UNIVERSIDADE DE SÃO PAULO
Instituto de Astronomia, Geofísica e Ciências Atmosféricas

UNIVERSITE DE BORDEAUX 1
Ecole Doctorale des Sciences Physiques et de l'Ingenieur
Laboratoire d'Astrophysique de Bordeaux

THESE DE DOCTORAT

Specialite : Astrophysique, plasmas, corpuscules

présentée par

Alberto Garcez de Oliveira Krone-Martins

Plus loin avec la mission spatiale Gaia grace a l'analyse des objets etendus

Preparé en régime de co-tutelle entre
l'Universidade de São Paulo et l'Université de Bordeaux I

Soutenue à l'Universidade de São Paulo

Directeurs de thèse :	Ramachrisna Teixeira	–	IAG-Universidade de São Paulo
	Caroline Soubiran	–	LAB-Université de Bordeaux I
	Christine Ducourant	–	LAB-Université de Bordeaux I
Examinateurs :	Ronaldo de Souza	–	IAG-Universidade de São Paulo
	Reinaldo de Carvalho	–	INPE-Inst. Nac. de Pesquisas Espaciais
Rapporteurs :	François Mignard	–	Observatoire de la Côte d'Azur
	Dimitri Pourbaix	–	Université Libre de Bruxelles

Lembrete
Se procurar bem você acaba encontrando,
não a explicação (duvidosa) da vida,
mas a poesia (inexplicável) da vida.

Rappel :
Si l'on a bien cherché on finit par trouver,
pas l'explication (douteuse) de la vie,
mais la poèsie (inexplicable) de la vie.
dans *Corpo*, Carlos Drummond de Andrade.

O astrônomo, mesmo na incapacidade de seu pequeno telescópio,
vê o passado das estrelas e o futuro da humanidade.

L'astronome, même dans l'incapacité de son petit télescope,
regarde le passé des étoiles et le futur de l'humanité.
dans *O taikonauta, o astrônomo e o espaço*, Elisa Andrade Buzzo.

Remerciements

D'abord, je remercie mon Epouse, Elisa, qui a été à mon coté pendant toutes ces années et qui m'a si souvent entendu dire « quand je finirai la thèse ». Merci pour son amour et sa patience.

Merci à mes Parents Alberto et Maria Auxiliadora, et à ma soeur Isabela, pour être des modèles d'humanité, de « dedicação », d'équilibre et d'efforts. Je vous remercie aussi pour votre amour, votre soutien et pour m'encourager à suivre mon propre chemin – même si parfois il fallait renoncer au repos.

Je remercie énormément mes Directeurs de thèse, Christine et Rama, pour être des modèles de professionnels, pour votre amitié et pour m'ouvrir de nouvelles portes amplifiant mes horizons dans l'univers de l'Astronomie et de la Science. Je vous remercie aussi pour comprendre mon besoin d'avoir des intérêts divers, et pour la sagesse de m'avoir alerté quand cela était nécessaire. Finalement, je vous remercie pour me rappeler que dans certaines occasions il est nécessaire de laisser les idées mûrir.

Merci à mon Beau-père Eliseu, à ma Belle-mère Célia et à toute ma famille pour votre soutien, votre confiance et votre amour, et aussi pour comprendre de mon manque de présence pendant ces dernières années.

Je remercie Caroline, pour les discussions scientifiques et pour me diriger dans un des chemins parallèles que j'ai suivi, tout en étendant mes connaissances en Astronomie.

De nombreuses personnes doivent recevoir ma reconnaissance et mes remerciements, pour cette thèse qui a été crée sur deux continents. Particulièrement, je remercie :

Xavier, Eduard et Carine pour m'avoir guidé dans les simulateurs de Gaia, et le programme de collaboration Alfa-LENAC pour mon séjour à l'*Universitat de Barcelona*. Albert pour des discussions éclairantes. Guylaine, Benoît et toute l'équipe du CNES pour leur compréhension que parfois les astronomes ne sont pas 100% *compliant*, et Christophe et Laurent pour nous aider a devenir un peu plus *compliant*. Jean-François pour les discussions sur l'algorithmes et pour les Shadoks. Diana pour les discussions sur les codes du CU5. Dimitris pour nos discussions à Athènes, et Panos pour son amitié,sa bonne humeur et pour MAGIL. Didier et Nelson pour vérifier que tout allait bien pendant le développement de cette thèse.

Gustavo, pour m'avoir initié à l'Astronomie professionnelle, pour la direction de mon mestrado (première thèse), et pour m'avoir servi de modèle de chercheur exemplaire.

Les professeurs et astronomes de São Paulo et de Bordeaux, pour l'enseignement des divers branches de l'Astronomie, et aussi d'autres sujets ? même s'ils ne s'en sont pas aperçu. Je vous remercie Paulo, Sylvio, Tatiana, Jorge, Boczko, Laerte, Augusto, Ronaldo, Roberto, Sandra, Alex, João, Amaury, Jane, Vera, Beatriz, Bete, Gastão, Jacques, Zulema, Sílvia, Ademir, Michel, Franck, Sean, Valentine, Didier,

Nathalie, Fabrice, Patrick. Et un remerciement special à Christine, Michel et Rama pour la révision détaillée des chapitres de cette thèse, dans ses versions portugaises et françaises.

Mes amis du Colégio Etapa, en particulier Fabio, Caio, Ulisses, Clayton, Goulart, Ricardo, Leo, Thamy, Sissy et mes amis de l'USP et de Bordeaux, pour nos discussions dans les couloirs, dans nos bureaux, pendant les déjeuners et cafés. Cette thèse n'aurai pas été la même sans nos discussions astronomiques et computationelles, dont certains deviendront des travaux, ou sans les conversations sur « la vie, l'univers et le reste ». Merci Rodolfo, Phillip, Rodrigo, Roberto, Mika, Pedro, Andressa, Tiago, Thiago, Rafaéis, Ulisses, Mônica, Rubens, Laerte, Grazi, Paulinho, Grzeg, Tatiana, Mairan, Carlos, Raí, Márcia, Reinaldo, Márcio, Adriana, Vinícius, Bel, Antoine, Arnaud, Geraldine, Aurelie, Thibauld, Mathieu, Cécile, Pierre, Yann, Anne-Sophie, Ming, Ileana, Adrian, Luis, Ugo.

Le CASP – Clube de Astronomia de São Paulo, et mes amis de vulgarisation de l'Astronomie, en particulier André, Tasso, Lucas, Lívia, Raquel, Bia, Denis, Chicão, Tony, Dorival et Márcio, pour encourager le gens a connaitre e apprendre chaque fois plus sur l'Univers ou ils habitent, tout en ouvrant des fenêtres pour qu'ils se motivent à transformer et améliorer le Brésil, et le monde.

Le Departamento de Astronomia de l'IAG de l'Universidade de São Paulo et le Laboratoire d'Astrophysique de Bordeaux de l'Université de Bordeaux I pour l'infrastructure et pour le soutien de ses équipes techniques et secrétaires. Merci Marina, Conceição, Regina, Cida, Cécile, Annick, Ulisses, Sylvie e Nadège. Le INCT-A pour la machine de calcul Gina.

La FAPESP pour la bourse de doctorat (2006/04251-4). La collaboration CAPES/COFECUB pour la bourse en France (BEX0812/07-2). Le CNES, CNRS, ELSA (FP6-EU) et à Action Spécifique Gaia pour des financement divers dans les réunions et conférences. Je remercie aussi l'ESA et le DPAC pour cet immense mission spatiale qui s'appelle Gaia.

Finalement, je tiens à remercier chaleureusement les rapporteurs et les membres du jury pour avoir accepté cette tâche.

Résumé

Ce travail a comme objectif principal de vérifier s'il est possible de faire de la science avec les observations d'objets étendus qui seront réalisées par la mission spatiale Gaia. Cette mission, l'un des plus ambitieux projets de l'Astronomie moderne, observera plus d'un milliard d'objets dans tout le ciel avec des précisions inédites, fournissant des données astrométriques, photométriques et spectroscopiques.

Naturellement, en fonction de sa priorité astrométrique, Gaia a été optimisé pour l'étude d'objets ponctuels. Néanmoins, diverses sources associées à des émissions étendues seront observées. Ces émissions peuvent avoir une origine intrinsèque, telles que les galaxies, ou extrinsèque, telles que les projections d'objets distincts sur la même ligne de visée, et présenteront probablement de solutions astrométriques moins bonnes.

Pour étudier ces émissions, leurs images bidimensionnelles doivent être analysées. Néanmoins, comme Gaia ne produit pas de telles données, nous avons commencé ce travail en vérifiant si à partir de ses observations unidimensionnelles il serait possible de reconstruire des images 2D d'objets dans tout le ciel.

Nous avons ainsi estimé la quantité de cas sujets à la présence d'émissions étendues extrinsèques, et nous avons présenté une méthode que nous avons développée pour analyser leurs images reconstruites. Nous avons montré que l'utilisation de cette méthode permettra d'étendre le catalogue final de façon fiable à des millions de sources ponctuelles dont beaucoup dépasseront la magnitude limite de l'instrument.

D'un autre coté, dans le cas d'émissions intrinsèques, nous avons premièrement obtenu une estimation supérieure du nombre de cas que Gaia pourra observer. Nous avons alors vérifié qu'après les reconstructions d'images, les codes que nous avons développés permettront de classifier morphologiquement des millions de galaxies dans les types précoce/tardif et elliptique/spirale/irrégulière. Nous avons de plus présenté une méthode que nous avons développée pour réaliser la décomposition bulbe/disque directement à partir des observations unidimensionnelles de Gaia de façon complètement automatique.

Finalement nous avons conclu qu'il est possible d'utiliser beaucoup de ces données qui pourraient être ignorées pour faire de la science. Et que le fait de les exploiter permettra aussi bien la détection de millions d'objets qui dépassent la limite de magnitude de Gaia, que de mener des études sur la morphologie de millions de galaxies dont les structures ne peuvent être révélées qu'à partir de l'espace ou au moyen d'optique adaptative, augmentant un peu plus les horizons de cette mission déjà immense.

Resumo

Este trabalho tem como objetivo principal verificar se é possível fazer ciência com as observações de objetos extensos que serão realizadas pela missão espacial Gaia. Um dos mais ambiciosos projetos da Astronomia moderna, essa missão observará mais de um bilhão de objetos em todo o céu com precisões inéditas, fornecendo dados astrométricos, fotométricos e espectroscópicos.

Naturalmente, devido à sua prioridade astrométrica o Gaia foi optimizado para o estudo de objetos pontuais. Contudo, diversas fontes associadas a emissões extensas serão observadas. Essas emissões podem ter origem intrínseca, como galáxias, ou extrínseca, como projeções de objetos distintos na mesma linha de visada, e deverão ter soluções astrométricas aquém do ideal.

Para estudar essas emissões suas imagens bidimensionais devem ser analisadas. Contudo, como o Gaia não obtém tais dados, iniciamos este trabalho verificando se a partir de suas observações unidimensionais seria possível reconstruir imagens de objetos em todo céu.

Dessa forma, por um lado, nós estimamos a quantidade de casos sujeitos à presença de emissões extensas extrínsecas, apresentamos um método que desenvolvemos para segregar fontes astronômicas em imagens reconstruídas, e mostramos que sua utilização possibilitará estender o catálogo final de forma confiável em milhões de fontes pontuais, muitas das quais estarão além da magnitude limite do instrumento.

Por outro lado, no caso de emissões intrínsecas, primeiro obtivemos uma estimativa superior para o número de casos que o Gaia poderá observar. Então verificamos que após reconstruções de imagens, os códigos aqui desenvolvidos permitirão classificar morfologicamente milhões de galáxias nos tipos precoce/tardio e elíptico/espiral/irregular. Mostramos ainda um método que construímos para realizar a decomposição bojo/disco diretamente a partir das observações unidimensionais do Gaia de forma completamente automática.

Finalmente concluímos que sim, é possível aproveitar muitos desses dados que poderiam ser ignorados para fazer ciência. E que salva-los possibilitará tanto a detecção de milhões de objetos além do limite de magnitude do Gaia, quanto estudos da morfologia de milhões de galáxias cujas estruturas podem ser apenas reveladas do espaço ou por meio de óptica adaptativa, expandindo um pouco mais os horizontes dessa já abrangente missão.

Abstract

The main objective of this work is to determine whether it is possible to do science from the observations of extended objects that will be performed by the Gaia space mission. One of the most ambitious projects of modern Astronomy, this mission will observe more than one billion objects throughout the sky, thus providing astrometric, photometric and spectroscopic data with unprecedented precision.

Naturally, Gaia has been optimized for the study of point-like sources due to its astrometrical priority. Nevertheless, many sources associated with extended emission will be observed. The origins of these extended sources can be either intrinsic, such as galaxies, or extrinsic, such as projections of objects in the same line of sight. In both cases, these sources will have less than optimal astrometric solutions.

In order to study those emissions, their two-dimensional images will be analyzed. Nonetheless, since Gaia will not acquire such images, we begin this work by checking whether it will be possible to reconstruct images anywhere in the sky from the satellite's one-dimensional observations.

Consequently, we, on the one hand, estimate the number of cases which will be subjected to the extrinsic extended emissions, present a method which we developed to analyze the reconstructed images by segregating the different sources and show that the adoption of this method will allow extending the catalogue reliably by millions of point sources, many of which are beyond the limiting magnitude of the instrument.

On the other hand, regarding intrinsic extended emissions, we first obtain an upper limit estimate for the number of cases which Gaia will be able to observe ; then, we verify that the combination of image reconstructions and the use of the codes introduced herein will allow performing the morphological classification of millions of galaxies in early/late types and elliptical/spiral/irregular classes. Afterward, we present a method which we developed to decompose those galaxies into their bulge/disk components directly from the one-dimensional Gaia data in a completely automatic way.

Finally, we conclude that it is possible to harness the data of many of the observations that might otherwise be ignored to do science. Saving these data will allow the detection of millions of objects beyond Gaia's limiting magnitude and the study of the morphology of millions of galaxies whose structures can only be probed from space or through the adoption of adaptive optics, thus somewhat expanding the horizons of this already comprehensive mission.

Table des matières

Chapitre 1

Introduction

"L'Astronomie, (...) c'est elle qui nous a fait une âme capable de comprendre la nature." H. Poincaré[1]

Sommaire

Dans ce Chapitre introductif sera présentée de manière brève la question centrale de ce travail de thèse, ainsi que le contexte dans lequel elle se situe et où elle s'est développée.

Nous présentons une description de l'Astrométrie spatiale, depuis ses débuts avec la mission Hipparcos, qui aidera à mieux comprendre la mission spatial Gaia dans laquelle ce travail s'insère. Cette mission est sans doute l'un des plus importants efforts de l'Astronomie moderne et la plus grande et la plus complexe entreprise de l'Astrométrie dans toute l'histoire.

L'essence de la présente thèse traite justement du traitement d'objets célestes observés par le satellite mais avec des solutions astrométriques problématiques ; en particulier nous avons cherché à montrer ici qu'il sera possible de faire de la science avec une partie des données qui ne seraient pas exploitée par le traitement des données de Gaia.

1. dans *La Valeur de La Science*, 1905.

1.1 Objectif

Un travail scientifique cherche, entre autres choses, à trouver des réponses ou des solutions à un problème, à contribuer à la solution d'un problème spécifique, etc. Dans le travail présenté dans ce document, nous nous sommes proposé d'étendre les horizons de la mission spatiale Gaia en développant des moyens pour répondre à la question suivante :

> *Est-il possible de faire de la science avec des objets observés par le satellite Gaia dont les solutions astrométriques seront problématiques du fait qu'ils possèdent un type quelconque d'émission étendue ?*

Comme nous avons essayé de le prouver clairement dans ce texte, à la fin de ce travail de thèse notre conclusion est oui, il est possible de tirer profit de beaucoup de ces objets dont les solutions astrométriques seront généralement mauvaises, pour faire de la science. Plus que cela encore, les résultats obtenus ici devront ouvrir une nouvelle fenêtre, repousser les horizons, sur l'étude de millions d'objets dont la structure n'a jamais été observée.

Le satellite Gaia a été développé pour étudier le contenu stellaire de notre Galaxie, cependant il observera de nombreux objets avec une structure étendue. Ces sources astronomiques, dont le nombre pourra être de quelques milliers à des dizaines de millions, risquaient d'être simplement ignorées. La raison pour cela est justement qu'ils ne rentrent pas dans le scope de la mission Gaia : ce ne sont pas des sources ponctuelles et l'attentions de la communauté était jusqu'alors, presque qu'exclusivement, concentrées sur l'étude des étoiles.

Les objets astronomiques que nous nous proposons de traiter dans ce travail sont toujours associés à un type quelconque d'émission étendue, étant entendu que celle-ci peut appartenir à deux types distincts : intrinsèques ou extrinsèques.

Les sources intrinsèques sont celles que nous désignons habituellement comme étendues, avec comme un bon exemple les galaxies non résolues dans des étoiles.

Les sources extrinsèques, de leur côté, résultent d'une « émission étendue »provoquée par la proximité angulaire de deux ou plusieurs sources.

L'intérêt de vouloir étudier ces objets pour faire de la science réside dans le fait que, bien que le satellite Gaia n'ait pas été projeté dans ce but, ses observations contiennent des informations de grande valeur, comme par exemple la structure d'un grand nombre de galaxies angulairement petites. Qu'il s'agisse d'objets extragalactiques ou de projections d'objets sur la même ligne de visée, dans tous les cas Gaia permettra d'atteindre une résolution seulement accessible depuis le sol en optique adaptative.

1.2 Comment travailler avec ces objets

En général, on utilise des images pour étudier ce type d'objets et, comme l'a fait son prédécesseur le satellite Hipparcos, Gaia n'a pas été conçu dans le but de produire des images mais au contraire de mesurer des angles entre les différents objets observés. Cette difficulté peut cependant être contournée par une

caractéristique des missions astrométriques spatiales : pour mesurer ces angles avec une grande précision il est nécessaire qu'un objet soit observé dans différents grands cercles, c'est-à-dire, des observations faites à partir de différentes directions de visée. Justement, ces observations à partir de différentes directions de visée sont suffisantes pour reconstruire des images bidimensionnelles. Comment cela peut être fait et les stratégies actuellement considérées pour l'application aux données de la mission Gaia permettant l'étude d'objets étendus, seront présentées dans le Chapitre 2.

Durant la mission Hipparcos (qui sera décrite en Section 1.4), ce type d'analyse a déjà été réalisée et un signal bidimensionnel des sources observées a été reconstruit avec l'utilisation des données de l'un de ses détecteurs même si pour ce satellite les observations étaient complètement unidimensionnelles – un exemple de ces reconstructions peut être vu sur la Figure 1.1. De manière différente de celle de son prédécesseur, Gaia possède une notion de « pixel »dans les données qu'il doit transférer vers la Terre, ce qui facilitera l'analogie de ce signal bidimensionnel reconstruit avec une image de l'objet.

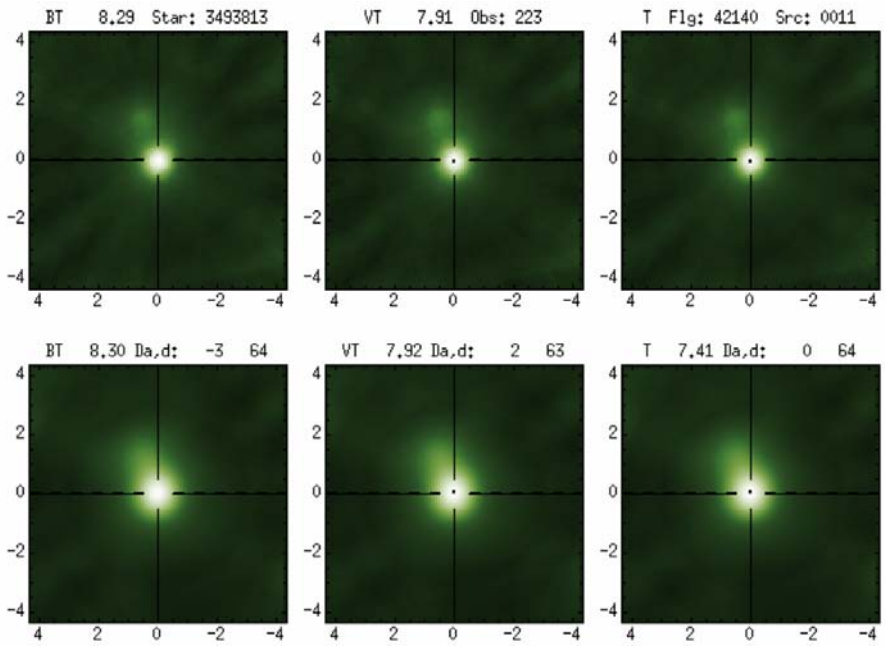


FIGURE 1.1 – Images de l'étoile double HIP76566 reconstruites à partir de la superposition de données des *Star Mappers* du satellite Hipparcos qui ont été produites pour la création du catalogue *Tycho Double Star Catalogue*. (De Fabricius, 2007)

Avec des images reconstruites de Gaia, il sera possible d'analyser tous les objets du ciel jusqu'à une magnitude $G \sim 20$[2] sur des échelles de ~ 50 mas/pixel et avec une *Point Spread Function* de ~ 180 mas. Au cas où il serait possible de reconstruire des images fiables avec les données de ce satellite, ceci représenterait des images significativement meilleures que celles qui peuvent être obtenues à partir de n'importe

2. Les magnitudes G sont obtenues sans filtre, couvrant la région entre 330nm et 1000nm.

quels sites terrestres sans l'utilisation d'optique adaptative.[3] Comme comparaison, le *seeing* de la plus grande partie des observatoires professionnels au sol tourne autour de ~ 1000 mas, pouvant atteindre ~ 600 dans des lieux exceptionnels[4] – un fait qui peut être vu sur la Figure 1.2 – et avec Gaia nous aurons ces images pour le ciel dans son entier.

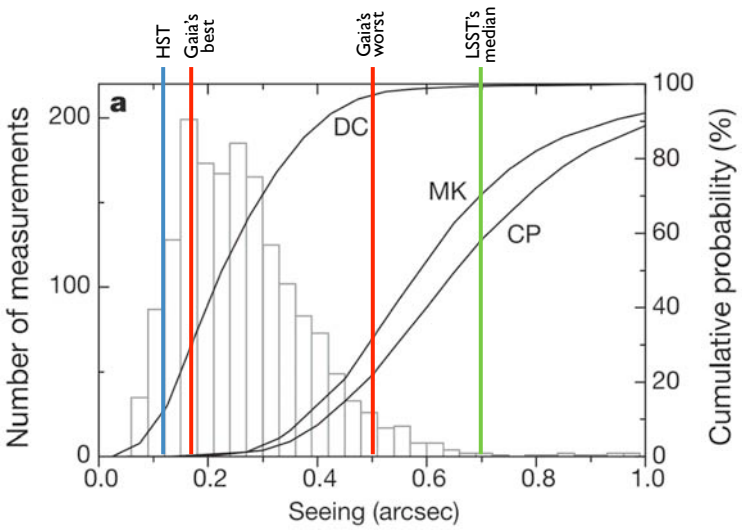


FIGURE 1.2 – Tailles des *Point Spread Functions* pour divers télescopes et sites. Signification des sigles : HST–*Hubble Space Telescope*, LSST–*Large Synoptic Survey Telescope*, DC–*Dome* C (Antártida), MK–Mauna Kea (EUA, Havaí), CP–Cerro Paranal (Antofagasta, Chile), *Gaia's best*–PSF dans la direction de plus grande résolution de Gaia, *Gaia's worst*–PSF dans la direction de la plus petite résolution. Un algorithme de reconstruction d'images parfait devrait être capable de toujours reconstruire des images avec la plus grande résolution de Gaia. (Adaptée de Lawrence et al, 2004)

1.3 Ce que nous pourrons apprendre

Un avantage direct d'analyser des images reconstruites est que, lorsque que l'on combine diverses observations d'une même région du ciel, le bruit de cette « observation combinée »est réduit à chaque nouvelle observation que l'on ajoute – principalement pour la zone dans laquelle l'intersection entre les différentes observations se produit, mais dépendant de l'algorithme de reconstruction, pour d'autres régions de l'image aussi.

Ce bruit plus bas permet que les sources plus faibles que celles qui pourraient être

3. Nous ne considérons pas l'utilisation d'optique adaptative étant donné que celle-ci n'est pas encore utilisée pour de grands relevés, tels que Gaia.

4. Des résultats récents de Lawrence et al (2004) indiquent que le meilleur *seeing* de la planète se trouve dans l'Antarctique (dans un endroit appelé Dome C), où il atteint une valeur intermédiaire de ~ 270.

détectées séparément dans chaque observation puissent être détectées dans l'image reconstruite, augmentant ainsi la magnitude limite de l'instrument.[5]

Durant la production du catalogue Tycho-2, des cartes bidimensionnelles ont été reconstruites avec les données de l'un des détecteurs de Hipparcos (dénommé *Star Mapper*) pour permettre des inspections de solutions douteuses, et non pas pour réaliser des mesures (Høg et al, 2000a). Cependant, par la suite, Fabricius et al (2002) a utilisé des cartes reconstruites pour découvrir 13251 nouvelles étoiles doubles non encore connues, créant ainsi une grande partie du *Tycho Double Star Catalogue*.

Des études sur la fréquence de systèmes multiples sont importantes pour la compréhension des mécanismes de formation et d'évolution stellaire (Burgasser et al, 2007) et, dans les images Gaia reconstruites, des millions de ces systèmes pourront être observés et découverts, et en particulier ceux avec de petites séparations angulaires.

Le faible bruit dans ces images permet que des cas d'objets multiples, mal ou pas résolus, et qui se trouvent sur la même ligne de visée soient analysés. Parmi ces cas, tous les types d'objets astronomiques pourront être observés.

Une classe d'intérêt scientifique particulier qui pourra être rencontrée, est celle des objets appelés les naines brunes. Ce sont des objets de faible masse, qui ne sont pas capables d'entretenir la fusion de l'hydrogène mais qui peuvent brûler le deutérium (ex: Basri, 2000). Il existe divers scénarios proposés pour la formation des naines brunes, une formation de type stellaire étant possible à partir du collapse du nuages moléculaire, ou un autre scénario établit que ces objets sont des embryons éjectés de systèmes proto-stellaires multiples (ex: Reipurth & Clarke, 2001; Bate et al, 2002). Des études de multiplicité des naines brunes (ex. Ahmic et al, 2007) sont importantes pour déterminer quel est le scénario le plus plausible pour leur formation, et pourront être réalisés à partir des données provenant d'analyses d'images reconstruites.

Des objets de types spectraux L et T (Kirkpatrick, 2005; Burgasser et al, 2006), encore plus froids que ceux du type M avec $T_{eff} < 2500$K, pourront peut-être être rencontrés sur ces images, aidant à augmenter leur échantillonnage et permettant de mieux comprendre le bas de la séquence principale.

Dans ces images, des projections d'étoiles ou des objets binaires et multiples de tous les types spectraux dans des amas stellaires, par exemple, pourront aussi être résolus, permettant des études plus détaillées de ces amas, y compris leur caractérisation et l'étude de leur évolution dynamique.[6] Le fait d'étudier des images reconstruites dans ces cas permettrait de résoudre des problèmes de contamination des diagrammes couleur-magnitude dus à la multiplicité et permettrait des études plus détaillées du déficit de naines blanches dans les amas (Fellhauer et al, 2003).

Des études de la multiplicité des systèmes de naines blanches–naines rouges, fournissent des informations sur l'évolution d'objets de faible masse et sur la fonction de masse initiale dans le bas de la séquence principale (Farihi et al, 2010), et pourront être réalisées avec les objets observés sur des images reconstruites. Les naines blanches

5. Comme on le verra dans la section 1.5.1, ce ne seront pas tous les pixels du plan focal de Gaia qui seront envoyés au sol, en fonction d'une sélection faite à bord du satellite.

6. Actuellement, ceci est principalement étudié au moyen de simulations, par exemple Parker & Goodwin (2010) et Ernst et al (2010).

peu lumineuses, éventuellement au delà de la limite de détection du satellite, pourront aussi être rencontrées avec le traitement proposé dont une fraction appartiendra au Halo de notre Galaxie.

Un dernier exemple que l'étude des images reconstruites de ces objets étendus extrinsèques permettra de traiter est les cas de lentilles gravitationnelles, car de multiples mirages de quasars doivent apparaître proches les uns des autres se perturbant mutuellement. Une étude analytique réalisé par Surdej et al (2010) a montré que dans les images Gaia on pourra rencontrer jusqu'à 3500 quasars soumis à un tel effet jusqu'à $V \sim 21$ (des images reconstruites peuvent atteindre une plus grande magnitude limite). Un exemple de ce type de phénomène est le quasar H1413 + 117 (Figure 1.3), découvert dans Magain et al (1988), lequel, en fonction de sa magnitude ($V \sim 17$), doit être observé par Gaia.

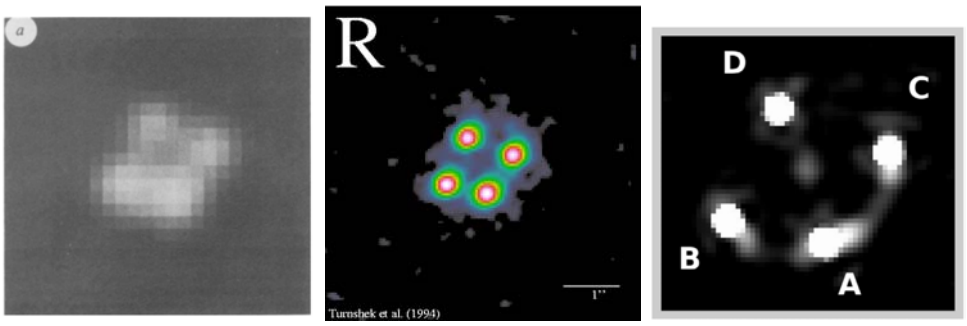


FIGURE 1.3 – Les multiples images du quasar H1413 + 117 (V∼ 17). À gauche : image de la découverte, obtenue par le télescope de 2.2m de l'ESO en 1988 (Magain et al, 1988). Au centre : image obtenue par le HST en 1994 (Turnshek et al, 1997, obtenue sur http://chandra.harvard.edu/photo/2004/h1413/). À droite : image du HST déconvoluée, révélant au centre la galaxie qui a produit le phénomène de lentille gravitationnelle (Chantry & Magain, 2007).

Les exemples d'études citées antérieurement devront être possibles en fonction des développements présentés dans les Chapitres 2, et 3 de cette thèse.

Les Chapitres restants traiteront de l'autre classe d'objets étendus dont l'analyse d'images reconstruites permettra l'obtention d'informations scientifiques : les objets avec une structure morphologique étendue intrinsèque. Parmi ces objets, les cas les plus nombreuses seront certainement les galaxies dont la population stellaire ne peut pas être résolue par Gaia.[7]

En fonction d'une caractéristique du système de détection des objets à bord du satellite, il existe une limite maximum à la taille angulaire des l'objet pour que le transfert des données ait lieu, 700 mas.[8] De cette manière, Gaia devra observer un grand nombre de petites galaxies proches, ce qui permettra d'obtenir des informations

7. D'autres cas seront ceux d'étoiles centrales de nébuleuses planétaires, régions HII, etc.

8. Nous trouverons plus d'informations dans la section 1.5.1.1. 700 mas ne correspond pas au rayon de la plus grande galaxie « non résolue », mais probablement à la taille de sa région la plus brillante

importantes qui peuvent corroborer ou invalider des modèles cosmologiques. Même un simple comptage de ces galaxies est une question qui permet des comparaisons avec les densités de ce type de galaxie prévues par des simulations numériques utilisant, par exemple, le modèle ΛCDM.[9] De plus, des études telles que Benson et al (2007) peuvent être réalisées permettant la détermination de la fonction de masse de trous noirs centraux de galaxies dans l'univers proche.

De plus grandes galaxies, cependant plus distantes et plus brillantes, pourront aussi être observées, permettant des études sur leur morphologie en haute résolution dans tout le ciel. Jusqu'à aujourd'hui, seulement une étude d'observation sur une plus grande échelle a été réalisée pour ce type de galaxie, le *Hubble Medium Deep Survey*, ou MDS (ex. Griffiths et al, 1994). Et bien qu'Abraham et al (1996b) ait utilisé des données du MDS pour des analyses de caractéristiques morphologiques de ces galaxies, seulement un petit pourcentage de la sphère céleste a été couvert, comptant sur une petite quantité d'objets plus brillants ($G \leq 20$) et donc avec de plus grandes incertitudes statistiques.

Bien que la mission spatiale Gaia ne soit pas en premier lieu destinée à l'observation d'objets extragalactiques, ses données pourront être à l'origine du premier relevé de ciel complet de galaxies avec des résolutions spatiales de ~ 180 mas, permettant d'observer des centaines de milliers ou même des millions de galaxies avec ce niveau de résolution.

Cela nous permettra de détailler la morphologie de ces objets, premièrement à partir de leur taxonomie – séparation en classes morphologiques – et ensuite à partir de la décomposition en leurs composants structurels principaux, tels que bulbe et disques, d'obtenir leur paramètres caractéristiques. Ces galaxies pourront être étudiées individuellement, servant aussi de cibles pour des observations dédiées futures. De plus, leurs données pourront aussi être analysées de façon globale (tel que fait par Allen et al, 2006), permettant par exemple de réaliser des études sur l'évolution de la morphologie de ces objets, et des études sur la formation, l'évolution et la stabilité des composants structuraux, comme ceci a été fait par exemple pour les galaxies elliptiques, les bulbe et pseudo-bulbe par Gadotti (2009).

Une complémentarité apparaît entre les deux relevés de ciel entier prévus dans un prochain futur : Gaia transférera le signal des objets ayant une émission centrale inférieure à ~ 700 mas. Il sera complétée par le Large Synoptic Survey Telescope (Le plus grand relevé du ciel du sud, prévu pour 2018) qui observera typiquement des objets au delà de 0.7–1 arcsec.

Nous verrons comment ces galaxies que nous avons citées peuvent être simulées pour des observations avec Gaia dans le Chapitre 4, et comment, à partir des données de cette mission, elles peuvent être classées morphologiquement dans le Chapitre 5 et comment leurs composants structurels sont analysés d'une manière plus détaillée dans le Chapitre 6.

9. Il s'agit du modèle d'Univers qui est actuellement assez accepté. Selon lui, l'univers est gouverné par la constante cosmologique qui est quelque chose qui le force à croître de façon accélérée et il possède de la matière sombre et froide (*Cold Dark Matter*). Pour une description plus détaillée, voir par exemple Freedman (2004).

1.4 Astrométrie et Hipparcos

Tel que commenté antérieurement, ce travail de thèse est inséré dans le contexte de la mission spatiale Gaia. Mais pour mieux comprendre le contexte de cette mission, nous dédions quelques lignes à son prédécesseur, la mission Hipparcos, et à l'Astrométrie d'une manière générale.

L'une des ramifications les plus anciennes de l'Astronomie, et qui durant une longue période de l'histoire de cette science a été pratiquement le synonyme de son côté observationnel, l'Astrométrie[10] est la branche la plus fondamentale de l'Astronomie : c'est sur elle que pratiquement tout la connaissance astronomique se repose. Ceci, car les mesures astronomiques ou sont purement astrométriques ou elles sont dépendantes d'une information astrométrique quelconque pour qu'elles puissent être transformées en connaissance physique : cette information est la distance des objets dans l'espace. La distance est l'une des principales questions traitées par ce domaine de la science.

L'une des questions les plus délicates de l'Astronomie sont les erreurs dans les paramètres physiques obtenus à partir de mesures observationnelles. L'une des composantes principales de ces erreurs provient des incertitudes de la détermination de la distance des objets qui sont étudiés, car ces incertitudes se propagent sur bien d'autres paramètres. La base de pratiquement toutes les méthodes de détermination de distance dans l'univers est la parallaxe trigonométrique,[11] qui durant longtemps a été considérée comme l'une des plus difficiles recherches observationnelles astronomiques – elle a été finalement mesurée par Bessel (1838a,b) sur l'étoile 61 du Cygne.[12]

Néanmoins, jusqu'à la moitié du siècle dernier, il n'existait seulement que quelques étoiles avec une distance déterminée par cette méthode, étant entendu que le nombre d'étoiles dont les mesures de distance pouvaient être considérées fiables était encore plus petit. Face à ce scénario, il était nécessaire augmenter la base sur laquelle repose la plus grande partie des indicateurs de distance.

La parallaxe est d'autant plus difficile à mesurer qu'elle est un effet périodique dans la variation de la position d'un objet céleste d'amplitude très modeste, car les distances astronomiques sont très grandes, et les principaux effets observationnels qui en rendent l'obtention difficile sont des effets causés par l'atmosphère de la Terre (comme le *seeing* et la réfraction), et les variations dans l'instrument utilisé pour réaliser la mesure et entre les instruments divers utilisés pour des mesures de parallaxe. Mais ces problèmes pourraient être résolus dans le cas où l'instrument ne serait pas sur Terre et au cas où un unique instrument serait utilisé pour observer tous les objets.

C'est pour cela qu'une mission spatiale astrométrique été proposé à l'Agence Spatiale Européenne (ESA), développée à partir d'un concept initial de Pierre La-

10. Du grec ἄστρον, étoile, et μέτρον, mesure, donc « mesure des étoiles ».

11. Plus spécifiquement, la parallaxe annuelle, qui est causée par le mouvement de la Terre autour du barycentre du système solaire.

12. Henderson a publié une parallaxe pour α du Centaure deux mois après, et Struve pour α de la Lyre (Vega) un an après Bessel.

croute présenté en 1967 durant la XIIIe Assemblée Générale de l'Union Astronomique Internationale.[13] Cette mission, appelée *High Precision Parallax Collecting Satellite* ou Hipparcos, devint le premier observatoire spatial consacré à l'Astrométrie, et il fut accepté vers 1980 dans l'objectif primaire de déterminer les paramètres astrométriques de ~ 100.000 étoiles (positions, mouvements propres et parallaxes) avec des précisions de 2 mas, inédites jusqu'alors.

Mais, après avoir été lancé, le 8 août 1989, un problème survint ne permettant pas que Hipparcos atteigne l'orbite géostationnaire prévue initialement : le moteur d'apogée ne fonctionna pas, laissant le satellite sur une orbite de transfert, qui obligeait l'instrument à croiser quatre fois par jour la ceinture de radiations de Van Allen. Après avoir confirmé que le moteur d'apogée ne fonctionnerait jamais, les espoirs scientifiques de la mission furent réduites à 5–10% des objectifs originaux (ESA, 1997), ce qui dans la meilleure des hypothèses, permettrait au moins de confirmer que le concept d'astrométrie spatiale fonctionnait. Même ainsi, le satellite résista pendant quatre ans, n'étant finalement mis hors service que le 15 août 1993.[14]

Les résultats obtenus après le traitement des données dépassèrent les attentes, et deux catalogues furent produits : Hipparcos, contenant 118.218 objets avec une précision médiane de 1 mas, en plus de données sur des objets doubles et multiples (y compris des solutions orbitales), et le catalogue Tycho, que contient plus de 10^6 objets, avec des précisions entre 7 et 25 mas. Les deux catalogues contiennent aussi des informations photométriques, avec des précisions qui dépendent de la magnitude du objet, variant entre $0.4 - 7$mmag pour Hipparcos et $10 - 60$ mmag pour Tycho (ESA, 1997).

1.4.1 L'instrument, la mission

Les principaux facteurs grâce auxquels les données du Hipparcos purent être aussi précises furent le fait que ses observations furent réalisées à partir de l'espace, les caractéristiques particulières de son instrumentation et un grand contrôle des erreurs associées aux observations. D'une manière générale, pour la réalisation de travaux astrométriques il existe trois avantages principaux quand on observe les objets hors de l'atmosphère terrestre :

– la position des objets n'est plus affectée par le seeing et par des phénomènes de réfraction. Le seeing est difficilement corrigible sur des grands angles,[15] et la correction de la réfraction est toujours limitée par notre connaissance de

13. Selon Kovalevsky (2005), Lacroute avait déjà émit l'idée en 1965, et la première proposition au Centre National d'Études Spatiales a été faite en 1966 selon Turon & Arenou (2008).

14. Plus de détails sur l'historique de cette mission peuvent être trouvés dans ESA (1997), Turon & Arenou (2008) et Perryman (2010).

15. Actuellement, les systèmes d'optique adaptative ne corrigent les effets atmosphériques que dans quelques secondes d'arc ($\sim 10" \times 10"$), et des systèmes avancés tels que le MCAO (*Multi-Conjugate Adaptive Optics*) du Gemini Sud ou le MAD dans l'ESO-VLT corrigent des champs de $\sim 1' \times 1'$ respectivement (ex: Bouy et al, 2008). Même des systèmes avancés projetés pour le futur, tels que le MOAO *Multiple-Object Adaptive Optics*, ne seront capables de corriger de tels effets que dans des champs de quelques minutes d'arc

l'atmosphère mutable ;
- l'instrument se déplace de manière facilement prévisible lorsqu'il réalise les mesures, différemment de ce qui se passe sur Terre.
- de l'espace il est possible d'observer le ciel entier avec un seul instrument, en éliminant le besoin de comprendre des comportements d'instruments distincts dispersés dans le monde pour la création de catalogues de ciel entier – permettant, de plus, l'obtention d'une astrométrie précise et uniforme sur la sphère.

Du point de vue de l'instrumentation astrométrique, Hipparcos, a introduit une grande innovation, permise justement par le fait qu'il se trouve dans l'espace : l'utilisation de deux miroirs observant des directions différentes du ciel, mais illuminant un même plan focal.[16] De cette manière, connaissant avec une grande précision l'angle entre les miroirs (appelé angle de base), il est possible de déterminer l'angle entre les objets dans le ciel observés par chacun de ces miroirs avec aussi une grande précision – un schéma de ce concept peut être vu sur la Figure 1.4.

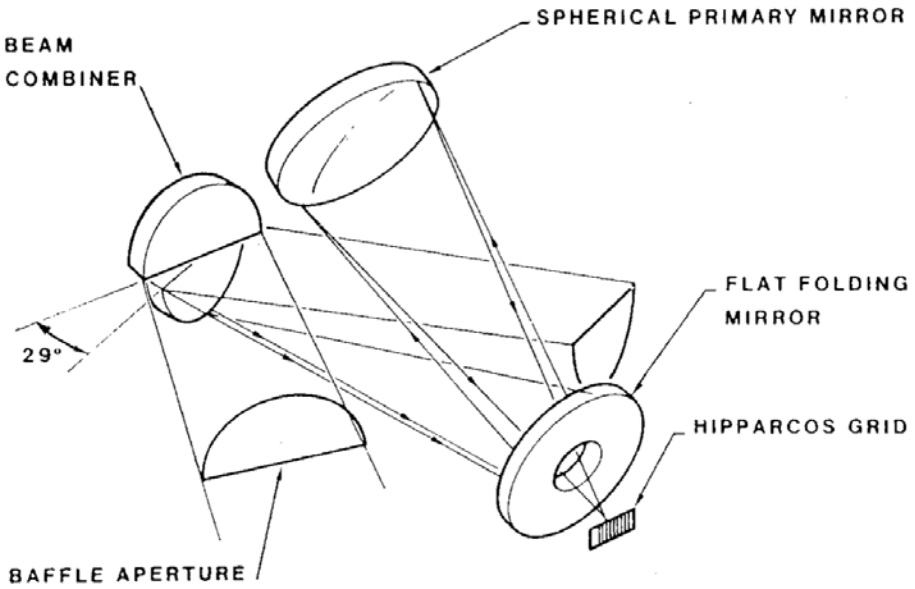
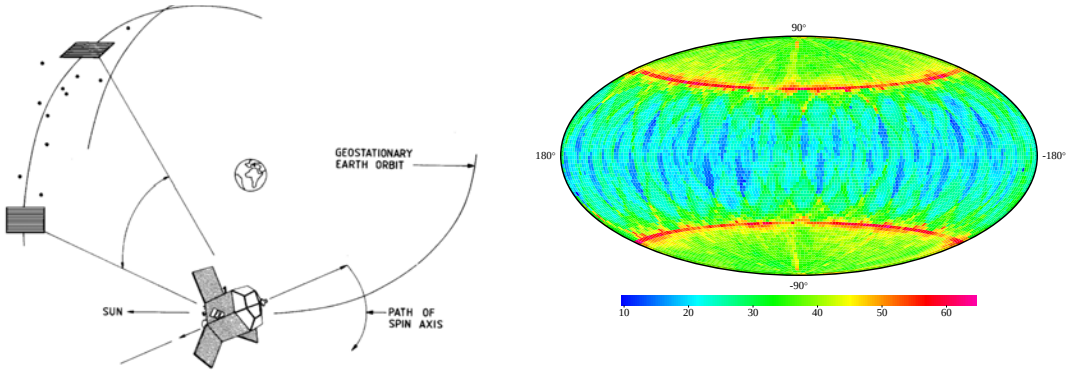


FIGURE 1.4 – Schéma optique du système de champ de vision double du satellite Hipparcos. Les images des deux champs de vision entrent par les *Baffle Apertures*, sont combinées dans une même direction par le *Beam Combiner* (qui possède la forme d'un correcteur Schmidt), réorientés par le *Flat Folding Mirror* et focalisées par le *Spherical Primary Mirror* dans *Hipparcos Grid* (le plan focal). (d' ESA, 1997)

16. Un concept similaire était présent dans l'instrument utilisé par Bessel durant la mesure de la parallaxe annuel, même si cet instrument (un héliomètre produit par Fraunhofer) avait été créé pour l'observation de petits angles. Le concept original de Lacroute utilisait en fait trois miroirs, mais le projet final n'en possédait que deux.

De plus, le principe d'observation d'Hipparcos était basé sur un mouvement de rotation constant du satellite autour de lui même, de manière à ce que le ciel était observé d'une façon analogue à un cercle Méridien sur la Terre, produisant des données qui consistent principalement dans les temps de passage des objets qui croisent le plan focal.[17] Et, pour que le ciel puisse être balayé entièrement, le satellite possédait aussi un mouvement de précession de son axe de rotation. Ces mouvements alliés à l'orbite de la Terre autour du Soleil permirent qu'Hipparcos observe une même région du ciel plusieurs fois, sous différents angles. Tout ce mouvement (appelé de loi de balayage, ou scanning law) a été optimisé pour éviter le Soleil.[18] La Figure 1.5 est une représentation du principe de l'observation et du nombre de fois que chaque région du ciel a été observée par le satellite Hipparcos.



(a) (b)

FIGURE 1.5 – (a) Concept d'observation de grands angles dans le ciel entier avec Hipparcos – l'orbite géostationnaire indiquée sur cette Figure n'a jamais été atteinte. (b) Nombre d'observations dans chaque région de la sphère céleste. (d' ESA, 1997)

La réduction de ces données et la transformation des mesures brutes et d'informations additionnelles dans des paramètres astrométriques (position, mouvement propre et parallaxe) pour toutes les sources a été réalisée en parallèle par deux consortiums indépendants, appelés FAST – *Fundamental Astrometry with Space Techniques* (Kovalevsky et al, 1992) et NDAC – *Northern Data Analysis Consortium* (Lindegren et al, 1992a). Comme il s'agissait de la première fois qu'un projet astrométrique aussi précis était réalisé, il n'existait pas de données indépendantes qui permettaient une vérification externe de la qualité astrométrique des résultats obtenus. La comparaison de deux réductions complètement indépendantes des mêmes données devait servir à détecter l'existence de problèmes graves.

Un troisième consortium, appelé INCA – *Input Catalogue Consortium*, fut responsable pour la préparation du catalogue initial des objets qui devaient être observés par Hipparcos, appelé *Hipparcos Input Catalogue* (Turon et al, 1992). Ce catalogue a

17. Un cercle Méridien détermine le temps de passage d'un objet au méridien du lieu.

18. Cette méthode a été utilisée pour la première fois dans l'espace par Hipparcos, et elle est actuellement utilisée par le satellite Planck, et sera aussi utilisée par Gaia.

été construit à partir de la sélection de propositions scientifiques.[19] En tout, plus de 200 scientifiques travaillèrent sur cette mission spatiale (ESA, 1997).

Durant la mission ont été accumulés $\sim$ 1Tbits de données brutes, ce qui constituait le plus grand problème d'analyse de données jamais rencontré par l'Astronomie.

1.4.2 Réduction des données

La mort du satellite, du télescope en lui-même, ne marqua pas la fin de la mission, et même si les consortiums de réduction avaient déjà travaillé avec les mesures du satellite durant les observations (ex Mignard et al, 1992; Lindegren et al, 1992b), la réduction dura plusieurs années. Des descriptions détaillées des procédures de réduction adoptées sont rapportées dans Perryman et al (1992), Kovalevsky et al (1992), Lindegren et al (1992a) et principalement ESA (1997).

Les deux consortiums utilisèrent un processus divisé en trois étapes, et la principale différence se trouvait dans les algorithmes et les méthodes adoptés pour l'implémentation de ce processus. La solution était fondamentalement divisée en :

1. Traitement initial des comptages de photons ;
2. Réduction de grand cercle ;
3. Solution de la sphère.

Le premier pas consistait dans la transformation entre le comptage des photons réalisé par les photomultiplicateurs des *Star Mappers* et par l'*Image Dissector Tube* de l'instrument principal dans un signal que puisse être utilisé pour le traitement astrométrique subséquent. Pour cela il était nécessaire de déterminer l'attitude du satellite simultanément. Ce pas est divisé en trois étapes :

- obtention de l'estimation de l'attitude et des temps de passage dans les *Star Mappers* ;
- reconstruction de l'attitude à partir des données de l'instrument principal (*Image Dissector Tube*) en utilisant les temps de passage obtenus dans l'étape antérieur et les positions de catalogue ;
- détermination de la phase et de l'amplitude du signal utilisant la reconstruction de l'attitude.

Le deuxième pas, qui est appelé réduction de grand cercle, consistait à transformer ces temps de passage en valeurs d'abscisse le long d'un grand cercle.[20] Ceci était fait au moyen de calibrage de distorsions des projections des objets et de l'amélioration de l'attitude du satellite dans la direction de ce grande cercle. Le résultat était le pôle du grande cercle, une phase préliminaire pour le signal et les abscisses.

Le troisième et dernier pas, appelé de solution de la sphère, est la détermination de phases cohérentes pour tous les grands cercles observés conjointement avec les paramètres astrométriques. Fondamentalement, pour un nombre donné de grands

19. Plus de 118.000 étoiles (quelques unes dans les Nuages de Magellan), un quasar (3C273), 48 astéroïdes et trois satellites (Europa, Iapetus et Titan) ont été sélectionnés pour être observés par Hipparcos (Turon et al, 1992).

20. Dans chaque grand cercle étaient observées environ deux mille étoiles.

cercles, ceci consistait dans la solution d'un problème de moindres carrés des paramètres de chaque objet$(\alpha, \delta, \varpi, \mu_\alpha \cos\delta, \mu_\delta)$, d'une correction pour le point zéro de chaque grand cercle et de variables additionnelles qui représentaient des erreurs systématiques, des problèmes d'harmoniques non traités, des corrections chromatiques, etc.[21] Le problème était représenté par un système matriciel qui équivalait à plus de 2 millions d'équations.

Un exemple avec des positions d'un objet observées dans différents grands cercles peut être vu sur la Figure 1.6 – les segments de ligne droite représentent les valeurs des abscisses déterminées à chaque passage et la ligne courbe est la trajectoire apparente de l'objet inférée à partir de la solution du problème de moindres carrés (le problème est résolu de façon simultanée pour tous les objets observés).

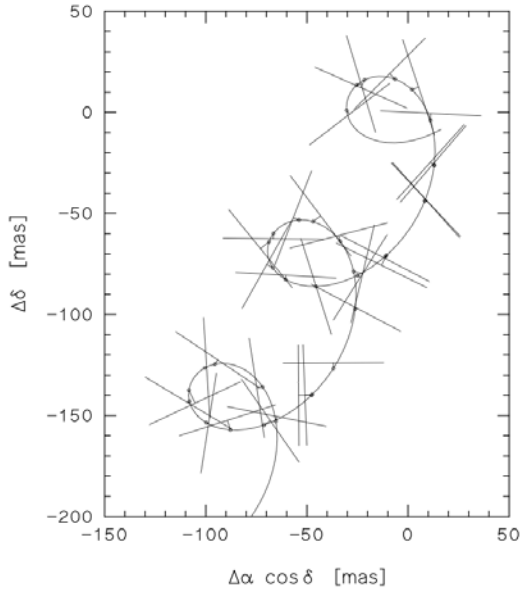


FIGURE 1.6 – Représentation des observations d'Hipparcos d'un objet donné et du résultat obtenu après la solution de la sphère. (De ESA, 1997)

Ce problème pourrait être résolu par un traitement direct du problème global, à partir de la phase et de l'amplitude du signal observé. Ceci néanmoins n'était pas faisable avec les ordinateurs disponibles à l'époque d'Hipparcos, en fonction du coût computationnel élevé : $\sim 10^{16}$ floating points operations et $\sim$ 500Gb de mémoire.[22] Néanmoins, déjà à cette époque une solution globale approchée fut tentée, appelée Global Iterative Solution, dont le principe a été proposé par P. Lacroute en 1978, mais qui a eu sa première mise en œuvre par L. Lindegren (ESA, 1997, vol.3, p.490).

21. Pour plus de détails, voir ESA (1997).
22. Actuellement ce problème est traitable même avec des ordinateurs de taille moyenne.

1.4.3 Tycho

En plus du catalogue Hipparcos, cette mission généra un deuxième catalogue, moins précis mais avec une quantité significativement supérieure d'objets – dix fois plus grande. Ce catalogue, pensé par E. Hog, a été appelé Tycho et a son origine dans l'utilisation scientifique des données des *Star Mappers* du satellite. Il a été construit par le consortium TDAC – *Tycho Data Analysis Consortium* à partir de la comparaison des observations des *Star Mappers* avec des prévisions de ces observations basées sur un catalogue d'entrée et des déterminations des corrections nécessaires. Les méthodes utilisées sont décrites dans ESA (1997), v.3.

1.4.4 La post-mission

Trois ans après la publication des catalogues Hipparcos et Tycho (ESA, 1997) et de la dissolution des consortiums, les données qui furent à l'origine du Tycho ont été re-analysées à partir de nouveaux algorithmes basés sur l'analyse de passages accumulés et de l'utilisation de meilleurs catalogues auxiliaires, générant Tycho-2 (Høg et al, 2000c,a,b) – celui-ci est devenu l'un des plus importants catalogues de l'Astronomie, et sert pour lier divers catalogues modernes au référentiel Hipparcos.[23]

Et même après plus d'une décennie de la dissolution des consortiums FAST et NDAC, les données brutes d'Hipparcos continuèrent à être analysées. En 2008, une nouvelle réduction a été publiée dans van Leeuwen (2007), dans laquelle à partir d'un meilleur modelage de la dynamique du satellite, d'impacts de micro-météorites et d'une solution globale moderne (comme celle qui sera utilisée pour Gaia) a permis d'améliorer les paramètres astrométriques, spécialement d'étoiles brillantes, atteignant des erreurs moyennes de $0.7 - 0.8$ mas, et arrivant à 0.09 mas pour certains cas (van Leeuwen, 2009a).

L'impact scientifique de la réduction originale des données de cette mission astrométrique a été gigantesque, touchant tous les secteurs de l'Astronomie jusqu'à aujourd'hui (plus de détails dans Perryman, 2009). Cependant une question scientifique encore à résoudre est la distance de l'amas stellaire des Pléiades : les résultats obtenus par Hipparcos pour la parallaxe de cet amas ne sont pas en accord avec les distances dérivées d'observations du HST et de l'ajustement de la séquence principale par isochrones théoriques.[24]

Déjà en 1976, presque deux décennies avant la fin de cette mission spatiale, selon Turon & Arenou (2008) le Groupe de Définition de Mission commentant ce qui n'était alors que la proposition du projet Hipparcos

> « recommande fortement, comme le progrès scientifiques le plus significatif, qu'un programme Spatial Astrométrique soit établi, menant au lancement d'un deuxième vaisseau spatial, identique ou similaire au premier, après un intervalle de temps d'une dizaine d'années. »[25]

23. Comme par exemple, des versions initiales du *Sloan Digital Sky Survey* qui utilisaient directement Tycho-2 (Pier et al, 2003), et de plus nouvelles versions utilisent l'UCAC2 qui de son côté est basé sur le matériel utilisé pour la création du Tycho-2 (Zacharias et al, 2004).

24. Pour un résumé actualisé de la littérature sur cette question, voir van Leeuwen (2009b).

25. *strongly recommend, as the most significant scientific improvements, that a Space Astrometry*

1.5 La mission Gaia

My reason for emphasizing precision space astrometry is that (...)
we need to be aware of our scientific roots.
Malcolm Longair, selon Perryman (2010).

Le grand succès de la mission spatiale Hipparcos était déjà accompagné par les préparatifs pour le second pas de l'Astrométrie spatiale. En décollant, Hipparcos était encore muni de photomultiplicateurs tandis qu'en parallèle des détecteurs bien plus efficace dénommés CCDs devenaient de plus en plus aptes à être utilisés dans les conditions extrêmes de missions spatiales.

Durant les années 90 surgirent diverses propositions de nouvelles missions astrométriques, la principale s'appelait GAIA (*Global Astrometric Interferometer for Astrophysics*), et était basée sur un concept d'interféromètre. Ce concept d'observation a été ensuite modifié pour suivre le même principe d'observation qu'Hipparcos, mais dont le plan focal était formé par des CCDs qui en plus d'effectuer des observations astrométriques permettait la mesure des couleurs et des vitesses radiales, donnant origine à l'actuelle mission Gaia.[26]

La Mission Gaia (Perryman et al, 2001; Mignard, 2010) est l'un des plus ambitieux projets de l'Astronomie moderne. Avec son lancement prévu en décembre 2012, le satellite doit réaliser des observations de plus d'un milliard d'objets, générant un catalogue avec des précisions astrométriques inédites : 7 μas jusqu'à une magnitude $G \sim 12$, 25 μas pour $12 < G < 15$, et 0.3 mas entre $15 < G < 20$.[27] En plus des informations astrométriques de qualité sans précédents, le satellite obtiendra des données spectrophotométriques équivalentes à des dizaines de bandes (~ 25) pour tous les objets observables et des données spectroscopiques dans la région du triplet du Ca II (R$\sim$11.500) qui permettront la détermination de la vitesse radiale de tout objet jusqu'à une magnitude $G \sim 17$. La mission prévoit l'observation de tous les objets avec une magnitude apparente allant jusqu'à $G \sim 20$, englobant entre un et deux pour cent de toutes les étoiles de la Galaxie, en plus de millions d'objets extragalactiques.

L'énorme impact de la mission Gaia peut être entrevu par une rapide analyse des résultats espérés pour la détermination de distances par la méthode de parallaxe trigonométrique. Actuellement, il existe moins de 500 étoiles avec des erreurs de parallaxe inférieures à 1%, provenant de la mission Hipparcos, avec Gaia ce nombre devra passer à 10 millions d'étoiles jusqu'à 2.5 kpc. Avec une erreur relative allant jusqu'à 10%, nous avons aujourd'hui des parallaxes d'un peu plus de 20 mille étoiles, et après Gaia celui-ci doit être plus grand que $\sim$ 100 millions pour des distances allant jusqu'à 25 kpc.

Cette base de données sans précédents, aussi bien du point de vue de la quantité que de la qualité, permettra une compréhension bien plus précise de la Galaxie sous

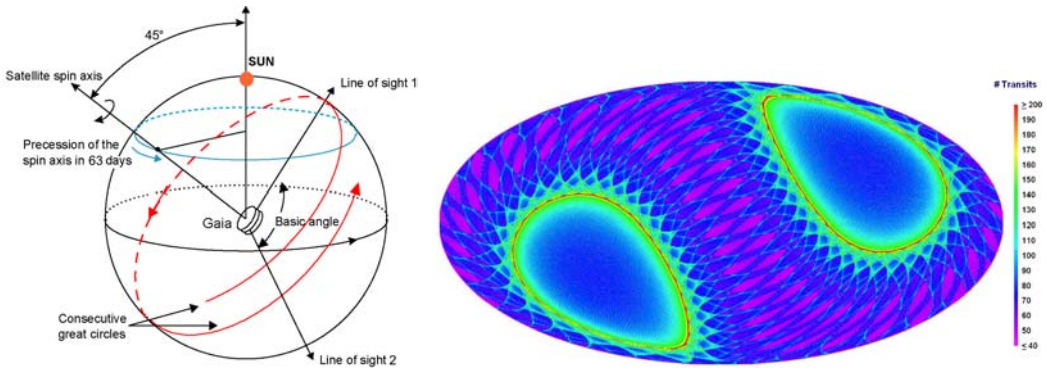
programme should be established, calling for the launching of a second spacecraft, identical or similar to the first one, after a time of some ten years.

26. Cette mission continue à s'appeler Gaia, cependant l'acronyme n'est plus valide.
27. G est la magnitude calculée sans filtres, entre 330 et 1000 nm (Jordi et al, 2010).

plusieurs aspects : origine, structure, formation, évolution, composition. Naturellement, ayant en vue tous les objets qui seront observés par Gaia, depuis des corps du Système Solaire et des planètes extra-solaires jusqu'aux quasars lointains, on peut dire que sa contribution scientifique va bien plus loin que des questions relatives à notre Galaxie. La quantité et la qualité de ses observations ainsi que l'étendue de son catalogue final, auront des implications très profondes et extrêmement importantes sur beaucoup d'aspects de l'étude de l'univers : depuis l'étude de la formation du Système Solaire, en passant par la recherche de signes de vie hors du Système Solaire, jusqu'à des implications cosmologiques issues du calibrage de l'échelle de distance et de mesures précises de constantes physiques fondamentales.

1.5.1 L'instrument

Tel que son prédécesseur Hipparcos, le satellite Gaia utilise le principe d'astrométrie globale à partir de la mesure d'angles entre les objets, méthode déjà testée et validée sur Hipparcos . Comme nous l'avons vu précédemment, ce principe est basé sur l'observation du ciel en son entier, simultanément à partir de deux champs de vision (séparées de 106.5°) d'un satellite qui tourne lentement autour d'un axe (perpendiculaire aux champs de vision) et qui possède un mouvement de précession, de manière à permettre une couverture complète du ciel. Dans le cas de Gaia, un grande cercle sera décrit dans le ciel toutes les 6 heures, et les objets qui passeront dans l'un des champs seront observés dans le second champ 106.5 minutes après l'observation dans le premier champ. La période de précession est de 63.12 jours.

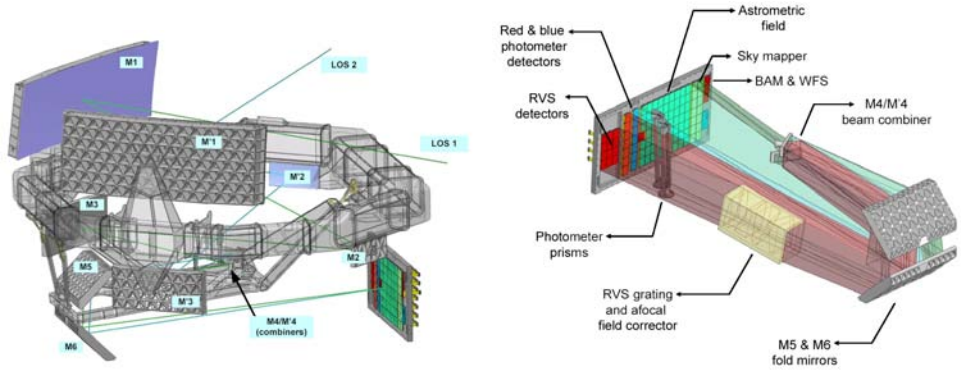


(a) (b)

FIGURE 1.7 – (a) Diagramme de la méthode d'observations de Gaia (De DPAC, 2007). (b) Carte représentant le nombre d'observations dans chaque région du ciel en coordonnées galactiques. (De Mignard, 2010)

Le dessin optique de l'instrument consiste en deux télescopes identiques, de format rectangulaire de 1.45m x 0.5m et une distance focale de 35m (voir Figure 1.8). Ces deux instruments partagent un même plan focal, composé par une mosaïque de 106 CCDs de 4500x1966 pixels (environ 4.7cm x 6cm chacun voir Figure 1.9), les pixels font 59x177 mas. Ce détecteur, le plus grand jamais lancé dans l'espace,

est divisé en régions de CCDs pour la cartographie du ciel (*Sky-Mapper*, ou SM), pour le champ astrométrique (*Astro-Field*, ou AF), pour le champ photométrique, pour le champ spectroscopique et pour le système de métrologie interne (plus détails sur la Figure 1.9). Le mode de fonctionnement de l'intégration et de la lecture de ce système est similaire à celui rencontré dans des CCDs installés sur des cercles méridiens, dénommé *Time Delay Integration* ou *drift-scan*.



(a) (b)

FIGURE 1.8 – (a) Schéma du payload scientifique du satellite Gaia montrant les miroirs primaires et les différents miroirs auxiliaires. (b) Schéma du plan focal avec la position des prismes de dispersion BP et RP en plus de l'instrument spectroscopique de vitesse radiale (RVS). (De EADS Astrium)

En fonction de la taille du plan focal ($\sim$ 1 gigapixel) et des limitations de bande de communication existantes au moment du développement du satellite, il serait complètement impossible de transférer vers la Terre le contenu de tous ces CCDs. Pour cela, il existe à bord un système de traitement et de sélection des données qui seront intégrées et transférées, une tâche qui est réalisée par la *Video Processing Unit* au moyen d'algorithmes dédiés (voir section 1.5.1.1).

La sélection de ces données est initiée au moment où des sources sont détectées dans les *Sky Mapper*, ainsi au cas où la détection serait confirmée dans la première colonne de CCDs du champ astrométrique, ces objets sont intégrés par les autres CCDs, qui ne seront lus qu'en « fenêtres »autour des objets. Ces fenêtres sont formées par des ensembles de pixels autour du pixel central de l'objet, étant entendu que le nombre de pixels lus et transférés est dépendent aussi bien de la colonne de CCDs où l'observation est réalisée que de la magnitude de l'objet observé (plus de détails sur le Tableau 2.1, dans le Chapitre 2).

À chaque colonne du plan focal astrométrique une nouvelle mesure complètement indépendante de l'objet observé est réalisée, de manière à ce qu'à chaque passage de Gaia, c'est-à-dire, toutes les fois que le satellite observe une certaine région du ciel, neuf observations astrométriques distinctes sont réalisées des mêmes objets.

Les données photométriques sont obtenues en utilisant deux prismes qui dispersent la lumière des objets pendant que ceux-ci transitent devant les CCDs photométriques.

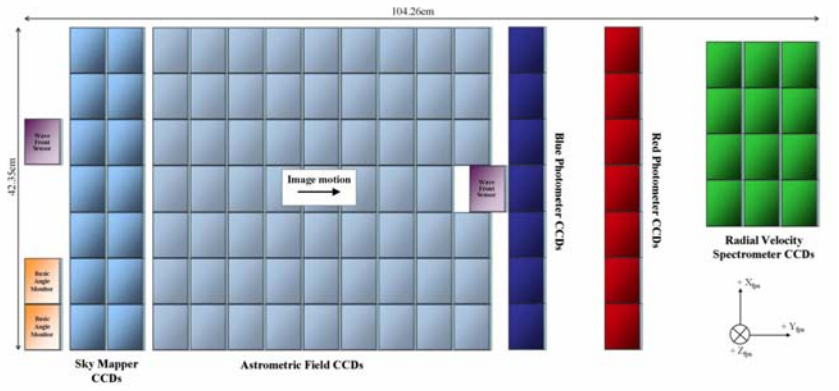
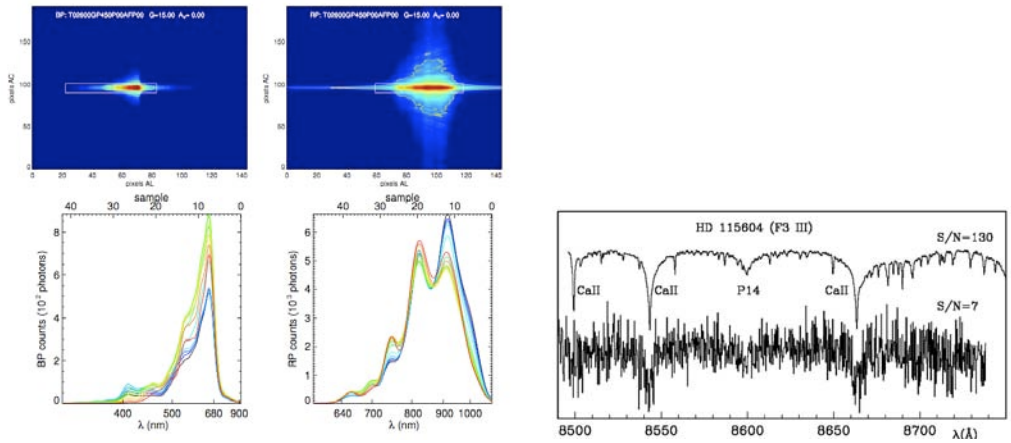


FIGURE 1.9 – Schéma du plan focal du satellite montrant les 106 CCDs qui forment un ensemble de presque 1 gigapixels. (De Alexander Short – ESA)

Ces prismes sont dénommés RP (*Red Photometer*) et BP (*Blue Photometer*), étant donné que l'un disperse la partie rouge du spectre (de 640–1050 nm) tandis que l'autre la bleue (de 330–680 nm). Ensemble, BP et RP couvrent toute la bande spectrale du visible en s'approchant même de l'infrarouge proche.

Le dernier instrument est un spectrographe dédié à la mesure de vitesse radiale, dénommé RVS ou *Radial Velocity Spectrometer*. Cet instrument possède une résolution de R 11.500, et couvre la région autour du triplet de Calcium II (entre 847 et 874 nm). Bien qu'il soit optimisé pour la détermination de la vitesse radiale des objets, il fournira des informations sur les vitesses de rotation, les paramètres atmosphériques, les abondances d'éléments et les traceurs du rougissement interstellaire.



(a) (b)

FIGURE 1.10 – Observations simulées (a) spectrophotométriques pour une étoile M6V dans les CCDs et intégrées dans la direction perpendiculaire à la dispersion pour un intervalle de valeurs de log(g) entre -1.0 et +5.0 et (b) spectroscopiques d'une étoile type F avec $G \sim 16$ pour une observation unique et pour la combinaison de toutes les observations à la fin de la mission. (De DPAC, 2007)

Le principe du RVS est similaire aux prismes BP/RP étant entendu qu'un réseau de diffraction disperse la lumière des objets qui passent dans le plan focal sans utilisation de fibres ou de fissures. Un exemple d'observation d'une étoile de magnitude $G \sim 15$ de type M6V par le BP et le RP, et de mesures spectroscopiques du RVS pour une étoile du type F avec $G \sim 16$ pour une unique observation et pour le spectre combiné à la fin de la mission, peuvent être vus sur la Figure 1.10.

1.5.1.1 La VPU – *Video Processing Unit*

Comme nous l'avons déjà dit, les données du plan focal complet ne peuvent pas être transférées vers la Terre pour une question de largeur de bande. Donc, déjà à bord du satellite, un traitement préliminaire des données est effectué, pour déterminer quels pixels seront transmis. C'est la tâche de la VPU, ou *Video Processing Unit*, qui est la centrale de traitement on-board des observations de Gaia. De plus, la VPU sert aussi pour alimenter le sous-système de contrôle d'attitude du satellite.

Pratiquement toutes les données scientifiques de Gaia passent à un moment ou un à autre par ce sous-système, et du point de vue de cette thèse, la VPU est un système essentiel, car c'est elle qui définit quels objets seront observés et comment les données de ces objets seront binnées avant leur envoi vers la Terre.

Les algorithmes de détection sont tels qu'ils inhibent le transfert des données de tout objet d'une taille angulaire supérieure à ~ 700 mas,[28] comme on peut le vérifier à partir de simulations de l'efficacité de la détection d'objets de différentes tailles présentée sur la Figure 1.11.

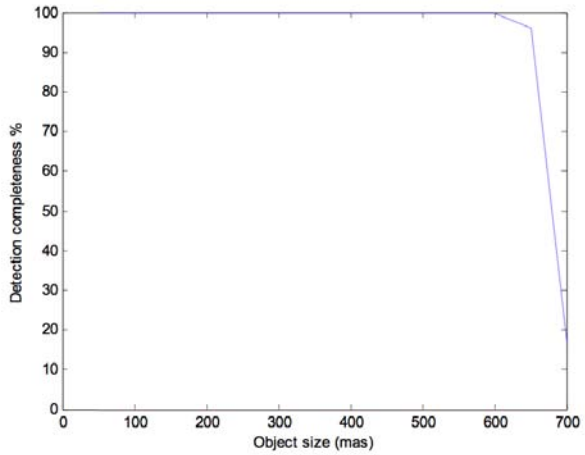


FIGURE 1.11 – Efficacité de la détection en fonction de la taille angulaire apparente de l'objet observé. (De Astrium, 2008)

La détection ou non d'un objet est réalisée après que l'objet soit passé par les colonnes des CCDs du *Sky Mapper*. Elle est basée sur le contraste entre la valeur d'un pixel et des pixels autour de ce dernier, s'agissant d'un processus en trois étapes. Le premier pas est le calcul d'une moyenne des valeurs des pixels qui forment le carré

28. Notez que c'est la taille que le système de détection attribue à l'objet, c'est-à-dire, le taille de la région la plus lumineuse de l'objet.

5x5 autour du pixel qui est analysé et la soustraction de cette valeur aux pixels dans le carré 3x3. Le second pas est le calcul de quelques attributs, comme par exemple les flux dans les directions parallèle et perpendiculaire au balayage, le flux total, si le pixel central est un maximum local, et quelques stratégies de rejet basées sur les flux. Le dernier pas est l'application de coupes principalement basées sur un flux limite (déterminé arbitrairement) et dans le fait que le pixel central est ou n'est pas un maximum dans les deux directions. Au cas où le pixel passerait par tous ces critères, une fenêtre de taille adéquate est créée, confirmée dans l'AF1 (la première colonne de CCDs de l'AF) et après avoir terminé l'intégration sur tout le plan focal, transférée vers la Terre. Plus de détails sur le fonctionnement de ce système peuvent être trouvés dans Astrium (2010).

1.5.2 Le DPAC – *Data Processing and Analysis Consortium*

Les données transférées par le satellite sont d'une grande complexité, et pour préparer le système de traitement des données Gaia, la communauté scientifique, principalement européenne, est organisée depuis 2006 en un consortium appelé Data Processing and Analysis Consortium, ou DPAC, dont la description complète peut être trouvée dans DPAC (2007).[29]

Ce consortium comprend actuellement plus de 400 chercheurs, de 23 nationalités différentes plus l'ESA (voir Figure 1.12), et a la responsabilité, envers l'Agence Spatiale Européenne, de préparer tout le système de traitement scientifique des données du satellite dans le but de transformer les données brutes qui arrivent sur Terre en données qui peuvent être utilisées scientifiquement.

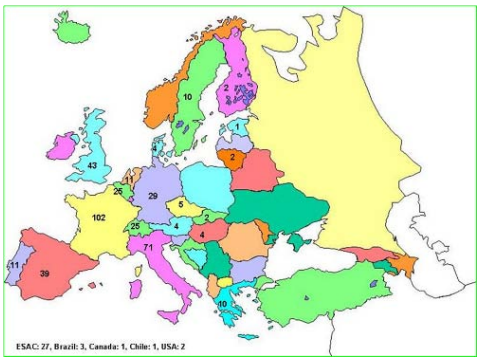


FIGURE 1.12 – Représentation du nombre de chercheurs membres du DPAC – situation en janvier 2010. Des informations actualisées peuvent être trouvées sur : `http://www.rssd.esa.int/index.php?project=GAIA&page=DPAC_Membership`.

La tâche comprend non seulement le traitement d'un grand nombre d'informations, mais principalement d'information d'une grande complexité. L'exemple le plus clair est le fait que Gaia est un système « auto-calibrable » : les positions mesurées des

29. Avant le DPAC, un comité appelé DACC (*Data Analysis Coordination Committee*) avait la tâche de créer le DPAC à partir de structures plus informelles préalablement existantes depuis 2001. Le DACC a été dissous après la formation du DPAC.

passages des étoiles sont utilisées à la fois pour déterminer l'attitude du satellite et les positions des étoiles dans le ciel.

Le DPAC est divisé en neuf groupes (Coordination Units – CU) qui traitent tous les aspects de la mission (par exemple CU7–*Variability Processing* est dédié à l'analyse de variabilité des objets). Ces Coordination Units sont elles-mêmes organisées en Development Units (DU), qui sont les groupes responsables des tâches scientifiques plus spécifiques du projet (par exemple, CU2–*Simulations* DU4–*Instrument Model* est responsable du modèle de simulation de l'instrument). Enfin, ces DUs sont subdivisés en WorkPackages, qui sont les sous-groupes responsables pour les spécialisations des tâches de ces DUs. Dans les sessions 1.5.5.2 et 1.5.5.1, nous verrons plus en détails les CU4–*Object Processing* et CU5–*Photometric Processing*, dans lesquels ce travail de thèse est inséré. Un diagramme avec les CUs peut être vu sur la Figure 1.13.

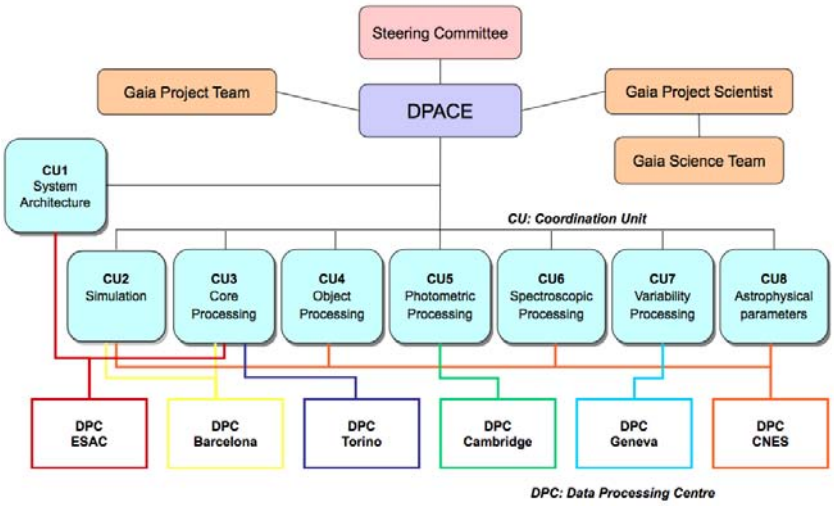


FIGURE 1.13 – Structure du DPAC. (De Mignard, 2010)

Les estimations présentées dans Mignard (2010) indiquent que seront produites par le satellite $\sim$ 250 Tb de télémétrie compressée, étant entendu que les données traitées doivent représenter entre $\sim$ 0.5 PB et $\sim$ 1 PB. Pour traiter ce volume d'informations, selon des estimations approximatives, plus de 1.5×10^{21} FLOPs seront nécessaires. Ceci est énorme mais parfaitement réalisable.[30]

Les données sont traitées de façon non centralisée pour que la réduction puisse être plus efficace. Pendant que le satellite observe le ciel, les données sont envoyées au centre d'opérations de la mission (MOC) et à l'ESAC (*ESA Center*) où elles sont stockées dans la base de données de Gaia, la MDB (*Main Data Base*). Elles sont alors distribuées dans les divers centres de traitement du DPAC où elles sont

30. Pour résoudre ce problème sur un délai d'un an, un ordinateur dédié de $\sim$ 30 TFLOP/s serait nécessaire, ce qui actuellement équivaut à deux ou trois racks d'un Cray XT6, d'un coût de $\sim$ 1.5 millions de dollars. Selon Mignard (2010), deux centres de traitement de données posséderont une infrastructure consacrée à Gaia de 10–20 TFLOP/s chacun à partir de 2012, et en temps partiel le Marenostrum de Barcelone qui est capable d'atteindre $\sim$95 TFLOP/s est déjà disponible.

transformées en paramètres astronomiques pour toutes les sources observées dans des cycles périodiques de 6 mois. Une fois traités, les résultats sont renvoyés à l'ESAC, où ils sont stockés dans la MDB. Durant le traitement des données, six centres de traitement (appelés DPC–*Data Processing Center*) seront impliqués (voir Figure 1.13). Les travaux développés durant cette thèse sont (dans le cas des simulations) ou seront (dans le cas des traitement de données) exécutés dans les centres du CNES[31] (Chap. 4, 5, 6), Cambridge (Chap. 3) et Barcelone (Chap. 4).

1.5.3 Simulateurs

Dans le but de donner de la cohérence au DPAC, ainsi que pour la préparation des logiciels qui seront utilisés dans l'analyse des données Gaia, les simulations Gaia sont centralisées les simulateurs sont accessibles à tous les membres de ce consortium. Les codes de simulation utilisent des routines communes englobées dans un ensemble dénommé GaiaSimu (Babusiaux et al, 2010), étant entendu qu'il existe trois codes principaux : GASS, GIBIS et GOG. Les fonctionnalités de ces codes seront décrites dans les prochaines sections, néanmoins, tous ces codes ont besoin d'une simulation de base du ciel, qui est fournie par le « Modèle d'Univers »Gaia.

1.5.3.1 Modèle d'Univers

Le « Modèle d'Univers »est un code de simulation du ciel : en fonction de la région de la sphère céleste considérée, ce code prédit quels objets astronomiques devraient y être observés. Ceci signifie que dans ce modèle, il existe des descriptions de comment est le Système Solaire (planètes et leurs satellites, astéroïdes, comètes), la Voie Lactée (étoiles, amas, étoiles binaires, radiation diffuse, nébuleuses), de comment les galaxies proches sont résolues en étoiles et de quelles sont les caractéristiques des galaxies non résolues mais étendues, en plus des quasars plus lointains et des distributions spatiales de tous ces objets.

Le modèle de galaxie adopté s'appelle modèle de Besançon.[32] Pour l'utilisation dans le simulateur, ce modèle a été adapté au langage Java, et modifié pour permettre la simulation d'objets variables aussi bien du point de vue photométrique que du point de vue astrométrique, en plus de l'inclusion d'objets multiples (et amas), radiation diffuse et nébuleuses.[33]

Des galaxies proches sont introduites en tant qu'étoiles individuelles, particulièrement les nuages de Magellan sont actuellement mis en place dans les simulateurs. Des galaxies plus distantes, qui ne sont pas résolues en objets individuels sont introduites dans le simulateur au moyen d'une méthode développée par E. Bertin appelée Stuff, qui a vu sa première mise en œuvre pour les paramètres de Gaia faite par C. Dollet

31. Le Centre National d'Études Spatiales.
32. Au début de la mission, un deuxième modèle était aussi utilisé, le modèle de Barcelone.
33. En fait, ceci n'est pas fait à partir d'une modification du modèle de Besançon, mais au moyen de l'inclusion de codes de simulation pour ces objets qui fonctionnent de façon plus ou moins indépendante du modèle de Besançon.

et dont le dessin final et la mise en œuvre en Java (JStuff), qui est utilisée dans les simulations de la mission, a été réalisée par nous, et sera décrite dans le Chapitre 4.

Des images des galaxies peuvent être produites de deux manières : soit en utilisant le Skymaker (autre code d'E. Bertin, dont l'implémentation actuelle est due à C. DelleLuche) qui forme des images à partir de la somme des distributions de brillance de deux composantes analytiques soit à partir de l'utilisation du code Magil, qui utilise des images réelles de galaxies proches comme prototypes pour produire des images plus réalistes (ce code, mis en place par Gavras et al, 2010, sera décrit avec plus de détails dans le Chapitre 4). Les supernovas extragalactiques sont aussi incluses, néanmoins de façon aléatoire. Selon Robin et al (2010), dans des versions futures du simulateur (post-2010) elles doivent être générées autour de galaxies avec des probabilités de survenue dépendantes du type de Hubble de leur galaxie hôte.

Les quasars sont introduits à partir d'une liste d'objets, décrite dans (Slezak & Mignard, 2007), qui est produite par un modèle statistique avec des propriétés similaires au SDSS extrapolées jusqu'à G=20.5. Ces listes sont complétées par le catalogue de quasars proches de Véron-Cetty & Véron (2006).

Plus de détails sur le modèle d'univers adopté peuvent être trouvés dans le document Robin et al (2010), qui est actualisé périodiquement.

1.5.3.2 GASS

GASS, ou *Gaia System Simulator*, est responsable de la simulation de la télémétrie du satellite et produit de grands volumes de données. Il est capable de reproduire sur ordinateur une mission complète dans toutes ses phases. Les simulations de GASS sont aussi réalistes que possible, étant statistiquement représentatives des observations du satellite. Une description actualisée de ce simulateur peut être rencontrée dans Gallardo & Masana (2010).

Les simulations générées par GASS sont utilisées pour des tests des chaînes de réduction et de traitement à tous les niveaux de la mission, elles seront utilisées pour la validation de tout le flux de données avant que la mission réelle ne soit initiée, elles sont utilisées pour tester et améliorer le Main Data Base et ont été utilisées pour des estimations du nombre d'objets qui seront observés et du nombre de données transmises vers le sol.

Le processus de base est divisé en étapes d'initialisation des générateurs de sources et de la simulation de la mission (mission complète, partielle, limite de magnitude, non linéarité, bruit dans l'attitude du satellite, etc.), de la simulation des sources astronomiques, de la simulation d'observations du satellite (reproduisant statistiquement les processus de détection, sélection, temps et position de passage, taille de la fenêtre d'observation, etc.), de la génération des paquets de télémétrie (mis en format exactement comme le satellite réel devra le faire, et augmentés d'erreurs de centrage, inefficacité dans le transfert de charges, etc.) et de l'archivage des résultats.

Cette thèse n'a pas fait un usage direct de ce simulateur, néanmoins GASS fait usage d'un code que nous avons mis en place durant le développement de cette thèse, le JStuff qui sera décrit dans le Chapitre 4.

1.5.3.3 GIBIS

GIBIS, ou *Gaia Instrument and Basic Image Simulator* (Babusiaux, 2006) est le simulateur des données de la mission Gaia au niveau du pixels des CCDs. De manière distincte de GASS, GIBIS simule les observations individuelles de Gaia de chaque objet de manière aussi réaliste que possible, étant entendu que des modèles extrêmement réalistes des objets et des instruments sont utilisés. Comme on peut imaginer, avec l'utilisation de ce simulateur la missions entière ne peut être simulée pour le ciel entier car le temps de calcul est prohibitif.

Néanmoins, toutes les observations d'un certain objet qui seront réalisées durant la mission complète peuvent être simulées, fournissant de cette manière des informations détaillées sur comment cet objet particulier sera observé. Ceci est important pour les études d'objets étendus, de multiples objets sur la même ligne de visée, pour la réduction photométrique de champs denses, pour le calibrage de l'instrument, pour le tests des algorithmes de détermination de fond de ciel, etc. (voir Babusiaux, 2006).

GIBIS peut être utilisé de deux manières : au moyen d'une interface web, mise à disposition par le CNES et qui peut être accédée à la carte, ou au moyen de demandes officielles à la CU2, quand des données qui normalement ne sont pas enregistrées par l'interface web sont nécessaires et qui sont utilisées par une petite fraction du DPAC.

Pour réaliser les simulations, après que l'utilisateur ait défini les paramètres du modèle d'univers, de l'instrument et de la mission via l'interface web, GIBIS simule la région du ciel demandée en faisant usage du Modèle d'Univers.

Donc, pour chacun des passages de Gaia dans cette direction du ciel, la réponse du plan focal de Gaia est simulée à partir de modèles détaillés de l'instrument (étant entendu que même la PSF peut être calculée à chaque observation), transformant la liste d'objets générée par le Modèle d'Univers dans des valeurs de pixels dans tous les CCDs du plan focal (exemple sur la Figure 1.14). Finalement, GIBIS exécute les algorithmes de détection et de création de fenêtres qui seront utilisées à bord du satellite, résultant dans les observations des objets à chaque passage. Le processus est donc répété pour tous les passages demandés, ou pour toute la mission.

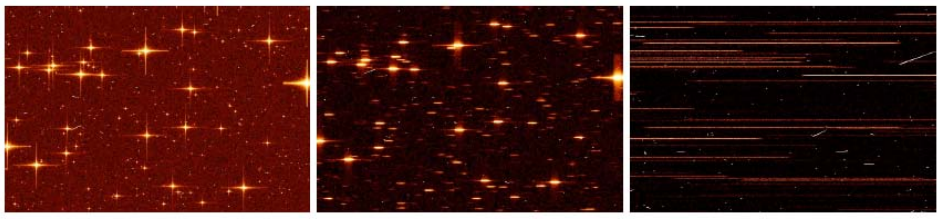


FIGURE 1.14 – Simulation d'un champ stellaire produit avec GIBIS. (De DPAC, 2007)

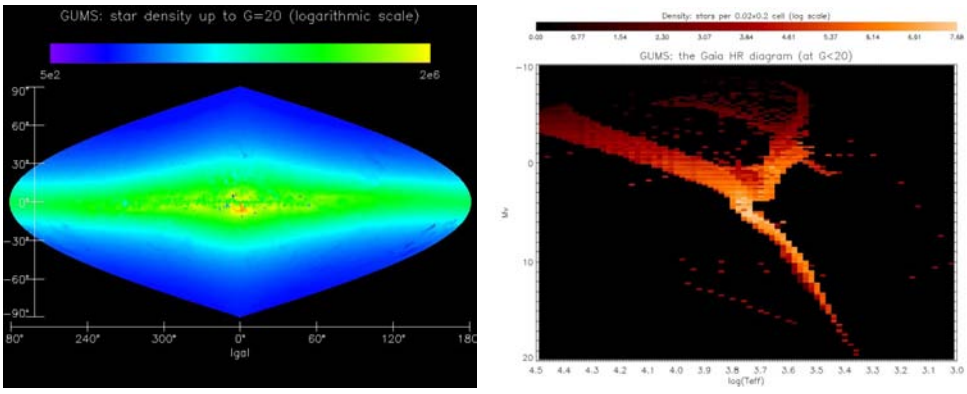
Celui-ci a été l'unique simulateur utilisé pour le développement de ce travail de thèse. De plus, GIBIS utilise les codes JStuff et Magil, qui seront décrits dans le Chapitre 4.

1.5.3.4 GOG

GOG, ou Gaia Object Generator, est un simulateur du catalogue Gaia. Pour cela il appelle le modèle d'Univers de Gaia dans la région désirée, il convolue les paramètres des objets simulés avec les fonctions d'erreurs nominales de la mission et stocke l'ensemble de ces données dans la base dans des tables similaires à la MDB de Gaia.

Nous avons réalisé le prototype de ce simulateur sous la direction de Xavier Luri à l'*Universitat de Barcelone*, et bien qu'il était relativement rudimentaire, il a été utilisé en 2006 pour réaliser la première simulation à large échelle de la mission Gaia complète, appelée GUMS (*Gaia Universe Model Snapshot*). GUMS a été produit sur le super-ordinateur Marenostrum à Barcelone (X. Luri, *pers. comm.*). Cette simulation permit de tester la distribution de densités d'objets sur tout le ciel, le diagramme Hertzprung-Russel, le diagramme magnitude-parallaxe, etc.

Le fonctionnement de GOG était le suivant : les paramètres du modèle d'Univers étaient obtenus à partir de fichiers XML[34] (ex. quel modèle de Galaxie utilisé, si des objets doubles et multiples et des variables devaient être inclus, si le modèle extragalactique devait être exécuté, etc.), alors le code instanciait les objets nécessaires dans la mémoire de l'ordinateur,[35] y compris en déterminant les régions HTM[36] correspondantes à la simulation et exécutait les codes de simulation, dont les résultats étaient mis en format et écrits dans un fichier de sortie.



(a) (b)

FIGURE 1.15 – Densité d'objets et diagramme Hertzprung-Russel générés à partir du Gaia Universe Model Snapshot simulé en 2006 avec le prototype de GOG que nous avons développé. Figures créées par X. Luri et C. Babusiaux.

Actuellement, le code qui est utilisé par le DPAC ne possède plus aucun héritage de celui qui a été mis en œuvre en 2006, étant donné que ce simulateur a considérablement

34. XML ou *eXtensible Markup Language* est un langage qui permet de décrire différents types de données de manière auto-contenue.

35. D'une certaine manière, ceci signifie que les codes nécessaires étaient chargés dans la mémoire.

36. Façon des diviser la sphère en régions triangulaires (voir section 4.2.1 pour plus de détails).

évolué, permettant non seulement la simulation de listes d'objets à la fin de la mission obtenues directement du modèle d'Univers, mais aussi de listes d'objets intermédiaires et de l'État de la MBD à n'importe quelle époque entre le début et la fin de la mission.

Ceci peut être fait grâce à des modèles statistiques de l'attitude et des erreurs intermédiaires aussi bien des observations que des procédures de réduction de données, qui sont fournies par les différents CUs participants du DPAC. De cette manière, GOG est capable de générer rapidement des simulations de grands volumes de données. Avec GOG nous avons aussi un ensemble d'outils qui permet de réaliser rapidement des analyses statistiques des résultats obtenus pour les simulations, en plus de pouvoir aussi être exécuté au moyen d'une interface web mise à disposition par le CNES ou être installé par le propre utilisateur. Une description de l'implémentation et des fonctionnalités de ce simulateur peut être rencontrée dans Isasi et al (2010).

Tel que dans le cas de GASS, cette thèse n'a pas fait un usage direct de ce simulateur, néanmoins GOG utilise le JStuff (décrit dans le Chapitre 4).

1.5.4 Réduction des données

La réduction de données Gaia surviendra sur deux niveaux de traitement, un exécuté quotidiennement et un semestriellement. Le traitement quotidien sera fait dès que les données du satellite arriveront sur Terre. Le premier pas est un ajustement d'une PSF/LSF[37] aux données de la fenêtre, et donc à partir de l'attitude approchée du satellite (qui est obtenue à partir de la télémétrie) et d'un catalogue intermédiaire, la fenêtre est associée à un objet donné (ou un nouvel objet est créé). Ce processus fait partie de la chaîne initiale de traitement des données Gaia, appelée IDT ou *Initial Data Treatment*, générant des données qui seront stockées quotidiennement dans la MDB. Durant l'IDT une reconstruction préliminaire de l'attitude du satellite basée sur des étoiles brillantes sera déjà réalisée.

De plus, dans des missions d'astrométrie globale comme celle-ci, pour que des résultats globaux et cohérents soient obtenus, plusieurs mois d'observation sont nécessaires. De cette manière, si quelque chose d'anormal survient, ceci ne serait détectable que bien des mois après l'évènement. Pour éviter ce type de situation, un traitement appelé *Detailed First Look* a été mis en place, et effectuera des tests de qualité scientifique des données 24h après leur réception. Des tests astrométriques, photométriques et spectroscopiques seront réalisés. L'alignement optique entre les télescopes et le centre sont quelques exemples de problèmes qui peuvent être rapidement détectés grâce à ce système. De plus, des alertes scientifiques seront également émises par le *First Look* (comme par exemple, des supernovas), de manière à permettre des réactions rapides de la communauté astronomique.

Donc, à chaque cycle de six mois, le traitement principal sera exécuté. La structure de ce traitement dépend de chaque CU et DU impliqué. Fondamentalement, pour

37. Comme les observations de Gaia sont essentiellement unidimensionnelles, une LSF est définie, ou *Line Spread Function*, qui est l'analogue unidimensionnelle d'une PSF classique. Voir Lindegren (2009).

que ce traitement survienne à chaque cycle, des copies de parties de la MDB seront transférées vers les DPCDBs (*Data Processing Center Data Base*), et à partir de ces copies les CUs/DUs exécutent leurs analyses de manière plus ou moins indépendante. Toutes les interdépendances entre les différentes CUs se fera à travers de la MBD, donc au cas où un CU dépend du résultat d'un autre CU, il sera nécessaire d'attendre un cycle pour un échange d'informations. Un schéma du traitement des données de Gaia peut être vu sur la Figure 1.16.

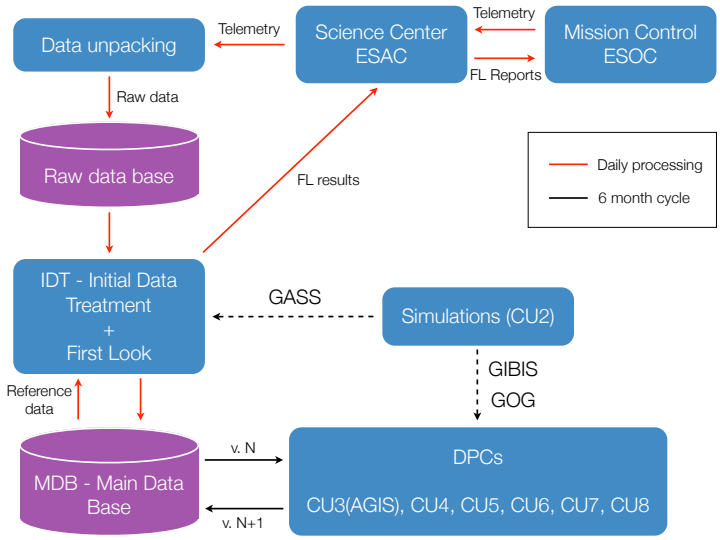


FIGURE 1.16 – Vision conceptuelle du traitement des données de Gaia.

Le traitement astrométrique fonctionnera de façon itérative, étant entendu qu'à chaque nouveau cycle des données seront incorporées à la solution et les paramètres astrométriques des objets, l'attitude du satellite, les calibrages de l'instrument et les corrections relativistes seront re-calculées. Durant le traitement des sources avec de bonnes solutions astrométriques $\sim 5 \times 10^8$ paramètres de sources, $\sim 4 \times 10^7$ paramètres d'attitude et $\sim 10^6$ paramètres de calibrage devront être résolus. Donc, les autres 5×10^9 paramètres astrométriques pour les autres sources seront déterminés.

1.5.5 Réduction des données d'objets problématiques

Dans le cas où il existerait des objets dont le Goodness of Fit obtenu pour l'ajustement de la PSF/LSF durant l'IDT à chaque passage observé serait incompatible avec celle qui est attendue lorsque l'on suppose l'existence d'une seule source ponctuelle dans les fenêtres observées (pour plus de détails, voir Burgon, 2010), l'objet est postérieurement analysé par deux DUs spécialisées du DPAC.

Ces deux DUs font partie des CU4 et CU5, et ont comme objectif d'analyser s'il existe d'autres objets angulairement proches de l'objet observé (qui peuvent se trouver dans la même fenêtre) ou si l'objet possède une émission étendue intrinsèque quelconque (comme par exemple ce sera le cas de galaxies non résolues en étoiles), et dans les deux cas extraire des informations additionnelles de ces objets.

Un schéma du concept du traitement effectué par ces deux DUs peut être vu sur la Figure 1.17. On remarquera que lors du traitement réel les données passeront par la MDB après le traitement qui est effectué dans le CU5, n'arrivant seulement qu'un semestre après dans le CU4.

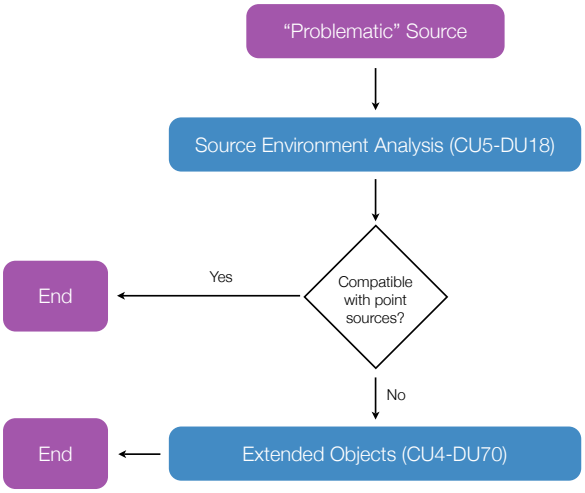


FIGURE 1.17 – Vision conceptuelle du traitement des objets problématiques durant la réduction des données de Gaia.

C'est dans les deux DUs, décrites globalement dans les deux prochaines sections, que la plus grande partie de ce travail de thèse a été réalisée.

1.5.5.1 CU5-DU18 – *Source Environment Analysis*

Les objets stockés dans la MBD qui possèdent une distribution de valeurs de χ^2 distincte de celle espérée pour des objets ponctuels sont analysés par la DU18 du CU5, appelée *Source Environment Analysis*. Dans cette DU, coordonnée par D. Harrison de l'*Institute of Astronomy* de l'*University of Cambridge*, est réalisée une analyse qui cherche à établir si le signal présent dans les observations est compatible ou non avec plus d'un objet ponctuel coexistant dans la même fenêtre. Un schéma de comment ce traitement se fait peut être vu sur la Figure 1.18.

Le premier pas est la réalisation d'une reconstruction d'image à partir des fenêtres observées durant les passages du satellite sur l'objet. Pour réaliser cette reconstruction, plusieurs algorithmes sont à l'étude chez d'autres chercheurs, et seront vus avec plus de détails dans le Chapitre 2. Dans ce chapitre démontrons également qu'il est possible de reconstruire ce type d'images 2D dans toutes les régions du ciel. Cette reconstruction, cependant, demande à être réalisée de la manière la plus rapide possible, en fonction du grand nombre d'objets (estimé dans des dizaines de millions) qui doivent être traités (plus de détails sur l'estimation que nous avons faite de leur nombre est présentée en Chapitre 3).

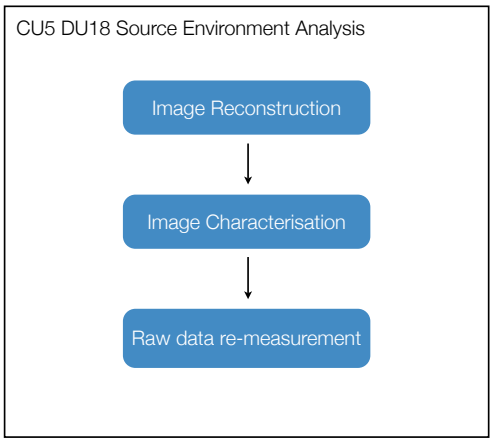


FIGURE 1.18 – Schéma du traitement qui a lieu dans le DU18 du CU5.

L'image reconstruite pour l'objet et son environnement est analysée et caractérisée en termes de nombre de sources existantes et de paramètres approximatifs de ces objets (position x, y, flux et erreurs associées) et d'une évaluation du type de fond de ciel. Les algorithmes qui ont été développés durant cette thèse pour réaliser une telle caractérisation seront détaillés dans le Chapitre 3.

Finalement, les transits individuels sont re-analysées de façon globale en considérant toutes les sources présentes dans l'image comme ponctuelles pour déterminer leur paramètres astrométriques $(\alpha, \delta, \mu_\alpha, \mu_\delta, \varpi)$. Au cas où cette analyse donne un résultat satisfaisant son résultat est stocké dans la MBD, dans le cas contraire, l'objet est marqué dans la MBD comme étant « non compatible avec des sources ponctuelles ». Ce travail est développé par D. Harrison au sein du CU5.

1.5.5.2 CU4-DU470 – *Extended Objects*

Les objets marqués dans la MBD comme étant « non compatibles avec des sources ponctuelles », ainsi que les objets classés par le CU8 comme étant des galaxies non résolues, sont traités par la DU470 du CU4, qui est coordonnée par C. Ducourant, du Laboratoire d'Astrophysique de l'Université de Bordeaux I. Le développement de la chaîne de traitement de ces sources non ponctuelles que nous appelons Objets Etendus (EO) a constitué une large part de mon travail de thèse. Ce traitement consiste à analyser le signal des fenêtres observées à chaque passage, à caractériser morphologique la source et à mesurer de ses paramètres morphologiques.

Comme un plus faible nombre d'objets est espéré (par rapport au CU5-DU18), autour de 10^6 sources, il est possible de consacrer un peu plus de temps au traitement de chaque source.[38] Néanmoins, comme nous le verrons dans le Chapitre 2, des algorithmes de reconstruction d'images complètement libres d'artefacts de reconstruction et qui possèdent un temps d'exécution compatible avec les besoins ne

38. Mais même ainsi, le temps de traitement nécessaire ne doit pas être négligé : si chaque source est analysée en 1 minute, 694 jours seront donc nécessaires pour analyser tous les objets en série.

sont pas encore disponibles. De cette manière, la structure du traitement des données pour ces objets étendus a été construite en prenant en considération depuis le début le fait que nous travaillerions sur des images reconstruites de qualité médiocres. Un schéma fonctionnel de ce traitement est donné en Figure 1.19.[39]

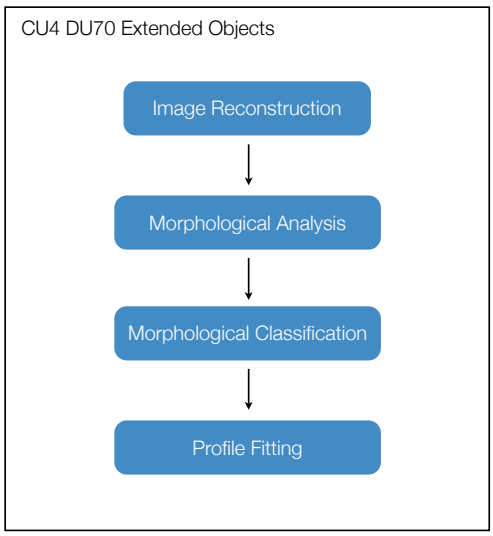


FIGURE 1.19 – Schéma général du traitement effectué dans la DU470 du CU4.

De la même manière que ce qui se passe dans le traitement du CU5-DU18, l'étape initiale ici est la reconstruction d'une image, un processus qui sera commenté en détails dans le Chapitre 2 de cette thèse. Donc, cette image est analysée du point de vue morphologique : des paramètres morphologiques robustes à la dégradation de l'information sont mesurés sur cette image. Ces paramètres concernent la répartition de la lumière dans l'image, et seront commentés en détails dans le Chapitre 5.

Les paramètres mesurés sont utilisés pour classer l'image en classes d'objets. Actuellement, la classification est implémentée pour différencier des types de galaxies (elliptiques, et spirales et une troisième classe pour des cas particuliers, irréguliers, etc.) qui sont les cas d'objets étendus intrinsèques le plus souvent observés. Dans le Chapitre 5, nous faisons quelques commentaires sur cette classification.

La dernière étape du traitement est un ajustement de profils de luminosité aux données unidimensionnelles (et non à l'image reconstruite). Pour cela, le profil théorique et ses paramètres sont initialement choisis sur la base de la classification obtenue dans l'étape précédente. Le processus d'ajustement est mis en œuvre d'une façon itérative, car au cas où le profil correspondant à la classification ne s'ajuste pas convenablement, d'autres profils sont essayés, et au cas où certains de ces autres profils s'ajustent mieux, la classification de l'objet est modifiée.[40] Une description de comment l'ajustement de chaque profil est réalisé sera présentée dans le Chapitre 6.

39. Cette représentation est schématique, car pour que le système puisse être exécuté au CNES les modules ne communiquent pas entre eux, et les données circulent via des banques de données, tandis que le traitement est contrôlé par un logiciel de gestion dénommé SAGA.

40. Cette étape est importante car, comme nous l'avons déjà mentionné, la qualité des images

1.6 Contenu de la thèse

Le corps de cette thèse comprends cinq autres chapitres, qui chacun présentent un bref résumé du sujet traité, plusieurs sections détaillent le travail effectué, et une conclusion est présentée à la fin.

Dans le Chapitre 2, nous présentons le principe de reconstruction d'images à partir de projections unidimensionnelles de données, nous discutons les algorithmes proposés dans la littérature pour leur utilisation avec les données Gaia et nous présentons quelques études que nous avons réalisées sur la couverture spatiale et angulaire de ces images reconstruites.

Dans le Chapitre 3 nous présentons une étude que nous avons réalisée sur la quantité de sources dans le ciel qui peuvent présenter des solutions astrométriques dégradées à cause des projections d'autres sources sur la même ligne de visée ; nous présentons une méthode que nous avons développée pour analyser les images reconstruites de ces sources problématiques et récupérer les signaux des diverses sources présentes sur l'image (y compris celles qui peuvent être au-delà de la magnitude limite du satellite) et nous montrons les résultats de l'analyse de simulations par cette méthode dans divers régimes de reconstruction d'images.

Dans le Chapitre 4, nous avons commencé des études sur un autre type de source astronomique qui présentera des solutions astrométriques perturbées mais, cette fois-ci, pour des raisons intrinsèques à sa structure : les galaxies non résolues en étoiles. Nous présentons un code de simulation de catalogues de galaxies appelé JStuff, que nous avons implémenté. Nous présentons également le code MAGIL responsable de l'introduction d'images de ces objets dans le simulateur GIBIS et pour lequel nous avons collaboré au design et à la spécification. Nous montrons aussi des résultats de simulations officielles du consortium d'analyse de données Gaia (DPAC) pour le ciel en son entier qui ont utilisé JStuff, et à partir de cette simulation nous présentons des commentaires sur le nombre de galaxies qui seront observées par Gaia.

Dans le Chapitre 5 nous présentons la méthode d'analyse des images 2D reconstruites qui mesure les paramètres de concentration de lumière, de smoothness et des moments de la distribution du flux et des asymétries produisant un espace 5-dimensionnel appelé CASGM20. Nous montrons une étude que nous avons réalisée sur comment ces paramètres varient avec la quantité d'informations disponibles pour leur estimation et comment ils varient selon le type de galaxie analysée. Nous présentons la méthode que nous avons utilisée pour classer les galaxies, appelée Support Vector Machines et nous présentons des tests empiriques que nous avons réalisés pour vérifier si le taux maximum de classification théorique est atteint dans les différents cas d'échantillonnage aléatoire de distributions gaussiennes. Enfin, nous présentons des résultats de l'application de cette méthode quand celle-ci est utilisée dans l'espace CASGM20 calculé à partir d'images de galaxies proches d'un catalogue d'images de la littérature, de galaxies reconstruites à partir de simulations de la mission Gaia, et de galaxies observées dans le *Hubble Deep Field*.

reconstruites avec les algorithmes disponibles n'est pas optimale et conduire à conduire à une classification erronée.

Dans le Chapitre 6, nous montrons la méthode que nous avons développée pour réaliser une analyse morphologique du profil de brillance des galaxies non résolues en étoiles basée sur la re-analyse des observations unidimensionnelles originales du satellite. Dans nos simulations de ces observations, nous prenons en compte une distribution réalistes des angles de passage du satellite. Nous présentons aussi des résultats obtenus à partir de l'application de cette méthode sur un ensemble d'observations simulées, et nous faisons quelques commentaires sur les possibilités de perfectionnement et d'application de la méthode sur des données autres que celles qui seront obtenues par Gaia.

Finalement, nous présentons les conclusions sur ce travail de thèse.

Pour finir ce chapitre, il me semble important de rappeler que la possibilité d'utiliser les données de la mission spatiale Gaia pour reconstruire des images et de les analyser augmente les capacités de cette remarquable mission, permettant d'entrer dans des domaines virtuellement inexplorés. Et, peut-être la plus grande découverte qui pourra être faite au moyen de ces images reconstruites est quelque chose que nous ne pouvons pas encore imaginer – mais comme dit Fabian (2009) citant Louis Pasteur, la chance favorise l'esprit préparé.

Il s'agit ici de la première occasion que l'humanité aura de voir tout le ciel jusqu'à une magnitude $G \sim 20$ dans des résolutions que seules des observations faites dans l'espace ou avec des télescopes de haute technologie sur Terre munis de systèmes d'optique adaptative, permettraient. Il n'existe pas encore de propositions pour l'utilisation de cette dernière technique pour des relevés de ciel complet. Cependant un *survey* sera réalisé à partir de l'espace dans quelques années, nous permettant de *naviguer sur des mers où l'on a jamais navigué*[41] – et il s'appelle Gaia.

41. *Canto I* de *Os Lusíadas*, de Luís Vaz de Camões

Chapitre 2

Reconstruction d'images

"*Le microphone n'est rien d'autre qu'une masse métallique. Il ne capte qu'à peine 30 pour cent du son original.*" S. Celibidache[1]

Sommaire

Nous avons vu dans le Chapitre 1 que Gaia ne transmettra pas vers la Terre des images à deux dimensions des sources détectées. Néanmoins, ces images 2D seront nécessaires dans certaines situations comme dans l'analyse des sources secondaires plus faibles que la magnitude limite de l'instrument, ou dans le cas d'objets étendus.

Des algorithmes qui permettront une reconstruction approchée de ce signal à partir des observations du satellite ont été proposés dans divers travaux. Dans ce Chapitre nous nous pencherons sur la problématique de la reconstruction, nous présenterons les algorithmes actuellement disponibles ainsi qu'une étude des couvertures spatiale et angulaire des images reconstruites sur la sphère céleste (Krone-Martins et al, in prep.b).

1. *in* Folha de São Paulo, 5 octobre 1993.

2.1 Introduction

En raison de la taille du plan focal du satellite Gaia et des limitations de bande de communication existantes pour le transfert de données entre le point L2 et la Terre, le transfert du contenu complet de l'information du plan focal de Gaia serait impraticable. De cette manière, il a été mis en place un système on-board qui sélectionne les données qui seront transférées (Astrium, 2006).

Durant le passage de l'objet dans le plan focal de Gaia, son signal croise diverses colonnes de CCDs (voir Figure 1.9, Chapitre 1). Les données relatives à l'objet seront transmises vers la Terre seulement dans le cas où l'objet serait détecté dans l'une quelconque des deux colonnes des *SkyMapper* et dans le cas où cette détection serait ensuite confirmée durant le passage par l'*AstroField 1*.

Ces données transmises consistent en surfaces dénommées « fenêtres » (*windows*), qui sont des régions autour de l'objet composées par un ensemble de *samples*. Ces *samples*, de leur côté, équivalent à la somme des signaux provenant des pixels des CCDs. Dans le langage astronomique courant un *sample* est un pixel binné. Une fenêtre distincte est générée toutes les fois que l'objet croise une colonne de CCDs ; de cette façon, ne considérant que les colonnes *SM* et *AF*, à chaque balayage du satellite un maximum de 10 fenêtres par objet sera transmis.[2]

Le taille exacte des fenêtres transmises dépend de la magnitude de l'objet et dans quelle colonne de CCDs elle a été générée. Le Tableau 2.1 explicite les tailles lues dans le CCD (en pixels physiques) et les tailles transmises vers la Terre (en *samples*) en fonction de la magnitude Gaia G.

Colonne de CCD	Magnitude G	Fenêtre lue (pixels)	Binning (pixels)	Fenêtre transmise (*samples*)	Taille angulaire (")
SM	$G < 13.0$	80×12	2×2	40×6	4.72×2.12
	$G > 13.0$	80×12	4×4	20×3	4.72×2.12
AF 1	$G < 13.0$	18×12	1×2	18×6	1.06×2.12
	$13.0 < G < 16.0$	12×12	1×12	12×1	0.71×2.12
	$G > 16.0$	6×12	1×12	6×1	0.35×2.12
AF 2, 5, 8	$G < 13.0$	18×12	1×1	18×12	1.06×2.12
	$13.0 < G < 16.0$	18×12	1×12	18×1	1.06×2.12
	$G > 16.0$	12×12	1×12	12×1	0.71×2.12
AF 3, 4, 6, 7, 9	$G < 13.0$	18×12	1×1	18×12	1.06×2.12
	$13.0 < G < 16.0$	12×12	1×12	12×1	0.71×2.12
	$G > 16.0$	6×12	1×12	6×1	0.35×2.12

TABLE 2.1 – Parametrès des fenêtres suivant la colonne de CCD et la magnitude estimée de l'objet. Les colonnes « Binning » et « Fenêtre transmise » ont été obtenues à partir de *Gaia Parameter Database* (GPDB_2010-03-02, 2010).

2. Ceci parce que, en plus des neuf colonnes *AF*, l'objet ne sera détecté que dans le *SM1* ou dans le *SM2*, vu qu'il ne peut être observé que par l'un des télescopes.

Pour permettre la détection d'objets mobiles et corriger tout problème de suivi et d'erreur de centrage, la fenêtre qui est lue à bord, possède dans certains cas, un binning différent, pouvant même avoir des tailles physiques distinctes (avec plus de pixels que ceux qui seront « binnés » et transmis). Des procédures de re-centrage et de re-échantillonnage sont utilisées par le système de traitement du satellite pour que seuls les samples significatifs soient envoyés vers la Terre.

Dans les colonnes AF 3, 4, 6, 7 et 9, par exemple, un objet avec G $>$ 16 aura une fenêtre lue à bord formée par 12×12 pixels, binés en 1×12 pixels, formant les 12×1 *samples* qui seront analysés par le sous-système de re-centrage. Ce sous-système détermine les 6 colonnes de pixels les plus importantes pour la définition de l'objet, lesquelles après avoir été sélectionnées sont transmises vers la Terre sous la forme d'une fenêtre de 6×1 *samples*, équivalents aux 6×12 pixels non-binés presentés dans le Tableau 2.1.

Donc, le problème de reconstruction d'image consiste dans la création d'une « image bi-dimensionnelle » de l'objet observé, à partir des données de ces fenêtres qui sont essentiellement unidimensionnelles – allié, autant que possible, à la connaissance du comportement du satellite Gaia en termes d'optique et d'attitude.

Cette image bi-dimensionnelle reconstruite sera spécialement utile dans l'analyse de possibles émissions originaires de l'environnement immédiat de la source observée – qu'elles soient d'origine intrinsèque, dans le cas d'objets étendus, ou extrinsèque dans le cas de sources secondaires.

2.2 Un *toy model* pour la reconstruction

Dans le but d'expliciter ce que l'on veut dire par reconstruction d'image dans le contexte de Gaia et introduire les idées de base d'algorithmes de reconstruction, nous allons considérer dans cette section un problème idéalisé basé sur l'observation de deux étoiles de même flux, projetées plus ou moins sur la même ligne de visée.

Ce système servira aussi pour que nous puissions montrer que pour la reconstruction exacte d'une image bi-dimensionnelle avec $k \times k$ pixels à partir de projections unidimensionnelles de k pixels, il est nécessaire, mais non suffisant, qu'au moins k projections soient disponibles.

Comme celui-ci n'est qu'un *toy model*, nous supposons que l'image est complètement définie par les sources, qu'il n'existe aucun type de bruit additionnel ou d'influence de l'instrument utilisé. Nous considérerons cependant, dans un deuxième temps, un bruitage du signal.

Nous avons commencé en supposant que le champs contenant les deux étoiles (on parlera de la source) est observé deux fois, avec deux angles de balayage qui forment un angle droit entre eux, et que les données disponibles pour la reconstruction ne sont que ces deux projections unidimensionnelles[3]. Un schéma du problème et de l'observation est donné en Figure 2.1.

3. Ce type d'observation « unidimensionnelle » est similaire aux fenêtres des colonnes AF, dont les *samples* sont formés par l'intégration des pixels dans la direction perpendiculaire au balayage

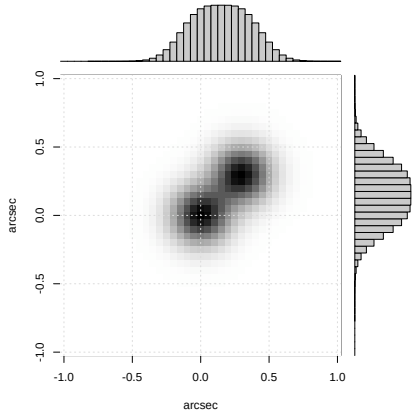


FIGURE 2.1 – *Toy model* avec la matrice $\mathbf{S}$ du signal de deux étoiles et leurs projections dans les directions perpendiculaire $v_\perp$ et parallèle $v_\parallel$ à l'axe x.

Donc, le problème direct qui consiste à produire l'observation à partir de la connaissance de l'image peut être facilement écrit : Soit la source donnée par un signal discrétisé dans l'espace bidimensionnel représenté par une matrice $\mathbf{S}$, avec (n, m) composants S_{ij}, le vecteur relatif à l'observation du balayage perpendiculaire[4] $\mathbf{v}_\perp$ à la direction définie par la ligne de la matrice s'écrit de la manière suivante :

$$\begin{cases} v_i = \sum_{j=1}^{m} S_{ij}, \text{ avec } S_{ij} \in S \\ \mathbf{v}_\perp = \{v_i : i < n\} \end{cases}$$

et de forme similaire, le vecteur relatif à l'observation du balayage parallèle $\mathbf{v}_\parallel$ est :

$$\begin{cases} v_j = \sum_{i=1}^{n} S_{ij}, \text{ avec } S_{ij} \in S \\ \mathbf{v}_\parallel = \{v_j : j < m\} \end{cases}$$

Donc, le problème inverse qui consiste dans la détermination des éléments de la matrice $\hat{\mathbf{S}} \sim \mathbf{S}$ à partir des vecteurs d'observations $\mathbf{v}_\parallel, \mathbf{v}_\perp$, peut être décrit par le système linéaire ci-dessous :

$$\begin{pmatrix} \sum_j \hat{S}_{1,j} \\ \vdots \\ \sum_j \hat{S}_{n,j} \\ \sum_i \hat{S}_{i,1} \\ \vdots \\ \sum_i \hat{S}_{m,1} \end{pmatrix} = \begin{pmatrix} v_{\perp,1} \\ \vdots \\ v_{\perp,n} \\ v_{\parallel,1} \\ \vdots \\ v_{\parallel,m} \end{pmatrix}$$

En observant le système ci-dessus, on s'aperçoit que ce *toy model* génère un

4. Un balayage perpendiculaire intégrera le signal dans la direction parallèle aux lignes de la matrice.

système sous déterminé (pour $n > 2$ ou $m > 2$) car le nombre d'inconnues $\hat{S}_{ij}$ est $(n \times m)$ tandis que le nombre d'équations est $(n + m)$.

En supposant une image rectangulaire, avec $n = m = k$, il existe k^2 inconnues à déterminer, et comme chaque observation fournit k équations, pour rendre le système possible et déterminé k projections $\mathbf{v}_\theta$[5] sont nécessaires. Cependant, bien que ce nombre d'équations soit nécessaire, nous savons que ce n'est pas une condition suffisante car rien n'est dit sur la dépendance entre ces équations.

L'idée de résoudre le problème inverse au moyen d'une résolution de systèmes linéaires tels que nous avons décrits de forme presque intuitive est la racine des diverses méthodes itératives, dont la plus connue est la Technique de Reconstruction Algébrique.[6] Dans ces méthodes, on ne recherche pas une solution exacte au système d'équations générées par les projections, mais ce que l'on fait c'est d'adopter un certain schéma de résolution itérative pour déterminer les éléments de la matrice $\hat{\mathbf{S}}$ qui optimisent un certain critère de qualité de la solution choisie de forme ad hoc.

Dans ces méthodes on re-écrit le problème sous la forme $\mathbf{AX} = \mathbf{B}$, où $\mathbf{A}$ est une matrice rectangulaire (généralement éparse) qui contient les poids des contributions de tous les pixels de $\mathbf{S}$ pour les éléments de $\mathbf{v}_\theta$, $\mathbf{X}$ est une matrice colonne avec les inconnues que l'on recherche à résoudre (les éléments de $\hat{\mathbf{S}}$ écrites sous la forme de colonne) et $\mathbf{B}$ est une matrice colonne avec les observations (les vecteurs $\mathbf{v}_\theta$). Une description étendue de ces méthodes et des algorithmes utilisés pour la résolution du système peut être trouvée par exemple chez Kak & Slaney (2001).

Comme ces méthodes sont itératives, il est possible d'inclure des informations connues à priori sur les éléments de la matrice, telles que par exemple non-négativité, forme permis pour l'objet que l'on essaye de reconstruire, conditions de *smoothness* de l'image (filtrant de façon naturelle le bruit du type poivre et sel, impulsif, ou les rayons cosmiques). Cependant, selon Marr (1979) ils sont très lourds en temps de calcul ce qui rend leur utilisation dans certains problèmes prohibitive. Ceci vient de ce qu'il est nécessaire de réaliser diverses itérations pour obtenir les images finales, étant entendu que chaque itération équivaut dans la réalité à une nouvelle reconstruction.

2.2.1 Reconstruction par la transformée de Radon

Une meilleure option que la Technique de Reconstruction Algébrique pour l'estimation de $\hat{\mathbf{S}}$ est l'utilisation de la transformée inverse de Radon (Radon, 1917,[7]). Une grande attention a été donnée à cette transformée durant les trois dernières décennies, parce qu'elle est parfaitement adaptées aux mesures de tomographie dans le domaine des Sciences de la Santé. Des descriptions de ses propriétés et applications (par exemple en reconstructions de tomographie par émission de positrons[8]) peuvent être trouvées dans Deans (1983), Toft (1996) ou Herman (2009).

5. Pour la détermination du vecteur d'observation $\mathbf{v}_\theta$ (angle de balayage ?), il suffit de réaliser la rotation de $\mathbf{S}$ de l'angle θ correspondant et de réaliser une somme pondérée (par la fraction de recouvrement du pixel de $\mathbf{v}_\theta$ par le signal du pixel de $\mathbf{S}$).

6. Ou ART, *Algebraic Reconstruction Technique.*

7. Traductions en anglais chez Deans (1983) et Parks (1986).

8. Ou PET, *Positron Emission Tomography.*

Cette transformée a déjà été utilisée antérieurement en Astronomie et des propositions pour son utilisation peuvent être rencontrées dès Aime et al (1978), et plus récemment chez Touma (1997) pour utilisation dans un type spécial de télescope appelé « télescope à fente ».[9]

Pour des données obtenues dans l'espace, la transformée de Radon a déjà été utilisée pour la reconstruction d'images du ciel entier en rayons-γ à partir d'observations avec le satellite BATSE d'occultations par la Terre (Zhang et al, 1993), et plus récemment chez Case et al (2009) où les auteurs annoncent d'appliquer cette technique pour la production d'une carte du ciel en γ avec les données du satellite Fermi.

Cette transformée correspond à une description du problème que nous essayons de résoudre en termes de fonctions continues au lieu de matrices discrètes, en supposant qu'il puisse être décrit par une fonction continue de deux variables $f(x, y)$.

Définition 1 *Soit $f(x, y)$ une fonction continue, la transformée de Radon est formée par toutes les intégrales de ligne de $f(x, y)$ calculées sous toutes les droites L existantes.*[10]

$$\check{f} = \int_L f(x, y) ds \tag{2.1}$$

Une représentation d'une fonction $f(x, y)$ et d'une droite L quelconques dans les systèmes de coordonnées concernés par la transformée sont présentées en Figure 2.2.

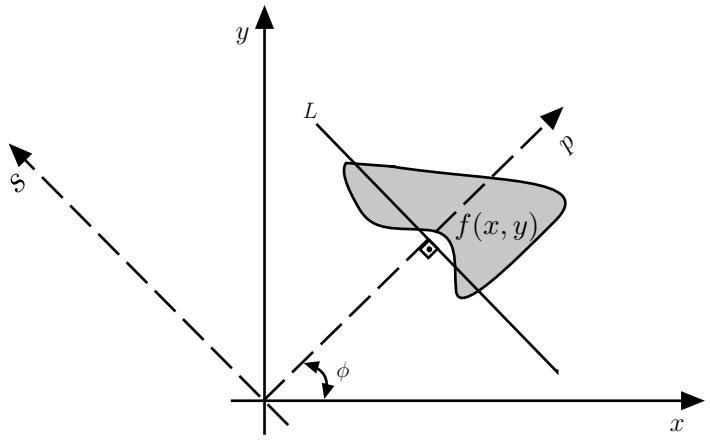


FIGURE 2.2 – Représentation graphique des grandeurs et des systèmes de coordonnées concernés dans la transformée de Radon.

À partir de la Définition 1 et de la Figure 2.2, on s'aperçoit qu'il est possible d'écrire explicitement la transformation $\check{f}$ de $f(x, y)$ sur une ligne L quelconque. En

9. Adopté par exemple par Aime et al (1978) pour des observations d'interferometrie *speckle* du phénomène de granulation solaire au *Kitt Peak National Observatory*, et proposé par Martin et al (1986) comme principe pour un télescope spatial équipé d'imagerie à haute résolution.

10. La définition la plus générale, pour une fonction dans $\mathbb{R}^n$, est l'intégrale de cette fonction sur tous les hyper-plans de dimension $(n-1)$, e.g. Deans (1983).

adoptant le système de coordonnées avec des axes ayant subi une rotation de l'angle ϕ (représenté dans le diagramme de la Figure 2.2 par les droites en pointillés) nous avons la transformation de variables suivante :

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} p \\ s \end{pmatrix}$$

Qui permet donc d'écrire la transformation sous la forme :

$$\check{f}(p,\phi) = \int_{-\infty}^{\infty} f(p\cos\phi - s\sin\phi, p\sin\phi + s\cos\phi)ds \tag{2.2}$$

Quand on connaît la valeur de la fonction $\check{f}$ pour n'importe quel (p,ϕ), on dit connaître la transformée de Radon de la fonction f. Cependant, au moment de son application, la transformée $\check{f}$ en elle-même n'est jamais obtenue, et ce que l'on connaît n'est seulement que sa valeur dans certains points de l'espace de (p,ϕ), dénommé espace de Radon. La représentation graphique de cet espace se dénomme Sinogramme.[11]

À partir de l'équation 2.2, nous voyons que $\check{f}(p,\phi)$ équivaut à une projection de $f(x,y)$ au point p et sur l'angle de balayage ϕ. De cette manière, fixer ϕ et balayer un sous-ensemble discrétisé dans l'axe p revient à déterminer le vecteur $\mathbf{v}_\phi$ d'une observation réalisée avec un angle ϕ. Ce que l'on désire donc est, à partir des vecteurs $\mathbf{v}_\phi$ qui forment un échantillon de l'espace de Radon, obtenir l'estimation de la fonction $f(x,y)$, en inversant le problème.

Dans son article original (Radon, 1917) la formule pour l'inversion était connue, mais comme il a été prouvé dans un théorème de Smith et al (1977) cité par Deans (1983) : *Une fonction f de support compact sur $\mathbb{R}^2$ est uniquement déterminée par un ensemble infini de ses projections, mais n'est pas uniquement déterminée par des ensembles finis, de ses projections.* Dans la pratique, cela signifie que toute solution obtenue à partir de l'inversion des observations sera seulement une approximation du signal original.

La technique non itérative la plus simple et directe pour déterminer une inversion approximative de cette transformée est appelée *Backprojection*, laquelle, comme son nom l'indique, n'est seulement qu'une projection dans la direction inverse des fonctions transformées. La Figure 2.3 représente le résultat obtenu par cette technique pour deux vecteurs d'observation et deux cas différents du *toy model*. Remarquez comme cette technique génère une version floue de l'image originale.

Bien que cette technique soit rarement utilisée seule, vu que la fonction $\check{f}$ obtenue après son application est une version très « floue » de la fonction originale f, elle représente un pas intermédiaire vers celle qui est probablement la technique la plus utilisée pour résoudre ce genre de problème : la *Filtered Backprojection*.

Il existe une étroite relation entre les transformées de Fourier et de Radon : réaliser une transformée de Fourier à deux dimensions $\mathcal{F}_2 f(x,y) = \tilde{f}(k_x,k_y)$ équivaut

11. La nomenclature a comme origine le fait qu'une fonction δ dans cet espace est une fonction trigonométrique.

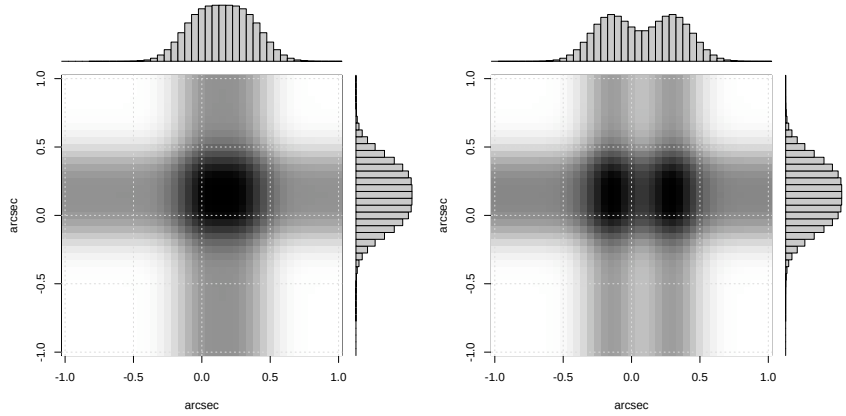


FIGURE 2.3 – *Toy model* avec la matrice $\hat{\mathbf{S}}$ récupérée par la technique de *Backprojection* appliquée sur deux observations ($v_\perp$ e $v_\parallel$). Le modèle à gauche est le même que la Figure 2.1. Sur le modèle de droite, les deux objets se trouvent dans les positions (-0,15" ; 0,0") et (0,3" ; 0,3").

à réaliser une transformée de Radon $\mathcal{R}f(x,y) = \check{f}(p,\phi)$ et, ensuite, une transformée de Fourier à une dimension dans la direction radiale $\mathcal{F}_1\check{f}(p,\phi) = \tilde{\check{f}}(k,\phi)$ (voir Deans, 1983, pour déduction). La différence entre les fonctions $\tilde{f}$ et $\tilde{\check{f}}$ obtenues n'est due qu'aux variables, de sorte qu'en adoptant une transformation de variables convenable[12] elles deviennent complètement équivalentes.[13] La Figure 2.4 schématise cette relation.

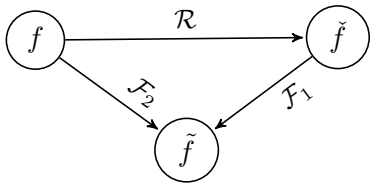


FIGURE 2.4 – Relation entre les transformées de Radon et de Fourier (adapté de Deans, 1983) : réaliser une transformée de Radon et, ensuite, une transformée de Fourier dans la dimension radiale d'une fonction quelconque, revient à réaliser une transformée de Fourier dans les deux dimensions originales.

En fonction de cette relation entre les deux transformées, on peut utiliser la transformée de Fourier pour l'inversion de $\check{f}$ (la fonction qui est estimée à partir des observations, ou vecteurs $\mathbf{v}_\phi$). Le diagramme de la Figure 2.5 schématise le processus d'inversion plus amplement adopté, qui est appelé *Filtered Backprojection*.

Dans le but de vérifier les résultats obtenus avec son utilisation, ainsi que pour

12. La transformation est $k = (k_x^2 + k_y^2)^{1/2}$ et $\phi = \tan^{-1}(k_y/k_x)$.
13. Dans le cas discrétisé, l'adoption d'une interpolation est nécessaire.

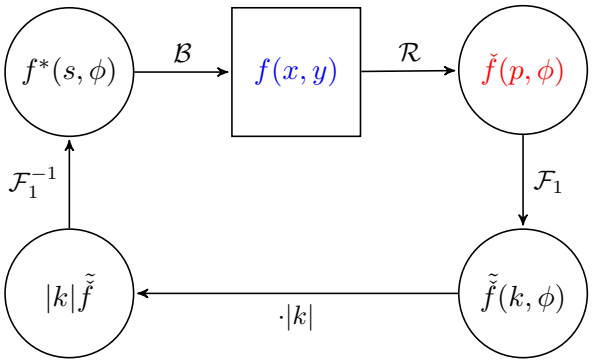


FIGURE 2.5 – Inversion de la transformée de Radon par *Filtered Backprojection*. En rouge, est indiquée la fonction que nous pouvons estimer à partir des observations, et en bleu, la fonction inversée à la fin du processus (adaptée de Deans, 1983).

analyser l' « efficacité » obtenue dans la reconstruction de l'image et sa susceptibilité à la présence de bruit dans les observations unidimensionnelles, nous appliquons cette méthode à notre *toy model*, en utilisant l'algorithme décrit par Toft (1996) et mis en œuvre dans R (R Development Core Team, 2009) par Schulz (2006).

Dans ce contexte, l'« efficacité » de la reconstruction doit être une grandeur qui nous servira pour comparer la qualité de reconstructions différentes et pour évaluer individuellement la qualité de chaque reconstruction. Intuitivement, le domaine inverse de cette fonction doit être $\{x \in \mathbb{R} | 0 \leq x \leq 1]\}$, en adoptant la valeur nulle dans le cas d'une reconstruction très éloignée de l'original et la valeur unitaire pour une reconstruction identique à l'original. De cette manière, en utilisant l'opération *modinv* définie dans l'Annexe « Module Inverse » , nous pouvons définir :

Définition 2 *Soient S_{ij} et $\hat{S}_{ij}$ les valeurs du signal original et du pixel reconstruit à la position (i,j), nous définissons l'efficacité E_{ij} de la reconstruction du pixel (i,j) par :*

$$E_{ij} = \begin{cases} \dfrac{\hat{S}_{ij}(modinv\ S_{ij})}{S_{ij}} & , \ pour\ \hat{S}_{ij} \leq 2S_{ij} \\ 0 & , \ pour\ \hat{S}_{ij} > 2S_{ij} \end{cases} \tag{2.3}$$

Nous avons réalisé diverses simulations de notre toy model pour vérifier le comportement de la fonction efficacité quand l'algorithme de *Filtered Backprojection* est appliqué à des cas avec diverses quantité de balayages et de niveaux de bruit dans le signal.

Comme on peut le vérifier qualitativement sur la Figure 2.6, un certain type d'information bi-dimensionnelle est récupéré avec seulement trois observations, au point de permettre la reconstruction du signal des deux objets avec un niveau d'efficacité légèrement supérieur à $\sim 60\%$. Clairement, en fonction de la petite

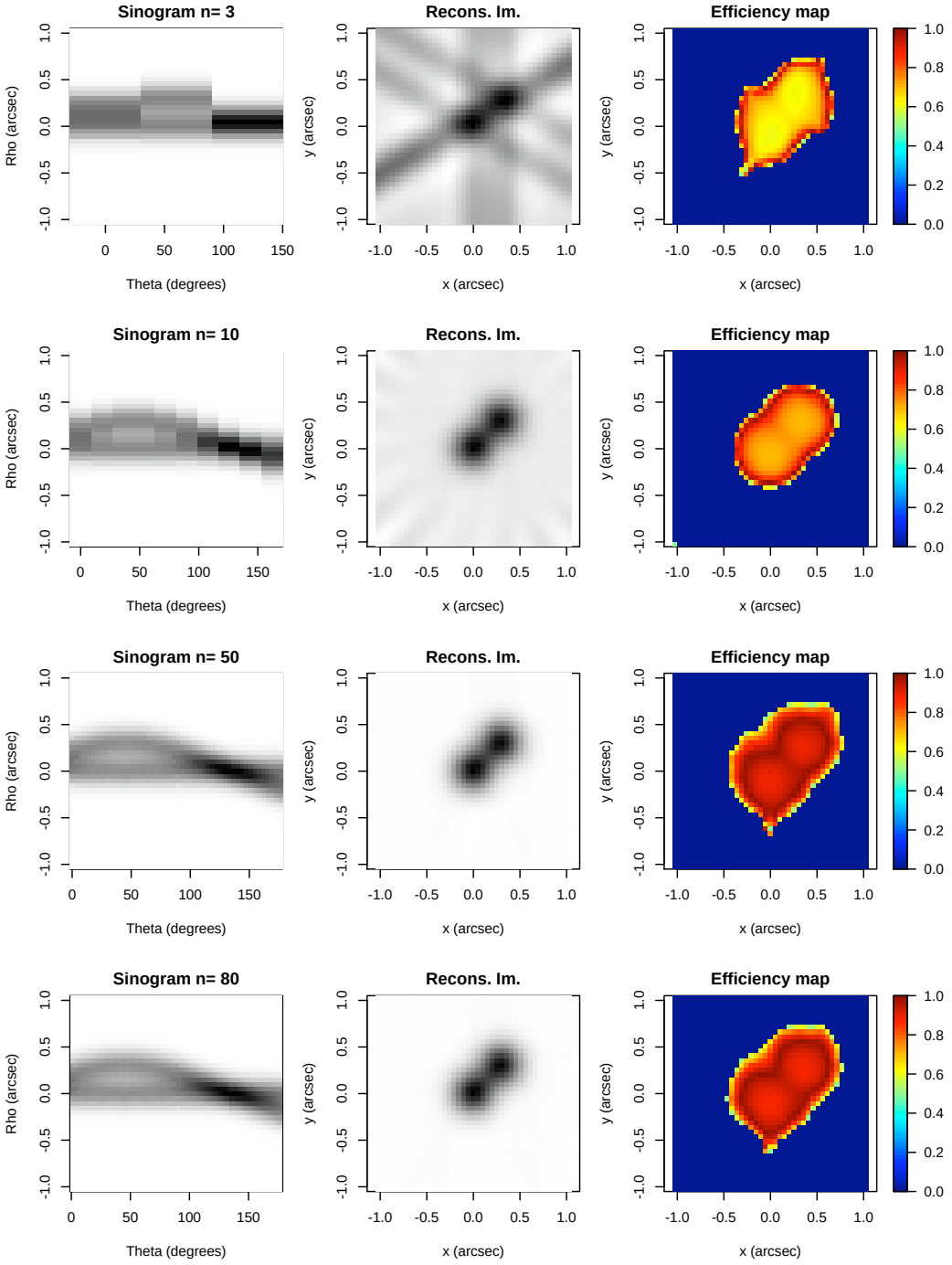


FIGURE 2.6 – Reconstruction de notre *toy model* par la transformée inverse de Radon en utilisant plusieurs quantités de projections ($n = 3, 10, 50, 80$). Les projections ont été distribuées régulièrement entre 0 et π radians et aucun bruit n'a été ajouté.

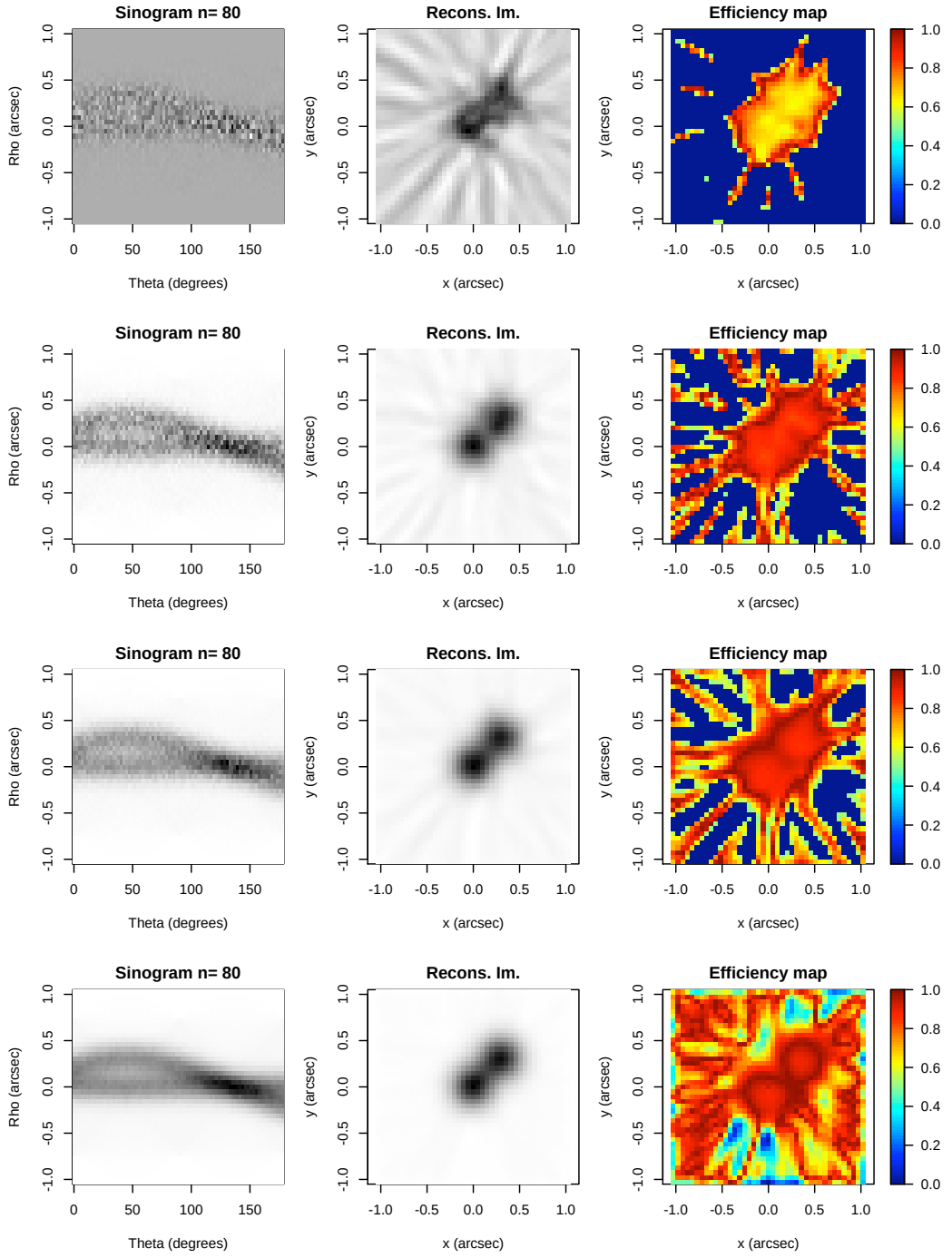


FIGURE 2.7 – Reconstruction de notre *toy model* par la transformée inverse de Radon en utilisant plusieurs niveaux de bruit dans le sinogramme ($S/B \sim 1, 5, 10, 40$). Toutes les reconstructions ont utilisé 80 projections régulièrement distribuées entre 0 et π radians. Un bias constant équivalent à 1% du signal moyen a été ajouté.

quantité d'observations, la reconstruction des deux objets n'est possible que grâce à l'existence d'au moins une projection adéquate des données.

Maintenant avec 10 observations, cette efficacité monte à $\sim 80\%$ et le format de l'objet devient essentiellement ponctuel, bien que le contraste avec le fond ne soit pas optimum. Avec 50 projections (un nombre de projections inférieur à celui espéré pour Gaia), l'efficacité des pixels de l'objet atteint des valeurs $> 90\%$.

Cependant, les résultats de l'analyse antérieure sont modifiés si l'on considère l'existence d'un niveau de bias et un bruit dans l'observation.[14] En Figure 2.7, qui présente des reconstructions pour différents niveaux de S/B simulés.

Considérant 80 projections (environ la quantité moyenne de balayages de Gaia dans n'importe quelle direction), pour un niveau de S/B égal à 1, nous voyons que bien que l'image reconstruite pour un objet double soit bien différente de celle qui serait produite pour un objet ponctuel simple, la distinction avec un objet étendu est essentiellement impossible. Cependant, pour $S/B = 5$ la situation devient plus satisfaisante, et bien que les objets soient encore légèrement déformés, il est déjà possible de percevoir la nature d'objet double du signal – résultat qui naturellement s'améliore encore plus si nous passons à $S/B > 10$.

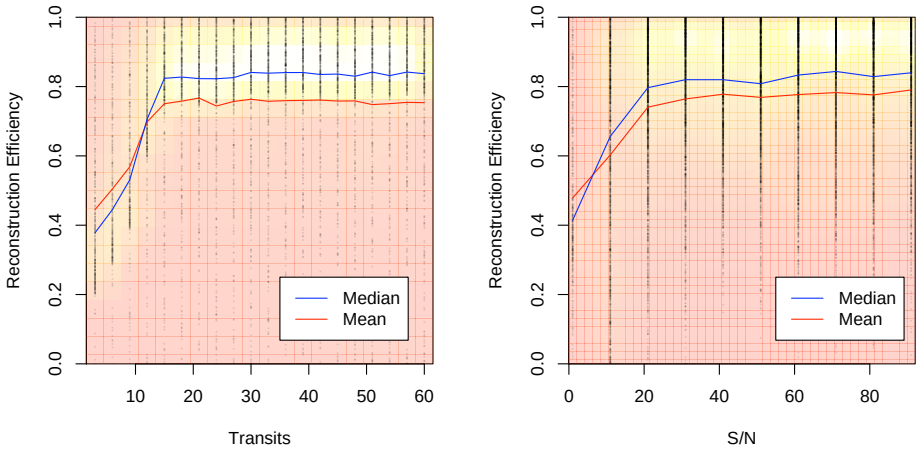


FIGURE 2.8 – Efficacité de la reconstruction par la transformée inverse de Radon en fonction du nombre de balayages (avec $S >> B$ à chaque observation) et du S/B (pour 80 balayages). Les points indiquent l'efficacité pour chaque pixel et la carte de couleurs utilisée comme fond représente une évaluation de densité des points.

En définissant les efficacités moyenne et médiane ainsi que les moyennes ou les médianes des efficacités de tous les pixels de l'image, et couvrant certains cas différents de nombres de balayages et de niveaux de S/B, nous pouvons analyser le comportement global de la fonction d'efficacité de l'algorithme de reconstruction par rapport à ces variables.

14. Le bruit a été simulé en tant qu'une indétermination du signal à chaque pixel, et la présence d'un composant aditif n'a pas été considérée.

La Figure 2.8 représente les résultats de ces simulations, et nous pouvons observer que le comportement des deux taux (moyen ou médian) est essentiellement le même : à partir d'environ 15 balayages ou $S/B \sim 20$ l'efficacité dans la reconstruction atteint un niveau stable autour de $\sim 80\%$.

Il est important de noter que ce taux a été défini, et est calculé, pour tous les pixels de l'image, et non seulement ceux sur lesquels le signal des objets est significatif. Dans le cas où l'on ne considérerait que les pixels au signal significatif (à partir des cartes des Figures 2.6 et 2.7) l'efficacité serait encore supérieure.

À partir des résultats obtenus dans cette sous-section, nous pouvons conclure que la *Filtered Backprojection* serait une technique candidate à être testée dans le futur pour la reconstruction des observations simulées de Gaia. Cependant elle ne peut pas être appliquée directement aux données de ce satellite pour des raisons que nous présentons dans la section suivante.

2.3 Reconstruction d'images pour Gaia

Le cas général de la reconstruction pour les fenêtres de Gaia est un problème plus compliqué que notre *toy model* pour plusieurs raisons. Le problème principal est que nous avons considéré dans la section antérieure que toutes les observations étaient similaires du point de vue du détecteur. Dans le cas de Gaia, chaque colonne de CCD du plan focal est et se comporte comme un détecteur différent où même l'échantillonnage du signal est différent. Un algorithme de reconstruction idéal applicable à Gaia doit au minimum prendre en compte les facteurs suivants :

1. *Fenêtres de différents types, avec samples de tailles angulaires différentes*

   Comme nous l'avons vu dans l'introduction de ce Chapitre, la fenêtre obtenue dans chaque colonne de CCDs est distincte aussi bien du point de vue de sa taille angulaire que de la stratégie d'échantillonnage adoptée. Cela signifie qu'un algorithme de reconstruction idéal serait capable de le prendre en compte, et reconstruire une image qui ne serait pas dégradée par la direction de plus petite résolution des *samples* des *AFs* ou par les samples (2x2 pixels) des *SM*.

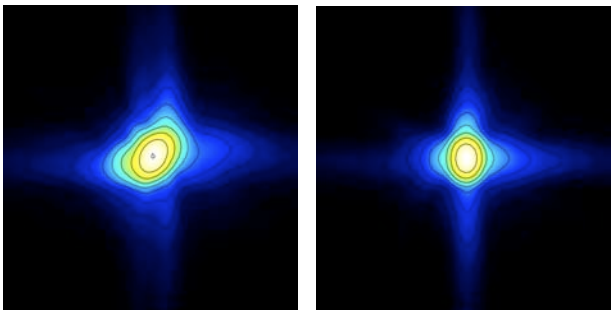
2. *Les fenêtres des AFs contiennent un signal provenant des deux champs de visée*

   Au cas où une reconstruction utiliserait le signal provenant des fenêtres obtenues par les *AFs*, il existe l'inconvénient que ces fenêtres possèdent un signal composé par l'addition des signaux des deux télescopes de Gaia. Un algorithme idéal pourrait, à partir des multiples observations du même objet, ainsi que des fenêtres des *SMs*, ignorer une partie du signal des *AFs*.

3. *Point Spread Function très asymétrique*

   Le miroir de Gaia est rectangulaire, de sortes que la *PSF* générée possède une forme de croix. Ceci complique la reconstruction, car au cas où l'algorithme adopté ne prendrait pas ce facteur en compte, en faisant tourner les observations durant la reconstruction de la matrice finale, l'effet de la *PSF* dégradera l'image reconstruite en la rendant floue. De plus la *PSF* varie en fonction de la position dans le plan focal, et selon le CCD dans lequel l'observation est réalisée la *PSF*

qui doit être considérée dans la reconstruction, possède une forme différente (voir Figure 2.9, par exemple).



(a) (b)

FIGURE 2.9 – PSFs simulées des *SMs*. (a) Colonne 1, Ligne 1. (b) Colonne 2, Ligne 3. (De Mary et al, 2006, , simulations de GIBIS 4).

4. *CTI – Inefficacité dans le transfert de charge*

   Comme nous l'avons vu dans le Chapitre 1, les observations de Gaia sont réalisées dans le mode TDI, où les charges des CCDs sont transférées d'un pixel à l'autre à la vitesse de rotation du satellite. Cependant le transfert de ces charges n'est pas parfait, et bien que la plus grande partie des électrons soit en effet transférée, l'efficacité n'est pas complète car il reste un peu de charge. Ceci est un problème pour la reconstruction, car les objets sont déformés dans le sens du transfert de charges (la position des objets est également légèrement altérée). La déformation des images peut être fortement dépendant de facteurs externes (comme d'autres objets brillants qui seraient passés avant ou des rayons cosmiques qui atteindraient la CCD). Un algorithme de reconstruction idéal devrait prendre ce fait en considération, probablement en corrigeant à priori les fenêtres observées avec l'adoption d'un modèle de distorsion, comme le CDM02, décrit par Short (2009).

5. *Effets temporaires*

   Les observations seront réalisées sur une durée d'environ 5 ans, durant lesquelles l'instrument sera exposé aux conditions spatiales, et de ce fait il subira des modifications (par exemple défauts en pixels de la CCD) qui influenceront les données des fenêtres transmises, et aussi le bruit instrumental ne sera pas strictement le même à chaque nouvelle observation d'une même source. De plus, il est possible que des variations intrinsèques surviennent dans les propres sources, de manière à ce qu'un algorithme idéal devrait être capable non seulement de reconstruire une bonne image de sources non variables, comme aussi devrait être capable de détecter qu'un type quelconque de variation temporelle est survenu, excluant cette observation de la reconstruction et lançant un type d'alarme quelconque.

2.3.1 Algorithmes de reconstruction

L'obtention d'une solution idéale[15] pour le problème de reconstruction est plutôt compliquée, et demande un temps de traitement relativement long (voir, par exemple Dollet, 2004). De plus, durant la mission on s'attend à avoir besoin de reconstruire des dizaines de millions d'images dans un but d'analyse de problèmes dus aux projections optiques de sources secondaires (Chapitre 3) et pour l'analyse de galaxies non résolues (Chapitres 5 et 6). Comme durant la mission, la reconstruction et l'analyse de toutes ces images doivent être réalisées relativement rapidement, il sera nécessaire d'adopter des méthodes simplifiées qui, bien qu'elles ignoreront beaucoup des effets décrits dans la Sous-section 2.3, conduiront à des images reconstruites de qualité suffisante pour être analysées.

Cependant, nous avons vu en section 2.2 qu'il est possible d'utiliser des algorithmes pour recréer des images bi-dimensionnelles à partir de données d'observations similaires aux fenêtres de Gaia. Et ceci, même si les fenêtres qui contiennent une information quelconque bi-dimensionnelle (comme celles produites par les *SkyMappers*) n'étaient pas disponibles.

Nous décrirons ici six algorithmes actuellement proposés pour la reconstruction d'images. Les quatre premiers ont été proposés et/ou mis en oeuvre par D. Harrison et ont été pensés depuis le début avec les contraintes de temps de traitement en tête. Les deux derniers sont des algorithmes plus détaillés et rigoureux, proposés et/ou mis en oeuvre par Dollet&Bijaoui et Pereyga&Bijaoui, qui bien qu'ils prennent en compte des facteurs de dégradation de l'image (comme ceux décrits dans la Sous-section 2.3), sont beaucoup plus exigeants du point de vue du temps de calcul, et sont probablement inapplicables durant la mission. Dans les descriptions ci-dessous, la matrice qui stocke l'image reconstruite finale sera toujours écrite comme étant $\hat{\mathbf{S}}$.

2.3.2 QuickStack

Le QuickStack a été développé par D. Harrison pour être l'algorithme le plus simple et rapide qui permette une reconstruction d'images à partir des fenêtres transmises par Gaia. Il est formé par un mélange entre la *Backprojection*, commentée antérieurement, et un empilage simple des fenêtres bi-dimensionnelles.

Le premier pas est l'initialisation de $\hat{\mathbf{S}}$ comme un réseau formé de pixels carrés d'une résolution donnée. Donc, pour les fenêtres unidimensionnelles originaires des *AFs*, une reconstruction est réalisée qui peut être vue comme une forme de *Backprojection* (avec re-échantillonnage), ne recouvrant que les pixels dont les données seraient physiquement dans ces fenêtres (ceci est important en fonction de la petite taille de ces fenêtres).

Pour les fenêtres bi-dimensionnelles des *SMs*, on réalise une rotation (avec re-échantillonnage). La valeur du *sample* est ajouté directement au pixel de $\hat{\mathbf{S}}$ qui est recouvert par ce *sample*, et la valeur finale est divisée par le nombre de *samples*, en

15. Ici définie comme une solution qui prend en compte tous les facteurs connus qui dégradent de l'image. Nous n'avons rien dit sur la qualité finale de l'image obtenue.

d'autre termes une moyenne est calculée. Finalement, l'image formée par l'empilage (somme) de toutes ces fenêtres est considéré comme l'image reconstruite.

Cet algorithme a la caractéristique de générer des images à l'apparence « floue » (ex. Figures 2.11 et 2.12). En plus du calcul de la moyenne des samples, un autre facteur qui contribue à cela est que l'existence de la *PSF* de l'instrument n'est pas prise en compte par l'algorithme : le format de la *PSF* de Gaia est similaire à une croix asymétrique alignée avec la fenêtre.

2.3.3 Drizzle

Un second algorithme qui a été mis en oeuvre est dénommé Drizzle, inventé par A. Fruchter et R. Hook (Fruchter & Hook, 2002) pour l'empilage d'images du télescope spatial Hubble, et qui a été appliqué avec succès dans l'obtention des images finales du *Hubble Deep Field* (Williams et al, 1996). Sa mise en oeuvre actuelle se doit à D. Harrison, bien qu'il ait été proposé afin d'être utilisé sur Gaia pour la première fois par Nurmi (2003).

De même que pour le QuickStack, le premier pas ici est l'initialisation de $\hat{\mathbf{S}}$ comme un réseau formé de pixels carrés, avec une résolution donnée.

L'idée de base de cet algorithme consiste dans l'idée que l'image reconstruite fonctionne comme un collecteur de photons des *samples* observés : on extrait la partie centrale (taille fixée arbitrairement) de chaque *sample* des fenêtre engendrant ce qui dans le langage de l'algorithme est dénommé « gouttes » , et ensuite chacune de ces gouttes est « versée » dans l'image finale $\hat{\mathbf{S}}$.

La valeur ajoutée à chaque pixel dépend de la fraction du recouvrement mutuel entre la goutte et le pixel du réseau, et cette addition est pondérée par un poids choisi par l'utilisateur. Les équations originales de Fruchter & Hook (2002) sont :

$$\begin{aligned} W'_{x_o y_o} &= a_{x_i y_i x_o y_o} w_{x_i y_i} + W_{x_o y_o} \\ I'_{x_o y_o} &= \frac{d_{x_i y_i} a_{x_i y_i x_o y_o} w_{x_i y_i} s^2 + I_{x_o y_o} W_{x_o y_o}}{W'_{x_o y_o}} \end{aligned} \tag{2.4}$$

où l'indices i désignent fenêtres observées et o l'image reconstruite, d la valeur du pixel (« goutte ») de la fenêtre, a la proportion de recouvrement entre les *samples* observés et les pixels de l'image finale ($0 \leq a \leq 1$), $\mathbf{W}$ la matrice des poids w_{ij}, et $\mathbf{I}$ la matrice de l'image reconstruite et $'$ indique que la procédure est itérative (l'inclusion de chaque observation d'effectue dans une itération différente). Dans notre notation, la matrice formée par les valeurs $I_{x_o y_o}$ après que toutes les observations aient été prises en compte est la matrice $\hat{\mathbf{S}}$. On peut dire que cet algorithme se réduit au QuickStack avec pour tous les pixels : $w = s = 1$ et $a = 1$ quand il existe un recouvrement ou $a = 0$ quand il n'existe pas.

Cet algorithme permet une amélioration de la résolution de l'image finale, grâce au re-échantillonnage (il peut être vu comme une technique de *dithering*). Cependant, il ne résout pas le problème de la dégradation de l'image finale par la rotation de la *PSF*, et bien que la reconstruction obtenue puisse atteindre de plus grandes

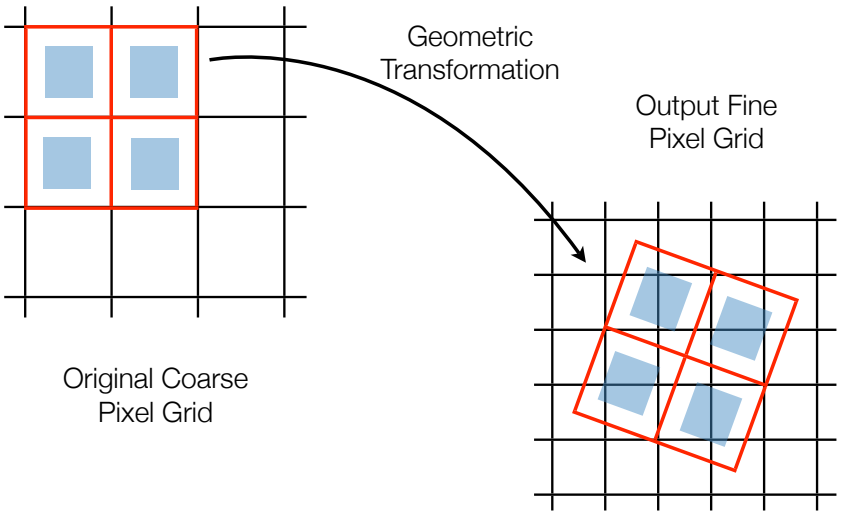


FIGURE 2.10 – Schéma de la technique de Drizzle. Notez que le pixel central de l'image reconstruite ne reçoit aucun signal.

résolutions que celles produites par le QuickStack, il génère des artefacts et prend plus de temps de calcul.

De plus, la méthode ne possède pas une description de la façon dont la taille de la partie centrale des samples (le paramètre d'échelle) doit être déterminé. Un choix empirique doit être fait afin que la taille des gouttes ne soit pas trop grande afin de gagner en résolution, ni trop petite pour éviter de dégrader l'image et pour recouvrir tous les pixels de l'image finale.

2.3.4 ShuffleStack

Un autre algorithme développé par D. Harrison, le ShuffleStack, est en fait un QuickStack plus sophistiqué. ShuffleStack fonctionne à partir d'un empilage simple des fenêtres observées, mais il possède trois subtilités importantes qui le différencient et lui permettent d'atteindre une meilleure qualité de reconstruction des images.

La première, et la plus importante, est que l'algorithme reconstruit en premier les pixels avec la plus grande quantité d'informations (c'est-à-dire, avec le plus grand nombre de *samples* contribuant à son estimation). Pour cela, une analyse préliminaire est réalisée qui, partant de chaque pixel de l'image finale, répertorie toutes les fenêtres (et les positions des *samples*) qui contribuent à ce pixel.

Chaque pixel de l'image finale est alors reconstruit en partant du pixel avec le plus grand nombre de samples vers le pixel avec le plus petit nombre. Cette reconstruction est réalisée à partir du calcul d'une moyenne avec un rejet itératif a 2σ (valeur définie empiriquement par D. Harrison). La valeur du pixel reconstruit est alors soustraite aux samples qui ont contribué à ce pixel – ce qui donne son nom à l'algorithme, car le signal est finalement transféré, des pixels avec moins d'observations vers les pixels avec plus d'observations.

Il est important de remarquer qu'il n'existe aucune garantie que cet algorithme ne reconstruise pas de pixels avec des valeurs négatives – ce qui se passe dans certains cas – et que cet algorithme possède la caractéristique de concentrer les signaux dans les pixels les plus observés.

2.3.5 BinOutliers

Dans la plus grande partie des cas où il sera nécessaire de reconstruire une image durant la mission, on cherchera à comprendre l'origine de perturbations de la solution astrométrique de la source primaire (située au centre de l'image). Pour cela D. Harrison a développé un algorithme qui prend ce facteur en considération.

A l'image de ShuffleStack, BinOutliers est un algorithme dérivé de QuickStack. Cependant, son point de départ est d'identifier quels samples contribuent aux pixels « centraux »[16] de $\hat{\mathbf{S}}$, et éviter qu'ils soient considérés dans la reconstruction des pixels adjacents. Ceci parce que, en général, ce que l'on cherche à observer, ce sont des objets plus faibles que l'objet central et la contribution de cet objet central peut masquer ces objets plus faibles.

De la même manière qu'avec ShuffleStack, les pixels de l'image finale sont reconstruits un à un, mais sans ordre préférentiel. Premièrement, on calcule la moyenne m_B et l'écart type σ_B à partir des valeurs des 70% de *samples* de plus petite valeur (pour exclure les *outliers*) qui contribuent à ce pixel dans l'image finale. Donc, la reconstruction du pixel est réalisée à partir du calcul de la moyenne des valeurs des *samples* après un rejet des valeurs des *samples* plus distants que $2.5\sigma_B$ de m_B (valeur définie empiriquement par D. Harrison). La valeur reconstruite n'est pas soustraite des *samples* qui ont contribué au pixel.

Les images obtenues en utilisant cet algorithme possèdent une structure similaire à celle des images produites par ShuffleStack, mais sont beaucoup plus lisses à l'exception des pixels centraux, qui tendent à concentrer le signal généré par l'étoile primaire. Ceci survient justement parce que l'algorithme empêche le déplacement libre du signal des samples comme dans ShuffleStack, ce qui entraîne moins de déformations des objets reconstruits (ex. Figure 2.11).

Comme on le verra dans le prochain Chapitre, cet algorithme est plus efficace que les trois autres pour l'analyse de la multiplicité dans les images reconstruites.

2.3.6 Régularisation Tikhonov

Originalement décrite par Tikhonov (1976) comme solution de problèmes mal posés (Demoment, 2002), son adaptation pour la reconstruction d'images pour Gaia a été proposée pour la première fois par Claire Dollet, durant son doctorat sous la direction d'Albert Bijaoui (Dollet, 2004; Dollet et al, 2005), et a ensuite été analysée par Mary et al (2006).

Elle peut être vue comme une technique de reconstruction algébrique. La première étape consiste à relier l'objet observé et le résultat de l'observation sous une forme

16. La valeur définie empiriquement par D. Harrison pour la distance maximum utilisée pour la considération d'un pixel de l'image reconstruite comme centrale est 50 mas

matricielle $\mathbf{D}_i = \mathbf{H}_i\mathbf{F}$, où $\mathbf{D}_i$ est un vecteur colonne avec les *samples* de la i-ème fenêtre, $\mathbf{H}_i$ une matrice qui décrit l'instrument durant l'observation de cette ième fenêtre et $\mathbf{F}$ un vecteur colonne qui décrit l'objet original. Ceci revient à écrire $\mathbf{S}$ sous la forme d'un vecteur colonne. Notons que la matrice $\mathbf{H}$ peut être aussi complexe que l'on veut et contenir la PSF, l'échantillonnage ou les déformations optiques.

De cette manière, connaissant la matrice inverse $\mathbf{H}_i^{-1}$ on pourrait obtenir $\hat{\mathbf{F}}$. Cependant, deux problèmes interdisent la résolution par ce moyen : $\mathbf{H}_i$ n'est pas carrée, donc il n'admet pas d'inverse, et $\mathbf{H}_i$ n'est pas une matrice de convolution, rendant impossible une résolution dans l'espace de Fourier. De plus, le problème est mal posé puisque le nombre d'éléments de $\mathbf{D}_i$ est très inférieur au nombre d'éléments de $\mathbf{F}$ (Dollet, 2004). Il faut donc rechercher une solution $\hat{\mathbf{F}}$ qui minimise la somme des différences quadratiques entre toutes les N observations :

$$\min\left\{\sum_{i=1}^{N}(\mathbf{D}_i - \mathbf{H}_i\hat{\mathbf{F}})^2\right\}$$

Comme l'algorithme de Tikhonov a été écrit pour résoudre des problèmes du type $y = Ax$, il est nécessaire de re-écrire le problème sous cette forme. En utilisant la matrice transposée de $\mathbf{H}_i$, la minimisation ci-dessus devient alors :

$$\sum_{i=1}^{N}\mathbf{H}_i^T\mathbf{D}_i = \sum_{i=1}^{N}\mathbf{H}_i^T\mathbf{H}_i\hat{\mathbf{F}}$$

Donc, le processus itératif dérivé en Dollet (2004) reconstruit l'image $\hat{\mathbf{F}}$ à l'étape $(n+1)$ par :

$$\hat{\mathbf{F}}^{n+1} = \hat{\mathbf{F}}^n + \alpha\sum_{i=1}^{N}\mathbf{H}_i^T\mathbf{D}_i - \alpha\left(\sum_{i=1}^{N}\mathbf{H}_i^T\mathbf{H}_i + \lambda\mathbf{L}\right)\hat{\mathbf{F}}^n \tag{2.5}$$

où L est l'opérateur Laplacien discret à deux dimensions, α est le paramètre de la convergence et λ est un paramètre de régularisation. Il est important de remarquer que ces deux paramètres doivent être déterminés de façon empirique.

La résolution du processus itératif est naturellement plus lent que tous les algorithmes décrits antérieurement, car diverses reconstructions intermédiaires sont réalisées. Mais le plus grand problème de cet algorithme est la détermination des matrices $\mathbf{H}$: en fonction du type de fenêtre, traiter 100 angles de balayage différents peut prendre plus de deux heures et demi pour seulement un objet (Dollet, 2004). En conséquence, l'utilisation de cet algorithme durant la mission a été abandonnée, même si Mary et al (2006) a montré qu'il est possible de restaurer des objets aussi faibles que $G \sim 23$. Actuellement, il n'existe aucune mise en oeuvre de cet algorithme directement compatible avec les données produites par les versions actuelles des systèmes de simulation utilisés par Gaia.

De plus, un problème des images générées par cet algorithme est l'introduction d'artefacts, comme on peut le voir sur la Figure 2.13. De plus cet algorithme ne garantit pas que le flux des objets soit préservé (Peyrega, 2007). Pour des objets

ponctuels, ces problèmes ont été minimisés ou résolus avec l'introduction de la famille d'algorithmes Clean, que nous verrons ci-dessous.

2.3.7 Clean & Cleanest

À l'origine développé pour la synthèse d'images de radiotélescopes, l'algorithme Clean (Högbom, 1974) a été adapté pour être utilisé sur Gaia par Charles Pereyga sous la direction d'Albert Bijaoui (Peyrega, 2007).

Dans le contexte de Gaia, cet algorithme peut être utilisé pour reconstruire les images d'objets ponctuels. Son principe de base considère les objets du champs comme des δs de Dirac et reconstruit leur image un à un, en soustrayant des observations leur contribution à chaque itération. L'algorithme Cleanest ne consiste qu'en une étape supplémentaire durant l'exécution de l'algorithme Clean, garantissant que la contribution des pixels les plus brillants de l'image soit prise en compte en premier.

Comme dans la reconstruction par la régularisation de Tikhonov, son point départ est la construction des matrices $\mathbf{D}_i$ et $\mathbf{H}_i$, contenant les *samples* et la modélisation de l'instrument, respectivement, pour la i-ème observation. Donc, deux nouvelles matrices, $\mathbf{D}_{global}$ et $\mathbf{H}_{global}$, sont créées à partir de la composition de toutes les N matrices $\mathbf{D}_i$ et $\mathbf{H}_i$ originales :

$$\mathbf{D}_{global} = \begin{pmatrix} D_{1,1} \\ \vdots \\ D_{1,\text{size}(\mathbf{D}_{1_1})} \\ D_{2,1} \\ \vdots \\ D_{N,\text{size}(\mathbf{D}_N)} \end{pmatrix} \qquad \mathbf{H}_{global} = \begin{pmatrix} H_{1,1} & \dots & H_{1,\text{size}(\hat{\mathbf{F}})} \\ \vdots & \vdots & \vdots \\ H_{1,\text{size}(\mathbf{D}_1)} & \dots & H_{1,\text{size}(\hat{\mathbf{F}})} \\ H_{2,1} & \dots & H_{2,\text{size}(\hat{\mathbf{F}})} \\ \vdots & \vdots & \vdots \\ H_{N,\text{size}(\mathbf{D}_N)} & \dots & H_{N,\text{size}(\hat{\mathbf{F}})} \end{pmatrix} \tag{2.6}$$

où size indique la taille des vecteurs $\mathbf{D}$ et $\hat{\mathbf{F}}$.

Donc, en considérant que chaque pixel de $\hat{\mathbf{F}}$ est constitué de fonctions δ de Dirac d'amplitude α, le problème se réduit à trouver les valeurs α_i telles que :

$$\min \left\{ \sum_{j=1}^{\text{size}(\mathbf{D}_{global})} \left(D_{global,j} - \alpha_i H_{global,j,i} \right)^2 \right\}$$

où i est un indice qui passe par les pixels de $\mathbf{F}$. Ce problème est donc résolu par :

$$\alpha_i = \frac{\sum_{j=1}^{\text{size}(\mathbf{D}_{global})} \left(H_{global,j,i} D_{global,j} \right)}{\sum_{j=1}^{\text{size}(\mathbf{D}_{global})} H_{global,j,i}^2}$$

Mais Clean possède une étape additionnelle, qui prête son nom à l'algorithme : après la détermination de α_i, la variance de cette valeur est aussi obtenue :

$$\sigma^2(\alpha_i) = \frac{\sum_{j=1}^{\text{size}(\mathbf{D}_{global})} H_{global,j,i}^2 \sigma^2(D_{global,j})}{\left(\sum_{j=1}^{\text{size}(\mathbf{D}_{global})} H_{global,j,i}^2 \right)^2}$$

de manière à ce que pour chaque α_i, le niveau de signal/bruit est estimé ($\alpha_i/\sigma(\alpha_i)$).

Finalement, à chaque itération n, α_i e $\sigma(\alpha_i)$ sont calculés pour tous les pixels de $\mathbf{F}$ mais seulement le pixel qui possède le plus grand niveau de S/B ($\alpha_i^{\max}$) est stocké dans $\hat{\mathbf{F}}$. Sa contribution est alors soustraite des observations (matrice $\mathbf{D}_{global}$) par un facteur γ (appelé facteur de nettoyage, ou *clean factor*), qui possède des valeurs entre 0 et 1. Donc, la nouvelle matrice $\mathbf{D}_{global}$ utilisée dans l'itération $(n+1)$ est construite avec les éléments j :

$$D_{global,j}^{(n+1)} = D_{global_j}^{n} - \gamma \alpha_i^{\max} H_{global,ji}$$

$H_{global,ji}$ ci-dessus est un vecteur colonne, car i est fixé par $\alpha_i^{\max}$.

Naturellement, ces algorithmes possèdent le même problème de la création des matrices $\mathbf{H}$ qui rendent l'utilisation sur une large échelle de la reconstruction par régularisation de Tikhonov impossible. Cependant, Peyrega (2007) conclut qu'en comparaison avec Tikhonov, la création d'artefacts durant la reconstruction est petite, que le signal de ces objets est aussi généralement plus petit que celui des objets, et que le problème d'altération du flux est inexistant. De cette manière son adoption serait préférentielle dans le cas d'étude d'objets ponctuels.

Il est clair que dans le cas d'objets étendus Clean/Cleanest ne seraient pas adaptés. Cependant Peyrega (2007) note qu'il serait possible de traiter de telles sources en adoptant un catalogue de modèles d'objets.[17] De cette manière, au lieu de reconstruire $\mathbf{F}$ par un $\hat{\mathbf{F}}$ créé à partir de fonctions δ, il serait adopté un « alphabet » de fonctions différentes, décrivant par exemple, des formes possibles pour les Galaxies. Cependant, aucun algorithme de ce type n'a été développé jusqu'à aujourd'hui.

2.3.8 Exemples de reconstructions

Pour présenter la qualité des images obtenues à partir de l'utilisation des algorithmes décrits dans la section antérieure, nous avons réalisé des simulations en utilisant un simulateur officiel du DPAC, le GIBIS 6 (Babusiaux et al, 2009). Nous avons simulé une région standard (avec 85 balayages), qui contenait cinq objets : un principal dans le centre de l'image avec des magnitudes $G_p = 15$ ou $G_p = 19$ et quatre objets secondaires, avec des magnitudes $G_s = G_p + 1.1, 1.5, 1.8$ ou 2.1.

Les balayages ont été utilisés pour reconstruire des images avec les algorithmes développés en Java (QuickStack, Drizzle, ShuffleStack et BinOutliers), étant donné que probablement l'un d'entre eux sera adopté pour le traitement des observations.

Sur les Figures 2.11 et 2.12 nous pouvons observer les différences entre les images obtenues avec ces quatre algorithmes. La reconstruction QuickStack, par le fait qu'elle est du type *Backprojection*, génère une image relativement floue (similaire à ce qui se passe avec le *toy model*), mais les objets ne sont pas déformés, comme ceci se produit avec les reconstructions ShuffeStack.

Cependant, dans les reconstructions réalisées avec la méthode ShuffleStack les objets secondaires apparaissent beaucoup plus clairement que dans les images obte-

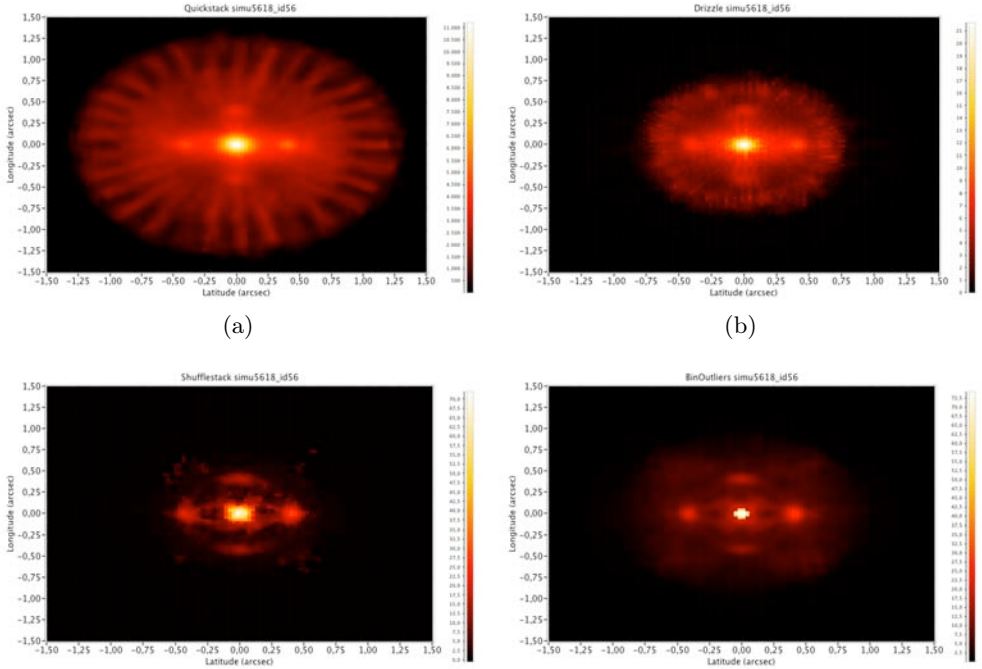
17. La définition de ce catalogue est probablement une tâche plus compliquée.

nues avec QuickStack, même si cette méthode crée des artefacts dans l'image finale, et qu'il existe la possibilité de reconstruire des pixels avec des valeurs négatives (ce qui n'a pas de sens physique).

La reconstruction par la méthode de Drizzle génère une image ou les objets ont une intensité égale ou supérieure à ce qu'elles devraient être, ce qui est un point extrêmement négatif pour l'adoption de cette méthode, en plus du fait que la reconstruction est fortement dépendante des paramètres dont le choix est empirique.

La méthode BinOutliers semble produire les images les plus fidèles. Le pic central est peu réaliste mais le « nettoyage » des artefacts est particulièrement efficace (contrairement à Drizzle et ShuffleStack). De plus les objets secondaires sont particulièrement bien mis en valeur (même avec une magnitude $G = 21.1$, sur la Figure 2.12). Dans cette méthode, le signal qui générait des artefacts dans les autres méthodes génère ici un fond diffus qui n'entrave pas l'analyse.

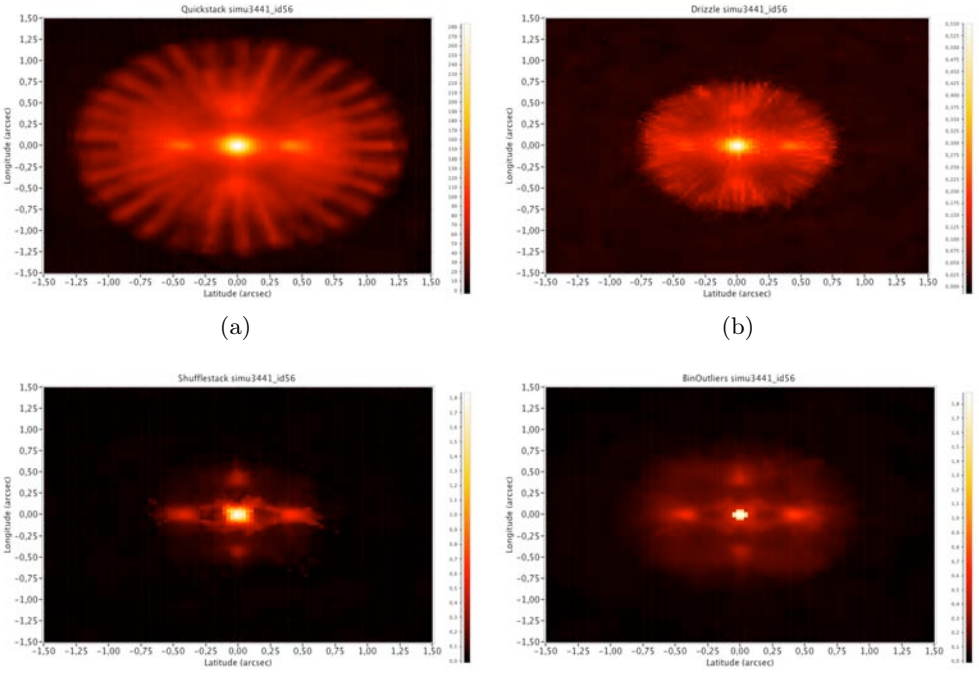
À titre de comparaison, sur la Figure 2.13 nous avons présenté deux reconstructions de Mary et al (2006) réalisées en utilisant un algorithme de régularisation de Tikhonov. Sur la Figure (a) l'objet central possède une magnitude $G = 18$ tandis que l'objet secondaire, dans le coin supérieur droit de l'image, $G = 20$. Il est intéressant de noter qu'en (b) cet algorithme a été capable de reconstruire des objets allant jusqu'à cinq magnitudes de différence.



(a) (b)

(c) (d)

FIGURE 2.11 – Exemples de reconstruction d'une même région contenant cinq objets (quatre secondaires). Objet primaire avec $G = 14$. (a) QuickStack, (b) Drizzle, (c) ShuffleStack et (d) BinOutliers. Pixels de 30 mas.



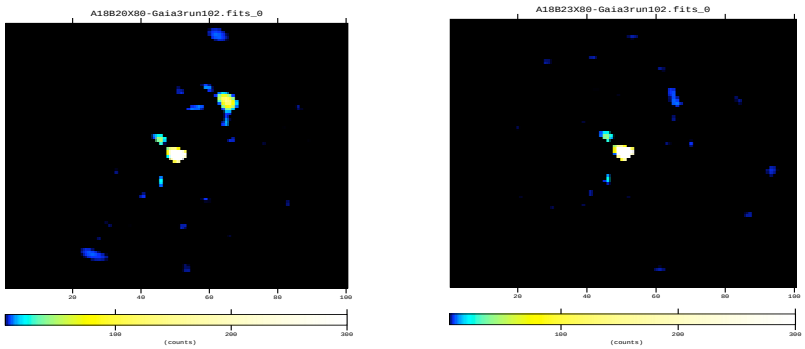
(a) (b)

(c) (d)

FIGURE 2.12 – Comme 2.11, mais avec un objet primaire de $G = 19$.

Selon Mary et al (2006) les différents artéfacts générés par l'algorithme de Tikhonov peuvent être nettoyés à partir de la soustraction d'une deuxième reconstruction avec les mêmes fenêtres, mais avec le signal de l'étoile centrale seul.

Nous ne présentons aucune reconstruction Clean, car elles ne génèrent que des images composées de δ de Dirac, où les pixels du centre de l'objet seront soit ON soit OFF – montrant si l'objet a été reconstruit ou non.



(a) (b)

FIGURE 2.13 – Images obtenues par la reconstruction par régularisation de Tikhonov. Objet central de magnitude $G = 18$ et secondaires de $G = 20$ (a) et $G = 23$ (b). Pixels de 50 mas. (De Mary et al, 2006)

2.4 Couverture spatiale et angulaire des reconstructions

La qualité de l'image reconstruite pourra être extrêmement dépendante de la position de la source sur la sphère céleste. Ceci, car à chaque position du ciel non seulement le nombre de balayages total sera différent (comme nous l'avons vu dans le Chapitre 1), mais aussi la distribution angulaire de ces balayages sera variable.

Cependant, par le fait que nous connaissons d'avance la taille de la fenêtre pour n'importe quelle observation et la loi de balayage du satellite, nous pouvons étudier cette configuration au moyen de simulations qui indiquent comment chaque coordonnée de la sphère céleste serait recouverte par des fenêtres au cas où une source serait détectée dans cette position. Pour cela, nous sommes partis de la définition de deux grandeurs qui vont nous aider pour une telle analyse :

Définition 3 *La couverture fractionnelle C_F est définie comme étant le rapport entre la surface du polygone résultant de l'union des fenêtres obtenues dans tous les balayages en un point donné et la surface d'un cercle d'un diamètre égal à la diagonale de la plus grande fenêtre observée en ce point.*

La grandeur ci-dessus décrit la fraction de la couverture spatiale normalisée par la couverture maximum (qui serait la surface du cercle définie par d'infinis balayages distribués sur des angles aléatoires) des données en un point donné. Néanmoins, elle ne nous dit rien explicitement sur la résolution d'une image reconstruite. Il est important de posséder un indicateur quantitatif de la résolution théorique accessible pour une reconstruction qui peut être atteinte dans une région donnée du ciel. Celle-ci est clairement limité par la résolution maximale des *samples* transmis pour cette coordonnée dans la direction du balayage.

Etant donné que les pixels de Gaia sont rectangulaires, la résolution maximum que l'image reconstruite peut avoir est liée à la distribution des angles d'observation, étant entendu que la résolution maximale ne peut être atteinte que dans le cas où il existerait des fenêtres obtenues avec des angles de $\pi/2$ entre eux. En utilisant l'opération module inverse (définie dans l'Annexe A), il est possible de définir une grandeur que nous appelons couverture angulaire qui permet de quantifier de combien nous seront proches de cet angle idéal :

Définition 4 *Soit $\mathbb{A}$ l'ensemble de n éléments formé par tous les n angles de balayage (en rad) de toutes les observations sur une coordonnée céleste donnée, nous avons défini l'ensemble des couvertures angulaires $\mathbb{V}$ comme étant l'ensemble avec $(n^2+n)/2$ éléments formé par $\{A_j, A_k \in A : |A_k - A_j|(modinv\ \pi/2)\}$. La couverture angulaire entre la k-ième et la j-ième observation est définie comme étant l'i-ième élément de $\mathbb{V}$.*

Corollaire 1 $\forall v \in \mathbb{V}, 0 \leq v \leq \pi/2$.

Définition 5 *C_A est la couverture angulaire maximum normalisée :*

$$C_A = \frac{\max(\mathbb{V})}{\pi/2} \tag{2.7}$$

La couverture angulaire maximum normalisée définie ci-dessus est un indicateur de l'existence de l'information angulaire nécessaire entre, au moins, deux balayages pour permettre la reconstruction d'images avec des pixels carrés d'une résolution donnée (égale à la résolution *along-scan*). Dans les cas extrêmes, si $C_A = 1$, on peut obtenir dans l'image reconstruite des pixels avec la même taille que celle des samples dans la direction *along-scan*, et si $C_A = 0$ la meilleure résolution que l'on peut obtenir sera celle des *samples* dans la direction *across-scan*. D'un autre côté, le calcul de la taille du pixel qui peut être reconstruit pour des cas intermédiaires de C_A est quelque chose qui dépend fortement de l'algorithme de reconstruction adopté.

Pour étudier ces deux grandeurs, C_F et C_A, et leurs respectives variations dans le ciel, nous devons réaliser des simulations des balayages sur le ciel entier. Cependant, la simulation des balayages de Gaia sur une coordonnée céleste donnée nous fournit les angles de chaque observation, mais pas les fenêtres observées,[18] bien que ceci soit suffisant pour calculer C_A, il ne nous permet pas de calculer C_F.

Néanmoins, comme nous connaissons les tailles de tous les types de fenêtres, nous pouvons calculer les coordonnées de leurs sommets en coordonnées galactiques, en définissant ainsi les polygones nécessaires pour le calcul de C_F. De plus, comme la couverture spatiale est complètement déterminée par la plus grande fenêtre disponible, nous pouvons considérer uniquement les fenêtres des *SkyMappers*, car elles sont toujours plus grandes que les fenêtres des *AstroFields* (section 2.1).

Donc, comme les dimensions l_1 et l_2 (*across-scan* et *along-scan* respectivement) des fenêtres des *SkyMappers* sont connues, connaissant l'angle de balayage γ nous pouvons écrire les quatre sommets (x_i, y_i) du rectangle résultant comme étant :

$$\begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ x_4 & y_4 \end{pmatrix} = d \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ \cos\phi & \sin\phi \end{pmatrix} \tag{2.8}$$

où,

$$\begin{aligned} d &= \sqrt{\left(\tfrac{l_1}{2}\right)^2 + \left(\tfrac{l_2}{2}\right)^2} \\ \theta &= \gamma + \arctan\left(\tfrac{l_1}{l_2}\right) \\ \phi &= \gamma - \arctan\left(\tfrac{l_1}{l_2}\right) \end{aligned} \tag{2.9}$$

Ainsi, nous pouvons obtenir C_A et C_F sur une coordonnée céleste donnée à partir de simulations utilisant la méthode suivante :

1. Exécuter des simulations des balayages du satellite, en obtenant tous les angles de balayage ;
2. Calculer C_A ;

18. Comme nous l'avons vu dans l'introduction de ce Chapitre, la taille d'une fenêtre est une information qui dépend uniquement de la magnitude de la source détectée sur cette coordonnée et de la CCD qui réalise l'observation, et non pas de l'attitude du satellite.

3. Pour chaque balayage calculer les sommets du rectangle défini par la fenêtre observée suivant l'Équation 2.8 ;
4. Réaliser une union booléenne entre tous les rectangles ;
5. Calculer C_F ;

Nous avons développé un code à partir de la méthode décrite ci-dessus, qui a été mis à disposition des membres de Gaia qui travaillent sur la reconstruction d'images (CU5-DU18).

Dans les prochaines trois sous-sections, nous verrons la description de comment sont réalisées les simulations des balayages (pas 1) et l'opération d'union booléenne entre polygones (pas 3), et finalement l'analyse de la distribution dans le ciel en son entier de C_A et C_F à la fin de la mission.

2.4.1 Simulation des balayages

Tel que nous l'avons vu dans le Chapitre 1, les deux champs de vision de Gaia balaient le ciel en son entier en fonction des mouvements de rotation et de précession du satellite et de son orbite autour du Soleil. De cette manière, pour qu'il soit possible de déterminer combien de balayages seront réalisés sur une coordonnée céleste donnée, il est nécessaire de réaliser une simulation aussi bien des mouvements du satellite autour de lui même que de ses mouvements autour du Soleil.

La loi de balayage (ou *Scanning Law*), présentée dans Lindegren (1998, 2001), est une description des coordonnées célestes de l'axe de rotation du satellite en fonction du temps, en plus d'une description du propre mouvement de rotation. En coordonnées écliptiques, la position dans la sphère céleste qui indique les coordonnées (λ_z, β_z) vers lesquelles l'axe de rotation du satellite $\mathbf{z_s}$ (défini sur la Figure 2.14) pointe au moment t est décrite par les équations :

$$\begin{cases} \lambda_z(t) = \lambda_\odot + \arctan\left|\tan\xi\cos\nu(t)\right| \\ \beta_z(t) = \arcsin\left|\sin\xi\sin\nu(t)\right| \end{cases} \tag{2.10}$$

où $\lambda_\odot$ est la longitude écliptique du Soleil, ξ est le rayon du petit cercle défini par le mouvement de précession de l'axe $\mathbf{z_s}$ autour de $\mathbf{x_0}$ et $\nu(t)$ est la phase de révolution.

Cette phase est spécifiée de manière à ce que la vitesse de l'axe $\mathbf{z_s}$ dans le ciel ($|\mathrm{d}\mathbf{z}/\mathrm{d}\lambda_\odot| = S$) soit constante, générant ce que l'on nomme « la loi de balayage uniforme ».[19] De plus, le taux de rotation autour de l'axe $\mathbf{z_s}$ doit être constant, décrit par la phase de rotation $\Omega(t)$. Ces deux phases sont décrites par :

$$\begin{cases} \sin\xi\dfrac{\mathrm{d}\nu}{\mathrm{d}\lambda_\odot} = \sin\nu\cos\xi + \sqrt{S^2 - \cos\nu^2} \\ \dfrac{\mathrm{d}\Omega}{\mathrm{d}t} = \omega_z - \dfrac{\mathrm{d}\lambda_\odot}{\mathrm{d}t}\sin\nu\sin\xi - \dfrac{\mathrm{d}\nu}{\mathrm{d}t}\cos\xi \end{cases} \tag{2.11}$$

19. Tel que commenté dans Lindegren (1998), cette loi n'est pas strictement uniforme, étant donné qu'il existe une variation de $\sim 1,67\%$ en $\mathrm{d}\lambda_\odot/\mathrm{d}t$ en fonction de l'excentricité non nulle de l'orbite de la Terre.

avec ω_s le taux de rotation du satellite autour de l'axe $\mathbf{z_s}$ et d'autres grandeurs telles qu'indiquées sur la Figure 2.14.

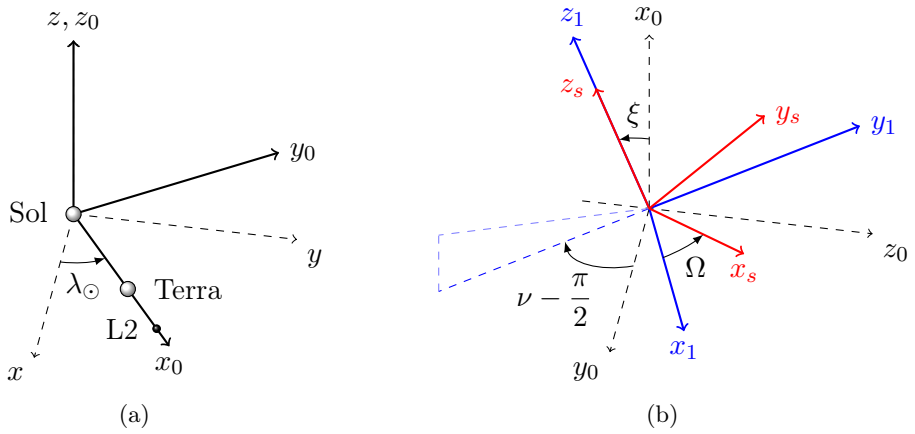


FIGURE 2.14 – Définition des systèmes de coordonnées et des angles utilisés pour la loi de balayage. En (a), $[x, y, z]$ indique un système inertiel centré dans le barycentre du Système Solaire, $[x_0, y_0, z_0]$ un système rotatif centré sur le Soleil avec l'axe x_0 pointant vers la Terre et $\lambda_\odot$ est la longitude écliptique apparente du Soleil (y compris aberration). En (b), ξ est l'angle de l'axe de rotation de Gaia, Ω est l'angle de la phase de rotation, ν est l'angle de la phase de révolution et les systèmes $[x_1, y_1, z_1]$ et $[x_s, y_s, z_s]$ sont centrés sur le satellite. (Adaptées de Luri, 2001; Astrium, 2006)

De cette manière, à partir de la détermination de $\lambda_\odot$ par des éphémérides, du choix de S, ξ et ω_z (pour optimisation de la couverture du ciel) et des conditions initiales $\nu(t = 0)$, $\Omega(t = 0)$, la loi de balayage est complètement déterminée, ne restant qu'à réaliser l'intégration des équations 2.11 pour la détermination de la position de l'axe de rotation dans le ciel (au moyen des équations 2.10).

Comme l'angle de base entre les deux miroirs de Gaia est connu (106.5°) et qu'il se trouve dans le plan des axes $\mathbf{x_s}$ et $\mathbf{y_s}$, nous pouvons calculer la position de l'axe de rotation pour la mission entière et déterminer combien de fois, et à quelle attitude, le satellite passera par un point quelconque de la sphère céleste. Pour obtenir les informations d'un point spécifique en coordonnées écliptiques, il suffit d'inverser les équations 2.10 et de déterminer les temps t durant lesquels le satellite observera cette position.

Nous avons utilisé la loi de balayage décrite dans Luri (2001) et mise en oeuvre dans la bibliothèque GaiaSimu 7 (Babusiaux et al, 2010) pour calculer les balayages dans le ciel en son entier. Etant donné que nous connaissons la quantité de ces balayages et les coordonnées écliptiques des champs de vision pour chacun de ces balayages, il suffit d'appliquer l'équation 2.8 pour la détermination des coordonnées des fenêtres observées. Nous verrons maintenant comment l'union des fenêtres observées pour le calcul de C_F a été réalisée.

2.4.2 Union de polygones

Bien qu'apparemment, il s'agisse de tâches insignifiantes, la réalisation d'opérations Booléennes entre des polygones génériques (non nécessairement convexes[20] ou monotones[21]) sont des processus d'un coût en temps de calcul élevé, réalisées par des bibliothèques spécialisées de géométrie computationnelle.

Tout d'abord, on peut penser que pour le calcul de C_F nous n'avons pas besoin de polygones génériques, étant donné que les fenêtres transmises par Gaia sont toujours des rectangles. Cependant, après l'union de deux fenêtres, on peut déjà avoir un polygone non monotone, par exemple.

Les algorithmes plus primitifs pour réaliser cette opération sont basés sur la rastérisation des polygones en *bitmap* binaires, et l'application postérieure d'opérations Booléennes entre les *bitmaps*. Bien que conceptuellement simple, cette technique présente plusieurs difficultés, spécialement liées à la résolution de la rastérisation et à la quantité de mémoire nécessaire pour l'obtention d'une bonne précision.

Les techniques plus modernes s'appuient, quand à elles, sur des constructions de géométrie computationnelle et sont principalement basées sur des algorithmes du type *plane* ou *line-sweep*. Dans ces algorithmes, une ligne (ou plan) imaginaire est déplacée dans l'espace rencontrant des régions « à gauche » ou « à droite » de chaque sommet ou bifurcations (pour une description détaillée, voir par exemple le Chapitre 2 dans De Berg et al, 2008 ou O'Rourke, 1998). Seulement récemment a été publié par Martinez et al (2009) un algorithme capable de connecter proprement des polygones quelconques.

Dans ce travail, nous avons utilisé la mise en oeuvre de l'union entre des polygones présente dans la bibliothèque *JTS – Java Topology Suite 1.8.1* (VividSolutions, 2006). La technique la plus directe serait une union itérative dans laquelle chaque fenêtre de chaque observation est unie avec la prochaine (avec un algorithme du type *plane-sweep*) et ainsi de suite jusqu'à la dernière observation disponible. Cependant cette méthode peut être inefficace pour des régions du ciel avec des dizaines ou des centaines de polygones. De cette manière, comme l'étude devrait être capable de couvrir le ciel en son entier, une solution plus optimisée a été choisie, et nous avons adopté le système d'union par buffer décrit dans Davis (2007).

Dans la terminologie de la *JTS*, et d'applications GIS[22] en général, l'opération de *buffer* représente le calcul de la région qui contient tous les points à une distance donnée (spécifiée par la taille du *buffer*) de la figure géométrique originale (VividSolutions, 2003). Cette opération est une somme de Minkowski[23] entre le polygone et un cercle de rayon égal à la distance requise.

20. Un polygone convexe est celui qui obéit à : toute angle interne est plus petit que 180° et tout segment de droite entre deux sommets est interne ou est un bord du polygone.

21. Un polygone est monotone en ce qui concerne la droite L, si toute orthogonale à L l'intersecte au maximum deux fois.

22. Geographic Information System

23. Pour deux sous-ensembles A et B d'un espace vectoriel, la somme de Minkowski de A avec B est définie comme étant $A + B = \{\mathbf{x} + \mathbf{y} | \mathbf{x} \in A, \mathbf{y} \in B\}$, ce qui est équivalent à une convolution entre les deux figures (Skiena, 2008).

Dans la définition ci-dessus on s'aperçoit qu'il est possible de réaliser une opération d'addition Booléenne de n'importe quels polygones entre eux en appliquant une opération de buffer de taille nulle sur cet ensemble. De cette manière, l'opération finale résultante est équivalente à une union entre les polygones, sans altération de leurs contours, et de forme très efficace (Davis, 2007).

2.4.3 Informatique dématérialisée

Pour la réalisation des études de couverture sur le ciel en son entier, il est nécessaire de modifier légèrement le code pour le paralléliser, sans quoi le temps d'exécution serait trop long pour des études sur une large échelle en n'utilisant qu'un seul processeur. Heureusement, le problème traité ici est hautement parallèle, et peut être divisé en divers nœuds de calcul qui travaillent de façon complètement indépendante, chaque nœud étant utilisé pour calculer une position distincte de la sphère céleste.

De plus, des paquets d'information avec diverses coordonnées peuvent être envoyés vers les nœuds, réduisant ainsi les exigences de communication et évitant tout embouteillage du réseau. De cette manière, nous pouvons adopter un modèle de calcul en *Grid* (Foster & Kesselman, 2004), qui est une forme de calcul distribué dans laquelle un processus est divisé entre plusieurs nœuds de calcul qui ne sont pas forcément connectés par une infrastructure de communication rapide et/ou stable, étant un paradigme qui cherche à utiliser Internet seulement pour stocker des informations, mais aussi pour les traiter. En plus, nous parlons ici d'informatique dématérialisée parce que les calculs ont été faites avec l'utilisation de machines virtuelles.

L'adoption de ce modèle permet de réaliser une même exécution du code en parallèle en utilisant des nœuds géographiquement dispersés, si besoin est, en profitant de l'infrastructure de calcul de divers laboratoires.[24] Comme infrastructure de logiciel pour notre *Grid*, la bibliothèque GridGain 2.1.1 (GridGain, 2010) a été utilisée. Elle réalise le parallélisme de tâches au moyen d'une mise en oeuvre de l'algorithme *Map-Reduce* (Dean & Ghemawat, 2004) et englobe déjà tous les outils nécessaires pour l'exécution dans des nœuds non homogènes (par *load-balancing*), transfert de code entre les nœuds, redondance de tâches, tolérance aux failles (au cas où un nœud ne serait plus disponible), découverte automatique de nouveaux nœuds et leur addition dans la topologie du *Grid* et dans la distribution des tâches, ce qui se montra bien utile, étant donné que les nœuds n'étaient pas seulement dédiés à l'exécution des simulations pour le calcul de C_A et C_F.

2.4.4 Résultats

La mise en oeuvre d'un code avec les caractéristiques décrites antérieurement permet de réaliser les analyses de C_A et C_F pour des coordonnées génériques dans la sphère céleste. L'exécution de ce code peut fournir une visualisation graphique des

24. Dans le cas de ce projet, aucune autre que celle disponible à l'IAG-USP n'a été nécessaire.

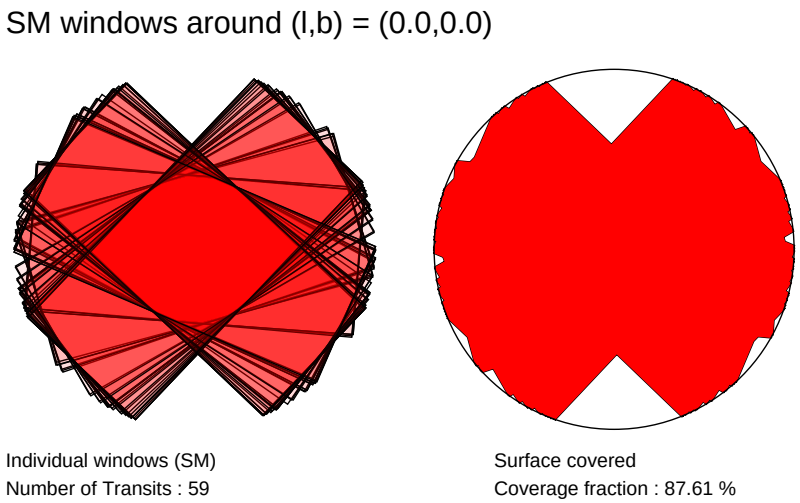


FIGURE 2.15 – Carte de couverture pour les coordonnées galactiques (l, b) = $(0.0°, 0.0°)$

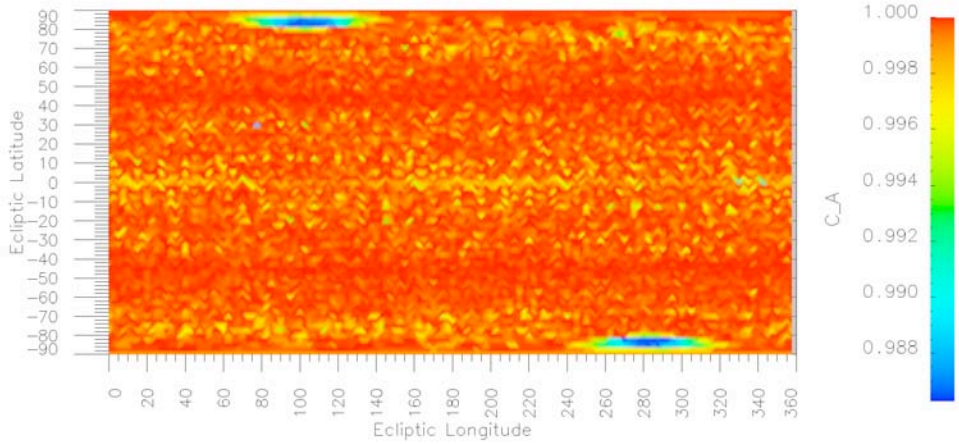
résultats pour une coordonnée unique ou des tableaux avec les valeurs calculées dans un réseau de coordonnées dans une région du ciel (ou dans le ciel en son entier).

Le calcul de la fraction de la surface couverte par les observations autour de chaque position dans le ciel est une information importante pour que l'on sache quel type d'information on pourra recouvrer à partir d'une image reconstruite. De plus, les deux valeurs C_A et C_F sont aussi extrêmement importantes pour la sélection des coordonnées utilisées durant les tests et les validations des codes et principalement, comme il sera vu dans le prochain Chapitre, dans la sélection de régions d'entraînement pour les algorithmes d'apprentissage numériques.

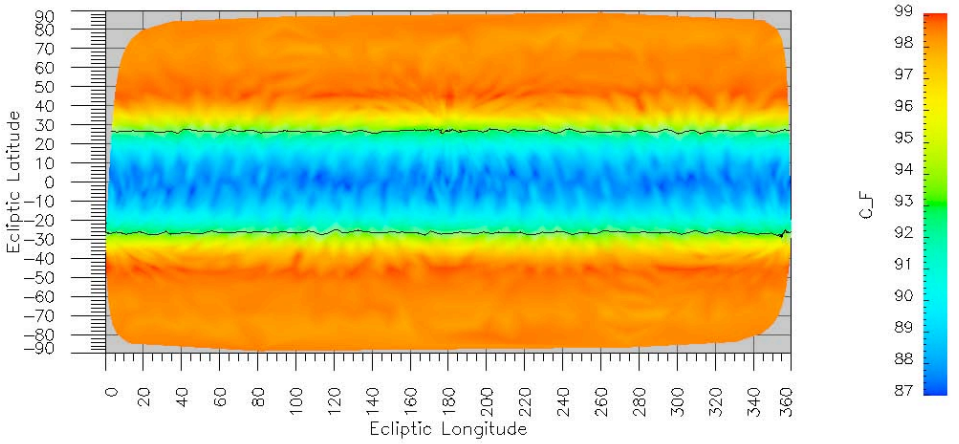
Pour réaliser cette sélection, il est nécessaire d'analyser la configuration géométrique des balayages puisque la distribution angulaire des observations a une forte influence sur les résultats des algorithmes de reconstruction d'images. La Figure 2.15 montre un exemple des informations obtenues pour une exécution aux coordonnées galactiques (l, b) = $(0.0°, 0.0°)$.

Dans le but de vérifier la structure sur une large échelle des paramètres C_A et C_F dans le ciel, nous avons réalisé une simulation pour le ciel en son entier en basse résolution de $150'$ et $300'$, respectivement, sur chacun des axes du système galactique. Durant l'analyse des résultats, nous nous sommes aperçus de l'existence d'une grande symétrie par rapport à l'écliptique, naturellement due à la loi de balayage du satellite. Les Figures 2.16 (a) et (b) présentent la distribution dans le ciel de C_A e C_F calculés.

Les simulations de C_A, Figure 2.16 (a), montrent que la couverture est importante (la valeur minimum obtenue pour C_A est de 0.986), et homogène (quasiment toutes les régions du ciel se trouve entre $C_A = 0.996$ et $CA = 1.000$). Ceci est un indicateur fort de ce qu'il sera possible de reconstruire des images avec des pixels carrés avec la

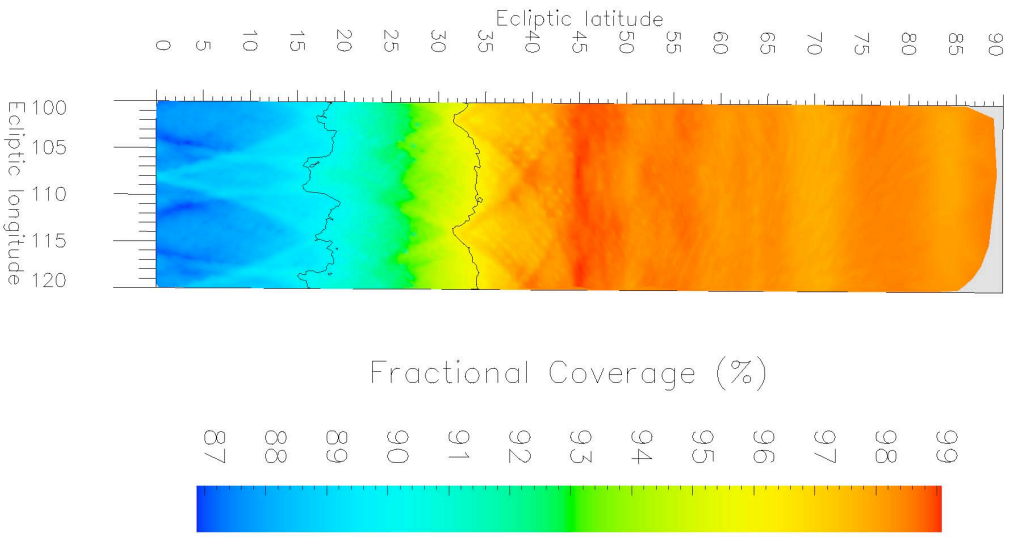


(a)

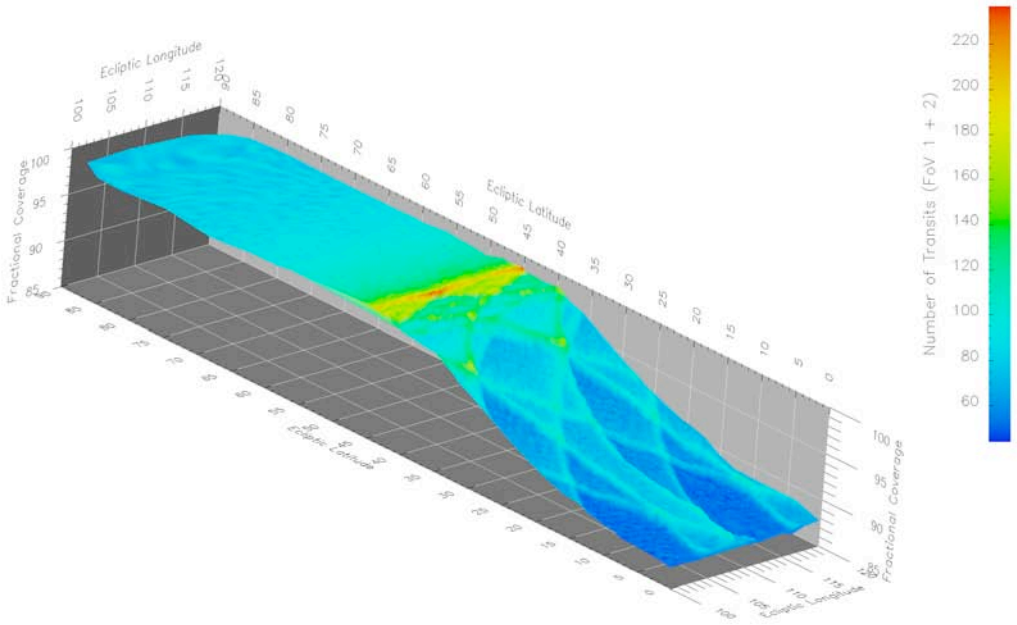


(b)

FIGURE 2.16 – Cartes de couvertures pour le ciel entier en coordonnées écliptiques. (a) Couverture Angulaire Normalisée – résolution du réseau régulier de 150′. (b) Couverture Fractionnelle – résolution du réseau régulier de 300′.



(a)



(b)

FIGURE 2.17 – (a) Carte de la Couverture Fractionnelle (C_F) en haute résolution ($25'$) pour une bande du ciel (le manque de données proche de $\beta \sim 90°$ est dû au fait que la simulation a été réalisée sur une grille de coordonnées galactiques). (b) Corrélation entre position dans le ciel, couverture fractionnelle (C_F) et nombre de balayages.

meilleure résolution angulaire possible pour presque toutes les source observées. La résolution finale ne sera alors limitée que par l'algorithme de reconstruction choisi.

La Figure 2.16 (b) montre les résultats pour les simulations de C_F. Ces simulations indiquent que la couverture fractionnelle est hautement symétrique dans le ciel par rapport au plan de l'écliptique, et qu'un pattern se répète avec deux comportements distincts : une région de couverture croissante entre $|\beta| = 0°$ et $|\beta| \sim 40°$ et une région avec C_F pratiquement stable au-dessus de 98% pour $|\beta| > 40°$. De plus, cette Figure indique que bien que la distribution de C_F soit apparemment homogène sur une large échelle, il est possible de noter l'existence de variations régulières sur une plus petite échelle.

De telles variations sur une plus petite échelle apparaissent plus clairement dans une simulation avec une plus grande résolution sur le réseau (de $25'$), dont le résultat peut être observé sur la Figure 2.17 (a). Nous pouvons remarquer l'existence d'un pattern tous les $\sim 13°$ de longitude, ce qui est très similaire à ce qui est observé dans la structure du nombre de balayages pour chaque région du ciel (Chapitre 1). Cependant, sur la Figure 2.17 (b), on observe qu'il n'existe pas une corrélation directe entre le nombre de balayages et l'indice C_F, en particulier des régions avec un nombre faible de balayages peuvent également avoir un C_F élevé.

Même dans des régions où le nombre de balayages n'est pas plus élevé que le nombre moyen pour les observations de Gaia, la valeur obtenue pour C_F se trouve au-dessus de 98% (ce qui est un fort indicateur que seulement certaines positions des bords du cercle représenté sur la Figure 2.15 n'ont pas été recouvertes). Il faut remarquer, cependant, que les régions avec le nombre minimum de balayages (un peu moins de 60), possèdent un C_F bas, indiquant l'absence de données pour la reconstruction de certaines régions de l'image – une de ces régions avec une valeur de C_F basse, a été présentée comme exemple sur la Figure 2.15.

Cependant, comme on peut le remarquer sur la Figure 2.15, même dans le cas de couverture $\sim 87\%$, la région dans laquelle l'information est perdue n'est pas très grande, de manière à ce que les valeurs de C_F obtenus pour le ciel entier (toujours supérieures à 87%), permettent de conclure qu'il existent des données suffisantes pour des reconstructions d'images qui permettent une analyse de pratiquement n'importe quel type d'émission qui surviendrait dans les alentours de tous les objets observés.

2.5 Conclusions

Dans cette section, nous avons vu au moyen d'un *toy model* les principes de reconstruction d'image tels qu'ils sont compris dans le contexte de Gaia. Nous avons présenté la transformée de Radon, qui peut être utilisée comme description mathématiques pour les fenêtres provenant des observations réalisées par le satellite, en plus d'introduire les idées générales de comment s'inverse cette transformation, et de comment cette inversion est sensible à la quantité des données disponibles et au bruit dans la détermination du signal.

Nous nous sommes penchés sur les difficultés pour réaliser la reconstruction d'images dans le contexte de la réduction de données de Gaia, et nous avons présenté

les algorithmes proposés jusqu'à aujourd'hui dans la littérature. Nous avons vu que les algorithmes basés sur la régularisation de Tikhonov et l'algorithme Clean/Cleanest, bien qu'ils permettent une reconstruction d'images avec une plus grande résolution, possèdent un coût en temps de calcul dissuasif pour leur utilisation sur une large échelle, et qu'à priori leur utilisation durant la mission est abandonnée. Nous avons vu que, qualitativement, le meilleur candidat est BinOutliers.

Nous avons présenté une étude que nous avons réalisée dans le cadre de ce doctorat sur la couverture spatiale des balayages, appelée C_F ou couverture fractionnelle, dont le résultat a démontré que pour n'importe quelle position du ciel, à la fin de la mission il y aura une couverture fractionnelle minimum de $\sim 87\%$, allant jusqu'à $\sim 99\%$ dans certaines régions, et étant supérieure à 98% sur une grande partie du ciel ($|\beta| > 45°$). Cette couverture fractionnelle élevée permettra que les informations de n'importe quelle émission qui surviendra autour d'une source observée soit analysée, même dans le cas de couverture fractionnelle minimum, avec $C_F \sim 87\%$.

Finalement, notre étude a aussi démontré que la distribution relative entre les angles de balayage, ici appelée C_A, ou couverture angulaire normalisée, pour une région du ciel quelconque possède une valeur pratiquement égale à 1, ce qui signifie qu'au moins deux balayages se retrouvent à $\sim 90°$ entre eux. Ceci signifie que pour une coordonnée quelconque de la sphère céleste il existe la possibilité théorique que la reconstruction puisse être réalisée avec des pixels aussi petits que la taille du pixel dans la direction de plus grande résolution, la résolution finale obtenue ne dépendant seulement que de l'algorithme adopté et non pas de l'absence d'information.

Chapitre 3

Sources Secondaires

« *Mais ce qui est vrai des maux de ce monde est vrai aussi de la peste. Cela peut servir à grandir quelques-uns.* » A. Camus[1]

Sommaire

Pour un instrument dont la magnitude limite est celle de Gaia, la quantité de sources de rayonnement dans le ciel est suffisamment grande pour que la présence de plusieurs objets sur la même ligne de visée ne soit pas négligée. De plus, divers objets de la Galaxie se révèlent être des systèmes multiples s'ils sont observés à de plus hauts niveaux de résolution angulaire.

Dans ce chapitre, nous présenterons des études quantitatives de ce phénomène (Krone-Martins et al, 2008a) qui perturbera le traitement astrométrique de diverses sources. Nous nous étendrons aussi sur la stratégie que nous avons développée, et qui sera utilisée durant le traitement des données de Gaia, pour récupérer les sources secondaires (Krone-Martins et al, 2010a). Ces sources sont intéressantes, car elles peuvent être au-delà de la magnitude limite du satellite, et il sera alors possible de compléter le catalogue Gaia.

1. La Peste, d'Albert Camus.

3.1 Introduction

Les structures étendues d'objets astronomiques conduiront à perturber les solutions astrométriques produites par le consortium de traitement de données de Gaia (Brown, 2007). Ces structures, définies ici comme un signal plus grand que la PSF, peuvent être générées par des raisons intrinsèques à l'objet (une galaxie, par exemple), ou extrinsèques, comme dans le cas de confusion entre des objets qui se trouvent sur la même ligne de visée. C'est sur cette deuxième définition, que nous appellerons ici « projections optiques », que nous nous pencherons dans ce chapitre.

Dans les sections 3.2 et 3.3 nous présenterons des analyses semi-numériques des résultats concernant la quantité de projections optiques rencontrées par Gaia. De telles analyses permettent d'estimer le volume de calcul nécessaire pour reconstruire les cas problématiques. Cependant, durant ces études nous ne considérerons pas seulement la quantité de sources secondaires qui pourront interférer directement avec les mesures astrométriques (ou sources avec $G \leq 20$), mais encore la quantité totale de ces projections, qui incluent des sources qui ne peuvent être observées qu'à partir d'images reconstruites ($G > 20$).

Comme nous l'avons vu dans le Chapitre 2, en utilisant des données des fenêtres transférées par Gaia, il est possible de reconstruire des cartes bidimensionnelles (ou images reconstruites) de l'objet et de ses abords. De telles images peuvent révéler les structures perturbatrices, que celles-ci soient des sources secondaires de magnitude quelconque ou des structures étendues intrinsèques à l'objet, étant entendu que le traitement de ces images reconstruites pour Gaia doit être fait de la manière suivante (Brown, 2007) :

1. Les statistiques de l'ajustement de la Point Spread Function / Line Spread Function (PSF/LSF) et la solution astrométrique seront examinées pour décider si une image doit être reconstruite ; dans ce cas , une carte bidimensionnelle de premier ordre sera reconstruite en utilisant probablement certains des algorithmes vus dans le Chapitre 2 (*QuickStack*, *Drizzle*, *ShuffleStack*, *BinOutliers*) ;
2. Cette carte sera analysée, en cherchant à identifier les sources astronomiques et leurs caractéristiques, en plus du fond de ciel ; Pour chaque source rencontrée, une tentative de détermination du flux et des cinq paramètres astrométriques (α, δ, μ_α, μ_δ e ϖ) sera faite en supposant que l'objet est ponctuel ;
3. Au cas où la méthode échouerait, un traitement plus sophistiqué devra être réalisé, dans lequel la nature physique de chaque source sera préalablement établie, de manière à adopter des méthodes et des profils adaptés à la nature de l'objet étudié (comme il sera vu pour les galaxies dans les prochains chapitres).

Dans la Section 3.4, nous présentons une description de la méthode d'analyse d'images (*Educated Image Segregation*) développée, et qui sera adoptée durant la réduction des données de Gaia (dans le CU5-DU18) pour la caractérisation de l'image reconstruite en termes des sources existantes, résolvant une partie de l'étape 2 ci-dessus.

3.2 Analyse semi-numérique – Catalogues

3.2.1 Procédure d'Analyse

Pour estimer le nombre de projections optiques à partir de catalogues connus d'objets astronomiques, deux options sont possibles. La plus directe est de chercher dans les catalogues le nombre de cas pour lesquels les limites de distance angulaire et les différences de magnitudes entre deux sources s'appliquent. Malheureusement cette option n'est pas seulement hautement coûteuse en temps de calcul, mais il est également difficile de disposer de catalogues complets jusqu'à la magnitude désirée (dans cette étude, $V \approx 22$) et avec la résolution angulaire recherchée (~ 0.1").

Une autre possibilité, et ce sera celle développée ici, est d'appliquer une méthode basée sur une estimation de densité d'objets sur des intervalles de magnitude a partir d'un catalogue dans des directions données du ciel. Pour cela, nous avons pour le moment, assumé une distribution homogène de sources dans le ciel et nous avons extrapolé le nombre de cas de projections optiques dérivé dans des directions-test à toute la sphère céleste.

3.2.1.1 Méthode

Le satellite Gaia va observer $N_G(m)$ sources de magnitude m sur la sphère céleste. Une source secondaire peut perturber la solution astrométrique et les ajustements de PSF/LSF, si elle se trouve à une distance$d \leq 2.53$" de la source principale (Brown, 2007) et si sa magnitude m_{sec} est dans l'intervalle $m < m_{sec} \leq (m + \Delta m)$, où Δm est encore une valeur à définir – dans cette section nous étudierons un cas extrême où $\Delta m = 5$ et un cas plus réaliste, avec $\Delta m = 3$.

Le nombre total d'objets secondaires autour de toutes les sources primaires de Gaia (pour une magnitude donné), peut être obtenu en supposant que la distribution des étoiles est homogène sur le ciel.

C'est intuitif considérer que le nombre de projections optiques N_{prob} que Gaia pourra rencontrer est limité supérieurement par le nombre de sources primaires, si les secondaires existent en quantité suffisante pour compenser les petites distances dans lesquelles elles doivent se rencontrer des primaires pour créer ce phénomène. Dans le cas contraire, nous pouvons calculer la quantité de cas problématiques pour des objets primaires de magnitude m par :

$$N_{prob}(m) \approx A_{esf}\rho(m) \int_{m+\delta m}^{m+\Delta m} \rho(m')A_G dm' \tag{3.1}$$

où A_{esf} est la surface de la sphère céleste, $\rho(m)$ est la densité surfacique d'étoiles de magnitude m et A_G est la surface (en degrés carrés) autour de chaque primaire pour laquelle une source secondaire peut être rencontrée. Cette relation se base sur l'hypothèse que la région du catalogue qui a été utilisé pour l'estimation de $\rho(m)$ est représentative de la densité d'étoiles dans toute la sphère.

Pour évaluer le nombre de projections optiques nous avons analysé ici le catalogue

GSC 2.3.2 (Lasker et al, 2008) en utilisant des données dans le filtre $V_{GSC2.3.2}$[2] pour les objets stellaires et non-stellaires.

Dans cette analyse nous n'avons pas séparé les objets stellaires des non stellaires, étant donné que les deux perturberont l'ajustement de la PSF/LSF et en conséquence la solution astrométrique de l'objet primaire.

Les données du catalogue n'ont été sélectionnées que pour quelques régions du ciel, étant entendu que dans chaque région la procédure adoptée pour l'obtention du nombre de projections optiques a été de :

1. Compter le nombre d'étoiles dans divers intervalles de magnitude ;
2. Analyser ces comptages et déterminer la magnitude limite du catalogue dans la région considérée (interruption de la loi exponentielle) ;
3. Pour extrapoler le catalogue aux grandes magnitudes, réaliser des ajustements linéaires de la loi exponentielle avec toutes les données et pour 100 sous-ensembles de ces données dans l'espace log(comptages) vs. $V_{GSC2.3.2}$[3] et obtenir les fonctions $\rho(m)$;
4. En utilisant la valeur moyenne obtenue à partir des lois $\rho(m)$ ajustées, extrapoler les comptages des données entre la magnitude limite du catalogue et $V_{GSC2.3.2} = 22$ de façon a obtenir $\rho(m)$;
5. Finalement, estimer pour chaque magnitude le nombre de projections optiques avec l'equation 3.1.

Il est important de remarquer que pour réaliser tous les calculs numériques de ce travail, nous avons supposé que le catalogue Gaia sera complet jusqu'à la magnitude $V_{GSC2.3.2} = 20$ pour les cibles principales, et que nous serons capables de détecter des objets jusqu'à la magnitude $V_{GSC2.3.2} = 22$ dans les images reconstruites.

3.2.1.2 Données

En utilisant les données du catalogue GSC v. 2.3.2, 10 régions circulaires (de rayon $r = 1°$) centrées sur des positions correspondant à des densités d'objets distinctes dans le filtre $V_{GSC2.3.2}$ ont été analysées.

La densité a une valeur minimale pour $(l; b) = (0°, 30°)$, où seulement 32 obj/dg^2 ont été rencontrés ; ceci est probablement lié à une très grande confusion (impossibilité de séparer les objets trop proches) lors de la construction du catalogue. La densité est maximale pour $(l; b) = (60°, 0°)$, avec 27779 obj/dg^2, probablement dû au fait que les objets ont été mieux séparés que dans la région centrale. Cette valeur est celle que nous pouvons espérer comme limite inférieure pour la densité dans le bulbe.

Comme notre analyse demande que les cibles primaires aient une magnitude limite $V_{GSC2.3.2} = 20$ tandis que les sources secondaires peuvent avoir une magnitude

2. Il est important de remarquer que le filtre V du GSC 2.3.2 n'est pas le filtre V de Johnson, mais le filtre V photographique. C'est la raison pour laquelle nous n'avons pas transformé ces magnitudes dans le système G de Gaia au moyen des formules publiées dans Jordi (2007). Cependant la magnitude donnée dans ce filtre est probablement assez proche de G.

3. Cette technique est un type de simulation de *bootstrap*.

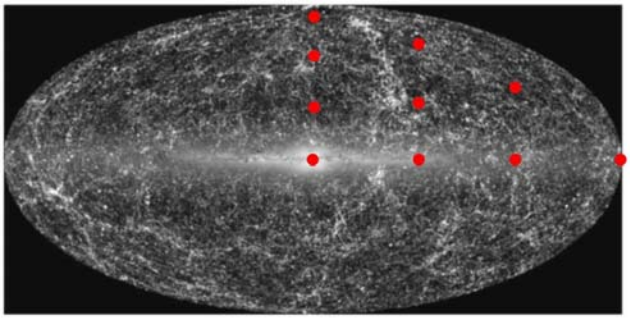


FIGURE 3.1 – Coordonnées galactiques des régions analysées. Les cercles rouges indiquent les positions approximatives d'étude du catalogue (adapté de Jarrett, 2004)).

l (°) ;b (°)	0 ;0	60 ;0	120 ;0	180 ;0	0 ;30	60 ;30	120 ;30
Mag. limite	17	18	18	18	11	19	18
Densidade	20666.6	27779.2	23787.9	14787.1	31.5	4626.9	2591.7

l (°) ;b (°)	0 ;60	60 ;60	0 ;85
Mag. limite	19	18	18
Densidade	2759.4	2269.2	1719.5

TABLE 3.1 – Magnitudes limite (telles que définies dans la section 3.2.1.1) et densités d'objets (en obj/dg^2) des régions analysées du catalogue GSC 2.3.2 dans le filtre $V_{GSC2.3.2}$.

limite $V_{GSC2.3.2} = 22$, nous avons extrapolé les comptages. Les comptages artificiels pour les magnitudes supérieures à la limite du catalogue, ainsi que les données du catalogues sont représentés sur la Figure 3.2 pour toutes les régions analysées.

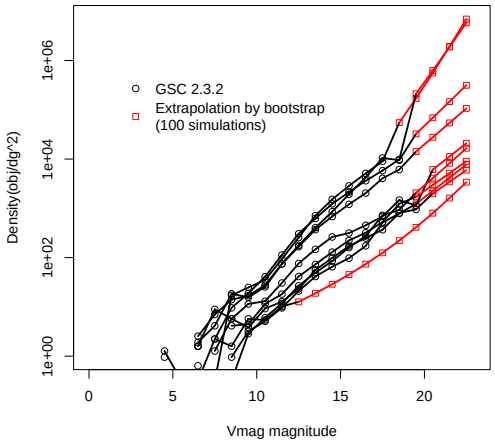


FIGURE 3.2 – Densités d'objets obtenues à partir de comptages du GSC 2.3.2 (en noir) et des lois d'extrapolation (en rouge).

Dans le cas particulier de la région $(l; b) = (0°, 30°)$, la plus grande partie des données est obtenue avec la loi d'extrapolation. Il est important de remarquer que même s'il s'agit d'une estimation peu fiable de la densité réelle de cette région (devant être considérée comme étant une limite inférieure).

3.2.2 Résultats

Après avoir traité chaque région au moyen de la procédure décrite antérieurement, nous avons obtenu le nombre de projections optiques sur l'ensemble du ciel, pour chaque intervalle de magnitude d'une source primaire. Ce nombre a été estimé à partir des données obtenues dans chaque région analysée, si bien qu'il varie beaucoup entre les régions en fonction de leurs densités. Le calcul a été réalisé aussi bien dans le cas où la différence maximale de magnitude entre primaire et secondaire est de cinq ou de trois.

La Figure 3.3 montre les estimations de projections optiques dans les 10 régions considérées en fonction de la magnitude de la primaire dans les cas $\Delta m = 3$ et $\Delta m = 5$. Ces résultats indiquent que le nombre de projections optiques que l'on peut espérer durant la mission doit être de l'ordre de quelques millions (même en considérant le cas le plus restrictif de $\Delta m = 3$), pouvant atteindre des centaines de millions.

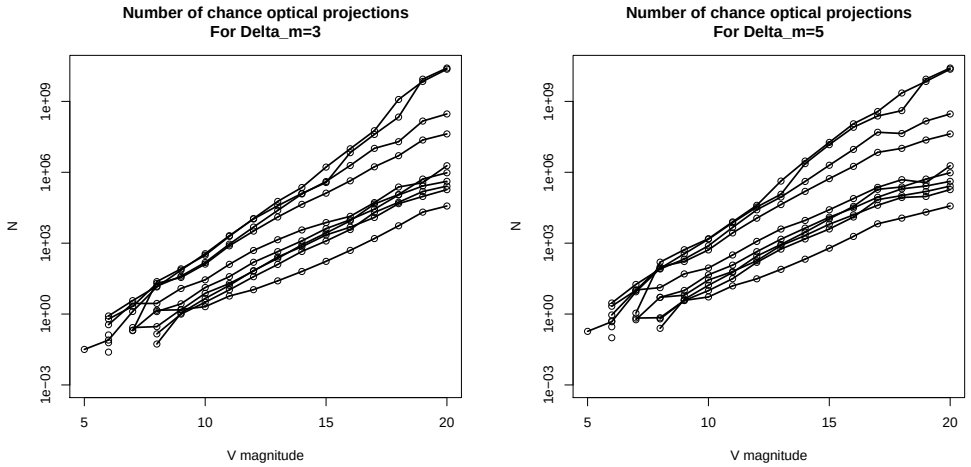


FIGURE 3.3 – Extrapolation pour le ciel entier du nombre de projections optiques si nous observons une source Gaia de magnitude V pour les 10 régions analysées. Le graphique de gauche représente le comportement en supposant $\Delta m = 5$ et celui de droite $\Delta m = 3$.

Nous pouvons aussi déterminer une probabilité de rencontrer une projection optique. Celle-ci peut être calculée simplement en prenant le rapport entre le nombre de projections optiques et le nombre de sources primaires. Les résultats obtenus pour ce calcul sont présentés sur la Figure 3.4. Une étude similaire a été réalisée avec le catalogue USNO-B1.0, et les conclusions obtenues ont été identiques.

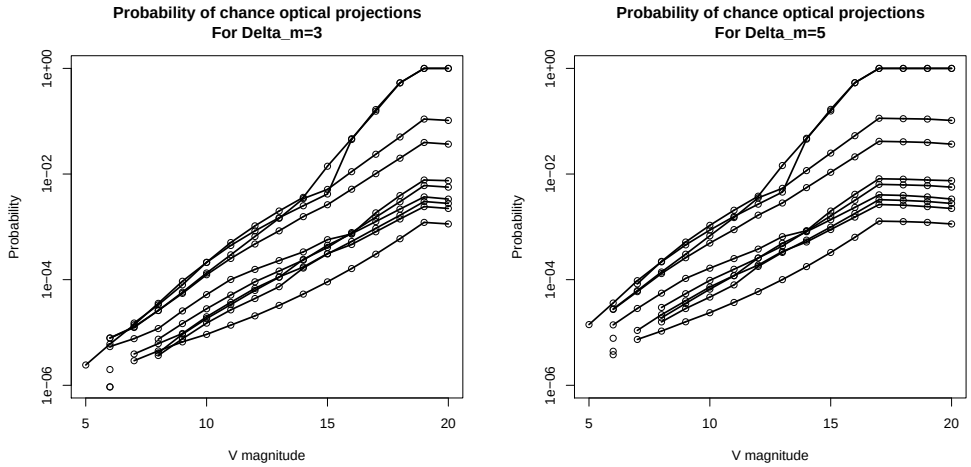


FIGURE 3.4 – Probabilité de rencontrer une projection optique en observant une source primaire de magnitude V pour toutes les régions analysées. Comme dans la dernière figure, à gauche sont représentés les résultats pour $\Delta m = 5$ et à droite pour $\Delta m = 3$.

3.2.3 Commentaires

Le nombre de projections optiques varie beaucoup en fonction de la région du ciel où l'analyse a été réalisée. Néanmoins, en prenant la moyenne de toutes les régions analysées, nous avons une première évaluation du nombre intermédiaire de cas qui doivent survenir durant la construction du catalogue Gaia.

Avec ce calcul, on estime que ce nombre est de l'ordre de $N \approx 2 \times 10^6$, pour la différence de magnitude la plus réaliste , $\Delta m = 3$. La plus petite valeur obtenue a été de $N \approx 7 \times 10^4$, alors que dans le cas d'une Galaxie uniquement constituée d'un bulbe, cette valeur est $N \approx 3 \times 10^{10}$.

Nous notons que, bien que ces chiffres et la Figure 3.3 nous donnent une impression qualitative de ce que l'on peut espérer pour tout le ciel en termes du nombre de projections optiques, les résultats obtenus ici ne limitent pas de manière très précise le nombre de projections optiques espéré pour tout le ciel. Ceci est dû à trois facteurs principaux : manque de réalisme dans l'extrapolation des catalogues, petit nombre de régions échantillonnées et manque d'une intégration interpolant le ciel entier. Pour résoudre de tels problèmes, nous avons résolu d'adopter une analyse un peu plus complexe basée sur des simulations, qui sera décrite dans la prochaine section.

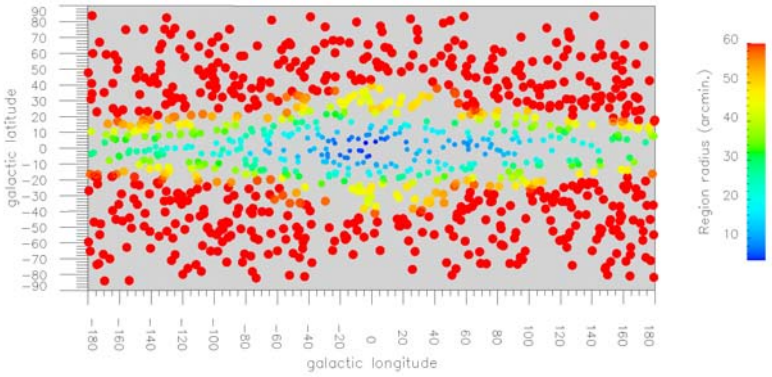
3.3 Analyse semi-numérique – Simulations

Dans le but de mieux contraindre le nombre de projections optiques qui pourront être rencontrées par Gaia, nous avons réalisé une estimation basé sur des données issues de simulations réalisées avec des routines de la bibliothèque officielle de simulations de Gaia, la GaiaSimu.

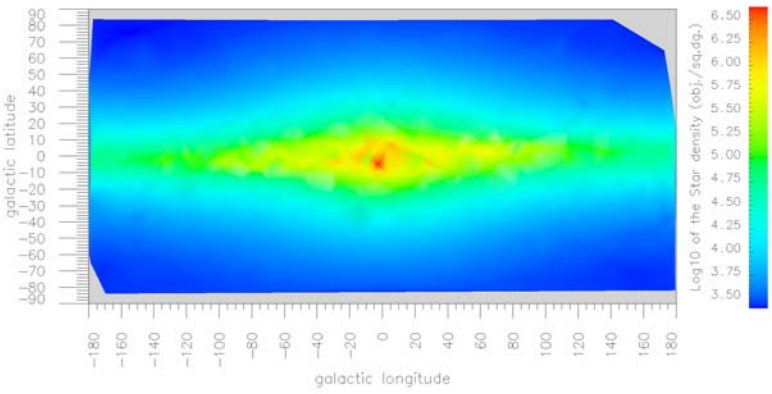
3.3.1 Procédure d'analyse

Nous avons simulé des catalogues d'étoiles complets jusqu'à la magnitude $G = 22$ dans mille régions de la sphère céleste sélectionnées de manière aléatoire avec une distribution uniforme sur la sphère. Ces simulations ont été réalisées en utilisant le modèle de galaxie de Besançon (Robin et al, 2007), tel qu'il est implanté dans le modèle d'univers de Gaia dans la bibliothèque GaiaSimu v.2.0 (Babusiaux et al, 2007) (on a utilisé la classe `BesanconStarGenerator`).

Les régions simulées sont circulaires avec un rayon variable, inférieur ou égal à 1° ; de cette manière nous avons donc une distribution spatiale homogène des sources dans ces régions. Le modèle de galaxie adopté inclut le modèle d'extinction de Drimmel (2002) et la binarité des étoiles d'Arenou (2003). Un exemple d'une telle simulation, y compris une carte du ciel entier avec des densités interpolées à partir des régions simulées est illustré sur la Figure 3.5.



(a)



(b)

FIGURE 3.5 – Sur (a) un exemple des régions simulées pour l'une des simulations réalisée (A1). Sur cette figure la taille et les couleurs des cercles indiquent les surfaces de chaque région simulée. Sur (b), diagramme interpolé à partir des régions simulées représentant des densités d'objets dans la sphère céleste.

Pour chacune de ces régions, la densité d'étoiles par intervalle de 0.1 magnitude a été calculée pour l'intervalle [10.0 ; 22.1[. Pour de telles densités, nous avons calculé le nombre N_{prob} de projections optiques par degré carré dans chaque région et dans chaque intervalle de magnitude en utilisant le relation :

$$N_{prob}(m) \approx \rho(m) \int_{m+\delta m}^{m+\Delta m} \rho(m') A_G dm' \tag{3.2}$$

où les variables et fonctions sont comme celles de l'equation 3.1, mais ici on ne considère pas une distribution homogène de sources dans la sphère céleste.

Nous avons découpé la sphère céleste en un certain nombre de triangles et nous avons interpolé les distributions obtenues au centre de ces triangles (dits HTM[4]). Durant le processus d'interpolation, le nombre de projections optiques par unité de surface et par intervalle de magnitude a été pondéré par la distance sur la sphère entre le centre du triangle HTM et les trois régions simulées les plus proches.

Finalement, nous avons intégré le nombre de cas dans tous les triangles HTM de toute la sphère de manière à estimer le nombre de cas qui pourront être rencontrés dans les données de Gaia pour tous les bins de magnitude. Ceci a été réalisé séparément pour la région du bulbe ($l < 15°$ ou $l > 345°$ et $-15° < b < 15°$) et pour le disque+halo, étant donné que les nombres de projections y sont très différentes. Pour estimer les erreurs associées à la sélection aléatoire des régions simulées, la procédure a été répétée 10 fois.

3.3.2 Résultats

Après avoir réalisé les 1000 simulations et calculé la moyenne et l'écart type des valeurs obtenues, nous avons conclu qu'un nombre total de $1.6 \times 10^8 \pm 5.1 \times 10^7$ projections optiques pourra être observé durant la mission. Cependant, nous devons remarquer que l'estimation diminue significativement si l'on ne considère que les régions de disque+halo de la galaxie. Le Tableau 3.2montre les résultats pour chacun des 10 tours indépendants de simulation+analyse.

	A1	A2	A3	A4	A5
n_{prob} disco+halo	1.6×10^6	1.8×10^6	1.5×10^6	1.7×10^6	1.5×10^6
n_{prob} bojo	1.6×10^8	1.1×10^8	2.7×10^8	1.1×10^8	1.8×10^8
	A6	A7	A8	A9	A10
n_{prob} disco+halo	1.5×10^6	1.5×10^6	1.6×10^6	1.4×10^6	1.8×10^6
n_{prob} bojo	1.4×10^8	1.9×10^8	9.0×10^8	1.7×10^8	1.6×10^8

TABLE 3.2 – Nombre de projections optiques pour les régions du bulbe et du halo+disque. Chaque colonne An représente une simulation+analyse indépendant.

Il est possible que le nombre de projections optiques évalué ici ne soit pas atteint au cours de la mission, car dans le bulbe le grand nombre d'objets peut saturer le

4. Plus de détails sur ces triangles peuvent être rencontrés dans la section 4.2.1.

système qui crée les fenêtres envoyées à la Terre.

Comme nous connaissons le nombre de projections optiques pour une magnitude donnée de la source primaire, ainsi que le nombre de sources, nous pouvons dériver les courbes de probabilité des projections optiques. En prenant en considération 10 simulations indépendantes et en calculant la courbe moyenne et une enveloppe de 1σ, nous obtenons le graphique présenté à droite de la Figure 3.6.

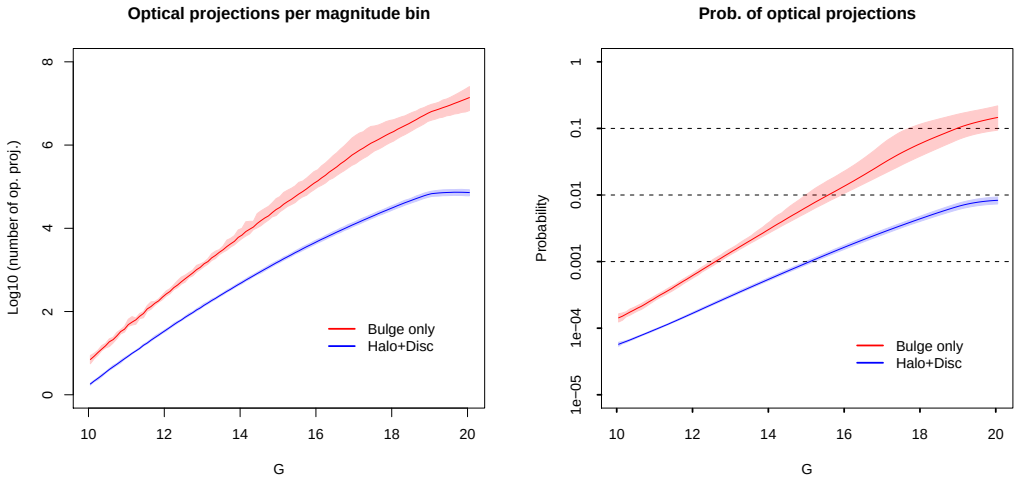


FIGURE 3.6 – À gauche, nombre de projections optiques. À droite, probabilité d'obtenir une projection optique dans un rayon de 2.3" autour d'une source primaire de magnitude G – les niveaux de probabilité de 0.1%, 1% et 10% sont indiqués.

Résolution des divergences avec Brown, 2008

Dans une autre estimation de projections optiques, basée sur une distribution homogène de sources sur tout le ciel, Brown (2008) a obtenu des probabilités beaucoup plus élevées que celles dérivées dans nos simulations. Les deux travaux ont été examinés, et nous avons conclu que les différences sont principalement dues à des pentes des fonctions de comptage d'étoiles, et à l'interruption de la loi exponentielle à partir de $G \sim 20$ (non considérée dans Brown, 2008).

Les pentes que nous avons obtenus dans nos simulations pour les relations $log(N)$ *vs* G sont plus petites en moyenne que celles obtenus à partir du GUMS[5], comme on peut le voir sur les graphiques gauche et droit de la Figure 3.7. A partir des simulations présentées dans la section antérieure, nous avons obtenu une valeur médian de 0.16, tandis que la pente utilisée dans Brown (2008) est de 0.23. La valeur

5. Le GUMS, ou *Gaia Universe Model Snapshot* est une simulation de toutes les sources présentes dans le ciel jusqu'à une certaine limite de magnitude. Pour produire cet ensemble de données, on a utilisé le modèle d'univers de Gaia tel que disponible en mi-2006.

inférieure obtenue dans nos simulations peut être due à l'incomplétude de notre stratégie d'échantillonnage aléatoire des régions simulées.

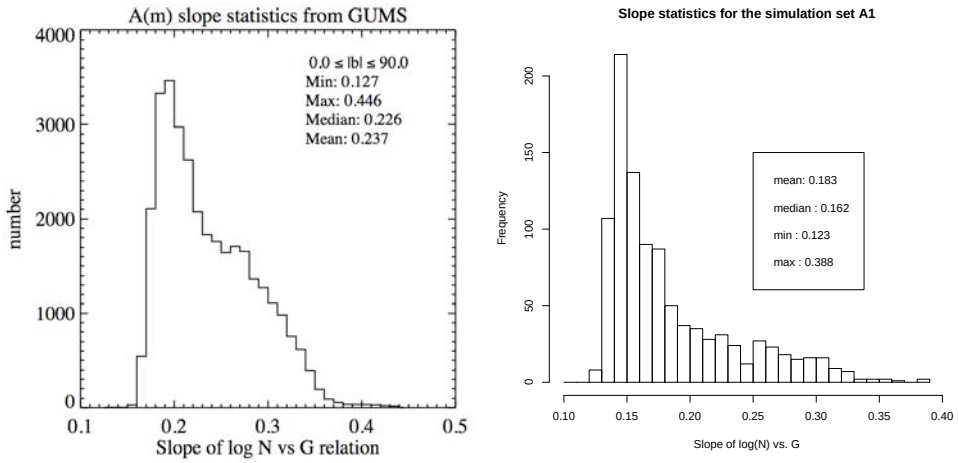


FIGURE 3.7 – À gauche un histogramme représentant les pentes des fonctions $log(N)$ *vs* G dérivées à partir du GUMS. À droite, les pentes dérivées à partir des 1000 régions circulaires sélectionnées aléatoirement dans le ciel et simulées avec la bibliothèque GaiaSimu v.2.0.

Une autre différence entre les deux travaux réside dans le fait que le nombre de sources diminue au delà de la magnitude $G \sim 20$, et la fonction de comptage d'étoiles commence à dévier significativement d'une droite (adoptée dans Brown, 2008). Ce comportement peut être constaté sur le graphique de gauche de la Figure 3.8, qui montre le nombre d'étoiles par intervalle de 0.1 magnitude et un ajustement robuste de droite sur les données (utilisant l'algorithme de Tukey, 1977).

Cette altération du comportement au-dessus de $G \sim 20$ mène à une surestimation du nombre de projections optiques si la croissance en $log(N)$ *vs* G est considérée linéaire. Cependant, nous pouvons obtenir le facteur de surestimation en réalisant un deuxième ajustement de droite n'utilisant que des données avec $G > 20$, et en calculant le rapport entre les intégrales des deux ajustements dans l'intervalle [20-23][6] ; en ignorant ce changement de comportement, l'estimation est en moyenne 1.4 fois la valeur réelle (voir Figure 3.8). Ce problème est particulièrement amplifié par le fait que dans Brown (2008), une fonction de comptage d'étoiles unique est extrapolée pour le ciel entier.

Un autre problème est le nombre d'étoiles adopté jusqu'à $G = 20$ – nécessaire pour réaliser l'estimation en utilisant la méthodologie présentée dans Brown (2008). Dans ce travail, à l'origine, une valeur de l'ordre de 10^{4-5}, a été adoptée, supérieure à que ce que l'on trouve dans la majorité des directions du ciel ($\sim 10^{3-4}$).

Après avoir pris les facteurs de correction en compte, spécialement l'altération du comportement de la fonction de comptage d'étoiles et en choisissant les régions de

6. Ici, nous avons utilisé G=23 comme limite supérieure, étant donné celle-ci a été la magnitude limite utilisée pour des sources secondaires par Brown (2008)

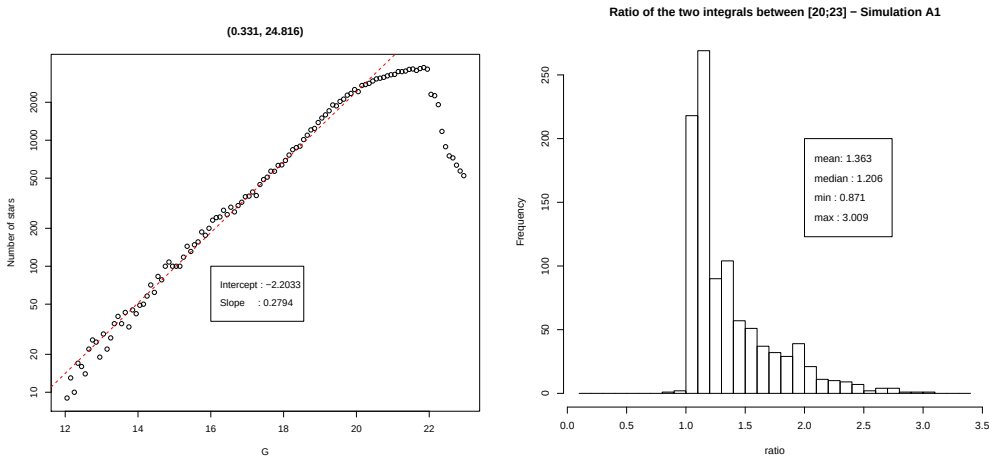


FIGURE 3.8 – Dans le graphique de gauche, sont présentés des comptages d'étoiles en bins de 0.1 magnitude pour l'une des régions simulées dans la section antérieure, ainsi qu'un ajustement robuste de droite. Sur le graphique de droite est présenté l'histogramme du taux de surestimation réalisée dans la fonction de comptage d'étoiles si un ajustement de droite simple est utilisée.

calcul de manière à échantillonner des niveaux différents de latitudes galactiques, les probabilités obtenues dans les deux travaux sont en accord, comme cela peut être vu en comparant la Figure 3.6 avec la Figure 3.9, qui représente les valeurs de Brown (2008) pour une probabilité de projections optiques avant et après des corrections.

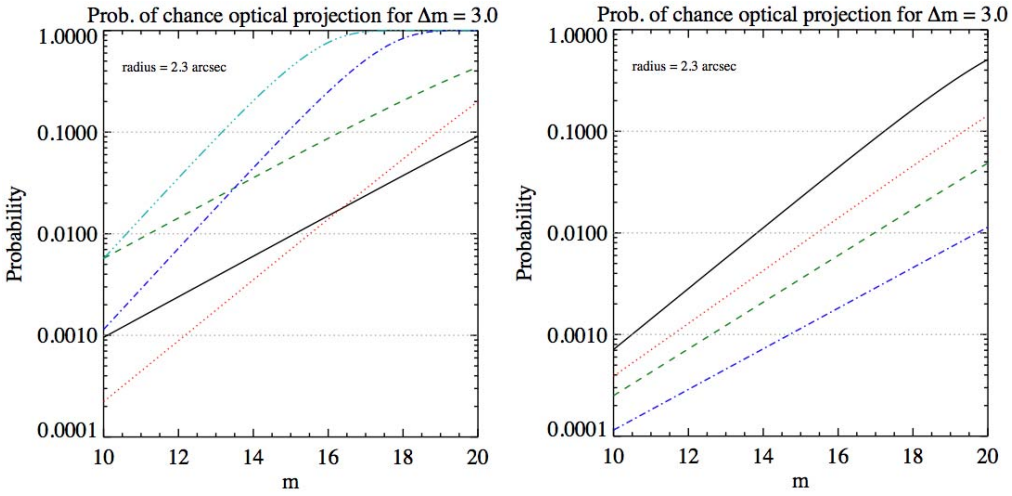


FIGURE 3.9 – Figures de Brown (2008) avec des estimations pour la probabilité de projection optique. Les lignes horizontales indiquent des niveaux de 0.1%, 1% et 10%. La figure de gauche représente des estimations originelles, et celle de droite des estimations corrigées.

3.4 La méthode d'analyse d'image – EIS

Dans n'importe quelle image astronomique, qu'elle soit normale ou reconstruite, pour que l'on puisse obtenir une information quantitative sur les sources présentes à partir de telles données, l'étape la plus importante est la réalisation d'une tâche dénommée ségrégation d'image.[7] Dans une première approximation il s'agit simplement de déterminer quels pixels appartiennent à quelle source dans l'image – en supposant que les signaux de toutes les sources sont complètement séparables.

Lorsque les sources ne sont pas séparables, la ségrégation devient un peu plus compliquée, car, idéalement, elle doit indiquer que dans certains pixels de l'image il existe un signal provenant d'une ou de plusieurs sources – il faut remarquer qu'il n'est pas nécessaire d'indiquer quel est le pourcentage de signal provenant de chaque source, car ceci exigerait une caractérisation détaillée du profil de brillance de chaque source individuelle présente dans l'image.

Comme nous avons vu dans le Chapitre 2, le cas des images reconstruites de Gaia est un cas particulièrement compliqué pour la ségrégation d'objets. Ces images reconstruites contiennent divers objets qui sont générés par les propres algorithmes utilisés pour leur reconstruction, ou par des effets causés par des directions préférentielles des angles de passage du satellite sur certaines coordonnées célestes. Le choix et/ou la création d'une méthode capable de résoudre ce problème de manière optimisée pour Gaia est rendu encore plus difficile par le fait que jusqu'à maintenant, on ne sait pas quel algorithme sera utilisé pour l'analyse des données.

Une difficulté additionnelle est le fait que la méthode utilisée pour la ségrégation d'image dans Gaia doit être efficace, étant donné que les estimations présentées dans les sections 3.2 et 3.3, permettent prévoir une limite supérieure d'une centaine de millions de cas pour lesquels une image aura besoin d'être reconstruite et analysée.

Afin de prendre en compte tous les facteurs commentés ci-dessus, nous avons développé une méthode de ségrégation décrite dans la présente section, qui est basée sur des algorithmes optimisés, en plus d'être capable d'apprendre automatiquement à reconnaître des objets à partir d'exemples.

3.4.1 Principes

L'idée de base de la méthode EIS (*Educated Image Segregator*) est qu'elle doit être capable d'apprendre et de s'auto adapter à la région analysée, étant donné que les caractéristiques d'un ensemble de pixels connexes (défini plus tard) formés par le signal d'une ou de plusieurs sources peuvent varier énormément en fonction de l'algorithme de reconstruction et de la distribution des angles de passage.

La mise en oeuvre suit une stratégie d'apprentissage computationnel supervisé, qui est un processus à deux étapes : la méthode est appliquée é sur un ensemble d'images de sources connues – de cette manière elle apprend quels sont les types différents de régions –, et devient donc capable de reconnaître à quel type de région la structure détectée appartient quand elle est appliquée sur une autre image.

7. Dans la littérature, la Ségrégation d'Image peut aussi être rencontrée comme Segmentation d'Image, ex. Gonzalez & Woods (2002).

Pour la détection de structures dans l'image, on utilise une stratégie de ségrégation d'image basée sur une coupe sur plusieurs niveaux avec deux raffinements supplémentaires : « étiquetage »de régions basé sur des graphes et la construction et l'analyse de graphes pour reconnaître quelles structures équivalent aux mêmes objets aux divers seuils (une altération d'une idée développée dans Starck et al, 2000, pour l'analyse d'image en multi-échelles).

Pour définir la ségrégation d'image, nous utilisons les définitions ci-dessous :

Définition 6 (Image) *Une image est une matrice bidimensionnelle I qui représente la version discrète d'un signal continu. Chaque position $I(i,j)$ de cette image est dénommée pixel avec indice (i,j), ou simplement pixel $I(i,j)$.*

Définition 7 (Fond) *Le fond F est un sous-ensemble de pixels de I qui ne contient que le bruit dû au fond de ciel réel et/ou dû à l'algorithme de reconstruction.*

Définition 8 (Objet) *Un objet est défini comme une source astronomique réelle.*

Définition 9 (Région) *Une région R est un sous-ensemble de pixels de I connexes qui en plus du signal du fond contient le signal d'au moins un objet, et/ou qui contient un signal provenant d'objets qui peuvent être créés aussi bien par l'algorithme de reconstruction que par le système optique.*

Définition 10 (Indépendance des Régions et du Fond) *Un pixel $I(i,j)$ d'une image peut appartenir à une seule région R_i ou au Fond F, soit : $\forall A, B \in \{F, R_1, ..., R_n\}, A \neq B \Leftrightarrow \bigcap \{A, B\} = \emptyset$.*

Corollaire 2 (Composition d'une Image) *Toute image I peut être complètement décomposée en termes de Régions ou Fond, c'est-à-dire, soit n Régions ($n \in \mathbb{N}$) et un Fond, l'image est formée le par l'ensemble $I = \bigcup \{F, R_1, ..., R_n\}$.*

Définition 11 (Ségrégation d'image) *La ségrégation d'une image I est la décomposition d'une image I en termes du fond F et des régions R_n ($n \in \mathbb{N}$). C'est-à-dire, la ségrégation d'image est la résolution du problème suivant : soit une image I, déterminer F et R_n.*

3.4.2 Mode avec Seuil-simple

Le premier mode de la méthode que nous avons développée est basé sur l'utilisation d'un seuil simple de détection d'objets. C'est la forme la plus simple possible pour réaliser une ségrégation d'objets sur une image, et elle revient à réaliser tout d'abord une binarisation de l'image qui est analysée, et donc à analyser l'image binarisée dans le but de rencontrer des pixels connexes.

Définition 12 (Binarisation d'image à un niveau η) *La binarisation d'une image I à un niveau η est la production d'une deuxième image I_B formée par le couple pixels $I_B(i,j) \in \{0,1\}$. Les valeurs des pixels de I_B sont données par :*

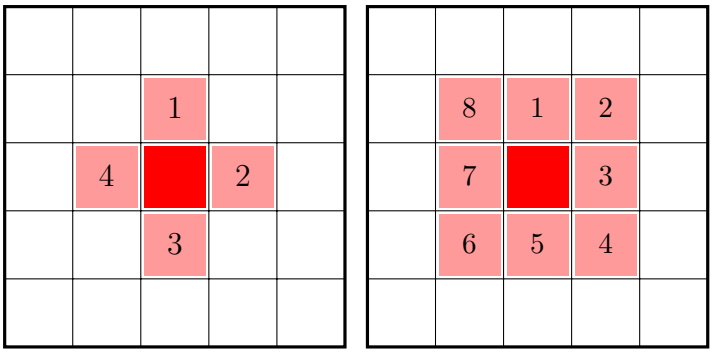
$$I_B(i,j) = \begin{cases} 0 & \text{si } I(i,j) < \eta \\ 1 & \text{si } I(i,j) \geq \eta \end{cases}$$

Définition 13 (Pixels 4-voisins) *Un pixel $I(k,l)$ est 4-voisin d'un pixel $I(i,j)$ si, et seulement si $(k,l) \in \{(i-1,j),(i,j-1),(i+1,j),(i,j+1)\}$.*

Définition 14 (Pixels 8-voisins) *Un pixel $I(k,l)$ est 8-voisin d'un pixel $I(i,j)$ si, et seulement, si $(k,l) \in \{(i-1,j-1),(i-1,j),(i-1,j+1),(i,j-1),(i,j+1),(i+1,j-1),(i+1,j),(i+1,j+1)\}$.*

Définition 15 (Pixels connexes) *Un couple de pixels dans une image binarisée est considéré connexe si les deux possèdent la valeur 1 et si les deux sont 4- ou 8-voisins – se dénommant 4-connexe ou 8-connexe, respectivement.*

Graphiquement, les critères de connectivité définis à partir des voisinages définis ci-dessus sont représentés sur la Figure 3.10. Bien que la méthode mise en oeuvre puisse être utilisée avec les deux critères de connectivité (les deux ont été mis en oeuvre), en général seul le mode 8-connexe est utilisé.



(a) (b)

FIGURE 3.10 – Représentation graphique de 4-connectivité (a) et 8-connectivité (b).

Chaque ensemble de pixels connexes définit donc une région R_n. À partir de l'image binarisée on détermine aussi l'ensemble des pixels appartenant au fond F, comme étant tous ceux qui possèdent une valeur 0, c'est-à-dire que sur l'image I, ils possèdent un signal d'intensité d'une valeur inférieure à un certaine nombre η. Pour la détermination des régions Rn, la méthode utilise un algorithme d'étiquetage (connected-component labeling)en deux itérations.[8]

Au premier passage, l'algorithme analyse l'image binarisée I_B en balayant chaque ligne de la matrice et crée une matrice I_E (de mêmes dimensions que I_B) et dans laquelle des étiquettes temporaires sont écrites. De plus, pour chaque étiquette créée durant ce premier passage sont créés des graphes qui gardent des informations sur l'équivalence des étiquettes , c'est-à-dire qui déterminent quelles étiquettes différentes

8. Les iterations sont realisés comme dans la méthode originale proposée dans Rosenfeld & Pfaltz (1966), cependant, nous avons introduit une modification basée sur la construction de graphes (inspirée par Fiorio & Gustedt, 1996) – qui doit être plus efficace, comme l'ont observé Chang et al (2004) pour l'algorithme de Fiorio & Gustedt (1996).

pourraient être données à un unique pixel par le fait qu'il est 8-connexe avec deux autres pixels d'étiquettes différentes.

Le deuxième passage survient dans la matrice I_E, et a pour objectif de remplacer chaque étiquette de cette matrice par l'étiquette avec la plus petite valeur stockée dans le graphe qui décrit ses équivalences – à chaque étiquette analysée le graphe est réduit : une nouvelle arête est créée en connectant l'étiquette en analyse directement au sommet avec l'étiquette de plus petite valeur, de manière à ce que pour chaque étiquette analysée le graphe demande á être parcouru une seule fois durant toute analyse. Un exemple graphique des deux passages de cet algorithme peut être rencontré sur la Figure 3.11.

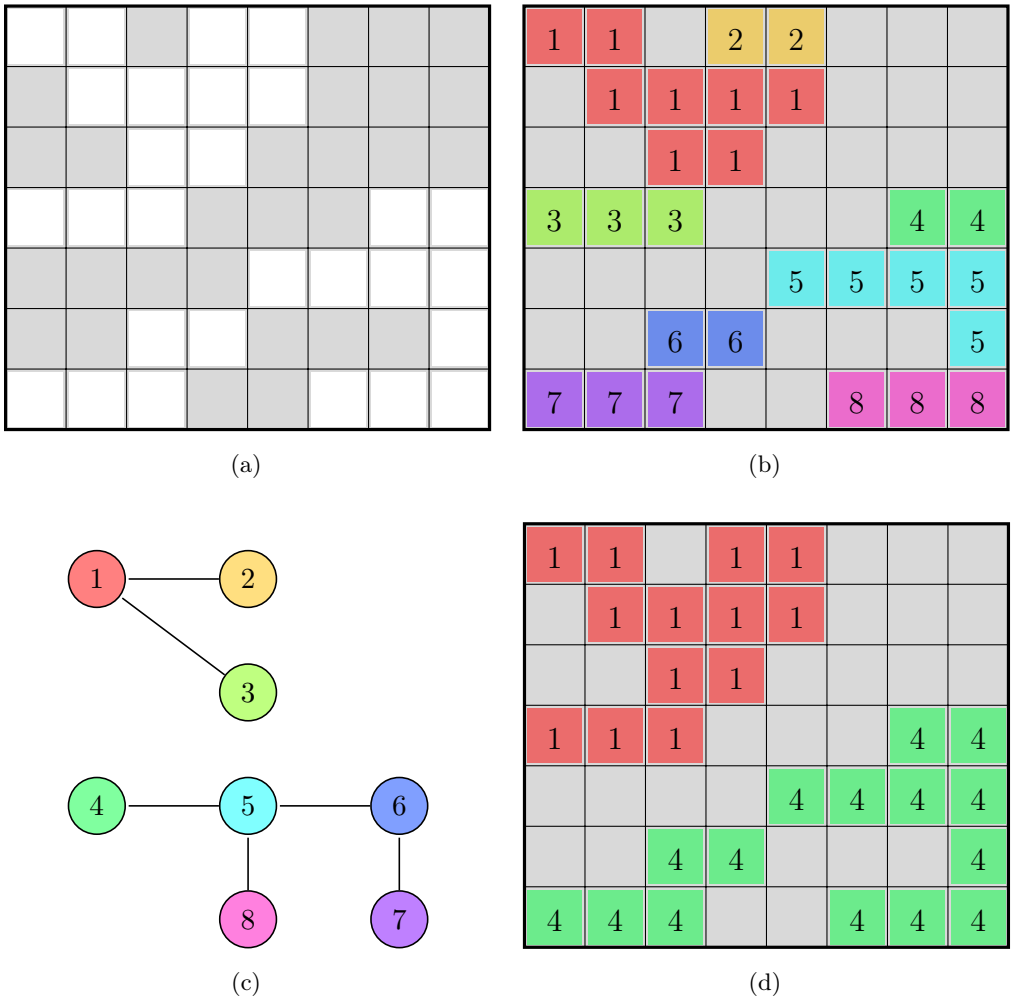


FIGURE 3.11 – Représentation schématique du processus d'étiquetage des régions en utilisant un algorithme de deux passages et un critère de 8-connectivité. Dans (a) un exemple d'image binarisée. Dans (b) le premier passage. Dans (c) les graphes d'équivalence de régions. Dans (d) le second passage.

Les résultats obtenus par ce mode à seuil simple seront sévèrement dépendants de la valeur choisie pour le seuil η durant la binarisation. Ceci peut être intéressant dans les cas où l'image est dominée par le fond de ciel, et lorsqu'une mesure statistique raisonnable (comme par exemple une moyenne calculée avec un rejet itératif à 3σ) ou une analyse de la distribution des pixels (comme dans Bijaoui, 1980) peut être appliquée pour la détermination d'une valeur optimum pour η.

Néanmoins, dans le cas des images reconstruites de Gaia, ce mode est inefficace étant donné la détermination qu'un η optimum est compliquée du fait qu'il existe des directions préférentielles pour les angles de passage (donc des régions de l'image avec des caractéristiques différentes du bruit) et que nous désirons analyser des sources avec des contrastes relativement grands ($\Delta m \sim 3$) et avec des séparations angulaires de quelques centaines de millisecondes d'arc (de l'ordre de deux ou trois PSFs). Tout ceci requiert une grande sophistication de la méthode, qui serait la détermination des η des sous régions de l'image, mais cette détermination exigerait que l'on sache (ou estime) préalablement s'il existe une source quelconque à analyser dans une position donnée. Dans le but de résoudre de tels problèmes sans l'introduction d'analyses locales, un deuxième mode a été développé, décrit dans la section suivante.

3.4.3 Mode avec Seuil-multiple

Le mode avec seuil multiple consiste à utiliser le mode de seuil simple globalement sur l'image pour différentes valeurs de η, produisant diverses listes de régions, chacune pour une valeur différente de η. Ces listes sont donc englobées dans un seul ensemble, la liste globale de régions, et à partir de cette liste globale nous déterminons quel est le meilleur ? pour chacune des régions rencontrées.

Définition 16 (Liste globale de régions) *Soit $L_\eta = \{R_1^\eta, ..., R_n^\eta\}$ un ensemble formé par toutes les régions déterminées par le mode de seuil simple en utilisant un seuil η, l'ensemble $P = \{L_{\eta_1}, ..., L_{\eta_j}\}$ est l'ensemble formé par toutes les régions déterminées pour j valeurs distinctes de η, et est dénommé liste globale de régions.*

Corollaire 3 *$\forall \eta_i > \eta_j$, pixel présent dans une région $R_n^{\eta_i}$ rencontrée par le mode de seuil simple appliqué sur l'image I avec un seuil η_i, sera rencontré dans une région quelconque $R_k^{\eta_j}$ déterminée en utilisant le seuil η_j sur la même image.*

Démonstration 1 (Corollaire 3) *Par la définition de région d'une image I, $R_n^{\eta_i}$ est formée par, au moins, p pixels $I(i,j)$ 4 ou 8-connexes dont les valeurs de leurs équivalents $I_B^{\eta_i}(i,j)$ dans l'image binarisée $I_B^{\eta_i}$ créée en utilisant le seuil η_i sont égales à 1. Mais, $I_B^{\eta_i}(i,j) = 1 \Leftrightarrow I(i,j) > \eta_i \Leftrightarrow I(i,j) > \eta_j, \forall \eta_j \leq \eta_i$*■

Corollaire 4 *Tous les pixels présents dans l'ensemble L_{η_i} seront présents dans l'ensemble L_{η_j}, $\forall \eta_j < \eta_i$.*

Corollaire 5 *$\forall \eta_i > \eta_j$, $\forall\ R_n^{\eta_i}$, $\exists R_n^{\eta_j}$: $R_n^{\eta_i} \subset R_n^{\eta_j}$, , c'est-à-dire, pour toute région détectée à un seuil η_i, il existe une région détectée à un seuil $\eta_j < \eta_i$ qui contient tous les pixels de la région détectée dans η_i.*

Les valeurs choisies pour les niveaux de seuils η peuvent être déterminées à partir d'un type quelconque de statistique ou choisies à partir d'un type quelconque de connaissance à priori des caractéristiques des images qui seront analysées (on peut par exemple inclure l'information sur le comportement d'images reconstruites dans certains régimes de distribution d'angles de passage).

La mesure statistique standard utilisée pour la détermination de la valeur du seuil inférieur de détection η_0 est une moyenne calculée avec un rejet itératif en 3σ : tout d'abord toutes les valeurs des $(n \times m)$ pixels de l'image I sont copiées dans un ensemble $K_0 = \{k_1^0, ..., k_{(n\times m)}^0\}$; on calcule alors la valeur moyenne $\langle K_0 \rangle$ et l'écart type $\sigma(K_0)$ qui servent à définir un sous-ensemble K_1 créé à partir des valeurs de K_0 passant le test à 3σ. L'algorithme est itéré et à la $n+1$ itération, $K_{(n+1)} = \{k_i^n : k_i^n < (\langle K_n \rangle + 3\sigma(K_n))\}$. Les itérations cessent lorsque $|\langle K_{(n+1}\rangle - \langle K_n \rangle| < qK_n$, où le critère de convergence $q = 10^{-3}$ est empirique.

Habituellement le choix des autres seuils est simplement réalisé à partir de la détermination de la valeur de plus grande intensité sur l'image (pris comme $\eta_{\max}$), et d'un échantillonnage d'un certain nombre n d'intervalles réguliers de taille $\Delta\eta = (\eta_{\max} - \eta_0)/(n-1)$. La Figure 3.12 montre une représentation unidimensionnelle du processus de construction de la liste globale de régions.

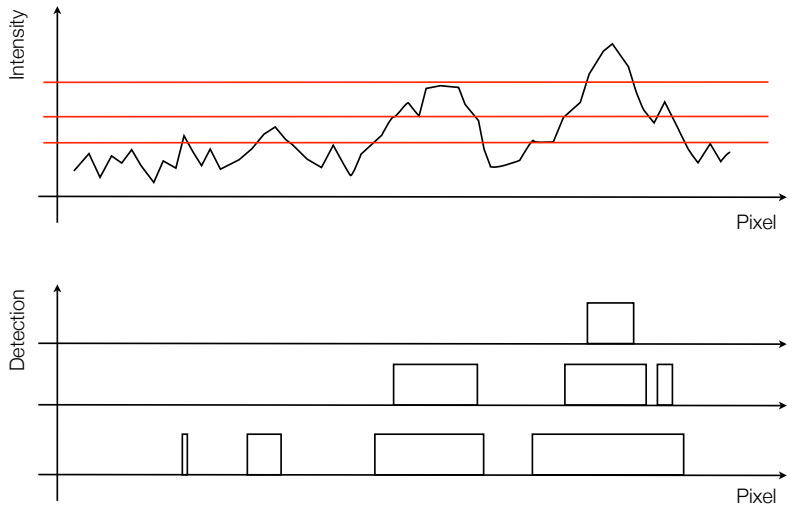


FIGURE 3.12 – La figure supérieure montre une abstraction unidimensionnelle d'une image qui sera séparée en trois niveaux de seuil. La figure inférieure montre les listes de régions séparées pour chaque seuil – la liste globale de régions est un ensemble qui inclut toutes les régions détectées à tous les seuils utilisés.

Etant donné que l'on connaît la liste globale de régions P, des graphes sont créés, que nous dénommons graphes de connectivité inter-seuil (*inter-threshold connectivity graph*) reliant les régions déterminées à différents seuils η (voir Figure 3.13). Chaque sommet des graphes (point rouge) indique l'existence d'une région rencontrée à un niveau quelconque de η. Pour déterminer les connexions, les régions correspondantes à ces nœuds sont vérifiées, et les régions qui possèdent des pixels en commun sont

liées par des arêtes – par le corollaire 5 il est garanti qu'un sommet déterminé à partir d'une région rencontrée sur un seuil η_n sera toujours connecté à un sommet créé à partir d'une région détectée sur un seuil $\eta_{(n-1)}$.

La connexion de ces régions est intéressante, car un même objet doit créer des régions sur divers seuils, et comme le corollaire 4 nous garantit que tout sommet déterminé au niveau η_i sera connecté à un sommet quelconque au niveau η_j , pour tout $\eta_j < \eta_i$, nous savons que de telles régions seront toujours connectées entre elles sur le graphe.

Une conséquence du corollaire 5, est qu'un sommet au niveau η_j peut se connecter avec un ou plusieurs sommets déterminés au seuil η_i indiquant une fusion de régions à un seuil η_{fusion} tel que $\eta_j \leq \eta_{fusion} < \eta_i$.

Une autre conséquence du corollaire 5 est qu'en supposant $\eta_0 = \min(I)$ il n'existera qu'un sommet, tous les graphes se connectent à ce niveau et que l'image peut donc être représentée par un graphe type arbre (ce que généralement nous ne faisons pas, étant donné dans un but pratique, on adopte toujours un $\eta_0 > 0$).

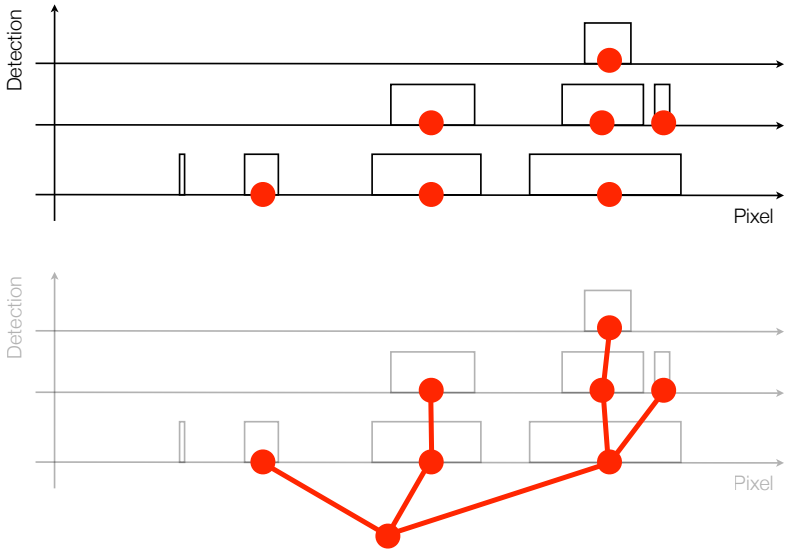


FIGURE 3.13 – La figure supérieure montre la détermination des sommets du graphe déterminé pour l'abstraction de la Figure 3.12. La figure inférieure montre le graphe créé à partir de la connexion des sommets avec les arêtes (en incluant un niveau pour $\eta = 0$ qui transforme le graphe en arbre).

Le problème à résoudre réside donc dans le choix de la coupure des arêtes de ces graphes, générant une liste finale de régions détectées. Évidemment, ici, on désire choisir les régions qui ne sont pas en état de fusion, de manière à ce qu'une solution triviale pour ce problème soit de couper le graphe sur les arêtes qui relient un sommet à seulement un autre sommet, en conservant les régions correspondantes aux sommets les plus élevés (feuilles du graphe arbre) dans l'ensemble final des régions. Une autre solution possible, celle qui est utilisée comme standard dans la méthode, est de couper le graphe sur des arêtes qui relient des sommets de degré plus grand que 2

(c'est-à-dire, qui possèdent des arêtes les reliant à plus de deux autres sommets). La Figure 3.14 montre la différence entre les régions incluses dans la liste finale des régions du mode de multiples seuils quand une ou l'autre solution est adoptée pour la coupe du graphe.

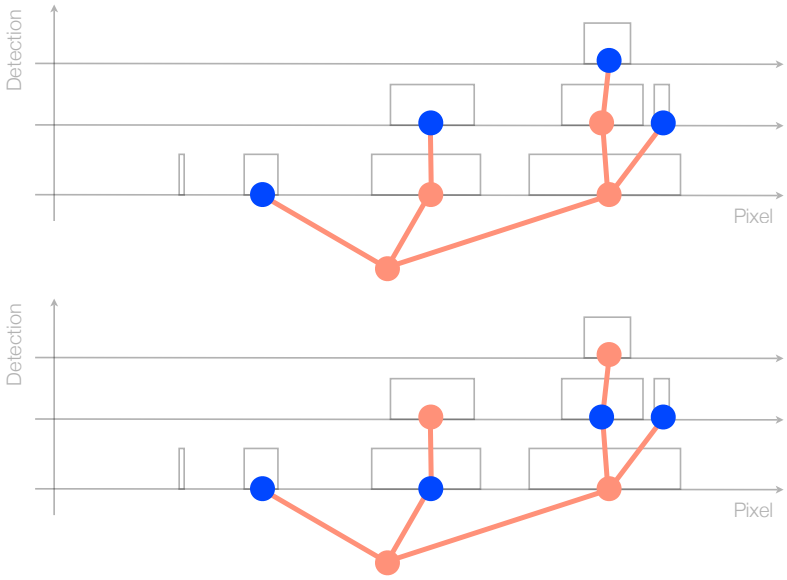


FIGURE 3.14 – Des graphes indiquant les régions formant le résultat du mode multi-seuil selon le critère utilisé pour sa coupe – les sommets bleus sont ceux qui composent l'ensemble final de régions. Dans la figure supérieure, la coupe par le critère de feuilles de l'arbre. Dans la figure inférieure avec un critère de degré plus grand que 2.

Il faut remarquer que ce mode est trivialement parallélisable, étant donné que l'exécution du mode de seuil simple en utilisant divers seuils peut être réalisée en même temps pour chaque seuil – ensuite, l'analyse et la coupe du graphe peuvent être réalisées aussi bien en parallèle que séquentiellement (mais le processus est relativement rapide et la parallélisation de cette partie n'apporte pas un grand gain de temps).

Cependant le mode peut être facilement optimisé même pour un traitement séquentiel : il suffit d'analyser l'image de bas en haut, en créant des sous-ensembles avec chaque fois moins de pixels, étant donné les pixels qui n'appartiennent à aucune région à un seuil donné η petit, continueront à n'appartenir à aucune région quand une valeur supérieure à η est utilisée. Finalement, ceci peut être généralisé à plusieurs dimensions, permettant par exemple l'analyse de cubes de données.

Le problème de ce mode est que, bien que son taux de détection d'objets soit plutôt élevé, il détecte toutes les structures possibles, y compris le bruit (au cas où on utilise des seuils de détection très bas et un nombre très petit de pixels connexes pour définir une détection), en plus de tous les artefacts générés par les algorithmes de reconstruction d'images. Dans la prochaine section, nous verrons comment nous avons traité ce problème.

3.4.4 Apprentissage supervisé – *Educated mode*

L'objectif principal de l'apprentissage supervisé est de générer une liste « propre » de régions détectées, c'est-à-dire, avec le nombre minimum, si possible zéro, d'artefacts de reconstruction. Comme son nom le suggère, ce mode doit être éduqué pour fonctionner. L'étape d'éducation, ou comme elle est plus communément appelée, d'entraînement, est réalisée par l'apprentissage par la méthode d'un ensemble d'exemples. Ainsi, le mode éduqué fonctionne en deux étapes distinctes :

1. **Entraînement** Durant l'entraînement, une série d'images « prototypes » est présentée à la méthode, sur lesquelles les objets sont préalablement connus (ces images peuvent être produites, par exemple, par des simulateurs). La méthode utilise alors le mode multi-seuil pour créer le graphe de détections, mais ne réalise pas la coupe sur ce graphe, produisant une liste de régions sur de multiples seuils. Pour chacune des régions détectées, certains paramètres morpho-photométriques sont mesurés à partir de ses pixels, et les données d'apprentissage sont produits avec de telles mesures et avec les types de région que l'on désire classer (objet de reconstruction, objet simple, fusion de multiples objets), qui est stockée et utilisée pour former un algorithme de classement supervisé.

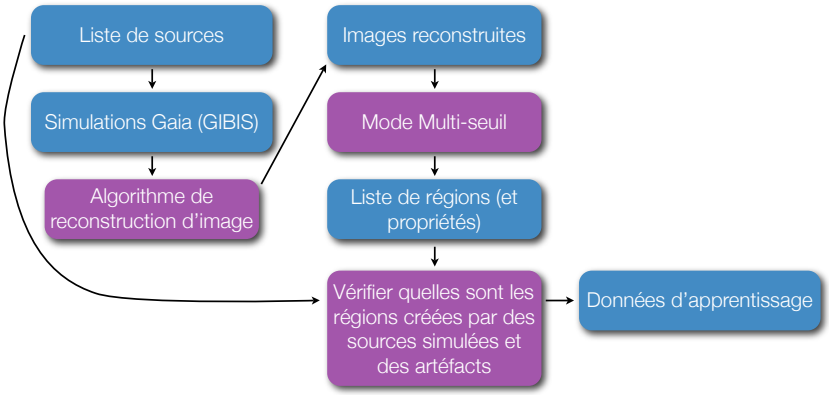


FIGURE 3.15 – Diagramme représentant de forme schématique l'entraînement supervisé.

2. **Détection** On appliquera le mode multi-seuil sur toute image que l'on désire analyser. Donc, la méthode mesurera les paramètres morpho-photométriques de chaque région de la liste produite par le mode multi seuils. Ce sont eux sur lesquels sera basée son classement. Une probabilité que les région appartiennent à chacune des classes pour laquelle la méthode a été entraînée sera ainsi estimée. La plus grande probabilité désignera à quelle classe appartient telle ou telle région. Finalement, une liste « propre » de régions sera produite au travers d'une sélection dans la liste renvoyée par le mode multi-seuil dans une

probabilité quelconque choisie à priori pour que la région soit formée par des régions qui ne correspondent qu'à des objets simples et/ou fusion de multiples objets.

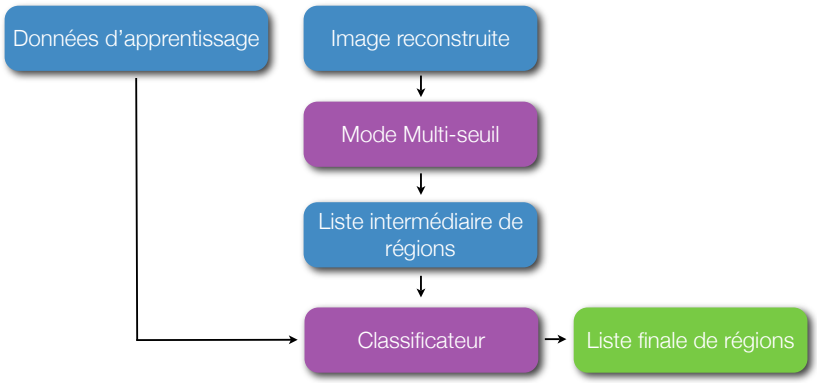


FIGURE 3.16 – Diagramme représentant l'étape de détection du *Educated mode*.

L'ensemble d'exemples, ou ensemble d'entraînement, peut être mieux défini dans le contexte du mode éduqué en utilisant $n-uplets$:

Définition 17 (Données d'apprentissage) *Les données d'apprentissage sont un ensemble formé par n-uplets* ($n \geq 2$)*, dans lequel* $n-1$ *éléments sont les paramètres morpho-photométriques mesurés pour les pixels de la région et l'élément restant stocke la classe de la région (généralement artefact, objet simple ou objets multiples non-résolus).*

Les paramètres morphologiques et photométriques utilisés ont été déterminés heuristiquement, et pour préserver l'efficacité computationnelle de la méthode, ils consistent en un petit nombre de grandeurs facilement calculables. Ces paramètres consistent en :

1. Nombre de pixels qui définissent la région ;
2. Niveau de détection de la région au-dessus du bruit ;
3. Rapport entre l'écart type de la distribution des intensités des pixels et leur moyenne ;

L'obtention de ces paramètres est une tâche réalisée en pratique par les algorithmes de détection quasi gratuits en temps de calcul pour le mode éduqué – cependant ils sont stockés comme des propriétés de la région seulement quand le mode supervisé est utilisé.

D'autres paramètres plus complexes peuvent être inclus si besoin est, en fonction des caractéristiques des images analysées, donc des algorithmes de reconstruction utilisés. Néanmoins pour les finalités de ségrégation d'images de Gaia, les populations que nous désirons classer, qui sont un artefact, un objet simple, une fusion de multiples

objets, se trouvent bien séparées, occupant des régions distinctes de l'espace défini par cette triade, comme on peut le voir sur la Figure 3.17.

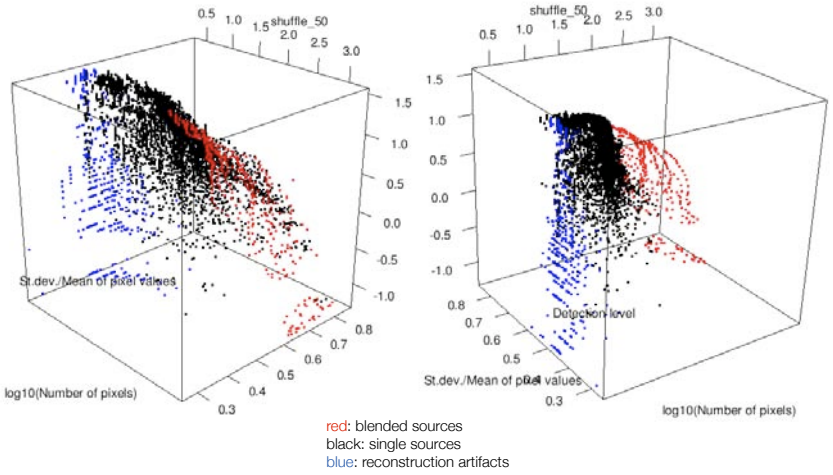


FIGURE 3.17 – Exemple de l'espace de paramètres utilisés pour le classement des régions détectées. Les points bleus représentent les régions correspondant aux artefacts, les points noirs aux objets réels, les points rouges aux multiples objets non résolus. Les graphes ont été créés en utilisant des images simulées et reconstruites par l'algorithme *Shufflestack*.

Pour classer les régions, tout algorithme de classement supervisé peut être utilisé. Le choix standard est un algorithme relativement simple, appelé k-NN (ou k-plus proches voisins). Bien qu'il s'agisse de l'un des plus simples algorithmes pour la reconnaissance de standards et bien qu'il soit un « classique », il est encore l'objet de recherches récentes, par exemple Hall et al (2008) pour le choix de la valeur optimum de k, ou Toussaint (2005) pour des optimisations basées sur la construction de graphes.

Intuitivement, il est simple de comprendre comment le k-NN fonctionne : si l'on désire classer un point de classe inconnue, on calcule une distance entre ce point et tous les points d'un ensemble de points préalablement classés. Le point inconnu sera classé dans la classe apparaissant le plus dans ces k voisins. Sa définition formelle peut être écrite comme :

Définition 18 (Classement k-NN) *Soit un ensemble de p classes $C = \{c^1, ..., c^p\}$, de manière que $\forall c^i$, $c^j \in C$, $c^i = c^j \Leftrightarrow i = j$ – nous appelons cet ensemble « ensemble des classes disponibles ». Soit un ensemble donné formé par m n-uplets $T = \{t_1, ..., t_m\}$. Soit un ensemble denominé d'ensemble des classes des n-uplets, un ensemble de taille m dont les éléments peuvent être seulement des éléments de l'ensemble des classes disponibles, $C_T = \{c_1, ..., c_m\}$, de manière que $\forall i$: $1 \leq i \leq m$, $c_i \in C$.*

Pour déterminer la classe d'un n-uplet x quelconque au moyen de la méthode de k-NN, on construit l'ensemble $D = \{d(x, t_1), ..., d(x, t_m)\}$ formé par les distances (généralement euclidiennes) entre x et tous les membres de T. Ensuite, on détermine en D les k éléments avec les plus petites valeurs, et à partir des indices de ces éléments on vérifie en C_T combien de fois chacune des p classes apparaissent, formant l'ensemble $N = \{N_1, ..., N_p\}$. Finalement la probabilité que x appartienne à la i-ième classe est donnée par N_i/k, étant entendu que la classe de x est définie comme étant la classe pour laquelle cette probabilité est maximale.

Disposant des paramètres mesurés et d'un algorithme de classement, il suffit maintenant de construire une base de donnée d'apprentissage. Pour construire cet ensemble dans le cas de Gaia, nous utilisons GIBIS (tel que présenté sur la Figure 3.15), qui est capable de simuler avec réalisme les fenêtres observées par Gaia, et nous reconstruisons donc une image qui sert de prototype pour toute autre source ayant les mêmes caractéristiques photométriques et position sur la sphère céleste.

Nous avons ainsi une méthode de ségrégation d'images qui possède la capacité de :

1. Ségréguer des régions créées par des objets avec de grands contrastes entre eux sur une même image (dû au mode multi-seuil) ;
2. Ségréguer des régions créées par des objets angulairement proches entre eux (dû au mode multi-seuil) ;
3. Ségréguer des régions sans se baser sur une morphologie pré-déterminée pour les objets (dû au mode de seuil-simple) ;
4. Discerner entre des régions créées par des objets uniques multiples et des objets générés par les algorithmes de reconstruction (dû au mode éduqué) ;
5. Auto-apprentissage – ce qui le rend utilisable sur des images normales ou reconstruites indépendante de l'algorithme de reconstruction (dû au mode éduqué).

Mais le mode éduqué présente une dernier difficulté lié au fait que les algorithmes de reconstruction d'image reconstruisent de façon différente les caractéristiques morpho-photométriques des régions en fonction du bruit généré par l'objet primaire, du nombre de passages disponibles pour la reconstruction et principalement de la configuration des angles de passage (dépendantes de la position du ciel dans laquelle se situe l'objet).

La solution idéale pour ce problème serait d'avoir accès à une simulation réaliste des observations de Gaia sur chaque coordonnée et pour chaque magnitude d'objet primaire, pour lequel on désire réaliser une ségrégation d'image. Cependant une telle solution n'est pas réaliste en raison du temps de calcul nécessaire pour réaliser de telles simulations : il dépasse de plusieurs ordres de grandeur la quantité de temps disponible pour le traitement de chaque image reconstruite. La solution adoptée a été de créer une bibliothèque avec un certain nombre de paramètres d'entraînement préparées pour des régions prototypes.

3.4.4.1 Base de données d'apprentissage

Pour que la méthode développée puisse être appliquée à n'importe quelle direction du ciel, nous avons développé une bibliothèque regroupant toutes les base de données d'apprentissage correspondant à ces régions. Pour cela nous avons développé des classes Java qui automatisent la génération des paramètres d'apprentissage à partir de simulations GIBIS et qui sélectionne les données d'apprentissage les plus adaptées à l'analyse d'une image donnée.

Pour la création d'une base de données d'apprentissage, la bibliothèque est alimentée directement avec les fichiers provenant des simulations GIBIS et lance automatiquement tous les codes IDL et Java nécessaires à la reconstruction des images, à l'entraînement de la méthode de ségrégation et à la génération des paramètres d'entraînement

L'utilisation de cette bibliothèque est complètement transparente, étant entendu que tous les objets sont créés et utilisés en interne par le code de ségrégation d'image. Au moment d'analyser une image le code de ségrégation d'image dans le mode éduqué demande automatiquement au code de gestion de la bibliothèque l'ensemble de données le plus adapté pour l'image qui est ségréguée. Pour cela le dispositif de gestion de la bibliothèque analyse :

1. L'algorithme de reconstruction ;
2. La magnitude de la source primaire ;
3. Les coordonnées de l'image dans le ciel ;
4. Le nombre de passages.

Une bibliothèque de base composée par une quantité minimale de paramètres d'apprentissage a été créée dans un but de tests et de démonstrations du concept. La Figure 3.18 montre les coordonnées présentes sur cette bibliothèque de base.

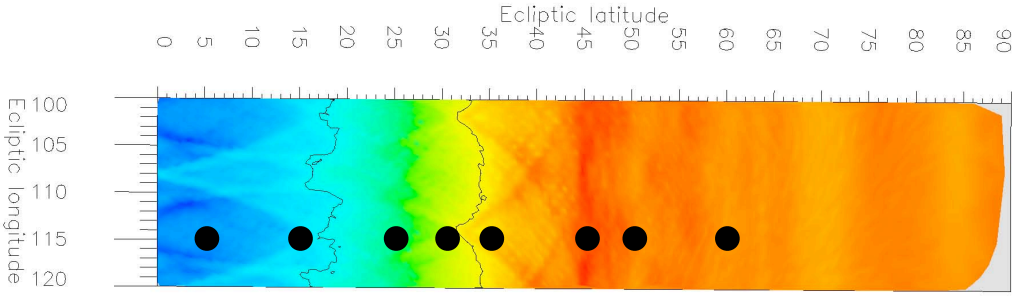


FIGURE 3.18 – Les coordonnées présentes dans la bibliothèque de base des paramètres d'apprentissage sont indiquées avec des cercles noirs – chaque coordonnée contient des paramètres pour trois magnitudes et quatre algorithmes de reconstruction.

Chaque coordonnée dans la bibliothèque de base (représentée sur la Figure 3.18) contient des paramètres pour trois magnitudes d'objets primaires (12, 15, 19), pour le nombre de passages à la fin de la mission et pour les 4 algorithmes de reconstruction

disponibles actuellement. Cependant, des bibliothèques avec une résolution beaucoup plus élevée dans l'espace des paramètres (spécialement en termes de positions sur la sphère céleste) seront créées.

3.4.5 Tests et résultats

Dans le but de déterminer la qualité des résultats obtenus en utilisant la méthode de ségrégation d'images décrite dans cette section, nous utilisons GIBIS pour générer des simulations de régions avec quatre sources secondaires autour d'une source primaire. Après avoir reconstruit les images avec les quatre algorithmes de reconstruction présentés dans le Chapitre 2, nous appliquons la méthode de ségrégation décrite antérieurement[9] et nous avons cherché pour quelles distances angulaires et différences de magnitude par rapport à l'objet primaire cette méthode est capable de récupérer les objets secondaires.

Nous avons aussi voulu évaluer le nombre de fausses détections, c'est-à-dire, de régions qui sont détectées mais ne correspondent à aucun objet simulé. Il est important de remarquer que ces fausse détections peuvent aussi bien être des artefacts générés par les algorithmes de reconstruction, que des régions générées par des objets simulés qui ont vu leurs positions significativement altérées par les algorithmes de reconstruction. En effet l'identification d'une fausse détection survient quand une région n'inclut pas les coordonnées de l'objet simulé. Ainsi le nombre de fausse détections doit être vu comme une limite supérieure du nombre d'artefacts de reconstruction détectés.

Comme nous l'avons indiqué précédemment, notre méthode de ségrégation exige une étape d'entraînement avant d'être capable de reconnaître les régions. Il est donc naturel de supposer que deux régimes distincts existeront pour lesquels leurs résultats doivent être testés : un premier régime dans lequel il existe dans la bibliothèque un ensemble de paramètres d'entraînement avec des caractéristiques égales ou très proches de celles de l'image qui sera ségréguée, et un second régime dans lequel les caractéristiques sont distantes. Voyons donc les résultats pour chacun de ces deux régimes.

3.4.5.1 Résultats dans le cas idéal

Le cas idéal d'application de notre méthode d'analyse d'images survient quand les caractéristiques principales de l'image sont très proches de certains des exemples présents dans la bibliothèque d'entraînement. Ces caractéristiques sont celles prises en considération au moment de sélectionner une base de données d'apprentissage, c'est-à-dire : l'algorithme de reconstruction, la magnitude de la source primaire, les coordonnées de l'image dans le ciel et le nombre de passages.

Naturellement, le premier test à réaliser est une comparaison entre les résultats obtenus avec les divers algorithmes de reconstruction d'image qui ont été présentés dans le Chapitre 2, car comme nous l'avons dit dans ce chapitre, l'algorithme qui

9. nous l'avons mise en oeuvre en Java selon les spécifications du DPAC (O'Mullane et al, 2008).

sera utilisé pour la reconstruction durant la réduction des données n'a pas encore été défini, et des résultats de tests tel que celui-ci doivent être pris en compte pour la sélection de l'algorithme qui sera adopté.

Sur les Figures 3.19 et 3.20 nous pouvons vérifier les résultats obtenus pour la détection de sources secondaires autour d'objets primaires brillants (G=12) et faibles (G=19). Dans les simulations, quatre sources secondaires ont été inclues autour de chaque objet primaire dans les positions nord, sud, est et ouest, à diverses distances angulaires de l'objet primaire et avec plusieurs magnitudes, de $\Delta G = 0.5$ à $\Delta G = 4.5$.

Les résultats obtenus sur la détection d'objets autour d'une primaire avec $G = 12$ (Figure 3.19), montrent que ceux obtenus avec l'algorithme *BinOutliers* et *ShuffleStack* sont bien supérieurs à ceux obtenus avec le *Quickstack* et *Drizzle* en termes de nombre de sources récupérées (nombre de carrés bleus dans les graphes (a), (c), (et) et (g)). De plus, au moins pour une primaire de magnitude $G = 12$, nous avons démontré que nous sommes capables de récupérer sur une image reconstruite le signal de sources avec un grand contraste (jusqu'à $\Delta G = 4.5$) et proches angulairement de la source primaire (jusqu'à $\Delta G = 3.0$ pour une source localisée à 300 mas de la primaire) – ce qui démontre l'efficacité de la ségrégation réalisée au moyen du mode multi-seuil.

En termes de nombre de fausses détections (résultat présenté sur les graphes (b), (d), (f) et (h)), il est essentiellement nul pour tous les algorithmes, avec une petite exception pour une détection réalisée sur une image *ShuffleStack*. Ceci démontre l'efficacité de la ségrégation éduquée d'images, et comment une pureté relativement grande dans la liste de régions détectées peut être obtenue.

On peut aussi noter que durant les tests réalisés, nous avons considéré seulement les objets avec une probabilité de 0.9 de provenir d'un objet réel. Si nous abaissions ce seuil, le nombre de sources récupérées augmenterait, mais et le nombre de fausses détections augmenterait aussi.

Un autre comportement intéressant observable sur la Figure 3.19 (a), est que la détection d'une secondaire proche de l'étoile centrale est liée à la différence de magnitude : plus elle est proche, moins cette différence est grande. Ce comportement s'explique par le fait que lorsque la secondaire est très proche de l'étoile primaire, son flux ne sera pas suffisant pour que son signal surpasse le signal de la primaire. Toutefois, même si ces objets ne sont pas dans la liste de détections en fonction de la coupe réalisée dans cette liste, dans beaucoup de cas, des régions ont été détectées et classées comme ayant été générées par des « objets multiples fusionnés » – ceci permettra (si nécessaire) le développement d'une méthode de *deblending* pour séparer ces objets de la source primaire.

Un comportement identique se répète en (c) pour l'algorithme de *Drizzle*, également en (e) et (g) pour *ShuffleStack* et pour *BinOutliers*, bien qu'avec un sensibilité supérieure à *Quickstack*.

Un autre comportement clair qui peut être observé sur la Figure 3.19 (a) et (b) est la coupe dans les détections d'objets secondaires survenant en $\Delta G \sim 2.25$ pour les algorithmes de *QuickStack* et *Drizzle*. Cette coupe peut être comprise comme une limitation de la méthode de reconstruction d'images pour des sources plus faibles de $2,25$ magnitudes par rapport à celle de l'objet primaire.

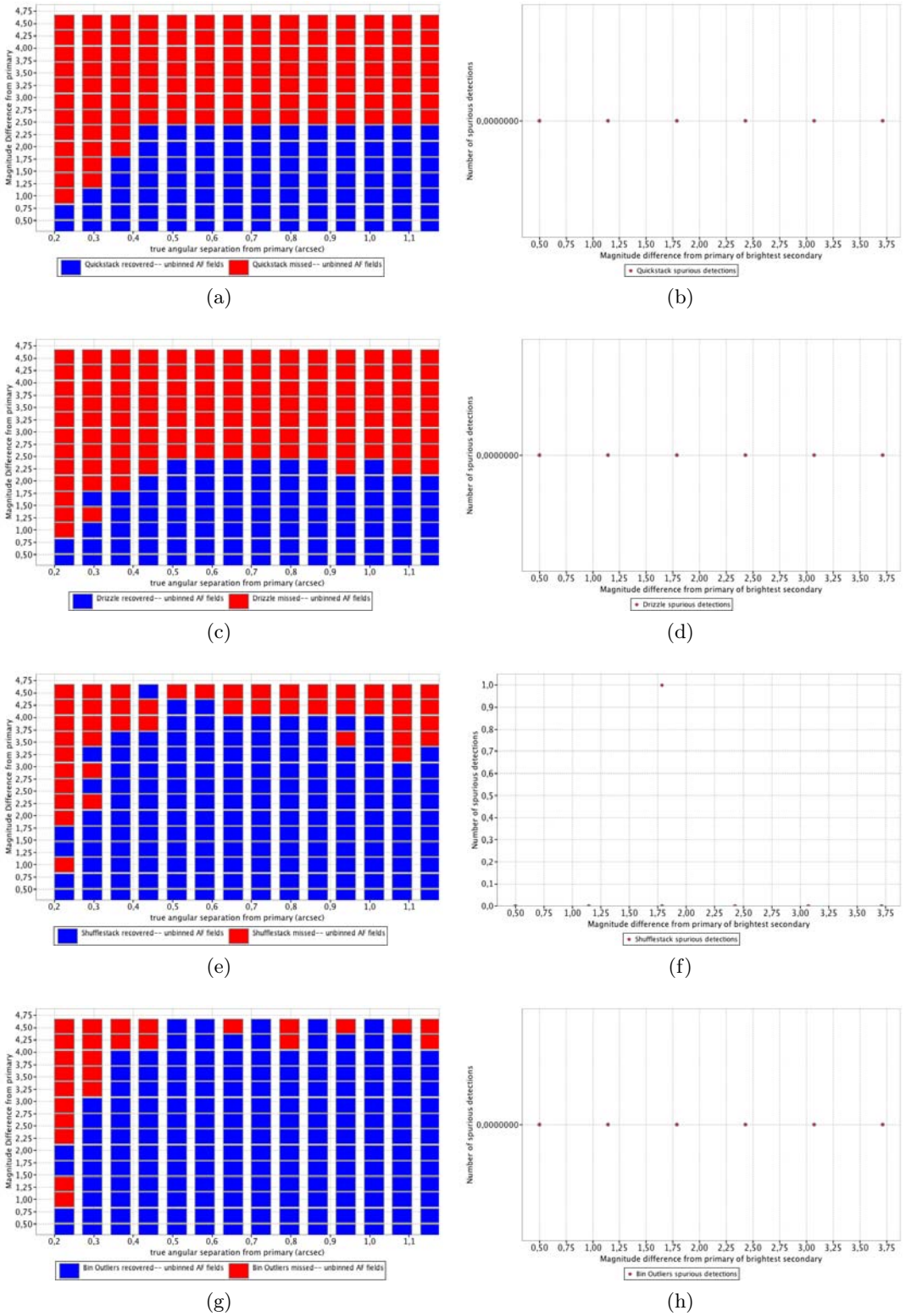


FIGURE 3.19 – Résultats obtenus par la méthode de ségrégation appliquée sur des images reconstruites pour une étoile primaire de magnitude G=12. En (a), (c), (et) et (g) les carrés bleus signifient qu'une région correspondante à l'objet secondaire a été récupérée, tandis qu'un carré rouge signifie qu'elle ne l'a pas été. En (b), (d), (f) et (h) est indiqué le nombre de fausses détections.

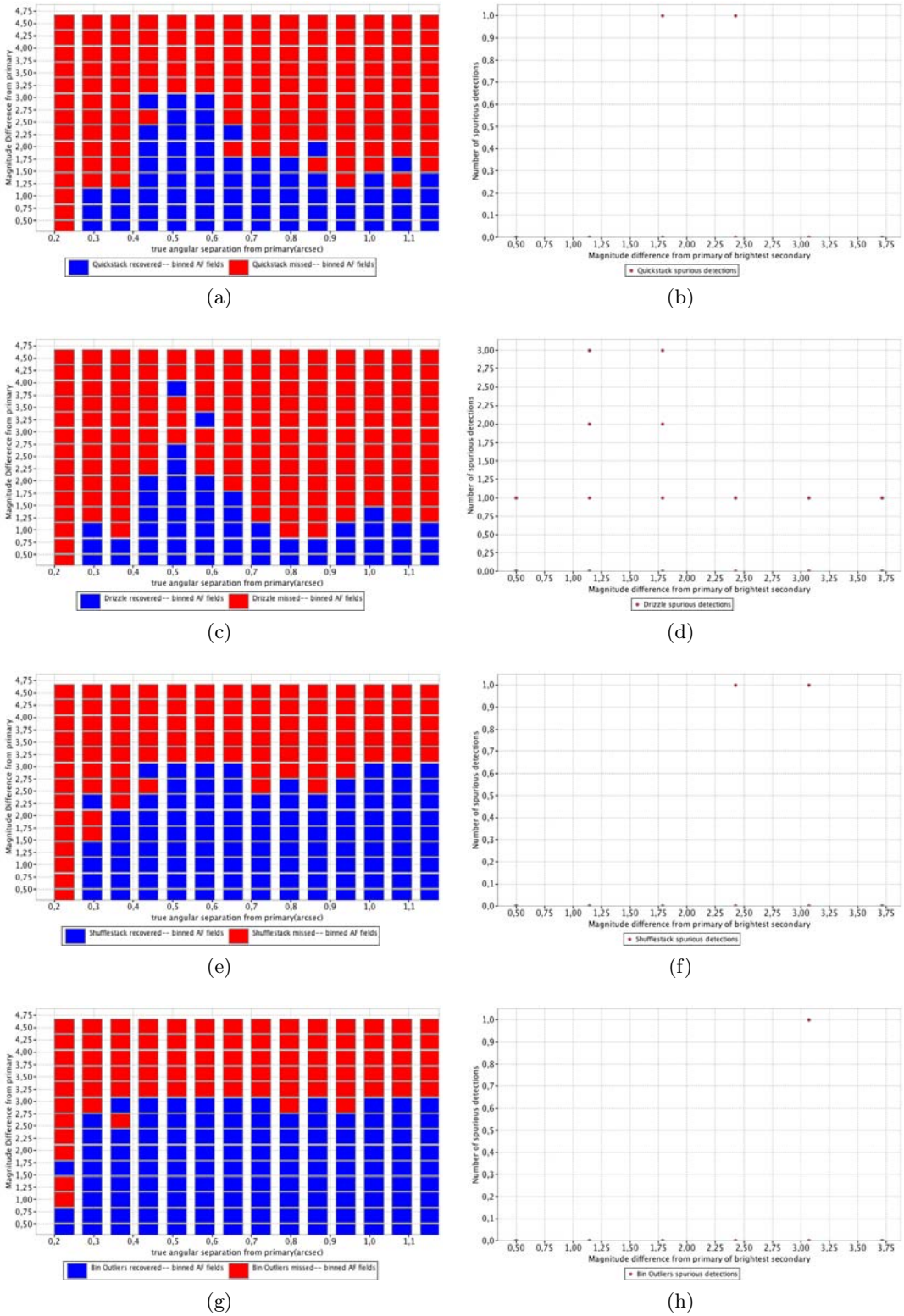


FIGURE 3.20 – Résultats obtenus par la méthode de ségrégation appliquée sur des images reconstruites pour une étoile primaire de magnitude G=19. Graphiques analogues à ceux de la Figure 3.19.

Néanmoins, les sources probablement les plus intéressantes du point de vue scientifique seront celles récupérées au delà de la limite de détection du satellite, c'est-à-dire, des sources secondaires avec $G > 20$. En analysant les résultats obtenus pour la détection de sources secondaires autour d'une primaire de magnitude $G = 19$ (présentés sur la Figure 3.20), nous avons constaté qu'il existe encore un pourcentage notable d'objets secondaires qui ont été détectés (spécialement par les derniers deux algorithmes), et que le nombre de fausses détections, bien que n'étant plus nul est relativement bas (à l'exception à l'algorithme *Drizzle*), et virtuellement nul (avec à peine une fausse détection) dans l'algorithme *BinOutliers*.

De plus, nous pouvons voir que les résultats obtenus par le code de ségrégation utilisé sur des images reconstruites par les quatre algorithmes disponibles, sont de nouveau divisés en deux classes, avec des comportements très similaires entre les algorithmes *QuickStack* et *Drizzle* d'une part, et entre *ShuffleStack* et *BinOutliers* d'autre part.

Les résultats obtenus en analysant des images reconstruites par les deux premiers algorithmes ont une efficacité faible ($\Delta G < 1$) jusqu'à ∼400 mas, une croissance élevée ($\Delta G < 3$) entre ∼400 mas et ∼650 mas et de nouveau une faible efficacité après ∼650 mas ($\Delta G < 1.75$ pour le *QuickStack* et $\Delta G < 1$ pour le *Drizzle*).

Dans des régions plus proches de la source primaire, l'inefficacité de notre algorithme de ségrégation quand il est appliqué à des images reconstruites par les algorithmes *QuickStack* et *Drizzle* est probablement due au même comportement que celui observé pour la primaire avec $G = 12$: l'émission de la source primaire « cache » les sources secondaires. D'un autre côté, le comportement inefficace sur des distances plus grandes peut être dû au fait que la source secondaire n'arrive pas à être distinguée du fond de détection (seuil η_0 qui a été déterminé), alors que dans le cas intermédiaire, sa distinction du bruit est aidée par le flux de la source primaire, qui est déjà suffisamment bas pour ne plus « cacher » la source secondaire mais est encore suffisamment fort et régulier pour aider le flux de l'objet secondaire à être plus grand que le seuil inférieur de détection.

Afin d'analyser les images reconstruites avec les deux derniers algorithmes (*ShuffleStack* e *BinOutliers*), nous pouvons voir qu'il est possible de récupérer des sources jusqu'à deux magnitudes plus faibles que la limite de détection de Gaia. Il est clair que dans la mesure où la magnitude de la primaire était $G = 19$, des sources jusqu'à $G = 22$ ont été récupérées de manière relativement stable ; le nombre de sources faibles que nous estimons pouvoir récupérer à travers du phénomène de projections optiques (présenté dans les premières sections de ce chapitre) pourra être atteint en théorie. Et ceci pour des séparations de jusqu'à ∼300 mas, dans le cas où l'algorithme *BinOutliers* serait adopté pour la reconstruction des images.

3.4.5.2 Résultats dans le cas non-idéal

L'autre régime d'utilisation de la méthode que nous avons développée est celui qui s'applique au cas où il n'existe aucun exemple correspondant à la région simulée dans la bibliothèque de données d'apprentissage. Dans ce cas, la bibliothèque sélectionnera l'ensemble de paramètres les plus proches de ceux de l'image que l'on désire analyser.

Sur la Figure 3.21 nous pouvons observer les résultats obtenus pour la ségrégation d'images générées par les reconstructions *BinOutliers* de simulations dans une direction du ciel où il n'y avait aucun exemple dans la bibliothèque. Un résultat digne d'intérêt est le fait que le nombre de fausses détections est de nouveau virtuellement nul, avec à peine une détection pour la source primaire simulée avec $G = 19$.

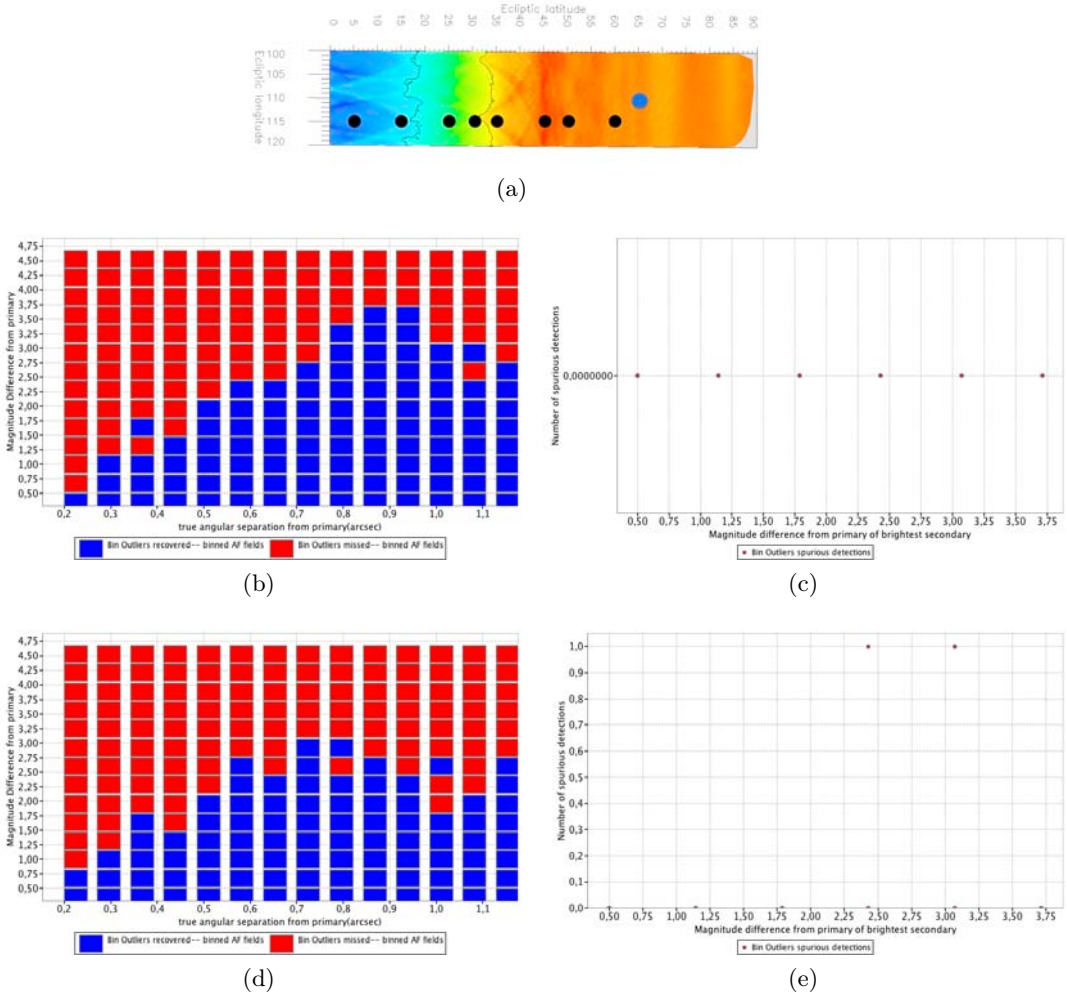


FIGURE 3.21 – Résultats obtenus par la méthode de ségrégation quand elle est appliquée sur des images reconstruites pour lesquelles il n'existe pas d'exemples proches dans la bibliothèque. Position $(\lambda, \beta) = (113°, 65°)$ indiquée en (a). Graphiques comme sur la Figure 3.19. (b) et (c) pour la primaire de $G = 15$, (d) et (e) pour $G = 19$.

Concernant les détections de sources secondaires réelles, le comportement obtenu pour les deux primaires est très similaire, avec une croissance continue suivant à peu près la même courbe jusqu'à une séparation de ∼700 mas. À partir de cette séparation, pour une primaire de magnitude $G = 15$ la détection continue à croître jusqu'à ∼950 mas pour ensuite décroître. Pendant ce temps, pour les sources secondaires autour de la primaire de magnitude $G = 19$ l'efficacité diminue légèrement après avoir atteint $\Delta G \sim 3.0$ – ceci peut être un réflet de la limite de magnitude pour laquelle *BinOutliers* est capable de reconstruire une source secondaire avec efficacité, car un seuil de magnitude $G \sim 22.0$ avait déjà été observé quand il existait un exemple identique ou très proche dans la bibliothèque (voir Figure 3.20 (g)).

Néanmoins, quand il n'existe pas dans la bibliothèque de paramètres d'apprentissage correspondants au cas étudié, comme dans le cas analysé sur la Figure 3.21, la détection des objets secondaires est bien moins efficace. Ceci n'est qu'un reflet de la coupure à 90% de probabilité pour que la région détectée soit une région formée par un objet simple – mais diminuer ce niveau augmente automatiquement les fausses détections, et comme généralement il est préférable d'avoir un plus petit nombre de sources détectées mais être sûr qu'elles soient des sources secondaires réelles, nous avons maintenu la coupe à ce niveau plus élevé.

Dans la direction du ciel que nous venons d'analyser, la couverture spatiale des observations est importante, avec une grande variété d'angles de passage. Comme on peut l'imaginer, les résultats doivent être différents s'il n'existe pas de bonnes données d'apprentissage dans la bibliothèque pour une direction correspondant une mauvaise couverture spatiale et avec des angles préférentiels pour les balayages. Sur la Figure 3.22, des résultats obtenus pour ce cas plus compliqué sont présentés – complétés avec deux exemples d'images reconstruites pour des sources localisées dans ces directions.

La première différence de comportement que l'on attend et qui apparaît ici est la croissance du taux de fausses détections.

Dans le cas d'une primaire de magnitude $G = 15$ on a eu jusqu'à 3 sources secondaires détectées parmi les 4 simulées mais sans garantie que les détections correspondent aux secondaires. Ceci signifie que la fiabilité durant une ségrégation aveugle (c'est-à-dire, une ségrégation où l'on ne connaît pas à priori les objets existants) doit être basse pour un cas comme celui-ci. Un comportement similaire peut être observé pour la primaire de magnitude $G = 19$.

Para a primária de magnitude $G = 15$ chegou-se a ter 3 fontes secundárias que podem ser detecções espúrias em imagens reconstruídas onde apenas 4 secundárias foram simuladas – isso significa que a confiabilidade durante uma segregação cega (ou seja, uma segregação onde não se conhece *a priori* quais são os objetos existentes) deve ser baixa para um caso como esse. Um comportamento similar pode ser observado para a primária de magnitude $G = 19$.

Toutefois, une analyse plus prudente prenant en considération les deux graphes, révèle que la fiabilité n'est pas si mauvaise, spécialement dans le cas de la primaire la plus faible : le graphique des fausses détections montre que peu de fausse sources sont de $\Delta G > 1.75$, mais le graphique des détections montre un nombre non négligeable

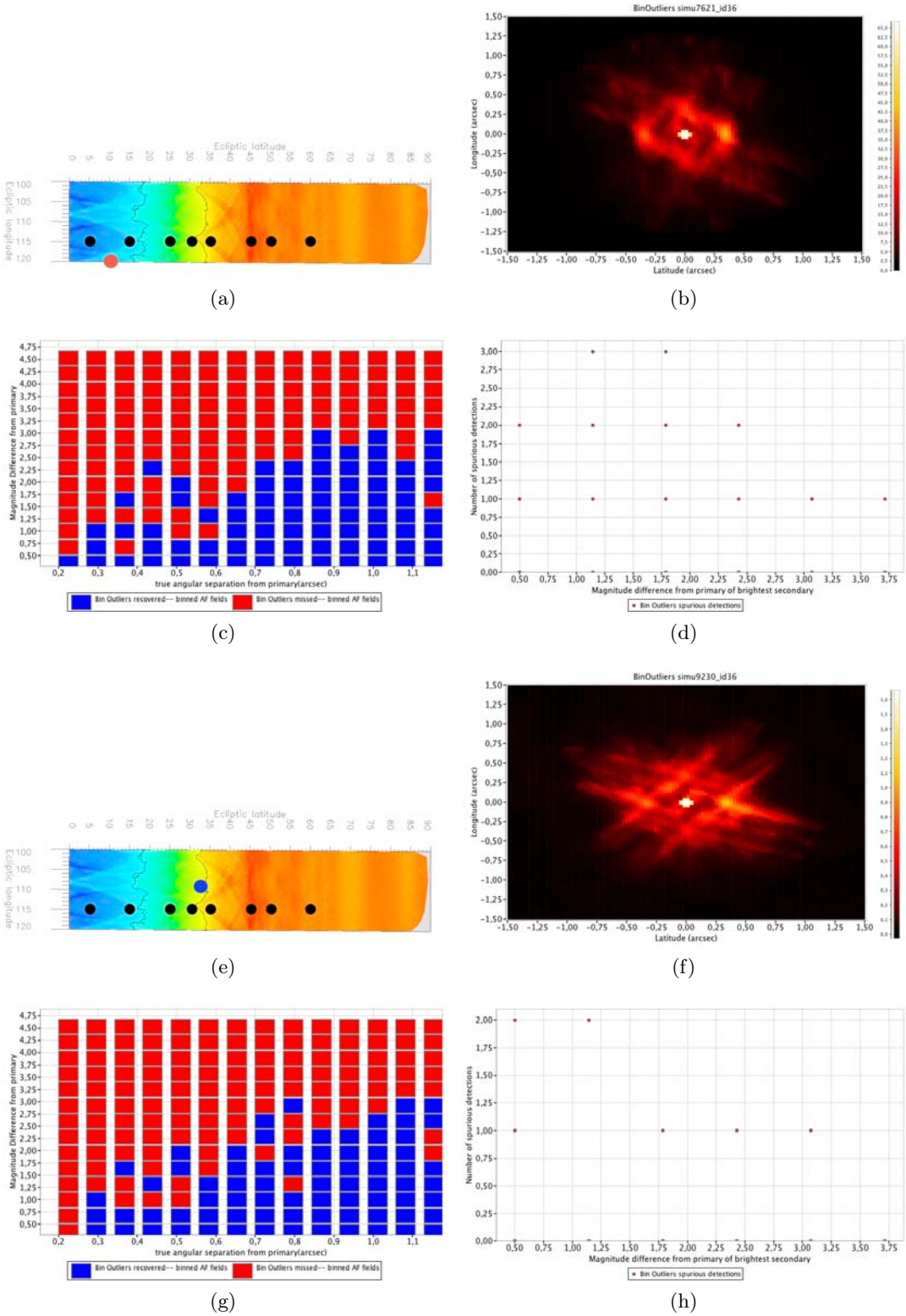


FIGURE 3.22 – Exemples de reconstructions (b) et (f) et résultats d'analyse (graphiques tels que sur la Figure 3.19) pour des projections optiques sur des primaires de magnitude $G = 15$, de coordonnées écliptiques $(\lambda, \beta) = (120°, 10°)$ et de magnitude $G = 19$, de coordonnées $(\lambda, \beta) = (110°, 32°)$.

de sources secondaires détectées sur ces magnitudes – donc, une plus grande fiabilité dans les détections.

Il est possible d'estimer à priori le taux de fausses détection de la méthode. On peut combiner cette information à priori avec la probabilité sortant du k-NN pour obtenir une probabilité finale qu'une détection soit réelle. Mais ce calcul ne se justifie que lorsqu'on ne peux pas augmenter la résolution de la bibliothèque de bases de données d'apprentissage ce qui n'est pas le cas de la mission Gaia.

3.5 Conclusions

Dans ce Chapitre, nous avons estimé le nombre de cas de projections optiques observables par Gaia dans le ciel entier. Pour cela nous nous sommes basés sur une distribution surfacique homogène des sources prises dans le catalogue GSC 2.3.2 et sur des extrapolations numériques du comptage d'objets dans des directions données du ciel, sélectionnées de manière à échantillonner des régions limites en densité, de la Galaxie. Avec ce calcul, nous estimons ce nombre à $\sim 2 \times 10^6$, pour une différence de magnitude maximum de $\Delta m = 3$ entre les sources primaire et secondaire. La plus petite valeur obtenue a été $\sim 7 \times 10^4$, tandis que dans le cas d'une Galaxie complètement d'un Bulbe, la valeur obtenue a été de $\sim 3 \times 10^{10}$ (les deux pour $\Delta m = 3$). Comme les valeurs obtenues à partir de l'analyse de catalogues n'ont pas permis de déterminer de façon précise limité de façon le nombre de projections optiques, (l'extrapolation linéaire peut être non physique, nombre faible de régions échantillonnées, hypothèse d'une distribution homogène dans le ciel), nous avons réalisé une analyse plus détaillée.

Notre analyse s'est basée sur des simulations de comptage d'étoiles effectuées avec le modèle de Galaxie de Besançon (tel que mis en oeuvre dans la bibliothèque GaiaSimu). Les simulations, complètes jusqu'à $G = 22$ ont été réalisées sur mille directions différentes de la galaxie, choisies aléatoirement avec une distribution uniforme sur la sphère et l'analyse a pris en compte une interpolation sur la sphère puis une intégration des cas de projections optiques. Le résultat moyen de dix répétitions de toute la procédure a permis de conclure que $1.6 \times 10^8 \pm 5.1 \times 10^7$ projections optiques pourront être observées dans le ciel. Des valeurs individuelles pour le bulbe et le disque en forme de courbes de probabilité et du nombre total de projections optiques à chaque magnitude de source primaire ont été aussi présentées.

Nous avons également décrit la méthode que nous avons développée pour analyser des images reconstruites en réalisant la ségrégation de cette image en termes de régions de pixels de l'image – chaque région peut être formée par une source unique, des sources multiples mélangées ou des artefacts générés par les algorithmes de reconstruction. Cette méthode, appelée Ségrégation Éduquée d'Images (*Educated Image Segregator*, ou EIS) est capable d'apprendre et de s'auto-adapter de manière automatique, à condition qu'elle soit entraînée avec des exemples des différents types de région que l'on désire ségréguer. Elle est basée sur des algorithmes robustes et rapides pour une binarisation sur de multiples seuils, une détermination de connecti-

vité dans un même seuil et de la connectivité entre les différents seuils basée sur des graphes, sur l'analyse de ces graphes pour la détermination du meilleur seuil pour la ségrégation d'une région donnée, d'un classificateur supervisé pour la production de la liste finale de régions ségrégées et d'une base de donnée d'apprentissage pour ce classificateur.

Nous avons présenté les résultats obtenus avec l'utilisation de la méthode EIS dans des simulations GIBIS de projections optiques et de reconstructions d'images avec les quatre algorithmes de reconstruction dont la mise en oeuvre est actuellement disponible. Ces résultats ont montré que le meilleur algorithme de reconstruction d'images en termes du nombre de sources secondaires récupérées et du minimum de fausses détections est le *BinOutliers*.

De plus, dans le cas où il existe un ensemble de données d'apprentissage dans la bibliothèque avec des caractéristiques (direction du ciel, principalement) similaires à celles de l'image que l'on veut ségréguer, nous avons montré que l'efficacité de la ségrégation est très bonne, arrivant à des valeurs de $\Delta m \sim 3.0$ pour une primaire de $G = 19$, et $\Delta m \sim 4.5$ pour une primaire de $G = 12$, et ceci jusqu'à de petites séparations, autour de ~ 300mas, virtuellement sans fausses détections.

Pour une ségrégation d'images sans données d'apprentissage d'example, pour une latitude écliptique de la source primaire $|\beta| < 45°$, le nombre de sources secondaires obtenues avec notre méthode est moins fiable à cause d'un plus grand nombre de fausses détections. Cependant, pour le reste du ciel ($|\beta| > 45°$), la détection de sources secondaires, bien que moins efficace pour de petites différences angulaires que dans le cas où il existerait de bonnes données d'apprentissage dans la bibliothèque, permet la détection de sources secondaires jusqu'à $\Delta m \sim 3.0$ avec une grande fiabilité (peu de fausses détections).

Les résultats obtenus dans ce Chapitre montrent, qu'en plus de pouvoir analyser des images pour comprendre l'origine de leur solutions astrométriques perturbées, il sera aussi possible de compléter le catalogue Gaia avec des millions de sources qui se trouvent au-dessus de la limite de détection du satellite.

Chapitre 4

Simulations de Galaxies dans le modèle d'Univers de Gaia

"*'No machine can lie,' said Father Brown;*
'nor can it tell the truth.'" G. K. Chesterton[1]

Sommaire

En plus des projections optiques, les galaxies avec une structure non résolue en étoiles sont un autre cas (et le second le plus fréquent) dans lequel les solutions astrométriques fournies par Gaia peuvent être perturbées, étant donné que leurs profils de brillance superficiels ne sont pas compatibles avec des sources ponctuelles. Pour que nous puissions estimer le nombre de galaxies de ce type qui seront observées et développer et tester les méthodes d'analyse pour cet type d'objet, il est nécessaire avant tout de pouvoir simuler à la fois la distribution de ces galaxies sur la sphère céleste et aussi leur observation par Gaia.

Dans ce Chapitre nous présenterons le simulateur de catalogues de galaxies appelé JStuff, que nous avons mis en oeuvre dans le modèle d'Univers de Gaia (Krone-Martins et al, 2008c) à partir du travail de (Bertin & Arnouts, 1996). Nous présenterons également brièvement le code MAGIL (Gavras et al, 2010), qui permet de générer des images de ces objets étendus dans le plan focal du satellite et auquel nous avons collaboré. Finalement, nous présenterons une estimation de la limite supérieure du nombre de galaxies non résolues qui pourront être observées par Gaia.

1. « « Aucune machine ne peut mentir, » dit le Prêtre Brown ; « mais encore moins dire la vérité. » », in *The Mistake of the Machine*, 1913.

4.1 Introduction

Comme nous l'avons vu dans l'introduction de cette thèse, toutes les sources du ciel jusqu'à une certaine magnitude seront observées part Gaia. Parmi ces sources se trouvent les galaxies qui ne peuvent pas être résolues en étoiles individuelles, c'est-à-dire, celles qui sont compactes et/ou qui sont suffisamment distantes de la Terre pour qu'il ne soit pas possible de distinguer séparément leurs étoiles, mais en même temps suffisamment proches pour qu'il soit possible de percevoir l'effet intégré de ces étoiles, qui forment leurs différentes composantes structurelles.

Ces galaxies non résolues sont, de la même façon que les sources secondaires présentées dans le Chapitre précédent, des objets qui vont présenter un *goodness of fit* anormal lors de leur traitement par le pipeline de Gaia étant donné que celui-ci a été développé pour traiter des sources ponctuelles. Comme nous l'avons vu dans le Chapitre 1, ces galaxies sont très intéressantes du point de vue scientifique, étant donné qu'il sera possible d'analyser leur morphologie pour la première fois dans l'histoire de l'humanité justement à partir des observations réalisées par Gaia.

Pour que nous puissions étudier ces objets, appelés dans le contexte de la mission Gaia des galaxies non résolues, il a été nécessaire de mettre en place dans le modèle d'Univers du satellite des codes permettant de produire des simulations. Ceci a été fait dans un premier temps à travers l'implémentation d'un simulateur de catalogues de galaxies cohérent, c'est-à-dire capable de reproduire les mêmes galaxies dans le ciel toutes les fois qu'une même région est observée, et qui statistiquement reproduise les fonctions de luminosités actuellement connues pour ces objets. Ce simulateur s'appelle JStuff (Krone-Martins et al, 2008c), et sera décrit avec plus de détails en section 4.2.

De plus, avec l'évolution du projet, nous avons noté le besoin d'utiliser des images réalistes de ces galaxies non résolues pour les reconstructions d'images 2D. Pour cela, nous avons collaboré à la spécification et au design d'un simulateur nommé MAGIL (Gavras et al, 2010) qui utilise et transforme des images prototypiques d'une bibliothèque d'images. Ce simulateur sera brièvement présenté en section 4.3.

4.2 Simulation de catalogues de galaxies

Le code JStuff est une implémentation en langage Java du code Stuff (Bertin & Arnouts, 1996). Son adaptation à Gaia s'est faite en deux phases distinctes : une adaptation pionnière pour une phase prototype du satellite a été réalisée par Dollet (2004) puis l'implémentation finale a été réalisée par Krone-Martins et al (2008c). Le code final est en Java, adapté aux normes de développement du DPAC, et il a été adapté pour permettre son intégration au modèle d'Univers de Gaia.

Les modifications importantes qui différencient JStuff de Stuff concernent les deux points suivants. Tout d'abord JStuff peut être appelé par n'importe lequel des simulateurs de Gaia mais en plus il réalise des simulations consistantes : quand on demande plusieurs fois à JStuff une simulation dans une région donnée du ciel, on obtient à chaque fois la même liste d'objets simulés (à condition que les paramètres

du simulateur et du modèle cosmologique soient maintenus sans altérations). Cette caractéristique de JSuff est très importante car elle permet que divers passages soient simulés sur les mêmes objets extragalactiques, quelque chose d'impossible avant notre implémentation.

4.2.1 Principe de fonctionnement

Le code JStuff a pour objectif de générer une liste de galaxies (et leurs paramètres) qui pourront être observées dans une direction donnée et jusqu'à une magnitude limite désirée. Dans la version actuelle du code, la distribution des positions des galaxies sur la sphère céleste est considérée comme aléatoire, étant générée à partir d'un échantillonnage uniforme sur deux dimensions. Le nombre total d'objets qui doivent être générés dans la région simulée est donc produit à partir d'une distribution poissonienne, tandis que le nombre de galaxies de chaque type de Hubble dans chaque intervalle de magnitude est tiré à partir de fonctions de luminosité de Schechter qui possèdent des coefficients indépendants pour chaque type de Hubble.

Le premier pas de la simulation est de découper la sphère en régions géométriques de forme quelconque dans lesquelles les listes d'objets générées sont statistiquement équivalentes. Pour des simulations sur de petites régions on peut utiliser des cercles ou des rectangles, mais des triangles sont utilisés quand on réalise des simulations sur une large échelle.

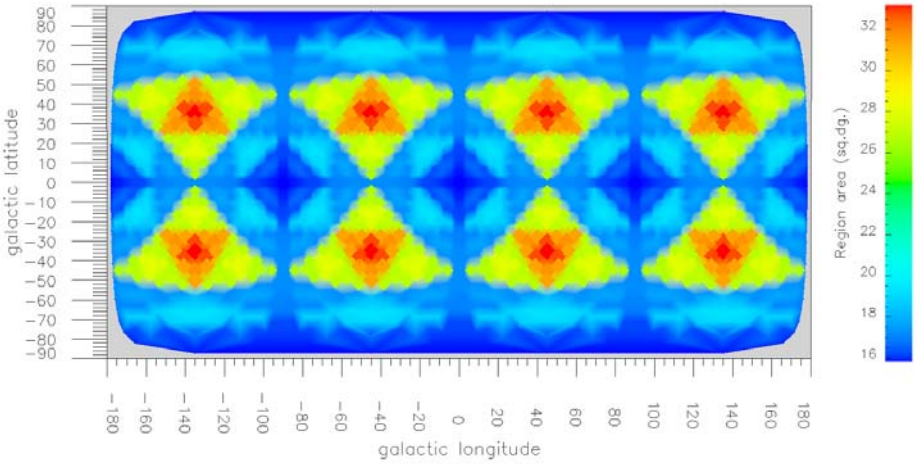


FIGURE 4.1 – Aires des triangles générés par une décomposition de la sphère par le *Hierarchical Triangular Mesh*.

La répartition triangulaire est réalisée à partir d'une sous-division de la sphère appelée *Hierarchical Triangular Mesh* (Barrett, 1995; Kunszt et al, 2000), qui est une sous-division en triangles sphériques de mêmes dimensions environ. Cette technique génère une structure de données du type quad-tree pour la sphère, c'est-à-dire, à

un premier niveau, la sphère se divise en huit triangles sphériques (ce niveau est formé par l'octaèdre inscrit dans la sphère). Puis à un second niveau, chacun de ces triangles se divise en quatre autres triangles. Au troisième niveau chaque triangle est de nouveau divisé en quatre, et ainsi de suite à chaque niveau additionnel jusqu'à la taille souhaitée par l'utilisateur. La Figure 4.1 montre la distribution des aires de ces triangles pour une décomposition au niveau 4.

Pour que le code JStuff génère toujours la même liste d'objets dans une région donnée du ciel, l'indice HTM de cette région est utilisé pour déterminer le germe du générateur de nombres aléatoires. Comme ce générateur est toujours initialisé à partir du même germe pour une même région, le catalogue d'objets sera toujours le même.

4.2.2 Modèle d'objets

Mais ce que JStuff doit générer dans ces régions sont des objets extragalactiques, et la représentation de ces corps célestes dans un ordinateur doit être décrite d'une manière ou d'une autre. Cette description, est réalisée à partir d' « objets » computationnels, étant donné que selon les normes du DPAC, tous les codes de Gaia doivent être écrits en langage Java. Un objet computationnel, dans ce cas, peut être compris comme une abstraction d'un objet réel dans la mémoire d'un ordinateur, c'est-à-dire, est quelque chose qui possède des « caractéristiques » (attributs), comme de la couleur, et qui peut réaliser des « actions » (mèthodes), tel qu'exploser.

Une galaxie, par exemple, n'est seulement qu'une des représentantes de la « classe » galaxie, et il existe des centaines de millards de galaxies dans l'univers qui, bien qu'elles puissent être du même type, elles ont eu des naissances et une évolutions différentes, donc elles possèdent des caractéristiques avec des valeurs différentes. Chacun de ces représentants est dénommé instance, c'est-à-dire, au moment de créer un objet dans la mémoire de l'ordinateur, on dit qu'une certaine classe est instanciée. Donc, ces représentations doivent être modelées avant que la programmation en elle-même commence.

Pour modeler les objets extragalactiques du simulateur de catalogues de galaxies non résolues nous nous sommes inspirés dans le modèle préalablement créé pour le simulateur d'étoiles, qui est une version en Java du modèle de Galaxie de Besançon (Robin et al, 2007). De cette manière il nous fut possible de proposer l'adoption d'une structure cohérente dans le modèle d'univers de Gaia, lequel jusqu'à l'époque de l'implémentation du JStuff, mi-2006, n'existait pas encore.

En suivant des modèles d'objets mis en oeuvre antérieurement, nous avons implémenté la classe `UnresolvedGalaxy` en adoptant des classes abstraites - en programmation orientée objets, ce type de classe représente des concepts abstraits, qui ne sont jamais instanciés (c'est-à-dire, créés dans la mémoire de l'ordinateur) et qui ne servent qu'à définir des modèles de comportements qui sont implémentés dans des classes spécifiques, héritières de la classe abstraite (ce concept est relativement simple, une classe « être vivant », par exemple, ne devrait jamais être instanciée, au contraire de la « tomate », du « rat », de l'« homme », etc.).

La Figure 4.2 représente la hiérarchie d'héritage qui a été adoptée dans le modèle d'Univers de Gaia, avec les classes que nous avons implémentées indiquées en rouge. La classe `AstroSource` dans ce diagramme est un bon exemple du concept d'une classe abstraite : bien qu'il existe des sources astronomiques, ce nom ne définit pas un objet réel mais au contraire un concept d'un objet réel, tandis que ce qui existe réellement ce sont les galaxies, les astroïdes, les quasars, les étoiles, etc.

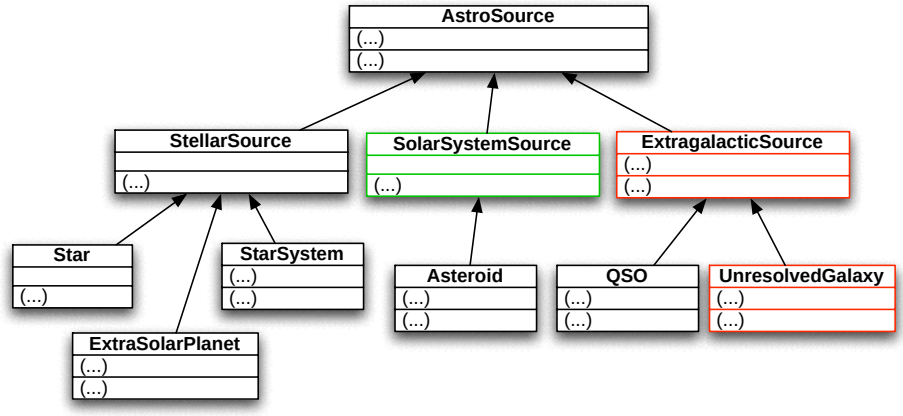


FIGURE 4.2 – Diagramme d'héritage de la classe AstroSource.

La classe `UnresolvedGalaxy` est celle qui décrit les galaxies non résolues, et qui contient toutes les méthodes et les attributs des galaxies qui sont simulées. Elle hérite de toutes les méthodes et attributs d'une classe abstraite, appelée `ExtragalacticSource`, qui est une classe père des classes qui travaillent avec des galaxies et avec des quasars, tel qu'indiqué sur le diagramme UML[2] de la Figure 4.2.[3]

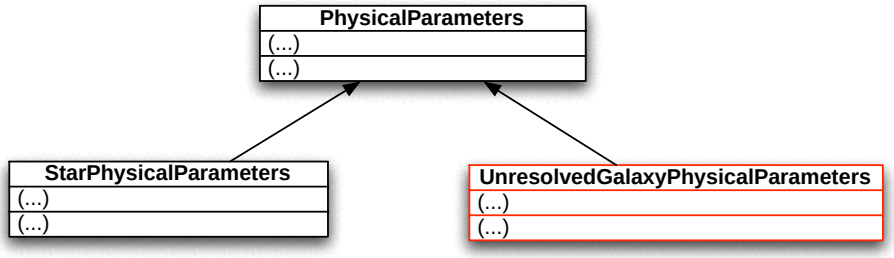


FIGURE 4.3 – Diagramme d'héritage de la classe PhysicalParameters.

Les caractéristiques physiques de chaque galaxie simulée sont manipulées (stockées et récupérées) au moyen d'une spécialisation de la classe `PhysicalParameters` que

2. Les diagrammes UML, ou *Unified Modeling Language* sont une forme standardisée pour indiquer la communication et l'interdépendance entre des objets.

3. Nous avons présenté premièrement les versions des diagrammes sans les noms des attributs et des méthodes de chaque classe pour faciliter une compréhension général de l'héritage.

nous dénommons `UnresolvedGalaxyPhysicalParameters`, et qui est représentée sur la Figure 4.3. C'est dans cette classe que nous stockons l'instance d'une classe qui est très importante pour n'importe quelle galaxie, dénommée `HubbleType`, et que verrons dans plus de détails dans la prochaine section.

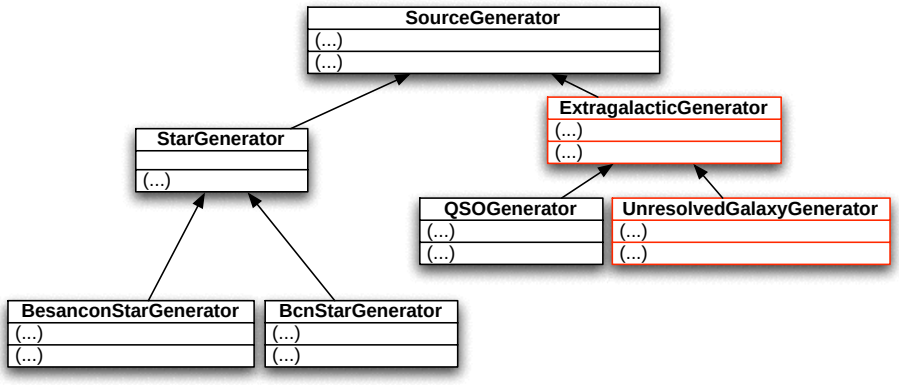


FIGURE 4.4 – Diagramme d'héritage de la classe `SourceGenerator`.

Enfin, la classe qui est responsable pour les simulations des catalogues proprement dits est dénommée `ExtragalacticGenerator`, et est représentée sur la Figure 4.4. La classe qui génère les catalogues au moyen du JStuff est l'`UnresolvedGalaxyGenerator`.

Diagrammes étendus

Les diagrammes de classe en UML peuvent aussi être utilisés pour indiquer quels attributs et méthodes les classes possèdent, en plus des relations entre les classes. Sur le diagramme présenté sur la Figure 4.5, nous voyons l'héritage de la classe `UnresolvedGalaxy` avec quelques uns des attributs et des méthodes représentés.

La classe abstraite `AstroSource`, sur ce diagramme, possède comme attributs aussi bien des caractéristiques simples que n'importe quel objet dans un simulateur doit posséder, que le code identificateur de l'objet (`id`) ou son type (`idType`), ainsi que des attributs plus complexes qui en fait sont des instances d'autres classes. Parmi ces attributs complexes, nous avons les paramètres astrométriques, photométriques, physiques, de forme, et spectroscopiques. Les instances de ces attributs plus complexes apportent aussi avec elles les méthodes, mais pour être utilisées extérieurement (c'est-à-dire par un code qui, hypothétiquement, pourrait instancer la classe `AstroSource`), celles-ci doivent être appelées par des méthodes de la classe elle même (un concept appelé encapsulement). Donc, des méthodes pour pouvoir traiter certains de ces attributs sont aussi incluses.

La classe abstraite `ExtragalacticSource`, qui est une spécialisation de la classe `AstroSource` étend cette dernière vers des paramètres plus caractéristiques des objets extragalactiques, comme par exemple, le *redshift*. Finalement, la classe qui est instancée pour représenter les galaxies non résolues, mais qui hérite des deux

dernières classes commentées, est l'`UnresolvedGalaxy`. Cette classe en instance une autre appelée `UnresolvedGalaxyPhysicalParameters`, qui traite des paramètres physiques de l'objet.

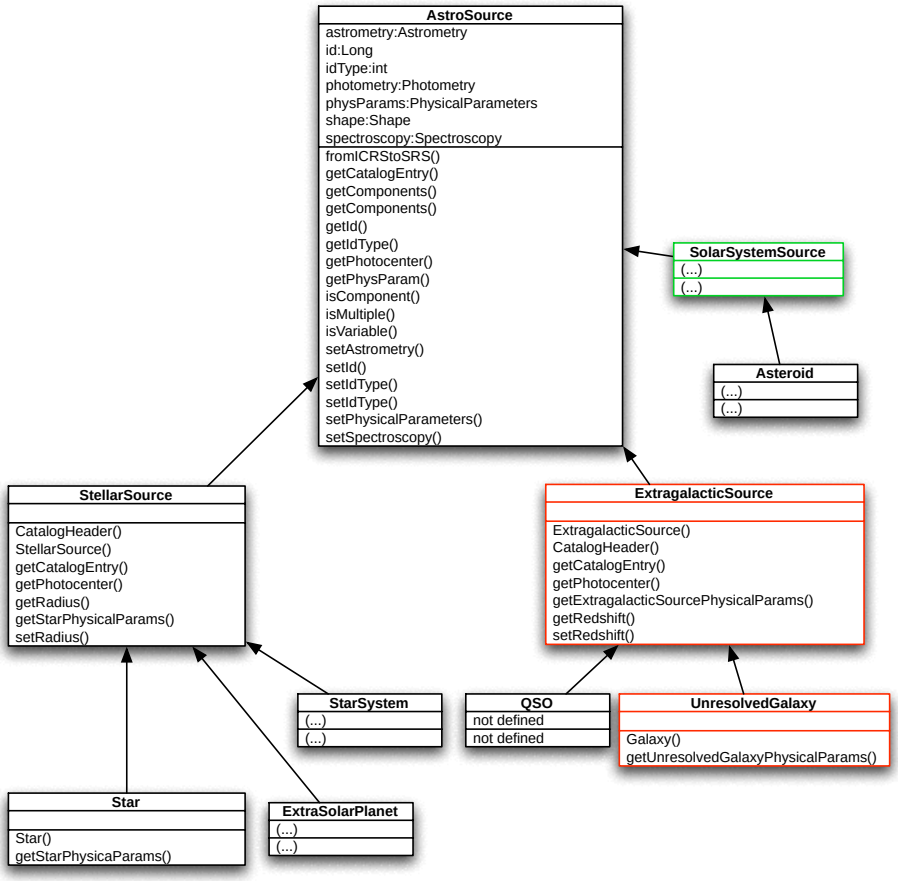


FIGURE 4.5 – Diagramme d'héritage de la classe `AstroSource`.

L'héritage de la classe `UnresolvedGalaxyPhysicalParameters` est présentée avec ses méthodes et attributs dans le diagramme de la Figure 4.6. Cette classe est responsable de la gestion des paramètres physiques des objets. La classe abstraite père `PhysicalParameters` ne possède que la masse de l'objet comme attribut, tous les autres attributs étant spécifiques à chaque type d'objet astronomique sont laissés pour les classes fils. Dans `UnresolvedGalaxyPhysicalParameters` nous avons comme attributs simples les tailles, l'éllipticité, l'angle de position (angle projeté sur le ciel), du bulbe et du disque de la galaxie. Cette classe possède aussi un attribut plus complexe, instance d'une autre classe, qui représente des paramètres définis spécifiquement par le type de Hubble de la galaxie. Cet attribut est une instance de la classe `HubbleType`, qui de son côté possède des attributs simples et complexes, représentant par exemple les fonctions de luminosité et de distributions spectrales d'énergie pour le disque et pour le bulbe.

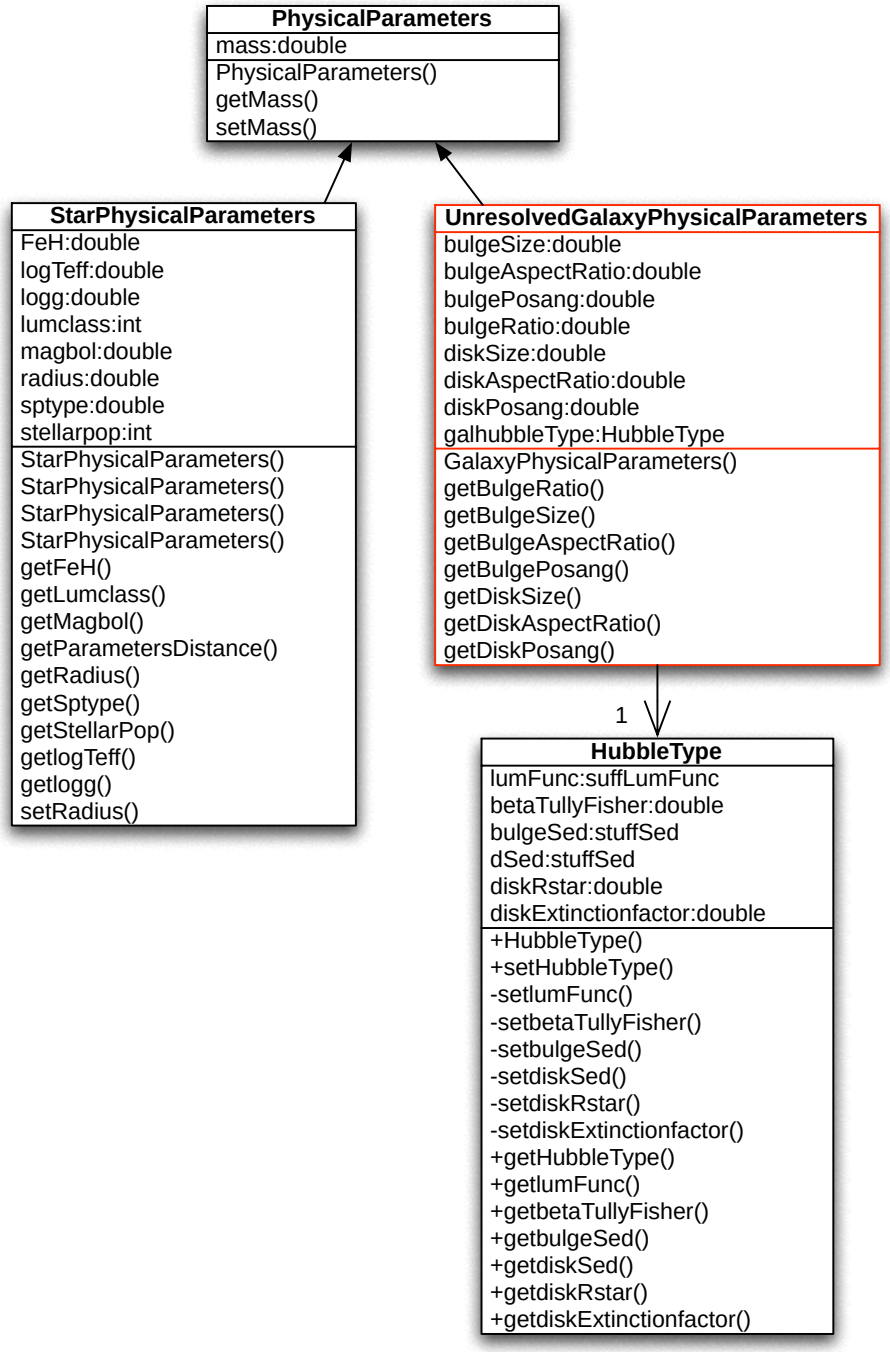


FIGURE 4.6 – Diagramme d'héritage de la classe `PhysicalParameters`.

La classe la plus centrale du simulateur est dite `ExtragalacticGenerator` car elle a la tâche de générer les instances de la classe `UnresolvedGalaxy` qui doivent être rencontrées dans la direction du ciel demandée dans la simulation. Comme nous l'avons vu dans la section précédente, cette classe est héritière de `ExtraGalacticGenerator` et `SourceGenerator`, desquels elle hérite l'interface de la méthode de simulation, qui est commune à tous les simulateurs implémentés dans le modèle d'Univers. Le diagramme pour cette classe peut être vu sur la Figure 4.7, tandis que la méthode de simulation proprement dite est présentée dans la prochaine section.

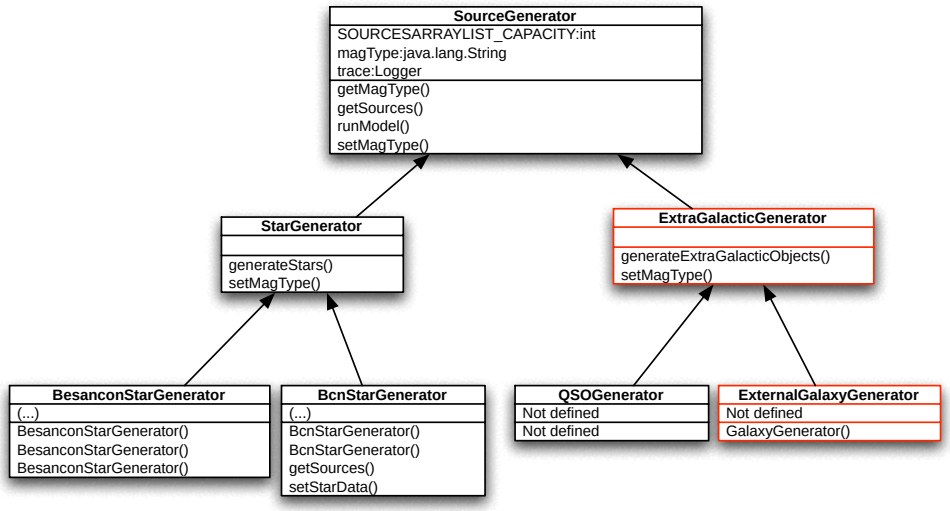


FIGURE 4.7 – Diagramme d'héritage de la classe `SourceGenerator`.

4.2.3 Méthode de simulation

La méthode adoptée pour la simulation de catalogues de galaxies non résolues est une adaptation du code Stuff, à l'origine développé par Bertin & Arnouts (1996) et Erben et al (2001). Notre re-implémentation de la méthode décrite dans ces travaux est appelée JStuff et a été basée sur les codes originaux écrits en C par E. Bertin. En plus de permettre l'inclusion de simulations de galaxies de façon cohérente dans le modèle d'Univers de Gaia, notre travail a également permis la détection et la correction d'une petite erreur existante dans le code original.[4]

Le code JStuff simule un catalogue de galaxies pour une certaine région du ciel au moyen d'un tirage réalisé à partir d'une distribution poissonienne du nombre de galaxies d'un type donné de Hubble (E, S0, Sab, Sbc, Scd, Irr, QSFG) supposant une fonction de luminosité[5] de Schechter (Schechter, 1976) du type :

4. E. Bertin, *private communication*, 13 nov. 2006.
5. Le code adopte une fonction de Schechter sans évolution, ce qui signifie qu'aucune dépendance du type morphologique au *redshift* n'a été introduite.

$$\phi(M)dM = \phi^* exp(0.92(\alpha+1)(M^*-M) - exp(0.92(M^*-M)))dM$$

Les coefficients de cette fonction de Schechter qui sont adoptés actuellement pour chaque type de galaxie ont été obtenus à partir de l'étude de Fioc & Rocca-Volmerange (1999), et peuvent être retrouvés dans la tableau 4.1.

Tipo	Φ* [Mpc−3]	M* [mag]	α	BT (bulge/total)
E2	1.91×10^{-3}	-20.02	-1.0	1.0
E-S0	1.91×10^{-3}	-20.02	-1.0	0.9
Sa	2.18×10^{-3}	-19.62	-1.0	0.57
Sb	2.18×10^{-3}	-19.62	-1.0	0.32
Sbc	2.18×10^{-3}	-19.62	-1.0	0.32
Sc	4.82×10^{-3}	-18.86	-1.0	0.16
Sd	9.65×10^{-3}	-18.86	-1.0	0.049
Im	9.65×10^{-3}	-18.86	-1.0	0.0
QSFG	1.03×10^{-2}	-16.99	-1.73	0.0

TABLE 4.1 – Paramètres des fonctions de luminosité utilisés actuellement dans JStuff. E2 représente une elliptique rouge et E-S0 des elliptiques normales ou S0s (Robin, 2010). Le rapport des luminosités BT (Bulbe/Total) est utilisé pour calculer le taux de luminosité entre le bulbe et la luminosité totale.

Ce tirage est réalisé dans une série d'éléments de volume identiques, définis par une figure géométrique, un *redshift* moyen et son extension dans l'espace de *redshift* de sorte à couvrir un domaine spécifié de *redshifts*. Généralement, comme JStuff est appelé à partir du modèle d'Univers de Gaia, des simulations sont adoptées avec des régions triangulaires, mais nous avons aussi mis en place des formes pour exécuter JStuff en mode *stand-alone*, et donc des formes rectangulaires et circulaires (formant un type de cône dans l'espace de *redshift*) peuvent être demandées.

Donc, chaque galaxie dans l'élément de volume actuel peut être décrit par somme de deux composants : un bulbe et un disque[6]. Le premier est décrit par une loi de Vaucouleurs :

$$\mu_S(r) = M_S + 8.3268 \left(\frac{r}{r_e}\right)^{1/4} + 5\log(r_e) + 16.6337$$

où M_S est la magnitude absolue dans la bande B du bulbe, et r_e son rayon effectif en parsecs. Pour des galaxies elliptiques, le rayon effectif du bulbe est calculé à partir de sa magnitude absolue au moyen d'un rapport décrit par Binggeli et al (1984) :

$$r_e = \begin{cases} \left(\frac{h}{0.5}\right)^{0.5} 10^{3.5-0.3(M_S+20.5)} & \text{si } M_S < -20.5 \\ \left(\frac{h}{0.5}\right)^{-0.5} 10^{3.5-0.1(M_S+20.5)} & \text{si } M_S \geq -20.5 \end{cases}$$

6. Une image peut être simulé par un autre logiciel de E. Bertin, le SkyMaker (voir section 4.3).

L'aplatissement q est sélectionné de manière aléatoire à partir d'une distribution normale entre 0.3 et 1.0, avec une moyenne $\langle q \rangle = 0.65$ et un écart type $\sigma_q = 0.18$. Ces valeurs ont été tirées de Sandage et al (1970).

Quant à la seconde composante, le disque, quand il existe (c'est-à-dire, quand la galaxie n'est pas une galaxie elliptique), il est décrit par un profil exponentiel :

$$\mu_D(r) = \mu_0 + 1.0857 \left(\frac{r}{r_h} \right)$$

où r_h est l'échelle et μ_0 la brillance superficielle centrale. Cette dernière valeur est sélectionnée de manière aléatoire à partir d'une distribution gaussienne avec une moyenne $\langle \mu_0 \rangle = 21.65 \text{mag.arcsec}^{-2}$ et un écart type $\sigma_{\mu_0} = 0.35$ si la magnitude absolue est $M \leq -17$, et une moyenne $\langle \mu_0 \rangle = 21.65 + 0.7(M + 17) \text{mag.arcsec}^{-2}$ sinon. L'échelle du disque r_h est obtenue à partir du rapport suivant :

$$r_h = exp(-4.713 + 0.2(\mu_0 - M_D))$$

où M_D est la magnitude absolue du disque vu *face-on* dans la bande B.

L'inclination et l'angle de visée sont sélectionnés de manière aléatoire à partir d'une distribution uniforme. Pour obtenir M_S pour une M donnée, on utilise un ajustement empirique rencontré chez Simien & de Vaucouleurs (1986) :

$$M_S = M + 0.80 + 0.145T + 0.0284T^2 + 0.00267T^3$$

où T est le type de la classification de Hubble.[7]

L'extinction interne dans le disque est simulée en utilisant une courbe d'extinction empirique dérivée pour le Grand Nuage de Magellan, et des corrections K+e sont appliquées en utilisant les ajustements polynomiaux de Metcalfe et al (1991).

Le *redshift* de chaque galaxie est alors choisi de façons aléatoire dans l'élément de volume qui est simulé. La distance diamètre angulaire est calculée par :

$$d_A = \frac{d_L}{(1+z)^2}$$

où

$$d_L = \frac{c}{H_0 q_0^2} \left(q_0 z + (q_0 - 1) \left(\sqrt{2q_0 z + 1} - 1 \right) \right)$$

Et ainsi on initie le calcul des magnitudes dans chacune des bandes demandées par l'utilisateur. Finalement, quand toutes les galaxies dans cet élément de volume en *redshift* ont été calculées on passe à l'élément suivant, étant entendu que le programme arrête la simulation quand le *redshift* $z = 0$ est atteint.

Un diagramme de flux de la méthode est présenté en Figure 4.8.

7. Pour $T \leq 4$, $M_S = M$.

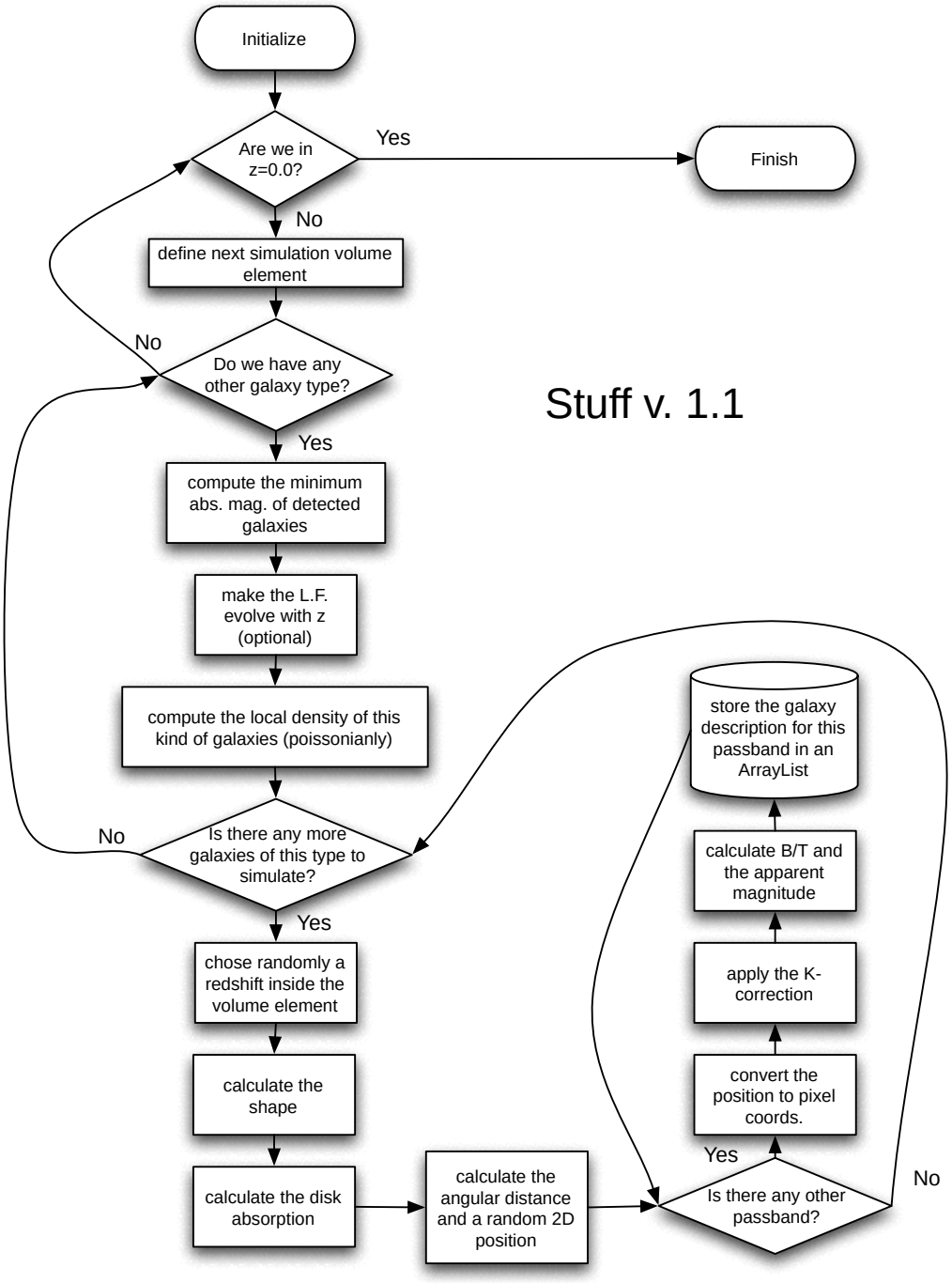


FIGURE 4.8 – Diagramme de flux de la méthode de simulation adoptée par le simulateur JStuff. Le *Volume Element* est une région du ciel tri-dimensionnelle qui est définie par une figure géométrique quelconque, un *redshift* central et un intervalle de *redshift*.

4.2.4 Validation de l'implémentation Java

Pour valider l'implémentation Java nous avons réalisé des tests comparant, pour les mêmes fichiers de paramètres d'entrée du code, les résultats obtenus avec les deux implémentations de la méthode de simulation (JStuff et Stuff). Ces comparaisons ont démontré que les listes d'objets obtenues à partir de l'utilisation des deux codes sont compatibles à 1σ, comme on peut le voir sur la Figure 4.9.

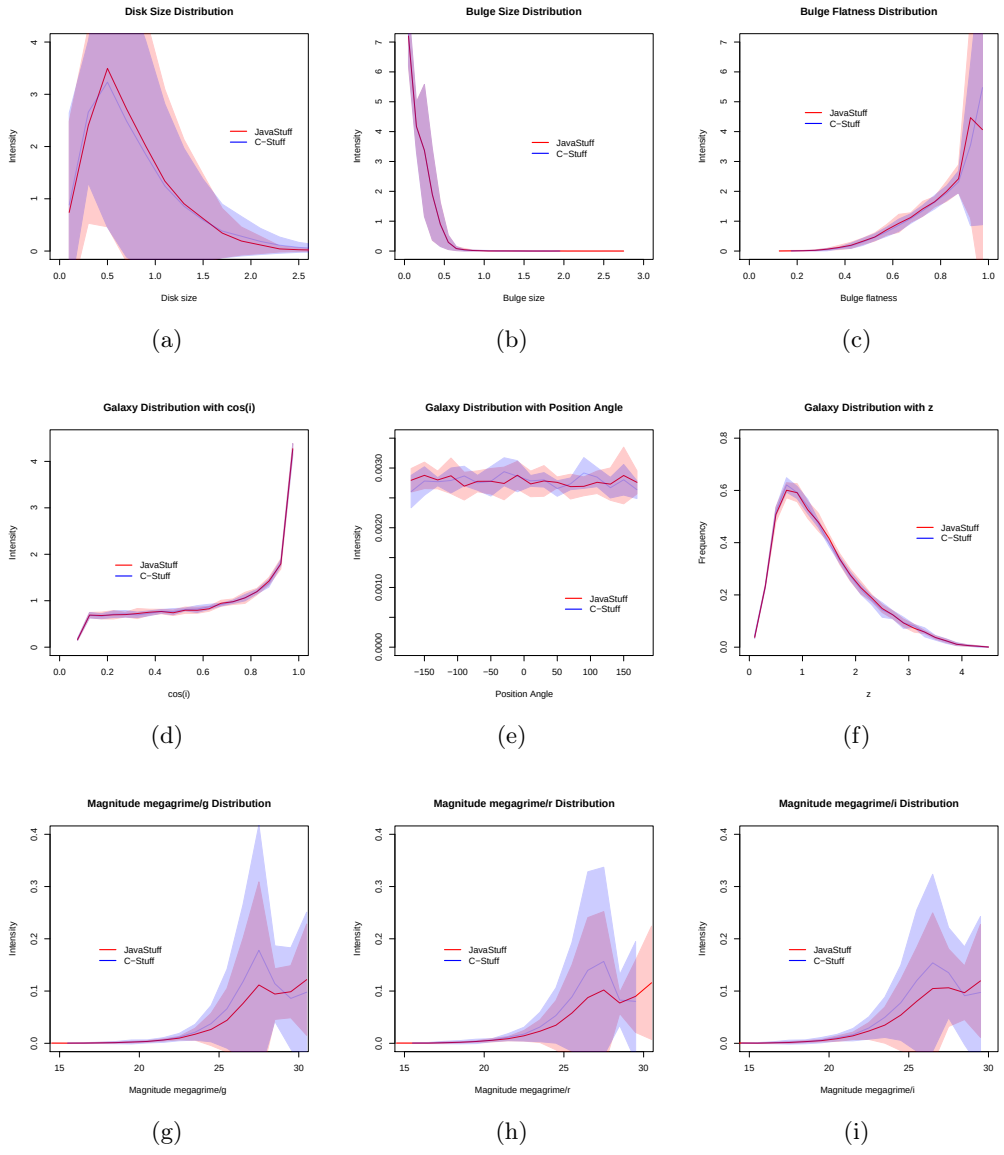


FIGURE 4.9 – Comparaison de distributions de paramètres obtenus avec JStuff et avec Stuff.

En plus de réaliser ce test de validation ponctuel, le code JStuff incorpore aussi des tests automatiques. Ces codes sont écrits en JUnits, c'est-à-dire, ce sont des tests unitaires : les méthodes individuelles des objets sont testées une à une de manière complètement automatique, et au cas où quelque chose serait modifié dans le code altérant les résultats espérés pour cette méthode, des avertissement et des alarmes seront déclenchés alertant sur le résultat inespéré – ceci est nécessaire en fonction de l'interdépendance entre les différents codes qui sont utilisés dans les simulateurs de Gaia, et du grand nombre de personnes que finissent par altérer ces codes.

4.3 Simulation d'images d'objets étendus – MAGIL

Plusieurs études demandent non seulement des catalogues de positions d'objets, mais aussi des images de ces objets. L'exemple le plus direct est la détermination de l'efficacité de détection des galaxies non résolues. Pour cela il est nécessaire, non seulement, que les catalogues de galaxies soient simulés, mais que leurs images puissent aussi être synthétisées au niveau du pixel sur plan focal du satellite. Pour cela, une première adaptation d'une méthode dénommée Skymaker, de E. Bertin a été réalisée par C. Dollet, pour être ensuite adaptée aux normes du DPAC et re-écrite en Java par C. Delle-luche. Ce code crée l'image des galaxies à partir d'une superposition d'un profil exponentiel pour le disque et d'un bulbe avec un profil de de Vaucouleurs.

Cependant pour permettre un plus grand réalisme dans les galaxies simulées permettant par exemple d'étudier la détection de clumps sur les disques des galaxies proches il est nécessaire de disposer d'un outil permettant d'introduire des images réelles de galaxies, extraites d'une bibliothèques d'images de galaxies proches (Babusiaux, 2005) .

Le simulateur qui a été mis en oeuvre et qui tient ce rôle s'appelle MAGIL (*MAnager of Gaia Image Library*), et ses spécifications et diagrammes de base du fonctionnement ont été créés par P. Gavras et A. Krone-Martins durant des discussions à *ELSA School on the Science of Gaia*, en 2007. La première mise en oeuvre fonctionnelle en Java a été créée par P. Gavras et a été délivrée à la mi-2008. Sa connexion à GIBIS a été réalisée dans la même année. Une description complète de ce simulateur peut être retrouvée dans Gavras et al (2010).

MA.G.I.L. est constitué de trois modules distincts : un contrôleur, un processeur et une bibliothèque. Un diagramme représentant la relation entre ces modules est présentée en Figure 4.10. La bibliothèque, est une base de données d'images qui, en plus de stocker des images des objets prototypes, stocke aussi des paramètres caractéristiques de ces objets (leurs types, rapports axiaux, tailles angulaires, etc.). Le processeur est responsable de la manipulation des images. Il modifier selon des desideratas de l'utilisateur, les rapports d'axes, la taille de l'objet en pixels, il redimensionne l'image en flux et éliminer les fonds de ciel significatifs.

Le troisième composant, le contrôleur, est responsable de la gestion du système complet. Il reçoit les requêtes de l'utilisateur (au moyen d'une interface GUI *stand-*

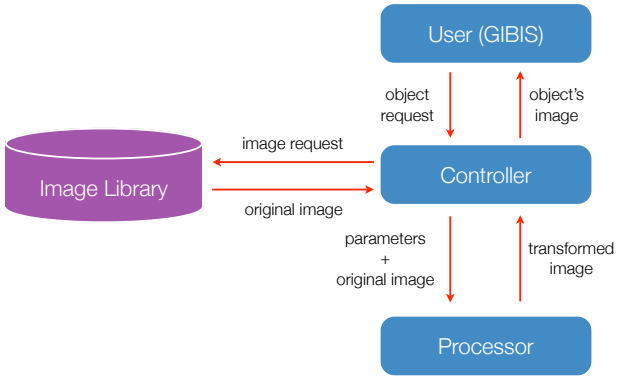


FIGURE 4.10 – Schéma général du traitement survenant dans MAGIL.

alone ou de GIBIS), demande à la bibliothèque l'image adéquate, détermine les modifications nécessaires à appliquer à cette image, demande ces modifications au processeur, et retourne l'image résultante à l'utilisateur.[8]

4.3.1 Exemples d'images simulées

Nous présentons ici deux exemples de galaxies générées par MAGIL conjointement avec les images prototypes qui ont été utilisées. La première galaxie demandée peut être vue sur la Figure 4.11. Il s'agit d'une galaxie spirale, de magnitude 17 et avec un rapport d'axes de 1. La matrice rectangulaire dans le coin supérieur droit est le résultat final.

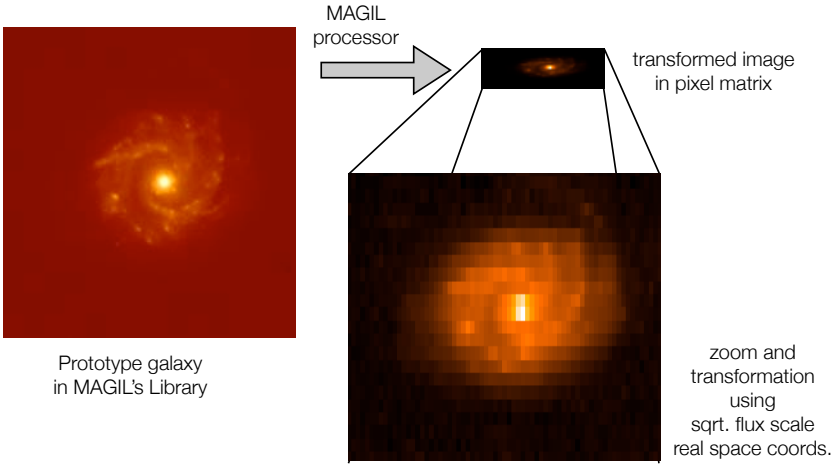


FIGURE 4.11 – Exemple d'image d'une galaxie spirale produite par MAGIL.

La galaxie spirale simulée ci-dessus, a une taille angulaire totale de 3″, mais en

8. Le traitement de la PSF de l'image est réalisé dans GIBIS, et il est fait de façon cohérente entre toutes les images qui sont formées dans le plan focal du satellite.

principe elle devrait avoir ses données transmises au sol en fonction de son profil de brillance centrale, ce qui en fait presque une source ponctuelle. Un deuxième exemple serait la galaxie Irrégulière montrée sur la Figure 4.12 qui, ayant un profil de brillance peu piqué et une étendue angulaire de ~ 700 mas, serait à la limite de la détection de Gaia.

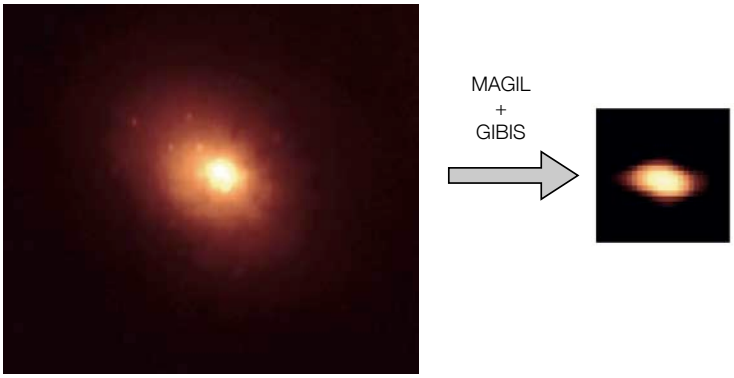


FIGURE 4.12 – Exemple d'image d'une galaxie irrégulière produite par MAGIL sur le plan focal simulé par le GIBIS. Un zoom de 4x a été appliqué, et une carte de couleurs sur une échelle logarithmique. (Figure de Gavras et al, 2010)

4.4 Estimations du nombre de galaxies observables

Dériver une estimation du nombre de galaxies qui seront observables par Gaia exige une connaissance détaillée de la réponse des algorithmes de détection à bord du satellite en ce qui concerne les objets étendus. Par exemple, même si l'on sait qu'il existe une coupe à ~ 700 mas, un objet plus grand que cette taille pourrait, en principe, être observé à la condition que son bulbe soit inférieur à cette taille, et que ce bulbe domine l'émission de la galaxie. Des exemples ont été montrés dans les images simulées des Figures 4.11 et 4.12.

Malheureusement jusqu'au début de la rédaction de cette thèse, les codes de détection qui seront utilisés au cours de la mission n'avaient pas été intégrés à GIBIS. Nous avons donc dérivé des estimations des limites supérieures de galaxies non résolues qui pourraient être observées par Gaia.

Les résultats des simulations officielles plus récentes (Mai 2010) réalisées avec GOG et qui ont utilisé JStuff, ont été présentés dans Robin (2010). Ceux-ci indiquent que la limite supérieure d'objets qui pourraient, en principe, être observés serait de $\sim 3 \times 10^7$ galaxies non résolues jusqu'à la magnitude $G \sim 20$. Néanmoins, il est réaliste de supposer que pour des objets étendus, tels que les galaxies, la magnitude limite sera un peu inférieure, si bien qu'en re-analysant le même lot de simulations et ne prenant en compte que les galaxies de magnitudes entre 11.5 et 18.5, le nombre d'objets existants dans le ciel en son entier serait de 6×10^6 galaxies, dont la distribution peut être vue sur la Figure 4.13.

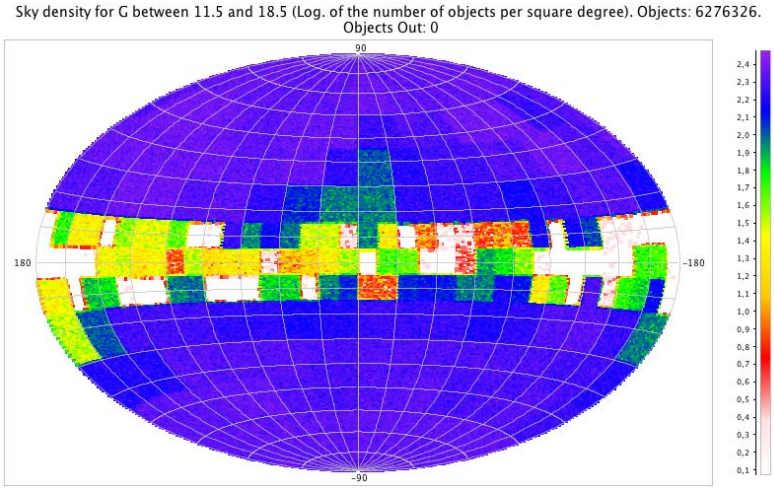


FIGURE 4.13 – Distribution des galaxies non résolues dans le ciel pour des magnitudes $11.5 \leq G \leq 18.5$. Les zones avec des comptages incorrects sur le plan de la galaxie sont causées par une erreur dans le modèle d'extinction. au moment de la simulation. Simulation du catalogue Gaia avec GOG (cycle 8, 05/2010).

En analysant la distribution de *redshift* des galaxies (statistique actuellement disponible uniquement pour $G \leq 20$) qui pourraient être observées par le satellite, nous estimons que probablement la valeur moyenne observée pour tous les types morphologiques sera de $\langle z \rangle \sim 0.17$, avec la plus grande partie des objets ayant une magnitude de $G > 17.0$. Ces résultats peuvent être vus graphiquement sur la Figure 4.14.

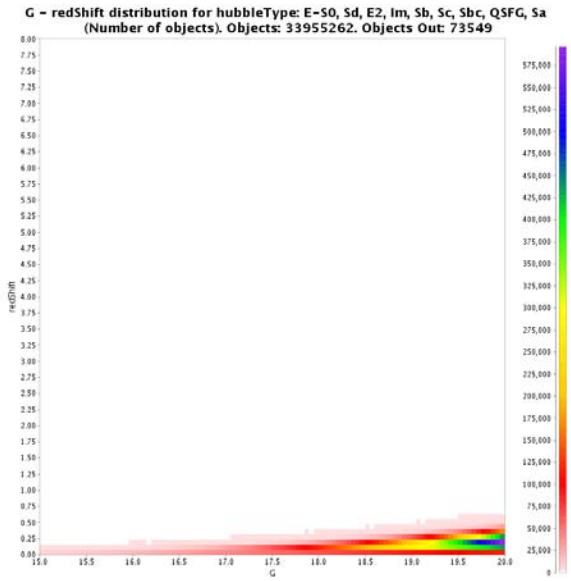


FIGURE 4.14 – Histogramme 2D des galaxies non résolues dans l'espace *redshift* vs. magnitude G obtenu par simulation du catalogue Gaia avec GOG (cycle 8, 05/2010).

4.4.1 Évaluation basée sur une extrapolation du Hubble-MDF

Une estimation complètement indépendante du nombre de galaxies non résolues observables peut être obtenue à partir d'une extrapolation pour la sphère céleste de la densité d'objets présentant les caractéristiques que nous espérons observer avec Gaia et qui ont déjà été observés dans le Hubble *Medium Deep Field* (Griffiths et al, 1994). Ces observations sont une série d'images réalisées en mode parallèle, c'est-à-dire, pendant que Hubble observait avec un instrument une certaine cible scientifique, la *Wide Field Camera* observait un champ proche quelconque.

Les données utilisés dans Griffiths et al (1994), sont spécifiquement originaires d'un champ proche du quasar 3C273. La taille moyenne obtenue pour les objets sur cette image était de $0''.4$, démontrant que beaucoup de ces objets passeraient effectivement par la coupe de 700 mas existante à bord du satellite Gaia, générant des objets problématiques durant la réduction des données. Nous avons noté que cet article montre que le meilleur modèle cosmologique pour expliquer le nombre de galaxies d'une certaine taille, est un modèle riche en galaxies naines.

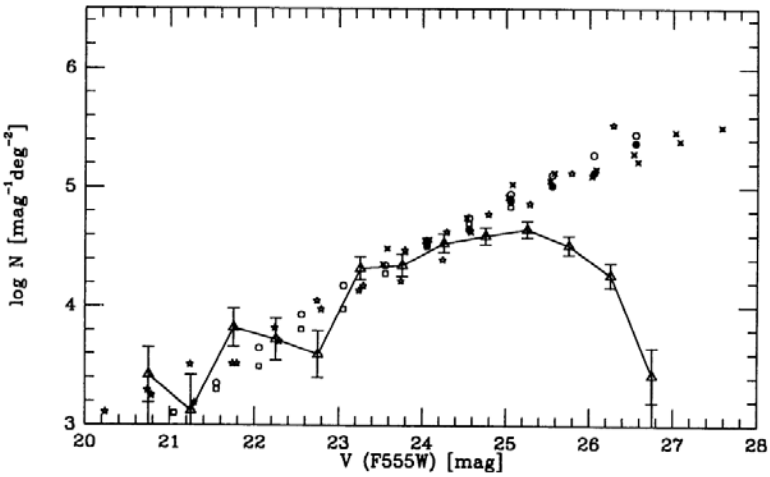


FIGURE 4.15 – Densités de galaxies dans la bande V obtenues à partir du *Hubble Medium Deep Field*. Figure de Griffiths et al (1994).

Pour une magnitude $V \sim 20$, la densité estimée est de $\sim 10^3$ [gal.mag^{-1}deg^{-1}], comme on peut le voir sur la Figure 4.15. La plus grande partie des galaxies observées dans le MDF possèdent des tailles angulaires inférieures à $0''.7$, donc elles seraient potentiellement observables par Gaia (au cas où la magnitude limite le permettrait). En extrapolant les valeurs de densité en $V \sim 20$ pour le nombre d'objets dans le ciel dans son entier, nous pouvons estimer que $\sim 3 \times 10^7$ galaxies est la limite supérieure du nombre de tel objets que Gaia pourrait observer (en excluant la bande de la Voie-Lactée). Il convient de remarquer que cette estimation de densité est basée sur un petit nombre d'objets, mais la valeur obtenue est parfaitement compatible avec celle obtenue à partir des simulations du GOG avec JStuff pour $G = 20$ présentées dans la section antérieure.

4.5 Conclusions

Dans ce Chapitre nous avons présenté le code JStuff, que nous avons mis en oeuvre dans le modèle d'Univers de la mission Gaia pour permettre la réalisation de simulations réalistes de catalogues de galaxies non résolues. Nous avons aussi présenté le code MAGIL développé par P. Gravas avec notre collaboration, qui permet l'introduction d'images de n'importe quel objet étendu, mais principalement des galaxies non résolues, dans le simulateur d'observations de Gaia au niveau du pixel (GIBIS).

Finalement, en utilisant des données simulées du GOG qui ont adopté JStuff et un modèle d'extinction dû à la Voie-Lactée, une étude préliminaire permet d'obtenir une estimation pour le nombre supérieur de galaxies non résolues qui pourra être observé par Gaia, donnant comme résultat $\sim 6 \times 10^6$ galaxies jusqu'à $G = 18.5$ et 3×10^7 jusqu'à $G = 20$, une valeur que nous avons montrée être compatible avec les résultats de l'extrapolation du Hubble *Medium Deep Field*. De plus, ces simulations aussi ont permis l'estimation du *redshift* moyen des objets que doivent être observés en $\langle z \rangle \sim 0.17$.

Chapitre 5

Analyse et classification d'Images de Galaxies

"*Crude classifications and false generalizations are the curse of organized life.*"[1]
G. B. Shaw

Sommaire

Bien que la mission spatiale Gaia ne soit pas destinée à l'observation d'objets extragalactiques un nombre considérable d'entre eux sera quand même observé. Cependant, en fonction du critère de sélection d'objets *on-board*, seules les données de galaxies angulairement petites seront transmises. De plus, seulement $\sim 10^2$ pixels seront disponibles, et comme nous l'avons vu dans le Chapitre 2, les reconstructions seront affectées par des artefacts et la qualité du signal reconstruit sera variable.

Dans ce chapitre nous présenterons des paramètres robustes qui seront utilisés durant l'analyse d'images reconstruites de Gaia pour effectuer une classification morphologique des galaxies observées (Krone-Martins et al, 2008b). Nous présentons ensuite la méthode de classification qui sera utilisée. Puis nous présentons l'application de cette méthode à des populations aléatoires, des images de galaxies proches, des images reconstruites à partir de simulations de la mission Gaia, et de galaxies observées dans le *Hubble Deep Field* (Krone-Martins et al, in prep.a).

1. « Des classifications rudimentaires et de fausses généralisations sont la malédiction de la vie organisée. »

5.1 Introduction

Le processus primaire dans les sciences observationnelles est la taxonomie, soit l'organisation d'objets en fonction de leurs similitudes et différences. C'est le premier pas pour dériver des relations entre les objets, leurs comportements individuels ou collectifs.

Dans le cas de galaxies ou de n'importe quel autre objet étendu, la morphologie est le traceur le plus accessible de la structure physique des objets, et ainsi la voie la plus basique et naturelle pour la réalisation de classifications. La compréhension de l'évolution de cette morphologie est l'une des principales clés pour la compréhension de l'évolution de l'Univers, en particulier de la distribution de la matière dans les structures observées aujourd'hui et dans la discussion des scénarios des formation de structures, comme par exemple le scénario hiérarchique.[2]

Traditionnellement la classification des galaxies est réalisée en utilisant le schéma qui les divise en Elliptiques, Lenticulaires, Spirales normales et Spirales barrées, proposé par Hubble (1926, 1936) et Sandage (1961), et en ajoutant les galaxies Irrégulières.[3] Bien que ce schéma de classification fonctionne très bien pour les galaxies brillantes proches, il peut être inefficace pour des objets compacts et/ou peu lumineux, pour la classification dans des amas riches et pour des objets de redshift élevé, tels que ceux observés par exemple dans Abraham et al (2003).

Une autre critique contre ce système est le fait qu'il repose sur une appréciation purement visuelle et qu'il existe entre astronomes des divergences systématiques dans la classification d'un même objet, spécialement dans le cas de galaxies distantes, compactes (ou avec peu d'information) et de galaxies particulières, comme on peut le constater dans les travaux de Naim et al (1995) et Abraham et al (1996b).

Finalement, les principaux paramètres qui définissent le type de Hubble, comme le rapport bulbe-disque, l'ouverture et le degré de résolution des bras spiraux, l'existence ou non d'une barre, sont peu résistants à la dégradation de la qualité spatiale de l'image et du bas rapport signal-bruit. Malheureusement ce sont ces facteurs qui seront prévalant dans les images reconstruites de Gaia. Ainsi, pour extraire des informations scientifiques à partir des images reconstruites avec les données de cette mission, il est nécessaire d'utiliser un schéma de classification basé sur des caractéristiques distinctes de celles citées ci-dessus. Néanmoins du fait de usage généralisé (Kormendy & Kennicutt, 2004; Sandage, 2005), il est souhaitable que quel

2. Cette dénomination est due au fait que dans ce scénario depuis le début de l'univers les plus grandes structures sont formées par la fusion de structures de plus petite taille (White & Frenk, 1991, et références) , et qui une théorie largement acceptée (Springel et al, 2005) . Cependant des travaux récents présentent des évidences non-expliquées par ce scenario, par exemple : Collins et al (2009) montre que les amás massives de galaxies se forment rapidement ($< 4 - 5$Gyr), et non lentement tel que le modèle hiérarchique le propose. Le travail de Delgado-Serrano et al (2010) montre que le nombre de galaxies particulières diminue tandis que celui de spirales augmente au cours du temps (le nombre d'elliptiques et lenticulaires se maintient constant).

3. En réalité, la classification est plus fine, contenant divers types de galaxies elliptiques, spirales et des critères visuels pour faire la différenciation entre de tels sous-types. Deux références importantes sur le sujet aussi bien du point de vue historique que du point de vue conceptuel sont Van den Bergh (1998) et Sandage (2005).

que soit le système utilisé pour la classification morphologique dans Gaia, il puisse être traduit dans un type de Hubble. Deux voies distinctes peuvent être adoptées pour atteindre cet objectif :

- la première, que nous dénommons paramétrique, consiste en la modélisation de l'objet à partir d'une définition à priori des structures attendues pour ces objets[4], dans l'ajustement profils, et donc dans la corrélation entre les valeurs ajustées et la classification des objets ;
- une seconde option, dénommée non paramétrique, consiste à définir des grandeur caractéristique des images (asymétrie, *smoothness*, etc.), à mesurer ces quantités sur les images et à classer les objets dontenus dans ces images à partir de ces quantités.

En général, on préfère utiliser la première voie comme Barden et al (2005), ou Delgado-Serrano et al (2010). En effet les paramètres des modèles sont en général directement traduits en aspects physiques des objets, comme par exemple les tailles du bulbe et du disque. Néanmoins, comme nous l'avons vu dans le Chapitre 2, les images reconstruites pour Gaia ne contiennent pas seulement l'objet étudié, mais contiennent différents signaux parasites et des artefacts qui proviennent ou de l'algorithme adopté pour la reconstruction ou de la distribution non homogène des angles de passage du satellite.

Ainsi, dans le cas particulier de la mission Gaia il est préférable d'utiliser des paramètres peu sensibles au petit nombre d'informations spatiales ($\sim 10^2$ pixels) et qui puissent être pondérés. Les algorithmes de reconstruction d'image disponibles actuellement ne fournissent pas d'évaluation fiable de l'erreur de reconstruction mais ceci est une nécessité pour les futurs d'algorithmes.

Nous commençons ce Chapitre avec une revue des travaux présentant les paramètres que nous avons choisis pour réaliser l'analyse des images Gaia reconstruites. Puis nous présenterons des analyses d'images que nous avons réalisées avec ces paramètres. Dans une deuxième partie, nous présenterons l'algorithme d'apprentissage computationnel choisi pour classifier les images reconstruites, ainsi que les résultats obtenus avec l'application de cette méthode sur des données simulées avec les simulateurs officiels du DPAC ainsi que sur des données réelles de galaxies proches et éloignées

5.2 L'espace CASGM20

Un ensemble de paramètres qui a été utilisé dans des cas similaires à celui de Gaia (particulièrement en l'absence de grandes quantités d'informations spatiales) est dénommé CASGM20. Ce sont essentiellement des mesures de la concentration centrale de la lumière de l'objet (C), de l'asymétrie du signal (A), de le smoothness

4. Des profils sont définis, en général analytiques, pour la distribution de brillance, comme celui de de Vaucouleurs (de Vaucouleurs, 1948), exponentiel ou de Sérsic (Sérsic, 1968).

(S), de la concentration non nécessairement centrale (G) et d'un moment de l'image ($M20$).

Doi et al (1993) ont démontré que la concentration centrale et la brillance superficielle moyenne d'une galaxie peuvent être utilisées pour distinguer effectivement entre des types de Hubble précoce et tardif, vu qu'elles ont une corrélation avec ces derniers. S'inspirant de ces résultats, Abraham et al (1994) ont décrit un système automatique de classification morphologique adéquat pour une utilisation dans l'étude de galaxies quand peu de pixels, ($\sim 10^2$) , sont disponibles (ce nombre de pixels est similaire à ce qui sera disponible pour classifier des images reconstruites de la mission Gaia). Un tel système est basé sur un paramètre qui trace aussi bien la relation disque-bulbe que le rayon effectif du bulbe. Dans ce même travail, ils démontrèrent que les classifications obtenues en utilisant ce système sont beaucoup moins sensibles à la dégradation de la résolution spatiale de l'image que par le système de Hubble.

Concentration centrale – C

Le paramètre C utilisé par Abraham et al (1994), décrit la concentration centrale de lumière dans l'image de la galaxie, et est directement calculé par le rapport du flux contenu dans l'isophote (E) de rayon $r = \alpha$ (dans cet article, la valeur 0.3 a été utilisée) et celui $r = 1$, définis par :

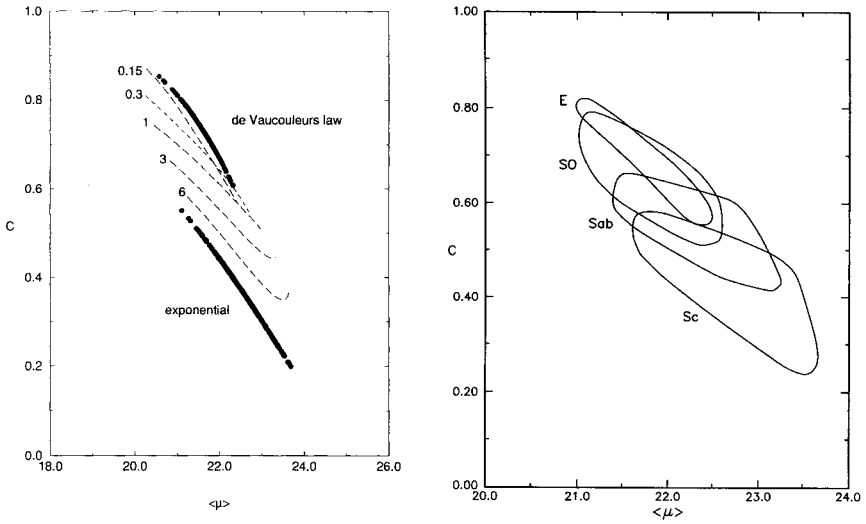
$$C = \frac{\sum \sum_{i,j \in E(\alpha)} I_{i,j}}{\sum \sum_{i,j \in E(1)} I_{i,j}} \tag{5.1}$$

La classification des galaxies étudiées est alors obtenue par Abraham et al (1994), en mesurant simplement la valeur de C et en la comparant aux valeurs moyenne de galaxies préalablement classées. Pour la classification de galaxies, on utilise aussi un plan construit par cette valeur C et la valeur de la brillance superficielle moyenne $\langle\mu\rangle$, calculée par le rapport entre la somme du signal de la galaxie et le nombre de pixels (où A est la région de l'image avec des pixels au-dessus du bruit – défini en 2σ dans l'article) :

$$\langle\mu\rangle = \frac{\sum \sum_{i,j \in A} I_{i,j}}{n_{pix}} \tag{5.2}$$

Avec ce système, les auteurs montrent que la classification des images de galaxies jusqu'à cinq classes distinctes peut être réalisée de façon complètement automatique quand de bonnes conditions de résolution spatiale sont disponibles, et qu'il est toujours possible de réaliser une classification entre les types précoce/tardif même dans des conditions de seeing dégradé. Le diagramme de la Figure 5.1a montre les régions dans le plan $C - \langle\mu\rangle$ occupées par des galaxies modelées comme des superpositions de disques de de Vaucouleurs et des bulbes exponentiels, pour diverses valeurs du rapport bulbe/disque.

Sur la figure 5.1 b les positions dans le plan $C - \langle\mu\rangle$ occupées par plusieurs types de Hubble sont représentées (en éliminant tous effets de dégradation de *seeing* ou d'échantillonnage). Les régions de classifications ont été déterminées sur la base



(a) (b)

FIGURE 5.1 – Positions dans le plan de classification de modèles créés à partir de superpositions de disques de de Vaucouleurs et de bulbes exponentiels. (a) Les lignes pointillées indiquent des régions occupées par diverses rapports bulbe/disque. (b) Positions dans le plan de classification pour plusieurs types de Hubble. Les contours ont été tracés à 2σ avec un isophote limite de 25.5 mag arcsec^{-2}. Les galaxies ont été classées en n'utilisant que les valeurs de la relation disque-bulbe : Sc entre [4 ;+inf], Sab entre [1.5 ; 6], S0 [0.1 ;2]. (tiré d' Abraham et al, 1994)

des valeurs bulbe/disque caractéristiques pour chaque type de galaxie et dans des contours en 2σ. Il faut noter que les positions occupées par des galaxies type précoce et tardif se trouvent relativement bien superposées dans ce plan, faisant que la classification dans ce plan soit incertaine.

Asymétrie – A

Dans un deuxième travail, Abraham et al (1996a) ont analysé la morphologie de galaxies distantes provenant de données du *Hubble Space Telescope Medium Deep Survey*, ou MDS (Griffiths et al, 1994) , et ont introduit une deuxième mesure A, dénommée Asymétrie. De cette manière, le plan $C - \langle\mu\rangle$ utilisé antérieurement pour réaliser la classification a été remplacé par le plan $C - A$.

L'asymétrie est obtenue en effectuant une rotation de l'image de 180° autour de son centre, en soustrayant cette image à +180° de l'image originale et en prenant la moitié du rapport entre la valeur absolue de la somme de la lumière présente dans cette image soustraite et la somme de la lumière présente dans l'image originale (après la soustraction du fond ciel). Le centre de rotation est déterminé au moyen d'un lissage de l'image avec un filtre gaussien de $\sigma = 1$ pixel et par la détermination du pixel de valeur maximum.

Les positions dans l'espace C - A pour diverses galaxies, qui ont été classifiées visuellement par Richard S. Ellis, sont présentées en Figure 5.2. Il faut noter que même s'il existe une grande superposition entre les types morphologiques, il y a de régions plus occupées par certains types de galaxies. Notons, cependant, que les séparations n'apparaissent pas naturellement.

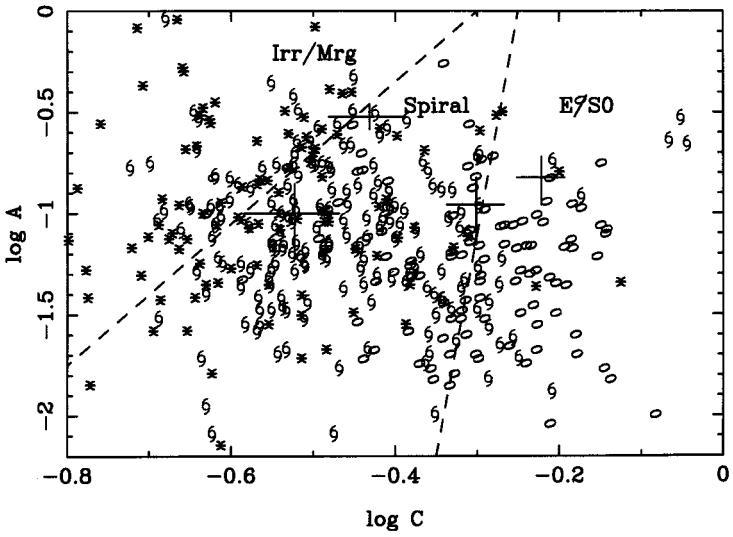


FIGURE 5.2 – Positions dans le plan $C - A$ (concentration-asymétrie) de galaxies du MDS classifiées visuellement par Richard S. Ellis. Les ellipses indiquent des galaxies compactes, elliptiques et S0s, des spirales indiquent des galaxies spirales, et les irrégulières ou les mergers sont indiquées par des astérisques. Les régions de séparation ont été créées à partir de l'analyse de galaxies proches et bien classifiées visuellement (originaires du catalogue de Frei et al, 1996) placées dans des redshifts plus élevés. (extrait d' Abraham et al, 1996a)

En conclusion ce travail montre que la classification basée seulement sur la concentration centrale n'est pas capable de séparer les divers types précoces, et qu'il n'est pas possible de distinguer les types intermédiaires-à-tardifs des tardifs et irréguliers. Néanmoins, le fait d'incorporer l'asymétrie permet d'améliorer la distinction entre les types intermédiaires et irréguliers, même s'il existe une superposition entre les types très tardifs/irréguliers/*mergers*.

D'autres travaux, comme Huertas-Company et al (2007) , incorporent de petites modifications dans ce système, de manière à rendre les mesures de C et de A plus robustes vis à vis de la sélection de brillance superficielle, des erreurs de détermination du centre de l'image et de dépendance avec le redshift.

Clumpiness – S

Le troisième paramètre que nous avons utilisé, appelé *clumpiness* (rugosité) S[5] a été proposé par Conselice (2003, 2006). Cette mesure est un indicateur de la granulation de l'image. Elle est intéressante étant donné que les galaxies elliptiques par exemple, ne possèdent pas de structures avec de hautes fréquences spatiales car leur distribution de lumière est moins piquée alors que les galaxies passant par une formation stellaire possèdent une quantité élevée de lumière émise dans des fréquences spatiales élevées.

Dans ce travail, S est défini comme étant le rapport entre la lumière contenue dans des hautes fréquences spatiales et la quantité totale de lumière dans la galaxie. Pour calculer cette valeur, l'image originale de l'objet doit être dégradée au moyen d'une atténuation, créant ainsi une image dont les structures de haute fréquence ont été éliminées.[6] Donc, l'image originale est soustraite de l'image dégradée, créant une carte qui ne contient que les composants responsables de la présence de structures de haute-fréquence dans la distribution spatiale de lumière de la galaxie. Le flux de cette image est alors additionné et divisé par la somme du flux de l'image originale, résultant dans la mesure de rugosité :

$$S = 10 \times \sum_{x=1}^{N} \sum_{y=1}^{N} \frac{(I_{x,y} - I^{\sigma}_{x,y}) - B_{x,y}}{I_{x,y}} \tag{5.3}$$

dans la définition ci-dessus, $I_{x,y}$ est le flux de la galaxie, $I^{\sigma}_{x,y}$ est la valeur de l'image atténuée et $B_{x,y}$ est la valeur du fond de ciel à la position (x, y).

Dans le calcul, la partie interne de la galaxie n'est pas prise en compte, car selon Conselice cette région contient une puissance en hautes fréquences qui n'est pas seulement liée à la distribution réelle de lumière, mais qui est conséquence de l'échantillonnage fini. Dans cette définition, tous les pixels d'intensité négative de l'image soustraite sont définis comme étant zéro, avant de calculer la valeur de S.

Dans ce même article, des mesures de C et de A sont utilisées, obtenues à partir d'une définition différente de celle adoptée par Abraham et al (1994, 1996a) :

- Pour mesurer l'asymétrie, une dégradation de l'image avant la rotation et la soustraction est réalisée ;
- Pour mesurer la concentration on utilise une méthode décrite en détails dans Bershady et al (2000), dans laquelle le rapport qui est pris en compte est celui entre les rayons de 80% et 20% d'émission (selon Conselice, 2003, l'émission totale est donnée par l'émission dans 1.5 rayons pétrosiens[7]) .

Une corrélation intéressante existe entre la valeur du paramètre de clumpiness S, l'indice de couleur $(B - V)$ et la largeur équivalente de la ligne de Hα (comme

5. Il est dénommé S, car C avait déjà été utilisé, et en même temps que ce paramètre indique la rugosité il indique aussi le *Smootheness*.

6. Il faut noter que ce type d'information aussi peut être déterminé à partir de coefficients d'une transformée en ondelettes.

7. Voir section 5.3.

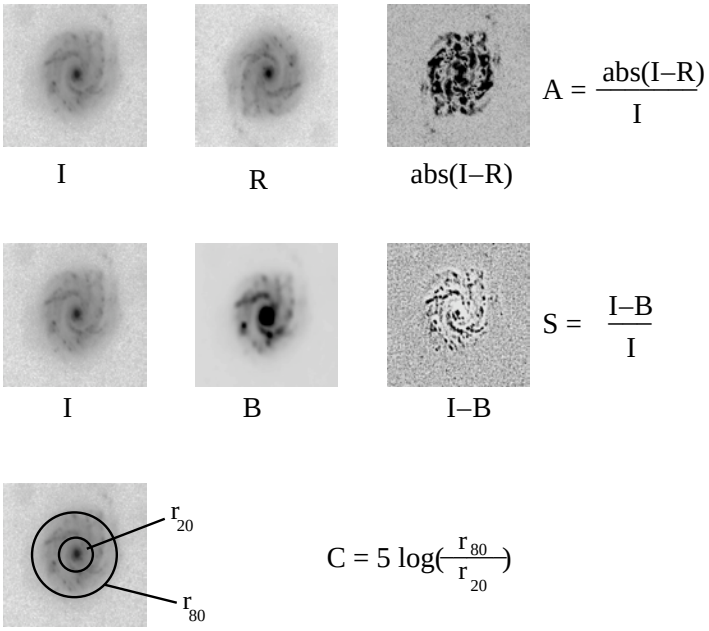


FIGURE 5.3 – Représentation des mesures CAS. (extrait de Conselice, 2003)

on peut le voir en figure 5.4). Cette relation était, d'une certaine manière attendue, car en fonction de la définition de S, des galaxies avec de nombreuses régions de formation stellaire, donc avec des valeurs de S qui réfléchissent une telle émission en hautes fréquences spatiales, sont aussi des galaxies avec une émission élevée en Hα et avec des valeurs de $(B-V)$ caractéristiques.

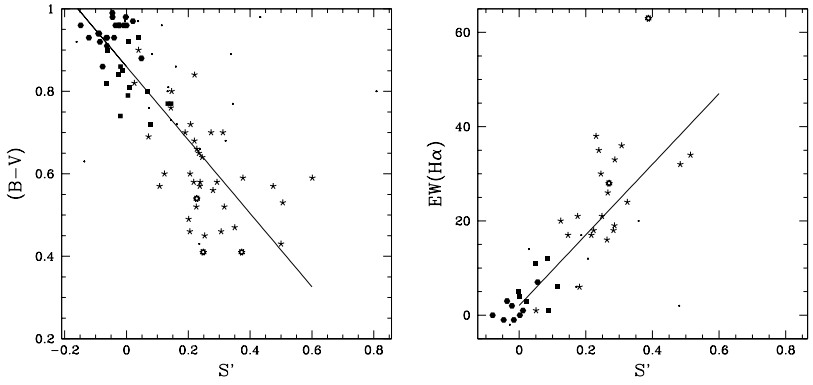


FIGURE 5.4 – Relations entre la valeur de S corrigée de l'inclination de la galaxie, l'indice $(B-V)$ et la largeur équivalente de la ligne Hα. (extrait de Conselice, 2003)

La dispersion existante est probablement due au fait que S est sensible à l'angle de visée sous lequel la galaxie est observée, et peut être comprise comme une incertitude additionnée en fonction de l'effet de projection du disque, et donc de la poussière présente dans ce disque.

Du point de vue observationnel, la ligne de Hα est importante par le fait d'être liée avec le taux de formation stellaire d'un objet (Kennicutt, 1998). De cette manière, une mesure purement morphologique, telle que la valeur de S, trace une caractéristique physique, qui est le taux de formation stellaire.

À partir des mesures de C, A et S, il est possible de construire les plans $C-A$, $S-A$ et $C-S$, dans lesquels Conselice réalise une classification. Ces plans sont présentés sur la Figure 5.5, avec des galaxies de diverses classes.

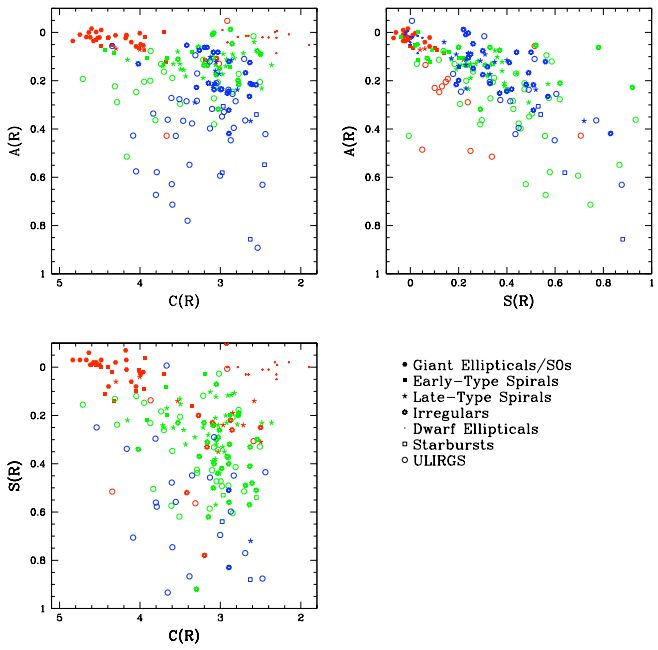


FIGURE 5.5 – Plans de classification. Les couleurs représentent le paramètre qui ne se trouve pas dans le plan représenté. Sur la première figure, le rouge représente $S < 0.1$, le vert $0.1 < S < 0.35$, le bleu $S > 0.35$. Sur $S-A$, le rouge représente $C > 4$, le vert $3 < C < 4$ et le bleu $C < 3$. Dans le plan $C-S$, le rouge représente $A < 0.1$, le vert $0.1 < A < 0.35$ et le bleu $A > 0.35$. (extrait de Conselice, 2003)

Coefficient de Gini – G

Un autre paramètre qui complète les volume CAS, est le coefficient de Gini, introduit en Astronomie par Abraham et al (2003). Ce coefficient est une mesure utilisée en économétrie pour décrire l'inégalité dans la distribution de richesse d'une population donnée. Selon la description dans cet article, la courbe de Lorenz (1905) est construite à partir de la proportion cumulative de richesse comme fonction du pourcentage de la population, de manière à ce que si la société est parfaitement égalitaire cette fonction est une ligne droite avec une pente égale à 1, et dans un cas extrême dans lequel seulement une petite proportion de la population détient presque toute la richesse, la courbe est d'abord plutôt horizontale et proche de zéro jusqu'au moment où ce elle monte rapidement en arrivant près de la fin de l'intervalle [0 ;1].

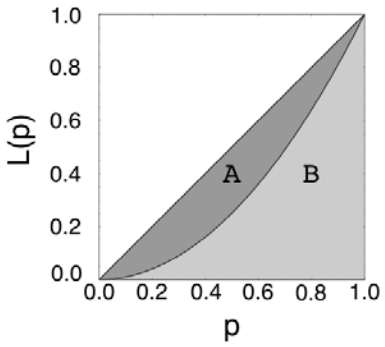


FIGURE 5.6 – Interprétation géométrique du coefficient de Gini. Le coefficient de Gini correspond au rapport entre la surface A et la surface totale A+B. (extrait d' Abraham et al, 2003) avec L= la part cumulée des revenus et p=la part cumulée de la population.

Selon Abraham et al (2003), une description plus formelle de ce coefficient serait : Soit X une variable aléatoire positive de fonction de distribution cumulative $F(x)$, et soit X_i un échantillonnage de X, la courbe de Lorenz est donnée par :

$$L(p) = \frac{1}{\overline{X}} \int_o^p F^{-1}(u)du \tag{5.4}$$

Le coefficient de Gini est défini alors comme étant la moyenne de la différence absolue entre toutes les combinaisons de X_i :

$$G = \frac{1}{2\overline{X}n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j| \tag{5.5}$$

Selon Abraham et al (2003), une forme particulièrement efficiente pour calculer G est d'ordonner les X_i dans un ordre croissant puis ensuite de réaliser une somme simple :

$$G = \frac{1}{\overline{X}n(n-1)} \sum_{i=1}^{n} (2i - n - 1)X_i \quad (n > 2) \tag{5.6}$$

Pour le calcul de G, on utilise 5.6 sur une liste ordonnée par intensité des valeurs des pixels de l'image de l'objet donné. Les auteurs remarquent que dans une première approximation, la valeur de G est une forme généralisée de la valeur de la mesure de concentration – ainsi, la Figure 5.7 montre que pour des valeurs élevées de C, le rapport est linéaire avec une pente relativement unitaire.

Cependant il est souligné que G ne peut pas être utilisé aveuglément comme substitut de C, étant donné que une modification de la distribution spatiale des pixels n'affecte pas la valeur de G alors qu'elle modifie celle de C. On peut percevoir intuitivement que G peut être lié à une concentration de lumière indépendante de la position sur l'image : même après avoir réordonné comme on veut les pixels, G n'est pas altéré.

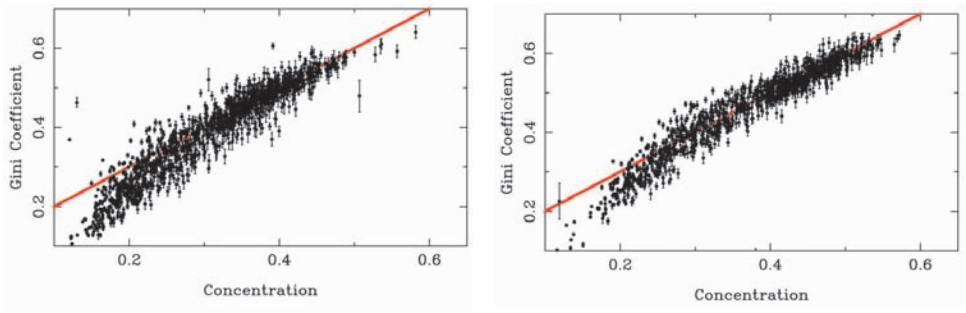


FIGURE 5.7 – Rapport entre le coefficient de Gini (G) et la mesure de concentration (C) pour un ensemble d'images dans les bandes g et i provenant du SDSS *Early Date Release*. (extrait d' Abraham et al, 2003)

Dans ce même travail les auteurs ont analysé les corrélations entre les valeurs de C, G et d'autres grandeurs, telles que la brillance superficielle moyenne, la magnitude absolue, la magnitude apparente, le *redshift*, la couleur et rapport des axes. Parmi toutes les grandeurs analysées, la corrélation rencontrée avec la brillance superficielle moyenne est intéressante. Des projections de la distribution des galaxies analysées dans un volume $C - G - \langle\mu\rangle$ (Figure 5.8) montrent que celles-ci se trouvent dans un plan très bien défini, avec une dispersion relativement faible. Ce point est intéressant, car cela lie une quantité purement photométrique à des quantités morphologiques.

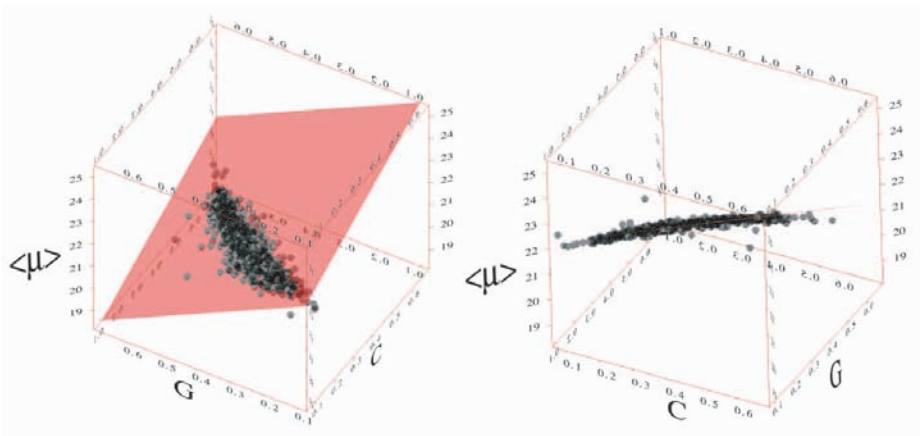


FIGURE 5.8 – Distribution des galaxies dans le volume $C - G - \langle\mu\rangle$. Le plan en rouge représente le meilleur ajustement, et montre comment les galaxies sont si peu dispersées dans ce volume. (extrait d' Abraham et al, 2003)

À partir de simulations, Abraham et al (2003) ont observé que la topologie et l'orientation de la superficie caractéristique dans laquelle les galaxies sont distribuées dans le volume $C - G - \langle\mu\rangle$ sont réelles, mais ils suggèrent que le fait que cette superficie soit un plan pour les galaxies observées pourrait être associé à une sélection dans leur distribution.

Moment M20

La dernière mesure que nous utiliserons est dénommée $M20$, et a été introduite par Lotz et al (2004). Celle-ci est le moment d'ordre 2 des pixels brillants responsables de 20% du flux de l'image.

Dans leur travail, les auteurs commentent aussi que le fait que l'inclusion de pixels du fond de ciel dans le calcul de G majorera systématiquement telle grandeur, tandis que la non inclusion des pixels de basse brillance superficielle de la galaxie la diminuera systématiquement, et c'est pourquoi ils redéfinissent aussi le calcul de G : ils créent une carte de segmentation des pixels de la galaxie, et seulement les valeurs de pixels qu'appartient à la galaxie sont utilisés dans le calcul de G. Une autre différence par rapport à la manière de calculer G est l'utilisation de la valeur absolue du flux car l'image est soustraite du ciel. Lotz et al (2004) notent que les corrections permettent d'obtenir la valeur de G avec un taux de confiance de 10% pour des images avec $S/B > 2$.

La mesure de $M20$, dépend du moment total d'ordre deux M_{tot}, qui est simplement le flux dans chaque pixel multiplié par la distance au carré au centre de la galaxie, additionné sur tous les pixels de la galaxie. Le centre de la galaxie ici est défini comme la position pour laquelle M_{tot} est minimisé. Donc, $M20$ est le moment normalisé de deuxième ordre des pixels jusqu'à atteindre 20% du flux total sur l'image. Pour calculer M20, les pixels de l'image sont ordonnés par flux (f_i), ajoutés jusqu'à atteindre 20% du flux total (f_{tot}) de la galaxie et sont alors normalisés :

$$M_{20} = \log_{10}\left(\frac{\sum_i M_i}{M_{tot}}\right), \text{ jusqu'a } \sum_i f_i < 0.2 f_{tot} \tag{5.7}$$

avec

$$M_{tot} = \sum_i^n M_i = \sum_i^n f_i|(x_i - x_c)^2 + (y_i - y_c)^2| \tag{5.8}$$

Lotz et al (2004) ont réalisé des mesures de C, A, S, G et $M20$ sur 170 galaxies proches (classes E, S0, Sa, Sb, Sc, Sd, dI), 73 ULIRGs (*Ultra Luminous Infrared Galaxies*[8]) et 49 galaxies Lyman-break[9], et à partir de ces données ils conclurent que les galaxies non URLIGs (voir Figure 5.10) suivent une séquence $G - M20 - C$ bien définie où les types précoces et les systèmes dominés par le bulbe devraient occuper la région de fort G, fort C, et bas $M20$, alors que les types tardifs occupent la région de bas G, bas C et fort $M20$.

Ils conclurent de plus qu'une combinaison de A et de S avec G forme un trio efficace pour discriminer entre des ULIRGs et des types de Hubble normaux : les ULIRGs occupent une région sur le plan $G - M20$ supérieure à celle occupée par les galaxies normales, et les ULIRGs avec des noyaux multiples occupent des régions

8. Galaxies extrêmement lumineuses dans l'infrarouge ($\sim 10^{12} L_\odot$).

9. Celles-ci sont des galaxies à hauts *redshifts* ($z > 2.5$) qui subissent une rapide formation stellaire. Une revue des propriétés de ces objets peut être trouvé dans Giavalisco (2002).

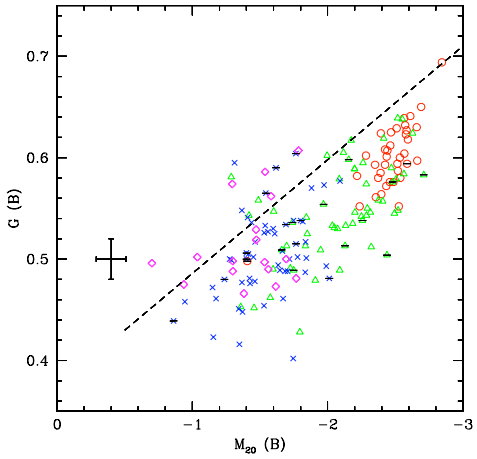


FIGURE 5.9 – M_{20} vs. G calculés pour des galaxies proches. Les symboles indiquent le type de la galaxie : E/S0–cercles, Sa/Sbc–triangles, Sc/Sd–croix, dI–losanges, Spirales edge-on–barres. (extrait de Lotz et al, 2004)

légèrement distinctes des ULIRGs simples dans l'espace morphologique (représentées graphiquement sur la Figure 5.10).

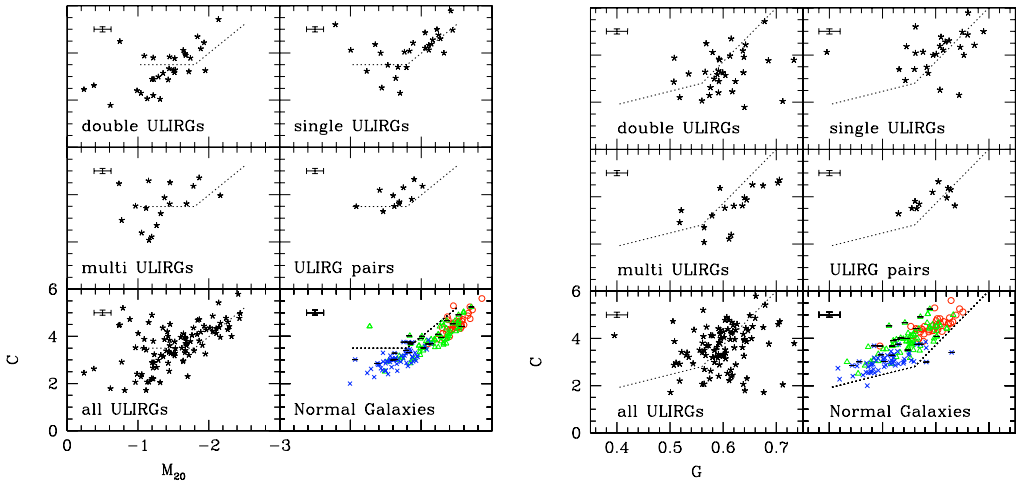


FIGURE 5.10 – C vs. $M20$ et C vs. G calculés pour des galaxies normales et des ULIRGs. Dans la plus grande partie des cas les galaxies normales se localisent dans une région bien définie, et suggèrent un critère multidimensionnel pour classification. (extrait de Lotz et al, 2004)

Ces paramètres ont été depuis quelques années utilisés pour des travaux de classification purement morphologiques de galaxies dont les images contenaient un petit nombre de pixels, comme dans Huertas-Company et al (2010) pour des galaxies avec $z \sim 2$.

5.2.1 Tests avec des données de la littérature—PCA

Dans le cadre de cette thèse, nous avons mené une première étude dans l'espace CASGM20 en effectuant une Analyse en Composantes Principales (*Principal Component Analysis*, ou PCA), dans le but de vérifier l'existence de corrélations significatives entre les paramètres permettant la classification des galaxies concernées.

Idéalement, on souhaiterait classer les galaxies à partir de coupes simples dans un plan bi-dimensionnel, en utilisant des algorithmes complexes d'apprentissage computationnels. Pour vérifier cette possibilité, nous avons réalisé une analyse à partir des paramètres CASGM20 calculés par Lotz et al (2004) (L04) des galaxies du catalogue de Frei et al (1996).

Les analyses de PCA sont basées sur l'idée de réduire le nombre de dimensions d'un ensemble de données quand il existe des corrélations entre les mesures. L'interprétation des composantes principales qui émergent de combinaisons des variables originales peut mener à des conclusions physiques intéressantes (Jolliffe, 2004), dans le cas où certaines composantes principales pourraient reproduire la plus grande partie de la variance présente dans les données, et au cas où ces composants seraient physiquement interprétables. Elles seraient alors susceptibles de fournir une description alternative et plus simple des données.

Cette réduction de dimension est obtenue par la transformation des variables originales en un nouvel ensemble de variables, appelées composantes principales (PCs), qui par définition ne sont pas de corrélées. De telles composantes sont ordonnés de manière à ce que les premières contiennent la plus grande partie de la variance. Dans une forme matricielle, les composantes principales peuvent s'écrire :

$$\mathbf{PCs} = \mathbf{A}^T \mathbf{x}^* \tag{5.9}$$

où $\mathbf{A}$ est la matrice des vecteurs propres de la matrice de corrélation des données et $\mathbf{x}^*$ représente les variables centrées.

L'utilisation de variables centrées tend à minimiser la dépendance d'échelle dans ce type d'analyse. Cependant il faut remarquer que l'utilisation de ce type de variable rend l'interprétation physique des composants obtenus plus difficile.

Les valeurs des composantes principales que nous avons calculées pour les données de L04 sont représentées sur la Figure 5.11. Nous pouvons remarquer que dans les plans *PCi vs. PCy* lorsque $i, j > 2$, les galaxies des différents types morphologiques sont difficilement séparables, démontrant qu'en principe une classification semble être rendue possible en n'utilisant que les deux premières composantes principales.

Cette impression, néanmoins, doit être évaluée sur un critère quantitatif . Même s'il existe des travaux dédiés à la détermination de critères et d'algorithmes pour la sélection du nombre de composantes principales qui doivent être préservés après l'analyse (très souvent par des moyens computationnellement intensifs, tels que le *resampling* et le *bootstrap*), la sélection est toujours un peu arbitraire, car il n'existe pas de critères d'application générale.

Une des règles d'utilisation la plus fréquente est la coupe suivant la proportion cumulative de la variance totale décrit par Jolliffe (2004). Cette règle se base sur

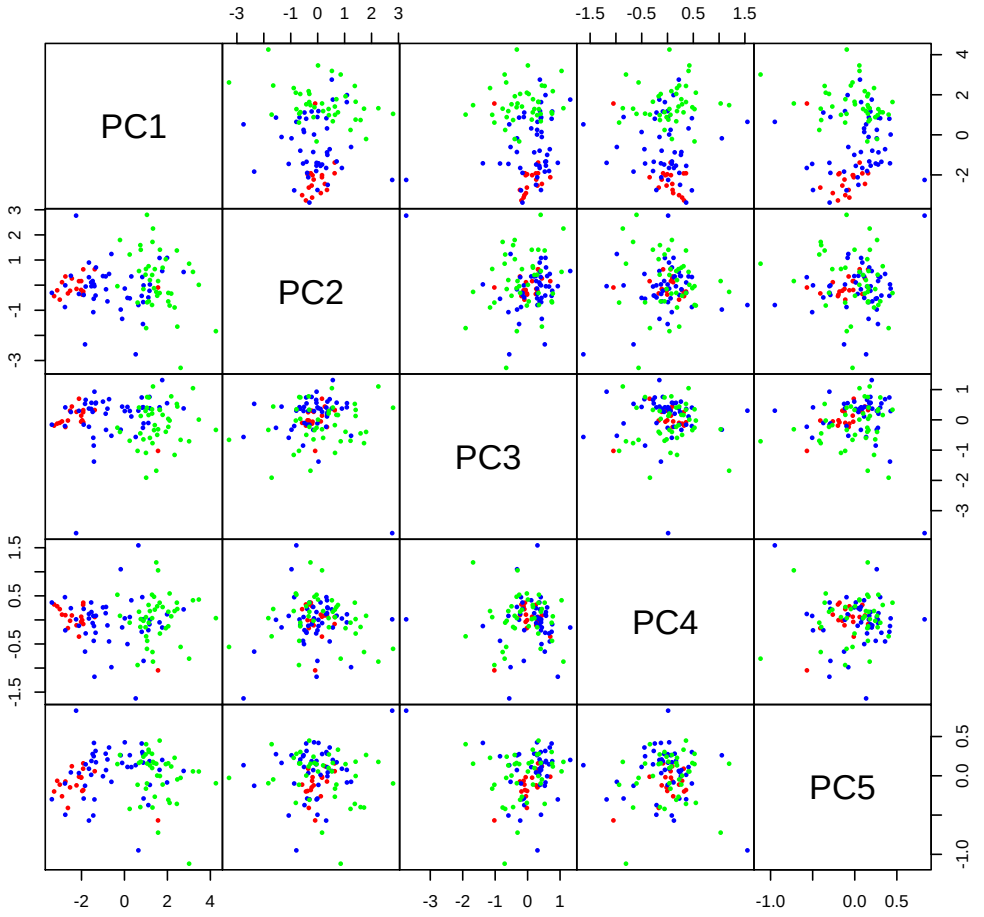


FIGURE 5.11 – Après un changement de base par analyse de composantes principales, les paramètres CASGM20 de L04 ont été transformés en les PCs. La couleur verte indique des galaxies spirales du type b/c et irrégulières, le bleu indique des galaxies spirales de types a/b et le rouge indique des galaxies elliptiques et lenticulaires. Des données d'images dans le filtre B ont été utilisées.

le choix d'une valeur de coupe pour laquelle on désire que la variance des données soit expliquée par les composantes principales – an niveau de 80% et 90%, selon l'auteur. Ceci signifie, que les n composantes principales capables d'expliquer jusqu'à un certain pourcentage de la variance des données doivent être maintenus pour des analyses futures.

Si nous observons les résultats de l'analyse de CASGM20 de L04, nous arrivons à la conclusion que l'espace d'analyse pourrait être réduit aux trois premières composantes principales (suivant la règle de coupe à 90%). Ceci peut être visualisé à partir de la proportion cumulative de la contribution à la variance totale de chacun

des composantes principales représentés dans le Tableau 5.1.

	PC1	PC2	PC3	PC4	PC5
Desvio padrão	1.813	0.971	0.6835	0.4633	0.2988
Proporção da variância	0.657	0.189	0.0934	0.0429	0.0179
Proporção cumulativa	0.657	0.846	0.9392	0.9821	1.0000

TABLE 5.1 – Importance des composantes principales calculés avec des données de L04. La ligne « Proportion de la variance » représente la contribution de la n-ième composante principale à la variance originale des mesures des paramètres CASGM20.

Comme nous possédons la classification visuelle pour ces galaxies, nous avons réalisé la PCA dans chacune de ces classes séparément, dans le but de comparer le résultat obtenu et de vérifier si l'une quelconque de ces classes possédait une particularité par rapport aux autres. Les résultats de la proportion cumulative pour chacune de ces analyses peut être observés sur le Tableau 5.2 ci-dessous.

	PC1	PC2	PC3	PC4	PC5
Proporção cumulativa (E/S0)	0.7240	0.8860	0.9682	0.9967	1.0000
Proporção cumulativa (Sa/b)	0.5560	0.7840	0.9120	0.9755	1.0000
Proporção cumulativa (Sb/c/Irr)	0.4180	0.7460	0.8910	0.9667	1.0000

TABLE 5.2 – Proportions cumulatives des PCs de L04 obtenues dans des PCAs séparées par groupes morphologiques.

Nous pouvons observer sur le Tableau 5.2 que pour les deux premières classes, nous pouvons réaliser la réduction dimensionnelle de trois PCs (en se basant sur la règle de la coupe à 90% de la proportion cumulative). Néanmoins, pour la troisième classe, qui englobe les galaxies du type Sb/c/Irr, nous ne pouvons réduire qu'à quatre dimensions. Du point de vue physique, le besoin d'un plus grand nombre de paramètres pour expliquer la variance dans la mesure où l'on avance dans la séquence de Hubble démontre que ces objets deviennent de plus en plus complexes du point de vue morphologique, et que, d'une certaine manière, les paramètres choisis reflètent cette augmentation de complexité.

Bien que la règle empirique de coupe à 90% de la proportion cumulative permet l'utilisation de seulement trois composantes principales, nous avons décidé d'utiliser les cinq paramètres dans l'analyse de galaxies de Gaia. Cette décision a été prise en tenant compte de la possibilité de dégradation de certains résultats concernant les galaxies Sb/c/Irr (qui demanderaient quatre composantes), en plus du fait que la confiance dans les paramètres CASGM20 mesurés à partir d'images reconstruites avec les données de Gaia ne peut pas encore être clairement déterminée.

5.3 Mise en œuvre pour Gaia

Comme nous l'avons vu dans l'introduction de ce Chapitre, des images Gaia reconstruites ne contiennent pas seulement l'objet étudié, mais aussi divers signaux parasites et des artefacts de reconstruction. Depuis 2007, la *Development Unit* 470 (DU470) est responsable du traitement des objets étendus dans Gaia. C'est dans ce cadre que nous avons proposé l'utilisation des mesures statistiques décrites dans les sections antérieures, alliées à la technique d'apprentissage computationnel supervisé SVM pour la réalisation d'une classification purement morphologique des galaxies non résolues qui seront observées par Gaia.

Nous avons aussi mis au point des codes en Java dédiés au calcul de tous les paramètres CASGM20 dans le pipeline de la *Coordination Unit 4* (Krone-Martins & Ducourant, 2008). Ces paramètres ont été choisis en tenant compte du besoin réduit d'information spatiale, de la robustesse à la dégradation et de la facilité de pouvoir être modifiés pour l'inclusion future de cartes de poids, permettant une utilisation plus fiable des images reconstruites.

Pour cette mise en place, nous avons supposé que l'objet étudié serait l'unique objet présent sur l'image reconstruite, et que toutes les autres contributions (rayons cosmiques, autres objets sur la ligne de visée, etc.) ont été pris en compte durant les procédures antérieures. Cette supposition est raisonnable car comme l'image est reconstruite à partir de diverses observations, les rayons cosmiques ont été exclus, et comme ces observations ont déjà été analysées préalablement par le *Source Environment Analysis* décrit dans le Chapitre 3, des sources secondaires ont déjà été détectées préliminairement.

Néanmoins, nous avons inclus la possibilité de traiter le fond de ciel. Le fond est déterminé à partir d'un rejet itératif à 3σ, et d'une binarisation de l'image dans laquelle les pixels au-dessus d'un certain niveau de coupe (configurable) sont considérés comme appartenant à l'objet. Sur cette carte de fond peuvent être inclus les pixels appartenant à des objets quelconques ou des signaux additionnels qui seraient présents sur l'image et qui ne doivent pas être pris en compte dans l'analyse.

Les définitions adoptées pour les paramètres qui sont mesurés par notre mise en œuvre sont résumées dans les équations suivantes :

$$\left(\begin{array}{ll} R_P(r) = \frac{\int_{0.8r}^{1.25r} dr' 2\pi r' I(r') / [\pi(1.25^2 - 0.8^2) r^2]}{\int_0^r dr' 2\pi r' I(r') / (\pi r^2)} & C = 5 \log_{10} \left(\frac{r_{80}}{r_{20}} \right) \\ M_{20} = \log_{10} \left(\frac{\sum_l M_l}{M_{tot}} \right), \text{for } \sum_l I_l \leq 0.2 I_{tot} & A = \frac{\sum_i \sum_j |I(i,j) - I_R(i,j)|}{\sum_i \sum_j I(i,j)} \\ G = \frac{\sum_k [2k - N - 1] I_k}{\langle I \rangle N(N-1)} & S = \frac{\sum_i \sum_j |I(i,j) - I_S(i,j)|}{\sum_i \sum_j I(i,j)} \end{array} \right) \quad (5.10)$$

Dans ces équations, $R_P(r)$ définit le taux pétrosien, tel qu'utilisé dans le Sloan Digital Sky Survey (Blanton et al, 2001) et le rayon pétrosien est adopté comme étant le Rayon r pour lequel $R_P(r) = 0.2$. $I(r)$ est le profil de brillance superficielle

moyen, r_{80} et r_{20} sont les rayons qui englobent 80% et 20% du flux total de la galaxie (défini comme le flux dans 1.5 rayons pétrosiens). M_l est le moment d'ordre 2 du l-ième pixel ordonné par flux décroissant. M_{tot} est le moment de deuxième ordre calculé pour tous les pixels. I_l le flux du l-ième pixel. $I(i,j)$ le flux de l'image dans le pixel (i,j). $I_R(i,j)$ le flux de l'image dans le pixel (i,j) après rotation de 180°. I_k le flux du k-ième pixel ordonné par flux croissant. N le nombre total de pixels dans l'image. $I_S(i,j)$ le flux de l'image convoluée par un filtre du type boxcar[10] avec une taille de $0.3R_P$ dans le pixel (i,j). Sur la Figure 5.12, ci-dessous, nous donnons l'exemple du calcul du paramètre S pour la galaxie NGC3351.

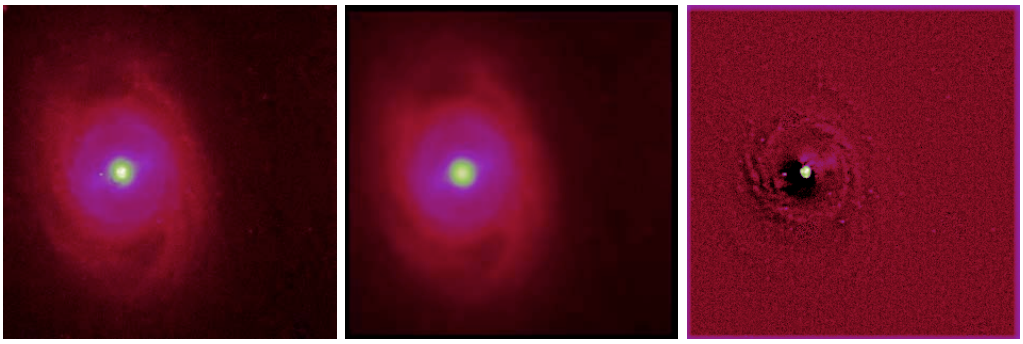


FIGURE 5.12 – Exemple du calcul du paramètre S pour la galaxie NGC 3351. **Gauche** : image originale I, du catalogue de Frei et al (1996). **Centre** : image lissée I_S. **Droite** : image résultant de la procédure de soustraction, à partir de laquelle S est calculé. La table de couleur utilisée dans dernière image n'est pas la même que celle des deux autres images I et I_S.

Les calculs de ces paramètres sont réalisés par des classes Java complètement indépendantes (dénommées *Measurers*) ; ainsi tout traitement pourrait, en principe, être réalisé de façon complètement parallèle. Cependant, comme le calcul est relativement rapide du fait de la petite quantité de pixels disponibles dans les images reconstruites, ces paramètres sont calculés dans la mise en œuvre actuelle de façon séquentielle. Ces valeurs de CASGM20 peuvent être stockées directement dans la base de données du centre de traitement où le code sera exécuté (en l'occurrence le CNES-Toulouse).

Il faut noter ici, que les images reconstruites à partir des données Gaia pourront être accompagnées d'une carte d'erreur estimée par l'algorithme de reconstruction pour chacun des pixels. Ceci signifie que le calcul des paramètres CASGM20 défini ci-dessus pourra être modifié pour tirer profit de cette information additionnelle. La façon la plus directe de le faire, et que nous envisageons utiliser, sera d'appliquer une pondération durant le calcul de tels paramètres.

10. Un filtre qui est défini par une fonction rectangulaire.

5.3.1 Tests avec des galaxies du catalogue de Frei

L'un des tests réalisé pour la mise en œuvre des codes que nous avons développés pour Gaia, consiste à vérifier la robustesse des paramètres CASGM20 face au petit nombre de pixels disponibles dans les images reconstruites.

Pour cela, nous avons calculé les valeurs de CASGM20 pour 82 galaxies proches représentantes des trois types morphologiques principaux (elliptiques, spirales et irrégulières), originaires du catalogue d'images de galaxies de Frei et al (1996). Ces images ont été re-échantillonnées à partir de l'intégration de *splines* bi-cubiques fittés sur l'image originale, pour diverses échantillonnages inférieurs à celui de l'image originale (entre 81x81 pixels et 11x11 pixels).

Les rayons pétrosiens et les magnitudes calculées sur les images résultantes varient entre $\sim$ 46.5 mas et $\sim$ 1.5″ (considérant 30 mas/pixel), et de 11.40 à 19.98 mag, respectivement, ce qui est comparable à ceux qui doivent être observés par Gaia. Un exemple des images utilisées pour cette étude, dans le cas de la galaxie NGC 3893, est présenté sur la Figure 5.13.

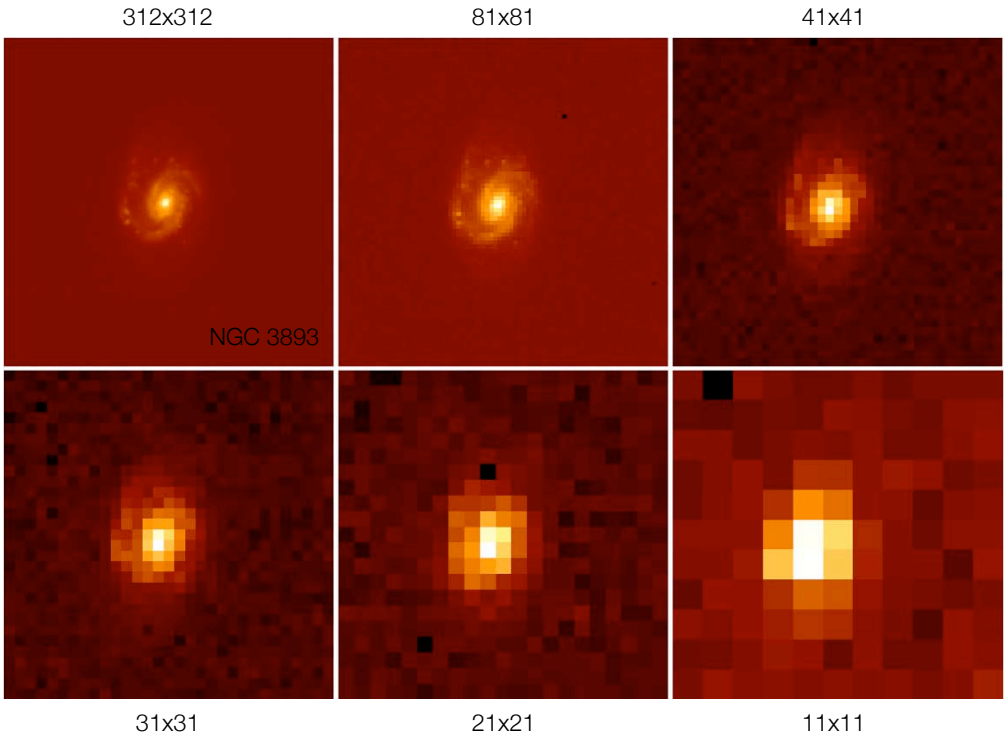


FIGURE 5.13 – Exemples de ré-échantillonnages d'une image de la galaxie NGC 3893 utilisés pour l'étude de la variation des paramètres CASGM20

Les paramètres CASGM20 ont été calculés pour toutes ces images, après avoir défini la grandeur (nommée MPV) pour quantifier la variation des valeurs calculées de

ces paramètres en fonction de la taille de l'image. Cette grandeur est la moyenne des variations en pourcentage pour chacun des paramètres calculé pour chaque galaxie individuelle. Pour N galaxies, et pour le paramètre C, elle peut être écrite de la manière suivante :

$$\mathrm{MPV}(C) = \frac{1}{N}\sum_{1}^{N}\frac{C(\text{Frei image}) - C(\text{subsampled image})}{C(\text{Frei image})} \tag{5.11}$$

Une comparaison entre les valeurs du catalogue original de Frei, et celles calculées à partir des images re-échantillonnées en utilisant la moyenne de la variation en pourcentage est montrée sur la Figure 5.14. Sur ce graphique, nous pouvons noter que le paramètre S diverge rapidement de sa valeur originale – ceci est prévisible, car moins il y a de pixels disponibles, plus la coupe est grande dans les fréquences spatiales élevées, et ce paramètre mesure justement la rugosité que doit être rencontrée dans des fréquences spatiales élevées.

Cependant, la comparaison pour tous les paramètres montre que l'espace CASGM20 comme un tout, est robuste à l'absence d'information, pouvant être effectivement utilisé dans des images composées par peu de pixels, telles que celles de la mission Gaia. Nous pouvons remarquer que la variation moyenne en pourcentage pour la plus grande partie des paramètres se situe au un niveau de 20%, stable même pour les petites images.

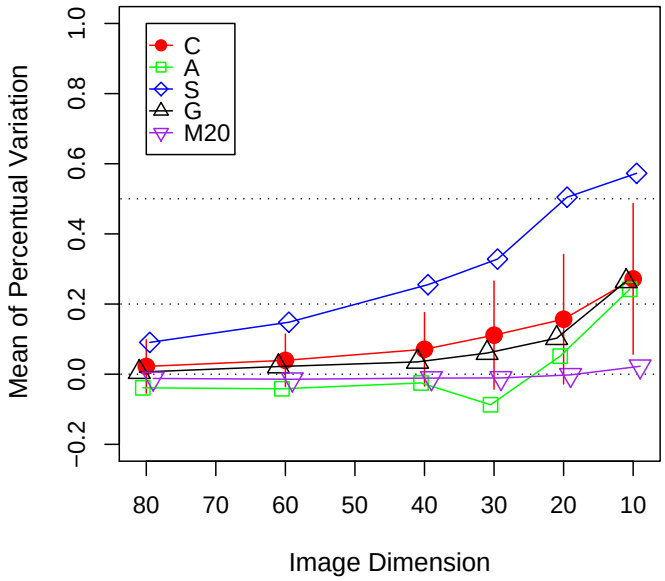


FIGURE 5.14 – Variation moyenne en pourcentage des paramètres CASGM20 en fonction de la dimension de l'image utilisée pour le calcul du paramètre. Les symboles sont légèrement déphasés entre eux pour faciliter la visualisation. De plus, des barres d'erreur de 1σ ont représentées seulement pour le paramètre C, vu qu'elles sont représentatives de toutes les autres.

Ces résultats (Krone-Martins et al, 2008b) démontrent que les paramètres CASGM20 peuvent être utilisés avec succès pour l'analyse d'images reconstruites, étant donné que ces images souffriront principalement de la faible quantité de pixels disponible pour l'analyse.

Néanmoins, bien que le test avec les galaxies proches re-échantillonnées démontre que les paramètres sont robustes même sur des images avec de petits échantillonnages, des effets associés à l'observation de galaxies angulairement petites peuvent ne pas encore être perçus. Pour cela un test additionnel, vérifiant si même pour des galaxies angulairement petites, des types morphologiques distincts occupent des régions distinctes de l'espace CASGM20, est nécessaire.

5.3.2 Calcul de CASGM20 sur des images du *Hubble Deep Field*

Le *Hubble Deep Field North* (HDF) représente une région du ciel mise en image par le télescope spatial Hubble grâce à de longues expositions ($\sim$ 120.000s de temps intégré par filtre), dont la description peut être trouvée dans Williams et al (1996) – il existe un champ équivalent pour l'hémisphère Sud.

Les galaxies observées dans ce champ ne seront pas observées par Gaia car la plus grande partie des objets présents ont une magnitude $G \gg 20$. Cependant Gaia doit observer un grand nombre de galaxies petites et de redshift faible, similaires morphologiquement aux galaxies du HDF. Ainsi, le HDF peut être considéré comme une bonne source galaxies angulairement petites sur lesquelles nous pourrons appliquer les codes que nous avons développés pour Gaia sur des.

Dans ce test, l'objectif est de vérifier comment les paramètres CASGM20 calculés avec notre code dans les images des galaxies qui peuvent être observées dans l'HDF se comportent en fonction du type morphologique de la galaxie, et principalement de vérifier si ces paramètres permettent une classification postérieure. Les galaxies utilisées pour le test possédaient une classification morphologique préliminaire de van den Bergh et al (1996).

Les images des galaxies individuelles que nous avons utilisées pour les calculs de CASGM20 (dénommées « timbres ») ont été construites à partir de la somme des images obtenues avec les différents filtres utilisés dans le HDF-N.[11] Après la construction de ces « images blanches », les objets présents dans les quatre CCDs différents ont été détectés et recoupés en timbres. Comme les positions des galaxies dans l'article van den Bergh et al (1996) ont été mesurées dans la première réduction du HDF, tandis que nous travaillons avec leur réduction finale, il a été nécessaire de réaliser une procédure d'identification croisée entre les objets classés et les objets détectés.[12] Certains exemples de timbres obtenus par cette procédure pour les trois types différents de galaxies peuvent être observés sur la Figure 5.15.

11. Les filtres F300, F450, F606 et F814 couvrent pratiquement tout l'intervalle spectral de Gaia.
12. La détection des objets dans les images originales, l'identification croisée et les coupes en timbres ont été réalisées par J.–F. le Campion, du Laboratoire d'Astrophysique de Bordeaux.

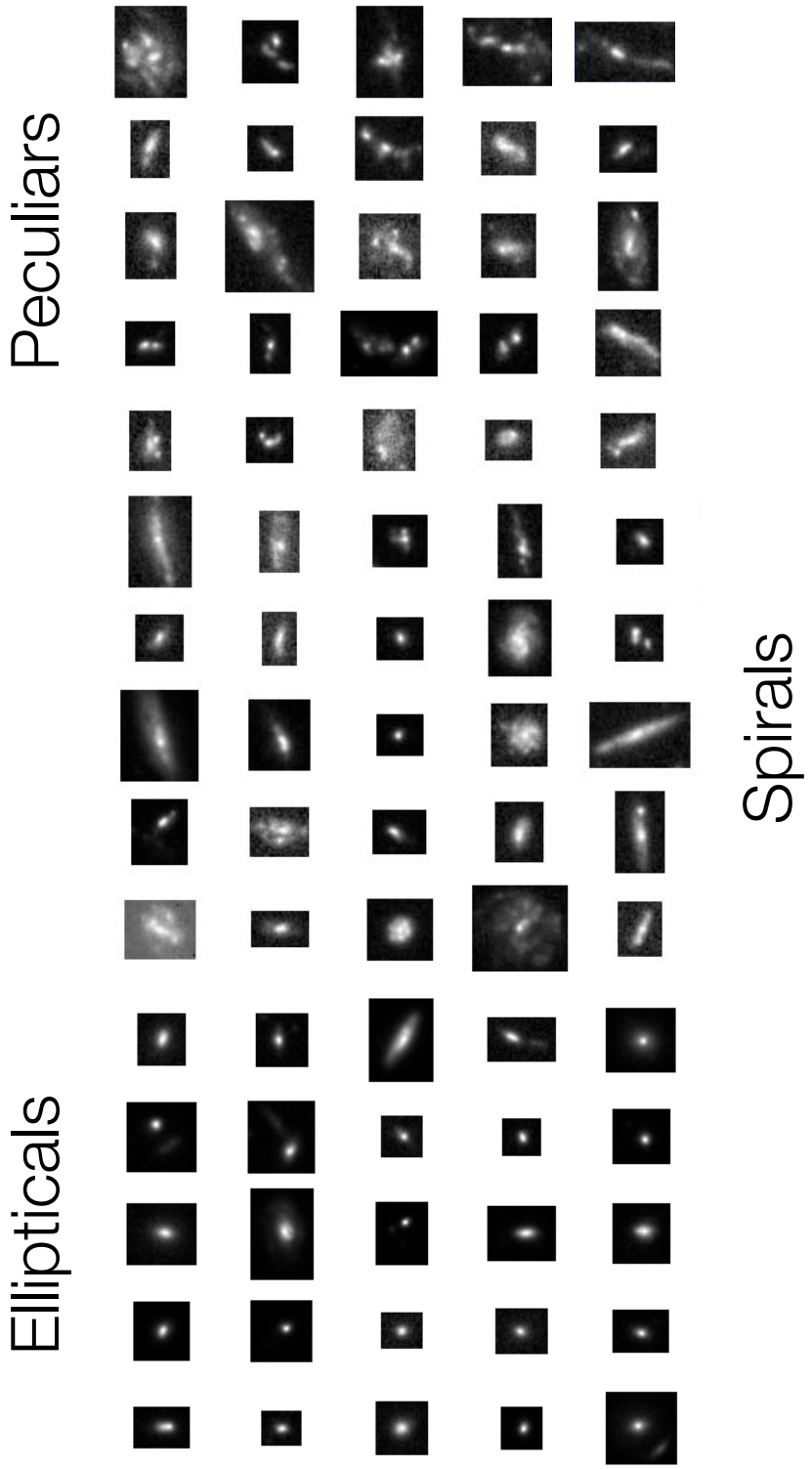


FIGURE 5.15 – Images de galaxies découpées du *Hubble Deep Field North*.

Les paramètres CASGM20 calculés pour les différents types de galaxies sont présentés sur la Figure 5.16. Nous pouvons vérifier qu'il est possible de séparer des régions des différents plans de l'espace CASGM20 où certains types de galaxies sont prédominants. Cependant, la superposition entre les différents types est aussi relativement grande, principalement parce que les galaxies spirales occupent aussi bien la région des galaxies elliptiques que celle de galaxies irrégulières.

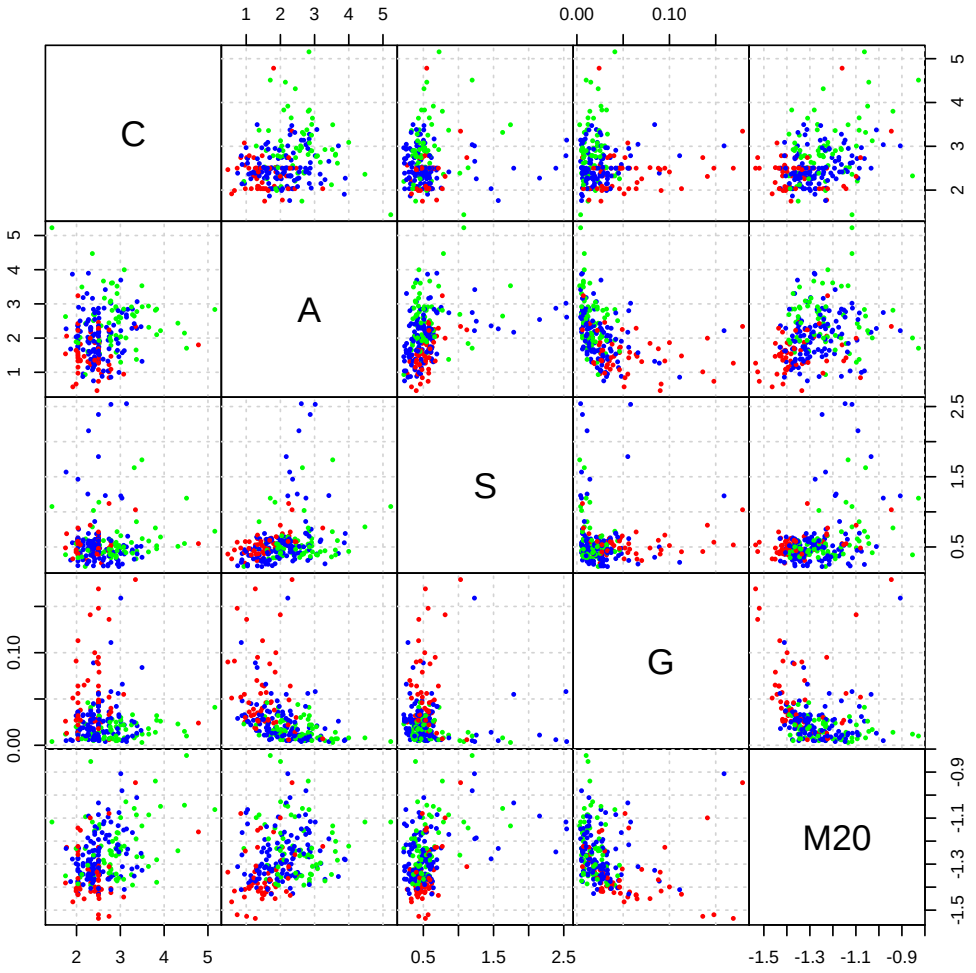


FIGURE 5.16 – Paramètres CASGM20 calculés pour les galaxies du HDF-N. La couleur verte indique des galaxies Irrégulières et Particulières, le bleu indique des galaxies Spirales et le rouge indique des galaxies Elliptiques.

Cette grande superposition peut indiquer des erreurs durant l'identification croisée entre la liste d'objets classés visuellement et la liste d'objets recoupés des images du HDF ou alors que dans le cas de ces galaxies angulairement petites, la distinction entre les différents types morphologiques devient très incertaine, car les interfaces entre les différents types sont moins bien définis (un fait qui peut être observé en examinant les exemples de la Figure 5.15).

5.3.2.1 Analyse en Composantes principales – PCA

Dans le but de vérifier la possibilité de simplifier l'espace de classification, une PCA a été appliquée aux données CASGM20 calculées sur les images des galaxies du HDF. Les valeurs obtenues pour les différents PCs peuvent être retrouvées sur la Figure 5.17.

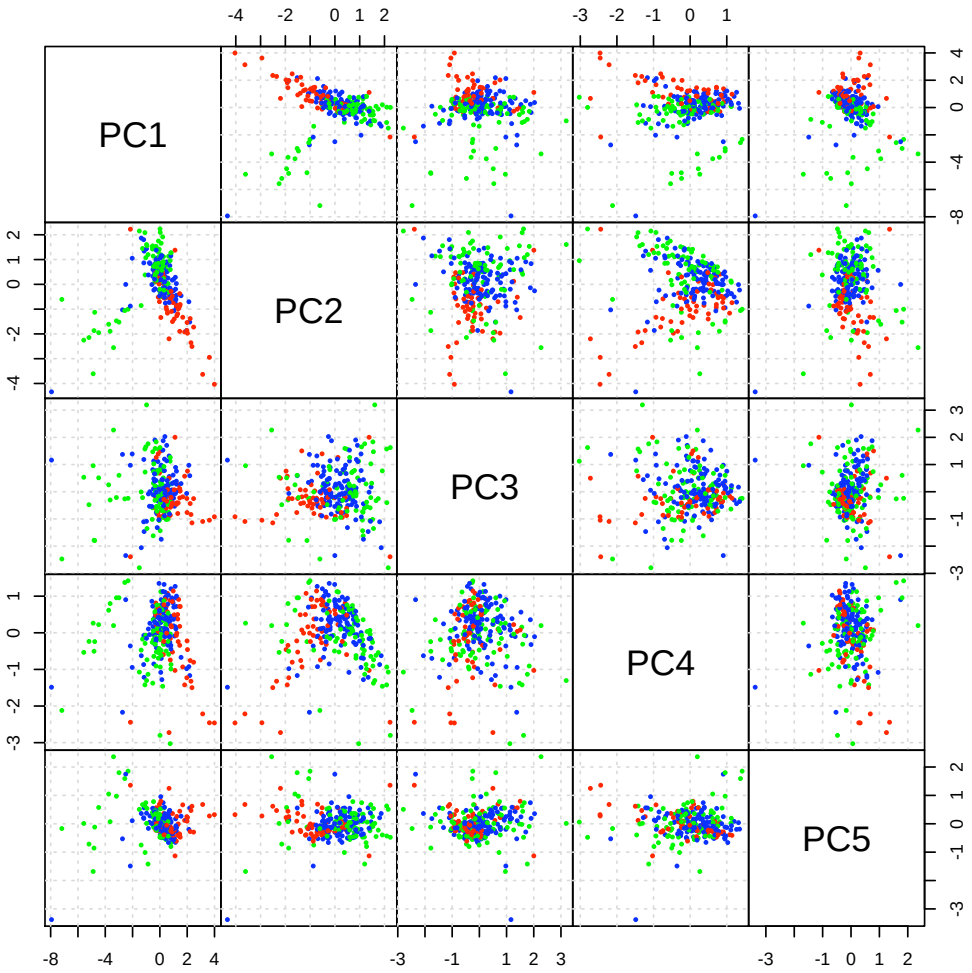


FIGURE 5.17 – Données CASGM20 calculées pour un sous-ensemble des galaxies du HDF dans l'espace des composantes principales. Verte = irrégulières et particulières, bleu = spirales et rouge = elliptiques.

Un résultat intéressant qui peut être observé dans le plan des deux premiers composantes principales est la présence d'une séquence qui va des galaxies elliptiques aux irrégulières, avec les objets classés comme Irréguliers et Particuliers (représentés en vert) qui définissent aussi un autre axe, qui coupe la séquence précédemment

décrite dans la région des galaxies spirales.

Ce comportement n'avait pas été observé durant l'analyse en PCA des galaxies proches, mais la comparaison de ces analyses est complexe car les PCs des deux analyses n'ont pas la même signification, car étant formés de combinaisons linéaires différentes des paramètres ; l'interprétation physique des PCs et la comparaison entre des PCs provenant d'analyses distinctes est une tâche relativement complexe (Jolliffe, 2004), nécessitant un travail indépendant.

	PC1	PC2	PC3	PC4	PC5
Écart type	1.457	1.106	0.859	0.791	0.539
Proportion de la variance	0.424	0.245	0.147	0.125	0.058
Proportion cumulative	0.424	0.669	0.817	0.942	1.000

TABLE 5.3 – Importance des composantes principales calculées à partir de paramètres CASGM20 mesurées dans des galaxies du HDF-N.

D'un point de vue qualitatif, l'analyse de la Figure 5.17 nous permet de penser qu'une bonne classification pourrait être obtenue en ne considérant que les deux, ou trois premières composantes principales. Néanmoins, la proportion cumulative des différents PCs (Tableau 5.3) montre qu'il serait nécessaire de conserver jusqu'à la quatrième composante si l'on veut que la variance soit expliquée à plus de 90%. De cette manière, la simplification d'un seul paramètre ne justifierait pas l'utilisation d'une PCA sous les paramètres CASGM20 dans le cas des galaxies du HDF.

5.3.3 Tests avec simulations

Un test final est la vérification du comportement de ces paramètres quand ils sont mesurés à partir d'images reconstruites par l'un des algorithmes de reconstruction d'image proposés et décrits dans le Chapitre 2. Pour cela, nous avons réalisé des simulations d'observations de galaxies avec GIBIS 7 et avons utilisé le code de reconstruction d'images qui met en place l'algorithme *ShuffleStack*.

Les coordonnées choisies pour la simulation correspondent à une région avec une bonne couverture angulaire, $(l, b) = (120°, 45°)$[13], générant un maximum de 82 passages individuels.[14] Les galaxies simulées balayent un large intervalle de paramètres, avec des magnitudes dans l'intervalle [13 ; 20], demi grand axe dans [200 ; 1200] mas et rapport axial dans [1 ; 10], générant 880 galaxies pour chaque type morphologique simulé (elliptiques, spirales et irrégulières). Un exemple avec les galaxies simulées dans une partie du plan focal est montré sur la Figure 5.18.

13. Voir Chapitre 2 pour une carte de couverture du ciel en son entier.

14. Le nombre exact de passages a été de 82, néanmoins toutes les galaxies ne sont pas observées à tous les passages à cause de la spécificité des systèmes de lecture de simulations GIBIS : les objets qui, au cours d'un passage donné occupent des colonnes de CCDs qui n'interceptent pas le centre de la simulation, ne peuvent pas avoir leurs données lues par les codes de conversion de simulations GIBIS sur des tableaux FITS.

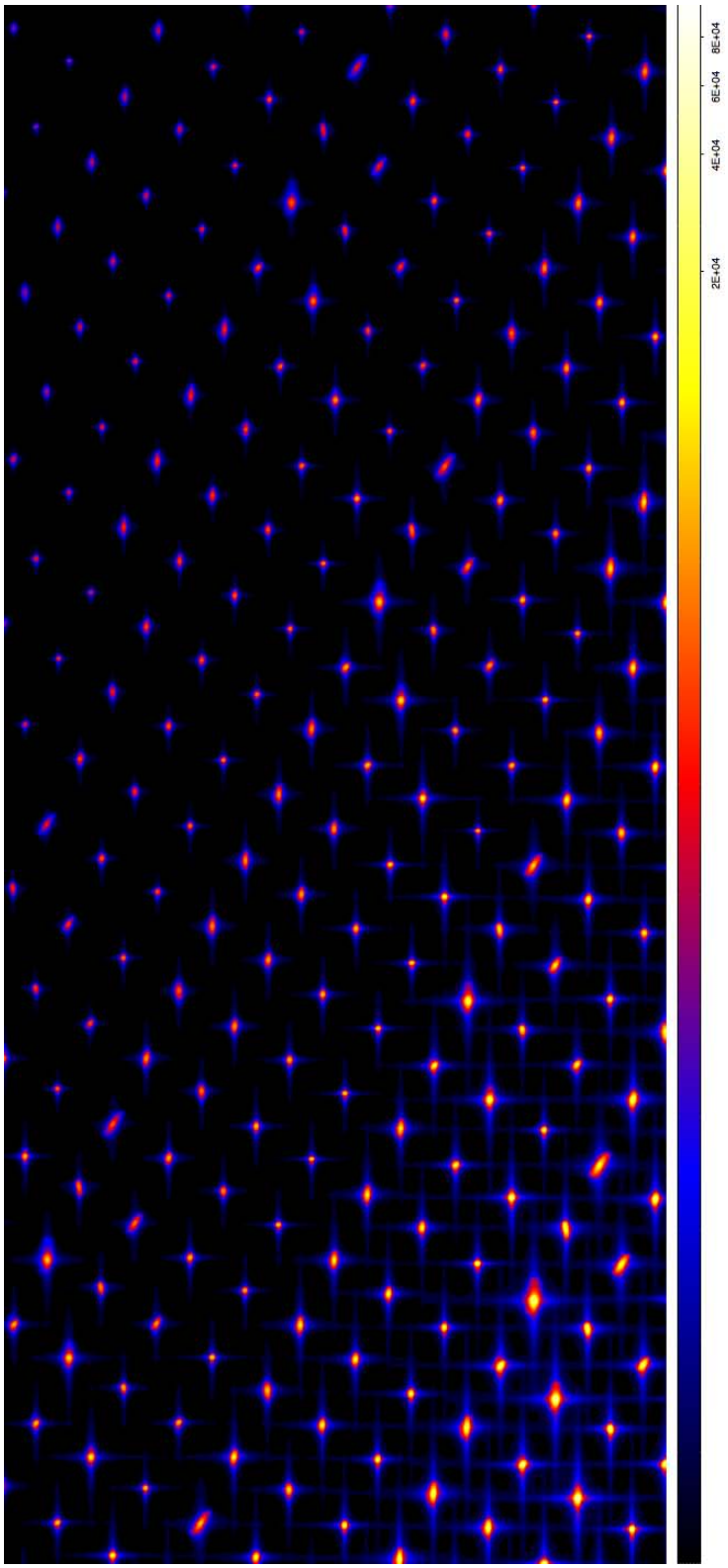


FIGURE 5.18 – Exemple de galaxies simulées dans le plan focal de Gaia. Les objets représentés sont des galaxies spirales couvrant une large bande de magnitudes.

Ces simulations ont été utilisées pour reconstruire des images. Pour cela, l'algorithme *ShuffleStack* a été adopté, avec une échantillonnage de 30mas/pixel. Les codes de mesures de paramètres CASGM20 ont été utilisés pour mesurer ces images reconstruites, et les résultats obtenus sont présentés sur la Figure 5.19.

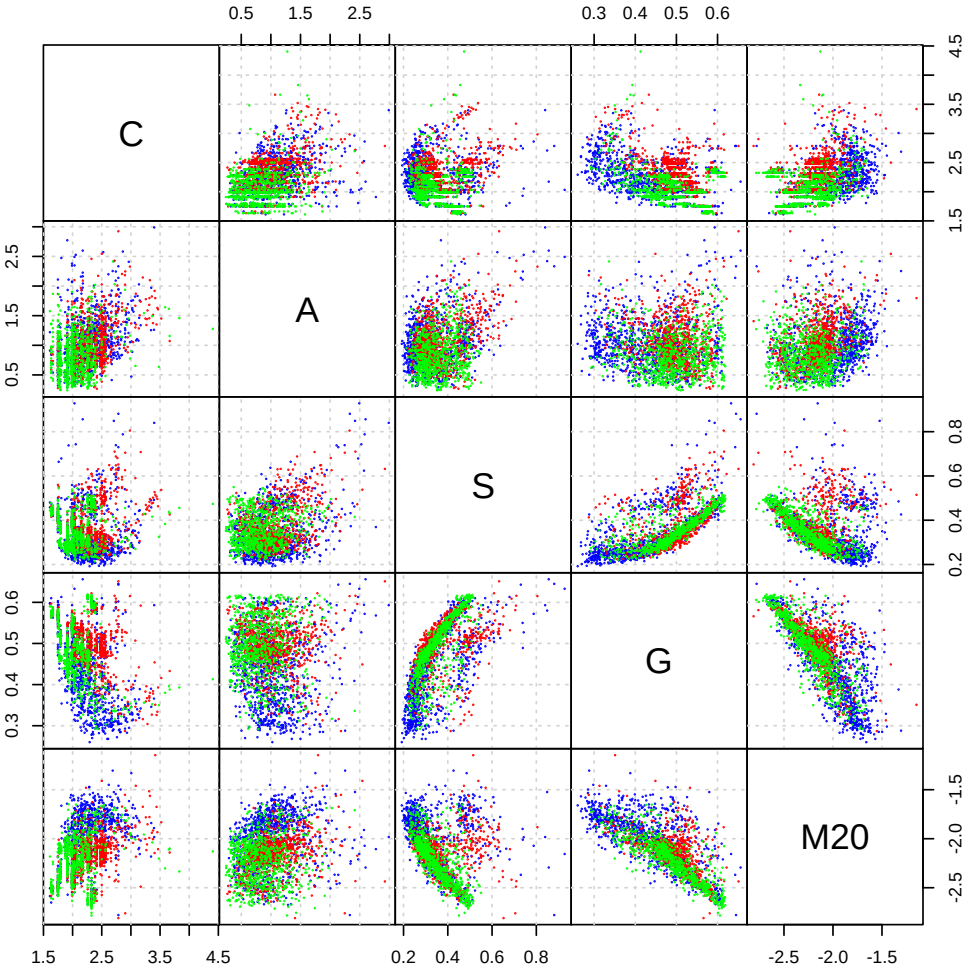


FIGURE 5.19 – Données CASGM20 calculées sur les images reconstruites pour 2600 galaxies de types elliptique (rouge), spirale (bleu) et irrégulière (vert). Les observations ont été simulées par GIBIS 7.1 avec l'utilisation de MAGIL, et la reconstruction utilise l'algorithme *ShuffleStack*.

On peut tout d'abord noter sur la Figure 5.19 que des types morphologiques distincts occupent des régions distinctes dans les différents plans représentés, ce qui doit permettre une classification entre ces types morphologiques, même si le recouvrement entre ces types est grand.

Contrairement au cas du HDF, l'interprétation physique des valeurs de CASGM20 pour des images reconstruites par *ShuffleStack* (qui n'a pas été créé pour des objets étendus) est quelque chose qui n'est pas simple. Ceci peut être constaté en comparant

les résultats obtenus dans cette section à ceux obtenus à partir des mesures de galaxies du HDF présentées dans la section antérieure : les types morphologiques distincts occupent des positions différentes de celles occupées par les données du HDF. C'est un effet de la reconstruction imparfaite d'image (des exemples de reconstructions peuvent être vus sur la Figure 5.20). Néanmoins, ces résultats démontrent qu'il est possible d'utiliser les mesures CASGM20 pour la réalisation de classifications morphologiques de galaxies avec les données de Gaia.

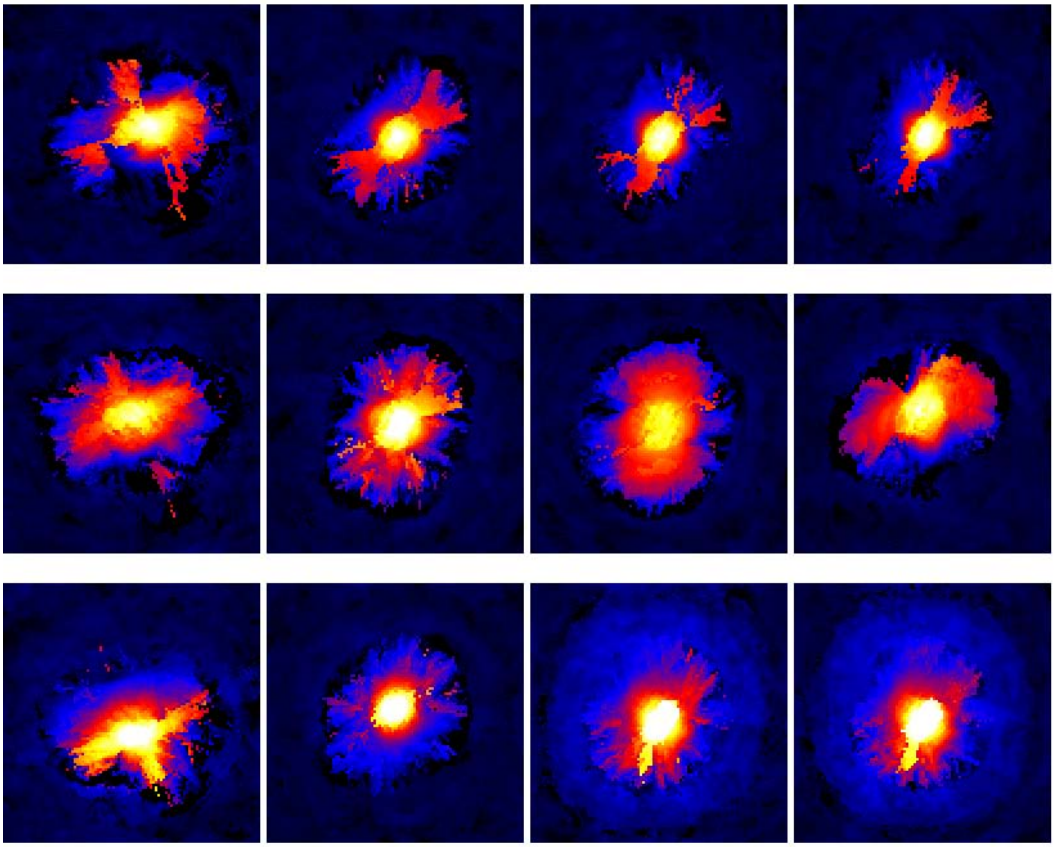


FIGURE 5.20 – Exemples de reconstructions de galaxies irrégulières, spirales et elliptiques (sur chaque ligne) avec l'utilisation de l'algorithme *ShuffleStack*. L'échantillonnage choisi pour la reconstruction est de 30mas/pixel, et les objets possèdent une magnitude $G \sim 17$.

Une analyse de PCA, réalisée dans l'intention de vérifier s'il serait possible d'adopter un système de classification simple, basé sur des coupes dans un plan ou dans un espace tridimensionnel, donne des résultats similaires à ceux obtenus antérieurement (seul le dernier composant pourrait être exclu). Des algorithmes de classification plus élaborés que de simples coupes doivent être donc appliqués dans le volume CASGM20 si on veut obtenir des estimations quantitatives des taux de succès possibles dans la classification morphologique à partir des données de Gaia.

5.4 Classification et Support Vector Machines

Après avoir vérifié que les paramètres CASGM20 peuvent être utilisés pour la classification morphologique de galaxies, car différentes classes de galaxies, même superposées, occupent des positions distinctes dans cet espace, il est maintenant nécessaire de choisir une méthode pour identifier automatiquement les régions caractéristiques des différentes classes à partir d'exemples, permettant une classification postérieure de tous objets dont la classe est inconnue.

Dans la littérature de classification ce problème est connu comme étant un problème d'apprentissage supervisé (*supervised learning*), car des exemples sont fournis au système. Du point de vue mathématique, il peut être vu comme la détermination de fonctions $D : X \rightarrow Y$, dénommées fonctions de décision de la classification à partir d'un ensemble d'entrainement $\{(x_1, y_1), ..., (x_n, y_n)\}$. La méthode de k-NN, vue dans le Chapitre 3, est un exemple d'algorithme d'apprentissage supervisé.

Il existe diverses méthodes construites pour résoudre ce problème, chacune d'entre-elles avec ses avantages et inconvénients. Quelques exemples sont le k-NN, les Réseaux Neuraux Artificiels, les *Support Vector Machines* (SVMs), et les *Revelance Vector Machines* (RVMs).[15]

Selon Graf & Wichmann (2004) les SVMs et les RVMS sont les algorithmes qui reproduisent le mieux les résultats obtenus par notre cerveau.[16] Dans cette étude la corrélation entre des objets classés correctement et incorrectement par le cerveau humain et par des algorithmes d'apprentissage computationnel atteignent ∼70% et sont les méthodes avec la plus grande corrélation.

Pour la classification morphologique de galaxies à partir d'images reconstruites de Gaia, nous avons décidé d'utiliser les SVMs en raison de trois caractéristiques très intéressantes de ce type de classificateur : absence de minimums locaux, robustesse aux *outliers* et maximisation de la capacité de généralisation.

Spécifiquement, les SVMs sont formées dans le but de rencontrer la fonction de décision qui maximise la capacité de généralisation. Ceci est réalisé en mappant l'espace original à l dimensions d'observations dans un hyperespace m-dimensionnel (avec m>l), dans lequel l'hyper-plan qui sépare les deux classes est déterminé au moyen de la solution d'un problème de programmation quadratique – et ainsi, l'existence d'une solution globale optimale est garantie.

De plus, l'hyperplan séparateur est calculé de façon à maximiser sa distance par rapport aux données, et par là la capacité de généralisation de la SVM. Finalement, la robustesse aux *outliers* est fournie par un paramètre de marge, qui équilibre un rapport entre la maximisation de la marge et la minimisation de l'erreur de classification. Pour obtenir des détails sur les SVMs, ainsi que des déductions des équations que nous présenterons dans cette section, voir Vapnik (2000) et Abe (2005).

15. La RVM, ou Relevance Vector Machine, est une nouvelle méthode d'apprentissage computationnel introduite par Tipping (2001), qui introduit un formalisme Bayésien durant la construction d'un SVM, et qui n'a pas encore été utilisée en Astronomie.

16. Il faut noter que ceci ne signifie pas que notre cerveau fonctionne comme ces algorithmes, mais que tous les deux donnent des résultats similaires pour un même problème de classification.

L'équation pour l'hyperplan qui sépare les classes, peut être écrite par :

$$D(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{5.12}$$

où $\mathbf{w}$ est un coefficient vectoriel et b est un coefficient scalaire. Ces deux coefficients doivent être déterminés à partir des données de l'ensemble de formation $\{(x_1, y_1), ..., (x_i, y_i))\}$, et y_i peut prendre les valeurs "1" ou "-1" selon que x_i appartient à la classe 1 ou 2.

Comme la fonction 5.12 doit prendre des valeurs positives pour $y_i = 1$ et négatives pour $y_i = -1$, nous pouvons écrire :

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \tag{5.13}$$

Donc, $D(x) = c$, avec $-1 < c < 1$ forme l'hyperplan de séparation recherché, étant entendu que la valeur de c (dénommée marge) traduit la distance entre l'hyperplan séparateur et les marges des données. Schématiquement, une représentation bidimensionnelle d'un hyperplan optimum conjointement avec ses vecteurs de support (représentés par des couleurs plus fortes), peut être vue sur la figure 5.21.

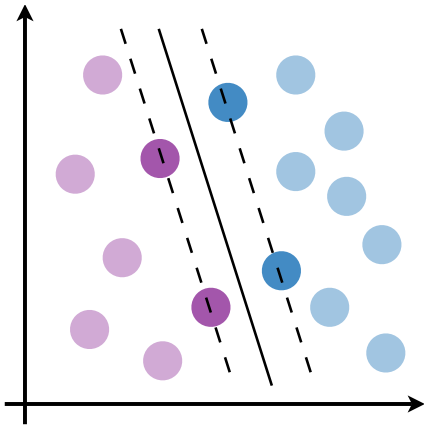


FIGURE 5.21 – Représentation de l'hyperplan optimum qui sépare les données de deux classes. Les *Support Vectors* sont les vecteurs coupés par les deux lignes pointillées.

Pour la détermination des paramètres de l'hyperplan optimum, une *Support Vector Machine* résout le problème d'optimisation suivant :

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{sujet à} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \\ & i = 1, ..., l. \end{aligned} \tag{5.14}$$

Cependant, la formulation exprimée ci-dessus n'est pas idéale dans le cas de données inséparables linéairement. Dans ce cas là, il est préférable d'inclure des variables de pénalité dans le problème. Ces variables peuvent être interprétées comme

les représentations de la distance entre les plans générés par les *Support Vectors* (les marges) et les données supplémentaires que l'on rencontre. Un exemple de ces variables, dénommées ξ, peut être vu sur la Figure 5.22.

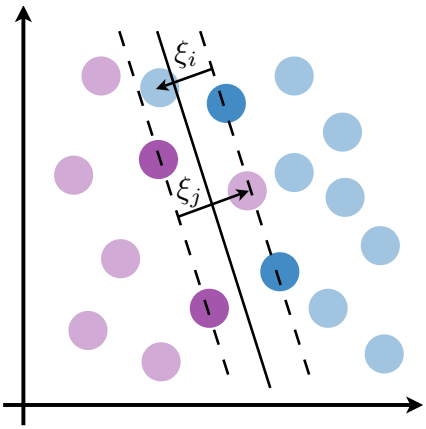


FIGURE 5.22 – Représentation de l'hyperplan optimum qui sépare les données de deux classes montrant les vecteurs qui pénalisent la classification et les paramètres ξ.

Ce type de *Support Vector Machine* est dénommé C-SVM, et est illustré par le problème d'optimisation suivant :

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{sujet à} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, ..., l. \end{aligned} \tag{5.15}$$

Dans le problème exprimé ci-dessus, C est une variable qui contrôle l'équilibre entre la maximisation de la marge et la minimisation de l'erreur de classification, et son choix définit un problème en lui-même.

Néanmoins, même avec l'addition des variables ξ, le problème continue à être inséparable linéairement. Ces variables permettent que le problème se comporte mieux en ajoutant une pénalité sur certains points. Cela augmente la valeur d'une fonction qui doit être minimisée. Pour permettre la séparation de données qui sont linéairement inséparables dans l'espace original, on utilise l'« astuce du noyau ».

L'astuce du Noyau (*Kernel Trick*)

Pour améliorer la capacité d'une Support Vector Machine à généraliser des données qui ne sont pas linéairement séparables, on applique une transformation des données originales de manière à les transporter dans un espace dans lequel elles sont séparables par hyperplans. Ceci est connu sous le nom d'« astuce du noyau » (ou *Kernel Trick*).

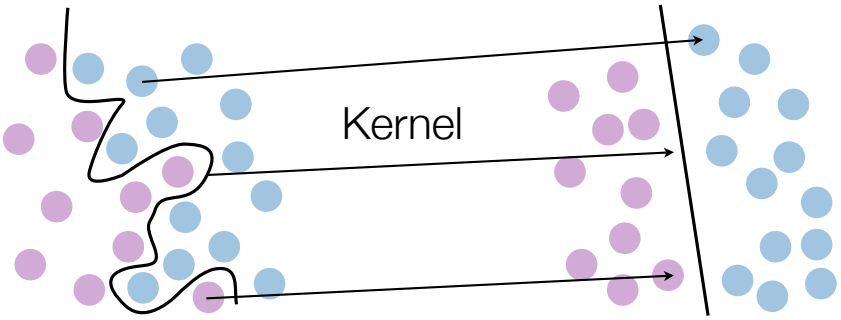


FIGURE 5.23 – Représentation de l'astuce du noyau. Des données initialement séparables seulement par des courbes complexes deviennent séparables par des hyperplans dans un autre espace après l'application d'une transformation.

Le noyau choisi pour la transformation peut être de plusieurs types : linéaire, polynomial, radial, *spline*, etc. Dans ce travail nous avons utilisé le noyau dénommé *Radial Basis Function*, en fonction de ses caractéristiques d'impulsion localisée. Cette fonction peut être écrite de la manière suivante :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||^2) \tag{5.16}$$

où γ est un paramètre qui contrôle le rayon du noyau.

Ainsi, le problème se résume à trouver la solution du problème :

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{sujeito a} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x_i}) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, ..., l. \end{aligned} \tag{5.17}$$

Le problème d'optimisation 5.17 n'est pas, en général, résoluble, et on cherche la solution du problème dual (voir Abe, 2005). Celui-ci est un problème équivalent au problème de départ (les équations 5.17), mais avec des équations qui sont exprimées à l'aide les multiplicateurs de Lagrange. Donc, l'équation qui est réellement minimisée est :

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \\ \text{sujet à} \quad & \mathbf{y}^T\boldsymbol{\alpha} = 0, \\ & 0 \leq \alpha_i \leq C, i = 1, ..., l. \end{aligned} \tag{5.18}$$

où α_i sont des multiplicateurs de Lagrange non-négatifs, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_l)^T$, $\mathbf{e}$ est un vecteur composé par des valeurs unitaires, Q est une matrice formée par des éléments $Q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$ et $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$, K est le noyau.

Finalement, la fonction de décision pour une donnée $\mathbf{x}$ est donnée par :

$$\mathbf{x} \in \begin{cases} \text{Classe 1} & \text{si} \quad D(\mathbf{x}) > 0, \\ \text{Classe 2} & \text{si} \quad D(\mathbf{x}) < 0. \end{cases} \tag{5.19}$$

où

$$D(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{5.20}$$

Séparant diverses classes

Après avoir résolu le problème pour la séparation de deux classes, il existe diverses manières d'appliquer la même méthode pour séparer un nombre arbitraire d'autres classes. Le code de *Support Vector Machines* que nous avons mis en œuvre dans le système de classification morphologique du DU470 utilise la méthode connue sous le nom de « un-contre-un »(*one-against-one*, ou *pairwise*).

Dans ce type de stratégie, pour un problème de classification avec k classes différentes, $k(k-1)/2$ SVMs sont construites, formées pour classer des échantillons d'une classe contre une autre classe, deux à deux. Un exemple pour trois classes peut être vu sur la Figure 5.24.

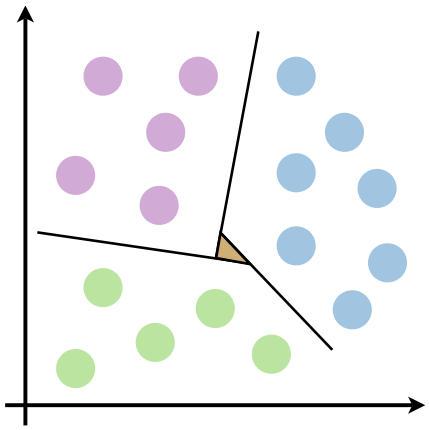


FIGURE 5.24 – Classification en trois classes montrant la région non classifiable au centre en marron.

Pour classer un certain point de données, une stratégie de « votation » est appliquée dans laquelle toutes les fonctions de décision sont calculées sur ce point, et la classe avec la plus grande quantité de votes est considérée comme étant la classe du point. Il faut dire que, dans ce type de formulation, des régions inclassables de l'espace peuvent exister, dans lesquelles il peut y avoir des égalités – et dans ce cas, la classe choisie pour un point dans cette région est prise arbitrairement comme étant celle de plus petit indice.

5.5 Mise en œuvre pour Gaia

Pour le développement des codes du DU470, nous avons utilisé une bibliothèque contenant des mises en oeuvre en Java de divers algorithmes de *Support Vector Machines* dénommée LIBSVM (Chang & Lin, 2001). La SVM choisie pour l'application dans ce problème est une C-SVM avec un noyau du type *Radial Basis Function*.

Cependant, la mise en œuvre du système de classification morphologique du DU470 ne sera pas exclusivement basée sur la classification de la *Support Vector Machine*, mais au contraire à partir d'une boucle dans laquelle la SVM est l'un des composants. Une représentation de ce processus peut être vue sur la Figure 5.25.

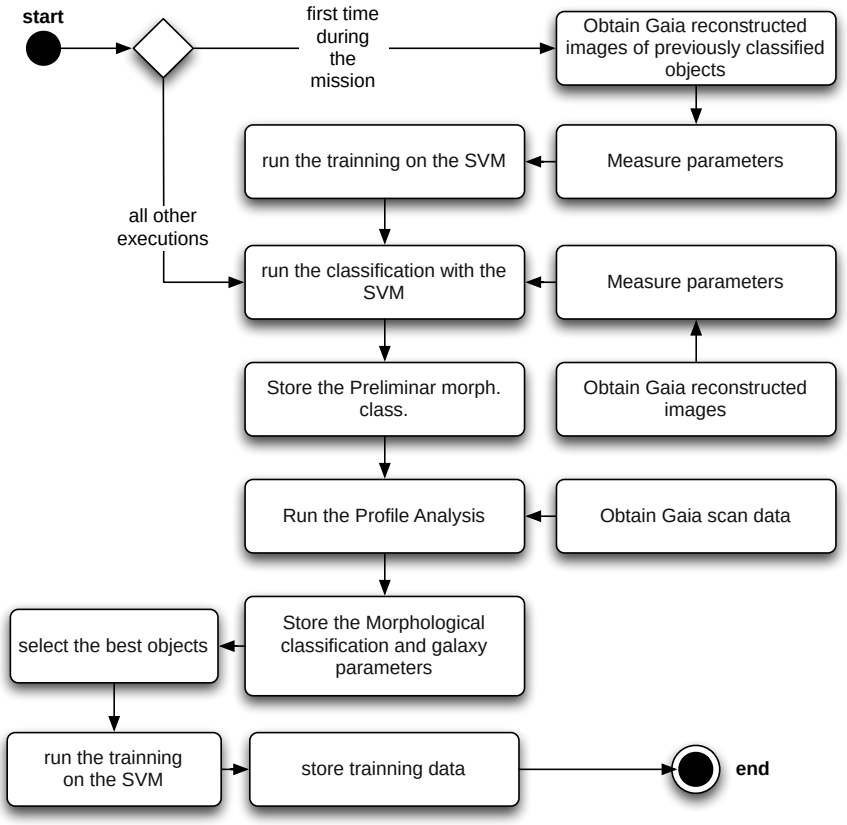


FIGURE 5.25 – Diagramme d'activité conceptuelle des processus impliqués dans le formation de la SVM.

Avant d'envisager de classer un objet, il est nécessaire de former la SVM. Comment se fera exactement la sélection des objets utilisés pour la formation initiale est un point qui doit encore être définie. Une possibilité est d'utiliser des objets possédant une classification préalablement connue, une autre est d'ignorer la classification et de réaliser une analyse de profil avec divers profils théoriques en sélectionnant les objets pour lesquels des profils correspondants à certains types morphologiques s'ajustent le mieux.

Après qu'une liste de galaxies de type morphologique connu est disponible, la SVM est formée pour la première fois, et est alors appliquée sur les images reconstruites d'une série d'autres galaxies de type inconnu, fournissant une classification morphologique pour ces objets. Après la classification morphologique, une analyse de profil est exécutée (qui sera vue de façon plus détaillée dans le Chapitre 6).

Lors cette analyse, un profil théorique est sélectionné à partir de la classification obtenue par la SVM et ajusté aux données du satellite. Dans le cas où l'ajustement est adéquat, la classification morphologique est confirmée, dans le cas contraire d'autres profils peuvent être testés, et la classification morphologique peut être altérée. Finalement, après avoir analysé toutes les galaxies de cette liste secondaire, les objets qui ont eu les meilleurs ajustements de profils sont sélectionnés, et sont utilisés pour former la SVM pour les cycles ultérieurs du traitement de la mission.

Pour les objets qui ne participeront pas à l'entraînement de la SVM, ou dans les cycles durant lesquels la formation de la SVM n'est pas exécutée, ce qui se passe est un traitement basé sur un enchaînement simple d'analyses. Premièrement une reconstruction d'images a lieu, ensuite les paramètres CASGM20 sont mesurés, la classification au moyen de la SVM est réalisée et finalement l'analyse de profil est exécutée. Une représentation de ce processus peut être rencontrée sur la Figure 5.26.

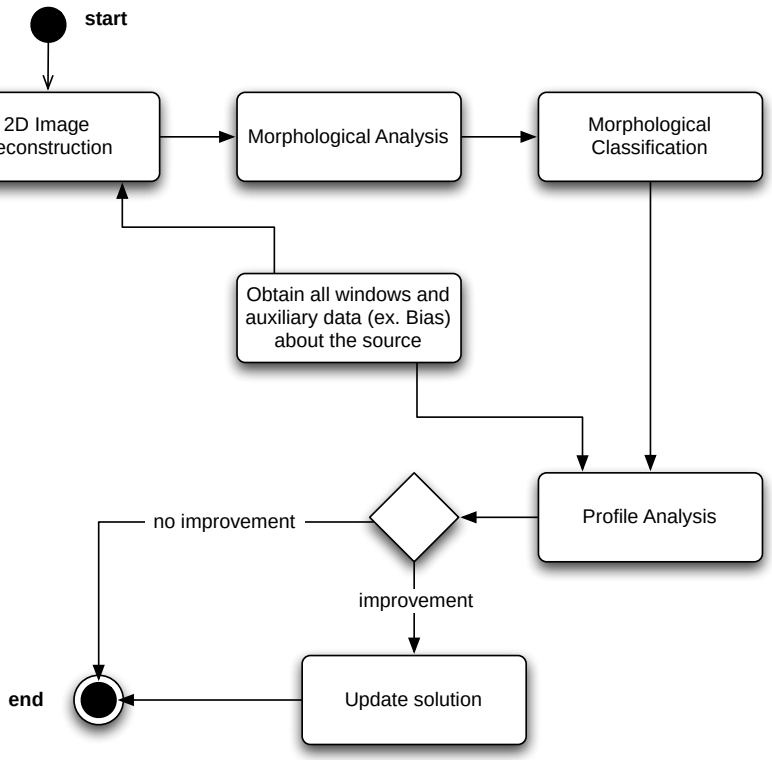


FIGURE 5.26 – Diagramme d'activité conceptuelle des processus impliqués dans la réduction des données du DU470.

5.5.1 Validation

Un premier test qui doit être réalisé est de vérifier la capacité que ce code a de classer correctement les points échantillonnés associés à des distributions de probabilité bien connues, pour lesquels les taux maximums de succès peuvent être estimés de façon complètement indépendante de la SVM. Une PDF (*Probability Distribution Function*) bien connue qui peut être appliquée dans ce type d'étude est la fonction gaussienne.

Soit un certain nombre de fonctions de densité de probabilités gaussiennes à deux variables ; nous désirons obtenir la meilleure séparation possible entre ces populations que tout classificateur pourrait obtenir avec des données aléatoires générées à partir de ces deux distributions. Après avoir obtenu cette évaluation, nous comparons le taux de succès dans la classification obtenue par une SVM et le calcul théorique pour le meilleur taux possible.

Le cas de deux gaussiennes

La meilleure classification possible est celle qui déterminera la fonction de décision dans la région de l'espace où les deux gaussiennes bidimensionnelles s'interceptent. Ainsi, le meilleur taux de classification possible est obtenu facilement (et en effet, nous verrons que ce problème peut être linéairement séparable), vu qu'il est suffisant d'intégrer les deux PDFs dans la région définie jusqu'à la fonction de décision. Ensuite, il suffit d'additionner ces deux valeurs, et diviser par la somme des intégrales des PDFs dans l'espace entier.

Malheureusement ce problème n'a pas de solution analytique, étant donné que les primitives de la gaussienne ne peut pas être obtenues sous forme explicite. Cependant, nous pouvons toujours obtenir une solution numérique. Dans notre cas, nous allons analyser la solution pour trois séparations distinctes entre les moyennes des deux gaussiennes: 0.5, 1.5 et 3.0 écarts type.

Pour simplifier encore l'analyse, nous prenons pour l'une des gaussiennes une moyenne $(\mu_x, \mu_y) = (0, 0)$ et pour l'autre une moyenne $\mu_y = 0$. Nous pouvons faire ceci sans perte de généralité, car pour deux gaussiennes il est facile de montrer que l'on peut toujours trouver une rotation permettant la réduction à ce cas pour toutes configurations des PDFs. Nous avons aussi considéré les mêmes écarts types pour les deux populations.

Mathématiquement, le problème prend la forme suivante: soit deux fonctions gaussiennes A et B ; il est faut trouver la solution de l'équation suivante (qui nous fournit la fonction de classification):

$$\frac{1}{2\pi\sigma_x^A\sigma_y^A} e^{-\left[\frac{(x-\mu_x^A)^2}{2(\sigma_x^A)^2}+\frac{(x-\mu_y^A)^2}{2(\sigma_y^A)^2}\right]} = \frac{1}{2\pi\sigma_x^B\sigma_y^B} e^{-\left[\frac{(x-\mu_x^B)^2}{2(\sigma_x^B)^2}+\frac{(x-\mu_y^B)^2}{2(\sigma_y^B)^2}\right]} \tag{5.21}$$

Comme dans le cas considéré, nous pouvons écrire:

$$\begin{cases} \sigma_x^A = \sigma_y^A = \sigma_x^B = \sigma_y^B = \sigma \\ \mu_x^A = \mu_y^A = \mu_y^B = 0 \end{cases}$$

La solution pour la fonction qui définit les limites de classification prend la forme:

$$\begin{cases} \forall \mu_x^B \in \mathbb{R}: & x = \frac{\mu_x^B}{2} \\ \forall y' \in \mathbb{R}: & y = y' \end{cases} \tag{5.22}$$

La courbe cherchée est une droite simple, et dans le cas ici traité, une droite verticale qui divise le plan x-y en deux régions facilement intégrables. Il suffit donc d'intégrer les PDFs sur les intervalles suivants:

$$\begin{cases} x \in]-\infty; \mu_x^B/2[\\ y \in]-\infty; +\infty[\end{cases}$$

Finalement, il suffit de résoudre l'équation suivante pour obtenir le taux maximum de succès possible pour la classification:

$$\Upsilon = \frac{\int_{-\infty}^{+\infty}\int_{-\infty}^{\mu_x^B/2} f_A(x,y)\,\mathrm{d}x\mathrm{d}y + \int_{-\infty}^{+\infty}\int_{\mu_x^B/2}^{+\infty} f_B(x,y)\,\mathrm{d}x\mathrm{d}y}{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_A(x,y)\,\mathrm{d}x\mathrm{d}y + \int_{-\infty}^{+\infty}\int_{-\infty}^{+-\infty} f_B(x,y)\,\mathrm{d}x\mathrm{d}y} \tag{5.23}$$

où

$$f_\xi(x,y) = f_\xi(\sigma_x^\xi, \sigma_y^\xi, \mu_x^\xi, \mu_y^\xi, x, y) = \frac{1}{2\pi\sigma_x^\xi\sigma_y^\xi} e^{-\left[\frac{(x-\mu_x^\xi)^2}{2(\sigma_x^\xi)^2} + \frac{(x-\mu_y^\xi)^2}{2(\sigma_y^\xi)^2}\right]}$$

La solution peut donc être obtenue numériquement (nous avons utilisé le système d'intégration mis en œuvre dans le logiciel Maple 11), étant entendu que pour les séparations étudiées nous avons obtenu les résultats présentés sur le Tableau 5.4.

Séparation	Taux maximale de succès
0.5σ	59.87%
1.5σ	77.34%
3σ	93.32%

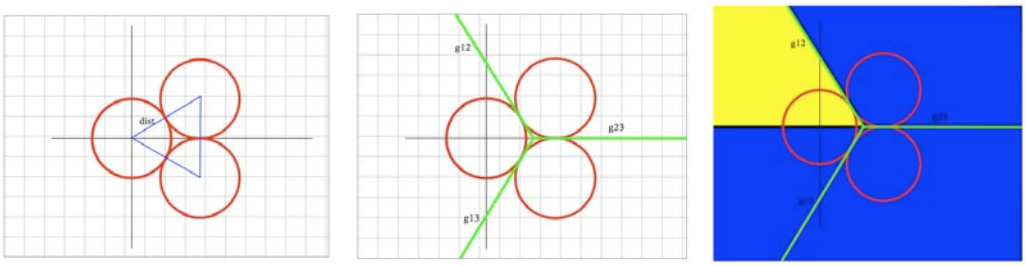
Tabela 5.4: Taux maximum théorique pour la classification dans la discrimination de deux populations gaussiennes.

Le cas de trois gaussiennes

Une analyse similaire permet l'obtention du taux maximum de succès théorique pour trois classes générées à partir de trois PDFs gaussiennes. Comme dans le cas de deux populations, la meilleure classification possible est celle qui utilise la fonction de décision dans laquelle deux ou plus gaussiennes se coupent. Comme dans le cas antérieur, le problème se réduit à intégrer les gaussiennes jusqu'à la fonction optimum de classification, à additionner ces valeurs et à diviser par l'intégrale dans tout l'espace.

Ici nous considérons le cas restreint d'une gaussienne de moyenne $(\mu_x, \mu_y) = (0, 0)$, les deux autres étant situées sur les sommets d'un triangle équilatéral (représenté en lignes bleus sur la figure 5.27a). Nous considérons aussi que les écarts types de

toutes les gaussiennes sont égaux. De plus, comme dans le cas de deux gaussiennes, le paramètre du problème est le côté du triangle, c'est-à-dire, la distance entre les centres des gaussiennes.



(a) (b) (c)

Figura 5.27: Diagrammes descriptifs de: (a) Indication des populations et de leurs moyennes; (b) les meilleures fonctions de discrimination possibles; (c) intervalles d'intégration.

Dans ce cas, la meilleure séparation possible est donnée par les fonctions de décision gij(x), représentées en vert sur la Figure 5.27b, où ij sont les index des gaussiennes que la fonction de décision est capable de discriminer. Dans le problème considéré ici, il n'y a pas de régions non classifiables, étant donné que les fonctions s'interceptent en un point.

Fondamentalement, pour résoudre le problème, nous avons besoin d'intégrer les gaussiennes dans les régions colorées sur la Figure 5.27c. Pour des raisons de symétrie, il suffit d'intégrer la première gaussienne dans la région jaune de la Figure 5.27c et de multiplier cette valeur par 6.

Ainsi, il nous faut trouver la fonction de décision gAB (g12 sur la Figure 5.27), pour pouvoir définir l'intervalle d'intégration. Ici, c'est une tâche simple, étant donné que comme dans le cas de deux populations, il suffit de trouver la solution de l'équation 5.21, mais restreinte aux conditions suivantes (dist est la distance entre les centres des gaussiennes, donnée en unités d'écart type):

$$\begin{cases} \sigma_x^A = \sigma_y^A = \sigma_x^B = \sigma_y^B = \sigma \\ \mu_x^A = \mu_y^A = 0 \\ \mu_x^B = \cos\left(\frac{\pi}{6}\right)\sigma \times \text{dist} \\ \mu_y^B = \sin\left(\frac{\pi}{6}\right)\sigma \times \text{dist} \end{cases}$$

De nouveau, la solution a une forme simple:

$$\begin{cases} \forall x' \in \mathbb{R}: & x = x' \\ \forall \sigma, \text{dist} \in \mathbb{R}: & y = -\cot\left(\frac{\pi}{6}\right)x + \sigma \times \text{dist} \end{cases} \tag{5.24}$$

Il nous faut donc intégrer la PDF jusqu'à la courbe ci-dessus, c'est à dire utiliser le domaine D d'intégration suivant:

$$\begin{cases} x \in]-\infty; (\sigma \times \text{dist})/\cot\left(\frac{\pi}{6}\right)] \\ y \in \ [0; -\cot\left(\frac{\pi}{6}\right) x + \sigma \times \text{dist}] \end{cases}$$

Ainsi, il suffit de résoudre:

$$\Upsilon = \frac{6 \int\int f_A(x,y)\, \mathrm{d}y\mathrm{d}x, \text{ sur le domaine D}}{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_A(x,y)\, \mathrm{d}x\mathrm{d}y + \int_{-\infty}^{+\infty}\int_{-\infty}^{+-\infty} f_B(x,y)\, \mathrm{d}x\mathrm{d}y + \int_{-\infty}^{+\infty}\int_{-\infty}^{+-\infty} f_C(x,y)\, \mathrm{d}x\mathrm{d}y}$$

où les fonctions f(x,y) sont définies comme dans le cas de deux populations. Le dénominateur prend la valeur 3, et l'équation se réduit à:

$$\Upsilon = 2 \int\int f_A(x,y)\, \mathrm{d}y\mathrm{d}x, \text{ sur le domaine D} \tag{5.25}$$

Nous obtenons donc, la solution numériquement (en utilisant de nouveau le logiciel Maple 11), pour les cas ici étudiés. Les résultats sont donnés sur le Tableau 5.5.

Séparation	Taux maximale de succès
0.5σ	43.77%
1.5σ	65.11%
3σ	88.47%

Tabela 5.5: Taux maximum théorique pour la classification dans la discrimination à trois populations gaussiennes.

Résultats de la SVM et comparaison

De manière à analyser le fonctionnement et l'efficacité de la séparation de ces populations gaussiennes en utilisant des SVMs, nous avons simulé des observations de populations aléatoires. L'étude a été réalisée en utilisant deux et trois classes différentes, et dans les deux cas nous avons analysé les scénarios où les moyennes des populations étaient séparées par 0.5, 1.5 et 3 écarts types, tel que nous l'avons fait dans l'étude présentée.

Dans chaque simulation, des distributions aléatoires de mille points ont été générées. Cette population a été divisée en deux fichiers, chacun avec 1000 points, contenant la même quantité de représentants de chaque population, choisis de façon aléatoire. Un des fichiers a été utilisé pour la formation et la validation croisée, et l'autre pour la réalisation d'un test de classification. Nous avons donc réalisé 10 formations distincts de la SVM (au moyen de prélèvements aléatoires différents des points utilisés pour la formation et le test), obtenant finalement les résultats pour la précision de la classification.

Les résultats obtenus pour les dix formations pour les trois cas de séparation entre les distributions sont indiqués sur le Tableau 5.6. En comparant les résultats obtenus par la SVM pour la classification avec les meilleurs résultats possibles analytiquement, on vérifie la haute efficacité de cette méthode pour la solution du problème de classification de populations : le taux de succès maximum théorique a été atteint pour toutes les séparations.

Un résultat similaire a été observé en utilisant trois classes de points différentes, de sorte que nous pouvons conclure que, pour le cas étudié ici, l'efficacité de la méthode est relativement élevée, sachant que c'est la nature du problème et non la méthode elle-même le principal facteur d'erreur de classification.

Populations	Séparation	Taux maximum théorique	Taux de succès de la SVM
2	0.5σ	59.87%	$(59.38 \pm 1.29)\%$
2	1.5σ	77.34%	$(77.50 \pm 1.10)\%$
2	3σ	93.32%	$(93.37 \pm 0.75)\%$
3	0.5σ	43.77%	$(43.61 \pm 1.19)\%$
3	1.5σ	65.11%	$(64.14 \pm 1.69)\%$
3	3σ	88.47%	$(87.19 \pm 0.98)\%$

Tabela 5.6: Taux maximum théorique et de succès de la SVM pour la classification dans la discrimination de deux et trois populations gaussiennes.

5.5.2 Tests avec des galaxies du catalogue de Frei

Après avoir vérifié l'efficacité de la méthode à classer correctement les données artificielles générées à partir de PDFs bien connues, il faut vérifier si elle est capable de classer correctement les galaxies qui seront observées par Gaia. Pour cela, un premier test bien contrôlé a été réalisé en utilisant des images de galaxies proches, dont les classifications étaient connues sans ambiguïté. De plus, il est nécessaire de vérifier le comportement de cette classification quand les quantités de pixels disponibles pour la classification décroissent.

Pour réaliser cet essai, nous avons utilisé des données des paramètres CASGM20 mesurées en images du catalogue de Frei et al (1996) (présentés dans la section 5.3.1). Ces mesures ont été introduites dans le code mise en place pour Gaia. Cependant, du fait du faible nombre de galaxies dans le catalogue est, et de leur utilisation aussi bien pour la formation du classificateur que pour le test de précision de la classification, nous avons adopté une procédure de re-échantillonnage, basée sur la création de sous-ensembles aléatoires de galaxies.

Cette procédure a été réalisée de la manière suivante : 1. l'ensemble total de mesures est divisé en deux sous ensembles, dans lesquels 65% des données sont utilisées pour former la SVM et le reste pour tester la précision de la classification 2. la SVM est formée et appliquée aux données, et le taux de succès dans la classification est calculé ; 3. de nouveaux ensembles sont créés et le processus est itéré.

Pour chaque taille d'image, 100 itérations ont été réalisées, et le résultat final de la précision globale obtenue est représenté sur la Figure 5.28. Sur cette figure, on peut noter la stabilité du résultat obtenu, à environ 85% de réussite pour toutes les tailles d'image.

Ces résultats démontrent que dans le cas d'une reconstruction d'image parfaite et de la détermination d'un ensemble d'entrainement sans aucun type d'erreur de classification, le meilleur taux de classification qui pourrait être obtenu, en principe, avec l'utilisation d'une SVM formée exclusivement en données de CASGM20 doit atteindre $85 \pm 3\%$.

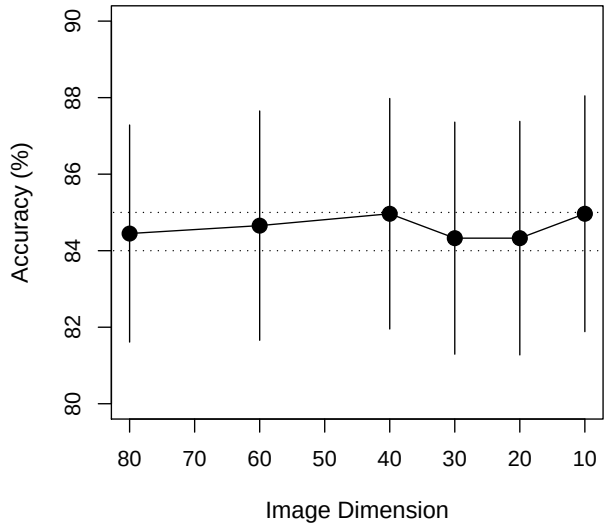


FIGURE 5.28 – Taux de précision globale obtenu pour la classification de galaxies du catalogue de Frei à partir de mesures CASGM20 et d'une *Support Vector Machine*. Les barres d'erreur ont été obtenues à partir de ré échantillonnages aléatoires des données.

5.5.3 Tests avec des simulations GIBIS

Un autre test qui a été réalisé avec ce code (toujours basé sur les paramètres CASGM20 et la SVM) a été son application aux images reconstruites de données simulées de la mission Gaia. Ce images sont à la fois fortement dégradées par la PSF mais les algorithmes de reconstruction introduisent des artefacts. Pour cela, les données observées de 2640 galaxies ont été simulées (GIBIS 7, voir section 5.3.3) sur les trois types morphologiques disponibles (elliptiques, spirales et irrégulières).

Nous avons testé plusieurs tailles pour l'ensemble de galaxies utilisé pour former la SVM (50, 100, 200, 400, 600 galaxies par type morphologique) et de forme similaire à la sous division antérieure ; une procédure de ré-échantillonnage a été utilisée pour permettre l'obtention des intervalles de confiance pour le taux de succès dans la classification – le nombre total d'itérations adopté a été 100. Dans ce cas, comme la formation de la SVM a impliqué de plus grandes capacités de traitement (pouvant

prendre des dizaines de minutes de temps de calcul pour chaque itération), divers processus Java locaux ont été utilisés en travaillant en parallèle pour la formation de la SVM.[17]

Le taux de succès obtenu dans la classification de chaque type distinct de galaxie a été calculé séparément, et la matrice de confusion entre les divers types de galaxie a aussi été déterminée. Les résultats pour une classification dans les trois types morphologiques sont montrés sur la Figure 5.29 et sur le Tableau 5.7.

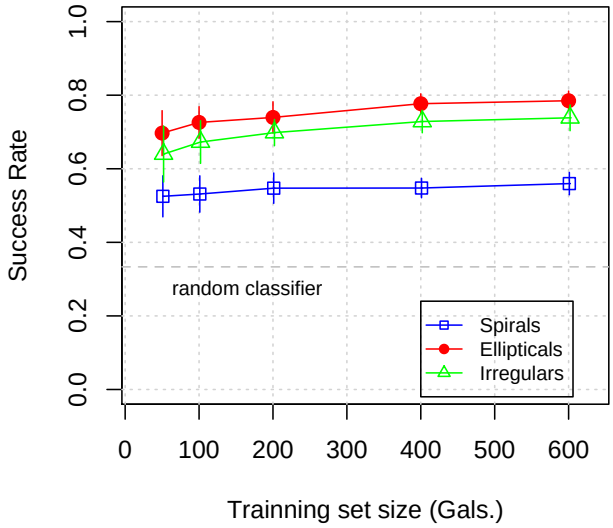


FIGURE 5.29 – Résultats du taux de succès obtenu dans la classification en trois classes de tailles variables pour l'ensemble de formation de la SVM. Sur l'axe x est représenté le numéro de galaxies par classe.

Morphologie réelle	Classée comme E	Classée comme S	Classée comme I
Elliptique (E)	0.79 ± 0.03	0.08 ± 0.02	0.14 ± 0.02
Spirale (S)	0.14 ± 0.02	0.56 ± 0.03	0.30 ± 0.03
Irrégulière (I)	0.11 ± 0.02	0.15 ± 0.03	0.74 ± 0.04

TABLE 5.7 – Matrice de confusion obtenue avec les paramètres CASGM20 et la SVM (formée sur 600 galaxies/type) pour classification en trois classes morphologiques.

Ces résultats démontrent que le gain dans la capacité de discernement entre les types n'est pas négligeable si l'on adopte des quantités supérieures de galaxies pour la formation de la SVM. De plus, ils démontrent aussi que la séparation entre les

17. Le nombre exact de processus a été de 12, et la formation a pris un total de ~ 10 de temps de traitement sur une machine avec 12 coeurs Xeon 2.66/3.05GHz, consommant un nombre total de ~ 1.5 TFLOP. Cette capacité doit être considérée très insuffisante s'il s'agit d'appliquer cette méthode sur les données réelles de la mission.

galaxies elliptiques et irrégulières a été relativement efficace, en restant au-dessus de 70% pour pratiquement toutes les tailles de l'ensemble de formation.

Néanmoins, les résultats pour la classification en trois classes montrent aussi qu'il existe une plus grande confusion dans la ségrégation entre galaxies spirales et irrégulières, avec environ 30% des spirales simulées étant classées comme irrégulières. Ceci démontre que, du point de vue des paramètres adoptés pour la classification, il existe une similitude relativement importante entre certains types de galaxies spirales et de galaxies irrégulières, qui doit être dû à des spirales du type tardif.

Une classification en deux classes a donc été réalisée pour analyser la différenciation entre les galaxies elliptiques (types précoces) et les galaxies spirales/irrégulières (types tardifs). Les résultats peuvent être vus sur la Figure 5.30 et sur le Tableau 5.8.

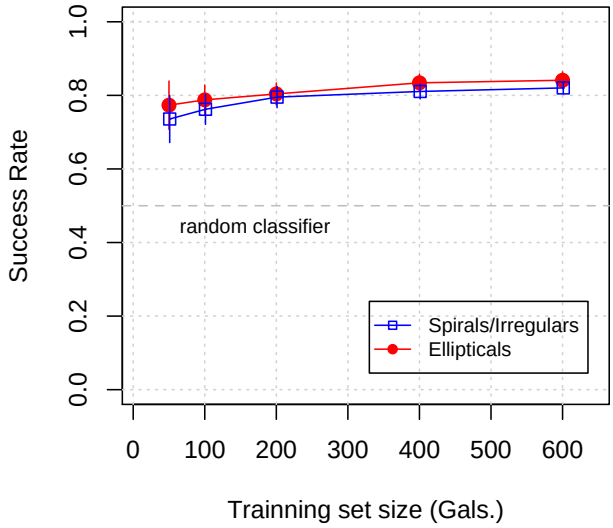


FIGURE 5.30 – Résultats du taux de succès obtenu dans la classification en deux classes avec des tailles variables pour l'ensemble de formation de la SVM. Sur l'axe x est représenté le nombre de galaxies par classe.

Morphologie réelle	Classée comme E	Classée comme S ou I
Elliptique (E)	0.84 ± 0.02	0.16 ± 0.02
Spirale (S) ou Irrégulière (I)	0.18 ± 0.02	0.82 ± 0.02

TABLE 5.8 – Matrice de confusion obtenue avec les paramètres CASGM20 et la SVM (formée sur 600 galaxies/type) pour une classification en deux classes morphologiques.

Les résultats pour la classification en deux classes démontrent que l'on peut obtenir un taux de succès de la différenciation entre ces deux types morphologiques stable, supérieur à 80% pour tous les ensembles de formation contenant plus de 200 galaxies par type morphologique. Cette capacité élevée de différenciation pourra

permettre que des profils bulbe+disque ou bulbe soient choisis correctement pour $\sim 80\%$ des galaxies dès la première tentative pour l'analyse de profil.

Cependant, la principale conclusion qui peut être obtenue à partir des analyses de cette section, est qu'à partir de galaxies simulées avec GIBIS 7 et passées par tout le processus de reconstruction, la mesure de paramètres CASGM20 et la classification avec les codes qui ont été mis en place pour Gaia, il sera possible des réaliser des études morphologiques sur des galaxies angulairement petites avec les données de cette mission spatiale. Et comme nous l'avons déjà dit antérieurement, la plus grande partie de ces galaxies seront observées pour la première fois.

5.6 Application aux données du *Hubble Deep Field*

Les tests réalisés avec des données simulées avec GIBIS 7 dépendent fortement de l'algorithme de reconstruction d'image qui est adopté, et nous savons que les algorithmes disponibles actuellement ne sont pas optimums. Nous savons aussi que des reconstructions plus détaillées peuvent être théoriquement réalisées, comme expliqué dans le Chapitre 2, et bien que nous ne puissions pas réaliser de tests avec des algorithmes qui n'existent pas encore, nous pouvons réaliser des tests en utilisant des images obtenues dans des régimes similaires à ceux des images de Gaia dans le cas d'une reconstruction parfaite.

Le *Hubble Deep Field* (HDF) peut être considéré comme un bon cas de test pour le système de classification morphologique qui sera appliqué à Gaia dans le cas d'une reconstruction d'image parfaite, mais d'une classification imparfaite. De cette manière, les résultats obtenus dans cette section pourront servir à déterminer une limite supérieure de ce qui peut être réalisé avec les méthodes décrites dans ce chapitre pour une classification purement morphologique par Gaia au cas où un algorithme de reconstruction d'images parfaite serait disponible.

Néanmoins, comme il est adéquat d'utiliser la plus grande quantité possible de galaxies correctement classées, et comme il existait déjà des suspicions sur la possibilité d'une contamination due à des identifications croisées inexactes entre les classifications visuelles de van den Bergh et al (1996) et notre extraction de timbres décrite dans la section 5.3.2, nous avons décidé de réaliser une nouvelle classification visuelle des objets directement à partir des timbres recoupés pour les objets individuels – les galaxies sélectionnées sont les 400 plus brillantes (à partir de la somme simple des flux de chaque pixel).

Cette nouvelle classification visuelle a été réalisée par deux collègues spécialistes (Dr. Sandra dos Anjos et Dr. Rubens Machado, IAG) de façon complètement indépendante. Les résultats des classifications obtenus individuellement ont été comparés et seules les galaxies dont les classifications visuelles étaient similaires (281 objets) ont été utilisées pour la construction de l'ensemble d'entraînement de la SVM et pour le test de la précision dans la classification.

La procédure d'analyse adoptée a été la même que celle décrite dans la section antérieure : les paramètres CASGM20 des images ont été mesurés et 120

ré-échantillonnages aléatoires ont été utilisés pour définir des ensembles de entraînement et de tests pour la SVM. Ces ensembles ont été utilisés pour la détermination de la précision de la classification des trois grands types morphologiques, étant entendu qu'une classification similaire en types précoce et tardif a aussi été réalisée. Les résultats obtenus peuvent être vus sur la Figure 5.31 et sur les Tableaux 5.9 et 5.10.

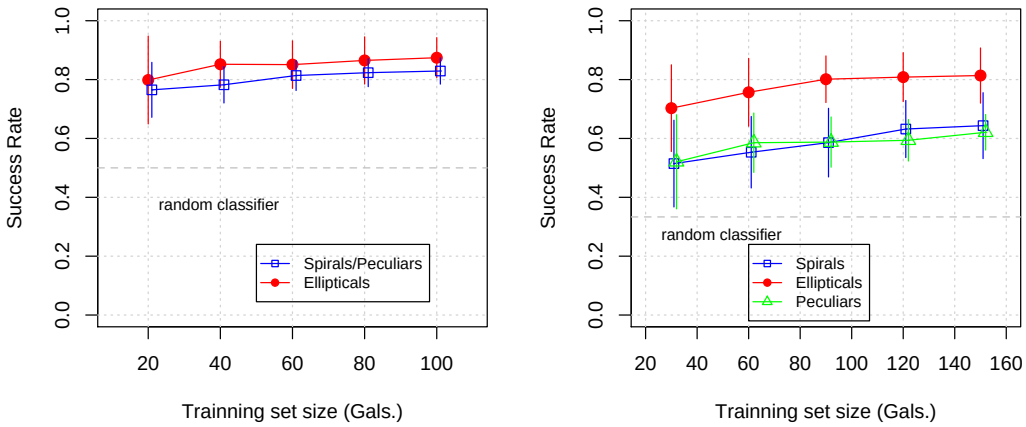


FIGURE 5.31 – Résultats du taux de succès obtenu dans la classification en deux et trois types avec des tailles variables pour l'ensemble de formation de la SVM des données du *Hubble Deep Field North*.

Morphologie réelle	Classée comme E	Classée comme S ou I
Elliptique (E)	0.87 ± 0.07	0.13 ± 0.07
Spirale (S) ou Irrégulière (I)	0.17 ± 0.05	0.83 ± 0.05

TABLE 5.9 – Matrice de confusion pour la classification des données du HDF-N en deux classes morphologiques (SVM treinada em 50 galáxias/tipo).

Morphologie réelle	Classée comme E	Classée comme S	Classée comme I
Elliptique (E)	0.81 ± 0.09	0.27 ± 0.07	0.07 ± 0.06
Spirale (S)	0.17 ± 0.08	0.64 ± 0.11	0.19 ± 0.11
Irrégulière (I)	0.09 ± 0.04	0.29 ± 0.06	0.62 ± 0.06

TABLE 5.10 – Matrice de confusion obtenue pour la classification des données du HDF-N en trois classes morphologiques (SVM treinada em 50 galáxias/tipo).

Les résultats obtenus pour la classification des images du HDF montrent qu'il est possible de distinguer des galaxies du type précoce de celles du type tardif avec une précision de $\sim 85\%$, qu'est le taux maximum de succès possible estimé a partir des galaxies proches. Ils montrent aussi qu'il est possible distinguer des galaxies

elliptiques, spirales et particulières (qui incluent des galaxies irrégulières ou des galaxies en interaction), peuvent être classées avec des taux de succès de $\sim 81\%$, $\sim 64\%$ e $\sim 62\%$, respectivement.

Comparant ces résultats avec ceux obtenus dans la section antérieure pour des simulations, nous avons observé qui même si les reconstructions d'image avec le *ShuffleStack* n'ont pas abouti dans des résultats trop mauvaises, c'est le processus de reconstruction que limite la qualité des classifications purement morphologiques. Ceci, car la classification en types tardif et précoce donne des résultats au-dessus $\sim 80\%$ pour une reconstruction d'image imparfaite (*ShuffleStack*) mais un echantillion d'entraînement parfait, tandis que pour des images du HDF, qui représenteraient des reconstructions parfaites, mais avec des types morphologiques mal déterminés, elle présente des taux de succès de $\sim 85\%$.

Les résultats obtenus ici à partir des données réelles du HDF montrent que face à des données similaires à celles qui seront rencontrées par Gaia, les méthodes que nous avons développées donnent de bons résultats pour la classification en trois classes (les types de Hubble), et donnent des résultats pour une classification en deux types (précoce/tardif) proche de celle obtenue par des galaxies proches.

5.7 Conclusions

Dans ce Chapitre, nous avons tout d'abord présenté des paramètres existant dans la littérature, qui définissent un volume dénommé CASGM20, dans lequel des morphologies distinctes de galaxies occupent des zones caractéristiques ; nous avons présenté la méthode que nous avons développée pour le calcul de ces paramètres sur des images reconstruites à partir de données du satellite Gaia. Nous avons montré que cette méthode, quand elle est utilisée pour calculer les paramètres d'images avec des échantillonnages divers, se comporte de façon plus ou moins stable même pour des images formées avec seulement 21x21 pixels.

Nous avons appliqué les codes construits pour l'analyse et la classification à des images extraites du *Hubble Deep Field North*, et nous avons vérifié que des types morphologiques distincts occupent bien des zones distinctes dans des plans de cet espace. Cette conclusion a également été obtenue lors de l'analyse d'images reconstruites issues de simulations GIBIS de 2640 galaxies de trois types morphologiques.

Nous avons montré, en utilisant ces ensembles de données CASGM20 calculées, qu'une PCA n'est pas capable de réduire significativement la dimension du problème rencontré. Nous en concluons que des techniques d'apprentissage computationnel plus avancées doivent être appliquées si l'on veut obtenir des résultats précis de façon automatique.

Nous avons alors présenté une méthode moderne de classification et d'apprentissage computationnel peu utilisée en astronomie jusqu'à aujourd'hui, appelée *Support Vector Machine*, qui trouve une application naturelle dans l'espace des paramètres CASGM20. Elle permet la prise en compte simultanée de tous les paramètres, et n'est pas sensible aux problèmes rencontrés par d'autres méthodes plus communes, comme par exemple des réseaux neuronaux.

Nous avons démontré que les codes que nous avons mis en place pour Gaia (qui utilisent une bibliothèque de SVMs disponible dans la littérature) sont capables de différencier des populations aléatoires gaussiennes avec succès, essentiellement en atteignant les taux maxima de précision qui peuvent être théoriquement obtenus.

Nous avons alors appliqué cette méthode sur des mesures d'images de galaxies proches, démontrant que le taux de précision global obtenu pour la classification en types précoce/tardif pour tout échantillonnage des images des galaxies (entre 81x81 et 11x11 pixels) reste stable entre 84% et 85%, de sorte que dans un scénario dans lequel une reconstruction d'image parfaite est disponible pour les données de Gaia et où l'ensemble d'entraînement est construit sans aucun type d'erreur de classification, le taux de réussite dans la classification obtenue par ce code est de $85\% \pm 3\%$.

Nous avons testé ce code sur des simulations réalisées avec GIBIS 7 et des reconstructions (loin d'être idéales) avec *ShuffleStack*. Ce test a montré que ce code peut être utilisé avec succès sur les données de Gaia, en présentant un taux de succès de différenciation entre les types précoce/tardifs supérieur à 80%, et dans les trois classes morphologiques principales, elliptiques, spirales et irrégulières, de 79%, 56% et 74%, respectivement, à condition que l'ensemble de formation de la SVM soit parfaitement construit.

Finalement, nous avons utilisé des images des galaxies du *Hubble Deep Field North* pour analyser le comportement de la méthode dans un régime dans lequel la reconstruction d'image est parfaite, mais dans lequel l'ensemble d'entraînement de la SVM souffre d'erreurs de classification. Cette dernière étude a démontré que même dans ce cas, on obtient des taux de classification entre les types précoce/tardifs supérieurs à $\sim 85\%$, et dans les trois classes morphologiques principales, elliptiques, spirales et irrégulières, de $\sim 81\%$, $\sim 64\%$ e $\sim 62\%$, respectivement.

Les résultats de ce Chapitre démontrent donc qu'avec l'utilisation des codes développés ici, des études morphologiques pourront être réalisées de manière satisfaisante avec les données provenant de Gaia, même si les ensembles d'entraînement contiennent des erreurs, et que les reconstructions d'images ne sont pas parfaites.

Chapitre 6

Estimation des paramètres morphologiques

"*Um retrato é apenas a idéia aproximada de uma pessoa.*", Florbela Espanca[1]

Sommaire

En plus du classement morphologique, l'autre objectif de l'analyse de la forme d'objets étendus est la détermination de leurs paramètres morphologiques, c'est-à-dire, les caractéristiques de leurs différents composants structurels. Ceci est fait au moyen d'ajustements de profils de brillance aux données d'observation, engendrant une décomposition des objets en un ou plusieurs composants. Néanmoins, la qualité des images reconstruites au moyen des algorithmes actuellement disponibles n'est pas suffisante pour permettre qu'une telle étude soit réalisée de manière fiable avec les données de la mission Gaia.

Dans ce Chapitre, nous parlons des profils de brillance les plus utilisés pour l'étude de galaxies et sur comment ces profils peuvent être ajustés aux données Gaia sans qu'il soit nécessaire de réaliser une reconstruction d'image bi-dimensionnelle. Nous montrerons la méthode que nous avons développée pour réaliser cet ajustement, qui est basé sur l'interprétation des observations Gaia comme une transformée de Radon de l'objet et sur la réalisation de simulations (Krone-Martins, 2010). Nous présenterons les résultats obtenus avec l'application de cette méthode sur un ensemble statistiquement significatif d'observations simulées. Finalement, nous commenterons les possibilités de perfectionnement et les applications de la méthode développée sur des images provenant de n'importe quel télescope.

1. « Un portrait n'est seulement que l'idée approchée d'une personne. »

6.1 Introduction

La première utilisation d'ajustement de profils analytiques pour étudier la distribution de brillance de galaxies a été faite par de Vaucouleurs (1948), qui montra que la brillance de galaxies elliptiques a tendance à suivre une loi du type $\exp(-r^{1/4})$. Pour cela, de Vaucouleurs (1959) (d'après Freeman, 1970), proposa que le profil de brillance de galaxies spirales pouvait être décomposé dans une fonction créée à partir de l'addition de deux composants, le bulbe et le disque.

Ces profils de brillance sont d'importants outils pour l'étude d'objets étendus (tels que les galaxies), car en plus de révéler les structures morphologiques de ces objets, ils sont des traceurs des processus physiques responsables de la formation et la stabilité (Freeman, 1970) des composants structurels de ces objets. Des bulbes classiques, des pseudo-bulbes et des barres, sont quelques exemples de composants centraux de galaxies du type spirales et S0, qui probablement possèdent des mécanismes de formation distincts (Kormendy & Kennicutt, 2004). Il existe toute une myriade d'études astronomiques qui sont rendues possibles grâce à ces décompositions, depuis des analyses individuelles de galaxies (ex. Gadotti, 2008), jusqu'à des études de populations de ces objets suivant l'âge de l'univers (ex. Marleau & Simard, 1998; Hathi et al, 2009).

Afin que l'étude de ces différents composants structurels des galaxies soit possible, il est nécessaire que les données soient décomposées dans les contributions de ces composants. C'est un problème habituellement résolu au moyen d'ajustements des fonctions aux images par des méthodes de minimisation de moindres carrés.

Initialement, les ajustements de ces fonctions analytiques étaient réalisés en utilisant des données obtenues par l'extraction du profil de brillance dans la direction du plus grand axe de la galaxie. Bien entendu les profils n'étaient ajustés séparément que dans les régions dans lesquelles ils étaient dominants, et ensuite les paramètres des deux composants étaient déterminés au moyen d'un processus itératif dans lequel soit le bulbe, soit le disque, étaient ajustés (Kormendy, 1977; Burstein, 1979). Une évolution, adoptée par Kent (1985), fut l'utilisation, d'ajustements unidimensionnels et simultanée pour les deux composants (bulbe et disque), en plus de considérer le profil de brillance mesuré sur les grands et petits axes de 105 galaxies.

Avec la mise à disposition d'une plus grande puissance de calcul, principalement durant cette dernière décennie, il est devenu possible de réaliser cet ajustement directement dans les données bi-dimensionnelles des images au moyen de codes tels que le GIM2D (Simard, 1998), le BUDDA (de Souza et al, 2004) et le GALFIT (Peng et al, 2002, 2010).

Dans ce Chapitre, nous avons introduit une troisième manière de réaliser cette décomposition, qui, en plus de permettre l'ajustement de profils de brillance et des décomposition de galaxies observées par Gaia en l'absence de reconstructions d'images bi-dimensionnelles, pourra aussi avoir une application plus ample, permettant la réalisation de ce type d'analyse sur toutes les galaxies dont les images ont été obtenues avec des bas niveaux de signal/bruit. Il faut remarquer que la méthode développée ici est générale, pouvant être appliquée sur des profils de brillance d'autres objets.

6.1.1 Profils de brillance

Pour réaliser la décomposition morphologique du profil de brillance de galaxies, il est nécessaire de connaître à priori la quantité de composants dans laquelle on désire décomposer l'objet.[2] Néanmoins du point de vue physique, il est intéressant de toujours adopter des critères parcimonieux dans ce choix, car numériquement il est possible d'obtenir de meilleurs ajustements si l'on considère des quantités chaque fois plus grandes de composants, qui ne sont pas forcément physiquement justifiés.

En général, les modèles de galaxies utilisent deux, trois ou quatre composants : une source nucléaire ponctuelle, un bulbe, une barre et un disque – habituellement les bras des galaxies spirales ne sont pas additionnés, bien que récemment Peng et al (2010) ait ajouté de telles structures au code GALFIT. La barre et la source centrale sont en général utilisées pour des galaxies particulières, et pour des analyses sur de petites échelles (comme dans Gadotti, 2008) et la plus grande partie des décompositions est réalisée ne considérant à peine que deux composants : le disque et le bulbe. cela, car ces derniers sont responsables pour la plus grande fraction de la luminosité d'une grande partie des galaxies. De plus, dans beaucoup de cas la résolution des données est inadéquate pour que d'autres composants soient résolus.

Les profils de brillance les plus fréquemment utilisés pour décrire les disques sont en général des profils exponentiels, qui peuvent être écrits de la manière suivante :

$$I_d(r) = I_{0d} \exp\left(-\frac{r}{r_d}\right) \tag{6.1}$$

où I_{0d} est l'intensité au centre du disque, r_d est l'échelle de longueur du disque et r est la distance à laquelle l'intensité est calculée.

Pour le bulbe, l'un des profils les plus utilisés est celui proposé par de Vaucouleurs (1948), commenté dans l'introduction de ce Chapitre. Néanmoins ces dernières années on a fréquemment adopté des profils suivant le paramétrage de Sérsic (Blanton & Moustakas, 2009). Ce profil, initialement décrit dans Sérsic (1963); Sérsic (1968) (d'après Graham & Driver, 2005) peut être écrit de la manière suivante :

$$I_b(r) = I_{0b} \exp\left(-b_n \left[\left(\frac{r}{r_b}\right)^{\frac{1}{n}} - 1\right]\right) \tag{6.2}$$

où r_b est le rayon effectif qui contient la moitié de la luminosité intégrée de la galaxie, I_{0b} est l'intensité du bulbe dans le rayon effectif, n est l'indice de Sérsic, b_n est une constante et r est la distance à laquelle l'intensité est calculée.

La valeur de la constante b_n dépend de la valeur choisie pour l'indice de Sérsic, pouvant être calculée numériquement à partir de l'intégrale du profil. Néanmoins, en général on adopte des approximations analytiques de ce paramètre, comme par exemple l'approximation $b = 1.9992n - 0.3271$ Capaccioli (1989) (apud Graham & Driver, 2005) qui est valable pour un large intervalle d'indices de Sérsic, $0.5 < n < 10$.

2. Pour cela, une classification morphologique basée sur des paramètres indépendants d'ajustements de profil (que nous avons vu dans le Chapitre 5 de cette thèse) est réalisée préliminairement dans le *pipeline* de Gaia.

Du point de vue physique, l'indice de Sérsic définit quel type de bulbe une galaxie donnée possède : un bulbe classique quand $n > 2$, ou un pseudo-bulbe quand $n < 2$. De plus, selon sa valeur, il peut aussi servir pour modéliser des barres, tel que réalisé par Gadotti (2008) – ceci, pour $n < 1$. Des composants avec des indices élevés de Sérsic ont tendance à générer un signal similaire à celui créé par une source ponctuelle, après la dégradation de l'image par la convolution avec la PSF.

Donc, le profil de brillance d'une galaxie qui possède les deux composants peut être écrit comme étant la somme des profils du disque et du bulbe, c'est-à-dire : $I(r) = I_d(r) + I_b(r)$. Il est important de remarquer, que dans la forme dans laquelle nous avons écrit les profils du bulbe et du disque, les significations dans l'expression pour le profil de brillance total $I(r)$ des paramètres I_{d0} et I_{b0} sont distincts entre eux, ainsi que ceux de r_d et r_b, et ne doivent pas être confondus.

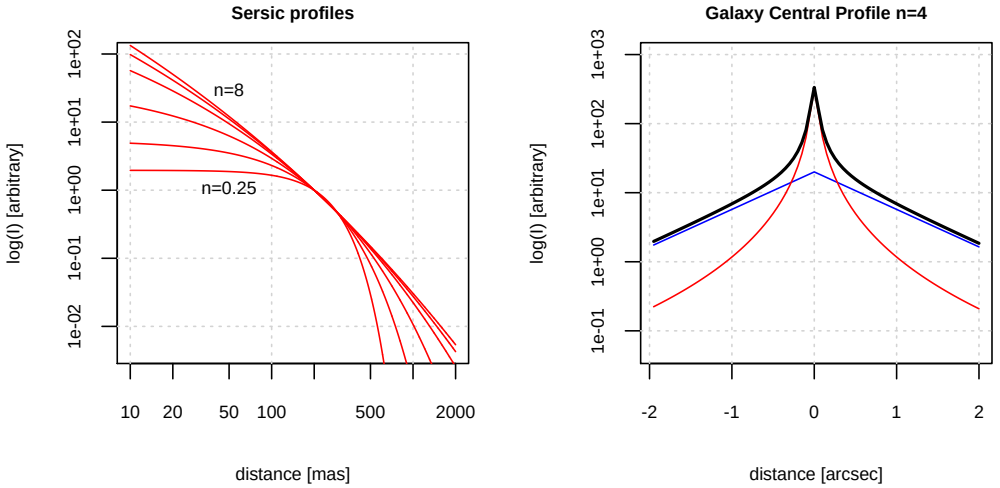


FIGURE 6.1 – Représentation de profils de galaxies considérés dans ce Chapitre. La Figure de gauche montre le profil de Sérsic ($I_{0b} = 1$ et $r_b = 0.2''$) pour différentes valeurs de n (0.25, 0.5, 1, 2, 4, 6, 8). La Figure de droite présente un profil exponentiel en bleu, un profil de Sérsic avec n = 4 en rouge et le profil composé en noir.

D'autres profils qui sont utilisés dans la littérature, comme dans Peng et al (2010), sont le profil de Ferrer, employé pour ajuster des barres et des lentilles, le profil de King, adopté pour les amas globulaires, le profil de Nuker, employé pour la distribution centrale de lumière de galaxies proches, et le profil de van der Kruit & Searle (1981), qui permet l'ajustement de galaxies edge-on. Cependant, il est intéressant que des décompositions soient toujours réalisées avec les profils les plus simples et avec le plus petit nombre possible de composants.

6.1.2 Ajustement de profils dans l'espace de Radon

Pour réaliser des études de décomposition bulbe/disque et de détermination de paramètres morphologiques de ces composants structurels à partir des données de Gaia, il est possible d'adopter deux voies distinctes. La première consiste en un ajustement des profils directement sur des images reconstruites avec les données

originales du satellite, en utilisant les codes déjà disponibles, tels que le BUDDA et le GALFIT, par exemple. Cependant, nous avons vu, dans le Chapitre 5, certains exemples de reconstructions de données simulées pour des galaxies à partir de l'un des meilleurs algorithmes disponibles. On peut rapidement en conclure qu'actuellement cette voie est impraticable.[3]

Une autre voie possible, est la réalisation de l'ajustement non pas du profil de luminosité de la galaxie dans l'espace réel, mais de la transformée de ce profil dans l'espace de Radon. Comme nous l'avons vu dans le Chapitre 2, les observations Gaia constituent fondamentalement un sous-ensemble échantillonné du signal bi-dimensionnel de la galaxie dans l'espace de Radon. De cette façon, lorsque nous connaissons les profils que nous désirons ajuster dans l'espace réel, et que nous savons comment calculer les transformées de ces profils pour l'espace de Radon, il suffit de les ajuster directement aux observations Gaia dans cet espace.

En principe, nous pourrions réaliser les transformations analytiques de ces profils, ce qui du point de vue de temps de calcul optimiserait significativement la procédure d'ajustement mais qui serait assez difficile à réaliser s'il s'agissait d'inclure des effets comme la convolution par exemple. C'est pourquoi notre système d'ajustement est basé sur des transformations numériques d'observations simulées de certains profils dans l'espace réel vers l'espace de Radon. Une représentation schématique du fonctionnement de ce système peut être vue sur la Figure 6.2.[4].

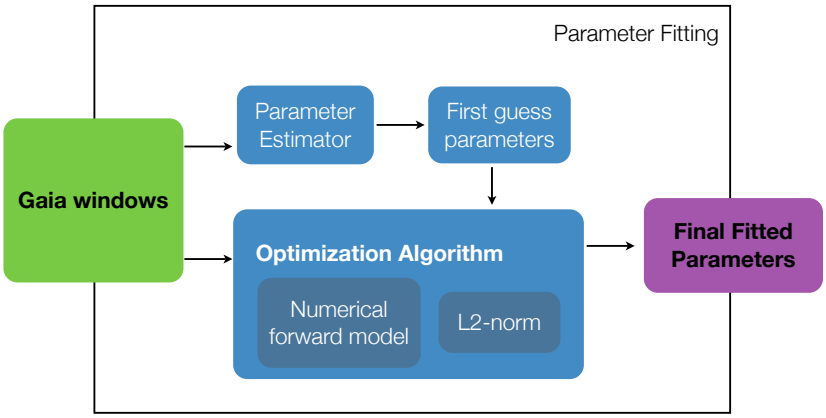


FIGURE 6.2 – Schéma du flux de données dans le système d'ajustement de profil construit pour la détermination de paramètres de profils morphologiques à partir des données de Gaia.

Ce système permet l'adoption de profils de luminosité aussi simples ou aussi complexes que ce que l'on veut dans la construction des modèles, même si les profils ne possèdent pas de représentation analytique simple dans l'espace de Radon.

3. Notons, cependant, qu'au cas où de meilleurs algorithmes de reconstruction seraient proposés/mis en œuvre dans le futur, cette voie pourrait commencer à être considérée, et servirait même comme test de consistance entre les diverses stratégies.

4. En vert, sont représentées les données qui entrent dans le système, en bleu, les processus ou les données internes et en en violet les données qui sortent – ces couleurs seront utilisées dans les diagrammes de ce Chapitre.

6.2 Construction du modèle

Une possibilité pour mettre en place la procédure d'ajustement commentée serait d'utiliser GIBIS comme modèle direct, c'est-à-dire, d'adopter ce simulateur pour produire un grand nombre de fenêtres simulées balayant l'espace des paramètres. Ces observations simulées qui pourraient alors être comparées avec les fenêtres observées par le satellite. Néanmoins, dans la pratique cela n'est pas possible, car les exigences computationnelles de ce simulateur (en termes de temps d'exécution et de mémoire) ne permettent pas l'analyse du nombre d'objets espéré.

De cette manière, nous avons construit un modèle direct simplifié, dont la représentation schématique est donnée en Figure 6.3.

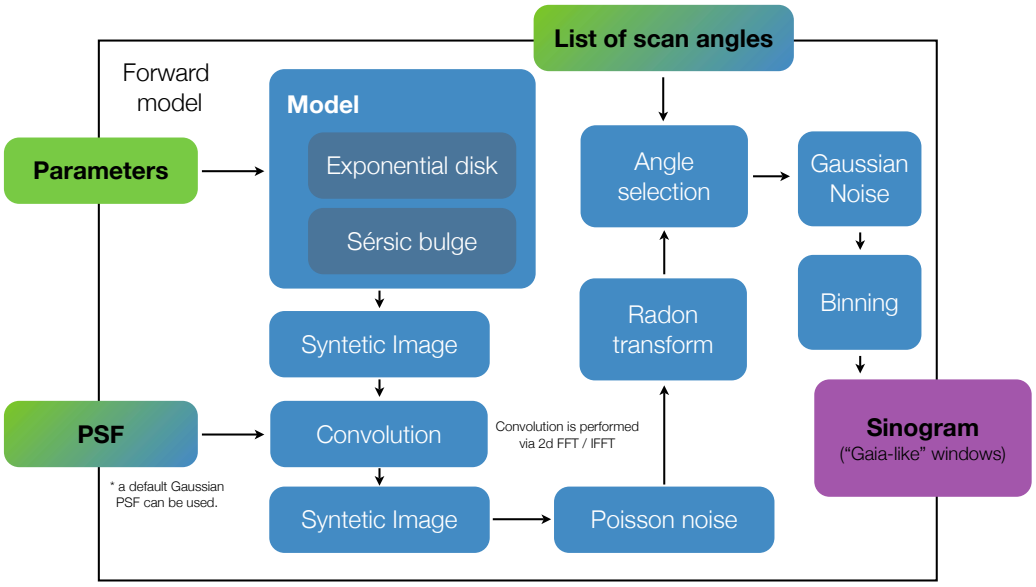


FIGURE 6.3 – Schéma de la construction du modèle direct du signal d'un profil de brillance d'un objet étendu. Le modèle peut être n'importe quelle fonction.

Le premier pas dans la construction du modèle direct est l'échantillonnage du profil de brillance de l'objet qui sera étudié dans une matrice. Pour cela, un modèle de l'objet qui est simulé est construit à partir d'un disque avec un profil exponentiel et d'un bulbe avec un profil de Sérsic. Durant cet échantillonnage, des transformations géométriques sont nécessaires pour inclure dans le modèle l'angle de position de l'objet dans le ciel (une rotation simple) et l'ellipticité. De plus, la mise en œuvre de ces transformations permet de prendre en compte des facteurs de *diskness/boxiness* de ces objets, comme dans Athanassoula et al (1990) :

$$r(x,y) = \left[|x - x_{gal}|^{C_0+2} + \left| \frac{y - y_{gal}}{1-e} \right|^{C_0+2} \right]^{\frac{1}{C_0+2}} \tag{6.3}$$

où e représente l'ellipticité et C_0 un paramètre qui représente de combien la galaxie est *disky/boxy* : $C_0 = 0$ ellipse normale, $C_0 < 0$ galaxie *disky* (avec un format similaire

à un diamant) et $C_0 > 0$ galaxie *boxy* (avec un format similaire à une boîte). La Figure 6.4 montre des exemples limites de galaxies simulées avec les trois possibilités de valeurs pour le paramètre C_0.

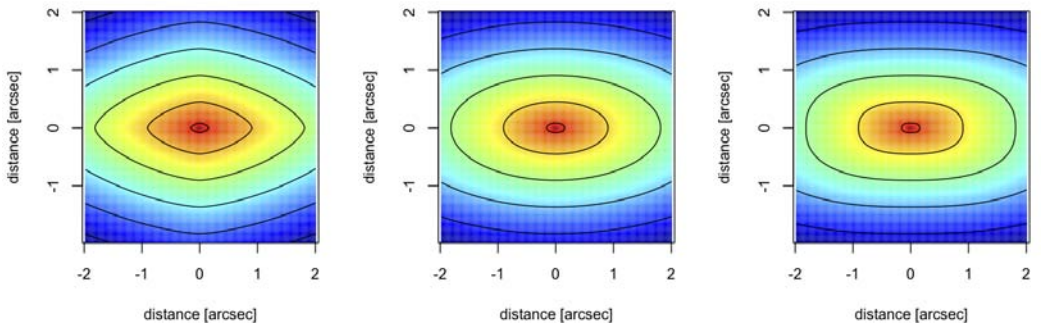


FIGURE 6.4 – Images synthétiques de galaxies produites à partir de l'échantillonnage du modèle et de l'application de transformations géométriques. Le paramètre d'ellipticité vaut $e = 0.5$ dans tous les profils, $C_0 = -0.5$ à gauche, $C_0 = 0.0$ au centre et $C_0 = +0.5$ à droite (carte de couleurs logarithmique allant du bleu au rouge).

L'échantillonnage est réalisé dans une matrice dont le nombre de pixels peut être configuré par l'utilisateur, ainsi que la taille de ces pixels. Dans la région interne à une certaine distance du centre, dont la configuration standard est 10 mas, le profil est montré comme s'il était à cette distance, évitant de trop grandes valeurs.

Après l'échantillonnage, une convolution avec une PSF est réalisée. Pour cela, l'image synthétique originale est transformée vers l'espace de Fourier, multipliée par la transformée de Fourier de la PSF et le signal résultant est ramené dans l'espace réel au moyen d'une transformée inverse, générant l'image synthétique convoluée. La PSF peut être adoptée comme une fonction gaussienne ou une PSF lue à partir d'une bibliothèque, telle que celle disponible dans GaiaSimu. Un exemple de ce modèle, avec une PSF gaussienne de FWHM=180 mas, peut être vu sur la Figure 6.5.

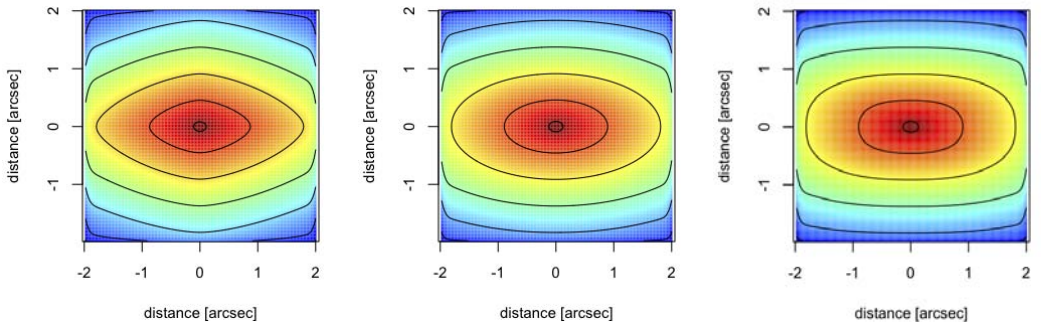


FIGURE 6.5 – Profils d'images synthétiques produites à partir de l'échantillonnage, de l'application de transformations géométriques et de convolution par une PSF gaussienne de FWHM de 180 mas. Les galaxies sont celles celles de la Figure 6.4.

L'image synthétique convoluée passe donc par un processus de re-échantillonnage poissonnien, où la valeur dans chaque pixel est remplacée par un échantillon aléatoire pris à partir d'une distribution de Poisson dont la moyenne est la valeur originale du pixel. Un exemple des résultats de ce modèle, pour les trois même galaxies que précédemment, peut être vu sur la Figure 6.6.

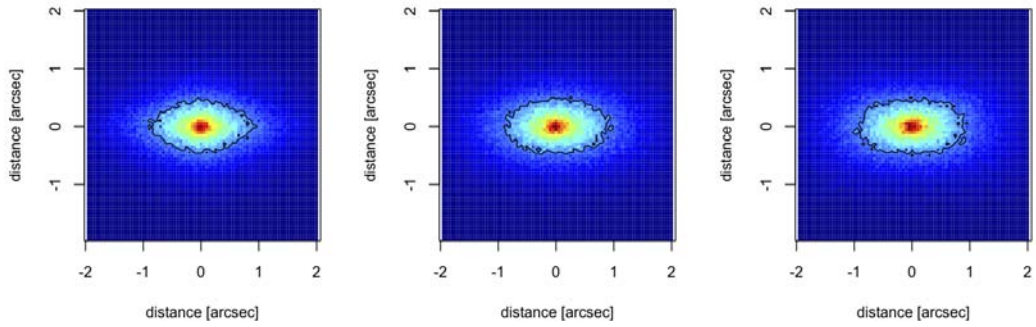


FIGURE 6.6 – Images synthétiques de galaxies générées à partir de l'échantillonnage du modèle, de l'application de transformations géométriques, de convolution par une PSF gaussienne de FWHM de 180 mas et un re-échantillonnage poissonnien. Les galaxies sont celles de la Figure 6.4, mais ici la carte de couleurs est linéaire.

Le pas suivant est de repasser l'image synthétique dans l'espace de Radon. Pour cela, l'image est remplie avec des valeurs nulles dans tous les points hors d'un certain rayon (la configuration standard est un rayon de 2", environ le rayon couvert par les fenêtres du *Sky-mapper* de Gaia). Alors, la transformée de Radon est calculée.[5] Le signal des galaxies utilisées comme exemple dans cette section dans l'espace de Radon peut être vu sur la Figure 6.7 – il est intéressant de remarquer que plus la valeur de C_0 est petite, mieux le signal est localisé dans cet espace.[6]

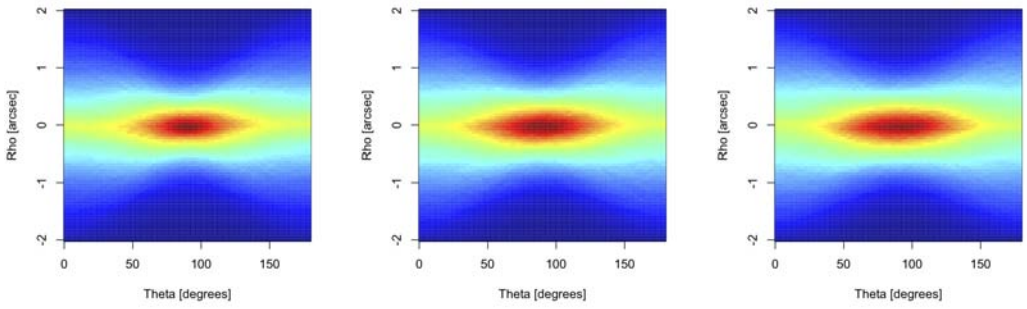


FIGURE 6.7 – Sinal das galáxias de exemplo desta seção no espaço de Radon.

5. Dans la mise en œuvre actuelle de la méthode ici décrite, qui a été écrite en langage R, la transformée est calculée au moyen du paquet PET (Schulz, 2006).

6. Un phénomène similaire survient pour de plus grandes ellipticités.

Comme Gaia n'échantillonne pas l'espace dans la direction θ de façon homogène, seuls les angles observés doivent être considérés et le modèle doit donc prendre ceci en compte. Pour cela, il est possible de fournir une liste d'angles au code, ou bien les coordonnées galactiques dans lesquelles l'objet est simulé. Dans ce second cas, le code utilise les routines en Java de la bibliothèque GaiaSimu pour calculer la loi de balayage nominale de la mission et déterminer la liste des angles de passage qui seront observés. Un exemple avec les angles échantillonnés dans les coordonnées $(l, b) = (160°, 50°)$ peut être vu sur la Figure 6.8.

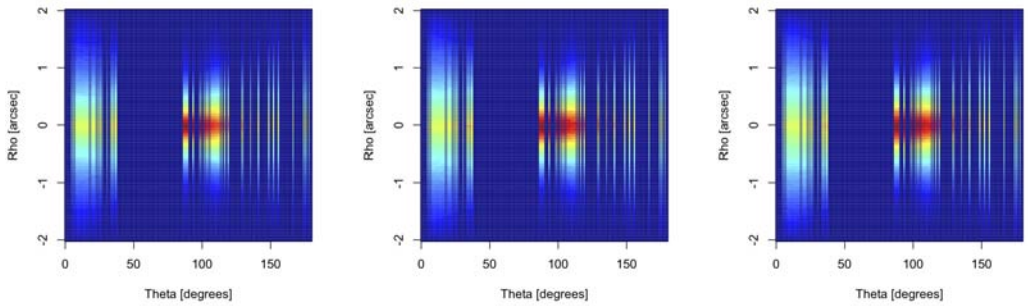


FIGURE 6.8 – Signal des galaxies en exemple dans cette section dans l'espace de Radon avec des passages observés par Gaia dans la direction $(l, b) = (160°, 50°)$.

Finalement, le bruit gaussien représentant le bruit de lecture est ajouté, en plus de la combinaison (binning) des pixels en *samples* des fenêtres. La Figure 6.9 montre les sinogrammes résultants de fenêtres des *Sky-mappers* et *Astro-fields 2, 5* et *8*.

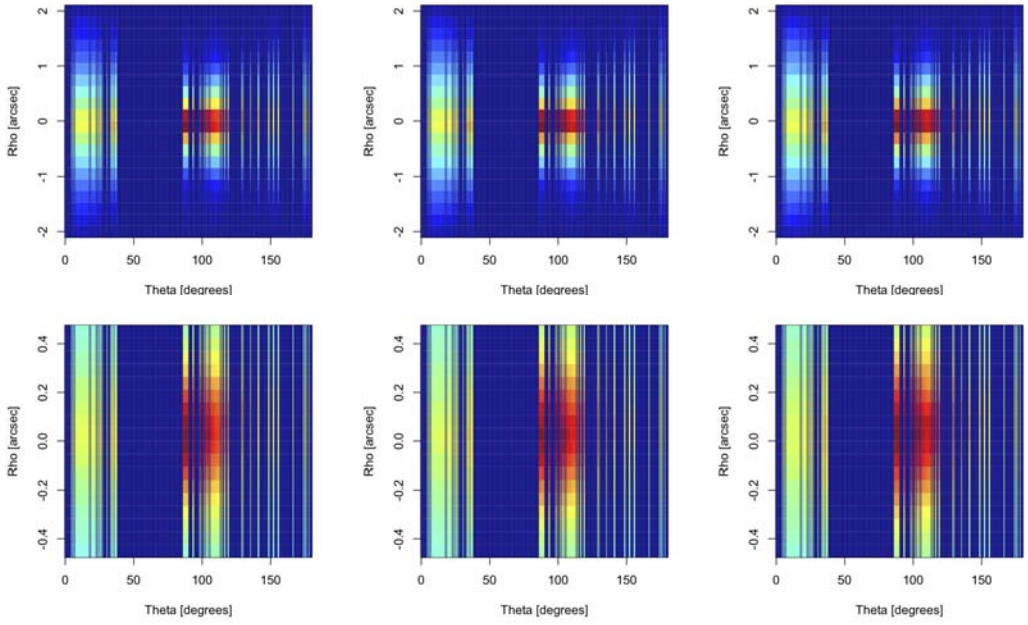


FIGURE 6.9 – Synogrammes du type SM et AF simulés dans $(l, b) = (160°, 50°)$.

Le modèle direct décrit ici permet la simulation de passages pour n'importe quels objets de façon efficace : chaque réalisation du processus direct complet demande $\sim 197 \pm 3$ ms dans $\sim 80\%$ des exécutions dans un centre Intel Core i7-620M 2.66 GHz (1000 exécutions ont été réalisées). De plus, cette efficacité permet que ce modèle soit utilisé pour réaliser des ajustements de paramètres de tous les profils de passages observés – pour cela les termes de bruit sont négligés. Comme nous l'avons déjà signalé, nous ne considérons ici que des galaxies formées par des bulbes de Sérsic et des disques exponentiels.[7] Cependant aussi bien la méthode décrite ici que le code mis en œuvre sont génériques et peuvent traiter d'autres profils de brillance.

6.3 Optimisation des paramètres

Une fois en possession d'un modèle direct, il est nécessaire de déterminer les paramètres pour lesquels le modèle et les données observées sont les plus proches, c'est-à-dire, résoudre un problème d'optimisation. L'une des possibilités pour résoudre ce problème est l'optimisation d'une norme euclidienne (aussi appelée norme L2) du vecteur formé par les différences entre les valeurs prévues par le modèle pour un certain vecteur $\mathbf{p}$ de paramètres ($f(\mathbf{p})$) et les valeurs observées ($\mathbf{d}$), pixel à pixel du sinogramme. C'est-à-dire, on désire résoudre le problème d'optimisation suivant :

$$\min_{\mathbf{p}} ||f(\mathbf{p}) - \mathbf{d}||_2 \tag{6.4}$$

La résolution idéale du problème décrit ci-dessus dépend beaucoup du format de la fonction f.[8] Cependant, afin de permettre l'utilisation de tous types de modèle, qu'ils aient des représentations analytiques ou non, nous avons décidé d'utiliser dans les codes que nous avons développés pour la mission Gaia, un optimisateur générique, basé sur une heuristique dénommée Algorithme Génétique (voir section 6.3.2), et un deuxième optimisateur basé sur des dérivées numériques de la fonction f.

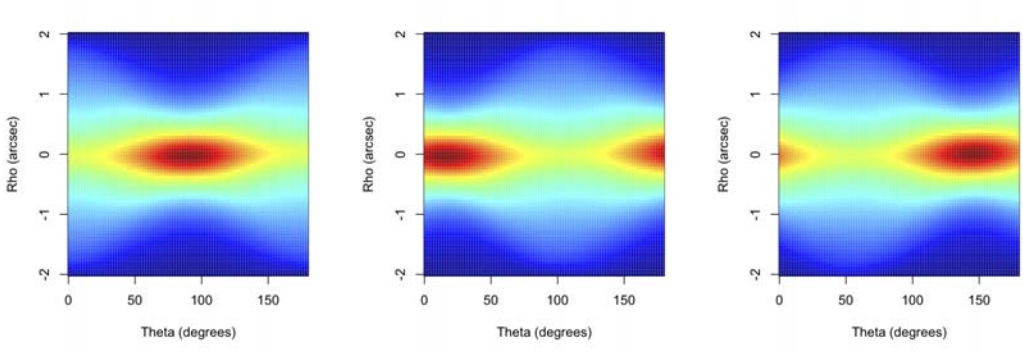
Le concept du fonctionnement du système mis en œuvre est basé sur trois pas intermédiaires. Le premier, consiste a effectuer une première détermination de paramètres à partir d'une analyse peu dépendante des modèles du signal du sinogramme. Le deuxième pas est une optimisation des paramètres avec l'utilisation d'une heuristique dénommée Algorithme Génétique. Le troisième, est une optimisation finale de ces paramètres avec l'utilisation d'une méthode basée sur des dérivées numériques de la fonction qui est optimisée. Il faut remarquer ici que des optimisations peuvent être exécutées sans l'utilisation des premier et troisième pas, néanmoins ceci conduirait soit à une moins bonne précision des paramètres finaux, ou alors à un temps de convergence beaucoup plus grand. De plus, en fonction de la nature stochastique des Algorithmes Génétiques, pour chaque objet analysé, ce système est exécuté cinq fois de façon indépendante, et la solution de plus petite norme L2 est adoptée comme étant la solution.

7. Un fond de ciel constant est aussi généralement ajouté.

8. Les théorèmes *No Free Lunch* (Wolpert et al, 1997) prouvent qu'il n'existe pas de méthode d'optimisation qui soit meilleure que toutes les autres quelque soit le problème.

6.3.1 Estimations initiales

Une analyse, indépendante des modèles du sinogramme, nous permet déjà d'obtenir des estimations initiales pour certains paramètres. L'un d'entre eux est l'angle de position de l'objet dans le ciel. Comme on peut le voir sur la Figure 6.10, la position du pixel avec le plus haut signal dans l'axe θ varie suivant l'angle de position de l'objet de façon très corrélée. Ceci se produit car la plus grande valeur d'un signal bi-dimensionnel intégré dans un intervalle de projections angulaires surviendra justement quand l'angle θ de la projection est tel qu'il maximise la quantité d'informations de l'objet par intervalle d'intégration.



(a) (b) (c)

FIGURE 6.10 – Synogrammes d'une même galaxie simulée dans divers angles de position : 0°, 75° e 125°.

Pour l'axe x du plan focal aligné dans la direction Nord-Sud, la relation entre la coordonnée θ et l'angle de position PA est donnée par

$$\mathrm{PA} = \begin{cases} 90° - \theta & \text{si } \theta \leq 90°, \\ 270° - \theta & \text{si } \theta > 90°. \end{cases} \tag{6.5}$$

Le sinogramme peut être préalablement filtré pour maximiser la possibilité d'application de cette analyse. Ce filtrage est destiné à minimiser les effets associés aux incertitudes dans la détermination du pixel de signal maximum et permettre l'application de cette analyse dans le cas de galaxies observées avec un bas niveau de signal/bruit. Cependant la plus grande partie de la littérature traite le filtrage du signal avec des bruits gaussiens, ce qui nous incite à nous ramener à ce cas en transformant le bruit de nos observations en un bruit gaussien.

La transformée d'Anscombe (Anscombe, 1948), est une transformée statistique qui à partir d'un signal qui suit une distribution poissonienne, génère un signal de forme presque gaussienne.[9] Originairement construite pour ce cas, elle a été généralisée par Bijaoui (1994) pour un cas de bruit mixte, formé par le bruit poissonnien original ajouté à un bruit gaussien, et publiée dans Murtagh et al (1995).

9. La transformation ne peut être appliquée qu'aux données des fenêtres de Gaia, parce que la somme de variables aléatoires poissonniennes indépendantes (pixels sur le plan focal de Gaia) est toujours une variable poissonienne (samples des fenêtres transmises).

Pour un processus tel qu'une image astronomique, dans lequel le signal total peut être écrit comme étant $\alpha n + \gamma$, où α représente un gain, γ une variable gaussienne additive de moyenne g et d'écart type σ et n une variable poissonienne, une telle transformation peut être écrite de la manière suivante :

$$I_n(x,y) = \frac{2}{\alpha}\sqrt{\alpha I(x,y) + \frac{3}{8}\alpha^2 + \sigma^2 - \alpha g} \tag{6.6}$$

où $I_n(x,y)$ représente le signal transformé dans la position (x,y) et $I(x,y)$ le signal original. Cette transformation est telle que la valeur de la variance de $I_n(x,y)$ est indépendante de la valeur de $I(x,y)$, et I_n suit une distribution gaussienne.

Étant donné que le signal a son bruit transformé dans un terme additif gaussien, il doit être filtré par un filtre qui élimine les fréquences les plus élevées. Le filtre mis en œuvre est basé sur un *soft-thresholding* du signal dans l'espace de *wavelets*.[10] Le *soft-thresholding* (multiplication des coefficients de *wavelets* par des poids de plus en plus petits) a été choisi du fait que l'on ne s'attend pas à des discontinuités dans le signal d'un objet étendu transformé dans l'espace de Radon alors que couper de façon brutale les coefficients de *wavelets* à partir d'une certaine échelle (*hard-thresholding*) peut générer de telles discontinuités. La *wavelet* choisie est la Daubechies D4, telle que décrite dans Sanz et al (1999). Cette base a été sélectionnée car elle permet de décrire convenablement des signaux constants et linéaires comme ceux qui sont attendus dans les diverses régions d'un sinogramme de structures étendues (telles que des galaxies, par exemple). De plus, celle-ci est numériquement efficace, ne comptant que sur quatre coefficients pour sa construction.

Un code pour la réalisation de l'évaluation initiale des paramètres a été construit en R (le filtre utilisé a été mise en place par Whitcher, 2010), et un schéma qui représente le flux des données est présenté en Figure 6.11.

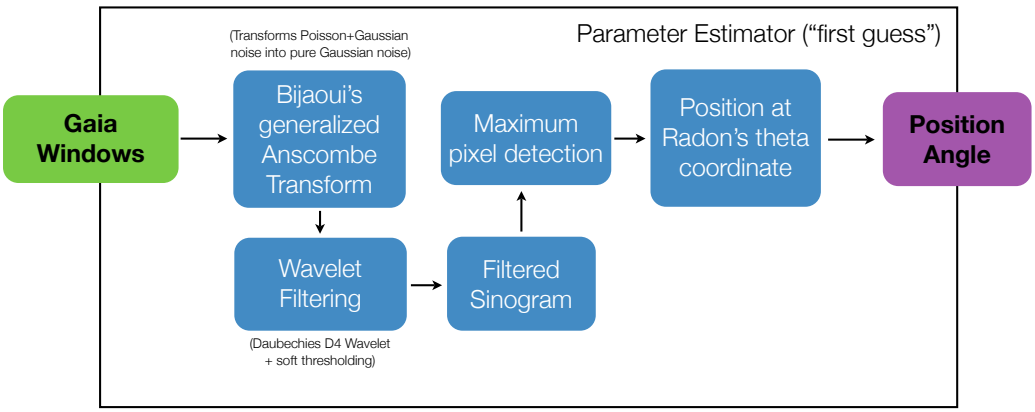


FIGURE 6.11 – Schéma du flux de données dans le système d'ajustement de profil construit pour la détermination des paramètres de profils morphologiques à partir des données de Gaia.

10. Une transformation de wavelet est la représentation d'un signal quelconque dans une base auto-similaire qui est localisée dans l'espace et dans le temps.

Le code a été utilisé pour tester l'efficacité de l'estimation de l'angle de position de 1000 galaxies simulées avec une ellipticité égale à 0.6, dans deux conditions de rapport signal sur bruit, aux coordonnées $(l, b) = (160°, 50°)$. La comparaison entre les valeurs simulées et obtenues peut être vue sur la Figure 6.12.

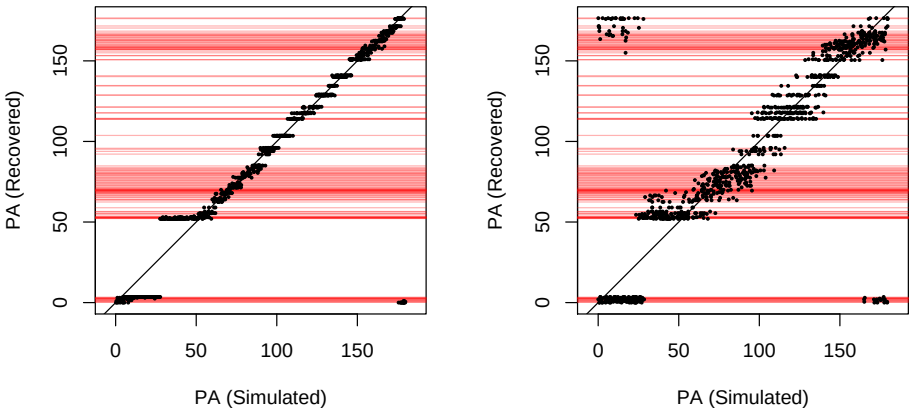


FIGURE 6.12 – Comparaison entre les angles de position simulés et estimés par la méthode décrite dans cette section pour 1000 galaxies générées dans les coordonnées $(l, b) = (160°, 50°)$. En (a) sans bruit gaussien ajouté et en (b) avec S/B∼0.2. Les lignes rouges représentent la loi de balayage de Gaia.

Les résultats démontrent qu'une estimation initiale de l'angle de position à partir de la méthode décrite ici est extrêmement robuste à condition que l'information soit disponible pour l'analyse. Naturellement la méthode se limite à déterminer des angles répartis selon la loi de balayage adoptée par le satellite Gaia, car l'échantillonnage du signal dans la direction θ de l'espace de Radon est réalisée de façon incomplète et non-homogène par ce satellite.

De plus, la loi de balayage de Gaia limite aussi la détermination d'autres paramètres, qui en principe peuvent être estimés en utilisant des techniques similaires. Par exemple, une évaluation de l'ellipticité de la galaxie peut être réalisée à partir de la détermination d'une dimension de l'objet (par binarisation par seuil, par exemple) qui est mesurée dans les colonnes du sinogramme relatives à la valeur maximum et à sa coordonnée θ complémentaire. Néanmoins, comme nous l'avons commenté à la fin de ce Chapitre, certaines nouvelles techniques pourront peut-être permettre la réalisation d'estimations initiales pour ces autres paramètres dans un futur proche.

Nous avons réalisé un dernier test en appliquant cette méthode à 1000 galaxies d'ellipticité 0.6 simulées avec des angles de position aléatoires et dans des régions aléatoires du ciel (suivant une distribution uniforme sur la sphère céleste). La distribution des passages pour chaque région simulée a été prise en compte dans la construction des sinogrammes. De la même manière que le test antérieur, celui-ci a été réalisé dans deux conditions de signal/bruit limites : sans bruit et avec S/B∼0.2. Les résultats de cette simulation peuvent être vus sur la Figure 6.13.

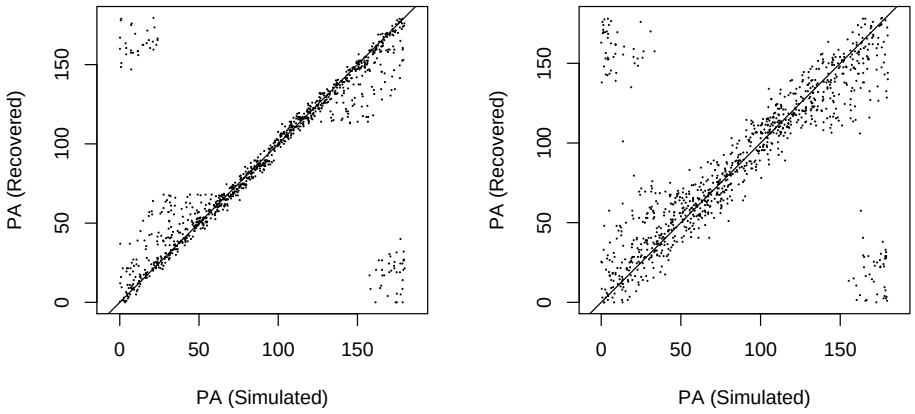


FIGURE 6.13 – Comparaison entre les angles de position simulés et estimés par la méthode décrite dans cette section pour 1000 galaxies générées à des coordonnées aléatoires dans le ciel. En (a) sans ajout de bruit et en (b) avec S/B∼0.2.

Ces résultats démontrent que même avec le problème de discontinuité dans les angles balayés par le satellite sur des coordonnées individuelles, du point de vue statistique, les angles de position des galaxies sont récupérés de manière fiable dans n'importe quelle position du ciel par la méthode décrite dans cette section, même dans des conditions défavorables du rapport signal sur bruit.

6.3.2 Algorithmes Génétiques

Un Algorithme Génétique (ou GA, *Genetic Algorithm*) est une heuristique[11] popularisée principalement par Holland (1975)[12] qui est utilisée pour l'optimisation générique globale de fonctions – ce qui veut dire qu'en principe un GA peut être appliqué pour trouver les points optimums globaux de n'importe quelles fonctions.[13]

Le GA est basé sur la transformation d'un problème mathématique d'optimisation dans un analogue du processus évolutif biologique. Initialement une population de solutions aléatoires est générée dans l'espace de paramètres de la fonction à optimiser : chacune de ces solutions candidates est considérée comme étant un individu. Puis cette population passe par divers processus imitant l'évolution biologique, et à chaque nouvelle génération, il est garanti par l'un de ces processus (clonage) qu'au moins la meilleure solution de la génération antérieure continuera à exister, et que de meilleures solutions peuvent apparaître.

Pour cela, chacune des solutions d'un GA est codifiée sous la forme de chromosomes, qui ne sont que des représentations du vecteur de paramètres (sous forme

11. Ceci signifie qu'un GA est une technique développée qui trouve une solution qui n'est pas prouvée être la solution correcte, mais qui en général (basé sur l'expérience) est une « bonne »solution.

12. Les premières utilisations de méthodes évolutives datent de Barricelli (1954) (d'après Fogel, 2006). Pour plus de détails sur le développement historique et la théorie des GAs, voir Reeves & Rowe (2002), Fogel (2006) et Sivanandam & Deepa (2007).

13. Cependant, rien ne garantit qu'il soit le moyen le plus rapide pour trouver une telle solution.

binaire ou de floating points) qui mène à cette solution. La fonction que l'on veut optimiser, est généralement décrite comme étant la fonction de *fitness*, et elle est calculée pour chaque individu, ce qui signifie que la fonction à optimiser doit être calculée autant de fois qu'il existe d'individus dans cette population. Dans le cas particulier du problème traité dans ce chapitre, la fonction de fitness calcule une norme L2 (en utilisant le modèle direct), les paramètres de ce modèle (qui sont les individus de la population) et les données.

Après le calcul de *fitness*, une nouvelle génération de la population est créée à partir de processus appliqués à la population actuelle. C'est à ce moment que les divers implémentations de GAs diffèrent entre elles. Pour optimiser les solutions telle ou telle opérateur génétique est ajoutée ou pas. La Figure 6.14 montre de forme schématique comment un GA fonctionne.

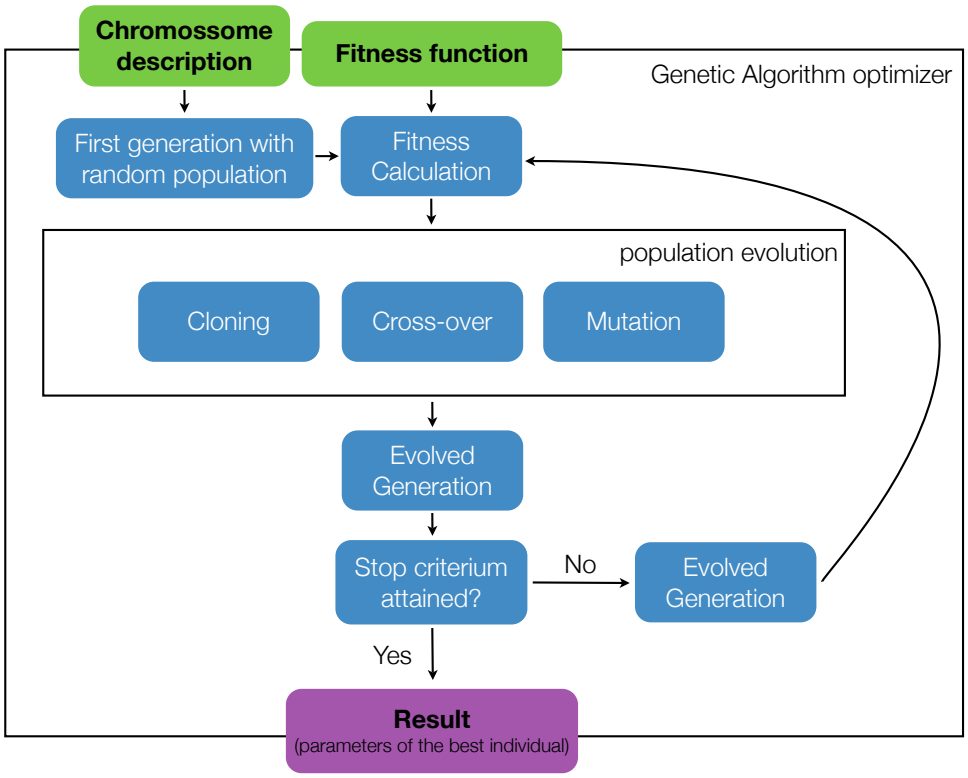


FIGURE 6.14 – Schéma conceptuel de fonctionnement d'un algorithme génétique. Les paramètres de configuration du GA ne sont pas représentés (taille de la population, taux de mutation, etc.).

Les processus ou opérations adoptés dans les GAs pour faire évoluer les générations sont le clonage, la mutation et la reproduction ou crossover. Comme leurs originaux biologiques, ces opérations peuvent agir sur un ou plusieurs individus de la population. Le clonage, par exemple, agit sur un unique individu, et ne représente qu'une copie d'un ou de plusieurs membres de la population entre une génération et autre. Il est

appliqué pour préserver la meilleure solution entre deux générations subséquentes.

Une autre opération essentielle dans les GAs est la reproduction, ou *crossover*. Les individus avec les meilleures valeurs de *fitness* (individus les mieux adaptés), sont ceux qui possèdent une plus grande probabilité de subir un *crossover* en se reproduisant avec d'autres individus et engendrant de nouveaux individus dont les génotypes sont formées par des parties des génotypes des deux individus « parents ». Les détails de comment ce processus s'opère, tel que la fonction de distribution de probabilité pour la sélection, si un individu peut se reproduire plus d'une fois par génération, le point du génotype qui est changé, etc., dépendent de l'implémentation du GA.

La dernière opération fondamentale est la mutation, qui peut être de deux types : locale ou globale. Dans une mutation locale un individu de la population peut voir son génotype altéré de manière complètement aléatoire en un point donné.[14] D'un autre côté, dans une mutation globale, tous les membres de la population sont affectés, mais quand ce processus est mis en œuvre, en général, l'individu le mieux adapté n'est pas affecté, ce qui garantit pour la nouvelle génération l'existence d'au moins un individu aussi bon ou même le meilleur de la génération antérieure.

Les GAs sont connus pour être efficaces pour trouver la région du point optimum, mais en général ils sont lents pour déterminer ce point. Selon Sivanandam & Deepa (2007), pour résoudre ce problème on adopte des heuristiques basées sur l'alliance d'un GA et d'une méthode plus rapide pour l'optimisation locale, telle qu'une méthode de Quasi-Newton (celles-ci sont appelés algorithmes mémétiques). Dans ce travail, nous avons adopté un optimisateur évolutif disponible dans le langage R dénommé `RGENOUD` (Sekhon & Mebane, 1998; Mebane & Sekhon, 2007), qui est basé sur l'alliance d'une implémentation d'Algorithme Génétique avec des méthodes mises en œuvre dans l'optimisateur `optim` (également en R). Pour ce travail nous avons choisi comme méthode basée sur des dérivées BFGS, présenté dans la prochaine section.

6.3.3 BFGS

Proposé de façon indépendante par Broyden (1970), Fletcher (1970), Goldfarb (1970) et Shanno (1970), le BFGS est l'un des algorithmes de la famille de Quasi-Newton. Ceux-ci sont des algorithmes basés sur une approximation de deuxième ordre de la fonction à optimiser, ainsi que sur l'algorithme de Newton. Néanmoins au lieu d'utiliser la matrice Hessienne, les algorithmes de Quasi-Newton calculent une approximation de cette matrice qui est mise à jour à chaque nouvelle itération. Une description détaillée de ces méthodes peut être retrouvée dans Nocedal & Wright (1999).

Comme dans la méthode de Newton, dans le BFGS une approximation de la fonction à optimiser est calculée par une expansion en série de Taylor de deuxième ordre :

14. Selon la représentation des paramètres du problème dans le génotype, ceci signifie une modification aléatoire d'un ou plusieurs paramètres.

$$f(\mathbf{x}_k + \Delta\mathbf{x}) \approx f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T \Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^T \nabla^2 f(\mathbf{x}_k)\Delta\mathbf{x} \tag{6.7}$$

dont la dérivée est exprimée par :

$$\nabla f(\mathbf{x}_k + \Delta\mathbf{x}) \approx \nabla f(\mathbf{x}_k) + \nabla^2 f(\mathbf{x}_k)\Delta\mathbf{x} \tag{6.8}$$

Comme au point extrême, le gradient de la fonction doit être nul, étant donné que si l'on est à un point $\mathbf{x}_k$, la direction $\mathbf{p}_k$ dans laquelle le prochain point $\mathbf{x}_{k+1}$, le plus proche de l'extrême, doit être recherchée est donnée par :

$$\Delta\mathbf{x} = \alpha_k \mathbf{p}_k = (-\nabla^2 f(x_k))^{-1} \nabla f(\mathbf{x}_k) \tag{6.9}$$

Pour résoudre l'équation ci-dessus sans qu'il soit nécessaire de calculer à chaque itération la matrice de dérivées secondes et leurs inverses, ce qui en général est computationnellement cher, l'algorithme BFGS calcule une approximation de cette matrice par un processus itératif, où $B_k \approx \nabla^2 f(\mathbf{x}_k)$. Les formules du BFGS pour la mise à jour de B_{k+1} et de son inverse, H_{k+1}, sont écrites de la manière suivante :

$$B_{k+1} = B_k + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \Delta\mathbf{x}_k} - \frac{B_k \Delta\mathbf{x}_k (B_k \Delta\mathbf{x}_k)^T}{\Delta\mathbf{x}_k^T B_k \Delta\mathbf{x}_k} \tag{6.10}$$

$$H_{k+1} = B_{k+1}^{-1} = \left(I - \frac{\mathbf{y}_k \Delta\mathbf{x}_k^T}{\mathbf{y}_k^T \Delta\mathbf{x}_k}\right)^T H_k \left(I - \frac{\mathbf{y}_k \Delta\mathbf{x}_k^T}{\mathbf{y}_k^T \Delta\mathbf{x}_k}\right) + \frac{\mathbf{y}_k \Delta\mathbf{x}_k^T}{\mathbf{y}_k^T \Delta\mathbf{x}_k} \tag{6.11}$$

où

$$\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k) \tag{6.12}$$

A partir du choix d'un point initial $\mathbf{x}_0$ et d'une approximation B_0 (qui peut être $B_0 = I$), le BFGS réalise les pas suivants jusqu'à la convergence en $\mathbf{x}$:

1. Il résout l'équation 6.9 ; Il réalise une recherche linéaire pour déterminer un pas α_k adéquat dans la direction de $\mathbf{p}_k$;[15]
2. Il actualise $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha\mathbf{p}_k$;
3. Il calcule l'équation 6.12 ;
4. Il calcule l'équation 6.11 et itère le processus.

L'optimisateur adopté dans ce travail de thèse utilise l'implémentation du BFGS disponible dans le R au moyen de la fonction `optim` (le code original a été implémenté par Nash Nash, 1990), qui, en absence de dérivées partielles analytiques pour la fonction f calcule numériquement le gradient.[16]

15. Pour plus de détails, voir Nocedal & Wright (1999).
16. Ceci est une caractéristique importante pour l'application que nous faisons dans ce Chapitre, car nous désirons que la méthode puisse traiter tous types de profils.

6.4 Tests avec simulations

Nous avons réalisé des tests avec des simulations d'observations Sky-Mapper d'objets étendus dans le but d'évaluer l'efficacité de la méthode mise en œuvre. Ces tests comprennent la simulation de 1600 galaxies dont les paramètres du bulbe, du disque, et du fond de ciel ont été choisis de manière complètement aléatoire. Les intervalles de variation de ces paramètres peuvent être vu sur le Tableau 6.1.

Paramètre	Minimum	Maximum	Unité
rd	400	2000	mas
rb	100	rd (max=2000)	mas
Ib	5∗Fundo (min=5)	100	de flux
Id	Ib	1000	de flux
e	0.1	0.9	–
Fond du ciel	1	10	de flux

TABLE 6.1 – Intervalle de variation des paramètres dans la simulation. Toutes les galaxies possèdent $C_0 = 0$ et $n = 4$.

Pour la réalisation des tests, des passages du type Sky-mapper ont été simulés suivant la distribution angulaire définie par la loi de balayage nominale de la mission Gaia à la position $(l, b) = (160°, 50°)$.[17] Les passages simulés ont été organisés en sinogrammes sur lesquels ont été appliquées les méthodes décrites dans ce Chapitre pour l'ajustement des profils théoriques dans l'espace de Radon.

Les comparaisons entre les résultats obtenus pour les paramètres des galaxies et les valeurs simulées sont présentés sur les Figures 6.15 et 6.16. Nous pouvons vérifier que du point de vue statistique, tous les paramètres sont retrouvés. Néanmoins, il est aussi possible d'observer sur les Figures 6.15 (a) et (c) que les paramètres Ib et rb du bulbe, présentent une dispersion qui est relativement grande. Les valeurs des médianes et de la déviation absolue de la médiane[18] pour ces paramètres sont : $\text{med}(Ib) = 11 \pm 53\%$ et $\text{med}(rb) = -9 \pm 36\%$.

La grande valeur des erreurs (aussi bien systématique que statistique) est probablement due au fait que le code d'optimisation actuellement mis en œuvre considère exclusivement des données du type *Sky-Mapper* dans l'ajustement, alors que le signal du bulbe sera prédominant dans les parties les plus centrales de la galaxie, qui sont justement mal échantillonnées dans les observations du *Sky-Mapper*. La plus grande partie du signal du bulbe, dans ces simulations, est contenue dans très peu de pixels (deux ou trois, car chaque *sample* type SM est de l'ordre de ∼200 mas), et même les bulbes simulés avec un plus grand paramètre rb, sont perturbés par le signal prédominant du disque dans leurs régions les plus externes. La Figure 6.16 (c) montre que l'hypothèse expliquant l'amplitude des erreurs est plausible, étant donné que

17. Les outils développés pour ce test permettent la réalisation de tests similaires dans n'importe quelles autres régions du ciel de manière complètement automatique.

18. 17 Il s'agit d'une quantité qui indique la dispersion statistique d'un ensemble de données de forme robuste aux outliers (Venables & Ripley, 2002), étant définie par : MAD = mediana($|x_i -$ mediana($\mathbf{x}$)|), pour un vecteur $\mathbf{x}$.

plus le rayon du bulbe est grand, plus la dispersion de l'erreur fractionnelle autour de zéro semble être petite, ce qui peut être aussi conclu en vérifiant sur la Figure 6.15 (c) qu'à partir de *sim*600 mas les valeurs ajustées tendent vers la valeur idéale (même si la dispersion continue à être grande).

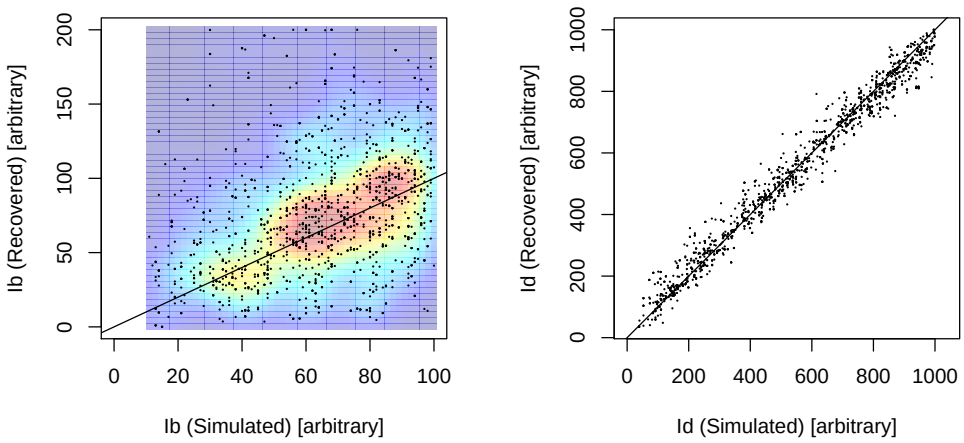
De manière à compenser l'erreur dans la valeur ajustée pour le paramètre *rb*, le code a tendance à surestimer systématiquement la valeur de l'intensité du bulbe dans le rayon ajusté, Ib, comme il peut être vu dans les Figures 6.15 (a) et 6.16 (a). Pour améliorer l'ajustement de la partie du profil théorique correspondante à ce composant structurel il faut inclure, dans l'optimisation de la décomposition de profil, les données d'observations du type *Astro-Field* qui possèdent une résolution spatiale trois fois plus grande, permettant alors un meilleur échantillonnage des régions les plus critiques pour cet ajustement.

Le profil *fitté* du disque n'est pas aussi bruité que celui du bulbe comme il peut être vu sur les Figures 6.15 (b) et (d). Ceci vient probablement du fait que le disque est bien échantillonné par les observations des fenêtres du type *Sky-Mapper*. Le rapport entre les valeurs simulées et récupérées par le code pour les paramètres des disques est très linéaire, avec une faible dispersion par rapport à la valeur idéale, en plus de l'absence d'effets systématiques significatifs : $\mathrm{med}(Id) = -1 \pm 7\%$ et $\mathrm{med}(rd) = 1 \pm 4\%$.

En observant la Figure 6.15 (d), on pourrait penser que les valeurs de rd seraient encore moins bien déterminées quand le disque devient plus grand, cependant, en vérifiant les erreurs relatifs de ce paramètre, représentées sur la Figure 6.16 (d), on s'aperçoit que ce n'est pas le cas, et que l'erreur fractionnelle est pratiquement constante sur tout l'intervalle étudié – toutefois, en termes absolus, la valeur de l'erreur augmente quand le disque augmente. De son côté, l'intensité centrale du disque, Id, présente une erreur en pourcentage qui se réduit quand ce paramètre croît. En effet plus le signal du disque est important, mieux seront déterminés ses paramètres (fait qui peut être également observé dans l'analyse des erreurs relatifs de Ib).

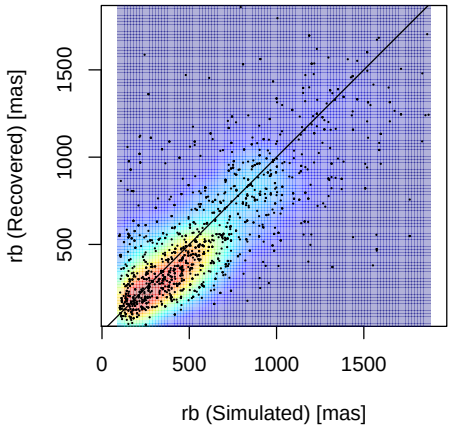
Les résultats montrent aussi que l'ellipticité de la galaxie est retrouvée de manière extrêmement précise, avec une dispersion essentiellement nulle. Ceci est dû principalement au fait que dans la version actuelle du code d'ajustement de profil nous considérons une résolution maximum de 0.1 pour ce paramètre, tandis que probablement l'erreur que ce code commet dans le calcul du paramètre est bien plus petite que la résolution qui a été considérée.

Finalement, les résultats obtenus pour l'ajustement du fond de ciel indiquent que le code possède une tendance systématique à surestimer les valeurs simulées, avec $\mathrm{med}(background) = 10 \pm 24\%$. Ceci peut survenir car les galaxies qui ont été simulées durant ce test sont des objets dont le signal est toujours beaucoup plus grand que le signal provenant du fond de ciel. La détermination de ce signal constant additif correspondant au fond de ciel peut absorber une petite fraction du signal des autres composants. Cet effet n'est finalement pas perçu dans l'analyse d'erreur des deux autres paramètres en raison de la très petite valeur du signal du fond de ciel par rapport au signal du bulbe et du disque.

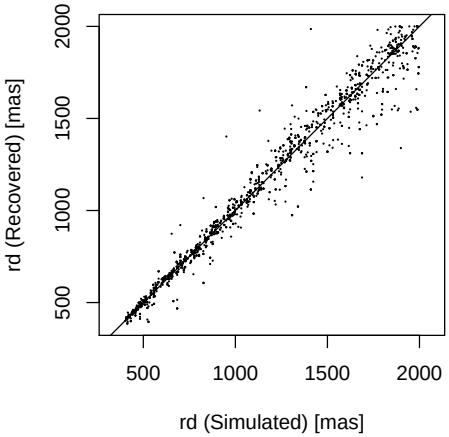


(a)

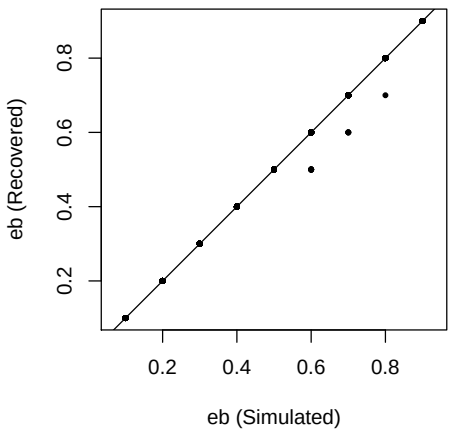
(b)



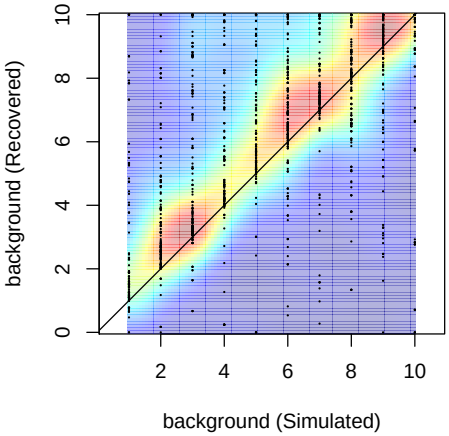
(c)



(d)



(e)



(f)

FIGURE 6.15 – Résultats obtenus à partir d'un test réalisé avec un ensemble de 1600 galaxies. Les simulations ont été réalisées dans la direction $(l, b) = (160°, 50°)$. Une droite x=y (en noir) indique la récupération idéale des paramètres. La carte de couleurs en fond des graphes plus dispersés indique la densité des points.

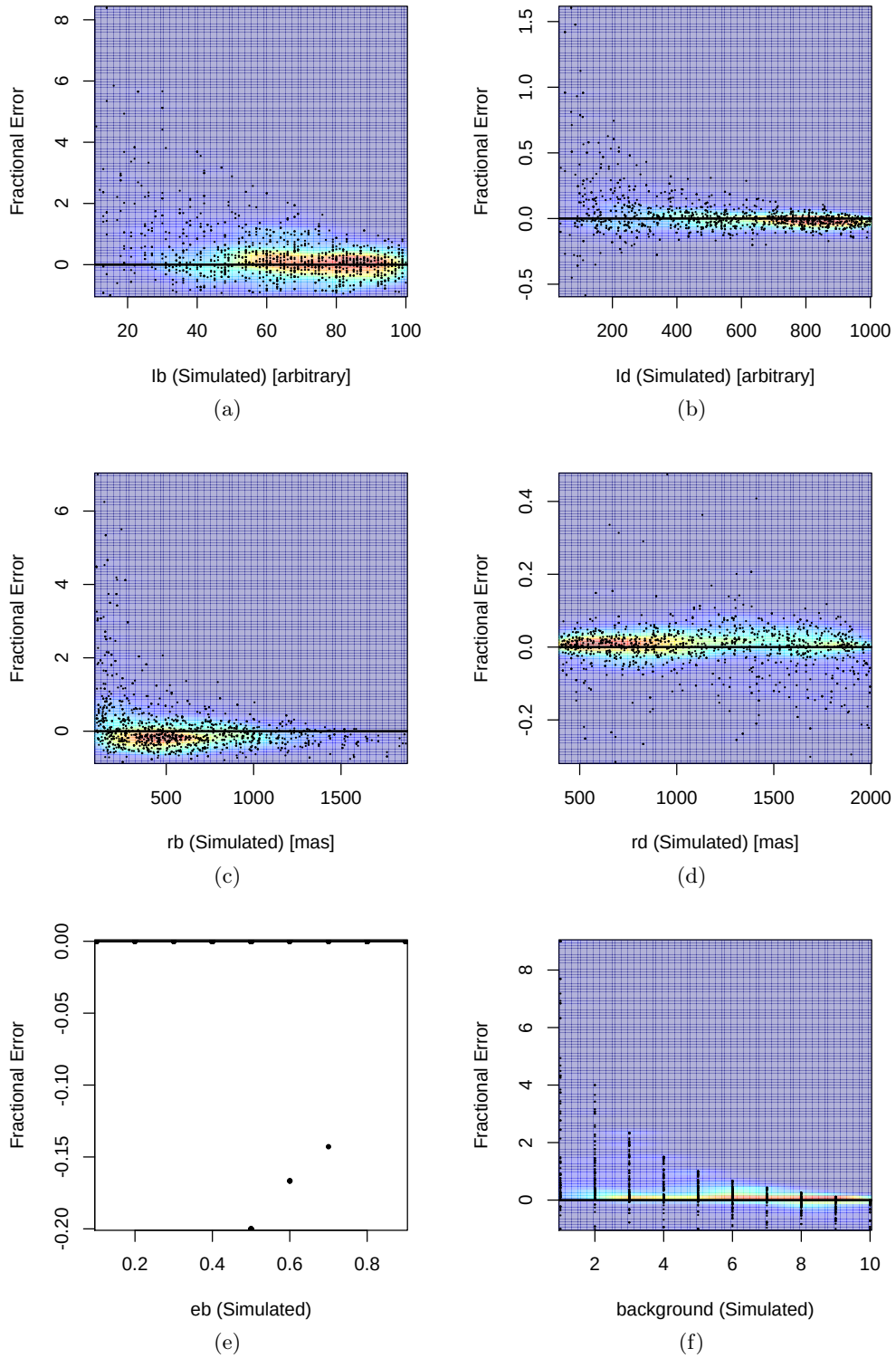


FIGURE 6.16 – Erreurs relatifs (dans le sens retrouvé moins simulé) des paramètres estimés par le code pour les 1600 galaxies simulées. La carte de couleurs en fond est représentative de la densité des points dans les régions respectives.

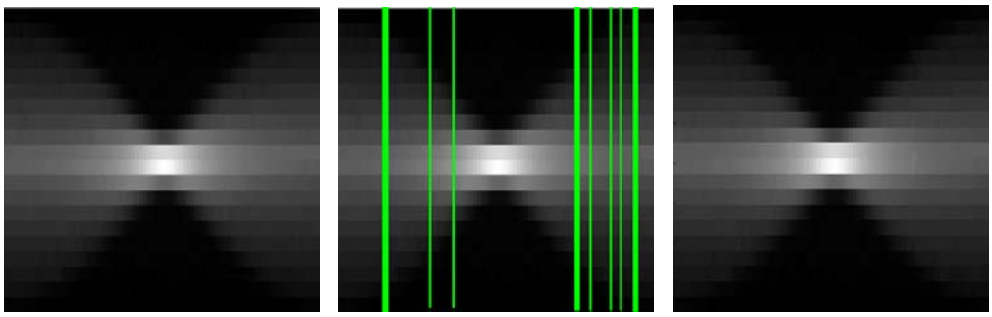
Ce test réalisé avec un ensemble statistiquement représentatif de galaxies, démontre qu'aussi bien les idées décrites ici pour l'analyse de profil des objets étendus, que le propre code prototype mis en œuvre dans R dans le but de tester ces idées, permettront l'analyse des d'objets étendus, révélant leurs structures morphologiques. Ceci permettra les études de décomposition bulbe/disque de millions de galaxies qui ne peuvent être observées qu'à partir de l'espace, sans qu'il soit nécessaire d'adopter des procédures de reconstruction détaillée d'images. Nous avons noté, cependant, que la reconstruction des images est toujours une étape nécessaire pour le choix du profil le plus apte à l'étude des objets (car plus il y a de paramètres libres, plus un signal donné est facile à ajuster), en plus du fait qu'une analyse et une reconstruction plus détaillées des objets qui ne présentent pas de bons ajustements avec des profils connus, pourront peut-être révéler de nouvelles caractéristiques inattendues.

6.5 Perfectionnements et possibilités

Les méthodes présentées dans ce chapitre pour résoudre le problème d'ajustement de profils théoriques de galaxies non résolues aux observations de Gaia peuvent être perfectionnées en certains points. Un premier perfectionnement qui pourra aider à l'ajustement des données morphologiques des régions les plus centrales des galaxies, est l'inclusion d'information de haute-résolution obtenue à partir de fenêtres venant d'observations des *Astro-Fields*. Avec cette inclusion, l'ajustement qui actuellement n'est réalisé qu'avec des données du type *Sky-Mapper*, serait réalisé de façon globale, prenant en compte toute l'information spatiale disponible dans les observations Gaia. Ceci pourra permettre non seulement un meilleur ajustement pour le bulbe, mais aussi pour tous types d'objets avec des tailles angulaires plus petites que quelques centaines de millisecondes d'arc.

Une autre possibilité est la réalisation d'un traitement du signal dans l'espace de Radon qui permet de mieux tirer profit de l'information contenue dans les observations. Tel que cela a déjà été montré antérieurement, les observations Gaia n'échantillonnent pas complètement cet espace dans la direction angulaire. Ces dernières années, certaines méthodes ont surgi pour résoudre un problème d'absence d'information (dénommé *inpainting*) dans des figures ou des séries temporaires, ou pour réaliser une extraction d'objets dans des images sans laisser de « trous »(ex. Criminisi et al, 2004; Fadili et al, 2009; Sato et al, 2010). Ce problème est équivalent au remplissage de certaines régions non observées par Gaia dans l'espace de Radon. Un exemple d'inpainting utilisant la méthode développée par Criminisi et al (2004) sur les données d'un sinogramme simulé, peut être vu sur la Figure 6.17.

En fonction de la régularité du signal d'une galaxie dans l'espace de Radon, il est possible que l'adoption de techniques similaires d'*inpainting* puisse permettre de compléter d'une manière satisfaisante une partie des angles non observés par le satellite en augmentant la quantité de données disponibles pour des analyses subsequentes, telles que les procédures d'évaluation initiale pour l'ajustement de profil commentées dans ce Chapitre – dans le cas de ce type d'utilisation, les pixels récupérés par *inpainting* devraient probablement être considérés avec un poids



(a) (b) (c)

FIGURE 6.17 – Exemple d'inpainting dans un sinogramme de données SM simulées. En (a), le sinogramme complet, en (b) le sinogramme avec les angles non observés indiqués en vert, et en (c) le sinogramme résultant de la reconstruction par la méthode d'*inpainting* décrite dans Criminisi et al (2004).

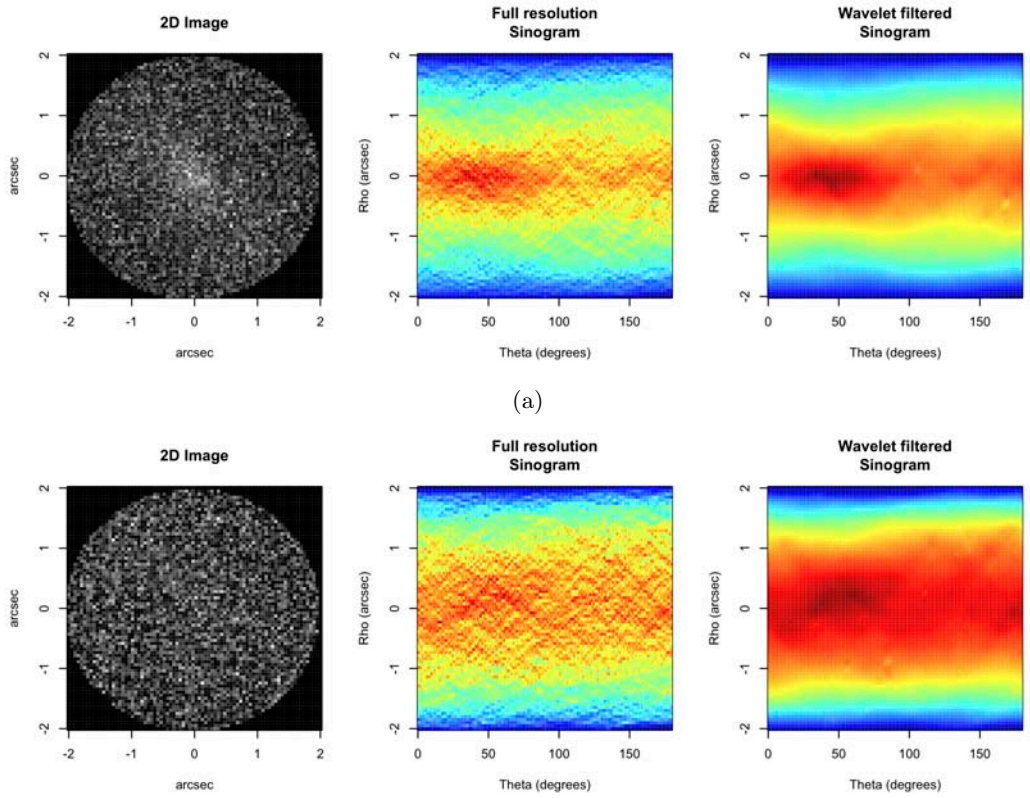
inférieur à celui des pixels qui représentent les données réelles.

Un autre perfectionnement qui doit être adopté est l'inclusion d'effets de CTI, ou *Charge Transfer Ineficiency* (décrit rapidement dans le Chapitre 2), probablement à partir de l'utilisation du modèle décrit dans Short (2009). C'est un effet important dans la mission Gaia, spécialement pour l'obtention de positions bien déterminées, et il peut être inclus dans la construction du modèle direct de manière relativement simple, avec l'application du modèle sur chaque angle séparément. Notre première approche de ce problème consisterait à utiliser les paramètres de CTI déterminés pour l'étoile la plus proche de l'objet traité. Cette stratégie évoluera probablement.

En plus de ces perfectionnements, il existe des possibilités d'exploitation de la méthode développée dans ce Chapitre qui transcendent leur application aux données du satellite Gaia. Un exemple est l'analyse d'observations ayant un bas rapport signal sur bruit : l'image d'une telle galaxie pourrait être transformée dans l'espace de Radon, et y être analysée. Un exemple peut être vu sur la Figure 6.18, où une galaxie a été simulée avec un signal/bruit égal à 1 et 0.2 (dans le pixel de signal maximum). Nous présentons son signal dans l'espace de Radon.

Cet exemple montre que même dans le cas d'un bruit dans l'observation originale cinq fois plus grand que le signal de l'objet, quand il est incorporé à l'espace de Radon, le signal de l'objet est légèrement amplifié, ce qui est dû au fait qu'il est « cohérent ». En principe, ceci pourrait être appliqué à des études de *weak-lensing*, où ce que l'on cherche justement est d'analyser les corrélations entre les angles de positionnement de diverses galaxies dispersées dans une certaine direction du ciel.

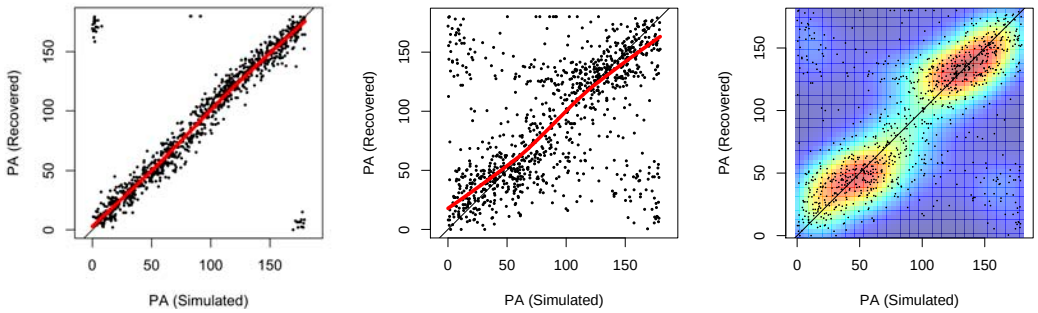
En utilisant de meilleurs traitements du signal de l'objet dans l'espace de Radon, il est peut-être même possible d'appliquer les méthodes d'évaluation et d'ajustement de paramètres commentés dans ce Chapitre pour analyser les profils de galaxies dans des conditions signal/bruit très bas. Sur la Figure 6.19, nous présentons un exemple avec des résultats pour l'angle de position de mille galaxies simulées avec un signal/bruit égal à 1 et 0.2.



(a)

(b)

FIGURE 6.18 – Exemple d'application de l'analyse de l'angle de position dans l'espace de Radon pour une galaxie simulée avec PA=45° et e=0.6. En (a) pour un cas avec S/B∼1. En (b) pour S/B∼0.2.



(a) (b) (c)

FIGURE 6.19 – Résultats pour des simulations de 1000 galaxies avec e=0.6 avec différents angles de position et les angles obtenus par l'analyse du sinogramme filtré. En (a) pour S/B∼ 1. En (b) et (c) pour S/B∼ 0.2, avec (c) présentant une carte de densité en fond.

6.6 Conclusions

Dans ce Chapitre nous avons présenté les profils de brillance les plus utilisés pour la décomposition structurelle de galaxies permettant l'estimation des paramètres morphologiques des principaux composants de ces objets : leur bulbe et leur disque. Nous avons d'abord discuté les profils analytiques et ensuite la méthode d'ajustement de ces profils dans l'espace de Radon.

Nous avons montré, pas à pas, la construction d'un modèle direct d'observation par le satellite Gaia, qui est construit depuis l'échantillonnage des profils et de transformations géométriques jusqu'à l'échantillonnage des fenêtres qui seront observées dans les différents CCDs du plan focal du satellite. Ce modèle peut être adapté de manière simple à tous profils de brillance.

Nous avons présenté une méthode d'optimisation de paramètres aux données d'observation, qui est basée sur trois étapes : évaluation initiale des paramètres, optimisation globale avec des Algorithmes Génétiques et optimisation locale avec l'algorithme BFGS. Nous avons présenté la stratégie que nous avons développée pour la détermination de paramètres initiaux appliquée à l'angle de position de l'objet, et nous avons montré des tests réalisés sur 1000 galaxies simulées dans une région et sur 1000 galaxies simulées dans des régions aléatoires dans le ciel – les résultats ont démontré que même si Gaia n'observe pas un objet sous tous les angles de visée, statistiquement la plus grande partie des objets auront une estimation correcte de leur l'angle de position même dans des conditions de rapport signal sur bruit bas. Les Algorithmes Génétiques et les BFGS ont brièvement été présentés.

Nous avons alors appliqué notre méthode mise en œuvre en R, à des simulations de fenêtres du type Sky-Mapper de 1600 galaxies dont les paramètres sélectionnés de façon aléatoire ont balayé un intervalle réaliste de conditions (les simulations ont été réalisées sur une coordonnée donnée du ciel : $(l, b) = (160°, 50°)$. Les résultats obtenus par notre méthode montrent que l'estimation des paramètres du bulbe présentent une valeur médiane de l'erreur de $\text{med}(Ib) = 11 \pm 53\%$ et $\text{med}(rb) = -9 \pm 36\%$, et du disque $\text{med}(Id) = -1 \pm 7\%$ et $\text{med}(rd) = 1 \pm 4\%$ – la moins bonne précision sur les paramètres du bulbe est probablement due au fait que ce composant structurel a été mal échantillonné dans les fenêtres du type Sky-Mapper.

Finalement, nous avons présenté des possibilités de perfectionnement de la méthode (tels que l'addition des fenêtres *Astro-Field*, ce qui devra permettre un meilleur échantillonnage du bulbe) et une possibilité d'application à des données en plus de celles qui seront obtenues par ce satellite, ce qui permettrait peut-être de profiter scientifiquement d'observations à très bas rapport signal sur bruit.

Les résultats obtenus dans ce Chapitre démontrent qu'à partir des données de la mission Gaia, il sera possible d'obtenir des informations sur les composants de galaxies angulairement petites (rayons de quelque centaines de mas, ou au maximum 2"), sans avoir recours à des reconstructions détaillées d'images 2D. Ceci permettra la réalisation de décompositions en bulbe / disque pour des centaines de milliers, voire des millions, d'objets extragalactiques dont la structure ne pourra être observée que par des missions spatiales ou par des télescopes terrestres munis d'optique adaptative.

Conclusions

En commençant cette thèse il existait un doute s'il était possible faire de la science avec les données des objets étendus qui seront observés par la mission spatiale Gaia. Après l'avoir terminée, nous pouvons dire que oui, ce sera possible.

De manière générale, c'est le résultat astronomique le plus important de ce travail, car cela veut dire qu'il sera possible d'étendre un peu plus l'objectif déjà immense de cette mission. Mais durant l'élaboration de ce doctorat, nous avons aussi progressivement transformé la "possibilité de faire de la science dans le futur" en les codes qui permettront que cette science soit faite. Nous avons mis en œuvre des codes computationnels dans les pipelines de deux *Coordination Units* de Gaia (CU4 et CU5), en suivant autant que possible les recommandations du DPAC et du CNES – et depuis la fin de l'année dernière, ces codes sont maintenus et étendus par un ingénieur (Laurent Galluccio, de l'Observatoire de la Côte d'Azur).

La mission Gaia permettra que pour la première fois l'humanité voie le ciel en entier jusqu'à des magnitudes relativement élevées ($G \sim 20$), avec une haute résolution (~ 180 mas) et nous nous devons profiter de ces données pour faire la meilleure science possible.

Dans la liste ci-dessous, nous présentons les principaux résultats obtenus durant ce travail de thèse :

- **Il est possible de reconstruire des images avec les données de Gaia dans tout le ciel**
  Dans le Chapitre 2, nous présentons la transformée de Radon, qui est une bonne description mathématique du mode d'observation du satellite. En utilisant un toy model, nous avons discuté la méthode d'inversion de cette transformée, appelée *Filtered Backprojection*, et nous avons analysé la sensibilité de cette méthode à la quantité de données disponibles et au bruit dans ces données. Nous avons vu qu'il est possible de reconstruire des images dans tout le ciel avec une couverture de surface minimum égale à $\sim 87\%$, pouvant atteindre 99% dans certaines régions et 98% dans n'importe quelles coordonnées avec $|\beta| > 45°$. Nous avons également montré que la distribution relative entre les angles de passage du satellite est telle que dans toutes les régions du ciel il existera au moins une paire de passages se recoupant à $\sim 90°$ ce qui est la condition nécessaire à la bonne restitution de l'information lors de la reconstruction d'une image 2D.

- **Le catalogue Gaia pourra être étendu à des dizaines de millions de sources ponctuelles, y compris celles qui dépassent la magnitude limite de l'instrument**
  Dans le Chapitre 3, nous avons réalisé une estimation préliminaire du nombre de projections optiques qui seraient présentes dans les observations Gaia. Ceci a été réalisé par comptage et extrapolation dans le catalogue GSC 2.3.2. Ensuite,

nous avons réalisé une analyse plus détaillée, basée sur des simulations faites avec le modèle de la Galaxie de Besançon, qui nous a permis de conclure que $1.6 \times 10^8 \pm 5.1 \times 10^7$ projections optiques pourront être observées dans le ciel entier lorsque l'on considère une différence de magnitude maximale de $\Delta m = 3$ entre la source primaire et la secondaire.

– **Nous avons développé et mis en place une méthode d'analyse d'image qui permet de séparer automatiquement les sources dans les images reconstruites avec peu de fausse détections**
Dans le Chapitre 3, nous avons présenté la méthode dénommée *Educated Image Segregator* ou EIS que nous avons élaborée pour analyser les images reconstruites. Cette méthode d'analyse d'image est capable d'apprendre et de s'adapter de façon automatique aux caractéristiques des images analysées. Les résultats obtenus à partir de reconstructions d'images de simulations réalisées avec l'un des simulateurs officiel de la mission (GIBIS) ont démontré que pour un apprentissage idéal du EIS et donc des résultats plus performants, c'est la méthode *BinOutliers*, qui a détecté le plus grand pourcentage de sources secondaires (pour une primaire de $G = 19$, $\Delta m_{max} \sim 3.0$ mag, et pour $G = 12$, $\Delta m_{max} \sim 4.5$ mag) et le plus petit nombre de fausses détections (jusqu'à des séparations angulaires de ~ 300 mas virtuellement sans détections d'artefacts de reconstruction). Des tests sur des régions non entrainées démontrèrent qu'il est nécessaire de posséder une bibliothèque d'entraînement balayant de façon plus serrée les régions de coordonnées $|\beta| < 45°$. Notre implémentation fait partie du pipeline de la DU18 – *Source Environment Analysis* de la *Coordination Unit* 5 du DPAC.

– **Gaia devra observer environ $\sim 10^7$ galaxies**
Dans le Chapitre 4, nous avons présenté le code JStuff, que nous avons construit en collaboration avec X. Luri de la *Coordination Unit* 2 pour simuler des catalogues de galaxies ainsi que leur propriétés en suivant des fonctions de luminosité et des paramètres cosmologiques. Ce code est connecté au modèle d'Univers de Gaia dans le DU3 de la *Coordination Unit* 2. Des simulations du DPAC qui ont utilisé ce code ont permis une estimation du nombre de galaxies qui pourront potentiellement être observées par le satellite : $\sim 6 \times 10^6$ galaxies jusqu'à $G = 18.5$ et 3×10^7 jusqu'à $G = 20$, avec $\langle z \rangle \sim 0.17$. Ces résultats, sont une limite supérieure car pour les produire nous n'avons pas pris en compte le « filtrage » des données transmises en fonction de la taille angulaire. Nous avons montré, de plus, que cette estimation est compatible avec les résultats de l'extrapolation des densités d'objets dans le *Hubble Medium Deep Field*.

– **Nous avons conçu un pipeline pour analyser les propriétés morphologiques des galaxies non résolues qui seront observées par Gaia**
Dans le Chapitre 1, nous avons présenté la stratégie que nous avons bâtie durant le développement de cette thèse pour analyser les données des galaxies que Gaia

observera. Ce pipeline est basé sur une reconstruction d'image, une mesure de paramètres morphologiques dans l'image reconstruite, une classification morphologique et une analyse de profil dans les données des fenêtres. Sa construction est sous la responsabilité de la DU470 – *Extended Objects* de la *Coordination Unit* 4 du DPAC.

– **Des classifications morphologiques de galaxies pourront être réalisées à partir d'images reconstruites**
Dans le Chapitre 5, nous avons présenté les paramètres caractéristiques que nous avons sélectionnés pour l'analyse des images reconstruites (CASGM20). Des tests ont démontré que les codes que nous avons développé sont capable de mesurer ces paramètres CASGM20 de façon stable même sur des petites images. À partir de l'analyse d'images extraites du *Hubble Deep Field North* et de d'images reconstruites de galaxies simulées avec GIBIS, nous avons vérifié que les trois types morphologiques (elliptiques, spirales, et irrégulières) occupent des zones distinctes de l'espace CASGM20, bien qu'il existe une superposition significative. Nous avons montré qu'une analyse alternative en composantes principales (PCA) n'est pas capable de séparer simplement les trois types d'objets. Notre implémentation fait partie du pipeline de la DU470 – *Extended Objects* de la *Coordination Unit* 4 du DPAC.
Nous avons présenté la méthode de classification appelée *Support Vector Machines*, et nous avons montré par des tests qu'elle est capable de différencier des populations gaussiennes en atteignant des précisions proches du maximum théorique. Nous avons appliqué ce système sur des paramètres obtenus sur des images de galaxies proches re-échantillonnées avec des tailles entre 81x81 et 11x11 pixels. Les résultats ont montré que dans un scénario dans lequel la reconstruction d'image est parfaite, et dans lequel l'ensemble d'apprentissage est construit sans erreurs, la valeur de la précision dans la classification atteint $85\% \pm 3\%$. Nous avons appliqué ce système sur des images reconstruites avec l'algorithme *ShuffleStack* à partir de simulations GIBIS 7, et les résultats ont montré que sous un régime de reconstruction en-dessous de l'idéal, mais avec un ensemble d'apprentissage sans erreurs, on obtient un taux de succès dans la différenciation entre les types précoce/tardif supérieur à 80%, et entre les trois classes morphologiques principales (elliptiques, spirales et irrégulières) de 79%, 56% et 74%. Finalement, à partir d'images de galaxies extraites du Hubble *Deep Field North* nous avons montré que dans une situation équivalente à un régime de reconstruction d'image idéale, mais avec des erreurs dans l'ensemble d'apprentissage, on obtient des taux de classification entre les types précoce/tardif de $\sim 85\%$, et dans les trois classes morphologiques principales de $\sim 81\%$, $\sim 64\%$ et $\sim 62\%$. L'implémentation finale de cette méthode est sous la responsabilité de la DU470 – *Extended Objects* de la *Coordination Unit* 4 du DPAC.

– **Il sera possible de réaliser des décompositions bulbe/disque directement à partir des fenêtres Gaia**
Dans le Chapitre 6, nous avons présenté comment il est possible de faire une décomposition structurelle des galaxies en bulbe et disque dans l'espace de Radon. Nous avons présenté une méthode d'ajustement des paramètres aux des données observationnelles basée sur un modèle que nous avons bâti pour les observations Gaia, une estimation initiale de paramètres, des Algorithmes Génétiques et un algorithme de Quasi-Newton. Des tests avec un prototype de nos codes d'estimation initiale des paramètres ont démontré qu'il permet la récupération de l'angle de position de l'objet dans toutes les directions de la sphère céleste, même dans des conditions de rapport signal sur bruit bas. Des tests sur des simulations de fenêtres du type *Sky-Mapper* de 1600 galaxies ont démontré que l'estimation des paramètres du bulbe présente une erreur médiane pour le bulbe de $\mathrm{med}(I_b) = 11 \pm 53\%$ et $\mathrm{med}(r_b) = -9 \pm 36\%$, et pour le disque $\mathrm{med}(I_d) = -1 \pm 7\%$ et $\mathrm{med}(r_d) = 1 \pm 4\%$. La faible précision dans l'estimation des paramètres du bulbe est probablement due au fait que ce composant structurel est mal échantillonné dans les fenêtres considérées. Finalement, nous avons présenté les possibilités de perfectionnement de la méthode développée (telle que l'addition de fenêtres *AstroField* qui possèdent une plus grande résolution), et commenté la possibilité de son application sur des données d'autres relevés. L'implémentation finale de cette méthode est sous la responsabilité de la DU470 – *Extended Objects* de la *Coordination Unit* 4 du DPAC.

Considérations finales

En arrivant à la fin d'un doctorat, il est intéressant de voir que son résultat réel n'est pas seulement le document qui a été produit, ni même seulement la recherche qui a été développée, mais aussi la matérialisation de tout un processus de construction de la recherche et de la formation du chercheur que l'a produite.

D'une certaine manière je sens que, bien que le plan de ce travail ait été élaboré dans ma tête depuis les phases initiales du doctorat, ce ne fut cependant qu'au moment d'écrire ce document, en transportant ce qui n'était initialement que des idées et qui sont ensuite devenues des résultats, sur le papier, que j'ai pu m'apercevoir que ce n'était pas là l'acte le plus important, mais que celui-ci était un processus nécessaire afin que l'acte de recherche se formalise et soit matérialisé dans une structure cohérente – et ceci même si le travail effectif de recherche était terminé avant d'être rédigé.

Durant son développement, les idées et même leur mises en œuvre dans des codes computationnels, ont été analysées et testées dans une grande partie au moyen de diagrammes purement graphiques, certains d'entre eux ont été inclus dans les chapitres de cette thèse. Ce ne fut seulement qu'à posteriori qu'elles ont été formalisées, générant le texte et finalement le document ici présenté.

Mais ce document ne correspond qu'à l'un des aspects que j'ai développé ces dernières années durant mon travail de recherche. Il représente sans doute le sujet sur lequel j'ai consacré la plus grande partie de mes efforts, mais qui n'a pas été, en aucune manière, l'unique. En plus du travail décrit ici, visant le futur de l'Astronomie au moyen du perfectionnement de la connaissance scientifique à partir d'une plus grande exploitation des données de la mission Gaia, ces dernières années je me suis aussi consacré aux questions du présent de cette science – dans certains cas, en appliquant même des techniques que j'ai appris justement pour permettre les travaux faits dans cette thèse. Comme dit Picasso : Plus vous vous avez de technique, moins vous aurez besoin de vous soucier d'elle.[19]

J'ai en particulier pu collaborer avec le Dr. Caroline Soubiran, et avec mes Directeurs de thèse Dr. Christine Ducourant et Dr. Ramachrisna Teixeira, sur l'étude d'amas ouverts à la recherche d'une meilleure caractérisation de ces objets au moyen de l'analyse des mouvements propres de leurs composants (Krone-Martins et al, 2010b). Pour la réalisation de cette étude, qui traite de quelques dizaines d'amas dans la région $+11° \leq \alpha \leq +18°$, nous avons utilisé des données du catalogue de mouvements propres précis PM2000 (Ducourant et al, 2006), sur lesquelles j'ai eu l'occasion d'appliquer diverses méthodes d'analyse. Bien qu'ayant choisi des optimisations similaires à celle présentée dans le Chapitre 6 de cette thèse pour l'obtention des résultats de l'article final, j'ai étudié beaucoup d'autres méthodes, en plus d'avoir initié le développement d'une méthode nouvelle basée sur des estimations

19. *The more technique you have, the less you have to worry about it.*

détaillées de densité. Cette étude devra être poursuivie et même étendue dans le futur, visant l'utilisation d'autres données conjointement aux mouvements propres.

J'ai pu aussi collaborer avec mes deux Directeurs de thèse, Dr. Ramachrisna Teixeira et Dr. Christine Ducourant, sur des études astrométriques qui transcendent celles que nous avons développées conjointement dans le contexte de la mission Gaia. En plus du travail sur les amas ouverts, nous avons aussi développé ensemble des recherches sur des associations stellaires et des objets jeunes. Pour cela j'ai développé des simulations numériques d'amas et de la méthode de *traceback* (Ducourant et al, 2008), des stratégies de point de convergence et des estimations d'erreur détaillées par la méthode de *bootstrap* (Teixeira et al, 2008) et des classifications par k-NN (Teixeira et al, 2009).

De plus, dans les derniers mois j'ai commencé à me tourner vers un futur encore plus lointain que celui de Gaia, ayant participé à des discussions sur le projet NEAT – *Nearby Earth Astrometric Telescope*, qui est une nouvelle mission spatiale astrométrique proposée pour la décennie de 2020. Actuellement, l'un des secteurs qui semble avoir une plus grande capacité de contribuer à la croissance de la connaissance humaine est la découverte et l'étude de planètes extra-solaires avec une masse et des rayons identiques ou très proches de ceux de la Terre et qui se trouvent dans la zone habitable d'étoiles telles que notre Soleil. Ceci est passionnant, car actuellement nous ne connaissons qu'un seul lieu de l'univers où la vie existe, et ce serait une avancée de grande valeur, même culturelle, que de découvrir aussi bien d'autres lieux habitables que (pourquoi pas ?) de la vie ailleurs. Le NEAT a comme objectif l'observation des déplacements astrométriques des 200 étoiles F, G et K les plus proches du Système Solaire causés par des planètes telluriques de masse jusqu'à une demi masse terrestre. Cette mission et a été proposée à l'ESA dans la Cosmic Vision 2, en décembre 2010.[20]

Un autre thème vers lequel je me suis tourné, moins lié à la recherche astronomique en elle-même, mais lié à son perfectionnement, fut celui d'aider au développement de l'infrastructure disponible pour le calcul astronomique au Brésil. Spécifiquement, j'ai participé activement (conjointement avec le Dr. Alex Carciofi) à la rédaction d'un projet du INCT-A pour la formation de la communauté astronomique brésilienne à l'utilisation de GPUs, et à des discussions à propos de la modernisation du parc de calcul du Département d'Astronomie du IAG-USP (qui durèrent presque un an). Cette modernisation (réalisée avec l'appui de la FAPESP), en plus de permettre la création d'un Centre de Traitement de Données, permettra l'installation d'un ordinateur avec une capacité de traitement de ∼20 TFlops (plus de 2300 cœus de traitement) apportant des bénéfices à toute la communauté astronomique brésilienne.[21]

De plus, durant cette thèse je n'ai pas (complètement) mis de côté l'un des points que je considère d'importance centrale dans la recherche scientifique, et spécialement

20. Une description du NEAT, ainsi que le résumé du projet qui a été envoyé à l'ESA, peuvent être retrouvés sur `http://neat.obs.ujf-grenoble.fr`.

21. Une estimation présentée dans de Carvalho et al (2010), montre que le nombre de cœurs d'usage exclusif de la communauté astronomique au Brésil n'était que de 423 en 2009 – ce chiffre n'inclut pas utilisation de ressources partagées au moyen de laboratoires nationaux.

au Brésil : la vulgarisation et l'éducation scientifique. J'ai participé aux activités du Club d'Astronomie de São Paulo, depuis l'organisation d'événements pour le public en général, tels que des soirées de poésie et d'astrophotographie, jusqu'à des cours grand public d'Astronomie et à la formation d'astronomes amateurs. J'ai eu également l'occasion de présenter une conférence à la Société Astronomique de Bordeaux et, conjointement avec Christine Ducourant et Caroline Soubiran nous avons rédigé un article pour Le Mensuel de l'Université, collaborant à la vulgarisation autour de la mission Gaia. A Bordeaux, j'ai aussi participé à des discussions sur l'élaboration de la sculpture Ville-satellite, de l'artiste plastique Joseph da Silva, installée en suspension sur l'Esplanade de la Cité Mondiale durant plusieurs mois. J'ai maintenu, de plus, des collaborations avec le projet « Télescopes à l'École », qui encourage l'utilisation de télescopes robotiques dans l'éducation des enfants et des adolescents, pour lequel j'ai aidé à l'élaboration de la spécification et des tests du logiciel PInE.[22] Finalement durant l'Année Mondiale de l'Astronomie, en 2009, j'ai pu collaborer avec le comité de pilotage du nœud national Brésilien à divers moments, mais principalement en organisant l'évènement « 100 Heures d'Astronomie », en écrivant quelques textes pour le blog *Cosmic Diary* et en rédigeant un petit chapitre du livre *Postcards from the Edge of the Universe* (Cosmic Diary et al, 2010).

Ce doctorat a permis que toutes ces activités soient développées avec une bien meilleure qualité, car il exigea une maturation de concepts principalement liés à l'Astronomie, mais aussi à d'autres disciplines (depuis la gestion de projets jusqu'aux sciences du calcul). Je souhaite aux candidats au doctorat qu'avec l'évolution de leurs projets et la rédactions de leurs thèses, ils apprennent autant que ce que j'ai pu apprendre durant ces quatre dernières années. Bien que j'ai pu proposer des réponses à certaines questions que nous avons abordées au début de la thèse il existe toute une infinité de problèmes astronomiques qui continuent d'être sans réponse. Comme disait le slogan mondial de l'Année Mondiale de l'Astronomie 2009 :

The Universe : Yours to Discover.[23]

22. Traitement d'Images à l'École : `http://www.telescopiosnaescola.pro.br/pine/PInE.html`
23. Ou dans la version du AMA09-France, *L'Univers, découvrez ses mysteres.*

Annexe A

Module inverse

A.1 Définition

L'opération de « module inverse »est définie comme :

Définition 19 *Soient $a \in \mathbb{R}$, $b \in \mathbb{R}^+$ et l le quotient de la division entière de a par b (de manière à ce que $l \in \mathbb{Z}^+$), nous définissons l'opération de module inverse de a sur b, écrite de la manière suivante $a(modinv b) = c$:*

$$c = \begin{cases} (l+1)b - |a| & si\ l\ impair \\ |a| - lb & si\ l\ pair \end{cases} \tag{A.1}$$

Pour des fins pratiques, cette opération ressemble à l'opération module, au cas où cette dernière serait définie pour $\mathbb{R}$. La principale différence conceptuelle est que dans l'opération module inverse, une fois que la valeur b est atteinte, on commence une « comptage »à rebours, tandis que dans l'opération module le comptage recommence à partir de 0. La Figure A.1 donne une exemple de la différence entre les deux opérations.

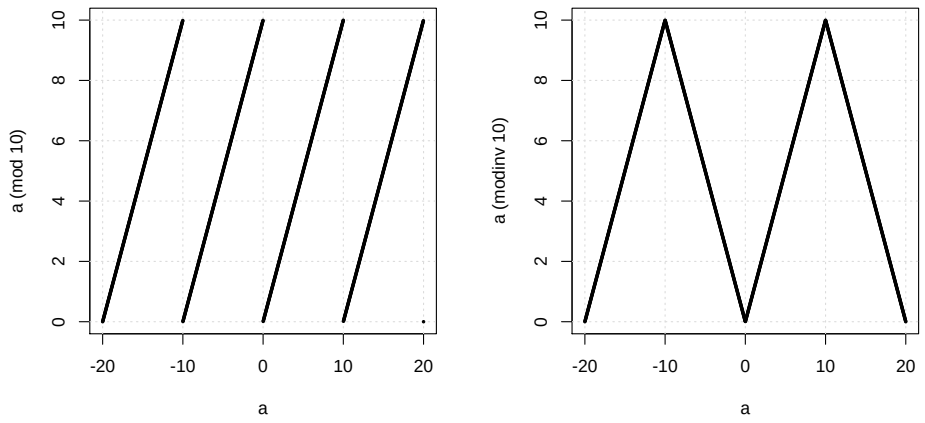


Figure A.1 – Exemple de « module »et « module inverse »de a sur 10, appliqué à des valeurs réelles sur l'intervalle $[-20; 20]$.

Annexe B

Amas ouverts

Comme nous l'avons décrit dans les considérations finales de ce travail de thèse, dans les dernières années j'ai développé un travail en collaboration avec Caroline Soubiran, Christine Ducourant et Ramachrisna Teixeira qui concerne l'étude d'amas ouverts comme prélude à l'exploitation des futures données du satellite Gaia. Notre objectif étant de mieux caractériser ces objets au moyen d'une analyse des mouvements propres de leurs étoiles membres (Krone-Martins et al, 2010b).

Les amas ouverts sont des outils extrêmement intéressants pour étudier notre Galaxie et l'Astrophysique stellaire, quand le problème de la sélection de ses membres est résolu. En effet, une des questions les plus basiques lors de l'étude des amas est la ségrégation des étoiles qui les définissent.

Pour cela il est nécessaire de réaliser l'hypothèse qu'un amas est un groupe d'étoiles localisés dans une région de l'espace relativement petite. De plus, que ses membres possèdent une vitesse spatiale commune qui, à cause des erreurs de mesures et d'une petite dispersion intrinsèque, apparaît comme une dispersion de valeurs autours d'une moyenne. En raison de la faible extension spatiale de la majorité des amas ouverts, la projection de ce critère sur la sphère céleste se traduit par des vecteurs de mouvements propres similaires pour les étoiles membres.

Dans l'étude faite par Krone-Martins et al (2010b), nous adoptons la définition décrite précédemment pour étudier plusieurs dizaines d'amas ouverts de la région $+11° \leq \alpha \leq +18°$ en utilisant les données du catalogue de mouvements propres PM2000 (Ducourant et al, 2006). Nous avons ici adopté le formalisme des PDFs (probability distribution functions) appliqué aux mouvements propres et nous avons optimisé les paramètres de ces PDFs au moyen de techniques similaires à celles présentées au Chapitre 6 de cette thèse, pour l'obtention des résultats présentés dans l'article final.

Cependant, pendant l'évolution de cette collaboration nous avons étudié diverses méthodes pour la détermination des probabilités d'appartenance des étoiles membres. Cette analyse nous a conduit à développer une méthode nouvelle basée sur l'estimation détaillée des densités dans toutes les dimensions de l'espace des données, qui est la base d'études que nous mènerons ultérieurement pour utiliser d'autres dimensions dans l'espace des données, comme les vitesses radiales, les parallaxes, les couleurs, etc., en plus des mouvements propres.

Dans les pages suivantes, nous annexons l'article qui présente ce travail sur les amas ouverts ainsi que ses conclusions.

A&A 516, A3 (2010)
DOI: 10.1051/0004-6361/200913881

Astronomy & Astrophysics

Kinematic parameters and membership probabilities of open clusters in the Bordeaux PM2000 catalogue⋆

A. Krone-Martins[1,2], C. Soubiran[2], C. Ducourant[2,1], R. Teixeira[1,2], and J. F. Le Campion[2]

[1] Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, Rua do Matão, 1226, Cidade Universitária, 05508-900 São Paulo-SP, Brazil
e-mail: algol@astro.iag.usp.br
[2] Observatoire Aquitain des Sciences de l'Univers, Laboratoire d'Astrophysique de Bordeaux, CNRS-UMR 5804, BP 89, 33271 Floirac Cedex, France

Received 15 December 2009 / Accepted 16 February 2010

ABSTRACT

Aims. We derive lists of proper-motions and kinematic membership probabilities for 49 open clusters and possible open clusters in the zone of the Bordeaux PM2000 proper motion catalogue ($+11^\circ \leq \delta \leq +18^\circ$). We test different parametrisations of the proper motion and position distribution functions and select the most successful one. In the light of those results, we analyse some objects individually.
Methods. We differenciate between cluster and field member stars, and assign membership probabilities, by applying a new and fully automated method based on both parametrisations of the proper motion and position distribution functions, and genetic algorithm optimization heuristics associated with a derivative-based hill climbing algorithm for the likelihood optimization.
Results. We present a catalogue comprising kinematic parameters and associated membership probability lists for 49 open clusters and possible open clusters in the Bordeaux PM2000 catalogue region. We note that this is the first determination of proper motions for five open clusters. We confirm the non-existence of two kinematic populations in the region of 15 previously suspected non-existent objects.

Key words. open clusters and associations: general – methods: data analysis – methods: statistical – proper motions

1. Introduction

Once the problem of selecting their physical members is resolved, open clusters are widely respected to be a most valuable tool for undertaking studies of our Galaxy and stellar astrophysics. These objects have been used, for example, to determine the spiral structure of the Galaxy and investigate star formation and evolution processes. They are particularly important as tracers of the dynamics (Frinchaboy & Majewski 2008) and the chemical evolution of our Galaxy's disk (Friel 1995). In the advent of high precision astrophysical surveys such as Gaia (Perryman et al. 2001), their contribution to astrophysical studies should become increasingly important.

However, to establish a coherent understanding of our Galaxy, one needs to use a significant number of open clusters with consistently measured astrophysical parameters, and there have been numerous efforts in this direction (Kharchenko et al. 2005; Bragaglia & Tosi 2006; Frinchaboy & Majewski 2008). However, as noted by Frinchaboy & Majewski (2008), the inaccuracy of current membership determinations poses difficulties in conducting these studies on a large scale.

The main advantage of using open clusters in these studies is that once the complex problem of stellar membership is resolved one can derive their main physical parameters such as distance, age, and metallicity. This membership determination is traditionally performed using the stellar kinematics, but in principle one could use a multidimensional space, using for example, spatial and kinematic information, as in Zhao et al. (2006), or CMD-isochrone information, as in Kharchenko et al. (2005). Nonetheless, we argue that, when analysing these objects, one should rely more on kinematics and as little as possible on a single age CMD-based model analysis as in the aforementioned study. This is because, as it has been reported in the literature, the star formation in some open clusters could be non-coeval, as seems to be true in NGC 3603 (Eisenhauer et al. 1998) and 14 other open clusters (Strobel 1992). An indication of an abundance spread has been reported for one of these objects (Frinchaboy et al. 2008).

Based on these results, we undertook a purely kinematic determination of open cluster membership probabilities for objects located in the zone covered by the high precision proper motion catalogue PM2000 (Ducourant et al. 2006). We used a fully automated optimization method and a set of modified parametrisations for the probability distribution functions based on Zhao et al. (1990, 2006).

This paper is organised as follows. In Sect. 2, we describe the data we used and its selection process. In Sect. 3, we present the methodology and algorithms chosen to obtain the membership lists and the cluster kinematic parameters. In Sect. 4, we describe the validation of the method. In Sect. 5 we present our results, and our comment on some individual objects. Finally, in Sect. 6 we present the conclusions of our current study.

⋆ Full Table 5 is also available in electronic form at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsweb.u-strasbg.fr/cgi-bin/qcat?J/A+A/516/A3

2. Data

2.1. The PM2000 catalogue

The PM2000 catalogue (Ducourant et al. 2006) is a proper motion catalogue that comprises about 2.6 million stars in the declination zone $+11^\circ \leq \delta \leq +18^\circ$ and contains positions and proper motions on the ICRS (International Celestial Reference System), as well as meridian magnitudes V_M. It was derived from the compilation of systematic drift-scan observations in the Bordeaux Carte du Ciel Zone with the Bordeaux automated meridian circle (Viateau et al. 1999) carried out over four years, the reduction of 512 *Carte du Ciel* plates (epoch $t \approx 1900$) of the Bordeaux zone (Rapaport et al. 2001) scanned at the APM Cambridge, and the catalogues AC2000.2 ($t \approx 1907$), USNO-A2.0 ($t \approx 1950$) and the unpublished USNO Yellow Sky (YS3, $t \approx 1978$). The positional precision ranges from 50 to 70 mas, while the proper-motion precision varies from 1.5 to 6 mas yr^{-1}, depending on the magnitude. All the data was analysed using a global iterative astrometric reduction (Teixeira et al. 1992; Benevides-Soares & Teixeira 1992; Ducourant & Rapaport 1991).

The catalogue is complete to $V_M = 15.4$, with a limiting magnitude of $V_M = 16.2$, and typical error of 0.03 mag ($9.5 \leq V_M \leq 13.5$). In addition, a cross identification between all sources in the PM2000 and 2MASS (Cutri et al. 2003) was performed, so the PM2000 catalogue also includes 2MASS photometry information for its objects.

2.2. The clusters sample

The starting point of our analysis is a list of 49 open clusters inside the PM2000 declination zone found in the D07 catalogue (Dias et al. 2002a). We performed a visual inspection of all these clusters by using the Aladin Sky Atlas (Bonnarel et al. 2000) to verify their coordinates. During this visual check, we noticed that the clusters Berkeley 29, 43, 45, and 47 needed to be slightly recentered about 3′25″, 3′39″, 1′13″, and 1′27″, respectively.

The objects were then separated into four different classes: *A-reference*, *B-known*, *C-known without proper-motion determination*, and *D-others* (including doubtful objects). An object was classified as *D* and not *C* when one or more of the following conditions applied:

1. no entry in the WEBDA database;
2. classification as *not found*, *dubious*, *no cluster*, or *non-existent NGC* in the D07 catalogue;
3. existence of some study that excludes the physical clustering of its stars (such as for NGC 1807).

For each cluster, we extracted from the PM2000 catalogue all the stars inside a circular area around the apparent cluster centre. We multiplied the cluster diameter by 1.5, to include most of the cluster members in the extracted zones. Two arcminutes were then added to ensure that even for a small cluster, located in a poorly populated region of the Galaxy, a sufficient number of field stars would be available in the extracted zone to enable us to discriminate between the cluster and the field (as in the case of NGC 7772). We visually inspected the DSS images of all the extracted zones, to ensure that the objects were inside the extracted regions if they were clearly identified. We note that this extended region could add some noise (in the form of field stars) to a small object located in a densely populated region. All the extractions were performed by using automated PERL scripts and CDS's vizquery tool. Table 1 shows the input list of clusters that we used, as well as the class that we assigned to each object.

The data used to obtain the cluster kinematical parameters were selected from the extracted data by rejecting on the basis of proper motion errors ($\epsilon_\mu \leq 6.0$ mas yr^{-1}) and the total proper motion ($|\mu| < 30.0$ mas yr^{-1}). This was justified since high μ field stars cause the proper motion distribution to be flattened, as previously noted in the literature (Balaguer-Núñez et al. 2004b, and references therein). Nonetheless, we computed membership probabilities for all the stars in the extracted zones.

3. Methods

3.1. Mathematical model

The traditional way of conducting membership assignment originates in the seminal works of Vasilevskis et al. (1958) and Sanders (1971). Those two works provided the basic ideas for developing parametric membership analysis. The derived mathematical model is based on the assumptions that there are two distinct kinematic populations in the observational field of the cluster and that the proper motion distributions of those two populations can be parametrised by bivariate Gaussians.

In these studies, an elliptical function was used to describe the field population's proper motion, while a circular one was used for the cluster. Nonetheless, if either an external gravitational influence or tidal effects were to act on different parts of the cluster, the cluster's proper motion distribution could be significantly affected, causing it to deviate from a circular function. For this reason, we tested four parametrisations of the probability distribution functions (PDF) in this study.

We adopted three variations of the general form of the PDF established by Zhao & He (1990) to take account of the observational errors in each individual point. The proper-motion dispersion parameters obtained from these PDFs are therefore the cluster and field intrinsic ones, which are independent of the observational errors in the proper motions. The first variation is a circular distribution, which is the most accurate representation of the PDF for a non-disturbed object with symmetric observation errors. The second variation is elliptical (allowing for the correlation coefficient) and the last a PDF in which we assume that the intrinsic dispersion cannot be observed because of the size of the errors (as adopted by Balaguer-Núñez et al. 2004b). We also tested a fourth PDF in which the (α, δ) positions of the individual stars are taken into account, as in Zhao et al. (2006). Nonetheless, unlike this study, which used the radial distance of the star to the centre of the cluster, we directly used the stars' coordinates and considered the cluster centre as a free parameter.

The adopted PDFs can therefore be written as a mixture of proper-motion (Φ) and position (Ψ) PDFs

$$\Phi_t = \Phi_c \Psi_c + \Phi_f \Psi_f, \tag{1}$$

where c and f correspond to cluster and field parameters.

The proper motion PDFs, Φ_c and Φ_f, are assumed to be Gaussians of the form

$$\begin{aligned}\Phi(\mu_{\alpha,i}, \mu_{\delta,i}) = & \frac{1}{2\pi(1-\rho^2)^{1/2}(\sigma^2_{\mu_\alpha} + \epsilon^2_{\mu_{\alpha,i}})^{1/2}(\sigma^2_{\mu_\delta} + \epsilon^2_{\mu_{\delta,i}})^{1/2}} \\ & \times \exp\left\{ -\frac{1}{2(1-\rho^2)}\left[\frac{(\mu_{\alpha,i}-\mu_\alpha)^2}{\sigma^2_{\mu_\alpha} + \epsilon^2_{\mu_{\alpha,i}}} + \frac{(\mu_{\delta,i}-\mu_\delta)^2}{\sigma^2_{\mu_\delta} + \epsilon^2_{\mu_{\delta,i}}} \right.\right. \\ & \left.\left. - \frac{2\rho(\mu_{\alpha,i}-\mu_\alpha)(\mu_{\delta,i}-\mu_\delta)}{(\sigma^2_{\mu_\alpha} + \epsilon^2_{\mu_{\alpha,i}})^{1/2}(\sigma^2_{\mu_\delta} + \epsilon^2_{\mu_{\delta,i}})^{1/2}} \right] \right\}. \end{aligned} \tag{2}$$

While the Ψ_c and Ψ_f functions depend on whether we take into account the position of the stars or not. We refer to as

Table 1. Our cluster sample sorted by class type.

Cluster	IAU number	α_{J2000}	δ_{J2000}	$l(^\circ)$	$b(^\circ)$	Diam ($'$)	Class	Remarks
NGC 2682	C 0847+120	08:51:18	+11:48:00	215.696	31.896	25.0	*A*	
NGC 1663	C 0445+130	04:48:58	+13:08:54	185.845	–19.735	12.0	*B*	pocr
NGC 1817	C 0509+166	05:12:15	+16:41:24	186.156	–13.096	16.0	*B*	
NGC 2169	C 0605+139	06:08:24	+13:57:54	195.608	–2.935	5.0	*B*	
NGC 2194	C 0611+128	06:13:45	+12:48:24	197.250	–2.350	9.0	*B*	
Berkeley 29*	C 0650+169	06:53:04	+16:55:41	197.948	7.980	6.0	*B*	
NGC 2304	C 0652+180	06:55:11	+17:59:18	197.207	8.897	3.0	*B*	
NGC 2355	C 0714+138	07:16:59	+13:45:00	203.390	11.803	7.0	*B*	
NGC 2395	C 0724+136	07:27:12	+13:36:30	204.605	13.988	14.0	*B*	
Chupina 1		08:50:07	+11:56:42	215.399	31.694	5.0	*B*	
Chupina 2		08:50:30	+12:17:42	215.070	31.925	2.1	*B*	
Chupina 3		08:51:27	+11:23:42	216.147	31.759	2.9	*B*	
Chupina 4		08:52:00	+12:22:42	215.159	32.293	5.0	*B*	
Chupina 5		08:53:01	+11:51:42	215.837	32.304	4.5	*B*	
Berkeley 82	C 1909+129	19:11:20	+13:07:06	46.853	1.624	2.0	*B*	
Roslund 1	C 1942+174	19:45:00	+17:31:00	54.600	–3.400	3.0	*B*	
NGC 7036		21:10:02	+15:31:00	64.544	–21.443	4.0	*B*	pocr
Skiff J0614+129		06:14:48	+12:52:24	197.314	–2.098	7.0	*C*	
Ivanov 2		06:15:53	+14:16:00	196.213	–1.198	2.0	*C*	
Berkeley 43*	C 1913+111	19:15:32	+11:16:31	45.696	–0.140	5.0	*C*	
Berkeley 45*	C 1916+156	19:19:07	+15:43:07	50.033	1.146	2.0	*C*	
Alessi 57		19:20:54	+15:40:36	50.197	0.765	2.5	*C*	
Berkeley 47*	C 1926+173	19:28:30	+17:21:52	52.546	–0.039	3.0	*C*	
King 26	C 1926+147	19:29:00	+14:52:00	50.410	–1.339	2.0	*C*	
Dias 8		19:52:07	+11:37:54	50.335	–7.826	2.3	*C*	
NGC 7772	C 2349+159	23:51:46	+16:14:48	102.739	–44.273	3.0	*C*	pocr
NGC 1807	C 0507+164	05:10:43	+16:31:18	186.088	–13.495	15.0	*D*	(c.3)
DolDzim 2		05:23:54	+11:28:00	192.240	–13.560	10.0	*D*	(c.2)
Teutsch 11		06:25:24	+13:51:59	197.654	0.649	2.3	*D*	(c.1)
Teutsch 12		06:25:40	+13:36:25	197.914	0.585	4.2	*D*	(c.1)
NGC 2224		06:27:32	+12:39:20	198.968	0.544	20.0	*D*	(c.2)
NGC 2234		06:29:29	+16:43:08	195.584	2.846	35.0	*D*	(c.2)
NGC 2265		06:41:41	+11:54:18	201.225	3.268	9.0	*D*	(c.2)
Dolidze 26	C 0727+120	07:30:06	+11:54:00	206.508	13.901	23.0	*D*	(c.2)
NGC 2678		08:50:02	+11:20:18	216.035	31.421	10.0	*D*	(c.1)
DolDzim 7		17:10:36	+15:32:00	36.294	29.166	5.0	*D*	(c.2)
NGC 6525		18:02:06	+11:01:24	37.378	15.890	8.0	*D*	(c.2)
NGC 6738	C 1859+115	19:01:21	+11:36:54	44.398	3.102	15.0	*D*	(c.2)
Riddle 15		19:11:09	+14:50:04	48.357	2.455	0.8	*D*	(c.1)
Juchert 1		19:22:32	+12:40:00	47.727	–1.001	3.2	*D*	(c.1)
Kronberger 13		19:25:15	+13:56:42	49.167	–0.979	1.5	*D*	(c.1)
Dolidze 35	C 1924+115	19:25:24	+11:39:30	47.170	–2.095	7.0	*D*	(c.2)
NGC 6837	C 1951+115	19:53:08	+11:41:54	50.519	–8.009	3.0	*D*	(c.2)
NGC 6839		19:54:33	+17:56:18	56.114	–5.152	6.0	*D*	(c.2)
NGC 6840		19:55:18	+12:07:36	51.162	–8.253	6.0	*D*	(c.2)
NGC 6843		19:56:06	+12:09:48	51.293	–8.404	5.0	*D*	(c.2)
NGC 6858		20:02:56	+11:15:30	51.360	–10.304	10.0	*D*	(c.2)
NGC 6950		20:41:04	+16:37:06	61.107	–15.198	15.0	*D*	(c.2)
NGC 7084		21:32:33	+17:30:30	69.963	–24.302	16.0	*D*	(c.2)

Notes. The parameter values are taken from the D07 catalogue at Vizier (Dias et al. 2002a), apart from the coordinates of the clusters Berkeley 29, 43, 45, and 47 since they were re-centred by means of a detailed visual examination performed with Aladin (*). The Galactic coordinates presented here were computed by VizieR. Classes: *A-reference*, *B-known*, *C-known without proper-motion* and *D-others*. In the remarks, pocr indicates the possible open cluster remnants while the "c." indicates which condition of the Sect. 2.2 applies. Note: the Chupina clusters are actually subclusters formed on the outskirts of NGC 2682 (Chupina & Vereshchagin 1998).

2D implementations those that do not take into account the positions, and 4D those that do. These parametrisations are given by

$$\Psi_c = \begin{cases} n_c & , \text{2D} \\ \dfrac{1}{1+g^{-1}\exp\left\{-\frac{1}{2(1-\rho^2_{\rm pos})}\left[\frac{(\alpha_i-\alpha)^2}{\sigma^2_\alpha}+\frac{(\delta_i-\delta)^2}{\sigma^2_\delta}-\frac{2\rho_{\rm pos}(\alpha_i-\alpha)(\delta_i-\delta)}{\sigma_\alpha\sigma_\delta}\right]\right\}} & , \text{4D} \end{cases} \quad (3)$$

$$\Psi_f = \begin{cases} (1-n_c) & , \text{2D} \\ \dfrac{1}{1+g\exp^{-1}\left\{-\frac{1}{2(1-\rho^2_{\rm pos})}\left[\frac{(\alpha_i-\alpha)^2}{\sigma^2_\alpha}+\frac{(\delta_i-\delta)^2}{\sigma^2_\delta}-\frac{2\rho_{\rm pos}(\alpha_i-\alpha)(\delta_i-\delta)}{\sigma_\alpha\sigma_\delta}\right]\right\}} & , \text{4D.} \end{cases} \quad (4)$$

In Eq. (2) above, $\mu_{\alpha,i}$, $\mu_{\delta,i}$ represents the proper motion of the ith star, $\epsilon_{\mu_{\alpha,i}}$, $\epsilon_{\mu_{\delta,i}}$ the proper motion errors of the ith star, μ_α, μ_δ the mean proper motions of the cluster and field stars (depending on the index of Φ), σ_{μ_α}, σ_{μ_δ} the intrinsic proper-motion dispersions of the cluster and the field, and ρ represents the correlation coefficients of the cluster and field proper motion distributions. In Eqs. (3) and (4), n_c, $1-n_c$ are the fraction of the cluster and field stars, g is the ratio of the cluster to the field stars, α_i, δ_i are the positions of the ith star, α, δ are the position of the cluster, σ_α,

σ_δ represent the dispersion, and $\rho_{\rm pos}$ the correlation coefficient of the position distribution.

The parametrisations presented above leave us with a vector $\mathbf{\Theta}$ of 11 or 16 parameters to be determined, depending on whether we consider 2D or 4D distributions.

3.2. Model parameter estimation

During the construction of the PM2000 catalogue, a global method was used to solve the star's astrometric parameters by selecting mean epochs such that the catalogue covariance matrices are diagonal, and the parameters obtained can be considered independent and identically distributed (iid). We are therefore allowed to write the likelihood function for a given region with N stars as

$$\mathcal{L} = \prod_{i=1}^{N} \Phi_t\left(\alpha_i, \delta_i, \mu_{\alpha,i}, \mu_{\delta,i}, \mathbf{\Theta}\right). \tag{5}$$

Following the maximum likelihood principle the most probable $\mathbf{\Theta}$ is that for which $\mathcal{L}$ takes its maximum value. A monotone transformation does not affect a maximum, and we can therefore write $\hat{\mathbf{\Theta}} = \arg_{\mathbf{\Theta}} \max \mathcal{L} = \arg_{\mathbf{\Theta}} \max \hat{\mathcal{L}}$, where

$$\hat{\mathcal{L}} = \sum_{i=1}^{N} \log \Phi_t\left(\alpha_i, \delta_i, \mu_{\alpha,i}, \mu_{\delta,i}, \mathbf{\Theta}\right). \tag{6}$$

Usually the $\hat{\mathbf{\Theta}}$ is found to be the solution of a non-linear system of equations constructed from $\partial\hat{\mathcal{L}}/\partial\mathbf{\Theta} = 0$, in which some sort of iterative procedure is used (as in Sanders 1971; Zhao & He 1990; Dias et al. 2006). However, this approach depends on the initial chosen value of each unknown parameter. Therefore, the solution to the problem will converge to the real, physical one, provided that there are no local optima, or the initial value is not far from the physical optimum. We note that the last condition may be cluster dependent, and for many objects, we will indeed have local optima. Thus, one cannot guarantee that the obtained solution corresponds to the physical one.

To help us obtain global optima, we chose to use a strategy based on evolutionary computing to solve the optimization problem, the genetic algorithm (GA). This is an adaptive search heuristic for finding optimal solutions, which has been used several times in astronomy (Hetem & Gregorio-Hetem 2007; Howley et al. 2008). However, the number of generations necessary to attain convergence can be quite large, since GAs are generally slow to converge. On the other hand, these algorithms are usually very good at finding the region of global maximum. In this study, we decided to use an optimiser available in the *R* environment (R Development Core Team 2008) called RGENOUD, which is a mixture of a GA and a derivative-based hill-climbing algorithm. The GA identifies the region of the global maximum, while the quasi-Newton hill-climbing algorithm Broyden-Fletcher-Goldfarb-Shanno (or BFGS) based on the fitness function's derivatives, is used to optimise the best (higher likelihood) solutions and, therefore, find the region's maximum (which may be a global maximum). For a complete description, we refer to Mebane & Sekhon (2007).

In principle we could allow the algorithm to search for a solution in $\mathbb{R}^{11}$ or $\mathbb{R}^{16}$ freely. Nonetheless, we know that, physically, the search region can be constrained for some of the parameters, such that if an optimal point were to exist outside this region, it would not have any physical meaning. For example, if most cluster stars were inside the extracted region, we could assume that the position of the cluster center would never be smaller or greater in value than the positions of the most extreme stars in that region. The minimum and maximum of the cluster and field proper-motion components should never be smaller or greater than the minimum and maximum proper-motion components of the stars themselves. We should also ensure that the correlation coefficient moduli never equals unity, otherwise we could encounter problems related to divisions by zero. We also constrain the cluster's proper motion dispersions to have a maximum of 5 mas yr^{-1}, which is an estimated upper limit to the average errors in the individual proper motions. Finally, we allow the field's proper motion (and cluster's position) dispersions to be almost free, adopting a very high limit of six times the standard deviation of the individual proper motions (and stellar positions).

Hence, given the four $\mathbb{R}^1$ sets $\boldsymbol{M}_{\alpha,\delta,\mu_\alpha,\mu_\delta}$ of all the stellar positions and proper motions in the region, we adopted the following constraining conditions for the unknown parameters:

$$\hat{\mathbf{\Theta}} \text{ is such that } \begin{cases} \min\left(\boldsymbol{M}_{\mu_\alpha}\right) \le \mu_{\alpha,c/f} \le \max\left(\boldsymbol{M}_{\mu_\alpha}\right) \\ \min\left(\boldsymbol{M}_{\mu_\delta}\right) \le \mu_{\delta,c/f} \le \max\left(\boldsymbol{M}_{\mu_\delta}\right) \\ 0.01 \le \sigma_{\mu_{\alpha,c}} \le 5.00 \\ 0.01 \le \sigma_{\mu_{\alpha,f}} \le 6 \cdot \sigma\left(\boldsymbol{M}_{\mu_\alpha}\right) \\ 0.01 \le \sigma_{\mu_{\delta,c}} \le 5.00 \\ 0.01 \le \sigma_{\mu_{\delta,f}} \le 6 \cdot \sigma\left(\boldsymbol{M}_{\mu_\delta}\right) \\ -0.99 \le \rho_{c/f} \le 0.99 \\ \\ 0 \le n_c \le 1, & \text{if 2D} \\ 0.01 \le g \le \text{length}\left(\boldsymbol{M}_\alpha\right), & \text{if 4D} \\ \min\left(\boldsymbol{M}_\alpha\right) \le \alpha \le \max\left(\boldsymbol{M}_\alpha\right) & '' \\ \min\left(\boldsymbol{M}_\delta\right) \le \delta \le \max\left(\boldsymbol{M}_\delta\right) & '' \\ 0.01 \le \sigma_\alpha \le 6 \cdot \sigma\left(\boldsymbol{M}_\alpha\right) & '' \\ 0.01 \le \sigma_\delta \le 6 \cdot \sigma\left(\boldsymbol{M}_\delta\right) & '' \\ -0.99 \le \rho_{\rm pos} \le 0.99 & '' \end{cases} \tag{7}$$

where $\sigma\left(\boldsymbol{M}_{\alpha,\delta,\mu_\alpha,\mu_\delta}\right)$ represents the standard deviation of the respective set of parameters.

When a solution is found, the Hessian matrix ($\mathcal{H}$) is computed for the solution parameters. We use this matrix to estimate the individual parameter's errors in the optimization procedure; this is necessary because for a regular problem (in the sense of Wald 1949), as considered here, when one maximises the log-likelihood function, the covariance matrix on the parameters is closely approximated by the negative of the inverse of the Hessian matrix at the solution point (see Pawitan 2001). Thus, the vector of the variances in the parameters is estimated to be

$$\hat{\mathbf{\Theta}}_{\rm var} \approx \text{diag}\left(-\mathcal{H}^{-1}\right), \tag{8}$$

and we use the square root of the elements of this vector to estimate the fitting procedure errors.

Once the cluster and field PDF's unknowns are found, the membership problem is solved because we are then able to compute p_i, the cluster membership probability of the ith star

$$p_{i(\mu_{\alpha,i},\mu_{\delta,i},\mathbf{\Theta})} = \frac{\Psi_c \Phi_c}{\Psi_c \Phi_c + \Psi_f \Phi_f}. \tag{9}$$

The number of apparent cluster members can be estimated from the mixture proportion n_c in the following way. First, from the list of stars used during the optimization process (the list with the quality cuts explained in Sect. 2.2), a sublist of the $N_{\rm opt} * n_c$ most probable stars (where $N_{\rm opt}$) is created. From this list of most probable stars, the probability of the least probable star $P_{\rm min}$ is

Table 2. Method validation for BN04's NGC 1817 data (units of μ and σ are mas yr^{-1}).

Work	n_c	$\mu_\alpha \cos\delta$	μ_δ	$\sigma_{\mu_\alpha \cos\delta}$	σ_{μ_δ}	ρ
Our	0.252	0.25	−0.94			
method	±0.02	±0.11	±0.07			
		2.33	−4.08	5.11	5.93	−0.04
		±0.24	±0.27	±0.18	±0.20	±0.04
BN04	0.261	0.29	−0.96			
	±0.02	±0.10	±0.07			
		2.29	−4.25	5.69	6.38	−0.08
		±0.02	±0.27	±0.02	±0.14	±0.03

obtained. Finally, the number of members in the extracted region (without any probability cut applied) can be estimated from the number of stars in that region with membership probabilities greater than $P_{\min}$.

4. Validation

We verify the reliability of our work in four steps. We firstly validate the automatic optimization method using very precise data. Secondly, we test the new 4D parametrisation by comparing it with another similar parametrisation and method used before in the literature. Thirdly, we validate the data by comparing our results for a well known cluster with the literature. Finally, we perform a comparison of individual members for one "A" and one "B" class cluster.

4.1. Optimisation method

Very precise proper motion data are available for the open cluster NGC 1817. These data were presented by Balaguer-Núñez et al. (2004b) (hereafter BN04), in which proper motions were determined from 25 plates covering a total time-span of 81 years. The mean error in the data of more than 80% of the stars is $\epsilon_\mu = 1.55$ mas yr^{-1}, while it is 0.97 mas yr^{-1} for 32% of the stars, for objects as faint as $V < 16$.

Since we use a new optimization method in this work, we tested it using these proper motion data. Since we consider the method itself, we used the same parametrisation as in BN04, and set the intrinsic dispersion of the cluster Gaussian PDF to zero. We also applied the same cuts in the data ($|\mu| < 30$ mas yr^{-1}). As can be noticed in Table 2, the agreement between our results is quite remarkable, indicating that the method adopted herein allows us to correctly determine the parameters in a fully automatic manner.

We also performed extensive tests using Monte Carlo randomly generated sets of Gaussian mixtures. In virtually all cases, the parameters of the distributions were recovered correctly.

4.2. 4D parametrisation

To test the 4D parametrisation, we compared our results for a set of open clusters with those obtained using another parametrisation and method. The Stochastic Expectation Maximization (SEM) method (Celeux & Diebolt 1986) can be used to deblend the two components corresponding to the field and cluster, assuming that the data can be described by a mixture of 4D Gaussians. The SEM algorithm is non-informative and iteratively solves the maximum likelihood equations, with a stochastic step, for a multivariate mixture of Gaussian distributions. It has been adopted to deconvolve the thin and thick disk populations from velocity distributions (Soubiran 1993; Soubiran et al. 2003). The algorithm initiates at a position decided at random, assuming that the sample is a mixture of 2 discrete components. We repeat the process 150 times. Most of the time, the algorithm converges to the same solution. When multiple solutions are found, the most frequent one is adopted.

We used SEM to analyse all the regions on our list. Afterward, we compared the solutions with those obtained by the 4D parametrisation described herein. The average compatibility between the results, presented in Fig. 1 is very good, with a clear distribution centred on $(0.25 \pm 0.9; -0.46 \pm 2.17)$ mas yr^{-1} before a simple 3σ rejection to eliminate outliers, and centred on $(0.12 \pm 0.64; -0.13 \pm 1.45)$ mas yr^{-1} after the rejection.

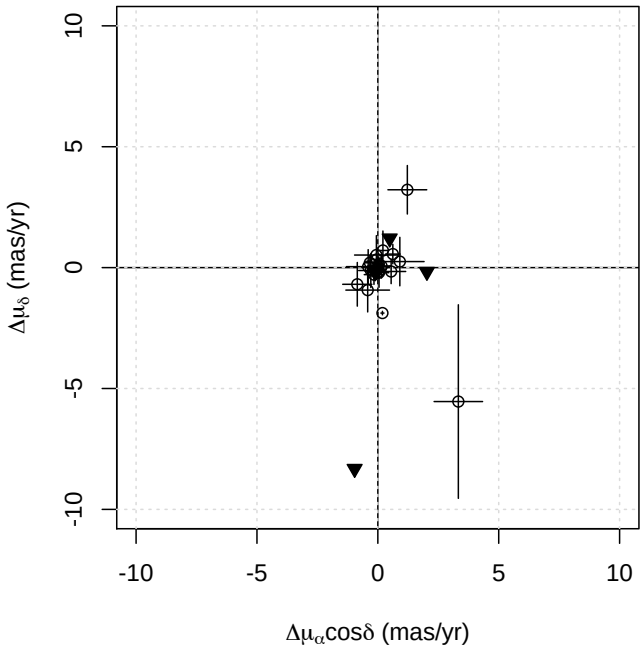


Fig. 1. Comparison between the result of the fit using the method described in this work and SEM. The three filled triangles indicate points with error bars greater than the plot axis.

4.3. Data

To test the data used throughout this work, we applied the method described above to the 4D parametrisation to obtain the kinematic parameters for the open cluster NGC 2682. This object, also called M67, is an old, ~4 Gyr and well known open cluster that has been extensively studied in the literature (Yadav et al. 2008, and references therein).

By comparing the results obtained for this open cluster with SEM and the new method described above, we find that they are in close agreement, generally less than the fitting errors computed for those parameters: $\Delta(\mu_\alpha \cos\delta, \mu_\delta) = (0.06, 0.02)$ mas yr^{-1}, $\Delta(\sigma_{\mu_\alpha \cos\delta}, \sigma_{\mu_\delta}) = (0.20, 0.10)$ mas yr^{-1}, and $\Delta\rho = 0.02$.

The comparison of the kinematic parameters derived from our method with those from previous studies in the literature is shown in Table 3. We observe very good agreement between all the results within the estimated errors, although we note that

Table 3. Kinematic parameters determined for NGC 2682.

Reference	$\mu_\alpha \cos\delta$ (mas yr^{-1})	μ_δ (mas yr^{-1})
This work	-8.32 ± 0.07	-5.65 ± 0.07
Frinchaboy & Majewski (2008)	-7.87 ± 0.61	-5.60 ± 0.59
Dias et al. (2002a)	-8.62 ± 0.28	-6.00 ± 0.28
Kharchenko et al. (2005)	-8.31 ± 0.26	-4.81 ± 0.22

some of the errors in that table are estimated from the standard deviations of member stars, and not from the fitting procedures.

4.4. Individual membership

We verify that the members determined by our method are compatible with those found in previous studies by comparing the members obtained for our "A" class cluster (NGC 2682) and one "B" class cluster (NGC 7036) with previously published membership lists. In performing this analysis, we adopted the parametrisation that does not take into account the cluster's internal dispersion.

For NGC 2682, Yadav et al. (2008) (hereafter Y08) computed the membership probabilities of 2410 stars in its vicinity. Their study was relatively deep in magnitude, to $V \sim 20$, although probably does not probe the entire cluster because of its large angular extension. Unfortunately, the authors did not indicate the number of members found, but a cut at 60% in membership probability was adopted in their work. For this value, 595 stars would be considered as members. When analysing PM2000 data, we obtained 502 members from a total of 1386 stars. We note that our catalog is much brighter than that of Y08, but covers a more extended area.

Between those two membership lists, there are 271 stars in common. Nonetheless, to perform comparison, we need to take into account that only 553 stars are common between our initial catalogues, and that among these common objects, 313 are considered to be members by ourselves, while 402 are Y08 members. This means that ~87% of our members that could be listed as members in Y08 were listed as such, while only ~68% of Y08 members were listed by ourselves.

Since PM2000 proper motions are precise in this particular field, with $\sigma_\mu \leq 2$ mas yr^{-1} for about 50% of them, but in the same magnitude range ($V \leq 16$) Y08 has only about 28% of its stars with such small errors, we attribute this lower value of ~68% to a possible contamination of field stars among Y08 members.

For NGC 7036, Dias et al. (2006, hereafter D06) analysed 69 stars in its vicinity, obtaining 20 members for this possible open cluster remnant. We analysed 91 objects close to NGC 7036, also obtaining 20 members. Adopting the same procedure as applied to NGC 2682, we found that ~65% of our members that could be listed as members in D06 were listed as such, while ~68% of D06 members were among our members.

Taking the results for these two clusters in to consideration, we conclude that there is relatively good agreement between the members obtained by our method and data with those previously published.

5. Results

We applied the optimization method and all variations in the parametrisation described in Sect. 3.1 to the extracted PM2000 data. We thus obtained four solutions for each cluster, and have to determine the most reliable. We first verify whether each solution is in agreement with all others. Next, we visually inspected probability histograms, the vector point diagrams (VPD) and star charts of members and non-members stars, and the DSS images of all clusters, and classified the results as "good", "intermediate", or "poor". To avoid any prejudice, we performed a blind classification: the names of the objects were not written on the diagrams and images while they were being analysed. During this check, we noticed that the probability histogram exhibited the expected field-cluster distribution with a dual peak distribution in most cases. Furthermore, in all cases, the VPD diagrams were indicative of accurate fits. Hence, we decided not to use those two diagrams when assessing the physical reliabilities of the four solutions.

Although this classification of the solutions was subjective, we adopted two clear criteria: some kind of structure or central concentration of stars in the member star chart (in contrast to an almost homogeneous distribution for non-members) and some correlation between a certain magnitude range and the member stars. The solutions were classified as "good" when a clear concentration of member stars was present (even if there was some contamination), and as "intermediate" when the members were not so clear concentrated (or when the concentration was very off-centred), but there was some correlation between the members and a certain magnitude range. Finally, they were classified as "poor" when the members were sparsely distributed without any apparent correlation with any magnitude range.

For NGC 2682, a reference open cluster for which we obtained "good" results, it is clear from the member diagrams that the vast majority of its stars were correctly classified, even if the inspection of its non-member diagrams indicates that some of this cluster's members were probably not classified as such. Moreover, we were able to obtain good results for the 4D parametrisation in this cluster, because it is a well-populated cluster in the magnitude range of PM2000, which allowed a correct determination of its centre and physical dispersion, and its kinematic parameters.

However, of the four parametrisations that we tested, the PDF which considers the internal dispersion of the clusters to be too small to be observed, provided the least number of results classified as "poor". This can be explained physically by the data of the distant clusters being affected by this dispersion in the proper motion far less significantly than the errors in the individual measurements, as noted by Balaguer-Núñez et al. (2004b).

After Gunn et al. (1988), the theoretical prediction for the intrinsic velocity dispersion for a Hyades like open cluster in a state of dynamical equilibrium is 0.23 km s^{-1}, and the observed value based on Hipparcos data for the Hyades cluster is 0.3 km s^{-1} (de Bruijne et al. 2001, and references therein). Mamajek (2010) reported an intrinsic dispersion of 1 km s^{-1} for several nearby objects, and an upper limit of 1.1 km s^{-1} in α Persei was set by Makarov (2006). Using the latest proper motion catalogues, it would be unrealistic to probe values similar to those for any relaxed open cluster that is neither located at very nearby distances nor is highly populated: at only a few hundred parsecs, the dispersion is already at μas yr^{-1} scales. Thus, the use of this additional degree of freedom in the analysis could allow the likelihood function to reach a better fitness value when modelling well the field distribution than when segregating field from cluster stars.

We present the results obtained by the PDF that considers the internal dispersion of the clusters to be too small to be observed, for our entire input list in Table 4. It includes all the fitted

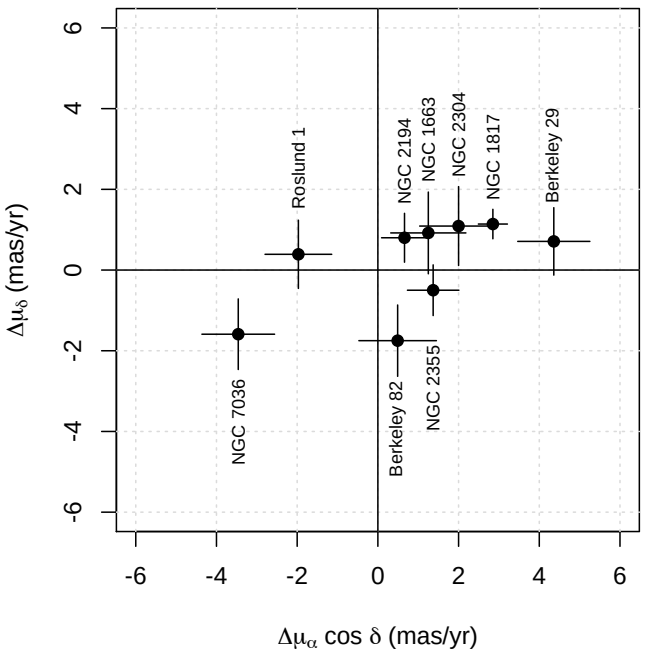


Fig. 2. Comparison between some results of this work and those published in the D06 catalogue. The differences in the open cluster proper motions are in the sense of hereof minus D06 results.

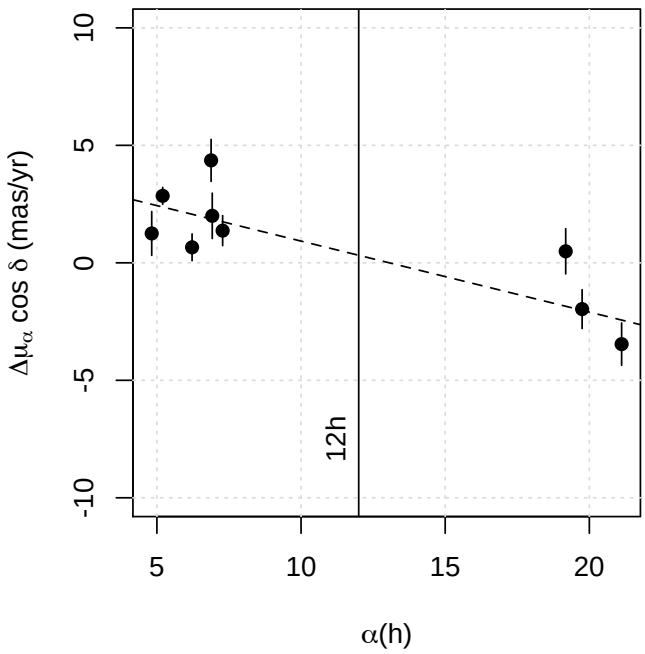


Fig. 3. Systematic effect in the results for the common clusters between this work and the D06 catalogue. The dashed line is a weighted least-squares fit on the points.

parameters from the cluster and the field PDFs, as well as their associated errors. In this table, we also present an estimate of the number of apparent members, and the reference class from Table 1. This table is divided into three sections, corresponding to the quality of the solutions ("good", "intermediate", "poor").

In Table 5 (fully available only through CDS), we present the membership probability information (computed from Eq. (9)) for all the stars in the extracted regions. In addition, we include some columns from the PM2000 catalogue, as well as the object's 2MASS identifier when available.

For some clusters on our list, D06 also obtained kinematic parameters by using UCAC2 data (Zacharias et al. 2004). For those objects, we compare our proper motions with those of D06 work in Fig. 2. We note that the objects tend to have systematically positive values of $\Delta\mu_\alpha \cos\delta$ and $\Delta\mu_\delta$ alike, which means that their proper motions determined by PM2000 data are somewhat higher than those determined by UCAC2 data, by $(0.8, 0.1)$ mas yr^{-1}. The same was found for all the other three parametrisations.

These results were surprising to us, so we performed several tests using the comparison data. During those tests, we found that the proper motion differences $\Delta\mu_\alpha \cos\delta$ were heavily correlated with the cluster's right ascension. This systematic effect can be clearly seen in Fig. 3, in which we present the differences between the results of the parametrisation that does not take into account the cluster's internal dispersion and the D06 catalogue.

We concluded that the origin of this systematic effect was the data itself, since the (α, δ) position of the object was not used in 3/4 of our parametrisations, it was not used in D06 reduction, and the results of all the four parametrisations exhibited exactly the same trends. We therefore directly compared the proper motions of the common stars between PM2000 and UCAC2 (the proper motion catalogue used by D06) inside the cluster regions, and noted a clear dependence on the right ascension, which can be seen in Fig. 4.

We then compared the entire PM2000, UCAC2, UCAC3 (Zacharias et al. 2010), and PPMX (Röser et al. 2008) catalogues (Fig. 5). For UCAC2, there appears to be a periodic systematic variation. However, when comparing each of them with Tycho-2, no similar effect was detected, indicating that this problem affects mostly faint stars, whose first epoch data for deriving proper motions were neither AC2000 nor Cartes du Ciel plates.

For UCAC3, we measure strong offsets in both proper motion components (4.3 ± 1.6 mas yr^{-1} in α and -3.1 ± 1.2 mas yr^{-1} in δ), while relative to PPMX there is a small offset in the right ascension component (1.2 ± 0.75 mas yr^{-1}), but no noticeable effect in declination. Since all these four catalogues use common material and cannot be considered to be completely independent, no firm conclusions could be drawn from a simple analysis and a detailed study is required to address the origin of these effects.

For clusters classified as "non-existent", "doubtful" or "not found" in the D07 catalogue, the chosen parametrisation failed to provide a good fit (our visual classification of the solutions was "poor"). This may indicate that those regions probably are not composed by two different kinematic components, thereby confirming the flag "non-existent" of the D07 catalogue. This is important, since the aforementioned flag is based mainly on a visual inspection performed by Sulentic & Tifft (1979). These objects are the NGCs 2224, 2234, 2265, 6525, 6738[1], 6837, 6839, 6840, 6843, 6858, 6950, and 7084. We note that this "poor" classification cannot be caused by the limiting magnitude of our catalogue since NGC objects should be relatively bright. The cluster Dol Dzim 2 and Dolidze 35, with the flags "not found" and "doubtful" in the D07 catalogue were also classified as presenting a "poor" solution.

Regarding all the other clusters, we obtained solutions classified as "good" for the NGCs 1817, 2169, 2194, 2355, 2682, and Berkeley 82. The clusters for which we obtained solutions considered as "intermediate", and therefore require further analysis, likely by using additional information from photometric,

[1] Confirming the results of Boeche et al. (2003), which was based on Tycho-2 data.

Table 4. The results obtained for all the extracted fields in the PM2000 zone.

Cluster	$\mu_{\alpha,c}$	$\mu_{\delta,c}$	$\mu_{\alpha,f}$	$\mu_{\delta,f}$	$\sigma_{\mu_{\alpha,f}}$	$\sigma_{\mu_{\delta,f}}$	ρ_f	n_c	$n_{\rm memb}$	Class
				Good						
NGC 2682	−8.23 ± 0.05	−5.72 ± 0.05	−4.15 ± 0.30	−5.84 ± 0.25	7.20 ± 0.22	6.29 ± 0.20	−0.17 ± 0.04	0.40 ± 0.02	502	A
Berkeley 82	3.02 ± 0.61	−1.64 ± 0.48	−0.13 ± 0.73	−2.15 ± 1.24	1.00 ± 0.00	3.50 ± 1.47	0.31 ± 0.22	0.33 ± 0.19	16	B
NGC 1817	2.66 ± 0.18	−3.72 ± 0.19	3.56 ± 0.56	−6.70 ± 0.67	7.16 ± 0.49	8.59 ± 0.54	−0.09 ± 0.06	0.72 ± 0.03	836	B
NGC 2169	−2.72 ± 0.40	−4.34 ± 0.41	−3.64 ± 3.58	−11.46 ± 4.61	9.30 ± 2.69	10.75 ± 3.14	−0.44 ± 0.29	0.86 ± 0.06	90	B
NGC 2194	0.07 ± 0.20	−2.69 ± 0.29	−0.31 ± 0.26	−3.60 ± 0.33	3.74 ± 0.20	4.79 ± 0.25	−0.08 ± 0.06	0.21 ± 0.04	65	B
NGC 2304	−1.78 ± 0.66	−5.77 ± 0.68	0.89 ± 1.28	−2.41 ± 1.51	3.54 ± 1.23	4.91 ± 1.31	−0.52 ± 0.19	0.64 ± 0.13	78	B
NGC 2355	−0.39 ± 0.32	−4.78 ± 0.30	1.03 ± 0.98	−7.73 ± 0.91	7.73 ± 0.79	6.75 ± 0.71	−0.22 ± 0.10	0.64 ± 0.06	213	B
				Intermediate						
Chupina 1	−10.54 ± 0.26	−3.52 ± 0.21	−3.59 ± 1.02	−3.17 ± 0.95	5.76 ± 0.71	5.75 ± 0.71	−0.09 ± 0.16	0.30 ± 0.07	22	B
Chupina 2	5.84 ± 0.59	−12.50 ± 0.69	−8.58 ± 1.09	−6.68 ± 1.19	3.61 ± 0.73	3.96 ± 1.05	−0.78 ± 0.10	0.13 ± 0.09	3	B
Chupina 3	−7.32 ± 0.25	−4.37 ± 0.22	3.18 ± 2.48	−6.55 ± 2.35	5.06 ± 2.05	5.90 ± 1.65	−0.38 ± 0.32	0.65 ± 0.12	17	B
Chupina 4	−9.91 ± 0.37	−5.05 ± 0.30	−2.06 ± 1.11	−5.13 ± 1.37	5.73 ± 0.81	7.86 ± 1.04	−0.27 ± 0.15	0.23 ± 0.08	12	B
Chupina 5	−8.68 ± 0.35	−6.90 ± 0.31	−3.12 ± 1.19	−8.28 ± 1.11	5.05 ± 0.81	5.34 ± 0.87	0.07 ± 0.19	0.32 ± 0.11	15	B
NGC 7036	−5.04 ± 0.41	−6.84 ± 0.35	0.06 ± 1.26	−8.15 ± 1.11	8.23 ± 0.98	7.33 ± 0.93	0.56 ± 0.09	0.26 ± 0.08	20	B
Roslund 1	−0.06 ± 0.52	−5.09 ± 0.55	−1.74 ± 1.06	−7.22 ± 1.17	7.18 ± 0.89	8.12 ± 0.99	−0.08 ± 0.12	0.47 ± 0.08	70	B
Berkeley 43	0.30 ± 0.78	−1.81 ± 0.68	−0.36 ± 1.02	−3.52 ± 1.65	4.38 ± 1.02	8.13 ± 1.31	0.08 ± 0.16	0.36 ± 0.15	23	C
Berkeley 45	−1.01 ± 0.96	−9.27 ± 0.94	6.28 ± 1.13	−3.94 ± 1.39	2.44 ± 1.17	3.75 ± 1.10	−0.37 ± 0.33	0.47 ± 0.12	42	C
Berkeley 47	0.90 ± 0.90	−4.14 ± 0.86	1.97 ± 1.93	−4.93 ± 2.46	6.79 ± 1.63	9.65 ± 2.05	0.65 ± 0.14	0.60 ± 0.12	35	C
King 26	3.80 ± 0.58	−5.72 ± 0.47	1.18 ± 1.99	−6.31 ± 1.73	7.68 ± 1.72	6.50 ± 1.58	0.04 ± 0.22	0.45 ± 0.15	21	C
Skiff J0614+129	−0.31 ± 0.29	−2.73 ± 0.37	−0.57 ± 0.44	−4.31 ± 0.63	4.48 ± 0.32	6.56 ± 0.47	−0.05 ± 0.09	0.18 ± 0.05	29	C
NGC 1807	2.06 ± 0.23	−4.56 ± 0.24	3.57 ± 0.68	−7.83 ± 0.75	7.80 ± 0.57	8.35 ± 0.59	−0.08 ± 0.07	0.66 ± 0.03	529	D
NGC 2678	−1.94 ± 0.32	−2.00 ± 0.28	−4.18 ± 0.75	−5.34 ± 0.57	8.29 ± 0.55	6.10 ± 0.43	−0.03 ± 0.09	0.11 ± 0.04	16	D
Teutsch 11	−2.24 ± 0.91	−2.61 ± 0.87	0.30 ± 1.28	−4.03 ± 1.41	5.60 ± 0.97	6.66 ± 1.13	0.34 ± 0.17	0.25 ± 0.14	16	D
				Poor						
Berkeley 29	0.56 ± 0.45	−4.29 ± 0.32	0.12 ± 0.96	−6.46 ± 0.88	7.43 ± 0.79	6.26 ± 0.68	−0.32 ± 0.10	0.53 ± 0.07	137	B
NGC 1663	1.13 ± 0.60	−1.32 ± 0.72	3.97 ± 0.65	−4.94 ± 0.72	6.83 ± 0.51	7.27 ± 0.55	−0.14 ± 0.08	0.23 ± 0.07	49	B
NGC 2395	1.22 ± 0.38	−1.96 ± 0.27	0.01 ± 0.40	−4.32 ± 0.48	5.80 ± 0.31	7.17 ± 0.36	0.01 ± 0.06	0.27 ± 0.05	142	B
Alessi 57	−0.92 ± 0.80	−7.89 ± 0.68	1.11 ± 1.36	−4.85 ± 1.94	5.81 ± 1.22	9.15 ± 1.55	−0.01 ± 0.19	0.30 ± 0.14	14	C
Dias 8	−4.51 ± 0.48	−2.85 ± 0.44	−1.13 ± 1.62	−4.26 ± 1.69	8.29 ± 1.24	9.02 ± 1.32	0.26 ± 0.16	0.40 ± 0.10	30	C
Ivanov 2	−1.02 ± 0.73	0.41 ± 0.89	−0.89 ± 0.98	−3.62 ± 1.33	3.29 ± 0.76	4.56 ± 0.88	−0.46 ± 0.21	0.22 ± 0.18	15	C
NGC 7772	17.50 ± 0.57	−10.00 ± 0.57	4.83 ± 1.85	−8.49 ± 1.63	7.59 ± 1.33	6.63 ± 1.21	0.13 ± 0.23	0.09 ± 0.06	2	C
DolDzim 2	1.92 ± 0.49	−3.25 ± 0.59	3.30 ± 0.92	−5.57 ± 1.13	7.23 ± 0.75	8.79 ± 0.86	−0.05 ± 0.11	0.48 ± 0.07	85	D
DolDzim 7	18.77 ± 0.75	−12.13 ± 0.70	−1.82 ± 1.02	−5.70 ± 1.23	5.79 ± 0.80	7.56 ± 0.94	−0.14 ± 0.15	0.03 ± 0.03	1	D
Dolidze 26	−0.64 ± 0.18	−2.24 ± 0.25	−0.96 ± 0.21	−4.13 ± 0.24	6.17 ± 0.17	6.84 ± 0.19	−0.13 ± 0.03	0.30 ± 0.02	536	D
Dolidze 35	3.87 ± 0.36	0.85 ± 0.37	0.32 ± 0.78	−5.61 ± 0.77	7.11 ± 0.57	6.22 ± 0.56	0.23 ± 0.09	0.18 ± 0.06	28	D
Juchert 1	−0.61 ± 0.69	0.95 ± 0.70	4.88 ± 0.77	2.17 ± 1.50	1.46 ± 0.72	5.75 ± 1.25	0.41 ± 0.17	0.17 ± 0.13	4	D
Kronberger 13	5.79 ± 1.57	−5.39 ± 1.57	−7.12 ± 5.16	4.26 ± 4.65	8.24 ± 3.87	7.52 ± 3.68	0.53 ± 0.43	0.70 ± 0.15	9	D
NGC 2224	0.66 ± 0.22	−0.91 ± 0.16	0.63 ± 0.16	−3.16 ± 0.20	4.24 ± 0.13	5.47 ± 0.15	−0.14 ± 0.03	0.19 ± 0.02	252	D
NGC 2234	0.14 ± 0.07	−3.65 ± 0.07	0.40 ± 0.16	−5.88 ± 0.18	5.48 ± 0.13	6.11 ± 0.14	−0.06 ± 0.03	0.36 ± 0.02	2221	D
NGC 2265	1.60 ± 0.16	−2.27 ± 0.19	1.26 ± 0.31	−3.82 ± 0.36	4.61 ± 0.25	5.24 ± 0.28	−0.12 ± 0.06	0.31 ± 0.04	149	D
NGC 6525	−2.39 ± 0.51	−3.05 ± 0.49	0.00 ± 0.55	−2.63 ± 0.75	5.56 ± 0.45	8.24 ± 0.61	0.06 ± 0.08	0.21 ± 0.06	41	D
NGC 6738	1.51 ± 0.17	−2.69 ± 0.17	−0.36 ± 0.23	−3.93 ± 0.24	6.44 ± 0.19	6.71 ± 0.20	0.22 ± 0.03	0.18 ± 0.02	255	D
NGC 6837	−23.18 ± 2.77	−4.49 ± 2.88	−0.13 ± 0.66	−3.63 ± 0.57	5.60 ± 0.55	4.48 ± 0.50	0.39 ± 0.09	0.02 ± 0.01	5	D

Table 4. continued.

Cluster	$\mu_{\alpha,c}$	$\mu_{\delta,c}$	$\mu_{\alpha,f}$	$\mu_{\delta,f}$	$\sigma_{\mu_{\alpha,f}}$	$\sigma_{\mu_{\delta,f}}$	ρ_f	n_c	$n_{\rm memb}$	Class
NGC 6839	0.11 ± 0.66	−4.88 ± 0.57	−0.05 ± 0.57	−6.04 ± 0.54	7.48 ± 0.48	7.12 ± 0.46	0.12 ± 0.06	0.29 ± 0.05	116	D
NGC 6840	1.08 ± 0.61	−2.49 ± 0.58	0.24 ± 0.47	−6.07 ± 0.51	6.20 ± 0.41	6.34 ± 0.40	0.06 ± 0.06	0.24 ± 0.06	91	D
NGC 6843	3.26 ± 0.68	−1.61 ± 0.66	0.70 ± 0.51	−4.69 ± 0.55	5.72 ± 0.44	6.24 ± 0.46	0.13 ± 0.07	0.14 ± 0.07	33	D
NGC 6858	−0.27 ± 0.43	−3.35 ± 0.34	0.46 ± 0.42	−4.43 ± 0.43	7.27 ± 0.33	7.71 ± 0.35	0.14 ± 0.05	0.36 ± 0.04	261	D
NGC 6950	−0.89 ± 0.34	−6.36 ± 0.39	−0.01 ± 0.27	−6.38 ± 0.27	6.44 ± 0.21	6.72 ± 0.22	0.42 ± 0.03	0.09 ± 0.03	80	D
NGC 7084	−5.50 ± 0.51	−12.09 ± 0.52	−0.04 ± 0.47	−9.43 ± 0.43	7.43 ± 0.32	7.31 ± 0.33	0.29 ± 0.05	0.19 ± 0.04	97	D
Riddle 15	2.73 ± 0.90	−8.10 ± 0.88	−2.42 ± 3.10	−2.21 ± 4.75	7.27 ± 2.27	11.87 ± 3.67	0.76 ± 0.18	0.24 ± 0.15	3	D
Teutsch 12	0.81 ± 0.37	−1.98 ± 0.35	−0.60 ± 0.79	−2.89 ± 0.96	4.97 ± 0.63	6.22 ± 0.78	−0.08 ± 0.13	0.31 ± 0.09	29	D

Notes. The *c* subscripts indicate the parameters for the clusters, while the *f* subscripts indicate the fields ones. The correlation coefficient (ρ_f), mixture proportion (n_c) and estimated number of members ($n_{\rm memb}$) are adimensionals, whereas all the other parameters are in mas yr^{-1} units. The "Class" column is the same as in Table 1. The solutions are divided following their visual classifications.

Table 5. Example of the membership probability table for the remnant NGC 7772.

Cluster	PM2000	h	m	α_{J2000} s	°	′	δ_{J2000} ″	σ_α arcsec	σ_δ arcsec	$\mu_\alpha \cos\delta$ mas yr^{-1}	μ_δ mas yr^{-1}	$\sigma_{\mu_\alpha \cos\delta}$ mas yr^{-1}	σ_{μ_δ} mas yr^{-1}	V_M mag	σ_{V_M} mag	2MASS	Prob. %
NGC 7772	664742	23	51	20.0035	16	14	31.344	0.039	0.043	2.5	–2.5	3.9	4.0	15.722	0.152	23512000+1614312	0.0
NGC 7772	664906	23	51	33.7205	16	20	33.280	0.055	0.057	4.4	–7.7	4.3	4.3	15.998	0.015	23513371+1620334	0.3
NGC 7772	664928	23	51	35.4557	16	12	49.718	0.024	0.025	–1.3	–19.3	0.8	0.8	13.660	0.054	23513545+1612494	0.0
NGC 7772	664949	23	51	37.3487	16	18	02.019	0.049	0.043	5.9	–7.7	3.9	3.9	15.543	0.117	23513734+1618020	0.4
NGC 7772	664966	23	51	38.9136	16	12	22.572	0.029	0.030	2.1	3.0	2.1	2.1	10.357	0.050	23513891+1612224	0.0
NGC 7772	664970	23	51	39.1210	16	18	13.397	0.030	0.031	–15.1	–37.1	3.7	3.7	13.266	0.065	23513911+1618132	0.0
NGC 7772	664998	23	51	41.3285	16	10	43.795	0.034	0.035	–6.4	–16.8	3.7	3.7	15.432	0.136	23514131+1610436	0.0
NGC 7772	665026	23	51	42.8851	16	14	06.988	0.023	0.024	11.2	–10.3	0.8	0.8	11.125	0.056	23514289+1614067	0.0
NGC 7772	665044	23	51	44.4787	16	20	17.501	0.039	0.040	4.8	–10.9	3.9	3.9	15.466	0.121	23514448+1620175	0.2
NGC 7772	665048	23	51	44.7867	16	14	43.252	0.024	0.025	15.6	3.4	0.8	0.8	14.376	0.082	23514479+1614431	0.0
NGC 7772	665067	23	51	46.2475	16	14	23.413	0.023	0.024	17.4	–10.1	0.8	0.8	12.578	0.059	23514626+1614231	97.2
NGC 7772	665069	23	51	46.3839	16	14	56.755	0.025	0.026	17.6	–9.9	0.8	0.8	13.514	0.136	23514640+1614564	97.3
NGC 7772	665087	23	51	47.5439	16	12	54.168	0.037	0.041	8.7	–20.7	3.8	3.9	15.753	0.185	23514753+1612539	0.3
NGC 7772	665088	23	51	47.5485	16	15	55.216	0.027	0.028	48.8	–6.0	0.8	0.8	12.748	0.066	23514760+1615549	0.0
NGC 7772	665096	23	51	48.0574	16	14	00.102	0.023	0.024	5.2	–3.8	0.8	0.8	13.723	0.067	23514805+1613598	0.0
NGC 7772	665101	23	51	48.1263	16	15	04.932	0.026	0.027	13.4	–12.3	0.8	0.8	13.436	0.071	23514814+1615045	0.0
NGC 7772	665113	23	51	48.9226	16	11	00.074	0.026	0.027	–2.0	–9.1	0.8	0.8	14.405	0.097	23514892+1610597	0.0
NGC 7772	665147	23	51	52.4475	16	11	54.192	0.035	0.033	39.7	1.5	5.7	5.7	15.326	0.148	23515244+1611540	1.8
NGC 7772	665148	23	51	52.5442	16	11	17.381	0.026	0.027	10.9	–40.4	0.8	0.8	12.733	0.061	23515255+1611165	0.0
NGC 7772	665203	23	51	57.9668	16	17	31.784	0.027	0.028	24.2	–11.9	0.9	0.8	14.148	0.079	23515799+1617313	0.0
NGC 7772	665223	23	51	59.5060	16	19	48.341	0.032	0.035	–9.3	–13.4	3.8	3.8	15.130	0.209	23515950+1619482	0.0
NGC 7772	665236	23	52	00.3693	16	09	44.603	0.029	0.030	3.3	–16.9	2.1	2.1	10.978	0.058	23520036+1609444	0.0
NGC 7772	665266	23	52	03.3627	16	17	41.938	0.026	0.027	1.2	–1.8	0.8	0.8	12.921	0.071	23520336+1617418	0.0
NGC 7772	665271	23	52	03.9432	16	14	13.038	0.027	0.030	–2.1	–3.5	3.7	3.7	15.152	0.141	23520393+1614129	0.0
NGC 7772	665273	23	52	04.3616	16	19	18.151	0.039	0.042	0.1	–7.9	3.9	4.0	15.670	0.212	23520435+1619180	0.0
NGC 7772	665279	23	52	04.7888	16	11	32.168	0.032	0.035	7.9	–0.3	3.7	3.7	15.342	0.096	23520479+1611321	0.1

Notes. The PM2000 and 2MASS columns points to the identifiers of the object in the respective catalogues. The Prob. column indicates the membership probability as calculated from Eq. (9).

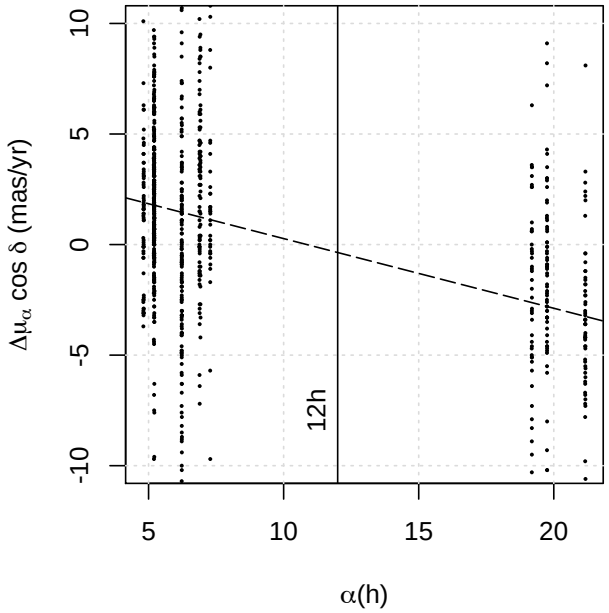


Fig. 4. Differences in proper motions between common PM2000 and UCAC2 stars in the sense PM2000 minus UCAC2. This comparison includes only stars in the regions of the common open clusters reduced by this work and D06. The dashed line is a least squares fit to the points.

radial velocity data and deeper or more precise proper motions, are: Berkeley 43, Berkeley 45, Berkeley 47, the Chupina series, King 26, the NGCs 1807, 2678, 7036, Roslund 1, Skiff J0614+129, and Teutsch 11.

In the case of NGC 2678, we note that the stars classified as members are concentrated in direction of NGC 2682. The brightest stars in this region (those that visually define the cluster) were also not classified as members. Since these two clusters are close to each other, this may indicate that these stars are members of NGC 2682 and that NGC 2678 does not exist.

The remaining objects on our list were classified as "poor". In the case of Berkeley 29, even though we have around 80% of the stars with membership probability greater than 51% in common between our and D06 analyses, the cluster's proper motions in right ascension are not compatible with those published in D06. We notice that both solutions are probably only good fittings of the field distribution, since from the photometric work of Tosi et al. (2004) we can see that the vast majority of this open cluster's stars are at $V > 18$, far beyond the reach of the PM2000 (used here) or UCAC2 (used in D06) catalogues. For NGC 2395, we could not distinguish a concentration of member stars, and the bright stars that define the cluster were not considered as members.

For all the other solutions classified as "poor", we examine with regions in which the difference between the number of field and the cluster stars was very high. As a result, the natural tendency of the likelihood function was to reach a better fitness value for the determination of a field only distribution. Nonetheless, we note that some clusters classified as "poor" can have a good determination of their proper motions, as in the case of NGC 7772, which we comment on the next section.

5.1. Comments on selected objects

Because of the physical interest of some open clusters in this study, we briefly discuss the properties of a small subset of them individually.

5.1.1. NGC 1807

The NGC 1807 open cluster is a concentration of stars close to NGC 1817. The solutions obtained for this object using all the four parametrisations indicate that its most probable members are concentrated in one border of the extracted field, mainly in the direction of NGC 1817. For the 4D parametrisation, the concentration in that part of the star chart is even larger than the other ones. In addition, the fitting parameters obtained for the cluster's PDF are compatible (within the fitting errors) with those obtained for NGC 1817 (as can be seen in Table 4).

We conclude that the existence of a separate cluster from NGC 1817 in the NGC 1807 region is not supported by our analysis of PM2000 kinematical data. This upholds Balaguer-Núñez et al. (2004a,b) assertion, that there is actually only one extended cluster in that region. However, we note that a radial velocity study would be of great value to definitely settle this issue.

5.1.2. NGC 2194

NGC 2194 is an open cluster located in a moderately rich field projected along the Galactic anti-centre direction, which has been studied by Kyeong et al. (2005), Piatti et al. (2003), and Sanner et al. (2000). Only the final of these three aforementioned studies deals with proper motions. Sanner et al. (2000) were the first to analyse this object's proper motion. In their study, proper motions were obtained by using photographic plates from the Bonn Doppelrefraktor (first plate from 14.02.1917) and from CCD observations performed at the 1.23 m telescope at the Calar Alto Observatory (15.10.1998). They determined the cluster's proper motion to be $(\mu_\alpha \cos\delta; \mu_\delta) = (-2.3 \pm 1.6; 0.2 \pm 1.5)$ mas yr^{-1} (the indicated errors are the fitted widths of the PDF).

This cluster was later analysed, automatically, in D06, in which its proper motion was computed from UCAC2 data $(-0.59 \pm 0.53; -3.49 \pm 0.53)$ mas yr^{-1}. As one can promptly see, there is a wide discrepancy between the proper motion values obtained by those two studies, even if compatible to within 3σ. It is even more confusing that the entry in the D07 catalogue, $(-0.31 \pm 0.64; -4.40 \pm 0.64)$ mas yr^{-1}, is once again different[2], even though it is compatible with the one obtained using UCAC2 proper motions.

We found herein the mean proper motion of this object to be more compatible with that of D06, with a value of $(0.07 \pm 0.20; -2.69 \pm 0.29)$ mas yr^{-1}. The almost 3σ discrepancy between the cluster's proper motion results obtained by this work, D06, and Sanner et al. (2000), may be caused by the different materializations of the HIPPARCOS reference system. As for UCAC2, PM2000 is linked to HIPPARCOS as materialized by Tycho-2. In constrast, Sanner et al. (2000) used ACT stars to perform the transformation from plate to celestial coordinates, ACT being linked to HIPPARCOS, but as materialized by Tycho-1.

It is also interesting that the proper motion of this cluster is compatible with that obtained for Skiff J0614+129. We can also identify some of the brightest stars in the region of Skiff J0614+129 (those that define the central concentration in its member star chart) among the list of NGC 2194 members, but a more detailed analysis using photometric and radial velocity data would be required to check whether they are somehow connected.

[2] The same value is quoted in the WEBDA database, and was computed from Tycho-2 data in Dias et al. (2002b).

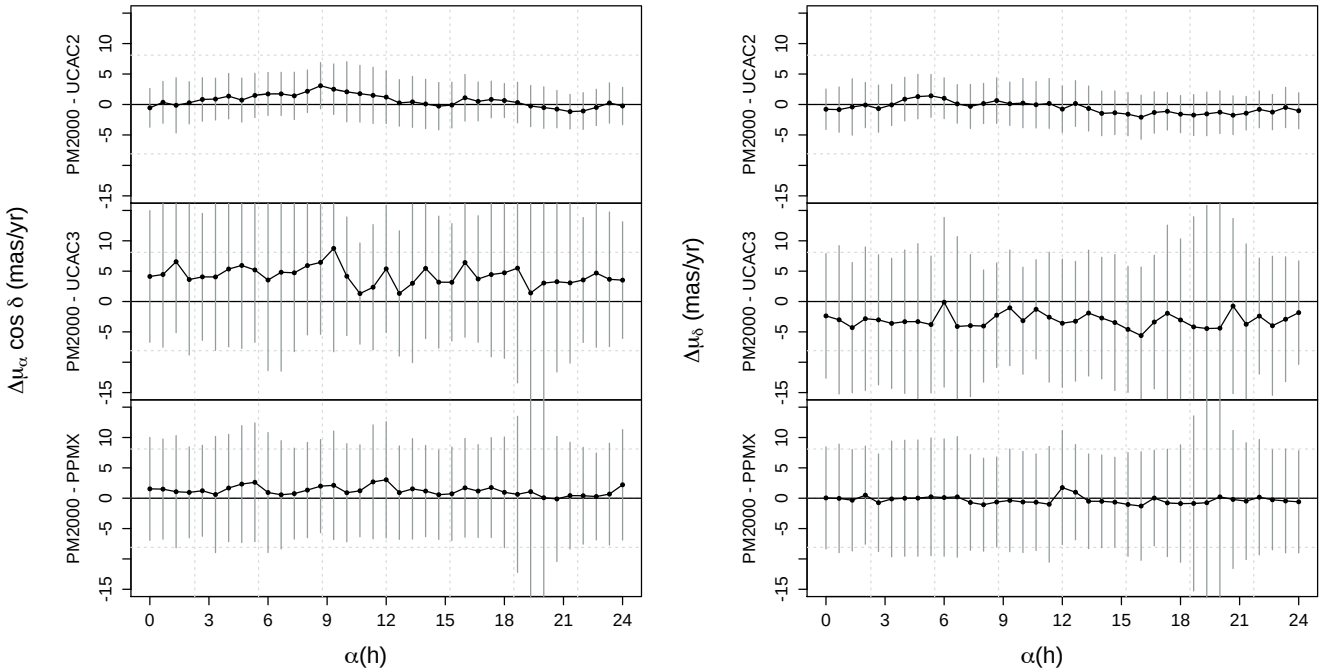


Fig. 5. Average differences in bins of α in the proper motion components of common stars between PM2000 and UCAC2, UCAC3, and PPMX catalogues.

5.1.3. NGC 7772

NGC 7772 was firstly proposed to be an open cluster remnant by Bica et al. (2001), and confirmed to be one by Carraro (2002). In the latter study, NGC 7772's age was estimated to be 1.5 Gyr, and its distance from the Sun to be 1.5 kpc, which assigned the membership to 14 stars. Nonetheless, since that work was purely based on photometric data, Carraro himself warned that his analysis needed to be constrained by radial velocity and proper motion determinations. Here we constrain the membership of the stars in this object by using proper motions.

We note that the fitting of two kinematic populations to this cluster VPD, using the PDF that does not take into account the internal dispersion, does not allow a highly reliable segregation between the remnant and the field, since only two stars were classified as members using the criteria for the number of members presented at the beginning of this section ($n = 2.34 \pm 1.56$). We believe that this to be caused by our choosing the PDF that does not take into account the internal dispersion between the member stars: since this is a remnant, the dispersion could be greater than the individual measurement errors, and we should be able to resolve it.

Since this is a sparsely populated region, we can, however, analyse a plot of the proper motions represented as vectors superimposed directly on the Star Chart, as can be seen in Fig. 6. This graphical representation allows us to observe what is happening from the point of view of the kinematic, spatial, and magnitude distributions. We realise that there exists a tightly packed spatial concentration of similar magnitude stars that share almost the same proper motion vector in this region.

On the basis primarily of their proper motions, we visually chose the four (compatible with n to within $\sim 1\sigma$) most similar stars in the high-concentration region, two of which are the member stars classified by our automatic reduction. We note that the other stars have higher membership probabilities in the 4D solution.

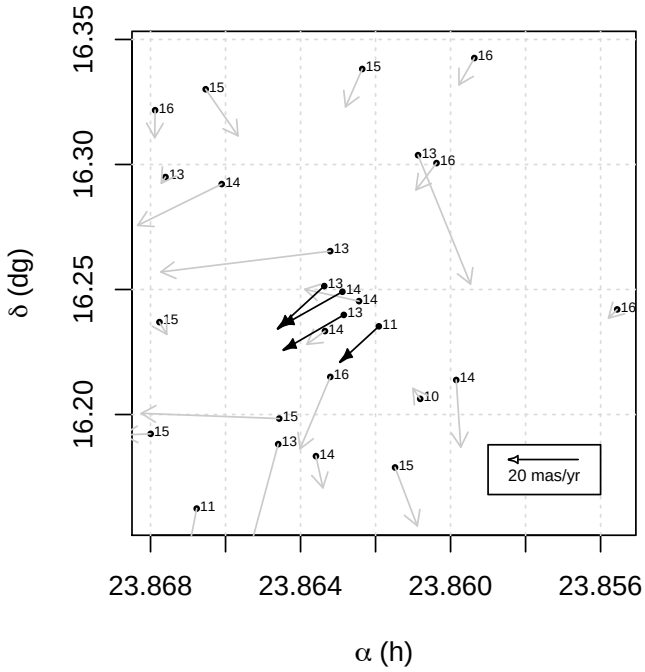


Fig. 6. Star chart of the field around NGC 7772. The non-member stars have their proper motions represented by the grey vectors, and the members by black ones. Meridian V magnitudes are also shown.

Since these are the constituent stars of this object, and whose errors in their individual proper motions are very small ($\langle \mu/e_\mu \rangle$ = 4%), the average proper motion is a very good estimate of this remnant's proper motion. The final average proper motion of the object is therefore $\langle \mu_\alpha \cos\delta \rangle = 14.90 \pm 3.13$, $\langle \mu_\delta \rangle = -10.65 \pm 1.11$ mas yr^{-1}, which is highly compatible with that determined automatically (see Table 4). This is the first ever determination of a proper motion for this object.

5.1.4. Berkeley 82

Berkeley 82 is a sparsely populated open cluster that has yet to be thoroughly studied. This cluster is projected in the direction of a HI supershell (Kim & Koo 2000) that has a kinematically derived distance of 1.4 kpc and a projected surface of 340 × 540 pc^2 (in the *l*, *b* directions). These authors suggested that this ∼5 Myr supershell may be physically related to some of the open clusters in its region: NGC 6738, Berkeley 43, and Berkeley 82. The existence of the first of these three objects was ruled out. Berkeley 43 perhaps is an interesting object without proper motion determination in the literature, but unfortunately our automatic solution for it was classified as "intermediate" probably because of the faintness of this cluster.

However, for Berkeley 82, we were able to assign membership of stars to this object and compute its proper motion, which was measured to be (3.02 ± 0.61; −1.64 ± 0.48) mas yr^{-1}. From D07, a Tycho-2 determination for the proper motion of this object was obtained by Loktin & Beshenov (2003) during a study of the rotation rate of the Galaxy (−0.06 ± 0.99; −4.38 ± 1.78) mas yr^{-1}. Nonetheless, no list has been published with individual membership probabilities, so one cannot explore the connection between the cluster and the shell on a star by star basis. According to WEBDA Berkeley 82 is located at a distance of 870 pc, indicating that this cluster could well lie at one of the borders of the shell (if the cloud size is taken into consideration).

Based on the quoted distance, our proper motion determination amounts to a tangential velocity of ∼13.8 ± 2.4 km s^{-1}. Kim & Koo (2000), determine the expansion velocity of the shell to be ∼15 km s^{-1}, and its central velocity to be 18 km s^{-1}. The similarities between the velocities that they obtained for the HI supershell and that which we obtained for the open cluster are quite remarkable. A radial velocity study would undoubtedly be of great value to determine the space velocity of this object, and finally reach a conclusion about the physical connection between this cluster and the supershell.

5.1.5. NGC 1663

During their photometric analysis, Baume et al. (2003) suggested that NGC 1663 is a ∼2 Gyr possible open cluster remnant. In that study, the authors derived a list of 2 members and 4 possible members, from a classification based on *UBVI* photometric data. They noted, however, that it was difficult to determine the true nature of this object.

Our automatic solution for NGC 1663 was classified as "poor", since we could observe neither a clear concentration of member stars, nor any correlation with the brightest stars, in the member star chart. Moreover, the member list obtained by our solution included only one star from the photometric member list.

Unfortunately, it is unclear whether members and possible members in Baume et al. (2003) show common proper motions. Because of the lack of ancient epoch data in this particular zone of PM2000, the proper motion errors are greater than the average of the catalogue. This lack of precision prevents us from drawing firm conclusions, and a more precise proper-motion study and/or a radial velocity analysis is needed to clarify the nature of this possible open cluster remnant.

6. Conclusions

We have developed a fully automatic system to determine the kinematic parameters and membership probability lists of open clusters using kinematic and spatial data. The adopted method permits us to take into account any additional parameter (such as radial velocity), given that an analytical form of its probability distribution functions for the cluster and field populations are known a priori. Using this tool, we visually compared the results for several possible parametrisations for the PDFs for all open clusters in PM2000. We concluded that the most reliable function is one that does not take into account the intrinsic dispersion of the cluster.

Based on PM2000 data, we obtained proper motions and kinematic membership lists for open clusters in the Bordeaux's PM2000 zone. For five of them, it was the first such measurement in the literature. We note, however, that for some of those objects, additional studies are necessary, if possible, based on radial velocity data to confirm the membership determination. Moreover, we confirmed from the kinematic point of view, the non-existence of 13 NGC objects, as well as Dol Dzim 2 and Dolidze 35, which are flaged as "not found" and "doubtful" in the D07 catalogue.

We concluded that the open cluster NGC 1807 probably does not exist, as previously suggested in the literature (Balaguer-Núñez et al. 2004a,b). Nonetheless, we note that a radial velocity study to finally settle this issue is still required.

We also determined the tangential velocity of Berkeley 82 to be remarkably similar to the velocities associated with the HI supershell located near it. Thus a physical connection between both objects remains plausible.

By comparing PM2000 and UCAC2 proper motions, we found a periodic systematic variation as a function of the object's right ascension. Strong offsets in both proper motions components were also detected on comparing PM2000 with UCAC3. However, a comparison of PM2000 with PPMX showed only a small offset in the $\mu_\alpha \cos\delta$ component, and no noticeable effect in the μ_δ component.

We finally conclude that the blind use of parametric fully automatic methods in present large-scale proper-motion catalogues is unreliable for most open clusters, since incorrect conclusions are often derived by inspecting VPD and probability histogram plots alone. When conducting studies based on these types of analyses, the results should be critically assessed using independent data; one should avoid using the aforementioned diagrams, since they are directly connected to the fitting itself. As a result, one only checks how mathematically successful the fit is, and not its physical reality. The use of additional data, such as readily available star charts (to identify the spatial clustering of member stars), is highly advisable in the absence of multi-colour photometry or large-scale radial velocity data.

Acknowledgements. Part of this work was supported by the Brazilian agencies FAPESP and CAPES and the Brazilian-French cooperation agreement CAPES-COFECUB. This research has made use of Aladin (Bonnarel et al. 2000), R (R Development Core Team 2008), Vizier (Ochsenbein et al. 2000), and the WEBDA database (Paunzen & Mermilliod 2008). The authors also wish to acknowledge the valuable comments of Prof. Sekhon (U.C. Berkeley) and Prof. Mebane (U. Michigan) regarding the error estimation in the optimization procedure.

References

Balaguer-Núñez, L., Jordi, C., Galadí-Enríquez, D., & Masana, E. 2004a, A&A, 426, 827
Balaguer-Núñez, L., Jordi, C., Galadí-Enríquez, D., & Zhao, J. L. 2004b, A&A, 426, 819
Baume, G., Villanova, S., & Carraro, G. 2003, A&A, 407, 527

Benevides-Soares, P., & Teixeira, R. 1992, A&A, 253, 307
Bica, E., Santiago, B. X., Dutra, C. M., et al. 2001, A&A, 366, 827
Boeche, C., Barbon, R., Henden, A., Munari, U., & Agnolin, P. 2003, A&A, 406, 893
Bonnarel, F., Fernique, P., Bienaymé, O., et al. 2000, A&AS, 143, 33
Bragaglia, A., & Tosi, M. 2006, AJ, 131, 1544
Carraro, G. 2002, A&A, 385, 471
Celeux, G., & Diebolt, J. 1986, Revue de Statistique Appliquée, 34, 35
Chupina, N. V., & Vereshchagin, S. V. 1998, A&A, 334, 552
Cutri, R., Skrutskie, M., Van Dyk, S., et al. 2003, 2MASS All Sky Catalog of point sources
de Bruijne, J. H. J., Hoogerwerf, R., & de Zeeuw, P. T. 2001, A&A, 367, 111
Dias, W. S., Alessi, B. S., Moitinho, A., & Lépine, J. R. D. 2002a, A&A, 389, 871, accessed through Vizier, v.2002-2007
Dias, W. S., Lépine, J. R. D., & Alessi, B. S. 2002b, A&A, 388, 168
Dias, W. S., Assafin, M., Flório, V., Alessi, B. S., & Líbero, V. 2006, A&A, 446, 949
Ducourant, C., & Rapaport, M. 1991, A&A, 241, 303
Ducourant, C., Le Campion, J. F., Rapaport, M., et al. 2006, A&A, 448, 1235
Eisenhauer, F., Quirrenbach, A., Zinnecker, H., & Genzel, R. 1998, ApJ, 498, 278
Friel, E. D. 1995, ARA&A, 33, 381
Frinchaboy, P. M., & Majewski, S. R. 2008, AJ, 136, 118
Frinchaboy, P. M., Marino, A. F., Villanova, S., et al. 2008, MNRAS, 391, 39
Gunn, J. E., Griffin, R. F., Griffin, R. E. M., & Zimmerman, B. A. 1988, AJ, 96, 198
Hetem, A., & Gregorio-Hetem, J. 2007, MNRAS, 382, 1707
Howley, K. M., Geha, M., Guhathakurta, P., et al. 2008, ApJ, 683, 722
Kharchenko, N. V., Piskunov, A. E., Röser, S., Schilbach, E., & Scholz, R.-D. 2005, A&A, 438, 1163
Kim, K.-T., & Koo, B.-C. 2000, ApJ, 529, 229
Kyeong, J., Byun, Y.-I., & Sung, E.-C. 2005, JKAS, 38, 415
Loktin, A. V., & Beshenov, G. V. 2003, Astron. Rep., 47, 6
Makarov, V. V. 2006, AJ, 131, 2967
Mamajek, E. E. 2010, Am. Astron. Soc. Meet., 215, 473
Mebane, W., & Sekhon, S. 2007, J. Stat. Software, to appear, http://sekhon.berkeley.edu
Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 221
Paunzen, E., & Mermilliod, J.-C. 2008, http://www.univie.ac.at/webda/
Pawitan, Y. 2001, In all likelihood: statistical modelling and inference using likelihood (Oxford University Press)
Perryman, M. A. C., de Boer, K. S., Gilmore, G., et al. 2001, A&A, 369, 339
Piatti, A. E., Clariá, J. J., & Ahumada, A. V. 2003, MNRAS, 340, 1249
R Development Core Team 2008, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org
Rapaport, M., Le Campion, J. F., Soubiran, C., et al. 2001, A&A, 376, 325
Röser, S., Schilbach, E., Schwan, H., et al. 2008, A&A, 488, 401
Sanders, W. L. 1971, A&A, 14, 226
Sanner, J., Altmann, M., Brunzendorf, J., & Geffert, M. 2000, A&A, 357, 471
Soubiran, C. 1993, A&A, 274, 181
Soubiran, C., Bienaymé, O., & Siebert, A. 2003, A&A, 398, 141
Strobel, A. 1992, A&A, 253, 374
Sulentic, J., & Tifft, W. 1979, The revised New General Catalogue of nonstellar astronomical objects (University of Arizona Pess)
Teixeira, R., Requieme, Y., Benevides-Soares, P., & Rapaport, M. 1992, A&A, 264, 307
Tosi, M., Fabrizio, L. D., Bragaglia, A., Carusillo, P. A., & Marconi, G. 2004, MNRAS, 354, 225
Vasilevskis, S., Klemola, A., & Preston, G. 1958, AJ, 63, 387
Viateau, B., Réquième, Y., Le Campion, J. F., et al. 1999, A&AS, 134, 173
Wald, A. 1949, Ann. Math. Stat., 20, 595
Yadav, R. K. S., Bedin, L. R., Piotto, G., et al. 2008, A&A, 484, 609
Zacharias, N., Urban, S. E., Zacharias, M. I., et al. 2004, AJ, 127, 3043
Zacharias, N., Finch, C., Girard, T., et al. 2010, AJ, 139, 2184
Zhao, J. L., & He, Y. P. 1990, A&A, 237, 54
Zhao, J. L., Chen, L., & Wen, W. 2006, Chinese J. Astron. Astrophys., 6, 435

Cette thèse a utilisé directement :

- **Hardware**
  Apple MacBook 13" (Intel Core2Duo, 2.16 GHz, 2Gb RAM), Apple MacBook Pro 15" (Intel Core i7 2.66 GHz, 4Gb RAM), Vanoise@Bordeaux (HP, AMD Opteron 1.8 GHz, 2x2 cores, 16 Gb RAM), Venus@Bordeaux (HP, AMD Opteron 2.8 GHz, 4x6 cores, 128 Gb RAM), Bach@IAGUSP (Intel P4 3 GHz HT, 2 Gb RAM), Hydra@IAGUSP (cluster Itautec, Intel Xeon 3.2 GHz, 21 nœuds de 2x1 cores, 42 Gb RAM total), GINA@INCT-A (cluster SGI, Intel Xeon 2.66 Ghz, 2 nœuds de 2x6 cores, 48GB RAM total), 2x Lacie Porsche HD 250Gb, 2x Lacie portable HD 500Gb, Maxtor OneTouch 1Tb, iRex DR1000S reader.

- **Software (OS)**
  Apple Mac OS X Tiger/Leopard/Snow Leopard (General Use), VmWare Fusion + Ubuntu 9 (Software testing), Apple TimeMachine (a sauvé cette thèse lors d'une pane de disque dur).

- **Software (Gestion de projet)**
  Apple iCal, OmniPlan, OmniOutliner.

- **Software (Réseau et communication)**
  Apple Safari, Firefox, FoxyProxy, SSH Tunnel Manager, ExpanDrive, Apple Mail, Skype, Marratech.

- **Software (Mise en page)**
  Apple Pages & KeyNote 06, Apple Preview, LaTeX, Tikz & PGF 2.0, tikz-3dplot, JpgfDraw, TeXShop, LaTeXiT, Papers (articles), BibDesk (livres), Skim, Adobe Acrobat 7 Professional, OmniGraffle.

- **Software (Recherche et développement)**
  ESO Virgo, Aladin, DS9, Eclipse Europa, Netbeans (pour le JStuff), Java 1.4, 1.5, 1.6 (v. Apple, Sun e SoyLatte), GridGain Open Cloud Computing (2.1.1), JFreeChart, JTS - Java Topology Suite, GaiaSimu Java libraries, GaiaTools Java libraries, SVN + Cornerstone (et svnX dans le début), TextWrangler, Apple FileMerge, R & R64, OpenDX, Maple 11, Wolfram's Mathematica 5 e 7, ITT IDL 6 e 7, Python, Perl, Ganglia Cluster Monitoring Toolkit.

- **Sites**
  Google, ADS, arXiv, Google Scholar, CDS (vizier, cdstools, etc.), Webda Open Cluster Database, Gibis@CNES, RSSD@ESA, Livelink@ESA, GaiaWiki@ESA, GaiaPeopleFinder@ESA, GaiaParameterDatabase@ESA, SpringerLink.

Bibliografia

Abe, S.: *Support Vector Machines for Pattern Classification*. Springer-Verlag, New York Inc., 2005. ISBN 1849960976. 151, 154

Abraham, R. G.; Valdes, F.; Yee, H. K. C. & van den Bergh, S.: The Morphologies of Distant Galaxies. I. An Automated Classification System. *Astrophysical Journal*, volume 432; p. 75, 1994; p. 75. doi:10.1086/174550. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1994ApJ...432...75A&link_type=ABSTRACT. 126, 127, 129

Abraham, R. G.; Tanvir, N. R.; Santiago, B. X.; Ellis, R. S.; Glazebrook, K. & van den Bergh, S.: Galaxy morphology to I=25 mag in the Hubble Deep Field. *Monthly Notices of the Royal Astronomical Society*, volume 279; p. L47, 1996a; p. L47. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1996MNRAS.279L..47A&link_type=ABSTRACT. 127, 128, 129

Abraham, R. G.; van den Bergh, S.; Glazebrook, K.; Ellis, R. S.; Santiago, B. X.; Surma, P. & Griffiths, R. E.: The Morphologies of Distant Galaxies. II. Classifications from the Hubble Space Telescope Medium Deep Survey. *Astrophysical Journal Supplement*, volume 107; p. 1, 1996b; p. 1. doi:10.1086/192352. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1996ApJS..107....1A&link_type=ABSTRACT. 7, 124

Abraham, R. G.; van den Bergh, S. & Nair, P.: A New Approach to Galaxy Morphology. I. Analysis of the Sloan Digital Sky Survey Early Data Release. *The Astrophysical Journal*, volume 588; p. 218, 2003; p. 218. doi:10.1086/373919. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2003ApJ...588..218A&link_type=ABSTRACT. 124, 131, 132, 133

Ahmic, M.; Jayawardhana, R.; Brandeker, A.; Scholz, A.; van Kerkwijk, M. H.; Delgado-Donate, E. & Froebrich, D.: Multiplicity among Young Brown Dwarfs and Very Low Mass Stars. *The Astrophysical Journal*, volume 671; p. 2074, 2007; p. 2074. doi:10.1086/522875. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2007ApJ...671.2074A&link_type=ABSTRACT. 5

Aime, C.; Ricort, G. & Harvey, J.: One-dimensional Speckle Interferometry of the Solar Granulation. *Astrophysical Journal*, volume 221; p. 362, 1978; p. 362. doi:10.1086/156034. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1978ApJ...221..362A&link_type=ABSTRACT. A&AA ID. AAA021.071.010. 38

Allen, P. D.; Driver, S. P.; Graham, A. W.; Cameron, E.; Liske, J. & de Propris, R.: The Millennium Galaxy Catalogue: bulge-disc de-

composition of 10095 nearby galaxies. *Monthly Notices of the Royal Astronomical Society*, volume 371; p. 2, 2006; p. 2. doi:10.1111/j.1365-2966.2006.10586.x. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2006MNRAS.371....2A&link_type=ABSTRACT. 7

Anscombe, F.: The Transformation of Poisson, Binomial and Negative-Binomial Data. *Biometrika*, volume 35, no. 3/4; pp. 246, 1948; pp. 246. URL http://www.jstor.org/stable/2332343. 181

Arenou, F.: Java Simulation of Binaries. *GAIA-C4-TN-OPM-FA-008-3*, 2003. 74

Astrium: Summary of Gaia Satellite Design. *GAIA-ASF-TCN-SAT-00007*, 2006. 34, 59

Astrium: Gaia Video Processing Algorithms Validation Test Report. *GAIA-ASF-BG-PLM-00015*, 2008. 19

Astrium: Gaia Video Processing Algorithm User Manual. *GAIA-ASF-UM-PLM-00022*, 2010. 20

Athanassoula, E.; Morin, S.; Wozniak, H.; Puy, D.; Pierce, M. J.; Lombard, J. & Bosma, A.: The shape of bars in early-type barred galaxies. *Monthly Notices of the Royal Astronomical Society*, volume 245; p. 130, 1990; p. 130. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1990MNRAS.245..130A&link_type=ABSTRACT. 176

Babusiaux, C.: The Gaia Instrument and Basic Image Simulator. *Proceedings of the Gaia Symposium "The Three-Dimensional Universe with Gaia" (ESA SP-576). Held at the Observatoire de Paris-Meudon*, volume 576; p. 417, 2005; p. 417. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2005ESASP.576..417B&link_type=ABSTRACT. 116

Babusiaux, C.: The Gaia Instrument and Basic Image Simulator - Overview. *GAIA-C2-SP-OPM-CB-007-D*, 2006. 24

Babusiaux, C.; Chéreau, F.; Grux, E.; Leclerc, N.; Luri, X.; Masana, E.; Robin, A. & Sartoretti, P.: GaiaSimu 2.0 User Guide. *GAIA-C2-TN-OPM-FC-0001-02*, 2007. 74

Babusiaux, C.; Sartoretti, P.; Leclerc, N.; Chéreau, F. & Weiler, M.: The Gaia Instrument and Basic Image Simulator - GIBIS 6.0 User Guide. *GAIA-C2-SP-OPM-CM-003-08*, 2009. 53

Babusiaux, C.; Grux, E.; Arenou, F.; Chéreau, F.; Leclerc, N.; Luri, X.; Masana, E.; Reylé, C.; Russo, F.; Sartoretti, P.; Robin, A. & Gardiol, D.: GaiaSimu 7.1 User Guide. *GAIA-C2-TN-OPM-FC-0001-07*, 2010. 22, 59

Barden, M.; Rix, H.-W.; Somerville, R. S.; Bell, E. F.; Häussler, B.; Peng, C. Y.; Borch, A.; Beckwith, S. V. W.; Caldwell, J. A. R.; Heymans, C.; Jahnke, K.; Jogee, S.; McIntosh, D. H.; Meisenheimer, K. et al: GEMS: The Surface Brightness and Surface Mass Density Evolution of Disk Galaxies. *The Astrophysical Journal*, volume 635; p. 959, 2005; p. 959. doi:10.1086/497679. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2005ApJ...635..959B&link_type=ABSTRACT. 125

Barrett, P.: Application of the Linear Quadtree to Astronomical Databases. *Astronomical Data Analysis Software and Systems IV*, volume 77; p. 472, 1995; p. 472. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1995ASPC...77..472B&link_type=ABSTRACT. 105

Barricelli, N.: Esempi numerici di processi di evoluzione. *Methodos*, 1954. 184

Basri, G.: Observations of Brown Dwarfs. *Annual Review of Astronomy and Astrophysics*, volume 38; p. 485, 2000; p. 485. doi:10.1146/annurev.astro.38.1.485. URL http://www.annualreviews.org/doi/pdf/10.1146/annurev.astro.38.1.485. 5

Bate, M. R.; Bonnell, I. A. & Bromm, V.: The formation mechanism of brown dwarfs. *Monthly Notices of the Royal Astronomical Society*, volume 332; p. L65, 2002; p. L65. doi:10.1046/j.1365-8711.2002.05539.x. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2002MNRAS.332L..65B&link_type=ABSTRACT. 5

Benson, A. J.; Džanović, D.; Frenk, C. S. & Sharples, R.: Luminosity and stellar mass functions of discs and spheroids in the SDSS and the supermassive black hole mass function. *Monthly Notices of the Royal Astronomical Society*, volume 379; p. 841, 2007; p. 841. doi:10.1111/j.1365-2966.2007.11923.x. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2007MNRAS.379..841B&link_type=ABSTRACT. 7

Bershady, M. A.; Jangren, A. & Conselice, C. J.: Structural and Photometric Classification of Galaxies. I. Calibration Based on a Nearby Galaxy Sample. *The Astronomical Journal*, volume 119; p. 2645, 2000; p. 2645. doi:10.1086/301386. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2000AJ....119.2645B&link_type=ABSTRACT. 129

Bertin, E. & Arnouts, S.: SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement*, volume 117; p. 393, 1996; p. 393. URL http://dx.doi.org/10.1051/aas:1996164. 103, 104, 111

Bessel, F. W.: Bestimmung der Entfernung des 61sten Sterns des Schwans. *Astronomische Nachrichten*, volume 16; p. 65, 1838a; p. 65. doi:10.1002/asna.18390160502. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1838AN....16...65B&link_type=GIF. 8

Bessel, F. W.: On the parallax of 61 Cygni. *Monthly Notices of the Royal Astronomical Society*, volume 4; p. 152, 1838b; p. 152. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1838MNRAS...4..152B&link_type=GIF. 8

Bijaoui, A.: Sky background estimation and application. *Astronomy & Astrophysics*, volume 84; p. 81, 1980; p. 81. URL http://adsabs.harvard.edu/abs/1980A%26A....84...81B. 83

Bijaoui, A.: Généralisation de la transformation d'Anscombe. Technical Report, 1994. 181

Binggeli, B.; Sandage, A. & Tarenghi, M.: Studies of the Virgo Cluster. I - Photometry of 109 galaxies near the cluster center to serve as standards. *Astronomical Journal*, volume 89; p. 64, 1984; p. 64. doi:10.1086/113484. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1984AJ....89...64B&link_type=ABSTRACT. 112

Blanton, M. R. & Moustakas, J.: Physical Properties and Environments of Nearby Galaxies. *Annual Review of Astronomy and Astrophysics*, volume 47; p. 159, 2009; p. 159. doi:10.1146/annurev-astro-082708-101734. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2009ARA%26A..47..159B&link_type=ABSTRACT. 173

Blanton, M. R.; Dalcanton, J.; Eisenstein, D.; Loveday, J.; Strauss, M. A.; SubbaRao, M.; Weinberg, D. H.; Anderson, J. E.; Annis, J.; Bahcall, N. A.; Bernardi, M.; Brinkmann, J.; Brunner, R. J.; Burles, S. et al: The Luminosity Function of Galaxies in SDSS Commissioning Data. *The Astronomical Journal*, volume 121; p. 2358, 2001; p. 2358. doi:10.1086/320405. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2001AJ....121.2358B&link_type=ABSTRACT. 139

Bouy, H.; Kolb, J.; Marchetti, E.; Martín, E. L.; Huélamo, N. & Navascués, D. B. Y.: Multi-conjugate adaptive optics images of the Trapezium cluster. *Astronomy & Astrophysics*, volume 477; p. 681, 2008; p. 681. doi:10.1051/0004-6361:20078599. URL http://adsabs.harvard.edu/abs/2008A%26A...477..681B. 9

Brown, A.: Minutes of the CU5/CU4 2D imaging meeting. *GAIA-C5-MN-LEI-AB-011-1*, 2007. 68, 69

Brown, A.: A simple estimate of chance optical projections on the sky. *GAIA-C5-TN-LEI-AB-013-1*, 2008. 76, 77, 78

Broyden, C.: The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics*, volume 6, no. 1; p. 76, 1970; p. 76. URL http://imamat.oxfordjournals.org/content/6/1/76.short. 186

Burgasser, A. J.; Kirkpatrick, J. D.; Cruz, K. L.; Reid, I. N.; Leggett, S. K.; Liebert, J.; Burrows, A. & Brown, M. E.: Hubble Space Telescope NICMOS Observations of T Dwarfs: Brown Dwarf Multiplicity and New Probes of the L/T Transition. *The Astrophysical Journal Supplement Series*, volume 166; p. 585, 2006; p. 585. doi:10.1086/506327. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2006ApJS..166..585B&link_type=ABSTRACT. 5

Burgasser, A. J.; Reid, I. N.; Siegler, N.; Close, L.; Allen, P.; Lowrance, P. & Gizis, J.: Not Alone: Tracing the Origins of Very-Low-Mass Stars and Brown Dwarfs Through Multiplicity Studies. *Protostars and Planets V*, p. 427, 2007; p. 427. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2007prpl.conf..427B&link_type=ABSTRACT. 5

Burgon, R.: Preliminary disturbance detection analysis for Gaia sources fainter than 13^{th} magnitude using extended AF windows. *GAIA-C5-TN-OU-RBG-002*, 2010. 27

Burstein, D.: Structure and origin of S0 galaxies. II - Disk-to-bulge ratios. *The Astrophysical Journal*, volume 234; p. 435, 1979; p. 435. doi:10.1086/157512. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1979ApJ...234..435B&link_type=ABSTRACT. A&AA ID. AAA026.158.172. 172

Capaccioli, M.: Photometry of early-type galaxies and the R exp 1/4 law. *Proceedings of the Conference "The world of galaxies"*, p. 208, 1989; p. 208. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1989woga.conf..208C&link_type=ABSTRACT. 173

Case, G. L.; Cherry, M. L.; Ling, J. C.; Shimizu, T. & Wheaton, W. A.: Initial Results of New Tomographic Imaging of the Gamma-Ray Sky with BATSE. *arXiv*, volume astro-ph.HE, 2009. URL http://arxiv.org/abs/0912.3815v1. 38

Chang, C.-C. & Lin, C.-J.: *LIBSVM: A Library for Support Vector Machines*. Software disponível em http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001. 156

Chang, F.; Chen, C. & Lu, C.: A linear-time component-labeling algorithm using contour tracing technique. *Computer Vision and Image Understanding*, volume 93, no. 2; pp. 206, 2004; pp. 206. URL http://dx.doi.org/10.1016/j.cviu.2003.09.002. 81

Chantry, V. & Magain, P.: Deconvolution of HST images of the Cloverleaf gravitational lens. Detection of the lensing galaxy and a partial Einstein ring. *Astronomy & Astrophysics*, volume 470; p. 467, 2007; p. 467. doi:10.1051/0004-6361:20066839. URL http://adsabs.harvard.edu/abs/2007A%26A...470..467C. 6

Collins, C. A.; Stott, J. P.; Hilton, M.; Kay, S. T.; Stanford, S. A.; Davidson, M.; Hosmer, M.; Hoyle, B.; Liddle, A.; Lloyd-Davies, E.;

Mann, R. G.; Mehrtens, N.; Miller, C. J.; Nichol, R. C. et al: Early assembly of the most massive galaxies. *Nature*, volume 458; p. 603, 2009; p. 603. doi:10.1038/nature07865. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2009Natur.458..603C&link_type=ABSTRACT. 124

Conselice, C. J.: The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories. *The Astrophysical Journal Supplement Series*, volume 147; p. 1, 2003; p. 1. doi:10.1086/375001. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2003ApJS..147....1C&link_type=ABSTRACT. 129, 130, 131

Conselice, C. J.: The fundamental properties of galaxies and a new galaxy classification system. *Monthly Notices of the Royal Astronomical Society*, volume 373; p. 1389, 2006; p. 1389. doi:10.1111/j.1365-2966.2006.11114.x. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2006MNRAS.373.1389C&link_type=ABSTRACT. 129

Cosmic Diary; Hearnshaw, J.; Sakamoto, S.; Consolmagno, B. G.; Ansari, S.; Mumpuni, E.; Korhonen, H.; Marchis, F.; Coustenis, A.; Krone-Martins, A.; Hekker, S.; Dall, T.; van Belle, G.; de Castro, A. I. G. et al: *Postcards from the Edge of the Universe*. ESO, 2010. URL http://www.postcardsfromuniverse.org/. 203

Criminisi, A.; Perez, P. & Toyama, K.: Region Filling and Object Removal by Exemplar-Based Image Inpainting. *IEEE Transactions on Image Processing*, volume 13, no. 9; pp. 1200, 2004; pp. 1200. doi:10.1109/TIP.2004.833105. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1323101&tag=1. 192, 193

Davis, M.: Secrets of the JTS Topology Suite. In *Free and Open Source Software for Geospatial Conference*. 2007. URL http://2007.foss4g.org/presentations/view.php?abstract_id=115. 60, 61

De Berg, M.; Cheong, O.; Van Kreveld, M. & Overmars, M.: *Computational Geometry: Algorithms and Applications*. Springer-Verlag, New York Inc., 3 edition, 2008. 60

de Carvalho, R.; Gal, R.; de Campos Velho, H.; Capelato, H.; Barbera, F. L.; Vasconcellos, E.; Ruiz, R.; Kohl-Moreira, J.; Lopes, P. & Soares-Santos, M.: The Brazilian Virtual Observatory–A New Paradigm for Astronomy. *Journal of Computational Interdisciplinary Sciences*, volume 1, no. 3; pp. 187, 2010; pp. 187. URL http://epacis.org/files/JCIS-v1n3a01.PDF. 202

de Souza, R. E.; Gadotti, D. A. & dos Anjos, S.: BUDDA: A New Two-dimensional Bulge/Disk Decomposition Code for Detailed Structural Analysis of Galaxies. *The Astrophysical Journal Supplement Series*, volume 153; p. 411,

2004; p. 411. doi:10.1086/421554. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2004ApJS..153..411D&link_type=ABSTRACT. 172

DE VAUCOULEURS, G.: Recherches sur les Nebuleuses Extragalactiques. *Annales d'Astrophysique*, volume 11; p. 247, 1948; p. 247. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1948AnAp...11..247D&link_type=GIF. 125, 172, 173

DE VAUCOULEURS, G.: General Physical Properties of External Galaxies. *Handbuch der Physik*, volume 53; p. 311, 1959; p. 311. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1959HDP....53..311D&link_type=ABSTRACT. 172

DEAN, J. & GHEMAWAT, S.: MapReduce: Simplified Data Processing on Large Clusters. *Proceedings of the OSDI '04: 6th Symposium on Operating Systems Design and Implementation*, pp. 137–149, 2004; pp. 137. doi:10.1145/1327452.1327492. URL http://portal.acm.org/citation.cfm?doid=1327452.1327492. 61

DEANS, S.: *The Radon transform and some of its applications*. Wiley-Interscience, New York, 1983. 37, 38, 39, 40, 41

DELGADO-SERRANO, R.; HAMMER, F.; YANG, Y. B.; PUECH, M.; FLORES, H. & RODRIGUES, M.: How was the Hubble sequence 6 Gyr ago? *Astronomy & Astrophysics*, volume 509; p. 78, 2010; p. 78. doi:10.1051/0004-6361/200912704. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010A%26A...509A..78D&link_type=ABSTRACT. 124, 125

DEMOMENT, G.: Problèmes inverses en traitement du signal et de l'image. In J.-P. R. et A. Bijaoui, editor, *Vois nouvelles pour l'Analyse des Données en Sciences de l'Univers, Écoles d'Astrophysique Solaire d'Oléron, Journal de Physique IV*. 2002. 50

DOI, M.; FUKUGITA, M. & OKAMURA, S.: Morphological Classification of Galaxies Using Simple Photometric Parameters. *Monthly Notices of the Royal Astronomical Society*, volume 264; p. 832, 1993; p. 832. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1993MNRAS.264..832D&link_type=ABSTRACT. 126

DOLLET, C.: *Compression et Restauration Pour L'Imagerie du Ciel a Haute Resolution Angulaire*. Tese de Doutorado. Université de Nice Sophia-Antipolis, École Docorale Sciences Fondamentales et Appliquées, 2004. 47, 50, 51, 104

DOLLET, C.; BIJAOUI, A. & MIGNARD, F.: The Windows Design and the Restoration of Object Environments. *Proceedings of the Gaia Symposium "The Three-Dimensional Universe with Gaia" (ESA SP-576). Held at the Observatoire de Paris-Meudon*, volume 576; p. 429, 2005; p. 429. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2005ESASP.576..429D&link_type=ABSTRACT. 50

DPAC: Proposal for the Gaia Data Processing - Response to the ESA's Announcement of Opportunity. *GAIA-CD-SP-DPAC-FM-030-2*, pp. 1–710, 2007; pp. 1. 16, 18, 20, 24

Drimmel, R.: Galactic absorption model: refined version. *GAIA-SWG-RD-03*, 2002. 74

Ducourant, C.; Campion, J. F. L.; Rapaport, M.; Camargo, J. I. B.; Soubiran, C.; Périe, J. P.; Teixeira, R.; Daigne, G.; Triaud, A.; Réquième, Y.; Fresneau, A. & Colin, J.: The PM2000 Bordeaux proper motion catalogue (+11dg < delta < +18dg). *A&A*, volume 448; p. 1235, 2006; p. 1235. doi:10.1051/0004-6361:20053220. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2006A%26A...448.1235D&link_type=ABSTRACT. 201, 207

Ducourant, C.; Teixeira, R.; Krone-Martins, A.; Campion, J. F. L. & Chauvin, G.: Parallax programs at sub-mas accuracy level. *IV Reunión sobre Astronomía Dinámica en Latino América*, volume 34; p. 29, 2008; p. 29. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2008RMxAC..34...29D&link_type=ABSTRACT. 202

Erben, T.; Waerbeke, L. V.; Bertin, E.; Mellier, Y. & Schneider, P.: How accurately can we measure weak gravitational shear? *Astronomy & Astrophysics*, volume 366; p. 717, 2001; p. 717. doi:10.1051/0004-6361:20010013. URL http://adsabs.harvard.edu/abs/2001A%26A...366..717E. 111

Ernst, A.; Just, A.; Berczik, P. & Petrov, M. I.: Calibration of radii and masses of open clusters with a simulation. *eprint arXiv*, volume 1009; p. 710, 2010; p. 710. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010arXiv1009.0710E&link_type=ABSTRACT. 5

ESA: *The Hipparcos and Tycho Catalogues*. ESA SP-1200, 1997. 9, 10, 11, 12, 13, 14

Fabian, A. C.: Serendipity in Astronomy. *arXiv*, volume physics.pop-ph, 2009. URL http://arxiv.org/abs/0908.2784v1. 32

Fabricius, C.: 2D Imaging Experiences from the Tycho project. *CU5 - CU4 2D Imaging meeting, Brussels, 28-29 June*, pp. 1–18, 2007; pp. 1. 3

Fabricius, C.; Høg, E.; Makarov, V. V.; Mason, B. D.; Wycoff, G. L. & Urban, S. E.: The Tycho double star catalogue. *Astronomy & Astrophysics*, volume 384; p. 180, 2002; p. 180. doi:10.1051/0004-6361:20011822. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2002A%26A...384..180F&link_type=ABSTRACT. 5

Fadili, M.; Starck, J. & Murtagh, F.: Inpainting and zooming using sparse representations. *The Computer Journal*, volume 52, no. 1; p. 64, 2009; p. 64. URL http://comjnl.oxfordjournals.org/content/52/1/64.abstract. 192

Farihi, J.; Hoard, D. W. & Wachter, S.: White Dwarf - Red Dwarf Systems Resolved with the Hubble Space Telescope. II. Full Snapshot Survey Results. *arXiv*, volume astro-ph.SR, 2010. URL http://arxiv.org/abs/1008.2545v2. 5

Fellhauer, M.; Lin, D. N. C.; Bolte, M.; Aarseth, S. J. & Williams, K. A.: The White Dwarf Deficit in Open Clusters: Dynamical Processes. *The Astrophysical Journal*, volume 595; p. L53, 2003; p. L53. doi:10.1086/379005. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2003ApJ...595L..53F&link_type=ABSTRACT. 5

Fioc, M. & Rocca-Volmerange, B.: Far-UV and deep surveys: bursting dwarfs versus normal galaxies. *Astronomy & Astrophysics*, volume 344; p. 393, 1999; p. 393. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1999A%26A...344..393F&link_type=ABSTRACT. 112

Fiorio, C. & Gustedt, J.: Two linear time Union-Find strategies for image processing. *Theoretical Computer Science*, volume 154, no. 2; pp. 165, 1996; pp. 165. URL http://dx.doi.org/10.1016/0304-3975(94)00262-2. 81

Fletcher, R.: A new approach to variable metric algorithms. *The Computer Journal*, volume 13, no. 3; p. 317, 1970; p. 317. URL http://comjnl.oxfordjournals.org/content/13/3/317.abstract. 186

Fogel, D.: Nils Barricelli–Artificial Life, Coevolution, Self-Adaptation. *IEEE Computational Intelligence Magazine*, volume 1, no. 1; pp. 41 , 2006; pp. 41 . doi:10.1109/MCI.2006.1597062. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1597062. 184

Foster, I. & Kesselman, C.: *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, 2 edition, 2004. 61

Freedman, W.: *Measuring and modeling the universe*. Cambridge University Press, 2004. 7

Freeman, K. C.: On the Disks of Spiral and S0 Galaxies. *Astrophysical Journal*, volume 160; p. 811, 1970; p. 811. doi:10.1086/150474. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1970ApJ...160..811F&link_type=ABSTRACT. A&AA ID. AAA003.151.062. 172

Frei, Z.; Guhathakurta, P.; Gunn, J. E. & Tyson, J. A.: A Catalog of Digital Images of 113 Nearby Galaxies. *The Astronomical Journal*, volume 111; p. 174, 1996; p. 174. doi:10.1086/117771. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1996AJ....111..174F&link_type=ABSTRACT. 128, 136, 140, 141, 162

Fruchter, A. S. & Hook, R. N.: Drizzle: A Method for the Linear Reconstruction of Undersampled Images. *The Publications of the Astronomical Society of the Pacific*, volume 114; p. 144, 2002; p. 144. doi:10.

1086/338393. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2002PASP..114..144F&link_type=ABSTRACT. 48

Gadotti, D. A.: Image decomposition of barred galaxies and AGN hosts. *Monthly Notices of the Royal Astronomical Society*, volume 384; p. 420, 2008; p. 420. doi: 10.1111/j.1365-2966.2007.12723.x. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2008MNRAS.384..420G&link_type=ABSTRACT. 172, 173, 174

Gadotti, D. A.: Structural properties of pseudo-bulges, classical bulges and elliptical galaxies: a Sloan Digital Sky Survey perspective. *Monthly Notices of the Royal Astronomical Society*, volume 393; p. 1531, 2009; p. 1531. doi: 10.1111/j.1365-2966.2008.14257.x. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2009MNRAS.393.1531G&link_type=ABSTRACT. 7

Gallardo, E. & Masana, E.: GASS User Manual. *GAIA-C2-TN-UB-002-1*, 2010. 23

Gavras, P.; Krone-Martins, A.; Ducourant, C. & Sinachopoulos, D.: MAnager of Gaia Image Library - MAGIL 1.0 - User Guide. *GAIA-C4-SP-NOA-PG-NUM-0D*, 2010. 23, 103, 104, 116, 118

Giavalisco, M.: Lyman-Break Galaxies. *Annual Review of Astronomy and Astrophysics*, volume 40; p. 579, 2002; p. 579. doi:10.1146/annurev.astro.40.121301.111837. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2002ARA%26A..40..579G&link_type=ABSTRACT. 134

Goldfarb, D.: A family of variable-metric methods derived by variational means. *Mathematics of Computation*, volume 24, no. 109; pp. 23, 1970; pp. 23. URL http://www.jstor.org/stable/2004873. 186

Gonzalez, R. & Woods, R.: *Digital Image Processing*. Prentice Hall, New York, 2 edition, 2002. 79

GPDB_2010-03-02: Gaia Parameter Database Contents, Version:-live-:2010-03-02T20:47:47. 2010. URL http://gaia.esac.esa.int/gpdb/. 34

Graf, A. B. A. & Wichmann, F. A.: Insights from Machine Learning Applied to Visual Human Classification. *Advances in Neural Information Processing Systems*, volume 16; pp. 905, 2004; pp. 905. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.67.8922. 151

Graham, A. W. & Driver, S. P.: A Concise Reference to (Projected) Sérsic R1/n Quantities, Including Concentration, Profile Slopes, Petrosian Indices, and Kron Magnitudes. *Publications of the Astronomical Society of Australia*, volume 22; p. 118, 2005; p. 118. doi:10.1071/AS05001. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2005PASA...22..118G&link_type=ABSTRACT. 173

GridGain: GridGain Open Cloud Platform. 2010. URL http://www.gridgain.com/. 61

Griffiths, R. E.; Ratnatunga, K. U.; Neuschaefer, L. W.; Casertano, S.; Im, M.; Wyckoff, E. W.; Ellis, R. S.; Gilmore, G. F.; Elson, R. A. W.; Glazebrook, K.; Schade, D. J.; Windhorst, R. A.; Schmidtke, P.; Gordon, J. et al: The Hubble Space Telescope Medium Deep Survey with the Wide Field and Planetary Camera. I: Methodology and results on the field near 3C 273. *The Astrophysical Journal*, volume 437; p. 67, 1994; p. 67. doi: 10.1086/174976. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1994ApJ...437...67G&link_type=ABSTRACT. 7, 120, 127

Hall, P.; Park, B. U. & Samworth, R. J.: Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, volume 36, no. 5; pp. 2135, 2008; pp. 2135. URL http://www.ams.org/mathscinet/search/publications.html?pg1=MR&s1=MR2458182. 89

Hathi, N. P.; Ferreras, I.; Pasquali, A.; Malhotra, S.; Rhoads, J. E.; Pirzkal, N.; Windhorst, R. A. & Xu, C.: Stellar Populations of Late-Type Bulges at z = 1 in the Hubble Ultra Deep Field. *The Astrophysical Journal*, volume 690; p. 1866, 2009; p. 1866. doi:10.1088/0004-637X/690/2/1866. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2009ApJ...690.1866H&link_type=ABSTRACT. 172

Herman, G. T.: *Fundamentals of Computerized Tomography: Image Reconstruction from Projection*. Springer-Verlag, 2 edition, 2009. 37

Høg, E.; Fabricius, C.; Makarov, V. V.; Bastian, U.; Schwekendiek, P.; Wicenec, A.; Urban, S.; Corbin, T. & Wycoff, G.: Construction and verification of the Tycho-2 Catalogue. *Astronomy & Astrophysics*, volume 357; p. 367, 2000a; p. 367. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2000A%26A...357..367H&link_type=ABSTRACT. 5, 14

Høg, E.; Fabricius, C.; Makarov, V. V.; Urban, S.; Corbin, T.; Wycoff, G.; Bastian, U.; Schwekendiek, P. & Wicenec, A.: (Erratum) Letter to the Editor - The Tycho-2 catalogue of the 2.5 million brightest stars. *Astronomy & Astrophysics*, volume 363; p. 385, 2000b; p. 385. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2000A%26A...363..385H&link_type=GIF. 14

Høg, E.; Fabricius, C.; Makarov, V. V.; Urban, S.; Corbin, T.; Wycoff, G.; Bastian, U.; Schwekendiek, P. & Wicenec, A.: The Tycho-2 catalogue of the 2.5 million brightest stars. *Astronomy & Astrophysics*, volume 355; p. L27, 2000c; p. L27. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2000A%26A...355L..27H&link_type=ABSTRACT. 14

Högbom, J. A.: Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines. *Astronomy & Astrophysics Supplement*, volume 15; p. 417, 1974; p. 417. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1974A%26AS...15..417H&link_type=ABSTRACT. 52

Holland, J.: *Adaptation in natural and artificial system: an introduction with application to biology, control and artificial intelligence*. University of Michigan Press, 1975. 184

Hubble, E.: *The Realm of the Nebulae*. Yale University Press, New Haven, 1936. 124

Hubble, E. P.: Extragalactic nebulae. *The Astrophysical Journal*, volume 64; p. 321, 1926; p. 321. doi:10.1086/143018. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1926ApJ....64..321H&link_type=ABSTRACT. 124

Huertas-Company, M.; Rouan, D.; Soucail, G.; Fèvre, O. L.; Tasca, L. & Contini, T.: Morphological evolution of z 1 galaxies from deep K-band AO imaging in the COSMOS deep field. *Astronomy & Astrophysics*, volume 468; p. 937, 2007; p. 937. doi:10.1051/0004-6361:20066673. URL http://adsabs.harvard.edu/abs/2007A%26A...468..937H. 128

Huertas-Company, M.; Tasca, L.; Rouan, D.; Kneib, J. P. & Fèvre, O. L.: Morphological Evolution from z 2 in the COSMOS Field from Ks-Band Imaging. *Highlights of Spanish Astrophysics V*, p. 301, 2010; p. 301. doi:10.1007/978-3-642-11250-8_50. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010hsa5.conf..301H&link_type=ABSTRACT. 135

Isasi, Y.; Zaldua, I.; Sartoretti, P.; Luri, X.; Babusiaux, C. & Martinez, O.: GOG v7.0 User Guide. *GAIA-C2-UG-UB-YI-003-7*, 2010. 26

Jarrett, T.: Infrared Universe. 2004. URL http://web.ipac.caltech.edu/staff/jarrett/galaxies/. 71

Jolliffe, I. T.: *Principal Component Analysis*. Springer Verlag, 2nd edition, 2004. 136, 147

Jordi, C.: Photometric relationships between Gaia photometry and existing photometric systems. *GAIA-C5-TN-UB-CJ-04-2*, 2007. 70

Jordi, C.; Gebran, M.; Carrasco, J. M.; Bruijne, J. D.; Voss, H.; Fabricius, C.; Knude, J.; Vallenari, A.; Kohley, R. & Mora, A.: Gaia broad band photometry. *Astronomy & Astrophysics*, volume 523; p. 48, 2010; p. 48. doi:10.1051/0004-6361/201015441. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010A%26A...523A..48J&link_type=ABSTRACT. 15

Kak, A. C. & Slaney, M.: *Principles of Computerized Tomographic Imaging*. Society of Industrial and Applied Mathematics, 2001. 37

Kennicutt, R. C.: Star Formation in Galaxies Along the Hubble Sequence. *Annual Review of Astronomy and Astrophysics*, volume 36; p. 189, 1998; p. 189. doi:10.1146/annurev.astro.36.1.189. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1998ARA%26A..36..189K&link_type=ABSTRACT. 131

Kent, S. M.: CCD surface photometry of field Galaxies. II - Bulge/disk decompositions. *The Astrophysical Journal Supplement Series*, volume 59; p. 115, 1985; p. 115. doi:10.1086/191066. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1985ApJS...59..115K&link_type=ABSTRACT. 172

Kirkpatrick, J. D.: New Spectral Types L and T. *Annual Review of Astronomy & Astrophysics*, volume 43; p. 195, 2005; p. 195. doi:10.1146/annurev.astro.42.053102.134017. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2005ARA%26A..43..195K&link_type=ABSTRACT. 5

Kormendy, J.: Brightness distributions in compact and normal galaxies. III - Decomposition of observed profiles into spheroid and disk components. *The Astrophysical Journal*, volume 217; p. 406, 1977; p. 406. doi:10.1086/155589. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1977ApJ...217..406K&link_type=ABSTRACT. A&AA ID. AAA020.158.040. 172

Kormendy, J. & Kennicutt, R. C.: Secular Evolution and the Formation of Pseudobulges in Disk Galaxies. *Annual Review of Astronomy and Astrophysics*, volume 42; p. 603, 2004; p. 603. doi:10.1146/annurev.astro.42.053102.134024. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2004ARA%26A..42..603K&link_type=ABSTRACT. 124, 172

Kovalevsky, J.: The Hipparcos project at Strasbourg Observatory. *The Multinational History of Strasbourg Astronomical Observatory. Edited by André Heck. Dordrecht: Springer*, volume 330; p. 215, 2005; p. 215. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2005ASSL..330..215K&link_type=ABSTRACT. 9

Kovalevsky, J.; Falin, J. L.; Pieplu, J. L.; Bernacca, P. L.; Donati, F.; Froeschle, M.; Galligani, I.; Mignard, F.; Morando, B.; Perryman, M. A. C.; Schrijver, H.; van Daalen, D. T.; van der Marel, H.; Villenave, M. et al: The FAST HIPPARCOS Data Reduction Consortium: Overview of the Adopted Reduction Software. *Astronomy & Astrophysics*, volume 258; p. 7, 1992; p. 7. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1992A%26A...258....7K&link_type=ABSTRACT. 11, 12

Krone-Martins, A.: Morphological analysis based on reconstructed images and Gaia window data. *ExtraGalactic Science with Gaia 2010*, 2010. URL http://www-n.oca.eu/rousset/EGSG/. 171

Krone-Martins, A. & Ducourant, C.: DU470 Software Design Description. *GAIA-C4-SP-LAB-AKM-002*, 2008. 139

Krone-Martins, A.; Ducourant, C. & Teixeira, R.: Semi-numerical estimate for the number of chance optical projections on the sky. *GAIA-C5-TN-LAB-AKM-001-1*, 2008a. 67

Krone-Martins, A.; Ducourant, C. & Teixeira, R.: Support Vector Machines and CASGM20 Parameters Applied to Morphological Classification of Reconstructed 2D Images of Extended Objects Within the ESA-Gaia Mission. *Proceedings of the International Conference: "Classification and Discovery in Large Astronomical Surveys". AIP Conference Proceedings.*, volume 1082; p. 151, 2008b; p. 151. doi:10.1063/1.3059030. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2008AIPC.1082..151K&link_type=ABSTRACT. (c) 2008: American Institute of Physics. 123, 143

Krone-Martins, A.; Ducourant, C.; Teixeira, R.; le Campion, J.-F. & Harrison, D.: Educated Image Segregation for Reconstructed Images from the Gaia satellite. *Invited talk at the 6th Astronomical Data Analysis Conference (Monastir, Tunisia)*, 2010a. URL http://www.aset.org.tn/conf/ADA6/online_presentations.php. 67

Krone-Martins, A.; Soubiran, C.; Ducourant, C.; Teixeira, R. & Campion, J. F. L.: Kinematic parameters and membership probabilities of open clusters in the Bordeaux PM2000 catalogue. *Astronomy & Astrophysics*, volume 516; p. 3, 2010b; p. 3. doi:10.1051/0004-6361/200913881. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010A%26A...516A...3K&link_type=ABSTRACT. 201, 207

Krone-Martins, A.; dos Anjos, S.; Ducourant, C.; Machado, R.; Teixeira, R.; le Campion, J.-F.; de Souza, R. & Gavras, P.: Purely morphological classification of galaxies from Gaia data - testing on simulations and real images. in prep.a. 123

Krone-Martins, A.; Ducourant, C. & Teixeira, R.: Fractional and Angular Coverage of Gaia reconstructed images in the celestial sphere. *GAIA-C4-TN-LAB-AKM*, in prep.b. 33

Krone-Martins, A. G. O.; Ducourant, C.; Teixeira, R. & Luri, X.: JStuff - a preliminary extragalactic model for the ESA-Gaia satellite simulation framework. *A Giant Step: from Milli- to Micro-arcsecond Astrometry*, volume 248; p. 276, 2008c; p. 276. doi:10.1017/S1743921308019297. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2008IAUS..248..276K&link_type=ABSTRACT. 103, 104

Kunszt, P. Z.; Szalay, A. S.; Csabai, I. & Thakar, A. R.: The Indexing of the SDSS Science Archive. *Astronomical Data Analysis Software and Systems IX*,

volume 216; p. 141, 2000; p. 141. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2000ASPC..216..141K&link_type=ABSTRACT. ISBN: 1-58381-047-1. 105

Lasker, B. M.; Lattanzi, M. G.; McLean, B. J.; Bucciarelli, B.; Drimmel, R.; Garcia, J.; Greene, G.; Guglielmetti, F.; Hanley, C.; Hawkins, G.; Laidler, V. G.; Loomis, C.; Meakes, M.; Mignani, R. et al: The Second-Generation Guide Star Catalog: Description and Properties. *The Astronomical Journal*, volume 136; p. 735, 2008; p. 735. doi:10.1088/0004-6256/136/2/735. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2008AJ....136..735L&link_type=ABSTRACT. 70

Lawrence, J. S.; Ashley, M. C. B.; Tokovinin, A. & Travouillon, T.: Exceptional astronomical seeing conditions above Dome C in Antarctica. *Nature*, volume 431; p. 278, 2004; p. 278. doi:10.1038/nature02929. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2004Natur.431..278L&link_type=ABSTRACT. 4

Lindegren, L.: The scanning law for GAIA. *SAG-LL-014*, 1998. 58

Lindegren, L.: Calculating the GAIA nominal scanning law. *SAG-LL-35*, 2001. 58

Lindegren, L.: A framework for consistent definition and use of LSFs in AF/BP/RP/RVS. *GAIA-C3-TN-LU-LL-080-02*, 2009. 26

Lindegren, L.; Hog, E.; van Leeuwen, F.; Murray, C. A.; Evans, D. W.; Penston, M. J.; Perryman, M. A. C.; Petersen, C.; Ramamani, N. & Snijders, M. A. J.: The NDAC HIPPARCOS data analysis consortium - Overview of the reduction methods. *Astronomy & Astrophysics*, volume 258; p. 18, 1992a; p. 18. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1992A%26A...258...18L&link_type=ABSTRACT. 11, 12

Lindegren, L.; van Leeuwen, F.; Petersen, C.; Perryman, M. A. C. & Soderhjelm, S.: Positions and parallaxes from the HIPPARCOS satellite - A first attempt at a global astrometric solution. *Astronomy & Astrophysics*, volume 258; p. 134, 1992b; p. 134. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1992A%26A...258..134L&link_type=ABSTRACT. 12

Lorenz, M.: Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association*, volume 9, no. 70; pp. 209, 1905; pp. 209. URL http://www.jstor.org/stable/2276207. 131

Lotz, J. M.; Primack, J. & Madau, P.: A New Nonparametric Approach to Galaxy Morphological Classification. *The Astronomical Journal*, volume 128; p. 163, 2004; p. 163. doi:10.1086/421849. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2004AJ....128..163L&link_type=ABSTRACT. 134, 135, 136

Luri, X.: Use of the GAIA Nominal Scanning Law in the GAIA simulator. *GAIA-XL-001*, 2001. 59

Magain, P.; Surdej, J.; Swings, J.-P.; Borgeest, U. & Kayser, R.: Discovery of a quadruply lensed quasar - The 'clover leaf' H1413 + 117. *Nature*, volume 334; p. 325, 1988; p. 325. doi:10.1038/334325a0. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1988Natur.334..325M&link_type=ABSTRACT. 6

Marleau, F. R. & Simard, L.: Quantitative Morphology of Galaxies in the Hubble Deep Field. *The Astrophysical Journal*, volume 507; p. 585, 1998; p. 585. doi:10.1086/306356. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1998ApJ...507..585M&link_type=ABSTRACT. 172

Marr, R. B.: An Overview Of Image Reconstruction. *Proceedings of the International Symposium on Ill-Posed Problems: Theory and Practice, Newark, Delaware*, pp. 1–59, 1979; pp. 1. URL http://www.osti.gov/bridge/purl.cover.jsp;jsessionid=E484FB678DEE258F7EE4F20D3E497F55?purl=/5404987-w4AHsA/. 37

Martin, F.; Bijaoui, A.; Touma, H. & Aime, C.: Astronomical image reconstruction via space slit aperture telescope. *Infrared, Adaptive, and Synthetic Aperture Optical Systems, Proceedings of SPIE*, volume 643; p. 189, 1986; p. 189. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1986SPIE..643..189M&link_type=ABSTRACT. 38

Martinez, F.; Rueda, A. J. & Feito, F. R.: A new algorithm for computing Boolean operations on polygons. *Computers & Geosciences*, volume 35, no. 6; pp. 1177, 2009; pp. 1177. doi:10.1016/j.cageo.2008.08.009. URL http://dx.doi.org/10.1016/j.cageo.2008.08.009. 60

Mary, D.; Hog, E.; Lindstroem, H. & Bastian, U.: Updated Simulation of SM Image Reconstruction. *GAIA-C5-TN-ARI-DM-002-1*, 2006. 46, 50, 51, 54, 55

Mebane, W. & Sekhon, J.: Genetic Optimization Using Derivatives: The rgenoud package for R. *Journal of Statistical Software*, 2007. URL http://sekhon.berkeley.edu/papers/rgenoudJSS.pdf. 186

Metcalfe, N.; Shanks, T.; Fong, R. & Jones, L. R.: Galaxy number counts. II - CCD observations to B = 25 mag. *Monthly Notices of the Royal Astronomical Society*, volume 249; p. 498, 1991; p. 498. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1991MNRAS.249..498M&link_type=ABSTRACT. 113

Mignard, F.: The Gaia mission objectives, description, data processing. *ADA 6 - Sixth Conference on Astronomical Data Analysis*, p. 10, 2010; p. 10. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010ada..confE..10M&link_type=ARTICLE. 15, 16, 21

Mignard, F.; Froeschle, M.; Badiali, M.; Cardini, D.; Emanuele, A.; Falin, J. L. & Kovalevsky, J.: HIPPARCOS double star recognition and processing within the FAST consortium. *Astronomy & Astrophysics*, volume 258; p. 165, 1992; p. 165. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1992A%26A...258..165M&link_type=ABSTRACT. 12

Murtagh, F.; Starck, J.-L. & Bijaoui, A.: Image restoration with noise suppression using a multiresolution support. *Astronomy & Astrophysics Supplement*, volume 112; p. 179, 1995; p. 179. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1995A%26AS..112..179M&link_type=ABSTRACT. 181

Naim, A.; Lahav, O.; Sodre, L. & Storrie-Lombardi, M. C.: Automated morphological classification of APM galaxies by supervised artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, volume 275; p. 567, 1995; p. 567. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1995MNRAS.275..567N&link_type=ABSTRACT. 124

Nash, J.: *Compact numerical methods for computers: linear algebra and function minimisation*. Adam Hilger, 1990. ISBN 085274319X. 187

Nocedal, J. & Wright, S.: *Numerical optimization*. Springer Verlag, 1999. ISBN 0387987932. 186, 187

Nurmi, P.: Combining GAIA windows I: SNR calculations of secondary sources in the ideal case for AF11. *DMS-PN-01*, 2003. 48

O'Mullane, W.; Hoar, J.; Levoir, T.; Angeli, F. D.; Nguyen, A.; Olias, A. & Lammers, U.: Software Engineering Guidelines for DPAC. *GAIA-C1-UG-ESAC-WOM-011-02*, 2008. 92

O'Rourke, J.: *Computational Geometry in C*. Cambridge University Press, 2 edition, 1998. 60

Parker, R. J. & Goodwin, S. P.: The dynamical evolution of very-low mass binaries in open clusters. *eprint arXiv*, volume 1009; p. 3110, 2010; p. 3110. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010arXiv1009.3110P&link_type=ABSTRACT. 5

Parks, P. C.: On the Determination of Functions from Their Integral Values Along Certain Manifolds. *IEEE Transactions on Medical Imaging*, volume MI-5, no. 4; pp. 170, 1986; pp. 170. doi:10.1109/TMI.1986.4307775. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4307775. 37

Peng, C. Y.; Ho, L. C.; Impey, C. D. & Rix, H.-W.: Detailed Structural Decomposition of Galaxy Images. *The Astronomical Journal*, volume 124; p. 266, 2002; p. 266. doi:10.1086/340952. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2002AJ....124..266P&link_type=ABSTRACT. 172

Peng, C. Y.; Ho, L. C.; Impey, C. D. & Rix, H.-W.: Detailed Decomposition of Galaxy Images. II. Beyond Axisymmetric Models. *The Astronomical Journal*, volume 139; p. 2097, 2010; p. 2097. doi:10.1088/0004-6256/139/6/2097. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010AJ....139.2097P&link_type=ABSTRACT. 172, 173, 174

Perryman, M.: *Astronomical Applications of Astrometry*. Cambridge University Press, 2009. 14

Perryman, M.: *The Making of History's Greatest Star Map*. Springer Verlag, 2010. 9, 15

Perryman, M. A. C.; Hog, E.; Kovalevsky, J.; Lindegren, L.; Turon, C.; Bernacca, P. L.; Creze, M.; Donati, F.; Grenon, M. & Grewing, M.: In-orbit performance of the HIPPARCOS astrometry satellite. *Astronomy & Astrophysics*, volume 258; p. 1, 1992; p. 1. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1992A%26A...258....1P&link_type=ABSTRACT. 12

Perryman, M. A. C.; de Boer, K. S.; Gilmore, G.; Høg, E.; Lattanzi, M. G.; Lindegren, L.; Luri, X.; Mignard, F.; Pace, O. & de Zeeuw, P. T.: GAIA: Composition, formation and evolution of the Galaxy. *Astronomy & Astrophysics*, volume 369; p. 339, 2001; p. 339. doi:10.1051/0004-6361:20010085. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2001A%26A...369..339P&link_type=ABSTRACT. 15

Peyrega, C.: *Reconstruction d'images des champs de la mission GAIA de l'ESA à partir des fenêtres transmises sur Terre*. Dissertação de Mestrado. Observatoire de la Côte d'Azur, 2007. 51, 52, 53

Pier, J. R.; Munn, J. A.; Hindsley, R. B.; Hennessy, G. S.; Kent, S. M.; Lupton, R. H. & Ivezić, Ž.: Astrometric Calibration of the Sloan Digital Sky Survey. *The Astronomical Journal*, volume 125; p. 1559, 2003; p. 1559. doi:10.1086/346138. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2003AJ....125.1559P&link_type=ABSTRACT. 14

R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL http://www.R-project.org. ISBN 3-900051-07-0. 41

Radon, J.: Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Berichte der Sächsischen Akadamie der Wissenschaft*, volume 69; pp. 262, 1917; pp. 262. 37, 39

Reeves, C. & Rowe, J.: *Genetic Algorithms: Principles and Perspectives: a guide to GA theory*. Kluwer Academic Publishers, 2002. ISBN 1402072406. 184

Reipurth, B. & Clarke, C.: The Formation of Brown Dwarfs as Ejected Stellar Embryos. *The Astronomical Journal*, volume 122; p. 432, 2001; p. 432. doi: 10.1086/321121. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2001AJ....122..432R&link_type=ABSTRACT. 5

Robin, A.: The Extragalactic Universe as simulated in Gaia simulator. *ExtraGalactic Science with Gaia 2010*, pp. 1–30, 2010; pp. 1. URL http://www-n.oca.eu/rousset/EGSG/. 112, 118

Robin, A.; Reylé, C.; Arenou, F.; Babusiaux, C.; i Musoll, A. L.; Luri, X. & Sartoretti, P.: Universe Model Overview. *GAIA-C2-TN-LAOB-AR-004-01*, 2007. 74, 106

Robin, A.; Reylé, C.; Arenou, F.; Babusiaux, C.; i Musoll, A. L.; Luri, X.; Sartoretti, P.; Tanga, P. & Grux, E.: Universe Model Overview. *GAIA-C2-TN-LAOB-AR-004-08*, 2010. 23

Rosenfeld, A. & Pfaltz, J.: Sequential Operations in Digital Picture Processing. *Journal of the ACM*, volume 13, no. 4, 1966. URL http://portal.acm.org/citation.cfm?id=321356.321357. 81

Sandage, A.: The Hubble Atlas of Galaxies. *Carnegie Institute of Washington Publications*, 1961. 124

Sandage, A.: The Classification of Galaxies: Early History and Ongoing Developments. *Annual Review of Astronomy and Astrophysics*, volume 43; p. 581, 2005; p. 581. doi:10.1146/annurev.astro.43.112904.104839. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2005ARA%26A..43..581S&link_type=ABSTRACT. 124

Sandage, A.; Freeman, K. C. & Stokes, N. R.: The Intrinsic Flattening of e, so, and Spiral Galaxies as Related to Galaxy Formation and Evolution. *The Astrophysical Journal*, volume 160; p. 831, 1970; p. 831. doi:10.1086/150475. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1970ApJ...160..831S&link_type=ABSTRACT. A&AA ID. AAA003.151.063. 113

Sanz, J. L.; Barreiro, R. B.; Cayón, L.; Martínez-González, E.; Ruiz, G. A.; Díaz, F. J.; Argüeso, F.; Silk, J. & Toffolatti, L.: Analysis of CMB maps with 2D wavelets. *Astronomy & Astrophysics Supplement*, volume 140; p. 99, 1999; p. 99. doi:10.1051/aas:1999119. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1999A%26AS..140...99S&link_type=ABSTRACT. 182

Sato, K. H.; Garcia, R. A.; Pires, S.; Ballot, J.; Mathur, S.; Mosser, B.; Rodriguez, E.; Starck, J. L. & Uytterhoeven, K.: Inpainting: A powerful interpolation technique for helio- and asteroseismic data. *eprint arXiv*, volume 1003;

p. 5178, 2010; p. 5178. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010arXiv1003.5178S&link_type=ABSTRACT. 192

Schechter, P.: An analytic expression for the luminosity function for galaxies. *The Astrophysical Journal*, volume 203; p. 297, 1976; p. 297. doi: 10.1086/154079. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1976ApJ...203..297S&link_type=GIF. 111

Schulz, J.: Diploma Thesis: Analyse von PET Daten unter Einsatz adaptiver Glaettungsverfahren. Humboldt-Universitaet zu Berlin, Institut fuer Mathematik, 2006. 41, 178

Sekhon, J. & Mebane, W.: Genetic optimization using derivatives. *Political Analysis*, volume 7, no. 1; p. 187, 1998; p. 187. URL http://pan.oxfordjournals.org/content/7/1/187.short. 186

Sérsic, J.: Atlas de galaxias australes. *Cordoba*, 1968. URL http://adsabs.harvard.edu/abs/1968adga.book.....S. 125, 173

Sérsic, J. L.: Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Boletin de la Asociación Argentina de Astronomía*, volume 6; p. 41, 1963; p. 41. URL http://adsabs.harvard.edu/abs/1963BAAA....6...41S. 173

Shanno, D.: Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, volume 24, no. 111; pp. 647, 1970; pp. 647. URL http://www.jstor.org/stable/2004840. 186

Short, A.: Charge Distortion Model 02 (CDM02). *GAIA-CH-TN-ESA-AS-015-1*, 2009. 46, 193

Simard, L.: GIM2D: an IRAF package for the Quantitative Morphology Analysis of Distant Galaxies. *Astronomical Data Analysis Software and Systems VII*, volume 145; p. 108, 1998; p. 108. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1998ASPC..145..108S&link_type=ABSTRACT. 172

Simien, F. & de Vaucouleurs, G.: Systematics of bulge-to-disk ratios. *The Astrophysical Journal*, volume 302; p. 564, 1986; p. 564. doi:10.1086/164015. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1986ApJ...302..564S&link_type=ABSTRACT. 113

Sivanandam, S. & Deepa, S.: *Introduction to Genetic Algorithms*. Springer Verlag, 2007. ISBN 354073189X. 184, 186

Skiena, S.: *The Algorithm Design Manual*. Springer Verlag London Limited, 2nd edition, 2008. 60

Slezak, E. & Mignard, F.: A realistic QSO catalogue for the Gaia Universe Model. *GAIA-C2-TN-OCA-ES-001-1*, 2007. 23

Smith, K. T.; Solmon, D. C. & Wagner, S. L.: Practical and mathematical aspects of the problem of reconstructing objects from radiographs. *Bulletin of the American Mathematical Society*, volume 83, no. 6; pp. 1227, 1977; pp. 1227. URL http://www.ams.org/mathscinet/search/publications.html?pg1=MR&s1=MR0490032. 39

Springel, V.; White, S. D. M.; Jenkins, A.; Frenk, C. S.; Yoshida, N.; Gao, L.; Navarro, J.; Thacker, R.; Croton, D.; Helly, J.; Peacock, J. A.; Cole, S.; Thomas, P.; Couchman, H. et al: Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, volume 435; p. 629, 2005; p. 629. doi:10.1038/nature03597. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2005Natur.435..629S&link_type=ABSTRACT. 124

Starck, J.-L.; Bijaoui, A.; Valtchanov, I. & Murtagh, F.: A combined approach for object detection and deconvolution. *Astronomy & Astrophysics Supplement*, volume 147; p. 139, 2000; p. 139. doi:10.1051/aas:2000293. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2000A%26AS..147..139S&link_type=ABSTRACT. 80

Surdej, J.; Claeskens, J.-F.; Smette, A.; Damerdji, Y.; Finet, F.; Poels, J. & Wertz, O.: Discovery of bright multiple imaged quasars with Gaia. *ExtraGalactic Science with Gaia 2010*, pp. 1–15, 2010; pp. 1. URL http://www-n.oca.eu/rousset/EGSG/. 6

Teixeira, R.; Ducourant, C.; Chauvin, G.; Krone-Martins, A.; Song, I. & Zuckerman, B.: SSSPM J1102-3431 brown dwarf characterization from accurate proper motion and trigonometric parallax. *Astronomy & Astrophysics*, volume 489; p. 825, 2008; p. 825. doi:10.1051/0004-6361:200810133. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2008A%26A...489..825T&link_type=ABSTRACT. 202

Teixeira, R.; Ducourant, C.; Chauvin, G.; Krone-Martins, A.; Bonnefoy, M. & Song, I.: Kinematic analysis and membership status of TWA22 AB. *Astronomy & Astrophysics*, volume 503; p. 281, 2009; p. 281. doi:10.1051/0004-6361/200912173. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2009A%26A...503..281T&link_type=ABSTRACT. 202

Tikhonov, A. N.: *Méthodes de résolution de problèmes mal posés*. Éditions Mir, 1976. 50

Tipping, M. E.: Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research 1*, volume 1; pp. 211, 2001; pp. 211. URL http://portal.acm.org/citation.cfm?id=944741. 151

Toft, P.: *The Radon Transform, Theory and Implementation*. Tese de Doutorado. Department of Mathematical Modelling, Section for Digital Signal Processing, Technical Univeristy of Denmark, 1996. URL http://eivind.imm.dtu.dk/publications/PhD_thesis/PeterToft_PhD_thesis_1.ps.gz. 37, 41

Touma, H.: Synthetic Aperture Technic in Astronomy Using Slit Aperture Telescope. *Astrophysics and Space Science*, volume 258; p. 47, 1997; p. 47. doi:10.1023/A:1001703602866. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1997Ap%26SS.258...47T&link_type=ABSTRACT. 38

Toussaint, G.: Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining. *International Journal of Computational Geometry and Applications*, volume 15, no. 2; pp. 101, 2005; pp. 101. URL http://www.ams.org/mathscinet/search/publications.html?pg1=MR&s1=MR2137722. 89

Tukey, J. W.: *Exploratory Data Analysis*. Addison-Wesley, 1977. 77

Turnshek, D. A.; Lupie, O. L.; Rao, S. M.; Espey, B. R. & Sirola, C. J.: Hubble Space Telescope Observations of the Gravitationally Lensed Cloverleaf Broad Absorption Line QSO H1413+1143: Imaging. *The Astrophysical Journal*, volume 485; p. 100, 1997; p. 100. doi:10.1086/304395. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1997ApJ...485..100T&link_type=ABSTRACT. 6

Turon, C. & Arenou, F.: The Hipparcos Catalogue: 10th anniversary and its legacy. *A Giant Step: from Milli- to Micro-arcsecond Astrometry*, volume 248; p. 1, 2008; p. 1. doi:10.1017/S1743921308018516. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2008IAUS..248....1T&link_type=ABSTRACT. 9, 14

Turon, C.; Crézé, M.; Egret, D.; Gómez, A.; Grenon, M.; Jahreiss, H.; Réquième, Y.; Argue, A. N.; Bec-Borsenberger, A.; Dommanget, J.; Mennessier, M. O.; Arenou, F.; Chareton, M.; Crifo, F. et al: The HIPPARCOS input catalogue. *ESA SP-1136 (7 volumes)*, volume 1136, 1992. URL http://adsabs.harvard.edu/abs/1992ESASP1136.....T. 11, 12

Van den Bergh, S.: *Galaxy morphology and classification*. Cambridge University Press, 1998. 124

van den Bergh, S.; Abraham, R. G.; Ellis, R. S.; Tanvir, N. R.; Santiago, B. X. & Glazebrook, K. G.: A Morphological Catalog of Galaxies in the Hubble deep Field. *The Astronomical Journal*, volume 112; p. 359, 1996; p. 359. doi:10.1086/118020. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1996AJ....112..359V&link_type=ABSTRACT. 143, 166

van der Kruit, P. C. & Searle, L.: Surface photometry of edge-on spiral galaxies. I - A model for the three-dimensional distribution of light in galactic disks. *Astronomy & Astrophysics*, volume 95; p. 105, 1981; p. 105. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1981A%26A....95..105V&link_type=ABSTRACT. A&AA ID. AAA029.158.016. 174

van Leeuwen, F.: *Hipparcos, the New Reduction of the Raw Data*. Springer Verlag, 2007. 14

van Leeuwen, F.: The Hipparcos catalog. Commentary on: Perryman M. A. C., Lindegren L., Kovalevsky J., et al., 1997, A&A, 323, L49. *Astronomy & Astrophysics*, volume 500; p. 505, 2009a; p. 505. doi:10.1051/0004-6361/200912202. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2009A%26A...500..505V&link_type=EJOURNAL. 14

van Leeuwen, F.: Parallaxes and proper motions for 20 open clusters as based on the new Hipparcos catalogue. *Astronomy & Astrophysics*, volume 497; p. 209, 2009b; p. 209. doi:10.1051/0004-6361/200811382. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2009A%26A...497..209V&link_type=ABSTRACT. 14

Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer Verlag, 2000. ISBN 0387987800. 151

Venables, W. & Ripley, B.: *Modern Applied Statistics with S*. Springer Verlag, 2002. ISBN 0387954570. 188

Véron-Cetty, M.-P. & Véron, P.: A catalogue of quasars and active nuclei: 12th edition. *Astronomy & Astrophysics*, volume 455; p. 773, 2006; p. 773. doi:10.1051/0004-6361:20065177. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2006A%26A...455..773V&link_type=ABSTRACT. 23

VividSolutions: JTS Topology Suite Technical Specifications (v. 1.4). 2003. URL http://www.vividsolutions.com/jts/bin/JTS%20Technical%20Specs.pdf. 60

VividSolutions: JTS Topology Suite v. 1.8.1. 2006. URL http://www.vividsolutions.com/jts/jtshome.htm. 60

Whitcher, B.: Package waveslim. 2010. URL http://www2.imperial.ac.uk/~bwhitche/. 182

White, S. D. M. & Frenk, C. S.: Galaxy formation through hierarchical clustering. *The Astrophysical Journal*, volume 379; p. 52, 1991; p. 52. doi:10.1086/170483. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1991ApJ...379...52W&link_type=ABSTRACT. 124

Williams, R. E.; Blacker, B.; Dickinson, M.; Dixon, W. V. D.; Ferguson, H. C.; Fruchter, A. S.; Giavalisco, M.; Gilliland, R. L.; Heyer, I.; Katsanis, R.; Levay, Z.; Lucas, R. A.; McElroy, D. B.; Petro, L. et al: The Hubble Deep Field: Observations, Data Reduction, and Galaxy Photometry. *The Astronomical Journal v.112*, volume 112; p. 1335, 1996; p. 1335. doi:10.1086/118105. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1996AJ....112.1335W&link_type=ABSTRACT. 48, 143

Wolpert, D.; Macready, W.; Center, I. & Jose, C. S.: No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, volume 1, no. 1; pp. 67, 1997; pp. 67. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=585893&tag=1. 180

Zacharias, N.; Urban, S. E.; Zacharias, M. I.; Wycoff, G. L.; Hall, D. M.; Monet, D. G. & Rafferty, T. J.: The Second US Naval Observatory CCD Astrograph Catalog (UCAC2). *The Astronomical Journal*, volume 127; p. 3043, 2004; p. 3043. doi:10.1086/386353. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2004AJ....127.3043Z&link_type=ABSTRACT. 14

Zhang, S. N.; Fishman, G. J.; Harmon, B. A. & Paciesas, W. S.: Imaging high-energy astrophysical sources using Earth occultation. *Nature*, volume 366; p. 245, 1993; p. 245. doi:10.1038/366245a0. URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1993Natur.366..245Z&link_type=ABSTRACT. 38