

機械学習を用いたCOVID-19新規感染者数の分析

Data Analysis on COVID-19 Newly Infected Patients 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（新型コロナウイルス）に関するオープンデータを、統計的な手法や機械学習の手法などを使い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、移動平均とそれを応用した分析、教師なし学習によるクラスタリング、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を深層学習のアルゴリズムを用いて学習した将来の新規感染者数予測について報告する。

1. はじめに

2020 年 1 月頃から現在も継続している COVID-19（新型コロナウイルス）のパンデミックは、人類の健康だけでなく社会的危機や経済的危機を生じさせ、大混乱が続いており収束がいまだ見通せない状況にある。これに関連する情報は、数値データのオープンデータとして、全世界の多くの機関がインターネット上に公開しており、それらをダウンロードし分析や解析を行うことが可能となっている。これらのサイトから新規感染者や現在患者数、累積死者数、対策病床数など多種類のデータがオープンデータとして CSV 形式や Excel 形式でダウンロード可能であり、これらのデータの推移をグラフ化するなどして可視化したり、基本的な統計分析したりした結果や知見は、日本でも厚生労働省のインターネットサイトなどにおいて公開されている⁽¹⁾⁽²⁾。厚生労働省では、全国のすべての都道府県毎のデータを集約したデータを公開しており、毎日データが追加されて常に最新のデータをダウンロードして取得することができるようになってきている⁽¹⁾。北海道においても道内の大都市や振興局単位での感染状況のデータが公開されており、ダウンロードし解析することを可能にしている⁽³⁾⁽⁴⁾。しかしながら、これらのデータの詳しい分析結果は、まだそれほど公開されていない。時系列の推移のグラフなどは多く公開されているが、新規感染者数の累積数や日次データその

もの、7 日移動平均などの単純なものが多く、本格的なデータサイエンスの視点からはまだ分析はなされていないようである。

本報告の目的は、COVID-19 の新規感染者数のデータの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何らかの知見を得ることとする。具体的には、ファイナンス分野⁽⁵⁾の時系列データ解析で多用されている移動平均とそれを応用した分析、機械学習の分野で教師なし学習と呼ばれている手法の一つである Time Series k-means 法⁽⁶⁾によるクラスタリング（自動グループ分け）、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）の発展形であり学習性能が高いとされる LSTM（Long Short-Term Memory）⁽⁷⁾を用いたニューラルネットワークを深層学習のアルゴリズムを用いて学習した将来の新規感染者数予測を試みる。

2. 北海道内の新規感染者数に関する分析

ここでは、まず北海道のデータについて分析してみる。今回は、2020 年 1 月 28 日から 2021 年 3 月 2 日まで北海道（保健福祉部地域保健課発表）のオープンデータ⁽⁴⁾をダウンロードし使用した。データには場所のデータも含まれているため、北海道内の主要都市や振興局単位での集計分析が可能である。

2.1 移動平均による傾向分析

新規感染者数の日次データは曜日依存した変化をする傾向があり、曜日毎のデータの比較をするようになってきた。例えば、休日明けに報告される新規感染者数は相対的に少なめであると言われている。日々の連続の時系列データの値の増減が大きく激しい変化の推移になる傾向があるため、まず時系列データを分析する手法として最も一般的に用いられている移動平均を使い分析してみる。一般的には、ある一定期間（ s 日とする、例えば 7 日など）の各日の値の総和を、この期間 s で割った値である単純移動平均（SMA, Simple Moving Average）が用いられている。COVID-19 の新規感染者数の 7 日移動平均であれば、7 日間の日次データの合計を期間（7 日）で割った値を 7 日の最終日にプロットする。 $s=7$ であれば短期移動平均といえるが、 $s=14, 28, 56, 84$ などと、 s の値を大きくしてゆくと、 s 日移動平均は長期移動平均になってゆく。ファイナンスの分野では、これらを用いて、予測や傾向を分析することが行われている。

図 1 は、北海道のデータから抽出した札幌市の新規感染者数の日次データを用いて、数種類の s 日移動平均線を描いたグラフである。ここでは、 $s=7, 14, 28, 56, 84$ についてプロットした。 s の値は、報道で頻繁に使用されてきている 7 日移動平均を基準にして、7 日の 2 倍の 14（感染から発症までの日数の目安）、14 の倍数として 28, 56, 84 を採用した。いわゆる第 3 波の開始時期に近い 2020 年 10 月中旬から、すべての移動平均線が急激な上昇をしており、その上昇の度合いは短期移動平均線ほど極端に大きいことがわかる。同様のことは、山のピークは低いものの第 1 波の開始時期の 2020 年 4 月上旬の移動平均線についても言える。このグラフはファイナンス分野（株取引など）で用いられているチャートに相当するが、時系列データのトレンド（上昇傾向か下降傾向か）を把握する手法として、移動平均を使ったテクニカル分析の考え方適用してみる。短期移動平均線が長期移動平均線を下から上に突き抜けることは上昇傾向が強くなったというサインであるとされ、2020 年 9 月から 10 月の頃や 2021 年 1 月上旬の移動平均線には、そのサインが表れていると言える。また、第 3 波のピークが過ぎ新規感染者数が減少してゆく 2020 年 11 月下旬から 12 月の頃や 2021 年 1 月下旬から 2 月の頃には、短期移動平

均線が長期移動平均線を逆に上から下に突き抜け、下降傾向が強くなったというサインが見られる。

さらに、別の指標を使って分析してみる。基本となるのは、MACD (moving average convergence/divergence, 移動平均収束拡散)⁽⁸⁾ である。これはファイナンスの分野で使われているトレンドの変化を判断するテクニカル分析の指標であり、以下の式で定義される。

$$\text{MACD} = (\text{短期 EMA}) - (\text{長期 EMA}) \quad (1)$$

$$\text{MACD シグナル} = \text{MACD の EMA} \quad (2)$$

ここで、EMA とは指数平滑移動平均 (Exponential Moving Average) のことであり、単純移動平均とは異なり加重移動平均の重みを過去に遡るにつれて値を半分にして平均をとるものである。EMA では直近の値を重視して、より重きを置くように加重されており、値の変動に対する感度が高い。MACD が MACD シグナルを突き抜けることでトレンドの変化を判断する。ファイナンスでの MACD は、式 (1) において短期 EMA は 12 日 EMA を、長期 EMA は 26 日 EMA を、式 (2) において MACD シグナルでは 9 日 EMA を使用している場合が多い。今回の COVID-19 新規感染者数の分析では、EMA の期間を変更し新たな指標として MACDc と MACDc シグナルを以下のように定義した。

$$\text{MACDc} = (14 \text{ 日 EMA}) - (28 \text{ 日 EMA}) \quad (3)$$

$$\text{MACDc シグナル} = \text{MACD の 7 日 EMA} \quad (4)$$

図 2 は、2020 年 10 月 1 日から 2021 年 3 月 2 日までの札幌市の新規感染者数の日次データとその 7/14/28 日 EMA のグラフと、それと同じ期間の対応する MACDc と MACDc シグナルをプロットしたグラフである。MACDc が MACDc シグナルを上から下に突き抜けたポイント (2020 年 11 月 15 日を過ぎた頃) が下降傾向の開始のサインになっていると見られる。また弱い上昇傾向であるが、下から上に突き抜けた 2020 年 12 月下旬のポイントは、その上昇傾向の開始のサインになっているように見られる。同様に、他の MACD の知見が MACDc に活かせるかもしれない。

また、ある一定期間の時系列データの標準偏差 σ を用いて、移動平均線に関して $\pm \sigma$, $\pm 2 \sigma$ の補助線を引いて、平均に対する実現値の動きを分析してみた。ファイナンス分野ではボリンジャーバンド⁽⁹⁾と呼ばれており、平均値に対して高過ぎか安過ぎかなどを判断する手法である。

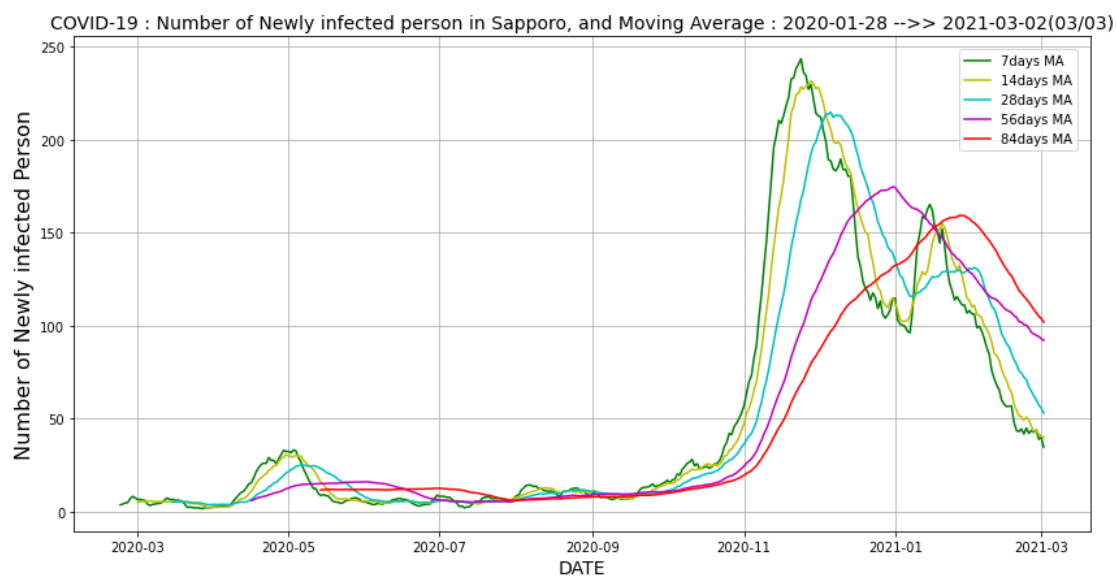


図 1 札幌市の新規感染者数の移動平均のグラフ (2020/1/28~2021/3/2)

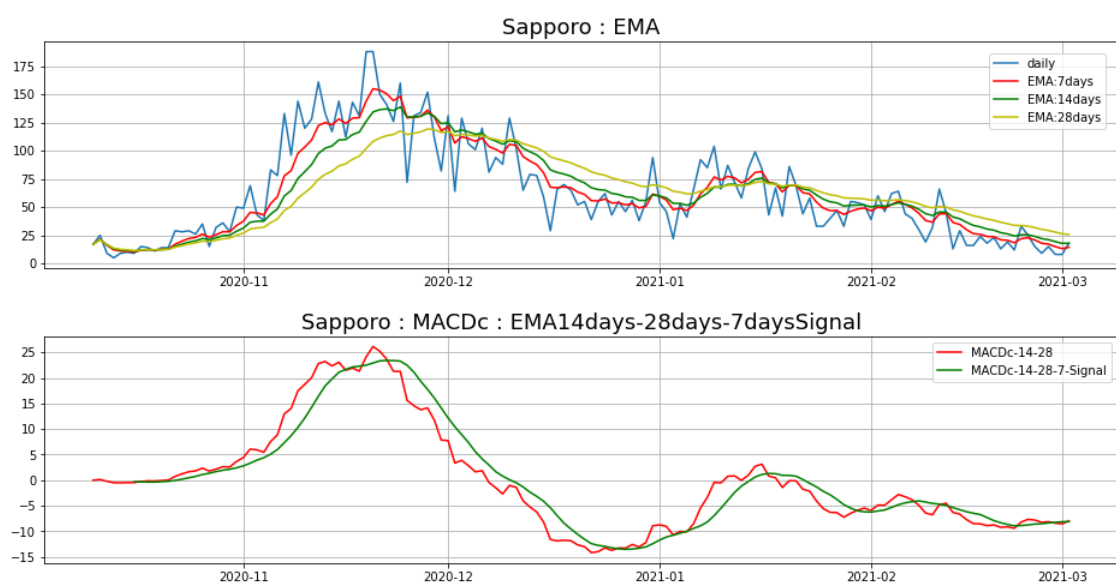


図 2 札幌市の新規感染者数の 14 日 EMA/MACDc/MACDc シグナルのグラフ (2020/10/1~2021/3/2)

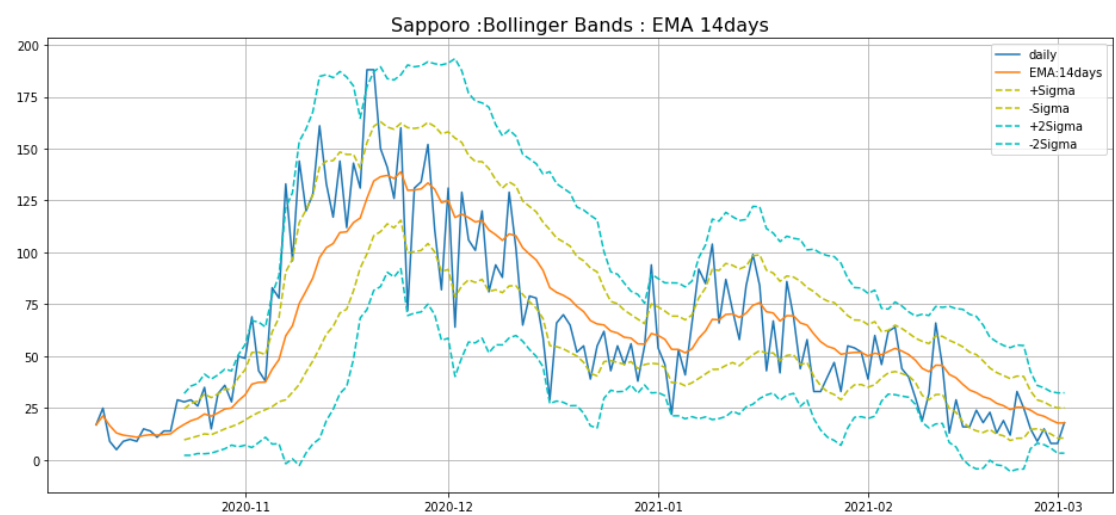


図 3 札幌市の新規感染者数の日次データと 14 日 EMA/ $\pm\sigma$ / $\pm2\sigma$ の各補助線 (2020/10/1~2021/3/2)

図3は、2020年10月1日から2021年3月2日までの札幌市の新規感染者数の日次データ、14日EMA、 $\pm\sigma$ 、 $\pm2\sigma$ の補助線を描いたグラフである。2020年10月後半から、日次データがEMAと $+2\sigma$ の間で遷移して増加傾向を示し、バンド幅($\pm2\sigma$ 補助線の上下幅)も増大していることが分かる。

2.2 地域（主要都市や振興局）間の相関分析

ここでは、道内の主要都市および振興局毎のデータの相関分析を行う。任意の2つの地域毎に全期間のデータで相関係数を算出し、その値を図4と図5でヒートマップ表現した。第1波の時期は地域間の相関はそれほど高くないが、第3波の時期には多くの地域間の相関が高くなっていることが分かる。

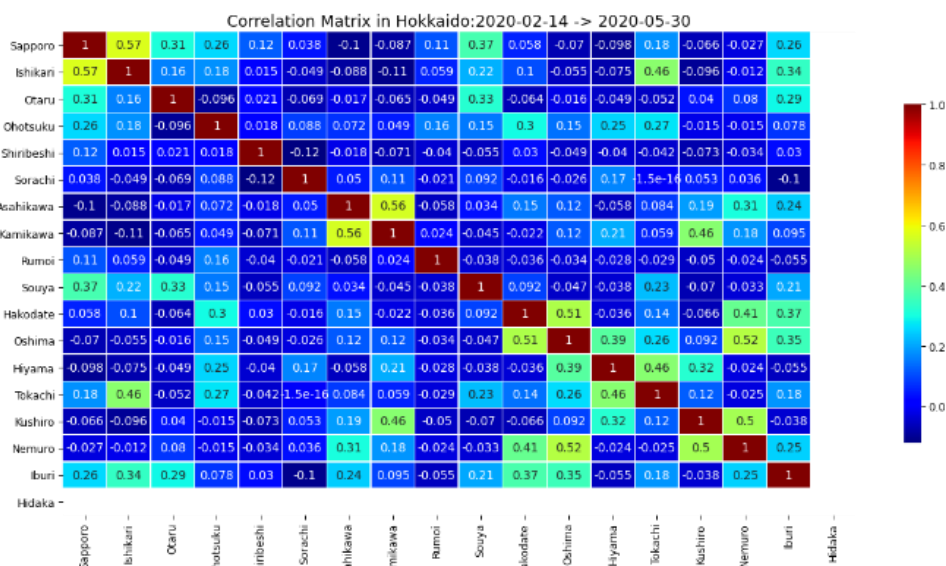


図4 北海道内の地域間の相関係数ヒートマップ（第1波：2020/2/14～2020/5/30）

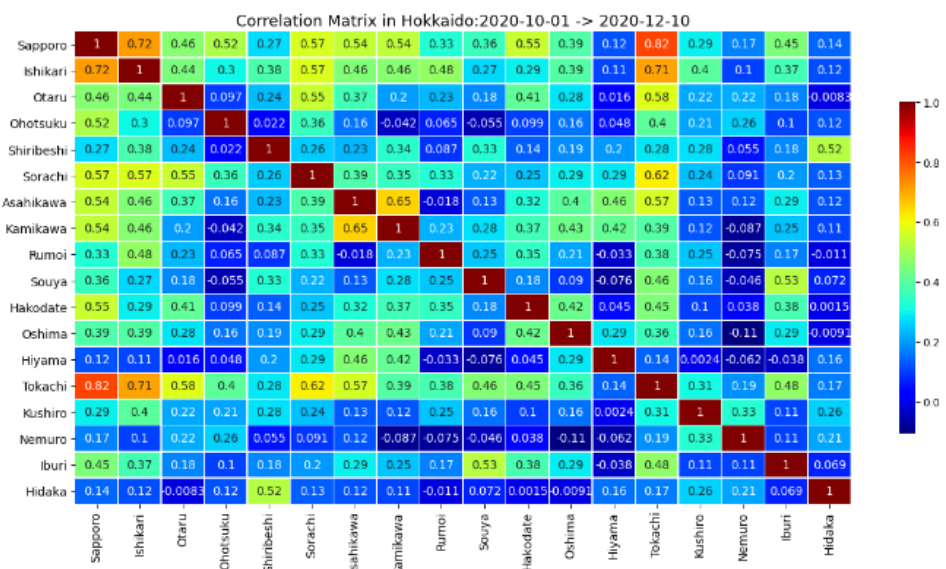


図5 北海道内の地域間の相関係数ヒートマップ（第3波：2020/10/1～2020/12/10）

2.3 地域（主要都市や振興局）のクラスタリング

ここでは、道内主要都市と振興局毎の人口10万人当たりの新規感染者数の推移の時系列データを、Times Series k-Means法⁽⁶⁾を用い、クラスタリングを試みる。クラスタリング数はエルボー法により7

とし、その結果を図6に示す。同じクラスタになったものを同じグラフにプロットしている。隣接地域が同クラスタであったり（札幌と石狩、函館と渡島など）、目立った感染クラスタが発生した地域は単独で分類されていることが分かる（旭川や桧山）。

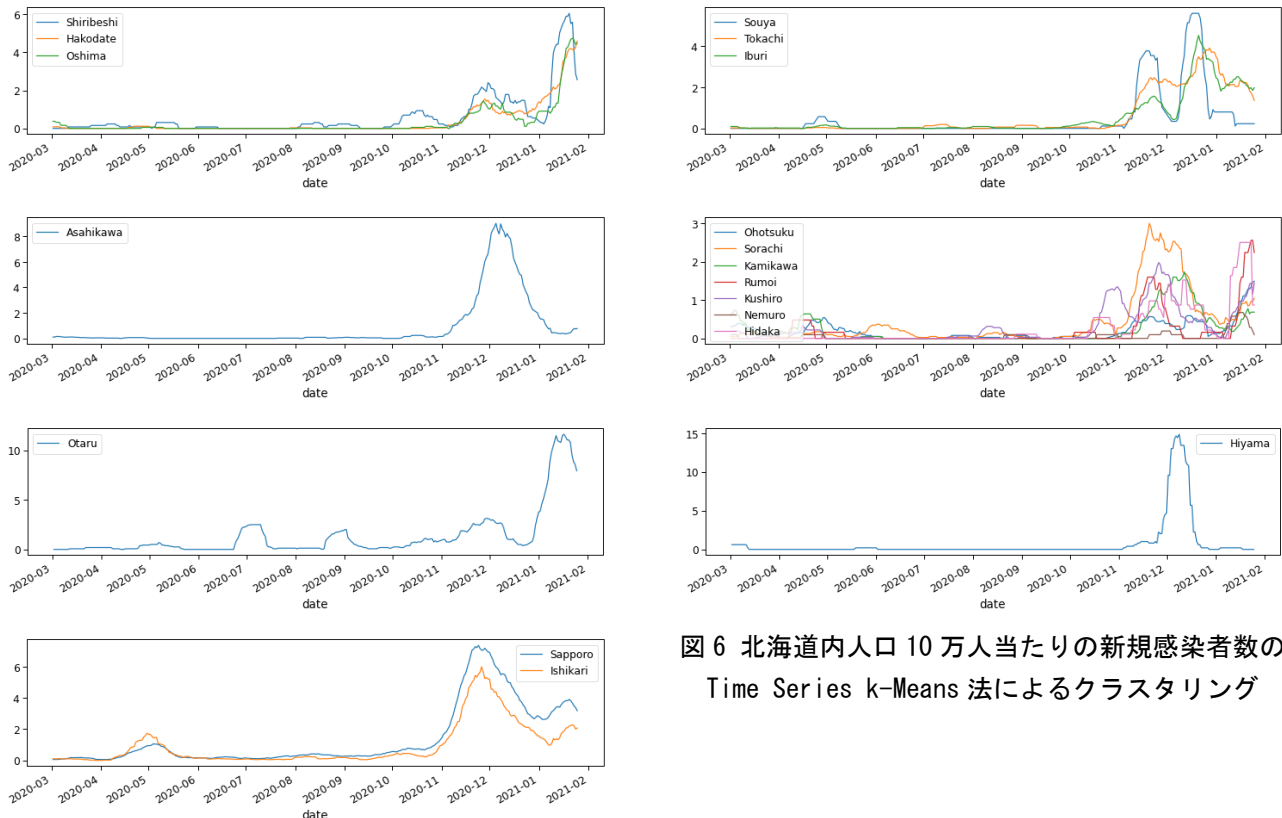


図 6 北海道内人口 10 万人当たりの新規感染者数の
Time Series k-Means 法によるクラスタリング

3. 全国の新規感染者数に関する分析

次に、全国のデータについて分析する。厚生労働省が都道府県毎に集計して発表しているオープンデータ⁽¹⁾を使用した。

3.1 移動平均による傾向分析

2.1 節と同様に、東京の第 3 波の時期の EMA/MACDc/MACDc シグナルのグラフを図 7 に示す。MACDc が MACDc シグナルを上から下に突き抜けたポイントが見られる 2021 年 1 月上旬の頃から、日次データの下降傾向が始まっていることが分かる。

3.2 全国の都道府県間の相関分析

次に相関分析を行う。関東 1 都 7 県（東京、神奈川、千葉、埼玉、群馬、栃木、茨城、山梨）の地域間の相関係数を第 1 波から第 3 波の期間毎に算出し、その値のヒートマップを図 8 に示す。第 1 波では首都圏内のみが相関が高かったが、第 2 波、第 3 波に進むにつれて、相関係数値も高くなり、北関東との相関も高くなった。図 9 は、全都道府県間の全期間での相関係数のヒートマップである。大都市がある地域間やそれらと隣接地域間の相関が高い傾向がある。

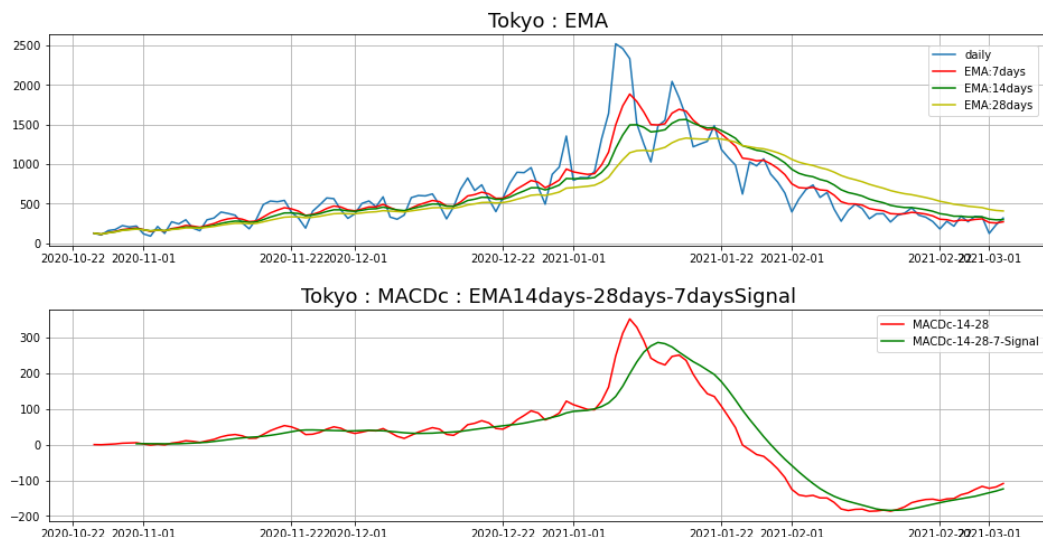


図 7 東京の新規感染者数の 14 日 EMA/MACDc/MACDc シグナルのグラフ（東京：2020/10/25～2021/3/3）

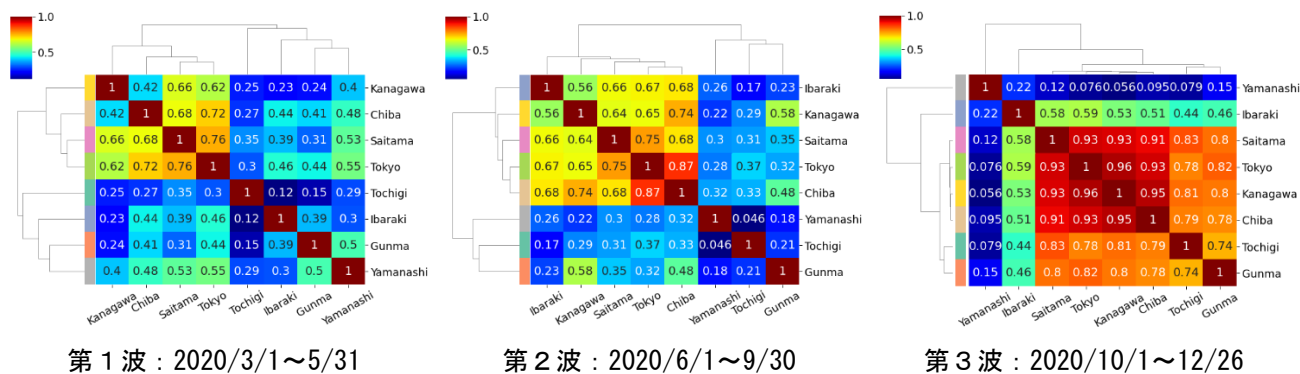
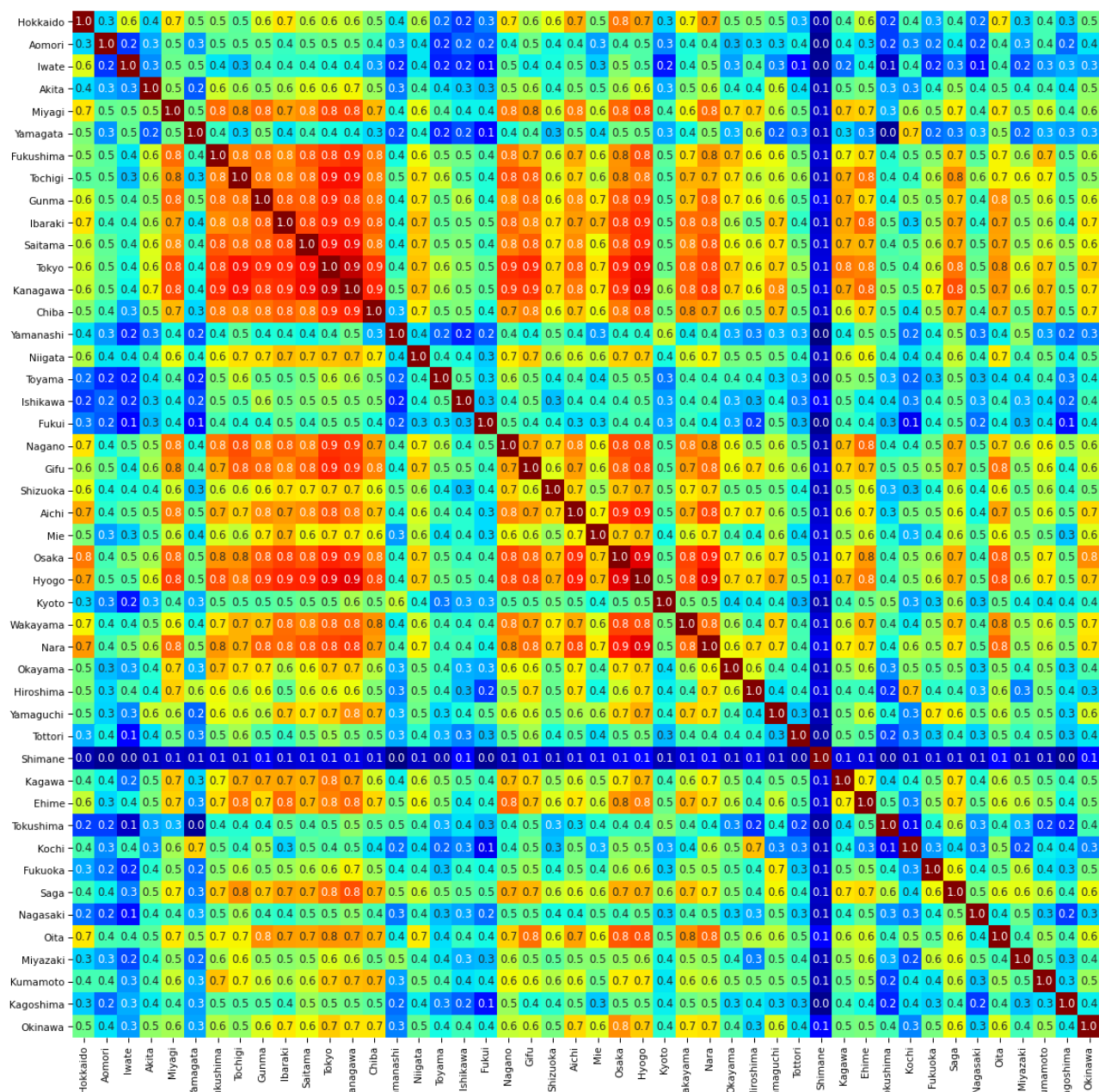


図8 関東1都7県（東京、神奈川、千葉、埼玉、群馬、栃木、茨城、山梨）の相関ヒートマップとデンドログラム



3.3 全国の都道府県間のクラスタリング

ここでは、全国 47 都道府県の 10 万人当たりの新規感染者数の推移を 2.3 節と同様に Time Series k-means 法⁽⁶⁾を用いてクラスタリングを行った。クラスタ数はエルボー法により 10 とした。図 10 は、クラスタ毎に同じクラスタと判断された都道府県の推移をプロットしたグラフである。大都市や遠隔地は単独（東京、大阪、北海道、沖縄など）、大都市周辺地域（千葉と埼玉、静岡と三重など）や隣接県（栃木と群馬、佐賀と長崎、四国など）は同じクラスタになっていることなどが分かる。

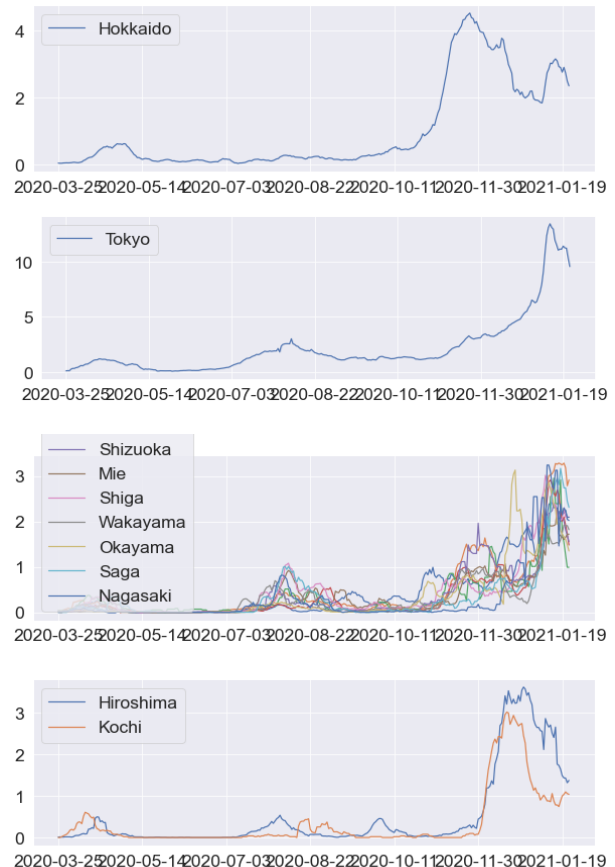
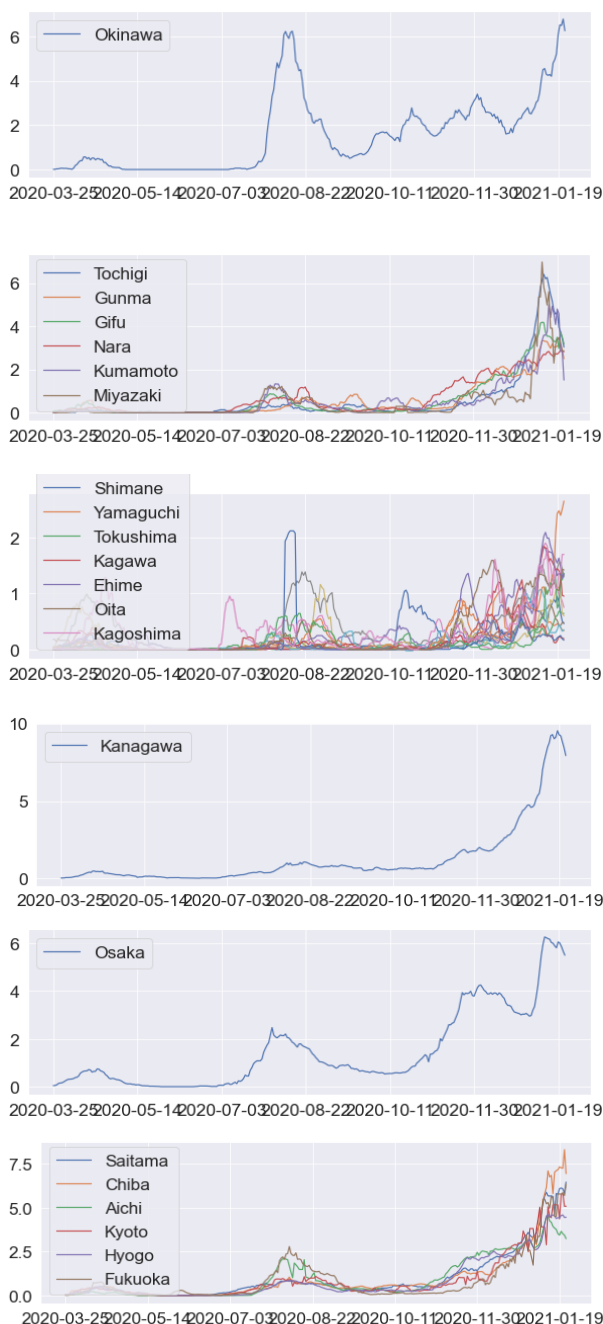


図 10 全国の人口 10 万人当たりの新規感染者数の Time Series k-Means 法によるクラスタリング

3.4 深層学習による予測

Google 社が新規感染者数の予測⁽¹⁰⁾⁽¹¹⁾を公開している。ここでは、東京の新規感染者のデータを用いて、深層学習を使った独自の予測モデルの構築を試みる。使用するデータは、2020 年 1 月 24 日～2021 年 1 月 26 日の 14 日移動平均のデータである。モデルは、ある連続する 9 日間の 14 日移動平均データからそれに続く翌日、すなわち 10 日目の 14 日移動平均の値を予測するものである。ネットワークは時系列データのモデリングに用いられる学習能力が高い LSTM (Long Short-Term Memory)⁽⁷⁾を多段につなげたモデルを用いる。モデルの構造は入力層→LSTM3 個→全結合層である。損失関数は平均二乗誤差とした。ここでは、LSTM のサイズを 98 から 110 までの値で離散的に変えながら学習と予測を試みた。学習データは、2020 年 1 月 24 日～2021 年 1 月 12 日の 14 日移動平均のデータを 10 日毎にデータセットとしてまとめ、それぞれ最初の 9 日の時系列データから 10 日目を予測するように学習する。その結果のモデルを用いて、2021 年 1 月 13 日～2021 年 1 月 26 日の 14 日移動平均のデータを予測した結

果を図 12 に示す。横軸は、日付であるが、ここでは day1 から day23 が 2021 年 1 月 4 日～2021 年 1 月 26 日に対応する。実線 (daily) は記録された実現値であり、day10 から day23 までの点線はそれぞれ LSTM のサイズを変えたモデルでの予測値である。LSTM のサイズを変えるにしたがって、予測の点線が変化してゆき、実現値の実線グラフは、ほぼこれらの点線でおおわれる領域内に収まっている。この結果から、LSTM のサイズを変えたモデルをそれぞれ学習し、それらから将来の値の予測範囲を推測することができる可能性を示唆していると考えられる。

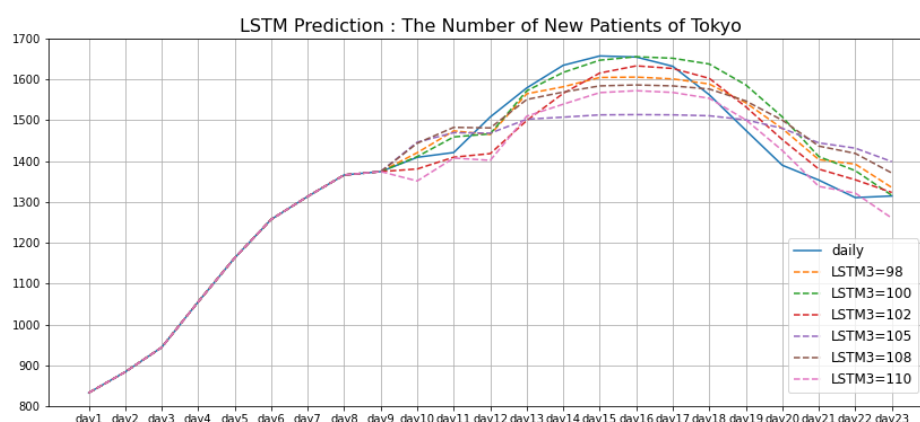


図 12 LSTM モデルを用いた深層学習による予測 (2021 年 1 月 13 日～2021 年 1 月 26 日を予測)

参考文献

- (1) 厚生労働省 : 2021 年 3 月 3 日, <https://www.mhlw.go.jp/stf/covid-19/open-data.html>.
- (2) COVID-19 Japan : 2021 年 3 月 3 日, <https://www.stopcovid19.jp/>.
- (3) 北海道庁:新型コロナウイルス感染症 (COVID-19) に関する情報 : 2021 年 3 月 3 日, <http://www.pref.hokkaido.lg.jp/ss/ssa/sin-gatakoronahaien.htm>.
- (4) 北海道庁:新型コロナ:道内の発生状況一覧, 2021 年 3 月 3 日, <http://www.pref.hokkaido.lg.jp/hf/kth/kak/hasseijoukyou.htm>.
- (5) Yahoo! Finance: 2021 年 3 月 3 日, <https://finance.yahoo.com/>.
- (6) Xiaohui Huang, Yunming Ye, Liyan Xiong, Raymond Y.K. Lau, Nan Jiang, Shaokai Wang: Time series k-means: A new k-means type smooth subspace clustering for time series data, Information Sciences, Volumes 367-368, 1 November, pp. 1-13, 2016.
- (7) S. Hochreiter, J. Schmidhuber: Long short-term memory, Neural computation, MIT Press, 1997.
- (8) Gerald Appel: Technical Analysis: Power Tools for Active Investors, Financial Times Prentice Hall, p.166, 2005.
- (9) John Bollinger's Official Bollinger Band Website: 2021 年 3 月 3 日, <https://www.bollingerbands.com/>.
- (10) 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>.
- (11) COVID-19 感染予測 (日本版): 2021 年 3 月 3 日, <https://datastudio.google.com/reporting/8224d512-a76e-4d38-91c1-935ba119eb8f/page/ncZpB>.