



HAL
open science

Un nouvel estimateur des niveaux de densité utilisant les graphes et complexes simpliciaux. Application à la détection des clusters de galaxies.

Louis Hauseux, Konstantin Avrachenkov, Josiane Zerubia

► To cite this version:

Louis Hauseux, Konstantin Avrachenkov, Josiane Zerubia. Un nouvel estimateur des niveaux de densité utilisant les graphes et complexes simpliciaux. Application à la détection des clusters de galaxies.. XVIe Journées de géostatistiques (Fontainebleau, 7-8 septembre 2023) organisé par les Mines de Paris - PSL, Sep 2023, Fontainebleau, France. . hal-04222280

HAL Id: hal-04222280

<https://inria.hal.science/hal-04222280>

Submitted on 29 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Un nouvel estimateur des niveaux de densité utilisant les graphes et complexes simpliciaux

Application à la détection des clusters de galaxies

Louis HAUSEUX Konstantin AVRACHENKOV Josiane ZERUBIA

Inria, Université Côte d'Azur, Sophia-Antipolis, France Prénom.NOM@Inria.fr

Présentation du problème

Les galaxies ne se répartissent pas uniformément au sein de l'univers mais se regroupent au sein de « structures à grandes échelles » :

- 1° des super-amas de galaxies (petits volumes hyper-denses de \mathbb{R}^3);
- 2° des feuillettes ou « murs » de galaxies (surfaces);
- 3° des « filaments » de galaxies (courbes).

Ces clusters délimitent de grandes régions quasiment vides de galaxies.

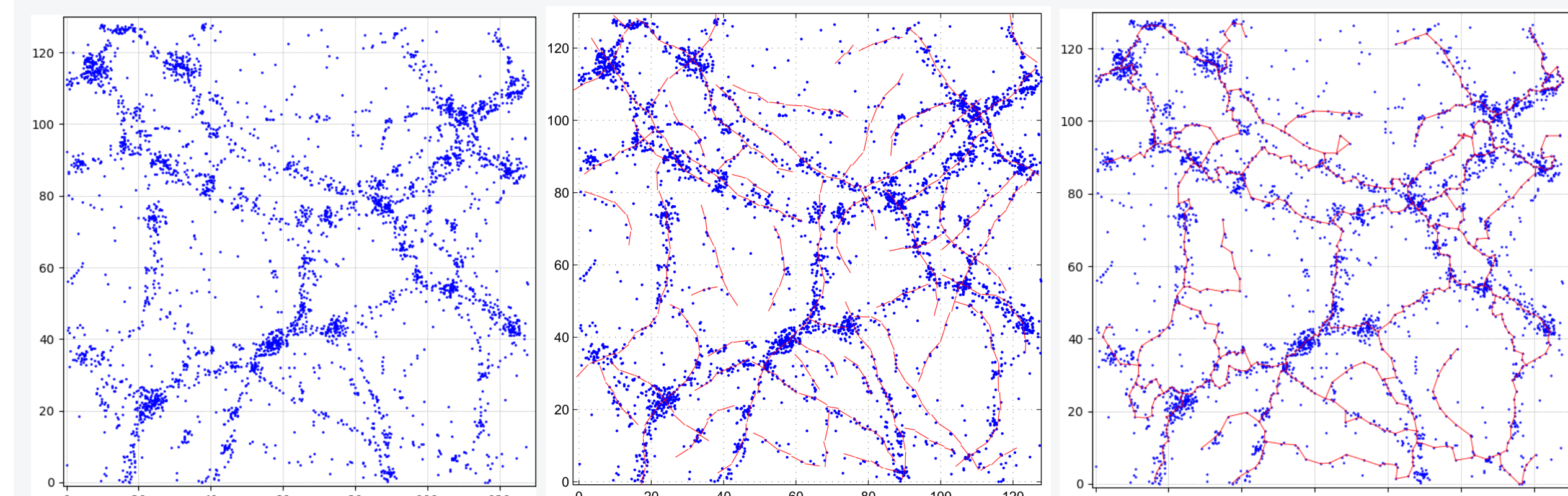


Figure 1. Nuage de points synthétique $\mathcal{X} \subset \mathbb{R}^2$ créé par les auteurs de l'article [4] et gracieusement fourni par R. Stoica, V. Martínez et E. Saar. Au milieu, le résultat de leur algorithme. À droite, le nôtre.

→ Clusters hiérarchiques ⇒ il est naturel de recourir à un **estimateur de densité**. Estimateurs de densité classiques pour notre problème [2] :

Delaunay [3] & **K-Plus-Proches-Voisins** [1].

Outils mathématiques

Graphes géométriques. Soit $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ un nuage de points, les x_i ayant été tirés i.i.d. selon une mesure à densité f par rapport à la mesure de Lebesgue. Soit $r > 0$ un rayon, le graphe géométrique $\mathcal{G}(\mathcal{X}, r)$ est le graphe (non orienté) :

$$\text{Sommets} := \mathcal{X}; \quad \text{Arêtes} := \{\{x_i, x_j\} \mid |x_i - x_j| < r\}.$$

Complexe simplicial est un ensemble $\mathcal{K} \subset \mathcal{P}(\mathcal{X})$ constitué de parties σ non vides d'un ensemble fini \mathcal{X} (= *Sommets*) vérifiant :

$$\forall \sigma \in \mathcal{K}, \forall \tau \subseteq \sigma, \tau \neq \emptyset \Rightarrow \tau \in \mathcal{K}.$$

Un élément $\sigma \in \mathcal{K}$ est un *simplexe*. Ses sous-ensembles $\tau \subset \sigma$ en sont ses *faces*.

→ permet de généraliser la notion de graphe ainsi que de **composante connexe**.

Percolation. Soit $\mathcal{X} := \mathcal{H}_1 \cup \{0\} \subset \mathbb{R}^d$ un processus ponctuel de Poisson homogène sur \mathbb{R}^d d'intensité $\lambda := 1$. Il existe un *rayon critique* r_c tel que :

- $\forall r < r_c$, avec probabilité 1, $\mathcal{G}(\mathcal{X}, r)$ n'a aucune composante connexe infinie;
- $\forall r > r_c$, avec probabilité 1, $\mathcal{G}(\mathcal{X}, r)$ possède une composante connexe infinie.

Percolation = composante ∞ présente (contenant une proportion $p_\infty(r)$ des points).

→ la **percolation est un phénomène rapide** : $p_\infty(r)$ se rapproche très vite de 1.

« Mesurons » cette vitesse. Soit $0 < \varepsilon < \frac{1}{2}$. Notons $r_{\min} := p_\infty^{-1}(\varepsilon)$ et $r_{\max} := p_\infty^{-1}(1 - \varepsilon)$.

La **vitesse de percolation** se mesure grâce à : $\frac{r_{\max}}{r_{\min}}$.

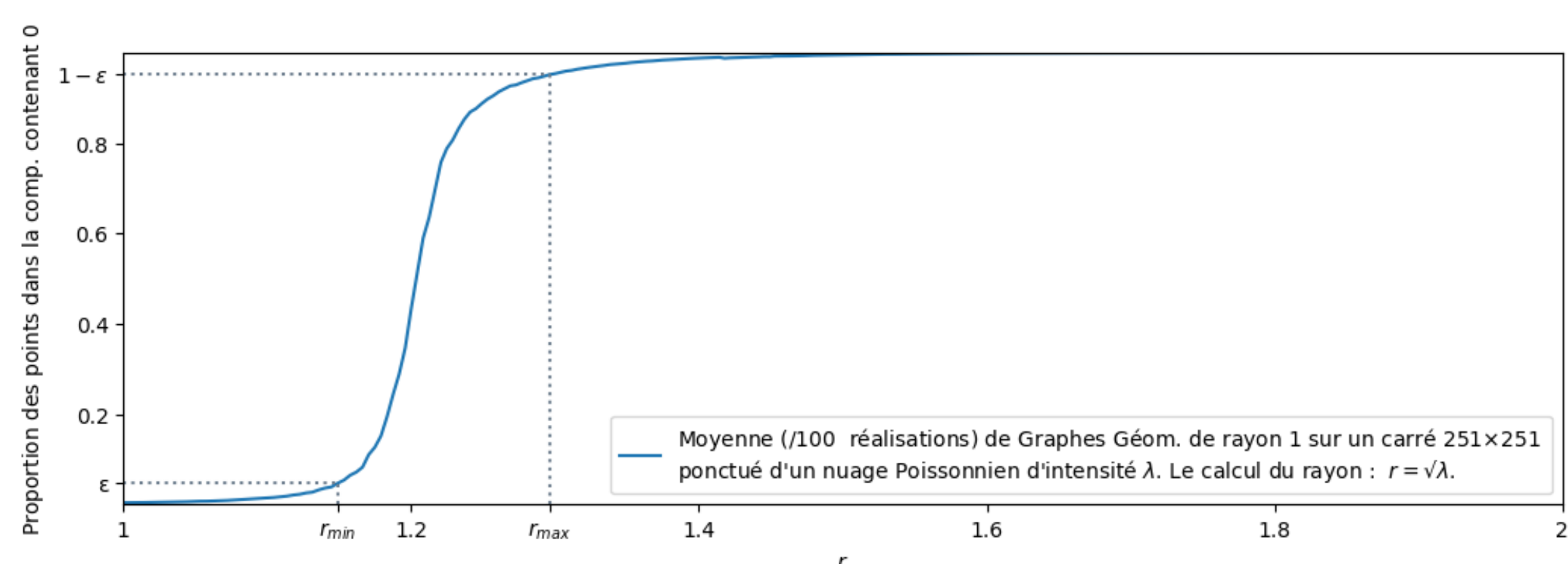


Figure 2. Avec $\varepsilon \leftarrow 0,05$, $r_{\min} = 1,15$ et $r_{\max} = 1,30$ soit $\frac{r_{\max}}{r_{\min}} \approx 1,13$.

→ **percolation rapide** ⇒ 2 clusters voisins de densités proches bien distingués.

→ Soit \mathcal{G}_l le graphe élagué des petites composantes (de taille $\leq \text{SeuilPerco} \leftarrow 50$).

Niveaux de densité. $\text{High-Density Clusters}(b) := f^{-1}([b, +\infty[)$ [5], [6].

Soit $P \in [0, 1]$, la *hauteur* h_P de densité associée : $h_P := \inf \{b \mid \int_{f \geq b} f(x) dx \leq P\}$.

Le *rayon* associé : $r_P := \inf \left\{ r \mid \frac{|\text{Nœuds}(\mathcal{G}_l(\mathcal{X}, r))|}{|\text{Nœuds}(\mathcal{G}(\mathcal{X}, r))|} \geq P \right\}$.

La **carte des niveaux** : $\mathcal{P} : x \mapsto \inf \{P \in [0, 1] \mid x \in \text{High-Density Clusters}(h_P)\}$.

La **carte empirique des niveaux** : $\hat{\mathcal{P}}(x) := \inf \{P \in [0, 1] \mid x \in \Sigma_P\}$ ($\inf(\emptyset) = 0$).

Le **cluster de niveau** P : $\Sigma_P := \bigcup_{x \in \text{Nœuds}(\mathcal{G}_l(r_P))} B(x, r_P)$.

Précision/Rappel sur la bonne identification d'un cluster.

Soit $P \in [0, 1]$ et \mathcal{C}_P un cluster de niveau P . Soit $\hat{\mathcal{C}}_P$ le cluster empirique associé.

$$\text{Précision} := \frac{|\mathcal{C}_P \cap \hat{\mathcal{C}}_P|}{|\hat{\mathcal{C}}_P|}, \quad \text{Rappel} := \frac{|\mathcal{C}_P \cap \hat{\mathcal{C}}_P|}{|\mathcal{C}_P|}.$$

Résultats

Données : densité f imitant des filaments plus ou moins fins/denses par des rectangles à densité homogène.

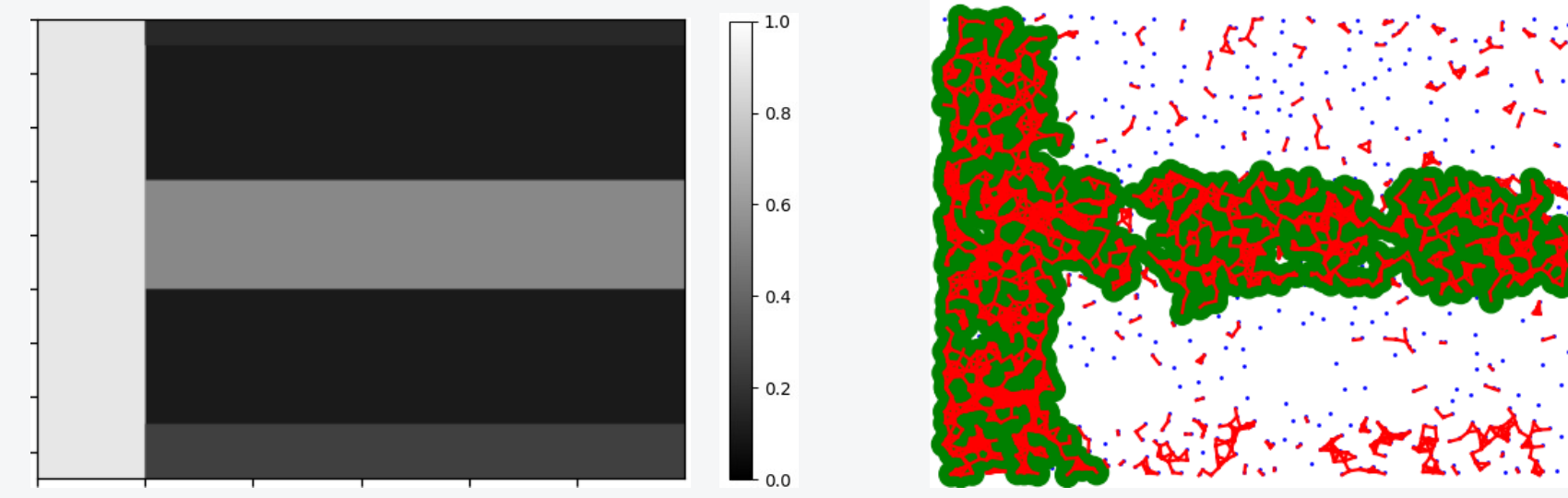


Figure 3. À gauche : les niveaux de densité théoriques. À droite : un nuage \mathcal{X} de 2000 points générés i.i.d. selon f . Un graphe géométrique $\mathcal{G}(\mathcal{X}, r)$ est dessiné avec arêtes en rouge et le cluster associé Σ_r en vert. $\text{SeuilPercolation} \leftarrow 50$.

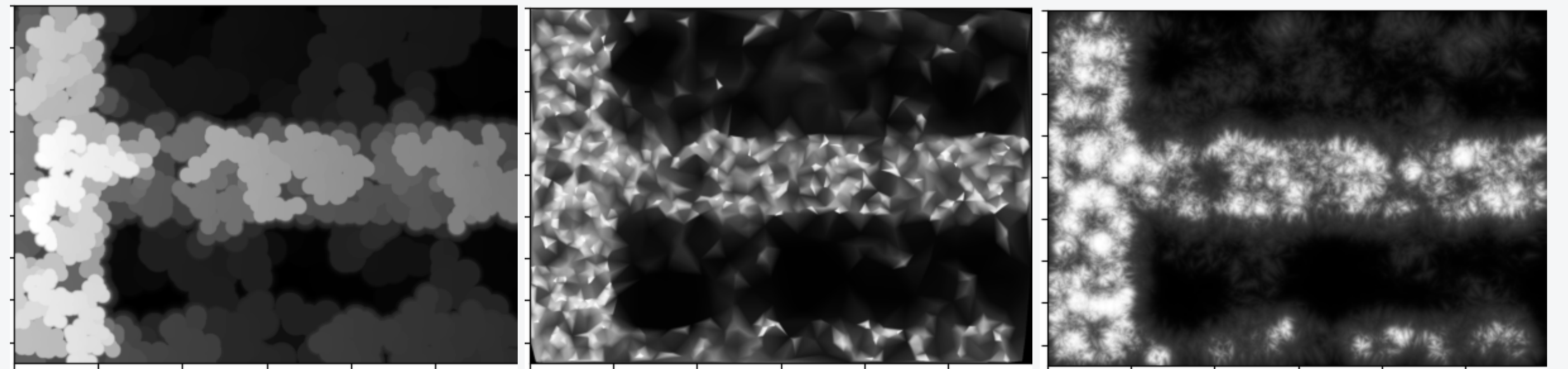


Figure 4. Trois estimateurs des niveaux de densité sur \mathcal{X} : notre **Graphe-Percol**; **Delaunay**; **10-Plus-Proches-Voisins**.

→ Visuellement, le nôtre est **plus homogène et moins sujet aux petites sur-détections**.

Quatre 'filaments' apparaissant successivement \iff Quatre niveaux de clusters :

Algorithme	$1 - \hat{P} > 0,617$		$1 - \hat{P} > 0,280$		$1 - \hat{P} > 0,169$		$1 - \hat{P} > 0,140$	
	Précision	Rappel	Précision	Rappel	Précision	Rappel	Précision	Rappel
Graphe-Percol	0,741	0,827	0,881	0,970	0,814	0,945	0,783	0,896
Delaunay	0,573	0,315	0,857	0,799	0,892	0,899	0,881	0,886
K-Plus-Proches-Voisins	0,600	0,528	0,899	0,864	0,861	0,945	0,802	0,893

Notre estimateur l'emporte haut-la-main sur le premier niveau (le 'super-amas' à gauche).

→ **Percolation rapide** ⇒ recouvre bien le 'super-amas' avant de commencer sur le 'filament' médian.

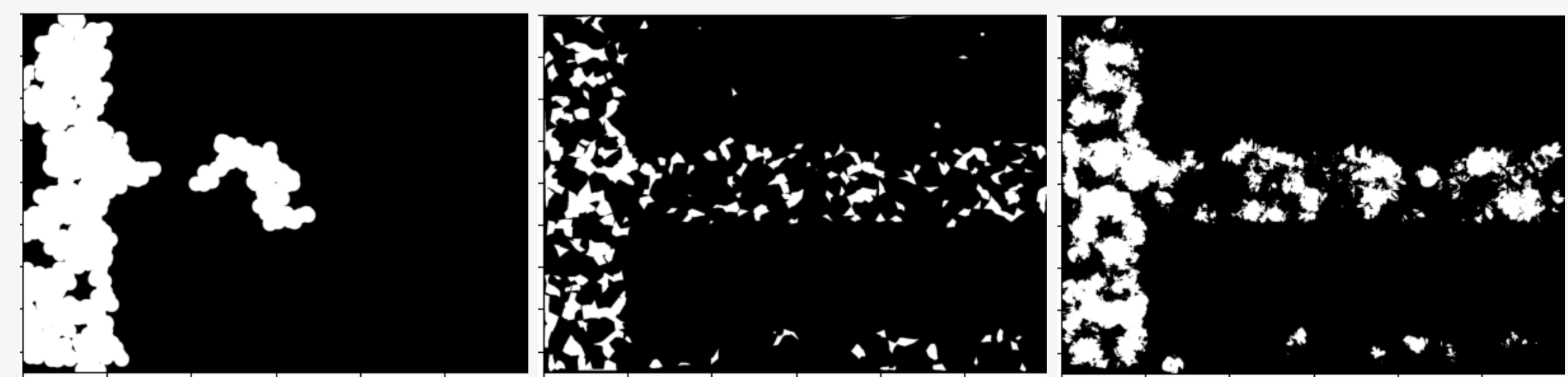


Figure 5. Clusters $\hat{\mathcal{C}}_P$ pour le seuil $1 - P = 0,617$. De gauche à droite : **Graphe-Percol**; **Delaunay**; **10-Plus-Proches-Voisins**.

→ Graphes/complexes simpliciaux ⇒ extraction plus facile de filaments. Cf. image de dr. Figure 1.

Algorithme d'extraction de filaments de galaxies

→ Partir d'une comp. connexe G de \mathcal{G}_l et de son *Filament* (sous-gr.) déjà construit.

→ Décider d'un nombre de *Centres* à ajouter selon la taille de G .

→ Itérativement, un nouveau centre x est choisi de façon à minimiser :

$$D[x] = \sum_{y \in \text{Nœuds}(G)} d(y, \text{Filament} \cup \text{Branche}_x)^2$$

(où d est la distance induite sur le graphe et Branche_x le chemin le plus court de x à *Filament*).

$$\text{Filament} \leftarrow \text{Filament} \cup \text{Branche}_x$$

→ Premier centre choisi : **moyenne de Fréchet** de G .

Conclusion

→ Avoir un regard *persistant* (= variationnel) permet d'observer les différentes phases de percolation.

→ La percolation est un phénomène très rapide permettant de distinguer deux niveaux de densité proches.

→ Comparé aux estimateurs de densité conventionnels pour ce problème, les premiers résultats sont très encourageants. Notamment pour les hauts niveaux de densité.

Perspectives

→ Considérer la masse (= luminosité) d'une galaxie. Introduire des rayons non uniformes.

→ Revenir sur le choix de Σ_P . Il y a certainement des améliorations à trouver dans cette voie.

→ **Regarder ≠ types de connexités** possibles avec les complexes simpliciaux. Il y a encore plus à faire sur ce plan : nos premiers résultats montrent que l'on peut ainsi **augmenter la vitesse de percolation**.

Bibliographie

- [1] G. BIAU et L. DEVROYE, *Lectures on the Nearest Neighbor Method*. Springer, 2015, t. 246, ISBN : 978-3-319-25386-2. DOI : 10.1007/978-3-319-25388-6.
- [2] B. DARVISH, B. MOBASHER, D. SOBRAL, N. SCOVILLE et M. ARAGON-CALVO, "A Comparative Study of Density Field Estimation for Galaxies," *The Astrophysical Journal*, t. 805, n° 2, p. 121, 2015. DOI : 10.1088/0004-637X/805/2/121.
- [3] W. E. SCHAAP, "DTFE : the Delaunay Tessellation Field Estimator," thèse de doct., Proefschrift Rijksuniversiteit Groningen, 2007, p. 287.
- [4] STOICA, R., MARTÍNEZ, V., MATEU, J. et SAAR, E., "Detection of cosmic filaments using the Candy model," *Astronomy & Astrophysics*, t. 434, n° 2, p. 423-432, 2005. DOI : 10.1051/0004-6361/20042409.
- [5] M. PENROSE, *Random Geometric Graphs*. Oxford University Press, 2003, t. 5, ISBN : 9780198506263. DOI : 10.1093/acprof:oso/9780198506263.001.0001.
- [6] J. A. HARTIGAN, *Clustering Algorithms*. John Wiley & Sons, Inc., 1975, ISBN : 9780471356455.