# ランダムフォレストによる排水ポンプ稼働時間のパターン分類と回帰分析

## Pattern classification and regression analysis of drainage pump operation time using Random Forests

○長野 峻介[*1]，近藤 浩一郎[*1]，藤原 洋一[*1]，田中健二[*2]，高瀬 恵次[*1]，一恩 英二[*1]

○CHONO Shunsuke[*1], KONDO Koichiro[*1], FUJIHARA Yoichi[*1], TANAKA Kenji[*2], TAKASE Keiji[*1], and ICHION Eiji[*1]

### 1. はじめに

石川県小松市と加賀市の今江潟，柴山潟，木場潟は“加賀三湖”と呼ばれ，国営加賀三湖干拓事業（昭和 27 年～昭和 44 年）により今江潟の全面と柴山潟の 2/3 の面積は干拓され圃場整備がなされてきた．現在，加賀三湖干拓地および周辺地区では，集中管理施設によって用排水路の水位状況などを把握し一元的に灌漑排水管理を実施している．ただし，豪雨時などにおいて排水施設の操作管理を適切に行うために，管理者は管轄する地区全体の状況を的確に判断する必要がある．そこで本研究では，加賀三湖干拓地の灌漑期における排水管理の意思決定を支援するため，灌漑期における排水ポンプの操作と降水量，ゲートの操作，用排水路の水位データ等との関係を考察し，近年様々な分野[1]の研究で用いられている機械学習手法の 1 種であるランダムフォレスト[2]を適用することによって，管理者の状況判断と排水機操作のモデル化を行った．

### 2. 研究対象地

加賀三湖干拓地（図 1）における主に柴山潟の干拓地からの排水を担う柴山潟排水機場における排水操作をモデル化の対象とする．柴山潟排水機場には排水量の異なる 4 台のポンプが設置されており，同型の 1，2 号ポンプは排水量が小さく自動化され交互に稼働しており，排水量が大きい 3，4 号ポンプは降水量が多い場合などに稼働している．

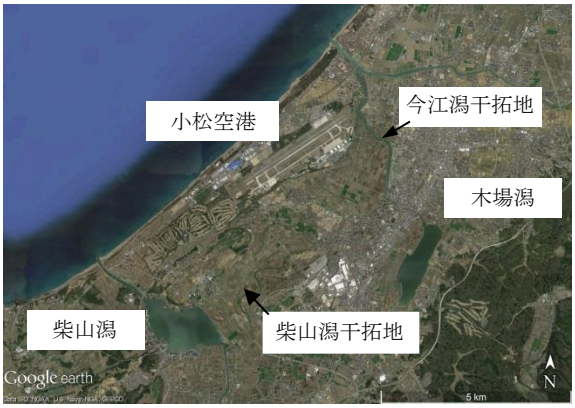


図 1　加賀三湖干拓地

### 3. 排水機操作へのランダムフォレストの適用

本研究では，ランダムフォレストを用いて排水機の稼働時間をパターン分類および回帰分析することにより解析を行った．ランダムフォレストは機械学習の手法の中で教師あり学習に分類され，ランダムな複数の決定木を組み合わせるアンサンブル手法である．その学習方式は計算速度が速く，外れ値やノイズに関して相対的に頑健であるとされる[3]．個々の決定木の成長には平均二乗誤差の条件が用いられ，予測される識別パターンはすべての決定木の予測から多数決により出力され，回帰分析では複数の決定木の予測結果から平均が出力される．また，各特徴がノード分割に使われた時の不純度（ジニ係数）の減少量を森全体で平均した量を用いて，各特徴の重要さ（変数重要度）を評価することができる．

加賀三湖干拓地の柴山潟排水機場にある 4 台のポンプ（1～4 号機）について，ポンプ稼働時間（分/1 時間）を識別対象および目的変数とし

*1 石川県立大学生物資源環境学部 Faculty of Bioresources and Environmental Sciences, Ishikawa Prefectural University
*2 国立研究開発法人土木研究所 寒地土木研究所 Civil Engineering Research Institute for Cold Region, Public Works Research Institute
キーワード：排水管理，排水施設，機械学習

て，各ポンプの稼働時間の予測を行った（1，2 号機は交互稼働しており，2 台合計の稼働時間）．なお，加賀三湖干拓地における降水量や各地点の水位データ，ゲート，ポンプなどの操作データを特徴ベクトル，説明変数として用いて，2014～2017 年の灌漑期間 4 月 1 日から 9 月 30 日における毎時データを訓練データとして使用した．精度評価には正答率，決定係数，OOB スコア，平均 2 乗誤差（MSE）を用いた．OOB スコアは未知のデータに対する予測精度を評価するものであり，パターン分類では正答率，回帰分析では決定係数で算出される．

## 4. 結果

ポンプ稼働時間のランダムフォレストによるパターン分類の解析結果を表 1 にまとめ，回帰分析の解析結果を表 2 にまとめる．パターン分類の解析結果では，1, 2 号機では正答率は高いが OOB スコアが低く，3, 4 号機の正答率および OOB スコアの値は高くなった．回帰分析の結果では，すべての決定係数は高く，1, 2 号機と 3 号機の OOB スコアは高くなったが，4 号機では若干低くなった．

表 1　パターン分類による解析結果

| | 正答率 | 正答率 (OOB) | MSE (min) | $R^2$ |
|---|---|---|---|---|
| 1, 2 号 | 0.942 | 0.177 | 12.988 | 0.977 |
| 3 号 | 0.999 | 0.986 | 1.105 | 0.981 |
| 4 号 | 1.000 | 0.998 | 0.054 | 0.984 |

表 2　回帰分析による解析結果

| | $R^2$ | $R^2$ (OOB) | MSE (min) |
|---|---|---|---|
| 1, 2 号 | 0.986 | 0.877 | 7.879 |
| 3 号 | 0.985 | 0.884 | 0.883 |
| 4 号 | 0.935 | 0.566 | 0.218 |

1, 2 号機では 2 台合計の稼働時間のパターン分類を行ったが，識別パターン増えたことで精緻に行うことは難しかった．3 号機および 4 号機では，稼働時のみを抽出した解析結果では，パターン分類の解析結果で予測値が実測値から大きく外れる場合が多く見られた．これらのことから今回の解析では，パターン分類よりも回帰分析のほうが適していると考えられる．

変数重要度については，1，2 号機では，主にポンプ稼働時間と柴山潟排水機場内水位および柴山潟排水機場内水位変化の変数重要度が大きかった．3 号機では 3 号機のポンプ稼働時間と降水量，4 号機では 4 号機のポンプ稼働時間と降水量の変数重要度が大きくなった．1, 2 号機は内水位を基準に交互稼働しており，変数重要度はこの運用操作の学習を示していると考えられる．また，3, 4 号機は排水能力が高く降水量が多い場合に稼働する機会が多く，柴山潟排水機場内水位や降水量の変数重要度が大きくなっていると考えられる．3, 4 号機のモデルは管理者の操作をほぼ再現できたといえるが，ポンプが稼働する機会が少なくより多くの訓練データが必要である．

## 5. おわりに

本研究では加賀三湖干拓地の灌漑期における排水管理にランダムフォレストを適用し，排水機場のポンプ稼働時間をパターン分類および回帰分析するモデル化を行った．稼働時における解析結果から，ランダムフォレストの回帰分析の有用性が確かめられた．今後の課題として，より安全な排水管理に向けて，訓練データを増やし学習精度を上げることや，他の機械学習手法の適用結果との比較などが挙げられる．

### 参考文献

[1]馬場真哉・松石隆(2015): ランダムフォレストを用いたサンマ来遊量の予測，日本水産学会誌，81(1)，pp.2-9. [2] Breiman, L. (2001) : Random Forests, Machine Learning, 45(1), pp.5-32. [3] 杉本知之・下川敏雄・後藤昌司(2007): 樹木構造接近法と最近の発展，計算機統計学，18(2), pp.123-164.