Ministério da
Ciência e Tecnologia

sid.inpe.br/mtc-m19/2011/04.27.19.23-TDI

TURBULÊNCIA EM COSMOLOGIA: ANÁLISE DE DADOS SIMULADOS E OBSERVACIONAIS USANDO COMPUTAÇÃO DE ALTO DESEMPENHO

Renata Sampaio da Rocha Ruiz

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Haroldo Fraga de Campos Velho, e César Augusto Caretta, aprovada em 26 de maio de 2011.

URL do documento original:
<http://urlib.net/8JMKD3MGP7W/39JJQ6E>

INPE
São José dos Campos
2011

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@sid.inpe.br

Ministério da
Ciência e Tecnologia

sid.inpe.br/mtc-m19/2011/04.27.19.23-TDI

TURBULÊNCIA EM COSMOLOGIA: ANÁLISE DE DADOS SIMULADOS E OBSERVACIONAIS USANDO COMPUTAÇÃO DE ALTO DESEMPENHO

Renata Sampaio da Rocha Ruiz

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Haroldo Fraga de Campos Velho, e César Augusto Caretta, aprovada em 26 de maio de 2011.

URL do documento original:
<http://urlib.net/8JMKD3MGP7W/39JJQ6E>

INPE
São José dos Campos
2011

Dados Internacionais de Catalogação na Publicação (CIP)

Ruiz, Renata Sampaio da Rocha.
R858t Turbulência em cosmologia: análise de dados simulados e observacionais usando computação de alto desempenho / Renata Sampaio da Rocha Ruiz. – São José dos Campos : INPE, 2011.
xxiv+121 p. ; (sid.inpe.br/mtc-m19/2011/04.27.19.23-TDI)

Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2011.
Orientadores : Drs. Haroldo Fraga de Campos Velho, e César Augusto Caretta.

1. Turbulência. 2. Cosmologia computacional. 3. Computação de alto desempenho. 4. Computação em grade. I.Título.

CDU 004.272.2:52-52:52-466

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de Doutor(a) em

Computação Aplicada

Dr. Fernando Manuel Ramos

Presidente / INPE / SJCampos - SP

Dr. Haroldo Fraga de Campos Velho

Orientador(a) / INPE / São José dos Campos - SP

Dr. César Augusto Caretta

Orientador(a) / UG / Mexico - ME

Dr. Stephan Stephany

Membro da Banca / INPE / SJCampos - SP

Dr. Jerônimo dos Santos Travelho

Membro da Banca / INPE / São José dos Campos - SP

Dr. João Braga

Membro da Banca / INPE / São José dos Campos - SP

Dr. Laerte Sodré Jr.

Convidado(a) / IAG/USP / São Paulo - SP

Dra. Andrea Schwertner Charao

Convidado(a) / UFSM / Santa Maria - RS

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

Aluno (a): Renata Sampaio da Rocha

São José dos Campos, 26 de maio de 2011

"Se não existe vida fora da Terra, então o Universo é um grande desperdício de espaço".

Carl Sagan

A meus pais Rosalvo Sampaio da Rocha e Maria Benevides da Rocha, a meus irmãos e ao meu querido esposo Paulo Roberto

AGRADECIMENTOS

A Deus por ter me dado saúde, perseverança e coragem para enfrentar os desafios.

À minha família pelo incentivo e apoio, em especial ao meu esposo Paulo Roberto pelo carinho e dedicação.

Aos meus orientadores Prof. Dr. Haroldo Fraga de Campos Velho e Prof. Dr. César Augusto Caretta pela paciência, pelo conhecimento compartilhado, pelas valiosas sugestões, pela orientação e apoio na realização deste trabalho.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela bolsa de doutorado no Brasil (Processo nº 2007/54133-0).

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de doutorado no exterior (Processo nº 4541/08-1).

Ao Instituto Nacional de Pesquisas Espaciais (INPE), pela oportunidade de estudos e utilização de suas instalações.

Ao Departamento de Astronomia (DA) da Universidade de Guanajuato, pelo suporte e apoio durante minha estadia no México.

Aos professores do INPE pelo conhecimento compartilhado.

Aos membros da banca examinadora pelas valiosas sugestões para o aprimoramento do trabalho.

Aos amigos que incentivaram e me acompanharam nessa jornada, em especial, o amigo e companheiro de sala no Inpe, Eduardo Luz, pelo carinho e pelas valiosas dicas em computação de alto desempenho, o amigo Wantuir pelas longas discussões, sugestões e esclarecimentos, a Fabiana, que sempre esteve me apoiando em todos os momentos difíceis, aos amigos do DA: Juan Pablo Papaqui, Hector Bravo, Carlos Rico, Irán, Tatiana, Isabel, Josué, René, Abdias, Daniel, Juan Luis, Juan Manuel, Carlos Arturo e Carlos Fernando, que me receberam com carinho e fizeram com que minha estadia em Guanajuato fosse muito agradável.

Aos cidadãos brasileiros que com seus impostos custearam todo o meu estudo, desde a alfabetização até a conclusão deste doutorado.

RESUMO

Estudos recentes sobre a evolução da estrutura em grande escala do Universo sugerem que o processo de formação de estruturas pode possuir similaridades com a dinâmica de um fluido turbulento. A lei de escala do espectro de energia para caracterizar um escoamento turbulento de fluidos apresenta um intervalo inercial com uma potência $-5/3$, conhecida como lei de Kolmorogov. Nesta tese, é apresentada uma avaliação do espectro de energia potencial gravitacional para verificar se a lei de escala desse espectro segue a lei de potência de Kolmogorov. Para esta análise, foram utilizados halos de matéria escura de galáxias em diferentes *redshifts*, provenientes de simulações de N-corpos do Consórcio Virgo, e também uma amostra observacional de galáxias, proveniente da base de dados do *Sloan Digital Sky Survey* (SDSS). Os resultados indicam que, para halos de matéria escura de galáxias, existe um intervalo ($0.01 < k < 0.07$, ou 14 a 100 h^{-1} Mpc) em que a inclinação do espectro se aproxima da lei de $-5/3$ de Kolmogorov. Para os dados observacionais, também é possível identificar um intervalo onde os espectros se aproximam de $-5/3$ (neste caso em escalas um pouco mais restritas, de 30 a 70 h^{-1} Mpc), evidenciando assim uma característica típica de processos turbulentos. Para viabilizar a análise e o processamento, técnicas de mineração de dados foram investigadas e implementadas. Também foram desenvolvidas versões paralelas dos algoritmos utilizados, para os quais se utilizou a biblioteca *Message Passing Interface* (MPI). O desempenho dos algoritmos paralelos foi avaliado por meio do cálculo de *speed-up* e apresentou bons resultados. Para aumentar a capacidade de computação, uma grade computacional foi estruturada para aplicações em Astrofísica, sendo utilizada nesta tese, para a identificação de halos de matéria escura (com a versão paralela do algoritmo de agrupamento *Friends of Friends*) e o cálculo do espectro de energia potencial gravitacional para diferentes *redshifts*.

TURBULENCE IN COSMOLOGY: ANALYSIS OF SIMULATED AND OBSERVATIONAL DATA USING HIGH PERFORMANCE COMPUTING

ABSTRACT

Recent results from the study of the Large-Scale Structure evolution suggest that the process of structure formation may have similarities with the dynamics of a turbulent fluid. The scaling law for the energy distribution in a turbulent fluid shows an inertial range of power -5/3, known as the Kolmogorov's law. This thesis presents an evaluation of the gravitational potential energy spectrum in order to verify if its scaling law follows the same power as Kolmogorov's law. For this analysis, we used galaxy dark matter halos in different redshifts, from Virgo Consortium's N-body simulations, as well as galaxy observational data, from the Sloan Digital Sky Survey (SDSS) database. We found that, for the galaxy dark matter halos, the spectrum follows closely the Kolmogorov's power law in the range $0.01 < k < 0.07$ (from 14 to 100 h^{-1} Mpc). The same is seen for the observational data, in a more restrict range from 30 to 70 h^{-1} Mpc. This means that the gravitational clustering of dark matter and galaxies may admit a turbulent-like representation. For optimizing the processing, data mining techniques have been investigated and implemented, and also parallel versions of the used algorithms, using the Message Passing Interface (MPI), were developed. The performance of the parallel algorithms was evaluated by speed-up measure and showed good results. To meet the demand of computational resources, we developed libraries and implement a computational grid for Astrophysical applications, used in this thesis for the identification of galaxy dark matter halos (with the parallel version of the Friends-of-Friends algorithm) and gravitational potential energy calculations, both for different redshifts.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE ABREVIATURAS E SIGLAS

AZ	–	Aproximação de Zel'dovich
BoT	–	Bag-of-Tasks
BDA	–	Brazilian Decimetric Array
BRAVO	–	Brazilian Virtual Observatory
CDM	–	Cold Dark Matter
CPU	–	Central Processing Unit
DA	–	Departamento de Astronimia
DAS	–	Divisão de Astrofísica
DPOSS	–	Digital Palomar Observatory Sky Survey
FPGA	–	Field Programmable Gate Arrays
FoF	–	Friends of Friends
Ga	–	Giga-anos
GPU	–	Graphics Processing Unit
GuMs	–	Grid Machines
HPF	–	High Performance Fortran
ICM	–	Intracluster Medium
INPE	–	Instituto Nacional de Pesquisas Espaciais
IVOA	–	Intenational Virtual Observatory Alliance
LAC	–	Laboratório Associado de Computação e Matemática Aplicada
LSST	–	Large Synoptic Survey Telescope
MCP	–	Modelo Cosmológico Padrão
MPI	–	Message Passing Interface
Mpc	–	Megaparsec
OV	–	Observatórios Virtuais
PAD	–	Processamento de Alto Desempenho
Pan-StaRRS	–	Panoramic Survey Telescope and Rapid Response System
PVM	–	Parallel Virtual Machine
RMI	–	Remote Method Invocation
SSL	–	Secure Socket Layer
SDSS	–	Sloan Digital Sky Survey
SOAR	–	Southern Observatory for Astrophysical Research
6dFGRS	–	Six Degree Field Galaxy Redshift Survey
TI	–	Tecnologia da Informação
2dFGRS	–	Two Degree Field Galaxy Redshift Survey
2MASS	–	Two Micron All Sky Survey
UFSM	–	Universidade Federal de Santa Maria
UKIDSS	–	UKIRT Infrared Deep Sky Survey
WMAP	–	Wilkinson Microwave Anisotropy Probe

LISTA DE SÍMBOLOS

F	–	força de atração gravitacional
G	–	constante gravitacional
M	–	massa
m	–	massa
d	–	distância
Λ	–	constante cosmológica
v	–	velocidade de recessão de galáxias
H	–	constante de Hubble
H_0	–	constante de Hubble na época atual
a e $a(t)$	–	fator de escala cósmica
t_H	–	tempo de Hubble
D_H	–	distância de Hubble
c	–	velocidade da luz
ρ	–	densidade de massa
ρ_c	–	densidade crítica
z	–	redshift
λ_e	–	comprimento de onda emitido
λ_o	–	comprimento de onda observado
D_c	–	distância co-móvel (linha de visada)
D_M	–	distância co-móvel (transversal)
D_A	–	distância do diâmetro angular
D_{A12}	–	distância entre dois objetos celestes
D_L	–	distância de luminosidade
K	–	energia cinética
U	–	energia potencial gravitacional
m_i	–	massa de uma partícula i
$J(r)$	–	densidade de luminosidade
W_i	–	função peso para uma galáxia i
$\ddot{r}_i$	–	aceleração da *i-ésima* partícula em um sistema
r_i	–	posição da partícula i
κ	–	curvatura do espaço
t	–	tempo
p	–	pressão
Φ	–	potencial gravitacional
r	–	coordenada de distância co-móvel
u	–	velocidade co-móvel
$\overline{u}$	–	velocidade média do fluido
D	–	longitude característica do fluxo ou o diâmetro para o fluxo
μ	–	viscosidade dinâmica do fluido

ν – viscosidade cinemática
R_e – número de Reynolds
R_c – número de Reynolds crítico
Υ_I – taxa de transferência de energia
ε – taxa de dissipação de energia
k – número de onda
C – constante de Kolmogorov
ϵ – parâmetro de suavização
M_T – massa total de uma halo de matéria escura
$\overline{V}$ – velocidade média de um halo de matéria escura

SUMÁRIO

Pág.

1 INTRODUÇÃO

O trabalho de pesquisa desenvolvido nesta tese está vinculado ao projeto *Brazilian Virtual Observatory* (BRAVO), um projeto interdisciplinar que conta com a participação de várias instituições brasileiras e colaboradores internacionais, visando a implementação de um Observatório Virtual (OV) no Brasil. Uma vez que a astronomia conta atualmente com uma quantidade gigantesca de dados, o conceito de Observatório Virtual foi proposto pela comunidade astronômica com o objetivo de atender a crescente demanda, tanto de recursos para disponibilização e acesso homogêneo destes dados, quanto de ferramentas e algoritmos capazes de permitir sua completa exploração.

O propósito principal do trabalho é verificar a possibilidade de existência de um processo similar à dinâmica turbulenta na formação das grandes estruturas observadas no Universo (galáxias, aglomerados, superaglomerados, filamentos). O entendimento sobre o processo de formação dessas estruturas é uma questão relevante da Cosmologia moderna.

O modelo padrão de formação de estruturas é baseado na teoria de instabilidades gravitacionais, que tem por base o modelo de colapso esférico proposto por Jeans (1902). Segundo esse modelo, perturbações de densidade (ou flutuações) tendem a aumentar sua amplitude com o tempo (ou seja, tornarem-se mais densas) em um processo conhecido como colapso gravitacional. Essas flutuações criam concentrações de matéria que evoluem para formar estrelas, galáxias, aglomerados de galáxias, superaglomerados de galáxias e assim por diante (PADMANABHAN, 1993; TATEKAWA, 2005; HAWLEY; HOLCOMB, 2005). Enquanto o contraste de densidade dessas flutuações permanecer relativamente pequeno, sua amplitude cresce linearmente. Desta forma, as equações que governam o movimento das partículas que compõem o "fluido cósmico" podem ser resolvidas analiticamente por meio da teoria da perturbação. No entanto, no cenário hierárquico em que as flutuações aumentam sucessivamente e entram em um regime não linear, ainda não existem soluções analíticas capazes de descreverem completamente tal comportamento(PADMANABHAN, 2006a).

A dinâmica do fluido cósmico pode ser descrita por uma Cosmologia semi-Newtoniana (uma formulação Newtoniana com hipóteses convenientes provenientes dos resultados gerais relativísticos), na qual o movimento da matéria, dominado pela expansão cósmica, é dado pelas equações de Friedmann (KOLB; TURNER, 1990;

PADMANABHAN, 1993; MADSEN, 1996; LIDDLE; LYTH, 2000; TATEKAWA, 2005). As equações de Friedman são um conjunto de equações em Cosmologia que governam a expansão métrica do espaço em modelos homogêneos e isotrópicos do Universo dentro do contexto da teoria da relatividade geral.

Matsubara (1995) destaca que existem duas formulações para as teorias da evolução dinâmica do Universo. Uma é a formulação Euleriana e a outra é a formulação Lagrangeana. Uma das formulações Lagrangeanas mais utilizada na simulação de N-corpos é a bem conhecida *Aproximação de Zel'dovich* (AZ), (ZEL' DOVICH, 1970). Com base na AZ outras aproximações foram propostas, como as versões otimizadas (COLES et al., 1993; MELOTT et al., 1994), a versão modificada (HUI; BERTSCHINGER, 1996), bem como, soluções perturbativas de alta ordem (CATELAN, 1995; MATSUBARA, 1995; SAHNI; COLES, 1995).

Uma abordagem pouco explorada no desenvolvimento de modelos para a evolução cósmica é a possibilidade de que a formação das grandes estruturas seja relacionada a hidrodinâmica, isto é, pela existência de um comportamento similar ao turbulento no processo de agrupamento da matéria.

Originalmente, o primeiro trabalho relacionando turbulência à formação de estruturas no Universo foi realizado por Von Weizsacker em 1951 (WEIZSACKER, 1951). Nesse trabalho, ele apresenta mecanismos para descrever de forma aproximada o movimento da matéria cósmica em seus diferentes estados (gás, poeiras, estrelas). Segundo o autor, o movimento do gás cósmico obedece as equações da hidrodinâmica, sendo, na maioria dos casos, turbulento e podendo ser incompressível. Nesse sentido, propõe um esquema geral de evolução para um corpo gasoso, em que uma nuvem turbulenta, gravitacionalmente estável é inicialmente achatada em um disco rotativo, que posteriormente é dissolvido em um corpo central que gira uniformemente e onde uma parte retorna ao espaço cósmico. Ainda, segundo o autor, as galáxias aparentemente foram formadas por uma competição entre a expansão e a turbulência. Dessa forma, a principal proposta do trabalho foi introduzir o conceito de hierarquia dos turbilhões e formular a lei de Kolmogorov de uma maneira mais apropriada para o uso em Astrofísica. Posteriormente outros trabalhos foram desenvolvidos, dos quais podemos citar: Ozernoy e Chernin (1968), Ozernoy e Chibisov (1971), Oort (1970).

Existem evidências provenientes tanto de trabalhos observacionais quanto numéricos

sugerindo que uma parte não desprezível de movimentos turbulentos está armazenada no meio intergaláctico de aglomerados de galáxias (ICM, do inglês *Intracluster Medium*). Vazza et al. (2006), apresentaram um estudo sobre os campos de velocidade turbulenta no ICM de uma amostra simulada de 21 aglomerados de galáxias. Os movimentos turbulentos no ICM dos aglomerados simulados foram detectados usando um método desenvolvido para melhor distinguir os movimentos laminares dos movimentos caóticos. O foco principal do trabalho foi encontrar uma relação de escala entre a componente de energia turbulenta das partículas do gás e sua energia cinética na forma de movimentos turbulentos. Seus resultados indicam uma lei de potência entre a energia cinética turbulenta e a massa dos aglomerados.

Simulações numéricas Eulerianas da fusão de aglomerados (RICKER; SARAZIN, 2001) ou da formação de estruturas menos massivas do que galáxias anãs têm fornecido boas representações da maneira pela qual a turbulência pode ter sido injetada no ICM. Além disso, resultados teóricos também sugerem uma quantidade não desprezível de movimentos turbulentos em aglomerados de galáxias (CASSANO; BRUNETTI, 2005).

Contudo, o conteúdo material do Universo não é composto apenas de matéria bariônica [1] mas também de matéria escura (matéria que só interage gravitacionalmente, a definição formal de matéria escura está na página 53), e se supõe que essa componente é a que controla a formação das estruturas. O problema é que a matéria escura é considerada acolisional, ou seja, além de não interagir eletromagneticamente ela também tem uma seção de choque muito pequena, fazendo com que suas "partículas" não interajam também por colisões. Por outro lado, a turbulência normalmente está associada a propriedades de fluidos bariônicos que são altamente colisionais. Nakamichi e Morikawa (2009) sugerem que o fluido acolisional de matéria escura passa a um estado turbulento após atravessar a superfície cáustica no estado não linear. Os autores propõem uma possível origem para o momento angular e as relações de escala apresentadas por sistemas auto-gravitantes a partir de movimentos turbulentos de matéria escura.

Shandarin e Zel'dovich (1989) sugerem que um simples cenário turbulento baseado na AZ é compatível com um Universo dominado por matéria escura quente (ou seja, relativística), no qual o processo de formação de estruturas é *top-down*, onde as gran-

[1] Se considera matéria bariônica todo tipo de matéria constituída por bárions e léptons, ou seja, é a matéria visível que forma tudo o que observamos.

des estruturas originais ("panquecas") correspondem à escala integral que se forma primeiro e depois se fragmenta em pequenos objetos, seguindo o processo de cascata. Entretanto, desde os anos 80 do século passado, acumulou-se uma grande quantidade de evidências, principalmente a partir de simulações de N-corpos cosmológicas (e.g. Davis et al. (1985)), de que a matéria escura do Universo é dominantemente fria.

A pergunta que tal raciocínio suscita é: Mesmo que não haja exatamente a turbulência que verificamos nos fluidos bariônicos, podem os processos não-lineares da formação de estruturas em escala cósmica ser similares aos processos turbulentos? Se a analogia entre a evolução cosmológica e a dinâmica da turbulência considerando intermitência está correta, alguma assinatura da turbulência deverá estar presente em dados da Cosmologia observacional e em dados simulados da evolução do Universo. Assim, é conveniente avaliar a validade de algumas propriedades do escoamento turbulento em dados da Cosmologia observacional ou em dados simulados da evolução do Universo. Deste modo, características da turbulência deveriam ser identificadas a partir de dados que estão disponíveis e determinar em que escalas tais características se manifestam. Para separar (ou identificar) objetos em diferentes escalas a partir de registros (dados) astronômicos, várias técnicas de mineração de dados podem ser empregadas, como algoritmos de agrupamento e redes neurais artificiais. Nesse sentido, foi realizada uma análise preliminar em busca de assinaturas da turbulência em dados simulados (CARETTA et al., 2008). Nesse estudo, foi analisado o espectro de energia potencial gravitacional de halos de matéria escura de galáxias e aglomerados de galáxias. Os resultados indicaram que, para halos de galáxias, este espectro pode ser bem descrito por uma lei de potência do tipo $-5/3$ em um intervalo de 15 a 50 $h^{-1}\ Mpc$, em que $0.6 \leq h \leq 0.9$ e Mpc é uma unidade de medida astronômica para comprimento, sendo que 1 *megaparsec* (Mpc) equivale a 3.26 milhões de anos-luz. Isto pode ser uma evidência de um comportamento turbulento, uma vez que esta potência coincide com a lei de Komolgorov de 1941 (FRISCH, 1995). Porém, nessa primeira análise, a simulação considerada tinha pouca resolução de massa ($M_{min} = 2 \times M_{part} \approx 1.4 \times 10^{11}\ h^{-1}\ Mpc$).

Para poder dar uma conclusão mais definitiva é necessário aprofundar os estudos, pelo menos em duas direções: uma maior resolução nas simulações e uma busca dos mesmos sinais em amostras de dados observacionais. Neste sentido, apresentamos nesta tese uma análise mais profunda considerando dados da Simulação *Millennium* ($M_{min} = 100 \times M_{part} \approx 8.6 \times 10^{10}\ h^{-1}\ Mpc$) e também dados reais (observacionais)

provenientes da base de dados do projeto *Sloan Digital Sky Survey* (SDSS). Este tipo de análise demanda recursos computacionais intensivos. Desta forma, versões paralelas dos algoritmos foram desenvolvidas e implementadas numa infra-estrutura de grade computacional para a Astrofísica.

Para os cálculos realizados neste trabalho utilizamos recursos computacionais do Laboratório Associado de Computação e Matemática Aplicada (LAC), da Divisão de Astrofísica (DAS), ambos do INPE, do Departamento de Eletrônica e Computação, da Universidade Federal de Santa Maria (UFSM) e do Departamento de Astronomia da Universidade de Guanajuato - México.

1.1 O Projeto BRAVO

Nos últimos anos o volume de dados astronômicos, tanto de observação quanto de simulação de N-corpos, tem crescido de maneira exponencial, o que demanda num desafio em tecnologia de informação para disponibilizar registros científicos, da ordem de *PetaBytes*, de maneira veloz, eficiente e amigável para o usuário (CARVALHO et al., 2009). Tais desafios incluem a homogeneização dos registros de dados, o desenvolvimento de ferramentas de análise eficientes, versáteis e intercompatíveis, e a capacidade de gerenciá-los e de atender solicitações num ambiente de uma rede mundial de computadores com enorme quantidade de informações.

Grandes levantamentos de dados fotométricos e espectroscópicos já foram ou estão em desenvolvimento em ambos os hemisférios do planeta, por exemplo: SuperCOSMOS (4 *TeraBytes* de dados, céu completo) (e.g. Hambly et al. (2001)), *Two-Micron All-Sky Survey* (2MASS), aproximadamente 1 *TeraByte*, céu completo (e.g. Skrutskie et al. (2006)), *Digital Palomar Observatory Sky Survey* (DPOSS), 3 *TeraBytes*, parte do céu (e.g. Gal et al. (2004)), *Panoramic Survey Telescope & Rapid Response System* (Pan-STaRRS), 3 *PetaBytes*, a partir de 2013 (alguns *TeraBytes* por noite!), *Large Synoptic Survey Telescope* (LSST), a partir de 2020 (30 *TeraBytes* por noite!), SDSS (64 *TeraBytes*, parte do céu) (e.g. Abazajian et al. (2009)), *Two-degree Field Galaxy Redshift Survey* (2dFGRS) (e.g. Colless et al. (2001)), *Six-degree Field Galaxy Redshift Survey* (6dFGRS) (e.g. Jones et al. (2009)), Dark Energy Survey (DES), entre outros. Como resultado, há uma quantidade extraordinária de dados de praticamente todas as bandas do espectro eletromagnético e que vem aumentando exponencialmente. A Figura 1.1 mostra o rápido crescimento do volume de dados astronômicos, considerando somente alguns dos levantamentos mais importantes.

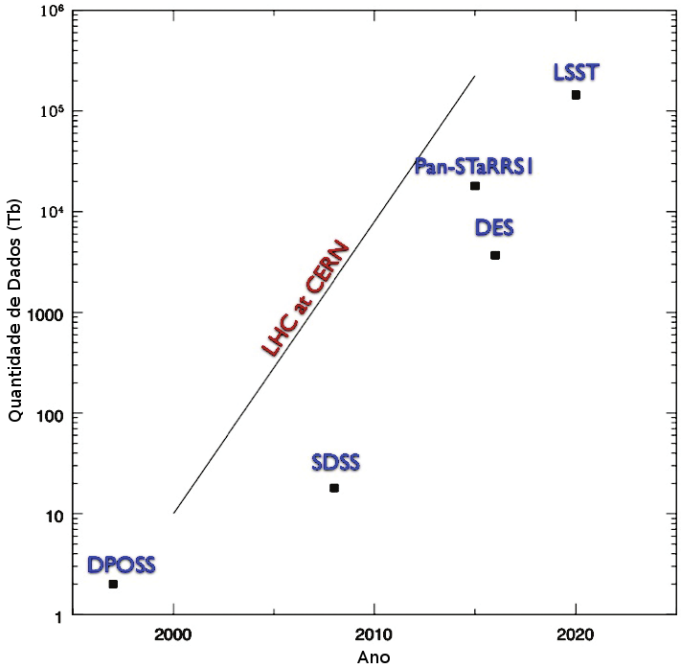

Figura 1.1 - Quantidade de dados astronômicos que serão disponibilizados até 2020. Fonte: Carvalho et al. (2009).

Para atender à crescente demanda de recursos computacionais de armazenamento e disponibilização desses dados, além do fornecimento de ferramentas eficientes de análise e exploração dos mesmos, a comunidade astronômica desenvolveu o chamado Observatório Virtual. Observatórios virtuais representam uma grande inovação na astronomia e englobam um esforço coletivo em que vários centros de pesquisas de diferentes partes do mundo trabalham em conjunto para prover recursos necessários a uma completa e efetiva exploração do gigantesco conjunto de dados disponíveis atualmente nas diversas bases de dados astronômicos (KIM et al., 2005; CARVALHO et al., 2009). Os observatórios virtuais existentes em diversos países estão agregados em uma organização internacional denominada *International Virtual Observatory Alliance* (IVOA)[2], responsável por definir protocolos, portais e padrões que permitirão a unificação das bases de dados astronômicas de todo o mundo em um único "telescópio virtual".

No Brasil, para preparar a comunidade astronômica brasileira para atender essa demanda, uma iniciativa muito recente é a criação do projeto BRAVO[3]. O obje-

[2] `http://www.ivoa.net`
[3] `http://www.lac.inpe.br/bravo/`

tivo principal do projeto é gerar investimento em Tecnologia da Informação (TI), com ênfase principalmente nas questões de infra-estrutura computacional, grades de processamento e de armazenamento de dados, bem como na implementação de ferramentas de mineração de dados astronômicos. Com a criação do BRAVO, o Brasil passou a fazer parte da IVOA em meados de 2009.

Do ponto de vista científico, os participantes do projeto são especialistas em vários campos da Cosmologia, da Astrofísica, sistemas computacionais e análise de dados, com significativa contribuição de parceiros internacionais, conduzindo pesquisas nas áreas de aglomerados e grupos de galáxias, galáxias elípticas, simulação de N-corpos e uma variedade de estudos em Astrofísica Estelar. Um dos principais aspectos do projeto é a incorporação desta ampla variedade de áreas da pesquisa em Astrofísica dentro de um contexto de desenvolvimento coerente da tecnologia de banco de dados (BRAZILIAN..., 2008).

O projeto BRAVO utilizará várias bases de dados astronômicos específicos, criados para armazenar dados gerados por vários instrumentos (incluindo os telescópios SOAR e Gemini e o rádio-telescópio BDA, entre outros), unindo-os num sistema comum e com interfaces padrão segundo o protocolo de um Observatório Virtual.

O procedimento padrão de um astrônomo profissional consiste em solicitar e obter (por competição internacional) um tempo de observação, que pode ser de dias ou semanas e, após operar um telescópio para aquisição dos dados desejados, reduzi-los, analisá-los e, finalmente, publicar os resultados de sua pesquisa. A crescente disponibilidade de dados e os modernos recursos computacionais e de telecomunicações permitem que os astrônomos atualmente tenham acesso a dados de maneira quase instantânea e realizem suas pesquisas com o uso desses dados disponíveis. Entretanto, estas inovações não caracterizam uma estrutura de OV, pois, grande parte desses dados são disponibilizados para a comunidade por meio de servidores públicos, geralmente em vários formatos diferentes e distribuídos em diversas instituições. A qualidade dos dados, metadados, interfaces e a acessibilidade são heterogêneas, uma vez que cada projeto geralmente trata seus próprios dados e os apresenta em um banco de dados personalizado.

Um conceito subjacente ao OV é que, proporcionando melhor acesso e dados homogêneos, combinados com ferramentas para explorar e manipular esses dados, a necessidade de novas observações será reduzida e a produção científica poderá ser

aumentada, beneficiando, de um modo geral, toda a comunidade científica

A solução da ciência da computação para tratar problemas que demandam recursos computacionais intensivos de processamento e armazenamento é a chamada Computação de Alto Desempenho, a qual está fundamentada no uso de sistemas computacionais como *clusters* e/ou supercomputadores por meio do paradigma da computação paralela. A computação paralela é uma técnica de programação em que várias instruções são executadas ao mesmo tempo (concorrentemente) visando a redução do tempo total de execução. Uma iniciativa mais recente é a proposta de computação em grade (FOSTER, 2003; FOSTER; KESSELMAN, 2004), cujo objetivo é conectar vários sistemas computacionais através da internet, de tal forma que o novo sistema integrado possa estar dedicado (em algum momento) para a execução de alguma tarefa importante, com uma capacidade de computação ampliada grandemente. Detalhes sobre esta abordagem serão tratados na Seção 4.2

1.2 Objetivos

O objetivo geral deste trabalho é identificar similaridades entre a evolução cósmica e a dinâmica de um fluido turbulento. A investigação consistiu no cálculo e análise do espectro de energia potencial gravitacional de halos de matéria escura de galáxias, provenientes da simulação de N-corpos e também de dados observacionais de galáxias provenientes do SDSS. No contexto do projeto BRAVO os objetivos específicos estão relacionados ao desenvolvimento de ferramentas de análise de dados astronômicos, bem como, na implementação de uma infra-estrutura de grade computacional dedicada a realização de tarefas importantes da astronomia que requeiram o uso de computação intensiva.

1.3 Organização da Tese

Esta tese de doutorado está organizada da seguinte maneira: No Capítulo 2 tem-se uma breve descrição de conceitos referente a Cosmologia. O Capítulo 3 apresenta, também de forma introdutória, conceitos de Cosmologia e Turbulência visando a contextualização do trabalho. No Capítulo 4 vamos apresentar os principais conceitos da Computação de Alto Desempenho e da abordagem em computação em grade que foi utilizada no desenvolvimento desta tese. Uma descrição dos dados que foram utilizados no trabalho é apresentada no Capítulo 5. A descrição dos algoritmos implementados para análise é feita no Capítulo 6. No Capítulo 7 discutimos os re-

sultados obtidos e ,finalmente, no Capítulo 8 apresentamos as conclusões e sugestões de trabalhos futuros.

2 CONCEITOS PRELIMINARES DE COSMOLOGIA

Cosmologia (do grego $\kappa o \sigma \mu o \lambda o \gamma \iota \alpha$, $\kappa o \sigma \mu o$ =“**ordem, tudo**” + $\lambda o \gamma \iota \alpha$ = “**discurso, estudo**”) é a ciência que estuda a origem, a estrutura, a composição e a evolução do Universo. O estudo do Universo tem uma longa história envolvendo a Física, a Astronomia, a Filosofia, o Esoterismo e a Religião, de modo que seu entendimento tem sido alterado ao longo do tempo.

Por um longo tempo, prevaleceu a visão aristotélica, ou seja, havia uma física para a Terra e outra física para os corpos celestiais. Grandes cientistas como Nicolau Copérnico, Galileu Galilei, Giordano Bruno, Tycho Brahe, Johannes Kepler, entre outros, contribuíram enormemente à compreensão da dinâmica dos corpos terrestres e celestes, cabendo a Isaac Newton a organização desses resultados em suas Leis de Movimento e sua Lei da Gravitação Universal, que culminaram numa completa mudança de paradigma. Segundo a Lei da Gravitação Universal, todos os objetos no Universo que possuem massa exercem uma força de atração gravitacional entre si, mesmo que estejam separados por uma grande distância. Essa força de atração é proporcional a massa de cada um deles e inversamente proporcional ao quadrado da distância que os separa. Matematicamente, a lei da gravitação universal pode ser escrita como:

$$F = G\frac{Mm}{d^2} \tag{2.1}$$

em que G é a constante gravitacional, M e m são as massas dos corpos que estão interagindo e d é a distância entre eles. Essa lei é capaz de explicar o movimento dos planetas, o movimento das marés e a queda de um objeto qualquer na superfície da terra.

O modelo cosmológico de Newton era de um Universo estático, homogêneo e infinito (se assim não fosse, o sistema todo iria colapsar sobre um centro, o que contrariava a observação de um Universo aparentemente estável). Mas, o século XX desafiou o modelo Newtoniano e uma nova interpretação surgiu com o desenvolvimento, por Albert Einstein em 1915, da teoria da relatividade geral, que relaciona a curvatura do espaço-tempo com o seu conteúdo de massa e energia. Posteriormente, em 1917, Einstein aplicou suas equações ao Universo como um todo, porém não conseguiu estabelecer uma configuração estática do mesmo. Assim, resolveu adicionar em suas equações um termo, a chamada constante cosmológica (Λ), que representa uma força repulsiva e com a qual era possível obter um Universo estático. Entretanto, a su-

posição de Universo estático de Einstein resultou não estar correta; alguns anos depois dados de observação aliados a novos desenvolvimentos teóricos permitiram uma das mais importantes descobertas da astronomia no século XX: o Universo está em expansão (MADSEN, 1996; LINDER, 2003; LONGAIR, 2008).

O primeiro modelo teórico satisfatório para um Universo em expansão foi proposto em 1922 por Alexander Friedman (FRIEDMAN, 1922). Nesse trabalho Friedman publicou um conjunto de soluções matemáticas para as equações da teoria relativística da gravitação, que mostravam que o Universo devia estar em expansão ou contração. As equações de Friedman governam a expansão métrica do espaço em modelos homogêneos e isotrópicos do Universo dentro do contexto da teoria da relatividade geral. Embora o trabalho de Friedman tenha sido publicado em uma importante revista, não causou grande impacto na sociedade astronômica na época. Em 1927, o astrofísico e matemático Georges Lemaître chegou de forma independente aos mesmos resultados de Friedman.

Como citado acima, a mudança de paradigma não se deu com o desenvolvimento de modelos teóricos e sim com a verificação observacional da expansão cósmica. Neste cenário existem 3 figuras que merecem destaque: Vesto Melvin Slipher, Milton Humason e Edwin Powell Hubble. Slipher foi o primeiro a obter, em 1912, medidas de velocidade de objetos que depois viriam a ser chamados de galáxias. Ele notou que os espectros dessas galáxias apresentavam grandes deslocamentos para regiões de maior comprimento de onda, ou seja, para o vermelho, o que significava que se moviam com grandes velocidades em relação a nós (Efeito Doppler). A galáxia de Andrômeda, por exemplo, mostrou estar aproximando-se da Via-Láctea. Na década seguinte, Milton Humason e Edwin Huble ampliaram bastante o trabalho de Slipher. Interessantemente, eles notaram que praticamente todas as outras galáxias estavam afastando-se de nós. Mais que isso, eles puderam obter distâncias para algumas dessas galáxias e constataram que havia uma relação linear entre esses dois parâmetros: velocidade de recessão e distância (HUBBLE, 1929), conhecida atualmente como Lei de Hubble. Cabe notar que a interpretação dos desvios para o vermelho como Efeito Doppler não é obvia, e muitos cientistas, ao longo dos mais de 80 anos que se passaram desde então, têm discutido e sugerido propostas alternativas. Nomes como Fred Hoyle, Halton Arp e Hannes Alfvén estão entre os importantes cientistas que criticaram essa interpretação e suas consequências para o Modelo do *Big Bang* Quente (veja, por exemplo, Assis (1992), Assis e Neves (1995), para uma discussão detalhada

sobre esse assunto).

Uma das consequências físicas mais importantes da descoberta da expansão do Universo é que, se as galáxias estão se afastando uma das outras, então no passado elas deveriam estar todas juntas, ou seja, tudo que existe teria partido de uma enorme concentração de matéria inicial. Seguindo esse raciocínio foi possível desenvolver a teoria do *Big Bang*, que afirma que o Universo surgiu a partir de um estado primordial extremamente denso e quente (KOLB; TURNER, 1990; LIDDLE; LYTH, 2000; LEVIN, 2007).

No final da década de 90, por meio de medições da relação magnitude-*redshift* para supernovas tipo *Ia*, os cosmólogos descobriram que o Universo está em uma expansão acelerada. Para explicar essa aceleração a teoria da relatividade geral exige que grande parte da energia no Universo consista de uma componente com grande pressão negativa e assim introduziram um novo conceito denominado energia escura (PEEBLES; RATRA, 2002; PADMANABHAN, 2006b), que seria uma forma de energia gravitacionalmente repulsiva.

A gravitação possibilita a criação de uma hierarquia de estruturas cosmológicas. As galáxias se juntam em grupos ou aglomerados, esses se juntam em filamentos e superaglomerados, numa escala de aglomeração que vai de dezenas a milhares de galáxias ligadas pela gravitação. A dinâmica composta pela força da gravidade e pela expansão do Universo descreve a história da formação destas estruturas (PADMANABHAN, 1993; MADSEN, 1996).

Antes de apresentarmos os dados das simulações e os dados observacionais que foram utilizados nesta tese é importante definir alguns conceitos e parâmetros que são utilizados nos atuais modelos cosmológicos.

2.1 Parâmetros Cosmológicos e Medidas de Distâncias

No trabalho realizado por Hubble (HUBBLE, 1929), que permitiu a comprovação observacional da expansão do Universo, foi descoberto que a velocidade de recessão ($\mathbf{v}$) de galáxias distantes é diretamente proporcional a sua distância (d), ou seja, $\mathbf{v} = H_0 d$. A constante de proporcionalidade H é chamada constante de Hubble e o subescrito "0" refere-se a época atual, uma vez que essa constante varia com o tempo. Normalmente esse parâmetro é escrito da seguinte maneira: $H_0 = 100\ h\ km\ s^{-1}\ Mpc^{-1}$. Segundo Freedman et al. (2001), o *Key Project* do

telescópio espacial Hubble, por meio do estudo do período de variação da luminosidade de estrelas cefeidas entre outros métodos, determinou $h = 0.72 \pm 0.08$. Um outro valor considerado frequentemente na literatura recente é $h = 0.71 \pm 0.03$, dos resultados de 7 anos do *Wilkinson Microwave Anisotropy Probe* (WMAP) (JAROSIK et al., 2011). A constante de Hubble está relacionada ao *fator de escala a* da seguinte forma:

$$H = \frac{\dot{a}}{a} \tag{2.2}$$

A partir da constante de Hubble podemos definir o tempo de Hubble (t_H) que está associado à idade do Universo:

$$t_H \equiv \frac{1}{H_0}. \tag{2.3}$$

Tomando os dois valores de H_0 (71 $km\ s^{-1}\ Mpc^{-1}$ e 72 $km\ s^{-1}\ Mpc^{-1}$) acima, temos que t_H é 13.8 Ga e 13.6 Ga, respectivamente. Porém, considerando também outros fatores que influenciam a estimação, o WMAP encontra a partir do valor de 13.8 Ga que a idade aproximada do Universo é 13.75 Ga.

Uma unidade cosmológica obtida a partir do tempo de Hubble é a distância de Hubble que é definida como:

$$D_H \equiv \frac{c}{H_0} \equiv 3000\ h^{-1}\ Mpc \tag{2.4}$$

em que c é a velocidade da luz.

Usando a presente taxa de expansão, podemos definir a densidade critica ρ_c, correspondente a um Universo plano (PADMANABHAN, 1993).

$$\rho_c \equiv \frac{3H_0^2}{8\pi G} = 11.26\ h^2\ protons/m^3$$

Um parâmetro cosmológico muito importante e muito utilizado na cosmologia é o desvio para o vermelho ou *redshift*. O *redshift* (z) de um objeto é medido como o deslocamento relativo do comprimento de onda emitido (λ_e) pela fonte e o comprimento de onda observado (λ_o), conforme Equação 2.5:

$$z = \frac{\lambda_o - \lambda_e}{\lambda_e} \tag{2.5}$$

Se $\mathbf{v} \ll c$ e o *redshift* é interpretado em termos da velocidade de recessão da galáxia,

então $\mathbf{v} = cz$.

2.1.1 Medidas de Distâncias Cosmológicas

Devido a expansão do Universo, a distância entre objetos é uma quantidade dinâmica e depende da geometria do espaço-tempo. Vamos descrever as principais distâncias utilizadas na Cosmologia e que foram utilizadas nesta tese, seguindo Hogg (2000).

Para dois objetos próximos em um *redshift* z, a distância co-móvel (linha de visada) entre eles é igual à sua separação própria dividida pelo fator de escala do Universo. Conforme Hogg (2000), a distância co-móvel total é obtida através da seguinte integral:

$$D_C = D_H \int_0^z \frac{dz'}{E(z')} \tag{2.6}$$

em que D_H é a distância de Hubble, $E(z) \equiv \sqrt{\Omega_M(1+z)^3 + \Omega_\kappa(1+z)^2 + \Omega_\Lambda}$, o parâmetro Ω_M representa a densidade de massa, o parâmetro de densidade Ω_Λ representa a densidade da constante cosmológica e Ω_κ é o parâmetro de densidade que mede a curvatura do espaço. Considerando o Universo homogêneo, isotrópico e dominado pela matéria, estes parâmetros determinam sua geometria e são definidos como segue:

$$\Omega_M \equiv \frac{8\pi G\rho}{3H^2} \tag{2.7}$$

$$\Omega_\Lambda \equiv \frac{\Lambda c^2}{3H^2} \tag{2.8}$$

$$\Omega_\kappa + \Omega_M + \Omega_\Lambda = 1 \tag{2.9}$$

A partir da distância co-móvel (D_C), podemos obter todas as outras distâncias utilizadas. A distância co-móvel (transversal) (D_M) pode ser obtida através da seguinte relação:

$$D_M = \begin{cases} D_H \frac{1}{\sqrt{\Omega_\kappa}} senh\left[\sqrt{\Omega_k} D_C/D_H\right] & \text{para } \Omega_\kappa > 0 \\ D_C & \text{para } \Omega_\kappa = 0 \\ D_H \frac{1}{\sqrt{|\Omega_\kappa|}} sen\left[\sqrt{|\Omega_k|} D_C/D_H\right] & \text{para } \Omega_\kappa < 0 \end{cases}$$

A distância do diâmetro angular (D_A) é definida como sendo a razão entre o tamanho transversal de um objeto pelo seu tamanho angular (em radianos). A distância do diâmetro angular está relacionada com a distância co-móvel (D_M) pela equação dada

a seguir:

$$D_A = \frac{D_M}{1+z} \tag{2.10}$$

Considerando dois objetos em diferentes *redshifts* a distância entre eles é obtida usando a Equação 2.11:

$$D_{A_{12}} = \frac{1}{1+z_2}\left[D_{M2}\sqrt{1+\Omega_\kappa \frac{D_{M_1}^2}{D_H^2}} - D_{M1}\sqrt{1+\Omega_\kappa \frac{D_{M_2}^2}{D_H^2}}\right] \tag{2.11}$$

em que D_{M1} e D_{M2} são as distâncias co-móveis para z_1 e z_2, D_H é a distância de Hubble e Ω_κ é o parâmetro densidade de curvatura. Por último vamos definir a distância de luminosidade, que é dada pela raiz quadrada da razão entre a luminosidade e o fluxo bolométrico.

$$D_L = \sqrt{\frac{L}{4\pi S}} \tag{2.12}$$

A distância D_L está relacionada a D_M e a D_A através da Equação 2.13:

$$D_L = (1+z)\, D_M = (1+z)^2\, D_A \tag{2.13}$$

2.1.2 Teorema do Virial e Estimação de Massas

Os métodos diretos de medição de massas em astronomia são dinâmicos. Assim, para sistemas tais como aglomerados de estrelas, galáxias e aglomerados de galáxias geralmente assume-se que o sistema já alcançou o equilibrio dinâmico e são considerados como configurações ligadas gravitacionalmente. A partir das dimensões do sistema e das velocidades dos objetos que o constitue pode-se estimar a sua massa. Neste sentido, um resultado chave para determinar as massas de galáxias e de aglomerados de galáxias muito utilizado é o teorema do Virial (e.g. Heisler et al. (1985)).

O teorema do virial estabelece que existe uma relação de equilíbrio entre as energias médias das partículas em um sistema, quando esse sistema esta relaxado dinamicamente, de modo que: $2K = U$, sendo K a energia cinética e U a energia potencial gravitacional. Vamos deduzir a forma do teorema do Virial para um sistema autogravitante de massas pontuais seguindo Longair (2008).

Considere um sistema de partículas (estrelas ou galáxias) cada uma com massa m_i, interagindo entre si somente pela força de atração gravitacional. Então, a aceleração da *i-ésima* partícula devido a todas as outras partículas pode ser escrita vetorialmente como segue:

$$\ddot{\mathbf{r}}_i = \sum_{i \neq j} \frac{G m_j (\mathbf{r}_j - \mathbf{r}_i)}{|\mathbf{r}_i - \mathbf{r}_j|^3} \tag{2.14}$$

Tomando o produto escalar de ambos os lados com $m_i \mathbf{r}_i$, tem-se:

$$m_i(\mathbf{r}_i \cdot \ddot{\mathbf{r}}_i) = \sum_{i \neq j} G m_i m_j \frac{\mathbf{r}_i \cdot (\mathbf{r}_j - \mathbf{r}_i)}{|\mathbf{r}_i - \mathbf{r}_j|^3} \tag{2.15}$$

Derivando $(\mathbf{r}_i \cdot \mathbf{r}_i)$ em relação ao tempo:

$$\frac{d}{dt}(\mathbf{r}_i \cdot \mathbf{r}_i) = 2\dot{\mathbf{r}}_i \cdot \mathbf{r}_i \tag{2.16}$$

e então, pegando a segunda derivada da Equação 2.16, tem-se:

$$\frac{1}{2}\frac{d^2}{dt^2}(\mathbf{r}_i^2) = \frac{d}{dt}(\dot{\mathbf{r}}_i \cdot r_i) = (\ddot{\mathbf{r}}_i \cdot \mathbf{r}_i + \dot{\mathbf{r}}_i \cdot \dot{\mathbf{r}}_i) = (\ddot{\mathbf{r}}_i \cdot \mathbf{r}_i + \dot{\mathbf{r}}_i^2) \tag{2.17}$$

Desta forma a Equação 2.15 pode ser escrita como:

$$\frac{1}{2}\frac{d^2}{dt^2}(m_i \mathbf{r}_i^2) - m_i \dot{\mathbf{r}}_i^2 = \sum_{i \neq j} G m_i m_j \frac{\mathbf{r}_i \cdot (\mathbf{r}_j - \mathbf{r}_i)}{|\mathbf{r}_i - \mathbf{r}_j|^3} \tag{2.18}$$

Agora vamos somar todas as partículas do sistema,

$$\frac{1}{2}\frac{d^2}{dt^2}\sum_i m_i \mathbf{r}_i^2 - \sum_i m_i \dot{\mathbf{r}}_i^2 = \sum_i \sum_{i \neq j} G m_i m_j \frac{\mathbf{r}_i \cdot (\mathbf{r}_j - \mathbf{r}_i)}{|\mathbf{r}_i - \mathbf{r}_j|^3} \tag{2.19}$$

A soma dupla do lado direito representa a soma sobre todos os elementos de uma matriz quadrada $n \times n$ com todos os termos da diagonal iguais a zero. Se somarmos os elementos ij e ji da matriz, encontramos:

$$G m_i m_j \left[\frac{\mathbf{r}_i \cdot (\mathbf{r}_j - \mathbf{r}_i)}{|\mathbf{r}_i - \mathbf{r}_j|^3} + \frac{\mathbf{r}_j \cdot (\mathbf{r}_i - \mathbf{r}_j)}{|\mathbf{r}_j - \mathbf{r}_i|^3} \right] = -\frac{G m_i m_j}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{2.20}$$

e desta forma:

$$\frac{1}{2}\frac{d^2}{dt^2}\sum_i m_i \mathbf{r}_i^2 - \sum_i m_i \dot{\mathbf{r}}_i^2 = -\frac{1}{2}\sum_{j \neq i} \frac{G m_i m_j}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{2.21}$$

Note que $\sum_i m_i \dot{\mathbf{r}}_i^2$ é duas vezes a energia cinética (K) de todas as partículas no sistema, isto é:

$$K = \frac{1}{2} \sum_i m_i \dot{\mathbf{r}}_i^2 \tag{2.22}$$

A energia potencial gravitacional do sistema é:

$$U = -\frac{1}{2} \sum_{j \neq i} \frac{G m_i m_j}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{2.23}$$

Desta forma,

$$\frac{1}{2} \frac{d^2}{dt^2} \sum m_i \mathbf{r}_i^2 = 2K - |U| \tag{2.24}$$

Se o sistema está em equilíbrio estatístico

$$\frac{d^2}{dt^2} \sum_i m_i \mathbf{r}_i^2 = 0 \tag{2.25}$$

E então:

$$K = \frac{1}{2}|U| \tag{2.26}$$

A igualdade da Equação 2.26 é conhecida como teorema do Virial. A partir desse teorema pode-se obter uma relação utilizada para a estimação de massas, conforme (HEISLER et al., 1985), que é dada pela equação .

$$M \approx \frac{3\pi N}{2G} \frac{\sum_i V_i^2}{\sum_{i<j} 1/R_{ij}} \tag{2.27}$$

2.2 O Modelo Cosmológico Padrão

Segundo o Modelo Cosmológico Padrão (MCP) ou modelo do *Big Bang*, o Universo se expande a partir de um estado primordial, onde sua massa era dominada por uma radiação térmica de corpo negro (PEEBLES, 1993). O MCP está baseado na Teoria da Relatividade Geral, que é a moderna teoria da gravitação, no princípio cosmológico que afirma que em grandes escalas o Universo é isotrópico e homogêneo e na lei de Hubble, descoberta por Edwin Powell Hubble em 1929, que estabelece a expansão do Universo.

O princípio cosmológico implica que o Universo deve ser descrito pela métrica de

Friedmann-Lemaître-Robertson-Walker (FLRW), dada pela seguinte equação:

$$ds^2 = c^2dt^2 - a^2(t)\left[\frac{dr^2}{1-\kappa r^2} + r^2d\theta^2 + r^2sin^2\theta d\phi^2\right] \tag{2.28}$$

onde r, θ, ϕ são coordenadas espaciais co-móveis, t é o tempo, $a(t)$ é o fator de escala cósmica que mede a taxa de expansão universal (Por convenção, $a(hoje) = 1$). A quantidade κ é a curvatura do espaço, sendo que $\kappa = -1$ corresponde a um Universo aberto, $\kappa = 0$ corresponde a um Universo plano e $\kappa = +1$ corresponde a um Universo fechado (PADMANABHAN, 1993; HAWLEY; HOLCOMB, 2005).

Neste contexto, as equações principais da cosmologia que descrevem a expansão do Universo são as equações de Friedmann, dadas por:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{\kappa c^2}{a^2} + \frac{\Lambda c^2}{3} \tag{2.29}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + \frac{3p}{c^2}) + \frac{\Lambda c^2}{3} \tag{2.30}$$

em que ρ é a densidade, p é a pressão, Λ é a constante cosmológica e c a velocidade da luz.

Segundo o MCP, o Universo foi criado há cerca de 14 bilhões de anos, evoluindo e se expandindo a partir de um estado inicial extremamente quente e denso. Durante sua expansão, a densidade e a temperatura média do plasma primordial diminuíram o suficiente para permitir que elétrons e prótons pudessem se combinar para dar origem aos primeiros átomos de hidrogênio. O "plasma primordial" era composto por: quarks, antiquarks, elétrons, pósitrons, neutrinos e fótons - antes da era da bariosíntese (formação de prótons e nêutrons) e de prótons, nêutrons, elétrons, neutrinos e fótons - antes da era da recombinação, ou seja, da formação de átomos neutros.

Entre essas duas etapas ocorreram muitos fenômenos, como a aniquilação de antiquarks, a bariosíntese, a nucleosíntese (formação de núcleos de *helio*) e o equilíbrio entre as densidades de matéria e radiação. Depois da nucleosíntese aconteceu o equilíbrio entre matéria e radiação porque as curvas de densidade das duas se cruzaram (a densidade de radiação antes era maior, mas cai mais rapidamente que a densidade de matéria). Depois disso, a matéria se viu livre para formar átomos neutros (juntar núcleos de *hidrogêneo* e *hélio* com seus respectivos elétrons). A radiação, por sua vez, se desacoplou da matéria e ficou "livre", formando a radiação cósmica de fundo.

Este processo, entretanto, não tornou o Universo transparente à radiação eletromagnética, ao contrário, tornou o Universo opaco (porque a radiação não podia avançar muito sem encontrar um átomo neutro e ionizá-lo).

Muito tempo depois, na época que chamamos de reionização ($z \sim 10$, em contraste com a recombinação que foi em $z \sim 1500$) os primeiros objetos (estrelas, quasares, etc) reuniram uma radiação suficiente para novamente ionizar parcialmente o Universo, tornando-o transparente, ou seja, depois disso os fótons podem viajar a grandes distâncias sem ser absorvidos.

Atualmente se pode ver os efeitos provocados pelos processos físicos que ocorreram neste estágio por meio das observações da radiação cósmica de fundo. A radiação cósmica de fundo correspondente a radiação de corpo negro em $T = 2.725\ K$, descoberta acidentalmente por Arno Penzias e Robert Wilson, em 1965, cuja isotropia média dá suporte ao princípio cosmológico e cujas pequenas anisotropias na distribuição espacial das temperaturas contêm uma riqueza de informações sobre os parâmetros cosmológicos. No modelo do *Big Bang*, quando houve o desacoplamento da radiação cósmica de fundo, sua temperatura era aproximadamente 3 mil vezes maior que agora. A temperatura de radiação cósmica de fundo varia com o *redshift* z da seguinte forma:

$$T_r = 2.728\,(1 + z) \tag{2.31}$$

Pela Equação 2.31 tem-se que no *redshift* $z = 1500$ a temperatura da radiação era aproximadamente $4000\ K$. Nesta época, conhecida como época da recombinação, as galáxias não haviam sido formadas e toda a matéria bariônica, que viria a se tornar a matéria visível de galáxias, ainda estava sob a forma de um gás pré-galáctico, extraordinariamente suave e parcialmente ionizado (LONGAIR, 2008).

2.3 Formação de Estruturas

O modelo padrão da Cosmologia para a formação e evolução das estruturas observadas atualmente (galáxias, aglomerados e superaglomerados de galáxias) está baseado na teoria da instabilidade gravitacional, cujo modelo básico é o chamado colapso esférico, que foi proposto por Jeans (1902). Neste modelo considera-se que tais estruturas surgiram a partir da evolução de pequenas perturbações de densidade. Com a expansão do espaço-tempo estas perturbações podem crescer ou desaparecer, dependendo de seu contraste de densidade, sua composição (em matéria escura,

matéria bariônica e radiação) e das propriedades cosmológicas do Universo em cada momento. Se uma destas instabilidades atinge um valor crítico ela deixa de expandir-se com o Universo e entra em colapso gravitacional. Sua evolução de aí em diante é dominada por processos não-lineares.

Os mecanismos para a geração destas perturbações ainda não são compreendidos. Acredita-se que o modelo inflacionário (modelo no qual o universo passa por um breve período de expansão exponencial no início de sua história) por si só seja capaz de explicar o surgimento destas perturbações. De acordo com o cenário inflacionário, tais perturbações tiveram origem em flutuações quânticas de um campo associado a algum tipo de partícula elementar (KOLB; TURNER, 1990; HAWLEY; HOLCOMB, 2005; LIDDLE; LYTH, 2000).

A busca pelo entendimento da evolução das instabilidades que surgiram no Universo primordial é feita através da teoria da perturbação, onde são consideradas pequenas flutuações nas grandezas físicas constituintes do Universo. Os efeitos relativísticos devido a curvatura do espaço-tempo podem ser desprezados em escalas de tamanho menores que a distância de Hubble e então, pode-se utilizar a cosmologia Newtoniana em que o movimento da matéria é descrito pelas equações da hidrodinâmica e a expansão cósmica é dada pelas equações de Friedmann (PADMANABHAN, 1993; TATEKAWA, 2005).

2.3.1 Equações Hidrodinâmicas e a Teoria da Perturbação

O Universo ao nosso redor é altamente heterogêneo, sendo constituído por estrelas, planetas, galáxias, super-aglomerados de galáxias, ao invés de um fluido harmoniosamente distribuído com densidade de massa ρ. Entender a formação dessas estruturas em todas as escalas do Universo é um dos mais facinantes desafios da astronomia moderna. Conforme Chabrier (2009), acredita-se que os blocos da contrução inicial das galáxias foram pequenos halos de matéria escura em colapso, produzidos por flutuações primordiais. Esses blocos se fundiram e galáxias maiores foram sendo formadas progressivamente em um esquema geralmente descrito como modelo hierárquico de formação de galáxias.

Suponhamos que em algum momento inicial, por exemplo na época do desacoplamento, existiram pequenas irregularidades na distribuição da matéria. As regiões com mais matéria vão exercer uma maior força gravitacional sobre as regiões vizi-

nhas e portanto tendem a atrair a materia dessas regiões. Essa materia extra torna cada vez mais densa a região inicialmente com mais matéria, aumentando a sua atração gravitacional de modo que cada vez mais será atraida mais matéria. Nesse sentido, uma distribuição irregular de matéria é instável sob a influência da gravidade, tornando-se cada vez mais irregular a medida que o tempo passa.

Essa instabilidade é exatamente o que seria necessário para explicar a observação de que o Universo é muito mais irregular agora do que na época do desacoplamento, portanto ela é quase universalmente aceita como sendo a principal influência que leva à formação de estruturas no Universo. No entanto, apesar de que a força da gravidade tem papel fundamental neste mecanismo, existem outros processos que também desempenham papéis importantes e que são caracterizados principalmente por processos não-lineares (LIDDLE, 2003; LONGAIR, 2008).

Para entender a natureza e a evoluçao das estruturas em grande escala é usual primeiro investigar a dinâmica do Universo como um todo, ou seja, considerando-o homogêneo, em seguida trata-se a não homogeneidade como pequenas flutuações, utilizando a teoria perturbativa.

Devido a sua simplicidade, o modelo Newtoniano é de grande importância para o estudo da formação de estruturas. Neste contexto, a matéria é tratada como um fluido cósmico que permeia todo o Universo e o movimento deste fluido é descrito pela equação da continuidade, equação de Euler e equação de Poisson, dadas respectivamente pelas Equações 2.32, 2.33 e 2.34 :

$$\frac{\partial \rho}{\partial r} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{2.32}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\, \mathbf{v} = -\frac{1}{\rho} \nabla p - \nabla \Phi \tag{2.33}$$

$$\nabla^2 \Phi = 4\pi G \rho \tag{2.34}$$

com: Φ é o potencial gravitacional Newtoniano, G é a constante universal de Newton, ρ a densidade de massa, p é a pressão e $\mathbf{v}$ é o vetor velocidade local do fluido

A derivada total em coordenadas langrangeanas é dada por:

$$\frac{d}{dt} = \frac{\partial}{\partial t} + (\mathbf{v} \cdot \nabla) \tag{2.35}$$

Usando a Equação 2.35 e a identidade $\nabla \cdot (\rho \mathbf{v}) = \rho \nabla \cdot \mathbf{v} + \mathbf{v} \cdot \nabla \rho$ as equações do movimento podem ser escritas em coordenadas Lagrangeanas da seguinte forma:

$$\frac{d\rho}{dt} = -\rho \nabla \cdot \mathbf{v} \tag{2.36}$$

$$\frac{d\mathbf{v}}{dt} = \frac{-1}{\rho} \nabla p - \nabla \Phi \tag{2.37}$$

$$\nabla^2 \Phi = 4\pi G \rho \tag{2.38}$$

Assumindo que $\mathbf{v_0}$, ρ_0, p_0, ϕ_0 são soluções das Equações 2.36 - 2.38 para um fluido isotrópico e homogêneo em expansão, teremos:

$$\frac{d\rho_0}{dt} = -\rho_0 \nabla \cdot \mathbf{v}_0 \tag{2.39}$$

$$\frac{d\mathbf{v}_0}{dt} = \frac{-1}{\rho_0} \nabla p_0 - \nabla \Phi_0 \tag{2.40}$$

$$\nabla^2 \Phi_0 = 4\pi G \rho_0 \tag{2.41}$$

Para obter a evolução das flutuações, vamos considerar pequenas perturbações nas variáveis ρ, p, Φ e $\mathbf{v}$, conforme segue:

$$\mathbf{v} = \mathbf{v}_0 + \delta \mathbf{v}; \rho = \rho_0 + \delta \rho; p = p_0 + \delta p; \Phi = \Phi_0 + \delta \Phi \tag{2.42}$$

Em seguida substituimos a solução perturbada nas Equações 2.36 - 2.38, desprezamos os termos não lineares e subtraimos as Equações 2.39 - 2.41. Para as Equações 2.36 e 2.39, temos:

$$\frac{d}{dt}\left(\frac{\delta \rho}{\rho_0}\right) = \frac{d\Delta}{dt} = -\nabla \cdot \delta \mathbf{v} \tag{2.43}$$

onde $\Delta = \delta\rho/\rho_0$ é o contraste de densidade. Esta equação é muito importante, uma vez que ela relaciona a taxa em que o contraste de densidade se desenvolve para a velocidade peculiar $\delta \mathbf{v}$ associada ao colapso da perturbação.

Para as Equações 2.37 e 2.40 teremos:

$$\frac{d(\mathbf{v}_0 + \delta \mathbf{v})}{dt} = \frac{\partial(\mathbf{v}_0 + \delta \mathbf{v})}{\partial t} + \mathbf{v} \cdot \nabla(\mathbf{v}_0 + \delta \mathbf{v}) \Longrightarrow$$

$$\frac{d(\mathbf{v}_0 + \delta \mathbf{v})}{dt} = \frac{\partial \mathbf{v_0}}{\partial t} + \frac{\partial \delta \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla(\mathbf{v}_0 + \delta \mathbf{v}) \Longrightarrow$$

$$\frac{d(\mathbf{v}_0+\delta\mathbf{v})}{dt}=\frac{\partial\mathbf{v}_0}{\partial t}+\frac{d\delta\mathbf{v}}{dt}-\mathbf{v}\cdot\nabla\delta\mathbf{v}+\mathbf{v}\cdot\nabla\mathbf{v_0}+\mathbf{v}\cdot\nabla\delta\mathbf{v}\Longrightarrow$$

$$\frac{d(\mathbf{v}_0+\delta\mathbf{v})}{dt}=\frac{\partial\mathbf{v}_0}{\partial t}+\frac{d\delta\mathbf{v}}{dt}+\mathbf{v}\cdot\nabla\mathbf{v}_0$$

lembrando que $\mathbf{v}=\mathbf{v_0}+\delta\mathbf{v}$, tem-se:

$$\frac{d(\mathbf{v}_0+\delta\mathbf{v})}{dt}=\frac{\partial\mathbf{v}_0}{\partial t}+\frac{d\delta\mathbf{v}}{dt}+(\mathbf{v}_0\cdot\nabla)\mathbf{v}_0+(\delta\mathbf{v}\cdot\nabla)\mathbf{v_0} \tag{2.44}$$

Logo, a Equação 2.37 torna-se:

$$\frac{\partial\mathbf{v}_0}{\partial t}+\frac{d\,(\delta\mathbf{v})}{dt}+(\mathbf{v}_0\cdot\nabla)\mathbf{v}_0+(\delta\mathbf{v}\cdot\nabla)\mathbf{v}_0=-\frac{1}{\rho_0+\delta\rho}\nabla\,(p_0+\delta p)-\nabla\,(\Phi_0+\delta\Phi) \tag{2.45}$$

Assumindo que no estado inicial o Universo é homogêneo e isotrópico, temos que $\nabla p_0=0$ e $\nabla\Phi_0=0$, lembrando que:

$$\frac{d\mathbf{v_0}}{dt}=\frac{\partial\mathbf{v_0}}{\partial t}+(\mathbf{v_0}\cdot\nabla)\mathbf{v_0}$$

logo, subtraindo 2.40 de 2.45 teremos:

$$\frac{d(\delta\mathbf{v})}{dt}+(\delta\mathbf{v}\cdot\nabla)\mathbf{v}_0=-\frac{1}{\rho_0}\nabla\delta p-\nabla\delta\Phi \tag{2.46}$$

Finalmente, a terceira equação resulta da subtração de 2.41 de 2.38 e é dada por:

$$\nabla^2\delta\Phi=4\pi G\delta\rho \tag{2.47}$$

As Equações 2.43, 2.46 e 2.47 são as equações diferenciais principais nesta análise. No caso cosmológico, como o Universo está em expansão, vamos introduzir coordenadas co-móveis, de modo que: $\mathbf{x}=a(t)\mathbf{r}$, onde $\mathbf{r}$ é a coordenada de distância co-móvel e $a(t)$ é o fator de escala. Escrevendo as derivadas em relação a coordenada co-móvel $\mathbf{r}$ em lugar de $\mathbf{x}$, de modo que $\frac{d}{dx}=\frac{1}{a}\frac{d}{dr}$, teremos:

$$\delta\mathbf{x}=\frac{1}{a(t)}\delta[a(t)\mathbf{r}]=\frac{1}{a(t)}\mathbf{r}\delta a(t)+\delta\mathbf{r} \tag{2.48}$$

A velocidade pode então ser escrita da seguinte forma:

$$\mathbf{v}=\frac{1}{a(t)}\frac{da}{dt}\mathbf{r}+\frac{d\mathbf{r}}{dt} \tag{2.49}$$

Assim, podemos identificar $\mathbf{v}_0$ com o termo da expansão de Hubble $\frac{\dot{a}}{a}\mathbf{r}$ e $\frac{d\mathbf{r}}{dt}$ com $\delta\mathbf{v}$. Em outras palavras, o segundo termo está associado a variaç ao na coordenada de distância co-móvel sob a influência da gravidade e do gradiente de pressão. É conveniente escrever a perturbação de velocidade como $a(t)\mathbf{u}$ de modo que $\mathbf{u}$ é a velocidade co-móvel. Logo, a Equação 2.46 torna-se:

$$\frac{d(a\mathbf{u})}{dt} + (a\mathbf{u}\cdot\nabla)\left(\frac{\dot{a}}{a}\right)\mathbf{r}_0 = -\frac{1}{\rho_0}\nabla\delta p - \nabla\delta\Phi \tag{2.50}$$

As derivadas em relação as coordenadas co-móveis serão indicadas por ∇_c. Desta forma, uma vez que $(a\mathbf{u}\cdot\nabla)\dot{a}\mathbf{r} = \mathbf{u}\dot{a}$, a Equação 2.50 torna-se:

$$\frac{d\mathbf{u}}{dt} + 2\left(\frac{\dot{a}}{a}\right)\mathbf{u} = -\frac{1}{\rho_0 a^2}\nabla_c\delta p - \frac{1}{a^2}\nabla_c\delta\Phi \tag{2.51}$$

Agora, consideramos perturbações adiabáticas em que as perturbações na pressão e densidade são relacionadas à velocidade adiabática do som, ou seja, $\partial p/\partial\rho = c_s^2$. Assim, δp pode ser substituido por $c_s^2\delta\rho$ em 2.51. Combinando as Equações 2.43 e 2.50 e pegando a divergência em coordenadas co-móveis de 2.51 e a derivada temporal de 2.43 tem-se:

$$\nabla_c\cdot\dot{\mathbf{u}} + 2\left(\frac{\dot{a}}{a}\right)\nabla_c\cdot\mathbf{u} = -\frac{c_s^2}{\rho_0 a^2}\nabla_c^2(\delta\rho) - \frac{1}{a^2}\nabla_c^2(\delta\Phi) \tag{2.52}$$

$$\frac{d^2}{dt^2}\left(\frac{\delta\rho}{\rho}\right) = -\nabla_c\cdot\dot{\mathbf{u}} \tag{2.53}$$

Desta forma,

$$\frac{d^2\Delta}{dt^2} + 2\left(\frac{\dot{a}}{a}\right)\frac{d\Delta}{dt} = \frac{c_s^2}{\rho_0 a^2}\nabla_c^2\delta\rho + 4\pi G\delta\rho \tag{2.54}$$

Buscando soluções do tipo onda para Δ de forma que $\Delta \propto exp\left[i(\mathbf{k}_c.\mathbf{r} - \omega t)\right]$, logo pode-se deduzir a equação da onda para Δ:

$$\frac{\partial^2\Delta}{\partial t^2} + 2H\frac{\partial\Delta}{\partial t} = \left(\frac{c_s^2}{\rho_0 a^2}\nabla_c^2\rho_0 + 4\pi G\rho_0\right)\Delta \tag{2.55}$$

com $\mathbf{k}_c$ o vetor onda em coordenadas co-móveis.

A Equação 2.55 é uma equação dinâmica para a evolução do contraste de densidade (Δ) em um Universo em expansão, que é válida para o modelo de Universo de Einstein-de Sitter. Muitas soluções foram desenvolvidas para o crescimento das

perturbações de densidade e uma análise intensiva pode ser encontrada em (PADMANABHAN, 1993). Para o caso do modelo cosmológico padrão com: $\Omega_0 = 0.3$ e $\Omega_\Lambda = 0.7$ o crescimento das perturbações de densidade durante a era pós-recombinação depende do fator de escala, ou *redshift*, da seguinte maneira:

$$\Delta = \frac{\delta\rho}{\rho} \propto \frac{1}{1+z} \tag{2.56}$$

No caso de modelos com $\Omega_\Lambda = 0$, o crescimento é muito menor. Em *redshifts* menores que $1/\Omega_0$, a instabilidade cresce muito mais lentamente e no limite $\Omega_0 = 0$ não cresce nada. No entanto, essas flutuações de densidade de Matéria Escura têm que ser maior que 10^{-3} porque senão as flutuações de bárions, que em $z \approx 1000$ tinham uma amplitude de $\approx 10^{-6}$, não poderiam hoje ter formado estrelas, galáxias, etc.

3 TURBULÊNCIA E COSMOLOGIA

A maioria dos fluxos que ocorrem na natureza são turbulentos e por isso a turbulência é um tópico de grande interesse na comunidade científica. Podemos dizer que o escoamento turbulento é um escoamento que é desordenado no tempo e no espaço. Lesieur (2008), destaca que um escoamento turbulento deve ser imprevisível, no sentido de que uma pequena incerteza quanto ao seu conhecimento em um determinado momento inicial irá se ampliar de forma a tornar impossível uma previsão determinística precisa de sua evolução.

3.1 Uma Introdução a Turbulência

Segundo Tennekes e Lumley (1972), as principais características dos escoamentos turbulentos são: Irregularidade, difusividade, altos números de Reynolds, flutuações de vorticidades tri-dimensionais e dissipação.

Irregularidade: Todos os escoamentos turbulentos são irregulares ou aleatórios. Isto faz com que seja impossível obter uma aproximação determinística para problemas turbulentos, sendo então, necessário a utilização de métodos estatísticos.

Difusividade: A difusividade da turbulência, que causa uma rápida mistura e aumenta a taxa de momento, de aquecimento e a transferência de massa é uma outra importante característica dos escoamentos turbulentos. Se um escoamento padrão tem aspecto aleatório, mas não exibe difusão de velocidade nos fluidos que estão ao seu redor, então não é um escoamento turbulento.

Altos números de Reynolds: Escoamentos turbulentos sempre ocorrem em altos números de Reynolds. A turbulência frequentemente se origina como uma instabilidade de um fluxo laminar se o número de Reynolds torna-se muito grande. As instabilidades são relacionadas a interação entre os termos viscosos e os termos inerciais não lineares na equação do momento. Deve ser enfatizado que o conceito de *transição para a turbulência* ainda não está bem compreendido. O mais famoso experimento sobre esta transição é o experimento de Reynolds de um fluxo em um tubo circular (fluxo de Poiseuille). Seja $\overline{v}$ uma velocidade média do fluxo através da seção do tubo, D o diâmetro do tubo e μ a viscosidade dinâmica do fluido. Reynolds introduziu em 1883 o parâmetro adimensional, conhecido como número de Reynolds

dado por:

$$R_e = \frac{\rho \overline{v} D}{\mu} \tag{3.1}$$

e mostrou experimentalmente que havia um valor crítico de R_e acima do qual o fluxo no interior do tubo tornava-se turbulento. Isto foi feito variando-se independentemente a velocidade $\overline{v}$, o diâmetro D do tubo ou considerando fluidos de várias viscosidades. O valor crítico (R_c) para este experimento foi da ordem de 2000. Para $R_e < R_c$ o escoamento permanece regular (laminar) e para $R_e > R_c$ o escoamento torna-se turbulento. O mesmo experimento repetido em um escoamento de Poiseuille plano (*plane channel*) mostra, como no experimento de Reynolds, a transição para a turbulência em um número de Reynolds crítico da ordem de 2000.

Flutuações da vorticidade tri-dimensional: A turbulência é rotacional, tridimensional e caracterizada por altos níveis de flutuações de vorticidade. A flutuação de vorticidade aleatória que caracteriza a turbulência não pode se manter por si só se as flutuações de velocidade forem bi-dimensionais.

Dissipação: Escoamentos turbulentos são sempre dissipativos. A turbulência necessita um contínuo suprimento de energia para compensar as perdas devido a viscosidade. Se energia não é suprida, a turbulência decai rapidamente.

As principais características dos escoamentos turbulentos não são controladas pelas propriedades moleculares do fluido em que a turbulência ocorre. Uma vez que as equações do movimento não são lineares, cada padrão individual do fluido tem certas características únicas que estão associadas com suas condições iniciais e de contorno. A turbulência não é uma característica do fluido, mas sim do escoamento do fluido. Geralmente é provocada pela não homogeneidade do escoamento, ou seja, se forma a partir de instabilidades do escoamento laminar, formando estruras chamadas vórtices. As condições de contorno são frequentemente responsáveis por tais heterogeneidades. Como regra geral pode-se afirmar que o ingrediente chave envolvido na geração de turbulencia é a heterogeneidade do fluxo médio, isto é, fortes gradientes de velocidades médias e a instabilidade desses gradientes (TENNEKES; LUMLEY, 1972).

Escoamentos turbulentos têm sido investigados por mais de um século, vários autores como, por exemplo, Kolmogorov com seus importantes trabalhos de 1941 e 1962 (KOLMOGOROV, 1991b; KOLMOGOROV, 1991a; KOLMOGOROV, 1962), Monin

e Yaglom (1975), Frisch (1995) e Lesieur (2008) têm dado importantes contribuições ao universo da turbulência, porém ainda não existe uma aproximação geral que possa ser aplicada aos diversos casos de geração e evolução deste fenômeno. Estudos estatísticos das equações do movimento sempre levam a uma situação em que se tem mais incógnitas do que equações. Esse é o chamado problema do fechamento da teoria da turbulência, para o qual as soluções propostas consideram suposições que permitem igualar o número de equações ao número de incógnitas.

3.1.1 Hipóteses de Kolmogorov

Em 1941, Andrey N. Kolmogorov publicou 3 trabalhos que forneceram alguns dos mais importantes resultados da teoria estatística da turbulência. Estes resultados englobam o que agora é conhecido na literatura como teoria K-41, que certamente é a teoria mais famosa e utilizada na análise da turbulência. Um importante resultado diz respeito ao processo de cascata de energia, cuja idéia original foi proposta por Richardson em 1922. Richardson notou que os grandes vórtices são instáveis e se quebram, transferindo sua energia para os pequenos vórtices. Esses pequenos vórtices passam por um processo similar, também se quebrando e transferindo sua energia para os vórtices menores. Esse processo é conhecido como cascata de energia, onde a energia é transferida sucessivamente dos vórtices maiores para os vórtices menores até que o número de Reynolds seja suficientemente pequeno e o movimento torna-se estável com a energia cinética sendo dissipada pelas forças viscosas. Kolmogorov parametrizou esse fenômeno através de uma equação que relaciona o espectro de energia cinética turbulenta com o número de onda.

A teoria K-41 tem por base as definições de turbulência localmente isotrópica e homogênea e as hipóteses de similaridade. Em geral, os grandes vórtices são anisotrópicos e afetados pelas condições de contorno do escoamento, porém, em números de Reynolds suficientemente altos, os movimentos turbulentos em pequenas escalas ($l \ll l_0$) são estatisticamente homogêneos e isotrópicos, onde l_0 é a escala integral (POPE, 2000).

Na cascata de energia os dois processos dominantes são a tranferência de energia para as escalas menores e a dissipação viscosa. Uma hipótese razoável então, é que os parâmetros importantes são a taxa em que as pequenas escalas recebem energia das grandes escalas (denotada por Υ_I) e a viscosidade cinemática ν. Como veremos a taxa de dissipação ε é determinada pela taxa de transferência de energia Υ_I, logo

$\varepsilon \approx \Upsilon_I$. Consequentemente, a hipótese de que o estado estatisticamente universal das pequenas escalas é determinado por ν e pela taxa de transferência Υ_I é a primeira hipótese de Kolmogorov e pode ser estabelecida como segue abaixo (KOLMOGOROV, 1991b; POPE, 2000).

Primeira hipótese de similaridade de Kolmogorov: Em todo escoamento turbulento com número de Reynolds suficientemente grande, a estatística dos movimentos em pequenas escalas ($l < l_0$) tem uma forma universal que é determinada apenas por ν e ε.

Assim, dados os parâmetros ε e ν, existem escalas únicas de tamanho, velocidade e tempo que podem ser formuladas. Estas são as escalas de Kolmogorov:

$$\eta \equiv \left(\nu^3/\varepsilon\right)^{1/4} \tag{3.2}$$

$$u_\eta \equiv (\varepsilon\nu)^{1/4} \tag{3.3}$$

$$\tau_\eta \equiv (\nu/\varepsilon)^{1/2} \tag{3.4}$$

Duas identidades derivadas dessas definições indicam claramente que as escalas de Kolmogorov caracterizam os menores vórtices dissipativos. Primeiro, o número de Reynolds baseado nas escalas de Kolmogorov é unitário, ou seja, $\eta u_\eta/\nu = 1$, o que é consistente com a idéia da cascata, onde a energia decai para menores e menores escalas até o número de Reynolds $u(l)l/\nu$ ser pequeno o suficiente para a dissipação ser efetiva. Segundo, a taxa de dissipação é dada por:

$$\varepsilon = \nu \left(u_\eta/\eta\right)^2 = \nu/\tau_\eta^2 \tag{3.5}$$

que mostra que $(u_\eta/\eta = 1/\tau_\eta)$ fornece uma consistente caracterização dos gradientes de velocidades dos vórtices dissipativos (POPE, 2000).

A razão das menores escalas para as grandes escalas são facilmente determinadas a partir da definição das escalas de Kolmogorov e da escala $\varepsilon \sim u_0^3/l_0$ - ver (POPE, 2000). Os resultados são:

$$\eta/l_0 \sim Re^{-3/4} \tag{3.6}$$

$$u_\eta/u_0 \sim Re^{-1/4} \tag{3.7}$$

$$\tau_\eta/\tau_0 \sim Re^{-1/2} \tag{3.8}$$

Observe que a taxa η/l_0 decresce com o aumento de Re. Consequentemente, em números de Reynolds suficientemente altos, existe um intervalo de escalas l que são muito pequenas quando comparadas com l_0 e muito grandes quando comparadas com η, isto é, $\eta \ll l \ll l_0$. Uma vez que os vórtices neste intervalo são muito maiores que os vórtices dissipativos, pode se supor que seu número de Reynolds $(lu(l)/\nu)$ é grande e, logo, seu movimento será pouco afetado pela viscosidade. Assim, a partir disso e da primeira hipótese de similaridade pode-se estabelecer a segunda hipótese de Kolmogorov (KOLMOGOROV, 1991b; POPE, 2000):

Segunda hipótese de similaridade de Kolmogorov: Em todo escoamento turbulento com número de Reynolds suficientemente grande, a estatística dos movimentos de escala l no intervalo $\eta \ll l \ll l_0$ tem uma forma universal que é unicamente determinada por ε e independente de ν.

De acordo com a segunda hipótese de similaridade, movimentos no subintervalo inercial são determinados pelos efeitos inerciais e os efeitos viscosos são desprezíveis, enquanto que os movimentos no intervalo de dissipação $(l < l_\eta)$ são determinados pelos efeitos viscosos. Dado um vórtice de tamanho l (no subintervalo inercial), as escalas características de velocidade e de tempo (*timescales*) para o vórtice são aquelas formadas a partir de ε e l, conforme segue:

$$u(l) = (\varepsilon l)^{1/3} = u_\eta(\eta)^{1/3} \sim u_0(l/l_0)^{1/3} \tag{3.9}$$

$$\tau(l) = (l^2/\varepsilon)^{1/3} = \tau_\eta(l/\eta)^{2/3} \sim \tau_0(l/l_0)^{2/3}. \tag{3.10}$$

Uma consequência da segunda hipótese de similaridade é que (no subintervalo inercial) as escalas de velocidade $u(l)$ e de tempo $\tau(l)$ descrescem quando l decresce.

Segundo Pope (2000), na concepção da cascata de energia a quantidade de principal importância, denotada por $\Upsilon(l)$, é a taxa em que a energia é transferida dos vórtices maiores que l para os vórtices menores que l. Se este processo de tranferência for completado principalmente por vórtices de tamanho comparáveis a l, então $\Upsilon(l)$ pode ser da ordem de $u(l)^2/\tau(l)$. A identidade

$$u(l)^2/\tau(l) = \varepsilon \tag{3.11}$$

derivada das Equações 3.9 e 3.10, sugere que $\Upsilon(l)$ é independente de l (para l no subintervalo inercial) e então, $\Upsilon(l) = \varepsilon$. Isto é, a taxa de transferência de energia das

grandes escalas (Υ_I) determina a taxa constante de transferência de energia através do subintervalo inercial ($\Upsilon(l)$) e a taxa em que a energia deixa o subintervalo inercial e entra no intervalo de dissipação (Υ_D), ou seja determina a taxa de dissipação ε. Isto mostra que ε é um parâmetro extremamente importante e que controla o fluxo de energia desde as grandes escalas onde ela é injetada até as pequenas escalas, onde ela será dissipada pela viscosidade.

Segundo a teoria K-41, uma maneira de se determinar como a energia cinética é distribuida entre os vórtices de diferentes tamanhos é através da utilização do espectro $E(k)$. Considerando que o número de onda da injeção é muito menor do que o número de onda da dissipação $k_i \ll k_d$, no intervalo intermediário de escalas $k_i < k < k_d$ o espectro de energia depende somente do fluxo de energia ε e do número de onda k. Logo, pode-se expressar a densidade de energia como:

$$E(k) = f(\varepsilon, k) \tag{3.12}$$

Usando análise dimensional tem-se:

Quantidade	**Dimensão**
Número de onda k	1/L
Energia por unidade de massa E	$E^2 \sim L^2/T^2$
Espectro de energia $E(k)$	$EL \sim L^3/T^2$
Fluxo de energia ε	$E/T \sim L^2/T^3$

A Equação 3.12 tem dimensionalidade L^3/T^2, a dimensão T^{-2} só pode ser balanceada por $\varepsilon^{2/3}$ porque k só depende de L e não tem dependência do tempo. Logo,

$$E(k) = \varepsilon^{2/3} g(k) \tag{3.13}$$

Seguindo com análise dimensional, $g(k)$ deve ter dimensão $L^{5/3}$ e portanto a equação para o espectro de energia cinética turbulenta torna-se:

$$E(k) = C\varepsilon^{2/3} k^{-5/3} \tag{3.14}$$

onde C é uma constante universal, a constante de Kolmogorov. Este é o famoso espectro de Kolmogorov, uma das pedras fundamentais da teoria da turbulência.

3.1.2 Intermitência

Logo após a formulação da teoria K-41 surgiram alguns trabalhos que apontaram discrepâncias entre os resultados teóricos e os experimentais (FRISCH, 1995). Assim, em 1962 Kolmogorov propõe uma correção de sua teoria (KOLMOGOROV, 1962). Enquanto na teoria K-41 a taxa de dissipação de energia cinética é considerada constante e o fenômeno da intermitência não é considerado, na teoria conhecida como K-62 é proposto uma distribuição *log normal* da taxa de dissipação de energia e a intermitência nas pequenas escalas é considerada.

Na Figura 3.1, tem-se um cenário que ilustra a cascata de energia conforme proposto pela teoria K-41 e um cenário segundo o modelo β que leva em conta a intermitência. Na Figura 5.1 -(a) os vórtices de vários tamanhos estão representados como bolhas empilhadas em tamanhos decrescentes. Os vórtices superiores têm uma escala $\sim l_0$, e as sucessivas gerações de vórtices têm escalas $l_n = l_0 r^n$ com $(n = 0, 1, 2, ...)$ onde $0 < r < 1$. Os vórtices menores têm a escala de dissipação de Kolmogorov (η). É assumido que o número de vórtices por unidade de volume, cresce com n como r^{-3n} para garantir que os pequenos vórtices completem todo o espaço (FRISCH, 1995).

A Figura 5.1 - (b) mostra a idéia do modelo β, que modifica o modelo da teoria K-41 para introduzir a intermitência. Basicamente, em cada estágio da cascata de energia, o número de *filhas* da vórtice *mãe* é escolhido tal que a função do volume ocupado é decrescente por um fator β com $(0 < \beta < 1)$. O fator β é um parâmetro ajustável no modelo e a fração p_l do espaço que está ativo, isto é, dentro de um vórtice *filho* de tamanho $l = r^n l_0$ decresce como uma lei de potência de l. Logo, tem-se:

$$p_l = \beta^n = \beta^{\frac{log(l/l_0)}{log\ r}} = \left(\frac{l}{l_0}\right)^{3-D} \tag{3.15}$$

onde $3 - D \equiv \frac{log\ \beta}{log\ r}$.

Uma abordagem multifractal para modelar a intermitência foi proposta por Parisi e Frisch (1985). De acordo com a estratégia de fechamento de primeira ordem, ou teoria-K, o fluxo turbulento pode ser representado pelo produto entre uma difusividade turbulenta e o gradiente da quantidade principal. As propriedades dos fluxos turbulentos segundo a teoria de Kolmogorov podem ser analisadas a partir da estatística do gradiente de velocidade $v_r(x) = v(x + r) - v(x)$ para várias escalas r. A partir da formulação multifractal, a função de estrutura $S_p(r)$, que é uma para-

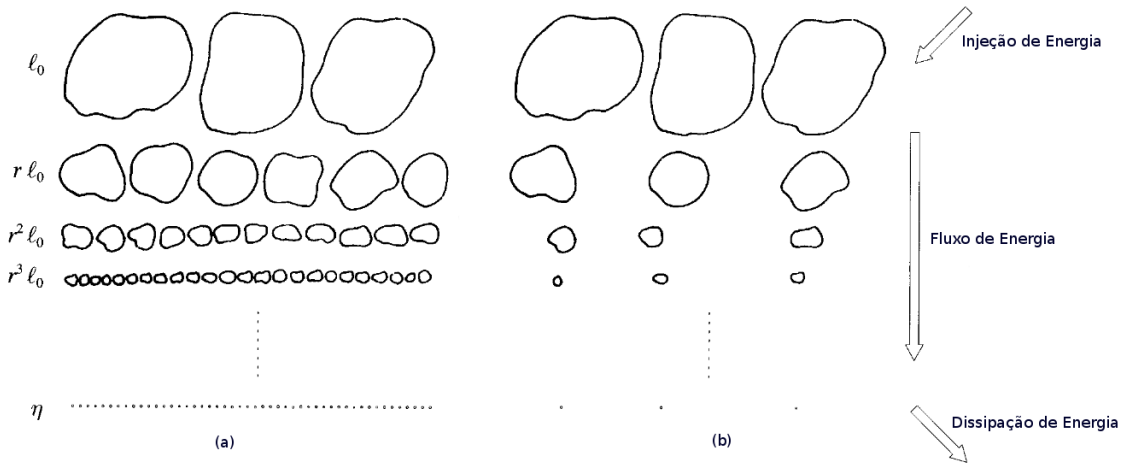

Figura 3.1 - (a) Cascata de energia de acordo com a teoria K-41. (b) Cascata de energia de acordo com o modelo β.
Fonte: Adaptada de Frisch (1995).

metrização dos momentos estatísticos do gradiente de velocidade formado na escala espacial $r = V_r(x)$, para o intervalo inercial ($\eta \ll r \ll L$) pode ser expressa por:

$$\frac{S_p(r)}{v_0^p} \equiv \frac{\langle v_r^p \rangle}{v_0^p} \sim \int_I d\mu(b) \left(\frac{r}{L}\right)^{pb+3-D_F(b)} \tag{3.16}$$

onde L e η são as escalas integrais e de Kolmogorov, $\langle . \rangle$ é o valor médio e $D_F(b)$ é a dimensão fractal. No domínio inercial, o menor expoente irá dominar a Equação 3.16 e $\zeta_p = inf\,\{pb + 3 - D_F(b)\}$ é o expoente para o escalonamento da função de estrutura, com $b \in (b_{min}, b_{max})$ (FRISCH, 1995; CAMPOS VELHO, 2010).

A transformada de Fourier da função de estrutura dada em 3.16 pode ser escrita da seguinte forma:

$$E(k) = c_2 \varepsilon^{2/3} k^{-5/3} \left(k/k_\eta\right)^{2/3-\zeta_2} \tag{3.17}$$

onde k_η é o número de onda de um comprimento característico do sistema físico. A Equação 3.17 pode ser considerada como uma lei de Kolmogorov generalizada para o domínio inercial do espectro. A partir desta equação vale ressaltar que, primeiro, conforme Frisch (1995), o expoente $\varsigma \equiv \zeta_2 - 2/3$ define o expoente de intermitência. Segundo, o menor comprimento geométrico dentro do sistema físico é representado através do número de onda k_η.

3.2 A Turbulência e a Evolução das Heterogeneidades no Universo

O estudo da dinâmica do Universo pode ser realizado considerando-se que seus elementos constituintes se comportam como um fluido. Este procedimento é adotado, por exemplo, nas simulações de matéria escura, onde a componente de massa dominante, a matéria escura fria é modelada como um fluido acolisional auto-gravitante. A partir de uma distribução inicial de matéria e um conjunto de condições iniciais, aplica-se uma pequena perturbação no sistema e este é evoluido no tempo. Os resultados reproduzem razoalvelmente a maioria das estruturas observadas atualmente, porém não explica completamente os fenômenos envolvidos neste processo. A teoria utilizada em simulações deste tipo é baseada na instabilidade gravitacional e na *Aproximação de Zel'dovich* (ZEL' DOVICH, 1970).

Uma outra abordagem para a explicação das heterogeneidades sugere um mecanismo turbulento de evolução. Esta idéia por si só é interessante, uma vez que, no regime turbulento, a própria dinâmica vai levar a uma magnificação da heterogeneidade dos campos dinâmicos, por exemplo, o campo de vorticidade num escoamento turbulento apresenta zonas de vorticidade intensa e zonas de vorticidade fraca ou quase inexistentes, padrão conhecido como intermitência (FRISCH, 1995). A Figura 3.2 exibe em (a) os resultados de uma simulação computacional $3D$ de turbulência em fluidos e em (b) uma simulação de matéria escura que apresenta estruturas filamentares e aglomerados.

As similaridades dos padrões observados na intermitência dos campo de vorticidade e na distribuição de massa, parecem sugerir que um mecanismo natural para a geração ou magnificação das heterogeneidades possa ser descrito por uma dinâmica do tipo da turbulência.

A intermitência seria o mecanismo que poderia explicar um pequeno desvio do expoente $-5/3$. Esta é uma das principais questões estudadas pelos teóricos da turbulência. Deduzir o expoente da intermitência por princípios básicos é um tema de pesquisa intensa (FRISCH et al., 1978; SHE; LéVêQUE, 1994; RAMOS et al., 2001; CAMPOS VELHO et al., 2001).

Uma questão relevante é uma análise qualitativa se a dinâmica da evolução cosmológica pode ser similar à dinâmica turbulenta identificada em fluidos. Este é o objetivo desta tese. Porém, aqui, este texto é voltado para um único aspecto da turbulência:

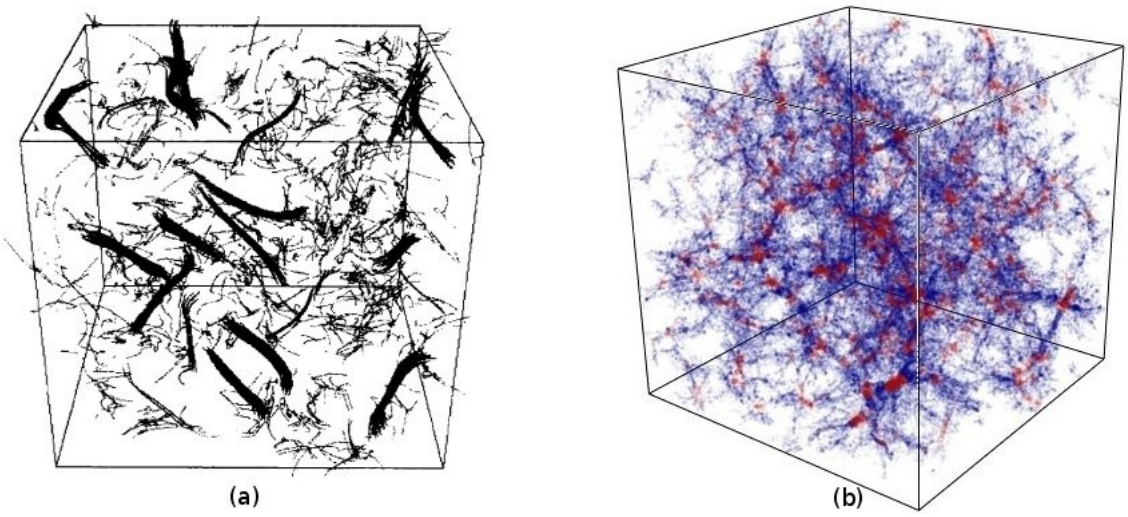

Figura 3.2 - (a) Filamentos de vórtices intermitentes em uma simulação computacional 3D de turbulência em fluidos. (b) Simulação de matéria escura com destaque em vermelho para os aglomerados de galáxias.
Fonte: (a) Adaptada de She et al. (1991) e (b) http://www.exp-astro.phys.ethz.ch/hahn/ .

se a distribuição de energia em relação ao número de onda segue a lei de Kolmogorov de -5/3. Contudo, a caracterização da turbulência se iniciou antes desta lei de espectro de energia.

Uma das caracterizações mais antigas foi dada por Osborne Reynolds, em que o fluxo turbulento ocorre quando uma certa grandeza adimensional (número de Reynolds: R_e) é maior do que um certo limiar. É correto dizer que o fluido apresentará um fluxo turbulento para altos números de Reynolds, mas um limiar quantitativo separando a dinâmica laminar da dinâmica turbulenta não é muito preciso. Alguns textos informam que para $R_e > 2500$ tem-se o escoamento turbulento. Entretanto, o mais correto é indicar um *intervalo* de transição do regime laminar para o regime turbulento. Em certas circunstâncias, é possível encontrar uma dinâmica turbulenta com $R_e = 500$, bem como identificar um regime laminar com $R_e = 4000$, mas estas são situações instáveis.

O número de Reynolds é uma grandeza adimensional que relaciona: massa específica, viscosidade, velocidade e um comprimento característico. Para uma dinâmica turbulenta da evolução cosmológica, quais seriam os parâmetros para o cálculo do R_e? A velocidade poderia ser a velocidade das galáxias e a massa específica seria a verificada para o Universo constituído predominantemente pela matéria escura. Mas,

o que seria o comprimento característico e a viscosidade para o *fluido cósmico*?

Para fluidos escoando dentro de um tubo, o comprimento característico é óbvio: o diâmetro do tubo, isto é, o maior diâmetro de um vórtice no escoamento turbulento dentro de um tubo será, no máximo, o diâmetro do tubo. Em outras situações, este comprimento característico não é tão óbvio. Por exemplo: para a atmosfera do planeta Terra, qual é o comprimento característico para turbulência atmosférica? De fato, a abordagem para o segundo caso pode ser vista como uma generalização do caso anterior e o conceito chave é a *teoria da camada limite*. A camada limite é uma camada próxima a uma fronteira sólida em que a turbulência (e viscosidade) não pode(m) ser desprezada(s). Neste caso, a espessura da camada limite é o comprimento característico (SCHLICHTING, 1979). Se a espessura da camada limite for maior que o raio de um tubo, a turbulência estará presente em qualquer ponto no interior do tubo (para um comprimento longe da borda de entrada do tubo) e o comprimento característico será então o diâmetro do tubo, como já mencionado.

Qual é o comprimento característico na evolução cosmológica regida por uma dinâmica turbulenta? Se o Universo partiu de uma região muito pequena, nessa fase, o próprio diâmetro do Universo seria o comprimento característico. Para uma fase em que as estruturas cósmicas já estão formadas, seria possível desenvolver uma teoria de camada limite (para o entorno das estruturas de grande escala) onde a turbulência não pode ser desprezada.

Por último, se para uma determinada escala a distribuição de objetos astronômicos (podem ser galáxias) pode ser vista como um fluido, seria possível deduzir uma quantidade análoga a uma viscosidade para tal fluido, usando as mesmas técnicas da mecânica estatística da dedução da viscosidade a partir do campo de velocidade das partículas. Na verdade, o cálculo da viscosidade pode ser feito com a teoria cinética dos gases, onde o gás é considerado diluído e os efeitos de colisões podem ser desconsiderados.

É pertinente destacar que, para o fluxo turbulento, a viscosidade laminar não é levada em consideração, pois na dinâmica turbulenta, o fluxo depende do regime dinâmico e não do tipo de fluido. Ou seja, no regime turbulento, não importa se o fluido é, por exemplo, o ar ou a água, a dinâmica da turbulência terá a *mesma* distribuição de energia (a lei dos $-5/3$) para ambos. O fluxo turbulento depende do regime do escoamento e não do tipo de fluido. Portanto, se a dinâmica turbulenta

rege a evolução cosmológica, por analogia, a viscosidade da matéria escura, seja ela qual for, não irá afetar a evolução (a dinâmica) do fluido cósmico e a lei de distribuição de energia deverá seguir a lei dos $-5/3$.

4 COMPUTAÇÃO DE ALTO DESEMPENHO

A computação de alto desempenho caracteriza-se principalmente pelo uso intensivo de recursos computacionais como supercomputadores, *clusters*, entre outros, visando aumentar o poder de computação e melhorar o desempenho de aplicações que demandam grande capacidade de processamento ou armazenamento.

Em geral, quando pretende-se melhorar o desempenho de aplicações que requerem grande poder computacional tem-se três possíveis alternativas: a utilização de processadores mais eficientes, a otimização dos algoritmos utilizados e o emprego da computação paralela. Quanto a utilização de processadores mais velozes e eficientes existem certas limitações físicas, além do elevado custo de fabricação que torna essa solução as vezes pouco acessível. A obtenção de algoritmos otimizados nem sempre é possível, além de muitas vezes também ter custos elevados. Assim, a computação paralela por meio da agregação de recursos locais (*clusters*) ou geograficamente distribuídos (*grids*) pode representar uma opção bastante adequada.

As máquinas paralelas podem ser divididas conforme o espaço de endereçamento de memória em duas grandes classes: máquinas com arquitetura de memória compartilhada e máquinas com arquitetura de memória distribuída:

- **Arquitetura de memória compartilhada:** Neste tipo de arquitetura tem-se um espaço único de memória em que cada processador compartilha seu uso, e assim, a comunicação entre os processadores ocorre por meio da memória global.
- **Arquitetura de memória distribuída:** Nesta arquitetura tem-se espaços múltiplos e independentes de memória. A comunicação entre os processadores é feita através de uma rede de conexão.

Cada uma dessas classes podem então ser divididas em máquinas vetoriais (onde executa-se a mesma instrução de forma sincronizada por todos os processadores sobre diferentes conjuntos de dados) e em máquinas escalares. As máquinas escalares atuais têm tipicamente mais de um núcleo, sendo que cada núcleo tem a funcionalidade de um processador, porém possui memória compartilhada (todos os núcleos enxergam a mesma memória). As máquinas paralelas escalares podem ter sua capacidade de processamento aumentada por meio de *Field Programmable Gate Arrays*

(FPGAs) e de *Graphics Processing Unit* (GPUs) que servem como uma espécie de co-processadores da *Central Processing Unit* (CPU). Exemplos desta configuração são CPU + FPGA, CPU + FPGA + GPU, CPU + GPU. A configuração híbrida (GPU + CPU) caracteriza-se como um modelo heterogêneo de co-processamento, em que a parte sequencial da aplicação é executada pela CPU e a parte computacionalmente intensiva é acelerada pela GPU. Esta abordagem tem se tornado muito popular entre a comunidade científica, sobretudo devido ao baixo custo em relação a outros sistemas de alto desempenho como supercomputadores e/ou *clusters*. O paralelismo usando uma configuração do tipo CPU + GPU pode ser obtido por meio de linguagens como OpenCL ou extensões do C/C++ para CUDA. Já o paralelismo para sistemas híbridos composto por CPU + FPGA utilizam linguagens como VHDL.

Uma das primeiras classificações de arquiteturas de computadores paralelos é a bem conhecida taxonomia de Flynn (FLYNN, 1966). Esta classificação leva em conta o número de instruções executadas em paralelo *versus* o fluxo de dados para os quais as instruções são submetidas. Assim, existem quatro diferentes modelos, a saber:

- **Single Instruction, Single Data (SISD):** Caracteriza-se por um computador seqüencial que executa uma única instrução sobre um único fluxo de dados. Esse modelo de arquitetura foi proposto por Von Newmann e como exemplo podemos citar as máquinas convencionais com uma CPU.
- **Single Instruction, Multiple Data (SIMD):** Nesse caso tem-se um tipo de arquitetura conhecida como arquitetura vetorial em que a mesma instrução é executada por múltiplos processadores em vários fluxos de dados. Exemplos de computadores com essa arquitetura são as máquinas Thinking Machine CM-2 e MASPAR MP-1216. As máquinas como NEC SX e Cray T90 possuem um modelo semelhante ao SIMD, porém possuem a vantagem de poderem realizar o processamento por partes.
- **Multiple Instruction, Single Data (MISD):** Este é um modelo que tem múltiplas instruções que são executadas simultaneamente sobre o mesmo fluxo de dados.
- **Multiple Instruction, Multiple Data (MIMD):** Máquinas com vários processadores trabalhando de forma independente, onde cada um pode executar diferentes instruções sobre diferentes subconjuntos de dados. Esta ar-

quitetura pode ser encontrada em computadores com memória distribuida e memória compartilhada. Como exemplo, podemos citar os computadores paralelos da linha SP da IBM, Intel Paragon, entre outros.

4.1 Ambientes de Programação Paralela

A utilização da computação paralela e distribuída é bastante atraente e assume um papel importante na busca da melhoria do desempenho de aplicações que requerem alto poder computacional. Sucintamente, podemos definir a computação paralela como sendo o uso simultâneo de vários processadores para resolver um determinado problema com o objetivo de reduzir o tempo total de execução.

Basicamente a implementação paralela de um determinado algoritmo consiste ou na partição das tarefas, onde as tarefas necessárias a resolução do problema são divididas entre os diversos nós de processamento ou na partição do domínio, onde os dados do problema são divididos entre os vários nós de processamento. O paralelismo pode ser obtido em diversos níves: paralelismo ao nível de instrução (realizado pelo próprio pipeline), paralelismo ao nível de *threads* (ideal para máquinas multi-cores), paralelismo ao nível de processos (sendo que cada processo pode ser responsável por vários *threads*), e, finalmente, paralelismo ao nível de *grids*.

Os principais modelos de paralelismo são a troca de mensagens entre processos, cujo modelo padrão é a biblioteca *Message-Passing Interface* (MPI) (PACHECO, 1997), que é utilizada amplamente em computadores com memória distribuída e o modelo de *threads fork-join*, cujo padrão é o OpenMP (CHAPMAN et al., 2007), amplamente utilizado em computadores com memória compartilhada. No contexto histórico vale destacar ainda o *High Performance Fortran* (HPF) e o *Parallel Virtual Machine* (PVM). HPF era uma extensão do fortran 90 em que o compilador é responsável pela distribuição dos dados e pelo controle da execução das tarefas, sendo então, uma linguagem explícita. Já o PVM é uma biblioteca (de *software*) capaz de interligar máquinas com arquiteturas homogêneas e heterogêneas, permitindo que elas sejam utilizadas como um único computador paralelo.

A máquina disponível para utilização nesta tese possui memória distribuida entre os vários nodos e memória compartilhada intra-nodo. Assim, para realizar as tarefas por meio de troca de mensagens entre os nodos, utilizamos o modelo MPI, que será descrito em mais detalhes na próxima seção.

4.1.1 Message Passing Interface

A interface MPI é um padrão que define a síntaxe e a semântica das funções contidas em uma biblioteca de envio e recebimento de mensagens que foi desenvolvida especificamente para ser usada em programas que explorem a existência de múltiplos processadores. É o padrão para a comunicação entre os nodos que executam um programa em um sistema de memória distribuida (El-REWINI; ABD-EL-BARR, 2005).

As implementações em MPI consistem de um conjunto de bibliotecas de rotinas que podem ser utilizadas em programas escritos nas linguagens de programação C, C++ e Fortran. A principal vantagem da MPI é que os programas são portáveis, já que ela foi implementada para quase todas as arquiteturas de memória distribuída, e rápidos, já que cada implementação da biblioteca foi otimizada para o *hardware* na qual se executa.

No padrão MPI, o envio de mensagens pode ser síncrono o assíncrono. Se o processo que envia a mensagem espera a mesma ser recebida para continuar sua execução, dizemos que o modo de envio é síncrono. Já se o processo que envia não espera a mensagem ser recebida e continua sua execução, podendo até enviar outras mensagens, dizemos que o modo de envio é assíncrono. A cada processo é designado uma variável denominada *rank*, a qual identifica o processo, no intervalo de 0 a $p-1$, onde p é o número total de processos. O controle da execução do programa é realizado mediante esta variável. Ela permite determinar qual processo executa determinado trecho do código.

A biblioteca MPI trabalha com o conceito de comunicadores para definir o conjunto de processos envolvidos em uma tarefa de comunicação. Essencialmente um comunicador é uma coleção de processos capazes de se comunicarem entre si. O comunicador padrão definido pela MPI é denominado MPI_COMM_WORLD.

As chamadas a MPI basicamente são divididas em quatro classes: chamadas utilizadas para inicializar, administrar e finalizar comunicações, chamadas utilizadas para transferir dados entre um par de processos, chamadas para transferir dados entre vários processos e, finalmente, chamadas utilizadas para criar tipos de dados definidos pelo usuário.

A primeira classe de chamadas permite inicializar a biblioteca de envio de mensagens, identificar o número de processos e o intervalo dos mesmos. Ela dispõe de 4

funções essenciais que são utilizadas em todo programa com MPI. Estas funções são: MPI_Init que permite inicializar uma sessão MPI. Deve ser utilizada antes de chamar qualquer outra função da MPI. A função $MPI\_Finalize$ que permite terminar uma sessão MPI e liberar a memória usada. Deve ser a última chamada a MPI realizada pelo programa. A função MPI_Comm_size que permite determinar o número total de processos que pertencem a um comunicador e a função MPI_Comm_rank que permite determinar o identificador (*rank*) do processo atual.

A segunda classe de chamadas inclui operações de comunicação ponto a ponto, para diferentes tipos de atividades de envio e recebimento. A transferência de dados entre dois processos é obtida através das funções MPI_Send, que permite enviar informação de um processo a outro e MPI_Recv, que permite receber informação de outro processo. Estas funções devolvem um código que indica seu sucesso ou fracaso e ambas são bloqueantes, ou seja, o processo que realiza a chamada se bloqueia até que a atividade de comunicação esteja completa.

A terceira classe de chamadas fornecem operações de comunicações entre grupos de processos. MPI possui funções para a comunicação grupal que incluem operações do tipo difusão (*broadcast*), re-coleta (*gather*), distribuição (*scatter*) e redução. A função MPI_Bcast permite a um processo enviar uma cópia de seus dados a outros processos dentro de um grupo definido por um comunicador. A função MPI_Gather estabelece uma operação de re-coleta, na qual os dados são re-coletados em um único proceso. $MPI\_Scatter$ estabelece uma operação de distribuição, na qual um dado é distribuido entre os diferentes processos. Por fim, a função MPI_Reduce permite que o processo mestre re-colete dados a partir de outros processos em um grupo, combinando-os em um único dado.

A última classe de chamadas fornece flexibilidade na construção de estruturas de dados complexos. Por meio da função MPI_Type é possível definir novos tipos de dados.

4.2 Computação em Grade

O surgimento da computação em grade (do inglês *grid computing*), está diretamente relacionado a crescente demanda de poder computacional. O conceito nasceu na comunidade de Processamento de Alto Desempenho (PAD), sendo apresentado pelos pesquisadores Ian Foster e Carl Kesselman em 1998 com o lançamento do livro

The Grid: Blueprint for a New Computing Infrastructure. De acordo com Foster e Kesselman (2004), uma grade computacional é uma infra-estrutura de *hardware* e *software* que fornece acesso a recursos computacionais de forma confiável, consistente, abrangente e de baixo custo.

A tecnologia de computação em grade fornece mecanismos para compartilhar e coordenar o uso de diversos recursos computacionais (supercomputadores, *clusters*, dados, espaço de armazenamento, entre outros) distribuídos geograficamente em diferentes instituições, de modo a criar um único *computador virtual* capaz de alcançar elevada taxa de processamento e armazenamento (FOSTER, 2003; FOSTER; KESSELMAN, 2004).

Em outras palavras, podemos dizer que os ambientes de *grid* fornecem uma tecnologia de acesso comum a uma grande quantidade de recursos e serviços que são disponibilizados pelas instituições que fazem parte do ambiente. Assim, conforme Dantas (2005), uma grade computacional é uma configuração distribuída que considera o uso de recursos e serviços geograficamente distribuídos com o objetivo de melhorar, viabilizar e possibilitar economicamente uma aplicação.

Diferente dos sistemas paralelos e distribuídos tradicionais, essa nova tecnologia precisa considerar questões importantes como segurança, acesso uniforme aos recursos, descoberta e agregação dinâmica, além da qualidade do serviço (ASADZADEH et al., 2005).

Quando pretende-se explorar de forma eficiente o ambiente de grade computacional, Asadzadeh et al. (2005) destaca alguns desafios que devem ser considerados, são eles: *Heterogeneidade*, que surge devido a multiplicidade de recursos e a grande variedade de tecnologias que constituem a grade. *Multiplicidade* de domínios administrativos, uma vez que os recursos da grade estão distribuídos geograficamente entre diferentes organizações e assim pertencem a diferentes proprietários. *Escalabilidade*, já que o aumento no tamanho da grade pode causar perda de desempenho na execução das tarefas. Por último, a *natureza dinâmica* desse ambiente, pois falhas na rede de comunicação alteram a disponibilidade dos recursos.

Levando-se em conta os desafios anteriormente citados, Foster et al. (2001) propõem um modelo de arquitetura em grade organizado em camadas, conforme indicado na Figura 4.1. Com base nesta proposta apresentamos uma breve descrição de cada

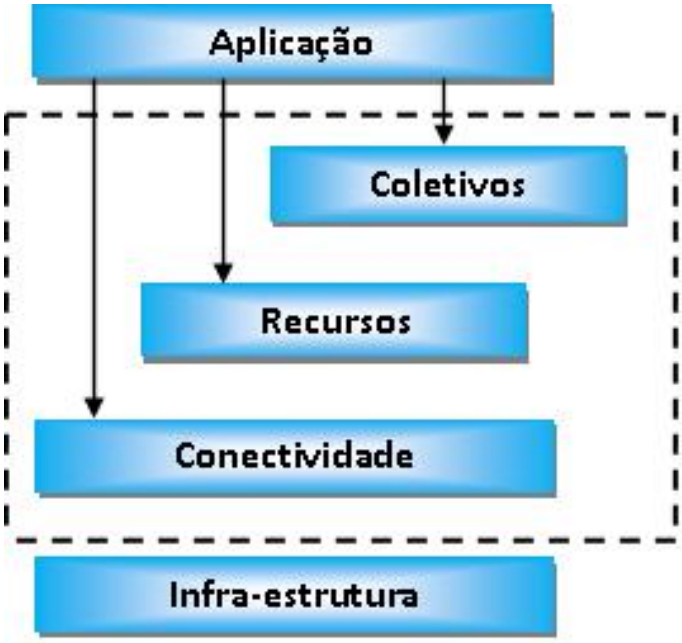

Figura 4.1 - Camadas que constituem a arquitetura da grade.
Fonte: Adaptada de Foster et al. (2001).

uma dessas camadas.

- A camada Infra-estrutura fornece os recursos para os quais o acesso compartilhado é mediado pelo protocolo da grade. Dentre os principais recursos fonecidos, podemos citar: recursos computacionais, sistemas de armazenamento, catálagos, entre outros.

- A camada Conectividade define o núcleo de protocolos de comunicação e autenticação necessários para transações na rede. Os protocolos de comunicação permitem a troca de dados entre os recursos da camada Infra-estrutura, enquanto que, os protocolos de autenticação criam, sobre os serviços de comunicação, mecanismos seguros para a verificação da identidade dos usuários e dos rescursos disponíveis.

- A camada Recursos é responsável pela criação de protocolos de comunicação e autenticação que servem para definir protocolos de negociação segura, inicialização, monitoramento e controle de operações de compartilhamento de recursos individuais.

- Enquanto a camada Recursos é focada sobre interações com um único recurso, a camada Coletivos contém protocolos e serviços que capturam interações por meio de coleções de recursos.

- Por último, a camada Aplicação, é responsável por viabilizar a execução das aplicações que utilizam e exploram os recursos disponíveis na grade.

As camadas Conectividade, Recursos e Coletivos, fazem parte de um nível da arquitetura de grade também conhecido como *middleware*. Os sistemas *middlewares* tem papel fundamental e devem ser adaptados dinamicamente de modo a prover o uso efetivo e eficiente dos serviços e recursos disponíveis na grade computacional, (ASADZADEH et al., 2005). Existem na literatura vários desenvolvimentos de *middlewares* aptos a explorar de forma eficaz o potencial da grade. Dentre estes desenvolvimentos podemos citar: *Globus Toolkit* (FOSTER; KESSELMAN, 1998), considerado um modelo padrão para a tecnologia de computação em grade, CIGRI/OAR (CAPIT et al., 2005) e a arquitetura *OurGrid* (ANDRADE et al., 2003), que foi utilizada nesta tese.

A arquitetura Globus Toolkit foi desenvolvida no contexto do projeto *Globus*, (FOSTER; KESSELMAN, 1998). Este projeto é um esforço em pesquisa multi-institucional que procura permitir a construção de computação em grade de forma eficiente e com acesso consistente aos recursos computacionais de alto desempenho, ainda que, recursos e usuários estejam distribuídos geograficamente. O *Globus toolkit* compreende um conjunto de componentes que implementam serviços básicos para projetos de segurança, alocação e gerenciamento de recursos, comunicação, entre outros (FOSTER; KESSELMAN, 1998; FOSTER et al., 2001).

Almeida (2007), avaliou o desempenho de uma aplicação de climatologia para o modelo BRAMS usando a grade do projeto G-BRAMS. Em seu estudo ele analisou o desempenho das 3 plataformas citadas anteriormente. A plataforma OurGrid foi a que apresentou melhor desempenho, porém ele observou que existiram outros fatores que provocaram a queda no desempenho da aplicação usando as plataformas Globus e CIGRI/OAR, como por exemplo: queda da Internet, indisponibilidade de recursos, falhas na rede e na administração dos recursos, entre outros. Além do projeto G-BRAMS, o LAC conta com outros projetos em grade como o RECLIRS, o CLIMARS e o STIC AmSud que envolvem aplicações de climatologia e o projeto de ótica hidrológica inversa, com aplicação em tomografia 3D de clorofila do oceano (água do tipo-1: longe da costa).

4.2.1 Plataforma OurGrid

De acordo com Andrade et al. (2003), a plataforma *OurGrid* é baseada em um modelo de recursos compartilhados que procura fornecer recursos às aplicações paralelas cujas tarefas são conhecidas como aplicações *Bag-of-Tasks* (BoT). BoT são aplicações paralelas compostas por um conjunto de tarefas independentes, ou seja, que não necessitam de comunicação durante a execução.

A plataforma *OurGrid* explora a idéia de que uma grade é composta de vários *sites* que têm interesse em trocar favores computacionais entre si, por isso ela foi desenvolvida para trabalhar como uma rede *Peer-to-Peer* de recursos pertencentes a comunidade de usuários da grade. Adicionando recursos na rede e compartilhando com a comunidade, o usuário ganha acesso a todos os recursos disponíveis (ANDRADE et al., 2003). Esta rede *Peer-to-Peer* de troca de favores permite que os recursos ociosos de um site seja fornecido. Podemos observar na Figura 4.2 um esquema da rede de favores, em que cada *Peer* controla um conjunto de recursos de um *site*. Quando surge uma demanda interna por recursos que o *Peer* de um determinado *site* não consegue suprir, ele fará requisições a comunidade. A plataforma *OurGrid* é formada por três componentes principais: *MyGrid, OurGrid Peer e OurGrid Worker*. O componente *MyGrid* é a interface para o uso da grade computacional, ou seja, é aquele que durante o processamento de tarefas (*jobs*) atua como o coordenador da grade, sincronizando a execução dessas tarefas e fazendo toda a transferência de dados necessária (CIRNE et al., 2003). O *OurGrid Peer* tem como principal função organizar e fornecer trabalhadores que pertencem ao mesmo domínio administrativo. Do ponto de vista de usuário, um *Peer* é um fornecedor de trabalhadores, ou seja, uma rede de serviços que dinamicamente fornece trabalhadores para a execução de tarefas. Do ponto de vista administrativo um *Peer* determina como e quais máquinas podem ser usadas como trabalhadores (CIRNE; SANTOS NETO, 2005). Finalmente, *OurGrid Worker* é responsável por executar tarefas para os usuários da grade. Quando um *Worker* recebe a tarefa para processar, ele deve criar uma máquina virtual para executá-la. Como resultado, todos os processos da grade são executados em uma máquina virtual e assim, as máquinas da grade real são protegidas do uso indevido ou malicioso (OURGRID..., 2008).

A seguir será apresentado uma breve descrição de como os componentes do *OurGrid* abordam os aspectos Autenticação, Descoberta e Alocação de Recursos e Transferência de Dados de acordo com Cirne e Santos Neto (2005).

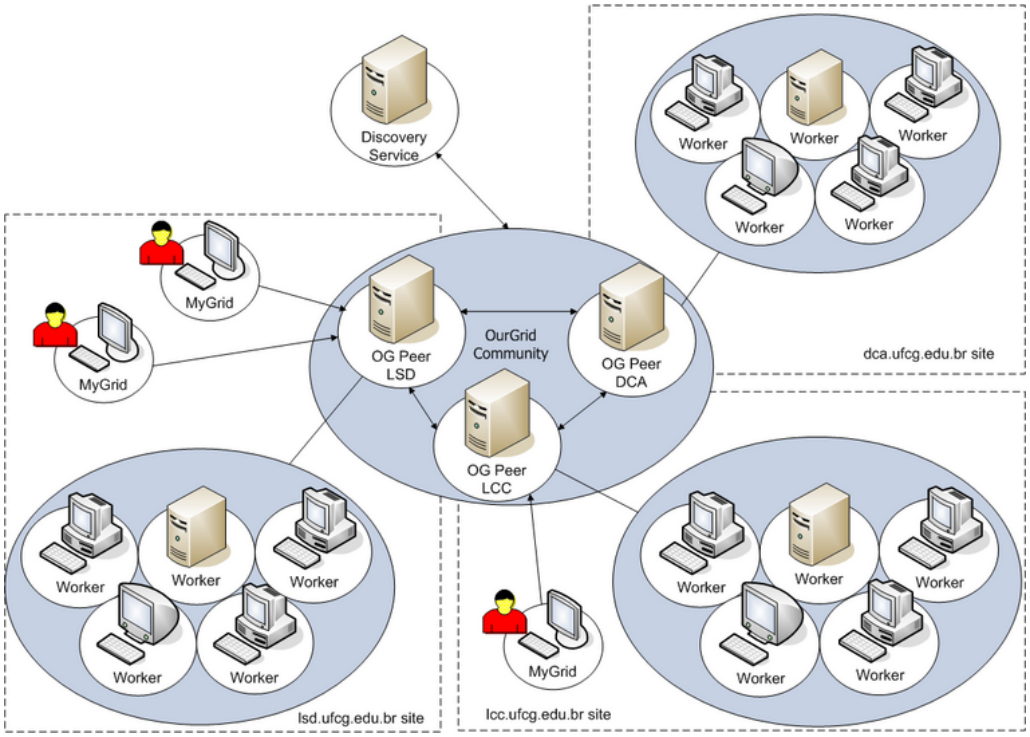

Figura 4.2 - Esquema de Funcionamento da Arquitetura OurGrid.
Fonte: OurGrid... (2008).

- Autenticação: A arquitetura *OurGrid* possui dois níveis de autenticação, que dependem da maneira pela qual o usuário obteve o recurso. No primeiro nível, o usuário pode acessar diretamente recursos por meio de *Grid Machines* - GuMs em sua rede local. Neste tipo de autenticação, geralmente usa-se a infra-estrutura de autenticação do sistema operacional do recurso. Vale ressaltar que o *OurGrid* possibilita ao usuário obter acesso a GuMs de outros *sites*.

- Descoberta e Alocação de Recursos: O primeiro passo na execução de uma aplicação, utilizando a arquitetura *OurGrid*, é a descrição da aplicação e do conjunto de recursos que são acessíveis ao usuário. Uma vez submetida a aplicação para execução no OurGrid, o *Scheduler* que é o componente interno do *MyGrid* encarregado de receber a submissão, requisita aos provedores de GuMs os recursos necessários para a execução da aplicação. Após o recurso ter sido descoberto, se ele estiver disponível será alocado e repassado ao *Scheduler* que o solicitou.

- Transferência de Dados: No aspecto de transferência de dados, o funcionamento do *OurGrid* baseia-se em operações de transferência de arquivos. Essas operações denominadas *put, store e get* são usadas para preparar o ambiente para a execução da tarefa. As operações *put* e *store* são utilizadas na transferência de arquivos para a GuM enquanto que a operação *get* é utilizada na coleta dos arquivos resultantes do processamento. A infra-estrutura de comunicação usada na transferência é a mesma usada na solicitação de serviços de execução, ou seja, *Remote Method Invocation (RMI)*. Uma vez que é possível ter segurança na comunicação com as GuMs de RMI sobre *Secure Socket Layer (SSL)*, as operações de transferências de dados também usufruem da segurança fornecida pela camada de comunicação baseada em certificados.

  Desta forma, o usuário ao descrever sua aplicação tem a sua disposição o uso de três operações de transferência de arquivos, *put, store e get*, que podem ser usadas para preparar o ambiente para execução da aplicação (colocando os arquivos nos *sites* onde a aplicação irá executar), como também coletar os dados resultantes do processamento. Tanto *put* quanto *store* são operações que permitem transferir arquivos para a GuM. A diferença entre as duas operações consiste apenas do fato que *store* evita a transferência do arquivo caso o arquivo já se encontre armazenado no lado remoto. Isso é útil, por exemplo, para executáveis da aplicação e dados que são reutilizados entre execuções sucessivas da aplicação. A terceira operação, *get*, fornece um método de coletar arquivos resultantes da execução das aplicações.

Basicamente, a implementação de um ambiente de grade computacional usando o *middleware* OurGrid consiste em instalar e configurar os três componentes *MyGrid, OurGrid Peer e OurGrid Worker* nas máquinas que farão parte da grade. Uma vez que todos esses componentes estão instalados e ativos na grade, a submissão de tarefas é feita através de *scripts*. Nestes *scripts* implementam-se diversas funções, como por exemplo, a leitura e execução dos respectivos programas do problema, transferência de dados, entre outras.

5 DADOS OBSERVACIONAIS E SIMULADOS UTILIZADOS

Atualmente, contamos com um grande avanço na compreensão dos fenômenos que ocorreram durante a evolução do Universo em virtude da enorme quantidade de dados observacionais disponíveis. Entretanto, como ainda não existem modelos analíticos completos para descrever a origem e a evolução da estrutura cósmica, especialmente a parte crucial que é não linear, uma forma eficiente para estudá-las é com o uso de simulações de N-corpos. O principal objetivo dessas simulações é fornecer previsões robustas, que quando comparadas com as observações permitem restringir os parâmetros cosmológicos.

É importante notar que os avanços obtidos com as simulações de N-corpos cosmológicas normalmente tomaram como base a formação de estruturas de matéria escura (chamadas genericamente de "halos"), sob a premissa de que essa é a componente dominante de massa do Universo. O quadro geral, considerando também processos dissipativos inerentes à matéria bariônica (matéria que interage com a luz, forma planetas, estrelas e galáxias, composta dominantemente, em termos de massa, de partículas chamadas bárions[1]), começa também a avançar com as simulações hidrodinâmicas, embora ainda limitadas em termos de resolução e do entendimento dos processos físicos associados (aquecimento e esfriamento do gas, taxa de formação estelar, retroalimentação por supernovas e núcleos ativos de galaxias, papel dos buracos negros supermassivos, acresção e perda de gás pelas galáxias, encontros e fusões entre galáxias, etc). Nesta tese utilizamos dados de simulações de N-corpos de matéria escura, que serão apresentados nesta seção. Em seguida apresentamos os dados observacionais utilizados.

5.1 Simulação de N-corpos

A interação gravitacional entre dois corpos tem uma forma analítica relativamente simples, o que não acontece quando consideramos um sistema de muitos corpos. Por isso o uso de simulações de N-corpos é fundamental para o entendimento da formação da estrutura cósmica, onde as escalas envolvidas implicam necessariamente o uso de um número assombrosamente grande de partículas. Graças aos enormes avanços tecnológicos dos últimos anos, tem-se meios para tratar o problema da instabilidade gravitacional com alta resolução de massa e espacial, que é a base do

[1] bárions são férmions (descritos pela estatística de Fermi-Dirac) compostos de 3 quarks e, por isso, susceptíveis à força nuclear forte. Os bárions mais importantes são os prótons e os nêutrons.

modelo de formação de estruturas. Os resultados obtidos são comparados com dados observacionais (EFSTATHIOU et al., 1985; BERTSCHINGER, 1998; JENKINS et al., 1998). Além disso, as ferramentas desenvolvidas são frequentemente utilizadas no estudo da formação de aglomerados de galáxias, das interações de galáxias isoladas e da evolução do gás intergaláctico. Sem estas ferramentas os imensos progressos obtidos nestes campos teriam sido quase impossíveis.

Enquanto o crescimento linear inicial das perturbações de densidade pode ser calculado analiticamente, o colapso das flutuações e a subsequente criação hierárquica de estruturas é um processo altamente não linear cuja solução somente pode ser obtida através de simulações numéricas (SPRINGEL et al., 2005).

A simulação de N-corpos gravitacional, nada mais é do que a obtenção de soluções numéricas para as equações do movimento considerando um conjunto de N partículas que interagem gravitacionalmente. A dinâmica que descreve esse comportamento é dada pela Lei de Newton para gravitação universal. Assim, a força $\mathbf{F}_i$ que atua sobre uma partícula de massa m_i gerada pelas partículas de massa m_j é obtida segundo a Equação 5.1:

$$\mathbf{F}_i = -m_i G \sum_{j \neq i} \frac{m_j(\mathbf{r}_i - \mathbf{r}_j)}{|\mathbf{r}_i - \mathbf{r}_j|^3} \tag{5.1}$$

Desta forma, o problema consiste num conjunto de equações diferenciais ordinárias de segunda ordem não linear que relaciona a aceleração com a posição de todas as partículas no sistema. Uma vez que um conjunto de condições iniciais é especificado (por exemplo, a posição e as velocidades iniciais de todas as partículas) existe uma solução única e analítica somente até 2 corpos, sendo que N grande requer integração numérica (AARSETH, 2003). Entretanto a força gravitacional apresenta uma singularidade quando a distância entre duas partículas se aproxima de zero. Dada a natureza não linear das equações, as singularidades dependem da escolha das condições iniciais. Desta forma, métodos de integração com passo de tempo constante não são capazes de garantir uma boa precisão no caso da dinâmica gravitacional (TRENTI; HUT, 2008).

A singularidade pode ser evitada introduzindo na Equação 5.1 um parâmetro de

suavização ϵ, da seguinte maneira:

$$\mathbf{F}_i = -G \sum \frac{m_i m_j (\mathbf{r}_i - \mathbf{r}_j)}{(|\mathbf{r}_i - \mathbf{r}_j|^2 + \epsilon^2)^{3/2}} \tag{5.2}$$

O fator de suavização ϵ serve para modificar a interação gravitacional em pequenas escalas, evitando a formação de sistemas binários não físicos e também para garantir o comportamento não colisional do sistema. Em cada passo de tempo a força gravitacional de cada partícula considerando as contribuições de todas as outras deve ser calculada. Neste caso, a complexidade computacional da solução numérica de um sistema de N corpos é da ordem de N^2. Assim, um conjunto de técnicas foi desenvolvido visando reduzir o tempo computacional (TRENTI; HUT, 2008), ver por exemplo os algoritmos em árvore (BARNES; HUT, 1986) e o clássico método em transformada de Fourier *particle-mesh* (HOCKNEY; EASTWOOD, 1981).

5.2 Simulações do Consórcio Virgo

O consórcio Virgo é uma iniciativa que começou em 1994 na Inglaterra com simulações cosmológicas usando computação de alto desempenho, tornando-se mais tarde um projeto internacional com participação de vários Países (entre eles, Alemanha, Canadá, Estados Unidos e Japão). Este consórcio realizou várias simulações ao longo de duas décadas, sendo composto por vários projetos que englobam áreas de pesquisa tais como: distribuição em larga escala da matéria escura, formação de halos de matéria escura, formação e evolução de galáxias e aglomerados, física do meio intergaláctico, entre outras (VIRGO..., 2008).

Nesta tese de doutorado trabalhamos com dados de dois projetos de simulação de matéria escura, a saber: Simulações de Escala Intermediária (JENKINS et al., 1998) e Simulação *Millennium*(SPRINGEL et al., 2005). Por definição a matéria escura é uma matéria que interage gravitacionalmente mas não eletromagneticamente, ou seja, não emite, não absorve e nem reflete radiação. Logo, somente podemos percebê-la pelos seus efeitos gravitacionais sobre a matéria bariônica. Contudo, sua origem ou natureza ainda é desconhecida (SAHNI; COLES, 1995; SAHNI, 2004). De acordo com as observações atuais, a matéria escura constitui a maior parte da massa do Universo, cerca de 23%. Apenas 4% constitui a matéria bariônica e os 73% restantes seriam constituídos de energia escura.

Sabe-se desde os anos 1930 que existe muito mais matéria no Universo do que a

matéria visível. Fritz Zwicky foi a primeira pessoa a sugerir a existência de uma matéria que não se pode ver. Em um trabalho de 1933, publicado em inglês em 1937 (ZWICKY, 1937), usando o teorema do Virial, estimou a massa do aglomerado de Coma necessária para manter as galáxias se movendo na velocidade observada. Os resultados foram comparados com a soma das massas individuais das galáxias. Zwicky encontrou que, se este aglomerado está ligado gravitacionalmente e em equilíbro, então sua massa é duas ordens de magnitude maior do que a soma das massas individuais de suas galáxias.

Muitas das evidências da matéria escura vêm do estudo das curvas de rotação das galáxias. De acordo com o teorema do virial, a energia cinética total deveria ser a metade da energia potencial gravitacional das galáxias. Porém, experimentalmente se tem encontrado que a energia cinética total é muito maior: em particular, assumindo que a massa gravitacional é devida somente a matéria visível, as estrelas distantes do centro da galáxia têm velocidades muito maiores do que as previstas pelo teorema do Virial. A curva de rotação galáctica que ilustra a velocidade de rotação em função da distância do centro da galáxia não pode ser explicada apenas pela matéria visível - ver Figura 5.1.

A maneira encontrada para explicar tal comportamento foi assumir que a matéria visível compõe apenas uma pequena parte das estruturas: as galáxias parecem possuir um halo de matéria escura, com simetria aproximadamente esférica, ao redor de sua estrutura visível (que inclui bulbo, disco e halo visível, no caso das galaxias espirais por exemplo). Em maiores escalas, os halos de matéria escura das galáxias em grupos e aglomerados se juntam para formar os grandes halos de matéria escura desses sistemas (BAHCALL et al., 1995).

Nas Simulações de Escala Intermediária, quatro modelos de evolução da matéria escura são apresentados, dos quais foi escolhido o modelo $\Lambda\ CDM$ (Λ : constante cosmológica; CDM: *Cold Dark Matter*). Assume-se que a componente de massa dominante, a matéria escura fria, é feita de partículas elementares que interagem somente gravitacionalmente, assim o fluido de matéria escura acolisional pode ser representado por um conjunto de partículas pontuais discretas. Quanto maior o número de partículas considerado, melhor será a precisão do modelo.

Os parâmetros numéricos que caracterizam essa simulação são: um cubo de aresta $L = 239.5\ h^{-1}\ Mpc$, número de partículas $N = 256^3$ (cada partícula

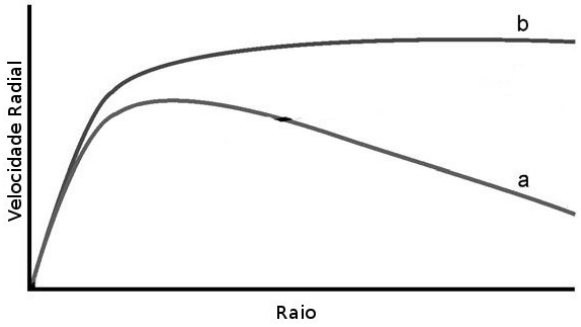

Figura 5.1 - Esquema das curvas de rotação esperada e observada da Via Lactea. (a) Esquema esperado a partir das estrelas e do gás. (b) Esquema da curva de rotação observada.
Fonte: Adaptada de Freeman e MacNamara (2006).

tem massa $6.86 \times 10^{10} M_{\odot}$) e os parâmetros cosmológicos são: $(\Omega_M, \Omega_\Lambda, h, \sigma_8) = (0.3, 0.7, 0.7, 0.9)$. Os dados estão disponíveis para *download* no seguinte endereço: `http://www.mpa-garching.mpg.de/Virgo/datadownload.html`.

A Simulação *Millennium* é uma das maiores simulações da evolução do Universo. Esta simulação emprega mais de 10 bilhões de partículas de matéria escura em uma região cúbica do Universo com 2 bilhões de anos-luz de lado a partir do *redshift* $z = 127$ até o presente ($z = 0$). Ela começa em aproximadamente 379 000 anos após o *Big Bang* (o qual ocorreu cerca de 13.8 bilhões de anos atrás), quando o Universo era extremamente quente e denso. Nesta época, o Universo estava atravessando a transição do domínio da radiação para o domínio da matéria. Com o Universo se expandindo e esfriando, ele chega a uma temperatura crítica de cerca de 3 000 K quando começa a desacoplar a radiação da matéria. Este evento produziu a radiação cósmica de fundo, que hoje apresenta uma temperatura de aproximadamente 2.7 K (MADSEN, 1996). Devido as observações detalhadas da radiação cósmica de fundo, os físicos/astrônomos têm uma boa idéia sobre o estado do Universo no instante do desacoplamento, essas informações são utilizadas como condições iniciais nas simulações.

Os parâmetros numéricos que caracterizam a Simulação *Millennium* são: um cubo de aresta $L = 500\ h^{-1}\ Mpc$, número de partículas $N = 2160^3$ (cada partícula tem massa $8.6 \times 10^8\ h^{-1}\ M_{\odot}$) e os parâmetros cosmológicos são: $(\Omega_M, \Omega_\Lambda, h, \sigma_8) = (0.25, 0.75, 0.73, 0.9)$, com σ_8 a amplitude da flutuação de densidade de massa em

uma esfera de 8 Mpc no *redshift* 0.

A Figura 5.2 apresenta o campo de densidade de matéria escura em várias escalas, produzido com a simulação *Millennium*. Nas grandes escalas existem poucas estruturas distinguíveis e a distribuição aparece homogênea e isotrópica. As imagens subsequentes têm um zoom com um fator de 4 em uma região ao redor de um aglomerado rico de galáxias. A imagem final revela centenas de subestruturas de matéria escura, ligando gravitacionalmente os objetos dentro do aglomerado. Estas subestruturas são remanescentes de halos de matéria escura menores que se fusionaram ao aglomerado em épocas primordiais (SPRINGEL et al., 2005).

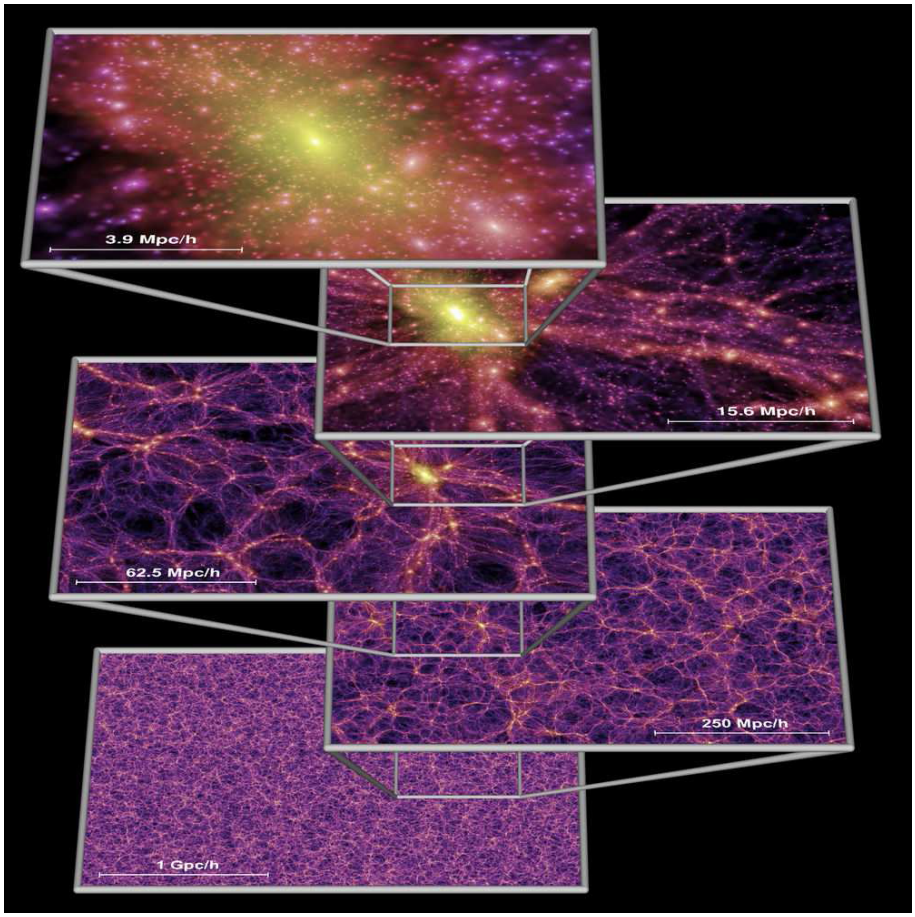

Figura 5.2 - Campo de densidade de matéria escura em várias escalas.
Fonte: Springel et al. (2005).

Os dados desta simulação podem ser obtidos em: `http://www.mpa-`

garching.mpg.de/millennium/, onde tem-se uma breve introdução sobre a simulação, links para imagens e animações e também para o site do projeto GAVO que descreve em detalhes a estrutura de dados implementada, o padrão SQL e os procedimentos para acesso e download dos dados. A simulação *Millennium* foi executada com uma versão otimizada do código GADGET2 (SPRINGEL et al., 2001), usando o metodo *TreePM* (XU, 1995) para avaliar as forças gravitacionais.

5.3 O Projeto *Sloan Digital Sky Survey* (SDSS)

Atualmente a astronomia conta com ambiciosos projetos tipo "varreduras" que fazem levantamentos dos objetos astronômicos em determinadas regiões do céu e disponibilizam os dados para a comunidade científica. Como exemplos deste tipo de projeto podemos citar: *Digital Palomar Observatory Sky Survey* (DPOSS), UKIRT *Infrared Deep Sky Survey* (UKIDSS) e o SDSS.

O SDSS é um dos projetos de varredura mais importante e influente da história da astronomia. Teve início em 2002 e está na sua terceira fase (SDSS-III, 2008-2014). Em oito anos de operações (SDSS-I, 2000-2005; SDSS-II, 2005-2008) foram obtidas imagens fotométricas que cobrem mais de um quarto do céu e um mapa tridimensional com informações espectroscópicas de 1.6 milhões de objetos, dos quais 930 000 são galáxias, 120 000 são quasares e 460 000 são estrelas.

Para o levantamento fotométrico e espectroscópico, o SDSS usa um telescópio de 2.5 metros localizado no *Apache Point Observatory*, Novo México - EUA, equipado com um conjunto de 30 câmeras CCDs de 2048×2048 *pixels* e um par de espectrógrafos. As câmeras[2] operam com um conjunto de 5 filtros (u, g, r, i, z) em que cada filtro cobre uma determinada banda do espectro eletromagnético, conforme se verifica na Figura 5.3. Cada espectrógrafo[3] contém 320 fibras óticas de 3" de diâmetro e coleta espectros em 2 CCDs, cobrindo um intervalo de comprimento de onda de 3800Å a 6100 Å e o outro de 5900 Å a 9100 Å. Na Tabela 5.1 tem-se o volume total de dados gerados nestes oito anos. Os dados do SDSS são os melhores dados de astronomia ótica que a comunidade científica tem disponível hoje. De acordo com Szalay (2007), o SDSS é um "projeto de genoma cósmico" para o céu do hemisfério norte.

[2]http://www.sdss.org/DR7/instruments/imager/
[3]http://www.sdss.org/DR7/instruments/spectrographs/

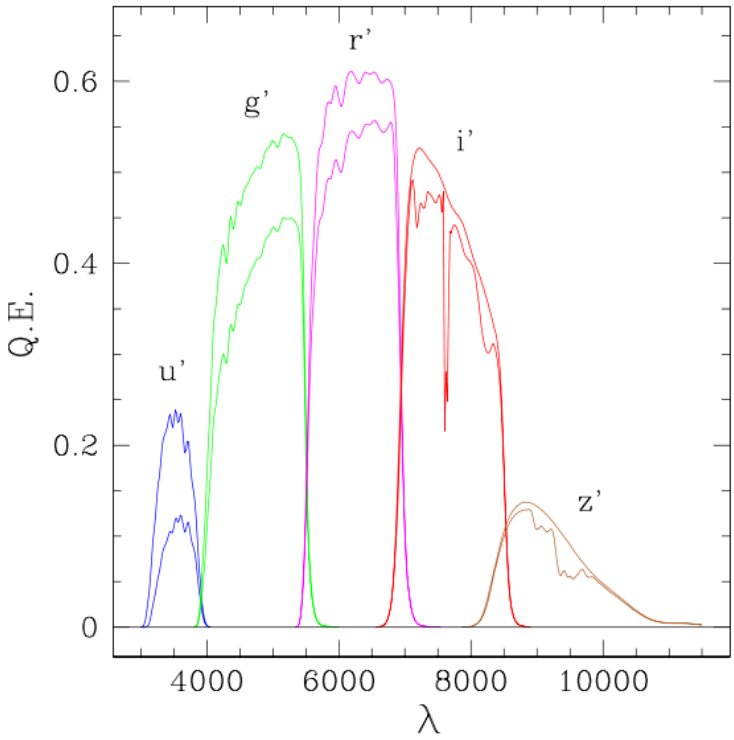

Figura 5.3 - Função resposta de cada filtro usado no SDSS e seus respectivos comprimentos de onda.
Fonte: Gunn et al. (1998).

Tabela 5.1 - Volume de dados gerados pelo catálago DR7

Imagens (Fits)	**Dados de Produtos**	**Catálogos (CAS, SQL database)**
15.7 TB	26.8 TB	18 TB

5.3.1 Amostra Espectroscópica Utilizada

Os dados observacionais utilizados neste trabalho são provenientes do *Data Release* 7 (ABAZAJIAN et al., 2009) do projeto SDSS. Esse catálogo contém aproximadamente 357 milhões de objetos (SEGUE: 127 milhões, Legacy: 230 milhões) em cinco bandas fotométricas diferentes. Os dados do DR7 são disponibilizados através da seguinte página: http://www.sdss.org/dr7/ e o *download* pode ser feito via submissão de *queries* no seguinte endereço: http://cas.sdss.org/astrodr7/en/.

Nesta tese trabalhamos com a amostra espectroscópica de galáxias do SDSS, porém

usando uma base de dados do Projeto SEAGal/Starlight[4], disponível para *download* em: http://casjobs.starlight.ufsc.br/casjobs/. Nesse projeto foi obtido a síntese espectral para todas as galáxias da amostra espectroscópica do DR7 com redshift $z > 0.002$, usando o método denominado Starlight (CID FERNANDES et al., 2005). A amostra gerada contém 926 246 galáxias, os resultados disponibilizados para essa amostra incluem idades médias, metalicidade, massas, magnitudes absolutas com correção do efeito da extinção galáctica e a correção-k, entre outros.

Na Figura 5.4 se pode observar a área projetada no plano do céu para o levantamento espectroscópico total. A partir do conjunto de dados com 926 246 galáxias, definimos

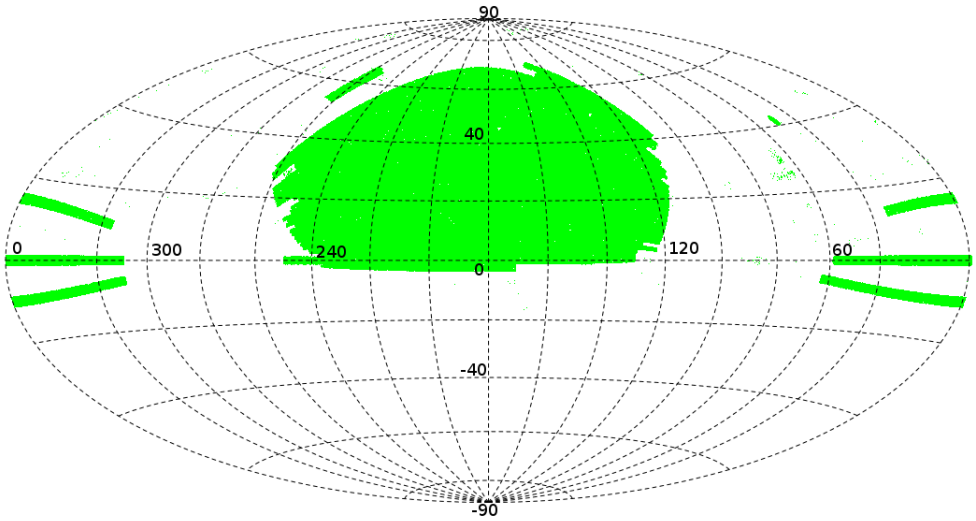

Figura 5.4 - Cobertura projetada do levantamento espectroscópico (Legacy DR7) no plano do céu (coordenadas equatoriais).

os limites da distribuição projetada da amostra a ser utilizada, excluindo os objetos que estão nas bordas, de modo a obtermos um volume bem definido, conforme se pode verificar na Figura 5.5. Em seguida construímos histogramas de magnitudes aparentes e *redshifts* (Figuras 5.6 e 5.7), que servem para indicar os limites de completeza da amostra. Como se verifica nestas figuras a amostra é provavelmente completa para $r \leq 17.75$ e $z \leq 0.1$.

Berlind et al. (2006), sugerem a construção de subamostras que sejam completas em determinados intervalos de *redshifts* e cortes em magnitudes absolutas para contor-

[4]http://www.starlight.ufsc.br/

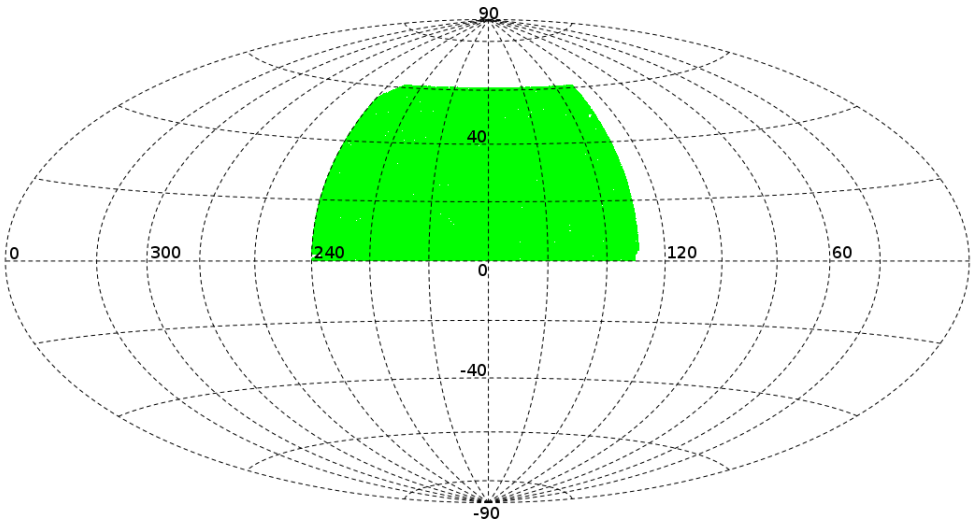

Figura 5.5 - Cobertura projetada no plano do céu da amostra selecionada para o presente trabalho.

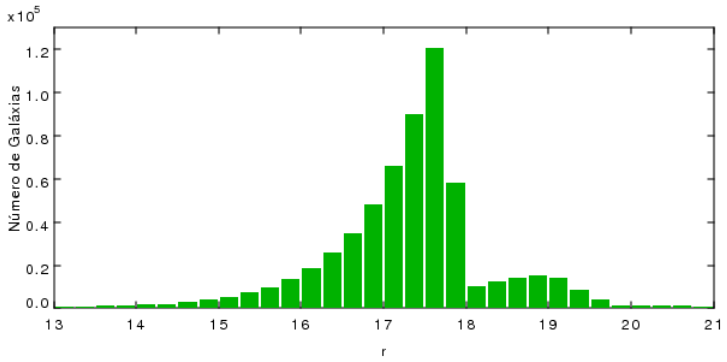

Figura 5.6 - Histograma de Magnitudes.

nar o efeito Malmquist. Assim, selecionamos três intervalos de *redshifts* para nossa amostra: $0.002-0.1, 0.002-0.068$ e $0.002-0.045$. Conforme se observa na Figura 5.8, essas amostras são aproximadamente completas para magnitudes absolutas na banda r respectivamente menores que $-20.6, -19.7$ e -18.8. Desse modo, conforme se verifica na Tabela 5.2, nossa amostra final é composta por 3 conjuntos de dados com 127 588, 78 391 e 39 204 galáxias, respectivamente.

Tabela 5.2 - Parâmetros das amostras limitadas em volume

Amostra	$Z_{min} - Z_{max}$	M_r	Nº de galáxias
Amostra 1	0.002 - 0.100	≤ -20.6	127 588
Amostra 2	0.002 - 0.068	≤ -19.7	78 391
Amostra 3	0.002 - 0.045	≤ -18.8	39 204

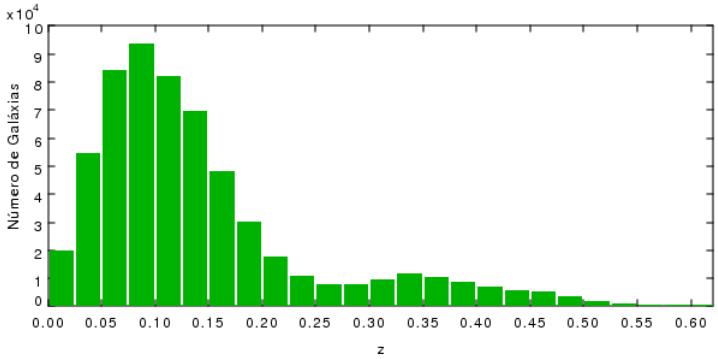

Figura 5.7 - Histograma de *Redshift*.

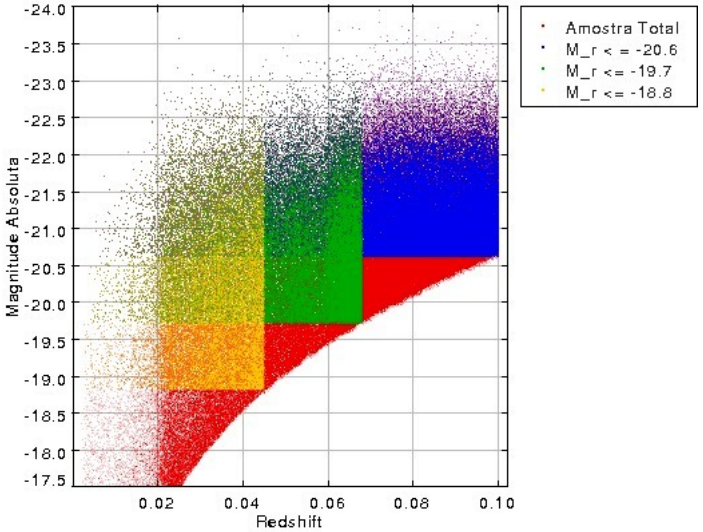

Figura 5.8 - Distribuição das magnitudes absolutas por *redshift* para as galáxias da amostra (pontos vermelhos). As caixas em amarelo, verde e azul representam, respectivamenteno, as subamostras 1, 2 e 3.

Porém, a essas amostras é necessário ainda aplicar uma correção para o efeito de incompleteza produzido pelo efeito de "colisão de fibras"(o chamado *fiber colision effect*). Esse efeito ocorre porque o sistema do espectrógrafo multi-fibras do SDSS não permite a colocação de fibras com separações angulares equivalentes menores que 55 segundos de arco (STRAUSS et al., 2002), levando a que regiões com maiores densidades de galáxias sejam subamostradas (parte das galáxias não é observada espectroscopicamente). Como existe uma relação entre a magnitude aparente das galáxias e sua densidade superficial no céu (há menos galáxias brilhantes e mais galáxias fracas por unidade de área), esse efeito produz maior perda em maiores profundidades do catálogo. Além disso, como galáxias com maiores magnitudes também têm,

em média, maiores *redshifts*, é necessária uma correção adicional com a distância entre nós e a galáxia (GAZTANAGA; YOKOYAMA, 1993; SELJAK et al., 2009). Para corrigir esses efeitos Costa-Duarte et al. (2011) propõem um modelo para a função seleção (S) que depende da magnitude aparente e da distância co-móvel e é definida da seguinte maneira:

$$S(m_r, d_c) = S_1(m_r)S_2(d_c) \tag{5.3}$$

em que $S_1(m_r) = 0.588605 - 1.941834m_r + 0.419142m_r^2 - 0.029956m_r^3 + 0.000724m_r^4$ e $S_2(d_c) = a * d_c + b$ com $a = 0.0025$ e $b = 0.1565$ para distâncias em Mpc. Assim, para corrigir as densidades locais no catálogo se pode aplicar uma função peso (W_i), que é a inversa da função seleção: $W = S(mr, dc)^{-1}$. Entretanto, em nosso caso não estamos lidando diretamente com densidades, mas com separações espaciais entre galáxias (para calcular suas energias potenciais gravitacionais). Se tomarmos $J(r)$ como a densidade de luminosidade num certo ponto r, teremos:

$$J(r) = \sum L_i W_i \tag{5.4}$$

onde L_i e W_i são, respectivamente, a luminosidade e a função peso da *i-ésima* galáxia. Assim, tem-se que $\sum W_i$ pode ser dado pelo inverso do volume, ou seja:

$$W_i = \frac{3}{4\pi R_i^3} \tag{5.5}$$

Logo,

$$R_i = \sqrt[3]{\frac{3}{4\pi W_i}} \tag{5.6}$$

onde R_i é a função peso a ser aplicada nas posições ou distâncias para corrigir a perda.

O segundo termo ($S2$) da função seleção proposta por Costa-Duarte et al. (2011), que depende da distância co-móvel das galáxias, foi calculado para um intervalo de *redshift* ($0.04 < z < 0.155$) que não coincide muito bem com o intervalo considerado neste trabalho ($0.002 < z < 0.1$) e, como representa uma correção adicional menor, decidimos nao aplicá-lo. Testes considerando ou não este segundo termo revelaram que não afeta significativamente os resultados.

5.3.2 Determinação das Distâncias entre Pares de Galáxias da Amostra

A separação física entre dois objetos celestes pode ser obtida a partir de sua posição em um sistema de coordenadas esférico centrado na Terra. As coordenadas angulares desse sistema nos dão a posição do objeto na esfera celeste e a terceira coordenada é dada pela distância entre nós e o objeto. No sistema de Coordenadas Equatoriais Celestes, as coordenadas angulares são chamadas respectivamente *ascensão reta* (A.R. ou α) e *declinação* (Dec. ou δ). Conforme mostra a Figura 5.9, a *separação angular* (ou *separação projetada*), $\Delta\Theta$, pode ser obtida a partir do Teorema de Pitágoras, utilizando as diferenças entre as coordenadas α e δ dos dois objetos:

$$(\Delta\Theta)^2 = (\Delta\delta)^2 + (\Delta\alpha\ cos\delta)^2 \tag{5.7}$$

Essa separação angular deve ser convertida em uma *separação física* (ou seja, em unidades que tenham dimensão), ΔS (Ver Figura 5.10). Para isso é preciso considerar as distâncias entre nós e os dois objetos, que são distâncias estimadas levando em conta os parâmetros Cosmológicos. A distância adequada nesse caso é a distância de diâmetro angular, D_A, conforme foi definida na Equação 2.10. Para a conversão podemos utilizar a distância média entre os dois objetos, ou seja:

$$\Delta S = \frac{(D_{A1} + D_{A2})}{2} \times \tan(\Delta\Theta)$$

Em seguida temos que estimar a distância de diâmetro angular para a separação entre os dois objetos na linha de visada, que não é simplesmente a diferença entre as distâncias de diâmetro angular dos dois objetos, conforme foi definido na Subseção 2.1.1 (Equação 2.11). Considerando $\Omega_k = 0$, a Equação 2.11 torna-se:

$$D_{A_{12}} = \frac{1}{1+z_2}[D_{M2} - D_{M1}]$$

E assim, podemos então combinar, utilizando outra vez o Teorema de Pitágoras, as separações física projetada (ΔS) e na linha de visada (D_{A12}):

$$(\Delta r)^2 = (\Delta S)^2 + (D_{A_{12}})^2$$

Substituindo ΔS e $D_{A_{12}}$ na equação acima encontramos a separação física espacial entre os dois objetos (Δr).

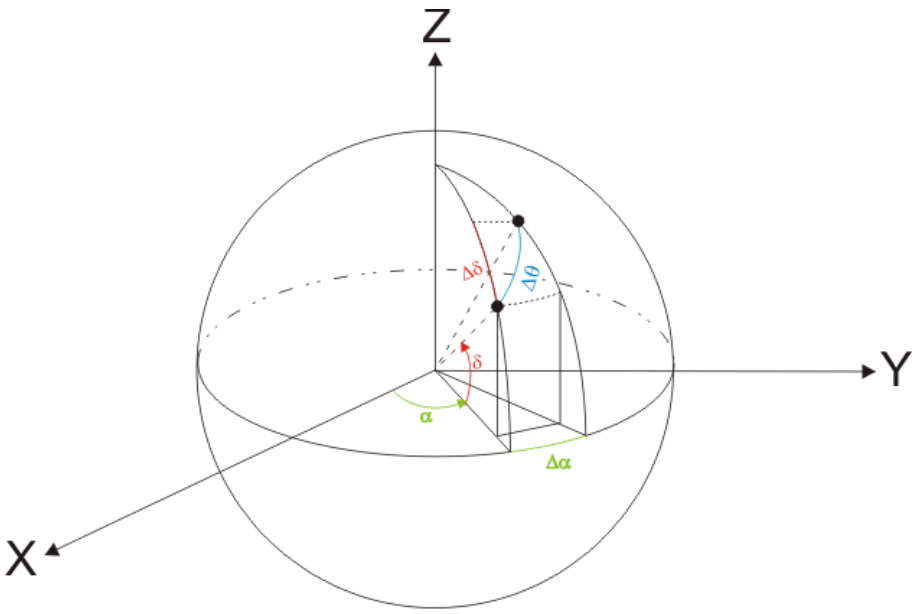

Figura 5.9 - Separação angular ($\Delta\Theta$) entre dois objetos .

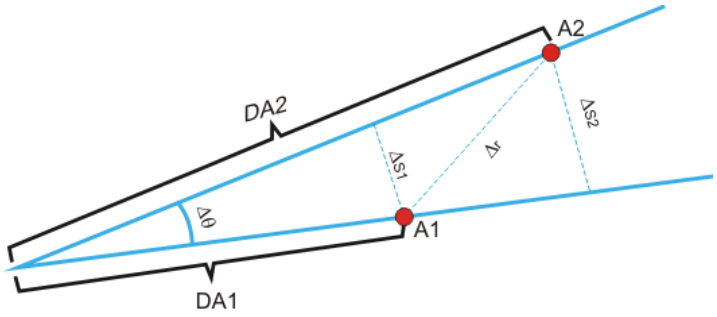

Figura 5.10 - Distância física (Δr) entre dois objetos celestes.

6 FERRAMENTAS DE ANÁLISE DE DADOS ASTRONÔMICOS

No Século 20, o grande desafio científico era encontrar soluções para as equações que governam as leis da natureza. A grande solução foi dada pela computação, por meio de algoritmos que permitiram a geração de soluções aproximadas, as chamadas soluções numéricas. Com isso, a computação se consolidou como o terceiro pilar no processo científico, ao lado da teoria e dos experimentos. Agora, no Século 21, o grande desafio é a análise de dados, ou seja, como extrair informações relevantes (conhecimentos) a partir da tsunami de dados que estamos vivenciando em todas as áreas. Esses dados podem ser experimentais, observacionais, simulados, entre outros. Formalmente, o termo designado para esta abordagem é "Ciência dos Dados" - do inglês: *Data Science* ou "O quarto paradigma", conforme foi definido por Hey et al. (2009). A computação, mais uma vez, está sendo a chave para a solução deste desafio científico.

Neste contexto, um conceito muito popular que surge é a mineração de dados. A mineração de dados pode ser entendida como um resultado natural da evolução da tecnologia da informação. Se refere a extrair ou "minerar" dados a partir de grandes quantidades. Ferramentas de mineração realizam análises nos dados buscando padrões importantes e capazes de gerar conhecimentos (HAN; KAMBER, 2001). O processo de mineração de dados envolve uma integração de técnicas vindas de várias áreas, a saber: Estatística, Reconhecimento de Padrões e Inteligência Artificial. O processo de mineração de dados geralmente envolve quatro classes de tarefas: Regras de associação, agrupamento, classificação e regressão. As regras de associação são usadas para descobrir elementos que ocorrem em comum dentro de um determinado conjunto de dados. A tarefa de agrupamento envolve técnicas que permitem descobrir grupos e estruturas que são similares nos dados, sem usar estruturas conhecidas *a priori*. Já a tarefa de classificação identifica(classifica) os elementos com base em classes já conhecidas ou pré-definidas. Finalmente, a regressão é aprender uma função que mapea um item dos dados para uma variável de predição real, com o menor erro possível (FAYYAD et al., 1996).

Uma vez que a astronomia é uma ciência que vem experimentando uma avalanche de dados nas últimas décadas, é imprescindível promover o desenvolvimento de eficientes ferramentas de análise destes dados. Neste sentido, abordamos duas ferramentas de análise de dados, a saber: árvore de decisão e o algoritmo *Friends of Friends*.

6.1 Árvore de Decisão na Classificação de Estrelas e Galáxias

Um correto mapeamento do céu necessita uma forma eficiente de separar objetos pontuais (estrelas individuais) de objetos extensos (galáxias, nebulosas, etc). Especificamente, para a Astronomía Extragalática é importante aumentar a completeza e reduzir a contaminação em amostras de galáxias até limites cada vez maiores de magnitude. A completeza se refere à fração de galaxias reais, dentro dos limites da amostra, que foram classificadas como tal, enquanto a contaminação se refere à fração de outros objetos, normalmente estrelas, que foram erroneamente classificadas como galáxias. Porém, em uma imagem astronômica, nem sempre é fácil fazer a separação entre uma estrela e uma galáxia, principalmente porque quanto maior a distância, cada vez vemos somente as partes mais brilhantes das galáxias, que são normalmente os seus núcleos, os quais se assemelham a estrelas nessas condições.

Os métodos de classificação, de um modo geral, usam um conjunto de parâmetros para classificar os objetos de acordo com suas características e similaridades. Basicamente o classificador é obtido a partir de um conjunto de amostras, denominado conjunto de treinamento. Quanto ao tipo de treinamento a classificação pode ser supervisionada ou não-supervisionada (THEODORIDIS; KOUTROUMBAS, 2003; WITTEN; FRANK, 2000)

Na classificação supervisionada, dado um conjunto de amostras previamente rotuladas, o desenvolvimento de um classificador pode ser visto basicamente como a tarefa de se obter uma função que realize o mapeamento de cada amostra desse conjunto a sua respectiva classe. Essa fase é denominada fase de treinamento. Após a fase de treinamento, tem-se a fase de validação e teste do modelo, na qual o desempenho do classificador é avaliado mediante sua aplicação na classificação de um novo conjunto de amostras que não foram utilizadas no treinamento. Um bom classificador deve ser capaz de prever, com certo grau de precisão, a classe a qual pertence cada amostra do conjunto de entrada. De um modo geral, em classificação supervisionada os dados utilizados no treinamento são padrões rotulados, ou seja, pré-classificados e o objetivo é classificar novos padrões ainda não-rotulados (BISHOP, 2006; KOTSIANTIS, 2007; THEODORIDIS; KOUTROUMBAS, 2003).

Por outro lado, na classificação não-supervisionada, cada amostra do conjunto de treinamento é associada a um grupo e não é necessário conhecer previamente o rótulo das diferentes classes presentes no conjunto. Geralmente, isto é feito através

de algoritmos de agrupamento. Um algoritmo de agrupamento divide um conjunto de padrões não-rotulados em grupos cujos elementos possuam alguma similaridade, de modo que os dados em cada grupo sejam mais similares entre si do que em relação aos dados de qualquer outro grupo (THEODORIDIS; KOUTROUMBAS, 2003).

Considere um conjunto de objetos que são descritos em termos de uma coleção de atributos. Estes objetos podem pertencer a diferentes classes. Cada atributo mede alguma característica importante de um objeto, considere também um conjunto de treinamento, cuja classe de cada objeto é conhecida. Conforme Quinlan (1986) é possível desenvolver uma regra de classificação que pode determinar a classe de qualquer objeto a partir dos valores dos seus atributos. Tal regra de classificação pode ser expressa como uma árvore de decisão. Uma árvore de decisão é uma estrutura simples em que as folhas contêm as classes, os nodos não-folhas representam atributos baseados em testes com um ramo para cada possível saída (QUINLAN, 1986; QUINLAN, 1993). Para classificar um objeto, começa-se com a raiz da árvore, aplica-se o teste e toma-se o ramo apropriado para aquela saída. O processo continua e quando uma folha é encontrada o objeto é classificado segundo a classe indicada naquela folha.

Se os atributos são adequados é sempre possível construir uma árvore de decisão que classifique corretamente os objetos no conjunto de treinamento. O interessante é ir além do conjunto de treinamento, isto é, classificar corretamente outros objetos. Para conseguir isso, a árvore de decisão deve capturar alguma relação significativa entre a classe do objeto e os valores de seus atributos. Quando se tem duas árvores de decisão que classificam corretamente um conjunto de treinamento, deve-se escolher a mais simples, uma vez que, ela é mais adequada para capturar a estrutura inerente do problema e assim, vai classificar corretamente mais objetos fora do conjunto de treinamento (QUINLAN, 1986).

São diversos os algoritmos de indução (construção) de árvores de decisão conhecidos na literatura. O algoritmo $C4.5$, desenvolvido por Quinlan (1986), é um dos algoritmos mais populares. A ideia básica deste algoritmo é iterativa. Um subconjunto do conjunto de treinamento chamado janela é escolhido aleatoriamente e uma árvore de decisão é formada a partir dele. Todos os outros objetos do conjunto de treinamento são classificados usando a árvore. Se esta árvore fornecer a resposta correta para todos os objetos o processo termina, se não, uma seleção dos objetos classificados incorretamente é adicionada a janela e o processo continua.

A essência do problema de classificação, usando árvores de decisão, está em como formar uma árvore para uma coleção C de objetos. Se C é vazio ou contém somente objetos de uma classe, a árvore de decisão mais simples é justamente uma folha classificada com aquela classe. Caso contrário, seja T qualquer teste sobre um objeto que tem os possíveis resultados $O_1, O_2, ..., O_w$. Cada objeto em C dá um desses resultados para T, portanto T produz uma partição $\{C_1, C_2, ..., C_w\}$ de C, com C_i contendo aqueles objetos que tem saída O_i. No pior caso essa estratégia fornecerá subconjuntos de um único objeto, que satisfaz a exigência de uma classe por folha. Assim, uma vez que um teste pode ser encontrado de uma divisão não trivial de qualquer conjunto de objetos, este procedimento sempre permite obter uma árvore de decisão que classifique corretamente os objetos em C (QUINLAN, 1986). A escolha do teste é crucial para a árvore de decisão ser simples. O $C4.5$ adota um critério baseado na teoria da informação que depende de duas hipóteses:

- Hipótese 1: Toda árvore de decisão correta para C classificará objetos na mesma proporção que sua representação em C. No caso de uma amostra de objetos que pertencem somente a duas classes, por exemplo, P e N, um objeto qualquer pertencerá a classe P com probabilidade $\frac{p}{p+n}$ e a classe N com probabilidade $\frac{n}{p+n}$, em que p é o número total de objetos pertencentes a classe P e n é o número total de objetos pertencentes a classe N.

- Hipótese 2: Quando uma árvore de decisão é usada para classificar um objeto, ela retorna uma classe. Assim, a árvore de decisão pode ser considerada como uma fonte de mensagem P ou N em que a informação necessária para gerar a mensagem é obtida conforme Equação 6.1:

$$I(p,n) = -\frac{p}{p+n} log_2 \left(\frac{p}{p+n}\right) - \frac{n}{p+n} log_2 \left(\frac{n}{p+n}\right) \tag{6.1}$$

Se o atributo A com os valores $[A_1, A_2, ..., A_v]$ é usado para a raiz da árvore de decisão, ela dividirá C em $[C_1, C_2, ..., C_v]$, em que C_i contém aqueles objetos em C que têm valores A_i de A. Considere C_i contendo p_i objetos da classe P e n_i objetos da classe N, a informação necessária para a subárvore em C_i é $I(p_i, n_i)$. A informação necessária para a árvore com A como raiz é obtida com a média ponderada:

$$E(A) = \sum_{i=1}^{v} \left(\frac{p_i + n_i}{p+n}\right) I(p_i, n_i) \tag{6.2}$$

O peso para o *i-ésimo* ramo é proporcional aos objetos que pertencem a C_i. O ganho de informação obtido por esse ramo usando o atributo A é dado pela Equação 6.3.

$$G(A) = I(p, n) - E(A) \tag{6.3}$$

O algoritmo $C4.5$ examina todos os atributos candidatos e escolhe A que maximiza o ganho de informação, forma as árvores como dito anteriormente e então usa o mesmo processo recursivamente para formar a árvore de decisão para os subconjuntos restantes.

6.1.1 Avaliação de desempenho do classificador

O desempenho do classificador geralmente é avaliado através da matriz de confusão. A matriz de confusão indica quais instâncias foram classificadas de forma correta e incorreta. Ela é quadrada e de dimensão igual ao número de classes avaliadas. Os resultados da classificação são colocados nas colunas da matriz, sendo que na diagonal principal encontra-se o número de observações classificadas corretamente. Se todo elemento fora das diagonais é igual a zero tem-se uma classificação 100% correta (WITTEN; FRANK, 2000).

A partir da matriz de confusão pode-se obter uma série de medidas estatísticas para o processo de validação, das quais vale destacar a exatidão total e o índice *kappa*. A exatidão total é a medida mais simples e relaciona os elementos da diagonal com o total de objetos, sendo dada pela Equação 6.4

$$E_{Total} = \sum_{i=1}^{r} X_{ii} \times \frac{100}{N} \tag{6.4}$$

em que X_{ii} são os elementos da diagonal, N é o número total de objetos e r é o número de linhas ou colunas da matriz.

O coeficiente *kappa* pode ser obtido conforme Equação 6.5

$$\hat{\kappa} = \frac{N \sum_{i=1}^{r} X_{ii} - \sum_{i=1}^{r} X_{i+} X_{+i}}{N^2 - \sum_{i=1}^{n} X_{i+} X_{+i}} \tag{6.5}$$

com $X_{i+} = \sum_j X_{ij}$ é a soma dos valores da linha i e $X_{+i} = \sum_j X_{ji}$ é a soma dos valores da coluna i. Quanto mais próximo de 1 for o valor do índice *kappa* melhor é

o desempenho do classificador.

Ruiz et al. (2009), usou o algoritmo de construção de árvore de decisão $C4.5$ no desenvolvimento de classificadores baseados em atributos fotométricos para classificação de estrelas e galáxias, utilizando dados da disponibilização SDSS-DR6. Os testes foram realizados utilizando o *software Waikato Environment for Knowledge Analysis* (WEKA). O WEKA tem por objetivo agregar algoritmos provenientes de diferentes abordagens na subárea da inteligência artificial dedicada ao estudo da aprendizagem de máquinas. Essa subárea pretende desenvolver algoritmos e técnicas que permitam a um computador "aprender" (no sentido de obter novo conhecimento) quer indutiva, quer dedutivamente (WITTEN; FRANK, 2000). Da amostra de dados do SDSS, composta por 43 289 estrelas e 452 400 galáxias (total 495 689 objetos), o treinamento (criação das árvores) foi realizado com 925 estrelas e 9 075 galáxias (total 10 000 objetos). Com essa amostra foram criadas 2 árvores que foram implementadas em linguagem C e testadas sobre a amostra total. A Tabela 6.1 apresenta o desempenho dos classificadores desenvolvidos. Uma descrição detalhada deste trabalho pode ser verificada no Apêndice A.

Tabela 6.1 - Resultados obtidos com as duas árvores de decisão sobre o conjunto de 495 689 objetos astronômicos.

	1ª **árvore**		2ª **árvore**	
	Estrelas	**Galáxias**	**Estrelas**	**Galáxias**
Estrelas	42 742	523	42 708	568
Galáxias	1 866	450 482	2 344	450 011
Objetos não classificados	24	52	13	45
Índice de acerto	98.79%	99.59%	98.69%	99.48%
Índice kappa	0.97		0.96	

Posteriormente, esse estudo foi ampliado por Vasconcellos (2011), Vasconcelos et al. (2011), que, em seu trabalho de mestrado, analisou a eficiência e o desempenho de 13 algoritmos baseados em árvore de decisão. O algoritmo que apresentou melhor desempenho e eficiência foi utilizado para desenvolver um classificador que foi aplicado na classificação de um conjunto de 69 545 326 objetos da amostra fotométrica de dados do SDSS-DR7 com magnitudes no intervalo $14 < r < 21$.

6.2 Algoritmo *Friends of Friends*

O primeiro passo na obtenção do espectro de energia potencial gravitacional, usando dados de simulação de matéria escura, é identificar no domínio em diferentes escalas do tempo e do espaço, os halos de matéria escura. Existem vários métodos para identificar halos de matéria escura em simulações de N-corpos, entre eles destacamos o algoritmo *Spherical Overdensity* (SO)(LACEY; COLE, 1994), DENMAX (GELB; BERTSCHINGER, 1994), SUBFIND (SPRINGEL et al., 2005) e o algoritmo *Friends of Friends*(HUCHRA; GELLER, 1982), o qual foi utilizado nesta tese.

O Algoritmo SO identifica picos de densidade e coloca esferas ao redor deles, aumentando o raio da esfera até que a densidade média atinge um determinado valor estabelecido. O método é baseado em encontrar regiões esféricas em uma simulação tendo uma certa sobredensidade média que é representada por $\kappa = \langle\rho\rangle / \bar{\rho}$. Primeiro, calcula-se a densidade local para cada partícula encontrando a distância r_N até o N-ésimo vizinho mais próximo. Toma-se a partícula com densidade mais alta como candidata ao centro da primeira esfera. A esfera cresce ao redor desse centro, com o raio sendo aumentado até a sobredensidade média cair abaixo do valor κ. O centro de massa das partículas nesta esfera é tomado como um novo centro e o processo de crescimento da esfera é repetido. Este processo é iterado até o deslocamento no centro entre sucessivas iterações, cair abaixo de $\epsilon\bar{n}^{\frac{-1}{3}}$. As partículas na esfera são rotuladas como pertencentes ao mesmo grupo e são então removidas da lista de partículas. O procedimento é repetido até encontrar todos os grupos, em seguida, quaisquer grupos que estejam dentro de grupos maiores são mesclados com o grupo maior. Uma desvantagem desse método, é que ele tende a perder as partes externas dos halos elipsoidais, devido à suposição de simetria esférica.

No algoritmo DENMAX, primeiro estima-se a densidade em cada ponto do espaço e em seguida, permitindo que cada partícula determine seu grupo, traça-se um caminho ao longo do gradiente desta superfície de densidade até alcançar um máximo local. Todas as partículas que terminam no mesmo máximo local são designadas para o mesmo grupo. Infelizmente, a resolução com a qual define-se o campo de densidade resulta em um compromisso entre a capacidade de reconhecer os grupos menores e a tendência de dividir os grupos devido a descoberta de vários máximos locais em uma resolução mais fina. Essas divisões podem ser indesejáveis e é difícil atribuir importância física à forma em que as partículas de baixa densidade são atribuídas a grupos com base nos detalhes da localização de vários máximos locais nas regiões

de alta densidade.

O algorito de percolação FoF é um dos métodos mais utilizados para se determinar estruturas no Universo (sejam elas de matéria escura ou de matéria visível). As principais vantagens do FoF são a simplicidade (possui apenas um parâmetro livre), a reprodutibilidade (para um certo tamanho de ligação produz um catálogo de grupos únicos), não assume ou aplica qualquer geometria particular para os grupos, mas identifica estruturas que são aproximadamente fechadas por uma superfície isodensa, cuja densidade está relacionada ao tamanho de ligação (HUCHRA; GELLER, 1982; BERLIND et al., 2006; CARETTA et al., 2008).

A ideia básica deste algoritmo é a seguinte: Considere uma esfera de raio R ao redor de cada partícula do conjunto total (a partícula pode ser uma estrela, uma galáxia, um grupo ou um aglomerado de galáxias). Se dentro dessa esfera existirem outras partículas, elas serão consideradas pertencentes ao mesmo grupo e serão chamadas de *amigas*. Em seguida, toma-se uma esfera ao redor de cada *amiga* e continua o procedimento usando a regra *qualquer amigo de meu amigo é meu amigo*. O procedimento pára quando nenhuma *amiga* nova pode ser adicionada ao grupo.

Em outras palavras, o algoritmo FoF agrupa partículas que são separadas por um certo tamanho de ligação l. Este tamanho, frequentemente, é dado por b vezes a separação média entre as partículas, sendo que os valores de b e l dependem da natureza da aplicação (CARETTA et al., 2008). Os grupos resultantes são limitados por uma superfície de densidade local constante de aproximadamente:

$$\frac{n}{\bar{n}} = \frac{2}{(4/3)\pi l^3}\frac{1}{n} = \frac{3}{2\pi l^3}\bar{l}^3 = \frac{3}{2\pi}\frac{1}{b^3} \approx \frac{1}{2b^3} \tag{6.6}$$

em que n é o número total de objetos e $\bar{n}$ é o número médio de objetos na região considerada.

Assumindo que o perfil de densidade destes grupos pode ser aproximado por esferas isotérmicas, a densidade média interna desta superfície é dada por aproximadamente 3 vezes a densidade de superfície. Quanto maior for l, menor é o contraste de densidade e maior é o número de partículas ligadas aos grupos. Em geral o valor de b é escolhido para dar uma sobredensidade média próxima ao valor esperado para um objeto virializado seguindo o modelo de colapso esférico. O valor estimado para identificar halos de aglomerados é $b = 0.2$ (LACEY; COLE, 1994). Nesta tese utilizamos o

valor de $l = 0.1\ Mpc$ para halos de galáxias.

Conforme Huchra e Geller (1982), os passos utilizados no algoritmo FoF para separar, por exemplo, galáxias de aglomerados de galáxias podem ser resumidos da seguinte maneira:

- Primeiro escolhe-se no catálogo uma galáxia i que ainda não faz parte de nenhum grupo.

- Procura-se ao redor dessa galáxia outra galáxia *amiga* ou galáxia j, usando as medidas de separação D_{ij} e $\mathbf{V}_{ij}$ dadas respectivamente por:

$$D_{ij} = 2sen\left(\frac{\Theta}{2}\right)\frac{\bar{v}}{H_0} \tag{6.7}$$

$$V_{ij} = |V_i - V_j| \tag{6.8}$$

  onde V_i e V_j são as velocidades radiais das galáxias i e j, m_i e m_j são suas magnitudes e Θ é sua separação angular.

- O próximo passo é verificar se as medidas D_{ij} e V_{ij} satisfazem a relação:

$$D_{ij} \leq D_{Lj} \tag{6.9}$$

$$V_{ij} \leq V_{Lj} \tag{6.10}$$

- Todas as galáxias j encontradas, ou seja, que satisfazem a relação acima, são adicionadas a lista de membros daquele grupo e o processo se repete recursivamente para cada uma delas.

- Se a galáxia i não possuir nenhuma *amiga*, ela é adicionada em uma lista de galáxias isoladas.

- O processo termina quando todas as galáxias pertencerem à algum grupo ou à lista de galáxias isoladas.

Neste caso, no fim, cada conjunto consiste ou de uma única galáxia isolada ou de um número de galáxias que tem cada uma no mínimo um vizinho cuja distância não exceda o tamanho do raio de percolação.

No caso de dados de simulação de matéria escura o procedimento é semelhante, porém, usa-se a distância euclidiana entre as partículas. Essa distância deve ser menor que o raio de percolação. No fim da aplicação do algoritmo FoF, todas as partículas recebem um rótulo (*Id Grupo*) que as identificam em seus respectivos grupos. Por exemplo, todas as partículas com $Id\ Grupo = 7$ pertencem ao mesmo grupo definido numericamente como "7". As partículas que possuem $Id\ Grupo = 0$ são partículas livres ou que permaneceram isoladas.

Após esta etapa é necessário obter algumas propriedades fundamentais dos halos para calcularmos o espectro de energia potencial gravitacional. As propriedades são: massa total, posição, velocidade média e dispersão de velocidades dos halos. A massa total M_T do halo é obtida somando-se a massa individual de cada partícula, ou seja: $M_T = \sum_{i=0}^{N} M_i$. As coordenadas são obtidas a partir da média das coordenadas das partículas do halo. Para obter a velocidade média $\overline{\mathbf{V}}$ somam-se as velocidades de todas as partículas e divide-se pelo número total de partículas do halo, conforme segue: $\overline{\mathbf{V}} = \frac{1}{N} \times \sum_{i=0}^{N} \mathbf{V}_i$. Finalmente, a dispersão de velocidades é obtida através do desvio padrão da velocidade das partículas de cada halo. O desvio padrão é uma medida do grau de dispersão dos dados em relação ao valor médio, ou seja, é a variação esperada em relação a média aritmética. Seja uma variável aleatória X que possui os valores $x_1, x_2 ... x_N$, o desvio padrão S é a raiz quadrada da variância e pode ser calculado pela Equação 6.11:

$$S = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{6.11}$$

em que $\bar{x}$ é a média da variável X.

6.2.1 Critério de *Boundedness*

Nem todos os halos identificados pelo algoritmo *Friends of Friends* estão ligados gravitacionalmente, assim, após a identificação dos mesmos e de suas propriedades, conforme descrito na seção anterior, é necessário aplicar um critério que possa garantir que apenas os halos virializados sejam selecionados. O Teorema do Virial estabelece que existe uma relação de equilíbrio entre as energias médias das partículas em um sistema quando esse sistema esta relaxado dinamicamente, de modo

que:

$$2K + W = 0 \Rightarrow 2K = -W \Rightarrow m\mathbf{V}^2 = \frac{GMm}{r} \Rightarrow \mathbf{V} \approx \sqrt{\frac{GM}{r}}$$

em que K é a energia cinética e W é a energia potencial gravitacional, $\mathbf{V}$ é a velocidade, m é a massa, r é o raio e G é a constante gravitacional. Considerando r como constante, podemos ver que a velocidade do sistema (halo) é proporcional a raiz quadrada de sua massa. A velocidade utilizada neste caso é a dispersão de velocidades, pois ela melhor representa o movimento de todas as partículas que fazem parte do halo. A Figura 6.1 mostra esta relação para os halos obtidos com o $redshift = 0$ da simulação em escala intermediária. Na curva vermelha temos um ajuste para $V \sim c_1\sqrt{M}$ e a curva verde representa este ajuste com um acréscimo de 2.5 rms. No trabalho foi considerado como virializados os halos abaixo da curva verde.

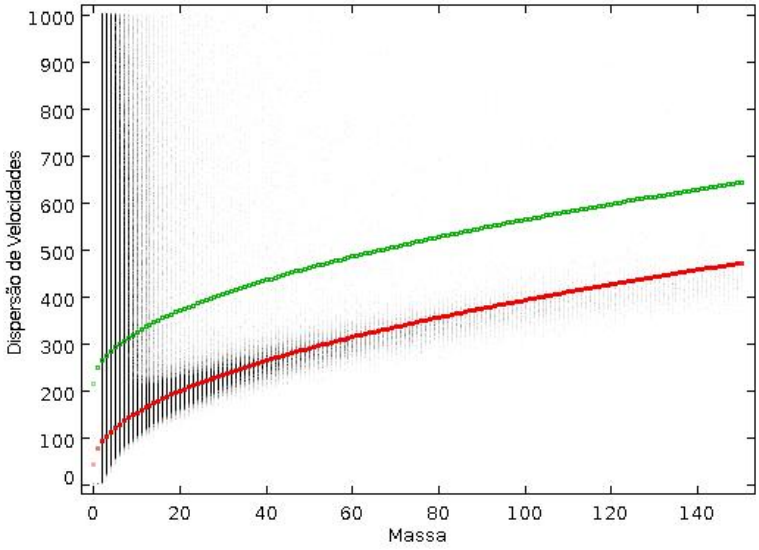

Figura 6.1 - Distribuição de massa de halos de galáxias com suas respectivas dispersão de velocidades.

6.2.2 Complexidade do Algoritmo FoF

A complexidade de um algoritmo indica a quantidade de trabalho requerida em sua execução para uma dada entrada de tamanho N. Em geral, é expressa em função do número de operações mais relevantes. As principais medidas de complexidade são: complexidade no melhor caso, complexidade no caso médio e complexidade no pior caso (TOSCANI; VELOSO, 2001). Na complexidade de pior caso considera-se o maior

número de operações realizadas, independente do tamanho da entrada, geralmente é a mais utilizada, pois representa um limite que não será ultrapassado, ou seja, garante a qualidade mínima do algoritmo.

Conforme se pode observar, no algoritmo FoF é inviável a utilização da complexidade de pior caso, uma vez que, esta ocorre quando o conjunto de dados não possui nenhum grupo, ou seja, todas as partículas estão isoladas. A alternativa mais coerente para o algoritmo FoF é a análise de complexidade de caso médio. Para realizar tal análise utilizamos uma ferramenta denominada *Vtune performance analyzer*(REINDERS, 2005). Por meio dessa ferramenta se pode analisar como varia a quantidade de execuções da operação dominante em função do tamanho da entrada. A Tabela 6.2 mostra o número de operações no algoritmo FoF em função do número de partículas (N).

Tabela 6.2 - Número de operações para o algoritmo FoF como uma função do número de partículas.

# Partículas	**# Operações**
20 000	1 966 000 000
40 000	7 771 500 032
60 000	17 493 999 616
80 000	31 125 000 192
100 000	48 421 498 880
120 000	69 568 798 720
140 000	93 066 403 840
160 000	121 513 598 976
180 000	153 710 395 392
200 000	189 676 797 952

A partir dos resultados da Tabela 6.2, podemos ver que a complexidade do algoritmo FoF é $\approx O(N^2)$. A Figura 6.2 apresenta um ajuste quadrático para os dados da Tabela 6.2. Para um exemplo prático, considerando um volume com 2 245 649 partículas, o algoritmo FoF levou 10.26 *horas*, assim, para o volume total da simulação em escala intermediária com 16 777 216 partículas (aproximadamente 7.5 vezes maior) seria gasto aproximadamente 577 *horas*. Com a versão paralela implementada nesta tese (ver próxima seção), usando 47 processadores, o tempo total gasto foi de 13.65 *minutos*.

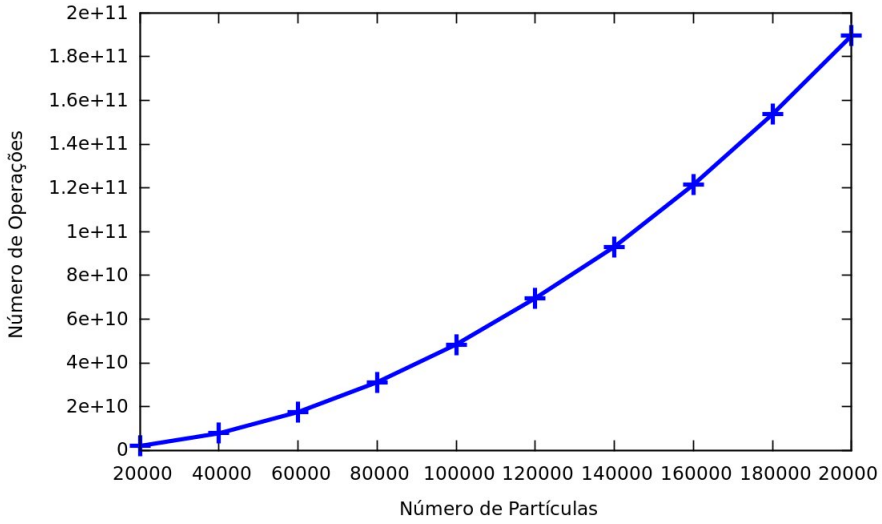

Figura 6.2 - Complexidade do Algoritmo FoF: Número de partículas x Número de operações

6.2.3 Estratégia de Paralelização para o FoF Paralelo (FoF-P)

Em trabalhos recentes (ver Fun et al. (2010) e Kwon et al. (2010)) foram desenvolvidas duas versões paralelas do algoritmo FoF, ambas utilizando o modelo de programação *MapReduce* desenvolvido pelo *google*. Uma abordagem baseada na partição do domínio é utilizada pelo consórcio *Dark Energy Universe Simulation* (DEUS). Nessa abordagem as partículas são distribuídas em subvolumes cúbicos e cada processador executa o FoF localmente. Para os grupos que contém partículas que estão localizadas próximas as bordas dos subvolumes o algoritmo verifica se estes grupos possuem partículas que estão no subvolume vizinho e as agrega ao grupo em questão. Esse código tem sido utilizado em Courtin et al. (2011) e no projeto *Halo Finder Comparison Project* (KNEBE; et al, 2011). Uma outra iniciativa é o projeto Ntropy [1]. Ntropy é uma biblioteca que permite agilizar o desenvolvimento de algoritmos paralelos que usem o algoritmo *kd-tree*. Nesse projeto foi desenvolvido uma versão paralela do algoritmo FoF que usa o algoritmo *Kd-tree* para identificar os grupos localmente e usa um procedimento baseado em grafos para conectar os grupos que são separados nas interfaces (GARDNER et al., 2007a; GARDNER et al., 2007b).

[1] http://www.phys.washington.edu/users/gardnerj/ntropy/

Nesta tese, foi implementada uma versão alternativa utilizando a biblioteca MPI e uma estratégia de pós-processamento nas interfaces. Na paralelização do algoritmo FoF a opção mais viável é a partição do domínio, ou seja, cada processo executa o FoF em uma determinada parte dos dados. Porém, como os dados são fortemente dependentes, a divisão aleatória do domínio entre os diversos processos pode ocasionar a separação de partículas que estão ligadas gravitacionalmente. Para resolver este problema, foi implementado uma estratégia de pós processamento que identifica nas interfaces da divisão do domínio quais partículas estão ligadas e que foram enviadas á processadores diferentes. Um novo cálculo é realizado nesse subdomínio e as partículas são associadas aos seus respectivos grupos, conforme Figura 6.3.

O algoritmo FoF foi paralelizado usando a biblioteca MPI da seguinte maneira: O processador mestre lê os dados de entrada, que basicamente é constituído pelo índice de cada partícula e suas posições e velocidades. O número de partículas é então dividido em tamanhos iguais e enviado para os processadores escravos. Cada processador calcula e envia os grupos parciais ao processador mestre, que também é usado para o cálculo de um determinado conjunto de partículas. Depois de receber todos o grupos o processador mestre calcula o pós-processamento e escreve o arquivo de saída com todos os grupos obtidos. Um esquema da implementação desta paralelização pode ser vista na Figura 6.4. Os códigos computacionais desenvolvidos foram implementados usando a linguagem de programação C.

6.3 Cálculo do Espectro de Energia Potencial Gravitacional

Uma das principais características da turbulência é a existência de uma cascata de energia cinética. Em um escoamento turbulento plenamente desenvolvido em fluidos, o comportamento do espectro de energia turbulenta para o subdomínio inercial (região do espectro que está afastada da região do ingresso de energia no sistema associada a baixos números de onda, bem como da região de dissipação viscosa associada aos mais altos números de onda) é expresso pela lei de potência de Komolgorov dada pela Equação 3.14.

Entretanto, é difícil obter um espectro de energia cinética para o caso de dados cosmológicos. Conforme Caretta et al. (2008), a energia que pode produzir um comportamento similar ao turbulento na evolução dinâmica do Universo é a energia potencial gravitacional. Uma vez identificados os halos de estruturas, usando o algoritmo FoF-P, o espectro de energia potencial gravitacional pode ser obtido estimando

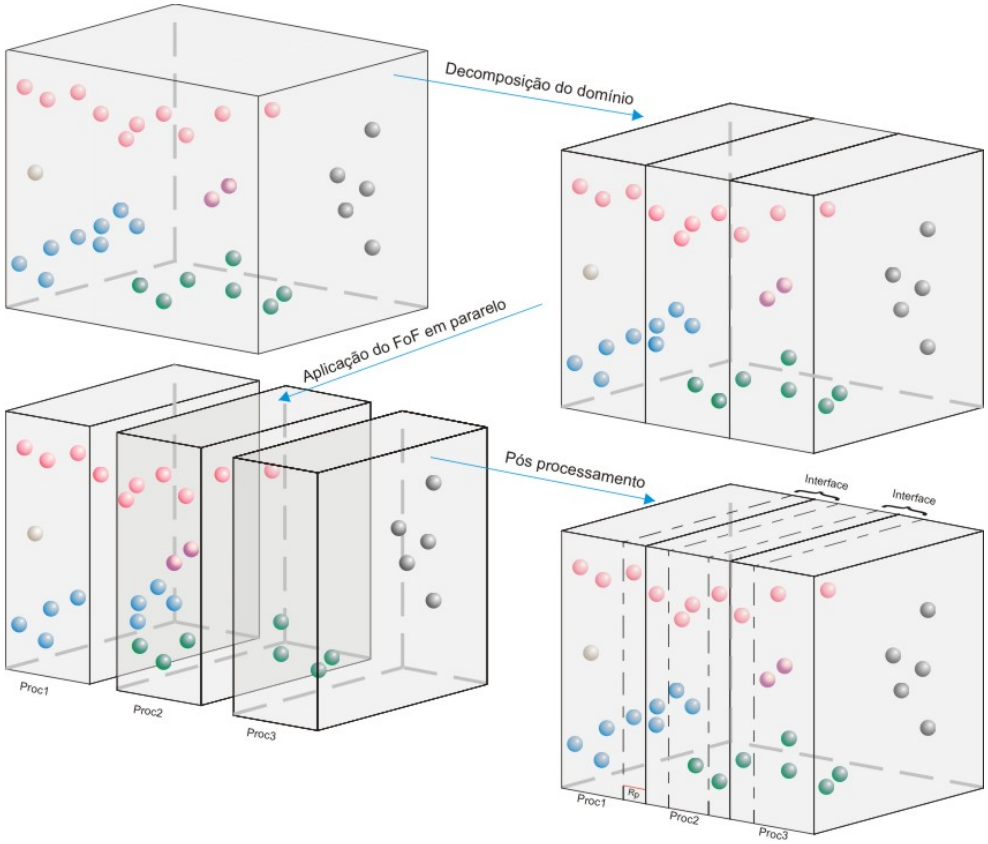

Figura 6.3 - Estratégia usada no FoF Paralelo: Decomposição do Domínio e Pós-Processamento nas interfaces dos subdomínios.

para cada halo, em esferas concêntricas, a energia potencial gravitacional (em módulo) devido aos outros halos, de acordo com a Equação 6.12.

$$U_j = \frac{1}{2} G m_j \sum \frac{m_i}{r_{ij}} \tag{6.12}$$

em que G é a constante gravitacional, r_{ij} é a distância entre os halos i e j, m_i e m_j são as massas dos halos i e j respectivamente. Para cada halo no domínio, varia-se o raio da esfera e estima-se o valor médio da energia em cada uma delas. O espectro é obtido com o valor cumulativo. Na Figura 6.5 temos uma ilustração de como é feito esta análise. Note que é preciso excluir os halos que estão próximos as bordas, de modo que o raio da esfera não ultrapasse os limites do volume. Confome se verifica na Figura 6.6 só serão considerados os halos que estão no volume interno.

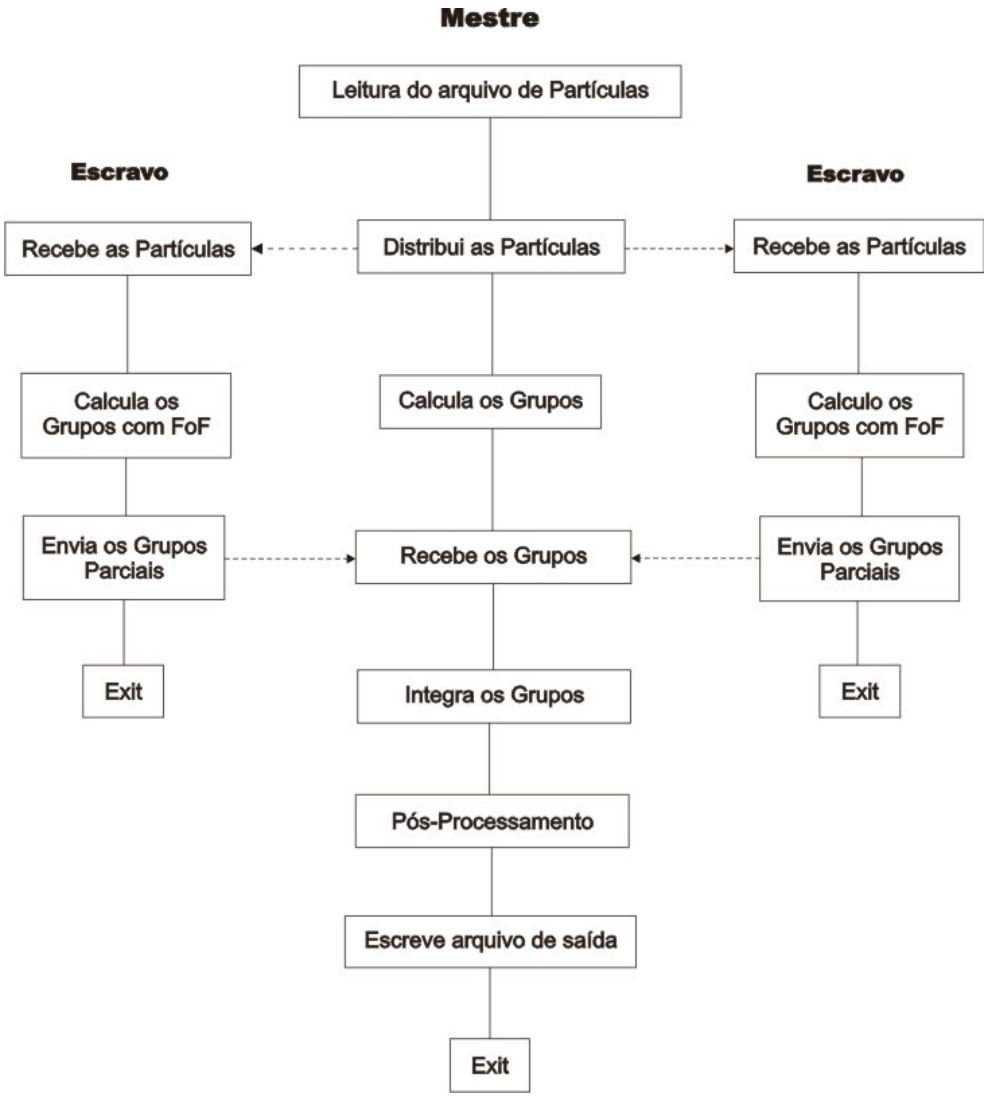

Figura 6.4 - Esquema da implementação parallela do FoF.

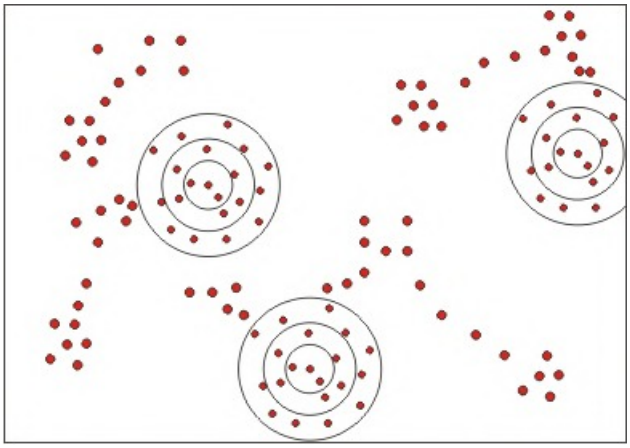

Figura 6.5 - Esquema de obtenção dos bins para o cálculo do espectro de energia.

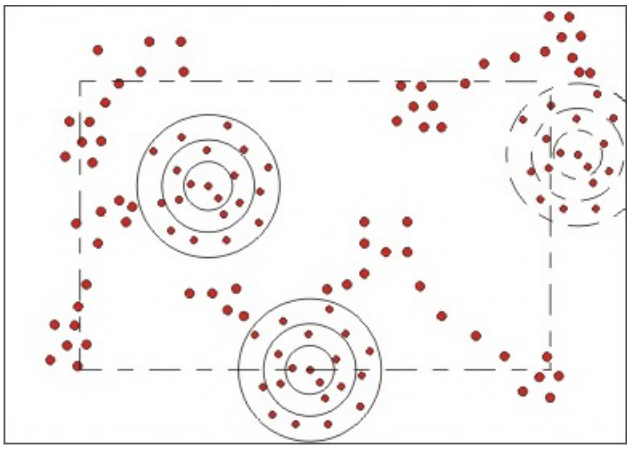

Figura 6.6 - Volume utilizado no cálculo do espectro de energia. A região pontilhada representa a região utilizada, excluindo-se os halos que estão na borda.

7 RESULTADOS OBTIDOS

Conforme já mencionado na Introdução, a proposta desta tese de doutorado está vinculada ao aprofundamento do estudo da caracterização da dinâmica cosmológica como sendo uma evolução de um sistema turbulento. A avaliação está voltada à análise de escala de uma lei do tipo de Kolmogorov - Equação 3.14. Outras características que são objetos de discussão na teoria moderna da turbulência, como a questão da intermitência são de grande relevância, mas com o compromisso de um período de tempo fixo de um trabalho de doutorado, esta será uma atividade de pesquisa que irá se efetuar em outra ocasião.

A verificação da possibilidade de existência de um processo similar a dinâmica turbulenta na formação das grandes estruturas no Universo pode ser feita por meio da análise do espectro de energia potencial gravitacional em diferentes *redshifts*, usando dados de simulação de matéria escura e, principalmente, dados observacionais disponibilizados em grandes bases de dados.

Neste sentido, trabalhamos com dados de simulações de N-corpos de matéria escura de dois projetos do Consorcio Virgo, a saber: Simulações de Escala Intermediária (JENKINS et al., 1998) e Simulação *Millennium*(SPRINGEL et al., 2005), conforme descrito na Seção 5.2. Trabalhamos também com 3 subamostras de dados observacionais de galáxias obtidas a partir do catálogo espectroscópico do projeto SDSS, conforme descrito na Seção 5.3. A análise de dados observacionais foi a etapa mais complexa do trabalho, além do maior desafio científico associado a incerteza na coordenada de profundidade dos objetos astronômicos - nem todas as galáxias possuem uma medida de *redshift*, os dados observacionais são mais limitados que os dados simulados por serem incompletos, não há observações em todas as direções do céu e parâmetros importantes são obtidos de forma indireta.

Como dito anteriormente, todos os códigos implementados/utilizados por Caretta et al. (2008) foram paralelizados. O código utilizado na obtenção dos espectros de energia para halos de matéria escura foi adaptado para uso com dados de observação, uma vez que o volume, neste caso, varia com o *redshift*. Na sequência apresentamos o desempenho das implementações paralelas, o desempenho da grade e uma análise dos espectros de energia potencial gravitacional obtidos. O sistema utilizado para avaliação de desempenho do algoritmo FoF paralelo e do algoritmo para cálculo do espectro de energia potencial é composto por uma máquina Cray XT6, que possui

13 nós de acesso interativo e 20 nós de processamento auxiliar (cada nó com 4 processadores AMD - Opteron quad core de 2.7 GHz e Redes Ethernet de 1 e 10 Gigabit.), 1272 nós computacionais Cray XT6 (cada nó com 2 processadores AMD Opteron 12-core de 2.1 GHz 1 módulo de interconexão Cray Seastar2+ por nó) e 32 nós de serviço (cada nó com 2 processadores AMD Opterom dual-core de 2.6GHz e 1 módulo de interconexão Cray Seastar2+ por lâmina)

7.1 Análise de Desempenho do FoF-P

Geralmente as duas medidas mais utilizadas na análise de desempenho de um programa paralelo são o *speed-up* e a eficiência (PACHECO, 1997). Basicamente, o *speed-up* é a razão entre o tempo de execução gasto no problema sequêncial e o tempo de execução gasto no programa paralelo. Se $T_\sigma(N)$ representa o tempo de execução sequencial e $T_\pi(N, P)$ o tempo de execução paralela com P processadores, o *speed-up* do algoritmo paralelo é definido conforme Equação 7.1

$$S(N, P) = \frac{T_\sigma(N)}{T_\pi(N, P)} \tag{7.1}$$

O *speed-up* é linear (ou ideal) quando $S(N, P) = P$. Em alguns casos, uma implementação paralela pode apresentar um desempenho acima do ideal ($S(N, P) > P$), ou seja, apresentar um *speed-up* maior que o linear, chamado *speed-up* super-linear.

De acordo com Pacheco (1997), a eficiência é uma medida da utilização de processadores em um programa paralelo, em relação ao programa serial. Ela é definida conforme Equação 7.2.

$$E(N, P) = \frac{S(N, P)}{P} = \frac{T_\sigma(N)}{PT_\pi(N, P)} \tag{7.2}$$

Para avaliar o desempenho de nossa implementação paralela foi realizado uma aplicação desta versão sobre uma região cúbica do projeto Virgo com 120 h^{-1} Mpc de lado e 2 245 649 partículas no *redshift* $z = 0$. Na Tabela 7.1 tem-se os tempos obtidos na execução paralela do algoritmo FoF variando-se o número de processadores e na Figura 7.1 um esquema gráfico desses tempos. A curva em vermelho representa o tempo gasto pelo FoF-P para realizar o agrupamento e mostra um desempenho muito acima do desempenho ideal. Já a curva em azul representa o tempo gasto pelo FoF-P mais o tempo gasto no pós processamento. Podemos ver que este tempo tem

um limite máximo de desempenho em aproximadamente 50 processadores, quando começa a aumentar, ficando em função do tempo gasto no pós processamento, representado no gráfico pela curva verde. A justificativa para este comportamento é que o pós processamento é realizado de maneira sequêncial, e pela lei de Amdahl o desempenho de uma aplicação em paralelo está limitada a porcentagem de execução serial do algoritmo. No nosso caso, isto se agrava porque o tempo gasto no pós processamento aumenta com o aumento de processadores. No entanto, quando comparamos o tempo gasto no FoF-P mais o tempo gasto no pós processamento com o tempo ideal (tempo esperado na paralelização) o desempenho do algoritmo FoF-P se torna inferior ao ideal somente a partir de 280 processadores.

Na Tabela 7.2 pode-se observar as medidas de *speed-up* para o FoF-P e *speed-up* e eficiência para o FoF-P + pós.

Tabela 7.1 - Tempo gasto na execução paralela do algoritmo FoF variando-se o número de processadores.

# Processadores	**Tempo FoF-P**	**Tempo Pós**	**Tempo FoF-P + Pós**
1	36944	0	36944
4	4169.99	9.06	4179.05
6	1529.52	9.92	1539.44
12	328.79	12.20	340.99
24	88.27	18.96	107.23
48	28.51	30.24	58.75
96	12.05	53.70	65.75
144	9.00	76.37	85.37
192	7.77	99.04	106.81
240	7.27	119.26	126.53
288	6.93	141.22	148.15
336	6.81	161.82	168.63
384	6.63	182.65	189.28

A Figura 7.2 exibe uma comparação entre o *speed-up* ideal e o *speed-up* obtido com nossa versão paralela. Conforme se pode observar o *speed-up* obtido, sem considerar o tempo de pós, apresenta um comportamento acima do ideal, ou seja, *speed-up* super linear. Quando consideramos também o tempo gasto no pós processamento, o *speedup* deixa de ser super-linear a partir de aproximadamente 287 processadores.

Tabela 7.2 - *Speed-up* e eficiência para a versão paralela do algoritmo FoF.

# Processadores	Speed-up FoF-P	Speed-up FoF-P + Pós	Eficiência
1	1.00	1.00	1.00
4	8.86	8.84	2.21
6	24.15	24.00	4.00
12	112.36	108.34	9.03
24	418.53	344.53	14.36
48	1295.82	628.83	13.10
96	3065.89	561.88	5.85
144	4104.89	432.75	3.00
192	4754.7	345.89	1.80
240	5081.71	291.98	1.22
288	5331	248.37	0.86
336	5424.96	219.08	0.65
384	5572.25	195.18	0.51

Em desenvolvimentos futuros vamos paralelizar o pós processamento para melhorar este escalonamento.

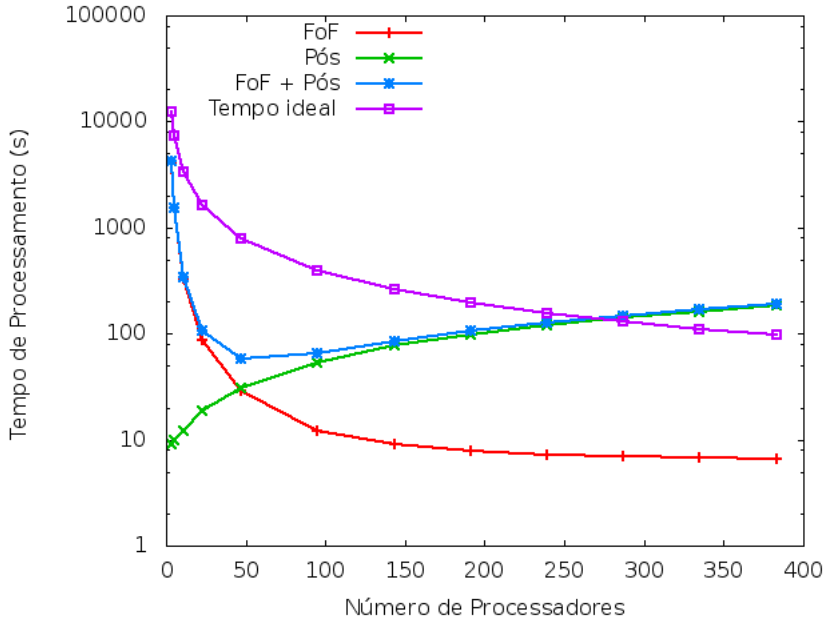

Figura 7.1 - Análise do tempo gasto, em função do número de processadores, para encontrar os halos de matéria escura de uma amostra do Virgo com 2 245 649 partículas, utilizando nossa versão paralela do FoF.

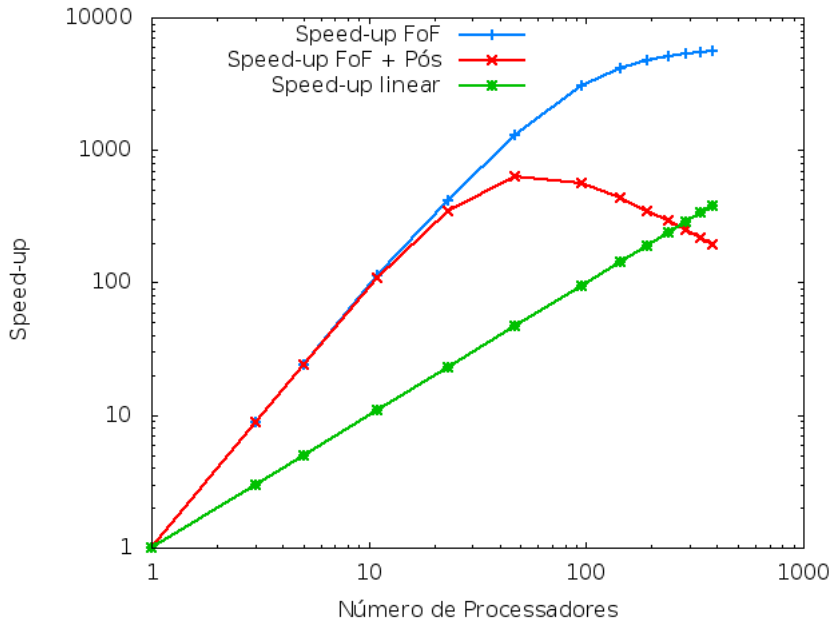

Figura 7.2 - Comparação entre o *speed-up* do FoF-P e o *speed-up* linear para a amostra de 2 245 649 partículas.

De acordo com Foster (1995), situações que podem favorecer um desempenho super linear são as seguintes:

- Efeitos de cache: Em geral, em máquinas paralelas, cada processador tem uma quantidade pequena de memória cache, e então só é possível armazenar nessa memória uma pequena quantidade de dados. Porém, quando um problema é resolvido usando um grande número de processadores, mais dados podem ser armazenados na memória cache e como o acesso aos dados que estão na memória cache é muito mais rápido do que o acesso aos dados que estão na memória principal, o tempo computacional total tende a diminuir. Deste modo, numa implementação paralela, se o ganho no tempo de execução devido aos efeitos de memória cache compensar o aumento no tempo de comunicação devido a utilização de processadores adicionais, isso pode levar a uma eficiência maior do que 1 e a um *speed-up* super linear.

- Anomalias em buscas: Se a árvore de busca contém soluções variando em níveis de profundidade, então, várias buscas no primeiro nível, mas em ramos diferentes, irão explorar menos nós da árvore antes de encontrar uma solução do que exploraria uma busca sequencial. Observe que, nesse

caso, o algoritmo paralelo executado é diferente do algoritmo sequencial.

Na nossa versão paralela do algoritmo FoF só existe comunicação entre o nó mestre e os processadores escravos, além disso, a divisão das partículas entre os diversos processadores reduziu o espaço de busca e isso também pode ter favorecido o desempenho super linear.

7.2 Análise de Desempenho do Algoritmo Paralelo que Calcula Energia Potencial Gravitacional

Também foi implementado, usando MPI, uma versão paralela do algoritmo que calcula o espectro de energia potencial gravitacional para os halos de matéria escura identificados pelo algoritmo FoF-P. A estratégia de paralelização utilizada foi distribuir os halos para os diferentes processadores e então obter a energia para cada esfera. Após o cálculo da energia, os processadores escravos enviam o resultado para o processador mestre, que calcula a energia média e cumulativa para todas as esferas. O *speed-up* obtido, em um teste com um conjunto de 905 141 halos de matéria escura é apresentado na Figura 7.3. Podemos verificar que o *speed-up* obtido é quase linear.

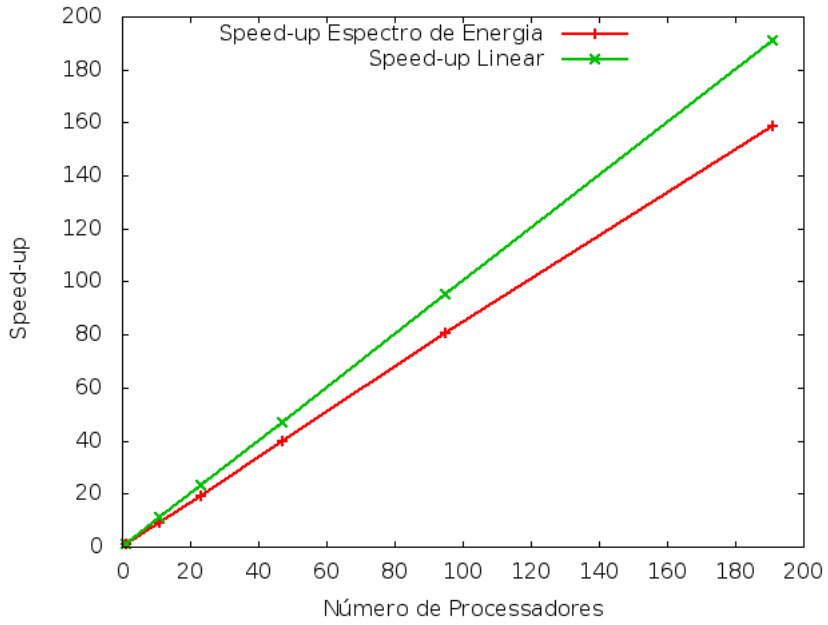

Figura 7.3 - Comparação entre o *speed-up* da versão paralela do algoritmo para cálculo do espectro de energia, para um conjunto de 905 141 halos de matéria escura, e o *speed-up* linear,

7.3 Análise de Desempenho da Grade

Outro objetivo da tese é a implementação de uma infra-estrutura de grade computacional de astrofísica e o desenvolvimento de ferramentas de análise de dados astronômicos a serem incorporadas na mesma. Nesse sentido foi desenvolvido uma versão paralela do algoritmo de percolação *Friends of Friends* e do algoritmo que calcula o espectro de energia potencial gravitacional, bem como, *scripts* em linguagem *shell script* para utilização dos mesmos na grade.

O ambiente de grade implementado nesta tese consistiu de 3 nós para processamento sendo composto de: Um *cluster* CRAY/XD1, com 12 processadores Opteron 2.6 GHz do LAC/INPE, um *cluster* HP XC do Centro de Processamento de Alto Desempenho (C-PAD) do INPE de São José dos Campos, com uma arquitetura escalar AMD-Opteron 2.2 GHz e interconexão de alta velocidade InfiniBand com uma largura de banda de 2.5 Gbps, contendo um total de 112 CPUs, mas que por se tratar de uma máquina Multi-usuários, para este experimento apenas 15 processadores foram disponibilizados. O terceiro nó, localizado no Departamento de Eletrônica e Computação da Universidade Federal de Santa Maria (UFSM) possui dois processadores Intel Xeon 2.0 GHz *quad-core*, com um total de 8 núcleos de processamento. Todas as máquinas possuem o sistema operacional linux e a biblioteca MPI usada para a execução em paralelo. A conexão aos *clusters* do INPE foi feita através de uma máquina Intel Xeon 2.0 GHz da Divisão de Astrofísica - DAS/INPE. Pode se observar que esta é uma grade heterogênea com diferentes arquiteturas e capacidade de processamento.

Um quarto nó de processamento localizado no departamento de Astronomia da Universidade de Guanajuato (México) composto por uma servidora com dois processadores Intel Xeon 2.8 GHz *six-core* será incorporado futuramente à grade.

O experimento realizado na grade para avaliação de seu desempenho consistiu na identificação dos halos de matéria escura para os dados de simulação em nove diferentes *redshifts*, utilizando a versão paralela do algoritmo de percolação *Friends of Friends* (RUIZ et al., 2009; RUIZ et al., 2011). A obtenção dos halos para diferentes *redshifts* são tarefas independentes, sendo então, totalmente adequadas para aplicações em grade. Deste modo, os 9 *redshifts* analisados foram distribuídos nos 3 *clusters* pertencentes a grade.

Definindo T_S como sendo o tempo total correspondente a execução sequencial do conjunto de tarefas e T_G o tempo total gasto na execução de todos os *jobs* simultaneamente na grade, o *speed-up* (S) é definido como sendo a razão entre T_S e T_G. Conforme se verifica na Tabela 7.3, considerando os 3 *clusters* utilizados, $T_S = 40.43\ h$ e $T_G = 14.53\ h$, logo:

$$S = \frac{T_S}{T_G} = \frac{40.43}{14.53} = 2.78. \tag{7.3}$$

Tabela 7.3 - Tempo total gasto por cada *cluster* da grade.

Cluster	**Jobs**	T_S **(hh:mm)**	$T_S/Jobs$ **(hh:mm)**
C-PAD/INPE	5	14:32	02:54
LAC/INPE	3	13:57	04:39
UFSM	1	11:57	11:57
Total	9	40:26	

Outra maneira de medir o desempenho da grade pode ser obtida se consideramos o melhor e o pior caso, em que se avalia o tempo médio gasto na execução de cada tarefa no *cluster* mais rápido e no *cluster* mais lento. Conforme Tabela 7.3, o tempo médio gasto para cada tarefa no *cluster* mais rápido foi de 2.9 horas, então para as 9 tarefas teremos $T_S = 26.1\ h$, portanto:

$$S = \frac{T_S}{T_G} = \frac{26.1}{14.53} = 1.80 \tag{7.4}$$

Já para o pior caso, o tempo médio gasto por cada tarefa foi de 11.95 horas, logo para as 9 tarefas teremos:

$$S = \frac{T_S}{T_G} = \frac{107.55}{14.53} = 7.40 \tag{7.5}$$

As medidas de tempo apresentadas na Tabela 7.3 foram extraídas do arquivo de *log* do *OurGrid* que registra o inicio e o fim de cada job.

Através da análise do tempo gasto na execução das tarefas podemos ver que realizar a tarefa em cada instituição separadamente demandará um tempo de 26.1 h no melhor caso (*cluster* mais rápido) e 107.55 h no pior caso (*cluster* mais lento). Por meio do uso compartilhado de recursos computacionais entre as instituições, utilizando a tecnologia da grade computacional, o tempo gasto foi de 14.53 h, uma redução

significativa no tempo total de execução. Outra grande vantagem no uso da grade é que se durante a execução das tarefas, algum nó da grade deixa de funcionar devido a uma falta de energia ou uma falha na rede, aquela tarefa automaticamente é enviada para outro nó que esteja ocioso, garantindo assim que o processamento não seja interrompido.

7.4 Análise do Espectro de Energia Potencial Gravitacional para 3 amostras

Nesta Seção apresentamos os espectros de energia obtidos para todos os conjuntos de dados, em diferentes *redshifts*. Vale lembrar que as medidas de *redshift* podem ser associadas a determinadas épocas da evolução cosmológica, assim, a análise feita com diversos *redshifts* visa mostrar o comportamento do espectro em diferentes escalas de tempo. Na Figura 7.4 temos o espectro obtido para os dados da simulação em escala intermediária. Na abscissa está representado o número de onda, k, que é o inverso da escala espacial (neste caso dada em Mpc), e na ordenada a energia potencial gravitacional (Equação 6.7) calculada. Os diferentes conjuntos de pontos representam os resultados obtidos para os diferentes "instantâneos" da simulação (diferentes *redshifts*). Nota-se claramente que os diferentes *redshifts* apresentam um comportamento bastante similar, com praticamente apenas uma diferença de ponto zero, ou seja, quanto mais recentemente, maior a energia potencial em qualquer escala. Esse resultado é o esperado porque as estruturas vão se tornando cada vez mais ligadas gravitacionalmente (ou, mais precisamente, cada vez mais as estruturas se colapsam e evoluem para um estado ligado). Além disso, se pode perceber também que o crescimento da energia potencial tem uma tendência a diminuir com o tempo, mais uma vez algo esperado porque, com a expansão acelerada do Universo, os objetos se afastam cada vez mais rapidamente e menos objetos entram em um regime de ligação e colapso gravitacional. Em relação à forma desses espectros, vemos que é aproximadamente "linear" no gráfico logarítmico (ou, mais precisamente, segue uma "lei de potência") ao menos em parte do intervalo de escalas considerado. Como verificado anteriormente por Caretta et al. (2008), há um intervalo ($0.01 < k < 0.07$, ou 14 a 100 $h^{-1}\ Mpc$) em que a inclinação do espectro se aproxima da lei de $-5/3$ de Kolmogorov, evidenciando uma característica típica de processos turbulentos.

Na Figura 7.5, um comportamento similar é apresentado pelos espectros obtidos para os dados da simulação *Millennium*, inclusive para um intervalo parecido. Considerando que esta simulação é completamente independente (a menos dos parâmetros

cosmológicos e das suposições fundamentais que são praticamente as mesmas), e que tem 80 vezes mais resolução em massa, esses resultados podem ser tomados como uma confirmação dos resultados anteriores.

A prova mais robusta, entretando, deve ser a dada pelos resultados com os espectros obtidos dos dados observacionais (Figura 7.6). Aqui não há suposições fortes como no caso dos dados simulados, a exceção de algumas considerações sobre completeza e acurácia das determinações de massa e outros parâmetros. Interessantemente, os espectros obtidos para as 3 amostras revelam novamente uma similaridade em relação à forma e, outra vez, podemos identificar um intervalo onde os espectros se aproximam de $-5/3$ (neste caso em escalas um pouco mais restritas, de 30 a 70 $h^{-1}\ Mpc$).

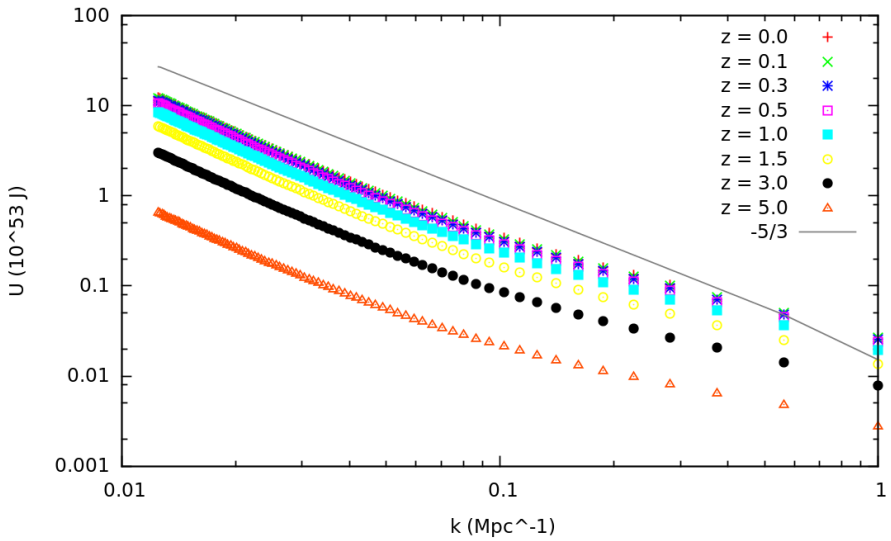

Figura 7.4 - Espectro de energia potencial gravitacional para halos de matéria escura considerando um cubo de aresta $L = 239\ h^{-1}Mpc$

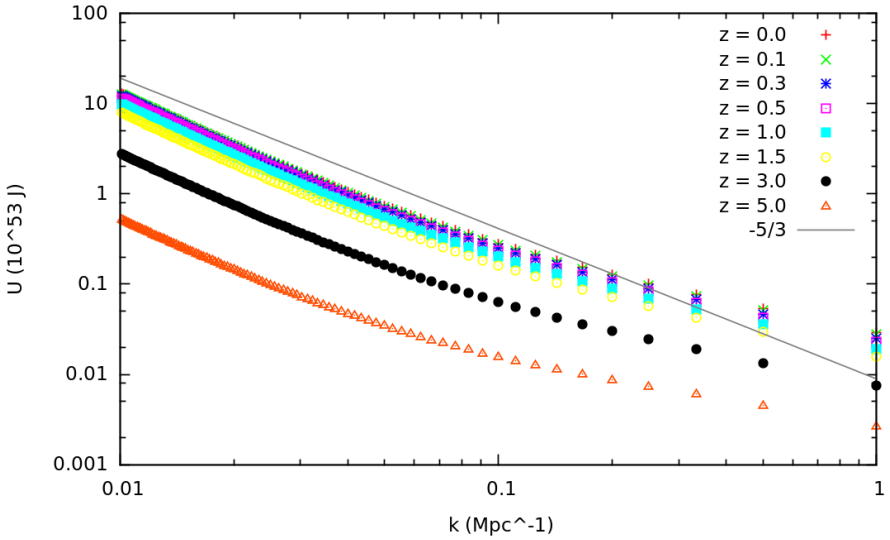

Figura 7.5 - Espectro de energia potencial gravitacional para halos de matéria escura de galáxias considerando um cubo de aresta $L = 500\ h^{-1}Mpc$

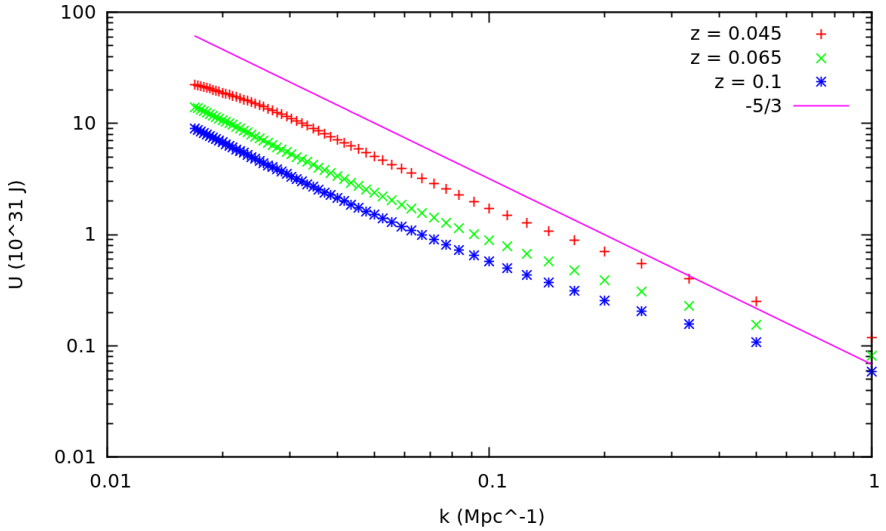

Figura 7.6 - Espectro de energia potencial gravitacional para dados de observação (galáxias), provenientes do projeto SDSS

8 CONCLUSÕES E CONSIDERAÇÕES FINAIS

Neste trabalho, é apresentado um aprofundamento no estudo da caracterização da formação de estruturas no Universo como um processo que apresenta similaridades com os processos turbulentos em fluidos. A análise mostrou que tanto para dados de simulação de matéria escura, quanto para dados de observação, o espectro de energia potencial gravitacional segue uma lei de potência do tipo $-5/3$ em uma parte do intervalo de escalas considerado.

Para chegar a esses resultados utilizamos computação de alto desempenho e a tecnologia de computação em grade. O experimento realizado mostrou a viabilidade na implementação de uma grade computacional para a Astronomia brasileira. Esta tese apresenta a primeira aplicação de cosmologia usando recursos de uma grade no contexto do projeto BRAVO @INPE.

As principais contribuições desta tese são:

- Implementação de uma nova versão paralela do algoritmo *Friends of Friends*. O FoF-P é uma das ferramentas de análise de dados (identificação de grupos em função do raio percolação) que será disponibilizada para a comunidade astronômica no portal do BRAVO @INPE.

- Implementação paralela do algoritmo que calcula o espectro de energia potencial gravitacional.

- Desenvolvimento do primeiro ambiente de grade para a astronomia do Brasil com 2 aplicações: identificação de halos de matéria escura com o FoF-P e cálculo do espectro de energia potencial gravitacional.

- Aplicação do algoritmo de construção de árvores de decisão J4.8 no desenvolvimento de classificadores baseados em atributos fotométricos para separar objetos astronômicos em estrelas e galáxias.

- Obtenção do espectro de energia potencial gravitacional para os halos de galáxias num cubo de dados de 239 h^{-1} Mpc de lado da Simulação de Escala Intermediária do consórcio Virgo, usando o volume total da simulação em escala intermediária de matéria escura (ampliando assim a análise de Caretta et al. (2008)) e confirmando seus resultados.

- Obtenção do espectro de energia potencial gravitacional para os halos de galáxias num cubo de dados de 500 $h^{-1}\ Mpc$ de lado da Simulação *Millennium*, que possui aproximadamente 80 vezes mais resolução que a anterior.
- Obtenção do espectro de energia potencial gravitacional para as galáxias de 3 amostras de diferentes volumes (limitadas a *redshifts* 0.045, 0.068 e 0.1) do levantamento SDSS, confirmando assim que os resultados obtidos para as simulações são reproduzidos pelos dados observacionais reais.

Como trabalhos futuros propomos os seguintes desenvolvimentos:

- Paralelização do pós processamento para melhorar o escalonamento do FoF-P.
- Implementação paralela da versão $nlog(n)$ do algoritmo FoF.
- Expandir a análise de $-5/3$ para aglomerados e superaglomerados dos dados da simulação *Millennium*.
- Formulação das equações de Navier-Stokes para o fluido cósmico, considerando a difusividade turbulenta.
- Busca da identificação do expoente de intermitência, que seria o fator responsável pela dinâmica da agregação das estruturas cósmicas.

REFERÊNCIAS BIBLIOGRÁFICAS

AARSETH, S. J. **Gravitational N-body simulations**: tools and algorithms. Cambridge, UK: Cambridge University Press, 2003. 413 p. 52

ABAZAJIAN, K. N.; ADELMAN-McCARTHY, J. K.; et al. The seventh data release of the sloan digital sky survey. **The Astrophysical Journal Supplement Series**, v. 182, p. 543 – 558, 2009. 5, 58

ALMEIDA, E. S. **Climatologia de mesoescala em grade computacional**. 152 p. Tese (Doutorado) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2007. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m17@80/2007/04.25.17.03>. Acesso em: 03 maio 2011. 46

ANDRADE, N.; CIRNE, W.; BRASILEIRO, F.; ROISENBERG, P. Ourgrid: An approach to easily assemble grids with equitable resource sharing. In: WORKSHOP ON JOB SCHEDULING STRATEGIES FOR PARALLEL PROCESSING, 9., 2003, Seattle, EUA. **Proceedings...** Seattle: Springer-Verlag LNCS, 2003. p. 61–86. ISBN 3-540-20405-9. 46, 47

ASADZADEH, P.; BUYYA, R.; KEI, C. L.; NAYAR, D.; VENUGOPAL, S. Global grids and software toolkits: A study of four grid middleware technologies. **High Performance Computing: Paradigm and Infrastructure**, p. 431 – 458, 2005. 44, 46

ASSIS, A. K. T. On hubble's law of redshifts, olbers' paradox and the cosmic background radiation. **Apeiron**, v. 12, p. 13 – 29, 1992. 12

ASSIS, A. K. T.; NEVES, M. C. D. The redshift revisited. **Astrophysics and Space Science**, v. 227, p. 13 – 24, 1995. 12

BAHCALL, N. A.; LUBIN, L. M.; DORMAN, V. Where is the dark matter. **The Astrophysical Journal**, v. 447, p. L81 – L85, 1995. 54

BARNES, J.; HUT, P. A hierarchical o(nlogn) force-calculation algorithm. **Nature**, v. 324, p. 446 – 449, 1986. 53

BERLIND, A. A.; FRIEMAN, J. A.; et al. Percolation galaxy groups and clusters in the sdss redshift survey:identification, catalogs and the multiplicity function. **The Astrophysical Journal Supplement Serie**, v. 167, p. 1 – 25, 2006. 72

BERTSCHINGER, E. Simulations of structure formation in the universe. **Annu. Rev. Astron. Astrophys**, v. 36, p. 599 – 654, 1998. 52

BISHOP, C. M. **Pattern recognition and machine learning**. New York: Springer, 2006. 738 p. 66

BRAZILIAN Virtual Observatory (BRAVO). 2008. Acesso em: 10 jan. 2011. Disponível em: <http://www.lac.inpe.br/bravo/>. 7

CAMPOS VELHO, H. F. **Modelagem matemática em turbulência atmosférica**. São Carlos: SBMAC, 2010. 87 p. (Notas em Matemática Aplicada; v. 48). 34

CAMPOS VELHO, H. F.; RAMOS, F. M.; ROSA, R. R.; et al. Multifractal model for eddy diffusivity and counter gradient term in atmospheric turbulence. **Physica A: Statistical Mechanics and its Applications**, v. 295, p. 219–223, 2001. 35

CAPIT, N.; COSTA, G.; GEORGIOU, Y.; HUARD, G.; MARTIN, C.; MOUNIE, G.; NEYRON, P.; RICHARD, O. A batch scheduler with high level components. In: CLUSTERS COMPUTING AND THE GRID, 5., 2005, Cardiff, UK. **Clusters computing and the grid - CCGrid 2005**. Cardiff, UK: IEEE Computer Society, 2005. p. 776–783. ISBN 0-7803-9074-1. 46

CARETTA, C. A.; ROSA, R. R.; CAMPOS VELHO, H. F. de; RAMOS, F.; MAKLER, M. Evidence of turbulence-like universality on the formation of galaxy-sized dark matter haloes. **Astronomy & Astrophysics**, v. 487, p. 445 – 451, 2008. 4, 72, 78, 83, 91, 95

CARVALHO, R. R.; GAL, R. R.; CAMPOS VELHO, H. F.; CAPELATO, H. V.; LA BARBERA, F.; VASCONCELLOS, E. C.; RUIZ, R. S. R.; KOHL MOREIRA, J. L.; LOPES, P. A.; SOARES-SANTOS, M. The brasilian virtual observatory - a new paradigm for astronomy. **Journal of Computational Interdisciplinary Sciences**, v. 1, p. 1 –1 20, 2009. 5, 6

CASSANO, R.; BRUNETTI, G. Cluster mergers non-thermal phenomena: a statistical magneto-turbulent model. **Monthly Notices of the Royal Astronomical Society**, v. 357, p. 1313 – 1329, 2005. 3

CATELAN, P. Lagrangian dynamics in non-flat universes and non-linear gravitational evolution. **Monthly Notices of the Royal Astronomical Society**, v. 276, p. 115–124, 1995. 2

CHABRIER, G. **Structure formation in astrophysics**. New York: Cambridge University Press, 2009. 454 p. 21

CHAPMAN, B.; JOST, G.; VAN DER PAS, R. **Using OpenMP**: portable shared memory parallel programming. Cambridge, MA: The MIT Press, 2007. 377 p. 41

CID FERNANDES, R.; MATEUS, A.; SODRé, L.; STASINSKA, G.; GOMES, J. M. Semi-empirical analysis of sloan digital sky survey galaxies - I. Spectral synthesis method. **Monthly Notices of the Royal Astronomical Society**, v. 358, p. 363–378, 2005. 59

CIRNE, W.; PARANHOS, D.; COSTA, L.; SANTOS-NETO, E.; BRASILEIRO, F.; SAUVé, J.; SILVA, F. A. B. da; BARROS, C. O.; SILVEIRA., C. Running bag-of-tasks applications on computational grids: The mygrid approach. In: INTERNATIONAL CONFERENCE ON PARALLEL PROCESSING, 2003. **Proceedings...** Kaohsiung: IEEE, 2003. p. 407–416. ISBN 0-7695-2017-0. 47

CIRNE, W.; SANTOS NETO, E. Grids computacionais: Da computação de alto desempenho a serviços sob demanda. In: SIMPóSIO BRASILEIRO DE REDES DE COMPUTADORES (SBRC), 23., 2005, Fortaleza, Ceará. **Anais...** Fortaleza: SBC, 2005. ISBN 85-7669-022-5. 47

COLES, P.; MELOTT, A. L.; SHANDARIN, S. F. Testing approximations for non-linear gravitational clustering. **Monthly Notices of the Royal Astronomical Society**, v. 260, p. 765–776, 1993. 2

COLLESS, M.; DALTON, G. B.; et al. The 2df galaxy redshift survey: spectra and redshifts. **Monthly Notices of the Royal Astronomical Society**, v. 328, p. 1039–1063, 2001. 5

COSTA-DUARTE, M. V.; SODRÉ Jr., L.; DURRET, F. Morphological properties of superclusters of galaxies. **Monthly Notices of the Royal Astronomical Society**, v. 000, p. 1–11, 2011. 62

COURTIN, J.; RASERA, Y.; ALIMI, J. M.; CORASANITI, P. S.; BOUCHER, V.; FUZFA, A. Imprints of dark energy on cosmic structure formation - ii. non-universality of the halo mass function. **Monthly Notices of the Royal Astronomical Society**, v. 410, p. 1911 – 1931, 2011. 77

DANTAS, M. **Computação distribuídas de alto desempenho**: redes, clusters e grids computacionais. Rio de Janeiro: Axcel Books do Brasil Editora, 2005. 288 p. 44

DAVIS, M.; EFSTATHIOU, G.; FRENK, C. S.; WHITE, S. D. M. The evolution of large-scale structure in a universe dominated by cold dark matter. **Astrophysical Journal**, v. 292, p. 371 – 394, 1985. 4

EFSTATHIOU, G.; DAVIS, M.; WHITE, S. D. M.; FRENK, C. S. Numerical techniques for large cosmological n-body simulations. **Astrophysical journal supplement series**, v. 57, p. 241–260, 1985. 52

El-REWINI, H.; ABD-EL-BARR, M. **Advanced computer architecture and parallel processing**. Hoboken: John Wiley and Sons, Inc, 2005. 273 p. 42

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **American Association for Artificial Intelligence**, v. 17, p. 37–54, 1996. 65

FLYNN, M. J. Very high-speed computing systems. **Proceedings of the IEEE**, v. 54, p. 1901–1909, 1966. 40

FOSTER, I. The grid: A new infrastructure for 21st century science. **Physics Today**, v. 55, p. 51–63, 2003. 8, 44

FOSTER, I.; KESSELMAN, C. The globus project: A status report. **IPPS/SPDP '98 Heterogeneous Computing Workshop**, v. 01, p. 4 – 18, 1998. 46

_____. **The grid**: blueprint for a new computing infrastructure. 2. ed. San Francisco: Morgan Kaufmann, 2004. 748 p. 8, 44

FOSTER, I.; KESSELMAN, C.; TUECKE, S. The anatomy of the grid: Enabling scalable virtual organizations. **International Journal of High Performance Computing Aplications**, v. 15, p. 200 – 222, 2001. 44, 45, 46

FOSTER, I. T. **Designing and building parallel programs**: concepts and tools for parallel software engineering. Boston: Addison-Wesley Longman Publishing Co., Inc., 1995. 381 p. 86

FREEDMAN, W. L.; MADORE, B. F.; et al. Final results from the hubble space telescope key project to measure the hubble constant. **Astrophys. J.**, v. 553, p. 47–72, 2001. 13

FREEMAN, K.; MACNAMARA, G. **In search of dark matter**. Chichester: Praxis Publishing, 2006. 158 p. 55

FRIEDMAN, A. On the curvature of space. **Zeitschrift für Physik**, v. 10, p. 377 – 386, 1922. 12

FRISCH, U. **Turbulence**: the legacy of a.n. kolmogorov. New York: Cambridge University Press, 1995. 296 p. 4, 29, 33, 34, 35

FRISCH, U.; SULEM, P.; NELKIN, M. A simple dynamical model of intermittent fully developed turbulence. **Journal of Fluid Mechanics**, v. 87, p. 719–736, 1978. 35

FUN, B.; REN, K.; LóPEZ, J.; FINK, E.; GIBSON, G. Discfinder: a data-intensive scalable cluster finder for astrophysics. In: INTERNATIONAL SYMPOSIUM ON HIGH PERFORMANCE DISTRIBUTED COMPUTING (HPDC), 19., 2010, Illinois, Chicago. **Proceedings...** Illinois: ACM, 2010. p. 348–351. ISBN 978-1-60558-942-8. 77

GAL, R. R.; CARVALHO, R. R.; et al. The digitized second palomar observatory sky survey (dposs). ii. photometric calibration. **The Astronomical Journal**, v. 128, p. 3082 – 3091, 2004. 5

GARDNER, J. P.; CONNOLLY, A.; MCBRIDE, C. Enabling knowledge discovery in a virtual universe. In: TERAGRID '07: BROADENING PARTICIPATION IN THE TERAGRID, 07., 2007, Madison, EUA. **Proceedings...** Madison: ACM Press, 2007. 77

_____. Enabling rapid developed of parallel tree search applications. In: SYMPOSIUM ON CHALLENGES OF LARGE APPLICATIONS IN DISTRIBUTED ENVIRONMENTS (CLADE), 5., 2007, Monterey, USA. **Proceedings...** Monterey: ACM, 2007. p. 1–10. ISBN 978-1-59593-714-8. 77

GAZTANAGA, E.; YOKOYAMA, J. Probing the statistics of primordial fluctuations and their evolution. **Astrophysical Journal**, v. 403, p. 450–465, 1993. 62

GELB, J. M.; BERTSCHINGER, E. Cold dark matter. i. the formation of dark halos. **The Astrophysical Journal**, v. 436, p. 467 – 490, 1994. 71

GUNN, J. E.; CARR, M.; et al. The sloan digital sky survey photometric camera. **The Astronomical Journal**, v. 116, p. 3040 – 3081, 1998. 58

HAMBLY, N. C.; MACGILLIVRAY, H. T.; et al. The supercosmos sky survey - i. introduction and description. **Monthly Notices of the Royal Astronomical Society**, v. 326, p. 1279 – 1294, 2001. 5

HAN, J.; KAMBER, M. **Data mining**: concepts and techniques. San Francisco: Academic Press, 2001. 770 p. 65

HAWLEY, J. F.; HOLCOMB, K. A. **Foundations of modern cosmology**. 2. ed. Oxford: Oxford University Press, 2005. 554 p. 1, 19, 21

HEISLER, J.; TREMAINE, S.; BAHCALL, J. N. Estimating the masses of galaxy groups: alternatives to the virial theorem. **The Astrophysical Journal**, v. 298, p. 08–17, 1985. 16, 18

HEY, T.; TANSLEY, S.; TENNEKES, K. T.; LUMLEY, J. **The fourth paradigm**: data-intensive scientific discovery. Redmond: Microsoft Research, 2009. 284 p. 65

HOCKNEY, R. W.; EASTWOOD, J. W. **Computer simulation using particles**. New York: McGraw-Hill, New York, 1981. 540 p. 53

HOGG, D. W. Distance measure in cosmology. **arXiv:astro-ph/9905116v4**, v. 4, p. 1–16, 2000. 15

HUBBLE, E. A relation between distance and radial velocity among extra-galactic nebulae. **Proceedings of the National Academy of Sciences of the United States of America**, v. 15, p. 168–173, 1929. 12, 13

HUCHRA, J. P.; GELLER, M. J. Groups of galaxies i. nearby groups. **The Astrophysical Journal**, v. 257, p. 423 – 437, 1982. 71, 72, 73

HUI, L.; BERTSCHINGER, E. Local approximations to the gravitational collapse of cold matter. **The Astrophysical Journal**, v. 471, p. 1–12, 1996. 2

JAROSIK, N.; BENNETT, C. L.; et al. Seven-year wilkinson microwave anisotropy probe (wmap) observations: Sky maps, systematic errors, and basic results. **The Astrophysical Journal Supplement Series**, v. 192, p. 1–42, 2011. 14

JEANS, J. H. The stability of a spherical nebula. **Philosophical Transactions of the Royal Society of London**, v. 199, p. 1 – 53, 1902. 1, 20

JENKINS, A.; PEARCE, C. S.; et al. Evolution of structure in cold dark matter universe. **The Astrophysical Journal**, v. 499, p. 20–40, 1998. 52, 53, 83

JONES, D. H.; READ, R. M. A.; et al. The 6df galaxy survey: Final redshift release (dr3) and southern large-scale structures. **arXiv:0903.5451v1 [astro-ph.CO]**, v. 399, p. 683–698, 2009. 5

KIM, S. C.; TAYLOR, J. D.; PANTER, B.; SOHN, S. T.; HEAVENS, A. F.; MANN, R. G. Using virtual observatory tools for astronomical research. **Journal of the korean astronomical society**, v. 38, p. 85–88, 2005. 6

KNEBE, A.; et al. Haloes gone mad: The halo-finder comparison project. **arXiv:1104.0949v1**, p. 1 – 27, 2011. 77

KOLB, E. W.; TURNER, M. S. **The early Universe**. New York: addison-wesley publishing company, 1990. 592 p. 1, 2, 13, 21

KOLMOGOROV, A. N. A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high reynolds number. **J. Fluid Mech.**, v. 13, p. 82 – 85, 1962. 28, 33

_____. Dissipation of energy in locally isotropic turbulence (1941-c). **reprinted in Proc. R. Soc. Lond. A**, v. 434, p. 15 – 17, 1991. 28

_____. The local sctructure of turbulence in incompressible viscous fluid for very large reynolds number (1941-a). **reprinted in Proc. R. Soc. Lond. A**, v. 434, p. 9 – 13, 1991. 28, 30, 31

KOTSIANTIS, S. B. Supervised machine learning: A review of classification techniques. **Informatica Journal**, v. 31, p. 249 – 268, 2007. 66

KWON, Y.; NUNLEY, D.; GARDNER, J.; BALAZINSKA, M.; HOWE, B.; LOEBMAN, S. Scalable clustering algorithm for n-body simulations in a shared-nothing cluster. In: INTERNATIONAL CONFERENCE ON SCIENTIFIC AND STATISTICAL DATABASE MANAGEMENT, 22., 2010, Heidelberg, Germany. **Proceedings...** Berlin: Springer-Verlag, 2010. p. 132–150. ISBN 3-642-13817-9 978-3-642-13817-1. 77

LACEY, C.; COLE, S. Merger rates in hierarchical models of galaxy formation - ii. comparison with n-body simulations. **Monthly Notices of the Royal Astronomical Society**, v. 271, p. 676 – 692, 1994. 71, 72

LESIEUR, M. **Turbulence in fluids**. 4. ed. Dordrecht: Springer, 2008. 558 p. 27, 29

LEVIN, F. S. **Calibrating the cosmos**: how cosmology explains our big bang universe. New York: Springer, 2007. 301 p. 13

LIDDLE, A. **An introduction to modern cosmology**. Chichester: John Wiley and Sons Ltd, 2003. 172 p. 22

LIDDLE, A. R.; LYTH, D. H. **Cosmological inflation and large-scale structure**. New York: Cambridge University Press, 2000. 400 p. 1, 2, 13, 21

LINDER, E. V. Exploring the expansion history of the universe. **Physical review letters**, v. 90, 2003. 12

LONGAIR, M. S. **Galaxy formation**. 2. ed. Berlin: Springer-Verlag, 2008. 735 p. 12, 16, 20, 22

MADSEN, M. **The dynamic Cosmos**: exploring the physical evolution of the universe. 1. ed. New York: Chapman & Hall, 1996. 144 p. 1, 2, 12, 13, 55

MATSUBARA, T. On second-order perturbation theories of gravitational instability in friedmann-lemaître models. **arXiv:astro-ph/9510137v1**, v. 94, p. 1151–1156, 1995. 2

MELOTT, A. L.; PELLMAN, T.; SHANDARIN, S. F. Optimizing the zel' dovich aproximation. **Monthly Notices of the Royal Astronomical Society**, v. 269, p. 626–638, 1994. 2

MONIN, A.; YAGLOM, A. M. **Statistical fluid mechanics**: mechanics of turbulence. 2. ed. Cambridge: MIT Press, 1975. 769 p. 29

NAKAMICHI, A.; MORIKAWA, M. Cosmic dark turbulence. **Astronomy & Astrophysics**, v. 498, p. 357–359, 2009. 3

OORT, J. H. The formation of galaxies and the origin of the high-velocity hydrogen. **Astronomy & Astrophysics**, v. 7, p. 381–404, 1970. 2

OURGRID Project. 2008. Acesso em: 21 nov. 2008. Disponível em: <www.ourgrid.org>. 47, 48

OZERNOY, L. M.; CHERNIN, A. D. The fragmentation of matter in a turbulent metagalactic medium. i. **Soviet Physics - Astronomy**, v. 11, p. 907 – 913, 1968. 2

OZERNOY, L. M.; CHIBISOV, G. V. Dynamical parameters of galaxies as a consequence of cosmological turbulence. **Soviet Astronomy - AJ**, v. 14, 1971. 2

PACHECO, P. **Parallel programming with MPI**. San Francisco: Morgan Kaufmann Publishers, Inc., 1997. 418 p. 41, 84

PADMANABHAN, T. **Structure formation in the universe**. New York: Cambridge University Press, 1993. 483 p. 1, 2, 13, 14, 19, 21, 26

_____. Challenges in non linear gravitational clustering. **C.R. Physique**, v. 7, p. 350 – 359, 2006. 1

_____. Dark energy: Mystery of the millennium. **arXiv:astro-ph/0603114v4**, v. 861, p. 179–196, 2006. 13

PARISI, G.; FRISCH, U. On the singularity structure of fully developed turbulence. In: INTERNATIONAL SCHOOL OF PHYSICS ENRICO FERMI, 1983, Varenna, Italy. **Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics**. North-Holland, Amsterdam: Elsevier Science Ltd, 1985. p. 84–87. ISBN 0444869360. 33

PEEBLES, P. **Principles of physical Cosmology**. New Jersey: Princeton University Press, 1993. 718 p. 18

PEEBLES, P. J. E.; RATRA, B. The cosmological constant and dark energy. **arXiv:astro-ph/0207347v2**, v. 2, p. 559–606, 2002. 13

POPE, S. B. **Turbulent flows**. 1. ed. New York: Cambridge University Press, 2000. 771 p. 29, 30, 31

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v. 1, p. 81–106, 1986. 67, 68

_____. **C4.5**: programs for machine learning. San Mateo: Morgan Kaufman, 1993. 302 p. 67

RAMOS, F. M.; ROSA, R. R.; NETO, C. R.; et al. Nonextensive statistics and three-dimensional fully developed turbulence. **Physica A: Statistical Mechanics and its Applications**, v. 295, p. 250–253, 2001. 35

REINDERS, J. **Vtune performance analyzer essentials**: measurement and tuning techniques for software developers. Hillsboro: Intel Press, 2005. 455 p. 76

RICKER, P. M.; SARAZIN, C. L. Off-axis clusters mergers: Effects of a strongly peaked dark matter profile. **The Astrophysical Journal**, v. 561, p. 621–644, 2001. 3

RUIZ, R. S. R.; CAMPOS VELHO, H. F.; CARETTA, C. A. Paralelização do algoritmo friends-of-friends para identificar halos de matéria escura. In: IX WORKSHOP DO CURSO DE COMPUTAÇÃO APLICADA, 2009, São José dos Campos. [S.l.], 2009. Acesso em: 18 jul. 2011. 89

RUIZ, R. S. R.; CAMPOS VELHO, H. F.; CARETTA, C. A.; CHARãO, A. S.; SOUTO, R. P. Grid environment for turbulent dynamics in cosmology. Aceito para ser publicado no Journal of Computational of Interdisciplinary Sciences. 2011. 89

RUIZ, R. S. R.; CAMPOS VELHO, H. F.; SANTOS, R. D. C.; TREVISAN, M. Árvore de decisão na classificação de dados astronômicos. **TEMA Tend. Mat. Apl. Comput.**, v. 10, p. 75–86, 2009. 70

SAHNI, V. Dark matter and dark energy. **Lect. Notes Phys**, v. 653, p. 141 – 180, 2004. 53

SAHNI, V.; COLES, P. Approximation methods for non-linear gravitational clustering. **Physics Reports**, v. 262, p. 1 – 135, 1995. 2, 53

SCHLICHTING, H. **Boundary layer theory**. New York: McGraw-Hill, 1979. 817 p. 37

SELJAK, U.; HAMAUS, N.; DESJACQUES, V. How to suppress the shot noise in galaxy surveys. **Physical Review Letters**, v. 103, p. 9–12, 2009. 62

SHANDARIN, S. F.; ZEL'DOVICH, Y. The large-scale of the universe: Turbulence, intermittency, structures in a self-gravitating medium. **Reviews of Modern Physics**, v. 61, p. 185 – 220, 1989. 3

SHE, Z.-S.; JACKSON, E.; ORSZAG, S. A. Structure and dynamics of homogeneous turbulence: models and simulations. **Proc. Royal Society London A**, v. 434, p. 101 – 124, 1991. 36

SHE, Z. S.; LéVêQUE, E. Universal scaling laws in fully developed turbulence. **Phys. Rev. Lett.**, v. 72, p. 336–339, 1994. 35

SKRUTSKIE, M. F.; CUTRI, R. M.; et al. The two micron all sky survey (2mass). **The Astronomical Journal**, v. 131, p. 1163 – 1183, 2006. 5

SPRINGEL, V.; SIMON, D. W.; JENKINS, A.; FRENK, C. S.; YOSHIDA, N.; GAO, L.; NAVARRO, J.; THACKER, R.; CROTON, D.; HELLY, J.; PEACOCK, J. A.; COLE, S.; THOMAS, P.; COUCHMAN, H.; EVRARD, A.; COLBERGM, J.; PEARCE, F. Simulating of the formation, evolution and clustering of galaxies and quasars. **Nature**, v. 435, p. 629 – 636, 2005. 52, 53, 56, 71, 83

SPRINGEL, V.; YOSHIDA, N.; WHITE, S. D. M. Gadget: a code for collisionless and gasdynamical cosmological simulations. **New Astronomy**, v. 6, p. 79 – 117, 2001. 57

STRAUSS, M. A.; WEINBERG, D. H.; LUPTON, R. H.; et al. Spectroscopic target selection in the sloan digital sky survey: The main galaxy sample. **Astronomical Journal**, v. 124, p. 1810–1824, 2002. 61

SZALAY, A. Science in an exponential world. In: **Australiasia eResearch Conference**. [S.l.: s.n.], 2007. 57

TATEKAWA, T. Langrangian perturbation theory in newtonian cosmology. **arXiv:astro-ph/0412025v4**, 2005. 1, 2, 21

TENNEKES, H.; LUMLEY, J. **A first course in turbulence**. Cambridge, MA: The MIT Press, 1972. 300 p. 27, 28

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern recognition**. 4. ed. Oxford: Academic Press, 2003. 837 p. 66, 67

TOSCANI, L.; VELOSO, P. A. S. **Complexidade de algoritmos**: análise, projeto e métodos. 2. ed. Porto Alegre: Bookman, 2001. 211 p. 75

TRENTI, M.; HUT, P. N-body simulations (gravitational). **Scholarpedia**, v. 3, p. 3930, 2008. 52, 53

VASCONCELLOS, E. C. **Árvores de decisão aplicadas ao problema da separação estrela/galáxia**. 101 p. Dissertação (Mestrado) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2011. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m19/2011/06.10.18.36>. Acesso em: 17 jul. 2011. 70

VASCONCELOS, E. C.; CARVALHO, R. R.; GAL, R. R.; LA BARBERA, F.; CAPELATO, H. V.; CAMPOS VELHO, H. F.; TREVISAN, M.; RUIZ, R. S. R. Decision tree applied in star/galaxy separation. **The Astronomical Journal**, v. 141, p. 1–30, 2011. 70

VAZZA, F.; TORMEN, G.; CASSANO, R.; BRUNETTI, G.; DOLAG, K. Turbulent velocity fields in smoothed particle hydrodynamics simulated galaxy clusters: scaling laws for the turbulent energy. **Monthly Notices of the Royal Astronomical Society**, v. 369, p. L14–L18, 2006. 3

VIRGO - The Virgo Consortium. 2008. Acesso em: 21 nov. 2008. Disponível em: <http://www.virgo.dur.ac.uk/>. 53

WEIZSACKER, C. F. V. The evolution of galaxies and stars. **The Astrophysical Journal**, v. 114, p. 165 – 186, 1951. 2

WITTEN, I. H.; FRANK, E. **Data mining**: practical machine learning tools and techniques with JAVA implementations. San Francisco: Morgan Kaufmann, 2000. 371 p. 66, 69, 70

XU, G. A new parallel n-body gravity solver: TPM. **Astrophys. J. Suppl.**, v. 98, p. 355 – 366, 1995. 57

ZEL' DOVICH, Y. B. Gravitational instability: An approximate theory for large densityperturbations. **Astronomy & Astrophysics**, v. 5, p. 84 – 89, 1970. 2, 35

ZWICKY, F. On the masses of nebulae and of clusters of nebulae. **The Astrophysical Journal**, v. 86, p. 217, 1937. 54

APÊNDICE A - ÁRVORES DE DECISÃO NA CLASSIFICAÇÃO DE DADOS ASTRONÔMICOS

TEMA Tend. Mat. Apl. Comput., **10**, No. 1 (2009), 75-86.

Árvores de Decisão na Classificação de Dados Astronômicos

R.S.R. RUIZ[1], H.F. DE CAMPOS VELHO [2], R.D.C. SANTOS [3], Laboratório Associado de Computação e Matemática Aplicada, LAC, INPE, 12227-010 São José dos Campos, SP, Brasil.

M. TREVISAN[4], Departamento de Astronomia, IAG, USP, 05508-900, São Paulo, SP, Brasil.

Resumo. Os registros de astronomia ótica constituem uma fonte de informação extremamente importante. Estas medidas são fundamentais para classificar estrelas e galáxias. Este trabalho descreve o algoritmo de construção de árvore de decisão (J4.8) e sua aplicação na construção de classificadores baseados em atributos fotométricos para classificar objetos astrônomicos em estrelas e galáxias. Dados do projeto Sloan Digital Sky Survey (SDSS) foram utilizados para treinamento e validação dos classificadores desenvolvidos. Os classificadores apresentaram índices de acerto, sobre o conjunto de teste, superiores a 98% para a classificação de estrelas e superiores a 99% para a classificação de galáxias.

Palavras-chave. Árvore de decisão, dados astronômicos, parâmetros fotométricos.

1. Introdução

O entendimento sobre a origem e evolução do Universo tem sido alterado ao longo dos tempos. Anteriormente, prevalecia a visão aristotélica: existia uma física para os fenômenos da Terra e outra física para os corpos celestiais. Isaac Newton alterou para sempre este paradigma e desenvolveu um modelo físico-matemático constituído de poucos postulados e algumas leis, como a formulação matemática da gravitação Universal. O modelo cosmológico de Newton é de um Universo infinito aparentemente estável e estático.

Em 1917, Albert Einstein propôs um modelo cosmológico relativístico, considerando ainda o Universo como estático. Alguns anos depois, dados de observação aliados a teoria permitiram uma das mais importantes descobertas da astronomia no século XX: o Universo está em expansão. Esta descoberta foi realizada por Hubble em 1929, quando descobriu que as galáxias estão se afastando. Tal descoberta marca o fim da era de um Universo estático e o início de uma nova era. Surge, então, o modelo cosmológico de um Universo em expansão [7, 14].

[1]renata@lac.inpe.br - A autora agradece a FAPESP bolsa de doutorado (2007/54133-0)
[2]haroldo@lac.inpe.br
[3]rafael.santos@lac.inpe.br
[4]marinatrevisan@gmail.com

Conforme [14], a dinâmica composta pela força da gravidade e pela expansão do Universo descreve a história da formação das grandes estruturas cosmológicas (galáxias, aglomerados de galáxias, super-aglomerados, etc.). As bases de dados astronômicos existentes hoje fornecem uma possibilidade de estudo dessas estruturas sem precedentes. Porém, o estudo dessas estruturas depende do correto mapeamento de galáxias, mas, numa imagem astronômica nem sempre é fácil fazer a distinção entre uma galáxia e uma estrela. As dificuldades para análise baseada em atributos fotométricos estão relacionadas a vários fatores, entre eles: baixa luminosidade, baixo brilho superficial, perfis extensos, diferentes resoluções angulares e a razão sinal-ruído. Por exemplo, quanto mais distante está uma galáxia do nosso planeta, menor é o seu tamanho na imagem e menor é a luminosidade observada. Quando se atinge um limite crítico de tamanho e luminosidade é difícil distinguir entre uma galáxia muito distante e uma estrela de baixa luminosidade da nossa própria galáxia [7].

A Figura 1 ilustra o tipo de dificuldade referente a luminosidade, na figura pode-se observar que na parte superior é simples realizar a classificação de um objeto, na parte central a classificação ainda é simples, mas o número de objetos nesse domínio de luminosidade é muito grande para ser feito visualmente, na parte inferior a identificação é muito complexa e necessita de métodos sofisticados. Tais métodos se baseiam em um conjunto de parâmetros que descrevem a imagem. Esses parâmetros podem ser fotométricos ou espectroscópicos. Porém, a aquisição de espectros em geral requer um tempo maior de observação e utilizar dados fotométricos tornam as observações mais eficientes.

Neste contexto, separar estrelas de galáxias a partir de dados fotométricos é um desafio interessante e o objetivo deste trabalho é aplicar a técnica de árvores de decisão a este problema. Há na literatura uma série de trabalhos utilizando árvores de decisão para classificar objetos astronômicos [4, 20, 26, 27]. Outras técnicas de classificação também muito utilizadas são as redes neurais artificiais [2, 8, 15] e o algoritmo *Friends of Friends* [10].

Em particular, para dados do projeto *Sloan Digital Sky Survey* (SDSS) uma abordagem em árvores de decisão foi utilizada por [22] na classificação de objetos fotométricos em estrelas, galáxias e Núcleos Galácticos Ativos (AGN). Para o desenvolvimento do modelo foi utilizado o projeto ClassX. Este projeto é um sistema online que foi originalmente desenvolvido para realizar a classificação de fontes de raio X. O algoritmo de criação da árvore de decisão utilizado pelo ClassX é o sistema OC1 de [16].

Árvore de decisão também foi utilizada por [3] para realizar a classificação de objetos da terceira divulgação de dados do SDSS em estrelas, galáxias ou "nem estrela nem galáxia". O classificador foi treinado sobre um conjunto de 477.068 objetos espectroscopicamente classificados. Posteriormente, o classificador desenvolvido foi utilizado em cerca de 143 milhões de objetos do projeto, sendo este o primeiro trabalho a utilizar árvores de decisão para classificar um conjunto inteiro de dados do SDSS.

No presente estudo, será utilizado algoritmo de construção de árvore de decisão C4.5 [18], implementado no software *Waikato Environment for Knowledge Analysis* (WEKA) como J4.8 [25], para desenvolver classificadores baseados em atributos

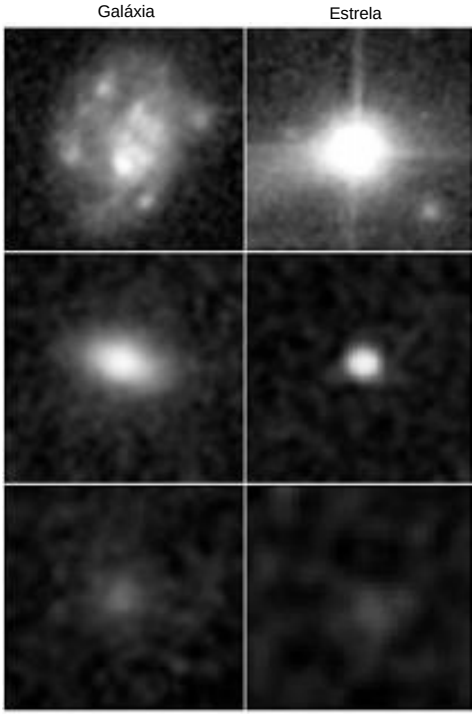

Figura 1: Exemplos de dificuldade crescente de separação estrela-galáxia – Fonte: [7].

fotométricos para serem utilizados na classificação de objetos astronômicos do projeto SDSS em estrelas ou galáxias. As demais seções deste trabalho estão divididas da seguinte maneira: Na Seção 2 tem-se a descrição do algoritmo C4.5, a Seção 3 apresenta os dados utilizados no treinamento e validação dos classificadores, os resultados obtidos são apresentados na Seção 4. Finalmente, a Seção 5 é reservada as considerações finais.

2. Árvores de Decisão e Classificação

Dado um conjunto de objetos descritos em termos de uma coleção de atributos, estes objetos podem pertencer a diferentes classes. Cada atributo expressa alguma característica importante de um objeto. Parte destes objetos, serão considerados para o treinamento e tem sua classificação previamente conhecida. Conforme [19] é possível desenvolver uma regra de classificação que pode determinar a classe de qualquer objeto a partir dos valores dos seus atributos. Tal regra de classificação pode ser expressa como uma árvore de decisão. Uma árvore de decisão é uma estru-

tura simples em que as folhas contêm as classes, os nodos não-folhas representam atributos baseados em testes com um ramo para cada possível saída [11, 18, 19]. Para classificar um objeto, começa-se com a raiz da árvore, aplica-se o teste em cada nodo e toma-se o ramo apropriado para aquela saída. O processo continua e quando uma folha é encontrada o objeto é classificado segundo a classe indicada naquela folha.

Com os atributos adequados, é sempre possível construir uma árvore de decisão que classifique corretamente os objetos no conjunto de treinamento e normalmente existem muitas árvores de decisão corretas. Mas o objetivo dos algoritmos de indução (construção) é ir além do conjunto de treinamento, isto é, criar árvores capazes de classificar corretamente outros objetos. Para conseguir isto, tais algoritmos devem capturar alguma relação significativa entre a classe do objeto e os valores de seus atributos.

São diversos os algoritmos de indução de árvores de decisão conhecidos na literatura, dos quais destacam-se: *Random Forest* [5, 6],*ADTree* [9], *NBTree* [12], C4.5 [18] e o ID3 [19]. O algoritmo ID3 e o C4.5 são os mais populares.

2.1. Algoritmo de Indução C4.5 ou J4.8

Como formar uma árvore de decisão para um conjunto C de objetos? Se C é vazio ou contém somente objetos de uma mesma classe, a árvore de decisão mais simples contém uma folha que representa essa classe. Caso contrário, seja T algum teste sobre um objeto que tem os possíveis resultados $O_1, O_2, \ldots, O_w$. Existe um mapeamento entre cada objeto em C associado aos resultados para T, portanto T produz uma partição $\{C_1, C_2, \ldots, C_w\}$ de C, com C_i contendo aqueles objetos que tem resultado O_i. Se cada subconjunto C_i pode ser substituido por uma árvore de decisão, o resultado será uma árvore de decisão para todos os elementos de C. No pior caso essa estratégia fornecerá subconjuntos de um único objeto. Assim, uma vez que, um teste que gera uma divisão não trivial de qualquer conjunto de objetos sempre pode ser encontrado, este procedimento produzirá uma árvore de decisão que classifica corretamente os objetos em C [19].

A escolha do teste é crucial para a árvore de decisão ser simples. O algoritmo C4.5 adota um critério baseado na teoria da informação que depende de duas hipóteses:

- No caso de uma amostra de objetos que pertencem somente a duas classes, por exemplo, P e N, um objeto qualquer pertencerá a classe P com probabilidade $p/(p+n)$ e a classe N com probabilidade $n/(p+n)$, em que p é o número total de objetos que pertencem a classe P e n o número total de objetos pertencentes a classe N.
- Quando uma árvore de decisão é usada para classificar um objeto, ela retorna uma classe. Árvore de decisão pode ser considerada como uma fonte de mensagem P ou N em que a informação necessária para gerar a mensagem é obtida conforme equação (2.1).

$$I(p,n) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right). \qquad (2.1)$$

Se o atributo A com os valores $[A_1, A_2, \ldots, A_v]$ é usado para a raíz da árvore de decisão, ela dividirá C em $\{C_1, C_2, \ldots, C_v\}$, onde C_i contém aqueles objetos em C que tem valores A_i de A. Considere C_i contendo p_i objetos da classe P e n_i da classe N. A informação necessária para a subárvore em C_i é $I(p_i, n_i)$. A informação necessária para a árvore com A como raiz é obtida com a média ponderada, conforme equação (2.2)

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i), \tag{2.2}$$

em que o peso para o i-ésimo ramo é proporcional aos objetos em C que pertencem a C_i. Portanto, o ganho de informação obtido por esse ramo usando o atributo A é dado pela equação (2.3)

$$G(A) = I(p, n) - E(A). \tag{2.3}$$

O algoritmo C4.5 examina todos os atributos candidatos e escolhe A que maximiza o ganho de informação. O processo é repetido recursivamente para obter os demais nós e formar a árvore de decisão com os subconjuntos restantes [18, 19]. Na Figura 2 pode-se observar um fluxograma do algoritmo de construção de árvore de decisão.

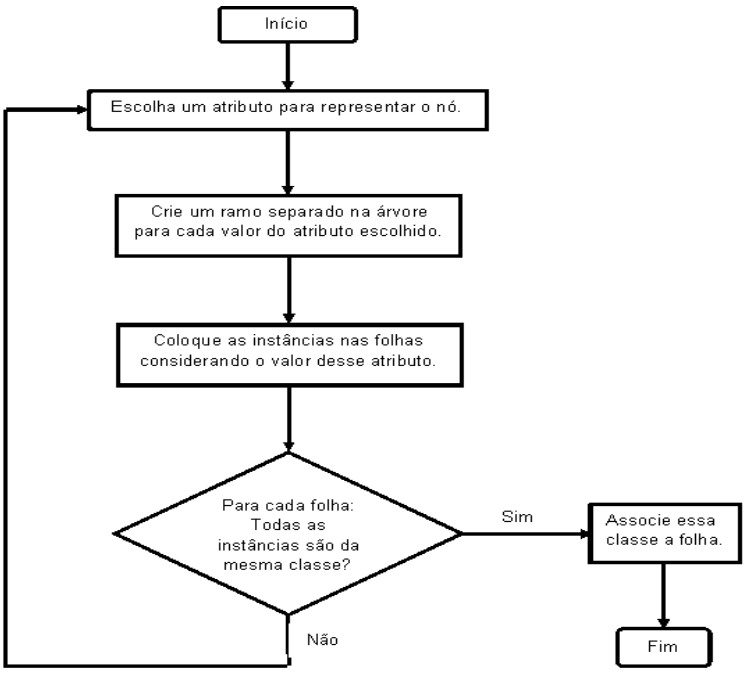

Figura 2: Fluxograma do algoritmo de árvore de decisão – Fonte: [18].

3. Aquisição dos Dados e Parâmetros Utilizados

Os dados utilizados neste trabalho são dados de seis anos do projeto SDSS [1]. Este levantamento cobre uma área de aproximadamente 10.000 graus quadrados do céu, contendo imagens de 287 milhões de objetos. A câmera do SDSS mede quão brilhantes são os objetos em cinco bandas fotométricas denominadas u, g, r, i, z. Além disso, há também o levantamento espectroscópico, que cobre uma área de aproximadamente 7.500 graus quadrados, com mais de um milhão de objetos catalogados.

A classificação de objetos baseada nos espectros é mais confiável. Sendo assim, os objetos do catálogo fotométrico foram selecionados levando em conta também as informações do catálogo espectroscópico, de forma a minimizar a falsa classificação de objetos nas amostras de treinamento e de testes.

Cada objeto identificado nas imagens é classificado pelo próprio *pipeline* do SDSS, em todas as cinco bandas. O mesmo é feito para os dados espectroscópicos, de forma que há seis parâmetros. Os parâmetros relacionados com esta classificação são *type-u*, *type-g*, *type-r*, *type-i*, *type-z* e *SpecClass*. Os cinco primeiros recebem os valores 3 se for galáxia e 6 se for estrela e *SpecClass* é 1 no caso de estrela e 2 se galáxia. Baseando-se nisso, o primeiro critério de seleção foi exigir que o objeto tenha a mesma classificação nestes seis parâmetros simultaneamente. Isso reduz a quase nula a probabilidade de ter na amostra um objeto classificado erroneamente (1.5 % dos objetos classificados nas bandas u, g, r, i e z como galáxias não possuem a mesma classificação quando considerado o espectro).

Entre estes objetos seguramente classificados como estrelas e galáxias, foram ainda impostas restrições aos *flags* de qualidade, gerados pelo *pipeline* do SDSS, relacionados à saturação do objeto e à detecção de mútiplos picos de intensidade nas imagens. A amostra final é composta por 43.289 estrelas e 452.400 galáxias. Os dados foram obtidos através do servidor CasJobs do SDSS [13, 23, 24].

Dos objetos da amostra final foram selecionados os parâmetros fotométricos considerados relevantes na distinção entre estrelas e galáxias, em todas as cinco bandas. A seguir uma breve descrição desses parâmetros.

- **nprof:** De cada objeto é extraído o perfil radial de brilho superficial. Este perfil é dado como a média azimutal do brilho em uma série de anéis, cujos raios podem ser encontrados na tabela 7 de [21]. O parâmetro *nprof* corresponde ao números de anéis para os quais ainda existe um sinal mensurável.

- **PetroR50, PetroR90:** Para cada objeto é definido o perfil de brilho superficial Petrosiano [17], e a partir deste são definidos os raios PetroR50 e PetroR90, que correspondem aos raios que compreendem 50% e 90% do fluxo Petrosiano, respectivamente. De uma maneira simplificada, estes podem ser entendidos como uma medida da "extensão"do objeto. Objetos mais difusos como galáxias tendem a ter o raio petrosiano maior.

- **isoA, isoB:** Os atributos isoA e isoB são definidos como o eixo maior e o eixo menor da figura geométrica representativa do objeto e são utilizados para encontrar a excentricidade. Logo, ambos se convertem em um único parâmetro a ser utilizado no treinamento.

- **Magnitudes:** Foram utilizadas as magnitudes Petromag, PSFmag, Fibermag, Modelmag. Magnitude é uma medida do brilho aparente do objeto e cada uma das quatro magnitudes são obtidas considerando modelos diferentes para o perfil de brilho: perfil petrosiano, perfil da *Point Spread Function*, da fibra ótica (dado espectroscópico) e a magnitude baseada no modelo que melhor se ajusta. Uma descrição detalhada das magnitudes pode ser encontrada em [21].

- **Redshift espectroscópico:** Este não é um dado fotométrico, mas sim obtido a partir dos espectros. Como é baseado em linhas de emissão e absorção, não apenas no fluxo em bandas, é uma medida mais precisa de distância do que o *redshift* fotométrico. O *redshift* de um objeto é medido como o deslocamento relativo do comprimento de onda emitido pela fonte e o observado:

$$z = \frac{\lambda_{\text{Observado}} - \lambda_{\text{Emitido}}}{\lambda_{\text{Emitido}}} \tag{3.1}$$

  o aumento no comprimento de onda é causado pela expansão do universo: quanto mais distante o objeto, maior é o seu *redshift* (z). Apesar de esperar-se que estrelas tenham sempre o *redshift* nulo, isso nem sempre é verdade, pois a mudança no comprimento de onda também pode ser causada por efeito Doppler devido ao movimento da fonte em relação ao observador.

4. Resultados Obtidos

O treinamento (criação da árvore) foi realizado com um conjunto de 10^4 objetos (925 estrelas e 9075 galáxias) utilizando-se o algoritmo J4.8. As árvores foram criadas, analisadas e modificadas (variando-se o número mínimo de objetos por folha, fator de confiança usado na poda, entre outros) para evitar que regras muito complexas mas que correspondem a poucos casos no conjunto de dados fossem criadas. Na verdade, uma árvore de decisão é semelhante a um sitema especialista, pois fornece um conjunto de regras que devem ser aplicadas a um determinado objeto para obter sua classe. Nesta seção, será apresentado o desempenho de duas árvores que foram implementadas em linguagem C e testadas sobre o conjunto total de amostras (495.689 objetos).

O atributo *redshift* permite classificar de imediato os objetos em estrelas e galáxias. Assim, a amostra de treinamento continha todos os atributos fotométricos descritos na Seção 3, juntamente com o *redshift* espectroscópico. Nesse caso, o objetivo do treinamento era verificar a robustez do algoritmo J4.8 na identificação do atributo mais importante para a separação das classes. A árvore de decisão obtida com esse conjunto de treinamento definiu um valor crítico para o *redshift*: se $\textit{redshift} \leq 0,001481$ o objeto é classificado como estrela, se $\textit{redshift} > 0,001481$ o objeto é uma galáxia. A aplicação dessa regra sobre a amostra total forneceu um índice de acerto de 99.99%. Este resultado demonstrou a eficiência do algoritmo na escolha do atributo hegemônico na classificação.

Após este teste, o próximo passo foi remover o parâmetro *redshift* do conjunto de treinamento e realizar a criação de árvores somente com os outros parâmetros

fotométricos, portanto, cada objeto do conjunto de treinamento é descrito em termos de 40 parâmetros (8 atributos x 5 bandas). A estratégia de seleção de atributos utilizada pelo algoritmo J4.8 permitiu a identificação do parâmetro PetroR50 como o atributo que fornece uma melhor separação entre as classes, ou seja, que contêm *a maior quantidade de informação*. A Figura 3 exibe a primeira árvore e seu desempenho sobre o próprio conjunto de treinamento (10.000 objetos). Os números dentro dos retângulos que representam as classes (estrela ou galáxia) indicam a quantidade de objetos classificados corretamente/erroneamente por aquela folha. Para a criação desta árvore foi usado poda, fator de confiança de 0,25 e valor mínimo de 5 objetos por folha. Na Figura 4 tem-se a segunda árvore e também seu desempenho sobre o conjunto de treinamento. Os parâmetros de configuração foram os mesmos da primeira com exceção do número mínimo de objetos por folha que, nesse caso, foi estipulado 20 objetos.

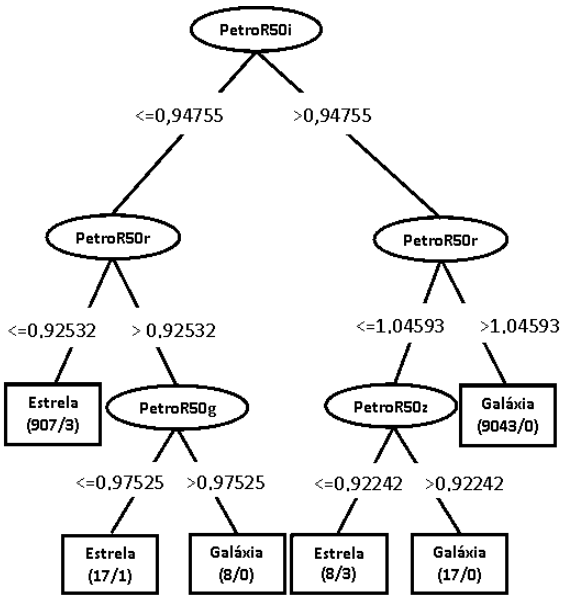

Figura 3: Esquema da primeira árvore de decisão

A Tabela 1 apresenta os resultados obtidos com as duas árvores sobre o conjunto de teste. Observa-se que os resultados usando a primeira árvore indicam que 523 estrelas do total de 43.289 foram classificadas erroneamente como galáxias e 24 estrelas não foram classificadas devido a ausência de algum atributo. Na classificação de galáxias, os resultados mostram que 1.866 galáxias do total de 452.400 foram classificadas erroneamente como estrelas e 52 galáxias não foram classificadas, também devido a ausência de algum atributo. O índice de acerto para a classificação de estrelas em termos de porcentagem foi de 98,79% e para a classificação de galáxias foi de 99,59%. Os resultados obtidos com a segunda árvore mostram que 568 estrelas foram classificadas erroneamente como galáxias e 13 estrelas não foram classificadas.

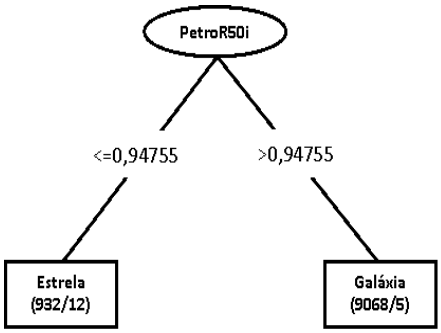

Figura 4: Esquema da segunda árvore de decisão

Na classificação de galáxias, 2.344 foram classificadas erroneamente como estrelas e 45 não foram classificadas. O índice de acerto para a classificação de estrelas foi de 98,69% e para a classificação de galáxias foi de 99,48%. Finalmente, na última linha da tabela tem-se o índice kappa, que é uma medida da acurácia da classificação, obtido por meio da matriz de confusão. Quanto mais próximo de 1 for o índice *kappa*, melhor é o desempenho do classificador.

Tabela 1: Resultados obtidos com as duas árvores de decisão sobre o conjunto de 495.689 objetos astronômicos.

	1ª **árvore**		2ª **árvore**	
	Estrelas	**Galáxias**	**Estrelas**	**Galáxias**
Estrelas	42.742	523	42.708	568
Galáxias	1.866	450.482	2.344	450.011
Objetos não classificados	24	52	13	45
Índice de acerto	98,79%	99,59%	98,69%	99,48%
Índice kappa	0,97		0,96	

Pode ser observado na Tabela 2 o índice de acertos referente a classificação de estrelas e galáxias do projeto SDSS obtidos com os classificadores desenvolvidos neste trabalho e também com os trabalhos de [22] e [3]. Os parâmetros fotométricos utilizados em ambos os trabalhos da literatura foram as cores dos objetos. A cor de um objeto pode ser medida através das diferenças de magnitudes entre os filtros. No trabalho de [22] foi usado as diferenças $u - g$, $g - r$, $r - i$, $i - z$ e $g - i$. Já no trabalho [3] foi utilizado as mesmas cores que [22] com exceção de $g - i$. Conforme pode ser observado, os classificadores baseados em árvores de decisão mencionados neste trabalho utilizando parâmetros fotométricos, apresentaram um desempenho similar ao obtido nos trabalhos de [22] e [3]. O índice de acerto na classificação de estrelas foi cerca de 0,60% superior a [22] e cerca de 5, 30% superior a [3]. Para a classificação de galáxias o índice de acerto foi cerca de 1,00% superior a ambos os

trabalhos da literatura.

Tabela 2: Comparação entre os resultados obtidos neste trabalho e os obtidos pelos trabalhos de Suchkov et al. [22] e Ball et al. [3].

Classificadores	**Atributos utilizados**	**Índice de acerto**	
		Estrelas	**Galáxias**
Suchkov et al.[22]	*Cores dos objetos*	98,10%	98,50%
Ball et al.[3]	*Cores dos objetos*	93,40%	98,20%
1ª **árvore**	*Raio PetroR50 (bandas i, r, g e z)*	98,79%	99,59%
2ª **árvore**	*Raio PetroR50 (banda i)*	98,69%	99,48%

5. Considerações Finais

A técnica de árvores de decisão foi empregada na classificação de estrelas e galáxias para dados do projeto SDSS, com base em parâmetros fotométricos, onde o algoritmo de construção empregado foi o C4.5, implementado no *software* WEKA como J4.8. A estratégia do sistema de árvores de decisão é um sistema automático de projeto de um classificador, baseado em aprendizado de máquina, que tornam explícitos os atributos mais relevantes. Existem algumas diferenças entre os índices de acerto obtidos com o presente trabalho e os resultados da literatura. Estas diferenças podem ser atribuídas a vários fatores, entre eles: (a) nos trabalhos anteriores outras classes foram consideradas (e não somente estrela/galáxia); (b) há diferenças nos parâmetros fotométricos analisados e/ou limiares considerados; (c) estratégias da configuração das árvores de decisão (número mínimo de objetos por folhas, dentre outros); (d) os autores citados utilizaram outros algorítimos para implementar seus classificadores baseado em árvores de decisão.

Os classificadores desenvolvidos com árvores de decisão no presente trabalho alcançaram desempenho similar aos classificadores desenvolvidos por Suchkov et al. [22] e Ball et al. [3]. Esses resultados mostram que o algoritmo de indução testado é robusto para o desenvolvimento de classificadores com base em atributos fotométricos dos dados do projeto SDSS.

Abstract. The optical measurement data constitute a source of very important information for the astronomy. Such measurements is fundamental to classify stars and galaxies. This work describes the algorithm to design decision trees (J4.8 algorithm). The classifiers were employed to the astronomical data from the project Sloan Digital Sky Survey (SDSS). The performance for the best classifiers for the test set was greater than 98% for stars classification, and greater than 99% for galaxies classification.

Referências

[1] J. Adelman-Mccarthy et al., The sixth data release of the sloan digital sky survey. *The Astrophysical Journal Supplement Series*, **175**, No. 2 (2008), 297–313.

[2] N. M. Ball et al., Galaxy types in the Sloan Digital Sky Survey using supervised artificial neural networks, *Monthly Notices of the Royal Astronomical Society*, **348** (2004), 1038–1046.

[3] N. M. Ball, R. J. Brunner, A. D. Myers, Robust machine learning applied to astronomical datasets I: star-galaxy classification of the sloan digital sky survey DR3 using decision trees. *The Astrophysical Journal*, **650** (2006), 497–509.

[4] D. Bazell, D. W. Aha, Ensembles of classifiers for morphological galaxy classification, *The Astrophysical Journal*, **548** (2001), 219–223.

[5] L. Breiman, Random forests, *Machine Learning*, **45**, No. 1 (2001), 5–32.

[6] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, "Classification and regression trees", U.S.A: Wadsworth Publishing Company, 1984.

[7] R.R Carvalho, H.V. Capelato, H.F. Campos Velho, Um universo escuro na era da tecnologia da informação, *Boletim da Sociedade Brasileira de Astronomia* - (submetido).

[8] F. Cortiglione, P. Mahonen, P. Hakala, T. Franti, Automated Star-Galaxy discrimination for large surveys, *The Astrophysical Journal*, **556** (2001), 937–943.

[9] Y. Freud, L. Mason, The alternating decision tree learning algorithm, *Proceedings of the Sixteenth International Conference on Machine Learning*, (1999), 124–133.

[10] J.P. Huchra, M.J. Geller, Groups of galaxies I. Nearby groups, *The Astrophysical Journal*, **257** (1982), 423–437.

[11] E.B. Hunt, J. Marin, P.J. Stone, "Experiments in Induction". New York: Academic Press, 1966.

[12] R. Kohavi, Scaling up the accuracy of naive - Bayes classifiers: a decision tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press*, (1996), 202–207.

[13] N. Lin, A.R. Thakar, CasJobs and MyDB: A batch query workbench, *Computing in Science and Engineering* , **10**, No. 1 (2008), 18–29.

[14] M.S. Madsen, "The Dynamic Cosmos - Exploring the Physical Evolution of the Universe", New York, NY, USA: Chapman e Hall, 1996.

[15] A.S. Miller, M.J. Coe, Star/galaxy classification using Kohonen self-organizing maps. *Monthly Notices of the Royal Astronomical Society*,**279**,(1996), 293–300.

[16] K.S. Murty, S. Kasif, S. Salzberg, A system for induction of oblique decision tree, *Journal of Artificial Intelligence Research*, **2**, (1994), 1–32.

[17] V. Petrosian, Surface brightness and evolution of galaxies, *The Astrophysical Journal*, **209**, No. 1 (1976).

[18] J.R. Quinlan, "C4.5: Programs for Machine Learning". San Mateo, CA: Morgan Kaufman, 1993.

[19] J.R. Quinlan, Induction of decision trees. *Machine Learning*, **1**, No. 1 (1986), 81–106.

[20] S. Salzberg et al., Decision trees for automated identification of cosmic ray hits in hubble space telescope images,*Publications of the Astronomical Society of the Pacific*, **107** (1995), 1–10.

[21] C. Stoughton, R.H. Lupton, M. Bernardi, M.R. Blanton, Sloan Digital Sky Survey: early data release. *The Astrophysical Journal*, **123**, (2002), 485–548.

[22] A. Suchkov, R.J. Hanisch, B. Margon, A Census of object types and redshift estimates in the SDSS photometric catalog from a trained decision tree classifier, *The Astronomical Journal*, **130**, (2005), 2439–2452.

[23] A.S. Szalay, A.R. Thakar, J. Gray, The sqlLoader data-loading pipeline, *Computing in Science and Engineering*, **10**, No. 1 (2008), 38–48.

[24] A.R Thakar, A.S. Szalay, G. Fekete, J. Gray, The catalog archive server database management system. *Computing in Science and Engineering*, **10**, No. 1 (2008), 30–37.

[25] I.H. Witten, E. Frank, "Data mining: Practical Machine Learning Tools and Techniques with JAVA Implementations". San Francisco: Morgan Kaufmann, 2000.

[26] Y. Zhang, Y. Zhao, A comparison of BBN, ADTree and MLP in separating quasars from large survey catalogues, *Chinese Journal of Astronomy and Astrophysics*, **7**, No. 2 (2007), 289–296.

[27] Y. Zhao, Y. Zhang, Comparison of decision tree methods for finding active objects, *Advances in Space Research*, **41**, No. 1 (2008), 1955–1959.

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.