配水システムの残留塩素濃度予測を目的とした LSTM モデルの提案

中岡　祐輔[1]・荒井　康裕[2]・小泉　明[3]

[1] 学生会員　東京都立大学大学院　都市環境科学研究科（〒192-0397 東京都八王子市南大沢 1-1）

E-mail: nakaoka-yusuke@ed.tmu.ac.jp

[2] 正会員　東京都立大学大学院　都市環境科学研究科（〒192-0397 東京都八王子市南大沢 1-1）

E-mail: y-arai@tmu.ac.jp

[3] フェロー会員　東京都立大学大学院　都市環境科学研究科（〒192-0397 東京都八王子市南大沢 1-1）

水道水の残留塩素（残塩）濃度は，安全でおいしい水を供給するうえで重要な水質項目のひとつである．したがって，浄水場では配水過程における残塩濃度の減少を考慮して，その注入量を決定しなければならない．こうした残塩濃度の予測は，主に動的管網解析を用いて行われてきたが，予測の簡便性やビッグデータの活用を期待して，ニューラルネットワークによる手法も提案されている．

本研究では，ニューラルネットワークのなかでも，時系列データの解析に広く使用されている LSTM に着目し，過去の情報を予測に反映することができるモデルの構築を試みる．モデルを構築する際の入力データとして，個人宅残塩濃度のほかに，浄水場送水流量，浄水場残塩濃度，個人宅水温を選択した結果，個人宅残塩濃度の挙動を細部まで再現することが可能となった．

Key Words: *residual chlorine, machine learning, Neural Network, Long Short Term Memory*

1. はじめに

我々の生活を支える水道インフラは，新たな施設の整備拡張を目的とした時代から，水質の向上や現状設備の維持管理に重点を置いた時代へと変化している．しかしながら，近年の少子高齢化による今後の人口減少を考慮すると，各水道事業体は給水量や，それに伴う料金収入の減少，さらには技術者不足から起こりうるノウハウの継承問題にも適切に対応し，安全な水の供給を持続できるよう，努めていかなければならない．したがって，今後の水道事業を運営する方法として，最新の情報通信技術やモニタリングデータを活用した施設運転の合理化が求められている [1]．

このモニタリングデータには，水道水の安全性を確保する目的で定められた水質基準項目や水質管理目標設定項目が該当し，浄水場や個人宅に設置された自動水質計器では，常に水質に関するデータが収集され，万が一の水質悪化事故に備えている．こうした常時監視によって得られた膨大なデータを解析することで，水道水の水質予測を行うことが可能であれば，上記に示した水道事業の合理化の一助になると考えられる．

水質基準項目や水質管理目標設定項目には様々な物質に関して，その基準値や目標値が設定されているが，なかでも残留塩素（残塩）は「安全でおいしい水」を実現するうえで重要な項目である．水道法では水道水の安全性を確保するため，給水栓における残塩濃度は 0.1 mg/L 以上（遊離残塩の場合）を保持するように定められているが，管路内の流下や滞留によって塩素が消費されると，残塩濃度の減少によって，上記の基準値を満たせない恐れが生じる．また，塩素処理は水道水のカルキ臭や異臭味の原因ともなりうることから，その目標値は 1.0 mg/L 以下と定められている．したがって，水道水の「安全性」と「おいしさ」を両立するためには，配水過程における残塩濃度の減少や，末端の基準値や目標値を考慮したうえで，浄水場での塩素注入量を適切にコントロールしなければならない．

こうした残塩濃度の予測を行う際には，動的管網解析（EPANET）を用いることが主流だが，前述の合理化に向けた手法として，ニューラルネットワーク（Neural Network：NN）を活用した予測モデルの提案が検討されてきた [2]．しかしながら，先行研究で用いられた全結合型 NN では，時系列データとしての特徴を十分に反映で

きないことから，残塩濃度の予測精度には課題が残る結果となった [3]．そこで，本研究では時系列データ解析に広く使用されている再帰 NN（Recurrent Neural Network : RNN）に着目し，さらに長期間の情報を記憶できるよう拡張された LSTM（Long Short Term Memory）を援用し，個人宅残塩濃度を予測するモデルの構築を試みる．

以下，2. では，本研究の対象データと分析手法である LSTM の概要について示す．次に 3. では，LSTM モデルを構築する際に必要となる入力時間幅と出力時間差に焦点を置きながら，本研究で用いるモデルの詳細についてまとめる．さらに 4. では，実際に構築したモデルを用いて個人宅残塩濃度の予測を行い，説明変数の差異がモデルの出力にどのような影響を与えるのかを明らかにする．最後に 5. では研究成果のまとめを示す．

2. 対象データと分析手法

(1) 本研究の対象データ

本研究では **図-1** に示す送配水ネットワークで得られたデータを対象として分析を行う．すなわち，K 浄水場計測データ（送水流量，濁度，残塩濃度，pH）と，各配水池計測データ（S 系第一配水流量，M 系–d 配水流量，M 系–o 配水流量），および個人宅計測データ（濁度，残塩濃度，pH，水温，色度，電気伝導率，水圧）が本研究の対象となる．なお，データの計測期間は 2016 年 4 月 1 日から 2017 年 3 月 31 日までの 1 年間であり，すべて時間単位データである．これらのデータのうち，モデルで扱う入力データ（説明変数）および，出力データ（目的変数）を設定する．ここで，全結合型 NN を用いた先行研究では，モデルの入力データに浄水場送水流量と浄水場残塩濃度，さらに残塩の消費に影響を与える水温を用いることで，末端の残塩濃度の挙動が再現可能になるとされている [2]．したがって，本研究もこの結果を踏まえ，K 浄水場送水流量 Q_t（m³/h），K 浄水場残塩濃度 B_t（mg/L），個人宅水温 W_t（℃），個人宅残塩濃度 C_t（mg/L）の 4 変数をモデルの構築に用いることとする．

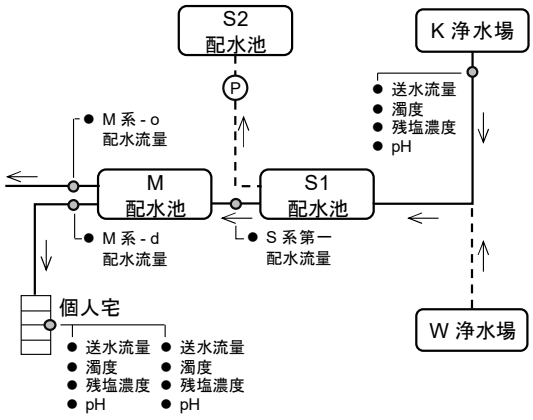

図-1 送配水ネットワーク

(2) LSTM の概要

本研究では機械学習のひとつである NN を用いて，残塩濃度を予測するモデルを構築する．先行研究で用いられた全結合型 NN の構造を **図-2** に示す．NN モデルは，脳の中に存在する神経細胞（ニューロン）のつながりをコンピュータ上で再現し，入力層から出力層へと信号を伝えることで，入出力データ間の関係を学習させるものである．

この全結合型 NN では，時刻 t の入力 $\boldsymbol{x}_t$ に対する出力 $\boldsymbol{h}_t$ は，重み（信号の伝わりやすさ）$\boldsymbol{W}_x$ とバイアス $\boldsymbol{b}$ を用いて，式(1)のように表せる．

$$\boldsymbol{h}_t = \boldsymbol{x}_t \boldsymbol{W}_x + \boldsymbol{b} \tag{1}$$

一方で，本研究で用いる LSTM の元となった RNN には，全結合型 NN とは異なり **図-3** で示すようなループ構造が存在する．そのため RNN では，出力 $\boldsymbol{h}_t$ を時

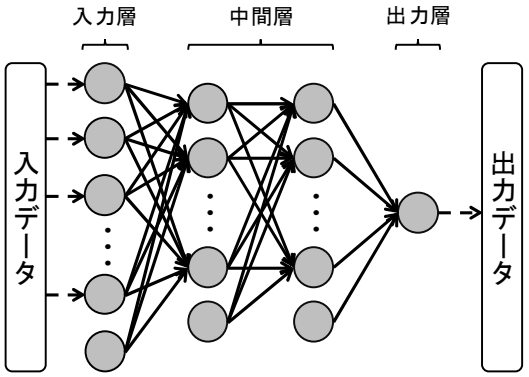

図-2 全結合型 NN の構造

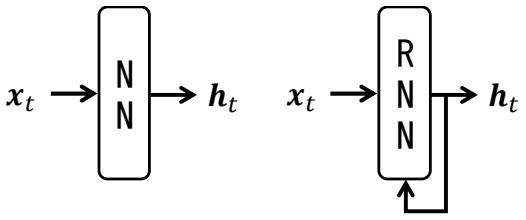

図-3 全結合型 NN と RNN の概略図

刻 $t+1$ の入力データとして再び用いることができる．すなわち，RNN の時刻 t の入力には $\boldsymbol{x}_t$ のほかに時刻 $t-1$ の出力 $\boldsymbol{h}_{t-1}$ が加わり，このときの出力 $\boldsymbol{h}_t$ は，重み $\boldsymbol{W}_x$，$\boldsymbol{W}_h$ を用いて，式(2)のように表せる．

$$\boldsymbol{h}_t = \tanh(\boldsymbol{h}_{t-1}\boldsymbol{W}_h + \boldsymbol{x}_t\boldsymbol{W}_x + \boldsymbol{b}) \tag{2}$$

全結合型 NN では，すべてのデータが互いに独立している（過去の情報が将来に影響を及ぼさない）と仮定したうえで学習が行われているのに対し，RNN ではこのループ構造を有することで，過去の情報を記憶しながら学習を行うことが可能である．このように，時系列データの学習と相性のよい RNN ではあるが，長期的なデータの関係を学習することが難しいという弱点がある．そこで，この弱点を補うように提案されたのが LSTM である．

図-4 に示すように，LSTM には記憶セル $\boldsymbol{c}$ と呼ばれる記憶部があり，時刻 t の記憶セル $\boldsymbol{c}_t$ には，過去から時刻 t までの必要な情報が記憶されている．この $\boldsymbol{c}_t$ は時刻 $t-1$ の記憶セル $\boldsymbol{c}_{t-1}$ のほかに，種々のゲートで計算を行った $\boldsymbol{x}_t$ と $\boldsymbol{h}_{t-1}$ を用いて求められる．また，時刻 t の出力 $\boldsymbol{h}_t$ についても，$\boldsymbol{c}_t$ のほかに，ゲートで計算を行った $\boldsymbol{x}_t$ と $\boldsymbol{h}_{t-1}$ を用いて求められる．これらのゲートは forget（忘却）ゲート，input（入力）ゲート，output（出力）ゲートから構成されており，ゲートを通過する情報量を調整することで，重要な情報のみを過去から将来へと引き継ぐことができる．

以上より，LSTM で時刻 t の出力 $\boldsymbol{h}_t$ を得るために行う計算は，次の式(3)から式(8)のように表せる[4]．ただし，式(3)から式(5)の $\boldsymbol{f}$, $\boldsymbol{i}$, $\boldsymbol{o}$ は，それぞれ forget, input, output ゲート通過時に行われる計算，式(6)の $\boldsymbol{g}$ は新しい情報を記憶セルに追加する際に行われる計算，σ は sigmoid 関数を表す．また，式(7)は時刻 t の記憶セル $\boldsymbol{c}_t$ を求める計算，式(8)は $\boldsymbol{c}_t$ をもとに出力 $\boldsymbol{h}_t$ を求める計算，$\odot$ は行列のアダマール積を表す．

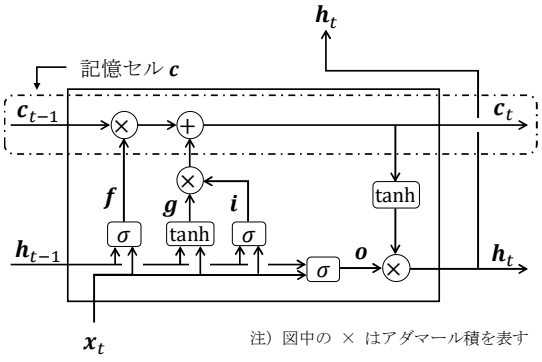

図-4 LSTM レイヤの内部構造

$$\boldsymbol{f} = \sigma\left(\boldsymbol{x}_t\boldsymbol{W}_x^{(f)} + \boldsymbol{h}_{t-1}\boldsymbol{W}_h^{(f)} + \boldsymbol{b}^{(f)}\right) \tag{3}$$

$$\boldsymbol{i} = \sigma\left(\boldsymbol{x}_t\boldsymbol{W}_x^{(i)} + \boldsymbol{h}_{t-1}\boldsymbol{W}_h^{(i)} + \boldsymbol{b}^{(i)}\right) \tag{4}$$

$$\boldsymbol{o} = \sigma\left(\boldsymbol{x}_t\boldsymbol{W}_x^{(o)} + \boldsymbol{h}_{t-1}\boldsymbol{W}_h^{(o)} + \boldsymbol{b}^{(o)}\right) \tag{5}$$

$$\boldsymbol{g} = \tanh\left(\boldsymbol{x}_t\boldsymbol{W}_x^{(g)} + \boldsymbol{h}_{t-1}\boldsymbol{W}_h^{(g)} + \boldsymbol{b}^{(g)}\right) \tag{6}$$

$$\boldsymbol{c}_t = \boldsymbol{f}\odot\boldsymbol{c}_{t-1} + \boldsymbol{g}\odot\boldsymbol{i} \tag{7}$$

$$\boldsymbol{h}_t = \boldsymbol{o}\odot\tanh(\boldsymbol{c}_t) \tag{8}$$

このように，LSTM では過去の情報の中から必要な情報を学習し，その結果を出力（将来の予測値）に反映させることができる．したがって，過去の値が将来に影響を与えることが多い時系列データにおいて，上記の特徴を有する LSTM を用いてモデル化を行う意義は十分にあり，本研究では LSTM による残塩濃度の予測モデルの構築を試みる．

3. LSTM モデルの詳細設定

(1) データの前処理

2.で述べたように，本研究では浄水場送水流量 Q_t，浄水場残塩濃度 B_t，個人宅水温 W_t，個人宅残塩濃度 C_t の 4 変数を用いてモデルを構築する．しかし，各変数の値をそのまま用いてモデルを構築してしまうと，時刻 $t+1$ の値を予測する際に，時刻 t の値（直前の値）を出力してしまう恐れがある．特に，平均二乗誤差を損失関数とする場合には注意が必要であり，直前の値をそのまま出力することで，モデルの学習が進んだと判断されてしまう．そこで，各変数の 1 階差分をとり，データの定常化を行うことで，上記の問題点を回避することにした．以降では，変数 x_t に対し，直前の値 x_{t-1} との 1 階差分をとったものを x'_t，その予測値を $\hat{x}'_t$ と表現する．

したがって，モデルの入力として用いる変数（説明変数）は Q'_t, B'_t, W'_t, C'_t の 4 つとなる．また，本研究では説明変数を C'_t のみとした場合（1 変数）の結果と比較することで，説明変数の差異がモデルの出力に影響を与えるか否かを検討する．

(2) 入力時間幅と出力時間差の設定

LSTM では「どれだけの時間幅を持ったデータを用いて，いくつ先の時刻の値を予測するか」をあらかじめ設定してから学習を行う必要がある．この「時間幅」と「いくつ先」の 2 つのパラメータに対し，前者を入力時間幅（TS），後者を出力時間差（OG）と呼ぶことにする．例えば「現時刻のデータを含めた過去 6 時間分のデータを用いて，次時刻（直後）の値を予測する」場合，$TS = 6, OG = 0$ となる．

(1)で述べたように，本研究では ${C'}_t$ のみを説明変数としたモデル（1 変数モデル）と，${Q'}_t$, ${B'}_t$, ${W'}_t$, ${C'}_t$ の 4 つを説明変数としたモデル（4 変数モデル）を作成する．このとき，それぞれのモデルについて TS と OG の設定を次のように行った．

1 変数モデルの場合は，現時刻を含めて過去 96 時間分のデータ（${C'}_t, {C'}_{t-1}, \cdots, {C'}_{t-95}$）から，次時刻の値 ${C'}_{t+1}$ の予測を行うように設定する．すなわち，$TS = 96, OG = 0$ となる．一方で，4 変数モデルの場合は，個人宅のデータだけではなく，浄水場のデータも扱う点に注意して TS と OG を設定しなければならない．まず，4 変数モデルでは 1 変数モデルと比較して，扱うデータの種類が 4 倍になっていることから，TS は 1 変数モデルの 1/4（$TS = 24$）とする．次に，OG について，この値を 1 変数モデルと同様に「0」とすることはできない．なぜなら，浄水場で計測された水質や流量に関するデータが，直ちに個人宅へ影響を与える可能性は低いからである．実際にこの地域における B_t と C_t の相互相関分析を行った結果，B_t は 11 時間後の C_t に最も影響を及ぼすことが明らかになっており [5]，$OG = 10$ とすれば，浄水場から個人宅への滞留時間を考慮したモデルが作成できる．

このように，1 変数モデルと 4 変数モデルでは同時刻の ${C'}_t$ を予測する場合でも，入力として用いるデータセットは異なる．例として 図-5 では，6 月 1 日 0 時の ${C'}_t$ を予測する場合に，どの期間データを用いればよいかを表している．図中で灰色に囲まれた部分は予測に用いるデータであり，逆に 4 変数モデルの斜線部は OG に含まれるデータであるから，こちらは予測には用いないことを意味する．

(3) 学習条件の設定

機械学習では訓練データを用いて学習を行ったのち，テストデータを用いてモデルの汎化能力を確認するといったプロセスをとる．前述の通り，本研究の 1 変数モデルと 4 変数モデルでは，訓練データとして与えるデータの期間が異なる．そのため，予測を行う期間を両モデルで同一とし，両者の比較を簡単に行えるようにした．以降では「訓練データを用いて予測を行った期間」を訓練期間と呼ぶことにする．なお，訓練期間とテスト期間を設定したのち，これらの期間に含まれるデータを 0 から 1 の範囲で正規化することで，変数のデータスケールの差が学習効果に影響を与えないようにした．

また，両モデルともに LSTM の中間層（隠れ層）数は 1 層，ユニット数は 512，バッチサイズは 24 とした．学習回数（エポック数）については，1 変数モデルは 1000，4 変数モデルは 500 をそれぞれ上限とし，テスト期間の損失関数の推移が収束した段階で学習を打ち切り，過学習（モデルが訓練データのみに対して過度に適合すること）が起こらないように配慮した．なお，モデルの最適化アルゴリズムには Adam を用いた．Adam は 2014 年に提案され，ハイパーパラメータの設定が容易であるなどの利点を持ち，従来のアルゴリズムより効率的なパラメータの探索が可能となっている [6]．

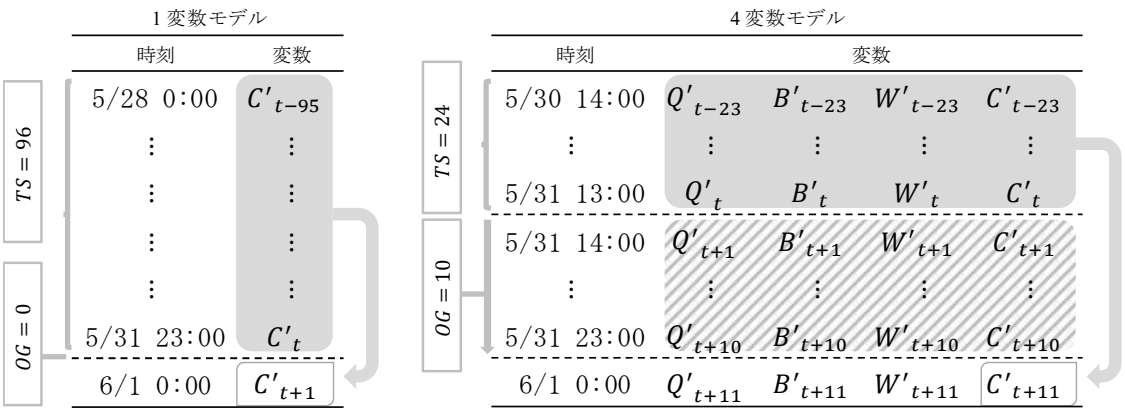

図-5 入力時間幅と出力時間差の設定例

4. 説明変数の差異がモデル出力へ与える影響

(1) データの後処理

本研究のモデルは個人宅残塩濃度の 1 階差分 C'_t を予測するものである．モデルが個人宅残塩濃度を正しく推定できたかを知るために，式(9)を用いてモデルの予測値 $\widehat{C'}_t$ から，個人宅残塩濃度の推定値 $\hat{C}_t$ を求める．

$$\hat{C}_t = C_{t-1} + \widehat{C'}_t \tag{9}$$

以降では，式(9)で求めた $\hat{C}_t$ をモデルによる推定値とし，時系列図の描画や推定誤差の計算も実測値 C_t と推定値 $\hat{C}_t$ を用いて行う．

(2) モデルの推定結果

1 変数モデルの推定結果時系列図を **図-6** に，4 変数モデルの推定結果時系列図を **図-7** に示す．なお，両モデルともに，訓練期間は 2016 年 6 月 1 日から 6 月 7 日までの 1 週間とし，続く 1 日（6 月 8 日）をテスト期間とした．**図-6** の 1 変数モデルでは，実測値の増減傾向を概略的に再現できているが，局所の「山」や「谷」の部分で推定精度が低下する現象がみられる．一方で，**図-7** の 4 変数モデルでは，テスト期間を含めて，実測値の挙動を細部まで再現しており，説明変数の種類を増やした効果がここに現れたと言える．

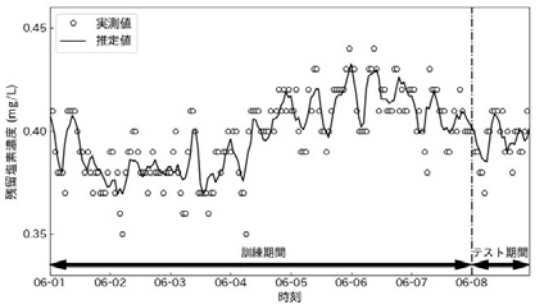

図-6 1 変数モデル推定結果

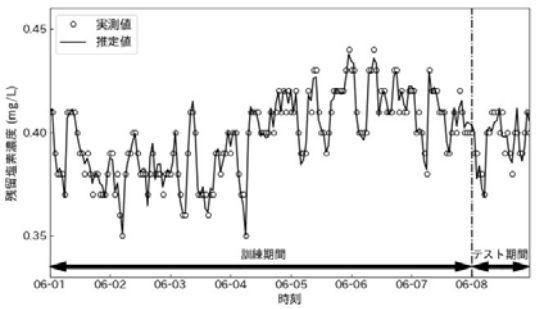

図-7 4 変数モデル推定結果

表-1 推定結果のまとめ

モデル	学習回数	平均絶対誤差 (mg/L)		最大誤差 (mg/L)	
		訓練	テスト	訓練	テスト
1 変数	427	0.0078	0.0075	0.0322	0.0156
4 変数	328	0.0024	0.0056	0.0107	0.0126

また，推定結果の精度をまとめた **表-1** より，4 変数モデルは 1 変数モデルと比較して，平均絶対誤差を訓練期間で 0.0054 mg/L，テスト期間で 0.0019 mg/L 改善することができた．さらに，最大誤差も同様に訓練期間で 0.0215 mg/L，テスト期間で 0.0030 mg/L 改善することができた．

以上より，本研究の LSTM モデルでは，説明変数（入力データ）として個人宅残塩濃度のほかに，浄水場送水流量，浄水場残塩濃度，個人宅水温を加えることで，良好な推定精度を持つモデルを得ることができた．残塩濃度予測を目的とした LSTM モデルを構築する際には，説明変数の差異がモデル出力に影響を与えることから，説明変数の選択はもとより，それらに関連する入力時間幅，出力時間差の設定を適切に行うことが必要である．

5. おわりに

本研究では機械学習のひとつである LSTM を用いて，個人宅残塩濃度の 1 階差分を予測するモデルを作成した．また，このモデルの予測結果をもとにした個人宅残塩濃度の推定も行った．本研究で得られた成果を以下に列記する．

1) 本研究で用いる LSTM の内部構造に着目し，先行研究で用いられてきた全結合型 NN との違いを明らかにしたうえで，時系列データである残塩濃度の予測を試みた．
2) LSTM モデルを構築する際には入力時間幅と出力時間差の設定が必要であり，用いる説明変数（入力データ）の種類によって，適切に設定しなければならない．特に，浄水場のデータを用いる場合には，個人宅までの配水過程における滞留時間を考慮したうえで，出力時間差を設定しなければならない．
3) モデルの説明変数を個人宅残塩濃度のみとした 1 変数モデルと，個人宅残塩濃度のほかに，浄水場送水流量，浄水場残塩濃度，個人宅水温を加えた 4 変数モデルを構築し，推定精度の比較を行った．その結果，4 変数モデルの推定結果では，実測値の時系列挙動を細部まで再現することが可能であり，1 変

数モデルと比べて，平均絶対誤差および最大誤差による予測精度を改善することができた．

今後の課題として，本研究で提案したアプローチを異なる送配水ネットワークに適用しながら，モデルの学習条件（中間層やそのユニット数）に関して詳細に検討することや，訓練期間の長さの違いによる予測精度の影響等について分析する必要があると考える．

参考文献

1) 厚生労働省：新水道ビジョン，2013.
2) 稲員とよの，小泉明：配水管網における残留塩素濃度推定に関するニューラルネットワークの応用，水道協会雑誌，第71巻，第8号，pp.2-10，2002.
3) 中岡祐輔，荒井康裕，酒井宏治，小泉明，佐々木史朗：送配水過程における残留塩素濃度減少の推定モデルに関する一考察，令和元年度全国会議（水道研究発表会）講演集，pp.836-837，2019.
4) 斎藤康毅：「ゼロから作るDeep Learning 2 自然言語処理編」，pp.236-244，オライリー・ジャパン，2018.
5) 荒井康裕，稲員とよの，堀口 幸菜，小泉明，佐々木史朗：配水管網の水質監視データを活用した残留塩素濃度シミュレーション，土木学会第 73 回年次学術講演会，pp.265-266，2018.
6) Diederik P. Kingma and Jimmy Lei Ba: Adam: A Method for Stochastic Optimization, arXiv:1412.6980 [cs.LG], 2014.

(Received June 19, 2020)

PROPOSAL OF LSTM MODEL FOR RESIDUAL CHLORINE CONCENTRATION PREDICTION IN WATER DISTRIBUTION SYSTEMS

Yusuke NAKAOKA, Yasuhiro ARAI and Akira KOIZUMI

The residual chlorine concentration in tap water is one of the most important water quality factors for supplying safe and delicious water. Therefore, it is necessary to determine the injection amount in the water treatment plant in consideration of the decrease in the residual chlorine concentration in the water distribution process. Traditionally, residual chlorine concentration predictions have been made primarily using pipe network analysis. Recently, research using neural networks has been developed because it is easier to use big data.

This study focused on LSTM, which is widely used for time series data analysis. It is expected to build a model that can reflect past information in a time series prediction. The input data of this model consists of residual chlorine concentration and water temperature at the demand point (end of the network), water flow rate and chlorine concentration at the water treatment plant. As a result, it became clear that the proposed model can reproduce the behavior of residual chlorine concentration at the demand point.