DP レポート

時系列解析 SARIMA モデルを用いた 日本の COVID-19 感染予測実証実験

江谷典子（全日本空輸（Peach・Aviation）（株））

概要　時系列解析 SARIMA モデルを用いた日本の COVID-19 感染予測実証実験を行っている．SARIMA モデルのオープンソースやオープンドキュメントは不十分なため，応用することが難しい．そこで，本実験で利用している Python モジュールによる本モデルのアルゴリズム構築を明確にし，実証実験における運用方法と評価を説明し，本実証実験から得た知見から SARIMA モデルの特徴を解説する．本プログラムは GitHub にて公開している．

1. はじめに

将来予測のためのビッグデータ利活用では，SARIMA モデルなどの時系列解析も行われている．時系列解析とは時間軸に沿って変化している現象の解析を行うことである. SARIMA モデルは, 統計的手法の機械学習による解析や予測を行う．本稿では，この SARIMA モデルを用いて，日本および人口密度の高い日本の都市「東京都」「大阪府」「神奈川県」を対象とした COVID-19 感染予測の実証実験を行いながら，このモデルの知見を蓄積している．現在，公開されているオープンソースやオープンドキュメントでは説明不足のため，SARIMA モデルを応用することが難しい．そこで，本稿は応用ができるようにアルゴリズムのパラメータ決定方法と根拠を解説する．また，利用している Python モジュールを用いたプログラムをオープンソースとして公開する．このプログラムは，下記の URL から取得できる．本プログラムを御参照頂きながら，本稿を読んで利用して頂くことを推奨する．

https://github.com/NorikoEtani/COVID-19

2. SARIMA モデルの概要

2.1 技術的側面

統計モデルの推定，統計検定や統計データ探索を実行するためのクラスと関数を提供する Python モジュールは StatsModels [1] を利用している．Google Summer of Code 2009 の期間中，StatsModels は新しいパッケージとしてリリースされた．それ以来，StatsModels 開発チームは，サポートを続けている[2].

2.2 理論的側面

後述する SARIMA モデルを構築するアルゴリズムにおけるパラメータ推定を行う上で必要となる理論的側面を説明する．時系列分析は過去の変数から目的変数を求める手法であり，SARIMA モデルは，時系列解析の一手法である[3]．このモデルの構成は次の通りである．

- SARIMA モデル = ARIMA モデル + 周期的変動

- ARIMA モデル = ARMA モデル + データの差分
- ARMA モデル = AR モデル + MA モデル

AR モデル（自己回帰モデル）は，時間の変化に対し規則的に値が変化する時系列モデルである．AR モデルはある時点のデータがそれ以前のデータで回帰的に推定できるモデルである．MA モデル（移動平均モデル）は，時間の変化に対し不規則に値が変化するが，ある区間での変動が一定であるようにモデルを考える．ARMA モデルは，AR モデルと MA モデルを組み合わせたモデルである．ARMA モデルは過去の値から回帰的に推定可能な要素と過去の誤差に影響を受け，推定が難しい要素が組み合わさったモデルである．AR モデル，MA モデル，ARMA モデルはいずれも定常性を対象とした時系列モデルである．定常性の時系列データは一定の周期で同程度の変動をしているということができる．ARIMA モデル（自己回帰和分移動平均モデル）は，ARMA モデルを非定常性に対応したものである．ARIMA モデルは ARMA モデルに加えて，前後のデータ間の差分を定義する．つまり，非定常データから差分をとって定常データになるような値とする．SARIMA モデル（季節自己回帰和分移動平均モデル）は，ARIMA モデルにさらに長期的な季節（あるいは周期）変動を取り入れたモデルである．

3. 準備

3.1 実行環境

Google Colaboratory の Colab ノートブック（*.ipynb）を利用する[4]．本サービスは，環境構築が不要，GPU への無料アクセス，簡単に共有ができる特徴を備えており，Colab ノートブックを利用してコードを記述して実行できる環境である．現時点にて利用しているのは python バージョン 3.7.11 である．

3.2 オープンデータ

日本の新規感染者数および新規死亡者数は日本国内の感染者数（NHK まとめ）（2021/01/16～）[5] の”nhk_new_covid19_domestic_daily_data.csv”から抽出し，日本のワクチン 2 回接種者数は Our World in Data Coronavirus (COVID-19) Vaccinations （2021/01/22～）[6], [7] の”owid-covid-data.csv”から抽出し，東京都・大阪府・神奈川県の新規感染者数および新規死亡者数は日本国内の感染者数（NHK まとめ）（2021/01/16～）[8] の”nhk_new_covid19_ prefectures _daily_data.csv”から抽出し，各々の csv ファイルを作成し保存する．

4. SARIMA モデルの構築と予測

日本の新規感染者数および新規死亡者数（2020/01/16~2021/09/21）のデータを用いてモデルを構築し，そのモデルを用いて予測した事例を解説する．

4.1 モデルの識別および選択

SARIMA モデルのアルゴリズムで設定するパラメータを決定している．扱うデータが同じである場合，予測毎に実行する必要はなく，最初のモデル構築の時のみ実行する．

4.1.1 加法モデル・乗法モデルの選択

日本の新規感染者数（2020/01/16~2021/09/21）の観測値（Observed）を基本成分「傾向動向（トレンド）：Trend」「季節変動：Seasonal」「残差（不規則変動 = 誤差変動 + 特異的変動）：Residual」に分解してみる（図 1）[9].

傾向動向とは，時間とともに単調に増加/減少する変動である．季節変動とは，季節によって左右される 1 年を周期として規則的に繰り返される変動である．ここでは，同じサイクルで繰り返される日変動のような固定的な変動を同様に扱う．残差とは，前述以外の説明がつかない不規則かつ短期間に起こる小変動である．加法モデルは観測値が大きいときも小さいときも，季節変動の幅は一定である．一方，乗法モデルとは観測値が大きいときは季節変動幅も大きくなり，観測値が小さいときは季節変動幅も小さくなる．日本の新規感染者数における観測値（2021/07/01~2021/07/31）に着目した季節動向を図 2 に示す．観測値が大きくても小さくても季節変動幅は一定であるので，今回は加法モデルを選択する．日本の新規死亡者数も同様であった．

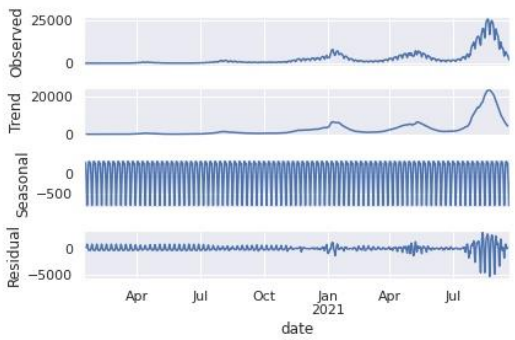

図 1　日本の新規感染者数における基本成分の分解

図 2　日本の新規感染者数における観測値（上）と季節変動（下）

4.1.2 周期性の確認：自己相関の検定

図 2 から 7 日間周期で繰り返されていることを確認できた．日本の新規死亡者数も同様であった．また，周期性の確認には自己相関の検定を用いることもできるので紹介しておく．時間差を考慮して，過去の値とどれほど似ているのか（相関があるのか）を表す指標を自己相関関係と呼ぶ．この時，時間差の度合いをラグと呼ぶ．ラグ 1 の自己相関関係がある場合，今日の値，先週の値，先々週の値が関係するという推移関係が成り立つ．推移関係を排除した，直接的な今週の値

と先々週の値の関係性を調べるには，先週の影響を除去した自己相関係数を調べる方法が必要となる．

図 3 は，日本の新規感染者数（2020/01/16~2021/09/21）の観測値から作成されたコレログラムである．左上は観測値の自己相関，左下はその偏自己相関，右上は残差の自己相関，右下はその偏自己相関を示す．青い帯は「相関がない」を帰無仮説とした 95% 信頼区間を表している．つまり，帯の外にある値は 5% の有意水準で相関があることになる．観測値の自己相関および残差の自己相関では，ラグ 1,7,14,21,28 の周期性の相関が表れている．一方，両者とも偏自己相関ではラグ 1 以外に相関がある箇所は見当たらない．つまり，残差では周期変動を取り除くことができていることになる．

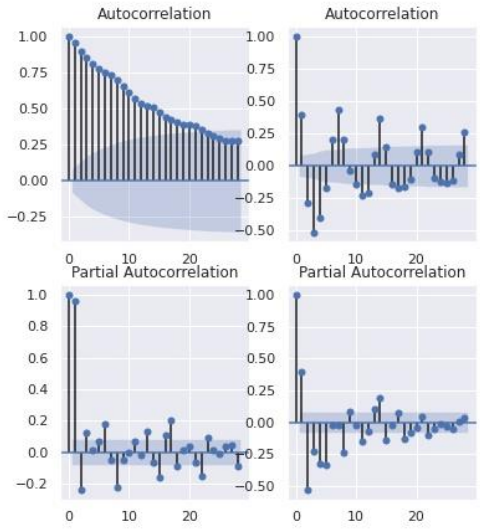

図 3　日本の新規感染者数における観測値および残差のコレログラム

4.1.3 ADF 検定：データの定常性・単位根の確認

図 4 から本データ（2020/01/16~2021/09/21）は，一定の周期で同程度の変動をしている定常性はなく，7 日間の平均値が時間的に常に一定ではない非定常性を示している．

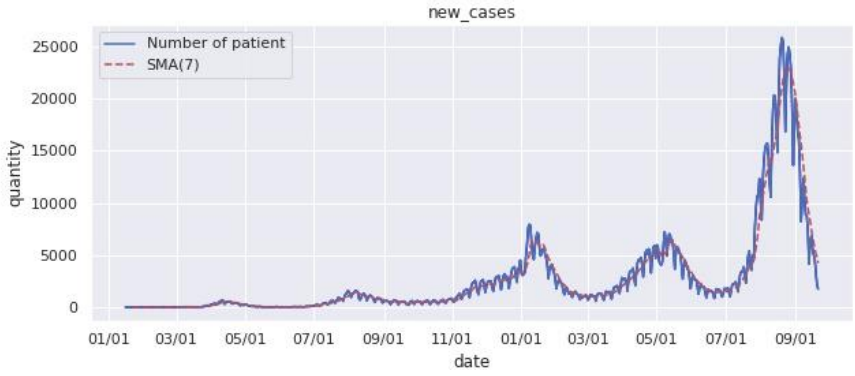

図 4　日本の新規感染者数推移（SMA(7):移動平均 7 日間）

非定常性の時系列データを対象とするために，データの観測値は非定常だが差分をとると定常になるようなデータであるという単位根の確認を行う．日本の新規感染者数および新規死亡者数から差分 1 回を行った場合の ADF 検定結果を表 1 に示す．新規感染者数および新規死亡者数において，「nc:トレンド項なし，定数項なし」は，帰無仮説『単位根（非定常性）である』を P 値 5%以下であるので放棄し，データに定常性があると判断できる．

表 1　日本の新規感染者数および新規死亡者数における ADF 検定

ADF 検定	新規感染者数 p 値	新規死亡者数 p 値
ct:トレンド項あり，定数項あり	0.009	0.052
c: トレンド項なし，定数項あり	0.001	0.011
nc: トレンド項なし，定数項なし	**0.00005**	**0.001**

4.1.4 評価法 aic および bic の選択

ラグの次数を赤池情報量基準（aic）あるいはベイズ情報量基準（bic）から選択を行う．aic は標本数が大きくなると，必要以上に複雑なモデルを過大評価する可能性が出てくるが，bic ではそのようなことは発生しないとある[3]．そこで，後述 4.2.1 の「StatsModels による ARMA モデルの最適パラメータ自動選択」を利用して，aic と bic の最適パラメータを出力した．両方ともに同じパラメータ値となった．そこで，ひとまず赤池情報量基準（aic）を選択した．

4.2 SARIMA モデルを構築するアルゴリズム

4.2.1 ARMA モデルのパラメータ推定

ARIMA モデルは，3 つのパラメータを持つ．

①AR モデルにおける回帰数

②差分を取る回数．1 回の差分で定常性を示すことができたので「1」回とする．

③MA モデルにおける平均を計算する際に考慮するデータ数．

次の手順により①③のパラメータを推定する．

- 差分計算

  差分は 1 回取ることで定常性のあるデータとなる．

  diff2 = ts2.diff()

- StatsModels による ARMA モデルの最適パラメータ自動選択

  下記プログラムを実行した結果，最適パラメータの出力を図 5 に示す．'aic_min_order': (2, 2) は上記①③のパラメータである．

  params2 = sm.tsa.arma_order_select_ic(diff2, ic = 'aic', trend = 'nc')

```
{'aic':          0         1         2
0       NaN  6047.136382  6047.073811
1  6047.190135  6045.508384  6045.534017
2  6047.931584  6044.894512  5971.789285
3  6046.007226  6042.253750  5973.678983
4  6038.292372  6037.458961  5975.800910, 'aic_min_order': (2, 2)}
```

図 5　ARMA モデルの最適パラメータ出力

4.2.2 ARIMA モデルのプログラム作成

ARIMA モデルを記述したプログラムを作成する．

arima_model2 = ARIMA(ts2, order = (2,1,2)).fit(dist = False)

4.2.3 SARIMA モデルのパラメータ推定

SARIMA モデルは，4 つのパラメータを持つ．

①AR 項の数

②定常系列に対して実行する必要がある差異

③MA 項の数

④周期性，データの周期性の長さ．「7」とする．

次の手順により①②③のパラメータを推定する．

- seasonal_order の最適パラメータ自動選択[10]

  前述の ARIMA モデル order = (2,1,2)の場合，①②③のパラメータを推定するシミュレーション結果を示す（図 6）．

```
Examples of parameter combinations for Seasonal ARIMA...
SARIMAX: (0, 1, 1) x (0, 1, 1, 7)
SARIMAX: (0, 1, 1) x (0, 1, 2, 7)
SARIMAX: (0, 1, 2) x (1, 1, 0, 7)
SARIMAX: (0, 1, 2) x (1, 1, 1, 7)
ARIMA(2, 1, 2)x(0, 1, 0, 7)7 - AIC:6050.587127908294
ARIMA(2, 1, 2)x(0, 1, 1, 7)7 - AIC:5763.327122126471
ARIMA(2, 1, 2)x(0, 1, 2, 7)7 - AIC:5698.487677416287
ARIMA(2, 1, 2)x(1, 1, 0, 7)7 - AIC:5835.758737252309
ARIMA(2, 1, 2)x(1, 1, 1, 7)7 - AIC:5775.685739080645
ARIMA(2, 1, 2)x(1, 1, 2, 7)7 - AIC:5698.082395493729
ARIMA(2, 1, 2)x(2, 1, 0, 7)7 - AIC:5734.230751208201
ARIMA(2, 1, 2)x(2, 1, 1, 7)7 - AIC:5709.065371169256
ARIMA(2, 1, 2)x(2, 1, 2, 7)7 - AIC:5699.834591604163
*BEST ARIMA(2, 1, 2)x(1, 1, 2, 7)7 - AIC:5698.082395493729
```

図 6　ARIMA モデルのパラメータシミュレーション結果

4.2.4 SARIMA モデルのプログラム作成

パラメータ推定結果から SARIMA モデルを記述したプログラムを作成する．

```
sarima_model2 = sm.tsa.SARIMAX(ts2, order = (2,1,2), seasonal_order = (1,1,2,7)).fit()
```

4.3 SARIMA モデルを用いた予測

4.3.1 28 日毎の予測期間の設定

作成したモデルに基づいて予測を行う．予測対象期間は，予測開始日から起算して将来 28 日間とした SARIMA モデルによる予測プログラムを作成する．

```
predict = sarima_model2.predict('2021-09-21', '2021-10-18')
```

周期性は「7（日間）」であるので，7 日間単位の予測も可能である．しかし，値の変化が激しい場合は，毎日モデルを構築することになり，将来の傾向を読み取ることが困難であった．モデルのメンテナンスが容易で，近未来の傾向を読み取るために 28 日間の予測とした．

4.3.2 予測期間 28 日間の最大値と最小値による予測

感染症の発症は，感染を受けやすい感受性のある人とない人の個体差がある[11]．そこで，毎日正しく新規感染者数や新規死亡者数を予測することは難しいと考え，この 28 日間における最大値と最小値を示すことで，感染状況の傾向を掴めるようにした．予測結果から抽出した予測最大値と最小値は，小数点以下を繰り上げて整数にしていることも個体差への配慮である．

5. 実証実験

5.1 運用方法

5.1.1 SARIMA モデルの管理

前述の SARIMA モデル構築におけるモデルの識別および選択について，日本および東京都・大阪府・神奈川県の新規感染者数と新規死亡者数は，同じ識別と選択を行うことができた．よって，計 8 モデルを管理する．

5.1.2 SARIMA モデルの構築手順

初回のモデル構築の場合は，前述の「4.1 モデルの識別および選択」を実行して基本的な設定項目を決定する．次回以降は，「4.2 SARIMA モデルを構築するアルゴリズム」から開始する．観察値が予測最大値と最小値を超えていない場合

は，4.2 を実行する必要はなく，その時点までの予測最大値と予測最小値，およびその時点までの観測値を用いて作図を行う．

5.1.3 SARIMA モデルの更新と予測

本モデルは過去のデータで回帰的に推定を行うという特性から，観測値が予測の最大値あるいは最小値を超えた場合，その時点までの観測値から SARIMA モデルを再構築するため，4.2 を実行する．また，その時点から 28 日間の新規感染者数予測および新規死亡者数予測を行う．

5.2 評価

2021/07/30～2021/09/21 までの日本の新規感染者数および新規死亡者数の予測結果を代表として示す．

5.2.1 結果

図 7 は日本の新規死亡者の観測値と予測，図 8 は日本の新規感染者数の観測値と予測を表示している．各図の青色は観測値，赤色は予測最大値と予測最小値を表示している．

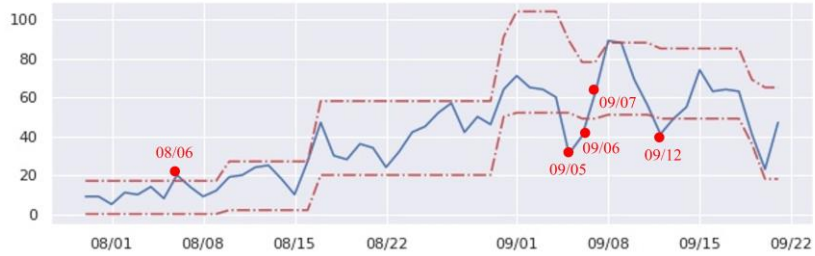

図 7　日本の新規死亡者数予測

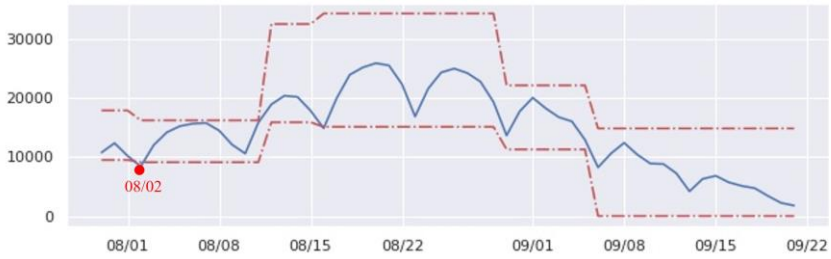

図 8　日本の新規感染者数予測

5.2.2 考察

予測が当たる場合と当たらない場合を考察し，SARIMA モデルの特徴について解説を行う．

- 予測が当たる場合

  図 9 は神奈川県の新規死亡者数の観測値と予測を示す．観測値の変化に追随が出来ている 2021/08/03 に着目する．この時点の前々日および前日の 3 日間の連続データ「2」「5」「0」と同じ値は，過去のデータでは 2021/05/13~2021/05/15 に存在する．

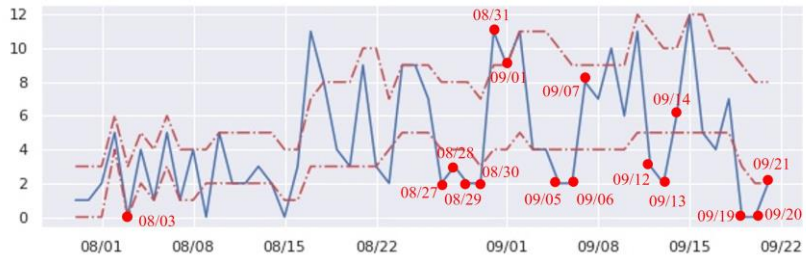

図 9　神奈川県の新規死亡者数予測

- 予測が当たらない場合

  図 7 で予測値が外れているのは，2021/08/06（値「20」）・2021/09/05（値「31」）・2021/09/06（値「40」）・2021/09/12（値「41」）である．図 8 で予測値が外れているのは，2021/08/02（値「8391」）である．各時点の前々日および前日の 3 日間の連続データと同じ値は，過去のデータには存在しない．

- 3 日後の予測復旧パターン

  図 7 の 2021/09/05~2021/09/07（値「31」「40」「62」）に着目すると予測が外れてから 3 日目には復旧している．また，図 9 の 2021/09/05~2021/09/07（値「2」「2」「8」）・2021/09/12~2021/09/14（値「3」「2」「6」）・2021/09/19~2021/09/21（値「0」「0」「2」）も同様である．2021/08/27~2021/09/01（値「2」「3」「2」「2」「11」「9」）のように 6 日後に復旧する例外もある．東京都および大阪府の新規死亡者数予測も同様に 3~4 日後には復旧している．予測が外れ続けていても，毎日，回帰的に推定を行うための過去のデータが蓄積され，復旧すると思われる．

以上より，時系列データは過去のデータからパターンを解析しているため，予測対象のデータが過去のデータのパターンから外れている場合は，予測や追随が難しくなる．また，SARIMA モデルによるモデリング自体に問題はないと考える．

6. おわりに

執筆者の researchmap にある研究ブログを活用して，日本の COVID-19 感染予測，東京都・大阪府・神奈川県の COVID-19 感染予測，の予測結果を適時アップロードし，更新した時は SNS twitter で呟いている．現在，２つのサービスを提供している．

- 時系列データの分析と予測その１：日本の COVID-19 感染予測[12]
- 時系列データの分析と予測その２：東京都・大阪府・神奈川県の COVID-19 感染予測[13]

COVID-19 対応を担っている医療機関や公的機関などの組織だけではなく，我々が個人として安全な生活を考える上で，将来の適切な行動や対応を検討や準備するための手がかりとなるように取り組んでいる．今回，利用しているプログラムを公開しているので，一人でも多くの方々がご自身でも予測を立てられて，安全な生活を行うためのツールとして利用して頂ければ幸いである．

参考文献

1) StatsModels, https://www.statsmodels.org/stable/index.html.
2) About StatsModels, https://www.statsmodels.org/stable/about.html.
3) 島田直希: 時系列解析-自己回帰型モデル・状態空間モデル・異常検知-, 共立出版株式会社 (2020).
4) Google Colaboratory, https://colab.research.google.com/notebooks/intro.ipynb.
5) 日本国内の感染者数（NHK まとめ）, https://www3.nhk.or.jp/news/special/coronavirus/data-all/.
6) Our World in Data Coronavirus (COVID-19) Vaccinations, https://ourworldindata.org/covid-vaccinations.
7) Mathieu, E., Ritchie, H., Ortiz-Ospina, E. et al., A global database of COVID-19 vaccinations, Nat Hum Behav (2021), https://www.nature.com/articles/s41562-021-01122-8.
8) 都道府県ごとの感染者数（NHK まとめ）, https://www3.nhk.or.jp/news/special/coronavirus/data/.
9) Python: statsmodels で時系列データを基本成分に分解する, https://blog.amedama.jp/entry/sm-decompose-series.
10) Model Creation: SARIMA Using Python – Forecast Seasonal Data, https://www.wisdomgeek.com/development/machine-learning/sarima-forecast-seasonal-data-using-python/.
11) 感染症の基礎知識, https://pro.saraya.com/fukushi/kansen/kiso/.
12) 時系列データの分析と予測その１：日本の COVID-19 感染予測, https://researchmap.jp/blogs/blog_entries/view/387354/49dc5d4309d202225ecf14670b29299e?frame_id=844568.
13) 時系列データの分析と予測その２：東京都・大阪府・神奈川県の COVID-19 感染予測, https://researchmap.jp/blogs/blog_entries/view/387354/206fc13e488a815b29abb72c3fb04f7e?frame_id=844568.

江谷 典子（正会員）dr.noriko.etani@ieee.org
全日本空輸（Peach・Aviation）株式会社．2001 年 3 月奈良先端科学技術大学院大学情報科学研究科博士後期課程修了 博士（工学）．2014 年 4 月～2016 年 3 月 JST CREST/京都大学 特定研究員（独立行政法人科学技術振興機構 CREST「科学的発見・社会的課題解決に向けた各分野のビッグデータ利活用推進のための次世代アプリケーション技術の創出・高度化」研究領域に従事）．

投稿受付：2021 年 9 月 23 日
採録決定：2021 年 9 月 29 日
編集担当：細野　繁（東京工科大学）