ウミガメの遊泳パターンを捉えた時系列のARIMAモデルに基づく分割とクラスタリング

Naoto K. Inoue*

Abstract

ウミガメの福祉評価を行うためには、遊泳パターンの継続時間と変化を定量化する必要がある。この目的で、これまでに筆者は形状解析によってウミガメの輪郭形状を二次元時系列に圧縮した[1]。本論では、取得した二次元時系列に対し時系列解析とクラスタリング手法を用いることで、遊泳パターンの客観的な分割と分類を試みた。このとき、計算が軽く柔軟なモデルである AR モデルを仮定するが、クラスタリングで用いられる時系列間の非類似度のうち AR モデルをベースとしたものには、Piccolo (1990), Maharaj (1996), そして Kalpakis and Putagunta (2001) によって提案された指標が存在する。そこで、まずこれらの指標のレビューを行った。そして、ウミガメの輪郭形状から取得した時系列データに対して、AR モデルを仮定した分割・分類を行い、形成されるクラスタの特徴を調べた。レビューの結果、クラスタリング手法と相性が良く、結果の解釈がしやすい Piccolo と Kalpakis and Putagunta の 2 つの指標が適当であると判断し、これらの指標を用いてクラスタリングを行った。結果として形成されたクラスタ間の遊泳パターンに明瞭な違いは見られず、現段階では分類精度が高くないと判断したが、これは二次元時系列が互いに独立という仮定が原因の一つである。それゆえ、二次元時系列間の相関を考慮した MAR モデルを解析のベースとすることで、分類精度を向上させることができると考えられる。

*Depertment of Science, Kobe University, Hyogo, Japan. Email:226s405s@stu.kobe-u.ac.jp

[1] この内容はまとめて論文にしているので、近日中に発表します

1 導入

時系列クラスタリングは、パターンや傾向の似た時系列を分類する手法であり、株取引や売上の動向を調べる、血液分析の結果の変動から健康状態を評価するといった目的で用いられている。時系列クラスタリングを行うには、時系列間の非類似度を決定する必要があるが、分類対象の時系列に応じた非類似度の選択はクラスタリングの精度に大きく影響を及ぼす。それゆえ、これまでにさまざまな非類似度が提案されてきた。

Liao (2005) [6] は、時系列間の非類似度を、生データベース、特徴ベース、モデルベースという 3 つのカテゴリに分けてレビューを行った。ここで、生データベースは二つの時系列の値の単純なユークリッド距離や、カルバックライブラー情報量[2]、ガウスノイズを仮定した時の誤差など、生データそのものから導出されるものである。また、特徴ベースは、時系列間で相互相関関数を計算したり、スペクトル変換、あるいはウェーブレット変換[3]を行った上でその距離を測るものである。そして、モデルベースは、後述する ARIMA モデルをはじめ、正規混合モデル、p 値など、時系列を生成したモデルを仮定し、その仮定のもとで非類似を測定するものである。

Liao (2005) [6] の 3 つの分類に対し、Montero and Vilar (2015) [8] は非類似度を、モデルフリー、モデルベース、複雑さベース、予測ベースの 4 つに分けてレビューを行った。ここで、モデルフリーを構成する要素は Liao の生データベース・特徴ベースと似ており、モデルベースもほとんど同じ定義である。一方、複雑さベースには、コルモゴロフの複雑性[4]や、パーミュテーション分布に基づく複雑指標[5]などが位置付けられている。また、予測ベースは Montero and Vilar (2015) [8] が提案した新たな非類似度であり、時系列間の予測値の距離を指標としたものである。

本論では、さまざまな非類似度の中で、ARIMA(自己回帰和分移動平均) モデルベースの手法に着目する。ARIMA モデルは、時系列の周期成分、トレンド、確率的変動、ランダム要

[2]情報理論におけるデータ間の距離や、確率論における確立分布間の距離の指標としてよく用いられる。

[3]一言で言うと、時間スケールを考慮したフーリエ変換。現象のスケールの変化を許容した上で周波数特性を抽出することができる。

[4]情報理論における文字列の複雑性の指標。プログラムにおいて、特定の文字列を記述可能な最小の長さ。

[5]二つの時系列の値を全てランダムに入れ替える作業を全通り繰り返した際に生成される分布。仮説検定をモチベーションとして考案された。

素を捉えることが可能な時系列モデルである。このモデルを利用する理由は以下の 2 点である。1 つ目は、本論で扱う時系列が持つ周期特性を検出する必要がある点である。本論で扱う時系列は、ウミガメの輪郭座標の変化を波として定量化したものであり、大まかには前肢と後肢の動きを反映したものである。前肢は遊泳において動力を得る役割を果たすため、振幅が大きく周期も長い運動を行う一方で、後肢は振幅が小さく周期も短い。このことから、本論で扱う時系列には遊泳パターンごとに異なる周期成分が卓越すると考えられ、それを分類可能なモデルを用いる必要がある。

2 つ目の理由は、時系列の分割を行う際に AR モデル (ARIMA の単純化) を利用するためである。筆者の研究目的はウミガメの遊泳パターンの分類であり、これを達成するにはクラスタリングを行う前に、数時間にわたる時系列を定常時系列[6]に分割する必要がある。この分割を行う際、各区間を異なる AR モデルに従う時系列と考えると、AR モデルの当てはまりの良さ (AIC; 赤池情報量基準) を用いて時系列を分割することが可能になる。それゆえ、クラスタリングにおいても時系列の背景に AR(IMA) モデルを仮定することで、解析手法に一貫性を持たせる必要がある。

本論では、まず 2.2 節において AR モデルを用いた時系列の分割手法を解説する。次に、2.3 節において ARIMA モデルを仮定したモデルベースの指標のレビューを行う。ここでは、提案された年度順に、Piccolo (1990) [9]、Maharaj (1996) [7]、Kalpakis and Putagunta (2001) [4] の指標を扱う。そして、実際に分割した短い時系列に対してレビューの結果選んだ非類似度を用いたクラスタリングを行い、その精度を評価した。

2 手法

2.1 材料

本論では、ウミガメの輪郭形状の変化に対して形状解析を行うことで抽出した 2 本の時系列 pc1、pc2 を用いる (動画 1 の timeseries を参照のこと)。ここで、pc1 が正の時は両前肢を広げた姿勢を表し、負のときは丸まっている姿勢を表す。一方、pc2 が正の時は右前肢だけ横に

[6]時間が経っても同じ確立分布に従う時系列のこと。厳密には強定常と言われる。

開いた姿勢を表し、負のときは左前肢だけ横に開いた姿勢を表す (動画 1 の pc plot を参照のこと、pc plot における薄い形状は、pc1 と pc2 の値のみを用いて復元した輪郭座標である)。それゆえ筆者は pc1 と pc2 の変化をそれぞれ、体軸を中心に非対称的な漕ぎ運動 (rowing) と、対称的な羽ばたき運動 (flapping) の特徴量として捉え、時系列データとして 2.2 の分割手法と、2.3 のクラスタリング手法を適用した。

2.2 AR モデルによる分割

筆者は、pc1 と pc2 の時系列に対して AR モデルを適用し、分割前に比べて分割後の時系列のモデルの当てはまりがよくなる点を網羅的に探すことで、各個体で 1 時間の時系列を短い時系列に分割した。以下の分割法は Kitagawa (2020) [5] を参考に実装した。

2.2.1 AR モデルと最小二乗法

y_n を n 番目の時系列データとすると、m 次の AR(自己回帰) モデルは、

$$y_n = \sum_{i=1}^{m} a_i y_{n-i} + \epsilon, \tag{1}$$

と書くことができる。ここで ϵ は平均 0、分散 σ^2 の正規分布に従う確率変数である。a_i は一般的な回帰における回帰係数に相当することから、AR モデルは自身の過去のデータの加重平均 $\sum_{i=1}^{m} a_i y_{n-i}$ に正規誤差を加えることで現在のデータが生成されることを仮定したモデルであるといえる。このとき時系列の $M(> m)$ 番目から N 番目までに注目すると、回帰式は $N - M$ 次元ベクトル y と $(N - M) \times m$ 行列 Z

$$y = \begin{bmatrix} y_{M+1} \\ y_{M+2} \\ \vdots \\ y_N \end{bmatrix}, \quad Z = \begin{bmatrix} y_M & y_{M-1} & \cdots & y_{M-m} \\ y_{M+1} & y_M & \cdots & y_{M-m+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N-1} & y_{N-2} & \cdots & y_{N-m} \end{bmatrix}, \tag{2}$$

を用いて、

$$y = Za + \epsilon, \tag{3}$$

と書くことができる。

この回帰モデルにおいて、回帰係数 a、分散 σ^2、説明変数 x が与えられている時、対数尤度[7]は、

$$l(a_1, a_2, \cdots, a_m, \sigma^2) = \sum_{n=M+1}^{N} \log p(y_n \mid a_1, a_2, \cdots, a_m, \sigma^2, y_{n-1}, y_{n-2}, \cdots, y_{n-m}), \tag{4}$$

と与えられる。ここで、$\theta = (a_1, a_2, \cdots, a_m, \sigma^2)$ とすると、データの正規性から、

$$l(\theta) = -\frac{N-M}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{n=M+1}^{N} \left(y_n - \sum_{i=1}^{m} a_i y_{n-i} \right)^2, \tag{5}$$

となる。この対数尤度が極大となる σ は、

$$\frac{\partial l(\theta)}{\partial \sigma^2} = -\frac{N-M}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=M+1}^{N} \left(y_n - \sum_{i=1}^{m} a_i y_{n-i} \right)^2, \tag{6}$$

を 0 とおくことにより、

$$\hat{\sigma}^2 = \frac{1}{N-M} \sum_{n=M+1}^{N} \left(y_n - \sum_{i=1}^{m} a_i y_{n-i} \right)^2 \tag{7}$$

となる。この分散の推定値を (5) 式に代入すると、対数尤度は

$$\begin{aligned} l(a_1, a_2 \cdots, a_m) &= -\frac{N-M}{2} \log\left(2\pi\hat{\sigma}^2\right) - \frac{N-M}{2} \\ &= \log \hat{\sigma}^2 + \text{Const}, \end{aligned} \tag{8}$$

と与えられる。このことから、対数尤度の最小化は分散の最小化によって達成されることがわかる。

2.2.2 ハウスホルダー法

分散は、(3) の行列表現より、

$$|\epsilon|^2 = |y - Za|^2 \tag{9}$$

[7]確率過程に基づく当てはまりの良さの指標である。

と書くことができる。ここで $|A|$ は A のユークリッドノルムを表す。分散の最小化のためにこの式を微分して 0 とすることで、回帰係数 a の推定値は、

$$\hat{a} = (Z^T Z)^{-1} Z^T y, \tag{10}$$

と得られる。

式 (10) はいわゆる最小二乗法であるが、この逆行列の計算は時系列が長くなればなるほど計算量が膨大となる。そこで、$(N-M)\times(N-M)$ の直行行列 U を用いた変換

$$|y - Za|^2 = |U(y - Za)|^2 = |Uy - UZa|^2 \tag{11}$$

を考える。a の最尤推定値には影響がないことに注意するとよい。ここで、UZ を扱いやすい形に変換してやることができれば、計算を簡単にすることができる。

以下では、U による**ハウスホルダー変換**を考える。まず、行列 Z の右側にデータのベクトル y を付加した $(N-M)\times(m+1)$ 行列

$$X = [Z \mid y], \tag{12}$$

を考える。この行列に対して適当な U をかけると、$(N-M)\times(m+1)$ の上三角行列に変換することができる。

$$UX = [UZ \mid Uy] = S = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,m} & s_{1,(m+1)} \\ 0 & s_{2,2} & \cdots & s_{2,m} & s_{2,(m+1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & s_{m,m} & s_{m,(m+1)} \\ 0 & 0 & \cdots & 0 & s_{(m+1),(m+1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & s_{(N-M),(m+1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \tag{13}$$

このとき、1からm列までがUZに、$m+1$列がUyに対応するため、

$$|Uy-UZa|^2=\left\|\begin{bmatrix} s_{1,(m+1)} \\ s_{2,(m+1)} \\ \vdots \\ s_{m,(m+1)} \\ s_{(m+1),(m+1)} \\ \vdots \\ s_{(N-M),(m+1)} \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,m} \\ 0 & s_{2,2} & \cdots & s_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{m,m} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \right\|^2 \tag{14}$$

となる。これは、

$$|Uy-UZa|^2=\left\|\begin{bmatrix} s_{1,(m+1)} \\ s_{2,(m+1)} \\ \vdots \\ s_{m,(m+1)} \end{bmatrix} - \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,m} \\ 0 & s_{2,2} & \cdots & s_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{m,m} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \right\|^2 + s^2_{(m+1),(m+1)} + \cdots + s^2_{(N-M),(m+1)}, \tag{15}$$

と書き直すことができる。ここでユークリッドノルムの外はaに関係しないため、ユークリッドノルムの中を0にすればよい。つまり、aの最尤推定値は

$$\begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,m} \\ 0 & s_{2,2} & \cdots & s_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{m,m} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} s_{1,(m+1)} \\ s_{2,(m+1)} \\ \vdots \\ s_{m,(m+1)} \end{bmatrix} \tag{16}$$

の解として求めることができる。例えば、下から2行は

$$s_{(m-1),(m-1)}a_{(m-1)} + s_{(m-1),m}a_m = s_{(m-1),(m+1)}, \quad s_{m,m}a_m = s_{m,(m+1)}, \tag{17}$$

となるため、$\hat{a}_m = s_{m,(m+1)}/s_{m,m}$ であり、これを代入することで、1 つ前を

$$\hat{a}_{(m-1)} = \frac{s_{(m-1),(m-1)} - s_{(m-1),m}\hat{a}_m}{s_{(m-1),(m-1)}}, \tag{18}$$

と求めることができる。これを一般化すると、

$$\hat{a}_i = \frac{s_{i,m+1} - s_{i,m}\hat{a}_m - \cdots - s_{i,(i+1)}\hat{a}_m}{s_{i,i}}, \tag{19}$$

と書くことができる。このとき、(15) は

$$\hat{\sigma}_m^2 = \frac{1}{N-M}\sum_{j=m+1}^{N-M} s_{j,m+1}^2 \tag{20}$$

となり、モデルの AIC(当てはまりの良さの指標) は、

$$\begin{aligned} \text{AIC}_m &= -2 \times l(a_1, a_2 \cdots, a_m) + 2 \times \text{モデルのパラメータ数} \\ &= (N-M)(\log 2\pi\hat{\sigma}^2 + 1) + 2(m+1), \end{aligned} \tag{21}$$

と計算することができる。

ハウスホルダー変換の優れた点は 2 つ存在する。1 つ目は、ひとたび m 次の AIC を求める際に行列 S を計算しておけば、

$$\hat{\sigma}_j^2 = \frac{1}{N-M}\sum_{i=j+1}^{N-M} s_{j,m+1}^2 \tag{22}$$

と j 次のモデルの分散を求めることで、j 次のモデルの AIC が計算できる点である。2 点目は、時系列を加えた際の AIC の変化は、新たなデータベクトル y' と行列 Z' により作成した X' を S に対して列方向に結合し、新たに得た S' に対してハウスホルダー変換を行うことで、逐次的に計算できる点である。これらの点から、ハウスホルダー変換を用いた最小二乗法は、他の手法に比べて計算が非常に軽い。

ハウスホルダー変換を用いた最小二乗法を用いて、pc1 と pc2 時系列を分割する。具体的には、pc1 や pc2 の区間時系列 $\boldsymbol{y_k} = y_{k,1}, y_{k,2}, \cdots, y_{k,N}$ に対し、候補点 $y_{k,c}, \cdots, y_{k,C}$ で時系列を分割した際の AIC を計算する。次に、候補点の中から最も AIC の低い、すなわち AR モデルの当てはまりの良い点 $y_{k,s}$ をこの区間の分割点として決定する。そして、次に分割を行

う区間時系列は、前の区間で選択された分割点が $y_{k,s}$ であることから、$y_{n+1,1} = y_{k,s} + 1$ と設定する。これを繰り返すことで、時系列全体を区間時系列に分割する。本論では、ウミガメの最小の遊泳行動継続時間を 10 秒以上と考え、N には 40 秒を設定し、分割候補点は 40 秒を (10,20,10) に分けた中間の 20 秒の中から選択した[8]。

2.3 AR モデルを仮定した時系列クラスタリング

2.2 節の操作によって分割した区間時系列 $\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_k}$ を、各区間の AR 係数を特徴量として分類することを考える。

2.3.1 Piccolo (1990) の指標

Piccolo は、時系列 $\boldsymbol{y_p}, \boldsymbol{y_q}$ から求められた AR 係数 $\boldsymbol{a_p} = a_{p,1}, a_{p,2}, \cdots$ と $\boldsymbol{a_q} = a_{q,1}, a_{q,2}, \cdots$ に関して、

$$d_{pic}(\boldsymbol{y_p}, \boldsymbol{y_q}) = \left\{ \sum_{j=1}^{\infty} (a_{p,j} - a_{q,j}) \right\}^2, \tag{23}$$

を時系列間の非類似度として提案した。この指標は以下に述べる特徴を持つ。

1-pic. 古典的な距離の定義 (非負性、対称性、三角不等式が成立する) を満たす点

2-pic. 定常時系列に対しては必ず導出することができる点

3-pic. AR 係数が 0 であっても適用可能である点

4-pic. 時系列のばらつきを考慮しないため、振幅のスケールに依存しない点

5-pic. 片方の時系列の AR 係数が全て 0 のとき (すなわちホワイトノイズであるとき)、距離はもう一方の時系列の係数の絶対値となる点

6-pic. AR 係数に基づくため、現象の時間スケールに依存しない点

7-pic. 任意の AR(j) モデルに従う時系列間の距離は有界である点

[8]本論では、1 秒 5 フレームに相当する

2.3.2 Maharaj (1996) の指標

Maharaj は、時系列 $\boldsymbol{y_p}, \boldsymbol{y_q}$ から求められた m 次の AR 係数 $\boldsymbol{a_p}$ と $\boldsymbol{a_q}$ に関して、自由度 m のカイ二乗分布に従う検定統計量を時系列間の非類似度として提案した。以下ではこの指標を導出し、特性を述べる。

2 つの時系列 $\boldsymbol{y_p}, \boldsymbol{y_q}$ を生成したプロセスの違いについての検定を行う場合に、以下の帰無仮説、対立仮説を考える。

H_0: 2 つの定常時系列を生成したプロセスが同じで $\boldsymbol{a_p} = \boldsymbol{a_q}$

H_1: 2 つの定常時系列を生成したプロセスが異なり $\boldsymbol{a_p} \neq \boldsymbol{a_q}$

Bhansall (1978) [3] は、長さが T の時系列に対し、AR 次数 m が

$$\frac{m^3}{T} \to 0, \tag{24}$$

を満たすとき、

$$\frac{1}{\sqrt{T}} \sum_{j=m+1}^{\infty} |a_{p,j}| \to 0, \tag{25}$$

となるという Berk (1974) [2] の結果を利用して、推定された AR 係数 $\hat{\boldsymbol{a_p}}$ と真の AR 係数 $\boldsymbol{a_p}$ の差が、

$$\sqrt{T}(\hat{\boldsymbol{a_p}} - \boldsymbol{a_p}) \sim \mathrm{N}(0, \sigma_p^2 R_p^{-1}(m)), \tag{26}$$

と漸近的に正規分布に従うことを導いた。σ_p^2 は時系列が含むホワイトノイズの分散であり、$R_p(m)$ は定常時系列の自己共分散行列における $m \times m$ 部分行列である。

$$R_p = \begin{bmatrix} C_0 & C_1 & \cdots & C_m & \cdots \\ C_1 & C_0 & \cdots & C_{m-1} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \cdot \\ C_m & C_{m-1} & \cdots & C_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{27}$$

ここで、

$$C_k = \frac{1}{T}\sum_{n=k+1}^{T}(a_{p,n}-\mu_p)(a_{p,n-k}-\mu_p), \tag{28}$$

はラグ k の自己共分散関数と呼ばれる。

この漸近的な性質から、2 本の時系列から推定された AR 係数は、

$$\hat{\boldsymbol{a_p}} \sim \mathrm{N}\left(\boldsymbol{a_p}, \frac{1}{T}\frac{\sigma_p^2}{R_p(m)}\right), \quad \hat{\boldsymbol{a_q}} \sim \mathrm{N}\left(\boldsymbol{a_q}, \frac{1}{T}\frac{\sigma_q^2}{R_q(m)}\right), \tag{29}$$

に従う。これより、

$$\hat{\boldsymbol{a_p}} - \hat{\boldsymbol{a_q}} \sim \mathrm{N}\left(\boldsymbol{a_p} - \boldsymbol{a_q}, \frac{1}{T}\left[\sigma_p^2 R_p^{-1}(m) + \sigma_q^2 R_q^{-1}(m)\right]\right) \tag{30}$$

となるため、

$$d_{mah}(\boldsymbol{y_p}, \boldsymbol{y_q}) = (\hat{\boldsymbol{a_p}} - \hat{\boldsymbol{a_q}})^T \left[\sigma_p^2 R_p^{-1}(m) + \sigma_q^2 R_q^{-1}(m)\right]^{-1} (\hat{\boldsymbol{a_p}} - \hat{\boldsymbol{a_q}}) \ \sim \ \chi^2(m), \tag{31}$$

が成り立つ。ここで、$\chi^2(m)$ は自由度 m のカイ二乗分布である。すなわち $d(\boldsymbol{y_p}, \boldsymbol{y_q})$ は、本節の初めに設定した仮設検定における検定統計量であり、Maharaj はこれを非類似度の指標として用いることを提案した。

Maharaj はいくつかの ARIMA モデルから生成した時系列に対してこの非類似度の指標を適用し、以下の特徴を強調した。

1-mah. 非負性と対称性を持つ点

2-mah. 全ての定常時系列に対し計算することが可能である点

3-mah. Piccolo の指標と異なり、時系列の振幅のスケールに影響を受ける点

4-mah. Piccolo の指標に比べ、クラスタリングの早い段階で類似したクラスタを形成することができる点

2.3.3 Kalpakis and Putagunta (2001) の指標

Kalpakis and Putagunta は、時系列のケプストラム (cepstral coefficients) を非類似度の指標として提案した。ケプストラムの導出は、Atal (1974) [1] を参考にした。

AR モデル、

$$y_n = \sum_{i=1}^{m} a_i y_{n-i} + \epsilon, \tag{32}$$

は $y_{n-1} = By_n$ を満たす時間シフトオペレータ B によって

$$y_n = \sum_{i=1}^{m} a_i B^i y_n + \epsilon \tag{33}$$

と表現することができる。時間シフトオペレータは、y がフーリエ級数展開できるとすると

$$y_n = \sum_{f=-\infty}^{\infty} c_f \exp(ifn) \tag{34}$$

となる。ここで i は虚数単位であり、f は周波数である。時間シフトオペレータは、

$$\sum_{f=-\infty}^{\infty} c_f \exp(ifn) = B \sum_{f=-\infty}^{\infty} c_f \exp(if(n-1)) \tag{35}$$

を満たすことから、

$$B = \exp(if) \tag{36}$$

とわかる。この時間シフトオペレータを用いて、y_n は

$$y_n = \frac{1}{1 - \sum_{i=1}^{m} a_i B^i} \cdot \epsilon = g(B) \cdot \epsilon \tag{37}$$

と書くことができる。定常 AR モデルにおいてこの式は、伝達関数と呼ばれる $g(B)$ によって重み付けされたホワイトノイズの過去から現在までの足し合わせが現在の時系列を生じていることを表す。ここで、対数を取った伝達関数のべき展開を考える。

$$\log g(B) = \sum_{j=1}^{\infty} \phi_j B^{-j} = \sum_{j=1}^{\infty} \phi_j \exp(-ifj), \tag{38}$$

この ϕ_j がケプストラムである。この式から、ケプストラムは $\log g(B)$ を逆フーリエ変換して得られる係数であると解釈することができる。つまり、ケプストラムは過去の時系列の現在への寄与を示す伝達関数の持つ周波数特性の指標と考えることができる。これは、ケプストラムがもともと音声の解析のために考案されたことと関連している。というのも、声は周

波数の大きな声帯成分と周波数の小さい声道成分からなり、その分離は伝達関数の周波数依存性を用いて行われるからだ。

ケプストラムと AR 係数の関係式を導出する。ケプストラムのべき式において、両辺を B で微分することを考える。

$$\frac{d}{dB}\log\left[\frac{1}{1-\sum_{i=1}^{m}a_iB^i}\right]=\frac{d}{dB}\sum_{i=1}^{\infty}\phi_iB^i, \tag{39}$$

これは、

$$\frac{\sum_{i=1}^{m}ia_iB^{i-1}}{1-\sum_{i=1}^{m}a_iB^i}=\sum_{j=1}^{\infty}j\phi_jB^{j-1} \tag{40}$$

となり、

$$\sum_{i=1}^{m}ia_iB^{i-1}=(1-\sum_{i=1}^{m}a_iB^i)(\sum_{j=1}^{\infty}j\phi_jB^{j-1}) \tag{41}$$

を得ることができる。この式を展開して整理すると、m 次の AR モデルのケプストラムは、

$$\begin{aligned}
&\phi_1=a_1,\\
&\phi_n=a_n+\sum_{j=1}^{n-1}\left(1-\frac{j}{n}\right)a_j\phi_{n-j},\quad \text{if } 1<n\leq m,\\
&\phi_n=\sum_{j=1}^{m}\left(1-\frac{j}{n}\right)a_j\phi_{n-j},\qquad\quad \text{if } m<n
\end{aligned} \tag{42}$$

と AR 係数によって与えられる。

Kalpakis and Puttagunta (2001) [4] は、二つの時系列間の非類似度を、ケプストラム間のユークリッド距離

$$d_{cep}(\boldsymbol{y_p},\boldsymbol{y_q})=\left\{\sum_{j=1}^{\infty}(\phi_{p,j}-\phi_{q,j})\right\}^2, \tag{43}$$

と定義した。また、シミュレーションした ARMIA 時系列に対して、この非類似度に基づいたクラスタリングを行い、以下の特徴を挙げた。

1-cep. 次数の小さい係数ほど重みが大きく設定されており、非類似度に大きく影響を及ぼす点

2-cep. 次数が大きくなればなるほど、影響が急速に 0 に近づくため、はじめの数個の係数がうまく現象を捉えている必要がある点

3-cep. 特徴ベースの非類似度と比較して AR モデルに対する分類精度が高い点

4-cep. 振幅の平均やスケールに対して不変である点

5-cep. 時系列の並行移動に対して不変である点

6-cep. 周期性に対して敏感である点

7-cep. 伝達関数の積算はケプストラムにおいては加法となるため、時系列の周期特性を減法によって検出できる点

2.4 非類似度の利用とクラスタリング

2.3 節で行った 3 つの非類似度に関するレビューから、本研究では d_{pic} と d_{cep} を用いることに決定した。これには 2 つの理由がある。1 つ目は、d_{mah} は時系列間の距離を計算することはできるが、相対座標が与えられない点である。距離だけが与えられた場合においてもクラスタリングを実行することはできるが、それにはクラスタ間の距離を近似しつつ更新する必要がある。さらに、クラスタの解釈を行う際には、後述するクラスタ重心が大きな役割を果たすが、相対座標が与えられないため、d_{mah} によって構築されたクラスタの特徴を可視化することができない。2 つ目は、d_{mah} が仮説検定に基づき構築された指標である点である。Maharaj の仮説検定においては、2 つの時系列間の距離が等しいという帰無仮説を利用するが、これは多次元時系列間の比較やクラスタという概念と整合的ではない。

非類似度 d_{pic} と d_{cep} を計算する際には、まず区間時系列 $\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_k}$ の AR 係数を *arfit* 関数 [5] を用いて求めた。d_{cep} の導出においては、式 (42) を用いてこれらの AR 係数からケプストラムを計算した。求めたこれらの相対座標に基づき、*hclust* 関数 [10] によって階層クラスタリングを行った。このとき、階層クラスタリングの手法として重心を用いるものには重心法と Ward 法と呼ばれるものが存在するが、重心法でクラスタリングを行った場合には、特定のクラスタ間の距離が、クラスタの統合ごとに変化し、樹形の逆転現象が起こることがある (図 2 のデンドログラムを参照のこと)。それゆえ、本論では Ward 法を用いる。

Ward 法は、分類対象同士のユークリッド距離を基礎とする分類法であり、クラスタ内での散らばりの指標を計算し、その増加が最も小さくなるようにクラスタ同士を結合する。ここで、区間時系列 $\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_k}$ に対して計算された AR 係数ベクトルが $\boldsymbol{a_1}, \boldsymbol{a_2}, \cdots, \boldsymbol{a_k}$ であ

るとする。このとき、$u(\leq k)$ 個のデータから構成されるクラスタの散らばりの指標として、偏差平方和行列

$$S = \sum_{i=1}^{u} (\boldsymbol{a_i} - \bar{\boldsymbol{a}})(\boldsymbol{a_i} - \bar{\boldsymbol{a}})^T, \quad \bar{\boldsymbol{a}} = \frac{1}{s} \sum_{i=1}^{u} \boldsymbol{a_i}, \tag{44}$$

を用いる。このとき、この行列の行列式を散らばりの指標とする場合と、対角成分のみを用いる場合があるが、*hclust* においては、対角成分のみを用いた、

$$\mathrm{tr}S = s_{11} + s_{22} + \cdots + s_{uu}, \tag{45}$$

が指標として用いられている。実際のクラスタリングでは、存在する v 個のクラスタ内の散らばりの指標の和

$$W_v = \mathrm{tr}S_1 + \mathrm{tr}S_2 + \cdots + \mathrm{tr}S_v, \tag{46}$$

を計算し、任意の 2 つのクラスタを 1 つに統合したときの W_{v-1} が最も小さくなるように、すなわち W の増加が最小となるようにクラスタの統合を行う。出力したデンドログラムの height は、W のことである。

作成されたクラスタの解釈には、形成されたクラスタの重心

$$\bar{\boldsymbol{a}} = \frac{1}{s} \sum_{i=1}^{u} \boldsymbol{a_i}, \tag{47}$$

を用いた。具体的には、各クラスタの重心の AR 係数、あるいは重心のケプストラムから再計算された AR 係数を用いて pc1 と pc2 のシミュレーションを行い、シミュレーション結果からウミガメの輪郭形状を復元することで、クラスタを特徴付ける遊泳パターンを調べた (動画 2、3 を参照のこと)。

3 結果

幼体のアオウミガメ 10 匹の輪郭形状の変化に関する時系列 pc1 と pc2 を AR モデルの AIC によって分割した。このとき、幼体のウミガメの運動は 10 フレーム (2 秒) 以下の周期運動によって構成されていたことから、AR モデルの最大次数は 20 に設定した。動画 1 の timeseries

は分割された pc1 と pc2 の時系列を表し、プロットした frame(表記上は time) に対応する水槽での相対座標 (Position)、輪郭形状 (Shape)、pc 平面上での値 (pc plot) を示す[9]。この動画から、AR モデルによる分割は、漕ぎ運動から羽ばたき運動、その逆、あるいはどちらとも取れない中間的な泳ぎへの変化点を抽出できたことがわかる。分割された泳ぎの細かな解釈については、クラスタリングの精度を確認する本論の趣旨から外れるため、次回の資料に記載する。

得られた区間時系列どうしの距離を非類似度 d_{pic}, d_{cep} を用いて計算し、Ward 法によるクラスタリングを行った。クラスタリングの結果、各非類似度に対してデンドログラム (図 1、2) を得た。非類似度間で生成されたデンドログラムには、3 つの違いが見られた。

1. d_{pic} を用いた方では、以降クラスタが急増する高さ 5 のあたりまで 4 つのクラスタが存在するのに対し、d_{cep} を用いた方は 3 つのクラスタが存在する点
2. d_{pic} を用いた方に比べ、d_{cep} を用いた方が高さの低いところでクラスタが急増している点
3. d_{pic} を用いた方に比べ、d_{cep} を用いた方が最も大きな 2 つのクラスタが統合されるまでの高さが高くなっている点。

1 点目から、d_{pic}, d_{cep} に対して、最適クラスタ数 (解釈上決定すべきクラスタ数) をそれぞれ 3、4 と設定した。2 点目と 3 点目から、d_{cep} の方は早い段階でまとまりを作っており、かつ最終的に大きなクラスタが形成される際には W が大きく増加していることがわかる。この結果は、[2-cep] 次数が大きくなると急速にその影響が少なくなる、という特徴を反映して、高さが低いところでは小さい次数の係数の値が同じ区間時系列同士でクラスタが形成される一方、大きなクラスタの併合にはすでに小さい次数同士でまとまったクラスタを併合する際に W が大幅に増加したためだと考えられる。

得られたクラスタ 1-3 の重心を、AR 係数 (piccolo)、あるいはケプストラムとして求めた。ケプストラムによって求めた重心は、式 (42) を用いることで AR 係数に変換した。得られた AR 係数を用いて、クラスタ 1-4 を特徴づける pc1 と pc2 をシミュレーションによって生成し

[9] クラスタリングの結果も示されているが、動画が重くなるので 1000 ステップにトリミングしている。今回は分割にのみ注目されたい。

た。そして、得られた pc1 と pc2 の値から輪郭形状を復元することで、動画 2-1, 2-2, 2-3, 2-4, 3-1, 3-2, 3-3 を作成した。

d_{pic} を用いたクラスタリングによって作成された動画において、クラスタ 1 から 4 の間に大きな違いは見受けられなかった。強いてあげれば、クラスタ 1 と 2 に比べ、3 と 4 は手足をばたつかせており、flapping が少ない印象を受けるが、明確な違いとはいえない。d_{cep} でも、クラスタ 1 に比べて 2 と 3 がばたついており、flapping が少ない印象を受けるが、これも明確な違いとはいえない。この結果から、クラスタ重心から復元された輪郭形状の変化は目視レベルでは特徴を判別できず、本論のクラスタリング手法の精度は高くないと言える。4. 議論では、この原因についての仮説の提案と検証を行った。

4 議論

AR モデルの当てはまりの良さを用いて、時系列を複数の小区間に分割した。本論で扱った動画では、目視によって、行動のレパートリーに休息や潜水はみられず、ほとんどが漕ぎ運動と羽ばたき運動から構成されていることが確認できる。AR モデルを用いた分割では、この結果と直感的に合う形で、漕ぎ運動と羽ばたき運動、さらにはこれらの中間的な泳ぎのパターン変化を検出可能であった。このことから、AR モデルは泳ぎのパターンの変化を検出するに十分であるといえる。

分割の結果に対し、クラスタリングの精度はあまり良くなかった。動画の目視観察から、漕ぎ運動と羽ばたきは明確に区別可能であるため、人間の視覚から分類可能なクラスタは 2 以上であると言える。それゆえ、実用性を考えると、少なくとも漕ぎ運動と羽ばたき運動が異なるクラスタとして区別されることが望ましい。しかし、d_{pic} と d_{cep} によって作成された各クラスタの重心を用いて復元した動画からは、クラスタの特徴を読み取ることができなかった。

クラスタリングの精度が高くないことは、pc1 と pc2 を独立な時系列と考え、AIC の計算、クラスタリング、シミュレーションを行ったためだと考えられる。pc1 と pc2 は主成分分析によって導出されるものであるが、時間方向には相関が存在する。例えば、羽ばたき運動が卓

越する際は、pc2 が大きく変動し、pc1 はほとんど変化しないはずであるが、このような関連性が捉えられていないために、不明瞭な輪郭形状の変化がクラスタとして抽出されたと考えられる。この点を検証するため、Ekahi の時系列のうち、目視により定常と判断した 10100 番目から 10400 番目の frame に対して、pc1 と pc2 が互いに独立と仮定して推定した AR 係数 (*arfit* によって推定 [5]) と、相関を考慮して推定した MAR 係数 (Multi dimensional AR: *marfit* によって推定 [5]) を用いて、シミュレーションを行った。そして、シミュレーションの結果生成された時系列 pc1 と pc2 から輪郭座標を復元した。この結果を、それぞれ動画 4-1 と 4-2 に示す。また比較のため、AR、MAR 係数の推定に用いた frame から計算した pc1 と pc2 を用いて復元した形状変化を動画 4-3 に示す。pc1 と pc2 を独立と仮定してシミュレーションした動画 4-1 では、例えば右手を出した状態からいきなり左手が伸びる、両手を広げた状態からいきなり縮こまるなど、不自然な形状変化がたびたびみられ、全体的にばたついている印象を受ける。一方、pc1 と pc2 の相関を考慮してシミュレーションした動画 4-2 では、形状が滑らかに変化しており、動画 4-3 と見分けがつかない程度に自然な運動が再現されている。また、時系列を見ても、pc1 の変動が大きい場合には pc2 の変動は小さい、あるいは pc2 の変動が大きい場合には pc1 の変動は小さい、といった漕ぎ運動と羽ばたき運動の基本的な特徴が再現されていることがわかる。

これらの結果から、直感に沿ったクラスタを形成するには、pc1 と pc2 の相関を考慮した MAR モデルの距離を用いてクラスタリングを行う必要があると考えられる。この際、元の時系列の分割も MAR モデルによって行い、クラスタ間距離には d_{pic} を MAR 係数に拡張したものを用いる必要がある。また、d_{cep} の拡張も可能であれば、MAR モデルをベースにしたクラスタリングにおいても d_{pic} との比較対象とすべきである。

本論では、視覚的に解釈が容易なことから、分割、クラスタリング、シミュレーションの全てにおいて pc1 と pc2 のみを用いた。しかし、抽出した形状変化そのものである動画 1 と、pc1 と pc2 のみの情報を用いて復元された動画 4-3 を比較すると、漕ぎ運動や羽ばたき運動を捉えることはできているが、詳細な動きは潰れてしまって見えない。それゆえ、MAR モデルを用いてもなお精度が良くない場合には、pc3 以降の軸を考慮することも視野に入れる必要があるだろう。本論で用いた手法はほとんどが線形計算からなるため、pc3 以降の軸を

考慮しても、計算時間が膨大になることはないと考えられるが、複雑性が増すことによる解釈性の低下には注意すべきである。

本論で行ったクラスタリングの精度が向上すれば、遊泳パターンの継続時間やその変化を定量化することが可能となる。図 4 に、d_{pic} によって作成したクラスタに代表される遊泳の継続時間の分布 (上) とその変化の割合 (下) を示す。図に記載されている 1-4 はそれぞれクラスタ 1-4 に対応する。行動の継続時間や変化を定量的に捉え、コルチコステロン濃度といった他のストレス指標との整合性を調べることで、ステレオタイプ行動の指標を提案することができれば、水族館において円滑なストレス評価が可能になると考えられる (こちらも詳しくは次回のゼミ資料を参照のこと)。

References

[1] Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. the Journal of the Acoustical Society of America, 55(6), 1304-1312.

[2] Berk, K. N. (1974). Consistent autoregressive spectral estimates. The Annals of Statistics, 489-502.

[3] Bhansali, R. J. (1978). Linear prediction by autoregressive model fitting in the time domain. The Annals of Statistics, 224-231.

[4] Kalpakis, K., Gada, D., & Puttagunta, V. (2001). Distance measures for effective clustering of ARIMA time-series. In Proceedings 2001 IEEE international conference on data mining (pp. 273-280). IEEE.

[5] Kitagawa, G. (2020) Introduction to Time Series Modeling with Applications in R. Chapman & Hall/CRC.

[6] Liao, T. W. (2005). Clustering of time series data—a survey. Pattern recognition, 38(11), 1857-1874.

[7] Maharaj, E. A. (1996). A significance test for classifying ARMA models. Journal of Statistical Computation and Simulation, 54(4), 305-331.

[8] Montero, P., & Vilar, J. A. (2015). TSclust: An R package for time series clustering. Journal of Statistical Software, 62, 1-43.

[9] Piccolo, D. (1990). A distance measure for classifying ARIMA models. Journal of time series analysis, 11(2), 153-164.

[10] R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

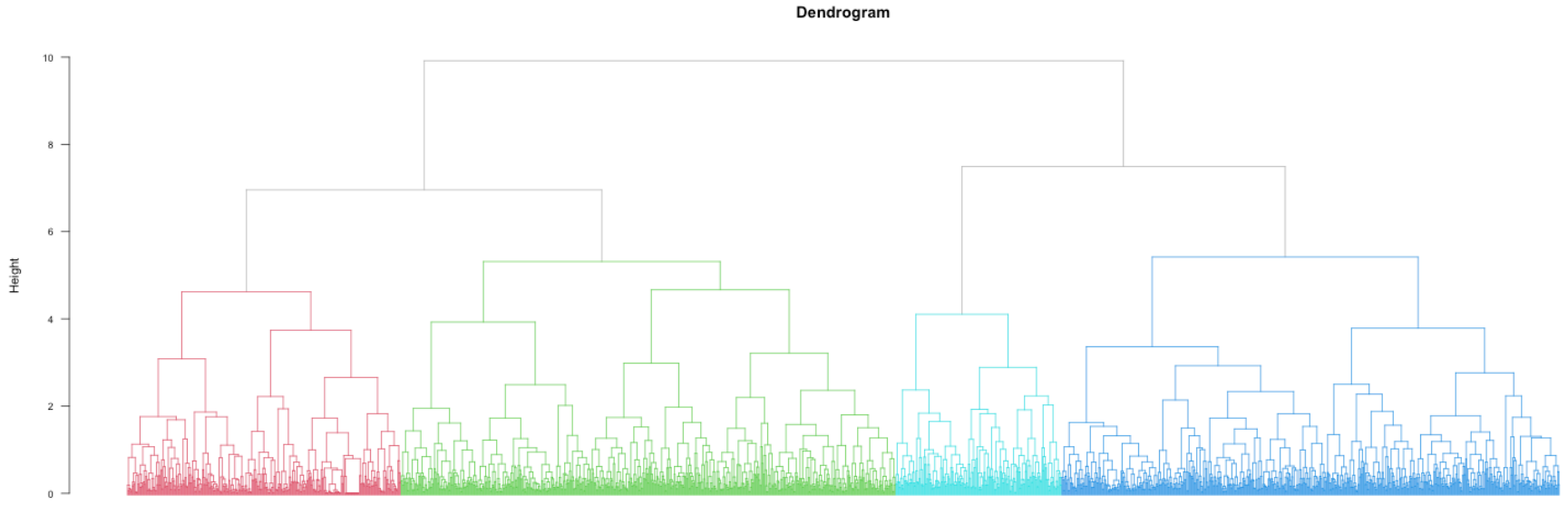

Figure 1: d_{pic} を用いて作成されたデンドログラム。最適クラスタ数 4 でプロットしており、赤色がクラスタ 1、緑色がクラスタ 2、青色がクラスタ 3、水色がクラスタ 4 に該当する。

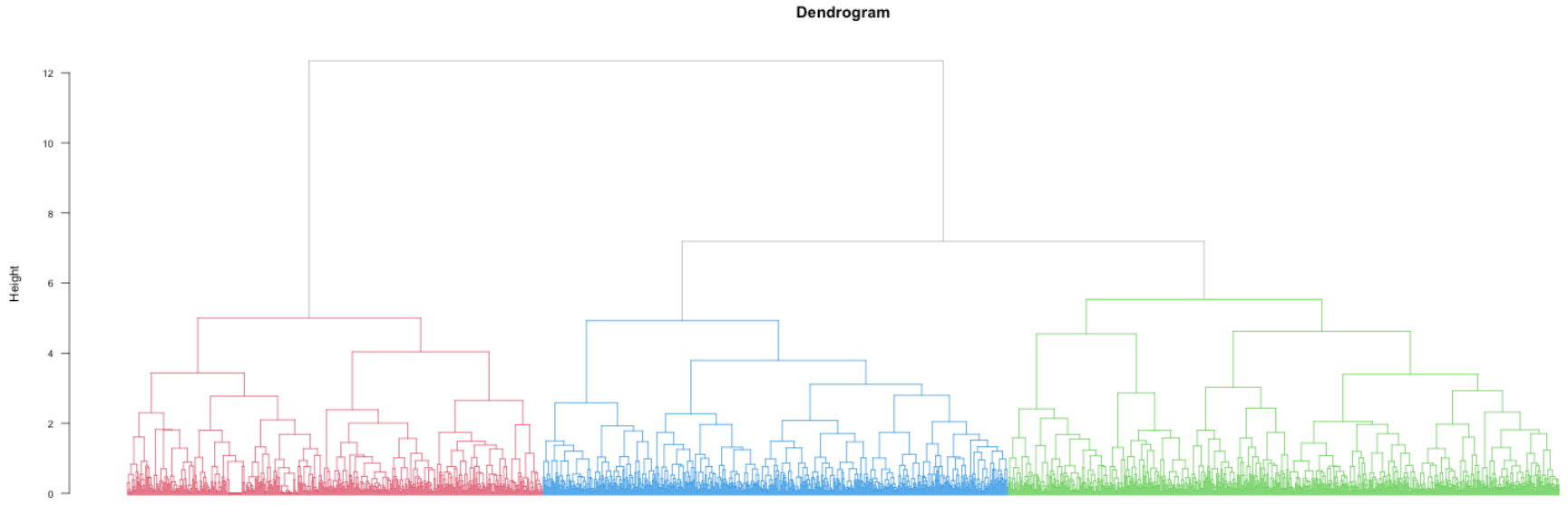

Figure 2: d_{cep} を用いて作成されたデンドログラム。最適クラスタ数 3 でプロットしており、赤色がクラスタ 1、緑色がクラスタ 2、青色がクラスタ 3 に該当する。

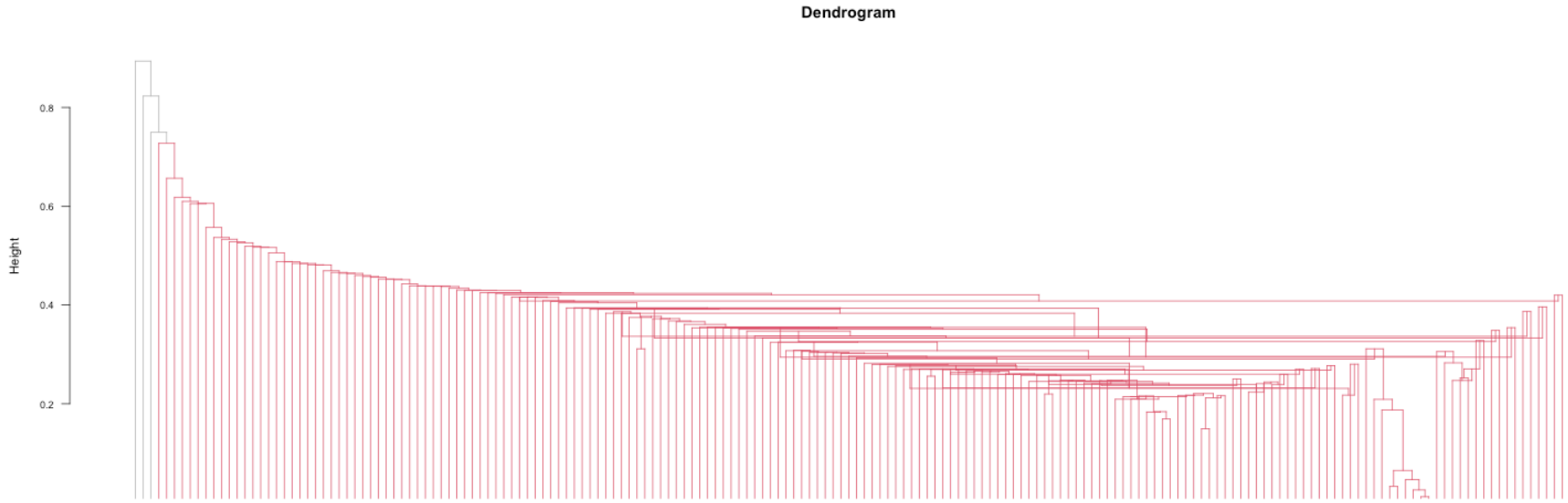

Figure 3: 重心法を用いて作成されたデンドログラムのサンプル。最適クラスタ数 4 でプロットしているが、樹形の逆転が発生していることに加え、すべての時系列が 1 つのクラスタに統合されている。

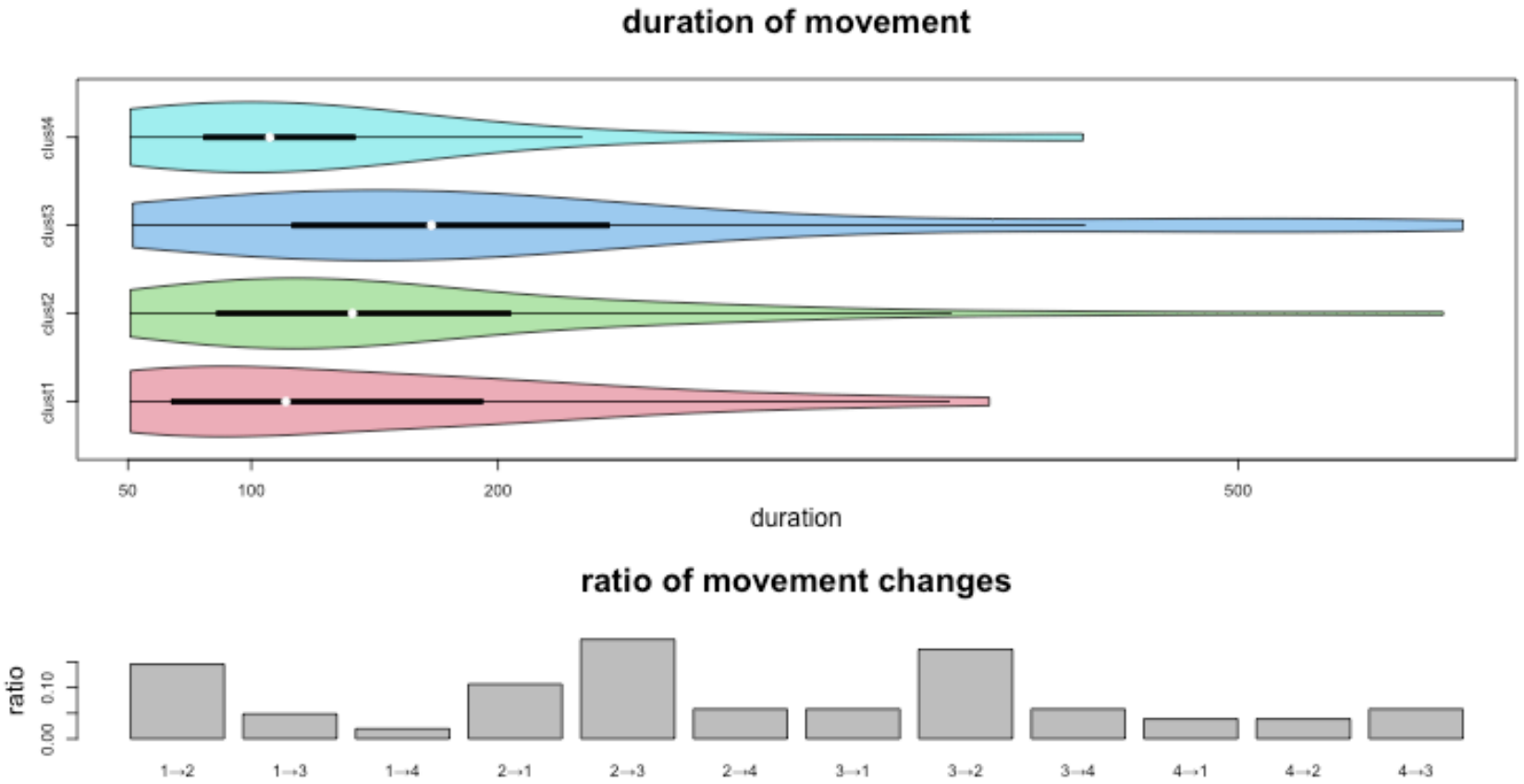

Figure 4: d_{pic} によって分類されたクラスタに代表される遊泳パターンの継続時間の分布 (上) とその変化の割合 (下)。