情報処理学会論文誌　Vol.0　No.0　1–8 (??? 1959)

Persistence Landscapesを用いた新型コロナウイルス感染症時系列データの位相的データ解析による分析

石原 優生[1]　佐藤 好久[2]

受付日 xxxx年0月xx日, 採録日 xxxx年0月xx日

概要：情報技術の発展に伴い，身の回りには膨大なデータが溢れている．その中から有益な情報を抽出するデータ解析にはさらなる発展が求められている．位相的データ解析（以下 TDA）は位相幾何学を基としており，データの「形」に着目した手法である．TDA はデータの分布モデルを必要としないため，あらゆるデータに対して分析を行うことができるとされている．本研究では，TDA の中でも Persistence Landscapes に着目し，日本国内の新型コロナウイルス感染推移の第 1 波から第 5 波までの特徴付けをすることができるかを検証した．また，Persistence Landscapes を用いて作成した曲線の時系列データから特徴量抽出が可能かどうかも検証した．

キーワード：情報数学，確率・統計，大規模データアルゴリズム

Analysis of novel coronavirus infection time series data by using topological data analysis with Persistence Landscapes

Yusei Ishihara[1]　Yoshihisa Sato[2]

Received: xx xx, xxxx, Accepted: xx xx, xxxx

Abstract: With the development of information technology, huge amounts of data are all around us. Further development is required in data analysis to extract useful information from the data. Topological data analysis (TDA) is based on topology and is a method that focuses on the ”shape” of data. Since TDA does not require a data distribution model, it is said to be able to perform analysis on any data. In this study, we focus on Persistence Landscapes among TDA and verify whether it is possible to characterize the transition of the new coronavirus infection in Japan from the first wave to the fifth wave. Furthermore we verify whether it is possible to extract features from time-series data of curves created by using Persistence Landscapes.

Keywords: Information Mathematics, Probability & Statistics, Large-scale Data Algorithms

1. はじめに

情報技術の発展に伴い，身の回りには膨大なデータが溢れている．その中から有益な情報を抽出するデータ解析にはさらなる発展が求められている．本論文では，データ解析の 1 つの手法である位相的データ解析（Topological Data Analysis）に着目し時系列データの分析に対する有用性について考えていく．

位相的データ解析（以下 TDA）は位相幾何学を基としており，データの「形」に着目した手法である．従来のデータ解析では，解析するデータを既知の分布モデルに当てはめて解析を行うため，分布に当てはめることのできないデータに対しては解析が難しいとされてきた．それに対し TDA はデータの分布モデルを必要としないため，あらゆるデータに対して分析を行うことができるとされている．

先行研究では，日本国内の新型コロナウイルス感染症時系列データから得られた第 1 波から第 5 波のパーシステント図を元に生成元推移曲線を作成し，その曲線の曲率，フーリエ解析により生成元推移曲線を分析することで第 1

[1] 九州工業大学情報工学部知能情報工学科
[2] 九州工業大学大学院情報工学研究院　知能情報工学研究系

波から第 5 波までの特徴付けを行った．

本研究では，Persistence Landscapes を用いて同データから得られたパーシステント図から生成される曲線の時系列データを作成し，先行研究と同様にフーリエ解析を行うことによって第 1 波から第 5 波までの特徴付けをすることが出来るか検証した．また，作成した曲線の時系列データから特徴量抽出が可能かどうかも検証し，自ら考えた特徴量についても検証した．

2. パーシステントダイアグラム

まず，パーシステントホモロジー群について説明する．パーシステントホモロジー群は TDA において重要な概念で，図形の連結成分や輪っか，空洞といった構造に注目することでデータの「形」を情報として抽出することができる．図 1 の $t = 0$ における点に対して，各点を中心とした円を考える．時刻につれ各点の円の半径を大きくしていくと，時刻 $t = 1$ で 2 つの輪っかが生成され，時刻 $t = 2$ では下の輪っかが消滅している．この消滅した輪っかについては発生時刻（birth）が $t = 1$，消滅時刻（death）が $t = 2$ と表すことができる．birth と death の 2 つの要素からなる集合 (birth,death) の集まりをパーシステントダイアグラムとよぶ．

3. Persistence Landscapes

Persistence Landscapes とは，TDA の分野において使用される手法で，特にパーシステントホモロジーから派生したデータのトポロジカル特徴を分析し、可視化する方法である。Persistence Landscapes を用いることでパーシステントダイアグラムの離散的情報を連続関数にすることできる．

パーシステントダイアグラムの離散的情報 $(1,5),(4,6)$ を図 2 に示す．図 2 の $(1,5),(4,6)$ から対角線に対して，図 3 のように縦軸と横軸方向に線を伸ばす．図 3 の対角線を横軸に取り直す（図 4）．図 4 の 2 つの曲線の交差がなくなるように曲線を変形し，高さが大きい順に λ_1, λ_2 とする（図 5）．このようにして Persistence Landscapes は作成される．

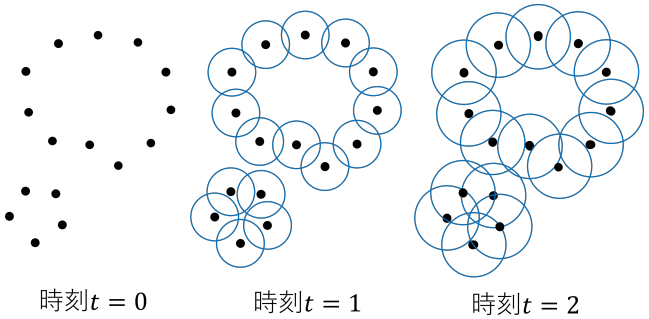

図 1　パーシステントホモロジー群の基本的な概念

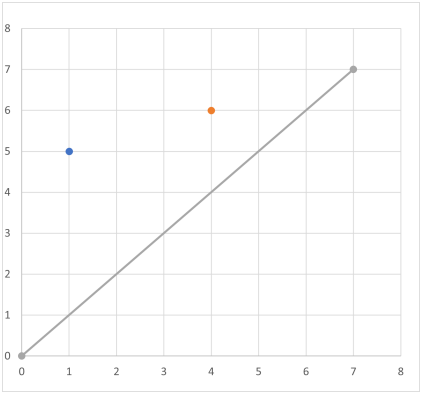

図 2　$(1,5),(4,6)$ のパーシステントダイアグラム

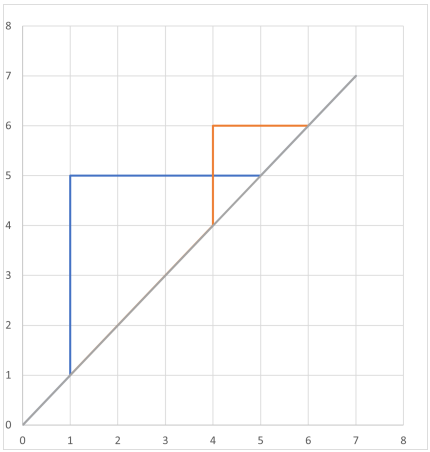

図 3

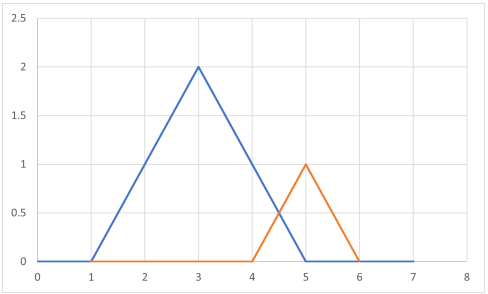

図 4　図 3 の対角線を横軸に取り直した曲線

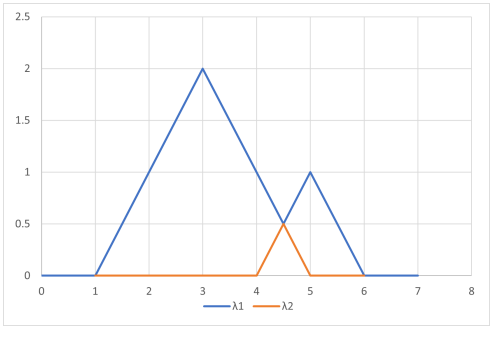

図 5　$(1,5),(4,6)$ の Persistence Landscapes

4. 新型コロナウイルス感染症時系列データの位相的データ解析による分析

今回の研究の手順を以下に示す.

(1) 元データ（新型コロナウイルス感染症時系列データ）の前処理

(2) パーシステントダイアグラム（1 日単位）の計算

(3) 実験 1：第 n 波単位での分析

(3-1) 第 n 波ごとのパーシステントダイアグラムの作成

(3-2) (3-1) のパーシステントダイアグラムの Persistence Landscapes（曲線）を作成

(3-3) (3-2) の曲線に対してフーリエ解析を行う

(4) 実験 2：週単位での分析

(4-1) 週ごとのパーシステントダイアグラムの作成

(4-2) (4-1) のパーシステントダイアグラムの Persistence Landscapes（曲線）を作成

(4-3) (4-2) の曲線に対してフーリエ解析を行う

(5) 実験 3：1 日単位の分析

(5-1) (2) のパーシステントダイアグラムの Persistence Landscapes（曲線）を作成

(5-2) tsfresh を用いて（5-1）の曲線の時系列データの特徴量抽出を行う.

(5-3) Persistence Landscapes の時系列データに対する特徴量の提案

手順 (1) の元データの前処理について説明する．日本国内（沖縄県を除く）での新型コロナウイルス感染症について，2020 年 1 月 16 日から 2022 年 4 月 12 日の都道府県ごとの感染者数データを元データとして使用する（[5] より引用）.

TDA では位置情報を必要とするが，上記のデータには位置情報が記述されていないため次のように位置情報を設定した.

各都道府県の感染者数を各都道府県の中で 1 番目に感染者数が多い市区町村，2 番目に多い市区町村，3 番目に多い市区町村，その他の市区町村の 4 つに分ける．各都道府県の感染者数を 4 つに分けた割合を表 1 に示す．各都道府県の 1 日ごとの感染者数データと表 1 を用いて 1〜3 番目に多い市区町村およびその他の市区町村に配置する点の数を決める．例えば北海道に 100 人の感染者がいた場合，札幌市に 38 人，旭川市に 11 人，函館市に 7 人，その他の市区町村に 44 人となる．点の配置は，まず各市区町村の役所の経緯度を中心とした円を考える（その他の市区町村については各都道府県の中央の経緯度を中心とした円を考える）. このとき感染者数は役所に近ければ近いほど多いと考え，正規分布を用いて円の中にプロットした．なお，このときの円の大きさは人口密度を使用し累乗近似を用いて設定した．これらの配置した点を 1 日目から $D_1, D_2, \ldots, D_n$ とする．図 6〜9 で地図上にプロットした例をいくつか示す.

表 1　各都道府県の市区町村ごとの感染者数の割合 (1〜3 番目とその他)

	1	2	3	その他
北海道	0.38	0.11	0.07	0.44
青森県	0.23	0.21	0.21	0.35
岩手県	0.24	0.22	0.15	0.54
⋮	⋮	⋮	⋮	⋮
宮崎県	0.55	0.09	0.05	0.31
鹿児島県	0.51	0.10	0.10	0.29

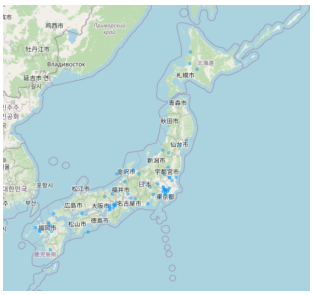

図 6　2020 年 4 月 1 日

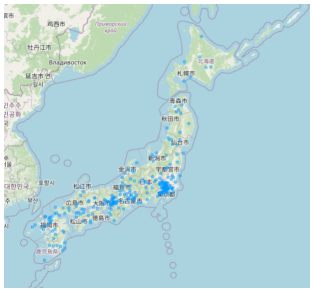

図 7　2020 年 12 月 30 日

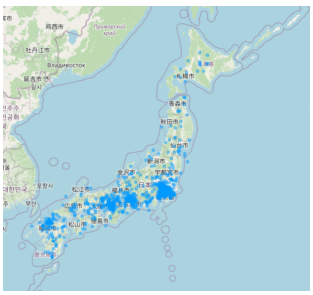

図 8　2021 年 8 月 27 日

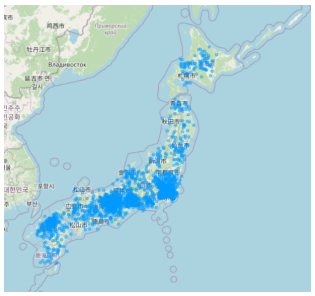

図 9　2022 年 2 月 23 日

手順 (2) では手順 (1) で作成した $D_1, D_2, \ldots, D_n$ に対しパーシステントダイアグラムを計算する．ここで，2020 年 1 月 16 日から 2021 年 9 月 30 日（第 5 波の終わり）の 624 日分のパーシステントダイアグラム $PD_1, PD_2, \ldots, PD_{624}$ を得られる.

5. 研究結果

5.1 実験 1：第 n 波単位での分析

実験 1 では第 1 波から第 5 波に着目して分析を行う．表 2 のようにそれぞれの波の始まりと終わりの区間を定める．表 2 の () の中の数字は 2020 年 1 月 16 日を 1 日目とした場合の経過日数である．手順 (2) で作成したパーシステントダイアグラムをそれぞれの波の区間ごとに合成して，第 1 波から第 5 波の 5 つのパーシステントダイアグラムを作成する．5 つのパーシステントダイアグラムを図 10〜14 に示す． 図 10〜14 のパーシステントダイアグラムの Persistence Landscapes（λ_1）を図 15 に示す．そして作成した Persistence Landscapes に対してフーリエ解析を行った結果を図 16 に示す.

表 2　第 1 波から第 5 波の区間

	開始	終了
第 1 波	2020/4/11(77)	2020/5/15(121)
第 2 波	2020/7/10(177)	2020/9/15(244)
第 3 波	2020/12/16(336)	2021/2/15(397)
第 4 波	2021/4/1(442)	2021/6/20(522)
第 5 波	2021/7/16(548)	2021/9/30(624)

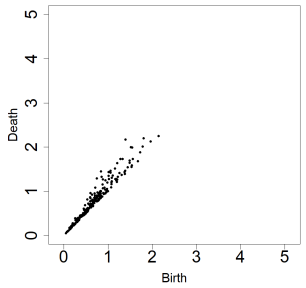

図 10　第 1 波

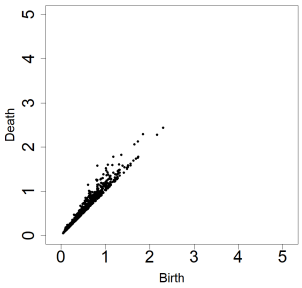

図 11　第 2 波

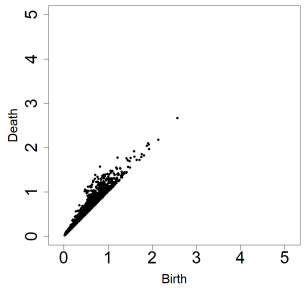

図 12　第 3 波

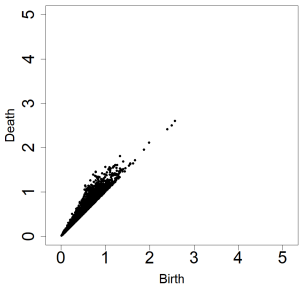

図 13　第 4 波

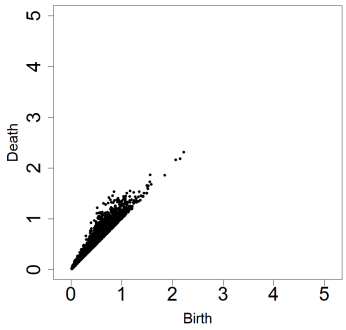

図 14　第 5 波

5.2　実験 2：週単位での分析

実験 2 では週単位でのパーシステントダイアグラムに着目して分析を行う．手順 (2) で作成したパーシステントダイアグラムを 7 日ごとに合成して，2020 年 1 月 16 日から 2021 年 9 月 20 日の計 89 週のパーシステントダイアグラム $PD_{w1}, PD_{w2}, \ldots, PD_{w89}$ を作成する．

89 週のパーシステントダイアグラムからそれぞれ Persistence Landscapes$PL_{w1}, PL_{w2}, \ldots, PL_{w89}$ を作成する．作

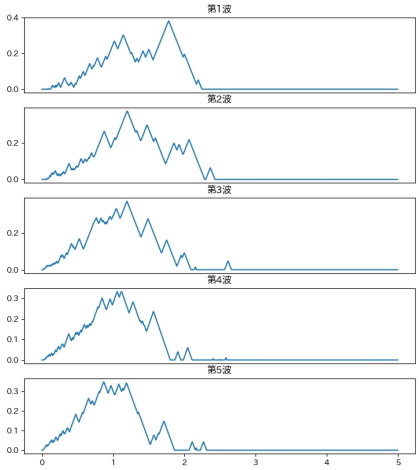

図 15　Persistence Landscapes

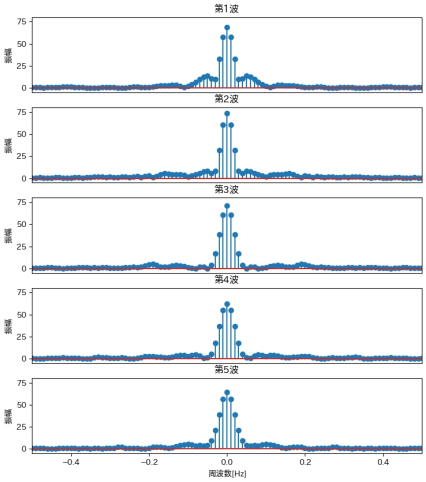

図 16　フーリエ解析

成した Persistence Landscapes の中で，第 n 波の区間に属するものを図 17〜25 で示す．　そして作成した Persistence Landscapes に対してフーリエ解析を行った結果を図 26〜34 に示す．

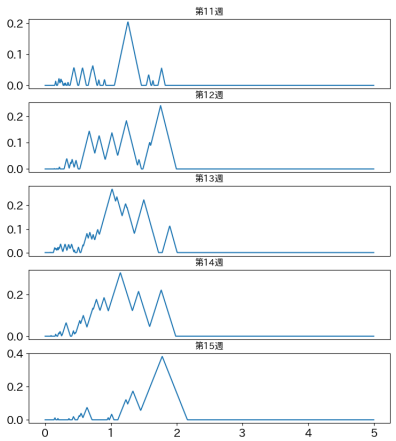

図 17　第 11〜15 週（第 1 波）

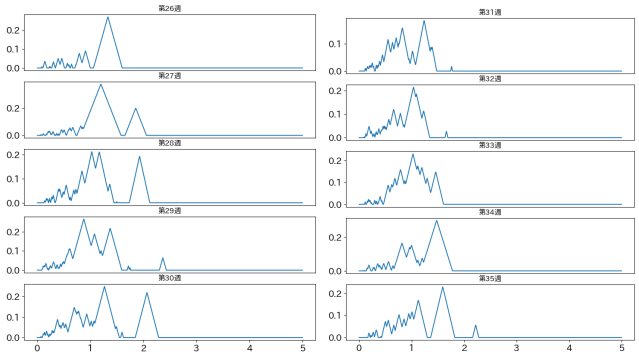

図 18　第 26〜30 週（第 2 波）

図 19　第 31〜35 週（第 2 波）

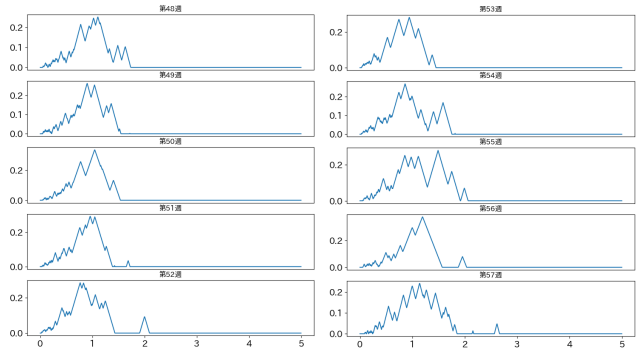

図 20　第 48〜52 週（第 3 波）

図 21　第 53〜57 週（第 3 波）

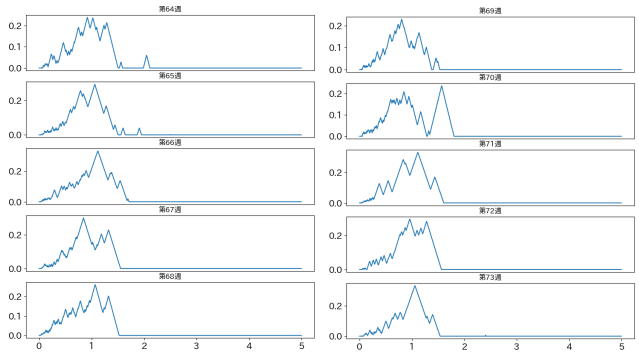

図 22　第 64〜68 週（第 4 波）

図 23　第 69〜73 週（第 4 波）

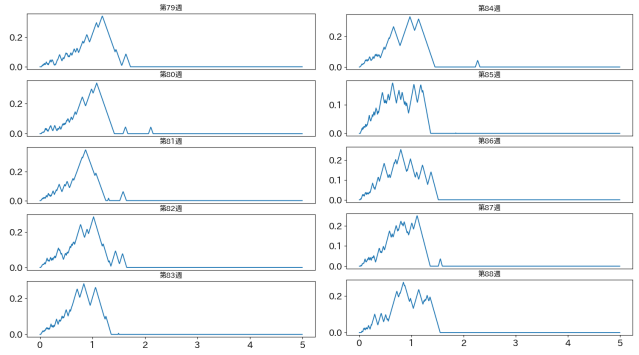

図 24　第 79〜83 週（第 5 波）

図 25　第 84〜88 週（第 5 波）

5.3　実験 3：1 日単位での分析

実験 3 では 1 日単位のパーシステントダイアグラムに着目する．手順 (2) で作成した $PD_1, PD_2, \ldots, PD_{624}$ に対して，Persistence Landscapes$PL_1, PL_2, \ldots, PL_{624}$ を作

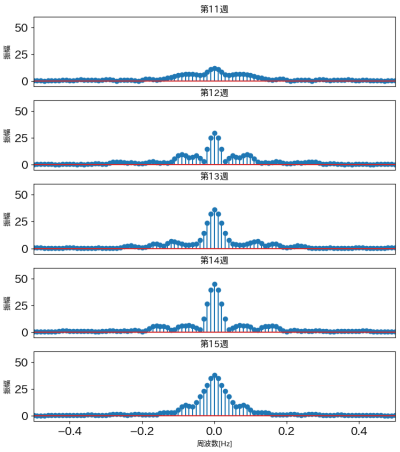

図 26　第 11〜15 週（第 1 波）

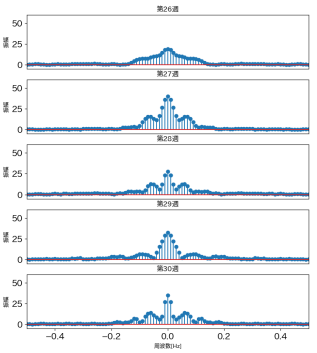

図 27　第 26〜30 週（第 2 波）

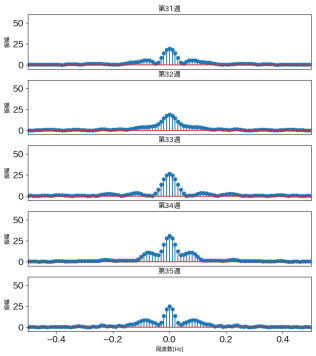

図 28　第 31〜35 週（第 2 波）

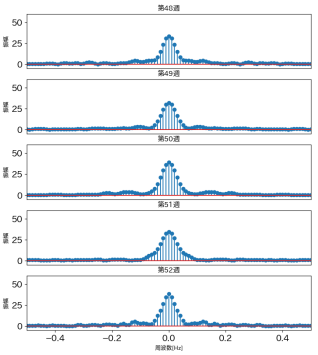

図 29　第 48〜52 週（第 3 波）

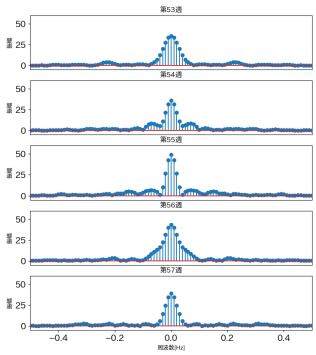

図 30　第 53〜57 週（第 3 波）

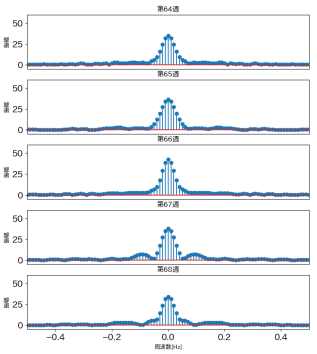

図 31　第 64〜68 週（第 4 波）

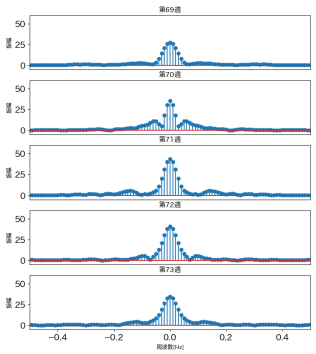

図 32　第 69〜73 週（第 4 波）

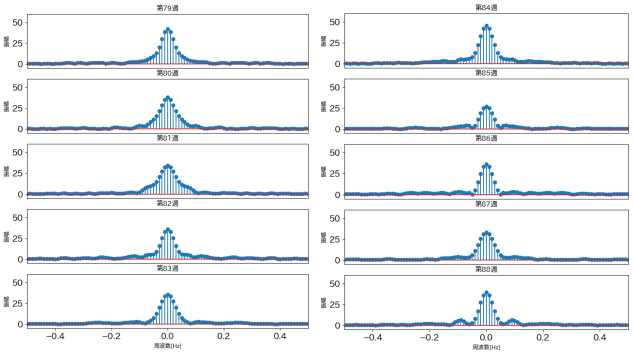

図 33　第 79〜83 週（第 5 波）　**図 34**　第 84〜88 週（第 5 波）

成する．ここでは 1 つの Persistence LandscapesPL_i は 1000 次元の数値データから構成されており，Persistence Landscapes$PL_1, PL_2, \ldots, PL_{624}$ は 1000 次元の数値データの時系列データとなっている．

まず，Python のライブラリである「tsfresh」を用いて，時系列データ $PL_1, PL_2, \ldots, PL_{624}$ から特徴量抽出が可能かどうかを検証する．

「tsfresh」で扱うデータとして計算量の都合上 Persistence Landscapes$PL_1, PL_2, \ldots, PL_{624}$ の中から第 1 波の区間にあたる $PL_{77}, \ldots, PL_{121}$ を抜き出して特徴量抽出を行う．使用したデータは表 3 に示す．「tsfresh」を用いて表 3 のデータの特徴量抽出を行った結果を表 4 に示す．783000 個の特徴量が抽出されたが，欠損値 (NaN) や全ての次元で同値である特徴量を削除した結果を表 5 に示す．第 1 波の区間にあたる $PL_{77}, \ldots, PL_{121}$ を抜き出して「tsfresh」により特徴量抽出を行った結果，1606 個の特徴が抽出された．「tsfresh」で抽出できる特徴量はデータの次元に依存しているので，全ての Persistence Landscapes$PL_1, PL_2, \ldots, PL_{624}$

表 3　使用したデータ

D	TIME	1	2	⋯	150	⋯	999	1000
D77	77	0.0	0.0	⋯	0.043699	⋯	0.0	0.0
D78	78	0.0	0.0	⋯	0.009278	⋯	0.0	0.0
D79	79	0.0	0.0	⋯	0.0	⋯	0.0	0.0
⋮	⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
D119	119	0.0	0.0	⋯	0.0	⋯	0.0	0.0
D120	120	0.0	0.0	⋯	0.0	⋯	0.0	0.0
D121	121	0.0	0.0	⋯	0.0	⋯	0.0	0.0

表 4　抽出した特徴量

D	1	2	3	⋯	782998	782999	783000
D77	0.0	0.0	0.0	⋯	NaN	NaN	NaN
D78	0.0	0.0	0.0	⋯	NaN	NaN	NaN
D79	0.0	0.0	0.0	⋯	NaN	NaN	NaN
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮
D119	0.0	0.0	0.0	⋯	NaN	NaN	NaN
D120	0.0	0.0	0.0	⋯	NaN	NaN	NaN
D121	0.0	0.0	0.0	⋯	NaN	NaN	NaN

表 5　削除後の特徴量

D	1	2	⋯	1605	1606
D77	0.0	0.0	⋯	0.0	0.0
D78	0.0	0.0	⋯	0.0	0.0
D79	0.0	0.0	⋯	0.057731	0.057731
⋮	⋮	⋮	⋱	⋮	⋮
D84	0.001119	0.000001	⋯	0.0	0.0
D85	0.004756	0.000023	⋯	0.0	0.0
⋮	⋮	⋮	⋱	⋮	⋮
D119	0.0	0.0	⋯	0.0	0.0
D120	0.0	0.0	⋯	0.0	0.0
D121	0.0	0.0	⋯	0.0	0.0

を用いて特徴量抽出を行うと，削除する前の特徴量の数である 783000 個の特徴量が抽出される．

「tsfresh」を用いて特徴量抽出が可能であることは確認されたが，実際に有効な特徴量を発見することは難しい．そこで，Persistence Landscapes$PL_1, PL_2, \ldots, PL_{624}$ の曲線の時系列データの特徴量として次のようなものを提案する．

提案する特徴量

Persistence Landscapes の傾きが「+」から「−」に変わるところ
（Persistence Landscapes を構成している数値データ $pl_1, pl_2, \cdots, pl_n$ の中で

$$pl_{i-1} < pl_i > pl_{i+1}$$

を満たす数値データ pl_i）
を Persistence Landscapes の頂点とよぶことにする．また pl_i の値を Persistence Landscapes の頂点の高さとする．提案する特徴量を以下に示す．

- num_pl　・・・　Persistence Landscapes の頂点の数
- sum_pl　・・・　Persistence Landscapes の頂点の高さの総和

同じ Persistence LandscapesPL_i の λ_1, λ_2 の曲線に対する特徴量 num_pl，sum_pl は λ_1 の曲線に対する特徴量 num_pl，sum_pl と λ_2 の曲線に対する特徴量 num_pl，sum_pl のそれぞれの和とする．

Persistence Landscapes$PL_1, PL_2, \ldots, PL_{624}$ に対する特徴量 num_pl，sum_pl を図 35〜40 に示す．　提案した特徴量を評価するために，今回の研究で使用した日本国内（沖縄県を除く）での感染数の推移を図 41 で示す．

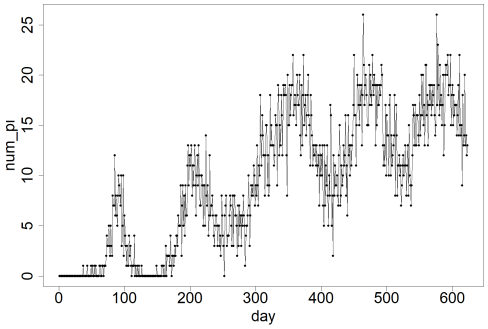

図 35 num_pl(λ_1)

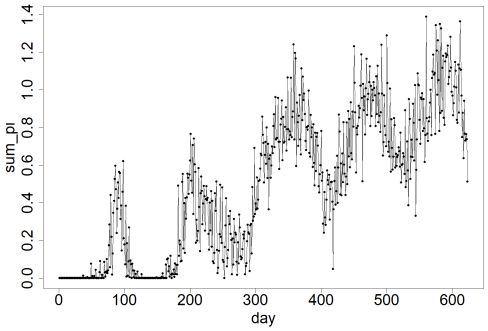

図 36 sum_pl(λ_1)

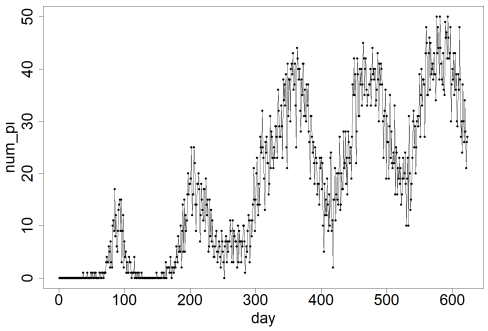

図 37 num_pl(λ_1, λ_2)

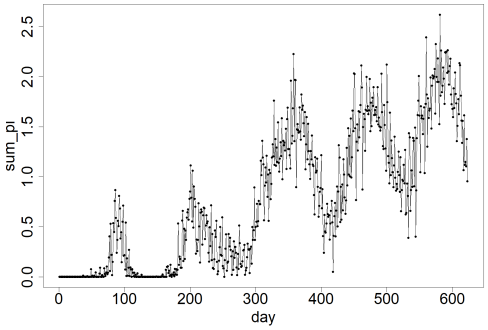

図 38 sum_pl(λ_1, λ_2)

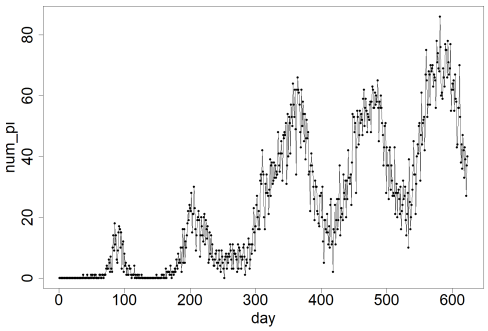

図 39 num_pl($\lambda_1, \lambda_2, \lambda_3$)

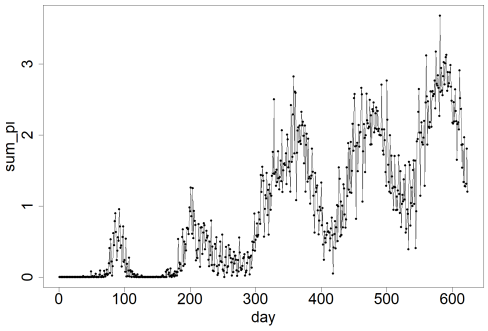

図 40 sum_pl($\lambda_1, \lambda_2, \lambda_3$)

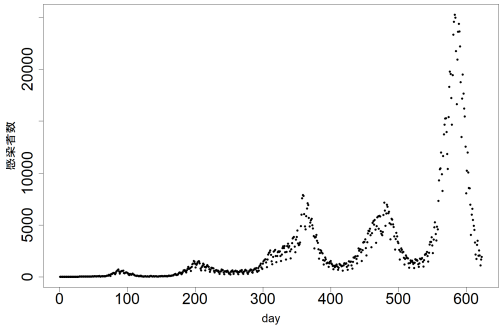

図 41 日本国内（沖縄県を除く）の感染者数の推移

6. 結論・考察

まず第 n 波単位での分析をまとめる．それぞれの Persistence Landscapes を比べると，曲線の中で最も高いところは第 1 波が横軸が 1.8，第 2 波，第 3 波が 1.2，第 4 波が 1，第 5 波が 0.8 であった．曲線の分布は第 1 波は 1 と 1.8 で 2 つのピークがあり，第 2 波は 0.8 から 2 の間に 1 つのピークがあり，第 3 波，第 4 波，第 5 波は 0.7 から 1.3 に 1 つのピークがあった．Persistence Landscapes を見ると第 3 波，第 4 波，第 5 波が似ていた．そこで曲線がどのような周波数で構成されているかを調べるためにフーリエ解析を行った．フーリエ解析の結果より相違点をあげると，第 1 波で振幅が 13 が，第 2 波では振幅が 8，第 3 波，第 4 波，

第 5 波では振幅が 4 以下である周波数（0.04 から 0.06[Hz]）が存在している．第 n 波単位での分析では第 1 波と第 2 波とその他での違いを確認することは出来たが，第 3 波，第 4 波，第 5 波の違いを明確に確認することが困難であった．

そこで，パーシステントダイアグラムを週ごとにまとめ，第 n 波の区間に属する Persistence Landscapes をそれぞれ作成した．第 n 波の区間内での Persistence Landscapes に対するフーリエ解析の結果の推移についてみると，第 2 波と第 5 波は構成している周波数は変わらず振幅の推移はあったが，第 3 波と第 4 波は構成している周波数と振幅ともに推移していた．第 1 波は区間が短いため比較はすることが難しかったが，周波数と振幅ともに推移していた．第 5 波については第 3 波，第 4 波との違いを確認することが出来た．第 3 波と第 4 波は感染者数や感染者数の推移が似ているため，Persistence Landscapes のフーリエ解析結果からは違いを確認することは出来なかった．

Persistence Landscapes はパーシステントダイアグラムに含まれる情報の要約を行っている．第 n 波単位での分析では，感染者の数によって違いが出たのではないかと考える．ただし感染者数が一定数以上の大きな情報（第 3 波から第 5 波）を要約して違いを確認することは難しいと考える．情報量を細分化して，Persistence Landscapes の推移について注目した週単位での分析では，感染者数の推移の仕方によって違いが出たのではないかと考える．そのため，感染者数の推移の仕方が似ている第 3 波と第 4 波については週単位での分析でも違いを確認することは出来なかった．

1 日単位での分析では，Persistence Landscapes の時系列データから特徴量抽出が可能かどうかを検証した．今回作成した Persistence Landscapes は 1000 次元の数値データで構成されている．計算量の都合上第 1 波の区間にあたる Persistence Landscapes の時系列データに対して,「tsfresh」を用いて特徴量抽出を行った．その結果 783000 個の特徴量が抽出され，欠損値や全ての次元同値であるものを削除すると 1606 個の特徴量が抽出された．「tsfresh」で抽出される特徴量の数はデータの次元数に依存しており，全ての Persistence Landscapes を用いて特徴量抽出を行うと，削除前の特徴量の数である 783000 個の特徴量が抽出されることとなる．「tsfresh」を用いて特徴量抽出が可能であることは確認されたが，実際に有効な特徴量を発見することは困難であった．そこで，新たな特徴量として num_pl と sum_pl を提案した．Persistence Landscapes に対する 2 つの特徴量の推移と感染者数の推移を比べると，λ_1 の曲線では 2 つの特徴量の推移のばらつきが大きいもののどちらの特徴量も感染者数の動向を捉えていた．λ_1, λ_2 の曲線，$\lambda_1, \lambda_2, \lambda_3$ の曲線と扱う情報量を増やしていくと，2 つの特徴量の推移のばらつきが小さくなることが確認できた．また，num_pl と sum_pl の特徴量の推移のばらつきを比べると，num_pl のほうが安定していた．Persistence Landscapes の高さは高いほど存在している時間が大きいため持っている情報量が大きい．そのため，sum_pl については，Persistence Landscapes の高さの総和と定義したが，高さに対して重み付けを行うとより安定した特徴量になるのではないかと考える．本研究は都道府県ごとの感染者数を元にして位置情報を持った感染者のデータを作成し，そのデータの 1 次元パーシステントホモロジー群について考えている．1 次元パーシステントホモロジー群について，データの点群が密である場合は存在している時間が短く，データの点群が広がっている場合は存在している時間が長くなる．考えている 2 つの特徴の定義により，これらの特徴量が大きいということは，そのデータの点群の個数が増加していることを表しているだけでなく，点群の広がり具合が拡大していることも表している．これらの特徴量が感染者の動向を捉えていたことは，感染者数の推移に対応して感染範囲が変化していることを表している．つまり，これらの特徴量はデータの点の個数だけでなく，データの点の広がりについても表している特徴量だと考える．

本研究では TDA の中でも Persistence Landscapes に着目をして，日本国内の新型コロナウイルス感染症時系列データの第 1 波から第 5 波の特徴付けを行った．ただし，第 3 波と第 4 波についての違いを確認することは出来なかった．Persistence Landscapes はパーシステントダイアグラムを視覚化や，第 n 波単位や週単位での実験で扱った λ_1 のように情報量の要約を行うことができる．しかし，作成元のパーシステントダイアグラムの情報が大きすぎると，Persistence Landscapes を作成し λ_1 等を抜き出して分析する際に，反映されない情報量が出てくるため，扱うパーシステントダイアグラムの情報量や抜き出す λ_i を適したものにすることが必要である．また，提案した 2 つの特徴量 num_pl と sum_pl については，最適化と本研究で使用したデータ以外でも適切な特徴量と言えるのかはさらなる検証が必要である．

参考文献

[1] 平岡裕章：タンパク質構造とトポロジー パーシステントホモロジー群入門，共立出版 (2013).

[2] 川原晃祐：位相的データ解析を利用した新型コロナウイルス感染症時系列データの分析 ，九州工業大学大学院, 修士論文 (2023).

[3] Peter Bubenik：Statistical topological data analysis using persistence landscapes ，Journal of Machine Learning Research 16 (2015) 77-102.

[4] NHK：都道府県ごとの感染者数 ，入手先 ⟨https://www3.nhk.or.jp/news/special/coronavirus/data/⟩

[5] 都道府県庁・市区町村役所の緯度経度ホームページ, 入手先 ⟨https://www.gaoshukai.com/20/15/0026/⟩