修士論文

バス到着時刻予測のための走行時間・停車時間の日周期性を考慮した欠損値補完手法の提案

丹羽 拓実
奈良先端科学技術大学院大学
先端科学技術研究科
情報理工学プログラム

主指導教員: 藤川 和利 教授
（情報科学領域）

令和 5 年 1 月 31 日提出

本論文は奈良先端科学技術大学院大学先端科学技術研究科に
修士 (工学) 授与の要件として提出した修士論文である。

丹羽 拓実

審査委員：

藤川 和利 教授	（主指導教員，情報科学領域）
安本 慶一 教授	（副指導教員，情報科学領域）
新井 イスマイル 准教授	（副指導教員，情報科学領域）

バス到着時刻予測のための走行時間・停車時間の日周期性を考慮した欠損値補完手法の提案*

丹羽 拓実

内容梗概

路線バスにおいて，遅延や早着により乱れやすいバス到着時刻を予測し，バスの利用者にその予測到着時刻を伝えることには，利用者が待ち時間の少ないルートを選択できる利点がある．既存のバス到着時刻予測の研究では走行時間や停車時間といったバス運行データを利用した予測を前提としているが，バス運行データの計測では頻繁にデータの欠損が発生する．既存研究では，バス運行データの欠損値補完手法として Last observation carried forward（LOCF）といった単純な手法を採用しており，バス運行データの特徴に着目した欠損値補完手法は検討されてこなかった．一方，バス到着時刻予測と近い研究分野である渋滞予測の分野では，データの特徴に着目した欠損値補間手法を適用することで予測誤差が減少したと報告されている．そこで本研究では，バス到着時刻予測の誤差削減を目指し，バス運行データの特徴に着目した時間的補完，パターン補完，およびそれらの組み合わせ補完の 3 つの欠損値補完方法を提案する．時間的補完は，直近の運行の乱れを反映できることを期待し，直前数運行分の走行時間や停車時間の平均値を使用する手法である．パターン補完は，日ごとの運行の周期性を反映できることを期待し，欠損したデータと同じ時間に発車するバスの走行時間や停車時間の平均値で補完する手法である．組み合わせ手法は，時間的補完，パターン補完の弱点を互いに補えるように，欠損が連続する部分ではパターン補完を，連続しない部分では時間的補完を行う手法である．評価では，既存の欠損値補完手法と提案手法それぞれを用いて補完したデータセットを用いてバス到着時刻予測を

*奈良先端科学技術大学院大学 先端科学技術研究科 修士論文, 令和 5 年 1 月 31 日.

行い，どの欠損値補完手法を用いた際に予測誤差が小さくなるかを検証した．評価の結果，1 運行先の予測では LOCF を用いた際にバス到着時刻予測の誤差が最も小さかった．また，2 運行先と 3 運行先の予測では提案手法であるパターン補完を用いた際にバス到着時刻予測の誤差が最も小さいことが確認できた．

キーワード

高度道路交通システム，バス到着時刻予測，欠損値補完，機械学習

Proposal of Missing Data Imputation Focusing on Daily Periodicity of Bus Running Time and Stopping Time for Bus Arrival Time Prediction*

Takumi Niwa

Abstract

Providing a predicted bus arrival time (BAT) allows bus users to choose a route with less wait time. Existing studies on BAT prediction use bus travel data such as bus running time and stopping time, which are frequently missing in their measurements. Although existing studies often adopt simple missing data imputation methods like last observation carried forward (LOCF), missing data imputation methods that focus on the characteristics of bus travel data have not been studied. On the other hand, in studies on traffic congestion prediction, which is similar to BAT prediction, it was reported that the prediction accuracy was improved by applying a missing data imputation method focusing on data characteristics. In this research, we aim to improve the accuracy of BAT prediction by using a missing data imputation method that focuses on characteristics of bus travel data. We propose three types of missing data imputation methods: temporal imputation, pattern imputation, and combined imputation. Temporal imputation uses means of running and stopping times of several travels before missing data to reflect recent disruptions in bus service. Pattern imputation uses means of running and stopping times at the same hour as a bus service where missing data occurs to reflect the daily periodicity. Combined imputation compensates for a weaknesses

*Master's Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology, January 31, 2023.

of temporal imputation and pattern imputation by using pattern imputation for consecutive missing data and temporal imputation for non-consecutive missing data. We compared the proposed methods with simple missing data imputation methods for BAT prediction. We found that the prediction using LOCF has the highest accuracy for BAT of next bus travel. We also found that the prediction using pattern imputation has the highest accuracy for BAT of the second and third bus travel from current one.

Keywords:

Intelligence Transport System, Bas Arrival Time Prediction, Missing Data Imputation, Machine Learning

目 次

図 目 次

表 目 次

1. はじめに

世界各地の多くの人々は，毎日の通勤や通学，観光など様々な用途で路線バスを利用している．路線バスのサービス品質向上の手段の 1 つとして，利用者へのバス到着時刻予測の提供があげられる．予測したバス到着時刻を提供することにより，利用者は遅延や早着を考慮した，より待ち時間の少ないルートを利用することが可能となる．またバス運行会社にとっても，バス到着時刻を予測することはバス運行スケジュールの管理や運行効率の評価に役立つ．このように，路線バスにおけるバス到着時刻予測は重要な意味を持つ．

研究者らによるバス到着時刻予測のこれまでの研究では，確率モデル，履歴モデル，統計モデル，人工知能ベースモデルといった様々な予測モデルが試みられてきた [1]．近年では，人工知能ベースモデルに該当する深層学習モデルを用いた研究が盛んである．そうした研究の多くは，Recurent neural network (RNN) や Long-shrot term memory (LSTM) といった時系列データの学習において有効とされるモデルの改善によって，バス到着時刻予測の誤差を削減することを目指している．

しかし，こうした時系列データを学習する深層学習モデルを用いるバス到着時刻予測においてバス運行データの欠損をいかに補完すべきかということはあまり議論されてこなかった．路線バスで行われるバス運行データの計測では，通信エラーによるデータの欠落や機器の不調によるデータの欠損は頻繁に発生する．時系列データの順序や時間的な傾向が重要となる LSTM 等の深層学習モデルではバス運行データの欠損を無視できないため，そうした欠損は，予測モデルの学習やそのモデルを用いた推論の前に補完する必要がある．こうした欠損値補完に関して，これまでの研究では，Last observation carried forward（LOCF）といった単純な手法を用いるにとどまっている [2, 3]．

一方，バス到着時刻予測と同じ高度交通システム（Intelligence transport system, ITS）の研究分野である渋滞予測において，Shin らはデータの特性に合わせた欠損値補完手法を適用することで予測誤差を削減できたと主張した [4]．このことからバス到着時刻予測においても，バス運行データの特徴に着目した欠損値補完手法を用いることで，予測誤差を削減できると考える．

そこで，本研究ではバス運行データの特徴に着目した欠損値補完手法によるバス到着時刻予測の誤差削減を目指す．バス運行データの特徴に着目した手法として，Shin らの手法を参考に時間的補完，パターン補完，およびそれらを組み合わせた手法の 3 つの欠損値補完手法を提案する．時間的補完は，直近の運行の乱れを反映することを考慮し，欠損が発生した運行の直前数運行分の走行時間や停車時間の平均値を使用する手法である．パターン補完は，日ごとの運行の傾向を反映することを考慮し，欠損が発生した運行と同じ時間に運行する便の走行時間や停車時間の平均値で補完する手法である．組み合わせ手法は，時間的補完，パターン補完の弱点を互いに補えるように，欠損が連続する部分ではパターン補完を，連続しない部分では時間的補完を行う手法である．評価として，兵庫県神戸市内のバス路線の実運行データに対しこれらの提案手法や単純な欠損値補完手法を適用し，石長らの手法で到着時刻予測をした際の予測誤差を比較した．評価の結果，複数運行先のバス到着時刻を予測する場合には，欠損値補完手法としてパターン補完が最も効果的であると確認できた．

本稿の構成は以下のとおりである．第 2 章では，バス到着時刻予測と時系列データ学習における欠損値補完手法の関連研究について説明する．第 3 章では，本研究で想定するバス到着時刻予測について説明する．第 4 章では，本研究で使用するバス運行データ，および気象データについて述べる．第 5 章では，本研究で提案する欠損値補完手法について述べる．第 6 章では，欠損値補完手法がバス到着時刻予測に与える影響を評価する．第 7 章では，評価結果を元に提案手法を考察し，今後の展望を述べる．第 8 章で本稿をまとめる．

2. 関連研究と課題

本章では，路線バスにおいて正確な到着時刻予測が必要な背景，およびバス到着時刻予測の関連研究について述べる．その後，時系列データ予測における欠損値補完の関連研究について紹介する．最後にバス到着時刻予測における欠損値補完について述べる．

2.1 路線バスにおいて正確な到着時刻予測が必要な背景

Chocholac らは，持続可能な都市物流の観点から都市圏内における自家用車の利用者数を減らし，路線バスなどの公共交通機関の利用者数を増やすことが必要であると主張した [5]．また Chocholac らは，公共交通機関の利用者数を増やすためにはそのサービス品質を向上させる必要があるとも述べている．公共交通機関のサービス品質を向上させる方法の 1 つとして，利用者への予測到着時刻の提供が挙げられる．予測到着時刻を提供することで，複数の公共交通機関を利用する際にもより待ち時間が少ないルートを提示できる．既に世界中の公共交通機関においてこのようなリアルタイム情報の提供は行われており，駅やバス停などのデジタルサイネージや Web サイト，地図アプリケーションなどを通じて，利用者は予測到着時刻を知ることができる．

一方で，Gooze らは路線バスにおいて大きな誤差を含む予測到着時刻の提示はかえって利用者の混乱を招き，一定数の利用者の乗車率が低下することを示した [6]．Gooze らはワシントン州シアトル広域の交通情報をリアルタイムに予測するツール群である OneBusAway の利用者を対象にアンケートを行った．このアンケートの回答者 5,074 人のうち 3,866 人（77%）が「過去に誤った予測到着時刻を提供された」と回答し，そのうちの 9%の人が「誤った予測到着時刻の提供を理由にバスの乗車回数を減らした」と回答した．また，Gooze らは同じアンケート内で「OneBusAway の予測到着時刻が何分異なる場合に，誤った予測到着時刻を提供したと判断するか」という質問をした．その結果を図 1 に示す．図に示されているように，最も多かった回答は 4〜5 分（37%）であり，次点が 2〜3 分（27%）である．以上のことから，サービス品質を向上させ路線バスの利用者を増やすた

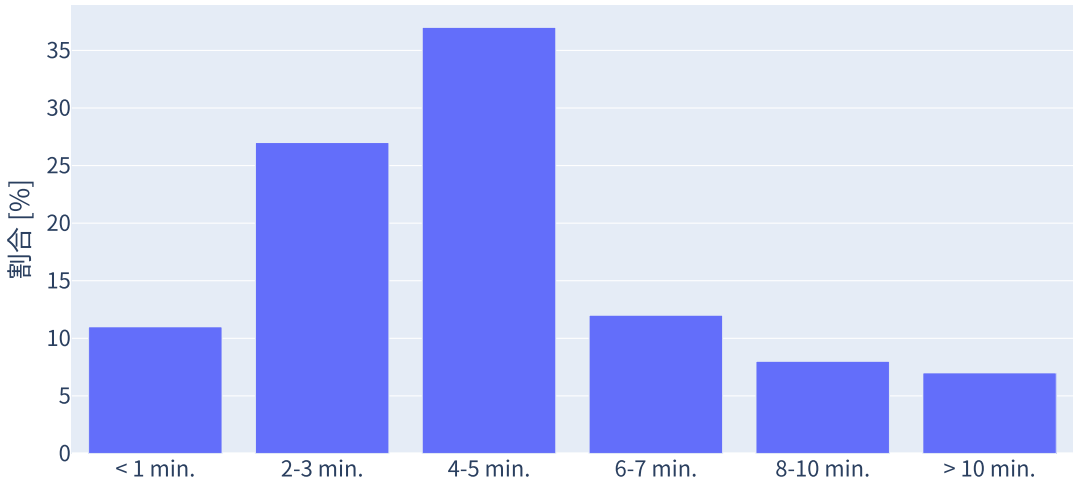


図 1: 到着時刻予測の誤差に対する意識調査の結果 [6]

めには，より予測誤差の小さいバス到着時刻予測手法が必要であると言える．

2.2 バス到着時刻予測の関連研究

路線バスは乗降者数，周辺の道路状況，天候など様々な要素が複雑に関係しているため，鉄道などの他の公共交通機関と比較して正確な到着時刻予測が難しい．そのためバス到着時刻予測の分野では様々な研究が行われてきた．

Nisha らは 2000 年〜2021 年に行われたバス到着時刻予測の研究 70 件のうち，45 件が人工知能を用いたものであり，そのうちの 16 件が深層学習を用いたものであることを指摘している [1]．また深層学習モデルは，その高い柔軟性，適応性，非線形性，データ駆動型のモデル構築により，確率論的モデルや統計モデルよりも高い性能を示すとも述べている．

深層学習を用いたアプローチでは，Recurent neural network (RNN) や Long short-term memory (LSTM) を用いた手法が注目されている．LSTM は過去数ヶ月〜1 年といった長期間の運行データから学習ができるとされ，LSTM を用いたバス到着時刻予測手法が Pang ら [7] や Han ら [8] によって提唱されている．

さらに，LSTM を拡張するモデルである Convolutional LSTM を用いたバス到着

時刻予測も検討されている．Convolutional LSTM は重み演算内の全結合を畳み込み演算に変更したもので，空間的な時系列の特徴を捉えることができる．Xie らは複数の RNN モデルによるバス到着時刻予測を比較した結果，Convolutional LSTM が最も小さい誤差をを示すと報告した [9]．Petersen らは Convolutional LSTM に対し過去の複数運行分のバス運行データを入力し，次以降の複数運行分のバス到着時刻を予測できる手法を提案した [2].

石長らは Petersen らの手法に改良を加えた [3]．まず石長らは Petersen らの手法に 2 つの問題点を指摘した．1 つ目は一定間隔の頻度で走行する路線を対象に評価したため早着によるバス停での意図的な停車など時刻表ベース特有の特徴が考慮されていない点，2 つ目は雨などの天候の影響によって運行が乱れた際に予測誤差が大きく増加する点である．そこで石長らは，バス運行データと気象データを結合して Convolutional LSTM に入力し，走行時間と停車時間を別々に予測することでこれらの問題を解決した．また，予測モデルの性能向上のために，中間層の出力を未来への順伝播と過去への逆伝播の両方向で伝播する Bidirectional LSTM を Convolutional LSTM に適用している．兵庫県神戸市内のバス路線の実運行データを使用し評価実験を行った結果，1 週間の予測において 1.36%の平均絶対パーセント誤差の減少が見られた．特に，夕方の時間帯や雨天時など，運行が大きく乱れた際に約 3%と大きく平均絶対パーセント誤差を減少させることができた．以上の結果から，Convolutional LSTM を用いて運行データと気象データを同時に畳み込み，走行時間と停車時間を別々に予測する手法が，バス到着時刻予測において有効であることを示した.

2.3 時系列データ予測における欠損値補完の関連研究

時系列の特徴を学習できる LSTM は，バス到着時刻予測に限らず幅広い分野の時系列予測に用いられている．しかし，現実の時系列データの計測においては，欠損のない完全なデータセットを得ることは非常に困難であり，少なからず欠損値を含んでしまう．一般に時系列データを用いた予測では，データに欠損が含まれると予測できなくなるため，データセットに含まれる欠損値を補完する必要がある.

単純な欠損値補完手法としては，Last observation carried forward (LOCF) や線形補間があげられる．LOCF は，ある欠損値を，その欠損が発生する以前でかつ最後に観測された値で置換する手法である．欠損が連続で発生した場合は，それらの欠損値が全て同じ値で置換される．線形補間は，ある欠損値を，その前後に観測した 2 つの値を用いた単回帰分析から得られる値で置換する手法である．これらの補完手法は単純であるため，分野を問わず多くの研究で利用されている．

単純な欠損値補完手法だけではなく，使用する時系列データの特徴に合わせた欠損値補完手法を提案する研究も行われている．一部の研究分野では，欠損値補完手法によって時系列予測の誤差を削減できたと報告されている．

Lee らは，LSTM に基づく EV 充電スタンド負荷予測モデルの予測誤差削減のために EV 充電データの欠損値補完手法を提案した [10]．この負荷予測モデルは，過去の EV 充電スタンドの使用状況から学習し将来の EV 充電スタンドの負荷を予測するものであるが，過去の EV 充電スタンドの使用状況のデータセットに欠損が生じてしまうという問題がある．Lee らは，スプライン補完と EM アルゴリズムを組み合わせた 2 段階の補完手法を提案した．評価実験として，従来の補完手法と提案手法を欠損率の異なるデータセットにそれぞれ適用し，それらを元にモデルを学習させ，モデルの予測誤差を比較した．結果として，欠損率が 16%を超えるデータセットにおいて提案手法の優位性が示された．

また，Shin らは LSTM に基づく渋滞予測手法において，空間的傾向・時間的傾向・パターンデータを用いた欠損値補完手法を提案した [4]．この渋滞予測手法においては，路上や沿道に設置された交通情報収集装置で収集した情報を元に過去の道路状況を学習し，将来の渋滞予測を行う．しかし収集する過去の道路状況には欠損が生じるという問題がある．Shin らは，空間的補完・時間的補完・パターン補完の 3 段階の補間手法を提案した．この補間手法では，まず欠測地点の隣接する道路状況を用いて補完する空間的補完を行う．隣接する道路のデータも欠損している場合は空間的補完が適用できないため，欠測地点での n 個前の欠測地の平均値を使用して補完する時間的補完を行う．これで理論上は全ての交通状況を補完可能であるが，連続で時間的補完を行うと，時間帯による動向が失われるため，長時間連続で欠損する場合は，曜日ごとに予め生成したパターンデータを当

てはめるパターン補完を行う．評価実験として，欠損率を 10%〜90%まで変化させたデータセットに対し従来の欠損値補完方法，提案手法による補完をそれぞれ適用し，それらを元にモデルを学習させ，モデルの予測誤差を比較した．結果として，提案手法が平均 2 乗パーセント誤差を最も低く抑えることができたと報告した．

2.4 バス到着時刻予測における欠損値補完

バス運行データは時系列データであるため，バス到着時刻予測においても欠損値補完が必要である．特に，予測モデルへ入力するバス運行データが欠損すると，そもそも予測できない問題がある．例えば石長らの手法は，過去複数運行分の連続したバス運行データが必要であり，石長らの評価実験では 8 運行分のバス運行データを入力に使用した．仮にバス運行データが 10%の確率で欠損するとした場合，予測モデルに入力できる確率は約 43.0%[1]となる．このように，バス運行データ自体の欠損が少ない場合でも予測モデルを使用できない場合がある．

調査した範囲のバス到着時刻予測の既存研究においては，具体的な欠損値補完手法には言及していないものがほとんどであった．また，Petersen らや石長らの研究のように欠損値補完に言及する研究もあるが，LOCF といった単純な手法を採用するにとどまっている．しかし，渋滞予測など他の分野と同様に，バス到着時刻予測の分野においても学習や予測に使用するデータの特徴を考慮した欠損値補完手法によって，予測誤差を削減できるのではないかと考える．そこで本研究では，バス運行データの特徴を考慮した欠損値補完手法を検討する．

[1] $(1-0.1)^8 \sim 0.430$

3. 想定するバス到着時刻予測

本研究では，バス到着時刻予測手法として石長らの手法 [3] を想定し，バス運行データの特徴に着目した欠損値補完によってバス到着時刻予測の誤差削減を目指す．本章では，まず石長らの手法で想定するバス路線，および使用するデータについて説明する．その後，石長らの手法でどのようにバス到着時刻を予測するか詳細を述べる．

3.1 想定する路線バス

石長らの手法では，以下のような特徴がある路線バスを想定して到着時刻を予測する．

始点バス停から終点バス停までの片道運行

一部のバス路線には始点や終点が存在せずループ運行するものも存在するが，本手法では考慮しない．また，折返しの運行は別路線として扱う．

単一の時刻表による運行

1 日に運行される便数が固定されており，異なる日付でも便番号 t が同じであれば同一時刻に運行されるものとする．1 日に運行される便数を T_{day} とする．なお，平日と休日などで異なる時刻表を用いる場合は別路線として扱う．

始点バス停を定刻どおりに出発

多くのバス路線では，始点のバス停には余裕を持って到着できるように運行ダイヤが設定されているため，始点バス停は定刻どおりに出発できるものとする．

図 2 に想定する路線バス運行における走行時間 r・停車時間 s・時刻表差分 d・所要時間 l を示す．石長らの手法では，バス運行データとして走行時間・停車時間・時刻表差分の 3 つを扱い，任意のバス停に到着するまでの所要時間を予測する．路線上にバス停が B 個存在する場合，図に示すように $B-1$ 個の走行区間が

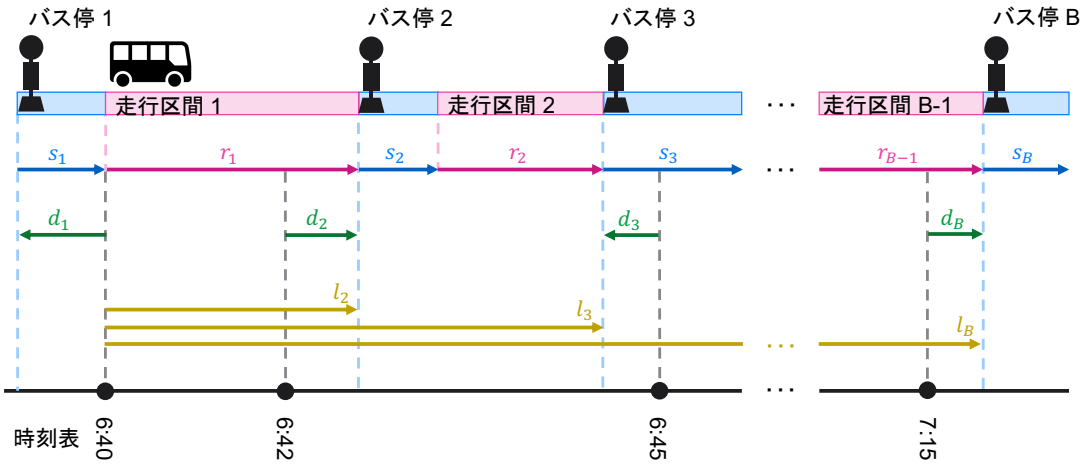


図 2: 想定する路線バス運行における走行時間 r・停車時間 s・時刻表差分 d・所要時間 l

ある．走行区間 b の走行にかかった時間を r_b，バス停 b に停車した時間を s_b，バス停 b に時刻表と比較して遅れた時間を d_b とする．時刻表差分に関して，時刻表の時刻より早く到着した場合は負数で表現する．始点バス停を出発しバス停 b に到着するまでの所要時間を l_b とする．l_b は式 (1) のように計算できる．

$$l_b = \begin{cases} r_1 & \text{if } b = 2, \\ r_1 + \sum_{i=2}^{b-1} (s_i + r_i) & \text{if } b > 2. \end{cases} \tag{1}$$

本手法では，バス停 1 を定刻どおりに出発できるものとし，バス停 2 からの到着時刻を予測対象とするため，バス停 1 への到着時刻は取り扱わない．

3.2 使用するデータセットの構造

石長らの手法では，バス運行データと気象データの 2 種類のデータを使用しバス到着時刻予測をする．本節では，それらのデータセットであるバス運行データセットと気象データセットの構造について述べる．

バス運行データセットの例を表 1 に示す．バス運行データセットは，各走行区間ごとの走行時間，各バス停ごとの停車時間と時刻表差分の 3 つの表からなる．

表 1: バス運行データセットの例（6 個のバス停がある路線の場合）

(a) 走行時間 r（秒）

index	日付	便番号 t	r_1	r_2	r_3	r_4	r_5
1	2022-06-01	1	93.5	128.5	72.0	1008.5	1355.5
2	2022-06-01	2	125.0	153.0	80.0	1276.5	456.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
26	2022-06-01	26	82.5	105.5	51.5	966.0	1331.0
27	2022-06-02	1	85.0	94.0	48.5	872.5	1621.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(b) 停車時間 s（秒）

index	日付	便番号 t	s_1	s_2	s_3	s_4	s_5	s_6
1	2022-06-01	1	28.0	56.5	13.5	26.5	28.0	36.0
2	2022-06-01	2	60.0	21.0	0.0	56.5	30.5	16.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(c) 時刻表差分 d（秒）

index	日付	便番号 t	s_1	s_2	s_3	s_4	s_5	s_6
1	2022-06-01	1	−18.5	−16.0	108.9	74.4	−90.5	692.9
2	2022-06-01	2	−58.9	16.1	130.1	90.1	223.1	109.6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

表 2: 気象データセットの例

日時	気温 (°C)	降水量 (mm)	晴れ フラグ	曇り フラグ	雨 フラグ
2022-06-03 06:00	21.0	0.0	1	0	0
2022-06-03 07:00	21.6	0.0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮
2022-06-05 19:00	20.5	0.5	0	0	1
2022-06-05 20:00	20.0	1.5	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮

各表には運行された日付，その日のうち何便目かを示す便番号，走行区間やバス停ごとの走行時間・停車時間・時刻表差分が入る．index は各データの通し番号であり，運行数と一致する．

石長らの手法で用いる気象データセットの例を表 2 に示す．このデータセットの各行は，日時，気温，降水量，晴れフラグ，曇りフラグ，雨フラグの 6 つで構成される気象データを表す．晴れフラグ，曇りフラグ，雨フラグの 3 つのフラグはその日時における天気を表し，その日時の天気に該当するフラグの値のみが 1 となり，その他の値は 0 となる．

3.3 バス到着時刻予測の流れ

本節では，石長らの手法におけるバス到着時刻予測の流れを説明する．石長らの手法では，過去のバス運行データとその日時に対応する気象データを結合して入力し，Convolutional LSTM を用いた予測モデルで走行時間と停車時間を別々に予測する．その後，予測された走行時間と停車時間から予測到着時刻を計算する．なおこの際に入力する過去の運行データ・気象データは過去複数運行分に対応するものであり，出力する予測到着時刻も複数運行分である．一度の予測で入力する運行データの運行数を N_{in}，出力する予測到着時刻の運行数を N_{out} とする．

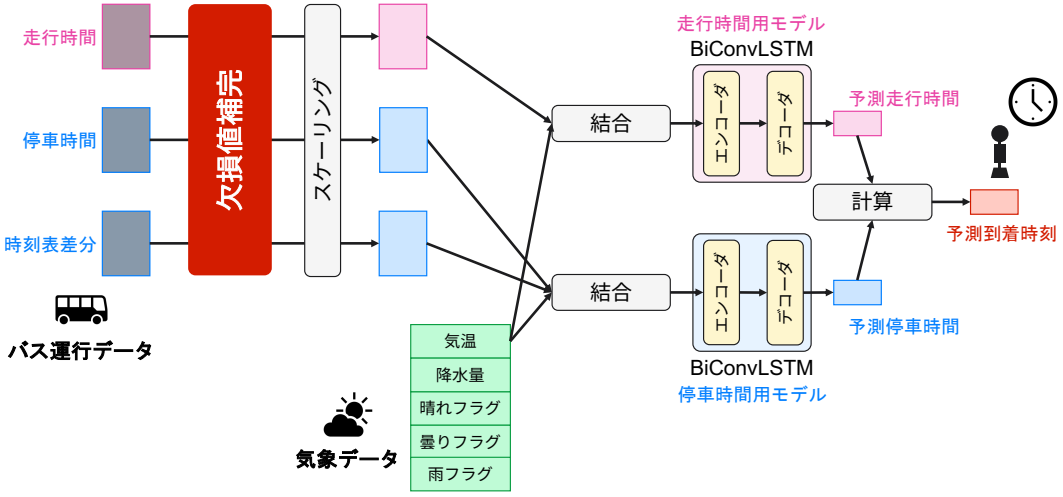


図 3: バス到着時刻予測の流れ

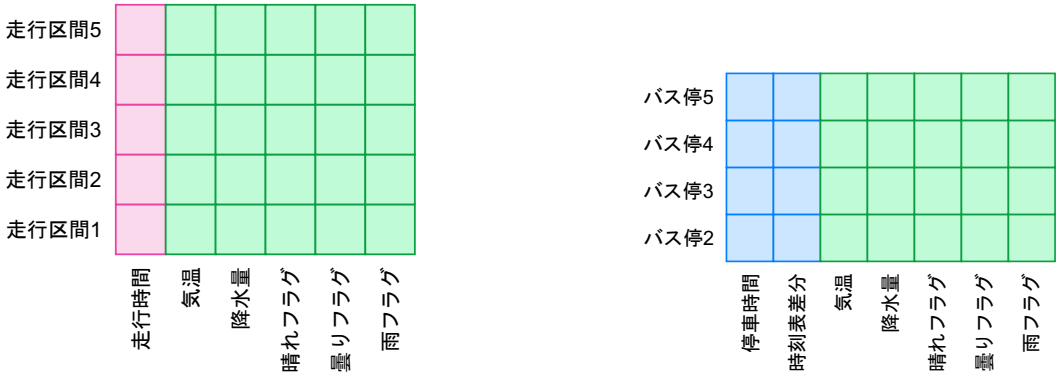
N_{in} はその路線の 1 日の運行数に応じて決定されるべきパラメータである．

バス到着時刻予測の流れを図 3 に示す．まず，バス運行データセットに含まれる欠損を補完する．次に，学習の前処理として各特徴量をスケーリングし，運行データと気象データを結合する．その後，走行時間・停車時間を Convolutional LSTM を用いたモデルを使って予測する．この際に使用するモデルは，走行時間・停車時間それぞれ別のモデルを使用する．走行区間ごと，バス停ごとに走行時間，停車時間を予測した後，式 (1) を用いて各バス停までの予測所要時間を計算する．その後，始点バス停を定刻どおりに出発するとして各バス停への予測到着時刻を得る．

本手法における予測対象のバスは，次以降に出発する複数台のバスであり，すでに走行中のバスは対象外とする．

3.4 使用する特徴量

本節では，石長らの手法で学習時や予測時に使用する特徴量について述べる．3.3 節で述べたように，本研究では過去のバス運行データとその時間に対応する気象データを用いる．走行時間には 6 次元，停車時間には 7 次元の時系列特徴量を Convolutional LSTM を用いて畳み込んで学習し予測する．図 4 に走行時間と



(a) 走行時間　　(b) 停車時間

図 4: 走行時間予測，停車時間予測の各特徴量（6 個のバス停がある路線の場合）

停車時間の予測に用いる特徴量を示す．桃色は走行時間予測にのみ使用する特徴量，水色は停車時間予測にのみ使用する特徴量，緑色は共通で使用する特徴量を示す．気象データに関してはどちらも同じ特徴量を使用するが，バス運行データに関しては，走行時間と停車時間で異なる特徴量を使用しており，停車時間の予測にのみ時刻表差分を使用する．これは，時刻表に基づいて運行している路線バスにおいて，停車時間は時刻表から受ける影響が大きいと考えられるためである．例えば，バス停に早着した場合は乗降客がいなくても時刻表の発車時刻前にバス停を出発しないよう停車時間を調整する．一方走行時間は，走行区間の長さや制限速度，道路状況の影響が大きく，時刻表差分が与える影響は小さいと考えられる．以上のことから，停車時間の予測にのみ時刻表差分を使用する．また，停車時間の予測は始点と終点のバス停を除いて行う．これは，本手法の到着時刻予測では始点バス停を出発した後から，終点バス停に到着するまでの間のみに着目するためである．そのため，停車時間予測の特徴量にも始点バス停と終点バス停の停車時間と時刻表差分を使用しない．

3.5 学習の前処理

学習前の前処理では，データセットの値のスケーリングおよび運行データと気象データの結合を行う．本節では，それらの前処理について述べる．

3.5.1 バス運行データセットのスケーリング

バス運行データセットは，各走行区間や各バス停の走行時間・停車時間・時刻表差分ごとに同じ手法を用いてスケーリングする．すなわち，表 1 の列ごとにスケーリングする．説明のため $x_{b,n}$ をスケーリング前，$x'_{b,n}$ をスケーリング後の値とする．b は走行区間やバス停の識別子，n はサンプルの識別子を示す．

まず，各走行区間（バス停）の便番号（t）に基づく平均値 $\bar{x}_{b,t}$ と，各走行区間（バス停）の標準偏差 σ_b を求める．この際，極端な外れ値による影響を避けるため，中央絶対偏差（Mean absolute deviation）[11] を用いた外れ値除去をする．

その後，式 (2) を用いてスケーリングする．

$$x'_{b,n} = \frac{x_{b,n} - \bar{x}_{b,t}}{\sigma_b} \tag{2}$$

このとき，n 番目のサンプルの便番号 t に対応する $\bar{x}_{b,t}$ を使用して計算する．

3.5.2 気象データセットのスケーリング

気象データのうち，気温と降水量は式 (3) のように四分位点を基準にしたスケーリングを行う．

$$x'_n = \frac{x_n - Q_2}{Q_3 - Q_1} \tag{3}$$

ここで，x_n はスケーリング前，x'_n はスケーリング後の値，n はサンプルの識別子を示す．Q_1，Q_2，Q_3 はそれぞれ第一四分位点，第二四分位点（中央値），第三四分位点を示す．

3.5.3 バス運行データと気象データの結合

各特徴量のスケーリング後，図 4 のようにバス運行データと気象データを結合する．バス運行データは走行区間（バス停）ごとに記録されるが，気象データは 1 時間単位のデータを使用しているため，各走行区間（バス停）の発車時刻（停車時刻）を元に一番近い日時の気象データを当てはめる．

3.6 到着時刻予測モデルの設計

図 5 に Convolutional LSTM を使用した到着時刻予測モデルの設計を示す．図では走行時間用の特徴量を使用し走行時間を予測しているが，停車時間の予測でも設計は同様であり，特徴量のみ異なる．大きくエンコーダとデコーダの 2 つに分

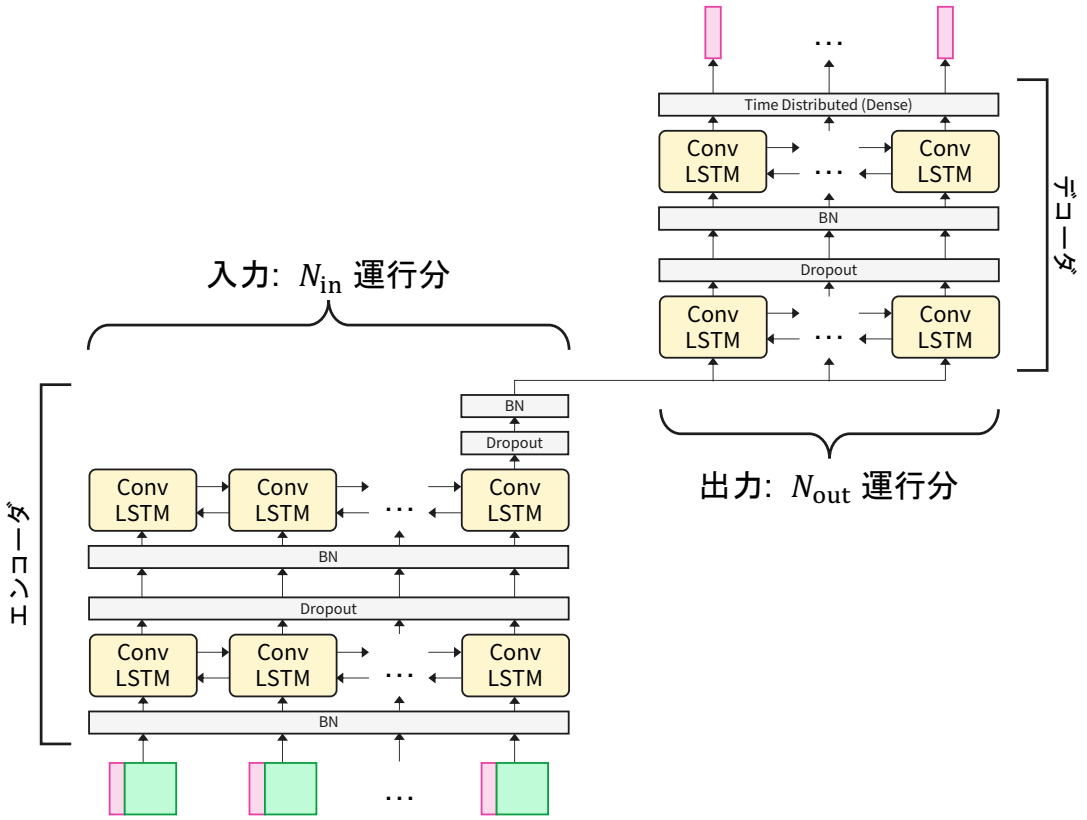


図 5: Convolutional LSTM を使用した到着時刻予測モデル

かれており，エンコーダとデコーダのそれぞれが Stacked 2-Layer Convolutional LSTM で構成されている．また予測モデルの性能向上のために，各 Convolutional LSTM 層（図中の Conv LSTM）には Bidirectional LSTM[12] を適用している．過学習を防ぐため，4 つの各 LSTM 層間に Dropout 層 [13] と Batch Normalization 層 [14]（図中の BN）を使用している．エンコーダは，入力として N_{in} 運行分の特徴量を受け取り，Batch Normalization 層を通して Convolutional LSTM 層に渡す．ここでは時系列データのまま処理する．その後 Dropout 層，Batch Normalization 層を通した後，2 つ目の Convolutional LSTM 層に渡す．2 つ目の Convolutional LSTM 層では入力末尾の運行に対応する出力のみを取り出し，Dropout 層，Batch Normalization 層を通してデコーダに渡す．デコーダはエンコーダから受け取った出力を N_{out} 回複製して入力する．こうすることによって N_{out} 運行先までの到着時刻予測を可能にしている．デコーダでも同様に 2 回 Convolutional LSTM 層を通した後，Time Distributed 層にデータを渡す．ここでは活性化関数に Rectified Linear Unit（ReLU）を用いて走行時間や停車時間の最終的な予測をする．オプティマイザーには RMSprop[15]，ロス関数には平均二乗誤差（MSE）を使用している．

4. 本研究で使用するデータ

本研究では，実際に運行されている路線バスにおいて，バス運行データを収集した．また，石長らの手法に必要となる気象データも収集した．これら 2 種類のデータから，本研究で使用するバス運行データセットと気象データセットを作成した．

本章では収集したバス運行データと気象データについて説明する．また，バス運行データの特徴を考慮した欠損値補完手法を提案するためのバス運行データの分析についても述べる．

4.1 バス運行データセットの対象路線・期間

本研究で利用するバス運行データセットの作成に必要なデータを収集する対象として，以下の 2 つの要件を満たす，兵庫県神戸市中心部を走行するみなと観光バス 21 系統上り路線を選んだ．

運行が乱れやすい路線

所要時間が安定している路線では過去の運行の平均値を利用するなど単純な予測手法が使用可能であり，補完手法によらず常に予測誤差が小さくなると予想される．そのため，評価対象としてふさわしい路線は，時間帯や天候などにより所要時間が変動する路線である．

データセットの欠損値がある程度少ない路線

データセットの欠損値処理が変化した場合の予測誤差を比較したいため，元のデータセットに大量の欠損が含まれる場合は，正解値が分からず予測誤差の比較できない．そのため，本研究では比較的欠損値が少ない路線のデータを使用し，そこに人為的に欠損を発生させることで欠損値補完による予測誤差の変化を確認する．

この路線において，2021 年 9 月 1 日から 2022 年 9 月 25 日の間に行われた計 9863 運行においてデータを収集した．

図 6: 評価に使用する路線の経路図

表 3: 21 系統上り路線のバス停一覧

No.	名称
1	神戸国際大学前
2	ウエストコート 7 番街前
3	六甲アイランド病院前
4	神戸ベイシェラトンホテル
5	神戸三宮
6	新神戸駅

表 4: 21 系統上り路線の時刻表

便番号 t	バス停					
	1	2	3	4	5	6
1	06:40	06:42	06:43	06:45	07:05	07:15
2	07:10	07:12	07:13	07:15	07:35	07:45
3	07:40	07:42	07:43	07:45	08:05	08:15
4	08:10	08:12	08:13	08:15	08:35	08:45
5	08:40	08:42	08:43	08:45	09:05	09:15
6	09:10	09:12	09:13	09:15	09:35	09:45
7	09:40	09:42	09:43	09:45	10:05	10:15
8	10:10	10:12	10:13	10:15	10:35	10:45
9	10:40	10:42	10:43	10:45	11:05	11:15
10	11:40	11:42	11:43	11:45	12:05	12:15
11	12:40	12:42	12:43	12:45	13:05	13:15
12	13:40	13:42	13:43	13:45	14:05	14:15
13	14:40	14:42	14:43	14:45	15:05	15:15
14	15:10	15:12	15:13	15:15	15:35	15:45
15	16:10	16:12	16:13	16:15	16:35	16:45
16	16:40	16:42	16:43	16:45	17:05	17:15
17	17:10	17:12	17:13	17:15	17:35	17:45
18	17:40	17:42	17:43	17:45	18:05	18:15
19	18:10	18:12	18:13	18:15	18:35	18:45
20	18:40	18:42	18:43	18:45	19:05	19:15
21	19:10	19:12	19:13	19:15	19:35	19:45
22	19:40	19:42	19:43	19:45	20:05	20:15
23	20:10	20:12	20:13	20:15	20:35	20:45
24	20:40	20:42	20:43	20:45	21:05	21:15
25	21:10	21:12	21:13	21:15	21:35	21:45
26	22:10	22:12	22:13	22:15	22:35	22:45

表 5: 21 系統上り路線の走行区間と予定所要時間

No.	出発バス停	到着バス停	予定所要時間 [分]
1	神戸国際大学前	ウエストコート 7 番街前	2
2	ウエストコート 7 番街前	六甲アイランド病院前	1
3	六甲アイランド病院前	神戸ベイシェラトンホテル	2
4	神戸ベイシェラトンホテル	神戸三宮	20
5	神戸三宮	新神戸駅	10
			合計: 35

この路線の経路を図 6 に示す．経路長は約 15km であり，主に都市部を走行する．この路線のバス停一覧を表 3，時刻表を表 4，走行区間一覧を表 5 に示す．この路線では，6 つのバス停が存在するため，5 つの走行区間を持つ．表 5 に示した各走行区間ごとの予定所要時間は時刻表で設定された所要時間である．この路線の時刻表では，到着時刻と出発時刻が同一時刻に設定されているため，最終区間である走行区間 5 を除き各走行区間の予定所要時間には到着バス停での停車時間も含まれる．また，平日と土日祝日で同じ時刻表が設定されている．

本研究で利用するバス運行データセットは，DOCOR システム [16] を用いて収集した GPS データ・速度データをもとに走行時間・停車時間・時刻表差分を計算し作成した．DOCOR システムでは，バス車内に設置された車載器によりバスの GPS データ・速度・エンジン回転数などのセンサデータを 0.5 秒間隔で収集する．収集したセンサデータは，セルラーネットワーク上を介し UDP プロトコルによりリアルタイムでサーバに送信される．サーバの GW がセンサデータを受信した後，MQTT によって DB にセンサデータが配信される．本研究では，DB に蓄積されたセンサデータを元にバス運行データセットを作成した．

4.2 気象データセットの対象地点・期間

気象データセットは，気象庁が公開している過去の気象データ [17] のうち，バス運行が行われている神戸市内の 1 時間ごとのデータから作成した．対象期間はバス運行データセットと同じく 2021 年 9 月 1 日から 2022 年 9 月 25 日までである．

気象庁の 1 時間ごとの観測では，気圧や風向など 16 項目が観測されており，本研究ではその中の天気，気温，降水量を使用する．天気は気象台の職員が目視で観測したものであり，気温と降水量は地域気象観測システム（アメダス）で観測したものである [18]．気象庁の観測する天気は 24 種類に分類されるが，本研究では，石長らの研究に従い表 6 のとおりに晴れ・曇り・雨の 3 種類に分類する．

アメダスの故障など，理論上は気象データも欠損する可能性があるが，2021 年 9 月〜2022 年 9 月のデータにおいて欠損は確認されなかった．そのため，本研究においては気象データの欠損は考慮しないものとする．

4.3 作成したバス運行データセットの特徴

バス運行データの特徴を考慮した欠損値補完手法を提案するため，本研究では収集したデータを分析した．本節では，分析結果を元にバス運行データの特徴を述べる．

4.3.1 所要時間の乱れやすさ

図 7 に，各便ごとの始点バス停から終点バス停までの平均所要時間とその標準偏差を示す．横軸は始点のバス停を出発する時刻，赤の補助線はこの路線の予定所要時間である 2100 秒（35 分），黒点が平均所要時間，その上下のバーは標準偏差を示している．特に午前中のピーク時間帯である 6〜10 便目はその前の便と比較して平均所要時間が長く標準偏差も大きいため，不安定な運行であることが読み取れる．

表 6: 観測された天気と分類の対応表

項番	天気	分類
1	快晴	晴れ
2	晴	晴れ
3	薄曇	晴れ
4	曇	曇り
5	煙霧	曇り
6	砂じん嵐	曇り
7	高い地ふぶき	曇り
8	霧	雨
9	霧雨	雨
10	しゅう雨または止み間のある雨	雨
11	降水	雨
12	雨	雨
13	みぞれ	雨
14	雪	雨
15	着氷性の雨	雨
16	着氷性の霧雨	雨
17	凍雨	雨
18	霧雪	雨
19	しゅう雪または止み間のある雪	雨
20	あられ	雨
21	ひょう	雨
22	もや	雨
23	細氷	雨
24	雷	雨

4.3.2 バス運行データセットの周期性

使用する 21 系統上り路線の走行時間，停車時間の自己相関について述べる．自己相関とは，元の時系列データが，それ自身を時間方向にずらしたデータとどの程度一致するかを表した尺度である．図 8〜12 に各走行区間ごとの走行時間の自己相関を，図 13〜18 に各バス停の停車時間の自己相関を示す．横軸のラグは，元のデータを何運行分ずらしたかを示す．なお，この分析に際しては欠損値を LOCF で補完している．

図 8〜18 から，ほとんどの走行区間の走行時間やバス停の停車時間に 26 運行ごとのピークが存在し，26 運行ごとの周期性があることを示している．これは，21 系統上り路線が 1 日に 26 運行されているためであると考えられる．ただし，図 8 と図 10 に示すとおり，走行区間 1 と走行区間 3 の走行時間には 26 運行ごとの自己相関が見られなかった．また，図 9 が示すとおり，走行区間 2 の走行時間は他の走行区間と比較して自己相関がかなり小さい．これらは，表 5 に示すとおり，走行区間 1〜3 の予定所要時間が 1〜2 分とかなり短く，運行ごとの差が発現しにくいためと考えられる．

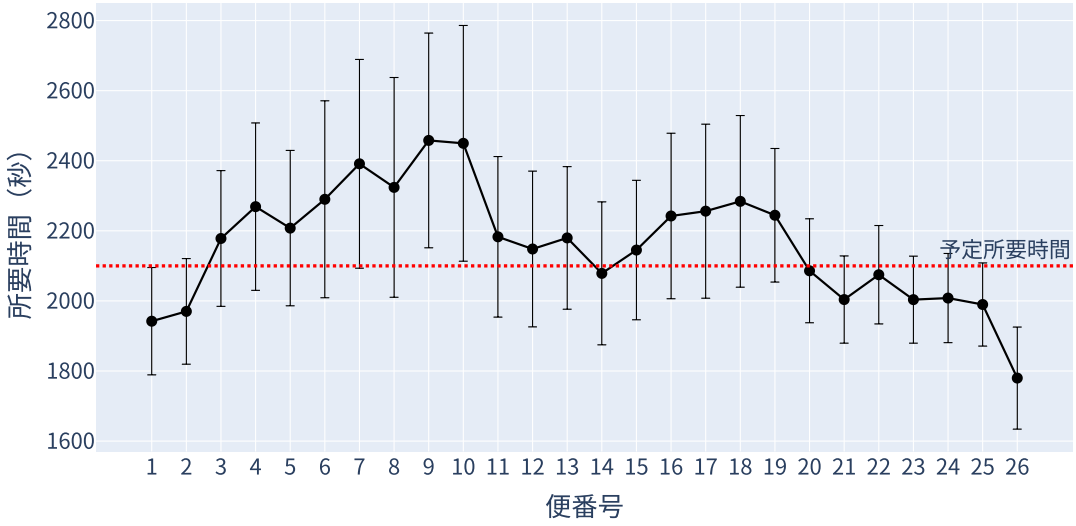


図 7: 21 系統上り路線の各便ごとの始点から終点までの平均所要時間と標準偏差

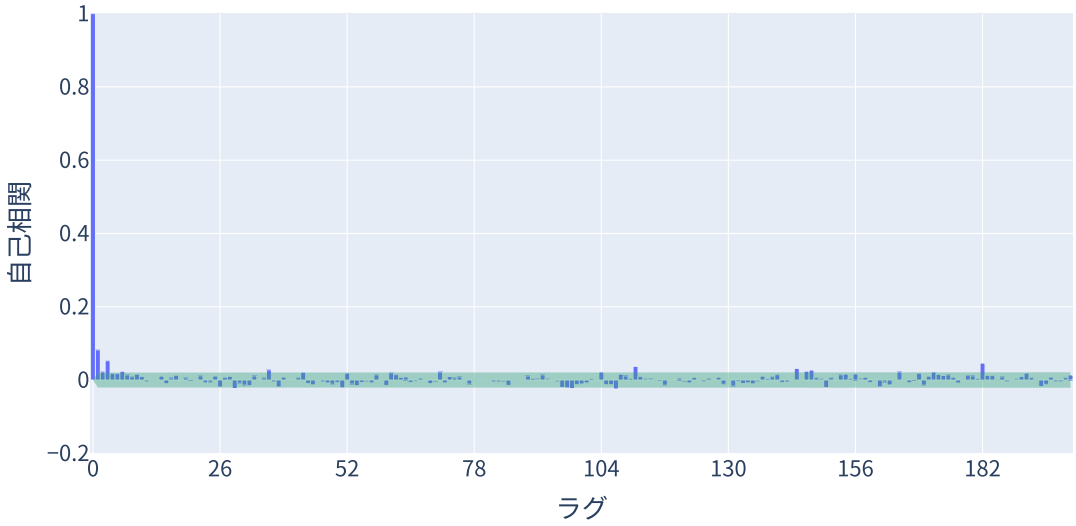


図 8: 走行区間 1 の走行時間の自己相関

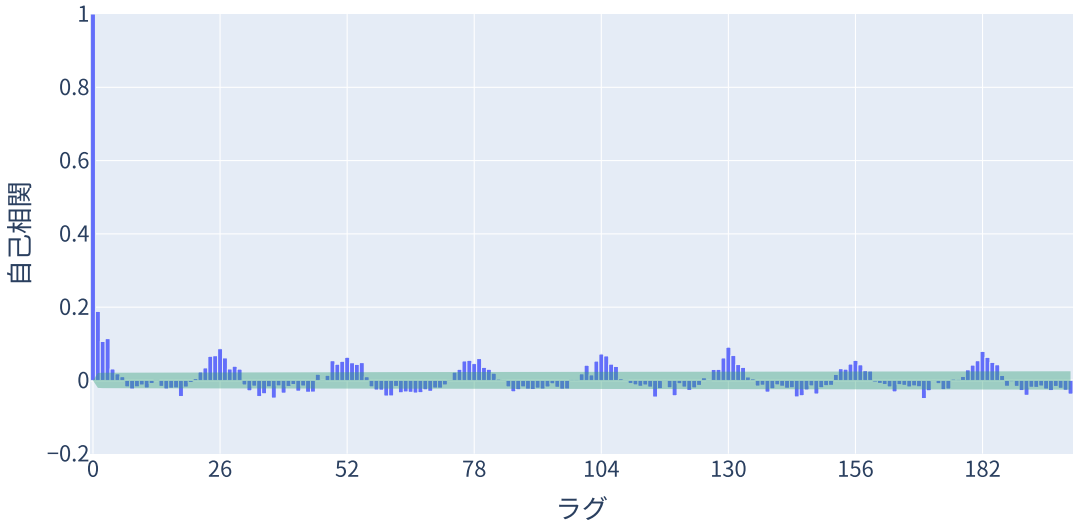


図 9: 走行区間 2 の走行時間の自己相関

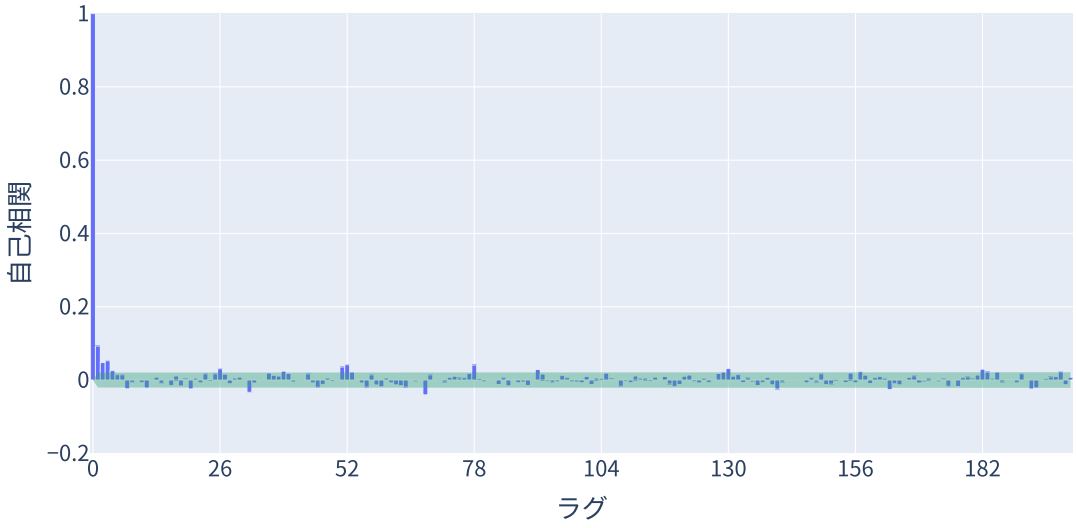


図 10: 走行区間 3 の走行時間の自己相関

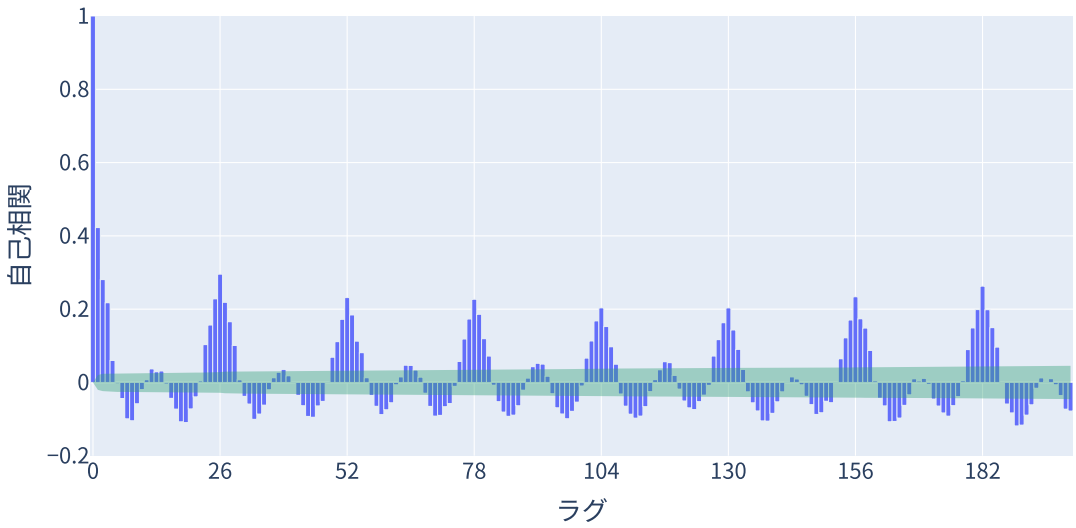


図 11: 走行区間 4 の走行時間の自己相関

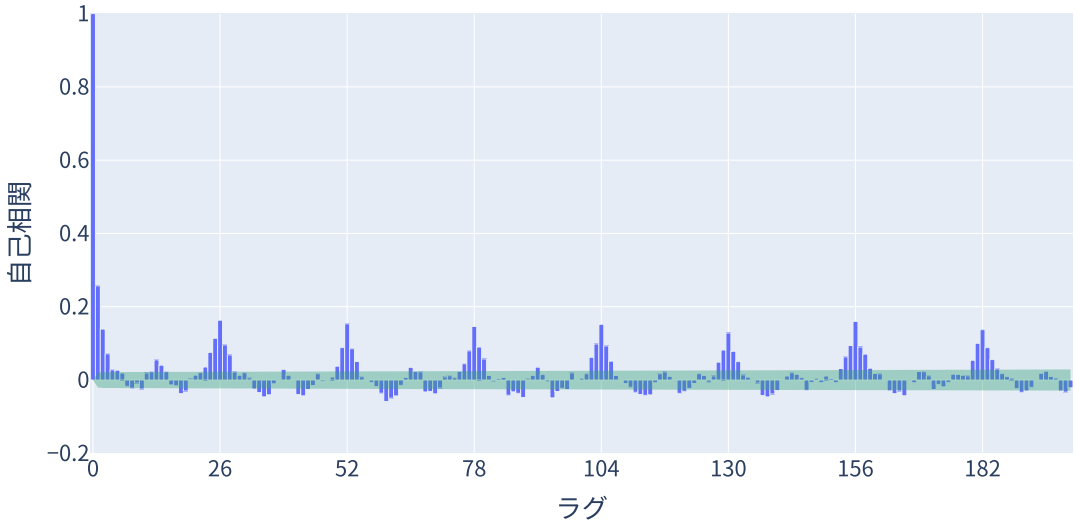


図 12: 走行区間 5 の走行時間の自己相関

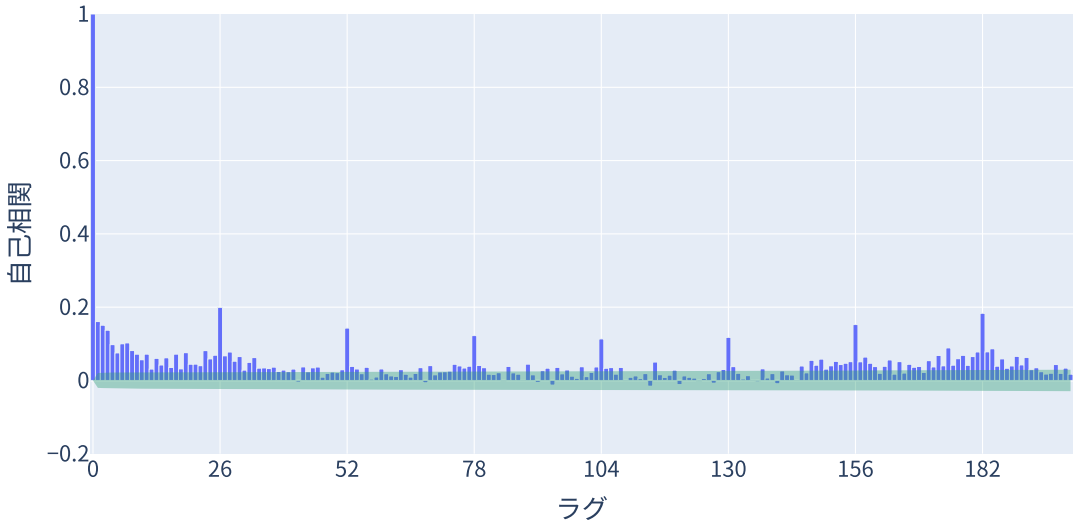


図 13: バス停 1 の停車時間の自己相関

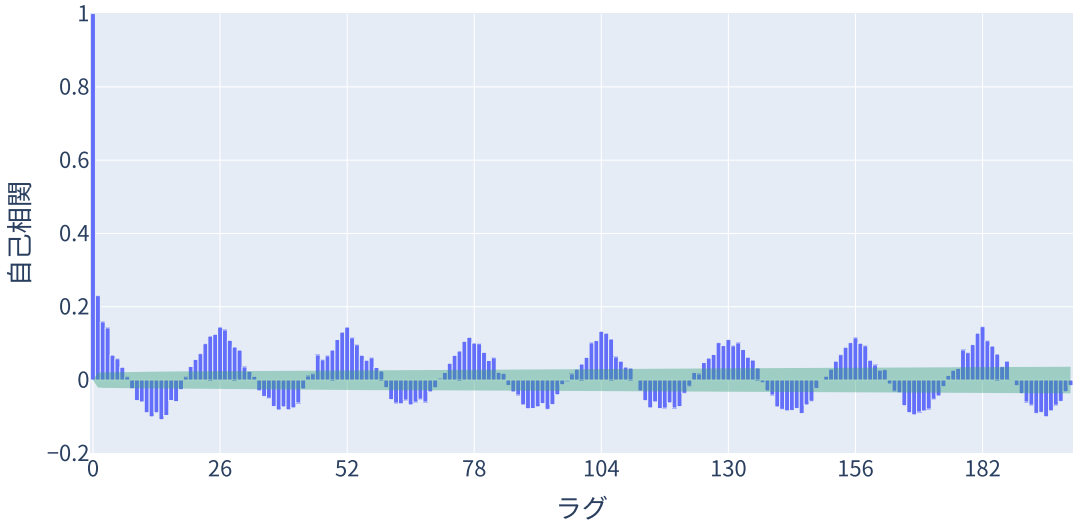


図 14: バス停 2 の停車時間の自己相関

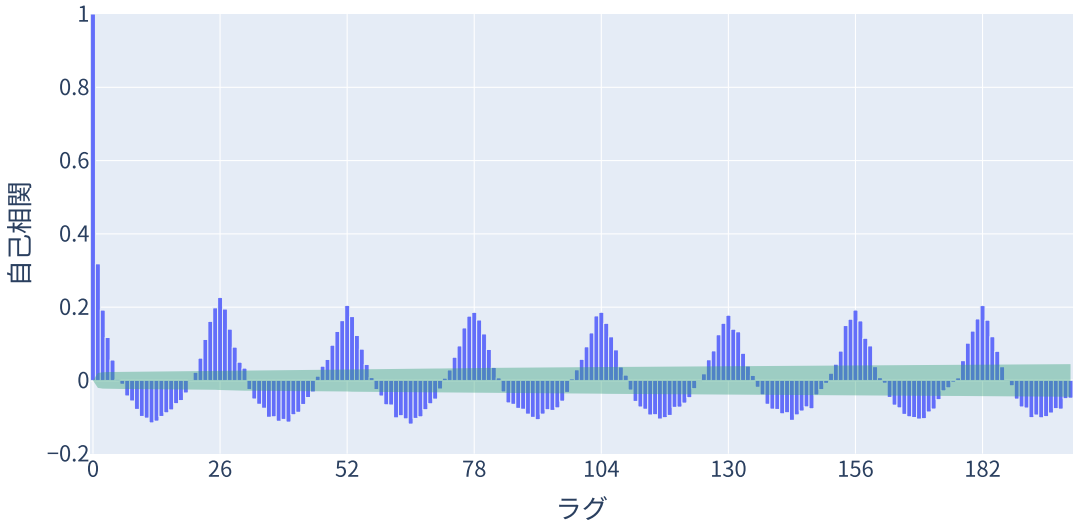


図 15: バス停 3 の停車時間の自己相関

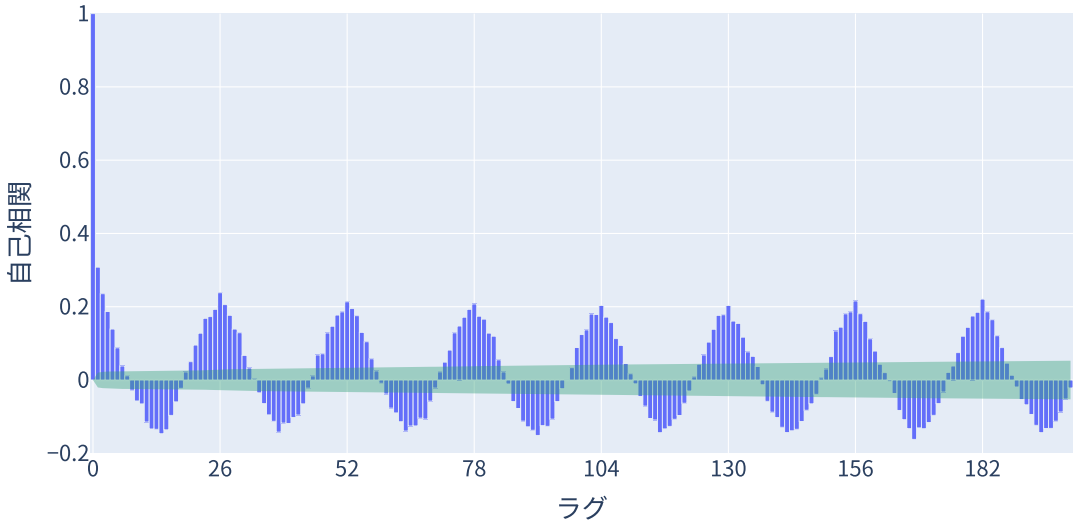


図 16: バス停 4 の停車時間の自己相関

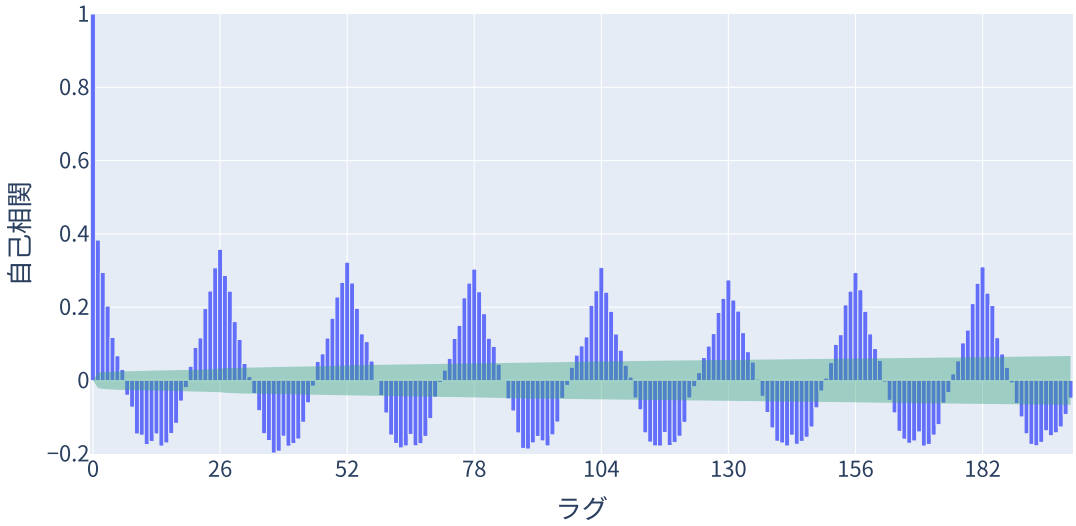


図 17: バス停 5 の停車時間の自己相関

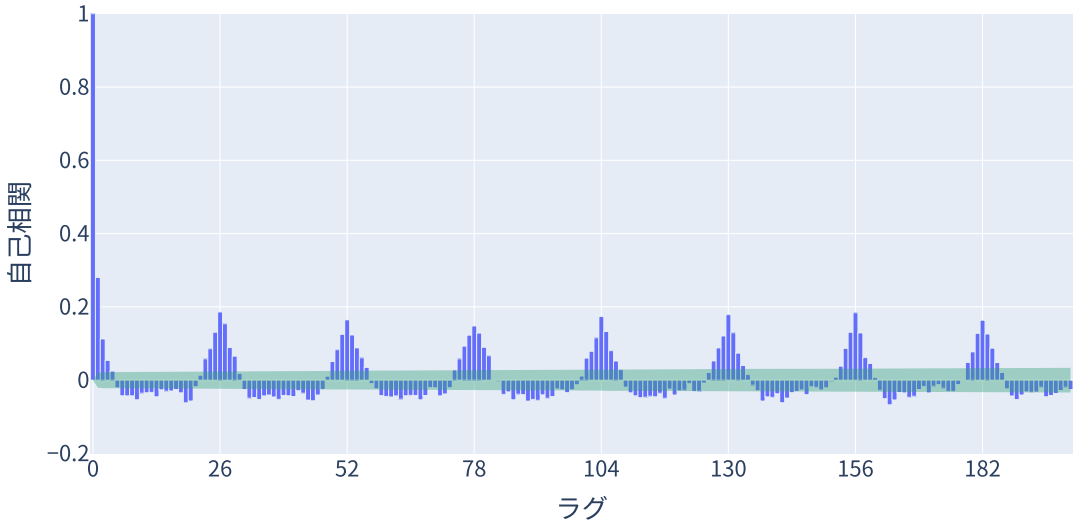


図 18: バス停 6 の停車時間の自己相関

4.3.3 バス運行データセットの欠損

3.4 節で述べたとおり，石長らの手法では特徴量として 1 運行の全区間の走行時間，停車時間，時刻表差分を使用する．そのため，1 運行の各区間ごとのデータのうち 1 つでも欠けているものを欠損として扱う．

時系列データの学習においては，欠損せずに連続でデータが取得できるかが重要になる．例えば，3.6 節で述べたとおり予測モデルの入力には N_{in} 運行分連続したデータが必要である．また，時系列データから学習をする際には，長期的な傾向を掴むために，長期間欠損せずに連続したデータが必要となる．そこで，どれだけ欠損せずに連続したデータを取得できるかに着目して分析した．欠損せずに連続したデータのことを連続データと呼ぶ．

今回作成したバス運行データセットは計 9863 運行が対象であり，そのうち 743 運行に欠損が生じた（欠損率 7.53%）．欠損の原因としては，バス車載器の起動に失敗するなどのトラブル，UDP プロトコルによる通信時のパケットロス，GPS センサの精度不足による走行時間・停車時間の計算失敗などがみられた．

図 19 に，作成したバス運行データセットに含まれる連続データが何運行分の長さであるかをヒストグラムに示す．連続データの長さが 1 運行分であるものは

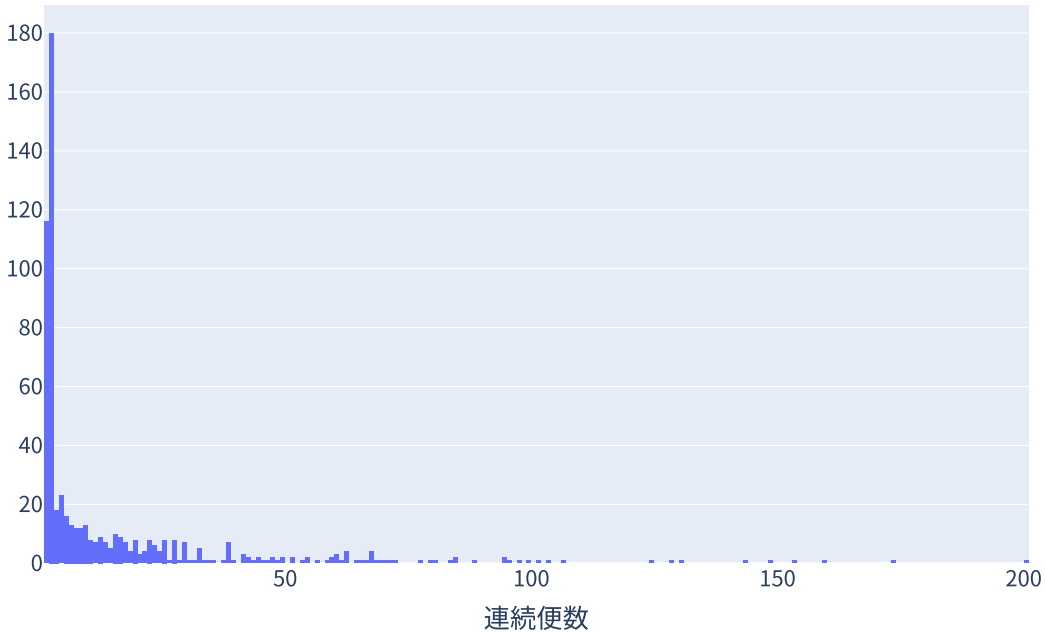


図 19: バス運行データセットの連続データの運行数のヒストグラム

116 サンプル，2 運行分であるものは 180 サンプルあり，ほとんどの連続データが数運行分の長さしか無いことが分かる．最長の連続データは 200 運行分のものが 1 サンプルだけ存在する．21 系統上り路線は 1 日 26 運行であるため，8 日分以上連続したデータがデータセット内に存在しないことが分かる．

また，欠損は全体的に散らばっており連続で欠損している場所はほとんど見られなかった．2.4 節と同様に 8 運行分の連続データを取得できる割合を実データから求めると，約 67.2%であった．すなわち，収集したデータにおいては 30%以上の運行で到着時刻を予測できないということになる．このことからも，バス運行データセットの欠損値補完が必要であると分かる．

5. バス運行データの特徴を考慮した欠損値補完手法

本研究ではバス運行データの特徴に着目した欠損値補完手法によってバス到着時刻予測の誤差削減を目指す．そうした手法として，時間的補完，パターン補完，およびそれらの組み合わせ手法の 3 つの欠損値補完手法を提案する．本章ではこれらの提案手法についてそれぞれ説明する．

5.1 提案手法の概要

バス運行データの特徴に着目した欠損値補完手法を提案するにあたって，同じ高度交通システムの研究分野である Shin らの手法 [4] と同様のアプローチが効果的であると考えた．ただし，Shin らの手法では空間的補完・時間的補完・パターン補完の 3 つを組み合わせていたが，そのままバス運行データの欠損値補完に適用することは難しいため，いくつかの変更を加えている．

1 つ目の変更点は，空間的補完を行わないことである．バス運行データの性質上欠損した運行データと同じ時刻に，欠損したデータのバスの近くを走行するバスの運行データ取得が困難である．そのため，本研究では空間的補完を行わず，時間的補完とパターン補完の 2 つを行う．

2 つ目の変更点は，パターンデータの取得方法である．3.2 節で述べたように路線バスは時刻表ベースで運行するため，運行データには日ごとの周期性が見られる．本研究では，この日ごとの周期性に着目してパターン補完を行う．

このような変更を加え，本研究では時間的補完とパターン補完，およびそれらの組み合わせ手法の 3 つの欠損値補完手法を提案する．

なお本提案手法は，時刻表差分に対してそのまま適用すると，場合によっては走行時間・停車時間との間に矛盾が生じる．そのため，時刻表差分に関しては始点バス停のみ本手法を適用し，残りのバス停の時刻表差分は本手法で求めた走行時間・停車時間を元に計算する．

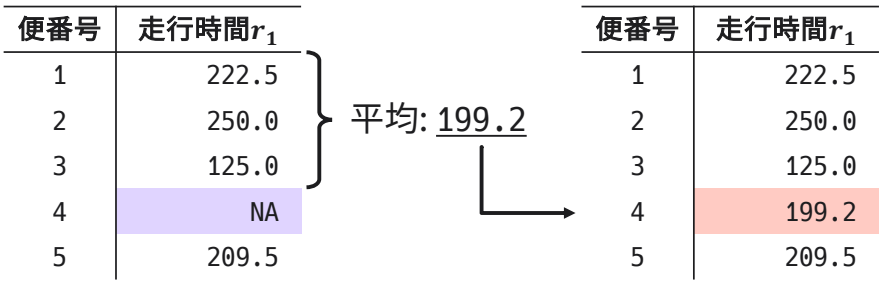


図 20: 時間的補完の例（走行区間 1 の走行時間，$N_{\text{mean}} = 3$ の場合）

5.2 時間的補完

図 20 に時間的補完の流れを示す．時間的補完においては，欠損部分の直前 N_{mean} 運行分の平均値を使用して補完する．時間的補完のみを使用する場合，直前 N_{mean} 運行の間に更に欠損が含まれると平均値を計算できない．その場合は，先頭から順に処理を行い補完された平均値を用いてさらに平均値を求めることで算出する．この手法では，雨天や渋滞などによる遅延など，直近のイベントを反映することが期待される．また，LOCF では外れ値の次の運行が欠損した場合に，外れ値を連続で適用する問題があったが，時間的補完の場合は過去の複数運行の平均を使用するため，外れ値の影響を受けにくいことも期待される．一方，欠損が連続した場合に朝夕の混雑時間帯のような短時間の変化を見落とす問題が生じると考えられる．また，データセットの先頭に欠損がある場合はこの手法では欠損値を補完できない．

5.3 パターン補完

パターン補完は，あらかじめバス運行データセットの欠損していない部分からパターンデータを生成しておき，欠損部分にそれを当てはめることで補完する手法である．図 21 にパターンデータ生成の流れを示す．パターンデータは，バス運行データセットの欠損していない部分のうち走行区間（バス停）ごと，便番号ごとの平均値を算出して作成する．その後，欠損した運行と便番号が一致するパターンデータを当てはめることで補完する．この手法では，直前の遅延などの情

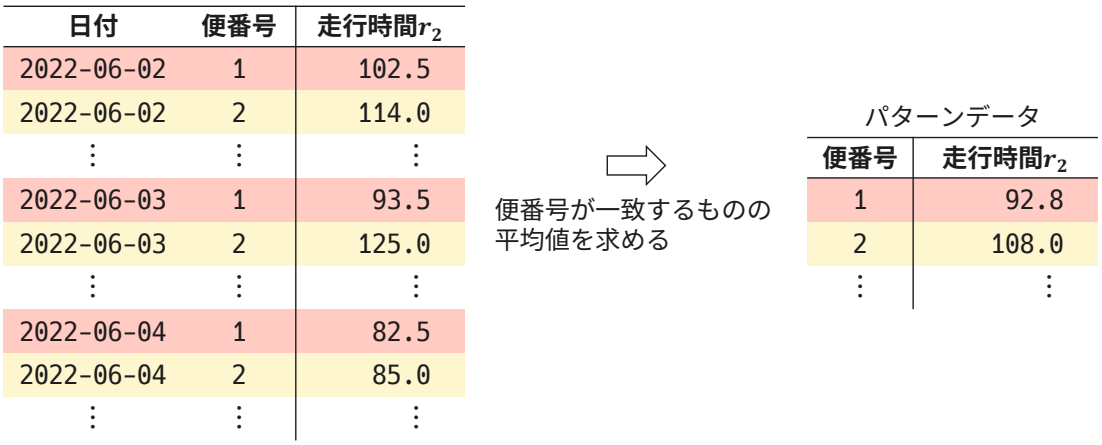


図 21: パターンデータ生成の例（走行区間 2 の走行時間の場合）

報を取り込めないが，朝夕のピーク時間帯などの周期性を取り込めると考えられる．また，LOCF・線形補間・時間的補完のいずれでも補完できなかったデータセットの先頭部分の欠損に関しても，便番号は分かるためパターン補完を適用できる．

5.4 時間的補完・パターン補完の組み合わせ手法

図 22 に時間的補完・パターン補完の組み合わせ手法の流れを示す．この手法においては，予めパターンデータを求めておいた後，時間的補完を行う．この際，欠損部分の直前 N_{mean} 運行の中にさらに欠損が存在する場合は時間的補完せず，パターン補完のみを行う．この手法においては，データがある程度揃っている場所であれば直前の遅延などの情報を取り込みつつ，データがまとめて欠損している部分には周期的な変化を取り入れることができるため，より実際の運行状況に近いデータ補完が期待できる．

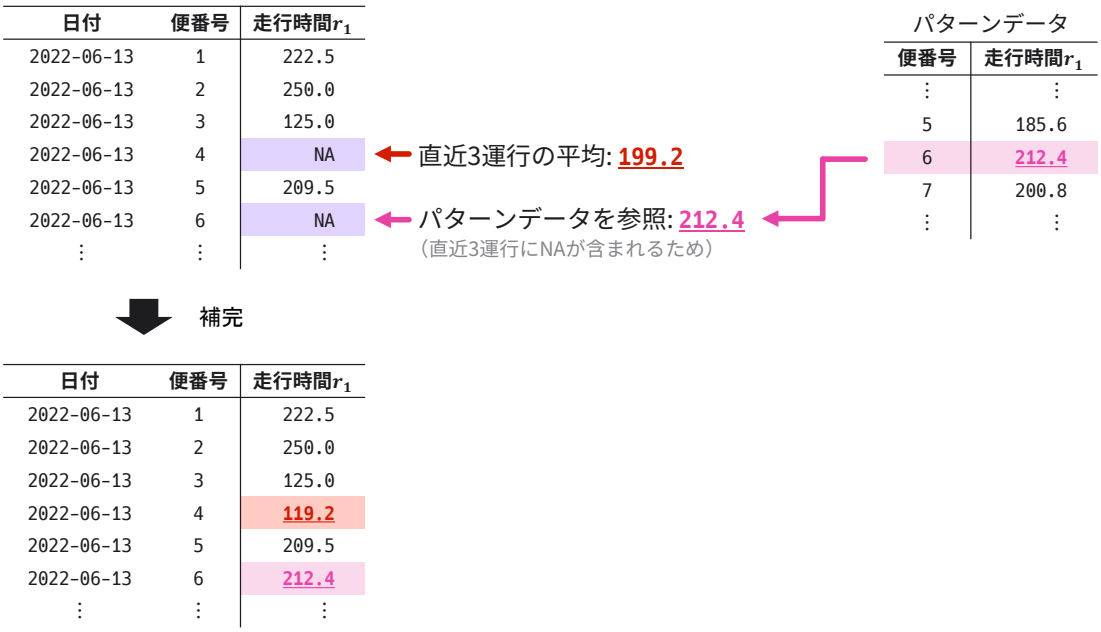


図 22: 組み合わせ手法の例（走行区間 1 の走行時間，$N_{\mathrm{mean}} = 3$ の場合）

6. 評価

各欠損値補完手法がバス到着時刻予測に適しているかを評価するため，本研究では 2 つの評価実験を行った．1 つ目は，各欠損値補完手法に対して訓練データの欠損率を変えながらモデルを学習させ，予測誤差の変化を確認する実験である．2 つ目は，テストデータの欠損率を変化させながら予測をした場合の予測誤差の変化を確認する実験である．

本章では，まず評価方法と比較対象，評価実験の方法について述べる．その後，実験でしたデータセットの分割方法とパラメータ調整について説明し，最後に評価結果を述べる．

6.1 評価方法

本評価では，既存の単純な欠損値補完手法と提案する欠損値補完手法それぞれをバス運行データセットへ適用した際にバス到着時刻予測の誤差がどのように変化するかを観察することで，各欠損値補完手法がバス到着時刻予測に適しているかを示す．そのために，欠損率を変化させたバス運行データセットに各欠損値補完手法を適用したデータセットからバス到着時刻を予測する実験を実施した．この実験の結果得られる各欠損値補完手法を用いた場合のバス到着時刻予測の誤差を比較することで，どの手法が最も小さい予測誤差を得られるかを観察する．実験の簡略化のために，実験では始発バス停から終点バス停までの所要時間を，3 運行先まで予測することとした．

実験は，訓練データの欠損率を変化させた場合と，テストデータの欠損率を変化させた場合の 2 通り実施した．前者の実験では，欠損率を変化させた訓練データに各欠損値補完手法を適用したもので予測モデルを学習し，欠損率を変更していない元のテストデータに対してバス到着時刻を予測した．後者の実験では，欠損率を変更していない元の訓練データで予測モデルを学習し，欠損率を変化させたテストデータに各欠損値補完手法を適用したものに対してバス到着時刻を予測した．なお，欠損率を変更していない元のデータに対する欠損値補完では，欠損率を変化させたデータで用いた手法と同じ欠損値補完手法を用いる．実験の詳細

は 6.4 節で述べる．

本研究では，バス到着時刻予測の誤差指標として平均絶対誤差（Mean Absolute Error, MAE）を使用した．MAE の計算方法を式 (4) に示す．

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{l}_{B,i} - l_{B,i}\right| \tag{4}$$

ここで，$\hat{l}_{B,i}$ は終点バス停までの所要時間の予測値，$l_{B,i}$ は終点バス停までの所要時間の正解値，n はテストデータのデータ数である．MAE は平均化された誤差に基づく評価指標であり，単位は秒である．バス到着時刻予測においては，誤差の平均が秒数で分かるため特に同一路線における評価の際に使用される．本評価においては，同一路線の同一区間に対する予測誤差を比較するため，MAE の結果に着目して評価した．MAE は石長らの手法 [3] でも使用されている．

6.2 データセットの分割

評価では，予測モデルの学習および予測モデルを使用した実験を行うため，データセットを訓練データ，検証データ，テストデータの 3 つに分割する必要がある．訓練データはモデルを学習するのに必要なデータである．検証データは，後述する予測モデルのハイパーパラメータを決定する際に使用するデータである．テストデータは，学習した予測モデルに入力として渡し，予測モデルの性能を評価するのに使用するデータである．

4.1 節で説明したとおり，本研究では 2021 年 9 月 1 日から 2022 年 9 月 25 日までのバス運行データセットと気象データセットを用意した．そのうち，2021 年 9

表 7: バス運行データセットの分割結果

種類	運行数	欠損運行数	欠損率（%）
訓練データ	9291	720	7.75
検証データ	182	11	6.04
テストデータ	390	12	3.08

月 1 日から 2022 年 9 月 3 日までを訓練データ，2022 年 9 月 4 日から 2022 年 9 月 10 日までを検証データ，2022 年 9 月 11 日から 2022 年 9 月 25 日までをテストデータとして使用した．データのリーケージを防ぐため，データの期間をもとに分割した．それぞれの運行数と欠損運行数，欠損率を表 7 に示す．

6.3 比較対象

本評価では，過去の運行の平均値をそのまま予測値として用いる Historical Average (HA) によるバス到着時刻予測の結果をベースラインとして使用する．また，欠損率が変更されたバス運行データセットに対して以下の 5 種類の欠損値補完を適用し，石長らの手法を用いてバス到着時刻を予測する．

- LOCF
- 線形補間
- 時間的補完
- パターン補完
- 組み合わせ手法

その後，それぞれの手法によるバス到着時刻予測の誤差を比較することで評価する．以降，それぞれのバス到着時刻予測の方法について説明する．

6.3.1 Historical Average (HA) によるバス到着時刻予測

訓練データの便番号ごとの走行時間・停車時間の平均値から所要時間を計算し，そのまま予測値として用いるバス到着時刻予測の手法である．ベースラインとして使用する予測方法で，この到着時刻予測のみ石長らの手法を使用しない．このような手法は交通量が静的であるなど，所要時間が安定している場合にのみ小さい誤差で予測可能となる．今回の路線のように所要時間が安定しない路線では使用が困難である．

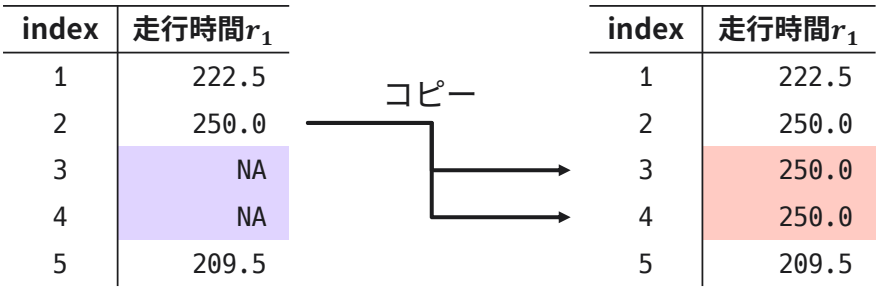


図 23: LOCF の例（走行区間 1 の走行時間の場合）

6.3.2 LOCF を適用したバス到着時刻予測

バス運行データセットに対し LOCF を適用し，石長らの手法でバス到着時刻を予測する手法である．走行時間の欠損に対して LOCF を適用する例を図 23 に示す．欠損した運行に対し，欠損部分より前の最後に観測された値を適用する．

単純な手法であるため多くの問題もある．1 つ目の問題は，特に連続で欠損した場合に時間的な傾向を無視してしまい，補完後の数値が不自然に一定値を示す点である．2 つ目の問題は，外れ値の次の値が欠損していた場合に外れ値を繰り返し適用してしまう点である．3 つ目の問題は，データセットの先頭部分が欠損している場合に適用できない点である．

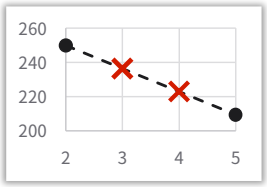
欠損部分の前の値が繰り返し適用されることで，データセットの日ごとの周期性が乱れると考えられるため，バス到着時刻予測の結果においても日ごとの周期性を無視した予測が行われ，予測誤差が増えると予想する．

6.3.3 線形補間を適用したバス到着時刻予測

バス運行データセットに対し線形補間を適用し，石長らの手法でバス到着時刻を予測する手法である．走行時間の欠損に対して線形補間を適用する例を図 24 に示す．走行時間の場合は，index を x，走行時間を y とし，欠損部分の前後で欠損していない 2 点をそれぞれ $(x_1, y_1), (x_2, y_2)$ とする．このとき，走行時間の関数 $f(x)$ を 2 点を通る式 (5) で近似して欠損値を求める．

$$f(x) = y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) \tag{5}$$

index	走行時間r_1
1	222.5
2	250.0
3	NA
4	NA
5	209.5



index	走行時間r_1
1	222.5
2	250.0
3	235.5
4	223.0
5	209.5

図 24: 線形補間の例（走行区間 1 の走行時間の場合）

線形補間は，欠損したデータより未来のデータを使用して補完を行う．そのため，訓練データの欠損値補完には使用できるが，予測時の入力の欠損には適用できない問題がある．また，LOCF と同じように連続で欠損した場合には時間的な傾向を無視してしまい，補完後の数値が不自然な傾きを示す問題がある．

LOCF と同様に，データセットの日ごとの周期性が乱れると考えられるため，バス到着時刻予測においても日ごとの周期性を無視した予測となり，予測誤差が増えると予想する．一方，欠損率が高くなった場合に，LOCF のように繰り返しの値は生じないため，LOCF より予測誤差が小さくなると予想する．

6.3.4 時間的補完を適用したバス到着時刻予測

バス運行データセットに対し 5.2 節で述べた時間的補完を適用し，石長らの手法でバス到着時刻を予測する手法である．

日ごとの周期性を無視した補完を行うため，バス到着時刻予測の結果でも日ごとの周期性を無視した予測が行われ，全体の予測誤差は増加すると予想する．一方，直近数運行分の運行の乱れを反映した補完を行うため，突発的な運行の乱れに対して予測が追従しやすいと予想する．

6.3.5 パターン補完を適用したバス到着時刻予測

バス運行データセットに対し 5.3 節で述べたパターン補完を適用し，石長らの手法でバス到着時刻を予測する手法である．パターンデータは訓練データの欠損

していない部分を用いて作成する．これは，テストデータを用いてパターンデータを作成すると，予測時に未来の情報を入力してしまうためである．

日ごとの周期性に沿った欠損値補完を行うため，バス到着時刻予測の結果でも日ごとの周期性に沿った予測が行われると予想し，予測誤差が小さくなると予想する．一方，突発的な運行の乱れに対しては予測が追従しづらいと予想する．

6.3.6 組み合わせ手法を適用したバス到着時刻予測

バス運行データセットに対し 5.4 節で述べた組み合わせ手法を適用し，石長らの手法でバス到着時刻を予測する手法である．

時間的補完，パターン補完のお互いの弱点を補うことで日毎の周期性に沿った予測をしつつ，突発的な運行の乱れに対しても追従できることを予想する．

6.4 評価実験方法

本評価では，訓練データの欠損率を変化させる実験と，テストデータの欠損率を変化させる実験を行った．本節では，それぞれの実験の想定環境と実験方法について説明する．

6.4.1 訓練データの欠損率を変化させる実験

学習時に使用する訓練データで欠損が発生することを想定し，欠損率を変化させた訓練データで予測モデルを学習させ，それぞれの予測モデルを使用した場合の予測誤差を欠損値補完手法ごとに比較する実験を行った．訓練データは，欠損率を変更していないものと，欠損率を 10%〜90%まで 10%ごとに変化させたのものを用意した．この際，欠損率は元のデータセットに含まれる欠損を含めて計算している．また，人為的に欠損を発生させる際は，乱数で選んだ運行のデータを欠落させることで欠損としたが，この乱数のシード値による上振れや下振れの影響を避けるため，各欠損率ごとにシード値を変化させながら 10 種類の訓練データを用意した．そのため訓練データは全部で 91 種類となる．これらの訓練データに対しそれぞれ欠損値補完を行い，モデルを学習させ，その予測誤差を比較する．

各欠損値補完手法を適用する際に，訓練データの先頭や末尾が欠損していて補完できない場合は，その部分のデータは使用せずに削除する．そのため，欠損値補完手法によっては訓練データの長さが短くなる．

この実験で使用するテストデータは欠損率を変更していないものであり，テストデータの欠損は，訓練データと同じ手法で欠損値補完した．

6.4.2 テストデータの欠損率を変化させる実験

予測時の入力データに欠損が生じる場合を想定し，訓練済みモデルに欠損率を変化させたテストデータを入力し，欠損値補完手法ごとに予測誤差を比較する実験を行った．テストデータは，欠損率を変更していないものと，欠損率を 10%〜90%まで 10%ごとに変化させたのものを用意した．訓練データの欠損率を変化させる実験と同様に，欠損率は元のデータセットに含まれる欠損を含めて計算し，各欠損率ごとに乱数のシード値を変化させながら 10 種類のテストデータを用意した．そのため，テストデータも 91 種類となる．それぞれのテストデータを各欠損値補完手法を用いて欠損値補完し，欠損率を変更していない状態で学習したモデルを用いて予測した際の誤差を比較する．

本実験は予測時の入力データが欠損していることを想定して実験を行うため，入力データの末尾が欠損している場合に利用できない線形補間は実験の対象外とした．また，LOCF と時間的補完は入力データの先頭が欠損している場合に利用できないが，過去のバス運行データは参照できることを考慮し，テストデータの範囲の前のデータ（2022 年 9 月 10 日の 26 便目）の値を参照して欠損値補完した．

この実験で使用する訓練データは欠損率を変更していないものであり，訓練データの欠損は，訓練データと同じ手法で欠損値補完した．

6.5 パラメータ設定

6.5.1 予測モデル入力運行数 N_{in}，出力運行数 N_{out}

$N_{\mathrm{in}}, N_{\mathrm{out}}$ は，対象となる路線ごとに設定されるべき値である．今回の評価実験では，午前中のピーク（7〜10 便目）の影響が午後のピーク（17〜19 便目）の予

測に利用できるよう，それぞれ $N_{\mathrm{in}} = 8, N_{\mathrm{out}} = 3$ に設定した．この決定方法は石長らの手法と同様である．

$N_{\mathrm{in}} = 8, N_{\mathrm{out}} = 3$ であるため，バス到着時刻予測では過去 8 運行分のデータを入力することで，次以降 3 運行分までの到着時刻予測が可能である．評価実験では，1〜3 運行先それぞれの終点バス停での到着時刻を予測し，それぞれの誤差を評価した．

6.5.2 時間的補完・組み合わせ手法の平均値計算に利用する運行数 $N_{\mathbf{mean}}$

今回の提案手法のうち，時間的補完と組み合わせ手法に使用される N_{mean} の値は，対象の路線バスによって異なる．この N_{mean} の値を決定するため，予備実験を行った．まず，欠損率を変えていない元のデータセットに対し時間的補完，組み合わせ手法の両方で N_{mean} の値を 1〜26 まで変化させながら補完を行った．その後，それぞれの場合で学習したモデルを使用し，1 運行先の終点バス停での到着時刻を予測した．N_{mean} の値の変化が 26 までである理由は，ちょうど 1 日分の平均を取るのに必要な N_{mean} の値が 26 であるためである．予測到着時刻の MAE をグラフにしたものを図 25，26 に示す．時間的補完における MAE の最小値は 149.89 秒，最大値は 157.13 秒であり，組み合わせ手法における MAE の最小値は 144.28 秒，最大値は 155.04 秒であった．ここから，N_{mean} の値による予測誤差の差は大きくないことが分かった．

そのため今回の評価においては，4.3.2 節で述べた自己相関を参考に $N_{\mathrm{mean}} = 5$ と設定した．表 5 で示したとおり，対象路線は走行区間 4 と 5 の予定所要時間が大きく，この部分が終点バス停での到着時刻予測に大きな影響を与えると予想される．図 11，12 に示す走行区間 4 と 5 の自己相関では，相関のピークの前後 5〜7 運行分までが信頼区間より高く相関が出ているため，$N_{\mathrm{mean}} = 5$ と設定した．

6.5.3 予測モデルのハイパーパラメータ

各モデルを学習する際に必要となるハイパーパラメータは訓練データごとに異なる．より予測モデルの性能を良くするため，欠損率を変えた各訓練データごと

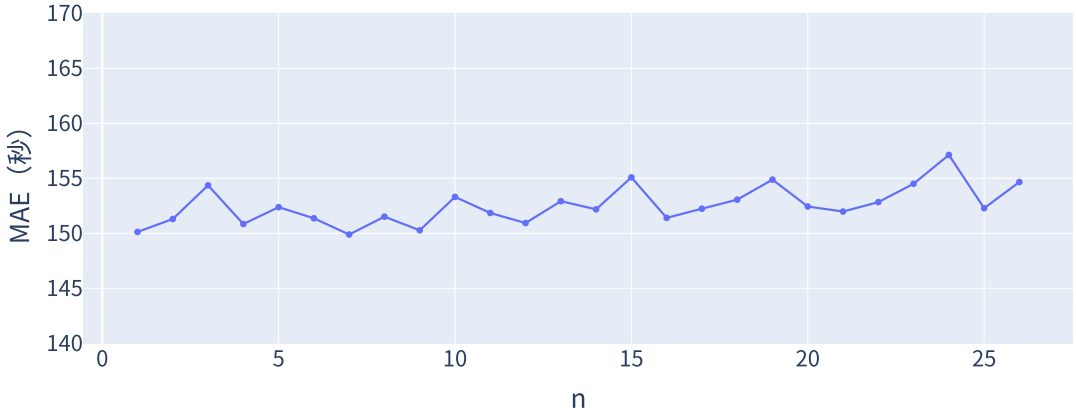


図 25: 時間的補完で N_{mean} を 1〜26 まで変化させた場合の予測到着時刻の MAE（1 運行先予測時）

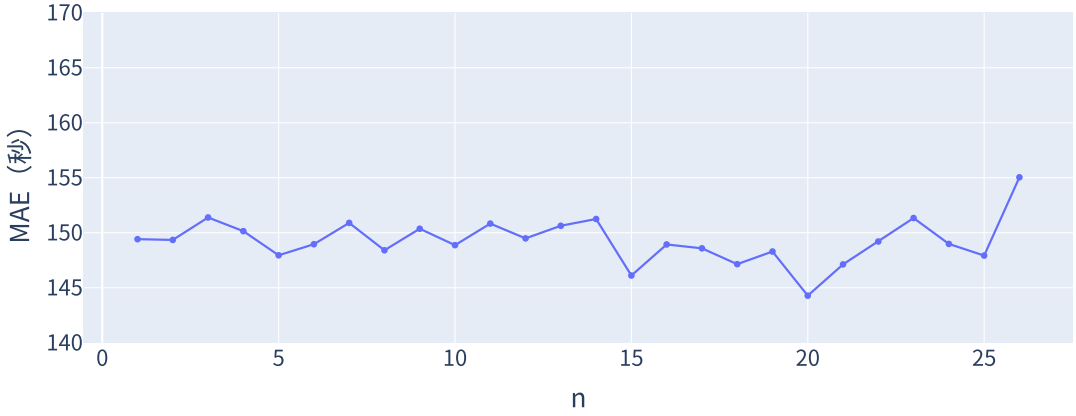


図 26: 組み合わせ手法で N_{mean} を 1〜26 まで変化させた場合の予測到着時刻の MAE（1 運行先予測時）

にハイパーパラメータチューニングを行った．ハイパーパラメータチューニングには HyperBand [19] を用いて，最も有効なパラメータを示す Best Trail のパラメータをそれぞれのモデルのパラメータとして使用した．このパラメータチューニングにおいては，検証データの平均二乗誤差を Validation Loss として探索した．探索したハイパーパラメータの種類と探索範囲は以下のとおりである．

Max epoch

1 つのモデルを学習させるための最大エポック数である．1〜100 の範囲で探索した．

Batch size

データセットをいくつかサブセットへ分割した際に，1 つのサブセットに含まれるデータ数を指す値である．[16, 32, 64, 128, 256] の中から選択した．

Dropout rate

過学習を防ぐための Dropout 層におけるパラメータであり，意図的に無効にするノードを選択する確率の値である．[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] の中から選択した．

Filters

Convolutional LSTM における畳み込みのフィルタ数である．[16, 32, 64, 128, 256] の中から選択した．

Kernel size 0, Kernel size 1

Kernel size 0 はエンコーダ，デコーダそれぞれの 1 層目のフィルタの行数である．同様に Kernel size 1 はエンコーダ，デコーダそれぞれの 2 層目のフィルタの行数である．走行時間の予測モデルでは 1〜5，停車時間の予測モデルでは 1〜4 の中から選択した．

Learning rate

1 回の学習でどの程度重みやバイアスを修正するかの割合である．[0.01, 0.001, 0.0001] の中から選択した．

6.6 評価結果

訓練データの欠損率を変化させた場合とテストデータの欠損率を変化させた場合の 2 つの実験結果を元に，各欠損値補完手法がバス到着時刻予測に適しているかを評価した．本節では，まずそれぞれの実験結果について述べた後，2 つの実験結果を総合してそれぞれの欠損値補完手法を評価した結果を述べる．

6.6.1 訓練データの欠損率を変化させた場合の実験結果

訓練データの欠損率を変化させて学習したモデルを使用し，テストデータに対して 1 運行先終点バス停の到着時刻を予測した際の MAE を図 27 に示す．同様に 2 運行先終点バス停の到着時刻を予測した際の MAE を図 28 に，3 運行先終点バス停の到着時刻を予測した際の MAE を図 29 に示す．各グラフの見方について述べる．横軸は訓練データの欠損率を示す．一番左のプロットは欠損率を変更していないデータセット（欠損率 7.75%）での結果である．それ以外の欠損率でのプロットには上下にバーがついている．これは，各欠損率において乱数のシードの値を変化させた 10 パターンのデータセットを使用したためである．プロット

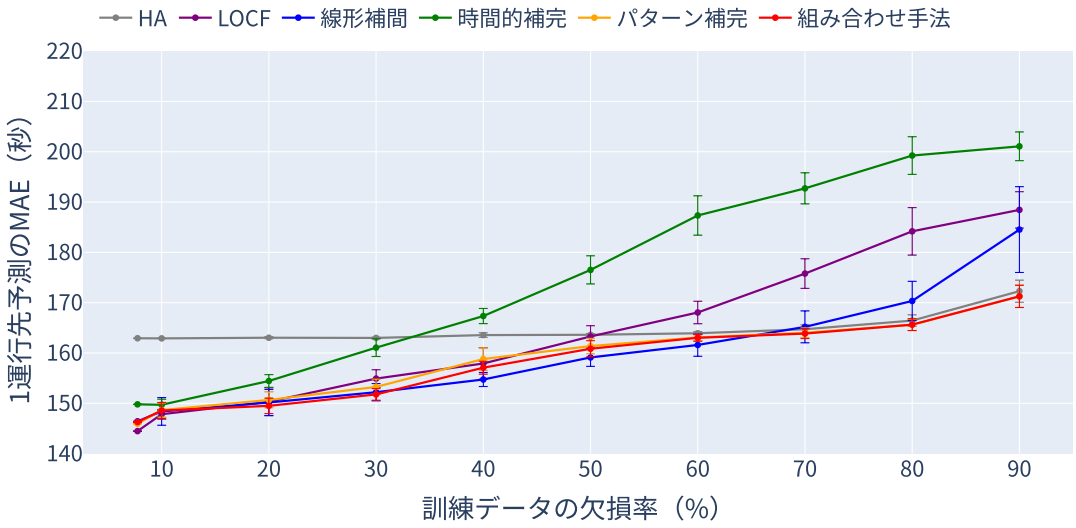


図 27: 訓練データ欠損率変更時の 1 運行先終点バス停の予測到着時刻 MAE

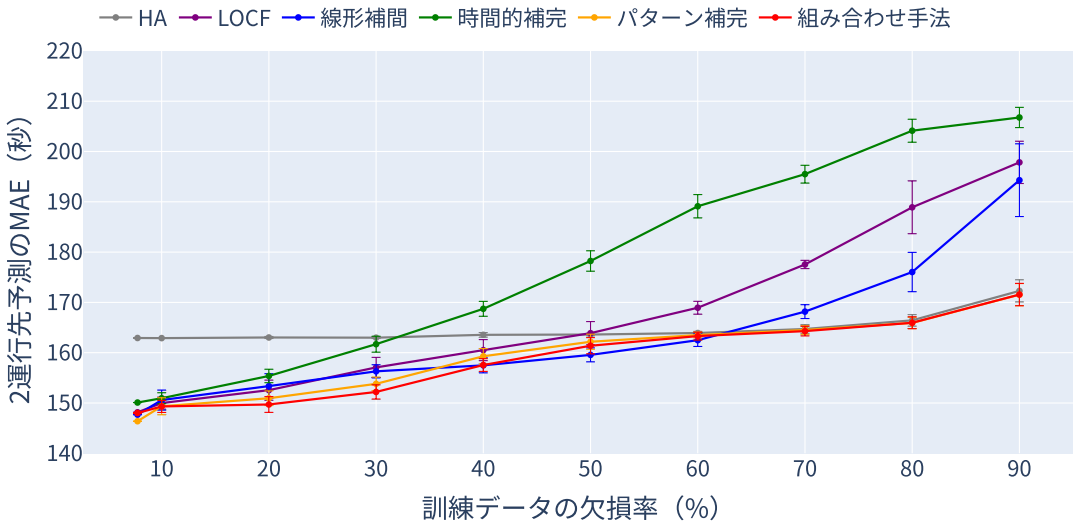


図 28: 訓練データ欠損率変更時の 2 運行先終点バス停の予測到着時刻 MAE

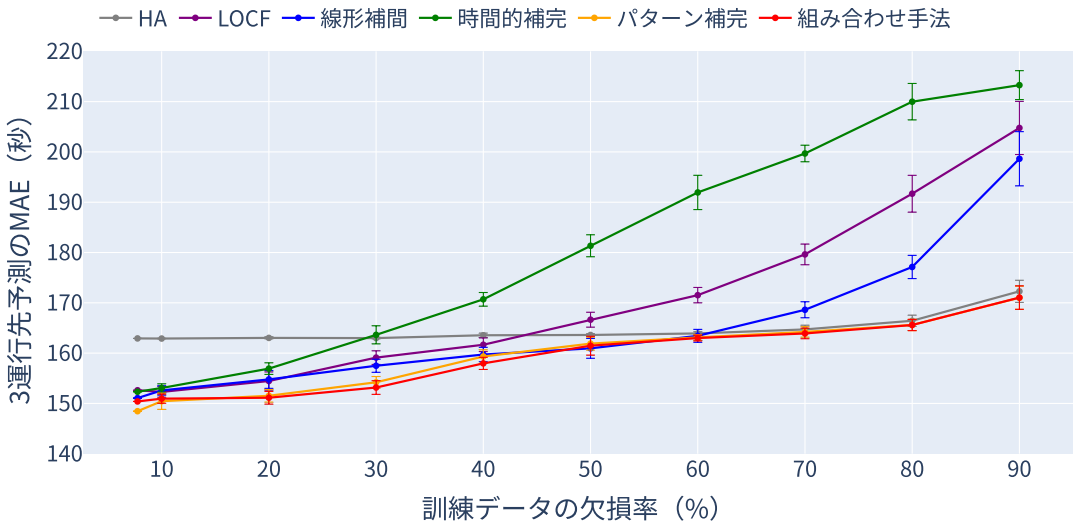


図 29: 訓練データ欠損率変更時の 3 運行先終点バス停の予測到着時刻 MAE

がその欠損率での MAE の平均値，上下のバーがその欠損率での MAE の標準偏差を示す．

欠損率を変更していないデータセットでの結果に着目すると，1〜3 運行先の全ての予測において HA を適用した場合に最も MAE が大きくなっている．このことから，今回の評価対象路線は過去の運行データを使用した単純な予測手法では予測が困難であると考えられ，本研究における評価対象路線としての妥当性が認められる．

ここから，予測した運行数別に結果の詳細について述べる．まずは 1 運行先予測の MAE（図 27）について述べる．全体を通じて，訓練データの欠損率が高くなると MAE は大きくなる傾向が見られる．このとき，時間的補完，LOCF，線形補間は HA よりも MAE が大きくなるが，パターン補完，組み合わせ手法は HA より MAE が大きくならなかった．欠損率が 30%以下のとき，時間的補完を適用した場合を除いてほとんど MAE に差が見られなかった．欠損率が 40%〜60%のとき，線形補間を適用した場合に MAE が小さいという結果になった．欠損率が 70%以上のとき，パターン補完と組み合わせ手法の MAE が小さいという結果になった．

次に 2 運行先予測の MAE（図 28）について述べる．HA 以外の手法においては，全体的に 1 運行先の予測より MAE が増加している．1 運行先予測の場合と全体的な傾向は似ているが，訓練データの欠損率が 20〜30%のとき，LOCF や線形補間と比べて，パターン補完や組み合わせ手法の MAE が小さかった．訓練データの欠損率が高い場合の結果は 1 運行先予測の場合と同様で，パターン補完と組み合わせ手法の MAE が小さいという結果になった．

最後に 3 運行先予測の MAE（図 29）について述べる．2 運行先予測の場合よりさらに全体的な MAE が大きくなっていることから，先の運行の予測は難しいことが分かる．おおむね 2 運行先予測の場合と同じような結果であるが，LOCF や線形補間を適用した場合の MAE がより大きくなっているため，相対的にパターン補完や組み合わせ手法の MAE が小さいという結果となった．

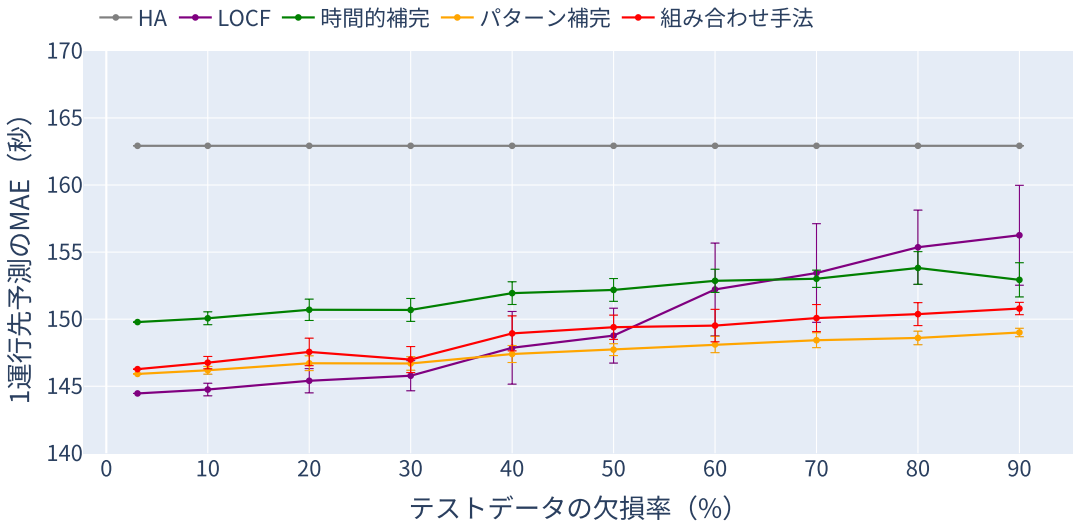


図 30: テストデータ欠損率変更時の 1 運行先終点バス停の予測到着時刻 MAE

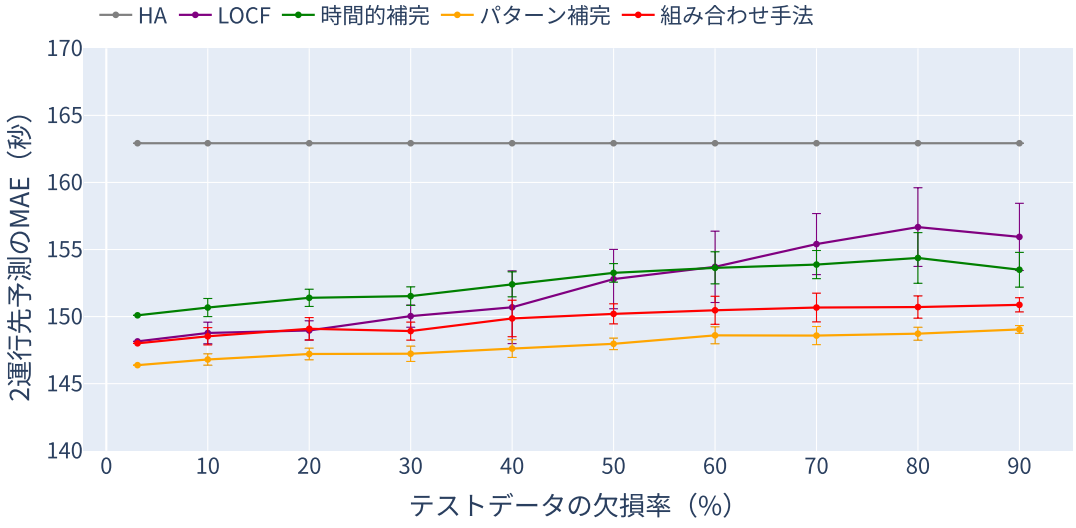


図 31: テストデータ欠損率変更時の 2 運行先終点バス停の予測到着時刻 MAE

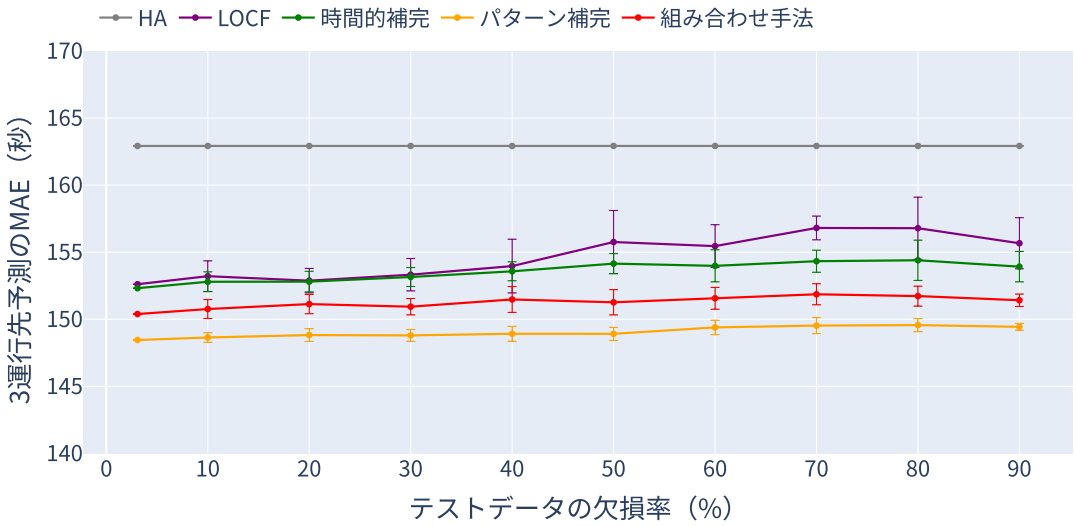


図 32: テストデータ欠損率変更時の 3 運行先終点バス停の予測到着時刻 MAE

6.6.2 テストデータの欠損率を変化させた場合の実験結果

欠損率を変更していないデータセットで学習した場合のモデルを利用し，テストデータの欠損率を変化させた場合の，1 運行先終点バス停での到着時刻を予測した際の MAE を図 30 に示す．同様に 2 運行先終点バス停での到着時刻を予測した際の MAE を図 31 に，3 運行先終点バス停での到着時刻を予測した際の MAE を図 32 に示す．各グラフの見方について述べる．横軸はテストデータセットの欠損率を示す．一番左のプロットは人為的に欠損を挿入していない元のテストデータ（欠損率 3.08%）での結果である．図 27〜29 と同様に，プロットがその欠損率での MAE の平均値，エラーバーがその欠損率での MAE の標準偏差を示す．

HA は入力データによらず同じ予測を出力するため，常に同じ MAE となる．訓練データの欠損率を変化させた場合の実験では HA を上回る MAE が示されたが，テストデータの欠損率を変化させた場合の実験では，常に HA より低い MAE を示した．

ここから，予測した運行数別に結果の詳細について述べる．まずは 1 運行先予測の MAE（図 30）について述べる．テストデータの欠損率が 30%以下の場合は LOCF が最も MAE が小さく，次いでパターン補完，組み合わせ手法が同程度の

MAE，時間的補完が最も大きいMAEであることが分かる．テストデータの欠損率が上昇するとLOCFのMAEは大きく増加し，時間的補完よりもMAEが大きくなる．

次に2運行先予測のMAE（図31）について述べる．訓練データの場合と同様に，全体的に1運行先の予測よりMAEが大きくなっていることが分かる．特に，LOCFを適用した場合のMAEは大きく増加しており，テストデータの欠損率が低い場合でもパターン補完，組み合わせ手法よりも大きいMAEとなっている．テストデータの欠損率に関わらず常にパターン補完が最も小さいMAEであることも分かる．

最後に3運行先予測のMAE（図32）について述べる．こちらも訓練データの場合と同様で，2運行先予測よりもさらにMAEが大きくなっている．LOCFはさらにMAEが増加しており，テストデータの欠損率によらず常に時間的補完よりも大きいMAEとなった．全体を通じパターン補完のMAEが小さく，欠損率に応じてMAEはあまり変化しないことが分かる．

6.6.3 各欠損値補完手法がバス到着時刻予測に適しているかの評価

2つの実験結果を元に，各欠損値補完手法がバス到着時刻予測に適しているか評価した結果を述べる．現実のバス運行データ欠損率は数%〜30%程度であると考えられるため，特にその部分に着目して評価する．

LOCFは，テストデータの欠損率が30%以下の場合に1運行先を予測した場合にMAEが小さくなっている．そのため，直近の運行を予測する場合の欠損理補完手法として適していると言える．

線形補間は，訓練データの欠損率が数%〜30%程度の場合のMAEはLOCFとほぼ同程度であった．また，訓練データの欠損率が40%〜60%の場合にMAEが小さくなっている．そのため，訓練データの欠損率が中程度の場合に適した欠損理補完手法であると考えられる．ただし6.3.3節で述べたとおり，線形補間は訓練データの場合のみ適用可能であり，予測時の入力には適用できないことに注意が必要である．

時間的補完は，訓練データとテストデータの欠損率や，予測先運行数の違いに

関わらず常にMAEが大きかった．このことから，時間的補完はバス到着時刻予測における欠損値補完手法として適していないと言える．

パターン補完は，1運行先予測の場合は他の欠損理補完手法を適用した場合のMAEと有意差が見られない．一方，2運行先や3運行先の予測の場合には，他の欠損理補完手法よりもMAEが小さかった．このことから，特に複数運行予測の場合に適した欠損理補完手法であると言える．

組み合わせ手法は，おおむねパターン補完と同じ傾向を示した．訓練データの欠損率を変更した場合にパターン補完よりわずかに小さいMAEである場合があった．一方，テストデータの欠損率を変更した場合はパターン補完の方が小さいMAEであった．このことから，パターン補完と同様に複数運行予測の場合に適した欠損理補完手法であるが，比較的訓練データの欠損理補完に向いている手法であると考えられる．

7. 考察

第 6 章では，各欠損値補完手法がバス到着時刻予測に適しているかをバス到着時刻予測の誤差から評価した．本章では，特に現実で起こりうる欠損率 10%〜30%の範囲について，バス運行データセットの補完結果などを元に実験結果を分析し，各提案手法の特徴について考察する．また，今後の展望についても述べる．

7.1 バス運行データセットの補完結果の誤差

各提案手法の特徴を分析するため，まずはじめに，各欠損値補完手法の補完結果に対する誤差（MAE）を調査した．元のデータセットで欠損している部分は正解値が分からず MAE を求めることができないため，人為的に欠損させた部分に対する補完結果の MAE を計算した．MAE は各走行区間ごとの走行時間と各バス停ごとの停車時間について，訓練データとテストデータを分けて調査した．また，調査対象は特に現実で起こりうる欠損率 10%〜30%の範囲に限定した．走行区間 b の走行時間に対する MAE（MAE_{r_b}）の計算方法を式 (6) に，バス停 b の停車時間に対する MAE（MAE_{s_b}）の計算方法を式 (7) に示す．

$$\mathrm{MAE}_{r_b} = \frac{1}{n}\sum_{i=1}^{n} |{r'}_{b,i} - r_{b,i}| \tag{6}$$

$$\mathrm{MAE}_{s_b} = \frac{1}{n}\sum_{i=1}^{n} |{s'}_{b,i} - s_{b,i}| \tag{7}$$

ここで，n はテストデータのデータ数，${r'}_{b,i}$（${s'}_{b,i}$）は走行時間（停車時間）の補完結果，$r_{b,i}$（$s_{b,i}$）は走行時間（停車時間）の正解値である．

図 33〜38 に訓練データの補完結果の MAE を，図 39〜44 にテストデータの補完結果の MAE を示す．それぞれの走行区間・バス停ごとに，走行時間・停車時間の補完結果 MAE を補完手法ごとにまとめている．

全体を通じてみると，ほとんどの場合において，パターン補完の補完結果 MAE が小さくなっていることが分かる．特に，他の欠損値補完方法と比較して走行区間 4・5 といった走行時間が長い区間での MAE が小さくなっている．時間的補完はパターン補完と比較すると，すべての走行区間・バス停において補完結果 MAE

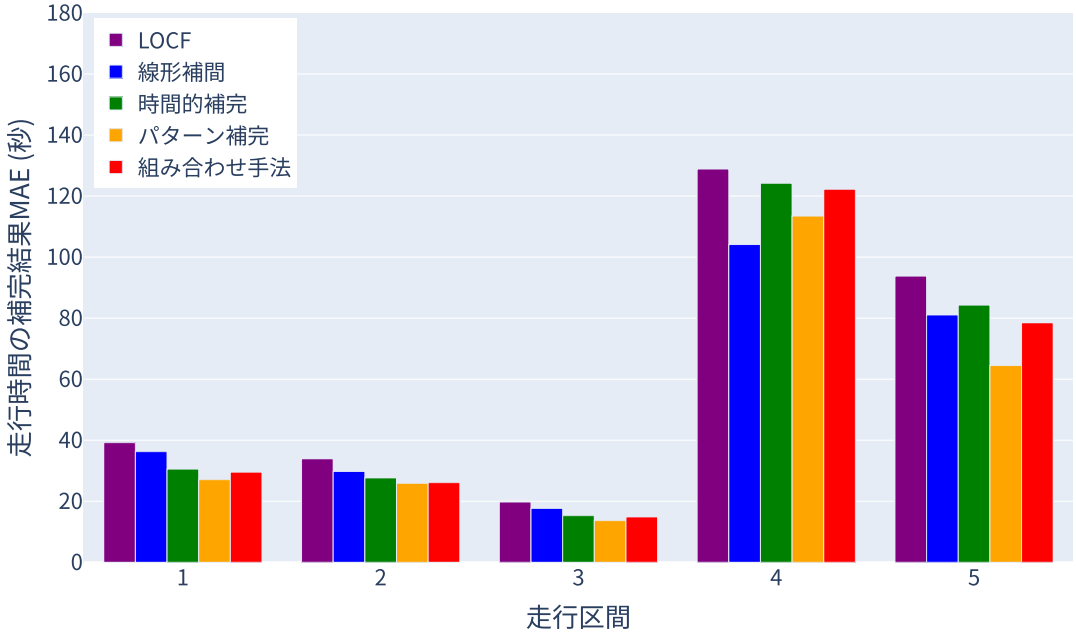


図 33: 訓練データ（欠損率 10%）の走行時間の補完結果 MAE

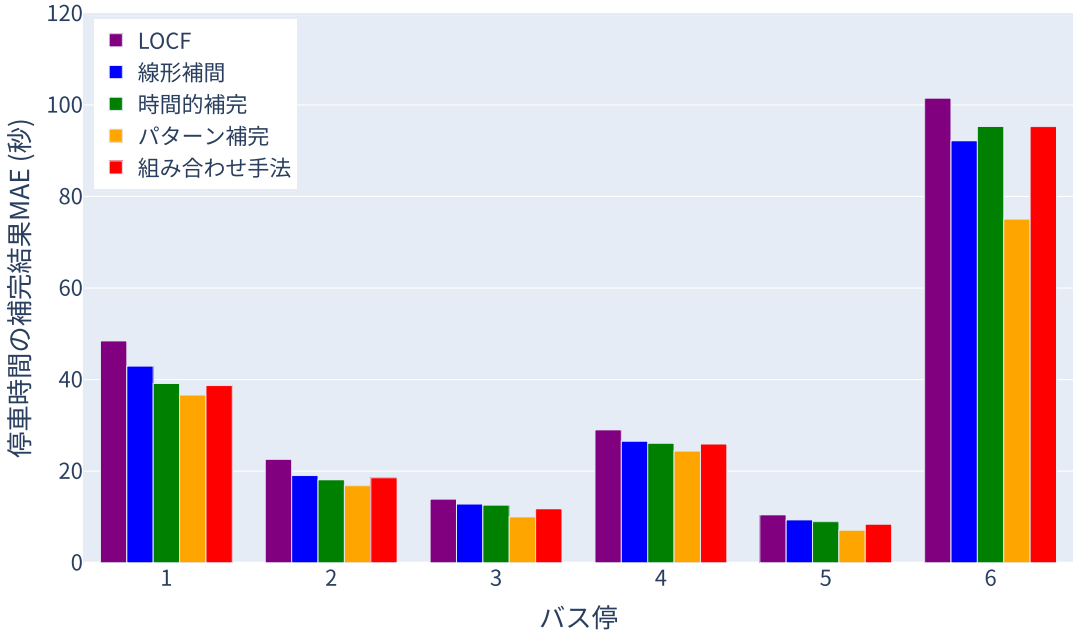


図 34: 訓練データ（欠損率 10%）の停車時間の補完結果 MAE

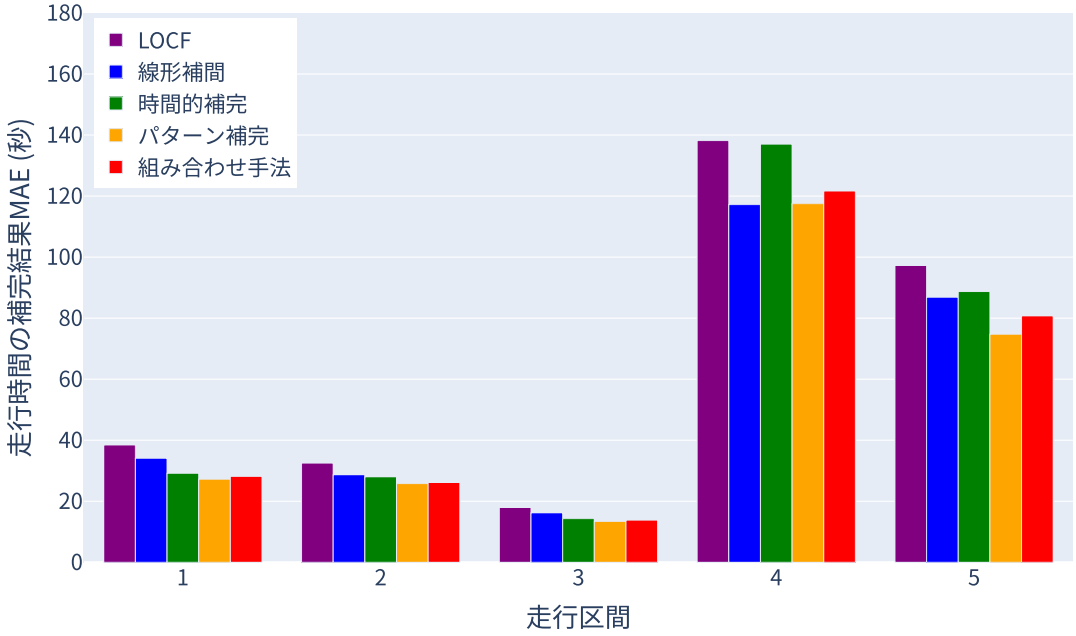


図 35: 訓練データ（欠損率 20%）の走行時間の補完結果 MAE

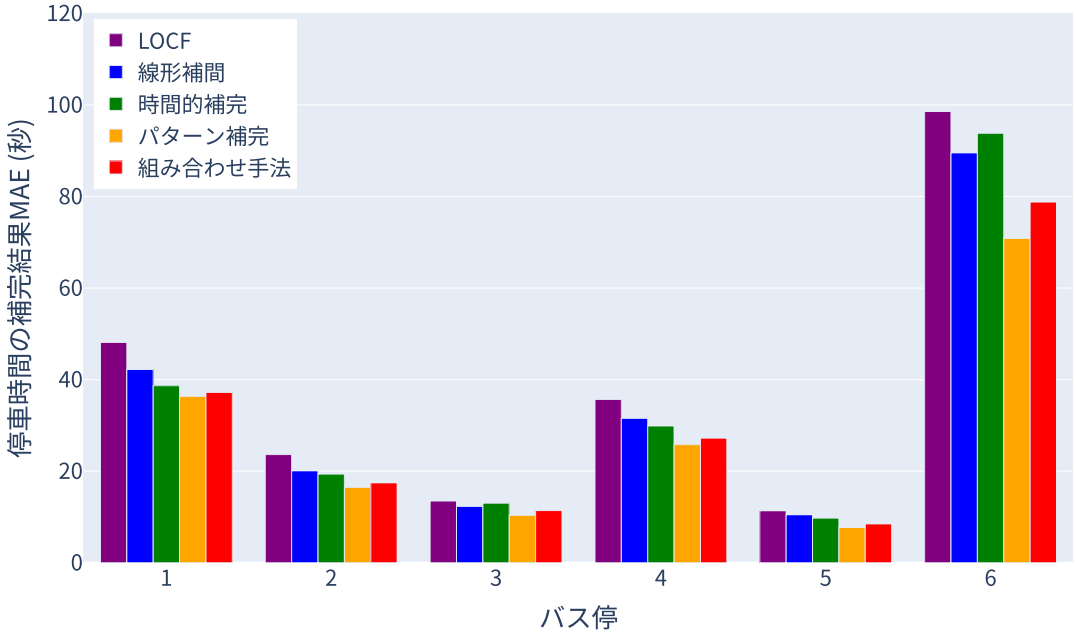


図 36: 訓練データ（欠損率 20%）の停車時間の補完結果 MAE

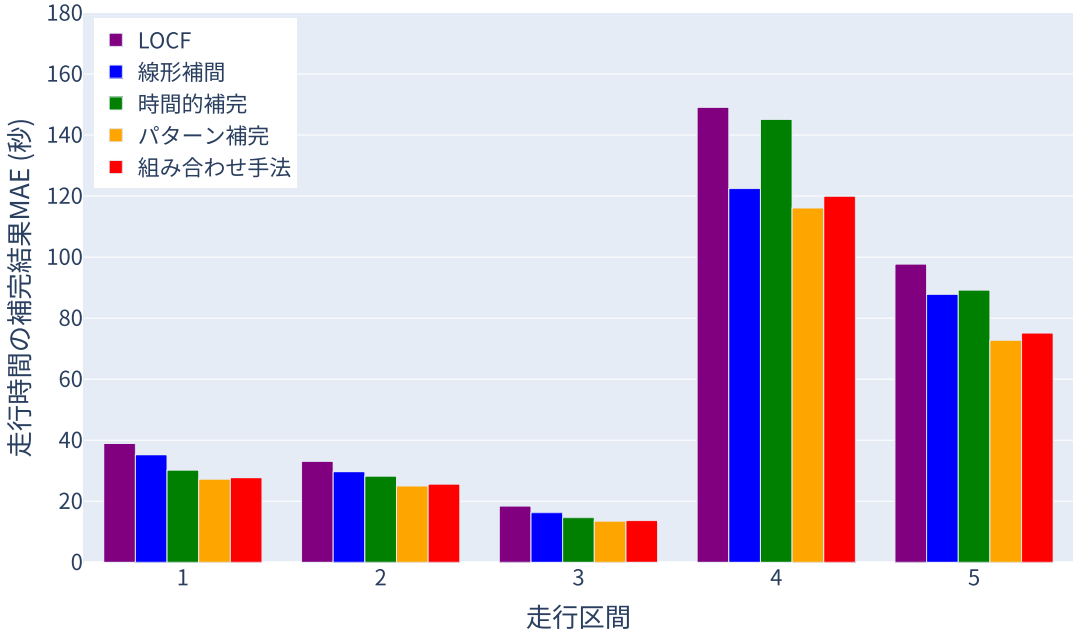


図 37: 訓練データ（欠損率 30%）の走行時間の補完結果 MAE

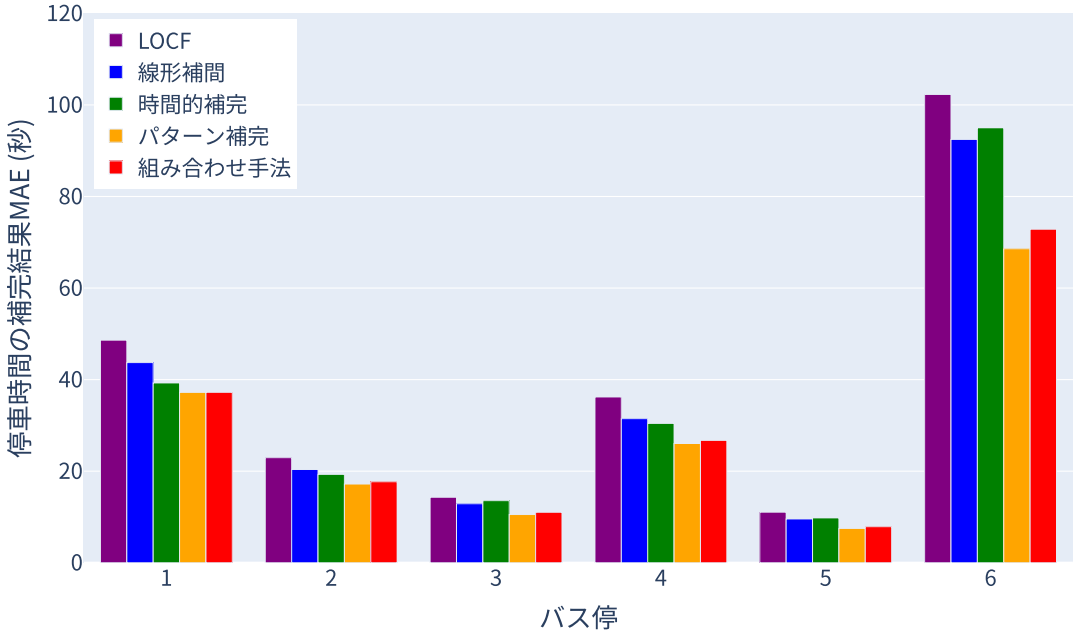


図 38: 訓練データ（欠損率 30%）の停車時間の補完結果 MAE

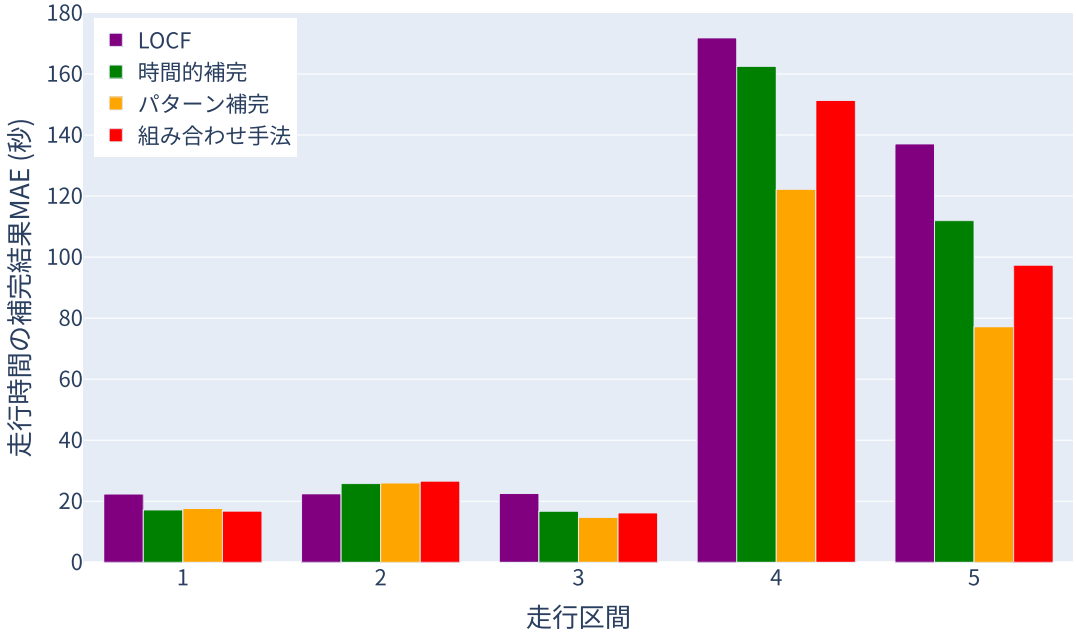


図 39: テストデータ（欠損率 10%）の走行時間の補完結果 MAE

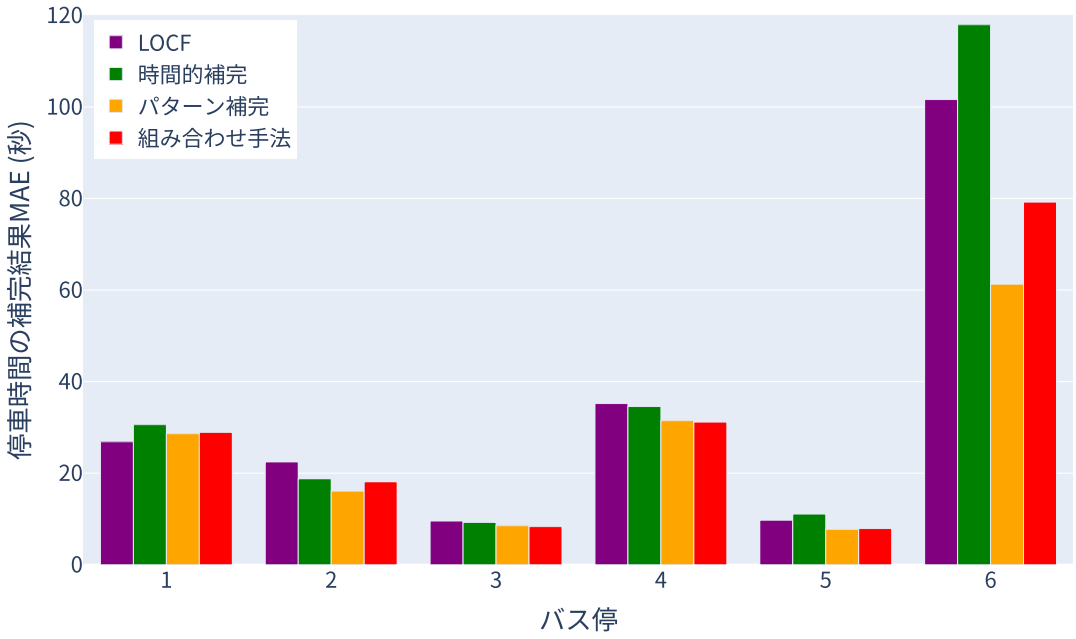


図 40: テストデータ（欠損率 10%）の停車時間の補完結果 MAE

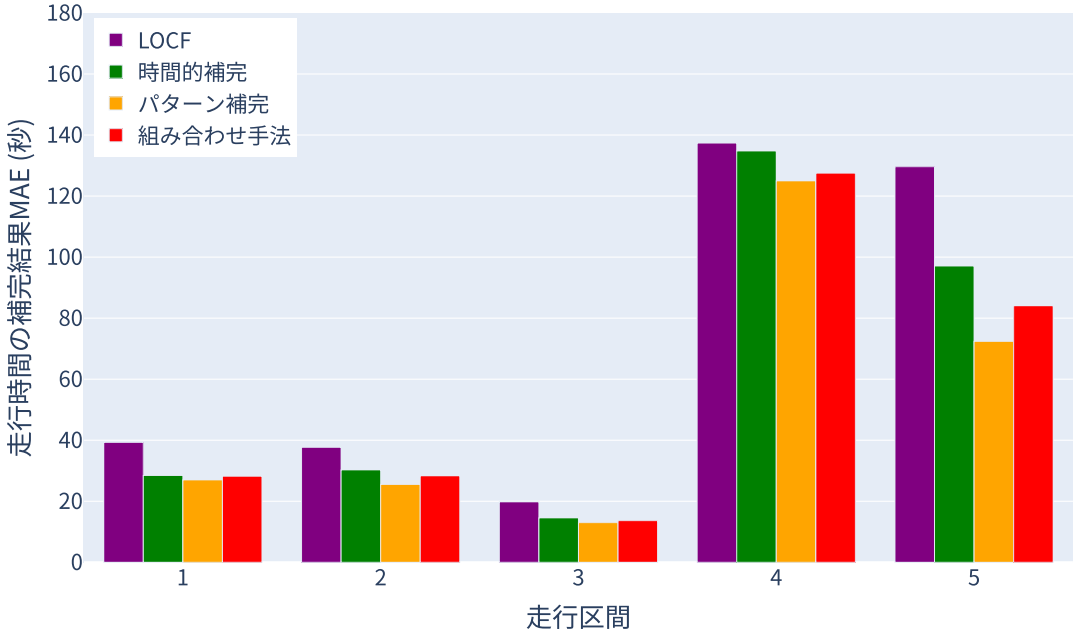


図 41: テストデータ（欠損率 20%）の走行時間の補完結果 MAE

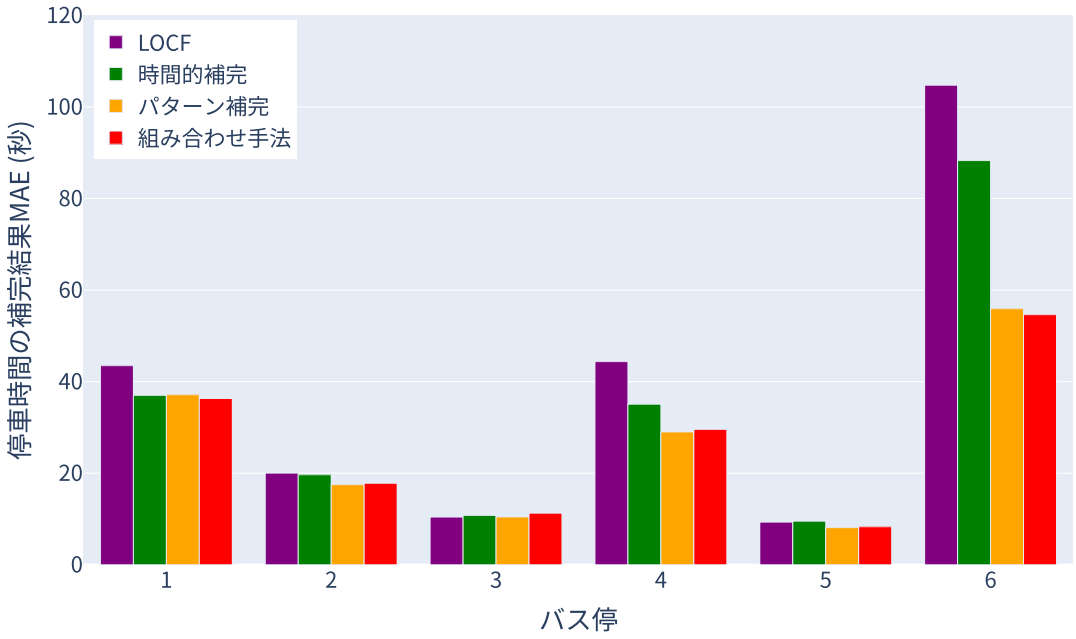


図 42: テストデータ（欠損率 20%）の停車時間の補完結果 MAE

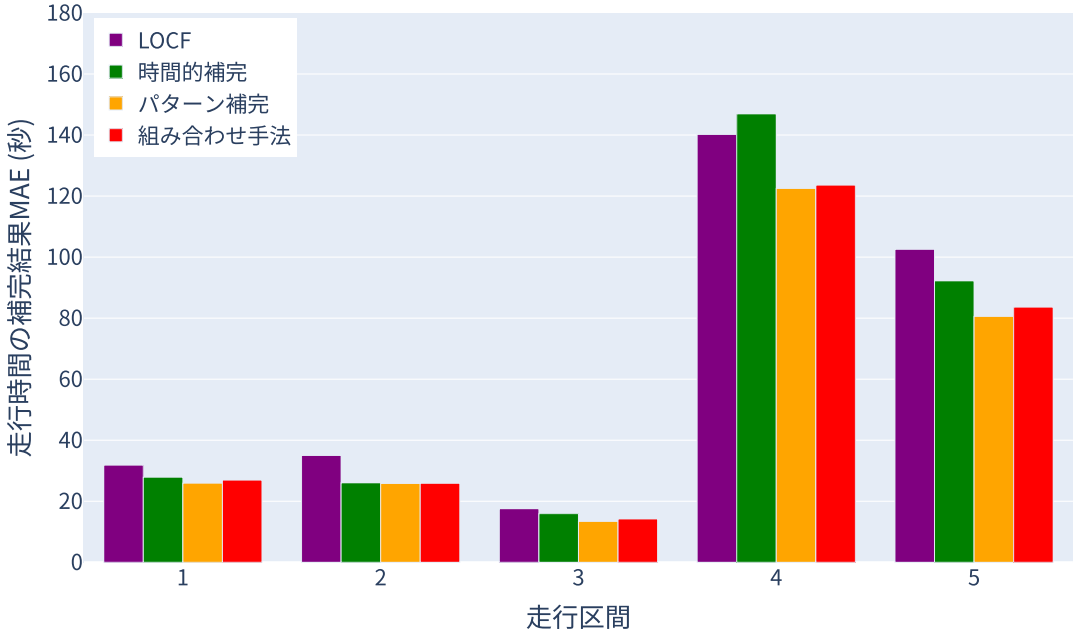


図 43: テストデータ（欠損率 30%）の走行時間の補完結果 MAE

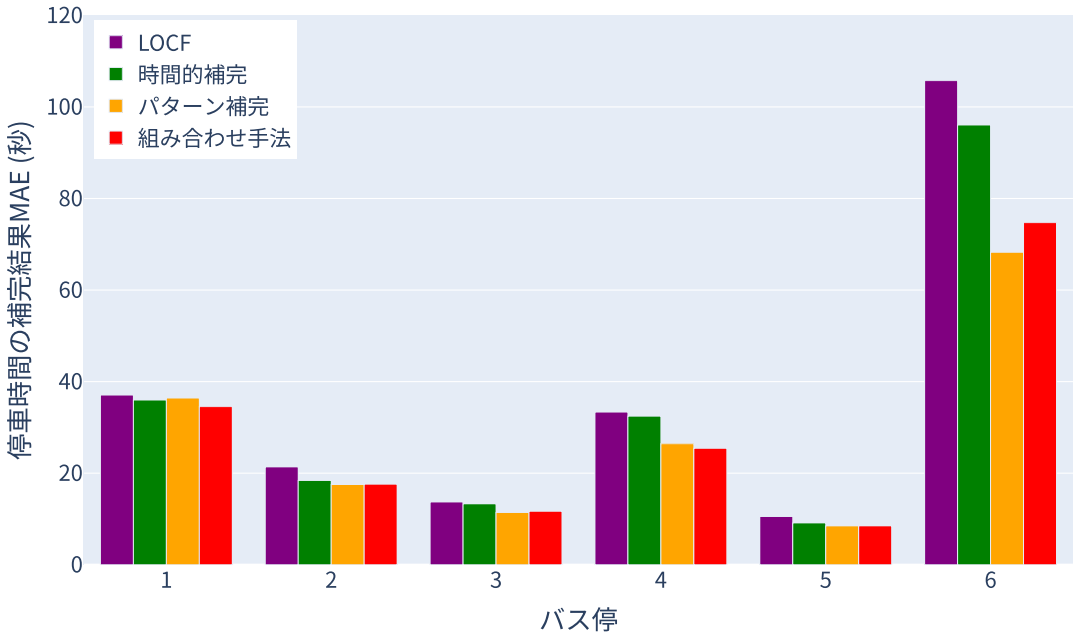


図 44: テストデータ（欠損率 30%）の停車時間の補完結果 MAE

が大きくなっており，組み合わせ手法は常にパターン補完と時間的補完の中間の補完結果 MAE であった．

7.2 各提案手法ごとの特徴

本節では，7.1 節で求めた各欠損値補完手法の補完結果 MAE に加え，補完結果の具体例，予測結果の具体例を元に各提案手法の特徴を考察する．

7.2.1 時間的補完

時間的補完は，欠損値補完においてバス運行データの特徴を考慮できると予想し，バス到着時刻予測の誤差も減少すると期待していた．しかし，評価結果では LOCF よりも予測誤差が大きいと分かり，バス到着時刻予測に適していないという結論になった．

図 33〜44 から LOCF と時間的補完の補完結果 MAE を比較してみると，多くの場合で補完結果 MAE は時間的補完が小さいものの，バス到着時刻予測の MAE は時間的補完を適用した場合に大きくなっている．このことから，必ずしも誤差の少ない補完方法が予測誤差の削減に繋がるわけではないと言える．この原因を分析するため，補完結果の具体例を調査した．

時間的補完による補完結果の具体例として，図 45 に 2021 年 9 月 17 日の走行時間 4 を時間的補完（$N_{\mathrm{mean}} = 5$）で補完した結果を示す．予想では直前数運行の平均値を欠損値に当てはめることで，直前の運行の乱れを反映できると考えていた．図に示す 10 便目や 15 便目などは予想どおり正解値に近い値で補完ができている．一方 12 便目と 13 便目では，正解値が減少しているのにも関わらず，補完結果が高い値を示している．また 17 便目と 18 便目では，正解値が増加しているのにも関わらず，補完結果が低い結果を示している．このように時間的補完では，正解値が急激な増減をした場合に，正解値と逆の増減を示す問題があるといえる．LOCF は欠損前の値を繰り返すため，正解値に存在する増減を再現はできなくなるものの，一定値を示すため逆の増減を示すことは無い．しかし，時間的

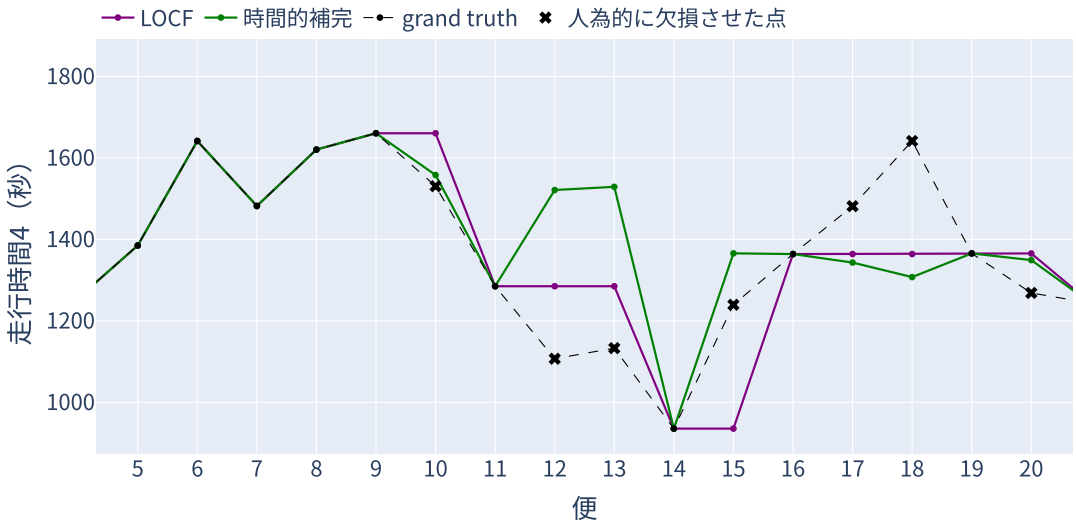


図 45: 2021 年 9 月 17 日の走行時間 4 を時間的補完と LOCF で補完した結果

補完では増減が再現できないどころか逆の増減を示してしまうことがあるため，この点が学習や予測に大きく影響を与えてしまったと考えられる．

7.2.2 パターン補完

パターン補完は，4.3.2 節で分析した日ごとの周期性に基づき補完をするため，誤差の少ない補完ができると予想した．7.1 節で述べたとおり，補完誤差は他の補完手法と比較して小さいことが確認できた．この原因を補完結果の具体例を元に分析する．

図 46 に，2021 年 9 月 17 日の走行時間 4 をパターン補完で補完した結果を示す．10，12，13，15 便目が比較的小さい誤差で補完できていることが分かる．図 45 に示した LOCF や時間的補完と比較しても全体的に誤差の少ない補完ができていると分かる．一方，17，18 便目に着目すると補完誤差が大きくなっている．パターンデータ（欠損していない運行の平均値）と比較すると，正解値が大きく外れていることから，この 17，18 便目は運行が遅れていたと言える．このように運行が乱れた際には，パターン補完による補完は追従できないことも分かる．このことから，運行が安定した路線であるほどパターン補完が有効であり，運行が

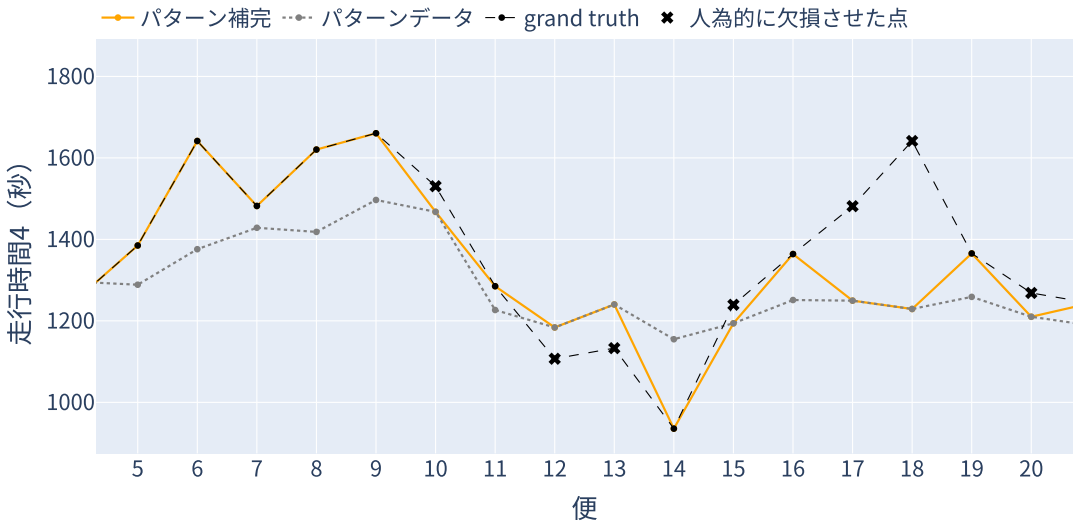


図 46: 2021 年 9 月 17 日の走行時間 4 をパターン補完で補完した結果

乱れやすい路線では補完誤差が増えていくことが考えられる．

次に，6.6.3 節で述べた，1 運行先予測の場合に他の欠損値補完手法との予測誤差の有意差が見られなかった原因について，予測結果の具体例を元に分析する．

テストデータの期間のうち運行が安定した日として 2022 年 9 月 23 日を，運行が乱れた日として 2022 年 9 月 17 日を選択し，その予測結果を比較する．運行が安定した日（2022 年 9 月 23 日）の予測結果を図 47 に示す．また，運行が乱れた日（2022 年 9 月 17 日）の予測結果を図 48 に示す．比較の都合上，縦軸は始点のバス停出発から終点バス停到着までの所要時間を表している．

運行が安定した日の予測結果に着目すると，予測運行数に関わらずほぼ同じ予測結果となっている．欠損値補完手法ごとにもほぼ変わらない結果となっているが，18 運行目の予測に着目すると，パターン補完が最も正解値に近い所要時間を予測していることが分かる．また，LOCF の予測所要時間は高めの値となっている．

運行が乱れた日の予測結果に着目すると，運行が大きく乱れた 7〜10 運行目の予測は，予測運行数，欠損値補完手法に関わらず大きく外している事がわかる．LOCF は 1 運行先予測においては 9 便目と 10 便目に少し追従できているが，2 運行先予測，3 運行先予測では運行の乱れのタイミングを外して予測しており，誤

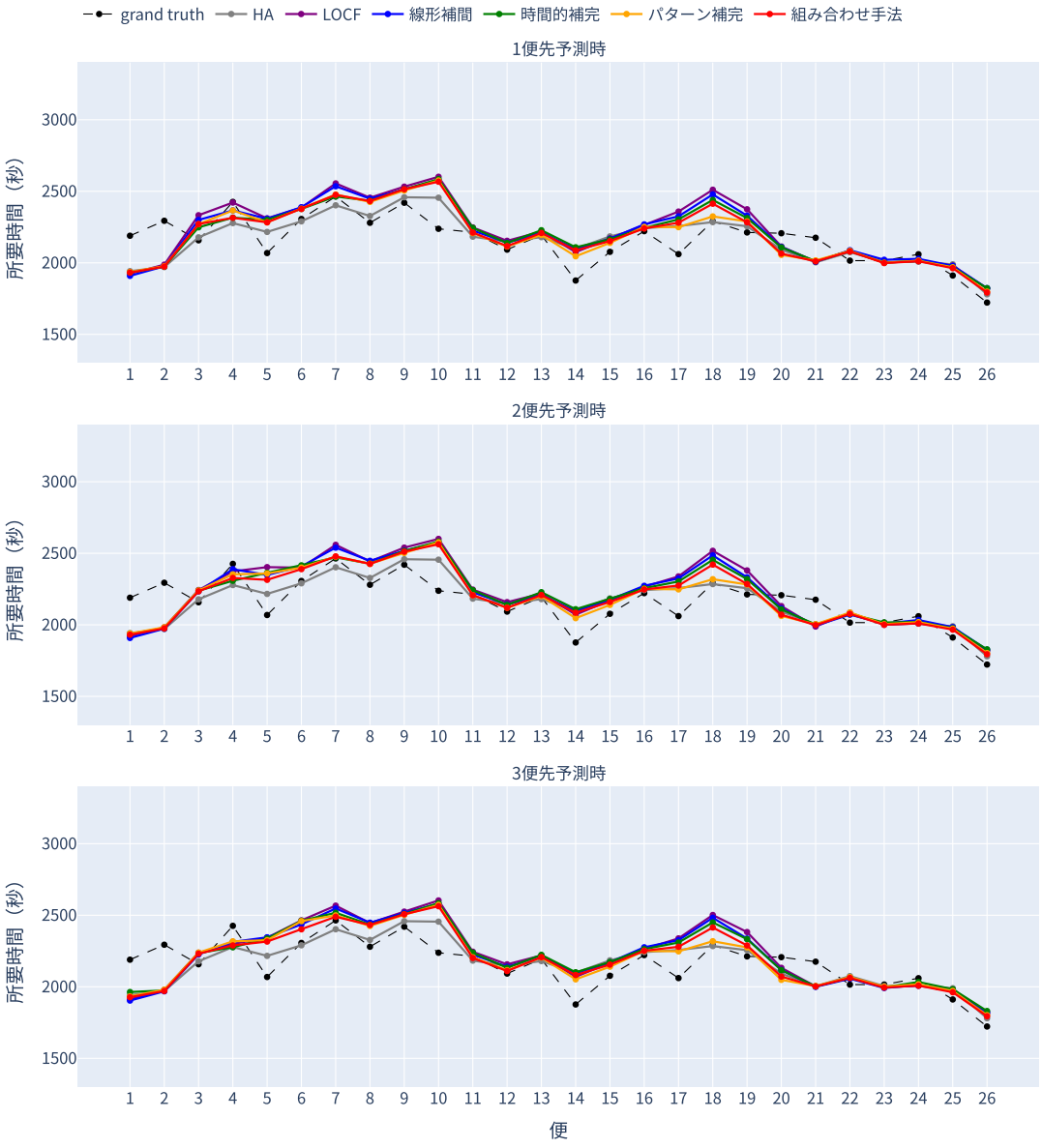


図 47: 予測運行数別の運行が安定した日（2022 年 9 月 23 日）の予測結果

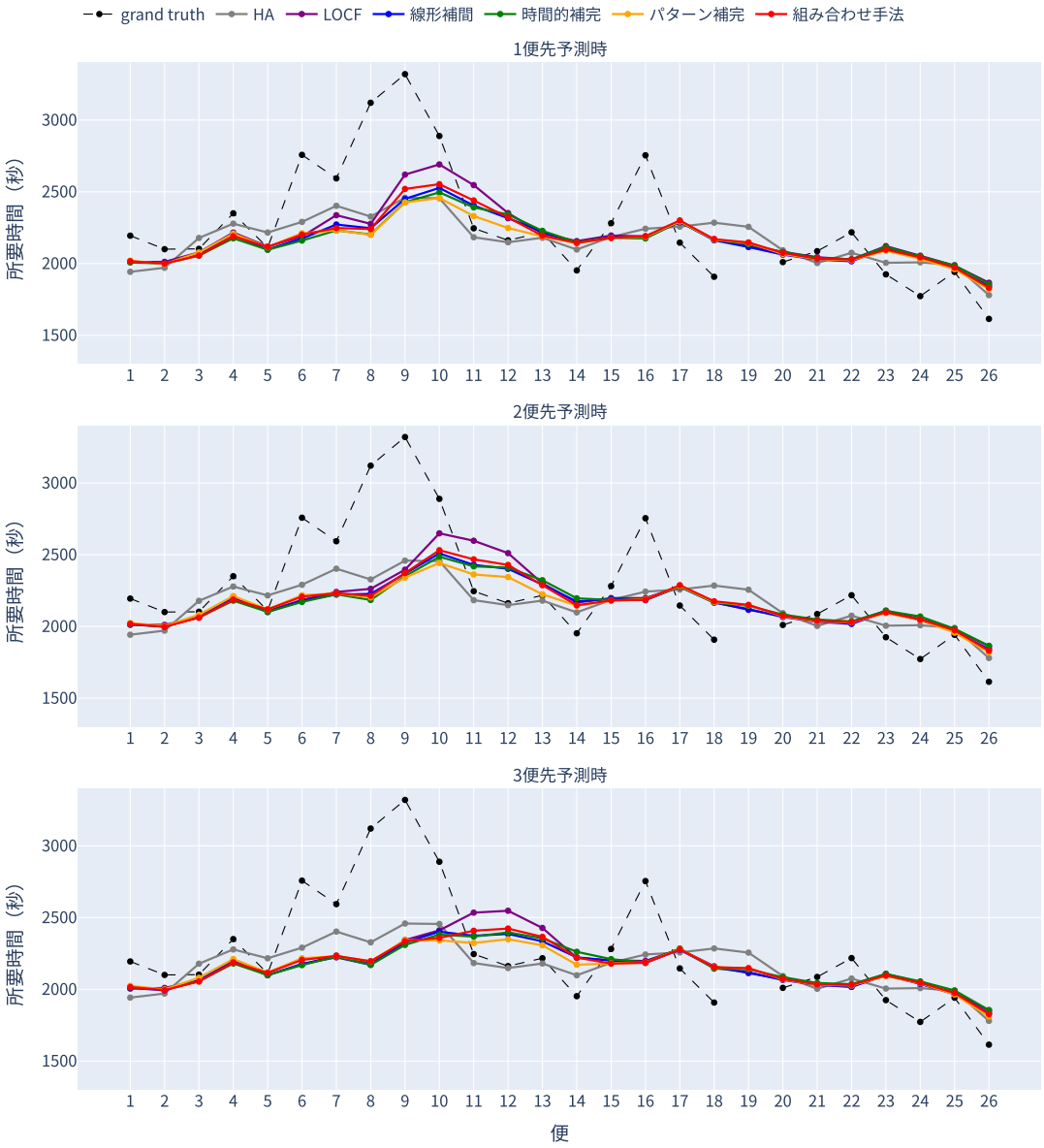


図 48: 予測運行数別の運行が乱れた日（2022 年 9 月 17 日）の予測結果

差が増えている事がわかる．一方，パターン補完は運行の乱れを予測できていないものの，11 便目以降で所要時間が短くなった際の予測で誤差が少なくなっている．

以上のことから，パターン補完は運行の乱れに対する追従性が低いが，パターンデータに近い予測結果を出すと言える．1 便先予測においては LOCF を使用した方が運行の乱れに追従できるため予測誤差が少なくなるが，複数運行先予測においては運行の乱れを正確に予測することが難しく，パターンに沿った予測をしたほうが，予測誤差が小さくなると考察される．

7.2.3 組み合わせ手法

組み合わせ手法は，時間的補完とパターン補完の欠点を補い合うことを期待していたが，図 33〜44 の補完結果 MAE はパターン補完より補完誤差が多い結果となった．また，テストデータの欠損率を変更させる実験ではパターン補完と時間的補完の中間の予測誤差であった．分析のため，組み合わせ手法による補完結果の具体例として，図 45 に 2021 年 9 月 15 日の走行時間 4 を組み合わせ手法

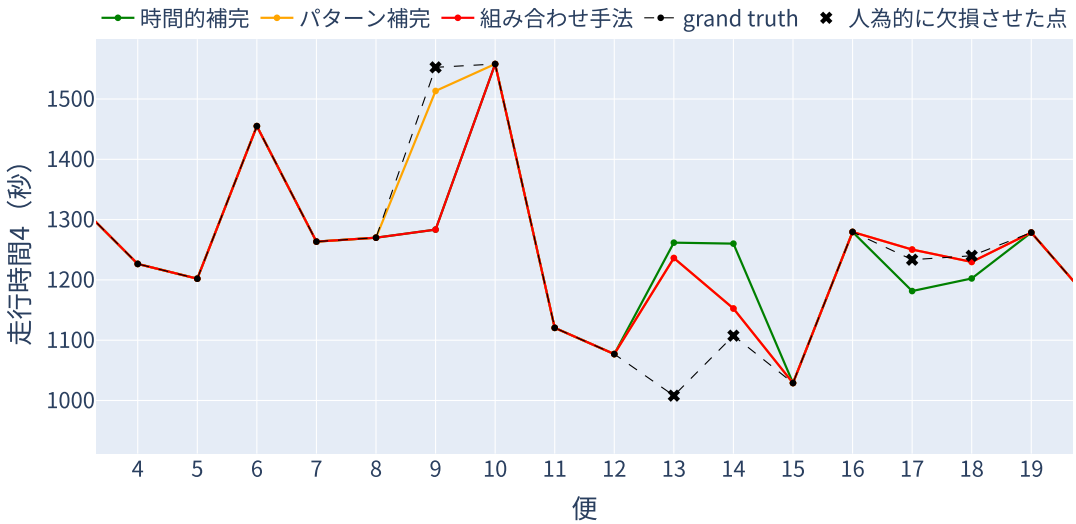


図 49: 2021 年 9 月 15 日の走行時間 4 を組み合わせ手法（$N_{\text{mean}} = 5$）で補完した結果

（$N_{\mathrm{mean}} = 5$）で補完した結果を示す．この例においては，9 便目の補完には直近 5 運行（4〜8 便目）に欠損が無いため時間的補完を使用しており，13，14，17，18 便目の補完には直近 5 運行に欠損があるためパターン補完を使用している．組み合わせ手法では，欠損部分の直近 N_{mean} 運行に対する欠損の有無で時間的補完とパターン補完を使い分けるが，9 便目のように時間的補完を選択した結果，補完誤差を増やしてしまう場合が多かったと考えられる．このことから，組み合わせ手法はパターン補完の利点を取り入れつつも，時間的補完の欠点を取り入れてしまったと予想される．

一方，訓練データの欠損率を変更させる実験では，訓練データの欠損率が 20%〜30%のときにパターン補完よりわずかに予測誤差の小さい部分が見られた．この原因について考察する．パターン補完は欠損部分を全てパターンで置き換えてしまうため，訓練データの分布に偏りが生じる．組み合わせ手法であれば，時間的補完を行う場所も含まれるため，パターン補完と比較すると分布の偏りが少なくなり，学習において良い効果が得られたと考えられる．例えば，図 48 の 1 運行先予測では，9 便目と 10 便目予測誤差が LOCF の次に少ない．このように，パターン補完と比較すると運行の乱れに対する追従性が向上すると考察される．

7.3 今後の展望

今後の展望として，まずはじめに時間的補完の改良が挙げられる．7.2.1 節でも述べたとおり，直近の運行の平均値を使用すると，実際の増減とは逆の増減を示すなど，正解値とは明らかに異なる補完をしてしまう点が，学習や予測に大きな影響を与えていると考えられる．また組み合わせ手法についても，時間的補完を使用することでパターン補完よりも補完誤差や予測誤差が増えていると推測される．一方でパターン補完では，直近の運行の乱れを取り入れることができないため，1 運行先予測では LOCF よりも予測誤差が大きくなっていると考えられる．そのため，直前の運行の影響を取り込む方法として時間的補完とは異なるアプローチが必要である．例えば，直前の運行が普段の運行よりどの程度遅れているかをパターンデータを元に計算し，バイアスをかけるといった方法が考えられる．

また，今回はバス運行データの特徴にのみ着目したが，欠損値補完の時点で気

象データを使用することも考えられる．過去の気象観測結果は 4.2 節で述べたとおり欠損がほとんどないため，活用しやすいと考えられる．他に活用の可能性があるデータとして，交差点や信号の数などの道路情報，周辺の道路の混雑状況などがあげられる．これらのデータも併用することでより効果的な欠損値補完ができると考える．

8. おわりに

路線バスのサービス品質向上の手段の１つとしてバス到着時刻予測があり，利便性向上やバス運行管理の面で重要な意味を持っている．これまでのバス到着時刻予測の研究では走行時間や停車時間といったバス運行データを前提としているが，バス運行データの計測では頻繁に欠損が発生する．既存のバス到着時刻予測の研究分野においては，バス運行データの特徴に着目した欠損値補完手法は検討されてこなかった．一方，バス到着時刻予測と近い研究分野である渋滞予測の分野では，データの特徴に着目した欠損値補完手法を適用することで予測誤差を削減できたと報告された．そこで本研究では，バス到着時刻予測の誤差削減を期待し，バス運行データの特徴に着目した時間的補完，パターン補完，およびそれらの組み合わせ補完の３つの欠損値補完方法を提案した．提案手法がバス到着時刻予測に適しているかを単純な欠損値補完手法と比較し評価した．評価の結果，バス運行データの日周期性に着目したパターン補完が，２運行先や３運行先を予測する場合に有効であると確認できた．欠損値補完手法のさらなるの改善のためには，日周期性だけではなく直前の運行の影響を取り込む手法が必要である．このため，パターンデータから運行の遅れを計算し反映する，気象データや道路状況データと組み合わせるなどが考えられる．

謝辞

主指導教員であり，研究の独自性や実用性の観点を中心に適切な研究指導をしていただきました本学情報基盤システム学研究室の藤川和利教授に心から感謝いたします．副指導教員であり，研究の方向性について的確に助言していただきました本学ユビキタスコンピューティングシステム研究室の安本慶一教授に心から感謝いたします．副指導教員であり，研究指導や論文添削をはじめ様々な側面からサポートしていただきました本学情報基盤システム学研究室の新井イスマイル准教授に心から感謝いたします．時間のない中，国際会議の原稿の添削をしていただき本当にありがとうございます．研究指導だけではなく，学内システムやネットワーク運用技術などについてもご教示いただきました本学情報基盤システム学研究室の垣内正年助教に心から感謝いたします．研究の目的意識や論文執筆に関して，具体的にアドバイスしていただきました本学情報基盤システム学研究室の遠藤新助教に心から感謝いたします．様々な事務手続きだけではなく，研究室内での生活の中でたくさんのコミュニケーションをとっていただきました本学総合情報基盤センターの辻元理恵女史に心から感謝いたします．研究で使用するデータを提供していただいたみなと観光バス株式会社の皆様に深く感謝いたします．研究活動だけにとどまらず，日々の生活，進路といった様々な場面で相談に乗っていただいた本学情報基盤システム学研究室の大平修慈氏，桂祐成氏に感謝いたします．そして，共に切磋琢磨し博士前期課程を過ごした情報基盤システム学研究室同期の福田匠氏，松永拓也氏，山村竜也氏，ならびに同研究室の学生の皆様に感謝いたします．最後に，博士前期課程への進学にあたり，私の意思を尊重し，経済面や生活面での援助や励ましとともに 2 年間見守ってくださいました家族に心から感謝いたします．

参考文献

[1] N. Singh, and K. Kumar, "A review of bus arrival time prediction using artificial intelligence," *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 4, p. e1457.

[2] N. C. Petersen, F. Rodrigues, and F. C. Pereira, "Multi-output bus travel time prediction with convolutional LSTM neural network," *Expert Systems with Applications*, vol. 120, pp. 426–435, 2019.

[3] 石長 篤人，新井 イスマイル，垣内 正年，藤川 和利，"運行情報と気象情報の畳み込みによるバス到着時刻予測手法"，研究報告高度交通システムとスマートコミュニティ（ITS），vol. 2021-ITS-84，no. 6，pp. 1–8，2021.

[4] D.-H. Shin, K. Chung, and R. C. Park, "Prediction of traffic congestion based on LSTM through correction of missing temporal and spatial data," *IEEE Access*, vol. 8, pp. 150 784–150 796, 2020.

[5] J. Chocholac, D. Sommerauerova, J. Hyrslova, T. Kucera, R. Hruska, and S. Machalik, "Service quality of the urban public transport companies and sustainable city logistics," *Open Engineering*, vol. 10, no. 1, pp. 86–97, 2020.

[6] A. Gooze, K. E. Watkins, and A. Borning, "Benefits of real-time transit information and impacts of data accuracy on rider experience," *Transportation Research Record*, vol. 2351, no. 1, pp. 95–103, 2013.

[7] J. Pang, J. Huang, Y. Du, H. Yu, Q. Huang, and B. Yin, "Learning to predict bus arrival time from heterogeneous measurements via recurrent neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 9, pp. 3283–3293, 2019.

[8] Q. Han, K. Liu, L. Zeng, G. He, L. Ye, and F. Li, "A bus arrival time prediction method based on position calibration and LSTM," *IEEE Access*, vol. 8, pp. 42 372–42 383, 2020.

[9] Z.-Y. Xie, Y.-R. He, C.-C. Chen, Q.-Q. Li, and C.-C. Wu, "Multistep prediction of bus arrival time with the recurrent neural network," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–14, 2021.

[10] B. Lee, H. Lee, and H. Ahn, "Improving load forecasting of electric vehicle charging stations through missing data imputation," *Energies*, vol. 13, no. 18, 2020.

[11] N. P. Olewuezi, "Note on the comparison of some outlier labeling techniques," *Journal of Mathematics and Statistics*, vol. 7, pp. 353–355, 2011.

[12] M. Schuster, and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[14] S. Ioffe, and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, pp. 448–456, 2015.

[15] G. Hinton, N. Srivastava, and K. Swersky, "Lecture 6d - a separate, adaptive learning rate for each connection." 2012, accessed on 2023-01-08. [Online]. Available: http://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf

[16] I. Arai, M. Kametani, N. Honda, and T. Akiyama, "DOCOR: Sensing everything from route buses," in *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, 2020, pp. 1–2.

[17] 気象庁，"過去の気象データ検索"，accessed on 2022/12/16．[Online]. Available: https://www.data.jma.go.jp/obd/stats/etrn/index.php

[18] 気象庁，“気象観測について”，accessed on 2022/1/13．[Online]. Available: https://www.jma.go.jp/jma/kishou/know/kansoku/weather_obs.html

[19] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, “Hyperband: A novel bandit-based approach to hyperparameter optimization,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6765–6816, 2017.