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

川原晃祐[1]　佐藤好久[2]

概要：現在ビッグデータの分析手法として TDA が注目されており、多くの企業が研究を行っている。位相的データ解析(以下 TDA)は位相幾何学を基としており、データの位相的形状に着目した手法である。また、新型コロナウイルスが猛威をふるい続けており、その分析も急務である。従来の統計的データ解析手法における感染症の分析では分布モデルを必要とするが、TDA はデータのもつ「形」に着目することでデータの分布モデルを必要としない。今回の研究では TDA を応用した手法にて特に新型コロナウイルスの第 1 波から第 5 波までの特徴付けを TDA を用いて行うことができないか検証した。

キーワード：情報数学，計算幾何学，大規模データアルゴリズム

Analysis of time series data of novel coronavirus infection using topological data analysis

KOSUKE KAWAHARA[1]　YOSHIHISA SATO[2]

Abstract: TDA is currently attracting attention as a method for analyzing big data, and many companies are conducting research on this topic. Topological data analysis (TDA) is based on topology and focuses on the shape of data. In addition, new coronaviruses continue to rage, and analysis of these viruses is an urgent task. Conventional statistical data analysis methods for analyzing infectious diseases require a distribution model, but TDA focuses on the "shape" of the data and does not require a distribution model. In this study, we examined the possibility of using TDA to characterize the first through fifth waves of new coronaviruses in particular.

Keywords: information mathematics, computational geometry, Large-scale data algorithms

1. はじめに

ビッグデータとは、人間では全体を把握することが難しい巨大なデータ群のことであり、近年、社会情勢の変化や関連技術の進化によってこれまで以上の注目を集めている。その進化や変化に伴い、ビッグデータから有益な情報を抽出するデータ解析技術の進化が求められている。近年ではデータ解析技術の新しい手法である位相的データ解析(Topological Data Analysis)が注目されており、参考文献[1]にもあげられているように多くの企業が研究を行っている。

位相的データ解析(以下 TDA)は位相幾何学を基としており、データの形状に着目した手法である。この手法を用いれば、従来の手法では知りえなかったデータの知見を取り出すことができる。また、高次元のデータにも解析を行うことができるといった特徴や、データのもつ「形」に着目することでデータの分布モデルを必要としないといった特徴もある。また、今回は TDA の手法の中でパーシステントホモロジー群を選択している。

パーシステントホモロジー群について説明する。n 次元離散点データを中心とする n 次元球体を考え、その球の半径 r を大きくしていくと、連結成分や輪っかが発生したり消滅したりする。その連結成分や輪っかの発生時刻(birth)、消滅時刻(death)、数、サイズなどの特徴付けを行うことを今回は TDA を実行すると表現している。なお、0 次元パーシステントホモロジー群は連結成分を生成元とする加群であり、1 次元パーシステントホモロジー群は輪っかを生成元とする加群である。今回は 1 次元パーシステントホモロジー群に着目して研究を行っている。

現在新型コロナウイルスは第 8 波を迎え、1 日の感染者数は 20 万人をこえる日もあるなど、国内で感染初確認から 3 年が経過した今もなお甚大な影響を及ぼし続けてい

1 九州工業大学 情報工学府 学際情報工学専攻
2 九州工業大学情報工学研究院　知能情報工学研究系

る。そんな中影響を少しでも抑えるため新型コロナウイルスのデータ分析が求められている。

感染症の研究では古くから SIR 数理モデルが多く用いられている。SIR とは、感染予備軍（Susceptible)、感染者（Infected)、感染症から回復した人（Recovered）の頭文字であり、SIR 数理モデルは感染者が予備軍と接触すると一定の確率で感染すると想定し、感染者数を予測するものである。

Zika ウイルスを TDA を用いて分析した参考文献[2]にもあるように TDA では分布モデルを必要としない。しかし、参考文献[2]の論文では TDA を実行しパーシステントダイアグラムを分析しただけであるが、著者が行った研究では機械学習への応用も考えパーシステントダイアグラムをベクトル化して分析を行い、その結果にて感染症の分野に TDA は有効であると確認した。

本論文は、参考文献[10]のシリカガラスの構造解析を行った論文より着想を得て、TDA の実行結果に対して曲線を作成することで分析を行った。今回の研究では特に新型コロナウイルスの第 1 波から第 5 波までの特徴付けを TDA を用いて行うことが出来ないか検証した。

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。下図のような点群があるとする。この点群に対しフィルトレーションを用い、各点に置いた円の半径を徐々に大きくすることを考える。

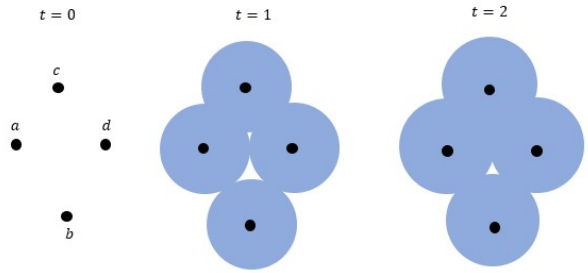


図 1：パーシステントホモロジー群

半径を徐々に大きくすると $t = 1$ で 2 つの穴が生じ、時刻 $t = 2$ で一方の穴が消滅していることがわかる。 $t = 2$ で消滅した穴については発生時刻(birth)が 1、消滅時刻(death)を 2 と表すことができ、これを対応付けた $(\text{birth}, \text{death}) = (1,2)$ を birth、death の時刻対とよぶ。この birth、death の時刻対の集まりをパーシステントダイアグラムとよぶ。パーシステントダイアグラムを可視化する際には横軸を発生時刻(birth)、縦軸を消滅時刻(death)とし、birth、death の時刻対の集まりを平面上にプロットする。これにより得られた散布図をパーシステント図とよぶことにする。

3. TDA の有用性の確認

まず、第一著者が TDA の感染症の分野への有用性を確かめるために行った研究の概要と結果について述べる。

目的は、第一著者が行った TDA を応用した手法により実際の感染者数の動向を捉えることで、感染症の分野への TDA の有用性を検証するものであった。

第一著者が行った卒業研究の手法について述べる。

- 元データ（感染症データ）の前処理(手順 1)
- TDA の実行、パーシステントダイアグラムの計算(手順 2)
- パーシステントダイアグラムのベクトル化(手順 3)
- k 近傍法により解析(手順 4)

各手順について詳細に説明する。

手順 1 の元データの前処理としては、九州地方の各県における RS ウイルス感染症の週ごとの感染者数を使用し（参考文献[3]より引用)、各県の県庁所在地を中心として感染者数と同じ数の点を各県にプロットする。 第 1 週から第 n 週のプロットした点群データを $D_1, D_2, \ldots, D_n$ とする。

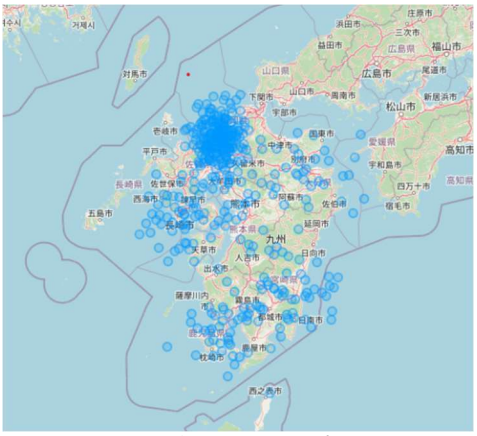


図 2：九州の地図上にプロット

手順 2 では上記の点群データ D_i に対し TDA を実行することによりパーシステントホモロジー群を計算し、パーシステントダイアグラム $PD_1, PD_2, \ldots, PD_n$ を得る。

手順 3 では、手順 2 で得られたパーシステントダイアグラムに含まれる birth、death の時刻対の個数であるベッチ数を求める。その際生存区間(birth-death の値)のデータをもとに階級値を設け、それぞれの階級に存在している birth、death の時刻対の数を並べてできる数列を特徴ベクトルとした。階級値が 5 つあれば 5 次元のベクトルということになる。

手順 4 ではベクトル同士のユークリッド距離を用い k 近傍法を行った。具体的には、第 4 週であった場合は第 4 週のベクトルと前後それぞれ 3 週分の第 1 週、第 2 週、第 3 週、第 5 週、第 6 週、第 7 週とのベクトルとの距離を計測し、その平均を求める。第 1 週から第 3 週、第 101 週から第 104 週については前後 3 週分のデータが確保できないため図 2 では便宜的に 0 としている。卒業研究の結果と、実際の RS ウイルス感染症の感染者数の推移のデータを次に示す。

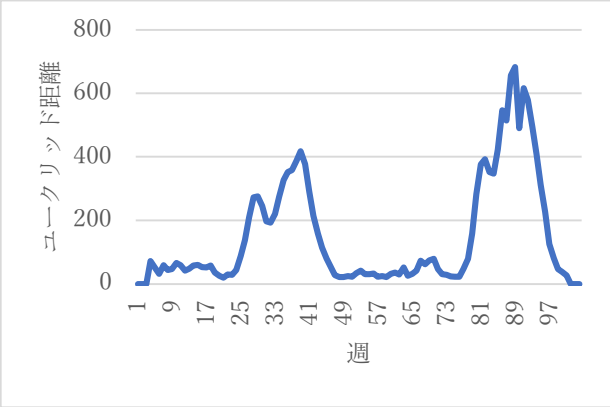


図 3：卒業研究の結果

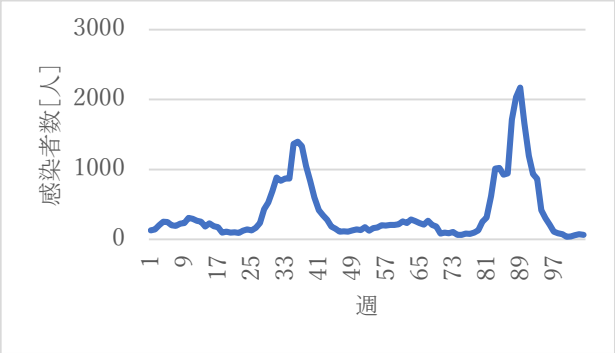


図 4：RS ウイルス感染症の週ごとの感染者数

上記の図 3 が図 4 の実際の感染者数の動向を捉えていると判断したため、TDA は感染症の分野に有用であると結論づけることができる。

4. 本研究の内容

卒業研究の内容により TDA の有用性がわかったので、今回の研究では TDA を用いて新型コロナウイルスの第 1 波から第 5 波までの特徴付けができないか検証した。

まず、今回自分が行った研究の手法について述べる。

- 元データ（感染症データ）の前処理(手順 1)
- TDA の実行、パーシステントダイアグラムの計算(手順 2)
- 第 n 波ごとのパーシステント図の作成(手順 3)
- 生成元推移曲線の作成(手順 4)
- 曲率、フーリエ変換により生成元推移曲線の特徴付けを行う(手順 5)

各手順について詳細に述べる。

手順 1 の元データの作成方法について説明する。

日本列島(沖縄県を除く)における新型コロナウイルス感染症の 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のデータを使用する(参考文献[5]より引用)。

TDA 解析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考えた。

各県の感染者数を各県の中で 1 番感染者数が多い市区町村、2 番目に感染者数が多い市区町村、3 番目に感染者数が多い市区町村、その他の市区町村の 4 つに分ける。その際の各県の感染者数を分ける割合は各県のニュースサイトなどを参照して設定した。割合については次の表 1 に示す。

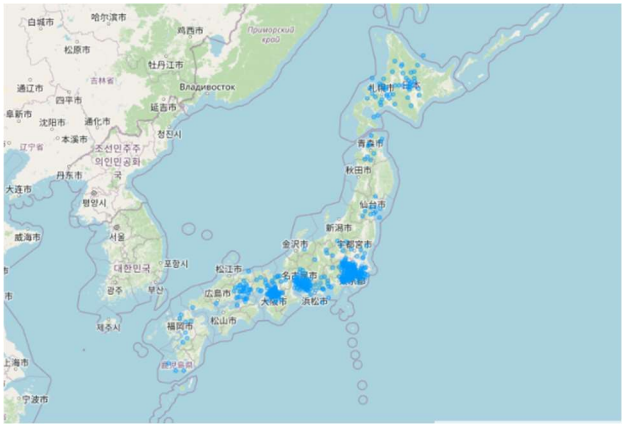
表 1：各県の感染者数の市区町村ごとの割合（1～3 番目とその他）

	1	2	3	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
宮城県	0.62	0.17	0.05	0.16
秋田県	0.43	0.11	0.11	0.35
…	…	…	…	…
宮崎県	0.55	0.09	0.05	0.31
鹿児島県	0.51	0.1	0.1	0.29

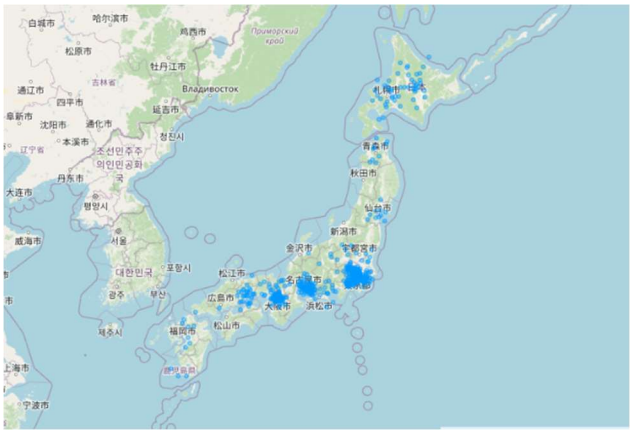
各県の日ごとの感染者数のデータと表 1 の市区町村ごとの感染者数の割合を用いることで各市区町村及びその他に配置する点の数を決める。例えば北海道に 100 人の感染者がいた場合、札幌市に 38 人、旭川市に 11 人、函館市に 7 人、その他の市区町村に 44 人と 4 つに分けることとする。

点の配置については、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（参考文献[6]より引用)。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。その他に分類された点の配置については各都道府県の中央部の経緯度を設定し、同様に正規分布を用い配置を行った。

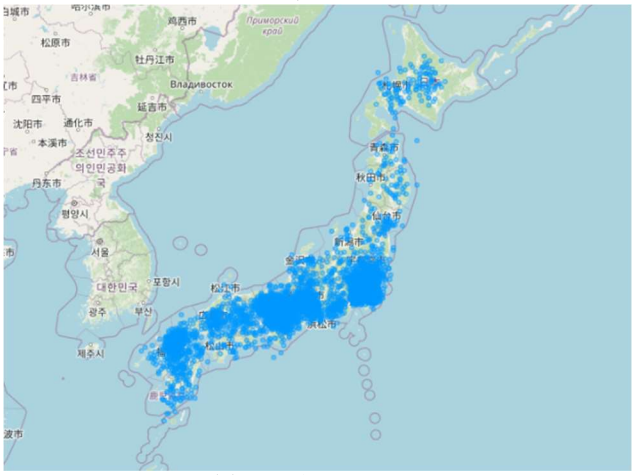
これらの 1 日目から n 日目までの配置した点群データを $D_1, D_2, \ldots, D_n$ とする。点群データを実際の日本地図上に配置した例を次の図 5 に示す。



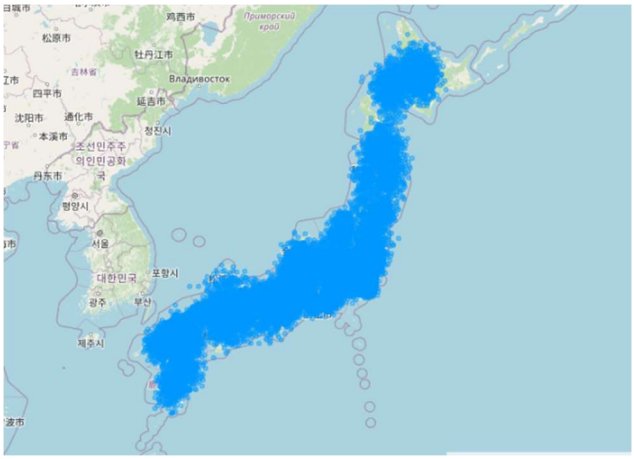
(a)2020 年 4 月 4 日



(b) 2020 年 10 月 29 日



(c) 2021 年 9 月 6 日



(d) 2022 年 3 月 25 日

図 5：日本地図上にプロット

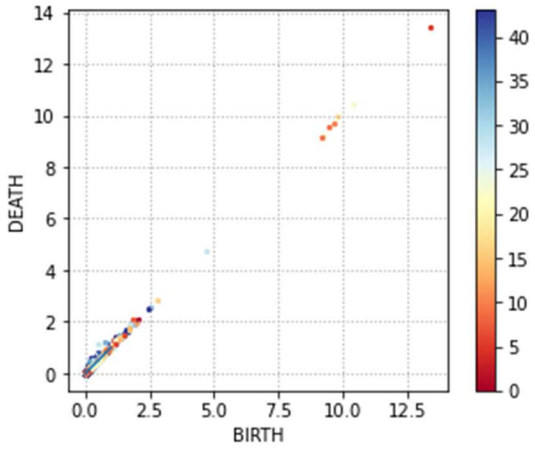
手順 2 では手順 1 で作成した$D_1, D_2, \ldots, D_n$ に対し TDA を実行する。TDA を適用することで 2020 年 1 月 16 日～2022 年 4 月 12 日の 818 日分のパーシステントダイアグラム PD_1, PD_2, PD_{818} が得られる。

手順 3 では、今回は新型コロナウイルスの第 1 波から第 5 波に着目するため、それぞれの波にあたる時期のパーシステントダイアグラムを抜き出し、合成する。それぞれの波の時期については次の表のように設定した。

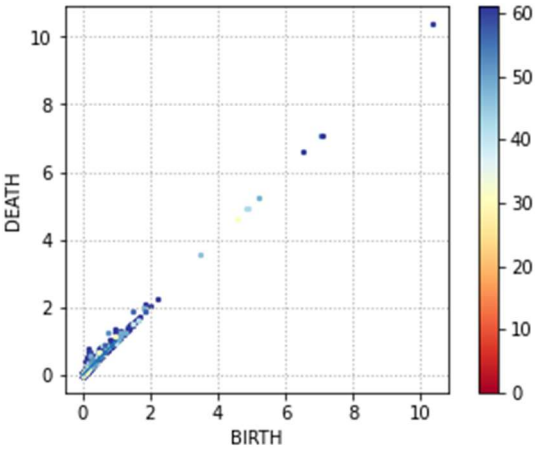
表 2：第 1 波～第 5 波の時期

	開始	終了
第 1 波	2020/4/1	2020/5/15
第 2 波	2020/7/10	2020/9/15
第 3 波	2020/12/16	2021/2/15
第 4 波	2021/4/1	2021/6/20
第 5 波	2021/7/16	2021/9/30

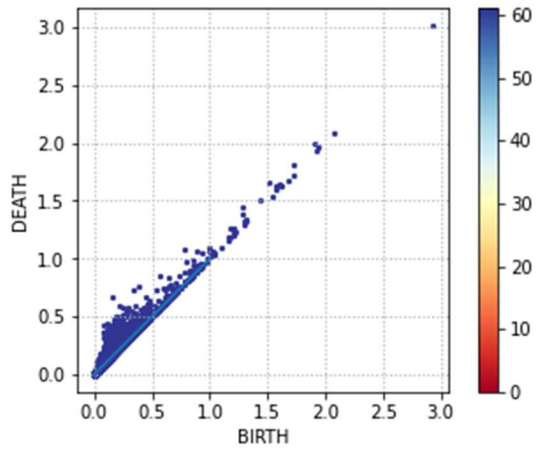
次の図 6 にそれぞれ第n 波の時期にあたるパーシステントダイアグラムを合成したものを示す。これらをパーシステントイメージとよび、$PI_1, PI_2, \ldots, PI_5$ とする。横軸は birth、縦軸は death、カラーバーは第 n 波内での日数を表している(赤から青になるにつれ日数が経過しており、0 は 1 日目である)。



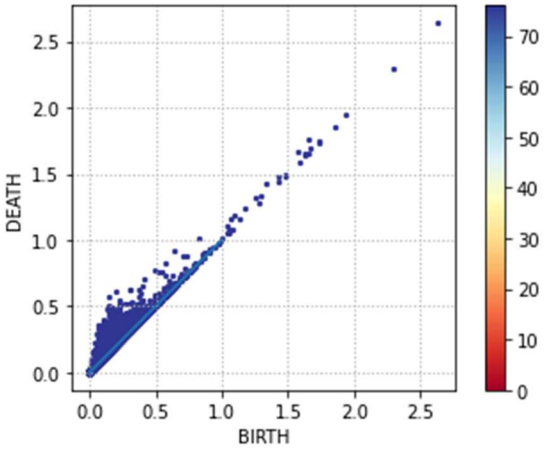
(a) 第 1 波



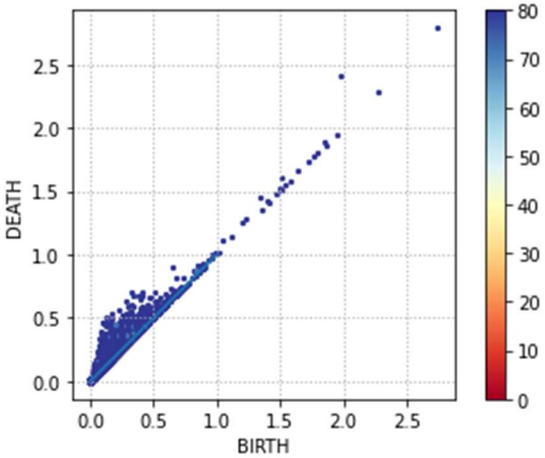
(b) 第 2 波



(c)第 3 波



(d)第 4 波



(e)第 5 波

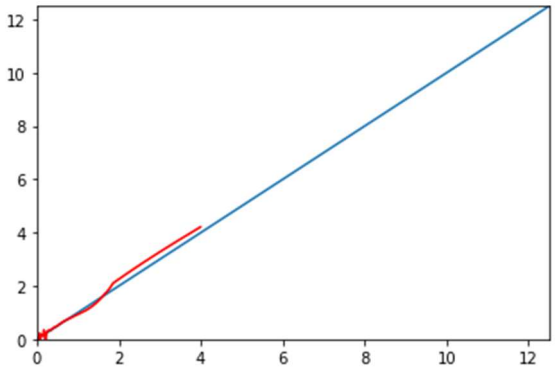
図 6：パーシステントイメージ

上記のパーシステントイメージでは、第 3 波以降については各 n 波内の後半に感染者数が特に多いため、実際には存在しているもの各 n 波内の前半部分の時刻対が確認出来ず、第 3 波、第 4 波、第 5 波については違いを確認することが難しい。そこで次の手順 3 を行った。

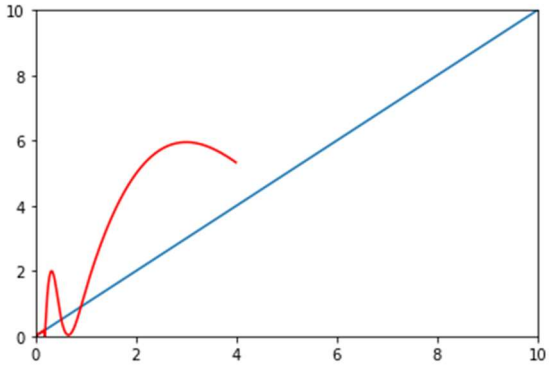
ここで、次の手順 3 では作成したパーシステントイメージに曲線を見出すのであるが、その契機となった論文(参考文献[7])を紹介する。この論文では材料工学の分野へ TDA を応用している。特にシリカガラスの分子構造に対し TDA を行い、得られたパーシステント図の中に特徴的な曲線を見出し、その曲線を解析することで構造分析を行っている。この論文の場合分子構造のパーシステントダイアグラムであるため特徴的な曲線を見出すのが比較的容易であったが、今回は感染者という実データに対して TDA を行うためパーシステント図が複雑であり特徴的な曲線を見出すことが難しい。そこで次のように考えた。

$PI_1, PI_2, \ldots, PI_5$ にそれぞれ存在する birth と death の時刻対の個数を要素数とする 3 次スプライン曲線を用い曲線を作成する。この曲線をパーシステント図から得られる曲線と考えることとし、これを生成元推移曲線をとよぶこととする。

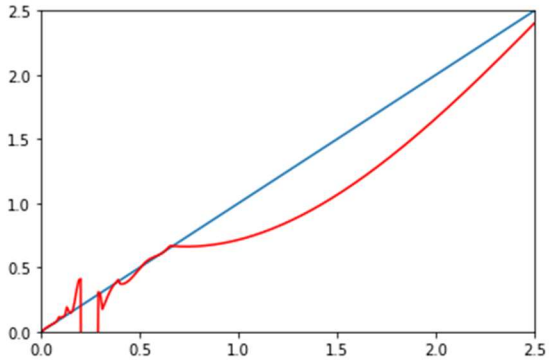
第 1 波から第 5 波それぞれのパーシステントイメージに対し作成した生成元推移曲線を次に示す。



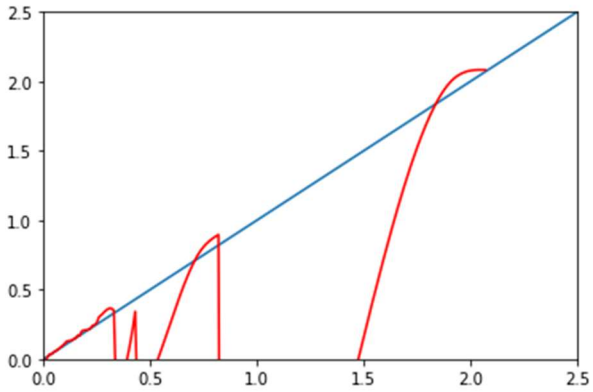
(a)第 1 波



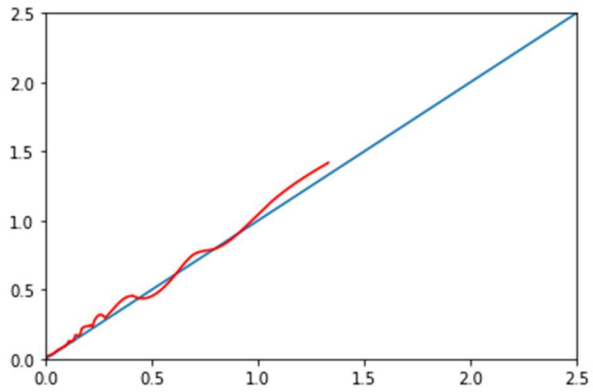
(b)第 2 波



(c)第 3 波



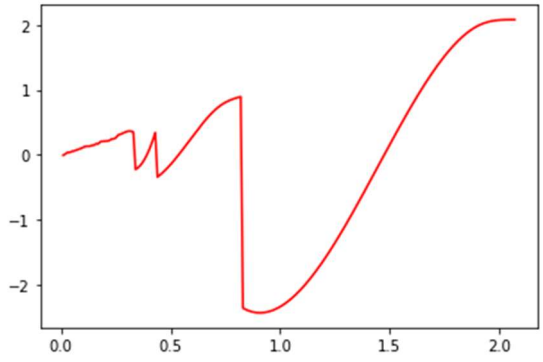
(d) 第 4 波



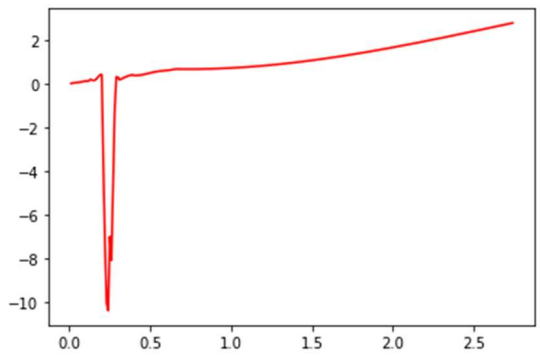
(e) 第 5 波

図 7：生成元推移曲線

第 3 波と第 4 波については軸の都合上見えていない部分があるため全体図を次に示す。



(a) 第 3 波



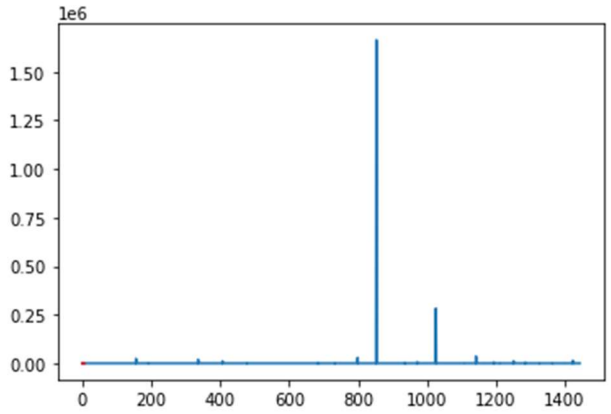
(b) 第 4 波

図 8：曲線の全体図

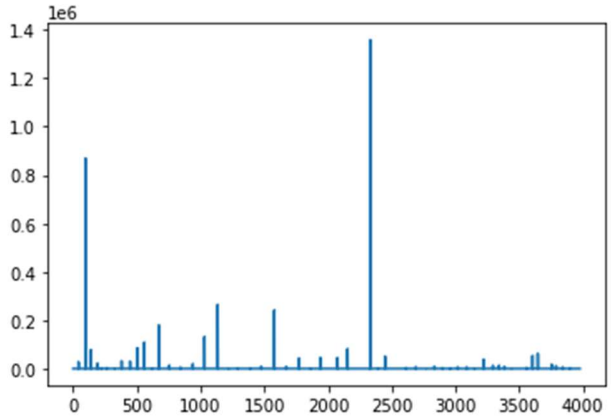
上記の曲線の違いを数量化するため、手順 4 では上記の生成元推移曲線に対し曲率とフーリエ変換を行う。結果については次に述べる。

5. 本研究の結果

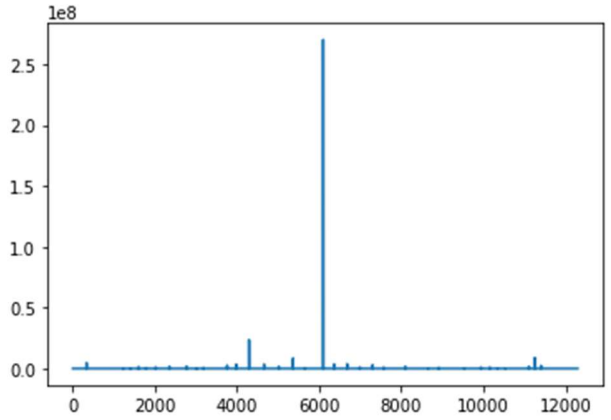
図 7 の各第 n 波の生成元推移曲線を曲率を用いて解析した結果を次の図 9 に示す。横軸は曲線の長さであり、縦軸はそのときの曲率である。また、縦軸上の 1e は 10 の累乗を表している。



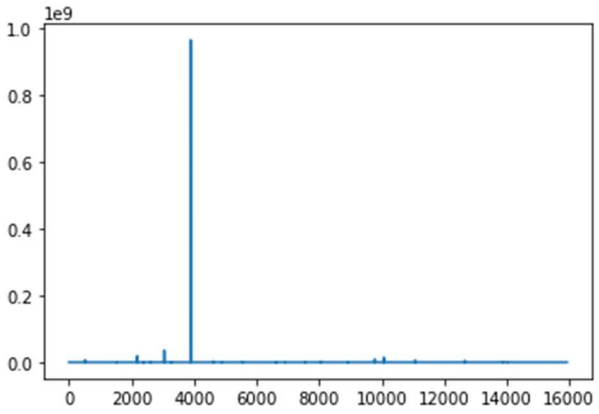
(a) 第 1 波



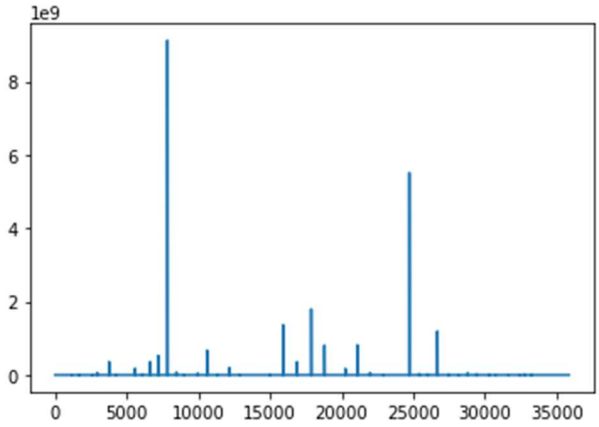
(b) 第 2 波



(c) 第 3 波



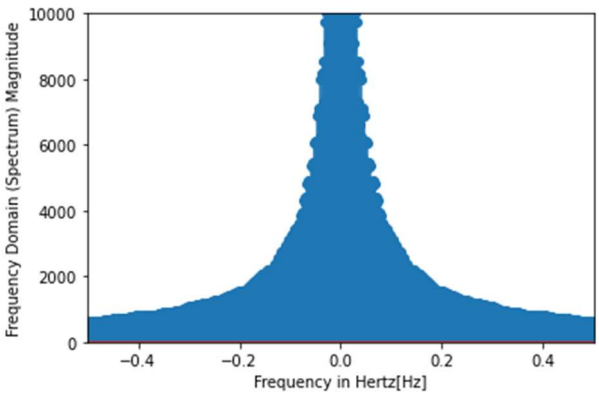
(d) 第 4 波



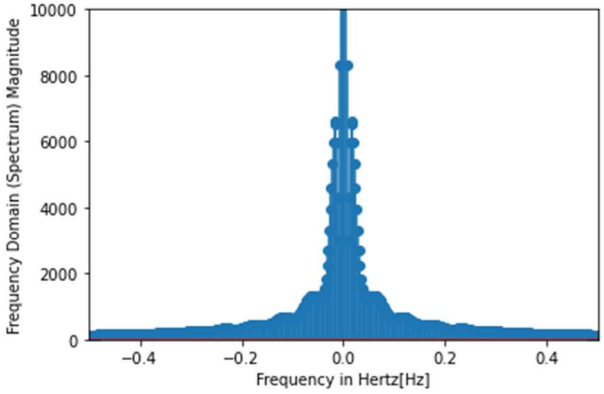
(e) 第 5 波

図 9：曲率

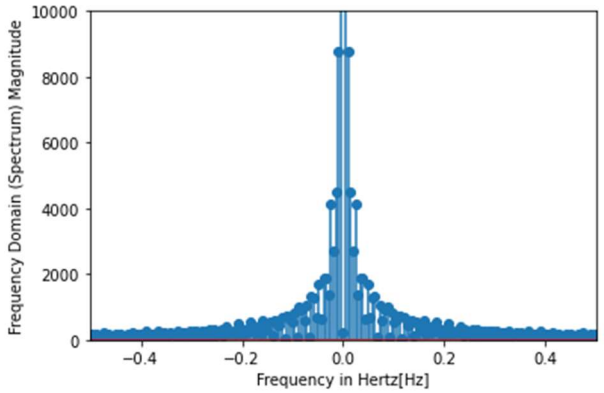
次に、各第 n 波の生成元推移曲線をフーリエ変換を用いて解析した結果を次の図 10 に示す。横軸は周波数、縦軸は振幅である。



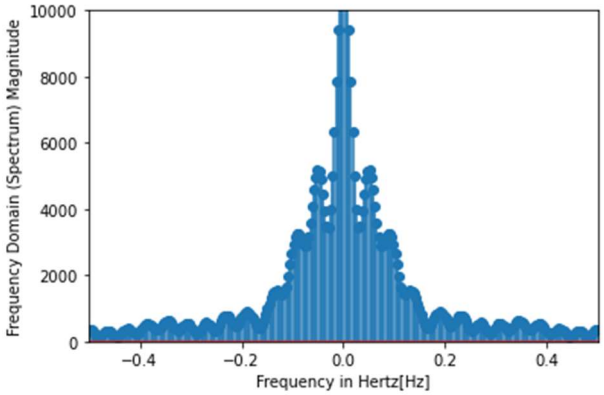
(a) 第 1 波



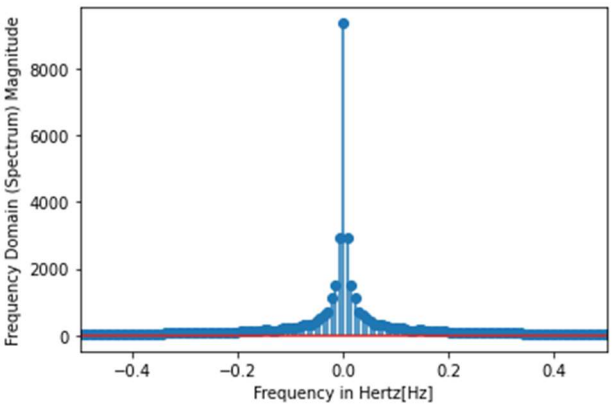
(b) 第 2 波



(c) 第 3 波



(d) 第 4 波



(e) 第 5 波

図 10：フーリエ変換

6. 考察

パーシステント図の結果について、第 1 波の birth、death がともに 10 と他の点と外れた位置にある点の集まりは 2020 年 4 月中旬の時期を示しているが、まだ新型コロナウイルスが流行する前段階であり、福岡県ですら感染者数が数人に過ぎず、首都圏と大阪に数十人ほどであったためパーシステントホモロジー群が形成されるのが遅かったからだと考えられる。前述の通り第 3 波以降については波の中盤以降に感染者数が多く、前半の感染者数が図からは確認できない状態となっている。

各第 n 波の累計の感染者数は第 1 波から 1444 人、3975 人、12295 人、15929 人、35866 人となっている。感染者数が増えると曲線は長くなり、曲率のピークは大きくなることがわかる。また、感染者数の少ない第 1 波、第 2 波においては曲線の後半部分にピークが出ていて、第 3 波以降感染者数が多くなるにつれ最も大きなピークは曲線の前半部分へと推移していることがわかる。また、累計の感染者数の似た第 3 波と第 4 波では第 4 波の方が 4 倍近く大きな値とり、ピークもより前半部分へ推移している。曲線の形が似ている第 1 波と第 5 波ではピークの回数や概形が似たものが出力されている。

フーリエ変換の結果より、ピークの値は第 1 波が 170 万ほどととびぬけて大きく、第 5 波が 1 万ほどで最も小さい。累計の感染者数が似ている第 3 波と第 4 波では第 4 波の方がピークの値は大きくなり、周波数 0 付近の振幅も第 4 波の方が大きくなっている。曲線の形が似ている第 1 波と第 5 波では第 5 波は周波数が 0 のとき以外の振幅は 0 に近い状態であるが、第 1 波は周波数が 0 のとき以外であっても振幅が見受けられた。またピークの値にも大きな差が生じた。第 2 波は第 1 波と比較すると周波数が 0 のとき以外の振幅が小さくなっている。

以上より、曲率においては感染者数が増えるほど曲線は長く曲率のピークの値は大きくなり、ピークは曲線の前半へ推移するものと考えられる。また、フーリエ変換においては感染者数が増えるほどピークの値が小さくなると考えられる。しかし、図 8 より第 3 波と第 4 波においては第 4 波の方が感染者数は多いもののフーリエ変換においては第 4 波の方がピークの値が大きくなっており、違った傾向が確認できたものもあった。また、第 1 波と第 5 波のおいては曲線の形や曲率のピークの回数などが似ていて判別が難しかったが、フーリエ変換の結果にて大きな違いが確認できた。

フーリエ変換において第 4 波の方がピークの値が大きくなった原因としては、他の曲線は birth=death である線に概ね沿うようになっているのに対し、第 3 波と第 4 波は前述の線から大きく外れており、中でも第 4 波は外れ具合が第 3 波よりも大きいので上述のような結果になったと考えられる。

また、第 2 波と第 5 波の曲率の結果に着目すると、他の波に比べピークが複数回あることが挙げられる。これは、繁華街を中心に広がった第 2 波、高齢者にワクチン接種が進んだ第 5 波は、帰省などで家庭内感染により広がったとみられる第 3 波、ワクチン接種が進む前に変異株であるアルファ株で爆発的に広がった第 4 波と比べて新規陽性者に占める高齢者の割合が比較的少なく、若者の割合が高かったことが原因と考えられる。

7. 参考文献

[1] VALUENEX ： 数学の時代到来の予感、パーシステントホモロジー -VALUENEX 技術トレンドレポート- , 2021

[2]Derek Lo and Briton Park ： Zika virus using TDA,preprint, 2018

[3]NIID 国立感染症研究所 IDWR 速報データ https://www.niid.go.jp/niid/ja/data.html

[4]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

[5]NHK 都道府県ごとの感染者数 https://www3.nhk.or.jp/news/special/coronavirus/data/

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

[7]Yasuaki Hiraoka,Takenobu Nakamura,Akihiko Hirata, Emerson G.Escolor,Kaname Matsue and Yasumasa Nishiura:Hierarchical structures of amorphous solids characterized by persistent homology Proc. Natl. Acid. Sci. USA , vol.113(2016),pp.7035-7040

[8]田中孝文 ： R による時系列分析入門,シーエーピー出版,2018

[9]川原晃祐 ： 位相的データ解析のベクトル化による感染症時系列データの分析 , 2021