

交通系 IC カードデータを用いた 路線バス利用者数の予測モデル

劉 強

熊本市都市政策研究所 研究員

キーワード：IC カード、バス利用者数、予測モデル、機械学習、深層学習

1 研究の背景と目的

公共交通の運行の効率化は、都市の交通政策において重要な課題である。特に、公共交通機関の利用を予測することは、各路線や時間帯ごとの利用率を把握でき、効率的な運行計画の立案に繋がるなど、都市の持続可能な交通の構築における重要な役割を果たす。代表的な公共交通機関の 1 つとして、路線バスが挙げられる。路線バス利用者数の正確な予測は、運行計画の最適化、バス利用者の利便性向上、そしてバスの遅延の把握に直接貢献することができる。こうした背景から、予測モデルの開発と精度の向上は、公共交通事業者やバス利用者にとって重要なこととなっている (Erfan ら, 2023)。

バス利用者数の予測を行うことによって、バス利用者側からみれば、バスの利用状況を事前に把握できることも大きな利点である (Li ら, 2014)。事前にバスの利用者数の情報を入手できるようになると、バス利用者は自身の移動の計画を柔軟かつ効率的に立てることができる。バス事業者にとって運営効率を向上させることが可能となり、適切なルートを提供することができる (Hu, 2011)。例えば、運行スケジュールの最適化や、走行距離の短縮が可能となる。さらに正確な利用者数の予測は、環境負荷を考慮した運営を行うための基盤ともなる。不要な運行を減らし、燃料消費の削減や CO₂ 排出量の低減を図ることができる (Carlos, 2024)。バス利用者数の予測は、持続可能な交通を構築するための重要なステップとなる (Alkhereibi ら, 2023)。

近年のデータサイエンス、情報技術、AI 技術の進歩は、交通流¹の予測方法に革命をもたらした (戸崎, 2022)。大規模な交通データの収集と分析により、精密な予測が可能になり、公共交通の運営効率が大幅に向上している。機械学習と深層学習などの AI 技術を活用した既往研究も過去 5 年のうちに急増している。機械学習や深層学習などのモデル

は、伝統的な統計モデル²と比較し、特に時系列予測の分野で性能向上が急速に進んでいる。機械学習モデル³は、データから複雑な関係性や隠れた傾向を発見し、予測の正確性を向上させるために利用される。さらに、深層学習モデル⁴は、より深いレベルでのデータ表現を学習し、特に大量のデータセットがある場合に高い予測精度を実現できる。そのため、本研究では、統計モデルのみならず、機械学習と深層学習のモデルを用いて熊本市のバス利用者数予測の可能性と精度を検証する。検証にあたっては、既存の予測モデルは種類が多く、どの予測モデルが熊本市の状況に合うのか明らかでないため、複数の予測モデルを比較し、熊本市に最も適切なモデルを特定する必要があると考えられる。

そこで本研究の目的は、IC カードデータを利用し、統計モデル、機械学習モデル、深層学習モデルを用いて、熊本市のバス停の利用者数を予測することとした。バス停は利用者数が最も多いものを選び、熊本駅前と通町筋の 2 つの主要バス停に焦点を当てる。なお桜町は、利用者数は多いものの 2019 年 9 月 14 日サクラマチクマモト開業前後で状況が大きく異なるため、研究対象から除外した。これらのバス停は熊本市の中心部に位置しており、多くの市民や観光客が利用しているため、交通流の変動が激しく予測が困難である。そのため、バス停周辺の POI⁵データを利用することで、バス利用者数の流れに影響を与える可能性のある周辺環境の変化をモデルに組み込む。最後に、これらのモデルの予測精度と性能を比較することにより、最も精度の高いモデルを特定する。

本研究のフローを図 1 に示す。第 2 章では関連する既往研究を紹介し、研究の基盤となる理論的背景を提供する。第 3 章では研究に使用するデータを紹介する。第 4 章では研究で使用される 3 つの予測モデルを紹介する。第 5 章ではこれらのモデルの予測精度を比較する。第 6 章では評価の最も

高いモデルを使用し、予測結果を示す。最後に、第7章では結論をまとめて考察する。

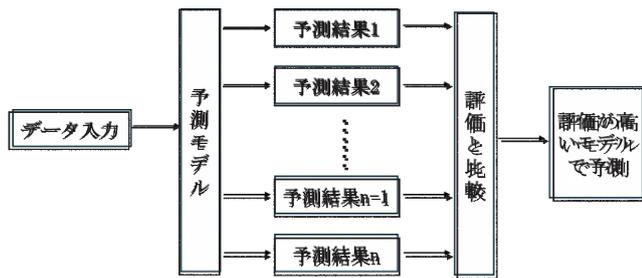


図1 研究のフロー

2 既往研究

既往研究では、バス利用者数の予測精度を高めるために様々なモデルが採用されている。多くの研究では統計モデル、機械学習モデル、深層学習モデルを用いて公共交通機関の利用者数を予測していた。角野ら (2014) はポテンシャル理論⁶を応用してICカードデータを利用したバスの需要予測を行なった。また、青地ら (2020a) は時間帯や天候の変化に応じて利用者数が変動することを考慮して、利用者数を目的変数とした予測モデルを構築した。さらに、青地ら (2020b) はランダムフォレスト⁷、一般化線形モデル⁸、一般化加法モデル⁹を用いて、利用者数の予測モデルを作成し、予測モデルの評価を行った。その結果、機械学習の手法の1つであるランダムフォレストが最も精度が高いことが示された。

海外の研究論文をみると、公共交通機関の利用者予測の精度を向上させるために、様々な手法を取り入れている。Yangら (2020) は中国北京の地下鉄のICカードデータに基づいて、深層学習のLSTM手法を利用し、短期的な地下鉄の利用者数を予測した。また、Lvら (2022) は、POIデータとXGBoost アルゴリズムを組み合わせたバス利用者予測モデルを提案した。このモデルは、まずバス停周辺のPOIデータを収集し、XGBoost アルゴリズムを使用して各バス路線のモデルを訓練した。

これらの研究では、機械学習モデルと深層学習モデルが高い予測精度を持つことを示している。しかし、国や都市によって状況が異なるため、一概に最も優れているモデルを決定することは困難である。また、空間的な影響を考慮している研究は少ない。そのため、本研究では空間的な影響を考慮し、統計モデル、機械学習モデル、そして深層学習モデルを用いて熊本市のバス利用者数の予測を行う。

3 本研究で使用するデータ

予測モデルごとに、各時間単位での予測の精度を検証するため、2019年5月7日から7月31日までのICカードデータを利用した。ICカードデータセットは、バス乗降車時に記録されたものである。表1に使用したICカードデータの概要を示す。

表1 ICカードデータの概要

データ名	内容
日付	バスを利用した日
乗車バス停番号	乗車したバス停に付いている番号
乗車バス停名	乗車したバス停名
乗車時間	乗車した時間 (分まで)
降車バス停番号	降車したバス停に付いている番号
降車バス停名	降車したバス停名
降車時間	降車した時間 (分まで)
車両番号	利用したバスの番号

また本研究では、バス停での乗車人数を利用者数と定義した。表1をみると、ICカードの乗降データにはバス利用者数の記録が含まれていないため、ICカードデータを集計し、熊本市のバス停から熊本駅前と通町筋のバス停におけるバス利用者数を抽出する必要がある。データの特徴を分析するために、2019年のICカードデータを用いて、1週間、1日、そして15分ごとに各バス停の利用者数を集計し、それぞれの時間スケールに対して週単位、日単位、15分単位の3種類の時系列を構築した。これにより、バス利用者の予測がこれらの時間間隔ごとにどのように精度が変化するかを詳細に分析するために時刻 t における日 n のバス利用者数を $d_n(t)$ と定義した。またバス利用者データには時間ラベルを付け、以下の3種類の時系列を構築した。

週間時系列 $T_{w}(t)$ は、同じ曜日または週末の同じ時間帯において、 $d_n(t)$ より n_w 週前のデータからなる。ここで、時間間隔は1週間と設定した。

日次時系列 $T_{d}(t)$ は、同じ時間帯において、 $d_n(t)$ より n_d 日前のデータからなる。ここで、時間間隔は1日と設定した。

15分時系列 $T_{15}(t)$ は、同じ日における $d_n(t)$ の前の n_m 時間間隔のデータからなる。ここで、時間間隔は15分と設定した。

週間時系列の例として、2019年5月20日の月曜日の午前

9 時から 10 時までのバス利用者数を分析する場合、週間時系列 $T_{im}(t)$ は、2019 年 5 月 13 日、5 月 6 日といった過去の同じ月曜日の午前 9 時から 10 時までのバス利用者数を含むことになる。これらのデータは全て同一の時間間隔（例えば毎週）かつ同じ平日または週末という条件を満たしている。

4 予測モデル

本研究で、使用する予測モデルを説明する。本研究は統計モデル、機械学習モデル、深層学習モデルを用いて、熊本市のバス停の利用者数を予測する。各モデルの詳細は以下の通りである。

4.1 ARIMA モデル

ARIMA モデル (Autoregressive Integrated Moving Average Model, 自己回帰和分移動平均モデルの略) は、時系列データの分析や予測に使用される統計モデルである。このモデルは、時系列データの中で発生するパターンを理解し、データの自己相関と移動平均部分を組み合わせる利用する。

$$(1 - \sum_{i=1}^p \phi_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^q \theta_i L^i) \varepsilon_t$$

ここで、 L はラグ演算子 (一期前の値を取る演算子)、 X_t は時刻 t での時系列データ、 ϕ_i は自己回帰パラメータ、 θ_i は移動平均パラメータ、 ε_t は時刻 t での誤差項 (ホワイトノイズ) である。

4.2 XGBoost モデル

XGBoost (eXtreme Gradient Boosting) は、勾配ブースティング (Gradient Boosting) フレームワークを実装した高性能な機械学習ライブラリで、分類、回帰などの多様な機械学習タスクに使用されている。XGBoost は、決定木を基にしたアンサンブル学習方法の 1 つで、複数の決定木を組み合わせ、1 つの予測モデルを構築する。特に大規模なデータセットでの予測精度と計算速度のバランスが良いことで知られている。

$$\text{Obj}(\Theta) = L(\Theta) + \Omega(\Theta)$$

ここで、 Θ はモデルのパラメータ、 $L(\Theta)$ は学習データに対する損失関数 (Loss Function)、 $\Omega(\Theta)$ はモデルの複雑さを測る正則化項である。

4.3 LSTM モデル

LSTM (Long Short-Term Memory) モデルは、深層学習モデルであり、長期的な依存関係を学習することができる。従来の RNN¹⁰ が短期的なコンテキストのみを扱うのに対し、LSTM は長期的な情報を保持し続けることができる。LSTM の主要な特徴は、情報を長期間保持するセル状態 (cell state) と、このセル状態を調整するためのゲート (gate) である。ゲートは情報の流れを調節する役割を持ち、それぞれ忘却ゲート (forget gate)、入力ゲート (input gate)、出力ゲート (output gate) がある。これらのゲートは、どの情報を保持し、どの情報を忘却するかを学習する。LSTM の計算式は以下の通りである。

忘却ゲート:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

入力ゲート:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

セル状態の更新:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

出力ゲート:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

隠れ状態の更新:

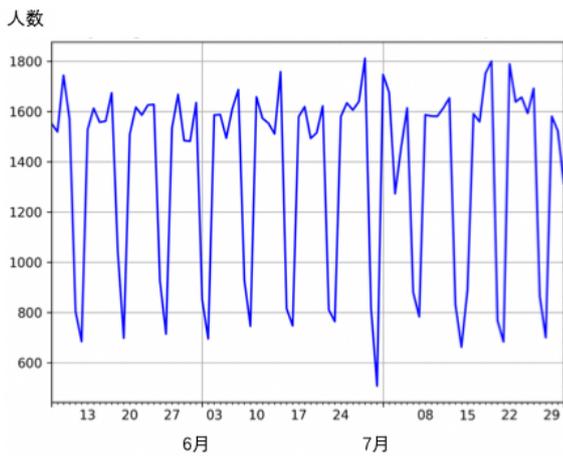
$$h_t = o_t * \tanh(C_t)$$

ここで、 t は時刻を表す。 x_t は時刻 t における入力。 h_t は時刻 t における隠れ状態 (出力)。 C_t は時刻 t におけるセル状態、 W と b は重みとバイアス。 σ はシグモイド関数であり、0 から 1 の間の値を出力し、ゲートの開閉を制御する。 $*$ は要素ごとの積を示す。 \tanh は双曲線正接関数であり、-1 から 1 の間の値を出力し、データの正規化を行う。

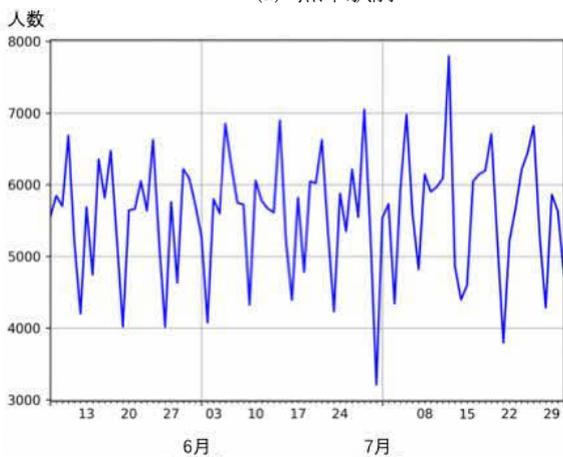
5 モデルの評価と比較

5.1 データの可視化

図 2 は、2019 年 5 月 7 日から 7 月 31 日までの期間における、熊本駅前と通町筋の 2 つのバス停の乗車人数の推移を示している。図 2(a) は熊本駅前のバス利用者数を、図 2(b) は通町筋のバス利用者数を示している。期間中のデータから、通町筋の乗車人数が熊本駅前に比べて多いことがわかる。また急激な増減が見られる部分は、週末の影響を反映していると考えられる。



(a) 熊本駅前



(b) 通町筋

図2 各バス停のバス利用者数

5.2 予測モデルの比較

5.2.1 適合性と予測精度の指標

各予測モデルの適合性を比較する。指標には、 R^2 (決定係

数)¹¹と AIC (赤池の情報量規準)¹²を使用する。これはモデルのあてはまり度を表す統計量である。また、各モデルの予測精度を評価し、それらを比較する。評価指標には、RMSE¹³ (Root Mean Squared Error、二乗平均平方根誤差)、MAE¹⁴ (Mean Absolute Error、平均絶対誤差)、MAPE¹⁵ (Mean Absolute Percentage Error、平均絶対パーセント誤差)を使用する。これらは、モデルの予測性能を定量的に測定するための標準的な指標であり、予測値と実際の値との差異を示すものである。これらの指標を用いることで、各予測モデルの結果が実際のデータにどの程度近いのか、またはどの程度の誤差があるかを詳細に分析することが可能になる。なお数値の計算に際しては、小数点以下2位で四捨五入した。

予測モデルの適合性の評価として、AIC が低い値、 R^2 が高い値が望ましい。予測モデルの精度の評価として、RMSE, MAE が低い値が望ましい。また MAPE は予測の平均的なパーセンテージ誤差を示すため、低い値が望ましい。これらの指標をもとに週単位、日単位、15分単位のデータに対するモデルの適合性と精度を比較し、それぞれの時間スケールに最も適したモデルを選択する。

5.2.2 比較の結果

モデルの適合性を比較した結果を表2に示す。またモデルの予測精度の評価を表3に示す。まず熊本駅前におけるARIMAモデルの結果について、 R^2 の値をみると、日単位と週単位のデータで適合性が低いことが示されている。特に、週単位のデータでは R^2 が0に近い値となり、適合性が低いことが示されている。

表2 モデルの適合性と予測実行時間

		R^2		AIC		実行時間 (分:秒)	
		熊本駅前	通町筋	熊本駅前	通町筋	熊本駅前	通町筋
1週間	ARIMA	0.00	0.00	45569.71	48683.78	00:00	00:00
	XGBoost	0.87	0.81	-	-	00:00	00:00
	LSTM	0.12	0.00	-	-	00:07	00:11
1日	ARIMA	0.01	0.00	1119.36	1251.66	00:00	00:00
	XGBoost	0.84	0.00	-	-	00:10	0.00
	LSTM	0.81	0.06	-	-	00:30	00:24
15分	ARIMA	0.02	0.00	173.02	499.60	00:00	00:01
	XGBoost	0.84	0.88	-	-	01:00	0.00
	LSTM	0.80	0.90	-	-	13:26	8:25

表3 モデルの精度

		MAE		RMSE		MAPE	
		熊本駅前	通町筋	熊本駅前	通町筋	熊本駅前	通町筋
1 週間	ARIMA	2475.54	11358.84	3442.50	14925.06	55.85%	69.42%
	XGBoost	5.21	17.28	8.21	23.47	>100%	>100%
	LSTM	13.25	1105.06	23.08	1464.40	51.57%	2.97%
1 日	ARIMA	289.51	608.92	343.06	750.17	26.88%	11.95%
	XGBoost	5.81	742.92	9.09	921.39	>100%	15.94%
	LSTM	5.08	531.37	8.28	1464.40	40.41%	2.97%
15 分	ARIMA	12.14	44.80	20.08	53.06	19.70%	>100%
	XGBoost	5.81	13.17	9.09	17.96	>100%	>100%
	LSTM	6.39	12.51	8.89	16.87	73.19%	>100%

XGBoost モデルの結果をみると、週単位、日単位、15分単位のデータに対して一貫して高いR²の値を示し、比較的良好な予測精度を達成している。XGBoost のR²、MAE、RMSE 結果は異なる時間単位であっても変化が少なく、最も安定している。また、LSTM モデルに対して、15分単位と日単位のデータに対して良好な適合性を示しているものの、週単位のデータではR²が低い値を示し、適合性が低下している。

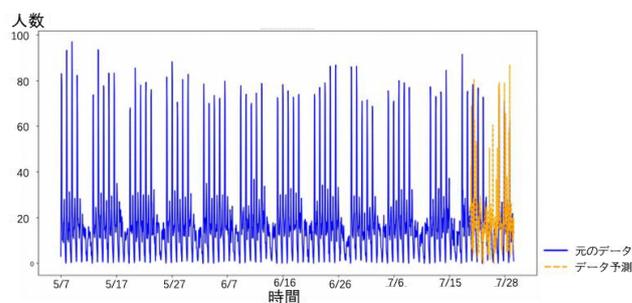
また通町筋における ARIMA モデルの結果は熊本駅前と同じく、特に日単位と週単位のデータで性能が低いことが示されている。週単位のデータではMAE と RMSE が高い値となり、予測の精度が悪いことが示されている。XGBoost モデルの結果をみると、15分単位、週単位のデータに対して一貫して高いR²の値を示し、良好な予測精度を達成している。XGBoost のR²、MAE、RMSE 結果は異なる時間単位においても安定しており、最も高い精度を示している。また、LSTM モデルは15分単位のデータに対して良好な適合性を示しているが、週単位のデータではR²の値が低下し、適合性が低くなる傾向が見られる。通町筋の予測では15分単位のXGBoost モデルが最も高い予測精度を示した。

全体として、熊本駅前の予測では、日単位の LSTM モデルが最も高い予測精度を示し、通町筋の予測では15分単位の XGBoost モデルが最も高い精度を示した。

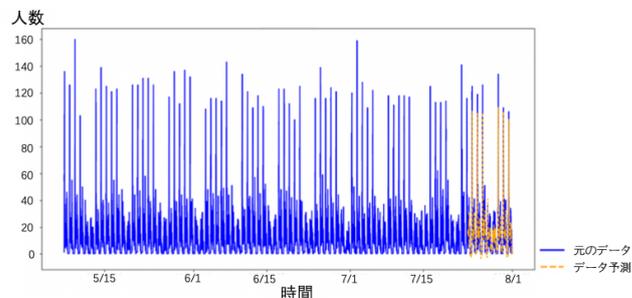
6 モデルの予測結果

R²、AIC、MAE、RMSE、MAPE の値を比較して、最も精度の高い予測を提供するモデルを利用し、バス利用者数の予測を行った。また評価が最も高いモデルを利用し、バス利用者

数の予測を行った。図3はモデルによるデータ予測と実際のバス利用者数との比較を示している。LSTM の予測結果を図3(a)に示し、XGBoost の予測結果を図3(b)に示す。図3をみると、予測結果は実際のバス利用者数と高い割合で一致していることがわかる。また、熊本駅の予測結果を詳しくみると、特にピーク時の利用者数の変動を正確に捉えている。通町筋の予測結果を詳しくみると、XGBoost モデルの15分単位の予測も高い精度を示している。しかし、ピーク時の利用者数の変動ではなく、オフピーク時の利用者数の変動を正確に捉えている。

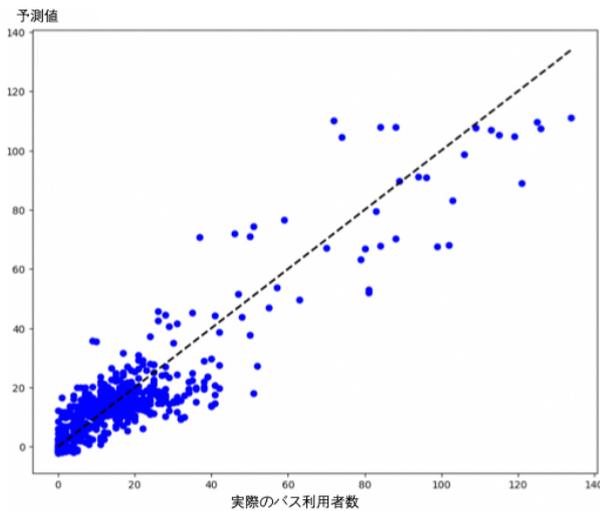


(a) 熊本駅 (LSTM)

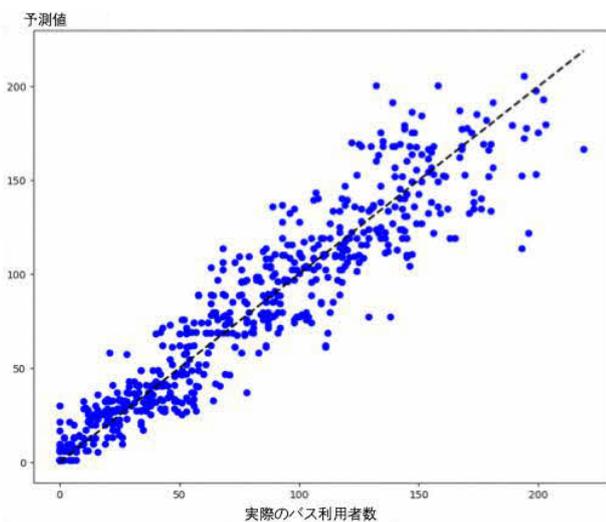


(b) 通町筋 (XGBoost)

図3 予測結果



(a) 熊本駅 (LSTM)



(b) 通町筋 (XGBoost)

図4 各モデルの予測精度

図4は各モデルの予測精度を示しており、モデルの予測が実際のバス利用者数から乖離する度合いを反映している。図4(a)では日単位のLSTMモデル、図4(b)では15分単位のXGBoostモデルの予測精度を比較している。図4(a)の点の分布がより密集している。図4(b)の点は、実際の利用者数の範囲全体にわたって広がっており、回帰直線に対して比較的一定の散布を示している。これは、モデルが全体として安定した予測を行っていることを示している。そのため、15分単位のモデルは予測精度が大幅に向上したことが分かった。日単位と15分単位の比較から、15分単位のモデルの方が予測性能に優れていることが分かった。これは日単位のデータでは、15分単位のデータと比べて、利用者数に大きな差がある日が存在し、予測が難しくなるため

と考えられる。

7 結論

本研究では、実際のバスの利用データを用いて、バス利用者数の予測を行い、熊本市に最も適切なモデルを特定した。熊本市に最も適切なモデルを1つに絞るのではなく、予測対象となるバス停や時間スケールに応じて適切なモデルを選択するアプローチが有効であることが示された。具体的には熊本市の熊本駅前、通町筋の2つのバス停をケーススタディとして取り上げ、短期のバス利用者数を予測した。予測にあたっては、過去のデータからバス利用者数のデータを抽出し、週単位、日単位、15分単位の3つの時系列モデルを構築した。時系列の定常性、周期性、および変動性¹⁶を調査し、適切な予測モデルを特定した。また、時系列の異なる特徴を捉えるために、ARMA、XGBoost、LSTMの予測モデルを利用した。

結果として、熊本駅前では日単位のLSTMモデル、通町筋では15分単位のXGBoostモデルが最も高い予測精度を示した。熊本市における最適な予測モデルは、バス停の特性や時間スケールに応じて選択されるべきであることが示された。

今後の研究の展開としては、モデルの性能と実用性を向上させ、より精度の高い予測について、取り組んでいきたい。具体的には、バス停ごとに乗降客数のパターンが異なるため、各バス停の特性に基づいたクラスタリングを行い、それに応じて最適な予測モデルを選択するアプローチが考えられる。また、短期間の予測精度を高めることを目指すとともに、リアルタイムデータの活用を積極的に行うことが望まれる。

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・資料)

- (1) Erfan, H., Ali, F., Alireza, A., Mehrdad, J., and Raghavendra, C. (2023) "A data analytics framework for reliable bus arrival time prediction using artificial neural networks", *Int J Data Sci Anal*, <https://doi.org/10.1007/s41060-023-00391-y>
- (2) Li, X., Lo, H.K. (2014) "An energy-efficient scheduling

and speed control approach for metro rail operations”
Transportation Research Part B: Methodological, Vol. 70,
pp 269–284.

- (3) Hu, M. (2011) “A Survey and Simulation of Passenger Flow Organization of the Shenzhen Urban Rail Transit Station”, *In ICCTP 2011: Towards Sustainable Transportation System*, pp. 2991–2997.
- (4) Carlos, A. R., Pablo, C., Edwan, A. A. E., Diego, V. (2024) “Strategies for Reducing Automobile Fuel Consumption” *Appl. Sci.*, Vol. 14, 910.
- (5) AlKhereibi, A.H. Wakjira, T.G. Kucukvar, M. Onat, N.C. (2023) “Predictive Machine Learning Algorithms for Metro Ridership Based on Urban Land Use Policies in Support of Transit-Oriented Development”, *Sustainability* 2023, 15, 1718. <https://doi.org/10.3390/su15021718>
- (6) 戸崎 肇 (2022) 「AI を中心とした 先端技術の交通分野における導入について」『産業総合研究』Vol. 30、17–32
- (7) 角野 惇、猪井 博登、土井 健司 (2014) 「交通系 IC カードデータを用いた路線バス需要予測」『土木計画学研究発表会・講演集』50、ROMBUNNO. 217
- (8) 青地 美桜、白石 陽 (2020a) 「交通系 IC カードの乗降データを用いた路線バスの乗客人数予測手法の提案」『第 82 回全国大会講演論文集』2020、353–354
- (9) 青地 美桜、白石 陽 (2020b) 「利用者の乗降データを用いたバス乗客人数予測モデルの評価」『研究報告モバイルコンピューティングとパーベイシブシステム (MBL)』2020-MBL-97、1–8
- (10) Yang, X., Xue, Q., Ding, M., Wu, J., Gao, Z. (2020) “Short-term prediction of passenger volume for urban rail systems: A deep learning approach based on smart-card data”, *International Journal of Production Economics*, Vol. 231, 107920.
- (11) Lv, W., Lv, Y., Ouyang Q., Ren Y. (2022) “A Bus Passenger Flow Prediction Model Fused with Point-of-Interest Data Based on Extreme Gradient Boosting”, *Applied Sciences*, Vol. 12, 940.

¹ 交通流 (traffic volume) とは、ある地点を単位時間当たりに通行する車両の量である。福田正『交通工学』(初版) 朝倉書店 (2024 年 7 月 20 日参照)

² 統計モデルは、データからパターンや関係性を見出すための数学的枠組みを提供するものであり、特定の仮説に基づいてデータを分析・予測するために用いられる。

³ 機械学習モデルは、データに基づいて予測や分類を行うために使用されるアルゴリズムや統計モデルの集合であり、経験的データから自動的にパターンを学習して性能を向上させることを目的としている。「Pattern Recognition and Machine Learning」by Christopher Bishop (2024 年 8 月 20 日参照)

⁴ 深層学習モデルは、多層にわたるニューラルネットワークを用いてデータから高度な特徴を抽出・学習する手法であり、特に画像認識や自然言語処理などの分野で顕著な成果を上げている。「Deep Learning」by Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2024 年 8 月 20 日参照)

⁵ POI とは、「Point of Interest」の略であり、直訳すると「興味のあるポイント」だが、「地図上の特定のポイント (地点)」のことを指す。ESRI ジャパン <https://www.esri.com/gis-guide/gis-other/point-of-interest/#:~:text=POI%20%E3%81%A8%E3%81%AF,POI%20%E3%81%AB%E3%81%AA%E3%82%8A%E3%81%88%E3%81%BE%E3%81%99%E3%80%82> (2024 年 7 月 8 日参照)

⁶ ポテンシャル理論は、数学において、ポテンシャル関数を用いて場の行動や分布を解析する理論である。「Real Analysis」by H. L. Royden (2024 年 8 月 20 日参照)

⁷ ランダムフォレストは、複数の決定木を組み合わせたアンサンブル学習手法であり、バギングとランダム特徴選択を用いることで、モデルの汎化性能を向上させることができる。「Pattern Recognition and Machine Learning」by Christopher Bishop (2024 年 8 月 20 日参照)

⁸ 一般化線形モデルは、線形回帰モデルを拡張したもので、応答変数が正規分布に従わない場合でも、異なる分布 (例えば二項分布やポアソン分布) に基づいた解析を行うことができる。

「Generalized Linear Models」by John Nelder and Robert Wedderburn (2024 年 8 月 20 日参照)

⁹ 一般化加法モデルは、一般化線形モデルの一種であり、線形部分に加え、スプラインや平滑化関数を用いて非線形な関係をモデル化することができる。「Generalized Additive Models」by Trevor Hastie and Robert Tibshirani (2024 年 8 月 20 日参照)

¹⁰ RNN (リカレントニューラルネットワーク) は、時系列データやシーケンスデータを扱うために設計されたニューラルネットワークで、過去の情報を保持しながら次のステップにフィードバックする構造を持ち、自然言語処理や音声認識に広く応用されている。

「Deep Learning」by Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2024 年 8 月 20 日参照)

¹¹ R^2 (決定係数) は、回帰モデルがどの程度データを説明できているかを示す指標であり、0 から 1 の間の値を取り、1 に近いほどモデルがデータに適合していることを示す。

¹² AIC (Akaike’s Information Criterion) は、赤池弘次博士が、数理統計理論を駆使して編み出したモデル選択指標のパラダイムである。 <https://www.kyoritsu-pub.co.jp/book/b10010581.html> (2024 年 7 月 8 日参照)

¹³ RMSE は誤差の平方を平均したものの平方根である。

¹⁴ MAE は絶対誤差の平均値である。

¹⁵ MAPE は予測値と実際の値との差異をパーセンテージで表したものである。

¹⁶ 変動性 (volatility) とは、時系列データにおける値の変動の大きさや不規則性を指す。

<https://www.tandfonline.com/doi/abs/10.1080/00031305.2017.1380080> (2024 年 7 月 20 日参照)

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)年度から令和6年3月まで熊本市都市政策研究所にて調査研究活動等に従事。令和6(2024)年度より至誠館大学現代社会学部(講師)。