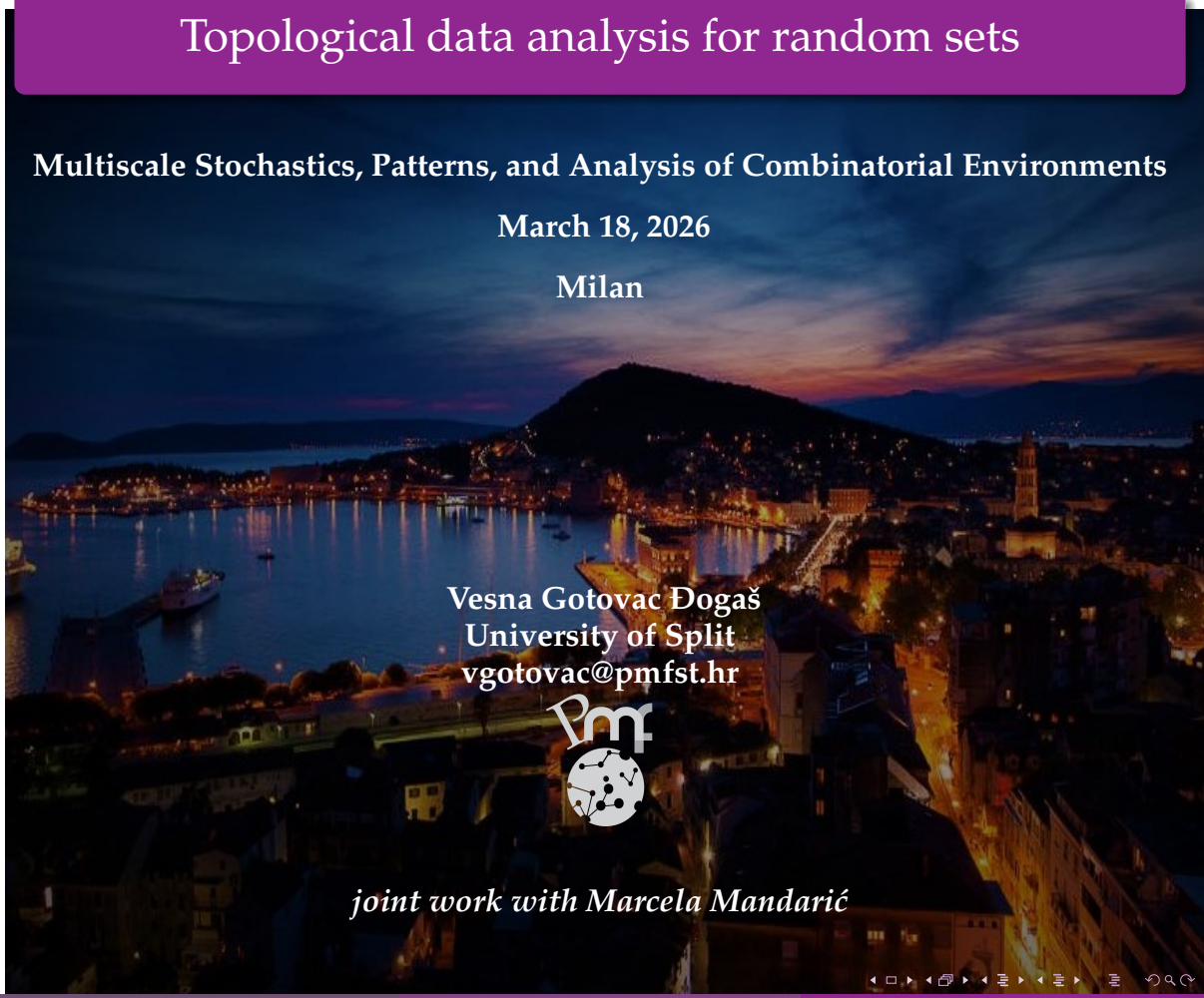
Topological data analysis for random sets

Multiscale Stochastics, Patterns, and Analysis of Combinatorial Environments

March 18, 2026

Milan



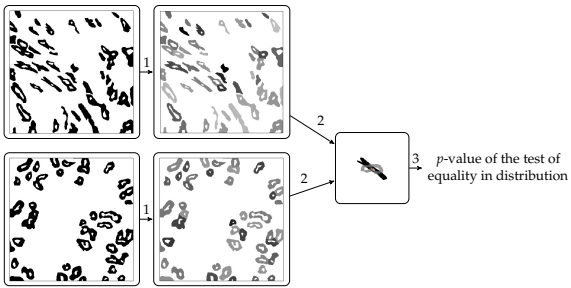
Vesna Gotovac Đogaš
University of Split
vgotovac@pmfst.hr

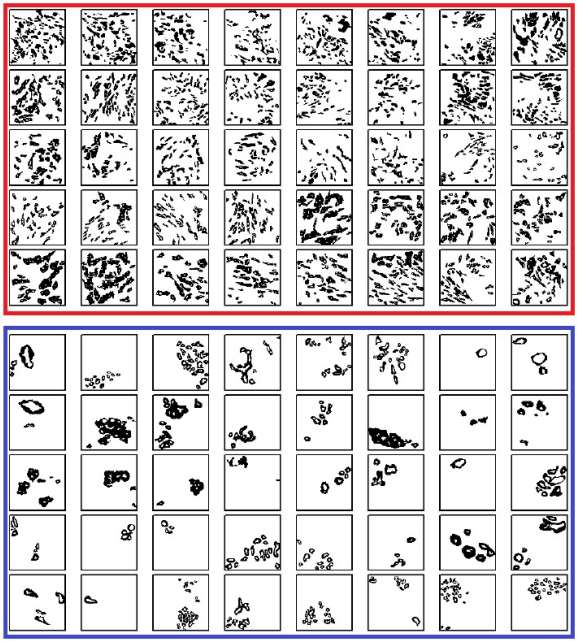
joint work with Marcela Mandarić

Data and problems considered

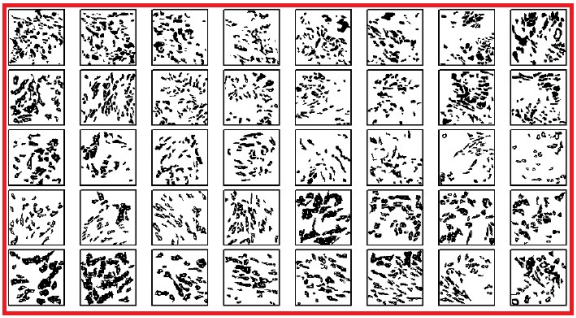
- Testing goodness of fit (comparing one realisation with the group of realisations)
- Classification (partitioning the group of realisations)

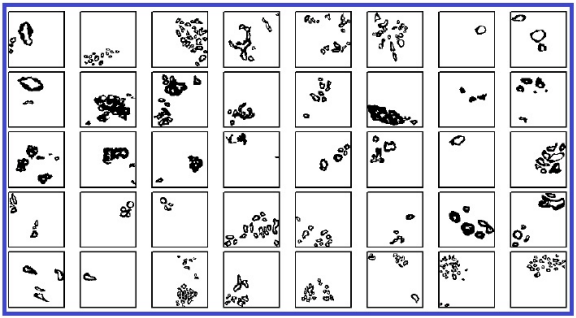
Two-sample problem (comparing just two realisations)





Motivation- Investigating properties of random sets in terms of topological properties of the shapes and interaction between components





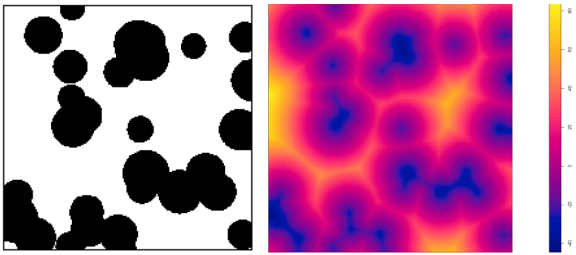
- Starting point: a sample or samples of realisations of random set.
- Statistical analysis concentrating on topological properties, size of the components and interactions between components (clustering, repulsion).
- Topological data analysis has proven successful in revealing the interactions and testing the goodness of fit of point processes [a]
- It seemed reasonable to generalise these techniques for random sets.

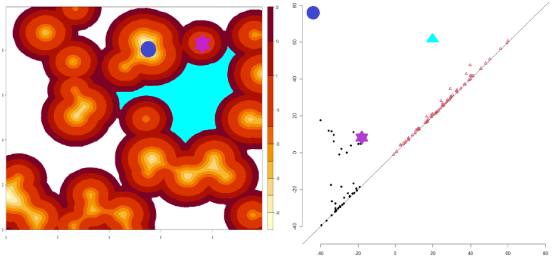
[a] BISCIO, C. A. N., CHENAVIER, N., HIRSCH, C. & SVANE, A. M. (2020). Testing goodness of fit for point processes via topological data analysis. *Electron J Stat* **14**(1), 1024–1074.

Overview

1. TDA for set data
2. Testing goodness of fit using summary functions
3. Application to real data

Topological data analysis





- Topological Data Analysis (TDA) studies the shape of data by turning it into a sequence of nested structures (filtration).
- It uses homology to detect features like holes, and persistent homology tracks when these features appear and disappear.
- These are visualized using persistence diagrams, where each point shows a feature's birth and death.

Construction of PD for set valued data

- Suppose we observe the realisations of our random set within the observation window $W \subseteq \mathbb{R}^2$.
- Let us consider the signed distance (to a set S) function defined as $f_d : W \to \mathbb{R}$,

$$f_d(x) = \begin{cases} d(x,S), & x \notin S, \\ -d(x, W \setminus S) & x \in S, \end{cases}$$

  where $d(x,S)$ is the distance from a point x to a set S.
- If we consider the sublevel sets of f_d,

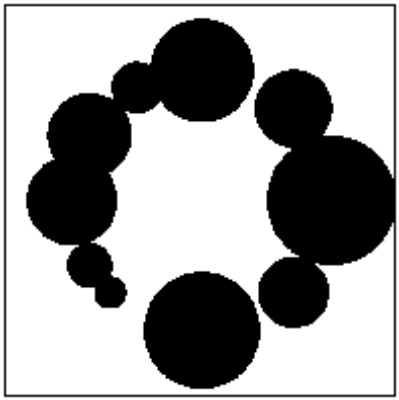
$$S_r = f_d^{-1}\left(\langle -\infty, r]\right),$$

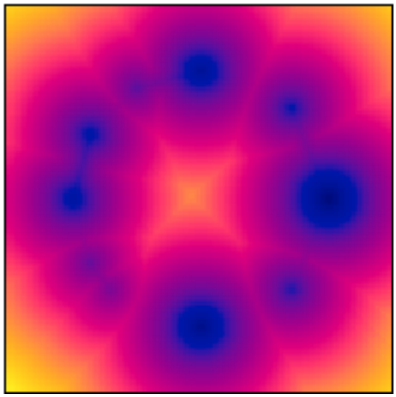
  we obtain a non-decreasing filtration $\{S_r\}_{r\in\mathbb{R}}$, from which we can construct a persistence diagram.
- Persistence diagram PD^q that captures features in dimension q is a multiset of points in the plane (b_i, d_i), with multiplicity $c_i \geq 1$, for $i \in \mathcal{I}_q$, where $\mathcal{I}_q$ is an index set of its distinct points.
- The persistence diagram can be regarded as an empirical measure

$$PD^q(X) = \sum_{i\in\mathcal{I}_q} c_i \delta_{(b_i,d_i)},$$

  where δ stands for Dirac delta measure which is equal to 1 if (b_i, d_i) is in persistence diagram and 0 otherwise.

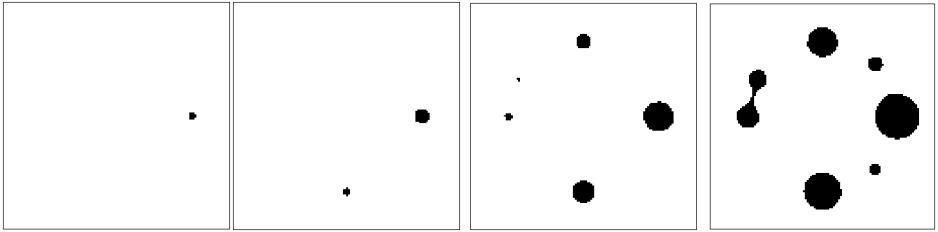
Example of construction of PD



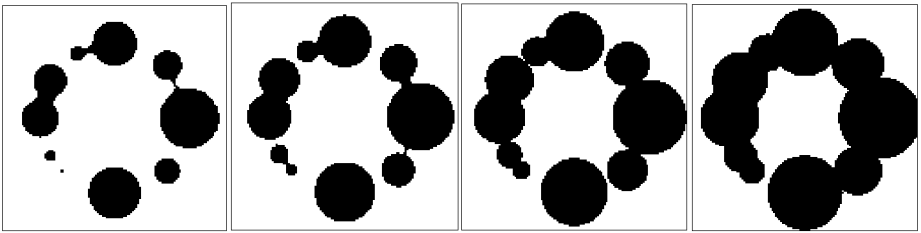


- Consider the union of the discs in Figure, whose centres lie on the circle around the origin with radius 200. The radii of the discs are
  $100, 90, 80, 70, 65, 60, 55, 40, 35, 25$.
- We can imagine these discs as the floor plans of tree crowns and assume that the growth of the trees is linear in time and that the same rate applies to all trees.
- Then the sublevel set S_r could be interpreted as the floor plan of the tree crowns at time r, and the points on the persistence diagram corresponding to 0-dimensional features capture the birth of each tree and the time at which the crown of that tree first touches a neighbouring tree that was born before.

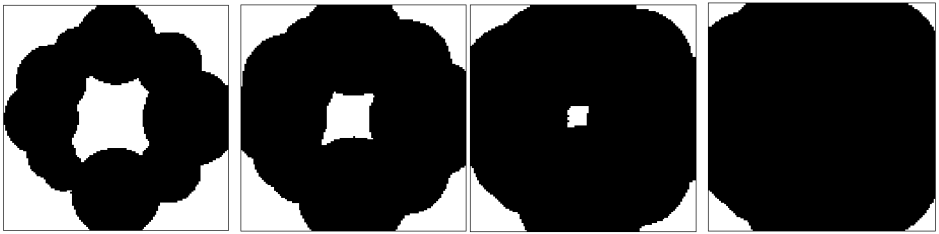
Sublevel set S_r



(a) $r = -90$ **(b)** $r = -80$ **(c)** $r = -60$ **(d)** $r = -40$

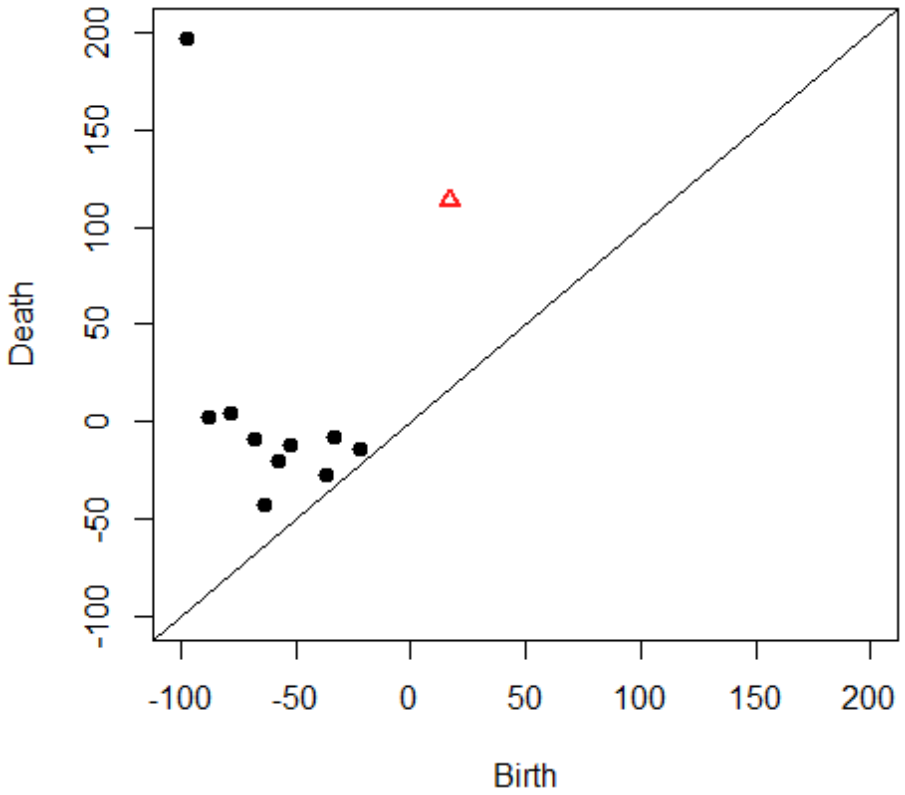


(e) $r = 20$ **(f)** $r = -10$ **(g)** $r = 0$ **(h)** $r = 10$

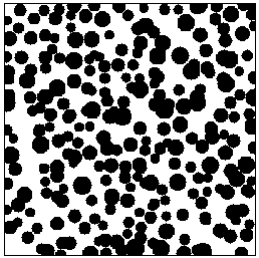


(i) $r = 30$ **(j)** $r = 60$ **(k)** $r = 90$ **(l)** $r = 120$

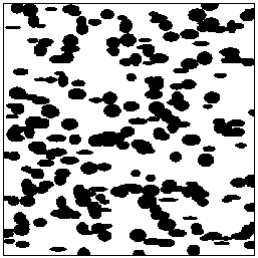
Perisistance diagram (black points for $q = 0$ and red triangle for $q = 1$)



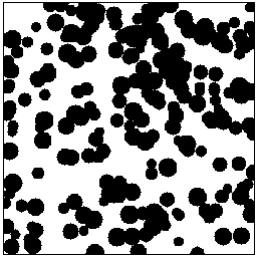
Simulation study: 7 random set models



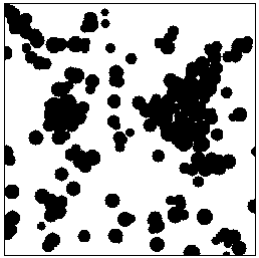
(a) Repulsive model



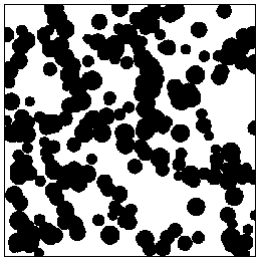
(b) Random-ellipse Boolean model



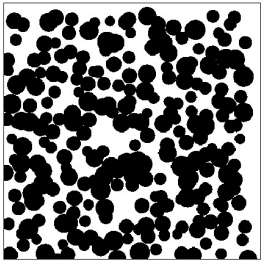
(c) Random-disc Boolean model



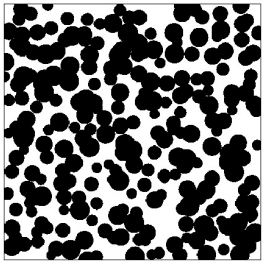
(d) Cluster model



(e) Matern cluster model



(f) Cell model



(g) DPP model

Standard summary functions: Capacity functional on the squares (CF) and extended empty space function (ESF)

Definition 1

A functional $T_{\mathbb{X}}: \mathcal{C} \to [0,1]$ given by

$$T_{\mathbb{X}}(K) = P(\{\omega : \mathbb{X}(\omega) \cap K \neq \emptyset\}), \quad K \in \mathcal{C},$$

is said to be *the capacity functional of random set* $\mathbb{X}$.

- The capacity functional uniquely determines the distribution of a random closed set.
- Since it is not possible in practice to derive the capacity functional for all $K \in \mathcal{C}$, we assume stationarity of the random set and consider the values of the capacity on the squares, $T_{\mathbb{X}}(r) = T_{\mathbb{X}}(rB)$, where B is a unit square.

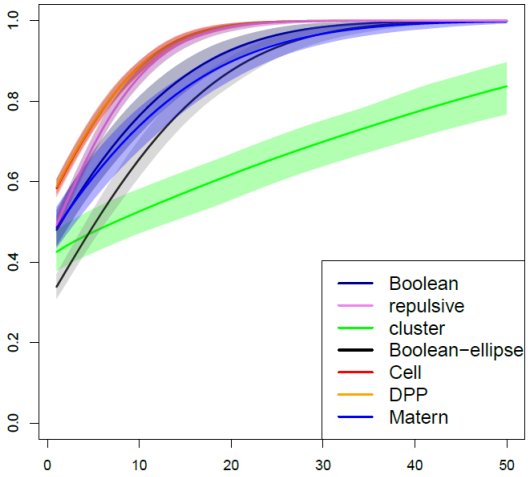
Definition 2

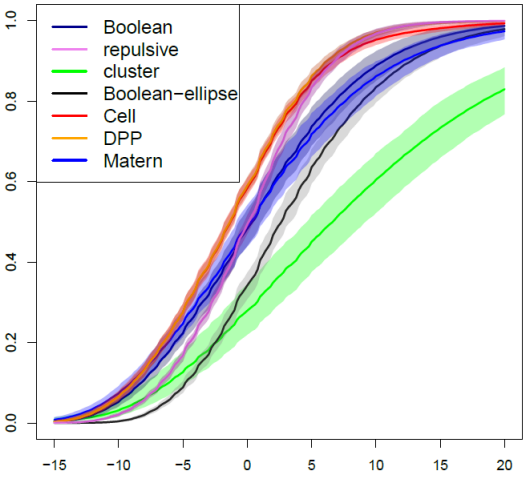
Function $F: \mathbb{R} \to [0,1]$ given by

$$F(r) = \begin{cases} \mathrm{P}(\mathbf{0} \in \mathbb{X} \oplus rB(\mathbf{0},1)) & \text{for } r \geq 0, \\ \mathrm{P}(\mathbf{0} \in \mathbb{X} \ominus |r|B(\mathbf{0},1)) & \text{for } r < 0, \end{cases}$$

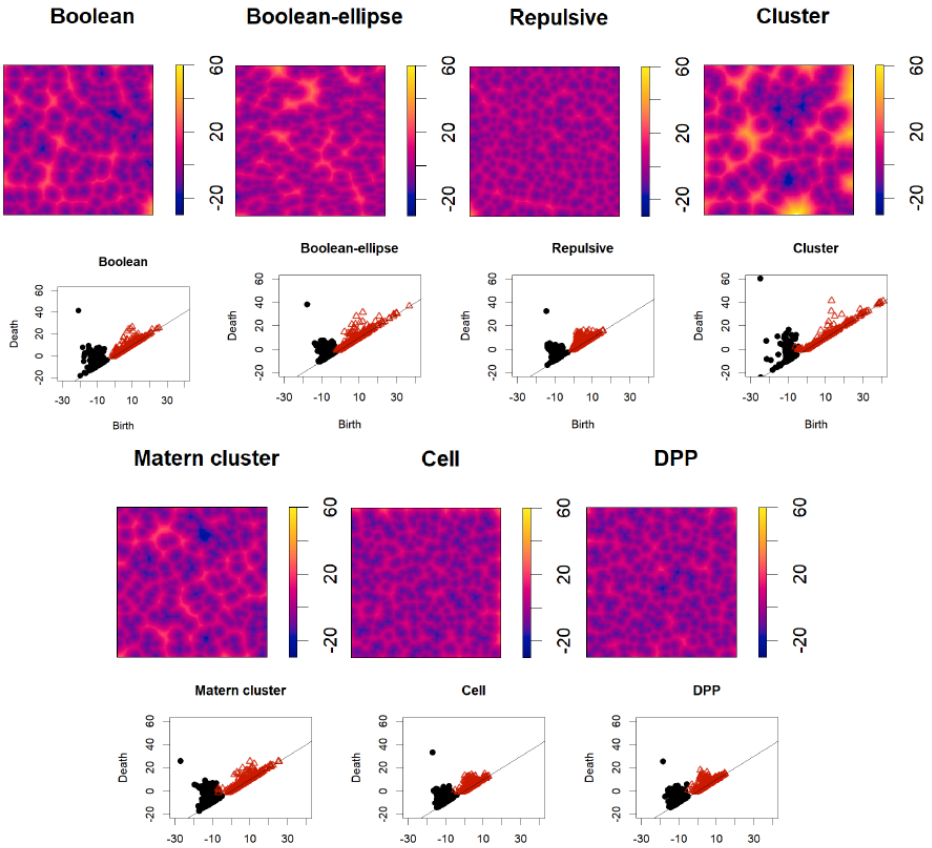
where $\mathbf{0}$ stands for the origin and $B(\mathbf{0},1)$ for the unit disc centred at the origin, is called *the extended empty space function* of the random set $\mathbb{X}$.

Mean capacity functional of each process at the left and mean extended empty space functions at the right together with their 95% envelopes obtained by simulating 100 realisations of each random set process.





The heat maps of the signed distance function of one realisation of each process with its persistence diagram



Accumulated persistence function (APF)

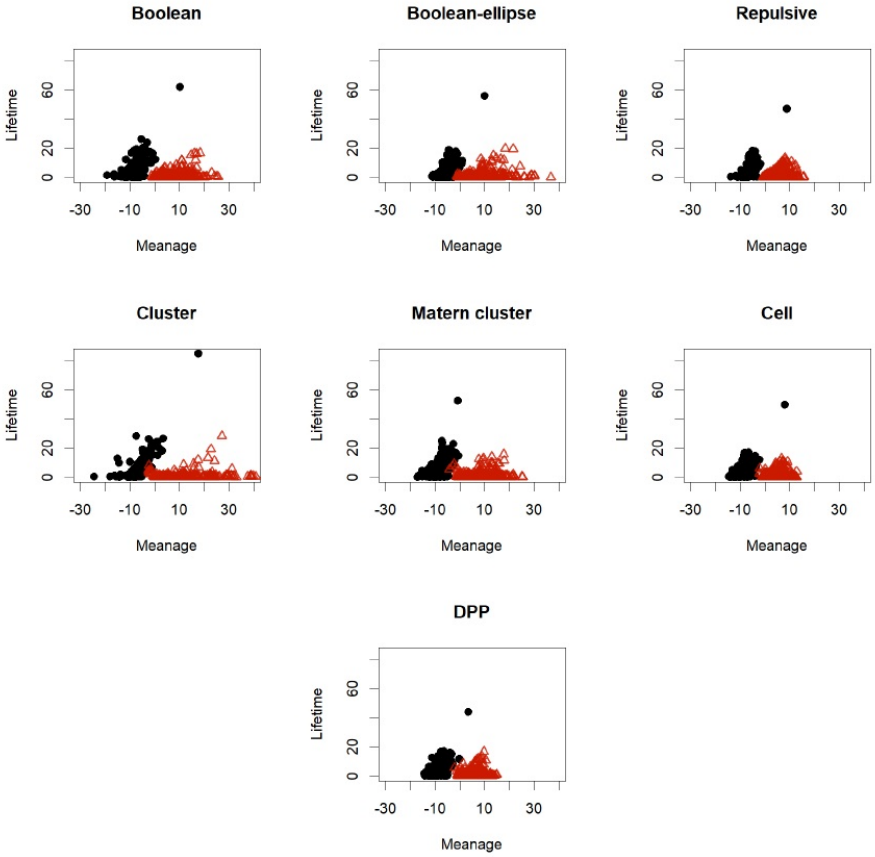
- We first consider the **accumulated persistence function** [1].
- Suppose that the persistence diagram PD^q, where q is the dimension of the topological features it captures, consists of n distinct points (b_i, d_i) with multiplicity c_i for $i = 1, \ldots, n$.
- We use $l_i = d_i - b_i$ to denote the lifetime and $m_i = \frac{b_i + d_i}{2}$ to denote the mean age of each feature.
- We consider the so-called rotated and rescaled persistence diagram ($RRPD_q$), which is a diagram consisting of multiset of distinct points (m_i, l_i) with multiplicity c_i for $i = 1, \ldots, n$.
- The accumulated persistence function (APF):

$$APF_q(m) = \sum_{i=1}^{n} c_i l_i 1(m_i \leq m), \quad m \in \mathbb{R}, \tag{1}$$

  where $1(\cdot)$ is the indicator function and q stands for the dimension of the topological features under consideration.
- APF_q cumulatively sums the lifetimes of the features with respect to their meanage.

[1] Biscio, C. A. N. & Møller, J. (2019). The accumulated persistence function, a new useful functional summary statistic for topological data analysis, with a view to brain artery trees and spatial point process applications. *J Comput Graph Statist*, **28**, 671–681.

The RRPDs of realisations of random set processes



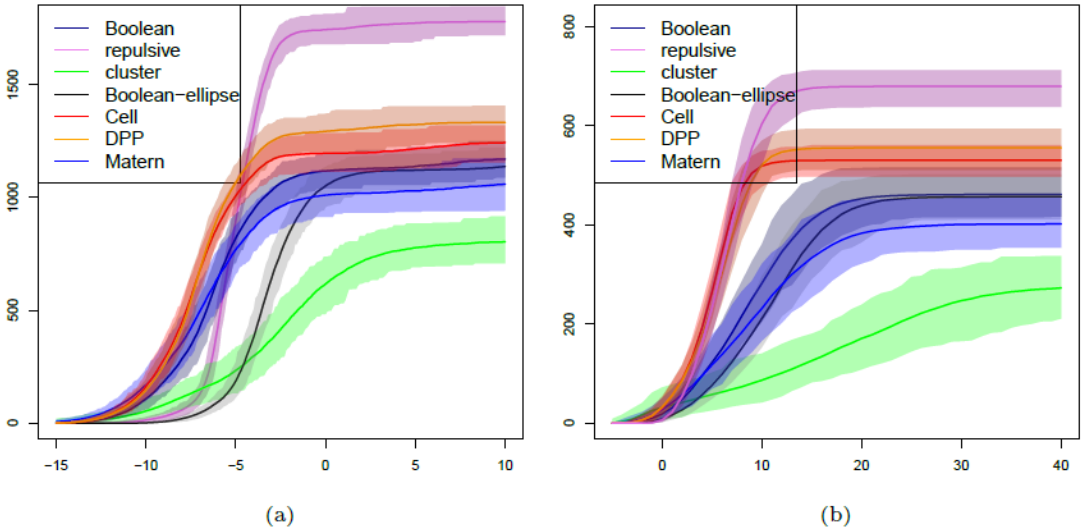


Figure: a) shows mean values of APF_0 obtained from 100 simulated realisations of all observed processes with its 95% envelope, figure b) shows mean values of APF_1 obtained from 100 simulated realisations of all observed processes with its 95% envelope.

Support function of lift zonotop of PD

- Consider the weighted empirical measure of the persistence diagram:

$$PD_W^q(X) = \sum_{i \in \mathcal{I}_q} (d_i - b_i) c_i \delta_{(b_i, d_i)},$$

  and has a mass $(d_i - b_i)c_i$ at the point (b_i, d_i), where c_i stands for the multiplicity of the point (b_i, d_i) in the PD.
- An integrable finite measure μ on $\mathbb{R}^2$ is uniquely identified by a convex set Z in $\mathbb{R}^3$ called a lift zonoid defined as the expected value of the random segment in $\mathbb{R}^3$ with one endpoint at the origin and the other at $(1, \eta)$, where η is distributed according to μ.
- The lift zonoid of $PD_W^q(X)$ is defined as:

$$Z = \bigoplus_{i \in \mathcal{I}_q} (d_i - b_i) c_i \cdot [\mathbf{0}, (1, b_i, d_i)]. \tag{2}$$

- Its support function $h_Z : S^2 \to \mathbb{R}$ is

$$h_Z^q(u) = \sum_{i \in \mathcal{I}_q} l_i c_i \max\{0, \langle u, (1, b_i, d_i) \rangle\}, \; u \in S^2. \tag{3}$$

- h_Z^q defined by (3) **uniquely** characterizes the lift zonotop in (2) and the $PD_W^q(X)$.

Interpretation of values of $h_Z^q(u)$

- Further on, we parametrise the domain of the support function S^2 in the usual way, i.e. we take $(\rho, \phi) \in [0, 2\pi] \times [0, \pi]$ and identify $u \in S^2$ with (ρ, ϕ) such that $u = (\sin(\rho)\cos(\phi), \sin(\rho)\sin(\phi), \cos(\rho))$.

ρ, u	$h_Z^q(u)$	$q = 0$	$q = 1$
$0, (0,0,1)$	$\sum_{i \in \mathcal{I}_q} l_i c_i \max\{0, d_i\}$	larger if the distance between the components and the distance from a component to the boundary is larger.	larger if the diameter and/or the number of holes within the connected components is larger and if empty spaces are larger.
$\frac{\pi}{4}, \left(0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$	$\frac{\sqrt{2}}{2} \sum_{i \in \mathcal{I}_q} l_i c_i \max\{0, d_i + b_i\}$	Smaller than $h_Z^0(0,0,1)$, since $b_i < 0$, with similar interpretation. Smaller if the disc diameters are larger.	Larger than $h_Z^1(0,0,1)$ if there are no holes within the connected components ($b_i > 0$). Can be smaller than $h_Z^1(0,0,1)$ if there are holes within the connected components ($b_i < 0$).
$\frac{\pi}{2}, (0,1,0)$	$\sum_{i \in \mathcal{I}_q} l_i c_i \max\{0, b_i\}$	Equal to 0, since $b_i < 0$.	Larger if there is more empty space between the components.
$\frac{3\pi}{4}, \left(0, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right)$	$\frac{\sqrt{2}}{2} \sum_{i \in \mathcal{I}_q} l_i c_i \max\{0, b_i - d_i\}$	Equal to 0, since $b_i < d_i$.	Equal to 0, since $b_i < d_i$.
$\pi, (0,0,-1)$	$\sum_{i \in \mathcal{I}_q} l_i c_i \max\{0, -d_i\}$	If the components are convex bodies, the value is strictly positive if the components overlap. The more the components overlap, the greater the value.	Equal to 0, since $d_i \geq 0$.
$\frac{5\pi}{4}, \left(0, -\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right)$	$\frac{\sqrt{2}}{2} \sum_{i \in \mathcal{I}_q} l_i c_i \max\{0, -b_i - d_i\}$	Larger values may indicate a larger number of components or clumps positioned close together such that the minima b_i inside them have a larger absolute value than d_i, which is the half-distance between them.	Can be strictly positive if there are holes within the connected components. Larger if the absolute value of the minima of the signed distance function in the components whose clump encloses the hole is larger than its maxima within the hole.
$\frac{3\pi}{2}, (0,-1,0)$	$\sum_{i \in \mathcal{I}_q} l_i c_i \max\{0, -b_i\}$	Larger if the sum of the radii or semi-minor axes is larger. This may be due to more components, leading to more local minima b_i of the signed distance function.	Positive if there are holes born before 0, e.g. in the case of holes within the components or clumps of connected components of the realisation.

Table: Table representing interpretation of values of $h_Z^q(u)$ for some specific values of u when $\phi = \frac{\pi}{2}$

Graphs of the average value of h_Z^0 and h_Z^1 for $\rho \in [0, 2\pi]$ and fixed $\phi = \frac{\pi}{2}$

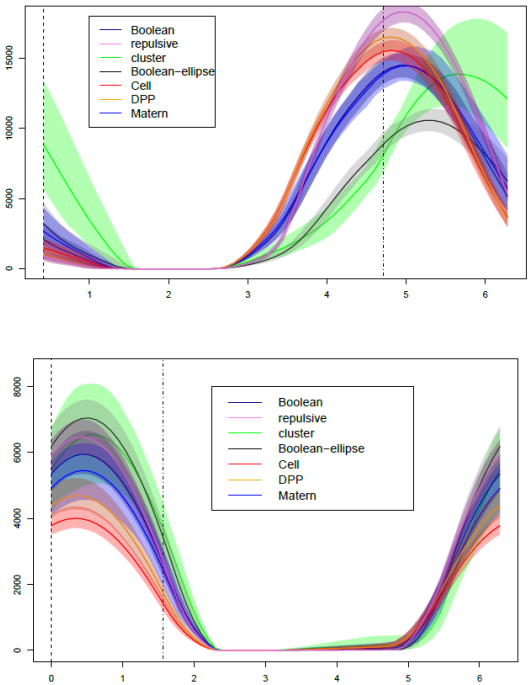


Figure: Top: Graphs of the average value of h_Z^0 for $\rho \in [0, 2\pi]$ and fixed $\phi = \frac{\pi}{2}$ with respect to 100 realisations of each random set process with a corresponding 95% envelopes.; below: Graphs of the average value of h_Z^1 for $\rho \in [0, 2\pi]$ and fixed $\phi = \frac{\pi}{2}$ with respect to 100 realisations of each random set process with corresponding 95% envelopes.

Testing goodness of fit using summary functions

- We consider different summary functions T (CF, ESF, APF_0, APF_1,h_Z^0 and h_Z^1) and use permutation version of global envelope test (Myllymäki and Mrkvička(2020) [2]).
- First, we simulate $n = 50$ realisations from the given random set model.
- For the chosen test summary function $T(x)$, $x \in D_T$, where D_T is the domain of T, we obtain its empirical estimate for the observed random set realisation, denoted by $T_1(x_i), i = 1, \ldots, m$
- In this way, we obtain $n + 1$ discretised curves, which are the input data for the global envelope test.
- The next step is ranking the obtained curves.
- We rank discretise curves $T_1(x_i), T_2(x_i), \ldots, T_{n+1}(x_i)$, $i = 1, \ldots, m$ obtaining ranks $r_1, \ldots, r_{n+1}$ using a extreme rank measure.

- The p-value is

$$p^{\text{erl}} = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbb{I}\left(r_j \leq r_1\right).$$

[2] Myllymäki M., Mrkvička T. (2020) *GET: Global Envelopes in R.* arXiv:1911.06583 [stat.ME], https://arxiv.org/abs/1906.09004

Tables showing percentage of rejected null hypothesis for each process and different test functions (the first number is the percentage of rejection for $p \leq 0.05$ and the second one is for $p \leq 0.1$)

Null hypothesis: Boolean

	M	R	C	Be	Cell	DPP
h_Z^0	4%, 42%	**100%**	**100%**	**100%**	40%, 74%	60%, 88%
h_Z^1	34%, 52%	42%, 54%	68%, 92%	48%, 60%	**100%**	72%, 98%
APF_0	**88%**, 94%	**100%**	**100%**	68%, 100%	4%, 64%	0%, 70%
APF_1	66%, 80%	**100%**	**100%**	10%, 72%	98%, 100%	50%, 96%
ESF	14%, 44%	98%, 100%	**100%**	10%, 100%	34%, 100%	**100%**
CF	26%, 42%	96% , 100%	**100%**	46%, 88%	**100%**	**100%**

Null hypothesis: Matern Cluster

	B	R	C	Be	Cell	DPP
h_Z^0	18%, 42%	**100%**	**100%**	**100%**	88%, 100%	**100%**
h_Z^1	0%, 86%	0%, 100%	64%, 100%	0%, 100%	0%, 100%	0%, 100%
APF_0	66%, 86%	**100%**	74%, 100%	76%, 100%	80%, 100%	**100%**
APF_1	**68%**, 72%	**100%**	80%, 100%	14%, 22%	98%, 100%	**100%**
ESF	22%, 24%	**100%**	**100%**	8%, 66%	**100%**	**100%**
CF	10%, 20%	**100%**	**100%**	6%, 100%	**100%**	**100%**

Null hypothesis: Repulsive

	B	M	C	Be	Cell	DPP
h_Z^0	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
h_Z^1	70%, 100%	88%, 100%	92%, 100%	**100%**	**100%**	90%, 100%
APF_0	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
APF_1	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
ESF	**100%**	**100%**	**100%**	**100%**	90%, 100%	16%, 100%
CF	98%, 100%	**100%**	**100%**	**100%**	18%, 86%	8%, 82%

Null hypothesis: Boolean Ellipse

	B	R	M	C	Cell	DPP
h_Z^0	**100%**	**100%**	**98%**, 100%	68%, 100%	**100%**	**100%**
h_Z^1	26%, 66%	**100%**	68%, 94%	**100%**	98%, 100%	94%, 100%
APF_0	26%, 100%	**100%**	96%, 100%	**100%**	40%, 100%	82%, 100%
APF_1	30%, 64%	**100%**	48%, 86%	98%, 100%	98%, 100%	**100%**
ESF	10%, 100%	**100%**	16%, 96%	**100%**	98%, 100%	**100%**
CF	50%, 96%	**100%**	18%, 88%	**100%**	**100%**	**100%**

Null hypothesis: Cluster

	B	R	M	Be	Cell	DPP
h_Z^0	**100%**	**100%**	**100%**	44%, 76%	**100%**	**100%**
h_Z^1	**100%**	**100%**	90%, 98%	98%, 100%	**100%**	**100%**
APF_0	**100%**	**100%**	70, 100%	**100%**	**100%**	**100%**
APF_1	**100%**	**100%**	44%, 100%	98%, 100%	**100%**	**100%**
ESF	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
CF	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

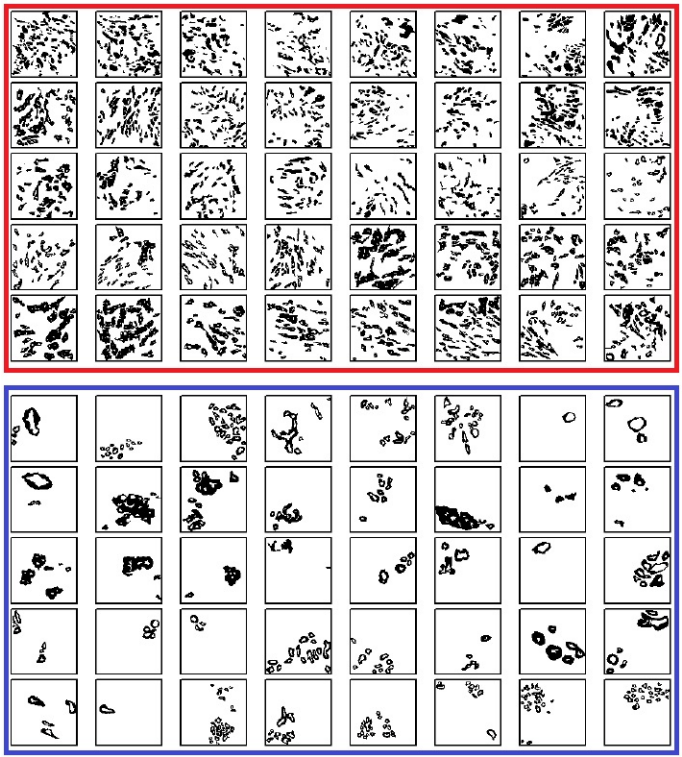
Null hypothesis: Cell process

	B	R	M	Be	C	DPP
h_Z^0	94%, 96%	**100%**	**100%**	**100%**	**100%**	12%, 38%
h_Z^1	**100%**	**100%**	**100%**	**100%**	**100%**	**60%**, 94%
APF_0	10%, 74%	**100%**	94%, 100%	14%, 100%	**100%**	18%, 40%
APF_1	34%, 100%	**100%**	96%, 100%	88%, 100%	**100%**	18%, 44%
ESF	96%, 100%	90%, 100%	**100%**	**100%**	**100%**	46%, 78%
CF	**100%**	22%, 88%	**100%**	**100%**	**100%**	0%, 0%

Null hypothesis: DPP

	B	R	M	Be	C	Cell
h_Z^0	98%, 100%	**100%**	**100%**	**100%**	**100%**	42%, 56%
h_Z^1	**100%**	**100%**	**100%**	**100%**	**100%**	10%, 56%
APF_0	64%, 100%	**100%**	**100%**	82%, 100%	**100%**	36%, 54%
APF_1	92%, 98%	**100%**	**100%**	**100%**	**100%**	26%, 28%
ESF	**100%**	84%, 96%	**100%**	**100%**	**100%**	**72%**, 72%
CF	**100%**	34%, 100%	**100%**	**100%**	**100%**	8%, 20%

Application to real data



Graphs of different summary functions for mammary cancer tissue samples (mamca, red lines) and mastopathy samples (masto, blue lines).

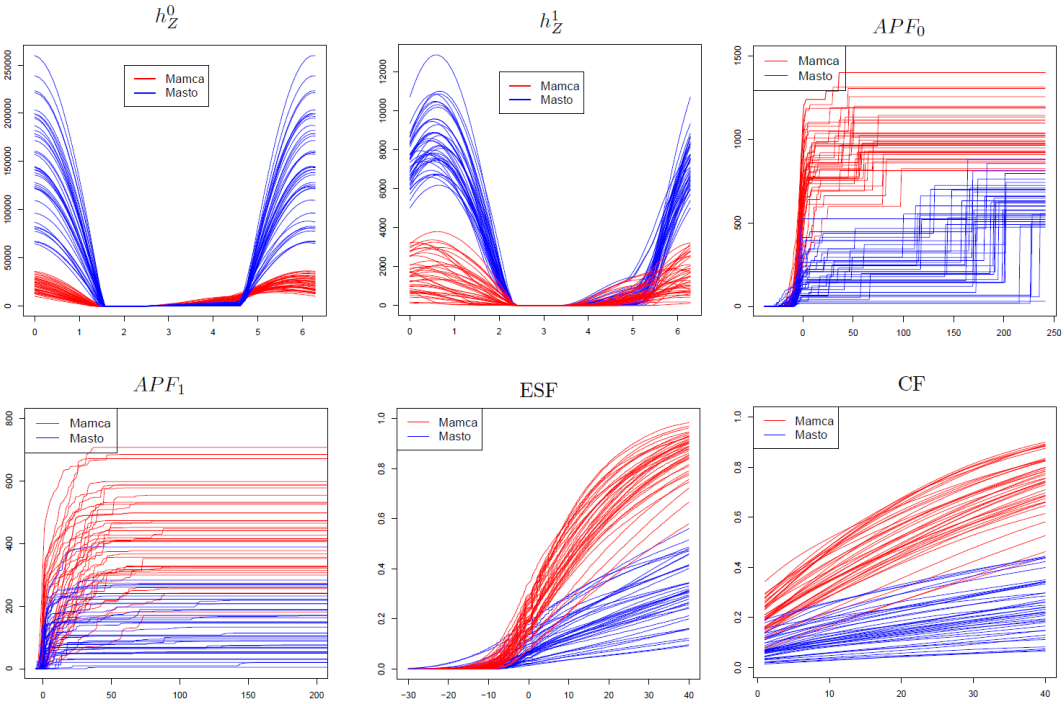


Table: Table showing the percentage of rejected null hypothesis in testing Mamca vs Masto, Masto vs Mamca for different test functions (the first number is the percentage of rejection for $p \leq 0.05$ and the second one is for $p \leq 0.1$).

	Mamca vs Masto	Masto vs Mamca
h_Z^0	**100%**	**100%**
h_Z^1	**100%**	**100%**
APF_0	**100%**	**100%**
APF_1	57.5%, 57.5%	67.5%, 67.5%
ESF	95%, 97.5%	90%, 100%
CF	70%, 72%	76%, 76%

Future plans:

- CLT for functionals of PD and applications to statistical testing,
- Apply methods to more examples of real data

References I

[1] V. Gotovac Đogaš (2024) *Depth for samples of sets with applications to testing equality in distribution of two samples of random sets* arXiv:2402.01861 [stat.ME]
`http://arxiv.org/abs/2402.01861`


Thank you for your attention!