The 35th Annual Conference of the Japanese Society for Artificial Intelligence, 2021

携帯電話人口統計データと新規陽性者数の相関に着目したCOVID-19の感染リスク地区の抽出

Extracting areas potentially spreading COVID-19 by focusing on correlation between mobile phone population statistics and the number of new positive cases

石田 展雅 [*1] Nobumasa Ishida　豊田 正史 [*2] Masashi Toyoda　梅本 和俊 [*2] Kazutoshi Umemoto　商 海川 [*2*3] Haichuan Shang　是津 耕司 [*3] Koji Zettsu

[*1]東京大学 The University of Tokyo

[*2]東京大学生産技術研究所 Institute of Industrial Science, The University of Tokyo

[*3]情報通信研究機構 National Institute of Information and Communications Technology

We propose a method to extract areas potentially spreading COVID-19 by focusing on correlation between mobile phone population statistics and the number of new positive cases. Our experiment showed that our method can successfully extract areas that are consistent with the government's views on infection sources.

1. はじめに

2020 年にパンデミックとして認められた新型コロナウイルス感染症（COVID-19）は世界中で感染が拡大しており，その抑制のために世界各国で市民の社会活動を制限する政策が施行された．日本でも 2021 年 2 月までに二度の緊急事態宣言が発され，対象となる都市全域で外出の自粛や飲食店の休業が要請されている．これらの介入政策は確かに COVID-19 の流行を抑えるものの [Dehning 20]，同時に多くの経済的な損失を生じており，国内でも多くの事業者が同期間の大幅な減益を報告している [*4]．そのため経済活動と感染拡大抑制を継続的に行っていくために，対象を限定した重点的な介入政策が求められている [Chang 20]．

局所的な介入政策を実施するための課題の 1 つとして，集中的に感染を生んでいる感染源を特定することの難しさが挙げられる．COVID-19 は感染しても無症状や軽症である場合が多く，実際の感染者数と陽性者数の間には乖離がある [Li 20, Perkins 20]．それゆえ陽性者の行動履歴に基づいて感染経路を調査しても感染経路が不明なケースが多い [*5]．また行動履歴が自己申告であることによる不確かさや発症件数の増加に伴い人手による追跡が困難になっていることも感染源の特定を妨げている．モバイル端末を活かして陽性者との接触を通知するシステムも導入されているものの，普及率が低く感染の追跡精度に課題がある [Munzert 21]．これらの事情から，感染源となっている可能性がある場所，すなわち**感染リスク地区**を効率的にスクリーニングする方法が求められている．

本研究では，メッシュ状の携帯電話人口統計データと市区町村の陽性者数の時系列を用いて，人口変動と陽性者数の変動との相関が高い地区及び時間帯を抽出する手法を提案する．公表されている日ごとの陽性者数の中で最も対象地域が狭いものは市区町村単位のデータである．そこで市区町村という空間的に粗いデータからメッシュ単位の推定を行うために，「感染リスク地区における人口の増減がその地区を含む市区町村の新規陽性者数の増減に反映されること」，及び「人口変動と感染症の実効再生産数の変動の関係が概ね線形であること」を仮定する．この仮定の下で人口と実効再生産数の相関が大きいメッシュを感染リスク地区とみなす．そして非線形な摂動を考慮して，直接的に相互相関関数を求める代わりに，望ましい性質を満たした距離を計算する．具体的にはロバストな量として時系列の極値に注目し，それらの対応付けを優先した動的時間伸縮法を用いる．

実験では 2020 年 7 月の東京都について提案手法を適用した．その結果，渋谷センター街などいくつかの飲食店街の夜の人口の時系列と渋谷区の実効再生産数の間に相関があることが見出され，政府による見解 [*6] と矛盾しない地区が抽出された．この結果は，提案手法によって得られた知見が専門家による人手調査の一助となる可能性を示唆している．

2. 関連研究

位置情報データを用いて COVID-19 の感染拡大を分析している既存研究は，感染症の流行過程をモデル化しシミュレートすることでデータを説明する方針と，データから統計的な手法で相関関係や因果関係を見出す方針に大別される．

前者の一例として，Chang らは施設単位の人口統計データを利用した感染症モデルによって superspreader と呼ばれる少数の施設が感染の大部分を占めていることを報告している [Chang 20]．彼らの手法は感染症流行のモデルである SEIR モデルを基にしており，行政区分ごとに人口を免疫を持たない者，潜伏期間中の者，発症者，回復した者の 4 つに区分けする．そして居住地と個々の施設を結ぶ二部グラフを構成し，様々な施設での接触を通じた感染拡大の時間発展をシミュレートする．ここで各施設での感染率は各時刻の人口統計データから得られる滞在時間及び人口密度によって与える．米国の 2 か月間のデータを用いた実験の結果，10 個の都市圏で陽性者数の時系列を回帰できることが確認され，また，施設の種別ごとの分析でフルサービスレストランにおいて最も多くの感染が発生したと報告されている．しかしながら Chang らの方法ではモデルの設計上，滞在時間と人口密度が大きい施設が必ず高いリスクを持つように評価されてしまうため，実際の感染状況とは解離している施設が存在する可能性がある．

連絡先: [*1*2]{ishida-n,toyoda,umemoto,shang}@tkl.iis.u-tokyo.ac.jp,[*3]zettsu@nict.go.jp

*4 https://www3.nhk.or.jp/news/html/20200730/k10012541151000.html

*5 https://www.bousai.metro.tokyo.lg.jp/_res/projects/default_project/_page_/001/012/788/2021011402.pdf

*6 https://www.cas.go.jp/jp/seisaku/ful/bunkakai/corona19.pdf

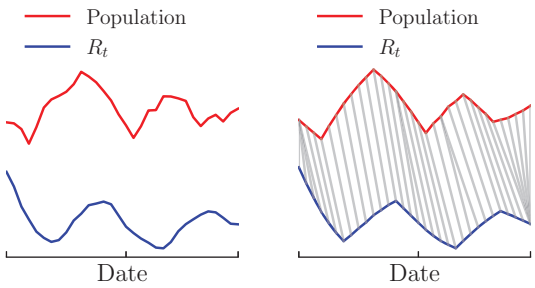


図 1: 一つのメッシュに対する提案手法の処理の流れ．(a) 処理前の人口及び実効再生産数 R_t．(b) 処理後の人口及び R_t．灰色の線は対応付けられた頂点同士を結んでいる．「混雑統計 ®」©ZENRIN DataCom CO., LTD.

別の例として，Loo らは集団感染が生じやすい場所の指標として，施設の集積度と訪問者の移動圏の広さを組み合わせた量を提案している [Loo 21]．実験でリスクが高いと評価された地区は COVID-19 の大規模な集団感染が生じた場所を含んでおり，指標としての妥当性が示唆されている．しかしながら Loo らの方法では指標の計算に実際の COVID-19 の感染状況を考慮していないため，現在感染が拡大している可能性がある場所を絞り込む用途には使えない．

後者の例として，Badr らは米国の各郡について人々の移動パターンと新規陽性者数の増加率に相関があることを見出した [Badr 20]．COVID-19 流行の初期に当たる 2020 年 1 月 1 日から 2020 年 4 月 20 日のデータを分析することで，ロックダウンにより人々の行動が制限された結果，11 日の遅延を持って新規陽性者数が減り始めたと報告している．この遅延は，感染後に発症し診断を受けて検査の結果が公表されるまでの平均的な日数と一致している．しかしながら Badr らの分析は都市のロックダウンという非常に大きな行動変容に関する分析であり，より小さなスケールでの移動パターンの変動に対して同手法をそのまま適用することは難しい．

本研究では位置情報データを用いて感染リスク地区を絞り込む問題に取り組む．Chang らが示唆した施設種別による感染発生数の偏りを，地区による偏りという異なる観点から評価する．また，Badr らが示した人々の移動パターンと感染症流行の相関を，ロックダウンよりも小規模な人口変動について検証できるようにする．

3. 手法：感染リスク地区の抽出

提案手法では市区町村単位の COVID-19 新規陽性者数の時系列と，同期間・同地域内の各メッシュにおける 1 時間単位の人口統計データを用いる．提案手法は 3 つのステップからなる．第 1 ステップでは，人口と近似的に線形な関係となる量として，新規陽性者数から実効再生産数 R_t を計算する．あるメッシュのある時間帯における人口と対応する市区町村の R_t の例を図 1a に示す．第 2 ステップでは，人口と R_t が線形な関係にあるか否かを定量化するために時系列間の距離を測る．このときノイズの影響を低減するために，それぞれの時系列の極値を保存するような平滑化を施す．その後，時系列の各点について極値同士の対応付けを優先したマッチングを行い，対応付けられた点間の距離の和として全体的な距離を求める．この処理を施した場合の例を図 1b に示す．第 3 ステップでは，メッシュ間で距離を比較することにより，人口と R_t の相関が大きいメッシュを感染リスク地区として抽出する．以降では第 1 ステップと第 2 ステップの詳細を述べる．

3.1 実効再生産数の計算

感染症の拡大を特徴付ける量として，各時点 t での実効再生産数 R_t を計算する．R_t は時点 t での疫学的な条件が維持された場合に一人の感染者が生む二次感染の件数の期待値として定義される [Thompson 19]．R_t は免疫を持っていない者と感染者の接触回数に比例するため [Lipsitch 03, Chang 20]，多くの感染が起こっている地区ではその地区の人口密度と線形な関係にあるとみなせる．さらに本研究で用いる人口統計データでは各メッシュの面積が固定されているため，人口密度の代わりに人口との相関に置き換えられる．

感染を直接観測することはできないため，R_t を求めるには統計的な推論が必要である．そこで提案手法では，感染症の発症日のデータからベイズ統計によって R_t を計算する Cori らの方法 [Cori 13, Thompson 19] を用いる．この手法では COVID-19 を特徴付ける量として，感染する個体と感染させる個体の発症日の間隔である serial interval の分布に関する情報が必要であるため，Nishiura らによって報告されている平均 4.8 日，標準偏差 2.3 日というパラメータを用いる [Nishiura 20]．

本研究で利用する陽性者数のデータは，保健所への報告日に基づいて集計した報告日ベースのものとなっている．そこで検査数の曜日による違いの影響を取り除くために前後 3 日間の移動平均を取り，さらに発症から報告までの平均的な遅延として 7 日 [Li 20, Chang 20] を減じることで近似的に発症日ベースのデータを作成する．また，Cori らの手法によって得られる R_t は潜伏期間の分だけ真の値よりも遅延している．そこで計算された値を COVID-19 の潜伏期間として 4 日 [Li 20, Chang 20] 早めることで最終的な R_t とする．

3.2 感染リスク地区としての尤もらしさの計算

COVID-19 の 2 次感染の件数は分散が大きく，感染の大部分が一部の施設における superspreading event という集団感染によって生じていると考えられている [Endo 20, Chang 20]．また日本国内においても集団感染の発生場所が一部の施設種別に偏っていることが指摘されている [*6]．このことから，感染リスク地区があるとすればそのメッシュだけで地域の感染をおおよそ説明できる可能性がある．そこで提案手法では個々のメッシュの人口変動と R_t の相関を評価することで，市区町村単位という空間的に粗いデータから詳細なメッシュ単位での感染リスク地区を抽出する．

人口変動と R_t の相関の計算方法として，時系列の相互相関関数を用いた方法が考えられる．しかしながらデータから計算した R_t には非線形な摂動が含まれているためこの方法はうまく働かない．摂動の具体的な要因は次の 2 点に大別される．

1. 感染してから陽性と診断され報告に至るまでの期間がばらつくため，人口と R_t の間の遅延が動的に変化する．
2. R_t の変動には人口で捉えられない要因が含まれており，R_t にノイズとして現れる．特に気温や湿度といった気候の影響は大きいとされている [Wang 20]．

以上の問題に対処するために，提案手法では人口と R_t の相関に代えて，これらの時系列間の距離を以下の 2 ステップで計算する．1 つ目のステップはノイズに強い特徴点として極値を残しながらノイズを除去する過程であり，2 つ目のステップは時系列間で動的な遅延を許しながら極値同士を対応させて時

系列間の距離を測る過程である．また分析に用いるデータの期間を短く設定することで上述の気候の影響は無視できるものとする．こうして求めた距離が小さいほど，人口によって R_t が説明できていることになり，当該メッシュで多くの感染が生じた可能性が高いと考えられる．以降で各ステップの詳細を説明する．なおそれぞれのステップの前に時系列の平均を 0，分散を 1 に規格化することでシフトやスケールの影響を除去する．

1 つ目のステップでは画像処理におけるエッジ保存平滑化方法として知られるバイラテラルフィルタ [Tomasi 98] と細線化の方法を用いる．バイラテラルフィルタとは畳み込みの重みが位置だけでなく対象の値に依存するフィルタであり，値が大きく変わる箇所が平滑化されずに維持される方法である．また近くにある極値をまとめるために，細線化の方法を流用する．各 i について，前後 l 点の中でその点が極大値または極小値であれば元の値を維持し，そうでない場合は隣接 2 点の平均値で置き換える．以上の 2 つの処理を k 回交互に繰り返すことで，時系列の極値を保存した平滑化が為される．

続いて 2 つ目のステップとして，動的時間伸縮法（Dynamic Time Warping; DTW）による時系列間のマッチングを行う．特に本研究では極値同士の対応付けを優先するために，DTW を形状が考慮されるように発展させた shape Dynamic Time Warping（shapeDTW）[Zhao 18] を用いる．この手法では，まず対象となる時系列について各点を特徴量に置き換えた多変量時系列に変換し，続いて通常の DTW により時系列間の距離を求める．ここでの特徴量とは各点の周囲の形状を反映した量であり，本研究では各点を中心とした 7 点における微分値の系列を用いる．

以上のステップにより，R_t に対する各メッシュの人口の時系列の距離を計算し，他のメッシュに比べて距離が小さいメッシュを感染リスク地区として尤もらしいとみなす．

4. 実験

区単位の新規陽性者数のデータに提案手法を適用しメッシュ及び時間帯ごとに感染リスク地区としての尤もらしさを求めた．

4.1 データセット

COVID-19 に関するデータセットとして，東京都の各区における報告日ベースの累積陽性者数を東京都のウェブサイト [*7] から取得し，日ごとの新規陽性者数の時系列を用意した．

また人口統計データとして，携帯電話の位置情報を集約した「混雑統計 ®」データ [*8] を使用した．このデータでは都市が 250 m 四方のメッシュに分割され，それぞれの領域について 1 時間ごとの滞在人数の推定値が利用できる．

4.2 実験設定

実験の対象は東京都 23 区とし，期間は 2020 年 7 月 1 日から 2020 年 7 月 31 日の 1 か月とした．これは日本の非常事態宣言が 5 月 25 日に解除 [*9] されてからおよそ 1 か月が経過し東京都の各区で陽性者が増加していた時期である．

各メッシュの人口には曜日による変動があるため，提案手法を適用する前に前後 3 日間の移動平均を適用した．また人口が少ない地区を候補から除外するために，期間中の最大人口が 2000 人を超えているメッシュのみを分析した．

提案手法中の距離計算におけるパラメータは $k = 5$ 及び $l = 2$ とした．

4.3 感染リスク地区の抽出結果

図 2 に渋谷区における結果を示す．図中のカラーバーは距離に対応しており，赤い方が距離が小さい．図 2a と図 2b はそれぞれ 12 時と 21 時の人口に適用した結果に対応する．これらの図を比べると，渋谷駅を中心とした繁華街や初台駅の北側，代々木駅周辺など飲食店が多く存在する地区において，距離が 21 時に小さくなっていることがわかる．このことはそれらの地区の夜の人口の時系列が，同時期における渋谷区の R_t の変動と類似していることを示しており，感染の多くが夜の飲食店で起こっていたという日本政府の見解 [*6] と矛盾しない．

時間帯による距離の変化を検証するために，渋谷センター街のメッシュにおける時系列を図 3 に示す．図 3a 及び図 3b はそれぞれ 12 時人口及び 21 時人口の場合に対応している．図より，21 時人口は 7 月 1 日から 7 月 16 日における R_t の増減に沿っている一方で，12 時人口は R_t と動きが乖離していることがわかる．

5. おわりに

本研究では市区町村単位の COVID-19 の新規陽性者数データから 250 m 四方メッシュ単位で感染リスク地区を絞り込む問題に取り組んだ．感染が一部の地区で集中的に生じていること，及びその地区の人口が実効再生産数と線形な関係にあることを仮定し，携帯電話人口統計データを用いて実際に渋谷区内の各地区の評価を行なった．

提案手法に関して注意すべき事項を 2 点述べる．第 1 に，実効再生産数の計算は統計的な不確かさが大きい．これは感染から発症までの期間や発症から検査結果が出るまでの期間が個人によって大きくばらつくためである．しかしながら，実効再生産数の代わりに文献 [Badr 20] で定義されている量を用いても波形の極値のパターンはおおよそ維持されることを確認している．第 2 に，本研究で感染リスクの可能性の評価に用いているのは因果関係ではなく相関関係である．それゆえ仮に尤もらしいと判断されたメッシュであっても，実際には感染が生じていない可能性がある．しかしながら，我々の手法によって感染リスク地区を絞り込むことで人手による調査・検証の負荷を低減することができると考えられる．

今後の課題として，人々が通勤などで区を超えて移動する影響を取り入れるための，区をまたいだ計算が挙げられる．また，距離の計算にも課題がある．現在の手法では shapeDTW において時間的に離れた点との対応も許容しているが，時間的に近い点同士を対応付ける方が尤もらしいと考えられるため，時間的なずれに応じてペナルティを加えることが考えられる．加えて，現在の手法では前処理として人口の平均を 0，分散を 1 に規格化しているため，人口の変動の規模の情報が失われている．感染症はより人が多い地点で広がりやすいため，人口の規模を考慮することが感染リスク地区の評価に有用であると思われる．

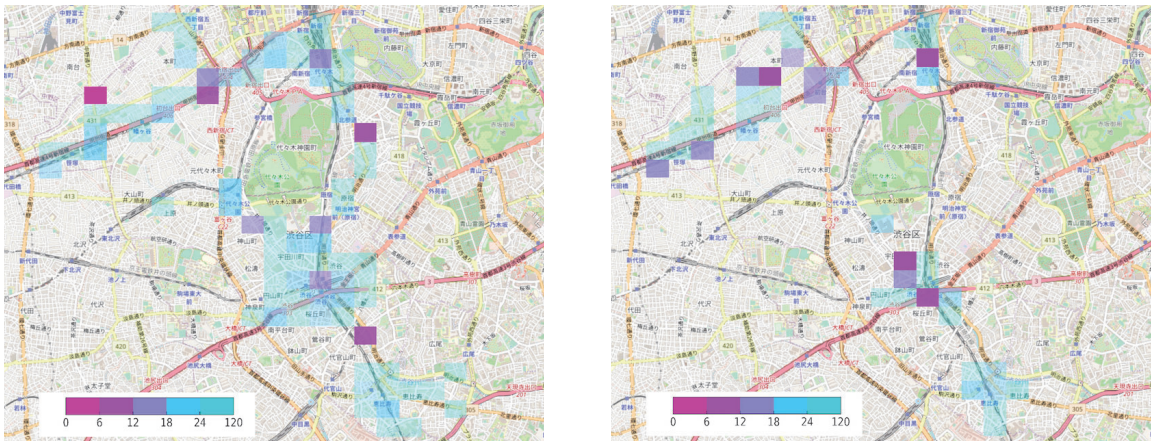
謝辞

本研究は，JST, CREST, JPMJCR19A4 の支援を受けたものである．

*7 https://stopcovid19.metro.tokyo.lg.jp/cards/number-of-confirmed-cases-by-municipalities/

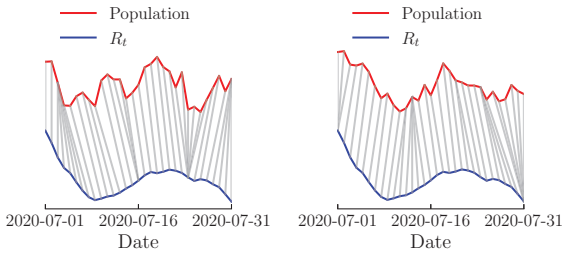
*8 「混雑統計 ®」データは，NTT ドコモが提供するアプリケーション（※）の利用者より，許諾を得た上で送信される携帯電話の位置情報を，NTT ドコモが総体的かつ統計的に加工を行ったデータ．位置情報は最短 5 分ごとに測位される GPS データ（緯度経度情報）であり，個人を特定する情報は含まれない．またデータの加工には「非特定化」「集計処理」「秘匿処理」がなされており個人が特定されることはない．※ドコモ地図ナビサービス（地図アプリ・ご当地ガイド）等の一部のアプリ．

*9 https://corona.go.jp/emergency/



(a) (b)

図 2: 渋谷区における感染リスク地区．赤い地区ほど距離が小さく，R_t との相関が大きい．(a)12 時，(b)21 時．期間は 2020 年 7 月 1 日から 2020 年 7 月 31 日．「混雑統計 ®」©ZENRIN DataCom CO., LTD.



(a) (b)

図 3: 渋谷センター街のメッシュにおける (a)12 時人口と R_t，及び (b)21 時人口と R_t．提案手法に従って処理した後，頂点の対応関係を処理前の時系列に反映させ，灰色の線で表している．「混雑統計 ®」©ZENRIN DataCom CO., LTD.

参考文献

[Badr 20] Badr, H. S. et al.: Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study, *Lancet Infect. Dis.*, Vol. 20, pp. 1247–1254 (2020)

[Chang 20] Chang, S. et al.: Mobility network models of COVID-19 explain inequities and inform reopening, *Nature*, Vol. 589, pp. 82–86 (2020)

[Cori 13] Cori, A. et al.: A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics, *Am. J. Epidemiol.*, Vol. 178, pp. 1505–1512 (2013)

[Dehning 20] Dehning, J. et al.: Inferring change points in the COVID-19 spreading reveals the effectiveness of interventions, *Science*, Vol. 369, p. eabb9789 (2020)

[Endo 20] Endo, A. et al.: Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China, *Wellcome Open Research*, Vol. 5, p. 67 (2020)

[Li 20] Li, R. et al.: Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2), *Science*, Vol. 368, pp. 489–493 (2020)

[Lipsitch 03] Lipsitch, M. et al.: Transmission dynamics and control of severe acute respiratory syndrome, *Science*, Vol. 300, pp. 1966–1970 (2003)

[Loo 21] Loo, B. P. Y. et al.: Identification of superspreading environment under COVID-19 through human mobility data, *Scientific Reports*, Vol. 11, p. 4699 (2021)

[Munzert 21] Munzert, S. et al.: Tracking and promoting the usage of a COVID-19 contact tracing app, *Nat. Hum. Behav.*, Vol. 5, pp. 247–255 (2021)

[Nishiura 20] Nishiura, H. et al.: Serial interval of novel coronavirus (COVID-19) infections, *Int. J. Infect. Dis.*, Vol. 93, pp. 284–286 (2020)

[Perkins 20] Perkins, T. A. et al.: Estimating unobserved SARS-CoV-2 infections in the United States, *Proc. Natl. Acad. Sci.*, Vol. 117, pp. 22597–22602 (2020)

[Thompson 19] Thompson, R. N. et al.: Improved inference of time-varying reproduction numbers during infectious disease outbreaks, *Epidemics*, Vol. 29, p. 100356 (2019)

[Tomasi 98] Tomasi, C. and Manduchi, R.: Bilateral filtering for gray and color images, in *Proceedings of IEEE Int. Conf. Comput. Vis.*, pp. 839–846, IEEE (1998)

[Wang 20] Wang, J. et al.: High temperature and high humidity reduce the transmission of COVID-19, SSRN 3551767 [Preprint] (2020)

[Zhao 18] Zhao, J. and Itti, L.: shapeDTW: Shape Dynamic Time Warping, *Pattern Recognit.*, Vol. 74, pp. 171–184 (2018)