

IC カードデータを用いた路線バス利用行動に関する研究

劉 強

熊本市都市政策研究所 研究員

キーワード：IC カードデータ、クラスタリング、K-means、サンキーダイアグラム

1 はじめに

1.1 研究の背景

近年、ICT 技術の進展に伴い、ビッグデータの利用が拡大している。スマートフォンや GPS 等を通じて収集されたデータは、都市問題の解決に向けた分析の材料として活用されている（西垣ら、2022）。具体的には、大量のデータを統合・分析することで、空気汚染や交通渋滞といった日常的な都市課題の解決へ向けて、データ駆動型の意思決定が注目されている。

特に交通分野において、路線バスのカードリーダーから得られる IC カードデータは、乗客の行動分析や、旅行需要の予測、交通サービスの最適化等に有効活用されている（有吉、2022）。従来の短期間のアンケート調査に比べ、IC カードデータは長期間にわたる公共交通の利用状況を幅広く捉えることができる。ただし、個人情報保護の観点から、年齢や性別等の個人情報は取り扱われていない。

世界中の公共交通では、IC カードが運賃支払いのために広く導入されている。具体的な例として、フランスのパリにある Navigo（ナヴィゴ）カード（2023）、イギリスのロンドンにある Oyster（オイスター）カード（2023）、中国の香港にある Octopus カード（2023）、チリのサンティアゴにある Trajeta Bip! カード（2023）等が挙げられる。IC カードは主に運賃収集を効率化する目的で開始されたが、そのデータは、人々の移動パターンを深く理解し、交通計画の策定等にも役立てられている。

乗客の時間的な移動習慣や旅行パターンの分析は、公共交通事業者や地方自治体にとっての重要な課題である（Etienne ら、2016）。これらについては、IC カードデータを活用することで、バス停の利用傾向や乗客の利用時間等の行動を詳細に把握することができる。特に、IC カードデータに基づいて各バス停の利用時間と利用者数の特徴を特

定し、類似の利用パターンを持つバス停をクラスター化する方法が考えられる。

既往研究においても、クラスタリングの様々な手法が提案されている。例えば、Mohamed ら（2017）は、類似の移動パターンを持つ乗客をクラスタリングする研究を行っており、主に朝や夕方、平日のみに特化したバス停の利用傾向を示している。Ceapa ら（2012）の研究では、IC カードデータを基にバス停の混雑パターンの分析を行い、階層的クラスタリングを用いてバス停の再グルーピングを試みた。また、Lathia ら（2013）の研究では、乗客の平日の移動パターンにクラスタリングを適用し、通勤や夜間移動等の異なる旅行行動を特定している。これらの研究は、乗客向けの交通情報サービスを提供する上で、IC カードデータの利用が不可欠であることを示している。

1.2 研究の目的

本研究の目的は、IC カードデータを用いて、熊本市内の路線バス停の利用行動をクラスタリングによって把握することである。具体的には、平日と土日、時間帯別のバス停の利用状況を特徴量としてクラスタリングを行う。2019 年と 2020 年の結果を比較することで、時系列の変化とクラスター間の対応関係を分析する。この分析により、各クラスターの特徴を理解し、バス停の時間帯別の利用者数の傾向を把握することができる。また、2019 年と 2020 年の各クラスター間の対応関係を分析し、どのような変化があるのかを把握する。以上のことにより、バス停の利用パターンを深く理解し、それに基づく具体的な改善策や最適化案の検討に貢献することが期待される。

本研究の構成は以下の通りである。研究に使用する IC カードのデータセットは 2 章で紹介し、記述統計に基づく予備的な分析も併せて説明している。バス停のカウント統計に基づき、バス停をクラスタリングする方法とその結果に

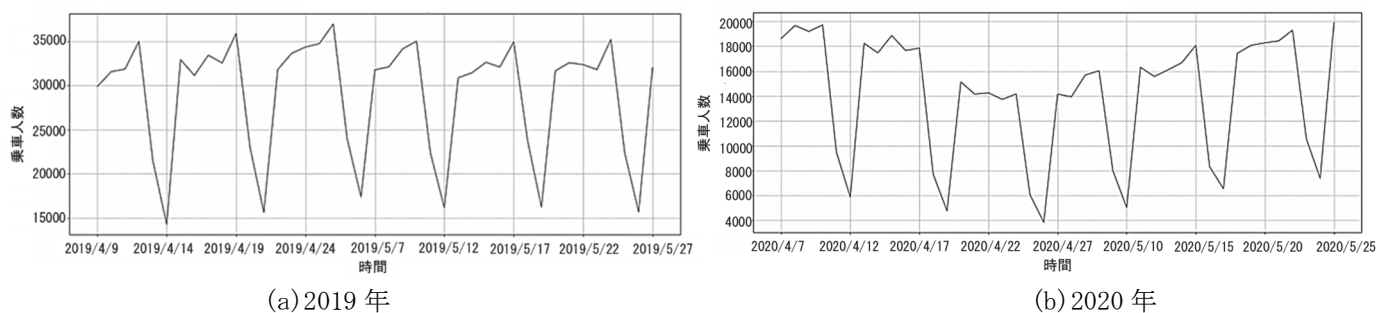


図1 乗降車人数の推移

については3章、4章で詳細に説明する。最後に、5章でまとめと今後の研究の方向性を述べる。

2 データの説明

この章では、分析に使用したICカードデータセットを説明し、データの記述統計を通じて移動の基本的な状況を紹介する。

2.1 データの内容

本研究は、2019年と2020年の路線バスのICカードデータを使用した。本市では、2014年3月に熊本市交通局の路面電車にて「でんでんニモカ」の運用を開始し、2015年4月には熊本県内の路線バス・電鉄電車にて「くまモンのIC CARD」の運用が開始された。

本研究に利用する路線バスのICカードデータは2019年4月9日から5月27日までの、合計2,240,102回のICカード利用と、2020年4月7日から5月25日までの、合計1,127,050回のICカード利用2つのデータセットである。なお、データ分析の精度を高めるため、祭日は除いた。

本研究で使用した2019年と2020年のデータは、バスのICカードタップ情報を基にしたものである。このデータは、利用者が乗降車する際のバス停のID、タッチの日付、及び時間を含んでいる。生データの処理において、以下の特徴を抽出した：

- バス停ID (StopID)：乗降車のバス停を識別。
- 日付 (Date)：タッチされた日付。
- 曜日タイプ (Day_Type)：それぞれの日付を「平日」または「土日」に分類したもの。
- 時間帯 (Hour)：24時間制でのタップ時間。

また、各利用履歴には、匿名化された乗客ID（利用したICカードのID）、カードの利用が発生したタイムスタンプ（分単位に四捨五入された日付と時刻）、乗降車のバス停の名称、乗車したバスの番号、及びバス会社に関する情報が

含まれる。

本研究では、各バス停の平日および土日の24時間毎の出発および到着のカウントを特徴量として使用した。これにより、各バス停の日常的な利用パターンを明確に把握することができる。しかし、各バス停の利用者数には大きなばらつきがあり、これをそのままクラスタリングに使用すると、利用者数が多いバス停の影響を大きく受ける可能性がある。そこで、データの正規化を行い、各特徴のスケールを統一した。本研究では、MinMaxScaler を使用して、各特徴の範囲を(0, 1)の間に正規化した。

2.2 データの記述統計

収集したデータは、多くの場合、欠損や異常値、表記ゆれ等が含まれ、そのまま分析させることができない。そのため、データサイエンスが行える形に整える必要がある。前処理では、収集したデータの内容を確認し、クラスタリングが可能な形にデータを整える。具体的には、不要なデータを削除する、欠損や異常値を取り除く、データを整える等の作業を行う。

この章ではICカードデータに前処理の記述統計を実施した。2019年と2020年の間における毎日の乗降車人数およびICカードの使用回数の変動を示した（図1・図2）。

図1を見ると、2019年の乗降車人数は金曜日にピークが見られる。土日に関連する週内の乗降車人数の変動も確認できる。また、2020年では乗降車人数は2019年と比較して大幅に減少したことが分かった。ピーク時の曜日にも変動が見られた。

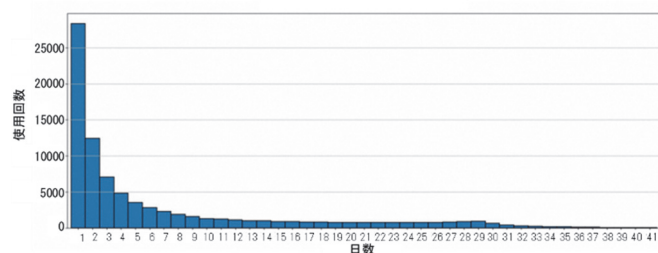
図2の横軸は、日数を示しており、これは月内のICカードの使用回数を示している。図2を見ると、2019年のICカードを使用する乗客の数は、20回で変化点を迎えるまで減少し、その後は増加傾向が見られる。しかし、29回に達すると、減少する傾向が再び現れる。次に、2020年の結果を見ると、この減少-増加-減少の現象は2019年ほど顕著では

ない。また、2019 年と 2020 年の間で、IC カードの使用回数の分布には大きな変動は確認されなかったが、特定の使用日数における利用回数の増減が存在することが分かった。そして、1 回だけの使用が多いことを確認した。これは、多くの利用者が IC カードを少回数で使用する傾向があることを示している。

2019 年と 2020 年の間の乗降車人数には顕著な変動があったものの、IC カードの使用回数の分布には大きな変動は見られなかった。



(a) 2019 年



(b) 2020 年

図2 IC カードの使用回数の変動

3 クラスタリングの方法

この章では、バス停のカウント統計に基づいてクラスタリングする方法を紹介する。

クラスタリングは、データマイニングや機械学習の中核を成す手法の一つであり、データの中から隠れた構造やパターンを抽出するために広く利用されている (Rokach ら, 2005)。特に、K-means は、その計算の効率と実用性から、クラスタリング手法として広く知られており、多くの実用的なアプリケーションに適用されている。しかし、K-means の性能は初期クラスター中心の選択に大きく依存しているため、この問題を改善するための多くの方法が提案されてきた。その中でも、K-means++ という改善されたクラスタリングの手法がある。そこで、本研究は K-means++ を利用して、熊本市のバス停のクラスタリングを行う。

K-means クラスタリングは、データを k 個のクラスターに分割するためのアルゴリズムである。アルゴリズムは、各

クラスター内の平方距離の合計を最小化するようにデータポイントを割り当てる (MacQueen, 1967)。初期のクラスター中心の選択方法によって、アルゴリズムの結果や収束速度が大きく影響を受けることが知られている。

与えられたデータセット $C = \{x_1, x_2, \dots, x_n\}$ に対して、次の目的関数 J を最小化する。

$$J = \arg \min_C \sum_{i=1}^k \sum_{x \in C_i} \|x_i - \mu_j\|^2$$

ここで、 C はクラスターの中心の集合であり、各 x_i は最も近い中心 μ_j に割り当てられる。

K-means++ は、K-means の初期化手順を改良する方法として提案された (Arthur と Vassilvitskii, 2007)。この手法は、初期クラスターの中心をデータセットから、より効果的に選択し、収束を高速化し、最終的なクラスタリングの質を向上させることが示されている。特に、各中心は、既存の中心からの距離の二乗に比例した確率で選択される。初期のクラスター中心を選択する際、データポイント x が次のクラスター中心として選択される確率 $D(x)$ は、既存の中心までの最短距離の二乗に比例する。

$$D(x_j) = \min_{i=1}^k \|x_j - C_i\|^2$$

この初期化手順により、K-means の収束が加速し、クラスタリングの質も向上することが報告されている (Bahmani ら, 2012)。

4 K-means++による分析

4.1 最適なクラスター数

K-means++ を行う前に、クラスター数 k を決める必要がある。エルボー法 (Elbow Method) を使用して最適なクラスター数を評価する。エルボー法は、異なるクラスター数でのクラスタリングの結果を比較し、クラスター内のばらつき (SSE: Sum of Squared Errors) が急激に減少する点を「エルボー」として特定する方法である。

クラスター数が増えるにつれて誤差平方和 (Inertia) が減少している。エルボー (肘) の位置を見ると、クラスター数が 3 から 5 の間にあることが示唆され (図 3)、今回は 2019 年と 2020 年の最適なクラスター数を 5 と特定した。

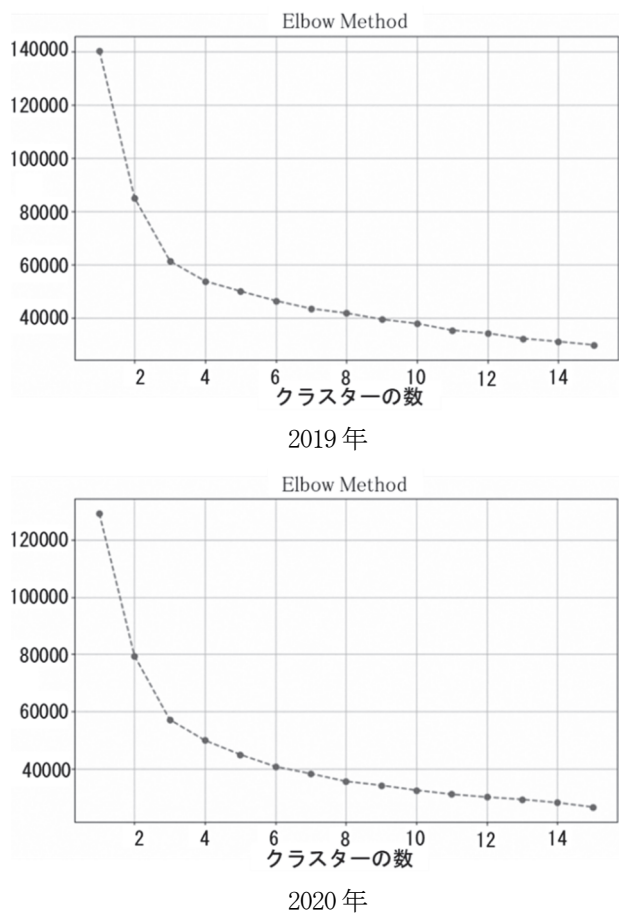


図3 エルボー法の結果

4.2 各クラスターの特徴

2019 年と 2020 年のバス停の利用者数に基づき、k-means クラスタリング (k=5) を実施した。エルボー法の結果、それぞれの年において5つのクラスターが識別された。まず、各年のクラスターごとの平均乗車・降車数の変動を確認する。

4.2.1 2019 年の特徴

2019 年の代表的な時間帯である 8:00-9:00 および 17:00-18:00 の平均乗降車の人数を分析した結果は図4で示す。2019_Cluster3 の二つの時間帯の差が大きい。また、図5を見ると、2019_Cluster2 および 2019_Cluster3 は、午前の時間帯に高い降車人数を示しており、これは主要なバス停や中心部に位置している可能性が高いことを示唆している。一方、2019_Cluster4 は、どちらの時間帯でも低い乗降車人数を示しており、これは人が少ないエリアに位置している可能性が考えられる。

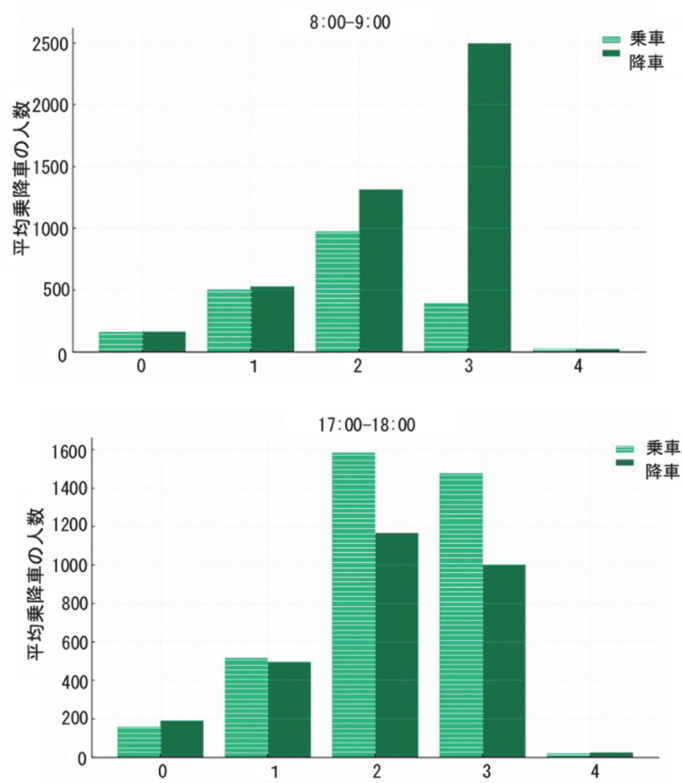


図4 2019 年代表的な時間帯の平均乗降車人数

図5を見ると、2019 年の各クラスターの特徴は以下の通りである。

- 2019_Cluster0 と 2019_Cluster1 は平日の乗車と降車の利用者数は似た傾向を示しており、特に朝と夕方にピークを迎えている。これは通勤や通学の時間帯に利用者数が増加することを示唆している。土日の利用者数は全体的に低いが、昼間にわずかな増加が見られる。でも 2019_Cluster1 の乗降者数は 2019_Cluster0 より多い。
- 2019_Cluster2 と 2019_Cluster3 は、他のクラスターと比較して、降車のピークは乗車のピークより多い。また、乗車と降車のピークは他のクラスターと逆である。そして 2019_Cluster2 の夜のピークは1時間早い。
- 2019_Cluster4 の夜の降車ピークは非常に小さい。土日について、乗車と降車の傾向が平日と類似しているが、全体的に利用者数が少なく、ピークは明確ではない。

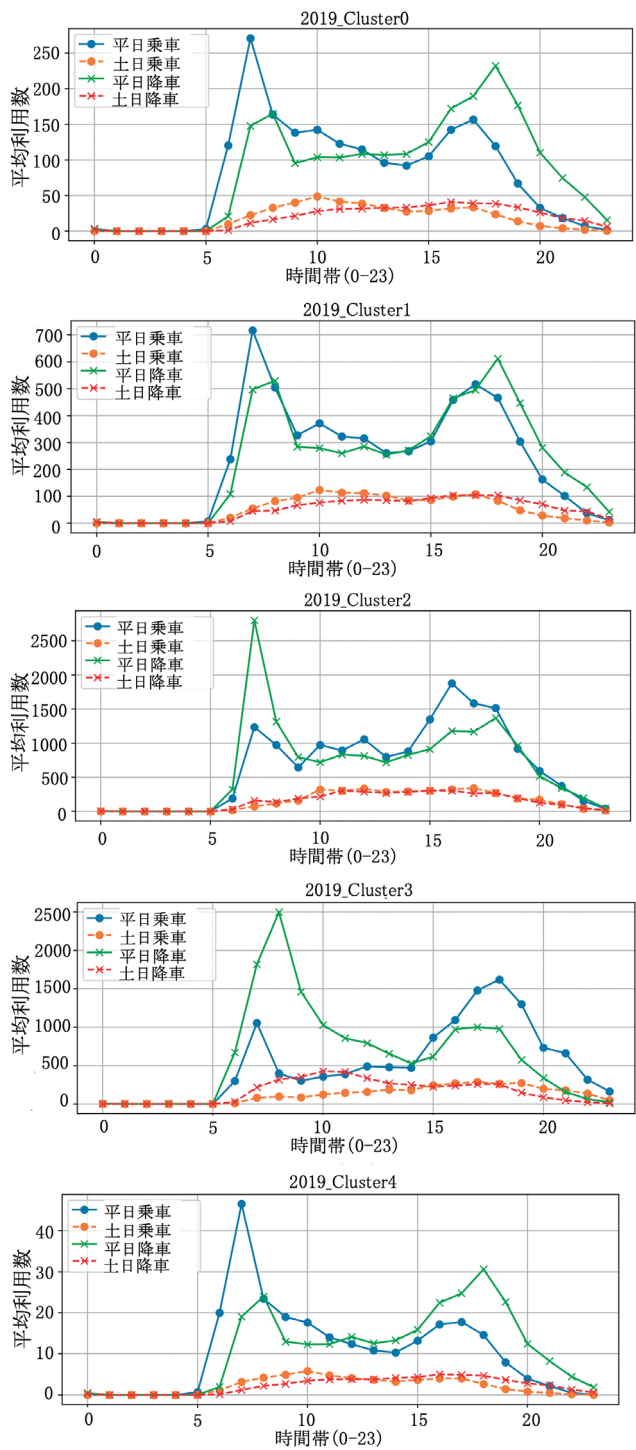


図5 2019年各クラスターの平均利用者数

4.2.2 2020年の特徴

2020年の代表的な時間帯の平均乗降車人数を分析した結果は図6で示す。2020_Cluster1はほかのクラスターと比較し、平均乗降車人数が高い。これは主要なターミナルや都市中心部に位置している可能性が考えられる。また、2020_Cluster2はどちらの時間帯でも低い乗降車人数を示

しており、これはあまり利用されていないエリアに位置している可能性が考えられる。

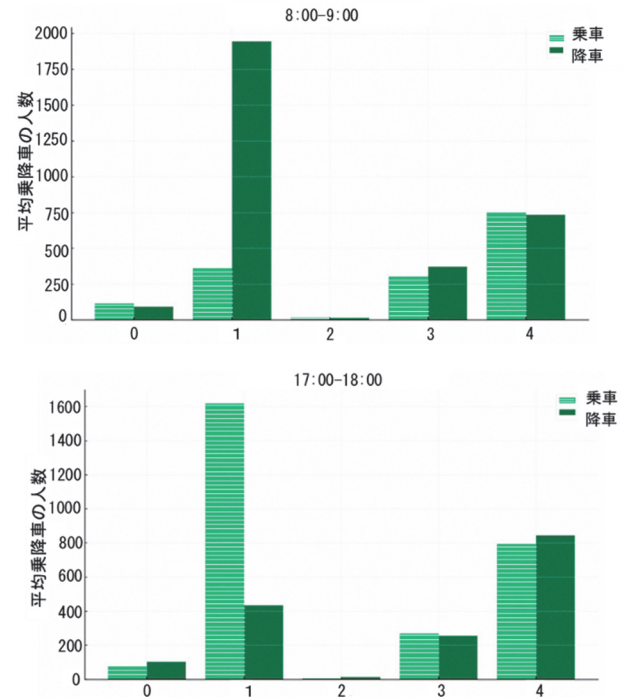


図6 2020年代表的な時間帯の平均乗降車人数

図7を見ると、2020年の各クラスターの特徴は以下の通りである。

- 2020_Cluster0 と 2020_Cluster2 の利用者数は似た傾向を示しており、朝に乗車ピークと夕方に降車ピークを迎えている。これは通勤や通学の利用者数が多いことを示唆している。
- 2020_Cluster1 の乗車と降車のピークは他のクラスターと逆である。また、土日の利用も顕著なピークが現れ、重要なバス停（例えばバスターミナル）や商業地に位置すると考える。
- 2020_Cluster3 と 2020_Cluster4 の乗降者数の傾向はほぼ一致している。同じ時間帯の乗降者数の差が小さい。

表1 各クラスターのバス停数

	Cluster0	Cluster1	Cluster2	Cluster3	Cluster4
2019	315	62	8	11	1358
2020	305	6	1270	68	9

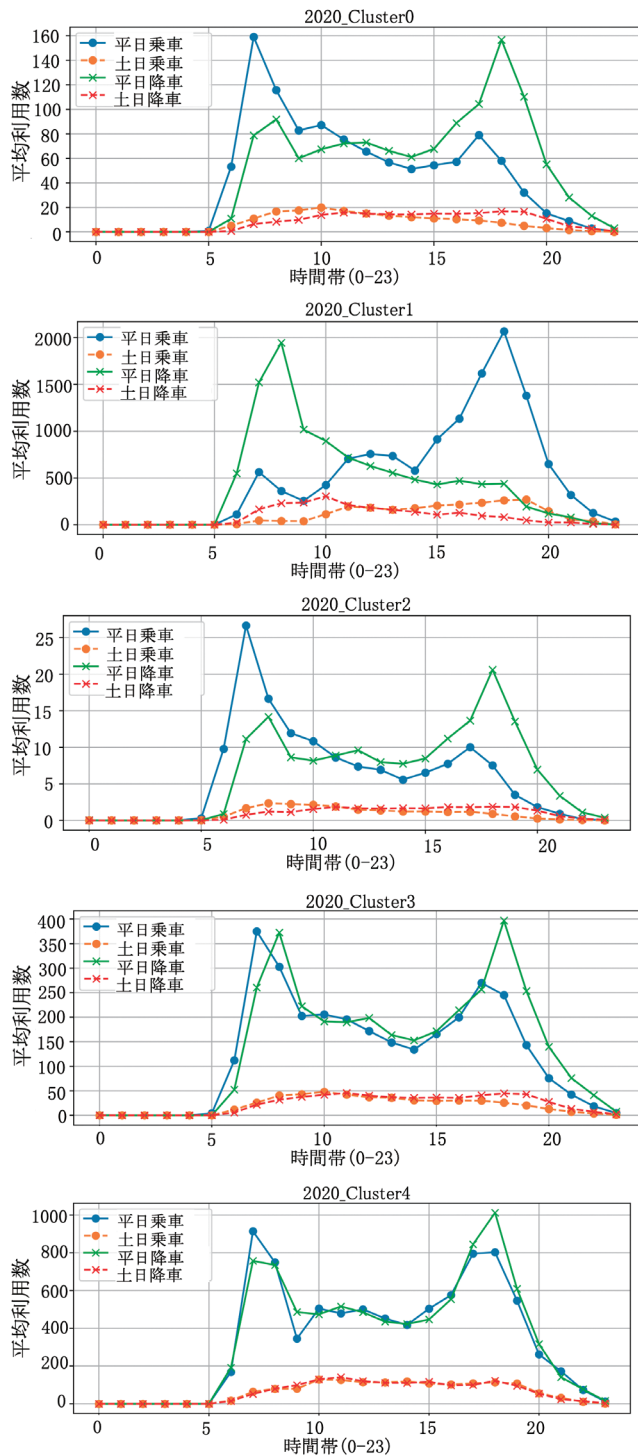


図7 2020年各クラスターの平均利用者数

4.2.3 2019年と2020年の比較

2019年の結果をベースとして2020年の結果と比較し、クラスター間の対応関係を探る。表1は各クラスターのバス停数を示している。4.2.1と4.2.2より、2019年と2020年のクラスターの対応関係を考えて。

- 2019_Cluster0 と 2020_Cluster0 は、朝夕の乗降人数が少なく、バス停の数も多い。

- 2019_Cluster1 と 2020_Cluster3 は、朝夕の乗降人数が中程度で、バス停の数は少ない。
- 2019_Cluster2 と 2020_Cluster1 は、乗降人数が多く、バス停の数が非常に少ない。また、両方のクラスターは全体の乗降車人数が最も多いといった特徴を持っている。
- 2019_Cluster3 と 2020_Cluster4 のバス停は、乗降車人数が多く、バス停の数も非常に少ない。また、両方のクラスターは、それぞれの年で全体の乗降者数が多い特徴を持っている。
- 2019_Cluster4 と 2020_Cluster2 は、乗降者数が最も少なく、バス停の数が最も多い。

以上のような似た特徴を持っていることから、この対応関係は合理的と考えられる。2019年と2020年のクラスターの対応関係は以下の通りである。

2019_Cluster0 ↔ 2020_Cluster0、
 2019_Cluster1 ↔ 2020_Cluster3、
 2019_Cluster2 ↔ 2020_Cluster1、
 2019_Cluster3 ↔ 2020_Cluster4、
 2019_Cluster4 ↔ 2020_Cluster2。

4.3 クラスター間の移動

前章の分析はやや表面的であるため、対応関係を確認するため、クラスター間のバス停の移動を通して見ていく。2019年と2020年の各バス停のID (StopID) と、そのバス停が所属するクラスターの情報 (Cluster) が含まれているデータを統合し、各バス停がどのクラスターからクラスターへ移動したかを明らかにするため、サンキーダイアグラムを作成した。サンキーダイアグラムは、流れの大小や移動の方向を視覚的に表現するのに適している。具体的には、Pythonのpychartsライブラリを使用してサンキーダイアグラムを作成した(図8)。また、クラスターごとに異なる色を割り当てることで、視覚的に区別しやすくした。

図8を見ると、最も多い移動は、2019_Cluster4 から 2020_Cluster2 への移動であった。その次は2019_Cluster0 から 2020_Cluster0 への移動であった。しかし、一部の移動の数が非常に大きく、他の移動が視覚的に小さくて見にくい問題があった。この問題を解決するため、流れの値を対数変換する方法を採用した。対数変換により、大きな値の影響が緩和され、サンキーダイアグラムがより均衡になる。実際の流れの値とは異なるが、相対的な大小関係をより明確に示すことができた(図9)。

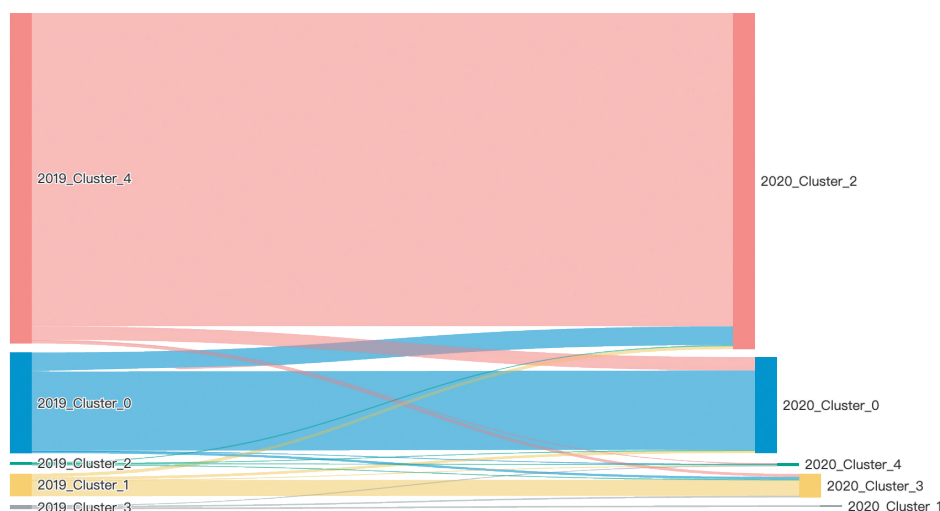


図8 クラスタ間の対応関係

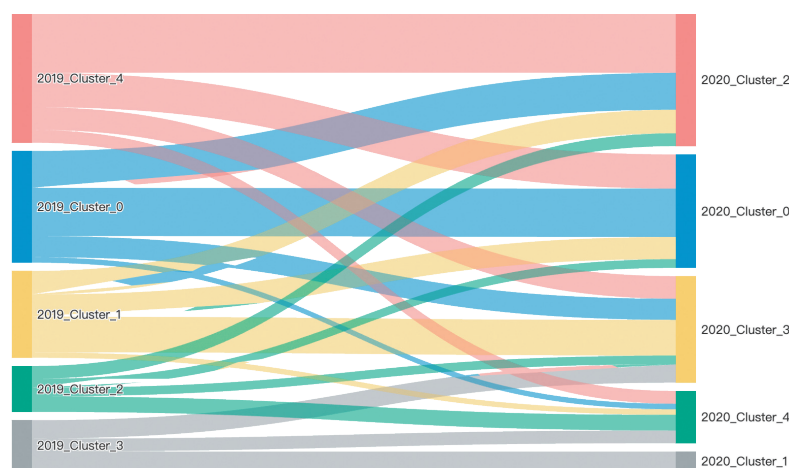


図9 クラスタの移動（対数変換）

相対的な大小関係を見ると、図8で視覚的に小さくて見にくいクラスタの移動関係も明確になった。2019_Cluster1から2020_Cluster3への移動、2019_Cluster2から2020_Cluster4への移動と2019_Cluster3から2020_Cluster1への移動が多いことが分かった。4.2のクラスタの対応関係は合理的と言える。また、各クラスタから他のクラスタへの移動は頻繁で、繋がりが複雑である。

4.4 クラスタの比較

2019年と2020年のクラスタを比較する。まず全ての2020年のクラスタは利用人数が減少した。2019_Cluster0↔2020_Cluster0の変化は、土日の降車ピーク時間が、2019年の16:00-18:00から、2020年は18:00-19:00に移動した。平日夜の降車ピークについて、2019年は朝の乗車より少なかったが、2020年は同程度であ

った。2020年には、平日の18:00の降車利用率が増加し、これは仕事後の飲食利用が減少した可能性がある。

2019_Cluster1↔2020_Cluster3の変化について、平日の利用者ピーク時間は、2019年と2020年ともに7:00と18:00で変わらず、土日のピーク時間は2019年の17:00-18:00から、2020年は18:00-19:00に移動した。2019年の平日朝夜の降車時の利用者数は、朝の乗車ピークより少なかったが、2020年はほぼ同じになった。

2019_Cluster2↔2020_Cluster1の変化について、2020年の平日の乗車ピークは、2019年と比べて増加しているが、平日の降車は低下している。

2019_Cluster4↔2020_Cluster2の平日の変化はあまり見られないが、土日の朝の利用者ピーク時間が、2019年の10:00から2020年は8:00に前倒しされた。

5 結論

5.1 まとめ

データの記述統計の結果から、多くの利用者が IC カードを少ない回数で利用されている傾向があることが分かった。また、K-means++法を用いたクラスタリングの結果から、バス停の利用行動には 5 つの特徴的なクラスターが存在することが判明した。特に、平日の朝夕の乗降人数が多いバス停、土日の利用も顕著なピークが現れるバス停等、それぞれ異なる特徴を持つクラスターが明確に識別された。以上の分析により、各クラスターのバス停がどのような目的で利用されているかが理解できる。そして、2019 年と 2020 年のクラスタリング結果を比較した。サンキーダイアグラムを利用して、2019 年と 2020 年のバス停のクラスター間の移動を視覚的に分析し、移動の多さや方向が分かった。これにより、2019 年と 2020 年のクラスターの対応関係を把握できた。この結果から、一部の顕著な移動が見られたが、全体としては各クラスターが一貫したことを保持していると言える。

以上の結果から、K-means++はバス停の利用パターンを識別するのに有効な手法であることが示された。このようなデータ駆動型のアプローチが路線バスの利用行動を理解するための洞察を提供すると考えられる。

5.2 今後の方向性

今後の研究について、以下のような分析が行える可能性がある。

時系列分析: 2019 年と 2020 年の間の乗降車人数やバス停の利用パターンの変化を追跡し、交通政策の影響、または他の外部要因の影響を評価できる。

将来予測: 過去のデータと現在のデータのトレンドを基に、将来の乗降人数やバス停の利用パターンの予測を行うことができる。

市民のニーズの理解: クラスタ分析を用いて、特定のエリアやバス停の利用者のニーズや行動パターンを理解し、より効果的なサービス提供やインフラの改善により、運行効率化が図れる可能性がある。

これらは、今後の研究の方向性についての一部の分析の例である。これらの分析を通じて、熊本市の路線バスの運行効率化や、バスの利便性を向上させること等を築くことが可能である。

謝辞

本稿では、共同経営推進室（九州産交バス・産交バス・熊本バス・熊本電鉄・熊本市都市バス）より貴重なデータをご提供いただきました。ここに記して謝意を表します。

(参考文献)

- (1) 西垣 友貴, Jan-Dirk SCHMÖCKER, 山田 忠史, 中尾 聡史 (2022) 「GPS データを用いたメッシュ人口に基づく観光者数の推定」、『土木計画学研究・論文集』 第 39 卷 (特集)、121-130
- (2) 有吉 亮 (2022) 「モビリティ関連分野における ICT 活用のこれまでとこれから」、『通信ソサイエティマガジン』 16(2)、121-130
- (3) Navigo (ナヴィゴ) カード, ホームページ URL : <https://www.iledefrance-mobilites.fr/> (閲覧日 : 2023 年 8 月)
- (4) Oyster (オイスター) カード, ホームページ URL : <https://oyster.tfl.gov.uk/> (閲覧日 : 2023 年 8 月)
- (5) Octopus カード, ホームページ URL : <https://www.octopus.com.hk/en/consumer/index.html> (閲覧日 : 2023 年 8 月)
- (6) Trajeta Bip! カード, ホームページ URL : <https://www.tarjetabip.cl/> (閲覧日 : 2023 年 8 月)
- (7) Etienne Côme, Mohamed K. El Mahrssi & Latifa Oukhellou, (2016), “A mixture model clustering approach for temporal passenger pattern characterization in public transport Anne-Sarah Briand”, International Journal of Data Science and Analytics, vol. 1, pp37-50.
- (8) Mohamed K. El Mahrssi, Etienne Côme, Latifa Oukhellou, and Michel Verleysen, (2017) ” Clustering Smart Card Data for Urban Mobility Analysis”, IEEE Transactions on Intelligent Transportation Systems, VOL. 18, NO. 3, pp712-728.
- (9) Ceapa I., C. Smith, and L. Capra, (2012) “Avoiding the crowds: Understanding tube station congestion patterns from trip data”, Proc. ACM SIGKDD Int. Workshop Urban Comput., pp134-141.
- (10) Lathia, N., Smith, C., Froehlich, J., and Capra, L. (2013) “Individuals among commuters: building personalised transport information services from fare collection systems”, Pervasive Mob. Comput. 9(5), 643-664. Special issue on Pervasive Urban Applications

- (11) Rokach, L., Maimon, O. (2005). Clustering Methods. In: Maimon, O., Rokach, L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA.
https://doi.org/10.1007/0-387-25465-X_15
- (12) MacQueen, J. (1967) “Some methods for classification and analysis of multivariate observations”, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No. 14, pp. 281-297.
- (13) Arthur, D., & Vassilvitskii, S. (2007) “k-means++: The advantages of careful seeding”, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1027-1035.
- (14) Bahmani, Bahman & Moseley, Benjamin & Vattani, Andrea & Kumar, Ravi & Vassilvitskii, and Sergei. (2012) “Scalable K-Means++”, Proc. VLDB Endow., 5.
10.14778/2180912.2180915.

劉 強 (熊本市都市政策研究所 博士研究員)

2016 年度に熊本大学自然科学研究科建築学専攻で修士 (工学) の学位を取得。2020 年度に熊本大学自然科学研究科の環境共生工学専攻で博士 (工学) の学位を取得。
令和2 (2020) 年度より現職。