Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

Identificação de candidatas a galáxias interagentes no infravermelho próximo a baixos redshifts

Stanley Miranda Cerqueira

Feira de Santana

Agosto 2016

Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

Stanley Miranda Cerqueira

Identificação de candidatas a galáxias interagentes no infravermelho próximo a baixos redshifts

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Orientador: Prof. Dr. Eduardo Brescansin de Amôres

Feira de Santana

Agosto 2016

Stanley Miranda Cerqueira

Identificação de candidatas a galáxias interagentes no infravermelho próximo a baixos redshifts

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Feira de Santana, 26 de agosto de 2016

BANCA EXAMINADORA

Prof. Dr. Eduardo Brescansin de Amôres (Orientador)
PGCA/UEFS

Profa. Dra. Priscila Freitas Lemes Lourenço
UNIVAP

Prof. Dr. Angelo Amâncio Duarte
PGCA/UEFS

Ficha Catalográfica – Biblioteca Central Julieta Carteado

Cerqueira, Stanley Miranda
C396i Identificação de candidatas a galáxias interagentes no infravermelho próximo a baixos *redshifts*./ Stanley Miranda Cerqueira. Feira de Santana, 2016.
98f.: il.

Orientador: Eduardo Brescansin de Amôres
Dissertação (mestrado) – Universidade Estadual de Feira de Santana, Programa de Pós-Graduação em Computação Aplicada, 2016.

1.Galáxias interagentes. 2.Software – Reconhecimento de padrões. I.Amôres, Eduardo Brescansin (orient.). II.Universidade Estadual de Feira de Santana. III. Titulo.

CDU : 004:524.77

Abstract

Interacting peculiar galaxies are objects that still require further studies because they play an important role in the processes of the evolution of galaxies. The knowledge of its location and properties constitutes a great benefit to the astronomical community. In this dissertation, we present an automatic method for identifying and classifying images of interacting galaxies at low redshifts for the Southern Hemisphere, based on the properties of stellar and interstellar extinction distribution as well as using a pattern recognition software called "Wndchrm" on images from the 2MASS survey, in the near infrared for the filters: J, H and K_S. The training phase was made with images of known interacting galaxies from the Arp & Madore Catalogue, Categories 1 and 2. After training, a validation was performed using images of a region of the sky with 573 square degrees, obtaining a hit of, approximately, 73% in the identification of galaxies identified by visual inspection as interacting. This rate can reach up 88% considering the comparison with previous know galaxy pairs of Category 2 of Arp & Madore Catalogue. The procedure was performed for an area of about 17.836 square degrees of the Southern Hemisphere, finding at least several hundred galaxy pairs as yet uncatalogued.

Keywords: interacting galaxies, large astronomical surveys, pattern recognition.

Resumo

As galáxias peculiares interagentes são objetos que ainda carecem de maiores estudos. Elas desempenham um papel importante nos processos de formação e de evolução das galáxias. O conhecimento de sua localização e propriedades, constitui um grande benefício para a comunidade astronômica. Nesta dissertação, apresentamos um método automático para identificar e classificar imagens de galáxias interagentes a baixos redshifts para o Hemisfério Sul. Para tal, estabelecemos critérios baseados na distribuição de estrelas e extinção interestelar, assim como de objetos identificados no Catálogo de fontes extensas do 2MASS. Usamos um software de reconhecimento de padrões chamado Wndchrm em imagens do grande levantamento 2MASS no infravermelho próximo para os filtros: J, H e Ks. A fase de treinamento foi feita com imagens de galáxias interagentes conhecidas, do Catálogo de Arp & Madore, das categorias 1 e 2. Após o treinamento foi realizada uma validação com imagens de uma região do céu de 573 graus quadrados, obtendo uma taxa de acerto de, aproximadamente 73% na identificação de galáxias previamente identificadas de forma visual como interagentes. Essa taxa aumenta para, aproximadamente 88% levando em conta pares da Categoria 2 previamente identificados no Catálogo de Arp & Madore como pertencentes à essa categoria. Executamos, o procedimento para uma área de, aproximadamente 17.836 graus quadrados do Hemisfério Sul, encontrando, ao menos, várias centenas de pares de galáxias ainda não catalogados.

Palavras-chave: galáxias interagentes, grandes levantamentos astronômicos, reconhecimento de padrões.

Prefácio

Esta dissertação de mestrado foi submetida a Universidade Estadual de Feira de Santana (UEFS) como requisito parcial para obtenção do grau de Mestre em Computação Aplicada.

A dissertação foi desenvolvida dentro do Programa de Pós-Graduação em Computação Aplicada (PGCA) tendo como orientador o Dr. Eduardo Brescansin de Amôres.

Agradecimentos

A Deus, pois é minha força, meu auxilio e minha salvação, em todos os momentos de minha vida, assim como foi durante a realização deste projeto.

A meu orientador, Prof. Dr. Eduardo Brescansin de Amôres, pela orientação, empenho, atenção, cuidados e incentivos dedicados a mim, durante todo esse tempo.

Ao PGCA, pela oportunidade de fazer o curso; em especial, ao Prof Dr. Ângelo Amâncio Duarte, que, em alguns momentos de dificuldades, me concedeu bons conselhos e incentivo, além das importantes sugestões feitas em meu Exame de Qualificação; e ao Prof. Dr. João Bosco Gertrudes, por sugestões dadas em relação ao projeto deste trabalho. Também agradeço à ambos por aceitarem em fazer parte de minha banca.

A minha querida e amada esposa, pela paciência, por acreditar em mim e pelo incentivo constante durante todo processo. A meus filhos, por esperar e compreender pacientemente a ausência nos vários momentos que isto foi necessário.

Minha mãe, por acreditar, incentivar e orar sempre por mim. Devo tudo a você, mãe!!

Minha tia, Profa. Vânia G. Miranda, pelo paciente trabalho de revisão da redação de boa parte deste trabalho.

A todos da Coordenação de Seleção e Admissão da UEFS, especialmente a pessoa do meu coordenador, Prof. Dr. Carlos Bichara, por concordar, incentivar e me aconselhar durante o tempo que durou esse trabalho.

Agradeço à L. Shamir por disponibilizar publicamente seu software, o Wndchrm para uso.

Agradeço ao INCT-A por possibilitar a utilização de seu cluster: SGI Altix ICE 8400. O que veio a permitir a conclusão deste trabalho em tempo hábil.

Ao colega George Oliveira Barros, que me dispensou atenção e ajuda quando necessário.

Ao Prof. Dr. Maximiliano Luis Faúndez-Abans, pelas discussões e sugestões que contribuíram com esse trabalho.

A todos que, direta ou indiretamente, fizeram parte da minha formação, o meu muito obrigado.

Este trabalho utilizou os equipamentos do Laboratório de Astroinformática (IAG/USP, NAT/Unicsul), cuja aquisição foi possibilitada pela agência brasileira FAPESP (processo 2009/54006-4) e pelo INCT-A.

This research has made use of the NASA/IPAC Extragalactic Database (NED) which

is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

This publication made use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation.

This research has made use of "Aladin sky atlas" developed at CDS, Strasbourg Observatory, France.

We acknowledge the usage of the HyperLeda database (http://leda.univ-lyon1.fr).

Sumário

Lista de Publicações

"Classificação de galáxias interagentes próximas no infravermelho a baixos redshifts"; apresentado e publicado nos anais do III WPOS/ERBASE 2016, ISSN 2177-4692. Autores: Cerqueira, S. M.; Amôres, E. B.; Faúndez-Abans, M. A.; da Rocha Poppe, P. C.; Martins, F. V. A.; Oliveira-Abans, M.

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Edwin Hubble propôs uma classificação para galáxias que leva em consideração a sua forma aparente [Hubble 1926]. De acordo com a classificação de Hubble Figura 1.1, uma galáxia pode ser elíptica (E), lenticular, espiral (S), espiral barrada (SB) ou irregular. Para cada uma dessas classificações, existem subgrupos. No grupo das elípticas, temos os subgrupos de E_0 a E_7[1], sendo as mais próximas de E_0 mais circulares na direção de um formato mais elipsoide até chegar a E_7. As lenticulares têm um bojo pequeno no formato de uma lentilha, dando a impressão que está envolta em uma nuvem. As espirais normais têm braços espiralados saindo diretamente do bojo central. Podem ser Sa, Sb ou Sc, de acordo com a formação definida de seus braços. As espirais barradas têm os seus braços espirais saindo da extremidade da barra [Lépine e Leroy 2000; López-Corredoira et al. 2001; Amôres et al. 2013, entre outros], uma estrutura composta de gás, poeira e estrelas predominantemente jovens que têm início no centro da Galáxia. De acordo com o tipo e definição da estrutura, uma galáxia espiral barrada pode ser SBa, SBb ou SBc também dependendo da definição dos seus braços espirais, conforme mostrado na Figura 1.1. Por fim, temos as galáxias irregulares que não têm uma forma definida. Exemplos típicos são a Pequena e a Grande Nuvem de Magalhães, que podem ser vistas a olho nu.

Segundo Naim e Lahav (1997), o conceito de galáxias peculiares está atrelado ao tipo de peculiaridade em questão. Por exemplo, se existe, ou não, um jato ou um grau de assimetria, um anel, uma quantidade atípica de poeira interestelar, etc. Por isso deve-se determinar que parâmetros devem ser usados, a fim de se definirem funções de morfologia de galáxias para todos os redshifts[2] (desvio para o vermelho, indicado por z) e todas as inclinações. No que diz respeito à morfologia, muita confusão ainda é feita com os termos "peculiar" e "irregular". De acordo com a Sequência de Hubble, galáxias irregulares encontram-se posteriores aos tipos tardios desta sequência. Elas são bastante fracas e distorcidas e têm pouco ou nenhum bojo.

As galáxias peculiares podem ser entendidas como todas aquelas cuja morfologia não se parece com nenhuma das do tipo Hubble. Por causa da grande quantidade de

[1] O subscrito n é calculado através da relação entre o semi eixo maior (a) e menor (b) da elipse por meio da relação: En = (a-b)/a

[2] É uma medida da velocidade de um objeto relativa a nós. Usado como indicador de distâncias. http://www.telescopiosnaescola.pro.br/hubble.pdf

características que podem diferenciar uma galáxia das do tipo Hubble, é difícil ter uma única definição. Uma saída utilizada por Naim e Lahav (1997), é definir medidas quantitativas para definir galáxias peculiares.

Em 1959, Vorontsov-Velyaminov (1959) catalogou galáxias que, na época, se acreditava, estavam interagindo. Desse trabalho, surgiu o Atlas e Catálogo de galáxias em interação contendo 355 sistemas, sendo a maioria dos sistemas encontrados pelo autor no levantamento do céu do Observatório de Palomar [Vorontsov-Velyaminov 1959]. Para Vorontsov-Velyaminov, um sistema de galáxias era considerado peculiar quando a forma regular de uma delas era alterada, aparentemente como consequência da interação com outra, ou quando galáxias estão envolvidas em uma mesma estrutura [Vorontsov-Velyaminov 1977].

Em 1962, Vorontsov compilou o Catalogo Morfológico de Galáxias, finalizado em 1974 com 4 volumes, (MCG) [Vorontsov-Velyaminov et al. 1962 - 1974], o qual contém muita informação sobre a descrição de pares de galáxias interagentes. No total 1.449 pares de galáxias interagentes foram catalogados e mensurados. Apesar de se obterem imagens por meio do Atlas do Céu de Palomar com uma excelente qualidade, tais só puderam ser publicadas em uma qualidade muito inferior, por conta do equipamento tipográfico disponível na época.

Entre as décadas de 1960 e 1980, Arp & Madore [Arp e Madore 1987], daqui por diante, AM87, publicaram dois catálogos que continham observações de galáxias peculiares para o Hemisfério Norte e Sul, respectivamente.

Apesar de muitos estudos isolados terem sido feitos com esses objetos, não existe até o presente momento um estudo abordando as características e propriedades globais de todas as categorias, como p.ex., as observações nos vários filtros e comprimentos de onda obtidos com os grandes levantamentos realizados para todo o céu nas últimas três décadas. Dentre alguns trabalhos que contém dados de redshifts, também para essas galáxias, podemos citar, o Levantamento Stromlo-APM Redshift versões I, II, III e IV [Loveday et al. 1992, 1995 e 1996].

Da análise do Altas AM87, em termos amplos parecem existir dois tipos de peculiaridades de morfologia:

- moderadas – desvios dos padrões normais esperados para uma galáxia de algum dos tipos da sequência de Hubble, mas que permite a classificação em alguns dos tipos, p.ex., faixas de poeiras nas elípticas e três braços espirais.
- peculiaridades fortes – que são graves perturbações na aparência das galáxias de modo que a tentativa de classificação na sequência de Hubble seja inconclusiva ou impossível de ser feita.

Na maioria dos casos, as galáxias peculiares não se enquadram em nenhum dos três tipos apresentados na Figura 1.1, contudo, em algum momento já foram, provavelmente, algum destes três tipos, sofrendo mudança de seu aspecto com a interação com outra galáxia ou por possuir aspecto peculiar, como as galáxias com extinção anômala. A Figura 1.2 apresenta alguns exemplos de galáxias peculiares. A imagem mais à esquerda mostra duas galáxias espirais, onde a principal é denominada como UGC 1810, possuindo um disco que é distorcido pela força de maré gravitacional da galáxia companheira abaixo dela, conhecida como UGC 1813. A faixa de pontos azuis na parte superior é a luz combinada de aglomerados de estrelas

jovens azuis intensamente brilhantes e quentes. A imagem do meio apresenta a interação entre duas galáxias espirais. A imagem mais à direita mostra duas galáxias aneladas interagindo, uma galáxia à esquerda, quase sem nenhuma alteração em seu anel; já a galáxia mais à direita exibe um anel azulado onde há intensa formação de estrelas; esse formato foi, provavelmente gerado pela passagem da galáxia da esquerda através da galáxia da direita.[3]

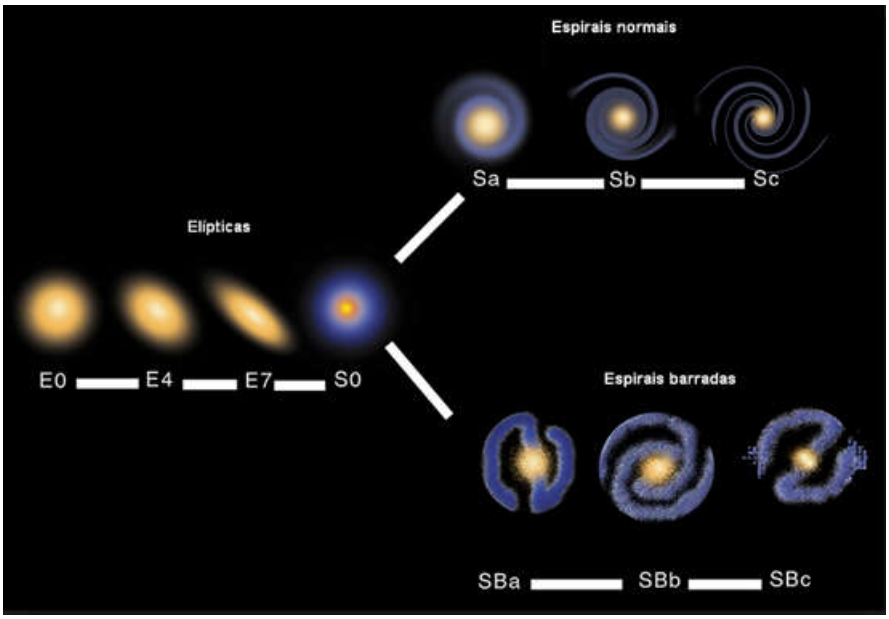

Figura 1.1: Classificação morfológica das galáxias – Sequência de Hubble. (Adaptado de Prof 2000 - http://www.prof2000.pt/).

Figura 1.2: Galáxias peculiares: ARP 273, NGC 6050 interagindo com IC 1179 e ARP 147. Fonte: NASA, ESA, e a Equipe Hubble Heritage (STScI / AURA).

Um método de detecção de fusões de galáxias por meio de fotometria de quatro cores foi estudado por Schombert et al. (1990). Nesse método, foi realizada a seleção de objetos com base em sua morfologia visual, por meio da análise de estruturas, tais como a secção transversal da cauda, pontes e envelopes, dentre outras. Nesse estudo, Schombert et al. notaram a existência de provas claras na amostra de uma correlação entre a grandeza de variação da cor e o tempo. As variações de cor são maiores em um curto período de tempo após o início da interação e diminuem a um nível muito baixo em sistemas que representam fusões. Essa correlação fornece uma alternativa para estimar a idade de interação; idade em sistemas nos quais a morfologia não

[3] http://hubblesite.org/newscenter/archive/releases/index/235/

permite conclusões precisas.

Um estudo sobre a formação de galáxias e a sua interação com as colisões entre elas foi feito por Struck (1999), no qual ele demonstra o fato de que as teorias da época vinham mudando de paradigmas por perceberem que a formação e evolução de galáxias são decorrentes de acúmulos de colisões e fusões seguidos de um longo período de relaxamento e reestruturação [Struck 1999]. Por meio de exemplos, ele mostra como modelos numéricos detalhados e observações em multi-banda têm permitido observar a sequência cronológica geral de morfologias colisionais, e como essas formas são produzidas pelos processos de cinemática de interações de marés, pela dinâmica dos gases hipersônicos, por fricção dinâmica coletiva e por um violento relaxamento.

As interações e fusões entre galáxias desempenham um papel importante nos processos de evolução das galáxias, incluindo a estruturação de massa, formação de estrelas, transformação morfológica e atividade no núcleo galáctico ativo (AGN). A fusão de galáxias pode levar à formação de galáxias maiores bem como outras estruturas galácticas [Kauffmann et al. 1993; Cole et al. 2001]. Por esse fato, interações/fusões entre pares ou grupos de galáxias são objetivos de pesquisas de muitos cientistas.

Geller et al. (2006), usaram espectroscopia e fotometria no infravermelho para analisar uma amostra completa de, aproximadamente 800 pares de galáxias próximas destacando suas propriedades. Woods et al. (2006), com base no Levantamento de redsfhit do CfA2 estudaram uma amostra de 345 galáxias, identificando 167 pares. Focardi et al. (2006), compilaram um catálogo com uma amostra com 178 campos com galáxias brilhantes usando o Catálogo UZC (Updated Zwicky catalog (UZC). Argudo-Fernandez et al. (2015), tendo como base os dados dos SDSS de fotometria e espectroscopia elaboraram um catálogo com 3.702 galáxias isoladas, 1.240 pares e 315 tripletos de galáxias isoladas. Cao et al. (2016), com base nos dados do Herschel, elaboraram um catálogo com 176 pares de galáxias. Os trabalhos mencionados acima foram todos para o Hemisfério Norte. Soares et al. (1995), obtiveram uma amostra 621 pares de galáxias para o Hemisfério Sul, tendo como base os dados do Catálogo ESO-Uppsala de galáxias.

Shalyapina et al. (2007), estudaram, por meio de espectrografia em duas dimensões, candidatas a galáxias de anel polar, em especial, o par de galáxias UGC 5600/09. Um estudo feito por Domingue et al. (2009), selecionaram uma amostra de pares de galáxias próximas "candidatas de maior-fusão", a fim de calcular a função de luminosidade (LF) e a dependência de massa de fusão. Uma outra pesquisa feita, mapeando duas regiões de 5 graus quadrados, uma em torno do par de galáxias NGC 7332/9 e outra da galáxia isolada NGC 1156, com o objetivo de investigar o ambiente dessas galáxias, identificando possíveis galáxias companheiras anãs e sinais remanescentes de outras interações. No total foram encontradas 87 galáxias nessa região e, dentre elas, três galáxias anãs [Minchin et al. 2010]. Num estudo sobre estrutura de grupos de galáxias, Kopylov e Kopylova (2015), usando dados de arquivo SDSS-DR7, relataram os resultados da medição e comparação de massas para uma amostra de 29 grupos e aglomerados de galáxias ($z < 0{,}1$).

1.1 O uso da computação no reconhecimento de padrões

Existe um grande volume de dados armazenados em formato de imagens e em catálogos astronômicos disponíveis em alguns centros de dados com informações detalhadas sobre vários tipos de objetos. Esses dados são provenientes da fotometria realizada sobre uma determinada imagem, ou seja, são utilizados softwares específicos, que levam em conta a característica das observações, determinando a magnitude, excentricidade, tamanho dos objetos, entre outros aspectos. Dentre essas características, algumas nos permitem classificar o objeto como estrela ou galáxia, entre outras categorias. Entretanto, como esses softwares realizam processamento em grandes áreas do céu, muitas vezes não é possível estimar com precisão as propriedades de determinado objeto. Isso se deve ao fato de que um dado objeto pode ter um baixo brilho devido à resolução de algumas imagens, ou, até mesmo, a algumas características específicas que são apenas encontradas em uma pequena fração de objetos.

Além desses dados armazenados em catálogos, também dispomos das imagens que foram utilizadas para realizar esse procedimento, e oriundas das observações. Isso é muito útil e importante, visto que os softwares de redução de dados, muitas vezes, não possuem rotinas específicas para determinar a forma e propriedades específicas dos objetos que se pretende estudar. Por isso, muitas vezes, tais objetos são classificados, simplesmente, como galáxias.

Por conta do elevado número de imagens disponíveis, o poder computacional de reconhecimento de padrões vem sendo uma ferramenta útil na detecção de estruturas específicas em imagens dos grandes levantamentos [Shamir e Wallin 2014]. O processo consiste em fazer com que um software aprenda sobre um determinado padrão e identifique ou classifique esse padrão dentro de um conjunto de imagens [Gonzalez e Woods 2010].

Um exemplo de trabalho que usa as técnicas de reconhecimento de padrões na Astronomia é o trabalho proposto por Dodd e MacGillivray (1986), que descreve um método para a detecção automática e parametrização de ricos aglomerados de galáxias em levantamentos de grande profundidade angular. A técnica foi aplicada a um campo do grande levantamento do céu do Hemisfério Sul (ESO/SERC Southern Sky Survey) e a campos com galáxias sintéticas obtidas via simulações de Monte Carlo [Dodd e MacGillivray 1986].

Alguns algoritmos foram concebidos para detecção de objetos em movimento, tais como cometas e asteroides. Toshifumi (2005), propõe um algoritmo que usa imagens, a fim de detectar objetos em movimento, muito escuros que são invisíveis em uma única imagem CCD; identificando, com sucesso, asteroides com magnitudes abaixo de 21 mag, em imagens obtidas por telescópios de pequeno porte.

Cotini et al. (2013), em seu trabalho sobre interação entre galáxias, investigaram uma possível ligação entre fusões e o aumento da atividade de buracos negros supermassivos no centro das galáxias. Eles compararam a fração de fusões de uma

amostra local de galáxias ativas com uma amostra de controle adequada extraída do catálogo HyperLeda com a mesma distribuição de redshift da amostra local. Foram detectados sistemas de interações nos dois exemplos baseados em algumas informações não parametrizadas, tais como: índices de concentrações estruturais e assimetria, sendo proposto um critério morfológico baseado na combinação desses valores. Nesse trabalho é proposto também um software para computação automática desses índices estruturais chamado PyCASSo (Python CAS Software). Os autores concluíram que a fração de galáxias interagentes dentre a população ativa excede a fração de fusão da amostra controlada.

Em 2008, foi proposto o projeto Zoológico de Galáxias (Galaxy Zoo)[4], que disponibiliza a classificação morfológica visual para aproximadamente um milhão de galáxias extraídas do Sloan Digital Sky Survey[5] (SDSS), convidando o público para classificá-las via internet. Esse público voluntário é chamado de cidadão cientista. O resultado deste processo foi uma classificação tão consistente quanto uma feita por astrônomos [Lintott et al. 2008, 2011]. O trabalho de classificação do Galaxy Zoo foi usado por Kuminski et al. (2014), que propõem um método que considera a classificação feita por cidadãos cientistas como base para o treinamento de um sistema de máquina de aprendizado. Nesse trabalho o software usado para análise de imagens foi o Wndchrm [Shamir et al. 2008, 2014].

Hocking et al. (2015), apresentam uma nova abordagem de aprendizagem não supervisionada para segmentar automaticamente e rotular imagens em pesquisas astronômicas. Foi usado o algoritmo de Gás Neural Evolutivo ou GNG (Growing Neural Gas) para codificar o espaço de características dos dados de imagens. Quando combinado com uma técnica chamada de agrupamento hierárquico, os dados de imagem podem ser, automaticamente, segmentados e rotulados por meio da organização de nós no GNG.

Como o método de aprendizado é "não supervisionado", os rótulos são indicados pelo próprio algoritmo, para provar a validade do trabalho, a técnica foi aplicada a dados de imagens de aglomerados de galáxias do Hubble Space Telescope Frontier Fields, contendo uma variação de tipos de galáxias que seriam facilmente reconhecidas e classificadas por uma inspeção humana. Ao treinar o algoritmo, usando o aglomerado de galáxia Abell 2744 e aplicando o resultado ao aglomerado MACS0416.1-2403 [Hocking et al. 2015], mostraram como o algoritmo pode separar corretamente as características de imagem que um ser humano iria associar com galáxias tipo precoces e tardias

du Buisson et al. (2015), apresentam um trabalho no qual máquinas de aprendizado baseado no PCA (Análise de Componentes Principais) são usadas para extrair características e classificar imagens transientes de supernovas do SDSS em artefatos ou objetos reais. O procedimento usado por du Buissson et al. (2015), mostrou ser equivalente a classificações feitas por seres humanos, com uma taxa de precisão de 84%. Nesse trabalho eles atentam para o fato de que futuros levantamentos, como o LSST (Large Synoptic Survey Telescope) exigem procedimento de reconhecimento e classificação totalmente baseado em máquinas de aprendizado, sendo o PCA importante por se igualar ao desempenho humano nesses processos [du Buisson et al.

[4] https://www.galaxyzoo.org/
[5] http://www.sdss.org/

2015].

Kuminski e Shamir (2015), chamam a atenção do crescente número de levantamentos atuais com grande cobertura do céu, com milhões de imagens de galáxias disponíveis para análise. Apesar do esforço que tem sido feito para envolver o público na tarefa de análise e classificação dessas galáxias como é o caso do Zooniverse, ao invés de concentrar esse trabalho nas mãos de poucos cientistas, ainda assim se faz necesária a automação desse processo de análise. Projetos futuros como o LSST irão precisar de reforço na tarefa de analisar as imagens de galáxias, dado o grande volume de imagens observadas. Kuminski e Shamir (2015), apresentam um método de visão computacional a fim de analisar automaticamente as imagens de galáxias e deduzir a sua morfologia. O método foi aplicado ao Galaxy Zoo II, mostrando um alto grau de concordância com a análise do público. Em alguns aspectos, por exemplo, a análise do tipo de galáxia espiral, o grau em questão chegou a 95%; já, em outros, como o número de braços espirais, o método não se mostrou tão eficiente assim, apresentando uma precisão de apenas 36%.

No entanto, mesmo contando com todo o trabalho dos cidadãos cientistas, uma vez que o LSST vai adquirir milhares de milhões de imagens de galáxias [Abell et al. 2009], esse esforço não será capaz de fornecer uma solução e variável para a análise de bases de dados adquiridos por esses levantamentos futuros. As classificações feitas por cidadãos cientistas são susceptíveis a erro por serem feitas por indivíduos que não possuem formação em Astronomia e ficam responsáveis por investigar dados a partir da análise de objetos nas imagens [Lintott et al. 2011]. Para amenizar fontes oriundas da análise humana, é necessário se aplicarem métodos estatísticos que ignoram objetos que não obtiveram consistência entre todos os dados analisados por diferentes pessoas.

Alguns geradores automáticos de catálogos de morfologia de galáxias foram propostos com a finalidade de superar os erros devido à inabilidade ou incapacidade de analisar de forma exaustiva uma enorme quantidade de dados das bases atuais. Huertas-Company et al. (2010), publicaram um catálogo de galáxias do SDSS gerado automaticamente com espectros, e os classificaram em quatro tipos básicos da sequência de Hubble. Kuminski e Shamir (2016), aplicaram uma metodologia de visão computacional e reconhecimento de padrões para analisar a ampla morfologia de 3 milhões de galáxias obtidas de SDSS-DR8[6], gerando um catálogo com informações, tais como o ID do objeto, seu par de coordenadas (RA e DEC) e a classificação automática em elíptica ou espiral. A acurácia do catálogo foi testada, usando galáxias que foram classificadas no Projeto GalaxyZoo.

[6] https://www.sdss3.org/dr8/

1.2 Objetivos

Este trabalho tem como objetivo principal:

- a identificação/classificação de novas galáxias peculiares interagentes nas categorias 1 e 2 do Catálogo de AM87 no infravermelho a baixos redshift, tendo como base os dados do grande levantamento no infravermelho de todo o céu, 2MASS.

E objetivos específicos:

- determinar características relativas à extinção interestelar e povoamento estelar nos campos das imagens com candidatas a galáxias peculiares, de forma que ainda seja possível identificarem-se galáxias peculiares na imagem;
- elaborar uma amostra de galáxias peculiares sem prévia identificação das categorias 1 e 2.

1.3 Aspectos de originalidade do trabalho

Conforme mencionado acima, significativos progressos têm sido realizados na identificação automática de galáxias tendo como base os dados de grandes levantamentos astronômicos. No entanto, esses trabalhos ainda são carentes em fazer análises de grandes regiões do Hemisfério Sul (apesar de grandes levantamentos com cobertura espacial reduzida, como o VISTA). Grande parte das análises são feitas, por exemplo, com o SDSS, que realizou observações de espectroscopia e fotometria no óptico em cinco bandas de uma área de aproximadamente 15.000 graus quadrados.

Por usar fotometria, espectroscopia e cobrir uma grande área do céu, os dados do SDSS têm sido explorados de forma bastante intensa nos últimos 15 anos, com várias atualizações e projetos de continuidade, estando já em sua 12ª atualização de dados e no projeto SDSS-3. Por outro lado, a identificação de galáxias peculiares, para o Hemisfério Sul, é uma área que requer maior atenção, pois existem certas categorias que tiveram novos objetos acrescidos, baseados em dados atuais, a exemplo, das galáxias aneladas. Moiseev et al. (2011), que identificaram, aproximadamente, 275 novas galáxias aneladas com os dados do SDSS.

Os dados do grande levantamento no infravermelho, 2MASS têm 15 anos que foram disponibilizados e nenhum trabalho similar foi realizado ainda. Para efetuar tal tarefa, utilizamos um software (Wndchrm) desenvolvido para reconhecimento de padrões, e utilizado em galáxias peculiares tendo como base os dados do SDSS, com bons resultados.

A nossa abordagem consistiu em utilizar esse software para todo o Hemisfério Sul, com as imagens do 2MASS, ao contrário do que foi utilizado por Shamir & Wallin (2014), com dados no óptico. Um aspecto importante é que desenvolvemos toda uma metodologia para selecionarmos uma dada imagem para análise, levando-se em conta o catálogo com dados de fotometria do 2MASS de fontes extensas, ou seja, em uma primeira etapa, selecionamos com base em um catálogo, as imagens com as quais, acreditamos, tenham candidatas a galáxias e inspecionamos se essa imagem

corresponde a de uma galáxia peculiar das categorias 1 e 2 de AM87. Levamos em conta, também, aspectos relativos à quantidade de estrelas e extinção interestelar, presentes nas imagens que serão utilizadas pelo software.

Nossa metodologia faz com que tenhamos um ganho de tempo, não somente de processamento das imagens, mas também na análise de dados, evitando inspecionar imagens as quais sabemos, com base em dados de fotometria, não conter sequer galáxias, muito menos, galáxias peculiares. Utilizando nosso método ao invés de fazermos a varredura em aproximadamente 22 milhões de imagens, reduzimos para algo em torno de 10 mil. Acreditamos que trabalhos similares possam ser feitos em outras categorias, utilizando-se do método desenvolvido no presente estudo.

1.4 Organização do Trabalho

No Capítulo 2, teremos uma breve revisão bibliográfica sobre Reconhecimento de Padrões. No Capítulo 3, temos uma descrição dos dados usados no trabalho, obtenção das imagens para compor o conjunto de dados, ou dataset, análise das imagens selecionadas como amostra de treinamento e as descartadas, assim como as propriedades dos campos onde esses objetos estão localizados. O Capítulo 4 descreve o método utilizado para identificar galáxias peculiares e o software (Wndcrhm) usado para esse fim, assim como resultados da classificação de imagens de uma região de 573 graus quadrados, usada como ferramenta de teste antes de se aplicar nosso método para todo o Hemisfério Sul. O Capítulo 5 apresenta a aplicação do método para todo o Hemisfério Sul, assim como as candidatas a galáxias encontradas para as categorias 1 e 2 e uma comparação com as encontradas no Catálogo de AM87 e em outros. Por fim, o Capítulo 6 aborda as conclusões e perspectivas do presente trabalho.

Capítulo 2

Reconhecimento de padrões em imagens

2.1 Conceitos

Para Gonzalez (2010), um padrão é um arranjo de descritores que, também, pode ser chamado de característica. Um conjunto desses arranjos que compartilham propriedades em comum é chamado de classe de padrões. Considerando-se esses conceitos, a tarefa de atribuição de padrões às suas respectivas classes é chamada de reconhecimento de padrão. O processo de reconhecimento de padrões pode ser automatizado quando se aplicam técnicas automáticas para atribuir padrões às suas classes e, nesse processo, quanto menor a intervenção humana, melhor.

Os arranjos de padrões podem ser representados por vetores de características, strings ou árvores. Os vetores de características têm seus descritores compostos por padrões, cuja natureza depende da metodologia que está sendo aplicada na análise do padrão físico. A escolha de descritores é importante, pois, pode ser que um tipo descritor não seja suficiente para separar duas classes analisadas. As classes de padrões podem ter seus vetores de características gerados com base em informações quantitativas, tais como comprimento, cor, área; porém, em algumas aplicações, as características são melhor descritas com base nas relações estruturais, tais como as relações entre as medidas quantitativas de cada uma. Outra forma, mais poderosa que as anteriores, é por meio de descrições por árvores, na qual os descritores de uma classe se relacionam por meio de uma hierarquia, onde a raiz da árvore é a imagem inteira; outras características marcantes dessa imagem compõem o próximo nível, e os níveis seguintes são compostos por características do nível anterior, e, assim, sucessivamente.

Um trabalho de reconhecimento automático de padrões em imagens digitais é composto pelas seguintes etapas: aquisição, pré-processamento, segmentação, análise e reconhecimento, conforme apresenta a Figura 2.1. A etapa de aquisição consiste no processo de obtenção das imagens que serão trabalhadas; no pré-processamento, a qualidade das imagens é melhorada, e essas são preparadas para as fases posteriores; a fase de segmentação implica encontrar semelhança entre regiões das imagens; análise e reconhecimento são a fase na qual as informações presentes nas imagens são interpretadas. Não, necessariamente, todas as quatro etapas estão presentes em uma

aplicação de reconhecimento de padrões [Siqueira et al. 2000].

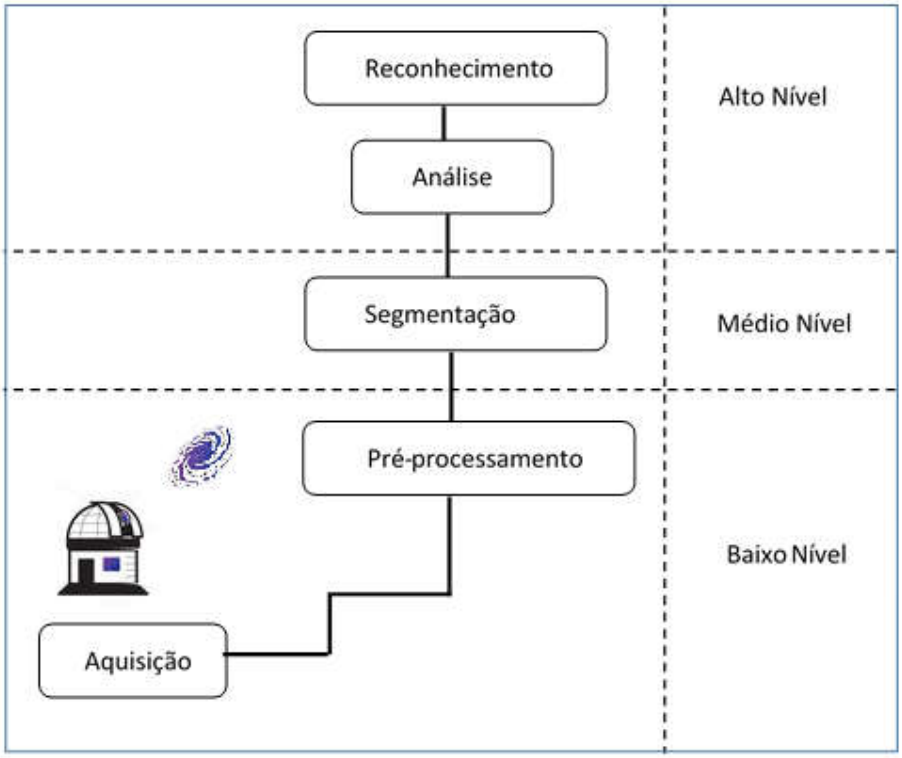

Figura 2.1: Fases típicas de um trabalho de reconhecimento de padrões.

Historicamente, são as seguintes técnicas utilizadas para se resolverem problemas de reconhecimento de padrões nas fases de alto nível: análise estatística, análise sintática ou estrutural, redes neurais e lógica fuzzy [Gonzalez e Woods 2010].

2.2 Reconhecimento de padrões por análise estatística

Nessa técnica, os vetores de características formam n-tuplas, sendo utilizadas regras de decisão, teoria de probabilidade, funções discriminantes e outros procedimentos estatísticos. É o tipo de reconhecimento mais tradicional.

Um dos métodos usados para reconhecimento de objetos na técnica da análise por estatística é o de decisão teórica. Essas abordagens são baseadas em uma função de decisão [Gonzalez e Woods 2010]. O ponto chave do método é encontrar uma função de decisão d(x) onde x é um determinado padrão, descrito a seguir:

$$d_i(x) > d_j(x)\ j = 1,2,\ldots,W; j \neq i \qquad (2.1)$$

W é a quantidade de funções de decisão. Isto quer dizer que um padrão desconhecido x pertence a uma determinada classe i em um conjunto de W classes de padrões, caso a substituição de x em todas as funções de decisão d(x) fizer com que $d_i(x)$ tenha o maior valor numérico. A chamada fronteira de decisão que separa as classes é dada pelos valores de x, para os quais $d_i(x) = d_j(x)$. Encontrar funções de decisão que satisfaçam a Equação 2.1 é a principal tarefa nesta abordagem para o reconhecimento de objetos.

Uma das técnicas utilizadas para encontrar funções satisfatórias é o casamento (matching). Nessa técnica, uma classe é representada por um vetor de características protótipo. Um padrão desconhecido é atribuído a uma classe, usando, por exemplo, a abordagem do classificador de distâncias mínimas (euclidiana) entre o padrão desconhecido e cada um dos vetores protótipos. A decisão é tomada considerando-se a menor distância entre todas as estabelecidas. Na prática, o classificador de distâncias mínimas é indicado quando a distância entre as médias é grande, em comparação com a dispersão, ou a aleatoriedade de cada classe em relação à sua média.

Considerações probabilísticas são especialmente importantes na tarefa de reconhecimento, tendo em vista que essa atividade envolve medida e interpretação de eventos físicos levando as classes de padrões a uma aleatoriedade. Com o objetivo de minimizar a probabilidade de erros de classificação, deve-se buscar uma abordagem que seja ótima [Gonzalez e Woods 2010].

Em uma classificação, quando um classificador decide que determinado padrão pertence a uma classe, quando, na verdade, pertence a outra, diz-se que aconteceu uma perda, que pode ser calculada pela equação de risco médio condicional ou perda.

$$r_j(x) = \sum_{k=1}^{W} L_{kj}\rho(x/\omega_k)P(\omega_k) \qquad (2.2)$$

L_{kj} é a perda cometida por um classificador, $\rho(x/\omega_k)$, a função de densidade de probabilidade dos padrões da classe ω_k e $P(\omega_k)$, a probabilidade de ocorrência da classe ω_k. Se o classificador calcular os $r_w(x)$ para as W classes possíveis e, para cada

padrão x, atribuir o padrão à classe com a menor perda, a perda média total com respeito a todas as decisões será mínima. Esse classificador é chamado de classificador Bayesiano [Gonzalez e Woods 2010]. A otimização de um classificador bayesiano está diretamente ligada à escolha de uma estimativa de função de densidade probabilística de forma conhecida. Na maioria dos casos, esses são métodos difíceis de ser aplicados. A função de densidade probabilística gaussiana é, geralmente, assumida para o classificador bayesiano. Essa premissa, quanto mais próxima for da realidade, mais o classificador bayesiano se aproxima da perda média mínima de classificação.

2.3 Reconhecimento de padrões por descrição estrutural e análise sintática

Nessa técnica de reconhecimento, os elementos têm suas características estruturais representadas sintaticamente por suas partes constituintes, propriedades e relacionamentos entre si. A ideia é tratar a complexidade de determinados padrões por meio de uma hierarquia de estruturas, possibilitando-se que um padrão seja descrito em funções de padrões mais simples, e estes sejam descritos em função de padrões mais simples ainda [Tanaka 1995]. Essa técnica é, frequentemente, usada no reconhecimento de figuras, formas em geral e análise de cenário. Em problemas desse tipo, a análise da forma ou da figura pode ser extremamente complexa resultando numa grande quantidade de características difíceis de ser analisadas. Uma decomposição do objeto principal em subcategorias hierárquicas facilitará essa análise revelando subpadrões mais simples. Um exemplo é a representação de uma forma humana na Figura 2.2.

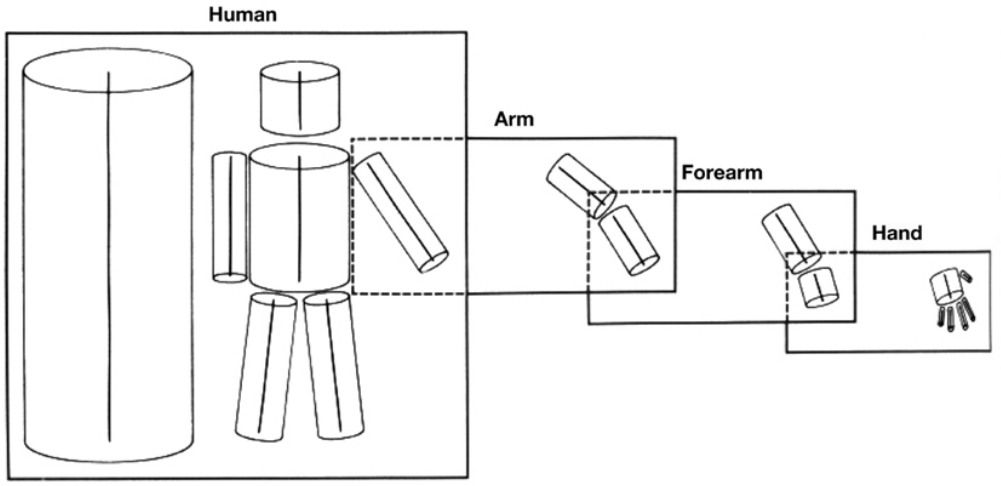

Figura 2.2: Exemplo de analise estrutural hierárquica de uma forma humana. Fonte: Adaptado de Marr e Nishihara (1978).

A mesma forma pode ser analisada através de uma árvore hierárquica, na qual a forma é decomposta em formas mais simples como mostra a Figura 2.3.

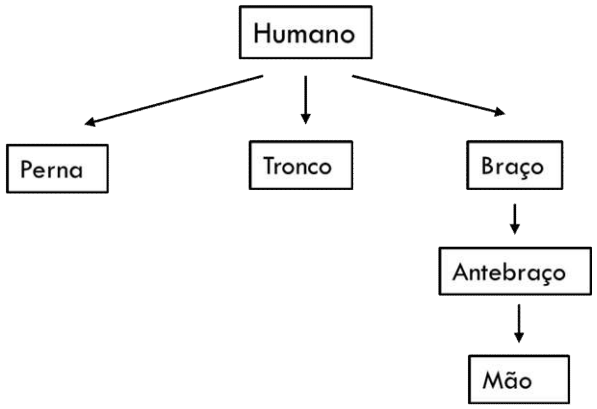

Figura 2.3: Árvore hierárquica da forma humana. Fonte: Adaptado de Marr e Nishihara (1978).

Nesse exemplo, uma forma humana representa um padrão, que é construído a partir de padrões mais simples, como as partes que compõem o corpo, exatamente como as frases são compostas através da concatenação de padrões mais simples, as palavras, e estas, por sua vez, compostas novamente pela concatenação de padrões ainda mais simples, os caracteres. Esse tipo de reconhecimento só funciona se os subpadrões forem mais simples que o padrão geral. O subpadrão mais simples é chamado de padrão primitivo e deve ser de fácil identificação, além de reunir características que permitam identificar um único padrão. Portanto, a seleção do padrão primitivo se torna uma tarefa muito importante na técnica da análise sintática, a qual varia de acordo com o caso em questão e com o padrão que se quer reconhecer [Tanaka 1995].

Uma maneira simplificada de se fazer a descrição estrutural do padrão, em termos de um conjunto de padrões primitivos e de suas operações de composição, é usar a Linguagem de Descrição de Padrão ou PDL, Pattern Description Language. A PDL possui uma gramática com regras para criação de composições de padrões. Essas operações podem ser expressas como operações lógicas e matemáticas. Um aspecto atraente nos métodos sintáticos é a sua fácil adequação à recursividade após se finalizar a escolha de uma série de regras para se descrever a relação entre as partes do objeto.

2.4 Redes Neurais

Em um problema de decisão teórica, em que as classes de padrões são, frequentemente, desconhecidas, ou de difícil estimativa, traz maiores benefícios usar métodos que utilizem treinamento, para produzir as funções de decisão necessárias. Os modelos resultantes da utilização de elementos básicos da computação não linear para o aprendizado, via treinamento, também conhecidos como neurônios, organizados em rede, são chamados redes neurais [Gonzalez e Woods 2010]. Esses são usadas no contexto de um desenvolvimento adaptativo aos coeficientes das funções de decisão por meio de um conjunto de padrões de treinamento apresentados sucessivamente.

MaCulloch e Pitt (1943) apresentaram, em um trabalho, modelos de neurônios na forma de dispositivos de limiarização binária e algoritmos estocásticos que alternavam subitamente de valores entre 0 e 1, representando os estados dos neurônios como uma base para os modelos dos sistemas neurais. Posteriormente, Hebb (1949), usou modelos matemáticos para representar o conceito de aprendizagem por reforço ou associação.

Em meados de 1950, a teoria do reconhecimento de padrões foi abalada pelo surgimento de uma classe de máquina especial chamada máquina de aprendizagem de Rosenblatt (1959, 1962). Ele provou matematicamente, que suas máquinas chamadas de perceptrons, quando preparadas por um conjunto de treinamento, separado por um hiperplano, convergem para uma solução em um número finito de passos iterativos, na forma de coeficientes de hiperplanos capazes de separar de forma correta as classes que são representadas pelos padrões no conjunto de treinamento. Com o tempo, a proposta de Rosenblatt se mostrou inadequada para a maioria das tarefas de reconhecimento de padrões de importância prática.

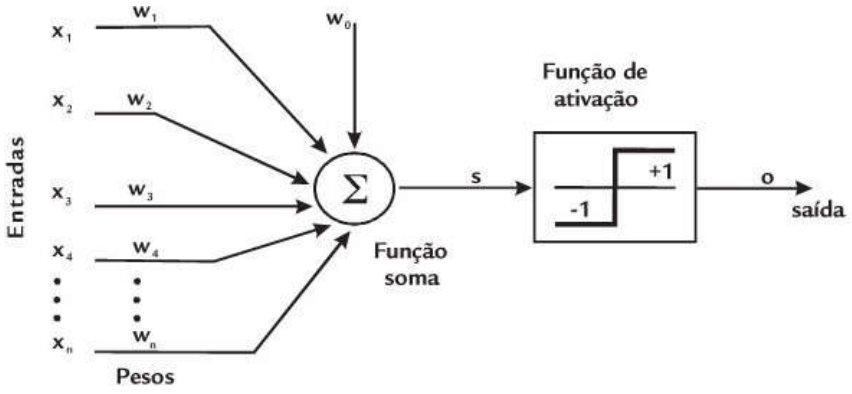

Figura 2.4: Exemplo de um perceptron para duas classes. Fonte: Adaptado de Medeiros (2006).

Tentativas posteriores para emular os perceptrons, considerando as suas múltiplas camadas, careciam de algoritmos de treinamento eficazes. Em 1969, Minsky e Papert apresentaram uma análise desanimadora das limitações dessas máquinas. Esse ponto de vista durou até meados da década de 80 quando, em 1986, Hinton e Williams

apresentaram novos algoritmos de treinamento para perceptrons de múltiplas camadas que mudaram as coisas, usando um método chamado de regra generalizada delta para o aprendizado por retroprogramação, que oferece um meio de treinamento eficaz para as máquinas de múltiplas camadas. Essa regra, apesar de não ter comprovação de solução no sentido de prova análoga para perceptron de uma camada, tem sido usada em diversos problemas com interesse prático, fazendo com que máquinas de múltiplas camadas com perceptron se tornassem um dos principais modelos de redes neurais utilizados atualmente.

Basicamente, um perceptron aprende uma função de decisão linear que dicotomiza dois conjuntos de treinamento, linearmente separáveis, baseando-se em uma soma ponderada de suas entradas, conforme pode ser visualizado na Figura 2.4. As entradas x_i são modificadas por pesos ou coeficientes w_i, antes de ser introduzidos no elemento de limiarização. Os pesos, nesse caso, são análogos às sinapses do sistema neural humano. A função de ativação mapeia a saída da soma na saída final do dispositivo.

2.5 Lógica Difusa

Reconhecimento de padrão é um processo inerentemente humano que exige respostas a estímulos de natureza imprecisa [Zadeh 1996]. Automatizar esse processo através de uma técnica que não leve em conta a característica de imprecisão do mundo real pode não retornar resultados úteis. A abordagem que leva em consideração o grau de incerteza, por vezes inerente às características e às classificações, é chamada de teoria dos conjuntos difusos ou, simplesmente, lógica difusa.

Na lógica booleana, temos apenas dois valores possíveis: verdadeiro ou falso, ou seja, 0 ou 1, que são os valores booleanos. A lógica difusa, ou fuzzi, trata de valores que variam entre 0 e 1. Assim, uma pertinência de 0,5 pode representar meia verdade, logo 0,9 e 0,1, representam quase verdade e quase falso, respectivamente. A lógica difusa é uma ferramenta capaz de capturar informações vagas, em geral, descritas em linguagem natural e convertê-las para um formato numérico, de fácil manipulação [Ortega 2001].

Bezdek (1993) diz que, ao contrário da lógica convencional, na lógica fuzzi, utiliza-se a ideia de que todas os processos admitem graus de pertinências. Por exemplo, em conjuntos fuzzi, um elemento pode estar, parcialmente, em mais de um conjunto. Cada elemento terá, então, um grau de pertinência a cada um dos conjuntos. Essa associação entre o valor de um dado elemento e o grau de possibilidade de pertencer a um conjunto é dado pela função de possibilidade, como mostrado na Figura 2.5.

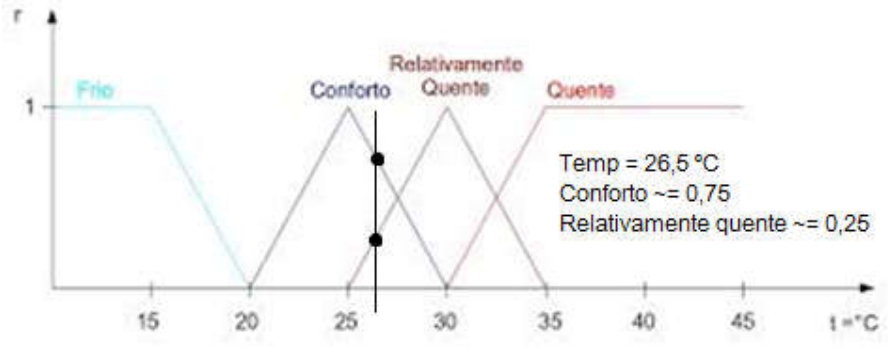

Figura 2.5 Função de possibilidade para um conjunto de temperaturas com destaque para temperatura de 26,5 °C. Fonte: UNICAMP – Faculdade de Tecnologia.

Os conjuntos fuzzi são apropriados para classificação de padrões, pois um determinado gesto ou padrão pode, de fato, ser associado parcialmente a muitas classes diferentes. O processo que consiste em converter informações vagas, na linguagem natural, em valores de pertinência facilmente compreensíveis numa escala de 0 a 1 é chamado de fuzzificação [Zadeh 1996].

O processo de criação de um conjunto fuzzi envolve a identificação das variáveis linguísticas fuzzi, que expressam de forma qualitativa um termo linguístico que atribui um conceito, sendo avaliadas quantitativamente por uma função de pertinência. Na Figura 2.5, temos as seguintes variáveis: frio, conforto, relativamente quente e quente. As regras fuzzi são um modo de relacionar um valor de entrada com uma saída, ou causa e resposta. Sendo, por exemplo, x a causa, e f(x) a resposta poderíamos ter a seguinte regra:

$$\text{Regra i : Se x é } A_i \text{ Então f(x) é } B_i \text{ , i = 1,..., N} \quad (2.3)$$

x representa a variável independente e f(x) a variável dependente, sendo A_i e B_i constantes linguísticas e N o número de dados experimentais que descreve a função.

Após a elaboração do conjunto de regras fuzzi, faz-se necessária a utilização de uma máquina de inferência para ser extraída a resposta final, que descreva situações específicas cuja inferência conduz a algum resultado desejado. Trata-se de capturar um conhecimento específico. Um conjunto de regras é capaz de descrever um sistema em suas várias possibilidades [Ortega 2001].

Além da aplicação em reconhecimento de padrões a lógica fuzzi é utilizada para outros fins, tais como: controle de tráfego [Pappis e Mamdani 1977], controlador de braços mecânicos [Tanscheit e Scharf 1988], controle de temperatura [Singhala et al. 2014].

Capítulo 3

O conjunto de dados

Para obter as imagens de candidatas a galáxias interagentes recorremos ao grande levantamento 2MASS. Nesse levantamento, existem imagens de todo o céu, ou seja, tanto do Hemisfério Norte (HN), quanto do Hemisfério Sul (HS). As informações do conteúdo das imagens, tais como: identificação dos objetos, brilho, redshift, entre outras, podem ser encontradas em catálogos de informações desses mesmos levantamentos [Cutri et al. 2003, Skrutskie et al. 2006]. Este capítulo descreve o processo de análise e obtenção das informações e imagens usadas a partir dos dados do 2MASS, assim como a análise das propriedades dos campos das imagens usados nas amostras de treinamento.

3.1 O Catálogo de Arp e Madore

Conforme mencionado anteriormente, não existem estudos sistemáticos para analisar as propriedades das galáxias peculiares de suas 25 categorias do Catálogo de Arp e Madore (1987, AM87), tendo em vista os dados de grandes levantamentos astronômicos. Amôres et al. (2016, em preparação) estão revisitando as propriedades de galáxias peculiares no HS, considerando os dados dos grandes levantamentos, tais como 2MASS (Skrutskie et al. 2006) e GSC General Star Catalogue, entre outros, assim como, os valores de redshifts, extinção interestelar Galáctica na direção desses objetos, entre outras propriedades.

Como esse artigo está em preparação, será resumido, o procedimento adotado por Amôres et al. (2016) para obter as coordenadas das galáxias peculiares, que foram obtidas por meio das coordenadas dos objetos fornecidos por AM87; como estas estavam com precessão B1950, foram transformadas em J2000, também verificou-se os objetos que se repetiam em várias categorias, dessa a forma é possível termos uma distribuição das galáxias peculiares do Catálogo de AM87, a qual é apresentada na Figura 3.1, nota-se regiões em branco, as quais estão na direção do plano de nossa Galáxia, que é uma região com grande quantidade de poeira e estrelas e por essa razão evitada na maioria dos estudos em Astronomia Extragaláctica.

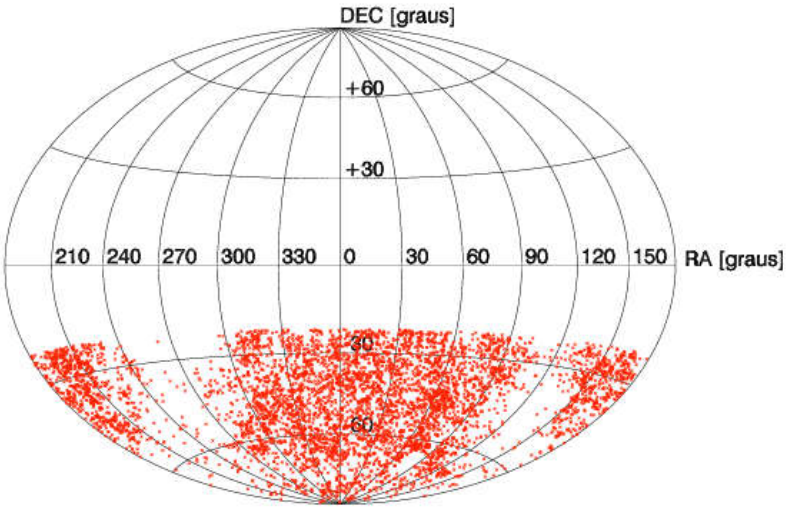

Figura 3.1: Distribuição das galáxias peculiares do Catálogo de AM87 em coordenadas equatoriais (J2000) em projeção Aitoff.

A Tabela 3.1, apresenta as categorias originais do Catálogo de AM87. Conforme pode ser visto nessa tabela, existem algumas categorias que possuem objetos interagindo mutuamente, notadamente das categorias 1 a 5, 8 e 9, 17 a 19, que, em uma etapa preliminar, pretendíamos usar em nosso estudo. Entretanto, devido ao elevado número de classes de treinamento, e, como estávamos nos familiarizando com o software Wndchrm (Capítulo 4), preferimos nos restringir apenas às categorias 1 e 2.

As categorias de 17 a 19 (grupos e/ou cadeias com mais de quatro galáxias), no Catálogo de AM87 também se mostraram inviáveis devido ao tamanho típico de nossos campos, da ordem de 2 minutos de arco, justificado pela diminuição do número de estrelas e objetos que não são os alvos principais desta análise, pois não aplicamos nenhum procedimento para retirar esses objetos de nossas imagens.

O Catálogo de AM87 não foi originado de um processo de análise com uma quantidade de categorias predeterminadas, mas sim, de uma observação empírica; listando todos os objetos que poderíamos ser classificados em uma mesma categoria. Dessa forma, uma nova categoria era elaborada e posteriormente outros objetos eram adicionados a ela.

A classificação do Catálogo de AM87, auxilia astrônomos, que estudam uma categoria ou tipo especial de objetos, a encontrarem galáxias que possam ser estudadas de forma mais detalhada em suas pesquisas; p.ex., o estudo das atividades nucleares não térmicas de galáxias com anel polar [Freitas-Lemes et al. 2013] ou investigar a natureza de galáxias que estão localizadas em um campo de uma determinada galaxia anelada [Faúndez-Abans et al. 2015]. Efeitos da interação entre cinemática e abundância química dos elementos foram estudados por Freitas-Lemes, et al. (2014).

Um motivo, em particular, do interesse em galáxias peculiares é a análise dos distúrbios em cada uma, causados pela interação entre esses objetos. As propriedades físicas das galáxias podem ser deduzidas observando-se essas interações. As

peculiaridades das galáxias podem ser atribuídas a três principais causas: vizinhas próximas, atividades internas e ambiente em geral.

Um dos aspectos do atual trabalho, é que evitaremos regiões próximas ao plano de nossa Galáxia, ou seja, latitudes (b) Galácticas, $|b| < 10°$, devido ao grande número de estrelas existentes (ver Seção 3.6), e, também, ao efeito da extinção interestelar que afeta, de maneira severa, a luz de objetos mais fracos, no caso galáxias [Amôres et al. 2012b]. Por outro lado, [Amôres et al. 2011] realizaram a estimativa do número de estrelas em comparação ao número de galáxias no plano Galáctico.

3.2 Descrição das categorias usadas

A seguir, descreveremos, com mais detalhes, as duas categorias do Catálogo de AM87 empregadas no presente estudo. A descrição é baseada no Catálogo de Galáxias Peculiares do Sul e Associações [Arp e Madore 1987].

A Categoria 1 representa as "Galáxias interagindo com companheiras". As galáxias dessa categoria parecem ter uma deformação significativa no equilíbrio de suas formas. Para cada galáxia, existe uma menor, com tamanho menor do que 50% da dominante do sistema, que aparenta estar causando alterações em sua forma. Quando a companheira é mais da metade da principal, o sistema é classificado na Categoria 2. Existirão casos em que um sistema é classificado em duas, três ou mais categorias, pelo fato de a classificação do AM87 ser feita de forma empírica.

Na Categoria 1, os principais casos tratados são as espirais com braços bem definidos. A interação das companheiras nos braços espirais da galáxia principal pode ser vista com clareza nas galáxias dessa categoria. Alguns casos mostrados aqui são de interesse especial, nos quais os braços espirais aparecem deformados pela proximidade das companheiras, como mostrados na Figura 3.2 (esquerda). No sistema AM 0320-372, mostrado pela Figura 3.2 (centro), temos uma galáxia principal elíptica interagindo com uma galáxia companheira espiral. Nesse catálogo, também existem sistemas onde os braços espirais parecem estar ligados ao redor da companheira, conforme visto no sistema AM 0244-302, mostrado na Figura 3.2 (direita).

Outro tipo de sistema é o em que as companheiras estão interagindo fortemente com toda a estrutura da principal. Posteriormente, nessa categoria encontram-se companheiras relativamente mais separadas da galáxia principal, o que faz com que essas sofram menos interações. Encontram-se também, algumas formas incomuns nas próprias companheiras, sugerindo que as companheiras têm forte atividade interna, mesmo que essas atividades não tenham sido causadas pelas interações.

Segundo AM87, no fim da categoria (para o caso de serem consideradas subcategorias), a evidência de interação é mais sutil. Esse arranjo enfatiza o importante ponto de que essa categoria é continuada na Categoria 8, "Galáxias com companheira aparente".

Tabela 3.1: Categorias originais do Catálogo de AM87. Fonte: AM87.

Número	Categoria	Percentual
1	Galaxies with interacting companion(s)	5.5
2	Interacting doubles (galaxies of comparable size)	12.6
3	Interacting triples	2.0
4	Interacting quartets	0.5
5	Interacting quintets	0.1
6	Ring galaxies (or morphologically similar objects)	3.1
7	Galaxies with (linear) jets	2.4
8	Galaxies with apparent companion(s)	11.5
9	M51-types (companion at end of spiral arm)	2.0
10	Galaxies with peculiar spiral arms	4.1
11	Three-armed spirals and multiple-armed spirals	0.5
12	Peculiar disks (major asymmetry or deformation)	2.8
13	Compact (very high-surface-brightness) galaxies	6.4
14	Galaxies with prominent or unusual dust absorption	1.6
15	Galaxies with tails, loops of material or debris	3.5
16	Irregular or disturbed,(apparently isolated) galaxies	4.2
17	Chains (four or more galaxies aligned)	4.0
18	Groups (four or more galaxies not aligned)	4.9
19	Clusters (only very conspicuous, rich systems)	1.6
20	Dwarf galaxies (low surface brightness)	6.8
21	Stellar objects with associated nebulosity	0.7
22	Miscellaneous (rare or distinctive objects)	1.4
23	Close pairs (not visibly interacting)	11.4
24	Close triples (not visibly interacting)	5.6
25	Planetary Nebulae	0.9

Figura 3.2: Galáxias peculiares Categoria 1: AM 2011-705, AM 0320-372 e AM 0244-302. Fonte: SPGA Atlas.

A Categoria 2 abrange as galáxias com "Interações Duplas". As galáxias desta categoria possuem tamanhos similares e mostram sinais de fortes interações, mais precisamente, quando a companheira é maior do que a metade da galáxia principal. São as mais numerosas, pois existe um grande número de galáxias identificadas, onde as companheiras e a principal são, a grosso modo, do mesmo tamanho, com pequena diferença entre os diâmetros, aparentando interações entre si. A Figura 3.3 contém 3 exemplos de interações dessa categoria. As galáxias dessa categoria podem ser divididas nas seguintes subcategorias: interações de elípticas com elípticas, elípticas com espirais e espirais com espirais (AM87).

Figura 3.3: Galáxias peculiares Categoria 2 AM 0053-353, AM 0058-311, ARP 148. Fonte: DSS

Aparentemente, todos os sistemas surgiram por causas similares às da Categoria 1 de AM87. É interessante notar que algumas galáxias com pequenas companheiras podem aparentar sofrer interações tão violentas como aquelas com companheiras de tamanho maior. Isso levanta a questão sobre como a interação acontece. Pode ser causada, estritamente, por interação gravitacional, principalmente, entre as estrelas, plasma, gás, poeira interestelar ou campos magnéticos, ou devido a efeitos de proximidade com outras galáxias (AM87).

3.3 O 2MASS

Os grandes levantamentos astronômicos são projetos que objetivam observar grandes regiões do céu, cobrindo significativas porções do mesmo em vários comprimentos de onda, com uma dada profundidade, que é denominada limite de completeza, ou seja, o quão completa é a amostra, em termos de sua magnitude, para uma dada região. Devido a sua grande cobertura, eles são ferramentas ideais para o estudo de estruturas em grande escala, para a obtenção de parâmetros em modelos, etc.

O 2MASS, Two Micron All-Sky Survey, ou Levantamento para Todo o Céu no Infravermelho Próximo, foi um projeto que cobriu 99,998% do céu observável, com as observações no Hemisfério Norte, no Monte Hopkins, Arizona, (EUA), quanto no Hemisfério Sul, com observações em Cerro Tololo (Chile). Cada telescópio foi equipado com uma câmera de três canais, cada canal composto por uma matriz de 256 × 256 e detectores HgCdTe, capaz de observar o céu simultaneamente nos filtros J (1,25 µm), H (1,65 µm), e K_S (2,17 µm). O volume de dados do 2MASS é composto por um atlas do céu cuja última versão apresentou 4.121.439 imagens (no formato FITS[7]) com resolução de 512 x 1024 pixels; um catálogo de fontes pontuais com informações de 470.992.970 objetos [Skrutskie et al. 2003] e um catálogo de fontes extensas com 1.647.599 objetos [Skrutskie et al. 2006].

A Figura 3.4 apresenta uma imagem em falsa cor com a observação das fontes pontuais de todo o céu, realizada nos filtros J, H e Ks. A figura está em coordenadas Galácticas, ou seja, representadas pelas coordenadas, longitude e latitude Galácticas. A região do plano Galáctico, na qual se concentra a maior parte da distribuição de gás, poeira e estrelas em nossa Galáxia [Amôres e Lépine 2005], pode ser vista no centro da imagem (considerando-se latitude $|b| < 10°$), por toda a faixa de longitude. A figura apresenta a emissão no infravermelho próximo de estrelas.

A parte central representa o bojo, assim como uma barra que se estende por, aproximadamente, 4,0 kpc [López-Corredoira 2001 e Amôres et al. 2013]. Nas partes mais externas, observa-se uma estrutura similar a um empenamento e esgarçamento do disco, que tem como origem a interação com o gás e a influência de objetos extragalácticos [Amôres, Robin e Reylé, 2016, aceito para publicação em Astronomy and Astrophysics]. Na parte inferior, do lado direito, são vistas duas estruturas, que são referentes à Pequena e Grande Nuvem de Magalhães.

[7] Flexible Image Transport System

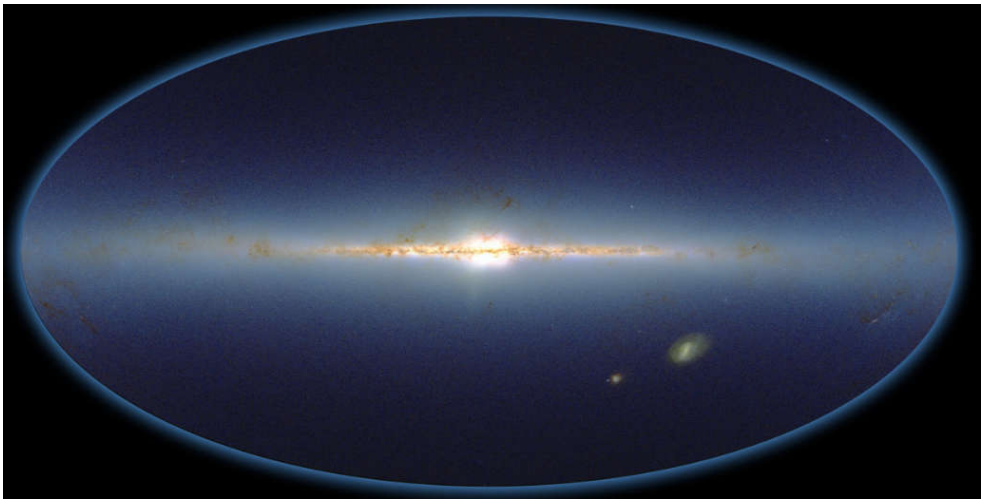

Figura 3.4: Imagem em falsa cor de todo o céu no infravermelho próximo observada pelo 2MASS nos filtros J, H e Ks. Fonte: IPAC/Caltech.

O levantamento anterior ao 2MASS, realizado no final da década de 60, cobriu, aproximadamente, 70% do céu, detectando aproximadamente 5.700 fontes celestes no infravermelho; era o TMSS, Levantamento do Céu em Dois Microns, Two Micron Sky Survey [Neugebauer e Leighton 1969]. Durante esse tempo houve muito avanço tecnológico no que diz respeito a detectores no infravermelho, surgindo equipamentos com a capacidade de detectar fontes mais fracas do que as detectadas pelo TMSS[8].

Com os resultados da exploração feita neste levantamento, informações importantes puderam ser oferecidas para entendimento do céu no infravermelho, bem como respostas a questões feitas sobre a estrutura em larga escala da Via Láctea e do Universo Local [Cutri et al. 2003].

Os dados do 2MASS estão disponíveis em alguns bancos de dados astronômicos. Destacamos os dois principais: CDS de Estrasburgo (França), IPAC-Caltech (Estados Unidos). O CDS (Centro de Dados Astronômicos de Estrasburgo)[9] é um centro dedicado à coleta, análise e distribuição mundial de dados astronômicos e informações relacionadas. Para facilitar a distribuição de informações e imagens dos levantamentos mais importantes, dentre eles o 2MASS, o CDS desenvolveu ferramentas especiais que permitem aos usuários obter, diretamente, dados astronômicos, no formato de catálogo ou de imagens.

Para a visualização de imagens, o CDS desenvolveu o Aladin[10], que permite a seus usuários, dentre outros serviços, visualizar imagens de objetos astronômicos de uma determinada parte do céu, para um dado comprimento de onda, bem como obter um conjunto de imagens de um determinado levantamento, necessitando-se apenas das coordenadas do objeto, ou conjunto de objetos desejados [Bonnarel et al. 2000].

Os catálogos fornecidos pelo 2MASS são resultado do que é denominado por fotometria, ou seja, a medida da quantidade de luz, para um dado objeto, seja ele, estrela, galáxia, etc. A equipe do 2MASS desenvolveu um software específico que

[8] http://www.ipac.caltech.edu/2mass/overview/about2mass.html
[9] http://cdsweb.u-strasbg.fr/CDS-description.gml
[10] http://aladin.u-strasbg.fr/aladin.gml

realiza essas medidas, fornecendo, entre outras propriedades, o tamanho do objeto, o brilho, assim como os erros envolvidos na determinação, a posição do objeto no céu, etc. Entretanto, essa fotometria é sujeita a erros, a exemplo de pixels com problemas na detecção, condições climáticas ruins durante a observação, número muito grande de objetos em um dado campo, etc.

Esses fatores dificultam a medida precisa do brilho de um objeto, assim como a profundidade, entre outros aspectos. Esses problemas são todos identificados por um atributo que indica a qualidade. Conforme mencionado anteriormente, o 2MASS possui, basicamente, dois catálogos, o de fontes pontuais e o de fontes extensas, que podem ser galáxias, ou outros objetos extensos; por vezes, a fotometria do 2MASS classifica como extenso um sistema composto por duas ou mais estrelas.

Em comparação com imagens adquiridas por telescópios na região do óptico, as imagens do 2MASS têm uma aparência notadamente diferente, porque o comprimento de onda no infravermelho próximo do 2MASS é mais sensível a diferentes tipos de estrelas, como as mais frias e as mais vermelhas. O infravermelho próximo (NIR) compreende uma faixa do infravermelho cujo comprimento de onda está entre 0,7–1 a 3 µm. Essa faixa de comprimento de onda permite a observação de estrelas mais distantes na direção do centro de nossa Galáxia, mesmo através da grande concentração de poeira que obscurece seu plano.

O 2MASS é menos sensível à emissão de brilho das nebulosas de hidrogênio e mais sensível ao brilho da poeira estelar nessas regiões [Skrutskie et al. 2006]. As figuras 3.5 e 3.6 apresentam uma comparação entre imagens de galáxias peculiares no óptico e no infravermelho. O contraste de imagens no óptico e no infravermelho pode ser ainda mais realçado quando comparados com regiões de formação estelar. A Figura 3.7 apresenta a imagem adquirida pelos telescópios ópticos do Observatório Anglo-Australiano[11] (AAO) e a mesma imagem adquirida pelos telescópios do 2MASS no infravermelho próximo da região de formação estelar de Carina.

Figura 3.5: AM 0052-321, e AM 0519-611 (no óptico). Fonte: DSS.

[11] https://www.aao.gov.au/about-us/anglo-australian-telescope

Figura 3.6: AM 0052-321, e AM 0519-611 (no infravermelho e falsa cor). Fonte: 2MASS.

Figura 3.7. Comparação entre imagens da nebulosa de Carina adquirida por telescópios óptico (esquerda) e no infravermelho próximo (direita). Fonte: AAO e 2MASS.

Além dos catálogos, existem as imagens que estão disponíveis, e que também podem ser utilizadas, por exemplo, em trabalhos que utilizam softwares de reconhecimento de padrões, usando-as para estudos mais detalhados sobre um dado objeto ou região. As imagens podem ser adquiridas nos três filtros observados pelo 2MASS e podem ser obtidas no tamanho desejado pelo usuário, sendo necessário fornecer um tamanho máximo e um par de coordenadas, p.ex., Galácticas ou equatoriais, a qual será considerado o centro da imagem.

3.3.1 O catálogo de fontes extensas do 2MASS

O catálogo de fontes extensas do 2MASS [Skrutskie et al. 2006] possui 1.647.599 fontes distribuídas por todo o céu e contém todos os objetos que não são fontes pontuais, p.ex., galáxias, nebulosas, etc.

Uma fração significativa desse catálogo contém objetos que foram analisados visualmente por meio das imagens do 2MASS nos filtros J, H e Ks, e classificados como: desconhecido (-2), sem verificação ou não analisado (-1), galáxia (1), objeto não extenso, sistema estelar simples, duplo, triplo ou artefato de imagem (2). Os números entre parênteses representam a flag chamada de visual confirmation (vc), ou confirmação visual.

O catálogo possui 389 colunas com dados relativos a cada fonte nos três filtros, J, H e

Ks. Selecionamos as propriedades mais relevantes para este estudo, tais como: o nome dos objetos, as coordenadas, o brilho em cada filtro, o tamanho e a elipticidade. Como nosso objetivo é trabalhar na análise dos objetos do Hemisfério Sul, foram elaboradas algumas estatísticas para tratar as propriedades globais desses objetos nessa região do céu. A Tabela 3.2, apresenta o número de objetos para declinações $< 0°$ para cada tipo de confirmação visual (vc), para objetos com valores válidos simultaneamente nos filtros J, H e Ks.

Cabe ressaltar que existem 50.926 objetos localizados em latitudes Galáticas, $|b| < 10°$ que serão evitados no presente estudo (não considerados nas contagens da Tabela 3.2), por duas razões básicas: tratar-se de regiões com um alto grau de estrelas de nossa Galáxia, com diferentes valores de brilho, e com alta extinção interestelar, ou seja, regiões com uma alta opacidade devido à distribuição de grãos de poeira interestelar [Amôres e Lépine 2005]. Esses dois efeitos fazem com que seja mais difícil encontrar galáxias atrás do plano de nossa Galáxia [Amôres et al. 2012b].

Tabela 3.2. Quantidade de objetos para cada tipo de confirmação visual (vc, ver texto) para declinação menor do que 0°.

vc	número de objetos
-2	12.403
-1	530.847
1	142.839
2	1.534

Conforme pode ser visto na Tabela 3.2, o maior número de objetos está categorizado como vc = -1, ou seja, objetos não analisados visualmente. Esses objetos podem ser de fato galáxias, estrelas duplas ou triplas, artefatos, etc. A Figura 3.8 apresenta dois mapas (em coordenadas equatoriais e Galácticas), nas quais contamos o número de objetos, considerando-se regiões com tamanho de 25 graus quadrados, ou seja, 5 graus tanto em ascensão reta (RA), como declinação (DEC), sem se levar em conta o limite de completeza, extinção, etc; mas, apenas o número de fontes extensas.

Na Figura 3.8 (parte inferior), nota-se que maior parte dos objetos está concentrada no plano Galáctico, o esgarçamento (mencionado anteriormente) pode ser visto para longitudes de +/- 90°, estendendo-se até latitudes intermediárias (b ~ +/- 30°). A Figura 3.8 (parte superior) que apresenta as contagens em coordenadas equatoriais, também fornece a quantidade de objetos que são utilizados no presente trabalho, para se verificar um dado par de objetos são galáxias em interação, da Categoria 1 ou 2 do Catálogo de AM87. Também pode-se notar uma faixa vertical que se estende por todo Hemisfério Sul para $270° < RA < 300°$, a qual corresponde à objetos localizados no plano Galáctico.

3.4 Obtendo imagens de galáxias peculiares

Obtivemos imagens em falsa cor do 2MASS nos filtros J, H e Ks, tendo como valores as coordenadas equatoriais (J2000) dos objetos das categorias 1 e 2, que não possuíam nenhuma duplicidade com nenhum outro objeto do Catálogo de AM87.

Para tal, elaboramos um script para ser executado dentro Aladin, de forma a obter as imagens em falsa cor, colored, em formato quadrangular, com cada lado tendo 2 minutos de arco, com centro na galáxia peculiar, sem grid de coordenadas e sem retículo, o qual é representado por uma cruz no centro da coordenada do referido objeto do 2MASS.

Como as imagens obtidas pelo Aladin continham marcações próprias para designar a orientação e o tamanho, principalmente nas bordas das imagens, foi necessário configurar um procedimento para retirar tais marcações. Para esse fim, devem-se desmarcar as seguintes opções, antes de executar o script: no menu "Overlay", opção "Overlayed info" e "Target arrow", e marcar a opção "no reticle".

O formato das imagens obtidas por padrão é o png, e o Wndchrm (Seção 4.1) não suporta esse formato; por essa razão foi necessário converter as imagens para o formato tif, com 72 dpi, 482 x 465 pixels de dimensão, padrão RGB, resultando em um arquivo com, aproximadamente 658 kB de tamanho. Para agilizar o processo de conversão de todas as imagens, foi utilizada a funcionalidade "Batch Processing" do Software XnView[12], usado para visualizar e editar imagens, que possui as seguintes configurações: no campo input, foram adicionadas todas as imagens a ser convertidas. No campo de output, foram indicados o diretório de destino e o formato de saída. A conversão de cerca de 800 imagens tem duração aproximada de 3 minutos, em um computador com processador Core I5 com 4GB de RAM.

Das categorias 1 e 2 abordadas no presente trabalho, AM87 faz uma divisão em subcategorias, porém, aqui, as subcategorias não foram consideradas, por conta da baixa quantidade de imagens que poderiam ser usadas para treinamento em cada uma delas.

Tendo como base as coordenadas dos objetos, obtidas no trabalho em preparação de Amôres et al. (2016), das categorias 1 e 2, foram elaboradas listas com o propósito de baixar as imagens do 2MASS dos objetos dessas categorias. Os objetos do Catálogo de AM87, por vezes, podem estar classificados em mais de uma categoria. Em uma estimativa inicial, aproximadamente 30% dos objetos do Catálogo de AM87, possuem duplicidade em relação a classificação em mais de uma categoria.

Incluir na amostra de treinamento galáxias que são classificadas em mais de uma categoria no Catálogo de AM87 pode levar o programa de classificação a resultados imprecisos. Dessa forma, foram obtidas imagens de cada uma das duas categorias evitando-se objetos com duplicidade de entradas, em um total de 212 e 548 objetos para as categorias 1 e 2, respectivamente.

As figuras 3.9 e 3.10 apresentam objetos classificados na categoria 1 e 2 com seus

[12] http://www.xnview.com/en/

respectivos nomes de acordo com uma codificação interna que foi feita, podendo-se, a qualquer momento, recuperar o nome original do objeto, tal como fornecido no Catálogo de AM87. Nessas figuras também pode-se ver imagens com muitas estrelas, objetos com brilho fraco, etc, que foram descartadas no processo de seleção de imagens descrito a seguir.

3.5 Seleção das imagens para a amostra de treinamento

Iniciamos um processo de seleção das melhores imagens de ambas as categorias, com o propósito de se ter imagens de galáxias que representassem de forma significativa, as galáxias de ambas as categorias, de forma a se elaborar um conjunto robusto dessas categorias que serão usadas como amostra de treinamento no Wndchrm.

As imagens foram todas analisadas visualmente, e a seleção foi realizada de forma a se considerarem imagens com o menor número possível de estrelas, pois esses objetos não fazem parte desse estudo, e, são, de forma geral, objetos mais brilhantes do que as galáxias, podendo-se tornar uma fonte de contaminação para a classificação das galáxias. Denominamos esses objetos de objetos de foreground, dos quais a maioria são estrelas dentro de nossa Galáxia.

Como a grande parte das imagens possuem, visualmente, até três ou quatro estrelas no campo, foi tolerada a presença de até essa quantidade de estrelas (não brilhantes e que aparecem na inspeção visual), desde que não estejam próximas das galáxias interagentes. Um aspecto importante é que o brilho das estrelas seja fraco. Foram descartadas imagens de galáxias não centralizadas, total ou parcialmente fora do campo visual ou com um grande tamanho angular, ou seja, maior ou aproximadamente igual a 2 minutos de arco. Como resultado, desse processo foram obtidas 65 e 73 imagens das categorias 1 e 2, perfazendo um total de 138 imagens.

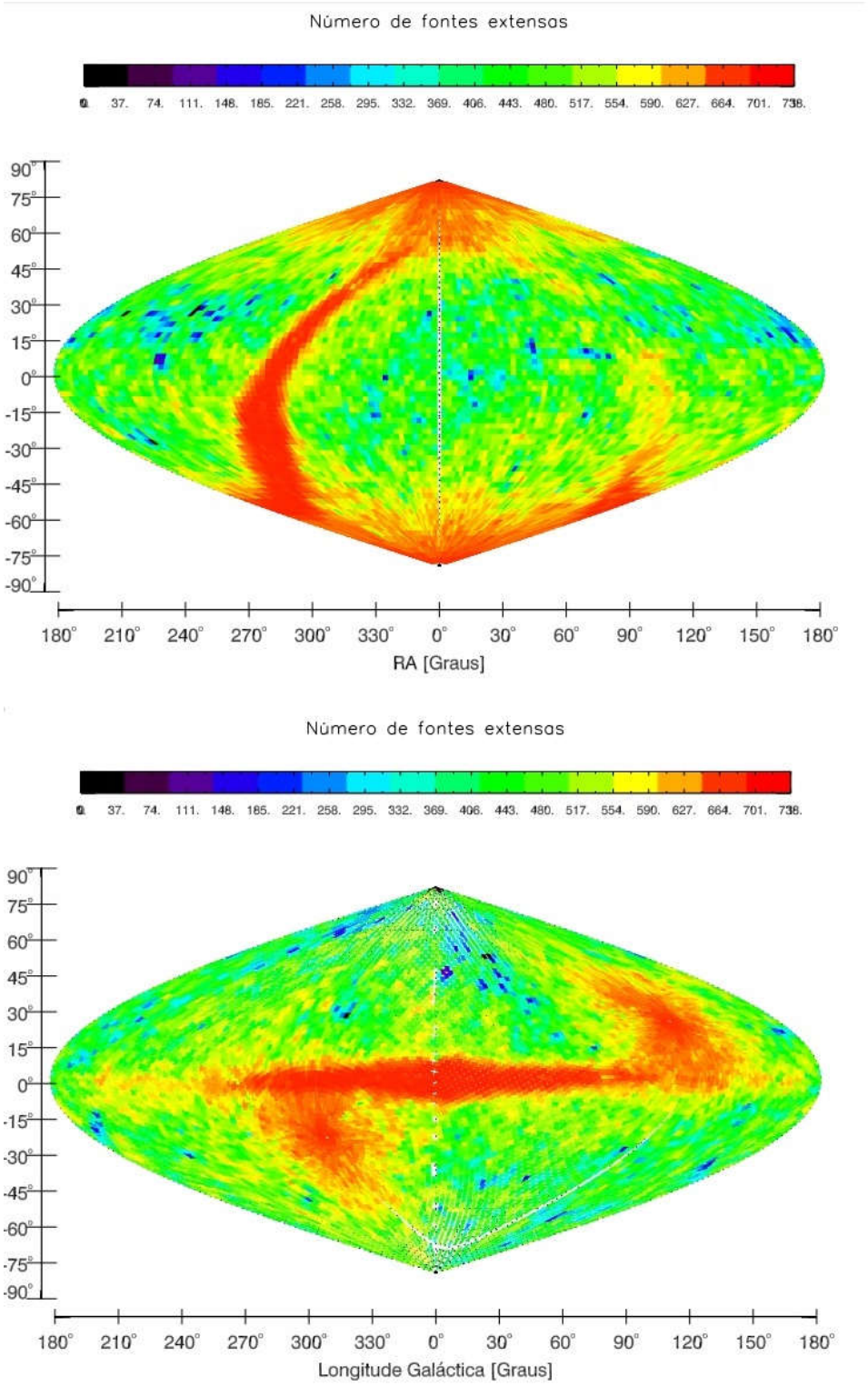

Figura 3.8. Distribuição das fontes extensas do 2MASS em projeção aitoff, para todo o céu que foi dividido em regiões, com tamanho com 25 graus quadrados. Parte superior (coordenadas equatoriais), parte inferior (coordenadas Galácticas).

Figura 3.9: Exemplos de imagens em falsa cor obtidas do 2MASS de galáxias da Categoria 1 do Atlas AM87.

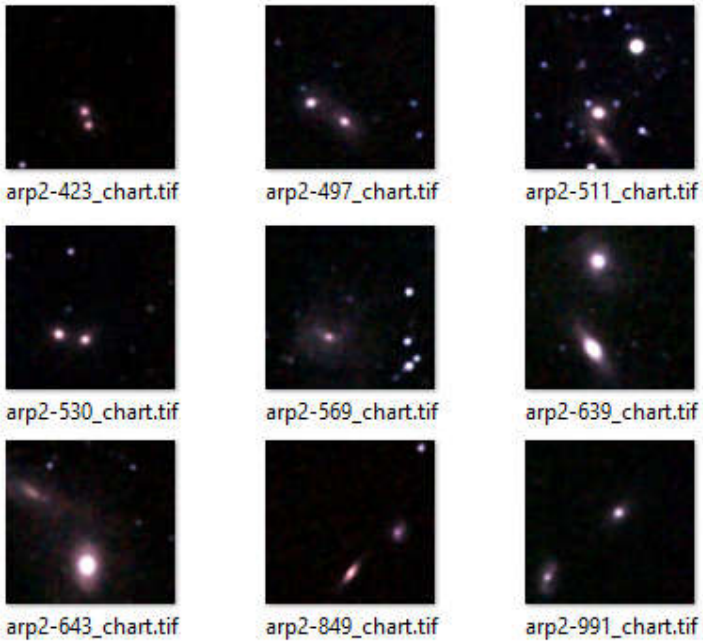

Figura 3.10: Exemplos de imagens em falsa cor obtidas do 2MASS de galáxias da Categoria 2 do Atlas AM87.

Na Figura 3.9, vemos exemplos de algumas imagens que foram obtidas nesse procedimento, dentre as quais, algumas que foram descartadas para compor a amostra de treinamento da Categoria 1. As imagens arp1-236_chart.tif (Figura 3.9, segunda linha da terceira coluna) e arp1-323_chart.tif (Figura 3.9, terceira linha da segunda coluna) foram descartadas por possuir muitas estrelas próximas às candidatas a galáxias interagentes. Já a imagem arp1-209_chart.tif (Figura 3.9, segunda linha da segunda coluna), além de possuir muitas estrelas, tem a galáxia interagente fora do campo visual.

Na Figura 3.10, vemos de maneira similar, imagens que foram descartadas para a amostra de treinamento da Categoria 2. Notadamente, a imagens arp2-511_chart.tif (Figura 3.10, primeira linha, terceira coluna) possui muitas estrelas e, na imagem arp2-569_chart.tif (Figura 3.10, segunda linha, segunda coluna), temos umas das galáxias interagentes fora do campo visual.

Figura 3.11: Exemplo de imagens da Categoria 1 selecionadas para a amostra de treinamento.

A Figura 3.11 apresenta exemplos de imagens da amostra de treinamento para a Categoria 1. São imagens nas quais as galáxias interagentes podem ser identificadas com certa facilidade, tendo a galáxia principal mais do que o dobro do tamanho da galáxia menor. A Figura 3.12, exibe algumas das imagens da amostra de treinamento da Categoria 2.

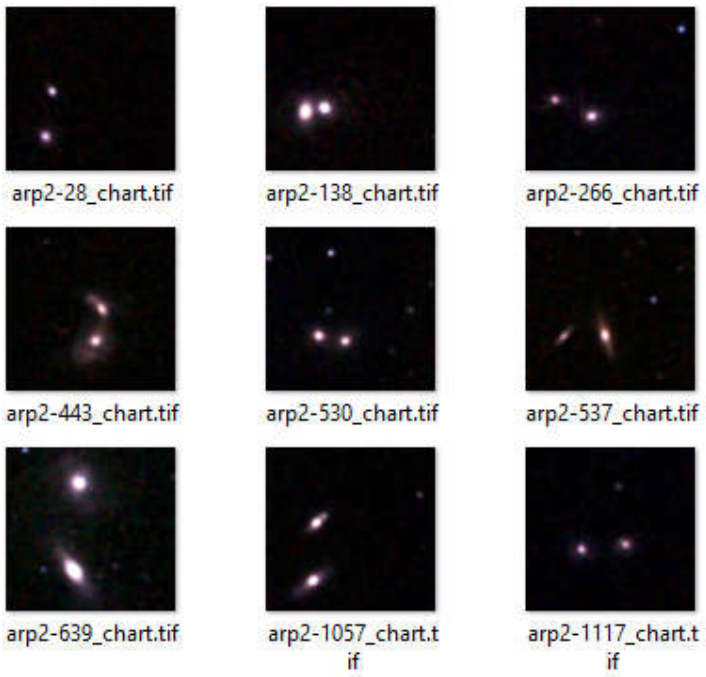

Figura 3.12: Exemplo de imagens da Categoria 2 selecionadas para a amostra de treinamento.

3.6 Análise das propriedades das imagens

A análise visual de imagens que fazem parte da amostra de treinamento para as categorias 1 e 2, assim como das descartadas, é pertinente, pois pode revelar suas características importantes, tais como, o número de estrelas das mesmas e, dessa forma, estimar um limite máximo de estrelas permitido por campo e/ou em função da magnitude para cada uma delas. Da mesma forma, pode-se verificar como a quantidade da extinção interestelar para um dado campo pode interferir na identificação dos objetos. Esta seção apresentará a comparação da extinção para as imagens selecionadas que contém objetos representativos das categorias 1 e 2, e das imagens recusadas no processo de inspeção visual, assim como da contagem de número de estrelas por imagens.

3.6.1 A extinção interestelar

A extinção interestelar é causada por grãos de poeira, compostos basicamente por grafite, silicato e carbono que estão localizados em nuvens moleculares bem misturadas ao gás. A estimativa da extinção é importante, tendo em vista que ela atenua a luz dos objetos, devido, basicamente, a dois processos físicos: a absorção e o espalhamento [Amôres e Lépine 2005 e referências contidas no artigo]. Dado ao fato de que seu valor pode variar de, aproximadamente, zero até 30 magnitudes no filtro V (a extinção nesse filtro é denominada por A_V), é extremamente importante sabermos seu valor na direção das candidatas a galáxias interagentes para fazermos uma correção no valor da magnitude das galáxias. Dessa forma, o valor da magnitude final ($mK_S´$) para um dado objeto, no filtro K_S, é o resultado da subtração da

extinção (A_{Ks}) da magnitude observada (mK_S) do objeto, ou seja:

$$mK_S' = mK_S - A_{KS} \quad (3.1)$$

A extinção varia com o comprimento de onda, sendo menor para comprimentos de ondas maiores. Dessa forma, a extinção no infravermelho próximo, no filtro K_S é, aproximadamente, 10% da encontrada no visível, no filtro V, ou seja, $A_{KS} = A_V$ x 0,112 [Schlegel et al. 1998, SFD]. Alguns modelos e/ou mapas fornecem a extinção em excesso de cor, ou seja, E(B-V), esse valor está relacionado com a extinção A_V, por meio da razão (R_V), ou seja, entre a extinção total e a seletiva, por meio da relação A_V = 3,05 x E(B-V). O valor de 3,05 é o valor mais comum [Whittet 1992] encontrado no meio interestelar em regiões difusas e usado em estimativas gerais. Esse valor pode alcançar valores de até 5,5 em densas regiões, mais notadamente em nuvens escuras, regiões com alta formação estelar.

A correção da extinção é também particularmente importante para definirmos de forma mais precisa, a magnitude limite das estrelas nas imagens, assim como das candidatas a galáxias interagentes, permitindo-nos inferir até qual magnitude funciona de maneira adequada. Existem vários modelos e mapas que descrevem a extinção em duas e/ou três dimensões [Amôres et al. 2012a], entretanto, um dos mais usados é o mapa de SFD, que está disponível para todo o céu e que fornece bons valores para a extinção interestelar. Nos mapas de SFD, temos a extinção integrada por toda a linha de visada, ou seja, toda a extinção de nossa Galáxia para um par de coordenadas, ideal para o estudo que pretendemos, que consiste de objetos que estão fora de nossa Galáxia. A extinção de um ponto imediatamente exterior da nossa Galáxia até os objetos de nosso estudo é considerada nula.

Esse mapa tem sérios problemas [Amôres & Lépine 2005 e 2007, entre outros] no plano de nossa Galáxia, devido a algumas simplificações feitas pelos autores, que fornecem resultados imprecisos para valores de extinção superiores à 0,5 mag. Fora do plano Galáctico, essas regiões com alta extinção estão, principalmente, localizadas em nuvens moleculares, onde a concentração de gás, poeira e estrelas é muito maior do que em outras direções fora do plano Galáctico. Entretanto, para as regiões de estudo no presente trabalho (fora do plano Galáctico), ele fornece bons valores e possui aproximadamente 10.000 citações.

Os problemas com os mapas de extinção de SFD se devem, basicamente, a algumas suposições feitas pelos autores na elaboração de seus mapas, dentre as quais: a resolução dos mapas do DIRBE/COBE que foram usados pelos autores, uma única temperatura de aproximadamente 18 K para todas as regiões, o que não reflete a distribuição real de nuvens moleculares com temperaturas maiores e localizadas em regiões com alta densidade de poeira e com formação estelar. Arce & Gordman (1998) em um estudo da extinção dos mapas de SFD em regiões de formação estelar, indicam que os valores dos autores começam a ficar discrepantes a partir de A_V = 0,5 mag, ou seja E(B-V) ~ 0,15 mag.

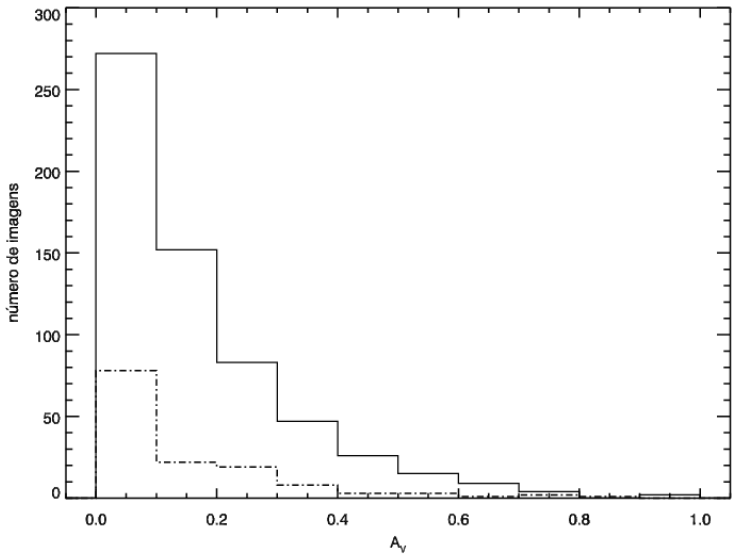

Figura 3.13: Distribuição da extinção interestelar (A_V) segundo os mapas de SFD para imagens selecionadas das categorias 1 e 2 (linha tracejada) e recusadas (linha sólida).

Tendo como base as coordenadas que representam o centro das imagens selecionadas (categorias 1 e 2) e recusadas, obtivemos a extinção interestelar usando o método de SFD na direção dessas imagens. Para tal, utilizamos a ferramenta de Observatório Virtual, GALExtin[13] [Amôres et al. 2012a] que, atualmente, fornece o valor da extinção para aproximadamente 10 modelos/mapas. Cabe ressaltar que, devido à resolução do mapa de SFD, a variação da extinção em dois minutos de arco é praticamente desprezível neste mapa.

A Figura 3.13 apresenta um histograma com a distribuição da extinção na direção das imagens selecionadas e recusadas. Nota-se, para ambas as distribuições, que a maior concentração é para $A_V < 0{,}1$ mag. Entretanto, para as imagens recusadas, o seu número por intervalo decresce de forma suave, até, aproximadamente, 0,8 mag, no que, para as imagens selecionadas, o valor mantém-se praticamente constante no intervalo de 0,1 mag $< A_V \leq 0{,}3$ mag, sendo quase desprezível para $A_V > 0{,}4$ mag. Os valores de extinção superiores a, aproximadamente, $A_V > 0{,}3$ mag, podem ser considerados elevados para regiões intermediárias e de elevadas latitudes Galácticas [Burstein 2003]. Da análise da Figura 3.13, vemos claramente que apenas uma pequena fração tem $A_V > 0{,}4$ mag e $A_V > 0{,}5$ mag, representando 6% e 8% das imagens com os objetos selecionados, respectivamente.

Podemos concluir, da análise da Figura 3.13, que a extinção usada de maneira isolada não é suficiente para discriminar imagens que possam ser examinadas para encontrarmos candidatas a galáxias peculiares interagentes, dada a grande concentração de objetos com extinção $A_V < 0{,}1$ mag para imagens selecionadas e descartadas. Entretanto, ela pode ser uma importante ferramenta para ser usada como auxiliar na seleção de imagens, particularmente neste caso de estudo. Dessa forma podemos descartar campos pertencentes a imagens com $A_V > 0{,}5$ mag.

[13] http://www.galextin.org

3.6.2 Análise da emissão estelar

Outro importante indicativo das características das imagens das candidatas a galáxias peculiares das categorias 1 e 2 que se espera encontrar, está relacionado ao número de estrelas identificadas próximos a esses objetos. Conforme mencionado na Seção 3.5, a escolha das imagens selecionadas, feita de maneira visual, foi realizada de forma a se evitar um grande número de estrelas, assim como de estrelas brilhantes. Esses objetos não fazem parte de nosso estudo.

Não fizemos um processamento das imagens de forma a tentar minimizar o efeito dessas estrelas nas imagens. Preferimos, neste método, escolher imagens com o menor número possível de estrelas e ter informações das magnitudes das mesmas, de forma a excluir imagens que tenham estrelas com essas características. Em um futuro estudo, esses objetos podem ser considerados em nossa análise com métodos apropriados para retirar as estrelas das imagens.

Para cada uma das imagens selecionadas das categorias 1 e 2 e das recusadas, obtivemos as estrelas existentes dentro de dois minutos de arco da imagem. Utilizamos o Catálogo de fontes pontuais do 2MASS[14] nos filtros J, H e K_S. Um aspecto importante é que, em estudos que visam à determinação, p.ex., de parâmetros de estrutura de nossa Galáxia [Robin et al. 2003 2012, 2014], tendo como base contagem de estrelas, é necessário fazer uma filtragem dos objetos, levando-se em conta, entre outras características relacionadas à qualidade da fotometria. Dadas as condições observacionais e/ou características do campo, um determinado objeto pode ser observado em um filtro e, não, em outro, o que pode ser devido à variação do seeing[15]. O catálogo, entre outros aspectos, possui informação sobre a qualidade da observação em cada filtro, a razão sinal/ruído, etc.

Em nosso estudo, esses aspectos não são relevantes, pois queremos ser capazes de identificar o maior número possível de estrelas, mesmo que estas não possuam boa qualidade de observação, não sejam observadas nos três filtros, etc. Uma estrela detectada no filtro J e, não, nos filtros H e K_S em uma imagem em falsa cor do 2MASS, certamente irá aparecer, mesmo que de maneira bem fraca, possivelmente imperceptível em uma análise visual, o mesmo sendo possível em outras combinações de filtros. A contribuição de tais fontes será detectada de alguma forma pelo Wndchrm e/ou outro programa de reconhecimento de padrões, sem que eliminemos esses objetos.

[14] http://www.ipac.caltech.edu/2mass/releases/second/doc/sec2_2.html
[15] efeitos de distorções produzidas pela atmosfera sobre a qualidade da imagem de um corpo celeste observado a partir da Terra (http://er.jsc.nasa.gov/seh/s.html).

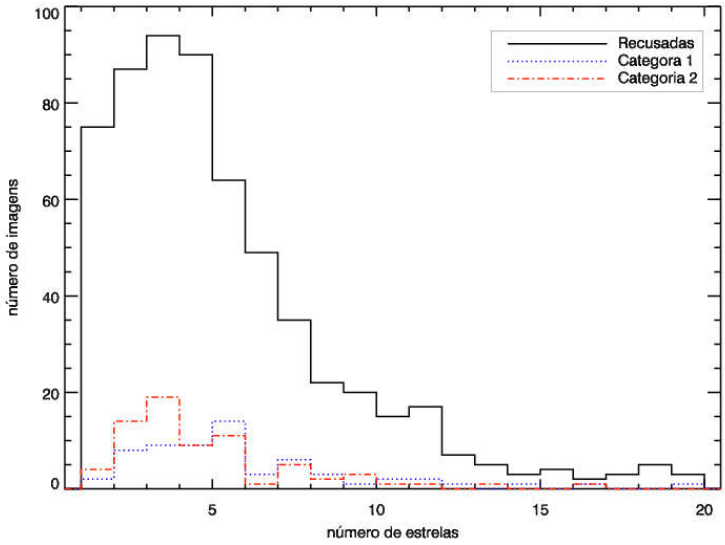

Figura 3.14: Distribuição do número de estrelas para as imagens das categorias 1 e 2 e recusadas.

A Figura 3.14 apresenta a frequência do número de objetos encontrados, independentemente de sua magnitude para cada imagem, considerando-se as categorias 1 e 2, e de imagens recusadas. Nota-se que, para as imagens selecionadas das categorias 1 e 2, representadas pelas linhas coloridas tracejadas e pontilhadas da Figura 3.14, o número de imagens selecionadas cai abruptamente após o número de estrelas ser seis por imagem. No entanto, para as imagens recusadas, ainda temos um número considerável de imagens que possuem mais de seis estrelas. Essa análise fornece uma estimativa inicial, existindo uma relação, nos dados da fotometria do 2MASS, entre o número de estrelas e as imagens selecionadas e recusadas.

Uma análise mais detalhada pode ser feita levando-se em conta a magnitude das estrelas observadas, de forma a ser feito um histograma do número de estrelas encontradas nas imagens que representam as categorias selecionadas e recusadas, respectivamente. Essa análise pode mostrar se existe uma dependência com a magnitude, fornecendo, dessa forma, um método mais sensível para se considerar se uma dada imagem tem ou não uma contaminação significativa de estrelas, diferentemente das imagens usadas como amostra de treinamento, devendo, desta forma, ser recusada por nosso método.

Para tal, realizamos contagem de estrelas para os objetos localizados nas três categorias de imagens para a faixa compreendida de $8{,}0 < K_S < 18{,}5$ com intervalos de 0,5 mag. Antes de realizar a contagem, efetuamos a correção da extinção, usando o mapa de SFD (valores apresentados na Figura 3.13) na magnitude no filtro K_S, fazendo uso da Expressão 3.1.

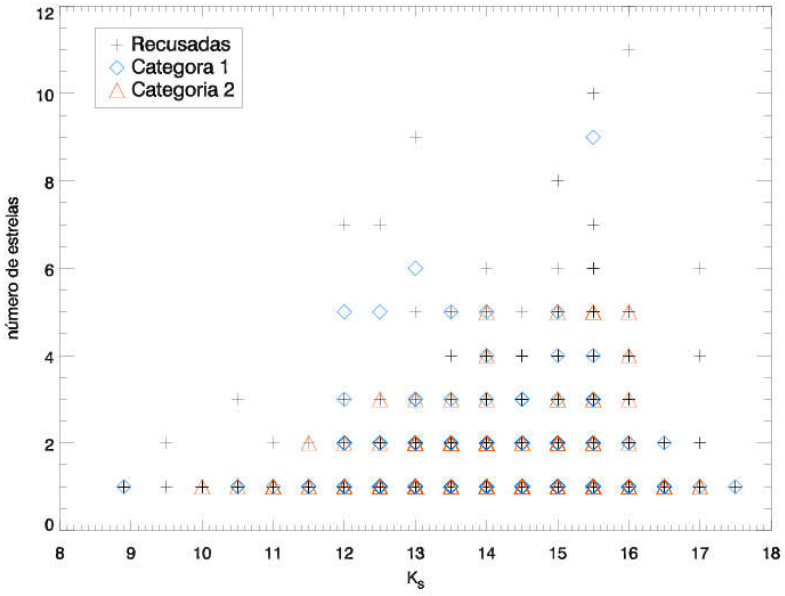

Figura 3.15: Distribuição do número de estrelas em função da magnitude para as estrelas das imagens selecionadas das categorias 1 e 2 e das recusadas. Os símbolos são indicados na legenda.

A Figura 3.15 apresenta a contagem de estrelas em função da magnitude para as três categorias de imagens. Nota-se que, a grande maioria de pontos é devida ao número de estrelas nas imagens descartadas, representadas por cruzes. Alguns símbolos possuem uma forma mais destacada, devido a se ter mais de um ponto para a contagem. Vê-se, claramente, um corte, ou seja, uma separação entre as contagens das imagens das categorias 1 e 2 e das recusadas, a partir de 6 estrelas, para a faixa de $K_S \geq 12$ mag, existindo apenas um ponto (Categoria 2) para magnitude $K_S = 14{,}0$ mag, o que é claramente uma exceção pois contém uma ou duas estrelas, o mesmo para objetos mais brilhantes do que $K_S = 12$ mag. Ressaltamos que, como o objetivo deste trabalho é considerar se uma estrela, com detecção pela fotometria, tenha uma contribuição, mesmo que fraca, na imagem, não estamos preocupados com o limite de completeza do 2MASS, que é de 14,3 mag para o filtro K_S.

Dessa forma, podemos sumarizar as condições para que, dada uma imagem qualquer, esta esteja compatível com as características da componente estelar:

i. $K_S < 12$ (até uma estrela)

ii. $12 <= K_S < 16$ (até cinco estrelas)

iii. $16 <= K_S <= 17$ (até uma estrela)

Capítulo 4

O Método para reconhecimento de galáxias interagentes

O capítulo apresenta o método para a identificação de candidatas a galáxias peculiares das categorias 1 e 2 do Catálogo de AM87. Em uma primeira etapa, é apresentado com detalhes, o software do qual fizemos uso, denominado Wndchrm. Posteriormente, é apresentada a elaboração das amostras de treinamento para as respectivas categorias. A validação do método, é realizada com imagens no infravermelho próximo de galáxias obtidas do 2MASS para uma área de 573 graus quadrados no Hemisfério Sul, baseado no Catálogo de fontes extensas do 2MASS.

4.1 O Wndchrm

Shamir e Wallin (2014) apresentaram um método para detecção automática de galáxias peculiares composto por algumas etapas, tais como: transformação para classificação de cores baseada na lógica fuzzi [Shamir 2006], transformação binária de Otsu [Otsu 1979], determinação de pixel de primeiro plano por contagem de pixel com vizinhança-4 [Shamir 2011], algoritmo de função de detecção de ponto de dispersão [Shamir e Nemiroff 2005] e a ferramenta de software de análise e classificação de imagem Wndchrm, adaptado para analisar a morfologia de galáxias peculiares [Shamir et al. 2013].

A princípio, o Wndchrm foi concebido para analisar e classificar imagens biológicas [Shamir et al. 2008]. Nessa época, análises em imagens biológicas eram um campo emergente, abrangendo um grande espectro de aplicações em pesquisas biológica e clínica. O Wndchrm permitiu a pesquisadores que não têm conhecimento profundo em informática usar o computador na tarefa de análise de seus bancos de imagens.

O Wndchrm foi aplicado também na análise de pinturas e associação das mesmas a um determinado pintor ou a uma escola de arte [Shamir et al. 2010]. Montou-se um conjunto de imagens de quadros formado por nove diferentes pintores, representando três escolas de artes com três pintores de cada uma. Nesse experimento, foram feitos dois estágios de treino e classificação, um para classificar cada imagem como obra de um dos nove pintores e outro para classificar as imagens em uma das três escolas de arte.

Em 2014, Shamir e Wallin aplicaram seu classificador de imagem na tarefa de identificar e classificar imagens de pares de galáxias peculiares do céu do Hemisfério Norte. Usando um conjunto de imagens do SDSS[16] com aproximadamente 400.000

[16] http://www.sdss.org/

imagens, encontraram, aproximadamente, 26.000 delas com morfologia similar à de fusão de galáxias.

4.1.1 Funcionamento

O Wndchrm é um aplicativo de código aberto, utilizado para análise e classificação de imagens, com possibilidade de processamento de grandes conjuntos dessas, de diversas áreas [Shamir et al. 2008]. Inicialmente desenvolvido para análises de imagens biológicas, o Wndchrm foi adaptado para analisar outros tipos de imagens como, por exemplo, imagens artísticas e astronômicas.

Desenvolvido em C++, o Wndchrm usa as bibliotecas fftw e libttif, ambas de livre acesso, para fazer o reconhecimento de padrões e extração de atributos. O software não tem interface gráfica; toda a interação com o usuário é feita através de linha de comando [Shamir et al. 2008].

O Wndchrm é baseado em três etapas. A primeira etapa analisa um conjunto inicial de imagens agrupadas de acordo com determinados aspectos, extraindo e salvando as características de cada grupo, criando-se assim um arquivo de atributos, essa fase é chamada de treinamento. Após essa fase, o Wndchrm oferece a opção de testar as características extraídas do conjunto de imagens na fase de treino, quando imagens de cada grupo são classificadas de acordo com os atributos extraídos. Esse processo permite verificar o grau de precisão em que as imagens são classificadas.

Por fim, um conjunto de características pode ser aplicado a todas as imagens que se deseja classificar, para que o aplicativo compare os atributos de cada uma com os atributos de cada categoria, indicando em qual categoria aquela imagem pode ser melhor enquadrada. O classificador faz as comparações utilizando o método da distância ponderada entre os vizinhos WND (Weighted Neighbor Distances) [Duda et al., 2000], mas pode ser alterado, via parâmetro, para usar o método da distância ponderada entre os vizinhos mais próximos WNN (Weighted Nearest Neighbor) [Cost e Salzberg, 1993].

4.1.2 Características

Para a extração de atributos e reconhecimento de padrões, o software utiliza os seguintes algoritmos:

- Transformação randômica de atributos - calculado para ângulos 0, 45, 90, 135 graus, proporcionando um total de 12 atributos de imagem [Lim, 1990].
- Desigualdade de Chebyshev – Um histograma de 32 partes de um vetor (1x400) produzido por Chebyshev [Gradshtein e Ryzhik, 1994]. Extrai 20 atributos de imagem.
- Filtros de Gabor – baseado nas funções harmônicas gaussianas [Gabor, 1946], com 7 atributos de imagem.
- Histogramas multi-scalas [Hadjidementriou et al., 2001] com 24 atributos de imagem.
- Primeiros quatro momentos [Shamir et al., 2008] com 48 atributos de imagem.

- Atributos de texturas de Tamura [Tamura et al., 1978] – 6 atributos de imagem.
- Atributos estatísticos de bordas com calculado pelo gradiente de Prewitt [Prewitt, 1970].
- Estatisticas de Objetos – baseado na máscara binária de Otsu para imagens [Otsu, 1979].
- Atributos de Zernike [Teague, 1980] com 72 atributos de imagem.
- Atributos de Halarick [Haralick et al., 1973] com 28 atributos de imagem.
- Atributos de Fourier-Chebyshev [Orlov et al., 2007] com 23 atributos de imagem.

4.1.3 Configuração do software Wndchrm

Para os testes iniciais, a instalação e a configuração do Wndchrm foram feitas no Linux Ubuntu 14.04. Os arquivos são baixados diretamente da página do aplicativo na internet indicado em Shamir e Wallin (2014). As bibliotecas libtiff, libtiff4 dev e fftw 3.1, necessárias para o funcionamento do Wndchrm, foram instaladas pela central de programas do Ubuntu. Foi necessário instalar, também, a biblioteca G++ e o autoconf para Ubuntu.

Para evitar um erro na configuração, foi feita a seguinte alteração no makefile: acrescentar a linha <AM_CXXFLAGS = -Wall -g -O2 –fpermissive>, logo abaixo da linha <AM_CPPFLAGS = -fPIC>. Após essa alteração, foi executado o ./configure e o make da pasta do Wndchrm.

4.1.4 Parâmetros utilizados

- r : altera o percentual de imagens usado para treino e teste. Padrão 75% para treino, 25% para teste. Ex: -r0.4 (40% para teste e 60% para treino).
- c: com esse parâmetro informado, o Wndchrm extrai características de cores das imagens.
- m: permite a execução de múltiplas instâncias do Wndchrm, simultaneamente, em diferentes processadores.
- l: parâmetro do tipo on/off; quando esse parâmetro é usado, o Wndchrm extrai uma quantidade maior de características das imagens, 2.659, contra 1.025 na execução padrão.
- fN: define o número máximo de atributos da imagens utilizados no processo de classificação. N pode variar de 0 a 1.
- dN: reduz o tamanho das imagens analisadas (downsample). Ex. –d50 reduz em 50% o tamanho da imagem.
- tN: divide as imagens em células para que cada uma tenha suas características extraídas independentemente.

- gN: que define uma faixa dinâmica de bits por pixel. Ex. -g16 define 16 bits por pixel.
- MN: faz com que todos os pixels com valores menores do que o limiar de Otsu vezes N sejam zerados. Ex. M0.6.

4.1.5 A Utilização

As execuções do Wndchrm foram feitas no sistema de computação de alto desempenho do INCT-A, localizado no IAG/USP, no Cluster SGI Altix ICE 8400 (Alphacrucis)[17], que utiliza processadores AMD Opteron 6172 Magny-Cours (12 núcleos), 2.1 GHz/12MB cache. O cluster possui 2304 cores (192 processadores), com 2 GB de memória por core, 48 GB por nó, com um total de 4.6 TB. Ao todo, o cluster possui 96 nós, divididos em 6 IRUs (Individual Rack Unit). O cluster ainda dispõe de uma capacidade de 32 TB para armazenamento de dados em discos.

Foram feitas execuções com 8 e 16 nós, sem diferença relevante no tempo de processamento estre elas. Optou-se, então, por usar apenas 8 nós do cluster, nas execuções futuras.

Na fase de treinamento, o Wndchrm examina um grupo de imagens e extrai as características e atributos das mesmas, baseados nos algoritmos citados acima. O usuário deve, antes, escolher as imagens que melhor representam uma classe que se deseja criar. A classe será representada por um diretório de arquivos, e todos os outros diretórios-classes deverão estar dentro de um diretório pai.

No final do processo de treinamento, as características e atributos extraídos das imagens de cada classe são armazenados em um arquivo chamado arquivo de atributos com extensão .fit.

Este é o comando para fazer o Wndchrm treinar a partir de um determinado conjunto de classes:

```
% Wndchrm train [opções] imagens arquivo_atributos
```

no qual:

- opções- são os parâmetros (opcional)
- imagens - caminho até o diretório onde se encontram os sub-diretórios que representam as classes
- arquivo_atributos- o nome do arquivo fit que será gerado (caminho inteiro até o nome)

Na fase de teste, o usuário do Wndchrm pode avaliar o treinamento feito. Na fase de treinamento, o Wndchrm usa (por padrão) 75% das imagens de uma determinada classe para extrair os atributos, e 25% (das imagens) são reservados para testar o treinamento. Por exemplo, se o usuário criou 3 classes de 10 imagens cada, o Wndchrm extrairá atributos de 8 imagens de cada classe na fase de treino. Na fase de teste, o Wndchrm tentará classificar as 6 imagens (2 de cada classe) não usadas, em

[17] https://lai.iag.usp.br/projects/lai/wiki/Cluster_SGI_Altix_ICE_8400

uma das classes, possibilitando, dessa forma, ao usuário, uma chance de medir a precisão da classificação com base no arquivo de atributos do treino.

Este é o comando para fazer um Wndchrm testar um treinamento:

```
% Wndchrm test [opções] arquivo_atributos [relatório]
```

no qual:

- opções- são os parâmetros (opcional).
- arquivo_atributos- o nome do arquivo fit gerado no treinamento a ser testado (caminho inteiro até o nome).
- Relatório – nome de um arquivo html que conterá informações sobre o resultado dos testes (opcional).

Após verificar os resultados dos testes, o usuário do Wndchrm poderá rever as classes ou partir para a classificação do seu banco de imagens. Nessa fase, o Wndchrm analisa cada imagem, extraindo seus atributos e comparando com os atributos do arquivo fit de cada uma das classes, gerando um número, que varia de 0 a 1, representando o grau de similaridade entre os atributos da imagem e os atributos de cada classe. O maior grau de similaridade será a classe escolhida para aquela imagem.

Este é o comando para classificar um conjunto de imagens no Wndchrm:

```
% Wndchrm classify arquivo_atributos imagens
```

no qual:

- arquivo_atributos- o nome do arquivo fit gerado no treinamento (caminho inteiro até o nome).
- imagens - caminho até o diretório onde se encontram as imagens a ser classificadas.

4.2 Determinação dos melhores parâmetros do Wndchrm para a amostra de treinamento

Conforme descrito na seção anterior, o Wndchrm possui parâmetros que afetam o resultado tanto da fase de treinamento, como na de classificação. Nosso próximo passo consistiu em analisar o impacto desses parâmetros com as imagens da amostra de treinamento escolhida (Seção 3.5). A partir da análise das características e efeitos de cada parâmetro, foram escolhidos os que seriam usados nos testes de execução do Wndchrm para o treinamento e classificação das imagens que formam o conjunto de imagens das categorias 1 e 2. A seguir, descrevemos os parâmetros considerados em nossa análise.

Devido ao fato das imagens usadas neste trabalho não terem formas complexas como polígonos, tendo um fundo escuro e os objetos na sua maioria circulares ou elípticos, o parâmetro f, permite um controle no sentido de reduzir os atributos de imagens, considerados na classificação, ficando apenas com os mais significativos. Neste sentido, foram realizados testes desse parâmetro, com os seguintes valores: 0,05; 0,1; 0,15; 0,20; 0,30; 0,40; 0,50. Esse parâmetro é usado apenas na fase de classificação.

O parâmetro l foi escolhido para verificar o comportamento do Wndchrm, usando diferentes quantidades de atributos. Com esse parâmetro informado são extraídos 2.659 atributos de uma imagem ou classe de imagens, sem ele são extraídos apenas 1.025 atributos. Essa verificação é feita, tanto na fase de treinamento como na de classificação.

Como as imagens utilizadas neste trabalho são resultados de observações no infravermelho nos filtros J, H e Ks, em uma imagem em falsa cor, o parâmetro c também foi considerado. Com esse parâmetro habilitado, o Wndchrm considera informações sobre as cores das imagens.

Embora nossas imagens não tenham uma dimensão muito grande (uma matriz típica de 500 x 500 elementos), a redução do tamanho por meio do parâmetro d, poderá trazer o retorno de um menor tempo de processamento. Valores de reduções em 70, 50 e 20 por cento foram escolhidos para testes.

As imagens com as quais trabalhamos têm 24 bits por pixel. Por meio do parâmetro g, que define uma faixa dinâmica de bits por pixel. Foram escolhidos os valores de 8, 16, 32 bits por pixel. Esse parâmetro ainda permite normalizar o histograma, sendo verificada esta possibilidade. As duas formas, com e sem a normalização foram testadas para cada valor utilizado.

Pelo fato das imagens que compõem a nossa amostra de treinamento, possuírem objetos como estrelas, que geralmente estão no plano de frente em relação às galáxias, que são o objeto de nosso estudo, o parâmetro M pode ter eficácia. Esse parâmetro faz com que todos os pixels com valores menores do que o limiar de Otsu vezes N sejam zerados (pixel preto). Foram escolhidos, os seguintes valores de teste para N, variando de 0,2 até 1,0; em intervalos de 0,2.

O parâmetro w foi escolhido para que se possa verificar a diferença entre o método de classificação pela distância ponderada entre os vizinhos mais próximos - WNN (Weighted Nearest Neighbor) e o método das distâncias ponderadas entre os vizinhos - WND (Weighted Neighbor Distances), na comparação entre um vetor de atributos de uma imagem e de uma classe.

Após a seleção para testes dos parâmetros e suas respectivas faixas de valores, o próximo passo foi gerar o arquivo (em extensão fit) de treinamento do Wndchrm, com os parâmetros que podem ser usados nesta fase. Sendo assim foi gerado um arquivo fit para cada um destes casos: c, l, d20, d50, d70, g8, g8#, g16, g16#, g32, g32#, M02, M04, M06, M08, M1, bem como um arquivo fit sem nenhum parâmetro previamente selecionado, ou seja, com os parâmetros default do Wndchrm.

Na fase de classificação, foi feita uma execução para cada um dos parâmetros escolhidos com seu respectivo arquivo fit. Os parâmetros que não são usados na fase de treinamento, e não geram arquivo fit correspondente, foram usados na classificação com base no arquivo fit gerado sem nenhum parâmetro. Os parâmetros f (com valores 0,05; 0,10; 0,15; 0,20; 0,30; 0,40; 0,50) e o w são usados apenas na fase de classificação.

Após a execução dos scripts com seus respectivos parâmetros, os resultados obtidos foram organizados na Tabela 4.1, que apresentam as simulações para cada vez que o Wndchrm foi utilizado para cada parâmetro. O tempo em minutos também é

fornecido, vemos que a simulação que levou menos tempo foi a simulação s5.1, por outro lado a que levou mais tempo foi a s2.1. As simulações que obtiveram a maior quantidade de acertos na Categoria 1, foram a s6.5 e s6.6 relativas ao parâmetro g32 e #g32, respectivamente.

Tabela 4.1: Resultado das simulações feitas com os parâmetros individuais.

Simulação (s)	Parâmetro	Minutos	Percentual de Acerto		
			Cat1	Cat2	Geral
1.1	nenhum	356	92%	96%	94%
2.1	c	701	90%	97%	94%
3.1	l	695	95%	97%	96%
4.1	f005	352	90%	100%	95%
4.2	f01	353	92%	95%	93%
4.3	f015	356	92%	96%	94%
4.4	f02	355	92%	97%	95%
4.5	f03	371	92%	97%	95%
4.6	f04	360	90%	97%	94%
4.7	f05	349	88%	96%	92%
5.1	d20	22	96%	96%	96%
5.2	d50	75	96%	96%	96%
5.3	d70	154	96%	97%	97%
6.1	g8	361	93%	96%	95%
6.2	g8#	355	92%	95%	93%
6.3	g16	319	96%	96%	96%
6.4	g16#	337	96%	96%	96%
6.5	g32	316	99%	93%	96%
6.6	g32#	323	99%	93%	96%
7.1	M02	362	93%	96%	95%
7.2	M04	344	93%	96%	95%
7.3	M06	338	92%	95%	93%
7.4	M08	357	96%	100%	98%
7.5	M1	358	96%	99%	97%
8.1	Vw	345	93%	96%	95%
9.1	W	348	88%	93%	91%
10.1	A	348	93%	95%	94%

Na Tabela 4.1, também temos o código e o nome do parâmetro e o valor; p.ex., f01 na s4.2 significa que o parâmetro f foi usado nesta simulação com valor de 0,1; na coluna seguinte temos o tempo de execução da classificação para a amostra de treinamento em minutos, o percentual de acerto para cada categoria individualmente e geral. A percentagem média de objetos com classificação menor do que 0,02 entre as categorias 1 e 2, foi de 23%.

Em cada uma das simulações foi testado apenas um único parâmetro, com apenas um valor, quando o parâmetro necessita de um valor numérico. Nesta etapa não foram feitas simulações com mais de um parâmetro. Como se pode observar na linha de comando (Figura 4.1) de classificação para a simulação 4.1 (parâmetro f com valor 0,05).

Figura 4.1: Linha de comando para execução do Wndchrm no modo de classificação com o parâmetro –

```
wndchrm classify -f0.05 /sto/home/smcerqueira/sims/IR/4/1/treinoIR.fit
/sto/home/smcerqueira/imagensIR/juntas >
/sto/home/smcerqueira/sims/IR/4/1/cl16cIR150_f005.txt
```

f em 0,05 gerando como resultado o arquivo cl16cIR150_f005.txt.

Neste caso, o Wndchrm foi executado com o parâmetro f, tendo seu valor igual a 0,05. Os atributos usados são baseados no arquivo de treinamento treinoIR.fit, classificando as imagens da pasta "juntas", gerando o resultado no arquivo cl16cIR150_f005.txt.

No intuito da elaboração de um gráfico para cada parâmetro, foi atribuído um código de duas partes separadas por um ponto. A primeira parte (à esquerda do ponto) representa o parâmetro a segunda parte (e direita) do ponto representa a variação de valor daquele parâmetro, quando houver. A Figura 4.2 apresenta o índice de acerto para cada simulação para cada uma das duas categorias.

De acordo com as informações da Figura 4.2 e da Tabela 4.1 foram escolhidas as simulações com os melhores percentuais de acertos nas categorias 1 e 2. A simulação s4.4 que foi realizada com o parâmetro f com o valor 0,2 por ter um percentual de acerto para a Categoria 2 de 97% e Categoria 1 com 92%; a simulação s5.2 que foi realizada com o parâmetro d, com uma redução de 50% da imagem, com taxa de acerto de 96% para as duas categorias, um tempo de execução reduzido para 75 minutos e um baixo valor de imagens com diferença das classificações menores que 0,02. A s6.3 que representa a simulação com o parâmetro g com um valor de 16 bits por pixel sem normalização do histograma; por fim a simulação s7.4 que foi realizada com o parâmetro M com valor 0,8 com taxas de acerto de 96% e 100% para as categorias 1 e 2 respectivamente. Para os parâmetros f e M além dos valores já testados foram atribuídos valores próximos para verificar a eficiência no uso daquele parâmetro. Para f usamos os valores 0,2; 0,25 e 0,3. Para M usamos os valores 0,75; 0,8 e 0,85.

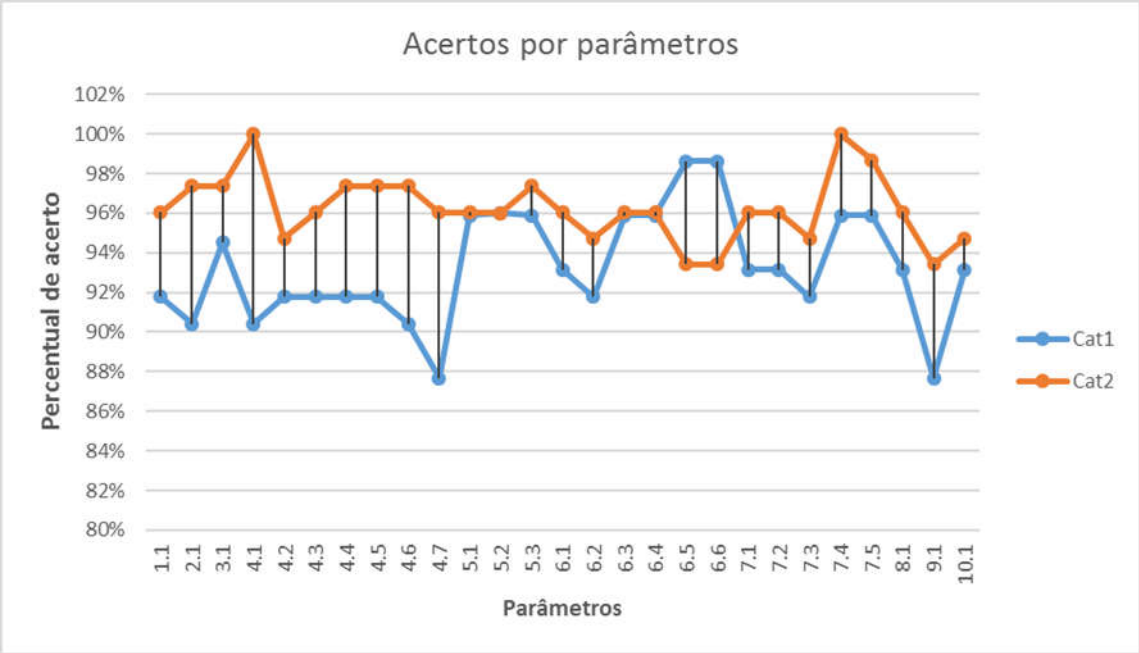

Figura 4.2: Variação dos acertos (em %) para as imagens referentes às categorias 1 e 2 para cada simulação com parâmetros e seus respectivos valores distintos, conforme apresentado na Tabela 4.1.

A partir destas escolhas fizemos uma combinação de parâmetros, de modo a avaliarmos qual seria o impacto usando os parâmetros de forma conjunta. Os arquivos fit (fase de treinamento) foram gerados com os parâmetros d, g e M combinando os valores dos parâmetros d e g com os valores do parâmetro M, conforme apresentado na Tabela 4.2.

Tabela 4.2: Combinação de parâmetros e seus respectivos arquivos fit gerados.

Combinação	Arquivo gerado
d50 g16 M075	treinoIR_d50g16m075.fit
d50 g16 M08	treinoIR_d50g16m08.fit
d50 g16 M085	treinoIR_d50g16m085.fit

Para cada um destes treinamentos foi feita uma classificação usando um dos arquivos fit gerados. Como o parâmetro f não é usado na fase de treinamento, então para cada valor de f um dos arquivos fit com valor de M correspondente foi utilizado, resultando então em nove classificações com a combinação de parâmetros conforme apresentados na Tabela 4.3.

Na Tabela 4.3 temos o código do parâmetro, a combinação dos parâmetros e seus respectivos valores, o tempo da execução da classificação para a amostra de treinamento em minutos, o percentual de acerto para cada categoria individualmente e geral.

Tabela 4.3: Resultado das simulações feitas na amostra de treinamento com a combinação de parâmetros.

Número	Parâmetros	Min	Acerto		
			Cat1	Cat2	Geral
13.1	f0.2 d50 g16 M075	64	97%	95%	96%
13.2	f0.2 d50 g16 M08	64	97%	93%	95%
13.3	f0.2 d50 g16 M085	65	97%	92%	95%
14.1	f0.25 d50 g16 M075	64	96%	92%	94%
14.2	f0.25 d50 g16 M08	67	97%	93%	95%
14.3	f0.25 d50 g16 M085	63	96%	96%	96%
15.1	f0.3 d50 g16 M075	64	95%	95%	95%
15.2	f0.3 d50 g16 M08	63	95%	93%	94%
15.3	f0.3 d50 g16 M085	63	96%	95%	95%

A Figura 4.3 apresenta o gráfico correspondente aos dados de taxa de acerto para as categorias 1 e 2 para as respectivas combinações de parâmetros.

Levando em conta as melhores taxas de acerto para a classificação, tanto da Categoria 1 quanto da Categoria 2, e com um tempo relativamente baixo de execução e uma quantidade grande de diferença de valores de classificação menores que 0,15 a primeira combinação de parâmetros escolhidos para executar com a próxima amostra de imagens, foi f0.25 d50 g16 M085, com 96% de acerto tanto para Categoria 1 quanto para a Categoria 2.

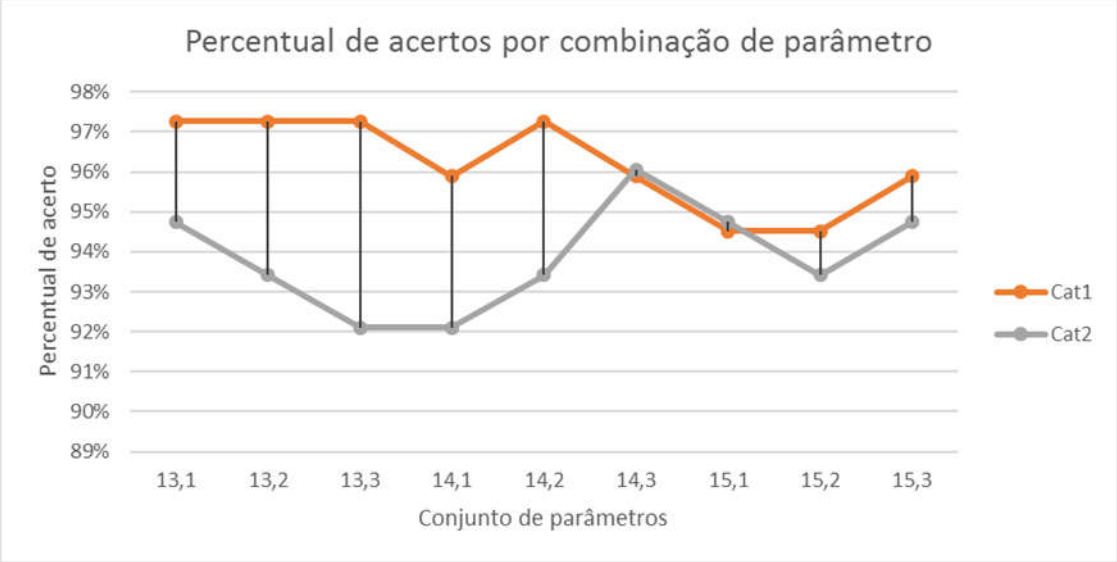

Figura 4.3: Variação dos acertos (em %) para as imagens referentes às categorias 1 e 2 para cada simulação com combinação de parâmetros, conforme apresentado na Tabela 4.3.

4.3 A verificação do método para uma área de 573 graus quadrados

Com o objetivo de verificarmos a validade do método, obtivemos imagens do 2MASS, com tamanho de dois minutos de arco, tendo como base os objetos identificados como galáxias e/ou possíveis candidatas no Catálogo de fontes extensas do 2MASS (2MASS-XSC, Seção 3.3) para uma região de 573 graus quadrados do Hemisfério Sul, cobrindo a faixa de $0 < \mathrm{RA} < 60$ e $-45 < \mathrm{DEC} < -30$.

Essa região foi escolhida por possuir uma maior quantidade de concentração de galáxias peculiares do Catálogo de AM87 e por estar completamente fora do plano de nossa Galáxia, onde a concentração de estrelas e material interestelar (gás e poeira) é muito alta, fazendo com que a identificação de objetos fracos (baixo brilho), como são as galáxias, seja de difícil detecção [Amôres et al. 2012b]. As propriedades da distribuição estelar e da poeira nessa região é descrita na Seção 3.6.

Na medida que identificamos os resultados do método para essa grande região, podemos verificar p. ex., quais parâmetros do Wndchrm são mais sensíveis à identificação de objetos que previamente não sabemos se podem ser enquadrados em uma das duas categorias e em qual exatamente.

Como mencionado no Capítulo 3, existe uma grande quantidade de objetos que não foram identificados e/ou são desconhecidos e mesmo os classificados como galáxias (vc = 1), podem ter sido erroneamente classificados, tendo em vista as características do 2MASS, tais como a resolução e a profundidade.

Considerando-se o tamanho típico (dois minutos de arco) de nossas imagens, em uma região de 573 graus quadrados, teríamos aproximadamente, 820 mil imagens para serem baixadas e verificadas. Apesar de dispormos de recursos computacionais para tal, deve-se notar que as imagens do 2MASS já foram submetidas à um processo de fotometria, que identifica com relativo sucesso, se o objeto é uma fonte extensa ou uma fonte pontual. Para o caso do 2MASS-XSC, inclusive existe uma verificação visual (vc) para saber se o objeto aparenta ser uma galáxia.

Com isso em mente, elaboramos um programa, que verifica no Catálogo de Fontes Extensas do 2MASS (ver Seção 3.3), para todos os objetos identificados com vc = 1, se possuem um outro objeto classificado como desconhecido, não confirmado, e galáxia, em uma região com 2 minutos de arco. Tendo como base, as coordenadas dos dois objetos, calculamos, um valor médio para as coordenadas (RA, DEC), que estão em uma posição central, entre os dois objetos. Esses valores de RA, DEC foram os considerados para buscarmos as imagens no Aladin.

Descartamos nesse procedimento todo objeto que possuía, mais de um vizinho em uma região com 2 minutos de arco, e/ou tinha como vizinho um objeto classificado como vc = 2, ou seja, estrela ou artefato, e também o objeto que já fora identificado como vizinho de um objeto anteriormente incluído na amostra.

A justificativa para considerarmos também objetos não confirmados (vc = -1) e desconhecidos (vc = -2) em nossa amostra, deve-se ao fato de que, com nosso método, podemos também ter uma análise que leve em consideração se o objeto, antes desconhecido ou não verificado, pode ser considerado como candidato até mesmo à

galáxia. No total, temos 807 imagens, sendo que, ao menos 199 "pares de objetos" (objetos que, a priori, apenas estão próximos em uma imagem de duas dimensões), identificados como galáxias no 2MASS e outras 608 imagens (com pares de candidatas), com ao menos um objeto identificado como galáxia e os outros objetos identificados como vc = -1, -2.

Para obter essas imagens, utilizamos o Aladin (Capítulo 3), por meio de um script para baixar as imagens a partir de suas coordenadas equatoriais, com tamanho de 2 minutos de arco. Executando o script em um computador Core I5, 4 MB de memória RAM, em uma rede com velocidade de banda de 10MB todo o processo de obtenção das imagens (download) demorou aproximadamente 15 minutos.

4.3.1 Análise das imagens

É importante efetuarmos uma boa estimativa sobre se os 807 objetos, nas imagens dos 573 graus quadrados, pertencem à uma das categorias 1 ou 2 do Catálogo de AM87. Dessa forma, fizemos uma classificação visual das imagens baixadas, nas seguintes classes: 1 e 2 para as imagens que julgamos ser das categorias 1 e 2, respectivamente do Catálogo de AM87. As imagens cujos os objetos não podem ser identificados como pertencentes a nenhuma das duas categorias; imagens sem a clara identificação de interação entre os pares de objetos; imagens que apresentavam galáxias de brilho fraco, muitas estrelas, mais de duas galáxias (provavelmente galáxias observadas pelo 2MASS, pois aparecem na imagem, mas sem fotometria) ou quando uma das galáxias se apresentavam fora ou parcialmente fora da área da imagem, foram classificadas com classe -1.

Na Figura 4.4, podemos ver imagens consideradas das categorias 1 e 2 usadas na classificação visual.

Figura 4.4: Duas imagens consideradas como Categoria 1 (primeira e segunda colunas) e duas consideradas Categoria 2 (terceira e quarta colunas). Fonte: 2MASS.

Na inspeção visual, alguns pares que poderiam ser classificados nas categorias 1 ou 2, tinham alguma das galáxias na borda das imagens, conforme apresentado na Figura 4.5, para algumas galáxias da Categoria 1. Isso deve-se ao fato, de que no programa elaborado, ele considera para verificar se dois objetos estão em uma mesma imagem de dois minutos de arco, o centro dos objetos e não inclui nessa verificação, o seu tamanho. Em nossa inspeção visual, evitamos classificar essas imagens como das categorias 1 ou 2 pois em nossa amostra de treinamento todas as galáxias estavam dentro do tamanho da imagem de dois minutos de arco.

Figura 4.5. Imagens consideradas Categoria 1 em nossa inspeção visual. Fonte: 2MASS.

4.3.2 Análise de propriedades gerais da área de 573 graus quadrados

Além da análise individual de cada imagem é importante analisar as propriedades dos campos de cada imagem, tais como a contagem de estrelas, efeito da extinção interestelar como abordado no Capítulo 3, sendo também importante analisar galáxias peculiares previamente identificadas para essa área e quais estão relacionadas com as 807 imagens, assim como verificar se algumas dessas imagens estão presentes na amostra de treinamento.

De maneira similar ao que foi feito na Seção 3.6, obtivemos a extinção interestelar, usando o GALExtin[18]. O mapa para a extinção interestelar, é apresentado na Figura 4.6, tendo como base os mapas de Schlegel et al. (1998, SFD), nota-se que os valores se situam na faixa de $0 < A_V < 0{,}1$ mag, o que também pode ser visto pelo histograma da Figura 4.7.

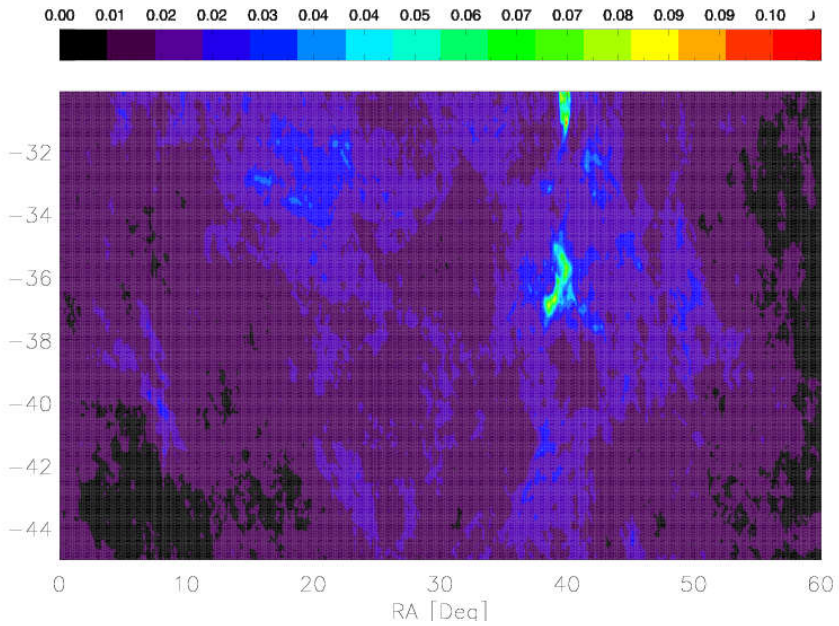

Figura 4.6: Distribuição da extinção interestelar (A_V) obtida com o mapa de SFD.

[18] http://www.galextin.org

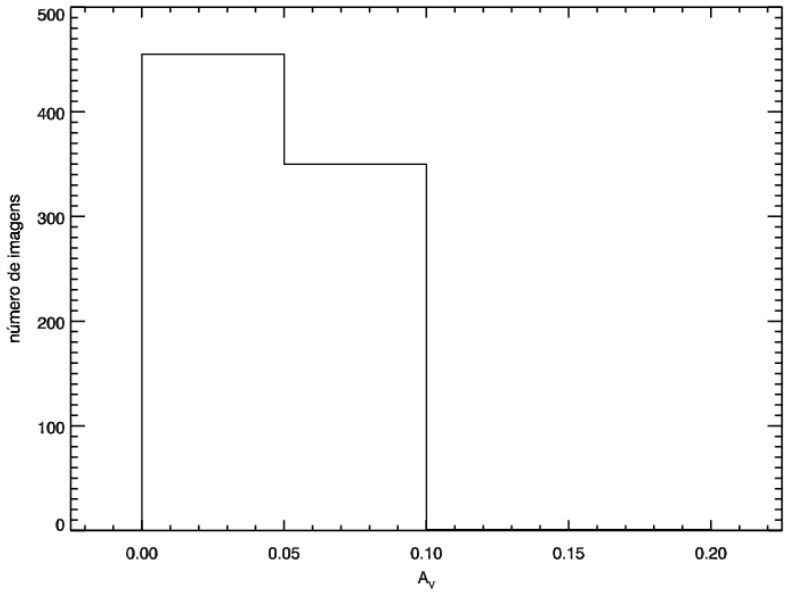

Figura 4.7: Histograma com a extinção na direção dos campos das imagens para 573 graus quadrados com o mapa apresentado na Figura 4.6.

A próxima etapa consistiu de contarmos quantas estrelas em função da magnitude existiam em cada imagem, sendo feita a correção da extinção, (Equação 3.1). Encontramos, um total de 126 imagens que foram descartadas seguindo o critério de seleção, apresentado na Seção 3.5.

4.3.3 Verificação dos parâmetros optimais

O conjunto de parâmetro da simulação s14.3, da Tabela 4.3, cujos valores são, f0.25; d50; g16 e M085, foi usado, a princípio, para classificar as imagens dos 573 graus quadrados, por apresentar uma taxa de acerto de 96% na fase de treinamento. Entretanto, utilizando esse conjunto de parâmetros, na classificação das imagens para os 573 graus quadrados, não obtivemos bons resultados, com probabilidades similares para as categorias 1 e 2.

De forma a termos uma amostra homogênea, utilizamos como amostra para a classificação apenas os pares classificados pelo 2MASS-XSC como vc = 1, ou seja, galáxias, sendo que de um total de 199 pares, 66 foram descartados em uma análise anterior (Seção 4.3), pelos motivos descrito anteriormente. Identificamos todos os pares que tinham ao menos uma ocorrência de nossa classificação igual a -1, que são imagens pertencentes à amostra de treinamento, ou com problema com o número de estrelas no campo, nos restando 108 imagens, as quais, foram usadas para determinamos os melhores parâmetros do Wndchrm para uma amostra de candidatas a galáxias peculiares das categorias 1 ou 2.

Dessa forma, com o intuito de melhorar a identificação para uma amostra compatível com o que iremos encontrar para o restante do Hemisfério Sul; realizamos a variação dos parâmetros, tendo como ponto de partida, os melhores resultados apresentados,

verificando o impacto dos parâmetros nas diferenças entre as probabilidades atribuídas a cada categoria. Os valores de cada parâmetro, assim como, das percentagens de acerto e o valor da mediana da probabilidade para cada categoria é apresentado na Tabela 4.4.

Nessa análise dos parâmetros, também verificamos os resultados obtidos com a utilização dos parâmetros na primeira fase da escolha de parâmetros, conforme descrito na Seção 4.1. Verificou-se que o parâmetro w, da simulação s9.1 na Tabela 4.1 apresenta bons valores para a quantidade de imagens, cuja diferença entre os pesos de classificação foi menor que 0,15, igual a 33 de um total de 138, e com taxa de acerto de 91%. Esse parâmetro foi escolhido para compor a combinação de parâmetros para as próximas classificação.

Além do parâmetro w, outras alterações nos valores dos parâmetros f, d e m foram feitas, com a intenção de verificar o resultado das quantidades de classificação com diferenças maiores do que 0,15; 0,05 e 0,02.

Outra simulação usada para uma segunda classificação dos 573 graus foi a simulação s13.1 na Tabela 4.3, a qual também não apresentou bons resultados. Por conta desse problema, consultamos a Tabela 4.1 para verificar o comportamento dos parâmetros quando usados de forma isolada. Verificamos que a simulação 9.1, com o parâmetro w apresentou bons resultados com relação à diferença entre os graus de similaridade das duas categorias.

Decidimos incluir o parâmetro w na combinação de parâmetros s13.1, para classificar as 807 imagens, criando assim a simulação com parâmetros f0.2, d50, g16, m0.75, w. Os resultados desta classificação foram melhores, no que diz respeito às diferenças entre as probabilidades atribuídas pelo Wndchrm. Obtivemos 244 imagens cuja a diferença foi maior que 0,15. Se consideramos o valor 0,02, na simulação sem o parâmetro w, obtemos 442 imagens. Percebemos assim, que o parâmetro w deve ser usado para as classificações pois aumenta a diferença entre as probabilidades das categorias.

Outras combinações foram feitas alterando-se os valores dos parâmetros f e d. Adotando-se um valor de 0,1 para o parâmetro f, verificou-se que não houve diferença nos resultados em combinação com os valores do parâmetro M 75 e 85, apresentando uma quantidade razoável de resultados com diferenças entre os graus de similaridade em 306 imagens.

Nesta fase, também fizemos um teste com o parâmetro d, colocando para esse parâmetro o valor de 80, o que significa um downsample de 20% na imagem. O acréscimo deste parâmetro nas simulações, representadas por s16, s17 e s18, na Tabela 4.4, não resultou em alterações nos valores, no entanto, fez com que o tempo de classificação elevasse a de uma média de 6 para 11 horas para classificar 807 imagens, sendo assim uma alteração de pouca valia para nosso trabalho.

No gráfico da Figura 4.8, bem como na Tabela 4.4, vemos que as simulações de s9 à s14 apresentam valores de taxa de acertos (em relação a classificação visual) mais próximos para as categorias 1 e 2. Estas simulações têm em comum o uso do parâmetro M, com valor de 0,6. No entanto, verificamos que a simulação s12 apresenta valores da mediana das probabilidades de acerto para a categoria 1 e 2 mais altos que as demais simulações; respectivamente 0,569 e 0,618. Por esse motivo

escolhemos os parâmetros dessa simulação para classificar as imagens do Hemisfério Sul.

Tendo em vista os 78 pares de galáxias classificados pelo Wndchrm nas categorias 1 e 2, considerando-se a simulação s12 da Tabela 4.4, e que coincidiram com a nossa classificação, realizamos algumas análises, tendo como base as propriedades desses objetos tais como magnitudes e tamanho. A Figura 4.9, apresenta um gráfico com a relação entre o número de galáxias e seus tamanhos. O tamanho é fornecido pelo raio de Kron em segundos de arco. Percebe-se no gráfico que não existe nenhuma galáxia da Categoria 2 e muito poucas da Categoria 1 com tamanho maior que 30"; por outro lado existe uma grande quantidade de galáxias, das duas categorias com tamanhos entre 5" e 15", isto é, a maior parte das imagens desta amostra apresenta galáxia de tamanho pequeno.

A Figura 4.10 apresenta a distribuição de magnitudes das galáxias, separando-se os pares de acordo com os mais brilhantes e com brilho mais fraco, nota-se que para as galáxias, com brilho mais fraco do par, temos que o pico no histograma para aproximadamente K_S = 14,0 mag, ou seja, podemos considerar esse um limite de completeza para nosso método.

4.4 Cruzamento com as galáxias do Catálogo de AM87

É importante verificarmos a quantidade de galáxias peculiares nessa região, o que foi feito usando os dados apresentados na Seção 3.1. Esse procedimento faz-se necessário para, não somente comparar os resultados obtidos de nosso método de identificação, mas para descartar os objetos que serão identificados, o que já estão em nossa amostra de treinamento, levando à identificação dos mesmos.

Foram encontradas 576 galáxias peculiares do catálogo de AM87, para essa região, em um total de 867 ocorrências, ou seja, algumas galáxias peculiares classificadas por AM87 em mais de uma categoria. Cabe ressaltar que, esses objetos não foram usados na composição de nossa amostra de treinamento. No total, temos 357 galáxias peculiares que são identificadas, apenas em uma categoria.

Seguimos o procedimento de identificar galáxias peculiares que estivessem contidas nas 807 imagens que compõem os 573 graus quadrados. Para cada imagem, tendo como base seu centro de coordenadas, verificamos as galáxias peculiares, analisando-se suas coordenadas contidas dentro de 2 minutos de arco da imagem. Em 20 imagens foram encontradas galáxias peculiares usadas em nossa amostra de treinamento, sendo 5 e 15 das categorias 1 e 2, respetivamente.

Em 78 imagens (com 132 ocorrências), encontramos galáxias peculiares previamente identificadas por AM87, sendo 7 e 19 nas categorias 1 e 2, respectivamente. A diferença entre esse número e o total de galáxias peculiares identificadas por AM87, para essa região, pode ser justificada pelo método utilizado no presente trabalho, que parte da identificação, na fotometria do 2MASS, com base em fontes extensas para identificar candidatas a peculiares, apenas para imagens com dois minutos de arco e com apenas duas candidatas a peculiares, caso existam mais de duas, para uma

imagem, a mesma é descartada de nosso objeto. Outro ponto, reside no fato que, procuramos apenas por objetos da categoria 1 e 2 de AM87, sem considerar o limite de completeza do 2MASS, que é de 14,3 mag, para o filtro Ks.

No Capítulo 5, será apresentada uma comparação para as galáxias do Catálogo de AM87, para todo o Hemisfério Sul e as classificações fornecidas por nosso método.

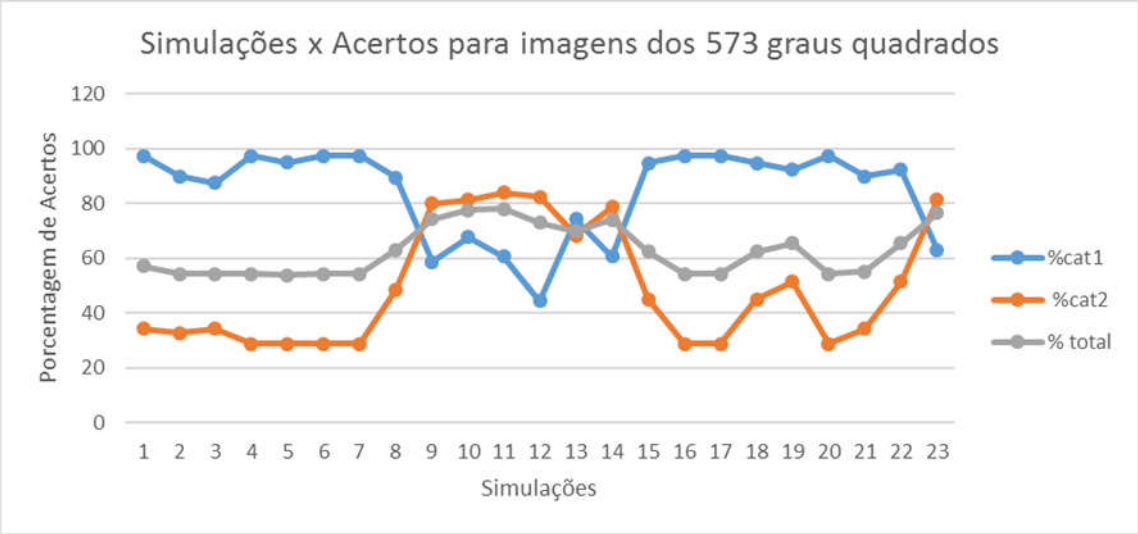

Figura 4.8: Gráfico com as comparações entre as simulações e a classificação visual para as categorias e 1 (cat1) 2 (cat2).

Tabela 4.4: Simulações feitas e comparadas com a classificação visual.

Simulação (s)	Parâmetros	%cat1	%cat2	% total	Mediana prob. 1	Mediana prob. 2
1	f025d50g16M085w	97,4	34,3	57,1	0,500	0,501
2	f025d50g16M075w	90,0	32,8	54,2	0,500	0,500
3	f025d50g16M065w	87,5	34,3	54,2	0,500	0,501
4	f015d50g16M075w	97,4	28,8	54,3	0,500	0,501
5	f015d50g16M065w	95,0	28,8	53,8	0,500	0,501
6	f01d50g16M085w	97,4	28,8	54,3	0,500	0,501
7	f01d50g16M075w	97,4	28,8	54,3	0,500	0,501
8	f01d50g16M065w	89,5	48,5	63,2	0,561	0,539
9	f015d50M06w	58,6	80,0	74,3	0,506	0,546
10	f005d50g16M06w	67,7	81,3	77,5	0,511	0,614
11	f004d50M06w	60,7	84,0	78,0	0,506	0,610
12	f003d50M06w	44,4	82,5	72,9	0,569	0,618
13	f02d50M06w	74,2	68,4	70,0	0,504	0,535
14	f01d50M06w	60,7	78,8	74,1	0,511	0,552
15	f01d50g16M090w	94,7	45,1	62,4	0,575	0,501
16	f01d80g16M075w	97,4	28,8	54,3	0,500	0,501
17	f01d80g16M085w	97,4	28,8	54,3	0,500	0,501
18	f01d80g16M090w	94,7	45,1	62,4	0,575	0,501
19	f02d50g16M075	92,3	51,4	65,5	0,500	0,500
20	f02d50g16M075w	97,4	28,8	54,3	0,500	0,501
21	f02d50g16M075wr0	90,0	34,3	55,1	0,500	0,501
22	f02d80g16M075w	92,3	51,4	65,5	0,501	0,501
23	f005d50M06	63,0	81,3	76,6	0,505	0,582

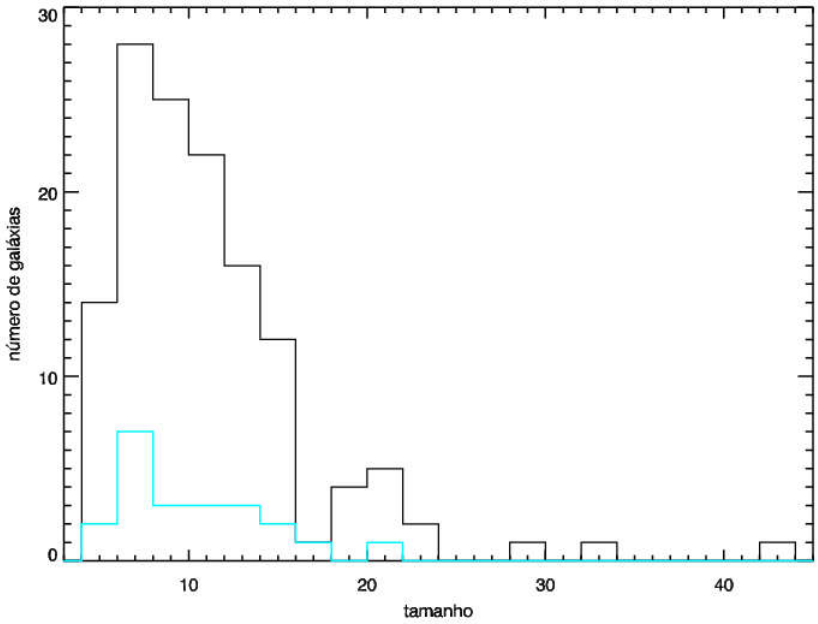

Figura 4.9. Distribuição do tamanho dos objetos, raio de Kron em segundos de arco. Categoria 1 (linha azul), Categoria 2 (linha preta).

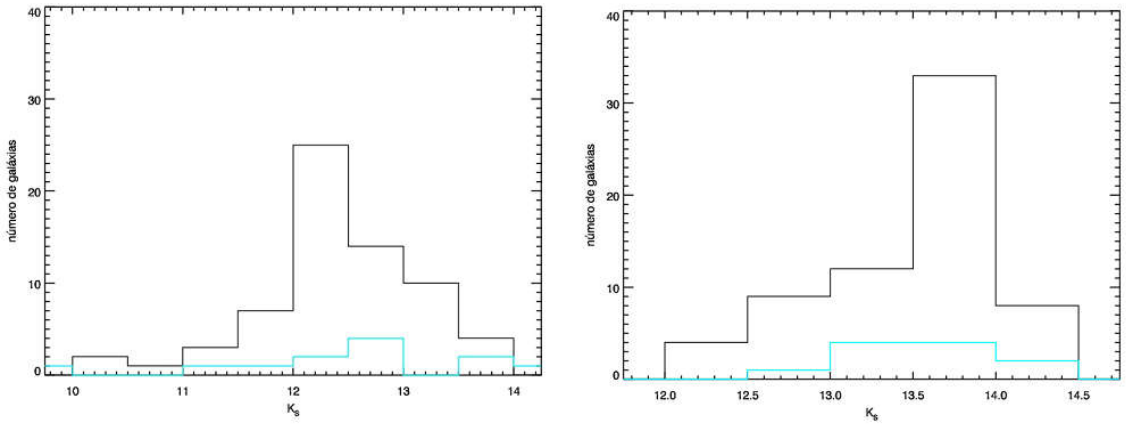

Figura 4.10: Distribuição de magnitudes das galáxias entre os pares: mais brilhantes (esquerda), com brilho mais fraco (direita). Categoria 1 (linha azul), Categoria 2 (linha preta).

Capítulo 5

Galáxias peculiares das categorias 1 e 2 no Hemisfério Sul

Este Capítulo apresentará a elaboração de nossa amostra de candidatas a galáxias peculiares das categorias 1 e 2 do Catálogo de AM87, o procedimento utilizado para rodar o programa para uma amostra com mais de 10 mil pares de candidatas, assim como os resultados encontrados e o cruzamento com dados existentes na literatura.

5.1 A amostra

Para analisar as possíveis candidatas a galáxias peculiares das categorias 1 e 2, fizemos uma seleção nos objetos do catálogo de fontes extensas do 2MASS com fotometria simultânea em J, H e Ks, de forma a obtermos objetos para o Hemisfério Sul, exceto para a região do plano Galáctico ($|b| < 10°$) e com vc igual a 2, sendo selecionados 686.089 objetos.

Tendo como base esses objetos, fizemos uma nova seleção, similar ao que foi feito para as candidatas para os 573 graus quadrados, de forma a termos, nas imagens, apenas duas candidatas a galáxias peculiares das categorias 1 e 2, na quais, ao menos um objeto fosse considerado (por inspeção visual prévia da equipe do 2MASS) como galáxia, ou seja, vc igual a 1. Dessa forma, após essa nova seleção, o número de imagens a serem analisadas foram de 21.843, cobrindo uma área de aproximadamente 17.772 graus quadrados, já levando em conta a não cobertura de regiões, no plano Galáctico, e da região de 573 graus quadrados (Seção 4.3).

A Figura 5.1 apresenta a distribuição do centro das imagens selecionadas acima. Nota-se áreas, com ausência de imagens, uma estendendo-se por toda uma grande região, o que é devido a termos eliminado imagens na direção do plano Galáctico, a outra característica e que pode ser notada na parte central da imagem, é a região de 573 graus quadrados, analisada no Capítulo 4.

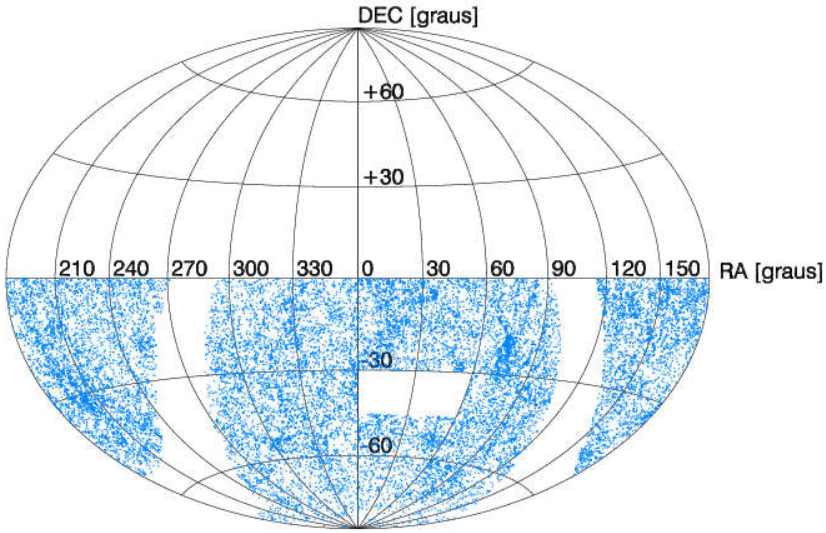

Figura 5.1: Distribuição em coordenadas equatoriais das imagens com candidatas a conterem galáxias peculiares para o Hemisfério Sul. As regiões com ausência de imagens são explicadas no texto.

De maneira similar ao que foi feito no Capítulo 3, obtivemos, por meio do GALExtin, a extinção interestelar fornecida pelos mapas de Schlegel, na direção das 21.843 imagens. A Figura 5.2 apresenta o histograma da distribuição do A_V. Nota-se, que a maior parte dos objetos estão localizados nas direções com baixa extinção. Levando-se em conta o critério estabelecido no Capítulo 3, identificamos que 18.379 imagens possuem extinção, $A_V \leq 0{,}4$ mag, ou seja, aproximadamente 85% das imagens. Das imagens descartadas, 823 possuem extinção, $A_V > 1{,}0$ mag, o que pode ser atribuído à nuvens de gás e poeira que estão fora do plano Galáctico, e são sítio de formação estelar [Amôres & Lépine 2005].

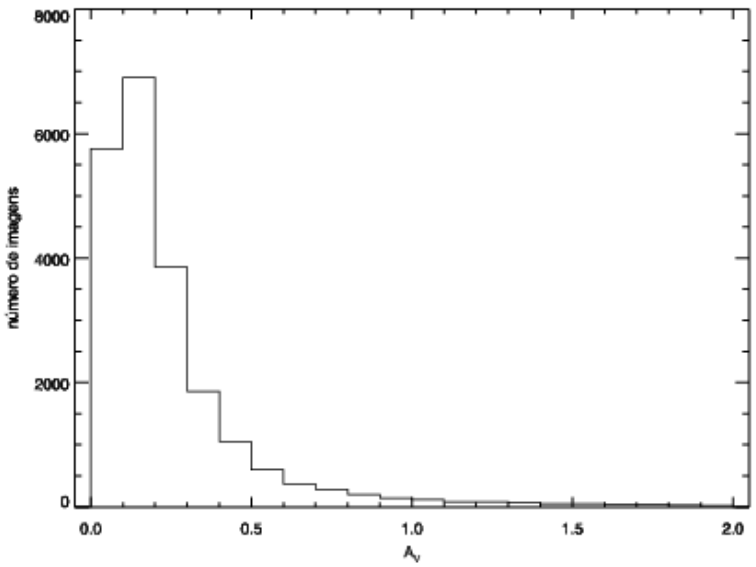

Figura 5.2: Distribuição da extinção interestelar (A_V) na direção das 21.843 imagens.

Tendo-se estimada a extinção interestelar na direção das imagens, a próxima etapa consiste em determinar o número de estrelas por imagens, de forma a aplicar o critério de seleção, definido na Seção 3.6. Com este propósito, foram obtidas, a partir do Catálogo de fontes pontuais do 2MASS, as estrelas que estavam no campo das imagens.

No total, foram obtidos dados, tais como coordenadas e magnitudes nos filtros J, H e Ks, para 165.036 estrelas. De forma a termos o número de estrelas, em função da magnitude para cada direção, é necessário fazer a correção da magnitude das estrelas devido ao efeito da extinção interestelar, conforme descrito pela Equação 3.1 (Seção 3.6). Após essa etapa, foi realizada a contagem de estrelas por faixa de magnitude, sendo descartadas as imagens que possuíam mais de um determinado número de estrelas por faixa de magnitude como apresentado no Capítulo 3.

A Tabela 5.1 apresenta o número de imagens que satisfazem as condições de descarte, para cada uma das faixas de magnitudes, e número de estrelas. Levando-se em conta as imagens descartadas devido a extinção e, que satisfaziam ao menos uma das condições de descarte fornecidas na Tabela 5.1, restaram 10.644 imagens para serem verificadas pelo Wndchrm, se poderiam ser classificadas na categoria 1 ou 2.

Tabela 5.1. Faixas de magnitudes no filtro K_S e número de estrelas que satisfazem as condições de descarte.

Faixa de magnitude – número de estrelas	Número de imagens
$K_S < 12$ (mais de uma estrela)	1.070
12 $K_S < 16$ (mais de cinco estrelas)	10.752
16 K_S 17 (mais de uma estrela)	988
$K_S > 17$ (qualquer estrela)	370

A Figura 5.3 apresenta a distribuição dos objetos selecionados. Em uma comparação com a Figura 5.1, vemos que a diminuição do número de candidatas a galáxias diminui, principalmente nas regiões próximas ao plano Galáctico, o que pode ser devido, além do ainda elevado número de estrelas, também aos efeitos da extinção interestelar. Uma discussão mais detalhada da distribuição dessas candidatas assim como de suas propriedades, tais como magnitudes, a verificação visual (vc) realizada pela equipe do 2MASS será efetuada na Seção 5.4.

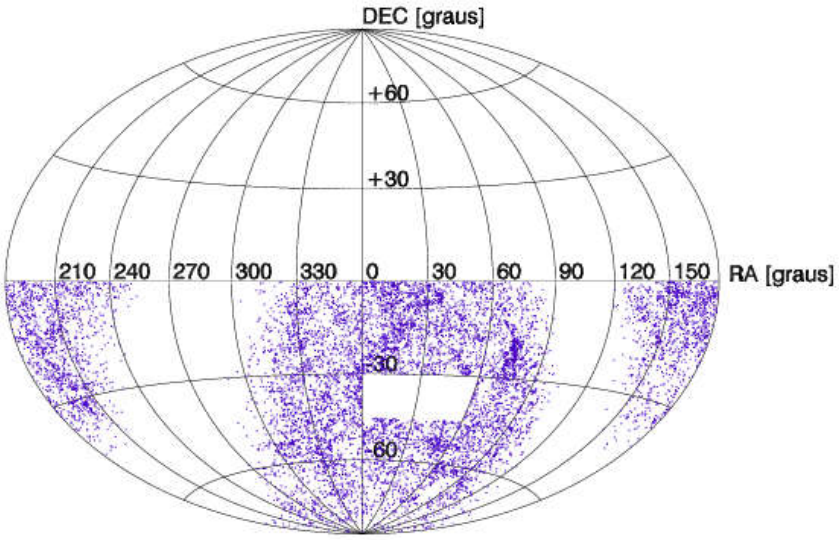

Figura 5.3: Distribuição das candidatas a galáxias peculiares da categoria 1 e 2 após a seleção levando em conta o número de estrelas e a extinção interestelar.

5.2 Rodando o Wndchrm para o Hemisfério Sul

O processo de obtenção das imagens teve início e durou aproximadamente 2 horas, em um computador com processador Core I5, com 4GB de RAM, conectado à internet por um rede de 10Mbps.

As imagens obtidas, ainda necessitaram de passar por um tratamento, no qual o formato do arquivo é alterado para tiff, além de serem redimensionadas de 511 x 525 pixels para 510 x 511 pixels. Para isso foi usado o XnView, essa alteração retira a marca d'agua deixada pelo Aladin nas imagens baixadas por meio de scripts. Esse procedimento de obtenção das imagens por meio do Aladin, e tratamento por meio do XnView é descrito na Seção 3.4.

Para classificar as 10.644 imagens do Hemisfério Sul, com o Wndchrm, dividimos o número total de imagens em 10 grupos de 1000 imagens e mais um grupo com 644 imagens, criando um diretório para cada grupo, nomeando-os como "parte1" até "parte11", devido a limitação do software de classificar apenas 1000 imagens por execução. Todos esses diretórios foram colocados em um diretório pai chamado "22k".

O arquivo fit usado para classificação foi o treinoIR_d50m06.fit, arquivo gerado na fase de treinamento, usando os seguintes parâmetros: d com valor 50 e M com valor 0,6. O comando usado para executar o Wndchrm no modo de classificação é mostrado na Figura 5.4.

```
wndchrm classify -f0.03 -d50 -M0.6 -w
/sto/home/smcerqueira/sims/IR/fits/treinoIR_d50g06p1.fit
/sto/home/smcerqueira/imagens/22k/parte1 >
/sto/home/smcerqueira/sims/IR/22k/cl8cIR_f003d50m06wp1.txt
```

Figura 5.4: Linha de comando para a classificação das imagens contidas no diretório "parte1", a partir do treinamento treinoIR_d50m06.fit.

Na linha de comando, mostrada na Figura 5.4, são usados os parâmetros f com valor 0,03, d com valor 50, M com valor 0,6 e o parâmetro w, para a execução do Wndchrm. Como dividimos o conjunto das imagens, em grupos de 1000, tendo para cada grupo um diretório, foram necessárias executar, o Wndchrm, 11 vezes. Para agilizar o processo, foram submetidas 5 execuções simultâneas de cada vez; e para isso foram feitas mais 5 cópias do arquivo treinoIR_d50m06.fit, uma para cada execução simultânea. Cada cópia recebeu uma terminação no nome do arquivo variando de p1 a p5. Na linha de comando na Figura 5.4, a cópia do arquivo fit usado foi treinoIR_d50m06p1.fit.

A classificação da linha de comando da Figura 5.4, analisou as 1.000 imagens do diretório "parte1" e os resultados da classificação foram salvos no arquivo cl8cIR_f003d50m06wp1.txt.

O procedimento de classificação de 1000 imagens, sendo executada de forma não dedicada, no o cluster SGI Altix ICE 8400 do sistema de computação de alto desempenho do INCT-A, localizado no IAG/USP (mais detalhes na Seção 4.1), durou, em média, 10h20.

5.3 Análise das galáxias peculiares das categorias 1 e 2 de AM87 nas imagens selecionadas

No total temos 5.319 galáxias peculiares do Catálogo de AM87 localizadas fora da área de 573 graus quadrados (Seção 4.3), sendo que, destas, 948 galáxias na direção das 21.843 imagens selecionadas por nosso método, em uma primeira etapa. Fazendo o descarte, levando em conta o número de estrelas e a extinção nos campos, de 10.644 restam 505 imagens, com 648 ocorrências de galáxias em outras categorias, em que temos, ao menos, uma galáxia peculiar. A visualização da quantidade de galáxias peculiares nas diferentes categorias é apresentada pela Figura 5.6. As 25 categorias do Catálogo de AM87 são apresentadas na Seção 3.1. Nota-se que as categorias com maior distribuição de objetos são a 2, 8 e 23, que representam as Categorias: "Galáxias com interação dupla", "Galáxias com companheiras aparente" e "Pares próximos (sem interação visível)", respectivamente (ver Tabela 3.1).

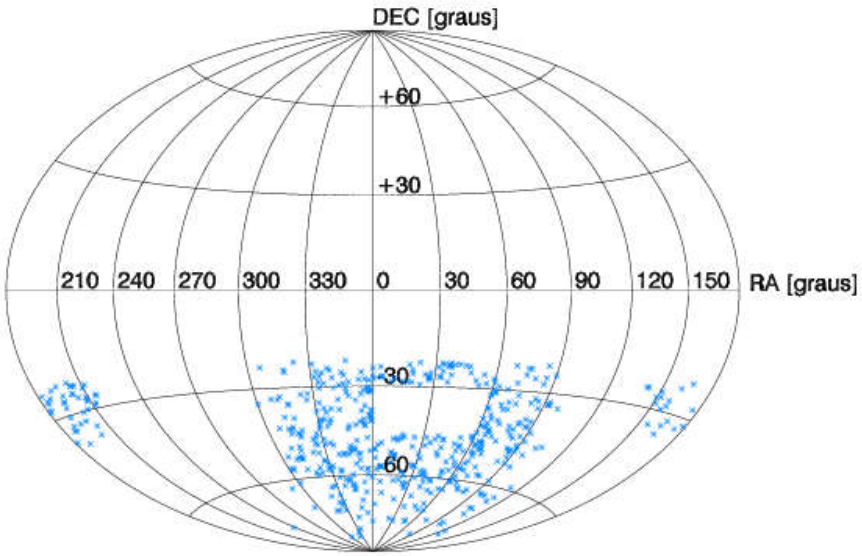

Figura 5.5: Distribuição das 399 imagens nas 10.644 imagens selecionadas para a busca de galáxias peculiares das categorias 1 e 2 de AM87, que continham objetos do Catálogo de AM87.

Em procedimento similar ao realizado no Capítulo 4, descartamos dessas 505 imagens todas as que foram usadas como amostra de treinamento, e/ou descartadas no procedimento de montagem da amostra de treinamento para compor os objetos representativos das categorias 1 e 2, dessa forma descartamos 106 imagens, restando 399 imagens a serem analisadas com galáxias peculiares. A distribuição desses objetos é apresentada na Figura 5.5.

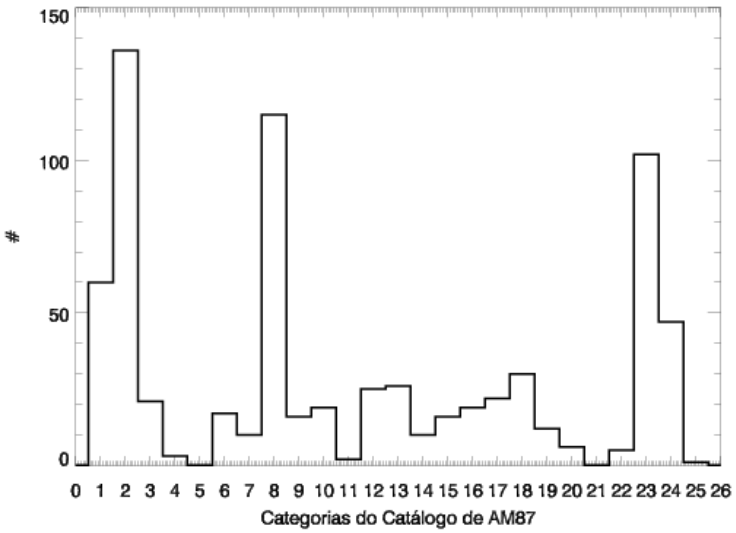

Figura 5.6: Distribuição das categorias das galáxias peculiares apresentadas na Figura 5.5.

Temos um total de 113 galáxias peculiares das categorias 1 e 2 para a nossa amostra restante de imagens, sendo que 38 e 75 das categorias 1 e 2, respectivamente. Tivemos um acerto de aproximadamente 54% e 88%, para os objetos das categorias 1

e 2. A Figura 5.7 e 5.8 apresentam exemplos de galáxias peculiares de ambas as categorias classificadas pelo Wndchrm, de acordo com a classificação de AM87.

Figura 5.7: Exemplo de galáxias classificadas pelo Wndchrm na Categoria 1.

Figura 5.8: Exemplo de galáxias classificadas pelo Wndchrm na Categoria 2.

Com o propósito de verificarmos os objetos da Categoria 1 que tiveram a pior probabilidade de classificação em comparação com o Catálogo de AM87, apresentamos na Figura 5.9 alguns desses objetos. Aproximadamente 28% dos pares classificados na Categoria 1 tiveram probabilidade de classificação entre 48-50%.

Na Figura 5.9, as imagens, em sua maioria, apresentam muitas estrelas no campo (algumas possivelmente não identificadas pela fotometria do 2MASS). Como nas imagens selecionadas para a amostra de treinamento, evitou-se esses objetos (Seção 3.5), a classificação do Wndchrm apresentou uma probabilidade próxima de 50%.

Na imagem com "número 37336" (primeira linha, segunda coluna), vemos a galáxia companheira fora do campo visual da imagem o mesmo para a imagem "417844" (segunda linha, terceira coluna. Na imagem com "número 393749", (segunda linha, segunda coluna), podemos ver uma estrela com brilho forte, aparentemente próxima a galáxia principal, o que pode levar parecer um par de galáxias da Categoria 2. Na imagem com "número 327810", (segunda linha, primeira coluna), a galáxia principal tem um tamanho angular grande, ocupando quase todo campo visual da imagem.

Figura 5.9: Exemplos de objetos da Categoria 1 de AM87, que tiveram probabilidade de classificação entre 48-50% na Categoria 1 usando o Wndchrm.

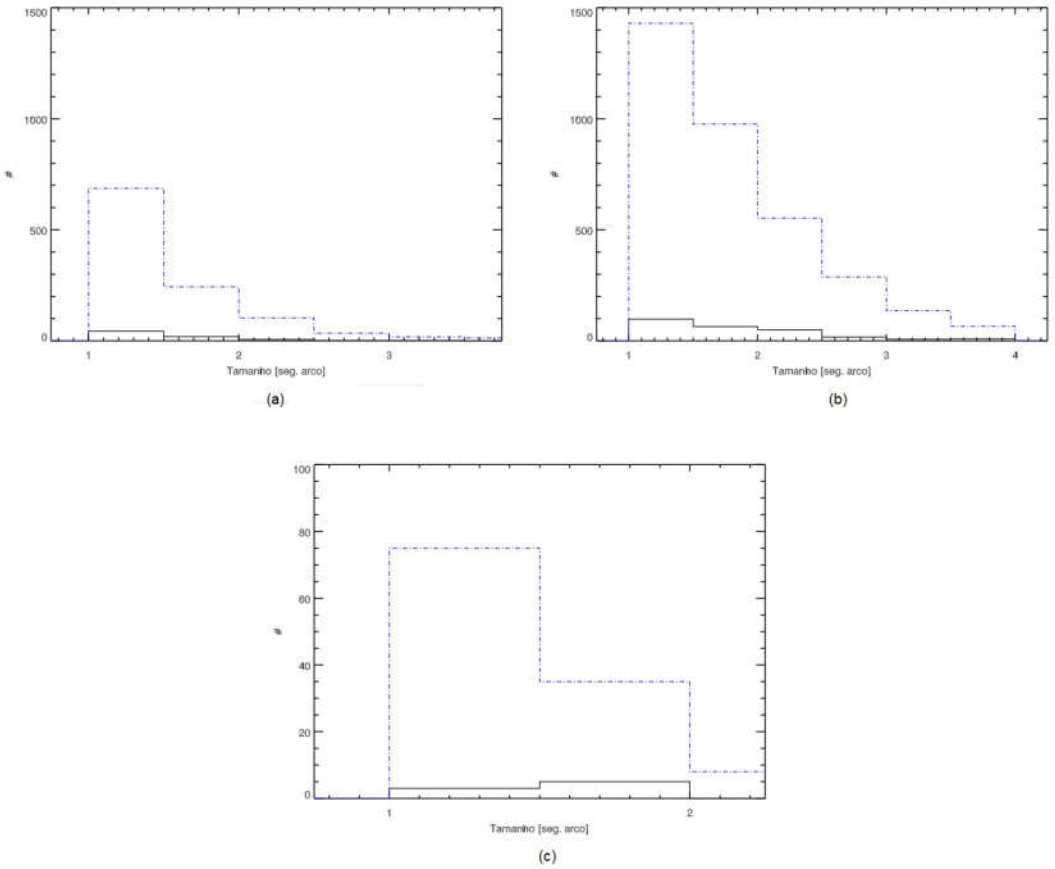

Figura 5.10: Relação entre o raio de Kron para pares classificados com probabilidade entre 50-51% para ambas as categorias. Nas três combinações de vc = 1 (a), -1 (b), -2 (c) apresentadas na Tabela 5.2, na qual as linhas contínuas e pontilhadas representam as categorias 1 e 2, respectivamente.

5.4 Análise das candidatas a galáxias peculiares das categorias 1 e 2

Com o propósito de verificarmos as classificações para os objetos restantes de nossa amostra, em uma etapa inicial dividimos as candidatas a galáxias interagentes de acordo com a classificação dos pares segundo a classificação visual (vc) do 2MASS, descartados objetos com raio de Kron > 30, que, conforme visto na Seção 4.3, existem apenas dois objetos com esses raios em nossa amostra. O número total de candidatas seguindo a classificação do 2MASS, é apresentada na Tabela 5.2. Elaboramos um histograma para as distribuições das categorias e nota-se que parte significativa dos objetos tanto para a Categoria 1 quanto para a Categoria 2, possui probabilidade entre 50% e 51% de classificação, os quais são apresentados entre parênteses na Tabela 5.2.

Tabela 5.2: Número de pares de candidatas a galáxias interagentes das categorias 1 e 2, tendo como base a flag vc, para cada galáxia (vc_1 e vc_2) par de galáxia, sendo valores de vc: 1 (galáxia), -1 (objeto não verificado), -2 (objeto desconhecido). Os números entre parênteses representam o número total de objetos. A coluna raio de Kron apresenta o número de objetos descartados por ter o raio de Kron > 30.

vc_1	vc_2	número de pares de candidatas	Categoria 1	Categoria 2	raio de Kron > 30 (Categ. 1)	raio de Kron > 30 (Categ.2)
1	1	2.433	278 (364)	932 (2069)	16	35
1	-1	7.522	828 (1095)	2812 (6427)	14	55
1	-2	290	37 (47)	110 (243)	2	4

Para verificarmos o alto número de objetos classificados com probabilidades entre 50% e 51% para ambas as categorias, analisamos o tamanho dos pares, usando o raio de Kron, fornecido em segundos de arco. No total, temos 5.122 pares de objetos com as probabilidades descritas acima para as três categorias de vc. Para cada par, verificamos qual o objeto com o maior raio de Kron.

A Figura 5.10 apresenta três histogramas com a razão entre o raio de Kron do maior objeto em relação ao menor. Os histogramas podem ser interpretados da seguinte forma: para candidatas a galáxias na Categoria 2, esperamos que a razão entre os tamanhos seja, aproximadamente, menor do que 1,5, e maior do que 1,5, para objetos da Categoria 1. Nota-se um grande número de objetos com tamanho superior a 1,5 para objetos da Categoria 2, por outro lado, vemos uma grande concentração de objetos com razão entre os tamanhos similares, para objetos identificados como Categoria 1. Possivelmente, são imagens compostas por galáxias que têm características de interagentes e/ou objetos extensos de nossa Galáxia.

Cabe ressaltar que, utilizamos imagens em falsa cor, as quais consideram três filtros e para a análise feita com o raio de Kron, utilizamos apenas o filtro K_S, galáxias podem ter características de tamanho diferentes, considerando-se outros filtros. Um aspecto

que se deve ter em mente é que, duas galáxias que possuem coordenadas próximas, em nossa análise, dois minutos de arco, não necessariamente precisam estar em interação, elas podem estar distantes entre si, não tendo nesse caso interação, e não sendo classificadas nas categorias 1 ou 2 do Catálogo de AM87.

5.5 Comparação com outros catálogos

Com o objetivo de verificar se nossos pares candidatos a galáxias interagentes das categorias 1 e 2 de AM87 já foram anteriormente identificados, fizemos uma busca em duas bases de dados contém um grande volume de dados de objetos extragalácticos, a saber HyperLeda[19] e NED (NASA/IPAC Extragalactic Database)[20]. Fizemos uma análise no VizieR do CDS e os catálogos com pares de galáxias encontrados estavam todos localizados no Hemisfério Norte, alguns destes catálogos indicados na Introdução. Outros, com dados já contidos no NED, como o trabalho de Soares et al. (1995).

O catálogo HyperLeda é um dos mais usados na literatura e, com grande cobertura espacial, fornecendo propriedades básicas para, aproximadamente, 2 milhões de galáxias. Existem outros catálogos com profundidade até maior que o HyperLeda, notadamente para regiões específicas, ou seja, não para todo o céu. Em sendo, nosso objetivo, em um primeiro momento, ter uma amostra que represente de maneira satisfatória o número de objetos e não a completeza, preferimos usar esse catálogo, como uma perspectiva pretendemos verificar outras bases.

Nossa consulta, com uma pesquisa SQL na base HyperLeda, teve como finalidade, obter todos os objetos do catálogo, que estão localizados no Hemisfério Sul ($\delta < 0°$) e com classificação de objeto como múltiplo, pois o HyperLeda não tinha um campo específico para pares de galáxias. No total, obtivemos 4.882 objetos com a distribuição espacial em coordenadas equatoriais apresentadas na Figura 5.11. Vemos uma grande concentração de objetos em torno de RA aproximadamente igual à zero, e regiões com ausências de galáxias, devido ao plano de nossa Galáxia.

[19] http://leda.univ-lyon1.fr/leda/fullsql.html
[20] https://ned.ipac.caltech.edu/

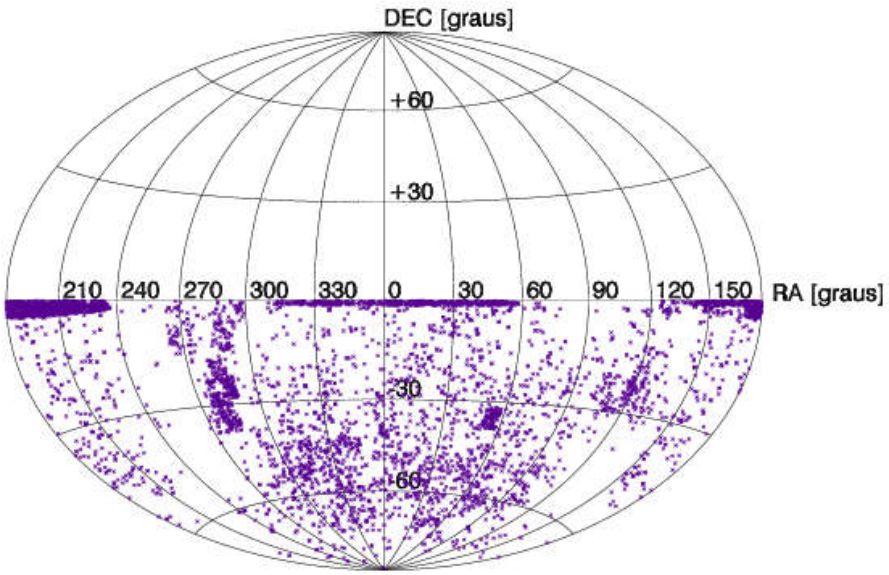

Figura 5.11: Distribuição dos objetos selecionados no Catálogo HyperLeda como galáxias e com multiplicidade.

Por seu lado, o NED em Caltech, disponibiliza uma excelente base de dados para uso em Astronomia, de maneira geral, e mais particularmente, no campo da Astronomia Extragaláctica. No tocante a pares de galáxias, utilizamos a base[21] que contém trabalhos com pares de galáxias, mencionados na introdução do presente trabalho.

O catálogo contém 2.318 entradas para 1.963 objetos, e além das coordenadas dos objetos, possui informação sobre tipo morfológico, redsfhit, e, claro, o nome do objeto. Em uma primeira etapa fizemos a conversão de coordenadas equatoriais (J2000), que estavam no formato sexagesimal para decimal. A distribuição dos objetos pode ser vista na Figura 5.12, com pares de galáxias também localizados no Hemisfério Norte e com presença praticamente nula de objetos próximos ao plano de nossa Galáxia.

Realizamos o cruzamento, simplesmente verificando se, no campo (dois minutos de arco) de nossas candidatas, existiam pares de galáxias previamente identificados, uma análise conservadora pois, o mesmo objeto pode ter coordenadas diferentes, nas várias bases, sendo necessário em uma etapa futura, verificar, sobrepondo as coordenadas dos pares identificados à nossas candidatas, se são o mesmo objeto. Por essa razão, temos um número elevado de pares para candidatas que não foram verificadas visualmente. Em todo o caso, o nosso objetivo é sobrestimar o número de pares já identificados de forma a termos um limite inferior em nossas candidatas.

[21] https://ned.ipac.caltech.edu/cgi-bin/INFatt?dom=H&id=4696

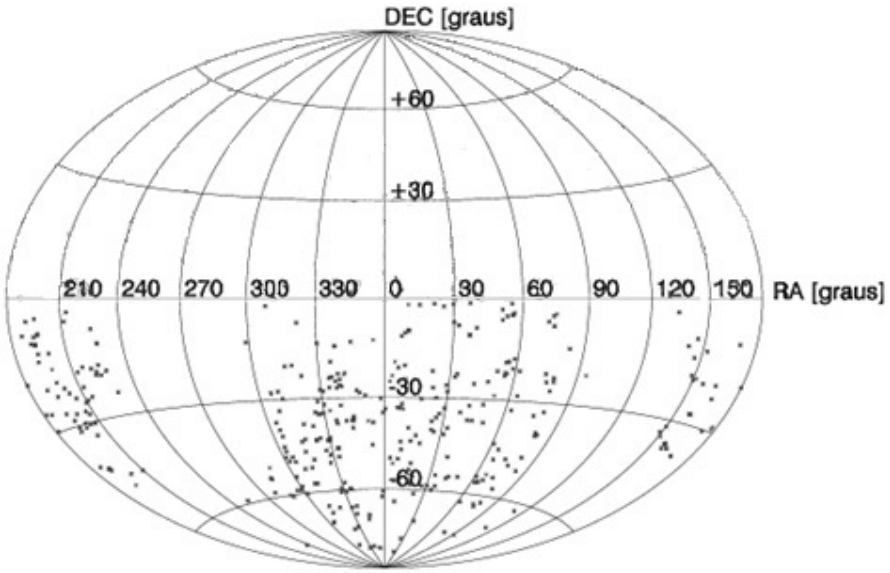

Figura 5.12: Distribuição em coordenadas equatoriais dos objetos identificados como pares de galáxias na base do NED.

A Tabela 5.3 apresenta nossas candidatas a galáxias interagentes, em relação às existentes nos catálogos obtidos da base de dados HyperLeda e NED. Vemos que o número de objetos coincidentes no HyperLeda é muito superior aos encontrados no NED, o que se deve ao fato da base HyperLeda ser mais atualizada do que a do NED, ao menos no tocante à classificação específica de alguns objetos, como pares de galáxias. Dentre os 9 objetos que identificamos no NED, apenas dois possuíam identificação nos 160 da base HyperLeda.

Tabela 5.3: Comparação entre as candidatas à pares de galáxias encontradas em nosso estudo e as identificadas pela base HyperLeda e NED.

vc_1	vc_2	HyperLeda	NED
1	1	64	5
1	-1	91	4
1	-2	5	0

A Tabela 5.4 apresenta o número de candidatas a galáxias interagentes, das categorias 1 e 2, para cada combinação da flag, vc, com probabilidade de classificação entre 51-60% e superior à 60%. Utilizando esse último valor de probabilidade, nos permite termos objetos com maiores chances de estarem classificados corretamente em uma das duas categorias, filtrando, dessa forma, ainda mais a nossa lista de candidatas a galáxias peculiares. Conforme, pode ser visto na Tabela 5.4, apesar da redução do número de candidatas, o número ainda é alto, em relação aos pares de galáxias conhecidos que temos, e serve, também, como uma estimativa de limite inferior no número de candidatas

Tabela 5.4: Distribuição das candidatas para ambas as categorias para cada combinação de vc e para duas diferentes faixas de probabilidade de classificação.

		Probabilidade: 51-60%		Probabilidade > 60%	
vc_1	vc_2	Categoria 1	Categoria 2	Categoria 1	Categoria 2
1	1	228	527	31	357
1	-1	705	1.651	101	1.088
1	-2	31	73	4	34

A Figura 5.13 apresenta a distribuição das candidatas a galáxias interagentes para as duas faixas de probabilidades, conforme apresentado na Tabela 5.4, para ambas as categorias, considerando duas combinações de verificação visual, que mais contém objetos, ou seja, as duas primeiras.

Os dois mapas da Figura 5.13 compreendem uma mesma região do céu, ambas cobrindo uma mesma faixa de coordenadas. Nota-se, porém, uma quantidade maior de candidatas a galáxias interagentes, na imagem superior; esta imagem mostra a distribuição das candidatas que obtiveram probabilidades entre 51-60%. Na imagem inferior, percebe-se uma redução na quantidade de candidatas distribuída pelo mapa, neste caso apenas estão sendo consideradas as que tiveram probabilidades maior do que 60%.

Analisamos, também, as propriedades dos objetos, no tocante à sua distribuição de brilho, separando, nos pares de galáxias, os objetos mais brilhantes (magnitudes menores) e os fracos (magnitudes maiores), para ambas as separações em probabilidade.

A Figura 5.14 apresenta a distribuição em magnitudes destes objetos. Nota-se que, para os objetos com brilhos mais fracos (lado esquerdo da Figura 5.14), um limite de completeza similar, exceto para objetos com vc = -1 (linha fina, painel inferior). Para os objetos mais fracos temos, novamente, um limite similar, exceto, novamente, para objetos com vc = -1, os quais podem, em sua maioria, serem estrelas. A grosso modo, podemos estimar que o limite de completeza para os objetos mais brilhantes esteja, entre 12,5 - 13,0 mag e, para os mais fracos, seja de, aproximadamente, 14,0 mag, esses são os valores de magnitude limite, que estimamos para nosso método, em concordância com o que foi encontrado no Capítulo 4.

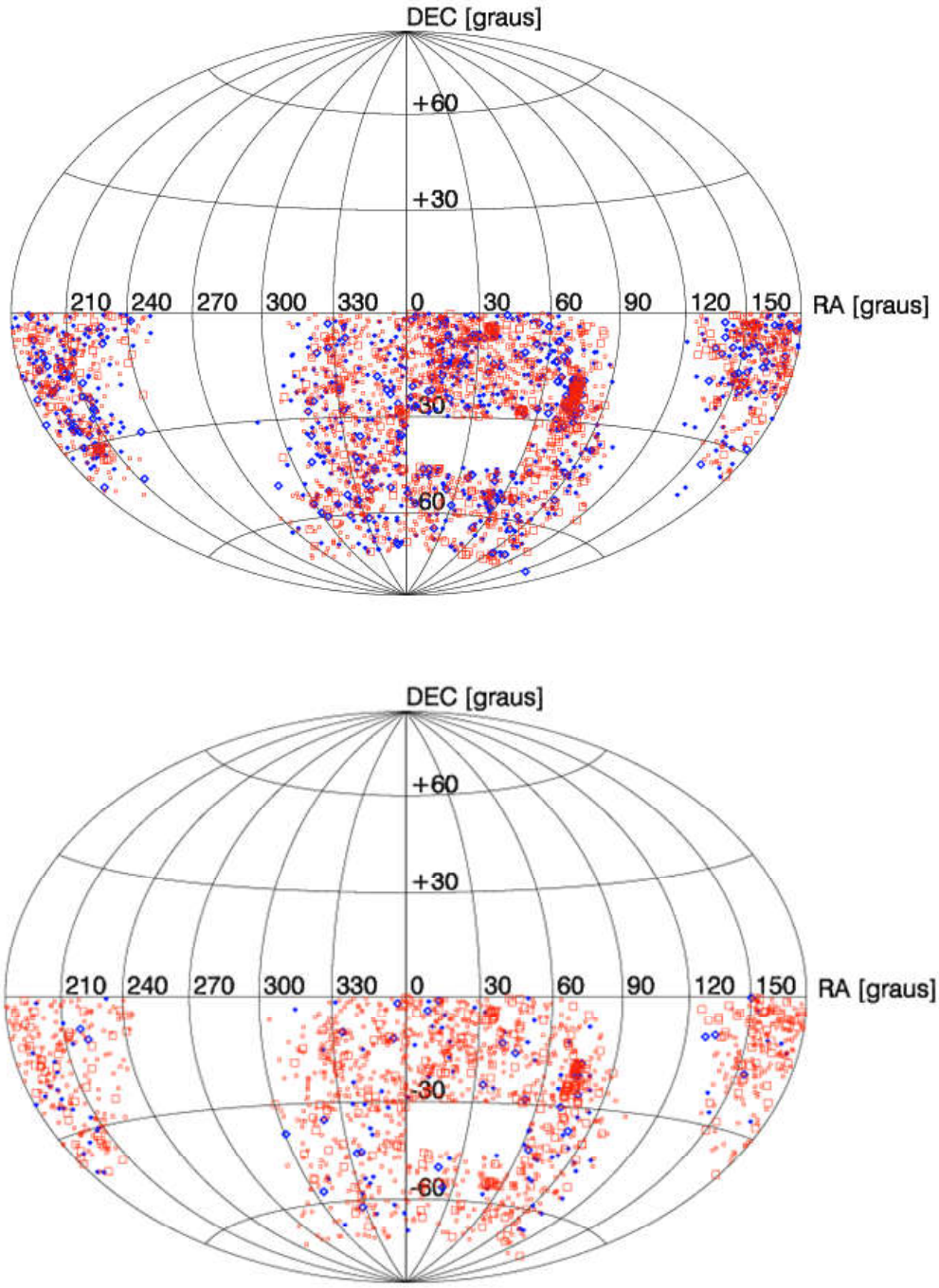

Figura 5.13: Distribuição, em coordenadas equatoriais, das galáxias, segundo duas faixas de probabilidades (ver Tabela 5.4), probabilidade entre 50-60% (parte superior), superior à 60% (parte inferior). Para ambas, os diamantes (em azul) e quadrados (em vermelho), representam categorias 1 e 2, respectivamente, sendo que os símbolos maiores e menores representam candidatas a galáxias com vc = 1 e -1, respectivamente.

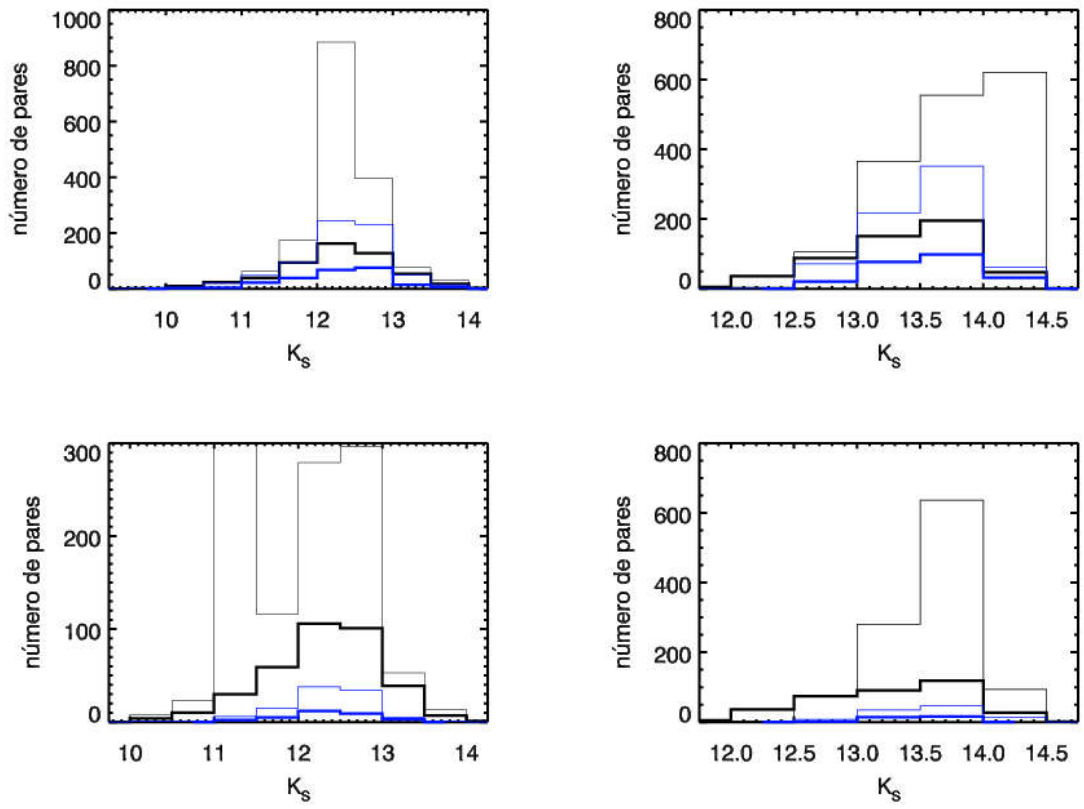

Figura 5.14: Distribuição de magnitudes no filtro K_S. Parte superior (probabilidade dentre 51-60%) e parte inferior (probabilidade maior do que 60%). Na esquerda, candidata à galáxia mais brilhante. Linhas azuis e pretas representam categoria 1 e 2, e linhas mais espessas vc = 1 enquanto as mais finas vc = -1.

Capítulo 6

Conclusão e perspectivas

No presente trabalho, foi elaborado um método para identificar e classificar objetos como candidatos a galáxias peculiares das categorias 1 e 2, do Catálogo de Arp & Madore (1987), tendo como base os dados do 2MASS no infravermelho próximo, nos filtros J, H e K_S, para todo o Hemisfério Sul.

Em uma primeira etapa, foi feito um estudo do Catálogo de AM87, tendo como objetivo conhecer as propriedades das galáxias peculiares existentes, assim como as propriedades das imagens do Catálogo de Fontes Extensas do 2MASS (XSC-2MASS), assim obtê-las automaticamente, utilizando-se o Aladin. Elaboramos amostras de treinamento no infravermelho. O programa mostrou-se capaz de recuperar (classificar), de uma grande lista de imagens, as imagens que foram incluídas na amostra de treinamento. Obtivemos um índice médio de acerto, com as imagens da amostra de treinamento, de 95%, considerando-se as simulações feitas com os "melhores" parâmetros.

Em nosso método, selecionamos imagens do 2MASS com tamanho de dois minutos de arco, tendo como base uma análise dos dados do XSC-2MASS, selecionando-se imagens que contivessem, ao menos, um objeto definido, visualmente, pela equipe do 2MASS como galáxia. Dessa forma, evitamos obter toda a base de imagens do 2MASS, o que resultaria em, aproximadamente 25 milhões de imagens de dois minutos de arco, sendo que a grande maioria com um elevado número de estrelas; em imagens que previamente, sabemos que não possuem galáxias, além de sistemas compostos por mais de duas galáxias, o que estaria em desacordo com as duas primeiras categorias do Catálogo de AM87.

Dessa forma, utilizamos imagens que tenham limites específicos do número de estrelas em função da magnitude, assim como de extinção interestelar, os quais foram determinados no presente estudo. Apenas, seguindo esses critérios uma imagem é analisada pelo Wndchrm, ou seja, com estas características similares às da amostra de treinamento.

Ao testarmos a eficiência do Wndchrm na amostra de imagens para uma determinada região, que compreendeu uma área de 573 graus quadrados, obtivemos uma baixa taxa de acerto para a Categoria 1 e uma taxa de acerto razoável para Categoria 2. Uma das possíveis razões é, ainda, a contaminação estelar de nossas imagens de treinamento, assim como das imagens que são usadas para se verificar a presença de pares das categorias 1 ou 2 de AM87. Outro aspecto reside no fato de, dentre as imagens classificadas em ambas as categorias, alguns objetos podem não corresponder a pares de galáxias.

a contaminação estelar de nossas imagens pode levar o software a erros, sendo que isso é mais frequente na tentativa de classificação das imagens na Categoria 1, pois

nessa categoria temos a galáxia companheira aparente, com tamanho menor que, pelo menos, 50% da galáxia principal, o que faz com que pequenos objetos, como estrelas, sejam confundidos com galáxias companheiras.

Realizamos uma comparação com catálogos de galáxias, tais como, os fornecidos pelas bases HyperLeda e pelo NED e encontramos, ao menos, 400 candidatas a pares, levando apenas em consideração objetos classificados visualmente pelo 2MASS como galáxias, e objetos com probabilidade fornecida pelo Wndchrm de ser classificado em uma das duas categorias superior a 60%. Essa estimativa praticamente dobra se considerarmos pares classificados pelo Wndchrm com probabilidade entre 51-60%. O aumento é ainda mais expressivo (3.500), quando consideramos pares que contenham, ao menos, um dos objetos que não foram inspecionados pelo 2MASS e, que podem ou não serem galáxias. Em todo o caso, os objetos classificados em uma das categorias, sem classificação visual pelo 2MASS, podem ao menos serem candidatas a galáxias, servindo nosso estudo de ferramenta para eventuais revisões da classificação destes objetos.

As candidatas encontradas nesse trabalho devem ser melhor estudadas, analisando suas cores, assim como usando-se levantamentos específicos, que não estão na base NED e HyperLeda, como p.ex., os do levantamento VISTA. Cabe ressaltar, que uma boa determinação, de que as galáxias estão ou não em interação, pode ser feita por meio da análise de seus redshifts, embora, mesmo galáxias em interação, possam ter redshifts discordantes, como o caso de galáxias com redshift anômalo, que estão claramente em interação, mas possuem redsfhits discrepantes.

Os programas para buscar as imagens, selecionar objetos de outros catálogos e analisar os resultados do programa Wndchrm se mostraram eficientes, permitindo analisar dados de catálogos, como o 2MASS, consultar coordenadas de objetos específicos, bem como fazer o download dos mesmos.

Dentre as melhorias que se pretende implementar no futuro, estão as que se referem aos parâmetros do Wndchrm, assim como ferramentas de processamento e tratamento de imagens. Apesar do Wndchrm tratar internamente as imagens, tanto no treinamento, quanto na classificação, p.ex., com o uso do parâmetro M, que aplica um filtro de limiarização de Otsu a uma imagem, pretendemos aplicar uma limiarização independente, antes de submeter as imagens ao Wndchrm; e, dessa forma, eliminar os objetos de brilho fraco dos campos visuais das imagens.

Um outro aspecto é que podemos melhorar, os acertos, não fazendo uso de uma amostra definida por nós, em 573 graus quadrados, mas, realizar comparações utilizando as galáxias do Catálogo de AM87 para as categorias 1 e 2.

Em um outro momento pretendemos aplicar nossa metodologia para outros comprimentos de onda, tais como o do óptico. Tendo em vista, os recentes grandes levantamentos e os que virão no futuro, como LSST, GAIA, ente outros, existe a grande perspectiva para a aplicação da metodologia empregada nesses levantamentos.

Referências Bibliográficas

[Amôres e Lépine 2005] Amôres, E.B. e Lépine, J.R.D. (2005). Models for Interstellar Extinction in the Galaxy. The Astronomical Journal, 130(2): 659-673.

[Amôres e Lépine 2007] Amôres, E.B. e Lépine, J.R.D. (2007). Comparing Extinction Models with a Sample of Elliptical Galaxies, Star Clusters, and the Extinction at the Galactic Center. The Astronomical Journal, 133(4): 1538-3881.

[Amôres et al. 2013] Amôres, E.B. et al. (2013). The long bar as seen by the VVV Survey - II. Star counts. Astronomy & Astrophysics, 559(1): A11 14.

[Amôres et al. 2012a] Amôres, E.B., Moitinho, A., Arsenijevic, V. e Sodré, L., (2012a). GALExtin: A VO-Service for Estimating Galactic Interstellar Extinction. In Star Clusters in the Era of Large Surveys, Astrophysics and Space Science Proceedings. Berlin: Springer-Verlag Berlin Heidelberg. p.93.

[Amôres et al. 2011] Amôres, E.B. et al., (2011). Simulations of Star Counts and Galaxies Towards Vista Variables in the via Láctea Survey Region. Environment and the Formation of Galaxies. In Astrophysics and Space Science Proceedings. Berlin: Springer. pp.141-143.

[Amôres et al. 2012b] Amôres, E.B. et al. (2012b). Galaxies Behind the Galactic Plane: First Results and Perspectives From the Vvv Survey. The Astronomical Journal, 144(5): 127.

[Anzai 1992] Anzai, Y., (1992). Pattern Recognition & Machine Learning. Academic Press. 1st edition. San Diego, CA.

[Arce e Goodman 1999] Arce, H.G. e Goodman, A.A. (1999). Measuring Galactic Extinction: A Test. The Astrophysical Journal, 512(2): L135-L138.

[Argudo-Fernandez et al. 2015] Argudo-Fernandez, M. et al. (2015). Catalogues of isolated galaxies, isolated pairs, and isolated triplets in the local Universe. Astronomical and Astrophysical Journal, 578A: 110.

[Arp e Madore 1987] Arp, H.C. e Madore, B., (1987). Catalogue of Southern Peculiar Galaxies and Associations, Vol I, Positions and Descriptions. Cambridge University Press. New York.

[Barden et al. 2012] Barden, M. et al. (2012). GALAPAGOS: Galaxy Analysis over Large Areas: Parameter Assessment by GALFITting Objects from SExtractor. Astrophysics Source Code Library, record ascl:1203.002.

[Bezdek 1993] Bezdek, J.C. (1993). Review of Probabilistic, Fuzzy, and Neural Models for Pattern Recognition. The Journal of Intelligent and Fuzzy Systems, vol.1, pgs. 1-25.

[Boch e Fernique 2014] Boch, T. e Fernique, P. (2014). Aladin Lite: Embed your Sky in the Browser. Astronomical Data Analysis Software and Systems XXIII, 485:

277-281.

[Bonnarel et al. 2000] Bonnarel, F. et al. (2000). The ALADIN interactive sky atlas - A reference tool for identification of astronomical sources. Astronomy & Astrophysics Supplement Series, Abril. 33-40.

[Burstein 2003] Burstein, D. (2003). Line-of-Sight Reddening Predictions: Zero Points, Accuracies, the Interstellar Medium, and the Stellar Populations of Elliptical Galaxies. The Astronomical Journal, 126(4): 1849-1860.

[Cao et al. 2016] Cao, C. et al. (2016). Herschel observations of Major merger pairs at z = 0: dust mass and star formation. The Astrophysical Journal Supplement Series, 222(2).

[Cole et al. 2001] Cole, S. et al. (2001). The 2dF galaxy redshift survey: near-infrared galaxy luminosity functions. Monthly Notices of the Royal Astronomical Society, 326(1): 255-273.

[Cost e Salzberg 1993] Cost, S. e Salzberg, S. (1993). A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. Machine Learning, 10(1): 57-78.

[Cotini et al. 2013] Cotini, S. et al. (2013). The merger fraction of active and inactive galaxies in the local Universe through an improved non-parametric classification. MNRAS, 431: 2661–2672.

[Cutri et al. 2003] Cutri, R.M. et al. (2003). 2MASS All Sky Catalog of point sources. NASA/IPAC Infrared Science Archive, June.

[Dodd e MacGillivray 1986] Dodd, R.J. e MacGillivray, H.T. (1986). Automated detection of clusters of galaxies. Astronomical Journal , 92: 706-712.

[Domingue et al. 2009] Domingue, D.L., Xu, C.K., Jarrett, T.H. e Cheng, Y. (2009). 2MASS/SDSS Close Major-Merger Galaxy Pairs: Luminosity Functions and Merger Mass Dependence. Astrophysical Journal, 695(2): 1559-1566.

[du Buisson et al. 2015] du Buisson, L., Sivanandam, N., Bassett, B.A. e Smith, M. (2015). Machine learning classification of SDSS transient survey images. Monthly Notices of the Royal Astronomical Society., 454(2): 2026-2038.

[Duda et al. 2000] Duda, R.O., Hart, P.E. e Stork, D.G., (2000). Pattern Classification. John Wiley & Son. New York.

[Faúndez-Abans et al. 2015] Faúndez-Abans, M. et al. (2015). Visiting two objects in the field of the ring galaxy HRG 2302. Astronomy and Astrophysics, 574: A70.

[Focardi et al. 2006] Focardi, P., Zitelli, V., Marinoni, S. e Kelm, B. (2006). A new sample of bright galaxy pairs in UZC. Astronomy and Astrophysics Journal, 456: 467-472.

[Frei 2001] Frei, Z. (2001). Mining 2D images: Automatic morphological classification of galaxies.. Mining the Sky, 344-346\r705.

[Freitas-Lemes et al. 2013] Freitas-Lemes, P., Rodrigues, I., Dors, O. e Faúndez-Abans, M. (2013). Análise da Atividade Nuclear de Dez Galáxias com Anel Polar. Univap Online, 19(34): 53-57.

[Freitas-Lemes et al. 2014] Freitas-Lemes, P. et al. (2014). The effects of interaction

on the kinematics and abundance of AM 2229-735. Monthly Notices of the Royal Astronomical Society, 441(2): 1086-1094.

[Friedman e Kandel 1999] Friedman, M. e Kandel, A., (1999). Introduction to Pattern Recognition : Statistical, Structural, Neural and Fuzzy Logic Approaches. World Scientific. 1st edition. Singapura.

[Gabor 1946] Gabor, D. (1946). Theory of communication. Journal of IEEE , 93: 429–457.

[Geller et al. 2006] Geller, M.J. et al. (2006). Infrared Properties of Close Pairs of Galaxies. The Astronomical Journal, 132(6): 2243-2259.

[Gonzalez e Woods 2010] Gonzalez, R.C. e Woods, R.E., (2010). Processamento Digital de Imagens. Pearson Education. 3rd edition.

[Gradshtein e Ryzhik 1994] Gradshtein, I. e Ryzhik, I., (1994). Table of integrals, series and products. Academic Press. 5th edition.

[Hadjidementriou et al. 2001] Hadjidementriou, E., Grossberg, M. e Nayar, S. (2001). Spatial information in multiresolution histograms. In IEEE Conference on Computer Vision., 2001.

[Haralick et al. 1973] Haralick, R.M., Shanmugam, K. e Dinstein, I. (1973). Textural features for image classification. IEEE Trans on Systems, Man, and Cybernetics, 6: 269-285.

[Hebb 1949] Hebb, D.O. (1949). The Organization of Behavior. Wiley.

[Hocking et al. 2015] Hocking, A., Geach, J.E., Davey, N. e Sun, Y. (2015). Teaching a machine to see: unsupervised image segmentation and categorisation using growing neural gas and hierarchical clustering. Monthly Notices of the Royal Astronomical Society, 000(July): 1-5.

[Hubble 1926] Hubble, E.P. (1926). Extragalactic nebulae. Astrophysical Journal, December. 321-369.

[Huertas-Company 2013] Huertas-Company, M. (2013). GALSVM: Automated Morphology Classification. Astrophysics Source Code Library, ascl:1304.003.

[Jie et al. 2012] Jie, L., Jigui, S. e Shengsheng, W. (2012). Pattern Recognition: an Overview. American Journal of Intelligent Systems, v. 2, n. 1, p. 23–27.

[Kauffmann et al. 1993] Kauffmann, G., White, S. e Guiderdoni, B. (1993). The formation and evolution of galaxies within merging dark matter haloes. Monthly Notices of the Royal Astronomical Society, 264(1): 201-218.

[Lenzen et al. 2013] Lenzen, F., Schindler, S. e Scherzer, O. (2013). Automatic detection of arcs and arclets formed by gravitational lensing. Astronomy & Astrophysics, 4662: 12.

[Lépine e Leroy 2000] Lépine, J.R.D. e Leroy, P. (2000). A new model for the infrared brightness of the Galaxy. Monthly Notices of the Royal Astronomical Society, 313(2): 263-270.

[Lim 1990] Lim, J.S., (1990). Two-Dimensional Signal and Image Processing. Prentice Hall.

[Lintott et al. 2011] Lintott, C. et al. (2011). Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. Monthly Notices of the Royal Astronomical Society, 410(1): 166-178.

[Lintott et al. 2008] Lintott, C.J. et al. (2008). Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, 389: 1179–1189.

[López-Corredoira et al. 2001] López-Corredoira, M. et al. (2001). Searching for the in-plane Galactic bar and ring in DENIS. Astronomy and Astrophysics, 373 (1): 139-152.

[Loveday et al. 1996] Loveday, J., Efstathiou, G., Maddox, S.J. e Peterson, B.A. (1996). The Stromlo-APM Redshift Survey. III. Redshift Space Distortions, Omega, and Bias. The Astrophysical Journal, 468: 1.

[Loveday et al. 1995] Loveday, J., Maddox, S.J., Efstathiou, G. e Peterson, B.A. (1995). The Stromlo-APM redshift survey. II. Variation of galaxy clustering with morphology and luminosity. Astrophysical Journal, 442(2): 457–468.

[Loveday et al. 1992] Loveday, J., Peterson, B.A., Efstathiou, G. e Maddox, S.J. (1992). The Stromlo-APM Redshift Survey. I - The luminosity function and space density of galaxies. The Astrophysical Journal, 390(2): 338.

[Loveday et al. 1996] Loveday, J., Peterson, B.A., Maddox, S.J. e Efstathiou, G. (1996). The Stromlo-APM Redshift Survey. IV. The Redshift Catalog. Astrophysical Journal, Supplement Series, 107(1): 201–214.

[Makarov et al. 2014] Makarov, D. et al. (2014). HyperLEDA. III. The catalogue of extragalactic distances. Astronomy & Astrophysics, 570: id.A13, 12 pp.

[Marr e Nishihara 1978] Marr, D. e Nishihara, H.K. (1978). Representation and recognition of the spatial organization of three dimensional structure. Proc. R. Soc. Lond. B Biol. Sci., 200: 269–294.

[McCulloch e Pitts 1943] McCulloch, W.S. e Pitts, W.H. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, May. 115-133.

[Medeiros 2006] Medeiros, L.F.d., (2006). Redes neurais em Delphi. Visual Books. 2nd edition. Florianópolis.

[Minchin et al. 2010] Minchin, R.F. et al. (2010). The Arecibo Galaxy Environment Survey. III. Observations Toward the Galaxy Pair NGC 7332/7339 and the Isolated Galaxy NGC 1156. The Astronomical Journal, 140(4): 1093-1118.

[Minsky e Papert 1969] Minsky, M. e Papert, S., (1969). Perceptrons. MIT Press. Cambridge, MA.

[Naim e Lahav 1997] Naim, A. e Lahav, O. (1997). What is a peculiar galaxy? MNRAS, March. 969-978.

[Neugebauer e Leighton 1969] Neugebauer, G. e Leighton, R.B., (1969). Two-micron sky survey. A preliminary catalogue. NASA SP. Washington.

[Orlov et al. 2007] Orlov, N. et al., (2007). Computer vision for microscopy

applications. In G. Obinata e A. Dutta, eds. Vision Systems – Segmentation and Pattern Recognition. Viena: ARS Press. pp.221-242.

[Ortega 2001] Ortega, N.R.S. (2001). Aplicação davTeoria dos Conjuntos Fuzzy a Problemas da Biomedicina. Tese Doutorado. São Paulo: Universidade de São Paulo.

[Otsu 1979] Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics - TSMC, v. 9, n. 1, pp. 62-66.

[Pappis e Mamdani 1977] Pappis, C.P. e Mamdani, H. (1977). A Fuzzy Logic Controller for Traffic Junctions. IEEE Transactions on Systems, Man and Cybernetics, SMC-7, Nº 10.

[Peng 2010] Peng, C. (2010). Using GALFIT to Classify Galaxies at High-z: Measuring Asymmetry Parametrically. Bulletin of the American Astronomical Society, v. 42, p.578.

[Prewitt 1970] Prewitt, J.M., (1970). Object enhancement and extraction. In Picture Processing and Psychopictoris. New York.

[Robin et al. 2012] Robin, A.C., Marshall, D.J., Schultheis, M. e Reylé, C. (2012). Stellar populations in the Milky Way bulge region: towards solving the Galactic bulge and bar shapes using 2MASS data. Astronomy & Astrophysics, 538: id.A106, 14 pp.

[Robin et al. 2003] Robin, A.C., Reylé, C., Derrière, S. e Picaud, S. (2003). A synthetic view on structure and evolution of the Milky Way. Astronomy and Astrophysics, 409: 523-540.

[Robin et al. 2014] Robin, A.C. et al. (2014). Constraining the thick disc formation scenario of the Milky Way. Astronomy & Astrophysics, 569: id.A13, 26 pp.

[Rosenblatt 1962] Rosenblatt, F., (1962). A comparison of several perceptron models. Spartan Books. Washington, DC.

[Rumelhart et al. 1986] Rumelhart, D.E., Hinton, G.E. e Williams, R.J. (1986). Learning Representations by Back-Propagating Errors. Nature, 323: 533 - 536.

[Schlegel et al. 1998] Schlegel, D.J., Finkbeiner, D.P. e Davis, M. (1998). Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds. The Astrophysical Journal, 500(2): 525-553.

[Schombert et al. 1990] Schombert, J.M., Wallin, J.F. e Struck-Marcell, C. (1990). A multicolor photometric study of the tidal features in interacting galaxies. The Astronomical Journal, 99: 497.

[Shalyapina et al. 2007] Shalyapina, L.V., Merkulova, O.A., Yakovleva, V.A. e Volkov, E.V. (2007). 2D spectroscopy of candidate polar-ring galaxies: I. The pair of galaxies UGC 5600/09. Astronomy Letters, 33(8): 520-530.

[Shamir 2005] Shamir, L..N.R.J. (2005). A Fuzzy Logic based algorithm for finding astronomical objects in wide-angle frames. Publications of the Astronomical Society of Australia, vol. 22(2), pp. 111-117.

[Shamir 2006] Shamir, L. (2006). Human perception-based color segmentation using Fuzzy Logic. In International Conference on Image Processing, Computer Vision and Pattern Recognition. Las Vegas, 2006.

[Shamir 2011] Shamir, L. (2011). Ganalyzer: A tool for automatic galaxy image analysis. Astrophysical Journal, v. 736, n. 2, pp. 141.

[Shamir et al. 2013] Shamir, L., Holincheck, A. e Wallin, J. (2013). Automatic quantitative morphological analysis of interacting galaxies. Astronomy & Computing, v. 2, pp. 67-73.

[Shamir et al. 2010] Shamir, L. et al. (2010). Impressionism, Expressionism, Surrealism: Automated Recognition of Painters and Schools of Art. ACM Transactions on Applied Perception, Fevereiro. 1-17.

[Shamir e Nemiroff 2005] Shamir, L. e Nemiroff, R.J. (2005). Photzip: A lossy FITS image compression algorithm that protects user-defined levels of photometric integrity. The Astronomical Journal, vol. 129(1), pp. 539-546.

[Shamir et al. 2008] Shamir, L. et al. (2008). Wndchrm – an open source utility for biological image analysis. Source Code for Biology and Medicine, 8 Julho. 1-13.

[Shamir e Wallin 2014] Shamir, L. e Wallin, J. (2014). Automatic detection and quantitative assessment of peculiar galaxy pairs in Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, v. 443, n. 4, 3528-3537.

[Singhala et al. 2014] Singhala, P., Shah, D.N. e Patel, B. (2014). Temperature Control using Fuzzy Logic. International Journal of Instrumentation and Control Systems, 4(1): 1-10.

[Siqueira et al. 2000] Siqueira, M.L. et al. (2000). Echocardiographic image sequence segmentation using self-organizing maps. Neural Networks for Signal Processing, 2: 594–603.

[Skrutskie et al. 2006] Skrutskie, M.F. et al. (2006). The Two Micron All Sky Survey (2MASS). The Astronomical Journal, v 131, n 2, pp. 1163-1183.

[Soares et al. 1995] Soares, D.S.L., de Souza, R.E., de Carvalho, R.R. e Couto Da Silva, T.C. (1995). Southern binary galaxies. I. A sample of isolated pairs. Astronomical and Astrophysical Journal, 110: 371-381.

[Struck 1999] Struck, C. (1999). Galaxy collisions. Physics Reports, 321: 1–137.

[Tamura et al. 1978] Tamura, H., Mori, S. e Yamavaki, T. (1978). Textural features corresponding to visual perception. IEEE Trans On Systems, Man and Cybernetics, 8: 460-472.

[Tanaka 1995] Tanaka, E. (1995). Theoretical Aspects of Syntactic Pattern Recognition. Pattern Recognition, v. 28, n. 7, p. 1053–1061.

[Tanscheit e Scharf 1988] Tanscheit, R. e Scharf, E.M. (1988). Experiments with the use of a Rule-Based Self-Organising Controller for Robotics Applications. Fuzzy Sets and Systems, 26(2): 195-214.

[Teague 1980] Teague, M. (1980). Image analysis via the general theory of moments. Journal of the Optical Society of America, 70: 920-930.

[Toshifumi et al. 2005] Toshifumi, Y. et al. (2005). Automatic Detection Algorithm for Small Moving Objects. Astronomical Soc of Japan, 57: 399-408.

[Vorontsov-Velyaminov 1959] Vorontsov-Velyaminov, B.A. (1959). Atlas and Catalogue of interacting galaxies. Part 1. Moscow University.

[Vorontsov-Velyaminov 1977] Vorontsov-Velyaminov, B.A. (1977). Atlas of interacting galaxies, part II and the concept of fragmentation of galaxies. Astronomy and Astrophysics , 28: 1 - 117.

[Vorontsov-Velyaminov et al. 1962 - 1974] Vorontsov-Velyaminov, B.A., Krasnogorskaya, A.A. e V.P., A., (1962 - 1974). Morphological catalogue of galaxies. Moscow State University. Moscow.

[Whittet 1992] Whittet, D.C.B., (1992). Dust in the Galactic Environment. Bristol: Inst. Physics Publ. 2nd edition. London.

[Woods et al. 2006] Woods, D.F., Geller, M.J. e Barton, E.J. (2006). Tidally triggered star formation in close pairs of galaxies: major and minor interactions. The Astronomical Journal, 132: 197-209.

[Zadeh 1996] Zadeh, L.A. (1996). Fuzzy Logic: Computing with Words. IEEE TransFuzzySys, 4(3): 103 - 111.