

Twitterと災害情報の関連性の同定

黒澤 僚祐^{1,a)} 松原 靖子² 櫻井 保志²

概要：ブログや Twitter をはじめとするソーシャルメディアの発展により、現実のイベントに関して情報が拡散するなどの現象が多く見られる。本論文では、大規模時系列データ間の関係抽出手法について述べる。本手法では、ある多次元時系列シーケンスと、それによって生じられる別の多次元シーケンスが与えられた時に、シーケンス間の関係を発見し、その情報をモデルとして要約、表現する。実データを用いた実験では、本手法が時系列シーケンス間の関係を直感的に示すことができる事を明らかにした。さらに、SNS ユーザがある単語について関心を持つ地域を示すことができた。

キーワード：システム同定、ソーシャルネットワーク、災害情報

Identifying the relationship between Twitter and disaster information

KUROSAWA RYOSUKE^{1,a)} MATSUBARA YASUKO² SAKURAI YASUSHI²

Abstract: Given an event composed scale of disaster and where it happen in real world, what reaction take place in the web? How do we go about describing an interest that people in the web have and classification it? In this paper we present the method of system identification between events and people in the web. Our method has following properties: (a)Intuitiveness: it shows meaningful and intuitive model; (b)Classification: it classify people under the influence of events; and (c)Periodic pattern: it extract periodic pattern. Experiment on real data demonstrate that our method can develop a connectivity model between event and people.

Keywords: System Identification, SNS, Disaster Information

1. 前書き

マーケティングなどの分野において、ユーザーがどのようなイベントに興味を持っているか、というのは非常に重要である。ユーザーは関心のあるイベントについてのブログを共有し、イベントについて Twitter^{*1}, Facebook^{*2}などで発言する。本論文では、あるイベントに対しての反応に地域差があるかどうかに着目する。

提案手法はイベントが発生した地点を元に分類した地域ごとのユーザーの関心を直感的に示す。

図 1 は Twitter における台風に関する日本の都道府県別のツイート数の推移と構成したモデルによってそれを再現したもの、台風のイベントシーケンス、都道府県での台風の位置に対する関心の大きさにより分類したもの、熊本県での台風の位置に対する関心の大きさである。提案手法はこのようにシーケンスをイベントを使用して再現でき、関心の大きさを直感的に表せるモデルを構成する。

本論文の目的は、日々発生する様々なイベントに対する人々の関心を表現することである。以下では台風と地震の例を言及するが、イベントは他の事象でもモデルを構成可能であり、データの分け方は都道府県でなく、ある基準によってユーザーを分類した際の一つ一つのクラスや年代、性別などでも可能と考えられる。本論文で取り組む問題は以下の通りである。

問題 1 事前情報として Twitter などのユーザーの特定

¹ 熊本大学 自然科学研究科
Computer Science and Electrical Engineering, University of Kumamoto, Japan

² 熊本大学

^{a)} ryosuke@st.cs.kumamoto-u.ac.jp

^{*1} <https://twitter.com/>

^{*2} <https://www.facebook.com/>

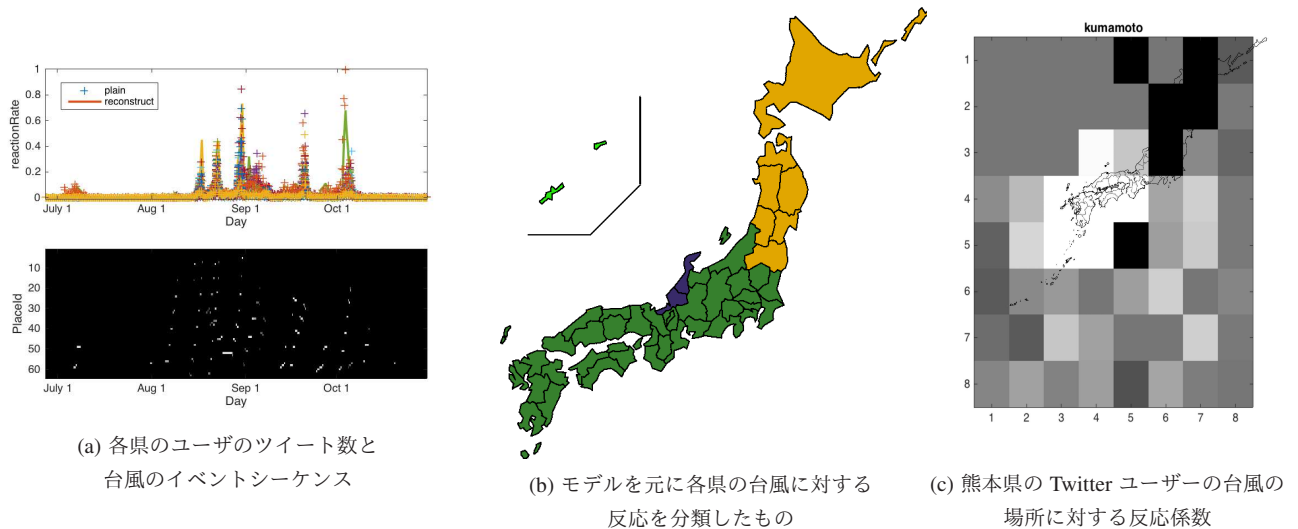


図1 提案手法の出力例

のワードについての発言数を地域で分けたシーケンス $\mathbf{Y}$ と、発生したイベントのシーケンス $\mathbf{U}$ が与えられた時、それぞれの地域がイベントに対してどのように反応するかを発見する。具体的には (1) $\mathbf{U}$ を入力とし $\mathbf{Y}$ を出力とするモデル $\mathcal{M}$ を発見し (2) $\mathcal{M}$ を元に地域ごとの反応の違いを示すパラメータを得る。

1.1 関連研究と本研究の位置付け

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 [1], [2], [11], [13], [15], [12]. 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 [6], [7], [16].

複数の時系列データの関連性に関する研究は様々な分野で行われている [10]. Evangelos E らは文献 [14] において、システム同定 (System Identification) [9][17][8] を用いて脳の各部位の関連性や外界の刺激との関連性を抽出している。

本研究で提案する手法は、(a) ある時系列データから他の時系列データに与える影響を抽出し、(b) 影響の度合いを元にシーケンスを分類し、(c) 周期性を持つ。また、本手法はシステム同定からなるため、シーケンスがパルス的でもモデルを構成することができる。Matsubara らの EcoWeb [10] は時系列データの双方向の関係性の抽出を行なっているが、パルス波の様なシーケンスはモデル化できず、Evangelos E らの GEBM [14] は周期性をモデル化できない。

2. 背景

ここで、現実でのイベント発生と Web 上での反応の関係について考えてみる。現実では事件、災害、政治、天気など多様なイベントが発生し、それらはマスコミなどによってニュースとして情報が与えられ、日本全土または一部の

地域の人々の話題となり、しばらく盛り上がりを見せたあと時間が経つにつれてイベントに関する新規の情報が少なくなり、人々が話題にしなくなる。イベントや、人々のイベントに対する関心が人々の生活する場所に依存しているとした時、その関係をどのようにモデル化すればいいだろうか。現実でのイベント発生と Web 上での反応の関係を表現するにあたり重要なのは次の要素である。(a) 潜在変数: 人々がそれぞれのタイミングで持っているイベントに対しての関心を表す。(b) 入力変数: 人々に与えられるイベントの場所や規模を表す。例えば大規模な山火事が長野県で発生する、等級3の台風が沖縄県に接近など。イベントの規模は他のイベントと比較できるように数値で表せるものが望ましい。(c) 出力変数: 関心を持った人々がどのように Web 上で発言するかを表す。

上記の要素を表現するための最もシンプルな方法として、状態方程式が挙げられる。まず、それぞれの要素をベクトルで表す。[18]

$$\mathbf{x}(t) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}, \mathbf{u}(t) = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_l \end{pmatrix}, \mathbf{y}(t) = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{pmatrix}$$

そして、それぞれは次のように表される。

$$\begin{cases} \mathbf{x}(t) = \mathbf{F}\{\mathbf{x}(t_0), \mathbf{u}(t_0, t)\} \\ \mathbf{y}(t) = \mathbf{G}\{\mathbf{x}(t_0, t), \mathbf{u}(t_0, t)\} \end{cases}$$

ここで、 $\mathbf{x}(t_0)$: 初期状態ベクトル、 $\mathbf{u}(t_0, t)$: t_0 から t までに加えられた入力ベクトルであり、 $\mathbf{F}, \mathbf{G}$ は初期状態ベクトル及び入力ベクトルの一価関数で、潜在変数の数はエネルギー蓄積素子の数に相当する。システムが線形微分方程式で表される場合には、状態方程式は状態変数の一階連立微分方程式の

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \quad (1)$$

出力方程式は

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \quad (2)$$

の形式で表される．ここで，行列 $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ は以下に示すものになる．

- $\mathbf{A}$: 係数行列 (coefficient matrix)
- $\mathbf{B}$: 制御行列 (control matrix)
- $\mathbf{C}$: 出力行列 (output matrix)
- $\mathbf{D}$: 伝達行列 (transfer matrix)

図 2 に状態方程式と出力方程式によるシステムのブロック線図を示す．

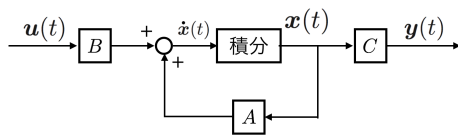


図 2 状態方程式と出力方程式による線形システムの表現 (ただし $\mathbf{x}(0_+) = \mathbf{0}$, $\mathbf{D} = \mathbf{0}$ とする)

入力ベクトル $\mathbf{u}(t)$ の各要素は，タイミング t でのイベントの発生状況を地域ごとに分け，イベントの強度を数値で表したものであり，制御行列をかけたのちに，係数行列を乗じた潜在変数 $\mathbf{x}(t)$ と足し合わせて潜在変数を更新する．タイミング t での各地域のユーザの反応 $\mathbf{y}(t)$ を得たい場合は潜在変数 $\mathbf{x}(t)$ に出力行列を乗じる．

状態方程式と出力方程式のそれぞれの行列 $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ を求めるために 4SID 法 (部分空間法: Sub Space-based State Space model IDentification method) を用いる．4SID 法は，通常の入出力データから状態空間モデルの同定を行うことができる有力なシステム同定法である．この手法は，入出力データからシステム同定を行うことができるため，特異値分解法と異なり，不安定システムに適用することができ，SVD(特異値分解)[3] や QR 分解 [4] などの数値的に安定なアルゴリズムを利用しているため，計算精度が高く，多項式ブラックボックスモデルに対する予測誤差法のように，非線形最適化計算を行うことなくシステム同定が行える．しかし予測誤差法のようにパラメータ推定値の信頼区間に関する情報を提供できないため，モデルの品質に関する情報が無いという欠点もある．[19]

3. 提案モデル

提案モデルについて述べる．

3.1 モデルの概要

提案モデルは， $\mathbf{U}$ を入力， $\mathbf{Y}$ を出力とした制御モデルになる． $\mathbf{Y}$ は時間ごとに区切った (1 時間ごとなど) 反応の数をそれぞれの地域の人口でわり，全体で正規化したものと

する． $\mathbf{U}$ はある分割方法で一つ一つのイベントの発生地点を元に分割し，さらに $\mathbf{Y}$ と同様に時間ごとに区切った時系列データに整形する． $\mathbf{U}$ は図 1(a) 下のようにスパースな行列となり，各要素はイベントの強度 (台風の等級や，地震の最大震度など) となる

3.2 ベースモデル

提案手法は次にあげる事前情報から構成される．

- $\mathbf{Y}$: 発生したイベントによって Twitter, Facebook などに投稿された関連ワード付きの発言数
- $\mathbf{U}$: 発生したイベントの位置とイベントの強さを示すシーケンス
- $\mathbf{X}$: 潜在変数

時刻 t での潜在変数 $\mathbf{x}(t)$ の状態は，時刻 t で発生したイベント $\mathbf{u}(t)$ と時刻 $t-1$ での潜在変数 $\mathbf{x}(t-1)$ によって決定され，発言数データは，行列 $\mathbf{C}$ によって潜在変数から復元される．

モデル 1 提案モデルは次の式から構成される．

$$\begin{cases} \mathbf{x}(t) = \mathbf{A}\mathbf{x}(t-1) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) \end{cases} \quad (3)$$

初期状態は $\mathbf{x}(0) = \mathbf{0}$, $|\mathbf{x}| = k$ である．

3.2.1 提案モデルの根拠

モデルのパラメータについての説明を行う，主な記号と定義を表 1 に示す．

- $\mathbf{A}$ は 1 つ前の潜在変数の状況が現在の潜在変数へどのような影響を与えるかを示す係数行列 ($k \times k$) である． k は任意に設定可能である．大抵の場合，イベントが起こった際に出力シーケンスはスパイクを示し，イベント発生後は減衰するため，潜在変数も類似の反応を示す．つまり $\mathbf{A}$ は減衰を表す行列となる．
- $\mathbf{B}$ は $\mathbf{U}$ を潜在変数 $\mathbf{X}$ に入力する時の制御行列 ($k \times l$) である．各要素の絶対値が大きいくほど，潜在変数に与える影響ないしは出力シーケンスへの影響が大きい．
- $\mathbf{C}$ は潜在変数を出力シーケンスへ変換するための出力行列 ($d \times k$) である．

ここで， $\mathbf{CB}$ は入力から出力への直接の影響を表す．これが今回の県ごとのイベント位置に対する関心を示す．

3.3 フルモデル

$\mathbf{Y}$ には周期性が存在する．例えば，イベントが夜に発生した場合と昼から夕方にかけて発生した場合では関心の大きさが同じでも後者の方が反応が小さくなる． $\mathbf{B}\mathbf{u}(t)$ によって発生したイベントによる反応は， $\mathbf{A}$ により減衰するが，これは指数的に変化するのみであり，周期性を表現することができない．

モデル 2 提案モデルは，以下の式によりユーザーの周期的な活動を表現する

表 1 主な記号と定義

記号	定義
$\mathbf{Y}$	出力シーケンス
$\mathbf{U}$	入力シーケンス
d	出力シーケンスの次元数
l	入力シーケンスの次元数
n	シーケンス長さ
$\mathcal{M}$	モデル: $\mathcal{M} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{P}\}$
$\mathbf{X}$	潜在変数
k	潜在変数の個数
$\mathbf{A}$	係数行列
$\mathbf{B}$	制御行列
$\mathbf{C}$	出力行列
$\mathbf{P}$	周期ベクトル
T	周期
$Cost_T(\mathbf{Y}, \mathbf{U}; \mathcal{M})$	$\mathcal{M}$ による $\mathbf{Y}$ の総コスト
$Cost_M(\mathcal{M})$	$\mathcal{M}$ のモデル表現コスト
$Cost_C(\mathbf{Y} \mathcal{M})$	$\mathcal{M}$ による $\mathbf{Y}$ の符号化コスト

$$\begin{cases} \mathbf{x}(t) = \mathbf{A}\mathbf{x}(t-1) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{p}(t \bmod T)\mathbf{C}\mathbf{x}(t) \end{cases} \quad (4)$$

ここで、 $\mathbf{U}, \mathbf{Y}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{X}$ はモデル 1 と同様である。

モデル 2 はモデル 1 に周期ベクトル $\mathbf{P}$ を加えたものである。 T は 1 日や 1 週間などのユーザの活動周期を示す。このモデルでは、反応を示す人の数が時刻によって変動する。深夜にイベントが発生しても反応する人が少ないためである。周期ベクトル $\mathbf{P}$ は時刻による反応の違いを反応が最大となる時間帯を 1 とした割合で示す T 個の要素を持つ行列であり、時刻ごとの補正を行う。モデル 1 の出力シーケンス $\mathbf{Y}$ に周期成分をかけることで元のシーケンスの周期性を再現する。

4. モデルパラメータの学習

与えられたシーケンス $\mathbf{Y}$ とイベント $\mathbf{U}$ からモデルを構成するために、2 節で紹介したシステム同定法である 4SID 法を用いる。

また、最適なモデルの導出のために最小記述長 (MDL: minimum description length) [5] の概念を用いる。MDL は情報理論に基づくモデル選択基準の一つであり、可逆圧縮を行うことができるが、そのものの概念だけでは本論文の目的を直接解決することはできない。そこで、与えられたシーケンス $\mathbf{Y}$ をイベント $\mathbf{U}$ で適切に表現するモデルを見つけるために、新しい符号体系を定義する。具体的には、新たな関数を用いて候補解 $\mathcal{M}$ のモデルコストを推定し、最適解を発見するためにパラメータを変更しながらモデルを計算する。

4.1 MDL

大規模時系列データを表現するための符号化スキームを導入する。直感的には、データを与えられた時のモデルの

良さは次の式で表現できる

$$Cost_T(\mathcal{M}) = cost(\mathcal{M}) + cost(\mathbf{Y}|\mathcal{M}) \quad (5)$$

$Cost(\mathcal{M})$ はモデル $\mathcal{M}$ を表現するためのコストを示し、 $Cost(\mathbf{Y}|\mathcal{M})$ は $\mathcal{M}$ が与えられた時のシーケンス $\mathbf{Y}$ の符号化のコストを示す。

4.1.1 モデル表現コスト

モデルは $\mathcal{M} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{P}\}$ で構成される。モデルの表現コストは以下の要素から構成される。

- 潜在変数の遷移行列 $\mathbf{A}(k \times k)$
- 潜在行列への入力行列 $\mathbf{B}(k \times l)$
- 潜在変数から再現シーケンスへの変換行列 $\mathbf{C}(d \times k)$
- 周期ベクトル $\mathbf{P}(T)$

それぞれの要素は浮動小数なので $Cost_M(\mathcal{M}) = c_F \times (k^2 + k \times l + d \times k + T)$ となる。 c_F は符号小数コストである。この中でユーザの地域分割数 d 、ユーザ活動周期 T はデータごとに固定であるので、 k, l つまりシステム同定の潜在要素数と発生イベントの位置依存の分割法を変動させる。

4.1.2 時系列シーケンスの符号化コスト

本手法ではシステム同定を用いてシーケンス $\mathbf{Y}$ の時系列パターンを表現するが、ここで重要なのは、推定したモデルが $\mathbf{Y}$ を正しく表現しているかを判断する指標の導入である。モデル $\mathcal{M}$ が与えられた際の $\mathbf{Y}$ 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

$$Cost_C(\mathbf{Y}, \mathbf{U}|\mathcal{M}) = \log_2 \frac{1}{P(\mathbf{Y}|\mathcal{M})} = -\log P(\mathbf{Y}|\mathcal{M}) \quad (6)$$

ここで $P(\mathbf{Y}|\mathcal{M})$ は $\mathbf{Y}$ の尤度を示す。

4.1.3 符号化コスト関数

候補解 $\mathcal{M}$ が与えられた時の $\mathbf{Y}$ の符号長は次のように表現される

$$\begin{aligned} Cost_T(\mathbf{Y}, \mathbf{U}; \mathcal{M}) &= Cost_T(\mathbf{Y}, \mathbf{U}; \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{P}) \\ &= Cost_M(\mathcal{M}) + Cost_C(\mathbf{Y}|\mathcal{M}) \end{aligned} \quad (7)$$

5. 実験

提案手法の有効性を検証するため、人工データおよび実データを用いた実験を行なった。本実験は、以下の諸問題に取り組む

- (1) 人工データを用いた構成モデル $\mathcal{M}$ の予測精度の検証
- (2) 実データを用いてモデル $\mathcal{M}$ を構成し、ユーザの反応の地域依存性とイベントの与える影響の地域依存性の抽出

5.1 データセット

本論文では以下のデータセットに対し評価実験を行なった

- 人工データ: 表 1 のデータおよび潜在パラメータの大きさとイベント発生頻度 r を元にデータを作成し、解

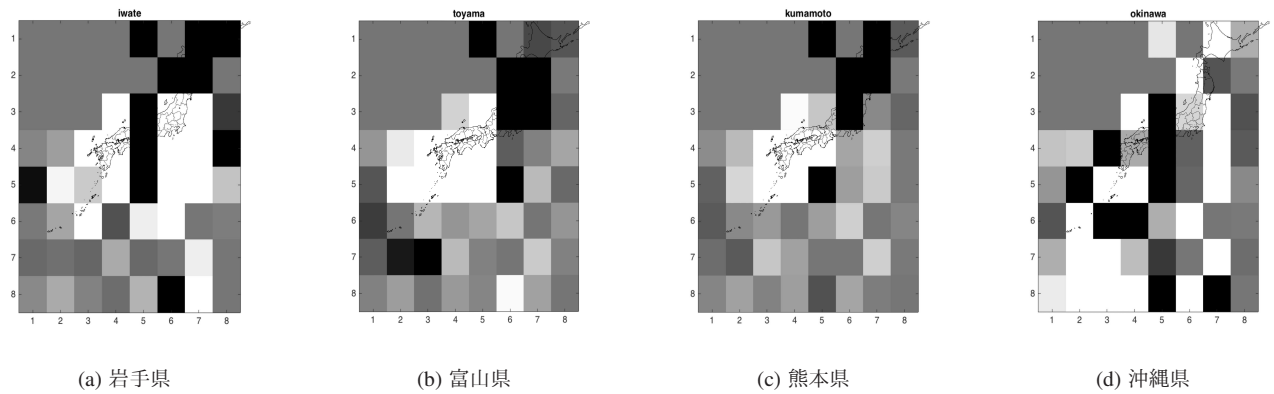


図3 各県の台風の位置による反応の違い

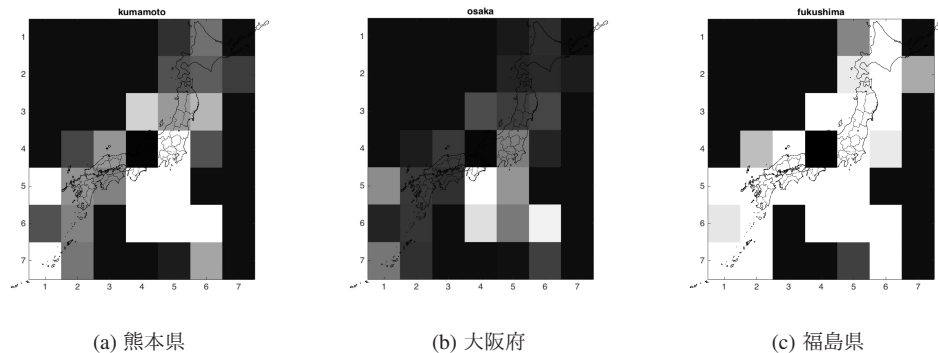


図4 各県の地震の位置による反応の違い

析を行う。解析方法は、長さ n のシーケンスを二分割し、前半部分の入力シーケンス U と出力シーケンス Y を用いて周期ベクトル P 、行列 A, B, C を求め、後半部分で入力シーケンス U と求めたモデルを用いて出力シーケンスの予測を行い、実データとの誤差を求める。比較手法として AR を用いる。

- 台風: ユーザの反応データ Y は Twitter の 2016 年 6 月 27 日から 10 月 30 日までの”台風”というワードが入ったツイートを取得し、集計したものを使用する。ユーザの所属地域は Twitter のユーザのプロフィール欄に記入された地名をもとに判断する。記入されていないものについては使用しない。イベント U は気象庁台風データベース^{*3}より 2016 年 6 月 27 日から 10 月 30 日に発生した台風のデータを使用する。イベントの所属地域の判定方法は、日本列島周囲の座標を指定し、四角形の領域を設定したのちに、その四角形領域を分割した何番目にイベントが所属するかによって判定する。その際に横の分割数を x 、縦の分割数を y とする ($l = xy$)。分割方法の例と各地域の台風の通過回数は図 6(a) に示す。集計間隔は 6 時間である。これは気象庁台風データベースの台風の観測間隔がそれぞれの台風や時間帯によって異なり、すべての台風で共通する間隔が 6 時間だったためである。
- 地震: 台風のデータと同様に集計、所属地域の判定を行

う。使用データは Twitter、地震データ共に 2016 年 6 月 10 日から 11 月 29 日まで^{*4}を使用し、集計間隔は 1 時間である。

イベントの発生地域の判定の際に、格子状に分けられたそれぞれの領域の中でイベントが存在しないものがある。その際には該当領域のシーケンスは使用せず、コスト計算の際にもシーケンスを削った後のモデルの規模を用いることとする。

5.2 精度の検証

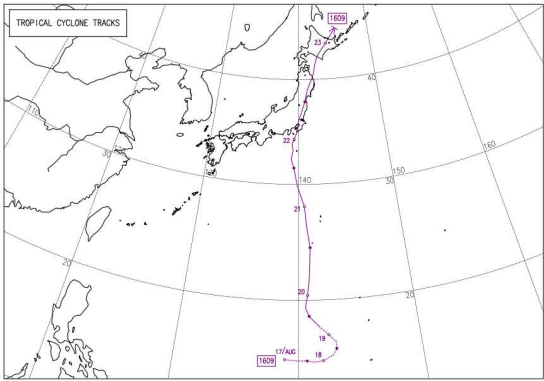
表2 精度検証

n	l	d	k	T	r	本手法	AR
						$R[\%]$	k $R[\%]$
1024	10	5	4	12	0.3	32.52	12 3.26
1024	10	5	4	12	0.3	32.52	24 8.28
2048	10	5	4	12	0.3	49.74	12 4.75
1024	20	5	4	12	0.3	43.34	12 19.59
1024	10	10	4	12	0.3	39.94	12 9.94
1024	10	5	8	12	0.3	35.59	12 8.75
1024	10	5	4	24	0.3	36.51	12 14.37
1024	10	5	4	12	0.6	52.56	12 24.65

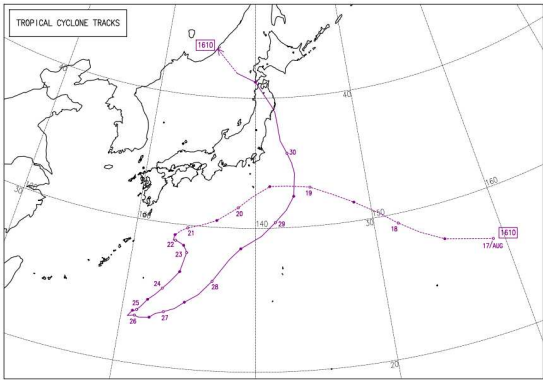
表2 は人工データ作成時の各パラメータと本手法を用いて予測したシーケンスの再現率 R 、AR の次数、AR を用い

^{*3} <http://www.jma.go.jp/jma/index.html>

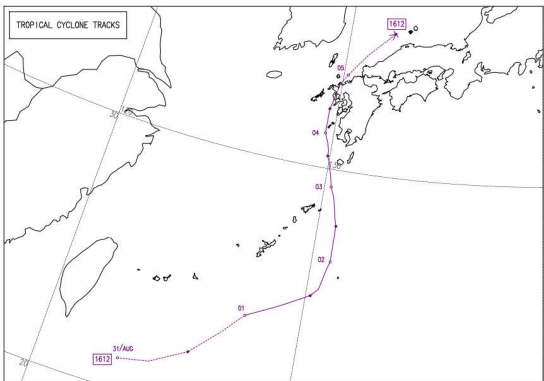
^{*4} <http://www.data.jma.go.jp/svd/eqdb/data/shindo/index.php>



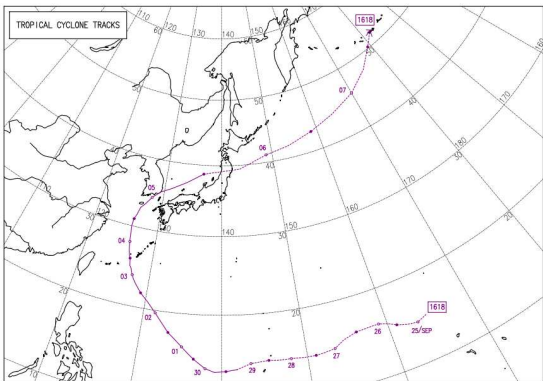
(a) 台風 9 号



(b) 台風 10 号



(a) 台風 12 号



(b) 台風 18 号

図 5 2016 年の台風の軌道

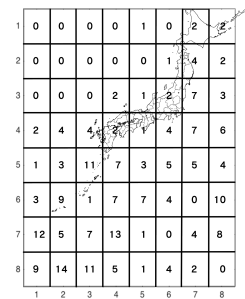
出典:<http://www.data.jma.go.jp/fcd/yoho/typhoon/route.map/bstv2016.html>

表 3 地震および台風の解析結果

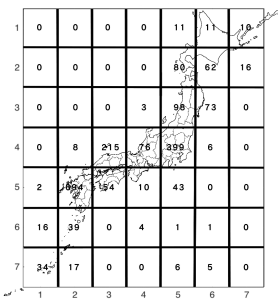
	n	l	d	k	T	r	$R[\%]$
地震	4128	32	47	1	24	0.0104	8.9
台風	504	48	47	3	4	0.0096	47.8

て予測したシーケンスの再現率である．表の中で再現率が高いのは出力シーケンス，入力シーケンス，入力シーケンスの次元，入力シーケンス内のイベントが増加した場合である．つまり，差分方程式内の行列 A, B, C を特定するための標本となるイベントや入力シーケンスの要素が多い時に予測の精度が上がることになり，より適切なモデルを構成できると言える．

表 3 は表 2 と同様に実データの内容と，構成したモデルと入力シーケンスを用いて再現した時の再現率 R を示したものである．実データではイベント発生頻度 r が人工データに比べ低く，予測をするにはデータが不足していたため，全ての入力シーケンスを用いて出力シーケンスを再現したものの再現率を示している．表 3 を見ると，地震および台風のイベント発生頻度は等しく，入力シーケンスが大きい (=実際のイベント発生回数が多い(図 6)) 地震の方が再現率は高くなると予想されるが，実際は 8.9%と低い．これは，日本列島に接近した台風はほぼ全て人々の目にニュースとして留まるのに対し，地震は発生回数は多いものの震度が

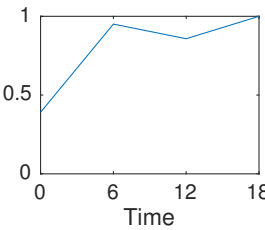


(a) 各地域の台風の通過回数
($x = 8, y = 8$)

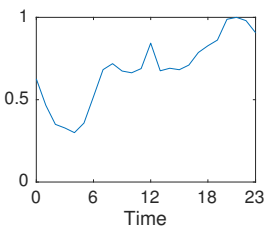


(b) 各地域の地震の発生回数
($x = 7, y = 7$)

図 6 各地域のイベント発生回数



(a) 台風



(b) 地震

図 7 ワードが含まれるツイート数の周期性

低い場合は人々の関心を集めず，ニュースにもならないためであると考えられる．

5.3 台風

事象の起こった地域に対する関心によって各ユーザの地域を分類した時、地域が近いものは共通の関心を持つことが多い。MDL で最適なモデルを選択したところ $(x, y, k) = (8, 8, 3)$ となった。図 1(b) は構成したモデルを節 3.2.1 の **CB** を計算したのち、各県のユーザの台風の位置と関心の関係の差を距離として分類し色分けしたものの、図 3 は分類したクラスタの台風の通過位置に対する反応の例である。図 5 は図 3 の各県に関連する 2016 年に発生した台風の軌道を示したものである。(a) 岩手県:台風 9 号, 台風 10 号 (7,11 号) が南や南東方向から接近したため、該当地域に対する関心が高い。(b) 富山県:台風 18 号が東シナ海方向から接近したため、該当地域に対する関心が高い。(c) 熊本県:台風 18 号, 台風 12 号 (16 号) が西や南西方向から接近し、該当地域に対する関心が高いが、熊本県以東にも関心を示している。これは通過した後の台風に対しても関心が残っていたためであると考えられる。(d) 沖縄県:他の県が主に自県の周囲に関心が高いのに対し、沖縄県では関心が高い範囲が特異なものになっている。これは九州地方の他の県よりも例年の台風上陸数が多く^{*5}、台風に対して敏感であるためと考えられる。

また、図 7(a) に台風に関するツイート数の周期を示す。0 時から 6 時までのツイートが少なく、18 時から 24 時までのツイートが多いことは 1 日のライフスタイルと合致する。

5.4 地震

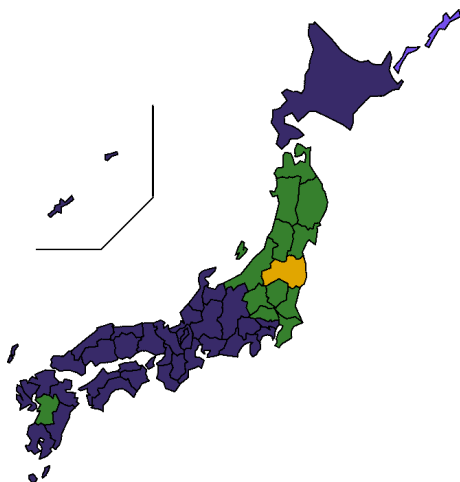


図 8 モデルを元に各県の地震に対する反応を分類したもの

MDL で最適なモデルを選択したところ $(x, y, k) = (7, 7, 1)$ となった。台風に対して潜在変数の数が少ないが、これは地震が発生してから地震に関する話題が収束す

るまでの時間が 1 時間未満であり、人々の地震に対する関心が長く続かないためであると考えられる。

図 8 は各県のユーザの地震の発生位置とツイート数の関係の差を距離として分類し色分けしたものの、図 4 は分類したクラスタの地震の発生位置に対する反応の例である。熊本県、福島県など最近地震が起こった地域やその周囲のユーザは地震に対して大きな関心を示すという結果が得られた。しかし鳥取県では 10 月 21 日に震度 6 の地震が発生したにもかかわらず関心が低い。これはモデルの再現率が低く、関心を表現しきれていないためであると考えられる。

また、図 7(b) に地震に関するツイート数の周期を示す。4 時のツイートが最も少なく、21 時のツイートが最も多いことは 1 日のライフスタイルと合致する。

6. おわりに

本論文では時系列データ間の関係抽出手法を提案した。本手法は SNS 発言数と現実のイベントとの関連性を発見し、SNS ユーザの地域やイベント発生地域ごとの関連性の違いを抽出し表現することができた。

今後は災害データ以外のデータや、複数のジャンルのイベントの SNS 発言数との関連性を抽出し、関連性を元にイベントの分類を行うなどの検証を行っていく。

参考文献

- [1] G. E. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, Englewood Cliffs, NJ, 3rd edition, 1994.
- [2] L. Chen and R. T. Ng. On the marriage of lp-norms and edit distance. In *VLDB*, pages 792–803, 2004.
- [3] G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische mathematik*, 14(5):403–420, 1970.
- [4] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.
- [5] P. D. Grünwald. *The minimum description length principle*. MIT press, 2007.
- [6] A. Jain, E. Y. Chang, and Y.-F. Wang. Adaptive stream resource management using kalman filters. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 11–22. ACM, 2004.
- [7] L. Li, B. A. Prakash, and C. Faloutsos. Parsimonious linear fingerprinting for time series. *PVLDB*, 3(1):385–396, 2010.
- [8] L. Ljung. System identification. In *Signal Analysis and Prediction*, pages 163–173. Springer, 1998.
- [9] L. Ljung. *System identification*. Wiley Online Library, 1999.
- [10] Y. Matsubara, Y. Sakurai, and C. Faloutsos. The web as a jungle: Non-linear dynamical systems for co-evolving online activities. In *Proceedings of the 24th International Conference on World Wide Web*, pages 721–731. ACM, 2015.
- [11] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *KDD*, pages 6–14, 2012.
- [12] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data*

^{*5} <http://www.data.jma.go.jp/fcd/yoho/typhoon/statistics/average/average.html>

- mining*, pages 6–14. ACM, 2012.
- [13] S. Papadimitriou, J. Sun, and C. Faloutsos. Streaming pattern discovery in multiple time-series. In *VLDB*, pages 697–708, 2005.
 - [14] E. E. Papalexakis, A. Fyshe, N. D. Sidiropoulos, P. P. Talukdar, T. M. Mitchell, and C. Faloutsos. Good-enough brain model: Challenges, algorithms, and discoveries in multisubject experiments. *Big Data*, 2(4):216–229, 2014.
 - [15] Y. Sakurai, S. Papadimitriou, and C. Faloutsos. Braid: Stream mining through group lag correlations. In *SIGMOD*, pages 599–610, 2005.
 - [16] Y. Tao, C. Faloutsos, D. Papadias, and B. Liu. Prediction and indexing of moving objects with unknown motion patterns. In *SIGMOD*, pages 611–622, 2004.
 - [17] M. Verhaegen and V. Verdult. *Filtering and system identification: a least squares approach*. Cambridge university press, 2007.
 - [18] 松瀬貞規. 基礎制御工学. 数理工学社, 2013.
 - [19] 足立修一. システム同定の基礎. 東京電機大学出版局, 2009.