Vol.2021-MPS-132 No.22
2021/3/2

ランダムフォレストを用いたインフルエンザ流行予測の試み

吉田　総太郎[1]　林　亮子[1,a)]

概要：現在新型コロナウイルス感染症 COVID-19 が世界中で流行しており，2021 年 1 月末ごろの段階で世界中の累積感染者数が 1 億人，累積死亡者数が 216 万人と言われている．COVID-19 は重症化すると死亡率が高く，犠牲者を減らすために日常生活を大きく制限している国が多いため，世界中で経済的な影響が出ている深刻な状況である．新型コロナウイルス感染症に関する蓄積データはまだ少ないために，現段階でデータマイニングを行うことは困難であるが，インフルエンザ感染症はすでに一定の蓄積データがあり，インフルエンザ感染症の理解が COVID-19 にもある程度参考になることが期待できる．そこで本稿では，統計プログラミング言語 R のランダムフォレストパッケージを用いてインフルエンザ新規感染者数の予測を試みた結果を報告する．

キーワード：データマイニング，ランダムフォレスト，インフルエンザ，ウイルス感染症，R

1. はじめに

データマイニング [1] は近年様々な分野で注目され，応用が試みられている．現在世界的に問題となっている新型コロナウイルス感染症（以後 COVID-19 と記す）に関しても，AI やデータサイエンスの立場から様々な調査が試みられている．しかし，COVID-19 は最初の感染例が報告されてから１年程度と経過時間が短いことから流行しやすい条件をデータから発見するのは困難が予想される．例えば季節変動の影響を調べるためには１０年程度の蓄積データが必要であるものと予想する．一方，近年のインフルエンザ [2] は特異な流行が少なく，感染者数が安定的かつある程度多くの件数がある [3] ため，データマイニングを用いた調査が可能なものと考えられる．そこで本稿では，予備的にインフルエンザの感染者数を調査した結果を報告する．

日本においては気象データも豊富に蓄積されており，国土交通省気象庁サイトから 1976 年以降の種々の気象データを得ることができる [4]．インフルエンザの流行は気象条件と相関があることが知られているため，本稿ではインフルエンザと気象条件の関係を調べる．今回は統計プログラミング言語 R[5] のランダムフォレストパッケージ randomForest を用いて，インフルエンザ新規感染者数に対する主な気象条件の重要度を調べ，さらに気象条件から新規感染者数の予測を試みる．

[1] 金沢工業大学
Yatsukaho 1-3, Hakusan, Ishikawa 921–8132, Japan
[a)] ryoko@neptune.kanazawa-it.ac.jp

2. インフルエンザの概要

インフルエンザ感染症（以後「インフルエンザ」と記す）はインフルエンザウイルスに感染して起こる気道感染症であり，年による変動はあるが，毎年決まって流行する感染症である [2]．インフルエンザウイルスには A，B，C の３つの型があって，特に大規模な流行が起こりやすいものは A 型である．そこで本稿では A 型を調査対象とする．

日本ではインフルエンザは定点報告対象であり，日本全国に 5500 箇所ある指定届出機関は週ごとに感染者数を保健所に届け出ることが感染症法で義務付けられている [2]．そのため，日本においてインフルエンザが発症して医療サービスを受けるとその多くが統計データとして集積されることが期待できる．

3. インフルエンザの感染者数予測

表 1 に今回使用したデータのデータ例とその数値範囲を示す．インフルエンザは冬季に流行するため，毎年 9 月から次年 6 月ごろ（その年の収束時期によって多少前後する）までを一つのシーズンの区切りとし，データが蓄積されている．今回は東京都の 2011 年 3 月から 2020 年 4 月（収束時期が 4 月であったため）までのデータを用いており，表 1 の数値はいずれも週ごとのデータである．表 1 で最初に示す新規感染者数はサイト [3] から得られるデータで，今回の目的変数である．数値範囲を見ると，新規感染者数は 0 人から 2.6 万人まで変化し，東京都だけでも多いときは週に 2 万人以上の新規感染者が発生することがわか

表 1　インフルエンザのデータ（東京都，2011 年 9 月から 2020 年 5 月まで．データ例は 2012 年 1 月 2 日 ∼1 月 8 日の週．）

Table 1　Influenza data used (Tokyo, from September 2011 to May 2020).

	データ例	数値範囲
新規感染者数 [人]	387	0 - 26635
平均気温 [℃]	5.9	2.4 - 31.7
最高気温 [℃]	11	6.5 - 37.3
最低気温 [℃]	1.9	-2.1 - 24.4
平均湿度 [%]	34	33 - 100
平均日照時間 [時間]	7.1	0 - 13.6

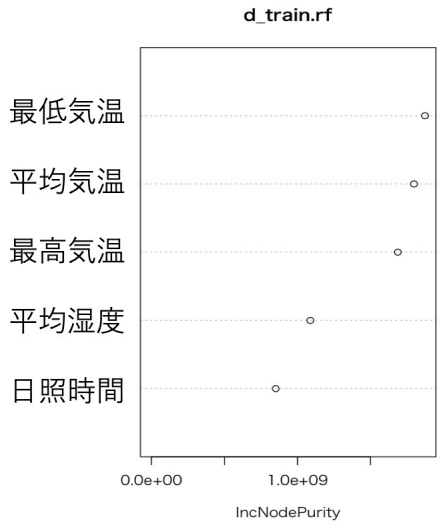

図 1　ランダムフォレストで得られる説明変数の重要度

Fig. 1　Importance of explanatory variables obtained in Random Forest.

る．表 1 に示す平均気温，最高気温，最低気温，平均湿度，平均日照時間を今回の説明変数とした．これらは全てサイト [4] から得られる．データを蓄積している期間には，真夏のような日もある 9 月や 6 月と，厳冬期である 1 月と 2 月を含むため，いずれの数値も東京の通年での最小値と最高値を含むような広い範囲となっている．

今回用意できたデータは全部で 338 件であったので，この中からテストデータを選び，残りを学習データとする．シーズン中での変動を予測したいので，2017 年 9 月から 2018 年 5 月のシーズンを選び，この期間中で各月 1∼2 件ずつとなるよう週を選んだところ 17 件となったのでこれをテストデータとし，残りの 321 件を学習に使用した．

図 1 にランダムフォレストで得られる説明変数の重要度を示す．図 1 では一番上に最低気温が表示されており，今回使用したデータでは最低気温の重要度が一番大きいことを示す．平均気温と最高気温が最低気温に次いで重要度が大きく，平均湿度と日照時間は 3 種類の気温データよりも重要度が低いことがわかる．「気温が低く，湿度が低いとインフルエンザが流行しやすい」とよく言われるが，図 1 では，湿度よりも最低気温との相関が大きいことを示す．

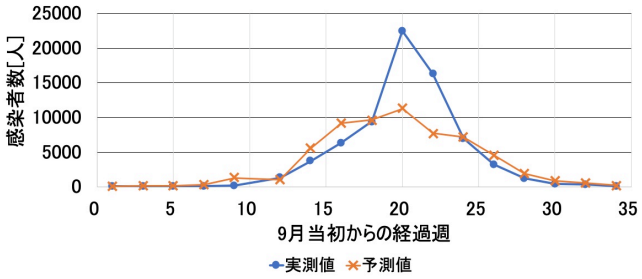

図 2　ランダムフォレストによる感染者数予測結果

Fig. 2　Prediction results of the number of infected people by random forest.

図 2 にランダムフォレストで予測した感染者数とその実測値を示す．図 2 の横軸はシーズン初めからの経過週を示し，縦軸は感染者数である．図 2 の実測値を見ると，シーズン初めはほとんど新規感染者がいないが，新規感染者は１１月に相当する第 9 週ごろから増え始め，１月末から２月初めに相当する第２０週ごろにピークとなり，その後は急激に減っていって 3 月末から 4 月初めに相当する第 30 週ごろにほとんど新規感染者がいなくなってシーズンが終わることがわかる．図 2 の予測値は感染者数を正確に予測することはできていないが，新規感染者の増え始める時期とピーク時期は実測値と同様である．

4.　おわりに

本稿では，インフルエンザ新規感染者数と気象条件の関係をランダムフォレストで調べた結果を報告した．その結果，気象条件では平均湿度よりも最低気温のほうが新規感染者数の予測で重要であることがわかった．さらに，シーズン中の新規感染者数の予測を試みた結果，新規感染者数の直接の予測は困難であったが，感染者の増え始める時期とピーク時期はおおむね予測できることがわかった．今後は，引き続きインフルエンザについてより詳しくデータマイニングを行っていくとともに，新型コロナウイルスについても調査を試みたい．

参考文献

[1] 平井有三：はじめてのパターン認識，森北出版株式会社 (2012).
[2] 国立感染症研究所：インフルエンザとは，入手先 ⟨https://www.niid.go.jp/niid/ja/kansennohanashi/219-about-flu.html⟩（参照 2021-01-27）.
[3] 厚生労働省：インフルエンザの発生状況，入手先 ⟨https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/kenkou/kekkaku-kansenshou01/houdou.html⟩（参照 2021-01-27）.
[4] 国土交通省気象庁：過去の気象データ・ダウンロード，入手先 ⟨http://www.data.jma.go.jp/gmd/risk/obsdl/⟩（参照 2021-01-27）.
[5] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 入手先 ⟨https://www.R-project.org/⟩.（参照 2021-01-27）.