

RAPPORT DU STAGE D'APPLICATION EN STATISTIQUE DE 2^E ANNEE

STRUCTURE D'ACCUEIL : Observatoire Midi-Pyrénées

THEME DU STAGE : Etude de la répartition spatiale des lentilles gravitationnelles en zones de surdensité projetée.

LIEU DE STAGE : Observatoire Midi-Pyrénées, 57 Ave d'Azereix, 65000 Tarbes - France



Promotion : 2022

Maître de stage : CABANAC Rémi

Tuteur pédagogique : PEPIN Rémi

Résumé

Ce rapport présente les objectifs, méthodes et résultats issus du stage d'application en statistique de deuxième année à l'ENSAI mené à l'Observatoire Midi-Pyrénées de Tarbes.

Ce stage traite de l'étude de la répartition spatiale des lentilles gravitationnelles. Dans ce rapport, nous détaillons :

1. l'analyse du sondage de lentilles issu des données du CFHTLS
2. une méthode d'estimation du redshift photométrique des lentilles
3. une étude de la corrélation spatiale et des regroupements en amas de surdensité projetée

Le travail mené nous a permis de conclure sur la bonne qualité des données issues du sondage spatial utilisé. Nous proposons des estimations de la valeur de redshift photométrique de bonnes qualités par deux méthodes statistiques distinctes : la régression linéaire par méthode des moindres carrés ordinaires et le modèle de mélanges Gaussiens. Enfin, notre analyse des zones de surdensité projetée et des pics de surdensité projetée nous permet de mettre en lumière une possible corrélation avec les zones spatiales présentant une forte signature de weak-lensing. De plus, l'étude de la fonction de corrélation à deux points sur l'échantillon de lentille est menée. Cette dernière est peu concluante car il s'avère difficile de bien estimer l'erreur du fait de la faible taille d'échantillons dont nous disposons.

Remerciements et citations particulières

J'aimerais exprimer mes remerciements à mon maître de stage, M. Rémi CABANAC, pour la confiance qu'il m'a témoigné au cours de ce stage, pour l'autonomie de travail dont j'ai pu disposer dans d'excellentes conditions d'accueil, ainsi que pour son aide avisée et toujours juste. Je remercie également la structure de l'Observatoire Midi-Pyrénées de m'avoir accueilli au cours des huit dernières semaines.

Au cours de ce stage, nous avons été amenés à réaliser des analyses statistiques informatiques. Le choix du langage de programmation s'est porté sur Python [VD09] plutôt que R pour ses packages permettant plus simplement l'utilisation de données astronomiques, stockées notamment au format FITS. Pour ne pas se montrer répétitif dans la suite de ce rapport, nous citons une fois pour les packages principaux qui nous ont permis de réaliser nos analyses.

1. Package matplotlib [Hun07]
2. Package numpy [Har+20]
3. Packages pandas et scipy [McK10]
4. Package astropy [Ast+18]

Les travaux réalisés au cours de ce stage s'inspirent des travaux réalisés par Masamune Oguri et son équipe [Ogu06] sur la distribution des lentilles gravitationnelles fortes en fonction des populations de halos.

Les codes informatiques réalisés au cours de ce travail sont disponibles sur

<https://github.com/clementgabas/StageOMP>

Environnement du stage¹

Ce stade d'application en statistique de deuxième année a été réalisé aux bureaux de l'Observatoire Midi-Pyrénées (OMP) situés dans le Pôle Universitaire Tarbes Pyrénées, accueillant aujourd'hui près de 6 000 étudiants avec des activités d'enseignement et de recherche, notamment dans des secteurs de haute technologie (industries céramiques, aéronautique, matériaux, etc.).

L'OMP est tout d'abord un observatoire des Sciences de L'univers. C'est également une école interne de l'Université Paul Sabatier - Toulouse III (UT3). L'observatoire rassemble les laboratoires des sciences de l'univers, de la planète et de l'environnement et constitue le noyau du pôle "Univers, Planète, Espace, Environnement" de l'UT3.

L'OMP est placé sous la tutelle du CNRS, de l'UT3, de Météo France, de l'IRD et du CNES. Les six laboratoires de l'OMP permettent de mener des recherches en physique et astrophysique, chimie, écologie-environnement, sciences de la terre, de l'océan et de l'atmosphère, pour couvrir un vaste champ de recherche allant de l'étude du Big Bang et de l'univers lointain au fonctionnement actuel des différentes enveloppes de la Terre et de leurs interactions, en passant par celles des planètes du système solaire et de la Terre interne.

L'OMP dispose également de l'Observatoire du Pic du Midi et de sa plateforme d'observation intrusmentalisée à 2877m d'altitude. La plateforme technique de l'Observatoire du Pic du Midi de Bigorre a pour mission d'apporter un soutien technique et logistique aux expériences scientifiques de l'OMP. Elle assure la bonne marche des observations et Travaux Pratiques d'enseignement supérieur implantés au Pic du Midi. Les expériences en cours couvrent tous les domaines thématiques des sciences de l'univers avec six coupole astronomiques, deux services d'observation en astronomie et douze expériences de laboratoires de l'OMP, et autres laboratoires de la communauté nationale et internationale. Les activités principales de l'observatoire

1. La plupart des explications de cette partie de présentation sont des copiés-collés presque mot pour mot des présentation officielle de l'OMP et du l'observatoire du pic du midi. Ces explications sont d'une grande clarté et j'ai donc fait le choix de les conserver comme telles.

sont les activités nocturnes en astrophysique, notamment l'étude du magnétisme des étoiles et la météorologie des planètes Jupiter et Saturne, l'étude des activités solaires et l'étude des activités atmosphériques.

Table des matières

	Page
Introduction	1
Éléments de physique et de cosmologie	1
Contexte de l'étude	2
Présentation des données CFHTLS-T0007	2
1 Analyse des biais d'observation des lentilles	4
1.1 Le seeing ne biaise pas l'échantillon	4
1.2 Le temps d'exposition ne biaise pas le sondage de lentilles	5
1.3 Ni le rayon d'Einstein, ni le redshift ne biaisaient les lentilles du sondage	6
2 Corrélations spatiales et photométriques	7
2.1 Différents quantificateurs de la corrélation	7
2.1.1 Théorie	7
2.1.2 Résultats empiriques	7
2.2 Estimation du redshift via les bandes photométriques	8
2.2.1 Estimation par régression linéaire par méthode des moindres carrés ordinaires	9
2.2.2 Estimation par un modèle de mélanges Gaussiens	9
3 Corrélation spatiale et regroupement en clusters	13
3.1 Surdensité locale de lentille	13
3.1.1 Zones de surdensité et sous-densité locales	14
3.1.2 Pics de surdensité	16
3.1.3 Corrélation au weak-lensing	18
3.2 Fonction de corrélation à 2 points	19
3.3 Densité de galaxies voisines	20
Conclusion	23
Bibliographie	26
A Compléments d'analyses statistiques	27
A.1 Seeing	27
A.2 Temps d'exposition	28
A.3 Autres biais d'observation	30
B Compléments sur les études de corrélations	32
B.1 Matrice de corrélation	32
B.2 Corrélation photométrique	32
B.3 Corrélation spatiale	35
C Bilan personnel de l'expérience et des compétences acquises	43

Liste des Figures, Tableaux et Listings

Introduction	1
1 Illustration de la LSS	1
2 Illustration d'une lentille gravitationnelle	2
3 Disposition des champs dans le ciel	3
1 Analyse des biais d'observation des lentilles	4
1.1 Histogramme de seeing	4
1.2 Histogramme de temps d'exposition	5
2 Corrélations spatiales et photométriques	7
2.1 Matrice de corrélation - r de Pearson	8
2.2 Différences entre bandes magnétiques	9
2.3 Comparaison AIC - BIC pour GMM	10
2.4 Différences entre bandes magnétiques - GMM	11
2.5 Différences entre bandes magnétiques - quartiles de redshift	11
2.6 Différences entre bandes magnétiques - redshift ± 0.5	12
3 Corrélation spatiale et regroupement en clusters	13
3.1 Densité de lentille par degré carré	13
3.2 Disposition des lentilles dans les champs	14
3.3 Surdensité locale de lentilles - champ W1 & D1	15
3.4 Surdensité locale de lentilles - champ W1 & D1 - écart type	16
3.5 Pics de surdensité locale de lentilles - champ W1	17
3.6 Signature de weak-lensing	18
3.7 Estimation de la corrélation à 2 points	20
3.8 Densité de galaxies voisines par lentille	21
3.9 Pourcentage de lentilles ayant au plus n galaxies voisines selon les champs.	21
3.10 Redshift Photométrique vs Redshift Spectroscopique	23
A Compléments d'analyses statistiques	27
A.1 QQ-plot du seeing	27
A.2 T-test de Welch - entrée python	27
A.3 T-test de Welch - sortie terminal	27
A.4 Densité de seeing pour les lentilles et les champs	28
A.5 QQ-plot de l'exposition	29
A.6 Test des rangs signés de Wilcoxon - entrée python	29
A.7 Test des rangs signés de Wilcoxon - sortie terminal	29
A.8 Histogrammes de seeing selon le rayon d'Einstein.	30
A.9 Histogrammes d'exposition selon le rayon d'Einstein.	30
A.10 Histogrammes de seeing selon le redshift.	31
A.11 Histogrammes d'exposition selon le redshift.	31
B Compléments sur les études de corrélations	32
B.1 Matrice de corrélation - ρ de Spearman	32
B.2 Régression par OLS pour estimation du redshift par les différences entre bandes magnétiques	33
B.3 Différences entre bandes magnétiques - OLS	34
B.4 Surdensité locale de lentilles - champ W2	36

B.5	Surdensité locale de lentilles - champ W2 - écart type	37
B.6	Surdensité locale de lentilles - champ W3 & D3	38
B.7	Surdensité locale de lentilles - champ W3 & D3 - écart type	39
B.8	Pics de surdensité locale de lentilles - champ W2	40
B.9	Pics de surdensité locale de lentilles - champ W3	41
B.10	Algorithme de type Monte-Carlo pour l'estimation de l'erreur sur ζ_{LS}	42
B.11	Nombre de galaxies voisines par lentille	42
B.12	Densité de galaxies par degré carré	42

Introduction

Éléments de physique et de cosmologie nécessaires à la compréhension des enjeux du sujet

Comment sommes-nous passés d'un univers pratiquement homogène à un âge de 380 000 ans, à un univers aujourd'hui structuré, via la seule force de gravitation, en seulement 15 milliard d'années ? Pour bien comprendre les enjeux de cette question, intéressons nous d'abord à la structure de l'univers.

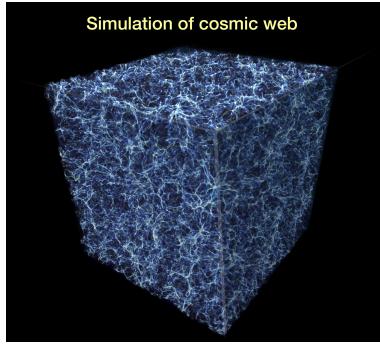


Fig. 1 – Illustration de la *Large Scale Structure de l'univers*. Crédits : NASA, ESA, and E. Hallman (University of Colorado, Boulder) ; [Shu]

La gravité est une propriété de l'espace-temps liée à sa courbure. Elle est à l'origine du regroupement des particules de gaz en étoiles, du regroupement des étoiles en galaxies, et du regroupement des galaxies en amas et en structures de tailles bien plus larges. Ces groupements de galaxies forment de longs filaments, parsemés de vide, ressemblant à une toile d'araignée. C'est ainsi qu'on se réfère couramment à la *Toile Cosmique* (Figure 1). En jargon astrophysique, nous parlons de **Large Scale Structure** (LSS) de l'univers, qui fait donc référence à cet ensemble de structures de galaxies, sur des échelles plus grandes que le simple amas de galaxies. Pour cartographier la LSS de l'univers observable, nous avons besoin de la position des galaxies sur la sphère céleste et de leur distance. Lorsque l'on observe la LSS de l'univers, nous observons les étoiles par leurs émissions lumineuses. De ces observations lumineuses, nous déduisons la distance des objets via leur rougissement. En effet, du fait de l'expansion de l'univers et de la dilatation de l'espace-temps induite, un décalage vers les grandes longueurs d'ondes (et donc vers le rouge pour la lumière visible) du spectre des objets lointains est observé. On parle

alors de **décalage vers le rouge** (ou **redshift**). Pour une galaxie lointaine, la mesure du redshift de l'objet permet d'avoir une idée de sa distance. Pour les galaxies les plus proches, leur mouvement propre, provoquant lui-même un effet Doppler bleuissant ou rougissant la lumière, est non négligeable devant leur rougissement induit par l'expansion de l'univers et il faut donc utiliser une autre méthode de calcul des distances.

Ainsi, pour observer les objets les plus anciens, nous regardons ceux qui ont le plus grand redshift. En particulier, lorsque l'on observe les objets ayant un redshift de 1000, nous observons en réalité 380 000 ans après le big bang. La lumière émise par ses objets les plus anciens provient d'une époque où l'espace est à une température de 3000K et est assez froid pour permettre aux premiers atomes de se former par la combinaison de protons, neutrons et électrons déjà pré-existants dans cet univers primitif. C'est le **rayonnement fossile**, ou **fond diffus cosmologique**, qui est donc composé des photons émis au moment où les premiers électrons sont capturés par les premiers protons.

Ce rayonnement correspond à la plus ancienne lumière observable depuis la Terre. Son observation nous renseigne sur l'état d'homogénéité et d'isotropie spatiale de l'univers primordial. Nous apprenons ainsi que l'univers primordial est très homogène mais présente de légères variations qui, au fil de son expansion, ont résultées en la création de la Toile Cosmique.

Aujourd'hui, notre connaissance du rayonnement cosmologique est très bonne. Cependant, les différents modèles physiques actuels tentant d'expliquer l'évolution de l'univers montrent que la gravitation n'a pas le temps, en 15 milliards d'années, de produire les galaxies, amas et autres structures que l'on peut observer autour de nous dans la LSS. La solution la plus communément envisagée est alors l'introduction de **matière noire**, une matière que l'on ne voit pas, mais qui permet d'introduire de la masse dans l'univers pour suffisamment augmenter l'intensité de la gravitation afin que les modèles théoriques rendent compte de la réalité observable aujourd'hui.

Contexte de l'étude

Dans la vie de tous les jours, lorsque deux objets sont alignés avec un observateur, ce dernier ne voit logiquement que l'objet le plus proche de lui, le second étant caché par le premier car aligné avec celui ci. Dans l'espace, le même principe s'applique. Lors de très rares occasions, il arrive que des objets, ici des galaxies ou des amas de galaxies, soient exactement alignés dans notre ligne de visée. Ils sont donc invisibles car alignés avec un autre objet qui cache alors leur lumière.

Cependant, nous travaillons ici avec des objets très massifs et denses. Comme leur masse est à l'origine d'une courbure locale de l'espace-temps, des rayons de lumière émis par l'objet caché et n'allant pas dans la direction de l'observateur (sur la Terre) sont déviés par cette courbure de l'espace temps, et nous parvennent. On observe alors, autour de l'objet le plus proche, un halo lumineux, preuve de la présence en arrière plan d'un objet caché. On appelle **lentille gravitationnelle** le couple objet de premier-plan et objet caché d'arrière-plan dont l'existence nous est confirmée par le halo lumineux ([Ein36], Figure 2).

La connaissance de la répartition des lentilles gravitationnelles est un enjeux clé de la recherche en astrophysique pour mieux comprendre les régions les plus denses de l'univers et notamment la répartition de la matière noire. La distribution spatiale des lentilles gravitationnelles reflète la répartition spatiale de toute la matière de l'univers, qu'elle soit visible ou noire. La répartition des lentilles gravitationnelles est utilisée pour créer des cartes de matière noire dans les amas de galaxies.

L'objectif de ce stage est l'étude de la répartition spatiale des lentilles gravitationnelles, et plus particulièrement leur regroupement en zones de surdensité projetée de lentilles.

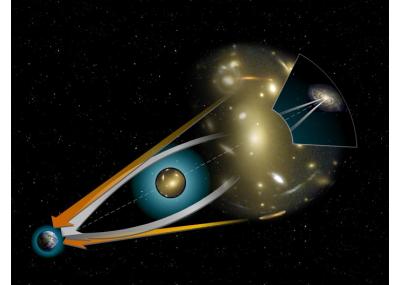


Fig. 2 – Illustration d'une lentille gravitationnelle. La lumière d'un objet distant est courbée autour d'un objet dense. Les flèches oranges montrent la position apparente de l'objet source. Les flèches blanches montrent le chemin suivi par la lumière et la vraie position de la source. Crédits : NASA, ESA, télescope Hubble

Présentation des données CFHTLS-T0007

Pour réaliser ce travail, nous disposons des données issues de la 7^{ème} et dernière version du sondage spatial² *Canada-France-Hawaï Telescope Legacy Survey*. Plus précisément, nous exploiterons les données du CFTHLS-T0007 Wide, constitué de 171 champs profonds photographiés avec l'instrument MegaCam, chaque MegaCam fournissant des mesures sur $1 \times 1 \text{ deg}^2$ ³. Cet ensemble de MegaCam produit une cartographie des objets célestes d'une taille effective (excluant les zones masquées par les étoiles brillantes d'avant plan), de 140 deg^2 répartis sur quatre zones du ciel distinctes (cf. W1-4 sur Figure 3).

2. On appelle *Sondage Spatail* un ensemble de données sur des objets spatiaux. Dans notre cas, il s'agit d'un ensemble de données de positions spatiales, de données photométriques et de diverses autres données sur les galaxies et lentilles gravitationnelles de quatre zones du ciel définies en Figure 3.

3. Chaque MegaCam observe sur un champs $1 \text{ deg} \times 1 \text{ deg}$ avec $0.186''/\text{pixel}$ (19354×19354 pixels)

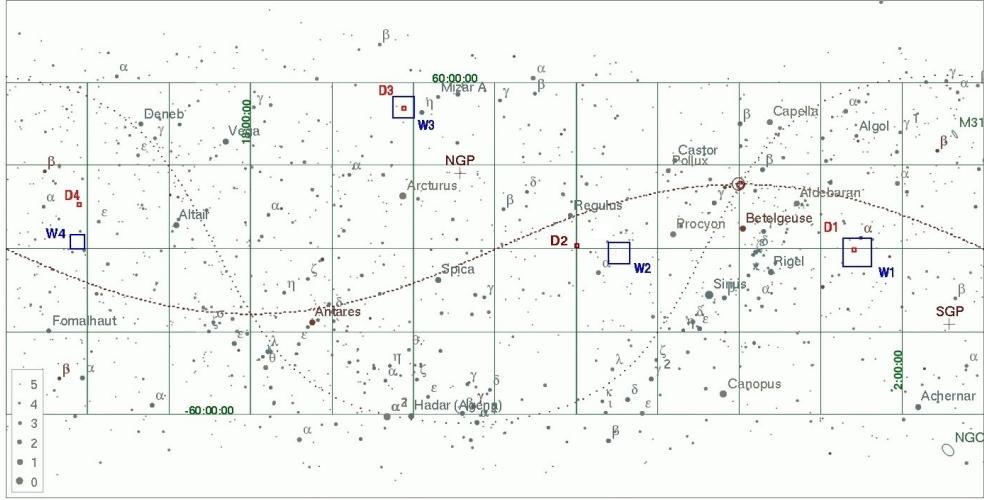


Fig. 3 – Disposition des 4 champs larges W1-4 dans le ciel. La localisation des champs a été choisie en dehors du plan de notre galaxie (ligne pointillée grise) et du plan du système solaire (ligne pointillée rouge), pour que l'on puisse observer un champ à n'importe quelle période de l'année.

Source : [Hud+12]

Les objets observés sont des galaxies occupant un cône de lumière projeté sur le ciel. Au premier ordre, ces galaxies sont divisibles en deux populations : une population d'avant plan et une population d'arrière-plan. Nous considérons les objets d'avant-plan comme ceux ayant un redshift inférieur à 0.5, et ceux d'arrière plan ayant un redshift supérieur à 0.8.

Nous disposons de 127 lentilles gravitationnelles [Mor+12] réparties sur les quatre champs W parmi les millions d'autres objets présents dans ces champs. Nous considérons que notre échantillon de lentilles est complet et pur : nous avons observé l'ensemble des lentilles gravitationnelles présentes dans le champ.

Pour chaque lentille, nous disposons des données de coordonnées spatiales selon le système de coordonnées galactiques international défini par l'Union Astronomique Internationale⁴. Nous nous servirons notamment des valeurs de position selon l'ascension droite et la déclinaison, qui sont identiques aux valeur de longitude et latitude terrestre sur la sphère céleste. Nous disposons également de données photométriques sur les bandes g (green), r (red) et i (infrared), ainsi que d'une mesure du redshift photométrique de chaque objet. Enfin, nous disposons des valeurs de rayon d'arc, ce qui nous permet d'estimer la valeur du rayon d'Einstein des lentilles, ce dernier étant un angle caractéristique pour les lentilles gravitationnelles, proportionnel à la masse totale, lumineuse et noire, de la lentille. La formule liant la masse M d'une lentille ponctuelle et son angle d'Einstein θ_E en arcsec est donné par

$$\theta_E = \left(\frac{M}{10^{11.09} M_\odot} \right)^{1/2} (d_L d_S / d_{LS})^{1/2} \quad (1)$$

avec d_L , d_S et d_{LS} les distances angulaires comobiles en Gigaparsec⁵ [Ein36].

Pour chaque degré carré issu de l'observation d'un champ, nous disposons des coordonnées spatiales du centre de la caméra, ainsi que des valeurs de seeing et de temps d'exposition. Le seeing est une valeur qui reflète la qualité optique du ciel au moment de la prise de données. On mesure ainsi la turbulence atmosphérique et la qualité du ciel. C'est une quantité variable et non prédictible qui peut dégrader la sensibilité de l'image et biaiser la détection de lentilles.

4. Bla+60.

5. Le parsec est une unité de longueur. Nous avons l'approximation suivante : 1 parsec (pc) \approx 3.26 années-lumières.

Chapitre 1

Analyse des biais d'observation des lentilles

Nous commençons cette étude par l'étude de potentiels biais d'observation sur le sondage de lentilles gravitationnelles. En effet, il est possible que l'observation de lentilles soit favorisée ou défavorisée par certaines conditions d'observation. Ainsi, nous commençons par vérifier si les conditions d'observations des différents champs du ciel W1-4 jouent un rôle sur la distribution observée de lentilles. Pour ce faire, nous étudions la répartition des lentilles en fonction de leurs valeurs de seeing¹ et de temps d'exposition. Nous comparons alors ces valeurs à la répartition du seeing et de l'exposition des différents champs observés.

1.1 Le seeing ne biaise pas l'échantillon

Ici, bien que les deux histogrammes semblent se chevaucher de façon plus ou moins homogène, on remarque une sur-densité aux alentours d'un seeing de 0.5, ainsi qu'une autre aux alentours de 0.7, ainsi qu'une sous-densité pour un seeing de 0.6. Notons également qu'il y a une coupure nette à 0.466 de seeing et une autre à 0.84.

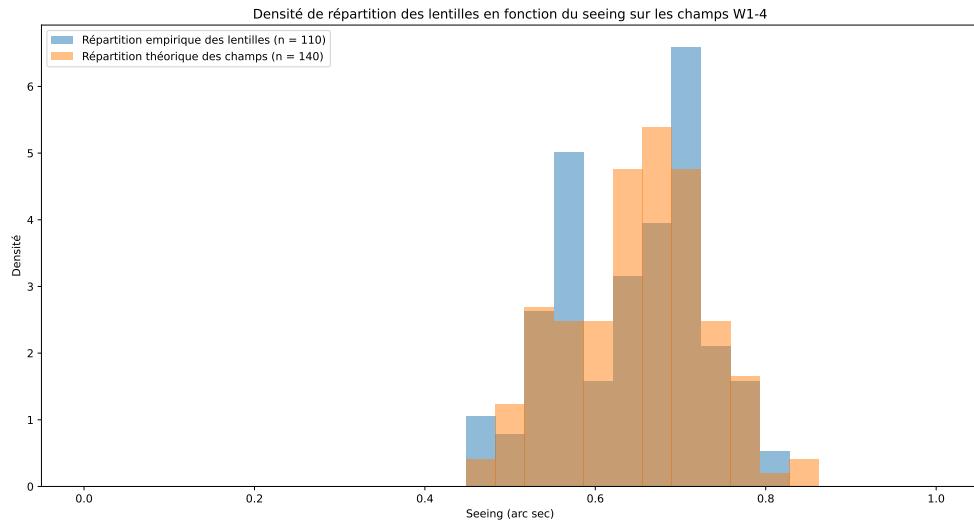


Fig. 1.1 – Histogramme des densités de répartitions en fonction du seeing pour les lentilles et les champs observés. Les deux échantillons semblent suivre la même loi de répartition. La loi des deux échantillons semble s'approcher d'une loi Gaussienne.

Source : [Hud+12]

L'objectif est de voir si la répartition des lentilles en fonction du seeing suit la même loi que la répartition théorique du seeing des mesures. Pour cela, on va utiliser un test-t de Welch [WEL47].

Après vérification de la normalité des échantillons (Figure A.1), on va tester l'hypothèse nulle suivante

$$H_0 : m_L = m_W \text{ vs } H_1 : m_L \neq m_W$$

1. Le seeing est la mesure intégrée dans la durée de la turbulence atmosphérique locale en secondes d'arc. Plus le seeing est élevé, moins les objets sont détectables pour un temps d'exposition fixé, car le même nombre de photons est réparti sur une surface plus grande dans les images

avec m_L et m_W les moyennes empiriques des échantillons de lentilles et de champs. Dans notre cas, il s'avère qu'on ne peut pas rejeter l'hypothèse nulle au seuil de 95%. On conclura alors que les deux échantillons suivent la même loi (Figure A.2, Figure A.3 Figure A.4). Ainsi, il ne semble pas y avoir de biais d'observation selon le seeing.

1.2 Le temps d'exposition ne biaise pas le sondage de lentilles

La détection des objets faibles sur le ciel est corrélée avec le temps d'exposition. Si les lentilles sont intrinsèquement faibles, nous allons en détecter davantage dans les champs observés longtemps que dans les champs à courte pose. Ainsi, nous estimons qu'il existe un possible biais.

Cette fois, on observe une légère sur-densité pour les expositions entre 80 et 90.

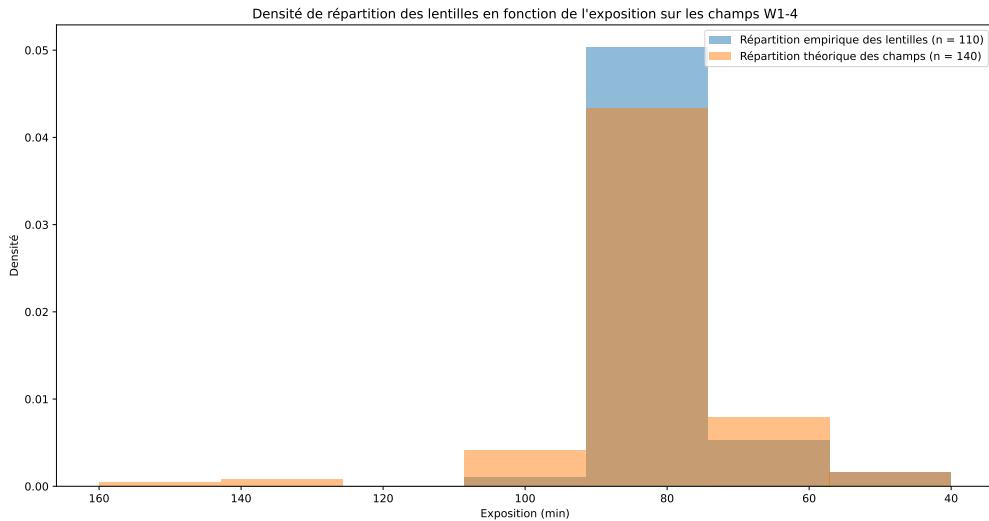


Fig. 1.2 – Histogramme des densités de répartitions en fonction de l'exposition pour les lentilles et les champs observés. Les deux échantillons semblent suivre la même loi. Cette loi ne peut évidemment pas être approchée par une loi Gaussienne car le temps de pose est fixe et est une fonction du filtre avec lequel on observe les objets.

Source : [Hud+12]

Ici, les analyses de nos échantillons (Figure A.5) montrent que ces derniers ne suivent pas une loi normale. Pour comparer les lois de répartitions, nous ne pouvons donc pas utiliser les tests paramétriques de Student ou de Welch qui presupposent de la répartition gaussienne des échantillons. Nous allons donc utiliser des tests non paramétriques, qui ne font pas d'hypothèse sur la forme des distributions.

En particulier, nous allons utiliser le test des rangs signés de Wilcoxon [Wil45] qui permet la comparaison de deux distributions de variables continues.

Plus précisément, pour 2 échantillons $X = (X_1, \dots, X_n)$ et $Y = (Y_1, \dots, Y_n)$, on calcule la différence $D = (X_1 - Y_1, \dots, X_n - Y_n)$. On note θ la médiane de l'échantillon D et on effectue le test

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta \neq 0$$

On ordonne alors les $|D_i|$ dans l'ordre croissant et on note R_i le rang de chaque D_i dans l'échantillon ordonné. On définit l'indicatrice de signe ψ_i qui vaut +1 si $D_i > 0$ et -1 si $D_i < 0$. Si $D_i = 0$, on l'exclut de l'échantillon. Après avoir effectué toutes les exclusions, notre échantillon est de taille $m \leq n$.

On a alors la statistique de test $T = \sum_i^n R_i \psi_i$. Sous l'hypothèse nulle, T ne suit pas une loi usuelle mais

une distribution spécifique d'espérance nulle et de variance $\frac{m \times (m+1) \times (2m+1)}{6}$.

Le test consiste alors en un rejet de H_0 si $T > T_{\text{critique}}$ avec T_{critique} disponible dans des tables de références². Dans notre cas, on ne peut pas rejeter l'hypothèse nulle au seuil de 95%. On conclura alors que les deux échantillons suivent la même loi (Figure A.6, Figure A.7). Ainsi, il ne semble pas y avoir de biais d'observation selon le temps d'exposition.

1.3 Ni le rayon d'Einstein, ni le redshift ne biaissent les lentilles du sondage

Le rayon d'Einstein est un angle caractéristique pour les lentilles gravitationnelles. On approxime le rayon d'Einstein (noté R_E , cf équation 1 pour une masse ponctuelle sphérique) par le rayon de l'arc de la lentille (noté R_A). Ce rayon d'arc est défini comme la distance de l'image par la lentille à la lentille (galactique) supposée.

Si l'on divise la population de lentilles en sous-échantillons selon leur rayon d'Einstein et que l'on étudie l'histogramme du temps d'exposition et des valeurs de seeing pour ces sous-échantillons (Figure A.9, Figure A.8), on remarque que le seeing et le temps d'exposition ne modifient pas (ou très peu), les valeurs de rayon d'Einstein des sous-populations. On observe le même résultat lorsque l'on étudie les lentilles selon leur redshift (Figure A.11, Figure A.10).

Finalement, nous concluons que notre sondage de lentilles est non biaisé. Nous allons donc pouvoir l'étudier sans plus de contraintes pour tenter d'obtenir des résultats sur la répartition spatiale des lentilles et leur regroupement en zones de surdensité projetée.

2. Low99.

Chapitre 2

Corrélations spatiales et photométriques

Nous rappelons l'objectif de ce stage : étudier la répartition spatiale - plus précisément la répartition de la projection angulaire 2D sur la voûte céleste - des lentilles gravitationnelles, et étudier leur distribution en amas de lentilles.

Nous venons de voir que notre échantillon de lentilles est non biaisé. Nous allons maintenant étudier la corrélation entre les différentes variables à notre disposition. En effet, cela nous permettra de mettre en lumière les liens de corrélations entre les variables et donc d'avoir des pistes à creuser lorsque nous chercherons les causes qui mènent à la distribution spatiale des lentilles.

2.1 Différents quantificateurs de la corrélation

2.1.1 Théorie

De manière générale, la méthode de quantification de la corrélation la plus communément utilisée est le coefficient **r de Pearson**. Pour deux échantillons $X = (X_1, \dots, X_n)$ et $Y = (Y_1, \dots, Y_n)$, on le définit comme suivant :

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Au travers du r de Pearson, on peut quantifier le niveau de relation linéaire entre deux variables. Le coefficient r est contenu dans le segment $[-1, +1]$. L'intensité de la corrélation est définie par la valeur absolue de r , avec une plus forte relation linéaire lorsque $|r|$ est proche de 1, et donc une corrélation nulle pour $r = 0$. Le signe du coefficient reflète le "sens" de la corrélation.

Cependant, toutes les formes de corrélations ne peuvent pas être décrites uniquement par la corrélation affine. C'est ainsi que l'on va également utiliser le coefficient **ρ de Spearman**. Ce dernier définit à quel point deux variables peuvent être liées par une fonction strictement monotone. Il se définit comme suivant :

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad \text{avec } d_i = \text{rang}(X_i) - \text{rang}(Y_i) = X_{(i)} - Y_{(i)}$$

2.1.2 Résultats empiriques

On calcule d'abord la matrice de corrélation sur nos différentes lentilles ([Figure 2.1](#)).

On observe d'abord une très légère anticorrélation entre la longitude et la latitude galactique. De manière générale, les lentilles étant regroupées en cluster sur quatre champs distincts, les corrélations dans les variables de coordonnées semblent logiques et sont même, au premier ordre, linéaires.

D'autre part, les valeurs de magnitude sur les bandes g, r et i sont très fortement corrélées entre elles. Cela s'explique par le fait que ces bandes sont contiguës et que l'on mesure en réalité le *continu thermique*¹ du spectre sur lequel s'ajoute un rougissement. On observe également une forte corrélation avec la valeur de redshift, ainsi qu'une légère corrélation négative avec la valeur de R_A . Cette anticorrélation entre magnitude et R_A peut s'expliquer par le fait que les objets les plus lointains sont, en moyenne, plus faibles, et donc de magnitude plus élevée. Une autre explication possible, également liée à l'effet distance, peut-être que le rougissement implique une distance plus élevée et donc des objets de diamètres angulaires plus faibles. Cependant, cet effet n'a jamais été observé et est sans doute non significatif ici.

Enfin, les valeurs de seeing et d'exposition révèlent une légère corrélation négative. Cette anticorrélation est

1. rayonnement de corps noir qui varie peu d'une bande photométrique à l'autre

étrange. Dans les faits, le seeing devrait être meilleur sur les bandes g, r et i et moins bon sur la bande u. Or, les temps de poses sont plus courts en g, r et i et plus longs en u et z.

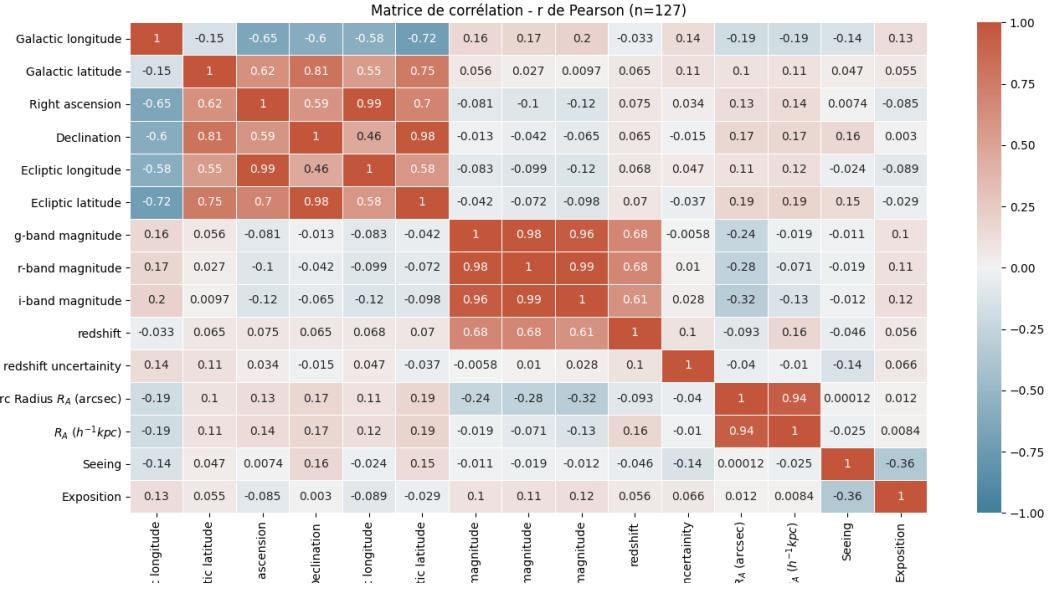


Fig. 2.1 – Matrice de corrélation des données lentilles selon le r de Pearson. L'intensité de la couleur représente l'intensité de la corrélation. La couleur représente le sens de la corrélation.

Source : [Hud+12]

On calcule ensuite la matrice de corrélation, cette fois avec le ρ de Spearman (Figure B.1).

On retrouve globalement les mêmes corrélations. On observe cependant une plus forte corrélation les bandes g, r et i et le redshift et le R_A . Cette fois, la corrélation entre seeing et exposition semble avoir complètement disparue.

2.2 Estimation du redshift via les bandes photométriques

Les valeurs de corrélation entre les bandes photométriques nous ont amené à nous demander comment ces dernières étaient liées avec le redshift et comment cela pouvait d'une part s'analyser, et d'autre part être utiliser dans la recherche de lentilles par exemples.

La figure ci-après (Figure 2.2) représente les nuages de points obtenus lorsque l'on affiche les différences entre bandes photométriques les unes en fonction des autres. On définit la différence de magnitude entre deux bandes comme ci-après :

$$\Delta mag = m_{\text{bande1}} - m_{\text{bande2}} = 2.5 \log \frac{F_{\text{bande2}}}{F_{\text{bande1}}} \quad (2.1)$$

bande1 et 2 sont des bandes photométriques (e.g. g et i), F_{bande1} , F_{bande2} sont les flux dans les bande1 et 2. La différence de magnitude est donc un rapport de flux. Dans l'exemple donné, si $\Delta mag = g - i > 0$ alors $F_i > F_g$. l'objet est davantage rouge.

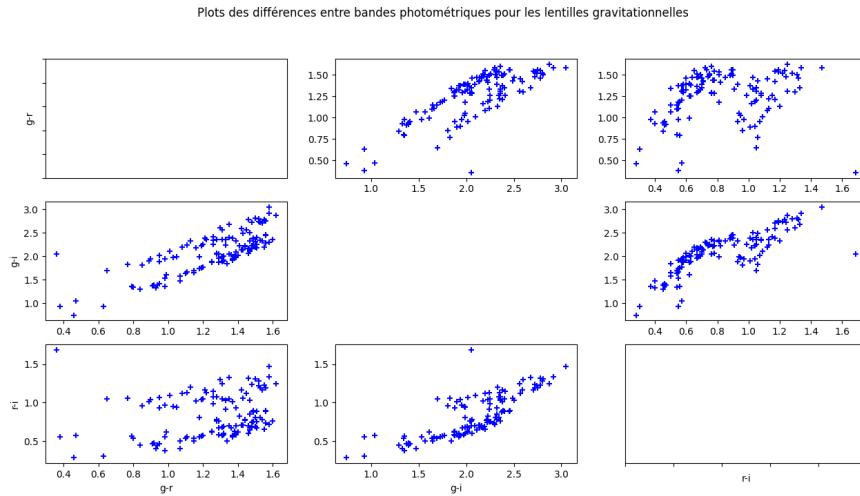


Fig. 2.2 – Nuages de points des différences entre les bandes bandes photométriques.

Source : [Hud+12]

2.2.1 Estimation par régression linéaire par méthode des moindres carrés ordinaires

Étudions maintenant les liens entre les bandes photométriques et les valeurs de redshift. Pour se faire, nous avons commencé par effectuer un modèle de régression linéaire par méthode des moindres carrés ordinaires. Notre objectif est d'expliquer une variable d'intérêt continue, ici le redshift, noté Y , par les variables explicatives continues X_{g-i} , X_{g-r} et X_{r-i} . Nous cherchons l'estimation $\hat{\beta}$ du vecteur β qui minimisera $\|Y - X\beta\|^2$ avec $Y = X\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 \text{Id})$.

Après vérification des hypothèses de base du modèle, nous avons utilisé le package **StatsModels**² et sa classe `api.OLS` pour réaliser notre modèle informatiquement.

La régression est très intéressante. La mesure de la proximité de la droite de régression aux données, fournit par le R^2 , est de 93,3%, ce qui traduit une bonne adéquation de la droite aux données (Tableau B.2). Cela fournit une méthode pour estimer le redshift selon les bandes photométriques (Figure B.3)³.

2.2.2 Estimation par un modèle de mélanges Gaussiens

Cependant, sur la Figure 2.2, une impression de découpe selon deux groupes semble sauter aux yeux, comme par exemple pour les nuages de points $g - i$ selon $g - r$ et $g - r$ selon $r - i$. Nous avons donc essayé de découper en sous-ensembles les lentilles par un modèle de mélange Gaussien (GMM). Le modèle GMM estime de façon paramétrique la distribution d'un groupe d'individus en considérant la densité comme une somme de variables Gaussiennes. Nous devons choisir le nombre de Gaussienne du modèle. Les paramètres de ces dernières sont estimées par la méthode du maximum de vraisemblance. La vraisemblance est elle-même estimée par l'algorithme d'espérance-maximisation⁴. Nous utilisons cette fois-ci le package **Scikit-Learn**⁵ et plus particulièrement la classe `GaussianMixture`. Avant de présenter les résultats obtenus, nous souhaitons mettre en avant les pour et les contre du modèle GMM qui nous ont poussés à choisir ce modèle. D'une part, le modèle GMM est très rapide. Cela pourra se montrer utile si cette méthode se voit être appliquée à de

2. SP10.

3. Note personnelle : au cours de mes travaux de recherche, je me suis rendu compte tout seul de cette relation. Cependant, il s'avère que c'est une méthode connue et qu'il existe des méthodes physiques rigoureuses pour faire cela et dériver également un modèle physique pour la galaxie observée (âge, contenu chimique, poussières, redshift, ...). [BMP11]

4. Des précisions sur le modèle GMM sont disponibles en annexe : item B.2.4

5. Ped+11.

plus larges sondages de lentilles plus tard. L'algorithme cherche uniquement à maximiser la vraisemblance, sans chercher à ramener la moyenne vers zéro par exemple. Cependant, il faut parfois repenser le nombre de sous-ensembles recherchés car si un mélange n'a pas assez de données, estimer la covariance devient difficile et l'algorithme peut diverger et fournir des vraisemblances infinies.

Pour choisir le nombre de composants du modèle GMM, nous avons eu à utiliser des méthodes pour mesurer la qualité du modèle. En particulier, nous avons utilisé le Critère d'information d'Akaike (AIC) et le Critère d'information Bayésien (BIC). Voici une définition de ces critères pour un modèle \mathcal{M} de vraisemblance maximisée \mathcal{L} avec k paramètres libres et n composants :

$$AIC(\mathcal{M}) = 2k - 2 \ln(\mathcal{L})$$

$$BIC(\mathcal{M}) = \ln(n)k - 2 \ln(\mathcal{L})$$

Nous obtenons le graphique suivant ([Figure 2.3](#)) pour les scores d'AIC et de BIC sur la population de lentilles avec le modèle GMM :

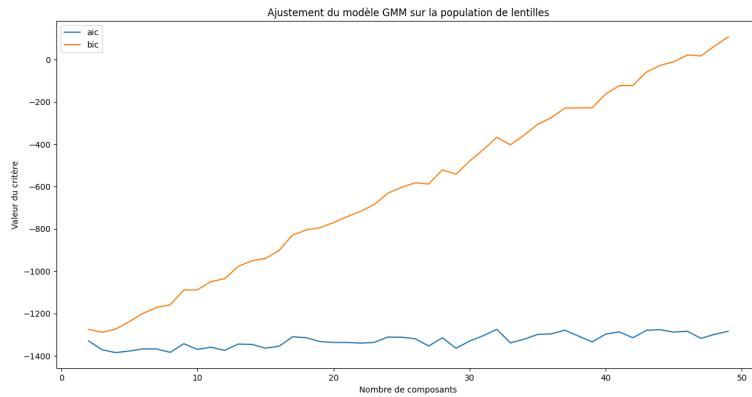


Fig. 2.3 – Valeurs des critères AIC et BIC pour différents nombres de composants dans le modèle GMM.
Package Scikit-Learn [[Ped+11](#)] ; Source : [[Hud+12](#)]

Le critère de l'AIC propose de choisir un modèle GMM à 3 composants, le BIC un modèle à 4 composants. Nous avons finalement choisi de prendre un modèle à 4 composants car la somme $AIC(\mathcal{M}) + BIC(\mathcal{M})$ est inférieure pour 4 composants que pour 3 composants.

Maintenant, si nous réalisons la [Figure 2.2](#) en colorant les lentilles selon le groupe de redshift selon le modèle GMM que l'on vient de réaliser, nous obtenons la figure suivante :

Plots des différences entre bandes photométriques pour les lentilles gravitationnelles en fonction du redshift, coloré selon le modèle GMM

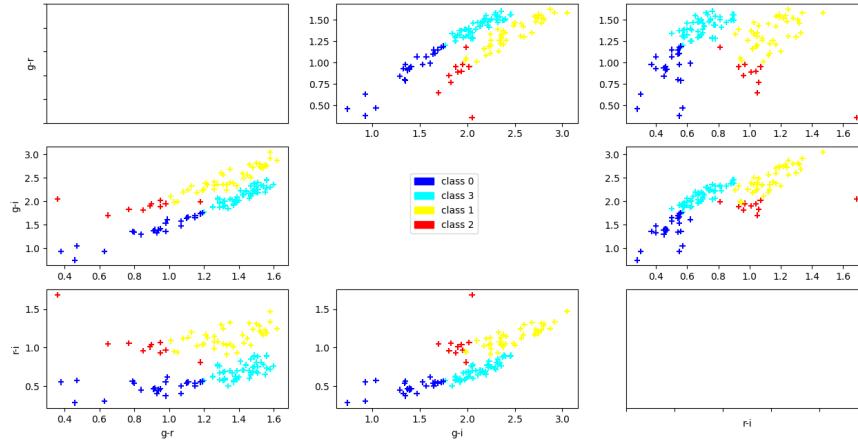


Fig. 2.4 – Nuages de points des différences entre les bandes photométriques. Les lentilles sont colorées selon le groupe de redshift choisi par le modèle GMM.
Package Scikit-Learn [Ped+11]; Source : [Hud+12]

Ce graphique peut être mis en parallèle avec le graphique des mêmes nuages de points, colorés cette fois ci en quatre classes de redshift selon les quartiles de redshift par exemple.

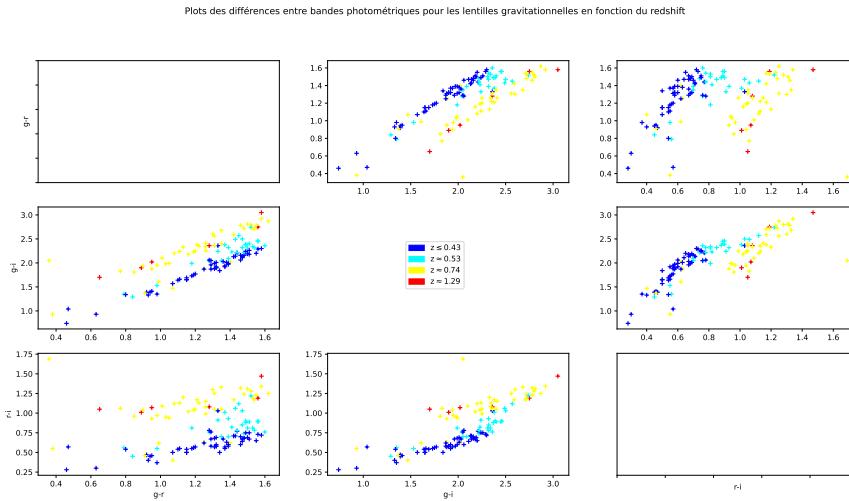


Fig. 2.5 – Nuages de points des différences entre les bandes photométriques. Les lentilles sont colorées selon le groupe de redshift choisi par quartile de redshift.
Package Scikit-Learn [Ped+11]; Source : [Hud+12]

Enfin, si nous revenons à l'observation initiale, à savoir un découpage en deux groupes, nous obtenons la Figure 2.6. Le choix d'un découpage à ± 0.5 est motivé par les valeurs de médianes dans Figure B.3 et Figure 2.5.

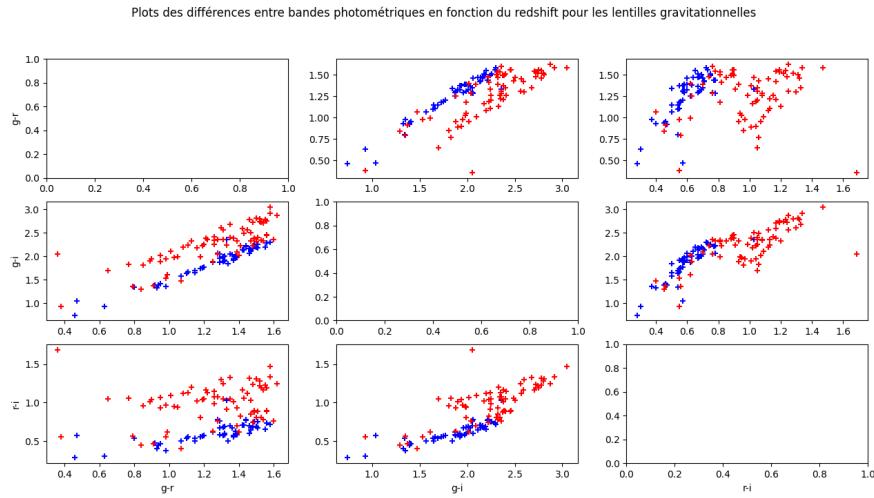


Fig. 2.6 – Nuages de points des différences entre les bandes photométriques. Les lentilles sont colorées selon leur valeur de redshift.

Source : [Hud+12]

Finalement, nous venons de montrer qu'il est possible d'estimer le redshift des lentilles en fonction des valeurs des bandes photométriques. On peut utiliser une multitude de méthodes pour extraire ces corrélations (GMM, OLS, PCA⁶, CNN⁷), mais en astrophysique, nous sommes intéressés aux causes physiques et aux effets. Ici, la cause des corrélations entre redshift et bandes photométriques peut être interprétée de façon robuste par l'âge de l'objet étudié, de sa composition (gaz, métaux, poussières), et par sa distance. C'est ce que les astrophysiciens peuvent faire de façon paramétrique en simulant les spectres attendus par des modèles bayésiens avec un *a priori* totalement défini par la physique de l'objet dans une fourchette de valeurs de paramètres assez réduite et une comparaison directe de l'*a priori* et de l'observation (par un Chi2 par exemple). Il est également possible d'utiliser des méthodes non paramétriques. Les plus utilisées en astrophysique sont les HCA⁸, les PCA, les GMM et aujourd'hui, l'apprentissage profond de réseaux de neurones (CNN, ANN⁹, GNN¹⁰). L'estimation du redshift des lentilles permet d'estimer la distance entre l'observateur, c'est à dire la Terre, et la lentille observée. Ainsi, cela donne une indication de la disposition 3D des lentilles gravitationnelles.

6. analyse en composantes principales

7. réseau neuronal convolutif

8. classification ascendante hiérarchique

9. réseau neuronal artificiel

10. réseau neuronal graphique

Chapitre 3

Corrélation spatiale et regroupement en clusters

L'étude des corrélations entre variables ([section 2.1](#)) a mis en évidence de fortes corrélations dans les données de positions des lentilles. Cependant, bien que leurs coordonnées de positions soient corrélées, nous ne savons pas dans quelle mesure cette corrélation est utilisable pour mesurer la tendance des lentilles à se regrouper en amas. L'objectif ici est donc de quantifier la tendance des lentilles gravitationnelles à se retrouver regroupées en amas.

3.1 Surdensité locale de lentille

Pour mesurer la tendance des lentilles à se regrouper en amas, on peut commencer par observer les régions dans lesquelles on retrouve une surdensité projetée anormale de lentilles. Il serait ensuite possible d'étudier ces régions pour tenter d'expliquer ces surdensités. De plus, les modèles de formation ne prédisent pas de surdensités au-delà d'un certain seuil. Mais les observations pourraient en montrer. Les plus fortes surdensités sont des tests sensibles des modèles. Détecter les plus fortes surdensités est donc intéressant.

En astrophysique, le modèle le plus simple de répartition des sources lumineuses suppose que chaque source est identique aux autres, ponctuelle (il est toujours possible d'adapter la taille des pixels pour que chaque pixel contienne au plus une source complète), et que les sources sont identiquement distribuées [[Pee20](#)]. Ces hypothèses définissent alors un processus de Poisson, c'est à dire une chaîne de Markov telle que le nombre d'occurrences dans un intervalle de longueur t suit une loi de Poisson. Nous présentons d'abord la représentation des quatre champs W1-4 étudiés et la disposition des lentilles en leur sein : [Figure 3.2](#).

Nous remarquons tout d'abord que le champs W1 est le plus dense, tandis que le champs W4 est plutôt vide. La densité moyenne de lentilles par degré carré ([Tableau 3.1](#)) souligne la sous-densité globale nette du champs W4.

Champ	Nombre de lentilles	deg^2	Densité	Bruit Poissonnien
W1	60	73	0.82	0.106
W2	17	25	0.68	0.165
W3	31	49	0.63	0.114
W4	10	25	0.40	0.125

TABLE 3.1 – Densité de lentilles par degrés carré selon les champs observés

Source : [[Hud+12](#)]

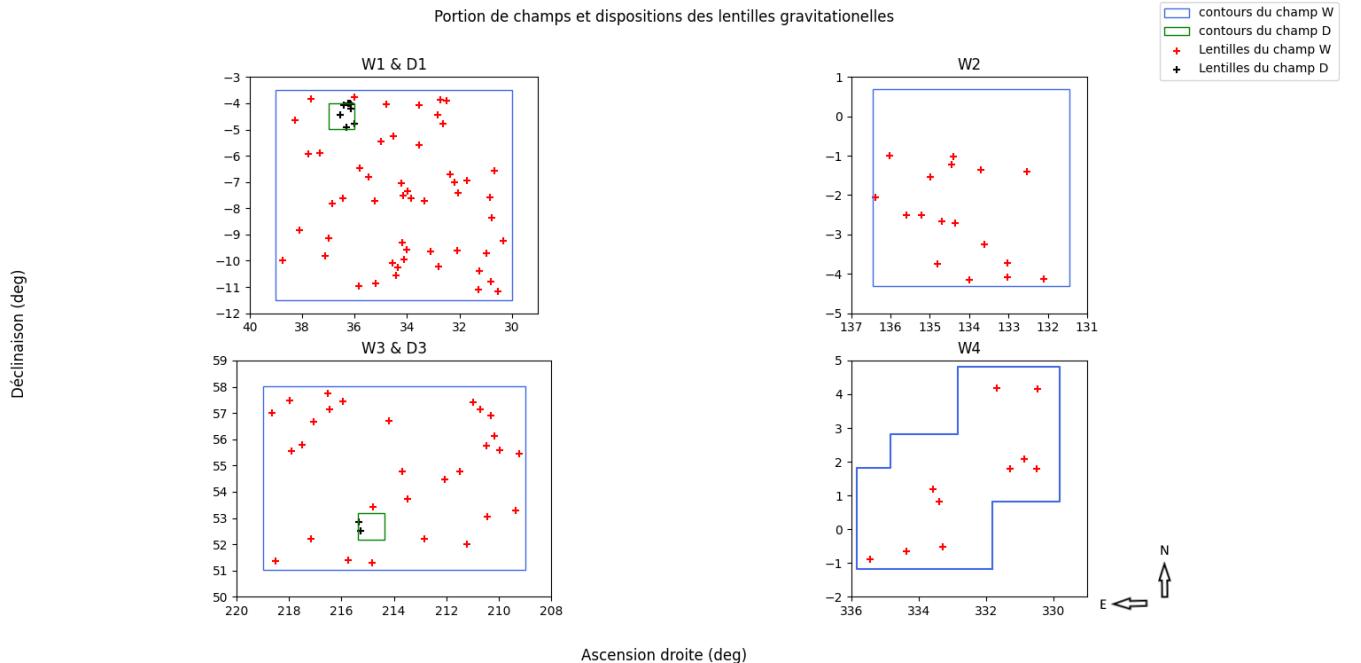


Fig. 3.2 – Disposition spatiale des lentilles gravitationnelles : chaque croix correspond à une unique lentille, les contours bleus délimitent les bords des champs W1-4. L'axe des abscisses est décroissant car les coordonnées célestes (ascensions droites) augmentent d'Ouest en Est, selon les mouvements du soleil. Source : [Hud+12]

3.1.1 Zones de surdensité et sous-densité locales

Pour étudier les surdensités relatives de lentilles dans les champs, nous allons appliquer un algorithme de *count-in-cells*. Nous commençons par diviser le champ W étudié en $l \times L$ subdivisions de tailles égales. En notant N le nombre réel de lentilles observées dans ce champ, nous tirons aléatoirement la position de N lentilles. Nous effectuons t tirages aléatoires comme celui-ci. Pour chaque tirage, nous mesurons la densité moyenne obtenue dans chaque subdivision. Nous comparons enfin cette densité obtenue par tirages aléatoires à la densité réelle observée dans le champ. Ainsi, nous avons un indicateur des zones de surdensité projetée de lentilles et des zones de sous-densité projetée.

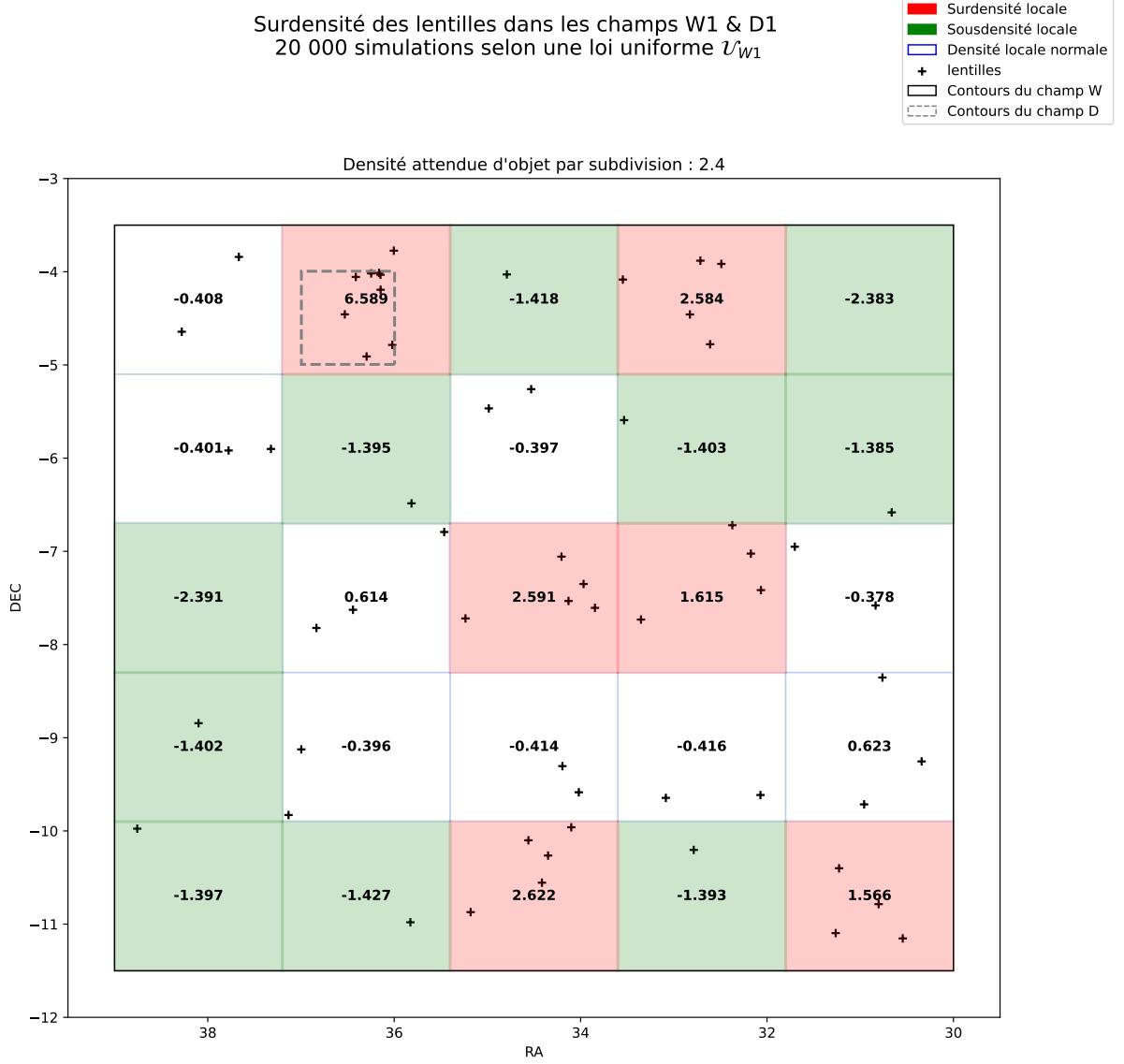


Fig. 3.3 – Graphique de surdensité et sous-densité locale de lentilles pour le champ W1. La valeur numérique inscrite au centre de chaque case correspond à la différence entre la densité réelle de lentille et la densité obtenu par moyennage de tirages aléatoires selon une loi uniforme \mathcal{U}_{W1} .
Source : [Hud+12]

Dans la Figure 3.3, nous comparons le nombre de lentilles en valeur absolue dans chaque subdivision du champ. Une autre façon de comparer la densité observée et la densité issue des tirages aléatoires serait d'étudier la distance de la densité observée en terme d'écart-type. En notant σ l'écart-type de l'échantillon $((n_{s,i})_{s,i \in [1,l] \times [1,t]})$ avec s une subdivision et i le numéro du tirage aléatoire, nous pouvons comparer le nombre de lentilles observées et σ . C'est ce que nous faisons dans la Figure 3.4.

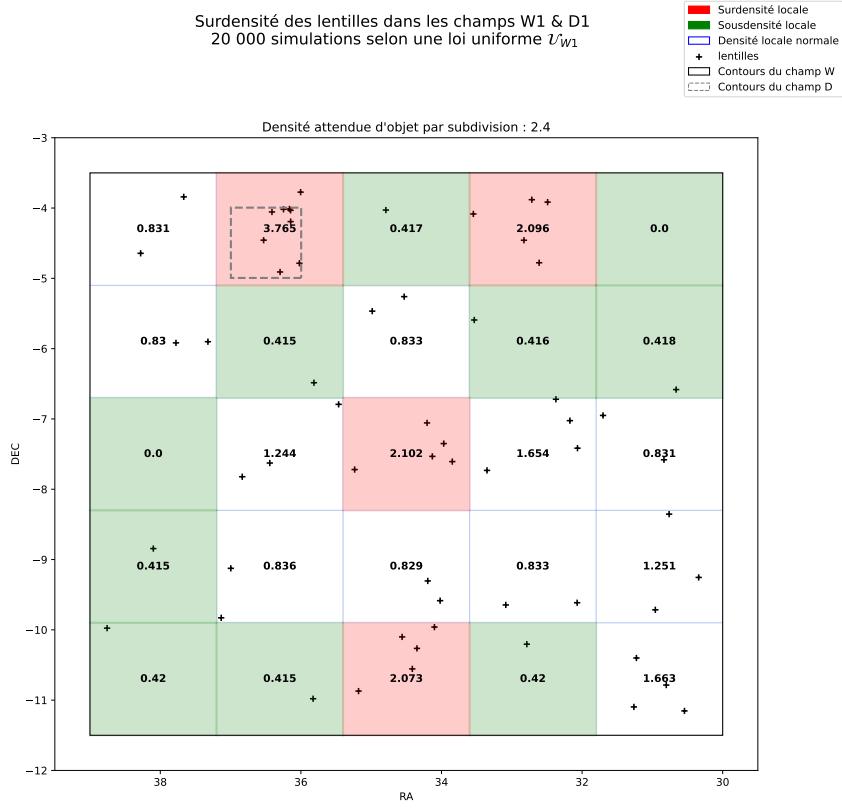


Fig. 3.4 – Graphique de la surdensité et sous-densité locale de lentilles pour le champ W1. La valeur numérique inscrite au centre de chaque case correspond au rapport entre le nombre réel de lentilles dans la subdivision et l'écart-type du nombre de lentilles obtenu par moyennage de tirages aléatoires selon une loi uniforme \mathcal{U}_{W1} .
Source : [Hud+12]

Que cela soit en terme de différence du nombre de lentilles ou d'écart à l'écart-type, il apparaît immédiatement que certaines zones sont des zones de sous-densité de lentilles, tandis que d'autres sont des zones de surdensité. Les mêmes zones sont caractérisées de la même manière par les deux méthodes.

Cela reste vrai pour les graphiques de surdensité projetée pour les champs W2 (Figure B.4, Figure B.5) et W3 (Figure B.6, Figure B.7).

3.1.2 Pics de surdensité

En appliquant ce même algorithme de *count-in-cell* et en mettant en place un découpage en subdivisions plus nombreuses et donc plus fines, nous pouvons mesurer cette fois-ci des pics de surdensité. Nous rajoutons une estimation de la densité locale par noyau Gaussien, avec une estimation de la largeur de bande suffisamment fine pour repérer les pics de surdensité.

L'existence de ces pics de surdensité mérite quelques approfondissements théoriques. Il est communément pensé que la LSS de l'univers observée aujourd'hui provient de perturbations Gaussiennes presque parfaites dans les tout premiers instants de l'univers. Cela signifie qu'il avait une symétrie parfaite dans l'abondance et l'amplitude des régions surdenses et sous-denses dans les premiers instants de l'univers. L'attraction gravi-

tationnelle a ensuite causé l'effondrement des zones de surdensité initiales en de petites structures largement surdenses, telles que des amas de galaxies. Les zones de sous-densité initiales se sont étendues mais sont restées des zones de sous-densité moyenne et sont devenues, par exemple, des cavités de vide¹. Par conséquent, la majorité du volume de l'univers récent est sous-dense. Cette large sous-densité est compensée par la présence de quelques zones de grande surdensité [Fri+18].

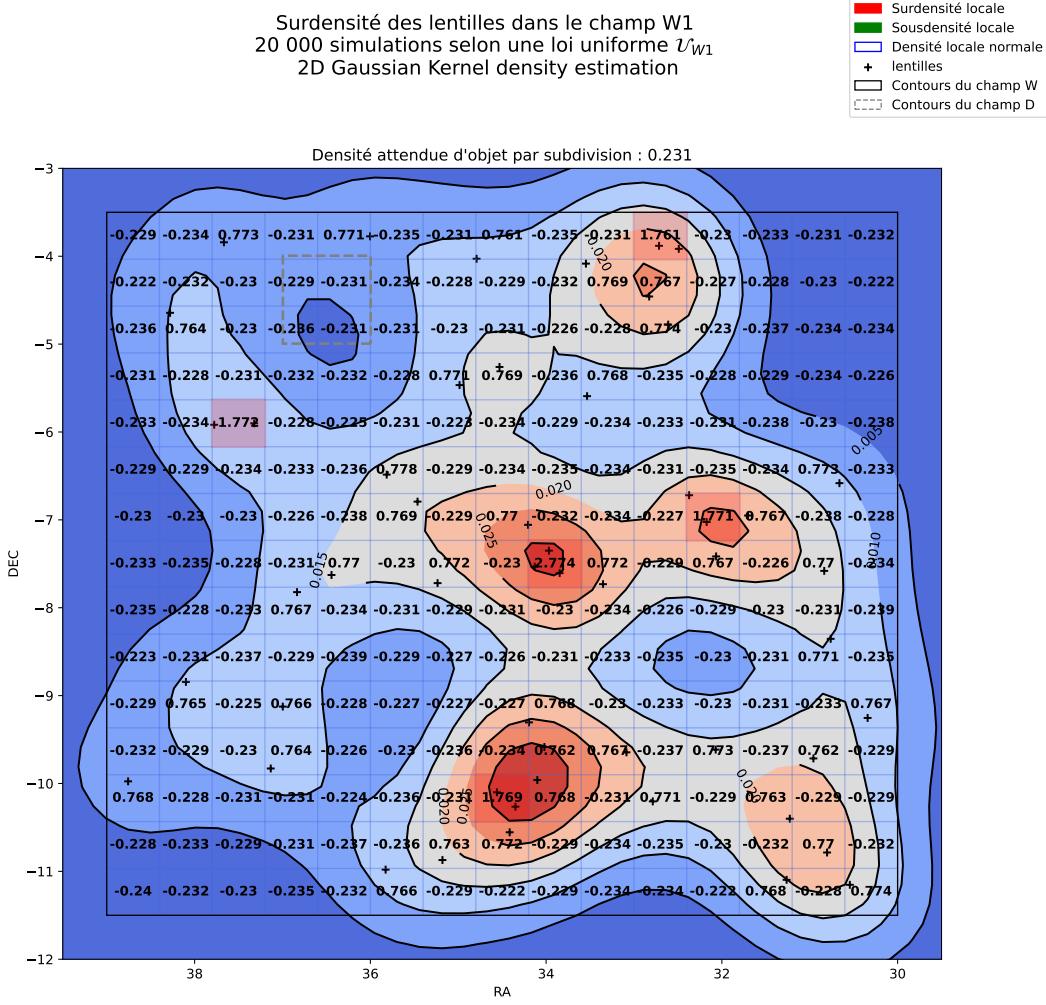


Fig. 3.5 – Graphique des pics de surdensité et sous-densité locale de lentilles pour le champ W1. La valeur numérique inscrite au centre de chaque case correspond à la différence entre la densité réelle de lentilles et la densité obtenue par moyennage de tirages aléatoires selon une loi uniforme \mathcal{U}_{W1} . La carte de couleur au dessus du quadrillage est obtenue par estimation non paramétrique de la densité de lentille par estimateur de Parzen-Rosenblatt avec un noyau Gaussien et une largeur de bande suffisamment fine pour rendre compte des pics de surdensité.

Source : [Hud+12]

Si nous nous concentrons sur les pics de surdensité dans le champ W1\|D1, nous comptons cinq pics de

1. En astrophysique, on appelle **vide** une zone de l'espace 3D de densité de matière très faible.

surdensité pour 11 lentilles réparties dans ses pics sur les $73 - 1 = 72$ degrés carrés du nouveau champ W1. Oguri propose une courbe permettant de prédire de façon théorique le nombre de lentilles jouant un rôle dans les pics de surdensité. Plus précisément, est proposée la contribution des différents types de halos sur la distribution de la séparation des images. [Ogu06]. Si l'on se réfère à cette courbe et que l'on applique les transformations adéquates pour se ramener à notre cas d'étude, nous sommes censés obtenir 0.18 lentille de pic de surdensité/degré carré.

Via notre algorithme, nous en observons 0.15/degré carré dans le champs W1. De même, nous obtenons 0.16 lentille/degré carré dans le champ W2 (Figure B.8) et 0.12 dans le champ W3 (Figure B.9). Finalement, nous obtenons des résultats en accord avec les prévisions théoriques de Oguri.

3.1.3 Corrélation au weak-lensing

Jusqu'à présent, nous étudions le lentillage gravitationnel fort, c'est à dire un lentillage suffisamment dense et massif pour produire une image comme présentée dans la Figure 2. Ce lentillage requiert, en plus d'une forte masse, un très bon alignement.

Cependant, il existe d'autres types de lentillages. En particulier, le lentillage faible se produit avec des objets massifs, mais dont l'alignement n'est pas parfait. Ce type de lentillage produit alors de faibles modifications dans notre perception visuelle des objets avoisinants. Le lentillage faible du champ W1 a été étudié pour développer une *mass map* par la mesure de la forme de 2.66 million de galaxies dans ce champ [Sha+12].

Du fait du grand nombre d'objet étudiés, les résultats obtenus sont largement bruités. Pour un meilleur lissage des résultats, l'article fournit différentes cartes convolées à différents niveaux (de 1 à 6 amin), ainsi qu'une produite par l'algorithme MRLens [Pir+10] qui effectue un filtrage non Gaussien des pics, mais résulte en un bruitage non Gaussien également qui complique la sélection.

Nous superposons nos résultats de densité de lentilles avec la carte de weak-lensing.

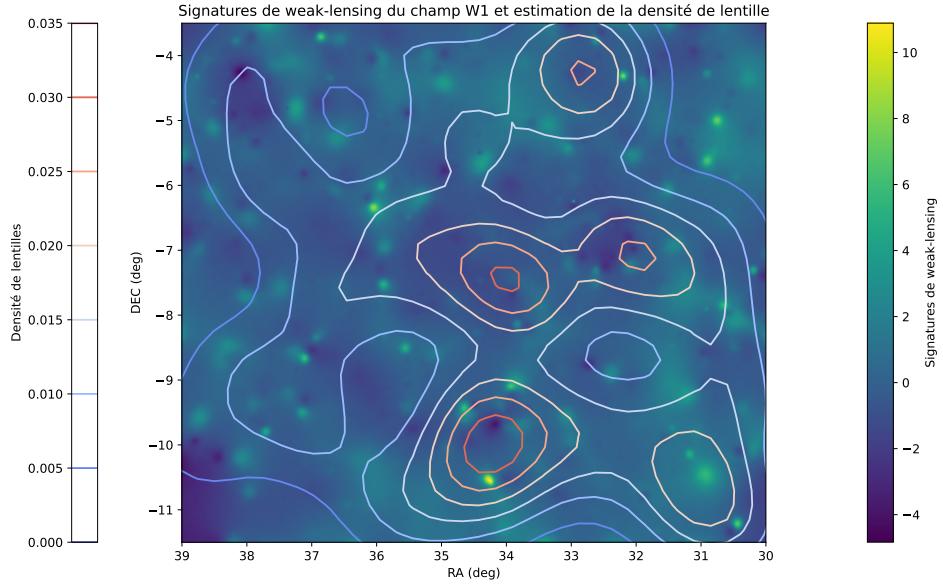


Fig. 3.6 – Carte de la signature du champ W1 en weak-lensing. Les données sont "smoothées" par l'algorithme MRLens. Nous superposons la densité de lentilles, identique à celle de la Figure 3.5. Source : [Hud+12] [Sha+12]

Nous n'avons malheureusement pas eu le temps d'étudier les corrélations entre les pics de weak-lensing et

les pics de surdensité de strong-lensing. Nous souhaitons cependant mettre en avant la corrélation visuelle, faible mais présente, entre les pics de strong et weak lensing. En effet, dans les zones de surdensité de strong-lensing, nous notons également la présence de pics jaunes, synonymes de pics de weak-lensing.

Il nous reste maintenant à comprendre et expliquer pourquoi certaines zones témoignent d'une surdensité ou d'une sous-densité de lentilles gravitationnelles.

3.2 Fonction de corrélation à 2 points

La mesure la plus communément utilisée pour mesurer la tendance des lentilles à se regrouper en amas est la fonction de corrélation à 2 points. On la note $\zeta(r)$. Cette fonction de corrélation rend compte de la quantité d'amas en fonction de leur échelle.

La fonction de corrélation à 2 points ζ est définie comme une mesure de la probabilité dP , au dessus de laquelle il est attendu, pour une distribution de Poisson, de trouver une autre lentille dans un volume dV à une distance r de notre lentille [DP83]. On a ainsi la relation :

$$dP = n(\zeta(r) + 1)dV \quad \text{avec } n \text{ la densité moyenne d'objet présent dans l'échantillon. [Pee20]}$$

La mesure de ζ reste pour autant le problème principal. En théorie, il suffit de compter les paires de lentilles comme une fonction de la distance r qui les sépare, et de diviser par la distribution attendue. Pour obtenir cette distribution attendue, il faut alors construire un catalogue aléatoire ayant exactement la même couverture tridimensionnelle (ascension droite, déclinaison et redshift) que les données de lentilles, mais ayant une population tirée aléatoirement. Comme nous ne connaissons pas parfaitement la distribution tridimensionnelle des données, notamment la distribution "lissée" du redshift, nous utiliserons des estimateurs non paramétriques pour mesurer ζ .

Historiquement, le premier estimateur à avoir été proposé est celui de Davis & Peebles [DP83]. En notant DD la quantité de paires de lentilles dans les données utilisées, et DR le nombre de paires entre les données et le catalogue aléatoire, et en notant n_D et n_R le nombre moyen d'objet dans les données et dans l'échantillon aléatoire, on obtient $\zeta_{DP} = \frac{n_R}{n_D} \frac{DD}{DR} - 1$.

Les estimateurs ont été améliorés au cours des années et aujourd'hui, c'est plutôt l'estimateur de Landy & Szalay [LS93] qui est utilisé :

$$\zeta_{LS} = \frac{1}{RR} \left[DD \left(\frac{n_R}{n_D} \right)^2 - 2DR \left(\frac{n_R}{n_D} \right) + RR \right]$$

L'objectif est également de mesurer l'incertitude de notre estimateur. Ce dernier est non-paramétrique car on ignore la loi réelle de l'échantillon. Il est donc impossible d'utiliser des méthodes habituelles de variance ou écart-type pour mesurer l'incertitude de façon classique. Nous allons alors approcher la distribution de l'estimateur par des simulations de type Monte-Carlo, selon une méthode de Bootstrapping².

Le bootstrapping est une méthode de ré-échantillonage où l'on substitue à la distribution de probabilité inconnue F , dont est issu l'échantillon d'apprentissage, la distribution empirique \hat{F} qui donne un poids $\frac{1}{n}$ à chaque réalisation. Ainsi on obtient un échantillon de taille n : l'échantillon bootstrap, qui suit la distribution empirique \hat{F} par n tirages aléatoires indépendants avec remise parmi les n observations initiales.

Dans notre cas particulier, on souhaite estimer l'écart-type par la méthode de bootstrapping car on souhaite mesurer la précision de notre estimateur ζ_{LS} . On note $X = (x_1, \dots, x_n)$ notre échantillon et $X^* = (x_1^*, \dots, x_n^*)$ un échantillon bootstrap de X de taille n , qui suit donc la loi \hat{F} , distribution empirique de X . L'estimation bootstrap de l'écart type $\widehat{\sigma}_{\zeta_{LS}}$ se calcule par une estimation plug-in. Cependant, il n'existe pas de formule connue pour le calcul direct de cet estimateur. On utilise donc une simulation de type Monte-Carlo via l'algorithme Figure B.10.

². ET93.

Afin de mettre en pratique la théorie expliquée ci-dessus, nous avons fait le choix d'utiliser les fonctions codées dans le package python **AstroML**³. Les fonctions codées en son sein sont une adaptation du livre de statistique⁴. Certaines fonctions ont du être adaptées à la marge pour un fonctionnement optimal dans notre cadre d'étude.

On rappelle que la disposition spatiale des lentilles est disponible sur la [Figure 3.2](#).

On applique ensuite l'algorithme décrit plus tôt pour estimer ζ_{LS} dans chacun des quatre champs. Nous faisons le choix d'une estimation séparée par champ pour que l'écart spatial entre les champs n'entre pas en compte dans le calcul de l'estimateur.

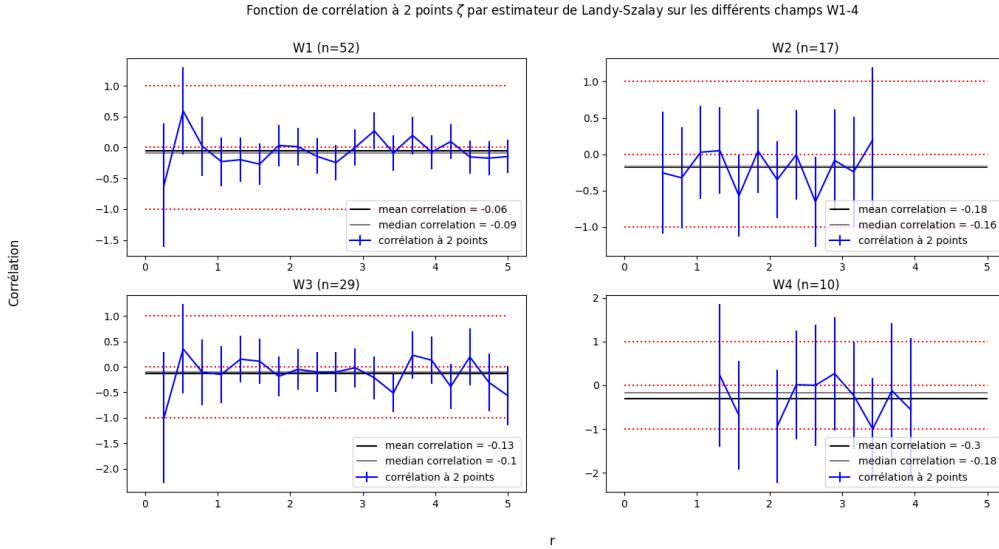


Fig. 3.7 – Estimateur ζ_{LS} sur chaque portion de champ. Le faible nombre d'individus dans les champs 2, 3 et 4 fait que l'erreur est parfois incontrôlée et rend l'analyse difficile.

Source : [Hud+12]

Les résultats obtenus sont assez peu probants ([Figure 3.7](#)). Pour les champs W1-3, la corrélation oscille dans les alentours de 0^- . Pour le champ W4, la corrélation à 2 points est plus forte avec une corrélation moyenne négative de -0.36. La plus grosse difficulté réside dans la petite taille de nos échantillons⁵. Du fait de cette petite taille, l'estimation des erreurs par méthode de bootstrap reste relativement grande devant les valeurs prises par la fonction de corrélation. Nous aurons donc du mal à exploiter ces résultats.

3.3 Densité de galaxies voisines

Une autre idée pour étudier la disposition géographique des lentilles gravitationnelles est d'étudier, pour chaque lentille, son nombre de galaxies voisines dans un rayon proportionnel au rayon d'Einstein R_E de la lentille (cf. [section 1.3](#) pour la définition du R_E). En effet, on estime que le lentillage à plus de chances d'avoir lieu lorsque de nombreuses galaxies sont proches. Ainsi, on peut espérer voir apparaître une relation dans la distance entre des lentilles proches et le nombre de galaxies voisines dans un rayon arbitraire. Nous faisons le choix de chercher dans des rayons de respectivement $3R_E$ et $4R_E$.

Nous mettons en place un algorithme permettant de compter, pour chaque lentille, son nombre de galaxies proches dans le rayon choisi. Cependant, pour les quelques 127 lentilles à notre disposition, nous disposons

3. [Van+12](#).

4. [Ive+14](#).

5. Pour des fonctions de corrélations à deux points faites avec des dizaines de milliers de galaxies, on peut se référer à [CdH00]

de presque 9 millions de galaxies. Calculer la distance de chaque lentille à chaque galaxie serait trop coûteux en ressources de calculs. Nous améliorons donc la complexité du calcul en ordonnant les galaxies selon une certaine coordonnée (RA ou DEC) puis en segmentant les galaxies selon leur champ d'observation. De plus, il est important de retirer du calcul les galaxies qui composent elles mêmes la lentille gravitationnelle. Pour se faire, nous retirons du calcul les galaxies dont le redshift est inférieur à $1.5 \times \text{redshift}_{\text{lentille}}$. Ainsi, nous ne prenons en compte dans le calcul que les galaxies d'arrière plan qui ne font pas partie de la lentille observée.

Nous dessinons les histogrammes du nombre de galaxies voisines pour chaque lentille ([Figure B.11](#)), puis nous faisons le même graphique en normalisant par la densité de galaxies dans le champ de la lentille ([Tableau 3.1](#), [Figure 3.8](#)).

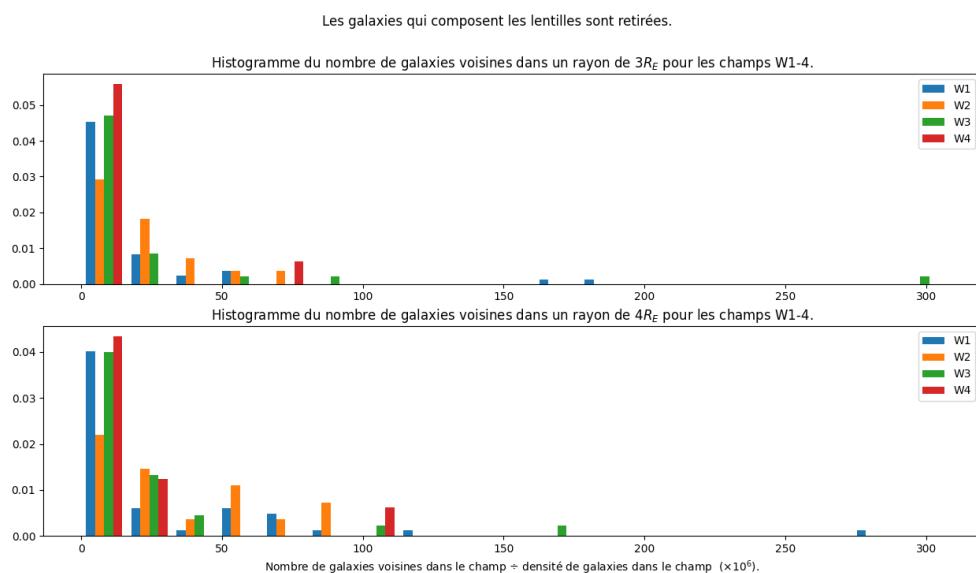


Fig. 3.8 – Nombre de galaxies voisines divisé par la densité de galaxies dans le champ dans un rayon de 3 et 4 R_E autour de la lentille.
Source : [[Hud+12](#)]

Nous observons que la distribution du nombre de galaxies voisines suit une loi qui décroît exponentiellement quand le nombre de voisines augmente. Cela est vrai dans les quatre champs W1-4. Il apparaît par ailleurs que 95 % des lentilles ont moins de cinq galaxies voisines et 99% des lentilles ont moins de quinze voisines.

Champ	0	≤ 1	≤ 2	≤ 5	≤ 10	≤ 15
W1	42	73	87	96	96	100
W2	24	47	76	100		
W3	59	76	90	93	97	97
W4	70	90	90	100		

TABLE 3.9 – Pourcentage de lentilles ayant au plus n galaxies voisines selon les champs.
42% des lentilles du champ W1 ont 0 galaxie voisine. 96% en ont 10 ou moins.

Source : [[Hud+12](#)]

Les lentilles ayant sensiblement plus de galaxies voisines sont d'autant plus rares qu'elles ont beaucoup de voisines. Il serait alors intéressant de comparer cette distribution avec celle liée à un tirage aléatoire de 127 lentilles prises n'importe où dans le champ. Ainsi, nous verrions si le nombre de galaxies voisines des lentilles se comporte comme pour une galaxie normale ou si une surpopulation de lentilles avec de nombreuses

voisines apparaît.

Cependant, pour éviter le bruit des résultats lié au tirage aléatoire, il faudrait réaliser cette expérience un grand nombre de fois. Or, comme nous l'avons vu, malgré nos tentatives pour réduire la complexité algorithmique du problème, chaque tirage prend plusieurs minutes à être calculé. Ainsi, il faudrait encore améliorer le code, en passant par exemple par des questions de parallélisation, d'optimisation et en utilisant des clusters de calculs plus puissants. Par manque de temps, nous ne pourrons réaliser ce projet.

Il aurait ensuite convenu de mettre cette répartition de galaxies voisines en écho avec la répartition de lentilles voisines. En effet, l'objectif est d'étudier si les lentilles ayant un grand nombre de voisines sont proches les unes des autres et ainsi essayer de mettre en lumière une potentielle liaison entre le nombre de galaxies voisines et la densité de lentilles dans cet espace local avec beaucoup de galaxies voisines. Il est également envisageable de corrélérer le tout avec la densité de galaxies et de lentilles par degré carré dans chaque champ ([Tableau 3.1](#), [Tableau B.12](#)).

Notons tout de même que l'étude du nombre de galaxies voisines est à prendre avec précaution. En effet, les groupements de lentilles sont très éloignés les uns des autres (de plusieurs centaines de R_E) et il n'y a pas de raison théorique selon laquelle nous devrions détecter des surdensités de galaxies autour des groupes de lentilles.

Conclusion

Résultats obtenus

Nous rappelons l'objectif de ce travail : étudier la répartition des lentilles gravitationnelles, notamment en densité projetée de lentilles.

Avant même de commencer à étudier la répartition des lentilles gravitationnelles, nous avons pris le temps d'étudier l'échantillon du CFHTLS à notre disposition et ses potentiels biais, notamment liés aux valeurs de temps d'exposition et de seeing. Nous avons pu conclure que l'échantillon était non biaisé. L'échantillon étant de plus annoncé comme complet, nous pouvons commencer l'analyse.

Après une étude des corrélations entre les différentes données de photométrie, de position et autres données des lentilles, nous avons proposé deux modèles pour estimer le redshift photométrique des lentilles. Le modèle de régression linéaire offre une très bonne estimation du redshift photométrique. Le modèle de mélanges Gaussiens, dont les paramètres sont choisis via les modèles BIC et AIC, nous permet une estimation en classe de redshift photométrique. La connaissance du redshift des objets permet d'estimer leur distance à la terre via la loi de Hubble, puis aux autres objets célestes. Ceci permet ensuite une meilleure analyse du regroupement en amas de lentilles dans un univers en 3D.

Enfin, nous étudions la surdensité projetée des lentilles. Nous mettons en avant l'existence de zones de surdensité et de sous-densité, ainsi que la présence de pics de surdensité dont les valeurs sont cohérentes avec les prédictions de Oguri⁶. Nous proposons également une ébauche pour mettre ces pics de surdensité en parallèle avec la signature de weak-lensing du champ. Nous tentons d'expliquer l'existence de zones de surdensité par l'étude de la fonction de corrélation à deux points mais les résultats sont peu probants, notamment du fait de la faible taille de nos échantillons qui ne permet pas d'obtenir une erreur satisfaisante.

Critiques et améliorations possibles

La première critique a émettre sur ce travail est la grande erreur de précision du redshift photométrique devant le redshift spectroscopique. En effet, là où le redshift spectroscopique propose une erreur de l'ordre de 10^{-4} sur les valeurs de redshift des objets, le redshift photométrique à une erreur de l'ordre de 10^{-1} . Ainsi, même si nos travaux permettent une très bonne estimation des valeurs de redshift photométrique, la mesure du redshift spectroscopique permettrait une amélioration de la précision de l'ordre de 1000. Sur la Figure 3.10, les traits bleus représentent l'estimation du redshift photométrique tandis que les points rouges, beaucoup plus précis, représentent l'estimation du redshift spectroscopique, pour les objets de W1.

Une seconde critique peut se porter sur la sous-section 3.1.3 qui étudie les liens entre la signature en weak-lensing du champ W1 et la disposition des pics de surdensité de strong-lensing. Cette partie, non terminée, aurait pu être menée en étudiant les corrélations entre la disposition des pics de weak-lensing et des pics de surdensité de strong-lensing.

De plus, les travaux sur la fonction de corrélation à deux points et sur la densité de galaxies voisines ne sont pas très probants. Pour la fonction de corrélation à deux points, la faible quantité d'individus dans les

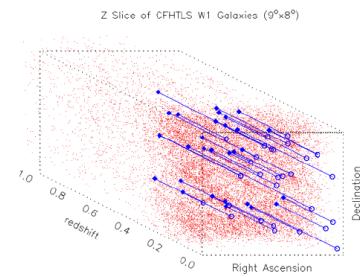


Fig. 3.10 – Illustration de l'erreur entre le redshift spectroscopique et le redshift photométrique. Crédits : R. CABANAC, private communication, 2012

6. Oguri06.

champs 2-4 est à l'origine d'une très mauvaise estimation de l'erreur.

Enfin, ma méconnaissance des notions d'astrophysiques nécessaires a sans doute été un frein à mon travail. En effet, je suppose qu'un astrophysicien serait plus à même de relever les interprétations métier des résultats obtenus.

Bibliographie

- [Ast+18] ASTROPY COLLABORATION et al. « The Astropy Project : Building an Open-science Project and Status of the v2.0 Core Package ». In : 156.3, 123 (sept. 2018), p. 123. DOI : [10.3847/1538-3881/aabc4f](https://doi.org/10.3847/1538-3881/aabc4f). arXiv : [1801.02634 \[astro-ph.IM\]](https://arxiv.org/abs/1801.02634).
- [Bla+60] A. BLAAUW et al. « The New I.A.U. System of Galactic Coordinates (1958 Revision) ». In : *Monthly Notices of the Royal Astronomical Society* 121.2 (août 1960), p. 123-131. ISSN : 0035-8711. DOI : [10.1093/mnras/121.2.123](https://doi.org/10.1093/mnras/121.2.123). eprint : <https://academic.oup.com/mnras/article-pdf/121/2/123/8078181/mnras121-0123.pdf>. URL : <https://doi.org/10.1093/mnras/121.2.123>.
- [BMP11] Micol BOLZONELLA, Joan-Marc MIRALLES et Roser PELLÓ. *Hyperz : Photometric Redshift Code*. Août 2011. ascl : [1108.010](https://ascl.net/1108.010).
- [CdH00] R. A. CABANAC, V. DE LAPPARENT et P. HICKSON. « Evolution of faint galaxy clustering. The 2-point angular correlation function of 20,000 galaxies to $V < 23.5$ and $I < 22.5$ ». In : *Astronomy and Astrophysics* 364 (déc. 2000), p. 349-368. arXiv : [astro-ph/0007184 \[astro-ph\]](https://arxiv.org/abs/astro-ph/0007184).
- [DP83] Marc DAVIS et P. PEEBLES. « A survey of galaxy redshifts. V. The two-point position and velocity correlations. » In : *The Astrophysical Journal* 267 (1983), p. 465-482.
- [Ein36] Albert EINSTEIN. « LENS-LIKE ACTION OF A STAR BY THE DEVIATION OF LIGHT IN THE GRAVITATIONAL FIELD ». In : *Science* 84.2188 (1936), p. 506-507. ISSN : 0036-8075. DOI : [10.1126/science.84.2188.506](https://doi.org/10.1126/science.84.2188.506). eprint : <https://science.sciencemag.org/content/84/2188/506.full.pdf>. URL : <https://science.sciencemag.org/content/84/2188/506>.
- [ET93] Bradley EFRON et Robert J. TIBSHIRANI. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. Boca Raton, Florida, USA : Chapman & Hall/CRC, 1993.
- [Fri+18] O. FRIEDRICH et al. « Density split statistics : Joint model of counts and lensing in cells ». In : *Physical Review D* 98.2 (juil. 2018). ISSN : 2470-0029. DOI : [10.1103/physrevd.98.023508](https://doi.org/10.1103/physrevd.98.023508). URL : [http://dx.doi.org/10.1103/PhysRevD.98.023508](https://dx.doi.org/10.1103/PhysRevD.98.023508).
- [Har+20] Charles R. HARRIS et al. « Array programming with NumPy ». In : *Nature* 585.7825 (sept. 2020), p. 357-362. DOI : [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2). URL : <https://doi.org/10.1038/s41586-020-2649-2>.
- [Hud+12] Patrick HUDELOT et al. *T0007 : The Final CFHTLS Release, Executive Summary*. Report. TERAPIX-CFHTLS, 2012. URL : <https://cfhtls.calet.org/T07/doc/T0007-doc.pdf>.
- [Hun07] J. D. HUNTER. « Matplotlib : A 2D graphics environment ». In : *Computing in Science & Engineering* 9.3 (2007), p. 90-95. DOI : [10.1109/MCSE.2007.55](https://doi.org/10.1109/MCSE.2007.55).
- [Ive+14] Ž. IVEZIĆ et al. *Statistics, Data Mining and Machine Learning in Astronomy*. Princeton, NJ : Princeton University Press, 2014.
- [Low99] Richard LOWRY. *Concepts & Applications of Inferential Statistics - The Wilcoxon Signed-Rank Test*. 1999. URL : <http://vassarstats.net/textbook/ch12a.html>.
- [LS93] S. LANDY et A. SZALAY. « Bias and variance of angular correlation functions ». In : *The Astrophysical Journal* 412 (1993), p. 64-71.
- [McK10] Wes McKINNEY. « Data Structures for Statistical Computing in Python ». In : *Proceedings of the 9th Python in Science Conference*. Sous la dir. de Stéfan van der WALT et Jarrod MILLMAN. 2010, p. 56-61. DOI : [10.25080/Majora-92bf1922-00a](https://doi.org/10.25080/Majora-92bf1922-00a).
- [Mor+12] A. MORE et al. « The CFHTLS-Strong Lensing Legacy Survey (SL2S) : Investigating the Group-scale Lenses with the SARCS Sample ». In : *ApJ* 749.1, 38 (avr. 2012), p. 38. DOI : [10.1088/0004-637X/749/1/38](https://doi.org/10.1088/0004-637X/749/1/38). arXiv : [1109.1821 \[astro-ph.CO\]](https://arxiv.org/abs/1109.1821).
- [Ogu06] M. OGURI. « The image separation distribution of strong lenses : halo versus subhalo populations ». In : *Monthly Notices of the Royal Astronomical Society* 367.3 (avr. 2006), p. 1241-1250. ISSN : 1365-2966. DOI : [10.1111/j.1365-2966.2006.10043.x](https://doi.org/10.1111/j.1365-2966.2006.10043.x). URL : [http://dx.doi.org/10.1111/j.1365-2966.2006.10043.x](https://doi.org/10.1111/j.1365-2966.2006.10043.x).

- [Ped+11] F. PEDREGOSA et al. « Scikit-learn : Machine Learning in Python ». In : *Journal of Machine Learning Research* 12 (2011), p. 2825-2830.
- [Pee20] P. J. E. PEEBLES. *The Large-Scale Structure of the Universe*. Princeton University Press, 2020. ISBN : 9780691206714. DOI : [doi:10.1515/9780691206714](https://doi.org/10.1515/9780691206714). URL : <https://doi.org/10.1515/9780691206714>.
- [Pir+10] S. PIRES et al. « Cosmological model discrimination from weak lensing data ». In : t. 1241. Juin 2010, p. 1118-1127. DOI : [10.1063/1.3462608](https://doi.org/10.1063/1.3462608).
- [Sha+12] HuanYuan SHAN et al. « WEAK LENSING MEASUREMENT OF GALAXY CLUSTERS IN THE CFHTLS-WIDE SURVEY ». In : *The Astrophysical Journal* 748.1 (mar. 2012), p. 56. ISSN : 1538-4357. DOI : [10.1088/0004-637x/748/1/56](https://doi.org/10.1088/0004-637x/748/1/56). URL : <http://dx.doi.org/10.1088/0004-637x/748/1/56>.
- [Shu] Michael SHULL. *Simulation of Cosmic Web*. <https://ecuip.lib.uchicago.edu/multiwavelength-astronomy/ultraviolet/science/08.html>. Accessed : 02/08/2021.
- [SP10] Skipper SEABOLD et Josef PERKTOLD. « statsmodels : Econometric and statistical modeling with python ». In : *9th Python in Science Conference*. 2010.
- [Van+12] J.T. VANDERPLAS et al. « Introduction to astroML : Machine learning for astrophysics ». In : *Conference on Intelligent Data Understanding (CIDU)*. Oct. 2012, p. 47-54. DOI : [10.1109/CIDU.2012.6382200](https://doi.org/10.1109/CIDU.2012.6382200).
- [VD09] Guido VAN ROSSUM et Fred L. DRAKE. *Python 3 Reference Manual*. Scotts Valley, CA : CreateSpace, 2009. ISBN : 1441412697.
- [WEL47] B. L. WELCH. « The Generalization of "Student's problem when several different population variances are involved ». In : *Biometrika* 34.1-2 (jan. 1947), p. 28-35. ISSN : 0006-3444. DOI : [10.1093/biomet/34.1-2.28](https://doi.org/10.1093/biomet/34.1-2.28). eprint : <https://academic.oup.com/biomet/article-pdf/34/1-2/28/553093/34-1-2-28.pdf>. URL : <https://doi.org/10.1093/biomet/34.1-2.28>.
- [Wil45] Frank WILCOXON. « Individual Comparisons by Ranking Methods ». In : *Biometrics Bulletin* 1.6 (1945), p. 80-83. ISSN : 00994987. URL : <http://www.jstor.org/stable/3001968>.

Annexe A

Compléments d'analyses statistiques

A.1 Seeing

A.1.1 Normalité de l'échantillon : Nous utilisons d'abord un diagramme quantile-quantile (ou QQ-plot) pour évaluer l'ajustement de nos échantillons à une loi normale centrée réduite $\mathcal{N}(0, 1)$. Nous avons évidemment centré et réduit nos échantillons pour pouvoir appliquer cette comparaison en soustrayant à chaque individu la moyenne de l'échantillon et en divisant par l'écart-type de l'échantillon.

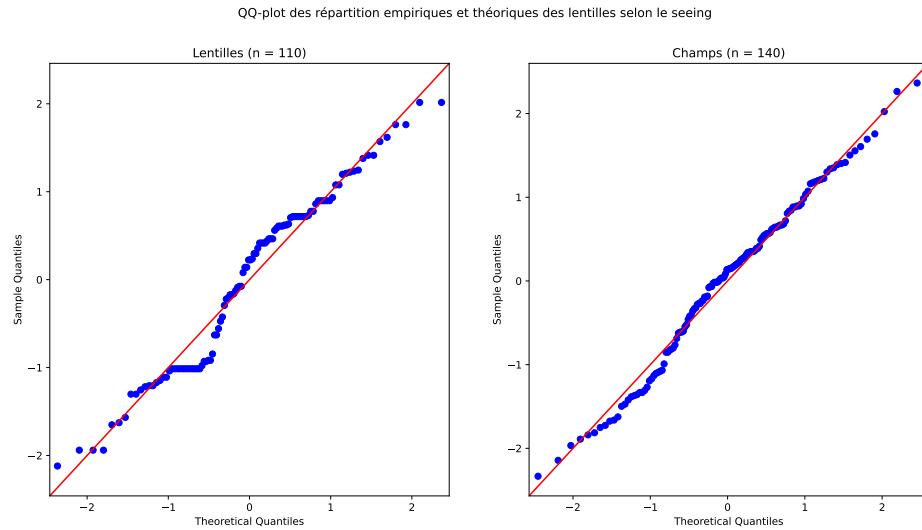


Fig. A.1 – Normalité des échantillons en fonction du seeing. Les quantiles empiriques et théoriques sont globalement alignés. On considère que l'échantillon suit une loi normale.

Source : [Hud+12]

A.1.2 Test-t de Welch : Nous appliquons ensuite l'algorithme du test-t de Welch comme définit dans [WEL47].

```
1 # T test de Welch
2 _, p_value = scipy.stats.ttest_ind(seeing_list_lentille_01,
3                                     seeing_list_obj_01, equal_var=False)
4 p_value
```

Listing A.2 – T-test de Welch

entrée python

```
1 >>> _, p_value = scs.ttest_ind(seeing_list_lentille_01,
2 ...                               seeing_list_obj_01, equal_var=False)
3 >>>
4 >>> p_value
5 0.9999999999999961
6 >>> # on ne rejette pas H_0
```

Listing A.3 – T-test de Welch

sortie terminal

A.1.3 Histogramme et densité : Finalement, nous traçons l'histogramme de nos distributions et rajoutons les densités théoriques des lois normales $\mathcal{N}(\mu_{\text{ech}_{\text{lense}}}, \sigma_{\text{ech}_{\text{lense}}})$ et $\mathcal{N}(\mu_{\text{ech}_W}, \sigma_{\text{ech}_W})$.

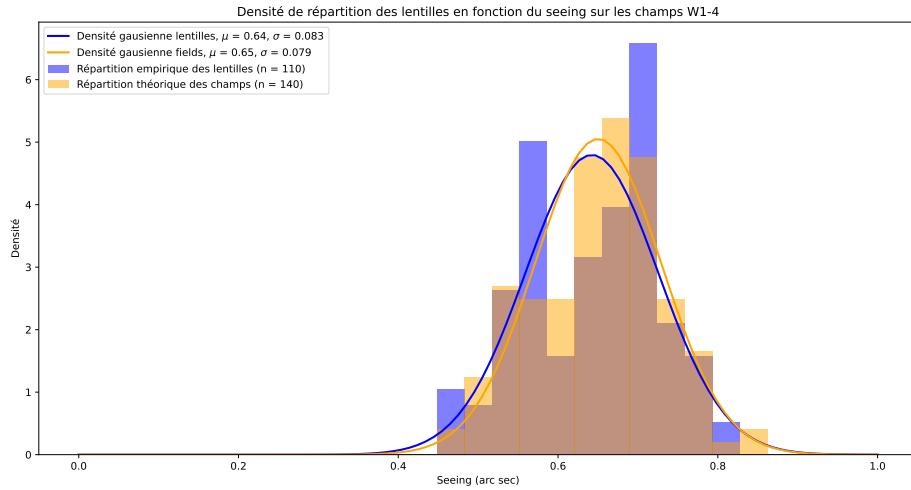
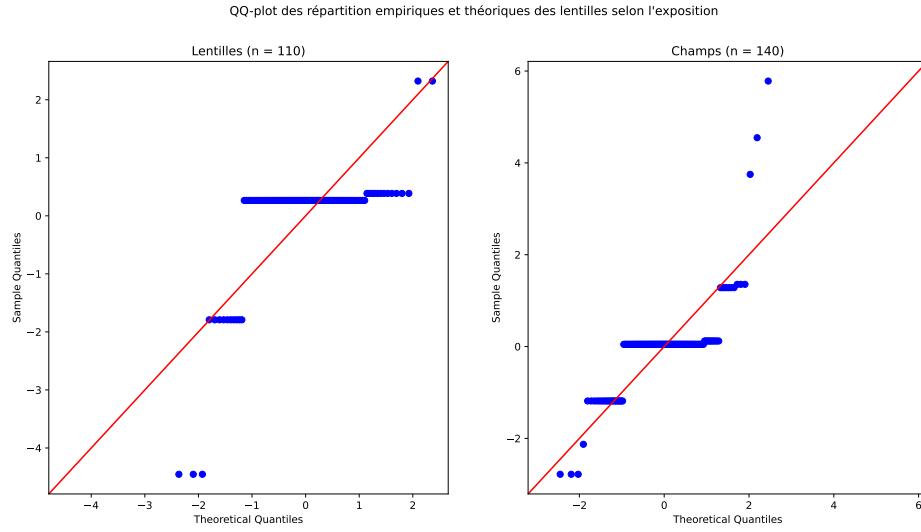


Fig. A.4 – Densités Gaussiennes des échantillons de lentille et de fields en fonction du seeing Source : [Hud+12]

Notons que l'histogramme de la répartition des lentilles apparaît comme bimodal, ce que l'on peut également relever dans le qq-plt précédent (Figure A.1) à travers les deux principaux écarts à la ligne droite. Cependant, nous faisons le choix d'apparenter ces deux modes à de la variance propre à notre échantillon. En d'autre termes, nous faisons le choix de croire que notre échantillon convergerait vers une loi normale si nous avions plus de données mais que la variance liée à la faible quantité de données cause cette impression de loi bimodale. Nous ne prenons pas le parti d'une loi bimodale Gaussienne par exemple.

A.2 Temps d'exposition

A.2.1 Normalité de l'échantillon : Nous commençons là encore par réaliser une diagramme quantile-quantile pour évaluer l'ajustement à la loi normale centrée réduite.



Listing A.5 – Normalité des échantillons en fonction du temps d'exposition. Les quantiles théoriques et empiriques ne sont pas alignés. Les deux échantillons ne suivent pas une loi normale. Source : [Hud+12]

Nous concluons qu'aucun des deux échantillons ne suit une loi normale.

A.2.2 Test des rangs signés : Nous vérifions tout de même si les deux échantillons suivent une même loi. Nous utilisons logiquement une alternative au test de Student qui suppose la normalité des échantillons.

```

1 N = 10_000
2 p_values = np.zeros(N)
3 for i in range(N):
4     np.random.shuffle(expo_list_obj_noNA)
5     expo_list_obj_noNA_len = expo_list_obj_noNA[0:len(expo_list_lentille_noNA)]
6
7     _, p_values[i] = scs.wilcoxon(x=expo_list_lentille_noNA, y=expo_list_obj_noNA_len)
8 np.mean(p_values)

```

Listing A.6 – Test non paramétrique des rangs signés de Wilcoxon

entrée python

```

1 >>> N = 10_000
2 >>> p_values = np.zeros(N)
3 >>> for i in range(N):
4     ...     np.random.shuffle(expo_list_obj_noNA)
5     ...     expo_list_obj_noNA_len = expo_list_obj_noNA[0:len(expo_list_lentille_noNA)]
6     ...     _, p_values[i] = scs.wilcoxon(x=expo_list_lentille_noNA, y=expo_list_obj_noNA_len,
7     ...                                         alternative = "two-sided")
...
8 >>> np.mean(p_values)
9 0.5645733799735497
10 >>> # on ne rejette pas H_0

```

Listing A.7 – Test non paramétrique des rangs signés de Wilcoxon

sortie terminal

Nous concluons que les deux échantillons suivent une même loi non Gaussienne.

A.3 Autres biais d'observation

A.3.1 De façon similaire à [section A.1](#) et [section A.2](#), nous étudions les possibles biais d'observations selon le seeing et le temps d'exposition en fonction des valeurs de rayon d'Einstein et de redshift des lentilles.

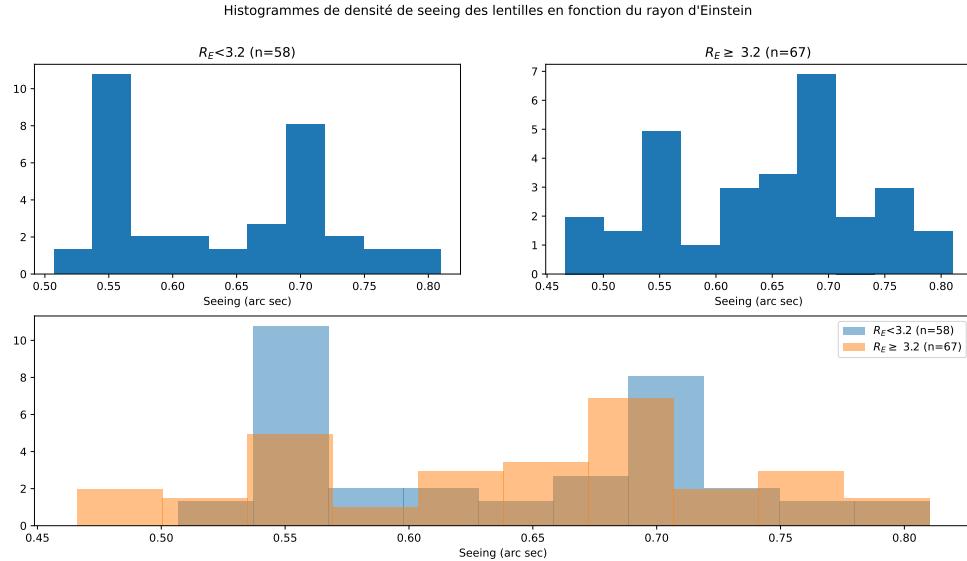


Fig. A.8 – Histogrammes de seeing selon le rayon d'Einstein. Les deux échantillons ont des histogrammes relativement semblables.

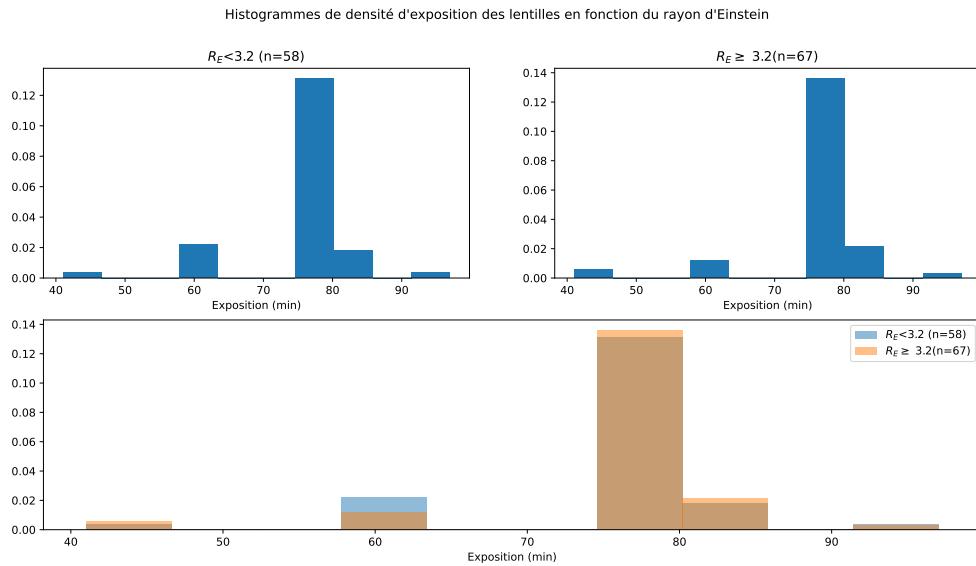


Fig. A.9 – Histogrammes d'exposition selon le rayon d'Einstein. Les deux échantillons ont des histogrammes relativement semblables.

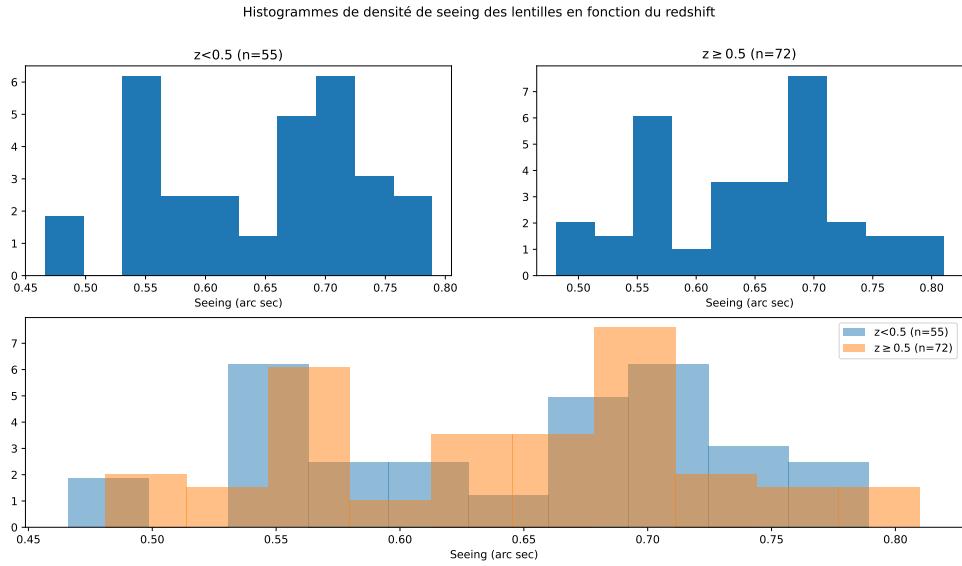


Fig. A.10 – Histogrammes de seeing selon le redshift. Les deux échantillons ont des histogrammes relativement semblables.
Source : [Hud+12]

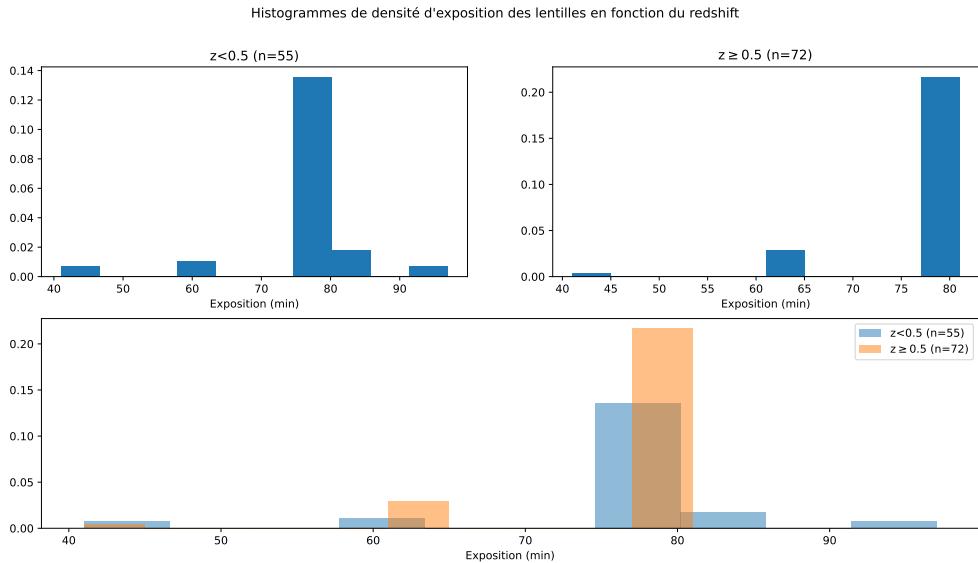


Fig. A.11 – Histogrammes d'exposition selon le redshift. Les deux échantillons ont des histogrammes relativement semblables.
Source : [Hud+12]

Source: Données CFHTLS [Hud+12]

Annexe B

Compléments sur les études de corrélations

B.1 Matrice de corrélation

B.1.1 Matrice de corrélation - ρ de Spearman : L'utilisation du ρ de Spearman pour calculer la corrélation entre deux variables plutôt que le r de Pearson permet de mettre en avant des corrélations monotones de types non affine. Par exemple, une relation de type puissance comme une relation ayant la forme de la fonction cube aura un ρ de Spearman proche de 1 tandis que le r de Pearson sera plus faible, sans doute entre 0.4 et 0.8. Cependant, il est important de noter qu'une relation non strictement monotone sera mal interprétée par les deux coefficients. C'est le cas par exemple d'une relation ayant la forme d'un binome (polynôme du second degré) pour lequel la corrélation est en réalité très forte mais que ni le coefficient ρ ni le coefficient r n'arrive à révéler.

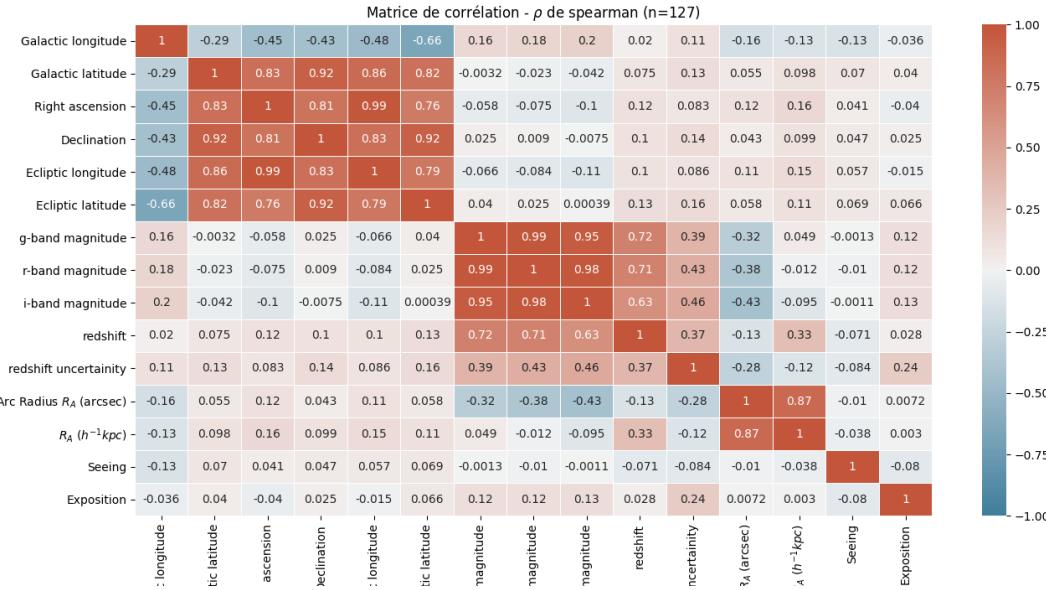


Fig. B.1 – Matrice de corrélation des données lentilles selon le ρ de Spearman. L'intensité de la couleur représente l'intensité de la corrélation. La couleur représente le sens de la corrélation. Source : [Hud+12]

B.2 Corrélation photométrique

B.2.1 Régression Linéaire : Nous effectuons une régression linéaire dont le résultat est disponible ci-après (Tableau B.2).

Nous rappelons d'abord les hypothèses de la régression linéaire qui ont été vérifiées avant la mise en place de la régression en elle-même :

- (a) Les variables explicatives du modèles sont non-colinéaires. Cela équivaut au fait que la matrice de design du modèle, notée X , soit de rang maximal.

- (b) Les erreurs sont supposées indépendantes. C'est ici le cas puisque nous utilisons un modèle avec un vecteur d'erreur $\epsilon \sim \mathcal{N}(0, \sigma^2 \text{Id})$.
- (c) Le terme d'erreur est de variance constante. On parle alors d'homoscédasticité. Cela est vérifié car $\epsilon \sim \mathcal{N}(0, \sigma^2 \text{Id})$.
- (d) Les variables explicatives sont exogènes, c'est-à-dire qu'elles sont décorrélées du terme d'erreur. On a donc $\mathbf{E}(\epsilon|X) = 0$.

B.2.2 Grâce à la vérification de ses hypothèses (en réalité, seule l'hypothèse 1 est obligatoire, les autres peuvent être corrigées par la suite), le théorème de Markov-Gauss garantit que notre estimateur est le meilleur estimateur linéaire non-biaisé. Par ailleurs, comme nous avons fait l'hypothèse de la normalité des erreurs, notre estimateur par méthode des moindres carrés ordinaires est également l'estimateur par maximum de vraisemblance.

Dep. Variable :	redshift	R-squared (uncentered) :	0.933			
Model :	OLS	Adj. R-squared (uncentered) :	0.931			
Method :	Least Squares	F-statistic :	574.8			
Date :	Thu, 05 Aug 2021	Prob (F-statistic) :	2.35e-49			
Time :	10 :10 :34	Log-Likelihood :	33.873			
No. Observations :	85	AIC :	-63.75			
Df Residuals :	83	BIC :	-58.86			
Df Model :	2					
	coef	std err	t	P> t	[0.025	0.975]
gMi	0.2363	0.008	28.054	0.000	0.220	0.253
gMr	-0.2524	0.044	-5.679	0.000	-0.341	-0.164
rMi	0.4887	0.051	9.650	0.000	0.388	0.589
Omnibus :	17.480				Durbin-Watson :	1.804
Prob(Omnibus) :	0.000				Jarque-Bera (JB) :	20.515
Skew :	1.100				Prob(JB) :	3.51e-05
Kurtosis :	3.974				Cond. No.	2.63e+16

Notes :

- [1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
- [2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

TABLE B.2 – Résultats de la régression par méthode des moindres carrés ordinaires sur l'échantillon d'apprentissage. Les tests de Student sur chaque variable explicative montre qu'elles sont toutes significatives. Package StatsModel [SP10]; Source : [Hud+12]

B.2.3 Nous découpons ensuite notre échantillon de lentilles en un échantillon d'apprentissage et un échantillon de test, de tailles relatives $\frac{2}{3} - \frac{1}{3}$. Nous apprenons le modèle de régression sur l'échantillon d'apprentissage et nous le testons sur l'échantillon de test, dont le graphique est disponible ci-après Figure B.3.

Plots des différences entre bandes photométriques pour les lentilles gravitationnelles en fonction du redshift, colorée selon les prévisions via la régression linéaire sur l'échantillon de test

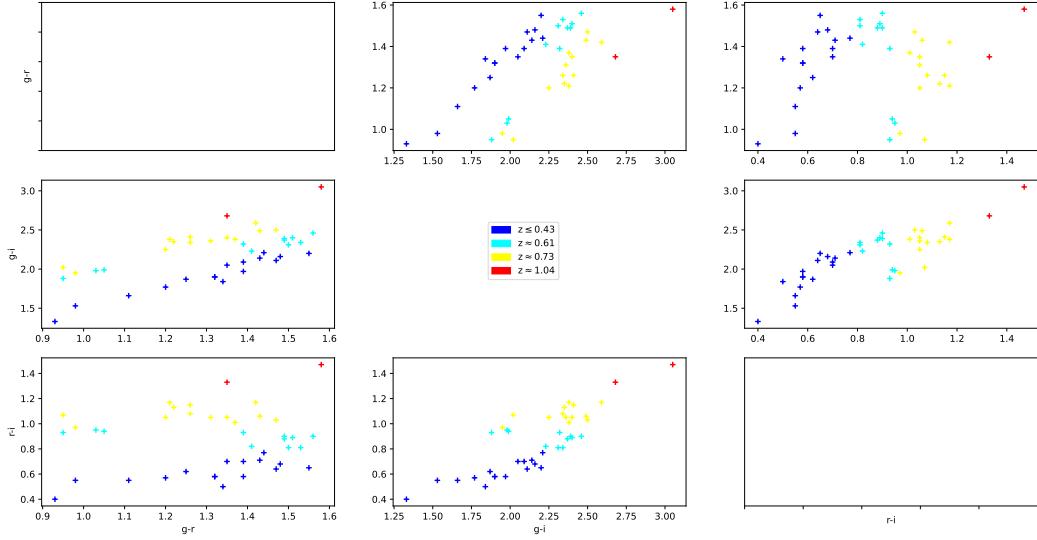


Fig. B.3 – Nuages de points des différences entre les bandes magnétiques. Les lentilles sont colorées selon le redshift estimé par la régression découpé en quartiles.

Package StatsModel [SP10]; Source : [Hud+12]

B.2.4 Précisions sur le modèle GMM : On parle de mélange Gaussien lorsque la loi de l'échantillon étudié ne suit pas une loi usuelle mais une densité mélange de lois Gaussiennes.

Une densité mélange est une fonction de densité issue d'une combinaison convexe de plusieurs fonctions de densité. Par exemple, on considère la densité \mathbf{D} d'une variable aléatoire X paramétrée par θ . Ainsi, avec la famille $(\theta_1, \dots, \theta_N)$, pondérée par les poids scalaires qui régissent le barycentre convexe (p_1, \dots, p_N) , on a notre densité mélange à N composantes donnée par

$$\mathbf{D}(X, (\theta_n)_{n \in \{1, \dots, N\}}) = \sum_{n=1}^N p_n \mathcal{D}(X, \theta_n) \quad (\text{B.1})$$

Maintenant, pour se placer dans notre cadre, le paramètre $\theta = (\mu, \sigma)$ est le paramètre d'une loi normale. Nous confondons volontairement les notations de la loi de probabilité de la fonction normale $\mathcal{N}(\mu, \sigma)$ et la fonction de densité d'une variable aléatoire X qui suivrait celle loi normale. Nous notons donc cette densité $\mathcal{N}(\mu, \sigma)$.

Ainsi, pour une famille de paramètres $(\theta_1, \dots, \theta_N) = ((\mu_1, \sigma_1), \dots, (\mu_N, \sigma_N))$, là encore pondérée par les poids (p_1, \dots, p_N) , nous avons notre densité de mélange Gaussien à N composantes donnée par

$$\mathbf{D}(X, (\mu_n, \sigma_n)_{n \in \{1, \dots, N\}}) = \sum_{n=1}^N p_n \mathcal{N}(\mu_n, \sigma_n) \quad (\text{B.2})$$

Notons que l'utilisation d'une combinaison convexe de densité assure que notre nouvelle fonction \mathbf{D} est une loi de densité.

Par la suite, on étend notre définition avec des lois normales multidimensionnelles de \mathbb{R}^p , notre paramètre θ devenant donc $\theta = (\mu, \Sigma)$, avec $\mu \in \mathbb{R}^p$ la moyenne et $\Sigma \in \mathcal{M}_p$ la matrice de variance-covariance.

Ainsi, l'objectif premier réside en la détermination du paramètre global du mélange, que nous noterons

$$\Theta = (p_1, \dots, p_N, \theta_1, \dots, \theta_N) = (p_1, \dots, p_N, \mu_1, \dots, \mu_N, \Sigma_1, \dots, \Sigma_N) \quad (\text{B.3})$$

C'est dans cet objectif que nous utilisons, comme précisé dans la [sous-section 2.2.2](#), la méthode du maximum de vraisemblance. Plus précisément, nous cherchons le paramètre Θ qui maximise la vraisemblance, notée \mathcal{L} .

Pour un échantillon de M individus $\mathcal{X} = (X_1, \dots, X_M)$, et sous hypothèse d'indépendance des individus, nous obtenons

$$\mathcal{L}(\mathcal{X}, \Theta) = \sum_{m=1}^M \ln \left(\sum_{n=1}^N p_n \mathcal{N}_{X_m}(\mu_n, \Sigma_n) \right) \quad (\text{B.4})$$

L'estimation de $\hat{\Theta}^{MV}$ s'effectue selon l'algorithme d'espérance-maximisation. Une fois l'estimateur $\hat{\Theta}^{MV}$ obtenu, il reste à classer les individus $X_m \in \mathcal{X}$ dans la classe la plus probable. On note les classes $(\mathcal{C}_1, \dots, \mathcal{C}_n)$. Pour se faire, on utilise la formule de Bayes, à savoir

$$\mathbb{P}(X_m \in \mathcal{C}_n) = \frac{p_n \mathcal{N}_{X_m}(\mu_n, \Sigma_n)}{\sum_{j=1}^N p_j \mathcal{N}_{X_m}(\mu_j, \Sigma_j)} \quad (\text{B.5})$$

On attribut finalement chaque individu X_m à la classe \mathcal{C}_n dont la probabilité à posteriori est la plus grande.

B.3 Corrélation spatiale

B.3.1 Étude de la surdensité projetée des lentilles : Nous commençons par étudier les zones de sous-densité de surdensité projetée de lentilles en écart absolu ([Figure 3.3](#), [Figure B.4](#), [Figure B.6](#)) et en distance à l'écart-type ([Figure 3.4](#), [Figure B.5](#), [Figure B.7](#)).

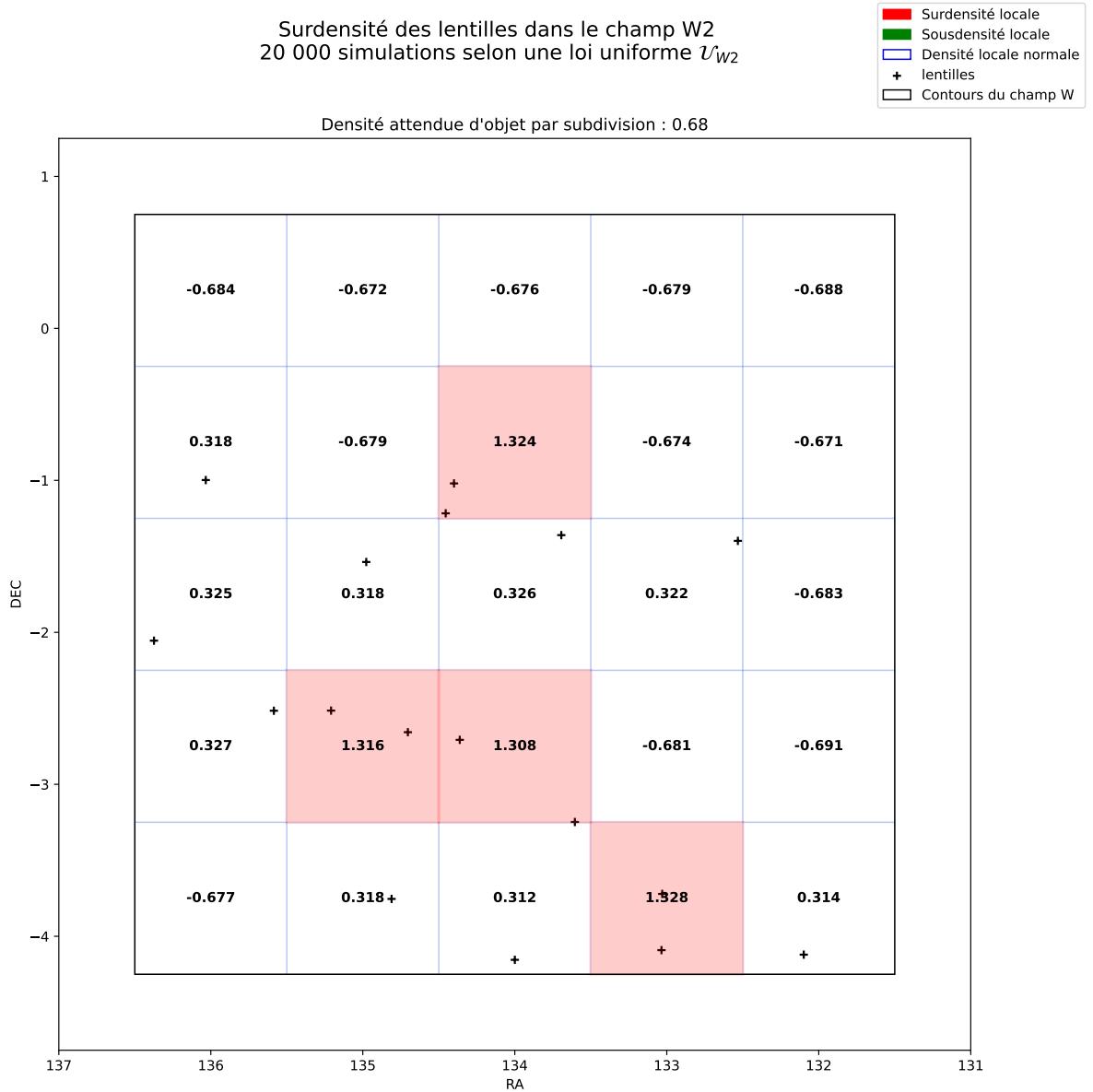


Fig. B.4 – Graphique de la surdensité et sous-densité locale de lentilles pour le champ W2. La valeur numérique inscrite au centre de chaque case correspond à l'écart densité réelle de lentille – la densité obtenue par moyennage de tirages aléatoires selon une loi uniforme \mathcal{U}_{W2} .

Source : [Hud+12]

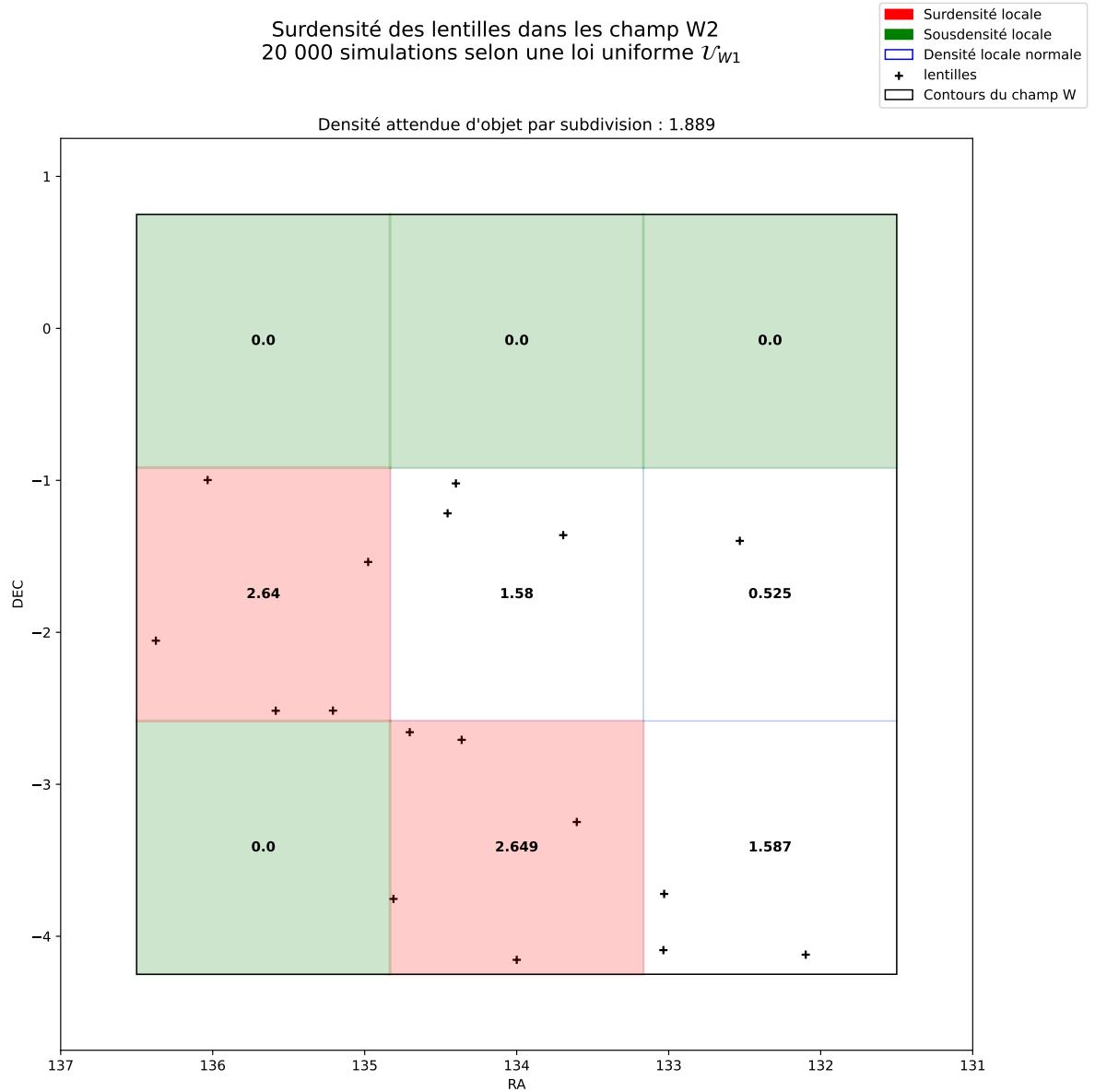


Fig. B.5 – Graphique de la surdensité et sous-densité locale de lentilles pour le champ W2. La valeur numérique inscrite au centre de chaque case correspond au rapport entre la le nombre réel de lentille dans la subdivision et l'écart-type du nombre de lentilles obtenu par moyennage de tirages aléatoires selon une loi uniforme \mathcal{U}_{W2} .

Source : [Hud+12]

Surdensité des lentilles dans le champ W3
20 000 simulations selon une loi uniforme \mathcal{U}_{W3}

- █ Surdensité locale
- █ Soudensité locale
- █ Densité locale normale
- + lentilles
- Contours du champ W
- Contours du champ D

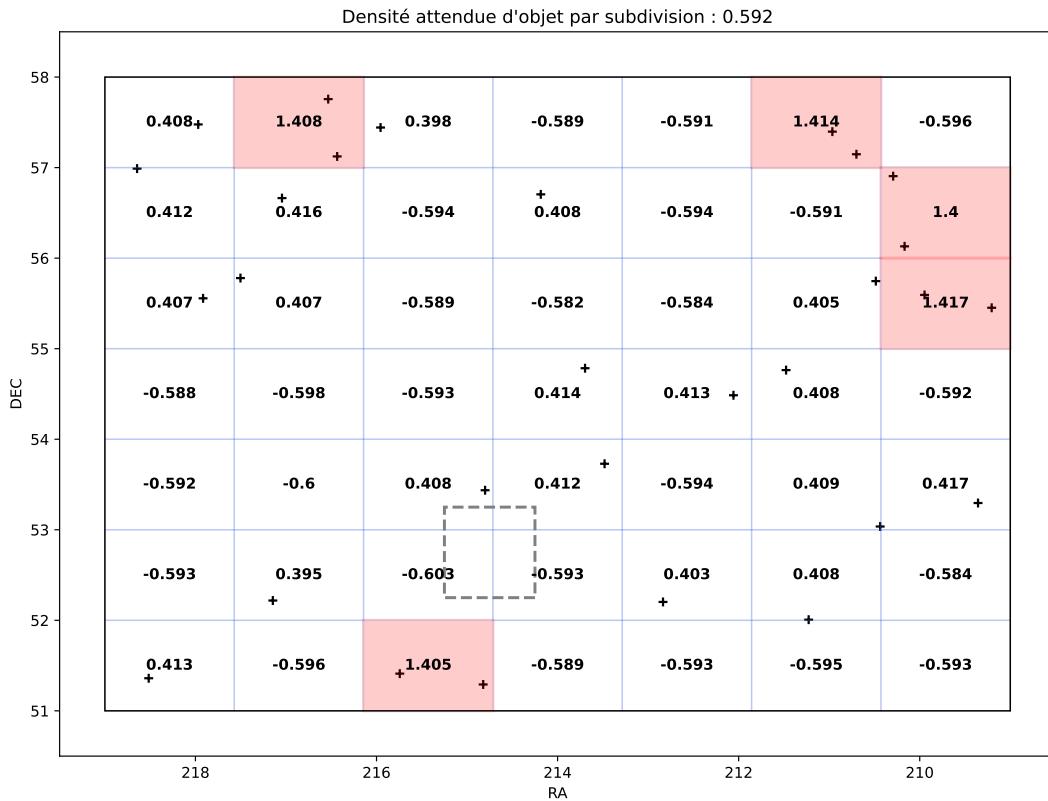


Fig. B.6 – Graphique de la surdensité et sous-densité locale de lentilles pour le champ W3. La valeur numérique inscrite au centre de chaque case correspond à l'écart densité réelle de lentille – la densité obtenue par moyennage de tirages aléatoires selon une loi uniforme \mathcal{U}_{W3} .

Source : [Hud+12]

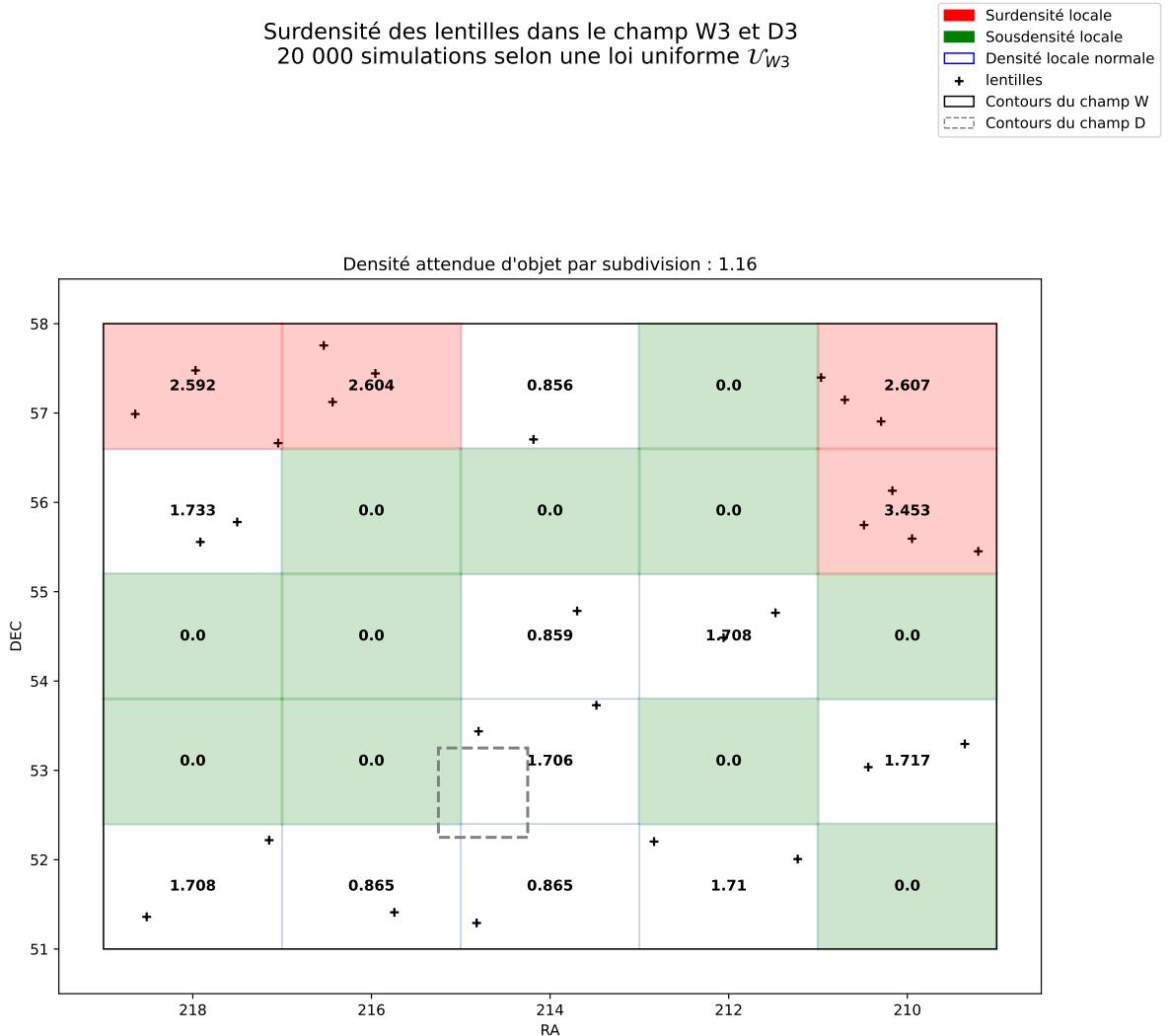


Fig. B.7 – Graphique de la surdensité et sous-densité locale de lentilles pour le champ W3. La valeur numérique inscrite au centre de chaque case correspond au rapport entre la le nombre réel de lentille dans la subdivision et l'écart-type du nombre de lentilles obtenu obtenu par moyennage de tirages aléatoires selon une loi uniforme \mathcal{U}_{W3} .

Source : [Hud+12]

B.3.2 Nous étudions ensuite les pics de surdensité projetée de lentilles (Figure 3.5, Figure B.8, Figure B.9).

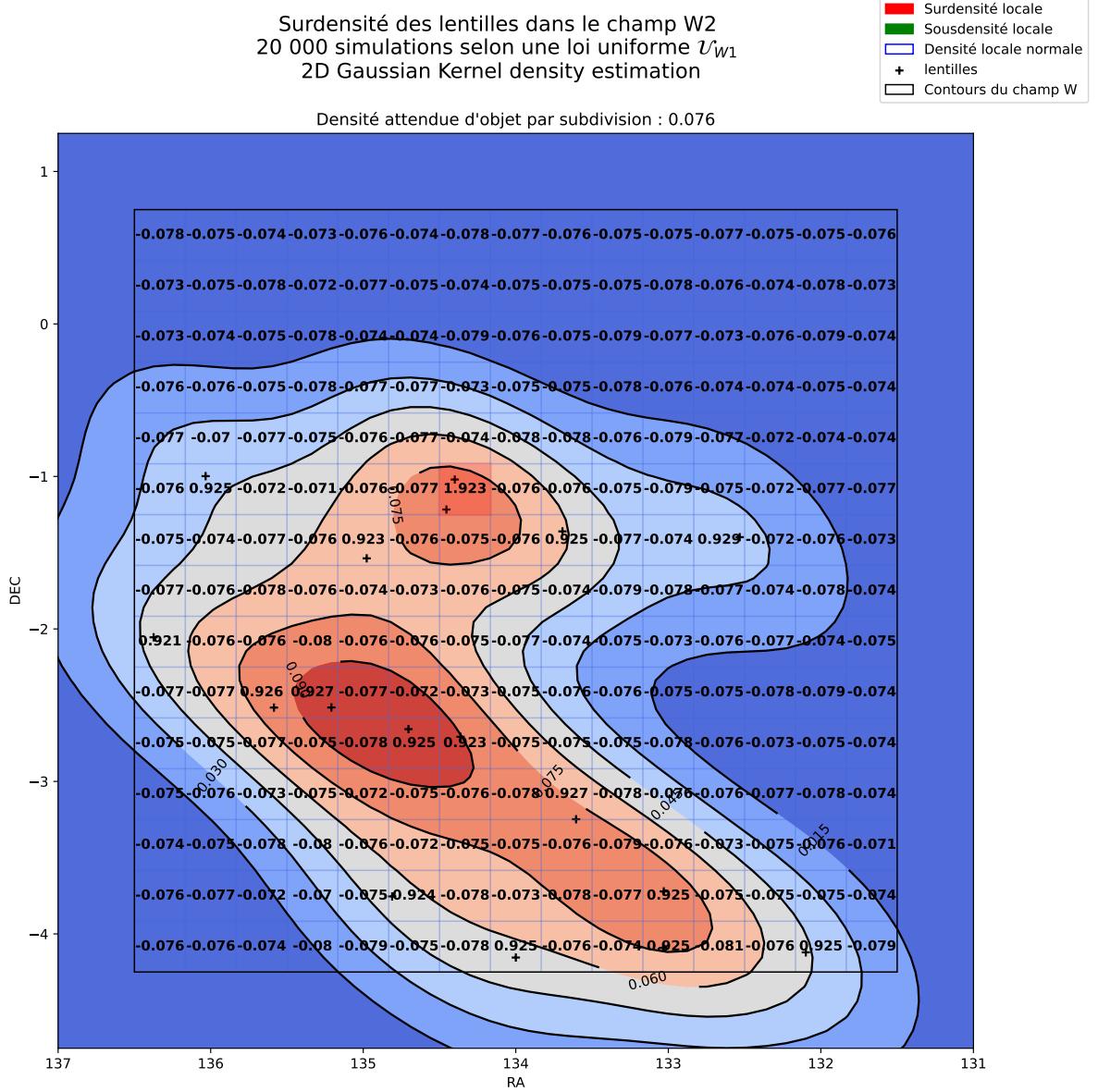


Fig. B.8 – Graphique des pics de surdensité et sous-densité locale de lentilles pour le champ W2. La valeur numérique inscrite au centre de chaque case correspond à l'écart densité réelle de lentille – la densité obtenue par moyennage de tirages aléatoires selon une loi uniforme \mathcal{U}_{W2} .

Source : [Hud+12]

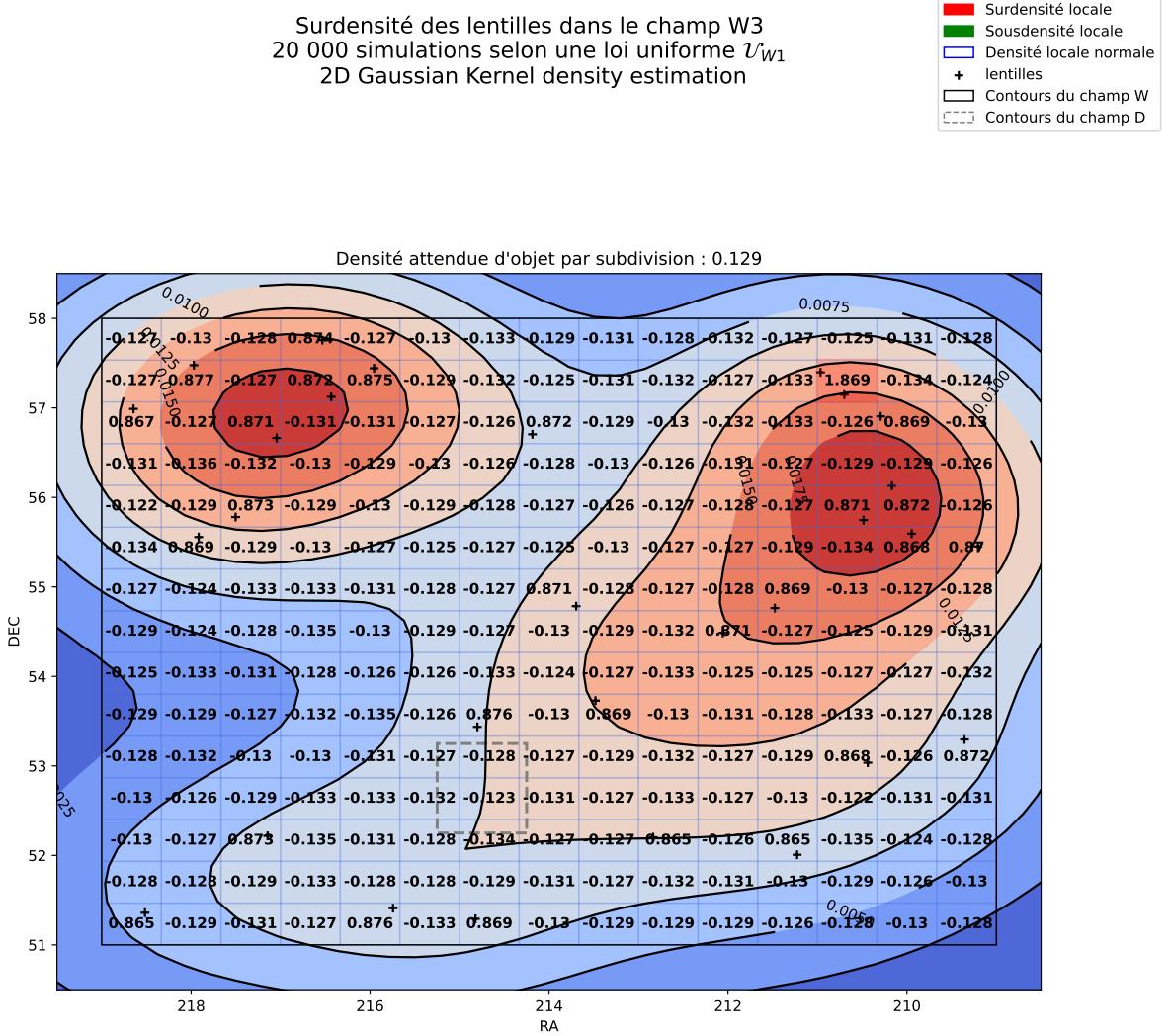


Fig. B.9 – Graphique des pics de surdensité et sous-densité locale de lentilles pour le champ W3. La valeur numérique inscrite au centre de chaque case correspond à l'écart densité réelle de lentille – la densité obtenue par moyennage de tirages aléatoires selon une loi uniforme \mathcal{U}_{W3} .

Source : [Hud+12]

B.3.3 Fonction de corrélation à 2 points : Nous commençons par donner l'algorithme mentionné dans la section 3.2 pour l'estimation de l'erreur sur l'estimateur ζ_{LS} .

```

1 let X: l'échantillon
2 let  $\zeta_{LS}$ : le paramètre à estimer
3 for  $b \in \{1, \dots, B\}$ :
4     sélectionner un échantillon bootstrap  $x_b^*$ 
5     estimer  $\zeta_b^*$  sur l'échantillon  $x_b^*$ 
6

```

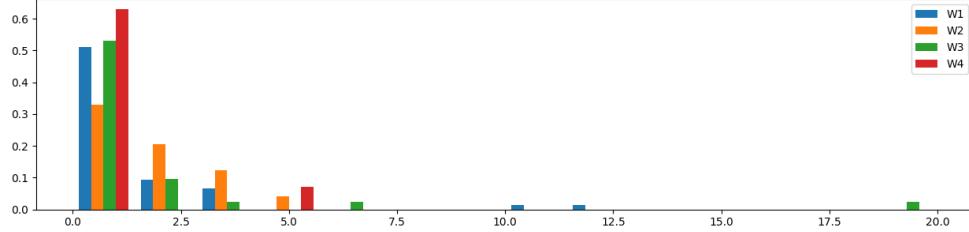
$$7 \text{ Calculer } \widehat{\sigma_{\zeta_{LS}}} = \frac{1}{B-1} \sum_{b=1}^B \left(\zeta_b^* - \frac{1}{B} \sum_{b=1}^B \zeta_b^* \right)^2$$

Listing B.10 – Algorithme de type Monte-Carlo pour l'estimation de l'erreur sur ζ_{LS} .

B.3.4 Étude des galaxies voisines : Enfin, nous concluons avec l'étude du nombre de galaxies voisines des lentilles.

Les galaxies qui composent les lentilles sont retirées.

Histogramme du nombre de galaxies voisines dans un rayon de $3R_E$ pour les champs W1-4.



Histogramme du nombre de galaxies voisines dans un rayon de $4R_E$ pour les champs W1-4.

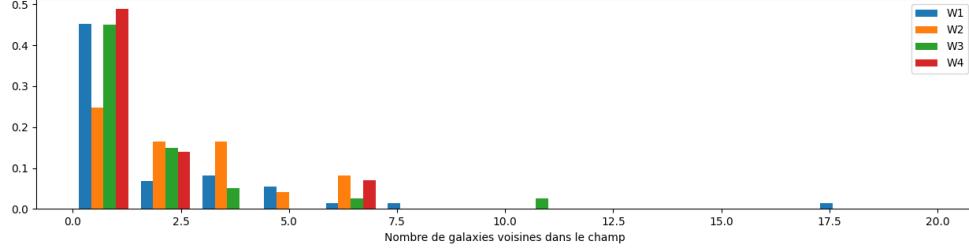


Fig. B.11 – Nombre de galaxies voisines dans un rayon de 3 et $4R_E$ autour de la lentille.

Source : [Hud+12]

Champ	Nb de lentilles	Nb de galaxies	Taille (deg ²)	Densité de galaxie/deg ²
W1	60	4 613 209	72	64072
W2	17	1 823 475	25	72939
W3	31	3 199 953	49	65305
W4	10	1 765 450	25	70618

TABLE B.12 – Tableau du nombre de lentille par champ ainsi que de la densité de galaxie par degré carré dans chaque champ. Tableau à mettre en parallèle avec le Tableau 3.1 pour la densité de lentille par degré carré.

Source : [Hud+12]

Annexe C

Bilan personnel de l'expérience et des compétences acquises

Connaissances statistiques, informatiques et cosmologiques

Au cours de ce stage, j'ai acquis des connaissances sur de nombreux sujets, mais tout particulièrement en astrophysique, domaine dans lequel mes quelques notions sur le big bang, l'expansion de l'univers ou la matière noire étaient très hasardeuses. En deux mois, sans être devenu un spécialiste de l'astrophysique moderne, mes connaissances se sont largement développées. J'ai désormais une compréhension comme fiable des bases et enjeux des différentes théories de l'expansion de l'univers, bien que la partie mathématique de ces théories n'ait pas été abordée en détails. Les nombreuses explications de mon maître de stage Rémi Cabanac, toutes aussi précises que complètes, m'ont permis de comprendre les enjeux cosmologiques des études statistiques que je menais. Ainsi, en plus d'avoir su m'éclairer tout au long de ce stage, il a su répondre à mes nombreuses interrogations théoriques, et je lui en suis encore une fois reconnaissant.

D'autre part, la réalisation de ce stage m'a permis de développer mes compétences informatiques en programmation sous python. J'étais déjà à l'aise en programmation de type algorithmique sous python, mais j'étais finalement assez novice en terme de programmation statistique sous python. Grâce à ses deux mois de stage, je me sens dorénavant serein lorsque j'utilise python pour réaliser des analyses statistiques. J'ai également eu à coder des algorithmes de manipulation de données, ce qui m'a permis de maintenir mon niveau en algorithmique. De plus, j'ai eu l'occasion de découvrir les packages python Astropy et AstroML, orientés pour les techniques de programmation pour l'astrophysique. Je suis très heureux d'avoir eu cette opportunité qui enrichit d'avantage mes compétences sous python.

Mon expérience

Je ne vais pas m'attarder sur la partie relations humaines que j'ai pu expérimenter au cours de ce stage, pour la simple et bonne raison que mes interactions avec d'autres personnes que mon maître de stage au cours de ces deux mois sont au nombre de zéro. En effet, au cours du moins de Juillet, le laboratoire était assez calme et, depuis mon bureau isolé au premier étage, je n'ai pas eu l'occasion de développer une quelconque interaction humaine. Au mois d'Août, j'ai passé deux semaines en télétravail pour des raisons de logistique personnelle, et au cours des deux semaines suivantes, seuls mon tuteur et moins même n'étions pas en vacances dans le laboratoire.

Cependant, si l'expérience humaine n'est pas des plus épanouissante, l'expérience intellectuelle se révèle être à la hauteur de mes attentes. Le stage de recherche m'a invité à développer mon sens de l'innovation et du travail en autonomie. En effet, de part la nature très particulière des travaux de recherche, nous ne savons pas toujours quoi faire ni comment faire les choses. Nous sortons souvent des sentiers battus et devons faire preuve d'ingéniosité et d'imagination. Ce fut le cas au cours du stage où j'ai été amené à réfléchir et à proposer des solutions par moi-même puis d'en discuter avec mon tuteur. C'est ainsi que j'ai lu de très nombreuses publications scientifiques pour m'inspirer de ce qui se fait dans le milieu et ainsi proposer des solutions les plus adaptées à notre problématique. La liberté intellectuelle qui m'a été proposée au cours de ce stage a été une grande richesse car je ne me suis pas contenté de suivre des directives et d'appliquer des méthodes toutes tracées, mais j'ai du imaginer le cheminement logique de mes travaux. Je ne doute pas une seule seconde que cette expérience me sera très utile pour la fin de mes études et pour mes futures expériences professionnelles.

Enfin, je tiens à insister sur la chance que j'ai eu d'effectuer ce stage. Tout d'abord, pour des raisons pratiques puisque ce stage, mon second de l'été après l'arrêt brutal de mon premier stage, a été trouvé en moins d'une semaine et m'a permis de valider mon stage obligatoire de deuxième année à l'ENSAI. D'autre part, le travail effectué était mon premier travail de recherche. En effet, mes différents travaux effectués jusqu'à présent étaient plus scolaires. J'ai ici appris à réfléchir par moi même aux solutions potentielles d'un problème dont je ne comprenais pas tout les tenants et aboutissants du point de vue de la physique théorique. Je suis enchanté d'avoir eu cette expérience. Je suis également ravi d'avoir eu l'occasion de visiter l'observatoire du Pic Du Midi, ses télescopes et divers appareils, ainsi que les "coulisses" et les salles de contrôle avec mon tuteur Rémi Cabanac.

A Rémi, encore une fois, merci pour tout.