Topological Data Analysis

Selected Examples

Advanced Mathematics for Scientific Challenges

28 November 2019

SCIENTIFIC REPORTS

OPEN

Lung Topology Characteristics in patients with Chronic Obstructive Pulmonary Disease

Francisco Belchi[1], Mariam Pirashvili[1], Joy Conway[2,3], Michael Bennett[3,4], Ratko Djukanovic[3,4] & Jacek Brodzki[1]

Quantitative features that can currently be obtained from medical imaging do not provide a complete picture of Chronic Obstructive Pulmonary Disease (COPD). In this paper, we introduce a novel analytical tool based on persistent homology that extracts quantitative features from chest CT scans to describe the geometric structure of the airways inside the lungs. We show that these new radiomic features stratify COPD patients in agreement with the GOLD guidelines for COPD and can distinguish between inspiratory and expiratory scans. These CT measurements are very different to those currently in use and we demonstrate that they convey significant medical information. The results of this study are a proof of concept that topological methods can enhance the standard methodology to create a finer classification of COPD and increase the possibilities of more personalized treatment.

Scientific Reports **8** (2018), 5341

Article 1

Background:

- **Chronic obstructive pulmonary disease** (COPD) is a progressive lung disease characterized by chronic inflammation of the bronchi and the lung parenchyma.
- High-resolution **computed tomography** (CT) scans are the most widely used form of imaging.

Objectives: To develop, by means of Topological Data Analysis, a set of new radiomic features that can distinguish between healthy non-smokers, healthy smokers, and patients with COPD.

Population: For 30 participants (8 healthy non-smokers, 9 healthy smokers, 8 mild COPD and 5 moderate COPD), both inspiratory and expiratory CT scans were obtained.

Article 1

Methodology: Persistent homology in degrees 0, 1 and 2 was used in different ways to obtain different kinds of clinical insight.

- In degree 0, it was used to define **upwards complexity.**
- Persistent homology in degree 1 was used to measure **branch-to-branch proximity.**
- The degree 2 was used to overcome the limitation of the low spatial resolution of CT scans by including information about the space between the airways and the outer boundary of the lobes.

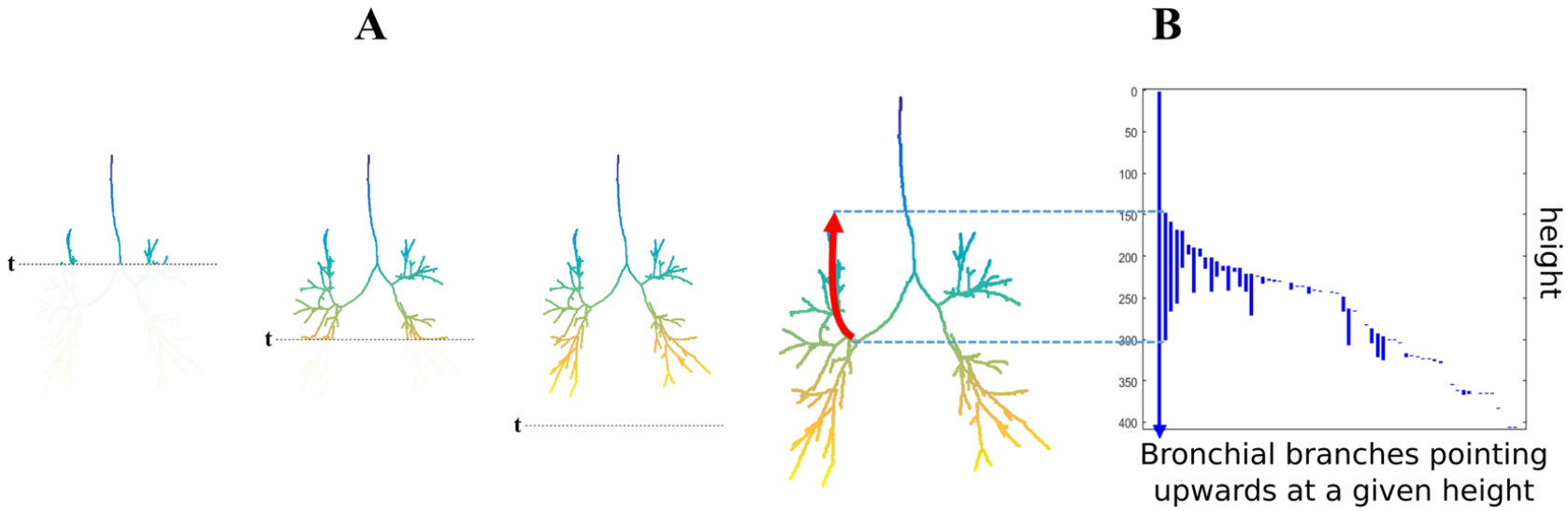


Figure 8. Explanation of upwards complexity. The color gradient indicates height. (**A**) To study upwards complexity, we slide a horizontal plane downwards. If we denote by X_h the part of the tree that sits above the horizontal plane at distance h from the top of the image, then $X_h \subseteq X_{h'}$ whenever $h \leq h'$, obtaining a sequence of nested graphs approximating the bronchial tree more accurately as we increase h. (**B**) The right part of the panel shows the degree-0 barcode of the sequence of nested graphs in (**A**). In this picture, the correspondence between bars in the barcode and branches that change trajectory upwards becomes apparent. In particular, the length of a bar indicates for how long a branch follows that upwards trajectory.

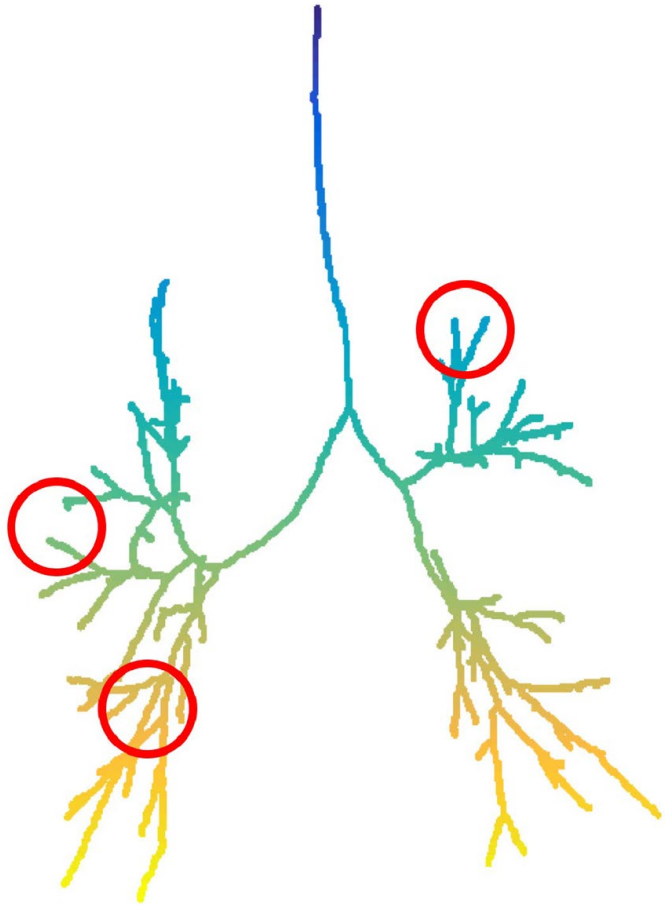


Figure 7. Calculations show that the lung function is better when more branches bend towards one another in the expiratory bronchial tree (such as the branches in the two circles on the left, in contrast with those in the circle on the right). See Fig. 3C.

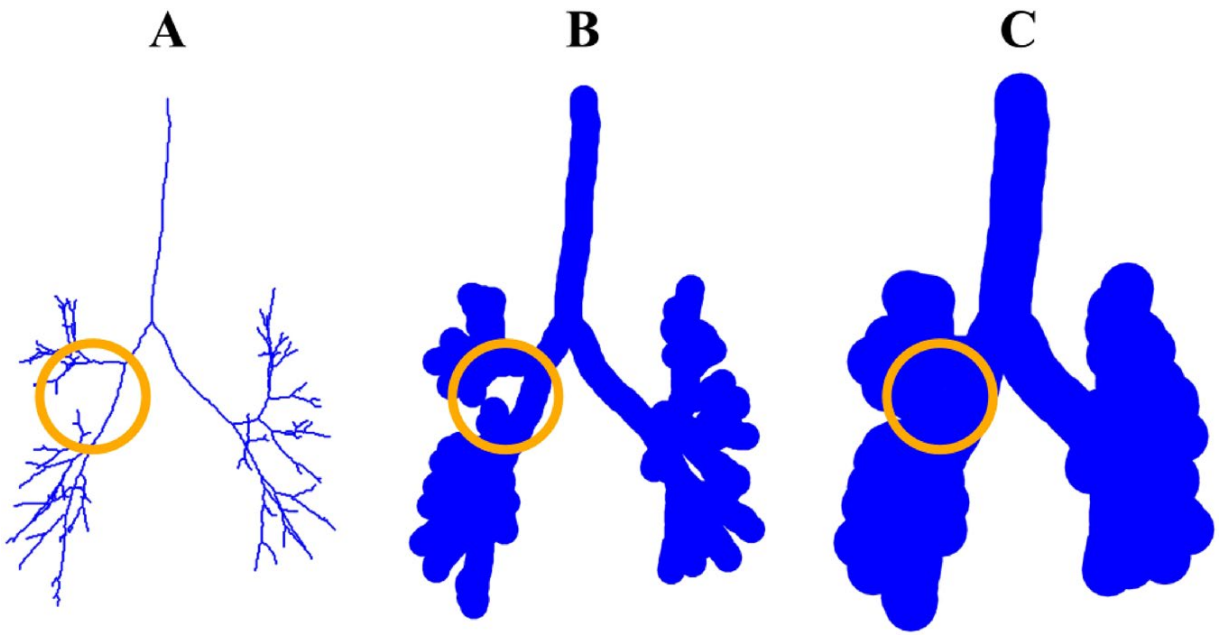


Figure 6. Computing of branch-to-branch proximity. Consider the graph representing the bronchial tree as explained in Methods (**A**). This graph is called a tree since it contains no loops, i.e., no branches that bifurcate and then merge. Of note, there are many nodes (up to 264) between any two consecutive bifurcations, so the nodes appear dense in the graph representation. Centered at each node of this graph, we virtually set a ball of a fixed radius, thickening the construction. As we keep thickening more and more, by increasing the radius of those balls, at some point we will find that some branches merge, creating a loop (**B**). We record the radius r_1 at which this happens. For a large enough radius r_2, though, this loop will be filled in (**C**). If a merging of branches creates a loop that appears for the value r_1 of the radius and disappears at r_2, we represent this merging as the positive number $r_2 - r_1$. Summing up all these terms, we obtain a number we call branch-to-branch proximity.

Article 1

Software:

- To generate barcodes, the package **TDATools** was used:

  https://github.com/ksian/ML2015FP/tree/master/3TDATools

- The resulting barcodes were compared using the **Wasserstein and bottleneck distances.**

- **Alpha complex** filtrations and their barcodes were computed using the **GUDHI** library.

Article 1

Conclusions:

- Topological methods can enhance the standard methodology to create a finer classification of COPD and increase the possibilities of more personalized treatment.
- The authors propose that the relation between lung diseases and the shape of the bronchial tree, including properties such as trajectory changes, are of value to advancing our understanding of the mechanisms of COPD.

Article 2

Physica A 491 (2018) 820–834

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Topological data analysis of financial time series: Landscapes of crashes

Marian Gidea [a,b,*], Yuri Katz [c]

[a] School of Civil Engineering and Architecture, Xiamen University of Technology, Fujian, China
[b] Department of Mathematical Sciences, Yeshiva University, New York, NY 10016, USA
[c] S&P Global Market Intelligence, 55 Water Str., New York, NY 10040, USA

HIGHLIGHTS

- We introduce a new method, based on topological data analysis (TDA), to analyze financial time series, and detect possible early signs prior to financial crashes.
- We analyze the time-series of daily log-returns of four major US stock market indices: S&P 500, DJIA, NASDAQ, and Russell 2000.
- We use persistence homology to detect and quantify topological patterns that appear in the multidimensional time series.
- We find that, in the vicinity of financial meltdowns, the L^p-norms of persistence landscapes exhibit strong growth prior to the primary peak, which ascends during a crash.
- Our method is very general and can be applied to any asset-types and mixtures of time series.

Article 2

ABSTRACT

We explore the evolution of daily returns of four major US stock market indices during the technology crash of 2000, and the financial crisis of 2007–2009. Our methodology is based on topological data analysis (TDA). We use persistence homology to detect and quantify topological patterns that appear in multidimensional time series. Using a sliding window, we extract time-dependent point cloud data sets, to which we associate a topological space. We detect transient loops that appear in this space, and we measure their persistence. This is encoded in real-valued functions referred to as a 'persistence landscapes'. We quantify the temporal changes in persistence landscapes via their L^p-norms. We test this procedure on multidimensional time series generated by various non-linear and non-equilibrium models. We find that, in the vicinity of financial meltdowns, the L^p-norms exhibit strong growth prior to the primary peak, which ascends during a crash. Remarkably, the average spectral density at low frequencies of the time series of L^p-norms of the persistence landscapes demonstrates a strong rising trend for 250 trading days prior to either dotcom crash on 03/10/2000, or to the Lehman bankruptcy on 09/15/2008. Our study suggests that TDA provides a new type of econometric analysis, which complements the standard statistical measures. The method can be used to detect early warning signals of imminent market crashes. We believe that this approach can be used beyond the analysis of financial time series presented here.

Article 2

Objectives: To quantify topological patterns that appear in **financial time series** using persistent homology, and detect possible early signs prior to financial crashes.

Population: Daily log-returns of four major US stock-market indices (S&P 500, DJIA, NASDAQ, and Russell 2000) during the technology crash of 2000 and the financial crisis of 2007–2009.

Methodology: Persistence of loops (1D persistence homology) was measured in a 4D point cloud formed by a single sliding window and four 1D time series of daily log-returns of the four stock-market indices. This was encoded in real-valued functions referred to as **persistence landscapes.** Temporal changes in persistence landscapes were quantified **via their L^p-norms.**

Article 2

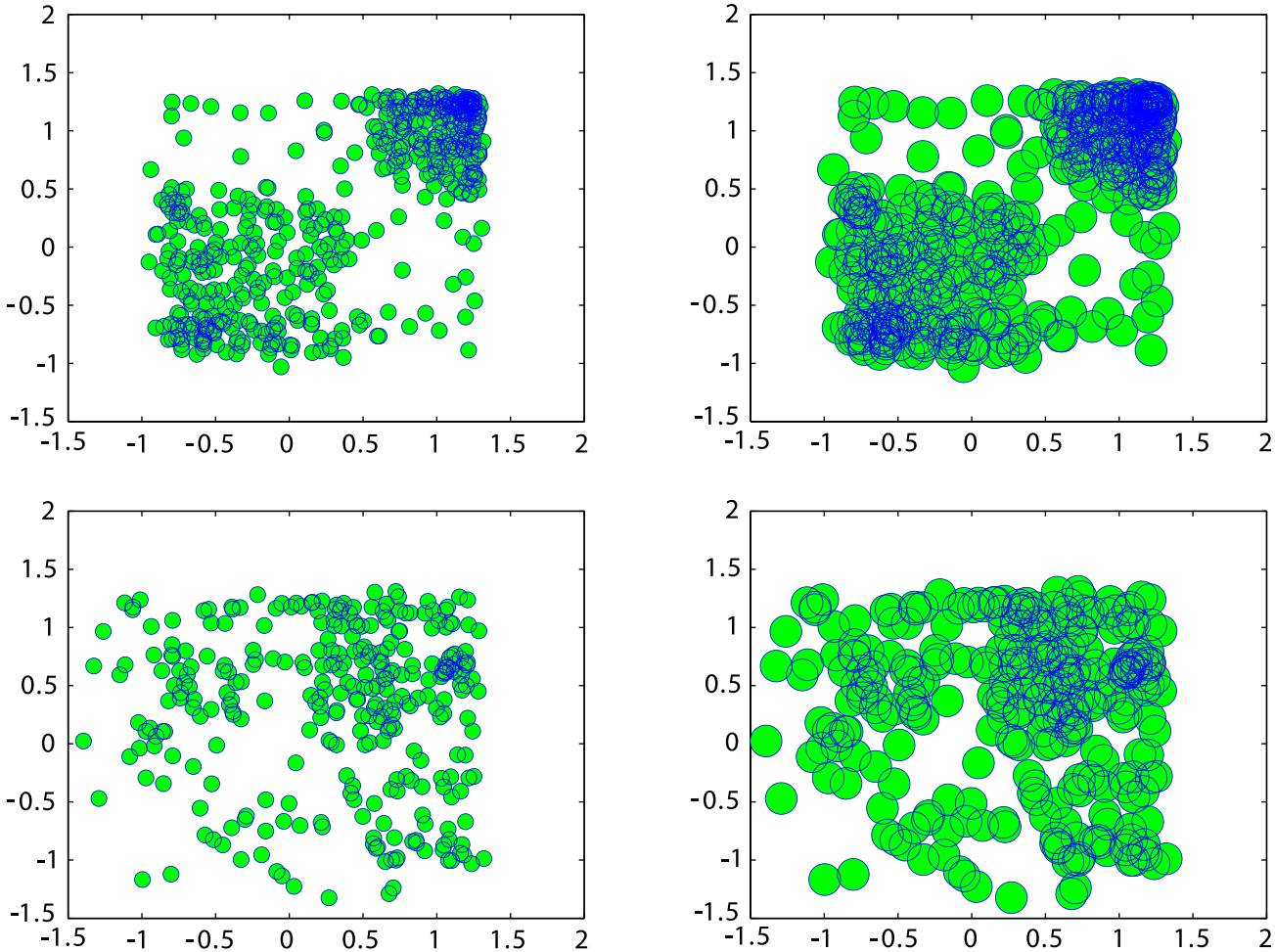


Fig. 3. Two-dimensional projections of point clouds corresponding to various instants of time; as time progresses, the value of the parameter *a* increases (from top to bottom). Around each point, we draw a blue circle of a small radius (left column), and of a bigger radius (right column), to illustrate the birth and death of various loops. Color online.

Article 2

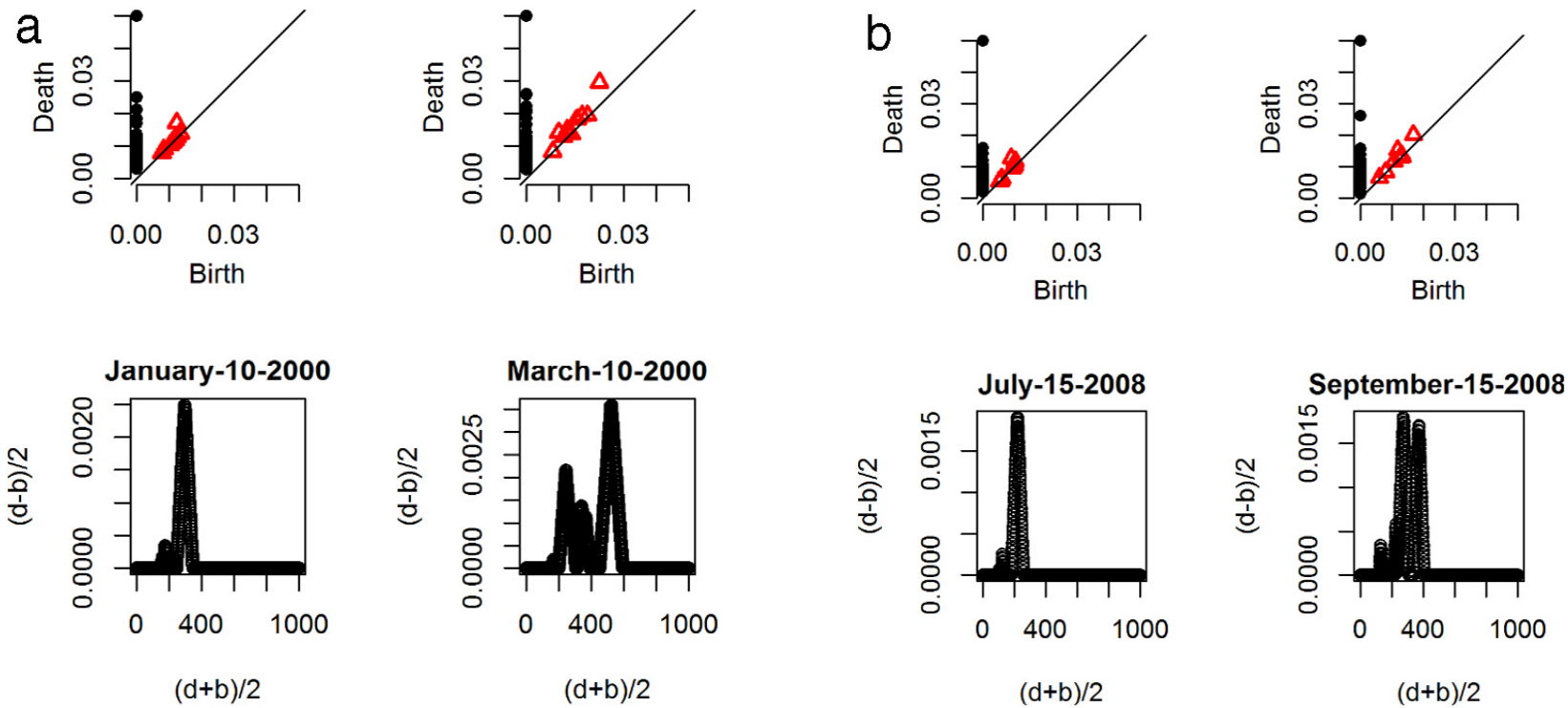


Fig. 8. The Rips persistence diagrams and the corresponding persistence landscapes calculated with the sliding window of 50 trading days on selected dates. The solid black dots represent connected components, red triangles represent loops. (a) Technology crash, 2000; (b) Financial Crisis, 2008. Color online.

Article 2

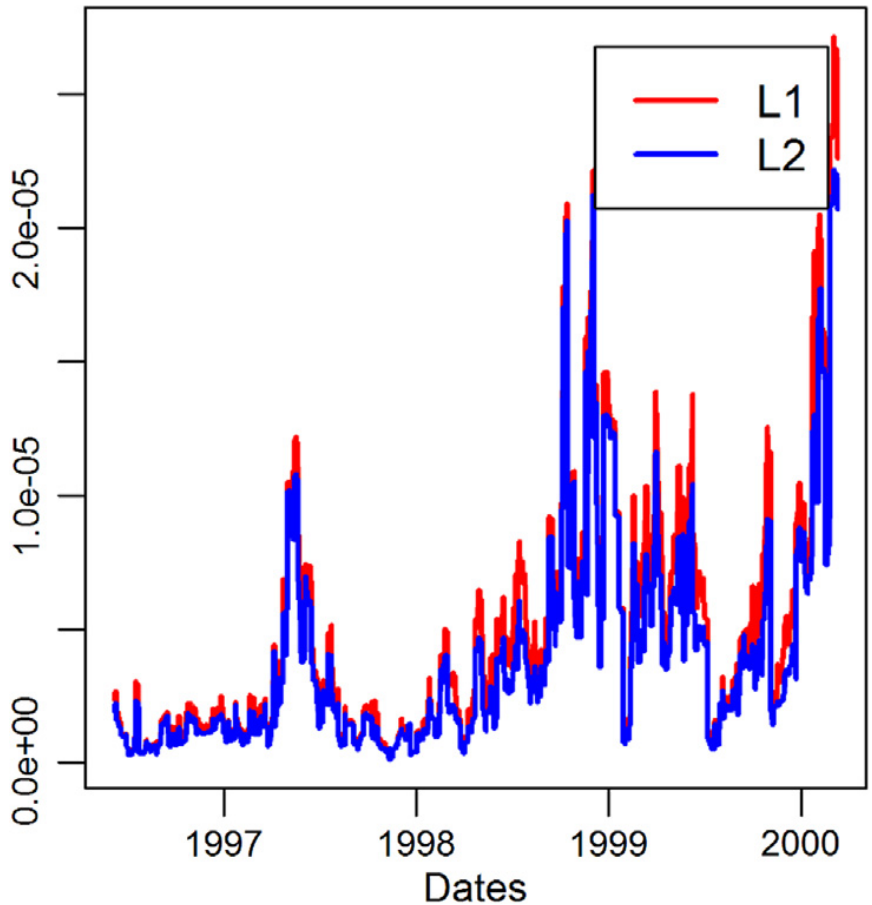


Fig. 9. The time series of normalized L^1 (blue line) and L^2 (red line) norms of persistence landscapes calculated with the sliding window of 50 days.

Article 2

Results:

- In the vicinity of financial meltdowns, the L^p-norms exhibited strong growth prior to the primary peak, which ascends during a crash such as the dotcom crash on 03/10/2000, or the Lehman bankruptcy on 09/15/2008.
- This behavior reflects an increased persistence of loops appearing in point clouds as the market undergoes transition from the ordinary to the “heated” state.

Conclusions: The study suggests that TDA provides a new type of econometric analysis, which complements the standard statistical measures. The method can be used to detect early warning signals of imminent market crashes.

Software: The **R-package TDA** was used for analyses.

Article 3

PERSISTENT HOMOLOGY MACHINE LEARNING FOR FINGERPRINT CLASSIFICATION

NOAH GIANSIRACUSA, ROBERT GIANSIRACUSA, AND CHUL MOON

ABSTRACT. The fingerprint classification problem is to sort fingerprints into pre-determined groups, such as arch, loop, and whorl. It was asserted in the literature that minutiae points, which are commonly used for fingerprint matching, are not useful for classification. We show that, to the contrary, near state-of-the-art classification accuracy rates can be achieved when applying topological data analysis (TDA) to 3-dimensional point clouds of oriented minutiae points. We also apply TDA to fingerprint ink-roll images, which yields a lower accuracy rate but still shows promise, particularly since the only preprocessing is cropping; moreover, combining the two approaches outperforms each one individually. These methods use supervised learning applied to persistent homology and allow us to explore feature selection on barcodes, an important topic at the interface between TDA and machine learning. We test our classification algorithms on the NIST fingerprint database SD-27.

arXiv:1711.09158v1 (November 2017)

Article 3

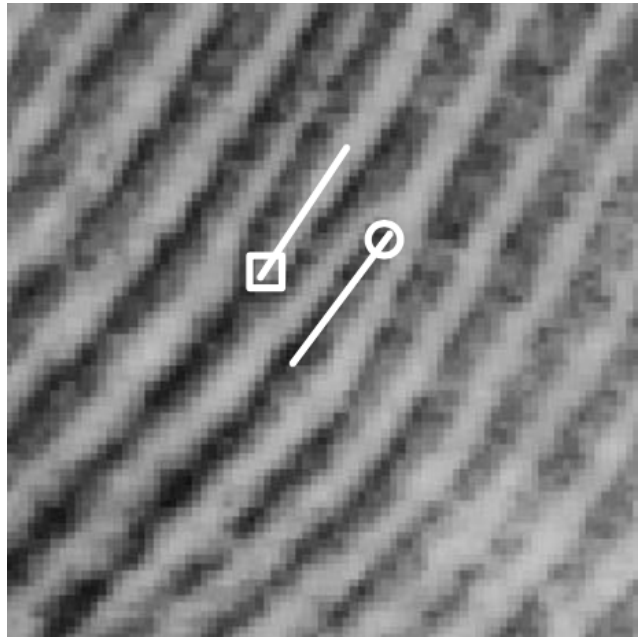


FIGURE 2. Illustration from the NIST SD-27 documentation of the orientation of the two types of minutiae points: bifurcation (square) and termination (circle).

Article 3

Background: Near the end of the 19th century, Sir Francis Galton introduced a systematic framework for **fingerprint analysis.** One component of his work was to divide all fingerprints into three classes: **arch, loop, and whorl.** This classification, often with refinements (such as subdividing arches into plain and tented types, and dividing whorls into singles and doubles) is still used by nearly every fingerprint classification scheme today.

Objectives: It was asserted in the literature that **minutiae points,** which are commonly used for fingerprint matching, are not useful for classification. The aim of the study is to prove that, on the contrary, standard classification accuracy rates can be achieved when applying Topological Data Analysis to 3-dimensional point clouds of oriented minutiae points.

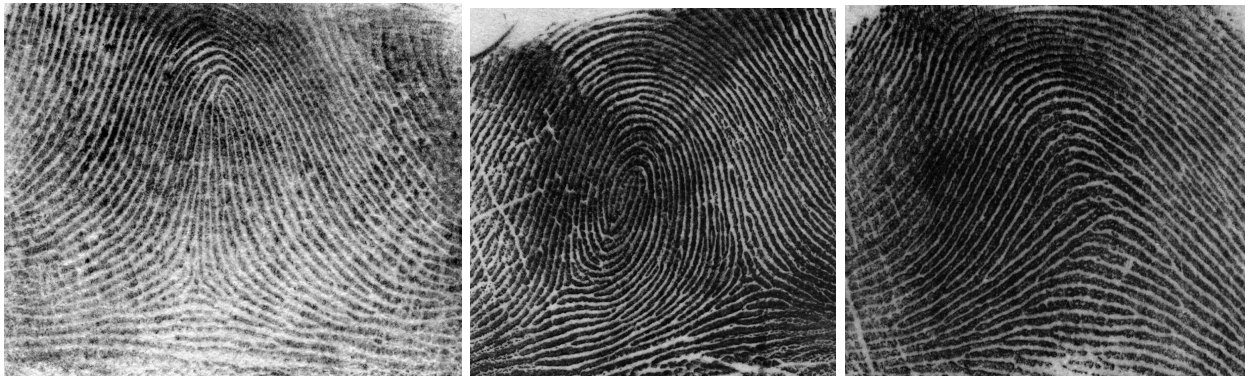
Article 3

Methodology:

- The authors develop an algorithm based on **machine learning procedures** for fingerprint classification and persistent homology.
- The algorithm is tested on the NIST database SD-27, which includes 245 fingerprints for which human experts have identified the minutiae point locations and orientations and determined the fingerprint class.
- The key insight is that persistent homology allows the point cloud to live in any metric space, not just Euclidean space. Minutiae point orientations are simply angles, so they are naturally viewed as points on the unit circle S^1. The minutiae points on a given fingerprint then form **a point cloud in the cylinder $\mathbb{R}^2 \times S^1$.**

Article 3

(a)



(b)

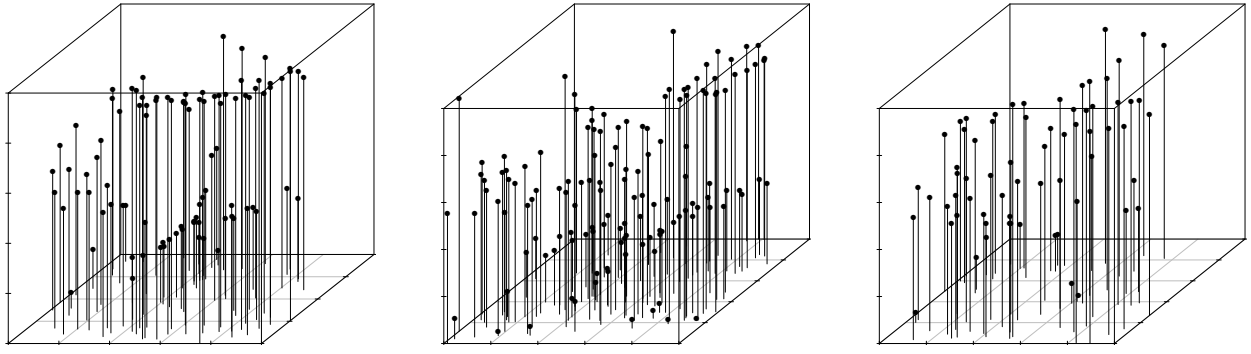


FIGURE 4. (a) Cropped images of a loop (left), whorl (middle), and arch (right). (b) Scatterplots of the corresponding normalized minutiae coordinates (the vertical axis is the orientation, so the top and bottom squares squares should be identified). (c) The 0-dimensional (gray) and 1-dimensional (black) barcodes for the unoriented minutiae point clouds in $\mathbb{R}^2$. (d) The barcodes for the minutiae point clouds in $\mathbb{R}^2 \times S^1$ with the metric d_2 defined earlier. Precisely interpreting these barcodes is not necessary; for the purposes of supervised learning we simply need that the barcodes reflect *some* relevant global geometric structure in the fingerprints.

Article 3

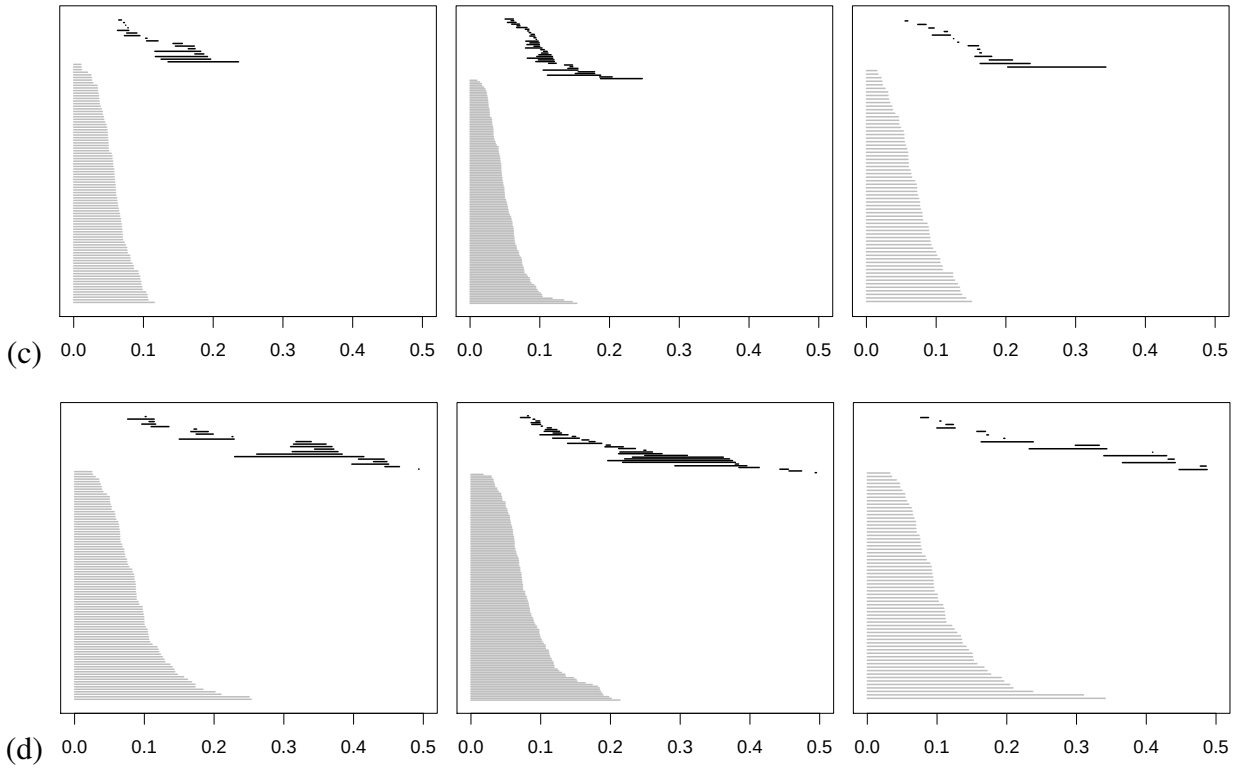


FIGURE 4. (a) Cropped images of a loop (left), whorl (middle), and arch (right). (b) Scatterplots of the corresponding normalized minutiae coordinates (the vertical axis is the orientation, so the top and bottom squares squares should be identified). (c) The 0-dimensional (gray) and 1-dimensional (black) barcodes for the unoriented minutiae point clouds in $\mathbb{R}^2$. (d) The barcodes for the minutiae point clouds in $\mathbb{R}^2 \times S^1$ with the metric d_2 defined earlier. Precisely interpreting these barcodes is not necessary; for the purposes of supervised learning we simply need that the barcodes reflect *some* relevant global geometric structure in the fingerprints.

Article 3

Conclusions:

- The assertion that minutiae points are not useful for classification is unfounded and untrue. Minutiae-based persistent homology appears to perform squarely within the range of accuracies demonstrated by other published fingerprint classification methods.
- While much remains to be understood regarding the interface between persistent homology and machine learning, the authors hope that this study helps provide some insight into feature selection on barcodes.

Article 4

PERSISTENT HOMOLOGY OF GEOSPATIAL DATA: A CASE STUDY WITH VOTING

MICHELLE FENG* AND MASON A. PORTER†

Abstract. A crucial step in the analysis of persistent homology is the transformation of data into an appropriate topological object (in our case, a simplicial complex). Modern packages for persistent homology often construct Vietoris–Rips or other distance-based simplicial complexes on point clouds because they are relatively easy to compute. We investigate alternative methods of constructing these complexes and the effects of making associated choices during simplicial-complex construction on the output of persistent-homology algorithms. We present two new methods for constructing simplicial complexes from two-dimensional geospatial data (such as maps). We apply these methods to a California precinct-level voting data set, demonstrating that our new constructions can capture geometric characteristics that are missed by distance-based constructions. Our new constructions can thus yield more interpretable persistence modules and barcodes for geospatial data. In particular, they are able to distinguish short-persistence features that occur only for a narrow range of distance scales (e.g., voting behaviors in densely populated cities) from short-persistence noise by incorporating information about other spatial relationships between precincts.

arXiv:1902.05911v2 (September 2019)

Article 4

Objectives: To develop new methods for constructing simplicial complexes from two-dimensional **geospatial data,** in such a way that the new constructions can capture geometric characteristics that **are missed by distance-based constructions.**

Population:

- Data compiled by the Los Angeles Times Data Visualization Team after the November 2016 elections were used. The data cover all of California's 24,626 precincts, which are organized into 58 counties.
- Red colour is used to indicate a voting preference for Donald Trump, and blue colour is used to indicate a preference for Hillary Clinton, with darker colors signifying a stronger voting preference in a particular precinct.

Article 4

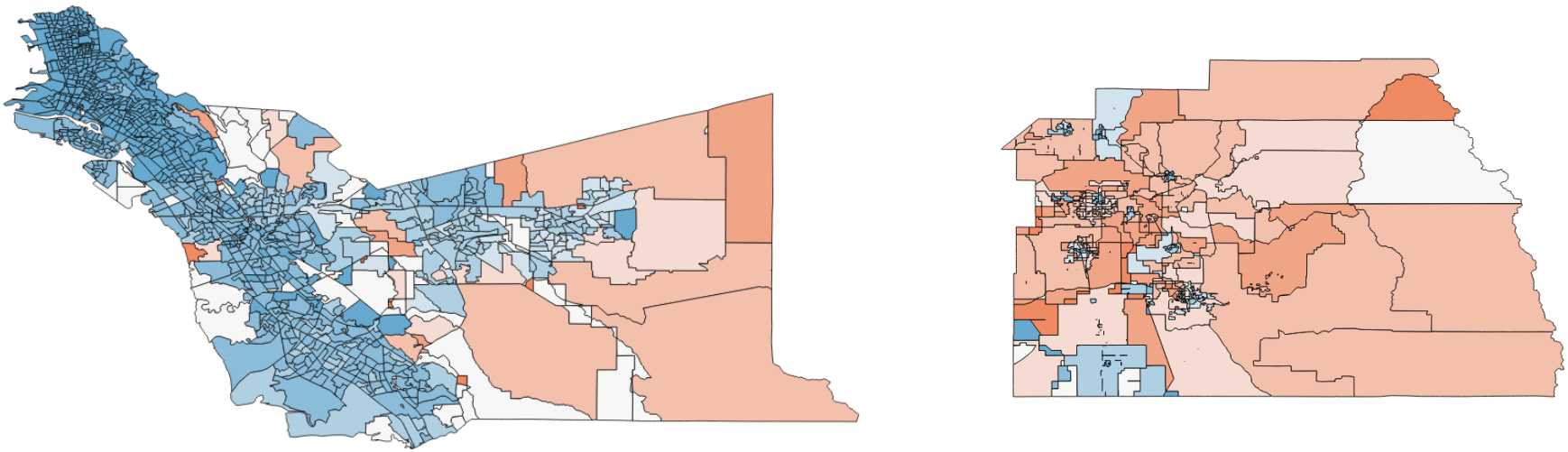


Fig. 1: The counties of (left) Alameda and (right) Tulare. Red precincts voted predominantly for Donald Trump, and blue ones voted predominantly for Hillary Clinton. Darker shading in a precinct indicates a stronger majority for the winning candidate, so Trump won dark-red precincts by a large margin and Clinton won dark-blue precincts by a large margin. We use the color white for precincts with a strictly equal number of votes for the two candidates.

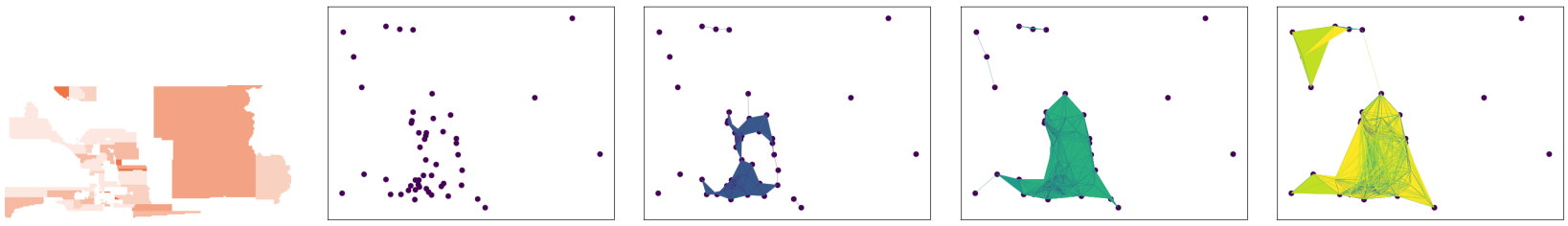
Article 4

Methodology:

- For two-dimensional geospatial data, holes are interpreted as geographical features such as lakes or deserts.
- The large number of precincts in several counties makes it intractable to compute Vietoris–Rips (VR) complexes for these counties. For county-candidate combinations with at least 151 precincts, **alpha complexes** were computed instead.

Software: For the construction of VR and alpha complexes, the Python package **GUDHI** was used. For the computation of persistent homology and its generators, a modified version of **PHAT** was used. The **adjacency and level-set constructions** were implemented using an adaptation of the fast incremental VR algorithm.

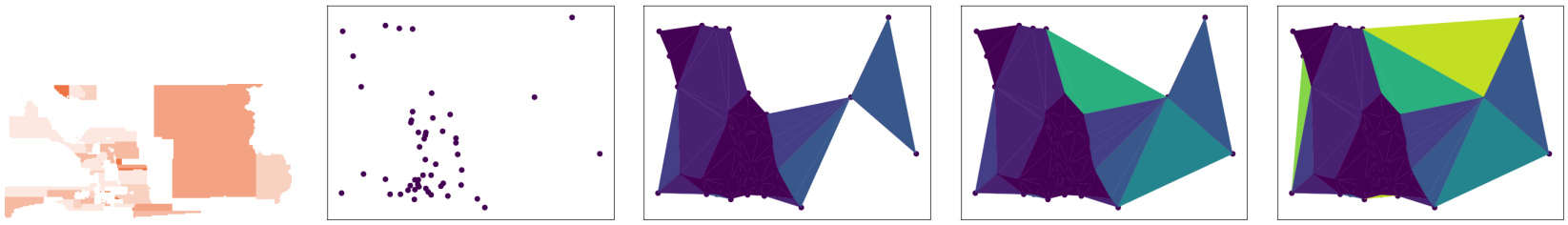
Article 4



(a) (b) (c) (d) (e)

Fig. 2: Illustration of a Vietoris–Rips complex on the *LA Times* voting data. (a) The red precincts (in which more people voted for Donald Trump than for Hillary Clinton) of Imperial County in 2016. In panels (b)–(e), we show the VR complex that approximates the county, with each successive image showing the VR complex as we increase ϵ. Observe that the contiguous region in the east of the county is not captured by this complex and that the western region includes a large number of 1-simplices and 2-simplices, despite the fact that there are relatively few precincts on the map. Both phenomena occur because the eastern precincts are much larger, so their centroids are much farther apart than the small (but not necessarily contiguous) precincts in the west.

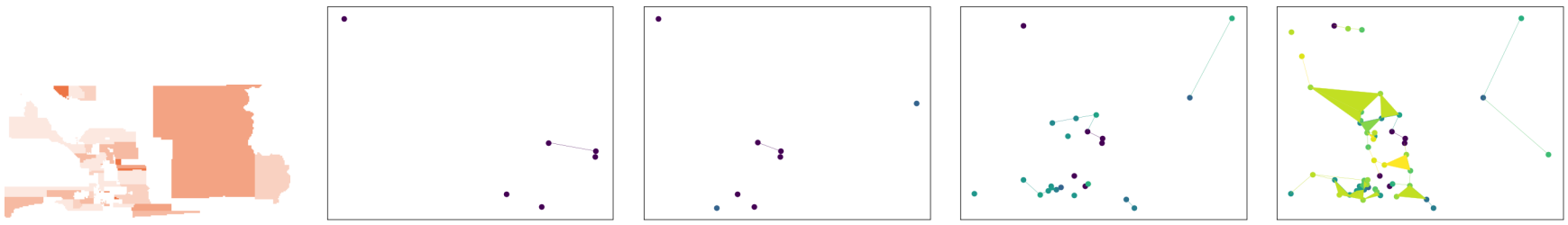
Article 4



(a) (b) (c) (d) (e)

Fig. 3: Illustration of an alpha complex on the *LA Times* voting data. (a) The red precincts (in which more people voted for Donald Trump than for Hillary Clinton) of Imperial County in 2016. In panels (b)–(e), we show the alpha complex that approximates the county, with each successive image showing the complex as we increase ϵ. Observe that the filtered simplicial complex has much larger 2-simplices than what we obtained for VR complexes (see Figure 2), and that (unlike in Figure 2) once the western region is covered by 2-simplices (which, as one can see in panel (c), occurs fairly early in the filtration), new 2-simplices do not arise as we increase ϵ. However, similar to what we observed in the VR complex, the resulting simplicial complex yields a simply-connected region in the west; this does not accurately reflect the underlying geographical map.

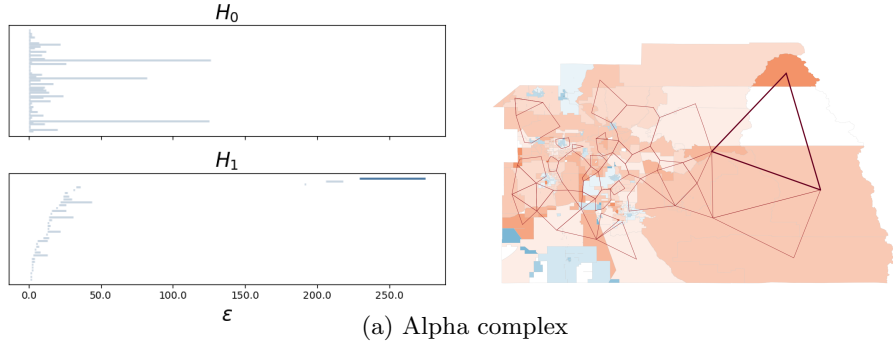
Article 4



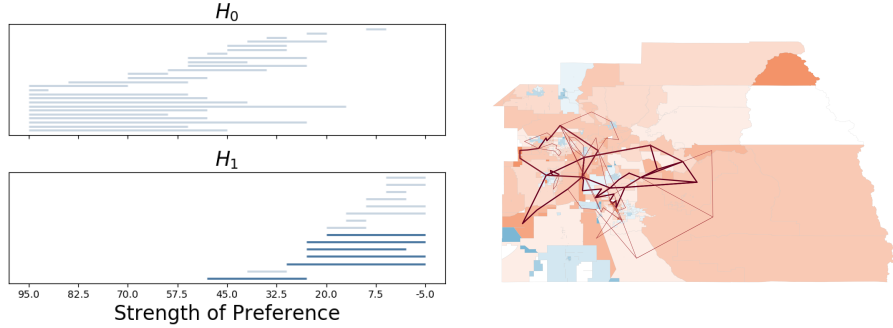
(a) (b) (c) (d) (e)

Fig. 4: Illustration of an adjacency complex on the *LA Times* voting data. (a) The red precincts (in which more people voted for Donald Trump than for Hillary Clinton) of Imperial County in 2016. In panels (b)–(e), we show an associated adjacency complex that approximates the county, where we order the panels based on decreasing strength of preference for Trump. In panel (e), we observe that the eastern region is simply connected and that the western region has many 1-simplices. However, the latter is not covered by 2-simplices, so it is not simply connected. Although the depicted filtered simplicial complex does not seem to closely resemble the geographical map in Figure 4a visually, its topological properties do appear to be similar.

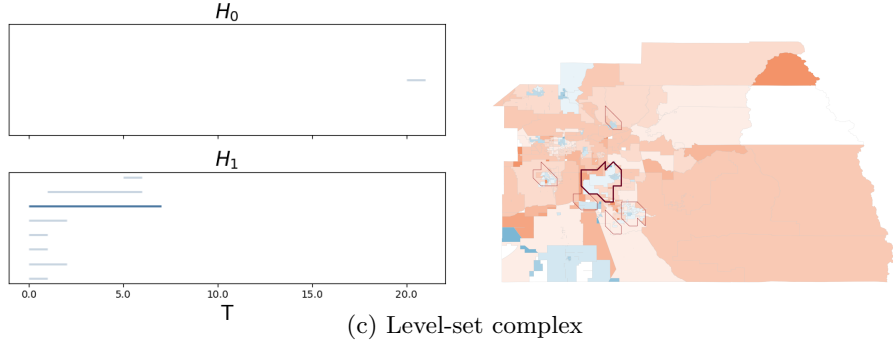
Article 4



(a) Alpha complex



(b) Adjacency complex



(c) Level-set complex

Fig. 10: Barcodes and generated loops for red precincts in Tulare County. We mark long-persistence features using darker loops with thicker line widths.

Article 4

Conclusions:

- A particularly difficult aspect of geospatial data is that barcodes of a similar length may represent either signal or noise, in stark contrast to the conventional wisdom that the features that persist the longest also carry the most meaningful information about a data set.
- In this article, two new methods were introduced for constructing a filtered simplicial complex encoding information about **contiguity** on maps: **adjacency complexes and level-set complexes.**
- Barcodes for adjacency and level-set complexes are more interpretable than those from traditional persistent homology constructions for geospatial data, allowing us to better understand the topology of voting patterns in counties.

Break Time!