

大規模疫病データのための非線形テンソル解析

松原 靖子[†] 櫻井 保志[†] Willem G. van Panhuis^{††} Christos Faloutsos^{†††}

[†] 熊本大学自然科学研究科情報電気電子工学専攻

^{††} Dept. of Epidemiology, University of Pittsburgh

^{†††} Dept. of Computer Science, Carnegie Mellon University

E-mail: [†]{yasuko,yasushi}@cs.kumamoto-u.ac.jp, ^{††}wav10@pitt.edu, ^{†††}christos@cs.cmu.edu

あらまし 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発見した。

キーワード 疫病、テンソル、時系列データ、特徴自動抽出

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 [4]。^(注1) より具体的には、以下の問題を扱う。

d 種の疫病、 l 箇所の地域、 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

本研究では、疫病に関するデータとして *Tycho* [10] を扱う。*Tycho* は、アメリカ合衆国における 125 年間に渡る 56 種の疫病の患者の報告件数で構成される。^(注2)

具体例. 図 1 は、FUNNEL による解析結果を示している。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、そして数年毎に大規模な感染（1941 年、1958 年）と小さな波（1940 年と 1947 年）が現れており [6]、これはスキップ現象として知られている [9]。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。例えば、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る舞いを表現し、図 1 に挙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のサイズに対して線形である。

(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. 提案モデル

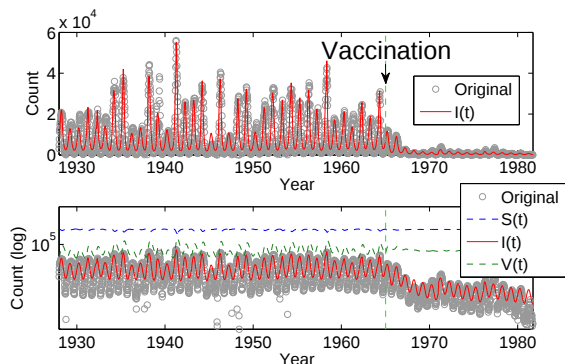
本章では、提案モデルである FUNNEL について述べる。

2.1 FUNNEL の概要

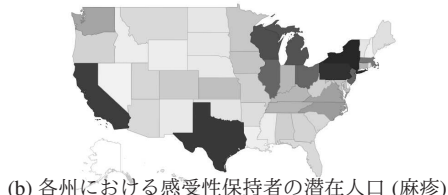
本研究で扱う *Tycho* データは、1888 年から現在にかけてのアメリカ合衆国 50 州における 56 種の疫病を患者の報告件数（計 87,950,807 件）で構成される。本データは (disease, location, timestamp) の三つ組で表現され、それぞれ、 d 種の疫病、 l 箇所の地域/州、 n の期間（一週間単位）から構成される。この疫病データは、3 階のテンソル $\mathcal{X} \in \mathbb{N}^{d \times l \times n}$ として表現することができ、 $\mathcal{X}$ の要素 $x_{ij}(t)$ は時刻 t 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。例えば、

(注1) : <http://www.cs.kumamoto-u.ac.jp/~yasuko/software.html>

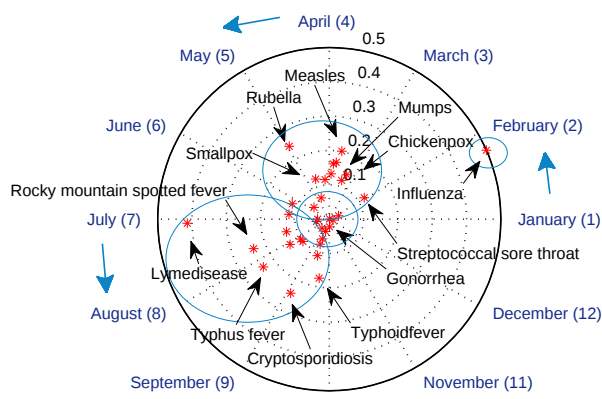
(注2) : Project Tycho at University of Pittsburgh: <http://www.tycho.pitt.edu/>



(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.

(‘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 $x_{ij} = \{x_{ij}(t)\}_{t=1}^n$ 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 $\bar{x}_i = \{\bar{x}_i(t)\}_{t=1}^n$ 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 1 (c) で示したように, 疫病には季節性を伴う場合が多く, 例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。同様に, 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。[知見 1] (疫病の季節性) 疫病の多くは年単位の周期性を伴い, 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り, 多くの疫病が, 数十年の間に大幅に減少している。例えば, 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。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。例えば, 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) がいることがわかる。

[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 年等において, 麻疹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 つの要素を全て表現する。

- (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

次章では, 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 (例えば, ニューヨーク州 (NY) の麻疹の感染者数など) 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 箇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。

2.2 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。

2.2.1 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): 既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 2 (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

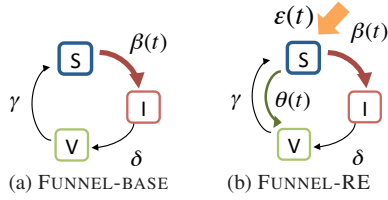


図2 FUNNEL の状態遷移の様子。

Fig.2 FUNNEL diagrams.

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^(注3)。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 における周期性を伴う関数として表現する。

[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、^(注4) 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： $\mathbf{b} = \{N, \beta_0, \delta, \gamma, P_a, P_s\}$ 。

- 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。

2.2.2 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 2 (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 $\mathbf{b}$ に加え、FUNNEL-R は次のパラメータを要する： $\mathbf{r} = \{t_\theta, \theta_0\}$ 。

- 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。

2.2.3 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。例えば、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 2 (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構成される：

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

ここで、 $\epsilon(t)$ は、一時的な感受性変化率を表す：

$$\epsilon(t) = 1 + \sum_{i=1}^k f(t; \mathbf{e}_i^{(T)}), f(t; \mathbf{e}^{(T)}) = \begin{cases} \epsilon_0 & (t_\mu - t_\sigma < t < t_\mu + t_\sigma) \\ 0 & (else) \end{cases}$$

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： $\mathbf{e}^{(T)} = \{t_\mu, t_\sigma, \epsilon_0\}$ 。

- 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。

2.3 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 の振る舞いについて述べた。ここからは、疫病テンソル $\mathcal{X}$ が与えられた上での d 種の疫病、 l 箇所の地域の個々の振る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挙げられるのが、テンソル $\mathcal{X}$ を $(d \times l)$ 個の長さ n のシーケンス集合： $\{\mathbf{x}_{ij}\}_{i,j=1}^{d,l}$ (つまり、ローカル (州) の疫病シーケンス) とみなし、各シーケンスに対し個別にモデルパラメータ集合： $\{\mathbf{b}, \mathbf{r}, \mathbf{e}^{(T)}\}$ 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、例えば、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) の疫病の振る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 箇所の全ての州で、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 3 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る舞い、(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) の振る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 $\mathcal{X}$

(注3) : γ は、出生率および死亡率も表現する。

(注4) : 1 年間は 52 週で構成される。

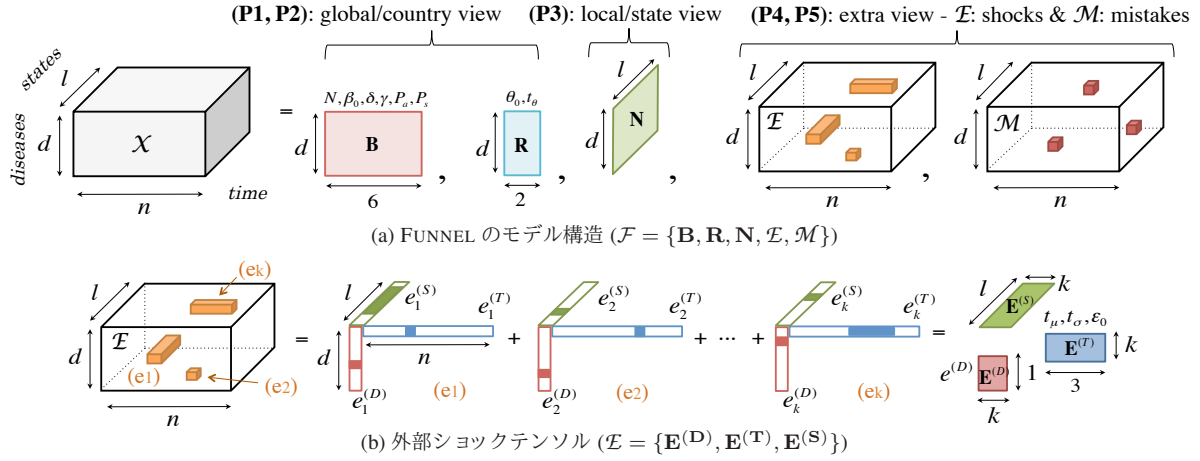


図3 FUNNEL の概要. 提案手法は疫病テンソル $\mathcal{X}$ の中から重要なパターンを自動抽出する.

Fig. 3 Illustration of FUNNEL structure: we extract the important behavior of epidemics from $\mathcal{X}$.

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、グローバルな振る舞いを表現し、 l 箇所の州で共有する.

[定義 2] (基本行列 $\mathbf{B}$ ($d \times 6$)) $\mathbf{B}$ を d 箇の疫病に関する基本パラメータ集合 $\mathbf{B} = \{b_1, \dots, b_d\}$ とする. ここで、 b_i は i 番目の疫病のパラメータを表す.

例えば、麻疹ウィルスの感染率及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 $\mathbf{R} = \{r_1, \dots, r_d\}$ とする. ここで、 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一体なんだろうか. 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 箇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 番目の州毎に導入する. 具体的には、モデル 3 において、次の式を満たす: $N_{ij} = S_{ij}(t) + I_{ij}(t) + V_{ij}(t)$. 例えば、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 箇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. 例えば、1946 年に国家規模で起きたインフルエンザの歴史的な大流行の場合は、(e1) “influenza, country-wide, 1946” として表現される. 同様に、2007 年にユタ州で起きた地域規模の

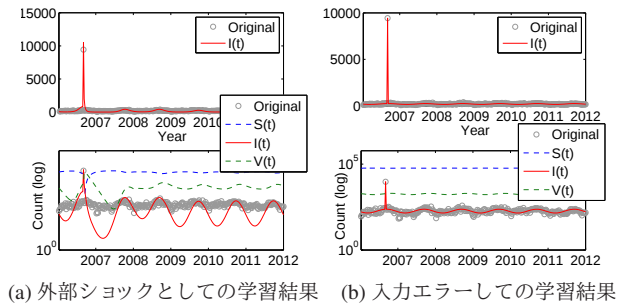


図4 ジアルジア症に対する学習結果の比較.

Fig. 4 External shock vs. mistake for giardiasis.

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 3 (b) に示すように、 $\mathcal{E}$ は、 k 箇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 $\{e^{(D)}, e^{(S)}, e^{(T)}\}$ として次のように表現される.

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

ここで、2.2.3 節で示したとおり、 $e^{(T)}$ はグローバルなパラメータであり、 $e^{(S)}$ はローカルなパラメータ: $e^{(S)} = \{e_j^{(S)}\}_{j=1}^l$ である. これは、モデル 3 におけるパラメータ ϵ_0 の l 箇所の州毎の変化率を表す. 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 箇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、3 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 4 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 (a) は外部ショックとして学習した結果であり、 図 (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わかる。

3. 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。

3.1 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 $\mathcal{X}$ を表現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、よりよいモデルであるとみなす。

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。

- 疫病的総数 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。 (注5)

- 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。 (注6)

同様に、外部ショックテンソル $\mathcal{E} = \{\mathbf{E}^{(D)}, \mathbf{E}^{(S)}, \mathbf{E}^{(T)}\}$ 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 $\mathbf{E}^{(D)}$: $k \log(d)$ 。
- 外部ショック (時間) 行列 $\mathbf{E}^{(T)}$ の中の各要素 $e^{(T)} = \{t_\mu, t_\sigma, \epsilon_0\}$: $\log(n), \log(n), c_F$ 。
- 外部ショック (地域) 行列 $\mathbf{E}^{(S)}$: $c_F \cdot kl$ 。

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。 $Cost_M(\mathcal{E}) = \log^*(k) + k(\log(d) + 2\log(n) + c_F \cdot (1 + l))$ 。

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、 $Cost_M(\mathcal{M}) = \log^*(|\mathcal{M}|) + \sum_{m_{ij}(t) > 0} (\log(d) + \log(l) + \log(n) + \log^*(m_{ij}(t)))$ 。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 [2]: $Cost_C(\mathcal{X}|\mathcal{F}) = \sum_{i,j,t=1}^{d,l,n} \log_2 p_{Gauss}^{-1}(x_{ij}(t) - m_{ij}(t) - I_{ij}(t))$ 。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。 (注7)

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。

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

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。

3.2 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 $\mathcal{F}$ が与えられた上でテンソル $\mathcal{X}$ を表現するための方法について述べた。次の課題は、どのようにして全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 2.3 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。例えば、基本行列 $\mathbf{B}$ と減少行列 $\mathbf{R}$ は、各疫病のグローバル (国) レベルの振る舞いを表現し、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 $\mathcal{E} = \{\mathbf{E}^{(D)}, \mathbf{E}^{(S)}, \mathbf{E}^{(T)}\}$ は、グローバル ($\mathbf{E}^{(D)}, \mathbf{E}^{(T)}$)、ローカル ($\mathbf{E}^{(S)}$) 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、 $\{\mathbf{M}^{(D)}, \mathbf{M}^{(S)}, \mathbf{M}^{(T)}\}$ 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、ローカルの二つの部分パラメータ集合 $\mathcal{F}_G, \mathcal{F}_L$ 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

(注5) : ここで、 $\log^*$ は整数のユニバーサル符号長を表す。

(注6) : 本論文では 4×8 ビットとする。

(注7) : 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。

$\{\bar{\mathbf{x}}_i\}_{i=1}^d$ に対し、最適なモデルパラメータ $\mathcal{F}_G = \{\mathbf{B}, \mathbf{R}, \mathbf{E}^{(D)}, \mathbf{E}^{(T)}, \mathbf{M}^{(D)}, \mathbf{M}^{(T)}\}$ を推定する。

- LOCALFIT: ローカル（州）の疫病シーケンス $\{\mathbf{x}_{ij}\}_{i,j=1}^{d,l}$ に対し、最適なモデルパラメータ $\mathcal{F}_L = \{\mathbf{N}, \mathbf{E}^{(S)}, \mathbf{M}^{(S)}\}$ 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 $\bar{\mathbf{x}}_i$ は l 箇所すべての州のローカルシーケンスの合計値とする： $\bar{\mathbf{x}}_i(t) = \sum_{j=1}^l \mathbf{x}_{ij}(t)$ 。アルゴリズム 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。

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

```

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

```

4. 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なった。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

4.1 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 5 は、9 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、感染者の推定値の他に、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) において既に示したように、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、水疱瘡などの小児呼吸器疾患は強い周期性を持ち、春にかけて感染者が増加する [7]。
- 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 [8]。
- 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。例えば、図 5 (a-e) において、多くの小児性疾患は疫病減少効果が見られる

ことがわかる。表 6 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わかる。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。例えば、図 1 (b) で示した通り、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。図に示す通り、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 5 に示すように、FUNNELFIT は次に挙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渡る流行パターンを持つ。

- (g) インフルエンザ (Influenza) は、1929 年、1941 年等到大規模な流行が確認できる [5]。

- (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 [3]。

- (f) 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 の強みの一つとして、ノイズに対する頑健性が挙げられる。図 5 (i) に示す通り、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。

4.2 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 [9] 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 $\{\bar{\mathbf{x}}_i(t)\}_{i,t}^{d,n}$ 、そして (b) はローカル（州）のシーケンス集合 $\{\mathbf{x}_{ij}(t)\}_{i,j,t}^{d,l,n}$ についてそれぞれ比較している。SIRS モデルは、季節性を伴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。図に示す通り、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上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

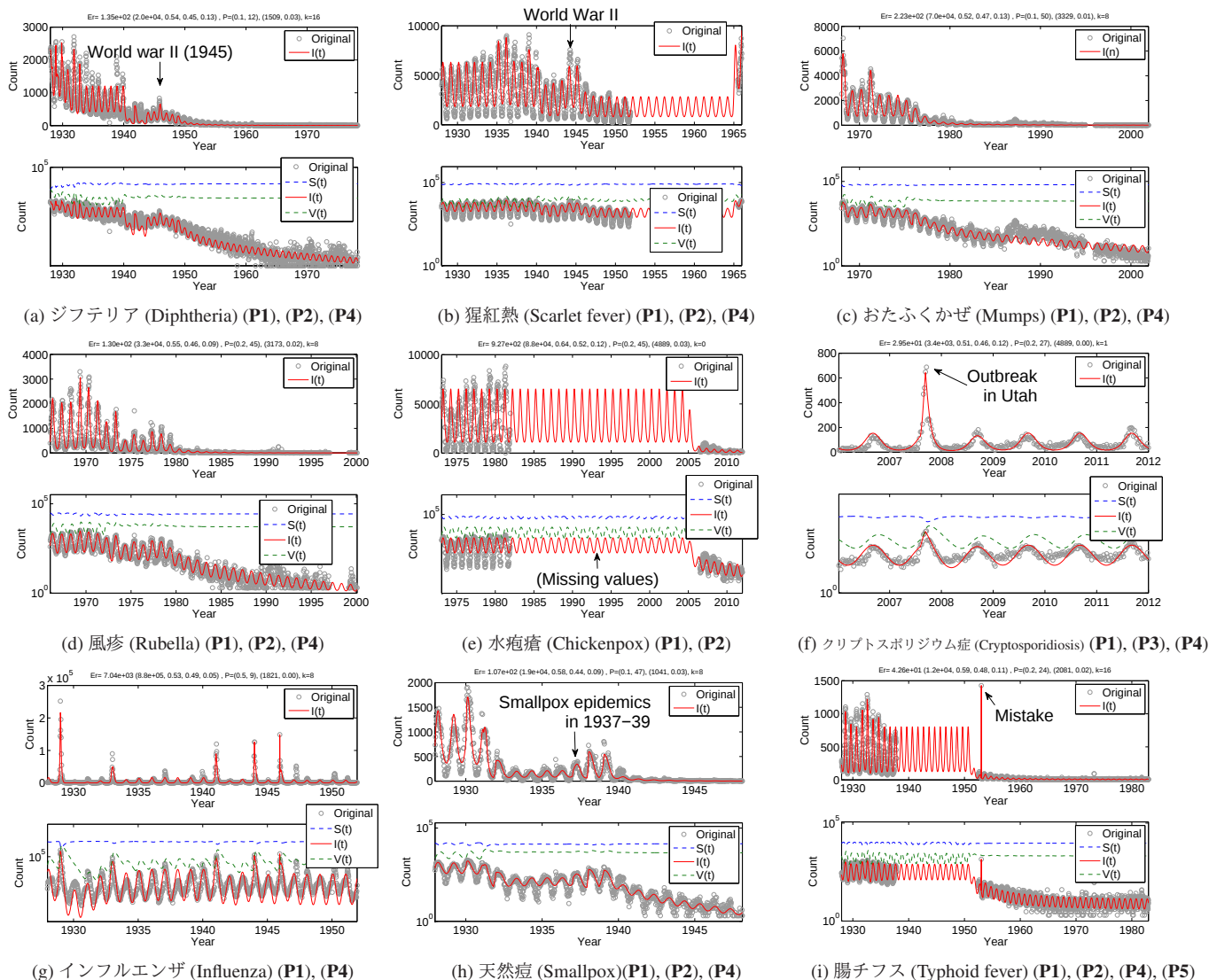


図5 主要な9種の疫病に対する FUNNEL の学習結果。

Fig. 5 Fitting results of FUNNEL for 9 diseases (global-level counts).

ケースの長さ n を変化させている。図に示すように、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。

5. 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、将来予測が挙げられる。図 10 は、2つの主要な疫病に対する予測結果を示している。ここでは、オリジナルデータの $2/3$ の長さ (黒線) を用いてモデルを学習し、その後の数年間を予測している (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 に示した通り、提案手法は AR と比較し高精度の予測を達成している。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 が外部ショックや外れ値に強い影響を受けているのに比べ、提案モデルは長期的な予測に成功している。

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、疫病のみならず、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。例えば、図 11 は、IPA^(注8) 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。例えば、(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

(注8) : IPA - IT security center: <https://www.ipa.go.jp/security/english/index.html>

図6 ワクチンのライセンス化[10]とFUNNELの結果.

Fig.6 The year of vaccine licensure [10] vs. detection.

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

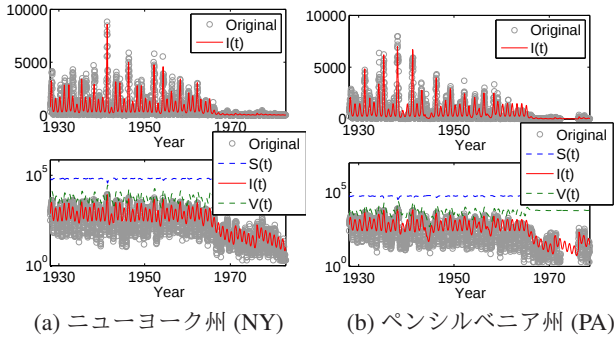


図7 麻疹のローカル(州)の疫病に対する学習結果.

Fig.7 Local-level fittings for measles.

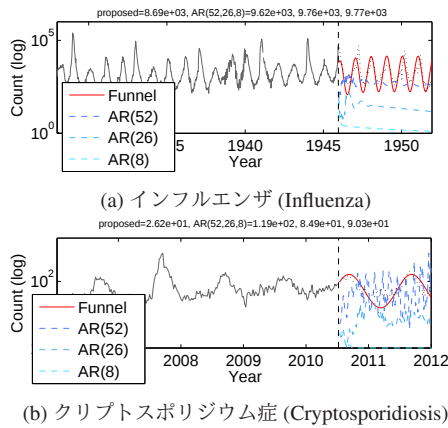


図10 2つの疫病に対する将来予測の結果.

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

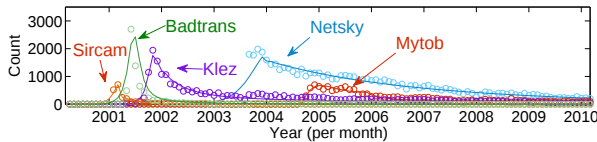


図11 コンピュータウイルスにおけるFUNNELの学習結果.

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

スで、2004年に大規模な感染が報告された。このウイルスは、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、このウイルスには、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)で伝搬するウイルスである。さらに、最近になり猛威をふるっているのが、“Koobface”、“Fbphotofake”などに代表されるソーシャルネットワークサイト上のウイルスである。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。

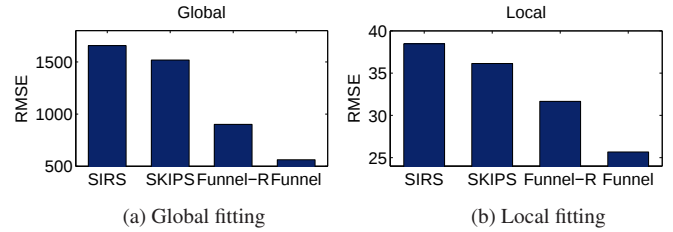


図8 FUNNELと既存手法の精度比較.

Fig.8 Fitting accuracy for the global and local sequences.

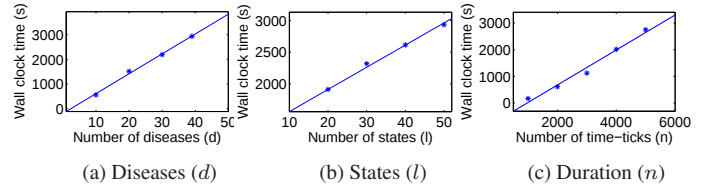


図9 FUNNELFITの計算コスト.

Fig.9 FUNNELFIT scales linearly: wall clock time vs. dataset size.

6. 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た。

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*, 61(19):348–52, 2012.
- [2] C. Böhm, C. Faloutsos, J.-Y. Pan, and C. Plant. Ric: Parameter-free noise-robust clustering. *TKDD*, 1(3), 2007.
- [3] D. CC. Smallpox in the united states: It’s decline and geographic distribution. *Public Health Reports*, 55(50):2303–2312, 1940.
- [4] Y. Matsubara, Y. Sakurai, W. G. van Panhuis, and C. Faloutsos. FUNNEL: automatic mining of spatially coevolving epidemics. In *KDD*, pages 105–114, 2014.
- [5] F. NM, G. AP, and B. RM. Ecological and immunological determinants of influenza evolution. *Nature*, 422(6930):428–33, 2003.
- [6] F. PE and C. JA. Measles in england and wales–i: An analysis of factors underlying seasonal patterns. *Epidemiol*, 11(1):5–14, 1982.
- [7] D. SF. Seasonal variation in host susceptibility and cycles of certain infectious diseases. *Emerg Infect Dis.*, 7(3):369–74, 2001.
- [8] M. SM, E. RJ, M. A, and M. P. Seasonality in six enterically transmitted diseases and ambient temperature. *Am J Trop Med Hyg.*, 2014.
- [9] L. Stone, R. Olinsky, and A. Huppert. Seasonal dynamics of recurrent epidemics. *Nature*, 446:533–536, March 2007.
- [10] W. G. van Panhuis, J. Grefenstette, S. Y. Jung, N. S. Chok, A. Cross, H. Eng, B. Y. Lee, V. Zadorozhny, S. Brown, D. Cummings, and D. S. Burke. Contagious diseases in the united states from 1888 to the present. *NEJM*, 369(22):2152–2158, 2013.