分担研究報告書４

衛星データと機械学習アルゴリズムを活用したダム湖の藻類異常発生予測モデルの構築

研究分担者　　西村　修
研究協力者　　佐野　大輔
研究協力者　　三浦　耀平

厚生労働科学研究費補助金 (健康安全・危機管理対策総合研究事業)
気候変動に伴う水道システムの生物障害等リスク評価とその適応性の強化に向けた研究
分担研究報告書

研究課題：衛星データと機械学習アルゴリズムを活用したダム湖の藻類異常発生予測モデルの構築

研究分担者　西村　修　　東北大学大学院　工学研究科　教授
研究協力者　佐野　大輔　東北大学大学院　工学研究科　教授
研究協力者　三浦　耀平　東北大学大学院　工学研究科　大学院生

研究要旨
水道水源で異常発生する浮遊性藻類が原因で浄水場において生物障害が確認されている。短期間先の障害生物の発生を予測することができれば，ダム湖の水質担当者や浄水場の技術者による事前対応が可能となり，安心・安全な水道供給の確保につながるだろう。藻類異常発生の既往の検知・予測技術の開発において，様々な種類の地球観測衛星により取得された情報が使用されてきたが，その技術開発の多くは海域や大規模な湖沼を対象とした研究であり，日本のダム湖のような比較的小規模な水域を対象とした生物発生予測モデルの構築は進んでいない。
本研究では日本国内の 4 つのダム湖を対象とし，代表的な浮遊藻類である *Dolichospermum* spp.の異常発生予測モデルを構築した。各ダム湖で定期的に計測された栄養塩データ，水理データ，最寄りのアメダス観測所の気象データ，及び空間解像度・回帰日数が異なる 2 つの衛星から取得されたデータを説明変数とし，automatic relevance determination により対象藻類の異常発生に関連のある要因の特定を行なった。各ダム湖で関連があると推定された変数を用いて，複数の機械学習アルゴリズムによる予測モデルの構築及び比較を行なった。室生ダムの 2 値分類における最良モデルは，正解率 91.7%，精度 100%，再現率 83.3%であった一方で，回帰モデルでは精度の良いモデルを構築することができなかった。ARD により選択された衛星データの種類から，小規模水域における藻類異常発生予測において，衛星センサーの空間解像度より衛星の回帰日数（時間解像度）が重要であることが示された。引き続き，様々な種類の衛星データの活用方法を探る予定である。

A. 研究目的

水道水源において藻類が異常発生する現象は，その水域から取水する浄水場の処理能力の低下を引き起こす。実際に日本全国の浄水施設でろ過閉塞障害，凝集沈澱処理障害等の生物障害が確認されている [1)]。水道利用者に対して安定した水道供給を実現するためには，藻類異常発生への早急な対応が求められる。早期警報システムとしての短期間先の藻類発生予測技術は有望な解決策の一つである。ダム湖において藻類を含む植物プランクトンを予測する手法としては，これまで数多くのモデルが提案されており，短期間先の藻類発生予測を行うため多様な機械学習アルゴリズムが適用されてきた [2)]。藻類の発生には様々な要因が影響し，その関係は複雑であることが知られており，これまでの藻類発生予測モデル構築においては，水質要因，水理的要因，気象的要因等が考慮されてきた。一方で，これらのデータを高頻度で取得することは人的・金銭的資源の制約から全ての水道水源で実施することは現実的ではない。そこで，近年藻類異常発生の検出及び予測において衛星データが大きな注目を集めている。地球観測衛星が定期的に取得する地表面の光や電波の反射の情報が大いに活用されている。各々の衛星は，観測する光の波長，空間解像度，回帰日数等の観点から異なる特徴を有している。藻類の異常発生を検出する際には，関心のある狭い水域のみの情報を取得するため，高い空間解像度の衛星センサーにより得られたデータが頻繁に使われてきた。一方で，藻類異常発生の予測には，回帰日数が短い，即ち，時間解像度の高い衛星が用いられてきた。時間解像度の高い衛星は一般的に空間解像度が劣る。そこで，時間解像度の高い衛星データが適用されてきたのは海域や広大な湖沼等の広い水域のみであった。異なる空間・時間解像度を有する衛星データを用いた比較的小規模な水域における藻類異常発生予測モデルの構築が進んでいないのが現状である。日本では，入り組んだ地形で比較的小規模なダム湖が多く存在しており，これらの水域における藻類発生予測モデルの構築が求められている。

本研究では，日本国内の 4 つのダム湖を対象とし，*Dolichospermum* spp.の異常発生を予測するモデルを構築した。藻類濃度を目的変数とし，栄養塩データ，水理データ，気象データ，衛星データを説明変数とし，automatic relevance determination（ARD）による関連のある説明変数の特定を行なった。関連があると推定された変数のみを用いて，5 つの機械学習アルゴリズムによる藻類発生・非発生の 2 値分類モデル及び回帰モデルを構築した。

B. 研究方法

本研究の全体の流れは，データ収集，変数選択，及び予測モデル構築の 3 段階に大きく分けられる。奈良県宇陀市の室生ダム（室生湖），岐阜県恵那市の阿木川ダム（阿木川湖），兵庫県川西市の一庫ダム（知明湖），福岡県朝倉氏の寺内ダム（美奈宜湖）の各 3〜5 地点（図 1，S1〜S14）において観測さ

れた，2013 年から 2017 年までの *Dolichospermum* spp.の濃度を目的変数として用いた。本研究では 7 日後の藻類異常発生予測を想定している。説明変数として，月 1 回の頻度で測定された全窒素濃度及び全リン濃度の過去 2 ヶ月間の平均値（TN2m, TP2m），ダム湖水の流入量及び流出量の過去 7 日間の平均値（Inflow7d，Discharge7d），最高気温及び風速の過去 7 日間の平均値（AveMaxTemp7d, AveWind7d），日射量及び降水量の過去 7 日間の合計値（Sun7d，Rain7d）を用いた。衛星データとして，Landsat-8 により測定された地表面温度及び 2 つの濁度指標の 31 日前から 1 週間前までの平均値（L8_LST，NDTI，red660nm），MODIS センサーによる地表面温度の日別データ及び 8 日間のコンポジットデータの最大 39 日前から 1 週間前までの平均値（MOD_LST_daily，MOD_LST_8-day composite）を使用した。コンポジットデータとは，雲の影響を最小限にするために，ある一定期間の中で最も品質の良いデータをその期間の代表値としたデータである。衛星データの取得には Google Earth Engine を使用した。

変数選択の段階においては，スパース推定の一つの ARD を用いて，各説明変数の目的変数に対する関連度（係数）を算出した。係数が 0 となった変数をデータセットから除外し，残りの変数を用いて，藻類異常発生予測の 2 値分類モデル及び回帰モデルを構築した。機械学習アルゴリズムとして，サポートベクターマシン（SVM），人工ニューラルネットワーク（ANN），ランダムフォレスト（RF），extreme gradient boosting （XGBoost），light gradient boosting machine （LightGBM）の 5 つのアルゴリズムをデータセットに適用し，2 値分類モデル及び回帰モデルの性能を比較した。2 値分類モデル構築の際には，*Dolichospermum* spp.の異常発生の濃度の閾値を 100 cells/ml と定義し，これらの値以上である場合に 1（発生），下回った場合に 0（日発生）とした。2 値分類のモデル性能の評価には，テストデータに対する正解率（正解数／データ数），精度（発生予測における正解数／発生予測のデータ数），再現率（発生予測における正解数／実際の正解が発生のデータ数）を用いた。回帰モデルに関しては，訓練データとテストデータの正解値と予測値の平均二乗誤差（MSE）を評価指標として採用した。Python 3.8.8 の scikit-learn ライブラリーを用いて ARD 及び機械学習アルゴリズムをデータセットに適用した。

C. 研究結果及び D. 考察

表 1 に ARD による説明変数の回帰係数の推定結果を示す。全窒素濃度，風速，日射量，降水量，MODIS の地表面温度が 3 つのダム湖で *Dolichospermum* spp.の異常発生に関連のある変数であると推定された。一方で，Landsat-8 の濁度指標の一つ（red660nm）は全てのダム湖で関連がないと推定された。

SVM，ANN，RF，XGBoost，LightGBM の 5 つの機械学習アルゴリズムにより構築した 2 値分類モデルの性能を表 2 に示す。各ダム湖のデータ数は，室生ダム 57，阿木川ダム 11，一庫ダム 23，寺内ダム 41 であった。阿木川ダム及び一庫ダムのデータを訓練データとテストデータに分割した際に十分にモデルを構築・評価可能なデータ量を確保できなかったため，これら 2 つのダム湖の結果は参考程度にされたい。室生ダムは，4 つのダム湖の中で最もデータ数が多く十分に学習をすることができたため，最も良いモデル性能となったと考える。5 つのアルゴリズムで同じデータセットを使用して構築した回帰モデルの性能を表 3 に示す。テストデータ MSE の比較から各ダム湖の回帰モデルの最良アルゴリズムは，室生，阿木川，一庫，寺内の順番に，ANN，ANN，SVM，SVM であった。藻類濃度を予測する回帰モデルに関しては，ANN と SVM の 2 つのアルゴリズムが有力な候補となることが示された。

室生ダムに関して正解率 91.7%のモデルを構築できたが，他のダム湖についてはデータ数の不足から十分な性能のモデルを構築できたとは言い難い。Landsat-8 のセンサーは空間解像度が 30m であり，MODIS センサーは 1km である。対象ダム湖の形状から 1km 四方の解像度では，対象水域以外の森林や道路等の要素が含まれるため，Landsat-8 の高解像度データも使用した。しかし，観測頻度の低い Landsat-8 のデータを含めたことで，使用できるデータ数が大幅に減少してしまった。十分な精度の予測モデル構築には，高解像度の衛星データを使用することだけでなく，十分な訓練データを確保できるように高いデータ頻度も求められる。本研究では，衛星データを含めた 2 値分類モデル及び回帰モデルに関して，データが少ないこともあり，有用な予測モデル構築ができたとは言えない。今後は，日本のような比較的小さなダム湖における藻類異常発生を予測するために，高頻度の衛星データ及び高解像度の衛星データの両方を活用できるように，データ処理や扱い方の工夫をしていく必要があるだろう。

E. 結論

日本国内の 4 つのダム湖を対象とし，現場（水質，水理，気象）データ，複数の衛星から取得されたデータを用いて，ARD による関連のある変数選択及び機械学習アルゴリズムによる藻類異常発生の予測モデルを構築した。衛星データが藻類発生予測に有用であり，特に時間解像度の高い（観測頻度が高い）衛星データがより重要であることが示された。今後も異なる種類の衛星データによる予測モデル性能の比較を行い，藻類発生予測に有用な衛星データの活用方法の探索を行う。

F. 健康危険情報

該当なし

G. 研究発表

1. 論文発表

該当なし

2.学会発表

三浦耀平，今本博臣，峠嘉哉，浅田安廣，下ケ橋雅樹，秋葉道宏，西村修，佐野大輔．衛星データを活用した水道水源における藻類異常発生予測モデルの開発．第 57 回日本水環境学会年会，2023.3，松山市．

H. 知的財産権の出願・登録状況 (予定も含む。)

1.特許取得

該当なし

2.実用新案登録

該当なし

3.その他

該当なし

I. 参考文献

1) 秋葉道宏，高梨啓和. 水道における異臭味問題の最新動向. J. Japan Association on Odor Environment 49, 101-108 (2018).

2) Rousso BZ, Bertone E, Stewart R, Hamilton DP,: A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes, Water Res., Vol.182, 115959, 2020.

J. 謝辞

本研究を進めるに当たり，水資源機構及び京都大学防災研究所水資源環境研究センター峠嘉哉特定准教授の協力を得ました。記して謝意を表します。

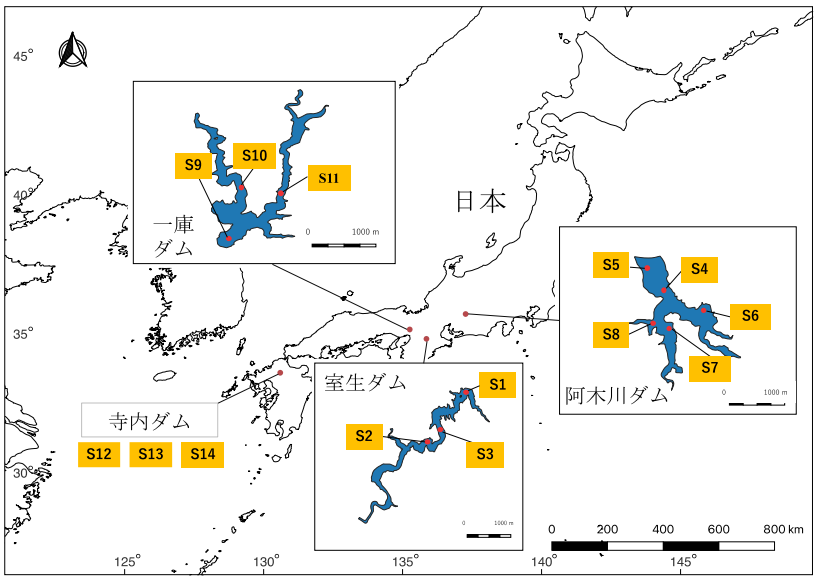

図 1　対象ダム湖及びサンプリング地点

表 1　ARD による説明変数の回帰係数の推定結果

	L8_LST	L8_NDTI	L8_red	MOD_LST_daily	MOD_LST_8-daycomposite	TN2m	TP2m
室生	0	0	0	0	-0.22	0.09	-0.11
阿木川	0	0.06	0	0	-0.15	0.04	0
一庫	0	0	0	0	0.07	0	0
寺内	0.09	-0.002	0	0.21	0	-0.05	0

	Inflow7d	Discharge7d	AveMaxTemp7d	Sun7d	AveWind7d	Rain7d
室生	0	0	0.27	0	-0.03	-0.05
阿木川	0	-0.10	0	-0.14	0	-0.07
一庫	-0.97	0	0	-0.17	0.40	0.50
寺内	0	0	0	0.32	0.26	0

表 2　各ダム湖における機械学習アルゴリズムによる 2 値分類モデルの予測結果

	機械学習アルゴリズム	カーネル	正解率（%）	精度（%）	再現率（%）
室生ダム	SVM	linear	75	100	50
		poly	75	80	66.7
		RBF	66.7	100	33.3
		sigmoid	75	100	50
	ANN	-	91.7	100	83.3
	RF	-	83.3	100	66.7
	XGBoost	-	75	80	66.7
	LightGBM	-	50	0	0
阿木川ダム	SVM	linear	×	×	×
		poly	×	×	×
		RBF	×	×	×
		sigmoid	×	×	×
	ANN	-	×	×	×
	RF	-	×	×	×
	XGBoost	-	×	×	×
	LightGBM	-	×	×	×
一庫ダム	SVM	linear	60	0	0
		poly	80	0	0
		RBF	80	0	0
		sigmoid	80	0	0
	ANN	-	60	0	0
	RF	-	80	0	0
	XGBoost	-	60	0	0
	LightGBM	-	80	0	0
寺内ダム	SVM	linear	55.6	50	25
		poly	44	0	0
		RBF	55.6	50	50
		sigmoid	55.6	50	25
	ANN	-	66.7	66.7	50
	RF	-	55.6	50	25
	XGBoost	-	33.3	0	0
	LightGBM	-	55.6	0	0

表 3　各ダム湖における機械学習アルゴリズムによる 2 値分類モデルの予測結果

	機械学習アルゴリズム	カーネル関数	訓練データMSE	テストデータMSE
室生ダム	SVM	linear	0.287	0.384
		poly	0.228	0.310
		RBF	0.009	0.289
		sigmoid	0.290	0.374
	ANN	-	0.253	0.277
	RF	-	0.049	0.344
	XGBoost	-	0.067	0.694
	LightGBM	-	0.342	0.424
阿木川ダム	SVM	linear	0.030	0.447
		poly	0.044	0.498
		RBF	0.044	0.496
		sigmoid	0.044	0.497
	ANN	-	0	0.316
	RF	-	0.012	0.450
	XGBoost	-	0.038	0.513
	LightGBM	-	0.044	0.487
一庫ダム	SVM	linear	0.492	0.050
		poly	0.576	0.063
		RBF	0.177	0.028
		sigmoid	0.492	0.047
	ANN	-	1.017	0.366
	RF	-	0.055	0.050
	XGBoost	-	0.104	0.136
	LightGBM	-	0.503	0.096
寺内ダム	SVM	linear	0.891	0.572
		poly	0.906	0.608
		RBF	0.847	0.563
		sigmoid	0.892	0.571
	ANN	-	1.065	0.683
	RF	-	0.265	0.803
	XGBoost	-	0.295	0.998
	LightGBM	-	1.08	0.747