SIAM J. MATH. DATA SCI.
Vol. 4, No. 3, pp. 1116–1144

Analysis of Spatial and Spatiotemporal Anomalies Using Persistent Homology: Case Studies with COVID-19 Data*

Abigail Hickok†, Deanna Needell†, and Mason A. Porter‡

Abstract. We develop a method for analyzing spatial and spatiotemporal anomalies in geospatial data using topological data analysis (TDA). To do this, we use persistent homology (PH), which allows one to algorithmically detect geometric voids in a data set and quantify the persistence of such voids. We construct an efficient filtered simplicial complex (FSC) such that the voids in our FSC are in one-to-one correspondence with the anomalies. Our approach goes beyond simply identifying anomalies; it also encodes information about the relationships between anomalies. We use vineyards, which one can interpret as time-varying persistence diagrams (which are an approach for visualizing PH), to track how the locations of the anomalies change with time. We conduct two case studies using spatially heterogeneous COVID-19 data. First, we examine vaccination rates in New York City by zip code at a single point in time. Second, we study a year-long data set of COVID-19 case rates in neighborhoods of the city of Los Angeles.

Key words. topological data analysis, persistent homology, spatiotemporal data, COVID-19, spatial data

MSC codes. 55N31, 68T09, 92D30

DOI. 10.1137/21M1435033

1. Introduction.

Many systems are spatial in nature. When working with spatial data sets, it is important to study the role of underlying spatial relationships [10]. To illustrate this importance, consider the spatiotemporal dynamics of Coronavirus disease 2019 (COVID-19) case rates, which is one of the key motivations for our work. The spatial adjacencies between the neighborhoods of a city affect the dynamics of disease spread [36], and it is important to account for them. Researchers have studied a wide variety of spatial data sets, such as gross domestic product and life expectancy by country [2, 46] and voting in elections across different regions of a state [19]. Such data sets often also include temporal information (e.g., daily COVID-19 case rates), and it is also important to take it into account.

We develop new methods for using *topological data analysis* (TDA) to analyze geospatial and geospatiotemporal data sets in a way that directly incorporates spatial information. TDA is a way to study the "shape" of a data set [6]. Using *persistent homology* (PH), which is

*Received by the editors July 20, 2021; accepted for publication (in revised form) April 7, 2022; published electronically August 22, 2022.

https://doi.org/10.1137/21M1435033

Funding: The second author received support from the National Science Foundation (grant DMS-2011140). The third author received support from the National Science Foundation (grant DMS-2027438) through the RAPID program. The first and third authors received support from the National Science Foundation (grant 1922952) through the Algorithms for Threat Detection (ATD) program.

†Department of Mathematics, University of California, Los Angeles, CA 90095 USA (ahickok@math.ucla.edu, deanna@math.ucla.edu).

‡Department of Mathematics, University of California, Los Angeles, CA 90095 USA and Santa Fe Institute, Santa Fe, NM 87501 USA (mason@math.ucla.edu).

an approach from algebraic topology, one can algorithmically find geometric voids of different dimensions in a data set and quantify the "persistence" of these voids [35]. Zero-dimensional (0D) voids are connected components and one-dimensional (1D) voids are holes. To quantify the persistence of holes and other voids, one constructs a *simplicial complex* (which is a combinatorial description of a topological space) and a *filtration function* (see section 2.1). In our work, we treat geographical data as two-dimensional (2D) data and construct a 2D filtered simplicial complex (FSC) to represent it. The computation of PH has yielded insights into a wide variety of areas, such as dynamical systems [28, 53], collective behavior [48], neuroscience [20, 40], materials science [5], and chemistry [29]. Spatial applications that have been examined as 2D data sets using PH include sensor networks [14], percolation [42], and city-street networks and other complex systems [18].

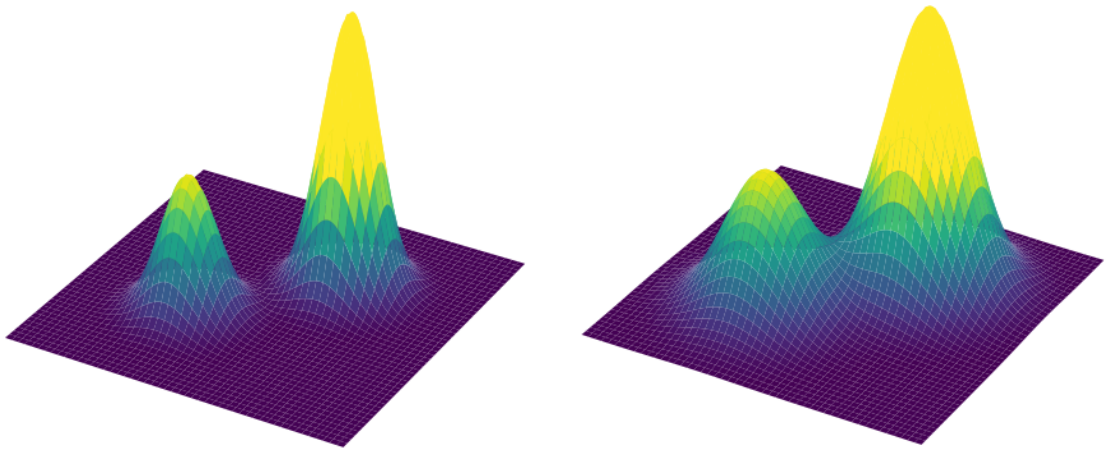
When we examine time-dependent data, we use *vineyards*, which were introduced in [12] as a way to represent time-varying PH, to incorporate temporal information. One can visualize a vineyard as a continuous stack of persistence diagrams (PDs), with one PD for each time point. The homology classes trace out curves, which are called *vines*, in $\mathbb{R}^3$. At any single point in time, a homology class in the PD at that time corresponds to a (birth simplex, death simplex) pair. The birth simplex creates the homology class, and the death simplex destroys the homology class (see section 2.1). In a vineyard, a vine corresponds to a sequence of (birth simplex, death simplex) pairs. See section 2.2 for the definition of a vineyard.

1.1. Our contributions. We use TDA to analyze local extrema of real-valued geospatial data.[1] Our approach captures both local information (specifically, the geographical locations and the values of the local extrema) and global information about the relationships between the extrema. The global information includes the extent to which extrema are "spatially separated" (see Figure 1).

To the best of our knowledge, existing methods of analyzing local extrema yield only local information. One can check whether or not a geographical region is an extremum by comparing its associated value to those of its neighboring regions. However, this approach does not provide any global information about the extrema. For example, it cannot distinguish between the two cases in Figure 1.

Examining vineyards allows us to measure the persistence of extrema with time, observe how spatial separations between extrema change with time, and track how geographical locations of extrema change with time. We accomplish the last of these by using vineyards to match the extrema at one time to the corresponding extrema at another time. (They may not be at the same geographical locations.) We identify the geographical locations of extrema by examining the sequence of (birth simplex, death simplex) pairs for each vine. To the best of our knowledge, the present paper is the first paper that uses information about the sequence of (birth simplex, death simplex) pairs for each vine, rather than using only the (birth, death) filtration values for each vine. A naive approach, such as comparing each region to its neighboring regions at each time step, does not come with a natural way to match the extrema that one identifies at different times and does not provide information about changes in global structure. With our approach, we are able to track how the global spatial structure of data changes with time.

[1] See section 4 for our definitions of a "local maximum" and a "local minimum" of a real-valued function on a discrete set of geographical regions.



(a) (b)

Figure 1. (a) *The graph of a function* $f : \mathbb{R}^2 \to \mathbb{R}$ *that has two "well-separated" local maxima.* (b) *The graph of a function* $g : \mathbb{R}^2 \to \mathbb{R}$ *whose two local maxima have the same locations and values as* f *but are not well-separated from each other.*

Another contribution of our paper is a new method to construct an "efficient" simplicial complex whose underlying space[2] is homeomorphic to a geographical space (which is a set of regions, as we will explain shortly).[3] In our applications, we use geographical data in the form of SHAPEFILEs. Each geographical region (e.g., a neighborhood or zip code) is represented in a SHAPEFILE by a polygon (or by multiple polygons, if the region is disconnected) with many vertices (about 100 to 1000 vertices, depending on the particular SHAPEFILE and the particular region). These polygons approximate the real-life boundaries of the geographical regions. A naive approach to building a simplicial complex is to simply triangulate each of the polygons. However, this approach has two issues. The first issue is that there are often small overlaps between the polygons or spurious gaps between the polygons because the polygon boundaries do not exactly match the real-life geographical boundaries. The vertices of a polygon often lie in the interior of another polygon. The second issue is that simply triangulating these polygons, which each have a very large number of vertices, would create orders-of-magnitude more simplices than are necessary to represent a geographical space. It is important to attempt to minimize the number of simplices in a simplicial complex because PH and vineyard computation times are very sensitive to the number of simplices.

Rather than naively triangulating the given polygons, we use the SHAPEFILE of a geographical space to infer adjacency information about the regions; we then use only this information to build a simplicial complex for that geographical space. In the resulting simplicial complex, each region is represented by a union of triangles. We use about 1 to 10 triangles per region, depending on the number of neighbors of the region. By contrast, the naive approach above requires about 100 to 1000 triangles per region. Two adjacent regions that have a connected

[2]The *underlying space* of a simplicial complex is the union of its simplices. We note that it is common in studies of TDA for authors to conflate the combinatorial and topological structures of a simplicial complex.

[3]The simplicial complex is "efficient" in the sense that it minimizes the number of simplices.

intersection share exactly one edge in our simplicial complex, except in rare special cases that we will discuss in section 3. Our simplicial complex for a geographical space satisfies the following "topological correctness" property: the union of any subset of the space's geographical regions is homeomorphic to the underlying space of the simplicial subcomplex (see section 3 for the definition of a simplicial subcomplex) that is induced by the union of the corresponding triangles. When the geographical regions satisfy the mild assumptions (A1)–(A4) that we define in section 3, our construction uses the minimum number of simplices that is possible for a simplicial complex with the topological-correctness property above. (See property (P) in section 3 for a precise statement of this property.)

As case studies, we apply our approach to two data sets. The first data set is a geospatial data set of per capita vaccination rates in New York City (NYC) by zip code [11]. The homology classes correspond to zip codes in which the vaccination rates are either lower or higher (depending on choices that one can make in our approach) than in the neighboring zip codes. The estimates of these rates are at a single point in time (23 February 2021). The second data set consists of 14-day mean per capita COVID-19 case rates in neighborhoods in the city of Los Angeles (LA) in the time period 25 April 2020–25 April 2021. Modeling the spatiotemporal spread of COVID-19 is a complex task [1, 51]. In this geospatiotemporal data set, the homology classes of our approach correspond to COVID-19 anomalies, which are regions whose case rates are higher than in the neighboring regions.[4] It is important to examine such anomalies, as COVID-19 spreads with significant spatial heterogeneity and thus has heterogeneous effects on different geographical areas.[5] Many factors (such as mobility, population density, socioeconomic differences, and racial demographics) play a role in how COVID-19 affects different regions in disparate ways [9, 21, 22, 26]. In our case study of COVID-19 case rates in LA, we construct a vineyard that (1) conveys which anomalies are most persistent in time and (2) reveals how the anomalies move geographically with time.

1.2. Related work. Our method addresses several limitations of previous efforts to combine TDA with geospatial analysis. In [44], Stolz, Harrington, and Porter studied the percentage of United Kingdom voters by electoral district that voted to leave the European Union in the "Brexit" referendum. The holes that they identified using PH corresponded to districts that voted differently than the neighboring districts. However, their approach does not distinguish between homology classes that are merely noise and homology classes that correspond to small geographical districts. In [19], Feng and Porter developed an approach to study PH by constructing FSCs using the level-set method [33] of front propagation from scientific computation.[6] Using their level-set complexes, they examined the percentage of voters in each

[4] We examine *local* maxima in the case-rate data. This contrasts to COVID-19 "hotspots," which the CDC has defined using an absolute threshold for the number of cases and criteria that are related to the temporal increase in the number of cases [34].

[5] Other scholars have studied contagions using TDA in ways that do not yield topological features with geographical meaning. For example, recent work used TDA to study the spatiotemporal spread of COVID-19 [39] and Zika [41]. These papers examined topological features in atmospheric data, which were then used to forecast case rates. In [45], TDA was used to study the Watts threshold model of a social contagion on noisy geometric networks.

[6] The name "level-set method" may cause confusion. Importantly, the level-set simplicial complex of [19] is not the simplicial subcomplex that has simplices with some prescribed filtration value (i.e., a level set of the filtration values of a simplicial complex).

precinct of California counties that voted for a given candidate (e.g., Hillary Clinton) in the 2016 United States presidential election. The homology classes represent precincts that voted more heavily for Clinton than the neighboring precincts. The level-set complexes in [19] have two key limitations. The first is that they cannot handle time-dependent data, as they are built to study either data at a single point in time or data that has been aggregated over some time window to yield time-independent data. The second limitation is that these simplicial complexes reduce real-valued data (e.g., the percentage of voters who voted for Clinton) to binary data (e.g., whether or not the majority voted for Clinton). Consequently, in this example, the level-set-based PH does not capture the extent to which a blue "political island" voted more heavily for Clinton. By contrast, our approach is designed specifically to capture such information. As a trade-off, we no longer capture the geographical sizes of the political islands. For further discussion, see Feng, Hickok, and Porter [17], who applied the level-set filtration to study the cumulative case count of COVID-19 infections in Los Angeles on one specific day.

Our new approach to compute PH is also able to resolve some other technical issues in [19]. In particular, some of the homology classes in the level-set approach of [19] are geographical artifacts that are indistinguishable from true features of a data set. By contrast, the finite 1D homology classes in our approach are either in one-to-one correspondence with the local maxima of a real-valued geospatial function (i.e., a real-valued function on a set of regions) or in one-to-one correspondence with its local minima, depending on the choices that one makes. Additionally, unlike the level-set approach in [19], we are able to detect extrema that are adjacent to the boundary of a geographical space.

Other methods to construct simplicial complexes from geospatial data, such as rasterization of a SHAPEFILE or treating the regions of the data as a point cloud, require a trade-off between the number of simplices and the accuracy of the representation of the geographical regions. For example, the level-set-based PH method of [19] uses orders-of-magnitude more simplices to achieve sufficient resolution of small geographical regions (e.g., densely populated urban centers that are important to analyze). See section 7 for further discussion.

We use vineyards in the present paper, but there are also other ways to study the topology of time-varying data. For example, zigzag PH [7] was used in [13] to analyze time-dependent point clouds (such as swarms) and in [50] to study time-delay embeddings of dynamical systems. Crocker plots and crocker stacks (i.e., stacks of crocker plots for different values of a smoothing parameter) illustrate how the Betti numbers of a time-dependent point cloud change with time and with a scale parameter ϵ [52]. Additionally, Kim and Mémoli [25] used multiparameter PH [8] to study time-dependent point clouds. In sections SM1.3 and SM1.4 of the accompanying supplementary materials, we show how one can use multiparameter PH [3, 8] and multiparameter zigzag PH [7] to study our spatiotemporal COVID-19 data sets.

1.3. Organization. Our paper proceeds as follows. In section 2, we briefly review relevant topological background. In section 3, we formulate how we construct simplicial complexes. In section 4, we define several filtration functions and discuss how to interpret the resulting PDs and vineyards. In section 5, we apply our method to the LA and NYC data sets. In section 6, we discuss our methodological choices. In section 7, we summarize our work and discuss

some of its implications. In Appendix A, we discuss technical details of the simplical-complex construction. In the accompanying supplementary material file supplement.pdf [local/web 779KB], we discuss alternative topological approaches for studying PH in geospatiotemporal data, provide further information about the LA results, compare our approach to an "all-but-one" statistical test, and show some demographic data. Our code is available at https://bitbucket.org/ahickok/vineyard/src/main/.

2. Background.

2.1. Persistent homology.

We give a brief introduction to PH. See [35] for a more thorough discussion of it.

A *k-simplex* is k-dimensional polytope that is the convex hull of $k+1$ vertices. The convex hull of a subset of these vertices is a *face* of the simplex. A *simplicial complex* $\mathcal{K}$ is a set of simplices that satisfies two requirements: (1) if $\sigma \in \mathcal{K}$ is a k-simplex, then every face of σ is in $\mathcal{K}$; (2) if σ and τ are simplices in K, then any nonempty $\sigma \cap \tau$ is a face of both σ and τ.

A *filtered simplicial complex* (FSC) is a nested sequence $\mathcal{K}_{\alpha_0} \subseteq \cdots \subseteq \mathcal{K}_{\alpha_n} = \mathcal{K}$ of simplicial complexes for some sequence $\{\alpha_0, \ldots, \alpha_n\}$ of indices. See Figure 2 for an example of an FSC. A *filtration function* (or simply a *filtration*) $f : \mathcal{K} \to \mathbb{R}$ is a function such that if the simplex $\tau \in \mathcal{K}$ is a face of $\sigma \in \mathcal{K}$, then $f(\tau) \leq f(\sigma)$. A pair $(\mathcal{K}, f)$ induces an FSC as follows. Let $\mathcal{K}_\alpha := \{\sigma \in \mathcal{K} \mid f(\sigma) \leq \alpha\}$ be the *α-sublevel simplicial complex*, and let $\{\alpha_0, \ldots, \alpha_n\}$ be the image of f. The sequence $\mathcal{K}_{\alpha_0} \subseteq \cdots \subseteq \mathcal{K}_{\alpha_n} = \mathcal{K}$ is a nested sequence of simplicial complexes. In our paper, we often refer to the pair $(\mathcal{K}, f)$ as the FSC itself. We do this because it is the most natural way to define the FSCs for our applications.

We compute the homology of each K_{α_i} over a field $\mathbb{F}$, which we set to $\mathbb{Z}/2\mathbb{Z}$ in the present paper. Let $H_p(\mathcal{K}_{\alpha_i}, \mathbb{F})$ denote the p-dimensional homology of $\mathcal{K}_{\alpha_i}$ over $\mathbb{F}$. Homology classes represent connected components, holes, and higher-dimensional voids in a simplicial complex; specifically, p-dimensional homology classes represent p-dimensional "holes." The inclusion relationship $\mathcal{K}_{\alpha_i} \hookrightarrow \mathcal{K}_{\alpha_{i+1}}$ between subcomplexes induces a map $\iota_i : H_p(\mathcal{K}_{\alpha_i}, \mathbb{F}) \to H_p(\mathcal{K}_{\alpha_{i+1}}, \mathbb{F})$ from the homology of $\mathcal{K}_{\alpha_i}$ to the homology of $\mathcal{K}_{\alpha_{i+1}}$. The inclusion map ι_i lets us track an element of $H_p(\mathcal{K}_{\alpha_i}, \mathbb{F})$ to an element of $H_p(\mathcal{K}_{\alpha_{i+1}}, \mathbb{F})$. The p-dimensional PH of an FSC is the pair

$$\left(\{H_p(\mathcal{K}_{\alpha_i}, \mathbb{F})\}_{0 \leq i \leq n}, \{\iota_i\}_{0 \leq i < n}\right). \tag{2.1}$$

We say that a homology class γ is *born* at filtration level α_i if i is the smallest index for which γ is an element of $H_p(\mathcal{K}_{\alpha_i}, \mathbb{F})$. We say that the homology class γ *dies* at filtration level α_j if α_{j-1} is the last filtration level at which γ exists. That is, $\iota_{j-1} \circ \cdots \circ \iota_i$ maps $\gamma \in H_p(\mathcal{K}_{\alpha_i}, \mathbb{F})$ to 0 in $H_p(\mathcal{K}_{\alpha_j}, \mathbb{F})$ and for all $k < j-1$, we have $\iota_k \circ \cdots \circ \iota_i(\gamma) \neq 0$. Not every homology class dies; we refer to classes that do die as *finite* and classes that do not die as *infinite*.

The Fundamental Theorem of Persistent Homology yields a set of generators for the PH of an FSC [15, 16]. Each generator is a homology class. A generator has a *birth simplex* σ_b that creates the homology class and (if finite) a *death simplex* σ_d that destroys the homology class. If one is computing homology in dimension p, then σ_b is a p-dimensional simplex and σ_d is a $(p+1)$-dimensional simplex. The simplex pair (σ_b, σ_d) represents the homology class. For example, in Figure 2, there is one 1D PH generator. Its birth simplex is the edge $(0, 3)$ because this is the edge that completes the loop that encircles the hole, and its death simplex

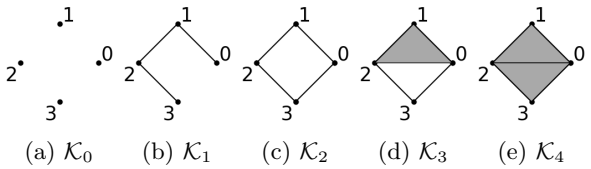


Figure 2. *An example of nested simplicial complexes in an FSC.*

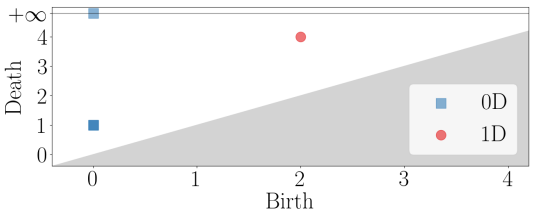


Figure 3. *The PD of the FSC in Figure 2.*

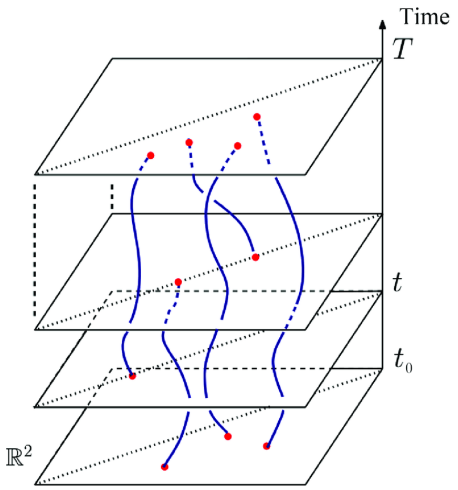


Figure 4. *An example of a vineyard. Each curve is a vine in the vineyard. [This figure is a slightly modified version of a figure that appeared originally in* [27]*.]*

is the triangle $(0, 2, 3)$ because this is the triangle that fills in the hole. The birth filtration level of the homology class is $f(\sigma_b)$ and the death filtration level (if finite) is $f(\sigma_d)$.

A *persistence diagram* (PD) is a way of representing PH as a multiset of points in the extended plane $\overline{\mathbb{R}}^2$. Each off-diagonal point represents a generator of the PH; the point's coordinates are the homology class's birth and death filtration levels. One includes the points on the diagonal for technical reasons; one can think of them as homology classes that die instantaneously upon birth. See Figure 3 for an example of a PD.

2.2. Vineyards. The examination of *vineyards* is one way to study time-varying PH [27]. A *time-dependent filtration function* on a simplicial complex $\mathcal{K}$ is a function $f : [t_0, T] \times \mathcal{K} \to \mathbb{R}$ such that $f(t, \cdot)$ is a filtration for all t. We compute the PH of $(\mathcal{K}, f(t, \cdot))$ for all times t. We visualize the vineyard in $\overline{\mathbb{R}}^2 \times [t_0, T]$ as a continuous stack of PDs (see Figure 4). The points in the PDs trace out curves with time; these curves are the *vines*. Each vine corresponds to a homology class; a vine is the graph of the birth and death filtration levels of a particular homology class as a function of time. The homology class that is represented by a vine has a *time-dependent birth simplex* $\sigma_b(t)$ and (if finite) a *time-dependent death simplex* $\sigma_d(t)$. At time t, the homology class is created by the simplex $\sigma_b(t)$ at filtration level $f(t, \sigma_b(t))$ and

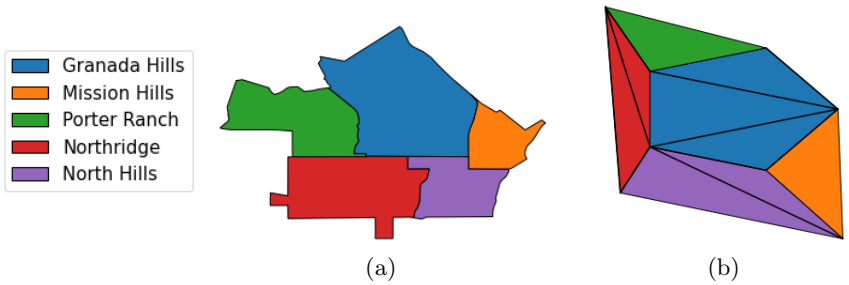


Figure 5. (a) *A set S of geographical regions, as given by a* SHAPEFILE [23]. (b) *The resulting simplicial complex $\mathcal{K}$.*

(if finite) destroyed by the simplex $\sigma_d(t)$ at filtration level $f(t, \sigma_d(t))$. The functions $\sigma_b(t)$ and $\sigma_d(t)$ are piecewise constant. We measure the overall *persistence* of a vine by calculating $\int_{t_0}^{T} [f(t, \sigma_d(t)) - f(t, \sigma_b(t))]\, \mathrm{d}t$.

Cohen-Steiner, Edelsbrunner, and Morozov [12] developed an algorithm for computing vineyards when they introduced them. One computes the initial PH at time $t = t_0$, and one then updates the pairings of birth and death simplices as the order of the simplices (as induced by $f(t, \cdot)$) changes with time. Each change in the order of the simplices occurs one transposition at a time. One can make these updates in $O(m)$ time (where m is the number of simplices) per transposition of simplices.

3. Constructing a simplicial complex. We now show how we construct a simplicial complex $\mathcal{K}$ from geographical data (e.g., a SHAPEFILE that specifies approximate geographical boundaries of a set of geographical regions). We partition a given geographical space into *regions*. In section 5.1, the regions are zip codes in NYC; in section 5.2, the regions are neighborhoods in the city of LA. Let S be the set of regions. We refer to the complement of $\bigcup_{R \in S} R$ as the *exterior region*. We construct a 2D simplicial complex $\mathcal{K}$ with the following property:

(P) There is an assignment of 2D simplices to regions such that the union of any subset of regions is homeomorphic to the underlying space of the *simplicial subcomplex*[7] that is induced by the union of the corresponding 2D simplices.

In Figure 5, we show an example of our construction, which we discuss in this section and present in more detail in Appendix A. Under the mild assumptions (A1)–(A4) that we define shortly, our simplicial complex has the minimum number of simplices that is possible for a simplicial complex that satisfies property (P). Constructing an efficient simplicial complex is important because the time that it takes for TDA computations depends sensitively on the number of simplices in a simplicial complex.

In our case studies, the geographical data take the form of SHAPEFILEs. In a SHAPEFILE, each region is represented by a *polygon with holes*[8] (or by multiple polygons with holes, if the

[7] The simplicial subcomplex that is induced by a set $E \subseteq \mathcal{K}$ is the smallest simplicial complex $\mathcal{K}'$ that contains the set E of simplices. That is, if $\mathcal{K}''$ is a simplicial complex that contains E, then $\mathcal{K}' \subseteq \mathcal{K}''$. When $\mathcal{K}$ is 1D, a simplicial subcomplex is equivalent to an induced subgraph.

[8] A polygon with holes is $P = Q - \bigcup_{i=1}^{h} \mathrm{int}(H_i)$, where Q is a polygon that encloses polygons $H_1, \ldots, H_h$ (the holes) [37] and $\mathrm{int}(H_i)$ denotes the interior of H_i. It is possible to have $h = 0$ holes. (In that case, $P = Q$.)

region is disconnected) that closely approximates the actual geographical region. (A SHAPEFILE stores the coordinates of the boundaries of the polygons.) For an example of SHAPEFILE data, see Figure 5(a). As we discussed in section 1.1, the polygon boundaries are not always aligned perfectly, so their interiors sometimes overlap and gaps can occur between them. Therefore, to construct a simplicial complex $\mathcal{K}$, we must do more than merely triangulate these polygons. Additionally, the polygons in our SHAPEFILEs have roughly between 100 and 1000 vertices, which is many more vertices per region than in the simplicial complex $\mathcal{K}$ that we will construct shortly.

We make the following assumptions about geographical regions:

(A1) There are a finite number of regions, and each region has a finite number of connected components.

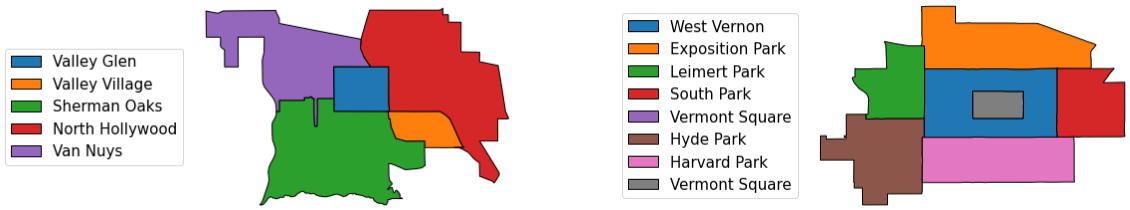
(A2) Each component of a region is homeomorphic to $D_0 - \bigcup_{i=1}^{h} \text{int}(D_i)$, where D_0 is a closed disk that encloses some number (which can be 0) of other closed disks $D_1, \ldots, D_h$ (i.e., the holes in the region). For all $i \neq j$, the intersection $D_i \cap D_j$ has at most one point. See, for example, the West Vernon region in Figure 6(b); it is homeomorphic to $D_0 - D_1$ (an annulus) for two disks D_0 and D_1 that do not intersect. (In our case studies, it is rare for any of the disks to intersect.)

(A3) The intersection between any two regions has a finite number of components, and the interiors of the regions do not intersect.

(A4) The intersection between three or more regions is either a point or $\emptyset$.

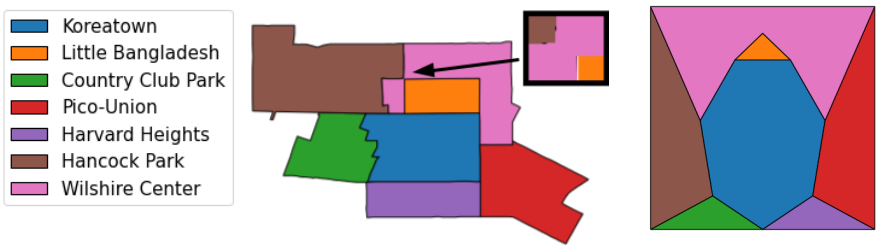
Assumptions (A1)–(A4) are very reasonable for human-made geographical boundaries. We do not even require the regions to be simply connected or the region intersections to be connected. In Figure 5(a), we illustrate the most typical situation that we encounter. In this example, LA neighborhood Granada Hills is homeomorphic to a disk and its boundary intersects the boundaries of five neighboring regions (counting the exterior region). In Figures 6 and 7(a), we illustrate a few other situations that can arise in geospatial applications.

We now outline our procedure for building a simplicial complex. For each region R, we construct a "reduced" polygon with holes P^R that has orders-of-magnitude fewer vertices than the polygons with holes in the associated SHAPEFILE. The number of holes in P^R is equal to the number of holes in the geographical region R. We glue the boundaries of $\{P^R \mid R \in S\}$ together in a way that respects the geographical region boundaries. We then



(a) Valley Glen (b) West Vernon

Figure 6. *Various neighborhoods of Los Angeles, as given by a* SHAPEFILE [23]. (a) *The four neighborhoods Valley Glen, Valley Village, Sherman Oaks, and North Hollywood intersect in a point.* (b) *The neighborhood West Vernon has a hole because of its neighbor Vermont Square.*



(a) (b)

Figure 7. (a) *A geographical set S that consists of the neighborhood Koreatown and its neighbors, as given by a* SHAPEFILE [23]. *Observe that the neighborhood Little Bangladesh has only two neighbors and that the intersection between Koreatown and Wilshire Center has two components.* (b) *The result of gluing Koreatown's polygon to the polygons of its neighbors.*

triangulate each of the polygons to obtain a 2D simplicial complex $\mathcal{K}$. We assign a 2D simplex $\sigma \in \mathcal{K}$ to the region R whose polygon P^R originally contained σ. In Figure 5, we show an example of the resulting $\mathcal{K}$. Our code for our simplicial-complex algorithm is available at https://bitbucket.org/ahickok/vineyard/src/main/.[9] In the remainder of this section, we discuss the details of this process.

3.1. Constructing a reduced polygon with holes for each region. Without loss of generality, we assume that each region is connected; if not, we treat each component of a region as if it were its own region. For each region R, we construct a reduced polygon with holes P^R using adjacency information that we infer from a SHAPEFILE. Let $D_0, D_1, \ldots, D_h$ be the disks in the statement of assumption (A2), and let $B_i = \partial D_i$. Under the geographical assumptions (A1)–(A4), the intersections of a region R with its neighbors are such that for each boundary B_i, one can order the neighbors in clockwise (or counterclockwise) fashion, possibly with repetition.[10] Let S_i denote this sequence of neighbors around B_i. We list intersections with the exterior region in the same manner as for any other neighboring region. We also record whether each intersection is 1D or 0D. For example, in Figure 6(a), the clockwise sequence of neighbors around the boundary of Valley Glen is {Van Nuys, North Hollywood, Valley Village, Sherman Oaks}. The intersection with Valley Village is 0D and the other intersections are 1D. For regions such as West Vernon in Figure 6(b), we obtain a sequence S_i for each boundary B_i. Each sequence is unique up to the choice of starting neighbor.

Given a sequence of neighbors for each boundary B_i (which, if necessary, we adjust as in Appendix A.1), we construct a polygon with holes P^R as follows. Let $(P')^R$ be a polygon that has one edge for each $N \in S_0$ for which the corresponding component of $N \cap B_0$ is 1D. Let

[9] This code has one limitation that the algorithm in the present paper does not. It requires that no interior region (i.e., a region that is contained within the outer boundary of another region) intersects any other interior region. This does not occur in our data, and we believe that it does not occur in most geographical spaces.

[10] Theoretically, several 0D intersections can be adjacent to each other, although this scenario does not occur in our data sets. That is, in principle, there can exist a sequence $\{N_i, \ldots, N_{i+k}\}$ of neighbors such that $N_j \cap R$ is the same point p for all j. The order of this sequence is not determined uniquely by the intersections of the neighbors with R. Instead, we order them in the order in which they appear clockwise (or counterclockwise) around the point p. This sequence must be finite because there are a finite number of regions and (A2) implies that $N_{j_1} \neq N_{j_2}$ if $j_1 \neq j_2$.

$\{H_i^R\}_{i=1}^h$ be a set of polygons that are contained in $(P')^R$ and satisfy the following properties:

1. H_i^R has one edge for each $N \in S_i$ for which the corresponding component of $N \cap B_i$ is 1D,
2. $H_i^R \cap H_j^R \neq \emptyset$ if and only if $D_i \cap D_j \neq \emptyset$,
3. $P^R \cap H_i^R \neq \emptyset$ if and only if $D_0 \cap D_i \neq \emptyset$, and
4. if the intersection of two polygons in $\{P^R, H_1^R, \ldots, H_h^R\}$ is nonempty, then the intersection is a vertex.

The locations of the vertices do not matter. We define P^R to be $(P')^R - \bigcup_{i=1}^h \operatorname{int}(H_i^R)$, which is homeomorphic to R by assumption (A2). Finally, we annotate each edge of P^R with the neighbor that corresponds to it. We also annotate each vertex with the sequence of its adjacent regions, which we list in clockwise order starting with R.

3.2. Gluing together the polygons with holes. We glue the polygons with holes $\{P^R \mid R \in S\}$ along their edges according to their edge and vertex annotations. More precisely, if P^{R_1} has n nonadjacent edges with the annotation R_2 (which is the typical situation when $R_1 \cap R_2$ has n components that are 1D), then P^{R_2} has exactly n nonadjacent edges with the annotation R_1. For example, in Figure 7, R_1 = Koreatown and the annotated polygon with holes P^{R_1} has two edges with the annotation R_2 = Wilshire Center. Let (u, v), with u and v in clockwise order, be the vertices of an edge in P^{R_1} with annotation R_2. Because the n edges are nonadjacent, u and v must each have at least three neighbors (including R_1 and R_2). For example, in Figure 7, again consider the two edges with the annotation Wilshire Center. The two vertices u_1 and v_1 of one edge have the adjacency sequences {Koreatown, Hancock Park, Wilshire Center} and {Koreatown, Wilshire Center, Little Bangladesh}, respectively. The two vertices u_2 and v_2 of the other edge have the adjacency sequences {Koreatown, Little Bangladesh, Wilshire Center} and {Koreatown, Wilshire Center, Pico-Union}, respectively. For a given (u, v), we seek an edge (x, y) (with x and y in clockwise order) in P^{R_2} with the annotation R_1 such that (1) u and y are annotated with the same sequences and (2) v and x are annotated with the same sequences. We know that there must be at least one such edge because (u, v) represents a component of $R_1 \cap R_2$ and there is some edge in P^{R_2} that represents the same component (so its vertices have the same sequences of adjacent regions as u and v). In Lemma A.2, we prove that there is a unique such edge. If there are $n > 1$ consecutive edges $e_0, \ldots, e_{n-1}$ on the boundary of $\mathcal{K}_{R_1}$ with annotation R_2, then there are n consecutive edges $e'_0, \ldots, e'_{n-1}$ on the boundary of $\mathcal{K}_{R_2}$ with annotation R_1. This situation arises precisely because of the adjustments that we discuss in Appendix A.1. We glue e_i to e'_{n-i} for all i. If $R_1 \cap R_2$ is homeomorphic to S^1, then the choice of e'_0 as the first edge in P^{R_2} is not unique, but all choices result in topologically equivalent spaces. In Figure 7(b), we show the result of the gluing process for Koreatown and its neighbors.

3.3. Triangulating the polygons with holes. We triangulate each polygon with holes P^R using the inductive algorithm in [37]. We show examples of triangulated polygons with holes in Figure 8. The result of this triangulation process is a 2D simplicial complex $\mathcal{K}$ with property (P). (The assignment in property (P) maps a 2D simplex in the polygon with holes P^R to the geographical region R.) The simplicial complex $\mathcal{K}$ is a minimal simplicial complex with property (P) because (1) each polygon with holes P^R has the minimum number of vertices and holes and (2) the number of triangles in the triangulation of P^R is determined by

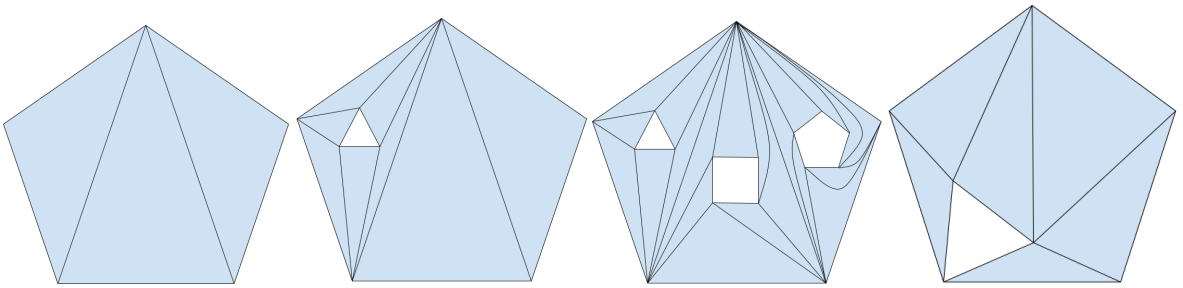


Figure 8. *Triangulation of a polygon with holes P^R for a region R when* (a) *R has no holes,* (b) *R has a single hole,* (c) *R has multiple holes, and* (d) *R has a hole that touches the exterior boundary of R.*

its number of vertices and its number of holes by Euler's theorem (see [37]). For an example, see Figure 5(b).

4. Our filtration functions. We define various filtrations that one can use with the simplicial complex $\mathcal{K}$ that we constructed in section 3, and we discuss how to interpret the resulting PDs and vineyards. Let S be the set of geographical regions R that the simplicial complex $\mathcal{K}$ represents, and let $F : S \to \mathbb{R}$ be a real-valued function on S. For example, in section 5.1, $F(R)$ is the per capita full-vaccination rate (i.e., having received all required doses of some vaccine) for COVID-19 in NYC zip code R. In sections 4.1 and 4.2, we define two filtration functions that are induced by F. Given a time-dependent and real-valued function $F(t, R)$, we define time-dependent filtration functions in section 4.3. For example, in section 5.2, $F(t, R)$ is the 14-day mean per capita COVID-19 case rate in neighborhood R on day t. From a time-dependent filtration function, we compute a vineyard.

4.1. The sublevel-set filtration. In this subsection, we define a sublevel-set filtration. In our applications, we use the 1D PH of the sublevel-set filtration to analyze local maxima in our data sets. We illustrate the idea of a sublevel-set filtration in Figure 9.

Definition 4.1 (sublevel-set filtration). *Let $\mathcal{K}$ be the simplicial complex that we obtain from our construction in section 3 for a set S of regions, and let g be the assignment of 2D simplices to the regions. Let $F : S \to \mathbb{R}$. We define the sublevel-set filtration function f by considering the sublevel sets of F. On the 2D simplices, we define the filtration function by*

$$f(\sigma) = F(g(\sigma)) .$$

We extend the filtration function to the remaining (lower-dimensional) simplices as follows. If σ is a vertex or edge on the boundary of $\mathcal{K}$, we set

$$f(\sigma) = \min_R F(R) .$$

Otherwise, we set

$$f(\sigma) = \min\{f(\tilde{\sigma}) \mid \tilde{\sigma} \text{ is a 2D simplex for which } \sigma \text{ is a vertex or edge of } \tilde{\sigma}\}. \tag{4.1}$$

At filtration level α, the simplicial complex $\mathcal{K}_\alpha$ is the simplicial subcomplex of $\mathcal{K}$ that is induced by the union of the set of 2D simplices σ such that $F(g(\sigma)) \leq \alpha$ and the set of vertices

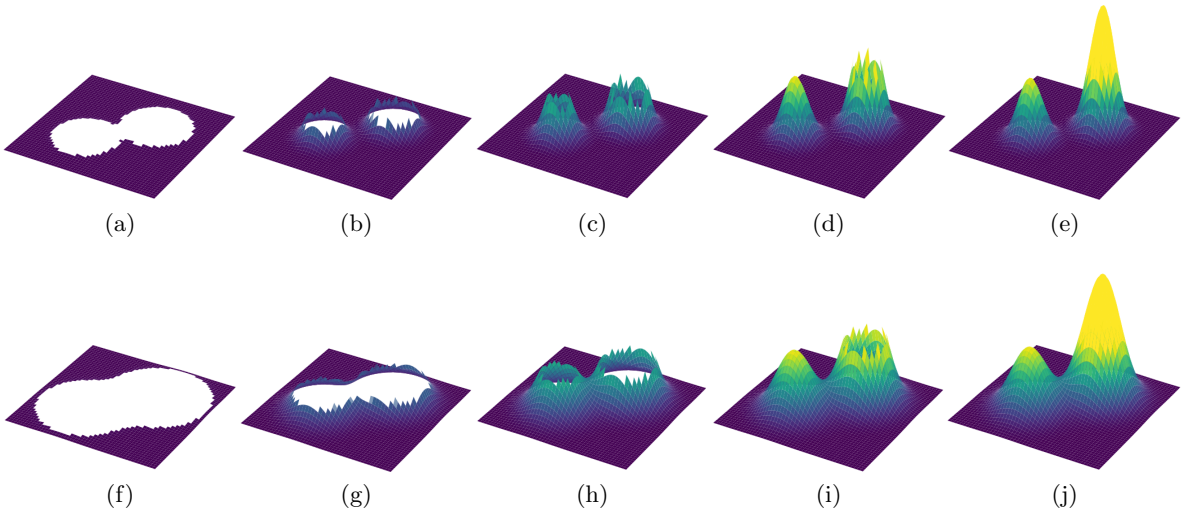


Figure 9. *In panels* (a)–(e)*, we show the α-sublevel sets for increasing α of a function $f : \mathbb{R}^2 \to \mathbb{R}$ that has two well-separated local maxima. In* (a)*, for the smallest value of α, there is one hole that corresponds to the global maximum. In* (b)*, a second hole appears; it corresponds to the other local maximum. In* (d)*, the second hole is filled in. In* (e)*, the first hole is filled in. In panels* (f)–(j)*, we show the α-sublevel sets for increasing α of a function $g : \mathbb{R}^2 \to \mathbb{R}$ whose two local maxima have the same locations and values as f but are not well-separated from each other. The second hole does not appear until the sublevel set in panel* (h)*. In all panels, the jagged edges are artifacts of the way that the Python package* MATPLOTLIB *plots surfaces.*

and edges that are on the boundary of $\mathcal{K}$. Henceforth, we say that the vertices and edges on the boundary of $\mathcal{K}$ are "exterior-adjacent." By construction, the underlying space of $\mathcal{K}_\alpha$ is homeomorphic to the union of all regions R such that $F(R) \leq \alpha$ and the exterior boundary. We set $f(\sigma) = \min_R F(R)$ for exterior-adjacent vertices and edges σ for technical reasons that we will explain in a few paragraphs. In section SM1.2 of the accompanying supplementary materials, we explore an alternative definition in which we set the filtration values of exterior-adjacent vertices and edges σ to $\min_R\{F(R) \mid R \subset C\}$, where C is the connected component that contains σ.

The 1D PH of the sublevel-set filtration encodes information about the structure of the local maxima of F. A region R of a geographical space is a local maximum if the value $F(R)$ is larger than the value $F(N)$ for all neighboring regions N of R for which $N \cap R$ is 1D. More generally, we consider a set $E \subseteq S$ of regions (where $|E| = 1$ is possible) to be a *local maximum* if

1. the interior of $\bigcup_{R\in E} E$ is connected,
2. the value of F is constant on E (we denote this value by $F(E)$), and
3. the value $F(E)$ is larger than the value $F(N)$ for all regions $N \notin E$ such that $N \cap R$ is 1D for some $R \in E$.

If E is a local maximum, there is a 1D homology class whose death simplex is one of the simplices in the preimage $g^{-1}(E)$, where g is the map from 2D simplices in $\mathcal{K}$ to geographical regions in S. The class dies at filtration level $\alpha = F(E)$. For example, if $F(R)$ is the COVID-19 case rate in region R, then the 1D homology classes correspond to COVID-19 anomalies and the death simplex of a 1D homology class indicates the epicenter of that anomaly. The larger

the value $F(E)$ in comparison to nearby regions (including regions that are not necessarily immediate neighbors), the more persistent the homology class is. If the union of all regions (excluding the exterior region) is not simply connected, then there is at least one 1D homology class with an infinite death time. See Figure 11(b) for an example. The infinite 1D homology classes correspond to the holes in the geographical space, rather than to local maxima. The local maxima of F are in one-to-one correspondence with the set of 1D homology classes with finite death times.[11] There is a canonical mapping from finite 1D homology classes to regions. A class that is represented by the simplex pair (σ_b, σ_d) is mapped to the region $g(\sigma_d)$ that includes σ_d. The region $g(\sigma_d)$ is the location of the local maximum of F that corresponds to the homology class,[12] and the death simplex's filtration value $f(\sigma_d)$ is the value of the local maximum. The death simplices of the finite 1D homology classes and their filtration values give the local-maximum locations R and their function values $F(R)$.

With the 1D PH, we can do more than simply identify local maxima and their locations; the 1D PH also reveals information about relationships between the local maxima. If the local maxima are well-separated from one another, then the corresponding homology classes all have early birth times. For example, the NYC data set has several connected components. One can think of the global maximum of each connected component as "totally separated" from each other because they are on different connected components. The corresponding 1D homology classes are all born at the earliest possible filtration time, which is $\min_R F(R)$ (see Figure 12(a)). We showed an example of well-separated local maxima in Figure 1(a). By contrast, the two local maxima in Figure 1(b) are not well-separated, so the homology class that corresponds to the lower peak in Figure 1(b) is born at a larger filtration value than the homology class in Figure 1(a). See Figure 9 for visualizations of the sublevel sets of the functions in Figure 1. The birth times of the 1D homology classes reflect structural information about the local maxima.

We set the filtration value of exterior-adjacent vertices and edges to the global minimum $\min_R F(R)$ so that 1D PH can detect local maxima on the boundary of a geographical space. (We consider an alternative approach in the accompanying supplementary material file supplement.pdf [local/web 779KB].) This is important for the LA data set of COVID-19 case rates. As we can see in Figure 17, many of the most-persistent COVID-19 anomalies are on the boundary of the geographical space; it is crucial that we are able to detect them. If we had not defined the exterior-adjacent filtration values in this way, then the filtration value of exterior-adjacent vertices and edges σ would be $F(R)$, where R is the unique region that is adjacent to σ. If R is a local maximum, its corresponding 1D homology class is born and dies at filtration level $\alpha = F(R)$. In the PD, it then appears as a point on the diagonal. Therefore, for 1D PH to detect local maxima on the boundary of a geographical space, we must adjust the filtration values of exterior-adjacent vertices and edges.

[11] Recall that in our definition of a local maximum, we only compare the value in a region R (or the constant value in a set E of regions) to the values in neighboring regions N that have 1D intersections with R (or with a region in E). It is possible for two local maxima, R_1 and R_2, to have a 0D intersection. In that case, we let N be the set of regions that are adjacent to $R_1 \cup R_2$. It is then the case that N is homotopy-equivalent to a figure-8, which has two 1D homology generators. One of the generators corresponds to R_1 and the other generator corresponds to R_2.

[12] Let $E \subseteq S$ be the local maximum that corresponds to the 1D homology class. If $E = \{R\}$, then $g(\sigma_d) = R$. However, if E contains multiple regions, then $g(\sigma_d)$ is only one of the regions in E.

The 0D homology classes correspond to local minima of F. However, unlike for the 1D homology classes, there is not a natural mapping from 0D homology classes to the locations of the minima. In the accompanying supplementary material file supplement.pdf [local/web 779KB], we discuss the interpretation and computation of 0D homology classes.

4.2. The superlevel-set filtration. An alternative to using the sublevel-set filtration from section 4.1 is to use superlevel sets of F to construct a superlevel-set filtration. In our case study of COVID-19 vaccination rates in NYC, we use a superlevel-set filtration to analyze local minima of the vaccination rate. We define a *local minimum* analogously to the way that we defined a local maximum in section 4.1. We illustrate the idea of the superlevel-set filtration in Figure 10.

Definition 4.2 (superlevel-set filtration). *Let $F : S \to \mathbb{R}$ for a set S of regions. The superlevel-set filtration function f is the sublevel-set filtration function that is induced by $-F$.*

At filtration level $-\alpha$, the simplicial complex $\mathcal{K}_{-\alpha}$ is the simplicial subcomplex of $\mathcal{K}$ that is induced by the union of the set of exterior-adjacent simplices and the set of 2D simplices σ for which $F(g(\sigma)) \geq \alpha$. By construction, the underlying space of $\mathcal{K}_{-\alpha}$ is homeomorphic to the union of regions R for which $F(R) \geq \alpha$ along with the exterior boundary. Local maxima of F now correspond to 0D homology classes, and local minima of F now correspond to 1D homology classes; this is the opposite situation from the sublevel-set filtration. Our discussion of local maxima for the sublevel-set filtration in section 4.1 applies to local minima for the superlevel-set filtration, and our discussion of local minima for the sublevel-set filtration in section 4.1 applies to local maxima for the superlevel-set filtration. The only difference is that the filtration values in the superlevel-set filtration are the additive inverses of the function values of F. This implies, for example, that the death filtration value of a 1D homology class that corresponds to a local minimum at region R is $\alpha = -F(R)$, rather than $\alpha = F(R)$.

4.3. A time-dependent filtration. Suppose that we have a time-dependent, real-valued function $F(t, R)$ whose domain is $\{t_0, t_1, \ldots, t_n\} \times S$, where $t_0 \in \mathbb{R}$ is the initial time and $t_n \in \mathbb{R}$ is the final time. For example, in section 5.2, $F(t, R)$ is the 14-day mean per capita COVID-19 case rate in Los Angeles in neighborhood R on day t. We seek to analyze the structure of local extrema as they change with time.

Definition 4.3 (time-dependent sublevel-set filtration). *Let $F : \{t_0, t_1, \ldots, t_n\} \times S \to \mathbb{R}$ be a time-dependent function on a set S of regions, and let $\mathcal{K}$ be the simplicial complex for S from the construction in section 3. At each time $t_i \in \{t_0, t_1, \ldots, t_n\}$, we define the time-dependent*

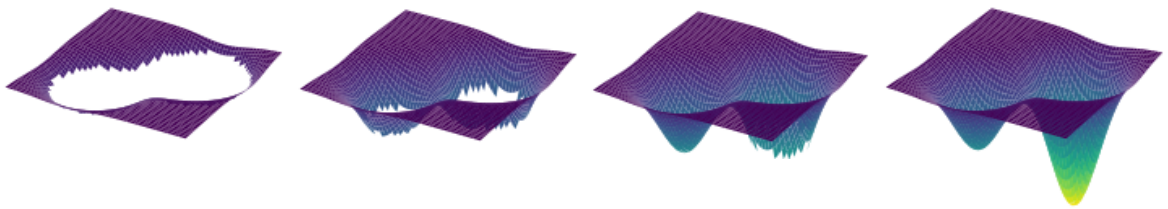
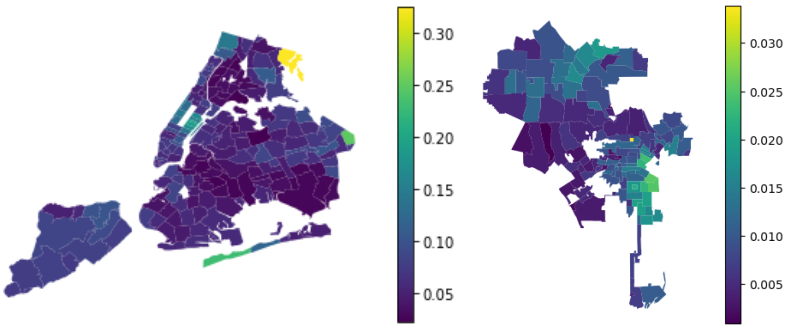


Figure 10. *The α-superlevel sets, with α decreasing from left to right, for the graph of a function $f : \mathbb{R}^2 \to \mathbb{R}$ with two local minima.*



(a) NYC zip codes (b) LA neighborhoods

Figure 11. *We show* (a) *the per capita COVID-19 full-vaccination rate in NYC by (modified) zip code on* 23 *February* 2021 *and* (b) *the* 14*-day mean per capita COVID-19 case rate in the city of LA by neighborhood on* 30 *June* 2020*. In both* (a) *and* (b)*, the white regions are geographical regions that do not belong to the depicted city.*

filtration function $f(t_i, \cdot)$ to be the sublevel-set filtration that is induced by $F(t_i, \cdot)$. To extend this filtration function to the entire interval $[t_0, t_n]$, we linearly interpolate $f(\cdot, \sigma)$ on each subinterval $[t_i, t_{i+1}]$ for all simplices $\sigma \in \mathcal{K}$.

In the present paper, we only use the time-dependent sublevel-set filtration, but one can analogously define a time-dependent superlevel-set filtration. We have implemented both of these filtrations in our code.

We use a time-dependent sublevel-set filtration to construct a vineyard. This allows us to track how the extrema move in both space and time. As in section 4.1, each finite vine corresponds to a local maximum whose location at time t is given by the region $g(\sigma_d(t))$ that contains the vine's time-dependent death simplex $\sigma_d(t)$.[13] The length of a vine corresponds to its persistence in time.

5. Case studies.

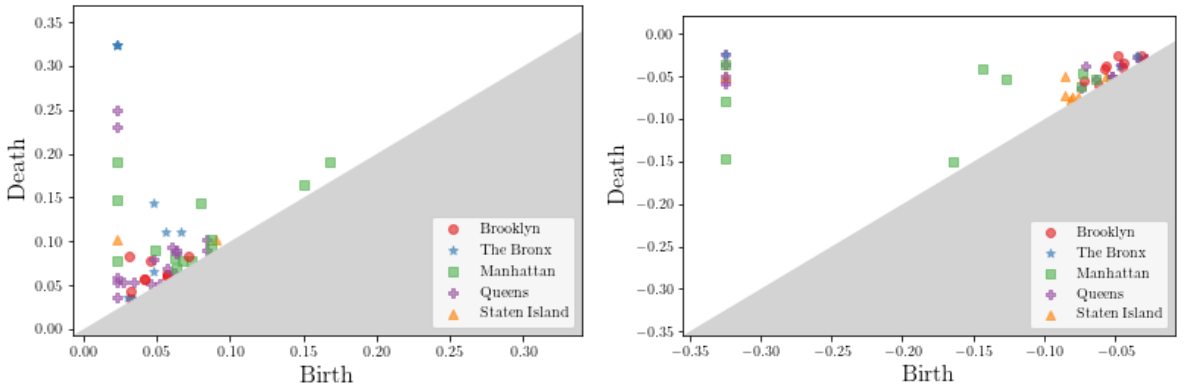
We now apply our methods to two data sets, which we illustrate in Figure 11.

5.1. COVID-19 vaccination rates in New York City.

We examine vaccination rates in (modified) zip codes of NYC[14] and we demonstrate the two filtrations that we defined in section 4. The geographical boundaries of the zip codes are specified by a SHAPEFILE [32]. From the SHAPEFILE, we construct a simplicial complex $\mathcal{K}$ in the manner that we described in section 3. Our vaccination data set, which we obtained from the NYC Department of Health and Mental Hygiene website [11], consists of the number of fully vaccinated people in each

[13] It is known that vineyards are not stable [52]. A small perturbation in filtration values can cause crossings of vines that previously did not cross (i.e., an "avoided crossing"). This, in turn, causes simplex pairings to change. Therefore, the geographical region $g(\sigma_d(t))$ that corresponds to a particular vine at time t is sensitive to small perturbations in filtration values.

[14] The NYC Department of Health and Mental Hygiene uses modified zip code tabulation areas (MODZCTA) for COVID-19 data [32]. In these modified zip codes, some zip codes with small populations are combined [31]. We henceforth refer to modified zip codes as simply "zip codes."



(a) Sublevel-set filtration (b) Superlevel-set filtration

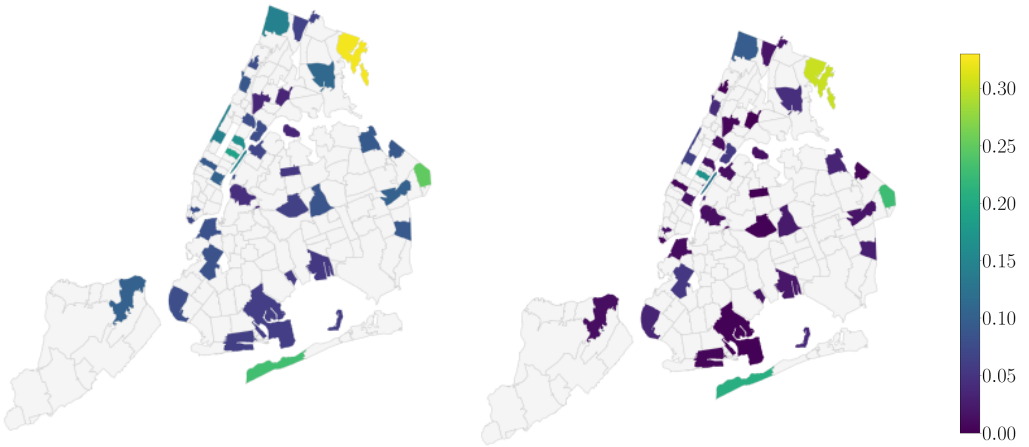
Figure 12. *PDs for the* 1*D PH of the NYC simplicial complex with filtrations that are induced by the per capita full-vaccination rate by zip code on* 23 *February* 2021*. We show only the finite homology classes. Each point in a PD corresponds to a zip code, which we label according to its borough* [30]*, that has* (a) *a higher vaccination rate than its neighboring zip codes or* (b) *a lower vaccination rate than its neighboring zip codes.*

zip code on 23 February 2021.[15] For each zip code, we divide this number by its population estimate in [11] to obtain a per capita vaccination rate. For zip code R, we define $F(R)$ to be the per capita vaccination rate in R on 23 February 2021.

We do not possess the daily vaccination-rate data that is necessary to compute a vineyard, so instead we calculate the PH of $\mathcal{K}$ with the sublevel-set and superlevel-set filtrations from sections 4.1 and 4.2. We show the resulting PDs for the 1D PH in Figure 12. As we described in section 4.1, the points in the PD from the sublevel-set filtration correspond to zip codes in which vaccination rates are higher than in the neighboring zip codes. The death filtration level of a homology class is the vaccination rate in that zip code, and the birth filtration level of a homology class reflects the extent of spatial isolation of that zip code from other local maxima. An earlier birth filtration implies more spatial isolation. Similarly, the points in the superlevel-set filtration PD correspond to zip codes in which the vaccination rates are lower than in the neighboring areas. As we discussed in section 4.1, we obtain the zip code that is associated with a homology class from its death simplex σ_d. We color the points in the PDs by the boroughs of their corresponding zip codes.

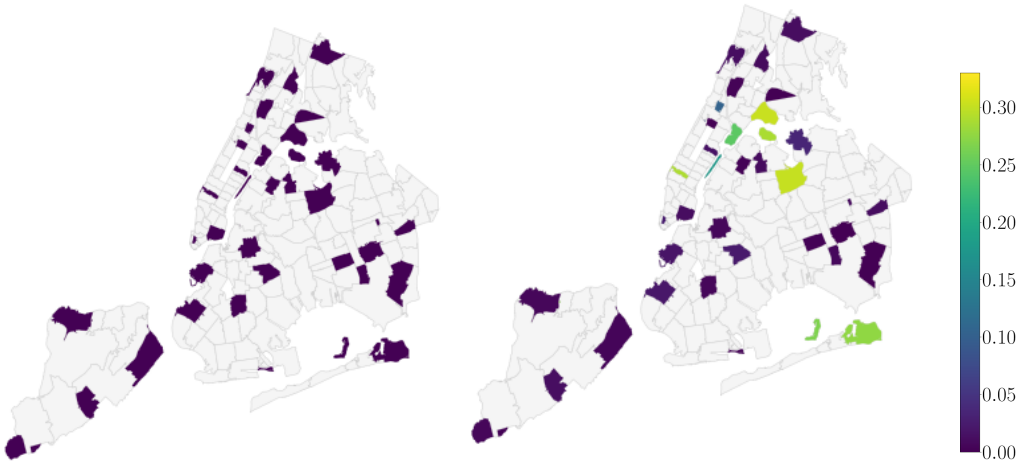
In Figures 13 and 14, we highlight the locations of the maxima and minima, respectively. In Figures 13(a) and 14(a), we color the extrema based on their vaccination rates. In Figure 14(a), we observe that the minima all have near-0 vaccination rates. In Figures 13(b) and 14(b), we color each zip code according to the persistence (i.e., the value death − birth) of its corresponding homology class. These two figure panels incorporate global information about the structure of the extrema, as we described in the paragraph above and in section 4. For example, in Figure 14(b), we observe that some of the minima (specifically, those with the largest values of persistence) are significantly more spatially separated than others,

[15] At the time, the NYC Department of Health and Mental Hygiene defined "fully vaccinated" people to be individuals who either had received both doses of the Pfizer or Moderna vaccine or had received one dose of the Johnson & Johnson vaccine. (This differs from common parlance at that time, in which people were sometimes considered to be "fully vaccinated" only after two weeks had passed since their final dose of a vaccine.)



(a) (b)

Figure 13. *Maps of the local maxima of the NYC vaccination-rate function.* (a) *Color corresponds to the vaccination rate of a zip code.* (b) *Color corresponds to the persistence (i.e., death − birth) of the corresponding homology class.*



(a) (b)

Figure 14. *Maps of the local minima of the NYC vaccination-rate function.* (a) *Color corresponds to the vaccination rate of a zip code.* (b) *Color corresponds to the persistence (i.e., death − birth) of the corresponding homology class.*

even though all of the minima have similar vaccination rates. A larger persistence of a local minimum indicates a greater difference between the vaccination rates of the minimum and those of the neighboring zip codes. A zip code that is a local minimum with a large persistence may have a greater inequity in vaccine access than its neighboring regions. Such insights may be useful for sociologists and policy makers.

An issue arises from the fact that several of the NYC zip codes are islands and thus are isolated. These islands are trivial extrema because they are not adjacent to any other zip codes.

One may wish to exclude these trivial extrema from a PD. In the accompanying supplementary material file supplement.pdf [local/web 779KB], we propose alternative methods for handling disconnected geographical spaces such as NYC.

The PDs in Figure 12 may be helpful for studies of inequities in vaccine access. For example, one may seek to discern patterns in demographic data that correspond to the most-persistent points in the PDs. For interested readers, we provide some demographic data in section SM4 of the accompanying supplementary material file supplement.pdf [local/web 779KB].

5.2. COVID-19 case rates in the city of Los Angeles. We now examine time-dependent COVID-19 case rates in neighborhoods of the city of LA.[16] The geographical boundaries of the neighborhoods are specified by a SHAPEFILE [23]. From the SHAPEFILE, we construct a simplicial complex $\mathcal{K}$ in the manner that we described in section 3. We also know the number of cases in each neighborhood on each day from 25 April 2020 to 25 April 2021. For each neighborhood, we divide the case count by the neighborhood population to obtain per capita case rates, and we calculate a running 14-day mean[17] on each day to smooth the data. For neighborhood R and time $t \in \{0, 1, \ldots, 365\}$, we define $F(t, R)$ to be the 14-day mean per capita case rate in R on day t after 25 April 2020. We compute the vineyard for a simplicial complex $\mathcal{K}$ using the time-dependent sublevel-set filtration that is induced by $F(t, R)$. We show the most important and interesting subsets of our vineyard in Figures 15 and 18. See Figure SM2 of the accompanying supplementary materials for the full vineyard.

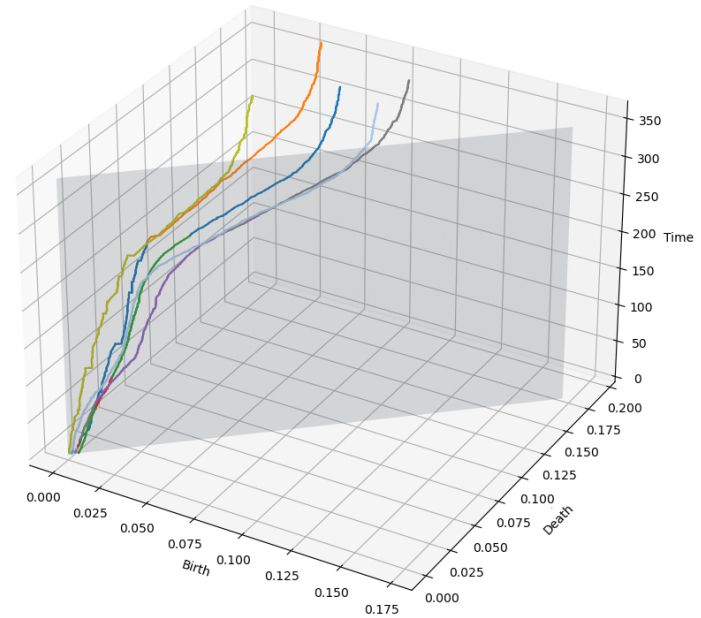
The vines in the vineyard correspond to COVID-19 anomalies, which we define to be neighborhoods that have higher running 14-day mean COVID-19 case rates than the surrounding neighborhoods for at least one day. Anomalies that are more spatially isolated yield vines with earlier birth-filtration levels, and anomalies with high case rates yield vines with late death-filtration levels. See section 4.1 for a detailed discussion. We color each vine according to the geographical location(s) of its anomaly. As we discussed in section 4.3, we obtain the anomaly location(s) from the time-dependent death simplex $\sigma_d(t)$ of a vine. The function $\sigma_d(t)$ is a piecewise-constant function; as it changes, so does the location of the associated anomaly. Therefore, the color of a vine can change with time. For example, consider Figure 15, where we show the five most-persistent vines.[18] The global maximum of the data set is initially in Little Armenia, but it moves to Vermont Square at about $t = 220$. In the vineyard, we see this from the vine that is initially blue (for Little Armenia) from time $t = 0$ until about $t = 220$ and then orange (for Vermont Square) starting from about time $t = 220$ through time $t = 365$. There are also other vines whose locations change with time. Such geographical location changes do not need to be adjacent, but they often are near each other. In Figure 17, we highlight these anomalies on a map.

A vineyard encodes the temporal persistence of anomalies. The length of time that a vine is not on the diagonal plane of a vineyard, which we henceforth call the "length" of a vine, is the amount of time that an anomaly exists in the vineyard. At the beginning of the COVID-19

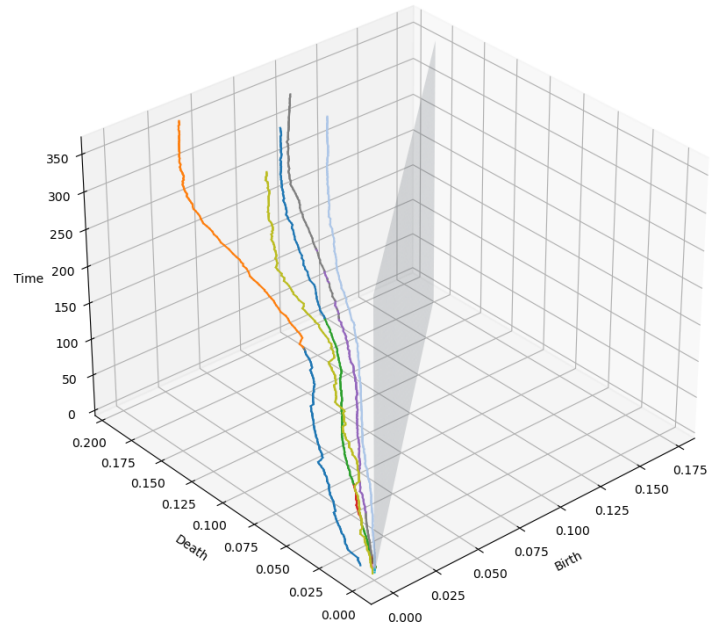
[16] We exclude Angeles National Forest because it has only 20 inhabitants.

[17] On day t, we take the mean of the case rates on days t, $t-1$, $\ldots$, $t-13$. Some outlets (e.g., [43]) report running 14-day means of COVID-19 case counts, and other outlets (e.g., [47]) report 14-day trends.

[18] In section 2.2, we defined the persistence of a vine to be $\int_{t_0}^{T} [f(t, \sigma_d(t)) - f(t, \sigma_b(t))]\, \mathrm{d}t$, where t_0 is the initial time and T is the final time.



(a)



(b)

Figure 15. (a) *The five most-persistent vines of the vineyard for the LA simplicial complex with a sublevel-set filtration from the* 14*-day mean per capita case rate during the period* 25 *April* 2020–25 *April* 2021. *(See Figure* SM2 *of the accompanying supplementary materials for the full vineyard.) Each vine corresponds to a COVID-*19 *anomaly. We color each vine according to the geographical locations of its associated anomaly. Because the geographical location of an anomaly can change with time, a single vine can have multiple colors. (See Figure* 16 *for the legend.)* (b) *A different view of the same five vines.*

Figure 16. *The legend for Figure 15. Each of the depicted regions is a local maximum of the COVID-19 case-rate function for some subset of the time period 25 April 2020–25 April 2021.*

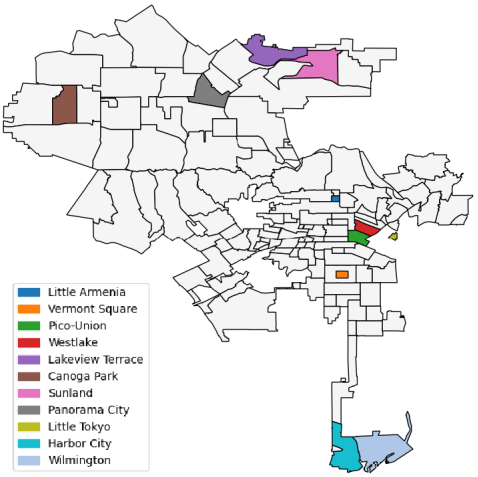
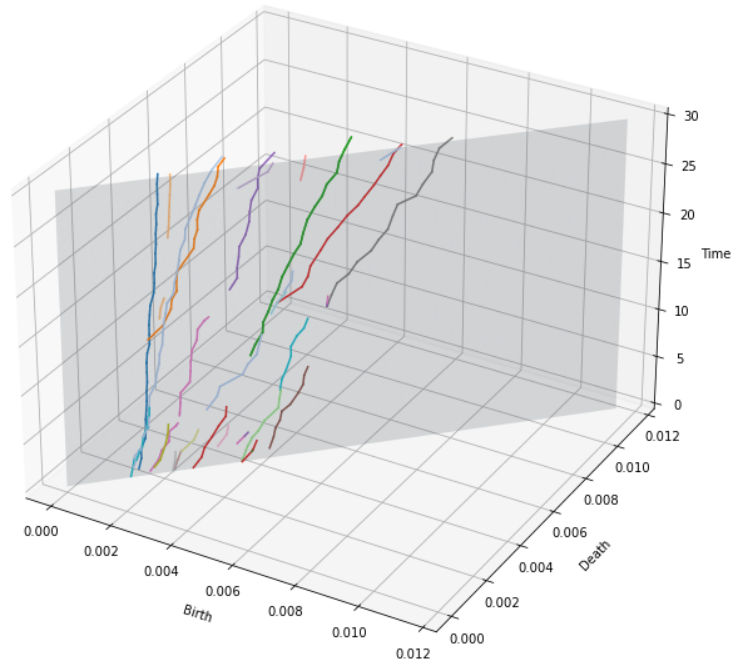


Figure 17. *A map of the most-persistent anomalies of the COVID-19 case-rate function in LA during the time period 25 April 2020–25 April 2021. Each of the highlighted regions is a local maximum of the COVID-19 case-rate function for some subset of the time period.*

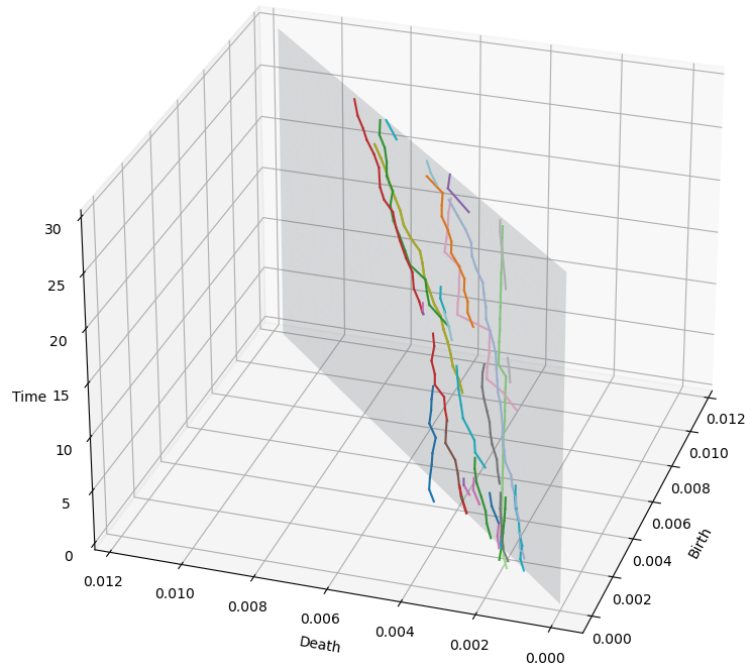
pandemic, all neighborhoods had low per capita case rates. We expect an emerging anomaly to have a low case rate for a long time and then for the case rate to grow rapidly starting at some later time. An emerging anomaly in the "low-case-rate" phase yields a vine that is close to the diagonal for a long time. By examining the lengths of vines, we hypothesize that one can distinguish between concerning emerging anomalies (i.e., those that may become major COVID-19 anomalies in the future) and anomalies of lesser concern, even when the anomalies have similar case rates.

In Figure 18, we show case rates early in the time period that we track (and close to the "beginning"[19] of the COVID-19 pandemic) by computing the vineyard for the period 25 April 2020–25 May 2020. In the depicted vineyard, we exclude the 20 most-persistent vines to more easily see the vines that are close to the diagonal plane. Many of these latter vines are short, so their associated anomalies are short-lived. The longer vines are anomalies that are longer-lived and thus of greater concern in the long run, even though they are close to the diagonal during the period 25 April 2020–25 May 2020. For example, there is an anomaly at Wilmington that we show with the light-blue vine. This vine is close to the diagonal plane, but it has a large temporal persistence during the period 25 April 2020–25 May 2020. In Figure 15,

[19]The COVID-19 pandemic was declared a national emergency in the United States on 13 March 2020 [49], and the city of LA closed its public schools and ordered the closure of restaurants, bars, and gyms on 16 March 2020 [24].



(a)



(b)

Figure 18. (a) *Vineyard for the LA simplicial complex with a sublevel-set filtration for the* 14*-day mean per capita case rate during the period* 25 *April* 2020–25 *May* 2020. *We exclude the* 20 *most-persistent vines to more easily see the vines that are near the diagonal plane. Each vine is associated with a COVID-*19 *anomaly, and we color each vine according to the geographical location(s) of its anomaly. See Figure* 19 *for the legend.* (b) *A different view of the same set of vines.*

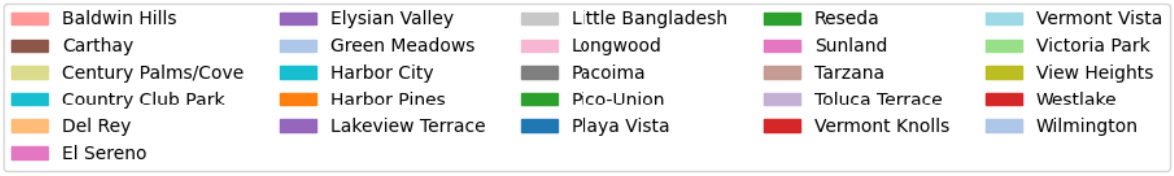


Figure 19. *The legend for Figure 18. Each of the depicted regions is a local maximum of the COVID-19 case-rate function for some subset of the time period 25 April 2020–25 May 2020.*

we see that Wilmington eventually becomes one of the most-persistent anomalies in LA.

6. Discussion. In our approach, we needed to make a variety of choices. There are other ways to construct a simplicial complex to represent a geographical space. There are also other choices of topological tools for analyzing time-varying data. We briefly discuss some of these possibilities in the next several paragraphs.

If one only cares about local information (specifically, the locations and values of the extrema) and not about global information (such as the spatial separation between extrema), then an alternative method to construct a simplicial complex $\mathcal{K}$ is to construct the dual graph of the set S of regions. That is, for each region R (and for each component of any region R with multiple components), there is a vertex $v_R \in \mathcal{K}$, and if regions R_1 and R_2 are adjacent, then there is an edge between v_{R_1} and v_{R_2}. If we wish to study local maxima of a function $F : S \to \mathbb{R}$, then we define the filtration of an edge $e = (v_{R_1}, v_{R_2})$ to be $f(e) = \max\{F(R_1), F(R_2)\}$ and we define the filtration of a vertex v_R to be $f(v_R) = 0$. (There is an analogous definition for studying local minima.) In the 0D PH of the FSC $(\mathcal{K}, f)$, the homology classes correspond to local maxima. If a homology class's birth simplex is the vertex v_R, then R is the corresponding local maximum and $F(R)$ is the death filtration level of the homology class. All of the 0D homology classes are born at 0, so the birth filtration level does not provide any additional information, as it did for our construction in section 3. Consequently, we do not obtain any global information from the PH of $(\mathcal{K}, f)$.

Rasterization gives another method to construct a simplicial complex from SHAPEFILE data. When one rasterizes a SHAPEFILE, one can transform the resulting image into a simplicial complex by imposing the pixels of the image onto a triangulation of the plane. However, our approach has several key advantages over rasterization. First, the number of simplices in the simplicial complex that one obtains by rasterizing a SHAPEFILE is orders-of-magnitude larger than the number of simplices in our construction. Computing the PH of a simplicial complex with fewer simplices allows significantly faster computations. Second, the simplicial complex that one obtains by rasterization has no guarantee of "topological correctness," as property (P) may not hold. The extent to which the resulting simplicial complex is topologically correct depends on the resolution of the rasterization, and using a higher resolution requires more simplices. Our construction of simplicial complexes also yields a natural way to map a 2D simplex to the geographical region that contains it. We use this preservation of geographical information to find the locations of the local extrema. Finally, our construction allows us to detect anomalies on the boundary of a geographical space.

Our construction uses geographical adjacencies, but one may instead wish to employ "effective" distances between regions. One can calculate effective distances using mobility and

transportation data. Two regions that are closely connected via transportation are effectively closer than they are based on direct geographical considerations; this affects phenomena such as the dynamics of infectious diseases [4, 38].

We used only 1D PH to study extrema, but one can alternatively use 0D PH if one is not interested in the geographical locations of the extrema; we discuss this in section SM1.1 of the accompanying supplementary materials. In section SM1.2 of the accompanying supplementary materials, we discuss alternative filtrations that one can apply to geographical spaces (such as NYC) that are disconnected. We used a time-dependent function on a geographical space to compute vineyards, but an alternative is to use an approach that is based on multiparameter PH. In section SM1.3 of the accompanying supplementary materials, we discuss how to do this when the time-dependent function $F(\cdot, R)$ is monotonic for all regions R. When $F(\cdot, R)$ is not monotonic for all R, we discuss how one can use an approach that is based on multiparameter zigzag PH. Both multiparameter PH and multiparameter zigzag PH are difficult to visualize, and they both suffer from a lack of easily interpretable invariants. Consequently, we chose to compute vineyards in our applications.

7. Conclusions. We developed methods to directly incorporate spatial structure into applications of topological data analysis (specifically, of persistent homology) to geospatiotemporal and geospatial data. We defined a way to construct a simplicial complex that efficiently and accurately represents a geographical space. Given a function on a geographical space, we defined filtration functions on a simplicial complex such that the homology classes are in one-to-one correspondence with either local minima or local maxima. By constructing a vineyard, one can track how the local extrema move and change with time.

We conducted case studies using COVID-19 vaccination and case-rate data. In one case study, we examined geospatial vaccination-rate structure in New York City on one day. In our other case study, in which we examined geospatiotemporal data, we constructed a vineyard to examine COVID-19 case-rate anomalies in the city of Los Angeles over the course of one year. From the vineyard, we identified the locations of these anomalies and measured the severity of the associated disease outbreaks. The vineyard also captures information about the relationships between anomalies, such as the extent to which they are separated from each other. We calculated the temporal persistence of each anomaly from the length of its corresponding vine.

There are several ways to build on our research. It is desirable to discover how to use a vineyard to produce systematic forecasts of how a disease (or something else) will spread in space and time. We hypothesized in section 5.2 that one can identify "emerging anomalies" in the COVID-19 case-rate data as vines that are long but close to the diagonal plane. In other applications, one may wish to forecast which locations of local extrema will have the largest data values and/or the largest temporal persistences. One may also wish to forecast how extrema will move in space. It will be valuable to investigate how to use the output of our approach as an input to forecasting models.

Our approach is useful for a wide variety of applications, and it seems possible to generalize it for many others. For example, given spatiotemporal voting data, one can identify regions that vote differently than the neighboring regions. This would allow one to generalize the work of [19] to track the intensity of voting differences and study spatial relationships between

different political islands. Our methodology is not restricted to geographical data. It is applicable whenever one has a surface that is partitioned into a finite number of regions and a real-valued function (or a sequence of real-valued functions) on those regions. For example, it may be possible to apply our approach to grayscale image data by partitioning an image into regions in which pixel values are close to each other. It also seems possible to extend our approach to higher dimensions; this would require constructing a higher-dimensional simplicial complex when one has adjacency information for the higher-dimensional regions. For example, in three dimensions, one can use such an extension of our approach to study atmospheric, oceanic, and video dynamics.

Appendix A. Details of our simplicial-complex construction.

A.1. Boundary-sequence adjustment. Before constructing the polygons with holes P^R for each region R, we adjust the boundary sequences if necessary. The adjustment procedure proceeds as follows. Let $D_0^R, D_1^R, \ldots, D_{h_R}^R$ be the disks in the statement of assumption (A2), let $B_i^R = \partial D_i^R$, and let S_i^R denote the sequences of neighbors around B_i^R. First, we adjust the sequences so that, for each region R and each B_i^R, the first element of S_i^R has a 1D intersection with R. We then adjust the sequences so that $|S_i^R| \geq 3$ for all R and i. When $|S_i^R| < 3$, there are two cases:

1. (Case 1) If $|S_i^R| = 1$, let N be the unique element of S_i^R. This situation occurs if R is an island, and it can also occur if R lies inside N or if N lies inside R. We adjust S_i^R to be the sequence $\{N, N, N\}$. If N is not the exterior region, let j be the index such that B_j^N intersects R. Adjust S_j^N to be the sequence $\{R, R, R\}$ to compensate for the adjustment that we made to S_i^R.
2. (Case 2) If $|S_i^R| = 2$, let N_1 and N_2 be the two elements of S_i^R. If B_i^R intersects R, then R is adjacent to the exterior; without loss of generality, let N_1 denote the exterior region. For example, in Figure 7(a), $S_0^{\text{Little Bangladesh}} = \{\text{Koreatown}, \text{Wilshire Center}\}$. We adjust S_i^R to be the sequence $\{N_1, N_1, N_2\}$. If N_1 is not the exterior region, which occurs if R is not adjacent to the exterior, then we also adjust $S_j^{N_1}$ to compensate, where j is the index of the boundary component of N_1 that intersects R. In this case, we adjust $S_j^{N_1}$ by repeating R an additional time.

A.2. Construction of $\mathcal{K}$ from the Set $\{P^R \mid R \in S\}$. We present two lemmas that we used in section 3 to construct $\mathcal{K}$ by gluing together the set $\{P^R \mid R \in S\}$ of polygons with holes.

Lemma A.1. *Let R_1 and R_2 be connected regions in a set S that satisfies assumptions (A1)–(A4). Let $D_0, \ldots, D_h$ be the disks in the statement of (A2) for R_1. It is then the case that exactly one of the following statements is true:*

1. *$R_2 \subseteq \text{int}(D_0)^c$ and $R_2 \cap \text{int}(D_i) = \emptyset$ for all $i > 0$; or*
2. *there is an $i > 0$ such that R_2 is enclosed in D_i and $R_2 \cap \text{int}(D_j) = \emptyset$ for all $j \neq i$.*

Proof. Because the interiors of R_1 and R_2 do not intersect, it must be true that $\text{int}(R_2) \subseteq$

$\text{int}(D_0)^c \cup (\bigcup_{i=1}^h \text{int}(D_i))$. Therefore,

$$\text{int}(R_2) = \left(\text{int}(D_0)^c \cap \text{int}(R_2)\right) \cup \left(\bigcup_{i=1}^{h} \text{int}(D_i) \cap \text{int}(R_2)\right).$$

The claim follows because $\text{int}(R_2)$ is connected and $\text{int}(D_0)^c, \text{int}(D_1), \ldots, \text{int}(D_h)$ are pairwise disjoint. ∎

Lemma A.2. *Let P^R be the annotated polygon with holes for a connected region R, let v be a vertex in P^R, and let $\{R, N_1, \ldots, N_n\}$ be the sequence of region adjacencies for v. If $n \geq 2$ and $N_1, \ldots, N_n$ are connected, then P^R has at most one other vertex w with the same set of region adjacencies. Additionally, if w exists, its sequence of region adjacencies must be $\{R, N_n, \ldots, N_1\}$, which is the mirror of the orientation of neighbors around v.*

Proof. Suppose that $w \neq v$ is a vertex in P^R with the same set of region adjacencies as v. Let v' and w' denote the points on the boundary of R that correspond, respectively, to v and w. Let R_0 be any connected region that is adjacent to both v' and w', let $D_0, D_1, \ldots, D_h$ denote the disks in the statement of (A2) for R_0, and let $B_i = \partial D_i$. Suppose that v' is in B_i. If $i = 0$, then there is a neighboring region N that is contained entirely in $\text{int}(D_0)^c$ (by Lemma A.1) and adjacent to v'. If $i > 0$, then there is a neighboring region N that is contained entirely in $\text{int}(D_i)$ (by Lemma A.1) and adjacent to v'. In both cases, $w' \in B_i$ because w' is also adjacent to N. Let $B_{i_1}, \ldots, B_{i_m}$ be the disk boundaries that contain v'. As we just showed, it must also be true that $w' \in B_{i_1}, \ldots, B_{i_m}$. If $m > 1$, then $w' \notin B_{i_1} \cap \cdots \cap B_{i_m}$ because $D_{i_1} \cap \cdots \cap D_{i_m}$ is a single point by assumption (A2); this is a contradiction. This argument shows that if v and w have the same set of region adjacencies, then there is a unique B_i that contains v', there is a unique B_j that contains w, and $B_i = B_j$.

Let B be the disk boundary of R that contains v and w. Either the interior of R is contained in the region that is bounded by B or it is contained in the complement of the region that is bounded by B. Without loss of generality, we suppose that the former is true. Let π be the permutation of $\{1, \ldots, n\}$ such that the sequence of region adjacencies around w is $\{R, N_{\pi(1)}, \ldots, N_{\pi(n)}\}$. Let $i_1, i_2 \in \{1, \ldots, n\}$, with $i_1 < i_2$, be a pair of indices. By the argument above (with $R_0 = N_{i_1}$), there is a unique disk boundary B_1 for N_{i_1} that contains v' and w'. Similarly, there is a unique disk boundary B_2 for N_{i_2} that contains v' and w'. We have that $v', w' \in B_1 \cap B_2$.

Because B_1 is homeomorphic to S^1, there exist paths γ_1 and γ_2 from v' to w' such that $\gamma_1 \cup \gamma_2 = B_1$. Because the interior of N_{i_1} does not intersect R, it follows that γ_1 and γ_2 are both in the complement of the region that is bounded by B'. There are two paths from v' to w' on B'. Let τ be the unique choice of path such that R is not contained in the region that is bounded by the closed curve $\tau \cup \gamma_1$. Either γ_1 is in the region that is bounded by the closed curve $\tau \cup \gamma_2$ or γ_2 is in the region that is bounded by the closed curve $\tau \cup \gamma_1$. Without loss of generality, we suppose that the latter is true.

Analogously to our argument above, there exist paths γ_3 and γ_4 from v' to w' such that $\gamma_3 \cup \gamma_4 = B_2$ and γ_3 and γ_4 are in the complement of the region that is bounded by B. Because B_2 is homeomorphic to S^1, the paths γ_3 and γ_4 are either both contained in the region that is bounded by $\gamma_1 \cup \tau$ or both contained in the complement of the region that is bounded by

$\gamma_2 \cup \tau$. Because $i_2 > i_1$, it must be the former case. Therefore, $\pi(i_2) < \pi(i_1)$. It follows that π is order-reversing. If there were another vertex x in B that is adjacent to the same set of regions, then the orientation of those regions around x would be the mirror of both the orientation of regions around v and the orientation of regions around w. This gives a contradiction when $n \geq 2$. ∎

To illustrate Lemma A.2, let R be the region Koreatown in Figure 7(a). The two vertices that are shared by Koreatown and Little Bangladesh have the same region adjacencies, but they have mirrored orientations.

Acknowledgments. We thank Henry Adams, Heather Zinn Brooks, Michelle Feng, Lara Kassab, and Nina Otter for helpful discussions. Additionally, we are grateful to Michelle Feng for teaching us how to work with geospatial data. We thank the Los Angeles County Department of Public Health for providing the LA city data on COVID-19 and the population estimates of LA neighborhoods.

REFERENCES

[1] J. Arino, *Describing, modelling and forecasting the spatial and temporal spread of COVID-19: A short review*, in Mathematics of Public Health, Fields Inst. Commun. 85, V. K. Murty and J. Wu, eds., Springer, Cham, Switzerland, 2022, pp. 25–51.

[2] A. Banman and L. Ziegelmeier, *Mind the gap: A study in global development through persistent homology*, in Research in Computational Topology, Assoc. Women Math. Ser. 13, E. W. Chambers, B. T. Fasy, and L. Ziegelmeier, eds., Springer, Cham, Switzerland, 2018, pp. 125–144.

[3] M. B. Botnan and M. Lesnik, *An Introduction to Multiparameter Persistence*, arXiv:2203.14289, 2022.

[4] D. Brockmann and D. Helbing, *The hidden geometry of complex, network-driven contagion phenomena*, Science, 342 (2013), pp. 1337–1342.

[5] M. Buchet, Y. Hiraoka, and I. Obayashi, *Persistent homology and materials informatics*, in Nanoinformatics, I. Tanaka, ed., Springer, Singapore, 2018, pp. 75–95.

[6] G. Carlson, *Topological methods for data modelling*, Nature Reviews Physics, 2 (2020), pp. 697–707.

[7] G. Carlsson and V. de Silva, *Zigzag persistence*, Found. Comput. Math., 10 (2010), pp. 367–405.

[8] G. Carlsson and A. Zomorodian, *The theory of multidimensional persistence*, Discrete Comput. Geom., 42 (2007), pp. 71–93.

[9] Centers for Disease Control and Prevention, *Risk for COVID-19 Infection, Hospitalization, and Death by Race/Ethnicity*, https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html (accessed 10 July 2022).

[10] Y. Chun and D. A. Griffith, *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*, SAGE, Thousand Oaks, CA, 2013.

[11] City of New York, *COVID-19: Data on Vaccines—NYC Health*, https://www1.nyc.gov/site/doh/covid/covid-19-data-vaccines.page (accessed 23 February 2021).

[12] D. Cohen-Steiner, H. Edelsbrunner, and D. Morozov, *Vines and vineyards by updating persistence in linear time*, in Proceedings of the Annual ACM Symposium on Computational Geometry, ACM, 2006, pp. 119–126.

[13] P. Corcoran and C. B. Jones, *Modelling topological features of swarm behaviour in space and time with persistence landscapes*, IEEE Access, 5 (2017), pp. 18534–18544.

[14] V. de Silva and R. Ghrist, *Coverage in sensor networks via persistent homology*, Algebr. Geom. Topol., 7 (2007), pp. 339–358.

[15] T. K. Dey and Y. Wang, *Computational Topology for Data Analysis*, Cambridge University Press, Cambridge, UK, 2022, https://www.cs.purdue.edu/homes/tamaldey/book/CTDAbook/CTDAbook.html.

[16] H. Edelsbrunner and J. Harer, *Computational Topology: An Introduction*, AMS, Providence, RI,

2010.
[17] M. Feng, A. Hickok, and M. A. Porter, *Topological data analysis of spatial systems*, in Higher-Order Systems, F. Battiston and G. Petri, eds., Springer, Cham, Switzerland, 2022, pp. 389–399.
[18] M. Feng and M. A. Porter, *Spatial applications of topological data analysis: Cities, snowflakes, random structures, and spiders spinning under the influence*, Phys. Rev. Res., 2 (2020), 033426.
[19] M. Feng and M. A. Porter, *Persistent homology of geospatial data: A case study with voting*, SIAM Review, 63 (2021), pp. 67–99.
[20] C. Giusti, R. Ghrist, and D. S. Bassett, *Two's company, three (or more) is a simplex*, J. Comput. Neurosci., 41 (2016), pp. 1–14.
[21] S. Hazarie, D. Soriano-Paños, A. Arenas, J. Gómez-Gardeñes, and G. Ghoshal, *Interplay between population density and mobility in determining the spread of epidemics in cities*, Commun. Phys., 4 (2021), 191.
[22] X. Hou, S. Gao, Q. Li, Y. Kang, N. Chen, K. Chen, J. Rao, J. S. Ellenberg, and J. A. Patz, *Intracounty modeling of COVID-19 infection with human mobility: Assessing spatial heterogeneity with business traffic, age, and race*, Proc. Nat. Acad. Sci. USA, 118 (2021), e2020524118.
[23] P. Jula, *COVID19 by Neighborhood*, City of Los Angeles Hub, https://geohub.lacity.org/datasets/covid19-by-neighborhood/about (accessed 3 June 2020).
[24] J. Kandel, *Timeline: A Look at Key Coronavirus Pandemic Events and Milestones in California*, NBC Los Angeles, 2021, https://www.nbclosangeles.com/news/coronavirus/2020-2021-california-coronavirus-pandemic-timeline-key-events/2334100 (accessed 10 July 2022).
[25] W. Kim and F. Mémoli, *Spatiotemporal persistent homology for dynamic metric spaces*, Discrete Comput. Geom., 66 (2021), pp. 831–875.
[26] B. L. Levy, K. Vachuska, S. V. Subramanian, and R. J. Sampson, *Neighborhood socioeconomic inequality based on everyday mobility predicts COVID-19 infection in San Francisco, Seattle, and Wisconsin*, Science Advances, 8 (2022), eabl3825.
[27] Y. Li, D. Wang, G. A. Ascoli, P. Mitra, and Y. Wang, *Metrics for comparing neuronal tree shapes based on persistent homology*, PLoS ONE, 12 (2017), e0182184.
[28] S. Maletić, Y. Zhao, and M. Rajković, *Persistent topological features of dynamical systems*, Chaos, 26 (2016), 053105.
[29] S. Martin, A. Thompson, E. A. Coutsias, and J.-P. Watson, *Topology of cyclo-octane energy landscape*, J. Chem. Phys., 132 (2010), 234115.
[30] NYC By Natives, *New York City Zip Codes*, https://www.nycbynatives.com/nyc_info/new_york_city_zip_codes.php (accessed 30 March 2021).
[31] NYC Department of Health and Mental Hygiene, *ZCTA vs MODZCTA*, https://github.com/nychealth/coronavirus-data/issues/64 (accessed 10 July 2022).
[32] NYC Open Data, *Modified Zip Code Tabulation Areas (MODZCTA)*, https://data.cityofnewyork.us/Health/Modified-Zip-Code-Tabulation-Areas-MODZCTA-/pri4-ifjk/data (accessed 23 February 2021).
[33] S. J. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Appl. Math. Sci. 153, Springer-Verlag, Heidelberg, 2003.
[34] A. M. Oster, G. J. Kang, A. E. Cha, V. Beresovsky, C. E. Rose, G. Rainisch, L. Porter, E. E. Valverde, E. B. Peterson, A. K. Driscoll, T. Norris, N. Wilson, M. Ritchey, H. T. Walke, D. A. Rose, N. L. Oussayef, M. E. Parise, Z. S. Moore, A. T. Fleischauer, M. A. Honein, E. Dirlikov, and J. Villanueva, *Trends in number and distribution of COVID-19 hotspot counties—United States, March 8–July 15, 2020*, Morbidity Mortality Weekly Report, 69 (2020), pp. 1127–1132.
[35] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington, *A roadmap for the computation of persistent homology*, Eur. Phys. J. Data Sci., 6 (2017), 17.
[36] M. A. Porter and J. P. Gleeson, *Dynamical Systems on Networks: A Tutorial*, Front. Appl. Dyn. Syst. 4, Springer, Cham, Switzerland, 2016.
[37] J. O'Rourke, Holes, in *Art Gallery Theorems and Algorithms*, Oxford University Press, Oxford, UK, 1987, pp. 125–145.
[38] O. Sadekar, M. Budamagunta, G. J. Sreejith, S. Jain, and M. S. Santhanam, *An infectious diseases hazard map for India based on mobility and transportation networks*, Current Science, 121 (2021), pp. 1208–1215.

[39] I. Segovia-Dominguez, Z. Zhen, R. Wagh, H. Lee, and Y. R. Gel, *TLife-LSTM: Forecasting future COVID-19 progression with topological signatures of atmospheric conditions*, in Advances in Knowledge Discovery and Data Mining, K. Karlapalem, H. Cheng, N. Ramakrishnan, R. K. Agrawal, P. K. Reddy, J. Srivastava, and T. Chakraborty, eds., Springer, Cham, Switzerland, 2021, pp. 201–212.

[40] A. E. Sizemore, J. E. Phillips-Cremins, R. Ghrist, and D. S. Bassett, *The importance of the whole: Topological data analysis for the network neuroscientist*, Network Neuroscience, 3 (2019), pp. 656–673.

[41] M. Soliman, V. Vyacheslav Lyubchich, and Y. R. Gel, *Ensemble forecasting of the Zika space-time spread with topological data analysis*, Environmetrics, 31 (2020), e2629.

[42] L. Speidel, H. A. Harrington, S. J. Chapman, and M. A. Porter, *Topological data analysis of continuum percolation with disks*, Phys. Rev. E, 98 (2018), 012318.

[43] STAT, *The Covid-19 Tracker*, https://www.statnews.com/feature/coronavirus/covid-19-tracker/ (accessed 16 June 2021).

[44] B. J. Stolz, H. A. Harrington, and M. A. Porter, *The Topological "Shape" of Brexit*, arXiv:1610.00752, 2016.

[45] D. Taylor, F. Klimm, H. A. Harrington, M. Kramár, K. Mischaikow, M. A. Porter, and P. J. Mucha, *Topological data analysis of contagion maps for examining spreading processes on networks*, Nature Commun., 6 (2015), 7723.

[46] The Gapminder Foundation, *Gapminder World*, http://www.gapminder.com/world (accessed 10 July 2022).

[47] The New York Times, *Coronavirus in the U.S.: Latest Map and Case Count*, https://www.nytimes.com/interactive/2021/us/covid-cases.html?action=click&module=Top%20Stories&pgtype=Homepage (accessed 10 July 2022).

[48] C. M. Topaz, L. Ziegelmeier, and T. Halverson, *Topological data analysis of biological aggregation models*, PLoS ONE, 10 (2015), e0126383.

[49] D. Trump, *Proclamation on Declaring a National Emergency Concerning the Novel Coronavirus Disease (COVID-19) Outbreak*, https://trumpwhitehouse.archives.gov/presidential-actions/proclamation-declaring-national-emergency-concerning-novel-coronavirus-disease-covid-19-outbreak/ (accessed 10 July 2022).

[50] S. Tymochko, E. Munch, and F. A. Khasawneh, *Using zigzag persistent homology to detect Hopf bifurcations in dynamical systems*, Algorithms, 13 (2020), 278.

[51] A. Vespignani, H. Tian, C. Dye, J. O. Lloyd-Smith, R. M. Eggo, M. Shrestha, S. V. Scarpino, B. Gutierrez, M. U. G. Kraemer, J. Wu, K. Leung, and G. M. Leung, *Modelling COVID-19*, Nature Reviews Physics, 2 (2020), pp. 279–281.

[52] L. Xian, H. Adams, C. M. Topaz, and L. Ziegelmeier, *Capturing dynamics of time-varying data via topology*, Found. Data Sci., 4 (2021), pp. 1–36.

[53] G. Yalniz and N. B. Budanur, *Inferring symbolic dynamics of chaotic flows from persistence*, Chaos, 30 (2020), 033109.

SUPPLEMENTARY MATERIALS: Analysis of Spatial and Spatiotemporal Anomalies Using Persistent Homology: Case Studies with COVID-19 Data*

Abigail Hickok†, Deanna Needell†, and Mason A. Porter‡

SM1. Alternative topological approaches.

SM1.1. 0D persistent homology. We do not compute 0D PH in the present paper. However, it is appropriate to use 0D PH to study the structure of local extrema when one is not interested in their geographical locations.

Let F be a real-valued function on a set S of geographical regions. In the main manuscript, we described how one can analyze the local maxima (respectively, local minima) of F by computing the 1D PH of the sublevel-set filtration (respectively, superlevel-set filtration). See sections 4.1 and 4.2 of the main manuscript for more details. We now discuss how the 0D PH of the sublevel-set filtration (respectively, superlevel-set filtration) yields information about local minima (respectively, local maxima) of F.

The 0D PH of the sublevel-set filtration encodes information about the structure of local minima of F in a way that is similar to how 1D PH encodes information about the structure of local maxima. One can imagine taking α-sublevel sets of the function in Figure 9 of the main manuscript (where we showed α-*super*level sets) to see why this is true. A region R is a *local minimum* if the value $F(R)$ is less than the value $F(N)$ in all neighboring regions N of R for which $N \cap R$ is 1D. If R is a local minimum, there is a 0D homology class whose birth simplex is one of the vertices in one of the triangles in the preimage $g^{-1}(R)$. The class is born at filtration level $\alpha = F(R)$. For the LA data set of COVID-19 case rates, 0D homology classes correspond to regions that have a lower case rate than neighboring regions. The smaller the value $F(R)$ in comparison to the values of F in the neighboring regions, the more persistent the homology class is. There is also one infinite 0D homology class for each connected component. One can think of these classes as corresponding to a "local minimum" in the exterior region. However, unlike for 1D homology classes, there is no canonical map from 0D homology classes to regions because the birth simplex of a 0D class is a vertex that belongs to several regions. Analogously, the 0D PH of the superlevel-set filtration encodes information about the structure of local maxima of F. However, as with a sublevel-set filtration, there is no canonical map from 0D homology classes to regions. Therefore, one cannot easily use the 0D PH of the sublevel-set filtration (respectively, superlevel-set filtration) to identify the geographical locations of the local minima (respectively, local maxima), so we did not examine 0D PH in our case studies.

*Supplementary material for SIMODS MS#M143503.
https://doi.org/10.1137/21M1435033

†Department of Mathematics, University of California, Los Angeles, CA 90095 USA (ahickok@math.ucla.edu, deanna@math.ucla.edu).

‡Department of Mathematics, University of California, Los Angeles, CA 90095 USA and Santa Fe Institute, Santa Fe, NM 87501 USA (mason@math.ucla.edu).

SM1.2. Alternative filtrations for disconnected geographical spaces. In section 4.1 (respectively, section 4.2) of the main manuscript, we defined a sublevel-set filtration (respectively, superlevel-set filtration) in which we set the filtration values of all exterior-adjacent vertices and edges to the global minimum (respectively, to the additive inverse of the global maximum) of F. In applications in which the union of all regions is not connected, such as for the NYC zip codes in section 5.1 of the main manuscript, an alternative definition is to consider extrema on each connected component separately, rather than on the entire geographical space at once. This solves the problem that an isolated region (i.e., a geographical island[1]) is trivially both a local maximum and a local minimum because it is not adjacent to any other regions. In Definitions 4.1 and 4.2 of the main manuscript, they appear as 1D homology classes that are born at the earliest filtration time; this may falsely emphasize the persistence of these trivial extrema.

Definition SM1.1 (alternative sublevel-set filtration). *Let $\mathcal{K}$ be the simplicial complex from section 3 of the main manuscript for a set S of regions, and let g be the assignment of $2D$ simplices to regions. Additionally, let $F : S \to \mathbb{R}$. If σ is a vertex or edge on the boundary of $\mathcal{K}$, let $\tilde{\sigma}$ be the $2D$ simplex with σ on the boundary of $\tilde{\sigma}$. On σ, we define the alternative sublevel-set filtration function f to be*

$$f(\sigma) = \min_R \{F(R) \mid R \subseteq C\}\,,$$

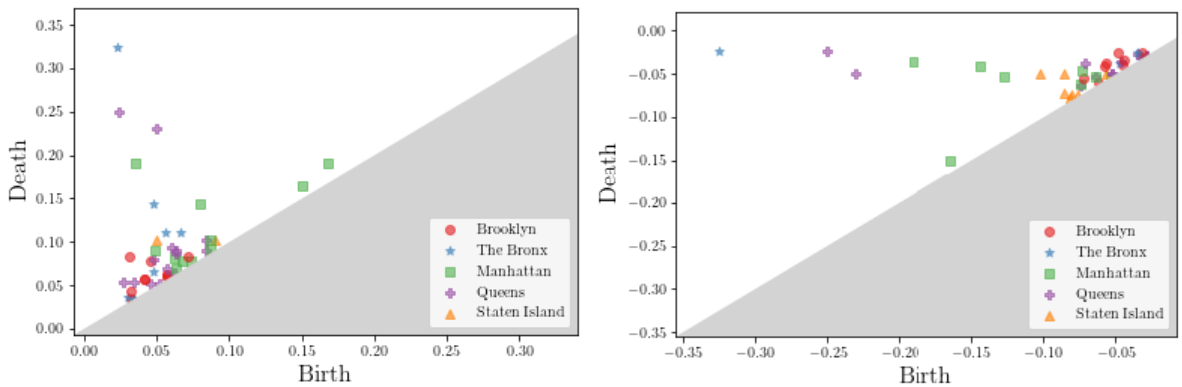
where C is the connected component that contains the region $g(\tilde{\sigma})$. On all other simplices, the filtration function f is equal to the sublevel-set filtration function.

Definition SM1.2 (alternative superlevel-set filtration). *Let $F : S \to \mathbb{R}$ for a set S of regions. The alternative superlevel-set filtration function f is the alternative sublevel-set filtration function that is induced by $-F$.*

Definitions SM1.1 and SM1.2 are appropriate options if one seeks to treat each connected component independently. In these alternative definitions, each connected component uses only information about other regions in the same component. One then compares region values $F(R)$ to global extremum values on their connected components. One consequence of using these definitions is that one ignores isolated regions, which are trivial extrema. In Definitions SM1.1 and SM1.2, these isolated extrema appear as points on the diagonal of a PD. This is often an appropriate way to handle isolated regions. However, when an isolated region is a global extremum of a data set, this may be undesirable. This situation never occurs in our data.

NYC has 14 connected components; several of them are zip codes that correspond to isolated islands. The alternative sublevel-set and superlevel-set filtrations effectively treat each connected component of NYC separately. In Figure SM1, we show the PDs that we compute using the alternative sublevel-set and superlevel-set filtrations that are induced by the vaccination-rate function that we defined in section 5.1 of the main manuscript. In these PDs, we compare a zip code's per capita vaccination rate to the global minimum or maximum rate on its connected component, rather than to the global minimum or maximum rate in all of NYC. More precisely, the birth time of a connected component's global extremum is

[1]These are literal islands, rather than "islands" from a PH computation.



(a) Alternative sublevel-set filtration (b) Alternative superlevel-set filtration

Figure SM1. *PDs for the 1D PH of the NYC simplicial complex with filtrations that are induced by the per capita full-vaccination rate by zip code on 23 February 2021. We show only the finite homology classes. Each point in a PD corresponds to a non-isolated zip code, which we label according to its borough [SM5], that has* (a) *a higher vaccination rate than its neighboring zip codes or* (b) *a lower vaccination rate than its neighboring zip codes.*

either the lowest per capita vaccination rate of that component (for the alternative sublevel-set filtration) or the additive inverse of the highest per capita vaccination rate of that component (for the alternative superlevel-set filtration). Consequently, the trivial island extrema yield homology classes on the diagonal of a PD.

The alternative sublevel-set filtration and the alternative superlevel-set filtration, along with time-dependent versions of them, are implemented in our code at https://bitbucket.org/ahickok/vineyard/src/main/.

SM1.3. Multiparameter persistent homology. One can use multiparameter persistent homology (MPH) to study how the topology of a data set changes as one varies multiple parameters. For extensive discussions of MPH, see [SM1, SM3].

One can use MPH to study local extrema of functions that are nondecreasing with time. To apply MPH to our COVID-19 case-rate data, two feasible parameters are (1) time and (2) the cumulative COVID-19 case rate. However, MPH is difficult to analyze. Although there are invariants (e.g., the rank invariant), there is no complete discrete invariant [SM3]. By contrast, one can use PDs for single-parameter PH.

Definition SM1.3. *Let $\mathcal{K}$ be the simplicial complex from the construction in section 3 of the main manuscript for a set S of regions. Let $F : \{t_0, \ldots, t_n\} \times S \to \mathbb{R}$ be a function such that $F(t, R) \geq F(s, R)$ for all $t \geq s$. Define the function $f(t_i, \sigma)$ to be the sublevel-set filtration that is induced by $F(t_i, \cdot)$. Let $\{\alpha_0, \ldots, \alpha_\ell\}$ be the image of F, where $\ell + 1$ is the number of elements in the image. We define the bifiltration*

$$\mathcal{K}_{i,j} := \begin{cases} \{\sigma \in \mathcal{K} \mid f(t_i, \sigma) \leq \alpha_j\}, & i \in \{0, \ldots, n\},\ j \in \{0, \ldots, \ell\} \\ \mathcal{K}, & j > \ell \text{ and } i \geq 0 \\ \mathcal{K}_{n,j}, & i > n \text{ and } j \geq 0 \\ \emptyset, & i < 0 \text{ or } j < 0. \end{cases}$$

One can use Definition SM1.3 to study cumulative COVID-19 case rates as a function of time.

SM1.4. Multiparameter zigzag persistent homology. One can use *multiparameter zigzag PH* (MZPH) to study how the topology of a data set changes as one varies multiple parameters nonmonotonically. See section 2.1 of [SM2] for a short discussion of MZPH.

To use MZPH to study our COVID-19 case-rate data, two feasible parameters are (1) time and (2) the current COVID-19 case rate. A diagram of simplicial complexes, such as the one in Equation SM1.1, induces a diagram of homology groups. This is a representation of a quiver. However, there are no known well-behaved statistical summaries (in contrast to single-parameter zigzag PH).

Definition SM1.4. *Let $\mathcal{K}$ be the simplicial complex from the construction in section 3 of the main manuscript for a set S of regions, and suppose that $F : \{t_0, \dots, t_n\} \times S \to \mathbb{R}$. Define half steps $t_{i+1/2} := t_i + (t_{i+1} - t_i)/2$ for $i \in \{0, \dots, m-1\}$, and let $s_i := t_{i/2}$. Define the function $G : \{s_0, \dots, s_{2n}\} \times S \to \mathbb{R}$ as follows:*

$$G(s_i, R) = \begin{cases} F(s_i, R)\,, & i \text{ is even} \\ \max\{F(t_{(i-1)/2}, R), F(t_{(i+1)/2}, R)\}\,, & i \text{ is odd}. \end{cases}$$

We define the function $h(s_i, \cdot)$ to be the sublevel-set filtration that is induced by $G(s_i, \cdot)$. Let $\{\alpha_0, \dots, \alpha_\ell\}$ be the image of G. We define

$$\mathcal{K}_{i,j} := \begin{cases} \{\sigma \in \mathcal{K} \mid h(s_i, \sigma) \leq \alpha_j\}\,, & i \in \{0, \dots, 2n\}\,,\ j \in \{0, \dots, \ell\} \\ \mathcal{K}\,, & j > \ell \text{ and } i \geq 0 \\ \mathcal{K}_{2n,j}\,, & i > 2n \text{ and } j \geq 0 \\ \emptyset\,, & i < 0 \text{ or } j < 0\,. \end{cases}$$

This yields the following diagram:

(SM1.1)

$$\begin{array}{ccccccc}
\uparrow & & \uparrow & & \uparrow & & \uparrow \\
K_{\alpha_0,s_3} & \hookrightarrow & K_{\alpha_1,s_3} & \hookrightarrow & K_{\alpha_2,s_3} & \hookrightarrow & K_{\alpha_3,s_3} \hookrightarrow \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
K_{\alpha_0,s_2} & \hookrightarrow & K_{\alpha_1,s_2} & \hookrightarrow & K_{\alpha_2,s_2} & \hookrightarrow & K_{\alpha_3,s_2} \hookrightarrow \\
\uparrow & & \uparrow & & \uparrow & & \uparrow \\
K_{\alpha_0,s_1} & \hookrightarrow & K_{\alpha_1,s_1} & \hookrightarrow & K_{\alpha_2,s_1} & \hookrightarrow & K_{\alpha_3,s_1} \hookrightarrow \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
K_{\alpha_0,s_0} & \hookrightarrow & K_{\alpha_1,s_0} & \hookrightarrow & K_{\alpha_2,s_0} & \hookrightarrow & K_{\alpha_3,s_0} \hookrightarrow
\end{array} .$$

The inclusion maps induce a corresponding diagram of homology groups.

One can use Definition SM1.4 to study non-cumulative COVID-19 case rates as a function of time.

SM2. The full LA vineyard. In Figure SM2, we show the full LA vineyard that we discussed in section 5.2 of the main manuscript.

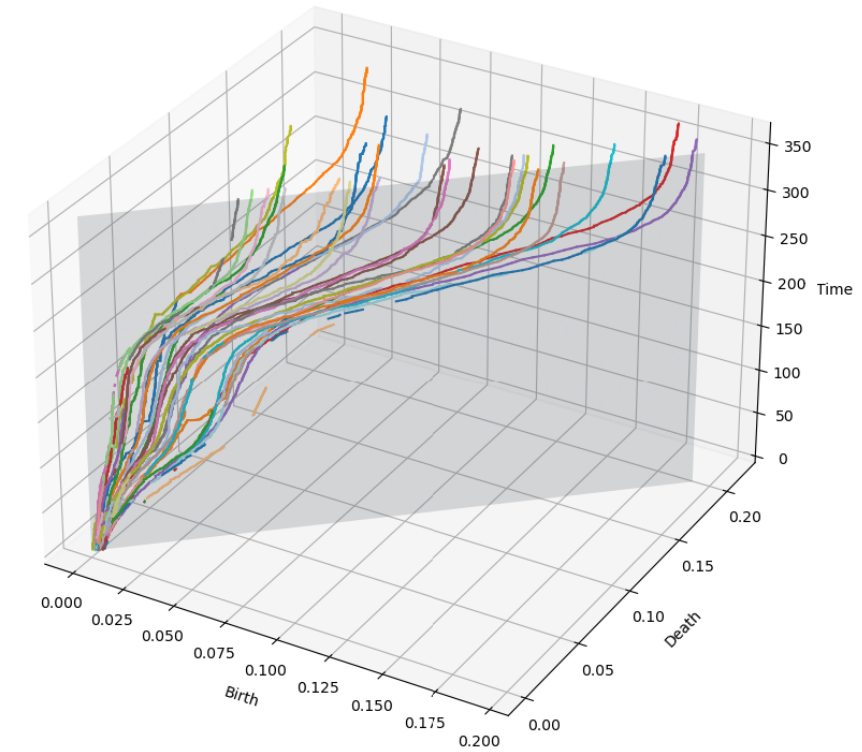
SM3. Results of an all-but-one statistical test. In the main manuscript, we examined local extrema of real-valued geospatial data; we called these "anomalies." For real-valued geospatiotemporal data, one can alternatively examine a different notion of anomaly. In this context, we say that a region is an *anomaly* if one is not able to infer its data successfully from the data of the other regions. More precisely, let X be the matrix whose (i,j)th entry is the value of a function in region j at time step i. In our case study of COVID-19 case rates in LA, the regions are the neighborhoods of LA and the (i,j)th entry of X is the 14-day mean per capita case rate in region j on the ith day after 25 April 2020. Let $\boldsymbol{x}^j$ denote the jth column of X, and let X^j denote the matrix that one obtains by deleting column $\boldsymbol{x}^j$. The vector $\boldsymbol{x}^j$ has the data for region j, and the matrix X^j has the data for all regions except for region j. We define our *prediction* of region j to be the least-squares solution $\boldsymbol{b}^*$ to $X^j\boldsymbol{b} = \boldsymbol{x}^j$, and we quantify the predictability of region j by calculating the relative residual norm $\left\|X^j\boldsymbol{b}^* - \boldsymbol{x}^j\right\|_2 / \left\|\boldsymbol{x}^j\right\|_2$. A smaller relative residual norm indicates greater predictability.

In Figure SM4, we show the result of this "all-but-one" statistical test for the LA COVID-19 data set. In this figure, we plot the relative residual norm for each neighborhood. All neighborhoods have near-0 relative residual norms, so the neighborhoods' case rates are very predictable when one knows the case rates of all other neighborhoods. The mean relative residual norm is only 5.970×10^{-7}, with a standard deviation of $\sigma \approx 7.558 \times 10^{-7}$. The neighborhoods that are least predictable (specifically, the ones whose relative residual norms have a z-score that is larger than 3) are Brookside, Little Armenia, Little Tokyo, Sycamore Square, and Toluca Terrace. We show their relative residual norms and z-scores in Table SM1.

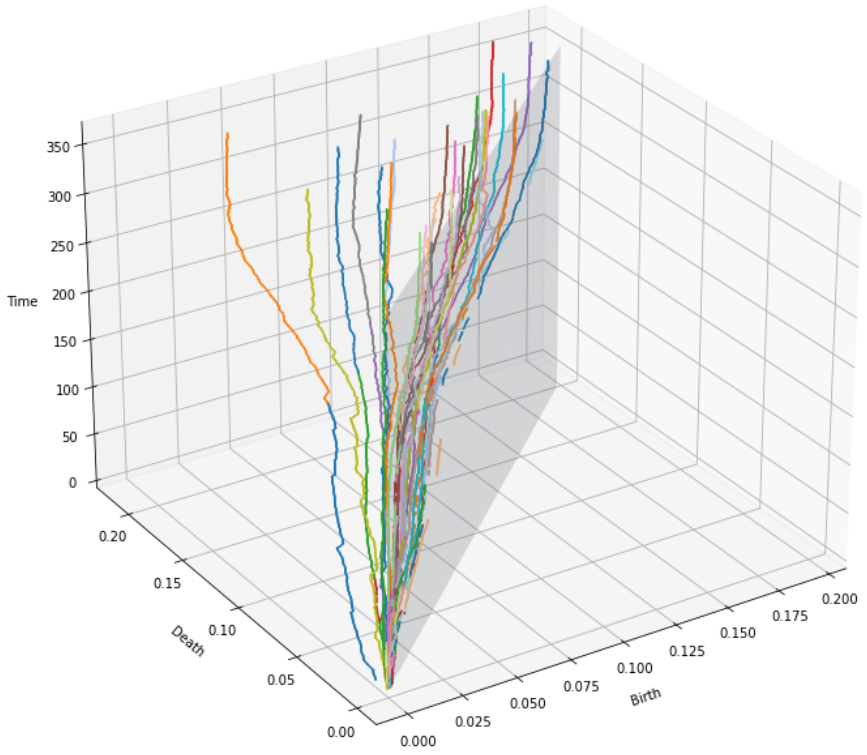
The difference between what we learn from the all-but-one statistical test and what we learn from our TDA approach is the following. Using our TDA approach, we identified local extrema (i.e., regions whose associated values are either all larger than or all smaller than those of all neighboring regions); this is a geographical notion of anomaly. By contrast, the all-but-one statistical test does not inherently capture local extrema because the test does not consider geographical adjacencies. Despite this conceptual difference, we observe some overlap in the anomalies that the two approaches identify. For example, the neighborhoods Little Tokyo and Little Armenia are identified as anomalies by both approaches. For further examples, compare Figure SM4 with Figure 17 in the main manuscript.

SM4. Demographic data. We provide some demographic data for NYC and LA for readers who are interested in comparing patterns in the PDs and demographic data, although an investigation of such patterns is beyond the scope of the present paper. In Figure SM5, we plot the median household income for each zip code[2] [SM7]. The geographical boundaries of the NYC and LA zip codes are specified by the SHAPEFILEs [SM6] and [SM4], respectively. It is worthwhile to examine and compare other demographic data (such as racial, religious, and political data) to the PDs.

[2]We do not possess median income data for LA zip codes 90073, 90089, 90095, 91330, 91522, and 91608. These zip codes are in non-residential areas. For example, 90073 corresponds to the Veterans Administration.



(a)



(b)

Figure SM2. (a) *The vineyard for the LA simplicial complex that we construct using the sublevel-set filtration from the* 14*-day mean per capita case rate during the period* 25 *April* 2020–25 *April* 2021*. Each vine is associated with a COVID-*19 *anomaly. We color each vine according to the geographical location(s) of its associated anomaly. (See Figure* SM3 *for the legend.) Because the geographical location of an anomaly can change with time, a single vine can have multiple colors.* (b) *A different view of the same vineyard.*

Figure SM3. *The legend for Figure* SM2. *Each of the depicted regions is a local maximum of the COVID-19 case-rate function for some subset of the time period* 25 *April* 2020–25 *April* 2021.

Table SM1

The relative residual norms and z-scores for the LA neighborhoods that are least predictable according to our all-but-one test.

Neighborhood	Relative Residual Norm	z-score
Brookside	3.973×10^{-6}	4.466
Little Armenia	3.220×10^{-6}	3.471
Little Tokyo	3.944×10^{-6}	4.429
Sycamore Square	3.944×10^{-6}	4.429
Toluca Terrace	2.873×10^{-6}	3.012

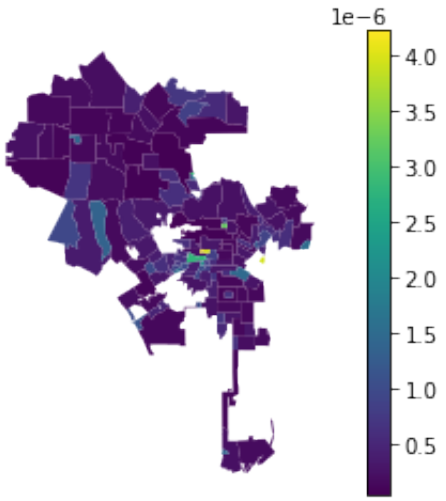
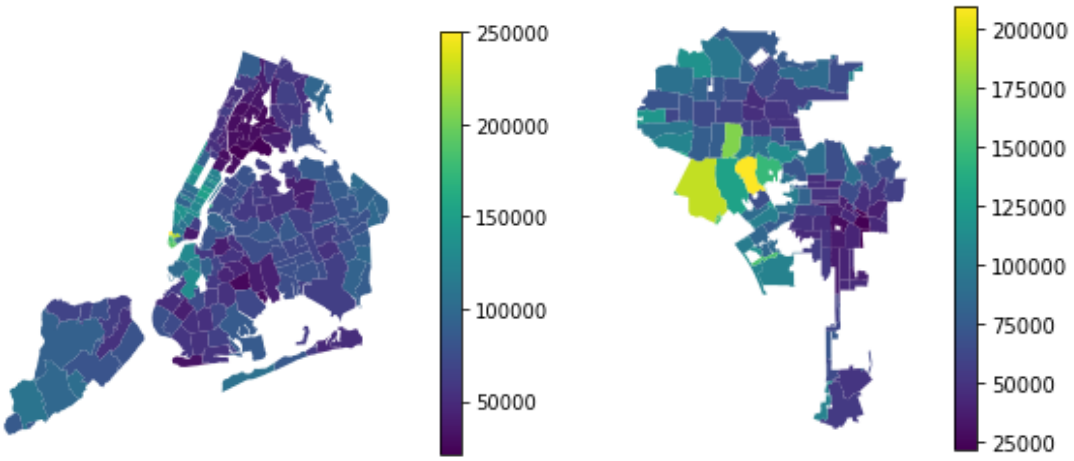


Figure SM4. *The results of an all-but-one statistical test for the LA COVID-19 case-rate data. We plot the relative residual norm for each neighborhood.*



(a) (b)

Figure SM5. (a) *Median household income (in U.S. dollars) by zip code in NYC.* (b) *Median household income by zip code in LA.*

REFERENCES

[1] M. B. Botnan and M. Lesnik, *An Introduction to Multiparameter Persistence*, arXiv:2203.14289, 2022.
[2] G. Carlsson and V. de Silva, *Zigzag persistence*, Found. Comput. Math., 10 (2010), pp. 367–405.
[3] G. Carlsson and A. Zomorodian, *The theory of multidimensional persistence*, Discrete Comput. Geom., 42 (2007), pp. 71–93.
[4] L. Coral, *Los Angeles City zip codes*, https://geohub.lacity.org/datasets/lahub::los-angeles-city-zip-codes/about (accessed 4 April 2020).
[5] NYC By Natives, *New York City Zip Codes*, https://www.nycbynatives.com/nyc_info/new_york_city_zip_codes.php (accessed 30 March 2021).
[6] NYC Open Data, *Modified Zip Code Tabulation Areas (MODZCTA)*, https://data.cityofnewyork.us/Health/Modified-Zip-Code-Tabulation-Areas-MODZCTA-/pri4-ifjk/data (accessed 23 February 2021).
[7] U.S. Census Bureau, *American Community Survey* 5-*Year Estimates* (2019), https://data.census.gov/cedsci/ (accessed 16 July 2022).