

コロナ禍における迅速サーベイランスを用いたビッグデータ解析 における機械学習・統計学的方法論の整備

聖路加国際大学 公衆衛生学研究科
准教授 米岡 大輔

(共同研究者)

慶應義塾大学医学部	准教授	野村 周平
聖路加国際大学 公衆衛生大学院	教授	スチュアート ギルモア
国立がん研究センター	研究員	斎藤 英子
早稲田大学	助教	田上 悠太
東京工業大学 情報理工学院	助教	川島 孝行
東京大学大学院 医学系研究科	助教	史 蕭逸
千葉大学 予防医学センター	助教	江口 哲史

はじめに

ここ数十年、グローバル化、災害、テロだけでなく、SARS、MERS、豚インフルエンザ、COVID-19などの新興感染症や、一般的感染症動向に関する公衆衛生問題に由来して「迅速サーベイランスシステム」の需要が高まっています。迅速サーベイランスシステムとは、現状をいち早く把握するためにインターネットや SNS を用いてビッグデータをリアルタイムに収集し、結果の社会還元までを行う機構の総称です。こういったシステムにより迅速な調査を行い、発生地域を迅速に絞り込むことが重要であり、それにより疾病の蔓延を抑制するための適時適切な介入方法を選択することができます。こういった有用性にも関わらず、これまで迅速サーベイランスシステムを用いた統計手法開発は非常に限定的なものでした。例えば、時系列データに対する自己回帰によるトレンドのモデル化、スキャン統計によるホットスポットの検出、点過程によるブレイクポイントの検出などです。

新興感染症のパンデミックをサーベイランスによって早期に発見するためには、通常の下況下（すなわち、パンデミックのない反実仮想状況下）で予想される死亡率と比較して急激な上昇を推定し、リアルタイムに監視することが一般的です。COVID-19パンデミックに関しては、多くの重篤な感染国においてCOVID-19の死亡負担が過小評価されていることを定量化するために、いくつかの超過死亡推定法が提案されています。例えば、FluMOMOとFarringtonという二つの有名なアルゴリズムがあり、これらは実際に、米国疾病対策予防センター（CDC）と世界保健機構が支援する欧州死亡率モニタリング活動（EuroMOMO）において正式に用いられています。FluMOMOアルゴリズムは、季節性とインフルエンザ活動に関するいくつかの非線形項を持つ単純かつロバストな準ポアソン回帰モデルに基づい

ており、特に欧州各国で広く使用されています。一方、Farringtonアルゴリズムは、CDCが超過死亡の推定に使用しており、毎週推定値を更新し、米国におけるCOVID-19対策の意思決定プロセスをサポートしています。Noufaily *et al.* (2013) ⁽¹⁾ らはFarringtonアルゴリズムを拡張し、ロバスト残差を導入して特異性を向上させるために大規模なシミュレーション実験を行っており、その一般化が本研究の目的の一つでもあります。また、同様のFarringtonアルゴリズムを用いて日本の超過死亡の現状を把握することも目的となっています。

現在、大規模なアウトブレイク検知の迅速サーベイランスシステムが抱える大きな課題の1つは、長期間のデータが利用できないケースがあることです。そのため、時点数が不足し推定・予測精度が悪くなり、結果としてアウトブレイク検出性能が悪くなるといった統計的欠点が指摘されています。この問題を解決するために、本研究では新しい統計アルゴリズム—地理的重み付き一般化Farrington (GWGF) アルゴリズム—を提案し、空間的に相関のあるプロセスからサンプリングした時系列カウントデータを解析して超過死亡の推定を行うことを提案しました。一般に、地理的近傍情報を用いることで、推定のバイアスを犠牲にすることで推定効率を向上させ、安定化を図ることができます。この性質を利用し、本研究では、時点数が少ない場合でも安定的にアウトブレイクを検出できるよう設計された手法を提案しています。この手法は、特に、最小限のデータしか蓄積されていない新しい疾患の調査や、過去のデータが蓄積されていないような監視体制が整備されていない地域でも有効です。

結 果

本研究では、地理的重み付きポアソン回帰を提案したNakaya *et al.* (2005) ⁽¹⁾と同様に、空間カーネルを用いた地理的重み付き回帰手法と、準ポアソン回帰モデルを含む準尤度アプローチを利用した超過死亡のための一般化Farringtonアルゴリズムを提案しました。これは、観測地点間の距離に対応する地理的重みを各観測地点ごとに割り当てることで、局所的な擬似尤度モデルを構築し、Farringtonアルゴリズムの安定化を図ったものです。また、同時に従来のFarringtonアルゴリズムでは、超過死亡を説明するための共変量をモデルに組み込む事ができませんでしたが、本研究では、Zhang *et al.* (2006) ⁽²⁾と同様の二段階推定方式を取ることで、空間に対して可変/不変な両方の共変量を同時に入れることに成功しています。詳細な数学的記述はYoneoka *et al.* (2021) ⁽³⁾を参照してください。

図1は2020年1-9月の県別死亡数(左)と我々のGWGFを用いて推定した超過死亡数(右)を表しています。これによると、特に東京、大阪、愛知などの大都市圏での超過死亡が比較的多かった可能性があることが示されています。ただし、全体の死亡数と比較すると0.2%程度であることにも注意が必要です。また、図2では、静岡県の時系列のデータを示しています。静岡県を選択した理由は、最もGWGFと従来のFarringtonアルゴリズムの差が明確に出るからです。これによると、GWGFの方が多くのアウトブレイクを検知していることがわかります。

図1 2020年1-9月の県別死亡数(左)と推定された超過死亡数(右)

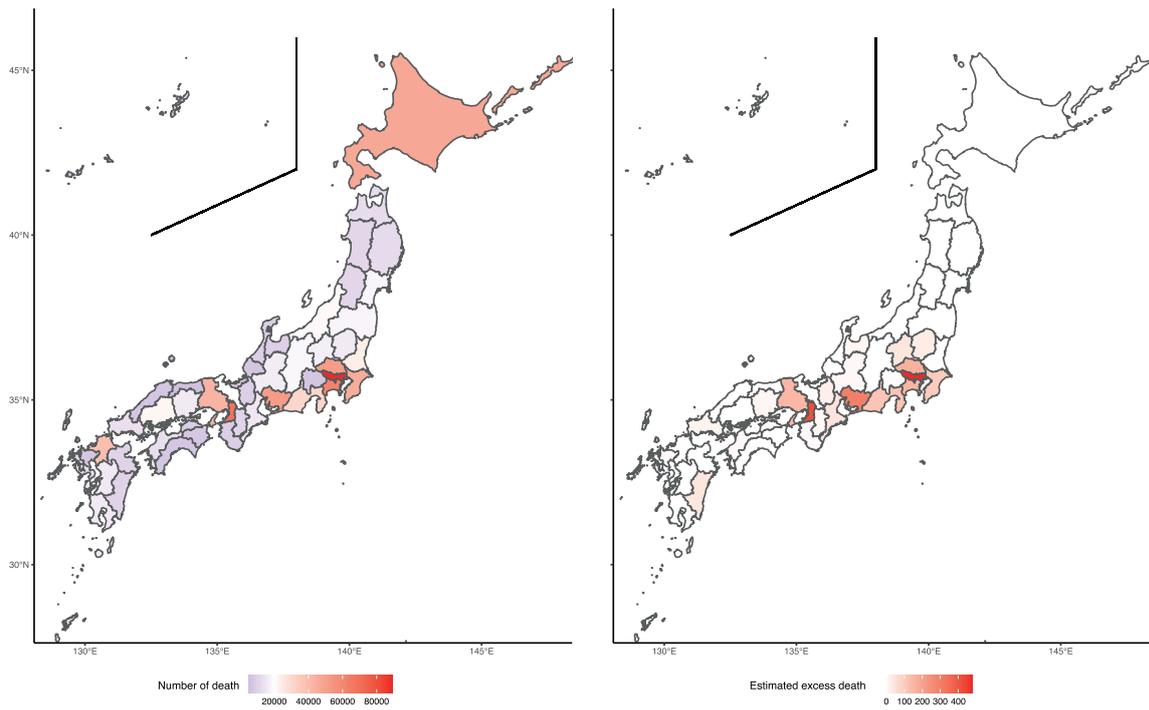
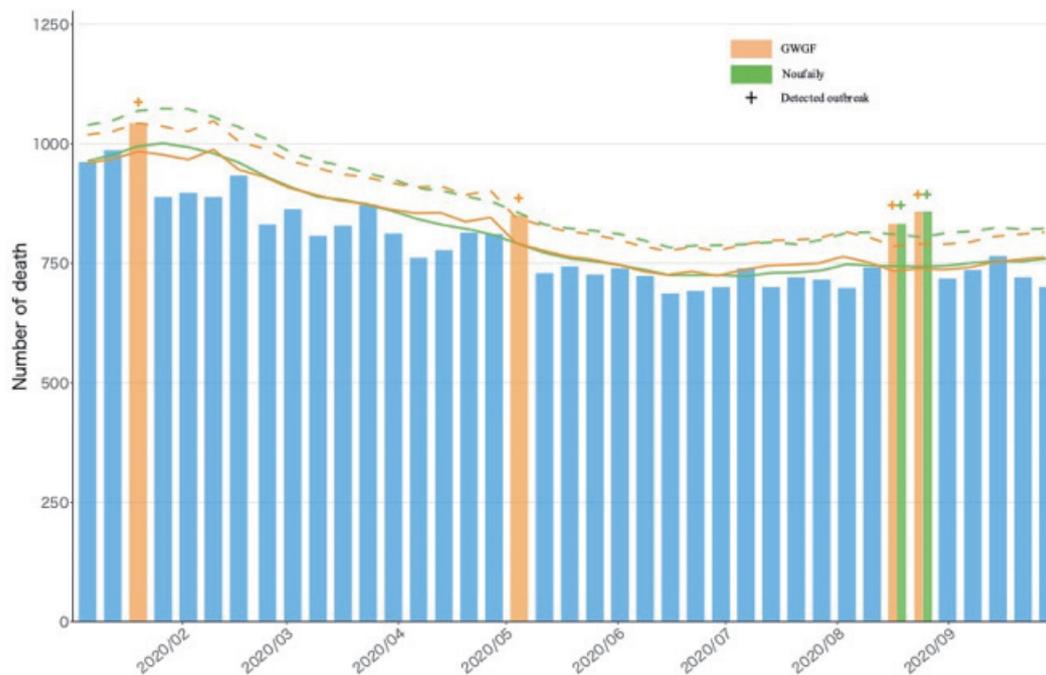


図2 GFGWと従来のFarringtonアルゴリズムを用いた超過死亡の時系列解析



考察

COVID-19に代表される新興感染症の迅速なサーベイランス体制が注目される中、より迅速かつ高精度なアウトブレイク検知のための予測手法の要求が高まっています。特に、利

用可能なデータセットの時点数がそれほど多くない場合、つまり、データが長期間蓄積されていない場合、回帰パラメータ推定値の標準誤差が大きくなるため、検出精度が制限される可能性があります。本研究では、準ポアソン回帰に地理的情報を組み込んで即時にアウトブレイクを検出する新しい手法、GWGFアルゴリズム、による予測精度向上を実証しています。この改善は、短期間のデータを使用した場合、より顕著に現れます。しかし、特異度は比較的低いことが予想されます。そのため、状況に応じてアルゴリズムを使い分けることが必要であると言えます。また、本研究の趣旨は新しい超過死亡のアルゴリズムを提案することであり、なぜ超過死亡が発生したのかなどに対する答えを提供するものではありません。また、本超過死亡の推定結果は、研究責任者も参画しているチームより、従来のFarringtonアルゴリズムを用いて毎月ダッシュボード (<https://exdeaths-japan.org/>) において報告している結果と大きい差はないことに留意し、最新版は以下のホームページより参照可能であることも最後に報告致します (<https://www.niid.go.jp/niid/ja/from-idsc/493-guidelines/10866-excess-mortality-all-211223.html>)。

要 約

近年、迅速なサーベイランスと地域的な流行の早期発見に対する需要が高まっています。COVID-19のような新興感染症の蔓延を抑制するために、迅速なサーベイランスは適時適切な介入を選択する一助となります。Farringtonアルゴリズムは、超過死亡の推定によく使用されていますが、大きな課題の1つは、学習に必要な過去の情報が不足していることで実際はよくある点が挙げられます。十分な学習データがない場合、このアルゴリズムの推定・予測精度は低下し、アウトブレイク検出がうまくいかなくなる可能性があります。この問題に対処すべく、本研究では新しい統計アルゴリズム、地理的重み付き一般化Farrington (GWGF)アルゴリズムを提案しました。このアルゴリズムでは、地理的に可変/不変な共変量、および地理的情報を組み込んで、空間的に相関するプロセスからサンプリングした時系列カウントデータを解析し、超過死亡の推定を行うことを可能にしています。また、同アルゴリズムを用いてシミュレーション及び、日本の2020年1-9月期の超過死亡も推定しました。

文 献

1. Noufaily, Angela, *et al.*: An improved algorithm for outbreak detection in multiple surveillance systems. *Statistics in medicine*, 1206-1222: 2013.
2. Zhang, Wenyang, *et al.*: Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis*. 166-188: 2002.
3. Yoneoka, Daisuke, *et al.*: Geographically weighted generalized Farrington algorithm for rapid outbreak detection over short data accumulation periods. *Statistics in Medicine*, 6277-6294: 2021.