##### 2.3 PT データと人流データの特徴

空間配分版 PT データは東京大学 CSIS が独自に加工したもので、従来は町丁目より粗いゾーンの代表点で再現していたトリ

ップの起終点が、住宅地図を元にした生起確率によりゾーン内部に再配分されている。

一方、人流データには 2 つの特徴があり、各ユーザー ID (UID) は午前 0 時にリセットされるため、UID を元に日を跨いで分析することができない。また、測位が低頻度であり、スパースなデータである (アンドロイド端末の場合 30 分/回)。

### 3. 推定手法

#### 3.1 手法概要

全 UID に対して、時間帯及び滞在時間を元に居住地を推定する。次に、転移学習として PT データによる学習済み職業属性分類モデルを人流データに適用し、各属性に分類する。そして、NHK 放送文化研究所 (2015) を参考に、就業者・学生に分類された UID に対して、居住地以外に最も長く滞在した地点 (閾値 2 時間以上滞在) を通勤通学地として推定する。

#### 3.2 居住地推定

PT データを用いて 1 時間毎の訪問地を居住地とみなした場合、正解率は 0 時から 6 時において 96% 以上であった。そこで、本研究では 0 時から 6 時における最頻出点を居住地として推定する。

#### 3.3 転移学習

転移学習とはドメインを跨いだ機械学習手法であり、あるドメインのデータで構築した学習済みモデルを別ドメインのデータに適用させることができる。本研究では元ドメイン( $S$ )の PT データに正解ラベル (職業属性情報) があり、目標ドメイン( $T$ )の人流データに正解ラベルがないため、特に Transductive Transfer Learning と呼ばれる。ここで、転移学習では両ドメイン間にどのような類似性があるか仮定する必要がある。

本研究では Nishimura et al. (2014) を参考に、人流データのサンプル数が大きく大数の法則より  $P[Y^{(S)}|X^{(S)}]$  と  $P[Y^{(T)}|X^{(T)}]$  の間に一定の類似性が見られると仮定し転移学習を行う。 $P[Y^{(D)}|X^{(D)}]$  は、データ集合を  $X$ 、ラベル空間を  $Y = \{y_1, \dots, y_n\}$  としたとき、ドメイン  $D$  において  $X$  が観測された際にラベルが  $Y$  である確率分布を表す。

#### 3.4 職業分類モデル構築

適切な通勤通学地推定のため、まず職業属性推定を行う。そこで、各 UID を就業者・学生・その他の計 3 クラスに分類する職業分類モデルを構築する。なお、PT データにはクラス間に不均衡性があるため、事前にアンダーサンプリングを行う。

分類手法として、Random Forest (RF)、勾配ブースティングアルゴリズムに基づいたフレームワークである LightGBM、3 層のニューラルネットワーク (NN)、時系列データ分析に優れた Gated Recurrent Unit (GRU) の計 4 手法を用いて比較を行う。

GRU は、再帰結合を持つ NN である Recurrent Neural Network (RNN) の拡張モデルであり、RNN における勾配消失問題をゲート機構により改善している。短期及び長期の時系列データを扱うことができ、株価予測や自然言語処理など広く用いられる。また、同じく RNN の拡張モデルである Long Short-Term Memory (LSTM) と同等の性能を有し、かつ簡易的な構造によって計算時間を削減することができる。

GRU におけるパラメータは、2 層の GRU、隠れ層のユニット数を 32、最適化手法を Adam (初期学習率 0.01)、バッチサイズを 32 としている。出力層の活性化関数には Softmax 関数を用い、各 UID で最も

高い値をとったクラスを職業属性として割り当てる．以上のパラメータ設定の元，50回の学習を行うことでモデルを構築する．

特徴量は Kobayashi et al. (2019) を参考に，1 日間における 15 分毎の推定居住地からの距離を UID 毎に正規化 (0~1) した，97 タイムステップのシーケンス情報を用いる．厳密には正確な緯度経度ではないという加工済み PT データの特性を考慮し，最小限の特徴量のみでモデルを構築する．

検証には Hold-out 法を用い，*Macro-F1* 値 (式(1)) を評価指標とする．*Macro-F1* 値は，各クラス  $i \in \{c1, c2, c3\}$  の *Precision* (適合率) と *Recall* (再現率) の調和平均である *F1* 値の平均値を表す．

*Macro-F1-Measure*

$$= \frac{1}{3} \sum_{i \in \{c1, c2, c3\}} \frac{2Recall_i \cdot Precision_i}{Recall_i + Precision_i} \quad (1)$$

### 3.5 職業分類モデル検証結果

表 2 に，PT テストデータに対するモデルの検証結果を示す．各手法を比較すると，*Macro-F1* 値から GRU の分類精度が最も高いことがわかった．これは GRU がシーケンスを入力として扱うことができる影響と考えられる．そこで，人流データへの適用時は GRU によるモデルを用いる．

また，参考として SHAP 指標を用いた LightGBM における特徴量重要度上位 10 個を図 1 に示す．9 時前後や 17 時前後といった通勤通学時間帯や帰宅時間帯における特徴量が職業分類に大きく影響していることがわかる．

表-2 各手法の精度検証結果 (\*:Macro)

Method	Acc.	Recall*	Precision*	F1*
RF	0.78	0.78	0.75	0.76
LightGBM	0.79	0.79	0.76	0.77
NN	0.76	0.75	0.72	0.73
GRU	0.81	0.81	0.78	0.79

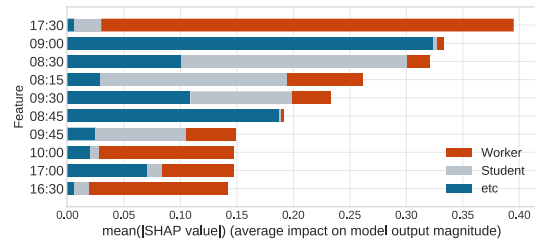


図-1 特徴量重要度

## 4. 人流データを用いた推定と検証

### 4.1 居住地推定と検証

3.2 で述べた通りに居住地を推定し，国勢調査による市区町村別の夜間人口統計データ (総務省統計局, 2015) によって推定結果を検証する．評価指標には相関係数を用いる．なお，国勢調査において夜間人口とは 2015 年調査時に各対象地域に常住している人口を指している．結果として，推定夜間ユーザー数 (居住地数) と夜間人口の相関係数は 0.982 となり，居住地推定は高い精度で行えたと言える．

### 4.2 モデルの適用結果

推定した居住地を元に特徴量を算出し，転移学習として学習済みモデルに入力する．なお，PT データの形式に合わせるため，事前に人流データに対して密度準拠型クラスタリングアルゴリズムである DBSCAN による緯度経度の整形などの前処理を行う．

モデルの適用後，就業者に分類されたのは 68.5%，学生は 11.2%，その他は 20.3% となった．本研究の対象地域 (関東) と異なるが参考として，全国を対象地域とした NHK 放送文化研究所 (2015) の調査によると，各職業属性の割合は就業者 55%，学生 12%，その他 32% であると明らかになっている．この割合の差の要因は，モデルの分類精度による影響，もしくは人流データ特有のバイアス (エリアや年齢層) による属性割合の偏りが考えられる．

### 4.3 通勤通学地推定と検証

4.2 で就業者・学生に分類された UID に対して、通勤通学地の推定を 3.1 の通りに行う。そして、市区町村別の昼間人口統計データ（総務省統計局，2015）と、より詳細な町域別（郵便番号単位）の昼間人口統計データ（日本統計センター，2015）によって、推定結果を検証する。昼間人口とは、4.1 で用いた夜間人口から通勤通学による流出・流入人口を足し引きした人口を指す。

図 2 左が市区町村別の相関図である。港区などのオフィスが多い都心エリアにおいては昼間ユーザー数が多く推定されているが、相関係数は 0.892 と強い正の相関を示した。また、職業属性推定を行わない場合は 0.867 であったため、職業属性分類によって精度が向上することがわかった。

次に、町域別に検証すると相関図は図 2 右の通りになり、相関係数は 0.899 であった。ここでは、推定昼間ユーザー数の上下 0.01% の外れ値を除いている（除去前の相関係数：0.777）。市区町村別と同様、町域別でも強い正の相関を示す結果となった。

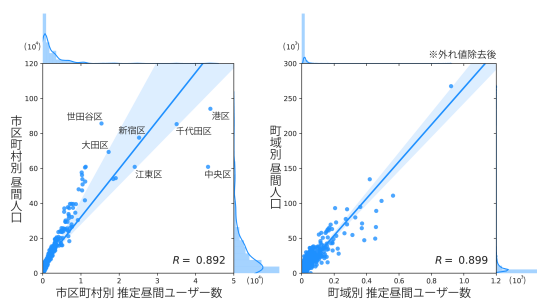


図-2 昼間人口との相関図（左：市区町村別、右：町域別）

### 5. 推定結果の活用展望

本研究による推定結果は、交通需要モデル構築をはじめとした多くの活用が考えられる。例えば、図 3 のように推定した通勤地の滞在時間から勤務時間の空間分布を把握することができる。これにより、行政機関

が人々の労働実態を低コストかつ即時的に把握することが可能となる。

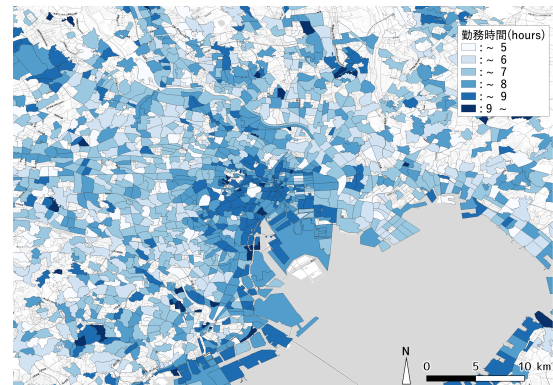


図-3 勤務時間の空間分布（東京都心部、町域別）

### 6. 結論

本研究では、1 日間の移動履歴に基づく学習済み職業属性分類モデルの転移学習により、職業属性を考慮した居住地及び通勤通学地推定手法を構築した。検証では、市区町村別に加え、町域別人口統計データとも強い相関を示し、ミクロ寄りのスケールでも推定が可能であることがわかった。

一方で、人流データにおいて一部のエリアや年齢層の偏りがあると考えられるため、これらの補正を行いロバストな職業属性分類モデルを構築することが今後の課題として挙げられる。

### 参考文献

- NHK 放送文化研究所, 2015. 国民生活時間調査報告書, [https://www.nhk.or.jp/bunken/research/yoron/pdf/20160217\\_1.pdf](https://www.nhk.or.jp/bunken/research/yoron/pdf/20160217_1.pdf).
- Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J. and Varshavsky, A., 2011. Identifying Important Places in People's Lives from Cellular Network Data, *Pervasive Computing*, pp.133-151.
- Kobayashi, R., Miyazawa, S., Akiyama, Y. and Shibasaki, R., 2019. Identification of the Homes, Offices, and Schools from Long-Interval Mobile Phone Big Data Using Mobility Pattern Clustering, *AGILE Conference on Geo-information Science*, #82.
- Nishimura, T., Akiyama, Y., Shibasaki, R. and Sekimoto, Y., 2014. Study of Estimate Human Demographic Attributes Using Person Flow Datasets and Apply It for GPS Log Data, *The International Symposium on City Planning 2014*, SS03, S03-12.
- Sari Aslam, N., Cheng, T. and Cheshire, J., 2018. A high-precision heuristic model to detect home and work locations from smart card data, *Geo-spatial Information Science*, Vol 22(1).