実績走行データを用いたランダムフォレストによる列車の遅延予測

○髙橋 司　　福田 卓海　　高橋 聖（日本大学）

中村 英夫（東京大学）

Prediction of Train Delays using Random Forest with Actual Running Data

○Tsukasa Takahashi,　Takumi Fukuda,　Sei Takahashi, (Nihon University)

Hideo Nakamura, (The University of Tokyo)

Trains in the metropolitan area have high congestion rates during rush hours. Congestion causes delays, and there is a lot of research and countermeasures to mitigate the delays. In order to evaluate the effect of the countermeasures, we need to evaluate the delays before and after the countermeasure. When the evaluation is done by simulation, it is necessary to predict the delays in response to changes in driving conditions. In this study, we use random forest, a method of machine learning, to predict delays at stations using actual train running data.

キーワード：遅延解消／改善，列車遅延，運行管理，機械学習，ランダムフォレスト
Key Words：Delay resolution/improvement, Train delay, Operations management, Machine learning, Random forest

１．はじめに

日本の首都圏における鉄道網は，都市インフラとしての重要な機関であり，通勤及び通学時の移動手段として多く利用されている．中でも，平日の通勤ラッシュ時には多くの人々が殺到し，世界に類を見ない混雑率で運行している．

このような混雑が引き起こす問題として遅延がある．特に通勤ラッシュでは，輸送量を重視した列車間隔の狭いダイヤで運行しているため，同一路線を走行する後続列車に遅延が伝搬しやすい．また，遅延は他路線にまで波及する特性を持っているため，遅延規模が拡大しやすく定時運行が困難である．そうした遅延を緩和するため，ミクロシミュレータを用いた列車の走行制御に関する研究[1)2)]や，列車運行の可視化による遅延の分析に関する研究[3)4)]が盛んに行われている．

その一つに，朝の混雑時間帯における遅延緩和対策の評価を目的とする，実際の運転を模擬した列車運行シミュレータが開発された[5)]．このシミュレータは走行区間のパラメータに応じて列車運行を詳細にシミュレートすることができる．そのため，路線内設備の改良による遅延への影響を検討することができる．しかし，駅構内で発生する遅延は実績データに基づく既知のデータを使用する必要がある．つまり，路線内設備を改良したことによる駅で発生する遅延までは分析することができず，あくまでも列車の走行をシミュレートするシミュレータであると言える．

他方で，各事業者は列車運行時の走行データを詳細に記録している．日本の首都圏において１日に走行する列車はおよそ 3000 本であり，それらの走行データは膨大な量になる．近年，このような膨大なデータに対して，ルールやパターンを発見することができる機械学習が注目されている．中でも機械学習の一手法である深層学習では，人間の脳を再現したニューラルネットワークを構築することで，複雑な特徴が学習できることから高精度な予測実績が多数報告されており，様々な場面での応用が期待されている．しかし深層学習では，学習した内容を説明することが難しく，予測結果の根拠が薄い．

そこで本研究では，シミュレータにおいて設備改良等の遅延対策効果をより詳細に検討できるよう，膨大な走行実績データをもとに，予測の根拠を分析することができるランダムフォレストと呼ばれる学習機を用いて，駅発生遅延を予測し分析することを目標とする．

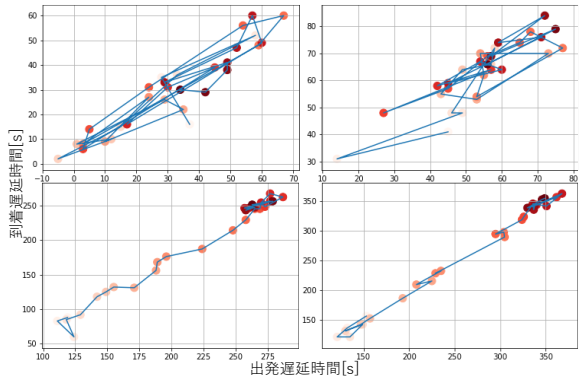

図 1　出発遅延と到着遅延の相関図
（左上から A 駅, D 駅, G 駅, J 駅）

2．実績走行データ

2.1 使用データ

本研究では，実績走行データを使用する．データは，駅名，列車識別番号，運行形態，到着時刻，出発時刻，到着遅延[秒]，出発遅延[秒]で構成されている．

使用した走行実績データは同一路線上にある連続した駅である A 駅（出発駅）～J 駅（到着駅）の 10 駅とし，予測対象駅は同一路線上の連続した D 駅～J 駅の 7 駅とする．通勤ラッシュ時間帯の遅延予測を目的とするため，学習データには A 駅に 7 時 30 分以降に到着する全 30 列車分のデータを使用した．

2.2 データ系列

遅延特性を把握しやすくさせるため，元データから駅間走行時間，駅間発生遅延，駅発生遅延，停車時間，到着間隔を導出する．各系列は以下の式で求められる（式(1)～(7)）．

$$到着遅延 = 実績到着時刻 - 予定到着時刻 \quad \cdots\cdots(1)$$
$$出発遅延 = 実績出発時刻 - 予定出発時刻 \quad \cdots\cdots(2)$$
$$駅間走行時間 = 当該駅到着時刻 - 前駅出発時刻 \quad \cdots(3)$$
$$駅間発生遅延 = 当該駅到着遅延 - 前駅出発遅延 \quad \cdots(4)$$
$$駅発生遅延 = 出発遅延 - 到着遅延 \quad \cdots\cdots\cdots(5)$$
$$停車時間 = 出発時刻 - 到着時刻 \quad \cdots\cdots\cdots(6)$$
$$到着間隔 = 当該列車到着時刻 - 前方列車出発時刻 \quad \cdots(7)$$

以上の計算を行い，データ系列を駅名，列車識別番号，運行形態，到着時刻，出発時刻，到着遅延，出発遅延，駅間走行時間，駅間発生遅延，駅発生遅延，停車時間，到着間隔とした．

3．遅延データの分析

相関分析により実績走行データ内から遅延発生に対して有効な規則性を分析し，規則性が見られた系列より回帰式を導出することで，遅延の予測を行うことを目標とした．

3.1 相関係数による分析

後述する 3.2 ではプロットによる詳細な分析を行うが，実績走行データは系列が多く全系列を分析することが困難であるため，分析対象を絞り込む必要がある．そこで，駅ごとに全ての系列に対し総当りで相関係数を導出した．

相関係数とは，2 系列のデータにおける線形関係の強度を示す指標である．線形関係の強度は-1～1 で示され，この数値が 1 もしくは-1 に近いほど，2 つの系列が強い線形関係にあり，0 に近いほど弱い線形性にあるとされる．相関係数を導出する式を式(8)に示す．

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-x)(y_i-y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad \cdots\cdots\cdots\cdots(8)$$

ここで，xとyはそれぞれの系列とする．

相関分析を行った結果，日ごとに多少のばらつきが見られるが，全ての駅において概ね同様の数値となった．強い相関が見られたのは，出発遅延と到着遅延で 0.713，駅間走行時間と駅間発生遅延で 0.860，駅発生遅延と停車時間で 0.997 となった．当初，到着間隔が広がるほど駅構内及び列車内が混雑し，乗降時間増加による遅延拡大が発生するとされる知見に基づいて，到着間隔と駅発生遅延で相関が見られると予想したが，相関は 0.012 であることから明確な規則性があるとは言えない結果となった．

3.2 プロットによる分析

3.1 の相関分析で高い相関が見られた組み合わせをグラフにプロットした．ここで，データの時間的規則性を分析するために，時間経過とともにプロットの色を濃くし，時間順にプロットを線で結んだ．

強い相関が見られた組み合わせのうち，出発遅延と到着遅延の分析を行う．出発遅延と到着遅延以外の組み合わせにおいても強い相関が見られたが，それらは駅構内で発生する遅延が含まれない．また，駅構内で発生した遅延をそれぞれの系列を導出する際に用いているため，分析の対象外とした．

プロットを行った結果, 時系列の線形性が確認された（図 1 参照）．この時系列の線形性は出発駅側では見られないが，到着駅側では顕著になっている．また，駅停車ごとに遅延が増加している上，後続列車ほど遅延が蓄積している特徴が見られたため, 駅構内で発生した遅延の伝搬影響が，停車回数を重ねるごとに増加していると言える．

このように遅延蓄積の規則性は得られたが，遅延発生に関する明確な規則性が見られなかったことから，遅延は 2 次元で分析できない複雑なメカニズムであることが考えられる．

4．遅延予測

3. の遅延データ分析より，遅延は複雑なメカニズムであり，手作業での分析が困難であることが考えられた．そこで，機械学習により単純な相関関係にない複雑な規則性を学習し，駅発生遅延の予測を行うことを目標とした．

4.1 学習方法

本稿では，機械学習の一手法である，ランダムフォレストを用いて学習を行った．ランダムフォレストは，決定木を用いた学習器であり，予測に至るまでの過程を出力することができるため，根拠のある予測結果を得ることができる．また，本学習機は複数の決定木を組み合わせて予測精

表 1　予測精度

	D 駅	E 駅	F 駅	G 駅	H 駅	I 駅	J 駅
±5 秒誤差[%]	47.7	42.5	39.5	45.6	55.0	60.3	50.4

表 2　重要度分析結果上位 10 系列

系列名（I 駅）	重要度（I 駅）	系列名（F 駅）	重要度（F 駅）
到着遅延	0.729337	到着遅延	0.98963
1 駅前　出発遅延	0.236677	1 列車前　出発遅延	0.000945
1 駅前　到着遅延	0.010059	1 列車前　停車時間	0.000475
2 駅前　出発遅延	0.009675	1 駅前 1 列車前　出発遅延	0.000198
2 駅前　到着遅延	0.006739	1 駅前 3 列車前　出発遅延	0.000185
3 駅前　出発遅延	0.003449	1 駅前　出発遅延	0.000178
3 駅前　到着遅延	0.000343	1 駅前　駅発生遅延	0.000173
1 駅前　駅発生遅延	0.00026	3 駅前　停車時間	0.000168
2 列車　前出発遅延	0.000231	2 駅前　出発遅延	0.000167
1 駅前　停車時間	0.000173	2 駅前　駅発生遅延	0.000159

度向上を行うアンサンブルモデルである．アンサンブルモデルは，個々の決定木のノイズに非常に強く，ハイパーパラメータの調整を行わずに，一定の精度が得られることが期待できる．そして，4.6 で後述する重要度分析を行うことで，各説明変数が予測結果に与える影響度合いを分析することができ，予測に必要な学習データを判別することができる．

4.2 学習条件

3.の遅延データの分析では，駅ごとに遅延特性が異なることが確認された．学習機では複数の規則性を学習することが可能であるが，ランダムフォレストで構築する木の深さや葉の数は有限であり，学習できる規則性の数には限りがあるため，駅ごとに予測モデルを構築する必要がある．

一方，列車走行実績データには学習時にノイズとなるデータが含まれている．例えば，人身事故や設備故障，突発遅延といったものであり，学習データ作成時に，これらのデータを日付単位で除外した．この処理を行ったデータを学習データ 9 割，テストデータ 1 割に分割した．遅延は季節や時間帯によって変動するため，データ分布の偏りを抑えて，満遍なく学習する必要がある．そこで，データを日付順にソートし，9 日分の学習データと 1 日分のテストデータを交互に読み込むことでデータの偏りを緩和した．

ハイパーパラメータのうち，構築する木の数を 100 とした．

4.3 説明変数

3.の遅延データの分析では，駅停車を重ねるごとにプロットの線形性が強くなる特徴と，後続列車ほど遅延が蓄積している特徴が見られた．このような遅延の伝搬特性を学習するため，当該列車，後方 3 駅及び前方 5 列車の各系列を説明変数に設定した．ただし，目的変数を出発遅延とするため，出発遅延を用いて導出した系列を除外する必要がある．したがって，当該駅に停車した予測対象列車の到着遅延以降の系列（駅発生遅延，停車時間）を除外した．また，学習機では数値以外のデータを学習できないことから，駅名，列車番号，運行形態，到着時刻，出発時刻の系列を除外した．

4.4 評価方法

一般的に，機械学習の評価手法では Mean Absolute Error などの統計指標や，正解データと予測データの完全一致割合を用いることが多い．統計指標を用いた評価方法では，正解値と予測値の分散などが正確に把握しやすい反面，予測誤差を実数値として把握することが困難である．また，完全一致割合を用いた評価方法では，扉の二重開閉や，天候による遅延の変動，運転士の個人差による出発までの待ち時間のような，乗客の乗降時間以外に生じる遅延の影響を受けた際に正確な評価を行うことができない．そこで，本研究では独自の評価指標(accuracy)を定義する(式(9))．

$$accuracy = \frac{予測誤差5\ 秒以内のデータ数}{データセット数} \times 100[\%] \quad \cdots\cdots (9)$$

定義した評価指標では，予測誤差を 5 秒まで許容しており，前述した遅延発生による分散を考慮することができる．

4.5 学習結果

予測結果より，H 駅，I 駅及び J 駅において，他駅よりも高い予測精度となった（表 1 参照）．その要因として，駅ごとに走行形態が異なる点が挙げられる．例えば G 駅～H 駅間は低速走行区間であるが，遅延緩和のために高速で走行し，駅間で遅延が緩和されている可能性がある．そこで，駅間発生遅延の各秒数のデータ数をヒストグラム化すると，G 駅まで増加傾向にあった駅間発生遅延が，H 駅では減少傾向になり，F 駅，G 駅ではほぼ遅延が変動しない結果となった．この結果により，H 駅～J 駅では遅延の変動が小さくなり，遅延発生の規則性が明確となり，高精度な予測が行えたと考えられる．

4.6 重要度分析

4.5 で行った学習では，I 駅において約 60%の学習精度が実現されたが，精度が 50%以下の駅もあり，精度の向上を行う必要がある．そこで，精度向上の 1 手法である重要

表 3　重要度分析結果を用いた説明変数削減

	D 駅	E 駅	F 駅	G 駅	H 駅	I 駅	J 駅
上位 5 変数[%]	47.1	42.4	32.3	40.4	55.6	58.1	45.7
上位 10 変数[%]	47.8	42.7	37.3	44.7	55.5	60.3	50.7
上位 20 変数[%]	49.3	43.3	36.3	47.9	56.1	61.2	52.5
上位 30 変数[%]	50.7	44.1	38.5	45.4	56.4	61.6	51.8
上位 40 変数[%]	49.9	42.2	38.2	46	54.8	62.7	50.4
上位 50 変数[%]	49.7	42.7	38.2	46.1	56.4	60.8	52.9

度分析を行う．重要度分析とは，データセット中の 1 系列の値を改変し，改変前後の予測精度を比較することで，改変した系列が予測精度に与える影響の大きさとなり，その値に準じて重要度を決定する分析手法である．機械学習では一般的に，説明変数が多いほど推論過程が細分化され，汎化しやすくなる．しかし，入力した説明変数は全て推論に必要であるとは限らず，こうしたデータが学習ノイズとなり，予測精度低下につながる可能性もある．また，計算時間は説明変数の数に比例するため，説明変数の削減はシミュレータ実装時の計算量削減にもつながる．

最も予測精度の高かった I 駅の重要度分析を行った結果，最大重要度は到着遅延となった（表 2 参照）．到着遅延は予測対象の出発遅延の直前に計測した系列であるため，他の系列に比べ，出発遅延との差が小さい可能性が高く，出発遅延の予測に大きな影響を与えていることが考えられる．また，2 番目～7 番目に重要度が高かった系列は，予測対象列車で 3 駅前までの間に発生した遅延であったことから，駅構内で発生する遅延は，過去に到着，出発した際の遅延が蓄積する規則性があると言える．

最も予測精度の低かった F 駅の重要度分析を行った結果，I 駅と同様に到着遅延が最大重要度となった（表 2 参照）．しかし，F 駅の到着遅延が占める重要度は全体の約 99%であるため，その他の系列が予測に大きく関与していないと言える．また，I 駅では予測対象列車の出発遅延と到着遅延が発生順に重要とみなされたが，F 駅ではそのような特徴は見られなかったことから，駅間において特殊な走行をしていたことが考えられる．

4.7 重要度が高い系列を用いた学習

4.6 で行った重要度分析結果の上位 10～50 変数を取り出し，採用した説明変数量に対する予測精度の変化を検証する．また，上位 10 変数以下の予測精度を検証するため，上位 5 変数の予測精度も検証する．

予測結果より，F 駅以外の駅において説明変数を削減することで，予測精度が向上することを確認した（表 3 参照）．上位 5 変数を採用した場合においては予測精度の低下が見られたが，上位 10 変数以上を採用した場合は予測精度が向上していることから，上位 10 変数未満では特徴量が足りず，全説明変数を入力すると，その多くが学習ノイズになっていたと言える．

F 駅では，先述した改善は見られず，説明変数を増やすほど精度が向上する結果となった（表 3 参照）．その要因として，F 駅では到着遅延が 99%の重要度を締めているが，その他の説明変数は説明変数ごとの重要度がほぼ変わらないためだと考えられる（表 2 参照）．

5．おわりに

本稿では，列車走行シミュレータにおいて，条件変更時のシミュレーション性能向上を目的とした，駅構内で発生する遅延の予測を行った．機械学習の一手法であるランダムフォレストを用いて学習を行った結果，最高予測精度は I 駅の 60.3%となった．さらに，重要度分析を行い，説明変数の削減を行うことで，I 駅において最高予測精度 62.7%を得られた．

列車運行における遅延は，運転士の個人差や天候など，乗客の乗降時間に起因しない要因もある．本研究で想定する朝ラッシュ時間帯ではこれらの遅延が頻発することが確認されたため，今後は学習ノイズとなりうるデータの抽出及び前処理方法の検討を行った後，より精度の高い学習機の検討を行っていく．

参　考　文　献

1) 福田卓海，渋谷明矢，高橋 聖，中村英夫：列車運行実績データに基づいた移動閉そくの遅延回復効果の検討，日本信頼性学会第 31 回秋季信頼性シンポジウム発表報文集，pp.99-102，2018

2) 福田卓海，渋谷明矢，高橋聖，中村英夫，是澤正人，米元和重，足立茂章：列車運行実績データに基づく移動閉そく導入効果の一検討，平成 30 年電気学会全国大会論文集，2018

3) 山村明義，足立茂章，牛田貢平，富井規雄：首都圏稠密運転路線における遅延改善策の検証，J-Rail2012-第 19 回鉄道技術連合シンポジウム，2012

4) 増間義樹，富井規雄，落合康文：軌道回路占有情報の可視化による詳細な列車運行状況の把握，平成 26 年電気学会全国大会，2014

5) 福田卓海，高橋聖，中村英夫，是澤正人，米元和重：列車遅延緩和を目的とした走行制御に対する一検討，平成 29 年電気学会全国大会論文集，2017