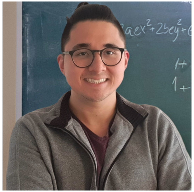
Topological methods for computational ecology

Stability of 0-dimensional persistent homology in enriched and sparsified point clouds [arxiv:2511.07093]



Jānis Lazovskis
University of Latvia

Ran Levi
University of Aberdeen

Juliano Morimoto
University of Aberdeen

June 3, 2026 / Inria DataShape seminar
Slides online at: `jlazovskis.com/talks`
Preprint and software linked at end

Funding for JL: Activity 1.1.1.9 "Post-doctoral Research" of the Specific Objective 1.1.1 "Strengthening research and innovative capacities and introduction of advanced technologies in the common R&D system" of the European Union's Cohesion Policy Programme for 2021-2027 research application No 1.1.1.9/LZP/1/24/125 "Efficient topological signatures for representation learning in medical imaging"

Līdzfinansē Eiropas Savienība
2027 Nacionālais attīstības plāns
Latvijas Zinātnes padome
LATVIJAS UNIVERSITĀTE
PostDoc Latvia

Overview

Ecologists are intereseted in **large data sets**

- On the order of 1 000 to 1 000 000, dimensions $\sim$ 20
- Meaning is inferred from the shape, but the shape is hard to infer
- Current statistical methods to infer shape are computationally intensive
  - Constructing distribution for $\sim$ 100 000 data points in 3 dimensions takes several hours on $\sim$1000 GB memory

Usual scenario: Topology (and geometry) can **simplify** the data (input and output)

- Well-suited for output, as disconnectedness, holes, voids in data are ecologically significant
- Topological computations also hard, simplify input with persistence bounds

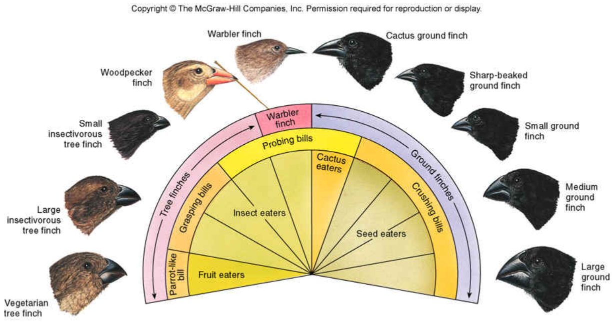
Work still in progress: *7 collaborators, 2 publications, 2 preprints, 1 GitHub repo...*

Ecological niches

The niche of a species is the set of environmental parameters in which it can exist.

- The **size** of a climatic niche is hypothesized to drive species diversification rates
- The **similarity** of species' environmental or functional trait hypervolumes measures niche divergence or packing, which may influence species coexistence and richness patterns

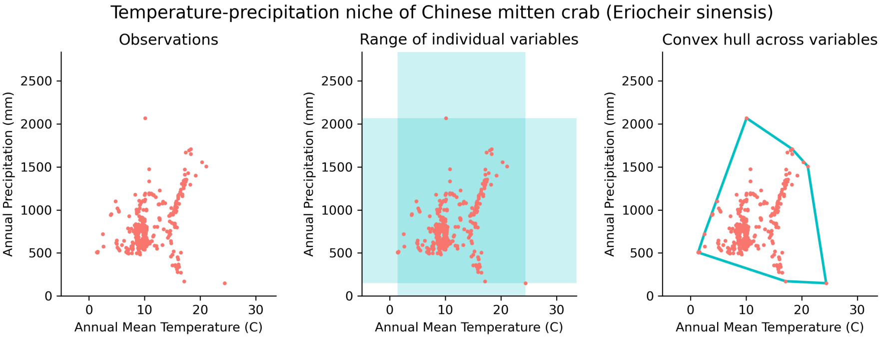
- Niche similarity also helps compare individuals within a species, assessing climate change **impacts** and niche **shifts** during invasions
- Introduced in 1957 by Hutchinson.



The shape of a niche

Prompt: What is the smallest, most reasonable space, in which observations are made?

- Box made from range of variables in each dimension
- Convex hull of all observations



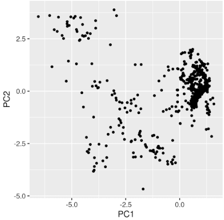
Issues: Sensitive to outliers, convex hull is difficult in high dimensions.

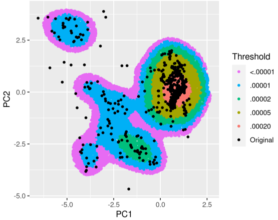
Current methods

The niche of a species as a computational object is called the **hypervolume**. Current statistical methods construct it as a superlevel set of a distribution.

GBIF
Global Biodiversity
Information Facility

```
> data_raw %>% head
  ID alt bio1 bio10 bio11 bio12 bio13 bio14 bio15 bio16 bio17 bio18 bio19 bio2
1  1  30  144   185    99   932   206     1    93   521     9    13   521  101
2  2  48  140   187    91   753   168     1    92   423     9    13   423  138
3  3   4  155   219    84   361    80     1    85   197     6     9   197  137
4  4  10  144   191    92   740   174     1    95   430     7    10   430  128
5  5   0  150   194   102   356    73     1    86   186     5     8   185  108
6  6   2  152   211    87   518   119     1    90   287     6    10   287  134
  bio3 bio4 bio5 bio6 bio7 bio8 bio9         x        y dcoast
1   52 3393  244   50  194   99  182 -122.4583 37.95833     -3
2   57 3812  273   33  240   91  183 -122.6250 38.20833     20
3   47 5333  315   27  288   84  218 -121.7917 38.04167     42
4   54 3936  270   37  233   92  188 -122.4583 38.12500      5
5   52 3613  253   49  204  111  191 -122.0417 37.45833      3
6   49 4891  299   30  269   87  209 -122.0417 38.20833     28
>
```



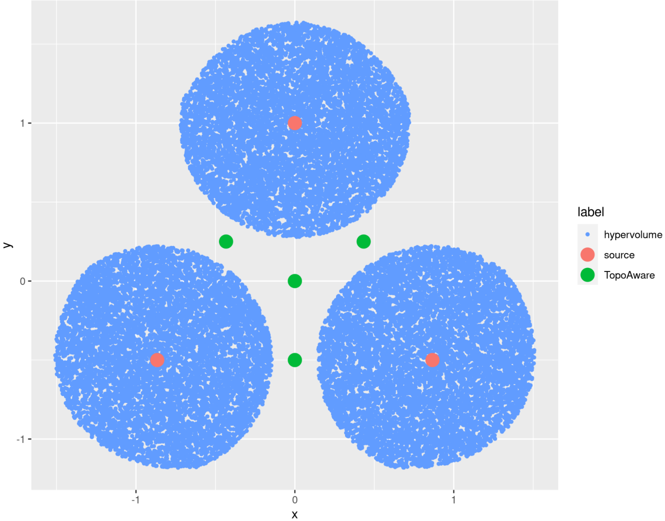


GBIF ⟶ observations ⟶ climate data ⟶ PCA ⟶ hypervolume

- Data is the observed niche, would like to get the realized niche: *unknown ground truth*
- Observations directly related to human activity density: *inherent sampling bias*
- Ecologists previously assumed hypervolumes were "blob-like:" *holes need justification*

Statistics vs topology

- Statistical method (Blonder, 2017) infers new samples **around** existing samples.
- Our topological method (TopoAware, 2025) infers new samples **between** existing samples.
- Methods may be considered as **complementary**, and both used in appropriate settings.



Our proposed pipeline

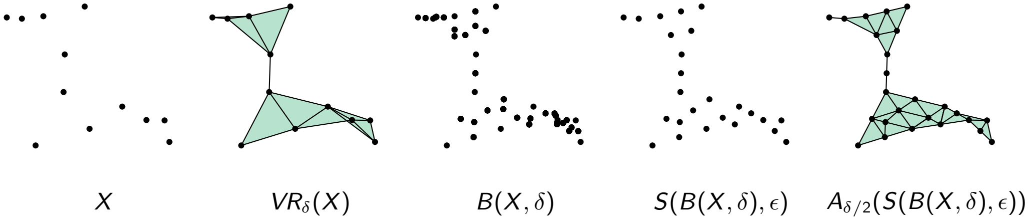
X $B(X,\delta)$ $S(B(X,\delta),\epsilon)$ $G(S(B(X,\delta),\epsilon),\mu)$ $\mu\mathbf{Z}^n \setminus G(S(B(X,\delta),\epsilon),\mu)$

- Use **barycentric subdivision** of Vietoris-Rips complex
  - Mimics KDE distribution by inferencing new points
  - Does not expand area / does not make boundary smoother, unlike KDE
- Use **sparsification** to overcome computational limits
  - Mimics KDE distribution by making density (more) uniform
  - Overcomes one type of sampling bias without making other type worse
- Use Alexander **duality** to find codimension 1 holes

Setup: Simplicial complexes

Let X be a data set, and $\delta, \epsilon > 0$.

- Let $B(-,\delta)$ be the 0-simplices of the barycentric subdivision of $VR_\delta^{\dim \leqslant d}(-)$.
- Let $S(-,\epsilon)$ be the sparsification produced by dropping all neighbors within ϵ.



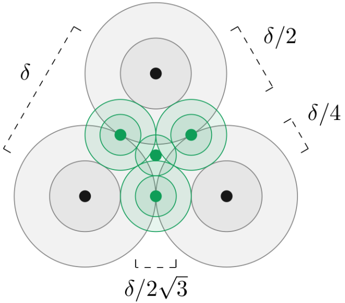
The output $S(B(-,\delta),\epsilon)$ has the same "shape" as the input, with a prescribed number and density of points. Use the **alpha complex** $A_{\delta/2}(-)$ for faster computation of the barcode and to retrieve geometric information in low dimensions.

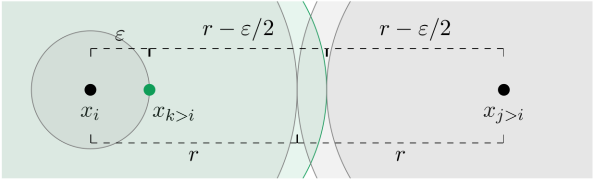
Results: Change in persistence with more and less points

Claim: Let $X \subseteq \mathbf{R}^N$ be a finite set, $B(X,\delta) \supseteq X$ the baycentric subdivision, $S(X,\epsilon) \subseteq X$ the sparsification. Let $D_\bullet$ be the persistence diagrams of the Vietoris–Rips filtrations in dimension 0. Then

$$d_b\left(D_X, D_{B(X,\delta)}\right) \leqslant \frac{\delta}{4}, \qquad d_b(D_X, D_{S(X,\varepsilon)}) \leqslant \frac{\varepsilon}{2}.$$

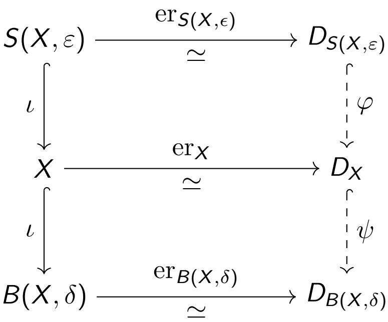
Proof: What's the worst that could happen?

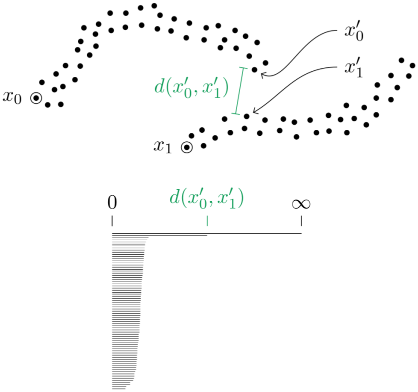




Proof of results: Taking advantage of the elder rule

In dimension 0, each point in an ordered X can be related to an interval by the elder rule. In these cases, that relation minimizes bottleneck distance.





Bottleneck distance is infimum of supremum of source-target distances in bijections between diagrams (with ephemeral intervals on the diagonal). Maps φ, ψ are injections. Bound determined by farthest new points that "appear." $\square$

Setup: Integer lattices

Let $\mu > 0$ and consider $\mu \mathbf{Z}^n$. Divide $x \in X$ as $x = q\mu + r$ with $\|r\|_\infty < \mu$, record only $q \in \mu \mathbf{Z}^n$.

The **grid** of X is $G(X, \mu) \subseteq \mu \mathbf{Z}^N$

The **thickening** of $G(X, \mu)$ is $G_t(-, \mu) \subseteq \frac{\mu}{2} \mathbf{Z}^N$

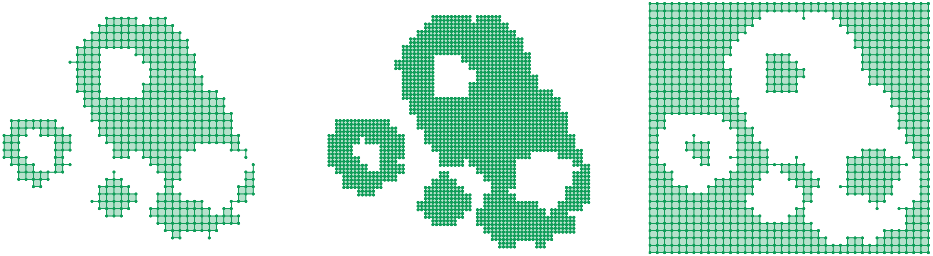
The **complement** of $G(X, \mu)$ is $\mu \mathbf{Z}^n \setminus G(-, \mu) \subseteq \mu \mathbf{Z}^N$

- Thickening is the union of all linear combinations of $G(X, \mu) \pm \frac{\mu}{2} e_i$ for basis vectors e_i.
- Complement is used to take advantage of degree 0 computations

Setup: Cubical complexes and filtrations

A cubical complex $\mathcal{C}$ may be constructed by the **V-construction** on any subset of $\mu\mathbf{Z}^N$.

$\mathcal{C}_\mu(G(X,\mu))$ $\quad$ $\mathcal{C}_{\mu/2}(G_t(X,\mu))$ $\quad$ $\mathcal{C}_\mu(\mu\mathbf{Z}^N \setminus G(X,\mu))$



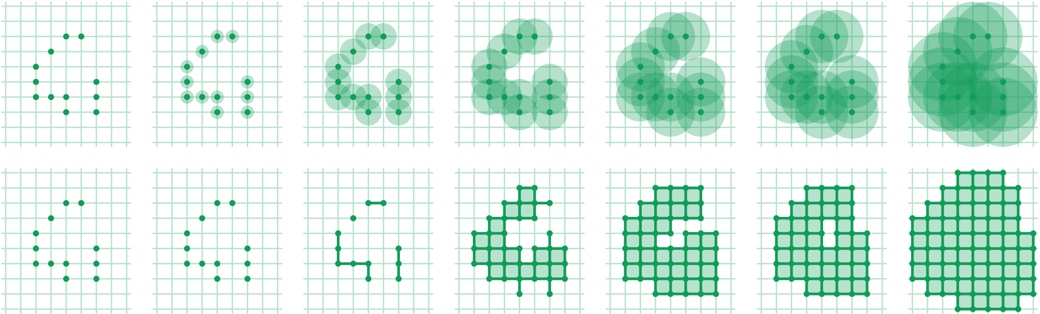
Every $G \subseteq \mu\mathbf{Z}^N$ induces a filtration of the complex of all cubes in $\mu\mathbf{Z}^N$ ("D-construction"?):

- *a cube σ enters at r if $\sigma \subseteq \bigcup_{g\in G} B(g,r)$, for $B(g,r) = \{p \in \mathbf{R}^N \ : \ d(g,p) \leqslant r\}$*

Setup: Distance construction

Top row: $\bigcup_{g \in G} B(g, r_i)$ Bottom row: $\mathcal{DC}(G, r_i)$

$r_0 < r_1 < r_2 < r_3 < r_4 < r_5 < r_6$



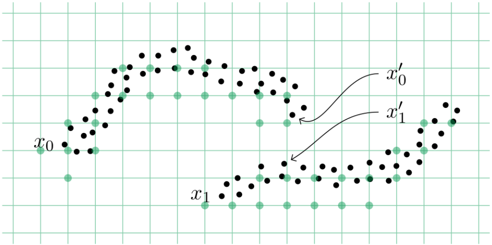
Cube σ has approx filtration value $\max_{s \in \sigma_0} d(s, G)$. Better approx with $\max_{\tau \subseteq \sigma} d\left(\frac{\sum_{t \in \tau_0} t}{2^{\dim(\tau)}}, G\right)$

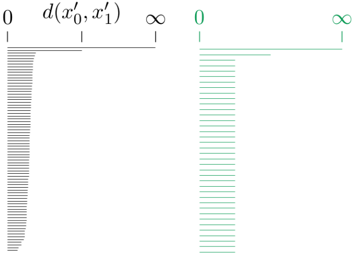
Results: Change in persistence when aligning to grid

Claim: Let $X \subseteq \mathbf{R}^N$ be a finite set and D_X the persistence diagram of its Vietoris–Rips filtration in dimension 0. Let $D_{G(X,\mu)}$ be the persistence diagram of the cubical filtration on $G(X,\mu)$ in dimension 0. Then

$$d_b\left(D_X, D_{G(X,\mu)}\right) \leqslant \frac{\sqrt{N}\mu}{2} \quad \text{and} \quad H_{N-1}(\mathcal{C}_{\mu/2}(G_t(X,\mu))) \oplus \mathbf{Z}_2 \simeq H_0(\mathcal{C}_\mu(\mu\mathbf{Z}^N \setminus G(X,\mu))).$$

Proof (of inequality): Similar to before (different inclusion map), triangle inequality.





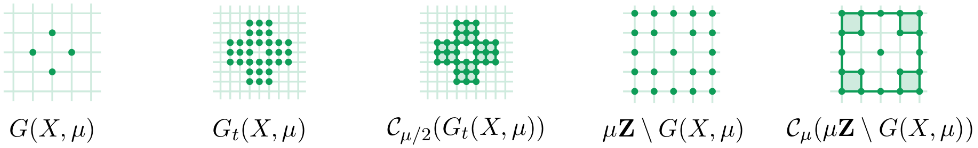
Proof of isomorphism

- Duality isomorphism is in (not persistent) homology, so a distance μ has to be chosen.
- After sparsifying a dense set, the "shape" of the data appears at the sparsification radius.

Proof (of isomorphism): Follows from Alexander duality, and $\mathbf{R}^N$ being disjoint union of

$$X' := \left\{x \in \mathbf{R}^N \ : \ \text{there exists } x' \in G(X,\mu) \text{ with } d_\infty(x,x') \leqslant \tfrac{\mu}{2}\right\},$$
$$X'' := \left\{x \in \mathbf{R}^N \ : \ \text{there exists } x' \in \mathcal{C}_\mu(\mu\mathbf{Z}^N \setminus G(X,\mu)) \text{ with } d_\infty(x,x') < \tfrac{\mu}{2}\right\},$$

as X' is the geometric realization of $\mathcal{C}_{\mu/2}(G_t(X,\mu))$.



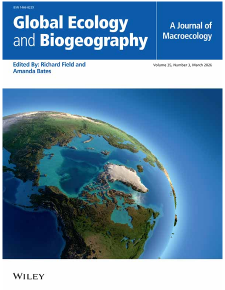
Implementation

Code is in C++, Python, R. Simplex tree and sparsification is taken from GUDHI.

function	input	output
`barycentric_subdivision`	$X \subseteq \mathbf{R}^N$, $\delta \in \mathbf{R}_{\geqslant 0}$	$B(X, \delta) \subseteq \mathbf{R}^N$
`sparsification`	$X \subseteq \mathbf{R}^N$, $\epsilon \in \mathbf{R}_{\geqslant 0}$	$S(X, \epsilon) \subseteq \mathbf{R}^N$
`gridification`	$X \subseteq \mathbf{R}^N$, $\mu \in \mathbf{R}_{>0}$, $z \in \mathbf{R}^N$	$G(X, \mu) \subseteq \mu\mathbf{Z}^N$
`complement`	$G \subseteq \mu\mathbf{Z}^N$	$\mu\mathbf{Z}^N \setminus G \ \subseteq \mu\mathbf{Z}^N$
`thickening`	$G \subseteq \mu\mathbf{Z}^N$	$G_t \subseteq \frac{\mu}{2}\mathbf{Z}^N$

```
exec(open("topoaware.py").read())
data = np.load("../examples/generated_data.npy")
bs = barycentric_subdivision(data, radius=0.25, max_dim=2)
sp = sparsification(bs, min_dist=0.01)
gr = gridification(sp, grid_interval=0.04, grid_origin=[0,0])
co = complement(gr, grid_interval=0.04, buffer=2)
th = thickening(gr, grid_interval=0.04)
```

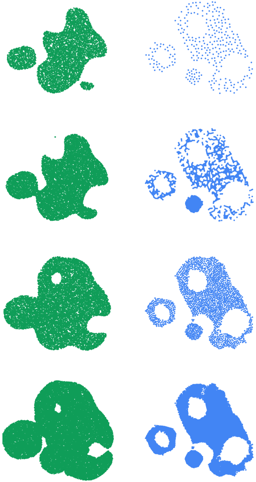
Applications



Recent work: *"The Evolution of Hutchinsonian Climatic Niche Hypervolumes in Gymnosperms"* compares global shapes (volume as superlevel set of KDE) of occupied environmental parameters of similar species

Initial application: Sparsify (geometry) before thinning, estimating (statistics). Compute PD of each species.

- *Ex: KDE of 300k samples impossible to estimate without sparsification on HPC*



Next steps: Compare with PD of genus, family, etc. Infer volume from simplicial or cubical complex constructions.

Thank you

- Blonder, Morrow, Maitner, Harris, Lamanna, Violle, Enquist, Kerkhoff. *New approaches for delineating n-dimensional hypervolumes.* Methods in Ecology and Evolution, 2017.
- Caron, Lazovskis, Burslem, Morimoto. *The Evolution of Hutchinsonian Climatic Niche Hypervolumes in Gymnosperms.* Global Ecology and Biogeography, 2026.
- Conceição, Morimoto. *'Holey' niche! finding holes in niche hypervolumes using persistence homology.* Journal of Mathematical Biology, 2022.
- G.E. Hutchinson. *Population studies-animal ecology and demography - concluding remarks.* Cold Spring Harbor symposia on quantitative biology, 1957.
- Lazovskis, Levi, Morimoto. *Stability of 0-dimensional persistent homology in enriched and sparsified point clouds.* arXiv, 2025.
- Lazovskis. *TopoAware: Topologically aware constructions for large and irregular datasets.* GitHub, 2026.
- Maria. *GUDHI: Filtered complexes.* GUDHI editorial board, 2025.
- Zhang, Mammola, McLay, Capinha, Yokota. *To invade or not to invade? Exploring the niche-based processes underlying the failure of a biological invasion using the invasive Chinese mitten crab.* Science of the Total Environment, 2020.