Computational Statistics and Data Analysis ••• (••••) ••••••

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Topology-based goodness-of-fit tests for sliced spatial data

Alessandra Cipriani [a], Christian Hirsch [b,c,d,1], Martina Vittorietti [e]

[a] Department of Statistical Science, UCL, United Kingdom of Great Britain and Northern Ireland
[b] Department of Mathematics, Aarhus University, Denmark
[c] Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Netherlands
[d] CogniGron (Groningen Cognitive Systems and Materials Center), University of Groningen, Netherlands
[e] Department of Economics, Business and Statistics, University of Palermo, Italy

ARTICLE INFO

Article history:
Received 17 January 2022
Received in revised form 3 November 2022
Accepted 3 November 2022
Available online xxxx

Keywords:
Topological data analysis
Persistence diagram
Materials science
Vineyards
Goodness-of-fit tests
Asymptotic normality

ABSTRACT

In materials science and many other application domains, 3D information can often only be obtained by extrapolating from 2D slices. In topological data analysis, persistence vineyards have emerged as a powerful tool to take into account topological features stretching over several slices. It is illustrated how persistence vineyards can be used to design rigorous statistical hypothesis tests for 3D microstructure models based on data from 2D slices. More precisely, by establishing the asymptotic normality of suitable longitudinal and cross-sectional summary statistics, goodness-of-fit tests that become asymptotically exact in large sampling windows are devised. The testing methodology is illustrated through a detailed simulation study and a prototypical example from materials science is provided.

1. Introduction

Topological data analysis (TDA) is an emerging branch within the domain of data science that holds the promise to unearth subtle properties of data by extracting shape-related characteristics. While TDA has its root in algebraic topology, which is often considered to be of the most theoretical fields of mathematical research, it is now applied in a variety of domains such as astronomy, biology and materials science (Wasserman, 2018).

While extracting refined topological information is already highly non-trivial, this task becomes even more difficult when the data varies over time. To address this challenge, the notion of *persistence vineyards* was developed, which makes it possible to track the evolution of topological features over time (Cohen-Steiner et al., 2006). These time-varying features are also known as *vines*. While the initial motivation for vineyards arose in a problem of protein folding, this method has now been successfully used in a variety of other contexts such as brain activity data gained via electroencephalography and functional magnetic resonance imaging (Salch et al., 2021a; Yoo et al., 2016). In essence, vineyards allow a researcher to synthesize the information inherent in different time slices into a global summary. In fact, persistence vineyards can be considered a tool from the theory of multi-parameter persistence, which deals with the challenge of extending the success of TDA to settings where the topology depends on multiple parameters. For a recent overview of this active field of research, we refer the reader to the survey paper (Botnan and Lesnick, 2022).

E-mail address: hirsch@math.au.dk (C. Hirsch).
[1] Ny Munkegade 118, 8000, Aarhus C, Denmark.

https://doi.org/10.1016/j.csda.2022.107655

Please cite this article as: A. Cipriani, C. Hirsch and M. Vittorietti, Topology-based goodness-of-fit tests for sliced spatial data, Computational Statistics and Data Analysis, https://doi.org/10.1016/j.csda.2022.107655

In this manuscript, we argue that the concept of vineyards is not by any means restricted to time-varying data but extends to far more general situations involving sliced data. As a prototypical example, we highlight a data set from materials science where the 3D microstructure of an extra low carbon strip steel is measured via 2D slices. Developing statistical tests to determine whether a given stochastic-geometry model provides a good fit to the given sample is an important task since the properties and performance of metals are intimately linked to their microstructure. In fact, unraveling the relation between microstructure features and mechanical properties can lead to the design of new materials with desired properties (Redenbach et al., 2012).

The first step in the investigation is material characterization. The most common way to observe microstructures is by 2D characterization techniques such as light microscopy, scanning electron microscopy, and electron backscatter diffraction. For opaque materials, serial slicing can be used for direct quantitative characterizations of 3D microstructures (Tewari and Gokhale, 2001). Conventional serial slicing involves photographing (or digitally recording) a microstructural field-of-view, polishing the specimen to remove small thickness and, in the second metallographic plane, photographing the field-of-view exactly below the first one. This procedure (i.e., photograph-polish-etch-photograph) is then repeated to generate a stack of aligned serial slices from which 3D microstructure information can be obtained.

However, this process causes material disruption and is time-consuming especially under misalignment problems (Pirgazi, 2019). Finding a stochastic model for microstructures is thus critical in the study of the relation between microstructure features and mechanical properties. There are also important conceptual contributions towards the reconstruction of 3D structures from 2D slices such as (Amini et al., 2013). However, in practice, it is difficult to assess to what extent the proposed density and transversality constraints are satisfied for a given dataset. Moreover, implementing the suggested method is not straightforward. Due to these difficulties, the current testing approaches for 3D data focus almost exclusively on test statistics computed from isolated 2D slices (Hahn and Lorz, 1993; Vittorietti et al., 2020). Therefore, we urgently need tests that can detect deviations of a model from the actual 3D topology without going through the process of a fully-fledged reconstruction.

Motivated by this example, we present a flexible testing framework for statistics derived from persistence vineyards. In its general form, this framework encompasses averages over certain scores associated with the individual vines. However, large spacing between slices may cause uncertainties when identifying the vines. Therefore, we also propose a simpler class of cross-sectional test statistics that can be computed entirely from aggregations over single slices. Our goodness-of-fit-tests are based on the asymptotic normality of these test statistics in large domains under suitable stabilization conditions. This asymptotic normality is the main conceptual contribution of our work. Our framework takes into account that when working with data from materials science, it is quite common to find samples in a highly imbalanced sampling window whose extension in the x-y-directions is far larger than in the z-direction. We stress that the asymptotic normality of the test statistics in large windows offers two decisive advantages. First, since a Gaussian random variable is determined by its first two moments, computing confidence intervals only requires the mean and variance under the null model. Second, it suffices to compute these quantities for moderately large sampling windows; the values obtained in this manner can then be used for data coming from arbitrarily large sampling windows.

After having established our testing framework, we illustrate the method in a simulation study of 2D slices taken from a 3D Voronoi tessellation. This setup illustrates that the asymptotic normality is already clearly visible for moderately large sampling windows. Moreover, by considering tessellations induced by different classes of point processes as generators, we provide indications of the statistical power of the testing framework. Finally, we apply the proposed method to the highlighted example of a 3D metallic microstructure. In that part, we also elucidate which practical obstacles need to be addressed when working with real datasets.

The rest of the manuscript is organized as follows. In Section 2, we define precisely our stochastic-geometry model and the considered test statistics. We also state our results on the asymptotic normality of the test statistics in large domains and explain the sufficient conditions. Among all the possible models used for representing materials microstructure, Voronoi tessellations stand out since this basic model has proven its power in approximating single-phase microstructures. Materials which contain only one phase consist of a single chemical compound. Thus, such materials are characterized by homogeneity from the morphological, chemical, and also mechanical viewpoints. The attractive mathematical properties and the availability of a wide range of sub-models make Voronoi tessellations the state of art for modeling also more complex microstructures (Madej, 2017). Therefore, we verify in Section 3 that the conditions of our main results are satisfied for the key example of 2D sections from a 3D Poisson-Voronoi tessellation. In Section 4, to make the presentation self-contained, we briefly review the most fundamental concepts related to persistence diagrams and vineyards. Next, in Section 5, we illustrate the proposed testing methodology on several simulated data sets. In Section 6, we analyze 2D sections taken from a sample of an extra low carbon strip steel. The proofs for the asymptotic normality of the test statistics are presented in the supplementary material. Finally, Section 7 summarizes the findings and provides an outlook for future research.

2. Methodology and main results

In this section, we explain the general testing methodology and present the main results. First, in Section 2.1, we describe the model assumption of a point cloud varying between sections. Second, in Section 2.2, we present the M-bounded persistence diagram as the main tool to extract topology-related information from data. Next, Section 2.3 describes a class of specific test statistics derived from the persistence diagram. Finally, Section 2.5 contains the precise statement of the

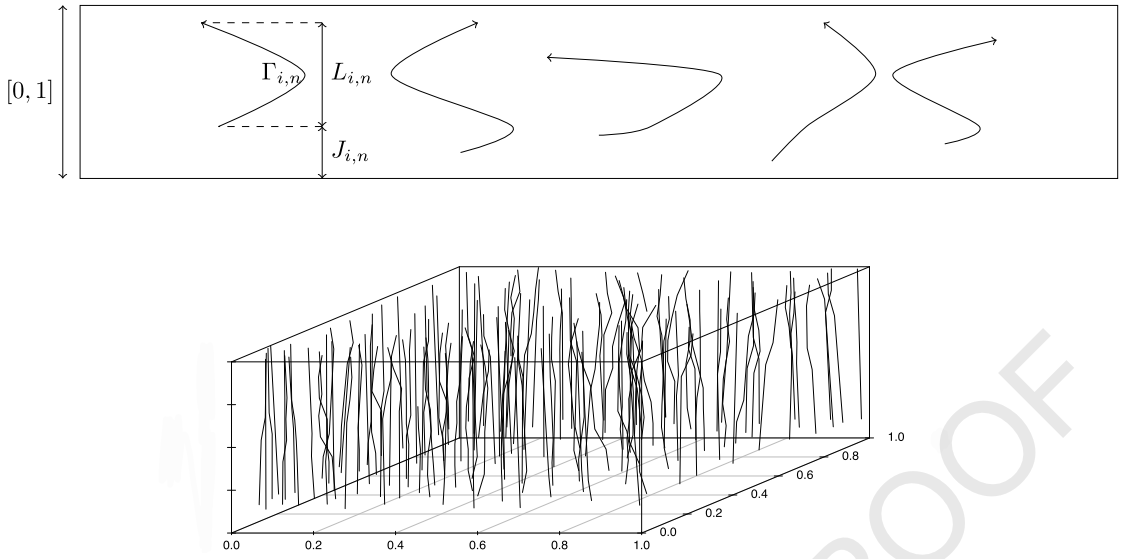

Fig. 1. Point process of trajectories within the window Q_n, $p' = 1$ (top) and $p' = 2$ (bottom).

results on the asymptotic normality of these test statistics provided that a crucial set of stabilization conditions, which are described in detail in Section 2.4, is satisfied.

2.1. Model

We consider statistical testing problems for data describing some geometric structure in the p-dimensional Euclidean space, which is measured along p'-dimensional slices with $p' < p$. In the example from materials science presented in Section 6, we consider a 3D structure measured via 2D slices, i.e., we assume that $p = 3$. However, for clarity of exposition, we present the overall framework in general dimensions.

We now describe more precisely the general purpose behind introducing a stochastic model. First, we note that one important research stream is to develop methods designed to extract 3D information from 2D slices. However, as mentioned in Section 1, reconstructing the 3D structure from 2D slices without distorting the topology is highly challenging. The conceptual approach described in Amini et al. (2013) is difficult to implement, and it is also not clear how to assess whether the set of conditions is satisfied for a given dataset. Hence, our motivation for introducing a stochastic model is not the 3D reconstruction of a dataset from 2D slices but rather to provide a statistically rigorous tool to decide whether a given dataset is in accordance with a null hypothesis. For instance, if the dataset is a 3D structure from materials science that is given as a collection of cells, we might suspect that these cells are not placed entirely at random but that there is some subtle interaction. To confirm this suspicion, we may start from a simple Poisson-Voronoi tessellation as null model and then test whether the hypothesis that the data stems from this model can be rejected at a 5% significance level. If the null hypothesis is indeed rejected, then we have achieved a statistically rigorous justification of our suspicion that the dataset exhibits nontrivial interactions between the cells. We will elaborate on this example in Section 6.

More precisely, in the null model, we let $\mathcal{X}_n = \{X_{i,n}(\cdot)\}_{i\geqslant 1}$ be a process of trajectories, which are observed (but not necessarily entirely contained) in a sampling window $Q_n := Q_n' \times [0,1]$, where $Q_n' := [-n/2, n/2]^{p'}$ is the p'-dimensional cube of side length n. Hence, we always think of the last coordinate as the one along which slices are taken. In the data example discussed in Section 6, the material is represented as a 3D tessellation, and the trajectories are formed by the centroids of the 2D cells in the slices. We now explain more precisely the reason for introducing the trajectory model. As mentioned in Section 1, persistence vineyards allow to track topological features over several slices. In order to achieve this goal, we crucially rely on a trajectory model which introduces an indexing of points so that we can follow individual points over several slices.

On the modeling side, the ith trajectory is determined by an *offset* $J_{i,n} \in [0,1]$, a *length* $L_{i,n} \in [0,1]$ and a *shape* $\Gamma_{i,n}$, where the latter is an element of $C([0,1], \mathbb{R}^p)$, the space of continuous curves from $[0,1]$ to $\mathbb{R}^p$. More precisely, for $\{(J_{i,n}, L_{i,n}, \Gamma_{i,n})\}_{i\geqslant 1} \subseteq [0,1] \times [0,1] \times C([0,1], \mathbb{R}^p)$ we set

$$X_{i,n}(h) := \big\{\big(\Gamma_{i,n}((h - J_{i,n})/L_{i,n}), h\big) \colon J_{i,n} \leqslant h \leqslant J_{i,n} + L_{i,n}\big\}.$$

Fig. 1 illustrates these definitions in 2D and 3D. More precisely, it shows a collection of fibers in the window Q_n, and illustrates for one of these fibers the definition of $J_{i,n}$, $L_{i,n}$, and $\Gamma_{i,n}$. We point out that when working with datasets stemming from a finite number of slices, the trajectories need to be reconstructed. We will discuss this issue in greater detail in Section 6.

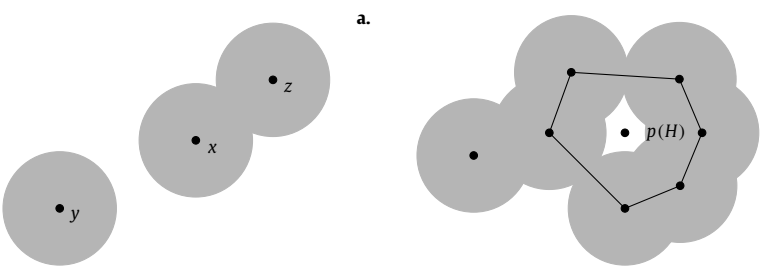

Fig. 2. a. M-bounded clusters, b. M-bounded loops.

2.2. Method: M-bounded persistence diagrams

Having set up the model, we rely on the tools of *persistent homology* in order to quantify the most striking topological aspects of the slices. For a superbly-written overview of this methodology intended for an audience with a background in statistics, we suggest (Wasserman, 2018) for further reading. However, for the presentation of the main results we do not need this abstract setting since we focus on the case $p = 3$ from now on. The restriction to the case of 2D slices of a 3D structure has the following reasons:

1. The data example in Section 6, which provides the motivation for the development of our goodness-of-fit tests, pertains precisely to the scenario of 2D slices of a 3D material.
2. When working with 2D slices, then persistent homology captures two types of topological features, zero-dimensional and one-dimensional homology generators in the Čech complex. In order to make the presentation more accessible for readers that are new to TDA, we will henceforth refer to them as connected components and loops. As we will elaborate further below, both of these features admit a concrete topological description not relying on the abstract toolkit of simplicial homology.
3. Long-range correlations coming from percolation theory render it difficult to keep track of features of an arbitrary size. Therefore, we proceed as in Biscio et al. (2020) and restrict our attention to features whose size admits a deterministic upper bound. Again, while in general it is not at all immediate how to measure the size of features represented in simplicial homology, the 2D setting has already been worked out in detail in Biscio et al. (2020).

Topologically, connected components correspond to zero-dimensional homology generators in the Čech, and loops correspond to one-dimensional homology generators. However, from the statistical perspective, one issue with these traditional notions is that these homology generators could potentially extend over large regions of the sampling window, thereby inducing long-range correlations, which make it difficult to establish asymptotic normality rigorously. To deal with this issue, we rely on the concepts of M-bounded clusters and M-bounded loops introduced in Biscio et al. (2020), and which will be explained in further detail below. The difference to the homology generators from the Čech complex is that the M-bounded variants do not induce long-range correlations, and are therefore more amenable to a rigorous asymptotic statistical analysis.

Let $M > 0$ and $\varphi \subseteq \mathbb{R}^2$ be a finite set of points. In order to make the manuscript self-contained, we now briefly recall the concepts of M-bounded clusters and M-bounded loops, and refer the reader to Biscio et al. (2020, Section 2) for the full technical details.

2.2.1. M-bounded clusters

To describe the M-bounded clusters at a level $r > 0$, we consider the connected components of the union set $\bigcup_{x\in\varphi} B_r(x)$, where $B_r(x)$ denotes the Euclidean disk of radius $r > 0$ centered at $x \in \varphi$. Note that at level $r = 0$ each $x \in \varphi$ forms its own connected component. We say that all of these components are born at $r = 0$. Each of these components disappears at a unique level r, which is called the *death time* of the component. This may occur for one of two reasons: i) at the level r the spatial diameter of the component exceeds M or ii) at the level r two growing components merge into a single one. Here, we stress that in case ii) only one of the two components "dies". For instance, we may proceed as in Biscio et al. (2020, Section 2.1) and decide this according to the lexicographic ordering of the two disks from the components that collide at level r. See Fig. 2 for an illustration (adapted from Biscio et al., 2020).

The collection of death times $\{D_i\}_{i\geqslant 1}$ defines the *M-bounded persistence diagram in degree 0*, which gives a summary of the M-bounded 0-features, i.e., of the merging pattern of M-clusters. To be consistent with the persistence diagram of M-bounded features to be introduced below, sometimes an additional first coordinate is added, which is identical to 0. We then write $\{(0, D_i)\}_{i\geqslant 1}$.

2.2.2. M-bounded loops

To describe the M-bounded loops, we proceed along similar lines with the difference that now we do not consider the connected components of $\bigcup_{x\in\varphi} B_r(x)$ but rather the connected components of the complement $\mathbb{R}^2 \setminus \bigcup_{x\in\varphi} B_r(x)$ that are of size at most M. To any such component H we may assign the *center* $p(H)$ to be the point which is covered last, i.e., which

has the largest distance to φ. If that point is covered at a level $r > 0$, then we refer to that level as the *death time* of that component. In contrast to the situation of the clusters, there are no M-bounded loops at level $r = 0$. A new M-bounded loop may arise at a level $r > 0$ if an existing (bounded or unbounded) component of $\mathbb{R}^2 \setminus \bigcup_{x\in\varphi} B_r(x)$ is split up into two. Next, we expound on how to measure the size of loops. More precisely, $B_r(x_1), \ldots, B_r(x_k)$ are the disks defining the boundary of a bounded connected component of $\mathbb{R}^2 \setminus \bigcup_{x\in\varphi} B_r(x)$, then we define the *size* of that component as the metric diameter of the set $\{x_1, \ldots, x_k\}$. An *M-bounded loop* is a connected component of size at most M. Finally, we define the *birth time* of a loop H to be the smallest level $r > 0$ for which there exists an M-bounded loop H' with $p(H') = p(H)$.

Similarly to the setting of M-bounded clusters, the collection of M-bounded birth times and death times $\{(B_i, D_i)\}_{i\geqslant 1}$ is a fingerprint for the appearance and disappearance of M-bounded 1-features. Hence, $\{(B_i, D_i)\}_{i\geqslant 1}$ is called the *M-bounded persistence diagram in degree 1*. In particular, both the birth time and the death time of M-bounded features admit a deterministic upper bound, and we may therefore assume that the persistence diagram is contained in $[0, \tau]^2$ for some $\tau > 0$.

2.3. Test statistics

The core of our testing methodology may be summarized in two key steps. First, in Section 2.2, we leveraged persistence diagrams to extract topological information from each of the slices. Now, we describe how to combine the information from individual slices into more powerful global test statistics.

To that end, in each slice $h \in [0, 1]$, we compute the M-bounded persistence diagram induced by points $\{X_{j,n}(h)\}_{j\geqslant 1}$, tacitly discarding trajectories with $h \notin [J_{j,n}, J_{j,n} + L_{j,n}]$. The persistence diagram in dimension q tracks the appearance and disappearance of q-dimensional features. Thus, it is represented as a collection $\{(B_i(h), D_i(h))\}_i = \{(B^q_{i,n}(h), D^q_{i,n}(h))\}_i$ of birth and death times in a fixed time interval $[0, \tau]$.

Since the persistence diagrams in different slices are intimately related through the trajectories of the underlying point pattern, it is attractive to move beyond considering the respective persistence diagrams in isolation. As indicated in Section 1, persistence vineyards or time-varying persistence diagrams are introduced in Cohen-Steiner et al. (2006) as a tool to study the persistence of dynamic point clouds. Although the vines as described in the previous paragraph are similar in spirit to this classical concept, our setting deviates from the time-varying vines in a few technical characteristics, on which we now elaborate in further detail. First, Cohen-Steiner et al. (2006, Section 4) computes the simplicial complex based on the lower-star filtration of an embedding of an abstract vertex set. Examples considered in the literature comprise distances resulting from the embedding a pair of points as single vertex, or the marking of 1-simplices through the negative correlation of two time series. In contrast, the vertices in our vines are already embedded in Euclidean space so that we may rely on the classical Čech filtration. Even though with a sufficiently fine spacing between the slices, we are in principle able to track the vertices over the slices, we still observe *switches* as in Edelsbrunner and Harer (2010, Chapter VIII). That is, when the filtration values of certain simplices swap their ordering, then the simplices causing the births and deaths of features may change from one slice to another.

After stacking the persistence diagrams along the h-axis, we now explain how to define a *vine* $(B_i(h), D_i(h)), h \in [0, 1]$ associated with a *fixed feature i*. Each point in the diagrams traces out a path called a *vine*, i.e., a line among points of the persistence diagrams. We start with the case of 0-features, i.e., M-bounded clusters, where features are in 1-1 correspondence with the original trajectories. That is, for $i \geqslant 1$ and $h \in [J_i, J_i + L_i]$, we define $D_i(h)$ to be the death time of the point $X_{i,n}(h)$ on $\{X_{j,n}(h)\}_{j\geqslant 1}$.

The case of 1-features is more involved since this one-to-one correspondence breaks down. In this case, we associate the ith feature with the simplex leading to the birth of that feature. That is, we mean the simplex appearing at the same time as the feature appears in the persistence diagram. If all simplices appear at different times, then this is uniquely determined. Here, we identify the simplices in different slices by their indices. That is, if $\{X_{i_0,n}(h), \ldots, X_{i_q,n}(h)\}$ is a q-simplex in the slice h, then the corresponding q-simplex in another slice h' is given by $\{X_{i_0,n}(h'), \ldots, X_{i_q,n}(h')\}$. Moreover, as observed in Biscio et al. (2020, Section 8), when working with M-bounded loops, it is possible that a single 1-simplex leads to the birth of two M-bounded loops. To retain the uniqueness of the identification we may associate one of the features with a "-", and the other with a "+" sign. However, as soon as M is chosen reasonably large, the simultaneous birth of two M-bounded loops is very rare, and therefore does not cause any concerns when evaluating simulated and real data in Sections 5 and 6, respectively. Note that whether or not a certain collection of centers defines a feature may change from one slice to another. We stress that this convention only leads to a certain approximation of the actual vines. Moreover, it may often be very natural to consider a loop as being associated with the same feature even if a different simplex causes the birth of this loop. Finding useful heuristics for carrying out these more subtle identifications is an instance of the problem of *cycle registration* and is an active research field (Reani and Bobrowski, 2021).

To test whether data gathered from slices comply with a null hypothesis on the distribution on the underlying full-dimensional geometric structure, we propose the following general class of *longitudinal test statistics*. In analogy to the vines from Cohen-Steiner et al. (2006), for each fixed $H \geqslant 1$, we can consider the discretized trajectory $\{(B_i(h), D_i(h))\}_{h\in\Xi_H}$ of the ith feature, where $\Xi_H := \{0, 1/H, \ldots, (H-1)/H\}$ describes a subdivision into slices of length $1/H$.

Now, we associate with each vine some statistic $\xi\big(\{(B_i(h), D_i(h))\}_{h\in\Xi_H}\big)$ and form a global test statistic by aggregating over all features in the window Q_n:

$$T_n = T(\mathcal{X}_n) := \sum_i \xi\big(\{(B_i(h), D_i(h))\}_h\big). \tag{1}$$

For instance, the score ξ could be the average number of slices in Ξ_H where the feature is present, or the life time averaged over those features. In Section 5.1.3, we define precisely test statistics of this flavor.

Example 2.1 *(Cross-sectional statistics).* Longitudinal statistics have the appeal of capturing the evolution of individual features over time. In practice, however, the trajectories might be observed only over a very limited number of slices so that vines may extend only of a handful of slices. Hence, computing general longitudinal statistics could be challenging. Nevertheless, by carefully choosing the score function ξ, we construct *cross-sectional test statistics* that can still be evaluated seamlessly in such cases.

More precisely, for any measurable function $\xi' \colon [0,\tau]^2 \to [0,\infty)$, we consider

$$\xi\big(\{(B_i(h), D_i(h))\}_{h\in\Xi_H}\big) := \frac{1}{H}\sum_{h\in\Xi_H} \xi'(B_i(h), D_i(h)),$$

implicitly setting $\xi'(B_i(h), D_i(h)) = 0$ if the ith feature is not present at height h. Then, forming the test statistic T_n as in (1), and exchanging sums, we arrive at

$$T_n = \frac{1}{H}\sum_{h\in\Xi_H}\sum_i \xi'(B_i(h), D_i(h)). \tag{2}$$

Here, we stress that to evaluate T_n it is not necessary to track features, since the inner sum depends only on the persistence diagram at time h. In fact, we can consider this estimator also as an integration with respect to the empirical measure associated with the persistence diagrams in the individual slices.

To make this precise, we may think of the persistence diagrams in a window as an empirical measures on $[0,\tau]^2$ averaged over the slices, i.e.,

$$\mathsf{PD}_n^q(\mathrm{d}b, \mathrm{d}d) := \frac{1}{H}\sum_{h\in\Xi_H}\sum_{i\geqslant 1} \delta_{(B_i^q(h), D_i^q(h))}(b,d).$$

Then, the test statistic (2) becomes

$$T_n := \langle \xi', \mathsf{PD}_n^q\rangle = \frac{1}{H}\sum_{h\in\Xi_H}\sum_{i\geqslant 1} \xi'(B_i(h), D_i(h)). \tag{3}$$

For instance, taking $\xi'(b,d) := d - b$ yields the total persistence averaged over the slices. As another example, taking the indicator $\xi'(b',d') := \mathbb{1}_{[0,b]\times[d,\tau]}(b',d')$ yields the *slice-averaged M-bounded persistent Betti numbers* $\beta_n^{b,d} := \mathsf{PD}_n^q([0,b],[d,\tau])$.

2.4. Stabilization conditions

To develop goodness-of-fit tests based on the longitudinal test statistics described above, we show that they are asymptotically normal in large domains. On a very general level, proofs of asymptotic normality are typically based on certain dependence assumptions, and we now elaborate on what form these assumptions take in the present setting.

Although we fix $p = 3$, we keep the abstract notation, writing $\mathbb{R}^p$, $\mathbb{R}^{p'}$ instead of $\mathbb{R}^3$ and $\mathbb{R}^2$. This makes the arguments more transparent, and simultaneously opens the door towards future generalizations.

We assume that $\mathcal{X}_n$ emerges from a homogeneous Poisson point process $\mathcal{P} = \{P_i\}$ on $\mathbb{R}^p$ with intensity $\lambda > 0$ in the background through a construction rule that is stabilizing in the spirit of Penrose and Yukich (2001). This means that changing $\mathcal{P}_n := \mathcal{P}\cap Q_n$ far away from a position $x \in Q_n$ does not change the trajectories in the vicinity of x. More precisely, we assume that the family $\mathcal{X}_n$ of trajectories is expressed as

$$\mathcal{X}_n = \{\mathcal{T}(P_i, \mathcal{P}_n)\}_{P_i\in\mathcal{P}_{n-\sqrt{n}}} \tag{4}$$

for some measurable transformation $\mathcal{T} \colon \mathbb{R}^p \times \mathbf{N} \to C([0,1],\mathbb{R}^p)$, where $\mathbf{N}$ denotes the family of all locally finite sets in $\mathbb{R}^p$. We assume that $\mathcal{T}$ is translation-covariant, in the sense that $\mathcal{T}(x+y,\varphi+y) = \mathcal{T}(x,\varphi) + y$ for every $y\in\mathbb{R}^p$, locally finite $\varphi\subseteq\mathbb{R}^p$, and $x\in\varphi$. In Section 3 below, we present a detailed example for $\mathcal{T}$ based on the Voronoi tessellation. Here, to control edge effects, we restrict to trajectories associated with Poisson points in the slightly eroded window $Q_{n-\sqrt{n}}$. We also note that in practice, the trajectories are not observed, whereas definition (4) relies on the index. Hence, when working with actual datasets, we need an additional preprocessing step that extracts trajectories from individual points in the slices. We will elaborate on this issue in detail in Section 5.4.3 below.

Moreover, we assume exponential stabilization in the vein of Błaszczyszyn et al. (2019); Penrose and Yukich (2001). This involves studying the effects of changing a point pattern φ close to a given location $x\in Q_n'$. More precisely, for any $r>0$ we put $Q_r(x) := (x+Q_r')\times[0,1]$, $Q(x) := Q_1(x)$ and for any finite $\mathcal{A}\subseteq Q_n\setminus Q_r(x)$, $r<n$, we let

$$\varphi_{r,\mathcal{A}} := \mathcal{A}\cup\big(\varphi\cap Q_r(x)\big)$$

be the configuration obtained from φ by placing $\mathcal{A}$ outside of $Q_r(x)$, and for $\mathcal{B} \subseteq Q(x)$ we let

$$\varphi_{r,\mathcal{A},\mathcal{B}} := \mathcal{A} \cup \mathcal{B} \cup \big(\varphi \cap (Q_r(x) \setminus Q(x))\big)$$

be the configuration where additionally $\mathcal{B}$ is placed in $Q(x)$. Moreover, for two trajectories $f, g \in C([0,1], \mathbb{R}^p)$, we define their distance through $\mathsf{dist}(f, g) := \inf_{h \leqslant 1} \|f(h) - g(h)\|$.

Definition 2.2 *(Stabilization).* A finite $\varphi \subseteq Q_n$ *stabilizes for* $x \in \mathbb{R}^{p'}$ *at* $r \geqslant 1$ *relative to* Q_n if

(**S1**) We have

$$\mathcal{T}(P, \varphi_{r,\mathcal{A},\mathcal{B}}) = \mathcal{T}(P, \varphi_{r,\mathcal{A},\varnothing})$$

for any finite $\mathcal{A} \subseteq Q_n \setminus Q_r(x)$, $\mathcal{B} \subseteq Q(x)$ and $P \in \varphi_{r,\mathcal{A}} \setminus Q_r(x)$. Loosely speaking, for the configuration outside $Q_r(x)$, what happens inside $Q(x)$ is not relevant.

(**S2**) We have

$$\mathsf{dist}\big(\mathcal{T}(P, \varphi_{2r,\mathcal{A},\mathcal{B}}), \mathcal{T}(P', \varphi_{2r,\mathcal{A},\mathcal{B}})\big) > M$$

for any finite $\mathcal{A} \subseteq Q_n \setminus Q_{2r}(x)$, $\mathcal{B} \subseteq Q(x)$, $P \in \mathcal{A}$ and $P' \in \varphi_{2r,\mathcal{A},\mathcal{B}} \cap Q_r(x)$. Loosely speaking, M-bounded features with points $P \in \mathcal{A}$ cannot contain points inside $Q_r(x)$.

(**S3**) We have $\mathcal{T}(P, \varphi_{4r,\mathcal{A}}) = \mathcal{T}(P, \varphi_{4r,\varnothing})$ for any finite $\mathcal{A} \subseteq Q_n \setminus Q_{4r}(x)$, and $P \in \varphi \cap Q_{2r}(x)$. Loosely speaking, trajectories in $Q_{2r}(x)$ will not be affected by the change to the configuration $\mathcal{A}$. Here, we changed from $2r$ to $4r$ in order to harmonize well with condition (**S2**).

Formally, we also set $Q_\infty := \mathbb{R}^{p'} \times [0,1]$, and impose in this case additionally that the entire trajectories from (**S3**) be contained in $Q_{4r}(x)$. That is,

(**S3′**) We have $\mathcal{T}(P, \varphi_{4r,\mathcal{A}}) \subseteq Q_{4r}(x)$ for any finite $\mathcal{A} \subseteq \mathbb{R}^p \setminus Q_{4r}(x)$, and $P \in \varphi \cap Q_{2r}(x)$.

Then, we let $R'(x, n; \varphi)$ denote the smallest integer $r \geqslant 1$ for which φ stabilizes at x relative to Q_n, and $R'(x, \infty; \varphi)$ the smallest integer $r \geqslant 1$ for which φ stabilizes relative to $\mathbb{R}^p$. We define the *stabilization radius* $R(x, n; \varphi) := R'(x, n; \varphi) \vee R'(x, \infty; \varphi)$. Note that while $R'(x, n; \varphi) \leqslant 2n$, it may happen that $R'(x, \infty; \varphi) = \infty$. In particular, we write $R(x, n) := R(x, n; \mathcal{P})$ for the stabilization radius associated with the input point process. Finally, we need the notion of exponential stabilization.

Definition 2.3 *(Exponential stabilization).* $\mathcal{T}$ is *exponentially stabilizing* if for some $c > 0$ one has

$$\limsup_{r \to \infty} \sup_{\substack{n \geqslant 1 \\ x \in Q_n}} \frac{\log(\mathbb{P}(R(x, n) > r))}{r^c} < 0.$$

2.5. Asymptotic normality

After having specified the conditions, we now state the asymptotic normality of scalar test statistics.

Theorem 2.4 *(Asymptotic normality; scalar level). Let $\mathcal{P} = \{P_i\}$ be a Poisson point process with intensity $\lambda > 0$. Furthermore, let $\mathcal{X}_n = \{\mathcal{T}(P_i, \mathcal{P}_n)\}_{P_i \in \mathcal{P}_{n-\sqrt{n}}}$, where $\mathcal{T}$ is a translation-covariant transformation. Assume that $\mathcal{T}$ is exponentially stabilizing as defined in (2.3) and that the score function ξ is bounded. Then, the test statistic*

$$\frac{T_n - \mathbb{E}[T_n]}{n^{p'/2}}$$

converges in distribution to a normal random variable.

Since it will be used in the proof of Theorem 2.5 below, we mention that Theorem 2.4 applies for M-bounded persistent Betti numbers.

When dealing with cross-sectional statistics, previous work from Biscio et al. (2020) already suggests that it should not only be possible to derive asymptotic normality separately for each scalar statistics F but obtain a functional statement. This is the content of our second main result, Theorem 2.5. The advantage of functional asymptotic normality of the persistent Betti numbers is that one can consider Kolmogorov-Smirnov-type statistics depending on a continuous range of birth/death-parameters. Moreover, in situations where the information loss of a simple Kolmogorov-Smirnov-type statistics is a concern, one can also consider combining them with other options such as Cramér-von Mises-type statistics.

To prove the functional CLT, we need to impose further moment conditions. More precisely, we assume that the factorial moment densities $\rho_{h,q,n}(x_0,\dots,x_q)$ of the points $\{X_{i,n}(h)\}_{i\geqslant 1}$ in the slice $h\in\Xi_H$ are bounded in the sense that for every $q\geqslant 1$,

$$\sup_{\substack{h\in\Xi_H\\ n\geqslant 1}}\sup_{x_0,\dots,x_q\in\mathbb{R}^p}\rho_{h,q,n}(x_0,\dots,x_q)\leqslant C_{\rho,q} \tag{M}$$

for some $C_{\rho,q}>0$. Here, the factorial moment densities are determined by the disintegration property

$$\mathbb{E}\Big[\sum_{X_{i_0,n},\dots,X_{i_q,n}\in\mathcal{X}_n} f(X_{i_0,n}(h),\dots,X_{i_q,n}(h))\Big]=\int_{\mathbb{R}^{dq}} f(x_0,\dots,x_q)\rho_{h,q,n}(x_0,\dots,x_q)\mathrm{d}(x_0,\dots,x_q) \tag{5}$$

where $X_{i_0,n},\dots,X_{i_q,n}$ must be pairwise distinct, see Last and Penrose (2016, Section 4.2).

Theorem 2.5 *(Asymptotic normality of M-bounded persistent Betti numbers; functional level). Assume that $\mathcal{T}$ is exponentially stabilizing and that condition (M) holds. Then, as a function on $[0,\tau]^2$, in the Skorohod topology, the recentered and rescaled M-bounded persistent Betti numbers*

$$\frac{\beta_n^{M,b,d}-\mathbb{E}[\beta_n^{M,b,d}]}{n^{p'/2}}$$

converge in distribution to a centered Gaussian process.

We note that the asymptotic normality M-bounded persistent Betti numbers is also the core theme of Biscio et al. (2020). However, this work deals with the setting where the data stems from a single planar point pattern. In contrast, Theorem 2.5 starts from the case where we have a 3D dataset that is observed in several slices that can potentially exhibit complex correlations.

We also note that by relying on the techniques from Błaszczyszyn et al. (2019), it is plausible that Theorems 2.4 and 2.5 can be established even in cases where $\mathcal{P}$ is not a Poisson point process. However, this generalization would come at the expense of introducing even further technical assumptions and making the proofs far less accessible. Since at the same time, the Poisson point process is the most fundamental model, we decided to present our theory in that setting.

3. Example

In this section, we present a specific model for the trajectories $\mathcal{X}_n=\{X_{i,n}\}_{i\geqslant 1}$, namely as perturbed centroids of the 2D slices of a 3D Poisson-Voronoi tessellation. We will present a detailed proof of the conditions from Theorem 2.5. This model will also form the basis for the simulation study in Section 5 below. The motivation for considering this model comes from materials science, where a 3D Poisson-Voronoi tessellation represents one of the most basic models for single-phase microstructures.

More precisely, let $\mathcal{P}_n=\{P_{i,n}\}\subseteq Q_n$ be a homogeneous Poisson point process and consider the associated 3D Voronoi tessellation $\{\mathcal{C}_{i,n}\}_i$ on $\mathcal{P}_n$. Then, for each $h\in[0,1]$ and each bounded cell $\mathcal{C}_{i,n}$, we let $G_{i,n}(h)$ be the centroid of the section $\mathcal{C}_{i,n}\cap(\{h\}\times\mathbb{R}^2)$ provided that the latter is non-empty. We denote the entire collection of centroids at level h as $G_n(h)$. To cast this example in the framework described in Section 2.1, we note that here the trajectories $\{\Gamma_i\}_{i\geqslant 1}$ correspond to the perturbed trajectories of the centroids through the slices.

We assume that, from the data, we only extract a slight perturbation of $G_{i,n}$. This added randomness is justified from measurement errors and is also technically useful in the proofs. Hence, we set

$$X_{i,n}(h):=G_{i,n}(h)+N_{i,n}(h),$$

where the $\{N_{i,n}(h)\}_{i,h}$ are assumed to be noise variables that are independent for different i but may be correlated in h. For instance, we take $N_{i,n}(h)\sim\mathsf{U}(B(o,\eta_0))$ to be uniformly distributed in a ball of a small radius $\eta_0<1/4$. Here, we use the notation $B(x,m)\subset\mathbb{R}^d$ for the Euclidean ball around $x\in\mathbb{R}^p$ of radius $m>0$. We now verify condition (M) and exponential stabilization. Henceforth, to ease notation, $|A|$ denotes either the cardinality of a discrete set A or the Lebesgue measure of A if A is a measurable subset of the Euclidean space.

Condition (M). Let $A_0,\dots,A_q\subseteq Q_n$ be Borel sets whose diameter and volume we may assume to be at most 1. Then, we claim that

$$\mathbb{E}\Big[\Big|\big\{i_0,\dots,i_q\geqslant 1\text{ pairwise distinct}\colon X_{i_0,n}(h)\in A_0,\dots,X_{i_q,n}(h)\in A_q\big\}\Big|\Big]\leqslant C_q\prod_{j\leqslant k}|A_j|, \tag{6}$$

with the constant C_q depending neither on h nor n.

To prove this claim, we observe that since the support of the noise variables is of diameter at most $\eta_0 \leqslant 1/4$, we can have $X_{i_j,n}(h) \in A_j$ only if $G_{i_j,n}(h)$ is contained in the 1-neighborhood

$$A_j^+ := \{x + y \colon x \in A_j,\ y \in Q_{1/2}\},$$

of A_j. Since the noise variables are uniformly distributed, there exists some $c_0 > 0$ such that conditioned on the locations $\{P_i\}_{i \geqslant 1}$,

$$\mathbb{P}\big(X_{i_0,n}(h) \in A_1, \ldots, X_{i_q,n}(h) \in A_q \big| \{P_i\}_i\big) \leqslant c_0^k \mathbb{1}_{\{G_{i_1,n}(h) \in A_1^+, \ldots, G_{i_q,n}(h) \in A_q^+\}} \prod_{j \leqslant q} |A_j|.$$

Hence, setting $A^+ := A_1^+ \times \cdots \times A_q^+$ and plugging in the right-hand side back into (6), it suffices to show that $\sup_{h \in [0,1]} \mathbb{E}\big[|G_n(h)^{q+1} \cap A^+|\big] < \infty$. Moreover, by the Hölder inequality, this further reduces to $\sup_{j \leqslant k,\, h \leqslant 1} \mathbb{E}\big[|G_n(h) \cap A_j^+|^k\big] < \infty$. To prove this claim, fix some point $x \in A_j$ and $h \in [0,1]$, and note that if A_j is hit by $\ell \geqslant 1$ Voronoi cells, then there exists some $2 \leqslant m \leqslant 3n$ such that i) there are no points in $B(x, (m-4)_+)$ and ii) there are at least ℓ Poisson points in $B(x, m) \setminus B(x, m-2)$ (Lemma C.1). Thus, for any $\ell \geqslant 2^8$,

$$\begin{aligned}
\mathbb{P}\big(|G_n(h) \cap A_j^+| = \ell\big) &\leqslant \sum_{2 \leqslant m \leqslant 3n} \mathbb{P}\big(\mathcal{P}_n\big(B(x, m) \setminus B(x, m-2)\big) \geqslant \ell, \mathcal{P}_n \cap B(x, (m-4)_+) = \varnothing\big) \\
&\leqslant \sum_{m \leqslant \ell^{1/4}} \mathbb{P}\big(\mathcal{P}_n\big(B(x, m) \setminus B(x, m-2)\big) \geqslant \ell\big) \\
&\quad + \mathbb{1}\{3n \geqslant \ell^{1/4}\} \sum_{\ell^{1/4} < m \leqslant 3n} \mathbb{P}\big(\mathcal{P}_n \cap B(x, m-4) = \varnothing\big),
\end{aligned}$$

where $\mathcal{P}_n(A) := |\mathcal{P}_n \cap A|$ denotes the number of points in a set A. Now, we bound the two summands separately. Since ℓ exceeds the expected number of Poisson points in $B(x, m) \setminus B(x, m-2)$, the first summand decays exponentially fast in ℓ by the Poisson concentration inequality. The probability in the second sum is at most $\exp(-c_1 m^3)$ for some universal constant $c_1 > 0$, so that the sum decays exponentially fast in $\ell^{3/4}$.

Exponential stabilization. We argue that exponential stabilization holds for the centroids, noting that a minor modification also gives the property after the perturbation. To achieve this goal, let

$$E_{r,x,n} := \Big\{ Q_{\sqrt{r}}(\sqrt{r}z) \cap \mathcal{P} \neq \varnothing \colon \text{for all } \sqrt{r}z \in \sqrt{r}\mathbb{Z}^{p-1} \cap Q_n \cap Q_r(x) \setminus Q_{r-8\sqrt{r}}(x) \Big\}$$

denote the event that for each point $\sqrt{r}z \in \sqrt{r}\mathbb{Z}^{p-1} \cap Q_n \cap Q_r(x) \setminus Q_{r-8\sqrt{r}}(x)$ the $\sqrt{r}$-box $Q_{\sqrt{r}}(\sqrt{r}z)$ contains at least one point from $\mathcal{P}$. The intuition for the event $E_{r,x,n}$ is the following. If each of the boxes contains at least one point from $\mathcal{P}$, then these boxes prevent the existence of large regions of void space so that they can be considered as barriers that prevent individual Voronoi cells from becoming very large. We claim that

$$\big\{E_{r/2,x,n} \cap E_{r,x,n} \cap E_{3r/2,x,n} \cap E_{5r/2,x,n}\big\} \subseteq \big\{R'(x, n; \mathcal{P}) \leqslant r\big\}, \tag{7}$$

and similarly with n replaced by ∞. Once we have proven this claim, the desired exponential stabilization will follow from a union bound in (7) and Lemma C.2 which gives, for two positive constants c_1, c_2

$$\mathbb{P}(E_{r,x,s}^c) \leqslant c_1 r^{p'/2} \exp\big(-c_2 r^{p'/2}\big).$$

To prove (7), we proceed in two steps: for all $s > 0$

(1) under the event $E_{s,x,n}$, every cell centered in $Q_{s-8\sqrt{s}}(x)$ is contained in $Q_{s-2\sqrt{s}}(x)$;
(2) under the event $E_{s,x,n}$, every cell centered outside $Q_{s+2\sqrt{s}}(x)$ does not intersect $Q_{s-2\sqrt{s}}(x)$.

In particular, for sufficiently large $s > 0$, under the event $E_{s/2,x,n}$ every cell centered in $Q_1(x)$ is contained in $Q_{s/2-\sqrt{2s}}(x)$, and every cell centered outside $Q_{s/2+\sqrt{2s}}(x)$ does not intersect $Q_{s/2-\sqrt{2s}}(x)$ by (2). Moreover, for sufficiently large $s > 0$, by (1) the event $E_{s,x,n}$ forces every cell centered inside $Q_{s/2+\sqrt{2s}}(x)$ to be contained in $Q_{s-2\sqrt{s}}(x)$. Therefore, under the event $E_{s/2,x,n} \cap E_{s,x,n}$, no cell centered in $Q_1(x)$ can intersect a cell with a point outside $Q_s(x)$, thereby establishing property (**S1**). Similar arguments show properties (**S2**), (**S3**) and (**S3′**).

We now provide the detailed derivations for claims (1) and (2). For claim (1), let C be a cell centered at $P \in Q_{s-8\sqrt{s}}(x)$, let P' be an arbitrary point outside $Q_{s-2\sqrt{s}}(x)$ and let P'' be its closest point on the boundary of $Q_{s-2\sqrt{s}}(x)$. In particular, $|P' - P| \geqslant |P' - P''| + 3\sqrt{s}$. On the other hand, under the event $E_{s,x,n}$ the distance from P' to one of the centers guaranteed by $E_{s,x,n}$ is at most $|P' - P''| + \sqrt{3s} < |P' - P|$, so that $P' \notin C$. The situation is sketched in Fig. 3.

For claim (2), let C be a cell centered at $P \in \mathbb{R}^p \setminus Q_{s+2\sqrt{s}}(x)$, let P' be an arbitrary point inside $Q_{s-2\sqrt{s}}(x)$ and let P'' be its closest point on the boundary of $Q_{s-2\sqrt{s}}(x)$. Then, as before, $|P' - P| \geqslant |P' - P''| + 2\sqrt{s}$. On the other hand, under the event $E_{s,x,n}$, again the distance from P' to one of the centers guaranteed by $E_{s,x,n}$ is at most $|P' - P''| + \sqrt{3s} < |P' - P|$.

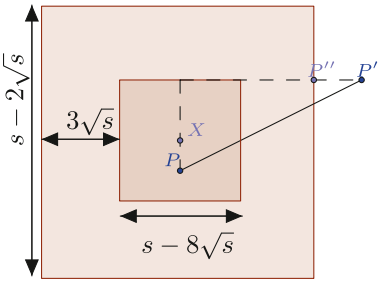

Fig. 3. Sketch for the proof of claim (1).

4. Preliminaries on persistence diagrams and vineyards

The use of topological data analysis in computational statistics has only recently become more widespread. Hence, to make our work accessible to researchers who have not worked in this domain so far, we now explain the key ideas behind persistence diagrams and vineyards. We refer the readers to Edelsbrunner and Harer (2010) for a more detailed exposition including formal definitions.

Some of the most qualitative topological features of data are encoded in the *homology groups*. Homology groups and their dimensions, called Betti numbers, are inherently global properties of a topological space that are insensitive to continuous geometric deformations. Such deformations may include shrinks, twists, etc. but they do not include tearing or gluing, which change the overall topology. In this paper, the focus is on:

H_0: homology group 0, the vector space encoding information about connected components;
H_1: homology group 1, the vector space encoding information about loops.

For illustrative reasons let us consider the 2D case, stressing that all arguments can be generalized to higher dimensions. The data often takes the form of a point cloud $\mathbf{X}$. Based on that, we build a *simplicial complex*, which is a special structure encoding information about the qualitative features discussed above. Here, a simplicial complex can be thought of a collection of geometric k-simplices, i.e., convex hulls of $k+1$ affinely independent points $v_0, v_1, \ldots, v_k$. More precisely the 0-simplex identifies vertices, the 1-simplex line segments and the 2-simplex triangles. Many choices of simplicial complexes are available in the literature, and we focus on the Čech complex as one of the most prominent ones. The Čech complex $\mathsf{Cech}_r(\mathbf{X})$ at a radius $r > 0$ consists of the following simplices:

1. a vertex set equal to $\mathbf{X}$;
2. simplices $\sigma \subseteq \mathbf{X}$ if and only if $\cap_{x\in\sigma} B(x, r) \neq \varnothing$.

One way for building the Čech complex starts off with $\mathbf{X}$. Then, disks are grown with increasing radius r, centered at the points in $\mathbf{X}$. For very small r, the Čech complex boils down to the set $\mathbf{X}$ of the generator points. Then, r grows and once two disks intersect, an edge between the two disk centers is added to $Cech_r(\mathbf{X})$.

The key idea in persistent homology is that rather than considering this structure for some fixed value of r, its evolution for growing $r > 0$ is registered. In particular, we keep track of the birth time b and a death time d of connected components and loops, where the 'time' is given by the radius of the disks corresponding to those events. One can think of the disks radii growing at constant rate. At time zero, the Čech complex equals $\mathbf{X}$. Hence, all individual points are separate connected components and are born at time zero. After some time, when the first two points get connected because their disks touch, one can say two connected components merge or one connected component 'dies'. Note that in a more general setting, where some of the birth times are non-zero, the edge that causes two components to merge is paired with the component that was born most recently, i.e. the one with the later birth 'time'. Increasing r further, more connected components will 'die' until only a single one remains.

During this growth process, also loops appear. This happens when a polygon appears that is not covered entirely by the r-disks around its vertices. At this time a loop is 'born', yielding a birth time b for this feature. It will also 'die' again, when r is further increased and the disks centered at the corners cover the entire polygon. Note that not all polygons that appear correspond to the birth of a loop. For instance, in the case of triangles, acute triangles always generate a loop while obtuse triangles do not. The resulting collection points $\{(b_i, d_i)\}_i$ can be used as coordinates and plotted in the plane, thereby giving rise to the *persistence diagram*. Since the topological features (connected components, loops) can only die 'after' they are born ($d \geqslant b$), necessarily each appears above the diagonal line in the first quadrant.

While the persistence diagram gives a concise representation for the topology of a single data set, in applications, we often deal with multiple measurements. Then, it is no longer sufficient to compute several persistence diagrams and study them in isolation since we have to understand the evolution over the course of the multiple measurements. This need was the key motivation leading to the concept of the *persistence vineyards* introduced in Cohen-Steiner et al. (2006). In Munch

Table 1
Parameter choices for the alternatives.

	HC_1	HC_2	CL_1	CL_2	CL_3
R	5.25	5.95	42.5	42.5	42.5
n_{cl}	/	/	10	5	4
λ_{cl}	/	/	10	20	25

(2013, Corollary 13), it is shown that if the persistence vineyard originates from a point cloud that varies continuously, then also the persistence vineyard evolves continuously in the bottleneck distance, which is one of the most widely used metrics for persistence diagrams.

In many applications, it is desirable to not only describe the evolution of the persistence diagrams in the persistence vineyard but also to track the motion of individual topological features in these diagrams. These evolving features are known as *vines*. However, in general, there is no universally applicable formal method to extract the individual vines from the vineyard. Although the conceptual idea of a vine is shared among all applications, the technical implementation may differ slightly depending on the context (Cohen-Steiner et al., 2006; Yoo et al., 2016; Salch et al., 2021b). In Section 2.3, we describe our approach to computing vines for 0-features and 1-features.

5. Simulation study

In this section, we elucidate through a simulation study how to design a goodness-of-fit test based on the test statistics from Section 2.3. To that end, in Section 5.1, we first describe the general set-up of this study including the considered null model, alternatives and test statistics. Then, Section 5.2 provides an exploratory analysis illustrating the distribution of the proposed test statistics under the null model and the alternatives. Finally, Section 5.3 contains a more detailed and structured investigation concerning the test power.

5.1. Simulation set-up

First, we describe the general set-up of the simulation study. That is, we present the null model, the alternatives as well as specific choices for test statistics.

5.1.1. Null model

Recalling the setting of Section 3, we study a null model for the trajectories traced by the centroids of the 2D-slices of a 3D-Poisson Voronoi tessellation. For this simulation study, the null model is a Poisson Voronoi tessellation with cell centers given by a Poisson process with intensity $\lambda = 2.18 \cdot 10^{-4}$ in a $170 \times 170 \times 85$-sampling window. The window height and the intensity of the point process are chosen to mimic the data set in Section 6. Due to the asymptotic normality, it suffices to compute mean and variances under the null model in the chosen sampling window, whose extent in the x-y-directions is far smaller than that found in the real data set. After constructing the tessellation, we take 9 slices parallel to the x-y-plane with a fixed spacing of 4 between them. The choice of the number of slices and of the spacing among them is guided by the exigency of mimicking a real serial slicing procedure, explained in more details in Section 6.

5.1.2. Alternatives

In order to design a set of alternatives to the null model, we retain the basic assumption of taking multiple 2D slices of a 3D Voronoi tessellation but vary the set of cell generators. More precisely, we deviate from the Poisson model by considering two classes of point processes with interactions, namely a more regular and a more clustered point pattern. We refer the reader to Chiu et al. (2013) for an overview of repulsive and clustered point processes.

As alternative to the null hypothesis of the Poisson process, we consider:

- ▷ A *Matérn hard-core process*, which is based on a Poisson point process where any two points at distance smaller than a threshold $R > 0$ are removed.
- ▷ A *Matérn cluster process*, where we first distribute a certain number n_{cl} of cluster centers in the window. Each cluster center generates offspring points according to a Poisson point process with intensity λ_{cl} inside a ball of radius R.

We work with two different parameter sets for the Matérn hard-core and three different parameter sets for the Matérn cluster point process, see Table 1. The parameters of the different point processes are chosen so as to lead to the same expected number of points in the sampling window.

Fig. 4 illustrates the effects of varying the cell generators on a single 2D slice of the 3D Voronoi tessellation. We see that this variation induces rather subtle changes in the sliced cells, which look fairly similar under the different alternatives.

Furthermore, we compute the persistence diagrams of the different 2D slices using the centroids of the different slices (shown in Fig. 5) as input point cloud. This illustrates that although typically the center locations are rather close between the different slices, occasionally larger movements can be noted.

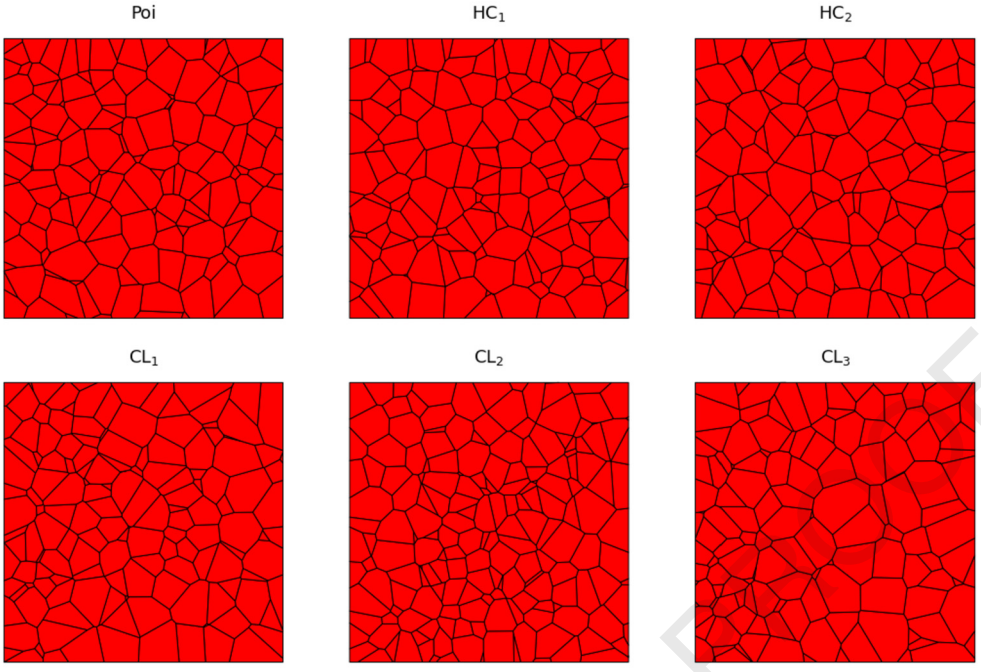

Fig. 4. 2D slices of Voronoi tessellations generated from different point processes for the generator points.

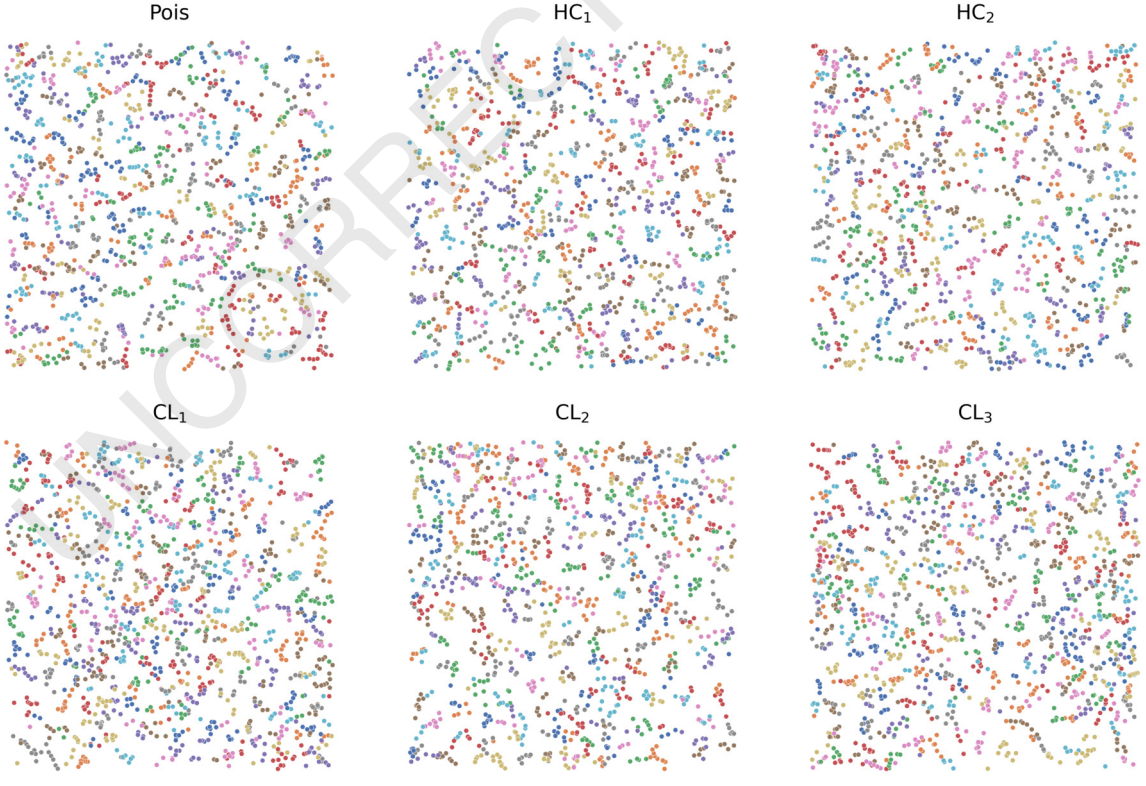

Fig. 5. x-y coordinates of the centroids of the 2D sliced grains of the 3D Voronoi tessellations on the considered point processes. Different colors represent the different slices.

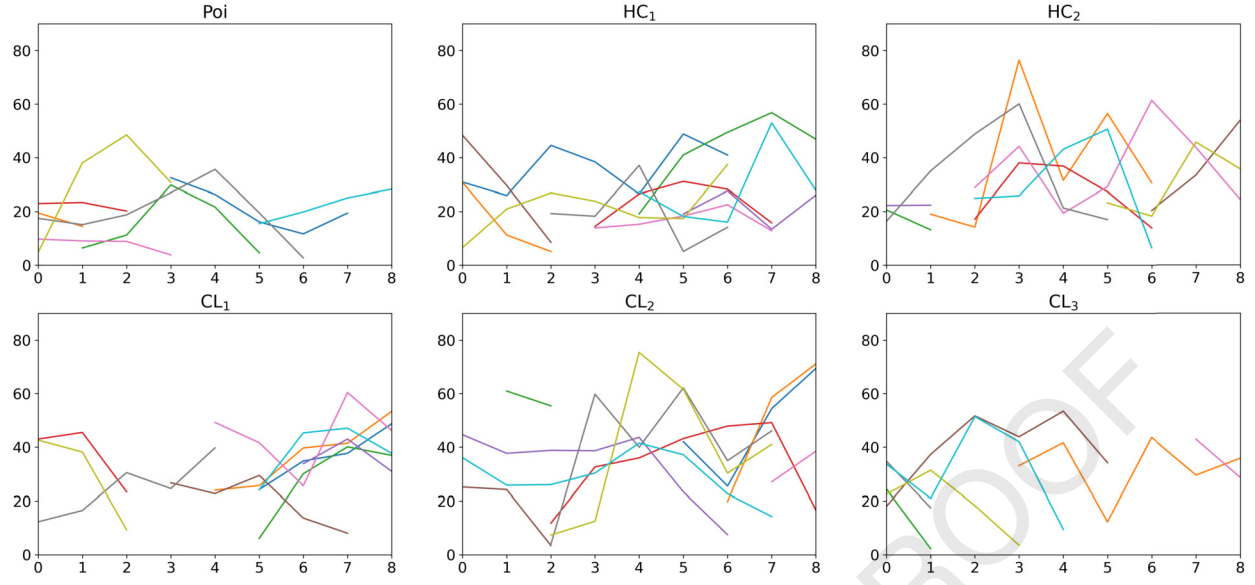

Fig. 6. Samples of vines in dimension 0. Lines indicating the same persistence point observed in more than one slice. The x-axis denotes the slice in which the diagram is computed. The y-axis denotes the death time of the considered feature.

5.1.3. Test statistics

To test deviations of data from the null model, we rely on three different test statistics.

1. **Cross-sectional total persistence.** To begin with, we present an example of a cross-sectional statistics in the sense of Example 2.1. In words, we extract global information from each of the persistence diagrams computed for every slice and then aggregate these quantities into a single characteristic for the dataset. More specifically, we put

$$T^q_{\mathsf{TP}} := \frac{1}{H}\sum_{h\leqslant H}\frac{1}{|W|}\sum_{i\leqslant n_h}\left(D^q_i(h) - B^q_i(h)\right),$$

where H is the total number of slices and n_h the total number of features in slice h. That is, in the definition (2), we take $\xi'(b,d) := d-b$ and then normalize the resulting statistic by the volume of the sampling window.

2. **Vine-based persistence.** Next, we discuss an example of a general longitudinal vine-statistic in the form of (1). Here, we average the life times of the features represented by the vine in the different slices over all slices where this vine is present. More precisely, we put

$$T^q_{\mathsf{M}} := \frac{1}{|W|}\sum_{i\leqslant N^q}\frac{1}{n^q_{h(i)}}\sum_{h\leqslant H}(D^q_i(h) - B^q_i(h)),$$

where N^q is the total number of unique features of dimension q observed in all the slices H and $n^q_{h(i)}$ is the number of slices in which the ith q-feature is visible. In order to represent T^q_{M} in the general form of (1), we may choose $\xi\left(\{(B^q_i(h), D^q_i(h))\}_h\right) := \frac{1}{n^q_{h(i)}}\sum_{h\leqslant H}(D^q_i(h) - B^q_i(h))$.

3. **Ripley K-function.** Finally, in order to compare TDA-based test statistics to classical test statistics from spatial statistics, we discuss an example derived from Ripley's K-function. More precisely, we let

$$T_{\mathsf{Rip}} := \int_0^{r_{\mathsf{Rip}}} \hat{K}_{\mathsf{pool}}(r)\mathrm{d}r,$$

where, $\hat{K}_{\mathsf{pool}}(r)$ is the pooled Ripley's K-function that takes the average of the estimates of the Ripley's K-function in each considered slice.

5.2. Exploratory analysis

Now, we consider the setup of Example 3. First, Figs. 6 and 7 respectively illustrate the 0-vines and 1-vines extracted from the persistence vineyard, obtained combining the persistence diagrams of each slice computed using the centroids

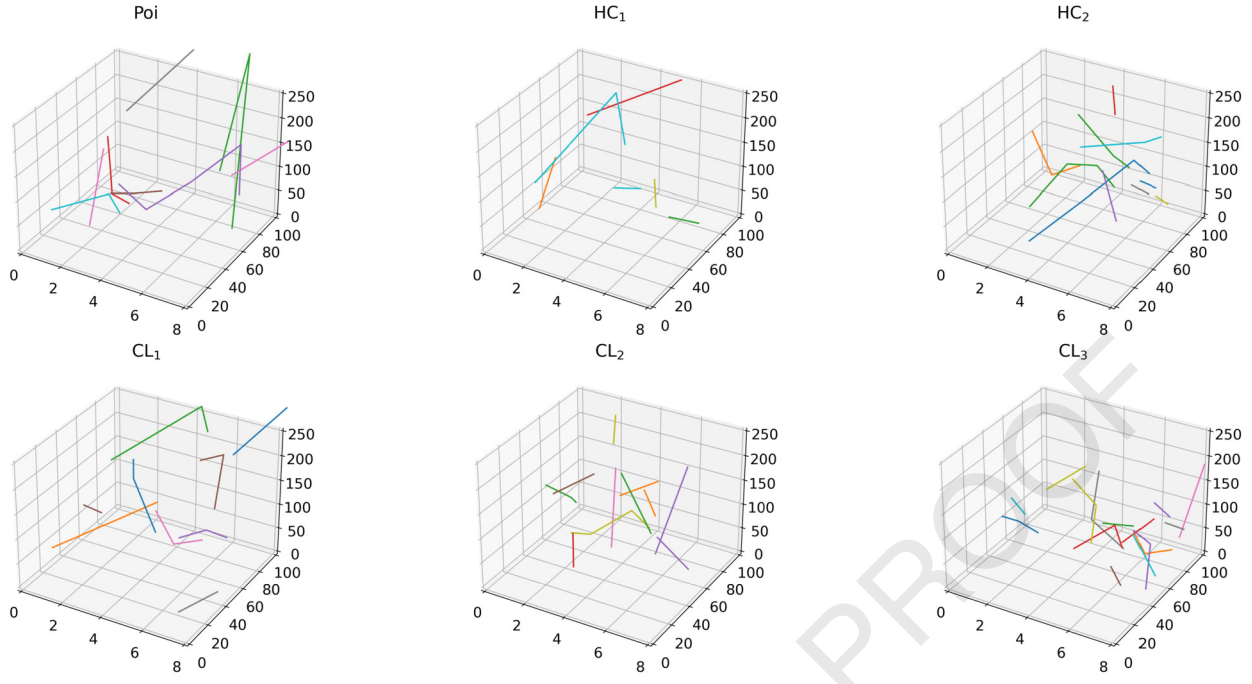

Fig. 7. Samples of vines in dimension 1. Lines indicating the same persistence point observed in more than one slice. The x-axis denotes the slice in which the diagram is computed. The y-axis denotes the birth time of the considered feature. The z-axis denotes the death time of the considered feature.

Table 2
Left: Mean (std. dev.) of the vine length for the different Voronoi tessellations (values based on 5,000 diagrams); Right: Mean (std. dev.) of the different test statistics in the Poisson-Voronoi model.

	$\dim = 0$	$\dim = 1$
PV	2.771 (1.836)	0.636 (0.878)
HC_1	2.768 (1.824)	0.643 (0.877)
HC_2	2.764 (1.821)	0.643 (0.876)
CL_1	2.785 (1.851)	0.635 (0.886)
CL_2	2.775 (1.854)	0.633 (0.888)
CL_3	2.771 (1.856)	0.630 (0.891)

	$170 \times 170 \times 85$	$140 \times 140 \times 140$
T^0_{TP}	1.618 (0.016)	1.629(0.021)
T^1_{TP}	0.899 (0.016)	0.844(0.020)
T^0_{M}	0.397 (0.008)	0.401(0.010)
T^1_{M}	0.503 (0.017)	0.474(0.020)
T_{Rip}	355.507 (6.137)	331.654(7.277)

shown in Fig. 5. We see that both in the null model and in the alternatives the 0-vines occur in a variety of different lengths and present different trends. On the other hand, 1-vines are much shorter.

This impression is also reinforced in Table 2, which shows the mean and the standard deviation of the vine length (L_v) for the different Voronoi models. The lengths are rather similar between the alternatives but the average length of 0-vines exceeds that of 1-vines.

5.2.1. Mean and standard deviation under the null model

The mean and variance of T^q_{TP} and T^q_{M} under the null model are computed using a simulation based on 5,000 Poisson-Voronoi tessellation for which 9 different slices with a spacing of 4 are considered.

By Theorem 2.4, the statistics T^q_{TP} and T^q_{M} are asymptotically normal so that knowing the mean and variance allows us to construct a deviation test whose nominal confidence level is asymptotically exact. Results on the asymptotic normality related to Ripley's K-function in the single section case are obtained in Heinrich (2015). We stress that our results on the asymptotic normality can be applied to a window of any fixed height as long as the size in the x- and y-directions are sufficiently large. On the other hand, it is also attractive to understand to what extent test statistics for windows of a given height generalize to other windows. This is not automatic since for fixed heights, the statistics may still be subject to finite-size effects. In Table 2 (right), we compare the test statistics in the original $170 \times 170 \times 85$-window to the one computed in a cube of side length 140. We see that when moving to the cubical window, the mean-values for the statistics corresponding to features deviate by around 7% or less from the values obtained in the original window.

As a conclusion, we note that the approximation resulting from ignoring effects coming from a finite height is already accurate for moderately large sampling windows. We also note that in a different context, edge effects and the dependence on the number of observed cells in 2D was observed in the analysis in Hahn and Lorz (1993); Vittorietti et al. (2020). There,

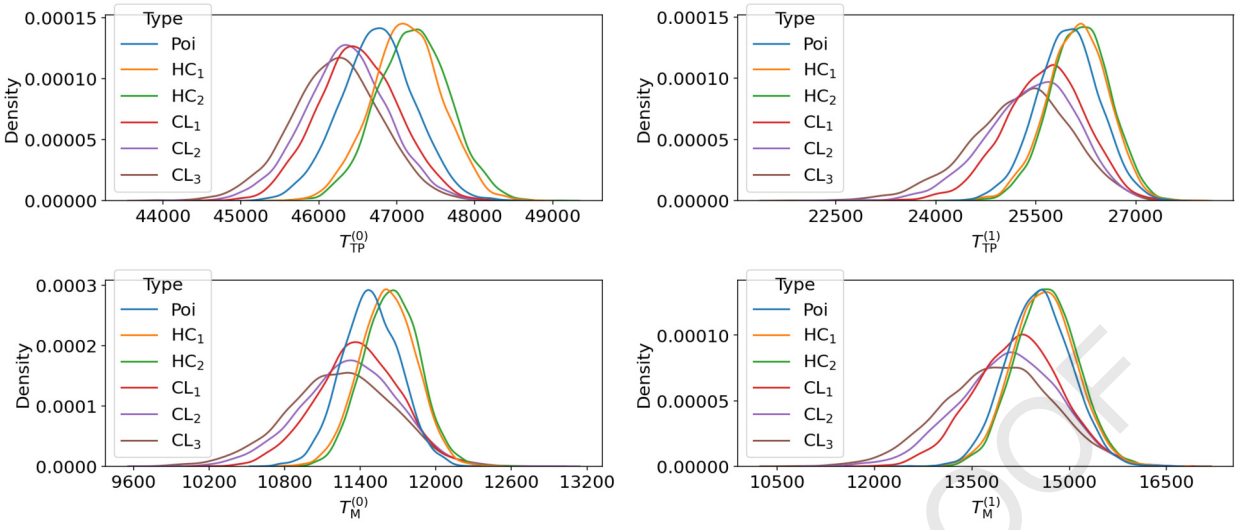

Fig. 8. Probability densities of the test statistics for the null model (blue) and the alternatives based on 5,000 realizations under the null model and the alternatives. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 3
Rejection rates for the multi-slice test statistics under the null model and the alternatives.

T	PV	HC_1	HC_2	CL_1	CL_2	CL3
T_{TP}^0	5.08%	11.96%	15.94%	10.70%	16.58%	24.50%
T_{TP}^1	5.30%	5.10%	5.64%	18.24%	28.77%	38.01%
T_{M}^0	5.00%	8.30%	9.46%	18.84%	27.47%	34.85%
T_{M}^1	4.74%	5.10%	5.02%	19.92%	27.83%	37.15%
T_{Rip}	4.94%	7.82%	9.76%	52.01%	70.17%	79.50%

this dependency is stated and the quantiles of the test statistics for different sample sizes (e.g. 50, 100, 150 visible cells in 2D) are reported.

Fig. 8 compares the probability density functions of the test statistics computed under the null and the alternative models. For all of the considered test statistics, the distributions of the Matérn cluster patterns differ clearly from the one of the Poisson null model. On the other hand, the situation for the Matérn hard-core patterns is more subtle. While we can observe differences for features in degree 0, there is very little difference when considering features in degree 1.

5.3. Power analysis

Before presenting the results, we explain in detail how to run the test. First, we choose a significance level of $\alpha = 0.05$. Second, we generate 5,000 realizations from the null model and from the alternatives. The Voronoi tessellations under the alternatives are the ones described at the beginning of the section. To analyze the type I and II errors, we draw Table 3 that contains the rejection rates of this test setup.

Under the null model the rejection rates are close to the nominal 5%-level thereby illustrating that already for moderately large point patterns, the approximation by the Gaussian limit is accurate. For the Matérn hard-core point pattern, we see that the rejection rates for T_{TP}^0 are higher than those for T_{TP}^1, whereas for the Matérn cluster point patterns, the situation is reversed. Moreover, the rejection rates for the T_{TP}-statistics are slightly higher than those for the T_M-statistics. We also observe that the pooled Ripley-statistics T_{Rip} is very powerful for detecting the clustered point patterns. On the other hand for the hard-core point patterns, the test statistic T_{TP}^0 has a bit higher rejection rate.

5.4. Extensions and variations

In this section, we discuss three possible extensions and variations of the testing methodology described above.

5.4.1. Single-slice testing

To highlight the difference between testing based on multiple slices and testing based just on one slice, in Table 4 the rejection rates obtained using just one slice are reported: the test based on multiple slices presents higher rejection rates and is not too conservative.

Table 4
Rejection rates for the single-slice test statistics under the null model and the alternatives.

T	PV	HC_1	HC_2	CL_1	CL_2	CL_3
T^0_{TP}	4.70%	5.54%	6.88%	6.58%	7.94%	9.92%
T^1_{TP}	4.96%	5.02%	5.62%	9.20%	12.90%	17.16%
T_{Rip}	5.38%	5.10%	5.10%	25.55%	40.79%	52.77%

Table 5
Reconstruction error Err_{Rec} for the actual vines under the reconstruction algorithm. Rejection rates for the test statistics based on the reconstructions the null model and the alternatives.

	PV	HC_1	HC_2	CL_1	CL_2	CL3
Err_{Rec}	21.4%	20.6%	20.4%	21.6%	21.8%	21.9%
T^0_M	5.14%	5.04%	5.04%	8.08%	9.52%	12.96%
T^1_M	4.64%	4.12%	4.66%	14.36%	19.96%	27.07%

5.4.2. Cell vertices instead of centroids

Moreover, in the present simulation study, we constructed the persistence diagrams based on the centroid of the cell slices. A concern of this methodology is that subtle differences between point patterns could be lost when moving from the sliced Voronoi cells to the centroids. Therefore, we also experimented with computing the test statistics T^i_{TP} when computing the persistence diagram on the basis of the cell vertices instead of the cell centroids. However, we found that this modification did not improve the testing power. We hypothesize that this is due to subtle dependencies induced by the slicing procedure. Even if for 3D cells the centroid may not be highly informative, this situation changes when working with the sequence of centroids obtained from the multiple 2D slices.

5.4.3. Labeling algorithm

In real data we have to deal with the problem of identifying to which grain the centroid belongs. Often, additional information on the orientation of the grains helps reconstructing the inner grain structures of the slices. Alternatively, we propose a simple algorithm for the labels assignment. We start by computing the pairwise distance matrix between the set of centroids that belong to adjacent slices. We then assign the label of the points of the first slice to the points of the second slice if and only if the distance between two points does not exceed a specific threshold and it is minimum with respect to all the other points.

To assess the effectiveness of this algorithm, in Table 5, we present the reconstruction error of the vines for the different models under consideration. Although the simple algorithm does not lead to a perfect reconstruction, it still succeeds in recovering a substantial proportion of the vines across the models. Moreover, in Table 5 we also note a decrease in the testing power of T^0_M and T^1_M when using the reconstructed labels instead of the ground truth. This illustrates that the test statistics T^0_M and T^1_M are more useful in settings featuring a relatively densely arranged set of slices. Their power may deteriorate if the slices are so far apart that a precise vine reconstruction may no longer be feasible. Note that T^0_{TP}, T^1_{TP} and T_{Rip} instead are not affected by the reconstruction algorithm.

6. Analysis of material data

In this section, the use of Poisson-Voronoi diagrams for representing the microstructure of extra low carbon strip steel is tested. By relying on sequential slicing via automated mechanical polishing, 3D electron backscatter diffraction (EBSD) measurements of this material are carried out.

In this technique, consecutive steps of sample preparation and EBSD scanning are employed to obtain 2D EBSD slices. More details about the material and the technique can be found in Galán López and Kestens (2021).

6.1. Exploratory analysis

The left panel in Fig. 9 (taken by J. G. López (Galán López and Kestens, 2021)) illustrates a sample of the steel data consisting of 20 EBSD scans with a spacing of 4 µm. On average $\approx 2,700$ grains ($sd \approx 120$) are observed in each 2D EBSD scan. In order to reduce edge effects, we proceed as in Section 5 and restrict the computation to nine 2D EBSD scans taken in the middle of the surface block. In Fig. 9 (right) the cell centroids of the grains visible in nine 2D EBSD scans are plotted (a reduced area is considered for visualization purposes). Each set of centroids is then used for building a persistence diagram (Fig. 10).

In total $24,008$ distinct features are observed in dimension 0 and $31,474$ in dimension 1. The mean "vine" length is 0.021243 and 0.0004, highlighting that only few features are actually observed in more than one slice.

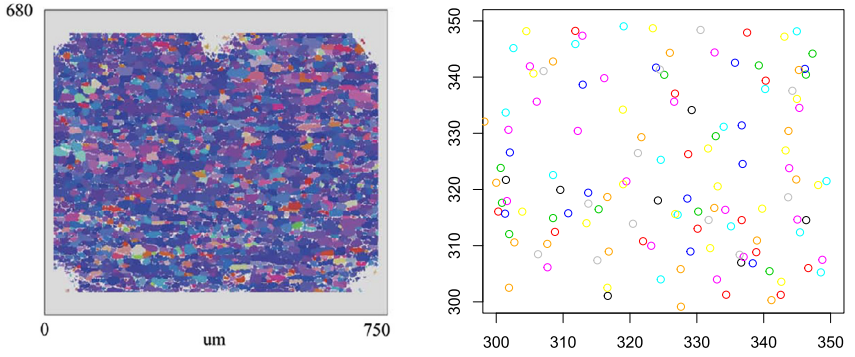

Fig. 9. Experimental EBSD data (left), and x–y coordinates of centroids of the 2D sliced cells of nine 2D EBSD scan (right). Different colors represent the different slices.

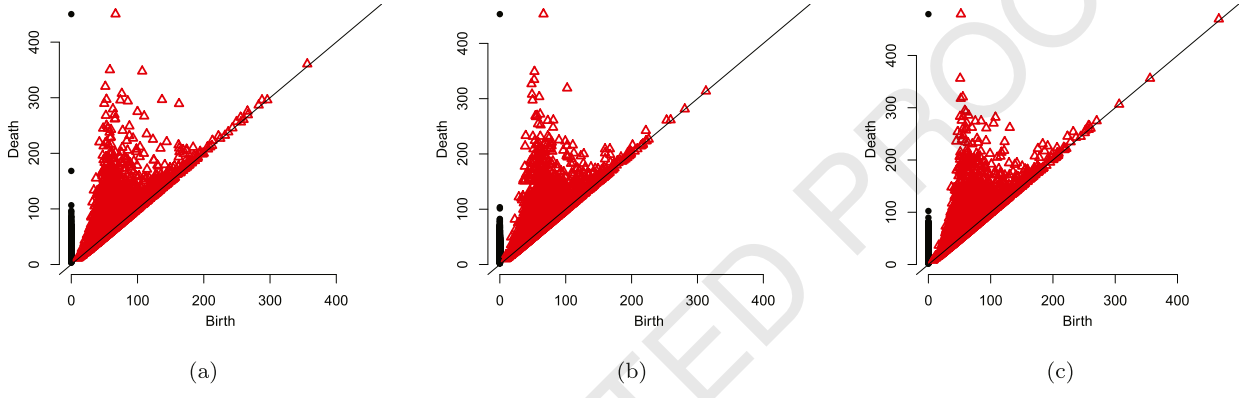

Fig. 10. Persistence diagrams corresponding to three different 2D EBSD slices based on the point clouds of centroids of the 2D sectional cells.

Table 6
z-scores associated with the different test statistics when the 2D EBSD slices are tested against a Poisson-Voronoi null model.

	T^0_{TP}	T^1_{TP}	T^0_M	T^1_M	T_{Rip}
z-score	23.00	5.40	94.75	30.05	81.8

6.2. Poisson-Voronoi testing

With the previous visualizations in mind, we now test the Poisson Voronoi hypothesis for the EBSD data. As in the previous section, we use the asymptotic normality of the test statistics under the null model. The test statistics are reported in Table 6.

First, we note that all considered test statistics suggest a highly significant deviation of the data from the 3D null model of a Poisson-Voronoi tessellation. Second, we see that the test statistics T^0_M and T^1_M lead to z-scores that exceed by far the test scores corresponding to T^0_{TP} and T^1_{TP}. This effect points to the potential that lies in the longitudinal statistics. Indeed, as observed at the beginning of the present section, on average, the vines in the data set are far shorter than the ones observed in the null model. Thus, the deviations from the null model are far more pronounced in the longitudinal test statistics T^0_M and T^1_M than they are in the cross-sectional test statistics T^0_{TP} and T^1_{TP}.

7. Conclusion and perspective

In this work, we applied the concept of persistence vineyards to develop goodness-of-fit tests for data given in the form of 2D slices of a 3D data set. These tests rely on the asymptotic normality of the considered test statistics, which we establish under suitable moment and stabilization conditions. The potential of the new methodology is illustrated through a simulation study and a dataset describing the microstructure of extra low carbon strip steel. We stress that the present work is to be considered only as the first step towards using the tools from TDA to develop statistical tests for slice-based datasets. We discuss now possible avenues for further research.

In this work, we have focused on the asymptotic normality of persistent homology information in the form of persistence vineyards. We remark that recently various new persistent functions have been developed that can capture geometric,

topological and spectral properties of the underlying dataset. Two of the most prominent examples are the *Ollivier persistent Ricci curvature* and the *persistent spectral graph* (Wee and Xia, 2021; Wang et al., 2020). We now expound on the prospects and possible first steps to extend our methodology to these examples.

In classical differential geometry, the concept of Ricci curvature encodes how far the neighborhood of a point on a manifold differs from a flat Euclidean space. This idea was extended to the setting of graphs through the concept of *Ollivier Ricci curvature (ORC)* (Ollivier, 2009), and has been successfully applied in the structural analysis of internet networks (Ni et al., 2015). Loosely speaking, the ORC evaluated at an edge provides an indication of whether this edge acts as a bridge connecting different clusters of a network. Since the ORC only relies on neighborhood information of the edge, it is amenable to stabilization methods, and therefore to asymptotic normality on a scalar level. However, the question of whether the normality holds on the functional level requires a more refined analysis of the probability of observing small perturbations.

The idea of studying refined properties of the graph Laplacian has developed into an extremely fruitful research direction, commonly referred to as *spectral graph theory* (Chung, 1997). Since higher-order interactions cannot be adequately captured in the setting of graphs, simplicial complexes and the associated q-combinatorial Laplacian have recently gained substantial attention in the research community (Horak and Jost, 2013). The observation that the qth Betti number can be recovered as the dimension of the kernel of the q-combinatorial Laplacian illustrates the intimate relation between this Laplacian and the topology of the underlying complex. This connection has motivated the introduction of p-persistent combinatorial Laplacians, which extend the aforementioned recovery of Betti numbers to persistent Betti numbers. In general, it is challenging to analyze the persistent spectral graph statistically because large components in the complex could induce long-range correlations. A starting point could be to work in a sparse setting, where the simplicial complex decomposes into a collection of small components so that it can be statistically analyzed by stabilization methods. While this heuristic makes it plausible that proving asymptotic normality is in reach on a scalar level, a functional version would again require a more detailed perturbation analysis.

Moreover, due to the generality of our method, we envision further applications to other sliced data. For example, in materials science, it could be used to study the appearance and development of defects occurring in materials during mechanical tests. In Hidalgo et al. (2020), 2D pictures of material's slices are observed at different levels of tensile deformation to study the development of necessary geometrical dislocations. Moreover, our methods could be used to study *microstructural banding*, which is a key phenomenon in steel manufacturing. Bands in a material can cause the mechanical properties to be anisotropic, that is, to be different in different material's regions. In particular, band connectivity plays an important role in the mechanical properties of steel. Information about connectivity is completely lost when only a 2D image is considered. Our method could be used to study the connectivity of the bands in the direction of slicing and to quantify and compare the band connectivity of different materials (McGarrity et al., 2012).

Motivated by the shape of the specific dataset, we established the asymptotic normality for domains that are large in the x- and y-direction but are fixed in the z-direction. For other datasets, it may be that all three dimensions are large, and it would be of interest to extend the proof of asymptotic normality for such settings. This would have the additional benefit that under a Poisson-Voronoi null model, it suffices to compute the mean and variance of the test statistics in Section 5 for a fixed point-process intensity. The values for other intensities can then be obtained through a suitable scaling by a power of the intensity. It is also interesting to extend the results to higher dimensions, power-law correlations and other stochastic models for representing microstructures.

Also our decision to choose the centroids of the sliced cells as a basis for the persistence computation is motivated from the constraints found in the considered dataset. As illustrated in Fig. 9, the cells in the material data can be shaped rather irregularly. Since the cell centroids are fairly robust with respect to small misspecifications of the cell shapes, they are attractive candidates for the data at hand. In Section 5, we also pointed out that computing the persistence diagram via the tessellation vertices did not improve the testing power – at least for the models considered in the simulation study. However, in other scenarios, the tests based on the centroids could be outperformed when replacing the cell centroids with other point clouds extracted from the slices.

Moreover, as pointed out in Section 1, instead of computing test statistics related to topological characteristics of 2D slices, one may also attempt to reconstruct faithfully certain topological characteristics of the 3D data set. However, this construction relies on potentially delicate density and transversality conditions. It is interesting to investigate whether the longitudinal and cross-sectional tests from our framework also become more powerful under such conditions.

Finally, it would be exciting to generalize to higher-dimensional situations such as snapshots of 3D data evolving in time. The main challenge here lies in generalizing the concept of M-bounded features which currently relies on the duality in two dimensions.

Acknowledgements

We thank the anonymous referees for their detailed reading of the manuscript. Their suggestions helped to improve the presentation substantially, and to make our work accessible to a broader readership. We thank J. G. López for providing the data. AC conducted this research while affiliated with TU Delft, and was supported by the Netherlands Organisation for Scientific Research (NWO) through grant 613.009.102. CH acknowledges the financial support of the CogniGron research center and the Ubbo Emmius Funds (University of Groningen). MV acknowledges the financial support of the Research Program of the Materials innovation institute (M2i) (www.m2i.nl) supported by the Dutch government.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2022.107655.

Uncited references

Baryshnikov and Yukich (2005) Bickel and Wichura (1971) Billingsley (1999) Davydov and Zitikis (2008) Divol and Polonik (2019) Eichelsbacher et al. (2015) Krebs and Hirsch (2022) McLeish (1974) Peccati and Taqqu (2011) Penrose (2003)

References

Amini, O., Boissonnat, J.D., Memari, P., 2013. Geometric tomography with topological guarantees. Discrete Comput. Geom. 50, 821–856.
Baryshnikov, Y., Yukich, J.E., 2005. Gaussian limits for random measures in geometric probability. Ann. Appl. Probab. 15, 213–253.
Bickel, P.J., Wichura, M.J., 1971. Convergence criteria for multiparameter stochastic processes and some applications. Ann. Math. Stat. 42, 1656–1670.
Billingsley, P., 1999. Convergence of Probability Measures, second ed. John Wiley & Sons, New York.
Biscio, C.A.N., Chenavier, N., Hirsch, C., Svane, A.M., 2020. Testing goodness of fit for point processes via topological data analysis. Electron. J. Stat. 14, 1024–1074.
Błaszczyszyn, B., Yogeshwaran, D., Yukich, J.E., 2019. Limit theory for geometric statistics of point processes having fast decay of correlations. Ann. Probab. 47, 835–895.
Botnan, M.B., Lesnick, M., 2022. An introduction to multiparameter persistence. arXiv preprint arXiv:2203.14289.
Chiu, S.N., Stoyan, D., Kendall, W.S., Mecke, J., 2013. Stochastic Geometry and Its Applications, third ed. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.
Chung, F.R.K., 1997. Spectral Graph Theory. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI.
Cohen-Steiner, D., Edelsbrunner, H., Morozov, D., 2006. Vines and vineyards by updating persistence in linear time. In: Computational Geometry (SCG'06). ACM, New York, pp. 119–126.
Davydov, Y., Zitikis, R., 2008. On weak convergence of random fields. Ann. Inst. Stat. Math. 60, 345–365.
Divol, V., Polonik, W., 2019. On the choice of weight functions for linear representations of persistence diagrams. J. Appl. Comput. Topol. 3, 249–283.
Edelsbrunner, H., Harer, J., 2010. Computational Topology. American Mathematical Society, Providence, RI.
Eichelsbacher, P., Raič, M., Schreiber, T., 2015. Moderate deviations for stabilizing functionals in geometric probability. Ann. Inst. Henri Poincaré Probab. Stat. 51, 89–128.
Galán López, J., Kestens, L.A., 2021. A multivariate grain size and orientation distribution function: derivation from electron backscatter diffraction data and applications. J. Appl. Crystallogr. 54.
Hahn, U., Lorz, U., 1993. Stereological model tests for the spatial Poisson-Voronoi tessellation II. Acta Stereol. 12, 131–140.
Heinrich, L., 2015. Gaussian limits of empirical multiparameter K-functions of homogeneous Poisson processes and tests for complete spatial randomness. Lith. Math. J. 55, 72–90.
Hidalgo, J., Vittorietti, M., Farahani, H., Vercruysse, F., Petrov, R., Sietsma, J., 2020. Influence of M23C6 carbides on the heterogeneous strain development in annealed 420 stainless steel. Acta Mater. 200, 74–90.
Horak, D., Jost, J., 2013. Spectra of combinatorial Laplace operators on simplicial complexes. Adv. Math. 244, 303–336.
Krebs, J.T.N., Hirsch, C., 2022. Functional central limit theorems for persistent Betti numbers on cylindrical networks. Scand. J. Stat. 49, 427–454.
Last, G., Penrose, M.D., 2016. Lectures on the Poisson Process. Cambridge University Press, Cambridge.
Madej, L., 2017. Digital/virtual microstructures in application to metals engineering–a review. Arch. Civ. Mech. Eng. 17, 839–854.
McGarrity, K.S., Sietsma, J., Jongbloed, G., 2012. Characterisation and quantification of microstructural banding in dual-phase steels part 2–case study extending to 3D. Mater. Sci. Technol. 28, 903–910.
McLeish, D.L., 1974. Dependent central limit theorems and invariance principles. Ann. Probab. 2, 620–628.
Munch, E., 2013. Applications of Persistent Homology to Time Varying Systems. Ph.D. thesis.
Ni, C.C., Lin, Y.Y., Gao, J., David Gu, X., Saucan, E., 2015. Ricci curvature of the Internet topology. In: 2015 IEEE Conference on Computer Communications (INFOCOM), pp. 2758–2766.
Ollivier, Y., 2009. Ricci curvature of Markov chains on metric spaces. J. Funct. Anal. 256, 810–864.
Peccati, G., Taqqu, M.S., 2011. Wiener Chaos: Moments, Cumulants and Diagrams. Springer/Bocconi University Press, Milan/Milan.
Penrose, M.D., 2003. Random Geometric Graphs. Oxford University Press, Oxford.
Penrose, M.D., Yukich, J.E., 2001. Central limit theorems for some graphs in computational geometry. Ann. Appl. Probab. 11, 1005–1041.
Pirgazi, H., 2019. On the alignment of 3D EBSD data collected by serial sectioning technique. Mater. Charact. 152, 223–229.
Reani, Y., Bobrowski, O., 2021. Cycle registration in persistent homology with applications in topological bootstrap. arXiv preprint arXiv:2101.00698.
Redenbach, C., Shklyar, I., Andrä, H., 2012. Laguerre tessellations for elastic stiffness simulations of closed foams with strongly varying cell sizes. Int. J. Eng. Sci. 50, 70–78.
Salch, A., Regalski, A., Abdallah, H., Suryadevara, R., Catanzaro, M.J., Diwadkar, V.A., 2021a. From mathematics to medicine: a practical primer on topological data analysis (TDA) and the development of related analytic tools for the functional discovery of latent structure in fMRI data. PLoS ONE 16, 1–33.
Salch, A., Regalski, A., Abdallah, H., Suryadevara, R., Catanzaro, M.J., Diwadkar, V.A., 2021b. From mathematics to medicine: a practical primer on topological data analysis (TDA) and the development of related analytic tools for the functional discovery of latent structure in fMRI data. PLoS ONE 16, e0255859.
Tewari, A., Gokhale, A.M., 2001. Estimation of three-dimensional grain size distribution from microstructural serial sections. Mater. Charact. 46, 329–335.
Vittorietti, M., Kok, P.J.J., Sietsma, J., Li, W., Jongbloed, G., 2020. General framework for testing Poisson-Voronoi assumption for real microstructures. Appl. Stoch. Models Bus. Ind. 36, 604–627.
Wang, R., Nguyen, D.D., Wei, G.W., 2020. Persistent spectral graph. Int. J. Numer. Methods Biomed. Eng. 36, e3376.
Wasserman, L., 2018. Topological data analysis. Annu. Rev. Stat. Appl. 5, 501–535.
Wee, J., Xia, K., 2021. Ollivier persistent Ricci curvature-based machine learning for the protein–ligand binding affinity prediction. J. Chem. Inf. Model. 61, 1617–1626.
Yoo, J., Kim, E.Y., Ahn, Y.M., Ye, J.C., 2016. Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages. J. Neurosci. Methods 267, 1–13.