Vol.2018-IS-146 No.7
2018/12/8

バスの到着時刻予測モデルの開発と移動手段提案システムの検討

佐藤孝大[†1]　大場みち子[†2]

概要：近年，地方圏では都市部への人口流出や自家用車の利用数増加を理由に，公共交通機関利用者が年々減少している．バス事業者は路線の廃止・減便をせざるを得なく，バスの利便性が低下している．この利便性低下の対策の一つにバスロケーションシステムがある．これは，バスの時刻表や接近情報を提供するシステムであり，接近情報からバスの遅延を把握することができるため,利便性向上の一助となる．しかし，ロケーションシステムの接近情報に大きな誤りが含まれる可能性があり，かえって利便性の低下を招いている．通勤や通学時など，利用者の遅れが許されない状況で接近情報に誤りが生じると，バスへの不信感からさらなる利用者の減少も考えられる．そこで，本研究では普段バスを利用する人々の移動方法選択支援を目的に，機械学習を用いて翌日におけるバスの到着時刻予測モデルと混雑予測モデルを開発し，予測モデルを用いた移動方法提案システムを検討する．過去の運行実績データや天候データ，人口データを元にした機械学習を行った結果を報告する．

キーワード：機械学習，ランダムフォレスト，バスロケーションシステム

Development of Bus Arrival Time Predictive Model and Investigation of Moving Method Proposal System

KODAI SATO[†1]　MICHIKO OBA[†2]

Abstract: In recent years, public transportation users in provincial areas are decreasing year by year because of the use of private cars increased and population outflow. Bus companies have to reduce or abolish bus routes, the convenience of the bus services are declining. There are bus location systems for measures to this reduction in convenience. These systems provide bus timetables and approach information. Since the bus users can confirm the delay of the buses by these systems, convenience of bus services are improved. However, the approach information of the these systems are not accurate. This problem leads to a decline in the convenience of the bus services. If an error of the approach information occurs in a situation where bus users can not be delayed such as commuting or going to school, there is a possibility that number of bus users may decrease due to distrust of the bus. In this paper, we develop the bus arrival time prediction model and the congestion prediction model on the next day using machine learning, with the aim of supporting the choice of movement method for people who usually use the bus. And, we consider the moving method proposal system using each prediction model. Here we report the result of machine learning based on past travel record data, weather data, and population data.

Keywords: Machine Learning, Random Forest, Bus Location System

1. はじめに

近年，地方圏を中心に路線バスをはじめとする公共交通利用者が減少しており[1]，路線の廃止や減便による利便性低下が大きな課題となっている．利便性低下の対策にはインターネットを用いた情報提供が広く実施され，路線バスではバスロケーションシステムが函館市[2]をはじめとする様々な地域で導入されている．バスロケーションシステムは，走行中のバスの到着時刻予測情報や，乗り場などの情報を受け取れるため，利便性向上の一助となっている．しかし，提供される到着予定時刻情報には，大きな誤りが含まれる場合がある．誤った予測結果を掲示した場合，バスが到着するまでの待機時間が増えることや，最悪の場合にはバスに乗り遅れるという状況が発生する．よって，利便性の低下を招いてしまうといえる．乗車率が高い路線では，乗車率によって乗車を拒否される場合もあるが，この乗車率について情報提供は行われていない．遅延や乗車率は経験的には予測できない場合があり，通勤や通学など，遅れが許されない場合に発生すると利用者への影響が甚大になる．バスを利用する前日に，予測到着時刻や予想乗車率を利用者に提供することにより，これらの事態は防ぐことが可能になると考える．

本研究の目的は，普段バスを利用する人々の目的地への移動方法選択を支援し，日常活動を円滑化させることである．そのために，翌日のバスの到着時刻予測モデルの開発とそれを元にした，移動方法提案システムの検討を目標とする．

†1 公立はこだて未来大学大学院
Future University Hakodate Graduate School
†2 公立はこだて未来大学
Future University Hakodate

2. 先行・関連研究と研究課題

著者らは先行研究として運行実績データを用いた到着時刻予測に取り組んできた[3]. 説明変数に運行実績データを設定した重回帰分析を行い，停留所別の遅延時間を予測するモデルを開発した．しかし，既存のバスロケーションシステムより予測精度が 1~3 分ほど悪化し，十分な予測精度が得られなかった．

辰巳らは，天気，月，曜日などの質的データを用いたバスの所要時間予測を行っている[4]. 質的データには天気，台風の有無，気温，月，曜日，五十日，時間帯がある．質的データを説明変数とした数量化 I 類による分析を行い，所要時間に与える影響が強い要素を用いて所要時間予測を行った．この結果，月・曜日・時間帯別と，月・曜日・便別の平均所要時間を用いた場合の精度が高いことが示された．しかし，この予測は始点の次の駅から，終点の手前の駅までの総所要時間を予測対象としているため，任意の停留所間でバスの所要時間予測を行うことができない. また，乗車人数などの混雑に関する指標を考慮していない.

前川らはバスの乗降者数データを用いた遅延予測を行っている[5]. バスの運行実績データと乗降者数データの分析結果を元にバス利用者が乗者と降車に要する平均時間を独自に設定し，乗降者人数に応じてバスの遅延時間を算出した．この結果，週ごとの遅延時間平均値や曜日別平均値を利用した予測よりも乗降者数を利用した予測モデルが良好な結果を示した．しかし，通過する停留所が多くなるほど実遅延時間との誤差が増加しており，精度に課題がある．

これらの関連研究では，曜日特性や乗降者数などの様々なバス運行時のデータに着目し，運行に影響を与える要因としての有効性を示した．しかし，これらは翌日など未来を予測することを考慮しておらず，翌日におけるバスの到着時刻予測手法には利用できない．文献[5]では予測モデルの予測精度が低い結果であったが，原因として，予測モデルで考慮した要因が少ないことが考えられる．バスは電車などと違い，道路状況や渋滞，乗車人数など，多種多様な要因が運行に影響を与えていると考えられる．乗降者数のみを遅延の要因と考慮したため，他の要因によって発生していた遅延を考慮できず，精度が低下したと考えられる．文献[4]では質的データにのみ着目し，良好な精度であったが，予測結果に生じた外れ値を分析すると，降水時や時間帯などの影響を受けていることが分かった．従って，考慮すべき要因が漏れていることが，予測精度の低下を招いていると考えられる．また，これら先行研究と関連研究では重回帰分析などの手法で分析を手作業で行い，バスに関係のある要因を検討していた．前述した通りバスに影響を与える要因は多岐に渡る可能性があるため，手作業による分析の漏れが予測精度の低下や外れ値の発生を招いたことも考えられる．

3. 提案手法

第 2 章で述べた先行研究・関連研究での到着時刻予測の課題から，これまでの研究では考慮する要因が少ないこと，手作業による分析の漏れのため精度が悪化していることが挙げられる．これらの課題を解決するため，本研究では機械学習を利用した分析及び予測モデルの開発を行う．機械学習を用いることにより，手作業で分析しきれていない要素の漏れを無くし，既存手法より正確な予測を行うことができると考えた．また，機械学習は一般的に予測に考慮可能な要素数が多く，様々な要因を同時に考慮できることから精度改善につながるとも考えた．機械学習により翌日の便・停留所ごとの乗車人数と到着時刻を予測するモデルをそれぞれ開発する．

3.1 機械学習の手法検討

最適な機械学習手法を検討するため，同一データに対し複数の手法を用いて学習を行い，予測精度を評価する．機械学習には主に分類と回帰の 2 種類に分類されるが，本研究では乗車人数や到着時刻などの具体的な数値を予測するため，回帰を対象とする．機械学習の手法にはランダムフォレスト[6], サポートベクター回帰, エラスティックネットの 3 手法を用いる．ランダムフォレストは決定木モデルが基のアルゴリズムであり，利用可能なデータの型が豊富で外れ値や欠損値に対応しやすいという特徴を持つ．乗車人数や気象条件など，様々なデータ型を持つ要素を考慮でき, 欠損値の多い過去の実績データを利用できることから，本研究には適している手法であると考えた．サポートベクター回帰はサポートベクターマシンを回帰分析に応用した手法であり，過学習の防止やノイズに強いという特徴を持つ. 過去の実績データにはヒューマンエラーなども含まれ，予測結果にはこれらノイズの影響も出ると考えられるため，適した手法であると考えた．エラスティックネットは一般化線形モデルの回帰に正則化項を加味するモデルであり，多重共線性に強い．バスに関連するデータには乗車人数と遅延時間など, 相関の考えられるものが多く存在するため，この影響を軽減することができると考えた．目的変数には翌日の便・停留所ごとの乗車人数と到着時刻のダイヤとの誤差を設定し，説明変数には節 3.3 で述べる要素を設定し，学習を行う．

3.2 利用ツール

機械学習には統計解析ソフトである R[7]を用いる．機械学習パッケージの一つである”randomForest”パッケージ[8]では，学習後に説明変数がどれほど目的変数を説明できているかという重要度を算出することができる．

3.3 説明変数の一覧

本研究で取り扱う説明変数を以下に示す.

(1) **停留所別の利用者数予測**

- 天候（天気予報の気温，降水確率）
- 人口（停留所周辺の世帯数，年齢別人口）
- 過去の利用者数（乗車数，降車数，通過数）

(2) **停留所別の到着時刻予測**

- 天候（天気予報の気温，降水確率）
- 過去の利用者数（乗車数，降車数，通過数）
- 過去の運行実績（到着時刻，ダイヤと到着時刻の誤差平均）
- 道路データ（停留所間距離，信号数（信号交差点密度[m/信号数]））

過去の実績値や天気予報，道路データを説明変数として設定し，学習を行う．学習結果から各説明編集の重要度を算出し，重要度が他と比べて明らかに低いものを除外していくことでチューニングを行う.

4. 実験

本章では, ランダムフォレスト, サポートベクター回帰, エラスティックネットの 3 手法の中から予測モデル開発に最も適している手法を調査するために行った実験の内容とその結果について述べる.

4.1 概要

停留所間の到着時刻予測及び停留所別の利用者予測における最適な手法を調査するため，短期間の運行実績データを用いてランダムフォレスト，サポートベクター回帰，エラスティックネットによる学習を行い，精度を評価した．モデルの精度評価は RMSE[9]を算出することで行う．RMSE は平均二乗誤差平方根とも言い，数値がゼロに近いほど予測精度が良いことを示す．RMSE の数式を式(1)に示す．N は予測対象の総データ件数であり，後述するテストデータのデータ件数がこれにあたる．y_iは実績値, $\hat{y}_i$は予測値である.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (1)$$

RMSE を算出後，実績値と予測値の誤差を計算し，グラフ化することで外れ値の発生具合も調査し，予測モデルの改善に繋げる.

4.2 対象路線と対象データ

函館における主要なバス 1 路線を予測対象とした．この路線は，函館駅前などの中心地を経由し，郊外の大学まで走行する路線であり，社会人や学生など，普段から多くの人々が利用する．路線図を図 1 に示す．なお，名称の長い停留所名は表記を省略している．始点から終点まで 29 カ所の停留所が存在する．中心地を通る「函館駅前」から「亀田支所」までの 16 停留所の区間は社会人の利用が普段多く，「富岡」から「未来大学」までの 12 停留所の区間は市街地から郊外の大学へ向かう学生の利用が多いという特徴がある.

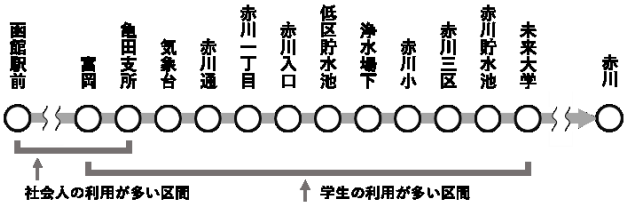

図 1　実験対象路線の路線図

社会人の利用が多い前半の区間では，他のバスや路面電車との乗り換えが頻繁に利用されており，この区間での予測には考慮すべき要因が非常に多いと言える．バス以外の交通のデータも必要となるが, 一般には公開されておらず, 様々な種類のデータの入手は困難であった．そのため，この実験では，普段からバスが主として学生に利用されている大学までの 12 停留所を予測対象とした．目的変数には 2016 年 12 月の実績データを利用し，データ件数は乗車人数の予測では 3,684 件, 到着時刻の予測では 827 件である．到着時刻の実績データにエラーデータが多く含まれていたため，到着時刻予測のデータ件数が少なくなっている．各データセットのうち 80%のデータを教師データ，20%のデータをテストデータとした．表 1 に使用した説明変数を示す．明日のバスを予測するため，明日における天気の指標として天気予報を説明変数に設定している．しかし，天気予報は一般的にリアルタイムでの利用を想定されているため，実際に公開された過去の天気予報を入手することができなかった．そのため，今回の予備実験では 1 日に何時間の降水・降雪が観測されたかの割合を降水確率とみなして利用した．具体的には，1 日に 10 時間降水記録があった日の降水確率を 10 時間 / 24 時間 ≒ 40% といったように算出した.

表 1 目的変数と説明変数

目的変数	便・停留所ごとの乗車人数 [人]	便・停留所ごとの到着時刻とダイヤの誤差 [秒]
説明変数	● 同便・同週の過去 1 か月乗車人数平均 [人] ● 各停留所周辺の青年人口（15 歳～24 歳）[人] ● 降水確率 [%] ● 気温 [℃]	● 同便・同週の過去 1 か月平均到着時刻誤差[秒] ● 乗車人数 [人] ● 降水確率 [%] ● 気温 [℃]

4.3 結果と考察

4.3.1 便・停留所ごとの乗車人数予測

ランダムフォレストの結果を図 2 に，サポートベクター回帰の結果を図 3 に，エラスティックネットの結果を図 4 に示す．グラフは実績値と予測値の誤差を示しており，予測結果を停留所の通過順に下からソートしている．グラフ左の文字は停留所名を示す．左に凸のグラフは予測値が実績値より多くなったデータを，右に凸のグラフは予測値が実績値より少なくなったデータを表している．

図 2 ランダムフォレストにおける実績値と予測値の差
(乗車人数予測，n=736)

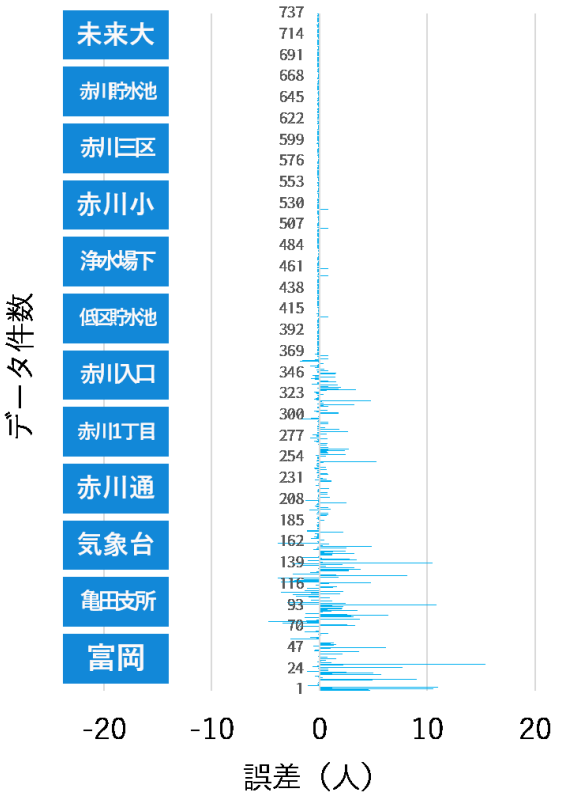

図 3 サポートベクター回帰における実績値と予測値の差
(乗車人数予測，n=736)

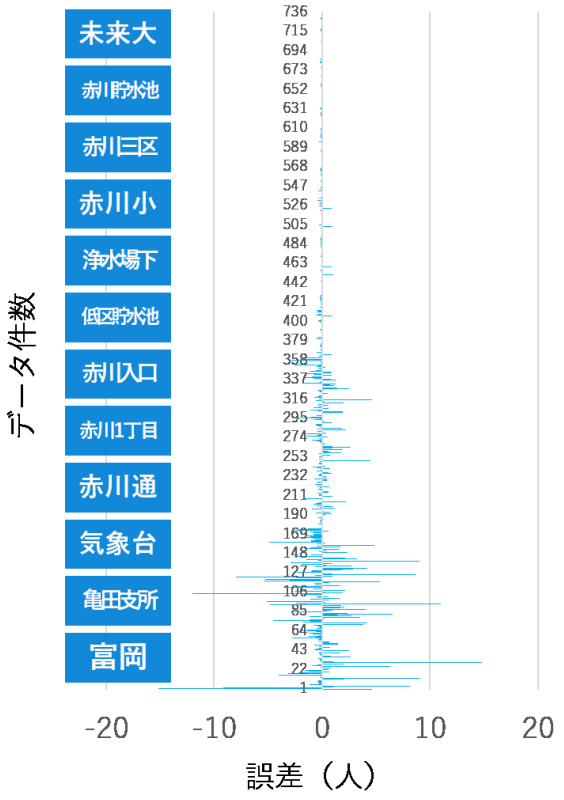

図 4 エラスティックネットにおける実績値と予測値の差
(乗車人数予測，n=736)

ランダムフォレストの RMSE は 0.43，サポートベクター回帰は 1.56，エラスティックネットは 1.67 となった．全手法で良好な結果を示したが，その中でもランダムフォレストが特に優れた精度となった．図 2 と図 3 を見ると，「赤川通」から「未来大」までの区間では予測誤差 6 人以内の結果となっており，特に良好な予測結果となっている．しかし，「富岡」から「気象台」までを見ると最大で 15 人の大きな外れ値が発生しており，この区間では必ずしも精度の高い予測を行えていない．図 4 のエラスティックネットの場合では，他の 2 手法と同様に「赤川通」から「未来大」までの区間は良好であったが，「富岡」から「気象台」にかけての精度が他の 2 手法よりも明確に悪化しており，10 人以上の予測誤差が多く発生した．

外れ値の発生した原因として時間帯の要因が考えられる．実績より 10 人前後少ない人数が予想された便では，各停留所での乗車人数が 10 人から 15 人と利用が極端に集中しているという特徴があった．大学で行われる講義の開始時刻に間に合う便に学生のバス利用が極端に集中したため，外れ値が生じやすかった可能性がある．予測結果が実績値よりはるかに大きくなった原因にはイベントの要因が考えられる．これらのデータの便では，普段から多くの利用者が見込まれる便だったが，実績の乗車人数が普段より少なく，1〜2 人の利用となっていた．12 月のデータのため，冬季休暇と重なり，学生の利用が普段と異なる状態であったため外れ値が生じやすかった可能性がある．これらのことから，時間帯における評価や，学生特有の長期休暇や定期試験などのイベントを学習に考慮することで精度が向上する可能性があることが示唆された．

4.3.2 便・停留所ごとの到着時刻誤差予測

ランダムフォレストの結果を図 5 に，サポートベクター回帰の結果を図 6 に，エラスティックネットの結果を図 7 に示す．項目等は誤差の単位以外は図 2 と同様である．

ランダムフォレストの RMSE は 28.05，サポートベクター回帰は 133.66, エラスティックネットは 160.38 となった．全結果において，乗車人数の予測と比べるとあまり良好な結果とはいえない．図 6 と図 7 を見ると，全体的に大きな外れ値が発生しており，8 分から 10 分程度の誤差も散見される．しかし，図 5 を見ると，他の手法と同様に全体的に外れ値は発生しているものの，誤差 5 分程度に収まっており，最も良好な結果といえる．ランダムフォレストの結果において「赤川三区」から「未来大」までの区間では他の区間と比べて 5 分前後の誤差が多く生じており，予測精度が低くなっている．

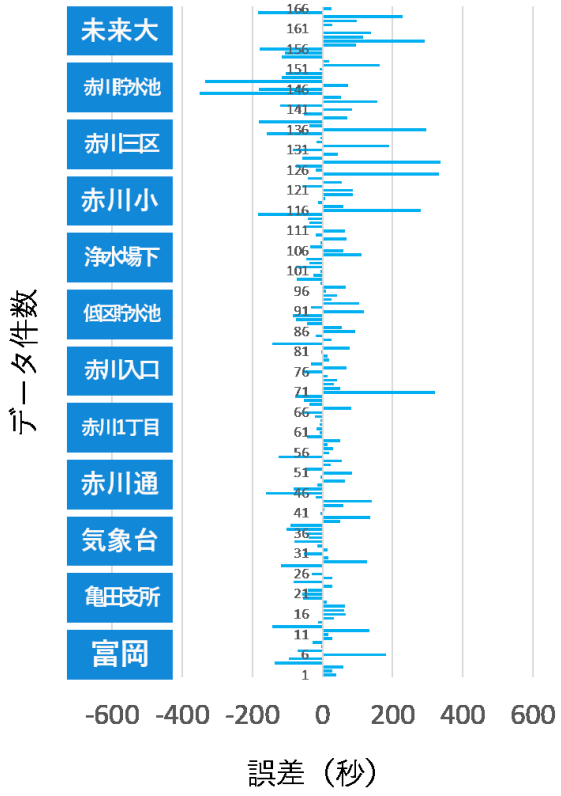

図 5 ランダムフォレストにおける実績値と予測値の差
(到着時刻誤差予測，n=166)

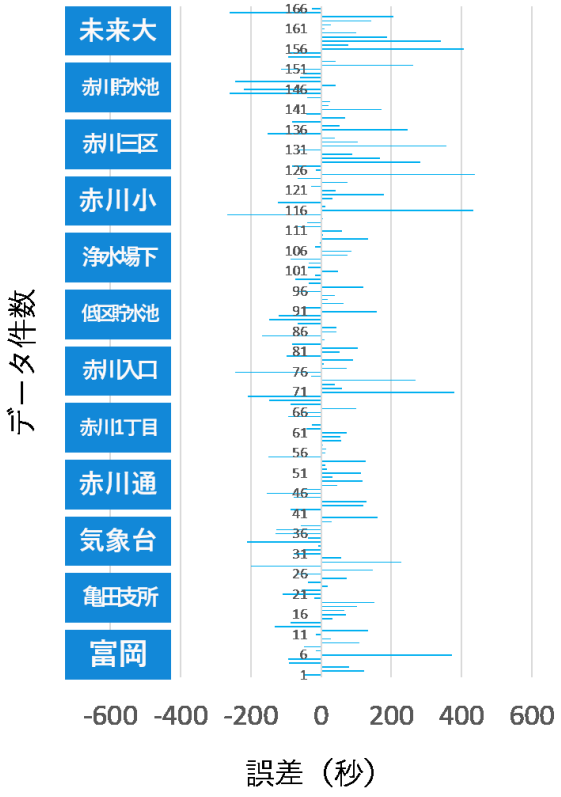

図 6 サポートベクター回帰における実績値と予測値の差
(到着時刻誤差予測，n=166)

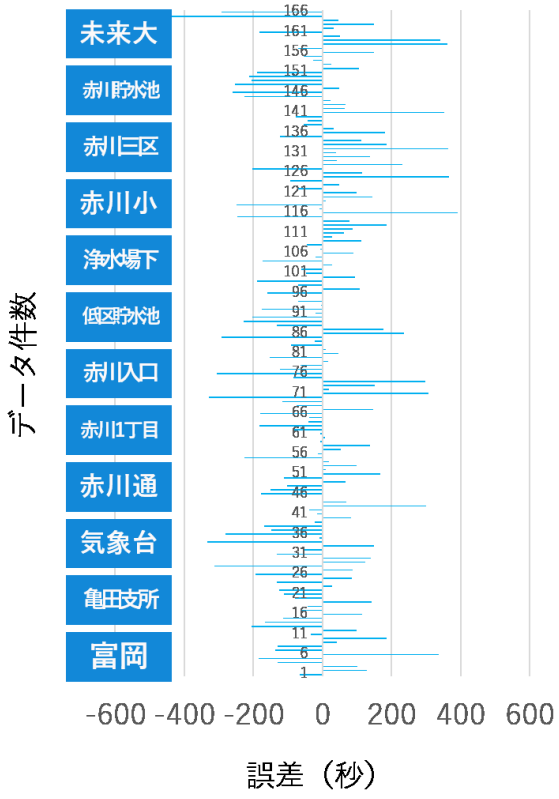

図 7 エラスティックネットにおける実績値と予測値の差
(到着時刻誤差予測，n=166)

外れ値が生じた要因として季節性の要因が考えられる．「赤川三区」から「未来大」の区間では，予測に利用した 11 月の運行記録ではではあまり遅れが生じていない区間だったが，予測対象の 12 月の記録では慢性的に大きな遅れが生じていた．2016 年の降雪は 12 月中旬より増えており，12 月の降雪による環境の変化に対応できず，精度が悪化した可能性がある．乗車人数の予測より RMSE が大きい要因としてデータ件数が少ないことが挙げられる．検証可能なデータ数が少ないことから降雪などによるバスの遅延をモデルが学習できず，外れ値が生じた可能性がある．これらのことから，利用可能なデータ数を増やし，予測対象と同時期のデータを充実させることで精度が向上する可能性が示唆された．

4.4 予備実験のまとめ

ランダムフォレスト，サポートベクター回帰，エラスティックネットの 3 手法の中で最も予測手法に適している手法を実験により調査した．RMSE の値は乗車人数予測と到着時刻誤差予測の両方においてランダムフォレストが最も小さくなり，到着時刻誤差予測においては他の手法より 100 以上小さかった．このことから，翌日におけるバスの予測にランダムフォレストが適している可能性が示唆された．しかし，12 月のみという利用可能なデータが少ない検証であったため，現状では夏季など他の季節・条件では利用できない予測モデルである可能性もある．他の季節への対応や予測精度改善のためには教師データ数を増やすことは勿論だが，学生の定期試験などのスケジュール情報や，時間帯などを考慮する必要が示唆された．

5. 移動方法提案システムの開発

5.1.1 システム概要

開発した予測モデルを元に，翌日の最適な移動方法を提案する Web システム（以下，提案システムと記す）を開発している．提案システムイメージを図 8 に示す．ターゲットユーザは普段からバスを利用する学生とする．①で学生は講義の時間や乗車する停留所を選択すると，②でサーバーサイドが予測モデルを用いた混雑率と到着時刻の計算が行われる．そして③の様に明日の移動に最適なバスの到着時刻と混雑率予測情報を受け取ることができる．スマートフォンでの画面 UI の例を図 9 に示す．検索画面やその結果画面ではシンプルな画面構成を意識し，操作性を重視した．結果画面では，バス情報の他に天気予報も表示し，運行予測以外の情報も確認できるようにすることで利用者のバス利用判断の一助にする．混雑率は選択した停留所に到達するまでの停留所の乗車人数を加算することで算出し，到着時刻は各停留所で予測した到着時刻誤差を運行ダイヤの時刻に加算することで算出する．

5.1.2 システム評価

提案システムの評価には，ユーザテストとアンケートによる定性的なシステム評価と提案システムの予測結果と実績値の誤差を比べる予測精度の定量的評価を予定している．

ユーザテストでは，普段からバスを利用する学生数名に数日間提案システムを日常生活の中で利用してもらう．利用後にアンケートを実施し，使いやすさ，わかりやすさ，掲示される情報の妥当性などについて 4 段階で評価してもらう．自由記述も用意し，提案システムに対する感想を広く収集する．予測精度評価では，提案システムが行った予測結果と実際のバス運行記録の誤差を取り，各便の誤差の大小によって精度を評価する．

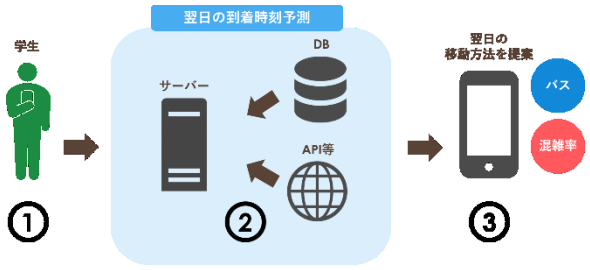

図 8 提案システムイメージ

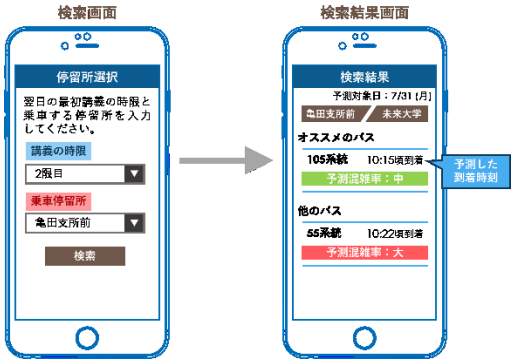

図 9 スマートフォンでの画面 UI の例

6. おわりに

地方圏では公共バスの減便によって低下している利便性を向上させるため，バスロケーションシステムを用いて到着時刻情報を提供している．しかし，予測結果に誤りが含まれる場合があり，かえって利便性を下げてしまっている場合がある．本研究では，普段バスを利用する人々の目的地への移動方法選択を支援し，日常活動を円滑化させることを目的に，複数の学習手法から妥当性の高い手法を検討し，機械学習を用いて翌日におけるバスの到着時刻予測モデルと混雑予測モデルを開発した．結果として，ランダムフォレストが翌日のバスを予測する手法として適している可能性が示唆された．しかし，利用可能なデータ件数が少ない検証であったため，現状では夏季など他の季節・条件に対応できない可能性もある．教師データ数を増やし，精度向上のため，学生の定期試験などのスケジュール情報や，時間帯などを考慮する必要も示唆された．

今後の展望としては，提案システム開発を中心に進め，ユーザテストとシステムの予測精度評価を行う．得られたフィードバックから UI などのシステム改善を進めるとともに，予測精度評価結果から予測モデルの改善をすすめていく．

謝辞 本研究を進めるにあたり，予測モデル開発に用いた乗降者数データや運行実績データは函館バス株式会社の協力によるものである．ここに深く感謝の意を表する．

参考文献

[1] “地域公共交通の現状”. http://wwwtb.mlit.go.jp/kinki/kansai/program/02.pdf, (参照 2018-11-06).
[2] “函館バスロケーション”. https://hakobus.bus-navigation.jp/wgsys/wgp/search.htm, (参照 2018-11-06).
[3] 佐藤孝大，大場みち子．運行実績データに基づくバス到着時刻予測モデルの開発．第 79 回全国大会講演論文集. 2017, vol. 2017, no. 1, pp. 409-410.
[4] 辰巳浩，大野雄作．バスプローブデータを用いた路線バスの予想所要時間に関する基礎的研究．都市政策研究. 2010, no.9, pp. 79-86.
[5] 前川裕一，中島秀之，白石陽．乗降者数データと運行実績データを用いたバス到着時刻予測手法の提案．第 76 回全国大会講演論文集. 2014, vol. 2014, no. 1, pp. 157-158.
[6] L. Breiman. Random Forests. Machine Learning. 2001, vol. 45, no. 1, pp. 5–32.
[7] “What is R?”. https://www.r-project.org/about.html, (参照 2018-11-06).
[8] “Package ‘randomForest’ “. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf, (参照 2018-11-06).
[9] “RMSE (Root Mean Squared Error)”. https://crowdsolving.jp-/node/1130, (参照 2018-11-06).