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

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

TESE DE DOUTORADO

Apresentada por
Alberto Garcez de Oliveira KRONE-MARTINS

AMPLIANDO HORIZONTES DA MISSÃO ESPACIAL GAIA GRAÇAS À ANÁLISE DE OBJETOS EXTENSOS

preparada em regime de co-tutela entre a
Universidade de São Paulo e a Université de Bordeaux I

defendida na Universidade de São Paulo

<i>Orientadores :</i>	Ramachrisna TEIXEIRA	–	IAG-Universidade de São Paulo
	Caroline SOUBIRAN	–	LAB-Université de Bordeaux I
	Christine DUCOURANT	–	LAB-Université de Bordeaux I
<i>Examinadores :</i>	Ronaldo DE SOUZA	–	IAG-Universidade de São Paulo
	Reinaldo DE CARVALHO	–	INPE-Inst. Nac. de Pesquisas Espaciais
<i>Rapporteurs :</i>	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.
in *Corpo*, Carlos Drummond de Andrade.

O astrônomo, mesmo na incapacidade de seu pequeno telescópio,
vê o passado das estrelas e o futuro da humanidade.
in *O taikonauta, o astrônomo e o espaço*, Elisa Andrade Buzzo.

Agradecimentos

Em primeiro lugar agradeço à Elisa, que esteve ao meu lado nestes últimos anos escutando diversos: “quando terminar a tese”. Obrigado pelo seu amor e paciência.

Aos meus pais Alberto e Maria Auxiliadora e minha irmã Isabela, pelo exemplo de esforço e dedicação, pelo amor, apoio e incentivo incondicional para que sempre perseguisse o caminho que escolhesse e lutasse por ele—não importando se às vezes fosse necessário renunciar ao descanso.

Aos meus orientadores Rama e Christine pelo exemplo, amizade e por abrirem novas portas, expandindo meus horizontes dentro do universo da Astronomia e da Ciência. Também agradeço por serem compreenderem com minha necessidade de manter amplos interesses, e pela sabedoria em colocar o barco de volta ao rumo sempre que necessário. E principalmente por me lembrarem que em algumas ocasiões precisamos diminuir o ritmo e deixar idéias assentarem.

Aos meus sogros Eliseu e Célia, e a toda minha família pelo apoio e pela confiança que sempre recebi, e por compreenderem minha ausência em muitos momentos.

À Caroline pelas discussões científicas e por me orientar e direcionar em um dos caminhos paralelos que tracei, estendendo meus conhecimentos em Astronomia.

Ao Xavier Luri, ao Eduard Masana e a Carine Babusiaux por me iniciarem nos simuladores do Gaia e ao programa Alfa-LENAC por me possibilitar trabalhar na *Universitat de Barcelona*. Ao Albert Bijaoui, por conversas inspiradoras. À Guylaine, ao Benoît e à equipe do CNES pela sua compreensão de que usualmente astrônomos não estarão sempre 100% *compliant*, e ao Christophe e ao Laurent por estarem nos ajudando a chegar lá. Ao Jean-François, pelas discussões sobre algoritmos e conversas. À Diana Harrison pelas discussões sobre os códigos do CU5. Ao Dimitris pelas conversas em Atenas, e ao Panos pela amizade, bom humor e pelo MAGIL. Ao Didier e ao Nelson por terem garantido que tudo estava correndo bem durante o desenvolvimento deste projeto.

Ao Gustavo Medina Tanco, por me iniciar na Astronomia profissional e pela orientação do mestrado, além de ser um grande exemplo de competência.

Aos professores e astrônomos de São Paulo e de Bordeaux por terem me ensinado muitas áreas diferentes da Astronomia, dentre outras coisas, talvez sem sequer perceberem. Obrigado Paulo, Sylvio, Tatiana, Jorge, Boczek, Laerte, Augusto, Ronaldo, Roberto, Sandra, Alex, João, Amaury, Jane, Vera, Beatriz, Bete, Gastão, Jacques, Zulema, Sílvia, Ademir, Michel, Frack, Sean, Valentine, Didier, Nathalie, Fabrice, Patrick. Um agradecimento especial ao Rama, à Christine e ao Michel pela cuidadosa revisão das versões desta tese tanto em Português quanto em Francês.

A todos meus amigos do Colégio Etapa, em especial ao Fabio, Caio, Ulisses, Clayton, Goulart, Ricardo, Leo, Thamy, Sissy, e aos meus amigos da USP e de Bordeaux pelas conversas nos corredores, nos almoços, nos cafés, e em nossas salas. Essa tese não seria a mesma sem discussões astronômicas e computacionais, algumas das quais com certeza se tornarão trabalhos no futuro, e também sem conversas sobre

“a vida, o universo e tudo mais”. Obrigado 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, Marina, Antoine, Arnaud, Geraldine, Aurelie, Thibauld, Mathieu, Cécile, Pierre, Yann, Anne-Sophie, Ming, Ileana, Adrian, Luis, Ugo.

A todo CASP – Clube de Astronomia de São Paulo e meus amigos na divulgação da Astronomia, em especial André, Tasso, Lucas, Livia, Raquel, Bia, Denis, Chicão, Tony, Dorival e Márcio, por incentivarem as pessoas a conhecer e aprender cada vez mais sobre o Universo em que vivem, abrindo janelas para que elas se inspirem a transformar o Brasil em um lugar melhor e o mundo como um todo.

Ao Departamento de Astronomia do IAG da Universidade de São Paulo e ao *Laboratoire d'Astrophysique de Bordeaux* da *Université de Bordeaux* I pela infraestrutura e pelo suporte recebido por meio das equipes técnicas e de secretariado. Em especial, obrigado Marina, Marcel, Conceição, Regina, Cida, Cécile, Annick, Ulisses, Sylvie e Nadège. Ao INCT-A pela máquina Gina.

À FAPESP pela bolsa de doutorado (2006/04251-4). À colaboração CAPES-Cofecub pela bolsa no exterior (BEX0812/07-2). Ao CNES, ao CNRS, à ELSA (FP6-EU) e à *Action Spécifique Gaia* por financiamentos em missões e participações em conferências. Finalmente, à ESA por essa incrível missão espacial chamada Gaia.

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 otimizado 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.

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 côté, 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.

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.

Conteúdo

1	Introdução	1
1.1	Objetivo	2
1.2	Como trabalhar com esses objetos	2
1.3	O que poderemos aprender	4
1.4	Astrometria e o Hipparcos	8
1.4.1	O instrumento, a missão	9
1.4.2	Redução dos dados	12
1.4.3	O Tycho	14
1.4.4	O pós-missão	14
1.5	A missão Gaia	15
1.5.1	O instrumento	16
1.5.2	O DPAC – <i>Data Processing and Analysis Consortium</i>	20
1.5.3	Simuladores	22
1.5.4	Redução de dados	26
1.5.5	Redução de dados de objetos problemáticos	27
1.6	Estrutura do texto	31
2	Reconstrução de imagens	33
2.1	Introdução	34
2.2	Um <i>toy model</i> para reconstrução	35
2.2.1	Reconstrução pela transformada de Radon	37
2.3	Reconstrução de imagens para o Gaia	45
2.3.1	Algoritmos de reconstrução	47
2.3.2	QuickStack	47
2.3.3	Drizzle	48
2.3.4	ShuffleStack	49
2.3.5	BinOutliers	50
2.3.6	Regularização de Tikhonov	50
2.3.7	Clean & Cleanest	52
2.3.8	Exemplos de reconstruções	53
2.4	Cobertura espacial e angular das reconstruções	56
2.4.1	Simulação dos trânsitos	58
2.4.2	União de polígonos	60
2.4.3	Cálculo na nuvem	61
2.4.4	Resultados	61
2.5	Conclusões	65

3	Fontes Secundárias	67
3.1	Introdução	68
3.2	Análise semi-numérica – Catálogos	69
3.2.1	Procedimento de Análise	69
3.2.2	Resultados	72
3.2.3	Comentários	73
3.3	Análise semi-numérica – Simulações	73
3.3.1	Procedimento de análise	74
3.3.2	Resultados	75
3.4	Segregação Educada de Imagem - EIS	79
3.4.1	Princípios	79
3.4.2	Modo com Limiar-Simples	80
3.4.3	Modo com Limiar-Múltiplo	83
3.4.4	Modo Educado	87
3.4.5	Testes e resultados	92
3.5	Conclusões	100
4	Simulações de Galáxias no modelo de Universo do Gaia	103
4.1	Introdução	104
4.2	Simulação de catálogos de galáxias	104
4.2.1	Princípio de funcionamento	105
4.2.2	Modelo de objetos	106
4.2.3	Método de simulação	111
4.2.4	Validação da implementação Java	115
4.3	Simulação de imagens de objetos extensos – MAGIL	116
4.3.1	Exemplos de imagens simuladas	117
4.4	Estimativas do número de galáxias observáveis	118
4.4.1	Estimativa baseada em extrapolação do Hubble-MDF	120
4.5	Conclusões	121
5	Análise e classificação de Imagens de Galáxias	123
5.1	Introdução	124
5.2	O espaço CASGM20	126
5.2.1	Testes com dados da literatura—PCA	136
5.3	Implementação para o Gaia – CASGM20	139
5.3.1	Testes com galáxias do catálogo de Frei	141
5.3.2	Cálculo de CASGM20 em imagens do <i>Hubble Deep Field</i>	143
5.3.3	Testes com simulações	147
5.4	Classificação e Support Vector Machines	151
5.5	Implementação para o Gaia – SVM	156
5.5.1	Validação com gaussianas	158
5.5.2	Testes com galáxias do catálogo de Frei	162
5.5.3	Testes com simulações GIBIS	163
5.6	Aplicação aos dados do <i>Hubble Deep Field</i>	166

5.7	Conclusões	168
6	Estimativa de parâmetros morfológicos	171
6.1	Introdução	172
6.1.1	Perfis de brilho	173
6.1.2	Ajuste de perfis no espaço de Radon	174
6.2	Construção do modelo	176
6.3	Optimização dos parâmetros	180
6.3.1	Estimativas iniciais	181
6.3.2	Algoritmos Genéticos	184
6.3.3	BFGS	186
6.4	Testes com simulações	188
6.5	Aprimoramentos e possibilidades	192
6.6	Conclusões	195
	Conclusões	197
	Considerações finais	201
A	Módulo inverso	205
A.1	Definição	205
B	Aglomerados abertos	207
	Bibliografia	223

CAPÍTULO 1

Introdução

“*L’Astronomie, (...) c’est elle qui nous a fait une âme capable de comprendre la nature.*” H. Poincaré¹

Conteúdo

1.1	Objetivo	2
1.2	Como trabalhar com esses objetos	2
1.3	O que poderemos aprender	4
1.4	Astrometria e o Hipparcos	8
1.4.1	O instrumento, a missão	9
1.4.2	Redução dos dados	12
1.4.3	O Tycho	14
1.4.4	O pós-missão	14
1.5	A missão Gaia	15
1.5.1	O instrumento	16
1.5.2	O DPAC – <i>Data Processing and Analysis Consortium</i>	20
1.5.3	Simuladores	22
1.5.4	Redução de dados	26
1.5.5	Redução de dados de objetos problemáticos	27
1.6	Estrutura do texto	31

Neste Capítulo introdutório será brevemente apresentada a questão central, motora deste trabalho bem como o contexto onde se situa e onde se desenvolveu.

Encontraremos uma descrição o mais detalhada possível da Astrometria espacial, incluindo seu início com a missão Hipparcos, que auxiliará a melhor compreender a missão espacial Gaia na qual este trabalho está inserido. Esta missão é sem dúvida um dos mais importantes esforços da Astronomia moderna e a maior e mais complexa empreitada da Astrometria em toda a história.

O núcleo da presente tese trata justamente do processamento de objetos celestes observados pelo satélite mas com soluções astrométricas problemáticas, ou em termos mais diretos, buscamos aqui mostrar que será possível fazer ciência com parte dos dados que poderiam ser rejeitados pelo processamento de dados do Gaia.

¹“A Astronomia, (...) foi ela que nos tornou uma alma capaz de compreender a natureza.”, in *La Valeur de La Science*, 1905.

1.1 Objetivo

Um trabalho científico visa entre outras coisas, encontrar respostas ou soluções para um problema em aberto, contribuir para a solução de um problema específico, etc. No trabalho apresentado neste documento, nos propusemos a expandir os horizontes da missão espacial Gaia, respondendo e/ou encontrando meios para resolver a seguinte questão:

É possível fazer ciência com objetos observados pelo satélite Gaia cujas soluções astrométricas serão problemáticas pelo fato de possuírem algum tipo de emissão extensa?

Como tentamos deixar claro neste texto, ao final deste trabalho de tese nossa conclusão é que sim, é possível aproveitar muitos daqueles objetos cujas soluções astrométricas previstas serão ruins, para fazer ciência. Mais do que isso, os resultados aqui obtidos deverão abrir uma nova janela, ampliar os horizontes, para o estudo de milhões de objetos que possuem estruturas ainda desconhecidas.

O satélite Gaia não foi planejado para estudar por exemplo, objetos extensos, mas ainda assim ele os observará. Essas fontes astronômicas, cujo número poderá ser de poucos milhares a dezenas de milhões, corriam o risco de serem simplesmente ignoradas. A razão para tal é justamente por não se encaixarem no escopo da missão Gaia: não são fontes pontuais e as atenções da comunidade estavam até então, quase que exclusivamente voltadas para o estudo das estrelas.

Os objetos astronômicos que nos propusemos a tratar neste trabalho estão sempre associados a algum tipo de emissão extensa, sendo que esta pode ser de dois tipos distintos: intrínsecas ou extrínsecas.

As fontes intrínsecas são aquelas que normalmente chamamos extensas tendo como um bom exemplo as galáxias não resolvidas em estrelas individuais.

As fontes extrínsecas por sua vez, resultam de uma “emissão extensa” provocada pela proximidade angular de duas fontes.

O interesse em recuperar esses objetos para fazer ciência reside no fato de que embora o satélite Gaia não tenha sido planejado para tal, suas observações contêm informações valiosíssimas, por exemplo sobre a estrutura de uma grande quantidade de galáxias angularmente pequenas. No caso de estrelas múltiplas e projeções de objetos na mesma linha de visada de um objeto Gaia esses objetos serão resolvidos como dificilmente o seriam a partir do solo sem óptica adaptativa.

1.2 Como trabalhar com esses objetos

Em geral, utilizam-se imagens para se estudar este tipo de objetos e tal como seu antecessor, o satélite Hipparcos, o Gaia não foi concebido com a finalidade de fazer imagens mas sim de medir ângulos entre os diversos objetos observados. Esta dificuldade, pode entretanto, ser contornada por uma característica das missões astrométricas espaciais: para medir esses ângulos com grandes precisões é necessário

que um objeto seja observado em diferentes grandes círculos, ou seja, observações a partir de diferentes direções de visada. Justamente, essas observações a partir de diferentes direções de visada são suficientes para se reconstruir imagens bi-dimensionais. Como isto pode ser feito e as estratégias atualmente consideradas para a aplicação aos dados da missão Gaia, possibilitando o estudo de objetos extensos, serão apresentadas no Capítulo 2.

Durante a missão Hipparcos (que será descrita com mais detalhes na seção 1.4), esse tipo de análise já foi realizado, e um sinal bi-dimensional das fontes observadas foi reconstruído com a utilização dos dados de um de seus detectores mesmo que para este satélite as observações fossem completamente unidimensionais—um exemplo destas reconstruções pode ser visto na Figura 1.1. De forma distinta de seu antecessor, o Gaia possui uma noção de “pixel” em alguns dos dados que ele deve transferir para a Terra, o que facilitará a analogia desse sinal bi-dimensional reconstruído com uma imagem do objeto.

Figura 1.1: Imagens da estrela dupla HIP76566 reconstruídas a partir do empilhamento de dados dos *Star Mappers* do satélite Hipparcos que foram produzidos para a criação do catálogo *Tycho Double Star Catalogue*. (De [Fabricius, 2007](#))

Com as imagens reconstruídas do Gaia será possível observar todos os objetos do céu até magnitude $G \sim 20^2$ em escalas de ~ 50 mas/pixel e com uma *Point Spread Function* de ~ 180 mas. Caso seja possível reconstruir imagens confiáveis com os dados desse satélite, isso representaria imagens significativamente melhores do que aquelas que podem ser obtidas a partir de quaisquer sítios terrestres sem o uso de

²Magnitudes G são obtidas sem filtro, cobrindo a região entre 330nm e 1000nm.

óptica adaptativa.³ Como comparação, o *seeing* da maior parte dos observatórios profissionais no solo gira em torno de ~ 1000 mas, podendo atingir ~ 600 mas em localidades excepcionais⁴—fato que pode ser visto na Figura 1.2—e com o Gaia teremos essas imagens para o céu inteiro.

Figura 1.2: Tamanhos das *Point Spread Functions* para diversos telescópios e sítios. Significado das siglas: 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 na direção de maior resolução do Gaia, *Gaia's worst*—PSF na direção de menor resolução. Um algoritmo de reconstrução de imagens perfeito deveria ser capaz de reconstruir imagens sempre com a maior resolução do Gaia. (Adaptada de Lawrence et al, 2004)

1.3 O que poderemos aprender

Uma vantagem direta de analisar as imagens reconstruídas é que ao combinar diversas observações de uma mesma região do céu o ruído dessa “observação combinada” é reduzido a cada nova observação adicionada—principalmente para a área na qual ocorre intersecção entre diferentes observações, mas dependendo do algoritmo de reconstrução, para outras regiões da imagem também.

Esse ruído mais baixo, permite com que fontes mais fracas do que aquelas que poderiam ser detectadas separadamente em cada observação possam ser detectadas na imagem reconstruída, elevando assim a magnitude limite do instrumento nas regiões do céu em que foram transferidos para a Terra.⁵

³Não estamos considerando o uso de óptica adaptativa uma vez que esta ainda não é utilizada para grandes levantamentos, como o Gaia.

⁴Resultados recentes de Lawrence et al (2004) indicam que o melhor *seeing* do planeta encontra-se na Antártida (em um local chamado de *Dome C*), onde atinge um valor mediano de ~ 270 mas.

⁵Como será visto na seção 1.5.1, não serão todos os pixels do plano focal do Gaia que serão

Durante a produção do catálogo Tycho-2, mapas bi-dimensionais foram originalmente reconstruídos com os dados de um dos detectores do Hipparcos (chamado de *Star Mapper*) para possibilitar inspeções de soluções duvidosas, e não para realizar medidas (Høg et al, 2000a). Contudo, posteriormente Fabricius et al (2002) utilizou mapas reconstruídos para descobrir 13.251 novas estrelas duplas ainda não conhecidas, originando grande parte do *Tycho Double Star Catalogue*.

Estudos sobre a frequência de sistemas múltiplos são importantes para a compreensão dos mecanismos de formação e evolução estelar (Burgasser et al, 2007), e nestas imagens reconstruídas do Gaia milhões destes sistemas poderão ser observados e descobertos, e em particular sistemas com separações angulares pequenas.

O menor ruído destas imagens permite que sejam analisados casos de múltiplas estrelas que se encontram na mesma linha de visada, e que serão mal resolvidos, ou não resolvidos, a partir dos dados originais. Dentre estes casos, poderão ser observados todos os tipos de objetos astronômicos.

Uma classe com interesse científico particular que poderá ser encontrada, são os objetos chamados de anãs marrons. Estes são objetos de baixa massa, que não são capazes de manter a fusão de hidrogênio de forma estável e que fundem deutério (ex: Basri, 2000). Existem diversos cenários propostos para a formação de anãs marrons, sendo uma possibilidade a formação normal como qualquer outra estrela a partir do colapso de nuvens moleculares, ou outro cenário de que estes objetos são embriões ejetados de sistemas proto-estelares múltiplos (ex: Reipurth & Clarke, 2001; Bate et al, 2002). Estudos de multiplicidade de anãs marrons (ex. Ahmic et al, 2007) são importantes para determinar qual é o cenário mais plausível para a sua formação, e poderão ser realizados a partir dos dados provenientes de análises de imagens reconstruídas do Gaia.

Objetos de tipos espectrais L e T (Kirkpatrick, 2005; Burgasser et al, 2006), ainda mais frios do que os de tipo M com $T_{eff} < 2500K$, talvez possam ser encontrados nestas imagens, auxiliando a aumentar sua amostra e possibilitando elucidar questões tão básicas sobre sua natureza como se são estrelas ou anãs marrons.

A partir destas imagens também poderão ser resolvidas projeções de estrelas, ou objetos binários e múltiplos de todos os tipos espectrais em aglomerados estelares, por exemplo, possibilitando estudos mais detalhados destes aglomerados, inclusive de sua caracterização e evolução dinâmica.⁶ Estudar as imagens reconstruídas nestes casos possibilitaria resolver problemas de contaminação dos diagramas cor-magnitude devido à multiplicidade e possibilitaria, por exemplo, estudos mais detalhados do déficit de anãs brancas em aglomerados (Fellhauer et al, 2003).

Ainda sobre anãs brancas, estudos da multiplicidade destas estrelas com anãs vermelhas, fornecem informações sobre a evolução de objetos de baixa massa na presença de envelopes estelares gigantes e sobre a função de massa inicial na parte de baixa massa da seqüência principal (Farihi et al, 2010), e poderão ser realizados com os objetos observados em imagens reconstruídas. Estrelas anãs brancas pouco

enviados ao solo, ocorrendo uma seleção a bordo do satélite.

⁶Atualmente isso é principalmente estudado por meio de simulações, por exemplo Parker & Goodwin (2010) e Ernst et al (2010).

luminosas, mesmo além do limite de detecção do satélite também poderão ser encontradas com o tratamento proposto—por exemplo aquelas pertencentes ao Halo de nossa Galáxia.

Um último exemplo que o estudo destes objetos extensos extrínsecos possibilitará é que dentre suas imagens reconstruídas poderão ser encontrados casos de quasares que estão sofrendo o fenômeno de lentes gravitacionais, pois múltiplas miragens destes objetos devem aparecer próximas entre si se perturbando mutuamente. Um estudo analítico realizado por [Surdej et al \(2010\)](#) mostrou que dentre as imagens do Gaia poderão ser encontrados até 3500 quasares sofrendo tal efeito até $V \sim 21$ (imagens reconstruídas podem atingir uma magnitude limite maior). Um exemplo deste tipo de fenômeno é o quasar H1413 + 117 (Figura 1.3), descoberto em [Magain et al \(1988\)](#), e que dado sua magnitude ($V \sim 17$) deve ser observado pelo Gaia.

Figura 1.3: As múltiplas imagens do quasar H1413 + 117 ($V \sim 17$). Esquerda: imagem da descoberta, obtida pelo telescópio de 2.2m do ESO em 1988 ([Magain et al, 1988](#)). Centro: imagem obtida pelo HST em 1994 ([Turnshek et al, 1997](#), obtida em <http://chandra.harvard.edu/photo/2004/h1413/>). Direita: imagem do HST deconvolvida, revelando a galáxia que criou a lente no centro além dos arcos gravitacionais ([Chantry & Magain, 2007](#)).

Os exemplos de estudos citados anteriormente deverão ser possibilitados devido aos desenvolvimentos apresentados nos Capítulos 2, e 3 desta tese.

Os Capítulos restantes tratarão da outra classe de objetos extensos cuja a análise de imagens reconstruídas possibilitará a obtenção informações científicas: os objetos com estrutura morfológica extensa intrínseca. Dentre estes objetos, os casos mais numerosos serão certamente as galáxias cuja população estelar não pode ser resolvida pelo Gaia.⁷

Devido a uma característica do sistema de detecção de objetos a bordo do satélite, existe um limite máximo do tamanho angular do objeto para transferência dos dados, de ~ 700 mas.⁸ Deste modo, o Gaia deverá observar um grande número de galáxias pequenas próximas, o que possibilitará obtermos informações importantes que podem corroborar com ou invalidar modelos cosmológicos. Até mesmo uma

⁷Outros casos serão estrelas centrais de nebulosas planetárias, regiões HII, etc.

⁸Veremos mais sobre isso na seção 1.5.1.1. Mas isso não equivale ao raio da maior galáxia “não resolvida” que poderá ter, mas provavelmente ao tamanho de sua região mais brilhante.

simples contagem destas galáxias é uma questão que permite comparações com as densidades deste tipo de galáxia previstas por simulações numéricas utilizando, por exemplo, o modelo Λ CDM.⁹ Além disso, estudos como [Benson et al \(2007\)](#) podem ser realizados, possibilitando a determinação da função de massa de buracos negros centrais de galáxias no universo próximo.

Galáxias maiores, porém mais distantes e mais brilhantes também deverão ser observadas, permitindo estudos de sua morfologia em alta resolução em todo o céu. Até o momento apenas um estudo observacional em maior escala foi realizado para este tipo de galáxia, o *Hubble Medium Deep Survey*, ou MDS (ex. [Griffiths et al, 1994](#)). E embora [Abraham et al \(1996b\)](#) tenha utilizado dados do MDS para análises de características morfológicas destas galáxias, apenas uma pequena porcentagem da esfera celeste foi coberta, contando com uma quantidade pequena de objetos mais brilhantes ($G \leq 20$) e logo com maiores incertezas estatísticas.

Embora a missão espacial Gaia não seja primariamente voltada à observação de objetos extragalácticos, uma janela ainda não mapeada será aberta por ela, pois seus dados poderão originar o primeiro levantamento de céu completo de galáxias com resoluções espaciais de ~ 180 mas, possibilitando observar centenas de milhares ou até mesmo milhões, de galáxias com a este nível de resolução.

Isso permitirá detalharmos a morfologia destes objetos, primeiro a partir de taxonomia—separação em classes morfológicas—e então a partir da decomposição em, e/ou obtenção de parâmetros de, seus componentes estruturais principais, como os bojos e os discos. Estas galáxias poderão ser estudadas individualmente, servindo também como alvos para observações dedicadas futuras. Além disso, seus dados também poderão ser analisados de forma global (como feito por [Allen et al, 2006](#)), possibilitando por exemplo contribuir com estudos da evolução da morfologia destes objetos, e de estudos sobre a formação, evolução e sustentação dos componentes estruturais, como por exemplo foi realizado para galáxias elípticas, bojos e pseudo-bojos por ([Gadotti, 2009](#)).

Uma convergência entre dois levantamentos de céu inteiro planejados para o futuro próximo que auxiliará nestes estudos, é o fato do corte na transferência de dados do Gaia para objetos com emissão central maior do que ~ 700 mas ser complementado pelo *seeing* mediano do *Large Synoptic Survey Telescope* (o maior levantamento do céu do sul, previsto para 2018): justo quando o Gaia deixa de transferir dados para estes objetos o LSST poderá continuar provendo os dados dos estudos morfológicos a partir de observações do solo, pois sendo seu *seeing* menor do que o tamanho dos objetos, informações estruturais poderão ser recuperadas.

Veremos como essas galáxias que comentamos podem ser simuladas para observações com o Gaia no Capítulo 4, e como, a partir dos dados dessa missão, elas podem ser classificadas morfológicamente no Capítulo 5 e ter seus componentes estruturais analisados com maiores detalhes no Capítulo 6.

⁹Este é o modelo de Universo atualmente apoiado pela maior parte dos cosmólogos. Segundo ele, o Universo é regido pela constante cosmológica que é algo que o força a crescer de forma acelerada (Λ) e possui matéria escura fria (*Cold Dark Matter*). Para uma descrição mais detalhada, ver por exemplo [Freedman \(2004\)](#).

1.4 Astrometria e o Hipparcos

Como comentado anteriormente, este trabalho de tese está inserido no contexto da missão espacial Gaia. Mas para melhor compreender o contexto desta missão, dedicamos algumas linhas à sua antecessora, a missão Hipparcos, e a Astrometria de uma maneira geral.

Uma das ramificações mais antigas da Astronomia, e que durante um longo período da história desta ciência foi praticamente o sinônimo de sua vertente observacional, a Astrometria¹⁰ é o braço mais fundamental da Astronomia: é nela que praticamente todo o conhecimento astronômico está ancorado. Isso, pois as medidas astronômicas ou são puramente astrométricas ou são dependentes de alguma informação astrométrica para que possam ser transformadas em conhecimento físico: essa informação é a distância dos objetos no espaço. A distância é uma das principais questões tratadas por essa área da ciência.

Uma das questões mais delicadas da Astronomia são os erros nos parâmetros físicos obtidos a partir de medidas observacionais. Um dos componentes principais destes erros são originados das incertezas da determinação de distância dos objetos em estudo, pois incertezas em determinações de distâncias se propagam por muitos outros parâmetros. A base de praticamente todos os métodos de determinação de distância no universo é a paralaxe trigonométrica¹¹, que durante muito tempo se considerou uma das mais árduas buscas observacionais astronômicas—ela foi finalmente observada em estrelas por Bessel (1838a,b) na estrela 61 do Cisne.¹²

Contudo, até meados do século passado, existiam apenas algumas poucas estrelas com distância determinada por este método, sendo que a quantidade de estrelas cujas medidas de distância podiam ser consideradas confiáveis era ainda menor. Em vista deste cenário, era necessário aumentar a base sob a qual repousa a maior parte dos indicadores de distância.

A paralaxe é tão difícil de medir por ser um efeito periódico na variação da posição de um objeto celeste que é muito pequeno, pois as distâncias astronômicas são muito grandes, e os principais efeitos observacionais que dificultam a obtenção são efeitos causados pela atmosfera da Terra (como o *seeing* e a refração), variações no instrumento utilizado para realizar a medida, e variações do próprio instrumento na elaboração de levantamentos de paralaxes (necessários para, por exemplo, trabalhos de calibração de outros métodos). Mas estes problemas poderiam ser resolvidos caso o instrumento não estivesse na Terra e caso um único instrumento fosse utilizado para observar todos os objetos.

Foi então proposta para a Agência Espacial Européia (ESA) uma missão espacial astrométrica, desenvolvida a partir de um conceito inicial de Pierre Lacroute apresentado em 1967 durante a XIII Assembléia Geral da União Astronômica In-

¹⁰Do grego ἄστρον, estrela, e μέτρον, medida, logo “medida das estrelas”.

¹¹Mais especificamente, a paralaxe anual, que é causada pelo movimento da Terra ao redor do baricentro do sistema solar.

¹²Henderson publicou uma paralaxe para α do Centauro dois meses depois, e Struve para α da Lira (Vega) um ano após Bessel.

ternacional.¹³ Essa missão, chamada *High Precision Parallax Collecting Satellite*, ou como é mais conhecida Hipparcos, consistiu no primeiro observatório espacial dedicado à Astrometria, e foi aceita para vôo em 1980 com o objetivo primário de determinar parâmetros astrométricos para ~ 100.000 estrelas (posições, movimentos próprios e paralaxes) com precisões de 2 mas, inéditas até então.

Mas, após ser lançado, em 8 de Agosto de 1989, ocorreu um problema que não permitiu com que o Hipparcos atingisse a órbita geostacionária prevista inicialmente: o motor de apogeu simplesmente não disparou, deixando o satélite em uma órbita de transferência, que obrigava o instrumento a cruzar quatro vezes ao dia o cinturão de radiação de Van Allen. Ao confirmar que o motor de apogeu não dispararia nunca, as expectativas científicas da missão foram reduzidas para 5 – 10% dos objetivos originais (ESA, 1997), o que na melhor das hipóteses, teria a capacidade de confirmar que o conceito de astrometria espacial funcionaria. Entretanto, o satélite resistiu quatro anos, vindo a ser desligado apenas no dia 15 de Agosto de 1993.¹⁴

Os resultados obtidos após o processo de tratamento de dados superaram as expectativas, e dois catálogos foram produzidos: o Hipparcos, contendo 118.218 objetos com precisão mediana de 1 mas, além de dados para objetos duplos e múltiplos (incluindo soluções orbitais), e o catálogo Tycho, que contém mais de 10^6 de objetos, com precisões entre 7 e 25 mas. Ambos os catálogos também incluem informações fotométricas, com precisões que dependem da magnitude do objeto, variando entre 0.4 – 7 mmag para o Hipparcos e 10 – 60 mmag para o Tycho (ESA, 1997).

1.4.1 O instrumento, a missão

Os principais fatores pelo qual os dados do Hipparcos puderam ser tão precisos foram o fato dele observar a partir do espaço, características particulares de sua instrumentação e um grande controle dos erros associados às observações. De um modo geral, para a realização de trabalhos astrométricos existem três vantagens principais ao se observar os objetos de fora da atmosfera terrestre:

- a posição dos objetos deixa de ser afetada pelo *seeing* e por fenômenos de refração. O *seeing* é dificilmente corrigível em grandes ângulos,¹⁵ e a correção da refração é sempre limitada pelo nosso conhecimento da mutável atmosfera;
- o instrumento se move de forma facilmente previsível ao realizar as medidas, ao contrário do que ocorre na Terra.

¹³De acordo com Kovalevsky (2005), Lacroute teve a idéia já em 1965, e a primeira proposta ao *Centre National d'Études Spatiales* foi realizada em 1966 segundo Turon & Arenou (2008).

¹⁴Mais detalhes sobre o histórico dessa missão podem ser encontrados em ESA (1997), Turon & Arenou (2008) e Perryman (2010).

¹⁵Atualmente, sistemas de óptica adaptativa corrigem efeitos atmosféricos apenas em torno de alguns poucos segundos de arco ($\sim 10''^2$), e sistemas avançados como o MCAO (*Multi-Conjugate Adaptive Optics*) do Gemini Sul ou o MAD no ESO-VLT corrigem campos de $\sim 1'2$ respectivamente (ex: Bouy et al, 2008). Mesmo avançados sistemas projetados para o futuro, como o MOAO *Multiple-Object Adaptive Optics*, serão capazes de corrigir tais efeitos apenas dentro de campos de alguns poucos minutos de arco.

- do espaço é possível observar o céu inteiro com um único instrumento, eliminando a necessidade de entender comportamentos de instrumentos distintos espalhados pelo mundo para criação de catálogos de céu inteiro—possibilitando ainda a obtenção de uma acurácia astrométrica uniforme na esfera.

Do ponto de vista de instrumentação astrométrica, o Hipparcos, introduziu uma grande inovação, possibilitada justamente pelo fato dele se encontrar no espaço: a utilização de dois espelhos observando direções diferentes do céu, mas iluminando um mesmo plano focal.¹⁶ Desse modo, conhecendo-se com grande precisão o ângulo entre os espelhos (chamado de ângulo básico), é possível determinar o ângulo entre os objetos no céu observados por cada um destes espelhos também com grande precisão—um esquema desse conceito pode ser visto na Figura 1.4.

Figura 1.4: Esquema óptico do sistema de campo de visão duplo do satélite Hipparcos. As imagens dos dois campos de visão entre pelas *Baffle Apertures*, são combinadas em uma mesma direção pelo *Beam Combiner* (que possui a forma de um corretor Schmidt), redirecionadas pelo *Flat Folding Mirror* e focalizadas pelo *Spherical Primary Mirror* no *Hipparcos Grid* (o plano focal). (De ESA, 1997)

Além disso, o princípio de observação do Hipparcos era baseado em um movimento de rotação constante do satélite em torno de si mesmo, de modo que o céu era observado de forma análoga ao feito por um círculo meridiano na Terra, produzindo dados que consistem principalmente nos tempos de trânsito dos objetos que cruzam o

¹⁶Um conceito parecido estava presente no instrumento utilizado por Bessel durante a descoberta paralaxe anual, ainda que aquele instrumento (um heliômetro produzido por Fraunhofer) houvesse sido criado para observação de ângulos pequenos. O conceito original de Lacroute utilizava na realidade três espelhos, mas o projeto final possuiu apenas dois.

plano focal.¹⁷ E, para que o céu pudesse ser varrido inteiramente, o satélite também possuía um movimento de precessão de seu eixo de rotação. Esses movimentos aliados à órbita da Terra ao redor do Sol permitiram com que o Hipparcos observasse uma mesma região do céu diversas vezes, em diferentes ângulos. Todo esse movimento (chamado de lei de varredura, ou *scanning law*) foi otimizado para permitir um eclipse constante dos campos de visão em relação ao Sol.¹⁸ A Figura 1.5 é uma representação da observação e da quantidade de vezes que cada região do céu foi observada pelo satélite Hipparcos.

Figura 1.5: (a) Conceito de observação de grandes ângulos pelo céu inteiro com o Hipparcos—a órbita geo-estacionária indicada nesta Figura nunca foi atingida. (b) Número de observações em cada região da esfera celeste. (De [ESA, 1997](#))

A redução destes dados e a transformação das medidas brutas e de informações adicionais em parâmetros astrométricos (posição, movimento próprio e paralaxe) para todas as fontes foi realizada em paralelo por dois consórcios independentes, chamados de FAST – *Fundamental Astrometry with Space Techniques* ([Kovalevsky et al, 1992](#)) e NDAC – *Northern Data Analysis Consortium* ([Lindgren et al, 1992a](#)). Como era a primeira vez que se realizava um projeto astrométrico tão preciso, não existiam dados independentes que possibilitassem uma verificação externa da qualidade astrométrica dos resultados obtidos, de modo que ao menos a comparação de duas reduções completamente independentes dos mesmos dados deveria ser capaz de indicar a existência de problemas graves.

Um terceiro consórcio, chamado INCA – *Input Catalogue Consortium*, ficou responsável pela preparação do catálogo inicial dos objetos que seriam observados pelo Hipparcos, chamado *Hipparcos Input Catalogue* ([Turon et al, 1992](#)). Este catálogo foi construído a partir da seleção de propostas científicas, re-observação dos

¹⁷Um círculo meridiano determina o tempo de trânsito de um objeto pelo meridiano do local.

¹⁸Este método foi utilizado pela primeira vez no espaço pelo Hipparcos, está sendo atualmente utilizado pelo satélite Plack, e será também utilizado pelo Gaia.

alvos e simulações das observações do satélite.¹⁹ Ao todo, mais de 200 cientistas trabalharam nesta missão espacial (ESA, 1997).

Durante a missão foram acumulados ~ 1 Tbits de dados brutos, o que consistia no maior problema de análise de dados já enfrentado pela Astronomia até então, e estas medidas precisavam ser transformadas em parâmetros astrométricos.

1.4.2 Redução dos dados

A morte do satélite, do telescópio em si, não marca o final da missão, e mesmo que os consórcios de redução já estivessem trabalhando com as medidas do satélite durante as observações (ex Mignard et al, 1992; Lindegren et al, 1992b), a redução durou diversos anos. Descrições detalhadas dos procedimentos de redução adotados estão relatadas em Perryman et al (1992), Kovalevsky et al (1992), Lindegren et al (1992a) e principalmente ESA (1997).

Ambos consórcios utilizaram um processo dividido em três etapas, e a principal diferença encontrava-se nos algoritmos e métodos adotados para a implementação desse processo. A solução era basicamente dividida em:

1. Processamento inicial das contagens de fótons;
2. Redução de grande círculo;
3. Solução da esfera.

O primeiro passo, consistia na transformação entre a contagem dos fótons realizada pelas fotomultiplicadoras dos *Star Mappers* e pelo *Image Dissector Tube* do instrumento principal em um sinal que pudesse ser utilizado para processamento astrométrico subsequente. Para isso era necessário determinar a atitude do satélite simultaneamente. Esse passo é dividido em três etapas:

- obtenção da estimativa da atitude e dos tempos de trânsito nos *star mappers*;
- reconstrução da atitude a partir dos dados do instrumento principal (do *Image Dissector Tube*) usando os tempos de trânsito obtidos na etapa anterior e posições de catálogo;
- determinação da fase e da amplitude do sinal usando a reconstrução da atitude.

O segundo passo, que é chamado de redução de grande círculo, consistia em transformar esses tempos de trânsito em valores de abscissa ao longo de um grande círculo.²⁰ Isso era realizado por meio de calibração de distorções das projeções dos objetos e do refinamento da atitude do satélite na direção deste grande círculo. O resultado era o polo do grande círculo, uma fase preliminar para o sinal e abscissas.

¹⁹Ao final mais de 118.000 estrelas (algumas nas Nuvens de Magalhães), um quasar (3C273), 48 asteróides e três satélites (Europa, Iapetus e Titã) foram selecionados para observação pelo Hipparcos (Turon et al, 1992).

²⁰A cada grande círculo eram observadas aproximadamente duas mil estrelas.

O terceiro e último passo, chamado de solução da esfera, é a determinação de fases coerentes para todos os grandes círculos observados conjuntamente com os parâmetros astrométricos. Basicamente, dado um número de grandes círculos, isso consistia na solução de um problema de mínimos quadrados dos parâmetros de cada objeto ($\alpha, \delta, \varpi, \mu_\alpha \cos \delta, \mu_\delta$), de uma correção para o ponto zero de cada grande círculo e de variáveis adicionais que representavam erros sistemáticos, problemas de harmônicos não tratados, correções cromáticas, etc.²¹ O problema era representado por um sistema matricial que equivalia a mais de 2 milhões de equações.

Um exemplo com posições de um objeto observadas em diversos grandes círculos pode ser vista na Figura 1.6—segmentos de reta representam valores das abscissas determinados a cada trânsito e a linha curva é a trajetória aparente do objeto inferida a partir da solução do problema de mínimos quadrados (o problema é resolvido de forma simultânea para todos os objetos observados).

Figura 1.6: Representação das observações do Hipparcos de um certo objeto e do resultado obtido após a solução da esfera. (De ESA, 1997)

Esse problema poderia ser resolvido por uma solução direta do problema global, a partir da fase e da amplitude do sinal observados. Isso contudo não era factível com computadores disponíveis na época do Hipparcos, pelo alto custo computacional: $\sim 10^{16}$ operações de ponto flutuante e $\sim 500\text{Gb}$ de memória.²² Contudo, já naquela época foi tentada uma solução global aproximada, chamada de *Global Iterative Solution*, cujo princípio foi proposto por P. Lacroute em 1978, mas que teve sua primeira implementação por L. Lindegren (ESA, 1997, vol.3, p.490).

²¹Para maiores detalhes, ver ESA (1997).

²²Atualmente esse problema se tornou factível até mesmo para computadores de médio porte.

1.4.3 O Tycho

Além do catálogo Hipparcos, essa missão gerou um segundo catálogo, menos preciso mas com uma quantidade significativamente maior de objetos—dez vezes maior. Esse catálogo, idealizado por E. Hog, foi chamado de Tycho e originou-se do aproveitamento científico dos dados dos *Star Mappers* do satélite, tendo sido construído pelo consórcio TDAC – *Tycho Data Analysis Consortium* a partir da comparação das observações dos *Star Mappers* com previsões destas observações baseadas em um catálogo de entrada e das determinações das correções necessárias. Os métodos utilizados para construção do Tycho encontram-se descritos em [ESA \(1997\)](#), v.3.

1.4.4 O pós-missão

Três anos após a publicação dos catálogos Hipparcos e Tycho ([ESA, 1997](#)) e da dissolução dos consórcios, os dados que deram origem ao Tycho foram re-analisados a partir de novos algoritmos baseados na análise de trânsitos acumulados e do uso de melhores catálogos auxiliares, gerando o Tycho-2 ([Høg et al, 2000c,a,b](#))—este se tornou um dos catálogos mais importantes da Astronomia, e serve como base para ligação de diversos catálogos modernos ao referencial Hipparcos.²³

E mesmo após mais de uma década da dissolução dos consórcios FAST e NDAC, os dados brutos do Hipparcos continuaram a ser analisados. Em 2008, uma nova redução foi publicada em [van Leeuwen \(2007\)](#), na qual a partir de uma melhor modelagem da dinâmica do satélite, de impactos de micro-meteoritos e de uma solução global moderna (como a que será utilizada para o Gaia) possibilitou aprimorar os parâmetros astrométricos em especial de estrelas brilhantes, atingindo erros médios de $0.7 - 0.8$ mas, e chegando em 0.09 mas para certos casos ([van Leeuwen, 2009a](#)).

O impacto científico da redução original dos dados dessa missão astrométrica foi gigantesco, tocando todas as áreas da Astronomia até hoje (mais detalhes em [Perryman, 2009](#)). Entretanto uma questão científica ainda aberta é a distância do aglomerado estelar das Pleiades: os resultados obtidos pelo Hipparcos para a paralaxe deste aglomerado não concordam com as distâncias derivadas de observações do HST e do ajuste da sequência principal por isócronas teóricas²⁴.

Já em 1976, quase duas décadas antes do final desta missão espacial, segundo [Turon & Arenou \(2008\)](#) o Grupo de Definição de Missão ao comentar sobre o que era então apenas a proposta do projeto Hipparcos

*recomenda fortemente, como os maiores avanços científicos, que um programa Espacial Astrométrico deva ser estabelecido, requisitando o lançamento de uma segunda espaçonave, idêntica ou similar à primeira, após um intervalo de tempo de uma dezena de anos.*²⁵

²³Como por exemplo, versões iniciais do *Sloan Digital Sky Survey* que utilizavam diretamente o Tycho-2 ([Pier et al, 2003](#)), e versões mais novas utilizam o UCAC2 que por sua vez é baseado no material utilizado para a criação do Tycho-2 ([Zacharias et al, 2004](#)).

²⁴Para um resumo atualizado da literatura sobre essa questão, ver [van Leeuwen \(2009b\)](#).

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

1.5 A missão Gaia

*My reason for emphasizing precision space astrometry is that (...)
we need to be aware of our scientific roots.*

Malcolm Longair, segundo Perryman (2010).

O grande sucesso da missão espacial Hipparcos já era acompanhado das preparações para o segundo passo da Astrometria espacial. Ao decolar, o Hipparcos ainda estava munido de fotomultiplicadoras enquanto em paralelo detectores muito mais eficientes denominados CCDs tornavam-se cada vez mais aptos para utilização nos ambientes extremos de missões espaciais.

Durante os anos 90 existiram diversas propostas de novas missões astrométricas, das quais a principal se chamava GAIA (*Global Astrometric Interferometer for Astrophysics*), e era baseada em um interferômetro. Esse conceito de observação foi posteriormente modificado para uma missão que seguia o mesmo princípio de observação do Hipparcos, mas cujo plano focal era formado por CCDs além de possuir a capacidade de realizar observações das cores dos objetos e de suas velocidades radiais, dando origem a atual missão Gaia.²⁶

A Missão Gaia (Perryman et al, 2001; Mignard, 2010) é um dos mais ambiciosos projetos da Astronomia moderna. Com lançamento previsto para dezembro de 2012, o satélite deve realizar observações de mais de um bilhão de objetos, originando um catálogo com precisões astrométricas inéditas: $7 \mu\text{as}$ até magnitude $G \sim 12$, $25 \mu\text{as}$ para $12 < G < 15$, e 0.3 mas entre $15 < G < 20$.²⁷ Além de informações astrométricas de qualidade sem precedentes, o satélite ainda obterá dados espectro-fotométricos equivalentes a dezenas de bandas (~ 25) para todos os objetos observáveis e dados espectroscópicos na região do tripleto de Ca II ($R \sim 11.500$) que possibilitarão a determinação da velocidade radial de quaisquer objetos até magnitude $G \sim 17$. A missão prevê a observação de todos os objetos com magnitude aparente até $G \sim 20$, englobando algo entre um e dois por cento de todas as estrelas da Galáxia, além de milhões de objetos extragalácticos.

O enorme impacto da missão Gaia pode ser sentido por exemplo, através de uma rápida análise nos resultados esperados para determinação de distâncias pelo método da paralaxe trigonométrica. Atualmente, existem menos de 500 estrelas com erros em paralaxe menores que 1%, provenientes da missão Hipparcos, com Gaia este número deve saltar para ~ 10 milhões de estrelas até 2.5 kpc. Com erro relativo até 10% temos hoje paralaxes de pouco mais de 20 mil estrelas, e após o Gaia esse deve ser maior do que de ~ 100 milhões para distâncias de até 25 kpc.

Essa base de dados sem precedentes, tanto do ponto de vista da quantidade quanto da qualidade, possibilitará uma compreensão muito mais precisa da Galáxia em vários aspectos: origem, estrutura, formação, evolução, composição. Naturalmente,

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.

²⁶Essa missão continua a se chamar Gaia, porém o acrônimo não é mais válido.

²⁷G é a magnitude calculada sem filtros, entre 330 e 1000 nm (Jordi et al, 2010).

tendo em vista todos os objetos que serão observados pelo Gaia, desde corpos do Sistema Solar e planetas extra-solares aos quasares longínquos, pode-se dizer que sua contribuição científica vai muito além das questões relativas à nossa Galáxia. A quantidade e a qualidade de suas observações bem como a abrangência de seu catálogo final, terão implicações muito profundas e extremamente importantes em muitos aspectos do estudo do Universo: desde o estudo da formação do Sistema Solar, passando por busca de sinais de vida fora do Sistema Solar, até implicações cosmológicas oriundas da calibração da escala de distância e de medidas precisas de constantes físicas fundamentais.

1.5.1 O instrumento

Assim como seu antecessor, o Hipparcos, o satélite Gaia utiliza o princípio de astrometria global a partir da medida de ângulos entre diversos objetos, já previamente testado por aquela missão. Como vimos anteriormente, este princípio é baseado na observação do céu inteiro, simultaneamente a partir de dois campos de visão (separados por 106.5°) de uma espaçonave que gira lentamente em torno de um eixo (perpendicular aos campos de visão) e que possui um movimento de precessão, de modo a proporcionar uma cobertura completa do céu. No caso do Gaia, um grande círculo será descrito no céu a cada 6 horas, e os objetos que estiverem em trânsito em um dos campos serão observados a partir do segundo campo 106.5 minutos após a observação no primeiro campo. A precessão é completada a cada 63.12 dias.

Figura 1.7: (a) Diagrama do método de observações do Gaia (De DPAC, 2007). (b) Mapa representando o número de observações em cada região do céu em coordenadas galácticas. (De Mignard, 2010)

O desenho óptico do instrumento consiste em dois telescópios idênticos, de formato retangular com $1.45\text{m} \times 0.5\text{m}$ e distância focal de 35m (ver Figura 1.8). Esses dois instrumentos compartilham um mesmo plano focal, composto por um mosaico de 106 CCDs de 4500×1966 pixels e aproximadamente 4.7cm por 6cm (cada uma), sendo que cada pixel ocupa um campo de 59×177 mas (ver Figura 1.9). Esse detector, o maior jamais lançado ao espaço, encontra-se dividido em regiões de CCDs para o

mapeador do céu (*Sky-Mapper*, ou SM), para o campo astrométrico (*Astro-Field*, ou AF), para o campo fotométrico, para o campo espectroscópico e para o sistema de metrologia interna (mais detalhes na Figura 1.9). O modo de funcionamento da integração e leitura desse sistema é similar aquele encontrado em CCDs instaladas em círculos meridianos, denominado *Time Delay Integration* ou *drift-scan*.

Figura 1.8: (a) Esquema do *payload* científico do satélite Gaia mostrando os espelhos primários e os diversos espelhos auxiliares. (b) Esquema do plano focal com a posição dos prismas de dispersão BP e RP além do instrumento espectroscópico de velocidade radial (RVS). (De EADS Astrium)

Devido ao tamanho do plano focal (~ 1 gigapixel) e às limitações de banda de comunicação existentes no momento de integração do satélite, seria completamente impossível a transferência para Terra do conteúdo de todas essas CCDs. Deste modo existe a bordo um sistema de processamento e de seleção dos dados que serão integrados e transferidos, uma tarefa que é realizada pela *Video Processing Unit* através de algoritmos dedicados (ver seção 1.5.1.1).

A seleção destes dados é iniciada no momento em que fontes são detectadas nos mapeadores do céu, então caso a detecção seja confirmada na primeira coluna de CCDs do campo astrométrico, esses objetos são integrados pelas outras CCDs, que serão lidas apenas em “janelas” ao redor dos objetos. Estas janelas são formadas por conjuntos de pixels ao redor do pixel central do objeto, sendo que a quantidade de pixels lidos e transferidos é dependente tanto da coluna de CCDs em que a observação está sendo realizada quando da magnitude do objeto observado (mais detalhes na Tabela 2.1, no Capítulo 2).

A cada coluna do plano focal astrométrico uma nova medida completamente independente do objeto observado é realizada, de modo que a cada trânsito do Gaia, ou seja, a cada vez que o satélite observa uma certa região do céu, nove observações astrométricas distintas são realizadas dos mesmos objetos.

Os dados fotométricos são obtidos utilizando dois prismas que dispersam a luz dos objetos enquanto estes transitam frente às CCDs fotométricas. Esses prismas são denominados RP (*Red Photometer*) e BP (*Blue Photometer*), uma vez que um

Figura 1.9: Esquema do plano focal do satélite mostrando as 106 CCDs que formam um conjunto com quase 1 gigapixels. (De Alexander Short – ESA)

dispersa a parte vermelha (de 640 até 1050 nm) enquanto o outro a azul (de 330 até 680 nm) do espectro. Juntos, os campos BP e RP cobrem toda a faixa espectral do visível, chegando a penetrar um pouco no infra-vermelho próximo.

O último instrumento é um espectrógrafo voltado para a medida de velocidade radial, denominado RVS ou *Radial Velocity Spectrometer*. Este instrumento possui uma resolução média, com $R \sim 11.500$, e cobre a região ao redor do tripleto de Cálcio II (entre 847 e 874 nm) e embora seja otimizado para a determinação da velocidade radial dos objetos observados, também fornecerá informações sobre velocidades de rotação, parâmetros atmosféricos, abundâncias de elementos e traçadores do avermelhamento interestelar.

Figura 1.10: Observações simuladas (a) espectro-fotométricas para uma estrela M6V diretamente nas CCDs e integradas na direção perpendicular a dispersão para um intervalo de valores de $\log(g)$ entre -1.0 e +5.0 e (b) espectroscópicas de uma estrela tipo F com $G \sim 16$ para uma observação única e para a combinação de todas as observações ao final da missão. (De DPAC, 2007)

O princípio do RVS é similar aos prismas BP/RP sendo que uma rede de difração dispersa toda a luz dos objetos que estão transitando no plano focal sem uso de fibras ou fendas. Um exemplo de observação de uma estrela com $G \sim 15$ de tipo espectral M6V pelo BP e o RP, e de medidas espectroscópicas do RVS para uma estrela tipo F com $G \sim 16$ para uma única observação e para o espectro combinado ao final da missão, podem ser vistos na Figura 1.10.

1.5.1.1 A VPU – *Video Processing Unit*

Como comentados anteriormente, os dados do plano focal completo não podem ser transferidos para a Terra por uma questão de largura de banda. Então, já a bordo do satélite ocorre um processamento preliminar dos dados para determinação de quais pixels serão transmitidos. Isso é a tarefa da VPU, ou *Video Processing Unit*, que é a central de processamento de video do Gaia. Além disso, a VPU também serve para alimentar o subsistema de controle de atitude do satélite.

Praticamente todos os dados científicos do Gaia passam em algum momento por esse subsistema, e do ponto de vista desta tese, a VPU é um sistema essencial, pois é ela quem define quais objetos serão observados e como os dados destes objetos serão combinados antes do envio à Terra.

Os algoritmos de detecção são tais que inibem a transferência dos dados de quaisquer objetos com tamanho angular maior do que ~ 700 mas,²⁸ como pode ser verificado a partir de simulações da eficiência na detecção de objetos de diversos tamanhos apresentada na Figura 1.11.

Figura 1.11: Eficiência na detecção em função do tamanho angular aparente do objeto observado. (De [Astrium, 2008](#))

A detecção ou não de um objeto é realizada após o objeto passar pelas colunas de CCDs do *Sky Mapper*. Ela é baseada no contraste entre o valor de um pixel e dos pixels ao redor deste, sendo um processo de três etapas. O primeiro passo é o cálculo de uma média dos valores dos pixels que formam o quadrado 5×5 ao redor do

²⁸Observe que este é o tamanho que o sistema de detecção acredita que o objeto possui, ou seja, o tamanho da região mais luminosa do objeto.

pixel sob análise e a subtração desse valor de todos os pixels dentro do quadrado 3x3. O segundo passo é o cálculo de alguns atributos, como por exemplo fluxos nas direções paralela e perpendicular à varredura, o fluxo total, se o pixel central é um máximo local, e algumas estratégias de rejeição baseadas em fluxos. O passo final é a aplicação de cortes principalmente baseados em um fluxo limite (determinado arbitrariamente) e no fato do pixel central ser ou não um máximo em ambas as direções. No caso do pixel passar por todos esses critérios, uma janela de tamanho adequado é criada, confirmada no AF1 (a primeira coluna de CCDs do AF) e após finalizar a integração em todo plano focal, transferida para a Terra. Maiores detalhes do funcionamento deste sistema podem ser encontrados em [Astrium \(2010\)](#).

1.5.2 O DPAC – *Data Processing and Analysis Consortium*

Os dados transferidos pelo satélite possuem uma grande complexidade, e para preparar o sistema de processamento de dados do Gaia a comunidade científica, principalmente européia, está organizada desde 2006 em um consórcio chamado *Data Processing and Analysis Consortium*, ou DPAC, cuja descrição estendida pode ser encontrada em [DPAC \(2007\)](#).²⁹

Esse consórcio envolve atualmente mais de 400 pesquisadores, de 23 nacionalidades diferentes mais a ESA (ver Figura 1.12), e possui a responsabilidade, perante a Agência Espacial Européia, de preparar todo o sistema de processamento científico de dados do satélite com a finalidade de tornar os dados brutos que chegam à Terra em dados científicos que possam ser explorados cientificamente.

Figura 1.12: Representação do número de pesquisadores membros do DPAC—situação em janeiro de 2010. Informações atualizadas podem ser encontradas em: http://www.rssd.esa.int/index.php?project=GAIA&page=DPAC_Membership.

A tarefa envolve não apenas o processamento de uma grande quantidade de informações, mas principalmente de informação de uma grande complexidade. O exemplo mais claro disso é o fato de que o Gaia é um sistema auto-calibrável: as

²⁹ Antes do DPAC, um comitê chamado DACC (*Data Analysis Coordination Committee*) tinha a tarefa de liderar a criação do DPAC a partir de estruturas mais informais previamente existentes desde 2001. O DACC foi dissolvido após a formação do DPAC.

posições medidas dos trânsitos das estrelas são utilizadas ao mesmo tempo para determinar a atitude do satélite e as posições das próprias estrelas no céu.

O DPAC é dividido em nove grupos (Coordination Units - CU) que desenvolvem os aspectos globais a serem abordados na missão (por exemplo CU7–*Variability Processing* é dedicado a análise de variabilidade dos objetos). Essas Coordination Units são por sua vez organizadas em Development Units (DU), que são os grupos responsáveis por tarefas científicas mais determinadas do projeto (por exemplo, CU2–*Simulations* DU4–*Instrument Model* é responsável dentre as simulações pelo modelo de simulação do instrumento). Enfim, esses DUs são subdivididos em WorkPackages, que são subgrupos responsáveis por especializações das tarefas destas DUs. Nas sessões 1.5.5.2 e 1.5.5.1 veremos em mais detalhes as CU4–*Object Processing* e CU5–*Photometric Processing*, dentro das quais este trabalho de tese encontra-se inserido. Um diagrama com todas as CUs pode ser encontrado na Figura 1.13.

Figura 1.13: Estrutura do DPAC. (De Mignard, 2010)

Estimativas apresentadas em Mignard (2010) indicam que serão gerados pelo satélite ~ 250 Tb de telemetria comprimida, sendo que os dados processados devem consumir entre ~ 0.5 PB e ~ 1 PB. Para processar esse volume de informações, segundo estimativas aproximadas, serão necessários $> 1.5 \times 10^{21}$ FLOPs, o que apesar de ser um valor grande, é perfeitamente atingível.³⁰

Os dados são processados de forma não centralizada para que a redução possa ocorrer de forma eficiente. Enquanto o satélite observa o céu, caso exista possibilidade de transferência, os dados seguem para o centro de operações da missão (MOC) e para o ESAC (*ESA Center*) onde são armazenados no Gaia MDB (*Main Data Base*).

³⁰Para solucionar esse problema em um ano, seria necessário um computador dedicado com ~ 30 TFLOP/s, o que atualmente equivale a dois ou três racks de um Cray XT6, com um custo de ~ 1.5 milhões de dólares. Segundo Mignard (2010), dois centros de processamento de dados possuirão uma infra-estrutura dedicada ao Gaia de 10–20 TFLOP/s cada um já em 2012, e tempo parcial já está disponível no Marenostrum em Barcelona, que é capaz de atingir ~ 95 TFLOP/s.

Eles são então distribuídos para os diversos centros de processamento do DPAC onde são transformados em parâmetros astronômicos para todas as fontes observadas em ciclos periódicos de 6 meses. Uma vez processados, os resultados são enviados de volta ao ESAC, sendo armazenados no MDB. Durante o processamento dos dados científicos, seis centros de processamento (chamados de DPC—*Data Processing Center*) estarão envolvidos (ver Figura 1.13). Os trabalhos desenvolvidos durante esta tese são (no caso de simulações) ou serão (no caso de processamento de dados) executados nos centros de processamento do CNES³¹ (Cap. 4, 5, 6), Cambridge (Cap. 3) e Barcelona (Cap. 4).

1.5.3 Simuladores

Com a finalidade de prover coerência para o DPAC, bem como para a preparação dos softwares que serão utilizados na análise dos dados do Gaia, as simulações do Gaia são centralizadas e são utilizados códigos que operam de forma comum para todos os membros deste consórcio. Os códigos de simulação utilizam rotinas comuns englobadas em um pacote denominado GaiaSimu (Babusiaux et al, 2010), sendo três códigos principais: GASS, GIBIS e GOG. Estes códigos terão suas funcionalidades descritas nas próximas seções, contudo, todos estes códigos necessitam de uma simulação básica do céu, que é provida pelo “Modelo de Universo” do Gaia.

1.5.3.1 Modelo de Universo

O denominado “Modelo de Universo” consiste em um código de simulação do céu: dada uma certa região da esfera celeste, este código simplesmente retorna quais objetos astronômicos deveriam ser observados. Isso significa que dentro deste modelo existem descrições de como é o Sistema Solar (planetas e seus satélites, asteróides, cometas), a Via-Láctea (estrelas, aglomerados, binárias, radiação difusa, nebulosas), de como são galáxias próximas resolvidas em estrelas e de como são as galáxias não resolvidas, mas extensas, além de quasares mais distantes e das distribuições espaciais de todos estes objetos.

O modelo de galáxia adotado é chamado modelo de Besançon.³² Para utilização no simulador, este modelo foi adaptado à linguagem Java, e modificado para possibilitar a simulação de objetos variáveis tanto do ponto de vista fotométrico quanto do ponto de vista astrométrico, além da inclusão de objetos múltiplos (e aglomerados), radiação difusa e nebulosas.³³

Galáxias próximas são introduzidas como estrelas individuais, particularmente as núvens de Magalhães estão sendo implementadas nos simuladores neste momento. Galáxias mais distantes, que não são resolvidas em objetos individuais são introduzidas no simulador por meio de um método desenvolvido por E. Bertin chamado Stuff,

³¹O *Centre National d'Études Spatiales* é a agência espacial francesa.

³²No início da missão, um segundo modelo também era utilizado, o modelo de Barcelona.

³³Na realidade isso não é feito a partir de uma modificação do modelo de Besançon, mas sim por meio da inclusão de códigos de simulação para estes objetos que funcionam de forma mais ou menos independente do modelo de Besançon.

que teve sua primeira implementação para os parâmetros do Gaia realizada por C. Dollet e cujo desenho final e implementação em Java (JStuff) que é utilizada nas simulações da missão foi realizada por nós, e será descrita no Capítulo 4.

As imagens das galáxias podem ser produzidas de duas formas distintas: seja utilizando o Skymaker (outro código de E. Bertin, cuja implementação atual é devida à C. DelleLuche) que forma as imagens a partir da soma de dois componentes analíticos para a distribuição de brilho dos objetos, ou a partir da utilização do código Magil, que utiliza imagens reais de galáxias próximas como protótipos para produzir imagens mais realistas (este código, implementado por Gavras et al, 2010, será descrito com mais detalhes no Capítulo 4). As supernovas extragalácticas são também incluídas, contudo de forma aleatória. Segundo Robin et al (2010), em versões futuras do simulador (pós-2010) elas devem ser geradas ao redor de certas galáxias com probabilidades de ocorrência dependentes do tipo de Hubble.

Os quasares são introduzidos a partir de uma lista de objetos, descrita em (Slezak & Mignard, 2007), que é gerada por um modelo estatístico com propriedades similares ao SDSS extrapoladas até $G=20.5$. Estas listas são complementadas pelo catálogo de quasares próximos de Véron-Cetty & Véron (2006).

Mais detalhes sobre o modelo de universo adotado podem ser encontrados no documento Robin et al (2010), que é atualizado periodicamente.

1.5.3.2 GASS

O GASS, ou *Gaia System Simulator*, é o responsável por simular grandes volumes de dados, simulando assim a telemetria do satélite e sendo capaz de reproduzir em computador uma missão completa em todas as suas fases. As simulações do GASS são tão realísticas quanto possível, sendo estatisticamente representativas das observações do satélite. Uma descrição atualizada deste simulador pode ser encontrada em Gallardo & Masana (2010).

As simulações geradas pelo GASS são utilizadas para testes das cadeias de redução e processamento em todos os níveis da missão, serão utilizadas para a validação de todo o fluxo de dados antes da missão real ser iniciada, são utilizadas para testar e aprimorar o *Main Data Base* e foram utilizadas para estimações do número de objetos que serão observados e da quantidade de dados transmitida para o solo.

O processo básico é dividido em etapas de inicialização dos geradores de fontes e da simulação da missão (missão completa, parcial, limite de magnitude, não linearidade, ruído na atitude do satélite, etc.), da simulação das fontes astronômicas, da simulação de observações do satélite (reproduzindo estatisticamente os processos de detecção, seleção, tempo e posição de trânsito, tamanho da janela de observação, etc.), da geração dos pacotes de telemetria (formatados exatamente como o satélite real deverá fazer, e adicionados de erros de centragem, ineficiência na transferência de cargas, etc.) e do arquivamento dos resultados.

Esta tese não fez uso direto deste simulador, contudo o GASS faz uso de um código implementado por nós durante o desenvolvimento desta tese, o JStuff que será descrito no Capítulo 4.

1.5.3.3 GIBIS

O GIBIS, ou *Gaia Instrument and Basic Image Simulator* (Babusiaux, 2006) é o simulador dos dados da missão Gaia ao nível dos pixels dos CCDs do plano focal do satélite. De forma distinta ao GASS, o GIBIS simula as observações individuais do Gaia de cada objeto de forma tão realística quanto possível, sendo que modelos dos objetos e dos instrumentos extremamente realistas são utilizados. Como se pode imaginar, com a utilização deste simulador missões completas não podem ser simuladas para o céu inteiro devido ao tempo de cálculo proibitivo.

Contudo, todas as observações de um certo objeto que serão realizadas durante a missão completa podem ser simuladas, provendo deste modo informações detalhadas sobre como este objeto particular será observado. Isso é importante para estudos de objetos extensos, de múltiplos objetos na mesma linha de visada, para a redução fotométrica de campos densos, para a calibração do instrumento, para testes dos algoritmos de determinação de fundo de céu, etc. (ver Babusiaux, 2006).

O GIBIS pode ser utilizado de duas formas: por uma interface web, disponibilizada pelo CNES e que pode ser acessada *à la carte*, ou por meio de requisições oficiais à CU2, quando são necessários dados não normalmente registrados pela interface web e que são utilizados por uma pequena fração do DPAC.

Para realizar as simulações, uma vez que o usuário tenha requisitado os parâmetros do modelo de universo, do instrumento e da missão via a interface web, o GIBIS simula a região do céu requisitada por meio do Modelo de Universo.

Então, para cada um dos trânsitos do Gaia naquela direção do céu, a resposta do plano focal do Gaia é simulada a partir de modelos detalhados do instrumento (sendo que até mesmo a PSF pode ser calculada a cada observação), transformando a lista de objetos gerada pelo Modelo de Universo nos valores de pixels em todas as CCDs do plano focal (exemplo na Figura 1.14). Finalmente, o GIBIS executa os algoritmos de deteção e de criação de janelas que serão utilizados a bordo do satélite, resultando nas observações dos objetos a cada trânsito. O processo é então repetido para todos os trânsitos requisitados, ou para toda a missão.

Figura 1.14: Simulação de um campo estelar produzida com o GIBIS. (De DPAC, 2007)

Este foi o único simulador utilizado diretamente para o desenvolvimento deste trabalho de tese. Além disso, o Gibis faz uso dos códigos JStuff e Magil, que serão descritos no Capítulo 4.

1.5.3.4 GOG

O GOG, ou *Gaia Object Generator*, é um código de simulação responsável por gerar listas de objetos observados cujos parâmetros devem estar convoluídos com as funções de erros esperadas de modo a estatisticamente reproduzir as observações presentes no MBD e o catálogo Gaia sem a necessidade de realizar simulações detalhadas dos processos de observação, como é feito pelo GIBIS, e nem executar os algoritmos de redução de dados da missão, como é necessário com o GASS.

A primeira implementação desse simulador foi criada por nós, e ainda que bastante rudimentar serviu em 2006 para realizar a primeira simulação de larga escala da missão Gaia completa, chamada de GUMS (*Gaia Universe Model Snapshot*), tendo sido executada no supercomputador Marenostrum em Barcelona (X. Luri, *pers. comm.*). Essa simulação permitiu confirmar, por exemplo, as distribuições de densidades de objetos em todo o céu além do diagrama Hertzsprung-Russel, magnitude-paralaxe, etc. que poderiam ser obtidas com o Gaia. O funcionamento era básico: os parâmetros do modelo de Universo eram obtidos a partir de arquivos XML³⁴ (ex. qual modelo de Galáxia utilizar, se objetos duplos e múltiplos e variáveis deveriam ser incluídos, se o modelo extragaláctico deveria ser executado, etc.), então o código instanciava os objetos necessários,³⁵ inclusive determinando as regiões HTM correspondentes a simulação³⁶ e executava os códigos de simulação, cujos resultados eram formatados e escritos em um arquivo de saída.

Figura 1.15: Densidade de objetos e diagrama Hertzsprung-Russel gerados a partir do *Gaia Universe Model Snapshot* simulado em 2006 com a versão 0.0 do GOG implementada por nós. Figuras criadas por X. Luri e C. Babusiaux.

Atualmente, o código que é utilizado pelo DPAC não possui mais herança daquele que foi implementado em 2006, sendo que esse simulador evoluiu consideravelmente,

³⁴XML ou *eXtensible Markup Language* é uma linguagem que permite descrever diversos tipos de dados de forma auto-contida.

³⁵De certo modo, isso significa que os códigos necessários eram carregados na memória.

³⁶Forma de divisão de esferas em regiões triangulares (ver seção 4.2.1 para detalhes).

permitindo não apenas a simulação de listas de objetos no final da missão obtidos diretamente do modelo de Universo, mas também de listas de objetos intermediárias e do estado do MBD em quaisquer épocas entre o início e o final da missão.

Isso pode ser realizado graças a modelos estatísticos da atitude e dos erros intermediários tanto das observações quanto dos procedimentos de redução de dados, que são fornecidos pelos diferentes CUs participantes do DPAC. Com isso, o GOG é capaz de rapidamente gerar simulações de grandes volumes de dados. Com o GOG também temos um conjunto de ferramentas que possibilita realizar análises estatísticas dos resultados obtidos para as simulações rapidamente, além de tanto poder ser executado por meio de uma interface web disponibilizada pelo CNES quanto instalado pelo próprio usuário. Uma descrição da implementação e das funcionalidades deste simulador pode ser encontrada em [Isasi et al \(2010\)](#).

Assim como no caso do GASS, esta tese não fez uso direto deste simulador, contudo o GOG utiliza o JStuff (descrito no Capítulo 4).

1.5.4 Redução de dados

A redução de dados do Gaia ocorrerá em dois níveis de processamento, um executado diariamente e um semestralmente. O processamento diário ocorrerá assim que os dados do satélite chegarem à Terra. O primeiro passo é um ajuste de uma PSF/LSF³⁷ aos dados da janela, e então a partir da atitude aproximada do satélite (que é obtida a partir da telemetria) e de um catálogo intermediário, a janela é associada a um determinado objeto (ou um novo objeto é criado). Esse processo faz parte da cadeia inicial de processamento de dados do Gaia, chamada de IDT ou *Initial Data Treatment*, gerando dados que serão armazenados diariamente no MDB. Durante o IDT uma reconstrução preliminar da atitude do satélite baseada em estrelas brilhantes já será realizada.

Além disso, em missões de astrometria global como esta, para que possam ser obtidos resultados globais e coerentes são necessários diversos meses de observação. Deste modo, se algo anormal ocorresse, isso seria percebido apenas muitos meses após o evento. Para prevenir este tipo de situação, um processamento chamado *Detailed First Look* foi implementado, e será responsável por realizar testes de sanidade científica dos dados 24h após sua recepção. Testes astrométricos, fotométricos e espectroscópicos serão realizados. O alinhamento óptico entre os telescópios e o foco são alguns exemplos de problemas que podem ser rapidamente detectados a partir deste sistema. Além disso, alertas científicos também serão executados no *First Look* (como por exemplo, supernovas), de modo a possibilitar reações rápidas da comunidade astronômica.

Então, a cada ciclo de seis meses ocorrerá o processamento principal, cuja estrutura é dependente de cada CU e DU envolvido. Basicamente, para que este processamento a cada ciclo ocorra, cópias de partes do MDB serão transferidas para os DPCDBs (*Data Processing Center Data Base*), e a partir destas cópias os

³⁷Como as observações do Gaia são essencialmente unidimensionais, está definida uma LSF, ou *Line Spread Function*, que é o análogo unidimensional de uma PSF clássica. Ver [Lindgren \(2009\)](#).

CUs/DUs executam suas análises de forma mais ou menos independente. Quaisquer inter-dependências entre CUs diferentes ocorrerão por meio do MBD, logo caso um CU dependa do resultado de outro CU, será necessário aguardar um ciclo para troca de informações. Um esquema do processamento dos dados do Gaia pode ser vista na Figura 1.16.

Figura 1.16: Visão conceitual do processamento de dados do Gaia.

O processamento astrométrico funcionará de forma iterativa, sendo que a cada ciclo novos dados serão incorporados à solução e os parâmetros astrométricos dos objetos, a atitude do satélite, as calibrações do instrumento e as correções relativísticas serão re-calculadas. Durante o processamento das fontes mais estáveis $\sim 5 \times 10^8$ de parâmetros de fontes, $\sim 4 \times 10^7$ parâmetros de atitude e $\sim 10^6$ parâmetros de calibração devem ser resolvidos. Então, os outros 5×10^9 parâmetros astrométricos para as outras fontes serão determinados.

1.5.5 Redução de dados de objetos problemáticos

No caso de existirem objetos cuja distribuição da estatística de χ^2 obtida pelo ajuste da PSF/LSF durante o IDT em cada trânsito observado seja incompatível com aquela que seria esperada pela existência de apenas uma fonte pontual dentro das janelas observadas (para detalhes, ver [Burgon, 2010](#)), o objeto é posteriormente analisado por duas DUs especializadas do DPAC.

Estas duas DUs fazem parte dos CU4 e CU5, e tem como objetivo analisar se existem outros objetos angularmente próximos do objeto observado (que podem se encontrar dentro da mesma janela) ou se o objeto possui alguma emissão extensa intrínseca (como por exemplo será o caso de galáxias não resolvidas em estrelas), e em ambos os casos extrair informações adicionais destes objetos.

Uma visão do conceito do processamento entre essas duas DUs pode ser encontrada na Figura 1.17, contudo o processamento exato não ocorre da maneira representada,

uma vez que os dados passam pelo MDB após o processamento que ocorre no CU5, chegando apenas um semestre depois no CU4.

Figura 1.17: Visão conceitual do processamento de objetos problemáticos durante a redução de dados do Gaia.

É dentro destas duas DUs, descritas globalmente nas próximas duas seções, que a maior parte deste trabalho de tese foi realizada.

1.5.5.1 CU5-DU18 – *Source Environment Analysis*

Os objetos armazenados no MBD que possuem uma distribuição de valores de χ^2 distinta daquela esperada para objetos pontuais são analisados pela DU18 do CU5, chamada de *Source Environment Analysis*. Nesta DU, coordenada por D. Harrison do *Institute of Astronomy* da *University of Cambridge*, é realizada uma análise que busca verificar se o sinal presente nos trânsitos é compatível ou não com mais de um objeto pontual existente dentro da mesma janela. Um esquema de como ocorre esse processamento pode ser visto na Figura 1.18.

O primeiro passo é a realização de uma reconstrução de imagem a partir das diversas janelas observadas durante os trânsitos do satélite sob o objeto. Para realizar esta reconstrução, diversos algoritmos encontram-se em estudo por outros pesquisadores, e serão vistos com mais detalhes no Capítulo 2, onde também demonstramos que é possível reconstruir essas imagens em quaisquer regiões do céu. Essa reconstrução, no entanto, precisa ser realizada da forma mais rápida possível, devido ao grande número de objetos (estimado em de dezenas de milhões) que devem sofrer este problema (maiores detalhes da estimativa realizada durante esta tese serão vistos no Capítulo 3).

Uma vez que uma imagem tenha sido reconstruída para o objeto e seus arredores, esta imagem é caracterizada em termos das fontes existentes e de parâmetros aproxi-

Figura 1.18: Esquema do processamento que ocorre no DU18 do CU5.

mados destes objetos (posição x , y , fluxo e erros associados) e de uma estimativa do tipo de fundo de céu. Os algoritmos que foram desenvolvidos durante esta tese para realizar tal caracterização serão detalhados no Capítulo 3.

Finalmente, os trânsitos individuais são re-analisados, buscando resolver os parâmetros astrométricos ($\alpha, \delta, \mu_\alpha, \mu_\delta, \varpi$) de todas as fontes presentes na imagem. Isso é tentativamente realizado por meio de uma re-análise dos dados brutos a partir de uma modelagem dos trânsitos dadas as fontes existentes na região, e é uma fase que ainda está em estudo por outros pesquisadores. Caso esta análise seja bem sucedida seu resultado é escrito no MBD, caso ela seja mal sucedida, o objeto é marcado no MBD como “não compatível com fontes pontuais”.

1.5.5.2 CU4-DU70 – *Extended Objects*

Os objetos marcados no MBD como “não compatíveis com fontes pontuais”, assim como os objetos classificados pelo CU8 como galáxias não resolvidas,³⁸ são re-processados pela DU70 do CU4, que é coordenada por C. Ducourant, do *Laboratoire d’Astrophysique de Bordeaux* da *Université de Bordeaux I*. Esse re-processamento, que foi desenhado durante o desenvolvimento deste trabalho de tese, é responsável por analisar o sinal das janelas observadas a cada trânsito em busca de objetos compatíveis com morfologias intrinsecamente extensas.

Como um menor número de objetos é esperado (em relação ao CU5-DU18), em torno de 10^6 de fontes, pode-se dedicar um pouco mais de tempo de processamento para cada fonte.³⁹ Contudo, como veremos no Capítulo 2, ainda não estão disponíveis algoritmos de reconstrução de imagens que sejam completamente livres de artefatos de reconstrução e que possuam tempo de execução compatível com o necessário. Deste modo, a estrutura de processamento de dados para estes objetos extensos foi

³⁸Não resolvidas em estrelas, mas com estrutura extensa.

³⁹Mas ainda assim o tempo de processamento necessário não deve ser negligenciado: se cada fonte for analisada em 1 minuto, serão necessários ~ 694 dias para analisar todos os objetos serialmente.

construída levando em consideração desde o início que as imagens reconstruídas eram intrinsecamente não plenamente confiáveis. Um esquema funcional de como ocorre esse processamento pode ser visto na Figura 1.19.⁴⁰

Figura 1.19: Esquema geral do processamento que ocorre dentro do DU70 do CU4.

Assim como ocorre no processamento do CU5-DU18, a etapa inicial é a reconstrução de uma imagem, um processo que será comentado com detalhes no Capítulo 2 desta tese. Então, essa imagem é analisada do ponto de vista morfológico, sendo que parâmetros morfológicos robustos à degradação da informação são mensurados nesta imagem. Estes parâmetros são medidas de como a luz encontra-se concentrada e distribuída na imagem, e serão comentados com detalhes no Capítulo 5.

Os parâmetros mensurados são utilizados para classificar a imagem em classes de objetos. No momento a classificação encontra-se implementada para diferenciar galáxias, que devem ser os casos de objetos extensos intrínsecos mais numerosos observados, nos tipos elípticos e espirais (com uma terceira classe para casos peculiares, irregulares, etc.). No Capítulo 5 comentamos mais sobre essa classificação.

A última etapa do processamento é um ajuste de perfil de luminosidade aos dados originais uni-dimensionais (e não à imagem reconstruída). Para isso, o perfil teórico e seus parâmetros são inicialmente escolhidos tendo em vista a classificação obtida na etapa anterior. O processo de ajuste é implementado de uma forma iterativa, pois caso o perfil correspondente à classificação não se ajuste adequadamente outros perfis são tentados, e caso algum destes outros perfis se ajuste melhor a classificação do objeto é modificada.⁴¹ Uma descrição de como o ajuste de cada perfil é realizado será apresentada no Capítulo 6.

⁴⁰Essa representação é funcional, pois para que o sistema possa ser executado no CNES os módulos não se comunicam entre si, e os dados fluem através de bancos de dados, enquanto o processamento é controlado por um software gestor chamado SAGA.

⁴¹Isso, pois como comentamos anteriormente, no momento não se deve confiar plenamente nas imagens reconstruídas.

1.6 Estrutura do texto

O corpo deste texto está dividido em mais cinco capítulos, cuja estrutura interna é tal que existem três partes bem definidas: na primeira página há um breve resumo em poucos parágrafos do assunto tratado, então múltiplas seções detalham o trabalho, e uma conclusão é apresentada ao final. Desse modo, é possível ter uma idéia global do que foi desenvolvido e concluído em cada um dos capítulos.

No Capítulo 2 apresentamos o princípio de reconstrução de imagens a partir de projeções unidimensionais de dados, comentamos sobre os algoritmos propostos na literatura para uso nos dados do Gaia e apresentamos alguns estudos que realizamos sobre a cobertura espacial e angular destas imagens reconstruídas.

No Capítulo 3 apresentamos um estudo que realizamos sobre a quantidade de fontes no céu que podem ter soluções astrométricas degradadas devido às projeções de outras fontes na mesma linha de visada, apresentamos um método que desenvolvemos para analisar imagens reconstruídas destas fontes problemáticas e recuperar os sinais das diversas fontes presentes na imagem (incluindo aquelas que podem estar acima da magnitude limite do satélite) e mostramos resultados de simulações analisadas por esse método em diferentes regimes de reconstrução de imagens.

No Capítulo 4 iniciamos estudos sobre outro tipo de fonte astronômica que possuirá soluções astrométricas perturbadas, mas agora por razões intrínsecas à sua estrutura: as galáxias não resolvidas em estrelas. Apresentamos um código de simulação de catálogos de galáxias chamado JStuff, que foi implementado por nós, além do código MAGIL responsável por gerar imagens destes objetos—e cujo desenho e especificação possuiu nossa participação. Mostramos também resultados de simulações oficiais do consórcio de análise de dados do Gaia (DPAC) para o céu inteiro que utilizaram o JStuff, e a partir desta simulação comentamos sobre a quantidade de galáxias que deve ser observada pelo Gaia.

No Capítulo 5 apresentamos o método de análise de imagens baseado em medidas que não dependem de perfis específicos para os objetos—medidas que indicam diversos tipos de concentração de luz, suavidade e momentos da distribuição de fluxo e assimetrias—, em um espaço 5-dimensional chamado CASGM20. Mostramos um estudo que realizamos sobre como esses parâmetros variam com a quantidade de informação disponível para sua estimativa e de como eles variam dependendo do tipo de galáxia sob análise. Discorremos sobre o método que utilizamos para classificar galáxias, chamado *Support Vector Machines* e apresentamos testes empíricos que realizamos para verificar se a máxima taxa de classificação teórica é atingida em diversos casos de amostragens aleatórias de distribuições gaussianas. Então, apresentamos resultados da aplicação deste método quando utilizado no volume CASGM20 calculado a partir de imagens de galáxias próximas de um catálogo de imagens da literatura, de galáxias reconstruídas a partir de simulações da missão Gaia, e de galáxias observadas no *Hubble Deep Field*.

No Capítulo 6 mostramos o método que desenvolvemos para realizar uma análise morfológica do perfil de brilho das galáxias não resolvidas em estrelas baseado na re-análise das observações originais do satélite e em simulações destas observações,

que levam em conta distribuições realísticas dos ângulos de trânsito do satélite a cada observação. Apresentamos também resultados obtidos a partir da aplicação deste método em um conjunto de observações simuladas, e comentamos brevemente sobre possibilidades de aprimoramento e aplicação do método em dados além dos que serão obtidos pelo Gaia.

Finalmente, apresentamos as conclusões sobre este trabalho de tese.

Para terminar esse Capítulo, é importante explicitar que a possibilidade de utilizar os dados da missão espacial Gaia para reconstruir imagens e de analisá-las estende as capacidades dessa notável missão para além do que ela foi originalmente concebida, abrindo *águas virtualmente ainda não mapeadas* (Mignard, *pers. comm.*). E, talvez a maior descoberta que poderá ser realizada por meio destas imagens reconstruídas seja algo que ainda não podemos imaginar—mas como Fabian (2009) diz citando Louis Pasteur, *a chance favorece a mente preparada*.⁴²

Esta poderá se tornar a primeira oportunidade que a humanidade terá de vislumbrar todo o céu até magnitude $G \sim 20$ em resoluções que apenas observações do espaço ou com avançados telescópios na Terra munidos de sistemas de óptica adaptativa, permitem. Ainda não existem propostas para uso desta última técnica em levantamentos de céu completo. Entretanto um *survey* será realizado a partir do espaço dentro de poucos anos, nos permitindo navegar por *por mares nunca de antes navegados*⁴³—e ele se chama Gaia.

⁴²*chance favours the prepared mind*

⁴³Do *Canto I* de *Os Lusíadas*, de Luís Vaz de Camões.

Reconstrução de imagens

*“O microfone não deixa de ser uma massa metálica.
Capta apenas 30 por cento do som original.” S. Celibidache¹*

Conteúdo

2.1	Introdução	34
2.2	Um <i>toy model</i> para reconstrução	35
2.2.1	Reconstrução pela transformada de Radon	37
2.3	Reconstrução de imagens para o Gaia	45
2.3.1	Algoritmos de reconstrução	47
2.3.2	QuickStack	47
2.3.3	Drizzle	48
2.3.4	ShuffleStack	49
2.3.5	BinOutliers	50
2.3.6	Regularização de Tikhonov	50
2.3.7	Clean & Cleanest	52
2.3.8	Exemplos de reconstruções	53
2.4	Cobertura espacial e angular das reconstruções	56
2.4.1	Simulação dos trânsitos	58
2.4.2	União de polígonos	60
2.4.3	Cálculo na nuvem	61
2.4.4	Resultados	61
2.5	Conclusões	65

Vimos no Capítulo 1 que o GAIA não transmitirá para a Terra imagens em duas dimensões. Entretanto, o conhecimento do sinal em duas dimensões será necessário em algumas situações, como na análise de fontes extensas extrínsecas (ex. fontes secundárias perturbadoras além da magnitude limite) ou intrínsecas (ex. galáxias).

Algoritmos que possibilitarão uma reconstrução aproximada desse sinal a partir das observações do satélite estão sendo propostos por vários pesquisadores. Devido à importância das imagens assim obtidas para este trabalho, neste Capítulo discorreremos sobre a problemática da reconstrução, os algoritmos atualmente disponíveis e apresentamos um estudo que realizamos sobre as coberturas espacial e angular dessas imagens reconstruídas na esfera celeste (Krone-Martins et al, in prep.b).

¹in Folha de São Paulo, 5 de outubro de 1993.

2.1 Introdução

Devido ao tamanho do plano focal do Gaia e às limitações de banda de comunicação existentes para transferência de dados entre a Terra e o satélite, seria impraticável a transferência do conteúdo completo da informação ali presente. Deste modo, foi implementado um sistema de processamento a bordo que seleciona os dados que serão transferidos (descrito em [Astrium, 2006](#)).

Durante o trânsito do objeto pelo plano focal do Gaia, seu sinal cruza diversas colunas de CCDs (ver Figura 1.9, Capítulo 1). Está definido que os dados referentes ao objeto serão transmitidos para a Terra apenas caso esse objeto seja detectado em qualquer uma das duas colunas dos *SkyMapper* e caso essa detecção seja posteriormente confirmada durante o trânsito pelo *AstroField 1*.

Esses dados transmitidos consistem em áreas denominadas “janelas” (*windows*), que são regiões em torno do objeto compostas por um conjunto de *samples*. Esses *samples*, por sua vez, equivalem a somas dos sinais provenientes dos pixels das CCDs. Na linguagem astronômica corrente um *sample* seria equivalente à um novo pixel formado por *binning*. Uma janela distinta é gerada cada vez que o objeto cruza uma coluna de CCDs, de modo que considerando apenas as colunas *SM* e *AF*, para cada trânsito do satélite estarão disponíveis um máximo de 10 janelas por objeto.²

O tamanho exato das janelas transmitidas depende da magnitude do objeto e em qual coluna de CCDs ela foi gerada. A Tabela 2.1 explicita os tamanhos lidos no CCD (em pixels) e os tamanhos transmitidos para Terra (em *samples*) em função da magnitude Gaia *G*.

Coluna de CCD	Magnitude <i>G</i>	Janela lida (pixels)	Binagem (pixels)	Janela transmitida (<i>samples</i>)	Tamanho angular (")
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

Tabela 2.1: Dados das janelas transmitidas para a Terra de acordo com a coluna de CCD e magnitude estimada do objeto. As colunas “Binagem” e “Janela transmitida” foram obtidas do *Gaia Parameter Database* ([GPDB_2010-03-02, 2010](#)).

²Isso, pois além das nove colunas *AF*, o objeto será detectado apenas no *SM1* ou no *SM2*, uma vez que ele pode ser observado por apenas um dos telescópios.

Para possibilitar a detecção de objetos móveis e corrigir qualquer problema de acompanhamento e erro de centragem, a “janela” que é lida a bordo, possui em certos casos, uma “binagem” diferente, podendo até mesmo ter tamanhos físicos distintos (com mais pixels do que aqueles que serão “binados” e transmitidos). Procedimentos de re-centragem e re-amostragem são utilizados pelo sistema de processamento de vídeo do satélite para que apenas *samples* significativos sejam enviados para a Terra.

Nas colunas *AF* 3, 4, 6, 7 e 9, por exemplo, um objeto com $G > 16$ terá uma janela total lida a bordo formada por 12×12 pixels, binados em 1×12 pixels, gerando os 12×1 *samples* que serão analisados pelo subsistema de re-centragem. Esse subsistema determina as 6 colunas de pixels mais importantes para a definição do objeto, que serão transmitidas para a Terra na forma de uma janela de 6×1 *samples*, equivalendo à janela lida de 6×12 pixels apresentada na Tabela 2.1 (os tamanhos totais lidos a bordo pode ser consultados no *Gaia Parameter Database*).

Então, o problema de reconstrução de imagem consiste na determinação de uma “imagem bi-dimensional” do objeto observado, a partir dos dados destas janelas essencialmente unidimensionais—aliado, tanto quanto possível, ao conhecimento do comportamento do satélite Gaia em termos de óptica e atitude.

Essa “imagem bi-dimensional” reconstruída é também chamada de “mapa dos arredores”, visto que será especialmente útil na análise de possíveis emissões originadas no entorno da fonte observada—sejam elas de origem intrínseca, no caso de objetos extesos, ou extrínseca no caso de fontes secundárias.

2.2 Um *toy model* para reconstrução

Com a finalidade de explicitar o que se quer dizer com reconstrução de imagem no contexto do Gaia e introduzir as idéias básicas de algoritmos de reconstrução, vamos considerar nesta seção um problema idealizado baseado na observação de duas estrelas de mesmo fluxo projetadas aproximadamente na mesma linha de visada.

Esse sistema também servirá para mostrarmos que para a reconstrução exata de uma imagem bi-dimensional com $k \times k$ pixels a partir de projeções unidimensionais com k pixels (mantendo assim a resolução angular), é condição necessária, mas não suficiente, que ao menos k projeções estejam disponíveis para análise.

Como este é apenas um *toy model*, assumimos que a imagem é completamente definida pelas fontes, não existindo nenhum tipo de ruído aditivo ou influência do instrumento utilizado—consideraremos no entanto, em um segundo momento, um ruído na determinação do sinal.

Iniciamos supondo que o objeto seja observado duas vezes, em dois ângulos de trânsito que formem um ângulo reto entre si, e que os dados disponíveis para análise sejam apenas estas duas projeções unidimensionais³. Um esquema do problema e da observação pode ser visto na Figura 2.1.

Então, o problema direto que consiste em dizer qual é a observação a partir do

³Esse tipo de observação “unidimensional” é similar às janelas das colunas *AF*, cujos *samples* são formados pela integração dos pixels na direção perpendicular ao trânsito.

Figura 2.1: *Toy model* com a matriz $\mathbf{S}$ de duas estrelas e suas projeções nas direções perpendicular $v_{\perp}$ e paralela $v_{\parallel}$ ao eixo x .

conhecimento da fonte pode ser facilmente escrito: Seja a fonte dada por um sinal discretizado no espaço bi-dimensional representado por uma matriz $\mathbf{S}$, com (n, m) componentes S_{ij} , o vetor referente à observação do trânsito perpendicular⁴ $\mathbf{v}_{\perp}$ à direção definida pela da linha da matriz é escrito como:

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

e de forma análoga, o vetor referente à observação do trânsito paralelo $\mathbf{v}_{\parallel}$ é:

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

Então, o problema inverso que consiste na determinação dos elementos da matriz $\hat{\mathbf{S}} \sim \mathbf{S}$ a partir dos vetores de observações $\mathbf{v}_{\parallel}, \mathbf{v}_{\perp}$, pode ser descrito pelo sistema linear a seguir:

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

Observando o sistema acima, percebe-se que esse *toy model* gera um sistema possível mas indeterminado para quaisquer n, m tais que $n > 2$ ou $m > 2$. Isso, pois

⁴Um trânsito perpendicular integrará o sinal na direção paralela às linhas da matriz.

o número de icógnitas (número de casas de $\hat{\mathbf{S}}$) é $(n \times m)$ enquanto o número de equações é $(n + m)$.

Assumindo uma imagem retangular, com $n = m = k$, existem k^2 icógnitas à resolver, e como cada observação fornece k equações, para tornar o sistema possível e determinado são necessárias k projeções $\mathbf{v}_\theta$ ⁵. No entanto, ainda que esse número de equações seja necessário, sabemos que isso não é uma condição suficiente pois nada se diz sobre a dependência entre essas equações.

A ideia de resolver o problema inverso por meio de uma resolução de sistemas lineares como descrevemos de forma quase intuitiva é a raiz dos diversos métodos iterativos, do qual o mais conhecido é a Técnica de Reconstrução Algébrica⁶. Nesses métodos, não se busca uma solução exata para o sistema das equações geradas pelas projeções, e o que se faz é adotar algum esquema de resolução iterativa para determinar os elementos da matriz $\hat{\mathbf{S}}$ que otimizem algum critério de qualidade da solução escolhido de forma *ad hoc*.

Nestes métodos se re-escreve o problema na forma $\mathbf{AX} = \mathbf{B}$, onde $\mathbf{A}$ é uma matriz retangular (geralmente esparsa) que contém os pesos das contribuições de todos os pixels de $\mathbf{S}$ para as casas de $\mathbf{v}_\theta$, $\mathbf{X}$ é uma matriz coluna com as icógnitas que se busca resolver (as casas de $\hat{\mathbf{S}}$ escritas na forma de coluna) e $\mathbf{B}$ é uma matriz coluna com as observações (os vetores $\mathbf{v}_\theta$ escritos todos juntos). Uma descrição extensa desses métodos e dos algoritmos utilizados para resolução do sistema pode ser encontrada por exemplo em [Kak & Slaney \(2001\)](#).

Como esses métodos são iterativos, a princípio é possível incorporar com facilidade, informações conhecidas *a-priori* sobre os elementos da matriz, como por exemplo não-negatividade, formatos permitidos para o objeto que se tenta reconstruir, condições de suavidade da imagem (filtrando de forma natural o ruído tipo *sal e pimenta*, ou raios cósmicos). Entretanto, segundo [Marr \(1979\)](#) eles são muito custosos computacionalmente, sendo que para certos problemas sua utilização chega a ser até mesmo proibitiva—o que também se deve a necessidade de realizar diversas iterações para se obter as imagens finais, sendo que cada iteração equivale na realidade a uma nova reconstrução de imagem.

2.2.1 Reconstrução pela transformada de Radon

Uma opção mais eficiente do que a reconstrução algébrica para determinação de $\hat{\mathbf{S}}$ é o uso da transformada inversa de Radon ([Radon, 1917](#))⁷. Grande atenção tem sido voltada à essa transformada nas três últimas décadas, em especial devido ao fato de que ela fornece uma descrição ideal para medidas tomográficas na área de Ciências da Saúde. Descrições de suas propriedades e aplicações (por exemplo em reconstruções de tomografia por emissão de pósitrons⁸) podem ser encontradas em

⁵Para determinação do vetor $\mathbf{v}_\theta$ de observação para qualquer ângulo de trânsito, basta realizar a rotação de $\mathbf{S}$ no ângulo θ correspondente e realizar uma soma ponderada das casas de $\mathbf{S}$ pela sua contribuição para a i -ésima casa de $\mathbf{v}_\theta$.

⁶Ou ART, *Algebraic Reconstruction Technique*.

⁷Traduções para o inglês em [Deans \(1983\)](#) e [Parks \(1986\)](#).

⁸Ou PET, *Positron Emission Tomography*.

Deans (1983), Toft (1996) ou Herman (2009).

Essa transformada já foi utilizada anteriormente em Astronomia, sendo que na região do óptico, propostas de sua utilização podem ser encontradas desde Aime et al (1978), e mais recentemente em Touma (1997) para uso em um tipo especial de telescópio chamado de “telescópio de fenda”.⁹

Em dados obtidos no espaço, a transformada de Radon já foi utilizada para reconstrução de imagens do céu inteiro em raios- γ a partir de observações com satélite BATSE de ocultações pela Terra (Zhang et al, 1993), e mais recentemente em Case et al (2009) onde os autores ainda reportam que esta técnica será adaptada para uso na produção de um mapa do céu em γ com os dados do satélite Fermi.

Essa transformada corresponde a uma descrição do problema que estamos tentando resolver em termos de funções contínuas ao invés de matrizes discretas, supondo que ele possa ser descrito por uma função contínua de duas variáveis $f(x, y)$.

Definição 1 *Seja $f(x, y)$ uma função contínua, a transformada de Radon é formada por todas as integrais de linha de $f(x, y)$ calculadas sob todas as retas L existentes*¹⁰.

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

Uma representação de uma função $f(x, y)$ e de uma reta L quaisquer nos sistemas de coordenadas envolvidos na transformada pode ser visto na Figura 2.2.

Figura 2.2: Representação gráfica das grandezas e dos sistemas de coordenadas envolvidos na transformada de Radon.

A partir da Definição 1 e da Figura 2.2, percebe-se que é possível escrever explicitamente a transformação $\check{f}$ de $f(x, y)$ sob uma linha L qualquer. Adotando

⁹Adotado por exemplo em Aime et al (1978) para observações via interferometria *speckle* do fenômeno de granulação solar no *Kitt Peak National Observatory*, e proposto por Martin et al (1986) como princípio para um telescópio espacial voltado ao imageamento de alta resolução

¹⁰A definição mais geral, para uma função em $\mathbb{R}^n$, é a integral dessa função sob todos os hiperplanos de dimensão $(n - 1)$, e.g. Deans (1983).

o sistema de coordenadas com eixos rotacionados pelo ângulo ϕ (representado no diagrama da Figura 2.2 pelas retas tracejadas), temos a seguinte transformação de variáveis:

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

Que permite então escrever a transformação na forma:

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

Quando se conhece o valor da função $\check{f}$ para quaisquer (p, ϕ) , diz-se conhecer a transformada de Radon da função f . Entretanto em sua aplicação, a transformada $\check{f}$ em si nunca é obtida, e o que se conhece é apenas seu valor amostrado em alguns pontos do espaço de (p, ϕ) , denominado espaço de Radon—a representação gráfica desse espaço denomina-se *Senograma*¹¹.

A partir da equação 2.2, vemos que $\check{f}(p, \phi)$ equivale a uma projeção de $f(x, y)$ no ponto p e no ângulo de trânsito ϕ . Desse modo, fixar ϕ e varrer um subconjunto discretizado no eixo p é similar a determinar o vetor $\mathbf{v}_\phi$ de uma observação realizada com ângulo ϕ . O que deseja-se então é a partir dos vetores $\mathbf{v}_\phi$ que formam uma amostra do espaço de Radon, obter a aproximação para a função $f(x, y)$, invertendo o problema.

Em seu artigo original (Radon, 1917) a fórmula para inversão era conhecida, mas como foi provado em um teorema de Smith et al (1977) citado por Deans (1983): *Uma função f de suporte compacto em $\mathbb{R}^2$ é unicamente determinada por qualquer conjunto infinito, mas não é unicamente determinada por nenhum conjunto finito, de suas projeções*. Na prática, isso significa que qualquer solução obtida a partir da inversão das observações será apenas uma aproximação possível para o sinal original.

A técnica não iterativa mais simples e direta para encontrar uma inversão aproximada para essa transformada é chamada de *Backprojection*, que como o próprio nome sugere, é apenas uma projeção na direção inversa das funções transformadas. A Figura 2.3 representa o resultado obtido por essa técnica para dois vetores de observação e dois casos diferentes do *toy model*. Note como essa técnica gera uma versão “borrada” da imagem original.

Embora essa técnica seja raramente utilizada sozinha, visto que a função $\check{f}$ obtida após sua aplicação é uma versão muito “borrada” da função original f , ela é um passo intermediário daquela que é provavelmente a técnica mais utilizada para resolver este tipo de problema: *Filtered Backprojection*.

Existe uma estreita relação entre as transformadas de Fourier e Radon: realizar uma transformada de Fourier em duas dimensões $\mathcal{F}_2 f(x, y) = \tilde{f}(k_x, k_y)$ equivale a realizar uma transformada de Radon $\mathcal{R}f(x, y) = \check{f}(p, \phi)$ e depois uma transformada de Fourier em uma dimensão na direção radial $\mathcal{F}_1 \check{f}(p, \phi) = \tilde{\tilde{f}}(k, \phi)$ (ver Deans, 1983, para dedução). A diferença entre as funções $\tilde{f}$ e $\tilde{\tilde{f}}$ obtidas é apenas devida às

¹¹ A nomenclatura tem origem no fato de uma função δ nesse espaço ser uma função trigonométrica.

Figura 2.3: *Toy model* com a matriz $\hat{\mathbf{S}}$ recuperada pela técnica de *Backprojection* aplicada em duas observações ($v_{\perp}$ e $v_{\parallel}$). O modelo à esquerda é o mesmo da Figura 2.1. No modelo da direita, os dois objetos encontram-se nas posições $(-0.15''; 0.0'')$ e $(0.3''; 0.3'')$.

variáveis, de modo que adotando uma transformação de variáveis conveniente¹² elas se tornam completamente equivalentes¹³. A Figura 2.4 esquematiza essa relação.

Figura 2.4: Relação entre as transformadas de Radon e de Fourier (adaptado de Deans, 1983): realizar uma transformada de Radon e após uma transformada de Fourier na dimensão radial de uma função qualquer, equivale a realizar uma transformada de Fourier nas duas dimensões originais.

Devido a essa relação entre as duas transformadas, pode-se utilizar a transformada de Fourier para a inversão de $\tilde{f}$ (a função que é estimada a partir das observações, ou vetores $\mathbf{v}_{\phi}$). O diagrama da Figura 2.5 esquematiza o processo de inversão mais amplamente adotado, que é chamado de *Filtered Backprojection*.

Com a finalidade de verificar os resultados obtidos com sua utilização, bem como para analisar a “eficiência” obtida na reconstrução da imagem e sua suscetibilidade à presença de ruído nas observações unidimensionais, aplicamos esse método ao nosso *toy model*, utilizando o algoritmo descrito em Toft (1996) e implementado em R (R Development Core Team, 2009) por Schulz (2006).

¹²A transformação é $k = (k_x^2 + k_y^2)^{1/2}$ e $\phi = \tan^{-1}(k_y/k_x)$.

¹³No caso discretizado, a adoção de uma interpolação é necessária.

Figura 2.5: Inversão da transformada de Radon por *Filtered Backprojection*. Em vermelho está indicada a função que podemos estimar a partir das observações, e em azul a função invertida ao final do processo (Adaptado de [Deans, 1983](#)).

Neste contexto a “eficiência” da reconstrução deve ser uma grandeza que servirá para realizarmos uma comparação entre a qualidade de reconstruções diferentes, além de mensurar individualmente o quão boa foi uma determinada reconstrução. Intuitivamente, o contradomínio dessa função deve ser $\{x \in \mathbb{R} | 0 \leq x \leq 1\}$, adotando o valor nulo no caso de uma reconstrução muito distante do original e o valor unitário para uma reconstrução perfeita (idêntica ao original). Deste modo, utilizando a operação *modinv* definida no Apêndice “Módulo Inverso”, podemos definir:

Definição 2 *Sejam S_{ij} e $\hat{S}_{ij}$ os valores do sinal original e do pixel reconstruído na posição (i, j) , definimos a eficiência E_{ij} da reconstrução do pixel (i, j) por:*

$$E_{ij} = \begin{cases} \frac{\hat{S}_{ij}(\text{modinv } S_{ij})}{S_{ij}} & , \text{ se } \hat{S}_{ij} \leq 2S_{ij} \\ 0 & , \text{ se } \hat{S}_{ij} > 2S_{ij} \end{cases} \quad (2.3)$$

Realizamos diversas simulações de nosso *toy model* para verificar o comportamento da função eficiência quando o algoritmo de *Filtered Backprojection* é exposto a casos com diferentes números de trânsito e níveis de ruído no sinal.

Como se pode verificar qualitativamente na Figura 2.6, algum tipo de informação bi-dimensional é recuperado com apenas três observações, ao ponto de possibilitar a reconstrução do sinal dos dois objetos em um nível de eficiência um pouco maior do que $\sim 60\%$. Claramente, devido ao pequeno número de observações, a reconstrução dos dois objetos somente foi possível devido a existência de ao menos uma projeção adequada dos dados.

Já com 10 observações, essa eficiência sobe para $\sim 80\%$ e o formato do objeto se torna essencialmente pontual, ainda que o contraste com o fundo não seja ótimo. Com 50 projeções, que é um número de projeções menor do que o esperado para as observações do Gaia, a eficiência para os pixels do objeto atinge valores $> 90\%$.

Figura 2.6: Reconstrução de nosso *toy model* pela transformada inversa de Radon usando diferentes números de projeções ($n=3, 10, 50, 80$). As projeções foram distribuídas regularmente entre 0 e π radianos e nenhum ruído foi adicionado.

Figura 2.7: Reconstrução de nosso *toy model* pela transformada inversa de Radon usando diferentes níveis de ruído no senograma ($S/R \sim 1, 5, 10, 40$). Todas as reconstruções utilizaram 80 projeções regularmente distribuídas entre 0 e π radianos. Um *bias* constante equivalente a 1% do sinal médio foi adicionado.

No entanto, os resultados da análise anterior são modificados ao se considerar a existência de um nível de *bias* e um ruído na observação¹⁴, como pode ser visto na Figura 2.7, que apresenta reconstruções para diferentes níveis de S/R simulados.

Considerando 80 projeções (aproximadamente a quantidade média de trânsitos do Gaia em qualquer coordenada), para um nível de S/R de 1, vemos que embora a imagem reconstruída para um objeto duplo seja bastante diferente daquela que seria gerada por um objeto pontual simples, fica essencialmente impossível a distinção com um objeto extenso. Entretanto, para $S/R = 5$ a situação se torna mais satisfatória, e embora os objetos ainda sejam levemente deformados, já é possível perceber a natureza de objeto duplo do sinal—resultado que naturalmente se torna ainda melhor ao elevarmos essa razão para $S/R \geq 10$.

Figura 2.8: Eficiência na reconstrução pela transformada inversa de Radon em função do número de trânsitos (com $S \gg R$ a cada observação) e em função do S/R (para 80 trânsitos). Os pontos indicam a eficiência para cada pixel e o mapa de cores usado como fundo representa uma estimativa de densidade dos pontos.

Definindo as eficiências média e mediana como as médias ou medianas das eficiências de todos os pixels da imagem, e varrendo diversos casos diferentes de números de trânsitos e níveis de S/R , podemos analisar o comportamento global da função de eficiência do algoritmo de reconstrução em relação àquelas variáveis.

A Figura 2.8 representa os resultados destas simulações, e podemos observar que o comportamento de ambas as taxas (média ou mediana) é essencialmente o mesmo: a partir de aproximadamente 15 trânsitos ou $S/R \sim 20$ a eficiência na reconstrução atinge um patamar estável em torno de $\sim 80\%$.

É importante notar que essa taxa foi definida, e está sendo calculada, para todos os pixels da imagem, e não apenas aqueles nos quais o sinal dos objetos é significativo—no caso de considerar apenas pixels, como podemos estimar pelos

¹⁴O ruído foi simulado como indeterminação do sinal a cada pixel, e a presença de uma componente aditiva não foi considerada.

mapas de eficiência das Figuras 2.6 e 2.7, a eficiência seria ainda maior.

A partir dos resultados obtidos nesta subseção, podemos concluir que o *Filtered Backprojection* seria uma técnica candidata a ser testada no futuro para reconstrução das observações simuladas do Gaia. Entretanto ela não pode ser aplicada diretamente aos dados deste satélite, devendo sofrer algumas adaptações. Na próxima seção veremos a razão para isso.

2.3 Reconstrução de imagens para o Gaia

O caso geral da reconstrução para as janelas do Gaia é um problema mais complicado do que o nosso *toy model* por diversos fatores. O fator complicante mais simples, é termos considerado na seção anterior, que todas as observações eram similares do ponto de vista do detector. No caso do Gaia, cada coluna de CCD do plano focal é, e se comporta, na realidade, como um detector diferente, onde até mesmo a amostragem do sinal é distinta. Um algoritmo de reconstrução ideal aplicável ao Gaia deveria levar em conta ao menos os seguintes fatores:

1. Janelas de diversos tipos, com samples de tamanhos angulares diferentes

Como visto na introdução deste Capítulo, a janela obtida em cada coluna de CCDs é distinta tanto do ponto de vista de tamanho angular quanto da estratégia de amostragem adotada. Isso significa que um algoritmo de reconstrução ideal seria capaz de levar isso em conta, e reconstruir uma imagem que não seja degradada pela direção de menor resolução dos *samples* dos *AFs* ou pelos *samples* espessos (2x2 pixels) dos *SMs*.

2. Janelas dos *AFs* possuem sinal proveniente de ambos os campos de visão

Além do problema dos tamanhos diferentes, as janelas possuem informações diferentes. No caso de uma reconstrução utilizar o sinal proveniente das janelas obtidas pelos *AFs*, existe o inconveniente de que essas janelas possuem um sinal composto pela adição dos sinais dos dois telescópios do Gaia. Um algoritmo ideal poderia, a partir das múltiplas observações do mesmo objeto, bem como das janelas dos *SMs*, desconsiderar parte do sinal dos *AFs*.

3. Point Spread Function muito assimétrica

O espelho do Gaia é retangular, de modo que a *PSF* gerada não é circularmente simétrica (formando um padrão de *Airy*) e possui um formato de cruz. Isso atrapalha a reconstrução, pois caso o algoritmo adotado não leve esse fator em conta, ao rotacionar as observações durante a reconstrução um certo pixel da matriz final, o sinal espalhado pela *PSF* irá degradar a imagem reconstruída, causando um efeito de “suavização”. Além desse fator, a *PSF* ainda varia de acordo com a posição no plano focal, e dependendo do CCD em que a observação é realizada a *PSF* que deve ser considerada na reconstrução possui uma forma diferente (ver Figura 2.9, por exemplo).

Figura 2.9: PSFs simuladas dos *SMs*. (a) Coluna 1, Linha 1. (b) Coluna 2, Linha 3. (De [Mary et al, 2006](#), , simulações de GIBIS 4).

4. *CTI - Ineficiência na transferência de carga*

Como visto no Capítulo 1, as observações do Gaia são realizadas no modo TDI, onde as cargas são transferidas de um pixel ao outro na mesma velocidade em que o satélite rotaciona. Entretanto, a transferência dessas cargas não é perfeita, e embora a maior parte dos fotoelétrons sejam de fato transferidos, a eficiência não é completa, restando um pouco de carga. Isso é um problema para a reconstrução, pois o formato dos objetos é modificado na direção ao longo do trânsito (até mesmo a posição dos objetos é levemente alterada), e o efeito pode ser fortemente dependente de fatores externos ao objeto (como outros objetos brilhantes que transitaram antes ou raios cósmicos que atingiram a CCD). Um algoritmo de reconstrução ideal deveria levar esse fato em consideração, provavelmente corrigindo *a priori* as janelas observadas com a adoção de algum modelo de distorção, como o CDM02, descrito em [Short \(2009\)](#).

5. *Efeitos temporais*

As observações serão realizadas ao longo de aproximadamente 5 anos, durante os quais o instrumento estará exposto ao ambiente espacial, e com isso deverá sofrer modificações (por exemplo defeitos em pixels da CCD) que influenciarão os dados das janelas transmitidas, e também o ruído instrumental não será estritamente o mesmo a cada nova observação de uma mesma fonte. Além disso, é possível ocorrer variações intrínsecas nas próprias fontes, de modo que um algoritmo ideal deveria ser capaz não apenas de reconstruir uma boa imagem de fontes não variáveis, como também seria capaz de detectar que algum tipo de variação temporal ocorreu, excluindo essa observação da reconstrução e lançando algum tipo de alarme.

2.3.1 Algoritmos de reconstrução

A obtenção de uma solução ideal¹⁵ para o problema de reconstrução é bastante complicada, e requer um tempo de processamento relativamente longo (ver, por exemplo Dollet, 2004). Além disso, durante a missão espera-se ter a necessidade de reconstruir dezenas de milhões de imagens para fins de análise de problemas causados por projeções ópticas de fontes secundárias (Capítulo 3) e para análise de galáxias não resolvidas (Capítulos 5 e 6). Como durante a missão, a reconstrução e análise de todas essas imagens deve ser realizada em um espaço de tempo de poucos dias, será necessária a adoção de soluções simplificadas, e que embora ignorem muitos dos efeitos descritos na Subseção 2.3, tornem a análise desses dados uma tarefa possível.

Entretanto, vimos na seção 2.2 que é possível utilizar algoritmos para recriar imagens bi-dimensionais de objetos e de qualquer informação que esteja em seus arredores a partir de dados de observações similares às “janelas” do Gaia. E isso, mesmo se as “janelas” que contêm alguma informação bi-dimensional (como aquelas geradas pelos *SkyMappers*) não estiverem disponíveis.

Descreveremos aqui seis algoritmos atualmente propostos para a reconstrução de imagens. Os quatro primeiros foram propostos e/ou implementados por D. Harrison, e desde o início foram pensados tendo as limitações de tempo de processamento em mente. Os dois últimos são algoritmos mais detalhados e rigorosos, propostos e/ou implementados por Dollet&Bijaoui e Pereyga&Bijaoui, que embora levem em conta fatores de degradação da imagem (como os descritos na Subseção 2.3), são muito mais exigentes do ponto de vista computacional, tendo sua aplicabilidade dificultada durante o decorrer da missão. Nas descrições a seguir, a matriz que armazena a imagem reconstruída final será sempre escrita como $\hat{\mathbf{S}}$.

2.3.2 QuickStack

O QuickStack foi desenvolvido por D. Harrison para ser o algoritmo mais simples e rápido que possibilite uma reconstrução de imagens a partir das janelas transmitidas pelo Gaia. Ele é formado por uma mistura entre a *Backprojection*, comentada anteriormente, e um empilhamento simples das janelas bi-dimensionais.

O primeiro passo é a inicialização de $\hat{\mathbf{S}}$ como uma rede formada por pixels quadrados com determinada resolução. Então, para as janelas unidimensionais originadas dos *AFs*, é realizada uma reconstrução que pode ser vista como uma forma de *Backprojection* (com re-amostragem implícita), recobrindo apenas os pixels cujos dados estariam fisicamente dentro daquelas janelas (isso é importante devido ao pequeno tamanho dessas janelas).

Para as janelas bi-dimensionais dos *SMs*, realiza-se uma rotação (com re-amostragem também implícita), o valor do *sample* é somado diretamente no pixel de $\hat{\mathbf{S}}$ que é recoberto por aquele *sample*, e o valor final é dividido pelo número de *samples*,

¹⁵Aqui definida como uma solução que leve em conta todos os fatores conhecidos que influenciem a degradação da imagem. Nada dizemos sobre a qualidade final da imagem obtida.

ou seja uma média é calculada. Finalmente, a imagem formada pelo empilhamento (soma) de todas essas janelas é considerada como a imagem reconstruída.

Esse algoritmo tem a característica de gerar imagens com aparência “borrada” (ex. Figuras 2.11 e 2.12). E além do cálculo da média dos *samples*, um outro fator que contribui para isso é que a existência da PSF do instrumento não é levada em conta pelo algoritmo: o formato da PSF do Gaia é similar a uma cruz assimétrica alinhada com a janela.

2.3.3 Drizzle

Um segundo algoritmo que foi implementado é chamado *Drizzle*, inventado por A. Fruchter e R. Hook (Fruchter & Hook, 2002) para empilhamento de imagens do telescópio espacial Hubble, e que foi aplicado com sucesso na obtenção das imagens finais do *Hubble Deep Field* (Williams et al, 1996). Sua implementação atual deve-se à D. Harrison, embora ele tenha sido proposto para uso no Gaia pela primeira vez em Nurmi (2003).

Assim como no QuickStack, o primeiro passo é a inicialização de $\hat{\mathbf{S}}$ como uma rede formada por pixels quadrados, com determinada resolução.

A idéia básica desse algoritmo consiste na abstração da imagem reconstruída como um coletor de fótons e das imagens observadas (no atual contexto, as janelas do Gaia) como fontes de fótons: os pixels de cada imagem observada são reduzidos até uma fração do tamanho original (um parâmetro livre), gerando o que na linguagem do algoritmo se denominam “gotas”, e posteriormente cada uma dessas “gotas” é “derrubada” em cima do grid final $\hat{\mathbf{S}}$.

O valor adicionado em cada pixel S_{ij} depende da fração do recobrimento mútuo entre a “gota” e o pixel da rede, e essa adição é ponderada por um peso escolhido pelo usuário. As equações originais de Fruchter & Hook (2002) são:

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

onde os índices i indicam as janelas observadas e o a imagem reconstruída, d o valor do *sample* (“gota”) da janela, a a proporção de recobrimento entre os *samples* e os pixels da imagem reconstruída ($0 \leq a \leq 1$), $\mathbf{W}$ a matriz dos pesos w_{ij} , e $\mathbf{I}$ a matriz da imagem reconstruída e $'$ indica que o procedimento é iterativo (a adição de cada janela equivale a uma iteração). Em nossa notação, a matriz formada pelos valores $I_{x_o y_o}$ após todas as observações serem levadas em conta é a matriz $\hat{\mathbf{S}}$. Pode-se dizer que esse algoritmo se reduz ao QuickStack quando se adota que para todos os pixels $w = s = 1$ e $a = 1$ quando existe recobrimento ou $a = 0$ quando não existe.

Esse algoritmo possibilita uma melhora na resolução da imagem reconstruída devido à correção da amostragem (ele pode ser visto como uma técnica de *dithering*). Entretanto, ele não resolve o problema da degradação pela rotação da PSF, e ainda que a reconstrução obtida possa atingir maiores resoluções do que aquelas geradas pelo QuickStack, ele gera artefatos e consome mais tempo de execução.

Figura 2.10: Esquema da técnica de Drizzle. Note como o pixel central da imagem reconstruída não recebe nenhum sinal.

Além disso, o método não possui uma descrição de como o valor para a fração de redução do pixel (parâmetro de escala) deve ser determinado, e uma escolha empírica deve ser realizada para que o tamanho das gotas não seja grande demais para que se ganhe resolução, nem pequeno demais para evitar degradar a imagem e para recobrir todos os pixels da imagem final.

2.3.4 ShuffleStack

Outro algoritmo desenvolvido por D. Harrison, o ShuffleStack é na realidade um QuickStack mais sofisticado. Da mesma forma que o primeiro algoritmo, o ShuffleStack funciona a partir de um empilhamento simples das janelas observadas, mas ele possui três sutilezas importantes que o diferenciam permitindo que ele atinja uma melhor qualidade na reconstrução das imagens.

A primeira, e mais importante, é que o algoritmo funciona de modo a reconstruir os pixels com maior quantidade de informação (ou seja, com maior número de *samples* contribuindo para sua estimativa) antes dos outros. Para isso, uma análise preliminar é realizada partindo de cada pixel da imagem final e determinando todas as janelas (e as posições de seus *samples*) que contribuem para aquele pixel.

Então, cada pixel da imagem final é reconstruído, do pixel com maior número de *samples* ao pixel com menor número. Essa reconstrução é realizada a partir do cálculo de uma média com rejeição iterativa a 2σ (valor definido empiricamente por D. Harrison). O valor médio é então subtraído dos *samples* que contribuíram para aquele pixel—o que batiza o algoritmo, pois o sinal acaba sendo movido, transferido, dos pixels com menos observações aos pixels com mais observações.

É importante notar que não existe nenhuma garantia que esse algoritmo não reconstrua pixels com valores negativos—o que ocorre em alguns casos—e que esse

algoritmo possui a característica de concentrar sinal naqueles pixels que forem observados um maior número de vezes.

2.3.5 BinOutliers

Na maior parte dos casos em que será necessário reconstruir uma imagem durante a missão, se busca analisar é algum tipo de perturbação da solução astrométrica da fonte primária, localizada aproximadamente ao centro da imagem reconstruída. Deste modo, D. Harrison, desenvolveu um algoritmo que leva este fator em consideração.

Assim como o ShuffleStack, o BinOutliers é um algoritmo derivado do QuickStack. Entretanto, seu ponto de partida é identificar quais *samples* contribuem para os pixels “centrais”¹⁶ de $\hat{\mathbf{S}}$, e evitar que eles sejam considerados na reconstrução dos pixels subjacentes. Isso, pois no geral o que se busca observar são objetos mais fracos do que o objeto central e a contribuição desse objeto central pode ofuscar esses objetos mais fracos.

Da mesma forma que o ShuffleStack, os pixels da imagem final são reconstruídos um a um, mas sem ordem preferencial. Primeiro calculam-se a média m_B e o desvio padrão σ_B a partir dos valores dos 70% *samples* de menor valor que contribuem para aquele pixel na imagem final (para excluir *outliers*). Então, a reconstrução do pixel é realizada a partir do cálculo da média dos valores dos samples após uma rejeição dos valores dos samples mais distantes do que $2.5\sigma_B$ de m_B (valor definido empiricamente por D. Harrison). Também de forma distinta do ShuffleStack, o valor reconstruído não é subtraído dos *samples* que contribuíram para o pixel.

As imagens obtidas em reconstruções que utilizam esse algoritmo possuem uma estrutura similar a que é gerada pelo ShuffleStack, mas são muito mais suaves—a não ser pelos pixels centrais, que tendem a concentrar o sinal gerado pela estrela primária. Isso ocorre justamente porque o algoritmo não faz o sinal dos *samples* se mover pela imagem, como ocorre no ShuffleStack, levando a menos deformações nos objetos reconstruídos (ex. Figura 2.11).

Como será visto no próximo Capítulo, esse algoritmo é, em geral, mais eficaz do que os outros três para análise da multiplicidade em imagens reconstruídas.

2.3.6 Regularização de Tikhonov

Originalmente descrita por Tikhonov (1976) para solução de problemas mal-postos (Demoment, 2002), sua adaptação para reconstrução de imagens para o Gaia foi proposta por Claire Dollet sob orientação de Albert Bijaoui (Dollet, 2004; Dollet et al, 2005), e posteriormente analisada em Mary et al (2006).

Ela pode ser vista como uma técnica de reconstrução algébrica. Seu primeiro passo é escrever o que se passa durante as observações do Gaia na forma matricial $\mathbf{D}_i = \mathbf{H}_i \mathbf{F}$, onde $\mathbf{D}_i$ é um vetor coluna com os samples da *i*-ésima janela observada, $\mathbf{H}_i$ uma matriz que descreve o instrumento durante a observação dessa *i*-ésima janela

¹⁶O valor definido empiricamente por D. Harrison para a distância máxima utilizada ao considerar um pixel da imagem reconstruída como “central” é 50 mas.

e $\mathbf{F}$ um vetor coluna que descreve o objeto original—equivalente a escrever $\mathbf{S}$ na forma de um vetor coluna. Note que a matriz $\mathbf{H}$ pode ser tão complexa quanto se queira: a PSF, amostragem, deformações ópticas, podem ser consideradas.

Desse modo, conhecendo-se a matriz inversa $\mathbf{H}_i^{-1}$ poderia-se obter $\hat{\mathbf{F}}$. Entretanto, dois problemas impossibilitam a resolução por esse caminho: $\mathbf{H}_i$ não é quadrada, logo não admite inversa, e $\mathbf{H}_i$ tão pouco é convolutiva, impossibilitando uma resolução no espaço de Fourier. Além disso, o problema é mal posto, visto que o número de elementos de $\mathbf{D}_i$ é muito menor do que o número de elementos de $\mathbf{F}$ (Dollet, 2004). Busca-se então uma solução $\hat{\mathbf{F}}$ que minimize a soma das diferenças quadráticas entre todas as N observações:

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

Como o algoritmo de Tikhonov foi escrito para solucionar problemas do tipo $y = Ax$, é necessário re-escrever o problema nesta forma. Utilizando a matriz transposta de $\mathbf{H}_i$, a minimização acima resulta em:

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

Então, o processo iterativo derivado em Dollet (2004) reconstrói a imagem $\hat{\mathbf{F}}$ no passo $(n + 1)$ por:

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

onde L é o operador Laplaciano discreto em duas dimensões, α é o parâmetro de convergência e λ é o parâmetro de regularização. É importante notar que esses dois parâmetros devem ser determinados de forma empírica.

A resolução do processo iterativo é naturalmente mais lenta do que todos os quatro algoritmos descritos anteriormente, visto que de certo modo diversas reconstruções intermediárias são realizadas. Mas o maior problema desse algoritmo é a determinação das matrizes $\mathbf{H}$: dependendo do tipo de janela, processar 100 ângulos de trânsito diferentes pode levar mais de duas horas e meia para apenas um objeto (Dollet, 2004). Devido a esse fator, sua utilização em larga escala durante a missão foi descartada, ainda que Mary et al (2006) tenha mostrado que é possível recuperar objetos tão fracos quanto $G \sim 23$. Atualmente não existe nenhuma implementação desse algoritmo diretamente compatível com os dados produzidos pelas versões atuais dos sistemas de simulação utilizados pelo Gaia.

Além disso, um problema das imagens geradas por esse algoritmo é a introdução de artefatos, como pode-se perceber na Figura 2.13, e ainda mais grave é o fato de que esse algoritmo não possui garantias de que o fluxo dos objetos seja preservado (Peyrega, 2007). Para objetos pontuais, esses problemas foram minimizados ou resolvidos com a introdução da família de algoritmos Clean, que veremos a seguir.

2.3.7 Clean & Cleanest

Originalmente desenvolvido para síntese de imagens de rádio telescópios, o algoritmo Clean (Högbom, 1974) foi adaptado para uso no Gaia por Charles Pereyga sob orientação de Albert Bijaoui (Peyrega, 2007).

No contexto do Gaia, esse algoritmo pode ser utilizado para reconstruir as imagens de objetos pontuais. Seu princípio básico consiste em considerar os objetos do campo como δ s de Dirac, e reconstruir estes objetos um a um e subtraindo das observações suas contribuições a cada iteração. O algoritmo Cleanest consiste apenas em uma etapa suplementar durante a execução do algoritmo Clean, garantindo que a contribuição dos pixels mais brilhantes de $\mathbf{F}$ seja levada em conta primeiro.

Como na reconstrução por regularização de Tikhonov, seu ponto inicial é a construção das matrizes $\mathbf{D}_i$ e $\mathbf{H}_i$, contendo os *samples* e a modelização do instrumento, respectivamente, para a i -ésima observação. Então, duas novas matrizes, $\mathbf{D}_{global}$ e $\mathbf{H}_{global}$, são criadas a partir da composição de todas as N matrizes $\mathbf{D}_i$ e $\mathbf{H}_i$ originais:

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

onde *size* indica o tamanho dos vetores $\mathbf{D}$ e $\hat{\mathbf{F}}$.

Então, considerando que cada pixel de $\hat{\mathbf{F}}$ é constituído de funções δ de Dirac de amplitude α , o problema se reduz a encontrarmos os valores α_i tais que:

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

onde i é um índice que corre por todos os pixels de $\mathbf{F}$. Esse problema é então resolvido por:

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

Mas o Clean possui uma etapa adicional, que justamente empresta o nome ao algoritmo: após a determinação de α_i , a variância desse valor também é obtida:

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

de modo que para cada α_i , o nível de sinal/ruído é estimado $(\alpha_i/\sigma(\alpha_i))$.

Finalmente, a cada iteração n , α_i e $\sigma(\alpha_i)$ são calculados para todos os pixels de $\mathbf{F}$ mas apenas o pixel que possuir o maior nível de S/R ($\alpha_i^{\max}$) é escrito em $\hat{\mathbf{F}}$. Sua contribuição é então subtraída das observações (matriz $\mathbf{D}_{global}$) por um fator γ (chamado de fator de limpeza, ou *clean factor*), que possui valores entre 0 e 1. Então, a nova matriz $\mathbf{D}_{global}$ utilizada na iteração $(n + 1)$ é construída com os elementos j :

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

$H_{global,ji}$ acima representa apenas um vetor coluna, pois i é fixo por $\alpha_i^{\max}$.

Naturalmente, esses algoritmos possuem o mesmo problema para a criação das matrizes $\mathbf{H}$ que impossibilita o uso em larga escala da reconstrução por regularização de Tikhonov. Entretanto, [Peyrega \(2007\)](#) conclui que em comparação com esse último, a criação de artefatos durante a reconstrução é pequena, que o sinal desses artefatos também é geralmente menor do que o dos objetos, e que o problema de alteração do fluxo é inexistente. Desse modo sua adoção seria preferencial no caso de estudo de objetos pontuais.

Já no caso de objetos extensos, obviamente Clean/Cleanest não seriam adequados do modo como foram desenhados. Entretanto como notado em [Peyrega \(2007\)](#), uma alteração não muito complexa possibilitaria o tratamento desse tipo de fonte astronômica: a adoção de um catálogo de modelos de objetos.¹⁷ Desse modo, ao invés de reconstruir $\mathbf{F}$ por um $\hat{\mathbf{F}}$ criado a partir de funções δ , seria adotado um “alfabeto” de funções diferentes, descrevendo por exemplo formas possíveis para Galáxias. No entanto, nenhum algoritmo desse tipo foi desenvolvido até o momento.

2.3.8 Exemplos de reconstruções

Para exemplificar a qualidade das imagens obtidas a partir da utilização dos algoritmos descritos na seção anterior, realizamos simulações utilizando um simulador oficial da colaboração Gaia, o GIBIS 6 ([Babusiaux et al, 2009](#)). Simulamos uma região padrão (com 85 trânsitos), que incluiu cinco objetos: um principal no centro da imagem com magnitudes $G_p = 15$ ou $G_p = 19$ e quatro objetos secundários, com magnitudes $G_s = G_p + 1.1, 1.5, 1.8$ ou 2.1 .

Os trânsitos foram utilizados para reconstruir imagens com quatro algoritmos implementados em Java (QuickStack, Drizzle, ShuffleStack e BinOutliers), visto que provavelmente um deles será adotado durante o decorrer da missão.

Nas Figuras 2.11 e 2.12 podemos observar as diferenças entre as imagens obtidas com o uso destes quatro algoritmos. A reconstrução QuickStack, uma vez que é apenas um tipo de *Backprojection*, gera uma imagem bastante “borrada” (semelhante ao que vimos ocorrer com o *toy model*), mas os objetos não são distorcidos, como ocorre com as reconstruções ShuffleStack.

Entretanto, nas reconstruções realizadas com o método ShuffleStack os objetos secundários são revelados de forma muito mais clara do que nas imagens obtidas com o QuickStack, mesmo que este método acabe por introduzir artefatos na imagem

¹⁷A definição desse catálogo, no entanto, talvez seja uma tarefa mais complicada.

final, e que exista a possibilidade de se reconstruir pixels com valores negativos (o que não possui sentido físico).

A reconstrução pelo método de Drizzle gera uma imagem com artefatos tão ou mais fortes do que os próprios objetos, o que é um ponto extremamente negativo à adoção desse método, além de possuir a grande desvantagem de que a reconstrução é fortemente dependente de parâmetros cuja escolha é empírica.

Já o método BinOutliers aparentemente é o melhor dentre todos, pois apesar de gerar um objeto central extremamente artificial, é muito bem sucedido tanto na limpeza em relação aos artefatos, que praticamente não existem, como no Drizzle e no ShuffleStack, quanto em sua capacidade em revelar os objetos secundários (até mesmo em magnitude $G = 21.1$, na Figura 2.12). Neste método, o sinal que gerava artefatos nos outros métodos gera um fundo difuso.

A título de comparação, na Figura 2.13 apresentamos duas reconstruções de [Mary et al \(2006\)](#) realizadas com o uso do algoritmos de regularização de Tikhonov. Na Figura (a) o objeto central possui magnitude $G = 18$ enquanto que o objeto secundário, no canto superior direito da imagem, $G = 20$. É interessante notar em (b) que esse algoritmo foi capaz de reconstruir objetos com até mesmo cinco magnitudes de diferença.

Figura 2.11: Exemplos de reconstrução de uma mesma região contendo cinco objetos (quatro secundários). Objeto primário com $G = 14$. (a) QuickStack, (b) Drizzle, (c) ShuffleStack e (d) BinOutliers. Pixels de 30 mas.

Figura 2.12: Como 2.11, mas com objeto primário com $G = 19$.

Segundo [Mary et al \(2006\)](#) os diversos artefatos gerados pelo algoritmo de Tikhonov podem ser limpos a partir da subtração de uma segunda reconstrução com as mesmas janelas, mas com apenas o sinal simulado da estrela central.

Não apresentamos nenhuma reconstrução Clean, pois elas geram apenas sinais δ , onde os pixels do centro do objeto estarão Ligados/Desligados—mostrando se o objeto foi reconstruído ou não.

Figura 2.13: Imagens obtidas pela reconstrução por regularização de Tikhonov. Objeto central com magnitude $G = 18$ e secundários com $G = 20$ (a) e $G = 23$ (b). Pixels de 50 mas. (De [Mary et al, 2006](#))

2.4 Cobertura espacial e angular das reconstruções

A princípio, a qualidade da imagem reconstruída pelos algoritmos propostos poderá ser extremamente dependente da posição da fonte na esfera celeste. Isso, pois em cada posição do céu não apenas o número de trânsitos no final da missão será diferente (como visto no Capítulo 1), mas também a distribuição angular desses trânsitos será variável.

Entretanto, uma vez que conhecemos de antemão o tamanho da janela para qualquer observação e a lei de varredura do céu pelo satélite, podemos estudar essa configuração por meio de simulações que indiquem como cada coordenada da esfera celeste seria recoberta por janelas caso uma fonte fosse detectada naquela posição. Para isso, partimos da definição de duas grandezas que vão nos auxiliar a realizar tal análise:

Definição 3 *A cobertura fractional C_F é definida como a razão entre a área do polígono resultante da união das janelas obtidas em todos os trânsitos em determinada coordenada e a área de um círculo com diâmetro igual à diagonal da maior janela observada.*

A grandeza acima definida descreve a fração da cobertura espacial normalizada pela cobertura máxima (que seria a área do círculo definido por infinitos trânsitos distribuídos em ângulos aleatórios) dos dados em uma determinada coordenada. Contudo, ela nada nos diz explicitamente sobre a resolução dessa informação. É importante possuir ao menos um indicativo quantitativo de qual resolução teórica pode-se atingir em determinada região do céu, o que claramente deve ser limitado superiormente pela máxima resolução dos *samples* transmitidos para aquela coordenada na direção ao longo do trânsito.

Uma vez que os pixels do Gaia são retangulares, a resolução máxima que a imagem reconstruída pode possuir está relacionada com a distribuição dos ângulos de observação, sendo que a máxima resolução só pode ser atingida no caso de existirem janelas obtidas com ângulos de $\pi/2$ ou $3\pi/2$ entre si. Usando a operação módulo inverso (definida no Apêndice A), é possível definir uma grandeza que denominamos de cobertura angular que permite quantificar o quão próximos estaremos desse ângulo ideal:

Definição 4 *Seja $\mathbb{A}$ o conjunto com n elementos formado por todos os n ângulos de trânsito (em rad) de todas as observações em uma determinada coordenada celeste, definimos o conjunto de coberturas angulares $\mathbb{V}$ como o conjunto com $(n^2 + n)/2$ elementos formado por $\{A_j, A_k \in \mathbb{A} : |A_k - A_j|(\text{mod inv } \pi/2)\}$. A cobertura angular entre a k -ésima e a j -ésima observação é definida ser o i -ésimo elemento de $\mathbb{V}$.*

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

Definição 5 C_A é a cobertura angular máxima normalizada:

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

A cobertura angular máxima normalizada definida acima é um indicador da existência da informação angular necessária entre pelo menos dois trânsitos para possibilitar a reconstrução de imagens com pixels quadrados de uma determinada resolução (igual à resolução *along-scan*). Nos casos extremos, se $C_A = 1$ pode-se obter na imagem reconstruída pixels com o mesmo tamanho dos *samples* na direção *along-scan*, e se $C_A = 0$ a melhor resolução que se pode obter será a dos *samples* na direção *across-scan*. Já o cálculo do tamanho do pixel que pode ser reconstruído para casos intermediários de C_A é algo fortemente dependente do algoritmo de reconstrução adotado.

Para estudar essas duas grandezas, C_F e C_A , e suas respectivas variações no céu, devemos realizar simulações dos trânsitos no céu inteiro. Entretanto, a simulação dos trânsitos do Gaia para uma determinada coordenada celeste nos fornece os ângulos de cada observação, mas não as janelas observadas¹⁸, o que embora seja suficiente para calcular C_A , não nos permite calcular C_F .

Contudo, como conhecemos os tamanhos de todos os tipos de janelas, podemos calcular as coordenadas de seus vértices em coordenadas galácticas, definindo assim os polígonos necessários para o cálculo de C_F . Além disso, como a cobertura espacial é completamente determinada pela maior janela disponível, podemos considerar apenas as janelas dos *SkyMappers*, pois elas são sempre maiores do que as janelas dos *AstroFields* (seção 2.1).

Então, como as dimensões l_1 e l_2 (*across-scan* e *along-scan* respectivamente) das janelas dos *SkyMappers* são conhecidas, sabendo o ângulo de trânsito γ podemos escrever os quatro vértices (x_i, y_i) do retângulo resultante como:

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

onde,

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

Logo, podemos obter C_A e C_F em uma determinada coordenada celeste a partir de simulações utilizando o seguinte método:

1. Executar simulações dos trânsitos do satélite, obtendo todos os ângulos de trânsito;
2. Calcular C_A ;

¹⁸Como visto na introdução deste Capítulo, o tamanho de uma janela é uma informação dependente apenas da magnitude da fonte detectada naquela coordenada e da CCD que realiza a observação, e não da atitude do satélite.

3. Para cada trânsito calcular os vértices do retângulo definido pela janela observada segundo a Equação 2.8;
4. Realizar uma união booleana entre todos os retângulos;
5. Calcular C_F ;

Desenvolvemos um código a partir da implementação do método descrito acima, que foi disponibilizado para uso comum entre os membros do Gaia que trabalham com a reconstrução de imagens (CU5-DU18).

Nas próximas três subseções veremos a descrição de como são realizadas as simulações dos trânsitos (passo 1) e a operação de união booleana entre polígonos (passo 3), e finalmente a análise da distribuição no céu inteiro de C_A e C_F no final da missão.

2.4.1 Simulação dos trânsitos

Como visto no Capítulo 1, os dois campos de visão do Gaia varrem o céu inteiro devido aos movimentos de rotação e precessão do satélite e de sua órbita ao redor do Sol. Desse modo, para que seja possível determinar quantos trânsitos serão realizados em uma determinada coordenada celeste, é necessário realizar uma simulação tanto dos movimentos do satélite em torno de si mesmo quanto de seus movimentos em torno do Sol.

A lei de varredura (ou *Scanning Law*), apresentada em Lindegren (1998, 2001), é uma descrição das coordenadas celestes do eixo de rotação do satélite em função do tempo, além de uma descrição do próprio movimento de rotação. Em coordenadas eclípticas a posição na esfera celeste que indica as coordenadas (λ_z, β_z) para onde o eixo de rotação do satélite $\mathbf{z}_s$ (definido na Figura 2.14) está apontando no momento t é descrita pelas equações:

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

onde $\lambda_\odot$ é a longitude eclíptica do Sol, ξ é o raio do pequeno círculo definido pelo movimento de precessão do eixo $\mathbf{z}_s$ em torno de $\mathbf{x}_0$ e $\nu(t)$ é a fase de revolução.

Essa fase é especificada de modo que a velocidade do eixo $\mathbf{z}_s$ no céu ($|\mathbf{dz}/d\lambda_\odot| = S$) seja constante, gerando o que se denomina “lei de varredura uniforme”¹⁹. Além disso, deve-se ter uma taxa de rotação constante ao redor do eixo $\mathbf{z}_s$, descrita pela fase da rotação $\Omega(t)$. Essas duas fases são descritas por:

$$\begin{cases} \sin \xi \frac{d\nu}{d\lambda_\odot} = \sin \nu \cos \xi + \sqrt{S^2 - \cos^2 \nu} \\ \frac{d\Omega}{dt} = \omega_z - \frac{d\lambda_\odot}{dt} \sin \nu \sin \xi - \frac{d\nu}{dt} \cos \xi \end{cases} \quad (2.11)$$

¹⁹Como comentado em Lindegren (1998), essa lei não é estritamente uniforme, uma vez que existe uma variação de $\sim 1.67\%$ em $d\lambda_\odot/dt$ devido à excentricidade não nula da órbita da Terra.

onde ω_s taxa de rotação do satélite ao redor do eixo $\mathbf{z}_s$ e outras grandezas como indicadas na Figura 2.14.

Figura 2.14: Definição dos sistemas de coordenadas e ângulos utilizados para a lei de varredura. Em (a), $[x, y, z]$ indica um sistema inercial centrado no baricentro do Sistema Solar, $[x_0, y_0, z_0]$ um sistema rotante centrado no Sol com o eixo x_0 apontando para a Terra e $\lambda_\odot$ é a longitude eclíptica aparente do Sol (incluindo aberração). Em (b), ξ é o ângulo do eixo de rotação do Gaia, Ω é o ângulo da fase de rotação, ν é o ângulo da fase de revolução e os sistemas $[x_1, y_1, z_1]$ e $[x_s, y_s, z_s]$ são centrados no satélite. (Adaptadas de Luri, 2001; Astrium, 2006)

Desse modo, a partir da obtenção de $\lambda_\odot$ via efemérides, da escolha de S , ξ e ω_z (para otimização da cobertura do céu) e das condições iniciais $\nu(t=0)$, $\Omega(t=0)$, a lei de varredura fica completamente determinada, bastando que se realize a integração das equações 2.11 para determinação da posição do eixo de rotação no céu (por meio das equações 2.10).

Como o ângulo básico entre os dois espelhos do Gaia é conhecido (106.5°) e encontra-se no plano dos eixos $\mathbf{x}_s$ e $\mathbf{y}_s$, podemos calcular a missão inteira e determinar quantas vezes, e com qual atitude, o satélite passará por qualquer ponto da esfera celeste. Para obter as informações de um ponto específico em coordenadas eclípticas, basta inverter as equações 2.10 e determinar os tempos t nos quais o satélite observará aquela posição.

Utilizamos a lei de varredura descrita em Luri (2001) e implementada na biblioteca GaiaSimu 7 (Babusiaux et al, 2010) para calcular os trânsitos no céu inteiro. Uma vez que conhecemos a quantidade desses trânsitos e as coordenadas eclípticas dos campos de visão em cada um destes trânsitos, basta aplicarmos a equação 2.8 para determinação das coordenadas das janelas observadas. Veremos agora como foi realizada a união das janelas observadas para o cálculo de C_F .

2.4.2 União de polígonos

Apesar de aparentemente serem tarefas triviais, a realização de operações Booleanas entre polígonos genéricos (não necessariamente convexos²⁰ ou monótonos²¹) são processos com alto custo computacional, realizadas por bibliotecas altamente especializadas de geometria computacional.

A princípio pode-se pensar que para o cálculo de C_F não necessitamos de polígonos genéricos, visto que as janelas transmitidas pelo GAIA são sempre retângulos. Entretanto, após a união entre duas janelas, já pode-se ter um polígono não monótono, por exemplo.

Os algoritmos mais primitivos para realizar essa operação são baseados na rasterização dos polígonos em *bitmap* binários, e a posterior aplicação de operações Booleanas entre os *bitmaps*. Apesar de conceitualmente simples, esta técnica possui diversos problemas, em especial ligados à resolução da rasterização e à quantidade de memória necessária para obtenção de boa precisão.

Já as técnicas mais modernas se apoiam em construções de geometria computacional, sendo principalmente baseadas em algoritmos tipo *plane* ou *line-sweep*. Nestes algoritmos, uma linha (ou plano) imaginária é movida no espaço encontrando regiões “à esquerda” ou “à direita” de cada vertice e bifurcações (para uma descrição detalhada ver por exemplo o Capítulo 2 em De Berg et al, 2008 ou O’Rourke, 1998). Apenas recentemente um algoritmo capaz de lidar de forma coerente com polígonos completamente genéricos foi publicado Martinez et al (2009).

Neste trabalho utilizamos a implementação da união entre polígonos presente na biblioteca *JTS – Java Topology Suite 1.8.1* (VividSolutions, 2006). A técnica mais direta seria uma união iterativa na qual cada janela de cada observação é unida com a próxima (com um algoritmo tipo *plane-sweep*) e assim por diante até a última observação disponível. Entretanto esse método pode ser ineficiente para regiões do céu com muitas dezenas ou centenas de polígonos. Desse modo, como o estudo deveria ser capaz de cobrir o céu inteiro uma solução mais otimizada foi escolhida, e adotamos o sistema de união por *buffer* descrito em Davis (2007).

Na terminologia da *JTS*, e de aplicações *GIS*²² em geral, a operação de *buffer* representa o cálculo da região que contém todos os pontos dentro de uma dada distância (especificada pelo tamanho do *buffer*) da figura geométrica original (VividSolutions, 2003). Essa operação é uma soma de Minkowski²³ entre o polígono e um círculo de raio igual a distância requerida.

Pela definição acima percebe-se que é possível realizar uma operação de adição Booleana entre quaisquer polígonos juntando todos os polígonos em um conjunto

²⁰Um polígono convexo é aquele que obedece: qualquer ângulo interno é menor do que 180° e todo segmento de reta entre dois vertices é interno ou é uma borda do polígono.

²¹Um polígono é monótono com respeito à reta L , se qualquer ortogonal a L intersectá-lo no máximo duas vezes.

²²Geographic Information System, ou Sistema de Informação Geográfica.

²³Para dois subconjuntos A e B de um espaço vetorial, a soma de Minkowski de A com B é definida como $A + B = \{\mathbf{x} + \mathbf{y} | \mathbf{x} \in A, \mathbf{y} \in B\}$, o que é equivalente a uma convolução entre as duas figuras (Skiena, 2008).

e aplicando uma operação de *buffer* com tamanho nulo sob esse conjunto. Dessa forma, a operação final resultante é equivalente a uma união entre os polígonos, sem alteração em seus contornos, e de forma muito eficiente (Davis, 2007).

2.4.3 Cálculo na nuvem

Para realização dos estudos de cobertura no céu inteiro, é necessário implementar uma leve modificação no código para possibilitar sua execução de forma paralela, visto que o tempo de execução seria demasiado alto para estudos em larga escala utilizando apenas um processador. Felizmente, o problema tratado aqui é altamente paralelo, e pode ser dividido em diferentes “nós” de cálculo que trabalham de forma completamente independente, com cada nó sendo utilizado para calcular uma posição distinta da esfera celeste.

Além disso, pacotes com diversas posições podem ser enviados para os nós, reduzindo requisitos de comunicação e aliviando quaisquer gargalos de rede existentes. Desse modo, podemos adotar um modelo de computação em *Grid* (Foster & Kesselman, 2004), que é uma forma de computação distribuída na qual um processo é espalhado entre diversos nós de cálculo que não necessariamente são conectados por uma infra-estrutura de comunicação rápida e/ou estável, sendo um paradigma que visa utilizar a Internet não apenas para armazenar informação, mas também para processamento. Além disso, falamos aqui em “cálculo na nuvem”, pois ele foi executado sob máquinas virtuais.

A adoção desse modelo possibilita realizar uma mesma execução do código em paralelo usando nós geograficamente espalhados, se necessário aproveitando a infra-estrutura computacional de diversos laboratórios²⁴. Como infra-estrutura de *software* para nosso *Grid*, a biblioteca *GridGain 2.1.1* (GridGain, 2010) foi utilizada. Ela realiza o paralelismo de tarefas através de uma implementação do algoritmo *Map-Reduce* (Dean & Ghemawat, 2004) e já engloba todas as ferramentas necessárias para execução em nós não homogêneos (por *load-balancing*), transferência de código entre os nós, redundância de tarefas, tolerância a falhas (caso algum nó deixe de estar disponível), descoberta automática de novos nós e sua adição na topologia do *Grid* e na distribuição de tarefas, o que se mostrou bastante útil, uma vez que os nós não eram dedicados apenas à execução das simulações para cálculo de C_A e C_F .

2.4.4 Resultados

A implementação de um código com as características descritas anteriormente permite realizar as análises de C_A e C_F para coordenadas genéricas na esfera celeste. A execução desse código pode fornecer visualização gráfica dos resultados para uma coordenada única ou tabelas com os valores calculados dentro de uma rede de coordenadas em uma certa região do céu (ou no céu inteiro).

O cálculo da fração da área coberta pelas observações em torno de qualquer posição do céu é uma informação importante para que se saiba qual tipo de in-

²⁴No caso desse projeto, não foi necessário mais do que aquela disponível no IAG-USP.

SM windows around $(l,b) = (0.0,0.0)$

Figura 2.15: Mapa de cobertura para as coordenadas galácticas $(l, b) = (0.0^\circ, 0.0^\circ)$

formação pode-se obter a partir de uma imagem reconstruída para uma fonte em determinada coordenada. Além disso, esses dois valores também são extremamente importantes para seleção das coordenadas usadas durante os testes e validações dos códigos desenvolvidos no decorrer deste trabalho, e principalmente, como será visto no próximo Capítulo, na seleção de regiões de treinamento para algoritmos de aprendizagem computacional.

Para realizar essa seleção é necessário analisar a configuração geométrica dos trânsitos, uma vez que a distribuição angular das observações possui forte influência nos resultados dos algoritmos de reconstrução de imagens. A Figura 2.15 mostra um exemplo das informações obtidas para uma rodada em uma coordenada específica, no caso as coordenadas galácticas $(l, b) = (0.0^\circ, 0.0^\circ)$.

Com a finalidade de verificar a estrutura em larga escala dos parâmetros C_A e C_F no céu, realizamos uma simulação para o céu inteiro em baixa resolução de $150'$ e $300'$, respectivamente, em cada um dos eixos do sistema galáctico. Durante a análise dos resultados, percebemos a existência de uma grande simetria em relação ao sistema eclíptico, naturalmente devida à lei de varredura do satélite (que como visto no Capítulo 1, é altamente simétrica ao plano da eclíptica)—nas Figuras 2.16 (a) e (b) apresentamos a distribuição no céu de C_A e C_F obtidas no céu.

As simulações de C_A , na Figura 2.16 (a), indicam que para qualquer ponto que se observe no céu a cobertura é bastante alta, sendo que o mínimo valor obtido para C_A foi de 0.986, além de ser bastante homogênea (visto que a variação para quase todas as regiões do céu encontra-se entre $C_A = 0.996$ e $C_A = 1.000$). Isso é um forte indicador de que será possível reconstruir imagens com pixels quadrados com a melhor resolução angular possível para qualquer fonte que seja observada—sendo a resolução final limitada apenas pelo algoritmo de reconstrução escolhido.

(a)

(b)

Figura 2.16: Mapas de coberturas para o céu inteiro em coordenadas eclípticas. (a) Cobertura Angular Normalizada—resolução da rede regular de $150'$. (b) Cobertura Fractional—resolução da rede regular de $300'$.

(a)

(b)

Figura 2.17: (a) Mapa de Cobertura Fracional (C_F) em alta resolução ($25'$) para uma faixa do céu (a falta de dados próximo de $\beta \sim 90^\circ$ é devido a simulação ter sido realizada em um grid de coordenadas galácticas). (b) Correlação entre posição no céu, cobertura fracional (C_F) e número de trânsitos.

A Figura 2.16 (b), mostra os resultados para as simulações de C_F . Essas simulações indicam que a cobertura fracional é altamente simétrica no céu em relação ao plano da eclíptica, e que se define um padrão homogêneo em larga escala se repetindo no céu em faixas de dois comportamentos distintos: uma região de cobertura crescente entre $|\beta| = 0^\circ$ e $|\beta| \sim 40^\circ$ e uma região com C_F praticamente estável acima de 98% para $|\beta| > 40^\circ$. Além disso, essa Figura indica que embora a distribuição de C_F seja aparentemente homogênea em larga escala, é possível notar a existência de variações regulares em menor escala.

Tais variações em menor escala são melhor reveladas em uma simulação com maior resolução na rede (de $25'$), cujo resultado pode ser observado na Figura 2.17 (a). Podemos notar a existência de um padrão a cada $\sim 13^\circ$ de longitude, que é bastante similar ao observado na estrutura do número de trânsitos para cada região do céu (Capítulo 1). Entretanto, na Figura 2.17 (b), observa-se que não existe uma correlação direta entre o número de trânsitos e o índice C_F , ainda que as regiões com um número de trânsitos muito grande possuam uma cobertura igualmente grande.

Mesmo em regiões onde o número de trânsitos não é maior do que o número médio para as observações do Gaia, o valor obtido para C_F encontra-se acima de 98% (o que é um forte indicador de que apenas certas posições das bordas do círculo representado na Figura 2.15 não foram recobertas). Cabe notar, no entanto, que as regiões com o número mínimo de trânsitos (pouco menos de 60), possuem também C_F baixo, indicando ausência de dados para reconstrução de certas regiões da imagem—uma destas regiões, que exemplifica um valor de C_F baixo, foi apresentada como exemplo na Figura 2.15.

Entretanto, como se pode notar na Figura 2.15, até mesmo no caso de cobertura $\sim 87\%$ a região na qual se perde informação não é muito grande, de modo que os valores para C_F obtidos para o céu inteiro (sempre maiores do que 87%), permitem concluir que existem dados suficientes para reconstruções de imagens que possibilitem uma análise de praticamente qualquer tipo de emissão que venha a ocorrer nos arredores de quaisquer objetos observados pelo Gaia.

2.5 Conclusões

Nesta seção vimos através de um *toy model* os princípios de reconstrução de imagem tal como isso é entendido no contexto do Gaia. Apresentamos a transformada de Radon, que pode ser utilizada como descrição matemática para as janelas originadas das observações realizadas pelo satélite, além de introduzirmos as idéias gerais de como se inverte essa transformação, e de como essa inversão é sensível à quantidade dos dados disponíveis e ao ruído na determinação do sinal.

Discorremos sobre as dificuldades de realizar a reconstrução de imagens no contexto da redução de dados do Gaia, e apresentamos os algoritmos propostos, até o momento na literatura, para uso na preparação desse catálogo. Vimos que os algoritmos baseados na regularização de Tikhonov e o algoritmo Clean/Cleanest, apesar de possibilitarem uma reconstrução de imagens com maior resolução, possuem

um custo computacional proibitivamente alto para uso em larga escala, e que até o momento sua utilização durante a missão está descartada. Vimos qualitativamente que o melhor candidato é o BinOutliers.

Apresentamos um estudo que realizamos no âmbito deste doutorado sobre a cobertura espacial dos trânsitos, chamada de C_F ou cobertura fracional, cujo resultado demonstrou que para qualquer posição do céu observada, no final da missão haverá uma cobertura mínima de $\sim 87\%$, atingindo até $\sim 99\%$ em determinadas regiões, e sendo maior do que 98% em grande parte do céu ($|\beta| > 45^\circ$). Essa alta cobertura permite que as informações de qualquer emissão que ocorra ao redor de qualquer fonte observada seja analisada, mesmo no caso de cobertura mínima, com $C_F \sim 87\%$.

Finalmente, nosso estudo também demonstrou que a distribuição relativa entre os ângulos de trânsito, aqui chamada de C_A , ou cobertura angular normalizada, para qualquer região do céu possui valor aproximadamente igual a 1, o que significa que ao menos dois trânsitos encontram-se a $\sim 90^\circ$ entre si. Isso significa que para qualquer coordenada da esfera celeste existe a possibilidade teórica de que a reconstrução possa ser realizada com pixels tão pequenos quanto o tamanho do pixel na direção de maior resolução, sendo a resolução final obtida dependente apenas do algoritmo adotado e não da ausência de informação.

Fontes Secundárias

*“Mais ce qui est vrai des maux de ce monde est vrai aussi de la peste.
Cela peut servir à grandir quelques-uns.”* A. Camus¹

Conteúdo

3.1	Introdução	68
3.2	Análise semi-numérica – Catálogos	69
3.2.1	Procedimento de Análise	69
3.2.2	Resultados	72
3.2.3	Comentários	73
3.3	Análise semi-numérica – Simulações	73
3.3.1	Procedimento de análise	74
3.3.2	Resultados	75
3.4	Segregação Educada de Imagem - EIS	79
3.4.1	Princípios	79
3.4.2	Modo com Limiar-Simples	80
3.4.3	Modo com Limiar-Múltiplo	83
3.4.4	Modo Educado	87
3.4.5	Testes e resultados	92
3.5	Conclusões	100

Para um instrumento com a magnitude limite do Gaia, a quantidade de fontes de radiação no céu é grande o suficiente para que a chance de ocorrer projeções entre objetos na mesma linha de visada não deva ser negligenciada. Além disso, diversos objetos da Galáxia se revelam sistemas múltiplos ao serem observados em maiores níveis de resolução angular.

Neste Capítulo apresentaremos estudos quantitativos que desenvolvemos sobre este fenômeno (Krone-Martins et al, 2008a), que perturbará a solução astrométrica de diversas fontes. Discorreremos também sobre a estratégia que desenvolvemos, e que será utilizada durante o processamento de dados do Gaia, para recuperar as fontes secundárias, e para aproveitar seus dados (Krone-Martins et al, 2010a). Essas fontes são interessantes, uma vez que podem se encontrar acima da magnitude limite do satélite, logo ao aproveitar seus dados será possível complementar o catálogo Gaia.

¹“Mas a verdade sobre todos os males do mundo também é verdade sobre a peste. Eles podem ajudar alguns homens a se superar.”, in *La Peste*.

3.1 Introdução

A presença de estruturas extensas em objetos astronômicos pode perturbar a solução astrométrica que será derivada pelo consórcio de processamento de dados do Gaia (Brown, 2007). Essas estruturas, definidas aqui como um sinal maior do que a PSF instrumental, podem ser originadas por razões intrínsecas ao objeto (uma galáxia, por exemplo), ou extrínsecas, como no caso de confusão entre objetos que se encontram na mesma linha de visada. É com esta segunda modalidade, denominada projeções ópticas ou fontes secundárias, que nos preocuparemos neste Capítulo.

Nas seções 3.2 e 3.3 apresentaremos análises semi-numéricas com resultados para o número de projeções ópticas às quais o Gaia estará sujeito. Em primeira instância, tais análises servem para estimar a quantidade de processamento necessária para reconstruir e processar os casos problemáticos. Entretanto, durante estes estudos não estaremos apenas considerando o número de fontes secundárias que poderão interferir diretamente nas medidas astrométricas (ou fontes com $G \leq 20$), mas sim o número total destas projeções, que incluem também fontes que podem ser observadas apenas a partir de imagens reconstruídas ($G > 20$).

Como vimos no Capítulo 2, utilizando dados das janelas transferidas pelo Gaia é possível reconstruir mapas bidimensionais (ou imagens reconstruídas) do objeto e de seus arredores. Tais imagens podem revelar as estruturas perturbadoras, sejam fontes secundárias de quaisquer magnitudes ou estruturas extensas intrínsecas ao objeto, sendo que o processamento dessas imagens reconstruídas no Gaia deve proceder da seguinte forma (Brown, 2007):

1. As estatísticas do ajuste da Point Spread Function / Line Spread Function (PSF/LSF) e a solução astrométrica serão examinadas para decidir se uma imagem deve ser reconstruída; Caso seja necessário, um mapa bidimensional de primeira ordem será reconstruído utilizando provavelmente algum dos algoritmos vistos no Capítulo 2 (*QuickStack*, *Drizzle*, *ShuffleStack*, *BinOutliers*);
2. Este mapa será então analisado, buscando encontrar as fontes astronômicas e suas características, além do fundo de céu; Para cada fonte encontrada, será realizada uma tentativa de determinação do fluxo e os cinco parâmetros astrométricos padrão (α , δ , μ_α , μ_δ e ϖ) assumindo sinais pontuais;
3. Caso a solução falhe, um processamento mais sofisticado deverá ser realizado, no qual cada fonte terá sua natureza física classificada previamente, de modo a adotar métodos e perfis adaptados à natureza do objeto sob análise (como será visto para galáxias nos próximos capítulos desta tese).

Na Seção 3.4 apresentamos uma descrição do método de segregação de imagens que foi desenvolvido por nós, e que será adotado durante a redução de dados do Gaia (dentro do CU5-DU18) para a caracterização da imagem reconstruídas em termos das fontes existentes, solucionando parte do passo 2 acima.

3.2 Análise semi-numérica – Catálogos

3.2.1 Procedimento de Análise

Para realizar uma estimativa do número de projeções ópticas a partir de catálogos previamente conhecidos duas opções são possíveis. A mais direta e conceitualmente simples é buscar diretamente o número de casos nos quais as restrições de distância angular e diferenças de magnitudes entre duas fontes se aplicam. Infelizmente esta opção não é apenas altamente custosa (do ponto de vista computacional), mas também é geralmente difícil encontrar catálogos completos até a magnitude desejada (no caso deste estudo, $V \approx 22$) e com a resolução angular requerida ($\sim 0.1''$).

Outra possibilidade, e a que será desenvolvida aqui, é aplicar um método baseado na estimativa da densidade de objetos a partir de catálogos para certos intervalos de magnitude e em determinadas direções do céu. Para isso, no momento assumimos uma distribuição homogênea de fontes no céu e extrapolamos o número de casos de projeções ópticas derivado em determinadas direções para toda a esfera celeste.

3.2.1.1 Método

O satélite Gaia vai observar $N_G(m)$ fontes com magnitude m na esfera celeste. Uma fonte secundária pode vir a perturbar a solução astrométrica, bem como os ajustes de PSF/LSF, caso ela esteja a uma distância $d \leq 2.53''$ da fonte principal (Brown, 2007). Além disso, a fonte secundária deve possuir uma magnitude m_{sec} no intervalo $m < m_{sec} \leq (m + \Delta m)$, onde Δm ainda é um valor a ser definido—nesta seção estudaremos um caso extremo onde $\Delta m = 5$ e um caso mais realista, em que $\Delta m = 3$.

Uma estimativa para o número total de objetos secundários ao redor das fontes primárias do Gaia (para uma dada magnitude), pode ser obtida supondo que estas estrelas estejam homogeneamente distribuídas ao redor de fontes primárias.

É intuitivo considerar que o número de projeções ópticas N_{prob} às quais o Gaia estará exposto será limitado superiormente pelo número de fontes primárias, se as secundárias existirem em número suficiente para compensar as pequenas distâncias em que podem ser encontradas ao redor das primárias. Caso contrário, podemos escrever o número de casos problemáticos para objetos primários de magnitude m como:

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

onde A_{esf} é a área da esfera celeste, $\rho(m)$ é a densidade superficial de estrelas de magnitude m e A_G é a área ao redor de cada objeto primário na qual uma fonte secundária pode ser encontrada. Esta relação assume que a região em que $\rho(m)$ é calculada é representativa de toda a esfera celeste.

Para estimar a densidade superficial de estrelas, nesta seção analisamos o catálogo GSC 2.3.2 (Lasker et al, 2008) usando dados no filtro $V_{GSC2.3.2}$ ² para objetos estelares

²É importante notar que o filtro V do GSC 2.3.2 não é o filtro V de Johnson, mas o filtro V

e não-estelares. Nesta análise não separamos objetos estelares de não estelares, uma vez que ambos devem perturbar o ajuste da PSF/LSF e consequentemente a solução astrométrica do objeto primário.

Os dados do catálogo foram selecionados para apenas algumas regiões do céu, sendo que em cada região o procedimento adotado para a obtenção do número de projeções ópticas foi:

1. Contar o número de estrelas dentro de diversos intervalos de magnitude;
2. Analisar tais contagens e determinar a magnitude limite no catálogo dentro da região como a magnitude na qual a lei exponencial é interrompida;
3. Para extrapolar o catálogo até magnitudes maiores, realizar ajustes lineares para 100 sub-conjuntos de dados obtidos a partir de diferentes combinações do conjunto de dados original no espaço $\log(\text{contagens})$ vs. $V_{GSC2.3.2}$ ³, obtendo funções $\rho(m)$;
4. Usando o valor médio obtido a partir das leis $\rho(m)$ ajustadas, extrapolar as contagens dos dados entre a magnitude limite do catálogo e $V_{GSC2.3.2} = 22$;
5. Então, estimar a cada magnitude o número de casos problemáticos na esfera pela equação 3.1.

É importante notar que para realizar todos os cálculos numéricos neste trabalho, supusemos que o catálogo Gaia será completo até magnitude $V_{GSC2.3.2} = 20$ para os alvos principais, e que seremos capazes de detectar objetos até magnitude $V_{GSC2.3.2} = 22$ nas imagens reconstruídas.

3.2.1.2 Dados

Usando dados do catálogo GSC v. 2.3.2, 10 regiões circulares (com raio $r = 1^\circ$) centradas em posições amostrando regiões do céu com densidades de objetos distintas no filtro $V_{GSC2.3.2}$ foram analisadas.

A densidade tem menor valor em $(l, b) = (0^\circ, 30^\circ)$, onde apenas $32 \text{ obj}/dg^2$ foram encontrados; isto pode ter sido ocasionado por algum problema de confusão, com a baixa densidade se originando na impossibilidade de separar os objetos na compilação do catálogo. A maior densidade foi encontrada nas coordenadas galácticas $(l, b) = (60^\circ, 0^\circ)$, com $27779 \text{ obj}/dg^2$, talvez pela separação dos objetos ter sido mais eficiente do que na região central, $(l, b) = (0^\circ, 0^\circ)$, onde uma grande quantidade de confusão era esperada; este valor é aquele que podemos esperar como um limite inferior para a densidade no bojo.

Como nossa análise requer que os alvos primários tenham uma magnitude limite $V_{GSC2.3.2} = 20$ enquanto as fontes secundárias podem ter magnitude limite

fotográfico. Está é a razão pela qual não transformamos as magnitudes para a banda G do Gaia por meio das fórmulas publicadas em [Jordi \(2007\)](#). Entretanto a magnitude dada neste filtro não deve estar muito distante de G .

³Esta técnica é um tipo de simulação de *bootstrap*.

Figura 3.1: Coordenadas das regiões analisadas. Círculos vermelhos indicam posições aproximadas de amostragem do catálogo (adaptado 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

Tabela 3.1: Magnitudes limite (como definidas na seção 3.2.1.1) e densidades de objetos (em obj/dg^2) das regiões analisadas do catálogo GSC 2.3.2 no filtro $V_{GSC2.3.2}$.

$V_{GSC2.3.2} = 22$, extrapolamos as contagens por bin de magnitude dentro das regiões estudadas usando ajustes lineares no espaço $\log(\text{contagens})-V_{GSC2.3.2}$ para 100 simulações *bootstrap*. As contagens artificiais para magnitudes maiores do que a limite do catálogo, tal qual os dados obtidos diretamente do catálogo, estão representadas na Figura 3.2 para todas as regiões analisadas.

Figura 3.2: Densidades de objetos obtidas a partir de contagens do GSC 2.3.2 (em negro) e das leis de extrapolação (em vermelho).

No caso especial da região $(l, b) = (0^\circ, 30^\circ)$, a maior parte da lei é obtida por meio da extrapolação. É importante notar que esta talvez seja uma estimativa pouco confiável da densidade desta região, devendo ser considerada um limite inferior.

3.2.2 Resultados

Após processar cada região obtivemos o número de projeções ópticas para o céu inteiro para cada intervalo de magnitude da fonte primária. Este número foi estimado a partir dos dados obtidos em cada região analisada, de modo que ele varia bastante entre as regiões dependendo de suas densidades—o que intuitivamente já era esperado. O cálculo foi realizado tanto para um caso em que a fonte secundária mais fraca possui um fluxo de até cinco magnitudes menor do que a primária quanto para o caso em que essa diferença é menor, de três magnitudes.

A Figura 3.3 mostra as estimativas obtidas para algumas magnitudes de fontes primárias, e para as regiões amostradas do catálogo. Estes resultados indicam que o número de projeções ópticas que pode ser esperado durante a missão deve ser da ordem de milhões (mesmo considerando o caso mais restritivo do $\Delta m = 3$), podendo atingir centenas de milhões.

Figura 3.3: Extrapolação para o céu inteiro do número de projeções ópticas ao observarmos uma fonte Gaia de magnitude V para todas as regiões. O gráfico da esquerda representa o comportamento assumindo $\Delta m = 5$ e o da direita $\Delta m = 3$.

Podemos também determinar uma probabilidade de se encontrar uma projeção óptica. Esta pode ser calculada simplesmente tomando a razão entre o número de projeções ópticas e o número de fontes primárias. Os resultados obtidos para este cálculo estão representados na Figura 3.4. Conclusões semelhantes foram obtidas a partir da análise do catálogo USNO-B1.0.

Figura 3.4: Probabilidade de encontrar uma projeção óptica ao observar uma fonte primária de magnitude V para todas as regiões analisadas. Como na última figura, a esquerda estão representados os resultados para $\Delta m = 5$ e a direita $\Delta m = 3$.

3.2.3 Comentários

O número de projeções ópticas varia bastante de acordo com a região do céu em que a análise foi realizada devido às diferentes densidades de objetos encontradas. Contudo, tomando a mediana de todas as regiões analisadas, pode-se ter uma primeira estimativa do número mediano de casos que devem ocorrer durante a construção do catálogo Gaia.

Com este cálculo, estima-se que tal número seja $N \approx 2 \times 10^6$, para a diferença de magnitude mais realista de $\Delta m = 3$. O menor valor obtido foi de $N \approx 7 \times 10^4$, enquanto que no caso de uma Galáxia completamente formada por um bojo, esse valor seria de $N \approx 3 \times 10^{10}$ (ambos considerando $\Delta m = 3$).

Podemos perceber que embora estes números e a Figura 3.3 nos dê um sentimento qualitativo do que se pode esperar para o céu inteiro em termos do número de projeções ópticas, os resultados obtidos aqui não restringem de forma muito precisa o número de projeções ópticas esperado para o céu inteiro—e isso se dá por três fatores principais: falta de realismo na extrapolação dos catálogos, número baixo de regiões amostradas e falta de uma integração interpolando o céu inteiro. Para resolver tais problemas, resolvemos adotar uma análise um pouco mais complexa baseada em simulações, que será descrita na próxima seção.

3.3 Análise semi-numérica – Simulações

Para melhor restringir o número de projeções ópticas que poderão ser encontradas pelo Gaia, realizamos uma estimativa destes casos utilizando dados oriundos de simulações com a GaiaSimu.

3.3.1 Procedimento de análise

Simulamos catálogos de estrelas completos até magnitude $G=22$ em mil regiões da esfera celeste selecionadas de forma completamente aleatória (com distribuição uniforme na esfera). Essas simulações foram realizadas através da utilização do modelo de galáxia de Besançon (Robin et al, 2007), como implementado no modelo de universo do Gaia na biblioteca GaiaSimu v.2.0 (Babusiaux et al, 2007) (foi utilizada a classe `BesanconStarGenerator`).

As regiões simuladas possuem geometria circular com raio variável, mas máximo de 1° —deste modo assumimos uma distribuição espacial homogênea para as fontes dentro destas regiões. O modelo de galáxia adotado inclui o modelo de extinção implementado em Drimmel (2002) e as estrelas duplas implementadas em Arenou (2003). Um exemplo de uma das rodadas de simulação realizada, e de um mapa de céu inteiro com densidades interpoladas a partir das regiões simuladas, pode ser encontrado na Figura 3.5.

Figura 3.5: Em (a) um exemplo das regiões simuladas para uma das rodadas de simulação realizada (rodada A1). Nesta figura o tamanho e as cores dos círculos indicam as áreas de cada região simulada. Em (b), diagrama interpolado a partir das regiões simuladas representando densidades de objetos na esfera celeste.

Para cada uma destas regiões, a densidade de estrelas por intervalo de 0.1 magnitudes foi calculada para o intervalo $[10.0; 22.1]$. Como na seção anterior, usamos tais densidades para calcular o número N_{prob} de projeções ópticas por grau quadrado em cada região e em cada intervalo de magnitude usando a relação:

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

onde as grandezas são como na equação 3.1—contudo aqui não se está assumindo uma distribuição homogênea na esfera celeste.

Então, interpolamos as distribuições no centro de triângulos que repartem a esfera celeste (chamados de HTM⁴). O número de projeções ópticas por unidade de área e por intervalo de magnitude foi ponderado pela distância na superfície da esfera entre o centro do triângulo HTM e as três regiões simuladas mais próximas.

Finalmente, integramos o número de casos dentro de todos os triângulos HTM de toda a esfera de modo a estimar o número de casos que poderão ser encontrados nos dados do Gaia para todos os bins de magnitude. Isso foi realizado de forma separada para a região do bojo ($l < 15^\circ$ ou $l > 345^\circ$ e $-15^\circ < b < 15^\circ$) e disco+halo, uma vez que os valores de projeções para tais regimes são muito diferentes (com o bojo dominando o número de projeções ópticas). Para estimar os erros associados com a seleção aleatória das regiões simuladas, o procedimento foi repetido 10 vezes.

3.3.2 Resultados

Após realizar todas as simulações e calcular a média e desvio padrão dos valores obtidos, concluímos que um número total de $1.6 \times 10^8 \pm 5.1 \times 10^7$ poderá ser observado durante a missão. Entretanto, devemos notar que a estimativa cai significativamente ao considerar apenas as regiões de disco+halo da galáxia, e que esse número é severamente dominado pelo bojo. A Tabela 3.2 mostra os resultados para cada uma das 10 rodadas independentes de simulação+análise.

	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

Tabela 3.2: Número de projeções ópticas para as regiões de bojo e halo+disco. Cada coluna A_n representa uma rodada independente de simulação+análise.

É possível que o número de projeções ópticas apresentado aqui não seja atingido, pois em algumas coordenadas na região do bojo o grande número de objetos pode

⁴Mais detalhes sobre esses triângulos podem ser encontrados na seção 4.2.1.

saturar o sistema que cria as janelas enviadas à Terra, deixando de registrar os dados para os objetos mais fracos (logo, mais sujeitos às projeções).

Uma vez que conhecemos o número de projeções ópticas para uma dada magnitude de fonte primária, assim como o número de fontes, podemos derivar as curvas de probabilidade de projeções ópticas. Levando em consideração todas as 10 rodadas independentes e calculando a curva média e um envelope de 1σ , obtemos o gráfico apresentado a direita da Figura 3.6.

Figura 3.6: A esquerda, número de projeções ópticas. A direita, probabilidade de ocorrer projeção óptica dentro de um raio de $2.3''$ ao redor de uma fonte primária de magnitude G —níveis de probabilidade de 0.1%, 1% and 10% são indicados.

Resolução das discrepâncias com Brown, 2008

Em outra estimativa de projeções ópticas, baseada em uma distribuição homogênea de fontes no céu inteiro, Brown (2008) obteve probabilidades muito mais altas do que as derivadas a partir das simulações que apresentamos na seção anterior (atingindo 100% em alguns cenários). Ambos trabalhos foram examinados, e concluímos que as diferenças se deram principalmente devido às pendentes das funções de contagem de estrelas, e na quebra do crescimento exponencial a partir de $G \sim 20$ (não considerada naquele trabalho).

Os coeficientes angulares das funções de contagem de estrelas que obtivemos em nossas simulações para relações $\log(N)$ vs G são menores em média do que os obtidos a partir do GUMS⁵, como pode ser visto nos gráficos esquerdo e direito da Figura 3.7. A partir das simulações apresentadas na seção anterior obtivemos um

⁵O GUMS, ou *Gaia Universe Model Snapshot* é uma simulação de céu completo de todas as fontes presentes no céu até um certo limite de magnitude. Para produzir esse conjunto de dados foi utilizado o modelo de universo do Gaia como disponível em meados de 2006.

valor mediano de 0.16, enquanto o coeficiente angular utilizado em Brown (2008) foi de 0.23. O valor menor obtido em nossas simulações pode ser devido à incompletude de nossa estratégia de amostragem aleatória das regiões simuladas.

Figura 3.7: A esquerda um histograma representando os coeficientes da relação $\log(N) vs G$ derivada a partir do GUMS. A direita, os coeficientes para a mesma relação, mas derivada a partir de 1000 regiões circulares selecionadas aleatoriamente no céu e simuladas com a biblioteca GaiaSimu v.2.0.

Outra diferença entre os dois trabalhos reside no fato que o número de fontes diminui após magnitude $G \sim 20$, e a função de contagem de estrelas começa a desviar significativamente de uma reta simples, como foi adotada em Brown (2008). Este comportamento pode ser verificado no gráfico da esquerda da Figura 3.8, que mostra o número de estrelas por intervalo de 0.1 magnitudes e um ajuste robusto de reta aos dados (usando o algoritmo de Tukey, 1977).

Esta alteração no comportamento acima de $G \sim 20$ leva a uma super-estimação no número de projeções ópticas se o crescimento em $\log(N) vs G$ for considerado linear. Entretanto, pode-se obter o fator de superestimação ao realizar um segundo ajuste de reta usando dados com $G > 20$ apenas, e calculando a razão entre as integrais de ambos os ajustes no intervalo $[20;23]$ ⁶—ao ignorar essa mudança de comportamento a estimativa é em média 1.4 vezes o valor real (ver Figura 3.8). Esse problema é particularmente amplificado pelo fato de que em Brown (2008), uma função de contagem de estrelas única é extrapolada para o céu inteiro.

Outro problema é o número de estrelas adotado até $G=20$ —necessário para realizar a estimativa utilizando a metodologia apresentada em Brown (2008). Naquele trabalho, originalmente foi adotado um valor da ordem de 10^{4-5} , que é maior do que o que se encontra na maioria das direções do céu ($\sim 10^{3-4}$).

Após levar os fatores de correção em conta, em especial a alteração no comportamento da função de contagem de estrelas e escolhendo as regiões de cálculo de

⁶Aqui usamos $G=23$ como limite superior, vez que esta foi a magnitude limite usada para fontes secundárias por Brown (2008).

Figura 3.8: No gráfico esquerdo, contagens de estrelas em bins de 0.1 magnitudes para uma das regiões simuladas na seção anterior, bem como um ajuste robusto de reta. No gráfico direito o histograma da taxa de superestimativa realizada na função de contagem de estrelas se um ajuste de reta simples é utilizado - existem 1000 pontos no histograma, referentes a cada região individual simulada.

forma a amostrar níveis diferentes de latitudes galácticas, as probabilidades entre ambos os trabalhos ficam em acordo, como pode ser visto ao comparar a Figura 3.6 com a Figura 3.9, que representa os valores de [Brown \(2008\)](#) para probabilidade de projeções ópticas antes e após correções.

Figura 3.9: Figuras de [Brown \(2008\)](#) com estimativas para a probabilidade de projeção óptica. Linhas horizontais indicam níveis de 0.1%, 1% and 10%. A figura da esquerda representa estimativas originais, e a da direita estimativas corrigidas.

3.4 Segregação Educada de Imagem - EIS

Em qualquer imagem astronômica, seja ela uma imagem normal ou reconstruída, para que se possa obter qualquer informação quantitativa sobre as fontes presentes em tais dados o mais importante passo é realização de uma tarefa denominada segregação de imagem.⁷ Em uma primeira aproximação é simplesmente a determinação de quais pixels pertencem a qual fonte na imagem—supondo que os sinais de todas as fontes são completamente separáveis.

No caso das fontes não serem separáveis, a segregação se torna um pouco mais complicada, pois idealmente ela deve indicar que em certos pixels da imagem existe sinal proveniente de uma ou mais fontes—notamos que não é necessário indicar qual o percentual de sinal proveniente de cada fonte, pois isso exigiria uma caracterização detalhada do perfil de brilho de cada fonte individual presente na imagem.

Como vimos no Capítulo 2, o caso das imagens reconstruídas do Gaia é um caso particularmente complicado para a segregação de objetos. Estas imagens reconstruídas contêm diversos artefatos que são gerados pelos próprios algoritmos utilizados para sua reconstrução, ou por efeitos causados por direções preferenciais dos ângulos de trânsito do satélite em determinadas coordenadas celestes. A escolha e/ou criação de um método capaz de resolver esse problema de forma ótima para o Gaia é dificultada pelo fato de que até o momento ainda não foi definido qual algoritmo será utilizado durante a análise de dados.

Uma dificuldade adicional é o fato de que o método utilizado para a segregação de imagem no Gaia deva ser eficiente, uma vez que pelas as estimativas apresentadas nas seções 3.2 e 3.3, podemos esperar um limite superior de uma centena de milhões de casos em que uma imagem precisará ser reconstruída e analisada.

De modo a levar em conta todos os fatores comentados acima, desenvolvemos um método de segregação de imagem que será descrito na presente seção, e que é baseado em algoritmos otimizados, além de ser capaz de aprender automaticamente a reconhecer objetos a partir de exemplos.

3.4.1 Princípios

A idéia básica do método é de que ele deve ser capaz de aprender e se auto-adaptar de forma automática à região sob análise, uma vez que as características morfológicas e fotométricas de um conjunto de pixels conexos (definido mais adiante) formados pelo sinal de uma ou mais fontes pode variar muito dependendo do algoritmo de reconstrução adotado e da distribuição dos ângulos de trânsito.

A implementação segue uma estratégia de aprendizado computacional supervisionado, que é um processo de dois passos: primeiro o método é treinado em um conjunto de imagens com fontes conhecidas—deste modo ele aprende quais são os tipos diferentes de regiões –, e então ele torna-se capaz de reconhecer a qual tipo de região a estrutura detectada pertence ao ser aplicado em uma outra imagem.

⁷Na literatura a Segregação de Imagem também pode ser encontrada como Segmentação de Imagem, ex. [Gonzalez & Woods \(2002\)](#).

Para a detecção de estruturas na imagem é utilizada uma estratégia de segregação de imagem baseada em corte em diversos níveis adicionada de dois refinamentos: “etiquetamento” de regiões baseada em grafos e construção e análise de grafos para reconhecer quais estruturas equivalem aos mesmos objetos entre os diferentes níveis de corte (uma alteração de uma idéia desenvolvida em [Starck et al, 2000](#), para análise de imagem em multi-escalas).

Para definir a segregação de imagem utilizamos as definições a seguir:

Definição 6 (Imagem) *Uma imagem é uma matriz bidimensional I que representa a versão discreta de um sinal contínuo. Cada posição $I(i, j)$ dessa imagem é chamada de pixel com índice (i, j) , ou simplesmente pixel $I(i, j)$.*

Definição 7 (Fundo) *O Fundo F é um subconjunto de pixels de I que contém apenas o ruído devido ao fundo de céu real e/ou devido ao algoritmo de reconstrução.*

Definição 8 (Objeto) *Um Objeto é definido como uma fonte astronômica real.*

Definição 9 (Região) *Uma região R é um subconjunto de pixels de I conexos que além do sinal do Fundo contém o sinal de pelo menos um Objeto, e/ou que contém um sinal proveniente de artefatos que podem ser criados tanto pelo algoritmo de reconstrução quanto pelo sistema óptico.*

Definição 10 (Independência de Regiões e do Fundo) *Um pixel $I(i, j)$ de uma imagem pode pertencer a uma e apenas uma região R_i ou ao Fundo F , ou seja: $\forall A, B \in \{F, R_1, \dots, R_n\}, A \neq B \Leftrightarrow \bigcap \{A, B\} = \emptyset$.*

Corolário 2 (Composição de uma Imagem) *Toda imagem I pode ser completamente decomposta em termos de Regiões ou Fundo, ou seja, dado n Regiões ($n \in \mathbb{N}$) e um Fundo, a imagem é formada pelo conjunto $I = \bigcup \{F, R_1, \dots, R_n\}$.*

Definição 11 (Segregação de imagem) *A segregação de uma imagem I é a decomposição de uma imagem I em termos do fundo F e das regiões R_n ($n \in \mathbb{N}$). Ou seja, a segregação de imagem é a resolução do seguinte problema: dada uma imagem I , determinar F e R_n .*

3.4.2 Modo com Limiar-Simples

O primeiro modo do método que desenvolvemos é baseado na utilização de um limiar simples de detecção de objetos. Esta é a forma mais simples possível para se realizar uma segregação de objetos em uma imagem, e equivale a primeiro realizar uma binarização da imagem que está sendo analisada, e então analisar a imagem binarizada com a finalidade de encontrar pixels conexos.

Definição 12 (Binarização de imagem em um nível η) *A binarização de uma imagem I em um nível η é a produção de uma segunda imagem I_B formada por pixels $I_B(i, j) \in \{0, 1\}$. Os valores dos pixels de I_B são dados por:*

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

Definição 13 (Pixels 4-vizinhos) Um pixel $I(k, l)$ é 4-vizinho de um pixel $I(i, j)$ se, e somente se $(k, l) \in \{(i-1, j), (i, j-1), (i+1, j), (i, j+1)\}$.

Definição 14 (Pixels 8-vizinhos) Um pixel $I(k, l)$ é 8-vizinho de um pixel $I(i, j)$ se, e somente se $(k, l) \in \{(i-1, j-1), (i-1, j), (i-1, j+1), (i, j-1), (i, j+1), (i+1, j-1), (i+1, j), (i+1, j+1)\}$.

Definição 15 (Pixels conexos) Um par de pixels em uma imagem binarizada é considerado conexo se ambos possuem valor 1 e se ambos são 4- ou 8-vizinhos—denominando-se 4-conexo ou 8-conexo, respectivamente.

Graficamente, os critérios de conectividade definidos acima são representados na Figura 3.10. Embora o método implementado possa ser utilizado com ambos critérios de conectividade (ambos foram implementados), em geral apenas o modo 8-conexo é utilizado.

Figura 3.10: Representação gráfica de 4-conectividade (a) e 8-conectividade (b).

Cada conjunto de pixels conexos define então uma região R_n . A partir da imagem binarizada se determina também o conjunto dos pixels pertencentes ao fundo F , como sendo todos aqueles que possuem valor 0, ou seja, que na imagem I possuem um sinal com intensidade menor do que um certo valor η . Para determinação das regiões R_n o método utiliza um algoritmo de etiquetamento (*connected-component labeling*) em duas-passagens.⁸

Na primeira passagem o algoritmo analisa a imagem binarizada I_B varrendo cada linha da matriz e cria uma matriz I_E (com as mesmas dimensões de I_B) na qual etiquetas temporárias são escritas. Além disso, para cada etiqueta criada durante essa primeira passagem são criados grafos que guardam informações sobre quais

⁸As duas passagens são realizadas da mesma forma que no método seminal proposto em Rosenfeld & Pfaltz (1966), mas introduzimos uma modificação baseada na construção de grafos (inspirada por Fiorio & Gustedt, 1996)—o que deve ser responsável por uma melhor performance, tal como observado por Chang et al (2004) para o algoritmo de Fiorio & Gustedt (1996).

etiquetas são equivalentes, ou seja, quais etiquetas diferentes poderiam ser dadas a um único pixel por ele estar 8-conexo com dois outros pixels de etiquetas diferentes.

A segunda passagem já ocorre na matriz I_E , e tem como objetivo substituir cada etiqueta dessa matriz pela etiqueta com menor valor armazenada no grafo que descreve suas equivalências—a cada etiqueta analisada o grafo é reduzido: uma nova aresta é criada conectando a etiqueta sob análise diretamente ao vértice com etiqueta de menor valor, de modo que para cada etiqueta analisada o grafo precise ser percorrido uma única vez durante toda análise. Um exemplo gráfico das duas passagens desse algoritmo pode ser encontrado na Figura 3.11.

Figura 3.11: Representação esquemática do processo de etiquetamento das regiões usando um algoritmo de duas passagens e critério de 8-conectividade. Em (a) um exemplo de imagem binarizada. Em (b) a primeira passagem. Em (c) os grafos de equivalência de regiões. Em (d) a segunda passagem.

Um fato que pode ser facilmente percebido é que os resultados obtidos por esse modo com limiar simples serão severamente dependentes do valor escolhido para o limiar η durante a binarização. Isso pode ser interessante em casos onde a imagem é dominada pelo fundo de céu, e que alguma medida estatística razoável (como por exemplo uma média calculada com rejeição iterativa em 3σ) ou uma análise da distribuição dos pixels (como em [Bijaoui, 1980](#)) possa ser aplicada para determinação de um valor ótimo para η .

Contudo, no caso das imagens reconstruídas do Gaia esse modo é ineficiente visto que a determinação de um η ótimo é bastante complicada pelo fato de existirem direções preferenciais para os ângulos de trânsito (logo regiões da imagem com diferentes características de ruído) e de desejarmos analisar fontes com contrastes relativamente grandes ($\Delta m \sim 3$) e com separações angulares de poucas centenas de milissegundos de arco (da ordem de duas ou três PSFs). Esses fatores levam a uma complicação grande no método, exigindo a determinação de η 's locais em sub-regiões da imagem. Entretanto, a determinação desses valores locais exigiria que se soubesse (ou estimasse) previamente que existe alguma fonte a ser analisada em determinada posição. Com a finalidade de resolver tais problemas sem a introdução de análises locais, foi desenvolvido um segundo modo, descrito na próxima seção.

3.4.3 Modo com Limiar-Múltiplo

O modo com limiar múltiplo consiste em utilizar o modo de limiar simples globalmente na imagem para diversos valores de η , produzindo diversas listas de regiões, cada uma para um valor diferente de η . Estas listas são então englobadas em um só conjunto, denominado lista global de regiões, e a partir desta lista global determinamos qual o melhor η para cada uma das regiões encontradas.

Definição 16 (Lista global de regiões) *Seja $L_\eta = \{R_1^\eta, \dots, R_n^\eta\}$ um conjunto formado por todas as regiões determinadas pelo modo de limiar simples ao se utilizar um limiar η , o conjunto $P = \{L_{\eta_1}, \dots, L_{\eta_j}\}$ é o conjunto formado por todas as regiões determinadas para j valores distintos de η , e é denominado lista global de regiões.*

Corolário 3 $\forall \eta_i > \eta_j$, um pixel presente em uma região $R_n^{\eta_i}$ encontrada pelo modo de limiar simples aplicado na imagem I com limiar η_i , será encontrado em alguma região $R_k^{\eta_j}$ determinada ao se utilizar o limiar η_j na mesma imagem.

Prova 1 (Corolário 3) *Pela definição de região de uma imagem I , $R_n^{\eta_i}$ é formada por ao menos p pixels $I(i, j)$ 4 ou 8-conexos cujos valores de seus equivalentes $I_B^{\eta_i}(i, j)$ na imagem binarizada $I_B^{\eta_i}$ criada ao se utilizar o limiar η_i são iguais a 1. Mas, $I_B^{\eta_i}(i, j) = 1 \Leftrightarrow I(i, j) > \eta_i \Leftrightarrow I(i, j) > \eta_j, \forall \eta_j \leq \eta_i$ ■*

Corolário 4 *Todos os pixels presentes no conjunto L_{η_i} estarão presentes no conjunto L_{η_j} , $\forall \eta_j < \eta_i$.*

Corolário 5 $\forall \eta_i > \eta_j, \forall R_n^{\eta_i}, \exists R_n^{\eta_j} : R_n^{\eta_i} \subset R_n^{\eta_j}$, ou seja, para toda região detectada em um limiar η_i , existe uma região detectada em um limiar $\eta_j < \eta_i$ que contém todos os pixels da região detectada em η_i .

Os valores escolhidos para os níveis de limiares η podem ser determinados a partir de algum tipo de estatística ou escolhidos a partir de algum tipo de conhecimento *a priori* das características das imagens que serão analisadas—algo interessante, visto que pode-se incluir conhecimento sobre o comportamento de imagens reconstruídas em certos regimes de distribuição de ângulos de trânsito, por exemplo.

A medida estatística padrão utilizada para a determinação do valor do limiar inferior de detecção η_0 é uma média calculada com rejeição iterativa em 3σ : primeiro todos os valores dos $(n \times m)$ pixels da imagem I são copiados em um conjunto $K_0 = \{k_1^0, \dots, k_{(n \times m)}^0\}$, então calcula-se um valor médio $\langle K_0 \rangle$ e o desvio-padrão $\sigma(K_0)$ que servem para definir um subconjunto K_1 criado a partir dos valores de K_0 abaixo da média menos 3σ e então o algoritmo é iterado, sendo que na $n + 1$ iteração, $K_{(n+1)} = \{k_i^n : k_i^n < (\langle K_n \rangle + 3\sigma(K_n))\}$, e as iterações seguem até que $|\langle K_{(n+1)} \rangle - \langle K_n \rangle| < qK_n$, onde o critério de convergência $q = 10^{-3}$ é empírico.

De forma padrão, a escolha dos outros limiares é simplesmente realizada através da determinação do valor de maior intensidade na imagem (tomado como $\eta_{\max}$), e de uma amostragem em um certo número n de intervalos regulares de tamanho $\Delta\eta = (\eta_{\max} - \eta_0)/(n - 1)$. A Figura 3.12 mostra uma representação unidimensional do processo de construção da lista global de regiões.

Figura 3.12: A figura superior mostra uma abstração unidimensional de uma imagem que será segregada em três níveis de limiar. A figura inferior mostra as listas de regiões segregadas em cada limiar—a lista global de regiões é um conjunto que inclui todas as regiões detectadas em todos os limiares utilizados.

Uma vez que se conheça a lista global de regiões P , criam-se grafos, que chamamos de grafos de conectividade inter-limiar (*inter-threshold connectivity graph*) ligando as regiões determinadas em diferentes limiares η —cada vértice dos grafos representa uma região encontrada em algum nível de η (ver Figura 3.13). Para determinar as conexões, são verificados as regiões correspondentes àqueles nós, e as regiões que possuírem pixels em comum são ligadas por arestas—pelo corolário 5 fica garantido

que um vértice determinado a partir de uma região encontrada em um limiar η_n sempre será conectado a um vértice criado a partir de uma região detectada em um limiar $\eta_{(n-1)}$.

A conexão destas regiões é interessante, pois um mesmo objeto deve criar regiões em diferentes limiares, e como o corolário 4 nos garante que qualquer vértice determinado no nível η_i será conectado a algum vértice no nível η_j , para qualquer $\eta_j < \eta_i$, sabemos que tais regiões estarão sempre conectadas entre elas no grafo.

Uma propriedade notória, e que segue do corolário 5, é que o fato de um vértice no nível η_j poder se conectar com um ou mais vértices determinados no limiar η_i transforma a análise desses grafos naturalmente em um indicador de fusão de regiões em algum limiar η_{fusion} tal que $\eta_j \leq \eta_{fusion} < \eta_i$.

Outra propriedade interessante que segue do corolário 5 e disso que acabamos de comentar é que assumindo $\eta_0 = \min(I)$ existirá apenas um vértice, de modo que todos os grafos se conectam neste nível e que a imagem pode então ser representada por um grafo tipo árvore (o que geralmente não fazemos, visto que para finalidades práticas, sempre é adotado um $\eta_0 > 0$).

Figura 3.13: A figura superior mostra a determinação dos vértices do grafo determinado para a abstração da Figura 3.12. A figura inferior mostra o grafo criado a partir da conexão dos vértices com as arestas (incluindo um nível para $\eta = 0$ que transforma o grafo em uma árvore).

O problema a ser resolvido, então, reside em escolher onde se deseja cortar as arestas desses grafos, gerando uma lista final de regiões detectadas. Obviamente, aqui deseja-se escolher as regiões que não estão em estado de fusão, de modo que uma solução trivial para esse problema é cortar o grafo nas arestas que ligam um vértice a apenas um outro vértice, guardando as regiões correspondentes aos vértices mais altos (folhas do grafo árvore) no conjunto final de regiões. Uma outra solução possível, e aquela que é utilizada como padrão no método, é cortar o grafo em arestas

que ligam vértices de grau maior do que 2 (ou seja, que possuem arestas ligando-os a mais do que dois outros vértices). A Figura 3.14 mostra a diferença entre as regiões que estarão na lista final de regiões do modo de múltiplos limiares quando uma ou a outra solução é adotada para o corte do grafo.

Figura 3.14: Grafos indicando quais regiões formam o resultado do modo multi-limiar dependendo do critério utilizado para o seu corte—os vértices azuis são aqueles que compõem o conjunto final de regiões. Na figura superior, corte pelo critério de folhas da árvore. Na figura abaixo com critério de grau maior do que 2.

Cabe notar que este modo é trivialmente paralelizável, uma vez que a execução do modo de limiar simples utilizando diversos limiares pode ser realizada ao mesmo tempo para cada limiar—após isso, a análise e corte do grafo pode ser realizada tanto em paralelo quanto sequencialmente (mas esse é um processo bastante rápido e não deve-se ganhar muito com sua paralelização).

Entretanto o modo pode ser facilmente otimizado mesmo para processamento sequencial: basta analisar a imagem de baixo para cima, criando sub-conjuntos com cada vez menos pixels, visto que os pixels que não pertencem a nenhuma região em um certo limiar η pequeno certamente continuarão não pertencendo a nenhuma região quando um valor maior do que η for utilizado. Finalmente, ele pode ser generalizado para muitas dimensões, possibilitando análise de cubos de dados, por exemplo, sejam eles astronômicos ou não.

O problema desse modo é que ainda que a sua taxa de detecção de objetos seja bastante alta, ele detecta todas as estruturas possíveis, incluindo ruído (caso se utilize limiares de detecção muito baixos e um número muito pequeno de pixels conexos para definir uma detecção), além de todos os possíveis artefatos gerados pelos algoritmos de reconstrução de imagens. Na próxima seção veremos como o efeito de tais problemas foi reduzido.

3.4.4 Modo Educado

O objetivo principal do modo educado é gerar uma lista “limpa” de regiões detectadas, ou seja, com o número mínimo, se possível zero, de artefatos de reconstrução. Como seu próprio nome sugere, esse modo precisa ser educado para funcionar. O passo de educação, ou como é mais comumente chamado, treinamento, é realizado através da apresentação de um conjunto de exemplos ao método—sendo um passo preliminar a segregação das imagens que se deseja analisar. Logo, o modo educado funciona em dois passos distintos:

1. **Treinamento** Durante o treinamento é apresentado ao método uma série de imagens de exemplo, nas quais os objetos são conhecidos previamente (essas imagens podem ser produzidas, por exemplo, por simuladores). O método então utiliza o modo multi-limiar para criar o grafo de detecções, mas não realiza o corte neste grafo, produzindo uma lista de regiões em múltiplos limiares. Para cada uma das regiões detectadas, alguns parâmetros morfo-fotométricos são mensurados a partir de seus pixels, e é produzido um “conjunto de treinamento” com tais medidas e com os tipos de região que se deseja classificar (artefato de reconstrução, objeto simples, fusão de múltiplos objetos), que é armazenada e utilizada para treinar algum algoritmo de classificação supervisionado.

Figura 3.15: Diagrama representando de forma esquemática como ocorre o treinamento do modo educado.

2. **Detecção** Quando se apresenta uma imagem que se deseja analisar ao método, ele irá aplicar o modo multi-limiar nesta imagem (incluindo o corte de grafo). Então, cada região presente na lista produzida pelo modo multi-limiar terá alguns parâmetros morfo-fotométricos mensurados, e baseada em tais medidas, será classificada. Uma probabilidade de que a região pertença a cada classe que o método foi treinado para reconhecer será estimada, e aquela que possuir maior probabilidade será escolhida como a classe da região. Finalmente, uma

lista “limpa” de regiões será produzida através de uma seleção na lista retornada pelo modo multi-limiar em alguma probabilidade escolhida *a priori* para que a região seja formada por regiões que correspondam apenas a objetos simples e/ou fusão de múltiplos objetos.

Figura 3.16: Diagrama representando o passo de detecção do modo educado.

O conjunto de exemplos, ou conjunto de treinamento, pode ser melhor definido no contexto do modo educado utilizando $n - \text{tuplas}$:

Definição 17 (Conjunto de treinamento do modo educado) *Um conjunto de treinamento para o modo educado é um conjunto formado por n -tuplas ($n \geq 2$), na qual $n - 1$ elementos são os parâmetros morfo-fotométricos medidos para os pixels da região e o elemento restante armazena a classe da região (geralmente artefato, objeto simples ou objetos múltiplos não-resolvidos).*

Os parâmetros morfológicos e fotométricos utilizados foram determinados heurísticamente, e para preservar a eficiência computacional do método consistem em um número pequeno de grandezas facilmente calculáveis. Estes parâmetros consistem na seguinte tríade:

1. Número de pixels que definem a região;
2. Nível de detecção da região acima do ruído;
3. Razão entre o desvio padrão da distribuição das intensidades dos pixels e sua média;

A obtenção destes parâmetros é uma tarefa realizada na prática pelos algoritmos de detecção—ou seja, são parâmetros praticamente gratuitos em tempo de cálculo para o modo educado –, entretanto eles são armazenados como propriedades da região apenas quando o modo educado é utilizado.

Outros parâmetros mais complexos podem ser incluídos, se necessário, dependendo das características das imagens sob análise, logo, dos algoritmos de reconstrução utilizados. Contudo para as finalidades de segregação de imagens do Gaia, as

populações que desejamos classificar, que são artefato de reconstrução, objeto simples, fusão de múltiplos objetos, se encontram bem separadas, ocupando regiões distintas do espaço definido por essa tríade, como pode ser visto na Figura 3.17.

Figura 3.17: Exemplo do espaço de parâmetros utilizados para classificação das regiões detectadas. Os pontos azuis representam as regiões correspondentes aos artefatos de reconstrução, os pontos negros aos objetos reais, os pontos vermelhos aos múltiplos objetos fundidos. Os gráficos foram criados utilizando imagens simuladas e reconstruídas pelo algoritmo *shufflstack*.

Para classificar as regiões, qualquer algoritmo de classificação supervisionada pode ser utilizado. A escolha padrão é utilizar um algoritmo bastante simples, chamado de k -NN (ou k -mais próximos vizinhos). Embora seja um dos mais simples algoritmos para reconhecimento de padrões (e de ser considerado como um “clássico”), ele continua sendo objeto de pesquisas recentes, por exemplo Hall et al (2008) para a escolha do valor de k ótimo, ou Toussaint (2005) para otimizações baseadas na construção de grafos.

Intuitivamente é bastante simples entender como o k -NN funciona: ao desejar classificar um ponto de classe desconhecida, calcula-se uma distância entre esse ponto e todos os pontos de um conjunto de pontos com classes previamente conhecidas; então, a classe do ponto desconhecido é assumida como sendo a classe que mais aparece entre esses k vizinhos. Sua definição formal pode ser escrita como:

Definição 18 (Classificação k -NN) *Seja um conjunto de p classes $C = \{c^1, \dots, c^p\}$, tal que $\forall c^i, c^j \in C, c^i = c^j \Leftrightarrow i = j$ —chamamos esse conjunto de conjunto das classes disponíveis. Seja um conjunto denominado formado por m n -tuplas $T = \{t_1, \dots, t_m\}$. Seja um conjunto denominado de conjunto das classes das n -tuplas, um conjunto de tamanho m cujos elementos podem ser apenas elementos do conjunto das classes disponíveis, $C_T = \{c_1, \dots, c_m\}$, tal que $\forall i : 1 \leq i \leq m, c_i \in C$.*

Para determinar a classe de uma n -tupla x qualquer através do método de k -NN, constroem-se o conjunto $D = \{d(x, t_1), \dots, d(x, t_m)\}$ formado pelas distâncias (geralmente euclidianas) entre x e todos os membros de T . Então determina-se em D os k elementos com menores valores, e a partir dos índices desses elementos verifica-se em C_T quantas vezes cada uma das p classes aparecem, formando o conjunto $N = \{N_1, \dots, N_p\}$. Finalmente a probabilidade de que x pertença à i -ésima classe é dada por N_i/k , sendo que a classe de x é definida como a classe para a qual essa probabilidade é máxima.

Uma vez que temos os parâmetros que são medidos para a construção das n -tuplas e um algoritmo de classificação, basta então construirmos um conjunto de treinamento. Para construir esse conjunto no caso do Gaia, utilizamos o GIBIS (como apresentado na Figura 3.15), que é capaz de simular realisticamente as janelas observadas pelo Gaia, e então reconstruímos uma imagem, que é utilizada no modo de treinamento e que serve como exemplo para toda e qualquer outra fonte com aproximadamente as mesmas características fotométricas e posição na esfera celeste.

Temos, então, um método de segregação de imagens que possui capacidade para:

1. Segregar regiões criadas por objetos com grandes contrastes entre si em uma mesma imagem (devido ao modo multi-limiar);
2. Segregar regiões criadas por objetos angularmente próximos entre si (devido ao modo multi-limiar);
3. Segregar regiões sem se basear em uma morfologia pré-determinada para os objetos (devido ao modo de limiar-simples);
4. Discernir entre regiões criadas por objetos simples, múltiplos e artefatos gerados pelos algoritmos de reconstrução (devido ao modo educado);
5. Auto-aprendizado—o que o torna utilizável em imagens normais ou reconstruídas independente do algoritmo de reconstrução (devido ao modo educado).

Mas o modo educado deixa um último problema em aberto que é atrelado ao fato de que os algoritmos de reconstrução de imagem reconstroem de forma diferente as características morfo-fotométricas das regiões dependendo do ruído gerado pelo objeto primário, do número de trânsitos disponíveis para a reconstrução e principalmente da configuração dos ângulos de trânsito—as duas últimas, grandezas dependentes da posição do céu no qual se encontra o objeto.

A solução ideal para esse problema seria ter acesso a uma simulação realista das observações do Gaia em cada coordenada, e para cada magnitude de objeto primário, na qual se deseje realizar uma segregação de imagem. Entretanto tal solução não é realista devido ao custo computacional necessário para realizar tais simulações: ele supera em diversas ordens de grandeza a quantidade de tempo disponível para o processamento de cada imagem reconstruída. A solução encontrada para esse problema foi então criar uma biblioteca com um certo número de conjunto de treinamentos preparados para regiões protótipo.

3.4.4.1 Biblioteca de conjuntos de treinamento

Para que o método desenvolvido pudesse ter uma aplicação geral (em qualquer direção do céu), desenvolvemos uma biblioteca de conjuntos de treinamento. Essa biblioteca consiste em classes que automatizam a geração de conjuntos de treinamento para o modo educado do método de segregação a partir de simulações GIBIS, e principalmente a recuperação do conjunto de treinamento mais adaptado para segregar uma determinada imagem.

Para criação dos conjuntos de treinamento, a biblioteca é alimentada diretamente com os arquivos provenientes das simulações GIBIS e automaticamente lança todos os processos de sistema e códigos em IDL e Java necessários para a reconstrução das imagens, para o treinamento do método de segregação e para a geração de sua estrutura de diretórios.

O uso dessa biblioteca é completamente transparente, sendo que todos os objetos são criados e utilizados internamente pelo código de segregação de imagem. Ao segregar uma imagem o código de segregação de imagem no modo educado requisita automaticamente ao código gerenciador da biblioteca o conjunto de dados mais adaptado para a imagem que está sendo segregada. Para isso o gerenciador da biblioteca analisa:

1. O algoritmo de reconstrução;
2. A magnitude da fonte primária;
3. As coordenadas da imagem no céu;
4. O número de trânsitos.

Uma biblioteca básica composta por uma quantidade mínima de conjuntos de treinamento foi criada com finalidade de testes e demonstrações do conceito. A Figura 3.18 mostra as coordenadas presentes nesta biblioteca básica.

Figura 3.18: Coordenadas presentes na biblioteca básica de conjuntos de treinamento são indicadas com círculos pretos—cada coordenada contém conjuntos de treinamento para três magnitudes e quatro algoritmos de reconstrução.

Cada coordenada amostrada na biblioteca básica (representadas na Figura 3.18) contém conjuntos de treinamento para três magnitudes de objetos primários (12, 15,

19), para número de trânsitos ao final da missão e para todos os 4 algoritmos de reconstrução disponíveis atualmente. Entretanto, bibliotecas com uma granulação muito mais alta nos parâmetros (em especial em termos de posições da esfera celeste) serão criadas.

3.4.5 Testes e resultados

Com a finalidade de determinar o quão bons são os resultados que podem ser obtidos utilizando o método de segregação de imagens descrito nesta seção, utilizamos o GIBIS para gerar simulações de regiões com quatro fontes secundárias ao redor de uma fonte primária. Após reconstruir as imagens utilizando os dados provenientes destas simulações e os quatro algoritmos de reconstrução apresentados no Capítulo 2, aplicamos o método de segregação descrito anteriormente⁹ e analisamos em quais regimes de distância angular e diferenças de magnitude em relação ao objeto primário esse método é capaz de recuperar os objetos secundários.

Adicionalmente também verificamos a quantidade de detecções espúrias, ou seja, de regiões que foram detectadas, mas que não correspondiam a nenhum objeto simulado. É importante notar que as detecções espúrias podem tanto ser artefatos gerados pelos algoritmos de reconstrução, como podem ser regiões geradas por objetos simulados que tiveram suas posições alteradas significativamente pelos algoritmos de reconstrução. Isso, pois a identificação de uma detecção espúria, ocorre quando uma região não inclui as coordenadas do objeto simulado em seu interior. Desse modo, a quantidade de detecções espúrias deve ser vista como um limite superior ao número de artefatos de reconstrução detectados.

Como comentado anteriormente, nosso método de segregação exige uma etapa de treinamento para que seja capaz de reconhecer as regiões. Logo, é natural supor que existirão dois regimes distintos em que seus resultados devam ser testados: um primeiro regime no qual existe na biblioteca um conjunto de treinamento com características iguais ou muito próximas às da imagem que será segregada, e um segundo regime no qual as características são distantes. Vejamos então os resultados para cada um destes dois regimes.

3.4.5.1 Resultados para o caso ideal

O caso ideal da aplicação do nosso método de segregação de imagens ocorre quando as características principais da imagem correspondem (ou são muito próximas de) a algum dos exemplos presentes na biblioteca de treinamento. Essas características são aquelas levadas em consideração ao se selecionar um conjunto de treinamento, ou seja: algoritmo de reconstrução, magnitude da fonte primária, coordenadas da imagem no céu e número de trânsitos.

Naturalmente, o primeiro teste a ser realizado é uma comparação entre os resultados obtidos para os diversos algoritmos de reconstrução de imagem que foram

⁹O implementamos em Java seguindo muitas das especificações requisitadas pelas normas do DPAC (O'Mullane et al, 2008).

apresentados no Capítulo 2, pois como comentamos naquele Capítulo, o algoritmo que será utilizado para a reconstrução durante a redução dos dados ainda não foi definido, e resultados de testes como esse devem ser levados em conta para a seleção.

Nas Figuras 3.19 e 3.20 podemos verificar os resultados obtidos para a recuperação de fontes secundárias ao redor de objetos primários brilhantes ($G=12$) e fracos ($G=19$). Nas simulações, quatro fontes secundárias foram incluídas ao redor de cada objeto primário nas posições norte, sul, leste e oeste, em variadas distâncias angulares em relação ao objeto primário e com várias magnitudes, entre $\Delta G = 0.5$ e $\Delta G = 4.5$.

Já pelos resultados obtidos para a recuperação de objetos em torno de uma primária com $G = 12$ (Figura 3.19), podemos observar que aqueles obtidos com o algoritmo *BinOutliers* e *ShuffleStack* são bem superiores àqueles obtidos com o *Quickstack* e *Drizzle* em termos da quantidade de fontes recuperadas (número de quadrados azuis nos gráficos (a), (c), (e) e (g)). Além disso, ao menos para uma primária de magnitude $G = 12$, demonstramos que somos capazes de recuperar em uma imagem reconstruída o sinal de fontes com grande contraste (até $\Delta G = 4.5$) e próximas angularmente da fonte primária (até $\Delta G = 3.0$ para uma fonte localizada a 300 mas da primária)—o que demonstra a eficiência da segregação realizada através do modo multi-limiar.

Em termos do número de detecções espúrias (resultado apresentado nos gráficos (b), (d), (f) e (h)), ele é essencialmente nulo para todos os algoritmos, com uma pequena exceção para uma detecção espúria realizada em uma imagem *ShuffleStack*. Isso demonstra a eficiência de realizar uma segregação de imagens educada, e em como uma pureza bastante grande na lista de regiões detectadas pode ser obtida.

Algo interessante a ser notado é que durante os testes que realizamos foi utilizado um corte em 90% de probabilidade que a região detectada seja proveniente de um objeto real. No caso de baixarmos tal limiar, o número de fontes recuperadas obviamente aumenta, contudo o número de detecções espúrias também cresce.

Outro comportamento interessante que pode ser percebido na Figura 3.19 (a), é que a região do espaço de parâmetros na qual as detecções de objetos foi bem sucedida possui uma borda a esquerda angulada. Esse comportamento se explica pelo fato de que ao chegar muito próximo da estrela primária, um objeto de certa magnitude não possuirá fluxo suficiente para que o seu sinal se sobressaia do sinal gerado pela primária, apenas deformando a emissão da primária. Entretanto, ainda que esses objetos não estejam na lista de detecções devido ao corte realizado nessa lista, em muitos casos regiões para eles foram detectadas e classificadas como tendo sido originadas por “objetos múltiplos fusionados”—isso permitirá (caso se mostre necessário) o desenvolvimento de um método de *deblending* para separar esses objetos da fonte primária.

O mesmo comportamento descrito acima se repete em (c) para o algoritmo de *Drizzle*, e também em (e) e (g) para o *ShuffleStack* e para o *BinOutliers*, embora com coeficiente angular significativamente mais inclinado para esses dois últimos.

Outro comportamento claro que pode ser observado na Figura 3.19 (a) e (b) é o corte nas detecções de objetos secundários que ocorre em $G \sim 2.25$ para os algoritmos de *QuickStack* e *Drizzle*. Esse corte pode ser entendido como uma limitação do

Figura 3.19: Resultados obtidos pelo método de segregação quando aplicado em imagens reconstruídas para uma estrela primária de magnitude $G=12$. Em (a), (c), (e) e (g) quadrados azuis representam que uma região correspondente ao objeto secundário foi recuperada, enquanto um quadrado vermelho significa que não foi. Em (b), (d), (f) e (h) está indicada a quantidade de detecções espúrias com uma certa magnitude obtida em cada imagem.

Figura 3.20: Resultados obtidos pelo método de segregação quando aplicado em imagens reconstruídas para uma estrela primária de magnitude $G=19$. Gráficos como na Figura 3.19.

método de reconstrução de imagens em reconstruir o sinal de fontes mais fracas do que 2.25 magnitudes em relação à magnitude do objeto primário.

Contudo, as fontes provavelmente mais interessantes do ponto de vista científico serão aquelas recuperadas além do limite de detecção do satélite, ou seja, fontes secundárias com $G > 20$. Ao se analisar os resultados obtidos para detecção de fontes secundárias ao redor de uma primária de magnitude $G = 19$ (apresentados na Figura 3.20), percebemos que ainda existe uma boa taxa de objetos secundários que foram detectados (em especial pelos últimos dois algoritmos), e que a quantidade de detecções espúrias, embora não seja mais completamente nula é bastante baixa (exceção ao algoritmo de *Drizzle*), e virtualmente nula (com apenas uma detecção espúria) no algoritmo *BinOutliers*.

Além disso, podemos perceber que os resultados obtidos pelo código de segregação ao ser utilizado em imagens reconstruídas pelos quatro algoritmos disponíveis novamente ficam divididos em duas classes, com comportamentos bastante similares entre os algoritmos *QuickStack* e *Drizzle* de um lado, e entre *ShuffleStack* e *BinOutliers* de outro.

Os resultados obtidos ao analisar imagens reconstruídas pelos dois primeiros algoritmos é significativamente modulado com uma baixa eficiência ($\Delta G < 1$) até ~ 400 mas, um crescimento grande ($\Delta G < 3$) entre ~ 400 mas e ~ 650 mas e novamente uma baixa eficiência após ~ 650 mas ($\Delta G < 1.75$ para o *QuickStack* e $\Delta G < 1$ para o *Drizzle*).

A ineficiência em regiões mais próximas da fonte primária de nosso algoritmo de segregação quando aplicado em imagens reconstruídas pelos algoritmos *QuickStack* e *Drizzle* se deve provavelmente ao mesmo comportamento que foi observado para a primária com $G = 12$: a emissão da fonte primária “esconde” as fontes secundárias. Já o comportamento ineficiente em distâncias maiores pode ser devido à fonte secundária não conseguir ser distinguida do fundo de detecção (limiar η_0 que foi determinado), enquanto que no caso intermediário sua distinção do ruído é auxiliada por parte do fluxo da fonte primária, que já é baixo o suficiente para não mais “esconder” a fonte secundária mas ainda é forte e regular o suficiente para ajudar o fluxo do objeto secundário ser maior do que o limiar inferior de detecção.

Ao analisar imagens reconstruídas com os dois últimos algoritmos, *ShuffleStack* e *BinOutliers*, podemos perceber que é possível recuperar fontes até duas magnitudes mais fracas do que o limite de detecção do Gaia. É notório que como a primária possuía $G = 19$, fontes até $G = 22$ foram recuperadas de forma aproximadamente estável, o número de fontes fracas que estimamos ser possível recuperar através do fenômeno de projeções ópticas (e que foi apresentado nas primeiras seções deste Capítulo) poderá ser atingido em teoria. No caso do algoritmo *BinOutliers* ser adotado para reconstrução das imagens, esta conclusão é válida para separações de até ~ 300 mas.

3.4.5.2 Resultados no caso não-ideal

O outro regime de utilização do método que desenvolvemos é quando não existe nenhum exemplo que corresponda à região simulada dentro da biblioteca de conjunto de treinamentos. Nesse caso, a biblioteca irá selecionar o conjunto de treinamento criado para uma imagem que tiver seus parâmetros mais próximos àqueles da imagem que se deseja segregar.

Na Figura 3.21 podemos observar os resultados obtidos para a segregação de imagens geradas por reconstruções *BinOutliers* de simulações em uma direção do céu onde não havia nenhum exemplo na biblioteca. Um resultado digno de nota é o fato de que a quantidade de detecções espúrias é novamente virtualmente nula, com apenas uma detecção para a fonte primária simulada com $G = 19$.

Figura 3.21: Resultados obtidos pelo método de segregação quando aplicado em imagens reconstruídas para as quais não existem exemplos próximos na biblioteca. Posição $(\lambda, \beta) = (113^\circ, 65^\circ)$ indicada em (a). Gráficos como na Figura 3.19. (b) e (c) para primária com $G = 15$, (d) e (e) para $G = 19$.

Em relação às detecções de fontes secundárias reais, o comportamento obtido para ambas as primárias é bastante similar, com um crescimento contínuo com aproximadamente a mesma pendente até uma separação de ~ 700 mas. A partir desta separação, para uma primária de magnitude $G = 15$ a detecção continua crescendo até ~ 950 mas para depois decrescer. Enquanto isso, para as fontes secundárias ao redor da primária de magnitude $G = 19$ a eficiência diminui levemente após atingir $\Delta G \sim 3.0$ —isso pode ser um reflexo do limite de magnitude para o qual o *BinOutliers* é capaz de reconstruir uma fonte secundária com eficiência, mesmo porque um limiar em magnitude $G \sim 22.0$ já havia sido observado para o caso de existir um exemplo igual ou muito próximo na biblioteca (ver Figura 3.20 (g)).

Contudo, quando não existe um bom conjunto de treinamento na biblioteca, como no caso analisado na Figura 3.21, a detecção dos objetos secundários é menos eficiente em distâncias pequenas. Isso é apenas um reflexo do corte em 90% de probabilidade para que a região detectada seja uma região formada por um objeto simples—mas diminuir esse nível automaticamente aumenta as detecções espúrias, e como geralmente é melhor possuir um menor número de fontes detectadas mas ter certeza de que elas são fontes reais, mantemos o corte nesse nível mais alto.

Na direção do céu que acabamos de analisar, a cobertura espacial das observações é grande, existindo uma alta variedade de ângulos de trânsito. Como se pode imaginar, os resultados devem ser diferentes caso não existam bons conjuntos de treinamento na biblioteca para uma direção com uma cobertura espacial ruim e com ângulos preferenciais para os trânsitos. Na Figura 3.22 são apresentados resultados obtidos para esse caso mais complicado—complementados com dois exemplos de imagens reconstruídas para fontes localizadas nestas direções.

A primeira diferença de comportamento, notável a primeira vista, no caso dessas duas direções do céu em relação a todos os testes apresentados até agora é o crescimento na taxa de detecções espúrias. Isso já era esperado, pois como direções “mal comportadas” do céu (como as duas que estamos analisando aqui) são mais diferentes entre si do que direções com boa cobertura espacial de observações, os artefatos gerados pelos algoritmos de reconstrução também possuirão características mais diferentes, de modo que a inexistência de conjuntos de treinamento para essas direções dificulta a identificação desses artefatos, gerando mais detecções espúrias.

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 $G = 19$.

Porém, uma análise mais cuidadosa levando em consideração ambos os gráficos, revela que a confiabilidade não é tão ruim, em especial no caso da primária mais fraca: o gráfico das detecções espúrias mostra que poucas das fontes espúrias estão a $\Delta G > 1.75$, mas o gráfico das detecções mostra um número não negligenciável de fontes secundárias detectadas nestas magnitudes—logo, fontes com uma maior confiabilidade nas detecções.

Figura 3.22: Exemplos de reconstruções (b) e (f) e resultados de análise (Gráficos como na Figura 3.19) para projeções ópticas em primárias de magnitude $G = 15$ nas coordenadas eclípticas $(\lambda, \beta) = (120^\circ, 10^\circ)$ e de magnitude $G = 19$ nas coordenadas $(\lambda, \beta) = (110^\circ, 32^\circ)$.

Ainda que seja possível realizar o cálculo de um *a priori* através de estimativas detalhadas da confiabilidade de detecção, e após esse cálculo combinar o *a priori* com a probabilidade calculada pelo *k*-NN de modo a estimar uma probabilidade final para que a região detectada seja um artefato, tal complicação apenas se justificaria no caso em que não fosse possível aumentar a densidade de conjuntos de treinamento na biblioteca para a região de latitudes eclípticas $|\beta| < 45^\circ$ —solução que será adotada para a biblioteca utilizada durante a redução e análise dos dados da missão.

3.5 Conclusões

Neste Capítulo, para estimar o número de casos de projeções ópticas que podem ser observadas pelo Gaia em todo o céu, inicialmente nos baseamos em um modelo no qual as fontes são distribuídas de forma completamente homogênea no céu. Para isso utilizamos dados do catálogo GSC 2.3.2 e extrapolações numéricas da contagem de objetos em determinadas direções do céu, selecionadas de forma a amostrar regiões limites de densidades da Galáxia. Com este cálculo, estimamos tal número em $\sim 2 \times 10^6$, para uma diferença de magnitude máxima de $\Delta m = 3$ entre a fonte primária e a secundária. O menor valor obtido foi $\sim 7 \times 10^4$, enquanto no caso de uma Galáxia completamente formada pela direção de maior densidade, o valor obtido foi de $\sim 3 \times 10^{10}$ (ambos para $\Delta m = 3$). Como os valores obtidos a partir da análise de catálogos não restringiram de forma muito precisa o número de projeções ópticas devido à falta de realismo na extrapolação dos catálogos, número baixo de regiões amostradas e de assumir uma distribuição homogênea de casos no céu, realizamos uma análise mais detalhada.

Apresentamos então tal análise, baseada em simulações de contagem de estrelas obtidas a partir do modelo de Galáxia de Besançon (como implementado na biblioteca GaiaSimu). As simulações, completas até $G = 22$, foram realizadas em mil direções diferentes na Galáxia selecionadas de forma aleatória (com distribuição uniforme na esfera), e a análise levou em conta uma interpolação na esfera e posterior integração dos casos de projeções ópticas. O resultado médio de dez repetições de todo o procedimento permitiu concluir que $1.6 \times 10^8 \pm 5.1 \times 10^7$ projeções ópticas poderão ser observadas no céu inteiro—foram também apresentados valores individuais para o bojo e disco em forma de curvas de probabilidade e do número total de projeções ópticas a cada magnitude de fonte primária.

Mostramos também o método que desenvolvemos para analisar as imagens reconstruídas realizando a segregação dessa imagem em termos de regiões de pixels da imagem—cada região pode ser formada por uma fonte simples, múltiplas fontes misturadas ou artefatos gerados pelos algoritmos de reconstrução. Esse método, chamado de Segregação Educada de Imagens (*Educated Image Segregator*, ou *EIS*) é capaz de aprender e se auto-adaptar de forma automática, desde que seja apresentado a exemplos dos diversos tipos de região que se deseja segregar. Ele é baseado em algoritmos robustos e rápidos para binarização em múltiplos limiares, determinação de conectividade dentro de um mesmo limiar e da conectividade entre os diferentes

limiares baseada em grafos, na análise desses grafos para determinação do melhor limiar para a segregação de uma determinada região, de um classificador supervisionado para a produção da lista final de regiões segregadas e de uma biblioteca de conjuntos de treinamento para esse classificador.

Apresentamos os resultados obtidos com a utilização do método EIS em simulações GIBIS de projeções ópticas e reconstruções de imagens com os quatro algoritmos de reconstrução com implementação disponível atualmente. Esses resultados mostraram que o melhor algoritmo de reconstrução de imagens em termos da quantidade de fontes secundárias recuperadas e mínimo de detecções espúrias é o *BinOutliers*.

Além disso, no caso de existir um conjunto de treinamento na biblioteca com características (principalmente direção do céu) similares àquelas da imagem que se deseja segregar, mostramos que a eficiência da segregação é muito boa, chegando em valores de $\Delta m \sim 3.0$ para uma primária com $G = 19$, e $\Delta m \sim 4.5$ para uma primária com $G = 12$, e isso até separações pequenas, em torno de ~ 300 mas, virtualmente sem detecções espúrias.

Para segregação de imagens sem exemplos com características próximas na biblioteca de conjuntos de treinamento, caso a latitude eclíptica da fonte primária seja $|\beta| < 45^\circ$, a solução ideal será aumentar a quantidade de exemplos na biblioteca de conjuntos de treinamento, pois a lista de fontes secundárias obtidas com nosso método é menos confiável devido a um maior número de detecções espúrias. Entretanto, para o restante do céu ($|\beta| > 45^\circ$), a detecção de fontes secundárias, apesar de menos eficiente para pequenas diferenças angulares do que no caso de existirem bons conjuntos de treinamento na biblioteca, ainda atinge boas diferenças de magnitude $\Delta m \sim 3.0$ com grande confiabilidade (poucas detecções espúrias).

Os resultados obtidos neste Capítulo mostram que além de ser possível segregar imagens para investigação das soluções astrométricas perturbadas, será também possível complementar o catálogo Gaia com muitos milhões de fontes que encontram-se acima do limite de detecção do satélite.

Simulações de Galáxias no modelo de Universo do Gaia

“‘No machine can lie,’ said Father Brown;
‘nor can it tell the truth.’” G. K. Chesterton¹

Conteúdo

4.1	Introdução	104
4.2	Simulação de catálogos de galáxias	104
4.2.1	Princípio de funcionamento	105
4.2.2	Modelo de objetos	106
4.2.3	Método de simulação	111
4.2.4	Validação da implementação Java	115
4.3	Simulação de imagens de objetos extensos – MAGIL	116
4.3.1	Exemplos de imagens simuladas	117
4.4	Estimativas do número de galáxias observáveis	118
4.4.1	Estimativa baseada em extrapolação do Hubble-MDF	120
4.5	Conclusões	121

Além das projeções ópticas, as galáxias com estrutura não resolvida em estrelas individuais são outro caso (e o segundo mais numeroso) no qual as soluções astrométricas providas pelo Gaia devem ser não-ótimas, visto que seus perfis de brilho superficiais não são compatíveis com fontes pontuais. Para que possamos estimar o número de galáxias deste tipo que serão observadas e, principalmente desenvolver e testar os métodos de análise para este tipo de objeto, é necessário de antemão que seja possível simular tanto a distribuição destes corpos na esfera celeste quanto as observações que serão realizadas pelo Gaia.

Neste Capítulo apresentaremos o simulador de catálogos de galáxias chamado JStuff, que implementamos no modelo de Universo do Gaia (Krone-Martins et al, 2008c). Além disso, apresentaremos brevemente o código MAGIL, que foi implementado para gerar as imagens destes objetos no plano focal do satélite e no qual tivemos participação no desenho e especificação (Gavras et al, 2010). Finalmente, apresentaremos uma estimativa para o limite superior do número de galáxias não resolvidas que poderão ser observadas pelo Gaia.

¹“ ‘Nenhuma máquina pode mentir,’ disse o Padre Brown; ‘mas tão pouco pode dizer a verdade.’ ”, in *The Mistake of the Machine*, 1913.

4.1 Introdução

Como vimos na introdução desta tese, devido às características do modo de observação do satélite Gaia, que ocorre em varredura, todas as fontes do céu até certa magnitude serão observadas. Dentre essas fontes encontram-se galáxias que não podem ser resolvidas em estrelas individuais, ou seja, aquelas que são compactas e/ou que estão distantes o suficiente da Terra para que não seja possível distinguir separadamente suas estrelas, mas ao mesmo tempo próximas o suficiente para que seja possível perceber o efeito integrado destas estrelas, que formam os seus diferentes componentes estruturais.

Estas galáxias não resolvidas, da mesma forma que as fontes secundárias apresentadas no Capítulo anterior, são objetos que devem apresentar problemas para o sistema de redução de dados do Gaia uma vez que este foi construído visando fontes puramente pontuais. Contudo, como vimos no Capítulo 1, estas galáxias são bastante interessantes do ponto de vista científico, uma vez que a maior parte delas terá indicações de suas estruturas morfológicas sendo revelada pela primeira vez na história da humanidade justamente a partir das observações que serão realizadas pelo Gaia.

Para que pudéssemos estudar estes objetos, chamados no contexto da missão Gaia de galáxias não resolvidas, foi necessário implementarmos no modelo de Universo deste satélite códigos que possibilitassem produzir simulações. Isso foi realizado em um primeiro momento através da implementação de um simulador de catálogos de galáxias coerente, ou seja que fosse capaz de reproduzir as mesmas galáxias no céu sempre que uma mesma região fosse observada, e que estatisticamente reproduzisse as funções de luminosidades atualmente conhecidas para estes objetos. Esse simulador se chama JStuff ([Krone-Martins et al, 2008c](#)), e será descrito com maiores detalhes na seção 4.2.

Além disso, com a evolução do projeto, percebemos a necessidade de utilizarmos imagens realistas destas galáxias não resolvidas. Para isso, trabalhamos na especificação e no desenho de um simulador que utiliza e transforma imagens prototípicas de uma biblioteca, chamado MAGIL, implementado por [Gavras et al \(2010\)](#), e que será brevemente descrito na seção 4.3.

4.2 Simulação de catálogos de galáxias

O código JStuff é uma implementação em linguagem Java do código Stuff ([Bertin & Arnouts, 1996](#)). A adaptação ao Gaia ocorreu em duas fases distintas: uma adaptação pioneira para uma fase protótipo do satélite foi realizada por [Dollet \(2004\)](#), sendo que a implementação final foi realizada por [Krone-Martins et al \(2008c\)](#). Esta nossa implementação foi realizada puramente em Java, seguindo as normas de desenvolvimento do DPAC, e já continham todas as alterações necessárias para possibilitar a integração ao Modelo de Universo.

Estas adaptações incluíram principalmente modificações estruturais no método de simulação que possibilitassem a adoção de um modelo de orientação a objetos

compatível com os simuladores do Gaia. Tais modificações permitiram que fossem utilizadas rotinas de cálculo presentes na biblioteca GaiaTools (GaiaToolsCommittee, 2010), de modo a prover uma consistência numérica entre todos os simuladores do Gaia, e principalmente possibilitaram a realização de simulações consistentes, ou seja, quando se requisita do JStuff uma mesma região do céu, se obtém a mesma lista de objetos simulados (desde que se mantenham os parâmetros do simulador e do modelo cosmológico sem alterações). Dessa forma foi possível simular diversos trânsitos sob os mesmos objetos, algo impossível antes de nossa implementação.

4.2.1 Princípio de funcionamento

Ao se requisitar uma dada região do céu, o código JStuff tem como objetivo gerar uma lista de galáxias que poderão ser observadas até a magnitude limite desejada. Na versão atual do código, a distribuição das projeções das posições das galáxias na esfera celeste é tomada como aleatória, sendo gerada a partir de uma amostragem uniforme em duas dimensões. Então, o número total de objetos que devem ser gerados naquela região é amostrado a partir de uma distribuição poissoniana, enquanto o número de galáxias de cada tipo de Hubble em cada magnitude é amostrado a partir de funções de luminosidade de Schechter que possuem coeficientes independentes para cada tipo de Hubble.

O primeiro passo da simulação é repartir a esfera em algum tipo de região geométrica para que dentro de cada uma destas regiões o JStuff gere uma lista de galáxias—estatisticamente, as listas de objetos de todas as regiões são equivalentes. Para simulações em pequenas regiões pode-se utilizar círculos ou retângulos, mas triângulos são utilizados quando se realizam simulações em larga escala.

Figura 4.1: Áreas dos triângulos gerados por uma composição da esfera pelo HTM.

A repartição triangular é realizada a partir de uma subdivisão da esfera chamada *Hierarchical Triangular Mesh* (Barrett, 1995; Kunszt et al, 2000), que é uma subdivisão em triângulos esféricos com aproximadamente as mesmas dimensões. Essa

técnica gera uma estrutura de dados tipo *quad-tree* para a esfera, ou seja, em um primeiro nível se divide a esfera em oito triângulos esféricos (esse nível é formado pelo octaedro inscrito na esfera), então em um segundo nível divide-se cada um destes triângulos em quatro outros triângulos, no terceiro, cada triângulo é novamente dividido em quatro, e assim por diante a cada nível adicional. A Figura 4.1 mostra a distribuição de áreas destes triângulos para uma decomposição no nível 4.

Para que o código JStuff gere sempre a mesma lista de objetos quando uma certa região do céu é analisada, o índice HTM dessa região é utilizado para determinação da semente do gerador de números aleatórios—logo, como esse gerador é inicializado sempre com a mesma semente para uma mesma região, o catálogo de objetos será sempre o mesmo para todos os trânsitos em uma determinada região.

4.2.2 Modelo de objetos

O que o JStuff deve gerar dentro destas regiões são objetos extragalácticos, e a representação destes corpos celestes dentro de um computador precisa ser descrita de algum modo. Essa descrição, é realizada a partir de “objetos” computacionais, uma vez que segundo as normas do DPAC, todos os códigos do Gaia devem ser escritos em linguagem Java. Um objeto computacional, neste caso, pode ser entendido como uma abstração de um objeto real dentro da memória de um computador, ou seja, é algo que possui “características”, como cor, e que pode realizar “ações”, como explodir. Em linguagens de programação modernas, estas características são denominadas de atributos, e as “ações” de métodos. Mas qualquer ente existente é na realidade apenas um representante específico de alguma “classe”.

Uma televisão, por exemplo, é apenas uma das milhares representantes da “classe” televisão, e existem milhões de televisões no mundo, que embora possam ser de modelos idênticos, tiveram histórias de construção e uso diferentes, logo possuem características com “valores” diferentes. A cada um destes representantes, denominamos de instância, ou seja, ao criar um objeto na memória do computador, diz-se que uma certa classe está sendo instanciada. Logo, estas representações devem ser modeladas antes da programação em si começar.

Para modelar os objetos extragalácticos do simulador de catálogos de galáxias não resolvidas nos inspiramos no modelo previamente criado para o simulador de estrelas, que é uma versão em Java do modelo de Galáxia de Besançon (Robin et al, 2007). Desse modo foi possível propormos a adoção de um modelo coerente no do modelo de universo do Gaia, que até a época da implementação do JStuff em meados de 2006, ainda não existia.

Seguindo modelos de objetos implementados anteriormente, implementamos a classe `UnresolvedGalaxy` adotando classes abstratas—em programação orientada a objetos, esse tipo de classe representa conceitos abstratos, que nunca são instanciados (ou seja, criados na memória do computador) e que servem apenas para definir modelos de comportamentos que são implementados em classes específicas, herdeiras da classe abstrata (este conceito é bastante simples, uma classe “ser vivo”, por exemplo, nunca deveria ser instanciada, mas sim “bromélia”, “rato”, “homem”, etc.).

O diagrama da Figura 4.2 representa a hierarquia de herança que foi adotada no modelo de Universo do Gaia, com as classes implementadas por nós indicadas em vermelho. A classe **AstroSource** naquele diagrama é um bom exemplo do conceito de uma classe abstrata: embora existam “fontes astronômicas”, esse nome não define um objeto do mundo real mas sim um “conceito de objeto do mundo real”, enquanto o que realmente existe são as galáxias, os asteróides, os quasares, as estrelas, etc.

Figura 4.2: Diagrama de herança da classe **AstroSource**—versão colapsada.

A classe **UnresolvedGalaxy** é aquela que descreve as galáxias não resolvidas em estrelas, e que guarda todos os métodos e atributos das galáxias que são simuladas. Ela herda todos os métodos e atributos de uma classe abstrata comum, chamada de **ExtragalacticSource**, que é uma classe pai para todas as classes que lidam com galáxias e com quasares, como indicado no diagrama UML² na Figura 4.2³.

Figura 4.3: Diagrama de herança da classe **PhysicalParameters**—versão colapsada.

As características físicas da galáxia simulada são manipuladas (armazenadas e recuperadas) por meio de uma especialização da classe **PhysicalParameters** que

²Diagramas UML, ou *Unified Modeling Language* são uma forma padronizada para se indicar a comunicação e interdependência entre objetos.

³Apresentamos primeiro as versões colapsadas dos diagramas, ou seja, sem os nomes dos atributos e métodos de cada classe, para facilitar uma compreensão geral da herança.

denominamos `UnresolvedGalaxyPhysicalParameters`, e que encontra-se representada na Figura 4.3. É nesta classe, por exemplo, que armazenamos, uma instância de uma classe que é muito importante para qualquer galáxia, chamada de `HubbleType`, e que veremos com maiores detalhes na próxima seção.

Figura 4.4: Diagrama de herança da classe `SourceGenerator`—versão colapsada.

Finalmente, a classe que é responsável pelas simulações dos catálogos propriamente ditos é denominada `ExtragalacticGenerator`, e encontra-se representada na Figura 4.4. A classe que gera os catálogos por meio do `JStuff` é a `UnresolvedGalaxyGenerator`.

Diagramas estendidos

Os diagramas de classe em UML também podem ser utilizados para indicar quais atributos e métodos as classes possuem, além das relações entre as classes. No diagrama apresentado na Figura 4.5 vemos a herança da classe `UnresolvedGalaxy` com alguns atributos e métodos representados.

A classe abstrata `AstroSource`, naquele diagrama, possui como atributos tanto características simples que quaisquer objetos dentro de um simulador devem possuir, como o código identificador do objeto (`id`) ou o seu tipo (`idType`), como também atributos mais complexos que na verdade se tratam de instâncias de outras classes. Dentre estes atributos complexos, temos os parâmetros astrométricos, fotométricos, físicos, de forma, e espectroscópicos. As instâncias destes atributos mais complexos trazem consigo os métodos também, mas para serem utilizados externamente (ou seja, por um código que hipoteticamente instanciaria a classe `AstroSource`), estes precisam ser chamados por métodos da própria classe (um conceito chamado encapsulamento). Logo, métodos para lidar com alguns destes atributos também são incluídos.

A classe abstrata `ExtragalacticSource`, que é uma especialização da classe `AstroSource` estende esta última para parâmetros mais característicos dos objetos extragalácticos, como por exemplo, o *redshift*. Finalmente, a classe que é instanciada para representar as galáxias não resolvidas, mas que herda as últimas duas

classes comentadas, é a `UnresolvedGalaxy`. Esta classe instancia outra chamada `UnresolvedGalaxyPhysicalParameters`, que lida com parâmetros físicos do objeto.

Figura 4.5: Diagrama de herança da classe `AstroSource`.

A herança da classe `UnresolvedGalaxyPhysicalParameters`, além de seus métodos e atributos, encontra-se explicitada no diagrama da Figura 4.6. Esta classe é responsável pelo gerenciamento dos parâmetros físicos dos objetos. A classe abstrata pai `PhysicalParameters`, da qual esta classe é herdeira, possui apenas a massa do objeto como atributo, e como considerou-se que todos os outros atributos seriam particulares de cada tipo de objeto astronômico, deixou-se mais definições para as classes filho. Dentro de `UnresolvedGalaxyPhysicalParameters` temos como atributos simples os tamanhos, a elipticidade, ângulo de posição (ângulo projetado no céu) dos bojo e do disco da galáxia, além da razão bojo/disco. Esta classe também possui um atributo mais complexo, que é uma instância de outra classe, e que representa parâmetros definidos especificamente pelo tipo de Hubble da galáxia. Este atributo é uma instância da classe `HubbleType`, que por sua vez possui atributos simples e complexos, representando por exemplo as funções de luminosidade e distribuições espectrais de energia para o disco e para o bojo.

Figura 4.6: Diagrama de herança da classe PhysicalParameters.

A classe mais central no simulador, é a denominada `ExtragalacticGenerator`, pois possui a tarefa de gerar as instâncias da classe `UnresolvedGalaxy` que devem ser encontradas na direção do céu requisitada da simulação. Como vimos na seção anterior essa classe é herdeira de `ExtraGalacticGenerator` e `SourceGenerator`, de quem ela herda a interface do método de simulação, que é comum para todos os simuladores implementados no modelo de Universo. O diagrama para esta classe pode ser visto na Figura 4.7, enquanto o método de simulação propriamente dito é apresentado na próxima seção.

Figura 4.7: Diagrama de herança da classe `SourceGenerator`.

4.2.3 Método de simulação

O método adotado para a simulação dos catálogos de galáxias não resolvidas é uma adaptação do código `Stuff`, primariamente descrito em Bertin & Arnouts (1996) e Erben et al (2001). Nossa re-implementação do método descrito naqueles trabalhos é chamada de `JStuff` e foi baseada nos códigos originais escritos em C por E. Bertin. Além de possibilitar a inclusão de simulações de galáxias de forma coerente no modelo de Universo do Gaia, nosso trabalho também permitiu a detecção e correção de um pequeno erro existente no código original.⁴

O código `JStuff` simula um catálogo de galáxias para uma certa região do céu através de uma amostragem realizada a partir de uma distribuição poissoniana do número de galáxias de um dado tipo de Hubble (E, S0, Sab, Sbc, Scd, Irr, QSFG) assumindo uma função de luminosidade⁵ de Schechter (Schechter, 1976) do tipo:

⁴E. Bertin, *private communication*, 13 nov. 2006.

⁵O código adota uma função de Schechter sem envolução, o que significa que nenhuma dependência com o *redshift* do tipo morfológico foi introduzida—mas isso é algo que poderia ser adicionado de forma simples caso necessário.

112 Capítulo 4. Simulações de Galáxias no modelo de Universo do Gaia

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

Os coeficientes desta função de Schechter que são adotados atualmente para cada tipo de galáxia foram obtidos do estudo de [FIOC & Rocca-Volmerange \(1999\)](#), e podem ser encontrados na tabela 4.1.

Tipo	Φ^* [Mpc ⁻³]	M^* [mag]	α	BT (bojo/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

Tabela 4.1: Parâmetros das funções de luminosidade utilizados atualmente no JStuff. E2 representa “elíptica vermelha” e “E-S0” elípticas normais ou S0s ([Robin, 2010](#)). BT indica as taxas típicas de luminosidade entre o bojo e a total.

Esta amostragem é realizada dentro de um elemento de volume, que é definido por uma figura geométrica, um *redshift* médio e sua extensão no espaço de *redshift*. Um *redshift* máximo é escolhido, e o simulador cria os elementos de volume nos quais as simulações são realizadas. Geralmente, como o JStuff é chamado a partir do modelo de Universo do Gaia, são adotadas simulações com regiões triangulares, mas implementamos também formas de executar o JStuff em modo *stand-alone*, e regiões retangulares e circulares (um cone no espaço de *redshift*) podem ser requisitadas.

Então, cada galáxia dentro do elemento de volume atual pode ser descrita por uma soma de dois componentes: um esferóide e um disco.⁶ O primeiro seguindo uma lei de de Vaucouleurs:

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

onde M_S é a magnitude absoluta na banda B do esferóide (componente do bojo), e r_e seu raio efetivo em parsecs. Para galáxias elípticas o raio efetivo do esferóide é calculado a partir de sua magnitude absoluta a partir de uma relação descrita por [Binggeli et al \(1984\)](#):

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

⁶A partir desta soma, uma imagem pode ser simulada, o que é realizado por um outro código de E. Bertin, o SkyMaker (ver seção 4.3).

O achatamento q do esferóide é aleatoriamente selecionado a partir de uma distribuição normal entre 0.3 e 1.0, com média $\langle q \rangle = 0.65$ e desvio padrão $\sigma_q = 0.18$. Estes valores foram obtidos de Sandage et al (1970).

Já o segundo componente, o disco, quando presente (ou seja, quando a galáxia não é uma galáxia elíptica), é descrito por um perfil exponencial:

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

onde r_h é a escala e μ_0 o brilho superficial central. O valor deste último é aleatoriamente selecionado de uma distribuição gaussiana com média $\langle \mu_0 \rangle = 21.65 \text{mag.arcsec}^{-2}$ e desvio-padrão $\sigma_{\mu_0} = 0.35$ para magnitude absolutas $M \leq -17$, e média $\langle \mu_0 \rangle = 21.65 + 0.7(M + 17) \text{mag.arcsec}^{-2}$ caso contrário. A escala do disco r_h é obtida a partir da seguinte relação:

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

onde M_D é a magnitude absoluta do disco visto em *face-on* na banda B.

A inclinação e o ângulo de visada são selecionados aleatoriamente de uma distribuição uniforme. Para obtermos M_S para uma dado M , é utilizado um ajuste empírico encontrado em Simien & de Vaucouleurs (1986):

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

onde T é o tipo de Hubble⁷.

A extinção interna no disco é simulada utilizando uma curva de extinção empírica derivada para a Grande Nuvem de Magalhães, e correções K+e são aplicadas utilizando ajustes polinomiais de Metcalfe et al (1991).

Cada galáxia é então colocada em seu *redshift* de forma aleatória dentro do elemento de volume que está sendo simulado, com sua luminosidade bolométrica. A distância de diâmetro angular é calculada por:

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

onde

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

E então é iniciado o cálculo das magnitudes em cada uma das bandas requisitadas pelo usuário. Finalmente, quando todas as galáxias dentro daquele elemento de volume em *redshift* foram calculadas parte-se para o próximo elemento, sendo que o programa interrompe a simulação ao atingir $z = 0$.

Um diagrama de fluxo do método que representa passo a passo o que ocorre durante a simulação pode ser visto na Figura 4.8.

⁷Para $T \leq 4$, $M_S = M$.

Figura 4.8: Diagrama de fluxo do método de simulação adotado pelo simulador JStuff. O *Volume Element* é uma região do céu tri-dimensional que é definida por uma figura geométrica qualquer, um *redshift* central e um intervalo de *redshift*.

4.2.4 Validação da implementação Java

Para validar a implementação Java realizamos testes comparando, para os mesmos arquivos de parâmetros de entrada do código, os resultados obtidos com ambas as implementações do método de simulação (JStuff e Stuff). Essas comparações demonstraram que as listas de objetos obtidas a partir da utilização de ambos os códigos são compatíveis dentro de 1σ , como pode ser visto na Figura 4.9.

Figura 4.9: Comparação de distribuições de parâmetros obtidos com o JStuff e com o Stuff.

Além de realizar este teste de validação, o código JStuff também incorpora testes automáticos que são executados periodicamente a cada integração dos simuladores do

Gaia. Estes códigos são escritos como JUnits, ou seja, são testes unitários: os métodos individuais dos objetos são verificados individualmente de forma completamente automática, e caso algo seja modificado no código que altere os resultados esperados de algum método, alarmes são executados alertando para o resultado inesperado—isso é necessário devido a interdependência entre os diversos códigos que são utilizados nos simuladores, e da grande quantidade de pessoas que os alteram.

4.3 Simulação de imagens de objetos extensos – MAGIL

Contudo, diversos estudos necessitam não apenas de catálogos de posições de objetos, mas sim das imagens destes objetos—o exemplo mais direto é a determinação da eficiência de detecção das galáxias não resolvidas. Para isso é necessário que não apenas os catálogos de galáxias sejam simulados, mas que também suas imagens possam ser sintetizadas ao nível de pixel no plano focal do satélite. Para isso, uma primeira adaptação de um método denominado Skymaker, de E. Bertin foi realizada por C. Dollet, sendo posteriormente adaptada para as normas do DPAC e re-escrita em Java por C. Delle-Luche. Este código cria a imagem das galáxias a partir de uma superposição de um perfil exponencial para o disco e de um esferóide com perfil de de Vaucouleurs.

Entretanto, para possibilitar um maior realismo nas galáxias simuladas, inclusive possibilitando estudar a detecção de clumps nos discos das galáxias próximas ainda que estas galáxias não sejam detectadas, desde o início do DPAC tem-se pensado criar um simulador de imagens de galáxias baseado no uso de bibliotecas de imagens de galáxias próximas previamente existentes ([Babusiaux, 2005](#)).

O simulador que foi implementado e que cumpre esse papel, chama-se MAGIL (*Manager of Gaia Image Library*), e suas especificações e diagramas básicos do funcionamento foram criadas por P. Gavras e A. Krone-Martins durante discussões na *ELSA School on the Science of Gaia*, em 2007. A primeira implementação funcional em Java foi criada por P. Gavras, tendo sido disponibilizada em meados de 2008 e acoplada ao GIBIS naquele mesmo ano. Uma descrição completa deste simulador pode ser encontrada em [Gavras et al \(2010\)](#).

O MAGIL funciona basicamente utilizando três módulos distintos: um controlador, um processador e uma biblioteca—um diagrama representando a relação entre estes módulos pode ser encontrado na Figura 4.10. Esta última, a biblioteca, é basicamente um banco de dados de imagens, que além de armazenar as imagens dos objetos protótipos também armazena parâmetros destes objetos (como seus tipos, razões axiais e tamanhos angulares, por exemplo). O processador é responsável por manipular imagens, sendo capaz de alterar a razão entre os eixos maior e menor e o tamanho absoluto do objeto em pixels, de re-escalar a imagem em fluxo e de remover fundos de céu significativos.

O terceiro componente, o controlador, é responsável por orquestrar o funcionamento do sistema, lidando com os requerimentos do usuário (através de uma interface GUI *stand-alone* ou do GIBIS), requisitando à biblioteca a imagem ade-

Figura 4.10: Esquema geral do processamento que ocorre no MAGIL.

quada, determinando as alterações necessárias para que a imagem do banco de dados possa ser alterada pelo processador, requisitando ao processador as alterações, e devolvendo a imagem resultante ao usuário.⁸

4.3.1 Exemplos de imagens simuladas

Apresentamos aqui dois exemplos de galáxias geradas pelo MAGIL junto com as imagens protótipos que foram utilizadas. A primeira galáxia requisitada pode ser vista na Figura 4.11, e é uma galáxia espiral, com magnitude 17 e razão axial unitária—a matriz retangular no canto superior direito é o resultado final.

Figura 4.11: Exemplo de imagem de uma galáxia espiral gerada pelo MAGIL.

A galáxia espiral simulada acima, tem tamanho angular total de $3''$, mas em princípio deveria ter seus dados transmitidos ao solo devido ao seu perfil de brilho

⁸O tratamento da PSF da imagem é realizado dentro do GIBIS, e ocorre de forma coerente entre todas as imagens que estão sendo formadas no plano focal do satélite.

central, o que a torna quase uma fonte pontual. Um segundo exemplo seria a galáxia Irregular mostrada na Figura 4.12, que possui um perfil de brilho que não decai rapidamente e que possui uma extensão angular de ~ 700 mas, ou seja, que estaria no limite de detecção.

Figura 4.12: Exemplo de imagem de uma galáxia irregular gerada pelo MAGIL no plano focal simulado pelo GIBIS. Foi aplicado um zoom de 4x, e um mapa de cores em escala logarítmica. (Figura de Gavras et al, 2010)

4.4 Estimativas do número de galáxias observáveis

Derivar uma estimativa para o número de galáxias que será observável pelo Gaia exige um conhecimento detalhado da resposta dos algoritmos de detecção de objetos a bordo do satélite aos formatos dos objetos extensos. Por exemplo, ainda que saibamos que existe um corte em 700 mas, um objeto maior do que este tamanho poderia, a princípio, ser observado desde que seu bojo tivesse menos do que esse tamanho, e que esse bojo dominasse a emissão da galáxia. Exemplos foram mostrados nas imagens simuladas das Figuras 4.11 e 4.12.

Contudo, infelizmente até o momento da redação desta tese, os códigos de detecção que serão utilizados na missão não se encontravam integrados ao GIBIS que é disponibilizado para uso no CNES, logo não foi possível determinarmos detalhadamente qual será esta resposta. Derivamos então estimativas dos limites superiores de galáxias não resolvidas que poderiam ser observadas pelo Gaia.

Os resultados das simulações oficiais mais recentes (05/2010) realizadas com o GOG e que utilizaram o JStuff, foram apresentados em Robin (2010). Estes indicam que o limite superior de objetos que poderiam a princípio ser observados seria de $\sim 3 \times 10^7$ galáxias não resolvidas até magnitude $G \sim 20$. Contudo, é realista supor que para objetos extensos, como galáxias, a magnitude limite será um pouco inferior, sendo que re-analisando o mesmo lote de simulações considerando apenas galáxias com magnitudes entre 11.5 e 18.5, o número de objetos existentes no céu inteiro seria de 6×10^6 galáxias, cuja distribuição pode ser vista na Figura 4.13.

Analisando a distribuição de *redshift* das galáxias (estatística atualmente disponível apenas para $G \leq 20$) que poderiam ser observadas pelo satélite, estimamos

Figura 4.13: Distribuição de galáxias não resolvidas no céu para magnitudes $11.5 \leq G \leq 18.5$. As zonas com contagens incorretas no plano da galáxia são causadas por um erro no modelo de extinção no momento da simulação. Simulação do catálogo Gaia com o GOG (ciclo 8, 05/2010).

que provavelmente o valor médio observado para todos os tipos morfológicos será de $\langle z \rangle \sim 0.17$, com a maior parte dos objetos sendo possuindo magnitudes $G > 17.0$. Estes resultados podem ser vistos graficamente na Figura 4.14.

Figura 4.14: Histograma 2D das galáxias não resolvidas no espaço *redshift* vs. magnitude *G* obtido em simulação do catálogo Gaia com o GOG (ciclo 8, 05/2010).

4.4.1 Estimativa baseada em extrapolação do Hubble-MDF

Uma estimativa completamente independente do número de galáxias não resolvidas observáveis pode ser obtida a partir de uma extrapolação para a esfera celeste da densidade de objetos com as características que esperamos observar com o Gaia e que já foram observados no Hubble *Medium Deep Field* (Griffiths et al, 1994). Estas observações são uma série de imagens realizadas em modo paralelo, ou seja, enquanto o Hubble observava com um instrumento um certo alvo científico, a *Wide Field Camera* observava um campo próximo qualquer.

Os dados utilizados em Griffiths et al (1994), são especificamente originados a partir de um campo próximo ao quasar 3C273. O tamanho mediano obtido para os objetos nesta imagem foi de $0''.4$, demonstrando que muitos destes objetos passariam efetivamente pelo corte de 700 mas existente a bordo do satélite, gerando “objetos problemáticos” durante a redução dos dados. Notamos que este artigo mostra que o melhor modelo cosmológico para explicar o número de galáxias de um certo tamanho, é um modelo rico em galáxias anãs.

Figura 4.15: Densidades de galáxias na banda V obtidas a partir do *Hubble Medium Deep Field*. Figura de Griffiths et al (1994).

Para magnitude $V \sim 20$, a densidade estimada é de $\sim 10^3$ [gal.mag $^{-1}$ deg $^{-1}$], como pode ser visto na Figura 4.15. A maior parte das galáxias observadas no MDF possuem tamanhos angulares menores do que $0''.7$, logo seriam possivelmente observadas pelo Gaia (caso a magnitude limite permitisse). Extrapolando os valores de densidade em $V \sim 20$ para o número de objetos no céu inteiro, podemos estimar como um limite superior $\sim 3 \times 10^7$ galáxias que o Gaia poderia observar (excluindo a faixa da Via-Láctea). Cabe comentar, no entanto, que por mais que este resultado seja incerto, pois esta estimativa de densidade foi baseada em um pequeno número de objetos, o valor obtido é perfeitamente compatível com aquele obtido a partir das simulações do GOG com o JStuff apresentadas na seção anterior para $G=20$.

4.5 Conclusões

Neste Capítulo apresentamos o código JStuff, que implementamos dentro do modelo de Universo da missão Gaia para possibilitar a realização de simulações coerentes de catálogos de galáxias não resolvidas em objetos individuais da missão (ou seja, simulações que possam se repetir a cada vez que uma mesma região do céu é simulada). Apresentamos também o código MAGIL desenvolvido por P. Gravas com nossa colaboração, que possibilita a introdução de imagens de quaisquer objetos extensos, mas principalmente de galáxias não resolvidas, no simulador de observações do Gaia a nível de pixel (GIBIS).

Finalmente, utilizando dados simulados do GOG que adotaram o JStuff e um modelo de extinção devido à Via-Láctea, um estudo preliminar permite obter uma estimativa para o número superior de galáxias não resolvidas que poderá ser observado pelo Gaia, resultando em $\sim 6 \times 10^6$ galáxias até $G = 18.5$ e 3×10^7 até $G = 20$, um valor que mostramos ser compatível com resultados da extrapolação do Hubble *Medium Deep Field*. Além disso, essas simulações também possibilitaram a estimativa do *redshift* médio dos objetos que devem ser observados em $\langle z \rangle \sim 0.17$.

Análise e classificação de Imagens de Galáxias

“Crude classifications and false generalizations are the curse of organized life.”¹
G. B. Shaw

Conteúdo

5.1	Introdução	124
5.2	O espaço CASGM20	126
5.2.1	Testes com dados da literatura—PCA	136
5.3	Implementação para o Gaia – CASGM20	139
5.3.1	Testes com galáxias do catálogo de Frei	141
5.3.2	Cálculo de CASGM20 em imagens do <i>Hubble Deep Field</i>	143
5.3.3	Testes com simulações	147
5.4	Classificação e Support Vector Machines	151
5.5	Implementação para o Gaia – SVM	156
5.5.1	Validação com gaussianas	158
5.5.2	Testes com galáxias do catálogo de Frei	162
5.5.3	Testes com simulações GIBIS	163
5.6	Aplicação aos dados do <i>Hubble Deep Field</i>	166
5.7	Conclusões	168

Embora a missão espacial Gaia não seja voltada à observação de objetos extragalácticos, uma janela ainda não mapeada será aberta por essa missão. Entretanto, devido ao critério de selecção de objetos a bordo, serão transmitidos dados apenas de galáxias angularmente pequenas e eventualmente de bojos. Além disso, apenas $\sim 10^2$ pixels estarão disponíveis, e como vimos no Capítulo 2 as imagens reconstruídas serão afetadas por artefatos e a qualidade da reconstrução será variável a cada pixel.

Neste capítulo apresentaremos e analisaremos parâmetros robustos que serão utilizados durante a análise de imagens reconstruídas para possibilitar uma posterior classificação morfológica das galáxias observadas (Krone-Martins et al, 2008b). Discorreremos ainda sobre o método de classificação que será utilizado e apresentaremos testes empíricos em distribuições gaussianas, imagens de galáxias próximas, imagens reconstruídas a partir de simulações da missão Gaia, e de galáxias observadas no *Hubble Deep Field* (Krone-Martins et al, in prep.a).

¹“Classificações rudimentares e generalizações falsas são a maldição da vida organizada.”

5.1 Introdução

O processo mais primário em ciências observacionais é a taxonomia, ou a organização de objetos de acordo com suas semelhanças e diferenças. Esse é o primeiro passo para que a partir destas características seja possível derivar relações entre os objetos, seus comportamentos individuais ou coletivos.

No caso de galáxias ou de quaisquer outros objetos extensos, a morfologia é o traçador mais acessível da estrutura física destes objetos, e logo o mais básico e natural caminho para realização de classificações. A compreensão da evolução desta morfologia é uma das principais chaves para a compreensão da evolução do próprio Universo, em particular da distribuição de matéria nas estruturas observadas nos dias atuais, e nas discussões sobre os cenários de formação de estruturas, como por exemplo o cenário hierárquico.²

Tradicionalmente a classificação de galáxias é realizada utilizando o esquema que as separa em Elípticas, Lenticulares, Espirais normais e Espirais barradas, originalmente proposto por Hubble (1926, 1936) e Sandage (1961), com as galáxias que não se encaixam nessas categorias sendo classificadas separadamente como Irregulares.³ Embora esse esquema de classificação funcione com sucesso para as galáxias brilhantes próximas, ele pode se tornar ineficiente para objetos compactos e/ou pouco luminosos, para classificação em aglomerados ricos e objetos em altos redshifts, como observado por exemplo em Abraham et al (2003).

Além disso, outra crítica a esse sistema é que ele possui um embasamento pouco quantitativo, sendo que mesmo entre astrônomos classificadores experientes, existem divergências sistemáticas na classificação de um mesmo objeto, em especial no caso de galáxias distantes, compactas (ou com pouca informação) e galáxias peculiares, como pode ser notado nos trabalhos de Naim et al (1995) e Abraham et al (1996b).

Finalmente, os principais parâmetros que definem o tipo de Hubble, como a razão bojo-disco, a abertura e o grau de resolução dos braços espirais, a existência ou não de uma barra, são pouco resistentes a degradação da qualidade espacial da imagem e a baixa relação sinal-ruído, fatores que devem prevalecer nas imagens reconstruídas para missão Gaia. Desse modo é necessário utilizar um esquema de classificação baseado em características distintas das supracitadas para se extrair informações científicas das imagens reconstruídas com os dados dessa missão. Contudo devido ao

²Esse denominação se deve ao fato de que neste cenário desde o início do Universo as estruturas maiores são formadas por fusão de estruturas menores (White & Frenk, 1991, e referências), e que constitui uma teoria com ampla aceitação (Springel et al, 2005). Entretanto, trabalhos recentes vem apresentado evidências não explicadas por esse cenário, por exemplo: Collins et al (2009) mostra que os aglomerados massivos de galáxias se formam rapidamente ($< 4 - 5$ Gyr), e não lentamente como o modelo hierárquico propõe. O trabalho Delgado-Serrano et al (2010) mostrou que o número de galáxias peculiares diminuiu enquanto o de espirais cresceu nos últimos 6 Gyr, e que este fato foi o único responsável pela evolução morfológica, pois o número de elípticas e lenticulares se manteve constante.

³Na realidade, a classificação é mais fina, contendo diversos tipos de galáxias elípticas, espirais e critérios visuais para a diferenciação entre tais subtipos. Duas referências abrangentes sobre o assunto tanto do ponto de vista histórico quando do ponto de vista conceitual são Van den Bergh (1998) e Sandage (2005).

seu apelo conceitual e astrofísico forte (Kormendy & Kennicutt, 2004; Sandage, 2005), é desejável que qualquer que seja o sistema utilizado para classificação morfológica no Gaia, este possa ser traduzido em um tipo de Hubble. Podem ser adotados dois caminhos distintos para atingir esse objetivo:

- o primeiro, que chamamos de paramétrico, consiste na modelização do objeto a partir de uma definição *a priori* de quais são as formas esperadas para esses objetos,⁴ no ajuste dos parâmetros destes perfis (por técnicas de minimização de mínimos quadrados, por exemplo), e então na correlação entre os valores ajustados e certas classes de objetos;
- uma segunda opção, chamada de não paramétrica, é a definição de um certo número de medidas estatísticas que podem ser mensuradas diretamente a partir da imagem destes objetos, levando em consideração o fluxo e/ou formato, sendo que tipos distintos de objetos possuiriam valores característicos de tais quantidades permitindo assim a classificação.

Em geral, prefere-se utilizar o primeiro caminho citado anteriormente, como em Barden et al (2005), ou Delgado-Serrano et al (2010). Isso, pois parâmetros dos modelos são em geral diretamente traduzidos em aspectos físicos dos objetos, como por exemplo os tamanhos do bojo e o disco. Contudo, como vimos no Capítulo 2, as imagens reconstruídas para o Gaia não refletem apenas o objeto em estudo, mas introduzem diversos sinais espúrios e artefatos que são provenientes ou do algoritmo adotado para a reconstrução ou da distribuição não homogênea dos ângulos de trânsito do satélite.

Logo, no caso particular do Gaia deve ser preferível utilizar parâmetros que sejam robustos à degradação, a baixa quantidade de informação espacial ($\sim 10^2$ pixels) e que possam ser redefinidos de modo a permitir a adoção de algum tipo de peso pixel-a-pixel. Embora os algoritmos de reconstrução de imagem disponíveis não sejam acompanhados de uma estimativa confiável do erro da reconstrução do sinal, isso é algo que deve ser uma característica de algoritmos futuros.

Iniciamos este Capítulo com uma revisão da literatura comentando e apresentando os parâmetros escolhidos para realizar a análise das imagens reconstruídas do Gaia, e então apresentaremos alguns estudos que realizamos sobre estes parâmetros. Em uma segunda parte do Capítulo apresentaremos o algoritmo de aprendizado computacional escolhido para classificar as imagens reconstruídas, e apresentaremos resultados obtidos com a aplicação deste método em dados reais de galáxias próximas e distantes além de dados simulados com os simuladores oficiais do DPAC.

⁴São definidos perfis, em geral analíticos, para a distribuição de brilho, como o de de Vaucouleurs (de Vaucouleurs, 1948), exponencial ou de Sérsic (Sérsic, 1968).

5.2 O espaço CASGM20

Um conjunto de parâmetros que tem sido utilizado em casos similares ao do Gaia (particularmente na ausência de grande quantidade de informação espacial) é chamado de CASGM20. Estes são essencialmente medidas da concentração central de luz do objeto (C), da assimetria do sinal (A), da suavidade (S), da concentração não necessariamente central (G) e de um momento da imagem ($M20$).

Inicialmente, [Doi et al \(1993\)](#) demonstraram que a concentração central e o brilho superficial médio de uma galáxia podem ser utilizados para distinguir efetivamente entre tipos de Hubble precoce e tardio uma vez que estão correlacionados com estes últimos. Inspirados nestes resultados, [Abraham et al \(1994\)](#) descreveram um sistema automático de classificação morfológica adequado para utilização no estudo de galáxias quando poucos pixels, da ordem de $\sim 10^2$, estão disponíveis (este número de pixels é similar ao que estará disponível para classificar as imagens reconstruídas da missão Gaia). Tal sistema é baseado em um parâmetro que traça tanto a taxa disco-bojo quanto o raio efetivo do bojo. Naquele mesmo trabalho, eles demonstraram que as classificações obtidas utilizando esse sistema são muito menos sensíveis à degradação da resolução espacial da imagem do que pelo sistema de Hubble.

Concentração central - C

O parâmetro utilizado por [Abraham et al \(1994\)](#), denominado C , descreve a concentração central de luz na imagem da galáxia, e é diretamente calculado pela taxa do fluxo entre isofotas internas e externas de raio normalizados $r = \alpha$ (onde α é escolhido, e para aquele artigo foi utilizado o valor 0.3) e $r = 1$, definidas por:

$$C = \frac{\sum \sum_{i,j \in E(\alpha)} I_{i,j}}{\sum \sum_{i,j \in E(1)} I_{i,j}} \quad (5.1)$$

A classificação é então obtida simplesmente especificando o valor de C e de sua incerteza. Para classificação de galáxias também é utilizado um plano construído por este valor C e o valor do brilho superficial médio $\langle \mu \rangle$, calculado pela razão entre a soma do sinal da galáxia e seu número de pixels n_{pix} (definindo A como a região da imagem com pixels acima do ruído—corte em 2σ no artigo):

$$\langle \mu \rangle = \frac{\sum \sum_{i,j \in A} I_{i,j}}{n_{pix}} \quad (5.2)$$

Com este sistema, os autores argumentam que a classificação das imagens de galáxias em até cinco bins distintos pode ser realizada de forma completamente automática quando boas condições de resolução espacial estão disponíveis, e que é sempre possível realizar uma classificação entre os tipos precoce/tardio mesmo em condições de *seeing* degradado. O diagrama da figura 5.1a mostra as regiões no plano $C - \langle \mu \rangle$ ocupadas por galáxias modeladas como superposições de discos de de Vaucouleurs e bojos exponenciais, para diversos valores da relação bojo/disco.

Figura 5.1: Posições no plano de classificação de modelos criados a partir de superposições de discos de de Vaucouleurs e bojos exponenciais. (a) Linhas tracejadas indicam regiões ocupadas por diversas razões bojo/disco. (b) Posições no plano de classificação para vários tipos de Hubble. Os contornos foram traçados em 2σ , com uma isofota limite de $25.5 \text{ mag arcsec}^{-2}$. As galáxias foram classificadas utilizando apenas os valores da relação disco-bojo: Sc entre $[4; +\infty]$, Sab entre $[1.5; 6]$, SO $[0.1; 2]$. (De Abraham et al, 1994)

Na figura 5.1b as posições no plano $C - \langle\mu\rangle$ ocupadas por vários tipos de Hubble estão representadas (descontados quaisquer efeitos de degradação de *seeing* ou amostragem). As regiões de classificações foram determinadas baseadas em valores bojo/disco característicos para cada tipo de galáxia e em contornos em 2σ . Nota-se que as posições ocupadas por galáxias tipo precoce e tardio encontram-se bastante sobrepostas neste plano, tornando a classificação neste plano incerta.

Assimetria - A

Em um segundo trabalho, Abraham et al (1996a) analisaram a morfologia de galáxias distantes provenientes de dados do *Hubble Space Telescope Medium Deep Survey*, ou MDS (Griffiths et al, 1994), e introduziram uma segunda medida A , denominada Assimetria. Deste modo, o plano $C - \langle\mu\rangle$ utilizado anteriormente para realizar a classificação foi substituído pelo plano $C - A$.

As medidas da concentração são realizadas como descrito na subseção anterior, enquanto a assimetria é obtida rotacionando a imagem em 180° a partir de seu centro, subtraindo essa imagem rotacionada da imagem original e tomando a metade da taxa entre o valor absoluto da soma da luz presente nessa imagem subtraída e a soma da luz presente na imagem original (após a subtração do céu). O centro

de rotação é determinado por meio de uma suavização da imagem com um núcleo gaussiano com $\sigma = 1$ pixel e pela determinação do pixel de valor máximo.

As posições no espaço $C - A$ para diversas galáxias, que foram classificadas visualmente por Richard S. Ellis, podem ser observadas na figura 5.2. Qualitativamente, nota-se que ainda que exista uma grande superposição entre tipos morfológicos, existem regiões mais ocupadas por certos tipos de galáxias. Notamos entretanto, que a separação não é clara sem a informação *a priori* de que existem galáxias de diferentes tipos presentes neste gráfico—ou seja, não existe uma separação natural entre as diferentes regiões do plano.

Figura 5.2: Posições no plano $C - A$ (concentração – assimetria) de galáxias do MDS classificadas visualmente por Richard S. Ellis. Elipses denotam galáxias compactas, elípticas e S0s, espirais indicam galáxias espirais, e asteriscos indicam irregulares, peculiares ou *mergers*. As regiões de separação foram criadas a partir da análise de galáxias próximas e bem classificadas visualmente (originadas do catálogo de [Frei et al, 1996](#)) colocadas em mais altos redshifts. (De [Abraham et al, 1996a](#))

O trabalho conclui que a classificação baseada somente na concentração central não é capaz de separar entre os diversos tipos precoces, e que não é possível distinguir tipos intermediários-a-tardios de tardios e irregulares. Contudo, incorporar a assimetria permite melhorar a distinção entre tipos intermediários e irregulares, ainda que exista superposição entre tipos muito tardios/irregulares/*mergers*.

Outros trabalhos, como [Huertas-Company et al \(2007\)](#), incorporaram modificações menores neste sistema, de modo a tornar as medidas de C e de A mais robustas à seleção de brilho superficial, erros de determinação do centro da imagem e dependência com o *redshift*.

Clumpiness - S

O terceiro parâmetro, denominado *clumpiness* ou rugosidade, S ,⁵ foi proposto por Conselice (2003, 2006). Esta medida, representada na Figura 5.3, indica a granulação da imagem, sendo interessante uma vez que galáxias elípticas não devem possuir estruturas em altas frequências espaciais. Isso, pois a distribuição de luz destas galáxias é mais suave, enquanto as galáxias passando por formação estelar possuem alta quantidade de luz sendo emitida em altas frequências espaciais.

Naquele trabalho, S é definido como a taxa entre a luz contida em altas frequências espaciais e a quantidade total de luz na galáxia. Para calcular esse valor, a imagem original do objeto deve ser degradada por meio de uma suavização, criando assim uma imagem cujas estruturas de alta frequência tenham sido removidas.⁶ Então, a imagem original é subtraída da imagem degradada, originando um mapa que contém apenas os componentes responsáveis pelas estruturas de alta-frequência na distribuição espacial de luz da galáxia. O fluxo desta imagem é então somado e dividido pela soma do fluxo da imagem original, resultando na medida de rugosidade:

$$S = 10 \times \sum_{x=1}^N \sum_{y=1}^N \frac{(I_{x,y} - I_{x,y}^{\sigma}) - B_{x,y}}{I_{x,y}} \quad (5.3)$$

na definição acima, $I_{x,y}$ é o fluxo da galáxia na posição (x, y) , $I_{x,y}^{\sigma}$ é o valor da imagem suavizada na posição (x, y) , $B_{x,y}$ é o valor do fundo na posição (x, y) .

Computacionalmente a parte interna da galáxia não é levada em conta, pois segundo Conselice essa região contém potência em altas-frequências que não são apenas relacionadas com distribuição real de luz, mas que são artefatos da amostragem finita—nesta definição, todos os pixels com contagem negativa da imagem subtraída são definidos como zero, antes de se calcular o valor de S .

Naquele mesmo artigo são utilizadas medidas de C e de A obtidas a partir de uma definição diferente daquela adotada por Abraham et al (1994, 1996a):

- Para medir a assimetria, é realizada uma degradação da imagem antes da rotação e da subtração;
- Para mensurar a concentração foi utilizado um método descrito em detalhes em Bershadsky et al (2000), no qual a razão que é levada em conta é aquela entre os raios de 80% e 20% de emissão (segundo Conselice, 2003, a emissão total é dada pela emissão dentro de 1.5 raios petrosianos⁷).

Uma correlação interessante existe entre o valor do parâmetro de clumpiness S , o índice de cor $(B - V)$ e a largura equivalente da linha de $H\alpha$ (esta pode ser observada na Figura 5.4). Essa relação era de certa forma esperada, pois tendo em

⁵Ele é chamado de S , pois C já havia sido utilizado, e ao mesmo tempo que este parâmetro indica a rugosidade ele também indica a Suavidade.

⁶Note que esse tipo de informação também pode ser determinado a partir de coeficientes de uma transformada de *wavelet*.

⁷Ver seção 5.3.

Figura 5.3: Representação das medidas *CAS*. (De Conselice, 2003)

vista a definição de S , galáxias com muitas regiões de formação estelar, logo com valores de S que refletem tal emissão em altas-frequências espaciais, são também galáxias com alta emissão em $H\alpha$ e com valores de $(B - V)$ característicos.

Figura 5.4: Relações entre o valor de S corrigido pela inclinação da galáxia, o índice $(B - V)$ e a largura equivalente da linha $H\alpha$. (De Conselice, 2003)

A dispersão existente nas relações é provavelmente devida ao fato de que S é sensível ao ângulo de visada no qual a galáxia é observada, e pode ser compreendida como uma incerteza adicionada devido ao efeito de projeção do disco, e logo da poeira presente neste disco.

Do ponto de vista observacional, a linha de $H\alpha$ é importante por ser intimamente relacionada com a taxa de formação estelar de um objeto (Kennicutt, 1998), seguindo uma relação linear. Desse modo, uma medida puramente morfológica, como o valor de S , traça uma característica física, que é a taxa de formação estelar.

A partir das medidas de C , A e S , é possível construir os planos $C - A$, $S - A$ e $C - S$, nos quais Conselice realiza classificação. Estes planos estão apresentados na figura 5.5, com galáxias de diversas classes.

Figura 5.5: Planos de classificação. As cores representam o parâmetro que não se encontra no plano representado. No primeiro painel, vermelho representa $S < 0.1$, verde $0.1 < S < 0.35$, azul $S > 0.35$. No painel $S - A$, vermelho representa $C > 4$, verde $3 < C < 4$ e azul $C < 3$. No plano $C - S$, vermelho representa $A < 0.1$, verde $0.1 < A < 0.35$ e azul $A > 0.35$. (De Conselice, 2003)

Coeficiente de Gini - G

Outro parâmetro, que complementa o volume CAS , é o coeficiente de Gini, introduzido na Astronomia por Abraham et al (2003). Este coeficiente é uma medida estatística utilizada em econometria para descrever a desigualdade na distribuição de riqueza de uma dada população. Segundo a descrição naquele artigo, a curva de Lorenz (1905) é construída a partir da proporção cumulativa de riqueza como função da porcentagem da população, de modo que se a sociedade é perfeitamente igualitária essa função é uma reta com coeficiente angular unitário, e em um caso extremo no qual somente uma pequena proporção da população detém quase toda a riqueza, a curva é principalmente plana e próxima de zero até que sobe rapidamente ao chegar perto do final do intervalo $[0;1]$.

Figura 5.6: Interpretação geométrica do coeficiente de Gini. O Coeficiente de Gini corresponde à taxa entre a área A e a área total A+B, com $L(p)$ proporção cumulativa dos recursos e p a proporção cumulativa da população. (De Abraham et al, 2003)

Segundo Abraham et al (2003), uma descrição mais formal deste coeficiente seria: Seja X uma variável positiva aleatória com uma função de distribuição cumulativa $F(x)$, e seja X_i uma amostragem de X , a curva de Lorenz é dada por:

$$L(p) = \frac{1}{\bar{X}} \int_0^p F^{-1}(u) du \quad (5.4)$$

O coeficiente de Gini é definido então como a média da diferença absoluta entre todas as combinações de X_i :

$$G = \frac{1}{2\bar{X}n(n-1)} \sum_{i=1}^n \sum_{j=1}^n |X_i - X_j| \quad (5.5)$$

Segundo Abraham et al (2003), uma forma bastante eficiente para calcular G é primeiro ordenar os X_i em ordem crescente e então realizar uma soma simples:

$$G = \frac{1}{\bar{X}n(n-1)} \sum_{i=1}^n (2i - n - 1)X_i \quad (n > 2) \quad (5.6)$$

Para o cálculo de G aplica-se a equação acima à uma lista ordenada por intensidade dos valores dos pixels da imagem daquele objeto. Os autores notam que em uma primeira aproximação, o valor de G é uma forma generalizada do valor da medida de concentração—de fato, a figura 5.7 mostra que para altos valores de C , a relação é linear com inclinação aproximadamente unitária.

No entanto é comentado que G não pode ser utilizado cegamente como um substituto para C , uma vez que o valor de G calculado em uma imagem se conserva mesmo após a distribuição espacial dos pixels ser completamente alterada, modificando assim a medida de C . Pode-se perceber intuitivamente que G pode ser relacionado à uma concentração de luz independente da posição na imagem: mesmo após reordenar como se queira as posições dos pixels, G não é alterado.

Naquele mesmo trabalho os autores analisaram as correlações entre os valores de C , G e outras grandezas, como o brilho superficial médio, a magnitude absoluta,

Figura 5.7: Relação entre o coeficiente de Gini (G) e a medida de concentração (C) para um conjunto de imagens nas bandas g e i provenientes do SDSS *Early Data Release*. (De Abraham et al, 2003)

a magnitude aparente, o *redshift*, a cor e a razão axial. Dentre todas as grandezas analisadas, é interessante a correlação encontrada com o brilho superficial médio. Projeções da distribuição das galáxias analisadas em um volume $C - G - \langle \mu \rangle$ (figura 5.8) demonstram que estas se encontram em um plano muito bem definido, com dispersão bastante pequena—isto é bastante interessante, pois liga uma quantidade puramente fotométrica à quantidades morfológicas.

Figura 5.8: Distribuição de galáxias no volume $C - G - \langle \mu \rangle$. O plano em vermelho representa o melhor ajuste, e ilustra o quão pouco dispersos encontram-se as galáxias neste volume. (De Abraham et al, 2003)

A partir de simulações, Abraham et al (2003) observaram que a topologia e a orientação da superfície característica na qual as galáxias estão distribuídas no volume $C - G - \langle \mu \rangle$ são reais, mas eles sugerem que o fato dessa superfície ser um plano para as galáxias observadas possa estar associado a uma seleção na distribuição das galáxias observadas.

Momento M20

A última medida que utilizaremos é denominada M20, e foi introduzida por [Lotz et al \(2004\)](#). Esta é o momento de segunda ordem dos pixels brilhantes responsáveis por 20% do fluxo da imagem. No trabalho pioneiro, os autores também comentam que a inclusão de pixels do céu no cálculo de G irá aumentar tal grandeza sistematicamente, enquanto que não incluir os pixels de baixo brilho superficial da galáxia irá diminuí-la sistematicamente, e por isso eles redefinem o cálculo de G .

Para contornar esse problema os autores criam um mapa de segmentação dos pixels da galáxia, sendo que para isso a imagem da galáxia é convoluída com um filtro gaussiano cujo σ é $1/5$ do raio Petrosiano da galáxia. Então o valor do brilho superficial é medido em um certo raio Petrosiano e os pixels na imagem convoluída cujo fluxo é maior do que o fluxo naquele raio são assumidos como pixels da galáxia e utilizados para o cálculo de G . Outra diferença em relação à forma de calcular G é a utilização do valor absoluto de X_i e não de X_i , pois a imagem é subtraída do céu, de modo que podem existir pixels com valores negativos. Os autores comentam que essa correção permite re-obter o valor de G com confiança de 10% para imagens com $S/N > 2$.

A medida de M_{20} , depende do momento total de segunda ordem M_{tot} , que é simplesmente o fluxo em cada pixel multiplicado pela distância quadrada ao centro da galáxia, somado sobre todos os pixels da galáxia. O centro da galáxia aqui é definido como a posição para a qual M_{tot} é minimizado. Então, M_{20} é o momento normalizado de segunda ordem dos pixels até atingir 20% do fluxo total na imagem. Para calcular M_{20} , os pixels da imagem são ordenados por fluxo, são somados até atingir 20% do fluxo total da galáxia e então são normalizados:

$$M_{20} = \log_{10} \left(\frac{\sum_i M_i}{M_{tot}} \right), \text{ enquanto } \sum_i f_i < 0.2 f_{tot} \quad (5.7)$$

com

$$M_{tot} = \sum_i^n M_i = \sum_i^n f_i |(x_i - x_c)^2 + (y_i - y_c)^2| \quad (5.8)$$

[Lotz et al \(2004\)](#) realizaram medidas de C , A , S , G e M_{20} para 170 galáxias próximas (classes E, S0, Sa, Sb, Sc, Sd, dI), 73 ULIRGs (*Ultra Luminous Infrared Galaxies*⁸) e 49 galáxias *Lyman-break*⁹, e a partir destes dados concluíram que as galáxias normais seguem uma sequência $G - M_{20} - C$ bem definida—tipos precoces e sistemas dominados pelo bojo deveriam ocupar a região de alto G , alto C , e baixo M_{20} , enquanto que tipos tardios a região de baixo G , baixo C e altos M_{20} .

Eles ainda concluíram que uma combinação de A e S com G forma um trio eficiente para discriminar entre ULIRGs e tipos de Hubble normais: ULIRGs ocupam uma região no plano $G - M_{20}$ superior àquela ocupada pelas galáxias normais, e

⁸Galáxias extremamente luminosas no infra-vermelho ($\sim 10^{12} L_{\odot}$).

⁹Estas são galáxias em altos *redshifts* ($z > 2.5$) que estão sofrendo um surto formação estelar. Um review sobre estes objetos pode ser encontrado em [Giavalisco \(2002\)](#).

Figura 5.9: M_{20} vs. G calculados para galáxias próximas. Os símbolos indicam o tipo da galáxia: E/S0–círculos, Sa/Sbc–triângulos, Sc/Sd–cruzes, dI– losangos, Espirais edge-on–barras. (De Lotz et al, 2004)

ULIRGs com núcleos múltiplos ocupam regiões ligeiramente distintas dos ULIRGs simples no espaço morfológico (representadas graficamente na Figura 5.10).

Figura 5.10: M_{20} vs. C , e G vs. C , calculados para galáxias normais e ULIRGs. Na maior parte dos casos as galáxias normais se localizam em uma região bem definida, e sugerem um critério multi-dimensional para classificação. (De Lotz et al, 2004)

Esses parâmetros têm utilização recente difundida nas tentativas de classificação morfológica de galáxias com baixo número de pixels a partir de dados puramente morfológicos, tendo sido aplicados por exemplo, em Huertas-Company et al (2010) durante uma análise de galáxias em $z \sim 2$.

5.2.1 Testes com dados da literatura—PCA

Dentro do contexto desta tese, um primeiro estudo que foi realizado com os dados de galáxias no espaço CASGM20 consistiu na Análise de Componentes Principais (*Principal Component Analysis*, ou PCA), sobretudo para verificação da existência de correlações significativas entre estes parâmetros que justifiquem a otimização do processo de classificação das galáxias quando seja necessário analisar uma grande quantidade de imagens.

Idealmente, essa otimização resultaria na possibilidade classificar as galáxias a partir de cortes simples em um plano bi-dimensional, dispensando a utilização de algoritmos complexos de aprendizagem computacional. Para verificar se isso seria possível, realizamos uma análise a partir dos parâmetros CASGM20 calculados por Lotz et al (2004) (L04) nas galáxias do catálogo de Frei et al (1996).

As análises de PCA são baseadas na idéia de reduzir o número de dimensões de um conjunto de dados quando existem correlações entre as medidas. Entretanto, como comentado em Jolliffe (2004), a interpretação dos componentes principais que emergem de combinações das variáveis originais pode levar a conclusões físicas interessantes. Isso, pois caso alguns poucos componentes principais possam reproduzir a maior parte da variância presente nos dados, e caso estes poucos componentes sejam fisicamente interpretáveis, eles podem fornecer uma descrição alternativa e mais simples dos dados do que as variáveis originalmente mensuradas.

Essa redução de dimensão é obtida a partir da transformação das variáveis originais em um novo conjunto de variáveis, chamados de componentes principais (PCs), que por definição não são correlacionados. Tais componentes são ordenados de modo com que os primeiros contenham a maior parte da variância. Em forma matricial, os componentes principais podem ser escritos como:

$$\mathbf{PCs} = \mathbf{A}^T \mathbf{x}^* \quad (5.9)$$

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

onde $\mathbf{A}$ é a matriz dos autovetores da matriz de correlação dos dados e $\mathbf{x}^*$ representa as variáveis padronizadas (escaladas à unidade e centralizadas em zero).

A utilização de variáveis padronizadas é realizada para minimizar a dependência de escala neste tipo de análise. Entretanto cabe notar que a utilização deste tipo de variável torna mais difícil a interpretação física dos componentes obtidos.

Os valores dos componentes principais que calculamos para os dados de L04 estão representados na Figura 5.11. Podemos notar que nos planos PCx vs. PCy com $x, y > 2$, as galáxias dos diversos tipos morfológicos encontram-se dificilmente separáveis, demonstrando que em princípio uma classificação parece ser possibilitada utilizando apenas os primeiros dois componentes principais.

Essa impressão, contudo, deve ser embasada por algum critério quantitativo. Ainda que existam trabalhos dedicados à determinação de critérios e de algoritmos para a seleção do número de componentes principais que devem ser preservados após a análise (muitas vezes por meios computacionalmente intensivos, como *resampling*

Figura 5.11: Após uma mudança de base por análise de componentes principais os parâmetros CASGM20 de L04 foram transformados para os PCs. A cor verde indica galáxias espirais a partir do tipo b/c e irregulares, azul indica galáxias espirais tipos a/b e vermelho indica galáxias elípticas e lenticulares. Foram usados dados de imagens no filtro B.

e *bootstrap*), a seleção sempre acaba sendo um pouco arbitrária, pois não existem critérios de aplicação geral.

Uma das regras de uso mais freqüente e que é descrita por Jolliffe (2004) como tendo um embasamento qualitativo forte, é o corte de acordo com a proporção cumulativa da variância total. Esta regra baseia-se na escolha de um valor de corte para o qual se deseja que a variância dos dados seja explicada pelos componentes principais—algo que se situa entre 80% e 90%, dependendo do autor. Isso significa, que os n componentes principais capazes de explicar até uma certa porcentagem da variância dos dados devem ser mantidos para análises futuras.

Se observarmos os resultados de importância dos componentes principais da análise de CASGM20 de L04, concluímos que o espaço de análise poderia em princípio ser reduzido para os três primeiros componentes (seguindo a regra de corte em 90%). Isso pode ser visualizado pela proporção cumulativa da contribuição à variância total de cada componente principal representado na Tabela 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

Tabela 5.1: Importância dos componentes principais calculados com dados de L04. A linha “Proporção da variância” representa o quanto o n-ésimo componente principal contribui para a variância original dos dados.

Como possuímos a classificação visual para essas galáxias, realizamos o PCA em cada uma dessas classes separadamente, com a finalidade de comparar o resultado obtido e verificar se alguma dessas classes possuía alguma particularidade em relação às outras. Os resultados da proporção cumulativa para cada uma dessas análises pode ser observado na Tabela 5.2 a seguir.

	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

Tabela 5.2: Proporções cumulativas dos PCs de L04 obtidas em PCAs separadas por grupos morfológicos.

Pode-se observar na Tabela 5.2 que para as duas primeiras classes podemos realizar a redução dimensional para três PCs (baseando-se na regra do corte em 90% da proporção cumulativa). Contudo, para a terceira classe, que engloba as galáxias tipo Sb/c/Irr, podemos apenas reduzir para quatro dimensões. Do ponto de vista físico, a necessidade de cada vez mais parâmetros para explicar a variância conforme se avança na sequência de Hubble demonstra que estes objetos se tornam cada vez mais complexos do ponto de vista morfológico, e que de algum modo os parâmetros escolhidos estão refletindo este aumento de complexidade.

Apesar da regra empírica do corte em 90% da proporção cumulativa possibilitar o uso de apenas três componentes principais, decidimos utilizar todos os cinco parâmetros na análise de galáxias do Gaia. Essa decisão foi tomada levando em conta a possibilidade de degradação em possíveis resultados de galáxias Sb/c/Irr (que exigiriam quatro componentes), além de que a confiança nos parâmetros CASGM20 mensurados a partir de imagens reconstruídas com os dados do Gaia ainda não pode ser determinada de forma clara neste momento.

5.3 Implementação para o Gaia – CASGM20

Como vimos na introdução deste Capítulo, as imagens reconstruídas do Gaia não refletem apenas o objeto em estudo, mas introduzem diversos sinais espúrios e artefatos da reconstrução. Desse modo, desde o início dos trabalhos da *Development Unit* 470 em 2007, que como vimos na introdução desta tese é responsável pelo tratamento de objetos extensos no Gaia, propusemos a utilização das medidas estatísticas comentadas nas seções anteriores medidas a partir das imagens reconstruídas, aliadas a técnicas de aprendizado computacional para realização de uma classificação puramente morfológica das galáxias não resolvidas que seriam observadas pelo Gaia.

Além de propormos essa redução, também implementamos códigos em Java voltados ao cálculo de todos os parâmetros CASGM20 no pipeline da *Coordination Unit* 4 (descritos em Krone-Martins & Ducourant, 2008). Esses parâmetros foram escolhidos devido a baixa necessidade de informação espacial, robustez à degradação e facilidade em serem modificados para inclusão futura de mapas de pesos, possibilitando utilização mais confiável das imagens reconstruídas.

Para essa implementação, assumimos que o objeto em estudo seria o único objeto presente na imagem reconstruída, e que todas as outras contribuições (tal como raios cósmicos, outros objetos na linha de visada, etc.) foram levadas em conta durante procedimentos anteriores—o que é razoável supor, pois como a imagem é reconstruída a partir de diversas observações, os raios cósmicos devem ter sido excluídos, e como estas observações já foram analisadas previamente pelo *Source Environment Analysis* descrito no Capítulo 3, fontes secundárias já devem ter sido detectadas preliminarmente.

Contudo, não assumimos que todo o sinal dos pixels sejam devidos ao objeto, e um tratamento de fundo de céu pode ou não ser aplicado (em geral, ele é aplicado). O fundo é determinado a partir de uma rejeição iterativa a 3σ , e de uma binarização da imagem na qual os pixels acima de um certo nível de corte (configurável) são considerados como pertencentes ao objeto. Neste mapa de fundo podem ser incluídos os pixels pertencentes a quaisquer objetos ou sinais adicionais que estejam presentes na imagem e que não devam ser levados em conta na análise.

As definições adotadas para os parâmetros que são mensurados pela nossa implementação são resumidas nas seguintes equações:

$$\left(\begin{array}{ll} R_P(r) = \frac{\int_{0.8r}^{1.25r} dr' 2\pi r' I(r') / [\pi(1.25^2 - 0.8^2)r^2]}{\int_0^r dr' 2\pi r' I(r') / (\pi r^2)} & C = 5 \log_{10} \left(\frac{r_{80}}{r_{20}} \right) \\ M_{20} = \log_{10} \left(\frac{\sum_l M_l}{M_{tot}} \right), \text{ for } \sum_l I_l \leq 0.2 I_{tot} & A = \frac{\sum_i \sum_j |I(i,j) - I_R(i,j)|}{\sum_i \sum_j I(i,j)} \\ G = \frac{\sum_k [2k - N - 1] I_k}{\langle I \rangle N(N-1)} & S = \frac{\sum_i \sum_j |I(i,j) - I_S(i,j)|}{\sum_i \sum_j I(i,j)} \end{array} \right) \quad (5.10)$$

Nestas equações, $R_P(r)$ define a taxa petrosiana, tal como utilizada no Sloan Digital Sky Survey (Blanton et al, 2001) e o raio petrosiano é adotado como o Raio r para qual $R_P(r) = 0.2$. $I(r)$ é o perfil de brilho superficial médio, r_{80} e r_{20} são

os raios que englobam 80% e 20% do fluxo total da galáxia (definido como o fluxo dentro de 1.5 raios petrosianos). M_l é o momento de segunda ordem do l -ésimo pixel ordenado por fluxo decrescente. M_{tot} é o momento de segunda ordem calculado para todos os pixels. I_l o fluxo do l -ésimo pixel. $I(i, j)$ o fluxo da imagem no pixel (i, j) . $I_R(i, j)$ o fluxo da imagem no pixel (i, j) após rotação de 180° . I_k o fluxo do k -ésimo pixel ordenado por fluxo crescente. N o número total de pixels na imagem. $I_S(i, j)$ o fluxo da imagem convoluída por um filtro tipo *boxcar*¹⁰ com tamanho de $0.3R_P$ no pixel (i, j) . Na Figura 5.12, a seguir, podemos observar o exemplo do cálculo do parâmetro S para a galáxia NGC3351.

Figura 5.12: Exemplo do cálculo do parâmetro S para a galáxia NGC 3351 originada do catálogo de Frei. **Esquerda:** imagem original I , do catálogo de Frei et al (1996). **Centro:** imagem convoluída I_S . **Direita:** imagem resultante do procedimento de subtração, a partir da qual S é calculado. A última imagem não compartilha do mesmo mapa de cores que I e I_S .

Os cálculos destes parâmetros são realizados por classes Java completamente independentes (denominadas *Measurers*), de modo que todo processamento poderia, em princípio, ser realizado de forma completamente paralela. Contudo, como o cálculo é bastante rápido devido à pequena quantidade de pixels disponíveis nas imagens reconstruídas, na implementação atual estes parâmetros são calculados de forma serial. Seus cálculos resultam em um objeto que armazena os valores de CASGM20, que pode ser guardado diretamente no banco de dados do centro de processamento em que o código será executado.

Cabe comentar aqui, que no futuro, as imagens reconstruídas do Gaia poderão ser acompanhadas de algum tipo de erro estimado pelo algoritmo de reconstrução para cada um de seus pixels. Isso significa que os parâmetros CASGM20 definidos acima poderão ser adaptados para tomar proveito dessa informação adicional. A forma mais direta de fazer isso, e aquela que pretendemos utilizar, será realizar algum tipo de ponderação durante o cálculo de tais parâmetros.

¹⁰Um filtro que é definido por uma função retangular.

5.3.1 Testes com galáxias do catálogo de Frei

Um dos testes que realizamos com a implementação dos códigos que desenvolvemos para o Gaia, e um dos mais necessários para assegurar que estes parâmetros podem ser utilizados para análise de imagens reconstruídas como esperadas com os dados deste satélite, é a verificação da robustez do volume CASGM20 frente a quantidade de pixels disponíveis nas imagens reconstruídas.

Para isso, calculamos os valores de CASGM20 para 82 galáxias próximas representantes dos três tipos morfológicos principais (elípticas, espirais e irregulares), originadas do catálogo de [Frei et al \(1996\)](#). Essas imagens foram re-amostradas a partir da integração de *splines* bi-cúbicas calculadas na imagem original, para diversas resoluções (em termos de quantidade de pixels) menores do que a original (312x312), sendo que o tamanho das imagens re-amostradas foi variado entre 81x81 pixels (esperado para o caso ideal de uma reconstrução de imagem do Gaia), até pequenas imagens formadas por apenas 11x11 pixels.

Os raios petrosianos e magnitudes dos objetos calculados nas imagens resultantes variam entre ~ 46.5 mas até $\sim 1.5''$ (considerando 30 mas/pixel), e de 11.40 até 19.98 mag, respectivamente, o que seriam casos comparáveis aqueles que devem ser observados pelo Gaia. Um exemplo das imagens utilizadas para este estudo, no caso da galáxia NGC 3893, pode ser visto na Figura 5.13.

Figura 5.13: Exemplos de re-amostragens de uma imagem da galáxia NGC 3893 que foram utilizadas para o estudo da variação dos parâmetros CASGM20.

Os parâmetros CASGM20 foram calculados para todas estas imagens, sendo que definimos uma grandeza para quantificar a variação dos valores calculados para estes parâmetros em função do tamanho da imagem. Esta grandeza é a média das variações percentuais para cada um dos parâmetros que é calculado para cada galáxia individual, ou seja, para N galáxias, e para o parâmetro C , ela pode ser escrita como:

$$\text{MPV}(C) = \frac{1}{N} \sum_1^N \frac{C(\text{Frei image}) - C(\text{subsampling image})}{C(\text{Frei image})} \quad (5.11)$$

Uma comparação entre os valores obtidos do catálogo original de Frei, e aqueles calculados a partir das imagens re-amostradas utilizando a média da variação percentual pode ser vista na Figura 5.14. Neste gráfico, notamos que o parâmetro S , diverge rapidamente de seu valor original—isso ocorre, pois quanto menos pixels estão disponíveis maior é o corte nas altas frequências espaciais, e esse parâmetro mede justamente a rugosidade que deve ser encontrada em altas frequências espaciais.

Entretanto a comparação de todos os parâmetros demonstra que o espaço CASGM20 como um todo, é robusto à ausência de informação, podendo ser de fato utilizado em imagens compostas por poucos pixels, tais como aquelas esperadas para a missão Gaia. Podemos notar que a variação percentual média para a maior parte dos parâmetros encontra-se dentro de um nível de 20%, se comportando de forma estável até mesmo para imagens pequenas.

Figura 5.14: Variação percentual média dos parâmetros CASGM20 em relação à dimensão da imagem utilizada para o cálculo do parâmetro. Os símbolos estão levemente deslocados entre si para facilitar a visualização. Além disso, barras de erro de 1σ foram graficadas apenas para o parâmetro C , uma vez que elas são representativas de todas as outras.

Estes resultados (originalmente apresentados em [Krone-Martins et al, 2008b](#)) demonstram que os parâmetros CASGM20 podem ser aplicados à análise de imagens reconstruídas com sucesso, uma vez que estas imagens sofrerão principalmente pelo baixo número de pixels disponível para a análise (apesar de possuírem grande resolução angular em comparação aos telescópios terrestres).

Contudo, embora o teste com as galáxias próximas re-amostradas demonstre que os parâmetros são robustos mesmo em imagens com amostragens pequenas, efeitos associados a observação de galáxias intrinsicamente pequenas do ponto de vista angular ainda não podem ser percebidos. Para isso um teste adicional que verifique que mesmo em galáxias que sejam observadas como sendo intrinsicamente pequenas do ponto de vista angular, tipos morfológicos distintos ocupam regiões distintas do espaço CASGM20, é necessário.

5.3.2 Cálculo de CASGM20 em imagens do *Hubble Deep Field*

O *Hubble Deep Field North* (HDF) consiste em uma região do céu que foi imageada pelo telescópio espacial Hubble em longas exposições (~ 120.000 s de tempo integrado por filtro), cuja descrição pode ser encontrada em [Williams et al \(1996\)](#)—existe também um campo equivalente realizado para o hemisfério Sul.

As galáxias observadas neste campo não serão observadas pelo Gaia por uma questão de limite de magnitude: a maior parte dos objetos ali presentes possuem $G \gg 20$. Entretanto, o Gaia deve observar um grande número de galáxias menores e em menores redshifts, e que devem possuir a mesma “quantidade de informação morfológica” que encontra-se presente nas galáxias do HDF. Desse modo, o HDF pode ser considerado um bom caso de teste para aplicarmos os códigos que estamos desenvolvendo para o Gaia em galáxias angularmente pequenas.

Neste teste, o objetivo é verificar como os parâmetros CASGM20 calculados com nosso código quando aplicado em imagens do HDF, e principalmente se estes parâmetros permitem uma posterior classificação morfológica. As galáxias utilizadas para o teste possuíam classificação visual preliminar de [van den Bergh et al \(1996\)](#).

As imagens das galáxias individuais que utilizamos para os cálculos de CASGM20 (chamadas de “selos”) foram construídas a partir da soma das imagens obtidas com os diferentes filtros utilizados no HDF-N.¹¹ Após a construção destas “imagens brancas”, os objetos presentes nas quatro diferentes CCDs foram detectados e recortados em “selos”. Como as posições das galáxias no artigo [van den Bergh et al \(1996\)](#) foram mensuradas na primeira redução do HDF, enquanto nós trabalhamos com sua redução final, foi necessário realizar um procedimento de identificação cruzada entre os objetos classificados e os objetos detectados.¹² Alguns exemplos de “selos” obtidos por esse procedimento para os três diferentes tipos de galáxias podem ser encontrados na Figura 5.15.

¹¹Os filtros F300, F450, F606 e F814 cobrem aproximadamente o intervalo espectral do Gaia.

¹²A detecção dos objetos nas imagens originais, a identificação cruzada e os recortes em selos foram realizados por J.-F. le Campion, do *Laboratoire d’Astrophysique de Bordeaux*.

Figura 5.15: Imagens de galáxias recortadas do Hubble *Deep Field North*.

Os parâmetros CASGM20 calculados para os diversos tipos de galáxias podem ser vistos na Figura 5.16. Podemos verificar que é possível separar regiões dos diversos planos do espaço CASGM20 onde certos tipos de galáxias são predominantes. Entretanto, a superposição entre os diferentes tipos também é bastante grande, principalmente devido às galáxias espirais ocuparem tanto a região de galáxias elípticas quanto aquela de galáxias irregulares.

Figura 5.16: Parâmetros CASGM20 calculados para as galáxias do HDF-N. A cor verde indica galáxias irregulares e peculiares, azul indica galáxias espirais e vermelho indica galáxias elípticas.

Essa grande superposição pode indicar que durante a identificação cruzada entre a lista de objetos classificados visualmente e a lista de objetos recortados das imagens do HDF alguns erros podem ter ocorrido, ou que no caso destas galáxias angularmente pequenas, se torna de fato muito incerta a distinção entre os diversos tipos morfológicos, pois as interfaces entre os diferentes tipos se torna pior definida (fato que pode ser observado ao examinar os exemplos da Figura 5.15).

5.3.2.1 Análise de Componentes Principais - PCA

Com a finalidade de verificar se seria possível simplificar o espaço de classificação, uma PCA foi aplicada aos dados CASGM20 calculados nas imagens das galáxias do HDF. Os valores obtidos para os diferentes PCs pode ser visto na Figura 5.17.

Figura 5.17: Dados CASGM20 calculados para um subconjunto das galáxias do HDF no espaço dos componentes principais. Código de cores como na Figura 5.16.

Um ponto interessante que pode ser observado no plano dos dois primeiros componentes principais com estes resultados, é a aparente formação de uma seqüência que caminha de elípticas para irregulares, com os objetos classificados como irregulares e peculiares (representados em verde) iniciando um outro eixo, mas intersectando a seqüência na região das galáxias espirais. Esse comportamento não havia sido observado durante a análise de PCA de galáxias próximas, mas é muito complicado

comparar estas análises uma vez que os PCs das duas análises não possuem o mesmo significado, pois eles são formados por combinações lineares diferentes entre os parâmetros. Como comentado em Jolliffe (2004), a interpretação física dos PCs e a comparação entre PCs originados de análises distintas é uma tarefa bastante complexa, configurando um trabalho independente.

	PC1	PC2	PC3	PC4	PC5
Desvio padrão	1.457	1.106	0.859	0.791	0.539
Proporção da variância	0.424	0.245	0.147	0.125	0.058
Proporção cumulativa	0.424	0.669	0.817	0.942	1.000

Tabela 5.3: Importância dos componentes principais calculados a partir de parâmetros CASGM20 medidos em galáxias do HDF-N.

De um ponto de vista qualitativo, a partir da Figura 5.17 temos a impressão de que uma boa classificação poderia ser obtida considerando apenas os primeiros dois, ou três componentes principais. Contudo, a proporção cumulativa dos diferentes PCs (representada na Tabela 5.3) demonstra que seria necessário manter até o quarto componente para que a variância fosse explicada em mais de 90%. Deste modo, a simplificação de apenas um parâmetro não justificaria a utilização de uma PCA sob os parâmetros CASGM20 no caso da classificação das galáxias do HDF.

5.3.3 Testes com simulações

Finalmente, um teste final é a verificação do comportamento desses parâmetros quando mensurados a partir de imagens reconstruídas por algum dos algoritmos de reconstrução de imagem propostos até o momento e descritos no Capítulo 2. Para isso realizamos simulações de observações de galáxias com o GIBIS 7 e utilizamos o código de reconstrução de imagens que implementa o algoritmo *ShuffleStack*.

As coordenadas escolhidas para a simulação são correspondentes a uma região com boa cobertura angular, $(l, b) = (120^\circ; 45^\circ)$ ¹³, gerando um máximo de 82 trânsitos individuais.¹⁴ As galáxias simuladas varrem um largo intervalo de parâmetros, com magnitudes entre [13; 20], semi-eixo maior entre [200; 1200] mas e razão axial entre [1;10], gerando 880 galáxias para cada tipo morfológico simulado (elípticas, espirais e irregulares). Um exemplo com as galáxias simuladas em uma parte do plano focal pode ser visto na Figura 5.18.

Essas simulações foram utilizadas para reconstruir imagens. Para isso foi adotado o algoritmo *ShuffleStack*, com uma escala de 30mas/pixel. Os códigos de medidas de

¹³Ver Capítulo 2 para um mapa de cobertura do céu inteiro.

¹⁴O número exato de trânsitos foi 82, contudo nem todas as galáxias são observadas em todos os trânsitos devido a especificidades dos sistemas de leitura de simulações GIBIS: objetos que em determinado trânsito ocupem colunas de CCDs que não interceptam o centro da simulação não podem ser ter seus dados lidos pelos códigos de conversão de simulações GIBIS em tabelas FITS.

Figura 5.18: Exemplo de galáxias simuladas no plano focal do Gaia. Os objetos aqui representados são galáxias espirais cobrindo uma larga faixa de magnitudes.

parâmetros CASGM20 foram utilizados para medir estas imagens reconstruídas, e os resultados obtidos podem ser vistos na Figura 5.19.

Figura 5.19: Dados CASGM20 calculados em imagens reconstruídas para ~ 2.6 mil galáxias de tipos elíptico (vermelho), espiral (azul) e irregular (verde). As observações foram simuladas pelo GIBIS 7.1 com o uso do MAGIL, e a reconstrução utilizou o algoritmo *ShuffleStack*.

A primeira característica notável na Figura 5.19 mostra que tipos morfológicos distintos ocupam regiões distintas nos diversos planos representados, o que deve possibilitar uma classificação entre estes tipos morfológicos, mesmo que o recobrimento entre estes tipos seja aparentemente grande.

De forma distinta do HDF, a interpretação física dos valores de CASGM20 no caso de seu cálculo em imagens reconstruídas a partir das simulações GIBIS, e em especial de imagens reconstruídas pelo algoritmo *ShuffleStack* (que não foi criado

pensando em objetos extensos) é algo não trivial. Isso pode ser notado ao comparar os resultados obtidos nesta seção àqueles obtidos a partir de medidas de galáxias do HDF apresentadas na seção anterior: os tipos morfológicos distintos ocupam posições diferentes daqueles ocupados pelos dados do HDF. Isso é um efeito da reconstrução imperfeita de imagem (exemplos de reconstruções podem ser vistos na Figura 5.20). Contudo, estes resultados demonstram que é possível utilizar as medidas CASGM20 para realização de classificações morfológicas de galáxias com os dados do Gaia.

Figura 5.20: Exemplos de reconstruções de galáxias irregulares, espirais e elípticas (a cada linha) com a utilização do algoritmo *ShuffleStack*. A amostragem escolhida para reconstrução é de 30mas/pixel, e os objetos possuem $G \sim 17$.

Uma análise de PCA, realizada com o intuito de verificar se seria possível adotar um sistema de classificação simples, baseado em cortes em um plano ou em um espaço tri-dimensional, apresentou resultados semelhantes aos obtidos anteriormente (apenas o último componente poderia ser excluído), de modo que algoritmos de classificação mais elaborados do que cortes simples devem ser aplicados no volume CASGM20 para obtenção de estimativas quantitativas das taxas de sucesso que poderão ser obtidas na classificação morfológica a partir dos dados do Gaia.

5.4 Classificação e Support Vector Machines

Uma vez que verificamos que os parâmetros CASGM20 podem ser utilizados para classificação morfológica de galáxias, pois diferentes classes de galáxias ocupam posições distintas (ainda que sobrepostas) naquele espaço, é necessário escolher um método para identificar as regiões características das diferentes classes automaticamente a partir de exemplos, possibilitando uma classificação posterior de quaisquer objetos cuja classe seja desconhecida.

Na literatura de classificação este problema é conhecido como um problema de aprendizado supervisionado (*supervised learning*), pois exemplos são fornecidos ao sistema. Do ponto de vista matemático, ele se pode ser visto como a determinação de funções $D : X \rightarrow Y$, chamadas de funções de decisão da classificação a partir de um conjunto de treinamento $\{(x_1, y_1), \dots, (x_n, y_n)\}$. O método de k -NN, visto no Capítulo 3, é um exemplo de algoritmo de aprendizado supervisionado.

Existem diversos métodos construídos para resolver este problema, cada qual com vantagens e desvantagens. Alguns exemplos são o próprio k -NN (que é o conceitualmente mais simples), as Redes Neurais Artificiais, as *Support Vector Machines* (SVMs), e as *Relevance Vector Machines* (RVMs).¹⁵

Segundo Graf & Wichmann (2004) as SVMs e as RVMs são os algoritmos que melhor reproduzem os resultados obtidos por nosso cérebro.¹⁶ Isso, pois naquele estudo a correlação entre objetos classificados corretamente e incorretamente entre o cérebro humano e algoritmos de aprendizado computacional foi analisada, e estes são os métodos que apresentaram maior correlação ($\sim 70\%$).

Para a classificação morfológica de galáxias a partir de imagens reconstruídas do Gaia, decidimos utilizar as SVMs por três características deste tipo de classificador que são muito interessantes: ausência de mínimos locais, robustez aos *outliers* e maximização da capacidade de generalização.

Especificamente, as SVMs são treinadas com a finalidade de encontrar a função de decisão que maximiza a capacidade de generalização. Isso é realizado mapeando o espaço original l -dimensional de observações em um hiper-espaço m -dimensional (com $m > l$), no qual o hiper-plano que separa as duas classes é determinado por meio da solução de um problema de programação quadrática—e logo, fica garantida a existência de uma solução global ótima.

Além disso, o hiper-plano separatriz é calculado de modo a maximizar sua distância em relação aos dados (margem), e logo a capacidade de generalização da SVM. Finalmente, a robustez aos *outliers* é fornecida por um parâmetro de margem, que balanceia uma relação entre a maximização da margem e a minimização do erro de classificação. Para detalhes sobre as SVMs, bem como deduções rigorosas das equações que apresentaremos nesta seção, ver Vapnik (2000) e Abe (2005).

¹⁵A RVM, ou *Relevance Vector Machine*, é um novo método de aprendizado computacional introduzido por Tipping (2001), que basicamente introduz um formalismo bayesiano durante a construção de uma SVM, e que não foi utilizado em Astronomia até o momento.

¹⁶Observe que isso não significa que o nosso cérebro funcione como estes algoritmos, mas que eles dão resultados mais ou menos similares para um mesmo problema de classificação.

A equação para o hiperplano que separa duas classes, pode ser escrita na forma:

$$D(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \quad (5.12)$$

onde $\mathbf{w}$ é um coeficiente vetorial e b é um coeficiente escalar. Estes devem ser determinados a partir dos dados do conjunto de treinamento $\{(x_1, y_1), \dots, (x_l, y_l)\}$, e y_i pode assumir os valores “1” ou “-1”, dependendo se x_i pertence a classe 1 ou 2.

Como a função 5.12 deve tomar valores positivos para $y_i = 1$ e negativos para $y_i = -1$, podemos escrever :

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 \quad (5.13)$$

Logo, $D(x) = c$, com $-1 < c < 1$ forma o hiper-plano de separação procurado, sendo que o valor de c (chamado de margem) especifica a distância entre o hiper-plano separatriz e as margens dos dados. Esquematicamente, uma representação bi-dimensional de um hiper-plano ótimo junto com seus vetores de suporte (representados por cores mais fortes), pode ser vista na figura 5.21.

Figura 5.21: Representação do hiperplano ótimo que separa os dados de duas classes. Os *Support Vectors* são os vetores cruzados pelas duas linhas pontilhadas.

Para a determinação dos parâmetros do hiperplano ótimo, uma *Support Vector Machine* resolve basicamente o seguinte problema de otimização:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ \text{sujeito a} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 \\ & i = 1, \dots, l. \end{aligned} \quad (5.14)$$

Entretanto, o problema expressado acima não é ideal no caso de dados inseparáveis linearmente. Uma melhor forma de tratar dados que originam o problema de separação no caso de dados inseparáveis linearmente, é incluir variáveis de penalização

no problema. Essas variáveis podem ser imaginadas como representações da distância entre os planos gerados pelos *Support Vectors* (as margens) e os dados que se encontrem além destes. Um exemplo destas variáveis, chamadas de ξ , pode ser visto na Figura 5.22.

Figura 5.22: Representação do hiperplano ótimo que separa os dados de duas classes mostrando os vetores que penalizam a classificação e os parâmetros ξ .

Este tipo de *Support Vector Machine* é chamado de C-SVM, e é expressado pelo seguinte problema de otimização:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \xi_i \\ \text{sujeito a} \quad & y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \dots, l. \end{aligned} \tag{5.15}$$

No problema expresso acima, C é uma variável que controla o balanço entre a maximização da margem e a minimização do erro de classificação, e sua escolha se configura um problema por si mesmo.

Contudo, mesmo com a adição das variáveis ξ o problema continua sendo inseparável linearmente, sendo que estas variáveis apenas o tornam melhor comportado por meio da adição de uma penalização em alguns pontos de dados que passam a contribuir aumentando o valor de uma função que deve ser minimizada. Para possibilitar a separação de dados que são linearmente inseparáveis no espaço original, é utilizado o “truque do núcleo”.

O truque do Núcleo

Para melhorar a capacidade de uma *Support Vector Machine* generalizar dados que não são linearmente separáveis, aplica-se uma transformação aos dados originais de

modo a levar os dados a um espaço no qual eles sejam separáveis por hiperplanos. Isto é conhecido como “truque do núcleo” (ou *Kernel Trick*).

Figura 5.23: Representação do truque do núcleo. Dados inicialmente separáveis apenas por curvas complexas passam a ser separáveis por hiperplanos em um outro espaço após a aplicação de uma transformação.

O núcleo escolhido para a transformação pode ser de diversos tipos: linear, polinomial, radial, *spline*, etc. Neste trabalho utilizamos o núcleo chamado de *Radial Basis Function*, devido às suas características de impulso localizado. Esta função pode ser escrita como:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (5.16)$$

onde γ é um parâmetro que controla o raio do núcleo.

Então, o trabalho se resume a encontrar a solução para o problema:

$$\min_{\mathbf{w}, b} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \xi_i \quad (5.17)$$

$$\text{sujeito a} \quad y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, i = 1, \dots, l.$$

O problema de otimização 5.17 não é em geral resolvido, e o que se faz é encontrar a solução de seu problema dual (ver Abe, 2005). Este é um problema equivalente ao problema primal (as equações 5.17), mas com equações que são expressadas por multiplicadores de Lagrange. Então, a equação que é realmente minimizada é:

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2} \boldsymbol{\alpha}^T Q \boldsymbol{\alpha} - \mathbf{e}^T \boldsymbol{\alpha} \quad (5.18)$$

$$\text{sujeito a} \quad \mathbf{y}^T \boldsymbol{\alpha} = 0, \\ 0 \leq \alpha_i \leq C, i = 1, \dots, l.$$

onde α_i são multiplicadores de Lagrange não-negativos, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_l)^T$, $\mathbf{e}$ é um vetor composto por valores unitários, Q é uma matriz formada por elementos $Q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$ e $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ é o núcleo.

Finalmente, a função de decisão para um certo dado $\mathbf{x}$ é dada por:

$$\mathbf{x} \in \begin{cases} \text{Classe 1} & \text{se } D(\mathbf{x}) > 0, \\ \text{Classe 2} & \text{se } D(\mathbf{x}) < 0. \end{cases} \quad (5.19)$$

onde

$$D(\mathbf{x}) = \sum_{i=1}^l \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \quad (5.20)$$

Separando diversas classes

Uma vez que se tenha resolvido o problema para a separação de duas classes, existem diversas maneiras de se aplicar o mesmo método para separar um número arbitrário de outras classes. O código de *Support Vector Machines* que implementamos no sistema de classificação morfológica do DU470 utiliza o método conhecido como “um-contra-um” (*one-against-one*, ou *pairwise*).

Neste tipo de estratégia, para um problema de classificação com k classes diferentes, são construídas $k(k-1)/2$ SVMs, que são treinadas para classificar amostras de uma classe contra outra classe, duas a duas. Um exemplo para três classes pode ser visto na Figura 5.24.

Figura 5.24: Classificação em três classes mostrando a região não classificável ao centro em marrom.

Para classificar um certo ponto de dados, uma estratégia de “votação” é aplicada na qual todas as funções de decisão são calculadas neste ponto, e a classe com maior quantidade de votos é considerada como sendo a classe do ponto. Cabe comentar que neste tipo de formulação podem existir regiões do espaço inclassificáveis, nas quais podem ocorrer empates—e neste caso, a classe escolhida para um ponto nesta região é tomada arbitrariamente como sendo aquela com menor índice.

5.5 Implementação para o Gaia – SVM

Para o desenvolvimento dos códigos do DU470, utilizamos uma biblioteca com implementações em Java de diversos algoritmos de *Support Vector Machines* chamada LIBSVM (Chang & Lin, 2001). Como já comentado, a SVM escolhida para aplicação neste problema é uma C-SVM com um núcleo tipo *Radial Basis Function*.

No entanto, a implementação do sistema de classificação morfológica do DU470 não será baseada exclusivamente na classificação obtida pela *Support Vector Machine*, mas sim a partir de um laço no qual a SVM é um dos componentes. Uma representação conceitual deste processo pode ser vista na Figura 5.25.

Figura 5.25: Diagrama de atividade conceitual dos processos envolvidos no treinamento da SVM.

Antes que se possa classificar qualquer objeto, é necessário treinar a SVM. Como ocorrerá exatamente a seleção dos objetos que serão utilizados para o treinamento inicial ainda é algo a ser definido. Uma possibilidade é utilizar objetos que possuam classificação previamente conhecida, outra é ignorar a classificação e realizar uma análise de perfil com diversos perfis teóricos selecionando os objetos para os quais perfis correspondentes a certos tipos morfológicos melhor se ajustarem.

Uma vez que uma lista de galáxias com tipo morfológico conhecido esteja disponível, a SVM é treinada pela primeira vez, e é então aplicada nas imagens reconstruídas de uma série de outras galáxias com tipo desconhecido, fornecendo uma classificação morfológica para estes objetos. Após a classificação morfológica, é executada uma análise de perfil (que será vista com maiores detalhes no Capítulo 6).

Durante essa análise, um perfil teórico é selecionado a partir da classificação obtida pela SVM e ajustado aos dados do satélite. Caso o ajuste seja adequado, a classificação morfológica é confirmada, caso contrário outros perfis podem vir a ser testados, e a classificação morfológica pode ser alterada. Finalmente, após todas as galáxias desta lista secundária serem analisadas, os objetos que tiveram os melhores ajustes de perfis são selecionados, e utilizados para treinar a SVM para os próximos ciclos de processamento da missão.

Para os objetos que não participarão do treinamento da SVM, ou nos ciclos nos quais o treinamento da SVM não é executado, o que ocorre é um processamento baseado em um encadeamento simples de análises. Primeiro ocorre uma reconstrução de imagens, então os parâmetros CASGM20 são mensurados, a classificação por meio da SVM é realizada e finalmente a análise de perfil é executada. Uma representação deste encadeamento pode ser encontrada na Figura 5.26.

Figura 5.26: Diagrama de atividade conceitual dos processos envolvidos na redução dos dados do DU470.

5.5.1 Validação com gaussianas

Um primeiro teste que deve ser realizado é verificar a capacidade deste código em classificar corretamente pontos amostrados de distribuições de probabilidade bem conhecidas, cujas taxas máximas de sucesso pudessem ser estimadas de forma completamente independente da SVM. Uma PDF (*Probability Distribution Function*) bem conhecida que pode ser aplicada neste tipo de estudo é a função gaussiana.

Dado um certo número de funções de densidade de probabilidades gaussianas de duas variáveis, gostaríamos de obter a melhor separação possível entre essas populações que qualquer classificador poderia obter se dados aleatórios são gerados a partir destas duas distribuições. Após obter esta estimativa, vamos comparar a taxa de sucesso na classificação obtida por uma SVM e o cálculo teórico para a melhor taxa possível.

O caso de duas gaussianas

A melhor classificação possível é aquela que determinará a função de decisão no local do espaço onde as duas gaussianas bi-dimensionais se interceptam. Isto, pois todos os dados depois (ou antes) deste local serão classificados erroneamente. Deste modo, a melhor taxa de classificação possível é obtida facilmente (e de fato, veremos que este problema pode ser linearmente separável), uma vez que é suficiente integrarmos as duas PDFs na região definida até a função de decisão. Então, basta somar esses dois valores, e dividir pela soma das integrais das PDFs no espaço inteiro.

Infelizmente esse problema não possui uma solução analítica, uma vez que a primitiva da gaussiana não pode ser obtida. Entretanto, podemos sempre obter uma solução numérica. Em nosso caso, vamos analisar a solução para três separações distintas entre as médias das duas gaussianas: 0.5, 1.5 e 3.0 desvios padrão.

De modo a tornar a análise ainda mais simples, tomamos uma das gaussianas como possuindo média $(\mu_x, \mu_y) = (0, 0)$ e a outra com média $\mu_y = 0$. Podemos fazer isso sem perda de generalidade, pois para duas gaussianas é sempre possível encontrar uma rotação que permite a redução a este caso para quaisquer configurações das PDFs. Consideramos também, que os desvios padrão são os mesmos para ambas as populações.

Matematicamente, o problema toma a seguinte forma: dadas duas funções gaussianas A e B, primeiro precisamos encontrar a solução para a seguinte equação (que nos fornece a função de classificação):

$$\frac{1}{2\pi\sigma_x^A\sigma_y^A}e^{-\left[\frac{(x-\mu_x^A)^2}{2(\sigma_x^A)^2}+\frac{(y-\mu_y^A)^2}{2(\sigma_y^A)^2}\right]} = \frac{1}{2\pi\sigma_x^B\sigma_y^B}e^{-\left[\frac{(x-\mu_x^B)^2}{2(\sigma_x^B)^2}+\frac{(y-\mu_y^B)^2}{2(\sigma_y^B)^2}\right]} \quad (5.21)$$

Como no caso que estamos considerando podemos escrever:

$$\begin{cases} \sigma_x^A = \sigma_y^A = \sigma_x^B = \sigma_y^B = \sigma \\ \mu_x^A = \mu_y^A = \mu_y^B = 0 \end{cases}$$

A solução para a função que define os limites de classificação toma a forma:

$$\begin{cases} \forall \mu_x^B \in \mathbb{R} : & x = \frac{\mu_x^B}{2} \\ \forall y' \in \mathbb{R} : & y = y' \end{cases} \quad (5.22)$$

Como podemos ver, a curva é uma reta simples, e no caso em questão é apenas uma reta vertical que divide o plano x-y inteiro em duas regiões facilmente integráveis. Basta então, integrar as PDFs nos seguintes intervalos:

$$\begin{cases} x \in] - \infty; \mu_x^B/2[\\ y \in] - \infty; +\infty[\end{cases}$$

Finalmente, basta solucionarmos a seguinte equação para obtermos a taxa máxima de sucesso possível para a classificação:

$$\Upsilon = \frac{\int_{-\infty}^{+\infty} \int_{-\infty}^{\mu_x^B/2} f_A(x,y) \, dx dy + \int_{-\infty}^{+\infty} \int_{\mu_x^B/2}^{+\infty} f_B(x,y) \, dx dy}{\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_A(x,y) \, dx dy + \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_B(x,y) \, dx dy} \quad (5.23)$$

onde

$$f_\xi(x, y) = f_\xi(\sigma_x^\xi, \sigma_y^\xi, \mu_x^\xi, \mu_y^\xi, x, y) = \frac{1}{2\pi\sigma_x^\xi\sigma_y^\xi} e^{-\left[\frac{(x-\mu_x^\xi)^2}{2(\sigma_x^\xi)^2} + \frac{(y-\mu_y^\xi)^2}{2(\sigma_y^\xi)^2}\right]}$$

A solução pode então ser obtida numericamente (utilizamos o sistema de integração implementado no software Maple 11), sendo que para as separações em estudo obtivemos os resultados apresentados na Tabela 5.4.

Separação	Taxa máxima de sucesso
0.5σ	59.87%
1.5σ	77.34%
3σ	93.32%

Tabela 5.4: Taxa máxima teórica para classificação na discriminação de duas populações gaussianas.

O caso de três gaussianas

Uma análise similar permite a obtenção da taxa máxima de sucesso teórica para três classes geradas a partir de três PDFs gaussianas. Como no caso de duas populações, a melhor classificação possível é aquela que utiliza a função de decisão na qual duas ou mais gaussianas se interceptam. Como no caso anterior, o problema se reduz a integrar as gaussianas até a função ótima de classificação, somar esses valores e dividir pela integral em todo espaço.

Aqui estamos considerando o caso restrito de uma gaussiana com média $(\mu_x, \mu_y) = (0, 0)$ e as outras duas nos vértices de um triângulo equilátero (representado em linhas azuis na Figura 5.27a). Estamos também considerando que os desvios padrões de

todas as gaussianas possuem os mesmos valores. Ainda de forma análoga à solução para duas gaussianas, o parâmetro do problema é o lado do triângulo, ou seja, a distância entre os centros das gaussianas.

Figura 5.27: Diagramas descritivos de: (a) Indicação das populações e de suas médias; (b) as melhores funções de discriminação possíveis; (c) intervalos de integração.

Neste caso, a melhor separação possível é dada pelas funções de decisão $g_{ij}(x)$, representadas em verde na Figura 5.27b, onde ij são os índices das gaussianas que a função de decisão é capaz de discriminar. No caso especial deste problema, não existem regiões não-classificáveis, uma vez que as funções se interceptam em um ponto.

Basicamente, de modo a resolver nosso problema, precisamos integrar as gaussianas nas regiões coloridas na Figura 5.27c. Contudo, devido à natureza altamente simétrica de nosso problema, é suficiente integrar a primeira gaussiana na região amarela da Figura 5.27c e multiplicar esse valor por 6.

Logo, precisamos descobrir a função de decisão g_{AB} (g_{12} na Figura 5.27), para permitir a definição do intervalo de integração. Em nosso caso, isso é uma tarefa simples, uma vez que como no caso de duas populações, nós precisamos apenas encontrar a solução da equação 5.21, mas agora restrita às seguintes condições (dist é a distância entre os centros das gaussianas, dada em unidades de desvio padrão):

$$\begin{cases} \sigma_x^A = \sigma_y^A = \sigma_x^B = \sigma_y^B = \sigma \\ \mu_x^A = \mu_y^A = 0 \\ \mu_x^B = \cos\left(\frac{\pi}{6}\right) \sigma \times \text{dist} \\ \mu_y^B = \sin\left(\frac{\pi}{6}\right) \sigma \times \text{dist} \end{cases}$$

novamente, a solução tem uma forma simples:

$$\begin{cases} \forall x' \in \mathbb{R} : & x = x' \\ \forall \sigma, \text{dist} \in \mathbb{R} : & y = -\cot\left(\frac{\pi}{6}\right) x + \sigma \times \text{dist} \end{cases} \quad (5.24)$$

Então, agora precisamos integrar a PDF até a curva acima, o que significa utilizar o seguinte domínio D de integração:

$$\begin{cases} x \in]-\infty; (\sigma \times \text{dist}) / \cot(\frac{\pi}{6})] \\ y \in [0; -\cot(\frac{\pi}{6})x + \sigma \times \text{dist}] \end{cases}$$

Logo, basta resolver:

$$\Upsilon = \frac{6 \int \int f_A(x,y) dy dx, \text{ sob o domínio D}}{\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_A(x,y) dx dy + \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_B(x,y) dx dy + \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_C(x,y) dx dy}$$

onde as funções $f(x,y)$ são definidas como no caso de duas populações. O denominador assume o valor 3, e a equação se reduz à:

$$\Upsilon = 2 \int \int f_A(x,y) dy dx, \text{ sob o domínio D} \quad (5.25)$$

Obtemos então, a solução numericamente (novamente utilizando o software Maple 11), para os casos em estudo. Os resultados podem ser encontrados na Tabela 5.5.

Separação	Taxa máxima de sucesso
0.5σ	43.77%
1.5σ	65.11%
3σ	88.47%

Tabela 5.5: Taxa máxima teórica para classificação na discriminação de três populações gaussianas.

Resultados da SVM e comparação

De modo a analisar o funcionamento e eficiência da separação dessas populações gaussianas utilizando SVMs, simulamos observações de populações aleatórias. O estudo foi realizado utilizando duas e três classes diferentes, e em ambos os casos analisamos os cenários em que as médias das populações estavam separadas por 0.5, 1.5 e 3 desvios-padrão, tal como procedemos no estudo analítico.

Em cada simulação foram geradas distribuições aleatórias de mil pontos cada. Esta população foi dividida em dois arquivos, cada um com 1000 pontos, contendo a mesma quantidade de representantes de cada população, tirados aleatoriamente. Um dos arquivos foi utilizado para treinamento e validação cruzada, enquanto o outro para realização de um teste de classificação. Realizamos então 10 treinamentos distintos da SVM (para tiragens aleatórias diferentes dos pontos utilizados para treinamento e teste), obtendo finalmente os resultados para a acurácia de classificação.

Os resultados obtidos para os dez treinamentos para todos os três casos de separação entre as distribuições podem ser encontrados na Tabela 5.6. Comparando os resultados obtidos pela SVM para a classificação com os melhores resultados

possíveis analiticamente, verifica-se a alta eficiência deste método para a solução do problema de classificação de populações: a taxa de sucesso máxima teórica foi atingida para todas as separações.

Um resultado similar foi observado utilizando três classes diferentes de pontos, de modo que podemos concluir que para o caso em questão a eficiência deste método mostra-se bastante alta, sendo em geral a natureza do problema (e não o método em si) o principal fator de um possível erro de classificação.

Populações	Separação	Taxa máxima teórica	Taxa de sucesso da 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: Taxas máxima teórica e de sucesso da SVM para classificação na discriminação de duas e três populações gaussianas.

5.5.2 Testes com galáxias do catálogo de Frei

Uma vez verificada a eficácia do método em classificar corretamente dados artificiais gerados a partir de PDFs bem conhecidas, é necessário verificar se este método é capaz de classificar corretamente as galáxias que serão observadas pelo Gaia. Para isso, um primeiro teste bem controlado foi realizado utilizando imagens de galáxias próximas, cujas classificações eram conhecidas sem nenhuma dúvida. Além disso, é necessário verificar o comportamento desta classificação quando quantidades cada vez menores de pixels estão disponíveis para classificação.

Para realizar este teste, dados dos parâmetros CASGM20 mensurados em imagens do catálogo de [Frei et al \(1996\)](#) foram utilizados (já apresentados na seção 5.3.1). Estas medidas foram introduzidas no código implementado para o Gaia. Entretanto, como o número de galáxias neste catálogo é pequeno, e estas galáxias devem ser utilizadas tanto para o treinamento do classificador quanto para o teste da acurácia da classificação, adotamos um procedimento de re-amostragem, baseado em criação de subconjuntos aleatórios de galáxias.

Este procedimento foi realizado da seguinte maneira: 1. o conjunto total de medidas é dividido em dois, nos quais 65% dos dados são utilizados para treinar a SVM e o resto para testar a acurácia de classificação, com as galáxias que estarão em um ou outro conjunto selecionadas de forma aleatória; 2. a SVM é treinada e aplicada aos dados, e a taxa de sucesso na classificação é calculada; 3. novos conjuntos são criados e o processo é iterado.

Para cada tamanho de imagem foram realizadas 100 iterações, e o resultado final da acurácia global obtida pode ser visto na Figura 5.28. Nesta figura podemos

observar que o resultado obtido é bastante estável, se encontrando próximo de 85% de acerto para todos os tamanhos de imagem.

Os resultados obtidos para este teste demonstram que no caso de uma reconstrução de imagem perfeita e da determinação de um conjunto de treinamento sem nenhum tipo de erro de classificação, a melhor taxa de classificação que poderia ser obtida, a princípio, com a utilização de uma SVM treinada exclusivamente em dados de CASGM20 deve atingir $85 \pm 3\%$.

Figura 5.28: Taxa de acurácia global obtida para a classificação de galáxias do catálogo de Frei a partir de medidas CASGM20 e de uma *Support Vector Machine*. As barras de erro foram obtidas a partir de re-amostragens aleatórias dos dados.

5.5.3 Testes com simulações GIBIS

Outro teste que foi realizado com esse código baseado nos parâmetros CASGM20 e SVM é sua aplicação às imagens reconstruídas com dados simulados da missão Gaia, que além da forte degradação devido a PSF e dos algoritmos de reconstrução, também incluem artefatos da reconstrução. Para isso, observações foram simuladas em 2640 galáxias dos três tipos morfológicos disponíveis (elípticas, espirais e irregulares) com a utilização do GIBIS 7 (descritas na seção 5.3.3).

Testamos diferentes tamanhos para o conjunto de galáxias que foi utilizado para treinar a SVM (50, 100, 200, 400, 600 galáxias por tipo morfológico) e de forma similar a subseção anterior, um procedimento de re-amostragem foi utilizado para possibilitar a obtenção dos intervalos de confiança para a taxa de sucesso na classificação—o número total de iterações adotado foi 100. Neste caso, como o treinamento da SVM envolveu maiores capacidades de processamento (podendo levar dezenas de minutos de tempo de cálculo para cada iteração), foram utilizados diversos processos Java locais rodando de forma paralela para o treinamento da SVM.¹⁷

¹⁷O número exato de processos foi igual a 12, e o treinamento consumiu um total de ~ 10 h de

A taxa de sucesso obtida na classificação de cada tipo distinto de galáxia foi calculada separadamente, e a matriz de confusão entre os diversos tipos de galáxia também foi determinada. Os resultados para uma classificação nos três tipos morfológicos podem ser vistos na Figura 5.29 e na Tabela 5.7.

Figura 5.29: Resultados da taxa de sucesso obtida na classificação em três classes com tamanhos variáveis para o conjunto de treinamento da SVM. No eixo x está representado o número de galáxias por classe.

Morfologia real	Classificada como E	Classificada como S	Classificada como I
Elíptica (E)	0.79 ± 0.03	0.08 ± 0.02	0.14 ± 0.02
Espiral (S)	0.14 ± 0.02	0.56 ± 0.03	0.30 ± 0.03
Irregular (I)	0.11 ± 0.02	0.15 ± 0.03	0.74 ± 0.04

Tabela 5.7: Matriz de confusão obtida com os parâmetros CASGM20 e a SVM (treinada em 600 galáxias/tipo) para classificação em três classes morfológicas.

Estes resultados demonstram que o ganho na capacidade de discernimento entre os tipos não é negligenciável ao se adotar quantidades maiores de galáxias para o treinamento da SVM. Além disso, eles também demonstram que a separação entre as galáxias elípticas e irregulares foi relativamente eficiente, se mantendo acima de 70% para praticamente quaisquer tamanhos do conjunto de treinamento.

Contudo, os resultados para classificação em três classes também demonstram que ocorre maior confusão na segregação entre galáxias espirais e irregulares, com

tempo de processamento em uma máquina com 12 *cores* Xeon 2.66/3.05GHz, consumindo uma quantidade total de ~ 1.5 TFLOP. Essa capacidade deve ser considerada muito pequena quando for necessário aplicar este método nos dados reais da missão.

aproximadamente 30% das espirais simuladas sendo classificadas como irregulares. Isso demonstra que, do ponto de vista dos parâmetros adotados para a classificação, existe uma similaridade bastante grande entre alguns tipos de galáxias espirais e de galáxias irregulares, que deve ser originada pelas espirais de tipo mais tardio.

Uma classificação em duas classes foi então realizada para analisar a diferenciação entre as galáxias elípticas (tipos precoces) e galáxias espirais/irregulares (tipos tardios). Os resultados podem ser vistos na Figura 5.30 e na Tabela 5.8.

Figura 5.30: Resultados da taxa de sucesso obtida na classificação em duas classes com tamanhos variáveis para o conjunto de treinamento da SVM. No eixo x está representado o número de galáxias por classe.

Morfologia real	Classificada como E	Classificada como S ou I
Elíptica (E)	0.84 ± 0.02	0.16 ± 0.02
Espiral (S) ou Irregular (I)	0.18 ± 0.02	0.82 ± 0.02

Tabela 5.8: Matriz de confusão obtida com os parâmetros CASGM20 e a SVM (treinada em 600 galáxias/tipo) para classificação em duas classes morfológicas.

Os resultados para classificação em duas classes demonstram que uma taxa de sucesso da diferenciação entre estes dois tipos morfológicos pode ser obtida de forma estável acima de 80% para quaisquer conjuntos de treinamento com mais do que 200 galáxias por tipo morfológico. Essa alta capacidade de diferenciação poderá permitir com que perfis bojo+disco ou bojo sejam escolhidos corretamente para $\sim 80\%$ das galáxias já em uma primeira tentativa para a análise de perfil.

Contudo, a principal conclusão que pode ser obtida a partir das análises desta seção, é de que a partir de galáxias simuladas com o GIBIS 7 que passaram por

todo o processo de reconstrução, medida de parâmetros CASGM20 e classificação já com os códigos que foram implementados para o Gaia será possível realizar estudos morfológicos em galáxias angularmente pequenas com os dados dessa missão espacial. E como já dissemos anteriormente, a maior parte destas galáxias nunca tiveram suas estruturas observadas anteriormente.

5.6 Aplicação aos dados do *Hubble Deep Field*

Os testes realizados com dados simulados com o GIBIS 7, no entanto, dependem fortemente do algoritmo de reconstrução de imagem que é adotado, e sabemos que os algoritmos disponíveis atualmente não são ótimos. Sabemos também que reconstruções mais detalhadas podem ser teoricamente realizadas, como visto no Capítulo 2, e embora não possamos realizar testes com algoritmos ainda não existentes, podemos realizar testes utilizando imagens obtidas em regimes similares aos de imagens do Gaia no caso de uma reconstrução perfeita.

O HDF pode ser considerado como um bom caso de teste para o sistema de classificação morfológica que será aplicado ao Gaia no caso de uma reconstrução de imagem perfeita, mas de uma classificação imperfeita. Deste modo, os resultados que serão obtidos nesta seção servem como um limite superior para aquilo que eventualmente pode ser realizado com os métodos descritos neste capítulo para uma classificação puramente morfológica pelo Gaia caso um algoritmo de reconstrução de imagens perfeito esteja disponível.

Contudo, como é adequado utilizar a maior quantidade de galáxias corretamente classificadas possível, e como já pairavam suspeitas sobre a possibilidade de existir alguma contaminação por identificações cruzadas incorretas entre as classificações visuais de [van den Bergh et al \(1996\)](#) e a nossa extração de selos descrita na seção 5.3.2, resolvemos realizar uma nova classificação visual diretamente a partir dos selos recortados para os objetos individuais—as galáxias selecionadas foram as 400 mais brilhantes (a partir da soma simples dos fluxos de cada pixel).

Essa nova classificação visual foi realizada por dois colegas especialistas do IAG-USP (Dra. Sandra dos Anjos e Dr. Rubens Machado) de forma completamente independente. Os resultados das classificações obtidos individualmente foram comparados e apenas as 281 galáxias cujas classificações visuais eram similares entre os dois classificadores foram utilizadas para construção do conjunto de treinamento da SVM e para a realização do teste da acurácia na classificação.

O procedimento de análise adotado foi o mesmo descrito na seção anterior: as imagens tiveram seus parâmetros CASGM20 mensurados e 120 re-amostragens aleatórias foram utilizadas para definir conjuntos de treinamento e testes para a SVM. Estes conjuntos foram utilizados para a determinação da acurácia na classificação dos três grandes tipos morfológicos, sendo que uma classificação similar em tipos precoce e tardio também foi realizada—o procedimento foi repetido para diversos tamanhos do conjunto de treinamento. Os resultados obtidos podem ser vistos na Figura 5.31 e nas Tabelas 5.9 e 5.10.

Figura 5.31: Resultados da taxa de sucesso obtida na classificação em duas e três classes com tamanhos variáveis para o conjunto de treinamento da SVM dos dados do *Hubble Deep Field North*.

Morfologia real	Classificada como E	Classificada como S ou I
Elíptica (E)	0.87 ± 0.07	0.13 ± 0.07
Espiral (S) ou Irregular (I)	0.17 ± 0.05	0.83 ± 0.05

Tabela 5.9: Matriz de confusão para classificação dos dados do HDF-N em duas classes morfológicas (SVM treinada em 50 galáxias/tipo).

Morfologia real	Classificada como E	Classificada como S	Classificada como I
Elíptica (E)	0.81 ± 0.09	0.27 ± 0.07	0.07 ± 0.06
Espiral (S)	0.17 ± 0.08	0.64 ± 0.11	0.19 ± 0.11
Irregular (I)	0.09 ± 0.04	0.29 ± 0.06	0.62 ± 0.06

Tabela 5.10: Matriz de confusão obtida para classificação dos dados do HDF-N em três classes morfológicas (SVM treinada em 50 galáxias/tipo).

Os resultados obtidos para a classificação das imagens do HDF mostram que é possível distinguir galáxias tipo precoce de tardio com uma acurácia de $\sim 85\%$, atingindo o que seria a máxima taxa possível estimada a partir de galáxias próximas. Além disso, esses resultados também demonstram que galáxias elípticas, espirais e peculiares (que incluem galáxias irregulares, em interação, etc.), podem ser classificadas com taxas de sucesso de $\sim 81\%$, $\sim 64\%$ e $\sim 62\%$, respectivamente.

Comparando estes resultados com aqueles obtidos na seção anterior para simulações, observamos que embora as reconstruções de imagem a partir da utilização do *ShuffleStack* não apresentem resultados muito ruins, ainda é o procedimento de

reconstrução que está limitando a qualidade destas classificações puramente morfológicas. Isso, pois a classificação em tipos tardio e precoce apresentou resultados acima de $\sim 80\%$ para conjuntos de treinamento perfeitos, enquanto que para imagens do HDF (que representam reconstruções perfeitas, com conjunto de treinamento imperfeito) apresentam taxas de sucesso superiores, de $\sim 85\%$.

Os resultados obtidos aqui a partir dos dados reais do HDF demonstram que mesmo em face de dados obtidos em condições similares àquelas que serão enfrentadas pelo Gaia, os métodos aqui desenvolvidos possuem uma boa performance na classificação morfológica de galáxias nos três tipos de Hubble, além de se mostrarem capazes de segregar eficientemente as galáxias em tipos precoce e tardio, com taxa de sucesso similares àquelas que são obtidas com sua aplicação em galáxias grandes e próximas.

5.7 Conclusões

Neste Capítulo apresentamos parâmetros coletados na literatura que definem um volume chamado CASGM20 no qual morfologias distintas de galáxias ocupam locais característicos, e apresentamos o sistema que construímos para o cálculo destes parâmetros em imagens reconstruídas a partir de dados do satélite Gaia. Mostramos que este sistema, quando utilizado para calcular os parâmetros de imagens com amostragens diversas, se comporta de forma mais ou menos estável até imagens formadas por apenas 21×21 pixels (ver Figura 5.14).

Utilizamos os códigos construídos para analisar as imagens reconstruídas do Gaia em imagens recortadas do Hubble Deep Field North, e verificamos que tipos morfológicos distintos de fato ocupam locais distintos em planos deste espaço. Essa conclusão também foi obtida a partir da análise de simulações de observações e de reconstruções de 2640 galáxias dos três tipos morfológicos de galáxias.

Mostramos com o uso de todos estes conjuntos de dados CASGM20 calculados, que uma PCA não é capaz de reduzir significativamente a dimensão do problema enfrentado, de modo que técnicas de aprendizado computacional mais avançadas devem ser aplicadas para se obter resultados acurados de forma automática.

Apresentamos então um método moderno de classificação e aprendizado computacional que foi utilizado poucas vezes em Astronomia até hoje, chamado *Support Vector Machines*, e que é adequado para ser aplicado em cenários como o que será enfrentado pelas imagens reconstruídas mensuradas pelos parâmetros CASGM20. Ele permite que todos os parâmetros sejam levados em conta paralelamente, e não está sujeito a problemas enfrentados por outros métodos mais comuns, como por exemplo redes neurais artificiais.

Primeiramente demonstramos que os códigos que implementamos para o Gaia (que utilizam uma biblioteca de SVMs disponível na literatura) é capaz de diferenciar populações aleatórias gaussianas com sucesso, essencialmente atingindo as máximas taxa de acurácias que podem ser teóricamente obtidas.

Aplicamos então este sistema em medidas de imagens de galáxias próximas, demonstrando que a taxa de acurácia global obtida para a classificação em tipos precoce/tardio para quaisquer amostragens das imagens das galáxias (entre 81x81 e 11x11 pixels) se mantém estável entre 84% e 85%, de modo que em um cenário no qual uma reconstrução de imagem perfeita estiver disponível para os dados do Gaia e em que o conjunto de treinamento seja construído sem nenhum tipo de erro de classificação, o valor de acurácia na classificação que pode ser obtida por este código deve ser de $85\% \pm 3\%$.

Baseados em simulações realizadas com o GIBIS 7 e reconstruções com o *ShuffleStack*, testamos esse código sob um regime de reconstrução de imagem abaixo do ideal. Esse teste mostrou que esse código pode ser utilizado com sucesso em dados do Gaia, apresentando uma taxa de sucesso para diferenciação entre tipos precoce/tardio superiores a 80%, e nas três classes morfológicas principais, elípticas, espirais e irregulares, de 79%, 56% e 74%, respectivamente, desde que o conjunto de treinamento da SVM seja perfeitamente construído.

Finalmente, utilizamos imagens de galáxias do *Hubble Deep Field North* para analisar o comportamento do sistema sob um regime no qual a reconstrução de imagem é perfeita, mas a preparação do conjunto de treinamento da SVM é sujeita a erros de classificação. Esse estudo demonstrou que mesmo neste caso são obtidas taxas de classificação entre tipos precoce/tardio de $\sim 85\%$, e nas classes morfológicas principais, elípticas, espirais e irregulares, de $\sim 81\%$, $\sim 64\%$ e $\sim 62\%$, respectivamente.

Os resultados deste capítulo demonstram que com a utilização dos códigos aqui desenvolvidos, estudos morfológicos com os dados provenientes do Gaia poderão ser realizados de forma satisfatória, mesmo quando apenas conjuntos de treinamento com erros estiverem disponíveis, ou quando as reconstruções de imagens forem realizadas sob regimes abaixo do ideal.

Estimativa de parâmetros morfológicos

“Um retrato é apenas a idéia aproximada de uma pessoa.”, Florbela Espanca

Conteúdo

6.1	Introdução	172
6.1.1	Perfis de brilho	173
6.1.2	Ajuste de perfis no espaço de Radon	174
6.2	Construção do modelo	176
6.3	Optimização dos parâmetros	180
6.3.1	Estimativas iniciais	181
6.3.2	Algoritmos Genéticos	184
6.3.3	BFGS	186
6.4	Testes com simulações	188
6.5	Aprimoramentos e possibilidades	192
6.6	Conclusões	195

Além da classificação morfológica, o outro objetivo da análise da forma de objetos extensos é a determinação de seus parâmetros morfológicos, ou seja as características de seus diversos componentes estruturais. Isso é realizado através de ajustes de perfis de brilho aos dados observacionais, gerando uma decomposição dos objetos em um ou mais componentes simples. Contudo, as imagens reconstruídas de objetos extensos por meio da aplicação dos algoritmos atualmente disponíveis não permitiriam que esse estudo fosse realizado de forma confiável.

Neste Capítulo discutiremos sobre os perfis de brilho mais utilizados para estudo de galáxias e sobre como esses perfis podem ser ajustados aos dados do Gaia sem que seja necessário realizar uma reconstrução de imagem bi-dimensional. Mostraremos o método que desenvolvemos para realizar esse ajuste, que é baseado na interpretação das observações do Gaia como uma transformada de Radon do objeto, e na realização de simulações (Krone-Martins, 2010). Apresentaremos também resultados obtidos com a aplicação deste método em um conjunto estatisticamente significativo de observações simuladas. Finalmente, comentaremos sobre possibilidades de aprimoramento e aplicações do método desenvolvido em imagens normais.

6.1 Introdução

A primeira utilização de ajuste de perfis analíticos para estudar a distribuição de brilho de galáxias foi realizada por [de Vaucouleurs \(1948\)](#), que mostrou que o brilho de galáxias elípticas tende a seguir uma lei tipo $\exp(-r^{1/4})$. Então, [de Vaucouleurs \(1959\)](#) (apud [Freeman, 1970](#)), propôs que o perfil de brilho de galáxias espirais possa ser decomposto em uma função criada a partir da adição de dois componentes, um para o bojo e um para o disco.

Estes perfis de brilho são importantes ferramentas no estudo de objetos extensos (como galáxias), pois além de revelar as estruturas morfológicas destes objetos, eles são traçadores dos processos físicos responsáveis pela formação e estabilidade ([Freeman, 1970](#)) dos componentes estruturais destes objetos. Bojos clássicos, pseudo-bojos e barras, são alguns exemplos de componentes centrais de galáxias tipo espirais e S0, que provavelmente possuem mecanismos de formação distintos ([Kormendy & Kennicutt, 2004](#)). Existe toda uma miríade de estudos astronômicos que são possibilitados por estas decomposições, desde análises individuais de galáxias (ex. [Gadotti, 2008](#)), até estudos de populações destes objetos de acordo com a idade do universo (ex. [Marleau & Simard, 1998](#); [Hathi et al, 2009](#)).

Para que o estudo destes diferentes componentes estruturais das galáxias seja possível, é necessário que os dados observacionais sejam decompostos nas contribuições destes componentes. Isso é um problema habitualmente resolvido através de ajustes das funções analíticas que descrevem estes perfis às imagens dos objetos por métodos de minimização de mínimos quadrados sob pontos amostrados em certas distâncias calculadas a partir do centro da galáxia.

Inicialmente, os ajustes destas funções analíticas eram realizados utilizando dados obtidos com a extração do perfil de brilho sob o semi-eixo maior da galáxia, sendo que primeiro os perfis eram ajustados separadamente apenas nas regiões nas quais eles eram dominantes, e então os parâmetros dos dois componentes eram determinados através de um processo iterativo no qual ora o bojo, ora o disco, eram ajustados ([Kormendy, 1977](#); [Burstein, 1979](#)). Uma evolução, adotada por [Kent \(1985\)](#), foi a utilização principalmente de ajustes unidimensionais, porém de forma simultânea para ambos os componentes, além de considerar o perfil de brilho amostrado sob os semi-eixos maiores e menores em um trabalho realizado com 105 galáxias.

Com a disponibilização de maior poder computacional, principalmente na última década, se torna possível realizar este ajuste diretamente nos dados bi-dimensionais das imagens através de códigos como o GIM2D ([Simard, 1998](#)), o BUDDA ([de Souza et al, 2004](#)) e o GALFIT ([Peng et al, 2002, 2010](#)).

Neste Capítulo introduzimos uma terceira forma de se realizar esta decomposição, que além de possibilitar o ajuste de perfis de brilho e decomposição de galáxias observadas pelo Gaia mesmo na ausência de reconstruções detalhadas de imagens bi-dimensionais, também poderá vir a ter aplicação mais ampla, possibilitando a realização deste tipo de análise em quaisquer galáxias cujas imagens tenham sido obtidas em baixos níveis de sinal/ruído. Cabe notar que o método desenvolvido aqui é geral, podendo ser aplicado em perfis de brilho de outros objetos extensos.

6.1.1 Perfis de brilho

Para realizar a decomposição morfológica do perfil de brilho de galáxias, é necessário assumir a priori a quantidade de componentes em que se deseja separar o objeto.¹ Contudo do ponto de vista físico é interessante sempre adotar critérios parcimoniosos nesta escolha, pois numericamente é possível obter melhores ajustes ao considerar quantidades cada vez maiores de componentes, que não necessariamente possuem justificativas físicas na análise de uma certa galáxia.

Em geral, modelos de galáxias utilizam dois, três ou quatro componentes: uma fonte nuclear pontual, um bojo, uma barra e um disco—usualmente os braços em galáxias espirais não são adicionados, embora recentemente Peng et al (2010) tenha adicionado tais estruturas no código GALFIT. A barra e a fonte central são em geral utilizadas em galáxias particulares, e em análises em pequenas escalas (como em Gadotti, 2008) e a maior parte das decomposições é realizada considerando apenas dois componentes: o disco e o bojo. Isso, pois estes são os responsáveis pela maior fração da luminosidade de grande parte das galáxias, além de que em muitos casos a resolução dos dados é inadequada para que outros componentes sejam resolvidos.

Os perfis de brilho mais freqüentemente utilizados para descrever os discos são em geral perfis exponenciais, que podem ser escritos como:

$$I_d(r) = I_{0d} \exp\left(-\frac{r}{r_d}\right) \quad (6.1)$$

onde I_{0d} é a intensidade no centro do disco, r_d é a escala de comprimento do disco e r é a distância na qual a intensidade é calculada.

Para o bojo, um dos perfis de brilho mais importantes é o proposto por de Vaucouleurs (1948), comentado na introdução deste Capítulo. Contudo nos últimos anos tem se adotado freqüentemente perfis descritos pela parametrização de Sérsic (Blanton & Moustakas, 2009). Este perfil, inicialmente descrito em Sérsic (1963); Sérsic (1968) (apud Graham & Driver, 2005) pode ser escrito como:

$$I_b(r) = I_{0b} \exp\left(-b_n \left[\left(\frac{r}{r_b}\right)^{\frac{1}{n}} - 1\right]\right) \quad (6.2)$$

onde r_b é o raio efetivo que contém metade da luminosidade integrada da galáxia, I_{0b} é a intensidade do bojo no raio efetivo, n é o índice de Sérsic, b_n é uma constante e r é a distância na qual a intensidade é calculada.

O valor da constante b_n depende do valor escolhido para o índice de Sérsic, podendo ser calculado numericamente a partir da integral do perfil. Contudo, em geral se adotam aproximações analíticas deste parâmetro, como por exemplo a aproximação $b = 1.9992n - 0.3271$ de Capaccioli (1989) (apud Graham & Driver, 2005) que é válida para um largo intervalo de índices de Sérsic, com $0.5 < n < 10$.

Do ponto de vista físico, o índice de Sérsic define qual tipo de bojo uma certa galáxia possui: um bojo clássico quando $n > 2$, ou um pseudo-bojo quando $n < 2$.

¹Para isso, uma classificação morfológica baseada em parâmetros independentes de ajustes de perfil (que vimos no Capítulo 5 desta tese) é realizada preliminarmente no *pipeline* do Gaia.

Além disso, dependendo de seu valor, o perfil de Sérsic pode também servir para modelar barras, como realizado por Gadotti (2008)—isso, para índices menores do que a unidade. Adicionalmente, componentes com altos índices de Sérsic tendem a gerar um sinal similar àquele criado por uma fonte pontual, após a degradação da imagem pela convolução com a PSF.

Então, o perfil de brilho de uma galáxia que possui ambos os componentes pode ser escrito como a soma dos perfis do disco e do bojo, ou seja: $I(r) = I_d(r) + I_b(r)$. É importante notar, que na forma em que escrevemos os perfis do bojo e do disco, os significados na expressão para o perfil de brilho total $I(r)$ dos parâmetros I_{d0} e I_{b0} são distintos entre si, bem como os de r_d e r_b , e não devem ser confundidos.

Figura 6.1: Representação de perfis de galáxias considerados neste Capítulo. A Figura da esquerda mostra o perfil de Sérsic (com $I_{0b} = 1$ e $r_b = 0.2''$) para diferentes valores de n (0.25, 0.5, 1, 2, 4, 6, 8). A Figura da direita um perfil exponencial em azul, um perfil de Sérsic com $n=4$ em vermelho e o perfil composto em preto.

Outros perfis que são utilizados na literatura, como em Peng et al (2010), são o perfil de Ferrer, usado para ajustar barras e lentes, o perfil de King, adotado em aglomerados globulares, o perfil de Nuker, usado para a distribuição central de luz de galáxias próximas, e o perfil de van der Kruit & Searle (1981), que possibilita o ajuste de galáxias *edge-on*. Contudo, como comentamos no início desta seção, é interessante que decomposições sejam sempre realizadas com os perfis mais simples e com o menor número possível de parâmetros.

6.1.2 Ajuste de perfis no espaço de Radon

Para realizar estudos de decomposição bojo/disco e de determinação de parâmetros morfológicos destes componentes estruturais a partir dos dados do Gaia, é possível adotar dois caminhos distintos. O primeiro consiste em um ajuste dos perfis comentados na seção anterior diretamente em imagens reconstruídas com os dados originais

do satélite, utilizando os códigos já disponíveis, como o BUDDA e o GALFIT, por exemplo. Contudo, vimos no Capítulo 5 alguns exemplos de reconstruções de dados simulados para galáxias a partir de um dos melhores algoritmos disponíveis, e pode-se rapidamente concluir que atualmente esse caminho encontra-se praticamente inviabilizado.²

Um outro caminho possível, é a realização do ajuste não do perfil de luminosidade da galáxia no espaço real, mas sim da transformada deste perfil no espaço de Radon. Como vimos no Capítulo 2, as observações do Gaia constituem basicamente um subconjunto amostrado do sinal bi-dimensional da galáxia no espaço de Radon, de modo que uma vez que conhecemos os perfis que desejamos ajustar no espaço real, e sabemos como calcular as transformadas destes perfis para o espaço de Radon, é suficiente que os perfis sejam transformados e ajustados diretamente às observações do Gaia neste espaço.

Em princípio poderíamos realizar as transformações analíticas destes perfis, o que do ponto de vista de tempo de cálculo optimizaria significativamente o procedimento de ajuste. Contudo, isso restringiria significativamente a possibilidade de realizarmos estudos mais genéricos a partir dos dados da missão, além de dificultar a adição de efeitos diversos, como convoluções, por exemplo. Desse modo, o sistema de ajuste construído é baseado em transformações numéricas de observações simuladas de certos perfis no espaço real ao espaço de Radon. Uma representação esquemática do funcionamento desse sistema pode ser encontrada na Figura 6.2.³

Figura 6.2: Esquema do fluxo de dados no sistema de ajuste de perfil construído para determinação de parâmetros de perfis morfológicos a partir dos dados do Gaia.

Esse sistema possibilita a adoção de perfis de luminosidade tão simples ou tão complexos quanto se queira na construção dos modelos, mesmo que os perfis não possuam representação analítica simples no espaço de Radon.

²Notamos, entretanto, que caso melhores algoritmos de reconstrução sejam propostos/implementados no futuro, esse caminho poderia vir a ser considerado, até mesmo como possível teste de consistência entre estratégias distintas de decomposição.

³Em verde estão representados dados que entram no sistema, em azul processos ou dados internos e em roxo dados que saem—estas cores serão utilizadas nos diagramas deste Capítulo.

6.2 Construção do modelo

Uma possibilidade para viabilizar o procedimento de ajuste comentado na seção anterior seria utilizar o GIBIS como processo direto, ou seja, adotar este simulador para produzir janelas simuladas para certos parâmetros que poderiam ser comparadas com as janelas observadas pelo satélite. Contudo, na prática isso não é possível, pois os requerimentos computacionais deste simulador (em termos de tempo de execução e memória) inviabilizariam a análise da quantidade de objetos esperada.

Desse modo construímos um modelo simplificado para o processo direto, cuja representação esquemática de todos seus passos pode ser vista na Figura 6.3.

Figura 6.3: Esquema da construção do modelo direto do sinal de um perfil de brilho de um objeto extenso. O modelo pode ser qualquer função no espaço real.

O primeiro passo na construção do processo direto é a amostragem do perfil de brilho do objeto a ser estudado em uma matriz regular. Para isso, um modelo do objeto que está sendo simulado é construído a partir de um disco com perfil exponencial e um bojo com perfil de Sérsic. Durante esta amostragem, transformações geométricas são necessárias para incluir no modelo o ângulo de posição do objeto no céu (uma rotação simples) e a elipticidade. Adicionalmente, a implementação destas transformações também é capaz de levar em fatores de *diskness/boxiness* destes objetos, como em Athanassoula et al (1990), por meio da transformação:

$$r(x, y) = \left[|x - x_{gal}|^{C_0+2} + \left| \frac{y - y_{gal}}{1 - e} \right|^{C_0+2} \right]^{\frac{1}{C_0+2}} \quad (6.3)$$

onde e representa a elipticidade e C_0 um parâmetro que representa o quão *disky/boxy* é a galáxia: $C_0 = 0$, elipse normal, $C_0 < 0$ galaxia *disky* (com um formato similar a

um diamante) e $C_0 > 0$ galáxia *boxy* (com formato similar a uma caixa). A Figura 6.4 mostra exemplos limites de galáxias simuladas com as três possibilidades de valores para o parâmetro C_0 .

Figura 6.4: Imagens sintéticas de galáxias geradas a partir da amostragem do modelo e da aplicação de transformações geométricas. O parâmetro de elipticidade vale $e = 0.5$ em todos os perfis, $C_0 = -0.5$ a esquerda, $C_0 = 0.0$ ao centro e $C_0 = +0.5$ a direita (mapa de cores logarítmico correndo do azul para o vermelho).

A amostragem é realizada em uma matriz cuja quantidade de pixels pode ser configurada pelo usuário, tal como o tamanho destes pixels. Na região interior a uma certa distância do centro, cuja configuração padrão é 10 mas, o perfil é amostrado como se estivesse a esta distância, evitando valores demasiadamente altos.

Após a amostragem é realizada uma convolução. Para isso, a imagem sintética original é levada ao espaço de Fourier, multiplicada pela transformada de Fourier da PSF e o sinal resultante é trazido de volta ao espaço real por meio de uma transformada inversa, gerando a imagem sintética convoluída. A PSF pode ser adotada como uma função gaussiana com tamanho configurável ou uma PSF lida de uma biblioteca, como a disponível no GaiaSimu. Um exemplo deste processo, com PSF gaussiana de FWHM=180mas, pode ser visto na Figura 6.5.

Figura 6.5: Perfis de imagens sintéticas de galáxias geradas a partir da amostragem do modelo e da aplicação de transformações geométricas e convolução por uma PSF gaussiana com FWHM de 180mas. As galáxias são as mesmas da Figura 6.4.

A imagem sintética convoluída passa então por um processo de re-amostragem poissoniana, onde o valor em cada pixel é substituído por uma amostra aleatória tomada a partir de uma distribuição de Poisson cuja média é o valor original do pixel. Um exemplo dos resultados deste processo, para as três galáxias de exemplo, pode ser visto na Figura 6.6.

Figura 6.6: Imagens sintéticas de galáxias geradas a partir da amostragem do modelo, da aplicação de transformações geométricas, convolução por uma PSF gaussiana com FWHM de 180mas e re-amostragem poissoniana. As galáxias são as mesmas da Figura 6.4, mas aqui o mapa de cores é linear.

O próximo passo é levar a imagem sintética para o espaço de Radon. Para isso, a imagem é preliminarmente preenchida com valores nulos em todos os pontos além de um certo Raio (a configuração padrão é um raio de 2'', aproximadamente o raio coberto pelas janelas do *Sky-Mapper* do Gaia). Então, a transformada de Radon é calculada.⁴ Um exemplo com o sinal das galáxias utilizadas nesta seção no espaço de Radon pode ser visto na Figura 6.7—é interessante notar que quanto menor o valor de C_0 , mais bem localizado é o sinal neste espaço.⁵

Figura 6.7: Sinal das galáxias de exemplo desta seção no espaço de Radon.

⁴Na implementação atual do método aqui descrito, que foi escrita em linguagem R, a transformada é calculada por meio do pacote PET (Schulz, 2006).

⁵Um fenômeno similar ocorre para maiores elipticidades.

Como o Gaia não amostra este espaço na direção θ de forma homogênea, apenas os ângulos observados em uma certa direção do céu devem ser mantidos, portanto o modelo deve levar isso em conta. Para isso, pode-se fornecer uma lista de ângulos ao código, ou alternativamente as coordenadas galácticas nas quais o objeto está sendo simulado. Neste segundo caso, o código utiliza as rotinas em Java da biblioteca GaiaSimu para calcular a lei de varredura nominal da missão e determinar a lista de ângulos de trânsito que serão observados. Um exemplo com os ângulos amostrados nas coordenadas $(l, b) = (160^\circ, 50^\circ)$ pode ser visto na Figura 6.8.

Figura 6.8: Sinal das galáxias de exemplo desta seção no espaço de Radon com trânsitos observados pelo Gaia na direção $(l, b) = (160^\circ, 50^\circ)$.

Finalmente, o ruído gaussiano representando o ruído de leitura é adicionado, além da combinação (*binning*) dos pixels em *samples* das janelas. A Figura 6.9 mostra os senogramas resultantes de janelas dos *Sky-Mappers* e *Astro-Fields 2*, 5 e 8.

Figura 6.9: Senogramas tipo SM e AF simulados em $(l, b) = (160^\circ, 50^\circ)$.

O modelo do processo direto descrito aqui possibilita a realização de simulações de trânsitos para quaisquer objetos de forma eficiente: cada chamada do processo direto completo levou $\sim 197 \pm 3$ ms em $\sim 80\%$ das execuções em um núcleo Intel Core i7-620M 2.66 GHz (1000 execuções foram realizadas). Além disso, essa eficiência permite que este modelo seja utilizado para realizar ajustes de parâmetros de quaisquer perfis à trânsitos observados—para isso os termos de ruído são desconsiderados. Como comentamos anteriormente, neste trabalho estamos considerando apenas galáxias formadas por bojos de Sérsic e discos exponenciais,⁶ contudo tanto o método descrito aqui quanto o código implementado são genéricos, podendo ser acoplados a outros perfis de brilho.

6.3 Optimização dos parâmetros

Uma vez de posse de um modelo para o processo direto, é necessário determinar os parâmetros que mais aproximem este modelo dos dados observados, ou seja, resolver um problema de optimização. Uma das possibilidades para resolver este problema é a optimização de uma norma euclidiana (também chamada de norma L2) do vetor formado pelas diferenças entre os valores previstos pelo modelo para um certo vetor $\mathbf{p}$ de parâmetros ($f(\mathbf{p})$) e os valores observados ($\mathbf{d}$), pixel a pixel do senograma. Ou seja, deseja-se resolver o seguinte problema de optimização:

$$\min_{\mathbf{p}} \|f(\mathbf{p}) - \mathbf{d}\|_2 \quad (6.4)$$

A resolução ideal do problema descrito acima é bastante dependente do formato da função f .⁷ Contudo, de modo a possibilitar a utilização de quaisquer modelos, tenham eles representações analíticas ou não, decidimos utilizar nos códigos que desenvolvemos para a missão Gaia, um primeiro optimizador genérico, baseado em uma heurística denominada Algoritmo Genético, e um segundo optimizador baseado em derivadas numéricas da função f que descreve o modelo.

O conceito do funcionamento do sistema implementado é baseado em três passos intermediários. O primeiro, consiste em uma determinação inicial de parâmetros a partir de uma análise pouco dependente de modelos do sinal do senograma. O segundo passo é uma optimização dos parâmetros com a utilização de uma heurística denominada Algoritmo Genético. O terceiro, é uma optimização final destes parâmetros com o uso de um método baseado em derivadas numéricas da função que está sendo optimizada. Cabe notar aqui que optimizações podem ser executadas sem a utilização dos primeiro e terceiro passos, contudo isso acarretaria ou em uma menor precisão dos parâmetros finais, ou em um tempo de convergência muito maior. Além disso, devido a natureza estocástica dos Algoritmos Genéticos, para cada objeto analisado, este sistema é executado cinco vezes de forma independente, e a solução com menor norma L2 é adotada como a solução para os parâmetros.

⁶Um fundo de céu constante também é geralmente adicionado.

⁷Os teoremas *No Free Lunch* (Wolpert et al, 1997) provam que não existe um método de optimização que seja melhor do que todos os outros em quaisquer problemas.

6.3.1 Estimativas iniciais

Uma análise livre de modelos do senograma já nos permite obter estimativas iniciais para alguns parâmetros. Um destes é o ângulo de posição do objeto no céu. Como pode ser visto na Figura 6.10, a posição no eixo θ que possui maior valor varia de acordo com o ângulo de posição de forma facilmente correlacionada. Isso ocorre pois o maior valor de um sinal bi-dimensional integrado em um intervalo de projeções angulares ocorrerá justamente quando o ângulo θ da projeção for tal que maximize a quantidade de informação do objeto por intervalo de integração (“pixel” na direção ρ do espaço de Radon).

Figura 6.10: Senogramas de uma mesma galáxia simulada em diferentes ângulos de posição: 0° , 75° e 125° .

Para o eixo x do plano focal alinhado na direção Norte-Sul, a relação entre a coordenada θ e o ângulo de posição PA é dada por:

$$\text{PA} = \begin{cases} 90^\circ - \theta & \text{se } \theta \leq 90^\circ, \\ 270^\circ - \theta & \text{se } \theta > 90^\circ. \end{cases} \quad (6.5)$$

Contudo, de modo a maximizar a possibilidade de aplicação desta análise, minimizando efeitos associados a incertezas na determinação do pixel de máximo sinal além de possibilitar sua aplicação em casos de galáxias observadas com baixo nível de sinal/ruído, o senograma pode ser filtrado previamente. Entretanto a maior parte dos estudos sobre filtragem de sinal é realizada com ruídos gaussianos, de modo que é desejável possuir um sinal com este tipo de ruído.

A transformada de Anscombe (Anscombe, 1948), é uma transformada estatística que a partir de um sinal que segue uma distribuição poissoniana, gera um sinal distribuído de forma aproximadamente gaussiana.⁸ Originalmente construída para este caso, ela foi generalizada por Bijaoui (1994) para um caso de ruído misto, formado

⁸A transformação somente pode ser aplicada aos dados das janelas do Gaia, porque a soma de variáveis aleatórias poissonianas independentes (pixels no plano focal do Gaia) continua sendo uma variável poissoniana (*samples* das janelas transmitidas).

pelo ruído poissoniano original adicionado de um ruído gaussiano, e publicada em (Murtagh et al, 1995). Para um processo como uma imagem astronômica, no qual o sinal total pode ser escrito como $\alpha n + \gamma$, onde α representa um ganho, γ uma variável gaussiana aditiva de média g e desvio padrão σ e n uma variável poissoniana, tal transformação pode ser escrita como:

$$I_n(x, y) = \frac{2}{\alpha} \sqrt{\alpha I(x, y) + \frac{3}{8} \alpha^2 + \sigma^2 - \alpha g} \quad (6.6)$$

onde $I_n(x, y)$ representa o sinal transformado na posição (x, y) e $I(x, y)$ o sinal original. Essa transformação é tal que o valor da variância de $I_n(x, y)$ é completamente independente do valor de $I(x, y)$, e I_n segue uma distribuição gaussiana.

Uma vez que o sinal tenha seu ruído transformado em um termo aditivo gaussiano, ele deve ser filtrado por um filtro que remova as frequências mais altas. O filtro implementado é baseado em *soft-thresholding* do sinal no espaço de *wavelets*,⁹ sendo que o *soft-thresholding* (multiplicação dos coeficientes de *wavelets* por pesos cada vez menores) foi escolhido pelo fato de que não são esperadas descontinuidades em um sinal de um objeto extensos transformado para o espaço de Radon—cortar de forma brusca coeficientes de *wavelets* a partir de uma certa escala (*hard-thresholding*) pode gerar tais descontinuidades. A *wavelet* escolhida é a Daubechies D4, tal como em Sanz et al (1999). Esta base foi selecionada por ser capaz de descrever adequadamente sinais constantes e lineares, tais quais aqueles que são esperados localmente nas diversas regiões de um senograma gerado a partir da transformada de Radon do perfil de estruturas extensas (como galáxias, por exemplo), além desta ser numericamente eficiente, contando com apenas quatro coeficientes para sua construção.

Um código para realização da estimativa inicial de parâmetros foi construído em R (o filtro utilizado foi implementado por Whitcher, 2010), e um esquema que representa o fluxo de dados dentro deste pode ser visto na Figura 6.11.

Figura 6.11: Esquema do fluxo de dados no sistema de ajuste de perfil construído para determinação de parâmetros de perfis morfológicos a partir dos dados do Gaia.

⁹Uma transformação de *wavelet* basicamente é a representação de um sinal qualquer em uma base auto-similar (que pode possuir diversos formatos) que é localizada no espaço e no tempo.

O código foi utilizado para testar a eficiência da estimativa para o ângulo de posição de 1000 galáxias simuladas com elipticidade igual a 0.6, em duas condições de sinal ruído nas coordenadas $(l, b) = (160^\circ, 50^\circ)$. A comparação entre os valores simulados e obtidos pode ser vista na Figura 6.12.

Figura 6.12: Comparação entre os ângulos de posição simulados e estimados pelo método descrito nesta seção para 1000 galáxias geradas nas coordenadas $(l, b) = (160^\circ, 50^\circ)$. Em (a) sem ruído gaussiano adicionado e em (b) com $\text{SNR} \sim 0.2$. As linhas vermelhas representam a lei de varredura do Gaia.

Os resultados demonstram que uma estimativa inicial para o ângulo de posição a partir do método descrito aqui é extremamente robusta desde que a informação esteja disponível para análise. Naturalmente o método está limitado a determinar ângulos de forma discretizada pela lei de varredura adotada pelo satélite Gaia, pois a amostragem do sinal na direção θ do espaço de Radon é realizada de forma incompleta e não-homogênea por este satélite.

Além disso, a lei de varredura do Gaia também limita a determinação de outros parâmetros, que em princípio podem ser estimados utilizando técnicas similares. Por exemplo, uma estimativa para a elipticidade da galáxia pode ser realizada a partir da determinação de uma dimensão para o objeto (por binarização por limiar, por exemplo) que seja mensurado nas colunas do senograma referentes ao máximo valor e à sua coordenada θ complementar. Contudo, como comentaremos ao final deste Capítulo, algumas novas técnicas de *inpainting* podem vir a possibilitar a realização de estimativas iniciais para estes outros parâmetros em um futuro próximo.

Realizamos um último teste aplicando este método em 1000 galáxias com elipticidade 0.6 simuladas com ângulos de posição aleatórios e em regiões aleatórias do céu (seguindo uma distribuição uniforme na esfera celeste), sendo que a distribuição nominal dos trânsitos para cada região simulada foi levada em conta na construção dos senogramas. Tal como o teste anterior, este foi realizado em duas condições de sinal/ruído limites: sem ruído e com $\text{SNR} \sim 0.2$. Os resultados desta simulação podem ser vistos na Figura 6.13.

Figura 6.13: Comparação entre os ângulos de posição simulados e estimados pelo método descrito nesta seção para 1000 galáxias geradas em coordenadas aleatórias no céu. Em (a) sem ruído adicionado e em (b) com $\text{SNR} \sim 0.2$.

Estes resultados demonstram que mesmo com o problema de descontinuidade nos ângulos varridos pelo satélite em coordenadas individuais, do ponto de vista estatístico os ângulos de posição das galáxias são recuperados de forma confiável em quaisquer posições do céu pelo método descrito nesta seção, até mesmo em condições de sinal/ruído desfavoráveis.

6.3.2 Algoritmos Genéticos

Um Algoritmo Genético (ou GA, *Genetic Algorithm*) é uma heurística¹⁰ popularizada principalmente por **Holland (1975)**¹¹ que é utilizada para otimização genérica global de funções—o que quer dizer que em princípio um GA pode ser aplicado para encontrar pontos ótimos globais de quaisquer funções.¹²

O GA é baseado em transformar um problema matemático de otimização em um análogo do processo evolutivo biológico. Inicialmente uma população de soluções aleatórias é gerada dentro do espaço de parâmetros da função a otimizar: cada uma destas soluções candidatas é considerada um indivíduo. Então essa população passa por diversos processos imitando a evolução biológica, e a cada nova geração é garantido por um destes processos (clonagem) que pelo menos a melhor solução da geração anterior continuará existindo, e que melhores soluções podem aparecer.

Para isso, cada uma das soluções em um GA é codificada na forma de cromossomos, que são apenas representações do vetor de parâmetros (em forma binária ou de ponto

¹⁰Isso significa que um GA é uma técnica desenvolvida que encontra uma solução que não é provada ser a solução correta, mas que em geral (baseado em ‘experiência’) é uma ‘boa’ solução.

¹¹As primeiras utilizações de métodos evolutivos datam de **Barricelli (1954)** (apud **Fogel, 2006**). Para mais detalhes sobre o desenvolvimento histórico e a teoria de GAs, ver **Reeves & Rowe (2002)**, **Fogel (2006)** e **Sivanandam & Deepa (2007)**.

¹²No entanto, nada garante que ele seja o meio mais rápido para encontrar tal solução.

flutuante) que leva àquela solução. A função que se quer otimizar, geralmente é descrita como a função de *fitness*, e ela é calculada para cada indivíduo, o que significa que a função a ser otimizada deve ser calculada tantas vezes quanto existirem indivíduos nesta população. No caso particular do problema tratado neste capítulo, a função de fitness calcula uma norma L2 utilizando o modelo direto, os parâmetros deste modelo (que são os indivíduos da população) e os dados.

Após o cálculo de *fitness*, uma nova população é criada a partir de processos aplicados à população atual, representando uma nova geração, sendo este o ponto em que as implementações de GAs diferem entre si, e que pode ser modificado para otimizar com maior ou menor eficiência problemas distintos. A Figura 6.14 mostra de forma esquemática como um GA funciona.

Figura 6.14: Esquema conceitual de funcionamento de um algoritmo genético. Os parâmetros de configuração do GA não estão representados (tamanho da população, taxa de mutação, etc.).

Os processos, ou operações, geralmente mais básicas adotadas em GAs para evoluir as gerações são a clonagem, a mutação e a reprodução ou *crossover*. Como seus originais biológicos, estas operações podem agir em um ou mais indivíduos da população. A clonagem, por exemplo, age sob um único indivíduo, e representa apenas uma cópia de um ou mais membros da população entre uma geração e outra. Ela é aplicada para preservar a melhor solução entre duas gerações subsequentes.

Outra operação essencial em GAs é a reprodução, ou *crossover*. Os indivíduos com melhores valores de *fitness*, também chamados de indivíduos melhor adaptados, são o que possuem uma maior probabilidade de sofrer *crossover* se reproduzindo com outros indivíduos e gerando novos indivíduos, cujo genótipos são formados por partes dos genótipos dos dois indivíduos “pais”. Os detalhes de como esse processo ocorre, como a função de distribuição de probabilidade para a seleção, se um indivíduo pode se reproduzir mais de uma vez por geração, o ponto do genótipo que é trocado, etc., são dependentes da implementação do GA.

A última operação fundamental é a mutação, que pode ser de dois tipos: local ou global. Em uma mutação local um indivíduo da população pode ter seu genótipo alterado de forma completamente aleatória em um determinado ponto.¹³ Já em uma mutação global, todos os membros da população são afetados, mas quando esse processo é implementado em geral o indivíduo melhor adaptado não é afetado de modo a garantir na nova geração a existência de pelo menos um indivíduo tão bom quanto o melhor da geração mais antiga.

Os GAs são conhecidos por serem eficazes em encontrar a região do ponto ótimo, mas em geral são lentos para determinar este ponto. Segundo [Sivanandam & Deepa \(2007\)](#), para se resolver este problema adotam-se heurísticas baseadas em aliar um GA e um método mais rápido para otimização local, como algum método quase-Newton (esses são chamados de algoritmos meméticos). Neste trabalho adotamos um otimizador evolutivo disponível na linguagem R denominado **RGENOUD** ([Sekhon & Mebane, 1998](#); [Mebane & Sekhon, 2007](#)), que é baseado em aliar uma implementação de Algoritmo Genético com métodos implementados no otimizador **optim** (também do R). Para este trabalho, como método baseado em derivadas escolhemos utilizar o BFGS, apresentado na próxima seção.

6.3.3 BFGS

Proposto de forma independente por [Broyden \(1970\)](#), [Fletcher \(1970\)](#), [Goldfarb \(1970\)](#) e [Shanno \(1970\)](#), o BFGS é um dos algoritmos da família quase-Newton. Estes, são algoritmos baseados em uma aproximação de segunda ordem da função a ser otimizada, assim como no algoritmo de Newton. Contudo ao invés de utilizarem a matriz Hessiana, os algoritmos quase-Newton calculam uma aproximação desta matriz que é atualizada a cada nova iteração. Uma descrição detalhada destes métodos pode ser encontrada em [Nocedal & Wright \(1999\)](#).

Como no método de Newton, no BFGS a função a ser otimizada é aproximada por uma expansão em série de Taylor de segunda ordem:

$$f(\mathbf{x}_k + \Delta\mathbf{x}) \approx f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T \Delta\mathbf{x} + \frac{1}{2} \Delta\mathbf{x}^T \nabla^2 f(\mathbf{x}_k) \Delta\mathbf{x} \quad (6.7)$$

cuja derivada é expressa por:

¹³Dependendo da representação dos parâmetros do problema no genótipo isso significa uma modificação aleatória em um ou mais parâmetros.

$$\nabla f(\mathbf{x}_k + \Delta \mathbf{x}) \approx \nabla f(\mathbf{x}_k) + \nabla^2 f(\mathbf{x}_k) \Delta \mathbf{x} \quad (6.8)$$

Como no ponto de extremo o gradiente da função deve ser nulo, uma vez que se esteja em um ponto $\mathbf{x}_k$, a direção $\mathbf{p}_k$ em que o próximo ponto $\mathbf{x}_{k+1}$, mais próximo do extremo, deve ser buscada é dada por:

$$\Delta \mathbf{x} = \alpha_k \mathbf{p}_k = (-\nabla^2 f(\mathbf{x}_k))^{-1} \nabla f(\mathbf{x}_k) \quad (6.9)$$

Para resolver a equação acima sem que seja necessário calcular a cada iteração a matriz de derivadas segundas e suas inversas, o que em geral é caro computacionalmente, o algoritmo BFGS aproxima esta matriz por um processo iterativo, onde $B_k \approx \nabla^2 f(\mathbf{x}_k)$. As fórmulas do BFGS para a atualização de B_{k+1} e sua inversa, H_{k+1} , são escritas como:

$$B_{k+1} = B_k + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \Delta \mathbf{x}_k} - \frac{B_k \Delta \mathbf{x}_k (B_k \Delta \mathbf{x}_k)^T}{\Delta \mathbf{x}_k^T B_k \Delta \mathbf{x}_k} \quad (6.10)$$

$$H_{k+1} = B_{k+1}^{-1} = \left(I - \frac{\mathbf{y}_k \Delta \mathbf{x}_k^T}{\mathbf{y}_k^T \Delta \mathbf{x}_k} \right)^T H_k \left(I - \frac{\mathbf{y}_k \Delta \mathbf{x}_k^T}{\mathbf{y}_k^T \Delta \mathbf{x}_k} \right) + \frac{\mathbf{y}_k \Delta \mathbf{x}_k^T}{\mathbf{y}_k^T \Delta \mathbf{x}_k} \quad (6.11)$$

onde

$$\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k) \quad (6.12)$$

Basicamente, a partir da escolha de um ponto inicial $\mathbf{x}_0$ e uma aproximação B_0 (que pode ser $B_0 = I$), o BFGS realiza os seguintes passos até a convergência em $\mathbf{x}$:

1. Resolva a equação 6.9;
2. Realize uma busca em linha para determinar um passo α_k adequado na direção de $\mathbf{p}_k$;¹⁴
3. Atualize $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha \mathbf{p}_k$;
4. Calcule a equação 6.12;
5. Calcule a equação 6.11 e itere o processo.

O otimizador adotado neste trabalho de tese utiliza a implementação do BFGS disponível no R por meio da função `optim` (o código original foi implementado por [Nash, 1990](#)), que na ausência de derivadas parciais analíticas para a função f calcula numericamente o gradiente.¹⁵

¹⁴Para detalhes, ver [Nocedal & Wright \(1999\)](#).

¹⁵Isso é uma característica importante para a aplicação que estamos fazendo neste Capítulo, pois desejamos que o método possa lidar com quaisquer perfis.

6.4 Testes com simulações

Realizamos testes com simulações de observações de objetos utilizando dados tipo *Sky-Mapper* com a finalidade de verificar o quão eficaz é o método implementado com as idéias descritas neste Capítulo para a recuperação dos parâmetros morfológicos de galáxias a partir da decomposição do bojo e do disco por meio de ajustes de perfis teóricos. Estes testes envolveram a simulação de 1600 galáxias cujos parâmetros do bojo, do disco, e do fundo foram sorteados de forma completamente aleatória. Os intervalos de variação destes parâmetros podem ser vistos na Tabela 6.1.

Parâmetro	Mínimo	Máximo	Unidade
rd	400	2000	mas
rb	100	rd (max=2000)	mas
Ib	5*Fundo (min=5)	100	arbitrária
Id	Ib	1000	arbitrária
e	0.1	0.9	NA
Fundo	1	10	arbitrária

Tabela 6.1: Intervalo de variação dos parâmetros na simulação. Todas as galáxias possuem $C_0 = 0$ e $n = 4$.

Para realização dos testes, trânsitos tipo *Sky-Mapper* foram simulados seguindo a distribuição angular definida pela lei de varredura nominal da missão Gaia sob a posição $(l, b) = (160^\circ, 50^\circ)$.¹⁶ Os trânsitos simulados foram organizados em Senogramas nos quais foram aplicados os métodos descritos neste Capítulo para o ajuste dos perfis teóricos no espaço de Radon.

Comparações entre os resultados obtidos para os parâmetros das galáxias e os valores simulados para os diversos parâmetros são apresentados nas Figuras 6.15 e 6.16. Podemos verificar que do ponto de vista estatístico, todos os parâmetros são recuperados de forma estável. Contudo, também é possível observar nas Figuras 6.15 (a) e (b) que os parâmetros *Ib* e *rb* do bojo, apresentam uma dispersão que é relativamente grande. Os valores das medianas e do desvio absoluto da mediana¹⁷ para estes parâmetros são: $\text{med}(Ib) = 11 \pm 53\%$ e $\text{med}(rb) = -9 \pm 36\%$.

O grande valor de erro (tanto sistemático quanto estatístico) é devido provavelmente ao fato do código de otimização atualmente implementado considerar exclusivamente dados tipo *Sky-Mapper* no ajuste, enquanto que o sinal do bojo será predominante nas partes mais centrais da galáxia, que são justamente mal amostradas nas observações do *Sky-Mapper*. A maior parte do sinal do bojo, nestas simulações, está contida em poucos pixels (dois ou três, pois cada *sample* tipo SM tem ~ 200 mas), e mesmo os bojos simulados com maior parâmetro *rb*, são perturbados

¹⁶As ferramentas desenvolvida para este teste permite a realização de testes similares em quaisquer outras regiões do céu de forma completamente automática.

¹⁷Este é uma quantidade que indica a dispersão estatística de um conjunto de dados de forma robusta aos outliers (Venables & Ripley, 2002), sendo definida por: $\text{MAD} = \text{mediana}(|x_i - \text{mediana}(\mathbf{x})|)$, para um vetor $\mathbf{x}$.

pelo sinal predominante do disco em suas regiões mais externas. A Figura 6.16 (c) mostra que essa hipótese é plausível, visto que quanto maior o raio do bojo, menor parece ser a dispersão do erro fracional em torno de zero, o que também pode ser concluído verificando na Figura 6.15 (c) que a partir de ~ 600 mas os valores ajustados aparentemente tendem ao valor ideal (ainda que a dispersão continue grande).

De modo a compensar o erro no valor ajustado para o parâmetro rb , o código tende a super-estimar sistematicamente o valor da intensidade do bojo no raio ajustado, Ib , como pode ser visto nas Figuras 6.15 (a) e 6.16 (a). Para melhorar o ajuste da parte do perfil teórico correspondente a este componente estrutural deve-se incluir na otimização os dados de observações tipo *Astro-Field*, que possuem uma resolução espacial três vezes maior, logo que deverão possibilitar uma melhor amostragem das regiões mais críticas para esse ajuste.

A maior dispersão observada nos parâmetros do bojo não foi detectada justamente nos parâmetros recuperados para o perfil do disco, como pode ser visto nas Figuras 6.15 (b) e (d). Isso ocorre provavelmente em razão do disco ser bem amostrado pelas observações das janelas tipo *Sky-Mapper*. A relação entre os valores simulados e recuperados pelo código para os parâmetros dos discos é bastante linear, com uma dispersão baixa em relação ao valor ideal, além de ausência de efeitos sistemáticos significativos: $\text{med}(Id) = -1 \pm 7\%$ e $\text{med}(rd) = 1 \pm 4\%$.

Observando a Figura 6.15 (d), poderia-se pensar que os valores de rd se tornariam um pouco pior determinados quando o disco se torna maior, contudo ao verificar os erros percentuais deste parâmetro, representados na Figura 6.16 (d), observa-se que isso não ocorre, e que o erro fracional é aproximadamente constante em todo o intervalo estudado—contudo, em termos absolutos o valor do erro é maior quanto maior o disco. Já a intensidade central do disco, Id , possui um erro percentual menor o quanto maior valor este parâmetro assume, o que ocorre porque quanto mais sinal disponível, mais facilmente estes parâmetro é determinado (fato que pode ser igualmente observado na análise dos erros percentuais de Ib).

Os resultados mostram também que a elipticidade das galáxias é recuperada de forma extremamente linear, com dispersão essencialmente nula. Isso ocorreu principalmente pelo fato de que na versão atual do código de ajuste de perfil consideramos uma resolução máxima de 0.1 neste parâmetro, enquanto provavelmente o erro que este código está cometendo na recuperação do parâmetro é muito menor do que a resolução que foi considerada.

Finalmente, os resultados obtidos para o ajuste do fundo indicam que o código possui uma tendência sistemática a super-estimar os valores simulados, com $\text{med}(\text{background}) = 10 \pm 24\%$. Isso pode estar ocorrendo pois as galáxias que foram simuladas durante este teste são objetos cujo sinal é sempre muito maior do que o sinal proveniente do fundo. Deste modo, a determinação deste sinal constante aditivo correspondente ao fundo pode estar incorporando uma pequena fração do sinal dos outros componentes—devido ao valor muito pequeno do sinal do fundo em relação ao sinal do bojo e do disco, este efeito acaba não sendo percebido na análise de erro dos outros dois parâmetros.

Figura 6.15: Resultados obtidos a partir de um teste realizado com um conjunto de 1600 galáxias. As simulações foram realizadas na direção $(l, b) = (160^\circ, 50^\circ)$. Uma reta $x=y$ (em preto) indica a recuperação ideal dos parâmetros. O mapa de cores ao fundo dos gráficos mais dispersos indica a densidade de pontos.

Figura 6.16: Erros percentuais (no sentido recuperado menos simulado) nos parâmetros recuperados pelo código a partir das 1600 galáxias simuladas. O mapa de cores ao fundo é representativo da densidade de pontos nas regiões respectivas.

Este teste realizado com um conjunto estatisticamente representativo de galáxias simuladas do ponto de vista estrutural, demonstra que tanto as idéias descritas aqui para análise de perfil dos objetos extensos, quanto o próprio código protótipo implementado em R com a finalidade de testar a aplicação destas idéias, possibilitarão a análise de milhões de objetos extensos, revelando suas estruturas morfológicas. Isso permitirá a realização de estudos da decomposição bojo/disco de milhões de galáxias que somente podem ser observadas a partir do espaço, sem que seja necessária adoção de procedimentos de reconstrução detalhada de imagens. Notamos, entretanto, que a reconstrução das imagens continua a ser um passo necessário para a escolha do perfil mais apto ao estudo dos objetos (pois quanto mais parâmetros livres, mais fácil é ajustar um determinado sinal), além de que uma análise e reconstrução mais detalhadas de objetos que não apresentem bons ajustes com perfis conhecidos, talvez possa vir a revelar novas características do universo em que vivemos.

6.5 Aprimoramentos e possibilidades

Os métodos apresentados neste capítulo para resolver o problema de ajuste de perfis teóricos em dados da missão Gaia de galáxias não resolvidas podem ser aprimorados em alguns pontos. Um primeiro aprimoramento, bastante natural, e que como comentamos anteriormente poderá auxiliar no ajuste dos dados morfológicos das regiões mais centrais das galáxias, é a inclusão de informação de alta-resolução obtida a partir de janelas originadas de observações dos *Astro-Fields*. Com essa inclusão, o ajuste que atualmente é realizado apenas com dados tipo *Sky Mapper*, seria realizado de forma global, levando em conta toda a informação espacial disponível nas observações do Gaia. Isso poderá possibilitar não apenas um melhor ajuste para o bojo, mas também para quaisquer objetos com tamanhos angulares menores do que algumas poucas centenas de milissegundos de arco.

Uma outra possibilidade é a realização de um tratamento do sinal do espaço de Radon que possibilite um melhor aproveitamento da informação contida nas observações. Como já foi mostrado anteriormente, as observações do Gaia não amostram completamente este espaço na direção angular. Nos últimos anos, alguns métodos tem surgido para resolver um problema de ausência de informação (chamado de *inpainting*) em figuras ou séries temporais, ou para realizar extração de objetos de imagens sem deixar “buracos” (ex. [Criminisi et al, 2004](#); [Fadili et al, 2009](#); [Sato et al, 2010](#)). Esse problema é completamente equivalente ao preenchimento das regiões não observadas pelo Gaia no espaço de Radon. Um exemplo de *inpainting* utilizando o método desenvolvido por [Criminisi et al \(2004\)](#) em dados de um senograma simulado, pode ser visto na Figura 6.17.

Devido a regularidade do sinal de uma galáxia no espaço de Radon, é possível que a adoção de técnicas similares de *inpainting* possa vir a completar satisfatoriamente parte dos ângulos não observados pelo satélite aumentando a quantidade de dados disponíveis para análises subsequentes, como os procedimentos de estimativa inicial para o ajuste de perfil comentados neste capítulo—no caso deste tipo de utilização,

Figura 6.17: Exemplo de *inpainting* em um senograma para dados SM simulados. Em (a), o senograma completo, em (b) o senograma com os ângulos não observados indicados em verde, e em (c) o senograma resultante da reconstrução pelo método de *inpainting* descrito em [Criminisi et al \(2004\)](#).

os pixels recuperados por *inpainting* provavelmente deveriam possuir um peso menor, em relação aos pixels que representam dados reais.

Outro aprimoramento que deve ser adotado é a inclusão de efeitos de CTI, ou *Charge Transfer Inefficiency* (descrito brevemente no Capítulo 2), provavelmente a partir da utilização do modelo descrito em [Short \(2009\)](#). Isso é um efeito importante no Gaia, em especial para a obtenção de posições bem determinadas, e pode ser incluído na construção do modelo direto de forma bastante simples, com a aplicação do modelo a cada ângulo separadamente. A determinação dos parâmetros do modelo de CTI não seria realizada durante o processamento, sendo obtida a priori.

Além destes aprimoramentos, existem possibilidades de exploração do método desenvolvido neste Capítulo que transcendem sua aplicação aos dados do satélite Gaia. Um exemplo é a análise de observações com baixo sinal/ruído: a imagem de uma galáxia observada com muito baixo sinal/ruído poderia ser levada ao espaço de Radon, e analisada naquele espaço. Um exemplo pode ser visto na Figura 6.18, onde uma galáxia foi simulada com sinal/ruído igual a 1 e 0.2 (no pixel de máximo sinal) e seu sinal do espaço de Radon.

Esse exemplo mostra que mesmo no caso de um ruído na observação original cinco vezes maior do que o sinal do objeto, quando levado ao espaço de Radon o sinal do objeto é levemente amplificado, o que ocorre por ele ser “coerente”. Em princípio, isso poderia ser aplicado em estudos de *weak-lensing*, onde o que se procura justamente é analisar as correlações entre os ângulos de posicionamento de diversas galáxias espalhadas em certa direção do céu.

Utilizando melhores tratamentos do sinal do objeto no espaço de Radon, talvez seja até mesmo possível aplicar os métodos de estimativa e ajuste de parâmetros comentados neste Capítulo para analisar os perfis de galáxias nestas condições de muito baixo sinal/ruído—na Figura 6.19, apresentamos um exemplo com resultados para o ângulo de posição de mil galáxias simuladas com sinal/ruído igual a 1 e 0.2.

Figura 6.18: Exemplo de aplicação da análise do ângulo de posição no espaço de Radon para uma galáxia simulada com $PA=45^\circ$ e $e=0.6$. Em (a) para um caso com $SNR \sim 1$. Em (b) para $SNR \sim 0.2$.

Figura 6.19: Resultados para simulações de 1000 galáxias com $e=0.6$ em diversos ângulos de posição e da recuperação destes ângulos pela análise do senograma filtrado. Em (a) para $SNR \sim 1$. Em (b) e (c) para $SNR \sim 0.2$, com (c) mostrando um mapa de densidade no fundo.

6.6 Conclusões

Neste Capítulo apresentamos os perfis de brilho mais utilizados para decomposição estrutural de galáxias visando a estimativa dos parâmetros morfológicos dos principais componentes destes objetos: seu bojo e seu disco. Iniciamos comentando sobre os perfis analíticos e então discorremos sobre o método de ajuste destes perfis no espaço de Radon (já apresentado no Capítulo 2).

Mostramos passo a passo a construção de um modelo para o processo direto de observação pelo satélite Gaia, que é construído desde a amostragem dos perfis e de transformações geométricas até a amostragem das janelas que serão observadas nos diversos CCDs do plano focal do satélite. Este modelo pode ser adaptado de forma simples para quaisquer perfis de brilho.

Apresentamos um método de otimização de parâmetros aos dados observacionais, que é baseado em três etapas: estimativa inicial de parâmetros, otimização global com Algoritmos Genéticos e otimização local com o algoritmo BFGS. Apresentamos a estratégia que desenvolvemos para a determinação de parâmetros iniciais aplicada ao ângulo de posição do objeto, e mostramos testes realizados em 1000 galáxias simuladas em uma região e em 1000 galáxias simuladas em regiões aleatórias no céu—os resultados demonstraram que mesmo que o Gaia não observe um objeto a partir de todos os ângulos de visada, estatisticamente a maior parte dos objetos terá o ângulo de posição bem estimado mesmo em condições de sinal/ruído baixo. Discorremos brevemente sobre os Algoritmos Genéticos e sobre o BFGS.

Aplicamos então este método de otimização, implementado em R, em simulações de janelas tipo *Sky-Mapper* de 1600 galáxias cujos parâmetros selecionados aleatoriamente varreram um intervalo realista de condições—as simulações foram realizadas em uma coordenada determinada do céu: $(l, b) = (160^\circ, 50^\circ)$. Os resultados obtidos por nosso método demonstram que a recuperação dos parâmetros do bojo possui um erro de $\text{med}(Ib) = 11 \pm 53\%$ e $\text{med}(rb) = -9 \pm 36\%$, e do disco $\text{med}(Id) = -1 \pm 7\%$ e $\text{med}(rd) = 1 \pm 4\%$ —a menor precisão na recuperação dos parâmetros do bojo é provavelmente devida a este componente estrutural ser mal amostrado nas janelas tipo *Sky-Mapper*.

Finalmente, apresentamos possibilidades de aprimoramento do método desenvolvido (como a adição das janelas *Astro-Field*, o que deverá possibilitar uma melhor amostragem do bojo) e uma possibilidade de aplicação à dados além daqueles que serão obtidos por esse satélite, o que permitiria um possível aproveitamento científico de observações com sinal/ruído extremamente baixos, que aliás já foram ser obtidas em diversos surveys existentes.

Os resultados obtidos neste Capítulo demonstram que a partir dos dados do Gaia será possível a obtenção de informações sobre as características estruturais de galáxias angularmente pequenas (raios de centenas de mas, ou no máximo $2''$), sem a necessidade de reconstruções detalhadas de imagens. Isso permitirá realização de decomposições bojo disco em centenas de milhares, ou milhões, de objetos extragalácticos que somente poderão ter sua estrutura observada por missões espaciais ou por telescópios terrestres munidos de óptica adaptativa.

Conclusões

Ao iniciar esta tese existia uma dúvida sobre se seria possível fazer ciência com os dados dos objetos extensos que serão observados pela missão espacial Gaia. Ao terminá-la, podemos dizer que sim, isso será possível.

De forma geral, este é o mais importante resultado astronômico deste trabalho, pois significa que será possível estender um pouco mais o escopo desta já abrangente missão. Além disso, durante o doutorado buscamos concretizar essa possibilidade através do desenvolvimento dos vários códigos necessários à obtenção de resultados científicos. Implementamos códigos computacionais dentro dos *pipelines* de duas *Coordination Units* do Gaia (CU4 e CU5), seguindo tanto quanto possível as recomendações do DPAC e do CNES—e desde o final de 2010, estes códigos estão sendo mantidos e estendidos por um engenheiro de software, Laurent Galluccio, do *Observatoire de la Côte d’Azur*.

Como comentamos ao final da introdução, essa deverá ser a primeira vez que a humanidade terá a oportunidade de vislumbrar o céu inteiro até magnitudes relativamente altas ($G \sim 20$) com alta resolução (~ 180 mas), e precisamos aproveitar estes dados para fazer a melhor ciência possível.

A seguir são resumidos os principais tópicos tratados nesse documento bem como os principais resultados obtidos durante este trabalho de tese:

- **É possível reconstruir imagens com os dados do Gaia em todo céu**

No Capítulo 2, apresentamos a transformada de Radon, que é uma boa descrição matemática do modo de observação do satélite quando olhado sob o ponto de vista de um objeto. Utilizando um *toy model* discurremos sobre um método de inversão desta transformada, chamado *Filtered Backprojection*, e analisamos o quanto esse método é sensível à quantidade de dados disponíveis e ao ruído na determinação do sinal. Vimos que é possível reconstruir imagens de quaisquer objetos em todo o céu com uma cobertura de área mínima igual a $\sim 87\%$, podendo atingir 99% em determinadas regiões e 98% em quaisquer coordenadas com $|\beta| > 45^\circ$. Mostramos também que a distribuição relativa entre os ângulos de trânsito do satélite é tal que em todas as regiões do céu existirão ao menos um par de trânsitos com $\sim 90^\circ$ entre si.

- **O catálogo Gaia poderá ser estendido em dezenas de milhões de fontes pontuais, inclusive além da magnitude limite do instrumento**

No Capítulo 3 primeiramente realizamos uma estimativa preliminar para a quantidade de projeções ópticas a partir de dados do catálogo GSC 2.3.2. Então, realizamos uma análise detalhada que foi baseada em simulações com o modelo da Galáxia de Besançon, e que permitiu concluirmos que $1.6 \times 10^8 \pm 5.1 \times 10^7$ projeções ópticas poderão ser observadas no céu inteiro, considerando uma diferença de magnitude máxima de $\Delta m = 3$ entre a fonte primária e a secundária.

- **Desenvolvemos e implementamos um método que possibilita segregar fontes nas imagens reconstruídas com poucas detecções espúrias**

No Capítulo 3, apresentamos o método que desenvolvemos para analisar as imagens reconstruídas, chamado de *Educated Image Segregator* ou EIS—este método é capaz de aprender e se auto-adaptar de forma automática, quando apresentado a exemplos de regiões que se deseja analisar. Resultados obtidos a partir de reconstruções de simulações realizadas com um dos simuladores oficiais da missão (GIBIS) demonstraram que para um treinamento ideal do EIS, dentre os algoritmos disponíveis para reconstrução atualmente o melhor é o *BinOutliers*. Este resultou na maior porcentagem de fontes secundárias recuperadas (para primária com $G = 19$, $\Delta m_{max} \sim 3.0$, e para $G = 12$, $\Delta m_{max} \sim 4.5$) e a menor quantidade de detecções espúrias (até separações angulares de ~ 300 mas virtualmente sem detecções espúrias). Testes para regiões não treinadas demonstraram que será necessário possuir uma alta granulação da biblioteca de treinamento em coordenadas com $|\beta| < 45^\circ$. Nossa implementação faz parte do *pipeline* da DU18 – *Source Environment Analysis* da *Coordination Unit 5* do DPAC.

- **O Gaia deverá observar em torno de $\sim 10^7$ milhões de galáxias**

No Capítulo 4, apresentamos o código JStuff, que construímos baseado em um código previamente existente de E. Bertin para simular catálogos de galáxias com dadas características como funções de luminosidade e parâmetros cosmológicos. Este código encontra-se acoplado ao modelo de Universo do Gaia, DU3 da *Coordination Unit 2*. Simulações do DPAC que utilizaram este código permitiram com que fosse realizada uma estimativa para o número de galáxias que potencialmente poderão ser observadas pelo satélite: $\sim 6 \times 10^6$ galáxias até $G = 18.5$ e 3×10^7 até $G = 20$, com $\langle z \rangle \sim 0.17$ —estes resultados, contudo, são um limite superior, pois não incluem cortes que existirão na transmissão de dados. Mostramos ainda que este valor é compatível com resultados da extrapolação das densidades de objetos no Hubble *Medium Deep Field*.

- **Desenhamos um *pipeline* para reduzir dados das galáxias não resolvidas em estrelas que serão observadas pelo Gaia**

No Capítulo 1, apresentamos a estratégia que construímos durante o desenvolvimento desta tese para analisar os dados das galáxias que o Gaia deverá observar. Este *pipeline* é baseado em uma reconstrução de imagem, medida de parâmetros morfológicos na imagem, classificação morfológica e análise de perfil nos dados das janelas. Sua construção está sob a responsabilidade da DU70 – *Extended Objects* da *Coordination Unit 4* do DPAC.

- **Classificações morfológicas de galáxias poderão ser realizadas a partir de imagens reconstruídas**

No Capítulo 5, apresentamos parâmetros morfológicos (CASGM20) que serão medidos em imagens reconstruídas, além do sistema que implementamos para

o cálculo destes parâmetros. Testes demonstraram que o sistema é capaz de recuperar parâmetros de forma estável até em imagens formadas por apenas 21x21 pixels. A partir da análise de imagens de galáxias do Hubble *Deep Field North* e de galáxias simuladas (com o GIBIS) e reconstruídas, verificamos que os três tipos morfológicos (elípticas, espirais, e irregulares) ocupam locais distintos do espaço CASGM20, embora exista uma sobreposição significativa—mostramos que uma PCA não possibilita uma separação simples entre os três tipos. Nossa implementação faz parte do *pipeline* da DU70 – *Extended Objects* da *Coordination Unit 4* do DPAC.

Apresentamos o método de classificação chamado *Support Vector Machines*, e mostramos com testes que ele é capaz de diferenciar populações gaussianas atingindo as taxas de acurácia máximas teóricas. Aplicamos este sistema em parâmetros obtidos em imagens de galáxias próximas re-amostradas para tamanhos entre 81x81 e 11x11 pixels. Os resultados mostraram que em um cenário no qual a reconstrução de imagem for perfeita, e em que o conjunto de treinamento for construído sem erros, o valor de acurácia na classificação deve atingir $85\% \pm 3\%$. Aplicamos este sistema em imagens reconstruídas com o algoritmo *ShuffleStack* a partir de simulações GIBIS 7, e os resultados mostraram que sob um regime de reconstrução abaixo do ideal, mas com um conjunto de treinamento sem erros, é obtida uma taxa de sucesso para diferenciação entre tipos precoce/tardio superior a 80%, e entre as três classes morfológicas principais (elípticas, espirais e irregulares) de 79%, 56% e 74%. Finalmente, a partir de imagens de galáxias do *Hubble Deep Field North* mostramos que sob um regime de reconstrução de imagem ideal, mas com erros no conjunto de treinamento, são obtidas taxas de classificação entre tipos precoce/tardio de $\sim 85\%$, e nas três classes morfológicas principais de $\sim 81\%$, $\sim 64\%$ e $\sim 62\%$. A implementação final deste método está sob a responsabilidade da DU70 – *Extended Objects* da *Coordination Unit 4* do DPAC.

- **Será possível realizar decomposições bojo/disco diretamente a partir das janelas do Gaia**

No Capítulo 6, apresentamos como se pode realizar uma decomposição estrutural de galáxias em bojo e disco no espaço de Radon. Apresentamos um método de otimização de parâmetros aos dados observacionais baseado em um modelo que construímos para as observações do Gaia, uma estimativa inicial de parâmetros, Algoritmos Genéticos e um algoritmo quase-Newton. Testes com um protótipo da estimativa inicial de parâmetros demonstraram que ela permite a recuperação do ângulo de posição do objeto em quaisquer coordenadas da esfera celeste, mesmo em condições de baixo sinal/ruído. Testes em simulações de janelas tipo *Sky-Mapper* de 1600 galáxias demonstram que a recuperação dos parâmetros do bojo possui um erro de $\text{med}(I_b) = 11 \pm 53\%$ e $\text{med}(r_b) = -9 \pm 36\%$, e do disco $\text{med}(I_d) = -1 \pm 7\%$ e $\text{med}(r_d) = 1 \pm 4\%$ —a menor precisão na recuperação dos parâmetros do bojo é provavelmente devida

a este componente estrutural ser mal amostrado nas janelas consideradas. Finalmente, apresentamos possibilidades de aprimoramento do método desenvolvido (como adição de janelas *AstroField*, com maior resolução, na otimização), e discutimos sobre a possibilidade de sua aplicação em dados de outros levantamentos. A implementação final deste método está sob a responsabilidade da DU70 – *Extended Objects* da *Coordination Unit 4* do DPAC.

Considerações finais

Ao chegar ao final de um doutorado, é interessante perceber que o seu real resultado não é apenas o documento que foi produzido, nem mesmo somente a pesquisa que foi desenvolvida, mas sim a materialização de todo um processo de construção da pesquisa e da formação do próprio pesquisador que a produziu.

De certo modo, sinto que embora o grande plano do trabalho tivesse sido delineado em minha cabeça desde as fases iniciais do doutoramento, foi apenas ao desenvolvê-lo e realmente escrever este documento, ao transportar o que eram inicialmente apenas idéias e que posteriormente se tornaram resultados para o papel, que pude vislumbrar que não era este o ato mais importante, mas que este era sim um processo necessário para que o ato de pesquisa se formalizasse e materializasse em um corpo coerente—e isso, mesmo que o trabalho efetivo de pesquisa tenha sido finalizado antes de ser redigido.

Durante seu desenvolvimento, as idéias e até mesmo suas implementações em códigos computacionais, foram analisadas e testadas em grande parte através de diagramas puramente gráficos, alguns dos quais foram incluídos nos capítulos desta tese. Somente *a posteriori* é que elas foram sendo formalizadas, gerando o texto e finalmente o documento apresentado aqui.

Mas este documento corresponde a apenas um dos aspectos que desenvolvi durante meu trabalho de pesquisa nos últimos anos—sem dúvida representa o assunto ao qual dediquei a maior parte de meus esforços, mas que não foi de modo algum o único. Além do trabalho descrito aqui, visando o futuro da Astronomia por meio do aprimoramento do conhecimento científico a partir de uma maior exploração dos dados da missão Gaia, nos últimos anos também me dediquei a questões do presente desta ciência—em alguns casos, até mesmo aplicando técnicas que aprendi justamente para possibilitar os trabalhos feitos nesta tese. Como disse Picasso: *Quanto mais técnica você tem, menos você precisa se preocupar com ela.*¹⁸

Particularmente, pude colaborar com a Dra. Caroline Soubiran, e com meus orientadores Dra. Christine Ducourant e Dr. Ramachrisna Teixeira, no estudo de aglomerados abertos em busca de uma melhor caracterização destes objetos por meio da análise dos movimentos próprios de seus componentes (Krone-Martins et al, 2010b). Para realização deste estudo, que tratou de algumas dezenas de aglomerados na região $+11^\circ \leq \delta \leq +18^\circ$, utilizamos dados do catálogo de movimentos próprios precisos PM2000 (Ducourant et al, 2006), aos quais tive a oportunidade de aplicar diversos métodos interessantes—apesar de escolher otimizações similares àquela apresentada no Capítulo 6 desta tese para a obtenção dos resultados do artigo final, estudei muitos outros métodos, além de ter iniciado o desenvolvimento de um método próprio baseado em estimações detalhadas de densidade. Este estudo deverá ser continuado e até mesmo estendido no futuro, visando a utilização de outros tipos de dados conjuntamente aos movimentos próprios.

¹⁸ *The more technique you have, the less you have to worry about it.*

Pude também colaborar com meus dois orientadores, Dr. Ramachrisna Teixeira e Dra. Christine Ducourant, em estudos astrométricos que transcendem aqueles que desenvolvemos conjuntamente no contexto da missão Gaia—e que encontram-se descritos nesta tese. Além do trabalho com aglomerados já comentado no parágrafo anterior, ainda temos desenvolvido juntos pesquisas sobre associações estelares e objetos jovens, sendo que para isso desenvolvi simulações numéricas de aglomerados e do método de *traceback* (Ducourant et al, 2008), estratégias de ponto de convergência e estimativas de erro detalhadas pelo método de *bootstrap* (Teixeira et al, 2008) e classificações por k-NN (Teixeira et al, 2009).

Além disso, nos últimos meses comecei a olhar para um futuro ainda mais distante do que o Gaia, tendo participado nas discussões sobre o projeto NEAT – *Nearby Earth Astrometric Telescope*, que é uma nova missão espacial astrométrica proposta para a década de 2020. Atualmente, uma das áreas que parecem ter maior capacidade de contribuir para o crescimento do conhecimento humano é a descoberta e o estudo de planetas extra-solares com massa e raios iguais ou muito próximos aos da Terra e que se encontrem dentro da zona habitável de estrelas como nosso Sol. Isso, pois conhecemos apenas um ponto do Universo onde existe vida, e seria de grande valor até mesmo cultural descobrir tanto outros locais habitáveis quanto (por que não?) outras formas de vida. O NEAT visa a observação dos deslocamentos astrométricos causados por planetas telúricos com até meia massa terrestre nas 200 estrelas F, G e K mais próximas do Sistema Solar, e foi enviado à ESA na chamada Cosmic Vision 2 em dezembro de 2010.¹⁹

Um outro tema no qual me envolvi, menos relacionado à pesquisa astronômica em si, mas relacionado ao seu aprimoramento, foi auxiliar no desenvolvimento da infra-estrutura disponível para computação astronômica no Brasil. Especificamente, participei ativamente (junto com o Dr. Alex Carciofi) da redação de um projeto do INCT-A para capacitação da comunidade astronômica brasileira no uso de GPUs, e das discussões da modernização do parque computacional do Departamento de Astronomia do IAG-USP (que duraram quase um ano). Esta modernização (realizada com o apoio da FAPESP), além de possibilitar a criação de um Centro de Processamento de Dados, resultará na instalação de um computador de médio porte, com capacidade de processamento de ~20 TFlops (mais de 2300 *cores* de processamento) trazendo benefícios para toda a comunidade astronômica brasileira.²⁰

Além disso, durante esta tese não deixei (completamente) de lado um dos pontos que considero de importância central na pesquisa científica, e em especial no Brasil: a divulgação e educação científica. Participei de atividades do Clube de Astronomia de São Paulo, desde a organização de eventos ao público geral, como saraus de poesia e astrofotografia, até cursos voltados a divulgação da Astronomia e formação de astrônomos amadores. Também ministrei uma palestra na *Société Astronomique de*

¹⁹Uma descrição do NEAT, bem como o sumário do projeto que foi enviado à ESA, podem ser encontrados em <http://neat.obs.ujf-grenoble.fr>.

²⁰Uma estimativa apresentada em de Carvalho et al (2010), mostra que a quantidade de *cores* de uso exclusivo da comunidade astronômica no Brasil era de apenas 423 em 2009—esse número não inclui utilização de recursos compartilhados por meio de laboratórios nacionais.

Bordeaux e, junto com Christine Ducourant e Caroline Soubiran redigimos um artigo para o *Le Mensuel de l'Université*, colaborando com a divulgação da missão Gaia. Em Bordeaux, também participei das discussões para a elaboração da escultura *Ville-Satellite*, do artista plástico Joseph da Silva, instalada em suspensão na esplanada da *Cité Mondiale* por vários meses. Mantive ainda colaborações com o projeto Telescópios na Escola, que incentiva a utilização de telescópios robóticos na educação de crianças e adolescentes, no qual ajudei na especificação e testes do programa PInE²¹. Finalmente durante o Ano Internacional da Astronomia, em 2009, pude colaborar com o nó nacional brasileiro em diversos momentos, mas principalmente organizando o evento 100 Horas de Astronomia, escrevendo alguns textos para o blog *Cosmic Diary* e redigindo um pequeno capítulo para o livro *Postcards from the Edge of the Universe* (Cosmic Diary et al, 2010).

O doutoramento possibilitou com que todas essas atividades fossem desenvolvidas com uma qualidade muito melhor, pois exigiu um amadurecimento de conceitos principalmente relacionados à Astronomia, mas também à outras disciplinas (desde gerenciamento de projetos até de ciências da computação). Espero que com a evolução de seus projetos de doutoramento e redações de suas teses, doutorandos sempre aprendam tanto quanto eu pude (e precisei) nos últimos quatro anos. Isso, pois embora alguns aspectos do trabalho e das questões que colocamos no início da tese possam ter sido respondidos dentro dos quatro anos dedicados ao seu desenvolvimento, existe toda uma infinidade de problemas astronômicos que continuam em aberto. Como dizia o slogan mundial do Ano Internacional da Astronomia 2009:

*The Universe: Yours to Discover.*²²

²¹Processamento de Imagens na Escola: <http://www.telescopiosnaescola.pro.br/pine/PInE.html>

²²Ou na versão do IYA2009-Brasil, *O Universo: Para você descobrir*.

Módulo inverso

A.1 Definição

A operação de “módulo inverso” é definida como:

Definição 19 *Sejam $a \in \mathbb{R}$, $b \in \mathbb{R}^+$ e l o quociente da divisão inteira de a por b (de modo que $l \in \mathbb{Z}^+$), definimos operação módulo inverso de a sob b , escrita como $c = a(\text{modinv } b)$:*

$$c = \begin{cases} (l+1)b - |a| & \text{se } l \text{ ímpar} \\ |a| - lb & \text{se } l \text{ par} \end{cases} \quad (\text{A.1})$$

Para finalidades práticas, essa operação é parecida com a operação módulo, caso esta última estivesse definida para $\mathbb{R}$. A principal diferença conceitual é de que na operação módulo inverso, após atingir-se o valor b , inicia-se uma “contagem” retrógrada, enquanto na operação módulo a contagem se reinicia de 0. A Figura A.1 exemplifica a diferença entre as duas operações.

Figura A.1: Exemplo de “módulo” e “módulo inverso” sob 10, aplicados à valores reais no intervalo $[-20; 20]$.

Aglomerados abertos

Como nós descrevemos nas considerações finais deste trabalho de tese, nos últimos anos desenvolvi um trabalho em colaboração com Caroline Soubiran, Christine Ducourant e Ramachrisna Teixeira, que consiste no estudo de aglomerados abertos em busca de uma melhor caracterização destes objetos por meio da análise dos movimentos próprios de suas estrelas membro (Krone-Martins et al, 2010b), além de servir como um prelúdio para uma exploração futura de dados do satélite Gaia.

Aglomerados abertos consistem em ferramentas extremamente interessantes para realizarmos estudos de nossa galáxia, bem como da Astrofísica estelar em geral, desde que o problema da seleção de seus membros físicos seja resolvido. Deste modo, uma das questões mais básicas no estudo de aglomerados é a determinação das estrelas que definem estes objetos.

Para isso, é necessário realizar alguma hipótese sobre o que são fisicamente tais objetos, sendo que costuma-se definir que um aglomerado é um grupo de estrelas localizado em uma região relativamente pequena do espaço e que possui um mesmo vetor velocidade espacial, salvo por erros de medida e uma pequena dispersão intrínseca. Devido à pequena extensão espacial da maior parte dos aglomerados, após projeção na esfera esse critério acaba sendo traduzido por vetores de movimento próprios similares entre as estrelas membro.

No estudo Krone-Martins et al (2010b) adotamos a definição comentada acima para estudar dezenas de aglomerados na região $+11^\circ \leq \delta \leq +18^\circ$ a partir de dados do catálogo de movimentos próprios precisos PM2000 (Ducourant et al, 2006). Neste estudo escolhemos adotar parametrizações de PDFs (*probability distribution functions*) para os movimentos próprios e otimizar os parâmetros destas PDFs por meio de técnicas similares àquelas apresentadas no Capítulo 6 desta tese para a obtenção dos resultados apresentados no artigo final.

Contudo, no decorrer desta colaboração analisamos a adoção de diversos métodos para a determinação de probabilidades de pertinência de estrelas aos aglomerados, chegando até mesmo a desenvolver um método novo baseado em estimações detalhadas de densidade em quaisquer espaços de dados. A partir deste novo método, este estudo deverá ser continuado e estendido no futuro, visando a utilização de outros tipos de dados como velocidades radiais, cores e paralaxes, por exemplo, conjuntamente aos movimentos próprios.

Nas páginas seguintes, anexamos o artigo que detalha o desenvolvimento deste trabalho sobre aglomerados abertos, bem como suas conclusões.

Kinematic parameters and membership probabilities of open clusters in the Bordeaux PM2000 catalogue[★]

A. Krone-Martins^{1,2}, C. Soubiran², C. Ducourant^{2,1}, R. Teixeira^{1,2}, and J. F. Le Campion²

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 differentiate 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](ftp://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

$$\Phi_t = \Phi_c \Psi_c + \Phi_f \Psi_f, \quad (1)$$

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

$$\begin{aligned} \Phi(\mu_{\alpha,i}, \mu_{\delta,i}) = & \frac{1}{2\pi(1-\rho^2)^{1/2}(\sigma_{\mu_\alpha}^2 + \epsilon_{\mu_{\alpha,i}}^2)^{1/2}(\sigma_{\mu_\delta}^2 + \epsilon_{\mu_{\delta,i}}^2)^{1/2}} \\ & \times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(\mu_{\alpha,i} - \mu_\alpha)^2}{\sigma_{\mu_\alpha}^2 + \epsilon_{\mu_{\alpha,i}}^2} + \frac{(\mu_{\delta,i} - \mu_\delta)^2}{\sigma_{\mu_\delta}^2 + \epsilon_{\mu_{\delta,i}}^2} \right. \right. \\ & \left. \left. - \frac{2\rho(\mu_{\alpha,i} - \mu_\alpha)(\mu_{\delta,i} - \mu_\delta)}{(\sigma_{\mu_\alpha}^2 + \epsilon_{\mu_{\alpha,i}}^2)^{1/2}(\sigma_{\mu_\delta}^2 + \epsilon_{\mu_{\delta,i}}^2)^{1/2}} \right] \right\}. \end{aligned} \quad (2)$$

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)
Dolidzim 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)
Dolidzim 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

$$\Psi_c = \begin{cases} n_c & , 2D \\ \frac{1}{1+g^{-1} \exp \left\{ -\frac{1}{2(1-\rho_{\text{pos}}^2)} \left[\frac{(\alpha_i - \alpha)^2}{\sigma_\alpha^2} + \frac{(\delta_i - \delta)^2}{\sigma_\delta^2} - \frac{2\rho_{\text{pos}}(\alpha_i - \alpha)(\delta_i - \delta)}{\sigma_\alpha \sigma_\delta} \right] \right\}} & , 4D \end{cases} \quad (3)$$

$$\Psi_f = \begin{cases} (1 - n_c) & , 2D \\ \frac{1}{1+g \exp^{-1} \left\{ -\frac{1}{2(1-\rho_{\text{pos}}^2)} \left[\frac{(\alpha_i - \alpha)^2}{\sigma_\alpha^2} + \frac{(\delta_i - \delta)^2}{\sigma_\delta^2} - \frac{2\rho_{\text{pos}}(\alpha_i - \alpha)(\delta_i - \delta)}{\sigma_\alpha \sigma_\delta} \right] \right\}} & , 4D. \end{cases} \quad (4)$$

In Eq. (2) above, $\mu_{\alpha,i}$, $\mu_{\delta,i}$ represents the proper motion of the i th star, $\epsilon_{\mu_{\alpha,i}}$, $\epsilon_{\mu_{\delta,i}}$ the proper motion errors of the i th 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 i th star, α , δ are the position of the cluster, σ_α ,

σ_δ represent the dispersion, and ρ_{pos} the correlation coefficient of the position distribution.

The parametrisations presented above leave us with a vector Θ 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

$$\mathcal{L} = \prod_{i=1}^N \Phi_t(\alpha_i, \delta_i, \mu_{\alpha,i}, \mu_{\delta,i}, \Theta). \quad (5)$$

Following the maximum likelihood principle the most probable Θ is that for which $\mathcal{L}$ takes its maximum value. A monotone transformation does not affect a maximum, and we can therefore write $\hat{\Theta} = \arg_{\Theta} \max \mathcal{L} = \arg_{\Theta} \max \hat{\mathcal{L}}$, where

$$\hat{\mathcal{L}} = \sum_{i=1}^N \log \Phi_t(\alpha_i, \delta_i, \mu_{\alpha,i}, \mu_{\delta,i}, \Theta). \quad (6)$$

Usually the $\hat{\Theta}$ is found to be the solution of a non-linear system of equations constructed from $\partial \hat{\mathcal{L}} / \partial \Theta = 0$, 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 $M_{\alpha,\delta,\mu_\alpha,\mu_\delta}$ of all the stellar positions and proper motions in the region, we adopted the following constraining conditions for the unknown parameters:

$$\hat{\Theta} \text{ is such that } \left\{ \begin{array}{l} \min(M_{\mu_\alpha}) \leq \mu_{\alpha,c/f} \leq \max(M_{\mu_\alpha}) \\ \min(M_{\mu_\delta}) \leq \mu_{\delta,c/f} \leq \max(M_{\mu_\delta}) \\ 0.01 \leq \sigma_{\mu_{\alpha,c}} \leq 5.00 \\ 0.01 \leq \sigma_{\mu_{\alpha,f}} \leq 6 \cdot \sigma(M_{\mu_\alpha}) \\ 0.01 \leq \sigma_{\mu_{\delta,c}} \leq 5.00 \\ 0.01 \leq \sigma_{\mu_{\delta,f}} \leq 6 \cdot \sigma(M_{\mu_\delta}) \\ -0.99 \leq \rho_{c/f} \leq 0.99 \end{array} \right. \quad (7)$$

$$\left\{ \begin{array}{ll} 0 \leq n_c \leq 1, & \text{if 2D} \\ 0.01 \leq g \leq \text{length}(M_\alpha), & \text{if 4D} \\ \min(M_\alpha) \leq \alpha \leq \max(M_\alpha) & '' \\ \min(M_\delta) \leq \delta \leq \max(M_\delta) & '' \\ 0.01 \leq \sigma_\alpha \leq 6 \cdot \sigma(M_\alpha) & '' \\ 0.01 \leq \sigma_\delta \leq 6 \cdot \sigma(M_\delta) & '' \\ -0.99 \leq \rho_{\text{pos}} \leq 0.99 & '' \end{array} \right.$$

where $\sigma(M_{\alpha,\delta,\mu_\alpha,\mu_\delta})$ 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

$$\hat{\Theta}_{\text{var}} \approx \text{diag}(-\mathcal{H}^{-1}), \quad (8)$$

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 i th star

$$p_{i(\mu_{\alpha,i}, \mu_{\delta,i}, \Theta)} = \frac{\Psi_c \Phi_c}{\Psi_c \Phi_c + \Psi_f \Phi_f}. \quad (9)$$

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 $N_{\text{opt}} * n_c$ most probable stars (where N_{opt} is created. From this list of most probable stars, the probability of the least probable star P_{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 method	0.252 ± 0.02	0.25	-0.94			
		± 0.11	± 0.07			
		2.33	-4.08	5.11	5.93	-0.04
BN04	0.261 ± 0.02	± 0.24	± 0.27	± 0.18	± 0.20	± 0.04
		0.29	-0.96			
		± 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

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.

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.

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 $\sim 87\%$ of our members that could be listed as members in Y08 were listed as such, while only $\sim 68\%$ 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 $\sim 68\%$ 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 $\sim 65\%$ of our members that could be listed as members in D06 were listed as such, while $\sim 68\%$ 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 $\mu\text{as yr}^{-1}$ 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.

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) cata-

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.

logues (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, 6738¹, 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_{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_{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_{memb}) are adimensional, 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	s	α_{J2000}	δ_{J2000}	''	σ_{α}	σ_{δ}	$\mu_{\alpha} \cos \delta$	μ_{δ}	$\sigma_{\mu_{\alpha} \cos \delta}$	$\sigma_{\mu_{\delta}}$	V_M	σ_{V_M}	2MASS	Prob.
								arcsec	arcsec	mas yr ⁻¹	mas yr ⁻¹	mas yr ⁻¹	mas yr ⁻¹	mag	mag		%
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 different², 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 contrast, 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 $340 \times 540 \text{ pc}^2$ (in the l, b directions). These authors suggested that this $\sim 5 \text{ Myr}$ 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 $(3.02 \pm 0.61; -1.64 \pm 0.48) \text{ mas yr}^{-1}$. 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 ($-0.06 \pm 0.99; -4.38 \pm 1.78) \text{ mas yr}^{-1}$. 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 $\sim 13.8 \pm 2.4 \text{ km s}^{-1}$. Kim & Koo (2000), determine the expansion velocity of the shell to be $\sim 15 \text{ km s}^{-1}$, 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 $\sim 2 \text{ Gyr}$ 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 flagged 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., & Lfbero, 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 Press)
- 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., Réquière, 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

Esta tese fez uso direto e recorrente dos seguintes recursos:

- **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 Itaotec, Intel Xeon 3.2 GHz, 21 nós de 2x1 cores, 42 Gb RAM total), GINA@INCT-A (cluster SGI, Intel Xeon 2.66 Ghz, 2 nós 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 (salvou este trabalho em certa ocasião).

- **Software (Organização)**

Apple iCal, Apple FileMerge, OmniPlan, OmniOutliner.

- **Software (Rede e comunicação)**

Apple Safari, Firefox, FoxyProxy, SSH Tunnel Manager, ExpanDrive, Apple Mail, Skype, Marratech.

- **Software (Editoração)**

Apple Pages & KeyNote 06, Apple Preview, L^AT_EX, Tikz & PGF 2.0, tikz-3dplot, JpgfDraw, TeXShop, LaTeXiT, Papers (indexar os artigos), BibDesk (indexar os livros), Skim, Adobe Acrobat 7 Professional, OmniGraffle.

- **Software (Pesquisa e desenvolvimento)**

ESO Virgo, Aladin, DS9, Eclipse Europa, Netbeans (durante o desenvolvimento do JStuff), Java 1.4, 1.5, 1.6 (versões Apple, Sun e SoyLatte), GridGain Open Cloud Computing (2.1.1), JFreeChart, JTS - Java Topology Suite, GaiaSimu Java libraries, GaiaTools Java libraries, SVN + Cornerstone (e svnX no início), TextWrangler, 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 PROPRIIS, 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>. 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
- BIJAOU, 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
- BIJAOU, 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 13th 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 Doctorale Sciences Fondamentales et Appliquées, 2004. 47, 50, 51, 104
- DOLLET, C.; BIJAOU, 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 ($+11\text{dg} < \Delta < +18\text{dg}$). *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](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
- GAIA TOOLS COMMITTEE: Gaia Tools Software User Manual. *GAIA-C1-UG-ESAC-MTL-015-01*, 2010. 105
- 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. 38
- 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 $z \sim 1$ 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 $z \sim 2$ 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, 137, 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&A 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. 125, 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. 69
- LAWRENCE, J. S.; ASHLEY, M. C. B.; TOKOVININ, A. & TRAVOUIL-LON, 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.; BERGEEST, 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.; BIJAOU, 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_](http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1992A%26A...258..165M&link_type=ABSTRACT)
[query?bibcode=1992A%26A...258..165M&link_type=ABSTRACT](http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1992A%26A...258..165M&link_type=ABSTRACT). 12
- MURTAGH, F.; STARCK, J.-L. & BIJAOU, 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/](http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1995A%26AS...112..179M&link_type=ABSTRACT)
[nph-data_query?bibcode=1995A%26AS...112..179M&link_type=ABSTRACT](http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1995A%26AS...112..179M&link_type=ABSTRACT). 182
- NAIM, A.; LAHAV, O.; SODRE, L. & STORRIE-LOMBARDI, M. C.: Automa-
ted 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?](http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1995MNRAS.275..567N&link_type=ABSTRACT)
[bibcode=1995MNRAS.275..567N&link_type=ABSTRACT](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=](http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010arXiv1009.3110P&link_type=ABSTRACT)
[2010arXiv1009.3110P&link_type=ABSTRACT](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://](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4307775)
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/](http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2002AJ....124..266P&link_type=ABSTRACT)
[nph-data_query?bibcode=2002AJ....124..266P&link_type=ABSTRACT](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. 40
- RADON, J.: Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Berichte der Sächsischen Akademie 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 On-going 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, 125
- 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. 40, 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. *Boletín 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.; BIJAOU, 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 University of Denmark, 1996. URL http://eivind.imm.dtu.dk/publications/PhD_thesis/PeterToft_PhD_thesis_1.ps.gz. 38, 40

- 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
- URNSHEK, 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-BORSENBARGER, 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