

ランダムフォレスト法を用いたシェアサイクルシステム利用目的の推定

Estimating the purpose of using the share cycle system using the random forest method

室蘭工業大学建築社会基盤系学科 ○学生員 野崎脩斗(Shuto Nozaki)
 室蘭工業大学大学院工学研究科 学生員 坂本信 (Shin Sakamoto)
 室蘭工業大学大学院工学研究科 正員 有村幹治(Mikiharu Arimura)

1. はじめに

現在、我が国では、コロナ感染対策、地球温暖化対策、コンパクトシティの形成、国民の健康増進等様々な観点から自転車利用が促進されている。中でも、放置自転車の抑制、モビリティ向上による都市内部の交流の活性化等を目的とした¹⁾シェアサイクルに注目されている。しかし、利用目的が明確ではなく、今後の発展や改善を考えると、利用目的の把握が必要である。

札幌のシェアサイクルであるポロクルは、都市内部のスムーズな移動、にぎわい創出、CO₂削減、路上駐輪の削減などへの寄与を目的²⁾として2011年にスタートし、札幌中心部に設置された48カ所のポート(図-1)で自転車の貸し借りを行うことができ、通勤・食事・ショッピング・観光などで利用されている。特に2019年度からは24時間利用可能になり、自転車に電動アシストが付くなど、利便性が増した。

このようなシェアサイクルに関する先行研究として、南部ら³⁾はシェアサイクルの利用データを使用し、我が国のCOVID-19の影響分析を行っている。また、橋本ら⁴⁾はシェアサイクルのアンケートデータを用いて、シェアサイクルの導入がまちの魅力に影響及ぼすことを示した。しかし、アンケートデータと利用データを組み合わせて目的等の情報を導き出す研究はまだ見られていない。そのため本研究では、上記で述べたポロクルを対象に、2020年度に実施した利用者アンケートデータを用いて、通勤や買い物等の利用目的種別の利用動向を推定できるモデルを機械学習の一つであるランダムフォレストを用いて構築する。そのうえで構築したモデルを2020年度の全利用データに適用し、従来把握することが困難であったサイクルシェアの利用目的別のトリップ数の推定を試みる。

2. データの概要

2.1 ポロクル2020年度全利用データ

ポロクルの2020年度全利用データは、多数の情報を含んだ323,766万パターンのビックデータであり、利用者数は19,090人であった。具体的には、ユーザーID、ユーザー種別、利用プラン、請求項目、請求額、利用開始日時、返却日時、利用時間(秒)、利用開始ポート名、返却ポート名、返却種別、車両情報が含まれている。なお冬季は積雪により営業が困難であるため、4月～10月までのデータである。本研究ではその中のユーザーID、請求項目、利用開始日時、返却日時、利用時間(秒)、利用開始ポート名、返却ポート名のデータを使用する。

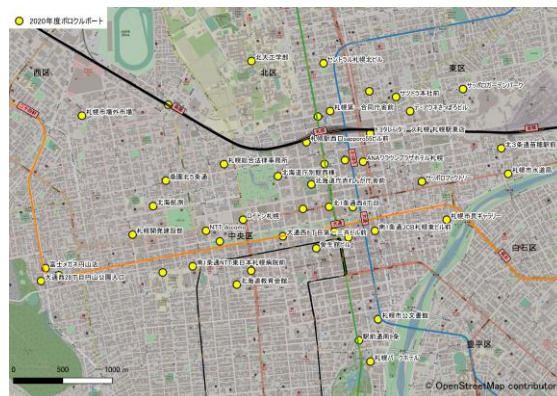


図-1 ポートマップ²⁾

2.2 ポロクル2020年度アンケートデータ

ポロクル2020年度アンケートデータは、ポロクル公式ホームページにて2020年度にポロクルを利用した人を対象に行ったアンケートデータであり、アンケート件数は1,194件である。アンケートデータには、年代・性別・住まいなどの個人属性・利用目的・公共交通機関との接続に関してなど計40個の回答データが含まれている。本研究ではまず「ポロクルの利用目的は?(2つまで選択可能)」という質問の回答で、無回答と2つ回答しているデータを除いた386件のデータ抽出した。そしてその中でも通勤・通学・ショッピング・食事の回答データを対象にして利用目的を推定する。抽出したデータの中で通勤目的であると回答した利用者は121人、通学目的であると回答した利用者は3人、ショッピング目的であると回答した利用者は96人、食事目的であると回答した利用者は17人であった。

2.3 都市計画基礎調査データ(建物用途別延床面積)

都市計画基礎調査データは市街化区域内における全建物(430,884棟)について位置情報、建物用途、階数、構造および延床面積などが含まれている。本研究では2019年度の札幌市における「都市計画基礎調査」を使用した。その中でポートから半径100m以内にある合計15種類に区分される建物用途(表-1)と延床面積を用いる。

3. 推定方法

3.1 RF(ランダムフォレスト)の概要

RFとは、複数の決定木から森を構成して、推定や識別することができる機械学習アルゴリズムである。特徴として、個々の決定木は高い識別性を持たないが、複数用いることで高い予測性能を得られること、目的変数を推定する際の説明変数の重要度（以下、寄与度と記す）を割り出すことができる、説明変数の正規化の必要性がないなどがあげられる。精度、性能、効率は決定木の個数、決定木の最大高さ、パラメータの調整、ノードの分割関数、学習時の目的関数、特徴選択によって左右される。また、過学習などにも注意しなければならない⁴⁾。

本研究では利用目的を推定するだけでなく、各説明変数の重要度を表す寄与度を求めることにより、要因分析も行う。

3.2 データセットの概要

本研究では、2.1で示した2020年度全利用データと2.2で示した2020年度アンケートデータをユーザーIDで紐づけして算出した11,390件のデータを使用する（以下、アンケート利用データと記す）。上記のユーザーIDによる紐づけイメージを示す（図-2）。アンケート利用データの中で通勤目的であると回答した利用パターンは8,988件、通学目的であると回答した利用パターンは53件、ショッピング目的であると回答した利用パターンは814件、食事目的であると回答した利用パターンは166件あった。それぞれの利用目的と平日か休日であるかをダミー変数を用いて分類し、全てのパターンで利用開始ポートと返却ポートの経度・緯度からポート間距離を算出する。これらのデータに利用開始ポートの半径100m以内の建物用途別延床面積と返却ポートの半径100m以内の建物用途別延床面積を利用開始ポートと返却ポートで紐づけすることで、データセットを構築した。

3.3 学習データとテストデータ

3.2で示したデータセットを学習モデルとして使用する。データセットを無作為に分割し、7割を学習データ、3割をテストデータとして使用する。学習と利用目的推定に使用する目的変数と説明変数を示す(表-2)。

表-1 建物分類

小分類	細分類
商業施設	－
官公庁施設	裁判所、道庁、税務署など
業務施設	会社、銀行、事務所など
宿泊施設	ホテル、旅館、モーテルなど
風俗娯楽施設	料理店、サウナ、ナイトクラブなど
遊技施設	カラオケ、パチンコなど
住宅	個人住宅
店舗併用住宅	－
共同住宅	公営
作業所併用住宅	－
文教厚生施設	学校、研究所、神社など
軽工業施設	－
サービス工業施設	－
運輸・倉庫施設	－
その他の施設	－

2020年ポロクル利用データ					アンケートデータ	
総数：323766 ID数：19090					総数：386（重複除く）	
ユーザーID	出発ポート	返却ポート	利用時間	利用開始時間・・・	ユーザーID	目的
ID1	ポート②	ポート⑧	5時間	4/12 07:04	ID1	通勤
ID1	ポート①	ポート⑦	1時間	5/19 16:32	ID3	買い物
ID1	ポート②	ポート⑧	5時間	7/7 07:08	ID5	食事
ID2	ポート①	ポート⑧	1時間	4/20 13:20	...	...
ID2	ポート④	ポート⑨	4時間	6/19 12:55	...	...
ID3	ポート⑦	ポート③	3時間	5/15 10:24	ID1194	買い物
...	...	...	...	...	...	...
ID 19090	ポート①	ポート⑧	2時間	6/29 07:02	...	...

図-2 ユーザーIDによる紐づけイメージ

表-2 目的変数と説明変数

変数	変数名	単位
目的変数1	利用目的ダミー（通勤）	無次元
目的変数2	利用目的ダミー（通学）	無次元
目的変数3	利用目的ダミー（ショッピング）	無次元
目的変数4	利用目的ダミー（食事）	無次元
説明変数	利用時間	秒
	平日祝日ダミー	無次元
	利用開始時時間	秒
	返却時間	秒
	ポート間距離	km
	利用開始ポート半径100m以内の建物用途別延べ床面積（表-1全て）	m ²
	返却ポート半径100m以内の建物用途別延べ床面積（表-1全て）	m ²

4. 結果

4.1 精度検証

テストデータの各利用目的の精度について検証した結果を示す(表-3)．精度の差は全ての利用目的で90%以上と高く，利用目的を十分に推定できる学習モデルを構築できたことがわかる．

4.2 各利用目的の比較

アンケート利用データの各利用目的と RF で推定した各利用目的の結果を示す(図-3)．通勤目的については，RF での推定結果がアンケート利用データより 14.2%低くなった．通学目的については，RF での推定結果がアンケート利用データより 0.2%高くなった．ショッピング目的については，RF での推定結果がアンケート利用データより 3.2%低くなった．食事目的については，RF での推定結果がアンケート利用データより 1.0%低くなった．したがって，通勤目的以外では，概ね正確に利用目的が推定出来ていることがいえる．

4.3 各利用目的の利用開始時間の比較

利用目的の中で利用開始時間の分布に特徴がある，通勤目的とショッピング目的を対象にして，アンケート利用データの利用開始時間を示す(図-4, 図-5)．また，2020年度全利用データの推定した各利用目的の利用時間を示す(図-6, 図-7)．利用目的が通勤である図-4と図-6を比較すると，通勤目的のピークはどちらも8時前後と17時前後にあり，通勤目的ではないピークはどちらも12時前後にあり，波形が似ていることがわかる．そのため特徴を抑えて通勤目的を推定できていることがいえる．同様に利用目的がショッピングである図-5と図-7も比較すると，ショッピング目的のピークはどちらも突出したピークは見られず，ショッピング特有の分布となった．

表-3 テストデータの精度

目的	テストデータの精度
通勤	92.2%
通学	99.8%
ショッピング	95.5%
食事	99.3%

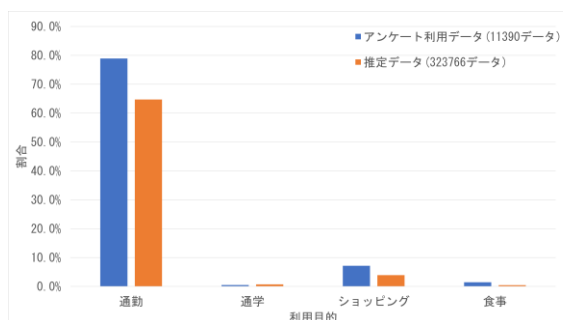


図-3 アンケート利用データの利用目的と全利用データの推定した利用目的の比較

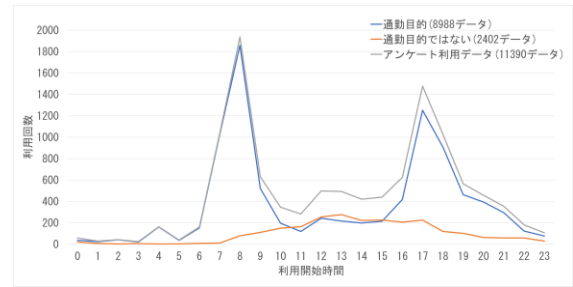


図-4 アンケート利用データの利用目的が通勤の場合

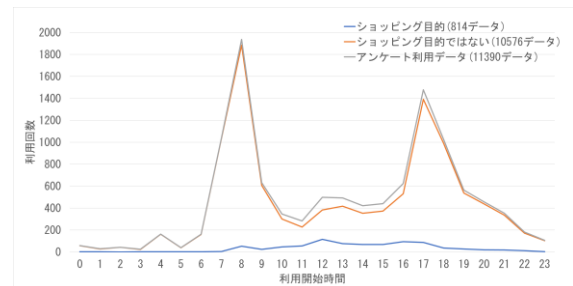


図-5 アンケート利用データの利用目的がショッピングの場合

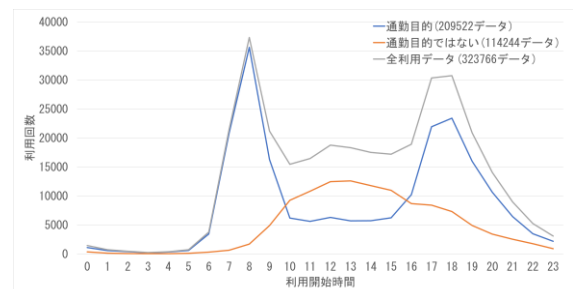


図-6 2020年度全利用データの推定した利用目的が通勤の場合

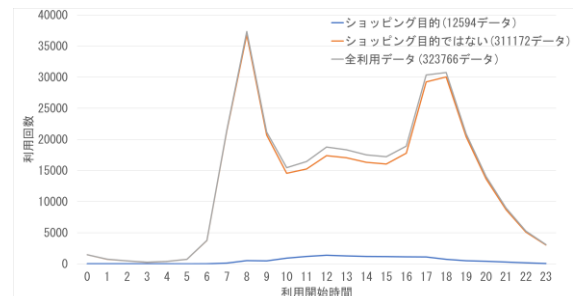


図-7 2020年度全利用データの推定した利用目的がショッピングの場合

4.4 各利用目的の寄与度

推定した 2020 年度全利用データの各利用目的の寄与度を示す(図-8, 図-9, 図-10, 図-11)。この結果より、利用目的が通勤・ショッピング・食事の場合の利用目的を推定するのに寄与する説明変数は、利用時間(秒)・ポート間距離・利用開始時間・返却時間であることが分かった。また、利用目的が通学の場合、利用開始ポートの半径 100m 以内の共同住宅の延床面積・ポート間距離・利用時間(秒)・返却ポートの半径 100m 以内の共同住宅の延床面積が寄与している。

5. まとめ

本研究では、RF を用いて、アンケートデータの利用目的を 2020 年度全利用データに拡大推定を行った結果を以下に示す。

- ・通勤目的は、アンケート利用データより低い結果となったが、通勤特有の朝と夕方のピークが結果に表れ、概ね正確に推定することができた。
- ・ショッピング目的は、アンケート利用データより低い結果となったが、突出したピークは見られず、ショッピング特有の分布が結果に表れ、概ね正確に推定することができた。
- ・通学・食事目的では、データ数が少なかったため、良い推定結果を得ることができなかった。

今後の課題として、RF 以外の分類方法や交差検証の見直し等を行い、精度を高めることが挙げられる。また、説明変数で使用する建物用途別延床面積の範囲の拡大や、アンケートデータの「ポロクルの利用目的は？(2 つまで選択可能)」という項目の扱い方を検討したい。

謝辞：本研究では、NPO 法人ポロクルからの貴重なデータを頂きました。ここに記して感謝の意を表します。

参考文献

- 1) 丸山翔太, 松田真宜, 長谷川裕修, 有村幹治: データマイニングアプローチによるコミュニティサイクルの利用動態の抽出, 土木学会論文集 D3(土木計画学), Vol.70, No.5, pp.671-680, 2014
- 2) ポロクルホームページ: <https://porocle.jp/> (2020.12.15 閲覧)
- 3) 南部浩之, 片岸将広, 熊谷美香子, 三浦 清洋, 成嶋良太, 門脇 照: 我が国のシェアサイクル事業における COVID-19 の影響分析, 土木計画学研究発表会・講演, 2020
- 4) 橋本成仁, 中島那枝: コミュニティサイクルの導入がまちの魅力に与える効果に関する研究, 都市計画論文集 Vol.52 No.2, 2017
- 5) 波部 齊: ランダムフォレスト, 情報処理学会研究報告, Vol.2012-CVIM-182 NO.31, 2012

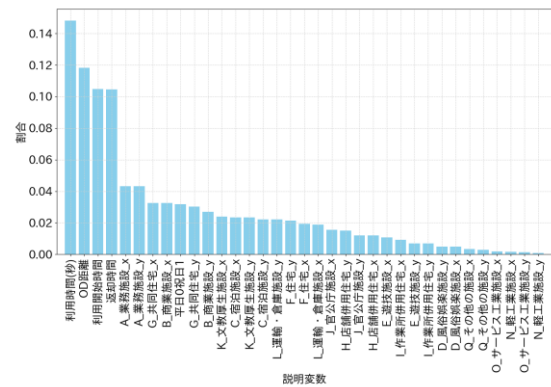


図-8 通勤目的の寄与度

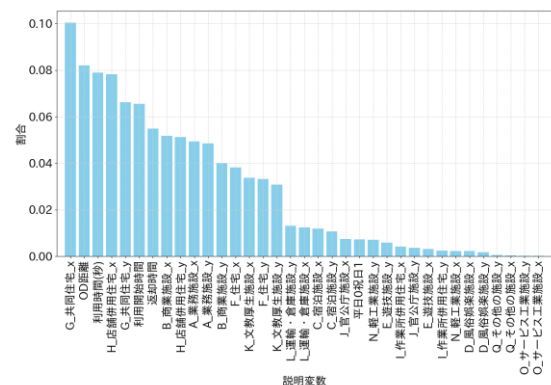


図-9 通学目的の寄与度

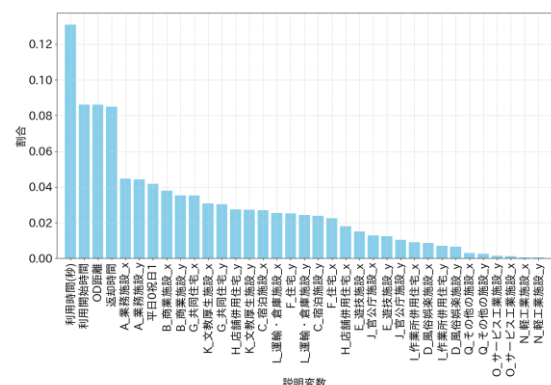


図-10 ショッピング目的の寄与度

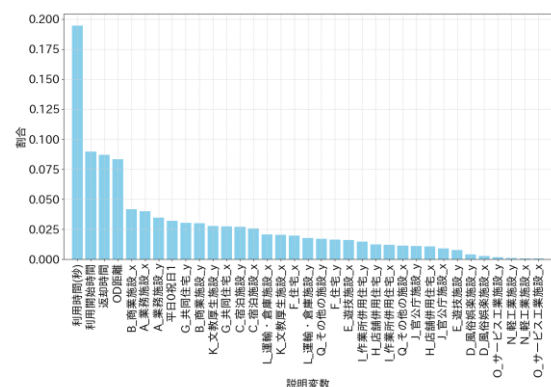


図-11 食事目的の寄与度