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

2019 年度
修　士　論　文

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

2020 年 1 月 20 日提出
指導教員 柴崎 亮介 教授

小林 稜介
Kobayashi, Ryosuke

目次

第 1 章 序論

1.1 研究背景

日本では，2007 年 4 月以降，携帯電話端末への GPS 機能搭載が義務化され[1]，携帯電話端末の位置情報に基づく人々の移動履歴「人流データ」の利活用が加速した．また現在は，2000 年代後半のスマートフォン普及，そしてそれに伴うスマートフォンアプリの登場により，Software Development Kit（SDK）を介して多くの事業者がユーザーの位置情報を収集している．これまで，人流データは電気通信事業者などの限られた事業者のみが得られるものであったが，大きく構図が変わることとなった．これらの動きと同時に，収集したデータを十分な匿名加工を施した上でデータ取引市場に開放し，データそのものや付加価値を与えたデータを他社に提供する取り組みも活発化している[2]．これまで自社内での活用に留められることが多かったマーケティング価値のある人流データであるが，近年では自社だけでなく他社へも開放されつつある．無論ここで，議題に挙がるのはプライバシーの問題である．人流データは個人の特定に繋がる情報を多く含むデータであり，人流/位置情報データの活用の是非について，これまでプライバシー保護の観点から国内外問わず議論が盛んにされてきた．例えば，米電気通信事業者がユーザーの位置情報に関するデータを売却し，大きな問題[3]となった．しかし，データ匿名加工技術は年々向上しており[4]，匿名性を高めることで，データ取引市場における人流データの流通もより一層勢いを増す可能性が高い．国内における人流データの代表的なものとしては，株式会社 NTT ドコモによるモバイル空間統計[5]がよく知られるが，近年はブログウォッチャー[6]や Agoop[7] などの企業が SDK を介したスマートフォンアプリユーザーの人流データ収集及びマーケティング活用を進めている．

また，これら人流データ活用の流れと同時に，高精度な測位を可能とする準天頂衛星システム（QZSS）の運用が始まった．2010 年に初号機が打ち上げられ，2018 年には 4 機体制の本格運用が開始されている[8]．同衛星は日本上空に長く留まる準天頂軌道を回り，GPS と同じ周波数信号を使うことによって，GPS のみの場合より安定した測位や高精度な位置情報取得を可能とする．特に，従来は測位精度が不安定であった高層ビルが建ち並ぶ都市エリアや山間エリアでの測位精度向上が期待されている[9]．同様に独自の衛星システムを運用している中国では，測位における位置情報の誤差が 5 m まで向上したとの報告もされている[10]．そのため，位置誤差に囚われない，よりミクロなスケールでの人流データ活用が今後可能となる．

このように，Information and Communication Technology（ICT）の発展に伴い，人流データ活用への期待が高まっている．同データは，商業施設の来訪者分析といったマーケティング目的だけでなく，観光客分析による地域活性化（沖縄県，2012）や帰宅困難者数調査（東京都，2011）など観光・防災分野にも活用されている[11]．

その他，活用方法は多岐にわたるが，中でもアドテクノロジーや都市・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．携帯端末から得られる位置情報データから居住地及び通勤通学地を明らかにすることで，低コストで即時性のある交通利用・労働実態を把握することや，効果的な広告配信が可能になる．例えば，人々のライフスタイルやエリア情報に基づくフットプリント方式のデジタル広告配信，鉄道駅のキャパシティを考慮した超高層マンションの開発計画及び規制検討が可能である．また，近年取り組まれている働き方改革の観点では，推定した通勤地に何時間滞在するかによって，対象エリアの企業群の労働時間を把握することも可能である．

これらの推定結果の有益性から，同推定には多くの既往研究が存在する．しかし，職業属性の分類を行わない，もしくは不十分なまま通勤・通学地を推定しているものが多い．そのため，職業属性を考慮した上で通勤通学地の推定を行う必要がある．

1.2 研究目的

本研究に関連する既往研究である Isaacman et al.（2011）[12]は，基地局通信履歴を用いて滞留時間などを元に居住地・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．NHK 放送文化研究所（2015）[47]によると，平日であっても仕事をする有職者の割合は 88% であり，平日に仕事を行わない人々に対して，通勤地推定を行うことは不適切である．また，Sari Aslam et al.（2018）[13]は交通機関の利用履歴データを用いて居住地・通勤推定を行っている．通勤者の抽出には定期的な行動，同一の交通ネットワークを利用しているかどうかにより判断している．しかし，このような定期・規則性を見出すために長期間データを用いる場合は，長期データの取得時と推定時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．森川ら（2015）[14]によると居住地・通勤地を定期・規則性から，安定した精度で推定するには 3 週間から 4 週間以上のデータが必要であると言われている．このように，短期の移動履歴を用いて職業属性推定を十分に考慮した上で居住地及び通勤通学地推定を行っている研究は見られない．

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，人々の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．

1.3 用語の定義

本研究で用いる用語の定義を以下に示す．ここでは，参考として総務省統計局（2010）[15]を用いる．

場所：
- 居住地：常住する場所
- 通勤地：就業者が勤務する場所．複数の勤務地を持つ就業者の場合，最も長く滞在する場所
- 通学地：学生が通学する場所

職業属性：
- 就業者：収入を伴う仕事を行っている人のうち，居住地以外の通勤地を持つ者
- 学生：小学生/中学生/高校生/大学生/大学院生/専修学校生/各種学校生のうち，一定の通学地を持つ者
- その他：上記以外の職業属性の者（ただし休日の就業者や学生も含む）

他：
- トリップ：人が何か目的をもって，ある地点から別の地点まで移る移動のこと

本研究における職業属性とは，あくまでもデータ取得した当日における職業属性を指し，人々の普遍的な職業属性を示すものではない．例えば前述したように，その他属性には休日の就業者や学生も含めている．

第 2 章 既往研究

本章では，居住地及び通勤通学地推定に関する既往研究について述べる．近年は，大規模データ処理技術の発展によってデータ収集が容易となったことで，様々なデータを用いて居住地や通勤地推定が行われている．

例えば，Twitter におけるジオタグ付きツイートを用いて居住地推定を行った Mahmud et al.（2012）[16]がある．ジオタグ付きツイートから得られた都市をユーザーの居住地都市（正解ラベル）とみなし，ハッシュタグやツイート文，ツイート文章中の地名情報などを用い，分類器によって居住地がある都市の予測を行っている．パブリックな投稿の場合，ツイートをはじめとした SNS 内のデータは入手が容易であるが，中でもジオタグ付き投稿の数は非常に少ない．これは要因の一つとして，時代とともにユーザーのインターネットリテラシーが高まり[17]，個人の特定に繋がりかねないジオタグ付き投稿は敬遠される傾向があることが考えられる．同研究によると全投稿（ツイート）のうち，ジオタグ付き投稿は 1% にも満たないと明らかにされている．また，Twitter 社は利用しているユーザー数が少ないことを理由に，正確なジオタグ付与の機能を廃止する意向を示しており（2019 年 6 月時点）[18]，今後は Twitter などの SNS から得られるジオタグ付きデータに制限がかかる可能性がある．

次に，本研究で用いる人流データに類似しているものとして，1.2 研究目的でも述べた基地局通信情報（CDR データ）を用いた研究，交通機関の利用履歴データを用いた研究を以下に挙げる．

CDR データを用いた既往研究 Isaacman et al.（2011）[12]は，Important Places Algorithms と Home and Work Algorithms による推定を提案している．Important Places Algorithms には，クラスタリングと回帰分析によって構成された 2 つのフローがある．まず，ユーザーの通信記録がある通信基地局を空間的にクラスター分割する．そして，回帰モデル（ボランティアから得たデータを元に構築）で各クラスターにスコアを付与し，重要度を明らかにすることで Important Clusters C_i （$i = 1, \cdots\cdots, n$）を抽出する．

次に Home and Work Algorithms にて時間帯に着目し，Important Clusters から居住地と勤務地を抽出する．Home and Work Algorithms は以下の通りである．

1. 「平日の午後 7 時から午前 7 時まで及び週末」を居住地時間帯,「平日の午後 1 時から午後 5 時まで」を勤務時間帯とする
2. 居住地時間帯に最頻出であるクラスターを居住地と推定する
3. a もしくは b の指標から勤務地推定を行う
   a. 勤務時間帯に最頻出であるクラスター C_l を勤務地とする
   b. C_i における居住地時間帯の割合 p を式 (1) で算出し，最も割合 p が低いクラスター C_l を勤務地とする

$$p = \frac{N_{C_i}}{\sum_{k=1}^{n} N_{C_k}} \tag{1}$$

＊N_{C_i}: クラスターC_iにおける居住地時間帯イベント数

最後に，ボランティアから得た正解ラベル付き CDR データにモデルを適用し，検証を行っている．同アルゴリズムは GPS データにも適用自体は可能であるが，職業属性推定をしておらず，データ内に通勤者以外の職業属性ユーザーが含まれる場合に関する考慮が十分になされていない．

また，CDR データ以外にも他のデータセットを用いた居住地・通勤地推定がある．Sari Aslam et al.（2018）[13]は．ロンドンのバスや鉄道といった交通機関で利用されるスマートカ

ードのデータを用いて居住地・通勤地の推定を行っている．データには通勤以外を目的とするユーザーのデータも含まれるため，長期間データを用いて固定されたユーザー ID（UID）情報を元に，同一の交通ネットワークを利用したかどうかによって通勤者を抽出している．そして，通勤者に対し訪問回数や滞在時間が閾値を超えたかを元に，居住地や通勤地推定を行っている．同様の規則性から通勤者を抽出するアプローチは，携帯電話端末の位置情報データに対しても行われることが多い．

しかし，研究目的にて述べた通り，長期データを利用する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．加えて本研究の人流データは同一の UID を持たないため，同アプローチは困難である．

このようにいずれの既往研究においても，職業属性推定が不足している，もしくは職業属性推定が長期データに依存するものであり，職業属性を十分に考慮したものが少ないという現状がある．

その他，筆者による研究 Kobayashi et al.（2019）[19]では，本研究で用いる人流データと同様のデータを用いて移動パターンのクラスタリングにより，居住地及び通勤通学地推定を行っている．また，スマートフォンアプリから得られるスパースな人流データを用いること，そして，従来は考慮されていなかった夜間労働者も推定対象であることに新規性を見出している．

居住地の推定は 1 日における滞在時間を元に行い，通勤通学地の推定は移動パターンと曜日割合に基づいて行っている．移動パターンとは，例えば昼間労働者や学生の場合，主に朝に通勤通学し，日中はオフィス・学校に滞在し夕方以降帰宅する，というものである．そのため，移動パターンを元にすると昼間労働者や学生の場合，自宅からの距離を縦軸とした 1 日間の時系列グラフを描くと山型の形状となる．一方で，夜間労働者は真逆の谷型の形状となる．これらの考えに基づき，K-means による移動パターンのクラスタリング（初期化：K-means++）を行って通勤通学者と思われるユーザーを抽出し，通勤通学地を推定している．

土日も含む 1 ヶ月のデータを対象にクラスタリングを行い，結果として図 1 の 20 クラスターを得ている．縦軸はユーザー毎に正規化した推定居住地からの距離を表し，横軸は時系列である．青色線がランダムに抽出されたサンプルであり，黒色線が中央値のグラフを表している．平日は通勤通学する人が多いことを念頭に置いて，各クラスターにおけるデータの平日と土日の比較を見てみると表 1 のようになっており，クラスター 1 から 10 の平日割合が多いことがわかる．これら移動パターンと曜日割合を元に，クラスター 1 から 10 を学生・昼間労働者，クラスター 20 を夜間労働者としている．なお，夜間労働者は飲食などのサービス業に従事する人が多い傾向があるため，クラスター 20 においては平日割合に偏りが見られないとしている．そして，得られたクラスター 1 から 10 と 20 のユーザーから居住地の次に長く滞在した場所を通勤通学地として推定している．

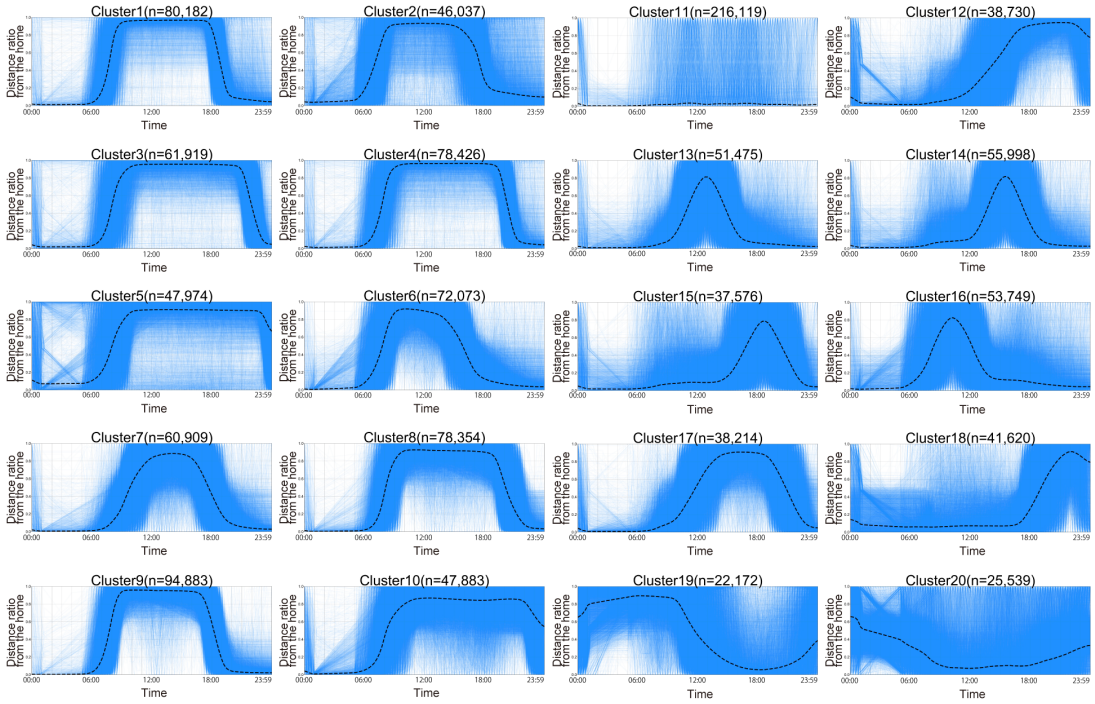

図 1　クラスタリング結果（出典：Kobayashi et al., 2019[a]）

表 1　クラスター毎 曜日割合（出典：Kobayashi et al., 2019[a]）

Cluster	Weekday（user）	Weekend（user）	Percentage of Weekday Users（%）
1	15,543.4	2,967.5	84.0
2	7,968.2	4,015.5	66.5
3	11,957.4	2,528.5	82.5
4	15,334.0	2,683.0	85.1
5	8,412.2	3,935.0	68.1
6	11,601.6	7,032.5	62.3
7	9,584.8	6,492.5	59.6
8	14,326.8	3,360.0	81.0
9	17,097.2	4,698.5	78.4
10	8,082.6	3,735.0	68.4
11	27,207.6	40,040.5	40.5
12	5,208.4	6,344.0	45.1
13	6,528.2	9,417.0	40.9
14	6,734.4	11,163.0	37.6
15	4,501.0	7,535.5	37.4
16	7,779.0	7,427.0	51.2
17	5,485.4	5,393.5	50.4
18	5,979.4	5,861.5	50.5
19	3,211.4	3,057.5	51.2
20	3,670.0	3,594.5	50.5

その後，市区町村別の人口統計データを用いて推定した居住地及び通勤通学地を検証し，強い相関があることを明らかにしている．

しかし，居住地の推定には滞在時間を用いて，基本的に最も長く過ごす場所を居住地と見なしていることから，長時間労働者への配慮が足りない．また，クラスタリングという手法の特性上，結果の解釈における分析者の主観的操作がみられることが問題として挙げられる．

第 3 章 データセット

3.1 データセット一覧

本研究で用いる主な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）
3. 市区町村別 昼間人口・夜間人口データ（総務省統計局，2015）[20]
4. 町丁目別 推計昼間人口データ（日本統計センター，2015）

3.2 PT データ

パーソントリップ調査（PT 調査）とは，人々の 1 日間の移動（トリップ）を把握するための調査である．調査手法は調査票や Web 上のアンケートによるものであり，トリップ時刻・トリップ目的・交通手段・起終点などの移動に関する情報やデモグラフィック属性情報を収集している．これらのデータから，都市圏における人々の移動を総合的に把握し，交通計画や防災対策，まちづくりの検討など様々な分野に活用している．

日本では 1967 年に広島都市圏で大規模に実施され，現在は京阪神都市圏や東京都市圏など全国各地で行われている．本研究において使用する“PT 調査”は，特に東京都市圏交通計画協議会によって行われる 10 年に 1 度の大規模な都市交通調査を指すこととする．同協議会は 1968 年に発足し，既に計 6 回（2020 年 1 月時点）の PT 調査を実施している[21]．本研究で用いる PT データは第 5 回 東京都市圏 PT 調査によるものであり，以下の概要の通り行われたものである[22]．

- 調査期間：2008 年 10 月 1 日から 11 月末のうち，平日 1 日
- 調査対象地域：東京都※ / 神奈川県 / 埼玉県 / 千葉県 / 茨城県南部 ※島嶼部除く
- 調査対象者：無作為に抽出された世帯の構成員（5 歳以上）全員

表 2，図 2 に PT データ（n = 587,434）における，性別と年齢別の内訳を示す．人流データの対象時期である 2016 年とは時差があるものの，男女比や年齢別の内訳は 2015 年の国勢調査（総務省統計局，2015）とある程度類似した傾向が見られるため 2008 年のデータを用いる[23]．

表 2 PT データ 性別内訳

Class	Sample
男性	278,996
女性	308,438
不明	0

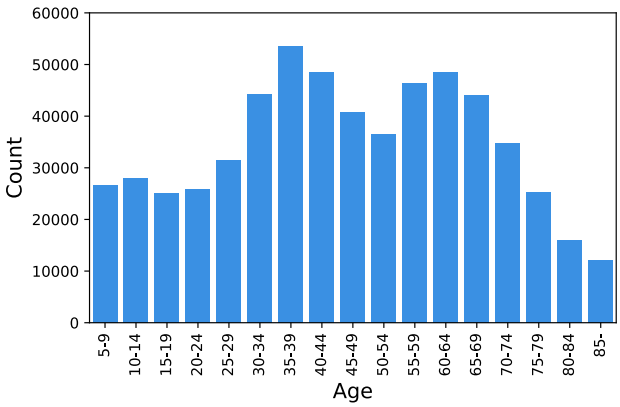

図 2 PT データ 年齢別内訳（5 歳階級）

居住地及び通勤通学地推定の観点からみたとき，（研究目的で利用できる）PT データにおける特徴は，記録されている緯度経度がゾーン単位の情報であるということである．調査票においては正確な場所名や名称を記入する PT 調査ではあるが，個人情報保護のため，公開されている PT データは各緯度経度がゾーンの代表点で表された匿名加工済みのデータとなっている．ゾーンとは，住民基本台帳などから居住人口・性別・年齢を元に分割されたエリアを指しており，そのため，同ゾーン内に居住地と通勤地がある場合は，いずれの地点も同一の緯度経度で再現されていることとなる．

そこで本研究では，より詳細に緯度経度が表現されている空間配分版 PT データを用いる．空間配分版 PT データとは，ゾーン内の建物の延床面積比率を各ゾーン内での起終点の生起確率と見なし，トリップの起終点を各建物に確率的に再配分したものである[24] [25]．無論，空間配分版データも厳密には正確な緯度経度を表していない．しかし，モデルに入力する特徴量を作成する上で，緯度経度が再配分されたデータが好ましく，かつ実データに近いと考え，本研究では空間配分版を採用する．なお，空間配分版 PT データは，東京大学空間情報科学研究センターによる加工済みデータであり，CSIS 共同研究利用システム（JoRAS）[26]から提供を受けたものである．

3.3 人流データ

人流データ（Agoop データ）は株式会社 Agoop[7] によって収集されたデータを用いる．同社では，提供するスマートフォンアプリ（iOS / Android で利用可能）をインストール済みのユーザーから同意の元，位置情報データを収集している．図 3 の示す通り，様々なアクティビティ時に身に着けていると想定されるスマートフォンの位置情報に基づくデータである．表 3 にデータ概要を示す．福田（2017）[27]によると，近年の Agoop データには独自に推定した各属性情報が含まれているが，2016 年時点の本データには属性ラベルがない．

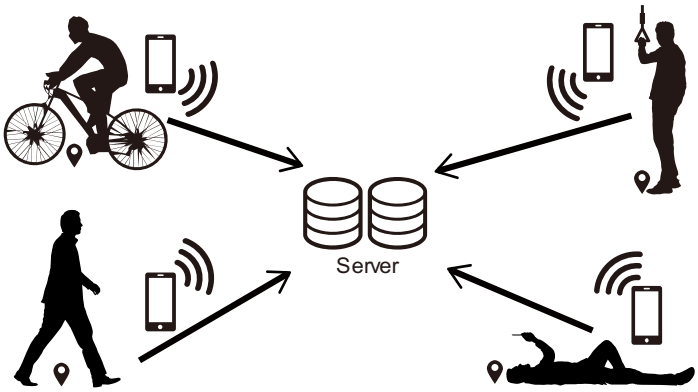

図 3　人流データ取得イメージ

表 3　人流データ概要

項目	詳細
形式	ポイント型流動人口データ
対象者	対象スマートフォンアプリをインストールした 関東地方全域の人々 （域内流入・域外流出含む）
対象期間	2016/06/01 ～ 2016/06/30 （推定には平日のみ利用）
データ項目	ユーザー ID，時刻，緯度経度，位置精度 など

人流データはデジタル広告マーケティングなどの商用のみならず，行政機関向けの観光分析や災害分析など幅広い分野で用いられている[28]．従来のアンケート調査によって把握できなかったような動的な人の流れを捉え，データを元にした定量的な施策を立てることを可能にする．例えば，観光分析では混雑時間を回避したツアー時間の最適化や近隣観光地との連携などが考えられる．

人流データには，主に 3 つの特徴がある．第一に，各 UID は午前 0 時にリセットされるということである．類似した株式会社ゼンリンデータコムの混雑統計®[29]をはじめとした人流データは長期間固定した UID を持つのに対して，特徴的なデータである．そのため，UID を元に日を跨いで分析することができず，長期の時系列や周期性に基づく分析が困難である．

第二に，深夜時間帯におけるログの欠如である．図 4 は時間帯別のログ数を示している．図 4 から分かる通り，午前 1 時から 5 時のデータが他の時間帯と比べ大幅に少ないことがわかる．これは匿名化加工のため，原則同時間帯の位置情報を記録していないことが原因である[30]．実際，後述する 5.4 でも示しているように，同時間帯は居住地に滞在する人が多いため，居住地推定では内挿処理をする必要がある．

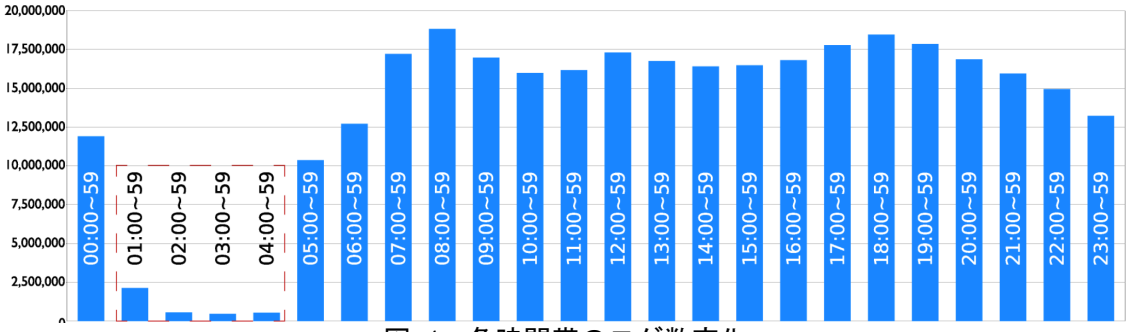

図 4　各時間帯のログ数変化

第三に，測位が低頻度，つまりスパースであるという特徴がある．測位頻度は携帯電話端末 OS によって異なるが，例えば Android 端末の場合 30 分に 1 回であり（バックグラウンド起動時），これは様々な測位間隔の人流データを研究した Alessandretti et al.（2017）[31]によると，低頻度な測位ということができる．

表 4 に土/日曜日も含めた 6 月の人流データの基礎統計量を示す．

表 4　6 月データの基礎統計量

項目	詳細
総ユーザー数	（延べ）3,070,439
総レコード数	145,861,354
ユーザー数平均 ／ 日	（延べ）102,347
ユーザー数標準偏差 ／ 日	1,986

図 5 は本研究で用いる 6 月のログ数を示している．概ね 1 ヶ月を通して，安定したサンプルサイズであるが，6 月 5 日から 1 週間単位でログ数が少なくなる傾向があることがわかる．これは日曜日に起こる現象であり，主に仕事が休みである休日に移動頻度が減ることに起因するものと考えられる．

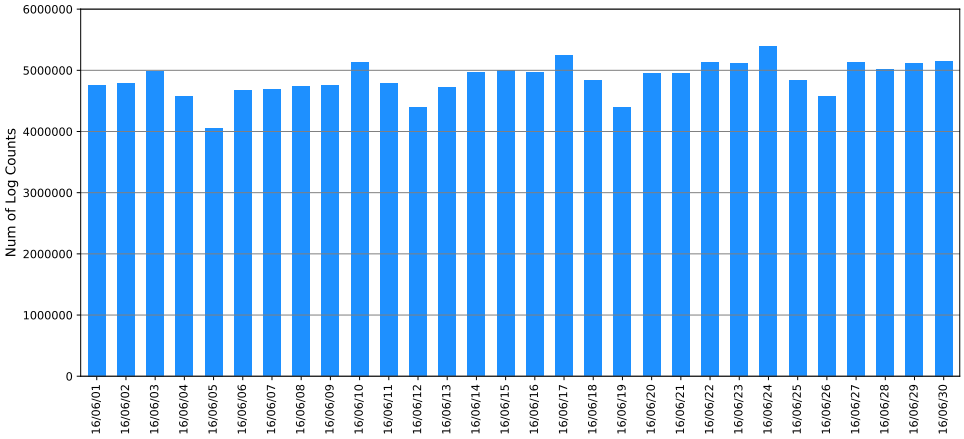

図 5　ログ数（6 月データ）

図 6 は各日のデイリー UID 数を示している．図 5 同様，日曜日に微減が見られるものの，1 ヶ月を通して安定したサンプル数を得られていることがわかる．

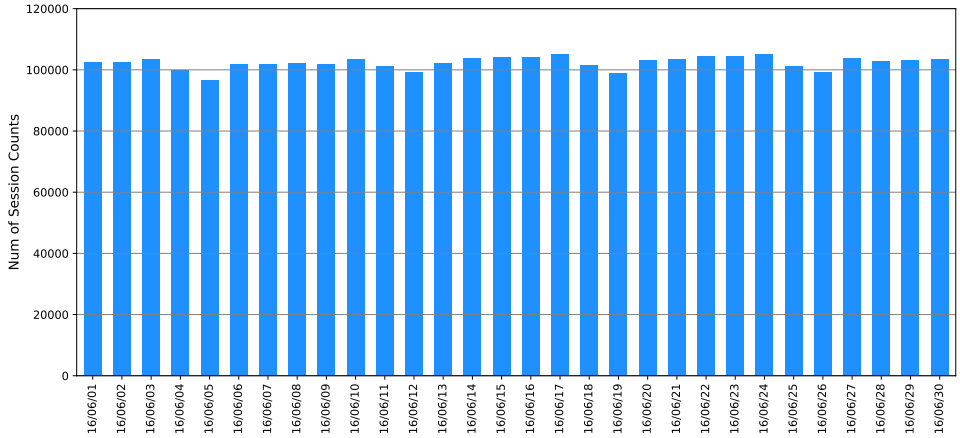

図 6　デイリー UID 数（6 月データ）

また，人流データには位置誤差という測位の精度を示す値が記録されている．これは値が小さいほど測位の精度が高いことを表している．同データにおける測位方法は GPS 測位，Wi-Fi 測位，通信基地局 3 点測位のいずれかによるものであるが，中でも 3 点測位によるログの場合は精度が低く位置誤差が大きくなる可能性が高い[30]．図 7 は 2016 年 1 年分の測位における位置誤差を図示したものであり，誤差が 1 km を超えるデータまであることがわかる．このような大きな誤差がある場合は，居住地や通勤通学地推定において問題が生じるため，いずれかの値を閾値として設定する必要がある．そこで本研究では，図を参照し位置誤差が収束する 365 m を閾値距離として利用した．

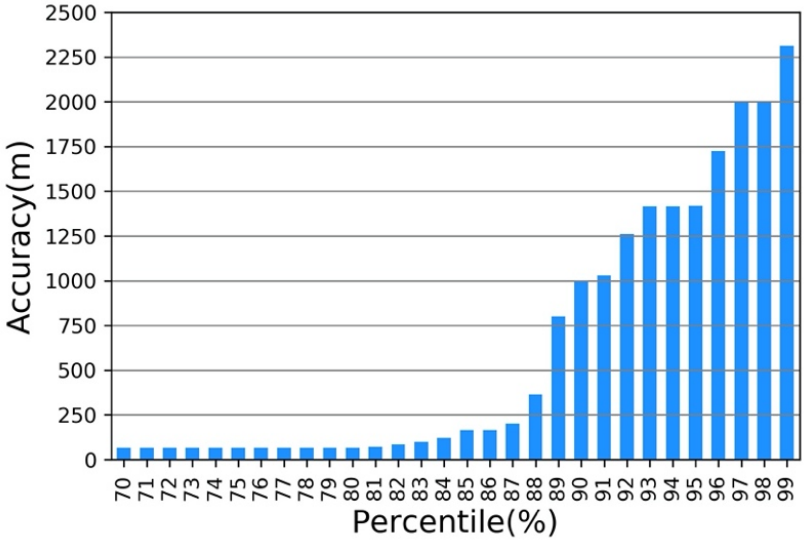

図 7　位置誤差パーセンタイル

次に，市区町村別 ログ数の上位 10 市区町村を表 5 に示す．いずれも東京都 23 区の市区町村となっており，都心のオフィスが多いエリアへの人口集中の様子が伺える．

表 5　ログ数 上位 10 市区町村

	都道府県	市区町村名	ログ数
1	東京都	港区	95,541,932
2	東京都	千代田区	72,282,725
3	東京都	中央区	69,266,077
4	東京都	渋谷区	67,988,283
5	東京都	新宿区	64,273,490
6	東京都	世田谷区	53,218,272
7	東京都	大田区	45,217,987
8	東京都	品川区	43,285,164
9	東京都	文京区	39,736,227
10	東京都	江東区	38,464,030

3.4 市区町村別 人口統計データ

平成 27 年度 国勢調査（総務省統計局, 2015）[20]による関東地方の市区町村別 夜間人口・昼間人口を推定結果の検証用データとして用いる．ここで，夜間人口とは国勢調査の調査時（2015 年 10 月 1 日）に各該当地域に常住している人々を指しており[32]，空間分布は図 8 の通りである．夜間人口の上位 10 市区町村は表 6 の通りであり，オフィス街ではなく住宅街エリアとして知られるような市区町村が多いことがわかる．

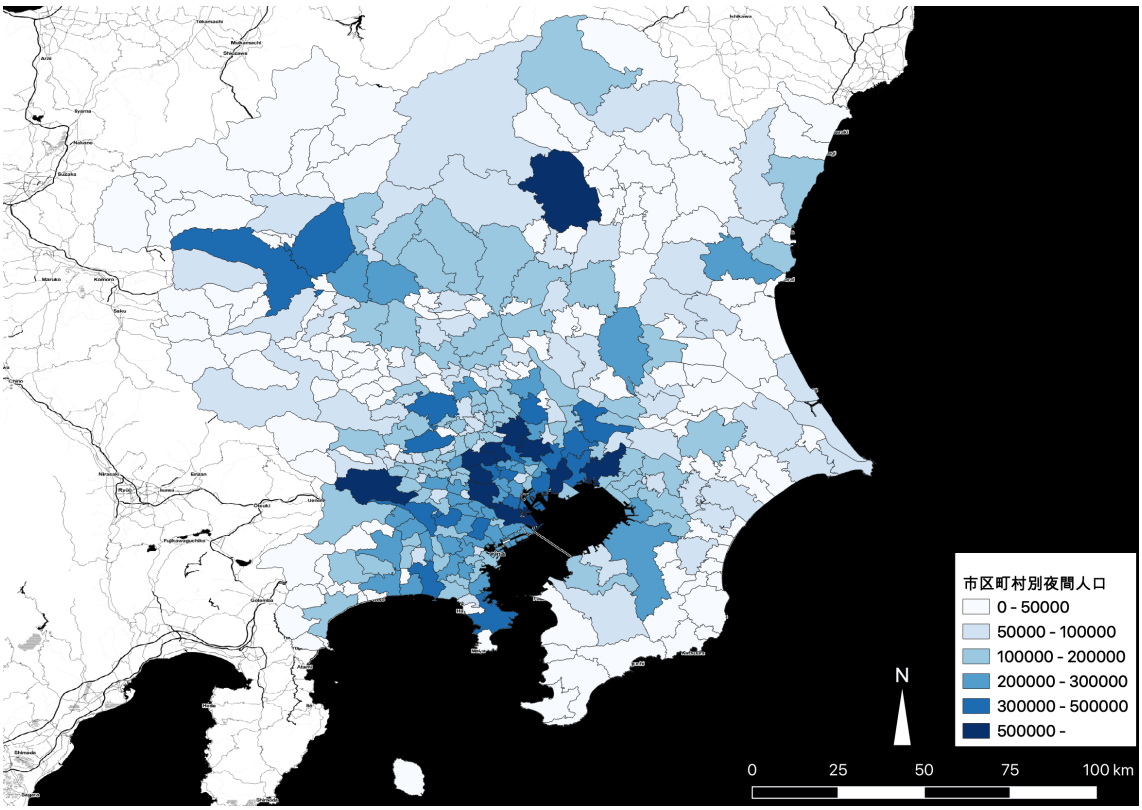

図 8　市区町村別 夜間人口分布

表 6　夜間人口 上位 10 市区町村

	県名	市区町村名	夜間人口
1	東京都	世田谷区	903, 346
2	東京都	練馬区	721, 722
3	東京都	大田区	717, 082
4	東京都	江戸川区	681, 298
5	東京都	足立区	670, 122
6	千葉県	船橋市	622, 890
7	埼玉県	川口市	578, 112
8	東京都	八王子市	577, 513
9	東京都	杉並区	563, 997
10	東京都	板橋区	561, 916

次に昼間人口について述べる．国勢調査では，夜間人口から流出人口と流入人口を足し引きした人口を昼間人口としている．具体的には，式 (2)（国勢調査より引用 [A]）で表される．なお，夜間労働者や夜間通学者も便宜的に昼間人口に含まれているが，買い物客や観光客などは含まれていない．空間分布については図 9 の通りである．

$$A\text{市の昼間人口} = A\text{市の夜間人口} - A\text{市からの流出人口}^{*1} + A\text{市への流入人口}^{*2} \quad (2)$$

*1: A 市からの流出人口：A 市から A 市以外への通勤通学者数

*2: A 市への流入人口　：A 市以外から A 市への通勤通学者数

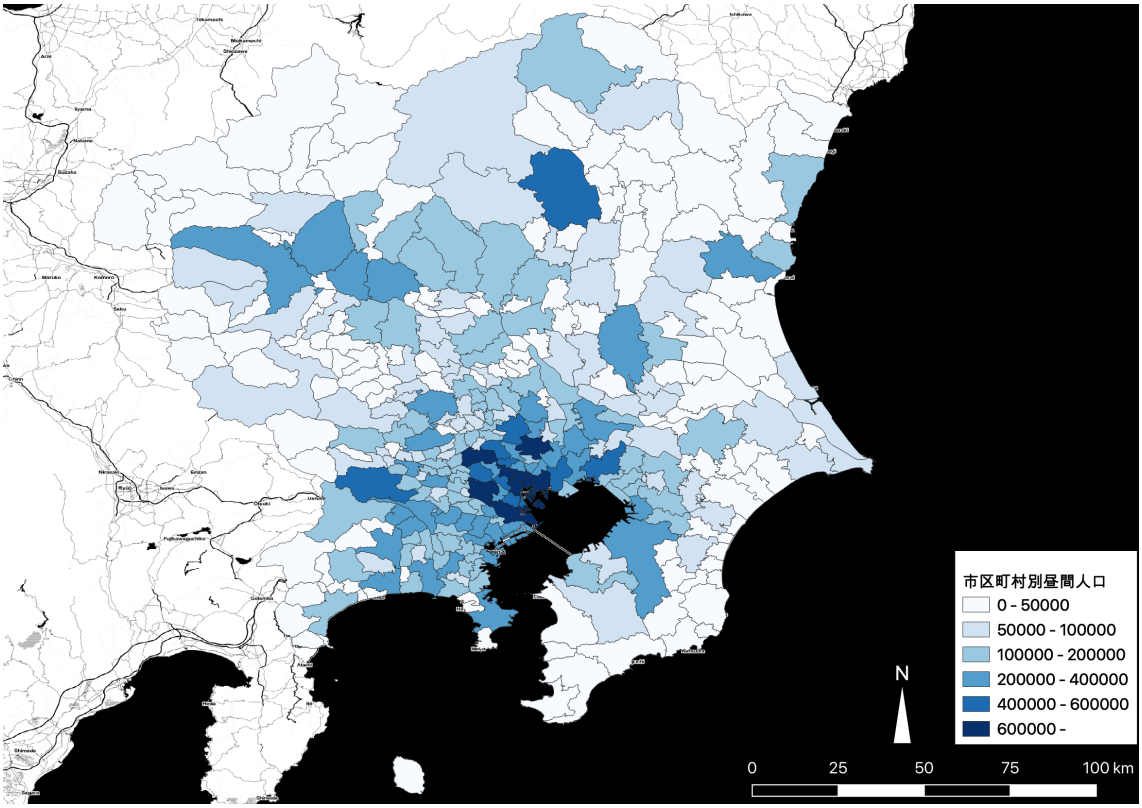

図 9　市区町村別 昼間人口分布

昼間人口の上位 10 市区町村は表 7 の通りであり，表 6 と比べるとオフィス街として知られるような東京都心部に位置する市区町村が多いことがわかる．

表 7　昼間人口 上位 10 市区町村

	県名	市区町村名	昼間人口
1	東京都	港区	940,785
2	東京都	世田谷区	856,870
3	東京都	千代田区	853,068
4	東京都	新宿区	775,549
5	東京都	大田区	693,865
6	東京都	足立区	608,968
7	東京都	中央区	608,603
8	東京都	江東区	608,532
9	東京都	練馬区	605,084
10	東京都	八王子市	576,240

3.5 町域別 人口統計データ

国勢調査データと同様に，関東地方の町域別人口統計データも検証用データとして用いる．ここで町域とは，郵便番号データ（日本郵便株式会社）による郵便番号単位の区域をいい，例えば”千葉県柏市柏の葉＊丁目＊番地＊号”における”千葉県柏市柏の葉”までを指す．そのため，市区町村よりは詳細であり，町丁目よりは粗いエリアである．データセットには，NSC データベース 国勢調査地図版の町丁目別 推計昼間人口データ（日本統計センター，2015）を利用する．同データは，同センターが独自の手法を用いて推計したデータ

である．本研究では，人流データにおける位置誤差の影響から町丁目単位では推定が困難であることが考えられるため，町域単位に集計して検証用データに用いる．図 10 に町域別昼間人口の空間分布を示す．

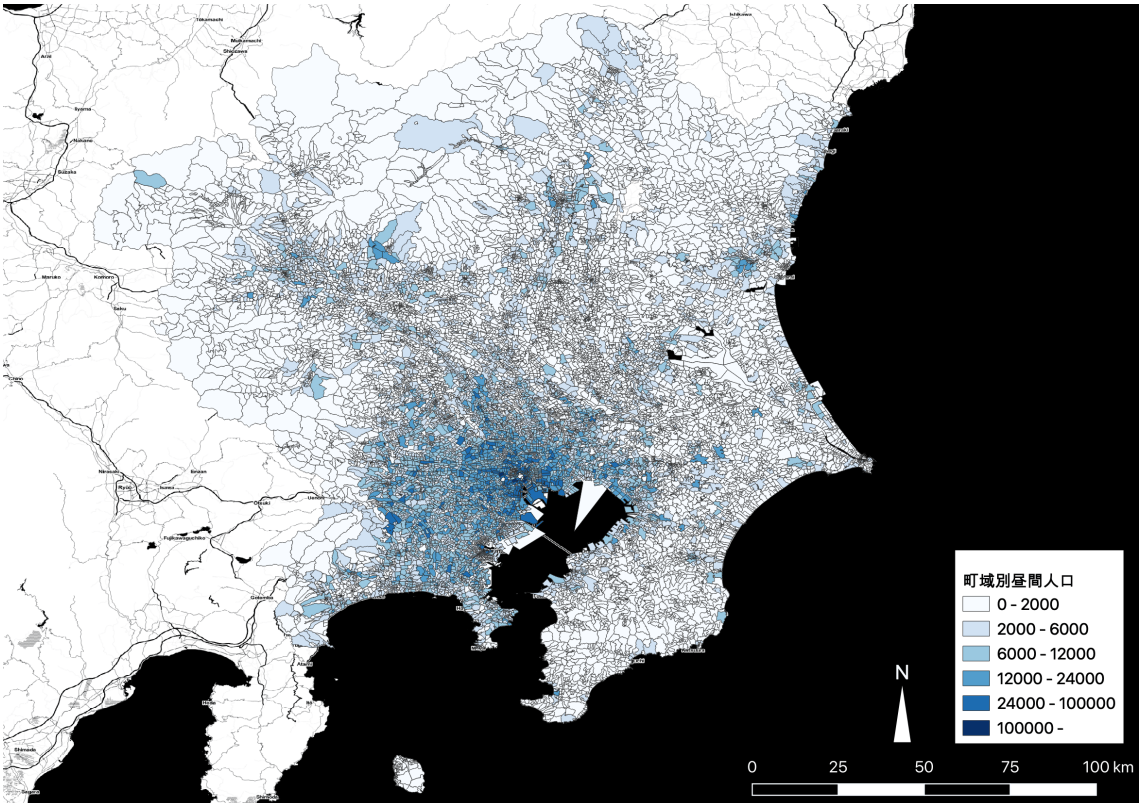

図 10　町域別 昼間人口分布

昼間人口の上位 10 町域は表 8 の通りである．西新宿が他町域に比べて特に多いのは，町域面積が大きいことが要因の一つである．

表 8　推計昼間人口 上位 10 町域

	県名	市区町村名	町域別	推計昼間人口
1	東京都	新宿区	西新宿	267,566
2	東京都	中央区	銀座	134,391
3	東京都	千代田区	丸の内	110,931
4	東京都	港区	赤坂	103,423
5	東京都	新宿区	新宿	94,786
6	東京都	豊島区	東池袋	93,112
7	東京都	千代田区	大手町	91,492
8	東京都	港区	港南	89,108
9	東京都	品川区	大崎	79,767
10	東京都	渋谷区	渋谷	77,556

第 4 章 本研究で用いる手法

本章では本研究で用いる手法を説明する．ここでは，第 2 章でも説明した本研究の前身でもある，筆者らによる Kobayashi et al.（2019）[19] で用いたクラスタリング関連手法も含めて説明する．

4. 1 K−means

K-means 法とは非階層型クラスタリングアルゴリズムの一つで，データを超球面上にクラスター分割することができる．

以下に MacQueen et al.（1967）[33]によるアルゴリズムを示す．

1. データ $x_i\ (i = 1, \ldots, n)$ をランダムに k 個に分割する（= 初期化）
2. 各クラスターのセントロイド $V_j\ (j = 1, \ldots, k)$ を求める
3. 全データ x_i と V_j の距離を計算し，x_i が所属するクラスターを最も近いクラスターに更新する
4. 更新されなくなるか，最大計算ステップに達したら終了

図 11 に実行例を示す．図 11 左のランダムに発生させた 2 次元のデータに対して，クラスタリングを実行させる（$k = 4$）．結果は図 11 右の通りであり，点の色がクラスターラベルを示しており，適切にクラスタリングされていることがわかる．

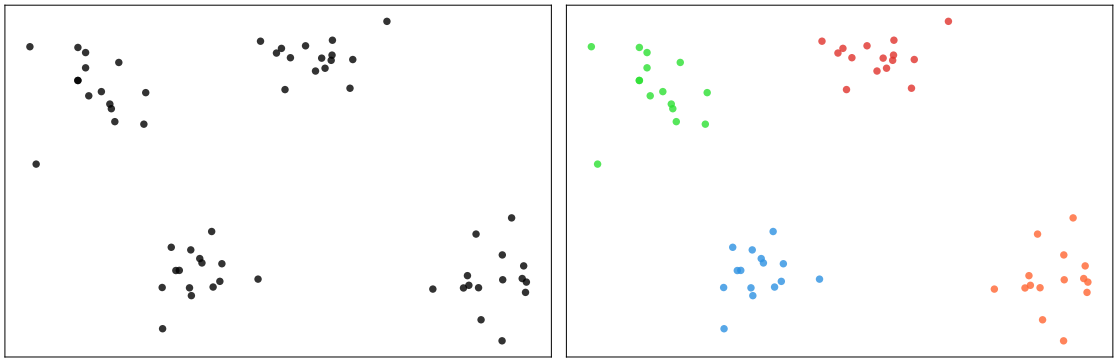

図 11 K-means クラスタリングの様子

このように，簡易的なアルゴリズムのためクラスタリングの際に多く用いられる K-means であるが，主に 2 つの欠点がある．

第一に，クラスター数（k 数）を事前に決定する必要がある点であり，このことはクラスタリング結果が分析者の主観・操作に影響されやすいという問題に繋がる．対処法としては，シルエット法やエルボー法による最適な k 数把握，もしくは自動的に最適な k 数に分割する X-means 法（Pelleg and Moore, 2000）[34]が挙げられる．

第二に，クラスタリング結果が初期値設定に依存するということである．これは，データをランダムに振る初期化プロセスにおける問題であり，初期値の設定が異なることで最終的に x_i へ割り当てられるクラスターが変わることに繋がる．この問題に対する改善手法は 4.3 の K-means++ に記す．

4. 2 Elbow Method

4.1 K-means における最適な k 数決定の手法としては，Elbow Method（エルボー法）が知られる．エルボー法は最適な k 数を決めるための手法であり，クラスター毎のクラスター内残差平方和（Sum of Squared Errors = SSE）を用いて，SSE が急降下するような k 数を最適な k 数とみなす．SSE は式 (3) で表される．

$$SSE = \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{3}$$

例えば図 12 の場合，$k = 3$ において急降下していることから，最適な k 数は 3 となる．前述の通り，K-means は分析者の決定する k 数に結果が左右されるため，このように何らかの指標を元に k 数の決定を行うことが望ましいとされる．ただし，必ずしも図 12 のような急降下が見られるわけではないため，最適な k 数検討の際はエルボー法の他にもシルエット法などの指標を用いることが好ましい．

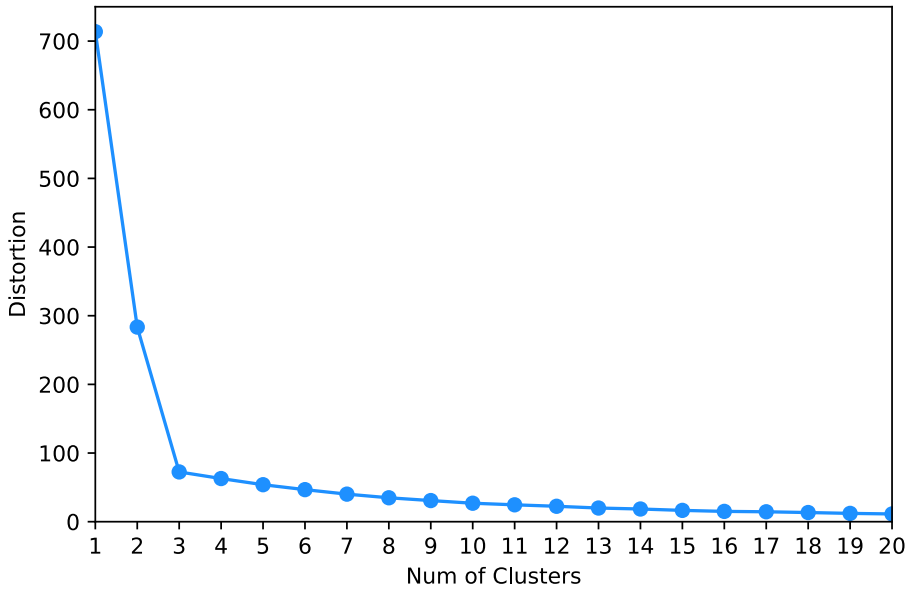

図 12　SSE の変化

4. 3 K-means++

4.1 の紹介した K-means では，初回にランダムにクラスターを割り当てる初期化に結果が左右される．そこで，初期化時のクラスター割り当ての改善を行った K-means++（Arthur et al., 2007）[35]という手法が対策として挙げられる．初回のセントロイドを割り当てる際に，互いに離れた位置を採用することで K-means よりも，一貫したクラスタリング結果が期待できる．

アルゴリズムは，データ点からランダムにセントロイドを 1 つ選択した後に，データ x と最近傍セントロイドの距離を $D(x)$ とし，データ x における重みつき確率分布$\frac{D(x)^2}{\sum D(x)^2}$を用いて，他のセントロイドをランダムに選択するというものである．

図 13 に両手法の実行結果を示す．なお，色はクラスターラベルを表している．K-means（図 13 左）ではセントロイドを近い位置に設定したことで，一部データに不適切にクラスターが割り当てられているが，K-means++（図 13 右）では適切にクラスタリングが実行されていることがわかる．

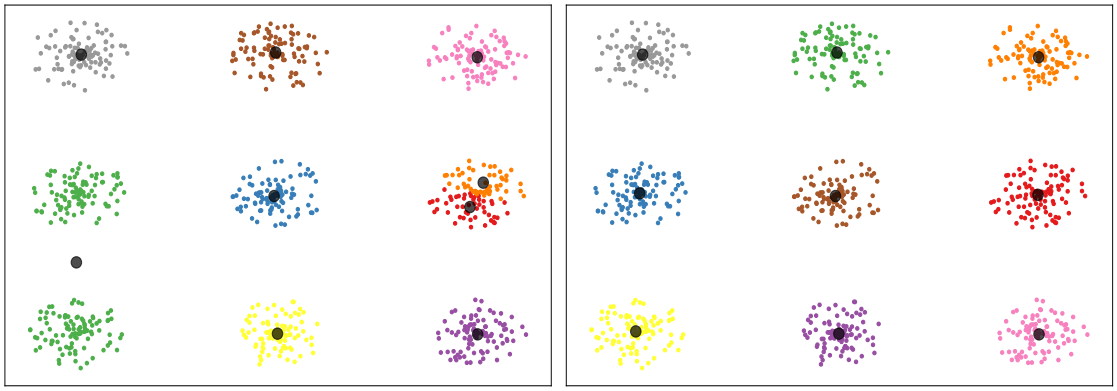

図 13　クラスタリング実行結果（左: K-means，右: K-means++）

4. 4 DBSCAN

Density-Based Spatial Clustering of Applications with Noise（DBSCAN）は Ester et al.（1996）[36]によって開発された密度準拠型のクラスタリングアルゴリズムである．DBSCAN のアルゴリズムは以下の通りであり，図 14 にイメージを示す（medium, 2019[37]を元に作成）．

1. 全データ点を Core，Border，Noise 点にわける
   Core: 半径 eps 以内に少なくとも $minPts$ 個の隣接点が位置する場合の点
   Border: 半径 eps 以内に $minPts$ 個の隣接点が存在しないが，半径 eps 以内に Core である p が存在する点
   Noise: 半径 eps 以内にも Core が存在しない点
2. 隣接する Core を接続しクラスター C を生成する
3. Core である C_i に隣接する Border も C_i に割り当てる

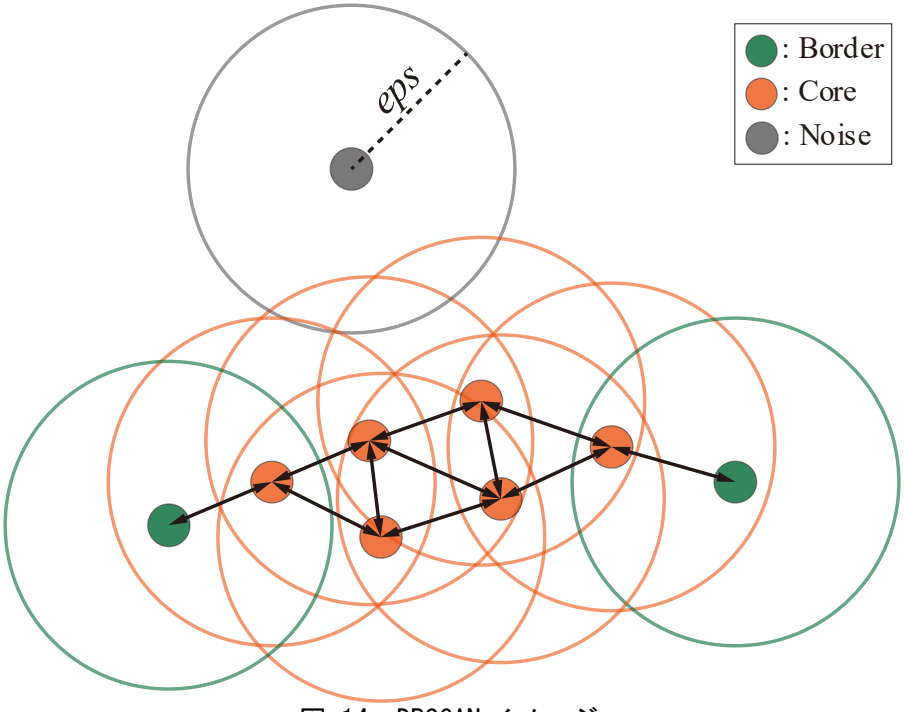

図 14　DBSCAN イメージ

アルゴリズムの通り，DBSCAN はデータ集合の密度・接続関係を考慮したクラスタリングが可能である．scikit-learn 0.22 documentation[38] を元に作成した図 15 に，他のクラスタリングアルゴリズムを含めた実行例を示す． 図 15 からも DBSCAN は他アルゴリズムに比べ，密度を元にしたクラスタリングを行っていることがわかる．

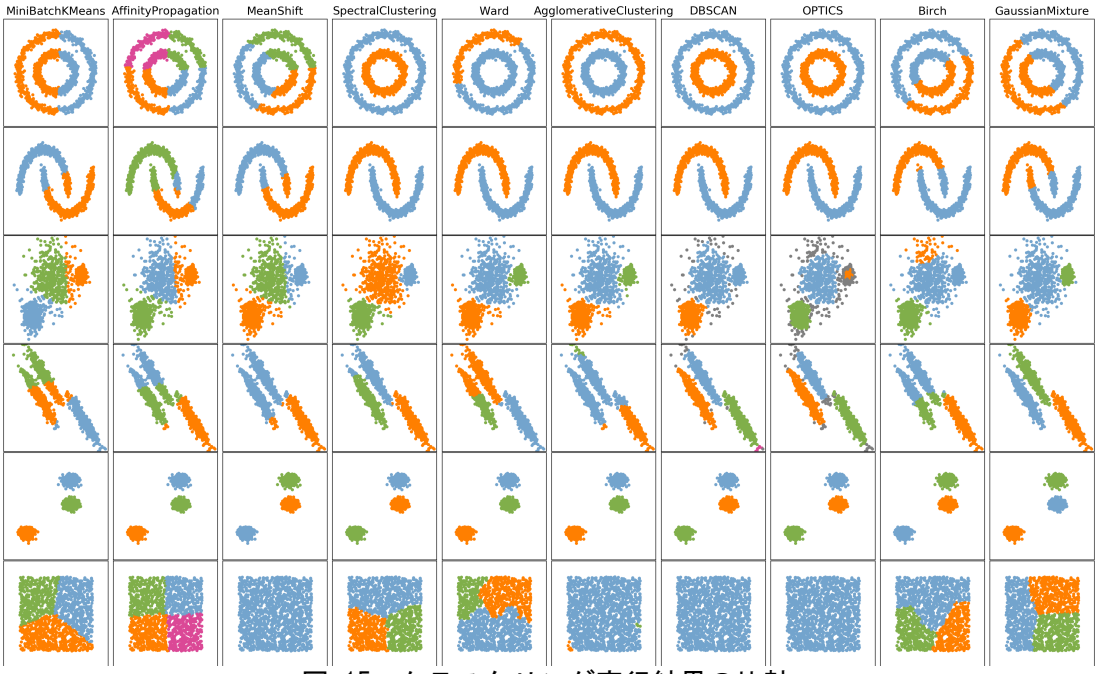

図 15　クラスタリング実行結果の比較

DBSCAN は前述した K-means とは異なり，事前にクラスター数を決定する必要がなく，非超球型のデータに対しても適用可能であること，そしてノイズに対して頑健であることが長所として挙げられ，GPS 移動履歴データに対して適用されることが多い．

ただ, DBSCAN では eps と $minPts$ という 2 つのパラメータを与える必要がある. eps は距離の閾値を，$minPts$ はクラスターを構成するために必要な最小点数を表している．そのため，$minPts$ を大きい値で設定すると，ノイズとして処理されるデータが増える可能性が高い．

本研究においては，人流データの位置精度が 365 m 以下のデータのみ扱っていることから，最大で 365 m の誤差が生じていると考え，$eps = 0.365$ とした．eps を変化させた場合のクラスターを集約的に抽出することができる OPTICS（Ankerst, 2000）[39]などがあるが，ここではドメインの知識を活かし，365 m で固定することとする．また $minPts$ については，本人流データはスパースなデータであり各ユーザーのログが比較的少ないため，ノイズとして多くのデータが除外されることのないよう $minPts = 2$ とした．

4. 5 Random Forest

Random Forest は代表的な機械学習アルゴリズムの一つで，決定木によって複数の弱学習器を組み合わせることで精度の高い分類や回帰を行う「アンサンブル学習アルゴリズム」である[40]．Random Forest のアーキテクチャは図 16 の通りである．一定の精度を得られること以外にも，学習速度の速さやシンプルさが長所として挙げられ，並列化により大量のデータに対しても有効である．アルゴリズムは以下の通りである．

1. データ集合 S から，ランダムサンプリングによって N 個のサンプルを抽出する
2. サンプルを用いて決定木 T_i 生成する
3. T_i に用いていないデータを用いて T_i の性能を測る
4. 多数決に基づいて T_i を結合することで優れた T_k を生成する

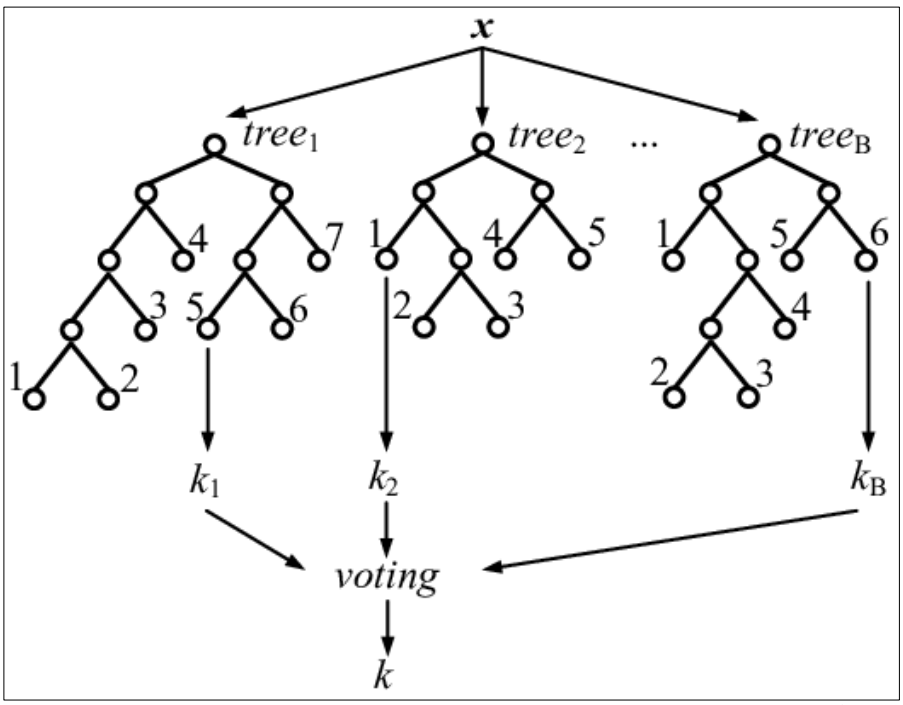

図 16　Random Forest アーキテクチャ（出典：Gelzinis et al., 2014[b]）

4. 6 LightGBM

LightGBM とは，MicroSoft（2017）[41] [42]によって開発された 勾配ブースティング機械学習アルゴリズムである．回帰分析やクラス分類のタスクにおいて高い精度を得られるため，Kaggle[43] などの競技プログラミングなどにも多く用いられる．

LightGBM の特徴は，Leaf-wise という方法で決定木を成長させる方法を採用していることである．Leaf-wise とは，決定木の葉（leaf）に準じて決定木を成長させる手法を指している．同じくブースティング系アルゴリズムの XGBoost が図 17 のように決定木の成長を水平方向に成長させる Level-wise tree growth であるのに対し，LightGBM は図 18 のように Leaf-wise を用いて決定木を葉に準じて成長させている．これにより損失の削減を実現でき，学習時間の短縮などのメリットがある．

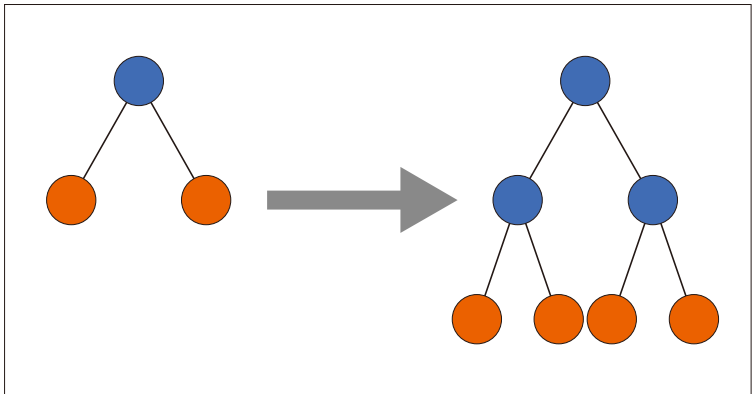

図 17　Level-wise tree growth イメージ

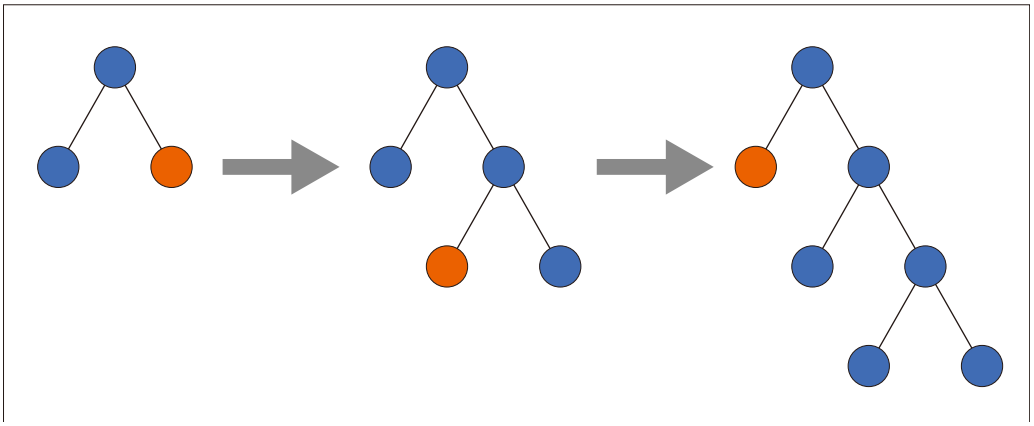

図 18　Leaf-wise tree growth イメージ

4.7 RNN 系

ここでは，まずニューラルネットワークについて説明し，次に RNN 及び RNN の拡張モデルである LSTM・GRU を説明する．

ニューラルネットワーク（NN）とは，人間の脳の神経回路網を模した数理モデルであり，図 19 のような構造を持つ．ここでは，例として入力層，隠れ層，出力層から構成される簡易的な 3 層パーセプトロンを挙げる．NN は，出力層で正解値が出るように，重みとバイアスを調整する学習を行う．重みは入力値毎に設定されシナプス結合の強さを表す．バイアスは入力値を一定範囲に偏らせる偏りを表す．入出力を繰り返し行うことで，これらの重みとバイアスを調整し精度を向上させ，予測結果を出力していく．

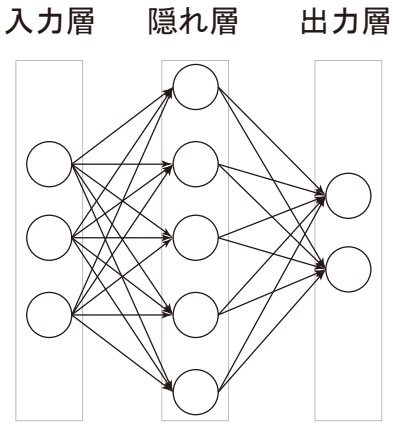

図 19　NN 概略図

Recurrent Neural Network（RNN）とは，再帰型の NN であり時系列データを適切に取り扱うことができる特徴をもつ．一般的に時系列データは強い自己相関をもつことが多く，シーケンシャルデータを入力できる RNN は有用である．従来の NN の構造は図 19 の通りで，入力値が互いに独立している．そのため，ある層の出力が次の層の入力に利用されるのみで互いにリンクおらず，時系列データにおける入力値の連続性について考慮されたモデルではない．

一方で，RNN は隠れ層に時間に依存した情報を埋め込むことで，次の入力時に利用し前回の状態を保持しながら学習することができる．ここで，簡略化した RNN の構造を図 20 に示す．図 20 のように，RNN では隠れ層に再帰結合を組み込むことで，前回の入力を記憶可能とし，前後の情報を元にした学習が行えるようになっている．例えば，自然言語処理において「私の名前は」という文章に続く，単語を予測する場合について考える．NN の場合は各入力が独立して扱われることから，”は” のみを利用して予測することになる．それに対して RNN の場合は，それ以前の情報も保持できることから ”私”，”の”，”名前”，”は” といった前後の入力も考慮した上で，予測することができるようになる．

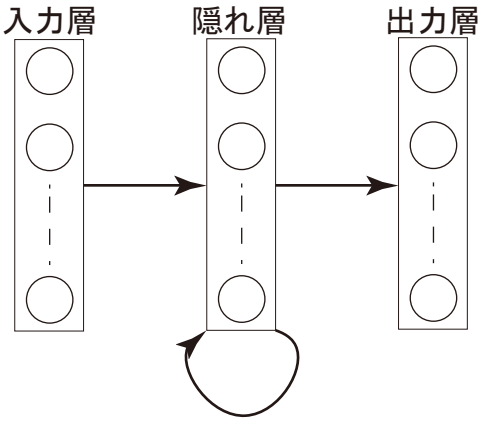

図 20　RNN 概略図

さらに，時刻 t の入力 x_t，隠れ層の値を h_t としたとき，時系列データ集合 $X = \{x_1, x_2,,, x_T\}$ の RNN への入力時イメージを図 21 に示す. 図 21 から分かる通り, RNN では前回の入力を記録し次の入力時に活かすため，時刻 t における隠れ層 h_t の値は時刻 $t-1$ における h_{t-1} に依存している．このようにして，RNN はシーケンシャルデータを

扱うことができ，その他にも可変長のデータも取り扱うことができる長所から，センサーから定期的に収集されるデータの予測や自然言語処理など多く用いられる．

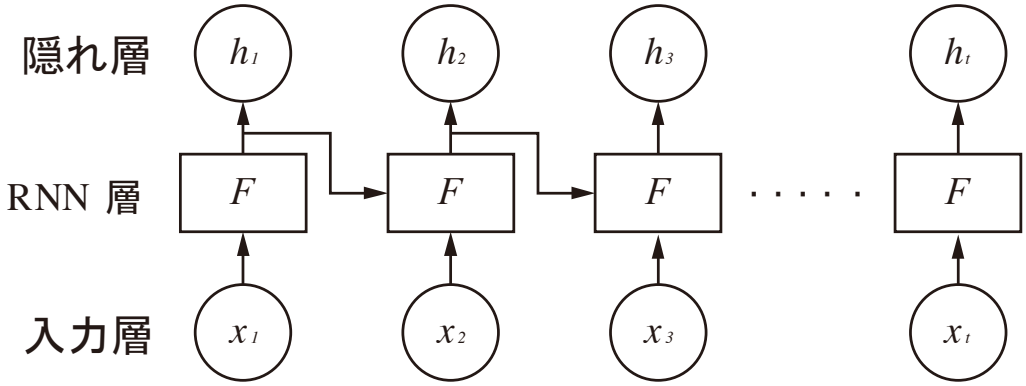

図 21　RNN 構造

次に Long Short-Term Memory（LSTM）について述べる．NN では重みとバイアスを調整する際に，出力ニューロンの誤差を出力層から入力層に遡って各ニューロンに伝える「誤差逆伝播」を行う．ここで，重み w が 1 未満の場合，繰り返し乗算することで勾配は 0 に近づき，勾配の消失が生じることとなる．また，重み w が大きい場合は繰り返し乗算することで勾配の発散が生じる．よって，過去データに対する重みが発散・消失する「勾配消失問題」が生じる．Sigmoid 関数（最大値 0.25）ではなく，Sigmoid 関数を線形変換した tanh 関数（最大値 1）を活性化関数として採用した場合でも，いずれ飽和する可能性がある．また，勾配消失に強い ReLU 関数に関しては重みが 0 以下の場合，誤差が伝達されず学習に困難性が生じる．

そのため，RNN では勾配消失問題が生じることが難点であり，理論上は長期の時系列データにも対応しているが，実際は 10 タイムステップ程度が限界であると言われる．そこで，勾配消失問題を解決する方法として Hochreiter & Schmidhuber（1997）[44]によって提唱された RNN の拡張モデルが LSTM であり，RNN における隠れ層ユニットが LSTM ブロックに置き換えられている．

LSTM ブロックは図 22 の通りであり，入力ゲート，忘却ゲート，出力ゲートの 3 つのゲートと記憶セルで構成される．記憶セルは，ゲートと組み合わせることで記憶するかどうかの判断を行い，不要な誤差を除く役割を担う．入力，出力ゲートでは必要な誤差のみを通すことで，不要な誤差で誤った重みに更新することを防いでいる．忘却ゲートでは，記憶セルに不要な誤差が停留することを防止するため，不要な情報を除去しリセットする．

このように LSTM では記憶セルと忘却ゲートを中心に勾配消失問題を解決し，長期の系列データを扱うことができる．前述した RNN では困難だった 1000 タイムステップのような長期の系列を学習することができる．

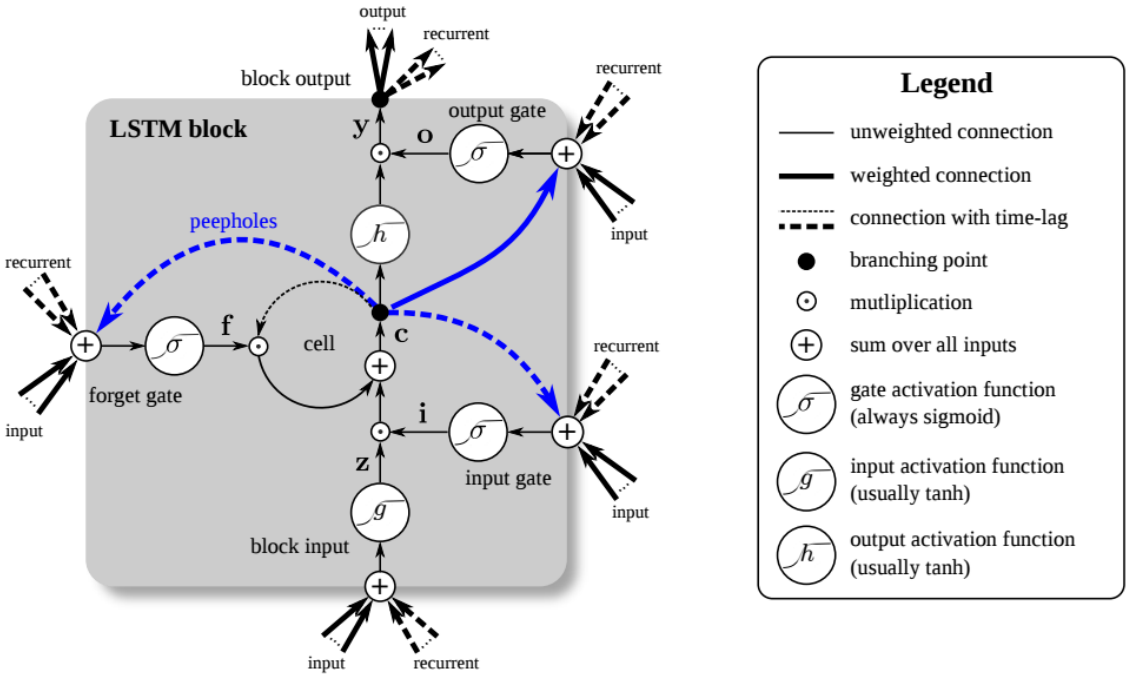

図 22　LSTM ブロック（出典：Greff et al., 2017[c]）

最後に，GRU について説明する．GRU は Gated Recurrent Unit の略称で，Kyunghyun et al.（2014）[45]によって開発された RNN の拡張モデルである．LSTM を簡略化した図 23 のようなアーキテクチャであり（Chung, 2014[46] を参考に作成），LSTM における入力ゲート，出力ゲート，忘却ゲートが，GRU ではリセットゲート，更新ゲートによって代替されている．GRU は，タスクにもよるが LSTM と同等の精度を発揮すること，そして何より計算時間の削減ができることから，データセットが大きい場合などに用いられる．

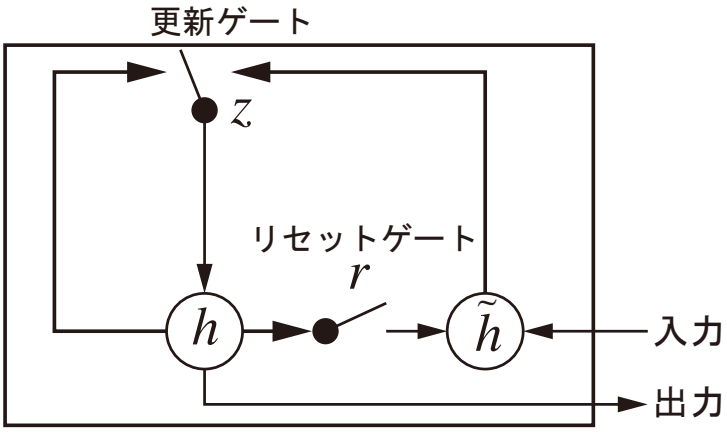

図 23　GRU ゲート概略図

第 5 章 居住地及び通勤通学地推定

5.1 推定フロー

フローは図 24 の通りであり，以下に要約する．まず PT データから職業属性分類モデルを構築する．人流データのフローに移り，全ユーザーに対して時間帯及び滞在時間を元に居住地推定を行う．次に転移学習として，PT データによって学習済みの職業分類モデルを人流データに適用し，就業者・学生・その他に分類する．そして，NHK 放送文化研究所による国民生活時間調査報告書（2015）[47]を参考に，就業者・学生に分類されたユーザーに対して，居住地以外で最も滞在時間が長い地点（閾値時間以上滞在）を通勤通学地として推定する．なお，推定においては 1 ヶ月の人流データのうち，平日データのみを利用する．

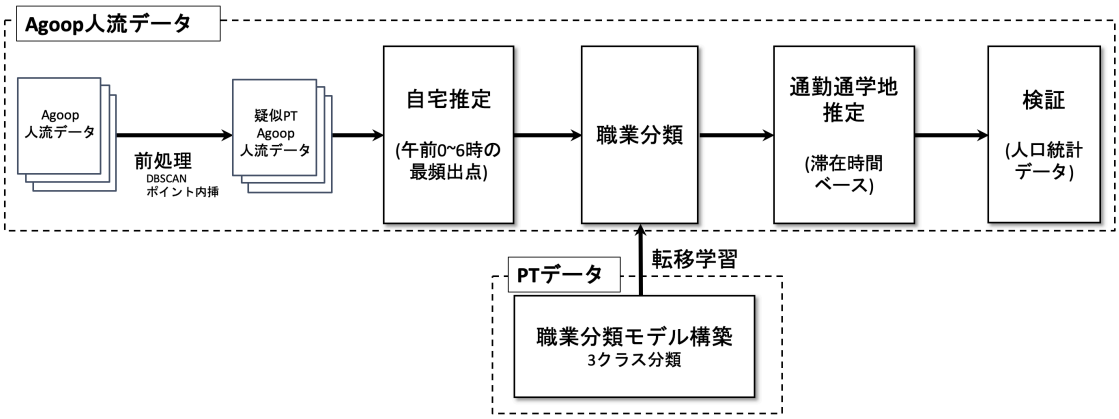

図 24　推定フロー

PT データを用いた職業属性モデルの構築，及び転移学習（Transductive Transfer Learning）については，既往研究に西村ら（2014）[48]がある．西村らは，PT データで得た知識を転移することで，人流データ内ユーザーのデモグラフィック属性（性別・年代・職業）の推定を試みている．しかし，同研究では株式会社ゼンリンデータコム（ZDC）による人流データ「混雑統計®[29]」を用いている．ZDC 人流データは，測位間隔が最短 5 分であり，UID が長期に渡って固定されている．そのため，西村らの構築するモデルは，長期の移動履歴を含み，密な人流データを想定したものとなっている．また，$F1$ 値は 0.711 であり，向上の余地がある．その点，本研究で用いるのはスパースな人流データであること，かつより高い $F1$ 値を求めることに意義がある．

5.2 職業属性分類モデル構築

5.2.1 PT データ前処理

PT データには回答者の調査票への記入に基づき，属性ラベルやトリップに関する情報が付与されている．今回は居住地及び通勤通学地推定のため，付与されているラベルのうち”職業属性”，”トリップ目的”を用いる．

まず，職業属性について述べる．職業属性は各回答者に対して，表 9 の Category, Occupation 列の通り付与されている．これらを，本研究では表 9 の Class 列に記載している通りに職業を分類し，簡易化した上で扱う（職業不明な回答者のデータは除去）．なお，例外的に行った分類について以下に示す．

・　就業者：その他属性であってもトリップに業務系目的トリップがある者
　　学生属性であっても通学トリップがなく業務系目的トリップがある者
・　その他：就業者属性であってもトリップに業務系目的トリップがない者
　　学生属性であってもトリップに通学及び業務目的トリップがない者

よって，1.3 でも述べた通り本研究で分類する職業属性とは，あくまでもデータ取得した日における職業属性を示しており，各人物の普遍的な職業属性を示すものではない．

表 9　PT データ内職業属性一覧

Category	Occupation	Class
1	農林水産業従事者	就業者
2	生産工程・労務作業者	
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	専門的・技術的職業従事者	
9	管理的職業従事者	
10	その他職業	
11	園児・小学生・中学生	学生
12	高校生	
13	大学生・短大生・各種専門学校生	
14	主婦・主夫（職業従事者を除く）	その他
15	無職	
16	その他	
99	不明	－

上記の通りに分類した職業属性別のライフスタイルを確認するため，職業属性別の居住地の出発・帰宅時間の分布を以下に示す．まず，図 25，図 26，図 27 は職業属性別の居住地の出発時間分布を示し，縦軸は回答者数を表している．就業者と学生では類似した傾向の分布ではあるが，午前 8 時前後におけるピーク時への偏りに違いがみられる．一方で，その他属性は午前の遅い時間以降の活動が多く，他の属性と異なる傾向を示している．

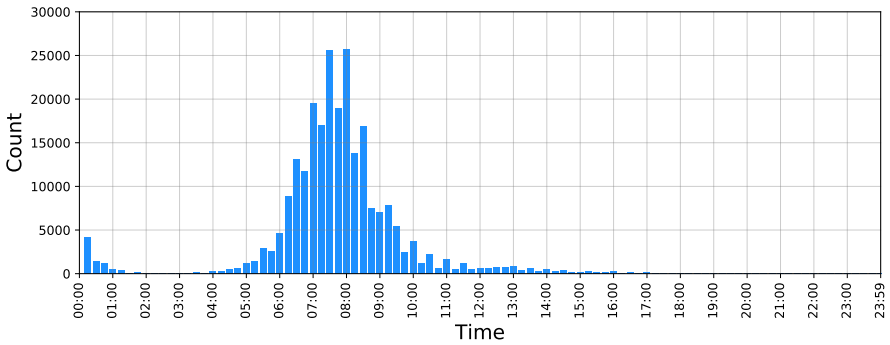

図 25　就業者の出発時間

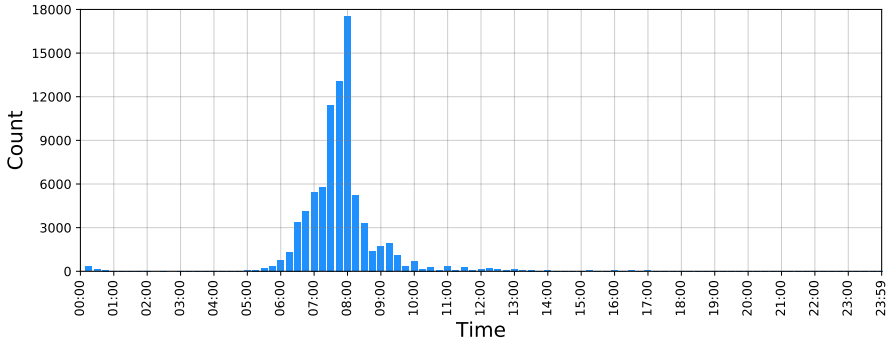

図 26　学生の出発時間

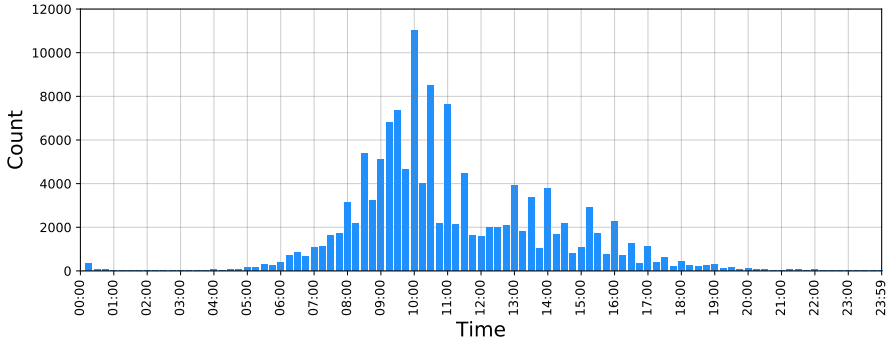

図 27　その他の出発

図 28，図 29，図 30 は，職業属性別の居住地の帰宅時間を表している．就業者の帰宅時間は定時で帰宅する人から深夜近くまで残業していると思われる人まで多様である．一方，学生は夕方頃に帰宅することが多く，大学生・専門学校生などの一部の属性が夜遅くに帰宅している．また，その他属性は昼から夕方頃の帰宅が多い傾向がある．このように，属性毎に居住地の出発・帰宅時間分布は異なり，かつ同一属性内においても一定傾向の範囲内で違いが見られることがわかる．

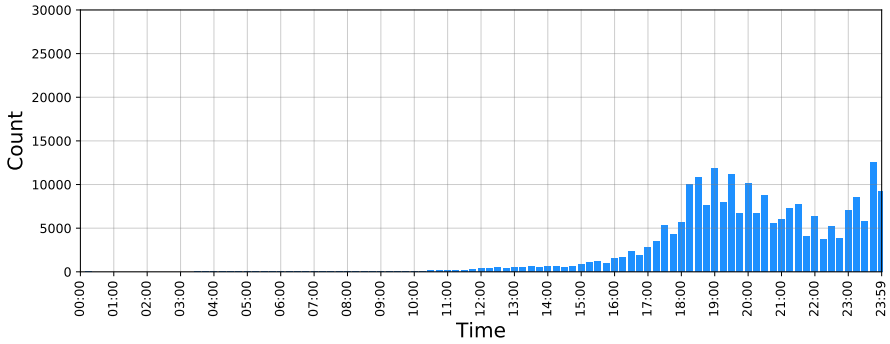

図 28　就業者の帰宅時間

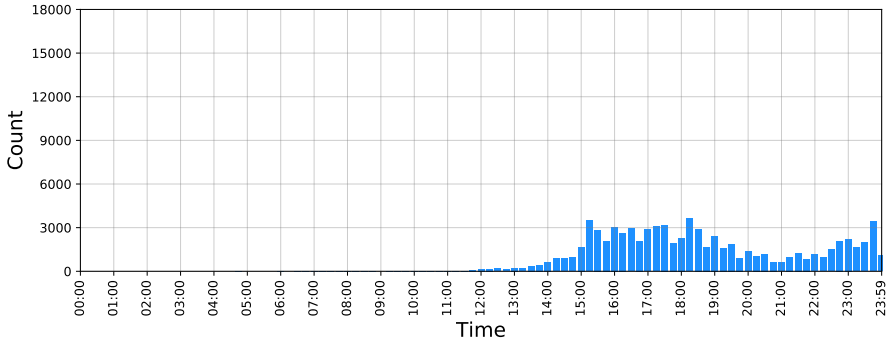

図 29　学生の帰宅時間

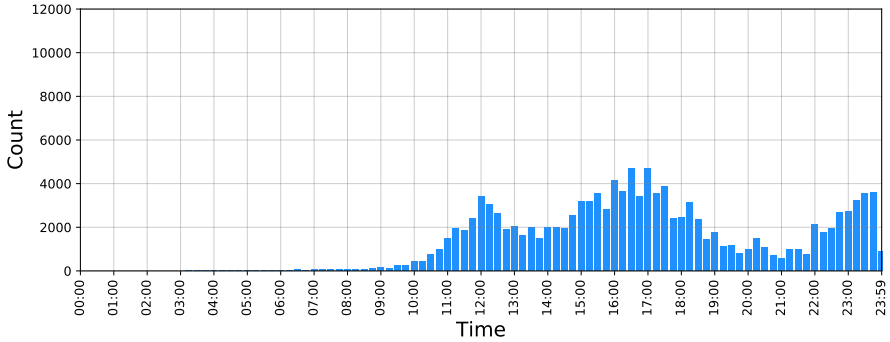

図 30　その他の帰宅時間

次に，トリップ目的について述べる．PT データでは表 10 のように各トリップに対して，トリップ目的情報が付与されている．そこで，トリップ目的から居住地及び通勤通学地を抽出する．例えば，”3 居住地へ”目的のトリップの終点は居住地であることを利用している．なお，移動しているにも関わらずトリップ目的が一切回答されていない（=99）場合，データを対象から除外する．

表 10　PT データ内トリップ目的一覧

Category	Content	本研究での扱い
1	勤務先へ（帰社を含む）	通勤/業務
2	通学先へ（帰校を含む）	通学
3	居住地へ	帰宅
4	買物へ	–
5	食事・社交・娯楽へ（日常生活圏内）	
6	観光・行楽・レジャーへ（日常生活圏外）	
7	通院	
8	その他の私用へ（塾・習い事など）	
9	送迎	
10	販売・配達・仕入・購入先へ	通勤/業務
11	打合せ・会議・集金・往診へ	
12	作業・修理へ	
13	農林漁業作業へ	
14	その他の業務へ	
99	その他	–

5. 2. 2 クラス不均衡性解消

対象となるデータにおいて，クラス毎のサンプル数に不均衡性がある場合，クラス分類に大きく影響がでる．例えば，表 11 のようなサンプルで構成されるデータについて考えてみる．Positive と Negative，各クラスのサンプル数に 1 : 99 と大きく差があることがわかる．この偏りを保ったままバイナリ分類モデルを構築した場合，全て Negative と判断してしまう可能性がある．実例として同様に不均衡データとなるケースは，発生率の極めて低い疾患の発生が挙げられる．

表 11　不均衡データ クラス内訳例

クラス	サンプル数
Positive	10
Negative	990

よって，不均衡データをクラス分類する場合，クラスの不均衡性の解消を行った上で学習させる必要がある．不均衡性の対処法としては，コスト考慮型学習を行う，もしくはデータレベルで加工を行う，計 2 つがある．

前者については各クラスに対してペナルティ課してコスト考慮型学習を行うことで，不均衡性を解消する方法である．

後者はオーバーサンプリングもしくはアンダーサンプリングを行うことで，不均衡性を解消する方法である．オーバーサンプリングとはサンプル数の少ないデータを増やす手法であり，アンダーサンプリングは逆にサンプル数の多いデータを減らす手法である．

これらを踏まえた上で，PT データのクラス別サンプル数を表 12 に示す．学生のサンプル数が少なく，不均衡なデータであることがわかる．

表 12　PT データ クラス内訳

クラス	サンプル数
就業者	244,860
学生	82,907
その他	131,572

そこで本研究では，PT データに対しデータレベルの加工を行うこととする．一般的にはデータ数を保持するためオーバーサンプリングが好ましいとされる．しかし，本研究で用いる特徴量はシーケンスデータであり，オーバーサンプリングに用いられる Synthetic Minority Over-sampling Technique（SMOTE）[49]はシーケンスデータに対応していない為，時系列的な意味のあるオーバーサンプリングを行うことができない．また，今回はアンダーサンプリングの場合でも精度が十分期待できることを理由に，アンダーサンプリングを用いて不均衡性の対処を行う．

5. 2. 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．ラベル付きデータが十分揃えることが困難なとき，半教師あり学習や能動学習と並んで多く用いられるのが，この転移学習である．機械学習分野において 1995 年頃から認識され始め[50]，現在も音声認識や自然言語処理をはじめとした多くの専門領域において用いられる．

転移元であるデータドメインを元ドメイン（Source Domain），転移先であるデータドメインを目標ドメイン（Target Domain）といい，　これを元に転移学習を表 13 のように細分化することができる．本研究では元ドメイン（S）の PT データに職業属性の正解ラベルがあ

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（神嶌，2010[51]）.

表 13　転移学習の細分化

		Target Domain	
		with Labels	without Labels
Source Domain	with Labels	Inductive Transfer Learning	Transductive Transfer Learning
	without Labels	Self-Taught Learning	Unsupervised Transfer Learning

ここで改めて，本研究における PT データと人流データの概要を表 14 に示す．転移学習を用いて，属性ラベルが付与されている PT データから構築したモデルを人流データに適用することで職業属性を推定する．PT データは属性ラベルがあり，人流データはサンプル数が多い．そこでドメインを跨いで知識の転移をすることで，両データの長所を活かして分析を行うことができる．

表 14　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（平日のみ）
総ユーザー数	587,434	（延べ）2,272,810
職業属性ラベル	あり	なし

転移学習は既往研究において一貫した定義がないのが現状であるが，サーベイ論文である神嶌（2010）[51] によると，両ドメイン間においてどのような類似性があるか仮定した上で転移を行う必要があると言われている（転移仮定）．それにより，転移モデルに転移仮定における設定を組み込む必要性が生じる．

今回のような，目標ドメインに正解ラベルがない Transductive Transfer Learning では，確率分布に関して議論することは一般的に難しいため暗黙的に仮定されることもあるが，ここでは Nishimura et al.（2014）[52]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 $P\left[Y^{(S)}|X^{(S)}\right]$ と $P\left[Y^{(T)}|X^{(T)}\right]$ 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．$P\left[Y^{(D)}|X^{(D)}\right]$ は，ドメインを $D \in \{S,T\}$，データ集合を $X = \{x_1, \dots, x_k\}$，ラベル空間を $Y = \{y_1, \dots, y_n\}$ としたとき，X が観測された際にラベルが Y である確率分布を表す．

5. 2. 4 モデル設定

分類モデルよって職業属性の多クラス分類を行う．クラスは下記の 3 つである．定義は 1.3 に記した通りであり，備考として具体例を示す．

1. 就業者（具体例：サラリーマン，パートする主婦/夫，アルバイトのみ行う学生）
2. 学生　（具体例：小学生，中学生，高校生，大学生）
3. その他（具体例：パートしない主婦/夫，休日のサラリーマンや学生）

就業者については，自社などの特定勤務地で働かず営業の外回りなどのトリップをしている就業者や，複数のオフィスで 1 日間働いている就業者も PT データにみられる．つまり，”マルチ勤務地”の就業者データも含まれる．職業属性分類モデルには，これらのデータも入力しており，マルチ勤務地な就業者に対しても就業者と分類するようにモデルを構築している．

分類手法としては以下の 4 つを用い，比較することで最適な手法を決定する．Random Forest はベンチマークとして，LightGBM は分類精度に優れることから利用し，GRU は時系列データに強くシークエンス入力が可能なことから利用する．NN は，再帰型 NN である GRU との比較として用いる．

1. Random Forest
2. LightGBM
3. NN
4. GRU

次に，各モデルのパラメータについて述べる．Random Forest のパラメータはデフォルト値を利用する．LightGBM は，学習回数 50，学習率 0.05 とし，その他はデフォルト値を用いる．

NN のモデルには，単純な 3 層パーセプトロンを用いる．最適化手法を Adam，初期学習率 0.01，バッチサイズ 32，学習回数は 30 回とする．出力層の活性化関数には Softmax 関数を選択する．ここで Softmax 関数は，n 個の出力層において a_i を入力値，$k \in \{1,,,n\}$ としたとき，式 (4) で表される．1 つの実数値入力のみでなく，n 次元の実数ベクトルを入力し，0 から 1 に正規化した n 次元の実数ベクトルを返すことができ，出力される各要素は最大値 1，全要素の合計値は 1 となる．そのため，出力を確率と解釈することができ，バイナリ分類はもちろん多クラス分類でも用いられる．よって，活性化関数には Softmax 関数を利用し，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．

$$y_k = \frac{exp(a_k)}{\sum_{i \in \{1,\ldots,n\}} exp(a_i)} \tag{4}$$

GRU のパラメータは，GRU の層数を $\{1, 2\}$，隠れ層のユニット数を $\{32, 128\}$，初期学習率を $\{0.01, 0.1\}$，バッチサイズを $\{32, 64, 128\}$とした組み合わせの中から複数試行し，隠れ層のユニット数 32，初期学習率 0.01，バッチサイズ 32 を採用する（dropout 無し）．学習回数は 50 回，最適化手法を Adam とし，モデルを構築する．NN と同様に活性化関数は Softmax 関数とする．

次に特徴量について述べる．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 毎に正規化した，97 タイムステップのシーケンス情報（時系列移動パターン）を用いる．ここでは，厳密に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用いてモデルを構築する．なお，Google Maps API[53] などを利用した交通ネットワークに沿ったネットワーク距離を利用するには多額のコストが発生するため，本研究では直線距離を利用する．

特徴量の具体例を図 31 に示す．これは可視化のために作成したダミーユーザーの移動軌跡である．時系列に沿って移動軌跡を追うと，推定された居住地に 0 時から 8 時まで滞在し，その後，電車である駅を経由し，勤務先へ移動している．勤務先には 9 時前から 19 時まで滞在し，その後，行きと同じ経路で 20 時に居住地に帰宅するという軌跡となっている．このとき，居住地からの距離を元にした時系列移動パターンの特徴量は，図 31 右下の

通りとなる．UID 毎に 0 から 1 に正規化を行うため，居住地から最も離れている勤務先の特徴量は 1 となり，乗換駅は中間の地点であることから特徴量は 0.5，居住地での特徴量は 0 となる．つまり，一切外出しない場合を除いて，いずれのユーザーも最大値 1 の特徴量をもつこととなる．このようにして，全ユーザーに対して特徴量を算出し，各モデルの学習を行う．

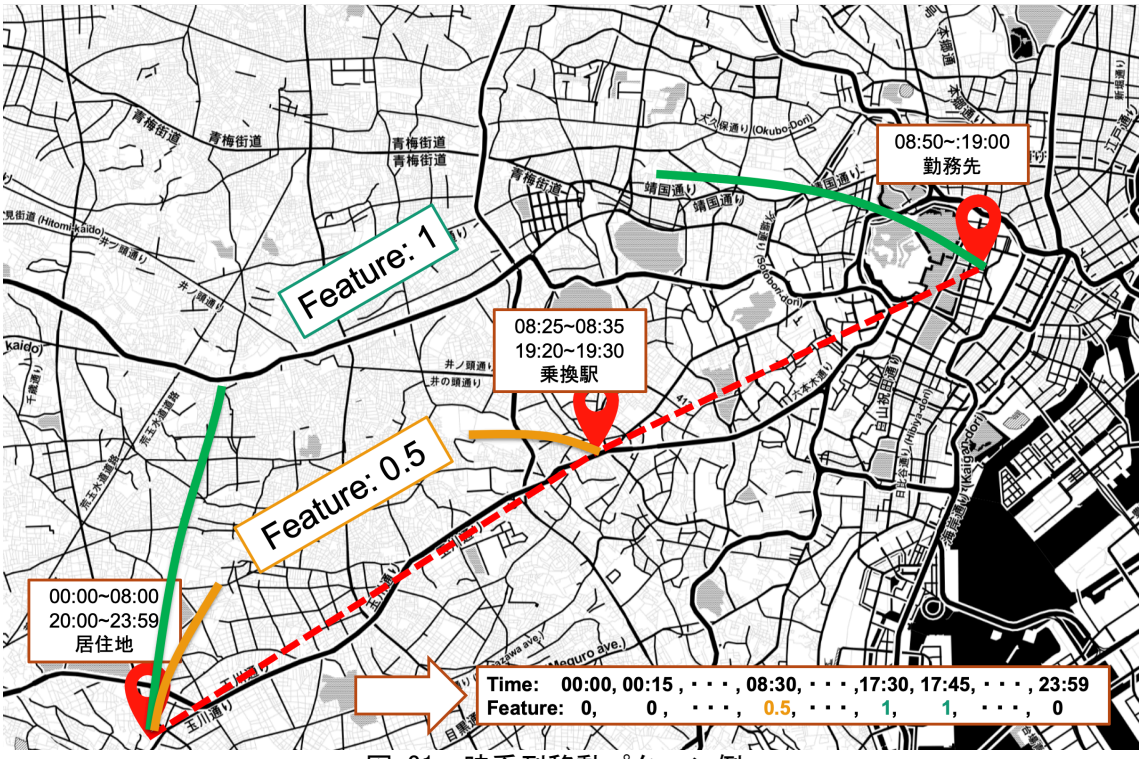

図 31　時系列移動パターン例

図 32，図 33，図 34 に属性別の特徴量，時系列移動パターンのグラフを示す．黒線は中央値，青線は無作為に抽出されたサンプル数 200 を表している．3 属性を比較すると，その他属性が大きく他と異なることがわかる．これは，その他属性の人々が同じ時間帯に一定した地点で長く滞在しないことや，特徴量 1 をとる地点で長く滞在しないことによる．就業者と学生を比較した場合，どちらも同じ時間帯に自宅から外出し，特徴量 1 に近い地点で長期滞在している．しかし，通勤通学先での滞在時間に差が顕著に現れており，就業者はオフィスと思われる場所に夕方以降も滞在しているのに対して，学生は夕方ごろには帰宅し始めるといった違いが見られる．このように職業属性毎に特徴量の時系列移動パターンの違いが見られる．

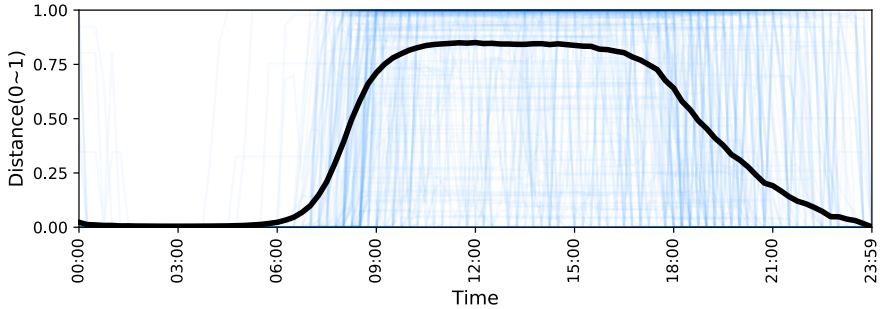

図 32　就業者の時系列移動パターン（n = 244,860）

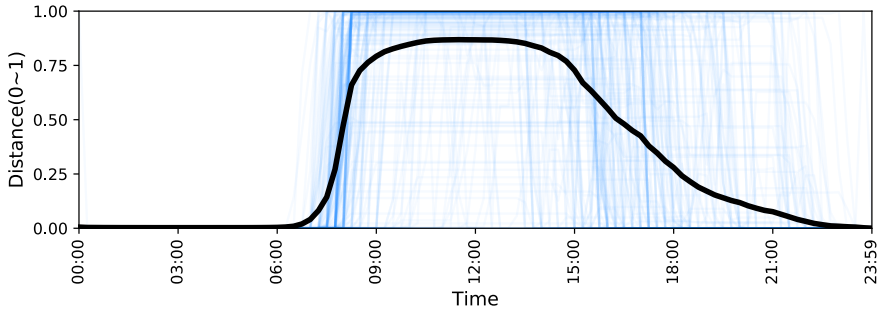

図 33　学生の時系列移動パターン（n = 82,907）

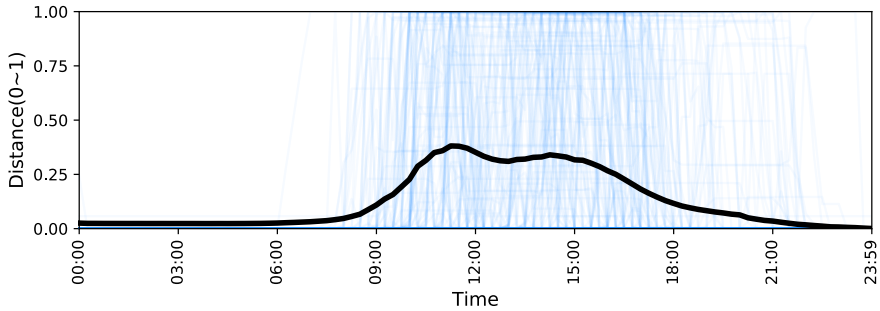

図 34　その他の時系列移動パターン（n = 131,604）

以上の 1 日の時系列移動パターンを利用し，モデリングする．モデルの検証には Hold-out 法（訓練データ 7 割，テストデータ 3 割）を利用し，訓練データで構築されたモデルをテストデータに適用した時の $Macro\text{-}F1\text{-}Measure$ を評価指標とする．ここで $Macro\text{-}F1\text{-}Measure$（$Macro\text{-}F1$ 値）とは，各クラスの $F1\text{-}Measure$（$F1$ 値）の平均値を指す. $F1$ 値をはじめとした各評価指標式は, 式 (5) から式 (10) で表される(表 15 参考). TP は True Positive の略称であり，クラス A，B，C は就業者, 学生, その他属性を表している.

表 15　Three-Classes Confusion Matrix

		True Class		
		A	B	C
Predicted Class	A	TP_A	E_{BA}	E_{CA}
	B	E_{AB}	TP_B	E_{CB}
	C	E_{AC}	E_{BC}	TP_C

$$Accuracy = \frac{\sum TP}{\sum TP + \sum E} \tag{5}$$

$$Precision_i = \frac{TP_i}{TP_i + \sum_{k\in\{p|p\in\{A,B,C\},p\neq i\}} E_{ki}} \tag{6}$$

$$Recall_i = \frac{TP_i}{TP_i + \sum_{k\in\{p|p\in\{A,B,C\},p\neq i\}} E_{ik}} \tag{7}$$

$Macro$：

$$Macro\text{-}Precision = \frac{1}{3}\sum_{i\in\{A,B,C\}} \frac{TP_i}{TP_i + \sum_{k\in\{p|p\in\{A,B,C\},p\neq i\}} E_{ki}} \tag{8}$$

$$Macro\text{-}Recall = \frac{1}{3}\sum_{i\in\{A,B,C\}} \frac{TP_i}{TP_i + \sum_{k\in\{p|p\in\{A,B,C\},p\neq i\}} E_{ik}} \tag{9}$$

$$Macro\text{-}F1\text{-}Measure = \frac{1}{3}\sum_{i\in\{A,B,C\}} \frac{2Recall_i \cdot Precision_i}{Recall_i + Precision_i} \tag{10}$$

5. 2. 5 モデル結果

表 16 に，PT テストデータに対するモデルの精度検証結果を示す．各手法を比較すると，$Macro\text{-}F1$ 値から GRU の分類精度が最も高いことがわかった．これは，他の手法が入力値を 97 次元として独立に扱う一方で，GRU が入力値をシーケンス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では GRU によるモデルを用いる．

表 16　各手法の精度検証結果

Method	$Accuracy$	$Macro\text{-}Recall$	$Macro\text{-}Precision$	$Macro\text{-}F1$
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 35，図 36 に，GRU によるモデルの学習曲線を示す．損失関数が大きく乖離し，過学習に陥ることのないよう学習回数は実験時と同じ 50 とする．

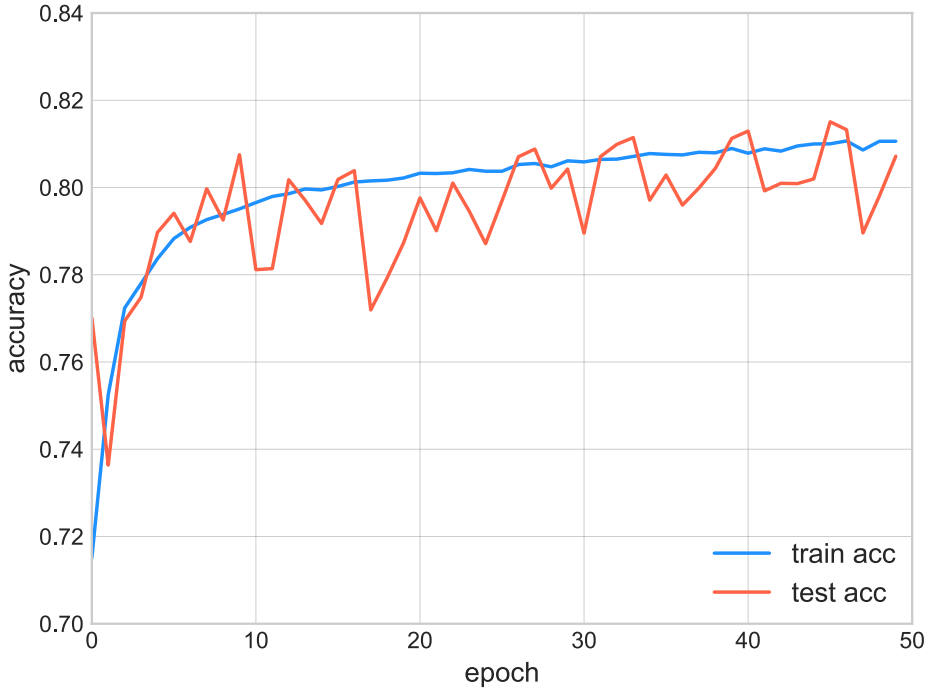

図 35　GRU の学習曲線 精度

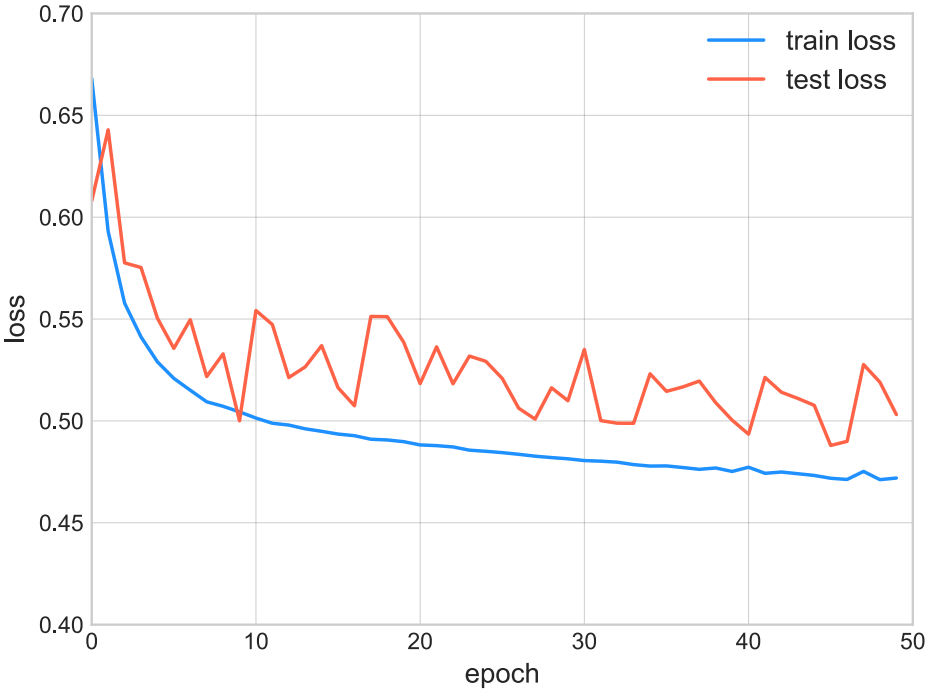

図 36　GRU の学習曲線 損失関数

また，モデルにおける各特徴量の重要度について確認する．参考として，LightGBM における特徴量重要度上位 40 個を図 37 に示す．重要度の評価には，SHapley Additive exPlanations（SHAP）指標[54]を用いる．多クラス分類を SHAP 指標で解釈した場合，各特徴量がどのクラスへの分類に影響を与えているのか把握することができる．ここでは図 37 から，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った．また，20 時や 22 時の特徴量は就業者に効いており，定時で勤務を終えるユーザーのみでなく残業などで帰宅が遅いと想定されるユーザーも就業者に分類されていることがわかった．

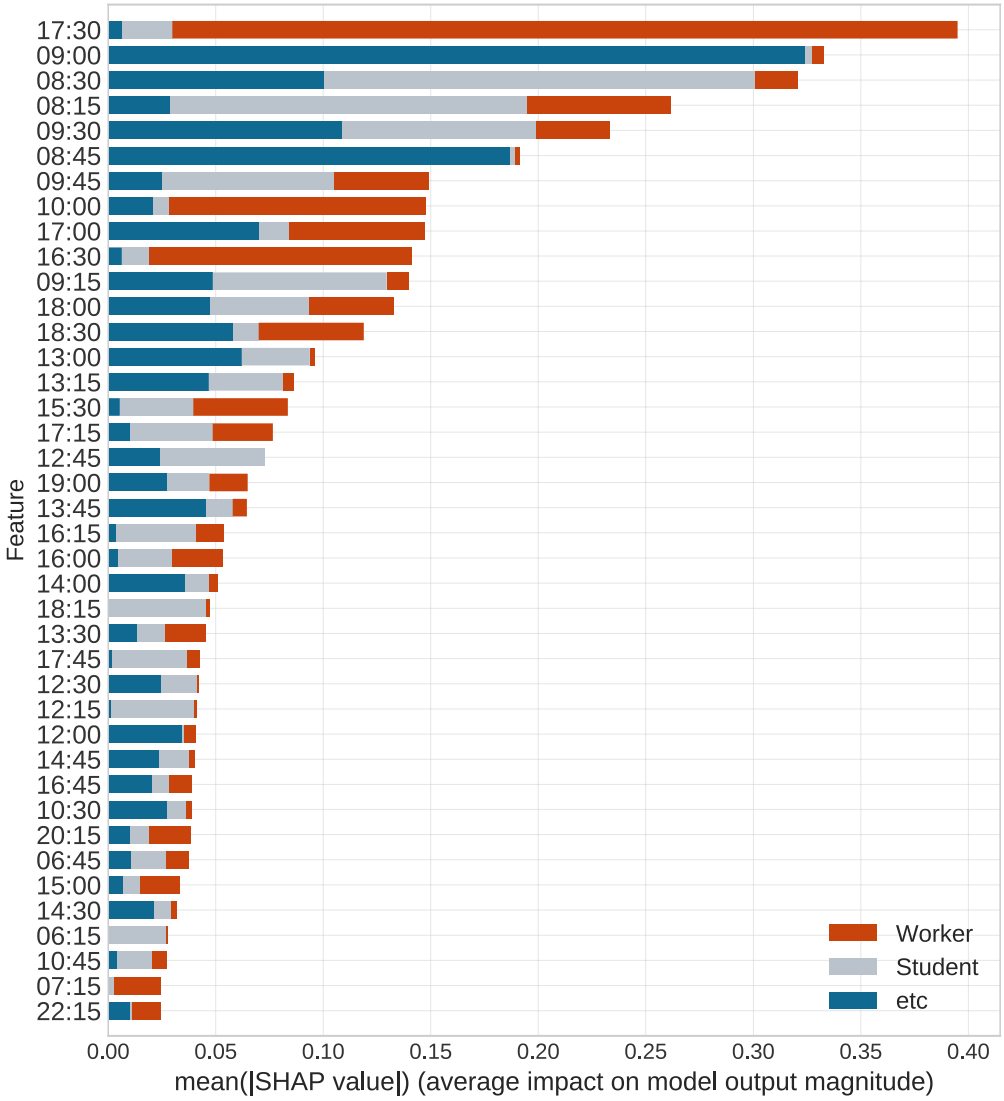

図 37　特徴量重要度上位 40 個

5. 3 人流データ前処理

5. 3. 1 長距離トリップ UID 除去

人流データには関東エリアに常住する人のみならず，関西などから旅行目的で訪れる（流入する）と想定されるユーザーや，出張が終わり東北へ帰宅（流出）すると想定されるユーザーなどの関東地方域外ユーザーも含む．そこで，1 トリップで 150 km 以上の直線距離を移動した UID を除去する．150 km の閾値については，以下の 2 つを参考に設定している．

第一に，通勤実態に関するアンケート調査（アットホーム株式会社，2018）[55]を参考にする．調査では都内勤務の会社員を対象とした場合，通勤時間の限界平均は 65 分であると明らかにしている．ここで，限界とは通勤時間が何分間まで身体的に耐えられるかということを示している．関東エリアにおいて在来線の表定速度が高い路線で約 60 km 前後であることから，通勤限界は少なくとも 150 km 未満といえる可能性が高い．

また，人流データ（6 月分）における最大移動距離ヒストグラム図 38 から，150 km 以上のトリップが著しく少ないことがわかる．そのため，本人流データでは 150 km は平均的なライフスタイルから逸脱した移動距離である可能性が高い．以上 2 点から関東地方の常住者を効率的に抽出するために，150 km を閾値として採用する．

この処理により，関東地方外に常住するユーザーを全て除去できる訳ではないが，続くフローで処理するデータの軽減を行うことができる．なお，検証においては関東エリアデータと空間結合するため，関東地方外に居住地が割り当てられているユーザーがいる場合でも検証時には影響しない．

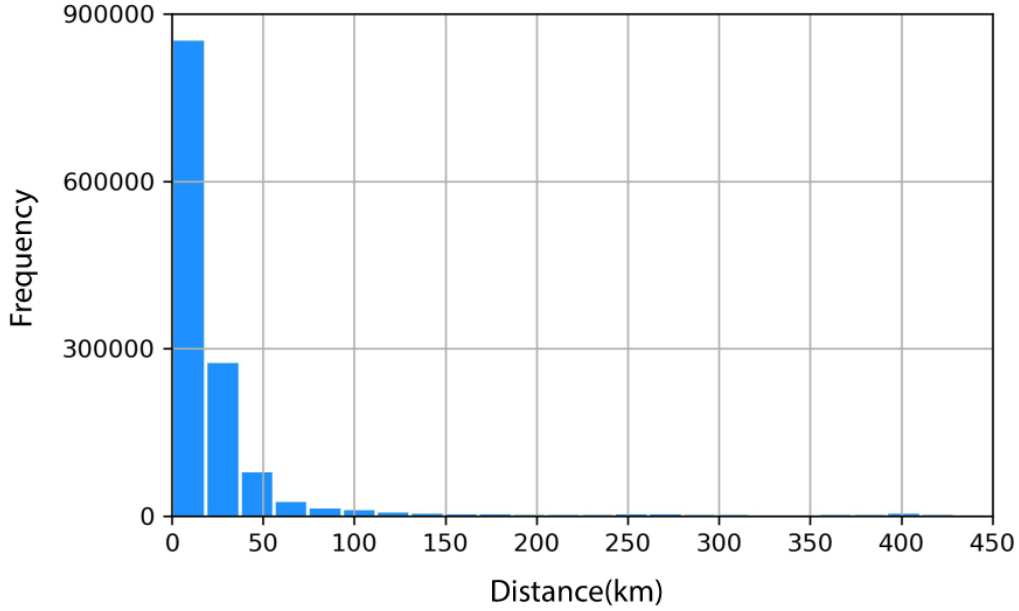

図 38　最大移動距離ヒストグラム

5. 3. 2 ログ不足 UID 除去

1 日におけるログが極端に少ない場合，モデルにおいて学習困難な状態に陥ることが考えられる．そこで，1 日 24 時間を 3 つのタイムフレーム（00 時 00 分 ～ 07 時 59 分，08 時 00 分 ～ 15 時 59 分，16 時 00 分 ～ 23 時 59 分）に分割し，各タイムフレームに 1 つもログがなかった場合，該当 UID のデータを省く処理を行う．

5. 3. 3 ポイント内挿

3.3 で述べた通り，人流データはスパースなデータであり，かつ個人情報保護の観点からバックグラウンド時の午前 1 時から 5 時（早朝時間帯とする）はデータ取得をしていない．そこで，前後の時間帯のポイントを参考にポイントの内挿を行う．また，早朝時間帯については PT データを元に「早朝時間帯」と「早朝時間帯を除く午前 0 時から 8 時のポイン

ト」は同じ地点にいる可能性が高いことから，午前 0 時から 8 時のポイントを参照して内挿する．以上のように，15 分毎にポイントを内挿し（線形補間），UID 毎にログ数が異なる人流データ内のデータ形式を揃える．

5.3.4 分散ポイント統一化

人流データに記録されている緯度経度は，衛星受信状況や GPS の精度，屋内などの環境要因から同一の緯度経度（同じ場所）にいたとしても，実際とは別の緯度経度を記録している場合がある．ここでは，これを「分散」と呼ぶことにする．整形された PT データと類似性を持たせるためには，同一の緯度経度にいたと思われる場合，分散ポイントを統一化する必要がある．

そこで，DBSCAN を用いて $eps = 0.365$，$minpts = 2$ のパラメータ設定でポイントのクラスタリングを行う．各ポイントが所属するクラスターの中心点を分散補正後のポイントとして割り当て，転移学習に適したデータに加工する．図 39 に具体例を示す．なお，可視化に用いるデータはダミーとして作成したものであり，特定の個人情報は含まない．また，アイコン色は所属するクラスターラベルを表しており，グレー色のアイコンはノイズとみなされた点であることを表している．図 39 から，それぞれ分散していた各ポイントがクラスター単位に集約されたことがわかる．

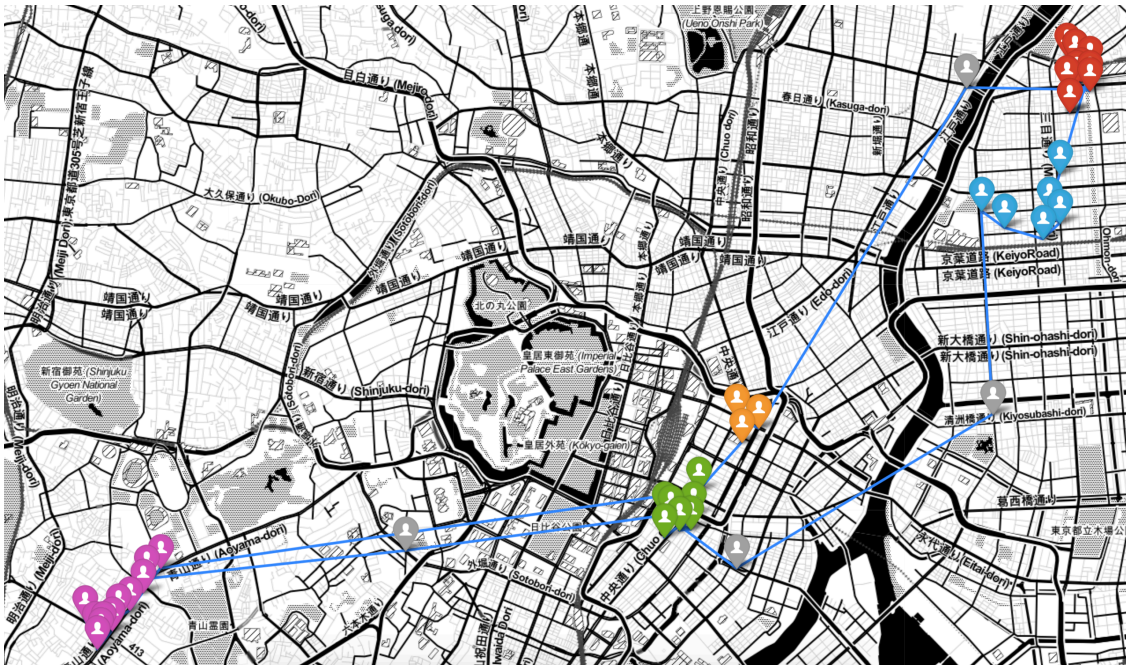

図 39　DBSCAN 実行例

5.4 居住地及び通勤通学地の推定手法

居住地推定について説明する．図 40 は 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の正解率を示している．日中は勤務地や学校などの居住地以外に，夜間は居住地に滞在するという一般的なライフスタイルに基づき，正解率が変化している．そのため早朝や深夜時間帯における正解率が高く，中でも 0 時から 6 時においては，いずれの時間も 96% 以上の正解率を示している（表 17）．

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．ただし，

この場合は夜間労働者を抽出することができない．本研究では，時刻別に仕事を行っている有職者率（30 分毎の平均行為者率）を表した図 41（出典：NHK 放送文化研究所，2015[d]）から夜間労働者の割合は 3% 前後と非常に低いこと，そして本研究の推定結果の活用は昼間労働者などの一般的なライフスタイルの人々を主にターゲットとしていることから，夜間労働者は特段考慮しないこととする．

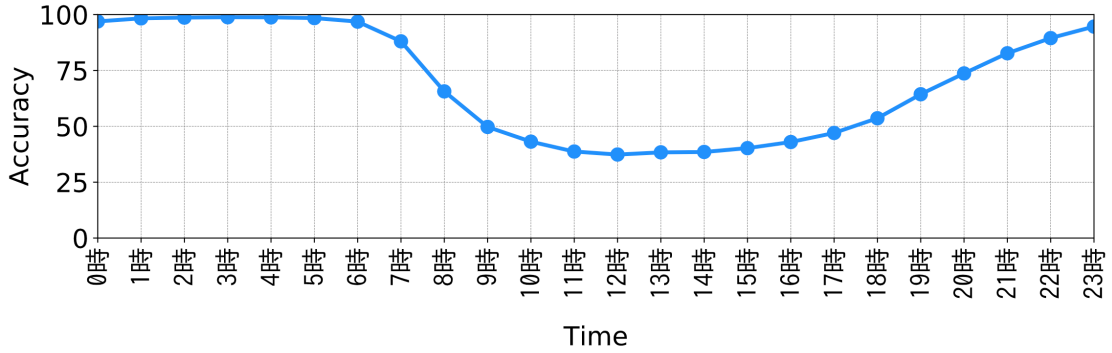

図 40　PT データ 時間帯別の居住地推定正解率

表 17　0 時から 6 時の居住地推定正解率詳細

時間帯	正解率（%）
0 時	96.95
1 時	98.31
2 時	98.66
3 時	98.82
4 時	98.75
5 時	98.45
6 時	96.85

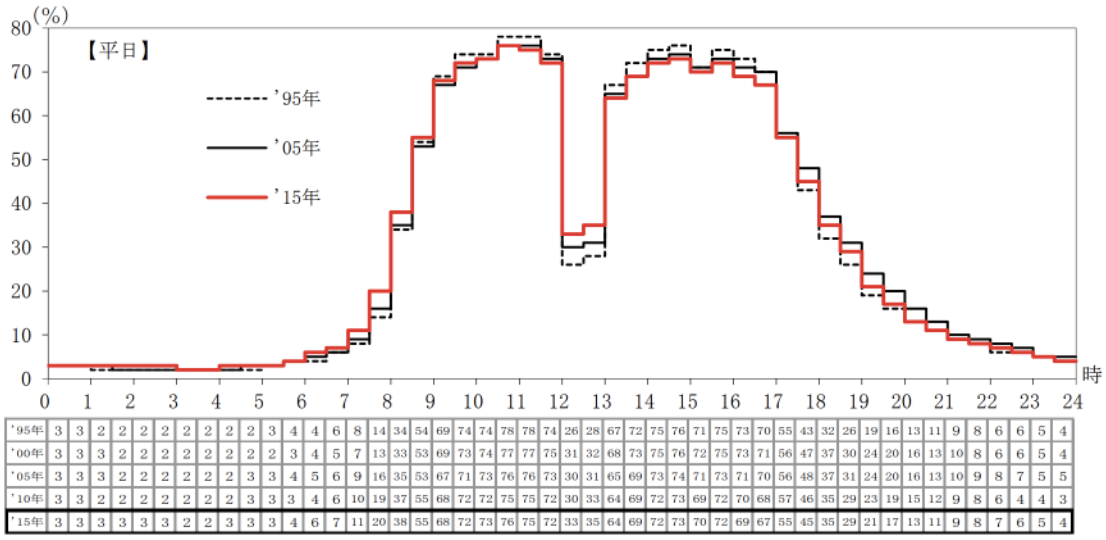

図 41　有職者の仕事行為割合（出典：NHK 放送文化研究所，2016[d]）

次に通勤通学地推定について説明する．5.1 で述べた通り，NHK 放送文化研究所による国民生活時間調査報告書（2015）を参考に，職業分類モデルによって就業者・学生に分類された UID に対し，居住地以外で最も滞在時間が長い地点（閾値 2 時間以上滞在）を通勤通学地として推定する．また，マルチな勤務地をもつ就業者にも同様な推定を行うため，1 日間で複数の勤務地がある場合でも，1 日で最も長く過ごした 1 地点を通勤地とみなすことになる．国勢調査においても，自動車運転従業者などの特定の場所で働かない人々に対して，所属している事業所 1 地点を便宜的に勤務地とみなしているため，検証用データを考慮した妥当な処理と考えられる．

5.5 推定と検証

5.5.1 居住地推定と検証

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．図 42 は推定の結果得られた居住地の空間分布（1 km グリッド毎）を示し，グラフの高さ及び色は分布度合いを表している．東京都や神奈川県エリアなどに多いが，局地的に集中しているエリアがあるわけではなく郊外にも分散しており，多様なエリアに推定居住地があることがわかった．

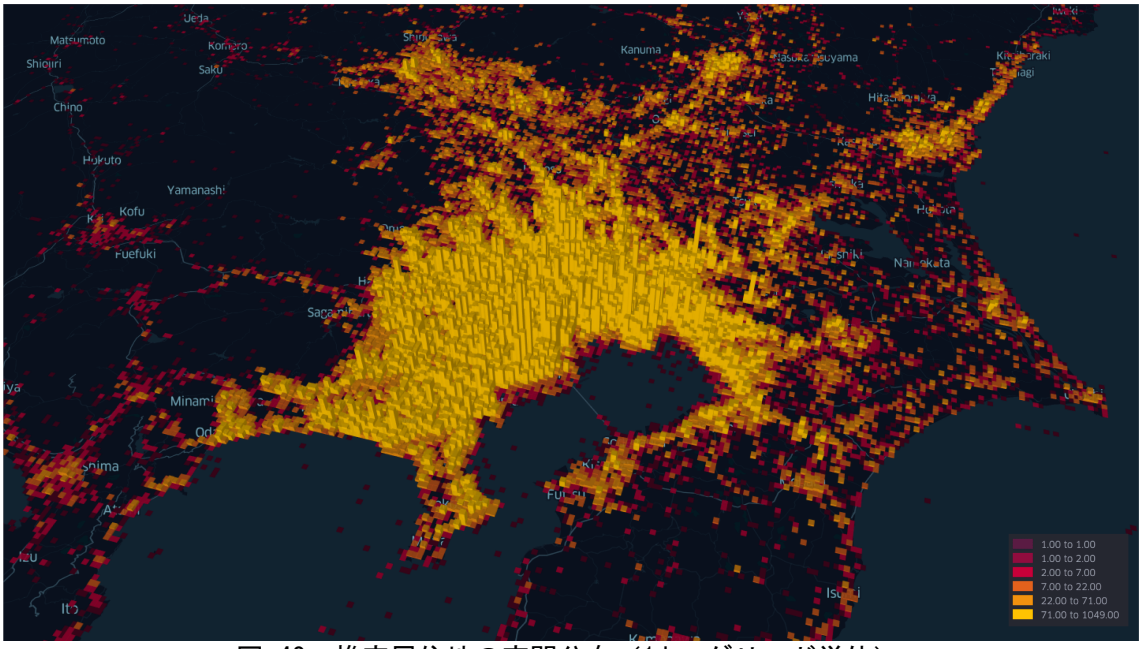

図 42　推定居住地の空間分布（1 km グリッド単位）

次に，推定夜間ユーザー数（推定居住地数）を夜間人口データによって検証する．評価指標には，式 (11) で表される相関係数を用いる．

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{11}$$

市区町村別の推定夜間ユーザー数と夜間人口データの相関図は図 43 の通りである．相関係数は 0.982 となり，強い相関を示した．相関係数は外れ値に弱い評価指標ではあるが，図 43 から大きく線形から外れたデータは見られない．よって，集計値による検証ではあるが，居住地推定は適切に行われたと判断し，フローの通り次に転移学習による職業属性分類モデルの適用を行う．

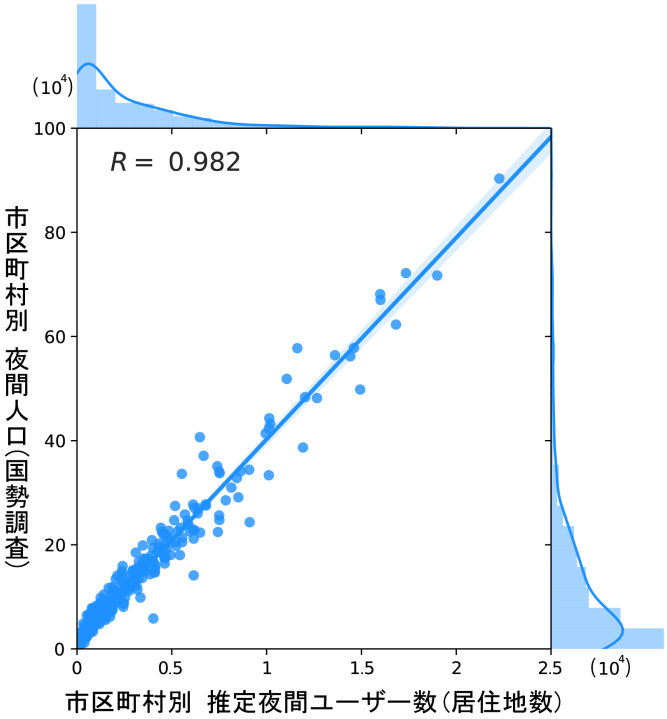

図 43　推定結果と夜間人口の相関図

5. 5. 2 職業属性分類モデル適用と検証

推定居住地を元に特徴量である時系列移動パターンを算出し，PT データによる学習済みモデルへ適用する．

結果として, 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．国民生活時間調査報告書（NHK 放送文化研究所，2015）[47] によると職業属性毎の割合は，就業者 55%，学生 12%，その他 32% である．この割合は全国を対象地域として行った統計調査によるものであり，本研究が対象としている関東地方とは厳密には異なるが参考として検証に用いる．

同調査の職業属性割合と本モデルの分類結果の割合を比較すると，差が見られることがわかった．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．具体的には，その他属性が多い傾向がある高齢者の割合が低いなどの影響である．高齢者のスマートフォン保有率は 図 44 のように年々上がってはいるものの，他の年代に比べると低い．そのため，人流データでも高齢者割合が低く，同時にその他属性も低くなってしまった可能性が要因の一つとして挙げられる．

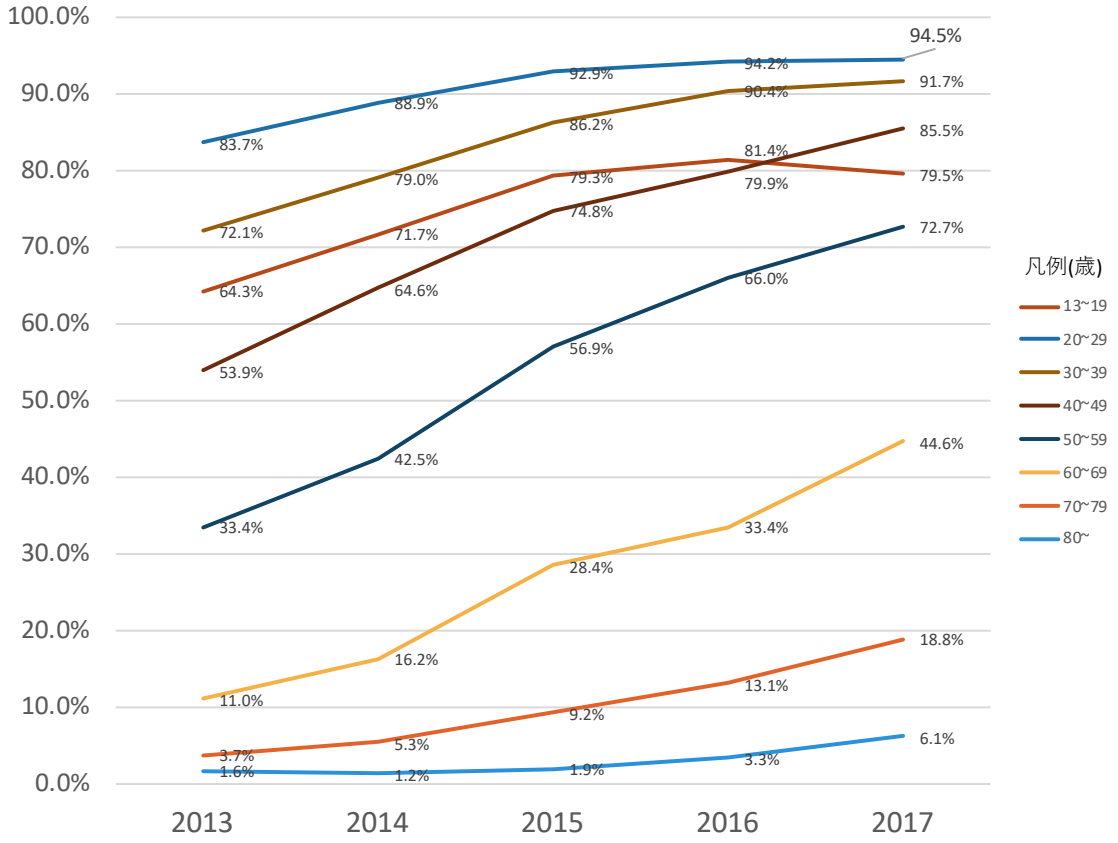

図 44　スマートフォンの年代別個人保有率推移
（通信利用動向調査（総務省，2013-2017）[56]を元に作成）

5. 5. 3 通勤通学地推定と検証

5.5.2 で就業者・学生と分類された UID に対して，通勤通学地推定を行う．推定結果の空間分布を図 45 に示す．図 45 は推定通勤通学地の空間分布（1 km グリッド毎）を示し，グラフの高さ及び色は分布の度合いを表している．居住地推定の分布と異なり，多くのユーザーの推定通勤通学地が港区や新宿区などのエリアに集中する一極集中の状態であることがわかった．

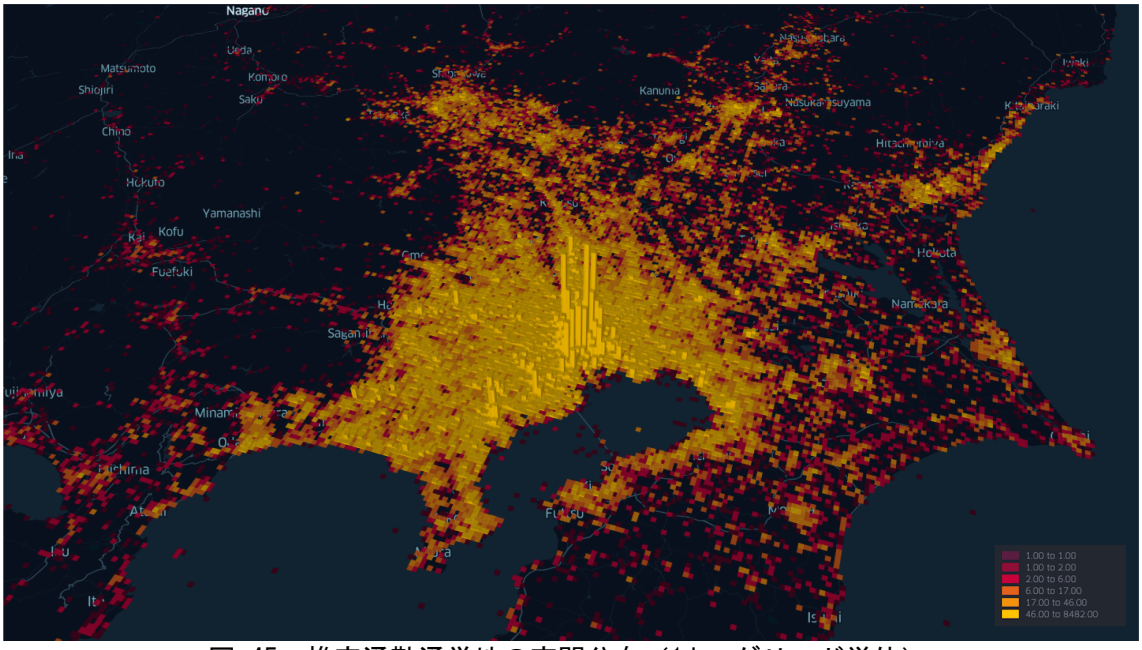

図 45　推定通勤通学地の空間分布（1 km グリッド単位）

次に，推定した通勤通学地結果の検証を行う．人流データには居住地及び通勤通学地の正解ラベルがないため，検証用データとして市区町村別の人口統計データと町域別の昼間人口統計データを用いる．評価指標には相関係数を利用する．

式 (2) によって推定昼間ユーザー数を算出し，昼間人口統計データとの相関を測る．結果として相関図は図 46 の通りとなった．相関係数は 0.892 となり，強い相関を示した．ただ，線形から大きく外れた市区町村がいくつか見られることがわかった．これら線形から大きく外れたエリアのラベルを見ると，図 46 の右上にあるエリアは港区・中央区・千代田区であった．この 3 区はオフィスや商業施設が多いエリアであり，人流データにこれらのエリアで働く人々が他に比べて多いことや，国勢調査に回答された通勤地ではなく一時的に営業トリップでこれらのエリアに訪問している人々が多いことなどが要因として考えられる．

一方，図 46 の左上にあるエリアは世田谷区・大田区であり，前述の 3 区と大きく異なる住宅街の印象が強いエリアとなった．このことから，本人流データにはエリアによるユーザーの偏りが見られる可能性が考えられる．つまり，人流データ内において，世田谷区などの住宅街エリアで勤務するユーザーよりも，港区などのオフィスエリアで勤務するユーザーの割合が国勢調査の割合と比べても多いことが示唆される．

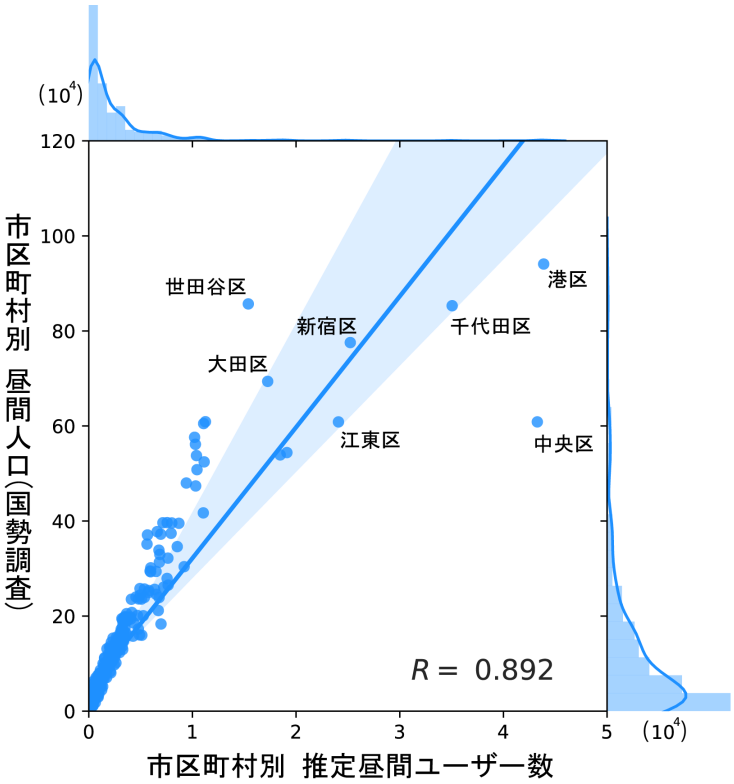

図 46　推定結果と市区町村別昼間人口の相関図

続いて同様に，町域別の人口統計データを用いた検証を行う．推定昼間ユーザー数と昼間人口データの相関図は図 47 の通りである．相関係数は 0.899 となり，強い相関を示した．

ただしここでは，推定昼間ユーザー数上下 0.01% の外れ値を除いている（除去した町域数: 2）．これはデータの特性上，東京都港区東新橋エリアのとある企業と思われる地点に通勤するユーザーが多く，エリアによる偏りがみられたためである．なお，除去前の相関係数は図 48 に示すように 0.777 となった．相関係数は外れ値に弱い指標であり，散布図と被観測個体の特殊性から外れ値として判断できる場合は除去するに値すると言われている（坂田，2017[57]）．そこで，一部町域を散布図及びデータの特性上から外れ値であると判断し除去した．

市区町村別と同様に，町域別スケールでも強い相関を示す結果となり，以上から両スケールにおける本手法の有効性が確かめられた．

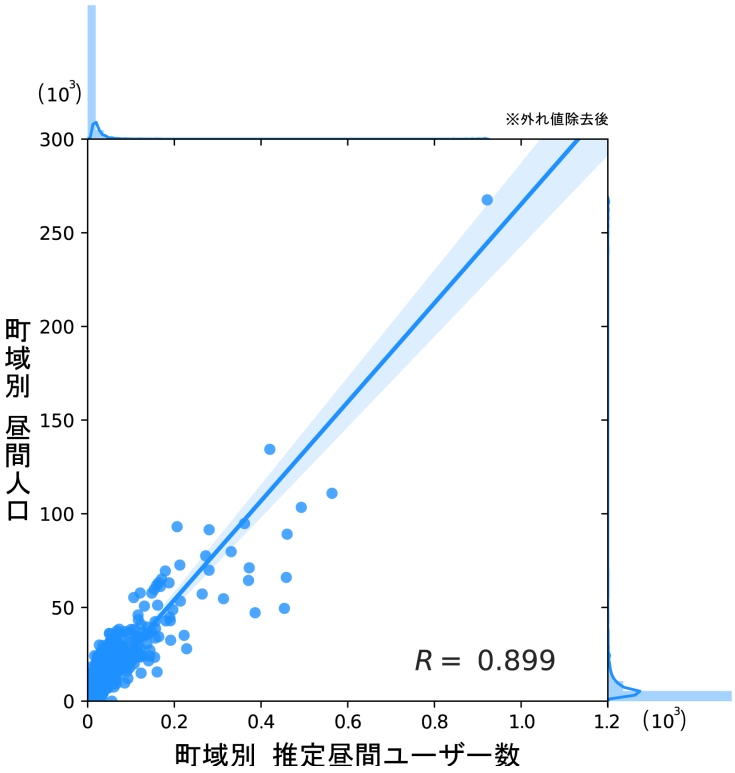

図 47 推定結果と町域別昼間人口の相関図（外れ値除去後）

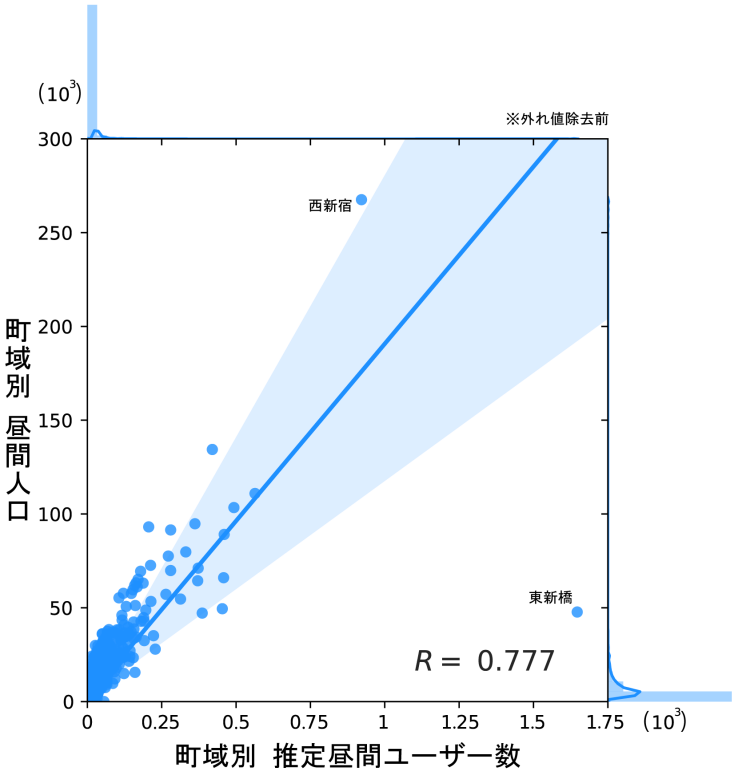

図 48 推定結果と町域別昼間人口の相関図（外れ値除去前）

第 6 章 推定結果の活用展望

居住地及び通勤通学地推定結果は，交通需要モデル構築など多くの活用が考えられる．ここでは，活用の例を挙げることで本研究による推定結果の活用展望を示す．

本研究で得られる推定結果から，通勤地における滞在時間，つまり勤務時間を把握することができる．図 49 に，今回得られた東京都心部における就業者の勤務時間の町域別 空間分布を示す．これは，各町域で勤務するユーザーの勤務時間の平均値を利用したもので，可視化においてサンプル数が 50 未満の町域は除外している．

図 49 から，山手線内の新宿や港区といったエリアおける勤務時間が長いことがわかる．これらのエリアはオフィスが多い都市部のエリアであり，残業を含む仕事をする人が多いことが要因として考えられる．なお，中央に位置する塗りつぶしのないエリアは皇居である．一方で，勤務時間の短い町域も存在する．これは本推定では就業者としてアルバイトやパート従業員も含んでいるためであり，5 時間以下の勤務という短時間勤務な町域も存在する．

しかし，図中の郊外部においても 9 時間近い勤務時間を平均値としている町域が存在する．これは本来居住地である地点を誤って通勤地として推定している可能性が考えられる．

本手法では一部改善の必要があるが，このように勤務時間を把握することができる．これにより例えば，行政機関が人々の労働実態を短期間かつ低コストで把握することが可能となる．将来的には，高精度な測位が可能な衛星システムによって位置誤差が小さいデータを収集することができれば，ビル毎のよりミクロなスケールでの労働実態を把握することも可能であると考えられる．

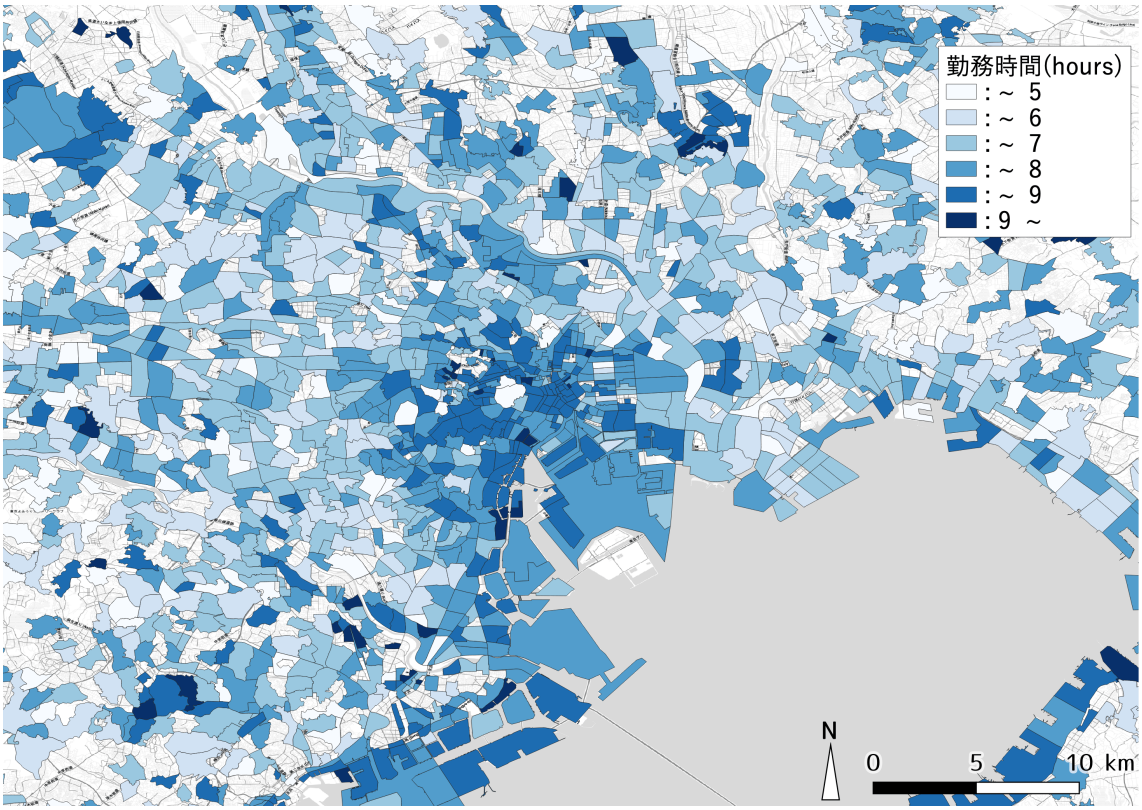

図 49　勤務時間の町域別空間分布

その他，通勤時間に関する調査・検討にも活かすことができると考えられる．UNSD（2011）[58]によると OECD 加盟国の中で通勤者の通勤時間を比較したところ，日本は 4 番目に通勤時間が長い国であると明らかにされている（図 50）．

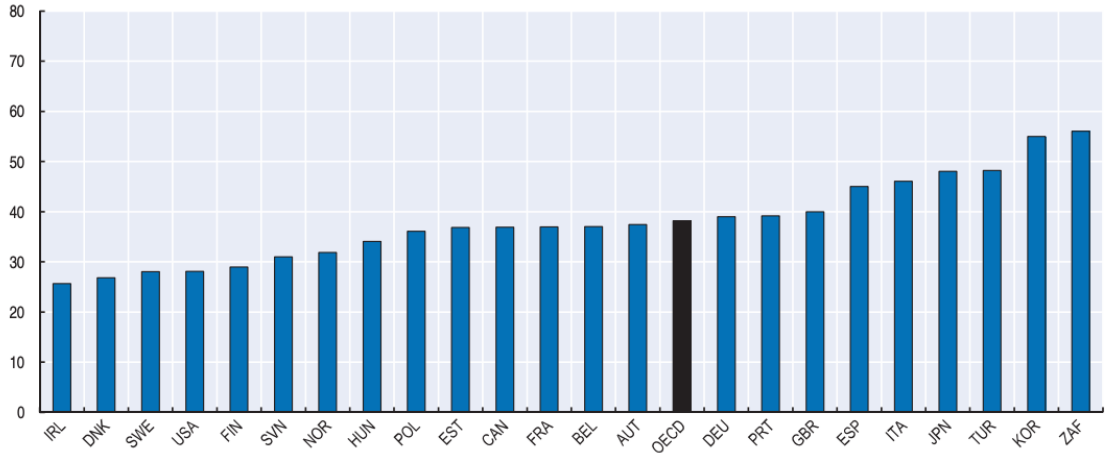

図 50　各国の通勤時間比較（出典：UNSD, 2011[e]）

それに加え，首都圏内の鉄道は通勤/帰宅時間帯の混雑が問題となっており，国土交通省（2018）[59]が発表した資料によると，首都圏の 11 路線が鉄道混雑率は 180% の混雑率を超えていることが明らかになっている．180% とは「折りたたむなど無理をすれば新聞を読める程度」の鉄道混雑率を示しており，この数値を超えた場合，体が触れ合い圧迫感を感じながら乗車すること（200%）となる．また，Google Maps における混雑度合いのデータから明らかにした分析結果でも示されている通り[60]，首都圏路線の混雑率は世界的に見ても高いといえる．

以上，通勤時間及び混雑率の観点から，通勤行動をより深く把握する必要がある．今後は，本研究で推定した居住地及び通勤通学地を利用し，人流データから見た通勤行動のより深い把握を行うことが期待される．

第 7 章 結論

7.1 本研究の成果

本研究では，1 日という短期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を用いて，職業属性を考慮した居住地及び通勤通学地推定手法の構築を行った．検証では，国勢調査による市区町村別の昼間人口，及び夜間人口データと強い相関を示した．これは職業属性を行わない場合よりも相関が強く，本研究で提案する職業属性による精度向上が見られることがわかった．また，町域別の昼間人口データとも強い相関を示し，ミクロ寄りのスケールでも居住地及び通勤通学地推定が可能であることがわかった．

このように，本研究では職業属性推定を事前に行うことで居住地及び通勤通学地推定の精度が向上することを示すことができた．同時に，PT データから得られる知識をスマートフォンアプリケーションから得られる人流データに転移する転移学習の有効性も示すことができた．

7.2 本研究の課題と展望

本研究における課題は主に 3 つある．第一に，推定結果に人流データ内の年代やエリアによると考えられる偏りが見られたため，これらを考慮した補正を行う必要がある．人流データはスマートフォンから得られるため，近年は高齢者のスマートフォン保有率が上がってはいるものの，他の年代に比べて高齢者割合が低いことは事実である．また，エリアに関しても都心のユーザーに有益なアプリケーションの場合，エリアに偏りが生じる可能性は高い．そこで，重み付けなどのモデルレベルでの対処，もしくはデータに拡大係数を適用するデータレベルでの対処をする必要がある．

第二に，職業属性分類モデルの改善である．まずは精度についてである．本研究では特徴量として時系列移動パターンのみ用いているが，両データの質の差を十分に考慮することで特徴量を増やすことができる．PT データが必ずしも正確な緯度経度を示していないということを念頭に置き，適切な特徴量を追加することで職業属性推定モデルにおける精度の向上がみられる可能性がある．次に検証及び最適化についてである．本研究では，処理時間の都合上，十分なパラメータチューニングや交差検証を行うことができなかった．そのため，精度や検証の改善のためには Optuna や HyperOpt によるパラメータの最適化や K-分割交差検証を行う必要がある．

第三に，人流データにおける位置誤差の補正手法や内挿手法を改善し，よりミクロなスケールでの居住地及び通勤通学地の推定を行うことである．内挿について，本研究では線形補間を用いているが，PT データは交通ネットワークを元に内挿されたデータである．両ドメイン間の転移学習をより適切に行うためには，人流データにも交通ネットワークに基づく内挿を行うことが望ましい．

これらを改善することで，より精度の高い居住地と通勤通学地の推定に繋がる．そして，活用フェーズとして生活スタイルを考慮した上で時空間的な通勤ニーズの把握，交通需要モデル構築などにつなげることができる．

また，本研究による手法はエリアに依存する特徴量を利用していないことから，今回対象地域とした関東地方以外のエリアにおける人流データに対しても，同手法を適用できる可能性が高い．その他，一般的な UID が固定された人流データに対して，1 日毎にユーザーの職業属性，居住地及び通勤通学地を推定し，最後に UID を元に集約することで，曜日毎もしくは月毎のライフパターンが抽出できる可能性もある．

参考文献

[1] 日経 xTECH，2006，“来年 4 月、すべての携帯電話に GPS”，https://tech.nikkeibp.co.jp/it/pc/article/NPC/20060925/248858/，アクセス日：2020-01-10.

[2] 日経クロストレンド，2018．“データは隠さず売る 人流データは月 100 万円から”，https://xtrend.nikkei.com/atcl/contents/technology/00003/00001/，アクセス日：2020−01−11.

[3] Washington Post, 2019. ”Congress should make it harder for cellphone carriers to sell your location data”. https://www.washingtonpost.com/opinions/congress-should-make-it-harder-for-cellphone-carriers-to-sell-your-location-data/2019/01/09/93602366-1451-11e9-803c-4ef28312c8b9_story.html，アクセス日：2020−01−11.

[4] Wei, R., Tian, H., and Shen, H., 2018, Improving k-anonymity based privacy preservation for collaborative filtering, Computers & Electrical Engineering, 67, 509-519.

[5] モバイル空間統計．https://mobaku.jp/，アクセス日：2020-01-11.

[6] ブログウォッチャー．https://www.blogwatcher.co.jp/，アクセス日：2020-01-11.

[7] 株式会社 Agoop．https://www.agoop.co.jp/，アクセス日：2020-01-11.

[8] 齋藤雅行，山岸敦，瀧口純一，浅里幸起，2016．準天頂衛星による高精度測位システムの紹介，MSS 技報，Vol.26-2.

[10] 共同通信，2019．”中国版 GPS の精度、誤差 5m に 運用開始から 1 年”，https://headlines.yahoo.co.jp/hl?a=20191227-00000125-kyodonews-int，アクセス日：2020-01-11.

[11] 総務省，2013．”社会・産業の発展に寄与するモバイル空間統計”，https://www.soumu.go.jp/main_content/000261299.pdf，アクセス日：2020-01-11.

[12] 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 2011, Pervasive Computing, pp. 133-151.

[13] 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).

[14] 森川高行，山本俊行，三輪富生，薄井智貴，2015．スマートフォン行動データとコンテクストデータを活用した活動・交通ログ自動生成手法，科研費 研究成果報告書.

[15] 総務省統計局，2010．“国勢調査の結果で用いる用語の解説”，https://www.stat.go.jp/data/kokusei/2010/users-g/pdf/04.pdf，アクセス日：2020-01-11.

[16] Mahmud, J., Nichols, J. and Drews, C., 2012. Where Is This Tweet From? Inferring Home Locations of Twitter Users, Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, pp. 511-514.

[17] 総務省，2014．“インターネットリテラシーの重要性”，
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h26/html/nc143120.html，
アクセス日：2020-01-11.

[18] Twitter, 2019. “Twitter Support”，
https://twitter.com/TwitterSupport/status/1141039841993355264，アクセス日：2020-01-11.

[19] 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.
https://agile-online.org/images/conference_2019/documents/short_papers/82_Upload_your_PDF_file.pdf，アクセス日：2020-01-10.

[20] 総務省統計局，2015．平成 27 年国勢調査，
https://www.stat.go.jp/data/kokusei/2015/kekka.html，アクセス日：2020-01-11.

[21] 横浜市，2019．“横浜市記者発表資料 第 6 回東京都市圏パーソントリップ調査の集計結果概要について”，https://www.city.yokohama.lg.jp/kurashi/machizukuri-kankyo/kotsu/toshikotsu/PT.files/R1.11.27press.pdf，アクセス日：2020-01-11.

[22] 東京都市圏交通計画協議会，2009．第 5 回東京都市圏パーソントリップ調査（交通実態調査）の集計結果について，
https://www.tokyo-pt.jp/static/hp/file/data/091130.pdf，アクセス日：2020-01-11.

[23] 総務省統計局，2015．平成 27 年国勢調査 抽出速報集計結果 結果の概要，
https://www.stat.go.jp/data/kokusei/2015/kekka/pdf/gaiyou1.pdf，アクセス日：2020-01-11.

[24] 金杉洋，2014．人の流れ研究会拡大版 人の流れデータチュートリアル 人の流れ研究会拡大版，https://pflow.csis.u-tokyo.ac.jp/wp-content/uploads/20140724.pflow-ws-tutorial.pdf，
アクセス日：2020-01-08.

[25] 金杉洋，関本義秀，瀬戸寿一，柴崎亮介，2014．“人の流れデータの整備状況と提供サービスについて”，人の流れ研究会拡大版．
https://pflow.csis.u-tokyo.ac.jp/wp-content/uploads/pflow_workshop11_poster10.pdf，アクセス日：2020-01-08.

[26] 東京大学 空間情報科学研究センター，“JoRAS”，https://joras.csis.u-tokyo.ac.jp/，アクセス日：2020-01-10.

[27] マイクロジオデータ研究会，2017．Agoop による位置情報分析について，
http://microgeodata.jp/contents/pdf/mgd11/mgd11_agoop.pdf，アクセス日：2020-01-08.

[28] G 空間情報センター，2018．流動人口データの活用について，
https://bit.ly/2ZGfNsk，アクセス日：2020-01-08.

[29] 株式会社ゼンリンデータコム，“混雑統計®”.
http://www.zenrin-datacom.net/business/congestion，アクセス日：2020-01-08.

[30] G 空間情報センター，2015．“ポイント型流動人口データ 注意事項（Q&A）”，
https://www.geospatial.jp/ckan/dataset/4889dcd9-18d8-4553-a436-365199d5bdf4/resource/e49c1ed0-597f-4807-b216-15193b51b3b1/download/cautionpoint.pdf，
アクセス日：2020-01-08.

[31] Alessandretti L., Sapiezynski P., Lehmann S. and Baronchelli A., 2017. Multi-scale spatio-temporal analysis of human mobility, PLOS One, 12(2), e0171686.

[32] 国勢調査，2015．平成 27 年国勢調査 従業地・通学地集計，
http://www.stat.go.jp/data/kokusei/2015/kekka/jyutsu1/pdf/gaiyou.pdf，アクセス日：2020-01-10.

[33] MacQueen, J. B., 1967. Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, pp. 281–297.

[34] Pelleg, D. and Moore, A., 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters, In Proceedings of the 17th International Conf. on Machine Learning, pp. 727-734.

[35] Arthur, D. and Vassilvitskii, S., 2007. k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, pp. 1027–1035.

[36] Ester, M., Kriegel, H., Sander, J., and Xu, X., 1996. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise, KDD'96 Proceedings of the Second International Conference, pp. 226-231.

[37] Medium, 2019. “Let’s cluster data points using DBSCAN”,
https://medium.com/@agarwalvibhor84/lets-cluster-data-points-using-dbscan-278c5459bee5,
アクセス日：2020-01-10.

[38] scikit-learn. “Comparing different clustering algorithms on toy datasets”,
https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py，アクセス日：2020-01-10.

[39] Ankerst, M., Breunig, M., Kriegel, H. and Sander, J., 1999. OPTICS: Ordering Points To Identify the Clustering Structure, Proc. ACM SIGMOD’99 Int. Conf. on Management of Data, Vol 28-2.

[40] Breiman, L., 2001. RANDOM FORESTS, Machine Learning, October 2001, Vol 45-1, pp. 5-32.

[41] Ke, G., Meng, Q., Finley, T., Wang, T., Wei Chen, Ma, W., Ye, Q. and Liu, T., 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree, 31st Conference on Neural Information Processing Systems.

[42] LightGBM. "Welcome to LightGBM's documentation!", https://lightgbm.readthedocs.io/en/latest/index.html#, アクセス日： 2020-01-10.

[43] Kaggle. https://www.kaggle.com/, アクセス日： 2020−01−10.

[44] Hochreiter, S. and Schmidhuber, J., 1997. Long Short-Term Memory, Neural Compu-tation Vol 9-8, pp. 1735-1780.

[45] Cho, K., Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y., 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1724-1734.

[46] Chung, J., Gulcehre, C., Cho, K. and Bengio, Y., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, arXiv:1412.3555 [cs.NE].

[47] 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, アクセス日： 2020-01-05.

[48] 西村隆宏，秋山祐樹，金杉洋，Horanont，T.，柴崎亮介，関本義秀，2014. 人の流れデータセットを用いたデモグラフィック属性の推定及び GPS データへの適用可能性に関する研究，地理情報システム学会講演論文集，Vol 23.

[49] Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W., 2002. SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16, pp. 321-357.

[50] Pan, S. J. and Yang, Q., 2008. A Survey on Transfer Learning, Technical Report HKUST-CS08-08, Dept. of Computer Science and Engineering, Hong Kong Univ. of Science and Technology.

[51] 神嶌敏弘，2010. 転移学習，人工知能学会誌，25 巻 4 号.

[52] 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.

[53] Google Maps Platform, Google Cloud, https://cloud.google.com/maps-platform/?hl=ja, アクセス日： 2020-01-10.

[54] Lundberg, S., Erion, G. and Lee, S., 2018. Consistent Individualized Feature Attribution for Tree Ensembles, arXiv:1802.03888 [cs.LG].

[55] アットホーム株式会社，2018, "「電車通勤実態」調査", https://athome-inc.jp/wp-content/uploads/2018/06/2018061301.pdf, アクセス日： 2020-01-10.

[56] 総務省，2013-2017，通信利用動向調査，
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h30/html/nd142110.html，
アクセス日：2020-01-10.

[57] 坂田綾香，2017．記述統計と確率変数・確率分布，
https://www.ism.ac.jp/~ayaka/2017_gairon_1.pdf，アクセス日：2020-01-10.

[58] UNSD, 2011. How's Life? MEASURING WELL-BEING,
https://unstats.un.org/unsd/broaderprogress/pdf/How's%20life%20-%20Measuring%20well-being.pdf，アクセス日：2020-01-10.

[59] 国土交通省，2018．東京圏で混雑率 189%超の路線が 12 路線から 11 路線へ，
https://www.mlit.go.jp/report/press/tetsudo04_hh_000076.html，アクセス日：2020-01-10.

[60] CNN. co. jp，2019．"世界一混雑がひどい列車は？　中央線 7 位、舎人ライナー9 位"，
https://www.cnn.co.jp/travel/35139401.html，アクセス日：2020-01-10.

引用文献

[A] 国勢調査，2015．平成 27 年国勢調査 従業地・通学地集計，
http://www.stat.go.jp/data/kokusei/2015/kekka/jyutsu1/pdf/gaiyou.pdf，アクセス日：2020-01-10.

図表出典元

[a] 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.
https://agile-online.org/images/conference_2019/documents/short_papers/82_Upload_your_PDF_file.pdf，アクセス日：2020-01-10.

[b] Gelzinis, A., Verikas, A. and Vaiciukynas,E., 2014. Exploring sustained phonation recorded with acoustic and contact microphones to screen for laryngeal disorders, 10.1109/CICARE.2014.7007844, pp. 125-132.

[c] Greff, K., Srivastava, R., Koutník, J., Steunebrink, B. and Schmidhuber, J., 2017. LSTM: A Search Space Odyssey, IEEE Transactions on Neural Networks and Learning Systems, Vol 28-10, pp. 2222-2232.

[d] 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，アクセス日：2020-01-05.

[e] UNSD, 2011. How's Life? MEASURING WELL-BEING,
https://unstats.un.org/unsd/broaderprogress/pdf/How's%20life%20-%20Measuring%20well-being.pdf，アクセス日：2020-01-10.

謝辞

本研究を行うにあたり，多くの方々にご指導．ご協力頂きました．心より感謝申し上げます．

指導教員である柴崎亮介先生には研究内容や論文執筆にあたり，丁寧なご指導をして頂きました．また，研究やプログラミング教育，学会参加などの環境面も整えていただいたこと，誠に感謝申し上げます．柴崎先生のお話は一語一語勉強になることばかりで，今でも TED × Tokyo などの映像をよく見ています．講義内容はもちろん，他の学生の発表に対するコメント・質問，あらゆるモノ・コトに対する視点が魅力的で，いつも聞いていて刺激を受けていました．

副指導教員である日下部貴彦先生には，研究テーマ設定の相談や梗概の添削をしていただきました．交通が専門で PT 調査に精通する日下部先生のアドバイスは的確であり，大変貴重なものでした．また，研究に息詰まっていた年明けの打ち合わせにて，本研究を興味深く聞いていただけたことはとても励みになりました．

秋山祐樹先生には，普段から研究テーマ相談や細かい論文添削をして頂き，お忙しい中大変お世話になりました．MGD（マイクロジオデータ）な皆さんと定期的に打ち合わせを行って頂いたことで，日々進捗を確認しながら研究を進めることができました．秋山先生とは，AGILE での場面が印象に残っています．初めての国際会議発表前に緊張気味な私を励まし，夜には発表練習もしていただいたことで，当日は上手く発表することができました．また，マイクロジオデータ研究会や議員会館での報告会など，貴重な場に参加する機会をいただきました．

小川芳樹先生には，打ち合わせや，時には忘年会にて多くのアドバイスを頂くことができました．常に論文を出し，若くして優秀な研究者の姿を見ることができました．いつも冗談を交えながら学生に近い視点でお話しいただき，時には研究以外にもプライベートな話も聞けてとても楽しかったです．オンもオフも楽しまれている姿がとても印象的でした．

宮澤聡特任研究員には，大変お世話になりました．学会提出前の M1 夏，頻繁に Slack しては，丁寧なご指導いただけたことをよく覚えています．それ以降も，研究のテーマ設定や手法検討，コーディングにおける工夫の仕方などアドバイス頂き，本当にお世話になりました．学会での対応など，時には厳しく指導頂き，自分の発表や質疑対応を見直す機会を頂きました．修士論文で苦戦した時も宮澤さんにアドバイスを伺いにいくと，いつも時間を空けて丁寧に指導していただきました．たまの雑談や同室の日下部研究室 三谷さんも交えて話す時間はリフレッシュにもなりました．宮澤さんのご指導なしでは，研究を進めることが難しかったと思います．語学が堪能で，ロジカルであり，そしてエンジニアリングスキルも持っている宮澤さんに近づけるよう，社会人になっても日々勉強し続けたいと思います．

日野智至先生にはプログラミング関連のご指導をしていただきました．それにより，プログラミングスキルが未熟であった入学時から，大きくスキルを伸ばすことができました．今では，毎回出される課題に苦戦していた頃が遠い遠い昔のように感じます．そして，スキルを伸ばせたことによって，かねてより志望していた企業・職種にも内定を頂くこともできました．

研究室の同期であるヨウ君，シンさん，種村君には公私ともにお世話になりました．学会に向けて頑張ったり，時には一緒に旅したりと，共に充実した時間を過ごすことができました．苦しい時にも 3 人がいることで乗り越えることができました．ヨウ君とは，普段からチャットで頻繁に連絡をとっていました．研究室のみならずアルバイト先が同じであるなど共通点が多く，プログラミング関連やプライベートの話もし，とても仲良くしてもらいました．シンさんはキプロスでの国際会議を一緒に参加したことが印象に残っています．初めての国際会議発表を終えた後，一緒にキプロスを回ることができて良い思い出になりまし

た．シンさんは来年度から職種が同分野なので，また情報交換できれば嬉しいです．種村君とは，研究における不安なことを打ち明けて話すことができました．その他，同じインターン先に行くなどの共通点もあり，研究のみならず色々な話をすることができました．また，種村君の運転で柴崎研の皆と四国を回り，とても楽しい思い出を作ることができました．私達は後輩と先輩がおらず，同期だけという特殊な状況でしたが，3 人が同期で本当に良かったと思っています．社会人になっても共に切磋琢磨し，また楽しい話をできればと思います．

スペースシンタックス・ジャパン株式会社 代表取締役 高松誠治さんには，学外においてお世話になりました．GIS や都市計画を始めとした知識・スキルを身につけることができました．また，昨年は金沢 21 世紀美術館にて展示企画する際，人物検出システム設計を担当させて頂きました．企画段階の打ち合わせから展示まで参加させて頂いたことで，大変貴重な経験ができました．焦りながらも楽しく，遅くまで美術館内やホテルで必死に作業したのは良い思い出となっています．

その他，お名前を挙げさせていただいた方々以外にも，東京大学 空間情報科学研究センターの先生/研究員や秘書の方々，アルバイト・インターンシップ先や学会でお声掛け頂いた企業の方々など，多くのお力添えがあったお陰で本研究を完成させることができました．心から感謝申し上げます．

最後に，これまで長い学生生活を常に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．

2020 年 1 月 20 日

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 修士課程
小林稜介
koba.csis@gmail.com