有効なデータが少量の事象に対するAIによる予測の検討
－土石流の発生に関する事例を対象として－

五大開発株式会社　正会員　○荒木光一　非会員　藤田達大
日本工営株式会社　正会員　伊藤隆郭　正会員　田方智
非会員　古木宏和　正会員　倉上健

1．はじめに

AI による時系列データの予測は盛んに研究されており，土木分野では現場のデータを用いた研究も行われている[1]．このような研究で扱う時系列データは，有効なサンプルが多量であることを前提としているため，少量の場合は予測精度の低下，または，予測不可能となる．実際に，桜島の野尻川では土石流時以外の流量はほぼ $0m^3$/秒であり，有効な流量は土石流発生時のみであるため，有効なサンプルは少なく，不連続である．

本稿では，有効なサンプルが少量の事象に対する AI の活用に関して，サンプル数の影響を確認することを目的とし，まず，野尻川の流量データで土石流発生時期とピーク流量の予測精度を確認する．次に，平均日気温のサンプル数と予測精度の関係から，AI 活用を検討するべきサンプル数について議論する．

2．有効なデータが少量の流量による土石流発生時期とピーク流量の予測の事例

有効なサンプル数が少量のデータからの予測事例として，土石流発生時期とピーク流量の予測を示す．本来，これらの予測は雨量などのデータを考慮する必要があるが，本稿の目的は，AI 活用時のサンプル数に関することであることを注意されたい．

野尻川において次回発生する土石流の時期とピーク流量の予測には，これまでの土石流発生時に観測したピーク流量を用いた．ピーク流量を観測した回数は 5 年間で 38 回(サンプル)であり，AI で扱うにはサンプル数が非常に少ないため，サンプル間を滑らかに補間できるエルミート補間を施した(図 1)．

予測手法には，時系列データに有効な Long Short Term Memory(LSTM)[2]を採用した．表 1 に，ネットワーク構成を示す．訓練データとテストデータは，エルミート補間したデータを 8:2 に分割して作成した．な

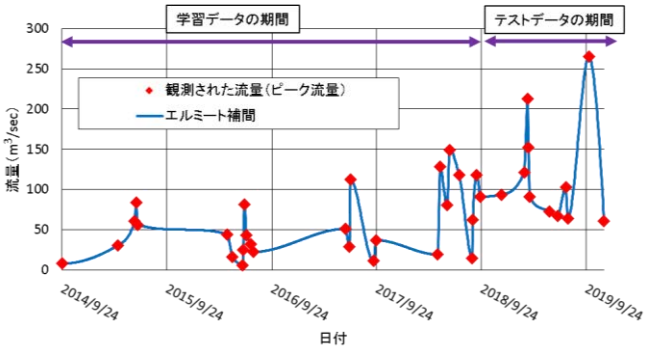

図 1 観測したピーク流量のエルミート補間

表 1 ネットワーク構成

1 層目	LSTM（8 ノード）
2 層目	全結合（1 ノード）
活性化関数	Linear 関数

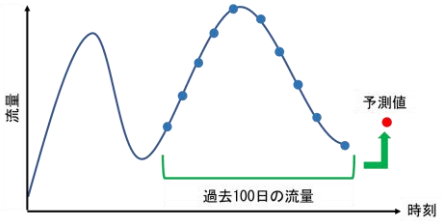

図 2 テストデータによる予測方法

お，過学習を確認するための検証データは，訓練データの 20%としてランダムに選択した．データの前処理は，訓練データの平均と分散による標準化とした．LSTM の入力は予測日の前日から過去 100 日間の流量とした．

図 2 に，テストデータによる流量の予測方法を示す．入力は予測日の前日から過去 100 日間の流量とした．したがって，自己回帰モデルのように，入力には予測した流量を含まない．図 3 に，LSTM によるテストデータの予測結果を示す．予測結果はピーク流量(菱形プロット)とは一致しなかったが，流量の上昇・下降のタイミングは概ね捉えている．ただし，土石流時の流量は急激に上昇するが，予測結果は緩やかに上昇しているため，学習で抽出した特徴は土石流の時期とピーク流量ではなく，エルミート補間の関数であると考えられ

キーワード　時系列データ，サンプル数，Long Short Term Memory （LSTM）
連絡先　〒921-8051　石川県金沢市黒田１丁目 35 番地　五大開発株式会社　ＴＥＬ076-240-6588

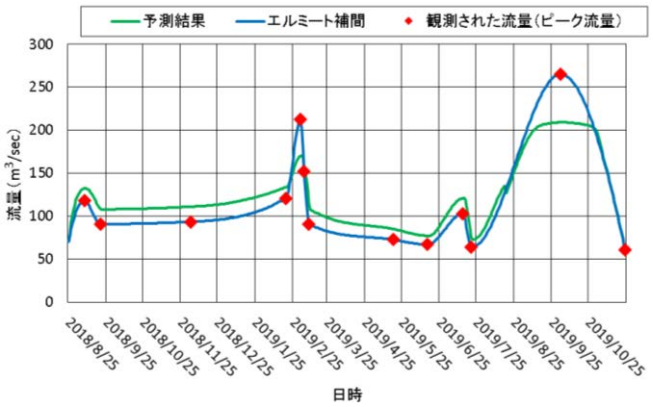

図 3 土石流に関するテストデータの予測結果

る．このことから，土石流発生時に観測された流量だけでは，有効なサンプル数が少量であるため，予測は困難であるといえる．したがって，AI で予測するためには，有効なサンプル数を増やす必要があるといえる．

3．平均日気温によるサンプル数と予測精度の関係

鹿児島市の平均日気温を例として，サンプルを間引くことで擬似的にサンプル数を変化させて予測精度の変化を確認する．

平均日気温は気象庁から取得し，取得期間を 2011 年 1 月 1 日～2020 年 12 月 31 日(3,653 サンプル)とした．サンプル数に対する予測精度の変化を確認するために，平均日気温を 7，14，30，60，200 と 300 の 6 パターンでサンプルを間引き，それぞれに対してエルミート補間を施した．

各間引きパターンにおいて訓練データとテストデータは，2011 年 1 月 1 日～2017 年 12 月 31 日(2,557 サンプル)と 2018 年 1 月 1 日～2020 年 12 月 31 日(1,096 サンプル)とした．検証データは，訓練データの 20%としてランダムに選択した．予想手法は LSTM とし，LSTM のノード数を 100 にした以外は表 1 と同様のネットワークである．データの前処理は訓練データの標準化とし，LSTM の入力は予測日の前日から過去 10 日間とした．

図 4 に，テストデータにおける各間引きパターンの平均誤差率を示す．間引いたサンプル数が多くなるにつれて，平均誤差率は増加する傾向がある．間引きなしと 7 サンプル間引いた平均誤差率の差は約 2%となった．図 5 に，代表的な間引きパターンの日平均気温の予測結果を示す．30 サンプル間引いた予測結果は，間引きなしより予測精度は低いが，正解の気温の起伏は捉えている．一方，平均誤差率が高い 300 サンプルの間引きでは気温を予測できていない．

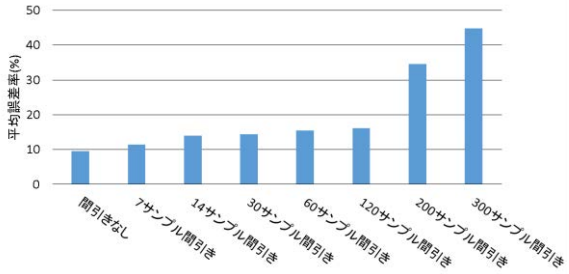

図 4 間引きパターンと平均誤差率の関係

図 5 各間引きパターンにおける日平均気温の予測結果

平均日気温の予測結果から，間引くサンプル数が小さい場合，予測精度の低下は数%である．したがって，予測誤差が許容範囲内であれば，観測の頻度は数日間隔であれば十分であると考えられる．

4．考察とまとめ

本稿では，サンプル数の観点から AI 活用の検討と確認を行った．平均日気温の結果を参考にすれば，本稿の土石流に関する予測には，少なくとも 30 日間隔で有効な流量が必要であると考えられる．ただし，本稿の土石流発生時期とピーク流量以外の事柄に対する予測で，予測誤差が許容範囲内であれば，サンプル数は十分である可能性はある．

今後は，不連続性を焦点としたサンプルの間隔や補間手法などに関して検討する予定である．

謝辞

本解析にあたり，桜島等のデータを参考にした．大隅河川国道事務所の諸兄に感謝申し上げます．

参考文献

[1]一言正之，澤谷拓海，植西清，“深層強化学習を用いたダム操作モデルのダム流入量予測誤差に対する影響評価”，AI・データサイエンス論文集，1 巻，J1 号，pp.459-464，2020 年 11 月

[2]Hochreiter. S. and Schmidhuber. J.，“Long Short-Term Memory”，Neural Computation，Vo.9，No.8，pp.1735-1780，Nov. 1997