北海道科学大学研究紀要
第50号（2022年）
研究報告

Bulletin of Hokkaido University
of Science, No.50(2022)
Research Reports

機械学習を用いたＣＯＶＩＤ－１９新規感染者数の予測

Predictions of the Number of COVID-19 Newly Infected People using Machine Learning

小松 隆行*

Takayuki Komatsu

概要

本稿では，インターネット上に公開されている COVID-19（新型コロナウィルス）に関するオープンデータを，時系列分析の手法や機械学習の手法などを使って，データサイエンス的な視点から分析することを試みる．今回は，都道府県毎に報告されている新規感染者数の日次データを時系列データと捉え，移動平均を算出し，これらを入力データとした時系列データを学習するニューラルネットワークの RNN（リカレントネットワーク）の発展形である LSTM（Long Short-Term Memory）を用いたニューラルネットワークで，深層学習のアルゴリズムを用いて学習させ，新規感染者数を予測した結果について報告する．

1. はじめに

COVID-19（新型コロナウィルス）のパンデミックは，感染者の増加減少の波を繰り返しながら未だに収束しない状況にある．全世界の多くの機関ではインターネット上に関連するデータを公開しており，日本でも厚生労働省などのサイトなどにおいて公開されていて，それらをダウンロードし解析できるようになっている[(1)(2)(3)]．また，短期的な感染者数の予測が様々な手法を用いて報告されている．例えば Google 社が新規感染者数予測をインターネット上に公開している[(4)]．それらにおいて用いられている手法は，疫学的な理論，社会統計学的な理論など，多岐に渡り，様々な人工知能（AI）の手法を用いた予測も行われている．

本報告の目的は，COVID-19 の新規感染者数のデータをある非定常の情報源から得られる実測値データと考え，これらの時系列データのみからデータサイエンスの機械学習などの手法を用いることによってモデルを学習させ予測をすることである．用いるデータは，北海道を含む全国 47 都道府県のオープンデータである．学習と予測を行うモデルは，RNN（リカレントニューラルネットワーク）の発展形であり学習性能が高いとされる LSTM（Long Short-Term Memory）[(5)]を用いたディープニューラルネットワーク（DNN）を，深層学習アルゴリズム[(6)]を用いて学習させたものである．入力データの基本となるのは，新規感染者数の日次データであるが，それらをファイナンス分野[(7)]の時系列データ解析で多用されている移動平均にした値も用いることにする．

2. 学習と予測のための深層学習モデルとデータ

新規感染者数のオープンデータ[(2)]を時系列データとして予測モデルを構築する．予測モデルには，時系列予測の深層学習モデル（以下では単にモデルと呼ぶ）の手法である性能の高い LSTM(Long Short-Term Memory)[(6)]を用いる．入力と出力に関しては，時系列データの 1 入力 1 出力のモデル，2 入力 1 出力のモデルとする．入力は，日次データ，それらの 7 日移動平均（以下では単に 7 日平均と呼ぶこともある），その組み合わせとし，出力（教師データ）は，日次データ，またはそれらの 7 日平均の値とする．また，感染者が初めて報告された 2020 年 1 月 16 日から 2022 年 2 月 15 日までのデータを使用した．次章以降では，北海道新規感染者数の日次データを用いて予測モデルを構築するが，用いる北海道のデータとそれらから算出される 7 日平均の値をプロットしたグラフを図 1 に示す．第６波は，第１波～第５波までのピークよりも数倍の感染者数となっているため，上下に分けて描画している．後半の章では，他の都府県毎のデータも同様にして使用する．

＊北海道科学大学未来デザイン学部メディアデザイン学科

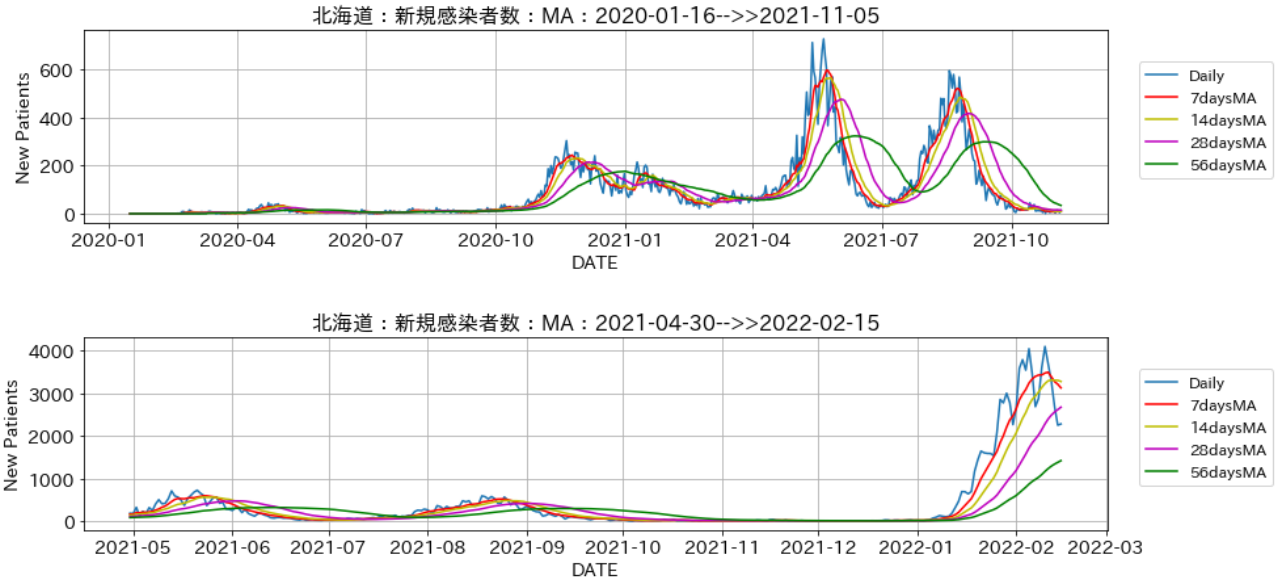


図 1 北海道の新規感染者数の推移と移動平均（上：2020/1/16～2021/11/5，下：2021/4/30～2022/2/15）

3. 深層学習モデルについて

今回は 2 種類の深層学習モデルを構築する．まず一つ目のモデルは，1 入力 1 出力のモデルである．この概略図を図 2 に示す．ここで，入力データ系列は連続する 7 日分の日次データ，または 7 日分の 7 日平均の値とし，続く 8 日目（7 日目の翌日）の日次データ，または 7 日平均の値を教師データとして，それらを予測できるように学習させる．LSTM のユニット数は 300 とし，学習における損失関数は平均二乗誤差（MSE），学習率を Adam[8] で変えながらエポック数 5000 で深層学習のアルゴリズムにより学習させる．なお，詳細は後の章で述べることにする．

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 300)	362400
dense_1 (Dense)	(None, 1)	301
activation_1 (Activation)	(None, 1)	0

Total params: 362,701
Trainable params: 362,701
Non-trainable params: 0

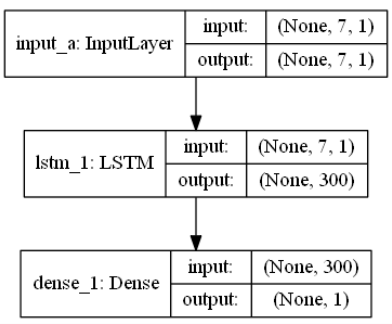


図 2　LSTM を用いた 1 入力 1 出力のモデル

二つ目は 2 入力 1 出力のモデルであり，概略図を図 3 に示す．予備実験のシミュレーションから，入力データそれぞれに LSTM ユニットを用意したほうが，予測精度が高かったことを確認できたので，このようなモデルとした．学習パラメータ数が増加すると，予測精度向上に有効となる場合と，逆に冗長になる場合があることに留意する．ここでは，入力データ系列は連続する 7 日分の日次データと，対応する同じ日付の 7 日分の 7 日平均の値とし，続く 8 日目の日次データ，または 7 日平均の値を教師データとして学習させる．これも，詳細は後述する．

Layer (type)	Output Shape	Param #	Connected to
input_a (InputLayer)	(None, 30, 1)	0	
input_b (InputLayer)	(None, 30, 1)	0	
lstm_5 (LSTM)	(None, 300)	362400	input_a[0][0]
lstm_6 (LSTM)	(None, 300)	362400	input_b[0][0]
concatenate_1 (Concatenate)	(None, 600)	0	lstm_5[0][0] lstm_6[0][0]
dense_5 (Dense)	(None, 1)	601	concatenate_1[0][0]

Total params: 725,401
Trainable params: 725,401
Non-trainable params: 0

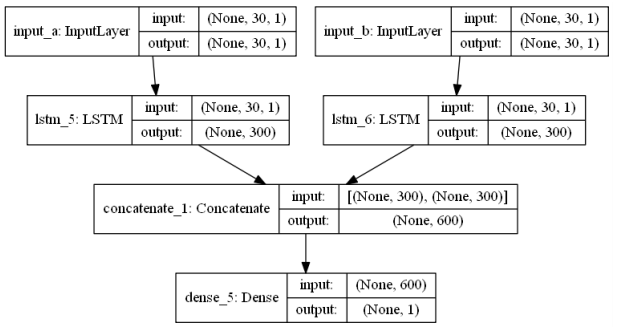


図 3　LSTM を用いた 2 入力 1 出力のモデル

4. 北海道の新規感染者数に関する学習と予測

ここでは，3 章で述べたモデルを順に構築し，2 章の図 1 で示した北海道のデータを用いて，学習と予測を行った結果を示す．モデルは，入力データと教師データの組み合わせで 4 種類となるが，モデル毎に学習曲線の例と第 6 波の期間の予測結果のグラフを示す．学習においては，用いる期間の日次データ全体の開始日付データから順に 95%の分量を訓練データとし，残りの 5%分を評価データとしている．すなわち，最新の 5%分のデータ（1000 日分のデータであれば，最新の 50 日分のデータ）を未知のデータとして，これらに対して誤差が最小のモデルを決定する．これは，未知データに対する汎化性能が高いモデルを決定することを意味する．なお学習では，用いるデータを標準化した値を使って学習し予測した後に，標準化の平均値と標準偏差の値を用いて逆算して実際の値のスケールに変換している．

4.1 日次データの予測：1 入力 1 出力モデル

まず，入力データも教師データも日次データとする 1 入力 1 出力のモデルを構築する．学習曲線の例を図 4 に示す．横軸はエポック数，縦軸は損失関数の値である．学習曲線では，青線は訓練誤差，橙色の線は評価誤差を表している．学習初期に評価損失が下げ止まり，その後上昇し続けており，早期に過学習が発生していることが分かる．図 5 は，評価損失が最小のモデル（以下では，最良モデルと呼ぶ）を使って予測を行った結果のグラフである．横軸は日次データの開始日を 0 とした日数，縦軸は感染者数である．黄線が日次データの実現値，赤線が予測値，緑線が実現値と予測値の差（予測誤差）である．予測値は実現値の増減にやや遅れて追従している．

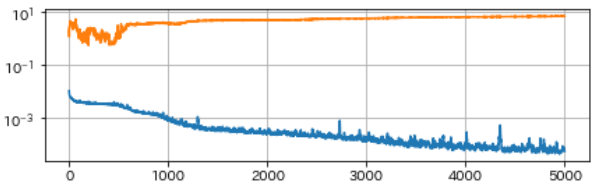


図 4　1 入力 1 出力モデルでの学習曲線の例

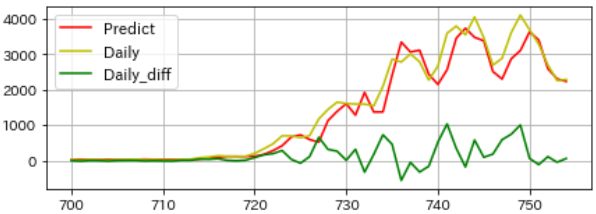


図 5　1 入力 1 出力モデルでの日次データ予測の例

データ全体の最新日付（最後の日付）から遡って 5%分が評価データであることから，この例では第 6 波の期間がほぼ評価データに相当する．この例では，701 日目以降では，予測誤差の平均は 202.76 人，最大は 1029.44 人，最小は 0.34 人と算出される．なお，予測誤差の最大と最小は誤差の絶対値である．

4.2 日次データの予測：2 入力 1 出力モデル

次に入力を日次データと 7 日平均の値とした 2 入力 1 出力のモデルを構築する．教師データは日次データとした．学習曲線の例を図 6 に示す．線種は図 4 と同じ意味のものである．やはり，早期に過学習が起きているが，評価誤差はエポック数が増えても低い値を保っている．図 7 は，最良モデルを使った予測結果である．線種は図 5 と同様である．予測値は実現値の短期の細かい増減の変化を予測できてはいないが，増減の大きな傾向（トレンド）は予測できている．この例において 701 日目以降では，予測誤差の平均は 250.12 人，最大は 1074.06 人，最小は 0.59 人と算出される．

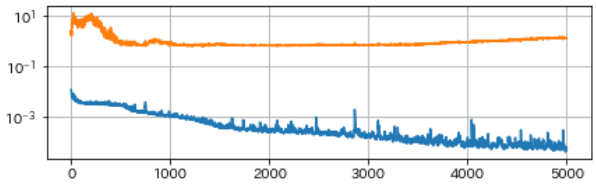


図 6　2 入力 1 出力モデルでの学習曲線の例

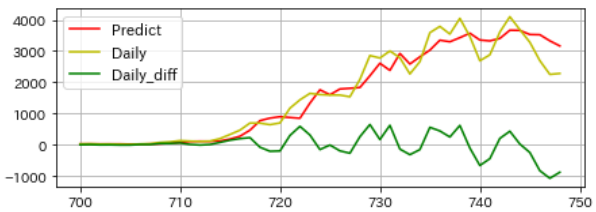


図 7　2 入力 1 出力モデルでの日次データ予測の例

4.3 7 日平均の予測：1 入力 1 出力モデル

次に，入力データも教師データも 7 日平均の値とした 1 入力 1 出力モデルを構築する．学習曲線の例を図 8 に示す．やはり，過学習が起きているが，その最小の評価誤差は先の 2 つのモデルよりもかなり小さい値であり，より良い汎化が実現していることが分かる．図 9 は，黄線が日次データの実現値，橙線が 7 日平均の予測値，赤紫線が 7 日平均の予測値，緑線が 7 日平均の実現値と予測値の差を表している．予測値は 7 日平均の実現値とほぼ一致している．この例では 701 日以降の予測誤差の平均は 29.46 人，最大は 124.00 人，最小は 0.08 人である．

これらは 7 日平均の予測値であるため，前節までの日次データの予測値の誤差と直接比較はできない．そこで，7 日平均の予測値と，日次データの実現値と 7 日平均の定義式を用い，7 日平均の変化率から逆算した値を日次データの予測値として，赤紫色線としてプロットしたものが図 10 である．図 10 から，日次データを直接予測した 4.1 節と 4.2 節の結果に比べ，誤差はあるものの日次データに対しよく追従しており，遅れも減っていることが分かる．ここでの 701 日目以降の予測誤差の平均は 206.19 人，最大は 867.98 人，最小は 0.53 人である．

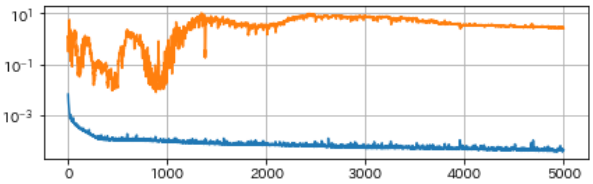


図 8　1 入力 1 出力モデルでの学習曲線の例

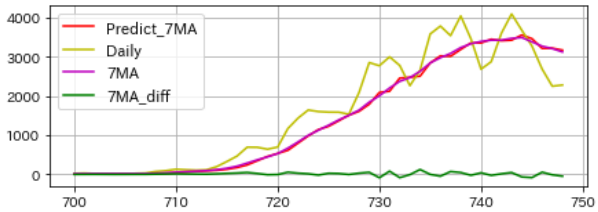


図 9　1 入力 1 出力モデルでの 7 日平均の予測の例

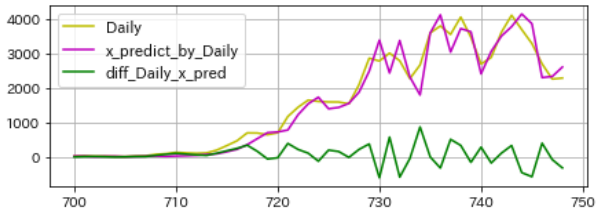


図 10　7 日平均の予測値による日次予測の例

4.4 7 日平均の予測：2 入力 1 出力モデル

最後に，入力は日次データと 7 日平均の値とし，教師データは 7 日平均の値とした 2 入力 1 出力モデルを構築する．図 11 に示す学習曲線では，学習開始直後と後半に最小の評価損失となっているが，前節のモデルの最小損失ほどは下がってはいない．また，図 12 の予測のグラフからも 7 日平均の予測がうまくいっていないことが分かる．この例では 701 日以降の予測誤差の平均は 101.72 人，最大は 246.52 人，最小は 0.27 人である．前節の図 10 と同様に，7 日平均の予測値を用いて逆算した日次予測の例を図 13 に示すが，誤差が大きく予測値は実現値とは大きくかけ離れていることが分かる．この例では 701 日目以降では，予測誤差の平均は 712.06 人，最大:1725.61 人，最小は 1.87 人と算出される．

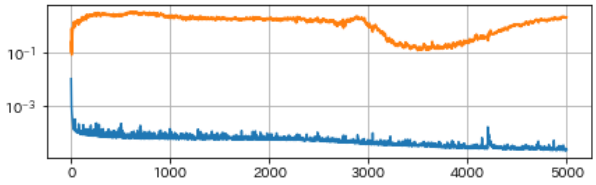


図 11　2 入力 1 出力モデルでの学習曲線の例

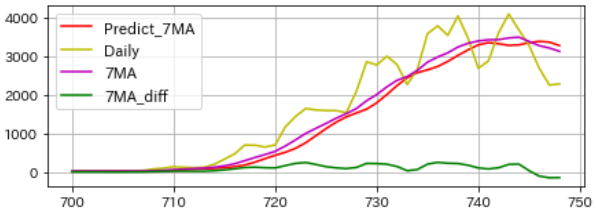


図 12　2 入力 1 出力モデルでの 7 日平均の予測の例

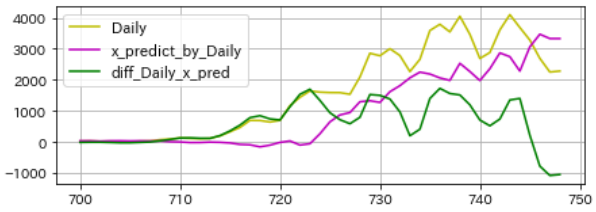


図 13　7 日平均の予測値による日次予測の例

4.5 予測が良かったモデルの結果の比較

ここで，予測結果が良かった 4.1 節の最良モデルの全期間での予測を図 14 に，4.3 節の最良モデルの全期間での 7 日平均の予測を図 15 に，この 7 日平均の予測結果から算出した日次予測を図 16 にそれぞれ示す．全期間を 4 つの期間に分割しているため，縦軸のスケールが異なっている．両方のモデルとも，日次予測は 200 日目までの予測は悪いが，200 日目以降は概ね予測値が実現値に近い値を示している．4.1 節のモデルでは，予測値のグラフが実現値より少し遅れることが多く，全期間での予測誤差の平均は 34.23 人，最大は 1029.44 人，最小は 0.04 人である（図 14）．一方，4.3 節のモデルでは，全期間の 7 日平均の予測が非常に良く，この 7 日平均の予測値から算出した日次予測（図 16）では，日次データの増減に対応した予測も多く見られ，全期間での予測誤差の平均は 33.82 人，最大は 867.98 人，最小は 0.04 人である．日次データでの学習と予測よりも，7 日平均での学習と予測から得られる日次予測のほうが，誤差の平均の意味でも誤差の最大が小さいという意味でも良い予測と言える．特に，400 日目以降のグラフを比較すると明らかである．

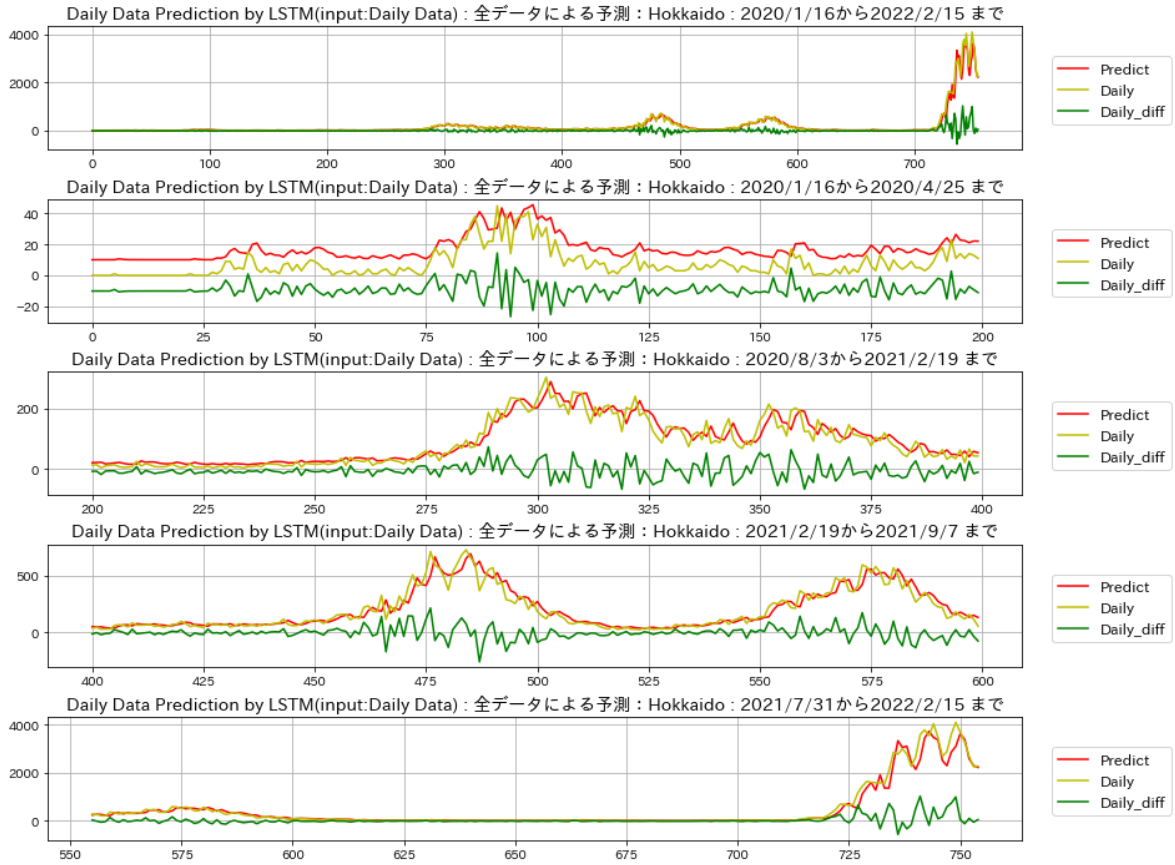


図 14　LSTM を用いた北海道の予測の例（入力と教師データは日次データ，2020/1/16～2022/2/15）

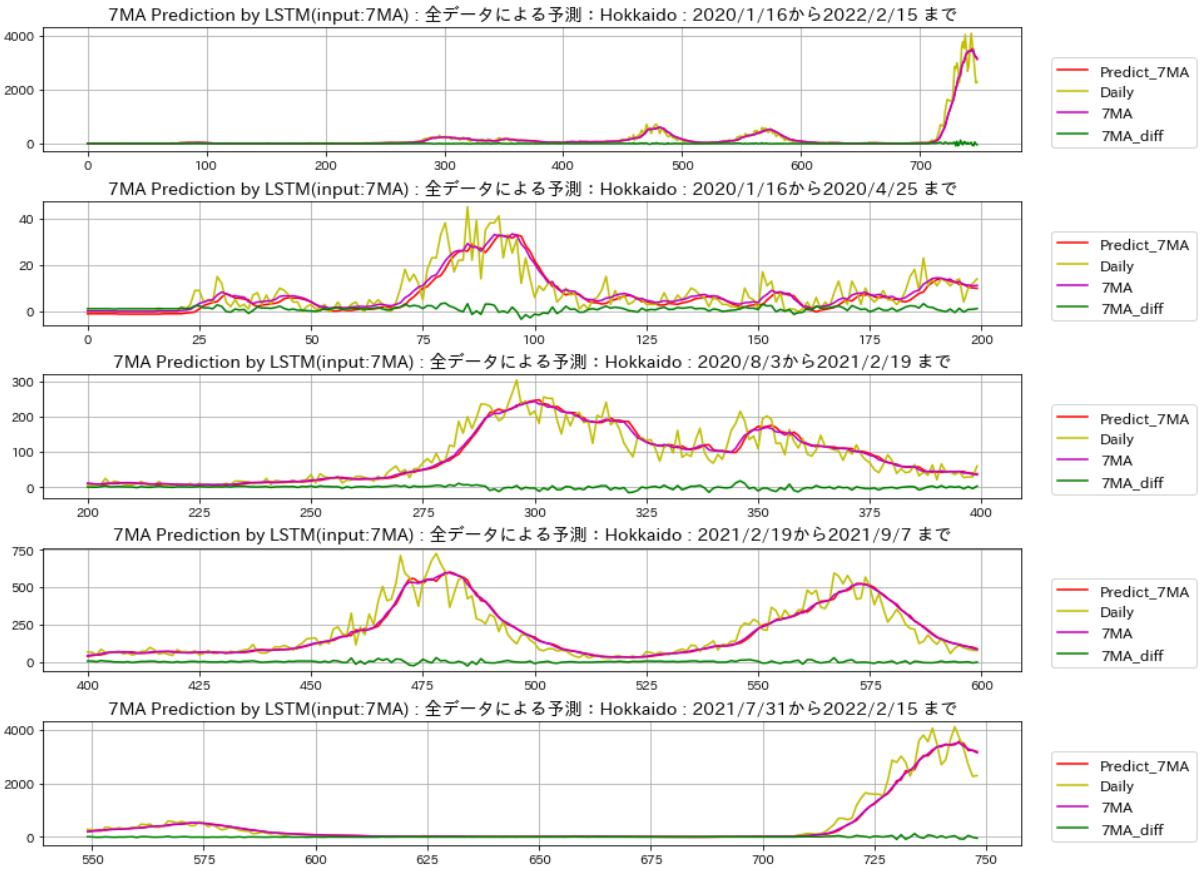


図 15　LSTM を用いた北海道の予測の例（入力と教師データは 7 日平均の値，2020/1/16～2022/2/15））

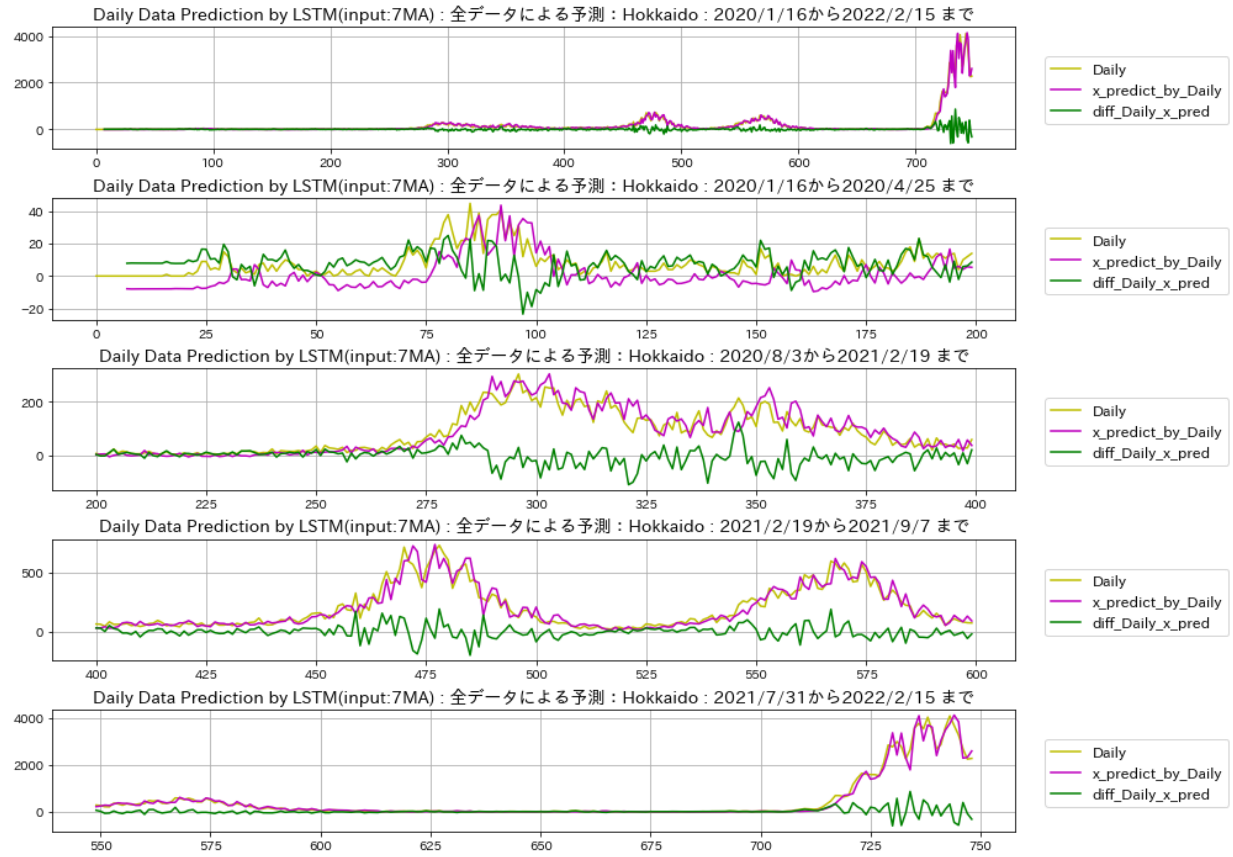
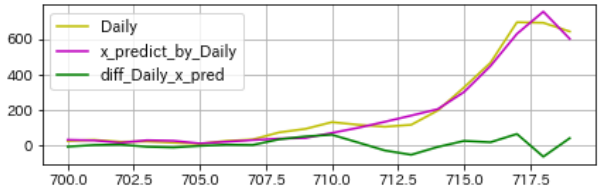


図 16　LSTM を用いた北海道の 7 日平均の予測値から算出した日次予測の例（2020/1/16～2022/2/15）

第６波は，第１波から第５波に比べ急拡大の速度が速く，かつ新規感染者数も数倍の規模となっていることから，訓練データには現れなかったような系列データが出現していると言える．したがって，モデルの高い汎化性能が要求される予測問題と言える．ここで先の比較で予測結果が良かったほうの 4.3 節の入力と教師データが 7 日平均である 1 入力 1 出力モデルを使い，用いるデータの最終日付を異なる 3 つの時期（日付）にして学習と予測を行ってみる．3 つの時期は，第６波の立ち上がり期（2022 年 1 月 17 日），急拡大期（2022 年 1 月 31 日），ピーク期（2022 年 2 月 15 日）とする．それぞれの学習と予測における予測結果を，図 17 で順に示す．

どの予測結果も日次データの実現値に近い値と動きをしていることが分かる．評価誤差は概ね第６波の時期のデータで求められているので，汎化能力がある程度得られていると考えられる．



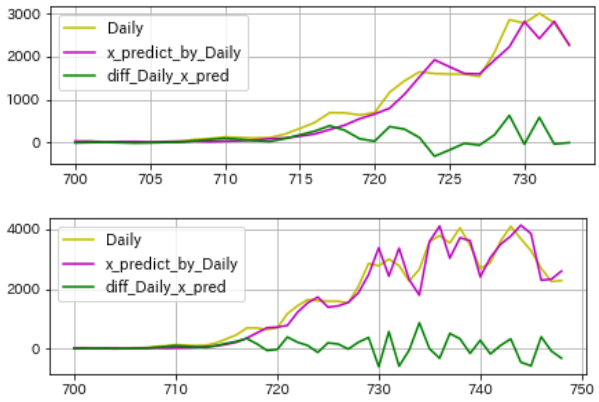


図 17 北海道の第６波の日次予測の例

5. 他の都府県の新規感染者数に関する学習と予測

これまでに述べたモデルによる学習と予測を他の都府県のデータを用いて行ってみる．47 都道府県毎に学習と予測は可能であるが，ここでは東京都と大阪府について学習と予測を行うことにする．

5.1 東京都と大阪府の新規感染者数の学習と予測

まず，東京都のデータ（2020 年 1 月 16 日から第６波のピーク期（2022 年 2 月 15 日））を使い，4.1 節から 4.4 節で述べたモデルによって，それぞれで学習と第６波の予測を行った結果を図 18 に示す．

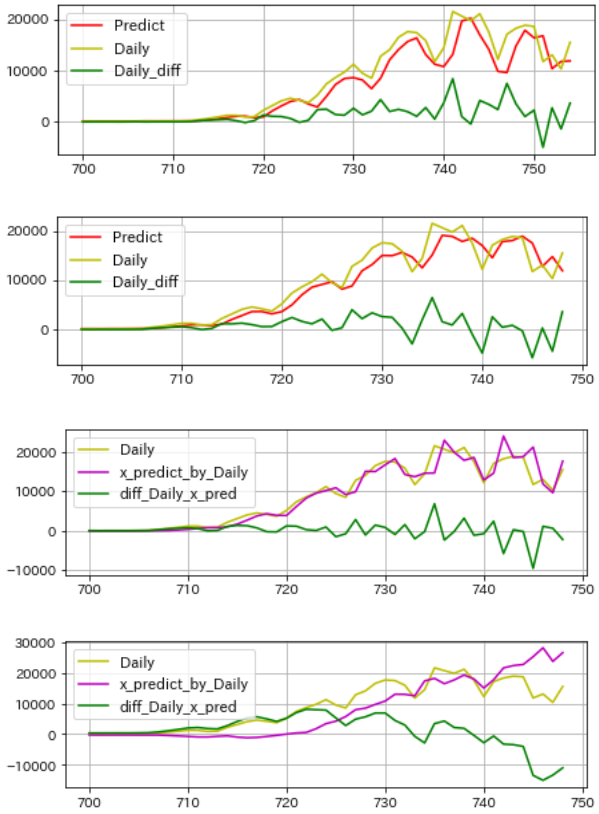


図 18 東京都の第６波の日次予測の例

図 18 から，4.3 節のモデルによる日次予測が最も良く，特に 730 日目頃までが良いことが分かる．

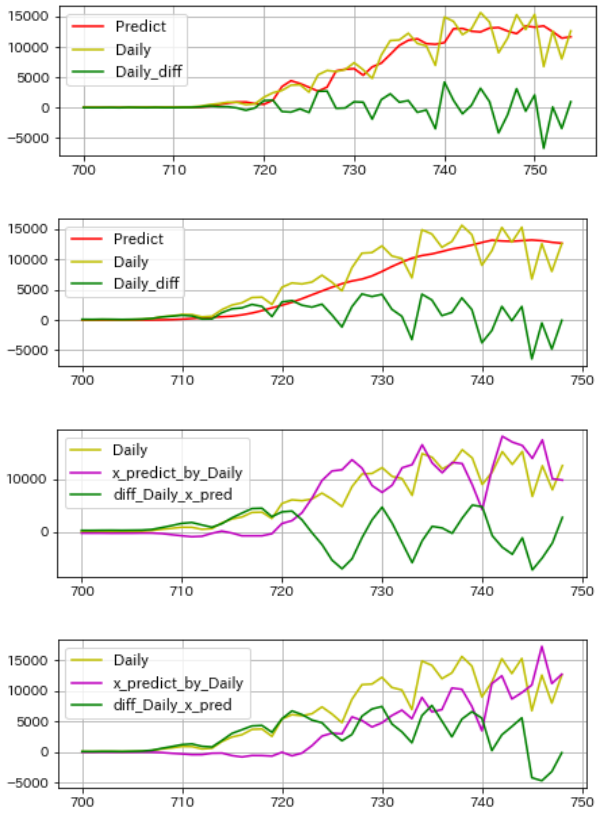


図 19 大阪府の第６波の日次予測の例

次に大阪府の学習と予測結果を図 19 に示した．4.1 節と 4.2 節のモデルが辛うじて日次データの実現値の動きの傾向（トレンド）を捉えているが，どのモデルも予測結果が悪い．そこで，4.5 節の後半と同様に，4.3 節の 7 日平均の 1 入力 1 出力モデルを使って，用いるデータの最終日付を第６波の立ち上がり期（2022 年 1 月 17 日），急拡大期（2022 年 1 月 31 日），ピーク期（2022 年 2 月 15 日）と変えて指定し，それぞれ学習と予測を試みた．その結果を図 20 に示し比較する．縦軸と横軸，および線種の意味は，図 17 と同じであるが，横軸の最大値と縦軸の最大値が各グラフで異なることに留意する．

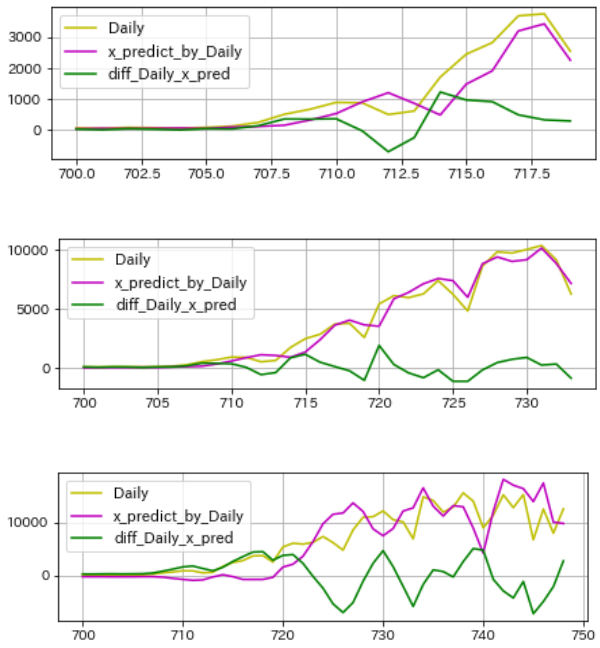


図 20 大阪の第６波の日次予測の例

図 20 の 2 番目のグラフから，急拡大期までのデータによる学習と予測はまずまずの結果と言えるが，他の場合では予測が失敗することが確認できた．汎化性能は，データに依存して大きく変化することもあることが確認できた．予測結果の悪化要因を厳密に特定はできないが，新規感染者数の短期間での急激な増加や，その規模の大きさや不規則さかもしれない．ここで LSTM のモデルを，異常検知を行うオートエンコーダー的な役割の視点から考えてみる．その場合に予測の破綻は，それまでに出現したデータには無かったようなデータが出現したということを意味するとも言える．したがって，予測結果の悪さや破綻にも，未知の感染拡大パターンの出現の検知という役割があると考えると，予測不良であってもそのモデルにも有用性はあるとも言える．

5.2 隣接県の新規感染者数に関する学習と予測

都道府県間の相関分析を行ってみると，大都市圏内での相関が高い傾向にある[9]．ここで，東京都の隣接県のデータを用いて4.3節のモデルで学習と予測を行い，さらに日次データ予測を行った結果を図21に示す．順に，神奈川県，埼玉県の予測結果である．両方とも日次の実現値の大きな増減傾向は予測できているが，細かい変化までは予測はできていない．両県の実現値は，スケールは異なるものの似た動きをしているので，学習データとして両県の日次データを結合したデータセットを用いることも有効かもしれない．

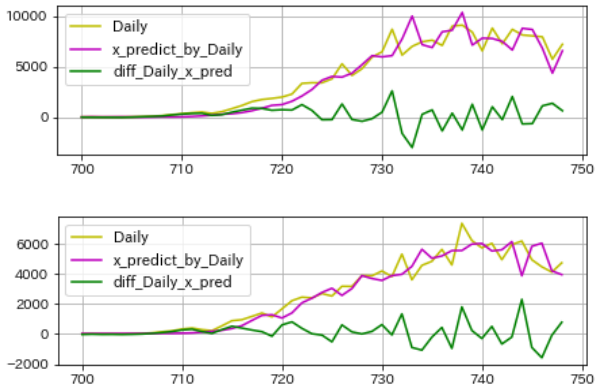


図21 東京都の隣接県（上：神奈川，下：埼玉）の第６波の予測の例

6. まとめ

COVID-19 の新規感染者数を時系列データと捉えて，機械学習の手法である LSTM を用いた深層学習による学習と予測を試みた．7日平均の学習と予測を行い，その予測値から日次データを予測する方法の有効性を確認できた．これは7日平均の学習と予測が良好であることに起因すると考えられる．

しかしながら，有効ではないケースもあることも確認できた．また，汎化性能を高めた最良のモデルでも全期間を通じてよい予測ができないことは，ノーフリーランチ定理の実例とも言えるかもしれない．より良い予測ができるようなモデルの構築のためには，LSTM のユニット数や学習時のバッチサイズなどのハイパーパラメータの最適化や，過学習対策などの様々な工夫，及び他の時系列データの学習モデルの導入が考えられる．特に LSTM は，長い系列のデータを学習することが可能という特徴を持っているので，系列長を長くした最適化も必要であろう．しかしながら，第６波のように，長い安定状態から短期間での急上昇を予測する問題では，系列長を長くするのは逆効果かもしれない．

本稿ではいくつかの知見は得られたが，まだ十分ではない．今後も増えてゆくデータを加えながら，モデルの学習を継続して行い，今回の分析手法の改良や他の様々な手法の応用を行い，新たな知見を増やしてゆく必要がある．またその知見をどのように有効利用してゆくかを，十分に検討し検証することも重要である．他の都道府県の学習と予測や，世界各国毎のモデルの構築，および感染拡大状況に応じた複数のモデルの構築も必要と考えられる．

参考文献

(1) 厚生労働省 ： 2022 年 3 月 3 日, https://www.mhlw.go.jp/stf/covid-19/open-data.html.
(2) ＮＨＫ特設サイト新型コロナウィルス:新型コロナデータ一覧 : 都道府県ごとの感染状況, 2021 年 6 月 30 日, https://www3.nhk.or.jp/news/special/coronavirus/data-widget/.
(3) World Health Organization: WHO Coronavirus (COVID-19) Dashboard, 2022 年 2 月 27 日, https://covid19.who.int/info/.
(4) COVID-19 Public Forecasts: 2021 年 3 月 3 日, https://cloud.google.com/blog/ja/products/ai-machine-learning/google-and-harvard-improve-covid-19-forecasts.
(5) S. Hochreiter, J. Schmidhuber: Long short-term memory, Neural computation, MIT Press, 1997.
(6) I Goodfellow, Y Bengio, A Courville: Deep Learning (Adaptive Computation and Machine Learning series), The MIT Press, 2016.
(7) Gerald Appel: Technical Analysis: Power Tools for Active Investors, Financial Times Prentice Hall, p.166, 2005.
(8) Diederik P. Kingma, Jimmy Ba:Adam Adam: A Method for Stochastic Optimization, https://arxiv.org/pdf/1412.6980.pdf.
(9) 機械学習を用いたＣＯＶＩＤ－１９新規感染者数の分析, 小松隆行, 北海道科学大学研究紀要, (49), 49-61 (2021-09-30).