

Observatoire de Paris-Meudon, Institut d'Astrophysique de Paris

---

# Thesis

---

Toward a new level of modeling of environmental effects on galaxies

---



Manuel DUARTE

---

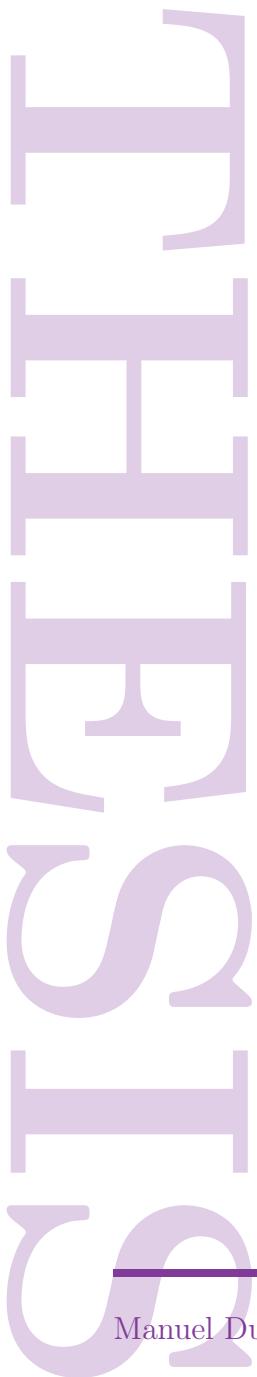
# THE TEST CASE

---

# Contents

<b>I Groupes dans le SDSS</b>	<b>5</b>
<b>1 Introduction . . . . .</b>	<b>7</b>
<b>2 Group Finder Algorithm . . . . .</b>	<b>9</b>
2.1 Introduction . . . . .	10
2.2 Problématique . . . . .	10
2.2.1 Contexte et objectifs . . . . .	10
2.2.2 Plan de travail . . . . .	11
2.3 Création du mock catalogue . . . . .	12
2.3.1 Pourquoi un mock catalogue? . . . . .	12
2.3.2 Obtention des données . . . . .	12
2.3.3 Réalisation du mock catalogue . . . . .	15
2.3.4 Limitations du mock catalogue . . . . .	22
2.4 L'algorithme de Yang <i>et al.</i> . . . . .	23
2.4.1 Description . . . . .	24
2.4.2 Algorithme de Friends-of-Friends . . . . .	27
2.4.3 Réalisation du Yang <i>et al.</i> . . . . .	30
2.4.4 Résultats et comparaison . . . . .	43
2.5 Notre algorithme . . . . .	46
2.5.1 Description . . . . .	46
2.5.2 Réalisation . . . . .	49
2.6 Conclusion . . . . .	66
2.7 Determine the LF . . . . .	70
2.7.1 Estimating parameters . . . . .	71
2.7.2 Tests on mock catalogues . . . . .	73
<b>3 Analyse du SDSS-DR8 . . . . .</b>	<b>83</b>
<b>4 Morphologies dans le SDSS . . . . .</b>	<b>85</b>
<b>5 Approche analytique . . . . .</b>	<b>87</b>
<b>6 Modélisations numériques . . . . .</b>	<b>89</b>

<b>A Analysing data on the SDSS-DR8 . . . . .</b>	<b>91</b>
A.1 Introduction . . . . .	91
A.2 Analysis . . . . .	91
A.2.1 Definitions . . . . .	91
A.2.2 Galaxies selection . . . . .	93
A.2.3 Fibre collision estimation . . . . .	95
<b>B How to generate mock catalogues? . . . . .</b>	<b>101</b>
<b>C Calcul du barycentre lumineux des groupes . . . . .</b>	<b>103</b>
<b>D Calcul de la fonction <math>\Gamma</math> incomplète . . . . .</b>	<b>105</b>



# Part I

## Groupes dans le SDSS

# THE TESTIMONY OF JESUS

# Chapter 1

## Introduction



THE  
TEST  
CASE

---

# Chapter 2

## Group Finder Algorithm

– " Le plus simple serait que tout se fasse le plus simplement possible "  
*Inconnu*

### Sommaire

<b>2.1</b>	<b>Introduction</b>	<b>10</b>
<b>2.2</b>	<b>Problématique</b>	<b>10</b>
2.2.1	Contexte et objectifs	10
2.2.2	Plan de travail	11
<b>2.3</b>	<b>Création du mock catalogue</b>	<b>12</b>
2.3.1	Pourquoi un mock catalogue?	12
2.3.2	Obtention des données	12
2.3.3	Réalisation du mock catalogue	15
2.3.4	Limitations du mock catalogue	22
<b>2.4</b>	<b>L'algorithme de Yang <i>et al.</i></b>	<b>23</b>
2.4.1	Description	24
2.4.2	Algorithme de Friends-of-Friends	27
2.4.3	Réalisation du Yang <i>et al.</i>	30
2.4.4	Résultats et comparaison	43
<b>2.5</b>	<b>Notre algorithme</b>	<b>46</b>
2.5.1	Description	46
2.5.2	Réalisation	49
<b>2.6</b>	<b>Conclusion</b>	<b>66</b>
<b>2.7</b>	<b>Determine the LF</b>	<b>70</b>
2.7.1	Estimating parameters	71
2.7.2	Tests on mock catalogues	73

## 2.1 Introduction

Les grands relevés du ciel sont utilisés de manière courante dans la recherche en astronomie et prennent de plus en plus de place dans la façon de travailler des chercheurs, surtout sur l'étude des structures des galaxies aux grandes échelles dans l'Univers. Un des plus fameux "surveys", est le SDSS (Sloan Digital Sky Survey) qui permet d'avoir accès aux spectres et aux mesures photométriques de presque un million de galaxies sur la voûte céleste. Ces données sont mises à jour annuellement et ne cessent donc de s'enrichir. Cette abondance de données a abouti à de nombreuses études qui ont permis de mesurer l'évolution des propriétés des galaxies (comme la couleur, le contenu en gaz, la morphologie, l'âge, l'abondance en métaux...) avec leur *masse en étoiles*, leurs *environnements global* (avec leur groupe parent) et *local* (distance au centre du groupe parent). Ces propriétés découlent de processus physiques qui dépendent fortement de la quantification des environnements des galaxies. Il est donc important de pouvoir réaliser cette quantification de manière optimale afin de comprendre au mieux ces processus physiques. C'est alors le but de ce stage: créer un programme qui permet de regrouper au mieux les galaxies du SDSS.

Mais ce regroupement n'est pas des plus simples à réaliser. Beaucoup d'effets sont à prendre en compte afin de quantifier les environnements et vont faire l'objet de la description du travail réalisé pendant le stage.

## 2.2 Problématique

### 2.2.1 Contexte et objectifs

Nous avons vu qu'il est nécessaire de regrouper de façon optimale les galaxies du relevé SDSS afin de comprendre quels sont les processus physiques qui permettent de donner les propriétés observées des galaxies. Mais pour aboutir à ce regroupement, les moyens utilisés peuvent être biaisés de plusieurs façons. Par exemple, si l'on souhaite obtenir une quantification à trois dimensions de l'environnement, la distance d'une galaxie à l'observateur  $D$  est évaluée à l'aide de son redshift  $z$

$$D = \frac{cz}{H_0} = D_{\text{vraie}} + \frac{v_p}{H_0} \quad (2.2.1)$$

On voit donc que la vitesse particulière  $v_p$  d'une galaxie nous donne une incertitude sur la distance qui est d'autant plus importante que cette vitesse est grande en norme sur la ligne de visée. On remarque cet effet sur la figure (2.1).

Les catalogues de groupe sur le SDSS qui sont utilisés le plus souvent sont eux aussi biaisés: les galaxies d'avant et arrière plan viennent contaminer les groupes et ceci est d'autant plus visible aux grands rayons projetés par rapport au centre du groupe, et il existe une ségrégation en masse des groupes et des galaxies où les groupes de faible masse et les galaxies naines sont manquantes.

## 2.2. PROBLÉMATIQUE

## CHAPTER 2. GROUP FINDER ALGORITHM

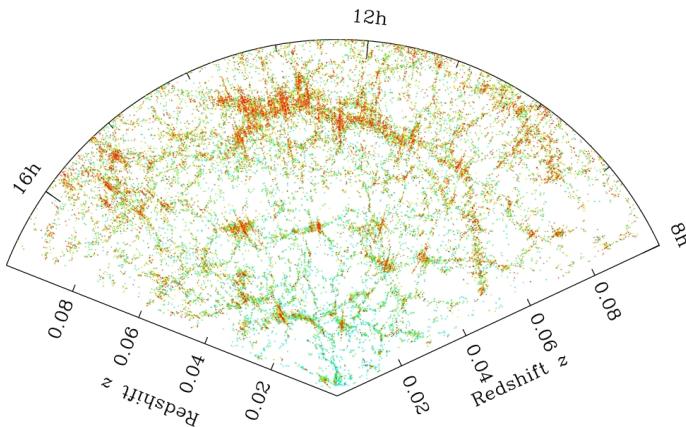


Figure 2.1: Effets de l'allongement des groupes visible sur le relevé du SDSS (la couleur sur le graphe traduit celle de la magnitude  $B - V$ ). (source: <http://www.planetastronomy.com>)

La technique habituellement utilisée pour déterminer les groupes est la méthode de la percolation appelée aussi Friends-of-Friends (FoF) dans l'espace des redshifts qui regroupe toutes les galaxies qui ont des voisines communes (voir la section (2.4.2)). Cette technique a également certains désavantages. C'est pourquoi durant le stage, des moyens différents ont été mis en œuvre afin de pouvoir aboutir au premier catalogue sur l'ensemble du SDSS-DR7 et qui sera optimal pour gérer les effets de projection et sélectionner les galaxies dans la sphère viruelle. L'idée générale est de remplacer la "simple" méthode de FoF par une sélection de surdensité dans l'espace projeté et corrigée par un filtrage des vitesses à l'aide de la méthode du maximum de vraisemblance, en prenant en compte la modélisation des effets de projection faite par Mamon et al. [1] sur des simulations cosmologiques. Ensuite une fois l'algorithme mis en place, le tester sur des simulations cosmologiques afin d'optimiser quelques paramètres de l'algorithme et finalement l'appliquer sur le SDSS.

Pour obtenir tous ces résultats, le travail réalisé s'est construit en plusieurs étapes qui vont être décrites dans ce rapport.

### 2.2.2 Plan de travail

Le but final étant de pouvoir réaliser un catalogue des groupes de galaxies et ceci de façon optimale, il est nécessaire de pouvoir réaliser quelques tests préliminaires du code réalisé pour vérifier qu'il est capable de retrouver des résultats déjà établis. C'est pourquoi il a fallu tester notre travail sur des simulations cosmologiques. Le choix s'est porté tout d'abord sur la simulation Millennium II. Une description brève de cette simulation sera faite afin d'expliquer certains choix effectués durant le travail. Les données issues de cette simulation ont été traitées et analysées dans le but de simplifier le travail qui devra être fait par la suite grâce au programme qui a été créé. Une fois les données de la simulation récoltées, un mock catalogue a été créé pour pouvoir tester par la suite le programme. Un mock catalogue est en quelque sorte une voûte céleste fabriquée de toute pièce à partir des données d'une simulation cosmologique. Pour être un bon mock catalogue, il faut donc retrouver les propriétés générales des galaxies enregistrées dans les surveys, tout en gardant également le plus possible les structures aux grandes échelles que forment les galaxies et les groupes de galaxies. D'une manière générale, on peut dire qu'il faut pouvoir ne pas perdre les informations contenues dans la simulation, et qu'il ne faut pas en introduire dans le mock catalogue par des biais causés par la méthode de création de ce mock (qui sera détaillée dans la suite du rapport).

Une fois ce "faux-ciel" créé, le programme de regroupement a été construit en plusieurs étapes lui aussi. Tout d'abord, nous avons souhaité mettre en place un algorithme similaire à Yang et al. [2]

pour pour pouvoir réaliser des comparaisons entre les programmes. Ensuite, notre propre algorithme a été développé.

Puis nos algorithmes ont été appliqués sur le mock catalogue qui a été créé, ceci afin d'optimiser les quelques paramètres des programmes qui ont besoin de l'être et de tester nos programmes.

Mais même ainsi d'autres effets, dus aux limitations mêmes du survey, sont à prendre en compte comme les limites en luminosité et les effets de bord des surveys (qui tronquent certains amas). Des moyens pour limiter ces effets et les prendre en compte dans le programme ont été aussi mis en œuvre pour pouvoir appliquer par la suite notre algorithme sur le SDSS-DR7 en entier, et alors obtenir les premiers résultats et les exploiter.

Dans la suite du rapport sera donc exposé en détail le travail effectué durant le stage pour aboutir aux résultats qui seront finalement décrits.

## 2.3 Création du mock catalogue

Dans cette section va maintenant être décrit le cheminement pour arriver à créer un mock catalogue qui va permettre de tester les différents programmes. Il a été construit à partir des données de la simulation Millennium-II et de façon plus précise du catalogue Guo2010a (Guo et al. [3]).

### 2.3.1 Pourquoi un mock catalogue?

Les programmes qui ont été réalisés durant le stage doivent pouvoir être testés avant de les lancer directement sur les données du SDSS DR7. C'est pourquoi le mock catalogue qui a été réalisé est d'une grande aide. En effet certains paramètres des programmes ont besoin d'être ajustés afin de regrouper au mieux les galaxies, et pour cela il est nécessaire de voir son "comportement" sur des données dont on connaît la structure à l'avance, c'est-à-dire avoir à disposition un ensemble de données contenant des positions de galaxies avec d'autres de leurs caractéristiques (comme leur vitesse, leur contenu en masse stellaire, leur magnitude absolue, etc...) et surtout le groupe auquel elles appartiennent pour pouvoir faire une comparaison avec les résultats donnés par notre programme. Il s'avère que les simulations cosmologiques permettent d'obtenir de telles informations qui sont ensuite mises à disposition du "public". C'est ce qui a été fait dans le but de créer le mock catalogue avec la simulation du Millennium-II.

### 2.3.2 Obtention des données

Pour le travail qui est à faire à partir des données du Millennium-II (MSII), le choix qui s'offre à nous est beaucoup trop grand. Nous n'avons pas intérêt à utiliser l'ensemble des informations qui sont disponibles, c'est pourquoi nous avons dû sélectionner les données et les analyser afin de se limiter au strict nécessaire en terme de volume de données. Ces choix vont être exposés dans cette sous-section.

#### La simulation Millennium-II

La simulation Millennium-II est une simulation à  $N$  corps de matière noire dans le cadre d'une cosmologie  $\Lambda$ CDM (Cold Dark Matter) Boylan-Kolchin et al. [4]. Elle permet d'étudier la formation et l'évolution des structures cosmiques de l'Univers à partir des résultats obtenus avec la simulation. Celle-ci hérite des caractéristiques de son "ancêtre" la simulation Millennium (MS), c'est-à-dire qu'elle possède les mêmes paramètres cosmologiques, ainsi que le même nombre de particules de matière noire. Les différences entre les deux simulations sont que la MSII est réalisée dans une boîte de simulation de côté cinq fois plus petite que dans le cas de MS avec une résolution spatiale cinq fois meilleure. Il se trouve aussi que la résolution en masse des particules est 125 fois meilleure elle aussi

### 2.3. CRÉATION DU MOCK CATALOGUE

### CHAPTER 2. GROUP FINDER ALGORITHM

que dans le cas du MS. Dans les deux cas, des conditions aux limites périodiques sont appliquées aux bords du cube. Tout ceci tend à nous inciter à l'utilisation des données obtenues à partir de cette simulation. Les paramètres de la simulation sont résumés dans la table (2.1).

$\Omega_{\text{tot}}$	$\Omega_m$	$\Omega_b$	$\Omega_\Lambda$	$h$	$\sigma_8$	$n_s$
1.0	0.25	0.045	0.75	0.73	0.9	1

Table 2.1: Les paramètres cosmologiques utilisés pour la simulation Millennium.

Lorsque cette simulation est réalisée, les données (positions, vitesses des particules...) sont régulièrement sauvegardées à différents "snapshots" ou redshifts, c'est-à-dire à des époques dans la simulation non séparées à des intervalles de temps réguliers. Pour chacun de ces snapshots on peut appliquer des modèles de formation des galaxies, qui nous seront utiles pour réaliser le mock. Sur ces snapshots sont appliqués des algorithmes de FoF pour déterminer les halos formés par les particules de matière noire dans la simulation, puis d'autres algorithmes sont également appliqués afin de raffiner les résultats du FoF et de détecter les sous-halos parmi les halos trouvés. À partir de cela et des caractéristiques de ces halos, on peut appliquer les modèles de population des galaxies. À la suite de ce traitement on a donc des données sur des galaxies avec leurs différentes propriétés ainsi que leur appartenance vraie à un halo de la simulation, résultant de la manière dont ces galaxies ont été obtenues.

La simulation de la population galactique que nous avons choisie pour réaliser le mock catalogue est celle du catalogue Guo2010a. Ce modèle était de l'art permet d'avoir accès à une grande gamme de type de galaxies, la résolution en masse étant assez faible, et donc d'avoir des galaxies naines dans ce modèle. Ces données sont elles aussi disponibles sur le site du MSII référencées sous forme de table dans le catalogue Guo2010a, avec les informations sur les halos dans un autre catalogue.

Dans le cadre de notre travail sur le mock, les données ont été filtrées de façon à obtenir que les informations sur les halos et les galaxies correspondant au dernier snapshot de la simulation à un redshift nul, c'est-à-dire correspondant à notre époque actuelle, en vue de notre future application du programme sur le SDSS DR7. Mais avant une quelconque utilisation des données à notre disposition, il est nécessaire de les analyser afin de voir quelles seront les limites de notre mock catalogue et de ne garder que les informations les plus utiles parmi le large choix offert par cette simulation et le modèle de population galactique choisi.

#### Analyse des données du catalogue Guo2010a

La première constatation faîte lors de l'analyse des données du catalogue Guo2010a est le nombre important des informations disponibles pour chaque galaxie et le nombre de ces galaxies (plus de 13 millions...). Ceci n'est pas sans problème pour la réalisation du mock catalogue, comme nos moyens informatiques en terme de mémoire sont limités et que le mock catalogue nécessite un certain nombre de réPLICATION de ces cubes de simulation, et donc un volume de données d'autant plus grand. Il a donc fallu trouver un moyen de réduire la taille de ces informations tout en évitant les pertes d'informations importantes.

Nom	Millennium II	Mini MSII
$L_{box}$ ( $h^{-1}\text{Mpc}$ )	100	100
$N_p$	10 077 696 000	80 621 568
$m_p$ ( $h^{-1}M_\odot$ )	$6.89 \times 10^6$	$8.61 \times 10^8$
Nombre de galaxies du Guo2010a	13 191 859	186 620

Table 2.2: Les différentes caractéristiques des simulations du Millennium.

Une autre simulation du MSII a été réalisée avec un nombre de particules plus faibles et une résolution en masse plus faible pour les particules de matière noire: il s'agit du mini-MSII. Elle conserve la même taille de côté pour le cube de simulation. Les différentes propriétés utiles des deux simulations sont résumées dans la table (2.2).

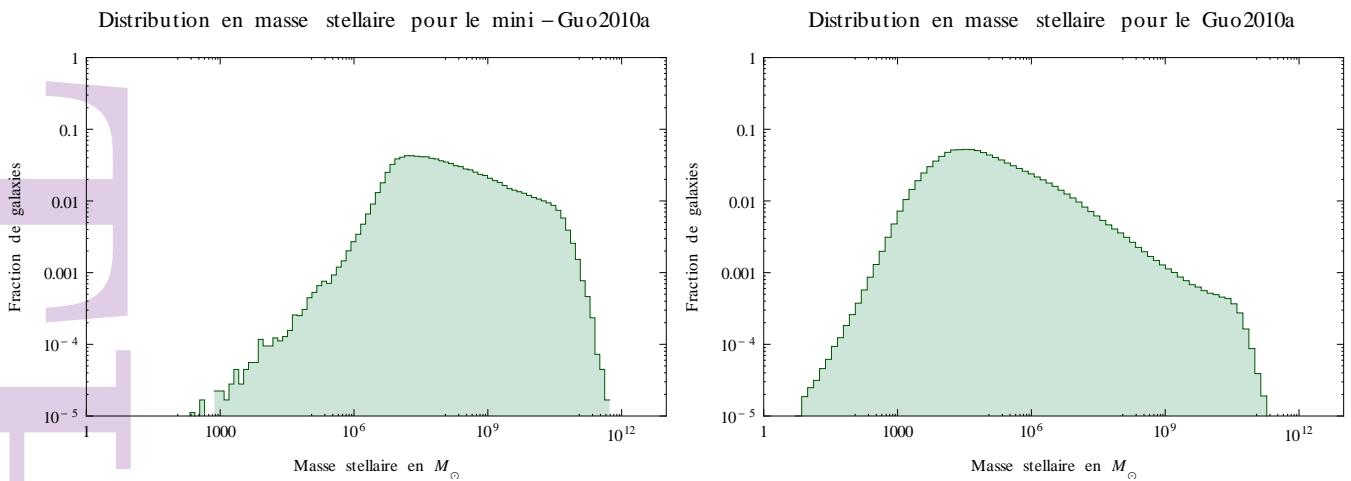


Figure 2.2: Distribution des masses stellaires pour les deux catalogues Guo2010a obtenus à partir du MSII et du mini-MSII.

Le fait que les deux simulations aient la même taille de boîte ( $L_{box}$ ) et un nombre de particules de matière noire beaucoup plus faible dans le cas du mini-MSII est un avantage pour notre mock car le nombre de données à prendre en compte est beaucoup plus faible et facilement gérable avec nos moyens informatiques. Cependant le fait que la résolution en masse des particules de matière noire soit plus faible est plus dérangeant car le modèle de population stellaire du catalogue Guo2010a appliqué dans le cas du mini-MSII donne alors lui aussi une résolution plus faible en terme de masse des galaxies. Celles qui ont une masse faible nous échappent alors. Ceci est visible lorsque l'on réalise la distribution de masse stellaire des catalogues Guo2010a dans le cas du MSII et du mini-MSII. Ces distributions sont visibles sur la figure (2.2). On voit bien que sur la distribution des masses dans le cas du MSII on a un "pic" à une masse stellaire presque 100 fois plus faible que dans le mini-MSII. En prenant donc le mini-MSII à la place du MSII on perd alors l'information dans notre mock sur les galaxies de faible masse.

Maintenant que nous avons fait le choix d'utiliser le catalogue Guo2010a du mini-MSII pour restreindre le volume de données à traiter pour notre mock catalogue, il faut encore essayer de diminuer le nombre d'informations qui est toujours un peu trop important. Quand on regarde la distribution des masses stellaires dans le cas du mini-MSII, on voit que les galaxies de masse stellaire inférieure au pic de la distribution représente une faible part de la population des galaxies. Pourquoi

### 2.3. CRÉATION DU MOCK CATALOGUE

### CHAPTER 2. GROUP FINDER ALGORITHM

alors ne pas supprimer de notre mock ces galaxies de faible masse? En regardant la part des galaxies qui ont une masse supérieure à une certaine masse stellaire, on remarque que cela permettrait de gagner en terme de volume de données tout en ne perdant pas trop d'informations dans notre mock qui est déjà biaisé par le manque de galaxies de faible masse par le choix de l'utilisation du mini-MSII. Cette répartition est visible sur la figure (2.3).

On voit alors qu'en faisant le choix de ne prendre en compte que les galaxies d'une masse supérieure à  $10^8 M_{\odot}$  on réduit fortement le nombre de galaxies à traiter dans le mock tout en ne perdant pas trop d'informations des catalogues car le choix du mini-MSII nous limite déjà en terme de résolution en masse des galaxies. On peut désormais réaliser le mock catalogue à partir de ces données avec la méthode qui est décrite dans la partie suivante.

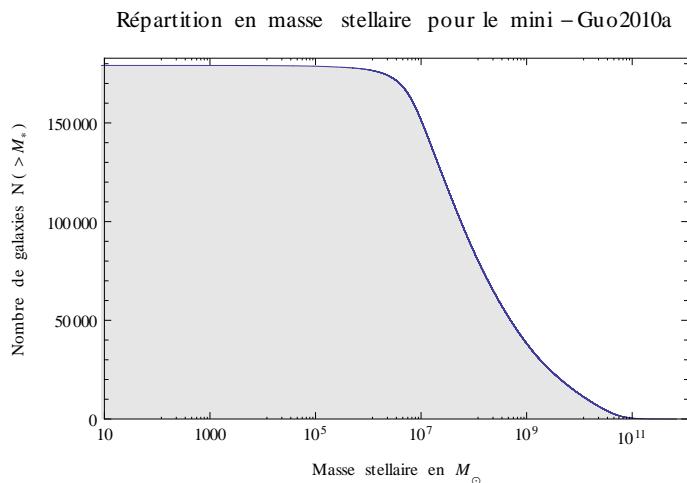


Figure 2.3: Fractions des galaxies du mini-MSII qui ont une masse stellaire supérieure à celle indiquée en abscisse.

#### 2.3.3 Réalisation du mock catalogue

La création du mock catalogue à partir des données de la simulation demande quelques pré-traitements pour éviter certains effets causés par sa formation à partir de données de simulation et en même temps pour pouvoir conserver les informations qui nous seront utiles plus tard pour tester notre programme. En effet pour simuler au mieux les données du SDSS certaines des propriétés de notre mock catalogue doivent correspondre à celles du SDSS comme la magnitude limite par exemple. Ces méthodes vont faire l'objet des sections suivantes.

##### Informations à conserver pour le mock catalogue

Avant de voir comment le mock catalogue a été réalisé, il faut connaître les données fournies par le MSII et qui vont nous permettre de simuler des observations.

Pour tester le programme de regroupement, il nous faut comparer les halos trouvés par notre algorithme avec ceux de la simulation du MSII. Deux catalogues sont disponibles à partir de cette simulation: un catalogue contenant des informations sur les halos de matière noire comme leurs positions, leurs vitesses, etc..., ainsi qu'un catalogue sur les galaxies qui peuplent ces halos avec de nombreuses caractéristiques. Le nombre de données étant assez élevé dans les deux catalogues, une simple comparaison une à une des données entre les deux pour associer à chaque galaxie les caractéristiques du halo auquel elle appartient dans la simulation donnerait un temps de calcul

prohibitif. On a donc mis en place un algorithme pour réaliser un matching plus rapide pour faire cette correspondance (non détaillé).

Nous avons désormais les données utiles pour la réalisation du mock catalogue, car nous avons la correspondance entre les données des halos et des galaxies.

### Méthodes de création du mock catalogue

Pour notre mock, nous avons besoin de données pouvant simuler des observations réalisées sur un intervalle de redshift compatible avec celui du relevé qu'on simule, c'est à dire dans notre cas le SDSS. On doit alors atteindre une profondeur en redshift de près de  $z = 0.3$ . Cependant on l'a vu la taille de la boîte de la simulation du MSII est de  $100 h^{-1}\text{Mpc}$  ce qui équivaut à une profondeur en redshift de  $z \sim 0.03$ , bien inférieure à la valeur nécessaire pour simuler le SDSS. C'est pourquoi pour créer le mock catalogue nous avons répliquer la boîte de simulation pour atteindre la profondeur souhaitée.

Mais pour cela il faut appliquer quelques transformations à cette boîte afin de limiter certains biais qui pourraient être introduits par cette réPLICATION.

#### RéPLICATION du cube de simulation

La façon de procéder pour aboutir à notre mock catalogue est assez simple sur le principe: on va juxtaposer des cubes de la simulation du MSII (comme celui de la figure (2.4)) les uns aux autres afin d'atteindre la profondeur souhaitée pour correspondre aux données du SDSS. En se contentant

Figure 2.4: Le cube du mini-MSII avec les galaxies du Guo2010a. La couleur d'une galaxie dépend du halo auquel elle est associé (à chaque couleur correspond un halo).

de faire ce travail, en translatant les positions des galaxies pour chaque cube que l'on accolé aux précédents, on introduit des biais dans le mock. En effet si on superpose ainsi les cubes les uns aux autres, on va répliquer une même image du ciel mais à différentes distances. Du coup, si on se place à un endroit dans le cube de notre mock et que l'on observe dans une direction du cube (comme dans le cas d'un télescope observant une portion du ciel), on remarque un effet de perspective dans le ciel causé par cette réPLICATION comme dans la figure (2.5).

Pour pallier à ce problème, une technique consiste à faire subir des transformations supplémentaires à chaque cube de manière aléatoire pour éviter les répliques parfaitement identiques comme décrit dans Blaizot et al. [5]. Pour cela on va effectuer trois transformations différentes sur les coordonnées du cube, en plus des simples translations suivant les axes ( $X, Y, Z$ ), pour pouvoir accoler les cubes les uns aux autres.

① La première transformation est une rotation d'un angle multiple du demi-entier de  $\pi$  autour de chacun des trois axes du cube, chaque angle étant éventuellement différent. Le choix du multiple pour l'angle de rotation doit se faire aléatoirement afin de ne pas introduire de nouveaux effets de perspective par ces rotations. On choisit un angle aléatoire multiple demi-entier de  $\pi$  et non pas un angle compris dans l'intervalle continu  $[0, 2\pi]$  pour ne pas perdre des informations sur les halos dans le cube de simulation. En choisissant un angle multiple de  $\frac{\pi}{2}$ , on permet à chaque galaxie de se retrouver à nouveau dans le "même" cube de simulation qu'auparavant mais avec des coordonnées différentes. Nous n'avons pas alors à récupérer des galaxies qui sortiraient de cette boîte pour les y remettre à l'aide de conditions périodiques aux limites qui "déstructurerait" complètement notre cube, et les informations sur l'appartenance d'une galaxie à un certain halo deviendraient totalement inutilisables. On peut voir ceci sur la figure (2.6).

### 2.3. CRÉATION DU MOCK CATALOGUE

### CHAPTER 2. GROUP FINDER ALGORITHM

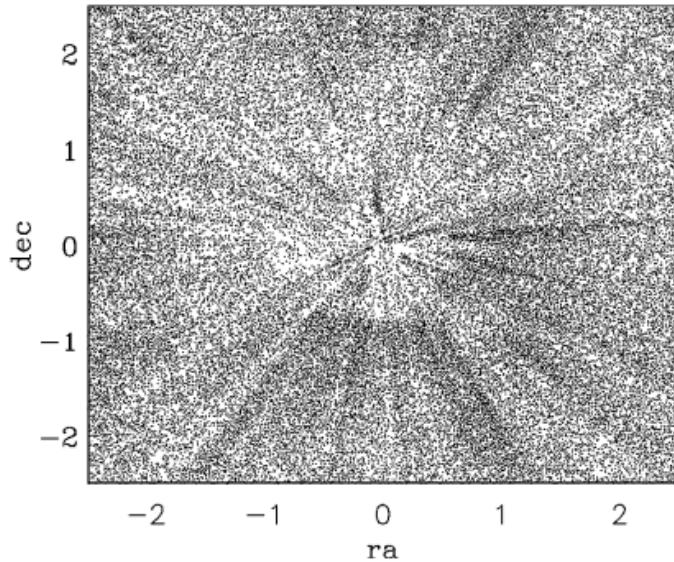


Figure 2.5: Effet de perspective dans un mock réalisé avec une simple superposition de cubes (projection angulaire), les ascensions droites (RA) et les déclinaisons (DEC) sont arbitraires et représentent seulement des directions orthogonales dans la "fausse voûte céleste", d'après Blaizot et al. [5].

② La deuxième transformation est une translation d'une longueur comprise dans l'intervalle continu de  $[0, L_{box}]$ . On pousse les galaxies du cube suivant une direction sur une longueur choisit aléatoirement dans l'intervalle précédent, et celles qui en sortent sont ramenées dans le cube de l'autre côté en appliquant des conditions aux limites périodiques.

③ La troisième et dernière transformation est une inversion choisit aléatoirement suivant l'un des trois axes du cube. Il peut aussi n'y avoir aucune inversion appliquée à l'ensemble des trois axes.

Avant de pouvoir réaliser ces transformations pour faciliter la suite des calculs, on décale l'origine de la boîte de simulation du MSII qui est située sur un coin pour la placer au centre. Ensuite on peut appliquer les transformations aléatoires sur les coordonnées des galaxies du cube et les translations sur les coordonnées pour placer chacun des cubes dans notre mock, tout ceci pouvant être résumé de la façon suivante:

$$\mathbf{X}' = \mathcal{I}_{XYZ} \mathcal{R}_X(\theta_X) \mathcal{R}_Y(\theta_Y) \mathcal{R}_Z(\theta_Z) \mathbf{X} + \mathbf{T} \quad (2.3.1)$$

où  $\mathbf{X}'$  représente les coordonnées  $\mathbf{X}$  suivant les trois axes du cube d'une galaxie une fois transformées,  $\mathcal{R}_i$  la rotation suivant l'axe  $i$  d'un angle multiple de  $\frac{\pi}{2}$  aléatoire  $\theta_i$ ,  $\mathcal{I}$  l'inversion suivant un des trois axes (qui peut ne pas être réalisé) et  $\mathbf{T}$  la translation qui va permettre de placer le cube dans la grande boîte du mock catalogue.

Une fois que ces transformations ont été réalisées sur le cube, celui-ci est juxtaposé aux autres cubes déjà placés en effectuant une translation des coordonnées des galaxies suivant les axes du cube de la longueur nécessaire pour placer correctement la boîte de la simulation parmi les autres. On obtient de cette façon un cube plus gros comme sur la figure (2.7). De cette façon, les effets de perspective ne peuvent plus être visibles sur notre fausse voûte céleste comme on peut le voir sur la figure (2.8) dans le cas d'un mock où cette méthode a été testée. Le mock catalogue ainsi créé et visible sur l'image ne présente plus d'effets de perspective causés par la réplication d'un même cube un certain nombre de fois. Le résultat est tout à fait semblable à ce que l'on pourrait obtenir en prenant directement une simulation réalisée dans une boîte de taille équivalente à la profondeur en redshift que l'on souhaite avoir pour notre mock catalogue.

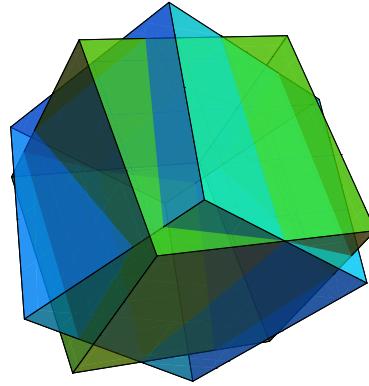


Figure 2.6: Illustration de l'effet d'une rotation quelconque du cube. On voit que l'application des conditions périodiques sur le cube bleu inscriraient des zones vides dans le cube vert ou des zones de sur-densités. On perdrait alors des informations du MSII.

Avec ce simple travail on ne peut pas dire que le mock créé est comparable au SDSS, c'est pourquoi il est nécessaire d'ajouter des caractéristiques et des biais qui vont permettre d'obtenir des données au comportement presque identique à celles du SDSS.

### Ajout de caractéristiques aux données

Une caractéristique importante du SDSS est le fait que l'on est sensible seulement à des galaxies qui ont une magnitude inférieure à une magnitude limite. Pour ajouter cet effet dans notre mock catalogue, il nous faut obtenir les magnitudes apparentes des galaxies de notre mock. Or les magnitudes apparentes sont des caractéristiques qui dépendent de l'observateur, ce qui nous amène à définir la position de cet observateur dans le cube de notre mock catalogue. L'origine des coordonnées de ce cube est située dans un de ses coins et semble être un bon choix pour la position de l'observateur car cela nous permettra de déterminer des coordonnées d'ascension droite et de déclinaison très facilement à partir des coordonnées cartésiennes des galaxies.

Dans notre mock catalogue nous avons à disposition les magnitudes absolues de chacune des galaxies qui découlent du modèle de population galactique appliqué sur la simulation du MSII. On souhaite obtenir à partir de cela les magnitudes apparentes vues depuis l'observateur (situé à l'origine de notre mock). Les effets qui contribuent à donner une certaine magnitude apparente  $m$  d'une galaxie d'une magnitude absolue  $M$  sont la distance à l'observateur ainsi que le décalage spectral du flux de la galaxie causé par l'expansion. L'extinction provoquée par les diverses poussières interstellaires sur la ligne de visée à l'observateur ne sont pas prises en compte dans notre mock étant donné le manque d'information pour appliquer ce type de correction dans les données de la simulation. Tout ceci peut être résumé par l'expression suivante:

$$m_X = M_X + 5 \log_{10} d_{\text{lum}}(\text{pc}) - 5 + K(z, m_X - m_{X'}) \quad (2.3.2)$$

où  $X$  représente une des bandes de longueurs d'onde dans lesquelles sont exprimées les magnitudes absolues du MSII qui correspondent à celles du SDSS ( $u, g, r, i, z$ ),  $K$  est la correction appliquée pour tenir compte du décalage spectral du flux de la galaxie (appelée "correction-K" dans la suite) et  $d_{\text{lum}}$

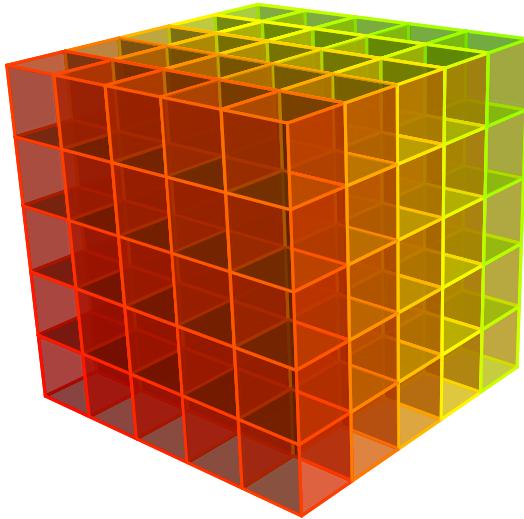


Figure 2.7: Une image du *gros* cube du mock réalisé en juxtaposant les différents cubes engendrés par transformation du cube initial.

la distance lumineuse de la galaxie basée sur la conservation de la loi de diminution de la luminosité en fonction de la distance comme définie dans un Univers euclidien.

Le calcul de  $K$  est basé sur les résultats de Chilingarian et al. [6] qui permet de calculer cette correction à partir d'une expression sous la forme d'un polynôme du redshift et de la couleur pour certaines bandes du SDSS. Les coefficients du polynôme dans les différents cas proposés sont tabulés et cette correction a pu être programmée avec ces coefficients. On remarque aussi un autre problème pour calculer la magnitude apparente avec l'équation (2.3.2): pour la correction-K il est nécessaire de connaître la magnitude apparente déjà dans une autre bande pour calculer la magnitude apparente dans une bande. La méthode qui a été utilisée tout d'abord est le calcul de cette magnitude de façon itérative. Dans un premier temps, on réalise le calcul à l'aide des magnitudes absolues dans la correction-K:

$$m_X^{(1)} = M_X + 5 \log_{10} d_{\text{lum}}(\text{pc}) - 5 + K(z, M_X - M_{X'}) \quad (2.3.3)$$

puis pour les itérations suivantes on injecte dans  $K$  les valeurs des magnitudes apparentes calculées précédemment, et tant que la différence des valeurs des magnitudes entre deux itérations successives  $\Delta m = |m^{(i)} - m^{(i+1)}|$  est supérieure à 0.01:

$$m_X^{(i+1)} = M_X + 5 \log_{10} d_{\text{lum}}(\text{pc}) - 5 + K(z, m_X^{(i)} - m_{X'}^{(i)}) \quad (2.3.4)$$

Le redshift  $z$  de la galaxie du mock est estimé à partir de la distance comobile  $D[h^{-1}\text{Mpc}]$  qui est donnée par les positions comobiles de la galaxie dans le MSII. Pour le calcul de la distance lumineuse, on utilise une approximation analytique donnée par Wickramasinghe and Ukwatta [7] dont la précision est suffisante aux redshifts où on travaille. La distance comobile étant égale à la distance de *mouvement propre* dans le cas d'un Univers plat et connaissant la relation entre distance de mouvement propre et la distance de luminosité, on peut alors en inversant la relation entre distance lumineuse et redshift, déterminer ce dernier à partir de la distance comobile. Pour cela, on détermine pour différents échantillons de  $z$  les valeurs de la distance comobile sur l'intervalle de  $z$  qui nous intéresse, puis en s'aidant d'une interpolation à splines cubiques, on peut connaître le  $z$  correspondant à une distance comobile donnée.

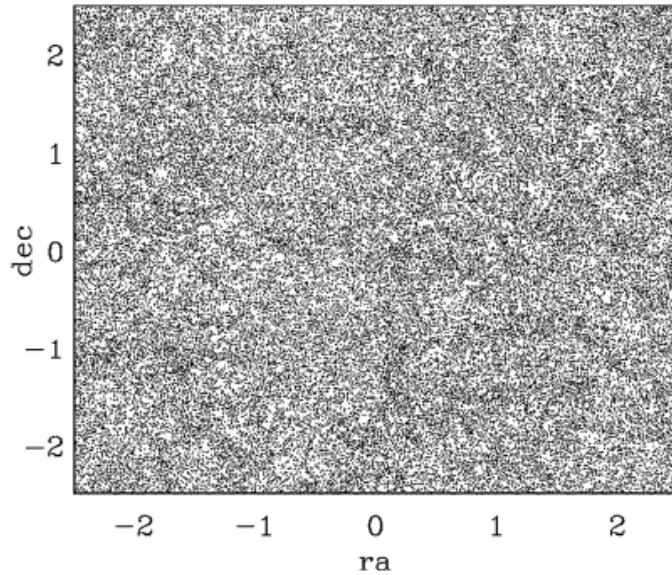


Figure 2.8: Cas d'un mock réalisé avec la méthode décrite dans l'article sur MoMaF: le faux-ciel semble homogène et les effets de perspective causés par la réPLICATION du cube ne sont plus visibles, d'après Blaizot et al. [5].

Pour avoir une autre vérification de nos résultats sur les magnitudes apparentes (car elles ne convergent pas toutes vers une valeur physique par la méthode de l'itération), on a tenté de résoudre le système cubique formé par les magnitudes apparentes avec l'approximation donnée par Chilingarian et al. [6]. Quand on regarde l'expression de la K-correction analytique, elle s'exprime comme:

$$K(z, m_X - m_{X'}) = \sum_{i=0}^{N_i} \sum_{j=0}^{N_j} a_{ij} z^i (m_X - m_{X'})^j \quad (2.3.5)$$

avec  $N_i$  et  $N_j$  qui représentent les ordres des polynômes utilisés pour l'approximation et  $a_{ij}$  les coefficients d'une matrice  $N_i \times N_j$ . Si on connaît le redshift, on remarque que déterminer les magnitudes apparentes revient à résoudre un système polynomial (et non linéaire) d'ordre  $N_j$ . Dans ce cas,  $N_j = 3$  donc le système est d'ordre 3. En utilisant un sous-programme de résolution de systèmes non linéaires de Press et al. [8], on remonte ainsi à chacune des magnitudes apparentes voulues. En faisant la comparaison avec la méthode par itération décrite précédemment, les résultats obtenus sont sensiblement identiques, les différences étant visibles à l'ordre  $10^{-3}$ . Malheureusement, de temps à autre, le calcul des magnitudes apparentes selon les deux méthodes peut ne pas aboutir à une solution physique (divergence vers des valeurs très grandes ou très négatives). On a alors choisi d'adopter une solution simple. La méthode est la suivante: on commence à chercher une solution numérique au système d'équation réalisée uniquement par les magnitudes apparentes en bande  $r$  et  $g$  pour limiter les problèmes de non-convergence plus fréquent quand on cherche à déterminer les solutions pour les cinq bandes  $u, g, r, i, z$ . Si aucune solution n'a été trouvée ou si la solution n'est pas réaliste, on tente à nouveau de déterminer les magnitudes apparentes par des itérations. Si pour les mêmes raisons que précédemment on ne trouve pas de *bonnes* solutions, on se contente de corriger la magnitude absolue du module de distance ( $m_X = M_X + 5 \log_{10} d_{\text{lum}}(\text{pc}) - 5$ ), et la galaxie se voit affecter un *flag* dans le mock catalogue. D'après les flags du mock catalogue, la combinaison des deux méthodes, l'une prenant le relais de l'autre quand elle a échoué, empêche d'aboutir à des galaxies non K-corrigée au final car aucun flag ne signale la présence d'une galaxie seulement corrigée du module de distance.

### 2.3. CRÉATION DU MOCK CATALOGUE

### CHAPTER 2. GROUP FINDER ALGORITHM

On peut alors de cette manière déterminer les magnitudes apparentes de chaque galaxie de notre mock. Pour pouvoir être réaliste, on rajoute alors un filtrage en magnitude dans le mock catalogue, en ne gardant que les galaxies qui ont une magnitude apparente  $m_r < 17.77$  dans la bande  $r$  pour correspondre à la limite en flux du SDSS.

Le mock catalogue doit permettre de simuler des observations sur une fausse voûte céleste, il faut donc attribuer des coordonnées équatoriales à chacune des galaxies. Pour cela on va utiliser les propriétés du cube de notre mock pour calculer ces coordonnées. On va supposer que le plan  $(X, Y)$  du cube est le plan équatorial et l'axe  $Z$  dans le sens positif donne l'axe du pôle céleste (en direction du pôle Nord). Les ascensions droites sont comptées positivement dans le sens indirect pour correspondre aux "véritables" coordonnées. L'origine est toujours l'observateur. Pour le calcul on va partir des coordonnées sphériques pour déterminer les expressions des ascensions droites  $\alpha$  et des déclinaisons  $\delta$ . Les coordonnées cartésiennes  $(X, Y, Z)$  s'écrivent à partir des sphériques  $(r, \phi, \theta)$ :

$$\begin{aligned} X &= r \cos \phi \sin \theta \\ Y &= r \sin \phi \sin \theta \\ Z &= r \cos \theta \end{aligned} \tag{2.3.6}$$

Pour les  $(\alpha, \delta)$ , on a  $\phi = 2\pi - \alpha$  et  $\theta = \frac{\pi}{2} - \delta$ . On peut donc réécrire:

$$\begin{aligned} X &= r \cos \alpha \cos \delta \\ Y &= -r \sin \alpha \cos \delta \\ Z &= r \sin \delta \end{aligned} \tag{2.3.7}$$

On détermine alors les coordonnées en appliquant une dernière transformation à  $\alpha$  pour ramener les valeurs dans l'intervalle  $[0, 2\pi]$ :

$$\begin{aligned} \alpha &= \begin{cases} -\arctan2(Y, X) + 2\pi & \text{si } Y > 0 \\ -\arctan2(Y, X) & \text{sinon} \end{cases} \\ \delta &= \text{signe}(Z) \arccos \left( \frac{\sqrt{X^2 + Y^2}}{\sqrt{X^2 + Y^2 + Z^2}} \right) \end{aligned} \tag{2.3.8}$$

Afin de réduire la taille des données et le temps de calcul de leur traitement, le calcul des magnitudes apparentes ainsi que le filtrage sont réalisés pendant la génération de chaque sous-cube du mock catalogue.

Le redshift que l'on détermine ici pour le calcul de la magnitude apparente est un redshift *vrai* c'est-à-dire le redshift causé seulement par l'expansion de l'Univers et qui traduit la distance à laquelle se trouve la galaxie. Cependant comme on l'a vu dans la section (2.2.1), il y a des effets d'élongation des groupes dans l'espace des redshifts causés par l'accumulation de la vitesse du *Hubble flow* (expansion) et de la vitesse particulière de la galaxie. On doit donc déterminer nous aussi un redshift "observationnel" pour notre mock catalogue afin de simuler correctement le SDSS. Pour cela on doit déterminer la vitesse particulière  $v_{\text{pec}}$  de chaque galaxie. Les composantes suivant les trois axes du cube de simulation sont données dans le Guo2010a. Il suffit alors de calculer la vitesse

projectée sur la ligne de visée de l'observateur pour avoir le redshift. Celui-ci est calculé de la façon suivante:

$$1 + z = \sqrt{\frac{1 + \frac{v}{c}}{1 - \frac{v}{c}}} \quad (2.3.9)$$

avec  $v$  la vitesse algébrique suivant la ligne de visée donc  $v = v_{\text{pec}} + v_{\text{flow}}$  où  $v_{\text{pec}}$  la vitesse particulière projetée de la galaxie et  $v_{\text{flow}} = 100 \times D [h^{-1} \text{Mpc}]$  la vitesse causée par l'expansion avec  $D$  la distance comobile de la galaxie accessible via le Guo2010a.

On rappelle aussi que l'intérêt de créer un mock catalogue est de posséder l'information déjà connue à l'avance de l'appartenance d'une galaxie à un halo. Après avoir réaliser le *matching* entre les galaxies et le halo, on a déterminé la zone où se trouve la galaxie afin de réaliser nos futures comparaisons. Il suffit de calculer pour cela  $r_{\text{3D}}/r_{\text{vir}}$  où  $r_{\text{3D}}$  est la distance au centre du halo de la galaxie et  $r_{\text{vir}}$  le rayon de viriel estimé à partir de la masse de la masse du halo. On en profite pour donner la relation entre la masse du halo et le rayon de viriel et poser une définition. Si on note  $M_{\text{vir}}$  la masse du halo, alors:

$$M_{\text{vir}} = \frac{\Delta}{2} \frac{H_0^2 r_{\text{vir}}^3}{G} \quad (2.3.10)$$

avec  $\Delta$  la valeur de la sur-densité. On définit alors  $r_{\Delta}$  le rayon de viriel à la valeur de la sur-densité  $\Delta$  (par exemple 180). C'est le rayon du halo où la densité est  $\Delta$  fois supérieure à la densité critique  $\rho_c$  avec:

$$\rho_c = \frac{3H_0^2}{8\pi G} \quad (2.3.11)$$

On dispose alors d'un mock catalogue contenant les données qui sont résumées dans la table (2.3).

### Vérification simple

Une vérification simple, pour voir si notre mock catalogue est cohérent avec notre idée de simuler le SDSS, doit être réalisée. D'une façon simple, on s'attend à avoir un rougissement des galaxies quand elles sont éloignées de l'observateur (voir Blaizot et al. [5]), et à cause de la limite en flux que l'on impose au mock catalogue on doit avoir moins de galaxies dans notre mock quand on s'éloigne de l'observateur. C'est ce qui semble être visible sur le mock catalogue quand on regarde la figure (2.9). On observe effectivement moins de galaxies quand le redshift devient plus grand, et un rougissement ( $(g - r)$ ) de plus en plus petit traduit par une couleur qui passe du vert au rouge sur la figure (2.9) vers les grands redshifts. L'autre effet de l'elongation des groupes est lui aussi bien modélisé car on voit des *doigts* pointés vers l'observateur à l'origine.

### 2.3.4 Limitations du mock catalogue

La raison pour laquelle on a créé ce mock catalogue est que l'on souhaite obtenir un faux catalogue possédant au mieux les propriétés du SDSS, tout en ayant à notre disposition les informations d'appartenance des galaxies à un halo de façon certaine, ce qui est fourni par les données du MSII. Pour avoir une profondeur en redshift suffisante, il nous a fallu répliquer la boîte de la simulation du MSII. Sur cette boîte des conditions aux limites périodiques sont appliquées afin de déterminer les halos de particules de matière noire, donc certaines galaxies sur un des bords de la boîte peuvent être regroupées avec des galaxies situées sur une autre face du cube. On voit alors qu'en juxtaposant ces

Numéro	Description
1	Numéro du cube
2	Identité dans le cube
3	Identité du halo
4	Ascension droite
5	Déclinaison
6	Position comobile $X$
7	Position comobile $Y$
8	Position comobile $Z$
9	Vitesse suivant $X$
10	Vitesse suivant $Y$
11	Vitesse suivant $Z$
12	$m_g$
13	$m_r$
14	$M_g$
15	$M_r$
16	Distance comobile
17	Redshift vrai $z$ de la galaxie
18	Redshift observé $z$ de la galaxie
19	$r_{\text{3D}}/r_{\text{vir}}$
...	Autres informations du Guo2010a

Table 2.3: Résumé des données fournies par notre mock catalogue pour chaque galaxie.

boîtes pour créer notre mock, si on cherche à y retrouver des groupes qui sont aux bords des cubes juxtaposés, on fera un lien entre des galaxies de cubes différents car nos transformations aléatoires auront créées ces structures, mais les informations d'appartenance ou non au halo données par le MSII seront faussées dans ce cas par la faute des conditions aux limites. Il faudra alors tenir compte de cet effet quand on appliquera les algorithmes de groupe sur ce mock, et peut être essayer de le quantifier pour corriger nos résultats.

Figure 2.9: Une *tranche* du mock catalogue sélectionnée avec  $\delta > 1.0$  et  $\delta < 1.3 \text{ radians}$  ainsi que  $z < 0.25$  pour les galaxies. On observe bien une évolution de la couleur des couleurs des galaxies ( $g-r$ ) avec le redshift, celles-ci devenant plus rouges quand on s'éloigne de l'observateur. Le mock étant vu sur un coin par l'observateur, les ascensions droites sont comprises entre  $3\pi/2$  et  $2\pi$  ce qui forme un angle droit sur l'image. On remarque aussi les effets d'élargissement qui ont bien été modélisés, le redshift étant celui observé sur la figure.

Il a fallu ensuite réaliser l'algorithme de Yang et al. [2] pour tester un autre algorithme en comparaison de celui que l'on va réaliser, et voir ainsi les différences entre les programmes appliqués sur le mock catalogue et sur le SDSS-DR7.

## 2.4 L'algorithme de Yang et al.

Cet algorithme va tout d'abord être décrit puis le travail réalisé pour l'implémenter également et finalement les résultats qui auront été obtenus seront discutés pour vérifier le bon fonctionnement de

ce programme.

### 2.4.1 Description

L'algorithme de Yang et al. [2] permet de déterminer les groupes de galaxies et d'obtenir certaines caractéristiques de ces groupes comme la masse stellaire par exemple. Le travail décrit dans Yang et al. [2] se base sur les données du SDSS-DR4 mais le principe général peut être également appliqué sur le SDSS-DR7. Les données utilisées sont les redshifts des galaxies lorsqu'ils étaient sûrs ainsi que les magnitudes absolues corrigées de l'extinction et de la correction-K d'une façon qui ne sera pas décrite ici mais que l'on peut voir dans Yang et al. [2]. Dans une première approximation, les masses stellaires sont déterminées à partir des magnitudes absolues selon l'expression suivante:

$$\log_{10} \left( \frac{M_*}{h^{-2} M_\odot} \right) = -0.306 + 1.097(g - r) - 0.1 - 0.4(M_r - 5 \log_{10} h - M_{r,\odot}) \quad (2.4.1)$$

avec  $M_*$  la masse stellaire de la galaxie,  $(g - r)$  la couleur corrigée de la galaxie,  $M_r$  la magnitude absolue K-corrigée en bande  $r$  et  $M_{r,\odot} = 4.64$  la magnitude absolue du Soleil en bande  $r$ .

La détermination des groupes avec cet algorithme s'organise en 5 étapes principales qui vont faire l'objet de ce qui suit.

① Tout d'abord trouver le centre potentiel des groupes. Pour cela, un algorithme de FoF est appliqué sur les galaxies pour déterminer des groupes potentiels, qui représenteraient plutôt la partie centrale des groupes "réels" car le paramètre qui définit le lien ou non entre les galaxies est choisi petit et différent selon la direction de la ligne de visée (espace de redshift) et selon la direction transverse. Il est plus grand dans le cas de la ligne de visée pour prendre les galaxies qui seraient éloignées du groupe par l'effet de l'elongation décrit dans la section (2.2.1). Ensuite le "barycentre lumineux" (voir la section (C)) est calculé pour chaque groupe potentiel et est défini comme le centre du groupe. Toutes les galaxies non liées à un groupe sont gardées et traitées comme des groupes potentiels elles aussi.

② Il faut maintenant déterminer la luminosité caractéristique de ces groupes potentiels. On ne garde dans ce cas là que les galaxies ayant une magnitude absolue en bande  $r$  telle que:

$$M_r \leq -19.5 \quad (2.4.2)$$

Avec ce choix de magnitude, tous les groupes avec un redshift  $z \leq 0.09$  et avec des galaxies qui répondent au critère de (2.4.2) sont visibles: on dit que le groupe est complet. Dans ce cas la luminosité caractéristique  $L_{19.5}$  ce calcul comme:

$$L_{19.5} = \sum_i \frac{L_i}{\mathcal{C}_i} \quad (2.4.3)$$

où  $\mathcal{C}_i$  est la complétude du survey à la position de la galaxie  $i$ , et  $i$  court sur les galaxies du groupe qui vérifient (2.4.2). Pour notre mock catalogue  $\mathcal{C}_i$  vaut 1 car toutes les galaxies sont comptées. Si le groupe n'est pas complet (i.e.  $z > 0.09$ ) alors la luminosité est:

$$L_{19.5} = \frac{1}{f(L_{19.5}, L_{lim})} \sum_i \frac{L_i}{\mathcal{C}_i} \quad (2.4.4)$$

## 2.4. L'ALGORITHME DE YANG ET AL.

## CHAPTER 2. GROUP FINDER ALGORITHM

avec  $L_{lim}$  la luminosité limite au redshift du groupe, et  $f(L_{19.5}, L_{lim})$  un facteur de correction pour prendre en compte les galaxies manquantes par le flux limite du SDSS, qui sera détaillé plus tard.

On calcule aussi les masses stellaires des groupes selon le même principe que pour les luminosités en sommant les masses stellaires des galaxies du groupe, et en corrigeant de la même façon quand le groupe n'est pas complet en utilisant un facteur de correction  $g(L_{19.5}, L_{lim})$  qui est calculé selon la même méthode que pour les luminosités mais pas pour la première itération. Les masses stellaires utilisées sont celles des galaxies qui répondent au critère de luminosité défini ci-dessus dans le groupe.  
③ On détermine maintenant les différentes caractéristiques de ces groupes. La masse du halo  $M_h$  est d'abord estimée à partir de la relation masse-luminosité suivante:

$$\frac{M_h}{L_{19.5}} = 500h \frac{M_\odot}{L_\odot} \quad (2.4.5)$$

Pour les itérations suivantes, on utilisera la relation  $\frac{M_h}{L_{19.5}} - L_{19.5}$  que l'on déterminera et qui sera détaillée dans la suite.

Le halo de matière noire est défini comme ayant une surdensité de 180 donc son rayon est calculé comme:

$$r_{180} = 1.26h^{-1}\text{Mpc} \left( \frac{M_h}{10^{14}h^{-1}M_\odot} \right)^{1/3} (1 + z_{group})^{-1} \quad (2.4.6)$$

où  $z_{group}$  est le redshift du centre du groupe.

La dispersion de vitesse sur la ligne de visée est calculée comme:

$$\sigma = 397.9 \text{ km} \cdot \text{s}^{-1} \left( \frac{M_h}{10^{14}h^{-1}M_\odot} \right)^{0.3214} \quad (2.4.7)$$

④ Mise à jour des groupes avec les informations du halo déterminées précédemment. Le nombre de contraste de densité aux alentours du centre du halo (confondu avec celui du groupe) peut s'écrire:

$$P_M(R, \Delta z) = \frac{H_0}{c} \frac{\Sigma(R)}{\bar{\rho}} p(\Delta z) \quad (2.4.8)$$

avec  $\Delta z = z - z_{group}$  où  $z$  est le redshift de la galaxie en question,  $\bar{\rho}$  la densité moyenne de l'Univers et  $\Sigma(R) = 2r_s \bar{\rho} f(R/r_s)$  la densité surfacique projetée pour un modèle de halo NFW (Navarro-Frenk-White, Navarro et al. [9]), avec  $r_s$  le rayon d'échelle. La fonction  $f$  s'écrit:

$$f(x) = \begin{cases} \frac{1}{x^2-1} \left\{ 1 - \frac{\ln[(1+\sqrt{1-x^2})/x]}{\sqrt{1-x^2}} \right\}, & x < 1 \\ 1/3, & x = 1 \\ \frac{1}{x^2-1} \left( 1 - \frac{\arctan \sqrt{x^2-1}}{\sqrt{x^2-1}} \right), & x > 1 \end{cases} \quad (2.4.9)$$

et avec aussi:

$$\bar{\delta} = \frac{180}{3} \frac{c_{180}^3}{\ln(1 + c_{180}) - c_{180}/(1 + c_{180})} \quad (2.4.10)$$

où  $c_{180} = r_{180}/r_s$ . La fonction  $p(\Delta z)d\Delta z$  décrit la distribution en redshift et est supposée avoir une forme gaussienne:

$$p(\Delta z) = \frac{1}{\sqrt{2\pi}} \frac{c}{\sigma(1 + z_{group})} \exp \left[ \frac{-(c\Delta z)^2}{2\sigma^2(1 + z_{group})^2} \right] \quad (2.4.11)$$

La façon de procéder est la suivante: pour chaque galaxie on boucle sur les groupes et on fait le calcul de  $R$  et de  $z$ , où  $R$  est la distance projetée au centre du groupe au redshift du groupe. Si  $P_M(R, \Delta z) \geq B$  avec  $B = 10$  choisit en fonction de précédents résultats, alors la galaxie appartient à ce groupe. Si elle peut être associée à plus d'un groupe par ce critère, alors on l'associe au groupe pour lequel  $P_M(R, \Delta z)$  est le plus grand.

⑤ Finalement en utilisant les nouveaux membres de groupe de l'étape précédente, on calcule à nouveau les centres de groupes puis on retourne à l'étape ②. Ceci se déroule tant qu'il y a des changements dans les groupes. À partir de ces groupes, on détermine la fonction de correction des luminosités  $f(L_{19.5}, L_{lim})$  ainsi que la relation masse-luminosité et on retourne à l'étape ①. Cette boucle s'achève une fois que la relation masse-luminosité a convergé.

Bien que les différents choix initiaux pour les différentes fonctions de correction comme  $f(L_{19.5}, L_{lim})$  ou la relation masse-luminosité soient grossiers, le fait de réaliser le calcul de manière itérative permet de gagner en précision car ces fonctions sont ensuite consistantes avec les données et les informations des groupes eux-mêmes. L'ensemble de ces étapes est résumé sur la figure (2.10).

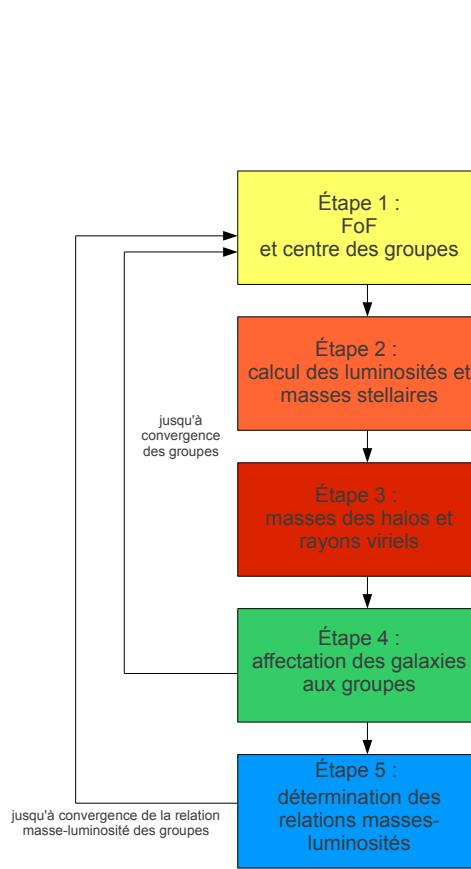


Figure 2.10: Diagramme résumant les étapes de l'algorithme de Yang et al. [2]. Les flèches représentent les itérations à réaliser et le texte à côté le critère d'arrêt de ces itérations.

Les méthodes développées pour correspondre à l'algorithme du Yang et al. [2] vont être décrites dans les parties suivantes.

## 2.4.2 Algorithme de Friends-of-Friends

L'étape ① du Yang et al. [2] est la réalisation d'un FoF qui va permettre de trouver des groupes potentiels de galaxies pour ensuite les mettre à jour par des itérations successives. Il a fallu alors réaliser l'algorithme du FoF. On va décrire dans la suite le principe du FoF et le travail réalisé pour le faire fonctionner.

### Principe

La technique du FoF n'est pas utilisée uniquement pour les groupes de galaxies de manière générale, mais essentiellement dans les simulations cosmologiques pour identifier les halos de particules de matière noire. Le principe va donc être expliqué à partir de ces simulations sans perte de généralité dans le cas des galaxies.

Dans les simulations, les halos sont identifiés d'abord de manière brutale par le FoF. Celui-ci considère que toutes les particules qui ont une voisine en commun de façon directe ou indirecte font partie du même halo. On peut voir ceci illustré sur la figure (2.11). Deux particules sont considérées voisines si la distance entre elles est plus petite qu'un seuil  $\epsilon$ .

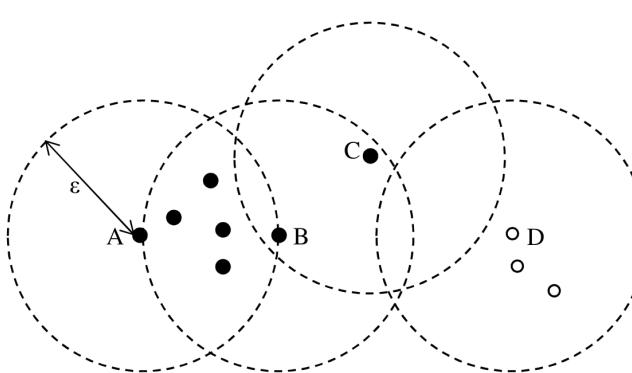


Figure 2.11: Relation du Friends-of-Friends: sur la figure,  $A$  est liée à  $C$  même si elles ne sont pas directement voisines, mais indirectement par l'intermédiaire de  $B$  qui est voisine de  $A$  et de  $C$ . Ces trois particules font partie du même groupe. Par contre  $D$  n'est liée d'aucune façon à  $A, B$  ou  $C$ :  $D$  appartient à un autre groupe.

La première idée qui vient à l'esprit pour mettre en œuvre un tel algorithme est de définir deux tableaux: le premier constitué des identités des  $N$  particules de la simulation allant donc de 1 à  $N$  et un second tableau de même dimension que le précédent contenant les identités des groupes des particules. Si on n'a aucun a priori sur la structure des particules, ce second tableau est initialisé de la même façon que le premier: chaque particule est son propre groupe. La méthode naïve qui suit est de calculer pour chaque particule lesquelles sont voisines de celle-ci (ce qui évolue en  $N^2$ ) et ensuite de changer les identités des groupes des particules qui sont voisines ou liées indirectement (celles qui ont la même identité de groupe que la particule qui est voisine) ce qui évolue, si on considère que  $M$  opérations d'union sont à réaliser, en  $MN$ . Pour le cas du SDSS ou de notre mock catalogue, cela peut prendre un temps considérable (plus d'un demi-million de galaxies). On a alors utilisé des méthodes qui permettent de gagner un important temps de calcul. Pour unir les groupes qui ont une galaxie commune, on a utilisé la méthode de l'Union-Find.

### L'Union-Find

Cette technique peut être bien comprise si on la visualise en une structure d'arbre. Dans les deux tableaux précédents le premier contient les identités des particules qui apparaissent sur les pastilles de

la figure (2.12) et les identités des groupes sont matérialisées par les liens entre les pastilles, l'identité de la particule à laquelle une autre est liée étant inscrite sur la pastille supérieure à laquelle elle est liée. Par exemple sur la figure (2.12), la particule 2 est liée à la 5 et possède comme numéro de groupe 5, et la 8 est liée à la 2 et possède comme numéro de groupe 2. La particule 1 est isolée.

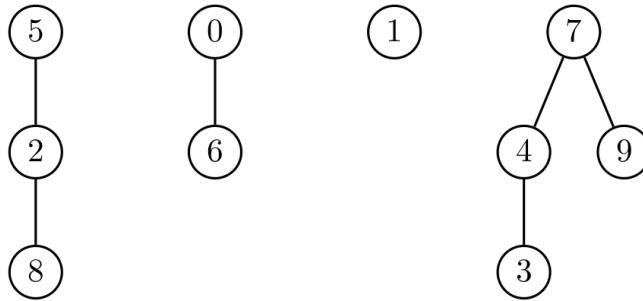


Figure 2.12: Représentation en arbre des liens entre les différentes particules.

Pour gagner du temps de calcul, on vérifie si deux particules sont déjà liées ou non en remontant à la racine de l'arbre pour voir si les identités diffèrent (pas encore liées) ou si elles sont identiques (déjà liées). On effectue la suite du travail si elles ne sont pas déjà liées. Ensuite pour lier entre eux deux groupes il est plus intéressant de lier celui qui contient le moins de membres à celui qui en contient le plus, c'est pourquoi on utilise un troisième tableau qui contient le nombre de membres de chaque groupe (initialisé à 1 au départ) et nous permet de choisir quel groupe il faut lier en pondérant par leur taille. On peut voir le résultat sur la figure (2.13) de l'union pondérée de 5 et 0 de la figure (2.12). Une fois que toute la recherche des voisins est terminée et que les liens ont

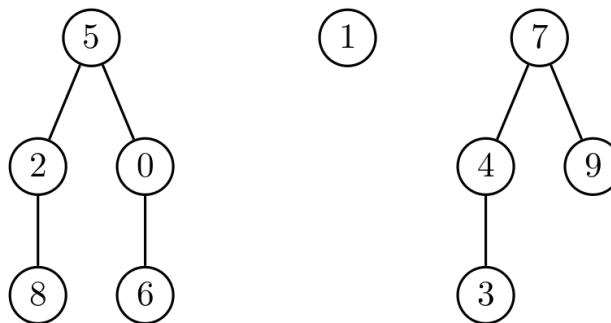


Figure 2.13: Résultat après l'union pondérée de 5 et 0 de la figure (2.12).

été établis, il est nécessaire d'aplanir les arbres en liant directement chaque particule à la racine à laquelle elle est liée directement ou indirectement. Pour cela on remonte chacune des branches des arbres en liant directement la particule à sa racine. On peut voir l'application de cette méthode sur la figure (2.14). L'implémentation de cette méthode permet de gagner un précieux temps de calcul car cet algorithme évolue en  $(M + N) \log N$  ce qui est quasi linéaire, le log étant presque comme une constante pour ces valeurs de  $N$ .

Bien qu'ayant diminué le temps de calcul par cette méthode, il reste le problème de la recherche des voisins qui se réalise toujours en  $N^2$  avec la façon simple.

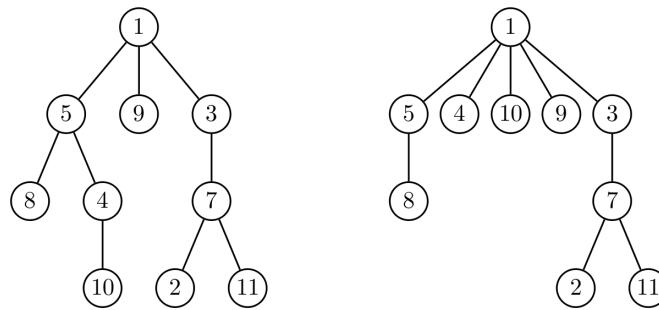


Figure 2.14: Résultat de la compression d'une branche en trouvant la racine de 10 et ensuite en liant (compressant) chaque noeud de la branche rencontré en le remontant à sa racine.

### Recherche des voisins

Le coût prohibitif de temps de calcul des voisins est causé par le fait que l'on a aucune connaissance du voisinage d'une particule qui pourrait nous faire diminuer le nombre d'opérations à effectuer, ce qui oblige à visiter chaque particule. On va donc essayer de structurer le voisinage des particules pour limiter le nombre de calculs à réaliser.

Une façon de le faire est de subdiviser la sphère céleste en boîtes délimitées par les méridiens comme on peut le voir sur la figure (2.15). On reproduit ce même découpage à différents niveaux en redshift, ce qui permet de créer des boîtes en 3D sur la sphère. Ce découpage est régulier sur  $(\alpha, \delta)$  et sur les redshift. On place ensuite chaque galaxie dans chaque boîte ainsi réalisée. Lors de la recherche des voisins, on prend chaque galaxie et on recherche ses voisins dans sa propre boîte et les boîtes voisines au cas où la galaxie serait proche d'un bord de cette boîte. Le choix de la taille de la boîte est un paramètre important car si elle plus petite que le seuil de distance qui définit que l'on est voisin ou non entre galaxies, il faut visiter plus d'une boîte dans le voisinage de celle de la galaxie dans une direction donnée. Pour être certain de ne pas rater de galaxies voisines, la taille des boîtes est définie à trois fois le seuil de distance à  $z$  proche de zéro. Sur la figure (2.15), on peut voir un autre problème de notre découpage en méridien: lorsqu'on s'approche du pôle céleste, si une galaxie s'y trouve, il faut rechercher sur l'ensemble des boîtes voisines car la distance angulaire entre deux boîtes est beaucoup plus faible. Pour prendre en compte cet effet de rapprochement entre les boîtes avec la proximité aux pôles, le nombre de boîtes à visiter est déterminé en fonction de  $\cos \delta$ .

Bien que cette méthode ait des lacunes, en pratique elle permet de diminuer fortement le temps de calcul alloué à la recherche des voisins, ce qui fait que le programme peut se dérouler assez rapidement malgré le nombre d'itérations qui peut être important. Il ne reste plus qu'à vérifier que le programme qui a été créé donne de bons résultats sur des simulations déjà analysées.

### Vérification

Pour faire la vérification de l'algorithme de FoF, nous avons utilisé un fichier contenant les positions de particules de matière noire issues d'une simulation de  $64^3$  particules. Ces données ont été fournies par Thierry SOUSBIE qui avait déjà réalisé un FoF dessus et on connaissait donc déjà les résultats à l'avance. Ce fichier contient aussi les informations permettant de retrouver les groupes de plus de 100 particules. Il nous a indiqué qu'il y 799 groupes de plus de 20 membres. Avec notre programme, on trouve 801 groupes dans ce cas là. Pour être plus sûr des résultats, on a réalisé la distribution en nombre de membres par groupe, c'est-à-dire l'histogramme du nombre de membres par groupes. La comparaison a indiqué que les distributions fournies par les données du fichier et notre

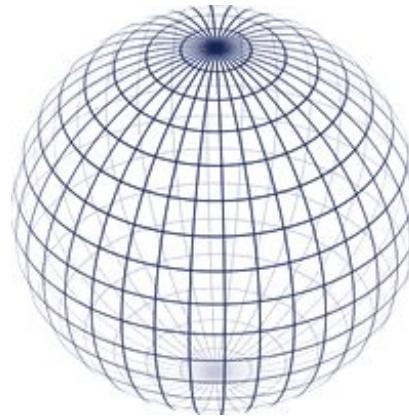


Figure 2.15: Image illustrant le découpage de la sphère en méridien, la même chose étant reproduite sur plusieurs couches de redshift.

programme sont sensiblement identiques: les plus gros groupes sont bien retrouvés mais quelques différences persistent avec des variations de un ou deux membres pour une quinzaine de groupes. Pour être certain de la méthode de l'Union-Find utilisée, on a utilisé aussi la méthode simple mais plus longue qui a été décrite plus haut. Les résultats obtenus sont identiques.

Après discussion avec Thierry SOUSBIE, il s'avère que le fichier fourni n'a pas la même précision pour les positions des particules que celle utilisée pour générer les données de ce fichier. Il est donc normal de trouver des résultats légèrement différents que ceux qui nous sont fournis. Si à cause de cette perte de précision dans les calculs, une particule ne se trouve plus voisine d'une autre, il se peut qu'elle soit voisine d'une autre particule avec laquelle elle n'est pas liée quand la précision des données est plus élevée. Ceci peut bouleverser complètement la structure déterminée par notre programme par rapport à celle réelle. Cela explique aussi les différences d'une ou deux particules observées pour les groupes trouvés par les deux programmes.

Du fait que la méthode simple, plus longue mais plus sûre, donne les mêmes résultats que la méthode rapide permet de dire que cet algorithme fonctionne bien, les différences étant imputables à la précision des données du fichier de test par rapport à la précision à laquelle elles ont été générées. Maintenant nous sommes certains du bon fonctionnement de cette partie du programme et on peut commencer à implémenter le reste de l'algorithme de Yang *et al.* [2].

### 2.4.3 Réalisation du Yang *et al.*

Les méthodes mises en œuvre pour chaque étape du Yang *et al.* vont maintenant être détaillées.

#### Étape 1: centre des groupes FoF préliminaires

L'algorithme de FoF a été mis en place, il reste à choisir le paramètre de lien pour le réaliser. Dans l'article il est indiqué qu'on choisit un lien de  $l_p=0.05$  selon la direction transverse à la ligne de visée et de  $l_z=0.3$  en unité de séparation moyenne des galaxies au redshift en question. En effet cette séparation varie avec le redshift et il faut donc la déterminer. On la calcule à l'aide de la densité comobile de galaxies. Elle s'exprime ainsi:

$$n(z) = \int_{L_{\min}(z)}^{\infty} \phi(L) dL \quad (2.4.12)$$

avec  $\phi(L)$  la fonction de luminosité différente selon le cas où on traite le mock catalogue ou le SDSS. Le choix de cette fonction de luminosité est important car elle doit traduire au mieux les propriétés

## 2.4. L'ALGORITHME DE YANG ET AL.

## CHAPTER 2. GROUP FINDER ALGORITHM

du catalogue de galaxies que l'on considère afin d'estimer correctement la séparation moyenne des galaxies. Dans Guo et al. [3], il est indiqué que la fonction de luminosité utilisée pour réaliser les simulations du mini-MSII correspondent assez bien aux données du SDSS utilisées dans Blanton et al. [10]. Afin de vérifier, que cela est vrai on a réalisé la fonction de luminosité du mini-MSII et on l'a comparée à celle donnée dans Blanton et al. [10]. Il s'avère qu'il y a quelques différences qui peuvent jouer de façon importante dans cette estimation de la séparation moyenne. C'est pourquoi on a essayé de modéliser la fonction de luminosité de Guo et al. [3] par une interpolation par splines cubiques, après avoir tenté différentes approches pour cette modélisation comme l'ajustement par des polynômes de différents ordres de la fonction de luminosité et également par des fonctions de Schechter avec des paramètres obtenus par la méthode du maximum de vraisemblance. Il résulte de cette modélisation par splines cubiques qu'elle permet d'ajuster de façon très précise les données du Guo2010a.

La fonction de luminosité des données du Guo2010a est tracée sur la figure (2.16) (en orange) avec celle obtenue par les splines cubiques (en rouge) et celle de Blanton et al. [10] (en vert). On remarque bien que l'ajustement à l'aide des splines cubiques donne de bons résultats, meilleurs que ceux de Blanton et al. [10]. L'intégrale de cette fonction de luminosité pour obtenir la densité de galaxies ne sera fera alors pas jusqu'à l'infini mais jusqu'à la magnitude qui correspond à la plus lumineuse du catalogue Guo2010a (qui est la valeur limite visible sur la figure (2.16)).

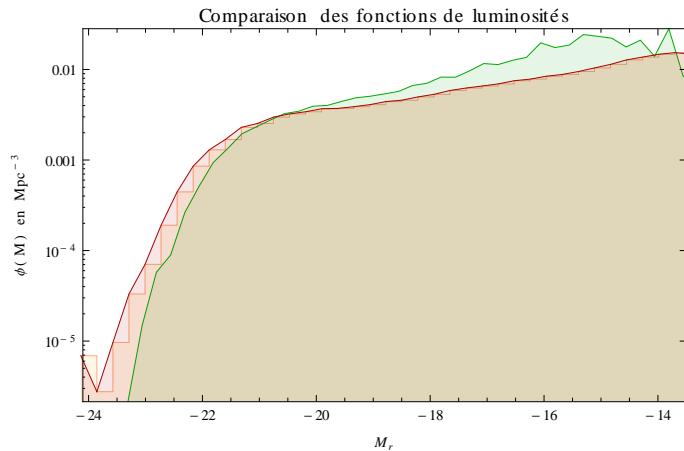


Figure 2.16: Comparaison des différentes fonctions de luminosités: la fonction représentée en orange est celle déterminée à partir des données du catalogue Guo2010a du mini-MSII, celle en vert les données de la fonction obtenues à partir de Blanton et al. [10] et celle en rouge l'ajustement par les splines cubiques que l'on a appliqué.

Dans le cas du SDSS, on va adopter comme fonction de luminosité celle de Blanton et al. [10] car elle permet de modéliser de façon optimale les données du survey SDSS. Il ne reste plus qu'à déterminer la luminosité limite en fonction du redshift pour estimer la séparation moyenne entre galaxies. On sait que la magnitude apparente limite du SDSS est 17.77 en bande  $r$ , et on la note  $m_{\text{lim}}$ . La luminosité limite à un redshift donné traduit la magnitude absolue maximale des galaxies observées à ce redshift  $M_{\text{lim}}$  et on peut écrire:

$$M_{\text{lim}} = m_{\text{lim}} - 5 \log_{10} \left( \frac{d_{\text{lum}}}{10pc} \right) \quad (2.4.13)$$

Donc la luminosité limite peut s'écrire:

$$L_{\text{lim}}(z) = \left( \frac{d_{\text{lum}}}{10pc} \right)^2 10^{0.4(M_{\odot} - m_{\text{lim}})} \quad (2.4.14)$$

où la dépendance en  $z$  se fait par l'intermédiaire de la distance lumineuse. Théoriquement pour réaliser les calculs précisément on devrait prendre en compte les effets de la correction-K sur cette luminosité limite qui dépend aussi de l'allure du spectre de la galaxie donc de son type et de sa couleur. Mais pour faciliter le travail et les calculs à réaliser, on se passera de cette précision sur la luminosité limite, sinon il aurait fallu déterminer une correction-K moyenne en fonction du redshift afin de prendre en compte cet effet, ce qui n'est pas précis également.

Pour vérifier que cette méthode donne de bons résultats, on va essayer de réaliser un comptage des galaxies dans notre mock catalogue et le comparer à un calcul théorique pour voir les différences qui peuvent apparaître et influencer le calcul de la séparation moyenne dans le FoF. La densité comobile de galaxies peut s'exprimer sous cette forme également:

$$n(z) = \frac{dN}{dz} \frac{dz}{dV} \quad (2.4.15)$$

où  $dN$  est le nombre de galaxies dans un intervalle de  $dz$  et  $dV$  le volume comobile élémentaire de calcul.  $dN/dz$  peut être déterminé par comptage des galaxies par intervalles de redshift directement à partir des données.  $dV$  peut s'exprimer comme:

$$dV = D_H \frac{(1+z)^2 D_A^2}{E(z)} d\Omega dz \quad (2.4.16)$$

avec  $D_A$  la distance angulaire,  $D_H = c/H_0$  et  $d\Omega$  l'angle solide dans lequel on réalise le comptage. La distance angulaire s'exprime en fonction de la distance lumineuse comme:

$$D_A = \frac{d_{\text{lum}}}{(1+z)^2} \quad (2.4.17)$$

La fonction  $E(z)$  est le rapport de la constante de Hubble à un  $z$  donné par rapport à  $z = 0$ . Elle s'exprime (dans le cas d'un Univers plat) comme:

$$E(z) = \sqrt{\Omega_m(1+z)^3 + \Omega_\Lambda} \quad (2.4.18)$$

où  $\Omega_m$  représente la fraction de la matière dans l'Univers et  $\Omega_\Lambda$  la part de l'énergie noire, avec  $\Omega_m + \Omega_\Lambda = 1$ . En utilisant à nouveau l'approximation pour la distance lumineuse donnée par Wickramasinghe and Ukwatta [7], on peut déterminer facilement  $dN/dz$  de façon théorique avec les expressions de  $n(z)$  basées sur les fonctions de luminosité. Les résultats sont présentés sur la figure (2.17) pour un mock catalogue où aucune correction-K n'a été appliquée sur les galaxies. Ainsi le calcul théorique de  $dN/dz$  doit coller parfaitement avec le comptage réalisé sur le mock catalogue, aucune perte de galaxies par un biais quelconque ne devant se produire dans le mock catalogue. Sur la figure (2.17) les résultats à partir du mock catalogue non K-corrigé et sans prise en compte des vitesses particulières pour les redshifts sont en bleu et notre calcul théorique en violet. Les résultats sont très similaires pour les deux courbes, les fluctuations visibles sur la courbe bleue étant causées par des variations locales de la densité de galaxies dans le mock (la répartition en filaments des galaxies y étant pour une grande partie responsable). On voit aussi une divergence entre les deux courbes aux grands redshifts, causées par une mauvaise estimation du nombre de galaxies avec une luminosité élevée (seules visibles à ces grandes distances) à cause de notre choix de magnitude limite pour la borne supérieure de l'intégrale de l'équation (2.4.12) qui est un peu arbitraire et du bruit dans la fonction de luminosité en orange sur la figure (2.16) qui donne une moins bonne évaluation par les splines cubiques et donc une valeur de l'intégrale moins précise.

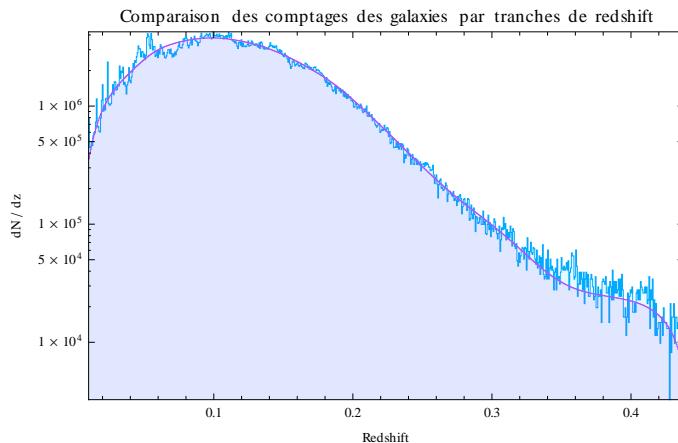


Figure 2.17: Comptage des galaxies par intervalle de redshift: en bleu les résultats obtenus directement à partir de notre mock catalogue (sans correction-K ni vitesse particulière), en mauve le calcul théorique détaillé dans le rapport.

On fait le même travail pour un mock catalogue avec cette fois une correction-K appliquées à chaque galaxie et des vitesses particulières pour calculer les redshifts comme décrit dans la section () pour voir l'effet de la correction-K par rapport à notre estimation théorique du nombre de galaxies. Le résultat se trouve sur la figure (2.18). Cette correction réduit un petit peu le nombre de galaxies par rapport au cas précédent et s'écarte donc un peu du calcul théorique mais pas avec des écarts importants ce qui nous permet de conserver notre estimation de la densité moyenne de galaxie par les splines cubiques dans le cas du mock catalogue.

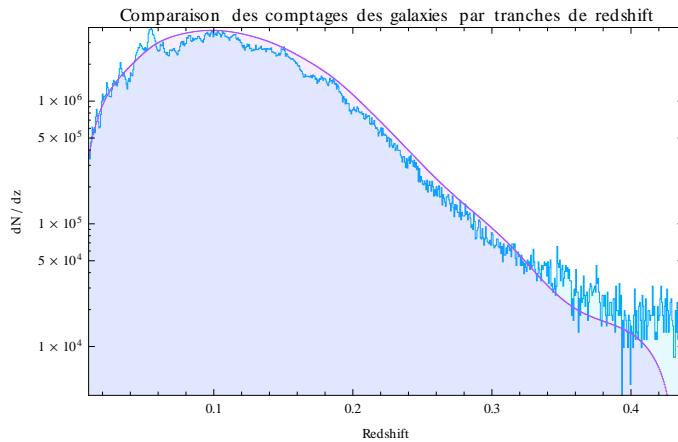


Figure 2.18: Comptage des galaxies par intervalle de redshift: les résultats sont présentés de la même manière que la figure (2.17) sauf que cette fois une correction-K a été appliquée sur les galaxies du mock et les vitesses particulières ont été prises en compte pour la calcul des redshifts. On voit que leur considération dans le mock catalogue atténue un peu le nombre de galaxies disponibles dans notre mock par rapport au cas théorique où on ne prend pas ces effets en compte. Les différences entre les courbes théoriques avec la figure (2.17) sont dues à de légères différences entre les bornes de l'intégration pour la magnitude minimale qui influent sur les grandes luminosités (donc les grands redshifts), pour des facilités de calcul et ont peu d'importance étant donnée l'imprécision du calcul à ces redshifts.

Nous pouvons maintenant déterminer la séparation moyenne entre galaxies, celle-ci étant égale à  $1/n(z)^{1/3}$ , pour connaître le paramètre de lien entre galaxies.

Pour savoir si deux galaxies sont voisines ou non, on utilise les critères définis dans Eke et al. [11]. En posant  $l_p$  et  $l_z$  les paramètres de lien entre galaxies dans une direction transverse à la ligne

de visée et suivant la ligne de visée respectivement, on considère que deux galaxies  $i, j$  sont voisines si la séparation angulaire entre les deux galaxies  $\theta_{ij}$  est telle que:

$$\theta_{ij} \leq \frac{1}{2} \left( \frac{l_{p,i}}{d_{c,i}} + \frac{l_{p,j}}{d_{c,j}} \right) \quad (2.4.19)$$

et si:

$$|d_{c,i} - d_{c,j}| \leq \frac{l_{z,i} + l_{z,j}}{2} \quad (2.4.20)$$

avec  $d_c$  la distance comobile de la galaxie correspondante.

Nous avons maintenant complété l'étape ① et on peut passer à la description de l'étape ②.

## Étape 2: correction des luminosités

Dans cette étape, il y a deux effets considérés: l'incomplétude de la luminosité causée par la magnitude limite du survey ainsi que la correction provoquée par les effets de bord. Il est tout d'abord nécessaire de déterminer la fonction de correction des luminosités pour la première itération. Elle est définie ainsi:

$$f(L_{19.5}, L_{lim}) = \frac{\int_{L_{lim}}^{\infty} L\phi(L)dL}{\int_{L_{cut}}^{\infty} L\phi(L)dL} \quad (2.4.21)$$

avec  $L_{cut}$  la luminosité correspondante à  $M_r = -19.5$ , luminosité de coupure de l'échantillon et  $\phi(L)$  la fonction de luminosité. Ceci permet de représenter la part de luminosité des membres effectivement dans le groupe par rapport à ce que cela serait si le groupe était complet. Donc pour corriger de la part manquante dans la luminosité caractéristique, on multiplie par l'inverse de cette fonction  $f$ .

Pour le choix de la fonction de luminosité nous avons choisie celle donnée par Blanton et al. [10] qui modélise au mieux les données du SDSS. La fonction s'exprime ainsi:

$$\phi(L)dL = \frac{dL}{L_*} \exp\left(-\frac{L}{L_*}\right) \left[ \phi_{*,1} \left(\frac{L}{L_*}\right)^{\alpha_1} + \phi_{*,2} \left(\frac{L}{L_*}\right)^{\alpha_2} \right] \quad (2.4.22)$$

Les valeurs des paramètres sont listées dans le tableau (2.4). Quand on injecte cette fonction de luminosité dans l'équation (2.4.21), on voit que la fonction de correction peut s'exprimer comme une somme de fonction gamma incomplète  $\Gamma(a, x) = \int_x^{\infty} t^{a-1} e^{-t} dt$  et on peut donc écrire:

$$f(L_{19.5}, L_{lim}) = \frac{\phi_{*,1}\Gamma(2 + \alpha_1, \frac{L_{lim}}{L_*}) + \phi_{*,2}\Gamma(2 + \alpha_2, \frac{L_{lim}}{L_*})}{\phi_{*,1}\Gamma(2 + \alpha_1, \frac{L_{cut}}{L_*}) + \phi_{*,2}\Gamma(2 + \alpha_2, \frac{L_{cut}}{L_*})} \quad (2.4.23)$$

$M_* - 5 \log_{10} h$	$\phi_{*,1}$	$\alpha_1$	$\phi_{*,1}$	$\alpha_2$
$(10^{-2}h^3 Mpc^{-3})$	$(10^{-2}h^3 Mpc^{-3})$		$(10^{-2}h^3 Mpc^{-3})$	

Table 2.4: Les paramètres de la fonction de luminosité de Blanton et al. [10] utilisés dans notre cas.

Pour réaliser le calcul de cette fonction  $\Gamma$  incomplète, on a utilisé la procédure qui est décrite en annexe (D) qui réalise le calcul de manière rapide.

Dans le cas du mock catalogue, on utilise aussi la fonction de luminosité calculée directement à partir des données du Guo2010a et on ajuste à l'aide des splines cubiques la fonction  $L\phi(L)$  qu'on doit intégrer de la même façon que précédemment. Ce type de calcul pour la fonction  $f$  n'est réalisé que pour la première itération suivant le FoF ou la détermination d'une relation masse-luminosité ainsi que celle de la fonction de correction elle-même. La façon dont on détermine cette fonction de correction pour les autres itération sera décrite plus loin.

Le Yang et al. [2] permet également de déterminer les masses stellaires des groupes de galaxies à partir des masses stellaires des galaxies déterminées par l'équation (2.4.1). Le principe de la détermination des masses stellaires des groupes est le même sur le principe que pour la luminosité des groupes sauf que la fonction de correction notée  $g$  dans ce cas là n'est pas déterminée pour la première itération mais on utilise seulement la relation déduite pour chaque itération décrite plus loin.

Un autre effet important doit être pris en compte à cette étape: c'est l'effet de bord. Certains groupes que l'on détermine sont situés au bord soit du mock catalogue soit du survey pour le SDSS. Ceci peut faire en sorte que certains groupes aient des galaxies manquantes dans leur membres, ce qui se traduit pas une mauvaise évaluation de leurs propriétés. Il faut donc prendre en compte cet effet. Pour cela le principe est à peu de choses près le même pour le mock et pour le SDSS. On recherche d'abord les groupes qui sont susceptibles d'avoir des galaxies manquantes parce qu'ils sont situés aux bords, donc les groupes qui ont la position de leur rayon de Viriel qui sort du mock peuvent avoir des galaxies manquantes (une partie de la sphère viruelle se retrouve en-dehors du mock), et pour le SDSS il s'agit des groupes qui sont à une complétude  $\mathcal{C} > 0.7$  dans leur sphère viruelle. Une fois ces groupes trouvés, on distribue aléatoirement 200 points dans la sphère viruelle du groupe et tous les points qui sortent alors du groupe (par leur position pour le mock et par la complétude pour le SDDS) sont supprimés. On peut voir cela sur la figure (2.19) où on a représenté un groupe dont la sphère viruelle sort de la zone du mock délimitée par le plan en vert. Les points qui sont placés aléatoirement sont en jaune s'ils sont sortis du mock et en bleu s'ils sont dans les limites du catalogue. On détermine pour chaque groupe le nombre de points restant parmi les 200 noté  $N_{\text{restant}}$  et on définit  $f_{\text{edge}} = N_{\text{restant}}/200$  comme la part du volume du groupe qui repose dans le catalogue. La fraction des galaxies manquantes dans les propriétés des groupes (luminosité et masse stellaire) est alors corrigée en ajoutant un facteur de correction  $1/f_{\text{edge}}$  quand on estime la luminosité et les masses stellaires comme dans le cas où le groupe est incomplet. Cependant ceci ne fonctionne pas bien pour les groupes avec un  $f_{\text{edge}}$  petit donc on élimine les groupes avec  $f_{\text{edge}} < 0.6$ .

### Étape 3: masses et rayons des groupes

Cette étape nécessite de déterminer la masse du halo associé au groupe trouvé précédemment. Pour cela on utilise une relation masse-luminosité fixée au départ, puis on utilise la relation que l'on détermine aux itérations suivantes. La méthode pour la déterminer sera décrite elle aussi plus loin.

### Étape 4: appartenance des galaxies

On doit calculer à cette étape la densité de contraste qui nous sert à définir un critère d'appartenance d'une galaxie à un groupe. Pour cela on doit pour chaque galaxie boucler sur les groupes et calculer la distance projetée au centre du groupe, au redshift du groupe. Si on réalise le calcul de cette façon, le même problème que celui décrit plus haut sur le temps nécessaire se pose. On procède alors là aussi à un découpage de la sphère céleste pour accélérer cette recherche. La taille des boîtes dans ce cas est choisie plus grande pour ne pas introduire de sélection des galaxies,

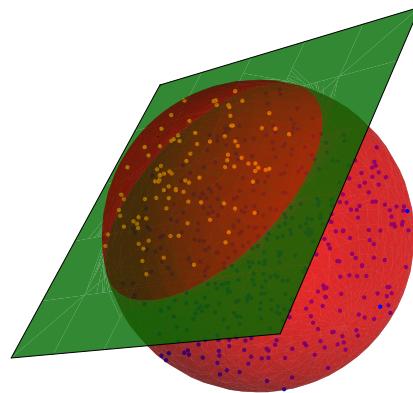


Figure 2.19: Représentation de l'effet de bord des groupes. La sphère virielle est représentée en rouge et la bordure du catalogue par le plan en vert. Les points placés aléatoirement dans la sphère virielle sont jaunes s'ils sont en dehors du catalogue et sont en bleu s'ils sont dans les limites du catalogue. Cette image n'est valable que pour le mock catalogue qui a des bordures nettes alors que pour le SDSS, il faut utiliser l'incomplétude du survey  $\mathcal{C}$ .

leur appartenance au groupe dépendant maintenant d'un critère qui n'est pas défini seulement par la distance au centre du groupe. En prenant des boîtes "larges" pour le découpage, on supprime les effets de sélection. Le découpage réalisé dans ce cas n'est pas le même que celui réalisé pour le FoF, la taille des boîtes étant conservées à peu près constante pour chaque bin en redshift et quand on fait varier la déclinaison  $\delta$  et l'ascension droite  $\alpha$ . La structure résultante n'est pas alors aussi simple que celle de la figure (2.15) et des précautions supplémentaires ont été appliquées pour éviter de manquer des boîtes dans la recherche des groupes auxquels une galaxie peut appartenir. Les boîtes ont été choisies larges dans le sens des redshifts afin de prendre en compte les effets d'élongation évoqués dans la section (2.2.1). Une image qui pourrait convenir est celle de la figure (2.20). On voit que dans ce cas le maillage n'est pas régulier. Les boîtes représentent le découpage dans l'espace projeté (sphère céleste) et les redshifts. Comme on essaie de garder des boîtes de taille constante dans l'espace physique, cela se traduit par un nombre de boîtes plus important aux grands redshifts qu'aux petits. Et étant donnée la géométrie sphérique du problème, pour une boîte où on cherche les groupes potentiels d'une galaxie, il faut chercher dans un plus grand nombre de boîtes sur l'extrémité en  $z$  où  $z$  est le plus grand que sur celle où  $z$  est plus faible. C'est ce qu'on tente de représenter sur la figure (2.20) où les différentes tranches de redshift ont des boîtes colorées différemment. Pour représenter le nombre de boîtes à visiter pour la galaxie située dans la boîte rouge qui est croissant avec le redshift, la taille des boîtes a été diminué quand  $z$  augmente (la taille des boîtes en angle diminue au fur et à mesure que l'on s'éloigne de l'observateur). Si on se place dans la boîte rouge en tant que galaxie et que l'on cherche les groupes voisins, il faut chercher dans les boîtes adjacentes seulement pour un  $z$  donné (les boîtes orange), par contre les boîtes juste inférieures et supérieures en redshift sont en contact de manières différentes selon les situations avec la boîte rouge. La recherche des groupes voisins dans les boîtes voisines n'est alors pas très aisée: on a donc décidé de rechercher dans les 5 boîtes voisines en  $\alpha$  et pour chaque étage en  $\delta$  dans le but de limiter le nombre de recherche tout en estimant que l'on ne manque aucun groupe en prenant de cette façon *large* dans les boîtes voisines (ce qu'on peut voir également sur la figure (2.20) sur la vue de face où on prend en compte toutes les boîtes voisines en appliquant ce traitement). On estime donc que la méthode mise en œuvre permet d'optimiser le programme en diminuant fortement le temps de calcul tout en permettant d'effectuer

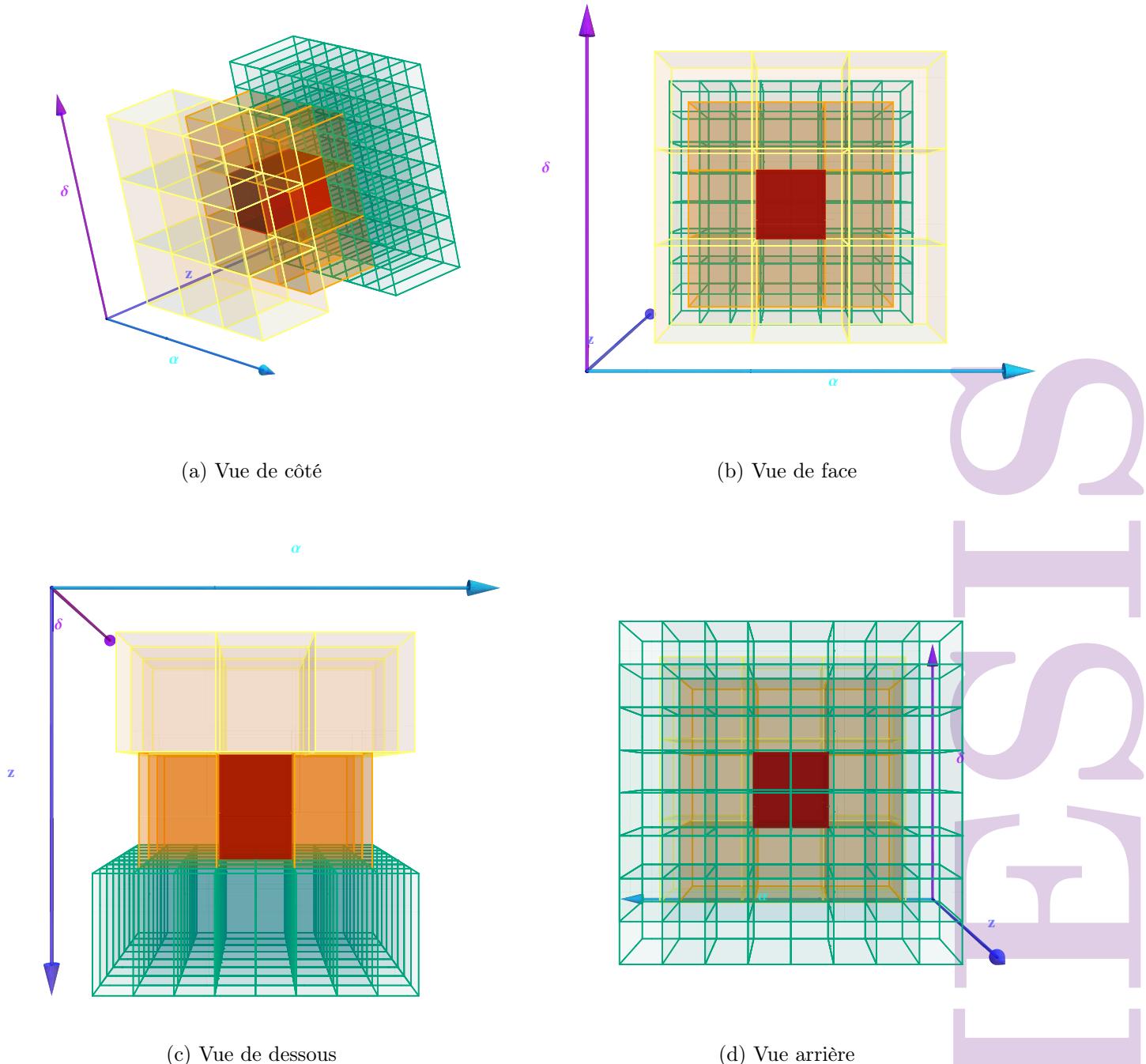


Figure 2.20: Illustration du découpage de la sphère céleste dans un maillage non régulier. La flèche bleue foncée représente l'axe croissant des redshifts, la flèche bleue claire l'ascension droite  $\alpha$  et la flèche violette la déclinaison  $\delta$ . La boîte rouge au centre représente la boîte où se trouve la galaxie dont on cherche les groupes auxquels elle pourrait appartenir. Les quatre figures représentent le même dessin mais vu de plusieurs points de vue différents pour aider à la visualisation en 3D du découpage. Il faut garder à l'esprit qu'il s'agit d'une représentation par des cubes de sections de cône, mais cela est sensiblement identique quand le nombre de boîtes est élevé pour le découpage et que l'on regarde seulement un petit morceau de la sphère.

une recherche des groupes voisins sûre.

Comme décrit dans la section (2.4.1), on doit calculer la concentration  $c_{180} = r_{180}/r_s$ , mais on ne connaît pas  $r_s$  a priori. Donc pour déterminer ce rayon caractéristique, on va utiliser les résultats établis par Macciò et al. [12] pour calculer la concentration. On se base sur les résultats obtenus pour une cosmologie WMAP5 qui donnent:

$$\log_{10} c_{200} = 0.830 - 0.098 \log_{10} \left( \frac{M_{200}}{10^{12} h^{-1} M_\odot} \right) \quad (2.4.24)$$

Or nous avons besoin des résultats avec  $c_{180}$ , on va donc appliquer quelques corrections à ces résultats pour déterminer  $r_s$ . D'une manière qui ne sera pas décrite ici, on peut déterminer le rapport  $r_{180}/r_{200}$  qui varie peu avec la concentration et de la même façon le rapport des masses  $M_{180}/M_{200}$ . On obtient pour ces rapports:

$$\begin{aligned} \log_{10} \left( \frac{M_{180}}{M_{200}} \right) &\approx 0.012 \\ \text{et } \log_{10} \left( \frac{r_{180}}{r_{200}} \right) &\approx 0.019 \end{aligned} \quad (2.4.25)$$

Il nous suffit maintenant de déterminer l'expression de  $M_{180}$  afin de pouvoir obtenir une valeur de  $c_{180}$ . En notant  $\Delta$  la sur-densité de l'Univers, et  $\rho_c$  la densité critique, on a:

$$M_{180} = \Delta \left( \frac{4}{3} \pi r_{180}^3 \right) \rho_c \quad (2.4.26)$$

Dans notre situation,  $\Delta = 180$  et en utilisant l'expression de la densité critique:

$$M_{180} = \frac{90 H_0^2 r_{180}^3}{G} \quad (2.4.27)$$

Tout ceci permet finalement d'écrire:

$$\log_{10} c_{180} = 0.849 - 0.098 \log_{10} \left( \frac{90 H_0^2 r_{180}^3}{G h^{-1} 10^{12} M_\odot} \right) \quad (2.4.28)$$

avec  $G$  exprimé dans les unités appropriées.

Il reste un dernier problème à régler si on réalise comme il est décrit dans l'article le regroupement des galaxies selon le paramètre de densité de contraste. En effet, si on considère deux galaxies proches et non encore attachées à un groupe et que le critère permet de lier la première galaxie à la seconde (considérée comme un groupe même si elle est seule dans ce groupe) et la seconde à la première, la simple application de la méthode décrite plus haut et dans Yang et al. [2] donnera toujours deux galaxies isolées (deux groupes distincts) car la première prendra l'identité du groupe de la seconde et la seconde l'ancien numéro de groupe de la première. Finalement elles seront considérées comme appartenant à deux groupes différents alors que selon le critère de regroupement elles peuvent appartenir au même groupe. On applique donc une vérification quand on lie une galaxie à un groupe afin de déterminer si cette galaxie n'a pas déjà été liée à une galaxie isolée pour prévenir ainsi ces oscillations de paires. Le même genre de situation problématique peut survenir avec des triplets (voire des multiplicités plus grandes) de galaxies isolées où chaque galaxie peut être liée ainsi successivement

## 2.4. L'ALGORITHME DE YANG ET AL.

## CHAPTER 2. GROUP FINDER ALGORITHM

à sa voisine dans une sorte de cercle vicieux. Mais on considère que ce genre de situation a peu de chances de se produire et aucune vérification n'est réalisée pendant le regroupement pour empêcher ce type "d'oscillations" d'appartenance aux groupes.

### Étape 5: itérations

Pour déterminer les groupes avant le calcul de la fonction de correction et de la relation masse-luminosité, il faut faire une itération sur la détermination des membres des groupes. Elle doit cesser lorsque le nombre de groupes ainsi que leur richesse en galaxie ont peu évolués par rapport à l'itération précédente. Pour définir un critère de convergence, on va utiliser la fonction de multiplicité des groupes, c'est-à-dire déterminer le nombre de groupes qui ont un certain nombre de galaxies comme membres. Le critère sera proche de celui du  $\chi^2$  en statistique.

Pour résumer, on cherche à savoir si le nombre de groupes avec un certain nombre de galaxies comme membres évolue peu. On dispose donc de deux distributions de deux itérations successives. On note  $N_{1i}$  le nombre de groupes de la première distribution (1) ayant  $i$  galaxies membres et  $N_{2i}$  la même chose pour la seconde distribution (2). La convergence est assurée si le nombre de groupes avec un certain nombre de galaxies membres n'évolue plus beaucoup, donc on définit le  $\chi^2$  comme:

$$\chi^2 = \frac{1}{N_{\text{tot}}} \sum_{i=1}^{N_{\text{tot}}} \frac{(N_{1i} - N_{2i})^2}{(N_{1i} + N_{2i})^2} \frac{1}{\sigma^2 \left( \frac{N_{1i} - N_{2i}}{N_{1i} + N_{2i}} \right)} \quad (2.4.29)$$

c'est-à-dire que l'on réalise une statistique sur les  $\Delta_i = \frac{|N_{1i} - N_{2i}|}{N_{1i} + N_{2i}}$ , ce qui permet d'évaluer une différence relative du nombre de groupes ayant un certain nombre de galaxies. Comme on réalise un comptage, on peut utiliser une statistique de Poisson, donc l'écart type  $\sigma$  peut se calculer comme:

$$\begin{aligned} \sigma^2(N_{1i}) &= N_{1i} \\ \sigma^2(N_{1i} \pm N_{2i}) &= N_{1i} + N_{2i} \\ \sigma^2 \left( \frac{A}{B} \right) &= \left( \frac{(A\sigma(B))}{B^2} \right)^2 + \left( \frac{\sigma(A)}{B} \right)^2 \end{aligned} \quad (2.4.30)$$

On trouve alors:

$$\sigma^2 \left( \frac{N_{1i} - N_{2i}}{N_{1i} + N_{2i}} \right) = \frac{(N_{1i} - N_{2i})^2}{(N_{1i} + N_{2i})^3} + \frac{1}{N_{1i} + N_{2i}} \approx \frac{1}{N_{1i} + N_{2i}} \quad (2.4.31)$$

Donc finalement le critère du  $\chi^2$  se résume à:

$$\chi^2 = \frac{1}{N_{\text{tot}}} \sum_{i=1}^{N_{\text{tot}}} \frac{(N_{1i} - N_{2i})^2}{(N_{1i} + N_{2i})^3} \quad (2.4.32)$$

Plus  $\chi^2$  devient petit et plus les distributions sont proches l'une de l'autre. À partir de cela on peut voir comment se comporte le  $\chi^2$  au fur et à mesure des itérations dans un cas pour déterminer à quel moment il converge vers une valeur limite qui servira à définir le seuil à partir duquel on stoppe les itérations. On peut voir son évolution avec les itérations sur la figure (2.21). Le  $\chi^2$  finit par converger aux alentours d'une valeur que l'on se fixera approximativement comme seuil pour l'arrêt des itérations sur la mise à jour des membres des groupes. On voit aussi, comme on s'y attendait pour les raisons évoquées plus haut sur les oscillations de certaines galaxies entre différents groupes,

une périodicité dans la valeur du  $\chi^2$  car au bout d'un certain nombre d'itérations, seules ces galaxies oscillantes empêchent une convergence complète des groupes ( $\chi^2 = 0$ ). Cela se traduit par un état des groupes qui se retrouve au bout d'un certain temps (car les galaxies oscillantes se retrouvent toutes dans les mêmes groupes à une itération qu'à une itération précédente) et donc une périodicité du  $\chi^2$  sur la figure (2.21).

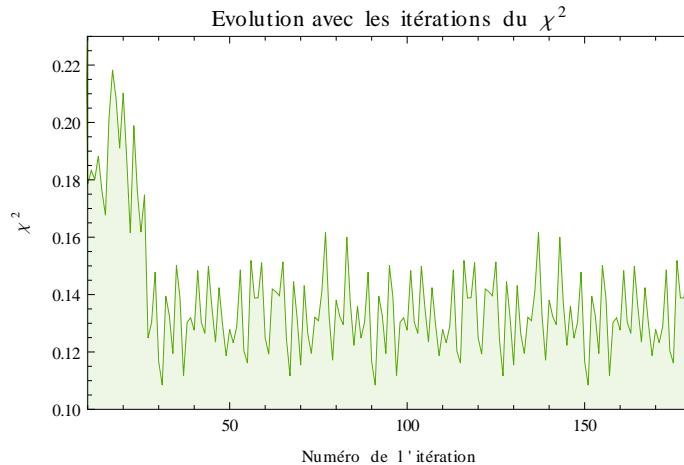


Figure 2.21: Évolution du  $\chi^2$  avec les itérations: on voit que les itérations convergent vers une valeur. On observe aussi une périodicité à partir de l'itération 30 traduisant des oscillations de certaines galaxies entre des groupes.

À cette étape on doit aussi déduire les fonctions de correction des luminosités et masses stellaires des groupes ainsi que la relation masse du halo-luminosité grâce aux propriétés des groupes obtenues aux itérations précédentes. Tout d'abord il nous faut déterminer les fonctions de correction. On rappelle que cette fonction a pour but de corriger les masses stellaires et les luminosités des groupes qui ne sont pas complets, c'est-à-dire de déterminer la part manquante des luminosités et masses stellaires dans les groupes trop éloignés où certaines galaxies ne sont pas visibles même avec une magnitude absolue inférieure à  $-19.5$ . On choisit donc les groupes que l'on considère complets c'est-à-dire avec un redshift inférieur à 0.09. Pour chacun d'eux on calcule pour une luminosité limite donnée  $L_{\text{lim}}$  le nombre de galaxies qui ont une luminosité supérieure à  $L_{\text{lim}}$ . On détermine la luminosité totale des galaxies qui vérifient cette condition dans le groupe et on la divise par la luminosité du groupe  $L_{19.5}$ , on définit ainsi la part de la luminosité due à la contribution de ces galaxies "complètes" par rapport à la luminosité  $L_{\text{lim}}$ . On fait le même genre de travail pour les masses stellaires et pour différentes valeurs de  $L_{\text{lim}}$  dans les deux cas. Pour une luminosité  $L_{\text{lim}}$  donnée, ces fractions de luminosité et de masse stellaire varient avec la luminosité  $L_{19.5}$  du groupe. Si on veut définir une fonction de correction qui dépend de  $L_{\text{lim}}$  et de  $L_{19.5}$  on va donc calculer une moyenne de ces fractions pour différents bins de  $L_{19.5}$ , pour chaque  $L_{\text{lim}}$  pour lesquelles on réalise le calcul. L'évolution de ces fractions avec  $L_{19.5}$  est ensuite modélisée avec une exponentielle décroissante en  $L_{19.5}$ . Les deux paramètres de cette exponentielle pour les différents  $L_{\text{lim}}$  sont ensuite ajustés à l'aide de splines cubiques pour pouvoir en déduire une fonction de correction qui dépend de  $L_{\text{lim}}$  par l'intermédiaire de l'évolution des coefficients de l'exponentielle avec  $L_{\text{lim}}$ .

Une fois ceci fait, il faut déterminer aussi la relation entre la luminosité des groupes  $L_{19.5}$  et la masse du halo. La procédure utilisée est un peu plus complexe sur le principe. À la fin de l'itération on dispose de la luminosité  $L_{19.5}$  de chaque groupe. On peut donc en déduire une fonction de répartition de la luminosité des groupes, en d'autres termes connaître le nombre de groupes qui ont une luminosité supérieure à une luminosité  $L_{19.5}$  donnée. Cela peut se faire très rapidement en réarrangeant dans

## 2.4. L'ALGORITHME DE YANG ET AL.

## CHAPTER 2. GROUP FINDER ALGORITHM

le programme le tableau contenant les luminosités des groupes par ordre décroissant, l'index (*i.e.* la position dans le tableau) du groupe donnant alors directement le nombre de groupes qui ont une luminosité supérieure à celle de ce groupe. On fait maintenant une grande hypothèse qui va introduire quelques erreurs dans la relation que l'on cherche à déterminer en supposant qu'il existe une relation une-à-une entre la masse du halo et la luminosité  $L_{19.5}$  du groupe, c'est-à-dire que pour chaque luminosité  $L_{19.5}$  on peut associer une masse du halo unique et inversement. À partir de ce postulat, on peut déterminer facilement la masse du halo à partir de l'index de la luminosité  $L_{19.5}$ . On choisit dans ce cas là d'utiliser seulement les groupes complets que l'on détermine en calculant la distribution en redshift de l'ensemble des groupes trouvés et en ne sélectionnant que les groupes qui ont un redshift inférieur à celui du maximum de la distribution. On s'assure ainsi d'avoir des groupes dont la distribution en luminosité n'est pas biaisée ce qui est nécessaire pour déterminer la masse du halo. Il faut pour cela connaître la fonction de masse des halos. Sa détermination repose sur la méthode décrite par Press and Schechter [13] mais légèrement modifiée par Warren et al. [14] pour correspondre au mieux aux résultats des simulations cosmologiques. Si on note  $n$  la densité de galaxies par unité de volume, on peut écrire la fonction de masse  $\phi(M) = dn/dM$  avec  $M$  la masse du halo (Tinker et al. [15]):

$$\phi(M) = \frac{dn}{dM} = f(\sigma) \frac{\bar{\rho}_m}{M} \frac{d\ln \sigma^{-1}}{dM} \quad (2.4.33)$$

avec  $\bar{\rho}_m$  la densité moyenne de matière dans l'Univers et  $\sigma(M)^2$  la variance en masse du champ de densité lissé. Si on connaît seulement  $\sigma$  pour  $z = 0$  alors on doit multiplier par le facteur de croissance  $D(z)$  qui peut s'exprimer comme (Carroll et al. [16]):

$$\begin{aligned} D(z) &= \frac{\delta_c(z)}{\delta_{c,0}} \\ &\approx \frac{\frac{5}{2}\Omega_m(z) \left[ \Omega_m(z)^{4/7} - \Omega_\Lambda(z) + (1 + \frac{1}{2}\Omega_m(z)) (1 + \frac{1}{70}\Omega_\Lambda(z)) \right]^{-1}}{\frac{5}{2}\Omega_{m,0} \left[ \Omega_{m,0}^{4/7} - \Omega_{\Lambda,0} + (1 + \frac{1}{2}\Omega_{m,0}) (1 + \frac{1}{70}\Omega_{\Lambda,0}) \right]^{-1}} \end{aligned} \quad (2.4.34)$$

avec pour l'évolution des paramètres cosmologiques avec le redshift:

$$\Omega_\Lambda(z) = \frac{\Omega_{\Lambda,0}}{E(z)^2} \quad \Omega_m(z) = \frac{\Omega_{m,0}(1+z)^3}{E(z)^2} \quad (2.4.35)$$

où la fonction  $E(z)$  est la même que celle définie précédemment dans la section (). On peut calculer la fonction  $\sigma(M)$  à partir d'une expression analytique issue d'une modélisation faîte par van den Bosch [17] sur des simulations cosmologiques:

$$\sigma(M) = \sigma_8 \frac{f(u)}{f(u_8)} \quad (2.4.36)$$

avec pour la fonction  $f$ :

$$f(u) = 64.087(1 + 1.074u^{0.3} - 1.581u^{0.4} + 0.954u^{0.5} - 0.185u^{0.6})^{-10} \quad (2.4.37)$$

et  $u$  et  $u_8$  qui s'expriment comme:

$$\begin{aligned} u &= 3.804 \times 10^{-4} \Gamma \left( \frac{Mh}{\Omega_{m,0}} \right)^{1/3} \\ u_8 &= 32\Gamma \\ \Gamma &= \Omega_{m,0}h \exp \left[ -\Omega_b(1 + \sqrt{2h}/\Omega_{m,0}) \right] \end{aligned} \quad (2.4.38)$$

Quant à elle la fonction  $f(\sigma)$  dans l'équation (2.4.33) s'exprime d'après Warren et al. [14] comme:

$$f(\sigma) = A(\sigma^{-a} + b)e^{-c/\sigma^2} \quad (2.4.39)$$

où les valeurs des coefficients sont détaillées dans la table (2.5). Il faut aussi calculer  $d \ln \sigma^{-1}/dM$

A	a	b	c
0.7234	1.625	0.2538	1.1982

Table 2.5: Valeurs des coefficients de la modélisation de  $f(\sigma)$  d'après Warren et al. [14].

mais comme on dispose d'une expression analytique de  $\sigma$  en fonction de  $M$  on peut facilement la déterminer. Il reste à calculer  $\bar{\rho}_m$  qui est la densité moyenne de matière dans l'Univers et qui s'écrit donc:

$$\bar{\rho}_m = \Omega_m(1+z)^3 \rho_c \quad (2.4.40)$$

où  $\rho_c$  est la densité critique de l'Univers et  $\Omega_m = \Omega_{m,0}$ . On obtient alors comme résultat la fonction de masse des halos visible sur la figure (2.22).

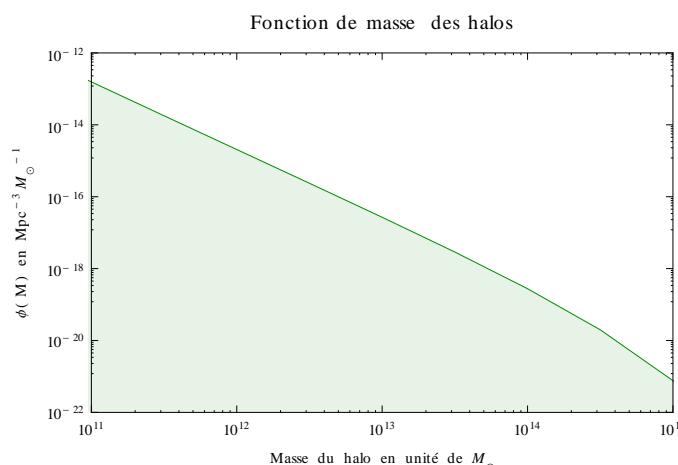


Figure 2.22: La fonction de masse des halos que l'on a calculé dans l'intervalle de valeurs physiques des masses des halos

Notre but de déterminer la masse du halo en fonction de la luminosité du groupe est alors rapide. En effet, si on note  $j$  l'index décrit précédemment du groupe, il s'agit du nombre de groupes ayant une luminosité supérieure à celle de ce groupe, donc on peut écrire:

$$j = \int_{L_{19.5}}^{\infty} \int_V \phi_L(L) dL dV = \int_{M_h}^{\infty} \int_V \phi(M) dM dV \quad (2.4.41)$$

où  $\phi_L$  est la fonction de luminosité des groupes et  $M_h$  la masse du halo qui a une luminosité  $L_{19.5}$  et un index  $j$ . Les intégrales portent sur le volume  $V$  mais en utilisant l'expression de  $dV/dz$  de la section (), on peut faire l'intégrale sur le redshift, les bornes de l'intégrale devenant plus simples car allant du  $z_{\min}$  du catalogue qui vaut 0.01 au  $z_{\max}$  déterminer en fonction du catalogue choisi (le mock ou le SDSS). L'intégrale se réécrit alors:

$$j = \int_{M_h}^{\infty} \int_{z_{\min}}^{z_{\max}} \frac{dV}{dz} \phi(M) dM dz \quad (2.4.42)$$

Il faut alors déterminer le  $M_h$  qui est solution de cette équation. On pourrait tenter de résoudre cette équation numériquement mais le plus simple est de calculer directement l'intégrale de l'équation (2.4.42) pour différentes valeurs de  $M_h$  réalistes avec nos données et d'ajuster par des splines cubiques les différentes valeurs de  $M_h$  en fonction de  $j$  (qui devient réel dans ce cas là car on ne tombe pas obligatoirement sur des entiers en calculant l'intégrale). Par la suite, si on veut déterminer  $M_h$  il suffit de trouver la valeur correspondante de la masse du halo à chacun des  $j$  par interpolation avec les splines cubiques.

On a ainsi déterminé la relation entre  $L_{19.5}$  et  $M_h$  pour les différents groupes obtenus à la suite des itérations et pour pouvoir appliquer la relation obtenue aux futures itérations, on ajuste à nouveau grâce aux splines cubiques la relation  $M_h/L_{19.5}$  versus  $L_{19.5}$  et la détermination de la masse du halo à l'étape ③ se fait par interpolation.

L'algorithme de Yang et al. [2] étant réalisé, on peut l'appliquer sur les différents catalogues à notre disposition pour voir les résultats que l'on obtient.

#### 2.4.4 Résultats et comparaison

On a d'abord appliqué cet algorithme sur le mock catalogue pour voir si les données obtenues étaient cohérentes avec celles indiquées par Yang et al. [2]. Tout d'abord la vérification que la première itération se déroule correctement. Pour cela on va réaliser les mêmes courbes que celles que l'on peut voir dans Yang et al. [2] pour les fractions de luminosités à cette itération. Elles sont présentées sur la figure (2.23). Ces fractions servent à déterminer la correction à appliquer aux luminosités des groupes pour compenser le manque de galaxies observées à cause de la magnitude limite d'observations de 17.77. Les résultats sont tout à fait semblables à ceux obtenus dans Yang et al. [2]. Les moyennes des fractions de luminosités pour les différentes luminosités limite sont bien modélisées par des exponentielles décroissantes en  $L_{19.5}$ . Donc, on estime que cette itération est bien réalisée. Cependant quand on regarde les mêmes graphes pour les itérations suivantes on ne retrouve pas le même genre de résultats. On a alors regardé le comportement de la relation masse-luminosité pour chaque itération (on en réalise cinq). Le résultat est présenté sur la figure (2.24). Pour les itérations impaires, le comportement de la relation est proche de celui auquel on s'attend, c'est-à-dire des masses proches de la centaine de fois la luminosité, les deux exprimés en unités solaires. Pour les autres itérations, le comportement est totalement différent de ce à quoi on s'attend. Pour l'instant, la cause de ces divergences n'a pas été élucidée, le problème n'apparaissant que d'une itération à l'autre. Il est fort probable que ce soit la cause d'un regroupement trop important des galaxies dans des groupes identiques, dû peut-être par une surestimation d'un des paramètres (rayon de viriel,...) qui pousse les galaxies à être affectées un même groupe. Cette idée est renforcée par le fait que pour les itérations *problématiques*, le nombre de groupes trouvé est très faible par rapport aux autres itérations, ce qui a tendance par la méthode de l'abundance matching à surestimer les masses des halos (ce que l'on voit car à une luminosité de  $10^{10}$  fois celle du Soleil, la masse pour les itérations à problèmes est de  $10^{14}$  donc bien plus élevée que dans le cas sans problèmes). La statistique des groupes est alors faussée et la méthode de l'abundance matching donne des masses élevées car il y a moins de groupes qui ont une luminosité supérieure à une luminosité donnée par rapport à ce que s'attend la fonction de masse des halos et considère donc qu'il s'agit de groupes très massifs car peu nombreux. Le travail pour déterminer la cause réelle de ce problème est en cours.

Depuis le problème a été résolu et est simple en fait à expliquer. La modélisation de la relation masse-luminosité s'effectuait à l'aide des splines cubiques ce qui semblait être le plus simple à mettre

Figure 2.23: Les fractions de luminosité pour des  $L_{\text{lim}}$  différents (c'est-à-dire des magnitudes  $M_r$  inférieures à une magnitude limite indiquée sur les figures). Il s'agit de la part des luminosités des groupes dont la contribution vient des galaxies avec  $M_r \leq M_{\text{lim}}$ . En orange, les fractions de luminosités pour chaque groupe, en rouge les moyennes de ces fractions dans des bins de  $L_{19.5}$  et en bleu la modélisation par une exponentielle décroissante de ces moyennes qui nous donne la fonction de correction des luminosités.

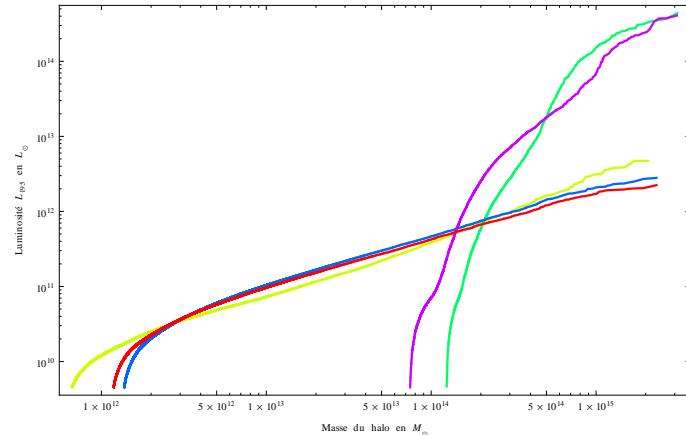


Figure 2.24: La relation  $L_{19.5}-M_h$  obtenue à partir du mock catalogue. En jaune la première itération, en vert la seconde, en bleu la troisième, en violet la quatrième et en rouge la dernière.

en œuvre et permettait de garder un maximum de précision sur la relation obtenue à partir des données. Mais en fait l'échantillonnage des points utilisés pour déterminer la relation n'était pas régulier en luminosité comme le nombre de groupes avec une luminosité donnée n'est pas le même suivant la luminosité considérée (il y a moins de groupes avec une forte luminosité, donc une grande masse). Dans ces domaines de luminosités sous-échantillonnées pour les splines cubiques, la relation obtenue à partir des coefficients trouvés dans ces régions produisait de fortes oscillations (des polynômes cubiques en fait) qui surestimaient grandement la masse des groupes quand on la détermine à partir de  $L_{19.5}$  dans les itérations suivantes (et pour certaines luminosités la sous-estimait aussi). Du coup des groupes avaient une masse trop grande par rapport à celle déterminée à partir de la *vraie* relation issue des données, ce qui avait tendance à *amener* beaucoup de galaxies dans ces groupes très massifs (donc à grand rayon de viriel). Le nombre de groupes obtenu était alors plus faible car des galaxies qui formaient peut-être des groupes de quatre ou cinq membres se retrouvaient incluses dans un seul et même groupe avec un halo très massif d'après la mauvaise modélisation de la relation masse-luminosité. L'abundance matching faussait alors complètement les masses des halos vu que les nouvelles luminosités des groupes étaient surestimées et que le nombre de groupes était faux lui aussi (ce que l'on voit sur la figure (2.24) où les points pour la deuxième itération ont tous des masses élevées).

Pour résoudre ce problème d'échantillonnage pour les splines cubiques, on a choisi d'utiliser une méthode complètement différente en modélisant directement la relation par un polynôme de degré 20. Un polynôme poussé à un tel degré est nécessaire pour modéliser au mieux les petites variations de la relation masse-luminosité engendrée par les données. Mais cette modélisation n'est pas parfaite et présente des fois des différences importantes à grand  $L_{19.5}$ . On pourrait pousser l'ordre du polynôme à des valeurs plus élevées mais cela pose des problèmes numériques qui empêchent la détermination des coefficients du polynôme. Même en se contentant de cela, en appliquant une simple méthode des moindres carrés sur les données pour trouver ces coefficients, des problèmes numériques persistent

## 2.4. L'ALGORITHME DE YANG ET AL.

## CHAPTER 2. GROUP FINDER ALGORITHM

car la matrice de dérivation partielle de la méthode des moindres carrés se retrouve mal conditionnée (si on calcule les valeurs propres de la matrice  $A$  de dérivation partielle dans ces cas là, des différences très importantes numériquement entre deux valeurs propres existent à cause du degré élevé choisi pour le polynôme qui peuvent amener à de grandes valeurs pour les coefficients de cette matrice  $A$ , et alors des pertes de précision *catastrophiques* peuvent survenir). On a donc fait appel à un sous-programme de la bibliothèque LAPACK qui utilise la méthode des moindres carrés mais qui élimine les points (donc les valeurs de la masse de halo et de la luminosité  $L_{19.5}$  dans notre cas) singuliers, ceci permettant de ne pas avoir une matrice mal conditionnée au final et d'obtenir des coefficients qui modélisent assez bien la relation masse-luminosité des données. On peut voir le résultat d'une telle modélisation sur la figure (2.25). Si on fait attention, on voit que pour les grandes luminosités la

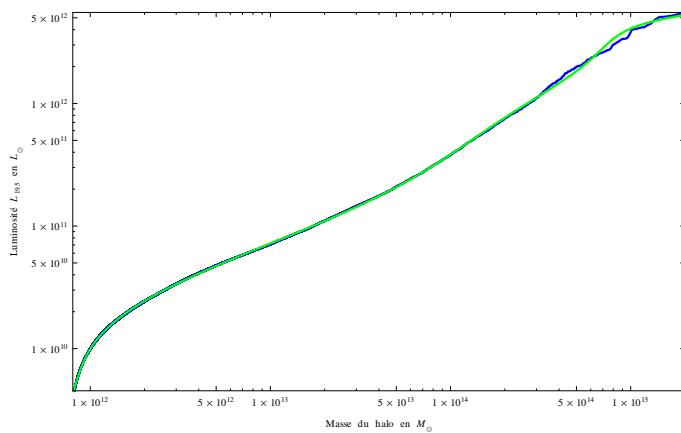


Figure 2.25: Résultat de la modélisation de la relation masse-luminosité par un polynôme de degré 20. En bleu la relation obtenue à partir des données de la première itération et en vert celle du polynôme. Les deux courbes sont assez proches l'une de l'autre sauf pour les grands  $L_{19.5}$  où des différences apparaissent.

modélisation n'est pas excellente mais il s'avère que pour notre situation cela suffit. Les masses des halos seront mal ajustées pour ces grandes luminosités mais ce n'est pas très grave car l'algorithme étant itératif, l'influence de ces écarts sera peu significatif au bout de quelques itérations, la relation masse-luminosité issue des données convergeant rapidement pour toutes les luminosités.

Une fois cette correction appliquée à l'algorithme de Yang et al. [2], on a appliqué à nouveau l'algorithme sur notre mock catalogue. Au niveau des fonctions de correction des luminosités le résultat est semblable à ceux présentés sur la figure (2.23). Par contre les résultats pour la relation masse-luminosité sont ceux auxquels on s'attendait pour l'algorithme.

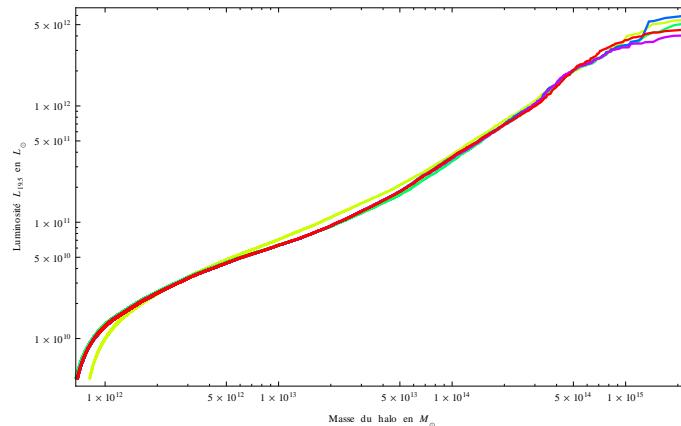


Figure 2.26: La relation  $L_{19.5}-M_h$  obtenue à partir du mock catalogue une fois changée la méthode de modélisation de cette relation. Le code couleur est le même que celui de la figure (2.24): en jaune la première itération, en vert la seconde, en bleu la troisième, en violet la quatrième et en rouge la dernière.

On observe sur la figure (2.26) que les itérations font que la relation masse-luminosité converge. En effet la courbe jaune de la première itération est celle qui s'écarte le plus des autres courbes ce qui est normal. La courbe verte de la seconde itération est elle déjà plus proche de la *vraie* relation. Finalement les trois dernières courbes des trois dernières itérations sont presque confondues car elles ont convergé. En regardant bien le graphe, on s'aperçoit quand même que la relation a du mal à converger pour les grandes luminosités (on distingue les courbes bleu et violette en dessous de la rouge qui est celle de la dernière itération pour les grandes masses). Ceci est due au fait que la relation masse-luminosité est mal modélisée par le polynôme de degré 20 pour les grandes luminosités (et donc les grandes masses). La convergence est alors plus difficile pour cette gamme de luminosité. Mais tout ceci s'estompe après quelques itérations (ici cinq itérations permettent d'avoir une bonne convergence et donc d'obtenir des masses de halos que l'on peut supposer fiables).

## 2.5 Notre algorithme

On vient de voir la méthode de Yang et al. [2] pour déterminer les groupes de galaxies dans un catalogue donné. Mais certaines des hypothèses réalisées pour faire cette détermination ne sont pas satisfaisantes, car trop simplistes ou ne collant pas particulièrement aux résultats des simulations cosmologiques ou des surveys. C'est pourquoi nous avons souhaité réaliser un algorithme différent qui gérera mieux certains aspects qui seront décrits plus loin.

### 2.5.1 Description

Nous allons tout d'abord voir quel est le principe du nouvel algorithme et ensuite quelles sont les différences entre cet algorithme et celui de Yang et al. [2].

#### Principe

La détermination des groupes sera organisée de la façon suivante:

- ① Déterminer les groupes potentiels. Comme dans le Yang et al. [2], on doit avoir des groupes initiaux afin de calculer les diverses caractéristiques qui vont nous aider à attribuer de façon sûre les groupes. Pour cela on va se servir des masses stellaires des galaxies du catalogue utilisé pour la détermination des groupes. La galaxie centrale d'un groupe a de fortes chances d'être la plus

## 2.5. NOTRE ALGORITHME

## CHAPTER 2. GROUP FINDER ALGORITHM

lumineuse également (comme dans le cas du Yang et al. [2] où on utilise le barycentre lumineux ce qui revient à peu près à la même définition). S'il s'agit de la plus lumineuse, il s'agit aussi de la plus massive en masse stellaire selon les relations masse-luminosité. Donc pour trouver les groupes potentiels on appliquera la méthode suivante: rechercher les galaxies qui ont les masses stellaires les plus grandes dans le catalogue, on considère alors qu'elles sont le centre d'un groupe potentiel. À partir de cette galaxie et de sa masse stellaire, on peut déterminer un rayon de viriel potentiel du groupe à l'aide de la masse du halo trouvée par abundance matching entre la fonction de distribution des masses de halo et de sous-halo et celle des masses des galaxies centrales des groupes, ce qui nous permettra de sélectionner les autres galaxies du catalogue qui appartiennent à ce groupe potentiel. À cause de l'imprécision de la détermination du rayon de viriel du fait que le groupe n'est *que potentiel*, on choisit de prendre en compte les galaxies qui sont à 2 ou 3 fois le rayon de viriel pour bien *récupérer* toutes les galaxies qui pourraient y appartenir.

② Affecter les galaxies aux groupes. Dans cet algorithme, on choisit d'associer une probabilité d'appartenance à un groupe pour chaque galaxie. Cette probabilité est calculée selon la méthode décrite dans la section (2.5.2). À partir de cette appartenance au groupe, on calcule les différentes propriétés des groupes comme la masse stellaire et la luminosité pour déterminer les différentes relations entre la masse du halo et les différentes caractéristiques des groupes (détaillée dans la suite).

③ À l'aide de ces groupes potentiels on calcule la relation entre la masse stellaire des groupes et celle du halo par l'abundance matching entre cette fois la distribution de masse des halos et la distribution de masse stellaire des groupes. Ainsi à partir de cette relation, on peut déterminer la masse du halo qui servira à fournir une estimation du rayon de viriel. Alors on peut itérer pour rechercher à nouveau les membres des groupes potentiels avec ce nouveau rayon de viriel et ainsi de suite jusqu'à obtenir une convergence dans la relation masse stellaire-masse du halo.

④ Ce travail est réalisé pour différents sous-échantillons du catalogue limités en flux et limités en volume afin de diminuer les effets d'éventuels biais liés à la luminosité et à la taille du catalogue, et obtenir ainsi des catalogues doublement complets comme sur la figure (2.27).

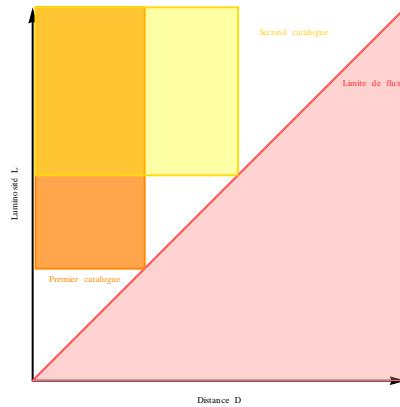


Figure 2.27: Représentation des catalogues doublement complets. La zone rouge correspond aux luminosités inaccessibles à une distance  $D$  donnée de l'observateur à cause de la limite de flux du survey. Deux catalogues doublement complets sont représentés en orange et en jaune. Par les limites de volume et de luminosité imposés, aucune galaxie ne doit manquer, c'est-à-dire entrer dans la zone en rouge et aucun biais n'est introduit.

On peut maintenant voir quelles sont les principales différences entre notre algorithme et celui de Yang et al. [2].

## Différences

On détaille maintenant les différences entre les deux algorithmes. Tout d'abord, pour les groupes potentiels on utilise un FoF pour les trouver dans Yang et al. [2] alors que l'on recherche les galaxies les plus massives qui sont éventuellement au centre d'un groupe potentiel et les galaxies qui lui sont associées dans une sphère virielle large afin de ne pas manquer de galaxies.

Dans le Yang et al. [2], la densité de contraste est ce qui sert de critère pour l'appartenance d'une galaxie à un groupe en définissant un seuil pour cette valeur qui indique si la galaxie est liée au groupe ou non. Même si une galaxie a des valeurs de cette densité de contraste pour d'autres groupes assez proches de celle du groupe avec lequel elle est liée, c'est la valeur maximale qui impose l'appartenance à un groupe unique. Ce choix est un peu arbitraire alors que cette galaxie pourrait appartenir éventuellement à un autre groupe. C'est pourquoi dans notre algorithme on calcule une probabilité d'appartenance au groupe. Ainsi lorsque l'on cherche à déterminer les propriétés des groupes comme dans le Yang et al. [2] pour la luminosité et la masse stellaire, on fera la sommation sur les membres d'un groupe non pas simplement mais avec des poids liés à la probabilité qu'une galaxie appartienne à ce groupe. Par exemple pour la masses stellaire du groupe  $M_{\text{groupe}}^*$  on aura:

$$M_{\text{groupe}}^* = \frac{\sum_i p_i M_i^*}{\sum_i p_i} \quad (2.5.1)$$

avec  $p_i$  la probabilité de la galaxie de masse stellaire  $M_i^*$  d'appartenir au groupe. On normalise par la somme des probabilités car elle n'est pas obligatoirement égale à 1.

Pour la détermination des relations qui permettent de d'obtenir les caractéristiques des groupes, dans le Yang et al. [2], on cherche la relation entre la masse du halo et la luminosité du groupe par la méthode de *l'abundance matching* comme décrit plus haut qui fait le lien entre la distribution des masses des halos et celle des luminosités des groupes. Cela repose sur l'hypothèse que pour chaque masse de halo est associée une unique luminosité de groupe, ce qui n'est pas vraiment le cas dans la réalité et d'après les simulations cosmologiques comme l'indiquent Yang et al. [2] dans leur article. Par contre la relation entre la masse stellaire des groupes et la masse des halos répond un peu mieux à ce critère. C'est pourquoi dans notre algorithme on déterminera la masse du halo à partir de la masse stellaire du groupe que l'on aura obtenue.

Et finalement dans Yang et al. [2], les catalogues utilisé et généré sont limités en flux ce qui évite d'obtenir des biais liés à la luminosité, et dans notre algorithme, on obtiendra des catalogues limités en flux et en volume ce qui devrait permettre de réduire encore plus les biais du même type dans notre estimation des groupes.

Les points sur lesquels notre algorithme devrait être meilleur sont les suivants:

- L'affectation des galaxies aux groupes car on définit une probabilité et non une appartenance unique à un groupe. De plus le critère de la densité de contraste du Yang et al. [2] est calculée à partir de l'idée d'une séparation des effets liés au rayon  $\Sigma(R)$  et des effets liés à la dispersion de vitesse  $p(\Delta z)$  pour la dé-projection, ainsi qu'une dispersion de vitesse  $\sigma$  indépendante du rayon dans le groupe. Le calcul de la probabilité d'appartenance au groupe d'une galaxie prend mieux en compte ces aspects avec les calculs de Mamon et al. [1] comme on le décrit plus loin.
- L'utilisation de la masse stellaire du groupe pour déterminer la masse du halo devrait aussi permettre d'obtenir de meilleurs résultats car la masse du halo sera mieux déterminée en principe qu'avec la luminosité du groupe comme indiqué dans Yang et al. [2].

On va maintenant décrire comment mettre en œuvre notre algorithme.

### 2.5.2 Réalisation

On va détailler le principe du calcul de la probabilité qui va nous servir à affecter les galaxies aux groupes. Quand on cherche la probabilité qu'une galaxie appartienne à un halo, en fait on cherche la probabilité qu'une galaxie que l'on voit dans le cône viriel associé au halo soit effectivement dans le halo et ne soit pas un *interloper*, c'est-à-dire une galaxie qui n'appartient pas au groupe mais qui vient polluer le champ de vision de l'observateur dans le cône comme dans la figure (2.28). Il se peut alors que cet interloper soit pris pour une galaxie qui appartient au groupe alors que non. C'est pourquoi on définit comme suit la probabilité  $p$  d'appartenir à un groupe:

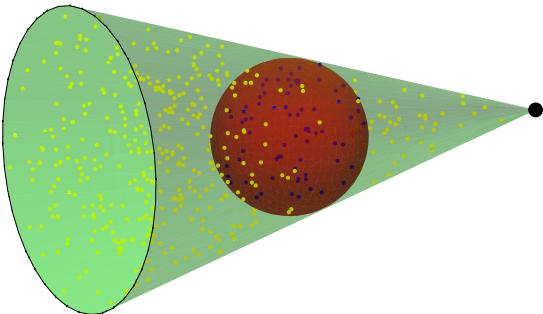


Figure 2.28: Représentation du cône viriel en vert associé à la sphère virielle en rouge. Les points colorés représentent des galaxies qui appartiennent au halo en bleu et qui sont des interlopers en jaune. L'observateur est situé à la place du point noir.

$$p(R, v_z) = \frac{g_h(R, v_z)}{g_h(R, v_z) + g_i(R, v_z)} \quad (2.5.2)$$

où  $g_h(R, v_z)$  est la densité de galaxie dans le halo et  $g_i(R, v_z)$  la densité d'interlopers dans l'espace projeté, avec  $R$  le rayon projeté et  $v_z$  la norme de la vitesse suivant la ligne de visée. D'après Mamon et al. [1] la meilleure modélisation possible de la densité des interlopers est:

$$g_i(R, |v_z|) = A \exp \left[ -\frac{1}{2} \left( \frac{v_z}{\sigma_i} \right)^2 \right] + B \quad (2.5.3)$$

Les coefficients ont été obtenus à l'aide d'un maximum de vraisemblance sur les données d'une simulation cosmologique (Mamon et al. [1]). Pour le cas du halo, la fonction peut s'exprimer comme:

$$\begin{aligned} g_h(R, v_z) &= \Sigma(R) \langle h(v_z|R, r) \rangle_{\text{los}} \\ &= 2 \int_R^{r_v} \frac{r \nu(r)}{\sqrt{r^2 - R^2}} h(v_z|R, r) dr \end{aligned} \quad (2.5.4)$$

avec  $h(v_z|R, r)$  la distribution des vitesses suivant la ligne de visée à la position  $(R, r)$  et  $r^2 = z^2 + R^2$  où  $z$  est le redshift. L'intégration ne se fait pas jusqu'à l'infini pour  $\Sigma(R)$  car pour rester dans le halo et ne pas prendre en compte les interlopers, on garde les bornes d'intégration sur la sphère virielle.

La distribution des vitesses est supposée suivre une distribution gaussienne en 3D. Ceci aboutit pour la distribution des vitesses suivant la ligne de visée à:

$$h(v_z|R, r) = \frac{1}{\sqrt{2\pi\sigma_z^2(R, r)}} \exp\left[-\frac{v_z^2}{2\sigma_z^2(R, r)}\right] \quad (2.5.5)$$

avec

$$\sigma_z^2(R, r) = \sigma_r^2(r) \left(1 - \beta(r) \left(\frac{R}{r}\right)^2\right) \quad (2.5.6)$$

où  $\beta(r)$  est l'anisotropie des vitesses. Pour la déterminer, on assume un modèle de Mamon and Łokas [18] (appelé ML) qui nous donne:

$$\beta(r) = \beta_{ML}(r) = \frac{1}{2} \frac{r}{r+a} \quad (2.5.7)$$

Il reste à déterminer  $\sigma_r(r)$  et qui vaut:

$$\sigma_r^2(r) = \frac{1}{\nu(r)} \int_r^{r_v} K_r(r, s) \nu(s) \frac{GM(s)}{s^2} ds \quad (2.5.8)$$

avec

$$K_r(r, s) = \exp\left[2 \int_r^s \beta(t) \frac{dt}{t}\right] = \frac{s+a}{r+a} \quad (2.5.9)$$

pour le modèle ML. Pour déterminer correctement  $\sigma_r(r)$ , on doit utiliser un modèle pour la densité de masse des halos. Or le problème est que les différents modèles qui ont été utilisés pour réaliser le calcul de la densité de galaxies dans le halo ne permettent pas de retrouver les résultats obtenus directement à partir des données de la simulation utilisée par [1] pour modéliser la densité des interlopers. On va détailler les différents modèles utilisés et visualiser les résultats que l'on obtient.

### Le modèle NFW

Il se trouve que pour les halos, la fonction qui la modélise au mieux est le modèle de Navarro et al. [9] (NFW) qui peut s'écrire sous la forme (Binney and Tremaine [19]):

$$\rho(r) = \frac{\rho_0}{(r/a)^\alpha (1+r/a)^{\beta-\alpha}} \quad (2.5.10)$$

avec  $(\alpha, \beta) = (1, 3)$  pour le modèle NFW et  $a$  le rayon caractéristique. On peut alors avoir facilement accès à la masse du halo à un  $r$  donné en intégrant cette fonction sur le volume:

$$\begin{aligned} M(r) &= \iiint_V \rho(r') dV \\ &= \int_0^{r/a} 4\pi \rho(r') r'^2 dr' = 4\pi \rho_0 \left( \ln(1+r/a) - \frac{r/a}{1+r/a} \right) \end{aligned} \quad (2.5.11)$$

De la même façon on peut calculer  $\nu(r)$  car  $\nu(r) \propto \rho(r)$  par définition et donc on peut écrire:

$$\nu(r) = \frac{cste}{r(1+r/a)^2} \quad (2.5.12)$$

## 2.5. NOTRE ALGORITHME

## CHAPTER 2. GROUP FINDER ALGORITHM

On peut alors réaliser tous les calculs pour pouvoir calculer la probabilité d'appartenance à un groupe pour toutes les galaxies.

On commence par le calcul de la dispersion de vitesse  $\sigma_r^2(r)$  qui s'avère être le plus coûteux en temps CPU si on n'a pas accès à une solution analytique pour le modèle choisi. On obtient  $\sigma_r^2(r)$  suivant l'expression (2.5.8) à partir de l'équation de Jeans appliquée pour la dispersion de vitesse:

$$\frac{d(\nu\sigma_r^2)}{dr} + \frac{2\beta(r)}{r} (\nu\sigma_r^2) = -\nu \frac{GM(r)}{r^2} \quad (2.5.13)$$

Pour obtenir une expression de la dispersion de vitesse facilement utilisable dans tous les cas sans devoir réaliser différentes intégrales, on va essayer d'obtenir une dispersion de vitesse normalisée, ne dépendant que de paramètres eux-mêmes normalisés. On va exprimer pour cela la fonction de la masse dans le halo à un rayon donné avec une fonction elle aussi normalisée et utiliser la même chose pour les fonctions de la densité de masse, densité en nombre et nombre de galaxies à un rayon donné dans le halo. On écrit donc:

$$\begin{aligned} M(r) &= M(a)\tilde{M}(r/a) \\ N(r) &= N(a)\tilde{N}(r/a) \\ \rho(r) &= \frac{M(a)}{4\pi a^3}\tilde{\rho}(r/a) \\ \nu(r) &= \frac{N(a)}{4\pi a^3}\tilde{\nu}(r/a) \end{aligned} \quad (2.5.14)$$

avec:

$$\tilde{M}(x) = \tilde{N}(x) = \frac{\ln(1+x) - \frac{x}{x+1}}{\ln 2 - (1/2)} \quad (2.5.15)$$

et pour les densités:

$$\tilde{\rho}(x) = \tilde{\nu}(x) = \frac{1}{\ln 2 - (1/2)} \frac{1}{x(1+x)^2} \quad (2.5.16)$$

En réinjectant l'ensemble de ces dernières équations dans l'équation (2.5.8), on obtient l'expression suivante:

$$\sigma_r^2(r) = \frac{GM(a)}{a} \tilde{\sigma}_r^2(r/a) \quad (2.5.17)$$

avec pour l'expression de la dispersion de vitesse *normalisée*:

$$\tilde{\sigma}_r^2(x) = \frac{1}{\tilde{\nu}(x)} \int_x^\infty \frac{S+1}{x+1} \frac{\tilde{\nu}(S)\tilde{M}(S)}{S^2} dS \quad (2.5.18)$$

On peut maintenant écrire une expression de forme semblable pour la dispersion de vitesse suivant la ligne de visée:

$$\sigma_z^2(R, r) = \frac{GM(a)}{a} \tilde{\sigma}_z^2(R/a, r/a) \quad (2.5.19)$$

avec:

$$\tilde{\sigma}_z^2(x_R, x) = \left(1 - \frac{1}{2} \frac{x_R^2}{(x(1+x))}\right) \tilde{\sigma}_r^2(x) \quad (2.5.20)$$

D'après *Mathematica*, la dispersion de vitesse normalisée admet une expression analytique:

$$\begin{aligned}\tilde{\sigma}_r^2(x) = & \frac{1}{3x(-1 + \ln 4)} \left( x \left( -3 + x \left( -9 + \pi^2(1 + x) \right) \right) \right. \\ & + 3x^3 \ln \left( 1 + \frac{1}{x} \right) \\ & + 3 \ln(1 + x) \left( 1 - x + x^2(1 + x) \ln(1 + x) \right) \\ & \left. - 3x^2 \ln(x(1 + x)) + 6x^2(1 + x)Li_2(-x) \right)\end{aligned}\quad (2.5.21)$$

où la fonction  $Li_2(x)$  est la fonction dilogarithme telle que:

$$Li_2(z) = - \int_0^1 \frac{\ln(1 - zt)}{t} dt \quad (2.5.22)$$

Du coup on peut calculer  $g_h$  avec les nouvelles expressions de ces fonctions. On peut alors effectuer un changement de variables dans l'équation (2.5.4) pour éviter des problèmes lors de l'intégration numérique qui sera réalisée dans le programme. On pose  $r = R \cosh u$  ce qui permet de récrire (2.5.4) ainsi:

$$g_h(R, r_{\text{vir}}, v_z) = \frac{N(a)}{4\pi a^2} \sqrt{\frac{2a}{\pi GM(a)}} \times I(r_{\text{vir}}/a, R/a, v_z/v_v) \quad (2.5.23)$$

avec:

$$\begin{aligned}I(c, x_R, \hat{v}_z) = & \int_0^{\text{acosh}(\frac{c}{x_R})} \frac{(x_R \cosh u) \tilde{\nu}(x_R \cosh u)}{\tilde{\sigma}_z(x_R, x_R \cosh u)} \\ & \times \exp \left( -\frac{\tilde{M}(c) \hat{v}_z^2}{2c \tilde{\sigma}_z^2(x_R, x_R \cosh u)} \right) du\end{aligned}\quad (2.5.24)$$

On a donc en notre possession l'expression de  $g_h$  pour calculer la probabilité d'appartenance à un halo d'une galaxie avec un rayon projeté  $R$  et une vitesse suivant la ligne de visée  $v_z$ . Mais si on examine bien l'expression (2.5.2), on voit que pour obtenir la probabilité on a juste besoin de connaître le rapport  $g_i/g_h$ , ce qui nous amène à obtenir l'expression de ce rapport.

D'après (2.5.3) tirée de Mamon et al. [1], on a:

$$g_i(R, v_z) = \frac{1}{2} g_i(R, |v_z|) = \frac{1}{2} \hat{g}_i(R/r_{\text{vir}}, |v_z|/v_{\text{vir}}) \frac{N_v}{r_{\text{vir}}^2 v_{\text{vir}}} \quad (2.5.25)$$

où  $\hat{g}_i$  a la même expression que  $g_i$  dans l'équation (2.5.3) et les mêmes coefficients mais cette fois sans unité.  $N_v$  est le nombre de galaxies contenues dans la sphère viruelle et  $v_{\text{vir}}$  la vitesse viruelle qui s'exprime comme  $GM_v/r_{\text{vir}}$  où  $M_v$  est la masse viruelle. On indexera désormais par  $v$  les propriétés relatives à la sphère viruelle. On peut simplifier le rapport  $g_i/g_h$  qui devient:

$$\frac{g_i}{g_h}(R, v_z) = \frac{N_v \hat{g}_i(R/r_v, |v_z|/v_v)}{2r_v^2 v_v} \frac{4\pi a^2}{N(a)} \sqrt{\frac{\pi GM(a)}{2a}} I(r_v/a, R/a, v_z/v_v)^{-1} \quad (2.5.26)$$

## 2.5. NOTRE ALGORITHME

## CHAPTER 2. GROUP FINDER ALGORITHM

Compte tenu du fait que  $N_v = N(a)\tilde{N}(c)$  avec  $c = r_v/r_{200}$  et que l'on peut assimiler  $r_{200}$  à  $a$  le rayon de pente  $-2$  dans le modèle NFW car les valeurs sont proches à moins de  $10\%$  près, alors:

$$\frac{g_i}{g_h}(R, v_z) = \sqrt{\frac{\tilde{N}(c)2\pi^3}{c^3}} \hat{g}_i(R/r_v, |v_z|/v_v) I(r_v/a, R/a, v_z/v_v)^{-1} \quad (2.5.27)$$

À partir de là, on peut déterminer la probabilité d'appartenir au halo si on est une galaxie en fonction de  $R$  et de  $v_z$ . Il est intéressant alors de comparer les résultats obtenus avec notre expression de la probabilité et ceux de la fraction des interlopers dans le cas de la simulation utilisée dans Mamon et al. [1]. Pour cela on va tracer les contours d'iso-probabilité dans le plan de l'espace des phases pour différentes valeurs de  $R$  et  $v_z$  contenues dans la sphère virielle. Les résultats obtenus sont visibles sur la figure (2.29).

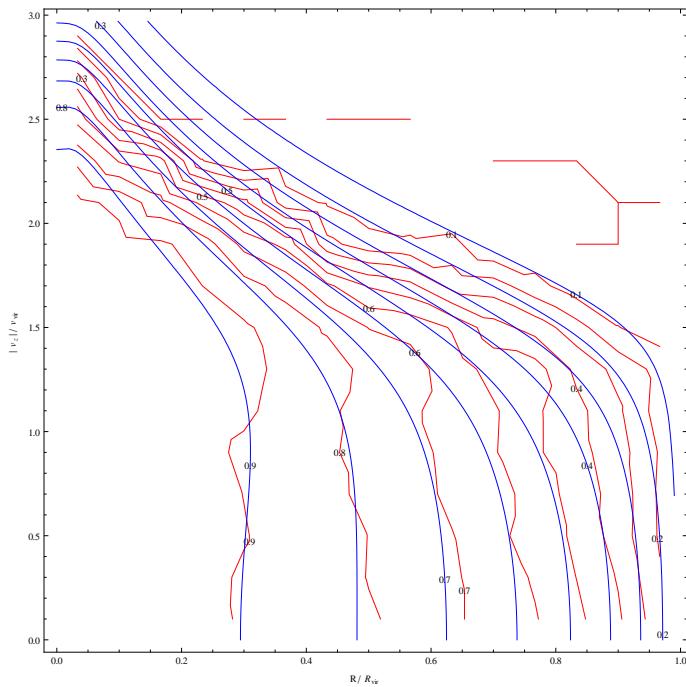


Figure 2.29: Comparaison des probabilités obtenues par le calcul réalisé avec le modèle NFW en bleu sur la figure et les données issues directement de la simulation utilisée dans Mamon et al. [1] en rouge. Le rayon projeté en abscisse est en unité de  $r_v$  et la vitesse sur la ligne de visée en unité de vitesse virielle  $v_v$ . Les différences existent principalement aux faibles rayons projetés et aux grandes vitesses.

Les différences jouent principalement sur les faibles rayons projetés et les grandes vitesses ce qui laisse supposer qu'il s'agit d'une mauvaise estimation de la dispersion de vitesse suivant la ligne de visée qui provoque ce genre de *déviation*. On a donc fait les mêmes comparaisons entre les expressions des paramètres selon le modèle et les données de la simulation. Les calculs sont faits pour différents rayons projetés et distances au centre du halo et présentés sur la figure (2.30).

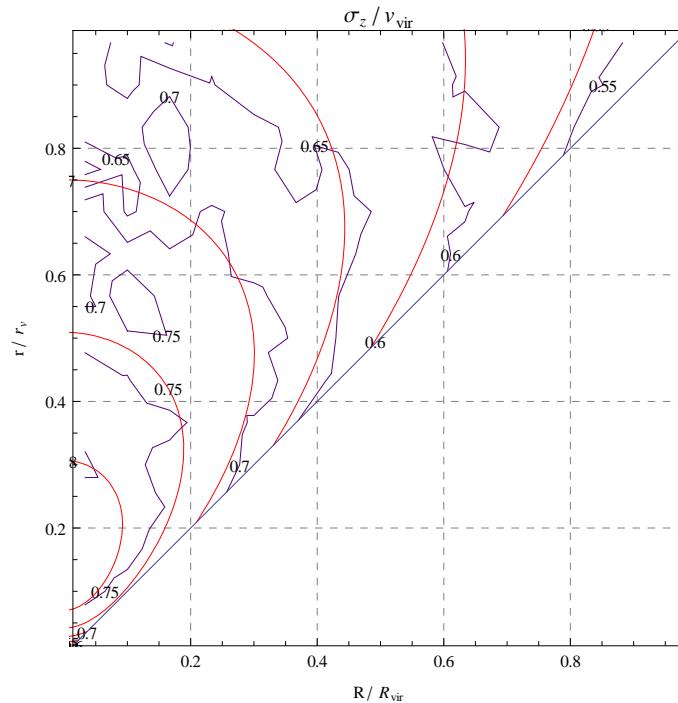


Figure 2.30: Comparaison de la dispersion de vitesse suivant la ligne de visée entre le résultat issu du modèle NFW en rouge et celui obtenu avec les données de la simulation utilisée par Mamon et al. [1] en violet. La dispersion de vitesse est exprimée en unité de vitesse virielle  $v_v$ . Les résultats sont semblables pour les petits rayons dans les deux cas mais divergent sur les bords de la sphère virielle.

Bien que très semblables, les résultats ne sont pas parfaitement identiques, ce qui nous pousse à tenter un autre modèle pour le calcul de la probabilité: le modèle d'Einasto.

### Le modèle d'Einasto

Ce modèle se présente sous cette forme initialement:

$$\rho(r) = \rho_0 \exp\left(-\left(\frac{r}{b}\right)^{1/m}\right) \quad (2.5.28)$$

Le paramètre  $m$  est ajustable selon le type de modèle d'Einasto que l'on souhaite utiliser. Le rayon caractéristique  $b$  n'a pas de sens particulier pour l'instant ce qui amène à l'exprimer en fonction du rayon de pente  $-2$  ce qui permet de le noter  $a$  comme dans le modèle NFW. En exprimant la dérivée en logarithmique de cette densité, on trouve la relation entre les deux rayons:

$$\left(\frac{1}{b}\right)^{1/m} = 2m \left(\frac{1}{a}\right)^{1/m} \quad (2.5.29)$$

ce qui permet de récrire:

$$\rho(r) = \rho_0 \exp\left(-2m \left(\frac{r}{a}\right)^{1/m}\right) \quad (2.5.30)$$

En intégrant cette relation sur le volume, on obtient la masse à un rayon donné qui s'exprime comme:

$$M(r) = \frac{M_\infty}{\Gamma(3m)} \gamma\left(3m, 2m \left(\frac{r}{a}\right)^{1/m}\right) \quad (2.5.31)$$

## 2.5. NOTRE ALGORITHME

## CHAPTER 2. GROUP FINDER ALGORITHM

où:

$$\gamma(a, x) = \int_0^x t^{a-1} \exp(-t) dt \quad (2.5.32)$$

est la fonction gamma incomplète qui s'exprime aussi:

$$\gamma(a, x) = \Gamma(a) - \Gamma(a, x) \quad (2.5.33)$$

en utilisant les fonctions déjà décrites précédemment.

La densité peut être réécrites en fonction de paramètres liés à  $a$ , ce qui nous donne alors:

$$\rho(r) = \frac{(2m)^{3m}}{m\gamma(3m, 2m)} \exp\left(-2m\left(\frac{r}{a}\right)^{1/m}\right) \frac{M(a)}{4\pi a^3} \quad (2.5.34)$$

Comme dans le cas du modèle NFW, on peut exprimer la dispersion de vitesse à l'aide de fonction adimensionnée et dépendant de paramètres sans dimension eux aussi. Avec les expressions précédentes de la masse à un rayon donné et de la densité (massique et en nombre ne différant que par le facteur  $M(a)$  ou  $N(a)$  selon le cas que l'on traite), on obtient:

$$\sigma_z^2(R, r) = \frac{GM_\infty}{a\Gamma(3m)} \tilde{\sigma}_z^2(R/a, r/a) \quad (2.5.35)$$

avec:

$$\begin{aligned} \tilde{\sigma}_z^2(x_R, x) &= e^{2mx^{1/m}} \left(1 - \frac{1}{2} \frac{x}{1+x} \left(\frac{x_R}{x}\right)^2\right) \\ &\times \int_x^\infty \left(\frac{S+1}{x+1}\right) \frac{\gamma(3m, 2mS^{1/m})}{S^2} e^{-2mS^{1/m}} dS \end{aligned} \quad (2.5.36)$$

En utilisant la même astuce pour l'intégration numérique dans le cas du modèle NFW, on fait un changement de variable dans l'intégrale de l'expression de  $g_h$ . On pose alors comme avant  $r = R \cosh u$  et aussi pour faciliter les écritures  $\alpha = GM_\infty / (a\Gamma(3m))$ . L'équation (2.5.4) s'écrit dans ce cas là:

$$g_h(R, v_z) = \sqrt{\frac{2}{\alpha\pi}} \frac{(2m)^{3m}}{m\gamma(3m, 2m)} \frac{N(a)}{4\pi a^2} I(R/a, r_v/a, v_z/v_v) \quad (2.5.37)$$

En remarquant que:

$$M_\infty = \frac{\Gamma(3m) M_v}{\gamma(3m, 2mc^{1/m})} \quad (2.5.38)$$

on trouve que  $\alpha$  se récrit:

$$\alpha = \frac{GM_v}{a\gamma(3m, 2mc^{1/m})} \quad (2.5.39)$$

toujours avec  $c = r_v/a$ . Donc:

$$\begin{aligned} I(c, x_R, \hat{v}_z) &= \int_0^{\text{acosh}(\frac{c}{x_R})} \frac{x_R \cosh u}{\tilde{\sigma}_z(x_R, x_R \cosh u)} \\ &\times \exp\left(-2m(x_R \cosh u)^{1/m}\right) \\ &\times \exp\left(-\frac{\gamma(3m, 2mc^{1/m}) \hat{v}_z^2}{2c\tilde{\sigma}_z^2(x_R, x_R \cosh u)}\right) du \end{aligned} \quad (2.5.40)$$

Le rapport des densités comme dans le modèle NFW peut alors s'écrire:

$$\frac{g_i}{g_h}(R, v_z) = \hat{g}_i\left(\frac{R}{r_v}, \frac{|v_z|}{v_v}\right) \frac{m}{(2m)^{3m}} \sqrt{\frac{2\pi^3 \gamma(3m, 2mc^{1/m})}{c^3}} \\ \times I(R/a, r_v/a, v_z/v_v)^{-1} \quad (2.5.41)$$

On peut maintenant réaliser les contours de la probabilité d'appartenance au halo comme pour le modèle NFW. Le problème est que la dispersion de vitesse suivant la ligne de visée n'a pas d'expression analytique comme pour NFW, et nous oblige donc à réaliser à chaque fois l'intégrale numériquement pour aboutir à un résultat. Ceci pouvant s'avérer très coûteux en temps CPU, il est nécessaire de trouver une modélisation pour  $\tilde{\sigma}_r^2(x)$  afin de rendre le calcul plus rapide. On choisit de la modéliser par une fonction rationnelle de polynômes d'ordre 16 au numérateur et d'ordre 17 au dénominateur. La modélisation résultante est précise à  $1 \times 10^{-5}\%$  près. Les résultats obtenus sont présentés sur la figure (2.31) en comparaison avec les résultats précédents de NFW.

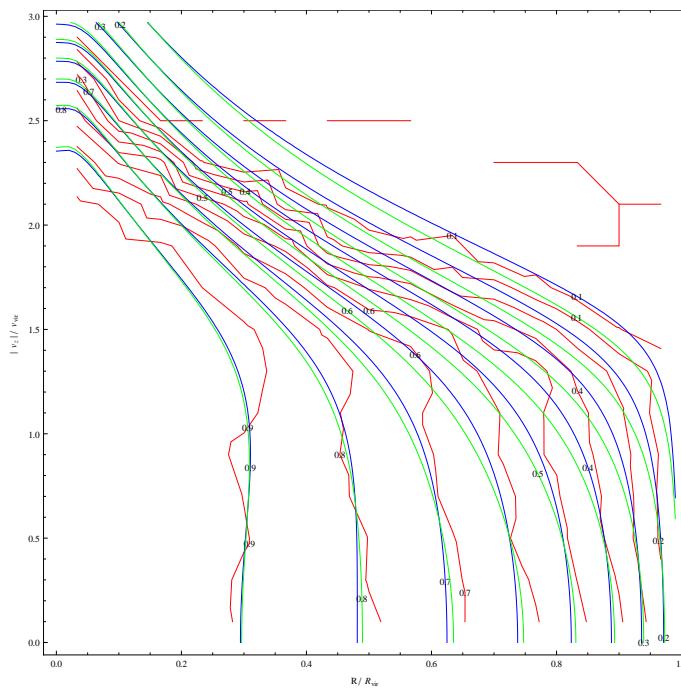


Figure 2.31: Contours de probabilité dans l'espace des phases avec le modèle d'Einasto ( $m = 5$ ) en vert, comparé avec les données de la simulation en rouge et le modèle NFW en bleu.

Les résultats n'étant toujours pas satisfaisant pour ce modèle, on a essayé de voir l'influence du choix du paramètre  $m$  du modèle d'Einasto sur le calcul de la probabilité. Sur la figure (2.32) sont visibles les contours de la probabilité d'appartenance au halo pour 4 valeurs différentes de  $m$ . L'influence de  $m$  sur la probabilité ne permet pas de correspondre mieux aux données de la simulation utilisée par Mamon et al. [1]. C'est pourquoi on essaie cette fois de modéliser directement la densité de particules de la simulation pour voir si les résultats sont meilleurs qu'avec les modèles décrits plus haut.

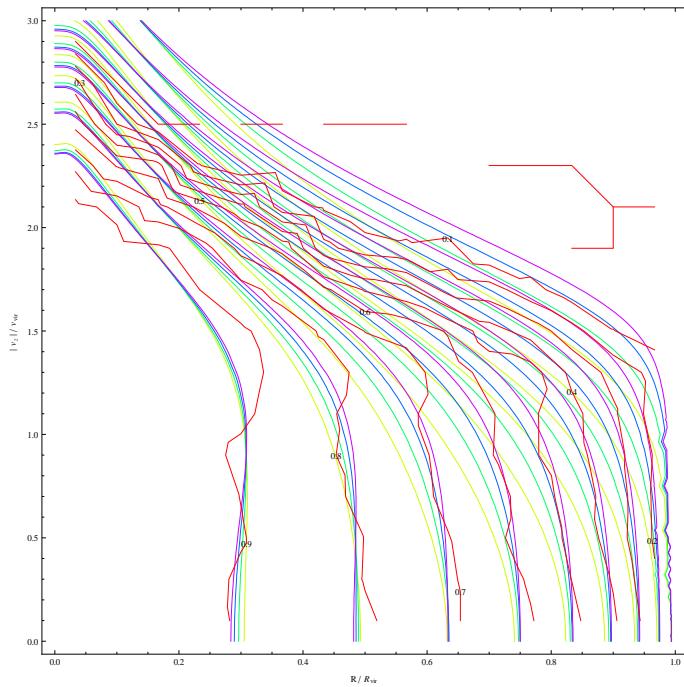


Figure 2.32: Contours de probabilité du modèle d'Einasto pour différentes valeurs du paramètre  $m$ . En jaune pour  $m = 4$ , en vert pour  $m = 5$ , en bleu pour  $m = 6$  et en violet pour  $m = 7$ . L'influence du paramètre est faible sur le calcul de la probabilité et n'aide pas à améliorer les résultats par rapport aux données de la simulation toujours en rouge sur la figure.

### Le modèle issu directement de la simulation

Ce que l'on peut faire à partir des données de la simulation c'est obtenir la densité en fonction de  $r/r_v$ . Donc on compte le nombre de galaxies présentes entre les deux sphères de rayon  $r/r_v$  et  $(r/r_v) + d(r/r_v)$ . On peut donc écrire:

$$dN = 4\pi \left(\frac{r}{r_v}\right)^2 \hat{\nu} \left(\frac{r}{r_v}\right) d\left(\frac{r}{r_v}\right) \quad (2.5.42)$$

où la fonction  $\hat{\nu}$  est la fonction que l'on détermine à partir des données. Son allure est celle de la figure (2.33). Pour modéliser cette densité, on choisit de prendre un modèle de la forme:

$$\hat{\nu}(x) = \frac{a}{x^\alpha(1+x)^{\beta-\alpha}} \quad (2.5.43)$$

en s'inspirant du modèle général dont est issu le modèle NFW. Il reste à ajuster les paramètres  $\alpha$  et  $\beta$  pour "coller" aux données de la simulation. Un ajustement issu de *Mathematica* est donné dans la table (2.6). La variance estimée pour cet ajustement est 0.013 302 1.

De l'égalité:

$$dN = 4\pi \left(\frac{r}{r_v}\right)^2 \hat{\nu} \left(\frac{r}{r_v}\right) d\left(\frac{r}{r_v}\right) = 4\pi r^2 \nu(r) dr \quad (2.5.44)$$

on peut déduire l'expression de la densité en nombre en fonction de  $r$ . Elle s'exprime alors:

$$\nu(r) = \frac{a/r_v^3}{(r/r_v)^\alpha (1+r/r_v)^{\beta-\alpha}} \quad (2.5.45)$$

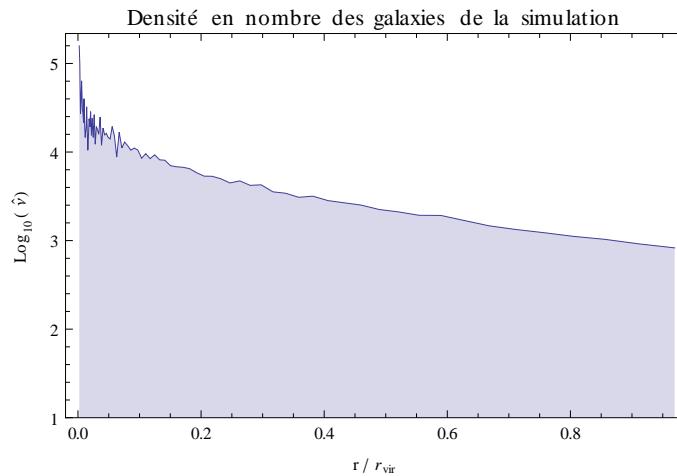


Figure 2.33: Densité en nombre issue de la simulation utilisée par Mamon et al. [1] en fonction de  $r/r_v$ .

Paramètre	Estimation	Erreur Standard
$a$	4505.31	672.116
$\beta$	2.8681	0.29134
$\alpha$	0.416376	0.0382949

Table 2.6: Estimation des paramètres du modèle utilisé pour la densité en nombre de la simulation.

En posant:

$$\tilde{\nu}(x) = \frac{1}{x^\alpha(1+x)^{\beta-\alpha}} \quad (2.5.46)$$

on peut écrire  $\nu(r) = C\tilde{\nu}(r/r_v)$ . En intégrant pour obtenir le nombre de galaxies comprises dans la sphère de rayon  $r$ , on obtient:

$$N(r) = 4\pi r_v^3 C \int_0^{\frac{r}{r_v}} \frac{x^{2-\alpha}}{(1+x)^{\beta-\alpha}} dx \quad (2.5.47)$$

D'après *Mathematica* on peut trouver une primitive à l'intégrand dans l'équation précédente qui s'exprime à l'aide de la fonction hypergéométrique 2F1 notée  $\mathcal{H}2F1$ . Techniquement, pour réaliser le calcul de cette fonction avec le programme en FORTRAN, on la modélise par une fonction rationnelle de polynômes qui permettent d'avoir une erreur relative de l'ordre de  $10^{-5}\%$ . Cette modélisation est réalisée dans un intervalle de valeurs compatibles avec les valeurs physiques que l'on peut rencontrer avec les données que l'on aura à traiter. Du fait qu'on utilise des paramètres sans dimension dans la plupart de nos calculs, il est facile de voir que les valeurs pour la fonction hypergéométrique est comprise dans l'intervalle  $[0, 1]$ . Finalement on peut écrire:

$$N(r) = \frac{4\pi r_v^3 C}{3-\alpha} \left(\frac{r}{r_v}\right)^{3-\alpha} \mathcal{H}2F1\left(3-\alpha, \beta-\alpha, 4-\alpha, -\frac{r}{r_v}\right) \quad (2.5.48)$$

Si comme précédemment on souhaite mettre  $N(r)$  sous la forme  $N_v \tilde{N}(r/r_v)$  et qu'on pose pour

## 2.5. NOTRE ALGORITHME

## CHAPTER 2. GROUP FINDER ALGORITHM

simplifier  $\mathcal{H}2F1(3 - \alpha, \beta - \alpha, 4 - \alpha, -x) = \mathcal{H}2F1(x)$  alors on obtient les formes suivantes:

$$\begin{aligned} N(r) &= N_v \tilde{N}(r/r_v) \\ \tilde{N}(x) &= x^{3-\alpha} \frac{\mathcal{H}2F1(x)}{\mathcal{H}2F1(1)} \\ \nu(r) &= \frac{N_v}{4\pi r_v^3 \mathcal{H}2F1(1)} \tilde{\nu}\left(\frac{r}{r_v}\right) \end{aligned} \quad (2.5.49)$$

La dispersion de vitesse elle s'écrit comme suit:

$$\sigma_z^2(R, r) = \frac{GM_v}{r_v} \tilde{\sigma}_z^2\left(\frac{R}{r_v}, \frac{r}{r_v}\right) \quad (2.5.50)$$

avec:

$$\tilde{\sigma}_z^2(x_R, x) = \left(1 - \frac{1}{2} \left(\frac{x}{x + 1/c}\right) \left(\frac{x_R}{x}\right)^2\right) \tilde{\sigma}_r^2(x) \quad (2.5.51)$$

et:

$$\tilde{\sigma}_r^2(x) = \frac{1}{\tilde{\nu}(x)} \int_x^\infty \left(\frac{S + 1/c}{x + 1/c}\right) \tilde{\nu}(S) \frac{\tilde{N}(S)}{S^2} dS \quad (2.5.52)$$

Du coup la densité de galaxies dans le halo (en utilisant toujours le même changement de variable pour éviter les problèmes numériques lors de l'intégration.

$$g_h(R, v_z) = \frac{N_v}{\sqrt{(2\pi r_v)^3 GM_v}} I\left(\frac{R}{r_v}, \frac{v_z}{v_v}\right) \quad (2.5.53)$$

avec:

$$\begin{aligned} I(x, \hat{v}_z) &= \int_0^{\text{acosh}(\frac{1}{x})} \tilde{\nu}(x \cosh u) \frac{x \cosh u}{\tilde{\sigma}_z(x, x \cosh u)} \\ &\times \exp\left(-\frac{\hat{v}_z^2}{\tilde{\sigma}_z^2(x, x \cosh u)}\right) du \end{aligned} \quad (2.5.54)$$

Le rapport des densités dans ce cas se trouve alors fortement simplifié et peut s'écrire:

$$\frac{g_i}{g_h}(R, v_z) = \frac{(2\pi)^{3/2} \hat{g}_i\left(\frac{R}{r_v}, \frac{|v_z|}{v_v}\right)}{2I(r_v/R, v_z/v_v)} \quad (2.5.55)$$

La probabilité résultante est présentée sur la figure (2.34) en violet en comparaison avec les données de la simulation elle même en rouge. Même si l'allure générale est plus proche de la réalité que dans le cas des autres modèles, les contours ne collent pas forcément mieux que les autres ce qui peut être causé par un mauvais ajustement de la densité calculée à partir des données.

Finalement d'autres améliorations sont sûrement à apporter à tout cela pour mieux coïncider avec les données de la simulation ou alors coller à la simulation ne sert peut être pas à grand chose... (les groupes sont mieux représentés par les modèles précédents sûrement).

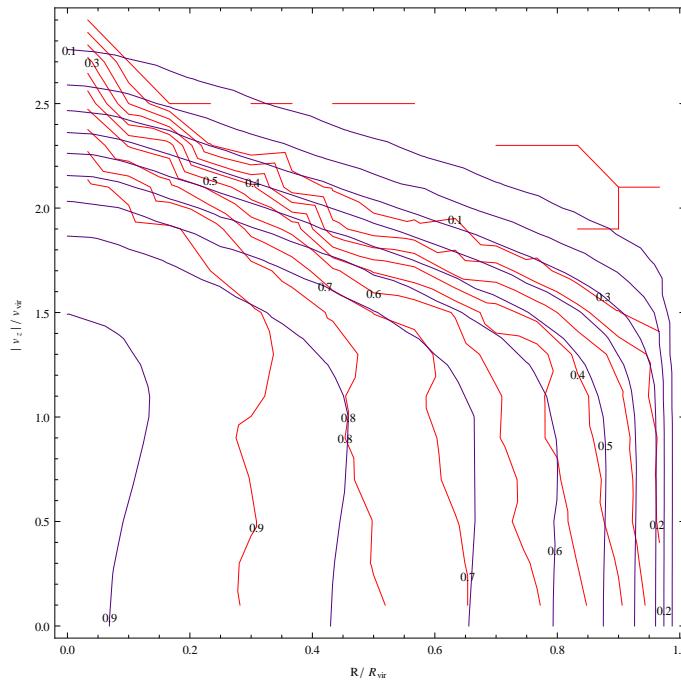


Figure 2.34: Contours de probabilité de la modélisation des données de la simulation en violet et des données de la simulation elles mêmes en rouge. L'allure des contours est plus proche que celle des modèles précédents mais coïncident moins avec les bonnes valeurs.

### Détermination de la cause des divergences des résultats par rapport aux données

Pour rechercher la source des divergences des résultats décrits dans la section précédente, la première chose qui vient à l'esprit est que le modèle de la densité des interlopers n'est pas parfait et qu'il doit provoquer ces différences. On va donc comparer la densité des interlopers dans le cas des données de la simulation et du modèle issu de Mamon et al. [1]. Pour retrouver les résultats dans le cas de la simulation il a fallu diviser la densité par le nombre de particules dans la sphère viruelle, c'est-à-dire le nombre de particules telles que  $r < 1$  avec  $r$  en unité viruelle. Les résultats sont visibles sur la figure (2.35). On ne peut pas incriminer le modèle qui semble coller suffisamment bien avec les résultats de la simulation. Il faut donc se tourner vers une autre cause.

Si on réalise le même genre de travail pour la densité de particules dans le halo, alors il faut faire la comparaison avec un modèle donné. On choisit le modèle NFW pour faire cela, de toute façon les résultats du modèle Einasto sont très proches. Les résultats du calcul de  $g_h(R, |v_z|)$  sont visibles sur la figure (2.36). On voit des différences apparaître quand les rayons projetés sont de plus en plus faibles et quand la vitesse sur la ligne de visée est élevée. Les courbes se croisent à environ un rayon de 15% du rayon de viriel, et cette même observation peut être faite pour les figures des contours de la probabilité dans les différents modèles. Comme les divergences apparaissent dans les mêmes régions dans le cas de la probabilité d'appartenance au halo et celle de la densité de particule dans le halo, on peut soupçonner que c'est le modèle qui est mauvais pour "fitter" les données.

Pour être sûr que le problème est bien le modèle, on a choisi de calculer la dispersion de vitesse sur la ligne de visée en fonction du rayon projeté. Pour la calculer, il faut partir du principe suivant: la dispersion de vitesse suivant la ligne de visée (LOS) est la moyenne sur la surface projeté des vitesses suivant la ligne de visée (la densité joue en quelque sorte le rôle d'une densité de probabilité

## 2.5. NOTRE ALGORITHME

## CHAPTER 2. GROUP FINDER ALGORITHM

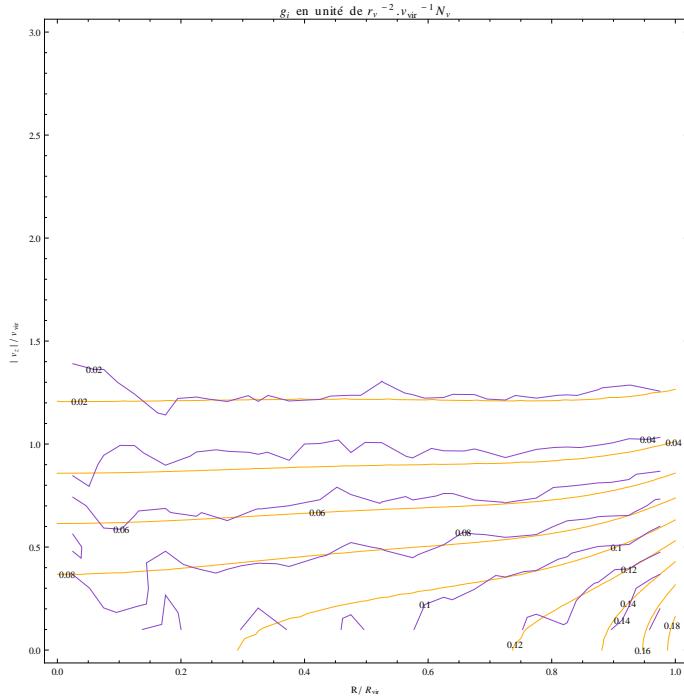


Figure 2.35: Graphe des contours de  $g_i(R, |v_z|)$  pour les données en violet et le modèle de Mamon et al. [1] en orange. Les résultats sont très semblables.

et donc on doit "normaliser" l'intégrale sur la ligne de visée). On peut donc écrire:

$$\Sigma(R)\sigma_{LOS}^2(R) = 2 \int_R^{r_v} \overline{v_{LOS}} \frac{\rho(r)r}{\sqrt{r^2 - R^2}} dr \quad (2.5.56)$$

En prenant en compte le fait que:

$$v_{LOS} = v_r \cos \theta - v_\theta \sin \theta \quad (2.5.57)$$

avec  $\sin \theta = R/r$ , alors on obtient après avoir appliqué la moyenne sur le développement du carré de l'équation (2.5.57) en sachant que les moyennes des composantes de la vitesse sont nulles:

$$\Sigma(R)\sigma_{LOS}^2(R) = 2 \int_R^{r_v} (\sigma_r^2(r) \cos^2 \theta + \sigma_\theta^2 \sin^2 \theta) \frac{\rho(r)r}{\sqrt{r^2 - R^2}} dr \quad (2.5.58)$$

et finalement avec le paramètre d'anisotropie  $\beta(r) = 1 - \sigma_\theta^2(r)/\sigma_r^2(r)$ :

$$\Sigma(R)\sigma_{LOS}^2(R) = 2 \int_R^{r_v} \left( 1 - \beta(r) \frac{R^2}{r^2} \right) \frac{\rho(r)\sigma_r^2(r)r}{\sqrt{r^2 - R^2}} dr \quad (2.5.59)$$

Or cette expression peut être simplifiée à partir de la solution de l'équation de Jeans (2.5.13) qui a pour solution l'expression (2.5.8) en remplaçant le  $\nu(r)$  par  $\rho(r)$  selon les cas choisis pour le calcul et donc il faut y prendre garde à l'aide des dimensions. Cela nous donne donc:

$$\begin{aligned} \Sigma(R)\sigma_{LOS}^2(R) &= 2 \int_R^{r_v} \int_r^\infty \frac{K_r(r,s)\rho(s)GM(s)}{s^2} ds \\ &\times \left( 1 - \beta(r) \frac{R^2}{r^2} \right) dr \end{aligned} \quad (2.5.60)$$

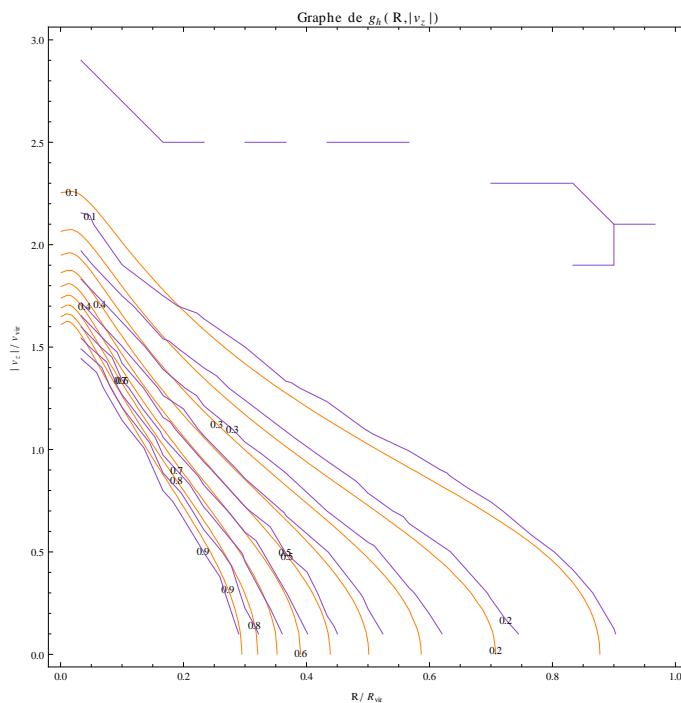


Figure 2.36: Graphe des contours de  $g_h(R, |v_z|)$  pour les données en violet et le modèle NFW en orange dans les mêmes unités que sur la figure (2.35). Les résultats sont problématiques dans certaines zones.

On voit que l'on a une intégrale double, donc on peut rechercher le domaine d'intégration pour voir si on peut aboutir à des intégrales plus simples. La variable  $r$  est comprise entre  $R$  et  $r_v$  et  $s$  entre  $r$  et l'infini, ceci peut se résumer de la façon suivante sur la figure (2.37) où le domaine de  $s$  est en jaune et celui de  $r$  en orange pâle. Le domaine d'intégration global est celui qui est le mélange des couleurs en orange foncé et jaune foncé qui représentent la subdivision de l'intégrale en deux domaines distincts qui vont faciliter la calcul de l'intégrale.

En décomposant l'intégrale suivant ces deux domaines et en développant l'expression, on obtient:

$$\begin{aligned}
 \Sigma(R)\sigma_{LOS}^2(R) &= 2 \int_R^{r_v} \frac{(s+a)}{s^2} \nu(s) GM(s) ds \\
 &\times \left( \int_R^s \left( \frac{r}{r+a} - \frac{1}{2} \left( \frac{R}{r+a} \right)^2 \right) \frac{1}{\sqrt{r^2 - R^2}} dr \right) \\
 &+ 2 \int_{r_v}^{\infty} \frac{(s+a)}{s^2} \nu(s) GM(s) ds \\
 &\times \left( \int_R^{r_v} \left( \frac{r}{r+a} - \frac{1}{2} \left( \frac{R}{r+a} \right)^2 \right) \frac{1}{\sqrt{r^2 - R^2}} dr \right)
 \end{aligned} \tag{2.5.61}$$

Si on réécrit la densité surfacique comme (dimension variable selon les cas)  $\Sigma(R) = \frac{M(a)}{\pi a^2} \tilde{\Sigma}(R/a, c) =$

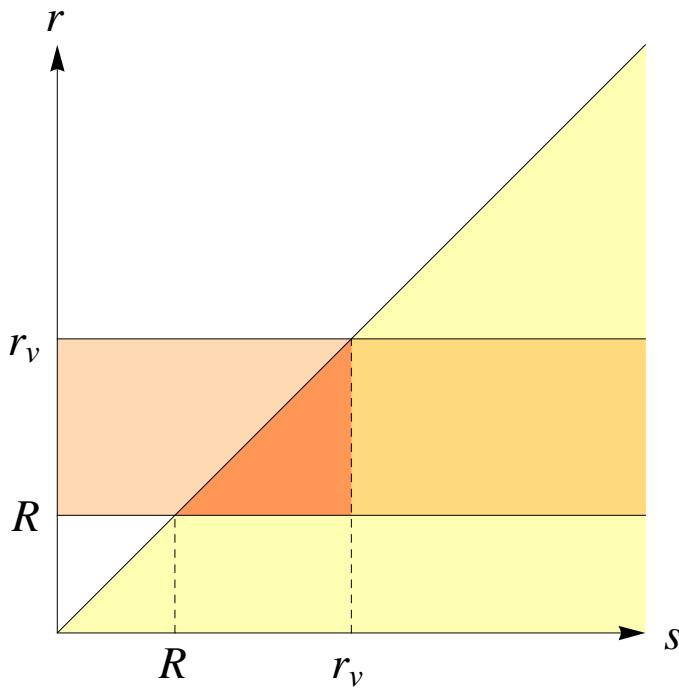


Figure 2.37: Le domaine d'intégration des différentes variables pour obtenir la dispersion de vitesse suivant la ligne de visée.

$4av\nu(a)\tilde{\Sigma}(R/a, c)$ , alors on peut aboutir en faisant le calcul des intégrales à:

$$\begin{aligned} \sigma_{LOS}^2(R) &= v_v^2 \frac{c/2}{\tilde{M}(c)\tilde{\Sigma}(R/a, c)} \\ &\times \left( \int_{R/a}^c K\left(x\frac{a}{r}, \frac{a}{r}\right) \tilde{\nu}(x) \frac{\tilde{M}(x)}{x} dx + I\left(c\frac{a}{r}, \frac{a}{r}\right) J(c) \right) \end{aligned} \quad (2.5.62)$$

avec:

$$I(u, u_a) = \begin{cases} -u_a \text{sign}(u_a - 1) \frac{u_a^{2-1/2}}{|u_a^2-1|^{3/2}} C^{-1}\left(\frac{1+uu_a}{u+u_a}\right) \\ + \text{acosh}u + \frac{1/2}{u_a+u} \frac{\sqrt{u^2-1}}{u_a^2-1}, & u_a \neq 1 \\ \text{acosh}u - \sqrt{\frac{u-1}{u+1}} \left(\frac{8+7u}{6(1+u)}\right), & u_a = 1 \end{cases} \quad (2.5.63)$$

$$K(u, u_a) = \left(1 + \frac{u_a}{u}\right) I(u, u_a) \quad (2.5.64)$$

et:

$$C^{-1}(X) = \begin{cases} \text{acosh}X & u_a < 1 \\ \text{acos}X & u_a > 1 \end{cases} \quad (2.5.65)$$

ainsi que:

$$J(y) = \int_y^\infty \frac{x+1}{x^2} \tilde{\nu}(x) \tilde{M}(x) dx \quad (2.5.66)$$

Cette dernière expression peut s'écrire également pour le modèle NFW:

$$\begin{aligned} J(y) = & \frac{2}{3y^2(1+y)(\ln 4 - 1)^2} \left( y(-3 + y(-9 + \pi^2(1+y))) \right) \\ & + 3y^3 \ln \left( 1 + \frac{1}{y} \right) + 3 \ln(1+y) (1 - y + y^2(1 + y \ln(1+y))) \\ & - 3y^2 \ln(y(1+y)) + 6y^2(1+y) Li_2(-y) \end{aligned} \quad (2.5.67)$$

Donc finalement on peut rapidement et facilement calculer la dispersion de vitesse sur la ligne de visée pour vérifier si nos modèles sont corrects pour être appliqués au calcul de la probabilité. Les premiers résultats étaient comme ceux décrits précédemment, c'est-à-dire que le modèle ne collait pas parfaitement avec la dispersion de vitesse issue directement des données de la simulation. Ceci était assez bizarre car dans Mamon et al. [1] ce calcul avait été fait aussi et les résultats étaient assez proches.

Mais un jour la lumière a jailli et mis en évidence le fait que les calculs précédents utilisaient la concentration calculée à partir de la relation de Macciò et al. [12] qui pour la masse viruelle utilisée de  $2 \times 10^{14} M_\odot$  donne  $\approx 5.8$  et non pas une concentration de  $c = 4$  comme pour la simulation ce qui explique les écarts observés.

On peut maintenant tracer la dispersion de vitesse dans le cas de la simulation en gardant les particules dans le halo et avec une coupure à  $2.7\sigma_{LOS}(R)$ . Les résultats sont sur la figure (2.38). On voit alors que le modèle utilisé peut être difficilement mis en cause pour nos écarts dans la probabilité avec les données. Même une fois la correction de la concentration apportée, des écarts existent toujours avec les contours de la probabilité hors on discrimine un peu le modèle avec la dispersion de vitesse qui marche bien. On peut étayer cette remarque par la nouvelle densité de particules dans les halos obtenue avec la concentration corrigée sur la figure (2.39). Si les écarts observés avec la densité issue des données était la cause du problème, alors les résultats seraient inversés. En effet le modèle sous-estime la vraie densité, donc  $g_i/g_h$  est sur-estimé et alors la probabilité est sous-estimée. Or les contours de la probabilité sont, comme on le voit sur la figure (2.40), au-dessus des contours de la simulation donc le modèle sur-estime la probabilité pour les petits  $R$  et les grands  $v_z$ . Le problème peut seulement venir alors de la densité des interlopers. Cette densité est calculée pour une coupure à  $2.7\sigma_{LOS}(R)$  pour le modèle et on voit sur la figure (2.40) que le modèle s'écarte des données au-dessus de cette coupure (la ligne verte sur le graphe) ce qui fait vraiment penser que le modèle des interlopers est en cause car mal évalué au-dessus de cette coupure.

Si le modèle des interlopers pose problème, il s'agit sans doute de la constante  $B$  dans le modèle qui est mal évaluée. Elle représente la composante plate de la densité dans le modèle mais on voit sur les graphes dans Mamon et al. [1] que cette composante fluctue assez au dessus de la coupure en vitesse sur la ligne de visée. Or c'est là que les probabilités calculées posent des soucis. Donc il faut voir l'influence de ce paramètre sur notre estimation de la probabilité.

Le fait de faire varier cette composante plate a pu à certains endroits permettre de faire coïncider un peu mieux les contours bleu et rouge, mais des fois accentue les écarts. En regardant Mamon et al. [1], on peut constater que sur les figures (2.41, pour des grands  $v_z$  les fluctuations sont importantes et peuvent expliquer les différences observées au-dessus de deux unités de vitesse viruelle. Sauf à paramétriser cette constante avec le rayon projeté et la vitesse suivant la ligne de visée, faire coïncider les données et le modèle est peine perdue. Il semble bien que dans cette région ce soit la densité des interlopers qui impose ses règles pour la probabilité. Il faut donc prendre mieux en compte la densité

## 2.5. NOTRE ALGORITHME

## CHAPTER 2. GROUP FINDER ALGORITHM

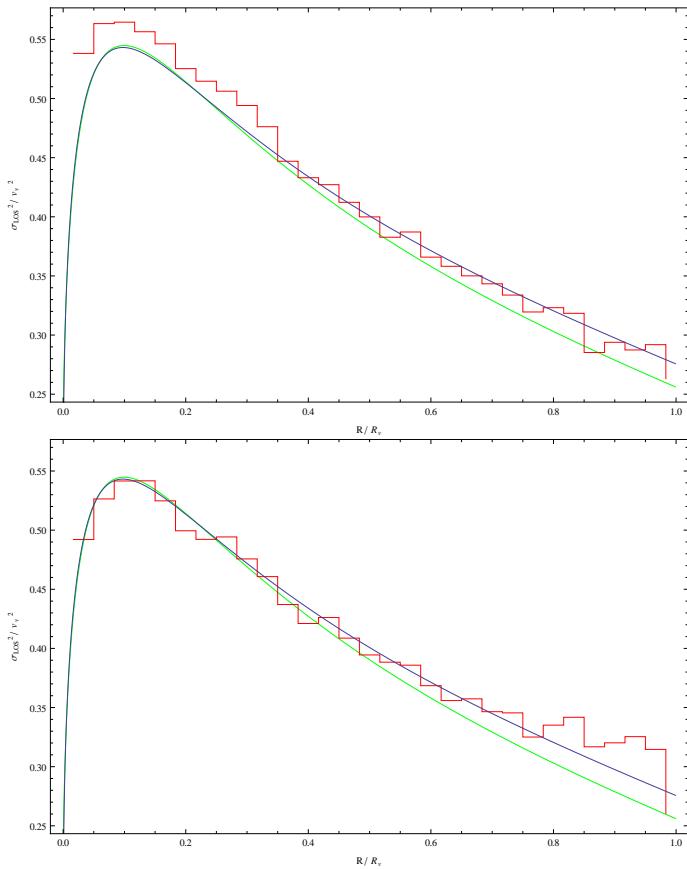


Figure 2.38: Graphe de la dispersion de vitesse calculée selon le modèle NFW en bleu et celle à partir des données de la simulation en rouge. En haut en prenant en compte les particules dans le halo et en bas en réalisant une coupure à  $2.7\sigma_{LOS}(R)$ .

des interlopers dans ce domaine de l'espace des phases projeté pour avoir une bonne correspondance. Mais on peut s'interroger sur l'intérêt de réaliser ce travail car les fluctuations observées dans la constante qui représente la composante plate de la densité des interlopers dépendent essentiellement de la simulation utilisée et vont sûrement varier d'une simulation à l'autre. Ne traduisant pas une propriété intrinsèque de la distribution des interlopers, ces fluctuations ne vont pas permettre de faire un meilleur calcul de la probabilité quand on le réalisera en situation sur les données du SDSS, la distribution des interlopers fluctuant sûrement aussi différemment dans ce domaine de l'espace des phases projeté.

Il faut alors se résoudre (peut-être) à se limiter à ce calcul de la probabilité d'appartenance au halo au-dessus de la coupure en vitesse de  $2.7\sigma_{LOS}(R)$ , au dessus la fluctuation de la densité des interlopers empêchant de réaliser un calcul propre et précis.

Il s'avère que si on trace les contours de  $g_h$  logarithmiquement, alors on voit que la densité de particules dans le halo du modèle s'éloigne des données aux grandes vitesses suivant la ligne de visée (voir la figure (2.42)). Et ceci de façon cohérente avec la surestimation expliquée plus haut. On ne peut plus incriminer la densité des interlopers mais bien la densité des particules dans le halo. Alors on peut se demander pourquoi la modélisation dans le cas du modèle ajusté directement ne marchait pas. Il est possible que les particules qui sont près du centre soient problématiques dans la détermination des coefficients du modèle choisi. Selon les recommandations de Gary, on va éliminer les particules en-dessous de 3% du rayon de viriel car la mauvaise détermination du centre des halos

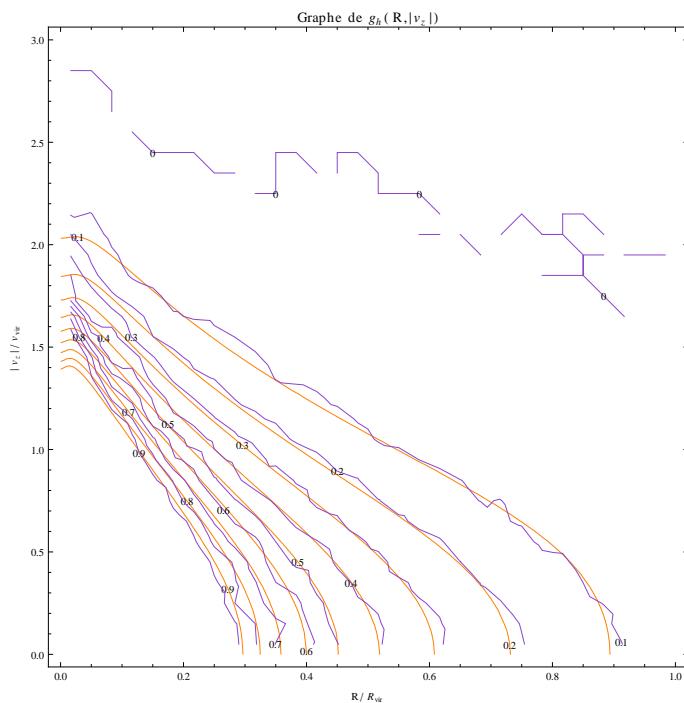


Figure 2.39: Nouveaux contours de la densité de particules dans les halos avec la concentration forcée cette fois à 4.

peut influencer sur le choix du modèle qui colle le mieux aux données et faire croire que par exemple NFW est meilleur que Einasto alors que non, et empêcher de trouver une modélisation.

## 2.6 Conclusion

Durant le stage, plusieurs points importants pour aider à la réalisation de notre algorithme ont été mis en place. On a réalisé un mock catalogue à partir des données de la simulation du Millennium-II qui nous permettra de tester les différents algorithmes réalisés. Des caractéristiques comme une magnitude limite, correction-K ont été appliquées dessus afin de le faire correspondre au mieux au SDSS.

L'algorithme de Yang et al. [2] a été programmé pour avoir à notre disposition un algorithme de groupe qui pourra être comparé au notre. Les différents outils utilisés pour y aboutir sont en place et pourront éventuellement être incorporés à notre algorithme s'ils s'avèrent utiles (comme par exemple le FoF, ou la méthode de l'abundance matching). Les premiers résultats sont satisfaisant à moitié, car il semble que le programme fonctionne bien pour certaines itérations et pas pour d'autres, la cause de ce problème étant en cours de recherche.

Le principe de notre algorithme a aussi été discuté, et amélioré à partir des remarques sur les défauts que semble présenter sur certains points l'algorithme de Yang et al. [2].

On estime que d'ici peu les deux algorithmes pourront fonctionner correctement et donner leur premier résultat sur le mock catalogue et ensuite être appliqués sur le SDSS-DR7 afin de pouvoir comparer les résultats sur les groupes obtenus et leurs propriétés.

Avec les améliorations que l'on apporte à notre algorithme par rapport à ceux déjà existants, les groupes devraient pouvoir être mieux quantifiés ce qui devrait permettre d'avoir accès à une meilleure connaissance des effets de l'environnement des galaxies sur leurs propriétés comme leur masse stellaire en fonction de la masse du halo, ou l'évolution de paramètres comme la métallicité

## 2.6. CONCLUSION

## CHAPTER 2. GROUP FINDER ALGORITHM

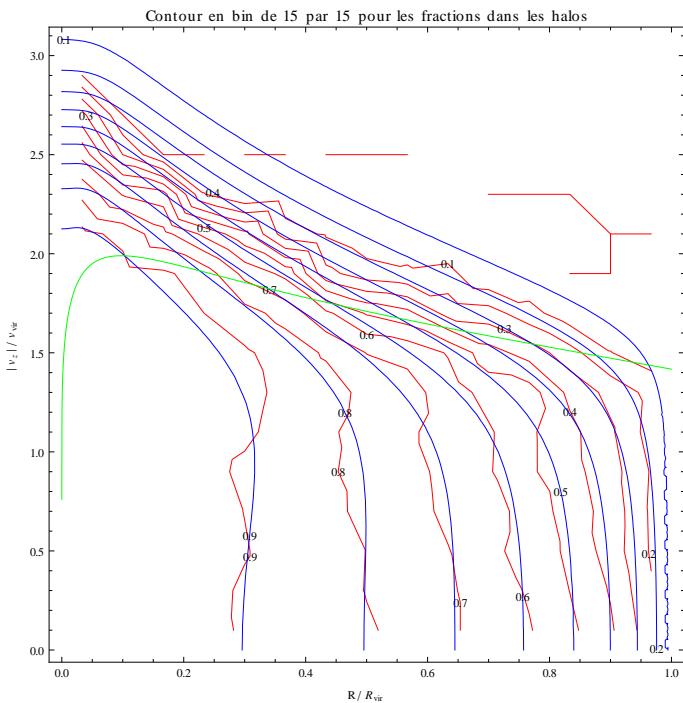


Figure 2.40: Contours de la probabilité issus des données en rouge, du modèle NFW en bleu et en vert la coupure à  $2.7\sigma_{\text{LOS}}(R)$  de la vitesse suivant la ligne de visée.

avec la position de la galaxie dans le groupe. Finalement, ces effets pourront alors être modélisés et intégrés sous forme de codes semi-analytiques dans les simulations cosmologiques avec les modèles de formation de galaxies pour prendre mieux en compte la physique de l'environnement, et réduire les écarts aux observations des données issues de ces simulations.

Je tiens à remercier Thierry SOUSBIE pour ces conseils sur la réalisation d'un FoF "rapide" qui a permis de faciliter le travail réalisé dans ce stage.

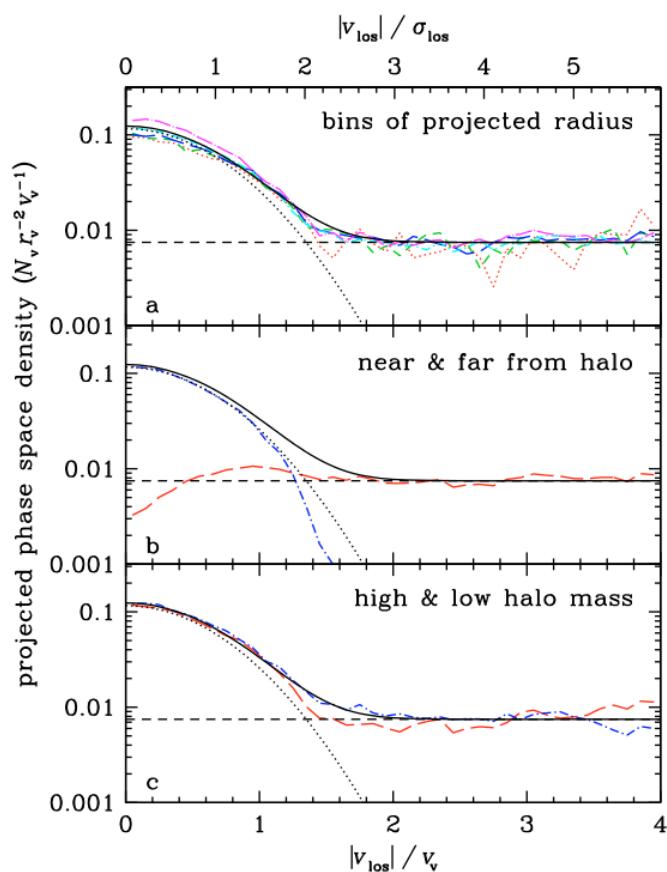


Figure 2.41: Densité des interloper en fonction de la vitesse suivant la ligne de visée. **a:** dépendance avec le rayon projeté en bins croissants du rayon (rouge, vert, bleu, magenta, cyan). **b:** dépendance en distance radiale. **c:** dépendance en masse du halo. (voir l'article pour plus de précision)

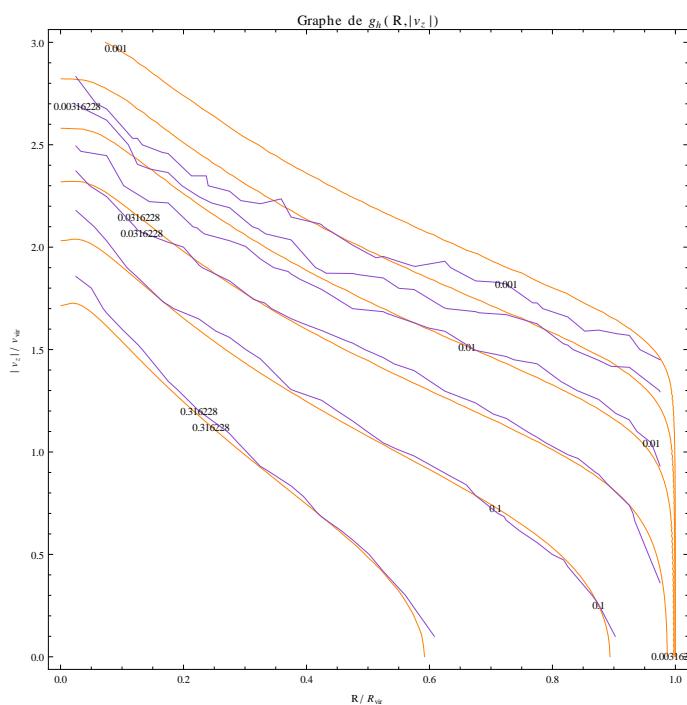


Figure 2.42: Contours de la densité de particules dans le halo comme dans la figure mais les contours espacés logarithmiquement.

## 2.7 Determine the LF

When we want to compare the results from our galaxy group finder to other existing algorithms, we have to compare a flux limited catalogue with our algorithm. But as said before, our algorithm works on a double complete sample of galaxies. So, we need to develop a flux limited version of our algorithm.

The problem when working with a flux limited sample of galaxies is that we must correct for missing galaxies. The galaxies observed by the survey can be seen just when they are brighter than a luminosity limit which depends on the redshift (the distance of the galaxy). We can determine this luminosity easily by the theory:

$$L_{\text{lim}}(z) = \left( \frac{d_{\text{lum}}(z)}{10pc} \right)^2 10^{0.4(M_{\odot} - m_{\text{lim}})} \quad (2.7.1)$$

with  $m_{\text{lim}}$  the magnitude limit of the survey and  $M_{\odot}$  the absolute magnitude of the sun, all in the same band filter. This luminosity is in unit of sun luminosity. In ours groups, when they are at a distance above the distance limit to see galaxies with the luminosity threshold of the catalogue, some galaxies are missing because they can't be seen. In order to correct for the number of missing galaxies, we have to know the distribution of galaxy luminosities. With this luminosity function (LF), we can calculate the "fraction" of galaxy luminosities in mean that we can see:

$$f(L_{\text{lim}}(z)) = \frac{\int_{L_{\text{lim}}(z)}^{\infty} L\phi(L)dL}{\int_{L_{\text{thres}}}^{\infty} L\phi(L)dL} \quad (2.7.2)$$

where  $\phi(L)$  is the LF. So determining this LF is useful to correct for missing galaxies.

But, as we expect with the goal of this thesis, it's clear for us that properties of galaxies depend on the environment. So the luminosity function may probably depend on the host halo. The LF have to depend on different characteristics of the halo. The unique "observable" property is the virial mass so we want that the LF depends on it. This is the better way to correct for the incompleteness of groups, with a particular correction for each group. **Manuel: Put something in order to justify physically this dependence on the halo mass!** Resulting from this idea, the LF used in (2.7.2) becomes a conditional LF, which is the LF in groups of a given halo mass:

$$\phi(L) \rightarrow \phi(L|M) \quad (2.7.3)$$

In galaxy groups, we separate galaxies in two classes: centrals and satellites. Centrals are expected to be the most massive galaxies in groups, and consequently, it's probable that the central is the brighter galaxy. A consequence is that if we can't see the central galaxy, we can't see other galaxies in the group and the correction is not needed because we don't know how to correct for incompleteness. So for the correction we just need to constrain the distribution of luminosities in satellite galaxies.

In practice, we have to choose a functional for this conditional luminosity function (CLF) which can be easily fitted and integrated to determine the correction factor in our group luminosities. In

## 2.7. DETERMINE THE LF

## CHAPTER 2. GROUP FINDER ALGORITHM

studies of the galaxy sample from the SDSS survey as in Blanton et al. [10], the LF has been well fitted by a double Schechter functional form which can be written:

$$\phi(L) = \left( \phi_1^* \left( \frac{L}{L_*} \right)^{\alpha_1} + \phi_2^* \left( \frac{L}{L_*} \right)^{\alpha_2} \right) \exp \left( - \left( \frac{L}{L_*} \right) \right) \quad (2.7.4)$$

Now we assume that the CLF have the same form that (2.7.4). The dependence on the halo mass  $M$  is taken with the parameters of the double Schechter (DS). For example  $\alpha_1 \rightarrow \alpha_1(M|\theta)$ , where the functional form of this dependence is not given explicitly here, and  $\theta$  is a set of parameters relative to the function used to describe the dependence with halo mass. The number of parameters in  $\theta$  can vary greatly, depending on the function used.

The form of this dependence can't be determine in advance when we want to fit the CLF on the data. For example in the SDSS, we have to know in advance the properties of the groups in order to choose a certain dependence for the parameters of the DS with the virial mass. So, for testing the viability of this method, we have to select a functional that describes correctly the modulation of the parameters with the halo, and samples of galaxies that can give us this information are present in outputs of semi-analytical models (SAM). In such samples, we know in which group a galaxy is, and the virial mass of the host halo is known too. To validate this method of correction for incompleteness, we can test it in mock galaxy catalogues.

### 2.7.1 Estimating parameters

We need to use a method for estimating the parameters that fit well the data (real or simulated). When working with distribution function, it is common and better to use the maximum likelihood estimation defined as:

$$\mathcal{L}(\theta|X) = \prod_i p_i(X_i|\theta) \quad (2.7.5)$$

where  $X$  is the set of data (in our case the luminosity  $L$ ) and  $\theta$  the set of parameters of the model that have to be estimated.

If we consider Bayesian statistics, the likelihood is defined as  $p(X|\theta)$  and it seems to be incoherent. But using the Bayes's theorem, we can see that *our* likelihood is in reality the posterior distribution which is proportional to the likelihood in the definition of Bayesian statistics, multiply by a prior. But we don't have any prior on the parameters distribution (which can be discussed...). So, if we take a constant for the prior (probability equal for each parameter), we get same results.

It's more convenient to use the logarithm of the likelihood in order to prevent numerical problems when calculating the likelihood. The product in (2.7.5) becomes at this moment a sum, and the computation is simplified. It's easier too to minimize a function numerically so we rewrite it.

$$-\log \mathcal{L}(\theta|X) = -\sum_i \log(p_i(X_i|\theta)) \quad (2.7.6)$$

We define  $p_i(X_i|\theta)$  as the probability to get the value  $X_i$  given the parameters  $\theta$ , so it's the probability density. To determine this density, we need to calculate the number of "points" in the sample which are between  $X_i$  and  $X_i + dX_i$  compare to the total number of points in the set  $X$ :

$$p_i(X_i|\theta) dX_i = \frac{dN_i}{N_{\text{tot}}} \quad (2.7.7)$$

By definition of the CLF, which is the number of galaxies by unit of volume comprised between  $L$  and  $L + dL$  at a given halo mass  $M$ , we can write:

$$d^2N = \phi(L|M) dL dV \quad (2.7.8)$$

Summing on all the volume in which we are working, we get:

$$dN = \phi(L|M) dL \quad (2.7.9)$$

**Manuel: Because we are working in limited volume region, we can rewrite the probability density as:**

$$p_i(L_i|\theta) dL_i dV = \frac{d^2N_i}{N_{\text{tot}}} \quad (2.7.10)$$

So we can write :

$$p_i(L_i|\theta) dL_i = \frac{\phi(L_i|M) dL_i}{N_{\text{tot}}} \quad (2.7.11)$$

and the total number of galaxies is just:

$$N_{\text{tot}} = \int_{L_{\text{thres}}}^{\infty} \phi(L|M) dL \quad (2.7.12)$$

In this way, the density probability for the DS can be written:

$$p_i\left(L_i \middle| \alpha_1, \alpha_2, M_*, \frac{\phi_2^*}{\phi_1^*}\right) = \frac{\left(\left(\frac{L}{L_*}\right)^{\alpha_1} + \frac{\phi_2^*}{\phi_1^*}\left(\frac{L}{L_*}\right)^{\alpha_2}\right) \exp\left(-\left(\frac{L}{L_*}\right)\right)}{\left(\Gamma\left(1 + \alpha_1, \frac{L_{\text{thres}}}{L_*}\right) + \frac{\phi_2^*}{\phi_1^*} \Gamma\left(1 + \alpha_2, \frac{L_{\text{thres}}}{L_*}\right)\right)} \quad (2.7.13)$$

where  $\Gamma(a, x) = \int_x^\infty e^{-t} t^{a-1} dt$  is the incomplete gamma function.

The principle of the estimation by the method of the maximum likelihood is that when we maximize the likelihood relatively to the parameters  $\theta$ , we get the maximum of probability, likelihood, of having the parameters that correspond denoted  $\hat{\theta}$ . The parameters  $\hat{\theta}$  are the parameters that best fit the data according to the functional form assumed for the CLF. Numerically we minimize the equation (2.7.6).

There are many ways of doing such a minimization. When the probability density isn't too complex,  $\hat{\theta}$  can be determined analytically. But in this case, with the DS, the incomplete gamma function prevent us to do it in this way. So we are constrained to use numerical methods in order to minimize the likelihood. Many algorithms exist to do this job like Powell's method, Newton-Raphson's method, etc..., but they share the same problem: when they find a minimum, we can't know if it is the global minimum or if it is a local minimum. The result depends on the initial starting point of the algorithm in the parameter space. Some other methods try, using Monte-Carlo methods, to do a better exploration of this parameter space, allowing some "jumps" to other regions in order to see if there isn't a best minimum near. An example of such an algorithm is the simulated annealing method which implement the cooling of a material where the function to minimize becomes the energy of the system and a fictive temperature  $T$  is introduced to allow some temperature jumps.

But it is not always sure that we get the global minimum. Moreover, we can't easily determine errors on the estimation of the parameters, except using bootstraps or jackknife techniques which need many estimation of the parameters varying the sample which may be expensive in calculation time.

We have chosen to use the Markov Chains Monte Carlo method (MCMC) to minimize our function. This method is the better in all the universe. **Manuel: Explain why!**. We can estimate easily with results of the algorithm the errors on the parameters.

We will now resume the result of works on mock catalogues of our method of estimating the CLF.

### 2.7.2 Tests on mock catalogues

There are two steps in order to determine the dependence on the halo mass of the parameters of the DS model. First, we have to determine what is the best functional form to fit this dependence which can be done on a complete sample of galaxy. Secondly, see if we can recover this parametrisation and modulation with a flux limited sample of this galaxies to know if the method works well when applied in a real survey.

#### Complete sample

In order to determine the dependence on the halo mass of the parameters, we use a complete sample of galaxies taken from the outputs of the SAM of Guo et al. [3] applied on dark matter halos from the Millennium II run. We limit our sample of galaxies from this catalogue to galaxies with a luminosity such that the absolute magnitude in the  $r$  band is  $M_r < -12$ . For each galaxy, we have the virial mass of the halo which contains this galaxy. Our complete sample is defined just as this galaxy catalogues from Guo et al. [3] with the truncation on the data through the  $r$  band magnitude.

First, we determine what is the best model for the "total" CLF, *i.e.* the LF when we don't do a segregation with the halo mass. We have tried to adjust a simple Schechter and a double Schechter. Results are shown on figure (2.43). We can see that the minimization works well because the fit seems to be good enough in the figure. The double Schechter fits better the data than the simple Schechter because we can constrain with this form the two populations of galaxies in the sample from the Guo et al. [3] SAM. We see that there is a low population with high slope and a brighter population with a slope more little. Differences with the data at luminous galaxies is due to the fact that the number of galaxies with  $M_r < -24$  is very low, in some bins there is just one galaxy. But we need a quantitative proof of this fit. We use for that the Kolmogorov-Smirnov test and the P-Value associated. **Manuel: TODO: KS test on the fit in order to get a good idea of the robustness of the fit.**

Happy to see that the minimization is good, we want to see the modulation of the parameters of the DS with the halo mass. For doing that we take galaxies in bins of the certain width in halo mass, and we compute the parameters that fit well the data in each, as previously. This modulation is represented in the figure (2.45). The resulting fit of the probability density function is shown for the double Schechter in the figure(2.44).

As we expect, we can see some physics process putted in the SAM in this figure. **Manuel: To detail.**

Therefore, **Manuel: Verify Cependant→Therefore.**, we don't see a particular modulation, *i.e.* a functional form to use for each parameter of the DS. So we have decided to use polynomials

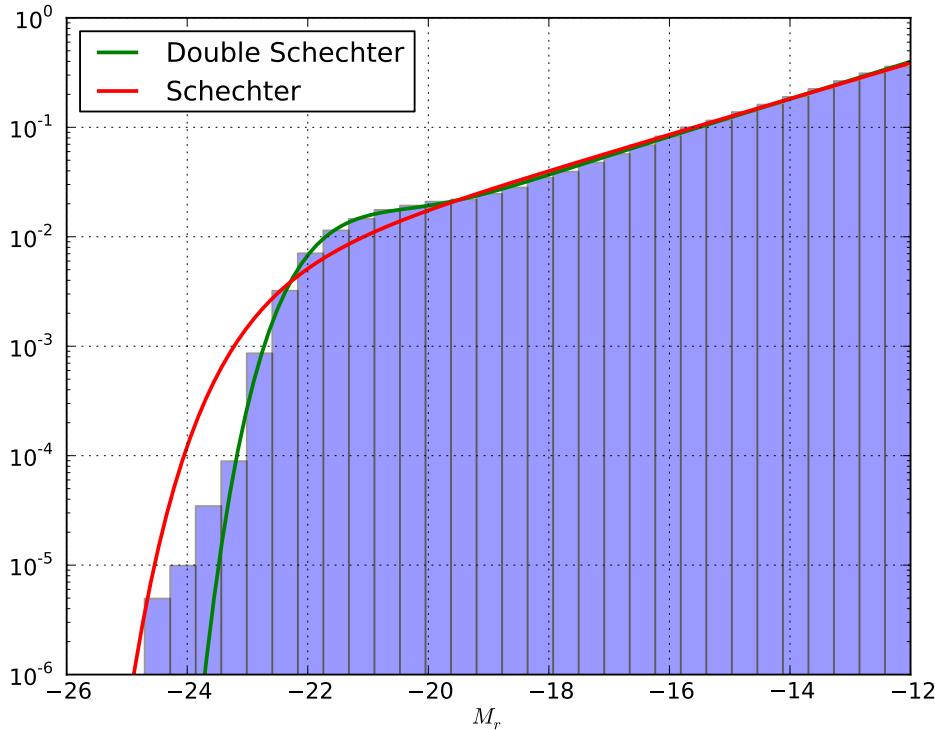


Figure 2.43: Results of the fit on the data in blue with a Schechter distribution in red and a double Schechter distribution in green. The double Schechter fit is better because this form allow to constrain the two galaxy populations we can see in our sample: a low population with high slope and a brighter population with a more little slope in absolute value.

functions to adjust this modulation. For a given parameter  $\theta_i$  we write:

$$\theta_i(M) = \sum_{j=0}^N a_{ij} M^j \quad (2.7.14)$$

where  $N$  is the order of the polynomial.

We now integrate this form of parametrization of the parameters of the DS into the minimization of the likelihood using directly the halo masses of the galaxies.

But it doesn't work!!! **Manuel: Some blabla!**

### Flux limited sample

Working with a flux limited sample, it's like having gaps into the set of galaxies. But we know why we are missing this galaxies: we can't see them at a given distance. Determining the luminosity limit which we can see at redshift  $z$  can be done analytically. So for that we use the STY method. **Manuel: Put Bibtex of STY and description related to this.**

We need to modify the probability density used in the likelihood to take into account missing data. The procedure is the following: in order to estimate the likelihood, we have to calculate the probability density for a galaxy  $i$  at the redshift  $z_i$  of having a magnitude between  $M_i$  and  $M_i + dM_i$ . The probability that a galaxy have a magnitude

## 2.7. DETERMINE THE LF

## CHAPTER 2. GROUP FINDER ALGORITHM

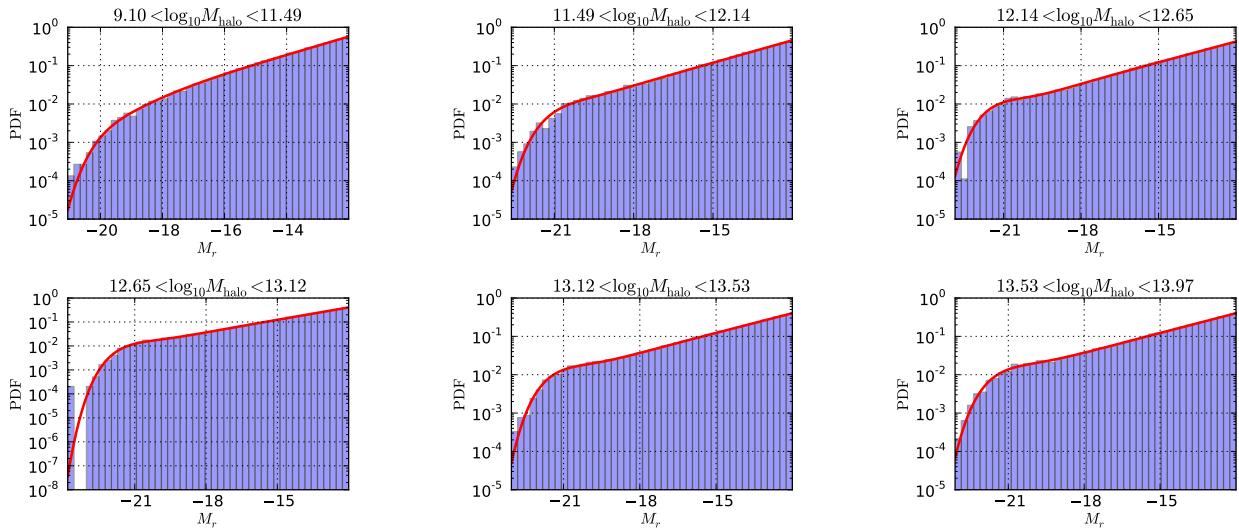


Figure 2.44: Results of the fit for the double Schechter distribution in different bins of halo mass. In blue the PDF of the data, and in red the fit with parameters obtained from the MCMC algorithm.

$\mathcal{M}$  superior to  $M$  is given by:

$$P(\mathcal{M} > M|z) = \frac{\int_{-\infty}^M \phi(M') \rho(z) f(M') dM'}{\int_{-\infty}^{\infty} \phi(M') \rho(z) f(M') dM'} \quad (2.7.15)$$

where  $f$  is the completeness function and  $\rho(z)$  is the redshift distribution. We can express it like this because the number of galaxies at a given redshift  $z$  with a given magnitude  $M$  is  $dN = \phi(M) dM$  in a given volume. So to avoid the volume dependence, we multiply by the number of galaxies in a given volume  $\rho(z) = dN/dV$ . So the number of galaxies with a magnitude between  $M$  and  $M + dM$  is  $d^2N = \phi(M) \rho(z) dM dV$ . But with the problem of the completeness, we can see just galaxies with a certain magnitude defined by  $f$  so  $d^2N = \phi(M) \rho(z) f(M) dM dV$ . The probability (2.7.15) results from this. Calculating the probability density is straightforward because we have:

$$P(\mathcal{M} > M|z) = \int_{-\infty}^M p(M'|z) dM' \quad (2.7.16)$$

and so:

$$p(M|z) = \frac{\partial P(\mathcal{M} > M|z)}{\partial M} \quad (2.7.17)$$

Finally:

$$p(M_i|z_i) = \frac{\phi(M_i)}{\int_{M_{\text{bright}}(z_i)}^{M_{\text{faint}}(z_i)} \phi(M') dM'} \quad (2.7.18)$$

and this defines the new likelihood in the case of a flux limited sample.

We apply this to the incomplete sample generated with the mock algorithm we have created. Firstly, we have tried to recover the parameters in the mock with the apparent magnitudes calculated applying a "K-decorrection". The results are very bad. Parameters can't be recovered correctly and with the two models chosen (simple Schechter and DS). We don't have the same estimations as in the complete sample, although the flux limited sample is the same as the complete sample but

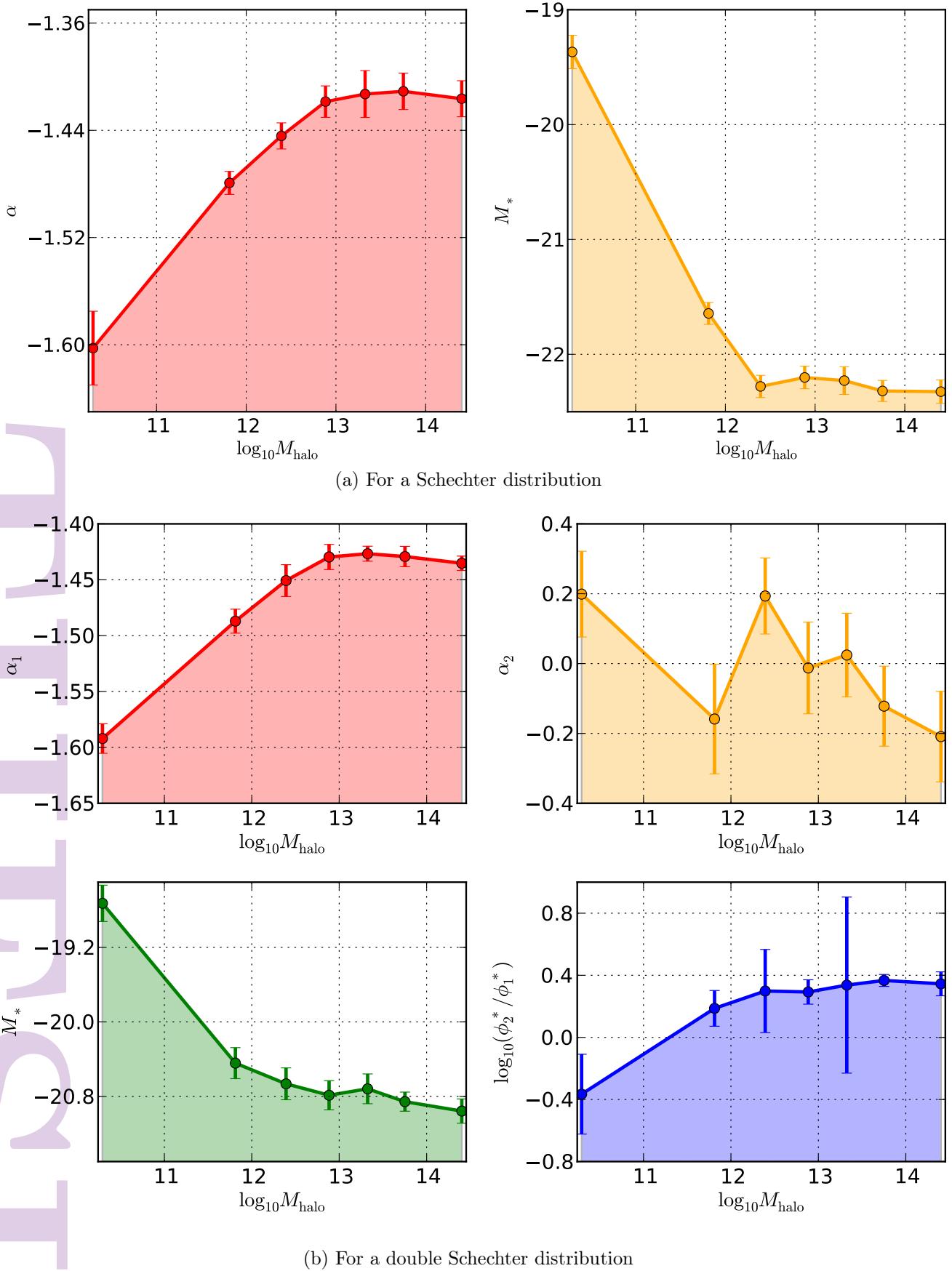


Figure 2.45: Modulation of the parameters of both Schechter and double Schechter luminosity distributions with the halo mass.

## 2.7. DETERMINE THE LF

## CHAPTER 2. GROUP FINDER ALGORITHM

"truncated". It's not clear why this parameters can't be found. The procedure isn't in cause, but it's possible that the data are the problem. We think that our method for estimating the K-decorrection have some troubles when the redshift is higher than 0.2.

For verifying this assumption, we have made more simple flux limited samples. We know how to generate random variables following a Schechter distribution and a DS distribution two. We have placed galaxies in a homogeneous universe, up to few hundred Mpc. We have assigned redshifts to this galaxies according to their distances  $D$  using simply  $z = v/c = H_0 D/c$ . We apply a Schechter distribution to the galaxies generated, without taking clusters effects into account. Calculating apparent magnitudes is done subtracting the distance modulus using galaxies redshifts. We have done the same applying a DS. Results are the followings:

- We are enable to recover the Schechter parameters used to generate the distribution in the flux limited sample. When data are perfect like in this situation, there are no troubles.
- Using a DS distribution, we have more difficulties in finding those parameters used to generate the flux limited sample. The slopes of both the low galaxies and brighter galaxies aren't near the true values.

We think that the number of low galaxies in the flux limited sample isn't not sufficient to well constrain the slopes in the DS. As a result, the maximum likelihood can't find real parameters. it's like a degeneracy is present and the algorithm can't decide to the true parameters.

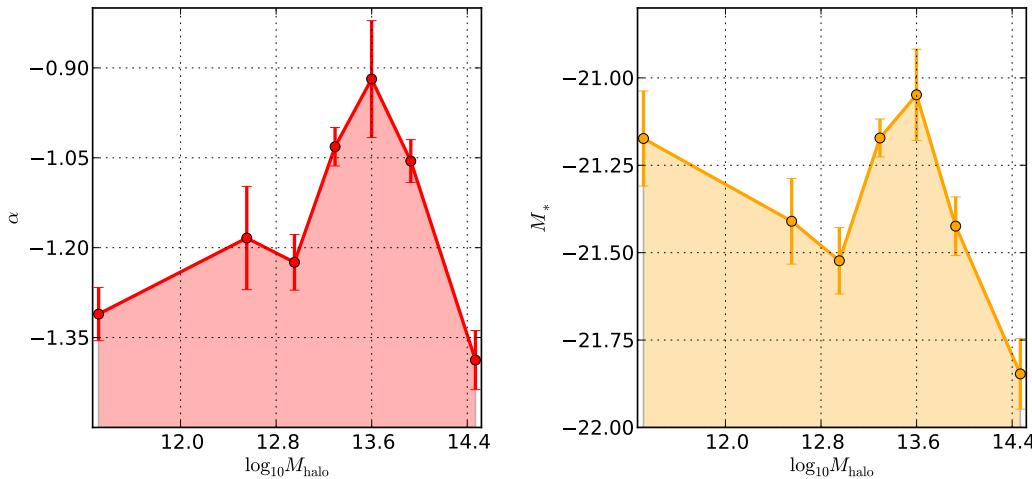


Figure 2.46: Modulation of the parameters for a Schechter distribution with the halo mass in a mock catalogue constructed with the galaxy catalogue from the Guo2010a, without taking a K-decoration for apparent magnitudes of galaxies.

So, it's a clue that the K-decoration isn't good at high redshift and underestimate the number of galaxies when we apply a flux limit using apparent magnitudes. We have made an other mock catalogue without this K-decoration in order to show that is the real problem. We have extended too the upper magnitude limit of galaxies in the sample in order to improve the number of low galaxies to estimate better slopes in the case of the DS. The former limit was  $-15$  in  $r$  band magnitude and now we go to  $-12$ . So we expect that the number of low luminosity galaxies increases and improves the parametrization. All results hereafter and before are for this new limit.

Unfortunately, the increasing number of low mass galaxies and taking no K-decoration doesn't improve the estimation of the parameters. Results are shown in figure (2.46).

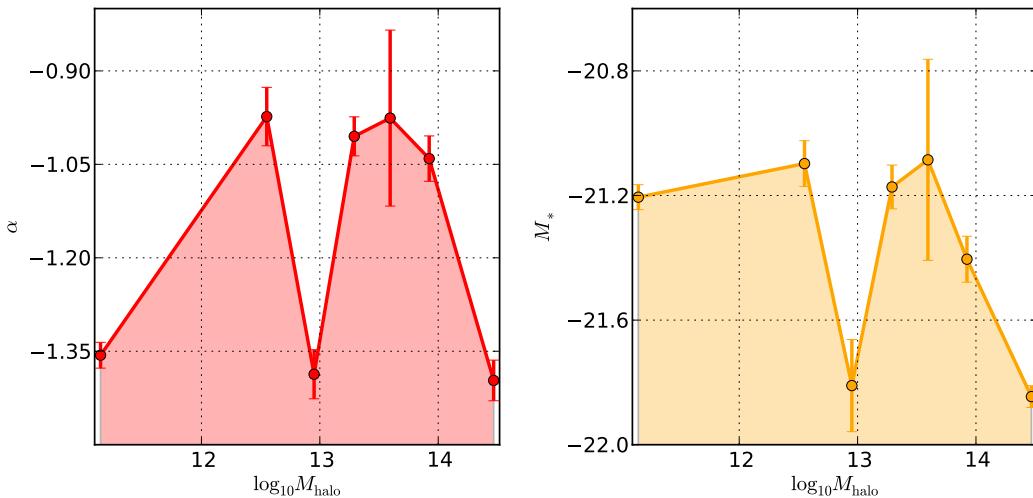


Figure 2.47: Modulation of the parameters for a Schechter distribution with the halo mass in a mock catalogue constructed with the galaxy catalogue from the Guo2010a, without taking a K-decoration for apparent magnitudes of galaxies and removing galaxies in groups that are closer to the border of cube simulation on the mock’s cube.

We can see that the modulation of this parameters with the halo mass isn’t the same as in the figure (2.45) for the Schechter distribution. It’s not very clear why when we construct the mock catalogue, we can’t recover the same parameters as for the data used to build this mock catalogue. It’s maybe a problem of spatial truncation of data in flux and redshift that may cause some variations on the intrinsic LF of the data from the galaxy catalogue used for the mock catalogue. The mock catalogue have a problem: periodic conditions in the box used in order to build this mock can move away from each other galaxies belonging to the same group. So in groups of a certain halo mass, we have more missing data than expected and the estimation of parameters is affected too. To see if this assumption is correct, we have removed from the sample of galaxies in the mock catalogue these ones that are in groups too close to the border of the box in the mock catalogue. The modulation is shown on the figure (2.47).

We can’t find parameters as in the galaxy catalogue from Guo et al. [3] both with modulation of the halo mass and for global data. The behaviour is the same as the latter situation.

The last test to understand why we can’t find the same parameters is described in what follows. We have used the algorithm described in the appendix **Manuel: add the description in the appendix to generate a galaxy population from a simulation** in order to create a sample of galaxy following a NFW profil in halos, with a velocity dispersion calculated using the Mamon and Lokas [18] model for the anisotropy factor. Galaxy luminosities in the halo are generated in order to have a linear modulation of the parameters of a Schechter distribution with the halo mass. We imposed this modulation in the data of the galaxy sample generated with the HOD model of [20]. The magnitude sample has  $-23 < M_r < -12$ . The result of this modulation in the complete sample is shown in figure (2.48).

We now construct an other mock catalogue using this galaxy sample in order to see what happened when we fix the modulation in the parameters with the halo mass in a flux limited sample. Results are shown on the figure (2.48)

The mock catalogue we have created goes to a redshift of 0.1 while the mock with the galaxy sample from Guo et al. [3] is limited to a redshift of 0.3. The modulation of the parameters we have imposed in the galaxy sample is well recovered. In the case of  $M_*$ , the estimation is always good,

## 2.7. DETERMINE THE LF

## CHAPTER 2. GROUP FINDER ALGORITHM

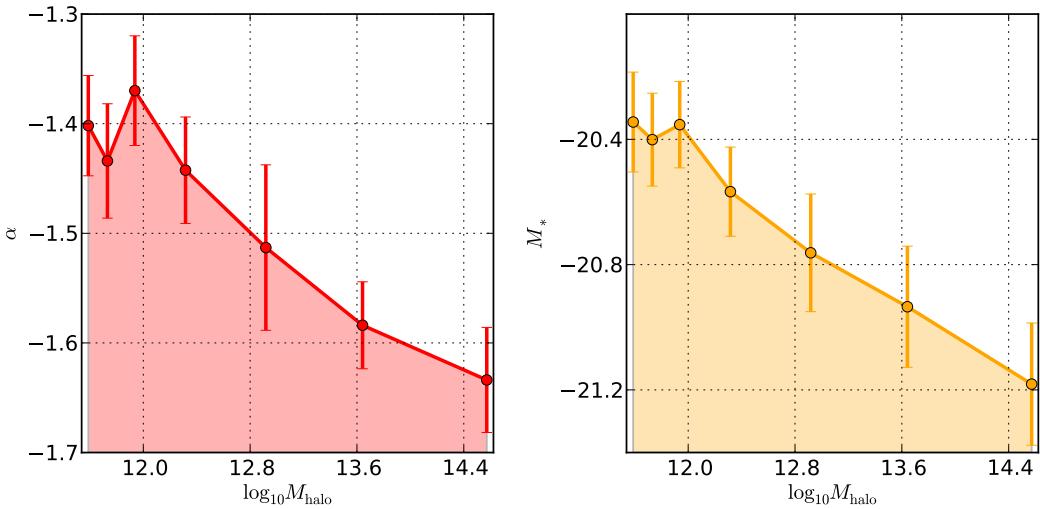


Figure 2.48: Modulation of the parameters for a Schechter distribution with the halo mass in the sample of galaxy constructed using a HOD model, and imposing a linear evolution of the parameters with halo mass. We have imposed that  $\alpha$  goes from  $-1.2$  to  $-1.7$  and  $M_*$  from  $-19.5$  and  $-21.5$  between the extreme halo mass of the simulation.

same with a DS. But there are more uncertainties in finding the slope of the LF with both Schechter and DS. Slopes are less constrained by the data when with have a flux limited sample.

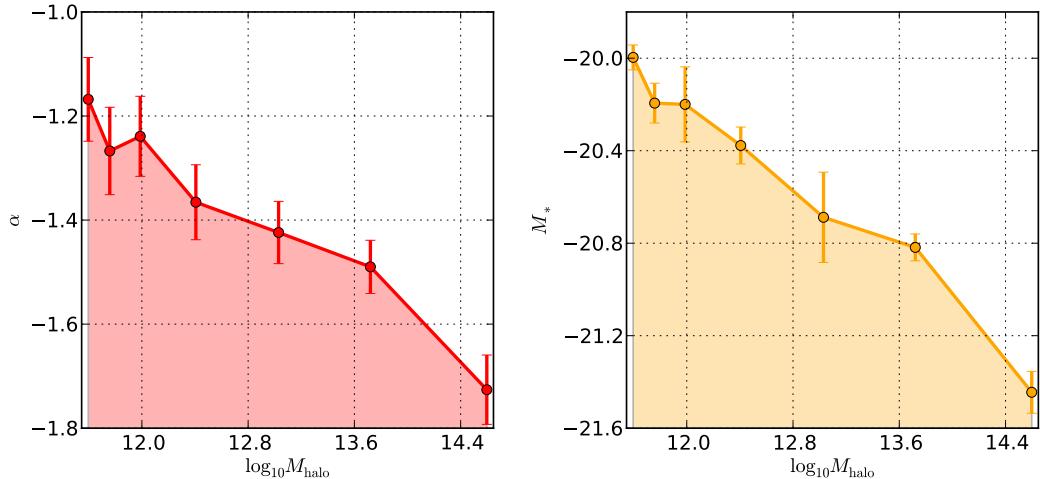


Figure 2.49: Modulation of the parameters for a Schechter distribution with the halo mass in the mock catalogue constructed as described in the previous figure. Parameters are recovered with a given uncertainties.



# Bibliography

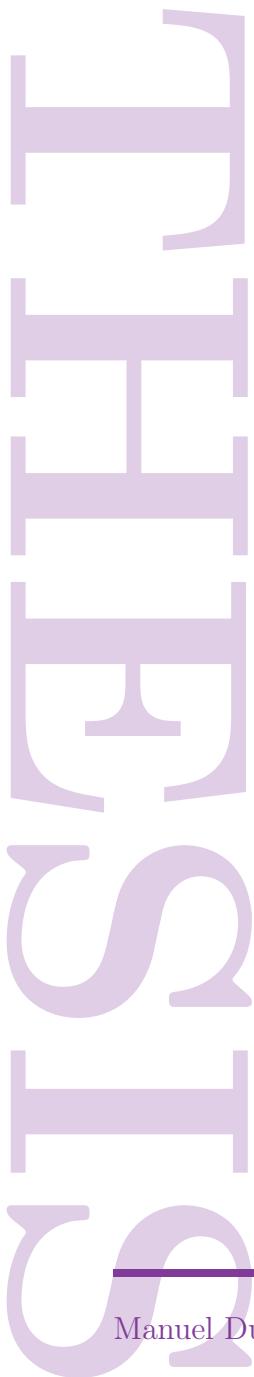
- [1] G. A. Mamon, A. Biviano, and G. Murante. The universal distribution of halo interlopers in projected phase space. Bias in galaxy cluster concentration and velocity anisotropy? *ApJ*, 520:A30+, September 2010.
- [2] X. Yang, H. J. Mo, F. C. van den Bosch, A. Pasquali, C. Li, and M. Barden. Galaxy Groups in the SDSS DR4. I. The Catalog and Basic Properties. *ApJ*, 671:153–170, December 2007.
- [3] Q. Guo, S. White, M. Boylan-Kolchin, G. De Lucia, G. Kauffmann, G. Lemson, C. Li, V. Springel, and S. Weinmann. From dwarf spheroidals to cD galaxies: simulating the galaxy population in a  $\Lambda$ CDM cosmology. *MNRAS*, pages 164–+, February 2011.
- [4] M. Boylan-Kolchin, V. Springel, S. D. M. White, A. Jenkins, and G. Lemson. Resolving cosmic structure formation with the Millennium-II Simulation. *MNRAS*, 398:1150–1164, September 2009.
- [5] J. Blaizot, Y. Wadadekar, B. Guiderdoni, S. T. Colombi, E. Bertin, F. R. Bouchet, J. E. G. Devriendt, and S. Hatton. MoMaF: the Mock Map Facility. *MNRAS*, 360:159–175, June 2005.
- [6] I. V. Chilingarian, A.-L. Melchior, and I. Y. Zolotukhin. Analytical approximations of K-corrections in optical and near-infrared bands. *MNRAS*, 405:1409–1420, July 2010.
- [7] T. Wickramasinghe and T. N. Ukwatta. An analytical approach for the determination of the luminosity distance in a flat universe with dark energy. *MNRAS*, 406:548–550, July 2010.
- [8] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, volume 1. Cambridge University Press, 2 edition, September 1992. ISBN 9780521430647.
- [9] J. F. Navarro, C. S. Frenk, and S. D. M. White. A universal density profile from hierarchical clustering. *ApJ*, 490:493–+, December 1997.
- [10] M. R. Blanton, R. H. Lupton, D. J. Schlegel, M. A. Strauss, J. Brinkmann, M. Fukugita, and J. Loveday. The Properties and Luminosity Function of Extremely Low Luminosity Galaxies. *ApJ*, 631:208–230, September 2005.
- [11] V. R. Eke, C. M. Baugh, S. Cole, C. S. Frenk, P. Norberg, J. A. Peacock, I. K. Baldry, J. Bland-Hawthorn, T. Bridges, R. Cannon, M. Colless, C. Collins, W. Couch, G. Dalton, R. de Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, K. Glazebrook, C. Jackson, O. Lahav, I. Lewis, S. Lumsden, S. Maddox, D. Madgwick, B. A. Peterson, W. Sutherland, and K. Taylor. Galaxy groups in the 2dFGRS: the group-finding algorithm and the 2PIGG catalogue. *MNRAS*, 348:866–878, March 2004.
- [12] A. V. Macciò, A. A. Dutton, and F. C. van den Bosch. Concentration, spin and shape of dark matter haloes as a function of the cosmological model: WMAP1, WMAP3 and WMAP5 results. *MNRAS*, 391:1940–1954, December 2008.
- [13] W. H. Press and P. Schechter. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *ApJ*, 187: 425–438, feb 1974.
- [14] M. S. Warren, K. Abazajian, D. E. Holz, and L. Teodoro. Precision determination of the mass function of dark matter halos. *ApJ*, 646:881–885, August 2006.
- [15] J. Tinker, A. V. Kravtsov, A. Klypin, K. Abazajian, M. Warren, G. Yepes, S. Gottlöber, and D. E. Holz. Toward a halo mass function for precision cosmology: The limits of universality. *ApJ*, 688:709–728, December 2008.
- [16] S. M. Carroll, W. H. Press, and E. L. Turner. The cosmological constant. *Annu. Rev. Astron. Astrophys.*, 30:499–542, 1992.
- [17] F. C. van den Bosch. The universal mass accretion history of cold dark matter haloes. *MNRAS*, 331:98–110, March 2002. doi: 10.1046/j.1365-8711.2002.05171.x.
- [18] G. A. Mamon and E. L. Łokas. Dark matter in elliptical galaxies - ii. estimating the mass within the virial radius. *MNRAS*, 363: 705–722, November 2005.
- [19] James Binney and Scott Tremaine. *Galactic Dynamics*. Princeton Series in Astrophysics, 1987.

## BIBLIOGRAPHY

BIBLIOGRAPHY

---

- [20] I. Zehavi, Z. Zheng, D. H. Weinberg, M. R. Blanton, N. A. Bahcall, A. A. Berlind, J. Brinkmann, J. A. Frieman, J. E. Gunn, R. H. Lupton, R. C. Nichol, W. J. Percival, D. P. Schneider, R. A. Skibba, M. A. Strauss, M. Tegmark, and D. G. York. Galaxy clustering in the completed sdss redshift survey: The dependence on color and luminosity. *ApJ*, 736:59, July 2011.
- [21] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth Dover printing, tenth GPO printing edition, 1964.
- [22] D. W. Hogg. Distance measures in cosmology. *ArXiv Astrophysics e-prints*, May 1999.
- [23] J. Berstel and J. É. Pin. Programmation et Algorithmique. École Polytechnique.



# Chapter 3

## Analyse du SDSS-DR8





## Chapter 4

# Morphologies dans le SDSS





# Chapter 5

## Approche analytique



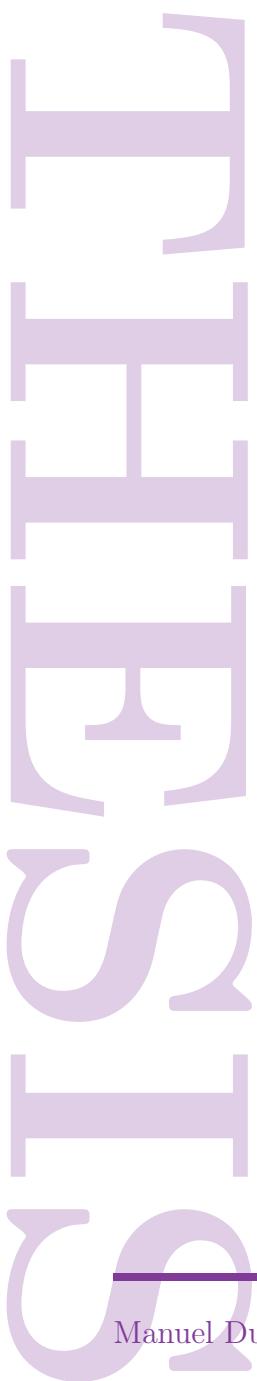
TESTING  
MANUAL  
DUARTE

---

# Chapter 6

## Modélisations numériques





# Appendix A

## Analysing data on the SDSS-DR8

### A.1 Introduction

In order to realize the mock catalogue for the group finder algorithm, we need to mimic the SDSS. This mock have to be realistic. But our algorithm needs to have the redshift for *all* galaxies in the volume selected and the SDSS provides spectroscopic redshifts just for galaxies that could have been targeted due to the problem of fibre collision. So for galaxies in this situation we use the photometric redshift. The problem is that dense regions on the survey are more susceptible to don't have a spectroscopic redshift than a galaxy in a lower dense region. So we have to determine how the fraction of photometric galaxies depends on the density of galaxies in the sky in order to apply this in our mock catalogue.

In the SDSS, there are different ways to estimate photometric so we list the methods here.  
**Manuel:** Add the list of method available in the SDSS.

We can select galaxies in the sample with an SQL query. All queries used will be summary here.

### A.2 Analysis

#### A.2.1 Definitions

In the SDSS there is something called "stripes" which is a band of observations in the sample. Those *bands* can overlap contrary to "chunks" which are similar bands but don't overlapping (they make a complete partition of the survey in their union). We can use this stripes in order to select galaxies in regions of interest for our studies. Data on the SDSS provide limits of this stripes, so we can use it to fix borders of the survey. In reality, in the region of the survey we consider, we don't see overlapping of the stripes. So it is more useful to use them in order to define limits of survey in region of our interest.

Following definitions given in the SDSS website, we can define two other coordinate systems in the survey which we can use to select galaxies.

**Great Circle:** This coordinates system is define with two angle  $(\mu, \nu)$ . Coordinates are relatives to one stripe so it can be use when working with galaxies in the region of the stripe we consider.

**Manuel:** More definitions of this.

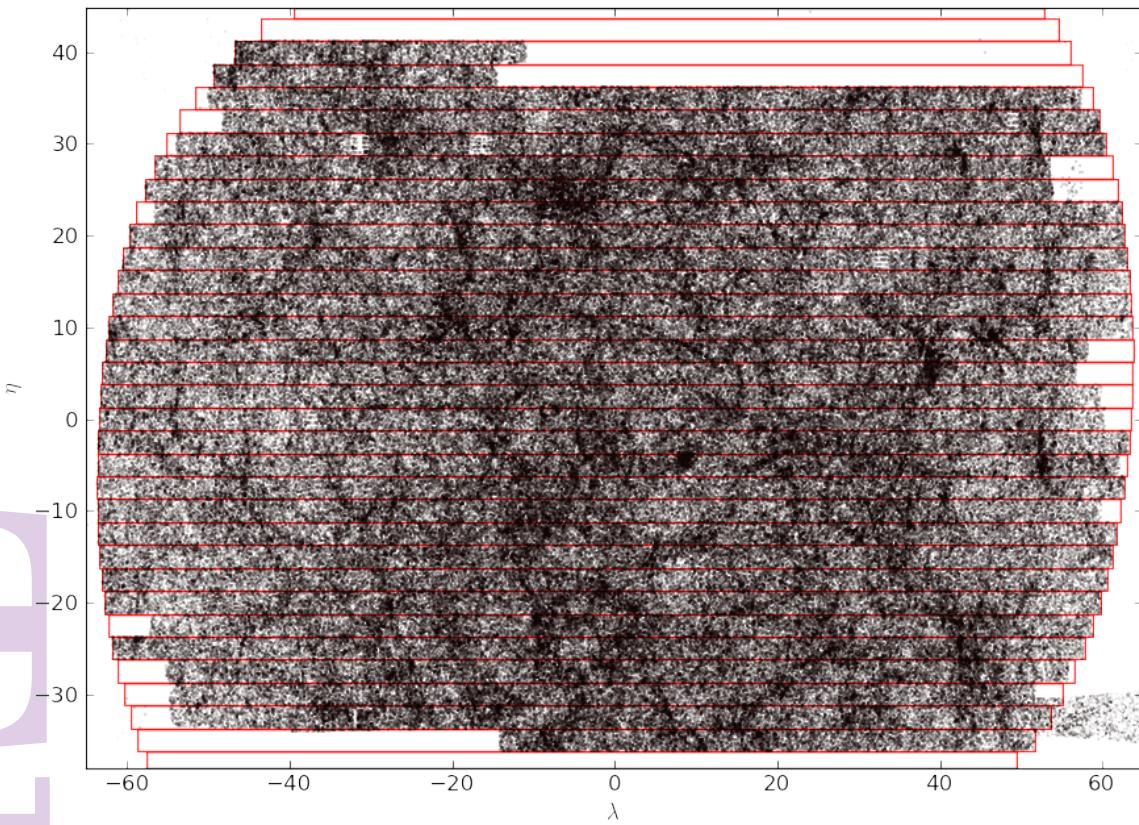


Figure A.1: Spectroscoped galaxies in the SDSS DR8 with stripes limits as planned. Coordinates are in degrees.

**Survey Coordinates:** It's an other system similar to celestial coordinates but "centred" on the "block" of galaxies of the survey that we can see in maps. Coordinates are written  $(\eta, \lambda)$ . If we use celestial coordinates, we have:

$$(0, 90^\circ)_{(\eta, \lambda)} = (275^\circ, 0)_{(\alpha, \delta)} \quad (57.5^\circ, 90^\circ)_{(\eta, \lambda)} = (0, 90^\circ)_{(\alpha, \delta)} \quad (\text{A.2.1})$$

It results from this that  $-\frac{\pi}{2} < \eta < \frac{\pi}{2}$  and  $-\pi < \lambda < \pi$ .

With this informations we can write the transformations between the different coordinate systems.

### Survey coordinates to celestial coordinates

From previous definitions, we see that the relation between those systems is just a rotation. So:

$$\begin{aligned} \delta &= \arcsin(\cos \lambda \sin(\eta + 32.5^\circ)) \\ \alpha &= \text{atan2}(\sin \lambda, \cos \lambda \cos(\eta + 32.5^\circ)) + 185^\circ \end{aligned} \quad (\text{A.2.2})$$

## A.2. ANALYSIS

## APPENDIX A. ANALYSING DATA ON THE SDSS-DR8

### Celestial coordinates to survey coordinates

The inverse transformation is in consequence:

$$\begin{aligned}\eta &= \text{atan2}(\sin \delta, \cos \delta \cos(\alpha - \alpha_0)) - \delta_0 \\ \lambda &= \arcsin(\cos \delta \sin(\alpha - \alpha_0))\end{aligned}\quad (\text{A.2.3})$$

with  $(\alpha_0, \delta_0)_{(\alpha, \delta)} = (0, 0)_{(\eta, \lambda)}$ . We have to apply too periodic conditions in the angles founded by the latter equation in order to have values in the correct range. So conditions are:

$$\begin{aligned}\text{Where } \eta < -90^\circ \text{ or } \eta > 90^\circ : \\ \eta &\rightarrow \eta + 180^\circ \\ \lambda &\rightarrow 180^\circ - \lambda\end{aligned}\quad (\text{A.2.4})$$

$$\begin{aligned}\text{Where } \eta > 180^\circ : \\ \eta &\rightarrow \eta - 360^\circ\end{aligned}\quad (\text{A.2.5})$$

$$\begin{aligned}\text{Where } \lambda > 180^\circ : \\ \lambda &\rightarrow \lambda - 360^\circ\end{aligned}\quad (\text{A.2.6})$$

Determining the number of a stripe to which a galaxy pertains is easy too because stripes are organized such they have a constant width along the  $\eta$  coordinate, with a width of  $2.5^\circ$ . The number of the stripe  $n$  of a galaxy with  $\eta$  position is:

$$n = \text{floor}\left(\frac{(\eta + 58.75^\circ)}{2.5^\circ}\right)\quad (\text{A.2.7})$$

### A.2.2 Galaxies selection

There are many tables in the SDSS saving galaxies and other objects properties extracted from images of the survey. Those tables are the results of different selections in objects detected in images. When crossing objects between images of the survey that overlap, there are some differences of positions between the same object in the two images. So there are possibilities that an object is observed twice or more. In many of those tables, there is no "double objects".

The Galaxy view is a selection from the PhotoPrimary for objects flagged as *galaxy*. The Galaxy view contains the photometric parameters (no redshifts or spectroscopic parameters) measured for resolved primary objects. But we have other useful informations to link with tables that give us photometric and spectroscopic redshifts. There is the `specobjid` to link with spectroscopic redshifts in the table `SpecObj` which doesn't contain duplicates (it's a clean table of `SpecObjAll` with clean

redshifts). If `specobjid=0`, the galaxy doesn't have a spectroscopic redshift. The `objid` is a link to the `Photoz` table which contains all photometric redshifts for galaxies in the `Galaxy` table. Estimation is based on a robust fit on spectroscopically observed objects with similar colors and inclination angle. There is also the `PhotozRF` where estimates are based on the Random Forest technique. Galaxies in the `SpecObj` are limited to  $m_r < 17.77$  and a surface brightness selection **Manuel: Add This !!**. So we need to do the same flux limitations when selecting galaxies on the `Galaxy` table. A possible SQL query for selecting galaxies in this table and link them with redshifts tables could be for spectroscoped galaxies:

```

1 select GG.ra, GG.dec, GG.petroMag_u, GG.petroMag_g, GG.petroMag_r,
2 GG.petroMag_i, GG.petroMag_z, GG.specobjid, GG.objid, Z.z, Z.Zerr
3 from Galaxy as GG, SpecObj as Z
4 where Z.specobjid=GG.specobjid and GG.specobjid!=0 and GG.petroMag_r<17.77
5 and GG.ra<275 and GG.ra>100 and GG.dec>-10 and GG.dec<75

```

and for galaxies which couldn't be spectroscoped:

```

1 select GG.ra, GG.dec, GG.petroMag_u, GG.petroMag_g, GG.petroMag_r,
2 GG.petroMag_i, GG.petroMag_z, GG.specobjid, GG.objid, Z.z, Z.Zerr
3 from Galaxy as GG, Photoz as Z
4 where GG.specobjid=0 and GG.objid=Z.objid and GG.petroMag_r<17.77
5 and GG.ra<275 and GG.ra>100 and GG.dec>-10 and GG.dec<75

```

Limits of stripes are given in the SDSS table `StripeDefs` but this limits aren't actual limits, they are planned limits when survey started. We can see it on the figure(A.1) where planned limits are shown in red and spectroscoped galaxies are the points.

We see that some planned regions aren't still observed (spectroscopically speaking). So we need to define other limits in  $\lambda$  coordinates for that stripes that aren't completes. We find by hand the new limits of stripes which contains spectroscoped galaxies. Now, the survey mask is like in figure(A.2). We will consider just galaxies in this mask in order to find groups in the SDSS. Other galaxies aren't easy in order to define borders of the survey and find groups.

For fibre collisions galaxies, we use galaxies selected in the table of the photometric redshifts and keep galaxies that are in the mask defined previously. Now we have a sample of galaxies in a region of the SDSS for which we can easily characterize borders and where all galaxies, given the flux limit of the SDSS, are presents. There is just the problem of fibre collisions galaxies for which the redshift in our possession is photometric, in consequence less precise than spectroscopic redshifts. But our algorithm is tested on a mock catalogue which is "perfect" if we don't take in account this problem of less robust photometric redshifts. In order to know the behaviour of the algorithm with those problematic redshifts, we need to implement this in our mock catalogue.

## Flags in the SDSS

Galaxies can have some troubles with photometry, resolve etc... due to fit and estimations in the SDSS. in the general case, those objects are flagged with the `clean` property which indicates by 1

## A.2. ANALYSIS

## APPENDIX A. ANALYSING DATA ON THE SDSS-DR8

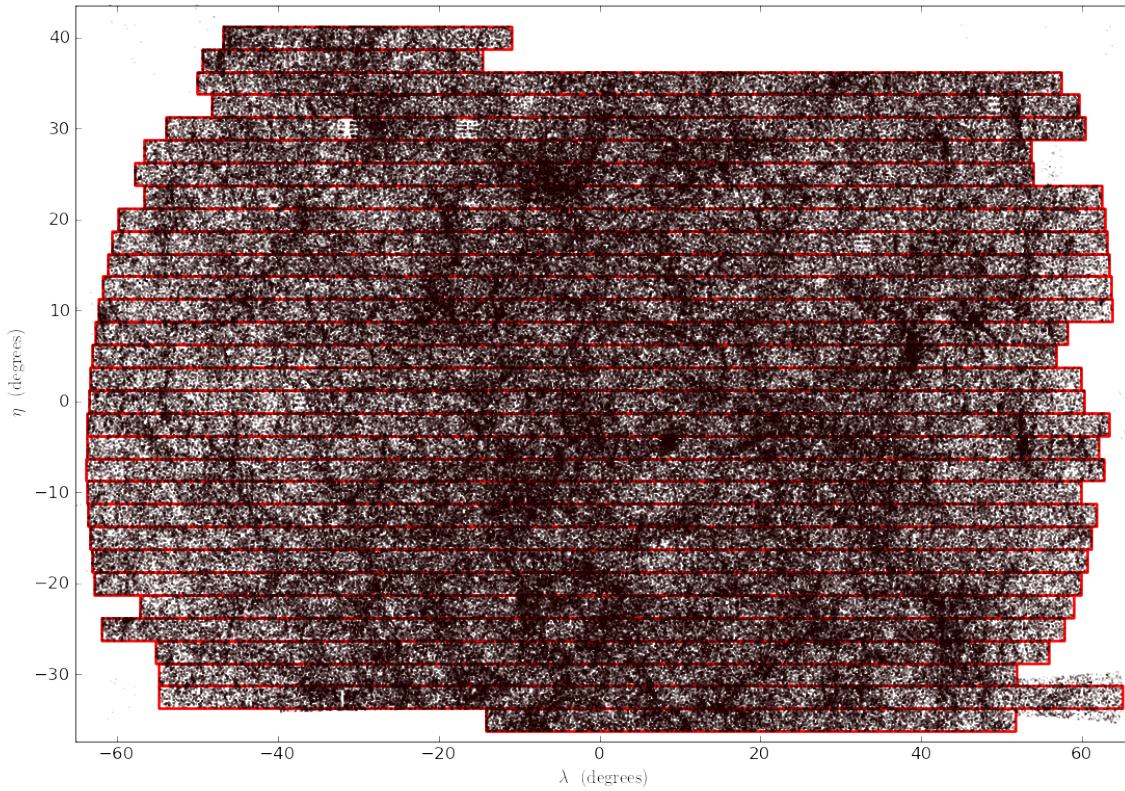


Figure A.2: Spectroscoped galaxies in the SDSS DR8 with stripes limits chosen in order to find easily groups at the border of the survey.

that the photometry is OK and by 0 when there is a problem. Details of the problems are in the bit flag. But for groups, we need to select all galaxies, whether there are not clean.

Galaxy table is a selection from `PhotoPrimary` view for objects with `type = 3` (galaxy). I think that we don't have to care of the "good" photometry of galaxies in the `Galaxy` view, but we can leave a flag in the group finder algorithm to say if a galaxy is in this case.

However, we have to take into account the error on the redshift estimation using the `zErr`. For photometric redshift I think that if the `zErr` is too high, we can use the `nnAvgZ` which is the average redshift of galaxies in the neighbourhood of the considered galaxy. It can be better too if the photometric redshift is too different from it.

The `SpecObjAll` contains duplicates and bad datas. But the `SpecObj` contains just clean spectras. We use `zWarning` to decide if we keep the redshift (`zWarning=0`) or not. In the latter case, we use the photometric redshift instead.

### A.2.3 Fibre collision estimation

In the SDSS, obtaining spectroscopic redshifts of galaxies is done using a plate of  $1.5^\circ$  diameter, in which there is a certain number of fibres in order to get spectrum of the galaxy. But in the field of the plate, the number of fibres is limited, and the number of coverings of a portion of the

sky is limited too because of the time needed to obtain a spectrum. Although runs may overlap, there is sometime galaxies that can't be spectroscopied. Moreover, fibres have a dimension of 55", so when galaxies are closer than this size, one (or more) of those galaxies aren't spectroscopied. We can see that in the figure(A.3) where we have taken the nearest neighbour of a galaxy and determined the differences in angular size and redshift between the two galaxies. As expected, the number of galaxies which are closer than 55" decreases dramatically. There are still some galaxies because the overlapping of runs can permit to get redshifts for galaxies behind this limit.

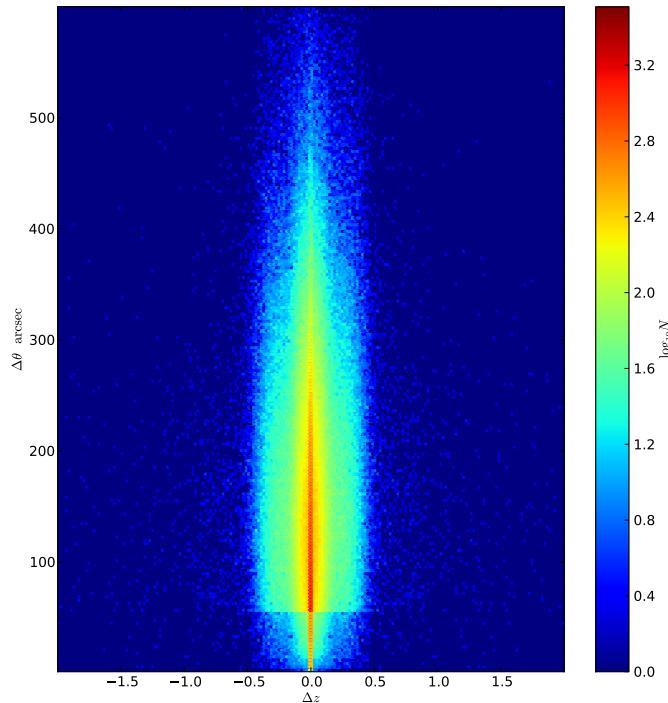


Figure A.3: Distribution of spectroscopic galaxies in the SDSS DR8 in angular size and redshift differences with the nearest neighbour galaxy.

A consequence of those problems is that in denser regions, the number of fibre collision increases, affecting more our groups analysis because the number of photometric redshifts is higher in those dense regions.

We need to implement this selection effect in our mock catalogue. For that we compute the local density in the field, taking all galaxies in the neighbourhood of  $1.5^\circ$  of a galaxy, and in the same time, we determine the fraction of galaxies that don't have a spectroscopic redshift. We deduce of this a relation between the density field and the fraction of fibre collisions. In the mock catalogue, we compute the same density field and we apply the relation estimated in the SDSS sample to the mock. **Manuel: Do it !!** We need for each galaxy to count the fraction of non spectroscopic galaxies in a region of  $1.5^\circ$  radius around. We have to remove galaxies that are to close of the border of the survey, because if we don't remove those galaxies, there are some regions with missing galaxies and the fraction estimation will be affected. The way of selecting those galaxies is to compute a circle of  $1.5^\circ$  around a galaxy, and if a generated point is out of the survey, the galaxy is defined as to be closer to the limits.

## A.2. ANALYSIS

## APPENDIX A. ANALYSING DATA ON THE SDSS-DR8

We can generate samples of points at an angular distance  $d$  to a point of coordinate  $(\alpha_0, \delta_0)$  using formulas of the spherical triangle. If we define a triangle by the pole, the point  $(\alpha_0, \delta_0)$  and the point whose we want coordinates  $(\alpha, \delta)$  denoted  $M$ , we can write the following relations:

$$\begin{aligned}\sin(\alpha - \alpha_0) &= \frac{\sin d \sin \gamma}{\cos \delta} \\ \sin \delta_0 \cos \gamma &= \cos \delta_0 \cot d - \sin \gamma \cot(\alpha - \alpha_0)\end{aligned}\quad (\text{A.2.8})$$

where  $\gamma$  is like a polar angle, which have all the values between 0 and  $2\pi$ . So we have now:

$$\begin{aligned}\alpha - \alpha_0 &= \arctan \left( \frac{\sin \gamma}{\cos \delta_0 \cot d - \sin \delta_0 \cos \gamma} \right) \\ \delta &= \arccos \left( \frac{\sin d \sin \gamma}{\sin(\alpha - \alpha_0)} \right)\end{aligned}\quad (\text{A.2.9})$$

There are problems in poles and equator with those formulas. For a  $\gamma$  limit, angles can't be recovered with those formulas. We have in those cases:

with  $\gamma_0 = \arccos \left( \frac{-\sin \delta_0 \cos d}{\cos \delta_0 \sin d} \right)$

Where  $\delta_0 - d < 0$  and  $\delta_0 > 0$  and  $\gamma_0 < \gamma < 2\pi - \gamma_0$  :  
 $\delta \rightarrow -\delta$

(A.2.10)

Where  $\delta_0 + d > 0$  and  $\delta_0 < 0$  and  $\gamma_0 > \gamma$  or  $\gamma > 2\pi - \gamma_0$  :  
 $\delta \rightarrow -\delta$

(A.2.11)

with  $\gamma_0 = \arccos \left( \frac{\cos \delta_0 \cot d}{\sin \delta_0} \right)$

Where  $\delta_0 + d > \frac{\pi}{2}$  and  $\gamma_0 > \gamma$  :  
 $\delta \rightarrow \alpha + \pi$   
 $\alpha \rightarrow \pi - \delta$

(A.2.12)

Where  $\delta_0 + d > \frac{\pi}{2}$  and  $\gamma > 2\pi - \gamma_0$  :  
 $\delta \rightarrow \alpha - \pi$   
 $\alpha \rightarrow \pi - \delta$

(A.2.13)

$$\begin{aligned} \text{Where } & \delta_0 - d < -\frac{\pi}{2} \text{ and } \gamma_0 < \gamma < 2\pi - \gamma_0 : \\ & \delta \rightarrow \alpha + \pi \\ & \alpha \rightarrow -\pi - \delta \end{aligned} \tag{A.2.14}$$

An other way to draw circles in the sphere is to consider the point for which we want to know celestial coordinates around a given angular distance as the pole of a new coordinate system. In this system, points at given distance of our central point are just points with  $\pi/2 - \delta$  and  $\alpha$  running between 0 and  $2\pi$ . We now can determine cartesian coordinates of those points in this system and apply a rotation to go from the "real" system and the system where the central point is the pole. In the new system we have:

$$\begin{aligned} X' &= r \cos \alpha' \cos \delta' \\ Y' &= -r \sin \alpha' \cos \delta' \\ Z' &= r \sin \delta' \end{aligned} \tag{A.2.15}$$

Then the rotation matrix to go from the "real" system to the new is:

$$R = \begin{pmatrix} \cos\left(\frac{\pi}{2} - \delta_0\right) \cos \alpha_0 & \sin \alpha_0 & \sin\left(\frac{\pi}{2} - \delta_0\right) \cos \alpha_0 \\ -\cos\left(\frac{\pi}{2} - \delta_0\right) \sin \alpha_0 & \cos \alpha_0 & -\sin\left(\frac{\pi}{2} - \delta_0\right) \sin \alpha_0 \\ -\sin\left(\frac{\pi}{2} - \delta_0\right) & 0 & \cos\left(\frac{\pi}{2} - \delta_0\right) \end{pmatrix} \tag{A.2.16}$$

with  $\vec{X} = R\vec{X}'$  where  $\vec{X} = (X, Y, Z)$ . Then we have just to convert those coordinates in celestial angles using:

$$\begin{aligned} \alpha &= \begin{cases} -\arctan2(Y, X) + 2\pi & \text{if } Y > 0 \\ -\arctan2(Y, X) & \text{else} \end{cases} \\ \delta &= \text{sign}(Z) \arccos\left(\frac{\sqrt{X^2 + Y^2}}{\sqrt{X^2 + Y^2 + Z^2}}\right) \end{aligned} \tag{A.2.17}$$

Fibre collisions are more probable in dense region of the sky in projection, but for the mock catalogue we need to quantify this. In order to do that, we have selected for all galaxies in the SDSS survey as defined previously galaxies that are closer than  $1.5^\circ$  in angular size, which is the radius of a plate used for spectroscopy in the SDSS. With that, we can estimate the local density field  $\Sigma_{1.5^\circ}$  in unit of number of galaxies per degree<sup>2</sup>. In this selection, we can determine which galaxies had been spectroscopied or not, and so we can estimate the fraction of non-spectroscopied galaxies. We remove for computing it galaxies that are to close of the border of the survey, and so galaxies which are closer than the radius selected can't be used to search neighbours because some galaxies may be missed and can affect our estimations. Results are shown in the figure (A.4). We can't see the trend we have expected with the density field, so we thought that it can be due to the large region in which we consider galaxies and we ran the same with a radius of  $0.3^\circ$ . Results are in figure (A.5).

## A.2. ANALYSIS

## APPENDIX A. ANALYSING DATA ON THE SDSS-DR8

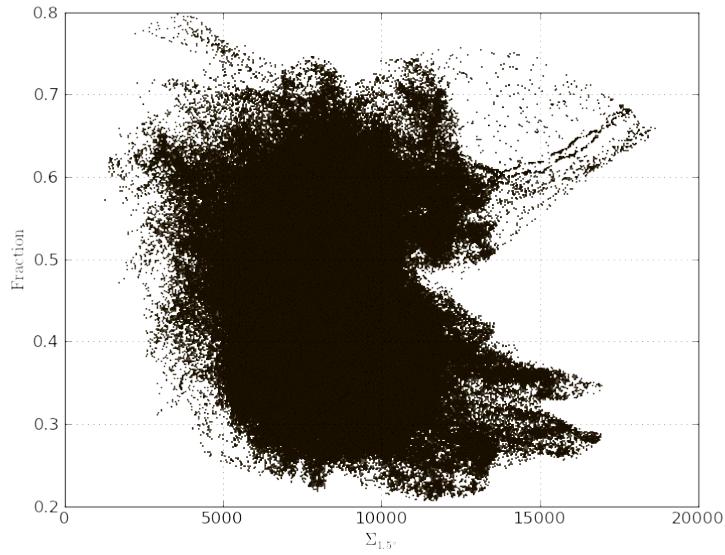


Figure A.4: Fraction of galaxies non-spectroscopic in the SDSS versus the local density field computed in a  $1.5^\circ$  radius region around galaxies not to close than this radius to the border of the survey. Density is in unit of galaxies per degree<sup>2</sup>.

In order to decide which redshift to assign to a galaxy in the mock catalogue which has been chosen to have a photometric redshift, we have estimated the distribution of photometric redshifts versus the spectroscopic redshift in the SDSS sample of spectroscopic galaxies. Results show that a normal distribution is a well fit for those distributions, so we get for this parameters in figure (A.6).

In the mock catalogue, we have interpolated this parameters and we assign a photometric redshift for a galaxy chosen to be in fibre collision.

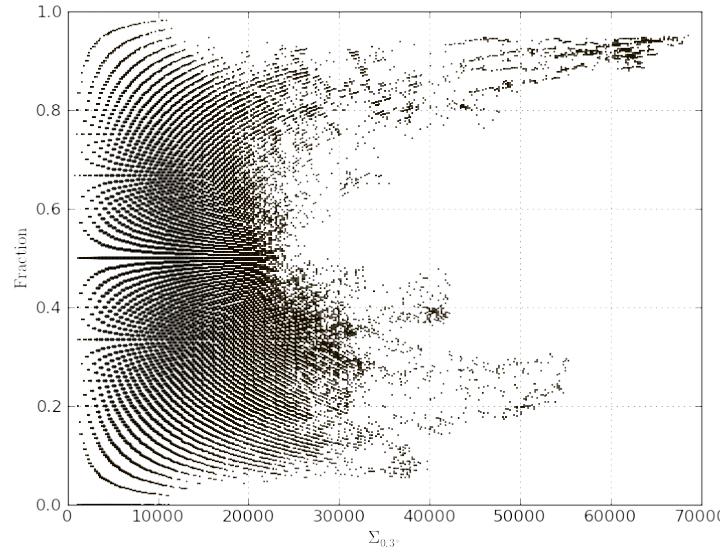


Figure A.5: Fraction of galaxies non-spectroscopic in the SDSS versus the local density field computed in a  $0.3^\circ$  radius region around galaxies not to close than this radius to the border of the survey. Density is in unit of galaxies per degree $^2$ .

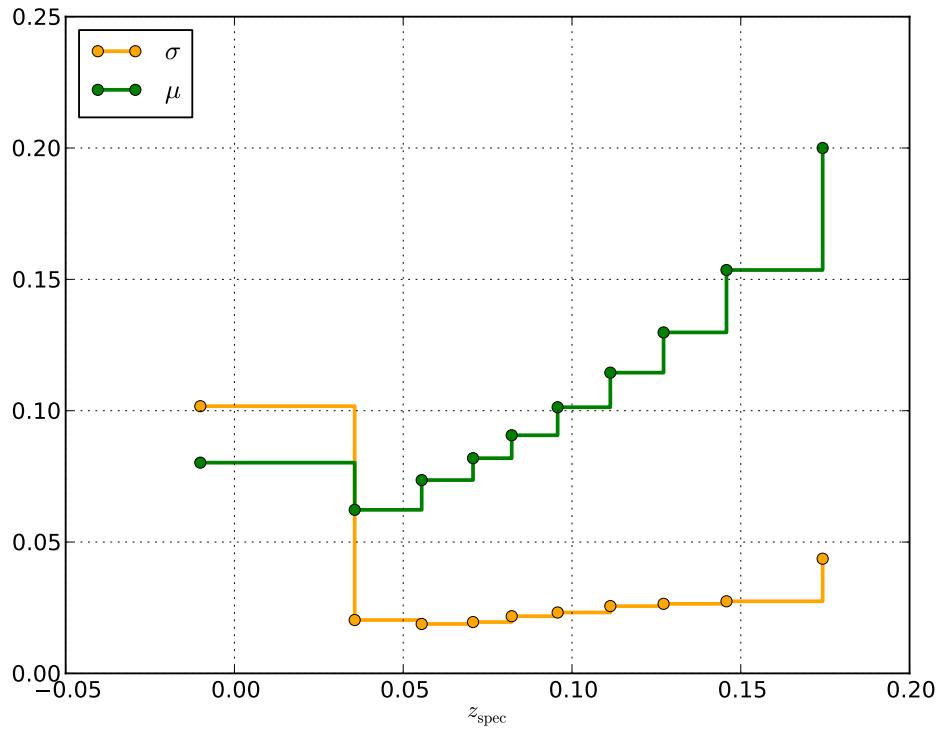


Figure A.6: Parameters of a normal distribution for photometric redshifts versus spectroscopic redshifts in the SDSS.

## Appendix B

### How to generate mock catalogues?



## Appendix C

# Calcul du barycentre lumineux des groupes

---

– "Mieux vaut *tar* que *gz*"  
*Inconnu*

---

On calcule le barycentre lumineux sur la sphère céleste ce qui n'est pas aussi simple que dans une géométrie cartésienne. On va donc adopter la méthode suivante qui est fréquente dans le domaine de la géographie. Comme il est plus simple de calculer un barycentre dans un espace cartésien, on va passer les coordonnées sphériques de la sphère céleste en cartésienne pour calculer le barycentre. Puis une fois ce barycentre trouvé, on le projette sur la sphère céleste en calculant les coordonnées célestes correspondantes.



## Appendix D

# Calcul de la fonction $\Gamma$ incomplète

Le calcul de la fonction  $\Gamma$  incomplète se fait d'une autre façon que celle qui vient à l'esprit en voyant la définition de la fonction. En s'inspirant de la relation fournie dans (Abramowitz and Stegun [21]), on voit que la fonction  $\Gamma$  incomplète peut s'écrire comme:

$$\Gamma(a, x) = \int_x^\infty e^{-t} t^{a-1} dt = \Gamma(a) (1 - P(a, x)) \quad (\text{D.0.1})$$

avec

$$P(a + 1, x) = P(a, x) - \frac{x^a e^{-x}}{\Gamma(a + 1)} \quad (\text{D.0.2})$$

On voit donc que l'on peut faire le calcul de  $P(a, x)$  par récurrence. Pour  $a > 0$ , Press et al. [8] fournit une fonction permettant de calculer la fonction  $P(a, x)$ . Donc il est très simple en utilisant cette *graine* de calculer la fonction  $P(a, x)$  pour toute valeur de  $a$  et de  $x$ . Il reste seulement à connaître la fonction  $\Gamma(a)$  qui peut être calculée à partir d'un développement limité pour tout  $a$ . Cette fonction est comprise dans le standard Fortran et permet en réalisant tout ce qui a été décrit plus haut de calculer  $\Gamma(a, x)$  très rapidement et sans passer par un calcul d'intégrale.

