Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

Uma aplicação alternativa para agrupamento de galáxias

Adilson Oliveira de Almirante

Feira de Santana

2017

Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

Adilson Oliveira de Almirante

Uma aplicação alternativa para agrupamento de galáxias

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

2017

Ficha Catalográfica – Biblioteca Central Julieta Carteado

A455 Almirante, Adilson Oliveira de
Uma aplicação alternativa para agrupamento de galáxicas/ Adilson Oliveira de Almirante . – 2017.
60 f.: il.

Orientador: Eduardo Brescansin de Amôres

Dissertação (mestrado) – Universidade Estadual de Feira de Santana, Programa de Pós-Graduação em Computação Aplicada.

1. Galáxias.2. Aatronomia. I. Amôres, Eduardo Brescansin de, orient. II. Universidade Estadual de Feira de Santana. III. Título.

CDU: 524.77

Adilson Oliveira de Almirante

Uma aplicação alternativa para agrupamento de galáxias

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, 29 de agosto de 2017

BANCA EXAMINADORA

Dr. Eduardo Brescansin de Amôres (Orientador)
Universidade Estadual de Feira de Santana

Dr. Marildo Geraldête Pereira
Universidade Estadual de Feira de Santana

Dra. Michele Fúlvia Angelo
Universidade Estadual de Feira de Santana

Abstract

In this work, we present an alternative method for clustering of galaxies, based on the data of galaxies available in large scale astronomical surveys. We have used the Friends of Friends (FoF) algorithm, as the basis of our clustering, with *LL* (Linking Length) and the minimum value of separation between galaxies, as our parameters. We have used a catalogue with approximately 44 thousand galaxies. After the simulation with six different values of *LL*, we analysed our clustering by means of the comparison with the obtained one by Tempel et al. (2016). In this comparison, we obtained a similarity of 79%. We also supplemented the validation of our method with an analysis of the silhouette coefficient, obtaining a general coefficient equal to 0.80. We present the characteristics of the groups obtained in our work, their comparison with the obtained one by Tempel et al. (2016) as well as the discussion with the non-grouped galaxies.

Keywords: algorithms, clustering, large structures, galaxies, large astronomical surveys.

Resumo

Neste trabalho apresentamos um método alternativo para agrupamento de galáxias, por meio do uso, dos dados de galáxias disponíveis em grandes levantamentos astronômicos. Utilizamos o algoritmo FoF, *Friends of Friends*, como base para o agrupamento, com os parâmetros de *LL* (*Linking Length*) e o valor mínimo de galáxias, em comum para definir a união entre grupos próximos. Utilizamos um catálogo com aproximadamente 44 mil galáxias e após a simulação com seis valores distintos para o *LL*, e com a comparação dos respectivos agrupamentos com os obtidos por Tempel et al. (2016), obtivemos, similaridade, de aproximadamente 79%. Complementamos a avaliação com uma análise do Coeficiente se Silhueta, e encontramos agrupamentos com coeficiente geral de 0,80. Apresentamos a discussão e características dos grupos obtidos, assim como das galáxias não agrupadas.

Palavras-chave: algoritmos, agrupamentos, grandes estruturas, galáxias, grandes levantamentos astronômicos.

Agradecimentos

Agradeço à minha querida mãe pelas palavras de apoio e todo seu carinho.

Ao meu orientador, Prof. Dr. Eduardo Brescansin de Amôres, pela paciência, dedicação e todo o trabalho em me conduzir na realização desse trabalho. Sua orientação foi fator determinante para a finalização do mesmo.

À Profª Drª Michele Fúlvia Angelo e ao Prof. Dr. Angelo Conrado Loula, membros da banca de qualificação, pelas recomendações preciosas que muito colaboraram para a melhoria desse trabalho.

À Universidade Estadual de Feira de Santana (UEFS) e ao Programa de Pós-Graduação em Computação Aplicada (PGCA) pela criação e manutenção do mesmo, oportunizando os profissionais da região.

A todos os professores do PGCA, pelo profissionalismo e presteza que dedicam a esse programa.

À minha querida esposa, Maiara Pereira da Cunha, pela compreensão e apoio e por muitas vezes suportar sozinha as tarefas, que deveriam ser divididas, a fim de que eu pudesse prosseguir nesse trabalho.

Aos meus amigos, colegas de trabalho e familiares, pelo apoio e palavras de incentivo que tanto me ajudaram.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

A compreensão da forma como as grandes estruturas de galáxias, estão distribuídas no Universo é também um fator altamente relevante para o entendimento de como este vem se expandindo, uma vez que as distâncias entre estas estruturas não são aleatórias, como mostram os grandes levantamentos cosmológicos publicados por Yadav (2008).

O estudo do *redshift* (desvio para o vermelho) das galáxias (Seção 2.1) é muito importante para determinar a velocidade de afastamento dessas galáxias, e ainda determinar quais delas formam uma dada estrutura. Dessa forma é possível contribuir entre outros aspectos, no entendimento da expansão e da distribuição das estruturas no Universo Local assim como na identificação de grupos e aglomerados de galáxias.

As galáxias costumam distribuir-se em agrupamentos e pesquisas nesse sentido começaram a serem feitas de forma sistemática a partir das décadas de 1920 e 1930, devido aos trabalhos de E. Hubble (Bierrenbach, 2016). O autor menciona que aproximadamente 60-70% das galáxias, podem ser identificadas em algum tipo de associação, e classifica estas associações da seguinte forma:

- Pares: duas galáxias;
- Grupos: aproximadamente 10 galáxias;
- Aglomerados pobres: aproximadamente 100 galáxias;
- Aglomerados ricos: aproximadamente 1000 galáxias.

Bierrenbach (2016) afirma ainda que os aglomerados ricos são raros, e contém apenas 7% das galáxias. Esses aglomerados são também denominados de estruturas, e a melhor forma de conhecer a sua natureza é identificando-as e estudando seus objetos.

A identificação de estruturas pode ser realizada por meio de levantamentos em grande escala. Tais levantamentos resultam em catálogos digitais de objetos astronômicos, que são utilizados por diversos pesquisadores com variadas finalidades, dentre a qual, a identificação de agrupamentos. Conforme mencionado por Yadav et al. (2005), os primeiros levantamentos sistemáticos da distribuição de galáxias, foram realizados por Shapley e seus colaboradores, em 1938. Atualmente contamos com diversos levantamentos astronômicos, com a finalidade do estudo de *redsfhits*, tais como o 2MRS (2MASS *Redshift Survey*[1]), 2dFGRS (*Two Degree Field Galaxy Redshift Survey*[2]), o

[1] https://www.cfa.harvard.edu/~dfabricant/huchra/2mass/
[2] http://www.2dfgrs.net/

6dFGS (*Six Degree Field instrument Galaxy Survey*[3]) entre outros.

O levantamento 2MRS tem a sua distribuição de objetos apresentada na Figura 1.1, na qual os objetos com *redsfhits* semelhantes são representados pelas mesmas cores. Nota-se ainda, a presença de grandes filamentos formando estruturas. Este levantamento foi produzido pela equipe de John Huchra[4] (Huchra et al., 2012) a partir do levantamento fotométrico no infravermelho próximo, denominado 2MASS (Skrutskie et al., 2006). O 2MRS contém 44.599 galáxias com magnitude limitada a $K_s \leq 11{,}75$, sendo que a média de *redshift* dos objetos é de 0,03, o que equivale a uma distância de aproximadamente, 115 Mpc, indicando objetos mais próximos do que em outros levantamentos, como 2dFGRS e 6dFGS, que veremos posteriormente.

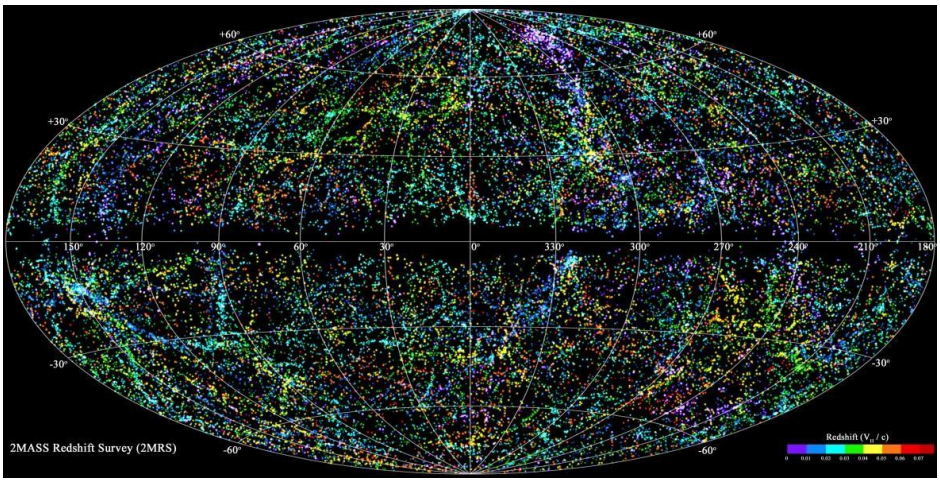

Figura 1.1 – Distribuição em coordenadas Galácticas das galáxias por faixa de *redshift* no levantamento 2MRS. A legenda no canto inferior direito representa os valores de *redshift*, e os grandes filamentos multicoloridos indicam estruturas. Adaptado de IPAC[5].

A legenda no canto inferior direito da Figura 1.1 representa a faixa de *redshift* que vai de 0,01 (roxo) a 0,07 (vermelho). O vermelho escuro indica valores acima de 0,07. Este levantamento teve como base, uma seleção de galáxias no infravermelho próximo do 2MASS, e teve como objetivo principal mapear a distribuição de galáxias e de matéria escura no Universo Local. No decorrer da Dissertação veremos que o mesmo foi muito importante e serviu como base para muitos trabalhos de agrupamento.

A parte central, com distribuição por toda a faixa de longitude Galáctica para valores de latitude Galáctica em módulo aproximadamente inferior a cinco graus, corresponde ao plano de nossa Galáxia (Robin et al., 2003; Amôres et al., 2013, 2017, entre outros), região com um elevado número de estrelas em relação ao restante das demais direções, assim como altos valores para a extinção interestelar, causada devido à poeira interestelar responsável pela redução do brilho dos objetos (Amôres e Lépine, 2005, 2007). Devido a esses fatores, a identificação de galáxias que de uma maneira geral possuem um brilho menor do que as estrelas são de difícil identificação nessas regiões (Amôres et al., 2012).

[3] http://www-wfau.roe.ac.uk/6dFGS/
[4] https://www.cfa.harvard.edu/~dfabricant/huchra/
[5] http://wise2.ipac.caltech.edu/staff/jarrett/2mrs/2MRS.allsky.png.

A Figura 1.2 apresenta a distribuição de objetos no levantamento 2dFGRS[6] (Colless, 2003), que foi conduzido pelo observatório Anglo-Australiano e possui 221.414 galáxias com magnitude no filtro de b_j de 19,45 mag. O levantamento cobriu uma área de aproximadamente 1.500 graus quadrados.

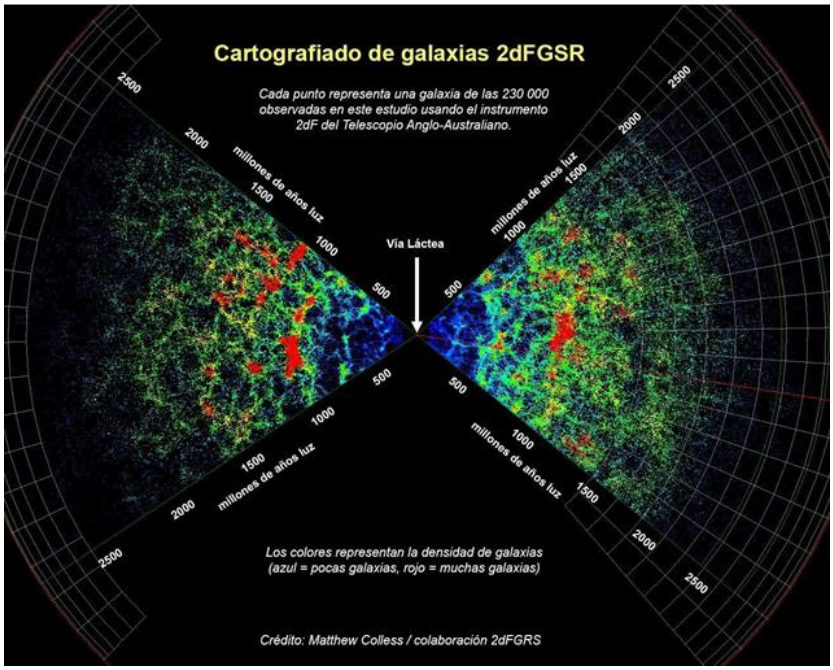

Figura 1.2 – Visão superior da distribuição de galáxias no levantamento 2dFGRS. Cada ponto representa uma galáxia no céu e as cores representam a densidade de galáxias, sendo que o azul, representa poucas galáxias e o vermelho, muitas galáxias.. Adaptado de: Diário de Códoba[7].

Este levantamento cobre uma grande área do Universo Local, incluindo *redshifts* de galáxias mais distantes do que a maioria dos outros levantamentos. O pico na distribuição de *redshifts* se dá em aproximadamente 0,1; conforme pode ser visto no histograma apresentado na Figura 1.3.

[6] http://magnum.anu.edu.au/~TDFgg/
[7] http://www.diariocordoba.com/noticias/zoco/mapeando-cosmos_1152457.html.

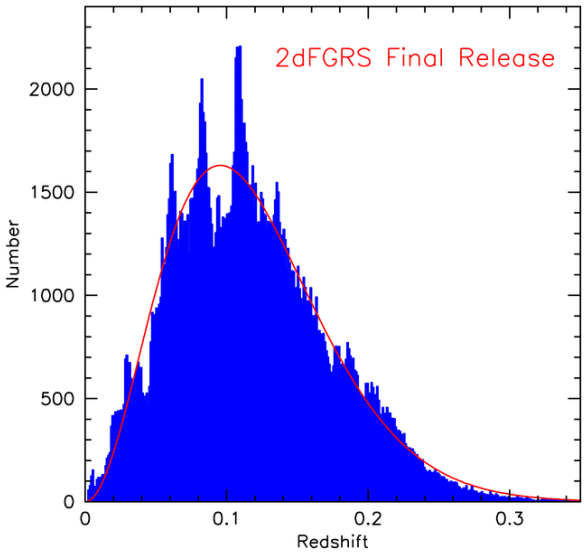

Figura 1.3 – Distribuição de *redshift* no levantamento 2dFGRS. Adapatado de 2dFGRS[8].

Finalmente, o último levantamento apresentado aqui, o 6dFGS (Jones et al., 2009). O nome 6dF refere-se a um dispositivo inteligente que utiliza fibra ótica e tecnologia de posicionamento robótico, aumentando o poder de observação do telescópio *Schmidt* britânico do observatório Anglo-Australiano. Esse levantamento foi produzido por uma equipe de astrônomos do projeto 6d *Galaxy Survey* e tem sua distribuição de objetos representada na Figura 1.4. O mesmo apresenta 125.071 galáxias no céu do Hemisfério Sul, com magnitude limitada a $b_j = 16{,}75$ mag e $K = 12{,}65$ mag e com valor médio de *redshift* em 0,054 (Jones et al., 2009).

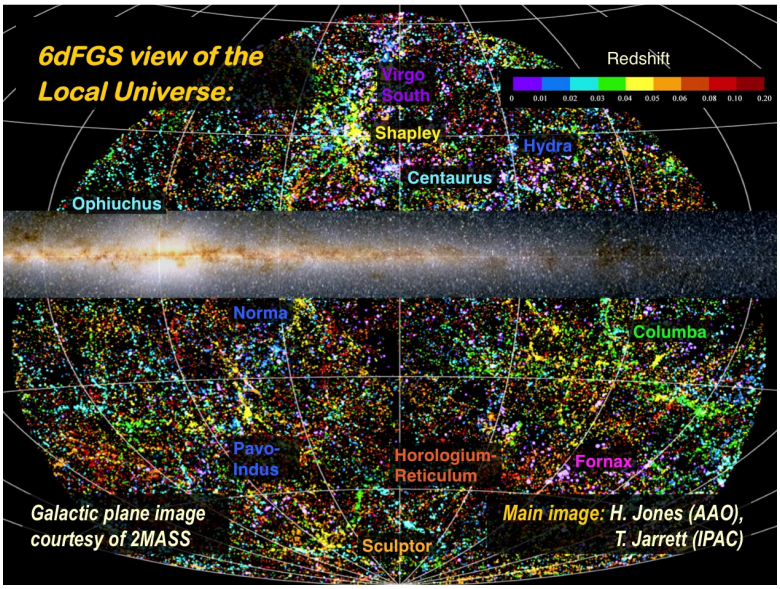

Figura 1.4 – Distribuição em coordenadas Galácticas de galáxias no levantamento 6dFGS. Algumas estruturas são identificadas, tais como Fornax, Norma, dentre outras. Na região central, temos uma imagem em falsa cor do plano de nossa Galáxia. Adaptado de: 6dFGS[9].

[8] http://www.2dfgrs.net/
[9] http://www.6dfgs.net/Gallery/Image_Pages/Slide10_index.html.

A legenda de cores no canto superior direito da Figura 1.4, indica uma faixa de *redshifts* desse levantamento, variando de 0,01 (azul) a 0,10 (vermelho). O vermelho escuro indica valores de *redshifts* acima de 0,20. De acordo com o histograma apresentado na Figura 1.5, podemos verificar que esse levantamento é composto por galáxias mais próximas do que as do Catálogo SDSS (*Sloan Digital Sky Survey*[10]) e 2dFGRS.

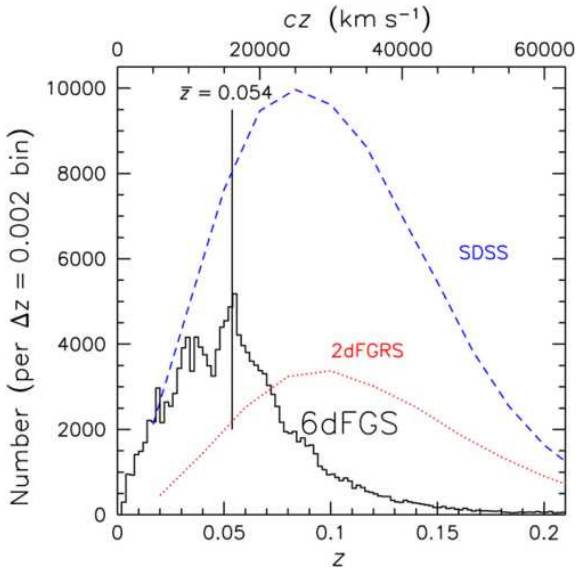

Figura 1.5 – Distribuição de *redshift* no levantamento 6dFGS em relação aos levantamentos 2dFGRS, representado pela linha pontilhada em vermelho, e o SDSS representado pela linha tracejada em azul. Adaptado de 6dFGS[11].

Os levantamentos de galáxias e *redshifts* mencionados até aqui, contribuíram fortemente para os trabalhos de identificação de agrupamentos de galáxias. A seguir, iremos apresentar alguns trabalhos de identificação de grandes estruturas.

A identificação de grandes estruturas, tendo como base os dados de grandes levantamentos de galáxias e *redshifts*, vem sendo estudada por diversos pesquisadores. Burbidge & Burbidge (1961), mencionam o trabalho de Lundmark (1926, 1927 e 1928) e Holmber (1937), que identificaram sistemas contendo mais do que duas galáxias. Outros estudos foram feitos com grupos de galáxias como o de Rood et al. (1970), e Turner (1976), que contribuíram com catálogos de pequenos grupos de galáxias, com densidade menor do que 14 galáxias, e propôs um método de agrupamento em um espaço bidimensional.

Posteriormente, Tully & Fischer (1978) apresentaram novo trabalho e trazem diversas questões sobre o tema, como se existem de fato estruturas em todas as escalas, e se existem galáxias verdadeiramente distribuídas aleatoriamente. Segundo Tully & Fischer (1978), olhando em duas dimensões é possível chegar a conclusões notavelmente diferentes. Ele questiona ainda as condições dinâmicas em quaisquer associações que determinemos existir, se há provas que essas entidades são estáveis.

Huchra & Geller (1982), apresentaram uma técnica para identificar estruturas em um espaço tridimensional, ou seja, as coordenadas dos objetos no céu e seu *redshift*, o que

[10] http://www.sdss.org
[11] http://www.6dfgs.net/Gallery/Image_pages/figure1_index.html.

não era utilizado nos métodos anteriores.

No final do Século XX, dois trabalhos importantes foram desenvolvidos, Ramella et al. (1997), realizaram agrupamento de galáxias no Hemisfério Norte do Catálogo CfA2 (*Center for Astrophysics 2*) utilizando o algoritmo FoF (*Friends of Friends)* (Huchra & Geller, 1982), culminando com a produção do Catálogo de grupos CfAN (CfA *Nothern*). Nesse trabalho, os autores relatam um método para escolha do parâmetro V_0, que consiste na velocidade fiducial, chegando ao valor de 350 km s^{-1} e concluindo que com este valor as propriedades dos grupos são estáveis.

Utilizando uma escolha semelhante de parâmetros, Diaferio et al. (1999), fizeram um trabalho de agrupamento em um catálogo simulado do CfA baseado em simulação de n-corpos. Os autores concluíram que 80% dos sistemas com quatro ou mais membros são sistemas verdadeiramente virializados, ou seja, as galáxias do grupo atingiram seu estado de equilíbrio orbital, enquanto que nos sistemas triplos, apenas 60% o são.

Em trabalhos mais recentes, Ramella et al. (2002) utilizaram os catálogos UZC (*Updated Zwicky Catalog*) e SSRS2 (*Southern Sky Redshift Survey 2*), compilando o catálogo de grupos UZC-SSRS2. O autor aplicou o algoritmo FoF, e obteve um total de 1.168 grupos, sendo que 411 com cinco ou mais membros.

Crook et al. (2007), realizaram agrupamento para um subconjunto do Catálogo 2MRS utilizando o algoritmo FoF. O autor reduziu a amostra do catálogo selecionando galáxias com magnitude $K < 11{,}25$ mag. A validação dos resultados, foi feita por meio da comparação com os catálogos de grupos UZC-SSRS2 (Ramella et al., 2002) e o CfAN (Ramella et al., 1997).

Tempel et al. (2014a) realizaram um trabalho de detecção de rede de filamentos no Catálogo SDSS. Para isto, eles procuraram por filamentos de galáxias em um raio aproximado de 0,5 h^{-1} Mpc utilizando o processo de pontos de objeto com interações, denominado de modelo *Bisous*. Postriormente Tempel et al. (2014b), ainda tomando como base o Catálogo SDSS, proveram um catálogo de grupos de galáxias com fluxo e volume limitados. Para este trabalho usaram o Algoritmo FoF com uma modificação para usar o *LL* nas direções transversal e radial, de forma a identificar grupos de forma mais realística possível.

Tempel et al. (2016) tendo como base catálogos anteriores de *redsfhit*, compilaram um catálogo com aproximadamente 80.000 galáxias, com uma cobertura de 430 Mpc. Os autores utilizaram um método que envolve o algoritmo FoF, e um posterior refinamento com análise multimodal. Para essa tarefa, foi utilizado o pacote *mclust* no ambiente de computação estatística R[12]. Foi fixado um número de subgrupos e usado o Algoritmo EM (*Expectation Maximisation*). O refinamento foi aplicado apenas para grupos com no mínimo sete galáxias, e esta tarefa resultou em um catálogo de 6.285 grupos com dois ou mais membros. Este trabalho foi validado com trabalho realizado por Tully (2015), no qual os autores encontraram mais que dois terços de grupos idênticos em ambos os trabalhos.

[12] https://www.r-project.org/

1.1 Objetivos

1.1.1 Objetivo Principal

Elaborar um método para realizar o agrupamento de galáxias, tendo como parâmetros de entrada suas coordenadas e *redshift*.

1.1.2 Objetivos Específicos

- Desenvolver ferramenta computacional capaz de identificar agrupamentos utilizando o algoritmo selecionado;
- Identificar um método de validação apropriado para verificar a qualidade do agrupamento;
- Validar método tendo com base um Catálogo conhecido;
- Identificar as propriedades e características dos grupos obtidos;
- Identificar grandes estruturas de galáxias no Universo Local.

1.2 Contribuições

As ferramentas computacionais têm grande importância no presente trabalho, seja como meio de armazenamento das informações, de maneira segura e eficiente por meio da utilização de bancos de dados, assim como pelo emprego de técnicas de mineração de dados para agrupamento e validação dos mesmos, por meio do desenvolvimento de aplicações que executem tais agrupamentos de forma rápida e precisa.

Apesar de todos os estudos e trabalhos elaborados nessa área, não encontramos uma ferramenta computacional, de código aberto, que possa ser utilizada por outros pesquisadores com interesse na área. Tal ferramenta, se testada e comprovada a sua eficácia, poderá trazer agilidade e confiabilidade na tarefa de agrupamento. Este trabalho pretende elaborar um método, validando-o e propondo melhorias, a fim de oferecer um resultado ainda melhor.

1.3 Organização do Trabalho

No Capítulo 2 faremos uma breve revisão sobre alguns pontos básicos de Astronomia que são abordados no trabalho, tais como distância e *redsfhit*, técnicas de agrupamento e de validação dos grupos obtidos. No Capítulo 3, apresentamos o conjunto de dados utilizado, que consistiu no catálogo compilado por Tempel et al. (2016), assim como uma breve discussão de suas propriedades básicas. No Capítulo 4, é apresentado o Algoritmo *Friends of Friends* (FoF), o nosso método de agrupamento, metodologia utilizada, assim como a implementação dos algoritmos e a respectiva validação do agrupamento. No Capítulo 5, são apresentados, os resultados obtidos para o agrupamento de galáxias, com a sua distribuição e propriedades básicas. No Capítulo 6, apresentamos as conclusões e perspectivas.

Capítulo 2

Revisão Bibliográfica

No presente capítulo, faremos uma breve revisão sobre alguns tópicos básicos em Astronomia e que são abordados na Dissertação, assim como sobre agrupamentos de dados. Na Seção 2.1, será feita uma breve introdução sobre medida de distância e o *redshift*. O restante do capítulo aborda técnicas de agrupamento conhecidas.

2.1 Distância e *redshift*

Uma noção básica de distâncias em Astronomia, e de como a mesma pode ser determinada, é de extrema importância para entender os resultados obtidos, uma vez que agrupamos as galáxias, tendo como base uma determinada distância. Definimos como Unidade Astronômica (*UA*), a distância média entre a Terra e o Sol, cujo valor é de aproximadamente 149,6 milhões de quilômetros (Oliveira & Saraiva, 2014).

A medida mais conhecida e intuitiva, talvez seja o Ano-Luz (AL). O AL é a distância percorrida pela luz, no vácuo, em um ano, aproximadamente igual a 9,46 x 1.0^{12} Km (Oliveira & Saraiva, 2014). A distância entre a nossa Galáxia, a Via Láctea e galáxia espiral mais próxima, que é Andrômeda é de aproximadamente 2,5 milhões de AL.

Outra unidade de medida, bastante utilizada em Astronomia para medir distâncias, é o parsec (pc). Um parsec, representa a distância, em que se situaria um observador, que observaria a diferença angular entre a Terra e o Sol como sendo de 1" (Oliveira & Saraiva, 2014). Esta medida é conhecida como paralaxe heliocêntrica, e é apresentada na Figura 2.1. Pode-se dizer que um objeto que esteja a 1 parsec de distância, apresenta paralaxe heliocêntrica de 1".

O parsec pode ser convertido em AL e vice-versa, sendo, 1 parsec aproximadamente igual a 3,26 AL (Oliveira & Saraiva, 2014). O parsec possui múltiplos em valores de 10^3, que são: Kiloparsec (kpc), Megaparsec (Mpc) e Gigaparsec (Gpc). Para melhor entendermos estas grandezas, podemos exemplificar que a distância entre a Via-Láctea e Andrômeda é de aproximadamente 0,96 Mpc. O diâmetro de nossa Galáxia, é de aproximadamente 40 kpc (Robin et al., 2003; Amôres et al. 2017), ou seja, aproximadamente 130 mil AL.

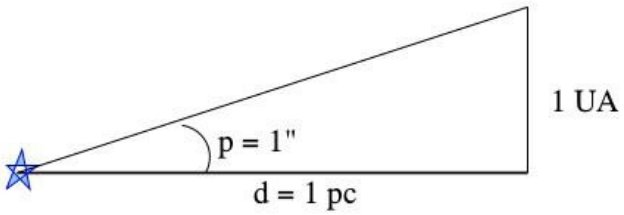

Figura 2.1 – Esquema que representa a paralaxe heliocêntrica, Um observador, representado pela estrela azul ao observar a Terra perceberia uma separação angular de $p = 1$" em relação ao Sol. Neste momento, a distância entre este observador e a Terra seria de 1 pc. Nesta figura, *p* representa a paralaxe, e *UA* a unidade astronômica, distância entre a Terra e o Sol. Adaptado de Oliveira & Saraiva (2014).

O conceito de *redshift* é extremamente importante, pois ele nos permite, dentre outros aspectos, determinar a distância de certos objetos astronômicos, assim como determinar algumas características e propriedades a partir deste.

Uma analogia interessante é a diferença entre o som emitido por um automóvel aproximando-se ou distanciando-se de um observador. Esse efeito é conhecido como efeito Doppler, descrito pelo físico Christian Doppler[13].

A Figura 2.2 apresenta o comprimento de onda que aumenta à medida que a seta se desloca para mais distante do ponto *P1* até o ponto *P4*. Nesse caso, temos um *redshift*. Quando a seta se aproxima, o comprimento de onda diminui, e temos então o *blueshift*.

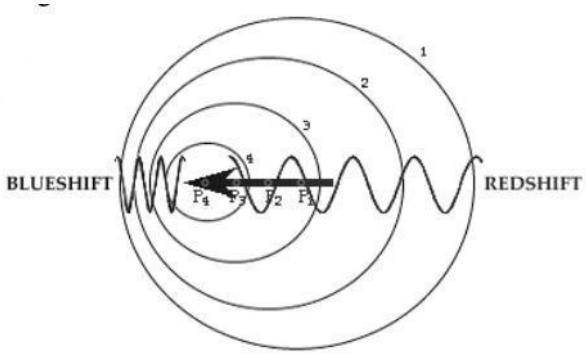

Figura 2.2 – *Redshift*. A seta indica o sentido do movimento, quando nos aproximamos de um objeto, as ondas de luz se comprimem, tendendo ao azul (*blueshift*). Porém, quando nos afastamos, estas ondas se expandem, tendendo ao vermelho. Para o significado dos números, ver o texto. Adaptado de: Oliveira & Saraiva (2014).

O espectro de emissão de uma galáxia fornece, o fluxo em função do comprimento de onda, e por meio da análise de suas linhas espectrais é possível calcular seu *redshift* ou *blueshift*. Quando uma linha espectral observada está em um comprimento maior do que se estivesse na Terra, pode-se concluir que a galáxia está deslocada para o vermelho, assim ela possui um *redshift*, e caso o comprimento seja menor, então a galáxia possui um *blueshift* e está, portanto, se aproximando. O *redshift* é representado nos catálogos pela letra *z*. O *redshift* obtido por meio da análise do espectro é chamado de *redshift* espectroscópico.

Outra técnica utilizada para obter o *redshift* de uma galáxia é a fotometria. O *redshift*

[13] http://cas.sdss.org/dr4/pt/proj/basic/universe/redshift.asp

fotométrico é obtido por meio da fotometria de um objeto em diversos filtros (Bierrenbach, 2016). Esta técnica é menos precisa do que a do *redshift* espectroscópico, porém permite estimar *redshifts* para um grande número de objetos (Bierrenbach, 2016), pois com ela é possível acessar mais objetos, inclusive aqueles com brilho mais fraco. Essa técnica é bastante útil para determinar o *redshift* de objetos mais distantes.

De acordo com a Lei de Hubble, a distância de uma galáxia em relação à Terra está relacionada com sua velocidade de recessão, e pode ser representada pela Equação 2.1[14] (Oliveira & Saraiva, 2014).

$$d = \frac{v}{H_0} \qquad (2.1)$$

Onde d é a distância da galáxia, v é a sua velocidade de recessão, e H_0 uma constante, conhecida como a constante de Hubble. Hubble, apontou que o Universo está em expansão, atribuindo um valor constante para essa taxa de expansão. A unidade de medida desta constante é km/s/Mpc, pois a velocidade das galáxias é medida em km/s e sua distância em Mpc (Megaparsec).

Sendo assim, dividindo a velocidade de recessão pela constante de Hubble, teremos a distância da galáxia em Mpc, de acordo com Equação 2.1. A velocidade de recessão (v) é encontrada por meio da Equação 2.2, que relaciona o produto entre o *redshift* (z) pela velocidade da luz (c).

Alguns trabalhos apresentam diferentes valores para a constante de Hubble (H_0). Crook et al. (2007) e Tempel et al. (2014) utilizaram em seus trabalhos $H_0 = 100\ h$ km s^{-1} Mpc^{-1}, com $h = 0{,}73$. Em outro trabalho, Tempel et al. (2016) utilizaram $H_0 = 67{,}8$ Km s^{-1} Mpc^{-1}. Adotamos o valor usado pelo Grande Levantamento *Planck* (Planck Collaboration XIII, 2016), é $H_0 = 67{,}8$ Km s^{-1} Mpc^{-1}, o qual é bem aceito e de larga utilização.

$$v = zc \qquad (2.2)$$

A obtenção do *redshift* das galáxias é de suma importância para estimarmos as propriedades de um objeto, dentre as quais, a sua distância, a sua localização e relacioná-la com outros objetos, a fim de determinar se pertencem a um mesmo grupo. O processo para determinar o *redshift* de uma galáxia é feito por diversos pesquisadores utilizando diversas técnicas, e o resultado disto é publicado em diversos trabalhos e também em catálogos que são disponibilizados para a comunidade científica. Alguns desses catálogos, os mais representativos para o nosso trabalho foram descritos no Capítulo 1.

2.2 Agrupamento ou Análise de grupos

Entende-se por *Clustering* (agrupamento) ou *Cluster Analysis* (análises de grupo), a tarefa de particionar um grupo de objetos, tendo como base os seus dados, de forma a garantir alta similaridade entre objetos de um mesmo grupo, e a de maior

[14] http://www.telescopionaescola.pro.br/hubble.pdf

dissimilaridade, entre objetos pertencentes a grupos diferentes (Han et al., 2012). No processo de agrupamento, o rótulo de cada grupo é desconhecido, e é descoberto durante o processo que envolve medidas de distância, entre os atributos das instâncias de dados. Por isto, o agrupamento é um processo conhecido como sendo, um aprendizado não supervisionado.

Esse processo é feito utilizando técnicas específicas, as quais podem variar, dependendo do objetivo do agrupamento, a ser efetuado. Técnicas distintas podem resultar em diferentes grupos, mesmo com o uso do mesmo conjunto de dados. Diversos algoritmos são descritos para cada técnica diferente, e são implementados em variadas ferramentas computacionais que permitem a manipulação de grandes conjuntos de dados, com rapidez e acurácia. Na Subseção 2.2.1, serão descritas sucintamente algumas técnicas conhecidas e detalhadas posteriormente nas seções seguintes.

Finalmente, precisamos avaliar a qualidade destes grupos, de forma a validar o processo como um todo, incluindo as tarefas de pré e pós-processamento, bem como as técnicas e algoritmos utilizados. Na Subseção 2.2.2 será discutido os métodos, bem como as medidas de avaliação empregadas.

2.2.1 Técnicas de Agrupamento

Dividir as técnicas de agrupamento em categorias, nos ajuda compreender o processo, e ainda auxilia na escolha adequada de uma ou outra técnica, a depender do que pretendemos com o agrupamento a ser feito. Iremos abordar a seguir três técnicas de agrupamento:

- Métodos de particionamento;
- Métodos hierárquicos;
- Métodos baseados em densidade.

O método de particionamento é o método mais simples de agrupamento (Han et al., 2012), que consiste em dividir um conjunto de dados em k grupos ou clusters, sendo que o valor de k é previamente conhecido, e é o parâmetro fundamental para o funcionamento do método.

A maioria dos métodos baseados em particionamento utiliza uma função de similaridade baseada em distância, e o critério de particionamento é garantir a maior similaridade *intracluster* e maior dissimilaridade *intercluster*.

Os métodos baseados em particionamento possuem algumas desvantagens:

- Trabalham com número fixo de grupos;
- Não são adequados para grupos de formato não globulares.

Existem duas abordagens heurísticas, que visam aperfeiçoar e minimizar a complexidade do método, que são os algoritmos *k-means* e o *k-medoids*. Esses algoritmos são eficientes para identificar *clusters* globulares em pequenos conjuntos de dados. As variações desses algoritmos podem ser usadas para grandes bancos de dados ou para clusters com formato aleatório.

O algoritmo *k-means* é um método de particionamento, que utiliza uma técnica baseada em centróides (Loula, 2015). Um centróide é o ponto médio de um grupo e representará este grupo *Ci* em um conjunto de dados *D*. Definiremos aqui cada centróide por c_i, onde ($1 \leq i \leq k$). Cada objeto *p* do conjunto de dados *D*, pertencerá ao grupo C_i que apresente maior proximidade entre *p* e c_i. Essa proximidade é dada por $dist(p,c_i)$, onde $dist(x,y)$ é a distância euclidiana entre os objetos *x* e *y*.

O funcionamento do algoritmo pode ser descrito da seguinte forma (Han et al., 2012): inicialmente, são calculados os *k* centróides. Há várias formas de calcular os centróides, e a maioria dos algoritmos o faz de forma aleatória. Após definir os centróides, o algoritmo inicia a identificação dos grupos, calculando a proximidade entre um objeto *p* com cada c_i. Ao final, teremos *k* grupos, e neste ponto o algoritmo irá recalcular os centróides tomando como novo centróide o elemento médio de cada grupo formado. A identificação de grupos é então refeita. Este processo é repetido até que os centróides estejam estáveis, de acordo com a Figura 2.3. Nesta figura temos uma fase inicial (a), a iteração (b) e a fase final (c). Na fase inicial os elementos centrais são escolhidos aleatoriamente, nas iterações, novos centros são escolhidos, e na fase final chegamos ao melhor agrupamento, com os objetos centrais determinados.

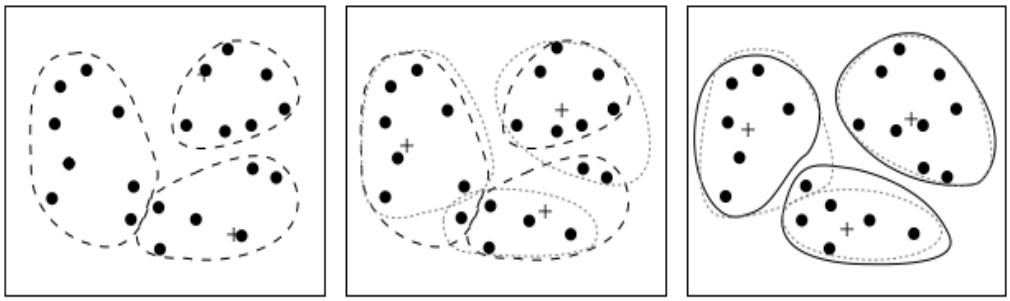

Figura 2.3 – Agrupamento utilizando o algoritmo k-means. Demonstração das três fases do agrupamento. Adaptado de: Han et al., (2012).

Algumas desvantagens *do k-means*, são o fato de não trabalhar com grupos não convexos e agrupamentos com grupos de tamanhos muito distintos, a obrigatoriedade em determinar previamente o *k* (número de grupos), e é sensível a ruídos e *outliers* (Loula, 2015).

O *k-medoids* difere do *k-means* por ter como objeto representativo aquele localizado mais próximo ao centro do grupo, ao invés de calcular o ponto médio. A principal vantagem do *k-medoids*, é ser menos sensível a ruídos e *outliers* (Loula, 2015).

Temos também os métodos hierárquicos, que são assim chamados por apresentarem os grupos organizados como uma árvore hierárquica. Nesse método, o agrupamento possui vários níveis, podendo aumentar a quantidade de grupos, cada vez que aprofundamos na hierarquia.

De acordo com Han et al. (2012), um exemplo interessante é o agrupamento dos funcionários de uma empresa. Inicialmente, teremos um único grupo que seria o topo da hierarquia, esse grupo representa todos os funcionários. Em um nível mais abaixo, teremos uma divisão maior destes grupos, onde estarão diretores, gerentes, *staff*, etc.

Em um nível ainda mais abaixo, podemos subdividir o *staff*, por exemplo, em *seniors*, *masters*, e *juniors*. Uma representação gráfica para o agrupamento hierárquico é o dendograma, apresentado na Figura 2.4. Nesta figura, é apresentado um diagrama em formato de árvore que mostra como os objetos foram agrupados ou separados.

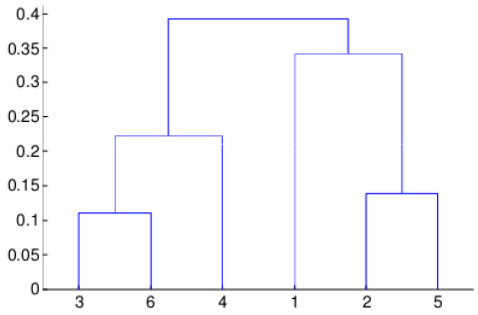

Figura 2.4 – Dendograma. A linha vertical representa a população de cada grupo, e a linha horizontal os diferentes grupos. Este gráfico representa o particionamento em diferentes níveis, onde o nível mais alto do gráfico representa um grupo com população equivalente ao total da amostra. Na medida em que nos aproximamos do eixo horizontal, aumentamos o particionamento e obtemos mais grupos, com a amostra total divida entre eles. Adaptado de: Han et al. (2012).

Efetuando cortes em diferentes níveis do dendograma, obtemos diferentes agrupamentos. Por exemplo, no dendograma ilustrado na Figura 2.5, traçamos dois cortes representados por uma linha em vermelho. O primeiro no nível 0,2 e o segundo no nível 0,35. No primeiro caso obtemos um agrupamento com dois grandes grupos e no segundo caso, obtemos quatro grupos diferentes. Isso mostra como podemos refinar um agrupamento hierárquico, por meio da observação do dendograma.

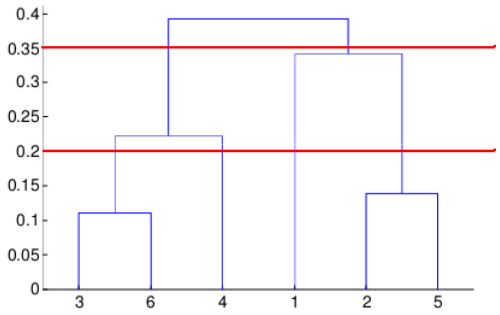

Figura 2.5 – Refinamento nos níveis 0,2 e 0,35. Adaptado de Han et al.(2012).

Segundo Han et al. (2012), algumas desvantagens do método hierárquico são:

- Se dois ou mais grupos forem separados não podem ser juntados novamente;
- Pode ter sensibilidade a ruídos e *outliers;*
- Pode ter dificuldades com grupos de tamanho diferente e convexos;
- Quebra grandes grupos.

Diferente dos métodos vistos até aqui, muitas vezes precisamos identificar os grupos, em regiões onde há uma maior densidade de objetos. O método que veremos a seguir propõe um agrupamento baseado nessa abordagem.

Os métodos baseados em densidade são projetados para encontrar grupos em regiões densas, separados por regiões esparsas, ou de baixa densidade. Dessa forma, os grupos encontrados podem apresentar formas arbitrárias como apresentado na Figura 2.6; diferente dos métodos de particionamento e métodos hierárquicos, que são projetados para encontrar grupos de formato esférico.

Figura 2.6 – Métodos baseados em densidade encontram grupos com formatos aleatórios. Esta figura mostra um grupo com formato em "S" e outro oval, e não apenas com formato esférico como nos métodos de particionamento ou hierárquicos. Adaptado de: Han et al. (2012).

Algumas técnicas são utilizadas para identificação desses grupos, abordaremos aqui o DBSCAN (*Density-Based Clustering Based on Connected Regions with High Density*) (Han et al., 2012).

O DBSCAN funciona identificando áreas de grande densidade, e conectando-as caso haja uma determinada quantidade de objetos em comum entre essas áreas. Para determinar se uma área é considerada densa, e se elas se interconectam, o método precisa de dois parâmetros iniciais, $\varepsilon > 0$, que representa o raio máximo de vizinhança de cada objeto *o*, e *MinPts*, que representa a quantidade mínima de objetos em uma vizinhança de *o* para que uma área seja considerada densa (Han et al., 2012).

Baseado nesses parâmetros, o algoritmo obtém aleatoriamente um objeto *o* de um conjunto de dados *D,* e que será considerado o objeto principal do grupo, também denominado de *core*. A seguir, é feita uma comparação deste objeto *o* com cada um dos outros objetos $p_1..p_n$, do mesmo conjunto, usando como padrão de comparação a medida da distância euclidiana *d* entre o objeto *o* cada outro objeto $p_1...p_n$. Caso $d \leq \varepsilon$, o objeto comparado é inserido no novo grupo.

Cada objeto inserido no grupo é chamado de vizinho, e a quantidade de vizinhos encontrados para *o* podemos chamar de *q*. Ao final, caso $q \geq MinPts$, esta será considerada uma área densa, e teremos um novo grupo. Cada objeto p_n pertencente ao

novo grupo passará a ser então um objeto principal, e sofrerá o mesmo processo submetido ao objeto inicial o. Para cada p_n será computado sua quantidade de objetos agrupados, baseados no raio $\boldsymbol{\varepsilon}$. Caso esta quantidade seja maior ou igual a *MinPts*, uma nova área densa será encontrada e os grupos encontrados por meio de p_n e o serão conectados.

Os objetos agrupados serão marcados como visitados, e o algoritmo deve continuar para cada objeto não visitado de D. No exemplo dado na Figura 2.7, consideremos $\varepsilon = 1$ e $MinPts = 3$. Os pontos o_1, o_2 e o_3 são os objetos principais dos seus respectivos grupos C_1, C_3 e C_4. O grupo C_2 tem como objeto principal o ponto p_2, que também é elemento do grupo C_1. Nesse caso, os grupos C_1 e C_2 irão se conectar formando um único grupo, pois possuem uma quantidade de objetos em comum igual ou superior a *MinPts*. Já os grupos C_3 e C_4 não se conectam, pois não possuem objetos em comum. Por sua vez, os pontos p_{10} e p_{11} são considerados *outliers*, pois não se agrupam.

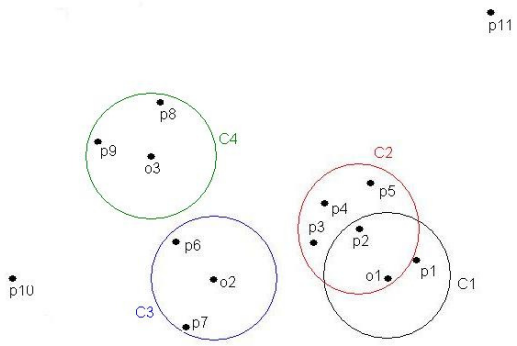

Figura 2.7 – Agrupamento com DBSCAN. Neste agrupamento os grupos *C1* e *C2* unem-se por meio dos pontos *p2*, *p2* e *o1*.

O algoritmo DBSCAN deixa a critério do usuário a escolha dos parâmetros ε e *MinPts*, o que pode ser um problema, pois uma escolha equivocada acarretaria em um agrupamento totalmente diferente (Han et al., 2012). A Tabela 2.1 faz uma relação entre os parâmetros e as características de agrupamento com o DBSCAN.

Tabela 2.1 – Relação entre os parâmetros e os resultados no agrupamento DBSCAN. Esta tabela mostra o que ocorre no agrupamento quando é feito o ajuste nos parâmetros de agrupamento, ε e MinPts.

ε	MinPts	Agrupamento
alto	alto	poucos grupos, grandes e densos
baixo	alto	mais grupos, pequenos e densos
alto	baixo	menos grupos, grandes e pouco densos
baixo	baixo	muitos grupos, pequenos e pouco densos

Uma proposta para esse problema é o método OPTICS(*Ordering Points to Identify the Clustering Structure*). Este método, o qual não é necessariamente um algoritmo de agrupamento, mas se propõe, a produzir uma base de dados ordenada de forma que se possam obter grupos baseados em densidade para uma grande variação de parâmetros (Ankerst et al., 1999). Por meio do OPTICS, podemos encontrar parâmetros mais adequados para identificar os grupos com o DBSCAN.

Como vimos, são vários os métodos de agrupamento, e cada qual pode ser usado de acordo com a área e aplicação. Para que o resultado seja satisfatório, é necessária a obtenção de uma boa fonte de dados e também um trabalho de pré-processamento. Após a tarefa concluída, a avaliação do agrupamento se faz crucial para determinar se houve o sucesso esperado.

2.2.2 Avaliação de Agrupamentos

Uma vez selecionado um método de agrupamento e aplicado o mesmo, é desejável conhecer a eficácia do trabalho realizado. Para isto, é necessária uma avaliação sistemática dos resultados.

De acordo com Han et al. (2012), podemos categorizar os métodos de avaliação em dois grupos, levando em conta se temos ou não um agrupamento ideal, previamente feito por especialistas. Quando dispomos de tal agrupamento ideal, é empregado o método extrínseco, que faz comparações entre o agrupamento obtido e aquele feito pelo especialista. Quando não dispomos do agrupamento ideal, é empregado o método intrínseco, que avalia a qualidade do agrupamento considerando a dissimilaridade entre grupos diferentes, ou seja, o quão melhor seja a separação entre grupos.

A tarefa básica do método extrínseco é associar uma pontuação $Q(C,C_g)$ a um determinado agrupamento, onde C é o agrupamento obtido e C_g o agrupamento ideal (Han et al., 2012). Alguns critérios são importantes para tornar efetiva a medida de Q, os quais destacaremos a seguir:

- Homogeneidade do grupo;
- completude do grupo;
- *rag Bag;*
- preservação de grupos pequenos.

A Homogeneidade do grupo diz respeito à pureza de um determinado grupo no agrupamento. Grupos com maior similaridade entre os seus objetos, de acordo com o grupo ideal a ser comparado, são considerados mais homogêneos e por isso devem receber uma maior pontuação $Q(C,C_g)$.

A completude é a contrapartida da homogeneidade. Objetos que possuam características em comum a dois grupos em um agrupamento, devem pertencer ao mesmo grupo. Dessa forma, os grupos, com os quais, o objeto possua características em comum, devem se unir.

Os *Rag Bag* são grupos, que contém objetos que não possuem semelhanças com outros objetos. Caso em um determinado agrupamento, objetos desta natureza são agrupados em um grupo válido, este agrupamento deve ser penalizado. Agrupamentos que identificam objetos sem características comuns e o colocam em um *Rag Bag*, devem possuir uma maior pontuação.

O critério da preservação de pequenos pedaços aponta que um grupo pequeno não deve ser dividido em pedaços ainda menores. Agrupamentos que dividem grupos maiores e preservam grupos pequenos devem ter uma pontuação maior, segundo este critério.

Segundo Han et al. (2012), muitas medidas de qualidade satisfazem a esses quatro

critérios. As métricas *BCubed Precision* e *recall* são exemplos de medidas que satisfazem os critérios expostos aqui.

Quando não dispomos de um agrupamento ideal realizado previamente para comparar com o nosso agrupamento, é usado o método intrínseco. Este método avalia os quão bem separados e compactos são os grupos em um agrupamento. Uma medida que utiliza este método é o Coeficiente de Silhueta. Calculando este coeficiente para cada objeto de um dado grupo de um agrupamento, podemos descobrir o quão próximo ou distante ele está dos outros objetos do mesmo grupo, verificando assim a qualidade do agrupamento.

De acordo com Han et al. (2012), para calcular o Coeficiente de Silhueta em uma base de dados D com n objetos que foi particionada em k clusters, $C_1,\ldots,C_k$ devemos calcular $a(o)$ e $b(o)$. Em cada objeto $o \in C_i (1 \leq i \leq k)$ calculamos *a(o)* como sendo a média das distâncias entre o e cada outro objeto pertencente ao mesmo cluster, descrita pela Equação 2.3, onde o' é um objeto de C_i diferente de o e *d(o,o')* a distância entre o e o'.

$$a(o) = \frac{d(o,o')}{|C_i| - 1} \{\forall\, o' |\, o' \in C_i; o \neq o'\} \qquad (2.3)$$

O valor de *b(o)* é calculado como sendo o valor mínimo da média entre as distâncias do objeto o a cada objeto o' pertencente a outro cluster no agrupamento. Este valor é calculado de acordo com a Equação 2.4, onde *min()* representa o menor valor, e j cada cluster diferente de i.

$$b(o) = min\left(\frac{d(o,o')}{C_j}\right) \{\forall\, o' |\, o' \in C_j; o \in C_i; 1 \leq j \leq k; j \neq i\} \qquad (2.4)$$

Calculados os valores de *a(o)* e *b(o)*, podemos obter o Coeficiente de Silhueta *s(o)*, como sendo a razão entre a diferença de *b(o)* e *a(o)* pelo valor máximo entre eles. O valor de *s(o)* é apresentado na Equação 2.5, onde *max()* representa o valor máximo entre *b(o) e a(o)*.

$$s(o) = \frac{b(o) - a(o)}{max\big(a(o), b(o)\big)} \qquad (2.5)$$

O valor de *a(o)* indica o quão compacto é um cluster para o objeto o. Quanto menor for esse valor, mais compacto é o cluster. O valor de *b(o)* indica o quão distante é um objeto o dos objetos em outros clusters. Quanto maior for este valor, mais distante está o objeto ***o***.

O Coeficiente de Silhueta pata cada objeto o varia entre -1 e 1, e quanto mais próximo de 1 for este valor, indica que o cluster é compacto, o que é uma situação esperada. Porém, quando valor do coeficiente é negativo, indica que o objeto o está mais próximo de algum objeto de outro cluster que dos objetos do seu próprio cluster, e neste caso devemos analisar a possibilidade de excluir este objeto.

Para medir a qualidade de cada cluster, podemos fazer a média dos coeficientes de silhueta de cada objeto contido nele, e da mesma forma, para medir qualidade de todo agrupamento, podemos fazer a média dos coeficientes de todos os clusters.

Capítulo 3

O Conjunto de Dados

O presente capítulo tem por objetivo descrever e discutir o catálogo utilizado na Dissertação, que contém as bases de dados necessárias para a execução do trabalho, sendo utilizado para a identificação do agrupamento de galáxias assim como para a validação do método que propusemos. Como vimos no Capítulo 1 existem muitos catálogos disponíveis com *redshifts* de galáxias.

O Catálogo *Friend of Friends Galaxy Group Finder*, daqui por diante chamado Catálogo FoF, foi elaborado por Tempel et al. (2016), em uma compilação dos Catálogos 2MRS, CF2 (Tully et al., 2013), e 2M++ (Lavaux & Hudson, 2011), que é uma combinação dos Catálogos 2MRS, 6dFGS e SDSS (Jones et al., 2009). Este catálogo pode ser obtido por meio da ferramenta *Vizier*, do CDS de *Strasbourg*[15]. O catálogo também pode ser obtido diretamente no sítio do pesquisador[16].

A escolha do Catálogo de Tempel et al. (2016), é devido ao fato do mesmo possuir uma grande amostra de objetos, distribuídos por todo o céu, e ser um dos mais recentes. Além de apresentar, o seu método, com toda uma contextualização, assim como o resultado do agrupamento obtido, também disponível em um catálogo, que será usado como validação externa para o nosso trabalho.

Os dados estão dispostos em quatro tabelas, sendo que duas delas com os dados de galáxias com diferentes amostras (tabelas 1 e 2 do Catálogo FoF), e duas com os resultados do agrupamento para ambos os conjuntos de dados (tabelas 3 e 4 do Catálogo FoF), sendo descritas nas Seções 3.1 e 3.2, respectivamente.

No presente trabalho, utilizamos as tabelas 1 e 3 do Catálogo FoF, que serão descritas a seguir. A Tabela 2 que contém os dados de 78.378 galáxias oriundas dos Catálogos CF2 e o 2M++, e a Tabela 4 que contém o agrupamento obtido por Tempel et al. (2016), não serão detalhadas.

Para efeito de cálculos nas propriedades das galáxias e dos agrupamentos, Tempel et al. (2016) utilizaram as constantes cosmológicas, fornecidas nos trabalhos de cosmologia do satélite Planck (Planck Collaboration XIII, 2016). A constante de Hubble tem valor H_0 =67,8 km s^{-1} Mpc^{-1}, a densidade de matéria de Ω_m=0,308 e densidade de matéria escura Ω_Λ=0,692.

[15] ftp://cdsarc.u-strasbg.fr/pub/cats/J/A%2BA/588/A14/
[16] http://cosmodb.to.ee/query

3.1 A tabela de galáxias 2MRS do Catálogo *FoF*

A Tabela 1 do Catálogo FoF, possui 43.480 galáxias do Catálogo 2MRS, todas, fora do plano de nossa Galáxia, ou seja, para latitudes Galácticas, $|b| > 5°$, com magnitude $K_s <$ 11,75 mag, com uma distribuição em distância (*comoving distance*) de até 430 Mpc.

Tabela 3.1 – Estrutura das tabelas de dados 1 e 2 do Catálogo FoF que contém dados de galáxias. As unidades, quando existentes estão entre parênteses.

Número do campo	Nome do campo	Descrição
1	*pgcid*	Identificador do objeto no catálogo de origem
2	*groupid*	Identificador do grupo
3	*ngal*	Número de membros do grupo
4	*groupdist*	"*comoving distance*[17]" para o centro do grupo (Mpc)
5	z_{obs}	*redshift* observado sem a correção CMB[18]
6	z_{CMB}	*redshift* com a correção para o CMB
7	z_{err}	Erro para o *redshift* observado
8	*dist*	*comoving distance* (Mpc). Calculada diretamente do *redshift* com correção CMB
9	*dist_cor*	"comoving distance" da galáxia após supressão do efeito *finger-of-god*
10-11	*RAJ2000, DEJ2000*	Ascensão Reta (*RA*) e Declinação (*DEC*) ambas em graus para equinóxio 2000
12-13	*glon,glat*	Coordenadas Galácticas: longitude e latitude da galáxia (graus)
14-15	*sglon-sglat*	Longitude e latitude supergaláctica (graus)
16-18	XYZ_{sg}	Coordenadas cartesianas supergalácticas tendo como base a dist_cor (Mpc)
19	*mag_*K_s	Magnitude no filtro K_s
20	*Source*	Catálogo de origem da galáxia: 2MRS (1), CF2 (2), 2M++ (3).

Os cálculos das coordenadas cartesianas supergaláctcias X_{sg}, Y_{sg} e Z_{sg}, apresentadas na Tabela 3.1 foram realizados por Tempel et al. (2016) tendo como base a *comoving*

[17] Distância entre dois objetos próximos no Universo, a qual permanece constante independente da época, caso esses objetos estejam se movendo com o fluxo Hubble. Fonte: http://astronomy.swin.edu.au/cosmos/C/Comoving+distance.

[18] *Cosmic Microwave Background*. É a radiação que preenche todo o Universo e pode ser detectado em toda direção. A CMB representa a radiação mais antiga detectada. Fonte: https://www.space.com/20330-cosmic-microwave-background-explained-infographic.html.

distance, e do *redshift* com a correção CMB (número de coluna 6 da Tabela 3.1). As equações usadas são fornecidas por Tempel et al. (2014) e descritas a seguir.

$$X_{sg} = -d_{gal}\sin(DEC) \quad (3.1)$$

$$Y_{sg} = d_{gal}\cos(DEC)\cos(RA) \quad (3.2)$$

$$Z_{sg} = d_{gal}\cos(DEC)\sin(RA) \quad (3.3)$$

nas quais, *RA* e *DEC* são as coordenadas angulares das galáxias fornecidas na Tabela 3.1, e d_{gal} é a *comoving distance* para uma dada galáxia (Tabela 3.1). A seguir, discutiremos algumas características dos dados descritos na Tabela 3.1.

A Figura 3.1 apresenta a distribuição no céu, em coordenadas equatoriais, *RA* e *DEC* das galáxias da Tabela 1 do Catálogo FoF, oriundas do Catálogo 2MRS. Percebe-se uma faixa branca, indicando ausência de galáxias nessa região. Isto representa o plano $|b| \leq 5°$, que foi a faixa excluída por Tempel et al. (2014) em sua seleção, por tratar-se do plano de nossa Galáxia. Conforme mencionado na Introdução, essa região possui uma grande quantidade de estrelas e poeira interestelar, o que dificulta a identificação de objetos com brilho mais fraco.

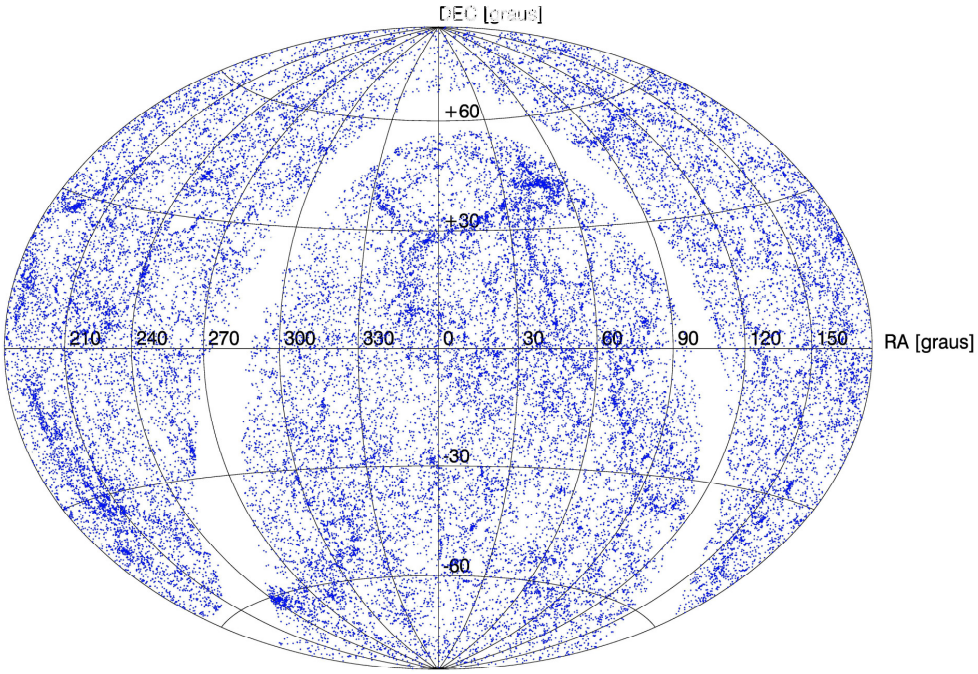

Figura 3.1 – Distribuição em coordenadas equatoriais (*RA*,*DEC*) para as 43.480 galáxias do Catálogo FoF (Tabela 1 dos autores). A região com ausência de galáxias representa o plano Galáctico para latitudes, $|b| \leq 5°$.

A Figura 3.2 apresenta a distribuição das magnitudes observadas das galáxias do Catálogo FoF (Tabela 2 dos autores). Os limites de completeza para os levantamentos

2MRS e 2M++ são 11,75 e 12,50; respectivamente, esse limite permite-nos estimar o quão completa é uma dada amostra.

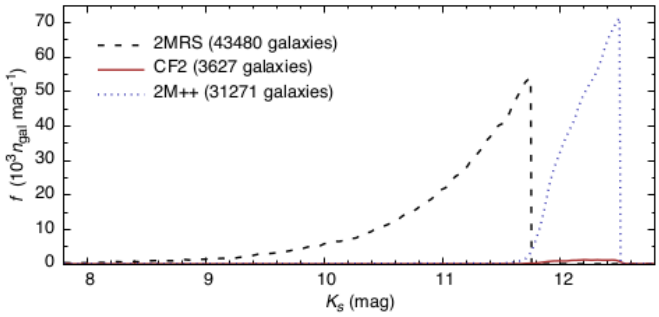

Figura 3.2 – Distribuição do número de galáxias em função da magnitude para o Catálogo FoF (Tabela 2 dos autores) para os três levantamentos do catálogo. Adaptado de: Tempel et al. (2016).

A Figura 3.3 apresenta a luminosidade das galáxias e a contribuição relativa de cada subconjunto de dados 2MRS, CF2 e 2M++, em função da distância. Nota-se que os dados do 2MRS, apresentam um grande número de galáxias próximas, o qual é complementado pelos dados do Catálogo CF2, enquanto os dados do Catálogo 2M++ passam a ser predominantes a distâncias de aproximadamente 200 Mpc. Conforme, apontado por Tempel et al. (2016), quanto maior a fração de galáxias para uma dada distância, melhor será a possibilidade de agrupamento e análise de suas propriedades.

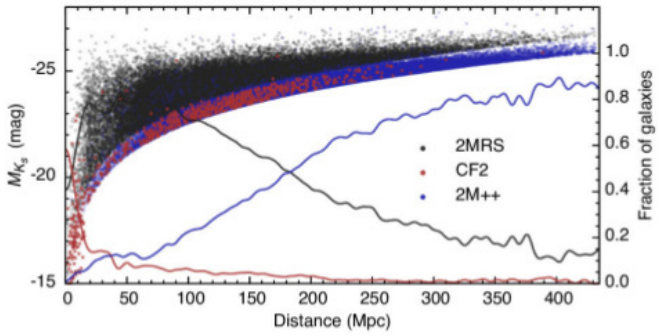

Figura 3.3 – Magnitude absoluta no filtro K_s em função da distância no Catálogo FoF (Tabela 2). As linhas sólidas indicam a quantidade de galáxias nos respectivos catálogos originais. Adaptado de: Tempel et al. (2016).

A maior parte das galáxias da Tabela 1 do Catálogo FoF, possuem *redshift* distribuídas na faixa de $0{,}01 \leq z \leq 0{,}04$ (Figura 3.4), que correspondem à objetos próximos no Universo Local. Como existe uma relação entre *redshift* e distância, a forma do gráfico é aproximadamente parecida, quando consideramos a distribuição da *comoving distance* (Figura 3.5). Nessa figura, nota-se que a maior parte das galáxias está distribuída em uma faixa de *comoving distance*, de 50 até 150 Mpc.

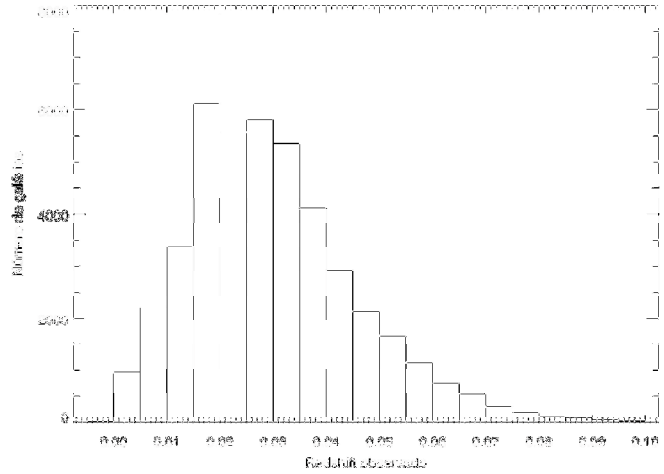

Figura 3.4 – Distribuição do número de galáxias em função do *redshift* observado do Catálogo FoF (Tabela 1 dos autores).

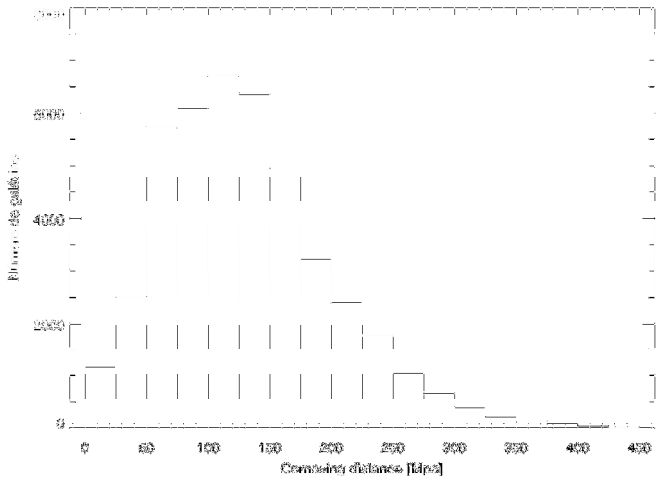

Figura 3.5 – Distribuição do número de galáxias em função da *comoving* distance do Catálogo FoF (Tabela 1 dos autores) .

3.2 A tabela de grupos do 2MRS do Catálogo *FoF*

Os resultados dos agrupamentos obtidos por Tempel et al. (2016) das galáxias de suas

Tabelas 1 e 2, são fornecidos, nas Tabelas 3 e 4, respectivamente dos mesmos autores. Da mesma forma que para as Tabelas 1 e 2, a estrutura das Tabelas 3 e 4 são idênticas e apresentada na Tabela 3.2.

Tabela 3.2 – Estrutura das tabelas de dados 3 e 4 do Catálogo FoF, que contém as informações dos agrupamentos de galáxias, obtido usando as tabelas 1 e 2, respectivamente. As unidades, quando existentes estão entre parênteses. Adaptada de Tempel et al. (2016).

Número do campo	Nome do campo	Descrição
1	*groupid*	Identificador do grupo, idêntico com o fornecido nas tabelas 1 ou 2
2	n_{gal}	Número de membros (galáxias) do grupo
3-4	*RAJ2000, DECJ2000*	Ascensão Reta e Declinação do centro do grupo (em graus)
5-6	*glon, glat*	Longitude e Latitude do centro do grupo (em graus)
7-8	*sglon, sglat*	Longitude e Latitude supergaláctica do centro do grupo (em graus)
9	z_{CMB}	*redshift* do grupo, que é calculado como sendo a média de todos os membros do grupo, já com correção CMB.
10	*groupdist*	"*comoving distance*" para o centro do grupo (em Mpc)
11	*sigma_v*	*rms* na velocidade radial do grupo
12	*sigma_sky*	*rms* das distâncias projectas no céu
13	r_{max}	Distância do centro do grupo para o seu membro mais distante no plano do céu (em Mpc)
14	$mass_{200}$	Massa estimada do grupo, assumindo o perfil de densidade NFW[19] (em unidades de 10^{12} $M_\odot$)
15	r_{200}	Raio da esfera em que a média de densidade do grupo é 200 vezes maior do que a densidade média do Universo
16	*mag_group*	Magnitude observada do grupo.

A Figura 3.6 apresenta a distribuição dos grupos conforme o *redsfhit*. Nota-se, que grupos com dez ou menos objetos, são os mais abundantes. Além disso, a maioria dos grupos, independente da riqueza, ocorre com maior intensidade em uma faixa de *redshifts*, de $0{,}01 \leq z \leq 0{,}04$, o que também pode ser explicado devido ao fato da maioria das galáxias do catálogo estar distribuída nessa faixa.

[19] O perfil Navarro-Frenk-White (NFW) é uma distribuição de massa espacial da matéria escura instalada em halos de matéria escura identificados em simulações de N-corpo por Julio Navarro, Carlos Frenk e Simon White.
(https://en.wikipedia.org/wiki/Navarro%E2%80%93Frenk%E2%80%93White_profile, acessado em 25/07/2017).

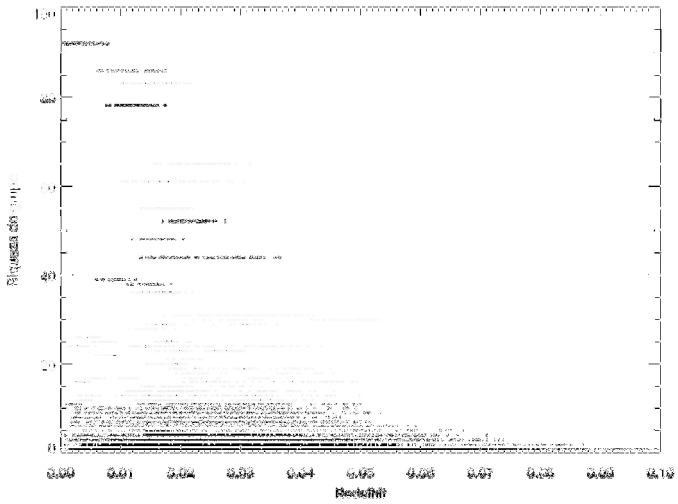

Figura 3.6 – Distribuição da riqueza do grupo de acordo com o *redsfhit*. Adaptado de: Tempel et al. (2016) (Figura 8 dos autores, painel superior).

A Figura 3.7(a) apresenta a distribuição da riqueza para a faixa dos grupos entre dois e nove membros. Podemos verificar mais uma vez, que os sistemas em pares são mais abundantes em relação aos sistemas mais populosos. Aqui, os pares representam aproximadamente 3.700 grupos, mais da metade de todo o agrupamento.

Na Figura 3.7(b) apresentamos a distribuição dos grupos para a faixa de dez até noventa ou mais membros. Nessa faixa, grupos com dez membros ocorrem com maior frequência que grupos maiores, e esse comportamento repete-se praticamente de forma sequencial nos grupos ligeiramente maiores. Esse fato ressalta um comportamento decrescente da quantidade de grupos em relação a sua riqueza, evidenciando que grupos maiores são mais raros. Segundo Bierrenbach (2016), apenas 7% das galáxias conhecidas no Universo encontram-se em super aglomerados, grupos com aproximadamente 1.000 galáxias.

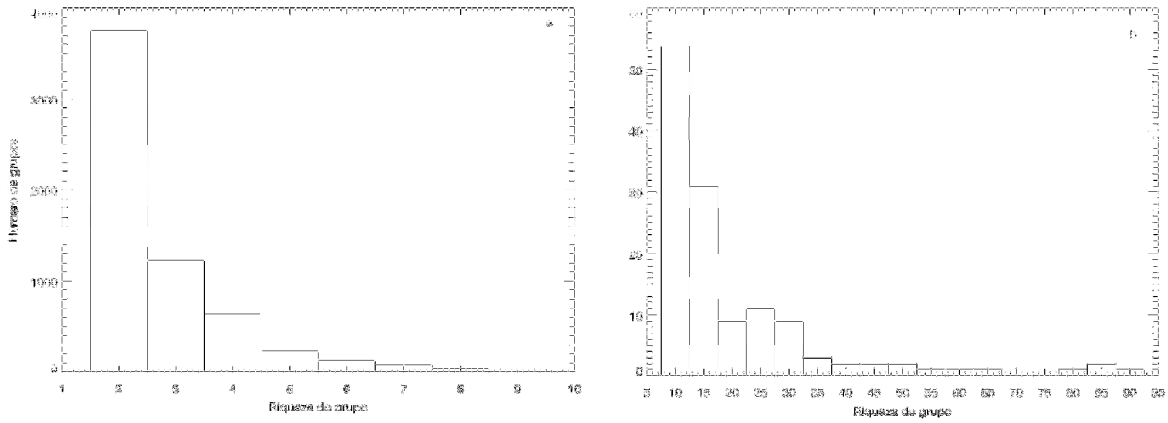

Figura 3.7 – Distribuição da riqueza dos grupos, para duas faixas: a-) grupos com até dez galáxias; b-) grupos de dez até 90 galáxias.

A Figura 3.8 apresenta a distribuição, em coordenadas supergalácticas, dos grupos obtidos por Tempel et al. (2016) e fornecidos em sua Tabela 3, para três faixas de

riqueza do grupo, conforme indicado na legenda da figura.

Em relação aos grupos com maior riqueza de objetos (diamantes em laranja), eles estão distribuídos predominantemente, na direção de grandes estruturas, dentre elas, a Concentração de Shapley, o Superaglomerado de Hydra-Centauros, o Superaglomerado Local e o Aglomerado de Virgo, na direção do anticentro. Na direção do centro, em superlongitudes negativas, nota-se uma distribuição na direção do Superaglomerado Perseus-Peixe. Uma distribuição significativa de grupos, com n_{gal} maior do que dez galáxias pode ser vista na direção do Grande Atrator (Amôres et al., 2012).

Apesar dos grupos com riqueza, na faixa de, $5 \leq n_{gal} < 10$ galáxias, (asteriscos em verde) estarem distribuídas por todo o céu, nota-se uma concentração relativamente maior na direção dos grupos discutidos anteriormente, com $n_{gal} > 10$ galáxias. Os grupos, com riqueza menor do que cinco galáxias são os mais abundantes, e estão distribuídos por todo o céu. A distribuição das estruturas pode ser obtida em diversos trabalhos na área, dentre os quais, no artigo de Courtois et al. (2004) e no sítio do WISE[20], dentre outros.

Na Figura 3.9 apresentamos a distribuição da distância, do membro mais distante de um determinado grupo, em relação ao respectivo centro de seu grupo. Como os grupos possuem formatos irregulares, o centro do grupo não é necessariamente a coordenada supergaláctica da galáxia identificada como *core*, ou seja, aquele a partir do qual são comparados os demais para efeito de agrupamento, e sim a média das coordenadas supergalácticas de todos os membros. Nota-se, que a grande maioria dos grupos possui membros próximos uns aos outros, até aproximadamente 0,8 Mpc. Essa proximidade *intra-cluster* mostra uma boa coesão dos grupos. A distância máxima é de 2,4 Mpc.

[20]http://wise2.ipac.caltech.edu/staff/jarrett/2mass/XSC/chart_SG.jpg

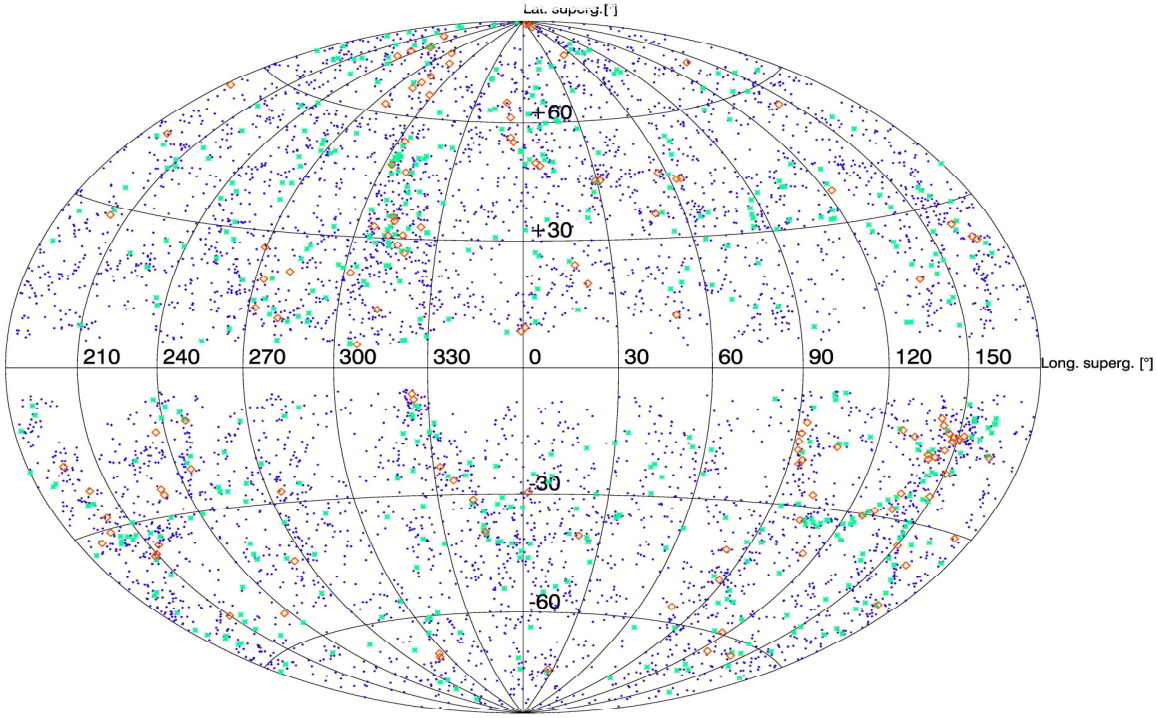

Figura 3.8 – Distribuição em coordenadas supergalácticas dos centros dos grupos referente ao agrupamento das galáxias do Catálogo FoF (Tabela 1) de acordo com a riqueza dos grupos, sendo que os grupos com menos de cinco galáxias (cruzes em azul), grupos com número igual à cinco e menor do que dez galáxias (asteriscos em verde), grupos com dez ou mais galáxias (diamantes em laranja). O tamanho dos símbolos segue três escalas distintas de acordo com número de galáxias por grupo e símbolos descritos acima.

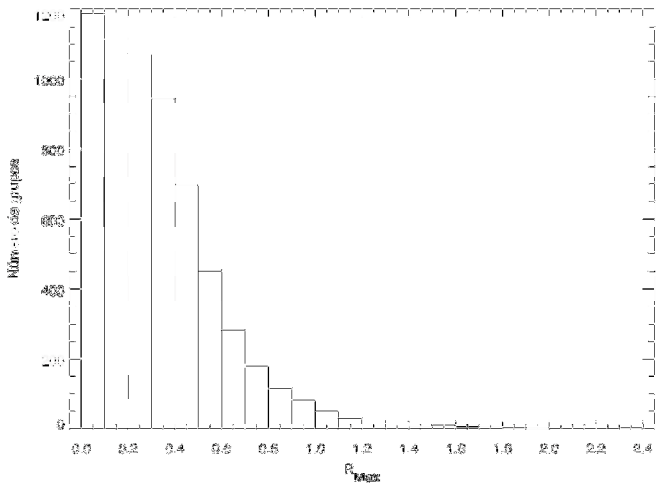

Figura 3.9 – Distribuição da distância máxima, em Mpc, do elemento mais distante do grupo até o respectivo centro de seu grupo, com base nos dados da Tabela 3 do Catálogo FoF.

Podemos fazer uma comparação entre os grupos das tabelas 3 e 4 do Catálogo FoF, de seus valores médios de algumas propriedades, tais como *redshift*, *comoving distance* para o centro do grupo, e membro mais distante. A Tabela 3.3 apresenta essa

comparação. Nota-se, uma leve diferença, de forma com que os grupos da Tabela 4 possuem valores maiores do que os da Tabela 3 desse respectivo catálogo. A maior diferença é do campo *group_dist* com diferença de aproximadamente 20% maior para os grupos da Tabela 4.

Tabela 3.3 – Comparação entre entre algumas propriedades dos grupos das tabelas 3 e 4 do Catálogo FoF. Na qual, valor médio de membros por grupo (n_{gal}), *redshift* (z_{CMB}), *comoving distance* para o centro do grupo (*group_dist*) e distância máxima, em Mpc, do membro mais distante para o centro do grupo.

Médias	n_{gal}	z_{CMB}	group_dist	r_{max}
Tabela 3	3,13	0,03	121,06	0,33
Tabela 4	3,24	0,03	150,03	0,37

Podemos observar que a diferença na quantidade de galáxias entre as tabelas 3 e 4 que representam os grupos, e que são obtidas das tabelas 1 e 2 das galáxias, não interfere, por exemplo, na quantidade média de galáxias (n_{gal}) por grupo. Esse dado está ligado, principalmente ao método empregado (Tempel et al., 2016), e ao *LL* utilizado. Isto porque o método vai dizer como as galáxias serão agrupadas, se haverá união de grupos ou não, e qual o tamanho mínimo de um grupo. Já o *LL*(*Link Lenght*), que será explicado na Seção 4.1, irá definir o raio máximo de abrangência de um grupo, limitando assim o seu tamanho.

Capítulo 4

O Método de Agrupamento

No presente capítulo iremos descrever o Algoritmo *Friends of Friends*, o método empregado na Dissertação, bem como as etapas para a obtenção do agrupamento e validação do mesmo, assim como uma apresentação dos valores do *Linking Length* (*LL*) usado. A base de dados que utilizamos foram as tabelas 1 e 2 do Catálogo FoF. Para verificarmos os resultados submetemos cada grupamento a dois métodos de avaliação, apresentados nas Seções 4.4 e 4.5, respectivamente.

4.1 O Algoritmo FoF (*Friends of Friends*)

O algoritmo *Friends of Friends* proposto por Huchra & Geller (1982), também conhecido por FoF, é considerado o método mais utilizado para identificar grandes estruturas no Universo (Caretta et al., 2008, Huchra & Geller, 1982).

O método consiste na seguinte idéia, considera-se uma esfera de um dado raio, que engloba um objeto pertencente ao conjunto total e que está posicionado ao centro desta esfera. Todos os objetos que estiverem dentro da mesma esfera pertencem ao mesmo subconjunto, e são chamados de amigos. O procedimento se repete para cada objeto considerado amigo, e seguindo a regra "amigos dos amigos", o grupo pode se expandir (Ruiz et al., 2009). O algoritmo se encerra tão logo não haja mais amigos a incorporar. A Figura 4.1 apresenta o fluxograma do algoritmo.

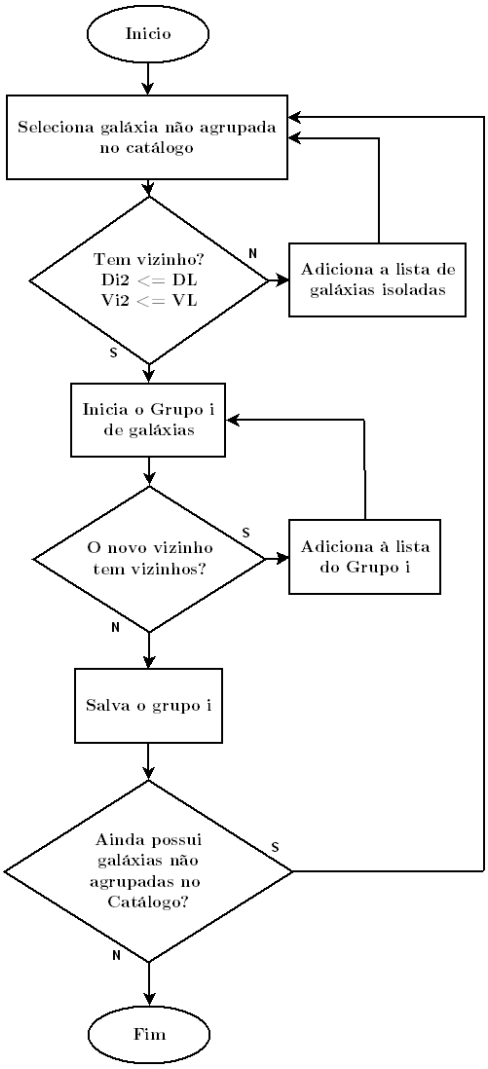

Figura 4.1 – Fuxograma do Algoritmo FoF, no qual podemos perceber na caixa "O novo vizinho tem vizinhos?", a principal característica do algoritmo que é procurar por amigos dos amigos de um mesmo grupo. Adaptado de: Huchra & Geller (1982).

O ponto fundamental do algoritmo é determinar o tamanho do raio de ligação. Huchra & Geller (1982), estabelecem dois parâmetros básicos que são D_L e V_L, no qual, D_L é a distância angular e V_L a velocidade de recessão, Dessa forma para verificar se um objeto é amigo, devemos comparar inicialmente, a sua distância angular e a sua velocidade de recessão. Ambos devem ser menor ou igual aos valores comparados para serem considerados amigos.

Os parâmetros de entrada do FoF, D_L e V_L são determinados pelo usuário, e o sucesso do agrupamento é sensível a seus valores. Farrens et al. (2011) propõe uma obtenção do *LL* de forma dinâmica, baseado na densidade de objetos para cada faixa de *redshift* no catálogo. Os autores denominaram o algoritmo de *Dynamic* FoF, ou DFoF. A possibilidade da escolha do parâmetro de entrada *LL*, traz uma semelhança com o método de agrupamento por densidade descrito no Capítulo 2.

4.2 A Aplicação

A aplicação proposta neste trabalho é baseada no Algoritmo FoF, descrito na seção anterior, e utiliza como entrada um arquivo que deve conter as seguintes propriedades, coordenadas equatoriais (*RA*, *DEC*) do objeto e seu *redshift* (*z*). Para o cálculo da distância, entre duas ou mais galáxias, precisamos determinar as suas respectivas coordenadas cartesianas X_{sg}, Y_{sg}, Z_{sg}, tomando como centro o Sol.

Um raio de percolação, também conhecido como *LL* (*Linking Length*), é previamente estabelecido, e a aplicação irá verificar cada galáxia e agrupar com todas as outras que estejam dentro do mesmo raio. O fluxograma da Figura 4.2 ilustra o funcionamento da aplicação.

Inicialmente o catálogo é carregado a partir de um arquivo de entrada. Posteriormente, percorre-se toda a lista carregada previamente e calcula-se a distância de cada galáxia com as galáxias subseqüentes. Caso a distância. entre uma galáxia e outra esteja dentro do raio de percolação previamente estabelecido, o agrupamento será feito. A medida que cada galáxia é agrupada é atribuída à mesma um identificador de grupo, que é um número seqüencial, e caso a galáxia comparada já faça parte de outro grupo, todos seus componentes passarão a fazer parte do novo grupo identificado, e seu identificador será alterado, seguindo a lógica do FoF.

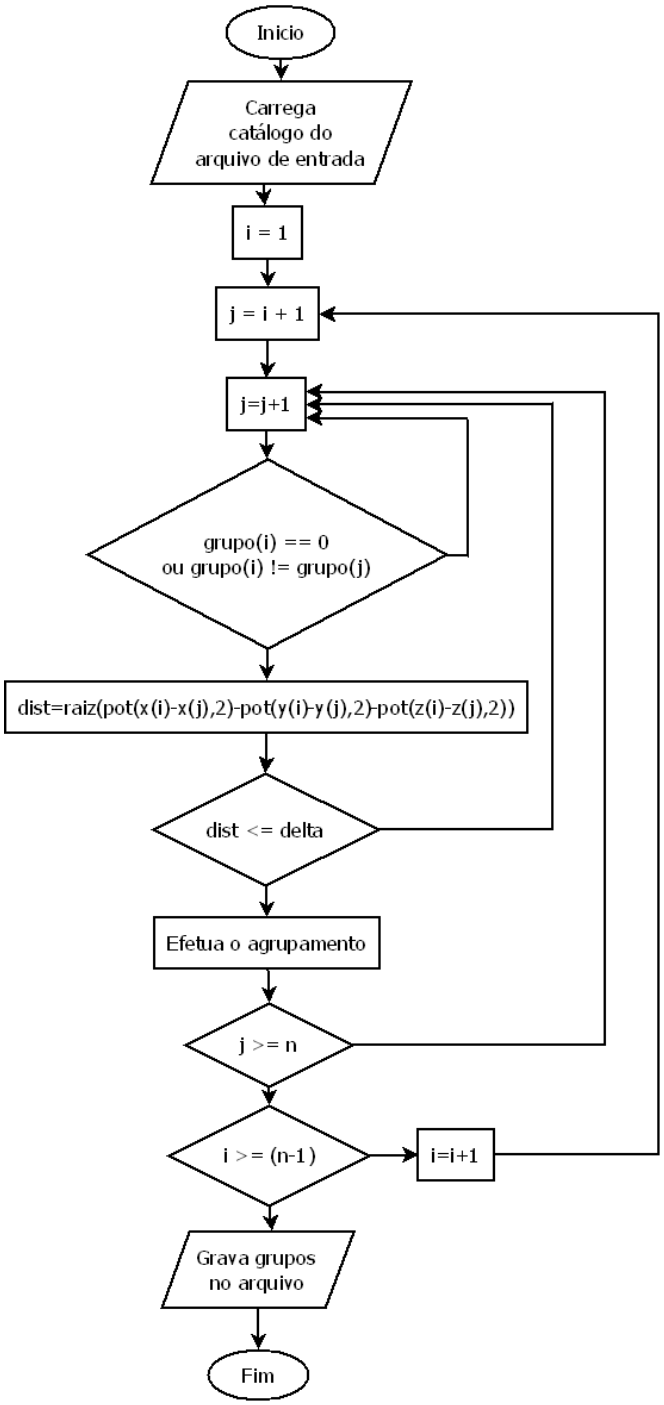

Figura 4.2 – Fluxograma da aplicação de agrupamento.

Para obter as coordenadas cartesianas de cada galáxia, consideramos seus valores de *RA* e *DEC* como sendo ângulos diretores e a distância da galáxia até nós, como sendo a norma do vetor. A distância da galáxia, que chamamos de *d*, pode ser obtida por meio da Lei de Hubble, e é representada pela Equação 4.1. Onde v é a velocidade de expansão da galáxia, e H_0 é a constante de Hubble, conforme mencionado no Capítulo 2. A velocidade de expansão v pode ser obtida por meio do seu *redshift*, e é representada pela Equação 4.2, na qual c é a velocidade da luz, e z o *redshift* que é fornecido no catálogo.

$$d = \frac{v}{H_0} \quad (4.1)$$

$$v = cz \quad (4.2)$$

Para calcular as coordenadas cartesianas utilizamos o método de transformação de coordenadas polares em coordenadas cartesianas em um sistema de coordenadas tridimensionais (Nadal, 2011). Consideramos o Sol como o centro das coordenadas.

A Figura 4.3 apresenta o plano cartesiano XYZ, onde *O* é o centro, *OP* é a distância da galáxia representada pela letra *d*, *d'* é a projeção do vetor *OP* no plano XY, β é a Declinação (*DEC*) e está localizado entre *OP'* e *d*, α é o complemento da Ascensão Reta ($360° - RA$) e está localizado entre *OP'* e *OY*, *p* o objeto observado e *p'* sua projeção no plano *XY*. As coordenadas cartesianas do ponto *p* são representadas por X_{sg}, Y_{sg}, Z_{sg}.

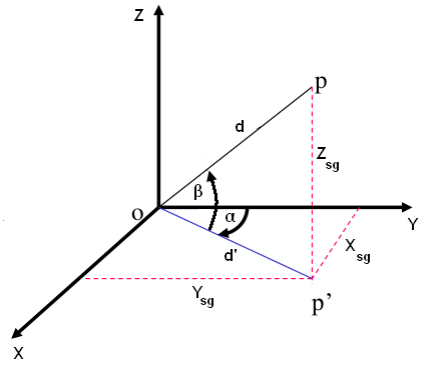

Figura 4.3 – Plano cartesiano *XYZ*. Onde *O* representa o centro, *d* representa a distância da galáxia, enquanto d' sua projeção no plano *XY*. Os ângulos *α* e *β* que se localizam entre *OP'* e *OY*, e *OP* e *OZ* respectivamente representam o complemento da *RA* e da *DEC*. Adaptado de: Nadal (2011).

Sabendo que a Declinação (*DEC*) varia de 0° a 90° para o Hemisfério Norte, e -90° a 0^0 para o Hemisfério Sul, e é representado por *β* obtemos a coordenada Z_{sg} de acordo com a Equação 4.3:

$$Z_{sg} = d\,\sin(\beta) \quad (4.3)$$

A projeção do vetor *d* no plano *XY* pode ser representada por *d'*, de acordo com a Figura 4.3. Para obter os pontos X_{sg} e Z_{sg}, precisamos do valor de *d'* e o valor de α, que é calculado como sendo o complemento da *RA*, uma vez que esta cresce em sentido anti-horário, partindo de *OY* para *OP'*. Tomando o plano *Zd'*, podemos representar *d'* pela Equação 4.4:

$$d' = d\,\cos(\beta) \quad (4.4)$$

Sendo assim, no plano *XY*, X_{sg} e Y_{sg} podem ser obtidos pelas Equações 4.5 e 4.6:

$$X_{sg} = d' \cos(\alpha) \quad (4.5)$$

$$Y_{sg} = d' \sin(\alpha) \quad (4.6)$$

Nas quais, o subscrito, *sg*, representa a abreviação para coordenadas supergalácticas. Substituindo *d'*, temos as Equações 4.7 e 4.8:

$$X_{sg} = d \cos(\beta) \cos(\alpha) \quad (4.7)$$

$$Y_{sg} = d \cos(\beta) \sin(\alpha) \quad (4.8)$$

Após os cálculos, a aplicação, gerará dois arquivos, o primeiro com informações a respeito de cada galáxia, um código identificador, a que grupo está relacionado, a quantidade de vizinhos, sua localização, seu *redshift*, sua distância em Mpc ao centro do grupo, suas coordenadas supergalácticas, e o código do objeto que representa o núcleo do seu grupo.

O segundo arquivo é um totalizador que armazena dados relativos aos grupos. O identificador de cada grupo, a quantidade de galáxias agrupadas nele, a distância média das galáxias ao centro do grupo, a distância da galáxia mais afastada, e as coordenadas supergalácticas do centro do grupo. Os campos desses arquivos estão descritos nas Tabela 4.1 eTabela 4.2, respectivamente.

Tabela 4.1 – Descrição dos campos do primeiro arquivo gerado pela aplicação. A primeira coluna da tabela informa a posição seqüencial do registro.

Número do campo	Nome do campo	Descrição
1	*Idobj*	Identificador do objeto. Número sequencial do objeto de acordo com sua posição no catálogo de origem
2	*Idgrp*	Identificado do grupo a qual pertence o objeto. Quando for zero indica que o objeto não está agrupado
3	*Qtgal*	Quantidade de galáxias no grupo
4	*Vizinhos*	Quando o objeto for considerado "core", este campo corresponde à quantidade de vizinhos que ele tem
5-6	*RA, DEC*	Coordenadas equatoriais *RA* e *DEC* da galáxia
7	*Red*	*Redhift* obtido no Catálogo FoF
8	*Distc*	Distância da galáxia ao centro do grupo (Mpc)
9-11	X_{SG}, Y_{SG} e Z_{SG}	Coordenadas supergalácticas da galáxia
12	*Core*	*Idobj* da galáxia tomada como referência para agrupamento. Em alguns casos, em um mesmo grupo duas ou mais galáxias podem ser "core", isto ocorre quando há união entre grupos.

Tabela 4.2 – Descrição dos campos no segundo arquivo gerado pela aplicação desenvolvida em nosso trabalho.

Número do campo	Nome do campo	Descrição
1	*Grupo*	Identificador do grupo
2	*Ngal*	Quantidade de galáxias no grupo
3	*Mdistc*	Média das distâncias (Mpc) de cada objeto ao centro do grupo
4	*Maxdistc*	Distância do objeto mais distante do grupo (Mpc)
5-7	X_C, Y_C, Z_C	Coordenadas cartesianas do centro do grupo.

A aplicação foi desenvolvida em Linguagem C usando o compilador G++ versão 6.4 no sistema operacional Debian 8. A plataforma de *hardware* foi um *notebook* HP *Pavillon*, Processador Intel i3 com 4 Gb de memória RAM. O tempo total de execução para agrupamento em um catálogo de 43.480 galáxias foi de 1 minuto e 22 segundos, considerando a quantidade de no mínimo quatro galáxias em comum para fusão entre grupos.

4.3 Os valores para o *Linking Length* (*LL*)

Para validarmos a nossa aplicação, utilizamos a Tabela 1 do Catálogo FoF descrita no capítulo anterior. Tendo como propósito, obter o valor de *LL* que em comparação com essa tabela, nos fornecesse o melhor valor para similaridade, fizemos seis simulações utilizando diferentes valores para o *LL* e limitamos em, no mínimo, quatro objetos em comum o critério para unir grupos diferentes, tendo em vista que sistemas com menos de quatro objetos possuem menos de 60% de galáxias considerados estáveis (Diaferio et al., 1999). A escolha dos valores para os *LLs* baseou-se no valor utilizado em Tempel et al. (2014a) e Tempel et al. (2016).

Escolhemos, como nosso menor valor de *LL*, o maior valor apresentado em Tempel et al. (2014). Na referida tabela, os autores apresentam, valores de *LL* variando, de aproximadamente 0,38 até 0,58, de acordo com diferentes amostras de volume, limitadas pela magnitude e *redsfhit*. O último valor de *redsfhit* na tabela dos autores, é o que está mais de acordo com a amostra utilizada no presente trabalho, e por essa razão adotamos esse valor como ponto de partida, para alguns testes, apresentados na Seção 4.4.

Utilizamos os seguintes valores, 0,5; 0,68; 0,85; 1.00; 1,50 e 2,30 h^{-1} Mpc. O maior valor da faixa de *LL* selecionado representa o objeto com maior distância para o centro do seu grupo encontrado na Tabela 1 do Catálogo FoF. Para fins de cálculo, os valores aqui apresentados em unidade Mpc h^{-1} foram convertidos para Mpc, sendo h[21] um parâmetro que varia entre $0{,}5 < h < 0{,}75$ e calculado de acordo com a Equação 4.9.

$$h = \frac{H_0}{100\,\frac{km}{s}/Mpc} \qquad (4.9)$$

na qual H_0 é a constante de Hubble. Os valores das constantes cosmológicas utilizados no presente trabalho são apresentados no início do Capítulo 3. Convertemos os valores para ficarem em unidades de Mpc, dessa forma, os valores de *LL* são 0,74; 1,00; 1,25; 1,48; 2,21 e 3,39 Mpc.

4.4 Avaliação do agrupamento pelo método extrínseco

Conforme visto no Capítulo 2, existem dois métodos para análise do agrupamento. O método extrínseco, quando temos um agrupamento prévio feito por um especialista, o qual pode ser usado para validar o nosso agrupamento, e o método intrínseco, no qual avaliamos a qualidade do agrupamento considerando a dissimilaridade entre os grupos. Nessa seção apresentamos a comparação, entre os resultados obtidos por nosso Método com vários valores de *LL* (mencionados na seção anterior), tendo como base a Tabela 1 do Catálogo FoF, com o agrupamento fornecido por Tempel et al. (2016), determinando dessa forma, um índice de similaridade entre o nosso agrupamento e o proposto pelos autores.

Para validar o nosso método realizamos seis simulações conforme mencionado na seção

[21] https://en.wikipedia.org/wiki/Parsec

anterior, e fizemos comparações tomando como referência os agrupamentos fornecidos por Tempel et al. (2016), levando em conta a similaridade entre os agrupamentos. Para determinar se um dos nossos grupos é similar a um grupo dos autores, é requerido que 80% ou mais dos objetos em nosso grupo sejam iguais ao do grupo correspondente no agrupamento que estamos comparando (Tempel et al., 2016), o qual pode ser calculado de acordo com a Equação 4.10.

$$N[G_{método} \cap G_{FoF}] \geq \|0{,}8([G_{método} \cup G_{FoF}])\| \qquad (4.10)$$

Na Equação, $N[G_{método}/G_{FoF}]$ é a quantidade de objetos no grupo $G_{método}/_{FoF}$. Em outras palavras, verificamos, se a quantidade de objetos resultante da interseção entre um grupo em nosso agrupamento, e um grupo no Catálogo FoF, é maior ou igual que 80% do total de objetos resultante da união entre estes dois grupos. Efetuamos os cálculos para cada objeto em cada agrupamento, para um dado valor de *LL*.

Os resultados das simulações com os diferentes valores de similaridade, entre o nosso agrupamento, com o obtido por Tempel et al. (2016) e disponíveis no Catálogo FoF (Tabela 3.1) utilizando os critérios descritos acima, são apresentados na Tabela 4.3.

Tabela 4.3 – Similaridade entre o nosso agrupamento e o obtido por Tempel et al. (2016). Para detalhes sobre a definição dos valores do *LL* (*Linking length*), ver o texto. Na coluna galáxias agrupadas são apresentadas as quantidades de galáxias agrupadas, assim como o % em relação ao número total de galáxias do Catálogo FoF, a coluna similaridade significa o percentual de similaridade entre nosso método e o método comparado em Tempel et al. (2016), e a coluna Idênticos representa percentual de grupos idênticos ao grupo correspondente no Catálogo FoF.

LL	*Galáxias agrupadas*	*Similaridade* (%)	*Idênticos* (Sim. 100%)
0,74	12.939 (30%)	62,95%	59,87%
1,00	16.573 (38,11%)	76,64%	73,18%
1,25	19.424 (44,67%)	79,00%	75,74%
1,48	21.312 (49%)	75,57%	72,36%
2,21	25.071 (58%)	56,65%	53,57%
3,39	29.883 (69%)	34,96%	32,34%

De acordo com a Tabela 4.3, podemos perceber que o valor do $LL = 1{,}25$, resultou em grupos com maior similaridade em relação ao agrupamento feito no Catálogo FoF, pois além de gerar 79% de similaridade, obteve mais de dois terços de grupos idênticos (100% de similaridade), apresentado na quarta coluna da Tabela 4.3. Outro aspecto consiste no fato de que, esse valor de *LL*, foi o que possibilitou agrupar a quantidade mais próxima de objetos ao agrupado por Tempel et al. (2016).

No trabalho de Tempel et al. (2016) foram agrupadas 19.636 galáxias, enquanto, que em nosso agrupamento com o $LL = 1{,}25$ foram agrupadas 19.424. Esta diferença pode ser explicada pelo fato de que, em nosso agrupamento, termos utilizado um valor mais alto para o *LL*.

O número similar de galáxias encontrado entre nosso agrupamento, e o utilizado para comparação, representa um bom resultado. Identificamos que mais de dois terços dos nossos grupos são grupos idênticos ao do agrupamento de referência, e o restante é similar. Percebemos que quanto maior for o valor de *LL*, maior será o percentual de galáxias agrupadas, porém isto reduz bastante a similaridade em comparação à Tempel et al. (2016), diminuindo assim a qualidade do agrupamento.

Para a melhor compreensão da Tabela 4.3 elaboramos a Figura 4.4, que apresenta a distribuição da percentagem de galáxias agrupadas (segunda coluna da Tabela 4.3) em relação ao total de galáxias da Tabela 1 do Catálogo FoF, assim como a similaridade (terceira coluna da Tabela 4.3) em comparação ao agrupamento feito por Tempel et al. (2016) para cada valor de *LL*. Nota-se, que existe uma tendência de crescimento da quantidade de objetos por grupo à medida que aumentamos o valor do *LL*. Isso se deve ao fato, de que quanto maior for o *LL*, maior será o raio máximo de separação entre as galáxias, de modo que sejam consideradas galáxias de um mesmo grupo.

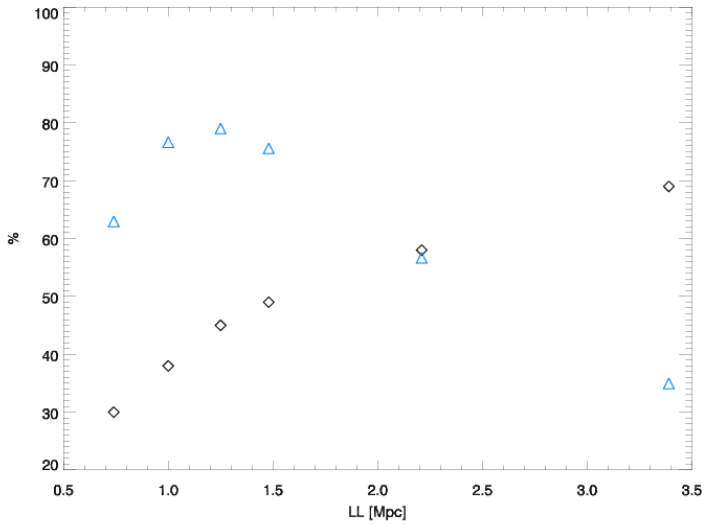

Figura 4.4 – Distribuição da percentagem versus *LL* [Mpc]: diamantes representam a % de agrupamento em relação ao número total de galáxias (Tabela 1 do Catálogo FoF), triângulos representam a % de similaridade em relação aos resultados obtidos por Tempel et al. (2016). Os valores são baseados na Tabela 4.3.

Em relação à similaridade, existe uma tendência decrescente à medida que aumentamos o valor do *LL*. Valores de *LL* muito altos acarretam grupos muito populosos. À medida que diminuímos o valor de *LL*, chegamos a uma quantidade maior de grupos similares, no entanto, existe um valor mais adequado, pois valores de *LL* muito baixos geram grupos muito pequenos, diminuindo a similaridade, como podemos verificar no gráfico para valores onde $0{,}5 \leq LL \leq 1{,}0$. Dessa forma, nota-se, uma relação inversa entre quantidade de galáxias agrupadas e o índice de similaridade, levando em conta o valor do *LL*.

A Tabela 4.4 apresenta os valores médios para algumas propriedades dos agrupamentos para seis valores distintos de *LL*. Nota-se que a medida que aumentamos o *LL*, existe um pequeno aumento na quantidade de objetos por grupo. Esse parâmetro é fundamental para o cálculo da distância euclidiana entre os objetos, e irá definir a quantidade de objetos em cada grupo, pois quanto maior for o *LL*, maior será o raio de abrangência e mais objetos serão incorporados a um mesmo grupo.

Tabela 4.4 – Valores obtidos para $<n_{gal}>$ (aqui, a média da quantidade de galáxias por grupo), *mdistc* (média das distâncias dos objetos ao centro do grupo), *mmaxdistc* (média dos objetos mais distantes) e *maxdistc* (maior distância encontrado entre um objeto e o centro do seu grupo), para diferentes valores de *LL*.

LL	$<n_{gal}>$	*mdistc* (Mpc)	*mmaxdistc* (Mpc)	*maxdistc* (Mpc)
0,74	3,26	0,28	0,34	2,19
1,00	3,31	0,36	0,44	2,89
1,25	3,32	0,43	0,53	3,16
1,48	3,34	0,49	0,61	4,81
2,21	3,62	0,69	0,91	6,78
3,39	4,34	1,12	1,69	16,62

Outro aspecto, que pode ser notado, é que a maior distância encontrada entre um objeto e o centro do seu grupo, representada na coluna *maxdistc*, excede o valor do *LL*. Isto pode ser explicado pelo fato dos grupos possuírem formato irregular, deslocando o centro do grupo.

Outro aspecto é que, durante o agrupamento pode ocorrer a união entre dois ou mais grupos, nesse caso, o novo grupo gerado terá o seu centro recalculado, fazendo com que os objetos fiquem ainda mais dispersos. No entanto, a média entre os objetos mais distantes, representado pela coluna *mmaxdistc*, apresenta um valor baixo, inferior ao próprio *LL*, indicando uma boa proximidade *intracluster*, o que pode significar boa coesão dos grupos.

Verificamos também, o comportamento da similaridade (Tabela 4.5) em relação à quantidade de objetos correspondentes por grupo, para $LL = 1{,}25$.

De acordo com a Tabela 4.5 mais da metade dos grupos similares, são encontrados, entre aqueles com quantidade pequena de objetos. Essa característica também é encontrada no agrupamento realizado por Tempel et al. (2016), que apresenta uma grande quantidade de grupos com três ou menos objetos. Ainda assim, foi possível encontrar similaridade, e até mesmo igualdade, em grupos com grande quantidade de objetos.

Tabela 4.5 – Comportamento da similaridade em relação a quantidade de objetos correspondentes por grupo para $LL = 1{,}25$. A primeira coluna representa a quantidade de objetos correspondentes, a segunda a quantidade de grupos que apresentaram similaridade para esta correspondência, e a terceira o percentual que representa em todos os grupos similares.

Correspondências	N_{grupos}	%
2	2.659	57,48%
3	967	20,90%
4 ou mais	999	21,60%

O comportamento apresentado na Tabela 4.5 pode também ser visto na Figura 4.5, onde nota-se que há uma maior concentração de grupos com a mesma quantidade de galáxias, entre aqueles com menos de dez objetos. Nessa figura pode-se notar os grupos similares para cada valor de LL. Quanto mais próximo da simetria entre os dois eixos estiver o símbolo, maior é sua similaridade entre os dois trabalhos.

Nota-se que os símbolos representados pelos diamantes azuis, que correspondem ao $LL = 1{,}25$, concentram-se próximos a uma região simétrica entre os dois eixos. Isso corrobora com o fato do agrupamento com o valor de $LL = 1{,}25$ ter alcançado o maior percentual de grupos idênticos, 75,74% (Tabela 4.3). Nota-se ainda que os símbolos dos agrupamentos com $LL \geq 2{,}12$ são raros para grupos com mais de 30 galáxias, indicando que com esses valores obtemos baixa similaridade em grupos grandes.

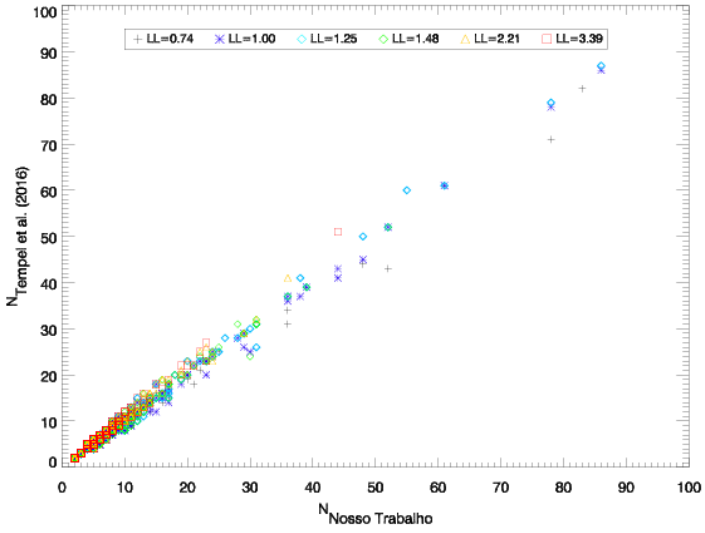

Figura 4.5 – Comparação entre a quantidade de galáxias agrupadas por agrupamento Tempel et al. (2016) e o nosso trabalho para diferentes valores de LL (em Mpc) conforme a legenda.

Na Figura 4.5 são apresentados apenas os grupos que possuem similaridade (Equação 4.10), o respectivo valor percentual para cada valor de LL é apresentado na Tabela 4.3. De forma a melhor visualizar a correlação entre o número de galáxias agrupadas por nosso método e o obtido por Tempel et al. (2016), apresentamos na Tabela 4.6 a diferença residual das galáxias mostradas na Tabela 4.5, ou seja, entre o número de galáxias por grupo obtidas por nosso método e o obtido por Tempel et al. (2016).

Na Figura 4.6 nota-se que o valor de $LL = 1{,}25$ é o que obtém grupos com maior riqueza de galáxias, além de ter uma dispersão menor para grupos com até 10 e 20 galáxias. A Tabela 4.6 apresenta o Xi^2 obtido na comparação do número de galáxias encontrado por nosso método e o obtido por Tempel et al. (2016) para essas duas faixas de galáxias. A última coluna representa o número de grupos encontrados.

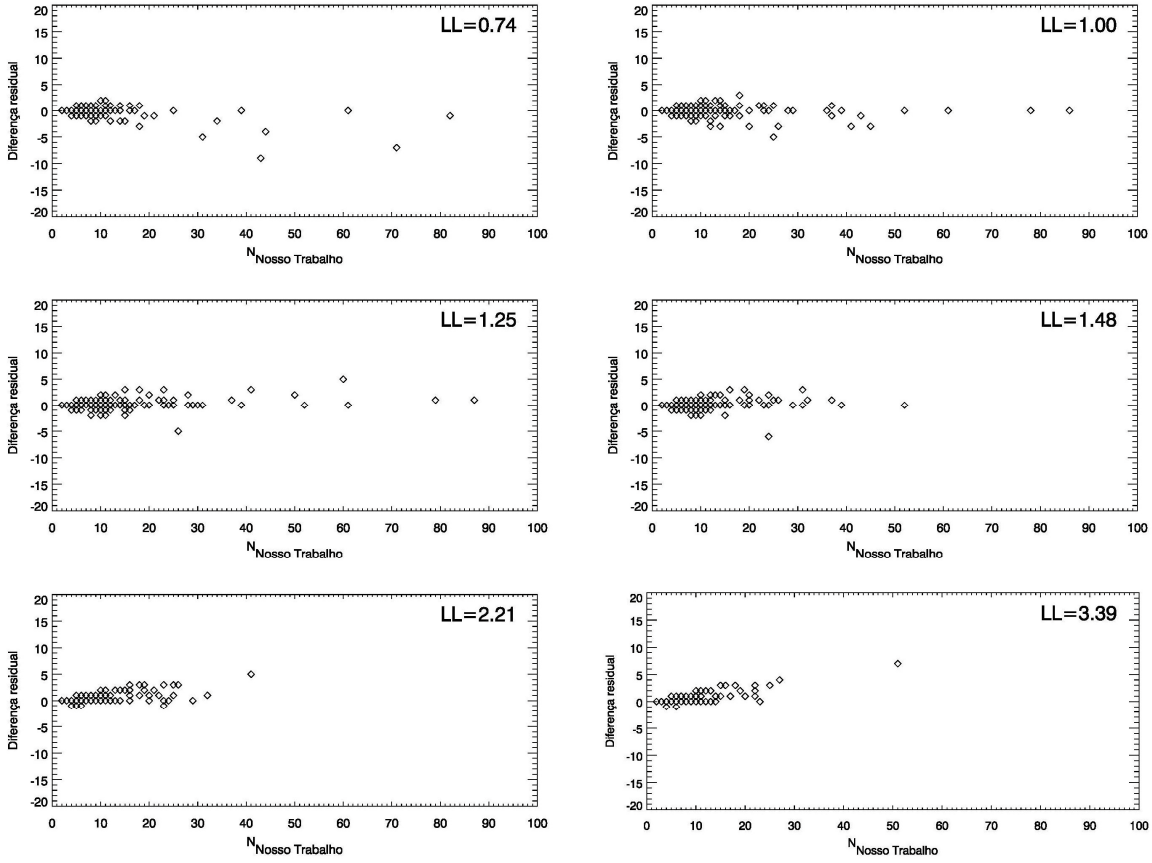

Figura 4.6 – Diferença residual, entre o número de galáxias obtidas por nosso método e o obtido por Tempel et al. (2016) e o número de galáxias obtidas por nosso método para cada valor de *LL*, apresentado na parte superior de cada figura.

Tabela 4.6 – Xi^2 obtido na comparação entre o número de galáxias obtido por nosso método e o de Tempel et al. (2016).

Até 10 galáxias	Até 20 galáxias	Número de grupos
0,218	0,245	2.463
0,205	0,243	3.767
0,196	0,225	4.545
0,206	0,231	4.760
0,221	0,265	3.869
0,261	0,310	2.375

De acordo com os resultados das simulações apresentados nessa seção podemos verificar que o valor de $LL = 1{,}25$, apresenta melhores resultados com um maior índice de similaridade com o método de validação, e agrupando uma quantidade mais próxima ao agrupamento feito por Tempel et al. (2016).

4.5 Avaliação do agrupamento pelo método intrínseco

Para a aplicação desse método escolhemos utilizar o Coeficiente de Silhueta, por este combinar duas avaliações ao mesmo tempo, a avaliação de coesão, que representa os quão ligados estão objetos em um mesmo grupo; e avaliação de separação, que representa, o quão distantes estão os objetos de outros objetos em grupos diferentes.

O Coeficiente de Silhueta (CS), descrito no Capítulo 2, é um valor que indica a qualidade do agrupamento. Sendo $-1 \leq CS \leq 1$. Quanto mais próximo de 1 for o valor de CS, melhor é o agrupamento. Os valores negativos indicam um agrupamento indesejável. Calculamos o CS para cada objeto de um grupo, bem como para cada grupo e, finalmente, para todo agrupamento. Os cálculos do CS são descritos em detalhes na Subseção 2.2.2. Aplicamos este método em nosso agrupamento e obtemos os valores apresentados na Tabela 4.7, levando em conta o LL empregado em cada caso.

Tabela 4.7 – *CS* para os agrupamentos considerando diferentes valores para *LL*.

LL	*CS*
0,74	0,80
1,00	0,80
1,25	0,80
1,48	0,79
2,21	0,76
3,39	0,69

Podemos perceber que os melhores valores foram obtidos para $0{,}74 \leq LL \leq 1{,}48$, o que indicam uma boa qualidade no agrupamento, significando boa coesão dentro dos grupos e boa dissimilaridade entre grupos diferentes. Além disso, está de acordo com a validação extrínseca feita com o agrupamento de Tempel et al. (2016), na qual, os melhores índices de similaridade também foram obtidos em agrupamentos feitos com $0{,}74 \leq LL \leq 1{,}48$, sendo que, naquele método, verificamos que o $LL = 1{,}25$ apresenta os melhores resultados.

Considerando os *LL*'s que apresentaram os quatro melhores valores de *CS*, foi feita uma avaliação para cada agrupamento em cada grupo separadamente, a fim de verificar a quantidade de grupos com *CS* negativo. O *CS* negativo não é desejável, e grupos com este comportamento devem ser considerados para uma possível exclusão do agrupamento.

De acordo com a Tabela 4.8, todos os agrupamentos apresentaram pouquíssimos grupos com *CS* negativo, sendo agrupamento com o $LL = 1{,}48$, o que apresentou a menor quantidade, porém a média do *CS* deste agrupamento ficou ligeiramente inferior aos outros. Isto se deve ao fato, de que, possivelmente, a maioria dos grupos com *CS* positivo neste agrupamento têm valores mais baixos que nos outros, influenciando em sua média geral.

Tabela 4.8 – Quantidade de grupos com *CS* negativo. A primeira coluna representa o valor do *LL*, a segunda coluna a quantidade de grupos obtidos neste agrupamento, a terceira o valor absoluto de grupos com *CS* negativo e a quarta coluna o valor relativo dos grupos com *CS* negativo em relação à quantidade total de grupos no agrupamento, que está na primeira coluna.

LL	Quantidade de grupos	Grupos com *CS* negativo	
		Valor absoluto	%
0,74	3.969	18	0,45
1,00	5.004	13	0,25
1,25	5.854	9	0,15
1,48	6,385	3	0,05

Os valores baixos para *LL* podem resultar em um agrupamento com alto índice de

coesão, uma vez que os objetos de um mesmo grupo estarão mais próximos entre si. Por outro lado, pode resultar em um baixo índice de dissimilaridade, o que pode explicar uma maior quantidade de grupos com CS negativo. Um agrupamento com esta característica pode não ser desejável, pois perderíamos vários grupos.

Os CS's consolidados por agrupamento para os quatro valores de LL se mantiveram iguais, com valor aproximado a 0,80. Contudo, de acordo com as avaliações anteriores, podemos concluir que o agrupamento feito com $LL = 1{,}25$ apresentou o melhor comportamento, pois além de apresentar o mesmo CS que os demais tem maior similaridade, de acordo com a análise feita quando utilizamos o método extrínseco.

Analisando isoladamente cada objeto do agrupamento com $LL = 1{,}25$, encontramos 207 com CS negativo. A distribuição destes objetos em relação à quantidade de objetos por grupo é apresentado na Tabela 4.9.

Tabela 4.9 – Objetos com CS negativo em relação a quantidade de objetos por grupo. A primeira coluna representa a faixa de grupos de acordo com sua população, a segunda coluna a quantidade de objetos com CS negativo, e a terceira coluna o percentual destes objetos com CS negativo em função do total de objetos com CS negativo do agrupamento, ou seja, 207 objetos.

Quantidade de Objetos por grupo	Objetos com CS negativo	
	Quantidade	%
2 – 5	46	22,22
6 – 10	51	24,63
11 – 15	17	8,00
16 – 20	8	3,00
21 ou mais	85	41,06

De acordo com a tabela acima, e tendo em vista que grupos com mais de 21 objetos sçao encontrados em menor quantidade, percebemos que quanto maior o grupo, mais objetos indesejáveis podem ocorrer. Estes objetos podem ser considerados como *outliers*.

Algumas abordagens podem ser feitas a fim de melhorar a qualidade de cada grupo, e consequentemente do agrupamento. A primeira seria remover do grupo todo objeto com CS negativo, e caso o grupo resultante fique com apenas um objeto, este último objeto deve ser removido também, removendo assim todo o grupo do agrupamento.

Uma segunda abordagem seria mover o objeto com CS negativo para outro grupo seguinte critério, caso a média da distância do objeto para outro grupo qualquer do agrupamento seja menor que a média da distância deste objeto dentro do seu próprio grupo, ele deve ser movido para o grupo de menor média e se o grupo resultante ficar com apenas um objeto, este deve ser removido.

Este processo deve ser repetido para todos os objetos com CS negativo no agrupamento, e após esta operação, um novo cálculo de CS deve ser feito no agrupamento. Caso o agrupamento ainda apresente objeto com CS negativo, todo o processo deve ser repetido até que sejam estes sejam removidos ou deixem de apresentar valor negativo.

Por não optar por excluir imediatamente os objetos com CS negativo, a segunda abordagem pode apresentar um melhor resultado em comparação com a primeira, caso os objetos possam ser remanejados com sucesso.

Analisando os métodos de validação extrínseco e intrínseco, é possível verificar que nosso agrupamento apresenta uma boa qualidade. Particularmente, o agrupamento feito com $LL = 1{,}25$ apresentou ótimos resultados tanto com o método de avaliação extrínseco, como com o método intrínseco. Desta forma, podemos concluir que este foi o melhor valor para o parâmetro LL utilizado no agrupamento.

A pontuação obtida pelo nosso agrupamento, no método extrínseco apresentou-se bastante próxima do Coeficiente de Silhueta obtido no método intrínseco, o que mostra a eficácia deste segundo. E sendo um método que não precisa de um agrupamento prévio para fazer comparações, pode ser utilizado em outros agrupamentos.

Capítulo 5

Discussão dos Resultados

O presente capítulo apresenta e analisa os resultados obtidos neste trabalho. Na Seção 5.1 é discutido o conjunto de galáxias que foram agrupadas, no tocante à riqueza dos grupos obtidos, a distância de suas galáxias em relação ao centro do seu grupo. Esses resultados nos ajudam a compreender, como essas galáxias estão agrupadas, assim como identificar as situações em que o método proposto neste trabalho apresentou melhor eficácia, e suas eventuais deficiências.

Na Seção 5.2 são discutidas as galáxias que não foram agrupadas, a sua localização e faixas de distâncias. Dessa forma, pretendemos analisar e identificar as possíveis razões do não agrupamento destas galáxias por nosso método.

5.1 Galáxias agrupadas

A quantidade de galáxias por grupo é o primeiro resultado que apresentaremos aqui. Este comportamento traz uma característica importante em nosso método, no que diz respeito à forma, com que o mesmo reconhece grandes grupos e aglomerados. Na Figura 5.1(a e b) apresentamos uma comparação entre nosso método e o de Tempel et al. (2016) quanto à riqueza dos grupos.

De acordo com a Figura 5.1(a) podemos observar que em grupos pequenos, entre quatro e 20 objetos, há grande concordância na distribuição geral de galáxias por grupos, entre os dois métodos. Nosso método, não considera um processamento pós-agrupamento, como por exemplo, a exclusão de *outliers* e uma possível realocação de objetos após a análise do agrupamento, o qual poderia resultar na divisão de grupos grandes ou mesmo na junção de outros grupos.

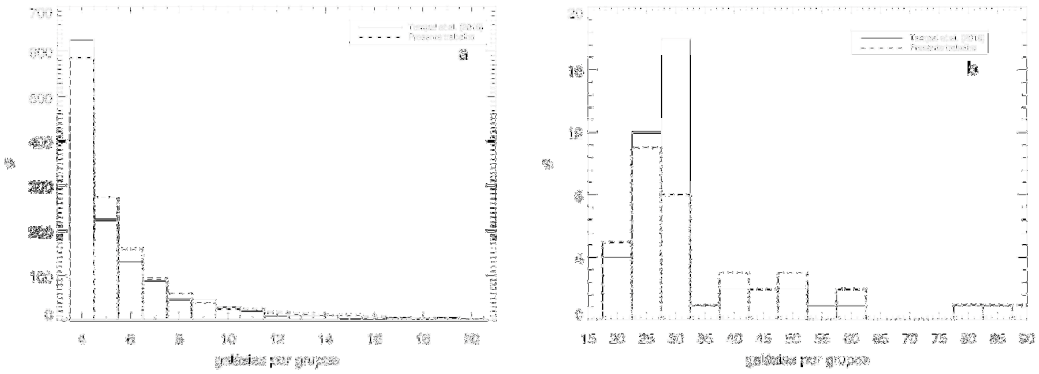

Figura 5.1 – Distribuição da quantidade de galáxias por grupo. Quantidade de galáxias agrupadas por nosso método: linhas tracejadas e por Tempel et al. (2016): linhas contínuas. A figura (a) apresenta a comparação de grupos com menos de 20 galáxias e a figura (b) de 20 a 90 galáxias.

Na Figura 5.1(b) podemos verificar a influência deste processamento pós-agrupamento, que no trabalho de Tempel et. al (2016) consistiu na utilização de uma análise multimodal com o Algoritmo EM (*expectation maximisation*), pois aí já podemos encontrar alguma diferença significativa, principalmente em grupos com 30 ou mais galáxias.

Por outro lado, em nosso trabalho preservamos os grupos grandes, com mais do que 90 galáxias, que pelo número de galáxias pode ser identificado como um aglomerado pobre (Bierrenbach, 2016), enquanto que no agrupamento de Tempel et al. (2016) o maior grupo possui 92 galáxias. Este resultado representa uma das diferenças entre os dois métodos. O nosso maior grupo que contém 152 galáxias, foi originado após a união de vários outros grupos, que é a característica principal do Algoritmo FoF. Em nosso trabalho usamos um parâmetro para limitar em no mínimo quatro galáxias em comum entre os grupos para que estes fossem unidos. A escolha deste parâmetro é explicada no Capítulo 4.

Muitas vezes, a união entre grupos muito grandes e dispersos pode gerar *outliers*, por essa razão, uma análise posterior é necessária para obter um índice de qualidade destes grupos. Em nosso caso utilizamos um método intrínseco, por meio da avaliação do Coeficiente de Silhueta (Seção 4.5). O grupo em questão obteve coeficiente 0,33, lembrando que quanto mais próximo de um melhor é o coeficiente. Neste caso, uma intervenção no grupo para tratar as galáxias com baixo coeficiente poderia melhorar o resultado final do mesmo.

De acordo com a Figura 5.2, a maioria dos grupos em nosso agrupamento apresentou uma média de distância entre cada galáxia ao centro entre 0,4 Mpc e 0,6 Mpc. Como o raio máximo utilizado foi de 1,25 Mpc, esse valor médio de distâncias mostra uma boa coesão dos grupos, ou seja, as galáxias se apresentam bastante próximas. As galáxias mais distantes podem ter sido agrupadas, devido a característica do Algoritmo FoF em unir grupos por meio de objetos em comum, conforme mencionado, no capítulo anterior. Dessa forma, quando há união entre grupos, o centro do novo grupo é alterado e passa a ficar mais distante de objetos limítrofes.

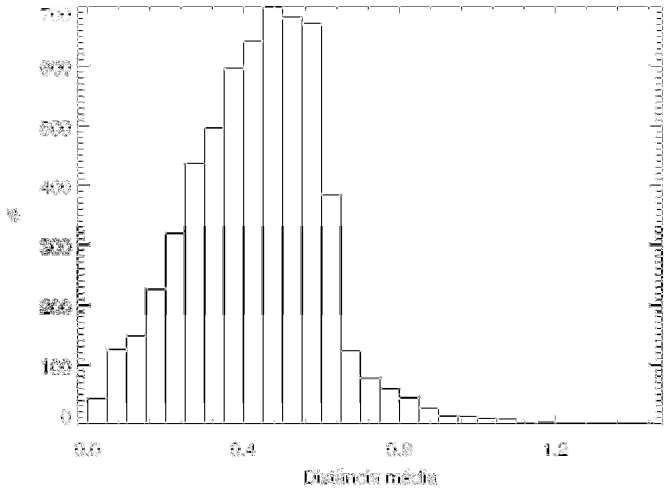

Figura 5.2 – Distância média das galáxias das galáxias ao centro do seu grupo em Mpc obtido por nosso método,.

Outra característica importante que podemos verificar na Figura 5.2, é que a quantidade de grupos com distância média superior a 0,8 Mpc é muito baixa. Isto indica que há uma boa separação entre os grupos, mais uma vez reforçando a boa qualidade do agrupamento.

De acordo com a Figura 5.3, poucas galáxias em nosso agrupamento estão situadas a uma distância máxima acima do raio utilizado como parâmetro para agrupamento (1,25 Mpc). O objeto mais distante do seu centro em todo agrupamento está a 3,16 Mpc. No entanto, na maioria dos grupos obtivemos objetos situados a distâncias entre 0,4 Mpc e 0,6 Mpc, como mencionado anteriormente, enquanto que no agrupamento de Tempel et al. (2016) a maioria das galáxias mais distantes, em relação ao centro do grupo, se encontra entre 0,2 Mpc e 0,4 Mpc. Isto significa, que em ambos os agrupamentos as galáxias de um mesmo grupo estão bastante próximas, sendo que em Tempel et. al (2016) muitos grupos aparentam conter objetos mais próximos entre si, possivelmente esta diferença está naqueles grupos que apresentam baixa similaridade, quando comparado os dois trabalhos.

Em alguns dos nossos grupos, as galáxias muito distantes podem ser consideradas *outliers*, o que podemos verificar na análise do Coeficiente de Silhueta. Aqueles objetos que possuem Coeficiente de Silhueta negativo ou muito próximo a zero, podem ser considerados *outliers* e, após avaliação podem ser descartados ou remanejados a outros grupos de acordo com sua proximidade a outros objetos.

Outro aspecto importante que pode influenciar diretamente na distância máxima dos objetos ao centro do grupo é a união de grupos durante o processo de agrupamento gerado pelo Algoritmo FoF. Em nosso agrupamento encontramos 610 grupos que foram originados por meio da união com outros grupos, o que representa 10,4% de todo agrupamento. Após a união, o centro do grupo é deslocado e seus objetos limítrofes passam a atingir distâncias maiores.

A união de grupos pequenos, com até 10 galáxias, não apresentou problemas significativos ao agrupamento, porém quando grupos muito grandes se unem, seus objetos limítrofes passam a atingir distâncias ainda maiores do centro, podendo resultar

em grupos com Coeficiente de Silhueta muito baixo ou até mesmo negativo. Uma forma de corrigir este comportamento seria, após a validação do agrupamento, utilizar abordagens como a proposta no final do Capítulo 4.

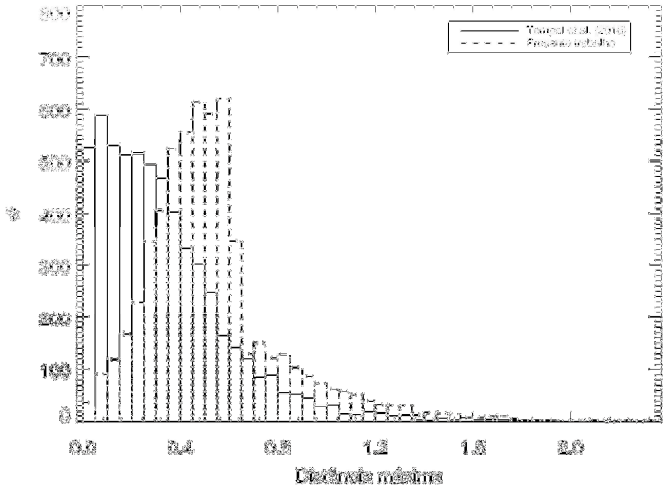

Figura 5.3 – Distância máxima de cada objeto ao centro do seu grupo em comparação com o agrupamento de Tempel et al. (2016).

De acordo com a Figura 5.4, percebe-se claramente que nosso método de agrupamento apresenta maior eficiência com galáxias na faixa que varia entre 50 e 150 Mpc. Em nosso trabalho, apenas 22,7% das galáxias com distância acima de 150 Mpc foram agrupadas. Segundo a lei Hubble as galáxias estão se afastando proporcionalmente às suas distâncias (Bierrenbach, 2016), e como usamos o mesmo LL para todo o conjunto de dados, galáxias muito distantes podem apresentar uma maior separação entre si. Uma possível solução seria utilizar um LL dinâmico de acordo com a faixa de distância dos objetos.

Ainda, de acordo com a Figura 5.4, o trabalho de Tempel et al. (2016) apresenta um agrupamento relativamente maior para galáxias com distâncias maiores do que 150 Mpc. Nesse agrupamento, aproximadamente 32,2% galáxias foram agrupadas. Isto pode ser explicado pelo fato de que no agrupamento de Tempel et al. (2016), o valor do LL é corrigido na medida em que aumenta a distância da galáxia.

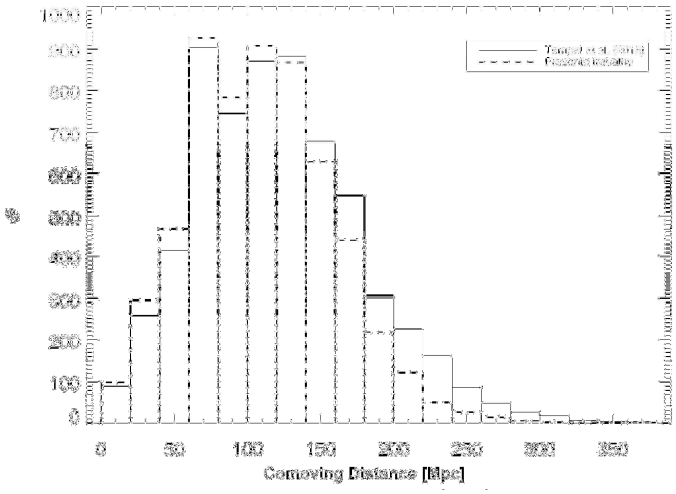

Figura 5.4 – *Comoving distance* (Mpc) das galáxias agrupadas em nosso trabalho em relação as do agrupamento de Tempel et al. (2016).

A Figura 5.5 apresenta a distribuição dos grupos encontrados na esfera celeste obtido no presente trabalho, em comparação com os obtidos por Tempel et al. (2016) em coordenadas Galácticas. Nessa figura nota-se que os símbolos menores ocorrem em maior freqüência, corroborando com Bierrenbach (2016), que diz que os grupos com dez ou menos objetos são mais freqüentes, e à medida que aumenta a riqueza dos grupos, estes se tornam mais raros. De fato, constatamos que em nosso agrupamento apenas 2,19% dos grupos possuem mais de dez objetos, enquanto que no agrupamento de Tempel et al. (2016) apenas 2% dos grupos possuem a mesma característica.

Nota-se também a semelhança entre os dois trabalhos, por meio da proximidade dos símbolos. Símbolos de cores diferentes que se sobrepõem representam grupos com praticamente os mesmos centros, ou com alto índice de similaridade em ambos os trabalhos, enquanto que a proximidade entre os símbolos representa o quão bom é definido o centro de nosso grupo em relação ao de Tempel et al. (2016). Quanto maior for a similaridade, mais próximos estarão os símbolos. Isto porque cada símbolo é localizado pelas coordenadas centrais dos grupos, portanto qualquer diferença na composição dos mesmos pode deslocá-los em relação ao seu similar.

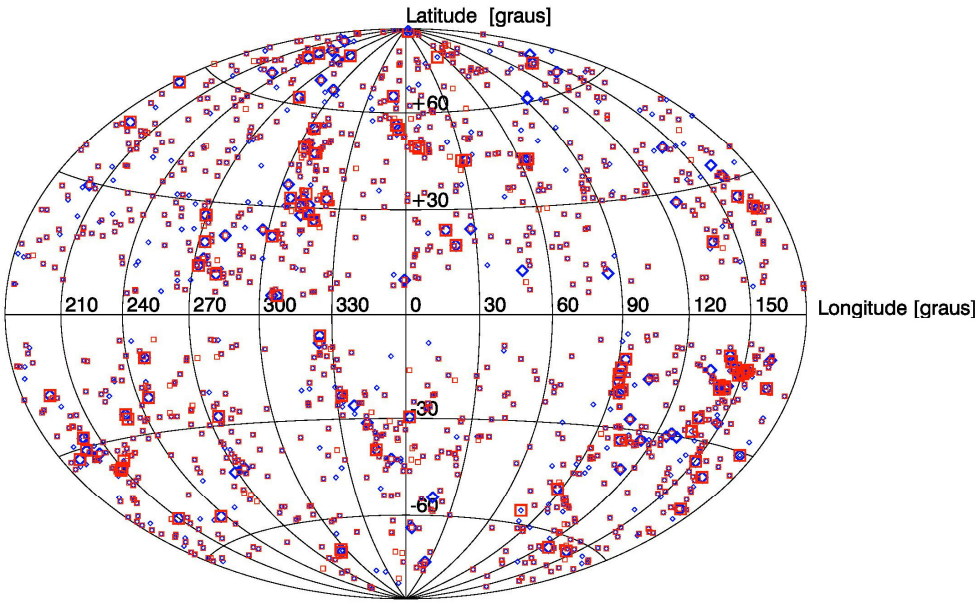

Figura 5.5 – Comparação entre as distribuições em coordenadas Galácticas dos grupos obtidos pelo presente trabalho (em azul) e Tempel et al. (2016) (em vermelho). Para ambos, o tamanho dos símbolos está dividido em grupos de quatro até dez galáxias, símbolos menores, e para mais de dez galáxias, símbolos maiores.

Na Figura 5.6 nota-se a abundância de grupos encontrados, com população igual ou inferior a dez, em relação aos grupos mais ricos. No painel superior, apesar de haver mais grupos há uma maior sobreposição dos símbolos que representam ambos os agrupamentos, o que indica que há uma maior similaridade entre o nosso agrupamento e o agrupamento de Tempel et al. (2016) em grupos na faixa de dez ou menos objetos.

Esse fato foi mostrado Capítulo 4 no qual constatamos que mais da metade dos grupos coincidentes possuem apenas dois objetos. Nota-se também, que as sobreposições dos símbolos ocorrem com maior frequência no centro da figura do que nas extremidades. Os grupos do centro da figura são compostos por galáxias mais próximas. Como vimos na Figura 5.4, nosso agrupamento apresentou melhor comportamento em galáxias até 150 Mpc.

No painel inferior, apesar de haver uma quantidade menor de sobreposições dos símbolos diferentes, nota-se que há uma proximidade dos mesmos. Neste caso, também representa similaridade, no entanto, a depender da proximidade do símbolo, em uma taxa menor. Quanto ao comportamento de maiores sobreposições e proximidades no centro da figura, repete-se aqui.

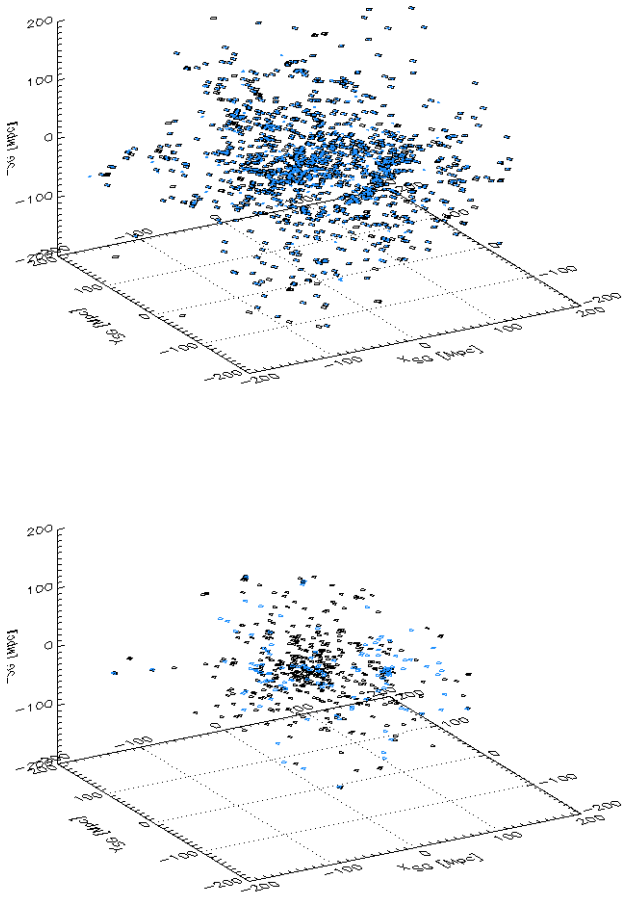

Figura 5.6 – Apresentação tridimensional dos grupos obtidos com nosso método (pontos azuis) e no trabalho de Tempel et al. (2016), diamante azul escuro. O painel superior representa os grupos com dez ou menos objetos, enquanto o painel inferior, os grupos com mais de dez objetos.

Nesta seção apresentamos os resultados das galáxias que foram agrupadas, com as principais características dos grupos encontrados, quanto à localização, a riqueza e a distância dos objetos.

5.2 Galáxias não agrupadas

Neste trabalho, em todo o conjunto de dados 55,33% das galáxias não foram identificadas em qualquer grupo, enquanto que no trabalho de Tempel et al. (2016) não foram agrupadas 54,84% das galáxias. Nesta seção iremos analisar e discutir as propriedades e características dessas galáxias.

De acordo com a Figura 5.7, notamos que não existe diferença entre o número de galáxias não agrupadas entre os hemisférios. Em ambos os trabalhos a quantidade de galáxias não agrupadas foi um pouco além da metade do catálogo. Uma possibilidade para este comportamento consiste no fato de que a maioria dessas galáxias se encontra mais distante do que as outras no catálogo, o que poderia dificultar o agrupamento das mesmas.

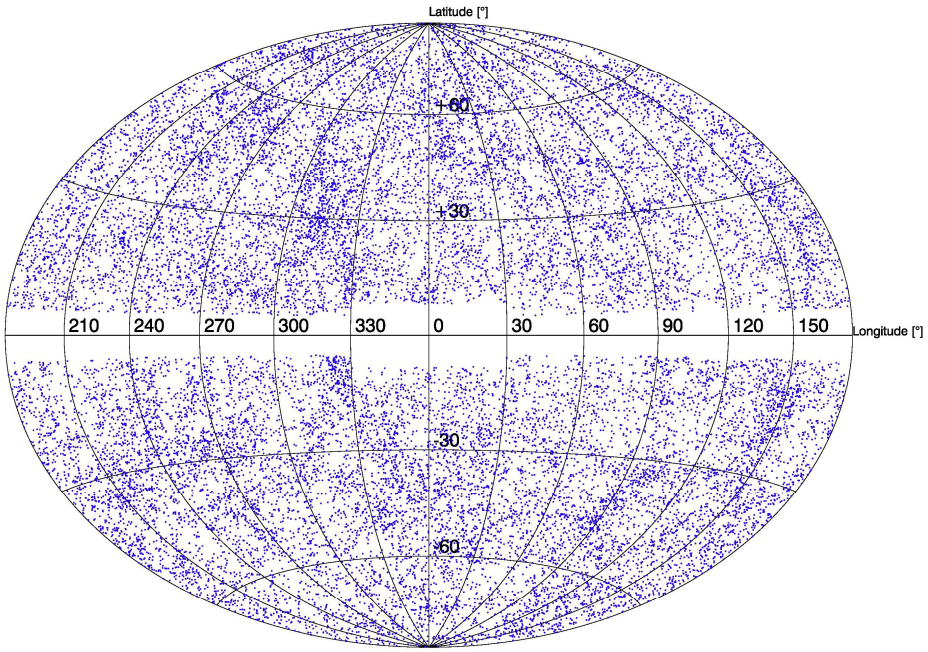

Figura 5.7 – Distribuição em coordenadas Galácticas das galáxias não agrupadas no presente trabalho.

De acordo com a Tabela 5.1 verificamos que 51,39% das galáxias não foram agrupadas em ambos os trabalhos, enquanto que um pequeno percentual foi agrupado apenas em um dos dois trabalhos de agrupamento. Esta é mais uma característica que evidencia a semelhança dos resultados obtidos em ambos os trabalhos. Esta pequena diferença pode ter sido originada pela diferença entre os *LL's* utilizados em cada agrupamento. Por se tratar de um percentual muito baixo, não influenciou de maneira significativa, entre os resultados obtidos nos dois trabalhos.

Tabela 5.1 – Galáxias agrupadas exclusivamente em cada trabalho, e não agrupadas, em nosso Método vs Tempel et al. 2016, galáxias agrupadas e não agrupadas em ambos.

Agrupamento	Quantidade	%
Apenas nosso agrupamento	1502	3,45
Apenas agrupamento de Tempel	1713	3,94
Agrupadas em ambos	17922	41,22
Não agrupadas em ambos	22343	51,39

De acordo com a Figura 5.8 nota-se a dificuldade em nosso método em agrupar galáxias muito distantes, principalmente acima de 150 Mpc. De fato, de acordo com a Tabela 5.2, a distância média entre as galáxias não agrupadas é de 219,79 Mpc, sendo que a maior distância entre estas galáxias é de 840,65 Mpc e a menor distância é de 1,25 Mpc. Isto

mostra que estão muito esparsas, ou seja, além de estarem distantes, estão distantes entre si e dificilmente formarão novos grupos dentro do parâmetro de *LL* que estabelecemos em nosso trabalho.

O fato de que 34,25% das galáxias do catálogo utilizado estão a 150 Mpc, ou valores superiores de distância, e que mais da metade deste catálogo não formaram grupos, corroboram com estas observações.

Ainda de acordo com a Tabela 5.2, as galáxias agrupadas estão em média, 148,76 Mpc distantes entre si, com a maior distância entre galáxias de 628,94 Mpc e a menor distância de 0,01 Mpc. Dessa forma, nota-se que as galáxias mais próximas também apresentam maior proximidade entre si, ou seja, maior densidade podendo assim formar novos grupos mais frequentemente.

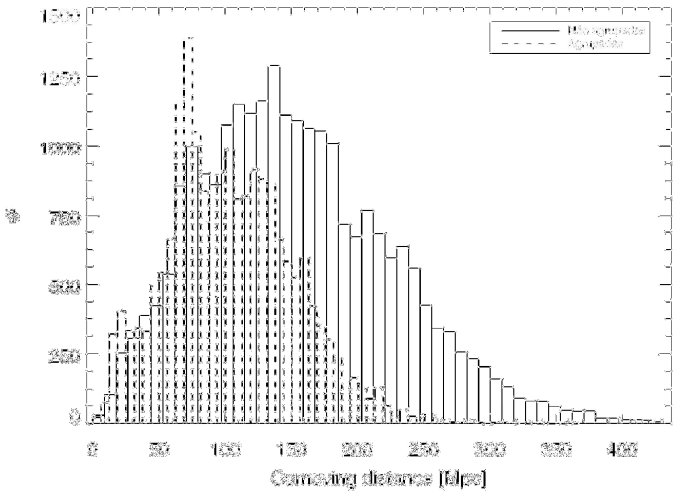

Figura 5.8 – Comoving distance das galáxias não agrupadas em nosso trabalho. As linhas tracejadas representam galáxias agrupadas, enquanto que as linhas contínuas representam as galáxias não agrupadas.

Tabela 5.2 – Distâncias média, mínima e máxima, em Mpc, entre galáxias não agrupadas e agrupadas por nosso Método.

	Distâncias em Mpc		
	Média	Mínima	Máxima
Galáxias não agrupadas	219,79	1,25	840,65
Galáxias agrupadas	148,76	0,01	628,94

Capítulo 6

Conclusão e perspectivas

No presente trabalho elaboramos um método alternativo para agrupamento de galáxias baseado no Algoritmo FoF, utilizando como parâmetros, o LL e a quantidade mínima de objetos em comum para definir a existência de um grupo. Com o propósito de realizar testes e aplicar o método, desenvolvemos algumas ferramentas computacionais para gerar e avaliar o agrupamento. Utilizamos como base de dados o catálogo disponibilizado publicamente por Tempel et al. (2016), contendo aproximadamente 44.000 galáxias.

Após a realização de pesquisa em trabalhos semelhantes (Crook et al., 2007; Diaferio et al., 1999; Tempel et al., 2014a,b, 2016) e outros, decidimos por usar seis valores diferentes para o LL, descritos na Seção 4.3 realizando simulações para seis valores diferentes com os dados do Catálogo FoF, o qual também utilizamos para validar os nossos resultados.

Verificamos que para o LL com valor 1,25 Mpc, dos 5.854 grupos de galáxias identificados, 79% atingiram 80% ou mais de similaridade, e desses, 75,74% atingiram 100% de similaridade, ou seja, mais de dois terços são idênticos ao do trabalho de referência. Dado que este valor de LL foi o qual resultou na maior quantidade de grupos idênticos em comparação ao identificado por Tempel et al. (2016), concluímos nesta análise que este valor seria o mais adequado.

Fizemos também uma análise de nosso agrupamento por meio do método intrínseco. Obtivemos resultados semelhantes do Coeficiente de Silhueta (CS) para os quatro primeiros valores de LL. O valor obtido foi aproximadamente de 0,80. Sabendo que este coeficiente varia em $-1 \leq CS \leq 1$, e quanto mais próximo de um, for esse valor, melhor é considerado o agrupamento, concluímos que a média que encontramos em nosso agrupamento, é boa; em concordância com as comparações que fizemos com o trabalho de Tempel et al. (2016).

Consideramos também, um bom resultado ter detectado em nosso agrupamento, uma baixa quantidade de grupos com $CS \leq 0$. Para quatro simulações, menos de 1% dos grupos apresentaram CS com valor negativo.

Após a comparação do nosso trabalho o de Tempel et al. (2016), e a análise do Coeficiente de Silhueta, concluímos que a simulação com $LL = 1{,}25$, foi a que apresentou melhor resultado. A análise do CS mostrou ser um método eficiente para determinar a qualidade do agrupamento. Por meio, deste, pode-se verificar, se cada objeto inserido em um grupo, não está tão próximo assim dos outros objetos do mesmo grupo, ou ainda,

se está muito próximo de um objeto em outro grupo.

Além da análise das galáxias que foram agrupadas, analisamos também as galáxias que não foram agrupadas. Verificamos que para o $LL = 1{,}25$, aproximadamente 55,33% das galáxias não foram incluídas em qualquer grupo, e dessas, 47,85% estão a uma distância igual ou superior a 150 Mpc.

Em uma análise mais detalhada, percebemos que nosso agrupamento decrescia na quantidade de grupos na medida em que aumentava a distância das galáxias, principalmente aquelas acima de 150 Mpc. Sendo assim, podemos concluir que nosso método mostrou menor eficácia no agrupamento de galáxias com distância superior a esse valor.

Apesar da dificuldade em agrupar galáxias com distâncias superiores a 150 Mpc, a partir das análises dos grupos encontrados, podemos concluir que nosso método apresentou resultado condizente com o apresentado em Tempel et al. (2016).

A fim de expandir a avaliação do comportamento do método proposto aqui, podem ser usados também outros catálogos fornecidos publicamente, submetendo-os à aplicação desenvolvida. Pretendemos expandir a nossa comparação, usando o Catálogo da Tabela 2 de Tempel et al. (2016), que possui dados para aproximadamente 80 mil galáxias.

Para melhorar o agrupamento em galáxias mais distantes, pode ser feito também um pré-processamento de vários catálogos e aferição da densidade dos objetos com distância superior ou igual a 150 Mpc no mesmo. Desta forma, pode ser feita alguma sugestão no método ou na escolha no LL para ampliar o agrupamento em distâncias superiores.

Pretendemos também elaborar mapas de contagens, com as galáxias agrupadas e não agrupadas em relação ao total de galáxias existentes para regiões com tamanhos específicos de forma a analisar se existe alguma distribuição preferencial para as mesmas, tendo também em conta, aspectos, como a extinção interestelar, entre outros.

Utilizando os valores do Coeficiente de Silhueta poderemos considerar excluir este objeto, ou ainda, inseri-lo em outro grupo, caso ele apresente maior proximidade. Desta forma, o agrupamento com um todo, pode receber um coeficiente ainda melhor. Tão logo essas etapas sejam implementadas, pretendemos deixar público nosso método.

Referências Bibliográficas

Amôres, E.B. & Lépine, J.R.D., 2005. Models for Interstellar Extinction in the Galaxy. *The Astronomical Journal*, (130:2), pp.659-73.

Amôres, E.B. & 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), pp.1538-3881.

Amôres, E.B. et al., 2013. The long bar as seen by the VVV Survey - II. Star counts. *Astronomy & Astrophysics*, (559(1)), pp.11-14.

Amôres, E.B., Robin, A.C. & Reylé, C., 2017. Evolution over time of the Milk Way's disc shape. *Astronomy & Astrophysics*, A67(602).

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

Ankerst, M., Breuning, M.M., Kriegel, H.-P. & Sander, J., 1999. OPTICS: Ordering Points To Identify the Clustering Structure. *SIGMOD '99 Proceedings of the 1999 ACM SIGMOD international conference of management of data*, 28(2), pp.49-60.

Bierrenbach, G., 2016. *Astronomia Extragalática*. [Notas de Aula] São Paulo, São Paulo, Brasil: Instituto de Astronomia, Geofísica e Ciências Atmosféricas (IAG).

Burbidge, E.M. & Burbidge, G.R., 1961. Recent investigations of groups and clusters of galaxies. *Astronomical Journal*, 66, pp.541-50.

Caretta, C.A. et al., 2008. Evidence of Turbulence-like universality in the formation of galaxy-sized dark matter haloes. *Astronomy and Astrophysics*, 87, pp.445-51.

Colless, M., 2003. *The 2dF Galaxy Redshift Survey*. [Online] Available at: http://www.2dfgrs.net/ [Accessed 27 June 2017].

Courtois, H., Paturel, G., Sousbie, T. & Labine, F.S., 2004. The LEDA galaxy distribution - I. Maps of the local universe. *Astronomy & Astrophysics*, (423), pp.27-32.

Crook, A.C. et al., 2007. GROUPS OF GALAXIES IN THE TWO MICRON ALL SKY REDSHIFT SURVEY. *The Astronomical Journal*, 01 Fevereiro. pp.790-813.

Date, C.J., 2012. *Introdução a Sistemas de Banco de Dados*. Oitava Edição ed. Rio de Janeiro: Campus.

Diaferio, A., Kauffmann, G., Colberg, J.M. & White, S.D.M., 1999. Clustering of galaxies in a hierarchical universe -III. Mock redshift surveys. *Monthly Notices of Royal Astronomical Society - MNRAS*, 307, pp.537-52.

Farrens, S. et al., 2011. Friends-of-Firends groups and clusters in the 2SLAQ catalogue. *Monthly Notices of the Royal Astronomical Society*, 417, pp.1402-16.

Han, J., KAMBER, M. & PEI, J., 2012. *DATA MINING - Concepts and Techniques*. 3rd ed. Walthen, MA, USA: Elsevier.

Huchra, J.P. & Geller, M.J., 1982. Groups of Galaxies I. nearby Groups. *The Astrophysical Journal*, pp.423-37.

Huchra, J.P. et al., 2011. The 2MASS Redshift Survey-Description and Data Release. *The Astrophysical Journal Supplement Series*, 199.

Huchra, J.P. et al., 2012. The 2MASS Redshift Survey - Descripition and Data Release. *The Astrophysical Jounal Supplement Series*, 14 March.

Jones, D.H. et al., 2009. The 6dF Galaxy Survey: Final Redshift Release (DR3) and Southern Large-Scale Structures. *Monthly Noticies of The Royal Astromical Society*, 09 outubro. pp.683-98.

Jones, D.H. et al., 2004. The 6dF Galaxy Survey: samples, observational techiques and the first data realese. *Monthly Notices Astronomical Society - MNRAS*, 355, pp.747-63.

Lavaux, G. & Hudson, M.J., 2011. The 2M++ galaxy redshift catalogue. *Monthly Notices of the Royal Astronomical Society - MNRAS*, 416, pp.2840 - 2856.

López-Corredoira, M. & Gutiérrez, C.M., 2004. The field surrounding NGC 7603: Cosmological or non-cosmological redshifts? *Astronomy & Astrophysics (A&A)*, pp.407-23.

Loula, A., 2015. *Material de aula*. [Online] Available at: http://aulas.artificial.eng.br/Home/pgca008-mineracao-de-dados/material [Accessed 25 Julho 2017].

Nadal, C.A., 2011. *UFPR - Engenharia Cartográfica e de Agrimensura*. [Online] Available at: http://www.cartografica.ufpr.br/home/wp-content/uploads/2011/10/Aula05-coordenadas-tridimensionais.pdf [Accessed 16 June 2017].

Oliveira, K.S. & Saraiva, M.d.F.O., 2014. *Astronomia e Astrofísica*. Porto Alegre, RS, Brasil: Departamento de Astronomia - Instituto de Física.

Planck Collaboration XIII, P., 2016. Plank 2015 results. XIII. Cosmological parameters. *Astronomy & Astrophysics*, 20 Sep. p.A13(63).

Ramella, M., Geller, M.J., Pisani, A. & da Costa, L.N., 2002. The UZC-SSRS2 Group Catalog. *The Astronomical Journal*, 123, pp.2976-84.

Ramella, M. & Pisani, A., 1997. Groups of galaxies in the northern Cfa redshift survey. *Astronomical Journal*, Feb. p.483.

Ramella, M., Pisani, A. & Geller, M.J., 1997. Groups of Galaxies in the Northern CfA Redshift Survey. *Astronomical Journal*, 113, p.483.

Robin, A.C., Reylé, C., Derrière, S. & Picaud, S., 2003. A Synthetic view on structure ad evolution of the Milk Way. *Astronomy & Astrophysics*, pp.523-40.

Rood, H.J., Rothman, V.A. & Turnrose, B.E., 1970. Empirical Properties of the Mass Discrepancy in Groups and Clusters of Galaxies. *Astrophysical Journal*, 162, p.411.

Ruiz, R.S., Velho, H.F.C. & Caretta, C.A., 2009. Parallel algorithm Friends-of-Friends to identify galaxies and cluster of galaxies for dark matter halos. In *IX Workshop do Curso de Computação Aplicada*. São José dos Campos, 2009. Instituto Nacional de

Pesquisas Espaciais-INPE.

Skrutskie, M.F. et al., 2006. The Two Micron All Sky Survey (2MASS). *The Astronomical Journal*, 131, pp.1163-83.

Tempel, E. et al., 2016. Friends-of-Friends galaxy group finder with membership refinement Application to the local Universe. *Astronomy & Astrophysics*, 10 Mar. p.11.

Tempel, E. et al., 2014a. Detecting filamentary pattern in the cosmic web: a catalogue of filaments for the SDSS. *Monthly Notices of the Royal Astronomical Society - MNRAS*, 21 Jan. pp.3465-82.

Tempel, E. et al., 2014b. Flux - and Volume - limited groups/clusters for the SDSS galaxies: catalogues and mass estimation. *Astronomy & Astrophysics*, 29 May. p.16.

Tully, R.B., 2015. Galaxy Groups: A 2MASS Catalog. *The Astronomical Journal*, 149(5), p.14.

Tully, B. et al., 2013. CosmicFlows-2: The Data. *The Astronomical Journal*, 146, pp.146 - 186.

Tully, R.B. & Fischer, R.J., 1978. Nearby Small Groups of Galaxies. *Astrophisycal Journal*, 79, pp.31-47.

Turner, E.L., 1976. Binary galaxies. I. A well-defined statistical sample. *Astrophysical Journal*, 208, pp.20-29.

Yadav, J.K., 2008. Nature of Clustering of Large Scale Structures.

Yadav, J.K., Somnath, B., Biswajt, P. & Seshadri, R.T., 2005. Testing homogeneity on large scales in the Sloan Digital Sky Survey Data Release One. *Monthly Notices of the Royal Astronomical Society*, 01 December. pp.601-06.