Programming by Exampleに基づくデータ補完手法APREP-Sにおける連続した欠損値を対象とした検証

永島 寛子[1,a)]　加藤 由花[1,b)]

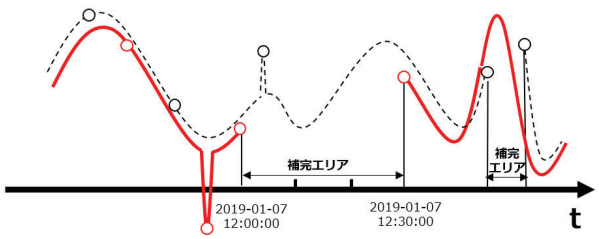
概要：近年，センサーやウェアラブルデバイスなどデータソースの種類の増加に伴い，分析に利用可能なデータの量が増えてきた．収集データを分析モデルに入力するためには，外れ値や欠損値の対処やセンサーの測定単位や表記の違いの統一など，分析者による「前処理」が必要である．しかしながら，分析者が前処理に費やす時間は分析フロー全体の 80 %と言われており，前処理は分析者の大きな負荷となっている．そのため，分析者の負荷を減らすための手法として，Programming by Example アプローチをベースとし，前処理で行われる処理のうち外れ値と欠損値をベイズ推論により自動補完する手法 “APREP-S（Automated PRE-Processing for Sensor data）” を提案してきた．APREP-S は，複数のデータ補完手法をあらかじめ定義しておくことにより，それぞれの手法の適正度を補完箇所ごとに出力する手法である．これまでの検証は補完対象箇所が 1 箇所ずつの点であることを前提とした方法となっていた．したがって本稿では補完対象が区間を対象としたエリア補完の場合を想定し，APREP-S 内の補完手法に時系列解析の一般化加法モデルと LSTM を定義した．APREP-S と既存手法の補完値の推測精度を比較し，区間のエリア補完に対しても APREP-S が有効であることを評価した．

1. はじめに

近年，データ分析で扱うデータは，基幹システムに蓄積されたデータだけではなく，センサーやウェアラブルデバイスなどのデータも対象となっており，分析に利用可能なデータの量と種類が増えてきた．活用分野も，ロボットの自律的行動，顧客の動向分析，農業の作物栽培支援など，多岐に渡る．しかしながら，収集データは欠損値や外れ値，センサーやウェアラブルデバイスの種類による測定単位や表記の違いなどを含んでおり，そのまま分析モデルに入力してしまうと，正しい結果が得られない [1]．とくにセンサーデータの場合，データの取得にネットワークを介するため，ネットワークの負荷状況によりデータが遅延する可能性や，転送時にタイムアウトや途中でデータをロストする可能性がある．さらに，電池で稼働しているセンサーならば，電池切れにより一定期間のデータを取得できない場合も起こり得る．また，データ分析のため，異なる種類のセンサーのデータを結合する場合には，センサーによる計測間隔や単位の違いや，センサーの計測誤差を考慮する必要がある．データ分析を行うためには，分析モデルに入力する前に，これらの外れ値や欠損値，計測間隔の違いを整えるための処理，すなわち「前処理」を行う必要があるが，この前処理にデータ分析処理の 80%のリソースが費やされている [2]．そのため，分析者の負担を減らす方法が求められている．

分析者の負担を減らす方法のひとつとして，前処理を自動化する方法がある．自動化の方法には，機械学習を利用する方法の他に，平均値代入法やスプライン補間など代入する値を 1 つのルールにより定める「単一代入法」や，データ集合からランダムに抽出したデータを入力とした異なる代入処理のシミュレーションを複数回行うことにより代入する値を定める「多重代入法」などがある [3]．しかしながら，自動で補完する方法は，1) データ品質が悪いと精度が悪くなる [1]， 2) 全体が汎化されるため補完したい箇所以外にも影響がある，という課題がある．これらの課題を解決するためには，人の知識が不可欠である．そのために，機械学習をベースとすることにより分析者の負担を最小限に留め，Programming by Example（PBE）アプローチにより人の知識を機械学習と融合させる手法として「APREP-S（Automated PRE-Processing for Sensor data）」を提案してきた [4]．当該研究では欠損値・外れ値などの補完箇所が連続していない，すなわち 1 箇所ずつを想定した検証になっているため，本稿では「APREP-S」が，補完対象が

[1] 東京女子大学
167–8585, 東京都杉並区善福寺 2-6-1
[a)] h18m001@cis.twcu.ac.jp
[b)] yuka@lab.twcu.ac.jp



APREP-S の出力例（一部）

Method	Proportion	val1	val2	val3
m_1 = Prophet	60%	25	27	26
m_2 = LSTM	10%	24	24	26
m_3 = スプライン補間	30%	26	28	26

図 1　欠損値・外れ値が連続しているデータを入力とした APREP-S のイメージ

区間であるエリア補完の場合にも有効であることを実験する．エリア補完のイメージを図 1 に示す．ある区間が外れ値や欠損値であるデータを入力とし，APREP-S に定義された手法それぞれの適正度と補完値を出力する．本稿はセンサーデータ分析を想定し，前処理の中でも外れ値・欠損値の対処を対象とする．

本稿の貢献は下記 2 点である．

- Programming by Example アプローチをベースとした APREP-S が，外れ値・欠損値の場所が連続している，すなわちある区間のデータがない場合のデータ補完手法としても有効であることを実験する．
- 本稿提案手法と既存の機械学習を含む自動補完手法と比較し，有効性を検証する．

本稿の構成は以下の通りである．まず 2 章で本稿の実験の関連研究について述べ，3 章で本稿の提案手法のベースである APREP-S について説明する．4 章で実験の概要を述べ，5 章では，提案手法と既存の手法の比較による評価を行い，6 章で本稿のまとめを行う．

2. 関連研究

2.1 スプライン補間

データ補完手法には，欠損データを削除する「リストワイズ法」，予め定めた 1 つの手法により取得したデータから代入値を一意に求める「単一代入法」，欠損データの分布を母集団としサンプリングしたデータから複数手法をシミュレーションを行うことにより代入値を求める「多重代入法」などがある．

単一代入法の一つに，スプライン補間がある．スプライン補間は，多数の等間隔のデータ点を通る滑らかな曲線を描くための手法である．指定された点を全て通る多項式であり，係数は点の位置を動かさず線の曲がり具合のみを変化させるため，各データ点の連続性は失われない．1 次スプライン補間は，データ点間を結ぶ式が 1 次式になるため折れ線となる．2 次以上のスプライン補間は曲線となり，例えば 3 次の多項式は以下の式で表わされる [5]．

$$S(x) = \begin{cases} s_1(x) & \text{if} \quad x_1 \le x < x_2 \\ s_2(x) & \text{if} \quad x_2 \le x < x_3 \\ \vdots & \\ s_{n-1}(x) & \text{if} \quad x_{n-1} \le x < x_n \end{cases} \tag{1}$$

$$s_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i \tag{2}$$

$i = 1, 2, ..., n-1$ である．スプライン補間をはじめとする単一代入法は，1 箇所の補完箇所に対する補完方法として有効であるが，データ点の間を滑らかに結ぶため，一定区間のデータ補完には適さない．

2.2 時系列解析（一般化加法モデル）による予測

複数の関数を足し合わせることにより非線形関数を表現する “一般化加法モデル” がある [6]．時系列解析を行うにあたり，人間の行動や季節性，時刻により変換するトレンドなどの非線形の傾向を周期性や変動点に当てはめる必要がある．Prophet は，一般化加法モデル（Generalized Additive Model, GAM）をベースとし，時系列データを予測するための回帰モデルである [7]．Prophet は以下の式で表され，f(·) が全て線形の場合は線形回帰と同じ式になる．

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{3}$$

$g(t)$ は時系列データの非周期性をモデル化するトレンド関数，$s(t)$ は季節などの周期的な変化を表す関数，$h(t)$ は休日の影響を示す関数であり，ϵ_t はモデルで対応できない特異な変化を示している．

2.3 機械学習（RNN）による予測

RNN（Recurrent Neural Network）は，ニューラルネットワークの出力を別のネットワークの入力値として利用する再帰的構造を持ったニューラルネットワークのことである．時系列データの予測において，入力データは全て独立ではなく，一連の流れとして考えることにより，精度を高める手法がある．RNN は，ニューラルネットワーク内の隠れ層の出力を，一般のニューラルネットワークの最後の層と同様に利用可能な出力とすることにより再帰性を持たせている．

現在広く使われている RNN 手法の一つとして，“Long Short Term Memory (LSTM)” がある．LSTM は，短い学習時間で長期的な時系列データを扱うことが可能な手法である [8]．

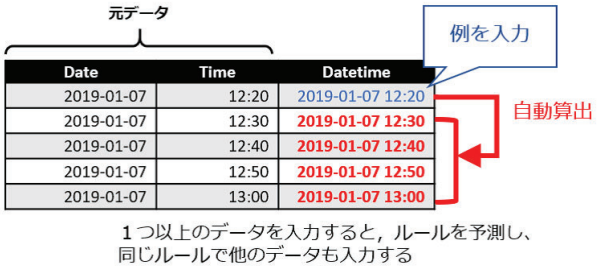


図 2　PBE により，*Date* 列と *Time* 列から *Datetime* 列を生成

2.4　PBE

PBE（Programming by Example）アプローチは，1 つ以上の入力と出力の組を元に変換ルールを予測し，自動で加工する手法によるアプローチである．近年，データの値に対する FlashFill[9] をはじめ，分析対象データ量の増加により，PBE アプローチはビッグデータのデータ変換手法として注目されている．PBE は 1) 分析者により入力された例と一致するルールを効率的に検索することができる検索アルゴリズム，2) 分析者により入力された例を満たす処理の中から最適な処理を選択するランク付け，3) ユーザビリティと使いやすさを促進するためのインタラクションモデル，という 3 つメイン処理を持つ手法である [2]．PBE の例を図 2 に示す．図 2 は，*Date* 列と *Time* 列を持つ表形式のデータに *Datetime* 列を追加するときの動きを表している．まず，分析者が *Datetime* 列の一番上の行に「2019-01-07 12:20」と入力する．分析者によるインタラクティブな入力を受け，PBE は入力されたデータを満たすルールを検索し，一致したルールの中から最適な処理を選択する．例えば，図 2 では「*Date* 列の後ろに *Time* 列を結合」「*Date* 列の後ろに文字列 “12:20” を追加」などのルールが考えられる．当該例では，予め定めた指標により「*Date* 列の後ろに *Time* 列を結合」というルールを最適と判断し，2 行目以降に同じルールでデータを入力している．このように，分析者は 1 番上の 1 行を入力することにより人の知識を入力することができ，それ以下の行は自動で補完されたデータを取得することができる．

3.　APREP-S

APREP-S（Automated PRE-Processing for Sensor data）は，Programming by Example アプローチをベースとし，分析者が補完したいデータを入力すると，APREP-S に定義されている補完手法に対する適正度と補完値ベイズ推論により算出する手法である [4]．ワークフローを図 3 に示す．APREP-S には複数のデータ補完手法が定義されており，補完対象箇所の特徴からどの補完手法が適しているかを推定する．入力は外れ値や欠損値を含むデータであり，出力は各補完対象箇所ごとに APREP-S に定義されたデータ補完手法の適正度と補完値である．これにより，分析者は APREP-S が算出した手法の適正度と補完値を確認しながら最適な手法を選択することが可能となる．APREP-S は，類似データを学習データとしてモデルを生成する「モデルトレーニングフェーズ」，モデルトレーニングフェーズで生成したモデルを使用して APREP-S に定義された手法の適正度と補完値を算出する「モデル運用フェーズ」がある．分析者がモデル運用フェーズで手法を 1 つ選択した後は，選択した手法を学習データとしモデルを更新するため，使い続けることにより精度の高いモデルとなる．

APREP-S に定義された手法の適正度は，ベイズ推論により算出する．APREP-S のモデルは 2 つのパラメータ α と $\boldsymbol{\beta}$ を持つソフトマックス関数である．

$$p(m_k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|m_k)p(m_k)}{\sum_{i=1}^{K} p(\boldsymbol{y}|m_i)p(m_i)} = \frac{\exp(\boldsymbol{y}(x_k))}{\sum_{i=1}^{K} \exp(\boldsymbol{y}(x_i))} \tag{4}$$

$$\boldsymbol{y}(x_q) = \alpha + \boldsymbol{\beta}\, x_q \quad (1 \le q \le Q) \tag{5}$$

x_q は正規化された特徴の値，$m_k \in M$ は APREP-S に定義した手法，α は APREP-S に定義した手法の数の配列，$\boldsymbol{\beta}$ は，$K \times Q$ の行列（K は特徴の数，Q は手法の数）を示す．

4.　実験の方針

本稿では，「APREP-S」を区間を対象としてエリア補完の場合でも有効であることを検証する．3 章で述べている通り，文献 [4] では欠損値・外れ値を PBE アプローチをベースに補完する方法を提案した．しかしながら，APREP-S で定義する手法に関しては「予め定義しておく」とのみ提案されており，どのような手法が良いか考慮されていない．評価方法も欠損値・外れ値をランダムで設定した単一の補完値に対する評価になっている．そのため，APREP-S と比較した既存のデータ補完手法も単一データ点のデータ補完に効果がある手法であった．したがって本稿では，より現実的なデータ補完が必要となる場面を想定し，ある区間が欠損値・外れ値がになっている場合，すなわち補完対象エリアが存在する場合にも有効であることを実験する．

4.1　APREP-S に定義する手法

ある区間が補完対象となるエリア補完の場合，点ではなくデータの流れを示す連続値による予想が必要になるため，単一欠損値とは異なる実験が必要である．具体的には，モデル生成部分は APREP-S のベイズ推論はそのまま利用し，APREP-S に定義する補完手法に時系列解析と機械学習を定義することにより，エリア補完の場合にも各手法の適正度を予測できることを評価する．

時系列解析と機械学習を定義した場合の学習データと補完エリアを図 4 に示す．時系列解析と機械学習の場合，予

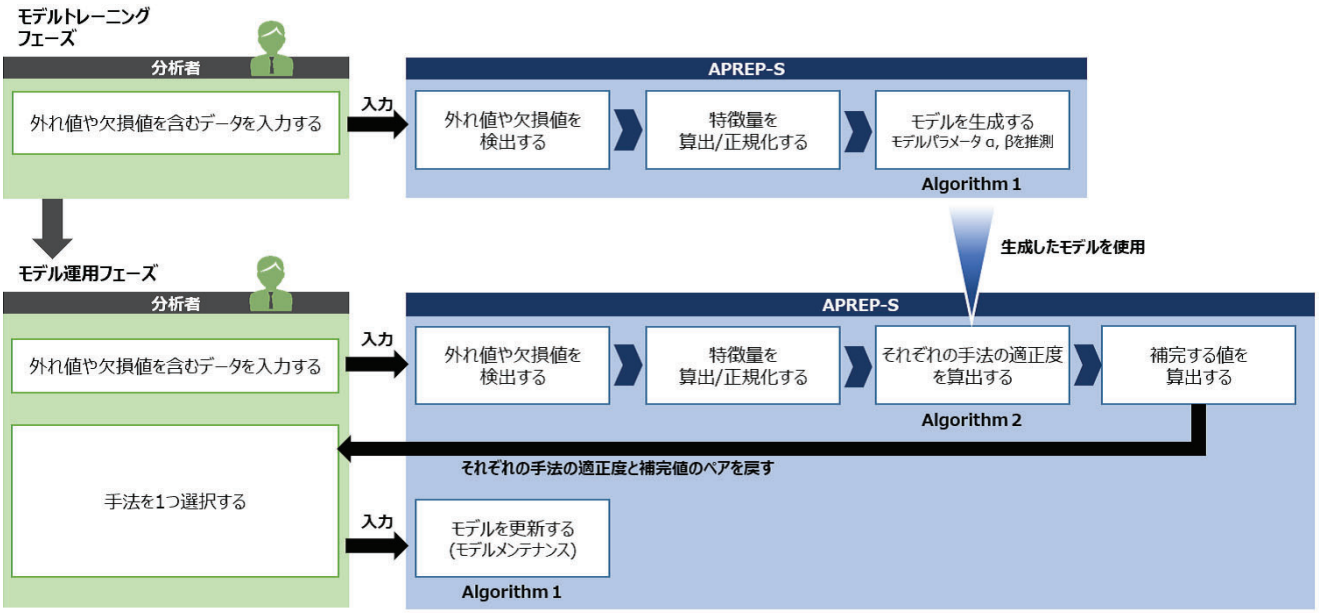


図 3 分析者と APREP-S のワークフロー：緑の箇所は分析者の作業を示し，青は APREP-S の作業を示している．（文献 [4] の図 2 を引用）

測を行うためにモデルを生成する必要がある．本稿では補完対象エリアごとに直前のデータを学習データとし，補完エリアごとにモデルを生成することとした．

時系列解析は，2.2 節で示した一般化加法モデルを採用する．式（3）で示した通り，非周期性を示すトレンド関数，周期性を示す関数，休日の影響を示す関数の和で表現される．使用するツールは，Python と R のライブラリの一つである Prophet をベースとし，facebook 社が OSS として公開している “fbprophet” とする [7][10]．非線形の傾向や分析に強く，外れ値や欠損値を含むデータをそのまま入力値として使用できる特徴がある．そのため，学習データに外れ値が含まれる場合もそのまま学習データとする．式（3）に対応する fbprophet の式は，

$$g(t) = \frac{C(t)}{1 + \exp(-(k + \boldsymbol{a}(t)^{\mathrm{T}}\boldsymbol{\delta})(t - (m + \boldsymbol{a}(t)^{\mathrm{T}}\boldsymbol{\gamma})))} \quad (6)$$

ただし，

$$\gamma_j = \left(s_j - m - \sum_{i<j} \gamma_l\right)\left(1 - \frac{k + \sum_{i<j} \delta_l}{k + \sum_{i\leq j} \delta_l}\right) \quad (7)$$

$$a_j(t) = \begin{cases} 1, & \text{if} \quad \mathrm{t} \leq \mathrm{s_j} \\ 0, & otherwise \end{cases} \quad (8)$$

とし，$C(t)$ は時間で変動する値，s_j はチェンジポイント，k は時刻 t でのベースとなる確率，δ は確率を調整するための変数である．

$$s(t) = \sum_{n-1}^{N}\left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right)\right) \quad (9)$$

ただし，a, b はパラメータ，P は周期の期間を示す．

$$h(t) = Z(t)\boldsymbol{\kappa} \quad (10)$$

$$Z(t) = [\mathbf{1}(t \in D_1, ..., \mathbf{1}(t \in D_L)] \quad (11)$$

となる．

機械学習は，2.3 節で示した時系列データを一連の流れとしてモデルを生成する “LSTM” を使用する．APREP-S は分析者とインタラクティブにデータ補完手法の選択とデータ補完モデルの更新を行う手法のため，短い時間でモデルを生成できる LSTM を採用した．fbprophet と同様に，学習データに外れ値が含まれる場合もそのまま学習データとする．

4.2 特徴量

本実験では補完対象データから算出できる特徴と補完対象データ以外の特徴を定める．補完対象データから算出できる特徴とは，例えば補完エリア前後の傾きや，補完エリアが存在する箇所など，補完対象データの値や傾向から抽出できる特徴である．一方，補完対象データ以外の特徴とは，別のデータセットから取得するデータであり，補完対象データと結合することにより関連性を持たせたデータである．特徴量は補完対象エリアにつき 1 つの補完方法ではなく，補完対象データのデータ間隔を同じ間隔ごとに特徴量を定める．それにより，複数の特徴量から導き出された最適手法は補完対象データのデータ間隔ごとに出力される．同じ手法がつながっているエリアは，補完エリアのサブエリアとし，該当する手法で 1 度に算出する．例えば，図 4 の 2 つ目の補完エリアにおいて，選択手法リストが (1,1,1,2,2) である場合，補完エリアの前半を method1，後

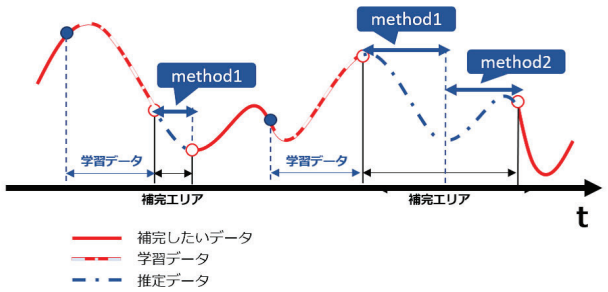


図 4　時系列解析と機械学習の学習データと補完エリア：赤い線が補完対象データ，赤い破線が学習データ，青い破線箇所が推定データを示す．補完エリアごとにモデルを生成し，生成したモデルによる予測を行う．

半を method2 で補完する．

4.3　実験データ

センサーデータの場合，人の流れや気温，加速度の変化など，時系列の状態で保持することが多いため，本実験でも時系列データを使用する．

補完対象データは，文献 [4] と同様にワイヤレスの気温と湿度センサー（DHT-22）のデータを含むデータセット [11] をオリジナルデータとし，外れ値と欠損値を挿入したデータを使用する．オリジナルデータセットは 29 列あり，計測時間，気温，湿度，気圧，風速などのデータが格納されている．気温と湿度は，キッチン，リビング，寝室，バスルームなど，室内外 9 箇所に設置されたセンサーにより収集している．本評価では，キッチンエリアの湿度データ $RH1$ とリビングエリアの湿度データ $RH2$ を選択し，モデルトレーニングフェーズの入力データを $RH2$，モデル運用フェーズの入力データを $RH1$ とする．データ件数は，137 日（4.5 ヶ月）のデータであり，各センサーごとに 19,735 件である．センサーからは 3.3 分ごとにデータが収集されるが，本データセットは 10 分ごとに集計したデータセットとなっている．また，センサー DHT-22 の湿度の誤差は ± 3% である．

$RH1$ と $RH2$ に挿入する欠損値と外れ値が起こる確率は指数分布

$$f(e) = \frac{1}{\epsilon}\exp\left(-\frac{e}{\epsilon}\right) \quad (500 \leq \epsilon \leq 1500) \tag{12}$$

に従うものとし，6 回に 1 回欠損値が生じるとした．外れ値の外れ度合いと，欠損値の場合の欠損箇所の連続数はガウス分布

$$f(e) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(e-org_i)^2}{2\sigma^2}\right\} \quad (0 \leq \sigma^2 \leq 10) \tag{13}$$

に従うとする．ここで，σ^2 はガウス分布の分散を示している．欠損値と外れ値のオリジナルデータと，欠損値と外れ値を挿入したデータを図 5 に示す．図 5 は，最初の 1 週間分のデータのみを対象としたグラフである．

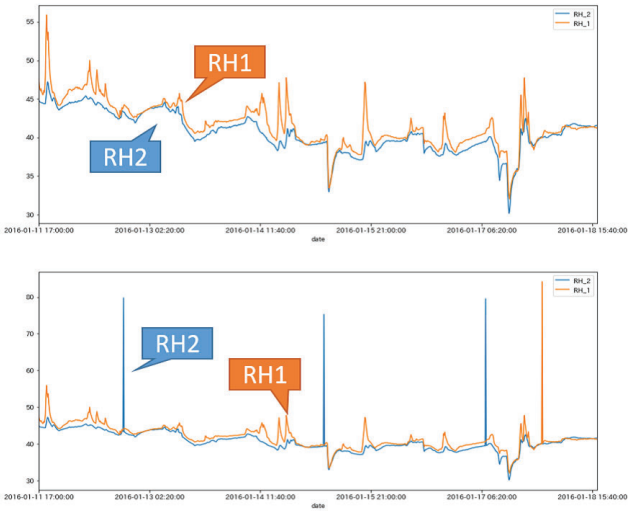


図 5　湿度（RH）データの 1 週間の時系列グラフ：オレンジが $RH1$，青を $RH2$ を示し，上のグラフはオリジナルデータ，下のグラフは欠損値と外れ値を挿入したデータ示す．

4.4　欠損値・外れ値検出

欠損値は，データが $Null$ のデータとする．

外れ値は，本稿では指数加重移動平均法

$$S_t = \alpha \times X_{t-1} + (1-\alpha) \times S_{t-1} \quad (t \geq 2) \tag{14}$$

S_1 は直前 n 日の平均，X_t は時刻 t での値，α は平滑定数（$0 \geq \alpha \geq 1$, $\alpha = 2/(n+1)$）で検出する．

4.5　評価方法

本評価では，モデル運用フェーズで出力した補完値と，既存の補完方法で求めた補完値を比較する．使用するデータは，オリジナルデータに外れ値と欠損値を挿入したデータを生成する．各手法の精度は，オリジナルデータと APREP-S または既存の補完方法で求めた値との二乗和誤差

$$ERR = \frac{1}{2}\sum_{n=1}^{N}(org_n - v_n)^2 \tag{15}$$

を比較する．ここで，org_n は n 番目のオリジナルデータ，v_n は n 番目の APREP-S，または既存手法で算出した補完値である．ERR が小さいほどオリジナルデータに近い値を補完できていることになるため，精度が高いと判断する．

5.　実験

本実験では，APREP-S に一般化加法モデル（fbprophet）と LSTM を APREP-S に定義し，補完エリアが存在する補完対象データに対するデータ補完を行い，既存の手法より APREP-S が有効であることを検証する．

5.1　データセット

4.3 節の方法で $RH1$ と $RH2$ をベースに実験用データセットを生成し，4.4 節の方法で欠損値と外れ値を検出し

た結果，$RH1$ は 16 行連続の補完エリアを含む合計 54 箇所の補完箇所を，$RH2$ は 12 行，9 行，4 行の補完エリアを含む 54 箇所の補完箇所を持つデータとなった．

さらに，補完対象データ以外の特徴として，実験用データが測定された地域（ベルギーのモンス市）の天気と湿度のデータセット [12] を利用した．2008/7/1 からデータをダウンロードした 2019/6/20 までの 1 時間ごとの気温，風速，湿度，気圧などが格納されている．本評価では，入力データセットの 2016/1/11 17:00 から 2016/5/27 18:00 までの気温，天気，湿度を利用する．当該データセットは 1 時間ごとのため，例えば，1:00:00〜1:59:59 までは 1:00:00 の天気データを使用するというように，「時」が等しいデータを取得する．

5.2 特徴量

本評価では，入力データから定まる補完対象箇所の特徴 5 つと，入力データ以外の特徴 3 つの合計 8 つの特徴を定めた．具体的には，入力データから定まる補完対象箇所の特徴として，5 項目を定義する．

(1) 補完エリアの行数 ：1 以上の整数
(2) 外れ値フラグ ：外れ値の場合 1, 欠損値の場合 0
(3) 補完対象箇所の時間帯：0 時〜6 時ならば 0，7 時〜12 時ならば 1，13 時〜18 時ならば 2，19 時〜24 時ならば 3
(4) 補完箇所前後の値の傾き：$(v_{behind} - v_{front})$/補完エリア行数
(5) 補完箇所前後の傾きの傾向 ： 補完箇所前後で比較し，前後の傾きが両方とも正もしくは負ならばならば 1，異なる傾きならば -1

の 5 つ，入力データ以外の特徴として地域の天気と湿度のデータセットから (1) 気温，(2) 天気，(3) 湿度の 3 つを定義した．

5.3 APREP-S 内の補完値算出手法

本評価では，4 つの補完手法 $m_1, m_2, m_3, m_4 \in \boldsymbol{M}$ を APREP-S に定義した． m_1 は補完エリア直前と直後の値の平均値，m_2 は fbprophet， m_3 は LSTM， m_4 はスプライン補間を定義した．

fbprophet は，使用するデータは室内の湿度データのため日ごとの周期性があると考え，一般化加法モデルの周期性に着目し，条件として日ごとの周期性を指定する．学習データは，15 日分のデータを使用した．

LSTM は，TensorFlow の API である Keras を使用し，1STEP を 6 時間分にあたる 36 個のデータとし，勾配の更新は 1 日分にあたる 144 個とした．隠れユニット数は 100，訓練数は 200 とし，学習データは fbprophet と同様に 15 日分のデータを用いることとする．

5.4 モデル生成と予測

使用する学習データは，補完エリアの行数が 1，すなわちエリア補完ではない場合は単一補完箇所への補完に適している m_1, または m_4 を選択し，補完エリアの行数が 2 以上，すなわちエリア補完の場合は m_2, または m_3 を選択したとし，選択手法リストを生成する．結果として，学習データとなる $RH2$ の補完対象箇所に対し選択した手法番号リストは，$(4, 4, 1, 4, 1, ..., 3, 3, 3, 4, 4)$ (54 箇所)，存在する補完エリア 3 箇所のうち 1 箇所が m_2，2 箇所が m_3，単一補完箇所のうち 10 箇所が m_1，19 箇所が m_4 となった．学習データによりパラメータ α，β を推定し，APREP-S モデルを生成する．

モデルトレーニングフェーズで生成した APREP-S モデルを使い，手法の適正度を推測する．$RH1$ の単一補完箇所・補完エリアに適した手法の適正度を推測し，適正度が一番高い手法を選択すると，選択手法リストは $(4, 4, 4, 4, 1, ..., 1, 1, 4, 4, 1)$ (54 箇所) となった．選択手法リストに対応する手法で値を算出し，補完値を定める．

5.5 APREP-S と比較する既存手法

APREP-S 内で選択できる手法として定義した 補完エリア前後の平均値，一般化加法モデル（fbprohet），LSTM，スプライン補間の 4 種類と比較した．既存手法は補完エリア全てに同じ手法で算出した値を補完する．fbprophet と LSTM の学習データは APREP-S と同様に補完エリア直前の 15 日分のデータとする．その他の設定値も APREP-S と同様に，fbprophet の周期性は “daily”，LSTM の 1STEP を 6 時間分にあたる 36 個のデータとし，勾配の更新は 1 日分にあたる 144 個とした．隠れユニット数は 100，訓練数は 200 とする．

5.6 結果

APREP-S と既存の手法を二乗和誤差（Eq.（15））の値で比較する．結果を表 1 に示す．上段が補完対象箇所全ての二乗和誤差の値，下段が補完エリアの長さが最大（本実験では 16 行分）の箇所の二乗和誤差の値である．

補完箇所全ての値の結果から，APREP-S が最も精度が高い結果となった．しかしながら，ある区間のエリア補完に有効であると想定していた fbprophet と LSTM は，単一補完箇所に有効と想定していた平均値，スプライン補間より低い結果となった．この理由として，評価データに使用した $RH1$ の補完エリアの長さが 2 行もしくは 3 行のデータが多かったことが考えられる．補完エリアの長さが最長の箇所のみに限定して APREP-S と各手法を比較すると，LSTM が最も精度が高い結果となっている．最長箇所のみに限定すると，APREP-S は LSTM より低い値となっているが，平均値，スプライン補間より良い結果となった．しかしながら，補完エリアの長さが最長箇所のみに限定して

表 1　二乗和誤差（Eq.（15））による評価結果

	APREP-S	平均値	fbprophet	LSTM	スプライン補間
補完箇所全ての和	365.0	664.0	1525.3	1764.7	391.1
補完エリア長が最長箇所のみの和	8.8	11.9	52.3	4.0	42.4

も fbprophet は本実験の方法では精度が低い結果となった．

以上 2 点より，は補完箇所が区間になっているエリア補完を対象とした場合でも機械学習を APREP-S 内の手法に定義することにより APREP-S は有効であることが検証できた．

6. おわりに

本稿では，センサーデータの様に外れ値や欠損値を含むデータの補完・変換手法のひとつ “APREP-S” が連続した補完箇所を含むデータにも有効であることを実験した．本稿の結論は以下の 2 点である．

- Programming by Example アプローチをベースとした APREP-S に定義する手法として時系列解析のひとつ “一般化加法モデル” と，機械学習のうち時系列予測に有効な RNN のひとつ “LSTM” を定義することにより，補完対象エリアが区間であるエリア補完にも既存手法より有効であることを示す実験した．
- 単一補完箇所に対する補完に有効な手法と組み合わせることにより，より現実的に起こり得る，補完対象が単一とエリア補完が混合している場合にも APREP-S が既存手法と比較し，有効であることが検証できた．

補完エリアの長さが最長箇所の結果より，LSTM が有効であることは評価できたが，学習データを毎回作成していたため，補完値の算出に時間を要してしまった．今後は APREP-S のインタラクティブ性を損なわないよう，学習データとモデル生成の最適なタイミングを研究・検討予定である．

参考文献

[1] Qi, Zhixin and Wang, Hongzhi and Li, Jianzhong and Gao, Hong: Impacts of Dirty Data: and Experimental Evaluation, *arXiv:1803.06071 [cs, stat]* (2018).

[2] Gulwani, Sumit and Jain, Prateek: Programming by Examples: PL meets ML, *Microsoft Research*, (online), available from ⟨https://www.microsoft.com/en-us/research/publication/programming-examples-pl-meets-ml/⟩ (2017).

[3] Graham, John W.: Missing Data Analysis: Making It Work in the Real World, *Annual Review of Psychology*, Vol. 60, No. 1, pp. 549–576 (online), DOI: 10.1146/annurev.psych.58.110405.085530 (2009).

[4] 永島寛子, 加藤由花：センサーデータのための Programming by Example に基づくデータ補完手法，マルチメディア，分散，協調とモバイル (DICOMO2019) シンポジウム，pp. 2–8 (2019).

[5] Mckinley, Sky and Levine, Megan: Cubic Spline Interpolation, *Coll. Redw.*, Vol. 45 (1999).

[6] Hastie, Trevor and Tibshirani, Robert: Generalized Additive Models, *Statistical Science*, Vol. 1, No. 3, pp. 297–310 (online), available from ⟨https://www.jstor.org/stable/2245459⟩.

[7] Taylor, Sean J and Letham, Benjamin: Forecasting at scale, (online), DOI: 10.7287/peerj.preprints.3190v2 (2017).

[8] Hochreiter, Sepp and Schmidhuber, Jurgen: Long Short-Term Memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780 (online), DOI: 10.1162/neco.1997.9.8.1735.

[9] Gulwani, Sumit: Automating String Processing in Spreadsheets Using Input-output Examples, *Proceedings of the 38th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL ’11, ACM, pp. 317–330 (online), DOI: 10.1145/1926385.1926423 (2011).

[10] Facebook’s Core Data Science team: Prophet, Facebook (online), available from ⟨https://facebook.github.io/prophet/⟩ (accessed 2019-07-15).

[11] Candanedo, Luis M. and Feldheim, Veronique and Deramaix, Dominique: Data driven prediction models of energy use of appliances in a low-energy house, *Energy and Buildings*, Vol. 140, pp. 81–97 (online), DOI: 10.1016/j.enbuild.2017.01.083 (2017).

[12] World Weather Online: World Weather Online, Facebook (online), available from ⟨https://www.worldweatheronline.com/⟩ (accessed 2019-07-15).