

非定常時系列の教師なしクラスタリング：位相的データ解析による分析

黒木 裕鷹

2018 年 4 月 27 日

1 はじめに

現在、世界の様々な工場では産業用ロボットが用いられている。そのような工場では、重量物の運搬を必要としたり、霧散している粉塵を吸い込んだりする危険性など、人間には負担の大きい作業の割合が多い。産業用ロボットはこのような安全性の問題を解決するだけでなく、経済性や効率性においてもメリットがある。しかし同時にロボットには故障のリスクがあり、ただひとつのロボットの故障が生産ライン全体に影響を及ぼし得る。故障を未然に回避するためにはメンテナンスなどの保守作業を行うことが重要であるが、メンテナンスや部品交換にも費用や時間などコストがかかるため、適切なタイミングで行われることが望ましい。

自動車メーカーの M では、生産ラインで使用するアームロボットの故障を未然に防ぐため、その減速機の交換を経験に基づくタイミングで行っている。交換のタイミングが遅すぎれば故障を招くが、早すぎても交換に時間がかかる分パフォーマンスが低下することとなる。本調査では、ロボットアームをモニタリングした振動のセンサーデータを用い、減速機交換前後のデータに明確な違いがあるかどうかを調査した。ロボットの故障により生産ラインを止めることはあってはならないため、故障直前のデータは得ることが出来ないことに留意しなければならず、教師無し学習によるクラスタリングを行うことを目的とした。

M より提供されたデータは一つのロボットアームにつき、減速機交換前後それぞれで 5 秒間の計測を 10 回行った 1 次元の振動データである。ロボットアームはそれぞれ挙動が異なるため、その主要な振動はアームごとに異なっている。また、各アームの行動 1 セットは 5 秒間ではないため、各計測ごとにより行動セットの中の計測する部分が異なっている。

クラスタ分析では、クラスタリングされる対象間の類似度もしくは非類似度が必要である。時系列を対象としたクラスタリングにおいて、最も広く使用されている尺度はユークリッド距離、Dynamic Time Warping(DTW) [1], CORT [2] などである。DTW は 2 つの時系列データの時刻を t_i, t_j とすると、 t_i, t_j 全ての組の誤差を計算し、それらの合計が最小になるような経路を求めるアルゴリズムである。そのため、DTW は二つの系列の周期や長さが異なる場合でも非類似度を算出することができる。DTW により挙動の計測開始時点が異なる問題は多少解決されるが、振動の主な挙動はアームによって異なるため、減速機交換前後の差異はクラスタリングの結果に表れない。また、未知音源分離で用いられる独立成分分析(ICA)は主要な振動と微細な振動を分けることを可能にするかもしれないが、同時点の複数観測を必要とするため本データには適用することはできない。つまり時系列の同時性を利用することなく、かつアームごとに異なる主要な振動に影響されない特徴量によって教師無し学習を行うことが課題である。

本調査では、以上のような課題を解決するため、データの位相的な特徴に注目した。近年着目されている Topological Data Analysis(TDA) は複雑なデータを扱う上で強力であり、パーシステント・ホモロジー [3] とその表現であるパーシステント・ダイアグラム [4] をはじめとするその手法はデータの位相的特徴を抽出し、新しい知見を与える。また TDA の分野で広く行われているように、時間遅れ座標を用いて時系列を多次元に埋め込むことで位相的特徴を抽出する足掛かりとした。時間遅れ座標による埋め込みはダイナミカルシステムの分野で、状態空間の復元を目的に広く用いられている。パーシステント・ダイアグラムを比較するために Wasserstein 距離 [5] がしばしば用いられるが、本調査で扱うようなデータでは計算量の観点で現実的ではない。また、[6] の提案する Betti sequence はデータの位相情報を 1 次元の系列として要約する。そこで本調査では、観測の上述した観測の長さや観測開始時点が異なる問題を解決するためにセンサーデータの位相情報を要約した Betti sequence をもとめ、DTW に基づく全結合型階層的クラスタリングを行うことにより、減速機交換前後の振動を分析した。

本レポートの構成は次のようである。まず 2 節では本調査で使った TDA の手法やクラスタリング手法について述べ、3 節では実際のデータ解析とその結果を示す。4 節では考察を行うと共に今後の展望について触れる。

2 Topological Data Analysis(TDA) と時系列クラスタリング

トポロジーを用いてデータの情報を抽出するものであり、その主要な手法は mapper [7] とパーシステント・ホモロジー [3] である。これらはノイズを含む複雑なデータセットから何か新たな知見を得る目的でしばしば用いられてきた。本調査ではパーシステント・ホモロジーを利用する。

2.1 パーシステント・ホモロジー

距離空間 (M, d_M) の有限点集合を X , X を中心とした半径 r の級の和集合を $B(X; r) := \bigcup_{i=1}^n B(x_i; r)$ とする。ただし、 $B(x; r) = \{y \in M | d_M(x, y) \leq r\}$ とする。級の和集合を半径パラメータ r で集めた集合 $\mathbb{B}(X) := \{B(X; r)\}_{r \geq 0}$ をここでは X のフィルトレーションという。 $r \leq a$ ならば包含関係 $B(X; r) \subset B(X; a)$ があるため、ホモロジー群間の射 $u_r^a : H_q(B(X; r)) \rightarrow H_q(B(X; a))$ を包含写像から誘導する。このとき、ホモロジー群の系列

$$H_q(\mathbb{B}(X)) : \cdots \rightarrow H_q(B(X; r)) \xrightarrow{u_r^a} H_q(B(X; a)) \rightarrow \cdots (r \leq a)$$

を X の q 次元パーシステントホモロジーという。パーシステントホモロジー係数多項式環や A_n 型籠の表現などの言葉で解釈することが出来、分解定理により適切な区間表現 $\mathbb{I}[b_i, d_i]$ を通じて

$$H_q(\mathbb{B}(X)) \cong \oplus \mathbb{I}[b_i, d_i] (b_i \leq d_i)$$

と分解される。これをユークリッド空間 \mathbf{R}^2 内に表示した多重集合

$$D_q(X) := \{(b_i, d_i) | i \in I\}$$

を X の q 次元パーシステント・ダイアグラムという。パーシステント・ダイアグラムの元 (b_i, d_i) はホモロジーの生成元の発生時間 (Birth time) を b_i , 消滅時間 (Death time) を d_i と記録しているものと解釈できる。

q 次のパーシステント・ダイアグラムは 2 次元の散布図として表され、その解釈や扱いが困難である。そのため、様々なダイアグラムの要約が提案されてきた。最も単純なものが最大パーシステンスであり、 $\max_i (b_i - d_i)$ で表される。

2.2 パーシステント・バーコード

前節で定義したダイアグラム $D_q(X)$ のホモロジー i に対して, 次の $s_i(r)$ を定義する.

$$s_i(r) = \begin{cases} 1 & (b_i \leq r \leq d_i) \\ 0 & \text{otherwise} \end{cases}$$

$s_i(r)$ を並べてプロットしたものをパーシステント・バーコードという.

また, [6] はバーコードを単純に扱い, 機械学習の枠組みで扱いやすくするために次で提案される Betti sequence, $BS(X)$ を提案している.

$$BS(X) = \sum_i s_i(r)$$

Betti sequence は半径 r のときにいくつのホモロジーが存在しているかを表す 1 次元の系列となる.

2.3 1次元時系列データに対する TDA

時系列から複数次元の有限点集合を構成する方法として, 遅れ時間座標を利用した埋め込み (embedding) がある. 埋め込みはアトラクタを再構成するために非線形ダイナミカルシステムの分野で盛んに利用されている. 長さ N の時系列 x_1, x_2, \dots, x_T から適当な遅れ時間 τ ごとの d 個の測定値を取り出し $V(t) = (x(t), x(t+\tau), \dots, x(t+d-1))$, ($t = 1, 2, \dots, T$) の d 次元有限点集合を得る. この有限点集合が元の k 次元力学系の埋め込みになるための十分条件として, $d \geq 2k + 1$ (ターケンスの埋め込み定理) が知られている.

また, 有限点集合の次元 d が大きくなったときに \mathbb{R}^d の n 個の点集合のドロネー三角形分割の計算量は $\mathcal{O}(n^{\lfloor \frac{d}{2} \rfloor})$ になりうる [8], [9]. そこで, 位相情報を保持しながら計算を簡略化するため, 十分多次元に埋め込んだ後主成分分析により 3 次元に次元削減するアプローチが提案されている [10].

2.4 時系列クラスタリング

時系列間の非類似度を算出する際, 用いられているのが [1] で提案された Dynamic Time Warping (DTW) である. DTW は三角不等式を満たさないため距離ではないが, 比較的少ない計算量で要素数の異なる系列同士の距離のような量を求める手法である.

また本調査では DTW に基づき, 完全連結方による階層型クラスタリングを行った. 完全連結法とは, クラスタ間で最も類似度が低いデータ間の距離をクラスタ間の距離にする方法である.

3 データ解析

3.1 データの概観

本調査では, 自動車メーカー M の産業用ロボットアームの振動データを扱う. 15 種類のロボットアームに対し, 減速機を交換する前と交換した後をそれぞれ 10 回モニタリングした合計 300 系列のデータである. 計測の基本単位は 100 ミリ秒であり, 計測の長さは 5 秒間である. ロボットアーム 15 種類の減速機交換前後のデータのそれぞれ

れ 10 回分の観測をつなぎ合わせ、以下の図 1 に示した。このように、いずれのロボットアームにおいてもその時系列は明らか非線形構造を持つと同時に交換前後の系列は似通っており、アームごとの挙動が主要な振動として現れていることがわかる。減速機交換前後の特徴を目視で判断することは出来ない。さらに交換前後で計測開始時間が異なるものもあり、単純な比較が難しいことがわかる。10 回の観測をつなぎ合わせたが、系列は概ね周期的になっている。しかし一部に周期的でなかったり、交換前後で明らかに異なった振動をしている系列が見て取れる。

3.2 データの位相的特徴

ロボットアームによる振動のスケールを調整するため、以下の分析は予め全系列を標準化した上で行った。1 次元の振動時系列から位相情報を取り出すために、遅れ時間座標への埋め込みを行った。その際、遅れ時間単位は $\tau = 1$ とした。またロボットアームの構造や状態変数の数が明らかでないことから、埋め込み次元にはロボットアームが 3 次元であることを考慮し、ターケンスの埋め込み定理を参考に $d = 3 \times 2 + 1 = 7$ とした。さらに [10] で提案されているように主成分分析を用い第 3 主成分までを取り出し 3 次元の有限点集合を抽出した。アルファ複体のフィルトレーションを用いて抽出した点集合の 0, 1, 2 次元パーシステント・ホモロジーを計算した。またその要約として各次元の Betti sequence を計算し、以下の図 2, 3 に示した

図 2, 3 より、ほとんどのロボットアームで Betti sequence が似通っていることがわかる。また、2 次の Betti sequence は 1 次のものに比べて観測ごとのばらつきが大きくなっていることが見て取れる。

3.3 DWT に基づいた時系列クラスタリング

ロボットアームごとに減速機交換前 10 系列、交換後 10 系列の計 20 系列をクラスタリングすることを考える。観測ごとのばらつきが少ない 1 次のベッチ数をもとに DTW を比類似度とした階層的クラスタリングを行った。クラスター間の距離は完全連結法を用いて算出した。そのデンドログラムを図 4 から図 18 に示した。

図 4 から図 18 より、いくつかのロボットアームでは明確に減速機交換前後のクラスターが構成できたことが分かる。またそうでないアームでも細かい部分では減速機交換前後のクラスターができていたことが分かる。

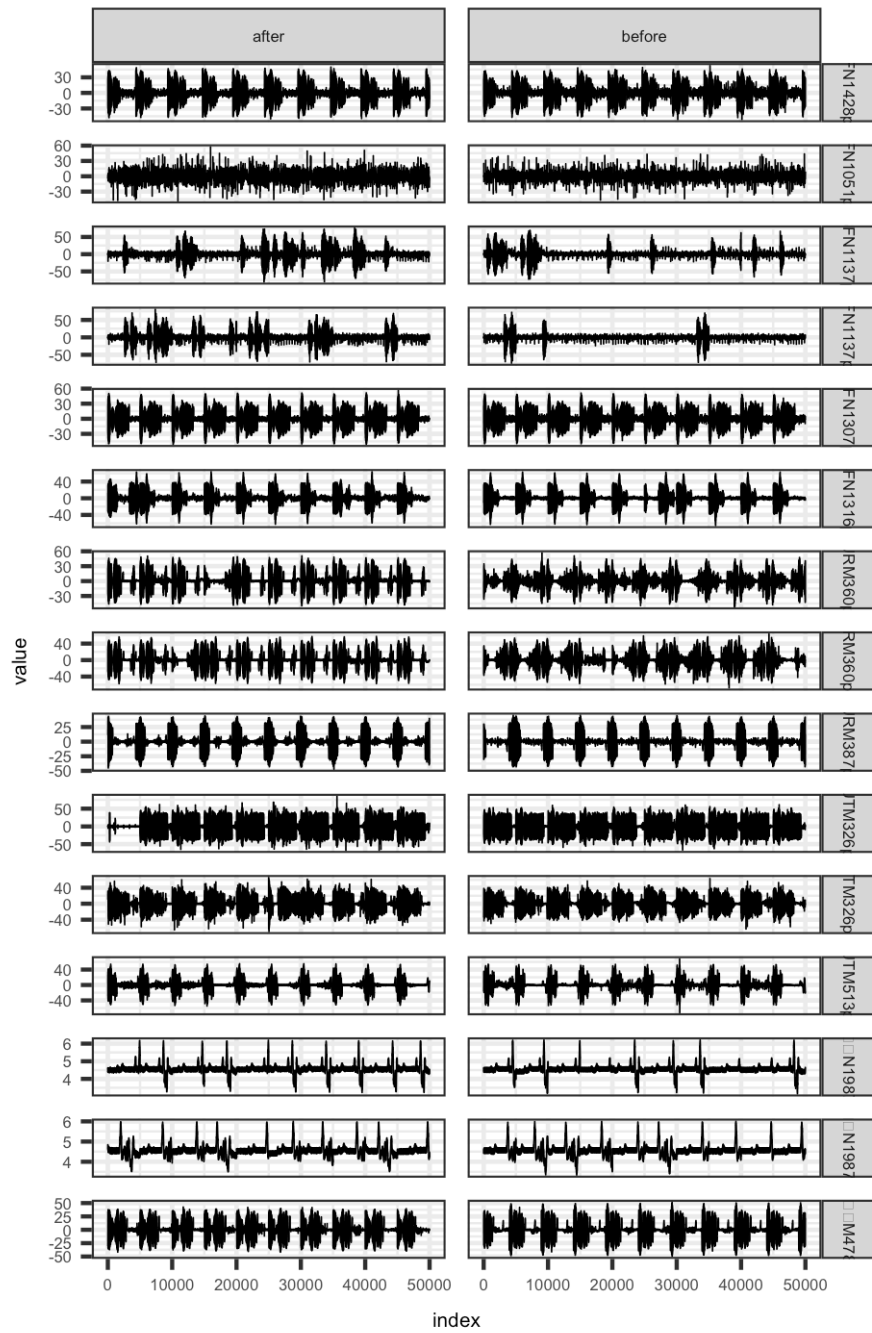


図 1: センサーデータの時系列プロット

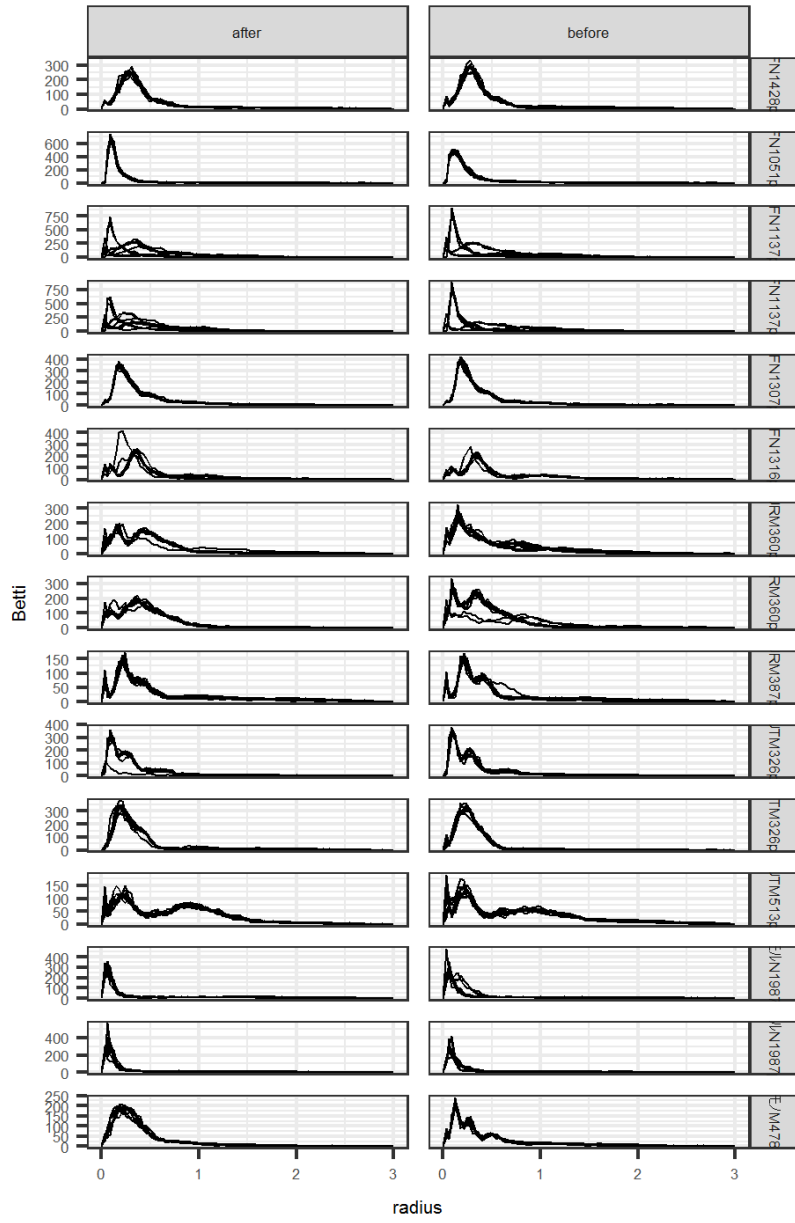


図 2: 1 次の Betti sequence

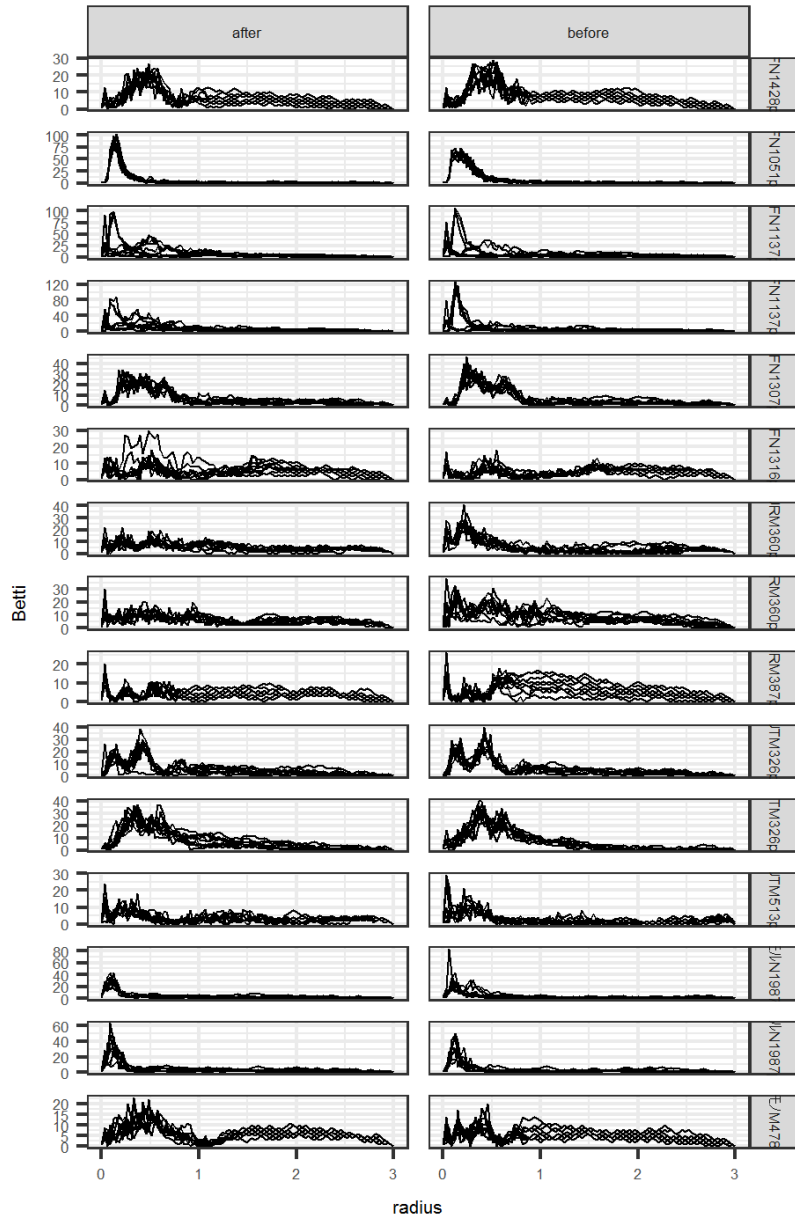


図 3: 2 次の Betti sequence

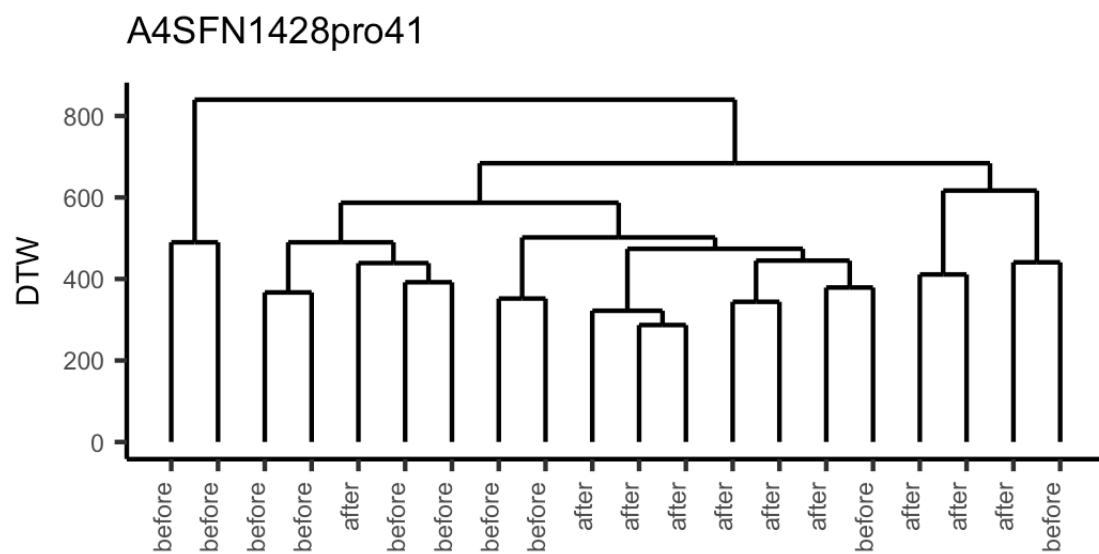


図 4: デンドログラム 1

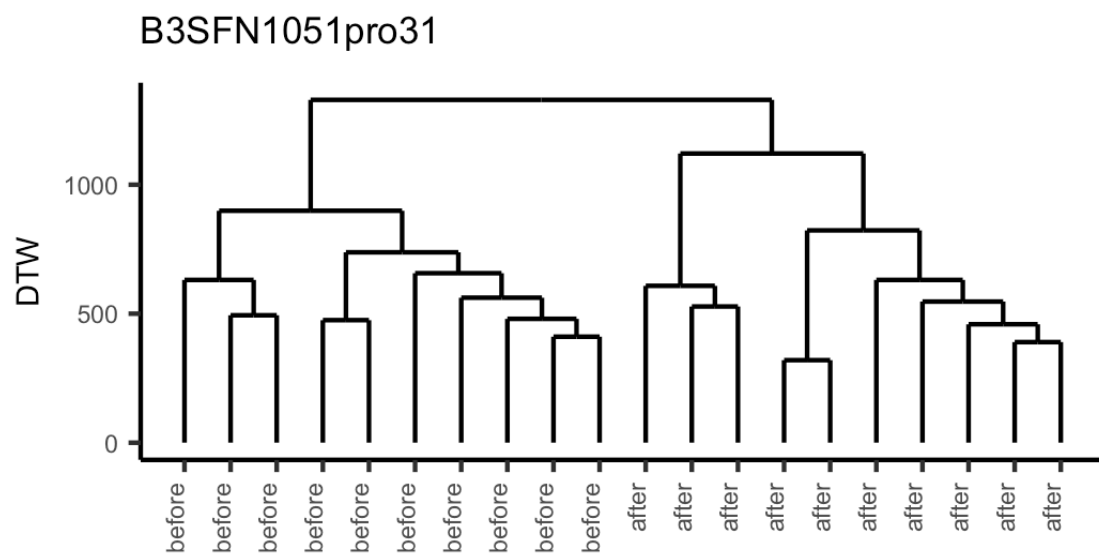


図 5: デンドログラム 2

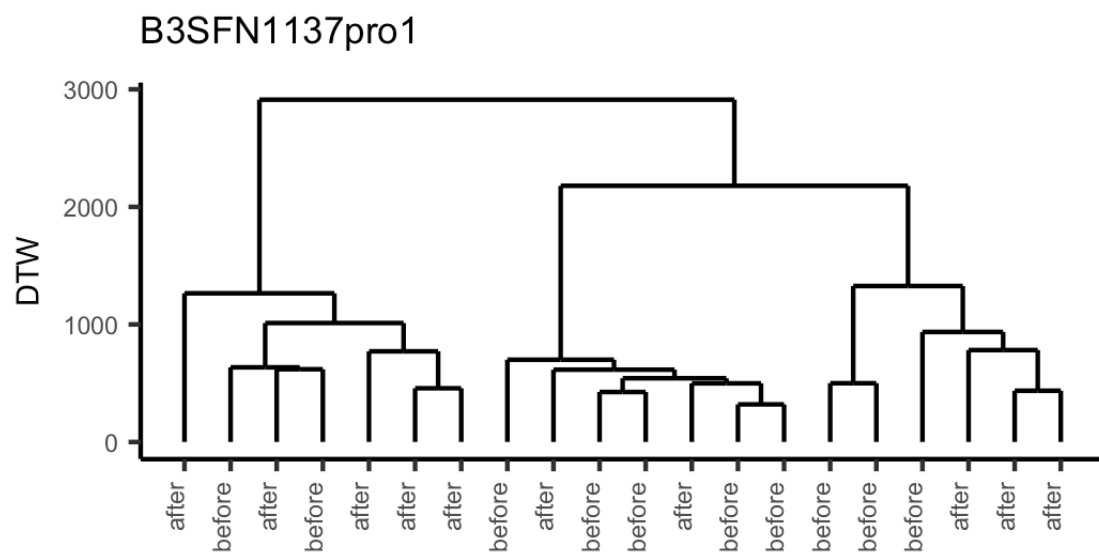


図 6: デンドログラム 3

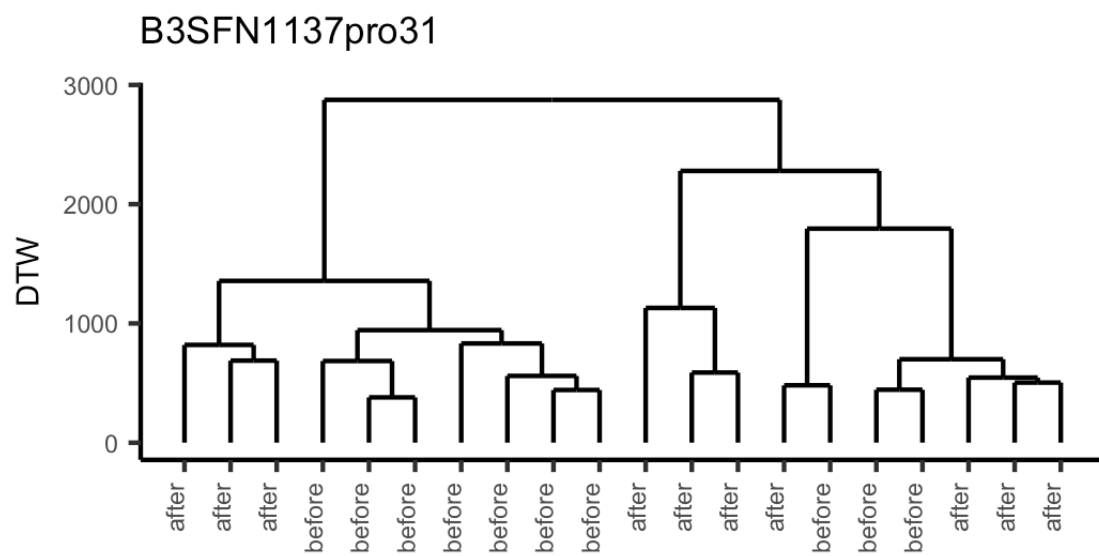


図 7: デンドログラム 4

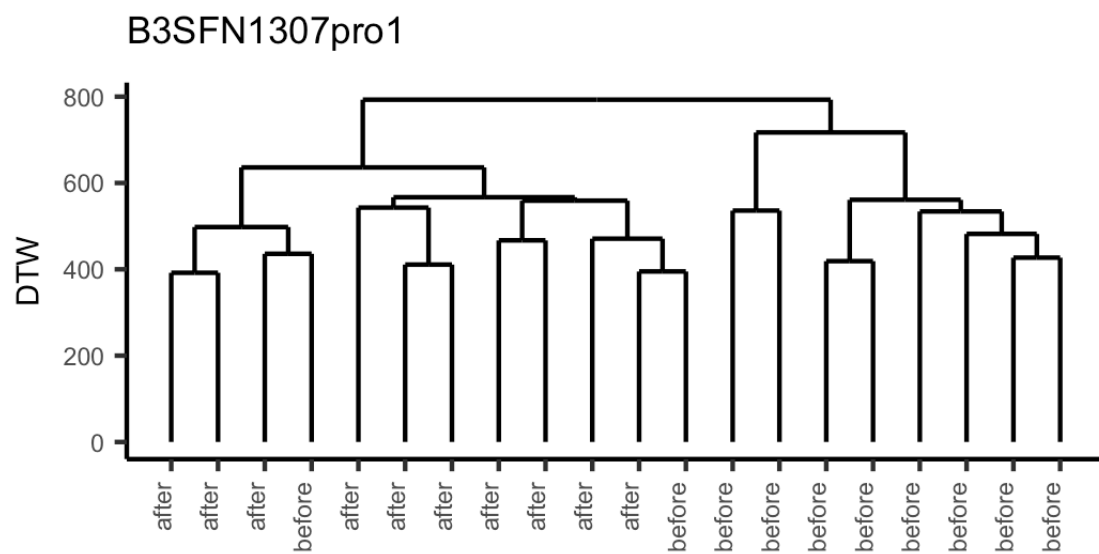


図 8: デンドログラム 5

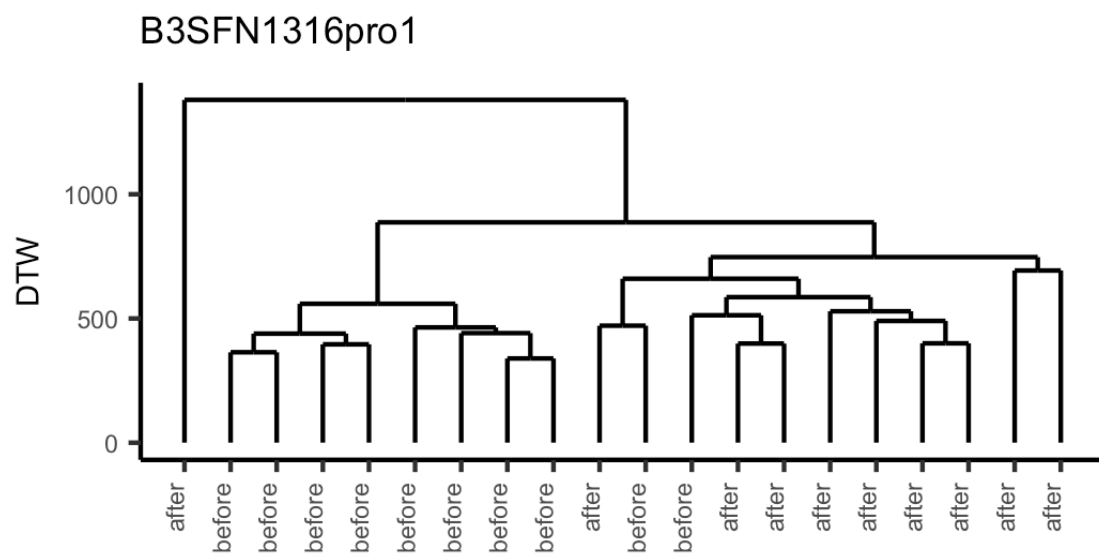


図 9: デンドログラム 6

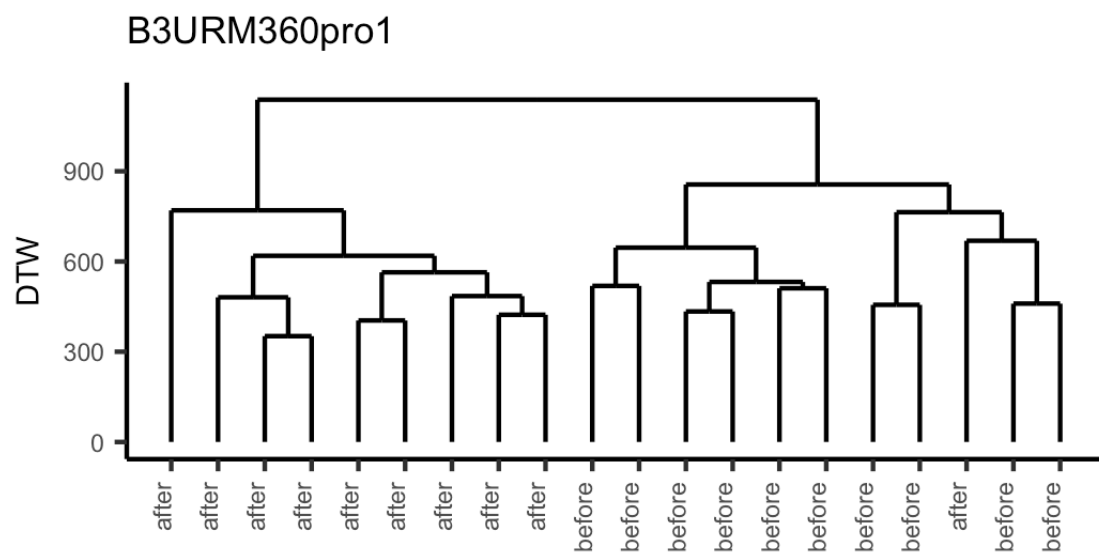


図 10: デンドログラム 7

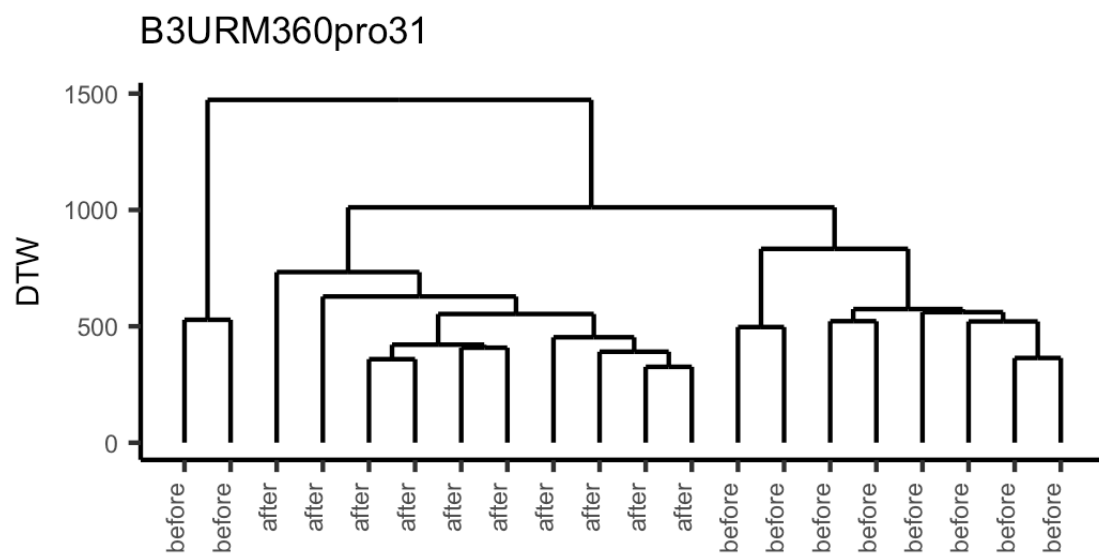


図 11: デンドログラム 8

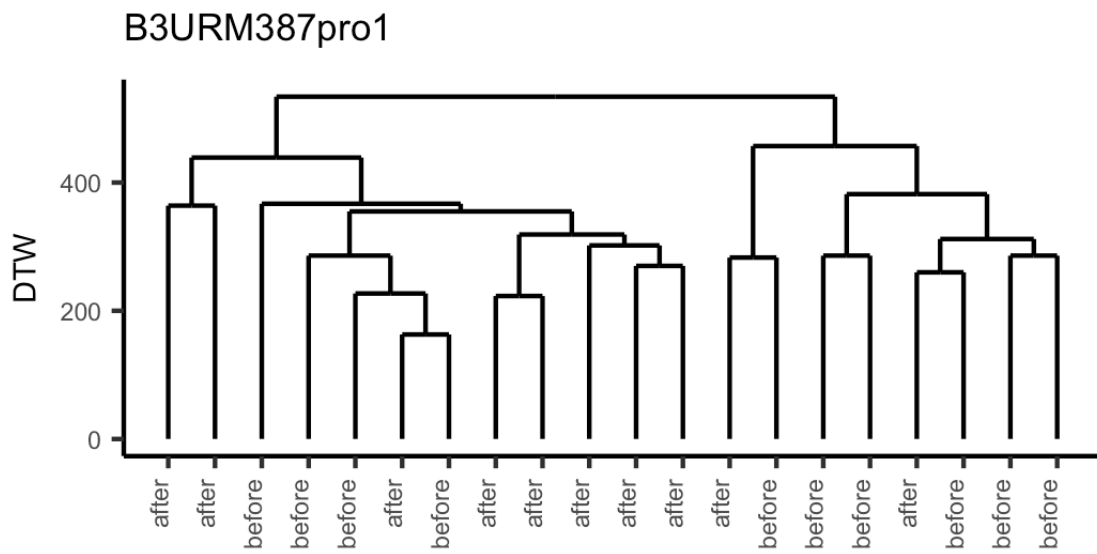


図 12: デンドログラム 9

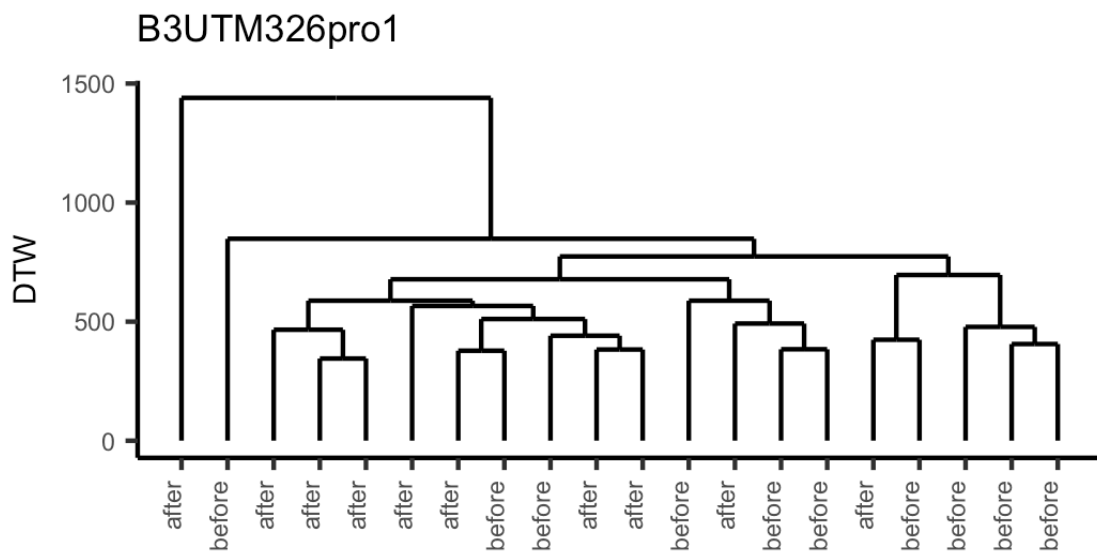


図 13: デンドログラム 10

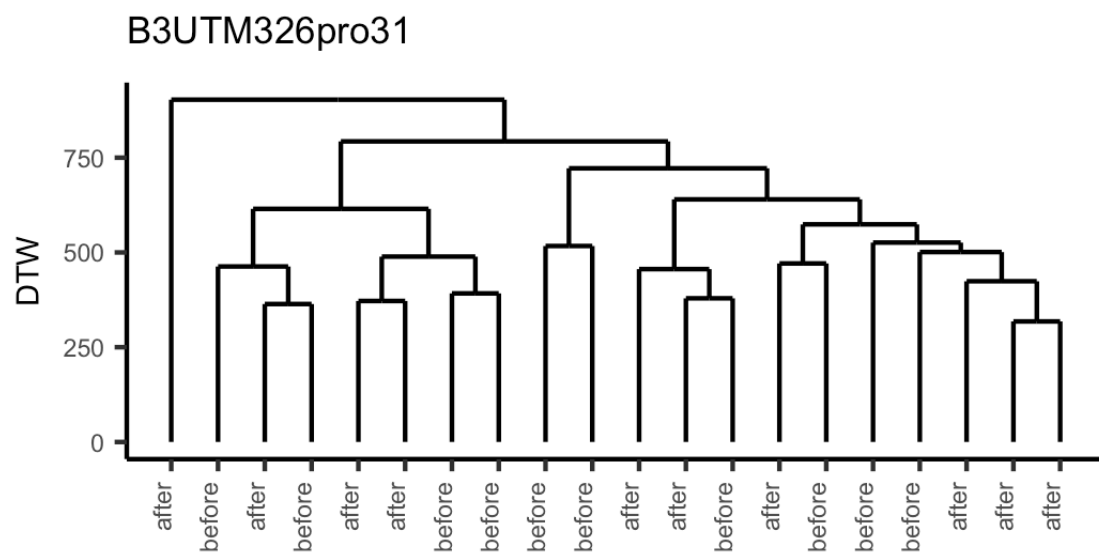


図 14: デンドログラム 11

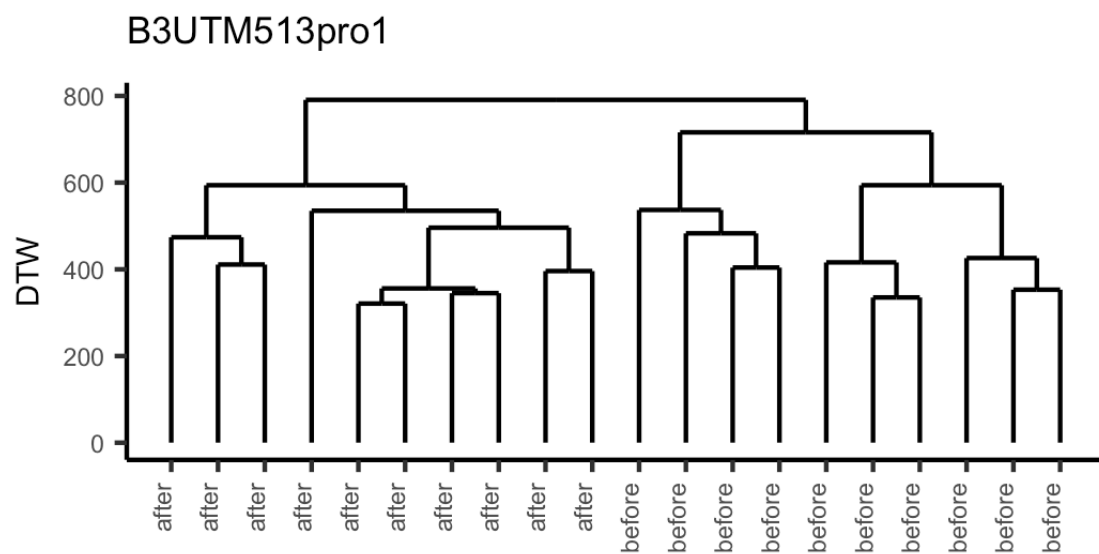


図 15: デンドログラム 12

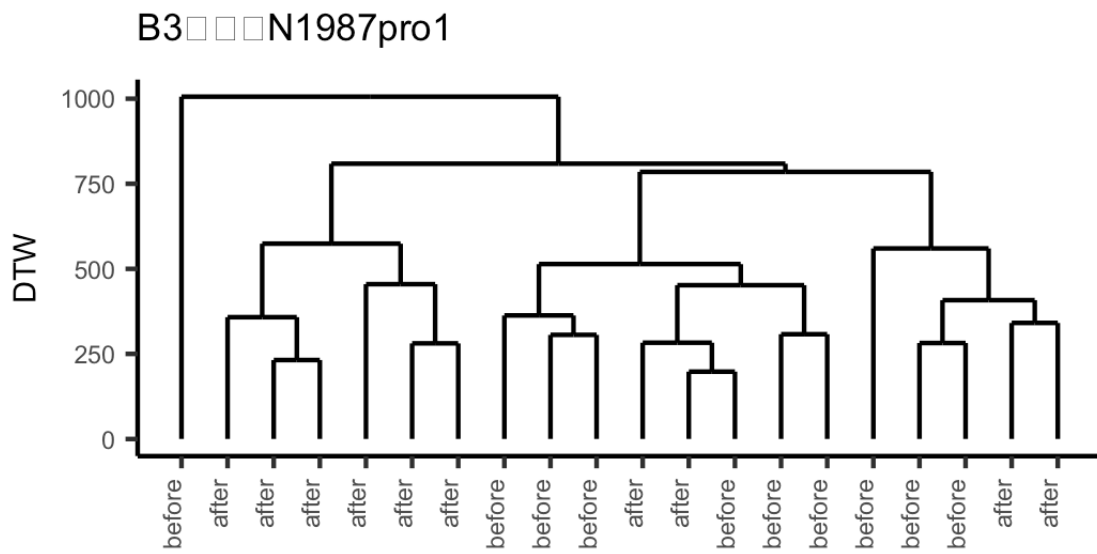


図 16: デンドログラム 13

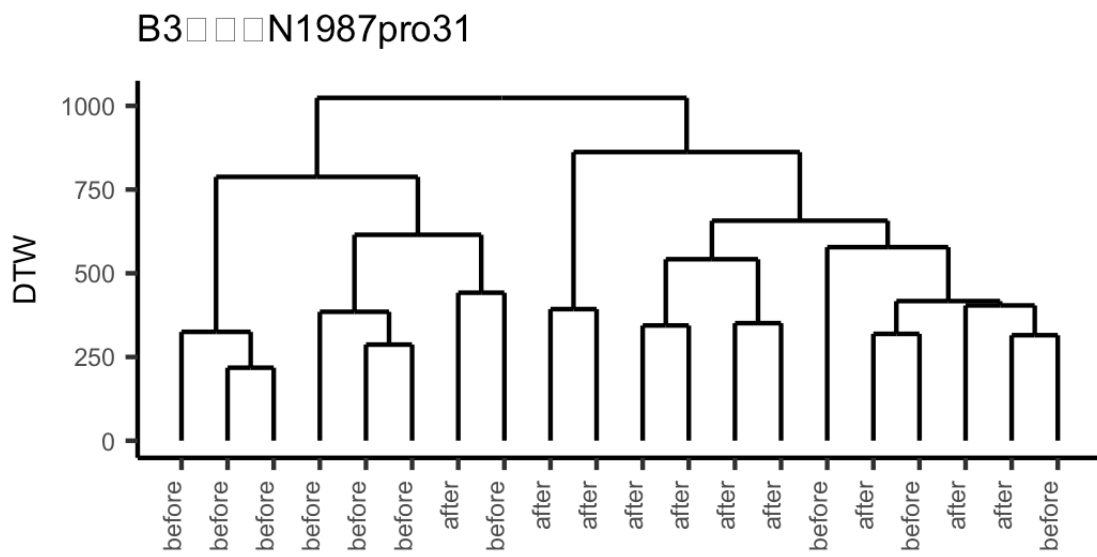


図 17: デンドログラム 14

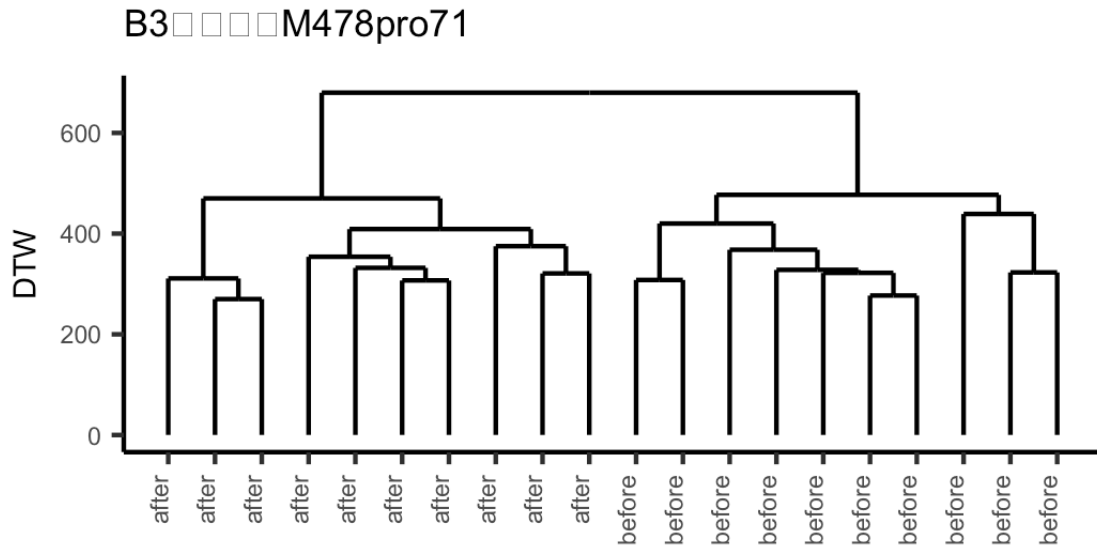


図 18: デンドログラム 15

4 考察・今後の課題

参考文献

- [1] Donald Berndt and James Clifford. Finding patterns in time series: a dynamic programming approach. pp. 229–248, 02 1996.
- [2] Ahlame Douzal Chouakria and Panduranga Naidu Nagabhushan. Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, Vol. 1, No. 1, pp. 5–21, Mar 2007.
- [3] Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, Vol. 28, No. 4, pp. 511–533, Nov 2002.
- [4] Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, Vol. 6, No. 1, p. 17, Aug 2017.
- [5] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, Vol. 27, No. 12, p. 124007, 2011.
- [6] Yuhei Umeda. Time series classification via topological data analysis. *人工知能学会論文誌*, Vol. 32, No. 3, pp. 1–12, 2017.

- [7] Gurjeet Singh, Facundo Mémoli, and Gunnar E. Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, 2007.
- [8] Nina Amenta, Dominique Attali, and Olivier Devillers. Complexity of delaunay triangulation for points on lower-dimensional polyhedra. In *SODA*, 2007.
- [9] Dominique Attali and Jean-Daniel Boissonnat. Complexity of the delaunay triangulation of points on polyhedral surfaces. *Discrete Comput. Geom.*, Vol. 30, No. 3, pp. 437–452, September 2003.
- [10] Patrick Truong. An exploration of topological properties of high-frequency one-dimensional financial time series data using tda. Master’s thesis, KTH, Mathematical Statistics, 2017.