センサーから取得した時系列データのためのデータ補完手法

永島 寛子[1]　加藤 由花[1]

概要：近年，センサーデータやウェアラブルデバイスのデータなど，分析に利用可能なデータの量と種類が増えてきた．Industry 4.0 のようなスマートファクトリーも分析例のひとつである．収集データは単位の統一や外れ値や欠損値の対処などを含んでおり，分析前の「前処理」が不可欠である．この前処理は分析プロジェクトのリソースのうち 80%を費やしているというデータもあり，分析者に多大なインパクトを与えている．そのため，私たちは分析者の前処理における負荷を削減し，かつ既存手法と同程度以上の精度をもつ前処理の自動化する手法として APREP-S（Automated Pre-Processing for Sensor Data）を提案してきた．APREP-S は，Programming by Example アプローチとベイズ推論を用いて外れ値・欠損値を自動で補完する手法である．しかしながら，従来の APREP-S は，初期モデル生成のためのトレーニングデータを分析者が生成する必要があった．そこで本稿では，APREP-S を拡張し，初期モデルのためのトレーニングデータ生成にクラスタリング手法を用いる手法を提案する．提案手法は補完精度の比較を，初期モデルのトレーニングデータ生成方法，既存のデータ補完手法，に関して行い，APREP-S が有効な補完手法であることを示した．

Data Imputation Method of Time-Series Data for Sensor Data

HIROKO NAGASHIMA[1]　YUKA KATO[1]

1. はじめに

近年，センサーデータやウェアラブルデバイスのデータなど，分析に利用可能なデータの量と種類が増えてきた．多種多様なデータを活用した分析の例として，ショッピングモールの顧客動向の分析やロボットの自律行動，クレジットカードの不正検知，ドイツの Industry 4.0 のようなスマートファクトリーなどがある．さらに，機械学習によるデータ分析は経験者の知恵やノウハウの代わりとして，工場における作業者の生産性向上や工場全体の稼働率を補うことができるため，管理する方法として着目されている [1]．しかしながら，収集データは欠損値や外れ値，データ計測機器の測定単位の違いや表記の揺れなどを含んでおり，そのまま分析モデルに入力してしまうと，正しい結果を得られない [2]．特にセンサーデータの場合，使用するセンサーによる計測間隔や単位の違いに加え，個々の機器の計測誤差や，収集にネットワークを介することによる転送タイムアウトや途中でロストする可能性がある．そのため，単位の統一や外れ値や欠損値の対処といった「前処理」が必要である．前処理を含むよく知られているデータマイニングのフレームワークに，cross industry standard process for data mining (CRISP-DM)[3] がある．私たちは，CRISP-DM をベースとし，前処理（pre-processing）プロセス内に自動化を定義した “APREP-DM (Automated Pre-Processing for Data Mining)[4]” を提案した．概要を図 1 に示す．前処理は「ビジネス知識（Business Understaning）」の後ろに位置しており，ビジネスの現場では，業務知識と機械学習の知識を兼ね備えていた分析者により作業されている．分析システムにおいて，データの前処理には全体の 80% のリソースが費やされており，分析者の作業量に大きなインパクトを与えている [5].

私たちはこれまで，APREP-DM の前処理のうち外れ値・欠損値の補完に着目し，Programming by Example（PBE）アプローチを使い，人の知識すなわち業務の知識と機械学習を融合したデータ補完手法 APREP-S (Automated Pre-Processing for Sensor Data) を提案してきた．PBE については，2.4 章で述べる．APREP-S は ベイズ推論をベー

[1] 東京女子大学 大学院理学研究科

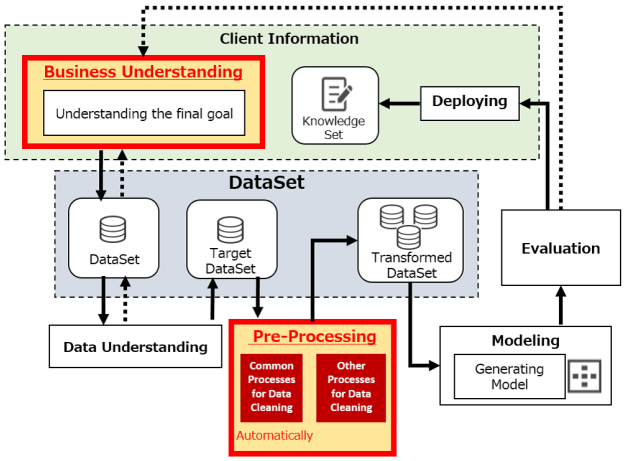

図 1　データマイニングフレームワーク (参考文献 [4] の Fig.5 をベースに作成)

スとしており，学習を繰り返すことにより，精度の高いモデルを生成する．しかしながら，従来の APREP-S は，初期モデル生成のためのトレーニングデータを分析者が生成する必要があった．そこで本稿では，APREP-S を拡張し，初期モデルのためのトレーニングデータ生成にクラスタリング手法を適用する．

本稿の貢献は下記 2 点である．

- データ分析の前処理のうち欠損値の補完に着目し，初期モデル生成時に使用するトレーニングデータの生成手法を明確化することにより，分析者が行うべき作業の一部を自動化できるため，分析者の負荷が軽減する．
- 提案手法と既存のデータ補完手法を比較し，提案手法が既存の手法と精度が同程度ないし高いことを検証する

本稿の構成は以下の通りである．まず 2 章で提案手法の関連研究について述べ，3 章で本稿提案手法全体のワークフローと特長について述べる．4 章では，提案手法と既存の手法を比較することによる評価を行い，5 章で本稿のまとめを行う．

2. 関連研究

本稿では，欠損値のデータ補完を，定められたルールにより計算された値で補完する方法，時系列解析により算出する方法，機械学習により推測する手法の 3 つに分類した．以下それぞれについて詳述する．また，提案手法 APREP-S で利用している Programming by Example についても詳述する．

2.1 ルールによるデータ補完

ルールによるデータ入力手法の代表例として，“リストワイズ法（完全ケース分析ともいう）”，“多重代入法”，“単一代入法” がある．リストワイズ法は，欠損データをすべて削除する手法である．欠損値の処理方法として有名であるが，当該手法はデータ補完を対象としているため，適さない．“多重代入法” はシミュレーションにより補完値を定める手法である．シミュレーションに利用するデータサンプルとして人の知識を融合する方法が考えられるが，いろいろなパターンのサンプルが必要であるため，生成が困難である．“単一代入法” は，決められた一つのルールによりすべての補完値を算出する方法である．例えば前後の平均値やある区間の中間値による補完や，離散値を滑らかに結ぶスプライン補間 [6] がある．ルールが決まっているため，人の知識は導入しにくい．

2.2 時系列解析により算出する方法

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 [7]．時系列解析では，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．一般化加法モデルは以下の式で表される．

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{1}$$

$g(t)$ 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 機械学習により推測する手法

RNN（Recurrent Neural Network）は，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)[8]” 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．

2.4 Programming by Example

Programming by Example（PBE）アプローチ [5] 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] やデータ構造を変換する Foofah[10] など，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 3 つメイン処理： 1) 分析者により入力された例と一致するプログラム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

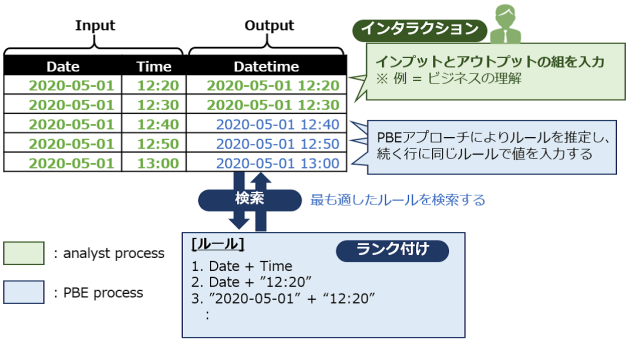

図 2 PBE により，*Date* 列と *Time* 列から *Datetime* 列を生成

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，を持つ．PBE の例を図 2 に示す．図 2 は，*Date* 列と *Time* 列をインプットデータとし，*Datetime* 列をアウトプットとするときの流れを表している．まず 1 行目に入力と出力の組として，*Date* = “2020-05-01”，*Time* = “12:20”，*Datetime* =“2020-05-01 12:20” と入力する．すると，PBE は *Datetime* 列に固定値 “2020-05-01 12:20” を挿入するというルールを学習し，以下に “2020-05-01 12:20” を挿入する．次に 2 行目に *Date* = “2020-05-01”，*Time* = “12:30”，*Datetime* =“2020-05-01 12:30” と入力すると，PBE は *Date* と *Time* を結合するというルールを学習する．1 行目，2 行目から，*Date* と *Time* を結合するルールが最適と判断し，3 行目以降のデータを同じルールで更新する．したがって，分析者は いくつかの入力と出力の組を入力するのみで，それ以下の行の値が補完されたデータを取得することができる．

3. 提案手法

前処理の自動化の方法として，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「APREP-S」を提案してきた [11][12]．本稿では，モデル生成時に利用するトレーニングデータの生成にクラスタリング手法を適用することにより，初期モデル生成において APREP-P を拡張する．本節では，まず既存の APREP-S の概要を説明し，その後提案手法について詳述する．本稿では，欠損値が連続しており，一つの手法で補完する区間を “補完対象エリア” と呼ぶこととする．

3.1 APREP-S

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 APREP-S により補完されたデータと，APREP-S に定義されたデータ補完手法の適正度である．これにより，分析者は更新時に 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，初期モデルを生成する「モデルトレーニングフェーズ」，モデルを更新する「モデル更新フェーズ」，モデルトレーニングフェーズとモデル更新フェーズで生成したモデルによりデータを補完する「モデル運用フェーズ」がある．

まず，モデルトレーニングフェーズについて説明する．APREP-S は，分析者からトレーニング用の外れ値や欠損値が含まれるデータと初期モデル用の手法選択リストを受け取った後，外れ値や欠損値箇所，すなわちデータ補完箇所を検索する．その後，各データ補完箇所に対し特徴量を定義する．定義した特徴量と分析者から受け取った初期モデル用の選択手法リストをトレーニングデータとし，APREP-S モデルを生成する．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．α と $\boldsymbol{\beta}$ はガウス分布である．各手法の事前分布を $p(m_k)$ とすると，

$$p(m_k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|m_k)p(m_k)}{\sum_{i=1}^{K} p(\boldsymbol{y}|m_i)p(m_i)} = \frac{\exp(\boldsymbol{y}(x_k))}{\sum_{i=1}^{K} \exp(\boldsymbol{y}(x_i))} \tag{2}$$

$$\boldsymbol{y}(x_q) = \alpha + \boldsymbol{\beta}\, x_q \quad (1 \le q \le Q) \tag{3}$$

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．尤度関数は，

$$C(\boldsymbol{M}|\boldsymbol{y}) = \prod_{k=1}^{K} (y_k^{u_k}) \tag{4}$$

で算出する．u_k は m_k が選択される確率である．$\sum_{i=1}^{K} p(m_i|\boldsymbol{y}) = \sum_{i=1}^{K} u_i = 1$ のため，$p(m_k|\boldsymbol{y})$ は正規化された指数関数である．したがって，各手法の適正度は $p(m_k|\boldsymbol{y})$ と同じになる．APREP-S モデル生成後は，APREP-S 内に定義する手法にモデル生成が必要な手法がある場合，モデルを生成する．

モデル運用フェーズは，まず分析者から受け取った補完したいデータ補完箇所を検索する．検索した特徴量を定義し，それぞれの補完箇所に対し，モデルトレーニングフェーズで生成したモデルにより APREP-S 内に定義された手法それぞれの適正度から最適な手法を選定する．選定した手法により，それぞれの補完箇所の補完値を算出し，補完したデータを分析者に返却する．

最後にモデル更新フェーズについて述べる．モデル更新フェーズでは，各手法の適正度と補完値を確認しながら選択された手法を変更することにより，変更データを元にし

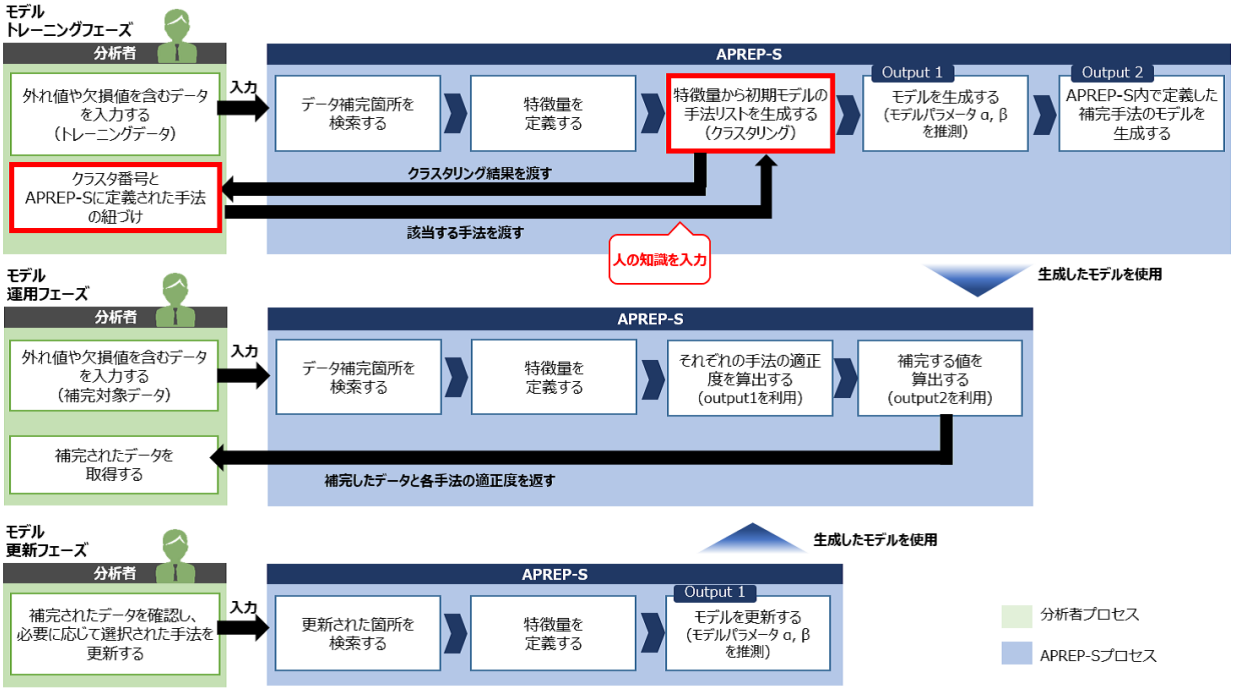

図 3　APREP-S のワークフロー図．赤枠は本稿の提案部分を示す

たモデルの更新が行われる．APREP-S は分析者から受け取った更新情報をもとに変更箇所を検索し，それぞれの補完箇所に対する特徴を定義する．定義した特徴と更新したデータを入力とし，APRPE-S モデルを更新する．更新したモデルは，モデル運用フェーズで使用する．

3.2　提案手法

APREP-S は，あらかじめ定義しておいたデータ補完手法の適正度と実際の補完値を機械学習により提示することにより，補完に適した手法を選択できる手法だが，初期モデル生成時に利用するトレーニングデータは分析者が生成する必要がある．そのため本稿では，データ補完箇所の特徴をクラスタリングすることにより，それぞれの補完対象エリアの手法選択を自動化する．

APREP-S のワークフローを図 3 に示す．モデル運用フェーズとモデル更新フェーズは既存の APREP-S と同様である．本提案は，モデルトレーニングフェーズの APREP-S が行っているプロセスに，初期モデルのための選択手法リストの生成を追加する．APREP-S は，分析者から欠損値などを補完対象データを受け取った後，既存の APREP-S 同様にデータの補完箇所を検索し，特徴量を定義する．その後，定義された特徴量を入力データとしたクラスタリングを行う．クラスタの数は APREP-S で定義している補完手法の数以上とする．そして，補完対象データのクラスタリング結果を分析者に返す．分析者は受け取った結果から分類されたクラスタの特徴と APREP-S に定義した手法の特徴を確認し，各クラスタに適した手法を紐づける．クラスタ番号と手法を紐づけるプロセスは，業務内容や所在地，気候など “ビジネスの理解” が必要になるプロセスであり，これによりモデル更新フェーズだけでなく，モデルトレーニングフェーズでも APREP-S に人の知識を入力することができる．ここで，クラスタリング手法は隠れマルコフモデル（HMM）[13] とする．HMM とは，モデル化されたシステムをマルコフ過程と仮定した統計的マルコフモデルであり，HMM は観測可能な変数 $\mathbf{X}$ の系列と，内部の隠れた状態 $\mathbf{Z}$ の系列を持っています．APREP-S は $\mathbf{X}$ は APREP-S で定義している補完手法である．

4.　評価

本評価では，APREP-S のモデルトレーニングフェーズで使用するトレーニングデータの生成方法による精度の比較を行う．

4.1　評価の準備

4.1.1　評価の準備

本評価では，ワイヤレス湿度センサー（DHT-22）のデータセット [14] を使用する．当該データセットに格納されている全ての湿度 RH データのグラフを図 4 に示す．データセットに含まれるデータのうち，特徴がある浴室エリアの湿度データ $RH5$ と屋外の湿度データ $RH6$，そしてその他のデータの中からキッチンエリアの湿度データ $RH1$ を選択した．

補完精度の検証のため，$RH1$，$RH5$，$RH6$ に欠損値を挿入する．欠損値は，データが $Null$ のデータとする．欠

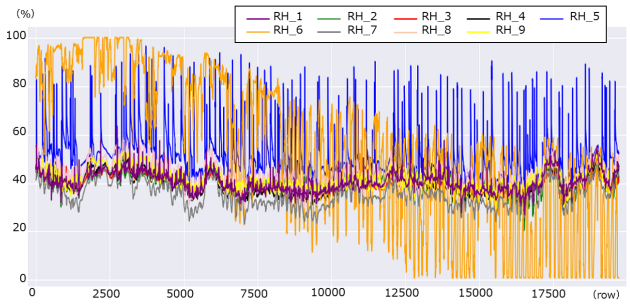

図 4　湿度（RH）データのグラフ：1row は 10 分を示す．

損値の出現確率と連増数はガウス分布

$$\mathcal{N}(e;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(e-\mu)^2}{2\sigma^2}\right\} \tag{5}$$

に従う．ここで，μ は平均値，σ^2 標準偏差とする．欠損値の出現確率は $\mathcal{N}(0,300)$，欠損値の連続数 $\mathcal{N}(0,200)$ とした．結果として，$RH1$ は 1,720 行，$RH5$ は 1,800 行，$RH6$ は 1,368 行の補完箇所を含むデータを生成した．

使用したデータセットは 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 の気温の誤差は ± 0.5°C，湿度の誤差は ± 3% である．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．

さらに，特徴に利用する天気データとして，補完対象データを観測したベルギーのモンス市から 24km 離れたスタンブルージュの公開データ [15] を使用する．気温や湿度、天気の種類などを含む 2008 年 6 月 1 日から 2019 年 6 月 20 日までの 1 時間ごとのデータを取得した．補完対象データは 10 分ごとのデータのため，本評価では，例えば 1:00:00 から 1:59:59 までのデータは 1:00:00 のデータを採用する．

4.2　APREP-S 内の補完値算出手法

本評価では，APREP-S に $m_1, m_2, m_3, m_4 \in \boldsymbol{M}$ の 4 つの補完手法を定義した．定義した手法リストを表 1 に示す．m_1 は補完対象エリアの前後の値の平均値である．v_b は補完対象エリアの直前の値，v_a は補完対象エリア直後の値としたとき，

$$f(v) = \frac{1}{2}(v_b + v_a). \tag{6}$$

で算出される値を補完する．m_2 は よく知られている GAM の手法で，Facebook 社により公開されている fbprophet[16] とする．m_3 は gated recurrent unit (GRU)[17] とする．GRU は RNN 構造を持つ手法ひとつであり，可変長のシーケンスを固定長のベクトル表現に符号化し，与えられた固定長のベクトル表現を可変長のシーケンスに復号化する手法である．m_4 はスプライン補間とする．

表 1　APREP-S に定義したデータ補完手法

m_1	補完対象エリア前後の平均値
m_2	fbprophet
m_3	GRU
m_4	spline interpolation

4.3　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，

(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 1，異なる傾きならば -1

の 5 項目，入力データ以外の特徴として地域の天気データセットから

(1) 現地の天気データセットの気温（°C）
(2) 現地の天気データセットの天気コード：快晴=113，曇り=119，霧雨=266，強い雨=308 など 48 種類
(3) 現地の天気データセットの湿度（%）

の 3 項目を定義した．

4.4　評価の方法

本評価では，それぞれのトレーニングデータにより生成したモデルで算出した補完値とオリジナルデータの二乗和誤差

$$ERR = \frac{1}{2}\sum_{i=1}^{I}(org_i - v_i)^2 \tag{7}$$

を比較する．ここで，org_i は i 番目のオリジナルデータ，v_i は i 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．

4.5　評価の手順

モデルトレーニングフェーズでは，初期モデルを生成する．トレーニングデータのクラスタリングで使用する HMM は，hmmlearn ライブラリの GaussianHMM とする．分類するクラスタ数は手法の数と同様 4 つとし，共分散パ

表 2　二乗和誤差 ERR の結果 (Eq.(7))

データ	補完対象行数	clustering	APREP-S	Mean	Fbprophet	GRU	Spline
$RH1$	1,720 行	HMM	3,519 (2.05)	3,327(1.93)	4,471(2.60)	89,760(52.19)	4,062(2.36)
		fixed	4,449 (2.59)				
		k-means	3,572 (2.08)				
$RH5$	1,800 行	HMM	66,661 (37.03)	66,821(37.12)	71,595(39.78)	259,690(144.27)	88,781(49.32)
		fixed	68,971 (38.32)				
		k-means	68,208 (37.89)				
$RH6$	1,368 行	HMM	167,168 (122.20)	186,014(135.98)	170,069(124.32)	628,169(459.19)	273,062(199.61)
		fixed	176,712 (129.18)				
		k-means	167,168 (122.20)				

* 括弧内の値は二乗和誤差の 1 行あたりの誤差．単位は%

ラメータは “full（完全な共分散行列を使用）”，繰り返し回数は 300 とした．クラスタリングの結果として出力された各クラスタに対し，どの手法で補完するかを確認しながら指定する．本評価では，補完箇所前後の値の傾きの絶対値が 0.01 以下のクラスタは平均値（m_1）をまず指定し，次に残ったクラスタのうち補完対象エリアの連続数が 50 以下のクラスタにはスプライン補間（m_4）を，最後に残ったクラスタには fbprophet（m_2）を指定した．

4.5.1　比較する選択手法の生成方法

APREP-S と 2 つの生成方法を比較する．1 つ目は補完対象エリアの連続数のみで手法を定義する方法とする．連続数が 20 以下の場合は平均値（m_1），連続数が 100 以下の場合はスプライン補間（m_4），それ以外の場合は fbprophet（m_2）とした．2 つ目は k-means 法とする．k-means 法は，入力データの重心を定めることによりクラスタリングを行う手法であり，よく知られているクラスタリング手法である．sklearn ライブラリの KMeans を使用し，クラスタ数は手法の数と同じ 4 つとした．クラスタとデータ補完手法の紐づけは HMM 同様に各クラスタにデータ補完手法を指定する．

4.5.2　比較する既存のデータ補完手法

APREP-S を 4 つの既存の補完手法と比較する．1 つ目は，補完対象箇所の前後の値の平均値を代入する方法である．2 つ目はよく知られている GAM の例として fbprophet とした．3 つ目は RNN の例として GRU，4 つ目は単一代入法の例としてスプライン補間を定義する．これからはすべて，APREP-S に定義した手法と同じである．

4.6　結果

二乗和誤差の結果を表 2 に示す．clustering 列は APREP-S の初期モデル生成時の手法を示している．“HMM” は HMM でクラスタリングした場合，“fixed” はデータ補完箇所の連続数のみで定めた場合，“k-means” は k-means 法でクラスタリングした場合の結果を表している．

クラスタリング手法が HMM のとき，fixed と k-means に比べ値が小さい，すなわち精度が高い結果なった．$RH6$ の HMM と k-means の結果が同じ値であるが，これはクラスタリングにより指定された手法の数が同じ結果だったためである．選択された手法ごとの行数を表 3 に示す．本評価では，精度が悪い GRU は選択していない．

表 3　$RH6$ のクラスタリングによる各手法選択行数

	m_1	m_2	m_3	m_4
HMM	243	1,125	0	0
k-means	243	1,125	0	0

他のデータ補完手法との比較において，$RH1$ では，mean の方が APREP-S より精度が高い結果になったが，変化量が大きい $RH5$ では同程度，$RH6$ では APREP-S の方が高い精度となっており，APREP-S の方が変化の大きいデータにも対応できている．また，$RH1$，$RH5$，$RH6$ 全てにおいて，GRU の結果が非常に悪い結果となった．

5.　おわりに

本稿では，APREP-S を拡張し，初期モデル用のトレーニングデータ生成にクラスタリング手法を適用する手法を提案した．本稿の結論は以下の 2 点である．

- APREP-S モデル生成時に，補完対象に定義した特徴を入力データとしてクラスタリングを行い，クラスタと手法を紐づけることにより，人の知識を入れ込んだモデルを生成できた
- 初期モデルのトレーニングデータとして，クラスタリング手法を用いた APREP-S の精度が，既存のデータ補完手法と比較し，同等ないし高いことを検証した

本稿では，APREP-S に複数のデータ補完手法を定め，APREP-S とオリジナルデータとの誤差が他のデータ補完手法より小さいことを検証した．しかしながら，$RH1$ 以外の結果の誤差は大きかった．APREP-S に定義している補完手法により左右されるものと推測している．次のステップとして，選定した補完手法による違いを研究・検証予定である．

謝辞　本研究の一部は，JSPS 科研費 20K11776 の助成を受けたものである．

参考文献

[1] 土谷宜弘：地域で活きる実践 IoT．自治体，農業，倉庫・工場の活用事例，リックテレコム，第 2 版 edition (2019).

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

[3] Cross Industry Standard Process for Data Mining Consortium: CRISP-DM by Smart Vision Europe, Cross Industry Standard Process for Data Mining Consortium (online), available from ⟨http://crisp-dm.eu/reference-model/⟩ (accessed 2019-12-14).

[4] Hiroko Nagashima and Yuka Kato: APREP-DM: a Framework for Automating the Pre-Processing of a Sensor Data Analysis based on CRISP-DM, *PerFoT'19 - International Workshop on Pervasive Flow of Things (PerFoT'19)*, Kyoto, Japan (2019).

[5] 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).

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

[7] 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⟩ (1986).

[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 (1997).

[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] Jin, Zhongjun and Anderson, Michael R. and Cafarella, Michael and Jagadish, H. V.: Foofah: Transforming Data By Example, *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, ACM, pp. 683–698 (online), DOI: 10.1145/3035918.3064034 (2017).

[11] Hiroko Nagashima, Yuka Kato: Data Imputation Method based on Programming by Example: APREP-S, *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 4412–4421 (2019).

[12] Hiroko Nagashima, Yuka Kato: Recommendation of Imputing Value for Sensor Data based on Programming by Example, *Journal of Information Processing*, Vol. 28 (2020).

[13] Bishop, Christopher M.: *Pattern recognition and machine learning*, Information science and statistics, Springer (2006).

[14] 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).

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

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

[17] Cho, Kyunghyun and van Merrienboer, Bart and Gulcehre, Caglar and Bahdanau, Dzmitry and Bougares, Fethi and Schwenk, Holger and Bengio, Yoshua: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, (online), available from ⟨http://arxiv.org/abs/1406.1078⟩ (2014).