Ⅱ．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）
「水道の基盤強化に資する技術の水道システムへの実装に向けた研究」
分担研究報告書

ビッグデータに基づく水質変動の早期予測手法の検討

研究分担者　山村 寛　　中央大学 教授

研究要旨

塩素注入量が多すぎた場合、トリハロメタンの生成につながることから、給配水系統における水道水の残留塩素濃度は 0.5 mg/L 程度になるように厳格に管理されている。現在は、高度な技術を持つ職員が経験に基づいて注入量を決定しているが、大量のベテラン職員の退職と職員数の減少などから、経験に依存しない、新しい塩素注入量管理手法が求められている。本研究では、浄水場が保有する残留塩素濃度の時系列データに着目し、時系列の濃度変化の傾向を学習することで、数時間先の残留塩素濃度を推測できる予測モデルの構築を目的とした。具体的には、長期短期記憶（LSTM）ネットワークモデルにより、3 時間、6 時間、12 時間、24 時間先の残留塩素濃度予測モデルの構築を試みた。続いて、アルゴリズム中のパラメーター類、及び入力データ数が予測精度に及ぼす影響を検討した。

予測に必要な入力項目を検討した結果、水温及び導電率による予測精度への影響は僅かであり、残留塩素濃度だけで信頼に足るモデルの構築が可能となることが明らかになった。モデル精度の向上を目指してタイムステップ及び予測時間を様々に変化させた結果、タイムステップを 24 時間に設定した際に最も精度が高く、誤差目標値±0.025 mg/L 以下に収めるには、予測時間を 6 時間以下にする必要があることが判明した。入力データ数を 10 年から 1 ヶ月にまで減少させながらモデルを構築した結果、最低 4 か月分の 1 時間間隔データがあれば、予測値と実測値の差を目標値以内に収めることができた。

A. 研究目的

日本では、水道法において蛇口における残留塩素濃度を 0.1mg/L 以上に維持することが義務づけられている。塩素消毒は高い消毒効果を長時間にわたって持続できる一方で、浄水処理施設で注入された塩素消毒剤が配水管・配水池ならびに給水管を経由して給水栓や一般家庭等の蛇口に到達する間に、水中のフミン質やアンモニア態窒素、配水管の管路壁面や表面に付着した生物膜などとの化学反応により、徐々に残留塩素濃度が希薄化する。給水栓や蛇口で残留塩素濃度 0.1mg/L を維持するためには、配水中に消費される残留塩素量を勘案した上で、浄水場での塩素注入量を決定する必要がある。

塩素消毒剤と水中の有機物が反応することで、トリハロメタンをはじめとする消毒副生成物が生成される。水質管理目標設定項目では残留塩素濃度が 1 mg/L 以下となるように設定されている他、総トリハロメタン濃度が 0.1 mg/L 以下となるように水質基準項目が定められている。神奈川県内広域水道企業団では、トリハロメタン抑制の観点から残留塩素濃度の管理目標値を 0.5 mg/L に定めており、現状、高度な技術を持つ職員が経験に基づいて適宜、塩素注入量を判断している。

日本は、2008 年に人口がピークに達した後、徐々に人口が減少する人口減少社会に突入した。浄水場の職員数も徐々に減少しており、2030 年までには、2000 年比 30%程度職員が減少すると予測されている。特に、高度な技術を持つベテラン職員の大量退職を控えており、これらの技術と経験の継承が重要な課題となっている。既存施設を持続的に維持・管理していくためにも、職員の技能や経験に依存しない、新しい浄水場の運転管理手法が求められている。そこで、本研究では配管内の残留塩素濃度の低減量を予測するモデルの構築に挑戦する。予測モデルが構築できれば、高度な技術を持った職員の判断を必要とせず、正確かつ迅速な塩素注入量の設定を自律的に最適化できるようになると期待する。

これまで、配管内の残留塩素予測を目的として、様々な物理モデルが構築されている。代表的なものとして、米国 EPA が提供する EPANET が世界中の水道事業者に利用されており、滞留が存在しない配管では比較的正確に残留塩素濃度の予測が可能とされる。一方で、貯水槽や滞留を伴う配管及び二次枝管などについては、既存モデルの適用が難しいことが指摘されている。Abokifa ら[1]は、既存の物理モデルに確率需要発生器を接続することで、水需要の変動による滞留時間の変化を組み込んだ確率モデルを開発したが、既存モデルよりも精度が向上した一方で、塩素注入制御に足る精度には至っていない。

神奈川県内広域水道企業団は、神奈川県内の４事業体（神奈川県営水道、横浜市水道局、川崎市上下水道局、横須賀市上下水道局）へ浄水を給水する特別地方公共団体である。浄水された水は、42 ヶ所の給水地点を経由して、各事業体に供給されており、各給水地点において、基本的な水質項目が連続的に監視されている。よって、神奈川県内広域水道企業団は、浄水場の出口に加え、各給水地点において、基本的な水質に関する連続監視データを保有していることになる。これらの膨大なデータセットを活用することで、給水地点における残留塩素濃度を高精度に予測しうるモデルが構築できるものと期待する。

ビッグデータを利用した制御方法として、ニューラルネットワークを用いたモデル構築が挙げられる。計算機の進化に伴って、ニューラルネットワークの中間層を時系列の前後で接続することで、時系列変化のパターンを学習するリカレントニューラルネットワーク（RNN）が開発され、様々な分野で将来予測に利用されている。Bowden ら[2]は、南オーストラリア、アデレード南部の配水システムを対象として、浄水場出口、ポンプ場、給水地点での残留塩素データを

使用して RNN による 72 時間後の残留塩素レベル予測モデルを構築した結果、R^2=0.96 の精度で予測することに成功している。本モデルは、浄水場出口及び給水地点前段の塩素濃度を把握する必要があるため、連続残留塩素系を多数備えた地域に限定されることが欠点として挙げられる。実用性の高いモデル構築には、さらに簡潔なデータセットによる予測モデルが必要とされる。

近年、特に長期間の予測を目的として、RNN の隠れ層に CEC (constant error carousel) を設置することで、入力ゲート、出力ゲート、忘却ゲートによって、過去から引き継いだデータを必要に応じて取得・修正・消去できる特徴を持つ長期短期記憶（LSTM）ネットワークモデルが開発された。Xuan-Hien Le ら[3]は、LSTM によってベトナムの洪水予測モデル構築に挑戦している。1961 年から 1984 年の 24 年間、上流地点の降雨量と流量を入力値として、下流地点における 3 日後の流量を予測するモデルを構築した結果、95%以上の予測精度が得られている。LSTM により精確な予測結果を得るには、タイムステップと予測時間を適切に設定する必要がある。タイムステップは 1 ユニットの時系列データの範囲を示す値であり、水質変動の周期に合わせて設定する必要がある。水質変動の周期よりタイムステップが短すぎると変化のパターンが単調になり、特徴を捉えられずに、モデル精度の低下を招く。一方で、水質変動の周期よりタイムステップが長すぎる場合、変化のパターンが多すぎるため、全てトレンドを学習するには膨大なデータ量が必要となる。

以上の背景から、本研究では神奈川県内広域水道企業団が保管する浄水場と給水地点における膨大なデータセットを活用して、3〜12 時間先の給水地点における残留塩素濃度を予測しうるモデルを構築する。モデル構築にあたって、時系列データの将来予測に有効なリカレントニューラルネットワークモデルのうち、長期間の予測モデル構築に優れる短長期記憶（LSTM）モデルを利用する。具体的には、データの前処理アルゴリズムを検討すると共に、モデルの各種パラメーター、データ種、データ数が予測精度に及ぼす影響について検討する。

B. 研究方法

モデル構築に利用したデータ

本研究で使用したデータは、水道技術研究センターの協力の下、神奈川県内広域水道企業団より提供頂いた。神奈川県内広域水道企業団は、相模川、酒匂川で取水した水を 6 か所の浄水場で浄水している。本研究では、相模川及び酒匂川の 2 河川を水源とする相模原浄水場を基点として、約 15km に位置する上和田給水地点、及び矢指調整池を経由して約 20km に位置する西谷給水地点における残留塩素予測モデルを構築する。本研究では、他の浄水場の影響を受けない地点として、上記の 2 給水地点を選定した。

相模原浄水場、上和田給水地点、西谷給水地点での 2018 年 7 月 1 日から 2020 年 6 月 30 日の 2 年間分の 30 分間隔の残留塩素濃度(mg / L)、水温(℃)、導電率(mS / m)（以下データセット①と表記する）と、2010 年 4 月 1 日から 2020 年 3 月 31 日の 10 年間分の 1 時間間隔の残留塩素濃度(mg / L)、水温(℃)、

導電率(mS / m)（以下データセット②と表記する）を使用した。データセット①、②ともに、以下の方法で異常値の除去及びノイズを除去したものを学習・検証に使用した。

異常値の除去は、箱ひげ図を用いて、上限＝95%点+四分位範囲（IQR）×3、下限＝95%点-IQR×3 として、極端にトレンドから外れた値を除外した後に、前後の値で線形補完した。

データノイズの除去は、下式を用いて前後 12 時間の 24 時間移動平均を算出することで、平滑化処理を実施した。

$$x_{t-6},\dots,x_{t-1},x_t,x_{t+1},\dots,x_{t+6} = \frac{x_{t-6}+\cdots+x_{t-1}+x_t}{7},\frac{x_{t-5}+\cdots+x_t+x_{t+1}}{7},\dots,\frac{x_t+x_{t+1}+\cdots+x_{t+6}}{7}$$

なお、x_t は、ある時刻 t における残留塩素濃度低減量とする。

モデル構築作業

管路内の滞留時間に時間周期性があると仮定し、滞留時間の周期変動も加味した残留塩素低減量を評価する。本研究では任意時刻における浄水場出口の残留塩素濃度と給水地点の残留塩素濃度の差を「残留塩素低減量」として、LSTM の入出力値に用いた。残留塩素低減量以外の２因子（導電率、水温）は、正規分布として平均 0、分散 1 になるように一般標準化したものを入力値として用いた。

LSTM は以下の式で表現される。

x_L={x_t, x_(t+1), x_(t+2), ⋯, +x_(t+L-1)}

ここで、L をタイムステップ、t を時間、x_t を時間 t における入力値、x_L をタイムステップ時間における入力値とする。バッチサイズは 32、活性化関数は relu 関数、エポック数は 50、ノード数は 1024 とした。学習関数、エポック数、ノード数は複数検討したが、どの組み合わせでもモデル精度の大幅な変化はなかった（データは非表示）。タイムステップとして、12、24、48、168 時間（7 日間）の 4 条件を検討した。

データセットを目的に応じて任意の割合で学習用のトレーニングデータとモデル精度検証用のテストデータに分割した。使用したデータセットを表にまとめる。

モデル精度の評価方法

モデルの精度は、二乗平均平方根誤差（RMSE）及び相関係数 R^2 により評価した。以下に RMSE 及び R^2 値の計算式を示す。

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum(\text{実測値} - \text{予測値})^2}$$

$$R^2 = 1 - \frac{\sum(\text{実測値—予測値})^2}{\sum(\text{実測値} - \text{実測値の平均値})^2}$$

残留塩素の連続測定装置の検出限界値 0.05 mg/L を勘案して、±0.025 mg/L を許容誤差範囲とし、この範囲内に予測結果の最大値が収まることをモデル構築にあたっての目標とした。

C. 結果と考察

タイムステップがモデル精度に与える影響

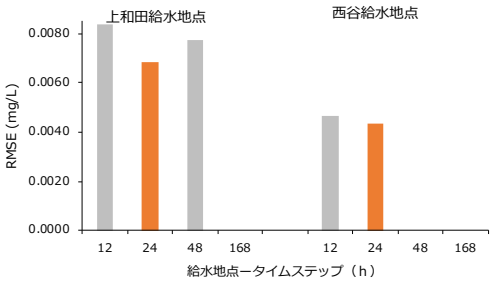

図-1　タイムステップと予測誤差の関係

データセット①中、入力を残留塩素低減量、導電率及び水温とし、出力を 6 時間後の残留塩素低減量とした際に、タイムステップがモデル精度に及ぼす影響を検討する。図-1 に上和田給水地点及び西谷給水地点におけるタイムステップと RMSE の関係を示す。図-1 を見ると、タイムステップ 12～48 時間において、すべてのプロットが目標精度である実測値±0.025 mg/L 以内に収まっていた。特に、タイムステップを 24 時間に設定した際に、最も高い精度が得られた。この結果は、対象とした地域では、水質変動パターンが 24 時間周期であることに起因すると考える。本研究では、以後の検討においてタイムステップを 24 時間に設定した。

図-1 中、上和田給水地点と西谷給水地点を比較すると、どの条件も上和田給水地点がより低い予測精度を示した。上和田給水地点は浄水場から比較的近く、残留塩素濃度低減量が低いことから測定誤差が大きくなったものと推測する。

上和田給水地点及び西谷給水地点のタイムステップを 168 時間に設定した際、及び西谷給水地点のタイムステップを 48 時間に設定した際に、モデルの出力が出来なかった。これはパラメーターが複雑になったことで、モデルが収束しなかったことが原因と考えられる。

予測期間がモデル精度に与える影響

モデルの出力となる予測期間は、浄水場職員からヒアリングした上で、3、6、12 及び 24 時間をそれぞれ検討した。本研究で対象とした給水地点は、平均的な滞留時間が 6 時間程度であることが事前調査で明らかになっている。予測結果に応じて浄水場出口での次亜塩素酸の注入制御を実施するには、6 時間先の予測結果が少なくとも必要となる。

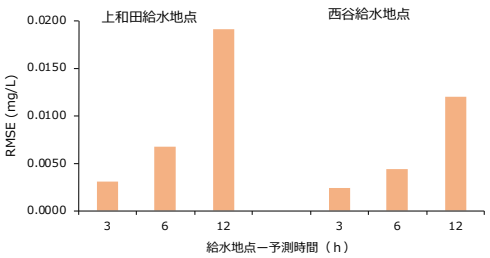

図-2　予測時間と予測誤差の関係

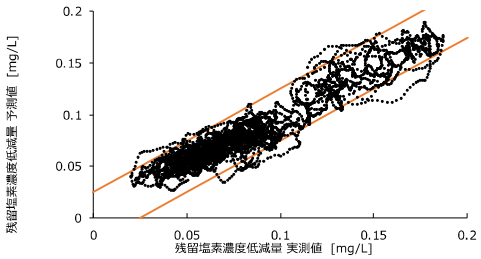

図-3　上和田給水地点における 12 時間後の残留塩素低減量　予測モデル精度

データセット①中、入力を残留塩素濃度低減量、導電率及び水温とし、出力を 3、6、及び 12 時間先の残留塩素濃度低減量とした際のモデル精度を図-2 に示す。両方の給水地点共に、予測時間の増加に伴って精度が徐々に悪化する傾向が得られた。予測時間を 12 時間に設定した際に、RMSE は 0.020 mg/L を示したが、実測値と予測値の散布図（図-3）から、一部のプロットにおいて、実測値±0.025 mg/L から逸脱する点が観察された。予測時間を 6 時間に設定した際（図-4）に、すべてのプロットが実測値±0.025

mg/L 以内に収まったことから、より高い精度で予測するには、予測時間 6 時間が望ましいことが分かる。

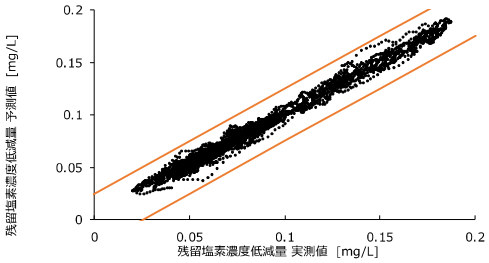

図-4　上和田給水地点における 6 時間後の残留塩素低減量　予測モデル精度

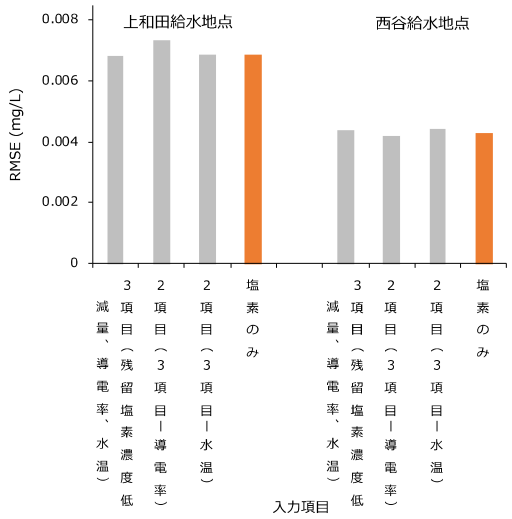

図-5　入力項目と予測精度の関係

また、図-1 と同様に、図-2 中、上和田給水地点は西谷給水地点と比較して、どの条件でもより低い予測精度を示した。

以上の結果から、現場で必要とされる 6 時間先の残留塩素濃度低減量を予測できることが示された。

入力項目が予測精度に与える影響

これまで、入力には残留塩素濃度低減量の他、一般的な水質項目である導電率と水温も使用してモデルを構築してきた。続いて、モデル構築に最小限必要となる項目を検討するために、各項目の予測精度に対する感度を分析した。

データセット①を対象として、入力項目中、残留塩素濃度低減量、導電率及び水温から、導電率を削除した場合、水温を削除した場合、導電率と水温を削除した場合の 4 条件で 6 時間後の残留塩素濃度低減量を予測するモデルを構築し、予測精度を比較した結果を図-5 に示す。

図-5 の通り、入力項目を 3 項目（残留塩素濃度低減量、水温、導電率）から 1 項目（残留塩素濃度低減量）に減らした際に、RMSE に大差が見られなかったことから、水温と導電率がモデル精度に及ぼす影響が小さいことが分かる。残留塩素濃度低減量を予測する物理モデルについて検討する既往研究において、水温や導電率を環境因子とする研究が散見されるが、LSTM により構築したモデルは、これらの環境因子を参照せずに、残留塩素濃度低減量の経時変化のトレンドを捉えることで、将来を予測していると推測される。本研究により、予測モデルの構築にあたって、残留塩素低減量の情報のみで、十分な精度のモデルが構築できることが明らかになった。

データ量が予測精度に与える影響

モデル構築に必要となるデータ量（データ蓄積期間）について検討する。データセット②を対象として、入力項目を残留塩素濃度低減量として、6 時間後の残留塩素濃度

低減量を予測するモデルを構築した。モデル構築にあたって、使用するデータを 12 ヶ月分から 1 ヶ月ずつ減少することで、データ量がモデル精度に及ぼす影響を検討した。なお、データの減量について、3 月から 4 月にかけて 1 ヶ月ずつ減らした P1 と 4 月から 3 月にかけて 1 ヶ月ずつ減らした P2 について、それぞれ検討した。2014 年から 2018 年の 5 年分実施した際の平均値を図-6 及び図-7 に示す。

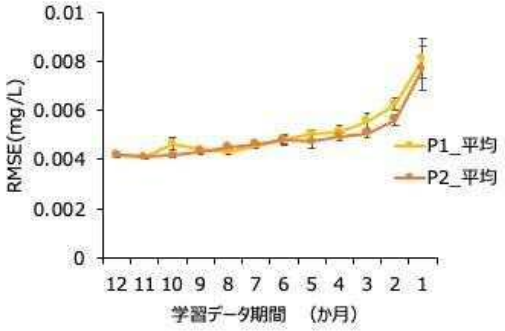

図-6　上和田給水地点におけるトレーニングデータの期間が予測精度に及ぼす影響

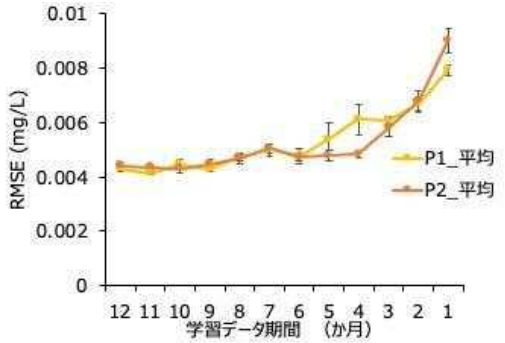

図-7　西谷給水地点におけるトレーニングデータの期間が予測精度に及ぼす影響

上和田給水地点及び西谷給水地点ともに、3 月から前年の 4 月まで 1 ヶ月ずつ遡ってトレーニングデータを減少した際(P2)に、4 か月分まで同定度の精度を維持した一方で、3 ヶ月分になると急激に精度が悪化した。この傾向は、3 月から 4 月まで 1 ヶ月毎データを減らした際（P1）でも同様の傾向が観察された。これらのことから、本研究で使用したデータについて、最低でも 4 か月間のトレーニングデータが最低限、必要となることが示唆された。

西谷給水地点において、P1 条件下で 4 か月間（12 月～3 月）学習した際のモデル精度が他の結果と比較して顕著に低いことがわかる。おそらく、12 月から 3 月のデータが他とは異なるトレンドを示したことが原因と考える。これらの結果から、データによってモデル精度が変化することが明らかになったと共に、モデル構築に用いるデータによっては、さらに長期間のデータセットを用いてモデル構築を行う必要性が示された。

トレーニングデータの質が予測精度に与える影響

本研究で用いたデータでは、精度の高いモデル構築に 4 ヶ月間のトレーニングデータが必要となることが明らかになった。4 か月間のトレーニングデータについて、開始月と終了月を変化させてモデル構築することで、精度の悪化を誘発するデータ群を探索した。

1 年のデータを 4 ヶ月毎に区切ることで 12 パターンのトレーニングデータを準備し、モデル構築に用いた。2014 年から 2018 年の 5 年間について年度毎にモデルを構築し、各パターンにおける精度の平均値を算出した。予測には、入力値、出力値共に残留塩素

濃度低減量を用いた。

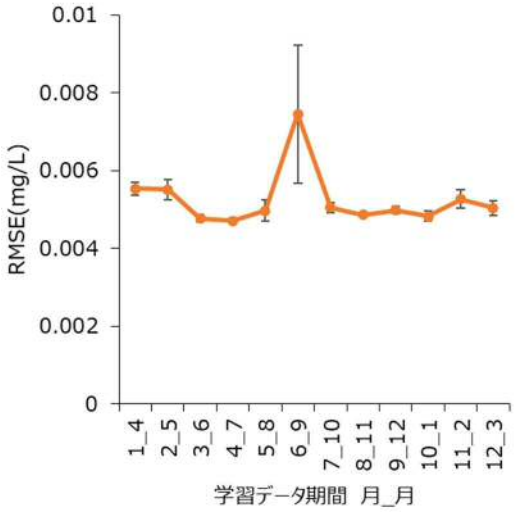

図-8　上和田給水地点におけるトレーニングデータの時期が予測精度に及ぼす影響

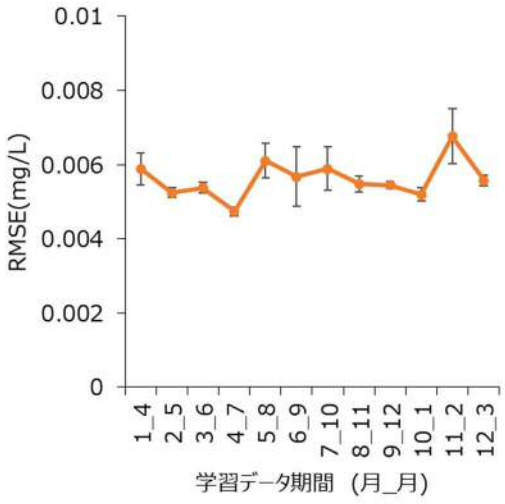

図-9　西谷給水地点におけるトレーニングデータの時期が予測精度に及ぼす影響

上和田給水地点の結果を図-8、西谷給水地点の結果を図-9 に示す。4 ヶ月間のトレーニングデータであっても、用いたデータ期間によって精度が異なることが明らかになった。本研究では、上和田給水地点、西谷給水地点共に、4 月から 7 月の 4 か月間のデータをトレーニングデータとして使用した際に、最も高い精度が得られた。一方で、上和田給水地点では 6 月から 9 月にかけて夏期間のデータ、西谷給水地点では 11 月から 2 月にかけて冬期間のデータをトレーニングデータとして使用した際に、モデル精度が低くなることが明らかになった。

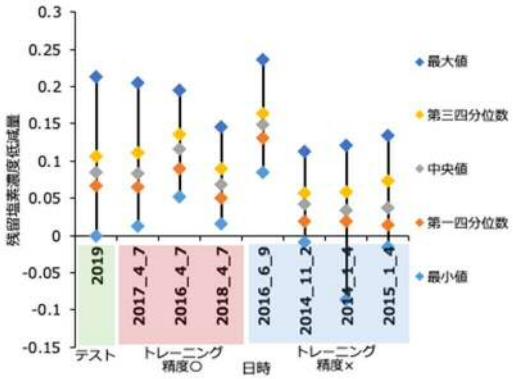

図-10　高精度及び低精度が得られたモデルのトレーニングデータのデータ特性

精度に影響するトレーニングデータ特性を検討するため、特に精度が高いモデル及び低いモデル構築に用いたトレーニングデータについて、最大値、最小値、中央値、四分位値をそれぞれ算出した結果を図-10 に示す。

テストデータとして、本研究では 2019 年のデータを用いたが、特に精度が高かった 4 月から 7 月のデータの分布とテストデータの中央値、四分位値がほぼ一致していることがわかる。一方で、精度が低かったトレーニングデータは、中央値、四分位値がテストデータと大きくずれていた。これらの結果から、テストデータの分布に近いデ

ータ群をトレーニングデータとして利用することで、精度が高いモデルを構築できると考えられる。

D. 結論

本研究では、時系列の濃度変化の傾向を長短期記憶ネットワーク（LSTM）アルゴリズムにより学習することで、数時間先の残留塩素濃度を推測できる予測モデルの構築を目的とした。また、他の自治体での適用可能性を示すためにモデルの構築に必要最小限のデータ量を検討した。

モデル構築にあたって最適なタイムステップは 24 時間であり、誤差目標値±0.025 以下に収めるには、予測時間を 6 時間以下にする必要があることが判明した。

モデル構築に必要最小限のデータ量は 4 月～7 月の 4 か月間の残留塩素濃度低減量であることがわかった。この期間のデータをトレーニングデータに用いたとき、高い精度のモデルが構築できた。以上より残留塩素濃度の 1 時間間隔の時系列データが 4 ヶ月分準備できれば､6 時間先の残留塩素濃度を LSTM により予測できることがわかった。

E. 研究発表

１．論文発表

なし

２．学会発表

なし

F. 知的所有権の取得状況

１．特許取得

なし

２．実用新案登録

なし

３．その他

なし

参考文献

1. Abokifa, A. A.; Yang, Y. J.; Lo, C. S.; Biswas, P., Water quality modeling in the dead end sections of drinking water distribution networks. *Water Res* **2016,** *89*, 107-17.

2. Bowden, G. J.; Nixon, J. B.; Dandy, G. C.; Maier, H. R.; Holmes, M., Forecasting chlorine residuals in a water distribution system using a general regression neural network. *Math Comput Model* **2006,** *44* (5-6), 469-484.

3. Le, X. H.; Ho, H. V.; Lee, G.; Jung, S., Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. *Water* **2019,** *11* (7).