Vol.2018-MPS-117 No.10
2018/3/1

多項分布型レジームスイッチング検出による周期的時系列データの単純化

山岸 祐己[1,a)]　岩﨑 清斗[2]　斉藤 和巳[1,b)]

概要：状態変化をともなう周期的時系列データは，変化が複雑な場合，無駄な情報量が多く，そのまま可視化しても解釈が難しいため，それらを単純化することを目的とした技術は重要であると言える．よって，本論文では，観測データの状態の確率分布はなんらかの理由で時間と共に変化していると考え，その変化をレジームスイッチングに基づくタイムラインとして表現する手法を提案する．提案手法は，各レジームにおける観測データは多項分布に従っていると仮定し，それらの尤度を最大化することによって，モデルパラメータとスイッチング時刻を推定する．また，提案手法の有効性を検証するため，平均的なモデルとの乖離が大きい時系列データを可視化する実験を行う．

キーワード：時系列データ，レジームスイッチング，最尤推定，貪欲法，局所探索法

Simplification of Periodic Time Series Data by Multinomial Distribution Type Regime Switching Detection

Yuki Yamagishi[1,a)]　Kiyoto Iwasaki[2]　Kazumi Saito[1,b)]

Abstract: Periodic time series data with categorical condition changing is hard to visualize due to lots of useless information when the condition fluctuations are complicated. Thus, we assume that the probability distribution of conditions in observed data usually changes over time due to several reasons, we propose a method for simplifying and visualizing such complicated time series data as timelines based on switching regimes. Namely, by assuming that fundamental condition changing in each regime obeys a multinomial distribution model, we first estimate the switching time steps and the model parameters by maximizing the likelihood of the observed time series data and then produce their timelines. For verification of the effectiveness of proposed method, we make a visualization experiment with the time series data which extremely deviated from the average model.

Keywords: time series data, regime switching, maximum likelihood estimation, greedy search, local search

1. はじめに

状態変化をともなう周期的時系列データは，変化が複雑な場合，無駄な情報量が多く，そのまま可視化しても解釈が難しいため，それらを単純化することを目的とした技術は重要であると言える．今回扱うような時系列データの研究では，現時点の状況解析や将来予測に焦点を当てているものもあるが，今回の研究内容は，過去に何が起き，どのような変化をしていたかということに焦点を当てた研究 [1], [2] と類似する．本研究では，レジームスイッチングの検出問題を定式化し，推定されたレジームスイッチングに基づいた時系列データのタイムラインを生成し，単純化して可視化する手法を提案する．ここで，各レジームにおける観測データの基本生成パターン多項分布に従ってい

[1] 静岡県立大学
University of Shizuoka, 52-1 Yada, Suruga-ku, Shizuoka 422–8526, Japan
[2] 静岡県工業技術研究所
Industrial Research Institute of Shizuoka Prefecture, 2078 Makigaya, Aoi-ku, Shizuoka 421-1221, Japan
a) yamagissy@gmail.com
b) k-saito@u-shizuoka-ken.ac.jp

ると仮定し，スイッチングが起こるタイムステップとモデルパラメータは，観測された時系列データの尤度を最大化することによって推定する．

本研究は，Kleinberg [1] や Swan と Allan [2] と同様に，回顧的 (retrospective) な枠組みによる時系列データからの構造抽出を目的としている．たとえば，Kleinberg の研究は，文書ストリーム内のトピックの出現をバーストとして表現し，その入れ子構造を推定することによって，ある期間におけるトピックのアクティビティを要約し，それらの分析を容易にしている．この Kleinberg の手法は，バーストが自然に状態遷移として現れる隠れマルコフモデルを使用しており，電子メールメッセージの階層構造を識別することができている．周期的時系列データにおける適応を考えると，観測時刻の間隔（データの時間密度）が変化しているものについては，既存のバースト検出技術 [1] とともに，ウィンドウに基づく手法 [3] や複数ストリームを対象とした手法 [4] なども適応可能であるが，観測時刻の間隔がほぼ一定のものについては，これら既存手法の有効性は低いことが予想される．さらに，既存のバースト検出技術は，単一情報のバーストを検出するものであり，複数情報とその分布の変化に着目していないため，状態変化などの複数情報の傾向変化は検出できるとは限らない．一方，Swan と Allan の研究は，仮説検定に基づいた時間経過による特徴出現モデルを使用し，コーパス内の主要トピックに対応する情報をクラスタとして生成することに成功している．本研究も同様に，過去に起こった現象を理解するという目的を持っているが，あくまでレジームスイッチングに基づく変化を仮定しているため，このような研究のモチベーションとも離れている．

ここで，今回扱うようなレジームスイッチング検出は，ノベルティ検出や外れ値検出 [5] で使用される技術のような，機械学習の分野で広く研究されている異常検出や変化点検出の典型的技術とは大きく異なることを強調しておく．たとえば，異常検出に使用される統計的手法は，与えられたデータに対して統計モデル（インスタンスの大多数は正常であるという仮定）を適合させ，統計的検定によって未知のインスタンスがこのモデルに属するか否かを決定するものである．このような手法では，適用された統計的検定に基づき，学習モデルから生成される確率が低いインスタンスは異常とされる．本研究は，時間で変化するモデルパラメータをレジームスイッチングとして扱っているため，これら典型的異常検出技術とは方向性が異なる．同様の方向性を持つ従来アプローチとしては，経済分野におけるレジームスイッチングモデルの研究 [6] があげられるが，これらの研究はガウシアンモデルに大きく依存している．意思決定支援の分野でも，オンラインレビューシステムにおける不正な評価を検出するための技術 [7] がいくつか開発されているが，これらの方法は明確に異常検出技術の領域に分類される．

2. 提案手法

2.1 問題設定

複数の状態に変化し得る時系列データを $\mathcal{D} = \{(s_1, t_1), \cdots, (s_N, t_N)\}$ とする．ここで，s_n と t_n は，J カテゴリの状態と n 番目の観測時刻をそれぞれ表す．$|\mathcal{D}| = N$ を観測数とすると，$t_1 \leq \cdots \leq t_n \leq \cdots \leq t_N$ となる．n はタイムステップとし，$\mathcal{N} = \{1, 2, \cdots, N\}$ をタイムステップ集合とする．また，k 番目のレジームの開始時刻を $T_k \in \mathcal{N}$ ，$\mathcal{T}_K = \{T_0, \cdots, T_k, \cdots, T_{K+1}\}$ をスイッチングタイムステップ集合とし，便宜上 $T_0 = 1$，$T_{K+1} = N + 1$ とする．すなわち，$T_1, \cdots, T_K$ は推定される個々のスイッチングタイムステップであり，$T_k < T_{k+1}$ を満たすとする．そして，$\mathcal{N}_k$ を k 番目のレジーム内のタイムステップ集合とし，各 $k \in \{0, \cdots, K\}$ に対して $\mathcal{N}_k = \{n \in \mathcal{N}; T_k \leq n < T_{K+1}\}$ のように定義する．なお，$\mathcal{N} = \mathcal{N}_0 \cup \cdots \cup \mathcal{N}_K$ である．

いま，各レジームの状態分布が J カテゴリの多項分布に従うと仮定する，$\boldsymbol{p}_k$ を k 番目のレジームにおける多項分布の確率ベクトルとし，$\mathcal{P}_K$ はそれら確率ベクトルの集合，つまり $\mathcal{P}_K = \{\boldsymbol{p}_0, \cdots, \boldsymbol{p}_K\}$ とすると，$\mathcal{T}_K$ が与えられたときの対数尤度関数は以下のように定義できる．

$$L(\mathcal{D}; \mathcal{P}_K, \mathcal{T}_K) = \sum_{k=0}^{K} \sum_{n \in \mathcal{N}_k} \sum_{j=1}^{J} s_{n,j} \log p_{k,j}. \tag{1}$$

ここで，$s_{n,j}$ は $s_n \in \{1, \cdots, J\}$ を

$$s_{n,j} = \begin{cases} 1 & \text{if } s_n = j; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

のように変換したダミー変数である．各レジーム $k = 0, \cdots, K$ と各状態 $j = 1, \cdots, J$ に対する式 (1) の最尤推定量は $\hat{p}_{k,j} = \sum_{n \in \mathcal{N}_k} s_{n,j} / |\mathcal{N}_k|$ のように与えられる．これらの推定量を式 (1) に代入すると以下の式が導ける．

$$L(\mathcal{D}; \hat{\mathcal{P}}_K, \mathcal{T}_K) = \sum_{k=0}^{K} \sum_{n \in \mathcal{N}_k} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{k,j}. \tag{3}$$

したがって，スイッチングタイムステップの検出問題は，式 (3) を最大化する $\mathcal{T}_K$ の探索問題に帰着できる．

しかし，式 (3) だけでは $\mathcal{T}_K$ の導入によってどれだけ尤度が改善したかという直接的な評価をすることができない．この問題において，レジームスイッチングを考慮しないときの尤度からの改善度合いを評価することは重要であるため，尤度比最大化問題として目的関数を構築し直す．もし，レジームスイッチングのような変化が存在しない，すなわち $\mathcal{T}_0 = \emptyset$ と仮定すると，式 (3) は

$$L(\mathcal{D}; \hat{\mathcal{P}}_0, \mathcal{T}_0) = \sum_{n \in \mathcal{N}} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{0,j}, \tag{4}$$

となる．ここで，$\hat{p}_{0,j} = \sum_{n \in \mathcal{N}} s_{n,j}/N$ である．よって，K 個のスイッチングを持つ場合と，スイッチングを持たない場合の対数尤度比は

$$LR(\mathcal{T}_K) = L(\mathcal{D}; \hat{\mathcal{P}}_K, \mathcal{T}_K) - L(\mathcal{D}; \hat{\mathcal{P}}_0, \mathcal{T}_0). \quad (5)$$

のように与えられる．最終的に，この問題は上記の $LR(\mathcal{T}_K)$ を最大化する $\mathcal{T}_K$ の探索問題に帰着できる．

2.2 解法

式 (5) を網羅的に解くと最適解が保証されるが，計算量が $O(N^K)$ となってしまうため，ある程度大きい N に対して $K \geq 3$ となってしまうと，実用的な計算時間で解くことができない．したがって，我々は任意の K について解くための高速な解法を提案する．以下では，まず貪欲法 (A1) と局所探索法 (A2) を説明し，更にそれらを組み合わせた提案解法について説明する．

2.2.1 貪欲法

まず，貪欲法 (A1) の手順について説明する．このアルゴリズムは，バックトラッキングをしないデータの 2 分割の繰り返しである．つまり，既に選択された $(k-1)$ 個のスイッチングタイムステップ $\mathcal{T}_{k-1}$ を固定したまま k 番目のスイッチングタイムステップ T_k を $\mathcal{T}_{k-1}$ に新たに追加することを繰り返す．ただし，モデルとしての適当なスイッチングタイムステップ数 K を選択するため，アルゴリズムの終了条件として最小記述長原理 (MDL) [8] を採用する．基本的には最小記述長原理に従って自動的に K の数を選択させるが，K の数を意図的に決めて出力することも可能である．貪欲法アルゴリズムの手順は以下となる．

A1-1. $k = 1, \mathcal{T}_0 = \emptyset$ のように初期化する．
A1-2. $T_k = \arg\max_{t_n \in \mathcal{T}} \{LR(\mathcal{T}_{k-1} \cup \{t_n\})\}$ を探索する．
A1-3. $\mathcal{T}_k = \mathcal{T}_{k-1} \cup \{T_k\}$ のように更新する．
A1-4. もし $-L(\mathcal{D}; \hat{\mathcal{P}}_k, \mathcal{T}_k) + (J-1)k \log N/2 > -L(\mathcal{D}; \hat{\mathcal{P}}_{k-1}, \mathcal{T}_{k-1}) + (J-1)(k-1) \log N/2$ なら $\mathcal{T}_{k-1}$ を $\mathcal{T}_K$ として出力して終了する（K が意図的に決まっている場合は $k = K$ のとき $\mathcal{T}_K$ を出力して終了する）．
A1-5. $k = k+1$ とし，A1-2 に戻る．

ここで，A1-3 での $\mathcal{T}_k$ の各スイッチングタイムステップは，$T_{k-1} < T_k$ を満たすように再インデックスする．明らかに，このアルゴリズムの計算量は $O(NK)$ と高速であるため，大規模な N に対しても実用的な計算時間で結果を得ることが可能である．しかし，先ほども説明したように，このアルゴリズムはバックトラッキングを行わないため，プアーな局所解に陥ってしまうことが危惧される．

2.2.2 局所探索法

次に，局所探索法 (A2) について説明する．このアルゴリズムは，A1 で得られた解 $\mathcal{T}_K$ から始まり，スイッチングタイムステップの改善を 1 つずつ試みるものである．つまり，k 番目のスイッチングタイムステップ T_k を一度取り去り，残った $\mathcal{T}_K \setminus \{T_k\}$ を固定して，よりよい尤度を得られる T'_k を探索することを $k = 1$ から K まで繰り返す．ここで，$\cdot \setminus \cdot$ は集合差を表す．もし，すべての k $(k = 1, \cdots, K)$ に対してスイッチングタイムステップの置換が行われない，すなわち，すべての k に対して $T'_k = T_k$ ならば，これ以上の改善は望めないとして処理を終了する．局所探索法のアルゴリズムは以下となる．

A2-1. $k = 1, h = 0$ のように初期化する．
A2-2. $T'_k = \arg\max_{t_n \in \mathcal{T}} \{LR(\mathcal{T}_K \setminus \{T_k\} \cup \{t_n\})\}$ を探索する．
A2-3. もし $T'_k = T_k$ ならば $h = h+1$ とし，さもなければ $h = 0$ として $\mathcal{T}_K = \mathcal{T}_K \setminus \{T_k\} \cup \{T'_k\}$ のように更新する．
A2-4. もし $h = K$ ならば $\mathcal{T}_K$ を出力して終了する．
A2-5. もし $k = K$ ならば $k = 1$, さもなければ $k = k+1$ とし，A2-2 に戻る．

明らかに，このアルゴリズムの計算量は改善が終わらない限り増え続けてしまうが，ある程度大規模な問題に対しても，せいぜい貪欲法アルゴリズムの計算量 $O(NK)$ の数倍程度で終了することを我々は既に実験によって示している [9]．

2.2.3 提案解法

もし，計算量を最低限に抑えることを目的として，単純に貪欲法アルゴリズムと局所探索法アルゴリズムを組み合わせると，

C1. A1 で $\mathcal{T}_K$ を得る．
C2. A2 で $\mathcal{T}_K$ を改善する．

となる．確かに，これだけでも十分な近似解が期待できるが，スイッチングタイムステップ数 K が貪欲法アルゴリズムによって決定されてしまうため，問題に対して不適切なスイッチングタイムステップ数のまま局所改善を行ってしまう恐れが大いにある．したがって我々は，不必要なスイッチングタイムステップは極力追加せず，且つ必要なスイッチングタイムステップは極力追加することを目的とした，アルゴリズムの反復的な組み合わせを提案する．提案解法の手順は以下となる．

P1. A1-1 から処理を開始する．
P2. A1-3 の処理後に $k \geq 2$ ならば，$\mathcal{T}_k$ を $\mathcal{T}_K$ として出力する．
P3. $\mathcal{T}_K$ を A2 で改善し，改善した $\mathcal{T}_K$ を $\mathcal{T}_k$ として出力する．

P4. A1-4 から処理を再開させ，P2 へ戻る．

この手順では，スイッチングタイムステップが追加される度に局所探索法アルゴリズムを行うため，更なる計算量の増加が予想されるが，ある程度大規模な問題に対しても，せいぜい貪欲法アルゴリズムの計算量 $O(NK)$ の数倍から十数倍程度で終了することを我々は既に実験によって示している [9]．また，上記の解法は，多項分布のレジームスイッチングを想定した人工データにおける実験で，極端に短いレジームの場合を除いて，真の分布に基いてパラメータを設定した Kleinberg の手法 [1] と同等，もしくはそれ以上の検出精度を示している [10].

2.3 可視化

上記の解法によって得られた推定タイムステップ集合を $\hat{\mathcal{T}}_K$ とし，各状態 j について，タイムステップ $n \in \mathcal{N}_k$ $(0 \leq k \leq K)$ における確率関数を $\hat{p}_j(n) = \hat{p}_{k,j}$ のように考える．そして，レジームスイッチングを視覚的に分析するために，J カテゴリの確率関数を同時にプロットしたタイムラインを生成することを考える．ただし，各カテゴリを同等に扱うために，実際の観測時刻 t_n ではなく，タイムステップ n に関する確率をプロットすることに注意されたい．

3. 実験

3.1 実験設定

実験で用いる現実データは，goo 天気 [*1] のデータである．今回，全国 57 箇所の地上気象観測所における 1961 年から 2016 年の天気情報を対象データとした．ただし，那覇，石垣島，宮古島，南大東島の観測所は 1964 年から 2016 年のみのデータ，舞鶴の観測所は 1961 年から 2012 年のみのデータとなっており，全ての観測所において閏年の 2 月 29 日は対象としていない．実験時には各観測所の 1 年ごとのデータを $\mathcal{D}$ として提案手法を適応しているが，観測所や年によって日ごとの観測回数が異なるため，観測数 $|\mathcal{D}| = N$ はデータごとに異なることに注意されたい．なお，各観測所において出現確率が 1% 未満の天気状態は欠損扱いとしており，カテゴリ J には含まれていない．この設定において，各観測所は $J = 3$（晴れ，曇，雨），$J = 4$（晴れ，曇，雨，雪），$J = 5$（晴れ，曇，雨，雪，霧）の 3 グループに分かれたため，グループごとに検証を行う．

ここで，今回のような 1 年周期のデータの状態確率の平均モデルを考える．いま，各年度 $y = 1, \cdots, Y$ の時系列データとそのタイムステップを $\mathcal{D}_y = \{\cdots, (s_{y,n}, t_{y,n}, d_{y,n}), \cdots\}$, $\mathcal{N}_y \subseteq \mathcal{N}$ とし，$s_{y,n}$ をダミー変数化したものを $s_{y,n,j}$ とする．$d_{y,n} \in \{1, 2, \cdots, X\}$ は日付を表しており，年度 y, 日付 x におけるタイムステップ集合を $\mathcal{N}_{y,x} = \{n \in \mathcal{N}_y; d_{y,n} = x\}$ とする．日付 x における状態 j の平均確率は

$$\bar{p}_{x,j} = \frac{\sum_{y=1}^{Y} \sum_{n \in \mathcal{N}_{y,x}} s_{y,n,j}}{\sum_{y=1}^{Y} |\mathcal{N}_{y,x}|}, \tag{6}$$

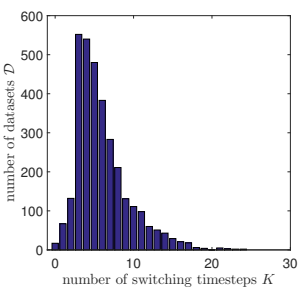
となるため，この $\bar{p}_{x,j}$ を状態確率の平均モデルとして扱う．このとき，年度 y における平均モデルとの平均確率誤差は

$$E_y = \frac{\sum_{n \in \mathcal{N}_y} 1.0 - \bar{p}_{d_{y,n},j}}{|\mathcal{N}_y|}, \tag{7}$$

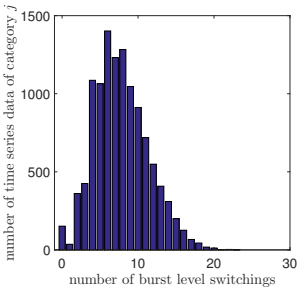
のように計算できる．すなわち，E_y が大きいほど，平均モデルとの乖離が大きいことを示す．

4. 実験結果

カテゴリ数 J のグループごとに E_y の降順ランキングを作成し，提案手法による上位の可視化結果を検証する．ここで，比較手法として，観測範囲（15 タイムステップ）ごとの状態出現数に基づく可視化結果と，Kleinberg [1] のバースト検出手法に基づく可視化結果を用いる．なお，Kleinberg のバースト検出手法は，1-カテゴリの時系列データにしか適応できないため，時系列データを状態ごとに J 個に分けて独立に適応している．また，可視化の尺度を統一するために，観測範囲ごとの状態出現数は範囲内の確率として変換し，Kleinberg のバーストレベルはレベルの持続範囲をレジームとみなして提案手法と同様に確率変換する．今回，提案手法が検出したスイッチングタイムステップ数 K の度数分布は図 1(a) のようになったため，Kleinberg のバーストレベルのスイッチング数も，これに近いものになるようパラメータを $s = 1.2, \gamma = 0.2$ に設定した（図 1(b)）．



(a) 提案手法のスイッチングタイムステップ数 K の度数分布



(b) Kleinberg のバーストレベルのスイッチング数の度数分布 $(s = 1.2, \gamma = 0.2)$

図 1 スイッチング数の分布比較

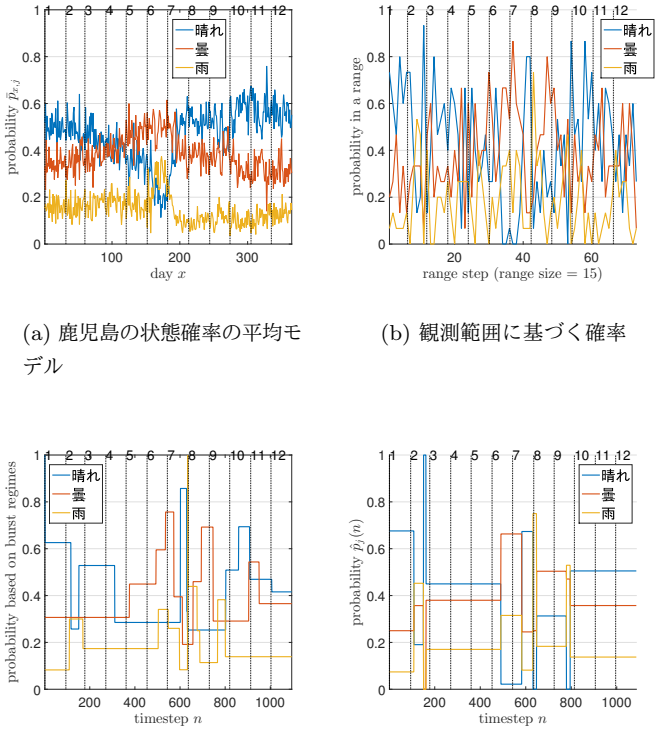
$J = 3$（晴れ，曇，雨）グループの観測所における平均確率誤差 E_y の上位を表 1 に示す．1 位の鹿児島の状態確率の平均モデルは図 2(a) のようになり，鹿児島 (2014) の観測範囲に基づく確率，Kleinberg の手法に基づく確率，提案手法に基づく確率は，それぞれ図 2(b), 2(c), 2(d) のよ

[*1] https://weather.goo.ne.jp/

うになった．なお，図上の縦の実線は月の区切りを意味している．図 2(b) より，観測範囲に基づく確率は，変動が激しく，全体的に乱雑であるため，平均モデルとの差異がどこで生じているかを把握することが容易ではない．図 2(c) より，Kleinberg の手法に基づく確率は，各状態が単純な階段関数になっているため，状態ごとで平均モデルとの差異を見つけることが容易になっている．しかし，それぞれ独立に生成された階段関数であるため，全体的に差異が生じた期間を把握することが容易ではない．図 2(d) より，提案手法に基づく確率は，Kleinberg の手法と同様，各状態が単純な階段関数になっているため，状態ごとで平均モデルとの差異を見つけることが容易であり，また，全ての状態確率がレジームごとに変動するため，全ての状態において差異が生じた期間を把握することも容易である．

rank	observatory	year y	E_y
1	鹿児島	2014	0.6063
2	石垣島	1989	0.6034
3	石垣島	1966	0.6013
4	鹿児島	1993	0.6006
5	銚子	1964	0.6002

表 1 $J = 3$ （晴れ，曇，雨）グループの平均確率誤差 E_y の上位



(a) 鹿児島の状態確率の平均モデル

(b) 観測範囲に基づく確率

(c) Kleinberg の手法に基づく確率

(d) 提案手法に基づく確率

図 2 $J = 3$ グループ 1 位:鹿児島 (2014) の可視化結果

$J = 4$（晴れ，曇，雨，雪）グループの観測所における平均確率誤差 E_y の上位を表 2 に示す．1 位の函館の状態確率の平均モデルは図 3(a) のようになり，函館 (2013) の観測範囲に基づく確率，Kleinberg の手法に基づく確率，提案手法に基づく確率は，それぞれ図 3(b), 3(c), 3(d) のようになった．これらの図より，おおむね $J = 3$ のときと同様の考察ができるが，提案手法のレジームスイッチングが長期に渡って起こらない期間があるため，この間においては詳細な比較ができなくなってしまっている．これは，提案手法の出力がモデルの尤度に依存しているため，基本的には状態確率分布に変化が無いと考えるべきだが，状態カテゴリ J が増えるとスイッチング感度が弱まることにも起因するため，一概に確率分布に変化が無いとは言い切れない．より詳細な比較を行いたい場合は，意図的にスイッチングタイムステップ数 K を設定する必要がある．例えば，図 3(e) のように $K = 15$ に設定して出力させれば，より詳細な状態確率の変化が分かるようになる．

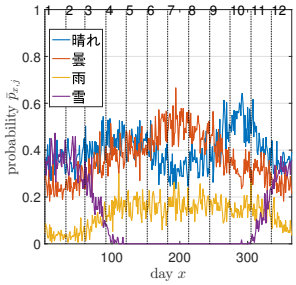
rank	observatory	year y	E_y
1	函館	2013	0.6591
2	富山	1972	0.6555
3	富山	1976	0.6541
4	富山	1983	0.6532
5	富山	1980	0.6528

表 2 $J = 4$（晴れ，曇，雨，雪）グループの平均確率誤差 E_y の上位

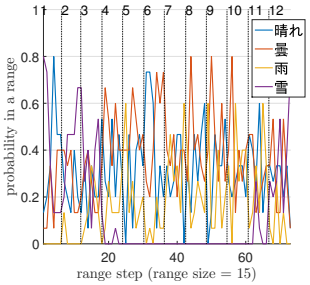
$J = 5$（晴れ，曇，雨，雪，霧）グループの観測所における平均確率誤差 E_y の上位を表 3 に示す．1 位の室蘭の状態確率の平均モデルは図 4(a) のようになり，室蘭 (2013) の観測範囲に基づく確率，Kleinberg の手法に基づく確率，提案手法に基づく確率は，それぞれ図 4(b), 4(c), 4(d) のようになった．これらの図より，おおむね $J = 3$ のときと同様の考察ができるが，先程と同様，状態カテゴリ J が増えると，提案手法のスイッチング感度が弱くなる恐れがあるため，より詳細な比較を行いたい場合は，図 4(e) のように $K = 15$ に設定するなどして出力する必要がある．

rank	observatory	year y	E_y
1	室蘭	2013	0.6621
2	室蘭	1966	0.6556
3	室蘭	2010	0.6516
4	室蘭	2002	0.6502
5	室蘭	2012	0.6502

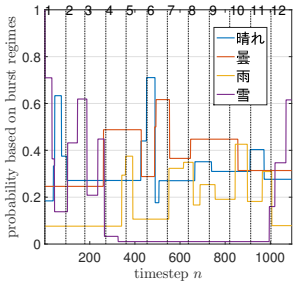
表 3 $J = 5$（晴れ，曇，雨，雪，霧）グループの平均確率誤差 E_y の上位



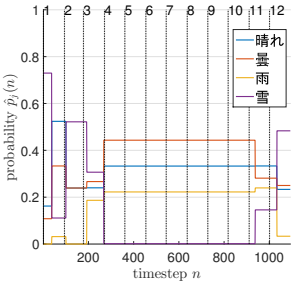
(a) 函館の状態確率の平均モデル



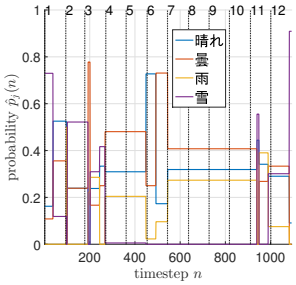
(b) 観測範囲に基づく確率



(c) Kleinberg の手法に基づく確率

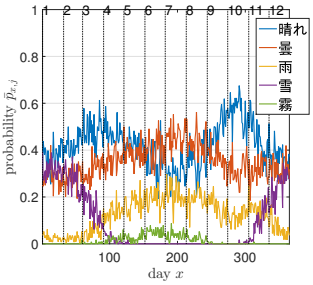


(d) 提案手法に基づく確率

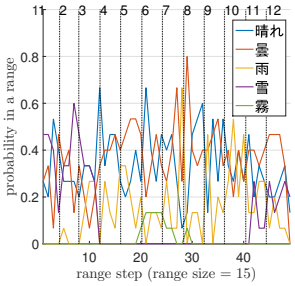


(e) 提案手法に基づく確率 ($K = 15$ に設定)

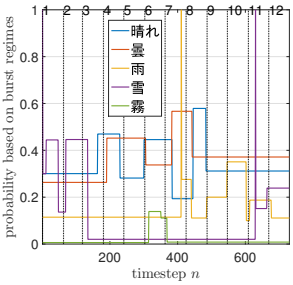
図 3 $J = 4$ グループ 1 位:函館 (2013) の可視化結果



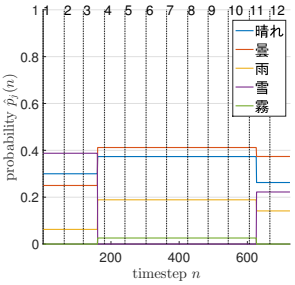
(a) 室蘭の状態確率の平均モデル



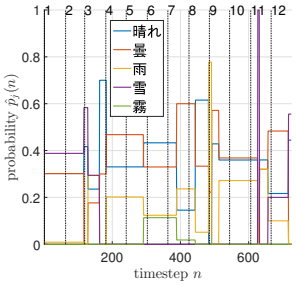
(b) 観測範囲に基づく確率



(c) Kleinberg の手法に基づく確率



(d) 提案手法に基づく確率



(e) 提案手法に基づく確率 ($K = 15$ に設定)

図 4 $J = 5$ グループ 1 位:室蘭 (2013) の可視化結果

5. おわりに

本論文では，多項分布型レジームスイッチング検出による周期的時系列データの単純化とその可視化を提案した．提案手法は，各レジームにおける観測データは多項分布に従っていると仮定し，観測された時系列データに対してレジームスイッチングモデルを適応することで単純な可視化結果を生成した．Kleinberg の手法に基づく確率との比較では，平均モデルとの差異を把握できるかという点において，状態ごとで平均モデルとの差異を見つけることが容易であるだけでなく，全ての状態において差異が生じた期間を把握することも容易であることを示した．また，スイッチングが検出されにくい場合でも，意図的にスイッチングタイムステップ数を設定することにより，出力結果を調整できることも示した．

謝辞

本研究は，科研費基盤研究 (C) 15K00429 の支援を受けて行ったものである．

参考文献

[1] Kleinberg, J.: Bursty and hierarchical structure in streams, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, pp. 91–101 (2002).

[2] Swan, R. and Allan, J.: Automatic generation of overview timelines, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pp. 49–56 (2000).
[3] Zhu, Y. and Shasha, D.: Efficient Elastic Burst Detection in Data Streams, *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, pp. 336–345 (2003).
[4] Sun, A., Zeng, D. and Chen, H.: Burst Detection from Multiple Data Streams: A Network-based Approach, *IEEE Transactions on Systems, Man, & Cybernetics Society, Part C*, Vol. 40, pp. 258–267 (2010).
[5] Chandola, V., Banerjee, A. and Kumar, V.: Anomaly Detection: A Survey, *ACM Comput. Surv.*, Vol. 41, No. 3, pp. 15:1–15:58 (2009).
[6] Kim, C. J., Piger, J. and Startz, R.: Estimation of Markov regime-switching regression models with endogenous switching, *Journal of Econometrics*, Vol. 143, pp. 263–273 (2008).
[7] Josang, A., Ismail, R. and Boyd, C.: A survey of trust and reputation systems for online service provision, *Decision support systems*, Vol. 43, pp. 618–644 (2007).
[8] Rissanen, J.: Modeling by Shortest Data Description, *Automatica*, Vol. 14, No. 5, pp. 465–471 (1978).
[9] Yamagishi, Y., Okubo, S., Saito, K., Ohara, K., Kimura, M. and Motoda, H.: A Method to Divide Stream Data of Scores over Review Sites, *Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence (PRICAI '14)*, pp. 791–800 (2014).
[10] Yamagishi, Y. and Saito, K.: Visualizing Switching Regimes Based on Multinomial Distribution in Buzz Marketing Sites, *Proceedings of the 23rd International Symposium on Methodologies for Intelligent Systems (ISMIS '17)*, pp. 385–395 (2017).