

大規模疫病データストリームのための将来予測アルゴリズム

木村 輔[†] 松原 靖子[†] 川畑 光希[†] 櫻井 保志[†]

[†] 大阪大学産業科学研究所産業科学 AI センター 〒 567-0047 大阪府茨木市美穂ヶ丘 8-1

E-mail: †{tasuku,yasuko,koki,yasushi}@sanken.osaka-u.ac.jp

あらまし SARS-CoV-2(COVID-19) のような感染症が急速に拡大している現在、世界中で収集された疫病データを用いて、将来起こりうる大流行やパンデミックをできるだけ早く予測することは重要である。本論文では、大規模疫病データストリームのための高速予測手法である EpiCAST について述べる。EpiCAST は、様々な地域を含む大規模疫病データストリームが与えられたときに、その中から疫病の特徴を表現、要約、共有し、長期的かつ継続的に将来の感染者数予測を行う。提案手法は (a) 疫病の複雑な拡散過程を非線形モデルで表現し、(b) それらの中に含まれる感染拡大のパターンやトレンドを各地域で共有し、適切なモデルを選択することで感染拡大予測を実現する。ここで、提案手法は (c) データストリームの長さに依存せず、一定の計算時間で感染者数を推定する。COVID-19 の実データを用いた実験では、EpiCAST が大規模疫病データストリームの中から疫病の重要な特徴を発見、共有することで感染者数を長期的に予測し、さらに、最新の既存手法と比較し大幅な精度、性能向上を達成していることを確認した。

キーワード 新型コロナウイルス, 時系列データストリーム, 非線形動的システム, 将来予測

1 まえがき

新型ウィルスである COVID-19 の感染は世界中で拡大しており [1], 働き方やコミュニケーションの取り方など、人々の生活に強い影響を与えている [2,3]. 将来起こりうるアウトブレイク (感染爆発) やパンデミック (世界的大流行) に対して、政府や人々が最良の意思決定を下すためには、疫病の感染状況について、より正確な予測を得ることが重要である。もし、今後の感染者数を把握できれば、通勤等の頻繁な社会活動を抑制したり、病床の稼働率をコントロールすることで、パンデミックのリスクを事前に管理することが可能となる [4].

強力な感染力を持つ COVID-19 が引き起こす、さらなるパンデミックを回避するには、感染の拡大や収束を既に経験した地域から、疫病の特徴を他の地域でも適用可能な形式で抽出し、蓄積することで、拡散過程に関する知識を互いに共有し、活用することが重要である。しかし、アウトブレイクを抑制するために施行された対策によって、疫病のダイナミクスは時間の経過と共に変化する可能性がある [5]. よって、拡散過程の動的な変化を捉え、今後の感染状況を推定可能なストリーミング手法の確立は極めて重要な研究課題である。

本論文では、大規模疫病データのための将来予測アルゴリズムである EpiCAST [6] について述べる。EpiCAST は、疫病テンソルストリームに潜む疫病の振る舞いを非線形モデルで表現し、それらの中に含まれる重要な特徴を各地域で共有することで複雑な拡散過程を柔軟に表現する。

より具体的には、次の問題を扱う。

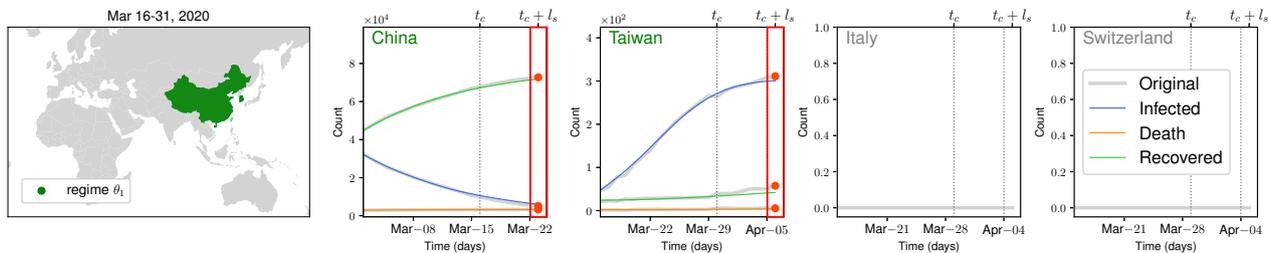
現在時刻 t_c において、 r カ所の場所で観測された d 次元の属性で構成される疫病テンソルストリーム $\mathcal{X} = \{x_{tij}\}_{t,i,j=1}^{t_c,d,r}$ が与えられたとき、 l_s ステップ先の未知の疫病テンソルストリームを継続的に予測する。

具体例. 図 1 は、COVID-19 の実データにおける提案手法の予測結果を示している。実データは、現在の感染者数、回復者数の総数および死者数の総数から構成されている。ここで、図の左側は EpiCAST-Map であり、各地域における典型的な感染パターン (θ_1, θ_2) を示す。以下、このようなパターンを“レジーム”と呼ぶ。図の右側は、各時刻 t_c における中国、台湾、イタリアおよびスイスの 7 日先の予測結果を示している。黒色の縦線は現在時刻 t_c を、赤色の丸印は現在時刻 t_c から 7 日先で観測される実際の値を表している。実データは灰色の実線で示されており、提案手法の各予測結果は、感染者数を青色、回復者数を緑色および死者数をオレンジ色の実線を表している。

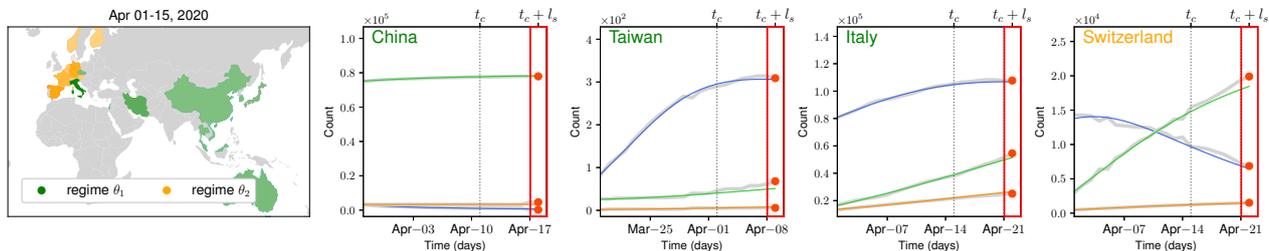
図 1 から、提案手法が疫病の拡散過程を捉え、長期予測に成功していることが確認できる。具体的には、EpiCAST は (a) 2020 年 3 月の 16 日から 31 日にかけてのアジアにおける COVID-19 の感染初期の広がり、および、(b) 2020 年 4 月の 1 日から 15 日にかけてのアジアからヨーロッパなどへの感染拡大の過程を捉えた。また感染者数、回復者数および死者数の予測結果から、提案手法が各地域の 7 日先の値を予測できていることがわかる。これは EpiCAST が、ストリーミング手法による感染予測を実現する上で望ましい性質である、(P1) 複雑な疫病テンソルストリームの非線形モデルリングと (P2) 複数の地域間におけるモデル共有の仕組みを持つためである。

本論文の貢献. 本論文では、大規模疫病データのための将来予測アルゴリズムである EpiCAST を提案する。EpiCAST は以下の特長を持つ。

- (1) 非線形方程式に基づいた疫病モデルを構築し、疫病の複雑な拡散過程をモデルパラメータとして抽出する。
- (2) 各地域の疫病モデルを共有し、適切なモデルを選択することで感染拡大予測を実現する。
- (3) 計算コストはデータストリームの長さに依存しない。



(a) 2020年3月16日から2020年3月31日における7日先の各予測結果



(b) 2020年4月1日から2020年4月15日における7日先の各予測結果

図1 COVID-19の感染者数、死亡者数および回復者数に対する提案手法の予測結果。

2 関連研究

この章では関連文献を、統計疫学および時系列解析とデータストリーミングの2つの観点から紹介する。

SARS-CoV-2 (COVID-19) は現在においても感染の拡大を繰り返している。これは COVID-19 の感染から発症までの潜伏期 (incubation period) が長く、かつ、感染した時点から感染性を有するまでの潜伏感染期 (latent period) が短いことが要因の1つである [7-9]。潜伏感染期が潜伏期より短ければ、感染者は何の兆候もなくウィルスを拡散させてしまう [10]。これまで COVID-19 について様々な研究が進められており、2020年2月にはその全ゲノムが特定されている [11,12]。COVID-19 は RNA の校正機能により変異しにくいとされてきたが [13]、感染期間の拡大により複製の試行回数が増加すると変異することが指摘されている [14]。そのため COVID-19 は変異を繰り返し、その性質を変化させ、拡散の過程をより複雑にしている。例えば、2020年3月から感染が拡大した D614G [15] は、アミノ酸配列の614番目の残基が変異した COVID-19 のミスセンス突然変異である。COVID-19 の変異株は世界中で出現し続けており [16]、例えば、2022年1月には、英国で初めて発見された B.1.1.7 が139カ国でその存在が確認され、同様に、87カ国において B.1.351 (南アフリカで初発見) が、54カ国において P.1 (ブラジルおよび日本で初発見) がそれぞれ報告されている。また血液型や民族的・遺伝的要因などの各個人が有する生物学的特性によって、ウィルスの影響 [17] やワクチンの効果 [18] が異なるため、同じ変異株が確認された地域でも流行パターンが異なる場合がある。これらの要因が重なることで、その感染過程はさらに複雑化している。結論として、COVID-19 は時間経過とともに変異を繰り返すため、その感染の拡散過程は変化しており、さらに感染が拡大している地域と変異株の組み合わせによっても拡散過程は変化する。

時系列データ解析は、多くの分野で関心を集めている重要なトピックである [19-24]。既存手法としては、自己回帰モデル (AR: autoregressive model) およびカルマンフィルタ (KF: Kalman filters) などがあり [25]、これらの手法は様々な拡張されてきた [26-28]。データセットに対する仮定が少ない、より一般的な問題設定のための非線形モデル [29,30] に加えて、適切な非線形微分方程式を選択することで、これまで観測されていない複雑なダイナミクスの予測を可能にするドメイン知識をモデルに適用できる [31,32]。また、計算時間や使用メモリの制約の下で、大量のデータを処理・解析するオンライン手法やストリーミング手法の重要性が高まっている [33-38]。一方で、新型コロナウイルスの出現により、パンデミック予測を重要な目標とする、リアルタイム性の高い手法も必要とされている [39,40]。しかしこれら既存手法による疫病モデルの学習において、そのスケーラビリティについては議論されていない。

Deep Neural Network (DNN) は、入力データから高次元の時間領域の特徴を取得し、様々なコンテキストやドメインにおける将来のイベントを予測する新たな手法として盛んに研究されている [41-46]。TCN [47] は、多階層の時間的因果関係を高速に学習できる時間方向の畳み込みニューラルネットワークモデルである。また最新手法である EpiDEEP [48] は、インフルエンザのデータセットにおいて、季節的に発生する感染症の拡散過程を DNN を用いてモデル化することに成功した。しかし、これら DNN ベースの手法は、数多くのモデルパラメータを推定するために膨大な量の学習データを必要とする。そのため、これまでのウィルスとは拡散過程が大きく異なり、かつ、学習データが少ない新型コロナウイルスのモデル化は困難である。

まとめると、いずれの既存手法も疫病テンソルストリームの非線形な振る舞いのモデル化、複数の地域間の拡散過程の共有および高速な感染者数予測に対応していない。本論文では、これらの要件を満たすことで、新型コロナウイルスの拡散過程を予測するためのストリーミングアルゴリズムを提案する。

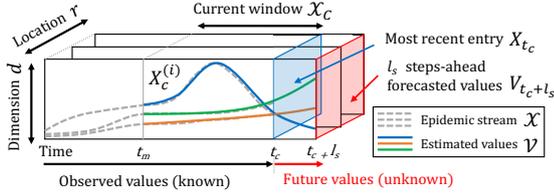


図2 EpiCASTの様子：提案手法は、疫病テンソルストリーム \mathcal{X} (灰点線) が与えられたとき、カレントウィンドウの時系列パターン (色太線) を推定することで、 l_s ステップ先の予測値を高速かつ継続的に出力する

3 提案モデル

本章では、疫病テンソルストリームのためのモデルを提案する。まず具体的な問題設定について述べ、その後、提案モデルについて詳細に説明する。

現在時刻 t_c までに観測された半無限長の疫病データ $\mathcal{X} = \{X_1, \dots, X_t, \dots, X_{t_c}, \dots\}$ を疫病テンソルストリームとする。毎時刻において新たに X_{t_c+1} が観測され \mathcal{X} の総量は増加する。ここで $X_t = \{x_{ij}(t)\}_{i,j=1}^{d,r}$ は、現在時刻 t_c に観測された r ヶ所の地域における d 次元の疫病データである。本論文では、1日の感染数、回復者数および死亡者数に対応するために次元数を $d=3$ とした。したがって $\mathcal{X} \in \mathbb{N}^{d \times r \times t_c}$ は、 r ヶ所の地域における d 次元の疫病データから構成される3階テンソルストリームとなる。

ストリーミング手法では処理速度が重要であるため、計算時間と使用メモリが小さくなるように制約され、多くの場合、これまで観測したデータの一部のみから将来を予測する必要がある。よって図2に示すように、最新の \mathcal{X} のみを用いて予測するための短い時間窓を定義する。具体的には、疫病テンソルストリームの最新の時系列を含む時間窓をカレントウィンドウ $\mathcal{X}_C = \{X_t\}_{t=t_m}^{t_c}$ と定義する。ここで t_m はカレントウィンドウの開始時刻である。同様に $V_{t_c+l_s} \in \mathbb{N}^{d \times r}$ は、推定値を生成したい期間に対応する l_s ステップ先の予測ウィンドウである。本論文で取り組む問題を以下のように定義する。

[Problem 1] 疫病テンソルストリーム $\mathcal{X} \in \mathbb{N}^{d \times r \times t_c}$ およびカレントウィンドウ $\mathcal{X}_C = \{X_t\}_{t=t_m}^{t_c}$ が与えられたとき、 l_s ステップ先の予測ウィンドウ $V_{t_c+l_s} \in \mathbb{N}^{d \times r}$ を生成する。

3.1 EpiCast – 単一地域の疫病ストリーム場合

1章のCOVID-19の疫病ストリームのスナップショットから確認できるように、疫病は非線形に拡散する傾向がある。まず解決すべき問題は、アウトブレイクからパンデミックまでの複雑なダイナミクスをモデル化する方法を見つけることである。ここで単純化のために、疫病テンソルストリームの単一の地域にのみ焦点を当て、SEIRベースのモデルであるEpiCAST-baseを提案する。提案モデルは次の5つの状態から構成される。

- **S**usceptible (感受性保持者)：疫病を患う可能性のある状態
- **E**xposed (潜伏者)：感染率 β で疫病に感染した状態
- **I**nfected (感染者)：発症率 σ で疫病を発症した状態
- **R**ecovered (回復者)：回復率 γ で罹患から回復した状態
- **D**eath (死亡者)：死亡率 δ で罹患から死亡した状態

これら5つのクラスの時間依存性は、次の式で表される。

$$\begin{aligned} \frac{dS}{dt} &= -\beta S(t)I(t), & \frac{dE}{dt} &= \beta S(t)I(t) - \sigma E(t), \\ \frac{dI}{dt} &= \sigma E(t) - \gamma I(t) - \delta I(t), & \frac{dR}{dt} &= \gamma I(t), & \frac{dD}{dt} &= \delta I(t). \end{aligned} \quad (1)$$

ここで、モデルは各推定値を生成するために、潜伏者数、感染者数、回復者数および死亡者数について、初期値 E_0, I_0, R_0, D_0 をそれぞれ必要とする。また、総人口の概念に従って、初期の感受性集団 S_0 は、疫病の潜在的人口 N を用いて、 $S_0 = N - (E_0 + I_0 + R_0 + D_0)$ と計算することで求めることができる。よって、推定したいパラメータ集合全体は以下のようなになる。

[Model 1] (EpiCAST-base) EpiCAST-baseのパラメータ集合を $\theta = \{\beta, \sigma, \gamma, \delta, N, E_0, I_0, R_0, D_0, t_0\}$ とする。疫病の感染率を β 、疫病の発症率を σ 、疫病的回復率を γ および疫病的死亡率を δ とし、それぞれ0から1の値をとる。ここで $E_0, I_0, R_0, D_0 (\geq 0)$ は、時刻 t_0 における潜伏者、感染者、回復者および死亡者の初期人口である。

またModel 1によって生成されるウィンドウとして、 $\mathcal{V} \in \mathbb{N}^{d \times r \times t}$ 、 $V_t \in \mathbb{N}^{d \times r}$ 、 $V^{(i)} \in \mathbb{N}^{d \times t}$ および $\mathbf{v}_i(t) \in \mathbb{N}^d$ ($i = 1, \dots, r$) を使用する。例えば i 番目の地域のウィンドウ $V^{(i)} \subset \mathcal{V}$ は、Model 1を用いて初期値 $\mathbf{v}_i(t_0) = \{E_0, I_0, R_0, D_0\}$ から値 $\mathbf{v}_i(t)$ を推定し続けることで生成される。

3.2 EpiCast – 複数地域の疫病ストリームの場合

提案手法においてより重要な目標は、異なる時刻において発生する可能性のある、地域間の類似するダイナミクスを特定することである。これは、感染の初期段階のような共通する感染傾向を、各地域が潜在的に保持していると考えられるためである。単一の地域でモデルを推定するために十分な観測値が得られない場合には、別の地域で得られたモデルを予測に適用すべきであるというのが、本手法が持つべき最も重要な性質である。

そこで本研究では、地域毎に人口や文化のような特徴に違いがあることを考慮し、 θ のパラメータを、疫病パラメータ $\theta^E = \{\beta, \sigma, \gamma, \delta\}$ および地域パラメータ $\theta^L = \{N, E_0, I_0, R_0, D_0, t_0\}$ の2つのグループに分ける。ここで、 $\theta = \theta^E \cup \theta^L$ である。疫病パラメータ θ^E は、どの地域でも共有可能であるのに対し、地域パラメータ θ^L は、地域毎に最適化されている。例えば、異なる地域で同じ感染対策が講じられていても、それを施行するタイミングが地域によって異なる場合、ダイナミクスが変化する時期は異なるが、共通のダイナミクスを保持する可能性が高い。よって複数のEpiCAST-baseのパラメータ集合を保持・共有することで、各地域で得られた知識を他の地域の予測に反映させることが可能となる。最終的に、EpiCAST-fullのパラメータ集合を以下のように定義する。

[Model 2] (EpiCAST-full) EpiCAST-fullのパラメータ集合を $\Theta = \{\Theta^E, \Theta^L\}$ とする。 $\Theta^E = \{\theta_1^E, \dots, \theta_g^E\}$ は g 種類の疫病パラメータを表し、 $\Theta^L = \{\theta_1^L, \dots, \theta_r^L\}$ は r カ所の地域パラメータを表す。ただし複数の地域で同じ疫病モデルが共有されるため $g \ll r$ である。

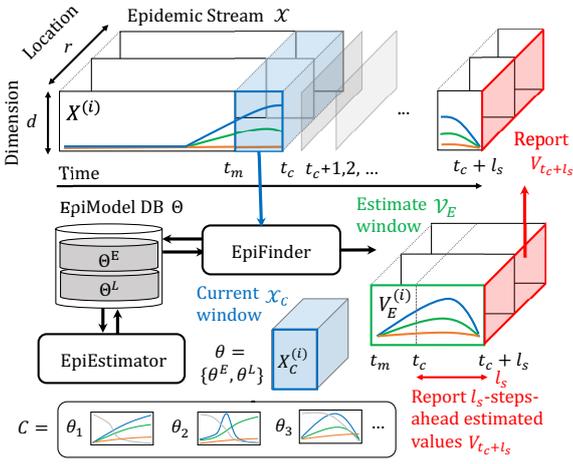


図3 EpiCASTのアルゴリズム概要

4 アルゴリズム

これまで、疫病テンソルストリームにおける非線形の拡散過程をモデル化する方法を述べた。本章では、複数の疫病モデルを時間発展と共に獲得するストリーミングアルゴリズムEpiCASTを提案する。本アルゴリズムは、以下に示す2つの重要な特性を満たす。

(P1) 複雑な疫病テンソルストリームの非線形モデリング

(P2) 複数の地域間におけるモデル共有の仕組み

まず、非線形に振る舞う感染データにおける、複雑なダイナミクスを捉える必要がある。(P1)に対して、非線形微分方程式を用いたストリーミングアルゴリズムを提案する。また、(P2)に対して、地域間の類似したダイナミクスを検出し、別の地域で得られた疫病モデルを、他の地域の予測に利用できるアルゴリズムを提案する。

図3は、以下の3つのアルゴリズムで構成されるEpiCASTの概要を示している。

- EpiESTIMATOR: i 番目の地域のカレントウィンドウ $X_C^{(i)} \subset \mathcal{X}_C$ が与えられたとき、新たなモデル $\theta = \{\theta^E, \theta^L\}$ を推定。
- EpiFINDER: i 番目の地域のカレントウィンドウ $X_C^{(i)} \subset \mathcal{X}_C$ およびモデルパラメータ集合 Θ が与えられたとき、 Θ から最良の疫病パラメータ θ^E および地域パラメータ θ^L を取得し、推定ウィンドウ $V_E^{(i)}$ を生成。
- EpiCAST: 疫病テンソルストリーム \mathcal{X} が与えられたとき、カレントウィンドウ \mathcal{X}_C を観測し、上記の2つのアルゴリズムを実行することで推定ウィンドウ V_E を得る。その後、 l_s ステップ先の予測ウィンドウ $V_{t_c+l_s} \in \mathbb{N}^{d \times r}$ を報告する。さらにモデルパラメータ集合 Θ を更新。

4.1 モデル推定 – EpiEstimator

議論の単純化のため、まず疫病ストリーム上の非線形モデルを効果的に推定する方法を説明する。具体的には、疫病テンソルストリーム \mathcal{X} 内の単一地域に焦点を当て、カレントウィンドウ X_C のモデルパラメータ $\theta = \{\theta^E, \theta^L\}$ を推定で

きるEpiESTIMATORを提案する。 $X_C^{(i)} \subset \mathcal{X}_C$ ($i = 1, \dots, r$) を i 番目の地域のカレントウィンドウとする。EpiESTIMATORは、 $X_C^{(i)}$ が与えられたとき、次の式を最小化することで最適な $\theta = \{\theta^E, \theta^L\}$ を発見する。

$$\{\theta^E, \theta^L\} = \arg \min_{\theta^E, \theta^L} \|X_C^{(i)} - V_C^{(i)}\|, V_C^{(i)} = f(\theta^E, \theta^L), \quad (2)$$

ここで、 $\|\cdot\|$ は平均二乗誤差を表し、 $V_C^{(i)} = f(\theta^E, \theta^L)$ は $X_C^{(i)}$ と同じ期間を持つ、式1で生成されたウィンドウを表す。最適な θ を得るためには、非線形最小二乗問題を解く必要がある。 θ のすべてのパラメータは、非線形性を有する学習に適したLevenberg-Marquardt (LM) アルゴリズム [49] によって最適化する。また、予測ウィンドウの各次元の値は、4次のルンゲ・クッタ法に基づき生成する。

次に疫病ストリームにおいて、各地域のアウトブレイクはランダムに発生する。そのためカレントウィンドウの初期値 $\{E_0, I_0, R_0, D_0\} \in \theta$ を推定する際に、適切な感染の開始時刻 t_0 を選択することは重要である。なぜならば、この開始時刻 t_0 の選択がモデルの品質に影響を与えるためである。そこで本研究では、最適な開始時刻 $t_0 \in [t_m : t_c]$ を探索する。 t_0 を $t_m, t_m + 1, t_m + 2, \dots$ と変化させながら $X[t_0 : t_c]$ を用いて新しいパラメータ θ を推定し、 $t_c - t_0 + 1$ 個の候補モデル θ を取得する。 $X_C^{(i)}$ と $V_C^{(i)}$ の二乗誤差にしたがって、 $X_C^{(i)}$ に最適なモデルを選択する。この処理において、探索の候補となる時刻の内、 $t'_m = \{t \mid X_t = 0\}$ となる範囲では、時刻 t において、アウトブレイクが発生しないことが保証されているので、開始時刻 t_0 を探索する範囲から排除することができる。これにより、より効率的に θ を推定することが可能となる。

4.2 モデル選択 – EpiFinder

カレントウィンドウ \mathcal{X}_C のダイナミクスが、時間と場所によって変化する疫病テンソルストリームでは、単一地域における疫病の振る舞いに関する情報が著しく不足しているため、任意の地域から得られる複数の非線形モデルを利用する必要がある。そこで本研究では、既存モデルのパラメータ集合 Θ からカレントウィンドウ \mathcal{X}_C に最適なモデル θ を選択するEpiFINDERを提案する。

Algorithm 1は、EpiFINDERの詳細を示している。ここでは、 i 番目の地域のカレントウィンドウ $X_C^{(i)} \subset \mathcal{X}_C$ について、既存モデルのパラメータ集合 $\Theta = \{\theta^E, \theta^L\}$ が与えられた場合を考える。ここで、 $\theta^E = \{\theta_1^E, \dots, \theta_g^E\}$ は g 種類の疫病パラメータ集合、 $\theta^L = \{\theta_1^L, \dots, \theta_r^L\}$ は r カ所の地域パラメータ集合を表す。 $X_C^{(i)}$ および Θ が与えられたとき、アルゴリズムは、 $X_C^{(i)}$ および式1で生成される $V_C^{(i)}$ の二乗誤差に従って最適なモデル $\theta = \{\theta^E, \theta^L\}$ を選択する。しかし、このステップにおいて、 i 番目の地域から推定されたモデルではなく、別の地域で推定された異なる人口規模等を持つモデル θ を使用する可能性がある。そのような場合、他の地域のモデルを適用するには効果的な微調整が必要である。

Algorithm 1 EPIFINDER($X_C^{(i)}, \Theta$)

```
1: Input: (a) Current epidemic sequence  $X_C^{(i)}$  in  $i$ -th location
           (b) Current parameter set  $\Theta = \{\Theta^E, \Theta^L\}$ 
2: Output: (a) Estimated values  $V_E^{(i)}$  in  $i$ -th location
           (b) Model parameter set  $\theta = \{\theta^E, \theta^L\}$ 
3:  $C = \emptyset$ ; // Candidate parameter set
4: for  $j = 1 : g$  do
5:   /* Model estimation with the  $j$ -th epidemic parameters and
      the  $i$ -th location parameters */
6:   Set  $\{\theta^{E'}, \theta^{L'}\} = \{\theta^{E_j}, \theta^{L_j}\}$  as initial condition
7:    $\{\theta^E, \theta^L\} = \arg \min_{\theta^{E'}, \theta^{L'}} \|X_C^{(i)} - V_C^{(i)}\|$ ; //  $V_C^{(i)} = f(\theta^{E'}, \theta^{L'})$ 
8:    $C = C \cup \{\theta^E, \theta^L\}$ ;
9: end for
10: /* Choose the best model in  $C$  */
11:  $\{\theta^E, \theta^L\} = \arg \min_{\theta^{E'}, \theta^{L'} \in C} \|X_C^{(i)} - V_C^{(i)}\|$ ; //  $V_C^{(i)} = f(\theta^{E'}, \theta^{L'})$ 
12: Compute  $V_E^{(i)} = f(\theta^E, \theta^L)$ ;  $\theta = \{\theta^E, \theta^L\}$ ;
13: return  $\{V_E^{(i)}, \theta\}$ ;
```

ここで EPIFINDER は, Θ ($j = 1, \dots, g$) から最適な疫病パラメータ θ_j^E を検索し, 現在の地域パラメータ θ_i^L を使用して, i 番目の地域に最適なパラメータを推定する. ここで, θ 全体が推定対象の地域に過適合しないように, 疫病パラメータ $\theta^E = \{\beta, \sigma, \gamma, \delta\}$ については特定の値の範囲¹ でパラメータを更新した. このようにすることで, 疫病パラメータ θ^E がもつ基本的な拡散過程を維持できることが期待される.

4.3 ストリームアルゴリズム – EpiCast

我々の最終目的は, 様々な地域で発生した疫病モデル $\Theta = \{\Theta^E, \Theta^L\}$ を捉え, すべての地域の l_s ステップ先の値 $V_{t_c+l_s} \in \mathbb{R}^{d \times r}$ を推定することである. そこで, 時間発展する疫病の非線形ダイナミクスを考慮した, 疫病テンソルストリームの高速な予測を実現するストリーミング手法 EPICAST を提案する.

Algorithm 2は, EPICAST の手順をまとめたものである. アルゴリズムは, カレントウィンドウ \mathcal{X}_C について, 次の3つの手順に従って各地域の最良な非線形モデルを得る.

まず, アルゴリズムは時刻 $t_c - 1$ に推定したパラメータ $\theta_i^{(t_c-1)}$ の適用を試みる. これは, 変異株の発生や拡散は稀に発生するという自然な仮定に基づいており, 効率化のためにアルゴリズムは $\theta_i^{(t_c-1)}$ を使い続ける.

次に, 時刻 $t_c - 1$ に推定したパラメータが最適でなかった場合, つまりは $X_C^{(i)}$ と $V_C^{(i)}$ の誤差が ϵ^2 以上のとき, アルゴリズムは EPIFINDER を用いて最適なパラメータを選択する. また時間経過に伴う Θ のモデル増加に対応するために, Θ を階層化することで検索を効率化する方法を用いた. 具体的には, Θ を地理情報に基づいてローカルエリアに分割する. これにより, 近隣の地域間の拡散過程の類似性を考慮することができる. ア

Algorithm 2 EPICAST(\mathcal{X}_C, Θ)

```
1: Input: (a) Current window  $\mathcal{X}_C = \mathcal{X}[t_m : t_c]$ 
           (b) Model parameter set  $\Theta = \{\Theta^E, \Theta^L\}$ 
2: Output: (a)  $l_s$ -steps-ahead values  $V_{t_c+l_s} = \mathcal{V}[t_c + l_s]$ 
           (b) Updated model parameter set  $\Theta' = \{\Theta^{E'}, \Theta^{L'}\}$ 
3: /* (I) Estimate optimal parameters for each  $i$ -th location
   *** */
4: for  $i = 1 : r$  do
5:   /* (1) Fitting by the previous best model at time  $t_c - 1$  */
6:    $\{\theta^E, \theta^L\} = \theta_i^{(t_c-1)}$ ; Compute  $V_C^{(i)} = f(\theta^E, \theta^L)$ ;
7:   if  $\|X_C^{(i)} - V_C^{(i)}\| > \epsilon$  then
8:     /* (2-1) Fitting by local area  $S$  for the  $i$ -th location */
9:      $\Theta^{(S)} \subset \Theta$ ; // Subset models of local area  $S$ 
10:     $\{V_E^{(i)}, \theta\} = \text{EPIFINDER}(X_C^{(i)}, \Theta^{(S)})$ ;
11:     $V_C^{(i)} = V_E^{(i)}[t_m : t_c]$ ; // Estimated values from  $t_m$  to  $t_c$ 
12:    end if
13:    if  $\|X_C^{(i)} - V_C^{(i)}\| > \epsilon$  then
14:      /* (2-2) Fitting by other local areas */
15:       $\Theta^{(\neg S)} \subset \Theta$ ; // Subset models of local area, except  $S$ 
16:       $\{V_E^{(i)}, \theta\} = \text{EPIFINDER}(X_C^{(i)}, \Theta^{(\neg S)})$ ;
17:       $V_C^{(i)} = V_E^{(i)}[t_m : t_c]$ ; // Estimated values from  $t_m$  to  $t_c$ 
18:      end if
19:     /* (3) Estimate new regimes (if required) */
20:     if  $\|X_C^{(i)} - V_C^{(i)}\| > \epsilon$  then
21:        $\{\theta^E, \theta^L\} = \text{EPIESTIMATOR}(X_C^{(i)})$ ;
22:        $\Theta^E = \Theta^E \cup \theta^E$ ;  $g = g + 1$ ; // Insert new model
23:       Compute  $V_E^{(i)} = f(\theta^E, \theta^L)$ ;
24:       end if
25:       Update  $\theta_i^L$  in  $\Theta^L$ ;
26:     end for
27: /* (II)  $l_s$ -steps-ahead future value generation *** */
28:  $V_{t_c+l_s} = \{V_E^{(i)}[t_c + l_s]\}_{i=1}^r$ ;
29: return  $\{V_{t_c+l_s}, \Theta'\}$ ;
```

ルゴリズムは, i 番目の地域がローカルエリア S に属する場合, $\Theta^{(S)} \subset \Theta$ のみを対象に EPIFINDER を用いて最良のパラメータを取得する. もし最適なモデルが存在しない場合, 他のローカルエリア $\Theta^{(\neg S)} \subset \Theta$ を対象に EPIFINDER を実行する.

最後に, Θ に適切なモデルがない場合, EPIESTIMATOR を用いて新しいモデル $\{\theta^E, \theta^L\}$ を推定する. その後, Θ^E に新たに推定された疫病パラメータ θ^E を追加し, Θ^L の地域パラメータを θ^L に更新する.

この手順を各地域に対して繰り返し行うことで, 最終的に $V_{t_c+l_s}$ を得ることができる.

[Lemma 1] 各時刻における EPICAST の計算時間は $O(g)$ となる.

[Proof 1] EPICAST のアルゴリズムにおいて, EPIFINDER および EPIESTIMATOR はパラメータの推定に $O(g \cdot d \cdot r \cdot l_c)$ の計算時間を要し, 最大で l_c 回繰り返される. d, r, l_c によって与えられるカレントウィンドウのサイズは無視できるほど小さい定数値であるので, EPICAST の計算時間は時間点あたりの $O(g)$ 時間である.

1: 本論文ではパラメータ更新時の値の範囲を ± 0.1 とする.

2: 本論文では, $\epsilon = 1/2 \|X_C^{(i)}\|$ とする.

5 評価実験

本章では EpiCAST の有効性を検証するため、COVID-19 の実データ³ を用いて実験を行った。本データセットは、600 日にもわたる疫病データの 3 次元のベクトル（感染者数、回復者数および死者数）で構成される。なお、感染拡大初期におけるデータの欠損に対応するために、GDP スコアの高い順に上位 50 カ国を用いた⁴。また実験は、16GB のメモリ、Intel Core i7 2.8GHz quad core の CPU を搭載したマシン上で実施した。本実験では、以下の項目について検証する。

Q1 疫病ストリームの予測に対する提案手法の有効性

Q2 疫病の拡散過程の予測に対する提案手法の精度の検証

Q3 疫病ストリームの予測に対する計算時間の検証

Q1. 提案手法の有効性. 本節では、疫病テンソルストリームに対する EpiCAST の予測能力を検証する。すでに 1 章の図 1 においても示したように、EpiCAST は、長期的な将来の感染者数の効果的な予測に成功している。ここでは図 1 の疫病ストリームにおいて、さらに先の期間について提案手法が予測した実験結果を図 4 に示す。図 4 (a) からわかるように、4 月末において中国では感染の拡大が抑えられていることがわかる。同様に他の国の流行状況も変化しており、図 4 (b) からわかるように、中国と同じレジームで予測された台湾も 5 月末には感染が収束していることが確認できる。一方で、図 4 (a) および (b) から、4 月末時点で中国と同じレジームだったイタリアは、5 月末においても感染を拡大させており、高い死亡率を維持したまま、異なる拡散過程をもつレジームへ遷移していることが確認できる。

図 5 は、チェコにおける EpiCAST の出力の詳細を示している。ここで上段の時系列データは、オリジナルデータであり、チェコにおける感染者数、回復者数および死者数から構成されている。それに対して下段の時系列データは、提案手法の 7 日先の予測結果を示している。また、オリジナルデータと予測結果について、線形（左列）と対数（右列）の両方のスケールで示している。図から、提案手法は感染の拡散過程が異なる 3 つのレジームを検出することで、将来の疫病データの振る舞いについて予測していることが確認できる。また、図 5 の右側の散布図は、提案手法が推定したレジームにおける、各モデルパラメータの相対的な特徴を示している。具体的には、黒色の丸印は Θ に格納されている各レジームに対応し、これらは EpiCAST によって推定される。図から 3 種類のレジーム間における拡散過程の変遷 (A)→(B)→(C) が確認できる。ここで、レジーム (A) およびレジーム (C) はレジーム (B) よりも感染率 β が高く、レジーム (B) は他の 2 種類のレジームより死亡率 γ が高い。実際、チェコにおいてレジーム (A) に対応する黄色の期間に最初のパンデミックが発生した。一時事態は収束し、レジーム (B) の期間に感染の拡大は落ち着いたが、感染者数に対して死者数が多い状況が続いていた。その後、レジーム (C) の期間において、最初のパンデミックよりも急速にウィルスが蔓延している。

図に示すように、EpiCAST は、複数の地点の異なる時点における様々なレジーム、すなわち、非線形な動的パターンを捉えることに成功している。このように、複雑かつ時間発展するダイナミクスの特徴をレジームとして要約することで、提案手法は、効果的かつ効率的に疫病の拡散過程を捉え、将来を予測することができる。

Q2. 提案手法の精度. 次に、本章では、EpiCAST の予測精度を検証するため、非線形の疫病モデルである SIRD、線形の時系列予測手法である ARIMA、再帰型 NN である GRU、多階層の時間的因果関係を高速に学習する畳み込み NN である TCN [47] および DNN を用いた感染予測の最新手法である EpiDEEP [48] と比較した。

図 6 は、疫病テンソルストリームにおける EpiCAST の予測精度を示している。具体的には、オリジナルデータと 7 日先の各推定値の平均二乗誤差 (RMSE: root mean square error) を示している。図に示すとおり、提案手法は感染者数、死亡者数および回復者数の 3 種類すべてで高い予測精度を達成した。

Q3. 提案手法の計算時間. 続いて、提案アルゴリズムの計算コストを検証する。図 7 の左側は、各時点 t_c における計算コストを、提案手法の EpiCAST と既存手法について比較した結果である。なお、グラフは対数スケールで表示されている。EpiCAST の実験結果において、いくつかのスパイクが見られるが、これは、EpiESTIMATOR が新しいモデルを生成したことを表しており、時間の経過とともに、レジームの数が増加していることがわかる。一方で、モデルの生成を除いた範囲において、提案手法の計算時間はデータストリームの終了まで一定であることが確認できる。この結果から、EpiFINDER が複数レジームのストリームマイニングに有効に働くことを示唆している。

図 7 の右側は、疫病ストリーム全体の計算時間の平均値を示している。図に示すとおり、EpiCAST は既存手法と比較し、長期的なイベント予測に対する大幅な性能向上を達成した。

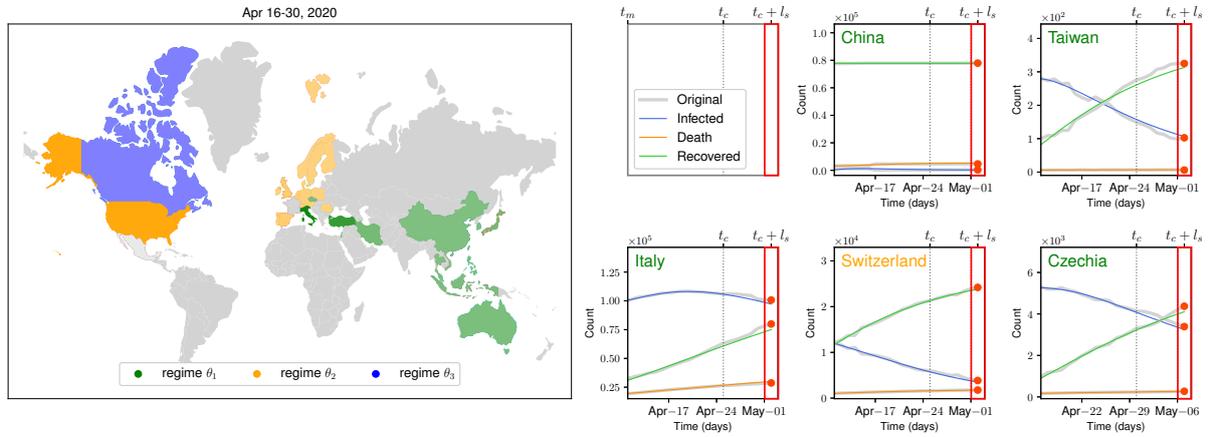
6 むすび

本論文では、大規模疫病データのための高速予測手法である EpiCAST について述べた。EpiCAST は、(a) 疫病の複雑な拡散過程を非線形モデルで表現し、(b) それらの中に含まれる重要な特徴を各地域で共有し、適切なモデルを選択することで、効果的な感染拡大予測を実現する。また、(c) データストリームの長さ依存せず、一定の計算時間で効率的に感染者数を推定する。COVID-19 の実データセットを用いて、大規模疫病データストリームの中から疫病の重要な特徴を発見、共有することで感染者数を長期的に予測し、さらに、最新の既存手法と比較し大幅な精度、性能向上を達成していることを確認した。

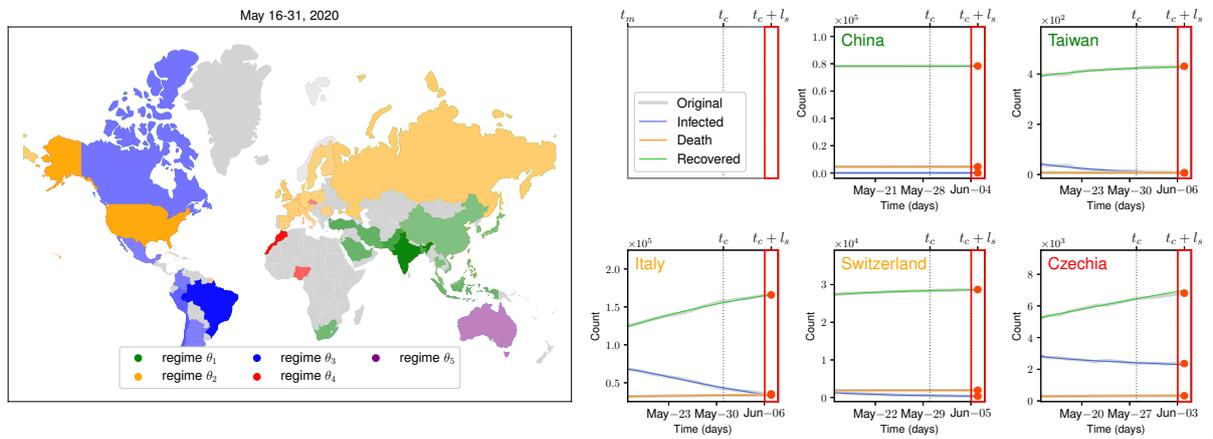
謝辞. 本研究の一部は JSPS 科研費、JP20H00585、JP21H03446、JP22K17896、国立研究開発法人情報通信研究機構委託研究 NICT 03501、総務省 SCOPE JP192107004、JSTAIIP 加速課題 JPMJCR21U4、ERCA 環境研究総合推進費 JPMEERF20201R02、の助成を受けたものです。

3 : <https://www.worldometers.info/coronavirus/>

4 : <https://www.imf.org/external/index.htm>



(a) 2020年4月16日から2020年4月30日における7日先の各予測結果。



(b) 2020年5月16日から2020年5月31日における7日先の各予測結果。

図4 COVID-19の感染者数、死亡者数および回復者数に対する提案手法の予測結果。

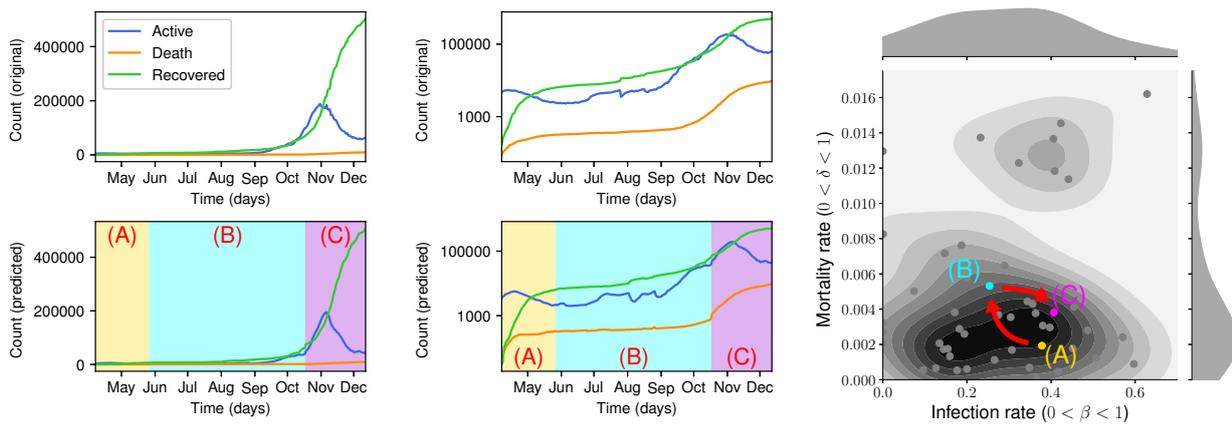


図5 チェコにおけるCOVID-19の提案手法の7日先の予測結果。

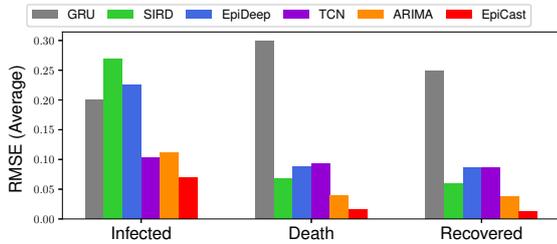


図 6 COVID-19 の感染者数、死亡者数および回復者数に対する 7 日先の予測精度の比較。

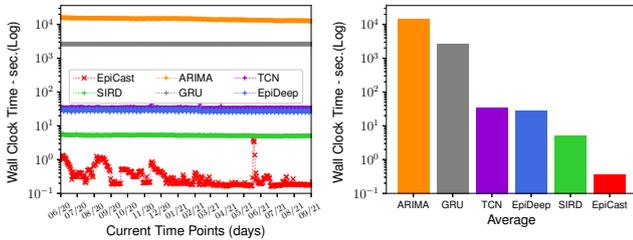


図 7 EpiCAST および既存手法の各時刻における計算時間と平均計算時間。

文 献

- [1] WHO: Coronavirus disease 2019 (COVID-19): situation report, 72 (2020).
- [2] Andreadis, Georgios and Quirós Gámez, Ana Isabel: Prospective analysis of the impact of a pandemic in Industry 4.0.
- [3] Prem, K., Liu, Y., Russell, T., Kucharski, A., Eggo, R., Davies, N., Jit, M., Klepac, P., Flasche, S., Clifford, S., Pearson, C., Munday, J., Abbott, S., Gibbs, H., Rosello, A., Quilty, B., Jombart, T., Sun, F., Diamond, C. and Hellewell, J.: The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study, *The Lancet Public Health*, Vol. 5 (2020).
- [4] Adhikari, B., Lewis, B. L., Vullikanti, A., Jiménez, J. M. and Prakash, B. A.: Fast and near-optimal monitoring for healthcare acquired infection outbreaks, *PLoS Comput. Biol.*, Vol. 15, No. 9 (2019).
- [5] Wang, C. J., Ng, C. Y. and Brook, R. H.: Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing, *JAMA*, Vol. 323, No. 14, pp. 1341–1342 (2020).
- [6] Kimura, T., Matsubara, Y., Kawabata, K. and Sakurai, Y.: Fast Mining and Forecasting of Co-Evolving Epidemiological Data Streams, *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, New York, NY, USA, Association for Computing Machinery, p. 3157–3167 (online), 10.1145/3534678.3539078 (2022).
- [7] Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., Hui, D. S., Du, B., Li, L.-j., Zeng, G., Yuen, K.-Y., Chen, R.-c., Tang, C.-l., Wang, T., Chen, P.-y., Xiang, J., Li, S.-y., Wang, J.-l., Liang, Z.-j., Peng, Y.-x., Wei, L., Liu, Y., Hu, Y.-h., Peng, P., Wang, J.-m., Liu, J.-y., Chen, Z., Li, G., Zheng, Z.-j., Qi, S.-q., Luo, J., Ye, C.-j., Zhu, S.-y. and Zhong, N.-s.: Clinical Characteristics of Coronavirus Disease 2019 in China, *New England Journal of Medicine*, Vol. 382, No. 18, pp. 1708–1720 (2020).
- [8] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q.,

- Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T. T., Wu, J. T., Gao, G. F., Cowling, B. J., Yang, B., Leung, G. M. and Feng, Z.: Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia, *New England Journal of Medicine*, Vol. 382, No. 13, pp. 1199–1207 (2020).
- [9] Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G. and Lessler, J.: The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application, *Annals of Internal Medicine*, Vol. 172, No. 9, pp. 577–582 (2020).
- [10] Nishiura, H., Linton, N. M. and Akhmetzhanov, A. R.: Serial interval of novel coronavirus (COVID-19) infections, *International Journal of Infectious Diseases*, Vol. 93, pp. 284–286 (2020).
- [11] Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C. and Zhang, Y.-Z.: A new coronavirus associated with human respiratory disease in China, *Nature*, Vol. 579, No. 7798, pp. 265–269 (2020).
- [12] Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., Zheng, X.-S., Zhao, K., Chen, Q.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F. and Shi, Z.-L.: A pneumonia outbreak associated with a new coronavirus of probable bat origin, *Nature*, Vol. 579, No. 7798, pp. 270–273 (2020).
- [13] Minskaia, E., Hertzog, T., Gorbalenya, A. E., Campanacci, V., Cambillau, C., Canard, B. and Ziebuhr, J.: Discovery of an RNA virus 3'→5' exoribonuclease that is critically involved in coronavirus RNA synthesis, Vol. 103, No. 13, pp. 5108–5113 (2006).
- [14] Holmes, E. C., Dudas, G., Rambaut, A. and Andersen, K. G.: The evolution of Ebola virus: Insights from the 2013–2016 epidemic, *Nature*, Vol. 538, No. 7624, pp. 193–200 (2016).
- [15] Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E. E., Bhattacharya, T., Foley, B., Hastie, K. M., Parker, M. D., Partridge, D. G., Evans, C. M., Freeman, T. M., de Silva, T. I., Angyal, A., Brown, R. L., Carrilero, L., Green, L. R., Groves, D. C., Johnson, K. J., Keeley, A. J., Lindsey, B. B., Parsons, P. J., Raza, M., Rowland-Jones, S., Smith, N., Tucker, R. M., Wang, D., Wyles, M. D., McDanal, C., Perez, L. G., Tang, H., Moon-Walker, A., Whelan, S. P., LaBranche, C. C., Saphire, E. O. and Montefiori, D. C.: Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus, *Cell*, Vol. 182, No. 4, pp. 812–827.e19 (2020).
- [16] WHO: COVID-19 weekly epidemiological update, 27 April 2021 (2021).
- [17] Group, S. C.-. G.: Genomewide Association Study of Severe Covid-19 with Respiratory Failure, *New England Journal of Medicine*, Vol. 383, No. 16, pp. 1522–1534 (2020).
- [18] Liu, G., Carter, B. and Gifford, D. K.: Predicted Cellular Immunity Population Coverage Gaps for SARS-CoV-2 Subunit Vaccines and Their Augmentation by Compact Peptide Sets, *Cell systems*, Vol. 12, No. 1, pp. 102–107.e4 (2021).
- [19] Sakurai, Y., Matsubara, Y. and Faloutsos, C.: Mining and Forecasting of Big Time-series Data, *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015* (Sellis, T. K., Davidson, S. B. and Ives, Z. G., eds.), ACM, pp. 919–922 (2015).

- [20] Faloutsos, C., Gasthaus, J., Januschowski, T. and Wang, Y.: Classical and Contemporary Approaches to Big Time Series Forecasting, *SIGMOD*, pp. 2042–2047 (2019).
- [21] Paparrizos, J., Liu, C., Elmore, A. J. and Franklin, M. J.: Debunking Four Long-Standing Misconceptions of Time-Series Distance Measures, *SIGMOD*, pp. 1887–1905 (2020).
- [22] Yuan, H., Li, G., Bao, Z. and Feng, L.: Effective Travel Time Estimation: When Historical Trajectories over Road Networks Matter, *SIGMOD*, pp. 2135–2149 (2020).
- [23] Wen, Q., He, K., Sun, L., Zhang, Y., Ke, M. and Xu, H.: RobustPeriod: Robust Time-Frequency Mining for Multiple Periodicity Detection, *SIGMOD*, pp. 2328–2337 (2021).
- [24] Ma, L., Aken, D. V., Hefny, A., Mezerhane, G., Pavlo, A. and Gordon, G. J.: Query-based Workload Forecasting for Self-Driving Database Management Systems, *SIGMOD*, pp. 631–645 (2018).
- [25] Durbin, J. and Koopman, S. J.: *Time Series Analysis by State Space Methods*, Oxford University Press, 2 edition (2012).
- [26] Shaman, J. and Karspeck, A.: Forecasting seasonal outbreaks of influenza, *Proceedings of the National Academy of Sciences*, Vol. 109, No. 50, pp. 20425–20430 (2012).
- [27] Rogers, M., Li, L. and Russell, S. J.: Multilinear Dynamical Systems for Tensor Time Series, *NIPS*, pp. 2634–2642 (2013).
- [28] Shi, Q., Yin, J., Cai, J., Cichocki, A., Yokota, T., Chen, L., Yuan, M. and Zeng, J.: Block Hankel Tensor ARIMA for Multiple Short Time Series Forecasting, *AAAI/IAAI*, pp. 5758–5766 (2020).
- [29] Matsubara, Y., Sakurai, Y., Prakash, B. A., Li, L. and Faloutsos, C.: Rise and fall patterns of information diffusion: model and implications, *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012* (Yang, Q., Agarwal, D. and Pei, J., eds.), ACM, pp. 6–14 (2012).
- [30] Taghvaei, A., de Wiljes, J., Mehta, P. G. and Reich, S.: Kalman Filter and its Modern Extensions for the Continuous-time Nonlinear Filtering Problem, *CoRR*, Vol. abs/1702.07241 (2017).
- [31] Matsubara, Y., Sakurai, Y., van Panhuis, W. G. and Faloutsos, C.: FUNNEL: automatic mining of spatially coevolving epidemics, *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014* (MacKassay, S. A., Perlich, C., Leskovec, J., Wang, W. and Ghani, R., eds.), ACM, pp. 105–114 (2014).
- [32] Song, H. A., Hooi, B., Jereminov, M., Pandey, A., Pileggi, L. T. and Faloutsos, C.: PowerCast: Mining and Forecasting Power Grid Sequences, *PKDD, Lecture Notes in Computer Science*, Vol. 10535, pp. 606–621 (2017).
- [33] Liu, C., Hoi, S. C. H., Zhao, P. and Sun, J.: Online ARIMA Algorithms for Time Series Prediction, *AAAI*, pp. 1867–1873 (2016).
- [34] Matsubara, Y. and Sakurai, Y.: Dynamic Modeling and Forecasting of Time-evolving Data Streams, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019* (Teresesai, A., Kumar, V., Li, Y., Rosales, R., Terzi, E. and Karypis, G., eds.), ACM, pp. 458–468 (2019).
- [35] Beyazit, E., Alagurajah, J. and Wu, X.: Online Learning from Data Streams with Varying Feature Spaces, *AAAI/IAAI*, pp. 3232–3239 (2019).
- [36] Yoon, S., Shin, Y., Lee, J. and Lee, B. S.: Multiple Dynamic Outlier-Detection from a Data Stream by Exploiting Duality of Data and Queries, *SIGMOD*, pp. 2063–2075 (2021).
- [37] Yoon, S., Lee, J. and Lee, B. S.: Ultrafast Local Outlier Detection from a Data Stream with Stationary Region Skipping, *KDD*, pp. 1181–1191 (2020).
- [38] Zhong, Z., Yan, S., Li, Z., Tan, D., Yang, T. and Cui, B.: BurstSketch: Finding Bursts in Data Streams, *SIGMOD*, pp. 2375–2383 (2021).
- [39] Tizzoni, M., Bajardi, P., Poletto, C., Ramasco, J. J., Balcan, D., Gonçalves, B., Perra, N., Colizza, V. and Vespignani, A.: Real-time numerical forecast of global epidemic spreading: Case study of 2009 A/H1N1pdm, *BMC medicine*, Vol. 10, p. 165 (2012).
- [40] Venna, S. R., Tavanaei, A., Gottumukkala, R. N., Raghavan, V. V., Maida, A. S. and Nichols, S.: A Novel Data-Driven Model for Real-Time Influenza Forecasting, *IEEE Access*, Vol. 7, pp. 7691–7701 (2019).
- [41] Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G. and Cottrell, G. W.: A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction, *IJCAI*, pp. 2627–2633 (2017).
- [42] Chen, P., Liu, S., Shi, C., Hooi, B., Wang, B. and Cheng, X.: NeuCast: Seasonal Neural Forecast of Power Grid Time Series, *IJCAI*, pp. 3315–3321 (2018).
- [43] Islam, M. R., Muthiah, S., Adhikari, B., Prakash, B. A. and Ramakrishnan, N.: DeepDiffuse: Predicting the ‘Who’ and ‘When’ in Cascades, *ICDM*, pp. 1055–1060 (2018).
- [44] Ye, J., Sun, L., Du, B., Fu, Y., Tong, X. and Xiong, H.: Co-Prediction of Multiple Transportation Demands Based on Deep Spatio-Temporal Neural Network, *KDD*, pp. 305–313 (2019).
- [45] Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H. and Chawla, N. V.: A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data, *AAAI/IAAI*, pp. 1409–1416 (2019).
- [46] Flunkert, V., Salinas, D. and Gasthaus, J.: DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks, *CoRR*, Vol. abs/1704.04110 (2017).
- [47] Lea, C., Flynn, M. D., Vidal, R., Reiter, A. and Hager, G. D.: Temporal Convolutional Networks for Action Segmentation and Detection, *CVPR*, pp. 1003–1012 (2017).
- [48] Adhikari, B., Xu, X., Ramakrishnan, N. and Prakash, B. A.: EpiDeep: Exploiting Embeddings for Epidemic Forecasting, *KDD*, pp. 577–586 (2019).
- [49] Moré, J. J.: The Levenberg-Marquardt algorithm: Implementation and theory, *Numerical Analysis*, pp. 105–116 (1978).