2B-04

空間情報ビッグデータを用いた新型コロナ感染予測と要因推定

水野貴之†‡　上坂大輔‡　幡容子‡　南川敦宣‡#
国立情報学研究所†　株式会社 KDDI 総合研究所‡　KDDI 株式会社#

１．はじめに

2019 年末に中国の武漢で発生したコロナウイルス感染症は，またたく間に，世界中に広がり，現在（2021 年 1 月 3 日）までに，全世界で 8422 万人以上が感染し，約 183 万人の命が失われている．感染者数の予測には短期的な流行過程を決定論的に記述する SIR タイプのモデルが数多く使われている[1]．SIR タイプのモデルはシンプルであり，主に感染症に対して免疫を持たない者，感染症が潜伏期間中の者，発症者，感染症から回復し免疫を獲得した者についての，それぞれの人数のダイナミクスを，感染率や発症率，回復率などをパラメータにして記述する．モデルのパラメータは，人口や日々の新規感染者数などを用いて推計することができる．推計されたパラメータを用いて，モデルから１人の感染者が平均何人の感染者を生むかを表す実効再生算数を算出することができ，日々の実効再生算数をモニタリングすることで，感染症対策の政策を実行するタイミングを決めたり，政策の効果を測ったりすることを科学的にサポートできる．

SIR モデルは感染状況を知ることには役立つが，感染要因はブラックボックス的にパラメータの中に入り込んでしまうために，どのような政策を実行するべきかを科学的にサポートすることには使いにくい．日本政府はこれまでに，全国一律の外出の自粛要請，飲食店の夜間営業の規制などを実行してきた．これらは，全国的な感染者数の増加や，感染経路の聞き取りにもとづいた政策であった．しかし，このような人間の気付きに頼る方法では，対策が後手にまわったり，感染要因を見落としたりする可能性がある．従って，過去の感染状況から感染要因を推定することを科学的にサポートするシステムが必要である．このようなシステムを開発することにより，効率的な感染症対策が可能になり，経済活動へのダメージを小さくすることができる．

本研究では，感染要因の候補に，地域の人口と，人々の地域間の移動，飲食街などの地域の属性を考える．各地の感染状況を，これらの感染要因で予測する機械学習モデルを構築することで，感染リスクの高い要因を絞り込む．機械学習による予測は，因果ではなく相関を利用するため，予測に寄与する要因は代理変数である可能性もあるが，多くの要因から主要因を絞り込み，人間による要因推定を科学的にサポートすることができる．

以降の節では，第２節で，感染要因を推定するシステムを開発するために用いる空間情報ビッグデータについて説明する．第３節では，人々の地域間の移動から生活圏を抽出する手法について述べる．第４節では，Lasso と LightGBM を用いた感染要因の推定システムを導入し，日本の 2020 年の春，夏，冬に発生した感染流行の要因を推計して考察する．

２．空間情報ビッグデータ

2020 年 2 月 14 日から 12 月 27 日までの日本各地の日々の新規感染者数を，JX 通信社が提供す罹患施設情報を用いて把握する．このデータには，日本国内における一般企業，病院，店舗，公的機関によるインターネット上への公開済み情報にもとづく新型コロナウイルス罹患者の発生場所の住所，感染者数，初回感染者報告日などが収録されている．データには，経路不明や家庭内感染は含まれないため，収録されている感染者数は，政府が公表している感染者数の約２割程度となっている．

人々の移動については，利用許諾のある全国数百万人の匿名加工された au のスマートフォン位置情報データの 2020 年 4 月 24 日，7 月 3 日，11 月 13 日分を用いる．このデータは推定された居住地情報，及び移動滞在の識別が可能な情報が付与された位置情報を，個人の履歴が追えな

Exploring Factors of COVID-19 Infection using Location-Based Big Data
Takayuki Mizuno †‡, Daisuke Kamisaka ‡, Yoko Hata ‡, Atsunori Minamikawa ‡#
†National Institute of Informatics
‡KDDI Research, Inc.
#KDDI CORPORATION

い形に加工されたデータとなっている．

地域の人口（夜間人口）については，2015 年の国勢調査を用いて把握する．また，昼間人口については，前述のスマートフォンの 9:00 から 18:00 までの位置の空間分布から把握する．

地域の属性については，ゼンリンが提供する 2016 年の座標付き電話帳 DB テレポイント法人版を用いて，法人施設の空間分布を把握する．このデータでは，法人施設の住所とともに，各法人施設の業種番号が大分類で３９分類付与されている．個人の家屋や共同住宅といった，法人以外の施設の空間分布は，ゼンリンが提供する 2020 年の建物統計データを利用して把握する．このデータには，住宅や商業ビル，オフィスビルなど全国約 3,800 万棟の建物について，基準地域メッシュ（500m メッシュ）ごとに，建物種別などでの件数が集計・統計化して収録されており，この中から個人の家屋と共同住宅の件数を利用する．

３．人流ネットワークによる生活圏の抽出

利用許諾のある全国数百万人の匿名加工された au のスマートフォン位置情報データを用いて居住地と滞在先を結ぶ人流ネットワークを作成し，DeepWalk を用いて各地域の生活圏を表す特徴量ベクトルを算出する．

人々の１日の行動は，居住地から外出して居住地に戻ってくる．従って，生活圏を表す人流ネットワークをえるために，各ユーザーの居住地と１時間毎の滞在地をリンクで結ぶ，各ユーザーの居住地と滞在地のネットワークを，全てのユーザーで集計することで，居住地と滞在地を結ぶ移動者の人数で重み付けられた２部グラフが得られる．居住地と滞在先の２分類をなくし，同じ地域を１つのノードにまとめることで，２部グラフを，地域間の移動を表す１部グラフ（人流ネットワーク）に変換する．このネットワークの各ノードは地域を表し，各地域からどの地域に流入出がおこなわれているかが，ネットワークで表現される．

DeepWalk を用いて，人流ネットワークにおける各地域（ノード）の L リンク先までの隣接地域（隣接ノード）の情報を N 次元まで集約する．人流ネットワークの各地域を日本の市区町村単位で設定し，各地域の $L=3$ リンク先までの隣接地域の情報を $N=32$ 次元まで集約する．各地域の 32 次元ベクトルを，k-means の逐次繰り返しと BIC による分割停止基準によりクラスター数 k を自動的に決定する x-means を用いてクラスタリングすると，各クラスターは人々の直感と合う各地の経済圏・生活圏を示すことが分かる．

４．感染要因の推定と考察

全国 1741 市区町村について，感染者数や業種分類別の物件数，人口，年齢別や性別の人口を数える．感染者数と人口を除き，各件数は人口１人あたりに規格化する．３節で作成した各市区町村の人流ベクトルを用いる．これらを感染要因の候補として，1741 市区町村の感染者数を Lasso 及び LightGBM で予測する．ハイパーパラメータは，Lasso は 0.05 に，LightGBM は Python Package の LightGBM[2]のデフォルト値を用いた．

第一波（2020 年 1 月-6 月），第二波（7 月-9 月），第三波（10 月-12 月）について，時期にそれほど依存せず，感染者数の実数と予測値の相関は，Lasso が 0.78，LightGBM が 0.65 であり，高い予測性能がえられることが分かった．

感染要因の候補についての Lasso の係数から，感染者数の予測に最も寄与する要因は人口であり，人が多い地域は感染者が多くなる．この人口の要因は，第一波→第二波→第三波で，0.32→0.41→0.49 と強くなっており，時間とともに感染が全国に拡大していることが分かる．また，18 歳未満の人口割合が高い地域ほど相対的に感染者は少なく，子供は感染しにくいことが分かる．一方で外国人の人口割合が高い地域は相対的に感染者が多く，対策が必要である．第二波では，娯楽施設と飲食店数が感染者数に強く正に働いていた．つまり，これらの施設への対策が有効であった．第三波でも，これらの施設の係数は引き続き正である．また，大型総合店舗の係数も感染者数に正に働いており，大型総合店舗における対策を強化すべきであると考えられる．関東，沖縄，北海道は，人口と地域属性を考慮しても感染確率が高い．全国一律ではなく，これらの地域を重点的に対策する必要があると言える．

本発表では，地域の解像度を高め，業種分類を増やした結果を合わせて示し，感染要因を推定する本システムの有効性について議論する．

謝辞

本研究の一部は、東京大学 CSIS 共同研究（No. 674）による成果である（利用データ: 座標付き電話帳 DB テレポイント 法人版（ゼンリン提供））

参考文献

[1] J. Li, et. al., (2020) Open Forum Infectious Diseases, ofaa442.

[2] https://github.com/microsoft/LightGBM (2021 年 1 月 3 日アクセス)