情報処理学会論文誌　データベース　Vol.9　No.4　17–31 (Dec. 2016)

大規模疫病データのための非線形モデル解析

松原 靖子[1,†1,a)]　櫻井 保志[1,b)]　Willem G. van Panhuis[2,c)]　Christos Faloutsos[3,d)]

受付日 2016年6月9日, 採録日 2016年10月12日

概要：本論文では，大規模疫病データのための非線形モデル解析手法である Funnel について述べる．Funnel は，(*disease, location, time*) の三つ組で構成された d の疫病と l の州を含む，長さ n のテンソルデータが与えられたときに，そのテンソルの中から非線形パターンや外れ値を統合的に解析，要約，表現する．提案手法は (a) 疫病の季節性，ワクチンの出現，外部要因による疫病の発生等の重要なパターンを発見し，(b) パラメータのチューニングを必要とせず，(c) 計算量は入力データのサイズに対して線形である．さらに，本手法は (d) 麻疹やインフルエンザ等の疫病だけでなく，コンピュータウィルスの感染パターンについても同様に解析することができる．実データを用いた実験では，Funnel が大規模疫病テンソルの中から (P1) 病気の季節性，(P2) ワクチン効果，(P3) 地域性，(P4) 外因性による疫病パターン，(P5) データ入力エラー等の重要パターンを正確に発見することを確認した．

キーワード：疫病，テンソル解析，時系列データ，特徴自動抽出

Non-linear Mining of Spatially Coevolving Epidemics

Yasuko Matsubara[1,†1,a)]　Yasushi Sakurai[1,b)]　Willem G. van Panhuis[2,c)]
Christos Faloutsos[3,d)]

Received: June 9, 2016, Accepted: October 12, 2016

Abstract: Given a large collection of epidemiological data consisting of the count of d contagious diseases for l locations of duration n, how can we find patterns, rules and outliers? For example, the Project Tycho provides open access to the count infections for U.S. states from 1888 to 2013, for 56 contagious diseases (e.g., measles, influenza), which include missing values, possible recording errors, sudden spikes (or dives) of infections, etc. So how can we find a combined model, for all these diseases, locations, and time-ticks? In this paper, we present Funnel, a unifying analytical model for large scale epidemiological data, as well as a novel fitting algorithm, FunnelFit, which solves the above problem. Our method has the following properties: (a) *Sense-making*: it detects important patterns of epidemics, such as periodicities, the appearance of vaccines, external shock events, and more; (b) *Parameter-free*: our modeling framework frees the user from providing parameter values; (c) *Scalable*: FunnelFit is carefully designed to be linear on the input size; (d) *General*: our model is general and practical, which can be applied to various types of epidemics, including computer-virus propagation, as well as human diseases. Extensive experiments on real data demonstrate that FunnelFit does indeed discover important properties of epidemics: (P1) disease seasonality, e.g., influenza spikes in January, Lyme disease spikes in July and the absence of yearly periodicity for gonorrhea; (P2) disease reduction effect, e.g., the appearance of vaccines; (P3) local/state-level sensitivity, e.g., many measles cases in NY; (P4) external shock events, e.g., historical flu pandemics; (P5) detect incongruous values, i.e., data reporting errors.

Keywords: epidemics, Tensor analysis, time-series, automatic mining

[1] 熊本大学大学院先端科学研究部
Faculty of Advanced Science and Technology, Kumamoto University, Kumamoto 860–8555 Japan
[2] Department of Epidemiology, University of Pittsburgh, Pittsburgh, PA 15261, USA
[3] Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213-3891, USA

[†1] 現在，国立研究開発法人科学技術振興機構，さきがけ
Presently with JST, PRESTO
[a)] yasuko@cs.kumamoto-u.ac.jp
[b)] yasushi@cs.kumamoto-u.ac.jp
[c)] wav10@pitt.edu
[d)] christos@cs.cmu.edu

1. まえがき

大規模な疫病データの中から有用なパターンを発見し，感染症の性質や法則を見つけ出すことは，医療分野や公衆衛生の研究，および政策や社会活動等において非常に重要な問題である．本研究では，大規模疫病データを対象とし，重要なパターンを自動抽出するための非線形モデル解析手法として FUNNEL を提案する [24] *1．より具体的には，以下の問題を扱う．

d 種の疫病，l カ所の地域，n の期間で構成される大規模疫病データ集合 $\mathcal{X}$ が与えられたとき，以下の重要なパターンを自動抽出する．

- 季節性や感染率等の各疫病の基本的な特徴
- 突発的な流行の外部要因
- 入力エラー等の外れ値

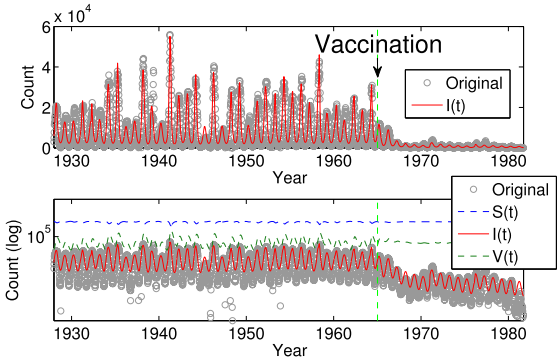
具体例．本研究では，疫病に関するデータとして *Tycho* [43] を扱う．*Tycho* は，アメリカ合衆国における 125 年間にわたる 56 種の疫病の患者の報告件数で構成される*2．**図 1** は，FUNNEL による解析結果の例を示している．具体的には，図 1 (a) において灰色の丸印は 1928 年から 1982 年までの，麻疹の感染者数を示しており，赤の実線は提案手法の学習結果を示している．シーケンスは年の周期性があり，そして数年ごとに大規模な感染（1941 年，1958 年）と小さな波（1940 年と 1947 年）が現れており [29]，これはスキップ現象として知られている [40]．また，感染者数は 1965 年において急激に減少しており，これはワクチン接種が 1963 年に開始されたことによるものである．考案手法とモデルはこのようなパターンを高精度で表現することができる．図 1 (b) は各州の麻疹に関する感受性保持者の潜在人口を示している．図 1 (c) は各疫病の季節性の強さ（半径）とそのピークの時期（角度）を示したものである．たとえば，1 月から 2 月にかけてインフルエンザの流行のピークがあり，麻疹のような小児呼吸器疾患は春，ライム病のようなダニ媒介疾患は夏がピークとなっている．また，淋病のような性感染症（STD）には周期性がないことが分かる．提案手法は，事前知識やパラメータ設定を要せず，すなわち，ユーザの介入なしに，様々な疫病に関する重要な特徴を自動的に抽出することができる．

本論文の貢献．FUNNEL は以下の特長がある．

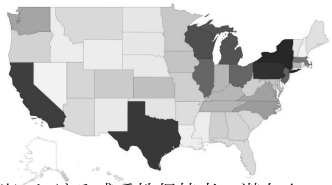
(1) 提案モデルは，麻疹，インフルエンザ，天然痘をはじめとする様々な感染症の振舞いを表現し，図 1 にあげられるような季節性，ワクチン効果，外部要因による突発的流行等の重要な時系列パターンを自動抽出する．

(2) 計算コストは入力データサイズに対し線形である．

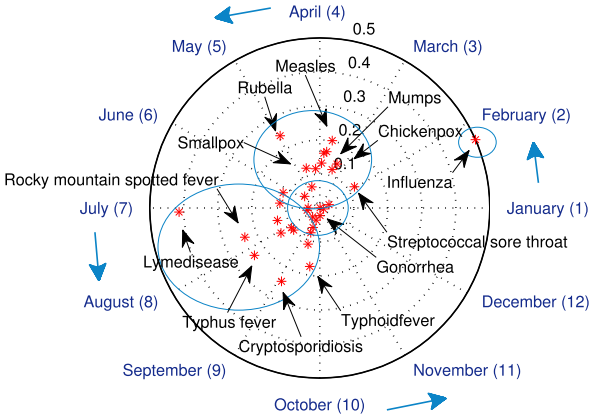
*1 http://www.cs.kumamoto-u.ac.jp/~yasuko/software.html
*2 Project Tycho at University of Pittsburgh: http://www.tycho.pitt.edu/



(a) FUNNELFIT の学習結果 (麻疹)



(b) 各州における感受性保持者の潜在人口 (麻疹)



(c) 各疫病の季節性: 半径は強さ，角度はピークを示す．

図 1　FUNNELFIT を用いたパターン発見と出力結果
Fig. 1　Modeling power of FUNNELFIT.

(3) 提案モデルは，SIRS モデル等の既存の疫病モデルを一般化すると同時に，コンピュータウィルスの感染パターン等の様々な種類の非線形伝染過程を柔軟に表現することができる．

2. 関連研究

関連研究は以下の 2 つに分類される．

疫学と医療データ解析．疫学分野において，SI モデル（susceptible-infected model）に代表される様々な疫病感染モデルが提案されている [2]．文献 [9], [10] では麻疹をはじめとする疫病の拡散過程を分析し，Stone ら [40] は季節性をともなう疫病における毎年の流行を予測するための新たな閾値を発見した．Van Panhuis ら [43] は，1888 年から 2013 年にかけてのアメリカ合衆国内のすべての疫病患者の報告件数を調査した．

時系列データ解析．時系列データの解析に関する研究は様々な分野で進められている [4], [7], [16], [28], [34], [35]．時系列シーケンスのための類似探索，パターン発見は重要な課題

表 1　既存手法との比較
Table 1　Capabilities of approaches.

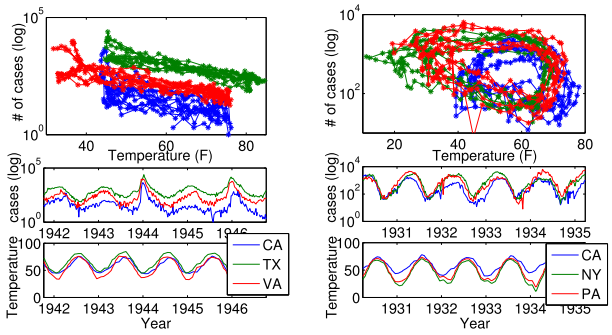
	SIRS	AR/*PLiF*	PARAFAC	FUNNELFIT
情報圧縮	√	√	√	√
ドメイン知識	√			√
欠損値	√			√
周期性		√		√
予測		√		√
自動化				√

である [6], [8], [13], [23], [25], [27], [33], [36], [37], [41], [44]．自己回帰モデル（AR: autoregressive model），線形動的システム（LDS: linear dynamical systems），カルマンフィルタ（KF: Kalman filters）は代表的な技術であり，これらに基づく時系列の解析と予測手法が数多く提案されている [11], [15], [42]．時系列ビッグデータの研究として，TriMine [21] は大規模複合時系列イベントデータのための高速な予測手法であり，文献 [18] では多次元時系列シーケンスのための特徴自動抽出手法を提案した．CompCube [20] は大規模複合テンソルデータのための非線形解析手法であり，RegimeCast [17] はデータストリームのリアルタイム将来予測に焦点を当てている．Rakthanmanon らは文献 [32] において，兆単位（“trillions”）の時系列シーケンスを対象としたDTW の類似探索問題を扱っている．疫病の拡散過程やソーシャルネットワーク上の情報伝搬に関する時系列データ解析の研究も活発に行われている [12], [14], [22], [30], [31]．FUNNEL [24] は大規模疫病テンソルデータのための非線形モデル解析手法であり，文献 [19] は Web 上におけるユーザの活動と，自然界における生態系モデルの類似性を分析した先駆的な取り組みである．

関連研究と本研究の位置づけ．表 1 は，既存手法と FUNNEL の能力の比較である．(a) SI，SIR，SIRS モデル等の既存の疫病モデルは疫病時系列データを圧縮し，非線形性を表現する能力を有するが，周期性，ワクチン効果やその他の複雑な特徴を表現できず，予測の能力も不十分である．(b) 自己回帰モデル（AR: autoregressive model）や *PLiF* [15] は圧縮と予測の能力を有しているが，非線形性を有する時系列パターンを表現できない．さらにこれらの手法はドメイン知識を考慮しないため，より高度な疫病パターンの抽出をすることができない．(c) 本研究で扱う疫病データはテンソルとして表現できる．PARAFAC はテンソルの圧縮能力を有するが，外れ値や周期性，ドメイン知識を表現できず，予測の能力も有していない．より重要な点として，上記の既存手法は基本的にすべて，パラメータの設定やチューニングが必要である．

3.　提案モデル

本章では，提案モデルである FUNNEL について述べる．



(a) インフルエンザ（CA，TX，VA）　(b) 麻疹（CA，NY，PA）
図 2　各州における気温と疫病感染者数の関係性
Fig. 2　The air temperature vs. # of cases.

3.1　FUNNEL の概要

本研究で扱う *Tycho* データは，1888 年から現在にかけてのアメリカ合衆国 50 州における 56 種の疫病患者の報告件数（計 87,950,807 件）で構成される．本データは (*disease, location, timestamp*) の三つ組で表現され，それぞれ，d 種の疫病，l カ所の地域/州，n の期間（1 週間単位）から構成される．この疫病データは，3 階のテンソル $\mathcal{X} \in \mathbb{N}^{d \times l \times n}$ として表現することができ，$\mathcal{X}$ の要素 $x_{ij}(t)$ は時刻 t において i 番目の疫病（disease）が j 番目の地域/州（location）が出現した頻度を示す．たとえば，(‘measles’, ‘PA’, ‘April 1-7, 1931’; 4740) の場合，麻疹（measles）がペンシルベニア州（PA）で 1931 年 4 月 1 日から 4 月 7 日の間に 4740 件報告されたことを表す．ここで，i 番目の疫病の j 番目の州におけるシーケンス $\boldsymbol{x}_{ij} = \{x_{ij}(t)\}_{t=1}^{n}$ を，ローカル（州）の疫病シーケンスと呼ぶ．同様に，各州のシーケンスの合計 $\bar{\boldsymbol{x}}_i = \{\bar{x}_i(t)\}_{t=1}^{n}$ を，グローバル（国）の疫病シーケンスと呼び，$\bar{x}_i(t)$ は i 番目の疫病の時刻 t における出現頻度の国内の総数（つまり，$\bar{x}_i(t) = \sum_{j=1}^{l} x_{ij}(t)$）とする．

疫病データの考察． ここでは，疫病データのモデル化にあたり重要となる要素について事前考察を行う．**図 2** は，インフルエンザと麻疹の 5 年間の感染者数の推移と気温の関係を示している[*3,*4]．図 2 (a) のように，インフルエンザは気温と強い負の相関があり，気温の低い季節に活発になる．一方，図 2 (b) のように，麻疹は位相のずれがあり，春に活発化する疫病であることが分かる．同様に，図 1 (c) で示したように，疫病には季節性をともなう場合が多く，たとえば麻疹やおたふくかぜに代表される小児呼吸器疾患は春に，ライム病のようにダニ媒介疾患は夏がピークとなっている．

知見 1（疫病の季節性）　疫病の多くは年単位の周期性をともない，気温や季節と強い相関を持つ．

ワクチンの導入，抗生物質の使用や公衆衛生の管理によ

[*3] CA: California, TX: Texas, VA: Virginia, NY: New York, PA: Pennsylvania.
[*4] National climate data center: http://www.ncdc.noaa.gov/cag/

り，多くの疫病が，数十年の間に大幅に減少している．たとえば，図 1 (a) で示したように，麻疹の感染者数はワクチン接種が 1963 年に開始されて以来，大幅な減少傾向にある．

知見 2（疫病減少効果） 疫病の多くは，ワクチン，抗生物質，公衆衛生管理により，大幅に減少あるいは撲滅している．

次に，各地域における傾向について述べる．図 2 に示すように，一般に，各州の感染者数は高い相関を持つが，同時に，患者数の割合は異なる．これは，各州での感受性保持者の潜在人口が異なることに起因する．たとえば，麻疹の主な感染対象は小児であるため，小児が多い地域（州）には，より多くの麻疹患者が見込まれる．図 1 (b) では，ニューヨーク（NY），ペンシルベニア（PA），カリフォルニア（CA），テキサス（TX）に，多くの感染対象者（つまり小児）がいることが分かる．

知見 3（地域性） 各州における感染者数の推移は正の相関を持つが，一方で，感受性保持者の潜在人口が異なる．

最後に，外因性による傾向について述べる． 図 1 (a) の 1941 年と 1958 年において，図 2 (a) の 1944 年と 1946 年において，それぞれ，麻疹とインフルエンザの感染者数が大幅に増加している．

知見 4（外部要因による突発的流行） 疫病の多くには，数年に一度現れる，外部要因に由来する大規模な流行パターンが存在する．

実データの中には，予期しないエラー値や入力ミスによる突発的な外れ値が存在する．

知見 5（入力エラー） 疫病の感染パターンに合致しない外れ値や入力エラーが存在する．

本研究では，大規模疫病データを対象とし，上記の重要な要素を表現する統合モデルとして，Funnel を提案する．具体的には，次の 5 つの要素をすべて表現する．

- (**P1**)：季節性
- (**P2**)：疫病減少効果
- (**P3**)：地域性
- (**P4**)：外部要因による突発的流行
- (**P5**)：外れ値，入力エラー

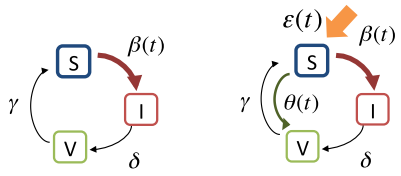
次節では，提案モデルの詳細を示す．まず，(a) 単一の疫病シーケンスの場合（たとえば，ニューヨーク州（NY）の麻疹の感染者数等）についてのモデルを述べ，次に発展として，(b) 複数シーケンスの場合（つまり，d 種の疫病，l カ所の地域に対する個々のパターン）のモデルについて述べる．

3.2 FUNNEL — 単一の疫病シーケンスの場合

まず最も簡単な場合として，単一の疫病シーケンスのモデル化について述べる．

表 2　主な記号と定義
Table 2　Symbols and definitions.

記号	定義
d	疫病の総数
l	州（地域）の総数
n	時系列の長さ
$\mathcal{X}$	3 階の疫病テンソル（$\mathcal{X} \in \mathbb{N}^{d \times l \times n}$）
$\boldsymbol{x}_{ij}$	i 番目の疫病の j 番目の州における ローカル疫病シーケンス
$\bar{\boldsymbol{x}}_i$	i 番目の疫病のグローバル 疫病シーケンス（国内総数）
$S_{ij}(t)$	i 番目の疫病の j 番目の州の時刻 t における 感受性保持者
$I_{ij}(t)$	i 番目の疫病の j 番目の州の時刻 t における 感染者
$V_{ij}(t)$	i 番目の疫病の j 番目の州の時刻 t における 免疫保持者
$\mathbf{B}$	基本行列（$d \times 6$）　i.e., $\mathbf{B} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d\}$
$\mathbf{R}$	減少行列（$d \times 2$）　i.e., $\mathbf{R} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_d\}$
$\mathbf{N}$	地域行列（$d \times l$）　i.e., $\mathbf{N} = \{N_{ij}\}_{i,j=1}^{d,l}$
$\mathcal{E}$	外部ショックテンソル i.e., $\mathcal{E} = \{\mathbf{E}^{(\mathrm{D})}, \mathbf{E}^{(\mathrm{T})}, \mathbf{E}^{(\mathrm{S})}\}$
$\mathcal{M}$	入力エラーテンソル i.e., $\mathcal{M} = \{m_{ij}(t)\}_{i,j,t=1}^{d,l,n}$
$\mathcal{F}$	Funnel の全パラメータ集合 i.e., $\mathcal{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathcal{E}, \mathcal{M}\}$



(a) Funnel-base　(b) Funnel-RE

図 3　Funnel の状態遷移の様子
Fig. 3　Funnel diagrams.

3.2.1 基本モデル — Funnel-base

提案モデルは以下の 3 つの状態から構成される．

- Susceptible（感受性保持者）：現時点では疫病に感染していないが，近隣の感染者から伝染する可能性のある状態
- Infected（感染者）：すでに疫病に感染しており，他者へ伝染させる可能性のある状態
- Vigilant（免疫保持者）：疫病から回復した状態，あるいはワクチン接種等により疫病に感染する可能性がない状態

図 3 (a) は，基本モデルである Funnel-base の状態遷移の様子を示す．ここで，$\beta(t)$ は感染者と感受性保持者の間の疫病の感染率を，δ は感染者の回復率を示す．γ は，免疫力の減少率を示す[*5]．さらに，先述の要素 (**P1**) を表現するために，本研究では，感染率 $\beta(t)$ は時刻 t における周期性をともなう関数として表現する．

[*5] γ は，出生率および死亡率も表現する．

モデル 1（**Funnel-**BASE）　$S(t)$，$I(t)$，$V(t)$ を時刻 t における感受性保持者，感染者，免疫保持者の総数とする．FUNNEL-BASE は，次の式で構成される．

$$\begin{aligned} S(t+1) &= S(t) - \beta(t)S(t)I(t) + \gamma V(t) \\ I(t+1) &= I(t) + \beta(t)S(t)I(t) - \delta I(t) \\ V(t+1) &= V(t) + \delta I(t) - \gamma V(t) \end{aligned} \quad (1)$$

ここで，$\beta(t) = \beta_0 \cdot \left(1 + P_a \cdot cos\left(\frac{2\pi}{P_p}(t + P_s)\right)\right)$，$P_p = 52$ とし[*6]，さらに，$N = S(t) + I(t) + V(t)$，$S(1) = N - 1$，$I(1) = 1$，$V(1) = 0$ とする．

まとめると，FUNNEL-BASE は次のパラメータ集合で構成される：$\boldsymbol{b} = \{N, \beta_0, \delta, \gamma, P_a, P_s\}$.

- N：疫病の潜在的人口．
- β_0：感染率の年間平均値．
- δ：疫病の回復率．
- γ：免疫力の減少率．
- P_a：季節性の強さ，振幅．
- P_s：季節性の位相．

3.2.2　疫病減少効果 — Funnel-R

2 つ目の要素 (**P2**) のために，モデルを次のように拡張する．

モデル 2（**Funnel-**R）　$\theta(t)$ は疫病減少効果の強さを示す．

$$\begin{aligned} S(t+1) &= S(t) - \beta(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t) \\ I(t+1) &= I(t) + \beta(t)S(t)I(t) - \delta I(t) \\ V(t+1) &= V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \end{aligned} \quad (2)$$

ここで，疫病減少効果の開始時刻を t_θ とし，次のように定義する：$\theta(t) = \begin{cases} 0 & (t < t_\theta) \\ \theta_0 & (t \geq t_\theta) \end{cases}$

このモデルは，FUNNEL-BASE に疫病減少効果の要素 $\theta(t)$ を加えたものである．これは，図 3 (b) に示すように，感受性保持者が，ワクチン接種や抗生物質投与等により，直接免疫保持者の状態に移行する．基本パラメータ $\boldsymbol{b}$ に加え，FUNNEL-R は次のパラメータを要する：$\boldsymbol{r} = \{t_\theta, \theta_0\}$.

- t_θ：疫病減少効果の開始時刻．
- θ_0：疫病減少効果の強さ，拡散率．

3.2.3　外部ショック — Funnel-RE

次に，3 つ目の要素：(**P4**) について述べる．疫病の多くは外部要因によって突発的に流行することがある．たとえば，豚インフルエンザの流行が発生した場合には，過去の状況よりも多くの感染者が見込まれることになる．

この現象を表現するために，本研究では一時的な感受性変化率 $\epsilon(t)$ を導入する．具体的には，図 3 (b) に示すように，時刻 t において外部ショックによるイベントが起きた場合，感受性保持者 $S(t)$ の感染者数が一時的に変化する．これにより，大規模な感染拡大パターンを表現する．

モデル 3（**Funnel-**RE）　提案モデルは次の式で表現される：

$$\begin{aligned} S(t+1) &= S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t) \\ I(t+1) &= I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t) \\ V(t+1) &= V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \end{aligned} \quad (3)$$

ここで，$\epsilon(t)$ は，一時的な感受性変化率を表す：

$$\epsilon(t) = 1 + \sum_{i=1}^{k} f(t; \boldsymbol{e}_i^{(\boldsymbol{T})}),$$

$$f(t; \boldsymbol{e}^{(\boldsymbol{T})}) = \begin{cases} \epsilon_0 & (t_\mu - t_\sigma < t < t_\mu + t_\sigma) \\ 0 & (else) \end{cases}$$

k は外部ショックの回数を示す（$k = 0$ の場合には $\epsilon(t) = 1$）．

各外部ショックは次のパラメータで構成される：$\boldsymbol{e}^{(\boldsymbol{T})} = \{t_\mu, t_\sigma, \epsilon_0\}$.

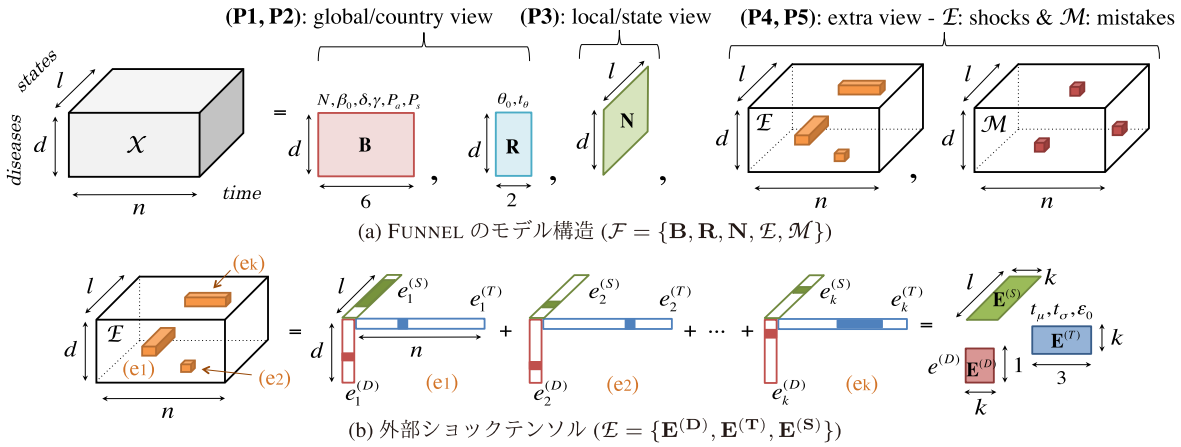
- t_μ：外部ショックイベントの中心時刻．
- t_σ：イベントの長さ．
- ϵ_0：外部ショック効果の強さ．

3.3　FUNNEL — 複数シーケンスの場合

前節では，単一の疫病シーケンスについての FUNNEL の振舞いについて述べた．ここからは，疫病テンソル $\mathcal{X}$ が与えられたうえでの d 種の疫病，l カ所の地域の個々の振舞いを表現するモデルについて考える．まず，最も簡易的な解決法としてあげられるのが，テンソル $\mathcal{X}$ を $(d \times l)$ 個の長さ n のシーケンス集合：$\{\boldsymbol{x}_{ij}\}_{i,j=1}^{d,l}$（つまり，ローカル（州）の疫病シーケンス）と見なし，各シーケンスに対し個別にモデルパラメータ集合：$\{\boldsymbol{b}, \boldsymbol{r}, \boldsymbol{e}^{(\boldsymbol{T})}\}$ を推定する場合である．しかしながら，これらのローカルの疫病シーケンスのペア (*disease, state*) の中には，たとえば，アラスカ州におけるライム病のように，感染者数が非常に少ないスパースな時系列データを含む場合が多く，直接的に個々のモデルパラメータを推定することが困難である．さらに，上記の方法では，グローバル（国）の疫病の振舞いをモデル化することができない．この問題を解決するために，本研究では，l カ所のすべての州で，一部のパラメータを共有する手法を用いる．

FUNNEL — パラメータ集合．本研究の目的は，大規模疫病データ $\mathcal{X} \in \mathbb{N}^{d \times l \times n}$ の中から重要なトレンドや特徴を抽出することである．**図 4** は提案モデルの概要を示す．疫病テンソル $\mathcal{X}$ が与えられたとき，提案手法は 5 つの重要な要素を発見する：(**P1**) $\mathbf{B}$：季節性等の疫病の基本的な振舞い，(**P2**) $\mathbf{R}$：疫病減少効果，(**P3**) $\mathbf{N}$：地域性，(**P4**) $\mathcal{E}$：外部ショックイベント集合，(**P5**) $\mathcal{M}$：入力エラー．ここで，(**P1**) と (**P2**) は，グローバル（国）の振舞いを表現するパラメータ集合，(**P3**) はローカル（州）の個々のパターン，そして，(**P4**) と (**P5**) は $\mathcal{X}$ の外部要因に関する

[*6] 1 年間は 52 週で構成される．



(a) FUNNEL のモデル構造 ($\mathcal{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathcal{E}, \mathcal{M}\}$)

(b) 外部ショックテンソル ($\mathcal{E} = \{\mathbf{E^{(D)}}, \mathbf{E^{(T)}}, \mathbf{E^{(S)}}\}$)

図 4　FUNNEL の概要．提案手法は疫病テンソル $\mathcal{X}$ の中から重要なパターンを自動抽出する
Fig. 4　Illustration of FUNNEL structure: we extract the important behavior of epidemics from $\mathcal{X}$.

パラメータ集合である．

定義 1 (FUNNEL のパラメータ集合)　$\mathcal{F}$ を疫病テンソル $\mathcal{X}$ を表現する全パラメータ集合 $\mathcal{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathcal{E}, \mathcal{M}\}$ とする．

次に，各パラメータの詳細を述べる．

(P1), (P2) グローバルパラメータ集合. 次のパラメータ集合は，グローバルな振舞いを表現し，l カ所の州で共有する．

定義 2 (基本行列 $\mathbf{B}$ $(d \times 6)$)　$\mathbf{B}$ を d 個の疫病に関する基本パラメータ集合 $\mathbf{B} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d\}$ とする．ここで，$\boldsymbol{b}_i$ は i 番目の疫病のパラメータを表す．

たとえば，麻疹ウィルスの感染率および回復率はニューヨーク州とフロリダ州で同じである．同様に，麻疹のワクチンが導入される時期は，どの州においても同じであり，導入されたワクチンの効果が出始める時期も一致する．

定義 3 (減少行列 $\mathbf{R}$ $(d \times 2)$)　$\mathbf{R}$ を疫病減少効果に関するパラメータ集合 $\mathbf{R} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_d\}$ とする．ここで，$\boldsymbol{r}_i$ は，i 番目の疫病のパラメータを表す．

(P3) ローカルパラメータ集合. 次の課題は，テンソル $\mathcal{X}$ の中の地域固有のパターンの表現方法である．ニューヨーク州とフロリダ州それぞれの麻疹の感染者数の推移に変化をもたらす原因はいったいなんであろうか．各州における麻疹のウィルスに地域性や大きな個体差はない．唯一の違いは，感受性保持者の潜在人口の差である．そこで本研究では，l カ所の州でグローバルパラメータ集合を共有し，疫病の潜在的人口 N_{ij} のみを i 番目の疫病，j 番目の州ごとに導入する．具体的には，モデル 3 において，次の式を満たす：$N_{ij} = S_{ij}(t) + I_{ij}(t) + V_{ij}(t)$．たとえば，小児呼吸器疾患である麻疹について考えると，ニューヨーク州にはフロリダ州よりも多くの小児がいるため（つまり，潜在的な麻疹感染対象者が多いため），より多くの感染者が見込まれる．

定義 4 (地域行列 $\mathbf{N}$ $(d \times l)$)　$\mathbf{N}$ を d 種の疫病，l カ所の地域における潜在的人口を表すパラメータ：$\mathbf{N} = \{N_{ij}\}_{i,j=1}^{d,l}$ とする．ここで，N_{ij} は，i 番目の疫病の j 番目の州における潜在的人口を表す．

(P4) 外部ショック.　本研究では，突発的に発生する疫病の流行イベントを 3 つの要素：(*disease, state, time*) として表現する．たとえば，1946 年に国家規模で起きたインフルエンザの歴史的大流行の場合は，(e1) “influenza, country-wide, 1946” として表現される．同様にして，2007 年にユタ州で起きた地域規模のクリプトスポリジウム症の流行は，(e2) “cryptosporidiosis, Utah, 2007” として表現される．これらの複数のイベントを表現するために，新たにパラメータ集合として外部ショックテンソル $\mathcal{E}$ を導入する．図 4 (b) に示すように，$\mathcal{E}$ は，k 個のイベント集合で構成される．外部ショックテンソル $\mathcal{E}$ は，3 つの要素行列 $\{\mathbf{E^{(D)}}, \mathbf{E^{(S)}}, \mathbf{E^{(T)}}\}$ として圧縮することもできる．さらに，単一の外部ショックイベントは，三つ組のベクトル集合 $\{\boldsymbol{e}^{(D)}, \boldsymbol{e}^{(S)}, \boldsymbol{e}^{(T)}\}$ として次のように表現される．

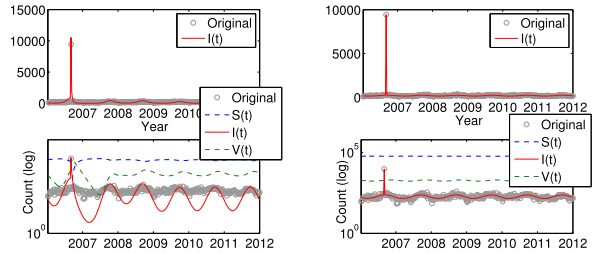
- 疫病ベクトル $\boldsymbol{e}^{(D)}$：疫病の ID
- 地域ベクトル $\boldsymbol{e}^{(S)}$：各地域のショックの強さ
- 時間ベクトル $\boldsymbol{e}^{(T)}$：ショックの発生期間

ここで，3.2.3 項で示したとおり，$\boldsymbol{e}^{(T)}$ はグローバルなパラメータであり，$\boldsymbol{e}^{(S)}$ はローカルなパラメータ：$\boldsymbol{e}^{(S)} = \{e_j^{(S)}\}_{j=1}^{l}$ である．これは，モデル 3 におけるパラメータ ϵ_0 の l カ所の州ごとの変化率を表す．より具体的には，j 番目の州での外部ショックの強さを $\epsilon_0 \cdot e_j^{(S)}$ とする．

定義 5 (外部ショックテンソル $\mathcal{E}$)　$\mathcal{E}$ を k 個の外部ショックイベントで構成される 3 階のテンソル：$\mathcal{E} = \{\mathbf{E^{(D)}}, \mathbf{E^{(S)}}, \mathbf{E^{(T)}}\}$ とする．ここで，三つ組の行列は，疫病，州，時間の各要素を表現するパラメータである．

(P5) 入力エラー.　実データには，多くのエラー値や報告の誤りが含まれる．提案モデルは，これらの入力エラーを外れ値として取り除く必要がある．

定義 6 (入力エラーテンソル $\mathcal{M}$)　$\mathcal{M}$ を入力エラーに関



(a) 外部ショックとしての学習結果　(b) 入力エラーとしての学習結果

図 5　ジアルジア症に対する学習結果の比較

Fig. 5　External shock vs. mistake for giardiasis.

する 3 階のテンソルとする．$\mathcal{M}$ の各要素 $m_{ij}(t)$ は，i 番目の疫病，j 番目の州，時刻 t の値を表す．
ここで，$\mathcal{M}$ は非常にスパースであり，多くの場合において，$m_{ij}(t)=0$ である．

図 5 は，ジアルジア症に対する 2 種類の学習結果の様子を示している．灰色の丸印はオリジナルデータ，赤線はフィッティング結果を示す．図に示すように，ジアルジア症は 2006 年において大きな外れ値（およそ 10,000）がある．図 5 (a) は外部ショックとして学習した結果であり，図 5 (b) は入力エラーとして学習した結果である．(a) において，近隣の点が外れ値に大きく影響されている一方で，(b) では外れ値が独立した点として認識され，より正しく学習できていることが分かる．

4. 最適化アルゴリズム

本章では，モデルの学習アルゴリズムである FunnelFit について述べる．提案アルゴリズムの目的は，与えられた疫病テンソル $\mathcal{X}$ に対し，重要なパターンを自動抽出することである．

問題 1　(*disease, state, time*) の三つ組で構成される疫病テンソル $\mathcal{X}$ が与えられたとき，$\mathcal{X}$ を表現する最適なモデルパラメータ集合 $\mathcal{F}=\{\mathbf{B},\mathbf{R},\mathbf{N},\mathcal{E},\mathcal{M}\}$ を発見する．

ここでの重要な課題は，(a) テンソル $\mathcal{X}$ の中の重要なパターンを表現するための最適なパラメータをどのように推定するか，(b) 外部ショックの最適な個数 k をどのように決定するか，そして (c) データ $\mathcal{X}$ の中に含まれる 入力エラーをどのように取り除くかである．

4.1 モデル化とデータ圧縮

本研究では，大規模疫病テンソルデータ $\mathcal{X}$ を適切に表現・モデル化するために，最小記述長（MDL: minimum description length）に基づく新たな符号化スキームを導入する．直感的に，データがより圧縮できれば，より良いモデルであると見なす．データが与えられたときのモデルの良さは次の式で表現できる：$Cost_T = Cost(\mathcal{M}) + Cost(\mathcal{X}|\mathcal{M})$．ここで，$Cost(\mathcal{M})$ はモデル $\mathcal{M}$ を表現するためのコストを示し，$Cost(\mathcal{X}|\mathcal{M})$ は，$\mathcal{M}$ が与えられたときのデータ $\mathcal{X}$ の符号化のコストを示す．

モデル表現コスト． Funnel のパラメータ表現コストは以下の要素から構成される．

- 疫病の総数 d，州の総数 l，および，時系列の長さ n：$\log^*(d)+\log^*(l)+\log^*(n)$ ビット[*7]．
- 基本行列 ($\mathbf{B}$)，減少行列 ($\mathbf{R}$)，地域行列 ($\mathbf{N}$) にそれぞれ $d\times 6$, $d\times 2$, $d\times l$ のパラメータを要する．まとめると，$Cost_M(\mathbf{B})+Cost_M(\mathbf{R})+Cost_M(\mathbf{N}) = c_F\cdot d(6+2+l)$．ここで，$c_F$ は浮動小数点のコストを示す[*8]．

同様に，外部ショックテンソル $\mathcal{E}=\{\mathbf{E}^{(\mathbf{D})},\mathbf{E}^{(\mathbf{S})},\mathbf{E}^{(\mathbf{T})}\}$) は次の要素で構成される．

- 外部ショックの個数 k：$\log^*(k)$ ビット．
- 外部ショック（疫病）行列 $\mathbf{E}^{(\mathbf{D})}$：$k\log(d)$．
- 外部ショック（時間）行列 $\mathbf{E}^{(\mathbf{T})}$ の中の各要素 $\boldsymbol{e}^{(\boldsymbol{T})}=\{t_\mu,t_\sigma,\epsilon_0\}$：$\log(n)$，$\log(n)$，$c_F$．
- 外部ショック（地域）行列 $\mathbf{E}^{(\mathbf{S})}$：$c_F\cdot kl$．

まとめると，外部ショックテンソル $\mathcal{E}$ のモデルコストは次のようになる．$Cost_M(\mathcal{E}) = \log^*(k)+k\big(\log(d)+2\log(n)+c_F\cdot(1+l)\big)$．入力エラーテンソル $\mathcal{M}$ は以下で構成される．

- テンソル $\mathcal{M}$ 内の非ゼロ要素の個数：$\log^*(|\mathcal{M}|)$．
- 非ゼロ要素の位置とその値 $m_{ij}(t)$：$\log(d)$，$\log(l)$，$\log(n)$，$\log^*(m_{ij}(t))$．

したがって，$Cost_M(\mathcal{M}) = \log^*(|\mathcal{M}|)+\sum_{m_{ij}(t)>0}^{|\mathcal{M}|}(\log(d)+\log(l)+\log(n)+\log^*(m_{ij}(t)))$．ここで $|\mathcal{M}|$ は $\mathcal{M}$ 内の非ゼロ要素の個数を示す．

疫病データの符号化コスト．　ハフマン符号を用いた情報圧縮では，モデルパラメータ集合 $\mathcal{F}$ が与えられた際の $\mathcal{X}$ の符号化コストを負の対数尤度を用いて次のように表現することができる [3]：$Cost_C(\mathcal{X}|\mathcal{F}) = \sum_{i,j,t=1}^{d,l,n}\log_2 p^{-1}_{Gauss(\mu,\sigma^2)}(x_{ij}(t)-m_{ij}(t)-I_{ij}(t))$．ここで，符号化にはガウス分布を用い，$x_{ij}(t)$，$m_{ij}(t)$ は，それぞれ，疫病テンソル $\mathcal{X}$，入力エラーテンソルの要素の値を示す．$I_{ij}(t)$ は，モデル 3 で示したように，感染者数の推定値を示す．さらに，μ，σ^2 はそれぞれ，オリジナルデータの値と，推定値の間の平均，分散を示す[*9]．

符号化コスト関数． まとめると，モデルパラメータ集合 $\mathcal{F}$ が与えられたときの $\mathcal{X}$ の符号長は次のように表現される．

$$\begin{aligned} Cost_T(\mathcal{X};\mathcal{F}) &= \log^*(d)+\log^*(l)+\log^*(n)\\ &+Cost_M(\mathbf{B})+Cost_M(\mathbf{R})+Cost_M(\mathbf{N})\\ &+Cost_M(\mathcal{E})+Cost_M(\mathcal{M})+Cost_C(\mathcal{X}|\mathcal{F}) \end{aligned} \tag{4}$$

したがって本論文の次の目標は，上記のコスト関数を最小化することである．

[*7] ここで，$\log^*$ は整数のユニバーサル符号長を表す．
[*8] 本論文では 4×8 ビットとする．
[*9] ここで，μ，σ^2 は $2c_F$ ビットを要するが，これらのコストは定数であるため，モデル推定の際には除外することができる．

Algorithm 1 FunnelFit ($\mathcal{X}$)

1: **Input:** Tensor $\mathcal{X}$ ($d \times l \times n$)
2: **Output:** Complete set of parameters,
i.e., $\mathcal{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathcal{E}, \mathcal{M}\}$
3: /* Parameter fitting for global-level sequences */
4: $\{\mathcal{F}_\mathcal{G}\}$ =GlobalFit ($\mathcal{X}$);
5: /* Parameter fitting for local-level sequences */
6: $\{\mathcal{F}_\mathcal{L}\}$ =LocalFit ($\mathcal{X}, \mathcal{F}_\mathcal{G}$);
7: **return** $\mathcal{F} = \{\mathcal{F}_\mathcal{G}, \mathcal{F}_\mathcal{L}\}$;

4.2　多階層最適化アルゴリズム

ここまでは，パラメータ集合の候補解 $\mathcal{F}$ が与えられたうえでテンソル $\mathcal{X}$ を表現するための方法について述べた．次の課題は，どのようにしてすべてのパラメータ集合 $\mathcal{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathcal{E}, \mathcal{M}\}$ を最適化するかである．

図 3 で示したように，Funnel は，複数のパラメータ集合から構成され，それらは，疫病テンソル $\mathcal{X}$ の中のグローバル（国）かローカル（州）のいずれかのパターンを表現する．たとえば，基本行列 $\mathbf{B}$ と減少行列 $\mathbf{R}$ は，各疫病のグローバル（国）レベルの振舞いを表現し，一方で，地域行列 $\mathbf{N}$ は，ローカル（州）なトレンドを表現する．また，外部ショックテンソル $\mathcal{E} = \{\mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{S})}, \mathbf{E}^{(\mathbf{T})}\}$) は，グローバル ($\mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{T})}$)，ローカル ($\mathbf{E}^{(\mathbf{S})}$) のパラメータとして分解できる．同様にして，入力エラーテンソル $\mathcal{M}$ は，$\{\mathbf{M}^{(\mathbf{D})}, \mathbf{M}^{(\mathbf{S})}, \mathbf{M}^{(\mathbf{T})}\}$ として分解できる．そこで本研究では，上記の性質を用いて，より効率的かつ効果的にパラメータの推定を行う手法として，多階層最適化アルゴリズムを提案する．より具体的には，全パラメータ集合 $\mathcal{F}$ を，グローバル，ローカルの 2 つの部分パラメータ集合 $\mathcal{F}_\mathcal{G}$，$\mathcal{F}_\mathcal{L}$ に分解しそれらのパラメータ集合を個別に学習し，高速にモデル推定を行う．

提案アルゴリズムは次の 2 つのステップで構成される．

- GlobalFit：グローバル（国）の疫病シーケンス $\{\bar{\boldsymbol{x}}_i\}_{i=1}^{d}$ に対し，最適なモデルパラメータ $\mathcal{F}_\mathcal{G} = \{\mathbf{B}, \mathbf{R}, \mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{T})}, \mathbf{M}^{(\mathbf{D})}, \mathbf{M}^{(\mathbf{T})}\}$ を推定する．
- LocalFit：ローカル（州）の疫病シーケンス $\{\boldsymbol{x}_{ij}\}_{i,j=1}^{d,l}$ に対し，最適なモデルパラメータ $\mathcal{F}_\mathcal{L} = \{\mathbf{N}, \mathbf{E}^{(\mathbf{S})}, \mathbf{M}^{(\mathbf{S})}\}$ を推定する．

ここで，i 番目の疫病に対するグローバル（国）のシーケンス $\bar{\boldsymbol{x}}_i$ は l カ所すべての州のローカルシーケンスの合計値とする：$\bar{\boldsymbol{x}}_i(t) = \sum_{j=1}^{l} \boldsymbol{x}_{ij}(t)$．Algorithm 1 は FunnelFit の概要を示す．疫病テンソル $\mathcal{X}$ が与えられたとき，提案アルゴリズムはモデルパラメータ集合 $\mathcal{F}$ を自動抽出する．

4.2.1　GlobalFit

本項の目的は，与えられたテンソル $\mathcal{X}$ に対し，コスト関数（式 (4)）が最小化するようなグローバルパラメータ集合 $\mathcal{F}_\mathcal{G}$ を発見することである．具体的には，各疫病に関する基本的なパラメータ（つまり，基本行列，減少行列），そして，最適な外部ショックと入力エラーの数を推定したい．ここで強調すべき点として，適切な外部ショックと入力エラーを発見することは非常に重要な課題である．なぜなら，図 5 (a) に示したとおり，パラメータの学習は，外れ値に対し非常に繊細であり，外部ショックと入力エラーを適切に発見し，取り除くことが，基本パラメータの学習のうえで重要となるからである．同時に，基本パラメータの適切な学習により，外部ショックと入力エラーの発見もより適切に行うことができる．この循環依存関係を避けるために，本研究では，反復法を用いたパラメータ学習を行う．具体的には，コスト関数が最小化するまで，基本パラメータと外部ショック，入力エラーパラメータの学習を交互に行う．

外部ショックと入力エラー． もう 1 つの重要な問題として，外部ショックと入力エラーの自動判別がある．たとえば，図 5 の場合には，入力エラーとして表現する方が適している．本論文で提案するコスト関数を用いれば，このような判別は自動的に行うことができる．具体的な方法としては，まず，データを外部ショックと入力エラーそれぞれの場合に分けてパラメータ学習を行い，次に，2 つの学習結果の符号化コストを比較し，適切なモデルを選択する．図 5 の例では，(b) のコストは (a) のコストより低くなるため，提案アルゴリズムは，この点を入力エラーとして判別する．

提案アルゴリズム． Algorithm 2 は GlobalFit の処理の流れを示している．テンソル $\mathcal{X}$ が与えられたとき，提案アルゴリズムは，d 個のグローバル（国）の疫病シーケンス：$\{\bar{\boldsymbol{x}}_i\}_{i=1}^{d}$ を作成し，基本パラメータと外部ショック，入力エラーをそれぞれ推定した．ここで，コスト関数の最小化には，非線形性を有する学習に適したレーベンバーグ・マルカート（LM: Levenberg-Marquardt）法を用いた．

ここで，外部ショックテンソル $\mathcal{E}$ と入力エラーテンソル $\mathcal{M}$ は三つ組 $(disease, state, time)$ から構成されるが，GlobalFit はグローバルなパラメータ（つまり，$(disease, time)$）のみを学習する．ローカルなパラメータ集合 $\mathbf{E}^{(\mathbf{S})}$，$\mathbf{M}^{(\mathbf{S})}$ は，LocalFit（Algorithm 3）において学習する．さらに，コスト関数（式 (4)）は，$\mathbf{N}$ 等のローカルなパラメータも含むが，これらのコストはグローバルの学習とは独立であるので，定数として扱う．

4.2.2　LocalFit

$d \times l$ 個で構成されるローカルな疫病シーケンス集合：$\{\boldsymbol{x}_{ij}\}_{i,j=1}^{d,l} \in \mathcal{X}$ と，グローバルパラメータ集合 $\mathcal{F}_\mathcal{G}$ が与えられたとき，LocalFit は，各疫病，各州の個々のパラメータ集合：$\mathcal{F}_\mathcal{L} = \{\mathbf{N}, \mathbf{E}^{(\mathbf{S})}, \mathbf{M}^{(\mathbf{S})}\}$ を学習する．

Algorithm 3 は LocalFit の処理を示す．LocalFit は反復法に基づくアルゴリズムであり，それぞれ，(a) 地域行列 $\mathbf{N}$，(b) ローカルな外部ショック $\mathbf{E}^{(\mathbf{S})}$，そして (c) ローカルな入力エラー $\mathbf{M}^{(\mathbf{S})}$ のパラメータを最適化し，コスト

Algorithm 2 GlobalFit ($\mathcal{X}$)

Input: Tensor $\mathcal{X}$
Output: Set of global-level parameters $\mathcal{F}_\mathcal{G}$
for $i = 1 : d$ **do**
  Create $\bar{\boldsymbol{x}}_i$ from $\mathcal{X}$; /* Global sequence $\bar{\boldsymbol{x}}_i$ of i-th disease */
  /* Initialize external shocks and mistakes for disease i */
  $\mathbf{E}_i^{(\mathbf{D})} = \mathbf{E}_i^{(\mathbf{T})} = \mathbf{M}_i^{(\mathbf{D})} = \mathbf{M}_i^{(\mathbf{T})} = \emptyset$;
  while improving the cost **do**
    $\boldsymbol{b}_i = \arg\min_{\boldsymbol{b}'_i} Cost_C(\bar{\boldsymbol{x}}_i | \boldsymbol{b}'_i, \boldsymbol{r}_i, \mathbf{E}_i^{(\mathbf{T})}, \mathbf{M}_i^{(\mathbf{T})})$; /* Base */
    $\boldsymbol{r}_i = \arg\min_{\boldsymbol{r}'_i} Cost_C(\bar{\boldsymbol{x}}_i | \boldsymbol{b}_i, \boldsymbol{r}'_i, \mathbf{E}_i^{(\mathbf{T})}, \mathbf{M}_i^{(\mathbf{T})})$; /* Reduction */
    $\mathbf{E}_i^{(\mathbf{D})} = \mathbf{E}_i^{(\mathbf{T})} = \mathbf{M}_i^{(\mathbf{D})} = \mathbf{M}_i^{(\mathbf{T})} = \emptyset$; /* Initialize values */
    /* Find external shocks and mistakes for disease i */
    while improving the cost **do**
      $\boldsymbol{e}^{(\boldsymbol{T})} = \arg\min_{\boldsymbol{e}'^{(\boldsymbol{T})}} Cost_C(\bar{\boldsymbol{x}}_i | \boldsymbol{b}_i, \boldsymbol{r}_i, \{\mathbf{E}_i^{(\mathbf{T})} \cup \boldsymbol{e}'^{(\boldsymbol{T})}\}, \mathbf{M}_i^{(\mathbf{T})})$;
      $\boldsymbol{m}^{(\boldsymbol{T})} = \arg\min_{\boldsymbol{m}'^{(\boldsymbol{T})}} Cost_C(\bar{\boldsymbol{x}}_i | \boldsymbol{b}_i, \boldsymbol{r}_i, \mathbf{E}_i^{(\mathbf{T})}, \{\mathbf{M}_i^{(\mathbf{T})} \cup \boldsymbol{m}'^{(\boldsymbol{T})}\})$;
      /* Compare external shock vs. mistake */
      if $Cost_T(\bar{\boldsymbol{x}}_i; \boldsymbol{e}^{(\boldsymbol{T})}) < Cost_T(\bar{\boldsymbol{x}}_i; \boldsymbol{m}^{(\boldsymbol{T})})$ **then**
        /* External shock wins - treat as an external shock */
        $\mathbf{E}_i^{(\mathbf{D})} = \{\mathbf{E}_i^{(\mathbf{D})} \cup i\}$; $\mathbf{E}_i^{(\mathbf{T})} = \{\mathbf{E}_i^{(\mathbf{T})} \cup \boldsymbol{e}^{(\boldsymbol{T})}\}$;
      else
        /* Mistake wins - treat as a mistake value */
        $\mathbf{M}_i^{(\mathbf{D})} = \{\mathbf{M}_i^{(\mathbf{D})} \cup i\}$; $\mathbf{M}_i^{(\mathbf{T})} = \{\mathbf{M}_i^{(\mathbf{T})} \cup \boldsymbol{m}^{(\boldsymbol{T})}\}$;
      end if
    end while
  end while
  /* Update parameter set of i-th disease */
  $\mathbf{B} = \mathbf{B} \cup \boldsymbol{b}_i$; $\mathbf{R} = \mathbf{R} \cup \boldsymbol{r}_i$;
  $\mathbf{E}^{(\mathbf{D})} = \mathbf{E}^{(\mathbf{D})} \cup \mathbf{E}_i^{(\mathbf{D})}$; $\mathbf{E}^{(\mathbf{T})} = \mathbf{E}^{(\mathbf{T})} \cup \mathbf{E}_i^{(\mathbf{T})}$;
  $\mathbf{M}^{(\mathbf{D})} = \mathbf{M}^{(\mathbf{D})} \cup \mathbf{M}_i^{(\mathbf{D})}$; $\mathbf{M}^{(\mathbf{T})} = \mathbf{M}^{(\mathbf{T})} \cup \mathbf{M}_i^{(\mathbf{T})}$;
end for
return $\mathcal{F}_\mathcal{G} = \{\mathbf{B}, \mathbf{R}, \mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{T})}, \mathbf{M}^{(\mathbf{D})}, \mathbf{M}^{(\mathbf{T})}\}$;

を最小化する．

補助定理 1　FunnelFit の計算量は $O(dln)$ である．

証明 1　テンソル $\mathcal{X}$ からグローバル（国）の疫病シーケンスを生成するために $O(dln)$ の計算量を要する．GlobalFit の計算量は $O(\#iter \cdot (k + |\mathcal{M}|) \cdot dn)$ であり，$\#iter$ は反復回数を示し，k と $|\mathcal{M}|$ はそれぞれ，外部ショックの個数，入力エラーテンソル $\mathcal{M}$ の中の非ゼロ要素の総数を示す．同様に，LocalFit は $O(\#iter \cdot (k + |\mathcal{M}|) \cdot dln)$ の計算量を要する．ここで，$\#iter$, k, $|\mathcal{M}|$ は非常に小さい定数であるため，無視することができる．したがって，計算量は $O(dln)$ である．

Algorithm 3 LocalFit ($\mathcal{X}, \mathbf{B}, \mathbf{R}, \mathbf{E}^{(\mathbf{D})}, \mathbf{E}^{(\mathbf{T})}, \mathbf{M}^{(\mathbf{D})}, \mathbf{M}^{(\mathbf{T})}$)

Input: (a) Tensor $\mathcal{X}$, (b) global-level parameter set $\mathcal{F}_\mathcal{G}$
Output: Set of local-level parameters, i.e., $\mathcal{F}_\mathcal{L}$
while improving the cost **do**
  /* For each local sequence $\boldsymbol{x}_{ij}$ of i-th disease in j-th state */
  for $i = 1 : d$ **do**
    for $j = 1 : l$ **do**
      $N_{ij} = \arg\min_{N'_{ij}} Cost_C(\boldsymbol{x}_{ij} | \mathbf{B}, \mathbf{R}, N'_{ij}, \mathcal{E}, \mathcal{M})$;
    end for
  end for
  for each external shock $(\boldsymbol{e}^{(\boldsymbol{D})}, \boldsymbol{e}^{(\boldsymbol{S})}, \boldsymbol{e}^{(\boldsymbol{T})}) \in \mathcal{E}$ **do**
    Update $\boldsymbol{e}^{(\boldsymbol{S})}$ to minimize the cost /* Local participation rate */
  end for
  for each mistake $(\boldsymbol{m}^{(\boldsymbol{D})}, \boldsymbol{m}^{(\boldsymbol{S})}, \boldsymbol{m}^{(\boldsymbol{T})}) \in \mathcal{M}$ **do**
    Update $\boldsymbol{m}^{(\boldsymbol{S})}$ to minimize the cost /* Mistake value */
  end for
end while
return $\mathcal{F}_\mathcal{L} = \{\mathbf{N}, \mathbf{E}^{(\mathbf{S})}, \mathbf{M}^{(\mathbf{S})}\}$;

5. 評価実験

本論文では Funnel の有効性を検証するため，実データを用いた実験を行った．具体的には，本章では以下の項目について検証する．

Q1　疫病のパターン抽出に関する提案手法の有効性
Q2　提案アルゴリズムの精度の検証
Q3　パターン抽出に対する計算時間の検証

5.1 疫病データからの重要パターンの発見

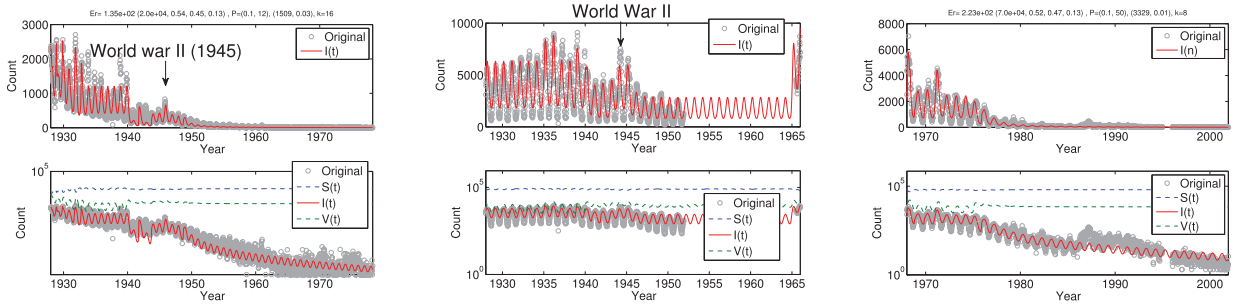
本節では，大規模疫病データに対する Funnel の情報抽出の効果を検証する．

図 6 は，15 種の主要な疫病に対するモデル学習の結果を示している． 灰色の丸印はオリジナルデータを，赤の実線は提案手法の学習結果 $I(t)$ を示している．各シーケンスはそれぞれ，線形スケール（上段），対数スケール（下段）で示しており，下段については，感染者の推定値のほかに，感受性保持者 $S(t)$ と免疫保持者 $V(t)$ の推定値も表している．

以下では，先述の 5 つの重要な疫病の要素に関して，提案手法で得られた知見について議論する．

(P1) 季節性． 図 1 (c) においてすでに示したように，Funnel は，季節性における次の 4 つのカテゴリを発見した．

- インフルエンザは 1 月から 2 月にかけてピークを持つ．
- 麻疹，おたふくかぜ，水疱瘡等の小児呼吸器疾患は強



(a) ジフテリア (Diphtheria) (**P1**), (**P2**), (**P4**)　(b) 猩紅熱 (Scarlet fever) (**P1**), (**P2**), (**P4**)　(c) おたふくかぜ (Mumps) (**P1**), (**P2**), (**P4**)

(d) 風疹 (Rubella) (**P1**), (**P2**), (**P4**)　(e) 水疱瘡 (Chickenpox) (**P1**), (**P2**)　(f) 百日咳 (Whooping cough) (**P1**), (**P2**), (**P4**)

(g) インフルエンザ (Influenza) (**P1**), (**P4**)　(h) 天然痘 (Smallpox)(**P1**), (**P2**), (**P4**)　(i) 腸チフス (Typhoid fever) (**P1**), (**P2**), (**P4**), (**P5**)

(j) クリプトスポリジウム症 (Cryptosporidiosis) (**P1**), (**P3**), (**P4**)　(k) ライム病 (Lyme disease) (**P1**)　(l) チフス熱 (Typhus fever) (**P1**)

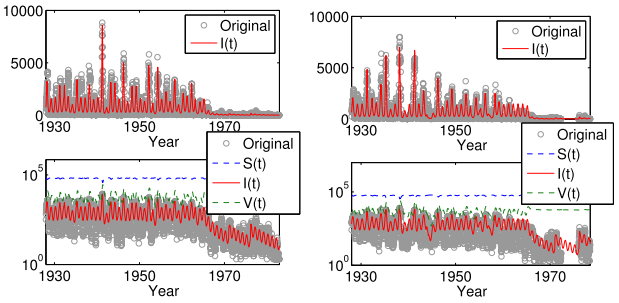
(m) 淋病 (Gonorrhea)　(n) 狂犬病 (Rabies in animals) (**P1**), (**P2**), (**P4**)　(o) ブルセラ症 (Brucellosis) (**P1**), (**P2**), (**P4**)

図 6　主要な 15 種の疫病に対する Funnel の学習結果

Fig. 6　Fitting results of Funnel for 15 diseases (global-level counts).

表 3　ワクチンのライセンス化 [43] と Funnel の結果
Table 3　The year of vaccine licensure [43] vs. detection.

Disease	licensure	detected
Measles	1963	**1965**
Mumps	1967	**1975**
Whooping cough (pertussis)	1948	**1951**
Rubella	1969	**1972**



(a) ニューヨーク州 (NY)　(b) ペンシルベニア州 (PA)

図 7　麻疹のローカル（州）の疫病に対する学習結果
Fig. 7　Local-level fittings for measles.

い周期性を持ち，春にかけて感染者が増加する [38].

- ライム病のようなダニ媒介疾患やクリプトスポリジウム症のような水媒性の疫病は，夏に強いピークを持ち，保菌生物やヒトの行動，あるいは気候等の要素と深い関連性を持つ [39].
- 淋病のような性感染症（STD）には周期性が見られない.

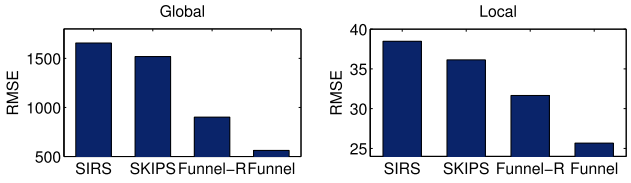
(P2) 疫病減少効果.　Funnel は，大規模疫病データの中から，自動的に疫病減少効果を発見することができる．たとえば，図 6 (a)–(f) において，多くの小児性疾患は疫病減少効果が見られることが分かる．**表 3** は，ワクチンがライセンス化された年と，Funnel において自動発見された疫病減少効果の開始年を比較したものである．基本的に，ライセンス化されて 2，3 年後に減少効果が見られることが分かる．これは，ワクチンプログラムの普及にラグが生じる影響と考えられる．一方で，インフルエンザに対する疫病減少効果は発見されなかった．これは，インフルエンザウィルスの変異が活発であり，ワクチンの効果が現れにくいことが原因である.

(P3) 地域性.　Funnel は，国家規模の疫病だけでなく，地域固有のパターンを発見することができる．たとえば，図 1 (b) で示したとおり，ニューヨーク州（NY）とペンシルベニア州（PA）には多くの麻疹の感受性保持者が存在する．**図 7** は，実際の感染者数とその学習結果である．図に示すとおり，Funnel は周期的な感染パターンや疫病減少効果を表現すると同時に，ローカルな外部ショックのパターンも柔軟に表現している.

(P4) 外部ショックイベント.　図 6 に示すように，FunnelFit は次にあげられるような重要な外部ショックイベントを自動発見した.

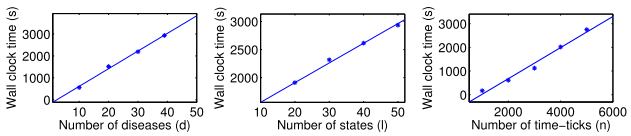
- 小児性疾患である (a) ジフテリア（Diphtheria）と (b) 猩紅熱（scarlet fever）は，第二次世界大戦前後に複数年にわたる流行パターンを持つ.
- (g) インフルエンザ（Influenza）は，1929 年，1941 年等に大規模な流行が確認できる [26].
- (h) ワクチンの普及が不十分だったことにより，天然痘（smallpox）の流行が 1937–39 年にかけて発生した [5].
- (j) 水媒性の疫病であるクリプトスポリジウム症（Cryptosporidiosis）が，2007 年におけるユタ州の公共プール汚染を原因として局所的に流行した [1].

(P5) 入力エラー.　提案手法である FunnelFit の強みの 1 つとして，ノイズに対する頑健性があげられる．図 6 (i) に示すとおり，FunnelFit は，腸チフス（typhoid fever）のデータに含まれる，1940 年代における欠損値を補完すると同時に，1953 年における外れ値を取り除くことができる.



(a) Global fitting　(b) Local fitting

図 8　Funnel と既存手法の精度比較
Fig. 8　Fitting accuracy for the global and local sequences.

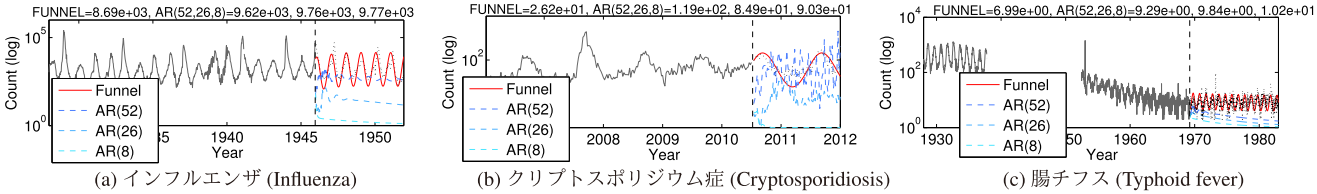


(a) Diseases (d)　(b) States (l)　(c) Duration (n)

図 9　FunnelFit の計算コスト
Fig. 9　FunnelFit scales linearly: wall clock time vs. dataset size.

5.2　提案手法の精度と学習時間

まず，提案モデルの精度を検証するため，既存手法である *SIRS* モデルと *SKIPS* [40] との比較を行った．本研究ではさらに，疫病減少効果の効果を検証するために，外部ショックと入力エラーに関するパラメータを取り除いた Funnel-R とも比較を行った．**図 8** (a) は，オリジナルデータと推定値との二乗平均誤差（RMSE: root mean square error）を示している．ここで，(a) は，グローバル（国）の疫病シーケンス集合 $\{\bar{\boldsymbol{x}}_i(t)\}_{i,t}^{d,n}$，そして (b) はローカル（州）のシーケンス集合 $\{\boldsymbol{x}_{ij}(t)\}_{i,j,t}^{d,l,n}$ についてそれぞれ比較している．*SIRS* モデルは，季節性をともなうパターンを表現できない．*SKIPS* は周期性を発見できるが，疫病減少効果や外部ショックイベントを表現できない．図に示



(a) インフルエンザ (Influenza)　(b) クリプトスポリジウム症 (Cryptosporidiosis)　(c) 腸チフス (Typhoid fever)

図 10　3 つの疫病に対する将来予測の結果

Fig. 10　Forecasting result: we train the model parameters using 2/3 of each sequence (i.e., solid black lines). We then start forecasting (at the vertical dotted line).

すとおり，*SIRS* モデルと *SKIPS* が疫病の複雑なパターンの表現に失敗したのに比べ，提案手法は高い精度でのデータの学習に成功した．

次に，FunnelFit の計算時間を検証する．**図 9** はデータのサイズを変化させたうえでの提案手法の計算時間を示している．ここでは，それぞれ，(a) 疫病の総数 d，(b) 州の総数 l，(c) シーケンスの長さ n を変化させている．補助定理 1 で示したように，提案アルゴリズムである FunnelFit は，データの入力サイズに対し線形である．

6.　ディスカッション

アプリケーション — 疫病の将来予測． Funnel は，様々な種類の疫病データを柔軟に表現可能なため，最も重要なアプリケーションとして，将来予測があげられる．**図 10** は，3 つの主要な疫病に対する予測結果を示している．具体的には，オリジナルデータの前半 2/3 の長さ（黒線）を用いてモデルパラメータを学習し，その後の 1/3 を予測している（赤線）．ここで，y 軸は対数スケールで表示している．提案モデルの精度を検証するため，自己回帰モデル（AR: autoregressive model）との比較を行った．AR の回帰係数はそれぞれ，$r = 52$（1 年間），26（半年間），8（Funnel の基本パラメータ数（6），および疫病減少効果パラメータ数（2）の合計）とした．さらに，学習の発散を抑えるため，AR については対数スケールに対し予測を行った．図 10 に示したとおり，提案手法は AR と比較し高精度の予測を達成している．図中上の数値は各手法における予測結果とオリジナルデータとの誤差（RMSE）を示している．より具体的には，(a) インフルエンザ（Influenza）と (b) クリプトスポリジウム症（Cryptosporidiosis）について，AR が突発的な外部ショックや外れ値に強い影響を受け，将来予測に失敗しているのに対し，提案モデルは複雑な時系列イベントを自動的に抽出することで潜在的な疫病パターンの長期予測に成功している． 同様にして，(c) 腸チフス（Typhoid fever）のような疫病減少効果や入力エラー，欠損値を多く含むデータに対しても，周期性を含む複雑なパターンを表現し，長期的な将来予測を行うことができた．

一般性 — コンピュータウィルスの流行パターン． Funnel は，その一般性と柔軟性により，疫病だけでなく，コンピュータウィルス等の流行パターンも学習することができる．たとえば，**図 11** は，IPA [*10] による日本国内の企業および教育機関におけるコンピュータウィルス感染数の報告データに対する Funnel の学習結果を示している．ここで，丸印はオリジナルの報告件数，Funnel の学習結果は線で表現され，期間は 2000 年から 2010 年である．たとえば，(a) “Badtrans”，“Klez” は Microsoft Outlook のセキュリティホールを利用したウィルスで，2001 年から 2002 年にかけて流行した．これらのウィルスは，強い感染力を持っていたが，アンチウィルスソフトウェアの普及により鎮火された．(b) “Netsky” は，メールの添付ファイルとして拡散されたウィルスで，2004 年に大規模な感染が報告された．このウィルスは，10 年間にわたり，徐々に減少傾向にあるが，依然として感染報告がある．さらに，このウィルスには，週単位の周期性が確認され，週末の感染数が少ないという特徴がある．(c) “Mytob” は，社内ネットワーク上（ローカル）で伝搬するウィルスである．さらに，最近になり猛威をふるっているのが，“Koobface”，“Fbphotofake” 等に代表されるソーシャルネットワークサイト上でのウィルスである．これらは，Facebook や Twitter 等で急速に伝搬し，潜在的な感染者数（ユーザ数）も多い．

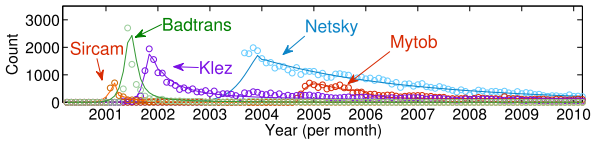


図 11　コンピュータウィルスにおける Funnel の学習結果

Fig. 11　Funnel is general: our model (solid lines) fits computer virus data (in circles) very well.

7.　むすび

本論文では，大規模疫病データのための非線形モデル解析手法として Funnel について述べた．Funnel は，大規模な疫病データの中から，季節性，ワクチン効果，地域性，外部ショックイベントや入力エラー等の重要な要素を自動抽出し，長期的な将来予測の能力を有する．様々な種類の疫病データを用いて実験を行い，Funnel が最新の疫病解析手法と比べてより高い精度と性能を持つことを示し

[*10] IPA - IT security center: https://www.ipa.go.jp/security/english/index.html

た．今後の課題として，他の地域や異なる疫病のデータ，あるいは気象をはじめとする外部要因に関するデータを統合し，疫病テンソルデータに対するより高度なモデル学習や解析をするための技術について検討していく予定である．

謝辞　本研究の一部は JSPS 科研費 JP15H02705，JP16K12430，JP26280112，JP26730060，JST さきがけおよび総務省 SCOPE（受付番号 162110003）の助成を受けたものです．

We thank Dr. Donald S. Burke, Dean of the Graduate School of Public Health, University of Pittsburgh for his support and expert opinion during this study. This material is based upon work supported by the National Science Foundation under Grant No.CNS-1314632. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

参考文献

[1] Promotion of healthy swimming after a statewide outbreak of cryptosporidiosis associated with recreational water venues–utah, 2008-2009, *MMWR Morb Mortal Wkly Rep*, Vol.61, No.19, pp.348–352 (2012).

[2] Anderson, R.M. and May, R.M.: *Infectious Diseases of Humans Dynamics and Control*, Oxford University Press (1992).

[3] Böhm, C., Faloutsos, C., Pan, J.-Y. and Plant, C.: Ric: Parameter-free noise-robust clustering, *TKDD*, Vol.1, No.3 (2007).

[4] Box, G.E., Jenkins, G.M. and Reinsel, G.C.: *Time Series Analysis: Forecasting and Control, 3rd edition*, Prentice Hall, Englewood Cliffs, NJ (1994).

[5] D. CC. Smallpox in the united states: It's decline and geographic distribution, *Public Health Reports*, Vol.55, No.50, pp.2303–2312 (1940).

[6] Chakrabarti, D., Papadimitriou, S., Modha, D.S. and Faloutsos, C.: Fully automatic cross-associations, *KDD*, pp.79–88 (2004).

[7] Chen, L. and Ng, R.T.: On the marriage of lp-norms and edit distance, *VLDB*, pp.792–803 (2004).

[8] Davidson, I.N., Gilpin, S., Carmichael, O.T. and Walker, P.B.: Network discovery via constrained tensor analysis of fmri data, *KDD*, pp.194–202 (2013).

[9] Earn, D.J., Rohani, P., Bolker, B.M. and Grenfell, B.T.: A simple model for complex dynamical transitions in epidemics, *Science*, Vol.287, No.5453, pp.667–670 (2000).

[10] Grenfell, B.T., Bjornstad, O.N. and Kappey, J.: Travelling waves and spatial hierarchies in measles epidemics, *Nature*, Vol.414, p.716 (2001).

[11] Jain, A., Chang, E.Y. and Wang, Y.-F.: Adaptive stream resource management using kalman filters, *SIGMOD*, pp.11–22 (2004).

[12] Kumar, R., Mahdian, M. and McGlohon, M.: Dynamics of conversations, *KDD*, pp.553–562 (2010).

[13] Lee, J.-G., Han, J. and Whang, K.-Y.: Trajectory clustering: A partition-and-group framework, *SIGMOD*, pp.593–604 (2007).

[14] Leskovec, J., Backstrom, L., Kumar, R. and Tomkins, A.: Microscopic evolution of social networks, *KDD*, pp.462–470 (2008).

[15] Li, L., Prakash, B.A. and Faloutsos, C.: Parsimonious linear fingerprinting for time series, *PVLDB*, Vol.3, No.1, pp.385–396 (2010).

[16] Matsubara, Y., Li, L., Papalexakis, E.E., Lo, D., Sakurai, Y. and Faloutsos, C.: F-trail: Finding patterns in taxi trajectories, *PAKDD*, pp.86–98 (2013).

[17] Matsubara, Y. and Sakurai, Y.: Regime shifts in streams: Real-time forecasting of co-evolving time sequences, *KDD*, pp.1045–1054 (2016).

[18] Matsubara, Y., Sakurai, Y. and Faloutsos, C.: Autoplait: Automatic mining of co-evolving time sequences, *SIGMOD*, pp.193–204 (2014).

[19] Matsubara, Y., Sakurai, Y. and Faloutsos, C.: The web as a jungle: Non-linear dynamical systems for co-evolving online activities, *WWW*, pp.721–731 (2015).

[20] Matsubara, Y., Sakurai, Y. and Faloutsos, C.: Non-linear mining of competing local activities, *WWW* (2016).

[21] Matsubara, Y., Sakurai, Y., Faloutsos, C., Iwata, T. and Yoshikawa, M.: Fast mining and forecasting of complex time-stamped events, *KDD*, pp.271–279 (2012).

[22] Matsubara, Y., Sakurai, Y., Prakash, B.A., Li, L. and Faloutsos, C.: Rise and fall patterns of information diffusion: Model and implications, *KDD*, pp.6–14 (2012).

[23] Matsubara, Y., Sakurai, Y., Ueda, N. and Yoshikawa, M.: Fast and exact monitoring of co-evolving data streams, *ICDM*, pp.390–399 (2014).

[24] Matsubara, Y., Sakurai, Y., van Panhuis, W.G. and Faloutsos, C.: FUNNEL: Automatic mining of spatially coevolving epidemics, *KDD*, pp.105–114 (2014).

[25] Matsubara, Y., Sakurai, Y. and Yoshikawa, M.: Scalable algorithms for distribution search, *ICDM*, pp.347–356 (2009).

[26] NM, F., AP, G. and RM, B.: Ecological and immunological determinants of influenza evolution, *Nature*, Vol.422, No.6930, pp.428–433 (2003).

[27] Papadimitriou, S., Sun, J. and Faloutsos, C.: Streaming pattern discovery in multiple time-series, *VLDB*, pp.697–708 (2005).

[28] Papadimitriou, S. and Yu, P.S.: Optimal multi-scale patterns in time series streams, *SIGMOD*, pp.647–658 (2006).

[29] PE, F. and JA, C.: Measles in england and wales–i: An analysis of factors underlying seasonal patterns, *Epidemiol*, Vol.11, No.1, pp.5–14 (1982).

[30] Prakash, B.A., Beutel, A., Rosenfeld, R. and Faloutsos, C.: Winner takes all: Competing viruses or ideas on fair-play networks, *WWW*, pp.1037–1046 (2012).

[31] Prakash, B.A., Chakrabarti, D., Faloutsos, M., Valler, N. and Faloutsos, C.: Threshold conditions for arbitrary cascade models on arbitrary networks, *ICDM*, pp.537–546 (2011).

[32] Rakthanmanon, T., Campana, B.J.L., Mueen, A., Batista, G.E.A.P.A., Westover, M.B., Zhu, Q., Zakaria, J. and Keogh, E.J.: Searching and mining trillions of time series subsequences under dynamic time warping, *KDD*, pp.262–270 (2012).

[33] Sakurai, Y., Faloutsos, C. and Yamamuro, M.: Stream monitoring under the time warping distance, *ICDE*, Istanbul, Turkey, pp.1046–1055 (April 2007).

[34] Sakurai, Y., Matsubara, Y. and Faloutsos, C.: Mining and forecasting of big time-series data, *SIGMOD, Tutorial*, pp.919–922 (2015).

[35] Sakurai, Y., Matsubara, Y. and Faloutsos, C.: Mining big time-series data on the web, *WWW, Tutorial* (2016).
[36] Sakurai, Y., Papadimitriou, S. and Faloutsos, C.: Braid: Stream mining through group lag correlations, *SIGMOD*, pp.599–610 (2005).
[37] Sakurai, Y., Yoshikawa, M. and Faloutsos, C.: Ftw: Fast similarity search under the time warping distance, *PODS*, Baltimore, Maryland, pp.326–337 (2005).
[38] SF, D.: Seasonal variation in host susceptibility and cycles of certain infectious diseases, *Emerg Infect Dis.*, Vol.7, No.3, pp.369–374 (2001).
[39] SM, M., RJ, E., A, M. and P, M.: Seasonality in six enterically transmitted diseases and ambient temperature, *Am. J. Trop. Med. Hyg.* (2014).
[40] Stone, L., Olinky, R. and Huppert, A.: Seasonal dynamics of recurrent epidemics, *Nature*, Vol.446, pp.533–536 (March 2007).
[41] Sun, J., Tao, D. and Faloutsos, C.: Beyond streams and graphs: Dynamic tensor analysis, *KDD*, pp.374–383 (2006).
[42] Tao, Y., Faloutsos, C., Papadias, D. and Liu, B.: Prediction and indexing of moving objects with unknown motion patterns, *SIGMOD*, pp.611–622 (2004).
[43] van Panhuis, W.G., Grefenstette, J., Jung, S.Y., Chok, N.S., Cross, A., Eng, H., Lee, B.Y., Zadorozhny, V., Brown, S., Cummings, D. and Burke, D.S.: Contagious diseases in the united states from 1888 to the present, *NEJM*, Vol.369, No.22, pp.2152–2158 (2013).
[44] Vlachos, M., Gunopulos, D. and Kollios, G.: Discovering similar multidimensional trajectories, *ICDE*, pp.673–684 (2002).

松原 靖子　（正会員）

2006 年お茶の水女子大学理学部情報科学科卒業．2009 年同大学院博士前期課程修了．2012 年京都大学大学院情報学研究科社会情報学専攻博士後期課程修了．博士（情報学）．2012 年 NTT コミュニケーション科学基礎研究所 RA．2013 年熊本大学大学院自然科学研究科日本学術振興会特別研究員（PD）．2014 年より同大学院助教．この間，カーネギーメロン大学客員研究員．2016 年 12 月より国立研究開発法人科学技術振興機構さきがけ研究員．2016 年度日本データベース学会上林奨励賞，山下記念研究賞受賞．大規模時系列データマイニングに関する研究に従事．ACM，日本データベース学会各会員．

櫻井 保志　（正会員）

1991 年同志社大学工学部電気工学科卒業．1991 年日本電信電話（株）入社．1999 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．2004～2005 年カーネギーメロン大学客員研究員．2013 年熊本大学大学院自然科学研究科教授．本会平成 18 年度長尾真記念特別賞，平成 16 年度および平成 19 年度論文賞，電子情報通信学会平成 19 年度論文賞，日本データベース学会上林奨励賞，ACM KDD best paper awards（2008，2010）等受賞．データマイニング，データストリーム処理，センサーデータ処理，Web 情報解析技術の研究に従事．ACM，電子情報通信学会，日本データベース学会各会員．

Willem G. van Panhuis

Wilbert van Panhuis, MD, PhD, is an assistant professor in the Department of Epidemiology and a faculty member of the Public Health Dynamics Laboratory, University of Pittsburgh Graduate School of Public Health. Dr. Van Panhuis is an infectious disease epidemiologist specializing in (inter)national disease surveillance systems, vector-borne and vaccine preventable diseases and global cooperation for disease data sharing and disease control. He is the lead scientist of the Tycho project which aims to provide open access to newly digitized, standardized US weekly disease surveillance data in a dynamic online user environment.

Christos Faloutsos

Christos Faloutsos is a Professor at Carnegie Mellon University. He has received the Presidential Young Investigator Award by the National Science Foundation (1989), the Research Contributions Award in ICDM 2006, the SIGKDD Innovations Award (2010), 24 “best paper” awards (including 5 “test of time” awards), and four teaching awards. Six of his advisees have attracted KDD or SCS dissertation awards, He is an ACM Fellow, he has served as a member of the executive committee of SIGKDD; he has published over 350 refereed articles, 17 book chapters and two monographs. He holds seven patents (and 2 pending), and he has given over 40 tutorials and over 20 invited distinguished lectures. His research interests include large-scale data mining with emphasis on graphs and time sequences; anomaly detection, tensors, and fractals.

（担当編集委員　渋谷 哲朗）