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

NEO

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 feuillets ou « murs » de galaxies (surfaces);
3°) des « **filaments** » de galaxies (courbes).

Ces clusters délimitent de grandes régions quasiment vides de galaxies.

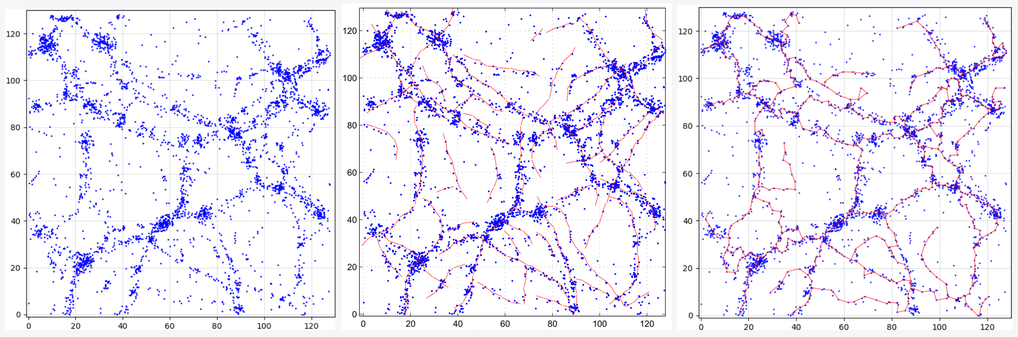

Figure 1. Nuage de points synthétique $\mathcal{X}_n \subset \mathbb{R}^2$ créé par les auteurs de l'article [4] et gracieusement fourni par R. Stoica, V. Martínez et E. Saar. Au milieu, le résultat de leur algorithme. À droite, le nôtre.

$\longrightarrow$ Clusters hiérarchiques $\implies$ 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 $\mathcal{X}_n = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ 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 $\mathcal{G}(\mathcal{X}_n, r)$ est le graphe (non orienté) :

$$\textit{Sommets} := \mathcal{X}_n; \qquad \textit{Arêtes} := \Big\{\{x_i, x_j\} \mid |x_i - x_j| < r\Big\}.$$

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

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

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

$\longrightarrow$ permet de généraliser la notion de graphe ainsi que de ***composante connexe***.

Percolation. Soit $\mathcal{X}_n := \mathcal{H}_1 \cup \{0\} \subset \mathbb{R}^d$ 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, $\mathcal{G}(\mathcal{X}_n, r)$ n'a aucune composante connexe infinie;
— $\forall r > r_c$, avec probabilité 1, $\mathcal{G}(\mathcal{X}_n, r)$ possède une composante connexe infinie.

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

$\longrightarrow$ **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 : $\dfrac{r_{\max}}{r_{\min}}$.

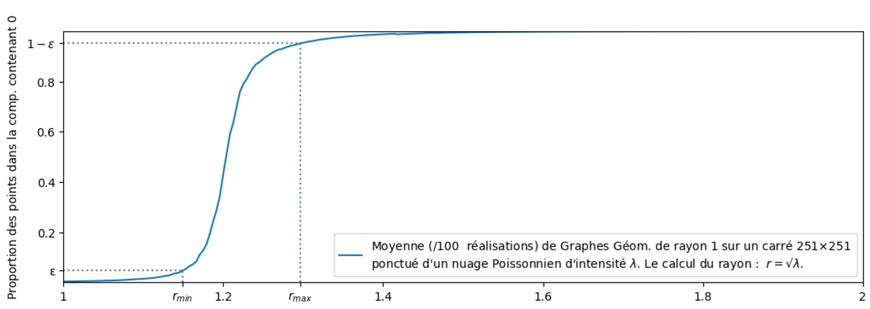

Figure 2. Avec $\varepsilon \leftarrow 0,05$, $r_{\min} = 1,15$ et $r_{\max} = 1,30$ soit $\frac{r_{\max}}{r_{\min}} \approx 1,13$.

$\longrightarrow$ **percolation rapide** $\implies$ 2 clusters voisins de densités proches bien distingués.
$\longrightarrow$ Soit $\mathcal{G}_{|}$ le graphe élagué des petites composantes (de taille $\leq$ *SeuilPerco* $\leftarrow 50$).

Niveaux de densité. *High-Density Clusters*$(h) := f^{-1}([h; +\infty[)$ [5], [6].
Soit $P \in [0;1]$, la *hauteur* h_P de densité associée : $h_P := \inf\left\{h \mid \int_{f \geq h} f(x)dx \leq P\right\}$.

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

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

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

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

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.

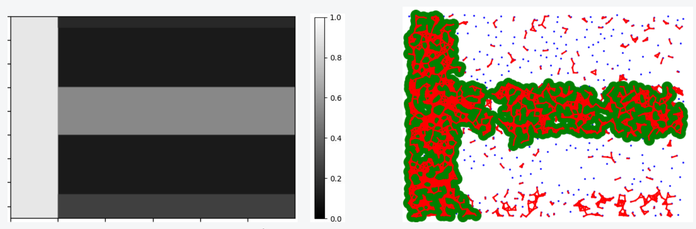

Figure 3. À gauche : les niveaux de densité théoriques. À droite : un nuage $\mathcal{X}_n$ de 2000 points générés i.i.d. selon f. Un graphe géométrique $\mathcal{G}(\mathcal{X}_n, r)$ est dessiné avec arêtes en rouge et le cluster associé Σ_r en vert. *SeuilPercolation* $\leftarrow 50$.

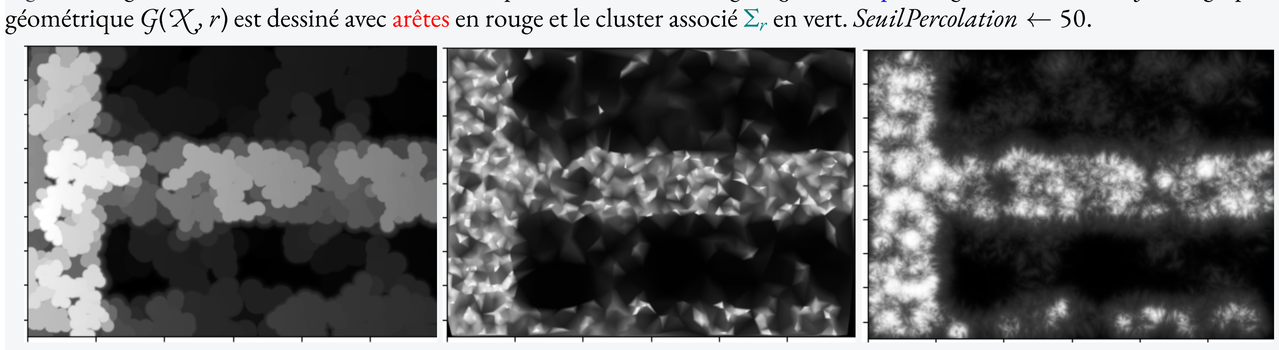

Figure 4. Trois estimateurs des niveaux de densité sur $\mathcal{X}_n$: notre **Graphe-Percol**; **Delaunay**; 10-**Plus-Proches-Voisins**.

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

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

Algorithme	$1-\hat{\mathcal{P}} > 0,617$		$1-\hat{\mathcal{P}} > 0,280$		$1-\hat{\mathcal{P}} > 0,169$		$1-\hat{\mathcal{P}} > 0,140$	
	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).

$\longrightarrow$ **Percolation rapide** $\implies$ recouvre bien le 'super-amas' avant de commencer sur le 'filament' médian.

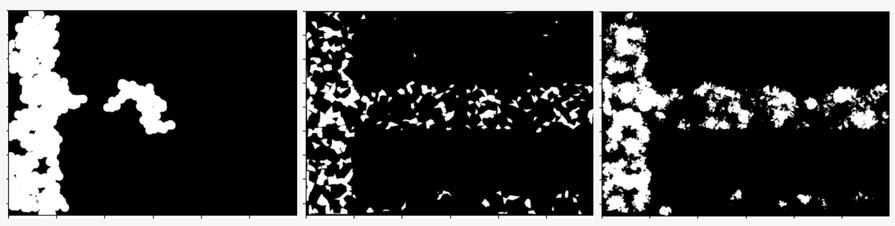

Figure 5. Clusters $\hat{\mathcal{C}}_P$ pour le seuil $1 - P = 0,617$. De gauche à droite : **Graphe-Percol**; **Delaunay**; 10-**Plus-Proches-Voisins**.

$\longrightarrow$ Graphes/complexes simpliciaux $\implies$ extraction plus facile de filaments. ***Cf.*** **image de dr. Figure 1**.

Algorithme d'extraction de filaments de galaxies

$\longrightarrow$ Partir d'une comp. connexe G de $\mathcal{G}_{|}$ et de son *Filament* (sous-gr.) déjà construit.
$\longrightarrow$ Décider d'un nombre de *Centres* à ajouter selon la taille de G.
$\longrightarrow$ Itérativement, un nouveau centre x est choisi de façon à minimiser :
$D[x] = \sum_{y \in \text{Nœuds}(G)} d\left(y,\ \textit{Filament} \cup \textit{Branche}_x\right)^2$
(où d est la distance induite sur le graphe et $\textit{Branche}_x$ le chemin le plus court de x à *Filament*).
$\textit{Filament} \leftarrow \textit{Filament} \cup \textit{Branche}_x$

$\longrightarrow$ Premier centre choisi : **moyenne de Fréchet** de G.

Conclusion

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

$\longrightarrow$ La percolation est un phénomène très rapide permettant de distinguer deux niveaux de densité proches.

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

Perspectives

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

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

$\longrightarrow$ Regarder $\neq$ 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.