機械学習を用いた異常値の検出と欠損値の修正方法 ―排水機場水位データを実例として―

Detection of abnormal values and correction of missing values using machine learning – A case study of water level data at a drainage pumping station

木村 延明*・吉永 育生*・関島 建志*・安瀬地 一作*・福重 雄大*・馬場 大地**
*農研機構　農村工学研究部門（〒305-8609 茨城県つくば市観音台 2-1-6）
**（株）アーク情報システム（〒102-0076 東京都千代田区五番町 4-2）

Nobuaki KIMURA*, Ikuo YOSHINAGA*, Kenji SEKIJIMA*, Issaku AZECHI*,
Yudai FUKUSHIGE*, and Daichi BABA**
*Institute for Rural Engineering, National Agriculture & Food Research Organization (NARO)
（2-1-6 Kannondai, Tukuba, Ibaraki 305-8609）
**ARK Information Systems, INC（4-2 Gobancho, Chiyoda-ku, Tokyo 102-0076）
(Received 4 December 2020, Accepted 22 February 2021)

Abstract

Data-driven technology in agriculture fields highly requires good quality data sets. Our study tried to find and correct error data (spike-noises and slide-gaps) using water level data, observed at the pond of a drainage pumping station. We employ One-Class Support Vector Machine (OCSVM) and specific conditions: discharge rate and continuity for anomaly detection of the error data. Additionally, Long Short-Term Memory (LSTM) and a mean-based simple method are utilized for missing data imputation. The results for anomaly detection show that spike-noises were feasibly detected although some truly normal data were regarded as abnormal data for wrong cases. Slide-gaps were successfully modified using the simple method, shifting to the mean water level, after detecting slide ranges by twice OCSVM trials with water level and their gaps. Missing data were corrected by LSTM. The corrected data were more realistic than those by linear interpolation.

Key words: machine learning, drainage pumping station, water level data, anomaly detection, missing data correction

要　旨

大量なデータを利活用するデータ駆動型農業において，データ品質を確保することは重要である．本研究では，低平地の排水機場調整池で観測された水位データから異常値（スパイクノイズとスライドずれ）を取り除き，それらの修正を試みる．異常検知は One-Class Support Vector Machine（OCSVM）と排水流量などの条件を用いて行い，また修正は平均値を使った簡易方法と Long Short-Term Memory（LSTM）を利用して行う．異常検知の結果として，スパイクノイズは概ね正確に検知できたものの，正常値を異常と検知した失敗例などが見られた．スライドずれは，水位に加え，水位差の OCSVM を導入することで，スライドずれの区間を検索し，平均水位を使う簡易的な方法で修正が可能であった．欠損値は LSTM を用いることで，2 点間の線形補間に比べて水位変化を再現することができた．

キーワード: One-Class SVM，LSTM，排水機場，水位データ，異常検知，欠損値修正

１．はじめに

近年，人工知能（以下，「AI」という）や情報通信技術（以下，「ICT」という）を使ったデータ駆動型の農業の普及が加速している．AI の導入について，ICT 器機などから収集されるデータに異常値や欠損値などが含まれていれば，AI は誤ったデータも学習してしまうので，精度の高い結果が得られない．従って，有用なデータ駆動型農業に向けて，ICT 器機で収集されたデータの確からしさなどを明確にする品質保証が求められている．一般的に，有用なデータにするために，冗長情報を修正するためのデータクレンジングや外れ値検出のためのフィルタリングが必要である．本研究では，スパイクノイズ（正常

値から外れた鋭い 1 点のみのピーク）やスライドずれ（連続した複数点が正常値から平行に上下にスライドする状態）の異常値を含むデータについて，それらの異常値を検出する方法（以下，「異常検知」という）に着目する．

異常検知には，一般的にホテリング理論（Hotelling, 1931）に基づく統計モデルが用いられるが，データが単一の正規分布に従うと仮定しているために，対象事象が非定常に変化するような複雑なパターンをもつデータに対しては，その有効性が制限される（ALBERT Inc）．他方，データ間の距離に基づいて分類・回帰ができる機械学習，例えば，K 近傍法（Altman, 1992），One-Class Support Vector Machine（以下，「OCSVM」という）（Schoelkopf et al., 1999），また，深層ニューラルネットワーク（Hinton et. al., 2006, LeCun et al., 2015）に異常検知の機能を付加した手法などを用いることで，非線形性が強い複雑なパターンを有するデータに対しても，異常検知が可能となる．人為的な操作が絡み複雑事象を伴うような製造過程で生じる製品異常や機械工作で起こるエラーなどを検知する場合に，OCSVM を使って異常検知を行った事例がある（平内ら 2019）．また，フィールドで観測される大量のデータにおいても，自然現象の不確実性に起因する複雑なパターンを有するデータが多く見られ，上記のような機械学習の導入が適していると考えられる（福島ら，2017）．

フィールド観測で収集された河川洪水の水位データについて，深層ニューラルネットワークを用いて，予測値と正常値のギャップの判断から，センサー異常や観測地点の環境変化で生じるスパイクノイズなどの異常検知に成功した既往研究がある（一言ら，2019）．しかし，低平地の排水管理では，排水機場調整池で観測される水位データについて，異常検知に機械学習を導入した事例は未だない．その理由は，豪雨などの自然現象による不確実性に加え，人為操作を伴う排水機場のポンプ運転で，より複雑なパターンを有する水位データになるからである．従って，水位変化が複雑になり観測データについて，単純なルールベースの正常・異常の判断が難しくなる．

データに品質保証を与えるためには，異常検知のみならず，その修正も必要である．観測データには，異常値に加えて，センサーやデータ送信システムの故障による欠損データも多く含まれる．そのデータの修正方法は，線形補間や多重代入法（Graham et al., 2007）などの手法が使われてきた．しかし，データのパターンを学習した深層ニューラルネットワークを利用して，より現実的な修正値を与える方法を提案する既往研究は，筆者らが調べる限りあまり見られない．例えば，加藤ら（2019）は，オートエンコーダを使って，賃貸物件データベースの欠損データを補完する手法を提案しているが，水位のような時系列データの欠損値を推定するものではない．

本研究は，排水機場調整池で観測され，スパイクノイズとスライドずれの異常値を含む水位データについて，正解データが不要の教師なし機械学習である OCSVM の利用と水位データの特性に関する条件を考慮し，異常値の検知手法を提案する．OCSVM で異常値とされたスライドずれについて，簡易的な修正方法も併せて提案する．スパイクノイズと判断された異常値は欠損値として扱い，観測データに元から含まれる欠損値と共に，深層ニューラルネットワークから得られる予測値を代入する．この導入される深層ニューラルネットワークについて，Hochreiter and Schmidhuber（1997）によって開発され，時系列データなどの連続性と長期傾向の記憶を保持し機械学習を行うことができる Long Short-Term Memory（LSTM）アーキテクチャ（以下，「LSTM モデル」という）を用いる．

２．方法

2.1　データ取得

本研究の対象地は，常時排水管理を行う低平農地である．この低平農地の概略を **Fig. 1** に示す．排水機場の調整池で観測され，欠損値や異常値が含まれる水位データを対象にする．1 時間間隔で観測された約 68,000 組の雨量（近隣のアメダス），水位，水位差，そして排水流量の各データを **Fig. 2** に示す．なお，水位データは，対象地を明確にできないため，敢えて観測データの最小値をゼロにした相対値を用いる．また，水位差は現時刻 t の値から 1 ステップ前の $t-1$ の値を差分して計算されたものである．観測データに含まれる異常値は，スパイクノイズ 3 箇所とスライドずれ 1 箇所である．具体例として，水位データに含まれる異常値（スライドずれ，スパイクノイズ），及び欠損値の事例を **Fig. 3** に示す．

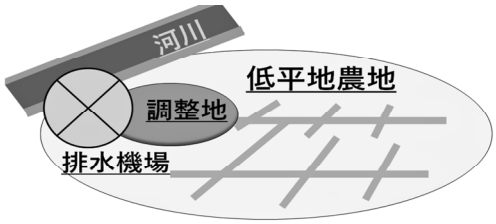


Fig. 1　**低平農地**の概略図

2.2　One-Class SVM（OCSVM）

本研究では，分類や回帰で用いられパターン認識が可能なサポートベクタマシン（以下，「SVM」という）（Vapnik and Lerner,1963; Vapnik, 2000）の発展型を利用する．正常値と異常値のラベルを用意して学習する方法を採用せずに，観測データを一度に読み込んで，異常の可能性がある外れ値を直接判別できる OCSVM を採用した．これは，SVM の一形態であり教師なし異常値の検知でしばしば利用されるものである．例えば，2 つの分類に分けられるデータの集合体（クラス）を考える．SVM では，2 つのクラス間を分ける線（決定境界）から一番近いデータまでの距離（マージン）が最大になるように決定境界の位置を決めるものである．クラスの分類について，次式で与えられる線形判別関数（f）を用いる．

$$f(x_1, x_2) = w_1 x_1 + w_2 x_2 - b, \qquad (1)$$

ここで，w_i は重み付き係数，x_i は変数（i=1, 2），b はバイアス（$b>0$）である．他のクラスのデータと最も近いところにあるデータをサポートベクターと呼び，2 つのクラスのサポートベクターと決定境界の間の距離（つまり，次式で示されるマージン）を最大化するように，f（x_1, x_2）を決める．

$$2/\|\boldsymbol{w}\| = 2/\sqrt{w_1^2 + w_2^2}. \qquad (2)$$

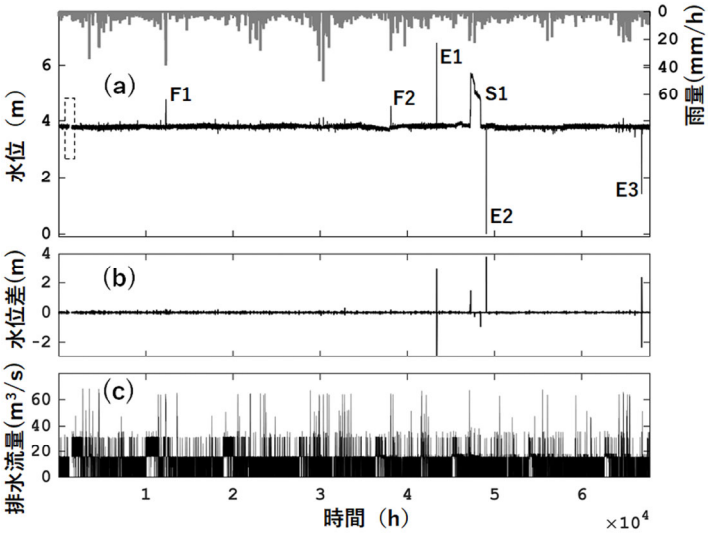


Fig. 2　時系列データ（雨量，水位，水位差，排出流量），図中の記号は異常検知の事例で扱う水位（正常値含む）．図中記号の E, F, S は，エラー，洪水イベント，スライドずれを示す．また，長方形ボックスは欠損値の区間を示す．

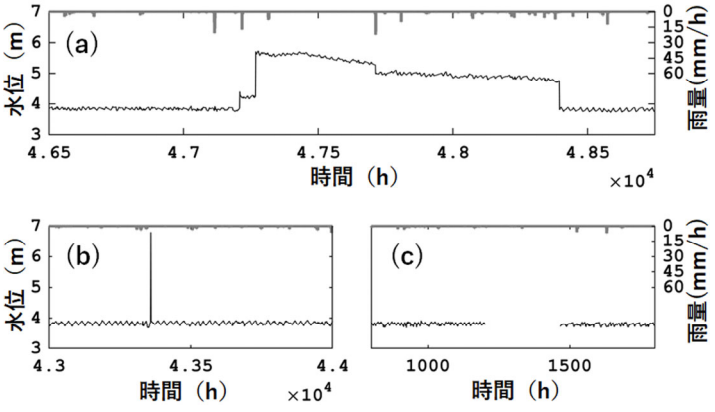


Fig. 3　異常値と欠損値の実例：(a) スライドずれ（Fig. 2a の S1），(b) スパイクノイズ（Fig. 2a の E1），(c) 欠損値（Fig. 2a の長方形ボックスの区間）．

OCSVM は，教師なし学習を行う機械学習なので，クラスの指定をせずに，データ集合体の中心部からの距離に応じて分類を行い，対象のサンプルを最小領域に押し込める境界面（以下，「超平面」という）を探すものである．この時，マージンは，原点からサポートベクターが乗る線上までの距離となる．データ集合体からの外れ値ほど原点の近くになるように高次元空間（ϕ）へ写像を行うならば，f とマージンは，$\boldsymbol{w}, \boldsymbol{x}$ のベクトル表記を用いて次式のようになる．

$$f\big(\phi(x_I), \phi(x_2)\big) = w_1\phi(x_I) + w_2\phi(x_2) - b = \boldsymbol{w^T}\phi(\boldsymbol{x}) - b, \quad (3)$$

$$b/\|\boldsymbol{w}\| = b/\sqrt{w_1^2 + w_2^2}, \quad (4)$$

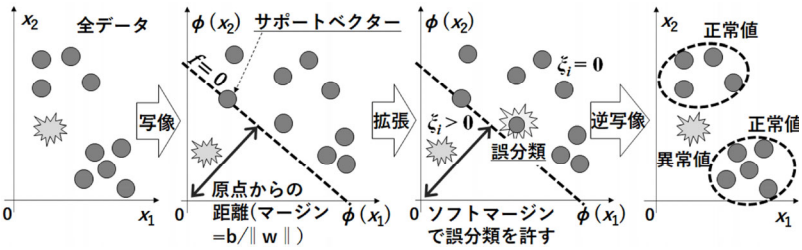
ここで，f=0 の時のマージン最大化は，計算の便宜上，$0.5\|\boldsymbol{w}\|^2 - b$を最小化することと同義と見なせる．さらに，マージンから拡張し，誤分類を許容するソフトマージンを持たせるためにスラック変数 ξ_iを導入することで，次式のような最適化問題として定義される．

$$\min_{w,\xi,b}(0.5\|\boldsymbol{w}\|^2 + \textstyle\sum_{i=1}^{n}\xi_i/(n\mu) - b), \quad (5)$$

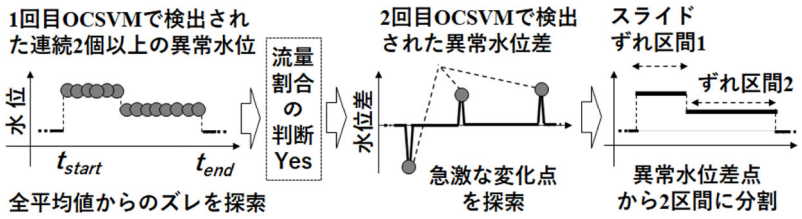
ここで，n はデータ数，μ はデータ数に対する異常値数の割合である．ただし，式（5）は任意の i について，以下の不等式を満たす．

$$\boldsymbol{w}^T\phi(\boldsymbol{x}) - b + \xi_i \geq 0,\ \xi_i \geq 0. \quad (6a,b)$$

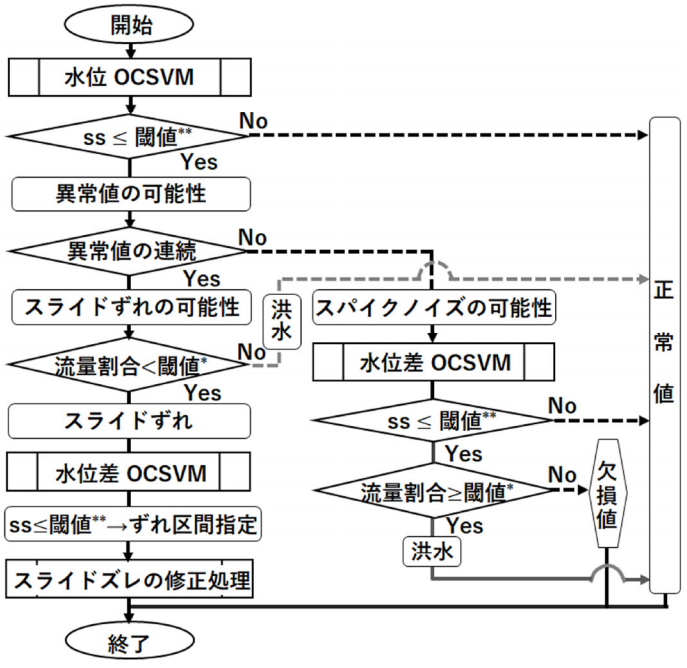
ϕ は陰関数なので，式 (5) の最適化問題は，ラグランジュの未定乗数法を用いて，乗数 α_i を導入すれば，次式のように，α_i に対して最大化する二次計画問題に帰結され，α_i を求めることができる．



(a) OCSVM を用いた異常検出の方法



(b)スライドずれの検出方法



(c) 異常検知のフローチャート（閾値*＝0.8，閾値**＝0.1）

Fig. 4　異常検知手法の概略図

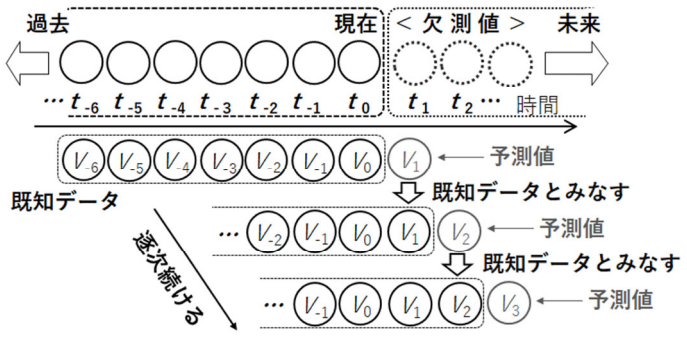


Fig. 5　LSTM モデルを使った欠損値の修正方法

$$\max_{\alpha}\left(0.5\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(x_i,x_j)\right), \tag{7}$$

ここで，カーネル関数 $K(x_i,x_j)=\phi(x_i)^T\cdot\phi(x_j)$ と定義され，写像したデータの内積である．また，α_i は二重変数である．式（7）は任意の i について，次式を満たす．

$$\sum_{i=1}^{n}\alpha_i=1,\ 0\le\alpha_i\le 1/(n\mu). \tag{8a,b}$$

本研究では，カーネル関数として次式のガウシアンカーネルを利用する．

$$K(x_i,x_j)=e\exp\left(-\|x_i-x_j\|^2/\sigma^2\right), \tag{9}$$

ここで，σ は ガウスカーネル幅で，境界の滑らかさを制御するものである．

$f(\boldsymbol{x})$ が 0 より小さい場合には，$\boldsymbol{x}$ は外れ値を示し，$f(\boldsymbol{x})$の値が大きいほど，データ密度が高いことを示す．これらの写像変換に関する概念図は，**Fig. 4a** で示される．

本研究で使用する OCSVM のモジュールは，python 機械学習ライブラリーの一種である scikit-learn（version : 0.23.2，URL : https://scikit-learn.org/stable/index.html）のオープンソースソフトウェアを利用する．このモジュールでは，超平面までの距離の決定に用いられ，異常度を示すスコアリング関数である score_sample（以下，「ss」という）を定義している．なお，本研究の異常度は，異常値から正常値までの範囲を 0≤ss≤1 で表すように規格化された．本研究では，OCSVM を水位・水位差データに実装する際に，ss の範囲と異常値の検出割合を考慮して，異常値と正常値を区別する閾値を 0.1 と設定した．この閾値が小さ過ぎれば，明らかな異常値以外は，異常検知が困難になる．一方，閾値が大き過ぎれば，ポンプの常時運転などで生じる比較的小さな水位変化でも異常値として検知される．式（5）の μ は，利用するデータの異常値個数を考慮して，0.02 を設定する．また，式（9）の σ^2 について，入力データから計算される分散値を採用する．

2.3　2 段階 OCSVM と連続性・排水流量条件

本研究で使用する水位データの特徴から，ss の閾値を人為的に操作して異常値の範疇を決める必要があるので，1 回目 OCSVM の実行のみでは異常検知が困難である．例えば，ss の閾値を厳しく設定すれば，洪水イベントを異常値として判別することもあり得る．これを改善するために，OCSVM を 2 回実行し（以下,「2 段階 OCSVM」という）並びに排水流量と異常値の連続性の条件付き判別を導入した（**Fig. 4b**）．最初に，1 回目 OCSVM で明らかな正常値を判別する．次に，1 回目の OCSVM で異常値と判別されたデータに対して，スパイクノイズとスライドずれの判別を行う．異常値の連続性を確認し，2 個以上連続する場合はスライドずれとする．この理由は，スライドずれの区間を水位差の変化で検索するために，異常値が少なくとも 2 個連続することが必要である．この時，スライドずれのほかに，豪雨時に発生する大きな規模の洪水ピークも含まれることがある．これを除外するために，ポンプ使用率を間接的に示す最大排水流量に対する実排水流量の割合（以下,「流量割合」という）を利用し洪水イベントかどうかを判別する．本対象地域のポンプ使用率が 8 割以上（つまり，流量割合 ≥ 0.8）になれば，大規模の洪水対策を実施しているので，スライドずれとの区別が可能である．最後に，洪水ピークが除外された上で, スライドずれの区間を指定するために, 水位差データを用いて 2 回目の OCSVM を実行する．この時，水位差が極端な場合を抽出できるので，スライドずれの区間が明確にできる．その後，区間毎に上下に区間を動かして修正を行う．

一方，異常値が不連続の場合は，スパイクノイズの可能性がある．しかし，通常範囲の水位変化も含まれることもあるので，水位差の極端事象を抽出するために，水位差データを使った 2 回目の OCSVM

を実行する．この時に，極端な水位変化の場合にスパイクノイズと判別され，欠測値として処理される．ただし，データ間隔が 1 時間と短いので，その前後の時間ステップで洪水ピークが発生することは考えられないが，念のために洪水ピークを除外するために，実際の流量割合条件を加えた．

Table 1　LSTM のハイパーパラメータなどの設定

パラメータなど	値・変数・式
バッチサイズ	100
エポック数	100
損失関数	二乗和誤差 (E= $0.5\sum_{i=1}^{N}(V_{o,i}-V_{c,i})^2$ *
オプティマイザー	確率的勾配降下法（SGD）
学習率	0.01
中間層（ノード数）	1 層（77 個）
入力変数	調整池の水位
出力変数	同上

*V_o = 観測データ, V_c = 予測計算データ

2.4　スライドずれの修正

上記 2.3 で検知された水位データ（V_o）のスライドずれについて，その修正方法は次のように行う．スライドずれの区間のうち，水位の最大値と最小値から求められる中間の値を全データ区間の平均値の位置に移動させる．つまり，次式のようになる．

$$0.5(V_{o,max}-V_{o,min}) \rightarrow \sum_{i=1}^{N} V_{o,i}/N \ , \qquad (10)$$

ここで，$V_{o,max}$, $V_{o,min}$ はスライドずれの区間の最大値と最小値，N は 全区間のデータ数，V_o は 水位の観測データである．なお，各区間のデータは上の方向にスライドしているが，スライドした高さを除けば，正常値であると仮定している．

2.5　Long Short-Term Memory (LSTM)

本研究では，木村ら（2019）が水位予測を行うために開発した LSTM モデルを利用する（詳細は，木村ら（2019）を参照されたい）．入出力は水位データのみで，入力は現時点から 6 ステップ前までの計 7 個のデータを使用し，また，出力は 1 ステップ後のデータとする．モデル精度の検証について，まず，異常値を取り除いた正常データのみを利用して，LSTM モデルを学習させた上で，1 ステップ後の予測精度を検証する．なお，モデルの学習時に，ある試行で 6 ステップ前までのデータに欠損値が含まれていれば，その試行はスキップされる（つまり，欠損値を含むデータ列の特徴は，学習されない）．全データのうち，時刻が新しい 100 個のデータを検証用，それ以外を学習用にして，LSTM モデルの検証を行った．予測値と観測値から求められる平均平方二乗誤差（RMSE）は，0.043 m となり，異常値を除いた観測値の水位差（最大値－最小値）に対して，約 4%の割合であった．このことから，LSTM モデルの予測精度は妥当であると考えられる．この時のハイパーパラメータなどの設定条件を **Table 1** に示す．欠測値の推定方法について説明する．**Fig. 5** に示すように，まず欠測位置の値を LSTM モデルを用いて推定する．次に，その推定値を既往データの観測値とみなし，次の時間ステップの推定を行う時に利用する．この手順を時間ステップ毎に繰り返す．

３．結果と考察

本異常検知のプロセス（**Fig. 4**）において， 1 回目の OCSVM では，ss の閾値（0.1）の緩い条件で異常値を判別したので，明らかな異常値は検知が可能なものの，洪水イベントのような本来正常と判別されるべきデータも異常と判断されている．それらの誤判断を訂正するために，水位差データで行う 2 回

目の OCSVM，流量割合，異常値の連続性の条件を導入した．

3.1　スパイクノイズ検知の成功例

スパイクノイズの検知について，2 段階 OCSVM 並びに流量割合や連続性条件で判別した後に成功した事例を **Fig. 6** に示す．例えば，**Fig. 6a** において，水位データを用いた 1 回目 OCSVM で異常値を検知できた（△印のピーク）．水位差データを用いた 2 回目の OCSVM でも異常値（水位差の変化が大きい）と判断され，かつ，流量割合が 0.8 以下（洪水イベントなし）だったので，その値はスパイクノイズとして認識された．この判断によって，異常値を取り除き，欠損値として処理する．**Fig. 6b** についても同様にスパイクノイズと判断され，その後，欠損値として処理された．

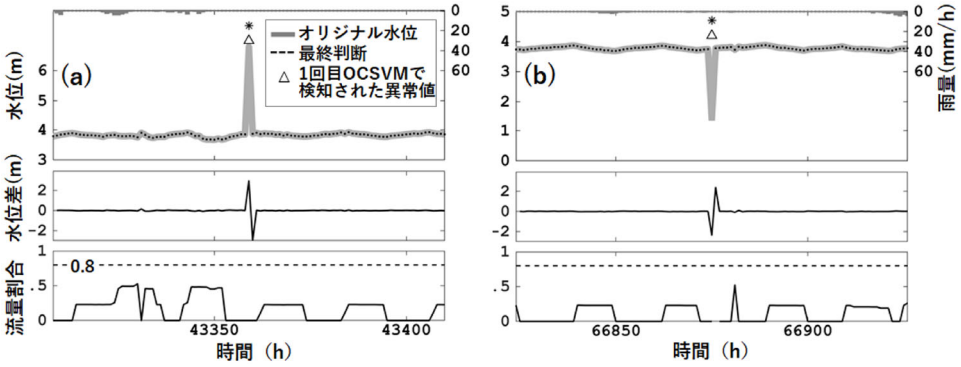


Fig. 6　スパイクノイズの検知成功例：(a) Fig. 2a の E1，(b) Fig. 2a の E3．「*」は最終的に異常値と判断されたデータを意味する．

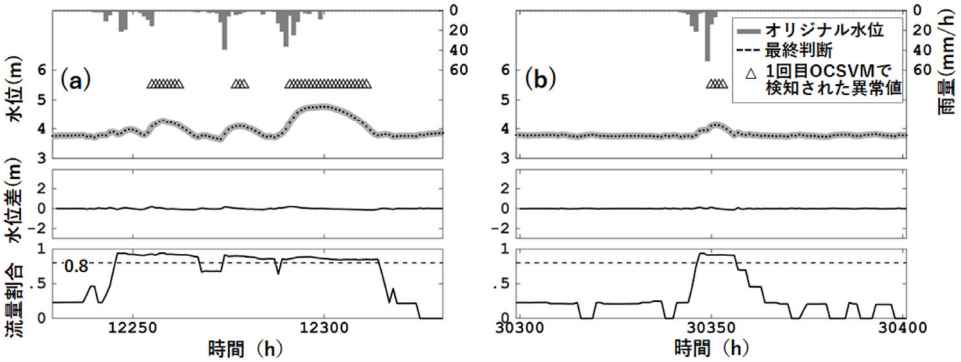


Fig. 7　洪水時の流量割合により最終的に正常値になった例：(a) Fig. 2a の F1，(b) Fig. 2a の F2．

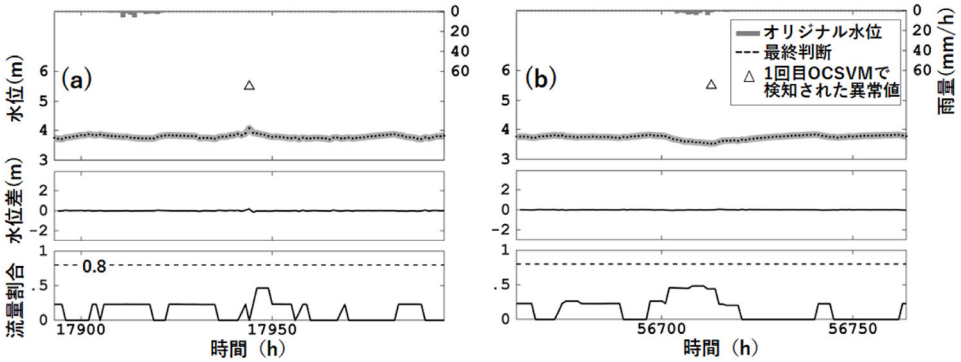


Fig. 8　水位差の判別により最終的に正常値になった例．

3.2　1 回目 OCSVM の異常値の判断から正常値に訂正された例

1 回目の OCSVM で異常値と判断したものの，実際には正常値である場合に，流量割合の条件と水位差による 2 回目の OCSVM で訂正された事

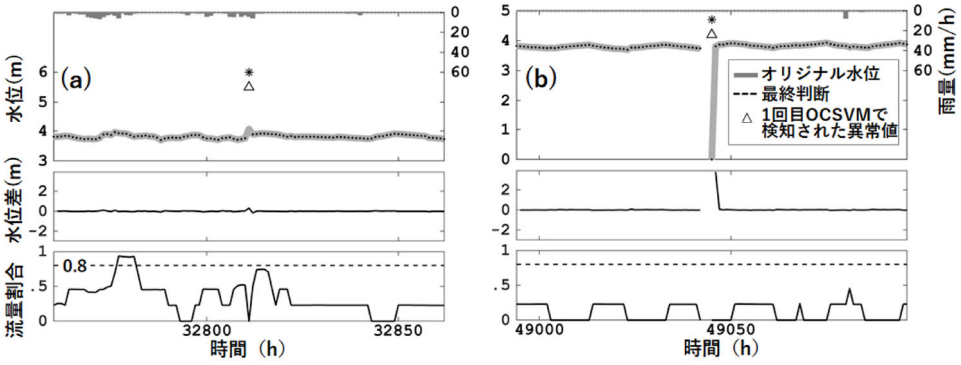


Fig. 9　異常検知の失敗例．「*」の意味は，Fig. 6 を参照．

例について検討する．洪水イベント期間の水位が 1 回目 OCSVM で連続した異常値（スライドずれ）と判断される．しかし，流量割合が 0.8 以上となる条件（ほぼ最大流量で河川に排出）を利用すれば，洪水イベント期間に異常値として判断される点（時刻）では，正常値として判断される（**Fig. 7**）．ただし，小規模の洪水イベントでは，必ずしも流量割合の条件のみでは，判断できない場合が考えられる．

さらに，1 回目 OCSVM で単独（1 点）の異常値（スパイクノイズ）と判断された場合に，水位差の極端な変化を捉えることができる 2 回目 OCSVM で，ss > 0.1 であれば，この場合も正常値として判断される（**Fig. 8**）．

3.3 異常検知の失敗例

Fig. 9 に異常検知の失敗例を示す．**Fig. 9a** は小規模な洪水イベントの場合を示しており，流量割合の条件が 0.8 以下，及び 2 回目 OCSVM で ss≤0.1 となり，最終判断で異常値のまま，正常値に訂正できていない．また，スパイクノイズの 1 ステップ前の値が，欠損値の場合（**Fig. 2a** の E2）に，異常値が連続しているのかどうかを判断できない．従って，スパイクノイズともスライドずれとも判断されなかった（**Fig. 9b**）．

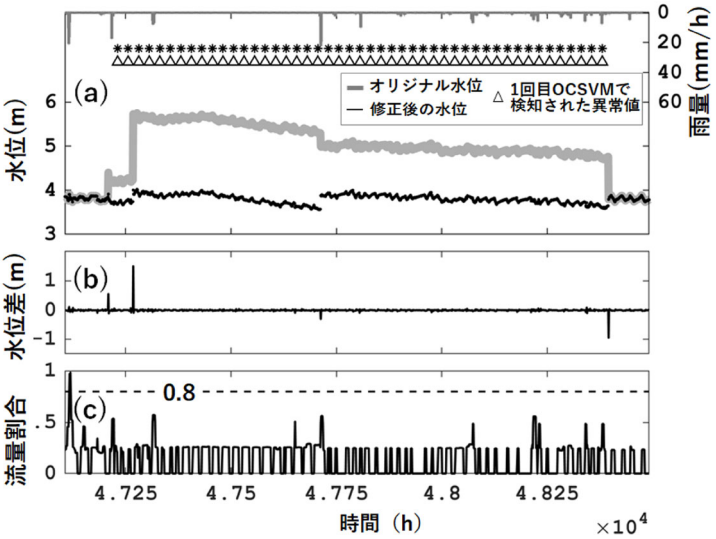


Fig. 10　スライドずれの検出とその修正：Fig. 2a の S1．「*」の意味は，Fig. 6 を参照．

3.4　スライドずれの検知と修正

Fig 4b に示す 1 回目 OCSVM の実行で，ss ≦ 0.1 の時，異常値と判断され，さらに異常値が連続し，かつ洪水イベント区間ではない場合（流量割合＜0.8）に，スライドずれの可能性がある．スライドずれの区間を検索するために，水位差を用いて 2 回目 OCSVM の実行を行い，極端に水位が変化する場合（ss ≦ 0.1）に，スライドずれを検知できる．この事例を **Fig. 10** に示す．オリジナル水位は 3 つのスライドずれ区間に分けられる．それぞれの区間の水位は，式（10）を使って修正される（**Fig. 10a** の細い黒線）．修正の結果，各区間の接続部分に多少のギャップが見られ，特に第 2 番目のスライドずれ区間は，減少傾向が見られる．より正確な接続部分の修正を行う場合には，スライドずれ区間に対してマイナス勾配のバイアスを考慮する必要がある．

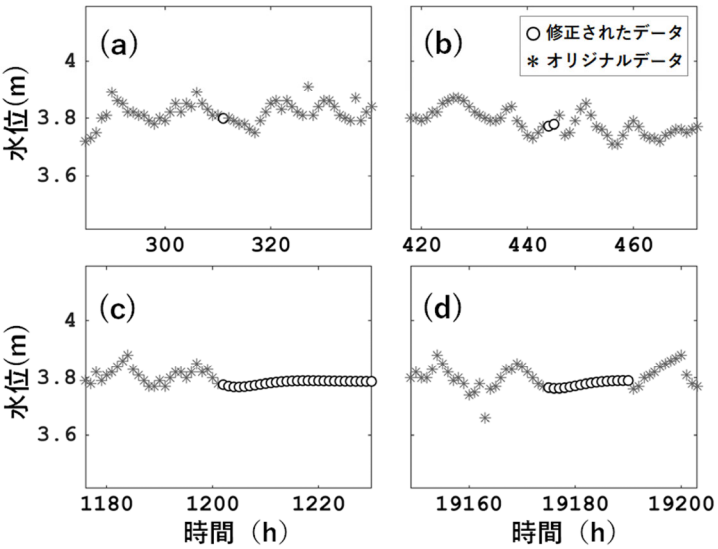


Fig. 11　欠損値の修正（代表的な例）．パネル (C) の推定値の右側は一定値が続くので省略した．

3.5 LSTM を利用した欠損値の修正

予め異常値を取り除いた正常値のみの水位データを用いて事前学習させた LSTM モ

デルを使って，欠損値を推定した．代表的な欠損値の修正事例を **Fig. 11** に示す．欠損値が，1～2 点の場合の修正結果は，単純な 2 点間の線形補間とほぼ違いは見られない（**Fig. 11a, b**）．しかし，10 点以上の連続な欠損値の修正結果は，線形補間に比べて，小さな幅であるものの水位変化を推定できた（**Fig. 11c, d**）．LSTM モデルの推定値を既知のデータとして取り込みながら，時間ステップ毎に逐次推定を長く行えば，最終的に水位変化がフラットになる（**Fig. 11c**）．この理由は，次の通りである．逐次推定に現れる水位変化が小さければ，次のステップの推定で使われる既往データにも小さな水位変化が追加される．これを繰り返すことで LSTM モデルは，水位変化の幅が徐々に小さくなると記憶するからである．これは，本研究で試行した LSTM モデルを用いる逐次推定手法の弱点である．

3.6　異常検知手法の適用範囲と課題

本研究では，水位・水位差の 2 段階 OCSVM と異常値の連続性・流量割合の条件を組み合わせて異常検知を行う本手法を，排水機場調整池で観測された水位データに初めて適用した．水位データには，洪水時の急激な水位変化も含まれ，異常値との区別が容易にできない場合もあり，**Fig. 9** のような検知ミスはあったが，概ね異常値の検知に成功したと考えられる．しかし，OCSVM の ss の閾値（0.1）は人為的に調整する必要があり，また，流量割合の閾値も，本排水機場のポンプ運転のルールによって決めたものなので，それぞれの閾値間の調整が必要である．従って，本手法を他の排水機場のデータに適用するためには，上記 2 つの閾値の調整が必須である．本手法の汎用化テストは今後の課題である．

４．まとめ

本研究では，低平地の排水機場調整池で観測された水位データに対して，水位・水位差の 2 段階 OCSVM と連続性・流量割合の条件を組み合わせた異常検知を行い，異常値（欠損値含む）の修正を行った．以下に，結果をまとめる．

- スパイクノイズは，非連続性の異常値，かつ 2 回目 OCSVM の実行で水位変化が極端と判断され，さらに流量割合が閾値よりも小さい場合に検知された．
- スライドずれは，連続な異常値，かつ 2 回目 OCSVM の実行による水位変化でスライドずれの区間が定まった場合に，簡易的な方法で修正された．
- 正常値を異常値と判断したり，直前の欠損のためにスパイクノイズが検知できなかったりする失敗例が一部見られた．
- LSTM を用いた推定値を既知データに組み込む方法で，逐次欠損値の修正を行い，2 点間の線形補間に比べて，多少なりとも水位の変化を表すような推定ができた．

謝辞：本研究は，（独）環境再生保全機構の環境研究総合推進費（JPMEERF20S11803）により実施させて頂いた．また，研究手法の助言を頂いた（株）アーク情報システムの木野由也氏，並びに本研究で利用したデータの一部を提供して頂いた団体に感謝を申し上げる．

引用文献

Altman, N. S. (1992): An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), pp.175-185.

LeCun, Y., Bengio, Y., and Hinton, G. (2015): Deep Learning. Nature, 521(7553), pp.436-444.

Geisser, S. (1993): Predictive inference: An introduction, Monographs on statistics and applied probability. 55, Chapman and Hall, NY (USA), 240p.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007): How many inputations are really needed? Some practical

clarificatons of multiple imputation theory. Prevention Science, 8(3), pp.206-213.
Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006): A fast learning algorithm for deep belief nets Neural Computation, 18(7), pp.1527-54.
Hochreiter, P. and Schmidhuber, J. (1997): Long short-term memory. Neural Computation, 9(8), pp.1735-1780.
Hotelling, H. (1931): The generalization of Student's ratio. Annals of Mathematical Statistics, 2(3), pp.360-378.
Schölkopf, B., Plattz, J. C., Shawe-Taylory, J., Smolax, A. J., and Williamson, R. C. (2001): Estimating the support of a high-dimensional distribution. Neural Computation, 13, pp.1443-1471.
Vapnik, V. N. (2000): The nature of statistical learning theory (2nd Ed). Springer, New York, 314p.
Vapnik, V. N. and Lerner, A. (1963): Pattern recognition using generalized portrait method. Automation and Remote Control, 24, pp. 774-780.
ALBERT Inc.：データ分析基礎知識（マシンラーニング>異常検知の基礎>時系列データに対する異常検知）
https://www.albert2005.co.jp/knowledge/machine_learning/anomaly_detection_basics/anomaly_detection_time
（確認日：2020/12/06）
加藤暢之・新妻弘崇・太田学（2019）：ニューラルネットワークによる賃貸物件データセットの欠損値補完の一手法. 第 11 回データ工学と情報マネジメントに関するフォーラム, H6-1.
木村延明・中田達・桐博英・関島建志・安瀬地一作・吉永育生・馬場大地（2019）：LSTM モデルを用いた低平地排水機場の水位予測，土木学会論文集 B1（水工学），75(2)，pp.I_139-I_144.
一言正之・川越典子・橋田創・清雄一・房前和朋（2019）：水位推定誤差の確率分布に基づく河川水位観測データのリアルタイム異常検知，土木学会論文集 B1(水工学)，75(2)，pp.I_193-I_198.
平内和樹・倉元昭季・瀬尾明彦（2019）：One-Class Support Vector Machine を用いた異常検知手法による作業性の評価 ― 身体近傍でのねじ締め作業への適用，人間工学，55(2)，pp.50-58.
福島俊一・藤巻遼平・岡野原大輔・杉山将（2017）：ビッグデータ×機械学習の展望　最先端の技術的チャレンジと広がる応用．情報管理 60(8), pp.543-554.