機械学習を用いたコロナ感染者数予測に用いるデータの検討

宮城県仙台二華高等学校　非会員　菅原玄晶
宮城県仙台二華高等学校　非会員　千葉幹
宮城県仙台二華高等学校　非会員　小関祐也
東北大学大学院工学研究科　正会員　佐野大輔
東北大学大学院工学研究科　非会員　飯塚勇仁
宮城県仙台二華高等学校　非会員　深沢恭子

１．はじめに

現在，新型コロナウイルス感染症（COVID-19，以下コロナ）が世界中で猛威を振るい，日本でも毎日多数の感染者や，それに起因する死者が発生している．また，その流行の程度に伴って，様々な行事の中止や実施方法の変更の検討がしばしば行われている．もし，コロナの流行をあらかじめ予測できれば，感染対策を事前に行うことで流行を抑制したり，行事実施可否の判断を根拠に基づいて行うことができ，様々な活動への影響を軽減することができると考えられる[1)]．そこで本研究では，コロナに対する人々の意識など，計測することのできない情報を Web 上のキーワード検索数などによって間接的に数値化できる点に着目し，これまであまり用いられてこなかった変数を入力データとして用いたコロナ感染者数予測の機械学習モデルを構築し，用いたデータの適用性を評価した．

２．方法

コロナ感染者数を予測するために用いたデータは以下のとおりである．すなわち，(a)コロナ感染者数，(b)下水中コロナ調査結果，(c)米ドル/円為替（為替の変動により出入国人数が変動すると考えた），(d)GoogleTrends における日本国内での「コロナ」検索数（人々のコロナに対する関心を示すものと考えられ，宮城県のデータがなかったため日本全体のデータを用いた）(e)人流データ（スマートフォンの位置データによる仙台駅周辺の午前七時の人口データ），及び(f)降水量（外出頻度に影響），である．なお，どのデータも期間は 2020/8/3(月)～2022/9/5(日)のものを使用した．以上のデータのうち，(a)及び(b)は常に使用し，(c)～(f)については，いずれか１つ，２つ，３つ，もしくは 4 つ全て用いることで，計 15 個のモデルを作成した．機械学習には人工ニューラルネットワークを用い，学習・検証の繰り返し回数は 3000 回とした．予測精度の比較には平均二乗誤差（MSE）を用いた．

３．結果および考察

予測精度の比較に用いた MSE は、予測と実際の値の差の二乗の和であり、MSE が小さいほど予測精度が高いことを示す．本研究で得られた MSE の値を図 1 に示した．結果から以下の２点が指摘できる．1 点目は、為替と GoogleTrends のデータの組み合わせが比較的高い予測精度を与えている点である．用いたデータの種類が２つ、３つ、及び４つのモデルでは、それぞれ為替と GoogleTrends が含まれているモデルが比較的小さな MSE を与えている．このことは、為替と GoogleTrends のデータが日本国内のコロナ患者数と比較的強く連動していることを示唆している．為替の場合、日本国内に新型コロナウイルスを持ち込む可能性のある外国からの来日人数を左右している可能性がある．また、GoogleTrends は、日本国内において Google を用いて「コロナ」がキーワードとして検索された頻度に関するデータであるが、症状等が出始めた感染者が当該キーワードを使用することが多い可能性が考えられる．２つ目は、人流データと降水量のデータをモデルに入れた場合、予測精度は向上しなかった点である．加えたデータの種類が２つ、３つ、及び４つのモデルにお

キーワード　新型コロナウイルス感染症，流行予測，機械学習，GoogleTrends，為替
連絡先　〒984-0052　宮城県仙台市若林区連坊 1-4-1　宮城県仙台二華高等学校　ＴＥＬ022-296-8101
〒980-8579　仙台市青葉区荒巻字青葉 6-6-06　東北大学大学院工学研究科　ＴＥＬ022-785-7481

いて、人流データと降水量が含まれているモデルにおけるMSEが最も大きくなっていた．これは、人流データが実際に外出した人口を表しているのに対し、降水量のデータも外出する人数に関連するものであり、同じような傾向を表すデータであるためであると考えらえた．

以上の考察から、15 個のモデルのうち一番 MSE が小さくなった、為替データ、GoogleTrends データ、及び人流データの３つを使ったモデルが、今回試したモデルの中では最も予測精度が高いモデルであると考えられた．このモデルを用いた予測結果を図２に示した．概ねよく予測されているが、コロナ感染者数の増加、減少するタイミングが、実際の値より少し遅れているとも見受けられる．さらなる精度向上のためには、用いるデータの種類の多様化や、他の機械学習手法の採用方法などが必要と考えられる．

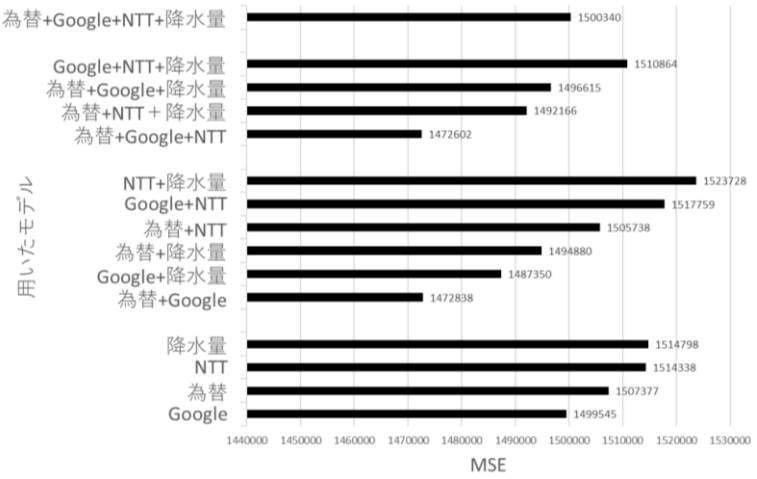

図１.MSE によるモデル間の予測精度評価

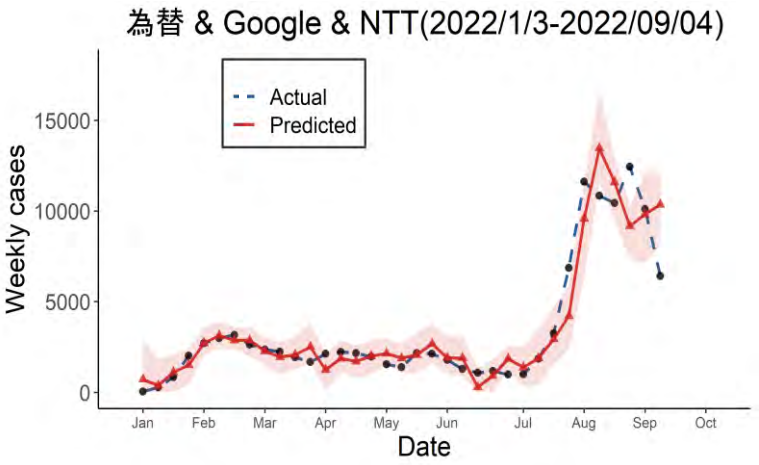

図２．コロナ感染者数予測結果

４．おわりに

本研究では、コロナ感染者数、下水中コロナ調査結果のほかに、為替、GoogleTrends、人流データ、降水量のデータを用いた機械学習によるコロナ感染者数予測モデルの精度の比較をすることで、用いたデータの妥当性を考察した．結果として、為替と GoogleTrends を組み合わせて使用することで予測精度が向上し、一方で人流データと降水量のデータの組み合わせは予測精度の向上に貢献しなかった．これらの結果は、為替と GoogleTrends はコロナ感染者数と比較的強い関連があったことに対し、人流データと降水量は共に同じような傾向を表すデータであったためと考えられる．

謝辞

本研究は、「三菱みらい育成財団」の助成を受けたものです。また、本研究で用いた人流データは、株式会社ドコモ・インサイトマーケティングにご提供いただき、下水中コロナ調査は仙台市の協力を得て国立研究開発法人日本医療研究開発機構 新興・再興感染症研究基盤創生事業（海外拠点研究領域）により行いました。ここに謝意を表します。

参考文献

１）佐野大輔、下水中新型コロナウイルス調査結果に基づく感染陽性判定者数予測、水環境学会誌、2021、vol. 44、no. 3、p. 383-386