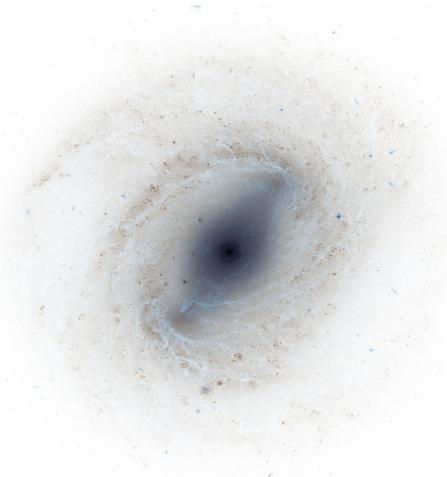


Fonctionnelles de Minkowski appliquées à la classification des galaxies

Julien Barrès, Sofiane Aïssani, Gabriel Aillet, Raphaël Seigmuller
Encadrant : Carlo Schimd (Laboratoire d'Astrophysique de Marseille)



Résumé

Nous présentons une analyse de l'utilisation des Fonctionnelles de Minkowski (MF) en tant qu'outil pour améliorer la classification des galaxies. Il s'agit d'une famille de mesures géométriques et topologiques permettant de caractériser la forme et la connectivité des sous-ensembles de \mathbb{R}^d .

La méthode mise en place repose sur l'Analyse en Composantes Principales (*Principal Component Analysis*, PCA) et la détection de *clusters* pour discriminer des groupes de galaxies de morphologies similaires parmi des jeux de données provenant du Hubble Space Telescope. Elle met en évidence de tels groupes sur un nombre assez restreint d'images de qualité variable. La comparaison visuelle des groupes obtenus montre qu'il est effectivement possible de classifier les images selon des critères morphologiques, mais elle montre également une grande sensibilité des fonctionnelles de Minkowski envers le bruit présent sur l'image. Les jeux de données choisis pour cette étude ne permettent pas de trouver de lien entre la classification obtenue et les caractéristiques physiques des galaxies.

Mots clés : galaxies : général, galaxies : structure, Fonctionnelles de Minkowski, méthodes : statistiques, méthode : analyse de données

Table des matières

Introduction	1
1 Contexte et objectifs	1
2 Données et méthode	2
2.1 Données	2
2.1.1 Images de simulation	2
2.1.2 HST NGC1512 High-res	2
2.1.3 COSMOS : HST-ACS Mosaic 1	2
2.1.4 COSMOS : HST-ACS Mosaic 2	3
2.2 Fonctionnelles de Minkowski	3
2.2.1 Définition et propriétés	3
2.2.2 Application des fonctionnelles sur \mathbb{R}^2	3
2.2.3 Implémentation en python	4
2.3 Traitement d'images	4
2.3.1 Altérations volontaires des images	4
2.3.2 Suppression des artefacts indésirables	6
2.4 Analyse en Composantes Principales	7
2.5 Classement des galaxies	9
2.5.1 “ k -means clustering” ou partitionnement par k -moyennes	10
2.5.2 Nombre de clusters optimal	10
3 Analyse et résultats	10
3.1 HST NGC1512 High-res : Analyse visuelle	10
3.2 Images de simulation : Impact des artefacts	11
3.3 COSMOS : HST-ACS Mosaic 2 : Analyse statistique	13
3.4 Discussion	14
Conclusion et perspectives	16
Références	17
A Produit de convolution	22
A.1 Définition	22
A.2 Filtrage d'images	22
B Programmes développés en python	23
C Images supplémentaires	24
C.1 Analyse de l'impact des altérations	24
C.2 Conclusion	26

Remerciements

Nous remercions notre encadrant Carlo Schimd pour avoir invité les étudiants de notre licence à prendre part à la recherche sur un sujet aussi ouvert et riche que celui de la classification des galaxies.

Nous remercions également notre professeur Xavier Bugaut pour son suivi et ses conseils au cours du semestre.

Enfin, nous remercions d'avance notre rapporteur M. Bethermin pour l'attention qu'il voudra bien porter à ce travail.

Introduction

La formation des galaxies occupe une place importante parmi les problématiques actuelles de la cosmologie. Il est désormais établi, depuis quelques décennies tout au plus, que la morphologie des galaxies, c'est-à-dire l'ensemble de leurs propriétés géométriques et topologiques, est porteuse d'informations physiques, notamment sur l'âge et le taux de formation stellaire des galaxies (BELL et al. 2012). Ces propriétés physiques sont précieuses car elles permettent d'affiner les modèles de formation de galaxies déjà existants ou d'en créer de nouveaux.

Cependant, les mécanismes régissant le lien entre morphologie et propriétés physiques des galaxies ne sont pas encore complètement compris. Cette relation reste un immense défi pour les campagnes d'observation de grande envergure en cours ou futures, conduites par des instruments comme WFC3 et ACS sur le Hubble Space Telescope, Hyper Suprime-Cam à Subaru, MUSE au VLT ou le très attendu James Webb Space Telescope. Les milliers ou millions d'images multi-couleur à haute résolution demanderont des méthodes de classification morphologique nouvelles et automatisées.

Nous proposons des recherches sur une nouvelle piste de classification morphologique des galaxies au travers de l'utilisation des Fonctionnelles de Minkowski (MF). Il s'agit d'une famille de mesures morphologiques issues du domaine de la géométrie intégrale, qui permettent de décrire les aspects géométriques et topologiques de parties de \mathbb{R}^d . Nous les définissons en sous-section 2.2.

La sous-section 2.3 est consacrée aux aspects de traitement d'images inhérents à notre méthode d'analyse morphologique. Elle expose des techniques permettant d'évaluer la robustesse des fonctionnelles de Minkowski face aux artefacts non désirés et de limiter l'impact de ces artefacts.

La méthode de classification des galaxies, que nous présentons en sous-section 2.4, couple le calcul des fonctionnelles de Minkowski avec des techniques issues du machine learning telle que l'Analyse en Composantes Principales (PCA) et le *clustering* pour discriminer des groupes de morphologies similaires parmi les ensembles d'images que nous exposons en sous-section 2.1.

Les résultats de cette démarche seront analysés et discutés en section 3.

1 Contexte et objectifs

Crée par Edwin Hubble entre 1925 et 1936, la *séquence de Hubble* range les galaxies selon quatre types visibles sur la figure 1. Ces quatre classes sont les galaxies *elliptiques* (de E0 à E7), *lenticulaires* (S0 et SB0), *spirales* normales ou barrées (respectivement de Sa à Sc et de SBa à SBc) et *irrégulières* (Irr). Cette première classification, qui est fondée sur des considérations purement visuelles, *ne garantit pas* que des galaxies d'un même groupe aient les mêmes propriétés physiques. Alors que la plupart des galaxies elliptiques sont des galaxies inactives ne formant pas d'étoiles (WUYTS et al. 2011), il existe des exemples de galaxies elliptiques actives. De même, alors que la plupart des galaxies spirales contiennent des zones de formation stellaire, il existe des galaxies spirales ne formant pas d'étoiles. Ces classes "de transition", bien que rares, montrent que la classification de Hubble *ne met pas en évidence* de séparation des galaxies selon des critères physiques. Il est possible que certaines classes doivent être découpées en plusieurs sous-classes afin de tenir compte de critères plus pertinents pour l'étude de l'évolution des galaxies.

Des tentatives de donner naissance à des classifications plus pertinentes ont vu le jour depuis. Ces tentatives emploient deux méthodes différentes : les méthodes *visuelles*, comme Galaxy Zoo qui s'appuie sur la participation d'internautes pour classer les galaxies (SIMMONS et al. 2016) et les méthodes *automatisées* qui s'appuient sur le traitement et l'analyse informatique des images.

Bien qu'elles permettent de distinguer des éléments plus subtils que les classifications automatiques, les méthodes de classification visuelles peuvent être subjectives et ne sont pas adaptées aux énormes jeux de données qui seront bientôt disponibles avec les futurs télescopes au sol ou spatiaux comme LSST, Nancy-Grace-Roman (anciennementWFIRST) ou SKA. Quant aux techniques automatisées, dont nous ne pourrons plus nous passer lorsque le flux de données deviendra trop important, certaines d'entre elles doivent faire usage de paramètres physiques qui ne sont pas précisément déterminés. De plus, elles éprouvent des difficultés à rendre compte des irrégularités telles que les barres

des spirales ou les amas stellaires au sein des galaxies en des temps de calcul raisonnables : c'est le cas de GALFIT (PENG et al. 2002).

La méthode utilisée couramment consiste à calculer des grandeurs morphologiques telles que *Concentration* (C), *Asymetry* (A) et *Clumpiness* (S) (CONSELICE 2003), dont la corrélation avec les propriétés physiques des galaxies a été démontrée (CONSELICE 2003 ; RODRIGUEZ-GOMEZ et al. 2018 ; PETH et al. 2016).

L'objectif de ce travail est donc de sonder la piste d'une méthode complémentaire à la méthode CAS, au travers de l'utilisation des fonctionnelles de Minkowski, qui ont déjà été appliquées à de nombreux problèmes de caractérisation morphologique dans divers domaines scientifiques comme la physique de l'état solide et la physique de la matière molle. Elles ont démontré leur efficacité en des temps de calcul corrects et sans supposition préalable sur les objets étudiés (K. R. MECKE 1997 ; EDER 2018 ; LEVCHENKO et al. 2016 ; PARKER et al. 2013 ; MANTZ, JACOBS et K. MECKE 2008). Il sera également envisagé de coupler les informations apportées par les fonctionnelles de Minkowski avec d'autres méthodes permettant de trouver les meilleurs critères de discrimination des galaxies.

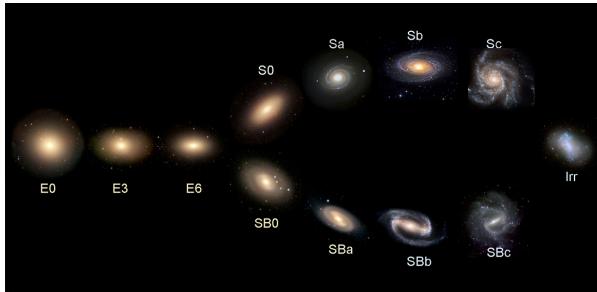


FIGURE 1 – Classification de Hubble de la morphologie des galaxies avant qu'elle ne soit complétée par De Vaucouleurs en 1959. Pour des raisons historiques, les galaxies elliptiques sont aussi appelées “galaxies primitives” et les galaxies restantes sont appelées “galaxies tardives”.

2 Données et méthode

2.1 Données

2.1.1 Images de simulation

Les huit images présentes dans ce jeu de données sont obtenues à partir de simulations *n*-corps de galaxies spirales vues de face, sans bruit, de taille 200×200 pixels², qui diffèrent par des détails morphologiques subtils (symétrie ou asymétrie, écartement des bras...). Elles ont été simulées par K. Kraljic (Institute for Astronomy, University of Edinburgh, Royal Observatory Edinburgh). Elles seront utilisées

en sous-section 3.2.

2.1.2 HST NGC1512 High-res

Ces images ont été prises par les caméras WFPC2 (Wide Field and Planetary Camera 2), NICMOS (Near Infrared Camera and Multi-Object Spectrometer) et FOS (Faint Object Spectrograph) du télescope spatial Hubble (HST) en plusieurs bandes spectrales (crédit : ESA/Hubble). Elles représentent la même galaxie NGC1512 en haute définition (1280×1280 pixels²), vue à travers les filtres de longueurs d'onde centrales 220 nm, 338 nm, 545 nm, 659 nm, 827 nm et 1600 nm. Elles sont brièvement utilisées pour faire une première analyse visuelle des fonctions de Minkowski sur des images de bonne qualité.

2.1.3 COSMOS : HST-ACS Mosaic 1

Les images de galaxies présentes dans ce jeu de données ont été prises par la caméra Advanced Camera for Surveys (ACS) du Hubble Space Telescope (HST) (KOEKEMOER et al. 2007 ; MASSEY et al. 2010). Elles ont été observées pendant les cycles 12 et 13 de l'étude, c'est-à-dire lors de la période juillet 2003 - juin 2005, et prises avec le filtre F814W centré sur la longueur d'onde 8211.2 nm. Les galaxies sont sélectionnées à partir du catalogue COSMOS 2015 (LAIGLE et al. 2016) répertoriant plus d'un demi-million d'objets dans une surface de deux degrés carré centré sur le point de coordonnées RA = $10^{\text{h}}00^{\text{m}}28.6^{\text{s}}$ et DEC = $+02^{\text{h}}12^{\text{m}}21.0^{\text{s}}$. Nous en extrayons les galaxies ayant un redshift photométrique $0.05 < z < 1$ et de magnitude apparente $i < 24$, de manière à concentrer l'analyse sur des galaxies assez bien visibles. Environ 0.35% aléatoires de ce total sont conservées, ce qui ramène le nombre d'objets à 808¹. Les 808 galaxies sont réparties en 4 intervalles de redshift photométrique ($[0.05, 0.25]$, $[0.25, 0.50]$, $[0.50, 0.75]$, $[0.75, 1.00]$) puis leurs images sont récupérées depuis le site internet https://irsa.ipac.caltech.edu/data/COSMOS/index_cutouts.html en choisissant un “Uniform Cutout Size” égal à 10 arcsec. Elles sont ensuite sélectionnées visuellement une par une et rognées à l'aide d'un programme python. Les images contenant des défauts ou un bruit trop important sont jetées : il reste au final 241 images de galaxies prêtes à être analysées. La grande proportion de galaxies jetées indique qu'il aurait été plus intéressant de choisir une taille de coupe plus petite pour avoir des images de meilleure résolution et moins dominées par le bruit. C'est ce qui a été fait pour le jeu de données suivant “COSMOS : HST-ACS Mosaic 2”.

¹. La proportion de galaxies conservée est choisie de manière à ne pas dépasser 1000 galaxies pour des raisons de débit de connexion.

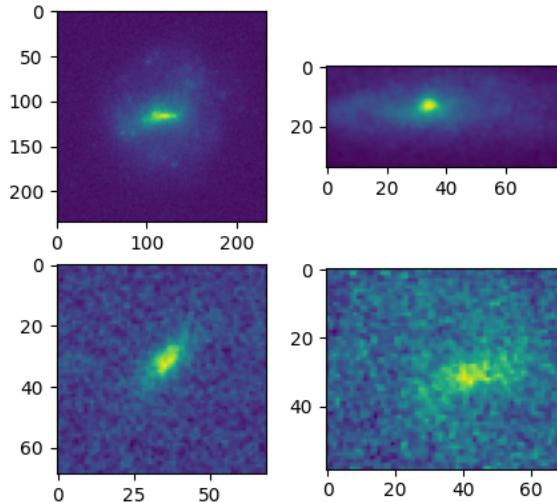


FIGURE 2 – Quatre images du jeu de données “COSMOS : HST-ACS Mosaic 2” montrant une quantité inégale de bruit.

2.1.4 COSMOS : HST-ACS Mosaic 2

Pour ce jeu de données, le protocole est exactement le même que pour le jeu précédent “COSMOS : HST-ACS Mosaic 1”. Depuis le même catalogue, 0.4% des galaxies vérifiant la condition de visibilité sont choisies aléatoirement, ce qui ramène le nombre d’objets à 932. Le “*Uniform Cutout Size*” est choisi égal à 7 arcsec, puis les images récupérées sont sélectionnées visuellement et rognées une par une. Au final, 804 images de galaxies sont conservées.

Dans ce jeu de données tout comme dans le précédent, les images sont inégalement bruitées comme la figure 2 le montre. Cette disparité risque d’influer sur la classification, car dans le cas où le bruit aurait un impact non négligeable sur les MF obtenues, il se pourrait que les galaxies soient classées par niveau de bruit présent sur leur image plutôt que par morphologie. Le choix du jeu de données est donc discuté en sous-section 3.4.

2.2 Fonctionnelles de Minkowski

2.2.1 Définition et propriétés

Dans un espace métrique (E, δ) de dimension d , où E est un espace vectoriel et δ une distance sur E , il existe $d+1$ fonctionnelles de Minkowski V_0, \dots, V_d qui décrivent la géométrie et la topologie des parties de E par des valeurs réelles. Ces fonctionnelles peuvent être définies par des intégrales de courbure en utilisant des notions de géométrie dif-

férentielle et de géométrie intégrale pourvu que l’espace métrique définisse une variété différentielle (MANTZ, JACOBS et K. MECKE 2008). Cependant, le calcul des fonctionnelles peut se généraliser à des parties non différentiables (K. R. MECKE, BUCHERT et WAGNER 1994).

Le théorème de caractérisation de Hadwiger (HADWIGER 1957) établit que toute fonctionnelle \mathcal{F} des parties de E dans \mathbb{R} additive, invariante par rotation et translation et continue est une combinaison linéaire des $d+1$ Fonctionnelles de Minkowski (MF), notées V_μ . En notation compacte, toute fonctionnelle s’écrit

$$\mathcal{F} = \sum_{\mu=0}^d \lambda_\mu V_\mu$$

si et seulement si elle vérifie les trois propriétés suivantes.

1. Additivité²

Pour tous sous-ensembles $A, B \subset E$,

$$V_\mu(A \cup B) = V_\mu(A) + V_\mu(B) - V_\mu(A \cap B)$$

2. Invariance par rotation et translation²

Pour tout sous-ensemble A de E , pour toute application g du groupe des rotations et des translations,

$$V_\mu(g(A)) = V_\mu(A)$$

3. Continuité²

Pour toute suite de sous-ensembles $(K_n)_{n \in \mathbb{N}}$ qui converge vers K pour la métrique de Hausdorff, la suite des $V_\mu(K_n)$ converge vers $V_\mu(K)$.

Grâce à ces trois propriétés, les applications V_μ apportent une caractérisation complète de la morphologie des parties de E .

2.2.2 Application des fonctionnelles sur \mathbb{R}^2

Dans le plan \mathbb{R}^2 muni de la distance euclidienne, les MF d’une région continue A correspondent à l’aire surfacique V_0 , le périmètre V_1 et la caractéristique d’Euler V_2 .

$$\begin{aligned} F = V_0(A) &= \frac{1}{4\pi} \int_A d\Omega \\ U = V_1(A) &= \frac{1}{16\pi} \int_{\partial A} dl \\ \chi = V_2(A) &= \frac{1}{8\pi^2} \int_{\partial A} \kappa dl = N_{\text{régions connexes}} - N_{\text{trous}} \end{aligned}$$

avec κ la courbure de la frontière ∂A de l’ensemble A . L’égalité sur la caractéristique d’Euler est donnée par le

2. Les trois propriétés sont vraies pour tout $\mu \in \{0, \dots, d\}$.

théorème de Gauss-Bonnet qui établit un lien entre la géométrie et la topologie des parties de \mathbb{R}^d .

En pratique, l'analyse des MF s'effectue sur une images plane, assimilable à un champ scalaire $f : \mathbb{R}^2 \mapsto \mathbb{R}$ discrétré qui représente l'intensité. Pour analyser un tel champ, on mesure les MF des ensembles $\Sigma_\nu := \{(i, j) \in \mathbb{R}^2 : f(i, j) > \nu\}$. Autrement dit, dans le cas d'une image, le paramètre d'analyse ν est un seuil de luminosité variable et les MF des ensembles Σ_ν sont calculées pour chaque valeur de ν . Une illustration est représentée sur la figure 3. Pour $\mu \in \{0, 1, 2\}$, on appellera “fonctionnelle de Minkowski au seuil ν ” la valeur $V_\mu(\Sigma_\nu)$ et “fonction de Minkowski” l'application $\nu \mapsto V_\mu(\Sigma_\nu)$. Dans la suite de ce rapport, l'analyse d'une image consistera à calculer les fonctions de Minkowski de l'image.

2.2.3 Implémentation en python

L'implémentation en python³ du calcul des MF est basée sur l'article MANTZ, JACOBS et K. MECKE 2008. Elle utilise une approche “diviser-pour-régner” en profitant de la propriété d'additivité des MF pour les calculer sur des cas simples et sommer les résultats afin d'obtenir le résultat total.

L'algorithme implémenté est l'algorithme du “marching square” (carré défilant). Pour chaque valeur de seuil ν , l'image binarisée représentant l'ensemble Σ_ν est découpée en morceaux de quatre pixels. Chaque morceau correspond à une configuration parmi les 16 possibles représentées sur la figure 4. Pour améliorer les performances de l'algorithme, chaque morceau est également paramétré par des valeurs de “poids” calculées en fonction des luminosités relatives des quatre pixels qui le composent. Les aires, périmètres et caractéristiques d'Euler de chaque morceau sont alors déterminées à l'aide de formules géométriques simples en utilisant les poids calculés. Enfin, les valeurs totales $V_\mu(\Sigma_\nu)$ sont calculées en sommant les contributions de chaque morceau (cf. propriété d'additivité).

2.3 Traitement d'images

Le calcul des fonctionnelles de Minkowski a été réalisé sur des ensembles d'images de qualité variable. Globalement, la résolution des images est souvent assez faible et les régions couvertes par le champ de vision contiennent des artefacts lumineux non désirés. De plus, qu'elles soient acquises par des télescopes terrestres ou spatiaux, ces images sont toujours contaminées par du bruit qui peuvent influer sur les valeurs des fonctionnelles de Minkowski.

Il a donc été nécessaire de réaliser des tests et des traitements sur les images d'entrée. Dans la suite de ce rapport, le terme image est utilisé pour décrire une matrice

$M \in \mathcal{M}_{n,m}(\mathbb{R})$ avec n et m ses dimensions en pixels et $M_{a,b}$ la valeur de l'intensité lumineuse du pixel en position (a, b) .

L'intervalle des valeurs de ces intensités lumineuses varie entre les jeux de données⁴. Ainsi, une uniformisation des valeurs a été réalisée afin que pour toutes les images, l'intensité lumineuse soit codée par des valeurs comprises entre 0 et 255. Ces valeurs normalisées sont d'unité arbitraire ; il y a donc une perte d'information sur la luminosité absolue de l'objet, mais elle n'influe pas sur les mesures morphologiques.

Les altérations d'images présentées dans la suite de cette section ont trois buts :

- quantifier l'impact de telles altérations sur le résultat du calcul des fonctionnelles de Minkowski ;
- harmoniser les caractéristiques des images d'un même jeu de données pour limiter la dispersion due aux inégalités de résolution, de bruit, d'inclinaison ;
- affranchir les images d'un maximum de leurs artefacts indésirables.

2.3.1 Altérations volontaires des images

Quantifier l'impact du bruit ou d'autres altérations sur les fonctionnelles de Minkowski est primordial pour que l'étude soit reproductible sur plusieurs jeux de données différents. Les tests suivants ont été réalisés sur des images de simulation de galaxies vues de face, sans bruit, ayant une haute résolution. Ce jeu d'images de test est présenté en section 2.1.1.

Dégradation de la résolution Pour dégrader une image $M_{n,m}$ d'un facteur d , l'image est scindée en $[n/d] \times [m/d]$ matrices de taille $d \times d$.

Chacune de ces matrices est ensuite traduite en un pixel en prenant la valeur moyenne de ses coefficients et une nouvelle image $M' \in \mathcal{M}_{[n/d] \times [m/d]}$ est construite par concaténation des pixels calculés. Une partie des pixels de la matrice originelle est perdue⁵, il a donc toujours été vérifié qu'aucune information nécessaire ne se trouvait dans cette zone éliminée.

À noter qu'une solution parallèle aurait été de lisser fortement les images à haute résolution afin de perdre suffisamment d'information pour pouvoir les comparer à des images de plus faible résolution, mais la solution de la dégradation a été choisie car elle est bien plus rapide.

Bruit Les images acquises par les caméras sont sujettes à différents types de bruit provenant de sources diverses

4. $I \in [0, 1]$ pour le jeu de données "Images de simulation", $I \in [0, 10^6]$ pour le jeu de données "COSMOS : HST-ACS Mosaic 2". Les jeux de données sont présentés en section 2.1

5. $[d \bmod n] \times m + n \times [d \bmod m] - [d \bmod m] \times [d \bmod n]$ pixels sont perdus sur le bord droit et bas de l'image

3. <https://github.com/moutazhaq/minkfncts2d>

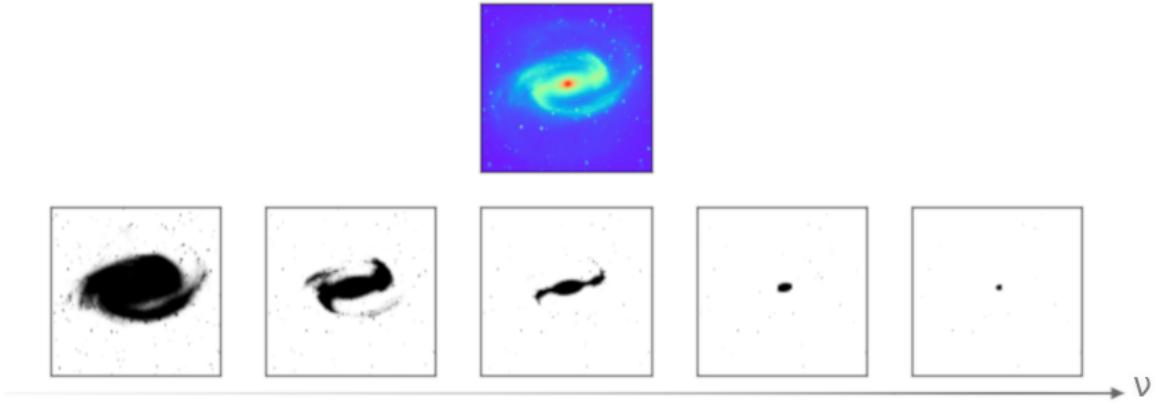


FIGURE 3 – Image de la galaxie NGC1300 prise par le télescope spatial Hubble (en haut) et allure des ensembles Σ_ν correspondant, représentés en noir, en fonction des cinq valeurs du seuil ν . Les MF sont calculées sur ces ensembles.

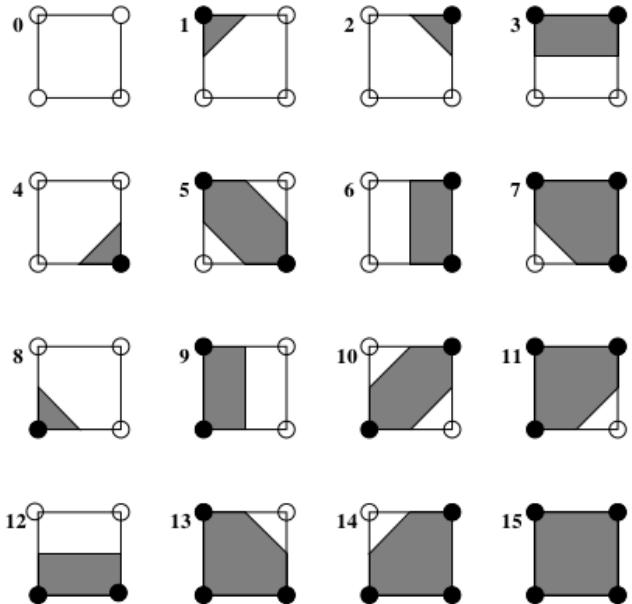


FIGURE 4 – Les seize configurations possibles pour un carré de 2×2 pixels² sur une image binarisée. Crédit : MANTZ, JACOBS et K. MECKE 2008

comme les effets quantiques dus à la nature corpusculaire des photons et des électrons, ou les effets thermiques. Dans tous les cas, le bruit consiste en des petites variations d'intensité lumineuse réparties de manière stochastique sur l'image et vérifiant certaines lois de probabilités. En tenant compte de ces lois, il a été possible de modéliser quelques types de bruit susceptibles de contaminer les images des jeux de données. Ces bruits sont représentés sur la figure 8.

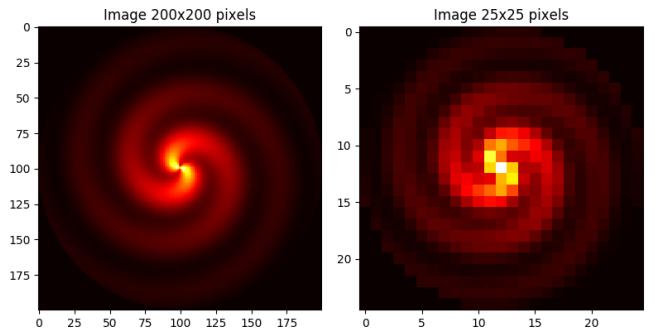


FIGURE 5 – Exemple de dégradation de facteur 8. La deuxième image correspond au résultat de l'application de la fonction sur la première. La sous-section 3.2 détaille le calcul des MF sur des images dégradées.

- **Bruit poivre et sel** Le principe du bruit poivre et sel est d'attribuer la valeur minimale (0) ou la valeur maximale (255) à certains pixels choisis aléatoirement sur l'image. Ces valeurs correspondent aux extrêmes d'intensité lumineuse (respectivement blanc et noir). Pour une probabilité p donnée par l'utilisateur, chaque pixel est visité indépendamment des autres et a une probabilité p de devenir noir ou blanc. Un pixel modifié aura autant de chance de prendre la valeur 0 ou que de prendre la valeur 255.

- **Bruit gaussien additif** Ce bruit suit une loi normale centrée d'écart-type σ choisi par l'utilisateur. La densité de probabilité est définie par $P(I) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{I^2}{2\sigma^2}}$. Un masque, c'est-à-dire une image d'intensité nulle de même dimension que l'image d'origine, est appliqué à l'image de bruit pour empêcher les pixels avec des intensités négatives de contribuer au calcul.

sions que l'image, est créé puis à chaque pixel du masque est attribuée une valeur tirée aléatoirement suivant la loi normale. Les valeurs décimales sont arrondies à l'unité. Enfin, le masque est ajouté à l'image et les pixels sortant de l'intervalle $\{0, \dots, 255\}$ ($I_{a,b} < 0$ ou $I_{a,b} > 255$) prennent respectivement les valeurs 0 et 255.

- **Bruit de Poisson homogène** Le bruit de Poisson homogène est généré en utilisant une loi de Poisson d'espérance \bar{I} , définie par $P(I = k) = \frac{\bar{I}^k}{k!} e^{-\bar{I}}$. De la même manière que pour le bruit gaussien, le bruit de Poisson se comporte comme un masque ajouté à l'image, dont chaque pixel prend une valeur aléatoire suivant la loi de Poisson recentrée. Dans le cas du bruit de Poisson homogène, l'espérance est la même pour chaque pixel et est déterminée au préalable par l'utilisateur.

- **Bruit de Poisson inhomogène** Le bruit de Poisson inhomogène est similaire au bruit de Poisson homogène, mais l'espérance dépend de la valeur de chaque pixel de l'image. Un exemple simple est d'associer à chaque pixel $M_{a,b}$ une espérance égale à $kM_{a,b}$, avec k un coefficient de proportionnalité. Cela a pour effet de favoriser la perturbation des pixels les plus lumineux par rapport aux pixels les moins lumineux.

Inclinaison de l'image Le plus souvent, les galaxies des jeux de données étudiés par la suite ne sont pas en face directe et peuvent être inclinées dans l'espace. Afin de vérifier l'impact de l'inclinaison sur les fonctionnelles de Minkowski, des images simulées ont été inclinées par rapport à plusieurs axes de l'espace.

La figure 6 montre l'exemple une rotation d'angle $\theta = 3\pi/8$ du plan de l'image par rapport à l'axe des abscisses. Pour chaque pixel (x, y) , sa nouvelle position est $(x', y') = (x, (y - h/2) \times \cos(\theta) + h/2)$, avec h la hauteur de l'image en pixels. La plupart du temps, la position en y devient décimale et il est donc nécessaire de faire une interpolation sur les pixels qui doivent être envoyés sur une même coordonnée.

2.3.2 Suppression des artefacts indésirables

Rognage La majorité des images acquises par les télescopes sont constituées d'une source primaire, la galaxie dont on cherche à mesurer la morphologie, et de sources secondaires qui sont en général des étoiles, des galaxies, des satellites ou des rayons cosmiques sur la ligne de vue. Le cumul des sources secondaires engendre une erreur importante sur les fonctionnelles de Minkowski calculées, puisque chaque source contribue à la variation du périmètre, de l'aire et de la caractéristique d'Euler.

La méthode utilisée pour supprimer la majeure partie des artefacts secondaires de l'image est un rognage simple qui

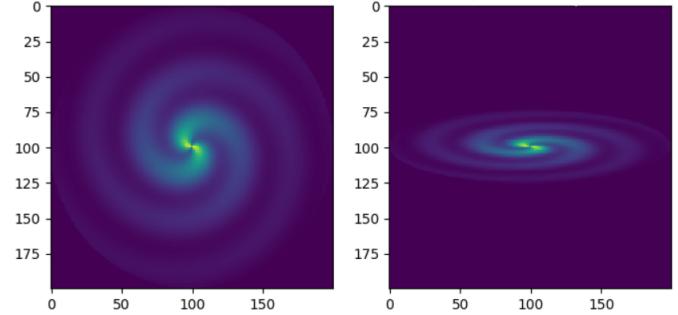


FIGURE 6 – Inclinaison d'angle $3\pi/8$ du plan de l'image par rapport à l'axe des abscisses. La sous-section 3.2 détaille le calcul des MF sur des images inclinées.

consiste à découper l'image pour ne garder que sa source primaire.

Il aurait été envisageable d'utiliser des méthodes plus avancées, comme la détection de sources secondaires proposée par le logiciel SExtractor (BERTIN et ARNOUTS 1996), mais le rognage visuel est apparu suffisant et moins complexe. Nous avons développé un programme python pour faciliter cette tâche (voir source en annexe B).

Lissage Le bruit est parmi les dégradations les plus récurrentes et les plus problématiques puisqu'il est créateur d'artefacts lumineux auxquels les fonctionnelles de Minkowski sont potentiellement très sensibles. Ces artefacts sont répartis sur l'image selon un processus stochastique, donc pas ou peu prévisible.

Pour diminuer l'impact du bruit, plusieurs méthodes existent. Deux techniques couramment utilisées sont les filtrages par convolution⁶ et le filtrage passe-bas, qui exploite la transformée de Fourier de l'image pour supprimer les composantes de haute fréquence, principales représentantes du bruit.

D'autres types de lissage sont très efficaces contre certains types de bruit : par exemple, le lissage médian et le lissage moyen éliminent respectivement le bruit poivre et sel et le bruit de Poisson. Dans cette étude, le lissage a été réalisé en calculant le produit de convolution de l'image avec un noyau gaussien⁷ à symétrie de révolution. Ce choix sera justifié par les jeux de données utilisés (sous-section 2.1), dont le bruit est majoritairement gaussien. La taille du noyau est égale à $8\sigma + 1$, où σ est l'écart-type choisi par l'utilisateur. La figure 7 affiche le noyau correspondant à $\sigma = 2$.

6. Voir en annexe A pour une explication plus complète sur la convolution.

7. Remarque : dans ce cas précis, le filtrage par convolution est strictement équivalent à un filtrage passe-bas. Mais nous distinguons tout de même les deux méthodes car elles ne sont pas toujours équi-

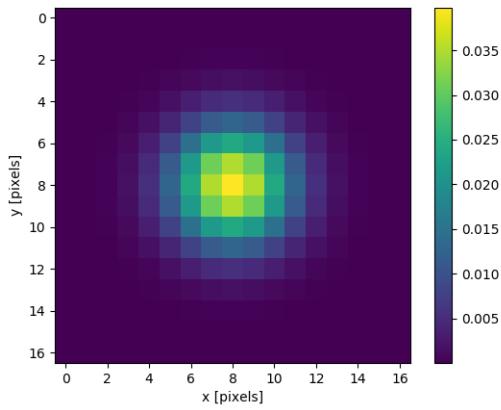


FIGURE 7 – Exemple d'un noyau gaussien de taille 17×17 pixels².

Augmenter la taille du noyau revient à faire un lissage plus fort. Cela permet de diminuer l'intensité des artefacts ponctuels au prix d'une perte de netteté des frontières de l'image, qui est floutée.

Contraste Toujours dans l'optique de réduire l'impact du bruit sur le calcul des fonctionnelles de Minkowski, une augmentation du contraste peut être judicieuse. Elle permet une meilleure distinction entre la galaxie qui sera d'intensité plus importante et le reste de l'image, dominé par le bruit, qui sera d'intensité moins importante. L'augmentation du contraste se fait en calculant l'image de chaque pixel par une application qui diminue les petites valeurs et augmente les grandes valeurs. De nombreuses applications peuvent convenir; un exemple simple et courant est la fonction tangente hyperbolique. Un résultat de l'augmentation du contraste est représenté dans le coin bas droite de la figure 8.

D'autres méthodes usuelles de modification du contraste utilisent la fonction racine carrée, la fonction carré ou le sinus hyperbolique.

2.4 Analyse en Composantes Principales

L'analyse visuelle des fonctions de Minkowski permet une première interprétation mais n'est pas suffisante pour détecter d'éventuelles familles de galaxies similaires. En effet, les trois courbes étant discrétisées en plus d'une centaine de points chacune, un grand nombre de quantités sont évaluées sur chaque image et il est difficile de savoir où regarder pour trouver des tendances et discriminer les galaxies.

L'Analyse en Composantes Principales (PCA) permet de déterminer, pour un jeu de n individus et p variables

valentes.

X_1, \dots, X_p évaluées sur chaque individu, les combinaisons linéaires des X_j ayant la plus grande variance. Ces nouvelles variables, appelées composantes principales (PC) permettent de discriminer au mieux les n individus.

La PCA a été réalisée en prenant comme individus les n images de galaxies du jeu de données et comme variables chaque valeur des fonctionnelles de Minkowski calculées pour des seuils de luminosité allant de 0 à 255. Il est à noter que le choix des variables aurait pu être fait différemment. En effet, puisque les fonctions de Minkowski sont généralement lisses, il est déjà clair que les variables représentant les fonctionnelles de Minkowski à des seuils voisins seront toujours proches et coderont donc une information redondante.

Pour déterminer les composantes principales, on réunit l'ensemble des données dans une matrice $D \in \mathcal{M}_{n,p}(\mathbb{R})$ dont le coefficient d_{ij} est égal à la valeur prise par la j -ème variable pour le i -ème individu.

La PCA repose sur le calcul des coefficients de corrélation linéaire entre chaque couple de variables. Le coefficient de corrélation linéaire entre la variable X_j et la variable X_k , noté r_{jk} , est défini de la manière suivante

$$r_{jk} = \frac{1}{n-1} \frac{\sum_{i=1}^n (d_{ij} - \bar{X}_j)(d_{ik} - \bar{X}_k)}{\sigma_j \sigma_k} \quad (1)$$

Le calcul de la totalité des coefficients r_{jk} se traduit en un unique produit matriciel en remarquant que

$$r_{jk} = \frac{1}{n-1} \sum_{i=1}^n \hat{d}_{ij} \hat{d}_{ik} \quad (2)$$

où \hat{d}_{ij} est défini comme $\frac{d_{ij} - \bar{X}_j}{\sigma_j}$. Les \hat{d}_{ij} sont les coefficients de la matrice de données "réduite" $\hat{D} \in \mathcal{M}_{n,p}(\mathbb{R})$. La réduction est nécessaire car elle permet d'effectuer la suite des calculs sur des grandeurs sans unité (et donc comparables). En nommant $R \in \mathcal{M}_p(\mathbb{R})$ la matrice contenant les coefficients r_{jk} , appelée matrice de corrélation, on a d'après l'équation 2

$$R = \frac{1}{n-1} \hat{D} {}^t \hat{D} \quad (3)$$

Un passage à la transposée dans l'équation 3 nous informe que la matrice R est une matrice symétrique réelle. D'après le théorème spectral, elle est donc diagonalisable dans une base orthonormée. De plus, les coefficients r_{jj} sont tous égaux à 1 d'après l'équation 1, donc la trace de R est égale à p .

Soient P une matrice orthogonale et $\Delta \in \mathcal{M}_p(\mathbb{R})$ diagonale telles que $\Delta = {}^t P R P$. Les vecteurs propres de R , qui sont les colonnes de P , sont des nouvelles variables qui sont combinaisons linéaires des anciennes variables et dont

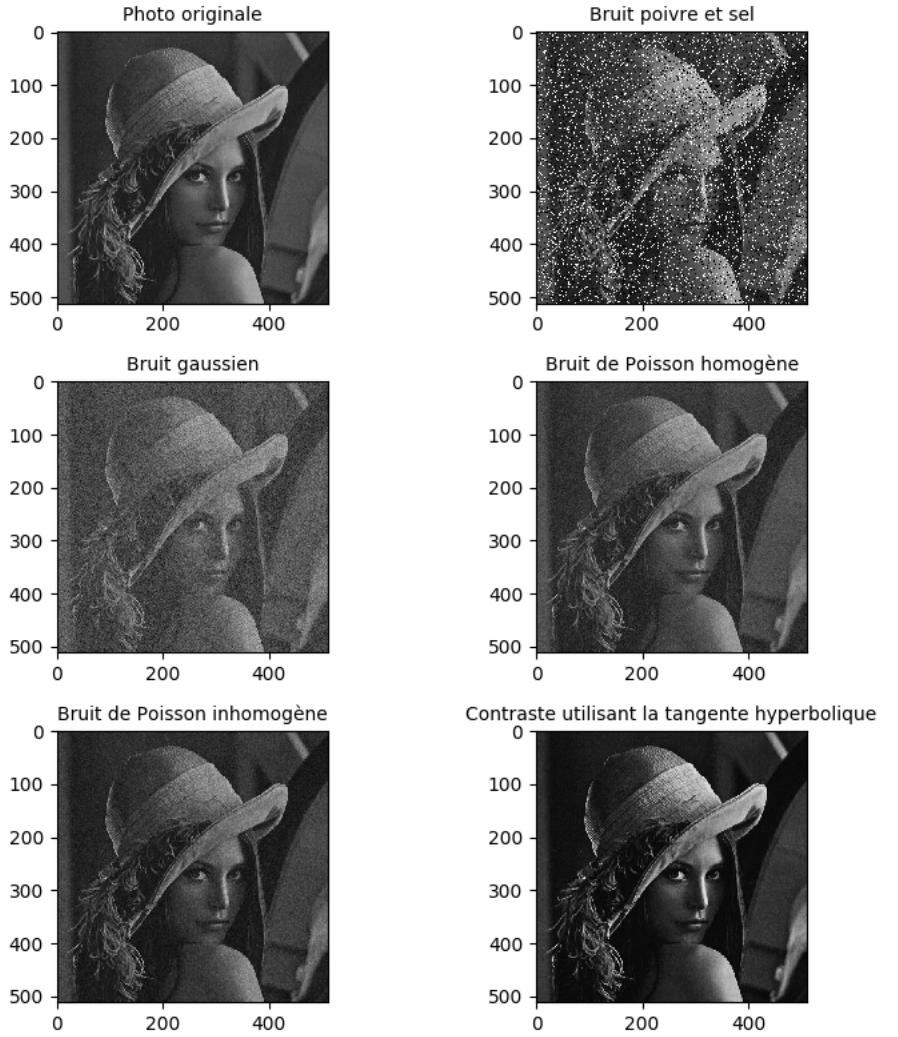


FIGURE 8 – Exemples d’application des différents bruits et du contraste appliqués à l’image originale.

la variance est maximale : on parle de "directions d'inertie maximale", ou composantes principales.

Chaque valeur propre δ_j de R , contenue dans Δ , représente la variance de son vecteur propre associé. Par invariance de la trace par changement de base, on a immédiatement

$$\sum_{j=1}^p \delta_j = p \quad (4)$$

Les p variables initiales donnent ainsi p nouvelles variables décorrélées dont la dispersion peut être quantifiée par $\frac{\delta_j}{p}$.

Très souvent, seules quelques composantes principales suffisent pour contenir la quasi-totalité de l'information. L'intérêt de la PCA est qu'on peut ainsi choisir de conserver autant de variables qu'il faut pour expliquer 90% de la

variance totale par exemple. Les variables conservées permettent alors de discriminer les individus selon des critères beaucoup moins nombreux et plus pertinents.

Exemple La PCA a été implémentée en python et testée sur les données présentées en figure 9.

On omet la variable "Modele" qui n'est pas numérique. Notons X_1, \dots, X_6 les six autres variables dans l'ordre où elles sont affichées. Après réduction de la matrice de données et diagonalisation de la matrice de corrélation, les valeurs propres obtenues sont triées par ordre décroissant et affichées sur l'histogramme de la figure 10. Notons X'_1, \dots, X'_6 les composantes principales, que nous confondrons avec les vecteurs propres de la matrice de passage.

Pour un individu quelconque i représenté par sa ligne L_i dans la matrice de données réduite \hat{D} , le produit scalaire

Modèle	CYL	PUISS	LONG	LARG	POIDS	V_MAX
Alfasud TI	1350	79	393	161	870	165
Audi 100	1588	85	468	177	1110	160
Simca 1300	1294	68	424	168	1050	152
Citroen GS Club	1222	59	412	161	930	151
Fiat 132	1585	98	439	164	1105	165
Lancia Beta	1297	82	429	169	1080	160
Peugeot 504	1796	79	449	169	1160	154
Renault 16 TL	1565	55	424	163	1010	140
Renault 30	2664	128	452	173	1320	180
Toyota Corolla	1166	55	399	157	815	140
Alfetta-1.66	1570	109	428	162	1060	175
Princess-1800	1798	82	445	172	1160	158
Datsun-200L	1998	115	469	169	1370	160
Taunus-2000	1993	98	438	170	1080	167
Rancho	1442	80	431	166	1129	144
Mazda-9295	1769	83	440	165	1095	165
Opel-Rekord	1979	100	459	173	1120	173
Lada-1300	1294	68	404	161	955	140

FIGURE 9 – Matrice de données choisie pour tester les algorithmes d’analyse en composantes principales.

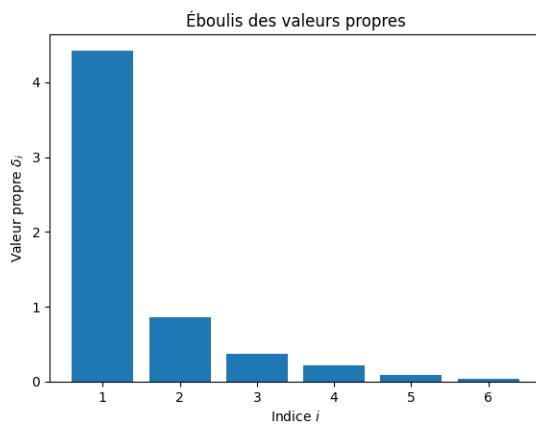


FIGURE 10 – Histogramme des valeurs propres de la matrice de corrélation associée aux données de la figure 9.

$L_i \cdot X'_j$ donne la valeur prise par la variable X'_j pour l’individu i . Faire le calcul du produit scalaire pour chaque individu et chaque variable revient à faire un unique produit matriciel

$$\hat{D}' = \hat{D}P \quad (5)$$

où \hat{D}' est la nouvelle matrice de données réduite, c'est-à-dire que son coefficient \hat{d}'_{ij} est égal à la valeur prise par la composante principale X'_j pour l’individu i .

Au vu de l’histogramme, il est clair que la dispersion totale du nuage est expliquée principalement par les variables X'_1 et X'_2 associées aux valeurs propres δ_1 et δ_2 . La somme $\delta_1 + \delta_2$ représente environ 88% de la variance totale : il est donc raisonnable, si l’on cherche à minimiser le nombre de variables, de projeter tous les individus sur le plan engendré par X'_1 et X'_2 . Cela revient à tronquer \hat{D}' en ne gardant

que ses deux premières colonnes. La figure 11 montre la projection de chaque individu sur X'_1 et X'_2 . Cette projection est “la plus proche de la réalité” au sens où son plan de projection minimise la somme des carrés des distances aux individus.

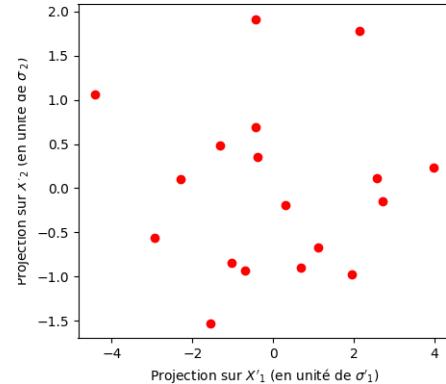


FIGURE 11 – Projections de chaque individu sur les deux composantes principales les plus dispersives.

Protocole

- Normaliser l’intensité des pixels des images entre 0 et 255,
- Faire le calcul des fonctions de Minkowski discrétisées en 100 valeurs régulièrement espacées de l’intervalle $[0, 255]$ sur chaque image,
- Construire la matrice D contenant, sur chaque ligne L_i , la concaténation des 300 valeurs des MF de l’image i ,
- Réaliser la PCA pour récupérer les variables les plus pertinentes,
- Évaluer les valeurs des composantes principales sur chaque image,
- Chercher à distinguer des groupes de points proches pour la distance euclidienne dans l’espace généré par les vecteurs propres donnés par la PCA.

Ce dernier point est décrit en section 2.5 : il s’agit de l’étape du classement (ou *clustering* des galaxies).

2.5 Classement des galaxies

L’étape suivant l’obtention des nouvelles coordonnées des individus dans les espaces propres donnés par la PCA est la séparation et discrimination de ces individus en plusieurs groupes. Cette étape est facilitée par le fait que la nouvelle base est optimale, au sens où elle maximise la variance des individus et permet donc un partitionnement plus efficace. Les paramètres définissant ces “clusters”⁸ d’individus dis-

8. Remarque : le terme “cluster” est un terme commun dans le vocabulaire du machine learning, mais il ne doit pas être confondu

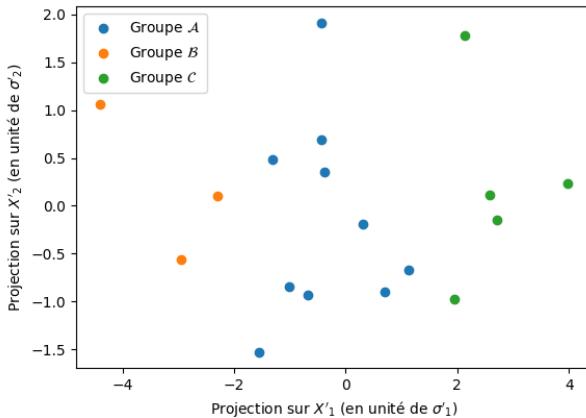


FIGURE 12 – Partitionnement des individus de la matrice d'exemple en trois clusters.

joints permettront à terme de définir les critères caractérisant une classification morphologique optimale des galaxies.

2.5.1 “ k -means clustering” ou partitionnement par k moyennes

L'approche qui a été utilisée pour partitionner les individus est une approche de machine learning, appelée “ k -means clustering”.

L'algorithme “ k -means”, tiré de la librairie scikit-learn⁹, divise l'espace des individus en k clusters chacun défini par son centroïde, un point de l'espace des individus. Chaque individu appartient au cluster défini par le centroïde dont il est le plus proche. De manière itérative, à partir d'une position aléatoire des k centroïdes, l'algorithme détermine un partitionnement optimal, c'est-à-dire un partitionnement qui minimise la somme des carrés des distances euclidiennes des individus aux centroïdes des clusters auxquels ils appartiennent. On appelle cette grandeur à minimiser l'*inertie* : dans le formalisme du *machine learning*, elle correspond à la fonction de coût de l'algorithme. La figure 12 montre le partitionnement obtenu pour un choix de trois clusters, sur l'ensemble des individus de l'exemple présenté en sous-section 2.4.

2.5.2 Nombre de clusters optimal

L'algorithme du “ k -means clustering” permet de trouver un partitionnement optimal pour n points de l'espace, en

avec les “clusters of galaxies” (amas de galaxies). Dans toute la suite, “cluster” signifiera toujours “classe” au sens du machine learning.

9. <https://scikit-learn.org/stable/>

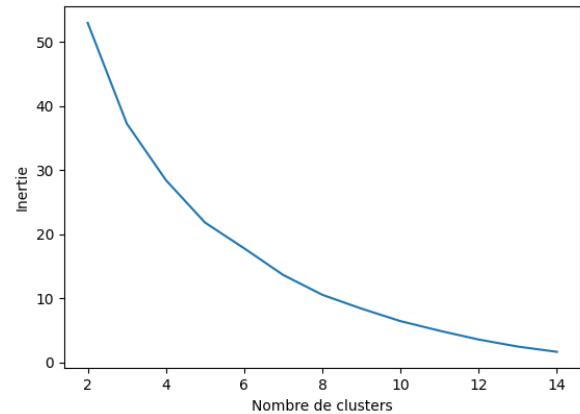


FIGURE 13 – Inertie en fonction du nombre de clusters k pour la matrice d'exemple.

les répartissant selon un nombre k de clusters choisi par l'utilisateur.

Une méthode courante pour déterminer un nombre de clusters pertinent (PETH et al. 2016) consiste à étudier la variation de l'inertie en fonction du nombre de clusters.

Augmenter le nombre de clusters fait nécessairement décroître l'inertie - le cas limite étant une inertie nulle atteinte lorsque chaque individu est le centroïde de son propre cluster - mais il est courant de déterminer le nombre de clusters idéal en cherchant un “coude” dans la représentation graphique de l'inertie en fonction de k . La figure 13 montre cette représentation graphique pour la matrice d'exemple présentée en sous-section 2.4. On observe deux coudes assez discrets en $k = 3$ et $k = 5$: compte tenu du nombre restreint d'individus, il semble donc pertinent de choisir $k = 5$.

3 Analyse et résultats

3.1 HST NGC1512 High-res : Analyse visuelle

Le calcul des fonctionnelles de Minkowski a été testé en premier lieu sur les images du jeu de données “HST NGC1512 High-res” (sous-sous-section 2.1.2). Dans la suite, les fonctions de Minkowski sont calculées pour des seuils allant de 0 à 255 et sont normalisées par le maximum de leur valeur absolue. La normalisation permet de s'affranchir de la dépendance des MF à la taille des images et à leur luminosité moyenne (qui peuvent varier d'une image à l'autre). De cette manière, les MF sont ramenées à un même ordre de grandeur, ce qui évite un biais évident dans la classification : la séparation en k clusters de galaxies distinguées seulement par l'ordre de grandeur de leurs MF.

La figure 14 montre un exemple de calcul des MF sur une galaxie du jeu de données. L'allure globale des courbes est assez typique. L'axe des seuils est tronqué à 70 car les trois courbes convergent vers 0 à partir de ce seuil. En effet, pour un seuil grandissant, l'image se rapproche généralement d'un signal nul pour lequel les trois MF sont nulles. Les images saturées, c'est-à-dire contenant des pixels de valeur maximale (255), sont les seules exceptions puisque les pixels de valeur maximale appartiennent à Σ_ν , quelle que soit la valeur du seuil ν .

On peut présumer que l'information importante est contenue dans les extrema locaux de ces courbes, dont la position et l'épaisseur dépendent de la morphologie de la galaxie. La PCA en apportera une confirmation.

3.2 Images de simulation : Impact des artefacts

Plusieurs images des jeux de données "COSMOS : HST-ACS Mosaic" sont fortement bruitées, de faible résolution, et/ou leurs galaxies sont inclinées en dehors du plan de projection (voir figure 2). Une étude de l'impact de ces altérations sur l'allure des fonctions de Minkowski a donc été réalisée afin de pouvoir apprécier l'importance de la qualité du jeu de données utilisé. L'étude est réalisée sur les images de simulation présentées en sous-sous-section 2.1.1 et les altérations utilisées sont celles présentées en sous-sous-section 2.3.1.

Bruit Pour analyser l'impact de chaque type de bruit sur les MF, nous calculons les MF de chacune des huit images de simulation avant et après application du bruit. Six des huit images sont représentées sur la figure 31 en annexe C.1. Le protocole est répété pour plusieurs valeurs des paramètres σ , λ et p . Les figures 15 et 16 montrent les fonctions de Minkowski de chaque image de simulation avant et après l'application des bruits gaussien et "poivre et sel" pour des valeurs σ et p qui donnent des images visuellement comparables à l'image la plus bruitée de la figure 2, voir la figure 32 en annexe pour une représentation de l'impact de ces bruits pour ces paramètres fixés. Les bruits de Poisson homogène et inhomogène ne sont pas discutés ici mais donnent des résultats très similaires, voir les figure 34 et 35 en annexe.

Aucun des bruits étudiés ne semble avoir de gros impact sur l'aire F mis à part le bruit gaussien qui décale la courbe. L'allure de la courbe est cependant plus ou moins conservée, y compris pour des intensités de bruit élevées.

Les fonctions du périmètre U et de la caractéristique d'Euler χ semblent quant à elles fortement impactées par tous les types de bruit. Pour rappel, les huit images de simulation du jeu de données sont très semblables, mais diffèrent par des détails subtiles. La dispersion des courbes bleues de

χ sur les figures 15 et 16 indique qu'initialement, les MF parviennent à distinguer le détail des images. Mais comme les bandes rouges le montrent, l'ajout du bruit engendre une uniformisation des fonctions de Minkowski de chaque image et rend impossible la distinction de celles-ci. Cette perte d'information a lieu même pour de faibles valeurs de l'écart-type σ et de la probabilité p .

L'effet du bruit est donc très fort, car en plus de modifier considérablement l'allure des fonctions de Minkowski de l'image, il rend des galaxies de morphologies différentes indistinguables aux yeux des MF. Ce dernier point pose problème pour classifier des galaxies contenues dans des images sujettes au bruit.

Dégénération de la résolution Un exemple de l'impact de la dégradation de la résolution est présenté avec la figure 17. L'image initiale de simulation de galaxie est de taille 200x200 pixels². Trois dégradations lui sont appliquées, une de facteur 2, une autre de facteur 8 et une dernière de facteur 20. Ainsi, la taille de l'image devient respectivement 100x100 pixels², 25x25 pixels² et 10x10 pixels².

Pour la fonction de l'aire F , on remarque que même lorsque le facteur de dégradation est élevé, la courbe reste très fidèle à celle de l'image originale. Bien sûr, ce résultat change si le facteur est excessivement grand : par exemple, si on prend un facteur égal à 200, l'image ne fait plus qu'un pixel et la valeur normalisée de l'aire passe subitement de 1 à 0 lorsque la valeur du seuil est égale à la valeur du pixel.

Même si l'on remarque de légères variations pour un facteur de dégradation relativement élevé, le périmètre U demeure tout de même fidèle à celui de l'image de base.

Contrairement aux deux autres fonctions, on remarque une nette différence entre la valeur de la caractéristique d'Euler χ , pour une image dégradée et celle de l'image initiale et ceci quel que soit le facteur de dégradation. La dégradation de la résolution de l'image a donc un impact important sur la caractéristique d'Euler. Cette analyse peut donner l'idée que, pour pouvoir effectuer une classification des galaxies, il est préférable d'avoir des résolutions d'images semblables.

Inclinaison La figure 18 représente les fonctions de Minkowski d'une galaxie de simulation à laquelle on applique une rotation d'angle variable entre 0 et $\pi/2$ par rapport au plan de projection. Les images associées sont en annexe C.1, sur la figure 33. Il est à noter que cette simulation d'inclinaison est limitée, car l'étalement de la galaxie selon la troisième dimension (portée par l'axe normal au plan de l'image) est complètement négligée. Ainsi, une simulation de rotation d'angle $\pi/2$ d'une galaxie contenue dans le plan de projection donne un disque complètement plat, alors qu'en réalité son épaisseur est non-négligeable.

L'impact de l'inclinaison sur l'allure de l'aire et du périmètre semble relativement faible, pour des inclinaisons rai-

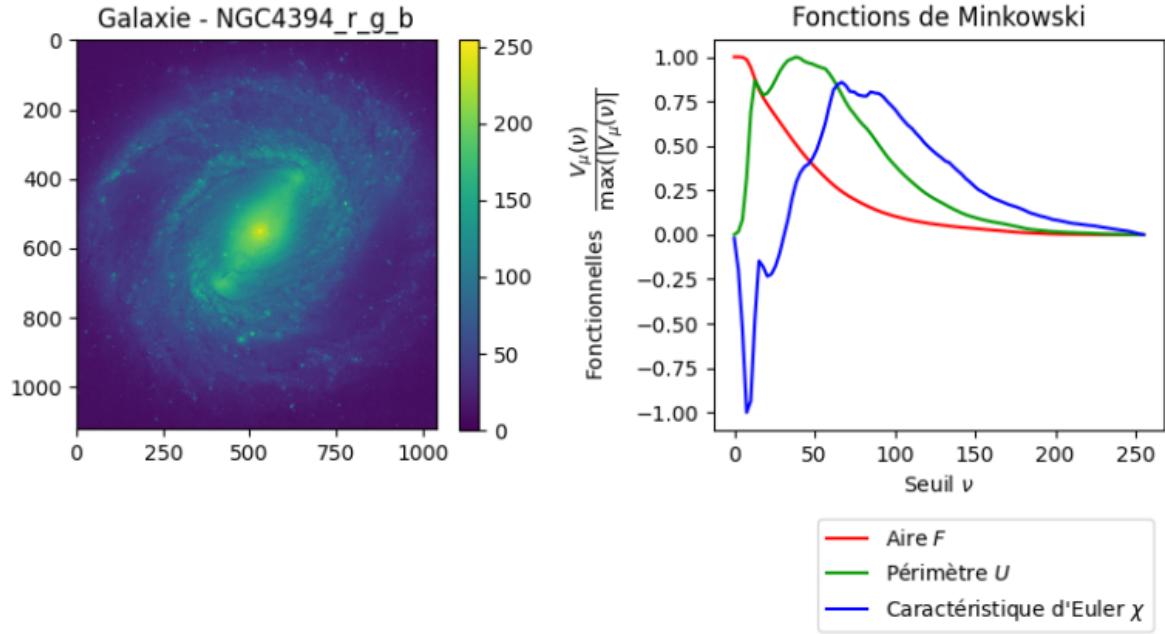


FIGURE 14 – Fonctions de Minkowski de la galaxie NGC4394.

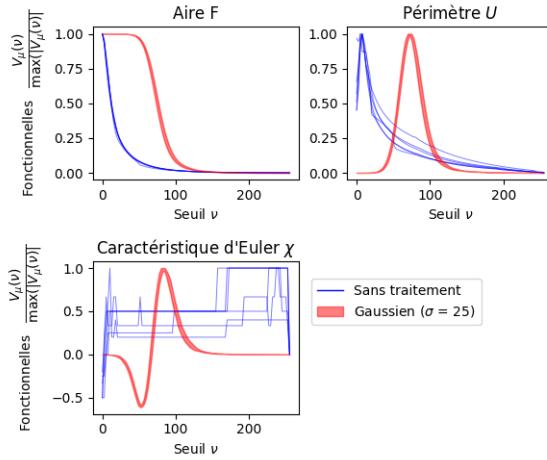


FIGURE 15 – Impact du bruit gaussien sur les MF images de galaxies simulées. Après l’application du bruit sur les images initiales (courbes bleues), les huit nouvelles courbes sont confondues à l’intérieur de la bande rouge.

sonnables. Si la valeur de ces fonctionnelles diminue largement avec l’inclinaison, la forme de leurs courbes est cependant conservée, à l’exception du cas critique $\theta = \pi/2$.

La caractéristique d’Euler semble par contre être plutôt sensible aux inclinaisons des galaxies : pour la plupart des galaxies spirales, une trop forte inclinaison finira par rendre indistinguables les différents bras de la galaxie, qui seront considérés comme un unique amas. Cette courbe est donc

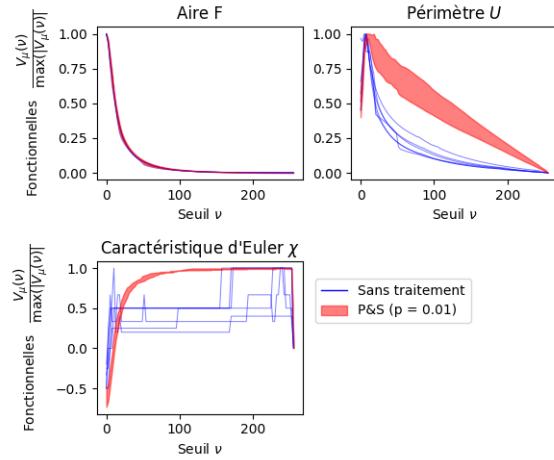


FIGURE 16 – Impact du bruit poivre et sel sur les MF images de galaxies simulées. Après l’application du bruit sur les images initiales (courbes bleues), les huit nouvelles courbes sont confondues à l’intérieur de la bande rouge.

plutôt altérée par l’inclinaison et ce même pour de faibles angles.

Bilan Ces résultats montrent que les fonctionnelles de Minkowski sont relativement sensibles aux altérations photométriques standard et en particulier au bruit. Ils montrent l’importance d’un choix d’images de qualité semblable afin

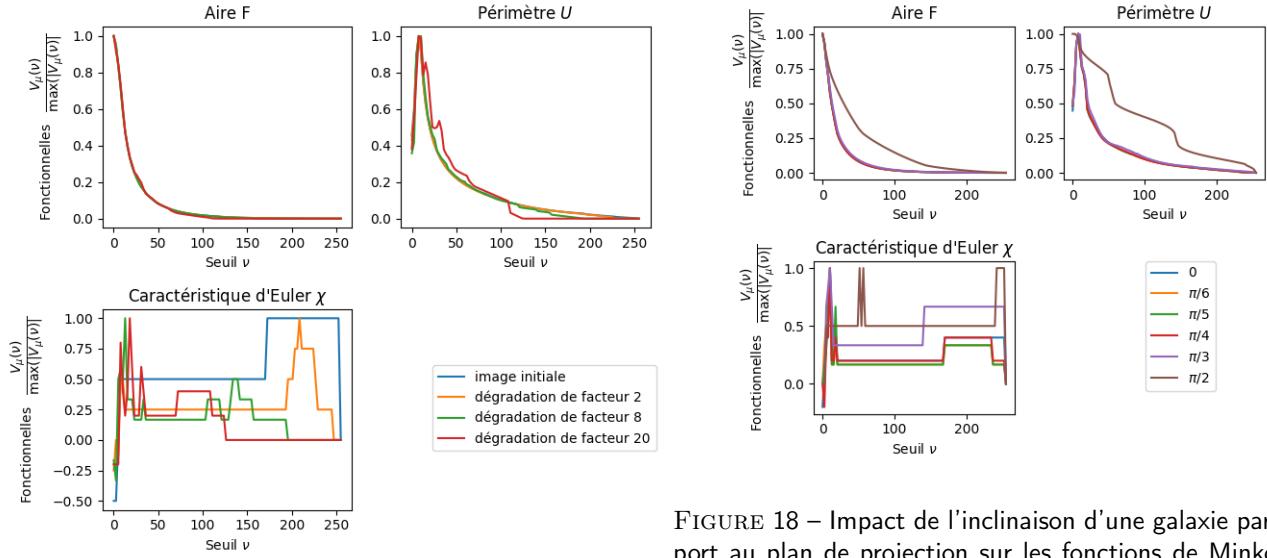


FIGURE 17 – Impact de la dégradation d'une image de simulation de galaxie sur les fonctionnelles de Minkowski.

de pouvoir comparer les courbes de leurs MF. Deux images de qualité différente d'une même galaxie peuvent en effet donner des valeurs de MF très différentes, ce qui complique le regroupement de galaxies par l'allure de leurs MF. L'analyse se doit donc d'être complétée par un traitement du bruit rigoureux et est d'autant plus compliquée que la distance et la magnitude des galaxies choisies est dispersée.

3.3 COSMOS : HST-ACS Mosaic 2 : Analyse statistique

La comparaison des MF et leur classification s'appuient sur les outils présentés en sous-parties 2.4 et 2.5. Le jeu de données "COSMOS : HST-ACS Mosaic 2" étant plus peuplé et contenant des images moins bruitées que le jeu "COSMOS : HST-ACS Mosaic 1", c'est lui qui a été choisi pour tenter une classification des galaxies.

Les trois fonctions de Minkowski de chacune des 804 images ont été calculées et discréétisées chacune en 100 points régulièrement espacés de l'intervalle allant de $\nu = 0$ à $\nu = 255$. Chaque individu est donc identifié par 300 variables décrivant sa morphologie. La matrice de données ainsi créée est de taille 804×300 . Après réduction de cette matrice et diagonalisation de la matrice de corrélation associée (équation 3), on obtient l'éboulis des valeurs propres affiché sur la figure 19. Cet histogramme nous indique que les huit premières composantes principales expliquent plus de 90% de la variance totale : la suite des calculs est donc effectuée en ayant projeté les individus sur ces huit premières composantes principales pour raccourcir les durées

FIGURE 18 – Impact de l'inclinaison d'une galaxie par rapport au plan de projection sur les fonctions de Minkowski d'une image du jeu de données de simulations de galaxies.

d'exécution.

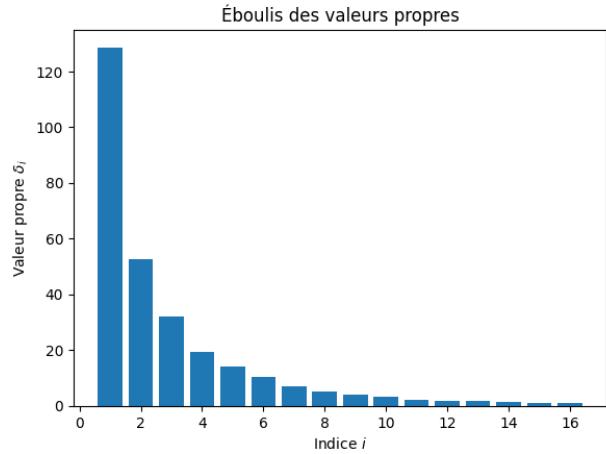


FIGURE 19 – Histogramme des valeurs propres de la matrice de corrélation associée à la matrice de données du jeu "COSMOS : HST-ACS Mosaic 2".

L'algorithme du k -means clustering est ensuite exécuté sur les 804 vecteurs de \mathbb{R}^8 représentant les individus. Pour maximiser les chances d'aboutir au minimum global d'inertie, l'algorithme est lancé 50 fois avec des conditions initiales aléatoires et la configuration d'inertie minimale est conservée. Au-delà d'un partitionnement en plus de cinq clusters, une diminution en valeur absolue du taux d'accroissement est observée (figure 20). Il a donc été décidé de fixer le nombre de clusters à cinq pour ce jeu de données. Dix ou onze clusters pourraient également être pertinents dans un second temps.

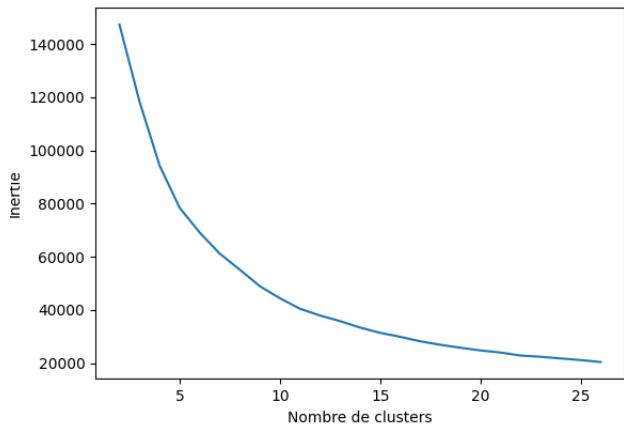


FIGURE 20 – Courbe de l'inertie d'un k -partitionnement en fonction de k pour le jeu de données "COSMOS : HST-ACS Mosaic 2"

Le partitionnement en cinq groupes est donc réalisé et représenté sur la figure 21, qui affiche la projection des individus sur les deux premières composantes principales. Il est à noter que ces deux premières composantes principales n'expliquent qu'environ 61% de la variance totale du nuage ; cette représentation est donc lacunaire puisque des dimensions supplémentaires permettraient de mieux comprendre le partitionnement (si toutefois ces dimensions supplémentaires pouvaient être représentées à l'écran de manière aussi lisible).

La classification obtenue est celle qui discrimine le mieux les images en cinq groupes vis-à-vis de leurs fonctionnelles de Minkowski calculées. Pour vérifier que les groupes obtenus ont des fonctions de Minkowski différentes, nous affichons les courbes médianes des fonctions de Minkowski des galaxies de chaque groupe sur la figure 22. Les courbes de U et de χ des groupes B et E montrent des extrema proches de ± 0.8 ou ± 0.9 tandis que les autres courbes sont d'amplitude plus faible. Cela témoigne d'une certaine dispersion sur la position des extrema dans les groupes A, C et D (si les maxima d'un groupe étaient tous atteints pour un même seuil ν_m , la courbe médiane du groupe atteindrait 1, ce qui n'est pas le cas). On peut en déduire que la position de l'extremum est une information importante pour caractériser les groupes B et E, mais elle l'est moins pour les groupes A, C et D qui sont caractérisés par d'autres critères. Par exemple, la caractéristique d'Euler médiane du groupe D pour ν proche de 255 nous informe que les images de ce groupe ont leur maximum de luminosité atteint en plusieurs régions non connectées, contrairement aux autres groupes.

Les sources d'erreur et les biais qui pourraient empêcher ce classement d'être efficace, c'est-à-dire de classifier

les galaxies *par morphologie*, sont multiples. Comme la sous-section 3.2 le montre, des images inégalement bruitées engendrent un biais sur l'allure des fonctions de Minkowski qui ne dépend pas de la morphologie de la galaxie : les caractéristiques de *l'image* passent avant les caractéristiques de *la galaxie représentée*. La faible résolution, le bruit et l'inclinaison des galaxies sont trois facteurs inhérents à l'image et non à la galaxie elle-même, qui jouent un rôle important dans le calcul des MF ; on peut donc imaginer que ces facteurs aient aussi une influence plus ou moins importante sur les clusters obtenus.

Informations physiques Toutes les galaxies du jeu de données "COSMOS : HST-ACS Mosaic 2" sont associées à des informations physiques les concernant¹⁰, comme le redshift, la masse ou l'âge (LAIGLE et al. 2016). La recherche d'un lien entre les clusters définis par la PCA sur les fonctionnelles de Minkowski et les informations physiques des galaxies de ces clusters peut avoir un intérêt certain : en cas de corrélation forte, on aurait réussi à retrouver des informations physiques d'une galaxie à partir de sa morphologie.

Cependant, aucune corrélation évidente n'a été observée entre les différentes informations physiques et les clusters définis par la méthode précédente. En prenant l'exemple de la masse (représenté en figure 23), la répartition de cette grandeur ne semble pas suivre le partitionnement défini par la méthode k -means. Une anticorrélation entre la seconde composante principale et la masse semble être possible, mais n'a pas été étudiée mathématiquement. La plupart des clusters contiennent des galaxies de masse très variable, sans claire distinction. Cette absence de corrélation visuelle est également observable pour tous les autres paramètres physiques connus pour les galaxies du jeu de données "COSMOS : HST-ACS Mosaic 2" : âge, redshift photométrique, magnitude, taux de formation stellaire...

Cela ne signifie cependant pas que la méthode utilisée ici ne permet pas d'obtenir de corrélation avec les informations physiques. Comme cela a été discuté dans cette partie d'analyse des résultats, le bruit des images a eu un impact très fort sur le classement des galaxies en clusters. Il est donc probable que ce biais ait brouillé toute information permettant d'exhiber des clusters en fonction de la réelle morphologie des galaxies et de faire apparaître une corrélation entre ces nouveaux clusters non bruités et les informations physiques des galaxies.

3.4 Discussion

Pour avoir une vision plus claire de ces clusters et vérifier visuellement l'efficacité de la classification donnée par la

10. Ces informations sont récupérées via le logiciel TOPCAT.

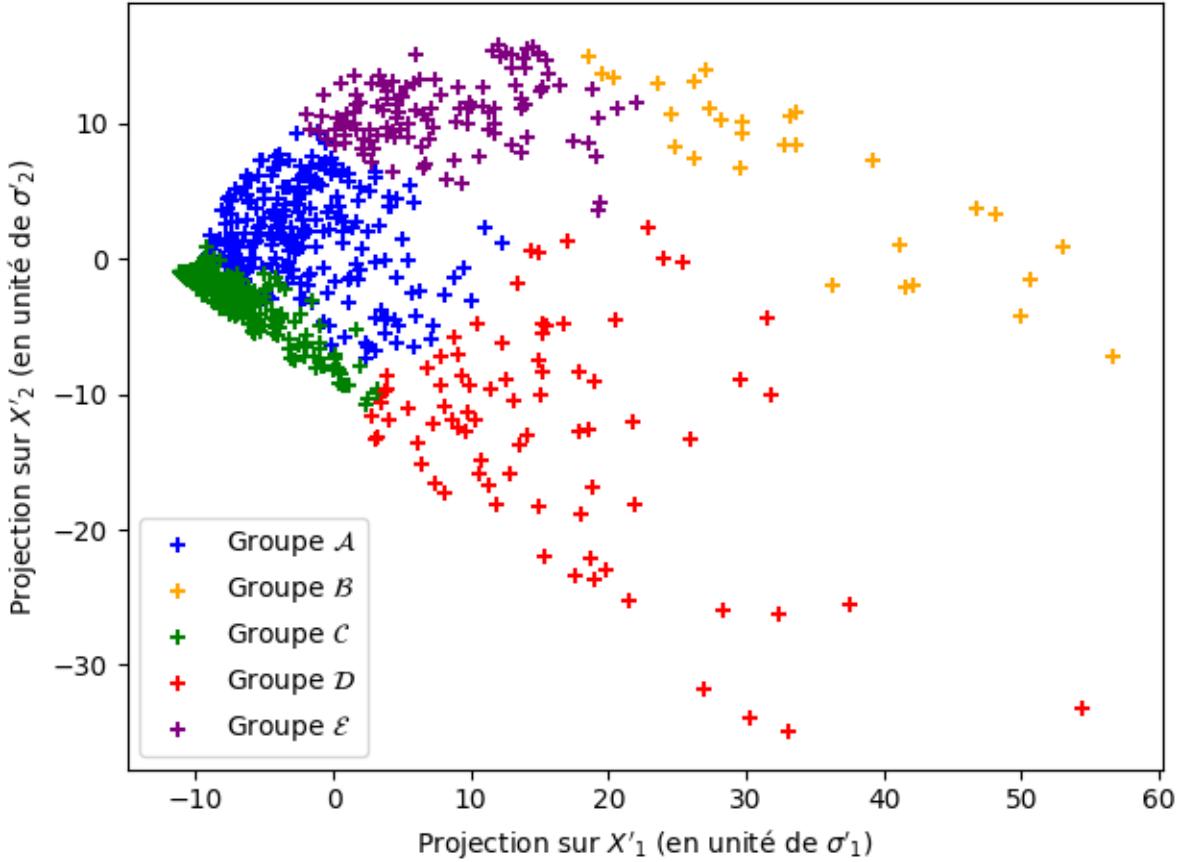


FIGURE 21 – 804 galaxies coloriées par groupe projetées sur les deux espaces propres principaux.

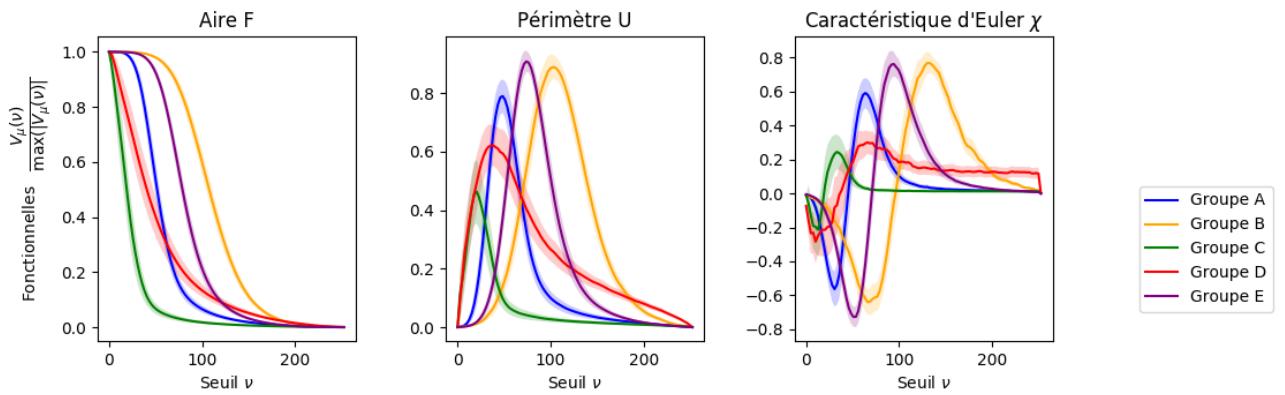


FIGURE 22 – Fonctions de Minkowski médianes des cinq groupes définis en figure 21. Les bandes d'erreur correspondent à $\pm 0.3\sigma_i(\nu)$ des fonctions F , U , χ , où $\sigma_i(\nu)$ est l'écart-type de la fonctionnelle de Minkowski au seuil ν pour le cluster i . Le coefficient 0.3 est choisi par souci de lisibilité.

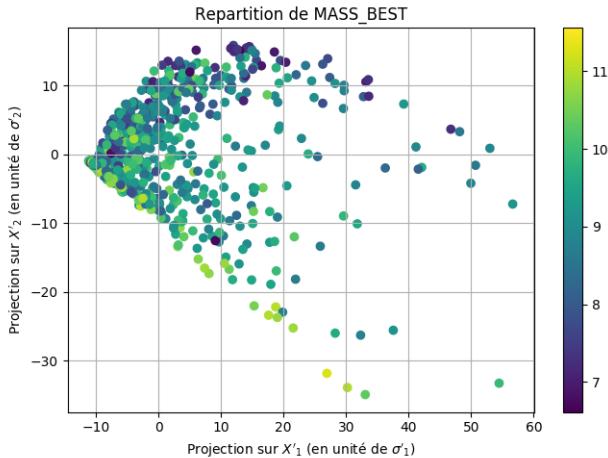


FIGURE 23 – Masse de chacune des galaxies du jeu de données “COSMOS : HST-ACS Mosaic 2”, par rapport aux deux premières composantes principales. MASS_BEST correspond au logarithme de la masse de la galaxie.

PCA et l'algorithme k -means, nous remontons aux images de départ du jeu de données et regroupons ces images selon cette classification. Les images correspondant à chaque groupe sont affichées sur les figures 25 à 29.

Comme ces figures le montrent, la classification obtenue exhibe effectivement une ressemblance entre les images de chaque groupe. Cependant, ce résultat n'est pas satisfaisant car il apparaît clairement que le bruit des images joue énormément sur les variables dominantes de la classification. Ceci est visible notamment en comparant le groupe B très bruité (figure 26) avec les autres groupes. Le groupe C (figure 27) contient quant à lui les images les moins bruitées, majoritairement sombres.

Aussi n'apparaît-il pas de distinction nette entre les morphologies des galaxies de chaque groupe. Le groupe A (figure 25) semble montrer des galaxies de morphologies très différentes, allant des galaxies elliptiques dominées par leur bulbe (ligne 5, colonne 3) aux galaxies très étalées et dominées par leur disque (ligne 4, colonne 4). Une mention du groupe E (figure 29) reste néanmoins nécessaire car ce groupe contient une majorité de galaxies de morphologie dominée par leur bulbe central. Ce groupe contient des images contaminées par du bruit d'intensité assez variable, ce qui est encourageant car cela montre que les MF ont été davantage sensibles à la concentration du bulbe qu'au bruit. Cependant, d'autres caractéristiques morphologiques plus subtiles, comme la taille du disque galactique et l'asymétrie, n'ont pas été détectées par les MF.

L'obstacle principal à la classification morphologique étant l'inégale répartition du bruit sur les images, une ten-

tative d'homogénéiser le jeu de données en utilisant les techniques de réduction du bruit présentées en sous-sous-section 2.3.2 a été réalisée. Cependant, ces techniques ne peuvent suffire à supprimer des quantités de bruit aussi importantes que celles dont sont contaminées les images du groupe B (figure 26). Supprimer les images du groupe B ne suffit pas non plus, car même sur les images modérément bruitées des groupes A, C, D et E, le filtrage se doit également d'être adapté à chaque image en fonction de son ratio Signal/Bruit. En effet, le même traitement ne pourrait être efficace à la fois sur une image de bonne qualité et sur une image bruitée. La tâche bien plus ardue que représente la suppression rigoureuse du bruit n'a donc pas pu être effectuée dans les temps, mais elle aurait permis de s'affranchir de la source d'erreur majoritaire et peut-être d'exhiber une classification discriminant les galaxies *par leur morphologie*.

Conclusion et perspectives

Le calcul et l'analyse des fonctionnelles de Minkowski ont été implémentées en python, puis couplées à des techniques d'*unsupervised learning* telles que la PCA et le k -means clustering, dans le but d'évaluer la possibilité d'en faire un outil efficace pour classer les galaxies par morphologie. Cet objectif n'a pas pu être atteint à cause du bruit présent sur certaines images du jeu de données utilisé. Nous avons ainsi montré que les fonctionnelles de Minkowski s'avéraient très sensibles aux altérations d'images et ouvert la voie vers une suite du projet avec un traitement plus minutieux du bruit ou un choix de jeux de données plus soigné.

Le travail que nous laissons aux prochains étudiants qui s'aventureront dans le vaste domaine de la classification des galaxies promet d'être aussi exigeant que le nôtre. Les perspectives pour améliorer ce travail sont nombreuses. Nous en listons quelques unes ci-dessous.

- Pour éviter d'aboutir à une classification des images *par niveau de bruit*, il pourrait être intéressant de choisir un jeu de données avec un intervalle de redshift photométrique plus restreint et des galaxies de magnitudes similaires de manière à comparer uniquement des images bruitées de manière homogène. Cela permettra une classification selon d'autres critères.
- Une autre possibilité est de caractériser chaque image par un rapport Signal/Bruit permettant de définir un seuil critique de rognage des MF. Nous avons nous-mêmes tenté d'estimer un seuil critique ν_c à partir des histogrammes de chaque image (en les lissant puis en calculant la position du second point d'inflexion), mais les contraintes de temps ne nous ont pas permis d'étudier cette pleinement cette piste.
- Une autre idée est de coupler les fonctionnelles de Minkowski avec d'autres quantités morphologiques comme C, A, S (CONSELICE 2003 ; RODRIGUEZ-

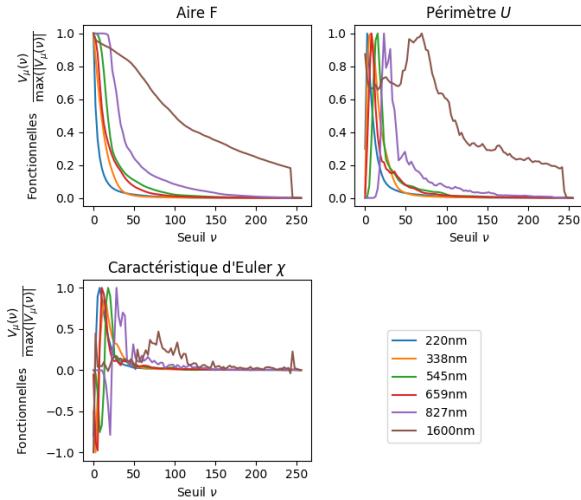


FIGURE 24 – MF de la galaxie NGC1512, prise dans plusieurs longueurs d’ondes. Les images de la galaxie NGC1512 sont affichées sur la figure 36.

GOMEZ et al. 2018) qui ont prouvé qu’elles étaient capables de caractériser la morphologie des galaxies sur des images issues du catalogue CANDELS (PETH et al. 2016).

— Enfin, pour approfondir l’analyse de la robustesse des MF dans des conditions photométriques diverses, on pourrait aussi évaluer l’impact de la longueur d’onde du filtre sur les MF calculées. Les images des galaxies du jeu de données “COSMOS : HST-ACS Mosaic 2” sont toutes prises avec le même filtre (F814W), qui est centré sur la longueur d’onde 8211.2 nm. Cependant, l’aspect visuel des galaxies peut changer en fonction de la longueur d’onde du filtre. Nous avons tenté d’explorer cette voie en calculant les MFs de la galaxie NGC1512 (voir figure 24). Une étude sur l’impact du filtre utilisé sur les groupes résultants du “*k-means clustering*” pourrait donc être pertinente.

Au cours de ses différentes phases, ce projet a été pour nous l’occasion de nous familiariser aussi bien avec les techniques et les logiciels standard en astronomie (DS9, TOPCAT, SExtractor, Astropy...), qu’avec la rédaction dans un style scientifique. Il nous a mené à synthétiser un ensemble de connaissances bâti à partir de recherches documentaires sur un sujet complexe au cœur de la recherche actuelle, pour en saisir les enjeux principaux et y apporter une touche innovante. Très ouvert, ce sujet a permis d’explorer de nombreux domaines telle que le traitement du signal, le machine learning ou encore la physique des galaxies. Il a permis à chaque membre du groupe de s’investir dans ce qui lui plaît : la programmation pour les uns, la documentation pour les autres...

Nous en tirons chacun des impressions et des résolutions

différentes quant à notre projet professionnel, mais des envies convergentes d’en apprendre plus sur les sujets que nous avons choisi de ne pas traiter en profondeur, comme la géométrie différentielle et intégrale, le *deep learning* et les modèles de formation des galaxies.

Références

- BELL, E. F. et al. (2012). “WHAT TURNS GALAXIES OFF ? THE DIFFERENT MORPHOLOGIES OF STAR-FORMING AND QUIESCENT GALAXIES SINCE $z \sim 2$ FROM CANDELS”. In : *The Astrophysical Journal* 753.2, p. 167.
- BERTIN, E. et S. ARNOUTS (1996). “SExtractor : Software for source extraction.” In : 117, p. 393-404.
- BUCHERT, T. (2010). “Morphological characterization using the Minkowski Functionals”.
- CONSELICE, C. J. (2003). “The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories”. In : *The Astrophysical Journal Supplement Series* 147.1, p. 1-28.
- DUBUSSON, S. (2010). “Bases du traitement des images : Filtrage d’images”.
- EDER, G. (2018). “The role of Minkowski functionals in the thermodynamics of two-phase systems”. In : *AIP Advances* 8.1.
- HADWIGER, H. (1957). *Vorlesungen über Inhalt, Oberfläche und Isoperimetrie*. T. 93.
- KOEKEMOER, A. M. et al. (2007). “The COSMOS Survey : Hubble Space Telescope Advanced Camera for Surveys Observations and Data Processing”. In : *The Astrophysical Journal Supplement Series* 172.1, p. 196-202.
- LAIGLE, C. et al. (2016). “THE COSMOS2015 CATALOG : EXPLORING THE $1 < z < 6$ UNIVERSE WITH HALF A MILLION GALAXIES”. In : *The Astrophysical Journal Supplement Series* 224.2, p. 24.
- LEVCHENKO, I. et al. (2016). “Morphological Characterization of Graphene Flake Networks Using Minkowski Functionals”. In : *Graphene* 05, p. 25-34.
- MANTZ, H., K. JACOBS et K. MECKE (2008). “Utilizing Minkowski functionals for image analysis : a marching square algorithm”. In : *Journal of Statistical Mechanics : Theory and Experiment* 2008.12, p. 12015.
- MASSEY, R. et al. (2010). “Pixel-based correction for Charge Transfer Inefficiency in the Hubble Space Telescope Advanced Camera for Surveys”. In : *Monthly Notices of the Royal Astronomical Society* 401, p. 371-384.
- MECKE, K. R. (1997). “Morphology of spatial patterns - porous media, spinodal decomposition and dissipative structures”. In : *Acta Physica Polonica Series B* 28, p. 1747-1782.
- MECKE, K. R., T. BUCHERT et H. WAGNER (1994). “Robust Morphological Measures for Large-scale Structure

- in the Universe". In : *Astronomy & Astrophysics* 288.3, p. 697-704.
- PARKER, J. M. et al. (2013). "Cluster Analysis of Protein Point Pattern Sets using Minkowski Functionals". In : *Biophysical Journal* 104.2.
- PENG, C. Y. et al. (2002). "Detailed Structural Decomposition of Galaxy Images". In : *The Astronomical Journal* 124.1, p. 266-293.
- PETH, M. A. et al. (2016). "Beyond spheroids and discs : classifications of CANDELS galaxy structure at $1.4 < z < 2$ via principal component analysis". In : *Monthly Notices of the Royal Astronomical Society* 458.1, p. 963-987.
- RAHMAN, N. et S. F. SHANDARIN (2003). "Measuring shapes of galaxy images — I. Ellipticity and orientation". In : *Monthly Notices of the Royal Astronomical Society* 343.3, p. 933-948.
- (2004). "Measuring shapes of galaxy images - II. Morphology of 2MASS galaxies". In : *Monthly Notices of the Royal Astronomical Society* 354.1, p. 235-251.
- RODRIGUEZ-GOMEZ, V. et al. (2018). "The optical morphologies of galaxies in the IllustrisTNG simulation : a comparison to Pan-STARRS observations". In : *Monthly Notices of the Royal Astronomical Society* 483.3, p. 4140-4159.
- ROUAUD, M. (2012). *Probabilités, statistiques et analyses multicritères*.
- SCHNEIDER, P. (2006). *Extragalactic Astronomy and Cosmology : An introduction*. Springer-Verlag.
- SIMMONS, B. D. et al. (2016). "Galaxy Zoo : quantitative visual morphological classifications for 48 000 galaxies from CANDELS". In : *Monthly Notices of the Royal Astronomical Society* 464.4, p. 4420-4447.
- WUYTS, S. et al. (2011). "GALAXY STRUCTURE AND MODE OF STAR FORMATION IN THE SFR-MASS PLANE FROM $z \sim 2.5$ TO $z \sim 0.1$ ". In : *The Astrophysical Journal* 742.2, p. 96.
- YOUNG, I. et al. (2004). "Fundamentals Of Image Processing".

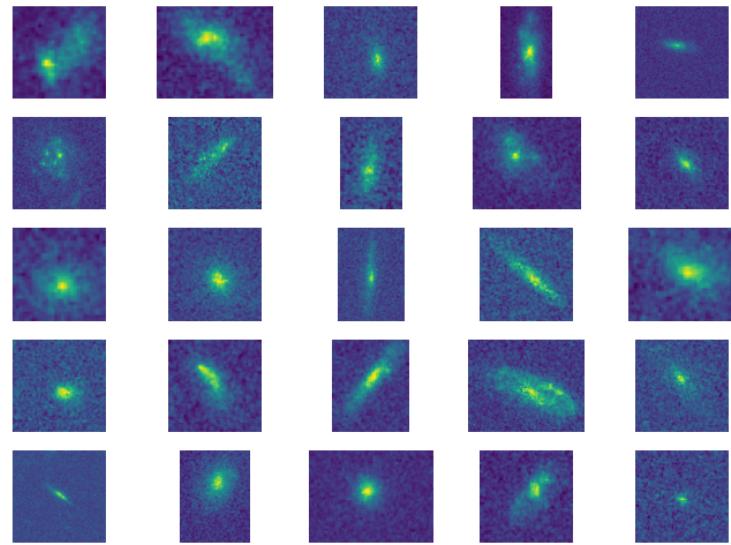


FIGURE 25 – Échantillon de 25 images parmi les images du groupe A représenté en bleu sur la figure 21.

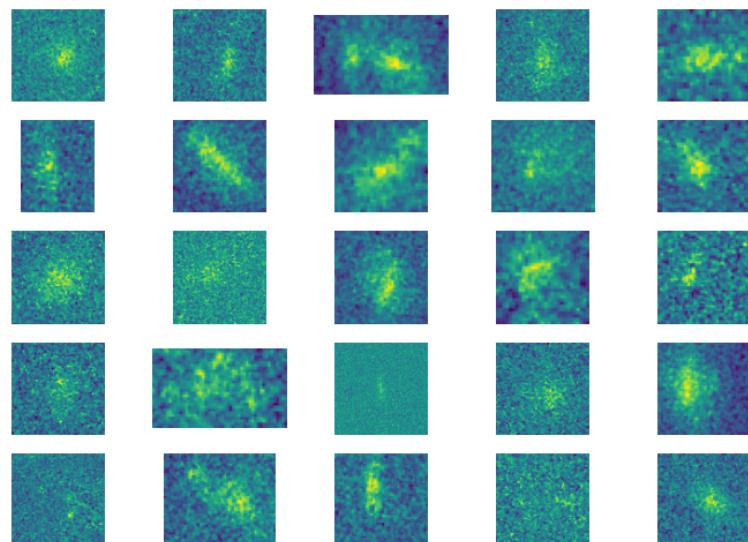


FIGURE 26 – Échantillon de 25 images parmi les images du groupe B représenté en orange sur la figure 21.

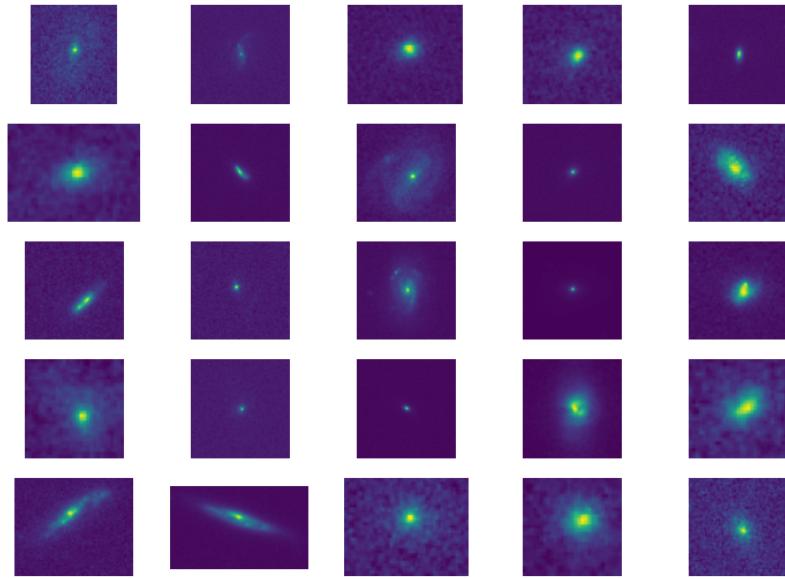


FIGURE 27 – Échantillon de 25 images parmi les images du groupe C représenté en vert sur la figure 21.

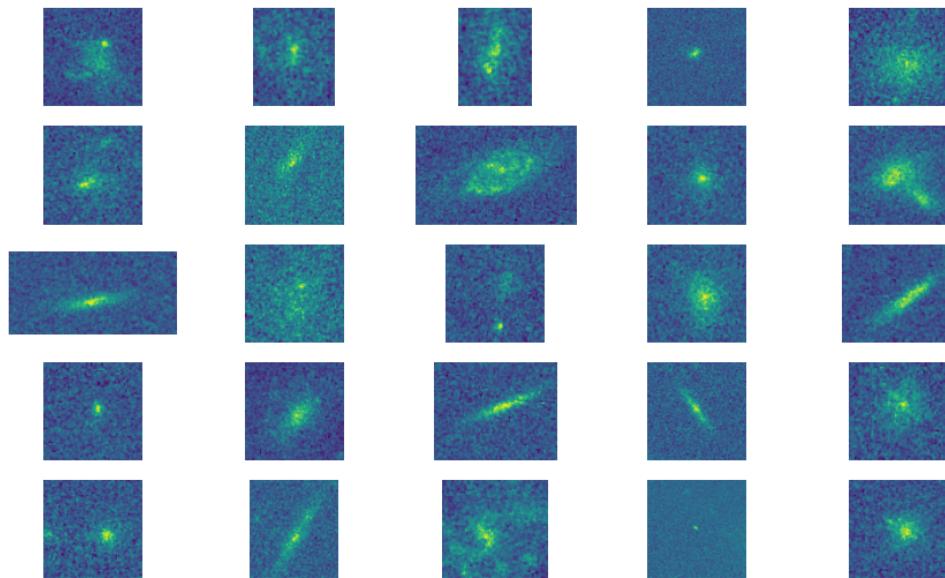


FIGURE 28 – Échantillon de 25 images parmi les images du groupe D représenté en rouge sur la figure 21.

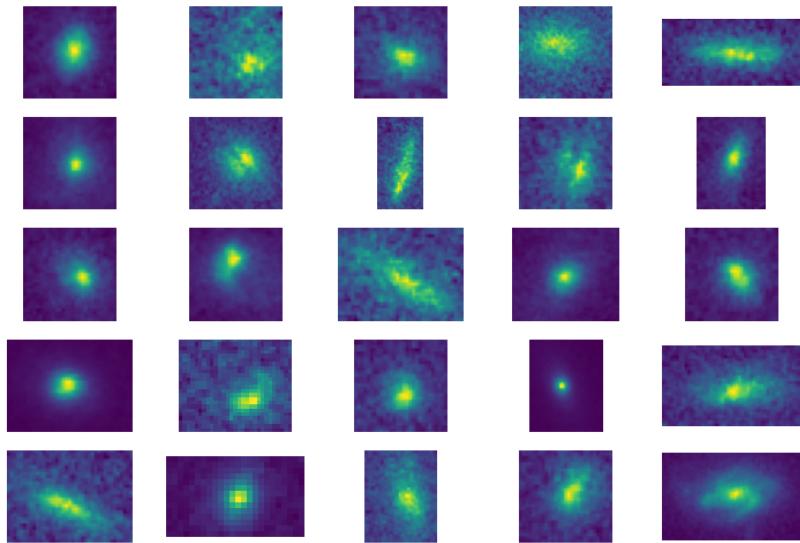


FIGURE 29 – Échantillon de 25 images parmi les images du groupe E représenté en violet sur la figure 21.

A Produit de convolution

A.1 Définition

Nous définissons dans cet annexe le produit de convolution pour des signaux discrets comme des pistes audio échantillonnées (1D) ou des images (2D).

1D Le produit de convolution d'un signal $x(n)$ avec un filtre $h(n)$ est un nouveau signal $(x \star h)(n)$ défini par :

$$(x \star h)(n) = \sum_{k=-\infty}^{+\infty} x(k)h(n-k) \quad (6)$$

On se limitera au cas où l'un des signaux, en général le filtre h , est de longueur finie d (comme il est d'usage en traitement du signal), c'est-à-dire $h(n) = 0$ pour tout $n < 0$ et pour tout $n > d$. Dans ce cas, la somme est bien définie sur tout le domaine.

2D Le produit de convolution d'un signal $f(i, j)$ (une image) avec un filtre $h(i, j)$ (aussi appelé "masque de convolution" ou "noyau") est un nouveau signal $(f \star h)(i, j)$ défini par :

$$(f \star h)(i, j) = \sum_{n=1}^N \sum_{m=1}^M f(i, j)h(n-i, m-j) \quad (7)$$

où N et M sont les dimensions de l'image. En général et dans tout ce rapport, le masque de convolution h est toujours carré de taille d impaire. On a donc :

$$(f \star h)(i, j) = \sum_{n=-\frac{d-1}{2}}^{\frac{d-1}{2}} \sum_{m=-\frac{d-1}{2}}^{\frac{d-1}{2}} f(i, j)h(n-i, m-j) \quad (8)$$

Propriétés Le produit de convolution est :

- Commutatif : $(f \star g)(n) = (g \star f)(n)$
- Distributif sur la somme : $(f \star (g + h))(n) = (f \star g)(n) + (f \star h)(n)$
- Associatif : $((f \star g) \star h)(n) = (f \star (g \star h))(n)$

A.2 Filtrage d'images

En pratique, le calcul du produit de convolution d'une image f par un noyau h au pixel $p = (i, j)$ s'interprète de la manière suivante :

1. Faire une rotation de π du noyau h par rapport à son centre,
2. Centrer le filtre sur p en le superposant à l'image,
3. Effectuer la somme pondérée entre les pixels de l'image et les coefficients du filtre qui leur sont superposés,
4. Le pixel p du produit de convolution a pour valeur cette somme pondérée.

Pour conserver la moyenne de f dans la nouvelle image $f \star h$, on normalise les coefficients du filtre pour que la somme de ces coefficients soit égale à 1. La convolution consiste donc à remplacer chaque pixel de l'image de départ par une moyenne des pixels de son voisinage pondérée par les coefficients du filtre, comme illustré sur la figure 30.

Le filtre gaussien, utilisé en sous-sous-section 2.3.2, est un exemple de noyau de convolution assez commun pour réduire le bruit. Il est construit à partir d'une courbe gaussienne d'équation $f(x, y) = Ae^{-(\frac{x^2+y^2}{2\sigma^2})}$ discrétisée en une matrice carrée de taille $8\sigma + 1$, où σ est choisi par l'utilisateur comme niveau de lissage. A est déterminé de manière à normaliser à 1 la somme des coefficients de la matrice. Ce noyau est affiché sur la figure 7.

Le cas où le masque recouvre des zones extérieures à l'image (pour les pixels du bord) peut se traiter de plusieurs manières :

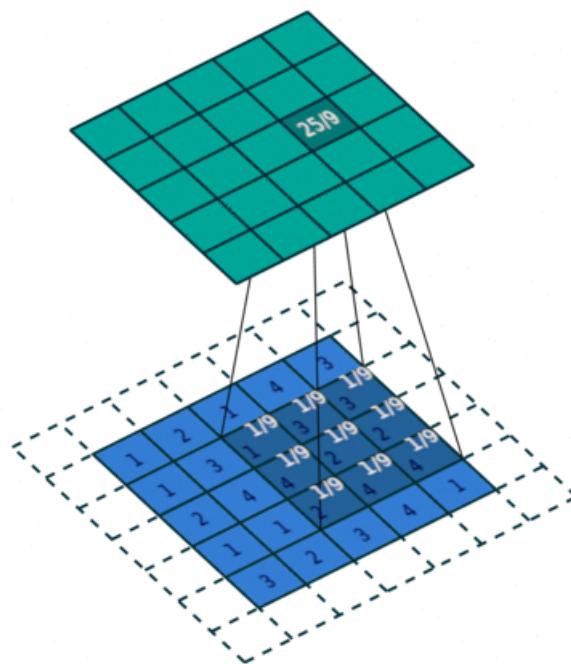


FIGURE 30 – Convolution d'une image 5×5 par un noyau uniforme 3×3 .

- **Convolution linéaire** : on considère que l'image est entourée de valeurs nulles (ce qui revient à s'en tenir à la définition, mais engendre une éventuelle perte de luminosité aux bords de l'image).
 - **Convolution circulante** : on considère que l'image est entourée d'elle-même autant de fois que nécessaire pour que le filtre n'en déborde pas.
- Nous utilisons la fonction `signal.convolve` de la librairie `scipy`¹¹, qui effectue une convolution linéaire.

B Programmes développés en python

Les programmes développés pour ce projet sont tous disponibles sur GitHub à l'adresse : <https://github.com/Anthys/Projet-Galaxie-2020>

¹¹. <https://www.scipy.org/>

C Images supplémentaires

C.1 Analyse de l'impact des altérations

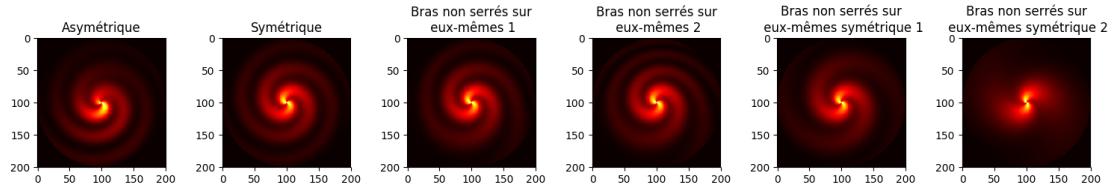


FIGURE 31 – Images de simulation utilisées pour analyser l'impact des altérations d'images sur les MF.

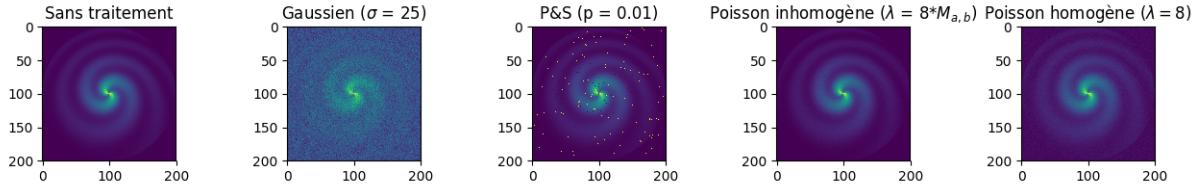


FIGURE 32 – Images représentant l'impact des bruits utilisés en figures 15, 16, 34 et 35.

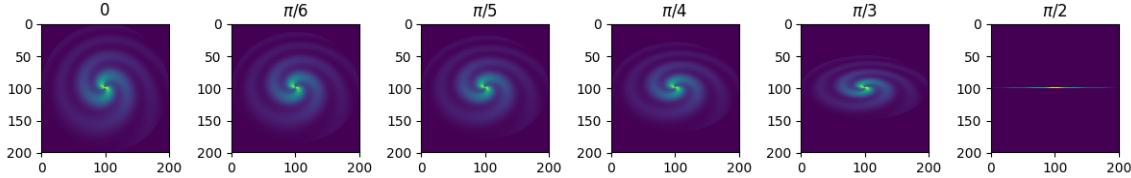


FIGURE 33 – Images de simulation utilisées pour la représentation des courbes en figure 18.

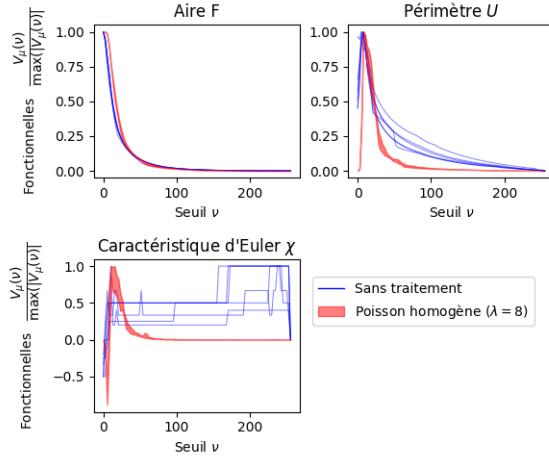


FIGURE 34 – Impact du bruit de Poisson homogène sur les MF images de galaxies simulées. Après l’application du bruit sur les images initiales (courbes bleues), les huit nouvelles courbes sont confondues à l’intérieur de la bande rouge.

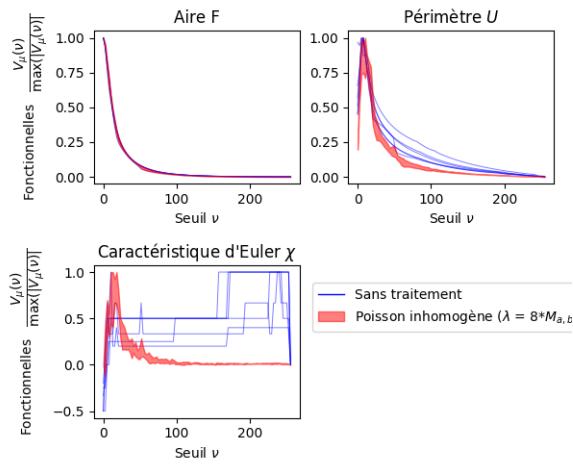


FIGURE 35 – Impact du bruit de Poisson inhomogène sur les MF images de galaxies simulées. Après l’application du bruit sur les images initiales (courbes bleues), les huit nouvelles courbes sont confondues à l’intérieur de la bande rouge.

C.2 Conclusion

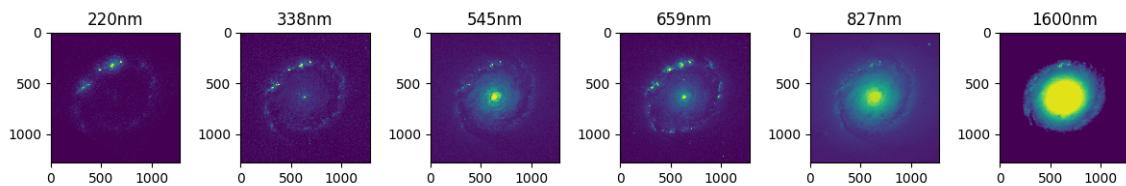


FIGURE 36 – Images du jeu de données “HST NGC1512 High-res” utilisées pour la représentation des courbes en figure 24.