UNIVERSIDADE ESTADUAL DE SANTA CRUZ
PRO-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL EM CIÊNCIA E TECNOLOGIA

SILAS SILVA SANTOS

CÓDIGO EM R PARA OBTENÇÃO DA DISTRIBUIÇÃO DE MASSA A PARTIR DE UM CATÁLOGO DE AGLOMERADOS DE GALÁXIAS

PPGMC – UESC

ILHÉUS-BA
2017

SILAS SILVA SANTOS

CÓDIGO EM R PARA OBTENÇÃO DA DISTRIBUIÇÃO DE MASSA A PARTIR DE UM CATÁLOGO DE AGLOMERADOS DE GALÁXIAS

PPGMC – UESC

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional em Ciência e Tecnologia da Universidade Estadual de Santa Cruz, como parte das exigências para obtenção do título de Mestre em Modelagem Computacional em Ciência e Tecnologia.

Orientador: Prof. Dr. André Luis Batista Ribeiro

ILHÉUS-BA
2017

S237 Santos, Silas Silva.
Código em R para obtenção da distribuição de massa a partir de um catálogo de aglomerados de galáxias PPGMC - UESC. – Ilhéus, BA: UESC, 2017.
87 f. il.

Orientador: André Luís Batista Ribeiro.
Dissertação (Mestrado) – Universidade Estadual de Santa Cruz. Programa de Pós-Graduação em Modelagem Computacional.
Referências e apêndices.

1. Galáxias. 2. Galáxias – Aglomerados. 3. Astrofísica. 4. Galáxias – Catálogos. 5. Modelagem – Computação. I.Título.

CDD 523.112

SILAS SILVA SANTOS

CÓDIGO EM R PARA OBTENÇÃO DA DISTRIBUIÇÃO DE MASSA A PARTIR DE UM CATÁLOGO DE AGLOMERADOS DE GALÁXIAS

PPGMC – UESC

Ilhéus-BA, 17/02/2017

Comissão Examinadora

Prof. Dr. André Luis Batista Ribeiro
UESC
(Orientador)

Prof. Dr. Francisco Bruno Souza Oliveira
UESC

Prof. Dr. Cássio Bruno Magalhães Pigozzo
UFBA

Dedico esse trabalho à minha mãe, Mayse, por ser minha provedora e estar sempre do meu lado quando precisei. Aos meus irmãos, Virgilio e Derlan, por estarem comigo nos momentos mais difíceis. Ao meu pai, Aristides (in memorian), por ter me posto neste caminho o qual está me levando a este dia, dedico a meu primo e irmão Emico(in memorian), por ter me mostrado como a carreira acadêmica poderia ser a realização dos meus sonhos.

Muito obrigado!

Agradecimentos

- Agradeço primeiramente ao meu orientador por ter tornado esse trabalho possível, e ter me auxiliado em todas as dúvidas que tive durante esses anos de pesquisa.
- Agradeço aos membros do LATO em especial para meu amigo Alisson que sempre esteve comigo durante as reuniões e se mostrou disponível para me ajudar a qualquer momento.
- Agradeço a CAPES por ter financiado todo o projeto durante esses dois anos, e a UESC pela toda infraestruturada fornecida.
- Agradeço aos meus amigos e familiares que sempre me apoiaram e me incentivaram a seguir a carreira acadêmica, em especial para Daniel, Gabriel, Zara, Laila e Ludmille que seguraram a barra nessa reta final do projeto, aos meus amigos Matheus, Rarana, Uendson, Júlio, Cel, Sara, Jayze, Firmino, Ícaro, Joel, Alina, Dayanna entre outros que estiveram presente nos momentos de descontração que foram de extrema importância nesses anos.

Código em R para obtenção da distribuição de massa a partir de um catálogo de aglomerados de galáxias

PPGMC – UESC

Resumo

Implementamos uma série de métodos para variadas análises em astrofísica extragaláctica com foco na obtenção da massa de aglomerados de galáxias. A integração destes métodos em um programa único tem como intuito elaborar e publicar um pacote no ambiente R. O programa contém métodos que oferecem soluções para remoção de galáxias intrusas, um problema inerente ao estudo de qualquer conjunto de dados nesse campo de trabalho. O programa contém ainda métodos estatísticos que realizam inferências sobre o estado dinâmico dos sistemas, assim como métodos que estimam a massa e o raio de aglomerados de galáxias, propriedades fundamentais tanto para a pesquisa em astrofísica como em cosmologia. O trabalho tem como resultado um *pipeline* flexível e otimizado combinando a utilização destes diferentes métodos. Aplicamos este *pipeline* a um catálogo simulado de aglomerados de galáxias e comparamos o desempenho dos diferentes métodos para recuperar as informações conhecidas do catálogo. Esta comparação permitirá ao usuário do programa escolhas mais eficientes de parâmetros e estimadores. Ao final, discutimos o trabalho e apresentamos algumas perspectivas futuras.

Palavras-chave: Obtenção da Massa. Aglomerados de Galáxias. Modelagem Computacional.

R-code for obtaining mass distribution from a catalog of galaxy clusters

PPGMC - UESC

Abstract

We implemented several methods for various analyses in extragalactic astrophysics with focus on obtaining the mass of galaxies clusters. The integration of these methods in a single program is intended to elaborate and publish a package for the R environment. The program contains methods that provide solutions for removal of interloper galaxies, an inherent problem in the study of any dataset in this research area. The program also contains statistical methods that estimate the radius and mass of galaxies clusters, fundamental properties both for the research in astrophysics as in cosmology. The work results in a flexible and optimized pipeline by combining the use of these different methods in each step of the analysis. We apply this pipeline to a simulated catalog of galaxies clusters and compar the performance of the different methods to recover the information that is known from the catalog. This study will allow the user of the program to choose more efficiently parameters and estimators..

Keywords: obtaining mass, galaxies clusters, computational modeling.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

UESC	Universidade Estadual de Santa Cruz
DCET	Departamento de Ciências Exatas e Tecnológicas
PPGMC	Programa de Pós-Graduação em Modelagem Computacional em Ciência e Tecnologia
SDSS	Sloan Digital Sky Survey
PS	Press & Schechter
FM	Função de Massa
HD	Distância de Hellinger
EM	Expectation-Maximization
HTB	Heisler, Tremaine and Bahcall
DS	Dressler and Shectman
MOCK	Palavra designada para catálogos construídos artificialmente
GNG	Gaussiano ou Não Gaussiano
Galclus	Rotina de analise do estado dinâmico do sistema
Galremov	Rotina de remoção de `outliers`
Massa	Rotina que estimativa de massa

Sumário

1 Introdução

Entre os objetos de maior interesse da astronomia moderna se encontram os aglomerados de galáxias, que são resultantes de um processo de agrupamento de galáxias, ou pequenos grupos delas, numa escala de alguns bilhões de anos (Dodelson 2003 e Ryden 2016). Aglomerados são considerados de extrema importância devido a sua formação estar diretamente ligada a estruturas de grande escala, como filamentos e superaglomerados, podendo assim fornecer informações de extrema importância para o entendimento dessas estruturas cósmicas. A taxa de formação desses sistemas ao longo do tempo também pode fornecer importantes informações sobre a cosmologia vigente no universo (vide Ryden 2016). A dinâmica de um aglomerado é dominada basicamente pela matéria escura (correspondendo a $\sim$80% de toda a massa do sistema), o que torna possível estudar as suas demais componentes (gás quente emissor de raios-X e galáxias) em resposta a um mesmo potencial gravitacional (vide Padmanabhan 1993). Neste trabalho, estaremos interessados na componente formada por galáxias.

Do ponto de vista observacional, um aglomerado é um objeto que deve ser primeiramente identificado em um catálogo de galáxias. Seu centro, seus membros e sua fronteira não são conhecidos a *priori*. São os variados métodos utilizados pelos astrônomos que permitem localizar e delimitar no espaço este tipo de sistema. A delimitação de um aglomerado , portanto, é o resultado da aplicação de um conjunto de técnicas que visam minimizar os erros na determinação de sua população de galáxias, assim como de suas propriedades globais. Ou seja, podemos dizer que um aglomerado observado é resultante de uma modelagem astroestatística implementada computacionalmente.

Atualmente, galáxias são observadas em grandes levantamentos conduzidos por consórcios internacionais, como por exemplo o Sloan Digital Sky Survey (SDSS). O SDSS (ir a <www.sdss.org> para informação geral) fornece dados para variadas investigações em astrofísica e corresponde a um extenso imageamento digital calibrado fotometricamente e astrometricamente de π esferorradianos [1] perto dos 30° de latitude Galáctica em cinco bandas ópticas a uma profundidade de $g' = 23^{m}$, e um estudo espectroscópico de aproximadamente 10^{6} galáxias encontradas previamente no catálogo de objetos fotométricos produzido pelo estudo de imageamento (vide York et al. 2000). Se um pesquisador deseja mapear aglomerados de galáxias em partes deste levantamento, precisará inicialmente de um código de agrupamento. O mais usado entre os astrônomos é o método conhecido como FoF (*friends-of-friends*) que baseia-se no algoritmo de linkagem simples (vide Mingoti 2005), onde o agrupamento se dá por uniões de objetos mais próximos entre si que um determinado comprimento, conhecido

[1] De acordo com a definição do dicionario Houaiss

como o comprimento de linkagem (*linking length*).

Esta etapa de identificar as regiões do céu onde provavelmente encontram-se os aglomerados não será incorporada neste trabalho, que se inicia justamente na etapa seguinte: a remoção de objetos indevidamente classificados como membros dos aglomerados pelos métodos de agrupamento. A inclusão de objetos chamados "intrusos" (`outliers`) nos aglomerados se deve ao fato de que os atributos utilizados para identificar os aglomerados levam em conta posições projetadas e uma única componente de velocidade das galáxias, aquela ao longo da linha de visada do observador. Esta incompleteza da informação espacial e da distribuição de velocidades das galáxias leva a erros na determinação da população de um aglomerado, levando à inclusão de objetos que podem estar localizados no plano da frente (`foreground`) e/ou de fundo (`background`) do céu. Numa tentativa de reduzir o grau de impureza e maximizar o índice de completeza foram desenvolvidos inúmeros métodos de remoção de galáxias intrusas.

Entre os primeiros trabalhos a identificar galáxias de campo (ou seja, que não estão nos aglomerados) como `outliers` está o de Yahil e Vidal (1977). Este autores propuseram remover iterativamente os `outliers` (galaxias mal classificadas como pertencentes ao aglomerado) que tivessem velocidades maiores do que três vezes a da dispersão de velocidades da linha de visada. Zabludoff et al. (1990), estudando o catálogo de Abell, desenvolveram um método semelhante, baseado em um histograma onde as galáxias, em cada intervalo, estivessem ordenadas em função das velocidades, e as galáxias que tivessem uma diferença de velocidade superior a $2000\ km\ s^{-1}$ em relação a seus vizinhos seriam consideradas `outliers`. Com o tempo foram surgindo métodos com diferentes abordagens (vide Wojtak et al. 2007 para uma extensa revisão sobre o assunto).

Com a evolução dos métodos de remoção de `outliers` tornou-se possível fazer o estudo dessas estruturas com uma precisão muito maior. Possivelmente, a quantidade mais importante referente a um aglomerado é a sua massa. Se conhecermos a massa de aglomerados, podemos entender melhor o processo de formação de estruturas, assim como restringir a quantidade de matéria e energia escura no universo. Mas para obter a massa de um aglomerado, a partir de sua componente de galaxias, não basta ter executado um processo eficiente de remoção de objetos intrusos – é preciso também ter algum conhecimento sobre o estado dinâmico desses sistemas. Se o sistema se encontra em equilíbrio, podemos utilizar estimadores de massa que se baseiam no teorema do virial e na análise de Jeans (vide Padmanabhan 1993). Caso não esteja em equilíbrio, técnicas alternativas devem ser empregadas, como a "gaussianização"da amostra (vide Ribeiro et al. 2011) ou o uso da cáustica (vide Serra et al. 2014).

Para acessar o estado dinâmico dos aglomerados são utilizados indicadores

indiretos. Usualmente, estes são baseados na distribuição de velocidades das galáxias dentro de uma certa distância ao centro do aglomerado. Diversos estudos mostram que devemos esperar uma distribuição de velocidades gaussiana (ou normal) quando o sistema se encontra em equilíbrio. Portanto, medir desvios de normalidade da distribuição de velocidades permite avaliar o estado dinâmico do sistema. Uma discussão sobre este tipo de análise pode ser encontrada, por exemplo, em Ribeiro et al. (2011). Adicionalmente, testes de multimodalidade também podem ser empregados. Estes visam determinar se a distribuição de velocidades é consistente ou não com uma unimodal. Recentemente, o uso de uma métrica definida no espaço de distribuições de probabilidade tem sido utilizada para medir desvios de normalidade (vide Ribeiro et al. 2013). Uma outra maneira de acessar o estado dinâmico de aglomerados é empregar testes de subestruturas que podem tanto ser 2D (levando-se em conta apenas as coordenadas de posição projetada), como 2D+1, quando se adicionam as velocidades radiais das galáxias. Uma extensa apresentação e revisão desses testes pode ser encontrada em Pinkney et al. (1996) para uma revisão.

Esta sequência de passos (remoção de `outliers` − análise dinâmica − cálculo da massa) deve ser considerada em qualquer estudo envolvendo aglomerados de galáxias. Muitos dos métodos de remoção de `outliers`, análise de subestruturas e de normalidade já possuem versões implementadas e difundidas no meio acadêmico. Contudo, devido à época em que foram desenvolvidos, sua utilização em conjunto torna-se difícil. Por exemplo, alguns deles foram implementados em diferentes versões do `Fortran`, enquanto outros deles foram implementados em C. O emprego de diferentes linguagens não apenas dificulta a sua utilização em conjunto, como também o manuseio do usuário, uma vez que linguagens como o `Fortran` deixaram de ser de amplo conhecimento para pesquisadores recém-formados. Outro problema associado ao uso de diferentes programas é a heterogeneidade dos formatos de entrada e saída dos dados, obrigando os pesquisadores a constantemente adaptarem os programas a cada conjunto de dados analisado.

Diante disto, decidimos estabelecer uma padronização desses métodos em uma linguagem atual que facilitasse o trabalho do usuário, como também o habilitasse a utilizar técnicas computacionais mais robustas como, por exemplo, a técnica do processamento paralelo, que pode ser extremamente importante para manipular as quantidades massivas de dados hoje disponíveis à comunidade acadêmica. Nosso trabalho será desenvolvido no ambiente R[2], uma conhecida linguagem de programação e ambiente computacional estatístico, contendo vários pacotes e bibliotecas integrados. O R provê uma ampla variedade de análises estatísticas (modelagem linear e não linear, testes estatísticos, análise de séries temporais, classificação, análise de agrupamentos, etc.) e

[2]R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

técnicas gráficas, além de ser altamente extensível, permitindo ao usuário desenvolver seus próprios programas. A linguagem S, desenvolvida por John Chambers, tem como objetivo facilitar e acelerar o desenvolvimento de *softwares*, e é com frequência a escolha feita para a investigação e desenvolvimento em metodologia estatística. O R, uma de suas implementações, provê uma rota de código aberto dentro deste contexto.

O R é pensado, mais do que como simples conjunto de pacotes estatísticos, como um ambiente no qual as técnicas estatísticas, algébricas e gráficas são implementadas. O R pode ser estendido (facilmente) via pacotes. Existem catorze pacotes na distribuição básica do R (**base**, **compiler**, **datasets**, **grDevices**, **graphics**, **grid**, **methods**, **parallel**, **splines**, **stats**, **stats4**, **tcltk**, **tools** e **utils**) e centenas de outros estão disponíveis através do repositório CRAN (*Comprehensive* R *Archive Network*) a partir de *sites* na internet cobrindo uma faixa muito ampla da estatística moderna.

1.1 Objetivo Geral

O principal objetivo deste projeto é desenvolver um pacote em R que (i) otimize a análise como um todo: modificando a linguagem de programação de alguns métodos; (ii) flexibilize o uso: permitindo aos usuários trabalhar com os métodos diferentes de forma complementar; (iii) integre as funções: tornando possível a união de métodos com diferentes objetivos e fazendo com que trabalhem em sequência; e (iv) implemente métodos que ainda não foram previamente implementados ou que possam ser implementados de forma mais eficiente. O pacote permitirá ao usuário realizar estudos sistemáticos sobre aglomerados de galáxias de maneira rápida e simples, produzindo resultados homogêneos e diretamente comparáveis entre si.

1.2 Objetivos Específicos

- Desenvolver, implementar e traduzir uma série de métodos que sejam utilizados para diminuir a contaminação dos dados de entrada através do processo de retirada de `outliers`, assim como desenvolver uma função que permita utilizar os diferentes métodos de remoção de `outliers` em conjunto, dando ao usuário um maior leque de possibilidades na tentativa de flexibilizar ao máximo esta etapa da análise;

- Desenvolver e implementar algoritmos para que se possa fazer uma análise do estado dinâmico do sistema em estudo utilizando métodos que possam ser integrados às outras etapas

- Implementar diferentes métodos de obtenção da massa e do raio dos aglomerados

- Realizar testes que ofereçam ao usuário sugestões de uso do pacote.

O trabalho está organizado da seguinte maneira: no Capítulo 2 apresentamos os métodos e algoritmos que serão utilizados, na ordem em que são acionados dentro do `pipeline` de análise que desenvolvemos; no Capítulo 3 apresentaremos o pacote de análise, dando ênfase ao seu funcionamento e possibilidades de uso integrado; no Capítulo 4 apresentamos uma aplicação do pacote sobre um conjunto de dados simulados, visando discutir a eficiência dos métodos e funções descritos nos capítulos anteriores; finalmente, no Capítulo 5 fazemos uma discussão geral do trabalho, apresentando algumas perspectivas para desenvolvimentos futuros.

2 Dados e Métodos

Nesta seção descrevemos os dados que são utilizados neste trabalho, assim como os métodos de análise, separando-os de acordo com cada etapa do estudo. Primeiro descreveremos os métodos de remoção de `outliers`, em seguida apresentaremos os métodos de análise de normalidade da distribuição de velocidades e testes de subestruturas; finalmente, discutiremos os métodos de cálculo da obtenção de massa.

2.1 Dados

Utilizamos dados de um catálogo MOCK de aglomerados de galáxias para realizar testes de aplicação do pacote desenvolvido. Em um catálogo MOCK os dados são gerados artificialmente, de forma que sabemos a priori quem são as galáxias membro de um aglomerado, assim como sua massa. Isto possibilita a realização de testes para estabelecer a eficiência em cada uma das alternativas de análise que o pacote oferece. O catálogo utilizado neste trabalho foi gentilmente cedido pelo Dr. Gary Mamon (IAP - França), e corresponde a uma amostra composta por 947 aglomerados, cuja construção é descrita em Duarte e Mamon (2015).

A técnica usada na construção do catálogo baseia-se em modelo semi-analítico (SAM) de formação e evolução de galáxias, que foi executado nas árvores de fusão de halos extraídas da simulação cosmológica não-dissipativa do `Millennium II`[1]. Uma caixa de tamanho aproximado de aresta $L_{box} = 100h^{-1}Mpc$ (Megaparsec), com parâmetros cosmológicos $\Omega_m = 0.25, \Omega_\Lambda = 0.75, \sigma_8 = 0.9$ e massa de partículas $9.5 \times 10^6 M_\odot$ (Massa Solar $M_\odot = 1.9891 \times 10^{30}$ kg). Halos foram identificados aplicando a técnica `Friends-of-Friends (FoF)` às partículas no espaço real 3D. O catálogo foi desenvolvido para ter a mesma extensão do céu e a profundidade da maior região contínua da amostra espectroscópica do SDSS.

2.2 O que são `outliers`?

A primeira etapa dentro do `pipeline` de análise é aquela em que são eliminados os objetos considerados `outliers`, também chamados de galáxias "intrusas". Estes são objetos incorretamente identificados como membros dos aglomerados. Quando observamos um aglomerado de galáxias tentamos determinar quais dentre os objetos coletados pertencem (ou têm maior probabilidade de pertencer) ao sistema. Contudo, somos fortemente sujeitos a erro devido à dificuldade de distinguir galáxias membro

[1] Vide https://wwwmpa.mpa-garching.mpg.de/galform/virgo/millennium/

daquelas nos planos de frente (`foregroud`) e de fundo (`background`) do céu. Isto se dá em virtude de acessarmos informação contendo apenas posições projetadas no plano do céu e velocidades na linha de visada das galáxias. Na Figura 1 ilustramos a perda de informação devido à visão em projeção do objeto: embora a distância real ao centro do aglomerado seja r, o observador terá acesso a R_p.

Quando coletamos os dados para os estudos dessas estruturas, executando consultas nos bancos de dados, fazemos um primeiro corte tentando retirar o máximo possível de "contaminação" dos dados. Neste primeiro corte, selecionamos apenas objetos dentro de um certo intervalo em torno do valor do `redshift` central do aglomerado, que geralmente é associado ao pico de um histograma de `redshifts`. O termo `redshift` é usado para indicar o desvio percentual das linhas espectrais observadas em relação a um espectro de referência verificado em laboratório, sendo positivo quando está deslocado para vermelho (comprimentos de onda maiores) ou negativo quando deslocado para o azul. No caso de galáxias o `redshift` em questão é de natureza cosmológica, decorrendo da taxa de expansão do espaço-tempo do universo, e estando relacionado à velocidade de recessão das galáxias, ou seja, à velocidade da galáxia no fluxo de Hubble (vide, por exemplo, Ryden 2016).

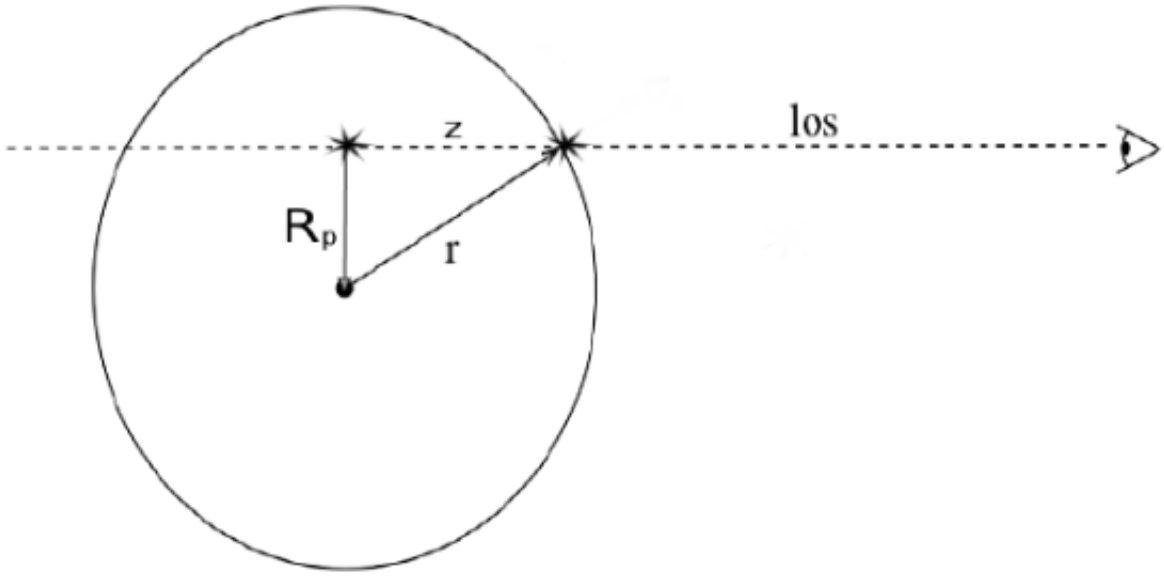

Figura 1 – Projeção de um corpo ao longo da linha de visada (los). As distâncias real (r) e projetada (R_p) são indicadas.

Com o primeiro corte, coletamos as galáxias dentro de um intervalo de velocidades radiais que tente de forma aproximada compreender todo o aglomerado (por exemplo, coletando galáxias dentro de ± 3000 km/s (± 0.01 em `redshift`) em torno do pico central). Esta coleta deve ser feita também dentro de um raio máximo em

torno de uma posição que tomamos como sendo o centro do aglomerado. Este raio em geral deve ser menor ou igual a 4 Mpc, para evitar a contaminação do sistema que se pretende estudar com outras estruturas vizinhas (Wojtak et al. 2007). Deve-se notar que a posição central de um aglomerado é de difícil determinação. Ela pode coincidir com a galáxia mais luminosa do aglomerado, ou com o pico da distribuição em raios-X, ou simplesmente corresponder ao baricentro da distribuição projetada. Neste trabalho, utilizaremos o baricentro não-ponderado como centro, por ser a escolha mais geral e de mais simples implementação que pode ser feita.

Mesmo seguindo os procedimentos descritos acima, um número variável de galáxias intrusas permanece no catálogo final do aglomerado. Na Figura 2 temos uma visão esquemática de um aglomerado. Uma escala de tamanho característica do sistema, de fundamental importância, é dada pelo seu raio virial, que corresponde ao raio onde esperamos que a distribuição de galáxias respeite o teorema do virial ($2T + V = 0$).[2] Este raio em geral é bem menor que a escala de 4 Mpc em que usualmente coletamos os dados. Notemos da figura que a projeção dos dados dentro do cone virial contém tanto galáxias da esfera virial (os membros), como ainda é possível conter galáxias intrusas. Quanto maior o cone utilizado, maior o número de intrusas. Como o raio virial é desconhecido no momento da coleta de dados, precisamos começar a análise a partir de um cone maior, o que torna necessária uma cuidadosa etapa de remoção de `outliers`.

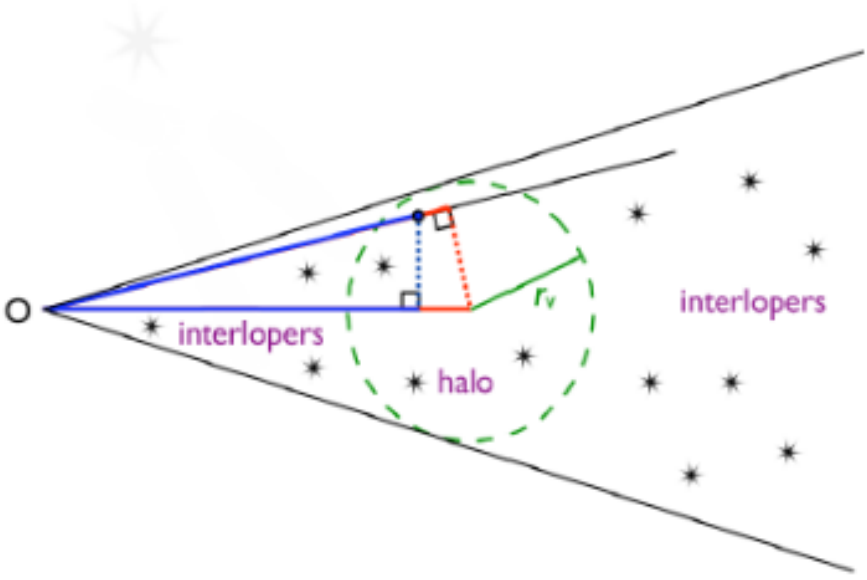

Figura 2 – Cone virial e presença de `outliers` (`Interlopers`) na região do aglomerado.

Fonte: Mamon et al. (2010)

[2] Onde T e V indicam as energias cinética e potencial, respectivamente.

Na próxima seção, descrevemos os quatro métodos utilizados neste trabalho para realizar a remoção dos `outliers`. Esses métodos foram escolhidos na tentativa de oferecer ao usuário diferentes escolhas de abordagens ao problema e, em consequência, aumentar a flexibilidade na tarefa de remover falsos membros dos aglomerados.

2.3 Métodos de remoção de `outliers`

Nesta seção, descrevemos brevemente os métodos empregados neste trabalho: Cáustica, Vmax, Gap Fixo e Gap Variável. Cada um destes métodos segue um algoritmo que identifica um `outlier` de maneira diferente e independente dos demais.

2.3.1 Cáustica

O método da Cáustica trabalha sobre a hipótese de que um aglomerado resulta de um processo de colapso esférico de matéria, seguido por uma acreção ou "queda" (o chamado `infall`) de galáxias sobre o sistema. Como descrito, por exemplo, em Svensmark et al. (2015) o movimento de `infall` produz um padrão em forma de cáustica sobre a distribuição de galáxias no espaço de fase projetado do sistema, que relaciona as distâncias projetadas às velocidades das galáxias no referencial do aglomerado.[3] Essa curva da cáustica envolve todas as galáxias para as quais o movimento de `infall` supera o movimento definido pelo fluxo de Hubble (ou seja, a expansão do universo). As curvas das cáusticas possuem uma forma similar a uma trombeta (vide Figura 3).

[3] As distâncias projetadas são calculadas a partir da separação angular entre cada galáxia e o centro do aglomerado, levando-se em conta a cosmologia utilizada.

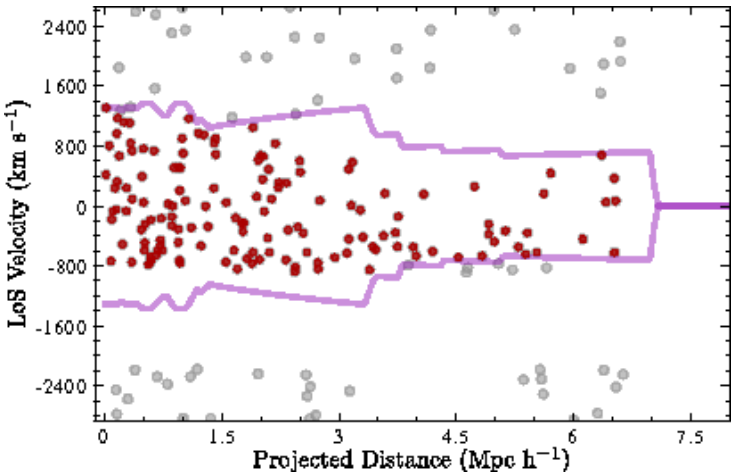

Figura 3 – O perfil da cáustica para A3888 (linhas rosas sólidas). Os pontos vermelhos mostram os membros do aglomerado definidos pela técnica da cáustica. Os pontos em cinza são as galáxia intrusas que provavelmente não são gravitacionalmente ligadas a A3888.

Fonte: Shakouri et al. (2016)

A determinação de membros e `outliers` consiste em saber se as velocidades dos objetos se encontram dentro ou fora da amplitude da cáustica para cada raio projetado. A amplitude da cáustica em um determinado raio projetado R está relacionada com a velocidade de escape do aglomerado da seguinte forma

$$\langle v^2_{esc}\rangle_{(R,\kappa)} = \int_0^R A^2_\kappa(r)\phi(r)dr \;/\; \int_0^R \phi(r)dr, \tag{1}$$

onde, $\phi(r) = \int \hat{f}(r,v)dv$. A distribuição de objetos no espaço de fase é dada pelo estimador $\hat{f}(r,v)$, sendo associada às amplitudes da cáustica contornos com $\hat{f}(r,v) = \kappa$ constantes, tal que $A(r)$ (amplitude da cáustica) seja escolhida pelo contorno $\hat{f}(r,v) = \kappa$ que minimize a função

$$S(\kappa, R) = |\langle v^2_{esc}\rangle_{R,\kappa} - 4\langle v^2\rangle_R|^2 \tag{2}$$

com o raio virial dado por $R = R_v$. Para obter-se $\langle v^2\rangle_R$ calcula-se a dispersão de velocidades média dentro de $R = R_v$. Deve-se notar que não sabemos a priori o raio virial R_v, portanto é utilizado um esquema iterativo que faz uma estimativa inicial do raio virial R_v que resulta numa M_v, massa da cáustica, que assim se converte ao novo raio virial proporcional à $M_v^{1/3}$, com o qual a técnica da cáustica pode continuar sendo aplicada de forma iterativa até que o valor da amplitude da cáustica final convirja (veja os detalhes em Alpaslan et al. 2012 e Gifford et al. 2013).

A rotina utilizada neste trabalho para a determinação da amplitude da cáustica (e posterior remoção dos `outliers`) é uma versão modificada do programa gentilmente cedido pelo Drs. Mehmet Alpaslan e Aaron Robotham (The University of Western Australia). O algoritmo básico segue os seguintes passos:

(i) Converter as posições das galáxias a partir dos `redshifts`, de acordo com a cosmologia especificada, e construir um histograma bidimensional de densidades numéricas de galáxias no espaço de fase do aglomerado.

(ii) Calcular $\hat{f}(r, v)$ para diferentes valores de κ.

(iii) Minimizar a função dada pela Eq. 2.

(iv) Ajustar a cáustica ao longo de R.

(v) Gerar as listas de membros e `outliers` ao final do processo.

2.3.2 Vmax

Nesta abordagem considera-se uma galáxia sendo um `outlier` caso ela exceda uma velocidade máxima para um determinado raio em que ela se encontra. Aparentemente, o método segue um conceito semelhante ao da cáustica, mas é mais simples computacionalmente. A ideia central do método é descrita a seguir.

Dado que conhecemos apenas distâncias projetadas ao centro do aglomerado, as galáxias podem estar em qualquer lugar sobre uma linha vertical numa distância projetada particular, dentro de um determinado círculo que defina o aglomerado (vide Figura 4). A cada galáxia do aglomerado deve ser atribuída uma velocidade limite. Neste método consideram-se duas atribuições, a velocidade circular e a velocidade de "queda" (`infall velocity`) das galáxias.

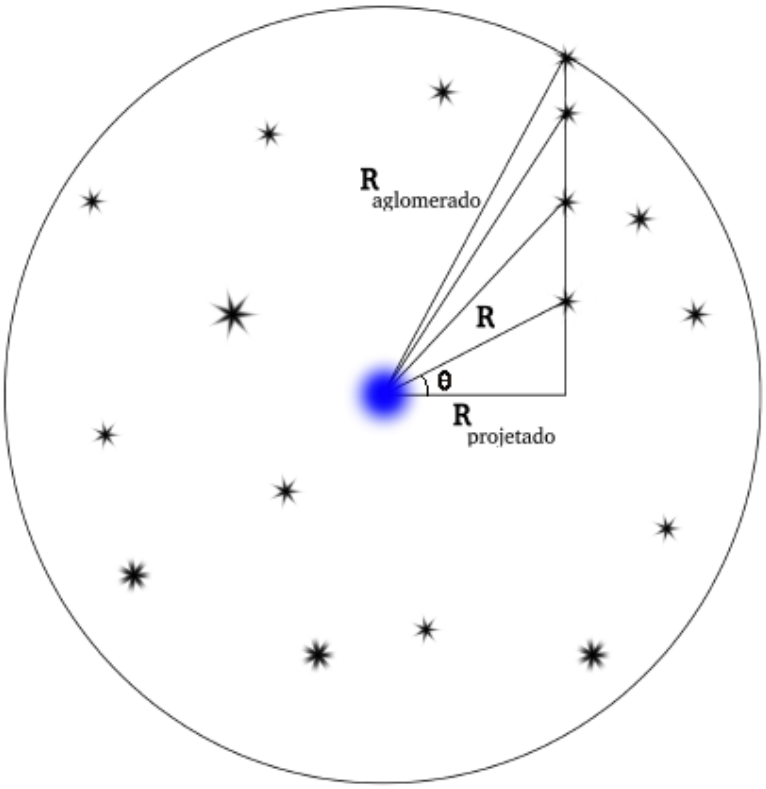

Figura 4 – Problema na determinação da posição das galáxias.

A partir dos teoremas de Newton da gravitação, sabemos que a atração gravitacional de uma distribuição de matéria dada por $\rho(r)$ sobre uma massa teste em um raio r é inteiramente determinada pela massa interior a r. Logo, para a velocidade circular, definida como a velocidade que uma partícula teste teria em uma órbita circular a uma distância r do centro, temos

$$v_{cir} = \sqrt{GM(R)/r}. \tag{3}$$

Assumimos órbitas circulares para as galáxias e a relação entre a energia gravitacional e cinética, postulada pelo teorema de virial $2T_{cir} = -V_{cir}$ (vide apêndice A). Ao mesmo tempo, velocidade de "queda" é definida por $T_{inf} = -V$, que define a separação entre galáxias ligadas e não-ligadas ao aglomerado. Neste limite temos

$$v_{inf} = \sqrt{2}v_{cir} \tag{4}$$

definimos a velocidade de queda como o limite superior para as velocidades das galáxias para as quais o teorema de virial é violado (Den Hartog e Katgert 1996 e Beers et al.

1982). Este limite tem origem no fato de que uma galáxia está ligada ao seu aglomerado se $T + V < 0$.

Sendo assim, den Hartog e Katgert (1996) e Wojtak et al. (2007) propuseram duas fórmulas para o perfil de máxima velocidade. A primeira, assumindo que a direção da velocidade da partícula no limite determinado por v_{inf} tem qualquer orientação

$$v_{max} = max_R(v_{inf}) \tag{5}$$

A segunda, com intuito de diminuir ao máximo a contaminação do aglomerado por `outliers`, utiliza um critério mais restritivo que nos dá limites mais precisos em regiões onde $R \sim r_v$ (r_v equivale ao raio virial do sistema).

$$v_{max} = max_R(v_{inf} cos\theta, v_{cir} sen\theta), \tag{6}$$

onde θ é o ângulo entre o vetor posição da partícula em relação ao centro do aglomerado e a linha de visada. Esta segunda versão do método é a que foi implementada neste trabalho. O algoritmo segue os seguintes passos:

(i) Define-se um raio máximo do aglomerado, R_{max}, dado pela distância à galáxia com maior raio projetado.

(ii) A linha vertical começando em cada raio projetado e terminando em R_{max} é incrementada em pequenos passos (aqui dados por 0.1 R_{max}). Os ângulos θ em cada passo são calculados (vide Figura 4).

(iii) Em seguida, são calculadas v_{cir} e v_{inf}, e é feita a maximização dada pela Eq. 6.

(iv) Galáxias com velocidades maiores que v_{max} são consideradas `outliers`, as demais serão consideradas membros, sendo escritas em arquivos de saída correspondentes.

Note que para v_{cir} e v_{inf} serem calculadas necessitamos definir um perfil de massa $M(r)$, ou seja, ordenando-se os dados em distância projetada ao centro, calculamos a massa virial correspondente ao raio de cada galáxia, procedendo o cálculo do centro para a borda. A massa virial é discutida na Seção 2.5.1.

2.3.3 Gap Fixo

Esta técnica foi introduzida por Fadda et al. (1996) e é também chamada de "`shifting gapper`". Esse procedimento funciona utilizando uma quantidade chamada `gap`. Ela se baseia em separar a amostra em intervalos (`bins`) de velocidade

para dados ordenados em distância projetada com relação ao centro do aglomerado. O tamanho de cada `bin` é uma escolha do pesquisador, mas Lopes et al. (2009) recomendam o uso de $0.4h^{-1}Mpc$ ($0.6Mpc$ para $h = 0.72$) ou maior para garantir que cada `bin` contenha ao menos 15 galáxias. Dentro de cada `bin` as galáxias são ordenadas em velocidades peculiares, ou seja, velocidades em relação ao centro do aglomerado, que são definidas por

$$v^i_{pec} = c(z_i - \bar{z})/(1 + \bar{z}), \tag{7}$$

onde v^i_{pec} é a velocidade peculiar da galáxia i, z_i é o `redshift` da galáxia i e $\bar{z}$ é o `redshift` médio do aglomerado.

Ao iniciar o procedimento, definimos um tamanho de "gap"máximo, ou seja, uma diferença de velocidade aceitável entre duas galáxias consecutivas. Então executamos a técnica do "`shifting gapper`" para as galáxias dentro do `bin`. Serão consideradas galáxias "intrusas"ao aglomerado aquelas que tiverem a diferença de velocidade maior do que 350 km/s (ou algum outro valor do gap definido pelo usuário) entre uma das suas vizinhas. Estas são consideradas como não pertencentes ao aglomerado, e partimos para o segundo `bin`, e assim sucessivamente. Ao concluir todos os `bins` retiramos todas as galáxias "intrusas" e recomeçamos o procedimento novamente, desde a separação do aglomerado em `bins` de intervalo de distância podemos observar esse processo pela Figura 5 .

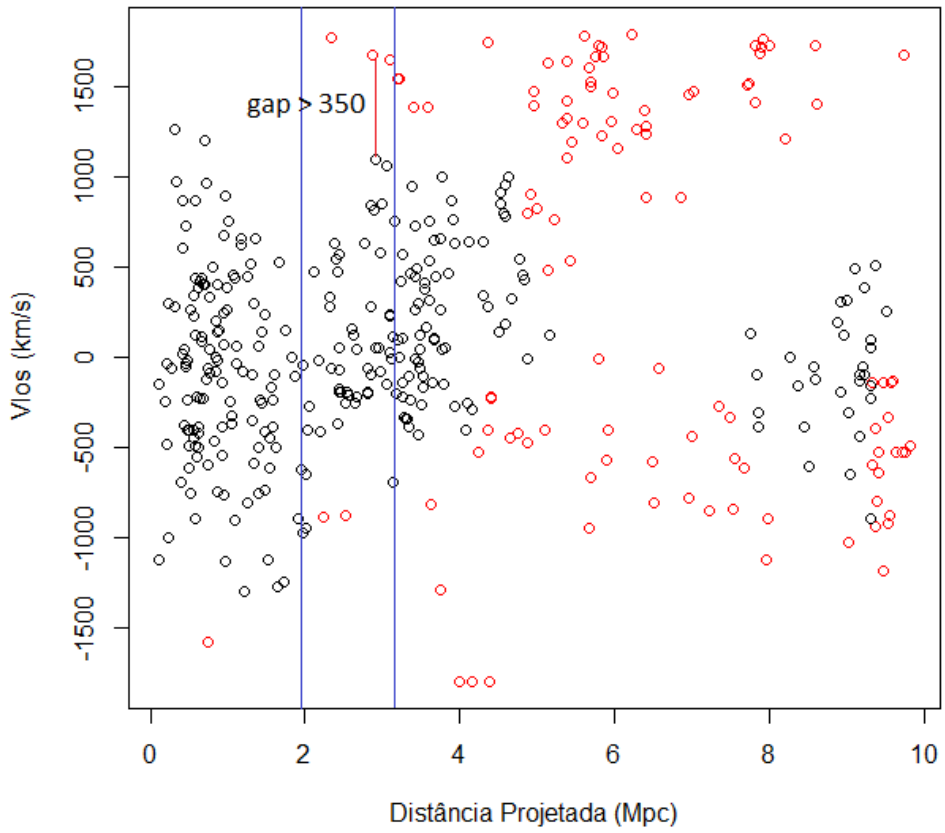

Figura 5 – Exemplo de uma execução do método do `shifting gapper` evidenciando a diferença entre duas galáxias em um `bin` especifico. O `bin` é indicado por linhas verticais azuis. Os pontos vermelhos são os `outliers`, enquanto os pretos são os membros ao final do processo de remoção.

As galáxias não associadas ao corpo principal do aglomerado são eliminadas da amostra, esse procedimento é repetido inúmeras vezes até que o número de membros se mantenha estável. É importante notar que este método não faz hipóteses sobre o estado dinâmico do aglomerado, enquanto o VMAX, por exemplo, baseia-se em suposições físicas sobre o perfil de massa do aglomerado.

2.3.4 Gap Variável

O `gap` variável é um método que pode ser considerado uma variante da técnica de "`shifting gapper`", sendo a diferença fundamental de que neste caso o tamanho do `gap` depende da própria distribuição de velocidades em cada `bin`, veja os detalhes em Beers et al. (1990).

Em cada bin, a um parâmetro 'f-pseudosigma' (vide Beers et al. 1990) que é determinado e usado como `gap` de velocidade para rejeitar `outliers`. O valor do

`f-pseudosigma` (S_f) corresponde à diferença normalizada entre os quartis superior (F_u) e inferior (F_i) de um conjunto de dados. Ele é calculado da seguinte forma:

$$S_f = (F_u - F_i)/1.349. \tag{8}$$

A constante 1.349 é a diferença esperada para $(F_u - F_i)$ se os dados têm distribuição de velocidade normal em cada `bin` (vide Beers et al. 1990). O procedimento de remoção segue exatamente o algoritmo do caso anterior, sendo repetido até que o número de membros se estabilize, ou o valor de `f-pseudosigma` caia abaixo de 250 km/s, ou o valor de `f-pseudosigma` comece a aumentar (vide Wing e Blanton 2013), podemos visualizar melhor o processo com a ajuda da Figura 6.

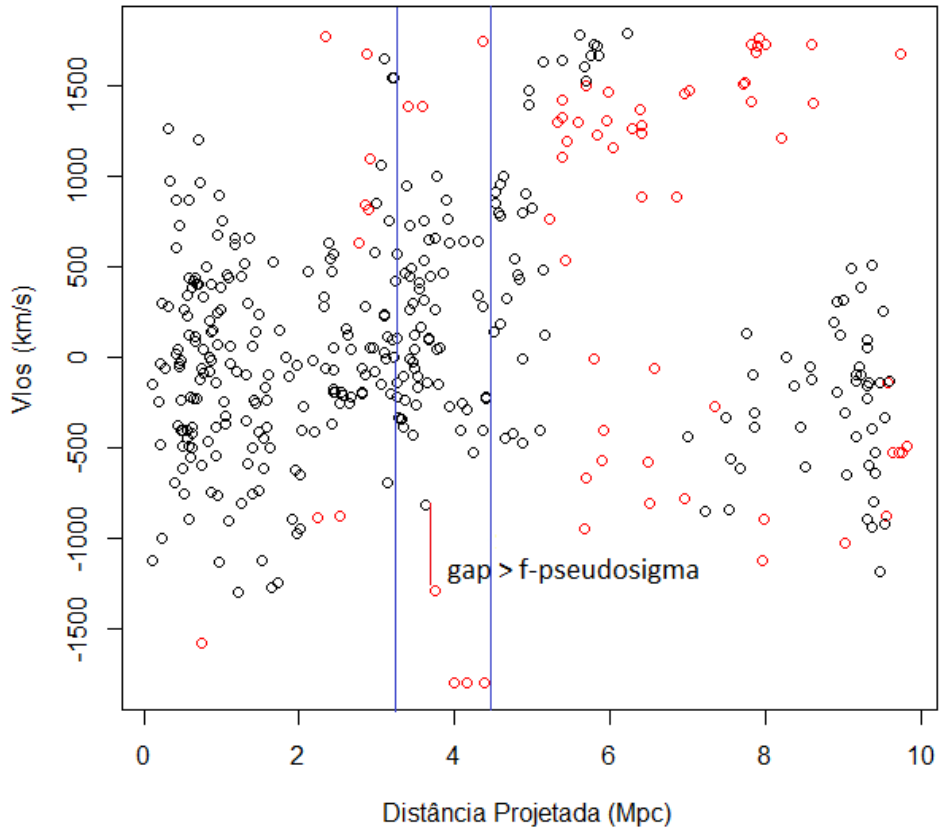

Figura 6 – Exemplo de uma execução do método do `shifting gapper` evidenciando a diferença entre duas galáxias em um `bin` específico só que no caso do `gap` variável a medida de corte entre as galáxias dentro do `bin` é determinada pelo cálculo do `f-pseudosigma`.

2.4 Análise Dinâmica

Uma vez que a remoção de `outliers` tenha sido feita através de um dos métodos descritos na seção anterior, o próximo passo é a realização da análise dinâmica do

aglomerado. Esta deve ser entendida como uma inferência indireta sobre o equilíbrio da componente de galáxias no potencial do aglomerado. Basicamente, dois aspectos são considerados: (i) a distribuição de velocidades das galáxias dentro de um certo raio do sistema; (ii) e a presença ou não de subestruturas, considerando-se desvios cinemáticos locais na distribuição de galáxias.

Consideramos um aglomerado em equilíbrio se a sua distribuição de velocidades das galáxias for consistente com uma distribuição normal, também conhecida como gaussiana. A razão para isto remonta aos estudos fundamentais de Ogorodnikov (1957) e Lynden-Bell (1967) que indicam que a distribuição de velocidades de partículas em um sistema gravitacional é consistente com a distribuição de Maxwell-Boltzmann que, considerada em coordenadas do espaço de fase do sistema, equivale a uma distribuição gaussiana. Experimentos numéricos dão suporte a este resultado teórico (vide Merral e Henriksen 2003; Hansen et al. 2005). Contudo, estudos recentes revelam a dificuldade, do ponto de vista observacional, em se determinar se a distribuição de velocidades de galáxias em aglomerados é ou não normal (vide Hou et al. 2009; Ribeiro et al. 2011; Ribeiro et al. 2013). Neste trabalho, utilizaremos alguns métodos descritos por esses autores.

Adicionalmente, considera-se um aglomerado em equilíbrio se ele não possui subestruturas, ou seja, sub-aglomerações internas. Muitos métodos têm sido utilizados para a determinação de subestruturas em aglomerados (vide uma abrangente revisão em Pinkney et al. 1996). Neste trabalho consideraremos apenas um deles, chamado "Delta de Dressler & Schectman" (vide Dressler e Shectman 1988), que mede o desvio cinemático local de grupos de galáxias em um aglomerado. Este método é o mais extensivamente usado pelos astrofísicos e o que ostenta melhores resultados nas comparações realizadas por Pinkney et al. (1996) e Knebe e Müller (2000).

A seguir, descrevemos os métodos que foram implementados no presente trabalho. Verificar se o sistema está ou não em equilíbrio só faz sentido dentro de um raio em que se espere de fato a virialização. Por esta razão, todos os métodos abaixo são aplicados dentro de um raio estimado, dado pela média harmônica dos raios projetados de todos os objetos, considerando massas iguais e órbitas isotrópicas.

2.4.1 GNG

A rotina GNG (Gaussiano / Não Gaussiano) reúne dois métodos recentemente introduzidos na astrofísica: (i) a distância de Hellinger, para medir desvios de gaussianidade (vide Ribeiro et al. 2013) ; e (ii) o MCLUST (Model-based clustering), que estuda a multimodalidade da distribuição de velocidades (vide Einasto et al. 2012 e Ribeiro et al. 2013). Ribeiro et al. (2013) utilizam essas duas técnicas para chegar a um diagnóstico sobre a distribuição de velocidades ser ou não consistente com uma normal. O resultado

de cada método pode ser combinado para o diagnóstico final. A seguir, descrevemos cada um desses métodos.

Distância de Hellinger

A Distância de `Hellinger` (ou HD) é uma aproximação estável para a matriz de informação de Fisher (ver, por exemplo, Amari 1985). Em nosso caso, utilizamos essa métrica para obter o quão distante é a distribuição de velocidades observada do aglomerado de uma distribuição gaussiana. Para um espaço discreto, a distância de `Hellinger` funciona da seguinte forma,

$$HD^2(p,q) = 2\sum_{x\in X}[\sqrt{p(x)} - \sqrt{q(x)}]^2 \tag{9}$$

Onde p e q são distribuições de probabilidade (observada e teórica) e x é uma variável aleatória, uma vez que a comparação se dar entre dados de natureza discreta e uma distribuição teórica que é uma função continua é feita antes da subtração uma suavização da distribuição observada por um *kernel* da largura do desvio absoluto da mediana dos dados. Os valores possíveis de HD estão na faixa $[0, \sqrt{2}]$, mas foi seguido o procedimento de Le Cam (1986), normalizando-se os valores possíveis para $[0, 1]$. Os códigos que usamos para determinar o HD estão disponíveis publicamente na linguagem e ambiente R (`R Development Core Team`) sob o pacote `distrEx` (vide Ruckdeschel et al. 2006).

Para que o estimador seja independente do número de objetos na amostra foi necessário introduzir uma calibração. Foram calculadas 1000 vezes a HD para distribuições normais com diferentes números de elementos (N variando entre 7 e 1000). O valor esperado em todos os casos é 0, no entanto o efeito do tamanho da amostra produz um viés forte e sistemático (veja Figura 7). O viés pode ser visto através da curva gerada pelas medianas de HD para cada N (curva verde da Figura 7). Calculamos ainda a dispersão dada pelo desvio absoluto da mediana dos valores de HD para cada N. Definimos como tolerância um valor igual a três vezes esta dispersão (curva vermelha da Figura 7). Consideramos desvios significativos de uma gaussiana aqueles que se encontram acima da curva vermelha para um determinado tamanho de amostra N.

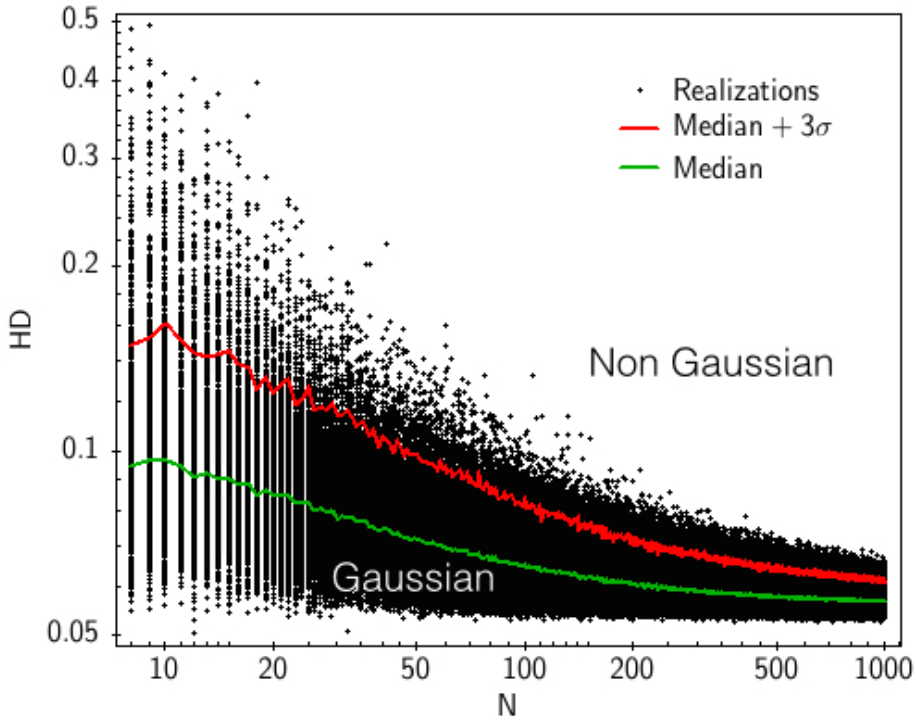

Figura 7 – Gráfico ilustrando o procedimento de calibração que é feito sobre o HD

MCLUST

O `MCLUST` (Model-based clustering) é um pacote construído em R para modelagem de misturas de gaussianas. Ele fornece funções para estimação de parâmetros através do algoritmo `Expectation-Maximization` (EM) para modelos de misturas com uma variedade de estruturas de covariância (vide Fraley e Raftery 2007), para o caso de dados multivariados. O método baseia-se na busca de um modelo ótimo para o agrupamento dos dados entre modelos com forma, orientação e volume variados. Encontra o número ótimo de componentes e a classificação correspondente (a composição de cada componente). Para o caso univariado, a mistura estima apenas os melhores parâmetros das dispersões em cada componente e o seu respectivo peso dentro da mistura. `MCLUST` associa ainda cada galáxia a uma das componentes da mistura.

Resumidamente, `MCLUST` determina se há subestruturas no espaço de velocidades em um aglomerado. Juntamente com o resultado de HD (que mede desvios gerais de uma normal) podemos determinar se o aglomerado tem uma distribuição de velocidades próximas a uma gaussiana e se ele contem subestruturas, fornecendo assim um diagnóstico sobre o estado dinâmico do aglomerado, ou seja, permitindo dizer se ele está ou não em equilíbrio. A Figura 8 ilustra os efeitos que a rotina GNG é capaz de detectar. Na parte de cima da figura, vemos à esquerda uma distribuição cujo valor de HD é 0.03, com diagnóstico de gaussianidade e de unimodalidade (dado por `MCLUST`). À direita, ainda na parte superior da figura, vemos uma distribuição que,

embora unimodal tem diagnóstico de não-gaussianidade, sendo portanto um caso onde HD e `MCLUST` divergem. Finalmente, na parte inferior da figura, vemos um caso de clara bimodalidade e de não-gaussianidade detectadas por ambos os métodos. HD, portanto, é capaz de medir desvios mais gerais da gaussianidade.

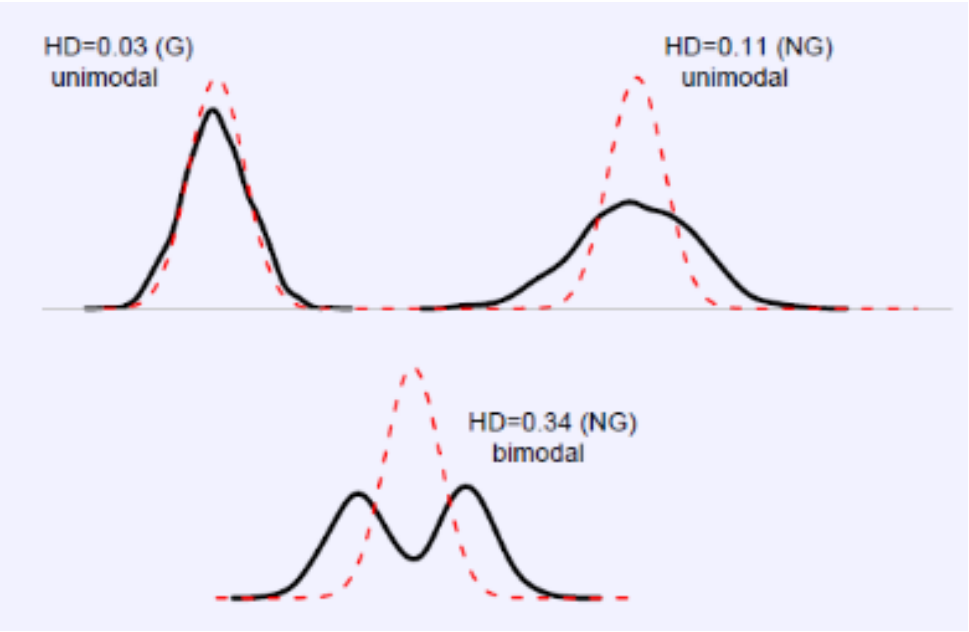

Figura 8 – Figura exemplificando a distância de `Hellinger` e `Mclust`. Em cada exemplo, mostramos a distribuição observada e a gaussiana de comparação. Os diagnósticos de HD e MCLUST também são indicados, onde G significa gaussianidade e NG não gaussianidade.

Cabe ressaltar novamente que, para cada aglomerado analisado, GNG deve ser executado dentro de um raio que esperamos equilíbrio. Efetivamente, esperamos equilíbrio apenas dentro da região virializada, o que torna necessário ter como entrada da rotina, não apenas o vetor de velocidades, mas uma estimativa do raio virial, que pode ser efetuada pelo raio harmônico do sistema.

Finalmente, a rotina GNG tem como saída não apenas os diagnósticos de HD e MCLUST, como também uma estimativa do grau de confiança em cada resultado. Esta medida de "confiabilidade" vem da execução de GNG em 1000 reamostragens do vetor de velocidades do aglomerado. A cada reamostragem, os diagnósticos são acumulados para que, ao final, possamos anexar aos diagnósticos de HD e MCLUST a confiabilidade dada pela porcentagem de casos que se repetem para o resultado mais frequente. Exemplo: HD indica gaussianidade com 75% de confiabilidade e MCLUST indica unimodalidade com 80% de confiabilidade. Isto significa que ao executar GNG sobre as 1000 reamostragens, em 750 casos HD indicou gaussianidade, enquanto MCLUST indicou unimodalidade em 800 casos. Neste trabalho, consideramos um resultado confiável, se o grau de confiança é maior ou igual a 70%.

2.4.2 GNG MIN

O GNG MIN é uma rotina em R desenvolvida integralmente neste trabalho. Ela é uma versão simplificada do GNG que usa testes estatísticos tradicionais de normalidade para avaliar a distribuição de velocidades das galáxias membro dos aglomerados. Os testes utilizados no GNG MIN são: Anderson-Darling, Jarque-Bera, Shapiro e D'Agostino (que podem ser encontrados no R sob o pacote `nortest`).

Estes testes foram escolhidos por serem de amplo uso em astrofísica. Em particular, o teste de Anderson-Darling (AD) vem sendo um dos mais utilizados após o estudo de Hou et al. (2009), em que os autores compararam diversos testes de normalidade, aplicados sobre dados controlados, e verificaram que o teste AD é o mais adequado para análise de distribuição de velocidades de galáxias, com performance significativamente superior aos demais. Basicamente, a rotina GNG MIN foi implementada para levar em conta o diagnóstico deste teste especificamente.

Em seguida, fazemos uma breve descrição dos métodos seguindo Lucambio (2008) e Kanji (2006)

Anderson-Darling

O teste de Anderson-Darling é definido pela estatística A:

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}[2i-1][ln(p_{(i)}) + ln(1-p_{(n-i+1)})], \tag{10}$$

onde $p_{(i)} = \Phi([x_{(i)} - \bar{x}]/\sigma)$ é a função de distribuição cumulativa da distribuição normal padronizada, e $\bar{x}$ e σ são a média e o desvio padrão, respectivamente. O valor-p é obtido da estatística modificada de $Z = (1.0 + 0.75/n + 2.25/n^2)A$ (vide (Lucambio, 2008)).

Jarque-Bera

Proposto por Bera & Jarque (1980), baseia-se na diferença entre os coeficientes de skewness e kurtosis dos dados $y_1, y_2, ..., y_n$ e àquelas da distribuição assumida normal.

As hipóteses nula e alternativa no teste Jarque-Bera são:

$$H_0 : y_1, y_2, ..., y_n \sim N(\mu, \sigma^2) \;\text{ vs }\; H_1 : \;\text{ não }\; H_0, \tag{11}$$

A estatística de teste é

$$JB = n(\frac{\alpha_3^2}{6} + \frac{(\alpha_4 - 3)^2}{24}), \tag{12}$$

onde

$$\alpha_3 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^3}{ns^3}, \tag{13}$$

$$\alpha_4 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^4}{ns^4}, \tag{14}$$

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{\hat{y}})^2}{n}, \tag{15}$$

Aqui, $\bar{y}$ é a média amostral e s^2, α_3 e α_4 o segundo, terceiro e quarto momentos centrais, respectivamente. A estatística JB têm distribuição assintótica $X^2(2)$ sob a hipótese nula.

Shapiro

Proposto por Shapiro & Wilk (1965) utiliza a estatística

$$W = \frac{(\sum_{i=1}^{n} a_i y_{(i)})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{16}$$

onde as constantes $a_1, a_2, ..., a_n$ são calculadas como a solução de

$$(a_1, a_2, ..., a_n) = \frac{m^\top V^{-1}}{(m^\top V^{-1} V^{-1} m)^{1/2}}, \tag{17}$$

sendo $m = (m_1, m_2, ..., m_n)^\top$ o vetor dos valores esperados das estatísticas de ordem da amostra e V a matriz de covariâncias dessas estatísticas. O p-valor deste teste é calculado exatamente para $n = 3$, em outras situações utilizam-se aproximações diferentes para $4 \leq n \leq 11$ e para $n \geq 12$, (Shapiro & Francia, 1972).

D'Agostino

Também conhecido como teste D foi proposto por D'Agostino (1970) e têm sido muito utilizado para verificar normalidade. Suponha que $y_1, y_2, ..., y_n$ é a amostra aleatória e que $y_(1), y_(2), ..., y_(n)$ é a amostra ordenada, isto é, $y_(1) \leq y_(2) \leq ... \leq y_(n)$. A estatística D de teste é

$$D = \frac{T}{n^2 s}, \tag{18}$$

onde s é o desvio padrão amostral, o que é calculado como a raiz quadrado positiva de s^2 segundo definido no contexto do teste Jarque-Bera e

$$T = \sum_{i=1}^{n}(i - \frac{n+1}{2}) y_(i), \tag{19}$$

Se a amostra é da distribuição normal, temos que

$$E\{D\} = \frac{(n-1)\Gamma(\frac{n}{2}-\frac{1}{2})}{2\sqrt{2n\pi\Gamma(\frac{n}{2})}} \approx \frac{1}{2\sqrt{\pi}} \approx 0.28209479, \tag{20}$$

o desvio padrão assintótico da estatística D é

$$s\{D\} = \sqrt{\frac{12\sqrt{3}-37+2\pi}{24n\pi}} \approx \frac{0.02998598}{\sqrt{n}}, \tag{21}$$

Utiliza-se a estatística D padronizada como

$$D* = \frac{D - E\{D\}}{s\{D\}}, \tag{22}$$

a qual têm distribuição normal aproximada sob hipótese nula. No R as funções `ad.test` no pacote `nortest` fornece a estatística dos testes acima assim como o valor-p e a função `adTest` no pacote `fBasics`.

Como na rotina GNG, em GNG MIN a cada um dos testes foi associada uma confiabilidade a partir de N reamostragens dos dados, definindo-se 70% como um valor mínimo de confiança para um diagnóstico final.

2.4.3 DS

O DS é um teste desenvolvido por Dressler e Shectman (1988). Este teste atua sobre as posições e velocidades das galáxias de um aglomerado e tem como objetivo identificar subestruturas. Ele utiliza as velocidades radiais para verificar se existem diferenças cinemáticas significativas ao longo do aglomerado. O teste opera da seguinte maneira, para cada galáxia selecionam-se as suas $\sqrt{N}$ vizinhas mais próximas (onde N é o número de galáxias do aglomerado). A partir desta amostra de $\sqrt{N}$, calcula-se a média e a dispersão de velocidades local e a comparamos com a média global e a dispersão determinada para toda a amostra do aglomerado. É definido o desvio cinemático δ como

$$\delta^2 = (\sqrt{N}/\sigma^2)((\bar{v}_{local} - \bar{v})^2 + (\sigma_{local} - \sigma)^2). \tag{23}$$

Também é definido um desvio cinemático cumulativo, Δ, que é a soma dos δ individuais para todos os membros do aglomerado. Se a distribuição das velocidades do aglomerado estiver próxima da gaussiana e as variações locais forem apenas flutuações aleatórias, então Δ será da ordem (ou menor) N. Usualmente, define-se $\Delta/N > 1.4$ como indicador de subestruturas (vide Knebe e Muller 1999). Contudo, como não há

garantia de que a distribuição de velocidades será gaussiana (e de fato em geral não será se o subaglomerado for importante), a estatística pode variar significativamente em torno de N, mesmo se não houver subaglomerado genuíno. Reamostragens da distribuição de velocidade podem ser usadas para associar um grau de confiança ao resultado. Executando 1000 reamostragens podemos calcular a fração de casos para os quais $\Delta > 1.4$. Neste trabalho, consideramos confiável um resultado com confiabilidade maior ou igual a 70%. A Figura 9 ilustra o procedimento seguido pelo teste DS através do chamado "gráfico de bolhas", onde cada bolha tem tamanho proporcional ao δ local.

A estatística Δ é insensível, obviamente, em situações onde os subaglomerados estão bem sobrepostos no plano do céu. Ela depende de algum deslocamento dos centróides dos subaglomerados.

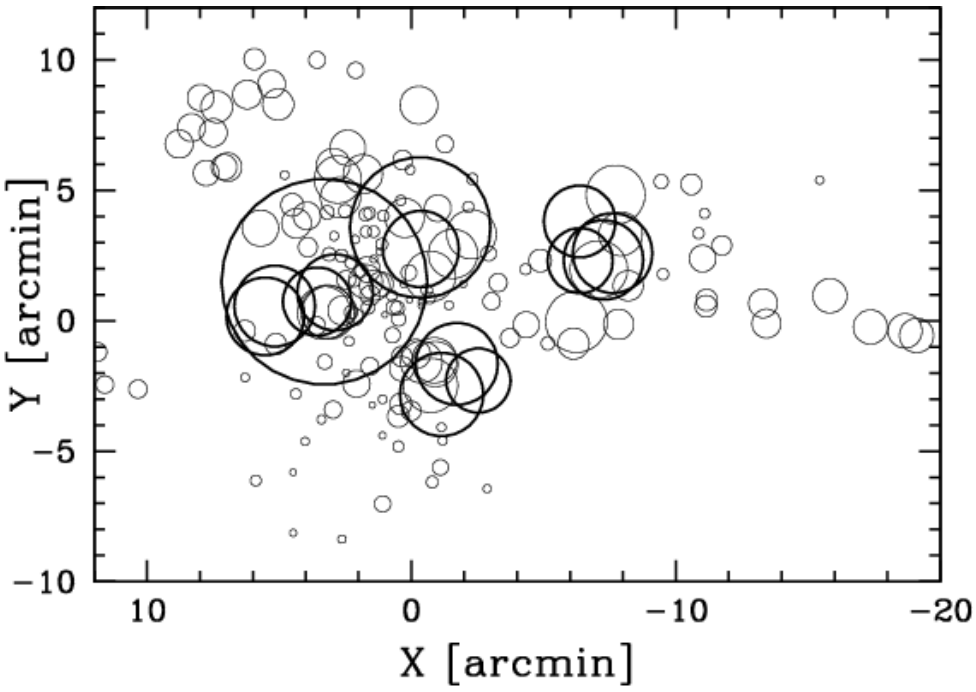

Figura 9 – Distribuição espacial dos 167 membros do aglomerado Abell 520, cada um marcado por um círculo: quanto maior o círculo, maior é o desvio δ_i dos parâmetros locais dos parâmetros do `cluster` global, ou seja, há mais evidências para a subestrutura Dressler & Shectman. Círculos grifados indicam aqueles com $\delta_i \geq 2.5$. Nos eixos arcmin sendo minuto de arco.

Fonte: Girardi et al. (2008)

2.5 Estimativa de massa

Após a remoção dos `outliers` de cada aglomerado, e de saber se o sistema está ou não em equilíbrio, precisamos fazer a estimativa da massa do aglomerado. Esta depende bastante do estado dinâmico do sistema, levando a métodos baseados ou não no teorema do virial (vide apêndice A). Neste trabalho, o único método não baseado no teorema do virial é desenvolvido sobre o conceito da cáustica (vide seção 2.5.5). Os demais métodos são variantes de estimadores da massa virial.

A obtenção da massa virial vem do uso do teorema do virial aplicado às posições e velocidades dos membros do aglomerado. Essa estimativa parte do pressuposto de que o aglomerado encontra-se em equilíbrio dinâmico (vide Binney e Tremaine 1987), com a população de galáxias membro em equilíbrio com o potencial do aglomerado, e que a distribuição da massa (total) segue a distribuição das galáxias observadas (vide Biviano et al. 1993). O teorema do virial é uma relação entre as energias cinética e potencial do sistema: $2T + V = 0$, que nos leva a (vide, por exemplo, Schneider et al. 2006)

$$M_V = \frac{R_G \langle v^2 \rangle}{G}, \tag{24}$$

onde G é a constante gravitacional, R_G é o raio gravitacional, e $\langle v^2 \rangle$ é a dispersão de velocidades ponderada do sistema (vide, por exemplo, Schneider et al. 2006). Os estimadores baseados no teorema do virial são diferentes maneiras de resolver esta equação com os dados observados.

2.5.1 Massa Virial

Para um sistema de N galáxias com separações projetadas entre elas denotadas por R_{ij} e onde v_{zi} é a componente de velocidade ao longo da linha de visada da galáxia i em relação ao centro de massa, assumindo simetria esférica e equilíbrio virial, a massa do sistema pode ser estimada por,

$$M_v = \frac{3\pi}{G} \sigma_z^2 R_H \tag{25}$$

(vide Heisler et al. 1985), onde σ_z representa o valor da dispersão de velocidades ao longo da linha de visada e R_h é o raio harmônico médio projetado. Estas quantidades são formalmente definidas por,

$$R_H^{-1} = 2 \sum_{i>j} \frac{m_i m_j}{R_{ij}} / (\sum_i m_i)^2 \tag{26}$$

$$\sigma_z^2 = \frac{\sum_i m_i v_{zi}^2}{\sum_i m_i} \tag{27}$$

Em geral, desconhecemos as massas individuais das galáxias, assumindo-se neste caso massas iguais $m_i = 1$.

2.5.2 Massa projetada

A massa projetada foi definida por Heisler et al. (1985) pela expressão

$$M_p = \frac{f_p}{GN} \sum_i v_{zi}^2 R_i, \tag{28}$$

onde f_p é um fator numérico que depende da distribuição das órbitas das galáxias ao redor do centro de massa. Verificou-se que, sob escolhas determinadas da geometria do aglomerado, dos perfis de densidade e dos tipos de galaxias que habitam o aglomerado, é possível dar uma forma explícita a este fator. Na verdade, vamos supor que o sistema seja esfericamente simétrico e, portanto, a condição de equilíbrio hidrostático do Jeans se aplica,

$$\frac{d}{dr}(\rho(r)\sigma_r^2(r)) + 2\frac{\rho(r)}{r}(\sigma_r^2(r) - \sigma_t^2(r)) = -\rho(r)\frac{d}{dr}\phi(r). \tag{29}$$

Nesta equação, $\rho(r)$ é a densidade de massa, $\sigma_r(r)$ e $\sigma_t(r)$ são os componentes radial e tangencial da dispersão de velocidades, e $\phi(r)$ é o potencial gravitacional, definindo-se ainda o parâmetro de anisotropia $\beta(r)$ como,

$$\beta(r) = 1 - \frac{\sigma_t^2(r)}{\sigma_r^2(r)}. \tag{30}$$

Agora, podemos multiplicar a Eq. 29 por r^4 e integrar sobre r:

$$\int \rho(r)\sigma_r^2 r^3(2\beta(r) - 4)dr = -\int \rho(r) r^4 \frac{d}{dr}\phi(r)dr. \tag{31}$$

Tendo em conta que σ_z, a dispersão de velocidades ao longo da linha de visão, é dada por $\sigma_z^2 = \sigma_r^2 \cos^2\theta + \sigma_t^2 \,\mathrm{sen}^2\theta$, onde θ é o ângulo de inclinação em relação ao plano do céu, e que a separação projetada R é $R = r\,\mathrm{sen}\theta$, pode-se obter

$$\langle v_z^2 R\rangle = \frac{1}{M}\int \rho(r)\sigma_z^2 R dr \tag{32}$$

$$\langle v_z^2 R\rangle = \frac{2\pi}{M}\int\int \rho(r)(\sigma_r^2(r) - \beta(r)\sigma_r^2 sen^2\theta) r^3 sen^2\theta d\theta dr \tag{33}$$

$$\langle v_z^2 R\rangle = \frac{\pi^2}{4M}\int \rho(r)\sigma_r^2(r)r^3(4-3\beta(r))dr. \tag{34}$$

Supondo β constante (o que nem sempre é bem justificado) e tendo em conta as relações entre a massa, a densidade e o potencial, encontra-se finalmente,

$$M_p = \frac{32}{\pi G}\frac{4-2\beta}{4-3\beta}\langle v_z^2 R\rangle. \tag{35}$$

Fica explícito na Eq. 35 que M_p depende criticamente do fator de anisotropia e pode mudar dependendo dos tipos dominantes de órbitas. Assim, o valor de M_p para órbitas radiais, isotrópicas e circulares, respectivamente, é dado por

$$M_p = \frac{64}{\pi G}\langle v_z^2 R\rangle \;\; (\beta = 1; \sigma_t = 0) \tag{36}$$

$$M_p = \frac{64}{2\pi G}\langle v_z^2 R\rangle \;\; (\beta = 0; \sigma_t = \sigma_r) \tag{37}$$

$$M_p = \frac{64}{3\pi G}\langle v_z^2 R\rangle \;\; (\beta \to -\infty; \sigma_t \to 0). \tag{38}$$

2.5.3 Massa Mediana

No caso de partículas (galáxias) de massas iguais, é possível aplicar a massa mediana. Este estimador é menos sensível a objetos intrusos e sua expressão é dada por

$$M_M = \frac{6.5}{G} Median_{i,j}[(v_{zi} - v_{zj})^2 R_{ij}]. \tag{39}$$

(vide Heisler et al. 1985).

2.5.4 M_{200}

De acordo com o modelo de colapso esférico, modelo que descreve a formação de sistemas esféricos na astronomia em diferentes escalas Padmanabhan (2002), a densidade média de matéria dentro do raio virial deve ser maior que a densidade cósmica média por um certo fator Δ_{vir}. A densidade média depende do modelo cosmológico e do `redshift` do aglomerado. Vários estudos mostram que a massa virializada de um aglomerado normalmente está contida dentro de uma superfície com densidade média igual a 200 vezes a densidade crítica do universo no `redshift` do objeto, ou seja, $\Delta_{vir} \approx 200$ (vide, por exemplo, Carlberg et al. 1997, Bartelmann et al. 2013). O raio

que delimita esta região é chamado de r_{200}. Usando o teorema do virial dentro deste raio, teremos:

$$M_{200} = \frac{3\sigma_v^2}{G} r_{200}, \tag{40}$$

onde G é a constante gravitacional e σ_v^2 é a dispersão de velocidades radiais.

Para determinarmos o r_{200}, devemos fazer a razão entre a densidade média do aglomerado e a densidade crítica do universo (considerando um universo espacialmente plano) no `redshift` do aglomerado. Referimos os detalhes matemáticos a Carlberg et al. (1997) e apresentamos o resultado final

$$r_{200} = \frac{\sqrt{3}\sigma_v}{10H(z)} \tag{41}$$

onde $H(z)$ é o parâmetro de Hubble que define a cosmologia que está sendo utilizada (vide Carlberg et al. 1997).

2.5.5 Massa da Cáustica

A distribuição no espaço de fase das galáxias de um aglomerado (velocidade na linha de visada contra a distância projetada em relação ao centro do aglomerado) apresenta um formato de trombeta (vide seção 2.3.1). Os limites dessa distribuição são chamados de cáusticas, e sua amplitude pode ser relacionada com a velocidade de escape relativa ao poço de potencial gravitacional do aglomerado, possibilitando deste modo a se estimar a massa do sistema (vide Diaferio et al. 2005). Sendo $A(R)$ a amplitude da cáustica em cada raio R, o perfil de massa cumulativo pode ser estimado através da seguinte equação (vide Diaferio 1999):

$$GM(< r) = \frac{1}{2}\int_0^r A^2(R)\, dR. \tag{42}$$

2.6 Consideração final

Os métodos apresentados neste capítulo permitem ao leitor (e ao potencial usuário do código) um melhor entendimento do que está sendo acionado a cada etapa de análise dos dados dentro do `pipeline` que estamos propondo. Obviamente, alguns dos métodos são mais complexos ou permitem variantes com maiores complexidades do que aquelas que foram discutidas aqui. Contudo, nosso objetivo foi dar o conjunto de informações necessário para que o programa como um todo seja compreendido. No próximo capítulo apresentaremos como o pacote funciona, de um ponto de vista mais operacional. Abordaremos como utilizar cada uma das rotinas já disponíveis, ilustrando

alguns casos, e mostrando assim como o programa será capaz de facilitar a análise de aglomerados de galáxias.

3 PACOTE

Neste capítulo, explicaremos como cada rotina, referente às diferentes etapas do pacote, foi desenvolvida, mostrando as entradas, saídas e comandos respectivos. Nosso objetivo é tornar a descrição do funcionamento do pacote clara o suficiente a qualquer usuário interessado em sua utilização.

Basicamente, o pacote oferece ao usuário uma sequência em três etapas: remoção de `outliers`, análise dinâmica, estimativas para os cálculos de massas e raios. Em cada passo desta sequência o usuário poderá escolher o método (ou mais de um método) que desejar, assim como definir os valores dos parâmetros livres associados a cada função particular.

Como ilustração, apresentamos um fluxograma (Figura. 10) contendo a ideia geral do pacote. Ao longo das próximas seções, mostraremos fluxogramas para cada etapa específica do `pipeline`.

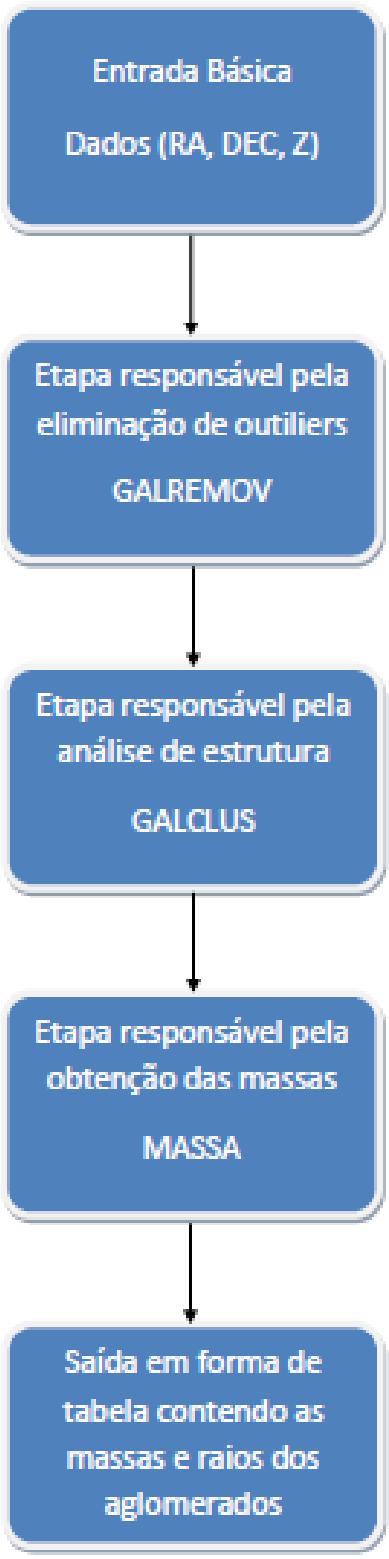

Figura 10 – Fluxograma representando o funcionamento geral do pacote.

3.1 Remoção de `outliers`

Nesta seção, descrevemos os 4 métodos que podem ser usados pela rotina `galremov` para realizar a remoção de `outliers` e, em seguida, como eles podem ser utilizados em conjunto. Na Figura 11 mostramos um fluxograma ilustrando o funcionamento geral da rotina.

Os exemplos apresentados neste capítulo referem-se a um mesmo aglomerado composto por 394 galáxias coletadas no SDSS. Este aglomerado é um dos Aglomerados de Abell, levantamento bastante conhecido e estudado na literatura astrofísica (vide Abell 1958). Na Figura 12 vemos a distribuição de galáxias neste aglomerado, assim como sua distribuição de `redshifts`.

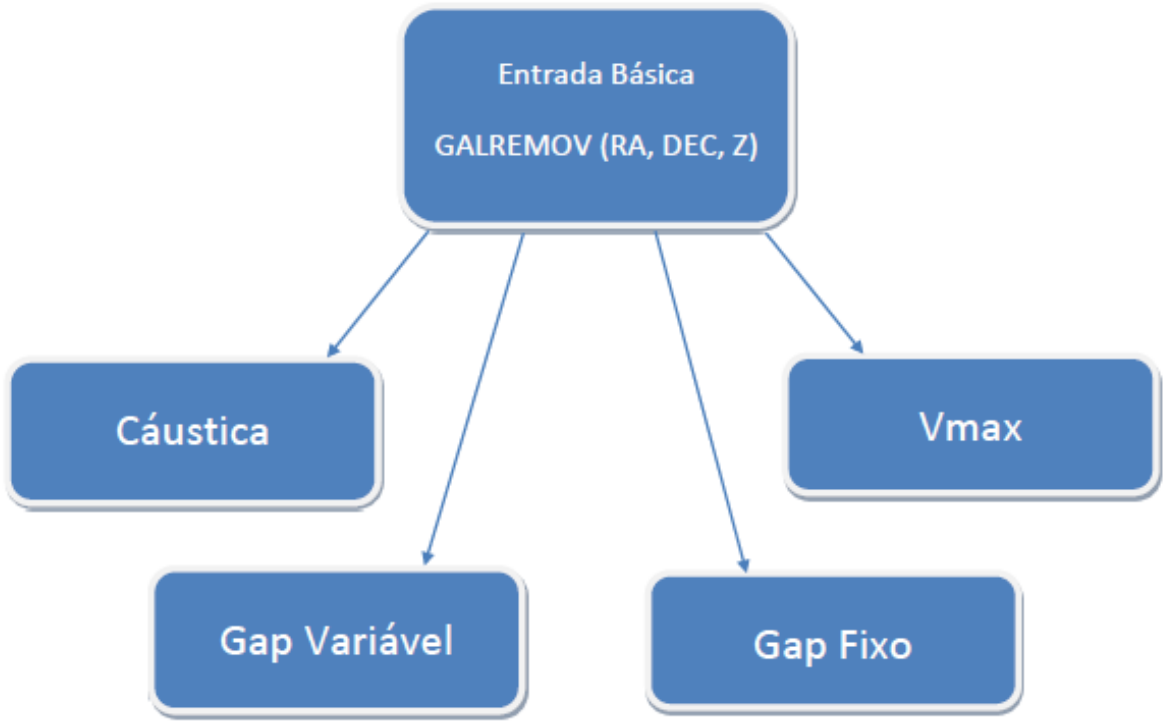

Figura 11 – Fluxograma mostrando os métodos disponíveis para remoção de `outliers` dentro do `galremov`.

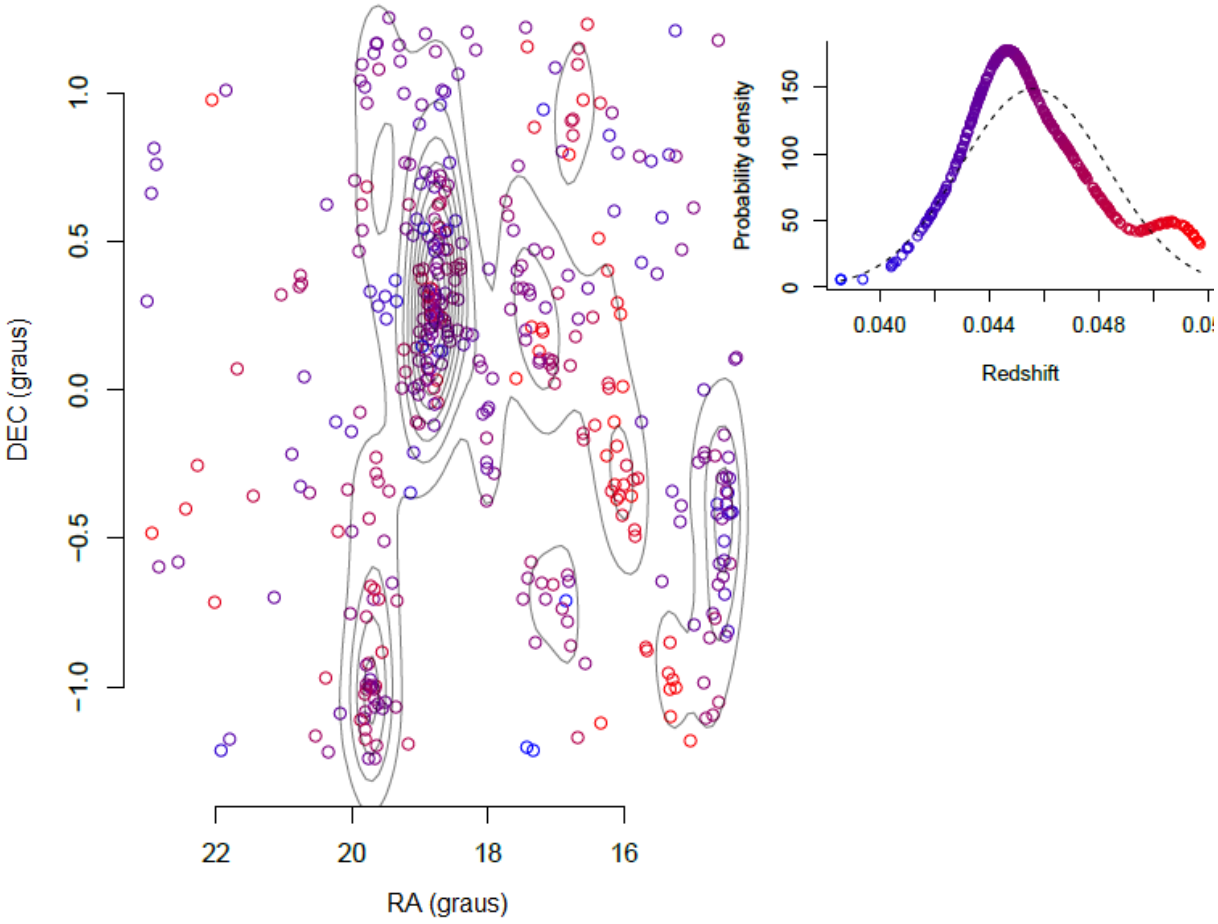

Figura 12 – Gráfico apresentando o aglomerado de Abell A168.

3.1.1 Cáustica

Para o método da cáustica utilizamos uma versão modificada da função escrita em R por Mehmet Alpaslan (com algumas sub-rotinas escritas por Aaron Robotham). A implementação desenvolvida por eles tem como intuito a obtenção da massa do aglomerado de galáxias, produzindo originalmente apenas esta saída. Isto nos obrigou a fazer algumas modificações para que, além do cálculo da massa, o programa também utilizasse a curva da cáustica como critério de remoção de `outliers`.

Os parâmetros mais gerais usados pelo programa[1] são aqueles relacionados à cosmologia particular que desejamos considerar: $H_0, \Omega_M, \Omega_\Lambda$. Estes são os parâmetros de Hubble, parâmetros de densidade de matéria e de densidade de energia escura do universo, que contabilizam a taxa de expansão do universo, a densidade de matéria e a densidade da componente ainda não compreendida totalmente e, portanto, chamada

[1] Na presente versão de pacote sempre consideramos o modelo ΛCMD

de 'escura' (vide, por exemplo Ryden 2016). Caso o usuário não defina a cosmologia, o programa assume os valores: $H_0 = 72$ km/s/Mpc, $\Omega_M = 0.25$ e $\Omega_\Lambda = 0.75$.

Além dos parâmetros cosmológicos, deve ser feita uma escolha para o centro do aglomerado e um raio projetado de abertura dentro do qual se estima a dispersão de velocidades inicial. Caso estes dois últimos valores não sejam dados, o programa estima o centro através do baricentro dentro de uma abertura dada pelo raio harmônico de todos os objetos do campo. Esta abertura também é utilizada para o cálculo da dispersão de velocidades inicial.

Adicionalmente, os dados de entrada necessários para a execução do programa correspondem às coordenadas angulares das galáxias (ascensão reta e declinação) e seus `redshifts`. Esta corresponderia a uma entrada mínima. Se o usuário tiver outros atributos associados às galáxias, estes são mantidos nas listas de saída após a execução da rotina. Um exemplo de comando executando o método da cáustica:

`remo_causticmass(RA, DEC, Z,` H_0, Ω_M, Ω_Λ, **`Ap, center, Plot)`**,

onde, além dos parâmetros cosmológicos, `RA, DEC, Z` são os vetores de posições angulares das galáxias e dos `redshifts`, `Ap` é a abertura dentro da qual se calcula a dispersão de velocidades inicial, e `center` é o centro do aglomerado. O parâmetro `Plot` ('T' ou 'F') é usado se desejamos uma saída gráfica após a execução do programa no espaço definido pela distância projetada ao centro (em unidades de Mpc) e velocidades na linha de visada no referencial do aglomerado (em km/s).

Tivemos que modificar a saída original da rotina para retornar a curva da cáustica, e assim podermos fazer a remoção das galáxias intrusas. A saída da nova rotina ficou sendo uma lista que contém a amostra limpa (com os intrusos eliminados), o vetor de velocidades e a massa do aglomerado. Modificamos também a exibição do gráfico para a melhor visualização da remoção dos `outliers`. Veja na Tabela 1 e na Figura 13 um exemplo de saída do programa. Na Tabela 1 vemos que além das coordenadas `RA, DEC e Z`, são relacionadas as distâncias projetadas ao centro do aglomerado (normalizadas pelo raio de abertura) e as velocidades na linha de visada (normalizada pela dispersão de velocidades final) no referencial do aglomerado. Na Figura 13, vemos em círculos pretos as galáxias que foram consideradas membros do aglomerado pela cáustica, sendo os círculos vermelhos aquelas galáxias que foram classificadas como `outliers`.[2]

[2]Embora a recomendação seja de que limitemos os distâncias projetadas em 4 Mpc, para efeito de melhor visualização do método, usamos dados até 10 Mpc.

Tabela 1 – As 5 primeiras linhas da primeira tabela resultante da saída do método da cáustica, nesse exemplo a tabela é referente a amostra livre das galáxias intrusas, essa tabela é similar a segunda tabela que é referente as galáxias que foram consideradas intrusas.

RA	DEC	redshift	projdist	vlos
18.8098	0.3387	0.0467	0.6551	441.7080
18.8180	0.3094	0.0433	0.6171	-500.682
18.7850	0.3036	0.0463	0.5589	343.6240
18.8080	0.2893	0.0459	0.5688	226.2086
18.7651	0.4052	0.0446	0.7314	-119.9943

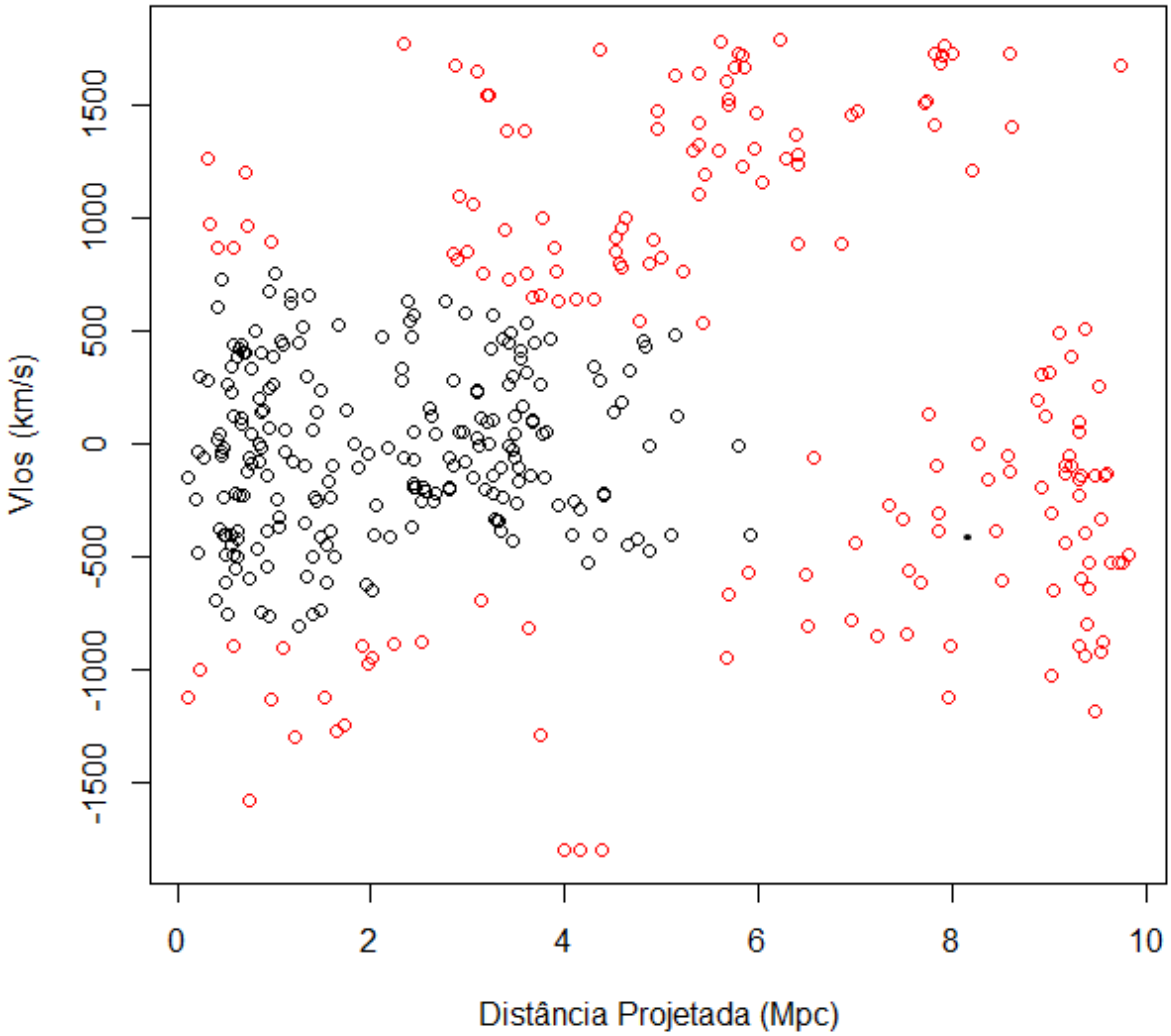

Figura 13 – Gráfico resultante da remoção de `outiers` executada pela cáustica. Os círculos vermelhos são as galáxias que foram classificadas como `outliers`.

3.1.2 Vmax

A rotina Vmax foi implementada inteiramente neste trabalho. Ela é executada de maneira semelhante à rotina da cáustica:

`remo_Vmax(RA, DEC, Z,` H_0**`,`** Ω_M**`,`** Ω_Λ**`, center, Plot)`**.

Mais uma vez, a rotina necessita obrigatoriamente dos valores de `RA, DEC, Z`, que são as coordenadas angulares das galáxias e seus `redshifts`. Os parâmetros cosmológicos, se não forem dados, assumem os valores indicados na seção 3.1.1. O centro, se não é dado, será definido pelo baricentro não-ponderado da distribuição.

A saída do programa é composta por uma lista com dois elementos: o primeiro deles é uma tabela contendo as galáxias consideradas pertencentes aos aglomerados e a segunda contém as galáxias que foram consideradas intrusas. Veja na Tabela 2 e na Figura 14 um exemplo de saída do programa.

Tabela 2 – As 5 primeiras linhas da primeira tabela da saída do método do Vmax que contém as galáxias que foram retiradas pelo método, sendo que a segunda tabela da saída segue este mesmo modelo.

RA	DEC	redshift	projdist	vlos
18.8098	0.3387	0.0467	0.6551	441.7080
18.8180	0.3094	0.0433	0.6171	-500.6826
18.7728	0.3222	0.0418	0.5766	-895.6634
18.7850	0.3036	0.0463	0.5589	343.6240
18.8080	0.2893	0.0459	0.5688	226.2086

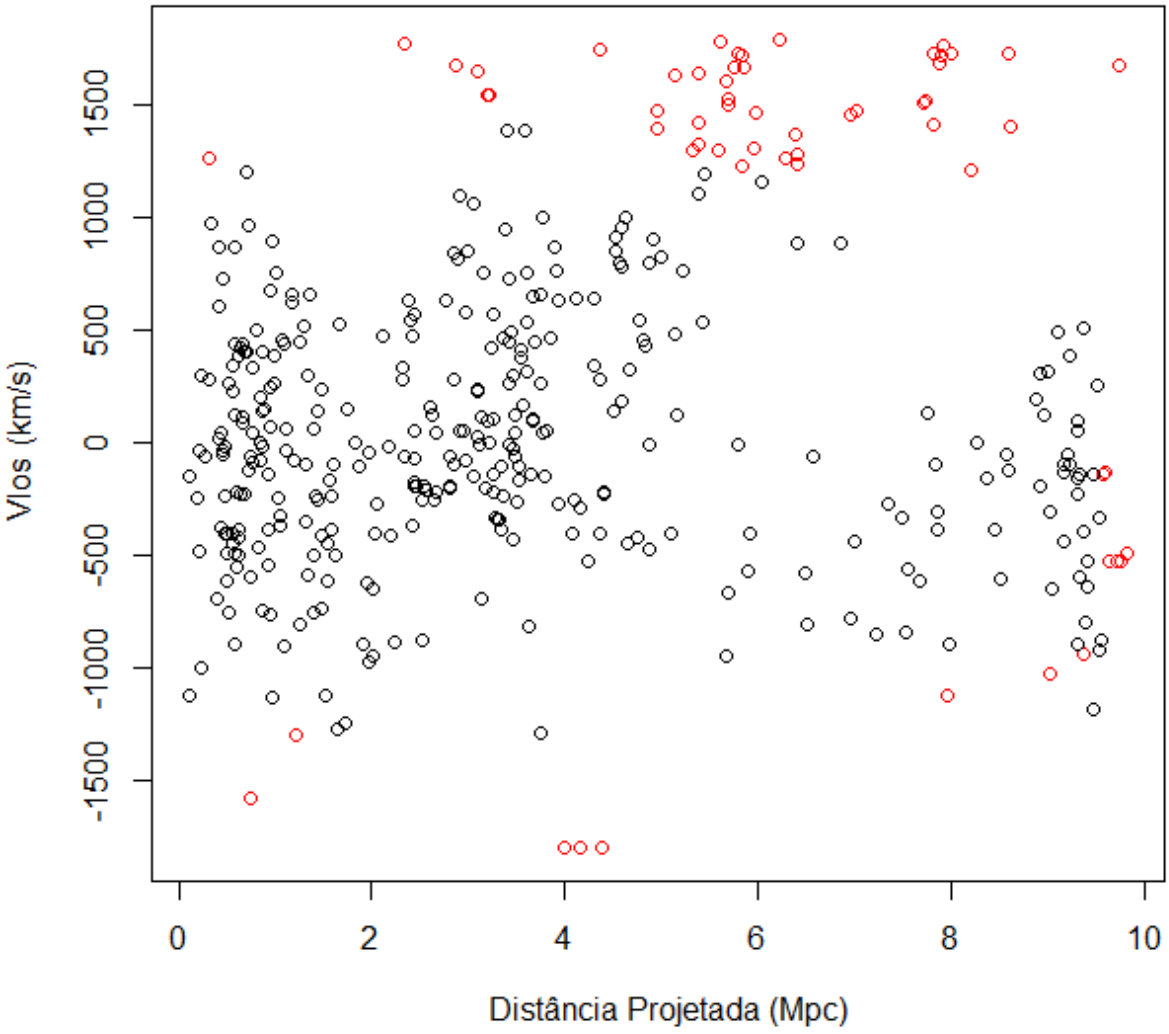

Figura 14 – Gráfico resultante da remoção de `outiers` executada pela VMAX. Os círculos vermelhos são as galáxias que foram classificadas como `outliers`.

3.1.3 Gap Fixo

A rotina do Gap Fixo, também implementada integralmente neste trabalho, é executada de forma semelhante aos casos anteriores:

`remo_GapFix(RA, DEC, Z,` H_0**,** Ω_M**,** Ω_Λ**, center, gap, Plot),**

sendo que, neste caso, temos o parâmetro `gap` que, se não for dado pelo usuário, será definido como `gap` = 350 km/s (vide Lopes et al. 2009). O tamanho dos `bins` é encontrado automaticamente pelo programa com o compromisso de ter aproximadamente o mesmo número de galáxias em cada um, com um mínimo de 15 galáxias em todos eles.

Sua saída é composta por uma lista com dois elementos: o primeiro deles é uma tabela contendo as galáxias consideradas pertencentes aos aglomerados e a segunda contém as galáxias que foram consideradas intrusas. Veja nas Tabelas 3, 4 e na Figura 15 um exemplo de saída do programa.

Tabela 3 – As 5 primeiras linhas da primeira tabela resultante da saída do método do Gap Fixo com relação as galáxias que foram consideradas pertencentes ao aglomerado.

RA	DEC	redshift	projdist	vlos
18.8098	0.3387	0.0467	0.6551	441.7080
18.8180	0.3094	0.0433	0.6171	-500.6826
18.7728	0.3222	0.0418	0.5766	-895.6634
18.7850	0.3036	0.0463	0.5589	343.6240
18.8080	0.2893	0.0459	0.5688	226.2086

Tabela 4 – As 5 primeiras linhas da segunda tabela resultante da saída do método do Gap Fixo com relação as galáxias que foram consideradas galáxias intrusas.

RA	DEC	redshift	projdist	vlos
18.9639	0.1469	0.0393	0.7523	-1578.1661
20.2352	-0.1055	0.0421	3.6273	-816.2408
17.3559	0.2118	0.0512	2.8650	1677.8126
19.6136	0.2854	0.0418	2.2385	-887.2803
19.7303	0.3311	0.0419	2.5147	-880.7114

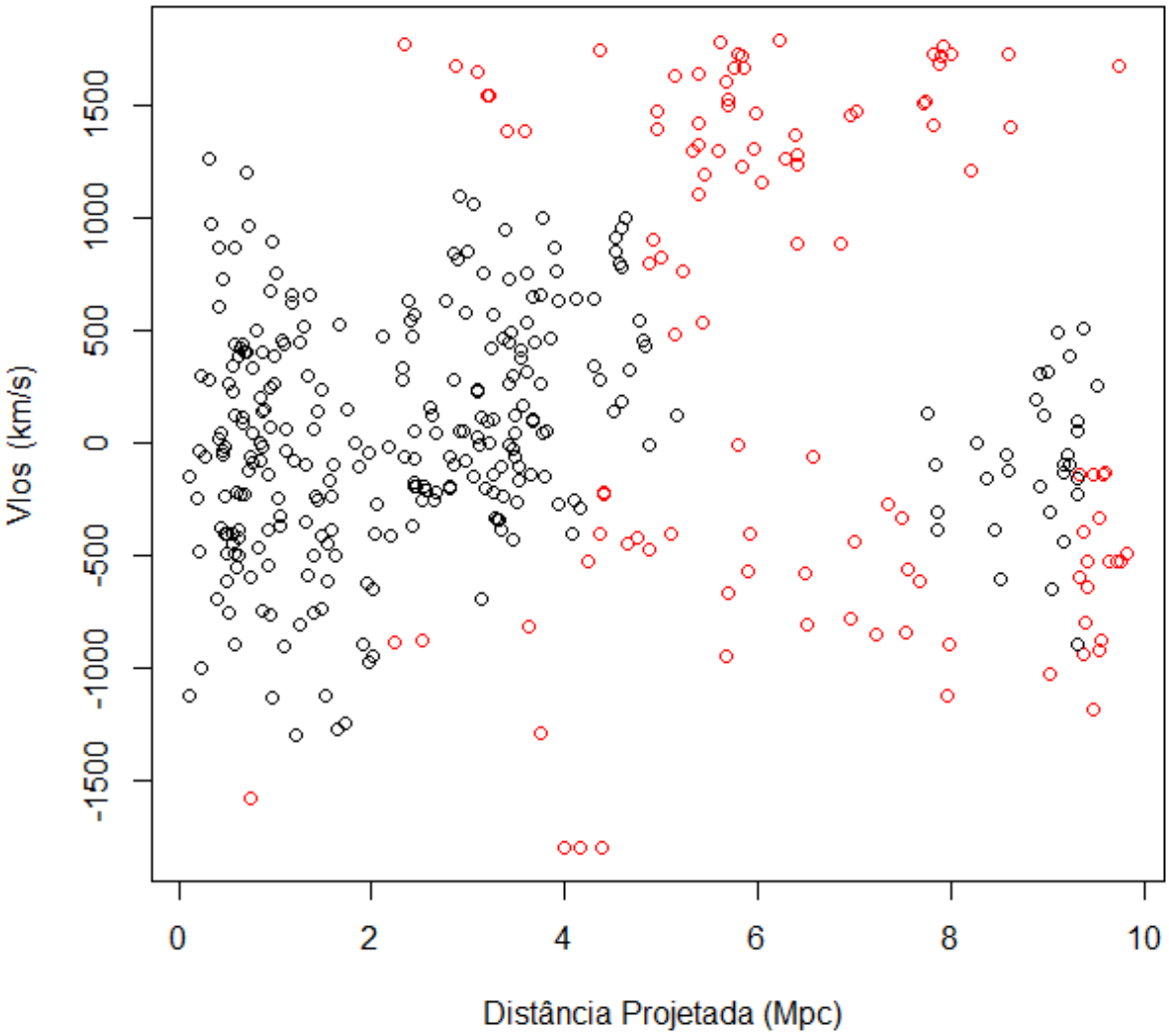

Figura 15 – Gráfico resultante da remoção de `outiers` executada pelo Gap Fixo. Os círculos vermelhos são as galaxias que foram classificadas como `outliers`.

3.1.4 Gap Variável

O Gap Variável é mais um caso de rotina que implementamos integralmente. Ele também é executado de forma semelhante aos anteriores:

`remo_GapVar(RA, DEC, Z,` H_0**`,`** Ω_M**`,`** Ω_Λ**`, center, Plot),**

sendo que neste caso o tamanho do gap é calculado dentro do programa, em cada `bin` de velocidades (vide seção 2.3.4). O tamanho dos `bins` é definido como no caso anterior.

A saída do programa é composta por uma lista com dois elementos: o primeiro deles é uma tabela contendo as galáxias consideradas pertencentes ao aglomerado e a segunda contém as galáxias que foram consideradas intrusas. Veja nas Tabelas 5, 6 e na Figura 16 um exemplo de saída do programa.

Tabela 5 – As 5 primeiras linhas da primeira tabela resultante da saída do método do Gap Variável com relação as galáxias que foram consideradas pertencentes ao aglomerado.

RA	DEC	redshift	projdist	vlos
18.8098	0.3387	0.0467	0.6551	441.7080
18.8180	0.3094	0.0433	0.6171	-500.6826
18.7728	0.3222	0.0418	0.5766	-895.6634
18.7850	0.3036	0.0463	0.5589	343.6240
18.8080	0.2893	0.0459	0.5688	226.2086

Tabela 6 – As 5 primeiras linhas da segunda tabela resultante da saída do método do Gap Variável com relação as galáxias que foram consideradas galáxias intrusas.

RA	DEC	redshift	projdist	vlos
18.9639	0.1469	0.0393	0.7523	-1578.1661
19.7441	-0.4351	0.0474	2.7761	628.3656
17.3559	0.2118	0.0512	2.8650	1677.8126
19.6136	0.2854	0.0418	2.2385	-887.2803
19.7303	0.3311	0.0419	2.5147	-880.7114

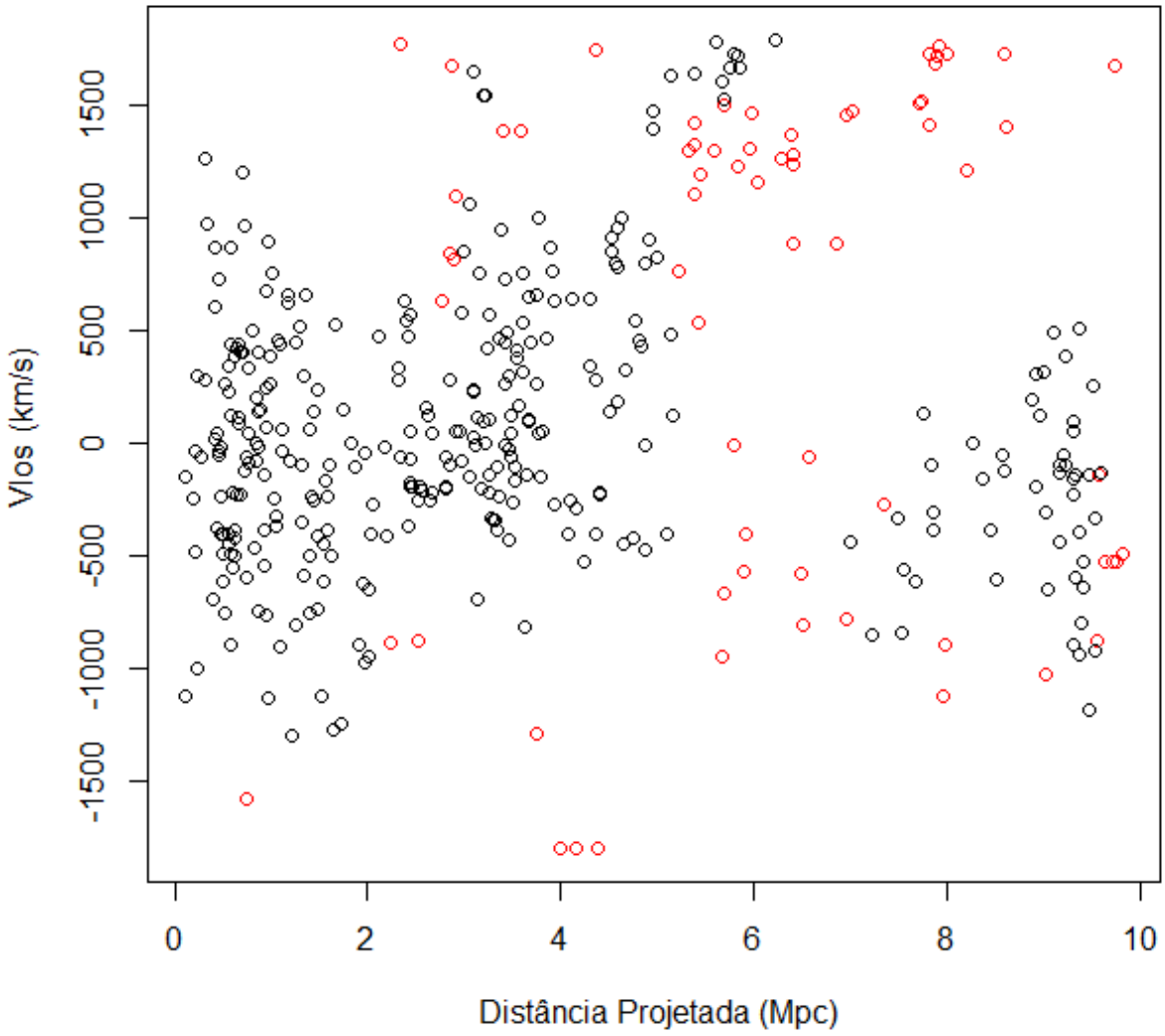

Figura 16 – Gráfico resultante da remoção de `outiers` executada pelo Gap Variável. Os círculos vermelhos são as galáxias que foram classificadas como `outiers`.

3.1.5 Galremov

Ao invés de executar cada rotina de remoção explicitamente através de comandos individuais, como descrevemos acima, definimos a rotina `Galremov`, que engloba toda a etapa de remoção de `outliers`. Além dos dados e parâmetros necessários para a execução de cada método, em `Galremov` é introduzida uma variável que é um vetor contendo os métodos escolhidos podendo ser eles `1-Cáustica, 2-Vmax, 3-Gap Fixo e 4- Gap Variável` outra variável que pode assumir os valores "INT", "AND" ou "UNI", referente aos modos em que os métodos vão trabalhar, significando: intersecção, sequência ou união, respectivamente podemos observar melhor na Figura 17. Inclui-se também uma variável `plot` para uma saída gráfica, essa variável pode assumir o valor de 'T' ou 'F'. Caso não seja dado o seu valor, o padrão é 'F'.

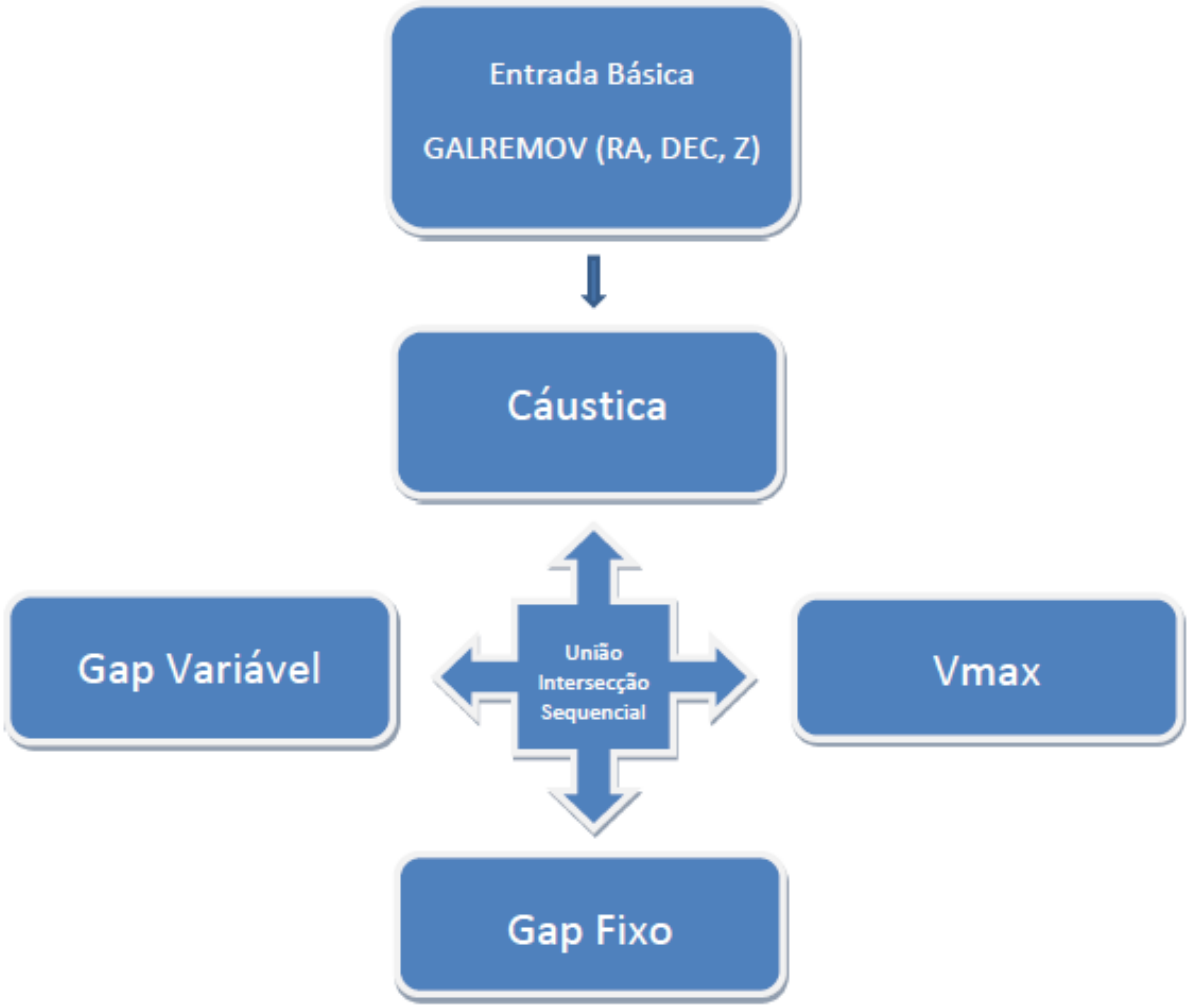

Figura 17 – Fluxograma mostrando os métodos disponíveis para remoção de `outliers` dentro do `galremov`, e os dois modos de união e interseção, sem esquecer também do modo sequencial.

Um exemplo de execução é dado por

`galremov(dados, method=c(1), plot=T)`,

nesse caso estamos utilizando o método da cáustica e exibindo o gráfico. Em seguida, detalhamos como são feitos os modos de execução.

Modos de execução

Existem três modos de execução. Apresentamos inicialmente o modo de união, que é escolhido enviando "UNI" na variável `type` de entrada. Isto é feito da seguinte maneira:

`galremov(dados, type=UNI, method=c(3,4), plot=T)`,

neste caso estamos usando a união dos métodos escolhidos. Ou seja, o resultado vem dos métodos escolhidos, `3 - Gap Fixo e 4 - Gap Variável`, fazendo-se uma comparação ao final para achar a união de todas as galáxias que foram consideradas intrusas ao aglomerado. Este modo é mais "agressivo" na eliminação das galáxias `outliers` e pode ser usado em estudos em que a a amostra final tenha um baixo nível de contaminação. Vide Figura 18 exibindo o resultado deste modo de execução.

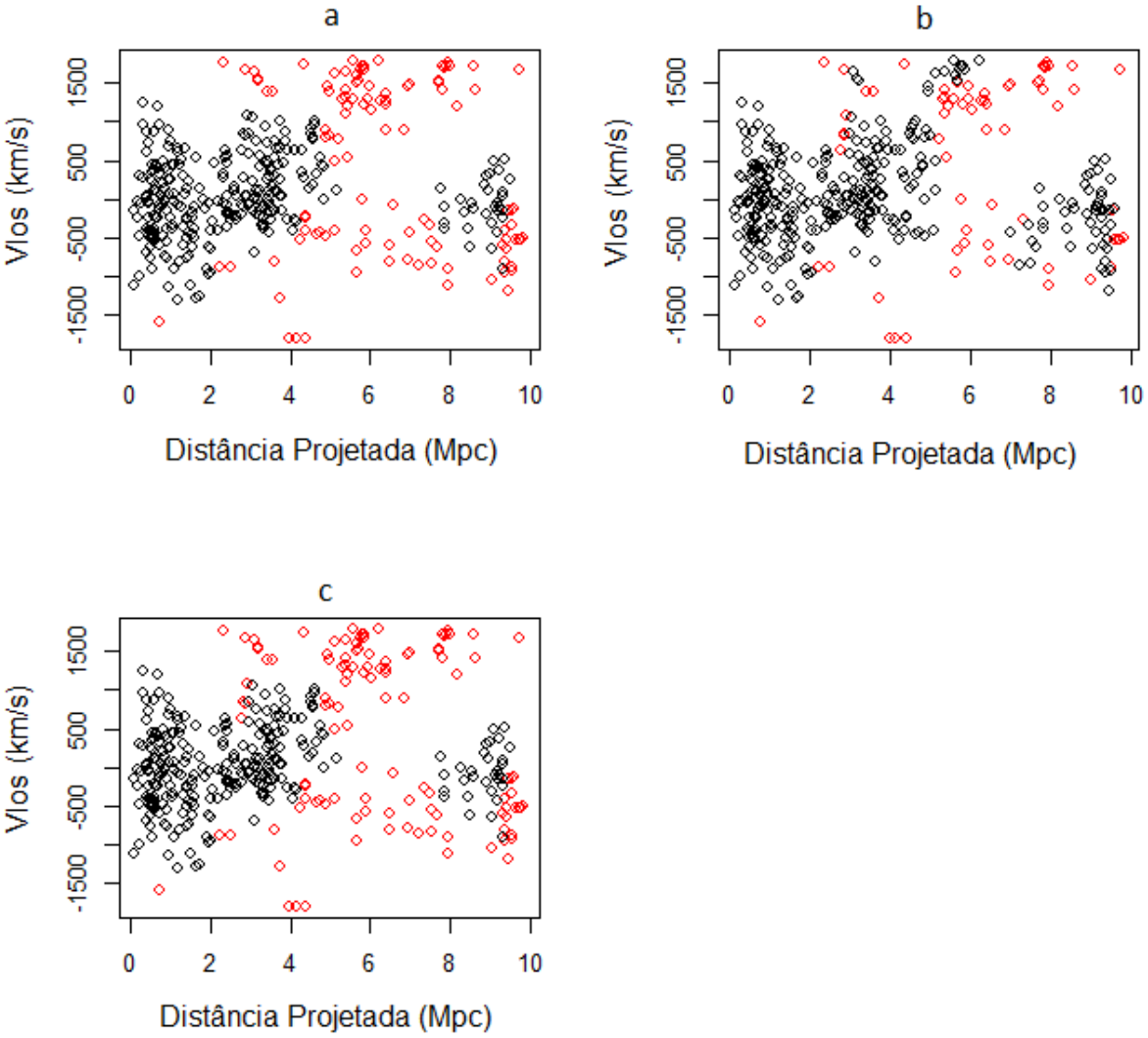

Figura 18 – Gráfico resultante da remoção de `outliers` executada pelo Gap Fixo e Gap Variável utilizando o modo "UNI". Os círculos vermelhos são as galáxias que foram classificadas como `outliers`. No gráfico 'a' e 'b' mostramos os resultados dos métodos Gap Fixe e Gap Variável respectivamente. No gráfico 'c' exibimos o resultado final.

Um outro modo de execução é o de interseção. Ele segue o mesmo princípio do de união e pode ser escolhido enviando "INT" na variável `type` de entrada:

```
galremov(dados, type=INT, method=c(3,4), plot=T),
```

a diferença é que neste caso os resultados dos métodos escolhidos são usados para se fazer uma interseção entre as galáxias que foram consideradas intrusas ao aglomerado, desse modo a remoção das galáxias intrusas passa por um critério mais "conservador", e pode ser utilizada na tentativa de evitar-se remoções indevidas dos dados. Vide Figura 19 exibindo o resultado deste modo de execução.

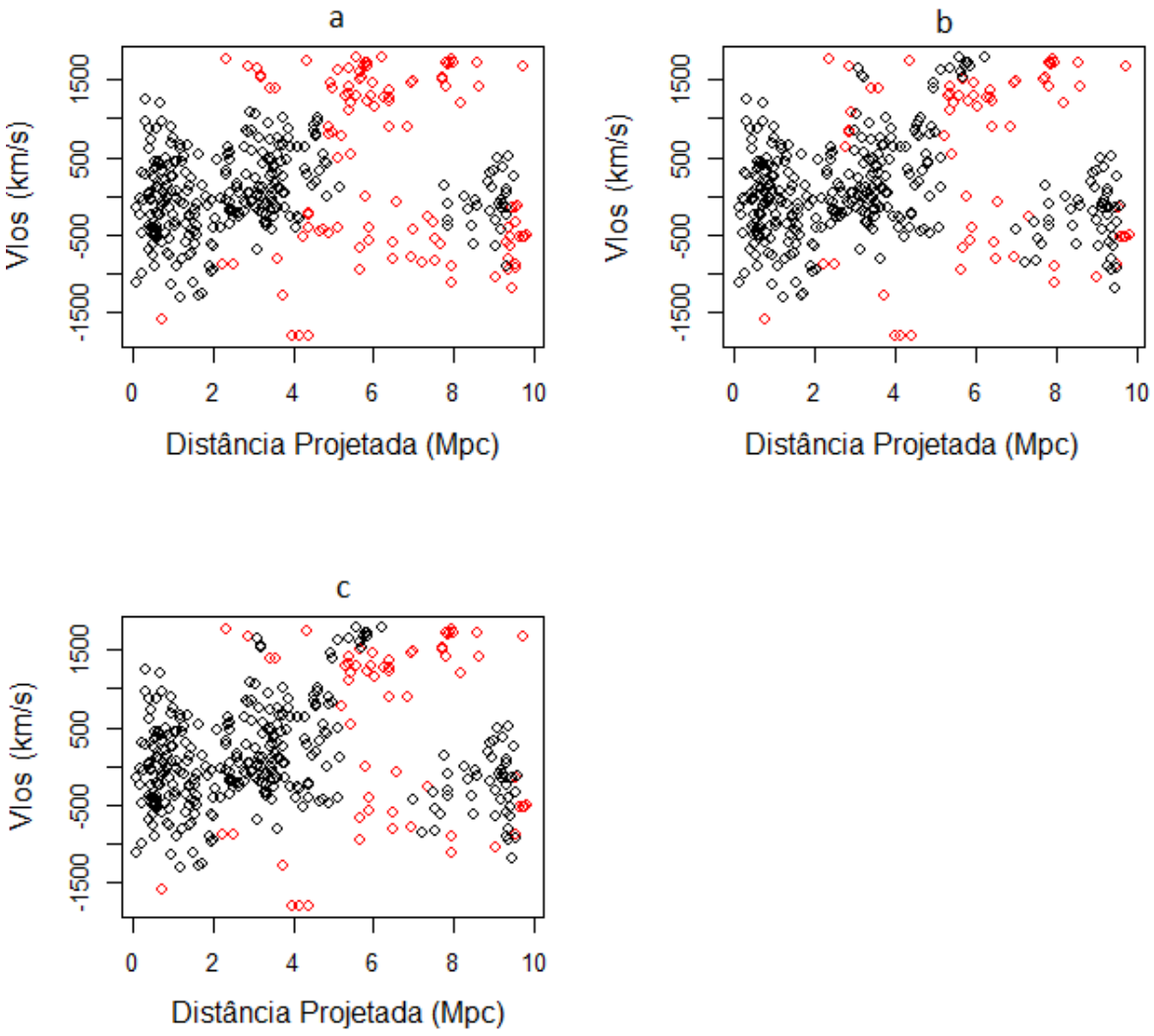

Figura 19 – Gráfico resultante da remoção de `outliers` executada pelo Gap Fixo e Gap Variável utilizando o modo "INT". Os círculos vermelhos são as galáxias que foram classificadas como `outliers`. No gráfico 'a' e 'b' mostramos os resultados dos métodos Gap Fixe e Gap Variável respectivamente. No gráfico 'c' exibimos o resultado final.

O último modo é o de sequência, que pode ser escolhido enviando "AND" na variável `type` de entrada:

```
galremov(dados, type=AND, method=c(3,4), plot=T).
```

Esse modo utiliza a lista de métodos escolhidos de forma sequencial. Ele executa o primeiro método, logo após utiliza o resultado do primeiro método e encaminha para o segundo, até acabar a lista de métodos. Este modo oferece grande liberdade ao usuário, podendo gerar resultados intermediários entre os casos anteriores, a depender exclusivamente dos métodos que foram escolhidos e da ordem que foi passada. Vide

Figura 20 exibindo o resultado deste modo de execução.

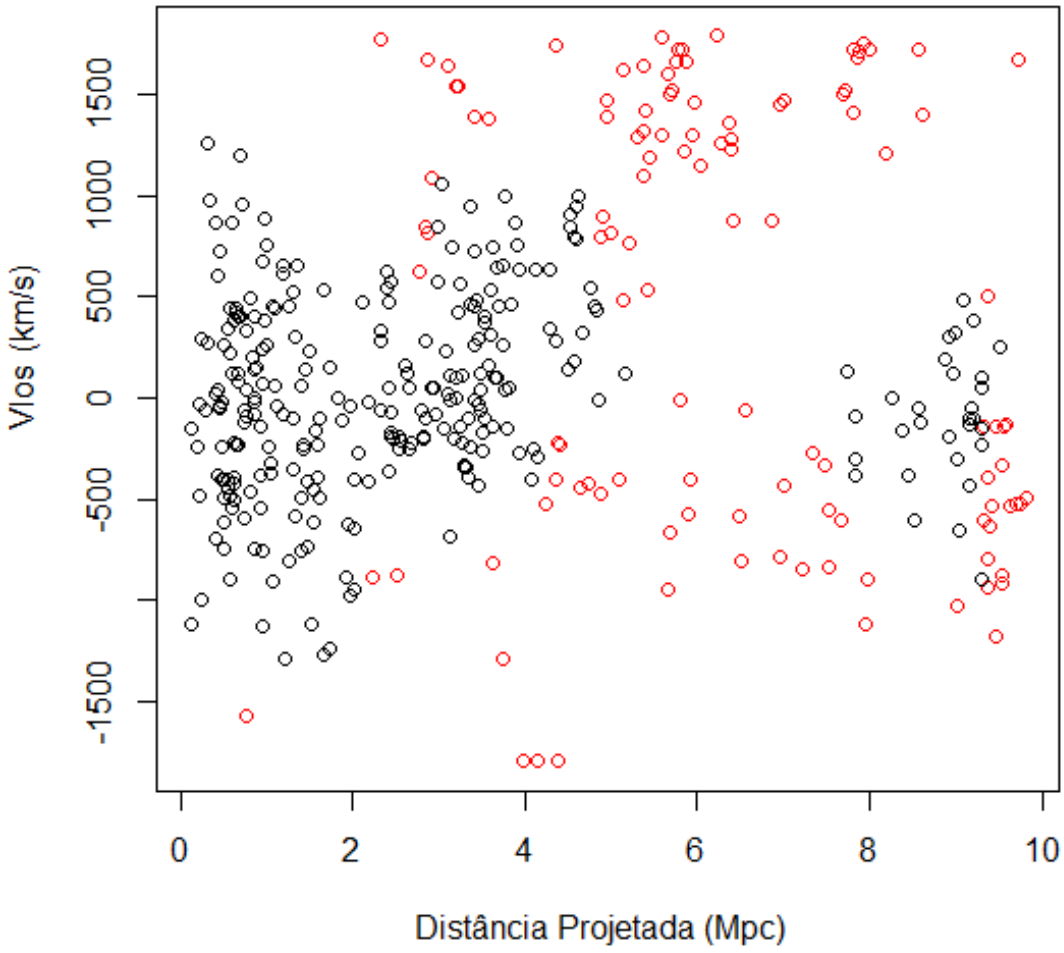

Figura 20 – Gráfico resultante da remoção de `outliers` executada pelo Gap Fixo e Gap Variável utilizando o modo "AND"na respectiva sequência. Os círculos vermelhos são as galáxias que foram classificadas como `outliers`.

Saída

A saída de `Galremov` é bem variável e dependerá da quantidade de métodos e do modo que foi escolhido. Caso tenha escolhido somente um método, a saída será duas tabelas, a primeira delas conterá todas as galáxias que foram consideradas pertencentes ao aglomerado, e a segunda tabela conterá todas as galáxias que foram consideradas intrusas. Caso tenha escolhido o método da cáustica haverá uma terceira tabela que é o valor que a cáustica encontrou como massa do aglomerado. Caso tenha escolhido mais de um método as saídas são as seguintes: a primeira tabela conterá as galáxias que foram consideradas pertencentes ao aglomerado, as tabelas seguintes conterão os

elementos que foram considerados `outliers` por método, e a último tabela conterá as galáxias que foram retiradas na junção do processo. Além das saídas em forma de listas, podemos ter saídas gráficas, caso o usuário tenha escolhido essa opção. Se tiver escolhido essa opção como modo de trabalhar com mais de um método, serão gerados vários gráficos seguindo a ordem de acordo com a sequência dos métodos escolhidos na entrada, e um outro gráfico no fim para o resultado final, podemos ver os exemplos de saída na Figura 21 e Tab 7 .

Tabela 7 – As 5 primeiras linhas de uma das tabelas resultantes da saída do `galremov`, todas as outras tabelas na lista são similares.

RA	DEC	redshift	projdist	vlos
18.8098	0.3387	0.0467	0.6551	441.7080
18.8180	0.3094	0.0433	0.6171	-500.6826
18.7728	0.3222	0.0418	0.5766	-895.6634
18.7850	0.3036	0.0463	0.5589	343.6240
18.8080	0.2893	0.0459	0.5688	226.2086

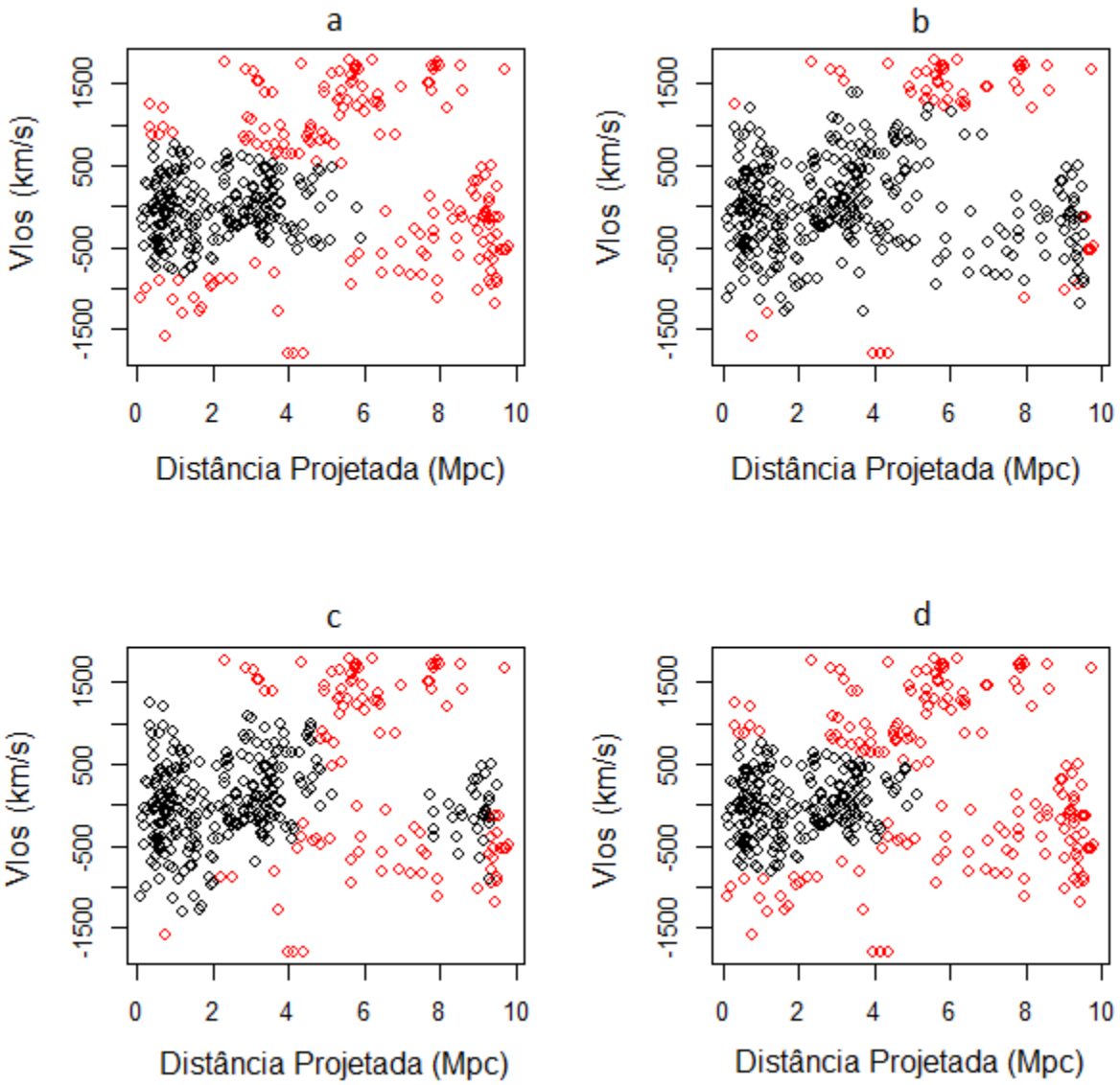

Figura 21 – Figura resultante da saída da função `galremov` utilizando como entrada o aglomerado de Abell A168, sendo o primeiro gráfico referente à eliminação feita pelo método da Cáustica, o segundo gráfico pelo método Vmax o terceiro pelo Gap Fixo, a último gráfico é o resultado da união dos métodos.

3.2 Análise Dinâmica

Nesta seção, descrevemos os 3 métodos que podem ser usados pela rotina `GalClus` para realizar a análise dinâmica dos aglomerados, uma vez que seus membros tenham sido determinados no passo anterior. Os resultados, após a execução desta função, são diversos diagnósticos que auxiliam o usuário a classificar o sistema como estando ou não em equilíbrio (lembrando sempre que estes diagnósticos são estimadores indiretos da dinâmica dos aglomerados). Na Figura 22 mostramos um fluxograma

ilustrando o funcionamento geral da rotina.

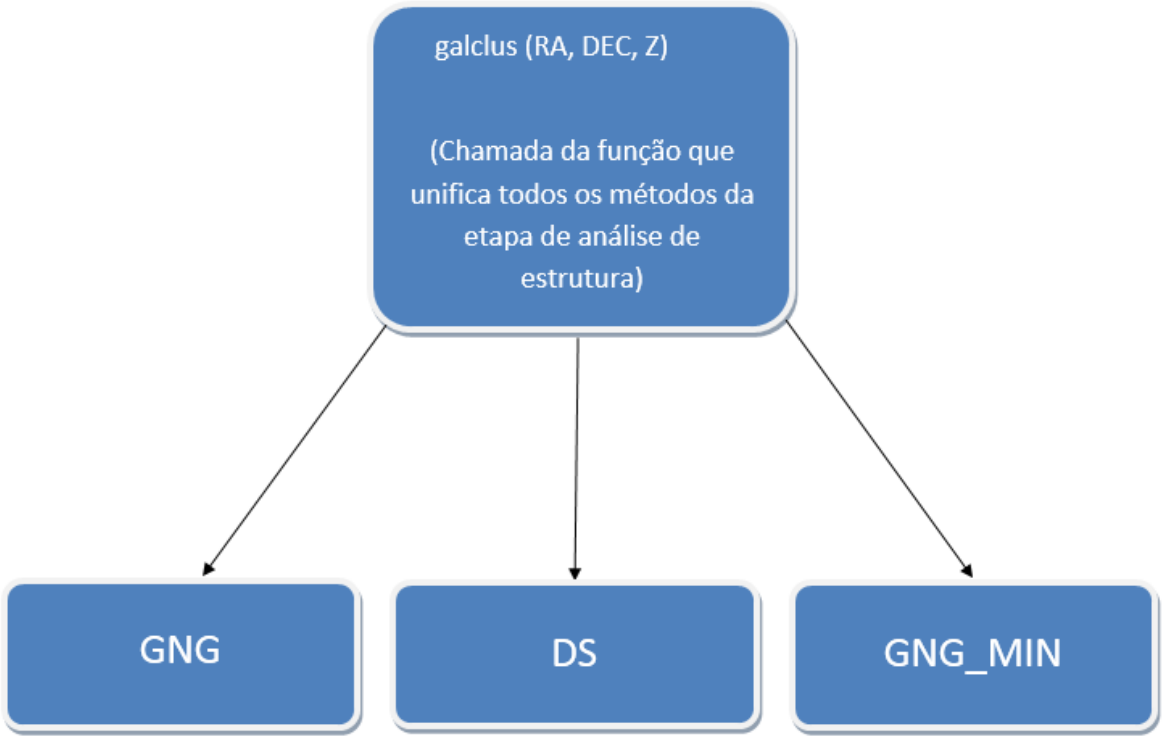

Figura 22 – Fluxograma representando os métodos internos da etapa de análise dinâmica.

3.2.1 GNG

O GNG é uma rotina desenvolvida por Ribeiro et al. (2013), modificada nesse projeto por motivos de controle de entrada e saída de dados, e para desenvolver uma versão que permita a utilização das bibliotecas para processamento paralelo. O GNG é um método que analisa a distribuição de velocidades das galáxias membro de um aglomerado. Ele utiliza dois métodos, já descritos na seção 2.4.1. O `Mclust` e a distância de Hellinger. Estes métodos fornecem diagnósticos sobre a distribuição de velocidades: se ela é gaussiana/não-gaussiana e unimodal/multimodal.

O método necessita como entrada os `redshifts Z`, agora referentes apenas aos objetos que foram considerados membros na etapa anterior de análise. Além disso, precisa de um parâmetro `Conflim` que define a confiabilidade mínima dos resultados para que seja obtida uma classificação 0 (sistema em equilíbrio), 1 (sistema fora do equilíbrio) ou 2 (quando não atinge a confiabilidade assim a confiança fica como -1). O padrão é `Conflim=70`, indicando que esperamos uma repetição dos resultados, ao longo de N reamostragens, em 70% das vezes. Em casos onde esta confiabilidade não for atingida, o diagnóstico final é 2, indicando que o estado dinâmico do sistema não tem

uma classificação estatisticamente significativa. Um exemplo de chamada do método é

```
GNG(Z, Conflim, N),
```

onde `N` é o número de reamostragens, cujo valor padrão é 1000. A execução desta rotina automaticamente ativa tanto o cálculo de HD como de Mclust. Um exemplo de saída da função GNG é:

Tabela 8 – Tabela contendo a saída do método do GNG para o aglomerado A168 de Abell sendo que a remoção de `outliers` foi feita utilizando o método da Cáustica

HD	HDConf	Mclus	MclustConf
1	100	2	-1

O programa retorna os diagnósticos e confiabilidades respectivas de HD e MClust vide Tabela 8. Para objetos dentro do raio harmônico de A168, HD indica não-gaussianidade da distribuição de velocidades com 100% de confiança, enquanto Mclust indica bimodalidade com 59% de confiança. Como o corte em confiabilidade é de 70%, temos um diagnóstico confirmado para HD (mantendo-se o 1) e um diagnóstico não confiável para Mclust (mudado para 2). Na Figura 23 exibimos a distribuição analisada, para esse exemplo foi utilizada a remoção de `outliers` pelo método da cáustica.

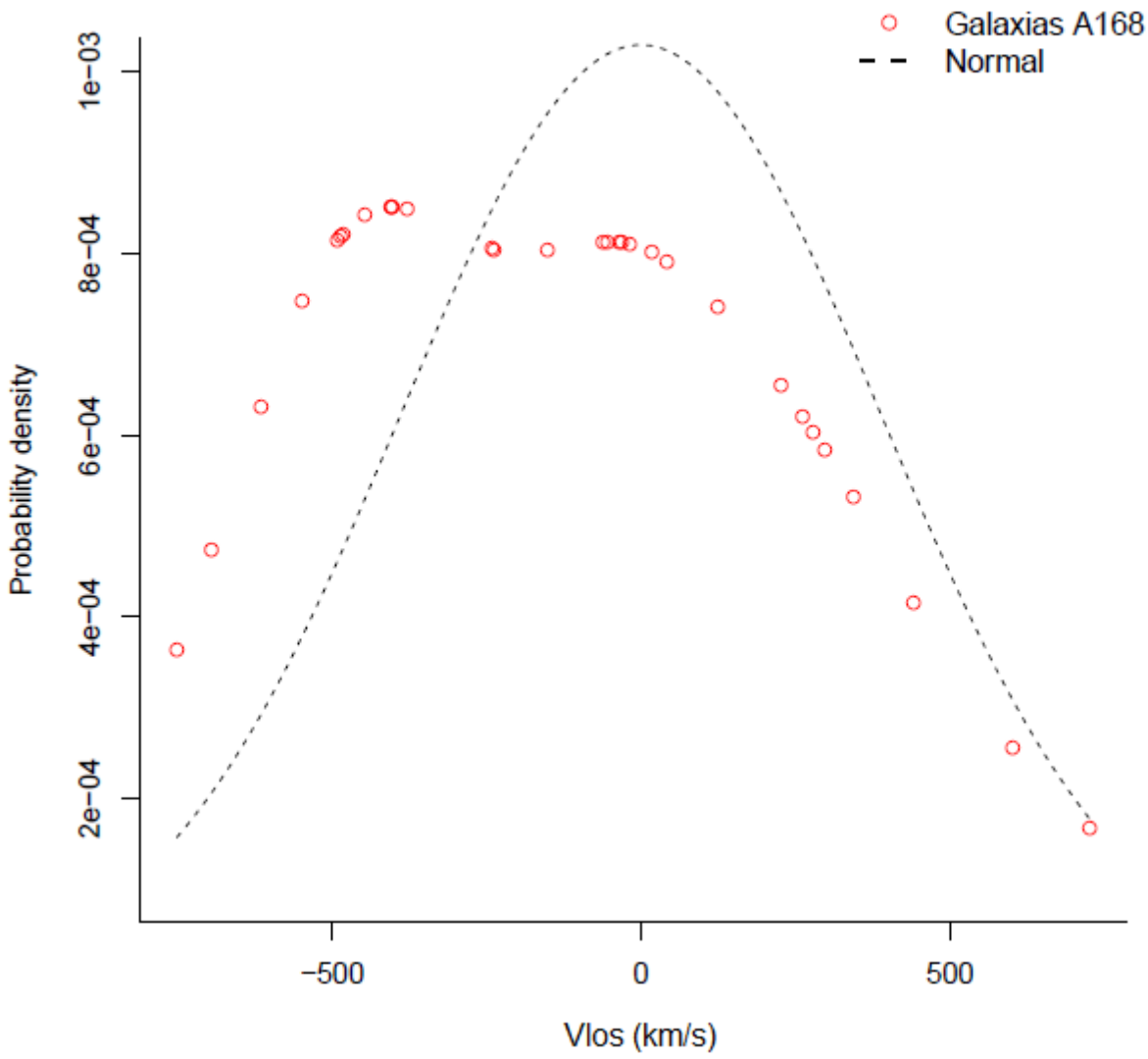

Figura 23 – Distribuição normal sendo comparada com a distribuição de velocidades padronizada do aglomerado A168 de Abell anteriormente submetido a remoção de `outliers` pelo método da Cáustica.

3.2.2 GNG MIN

O GNG MIN é uma rotina em R desenvolvida integralmente neste trabalho. Ela é uma versão simplificada do GNG que usa testes estatísticos tradicionais de normalidade para avaliar a distribuição de velocidades das galáxias membro dos aglomerados, o GNG MIN faz uma análise 1D do perfil de velocidades da linha de visada. Os testes foram apresentados na seção 2.4.2: Anderson-Darling, Jarque-Bera, Shapiro, Anscombe e D'Agostino (vide seção 2.4.2). A execução do método é semelhante ao caso anterior:

```
GNG_MIM(Z, Conflim, N),
```

retornando uma matriz contendo os diagnósticos (definidos também em N reamostragens) `ADc, SHAPc, JBc, ANSc, DAGc`, e suas respectivas confiabilidades. Na Tabela 9 exibimos o resultado do GNG MIN para o aglomerado A168 de Abell, o qual passou pelo processo de remoção de `outliers` pelo método da `cáustica`.

Tabela 9 – Tabela contendo a saída do método GNG MIN para o aglomerado A168 de Abell sendo que a remoção de `outliers` foi feita utilizando o método da Cáustica, o resultado 2 para método do Shapiro significa que o resultado foi inconsistente pois nenhum dos dois casos obtiveram a porcentagem de confiança necessária.

ADc	ADco	SHAPc	SHAco	JBc	JBco	ANSc	ANSco	DAGc	DAGco	class
0	74.09	2	-1	0	99.59	0	75.90	0	99.7	0

Vemos que o teste AD indica normalidade da distribuição com 74% de confiabilidade. Os demais testes, exceto o de Shapiro, apresentam resultado semelhante. O diagnóstico final é dado em função do teste AD, portanto, indicaria gaussianidade.

3.2.3 DS

DS é um teste desenvolvido originalmente por Dressler e Shectman (1988), com diversas implementações disponíveis em C e Fortran. Neste trabalho, implementamos o algoritmo básico integralmente em R. O teste foi criado para identificar subestruturas em aglomerados, através da análise de desvios cinemáticos de pequenos grupos de galáxias vizinhas em relação ao aglomerado como um todo (vide descrição feita na seção 2.4.3).

O método tem como entrada uma matriz contendo as coordenadas de cada galáxia membro do aglomerado `RA, DEC e Z`. Um exemplo de chamada de execução para o método seria

```
DS(c(RA, DEC, Z), N, conflim, plot=T),
```

onde `N` é o número de reamostragens e `lim` é o limite para a razão Δ/N_{gal} ser considerada um indicador da presença de subestruturas, ou seja, quando esta razão excede

`lim`, o diagnóstico final é de que o sistema possui subestruturas. O valor padrão de N é 1000 e de `lim` é 1.4. A saída do programa é uma lista que contém dois elementos: o primeiro deles contendo o resultado geral da análise (ou seja, os valores de δ em torno de cada galáxia) e o segundo contém Δ/N_{gal}, o diagnóstico final, 1 ou 0, indicando se se o sistema possui subestruturas ou não, e finalmente a confiabilidade dada pela fração de vezes que o resultado se repetiu (como no caso de GNG e GNG_MIN). A opção plot propicia ao usuário um gráfico de bolhas, que é amplamente usado quando se analisa estudos com o teste DS. Na Figura 24 ilustramos esta saída. Na Tabela 10 se encontra o resultado final do diagnóstico e na Tabela 11 o resultado geral. Na Tabela 10 vemos que $\Delta/N = 1.06$ e, de acordo com nosso critério não indica subestruturas na região central. De fato, na Figura 24 vemos que as prováveis subestruturas encontram-se fora do raio harmônico do sistema.

Tabela 10 – Primeira tabela da saída do método DS, referente a análise dinâmica do aglomerado A168 de Abell anteriormente submetido ao método de remoção de `outliers` da cáustica.

N	Δ	Δ/N	Class	Conf
31	32.8169	1.0586	0	79%

Tabela 11 – Parte da segunda tabela da saída do método DS, referente ao resultado de cada galáxia do aglomerado A168 de Abell anteriormente submetido ao método de remoção de `outliers` da cáustica.

Ra	Dec	Z	δ_i
18.7850	0.3036	0.0463	1.1647
18.8080	0.2893	0.0459	0.7850
18.7260	0.3032	0.0442	1.3490
18.7596	0.2629	0.0477	1.0542
18.7866	0.2658	0.0428	0.9253

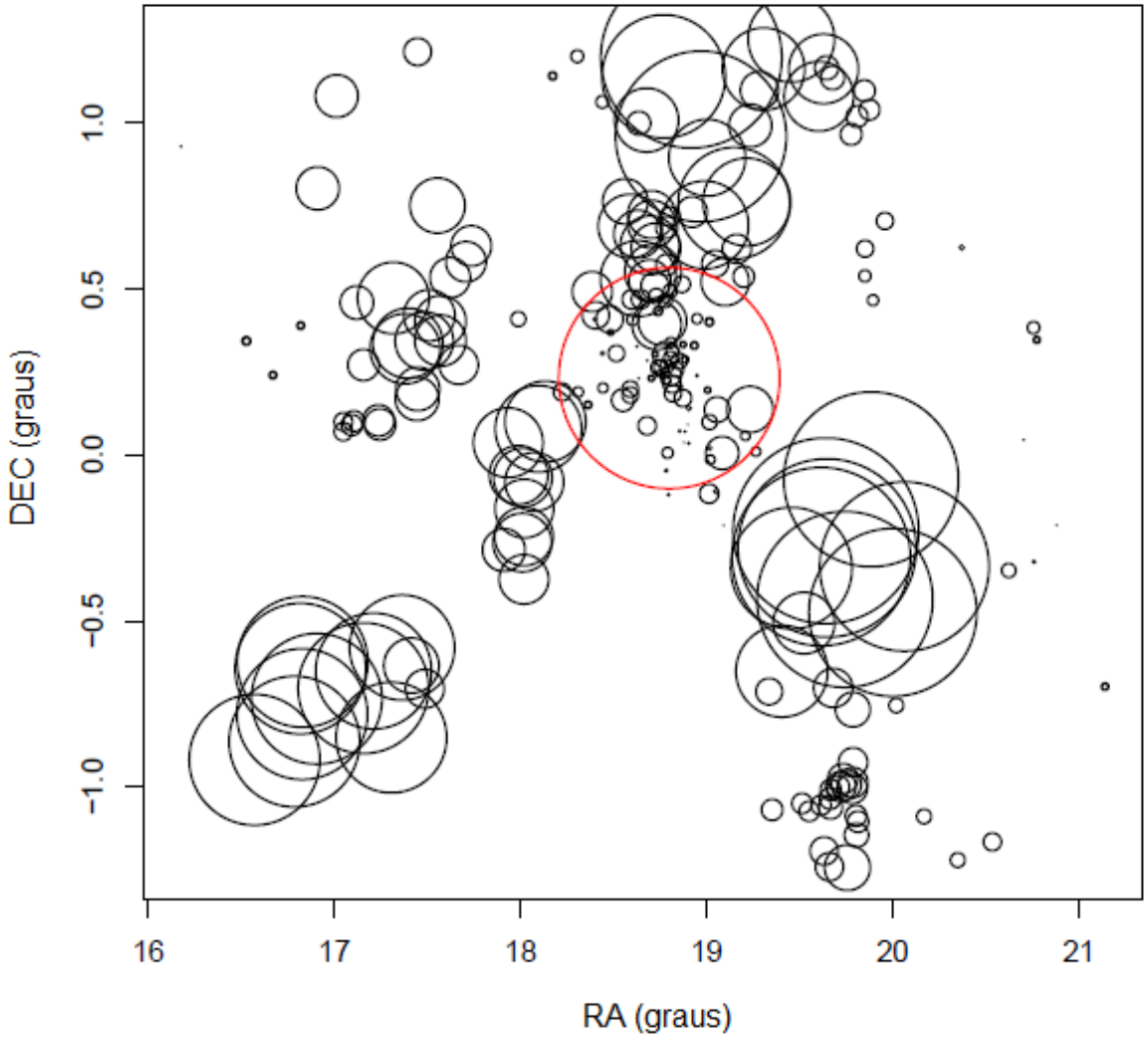

Figura 24 – Distribuição espacial dos membros selecionados pelo método da Cáustica do aglomerado A168 de Abell, cada um marcado por um círculo: quanto maior o círculo, maior é o desvio δ_i dos parâmetros locais dos parâmetros do `cluster` global. Em vermelho indicamos a região contida no raio harmônico.

3.2.4 Galclus

A função `Galclus` foi desenvolvida em R nesse projeto para facilitar a análise da estrutura dinâmica dos aglomerados utilizando todos os três métodos anteriores, `GNG, DS, GNG_MIN`. `Galclus` recebe como entrada as coordenadas `RA, DEC, Z` das galáxias consideradas membros após a etapa de remoção de `outliers`. Os parâmetros cosmológicos são fixados, como em `Galremov`, tendo os mesmos valores padrão,

caso o usuário não os indique. Da mesma forma, o centro do aglomerado pode ser dado pelo usuário, ou tomado como o baricentro da distribuição. A função tem ainda como entrada uma variável que informa quais métodos serão utilizados na análise, sendo eles `1-GNG, 2-DS e 3-GNG_MIN`. Finalmente, a função recebe ainda um valor de confiança, `Conflim`, que tem como padrão o valor 70 (como em `Galremov`). Na Figura 22 mostramos um fluxograma ilustrando a estrutura da função. Um exemplo de entrada seria

```
galclus(RA,DEC,Z,c(1,2,3)),
```

indicando que nesse caso ele usaria os três métodos para fazer a análise.

A saída do `galclus` é uma lista que contém um elemento para cada método escolhido no vetor de entrada. As tabelas 12, 13, 14 ilustram exemplos de saída da análise feita por `Galclus` sobre o aglomerado A168 do catalogo de Abell.

Tabela 12 – Primeira tabela da saída do método `galclus`, contendo a saída do método do GNG para o aglomerado A168 de Abell sendo que a remoção de `outliers` foi feita utilizando o método da Cáustica

Método	Valor
HD	1
HDConf	100
Mclus	2
MclustConf	-1

Tabela 13 – Segunda tabela da saída do método `galclus`, referente a análise dinâmica do aglomerado A168 de Abell anteriormente submetido ao método de remoção de `outliers` da cáustica.

N	Deltas	Δ/N	Class	Conf
31	32.8169	1.0586	0	79%

Tabela 14 – Terceira tabela da saída do método `galclus`, para o aglomerado A168 de Abell sendo que a remoção de `outliers` foi feita utilizando o método da Cáustica, o resultado 2 para método do Shapiro significa que o resultado foi inconsistente pois nenhum dos dois casos obtiveram a porcentagem de confiança necessária.

ADc	ADco	SHAPc	SHAco	JBc	JBco	ANSc	ANSco	DAGc	DAGco	class
0	74.09	2	-1	0	99.59	0	75.90	0	99.7	0

3.3 Estimativa de Massa e Raio

Para a estimativa da massa dos aglomerados, desenvolvemos uma rotina chamada `Massa`. Esta rotina utiliza cinco métodos para estimar a massa (vide Figura 25),

descritos previamente na seção 2.5. São eles: a massa virial, a massa projetada, a massa mediana, M200 e massa da cáustica. Além de estimar a massa, a função `Massa` também fornece a dispersão de velocidades dada pelo estimador robusto SBI (dispersão de velocidades) (Beers et al., 1990) e duas estimativas de raio: raio harmônico e R_{200}, que são aproximações ao raio virial real do sistema (vide seção 2.5).

Essa rotina recebe como entrada uma matriz contendo as coordenadas `RA, DEC, Z` e como opcionais temos os parâmetros cosmológicos (como em `Galremov` e `Galclus`). Ainda como parâmetro opcional temos `b`, indicando a anisotropia das órbitas das galáxias (vide seção 2.5); o valor padrão é 0, fixando órbitas isotrópicas (o usuário pode ainda escolher os valores '1' para órbitas radiais ou '-1' para órbitas circulares). Por fim, a função recebe também como entrada opcional o centro do aglomerado, que está predefinido como o baricentro. Um exemplo de entrada mínima seria

`Massa(RA,DEC,Z)`.

A sua saída é uma matriz contendo o resultado de cada método. A Tabela 15 exibe a saída para a execução de `Massa` sobre o aglomerado A168 do catalogo de Abell.

Tabela 15 – Tabela contendo o resultado do método da `massa`, utilizando como entrada o aglomerado A168 de Abell que foi submetido a retirada de `outliers` pelo método da Cáustica.

Método	Valor	Unidade
Massa Projetada	1.2990 x 10^{14}	Massa Solar
Raio Harmônico	1.4607	Megaparsecs
M200	5.4379 x 10^{13}	Massa Solar
R200	0.6075	Megaparsecs
Numero de membros	220	
Dispersão de velocidade	358.7745	km/s

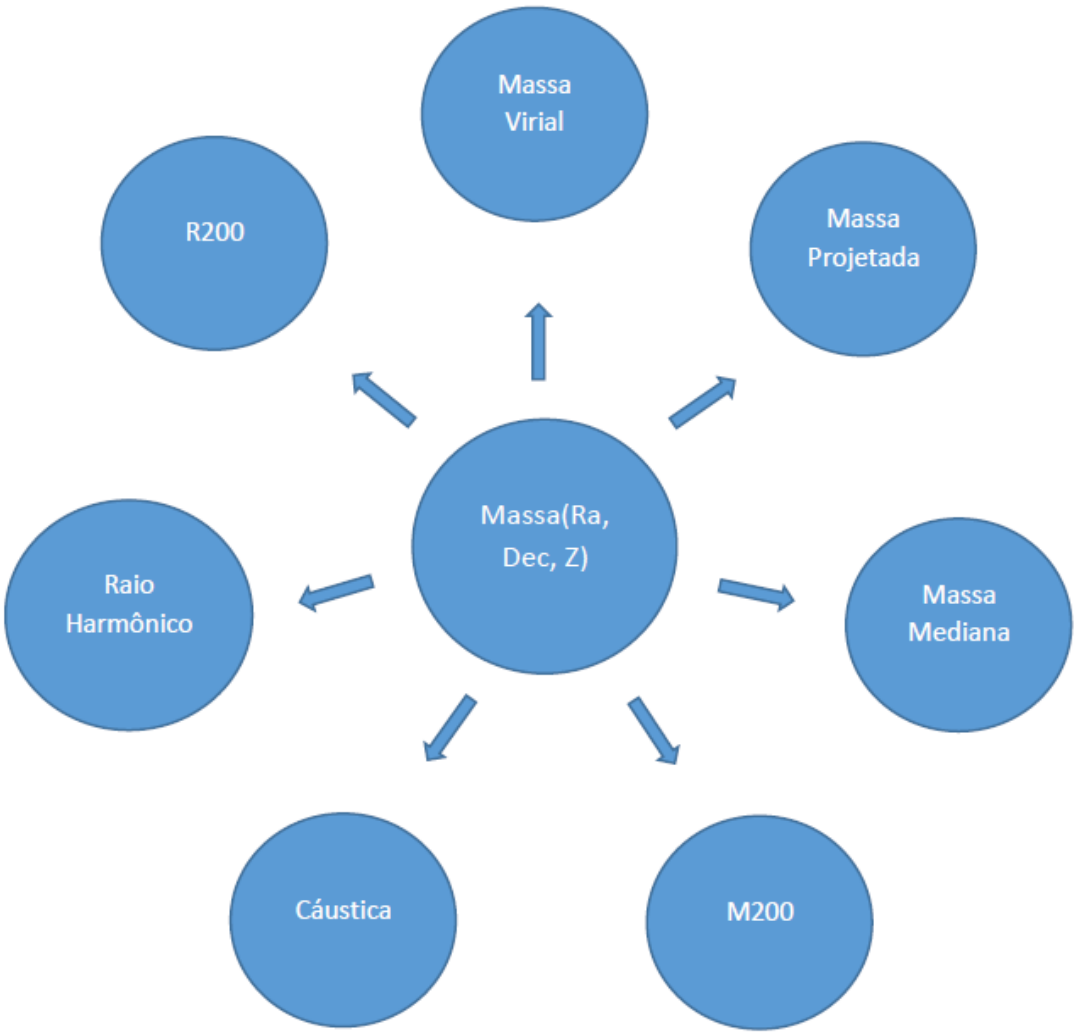

Figura 25 – Fluxograma mostrando os métodos para obtenção de massa disponíveis na função `massa`.

4 Resultados

Neste capítulo apresentamos os resultados da aplicação de nosso pacote sobre o catálogo MOCK composto por 947 aglomerados cuja construção é descrita em Duarte e Mamon (2015) (vide Seção 2.1). Nosso objetivo aqui é ilustrar todo o procedimento sobre um catálogo completo de aglomerados, uma vez que a automatização dos métodos descritos nos capítulos anteriores destina-se a aplicações em grandes volumes de dados. Além disso, o catálogo MOCK, uma vez que é construído, permite com que saibamos quem são seus membros, seus raios, suas massas. Portanto, podemos quantificar os erros que são cometidos ao longo da análise feita com as funções que nosso pacote disponibiliza. Esta verificação de erros possibilita ao usuário efetuar escolhas eficientes de uso do pacote.

4.1 Resultados da remoção de `outliers`

Nesta seção mostramos uma análise comparativa de desempenho das rotinas de remoção de `outliers`. Para efeito de comparação, definimos duas quantidades importantes: a completeza e a pureza do catálogo final, ou seja, a lista de objetos considerados membros, após a remoção dos intrusos. A completeza é definida pela razão do número de membros identificados por um método particular em relação aos membros dados *a priori*, pela construção do catálogo MOCK. A pureza é definida pela razão do número de membros corretos em relação ao catálogo MOCK. Por exemplo, se temos para um determinado aglomerado MOCK 100 membros, e um dos métodos obteve 90 membros, então ele é 90% completo. Mas, se apenas 80 eram membros idênticos aos do MOCK, então ele é 80% puro. Um objeto muito completo terá a propriedade riqueza (que dependende essencialmente do número de galáxias associadas a um aglomerado) bem estimada. Contudo se, ao mesmo tempo ele for impuro, terá as propriedades dinâmicas mal estimadas. Por exemplo, sua distribuição de velocidades pode estar comprometida e, em consequência, todos os cálculos dela decorrentes. Um estudo recente de Aguena e Lima (2016) discute as várias implicações sobre os efeitos de incompleteza e impureza em aglomerados de galáxias.

4.1.1 Completeza

Para a completeza, portanto, comparamos quantas galáxias foram selecionadas como membros em relação ao "gabarito" do catálogo `MOCK`. Para melhor avaliar o desempenho dos métodos de remoção, fizemos um gráfico comparando as completezas em função da massa dos aglomerados.

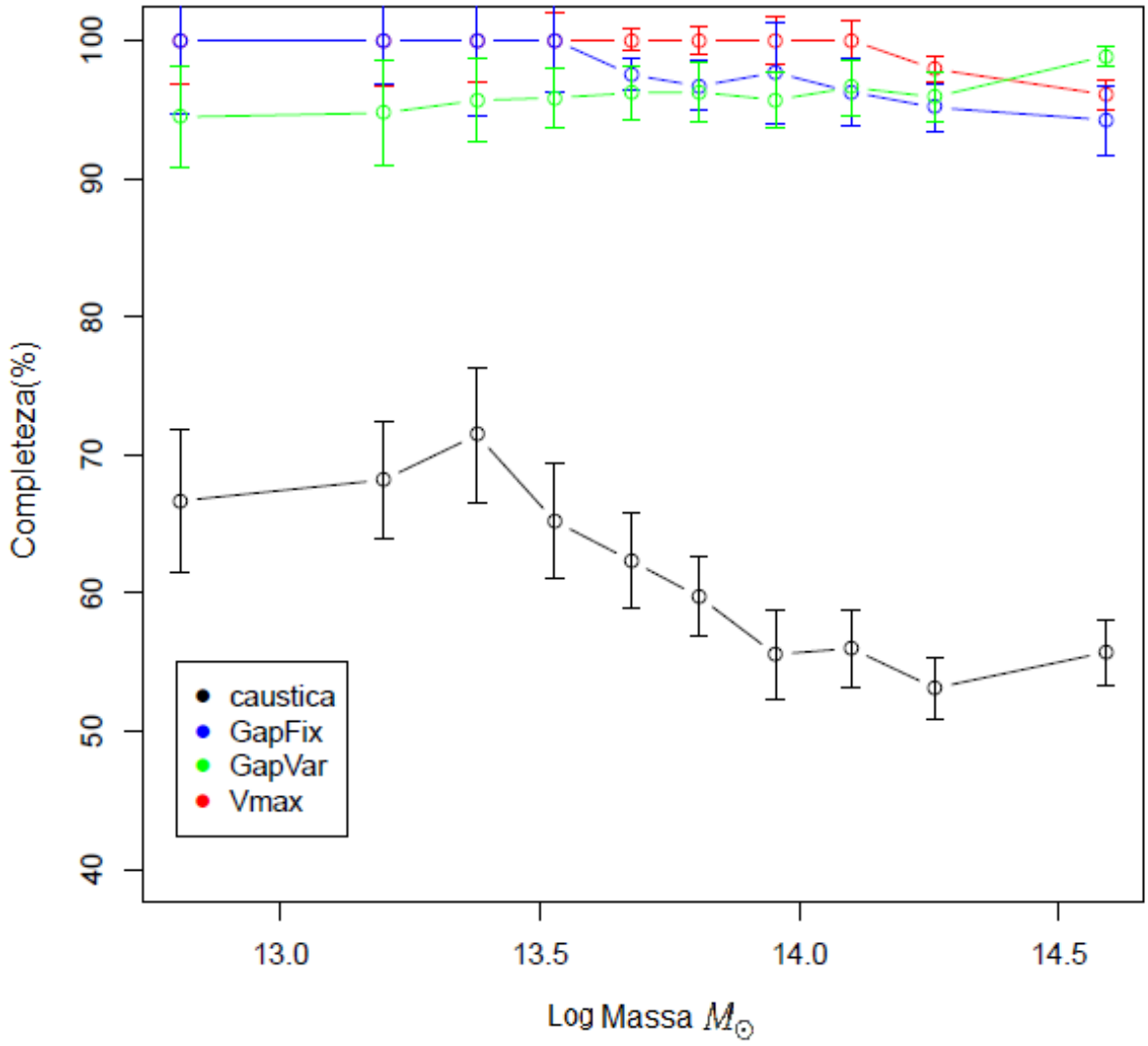

Figura 26 – Gráfico que compara a completeza dos métodos de remoção de `outliers` em função do logaritmo da massa dos aglomerados dada em unidade de massas solares.

Na Figura 26 vemos que os métodos **G**apFix, **G**apVar e **V**max possuem completeza acima de 90% para todo o intervalo de massa disponível, com uma ligeira queda de desempenho para altas massas. Por outro lado, o método da **C**austica tem completeza entre 55% e 70%, exibindo um comportamento que sugere uma piora de desempenho quando avançamos para massas maiores. Este resultado aparentemente pior da cáustica

tem que ser analisado em conjunto com o seu desempenho em termos de pureza, que será apresentado a seguir.

4.1.2 Pureza

No teste de pureza, verificamos quantas galáxias os métodos selecionam como membros que de fato são membros de acordo com o `MOCK` (palavra designada para catálogos construídos artificialmente). O resultado é apresentado na Figura 27. Neste caso, percebe-se que todos os métodos apresentam desempenhos melhores à medida que as massas aumentam. Contudo, neste quesito a **C**áustica é quem possui o melhor desempenho, com valores de pureza sempre acima de 90%, enquanto os demais métodos somente atingem este patamar para massas maiores que 10^{14} $M_{\odot}$ (Massa Solar). Por outro lado, para massas menores que $10^{13.5}$ $M_{\odot}$, os métodos **G**apVar e **V**max apresentam valores de pureza abaixo de 50%, significando que são bastante contaminados, no caso de aglomerados menos massivos.

Não apresentamos nesta seção nenhuma combinação dos métodos de remoção. Além de haver um número relativamente grande de combinações possíveis, este ainda seria aumentado significativamente a cada etapa, na sequência da análise, quando novas opções são oferecidas. Desta forma, o objetivo nesta seção se restringe apenas a identificar a perfomance individual dos métodos que podem ser empregados. Este conhecimento ajudará o usuário a fazer suas escolhas e possíveis combinações de métodos.

Em seguida, já conhecendo o nível de completeza e pureza de cada método de remoção, verificaremos como as escolhas feitas nesta etapa podem afetar a caracterização dinâmica dos sistemas.

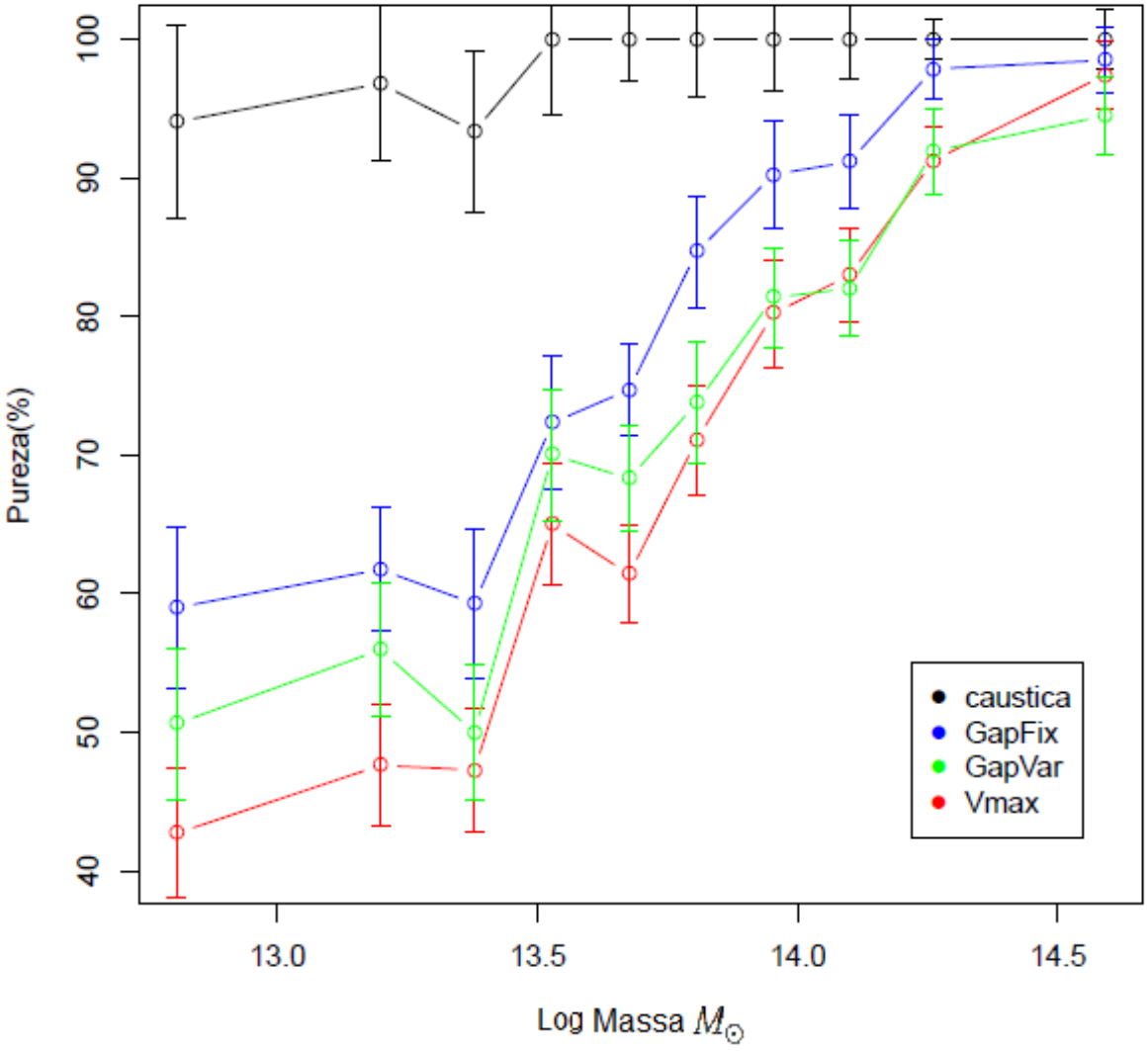

Figura 27 – Gráfico representando a comparação da pureza dos resultados dos métodos de remoção de `outliers`.

4.2 Análise Dinâmica

Finalizada a primeira etapa do `pipeline`, podemos realizar a análise dinâmica dos sistemas. Como o objetivo é fazer uma comparação dos diferentes métodos, procedemos da seguinte maneira: executamos inicialmente **G**alClus (incluindo todos os métodos disponíveis: GNG, GNG_MIN e DS) para os membros do `MOCK` e, em seguida, para os membros definidos por cada método de remoção. Então, os diagnósticos são comparados.

Vimos nas seções 2.4 e 3.2.4 que os métodos usados em `GalClus` deve avaliar a dinâmica do aglomerado dentro de um raio onde o equilíbrio é esperado, sendo que diversos testes feitos por Ribeiro et al. (2013) sugerem que pelo menos 20 galáxias são

necessárias para que resultados confiáveis sejam obtidos com o GNG. Nesta aplicação, consideramos este limite para que os dados sejam analisados.

No caso dos aglomerados MOCK, dos 947 inicialmente presentes no catálogo, apenas 630 possuem 20 ou mais galáxias membro dentro do raio R_{200} dado pelo catálogo. Executamos `GalClus` para estes sistemas e definimos o que seria o "gabarito" para cada método de análise, dado que antes um método de remoção tenha sido executado. Em seguida, executamos novamente `GalClus` agora seguindo as regras de análise definidas nos capítulos anteriores. Portanto, simulamos desconhecer R_{200}, e usamos o raio harmônico dos membros definidos pelos métodos de remoção. Os resultados referentes ao gabarito estão reunidos na Tabela 16, onde apresentamos a fração de aglomerados em cada classificação (diagnóstico) dada pelo programa.

Devemos notar nesta tabela a considerável amplitude de valores em cada diagnóstico possível. Destacamos o fato de que DS produz a maior fração de casos indefinidos, AD a maior fração de sistemas gaussianos e também a maior fração de sistemas não-gaussianos, possivelmente uma razão para esse resultado seria o fato que o teste DS trabalha com uma dimensionalidade mais alta potencializando a frequência de erros enquanto o teste AD tem uma arquitetura conservadora visando cometer o menor tipos de erros tipo 1. Contudo, seguiremos o estudo de Ribeiro et al. (2013) e consideraremos, entre os métodos disponibilizados por `GalClus`, HD como sendo o mais confiável. Esta escolha facilitará a comparação com a execução de `GalClus` sobre membros obtidos pelos métodos de remoção e raio de análise definido pelo raio harmônico. Os resultados são apresentados nas Tabelas 17 18 19 e 20.

Na Tabela 17 vemos os resultados para `Galremov` $\rightarrow$ `GalClus` quando o método de remoção é o da Cáustica. Antes de discutir esta tabela, um ponto a se considerar é que o número de sistemas com pelo menos 20 galáxias dentro do raio harmônico diminui, passando para 548. Isto se deve ao fato de que a remoção via cáustica é a menos completa entre todos os métodos, diminuindo as amostras de galáxias em cada aglomerado. Os resultados indicam que a maior fração de indefinidos continua sendo a do teste DS, enquanto novamente o teste AD gerou a maior fração de sistemas gaussianos e também a maior fração de sistemas não-gaussianos. No que se refere ao HD, que vamos tomar como o indicador mais confiável de não-gaussianidade, ocorre uma pequena redução de casos G (Gaussianos) e NG (Não Gaussianos), e um aumento de casos indefinidos. A flutuação geral de resultados é de 5,3%, sendo um pouco menor no que refere a HD, 4,7%. Interessante notar que a incompleteza das amostras parece não causar diferenças muitos grandes em relação a resultado esperado.

Na Tabela 18 vemos os resultados para `Galremov` $\rightarrow$ `GalClus` quando o método de remoção é o Vmax. Neste caso o número de sistemas com pelo menos 20 galáxias dentro do raio harmônico aumenta, passando para 803. Isto se deve ao fato de

que a remoção via Vmax ser a mais completa (e também a mais impura) entre todos os métodos, aumentando o tamanho da amostra de galáxias em cada aglomerado.

Notemos na Tabela 18 que agora todos os métodos tiveram uma redução da fração de sistemas em equilíbrio. Isto pode resultar das amostras contaminadas (embora completas) produzidas por Vmax. No total, em todos os métodos há um aumento de casos NG e indefinidos, que parecem favorecidos pela impureza das amostras. A flutuação geral é de 16,8%, e em relação a HD, nosso método-referência, a flutuação é de 14,7%.

Na Tabela 19 vemos os resultados para `Galremov` $\rightarrow$ `GalClus` quando o método de remoção é o Gap Fixo. Neste caso o número de sistemas com pelo menos 20 galáxias dentro do raio harmônico também aumenta com relação ao gabarito, passando para 687 (um aumento menor que no caso anterior). Isto também está associado ao fato de que o método do Gap Fixo também produzir amostras muito completas, porém menos impuras que o Vmax.

Na Tabela 19 vemos que, de modo geral, resultados razoavelmente semelhantes àqueles encontrados no gabarito. Em média, ocorre uma flutuação de 4,3% considerando todos os resultados, sendo a flutuação em relação a HD de apenas 2,7%. Além disso, devemos notar que todas as tendências observadas no gabarito foram mantidas agora. Este resultado pode indicar um certo equilíbrio entre completeza e pureza alcançado por este método.

Finalmente, na Tabela 20 vemos os resultados para `Galremov` $\rightarrow$ `GalClus` quando o método de remoção é o Gap Variável. Neste caso o número de sistemas com pelo menos 20 galáxias dentro do raio harmônico também aumenta, passando para 745 (um aumento menor que no caso do Vmax porém maior que no caso do Gap Fixo). Isto provavelmente está relacionado à alta completeza do método associada com uma impureza maior que no caso do Gap Fixo e menor do que no caso do Vmax.

Tabela 16 – Catálogo MOCK - Gabarito.

Classe	HD	MCLUST	AD	DS
0	50%	36%	67%	21%
1	19%	15%	31%	6%
2	31%	48%	2%	63%

Tabela 17 – Catálogo MOCK - Remoção via Cáustica.

Classe	HD	MCLUST	AD	DS
0	47%	35%	61%	18%
1	15%	10%	28%	5%
2	38%	55%	11%	77%

Tabela 18 – Catálogo MOCK - Remoção via Vmax.

Classe	HD	MCLUST	AD	DS
0	28%	10%	35%	11%
1	30%	36%	50%	23%
2	42%	64%	15%	66%

Tabela 19 – Catálogo MOCK - Remoção via Gap Fixo.

Classe	HD	MCLUST	AD	DS
0	48%	25%	62%	18%
1	23%	20%	31%	12%
2	29%	55%	7%	70%

Tabela 20 – Catálogo MOCK - Remoção via Gap Variável.

Class	HD	MCLUST	AD	DS
0	43%	30%	60%	15%
1	28%	22%	37%	18%
2	29%	48%	3%	67%

Na Tabela 20 vemos que os resultados, mais uma vez, flutuam em torno daqueles obtidos no "gabarito". Desta vez, a flutuação média é de 5,6%, sendo um pouco maior para HD, 6%. Sendo um resultado melhor que o Vmax, mas pior que os métodos da cáustica e do Gap Fixo.

Conclusão

Estes resultados sugerem que a sequência do melhor para o pior método de remoção seria esta:

Gap Fixo $\rightarrow$ Cáustica $\rightarrow$ Gap Variável $\rightarrow$ Vmax

embora nossas comparações tenham um caráter generalista (número de aglomerados com determinado diagnóstico) e não específico (fração de sistemas cuja classificação coincide com a do gabarito). Esta comparação mais detalhada não foi feita em virtude de o número de objetos considerados por cada método (aqueles que possuem 20 galáxias dentro do raio harmônico) variar de caso a caso, dificultando as comparações finais.

4.3 Comparando Massas e Raios

Para realizar comparações entre massas e raios do catálogo MOCK, ou seja, que conhecemos *a priori*, e massas e raios oriundos das amostras de membros geradas por cada método de remoção de `outliers`, utilizamos histogramas para a visualização das distribuições, assim como os testes estatísticos t de Student (para comparação de

médias) e de Kolmogorov-Smirnov (KS) (para as distribuições), para caracterizar as semelhanças e dessemelhanças em cada caso. O teste t bilateral da média é usado para testar a hipótese de que a média gerada pelos métodos é igual à média do gabarito. Já o teste KS é usado para testar se as distribuições dos métodos e do gabarito (para massas e raios) diferem signficativamente ou não. Ambos os testes são amplamente usados em todas as áreas do conhecimento. Para maiores detalhes, veja Kanji (2006). Os testes estão diponíveis na distribuição básica do R e podem ser executados através de `t.test` e `ks.test`, respectivamente.

Os resultados de todos estes testes estão organizados nas Tabelas 21 a 28.

4.3.1 Comparação de Raios

Uma vez que o catálogo MOCK disponibiliza o valor de R_{200} como *proxy* (substituto) do raio virial, fizemos comparações dos estimadores disponíveis em `MASSA` em relação a este valor catalogado. Utilizamos os estimadores seguindo-se aos quatro métodos de remoção de `outliers`, independentemente da classificação dinâmica obtida em **G**alClus.

Nos painéis da Figura 28, onde comparamos os raios harmônicos estimados para os sistemas, vemos que as amostras geradas por Vmax, Gap Fixo e Gap Variável sistematicamente subestimam os raios dos aglomerados. Os testes t e KS rejeitam a hipótese de que as médias e distribuições de raio sejam retiradas de uma mesma população a um nível de 99% de confiança (vide Tabelas 21 a 23). Ao mesmo tempo, os testes não rejeitam a semelhança entre as médias e distribuições de raios gerada pela cáustica e a do catálogo, a um nível de 95% de confiança (vide Tabela 24). A comparação favorece, portanto, a remoção feita pela cáustica. Este resultado vem a ser reforçado pela comparação dos métodos para o estimador de R_{200}. Na Figura 29, vemos que, mais uma vez, apenas a remoção feita pela cáustica gera uma distribuição de raios, agora estimados pelo R_{200}, semelhante à distribuição do gabarito (vide também a Tabela 24).

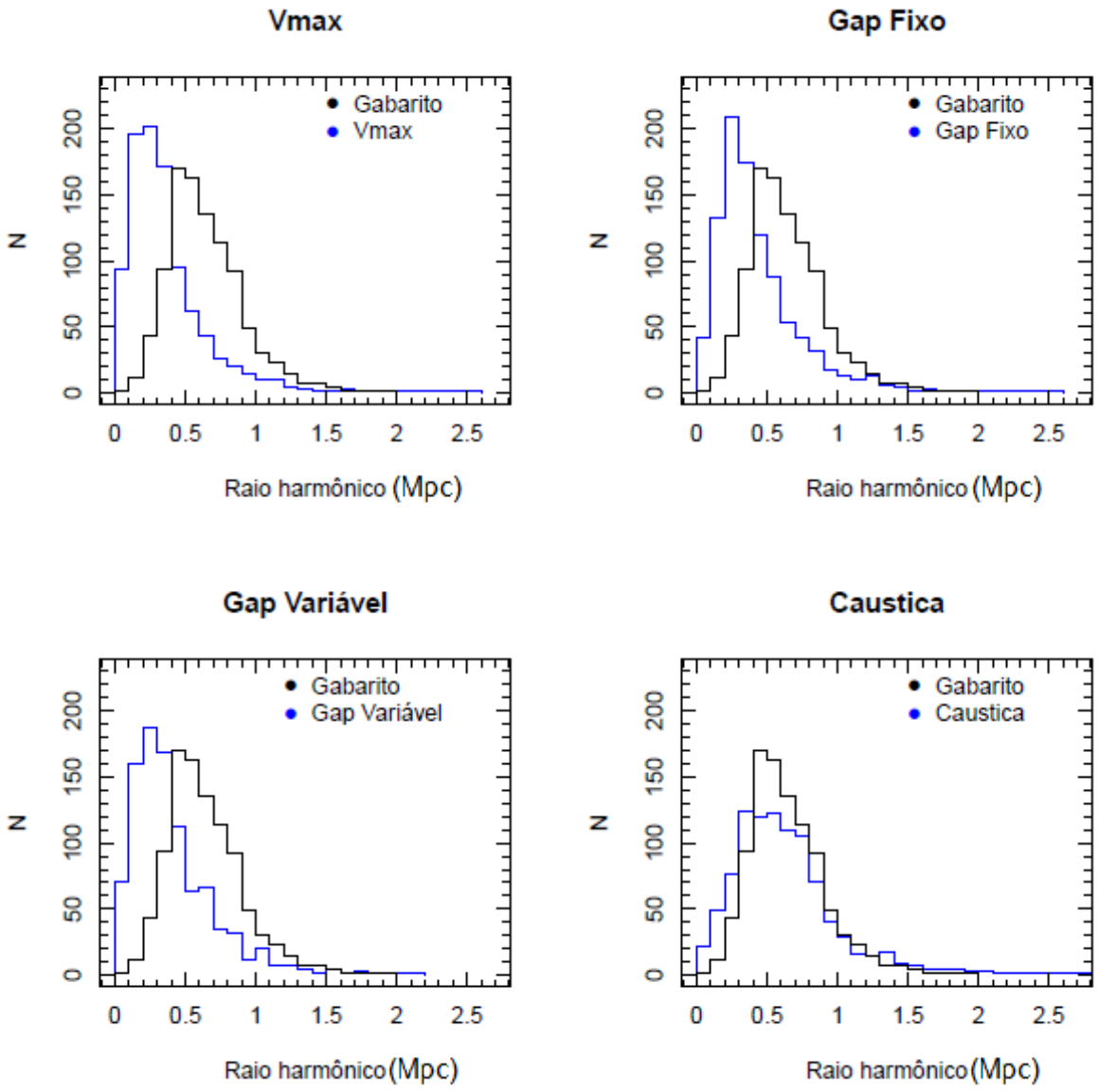

Figura 28 – Histogramas comparando os Raios Harmônicos provenientes dos resultados dos métodos de remoção de `outliers`.

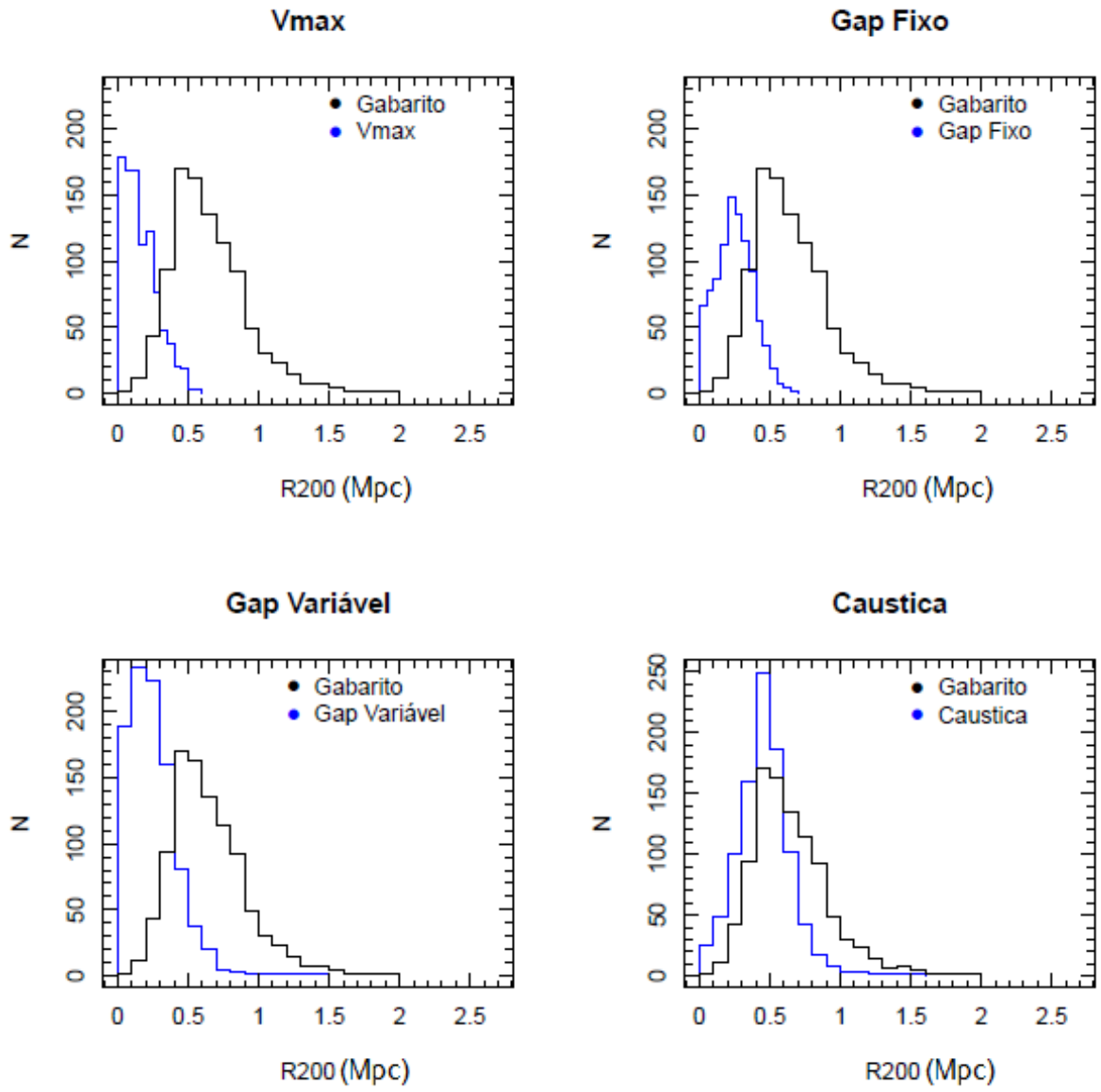

Figura 29 – Histogramas comparando os R200 provenientes dos resultados dos métodos de remoção de `outliers`.

Tabela 21 – Resultados dos testes t e KS para comparação de raios, usando o método Vmax para remover `outliers`.

Raio	p-valor (t)	p-valor (KS)
R_H	< 0.001	< 0.001
R_{200}	< 0.001	< 0.001

Tabela 22 – Resultados dos testes t e KS para comparação de raios, usando o método Gap Fixo para remover `outliers`.

Raio	p-valor (t)	p-valor (KS)
R_H	< 0.001	< 0.001
R_{200}	< 0.001	< 0.001

Tabela 23 – Resultados dos testes t e KS para comparação de raios, usando o método Gap Variável para remover `outliers`.

Raio	p-valor (t)	p-valor (KS)
R_H	< 0.001	< 0.001
R_{200}	< 0.001	< 0.001

Tabela 24 – Resultados dos testes t e KS para comparação de raios, usando o método da Cáustica para remover `outliers`.

Raio	p-valor (t)	p-valor (KS)
R_H	0.235	0.142
R_{200}	0.417	0.231

4.3.2 Comparação de Massas

De forma semelhante, comparamos a massa M_{200} do gabarito com as distribuições de massa das amostras de membros dos aglomerados obtidas por cada método de remoção. No caso da massa virial, os testes t e KS (relacionados nas Tabelas 25 a 28) rejeitam, a um nível de confiança de 99%, a hipótese de que as amostras de membros geradas por todos os métodos produzam médias e distribuições de massa semelhantes à distribuição de massas catalogadas. Na Figura 30 vemos que os métodos Vmax, Gap Fixo e Gap Variável superestimam a massa, enquanto a cáustica a subestima. A inspeção visual desta figura sugere que os métodos menos confiáveis para este estimador seriam o Vmax e o Gap Variável.

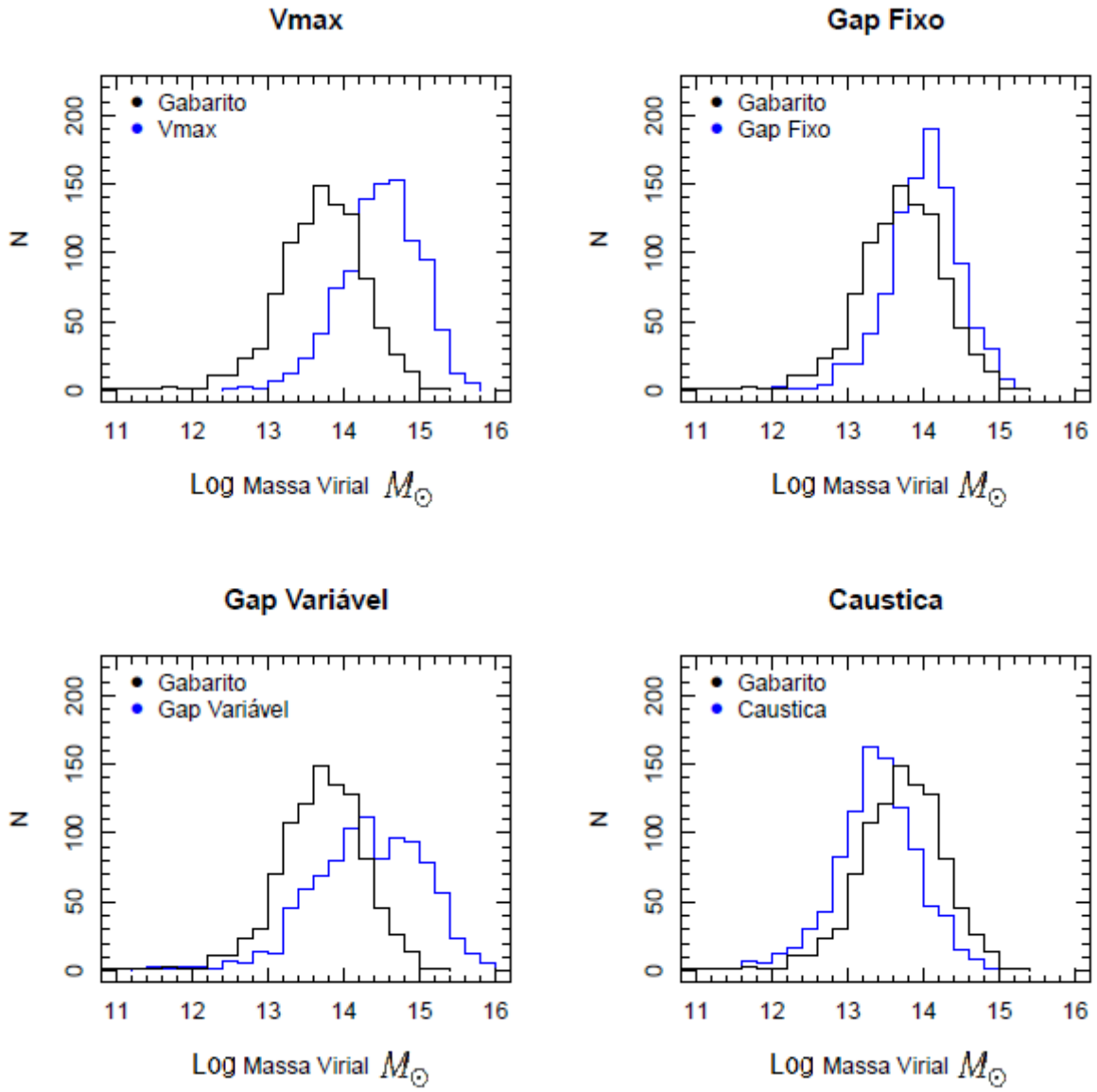

Figura 30 – Histogramas comparando as massas viriais provenientes dos resultados dos métodos de remoção de `outliers`.

Considerando agora o estimador de massa projetada (vide Figura 31) (para órbitas isotrópicas), o teste KS rejeita a hipótese, a um nível de confiança de 99%, que as amostras originadas em cada caso possam ter sido retiradas da mesma distribuição de massas do gabarito (vide Tabelas 25 a 28). Contudo o teste t não rejeita a hipótese de que a média da amostra gerada pela cáustica seja semelhante à média das massas do catálogo MOCK, a um nível de confiança de 90% (vide Tabela 28). Os demais métodos superestimam a média. Este é mais um resultado que favorece o método de remoção da cáustica. Este resultado é muito próximo do que encontramos para o estimador de massa mediana (vide Figura 32), exceto pelo fato de que agora, para a distribuição de massas gerada após a remoção de `outliers` feita pelo método da cáustica, ambos os

testes t e KS não rejeitam as hipóteses de que as distribuições comparadas venham da mesma população e tenham mesma média, a um nível de confiança de 90% (vide Tabela 28).

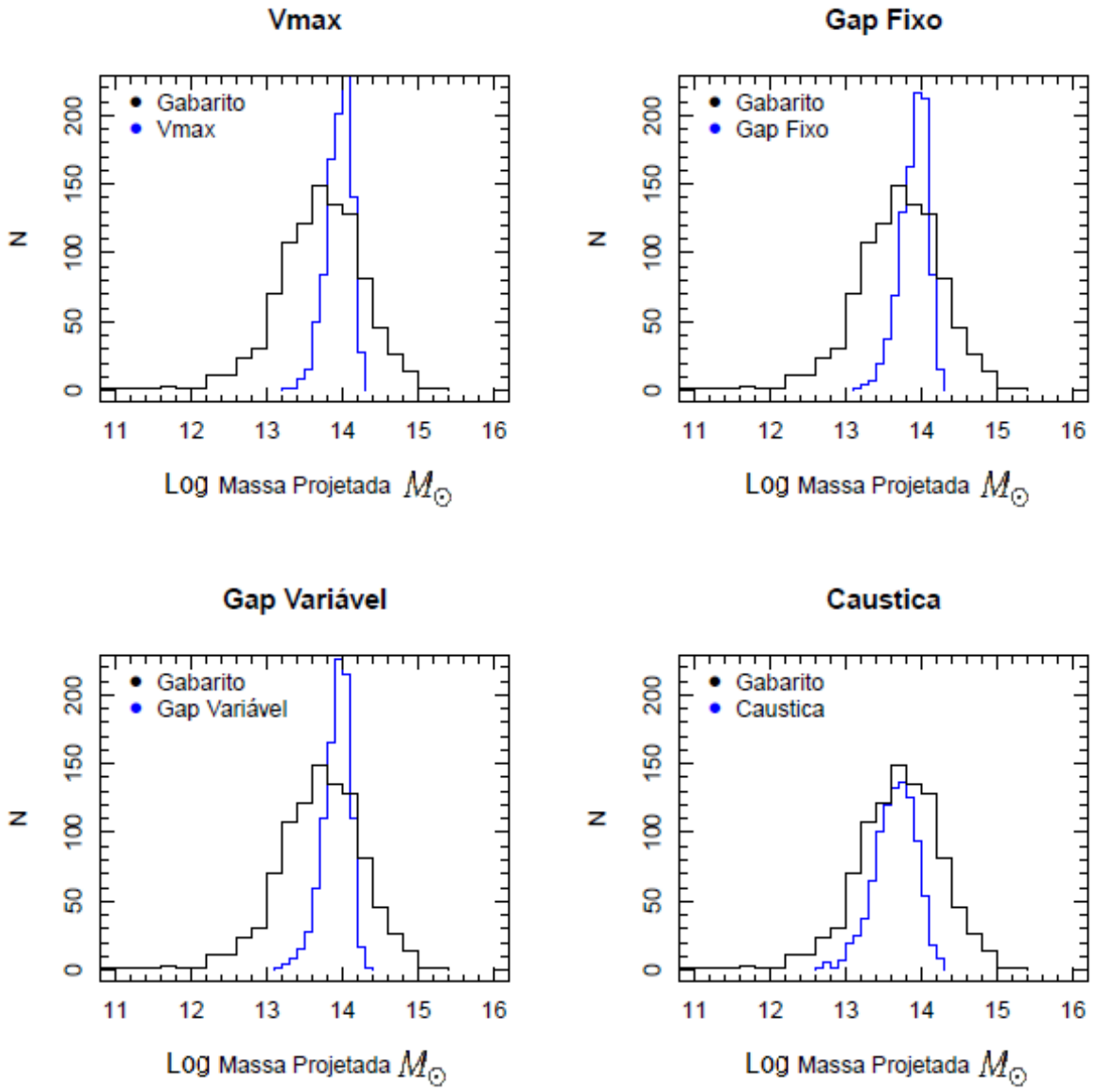

Figura 31 – Histogramas comparando as massas projetadas provenientes dos resultados dos métodos de remoção de `outliers`.

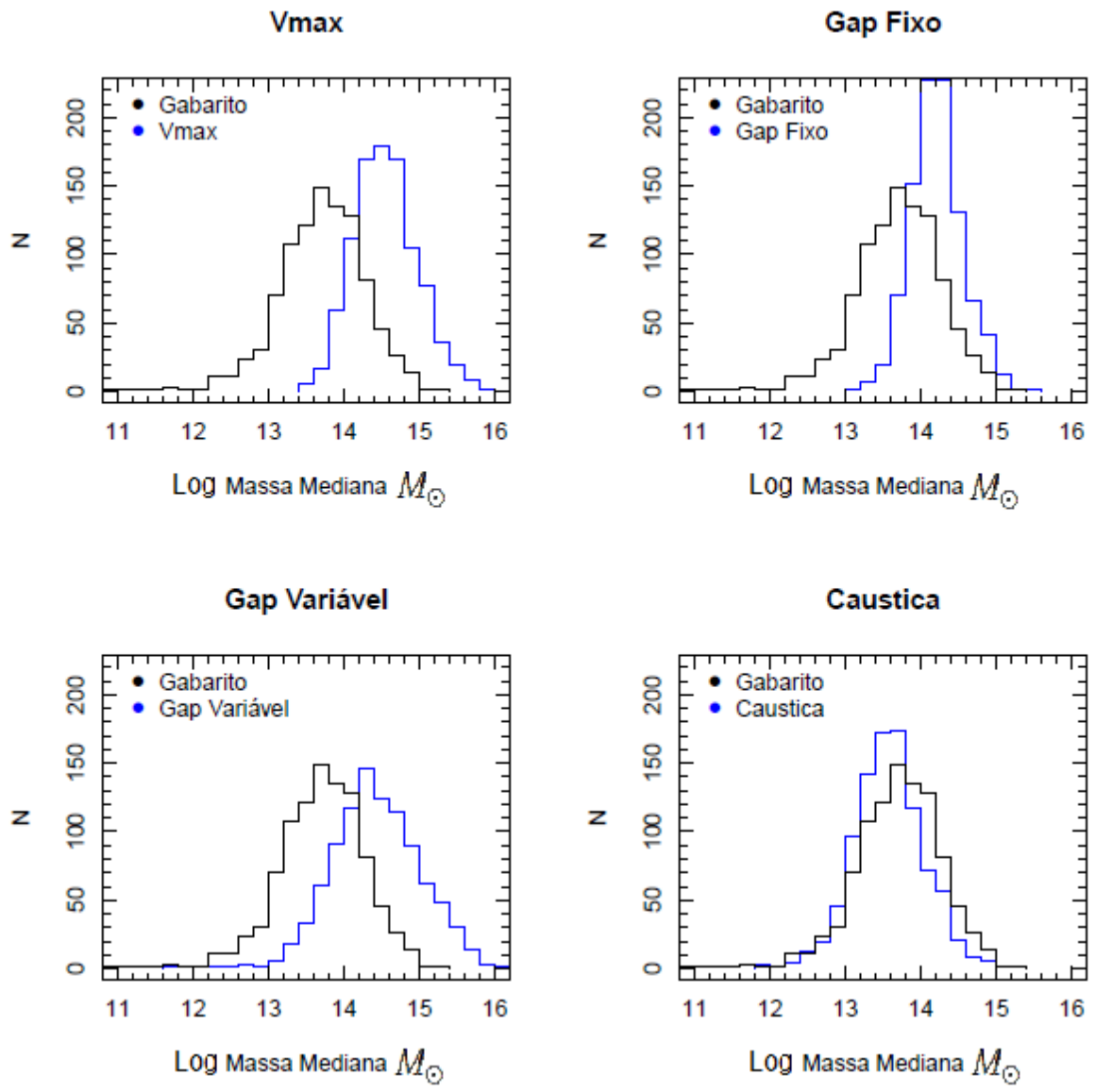

Figura 32 – Histogramas comparando as massas medianas provenientes dos resultados dos métodos de remoção de `outliers`.

Tabela 25 – Resultados dos testes t e KS para comparação de massas, usando o método Vmax para remover `outliers`.

Raio	p-valor (t)	p-valor (KS)
M_V	< 0.001	< 0.001
M_P	< 0.001	< 0.001
M_M	< 0.001	< 0.001
M_{200}	< 0.001	< 0.001

Continuando a análise, para o estimador de M_{200}, o método com melhor performance é o Gap Fixo, cujas amostras resultantes geram massas cuja distribuição é

Tabela 26 – Resultados dos testes t e KS para comparação de massas, usando o método Gap Fixo para remover `outliers`.

Raio	p-valor (t)	p-valor (KS)
M_V	< 0.001	< 0.001
M_P	< 0.001	< 0.001
M_M	< 0.001	< 0.001
M_{200}	0.061	0.087

Tabela 27 – Resultados dos testes t e KS para comparação de raios, usando o método Gap Variável para remover `outliers`.

Raio	p-valor (t)	p-valor (KS)
M_V	< 0.001	< 0.001
M_P	< 0.001	< 0.001
M_M	< 0.001	< 0.001
M_{200}	< 0.001	< 0.001

Tabela 28 – Resultados dos testes t e KS para comparação de raios, usando o método da Cáustica para remover `outliers`.

Raio	p-valor (t)	p-valor (KS)
M_V	< 0.001	< 0.001
M_P	0.098	< 0.001
M_M	0.091	0.089
M_{200}	< 0.001	< 0.001

semelhante à distribuição do gabarito e também possuem a mesma média das massas do catálogo MOCK, a um nível de 95% de confiança, de acordo com os testes t e KS, respectivamente (vide Tabela 26). Neste caso, Os métodos Vmax e Gap Variável superestimam a média, enquanto o método da cáustica a subestima (vide Figura 33).

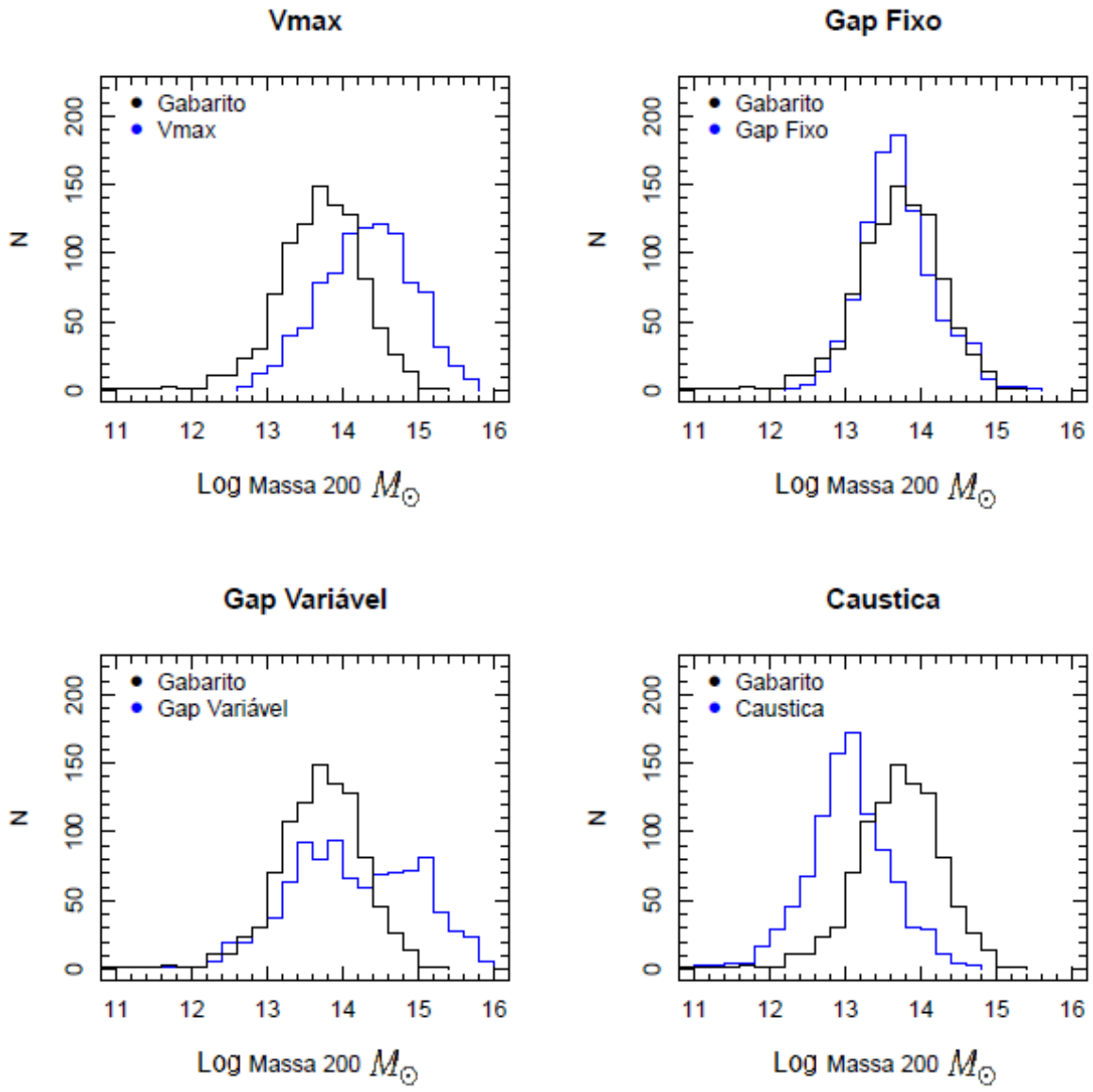

Figura 33 – Histogramas comparando as massas M200 provenientes dos resultados dos métodos de remoção de `outliers`.

4.3.3 Considerações Finais

Os resultados discutidos neste capítulo indicam claras vantagens do método da cáustica e do Gap Fixo para gerar os membros do aglomerado. Verificamos isto tanto na comparação de raios como massas. Ambos os estimadores de raio produzem melhores resultados quando operam sobre as galáxias membro após a remoção de `outliers` efetuada pela cáustica. Para os estimadores de massa, a performance geral dos métodos é pior. Mas a cáustica tem melhor resultado para as massas projetada e mediana, enquanto o Gap Fixo obtem melhores resultados para o estimador de M_{200}.

O trabalho como um todo será discutido no próximo capítulo, sendo algumas perspectivas de desenvolvimentos futuros apresentadas em seguida.

5 Discussão e Perspectivas

5.1 Aspectos gerais

No trabalho realizado e apresentado nesta dissertação, desenvolvemos e conectamos uma série de rotinas em R com o principal objetivo facilitar e otimizar uma análise frequente e bastante específica da astrofísica extragaláctica, que se refere (i) à identificação da fração de galáxias na região de um aglomerado que tenham maior probabilidade de pertencer ao aglomerado dinamicamente; e (ii) à obtenção de algumas informações sobre esses sistemas, como o seu estado dinâmico e propriedades globais como raio e massa. A primeira etapa consiste em uma modelagem astroestatística (no sentido em que a astrofísica determina uma amostra estatística que será utilizada na análise subsequente) que depende da escolha dos métodos. Vimos ao longo do trabalho que um aglomerado pode ter aparências bastante distintas dado que realizamos a etapa de remoção de `outliers` de diferentes maneiras (vide Capítulos 2, 3 e 4). A segunda etapa também tem forte caráter astroestatístico, uma vez que congrega métodos tanto astrofísicos como estatísticos para extrair informações sobre o equilíbrio do sistema e suas propriedades. O trabalho se insere na era do *big data*, onde dados em quantidades cada vez maiores desafiam os pesquisadores em termos de organização, manipulação e análise dos dados.

Ao longo da execução do projeto foram desenvolvidas várias rotinas para trabalhar com cada etapa do estudo desses objetos. As etapas estudadas foram a remoção de `outliers`, a análise dinâmica e a obtenção de massa e raio. Como podemos ver no capítulo 2, para cada uma dessas etapas foram escolhidas abordagens distintas, possibilitando maior flexibilidade no objetivo do estudo (vide capítulo 2). O pipeline final integra as várias rotinas de forma eficiente e abrangente em termos de escolhas. O conjunto de rotinas implementadas em R possibilita ao pesquisador realizar um estudo simples de executar, mas sendo bem detalhado e flexível em cada um de seus estágios. Esta característica do programa foi atingida com a definição de funções que unificam os diferentes métodos, padronizando suas entradas e saídas de tal forma que as várias rotinas possam ser executadas em conjunto com considerável fluidez.

Para a execução do programa do início ao fim para o conjunto de dados do catálogo MOCK (947 aglomerados), o tempo de máquina (usando um computador pessoal comum[1]) não excedeu um total de aproximadamente 12 horas em modo serial. As etapas mais custosas são aquelas incluídas em `GalClus`, uma vez que cada cálculo é repetido 1000 vezes para gerar os fatores de confiabilidade (vide Capítulos 2 e 3).

[1] Processador de 2 núcleos, 3,4 GHz, 8 MB de cache, Memória RAM de 4 GB, padrão DDR3 1333 MHz.

Este aspecto da performance do código (assim como os demais resultados referentes à análise do catálogo MOCK apresentados no Capítulo 4) sugerem que nosso pacote em R será competitivo e útil para a pesquisa em astrofísica extragaláctica.

Em seguida, discutiremos os resultados obtidos no Capítulo 4.

5.2 Principais resultados

Nesta seção enfatizaremos alguns resultados obtidos no capítulo 4 e sobre algumas peculiaridades dos métodos.

Remoção de `Outliers`

Na remoção de `outliers` destacamos o melhor desempenho dos métodos da cáustica (no quesito pureza) e do Gap Fixo, por ter mostrado uma combinação entre completeza e pureza que se mostrou razoavelmente eficiente nas etapas seguintes.

Análise dinâmica

Na etapa de analise dinâmica observamos que os resultados são significativamente dependentes do método de remoção de `outliers`. Isto se dá em virtude de os métodos dependerem da quantidade (e qualidade, isto é, o grau de pureza) das galáxias dentro do raio harmônico do sistema. A comparação entre os resultados por método e o gabarito indicaram que a sequência do melhor para o pior método de remoção seria esta:

Gap Fixo $\rightarrow$ Cáustica $\rightarrow$ Gap Variável $\rightarrow$ Vmax

Massa e Raio

Nesta etapa os resultados também mostraram considerável dependência do método de remoção de `outliers`. Na comparação de raios o único método de remoção de `outliers` que obteve um resultado de raio similar ao do gabarito foi o da Cáustica (tanto para R_H como para R_{200}. No caso da comparação de massas o método da Cáustica apresenta bons resultados para M_P e M_M, enquanto o método do Gap Fixo apresenta melhor resultado para M_{200}.

5.3 Conclusão

É possível fazer uma análise de uma grande amostra de aglomerados, a partir de catálogos observados ou simulados, com rapidez e flexibilidade no uso de ferramentas que prospectam as propriedades desses sistema. Contudo, verifica-se também que, a despeito dos detalhes e da complexidade que cada escolha de métodos possam introduzir nos resultados, a determinação da completeza e, sobretudo, da pureza das

amostras de galáxias em aglomerados é de fundamental importância para a astrofísica de aglomerados, assim como para estudos de cosmologia, como indica o trabalho de Aguena e Lima (2016). Sem dúvida, um grande desafio para pesquisadores nesta área, é o de aumentar a pureza de suas amostras sem reduzir significativamente a sua completeza. A convergência para valores ótimos desses dois fatores será determinante para trabalhos com maior impacto científico nos próximos anos.

5.4 Perspectivas

Como perspectivas para esse projeto temos primeiramente a publicação do código contendo o *pipeline* de análise como um pacote no repositório online do R. O pacote deverá ser acompanhado de manual para o `CRAN` (The Comprehensive R Archive Network - https://cran.r-project.org/) e possivelmente dois artigos científicos: um destinado ao `Journal of Statistical Software` (https://www.jstatsoft.org/index); e outro destinado a uma revista de astrofísica.

Logo após a liberação do pacote para uso, pretendemos desenvolver uma segunda versão do código adicionando novas funcionalidades, acrescentando mais métodos em cada uma das etapas, otimizando suas saídas gráficas, inserindo técnicas de programação paralela em mais métodos.[2].

Contudo, a principal expansão do pacote seria a adição de uma nova etapa dentro do *pipeline* que seria a obtenção da função de massa de aglomerados de galáxias. Esta é uma função de extrema importância para a ligação entre a astrofísica de aglomerados e a cosmologia. Uma vez que a função de massa depende basicamente da distribuição de massas de aglomerados oriundos de um determinado levantamento, o pacote na forma atual já produziria um *input* significativo para a nova função `FMass`.

A seguir, apresentamos alguns pontos sobre a função de massa que pretendemos anexar ao código atual.

Função de Massa - o que é?

A função de massa de aglomerados de galáxias é uma descrição estatística importante da população de aglomerados de galáxias tanto do ponto de vista astrofísico como cosmológico. Em termos simples, ela calcula a densidade numérica de aglomerados por intervalo de massa e por volume numa dada época do universo. Originalmente, o cálculo da função de massa foi proposto por Press e Schechter (1974), usando uma estatística gaussiana para descrever o campo primordial de flutuação de densidades do universo. Outras formulações foram propostas ao longo do tempo (vide Padmanabhan 1993).

[2] Atualmente, apenas a rotina GNG permite a execução em paralelo.

Dada sua forte dependência com os parâmetros cosmológicos, a função de massa pode ajudar a estabelecer limites sobre a natureza do espectro primordial das flutuações e pode ser usada até mesmo como um teste da teoria da gravitação. Ela também pode impor restrições ao parâmetro que controla a densidade da energia escura através de comparações da função de massa teórica com os dados observacionais. Com a função de massa é possível ainda estimar a taxa de fusão dos halos virializados, o tempo de formação e o tempo de vida dessas estruturas (vide Borgani 2008 para uma boa revisão sobre função de massa de aglomerados). Por todas estas razões, implementar a função de massa é aumentar a aplicabilidade e a utilidade do código que estamos desenvolvendo.

Resumidamente, a função de massa tem como objetivo descrever o número de aglomerados dentro de um certo intervalo de massa por unidade de volume comóvel. Ela pode ser derivada analiticamente, como no trabalho de PS (Press & Schechter) , mas mais recentemente ela vem sendo medida a partir de simulações numéricas de grandes volumes no universo (vide Jenkins et al. 2001). A FM é uma função decrescente e cai rapidamente com valores de massa mais altos (aglomerados de massa muito alta são extremamente raros), conforme vemos na Figura 34.

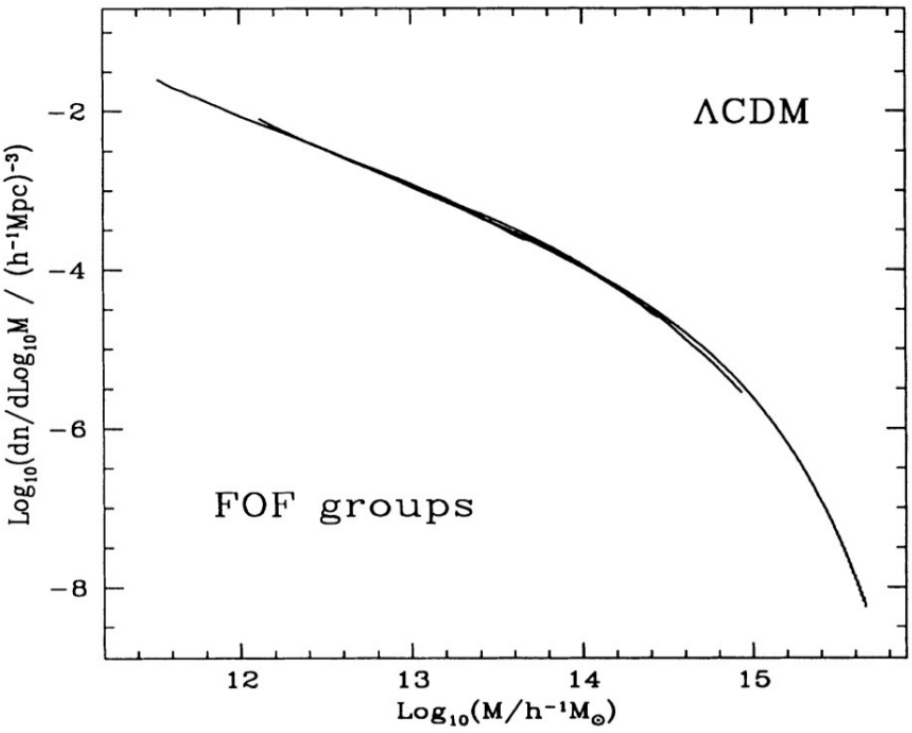

Figura 34 – Função de massa diferencial para halos de matéria escura em simulações ΛCDM.

Fonte: (Jenkins et al., 2001)

Calculando a função de massa

Do ponto de vista operacional, o processo para a construção da função de massa de aglomerados de galáxias a partir de dados observados requer basicamente os seguintes estágios:

1. Detectar e contar os aglomerados em levantamentos de galáxias.
2. Determinar o volume amostrado.
3. Estimar a massa dos aglomerados.
4. Organizar os dados para o ajuste da função de massa.

Portanto, pelo menos dois passos do trabalho já estão implementados na versão atual do código (os passos 1 e 3). Precisaríamos incluir os passos 2 e 4 para termos a função de massa para um dado catálogo de aglomerados. Acreditamos ser possível realizar este *upgrade* numa escala de tempo inferior a 12 meses.

Referências

Abell, G. O. The Distribution of Rich Clusters of Galaxies. **apjs**, v. 3, p. 211, maio 1958.

Aguena, M.; Lima, M. Effects of Completeness and Purity on Cluster Dark Energy Constraints. **ArXiv e-prints**, nov. 2016.

Alpaslan, M.; Robotham, A. S. G.; Driver, S.; Norberg, P.; Peacock, J. A.; Baldry, I.; Bland-Hawthorn, J.; Brough, S.; Hopkins, A. M.; Kelvin, L. S.; Liske, J.; Loveday, J.; Merson, A.; Nichol, R. C.; Pimbblet, K. Galaxy And Mass Assembly (GAMA): estimating galaxy group masses via caustic analysis. **mnras**, v. 426, p. 2832–2846, nov. 2012.

Amari, S. **Differential-geometrical Methods in Statistics.** [S.l.]: Lecture Notes in Statistics, 1985.

Bartelmann, M.; Limousin, M.; Meneghetti, M.; Schmidt, R. Internal Cluster Structure. **ssr**, v. 177, p. 3–29, ago. 2013.

Beers, T. C.; Flynn, K.; Gebhardt, K. Measures of location and scale for velocities in clusters of galaxies - A robust approach. **aj**, v. 100, p. 32–46, jul. 1990.

Beers, T. C.; Geller, M. J.; Huchra, J. P. Galaxy clusters with multiple components. I - The dynamics of Abell 98. **apj**, v. 257, p. 23–32, jun. 1982.

Binney, J.; Tremaine, S. Book Review: Galactic dynamics. / Princeton U Press, 1988. **nat**, v. 326, p. 219, mar. 1987.

Biviano, A.; Girardi, M.; Giuricin, G.; Mardirossian, F.; Mezzetti, M. The mass function of nearby galaxy clusters. **apjl**, v. 411, p. L13–L16, jul. 1993.

Borgani, S. Cosmology with Clusters of Galaxies. In: Plionis, M.; López-Cruz, O.; Hughes, D. (Ed.). **A Pan-Chromatic View of Clusters of Galaxies and the Large-Scale Structure**. [S.l.: s.n.], 2008. (Lecture Notes in Physics, Berlin Springer Verlag, v. 740), p. 24.

Carlberg, R. G.; Yee, H. K. C.; Ellingson, E.; Morris, S. L.; Abraham, R.; Gravel, P.; Pritchet, C. J.; Smecker-Hane, T.; Hartwick, F. D. A.; Hesser, J. E.; Hutchings, J. B.; Oke, J. B. The Average Mass Profile of Galaxy Clusters. **apjl**, v. 485, p. L13–L16, ago. 1997.

den Hartog, R.; Katgert, P. On the dynamics of the cores of galaxy clusters. **mnras**, v. 279, p. 349–388, mar. 1996.

Diaferio, A. Mass estimation in the outer regions of galaxy clusters. **mnras**, v. 309, p. 610–622, nov. 1999.

Diaferio, A.; Geller, M. J.; Rines, K. J. Caustic and Weak-Lensing Estimators of Galaxy Cluster Masses. **apjl**, v. 628, p. L97–L100, ago. 2005.

Dodelson, S. **Modern cosmology**. [S.l.: s.n.], 2003.

Dressler, A.; Shectman, S. A. Evidence for substructure in rich clusters of galaxies from radial-velocity measurements. **aj**, v. 95, p. 985–995, abr. 1988.

Duarte, M.; Mamon, G. A. MAGGIE: Models and Algorithms for Galaxy Groups, Interlopers and Environment. **mnras**, v. 453, p. 3848–3874, nov. 2015.

Einasto, M.; Vennik, J.; Nurmi, P.; Tempel, E.; Ahvensalmi, A.; Tago, E.; Liivamägi, L. J.; Saar, E.; Heinämäki, P.; Einasto, J.; Martínez, V. J. Multimodality in galaxy clusters from SDSS DR8: substructure and velocity distribution. **aap**, v. 540, p. A123, abr. 2012.

Fadda, D.; Girardi, M.; Giuricin, G.; Mardirossian, F.; Mezzetti, M. The Observational Distribution of Internal Velocity Dispersions in Nearby Galaxy Clusters. **apj**, v. 473, p. 670, dez. 1996.

Fraley, C.; Raftery, A. E. Bayesian regularization for normal mixture estimation and model-based clustering. **Journal of Classification.**, v. 24, p. 155–181, 2007.

Gifford, D.; Miller, C.; Kern, N. A Systematic Analysis of Caustic Methods for Galaxy Cluster Masses. **apj**, v. 773, p. 116, ago. 2013.

Girardi, M.; Barrena, R.; Boschin, W.; Ellingson, E. Cluster Abell 520: a perspective based on member galaxies. A cluster forming at the crossing of three filaments? **aap**, v. 491, p. 379–395, nov. 2008.

Hansen, S. M.; McKay, T. A.; Wechsler, R. H.; Annis, J.; Sheldon, E. S.; Kimball, A. Measurement of Galaxy Cluster Sizes, Radial Profiles, and Luminosity Functions from SDSS Photometric Data. **apj**, v. 633, p. 122–137, nov. 2005.

Heisler, J.; Tremaine, S.; Bahcall, J. N. Estimating the masses of galaxy groups - Alternatives to the virial theorem. **apj**, v. 298, p. 8–17, nov. 1985.

Hou, A.; Parker, L. C.; Harris, W. E.; Wilman, D. J. Statistical Tools for Classifying Galaxy Group Dynamics. **apj**, v. 702, p. 1199–1210, set. 2009.

Jenkins, A.; Frenk, C. S.; White, S. D. M.; Colberg, J. M.; Cole, S.; Evrard, A. E.; Couchman, H. M. P.; Yoshida, N. The mass function of dark matter haloes. **mnras**, v. 321, p. 372–384, fev. 2001.

Kanji, G. K. **100 Statistical Tests.** London: SAGE Publications, 2006.

Knebe, A.; Müller, V. Formation of groups and clusters of galaxies. **aap**, v. 341, p. 1–7, jan. 1999.

Knebe, A.; Müller, V. Quantifying substructure in galaxy clusters. **aap**, v. 354, p. 761–766, fev. 2000.

Le Cam, L. M. **Asymptotic Methods in Statistical Decision Theory.** New York: Springer-Verlag, 1986.

Lopes, P. A. A.; de Carvalho, R. R.; Kohl-Moreira, J. L.; Jones, C. VizieR Online Data Catalog: Galaxy clusters from SDSS (Lopes+, 2009). **VizieR Online Data Catalog**, v. 739, maio 2009.

Lucambio, F. Diferentes testes para verificar normalidade de uma amostra aleatória. **Universidade Federal do Paraná**, p. 1–5, 2008.

Lynden-Bell, D. Galaxy Formation and the Statistical Mechanics of Violent Relaxation. In: **Liege International Astrophysical Colloquia**. [S.l.: s.n.], 1967. (Liege International Astrophysical Colloquia, v. 14), p. 143.

Mamon, G. A.; Biviano, A.; Murante, G. The universal distribution of halo interlopers in projected phase space. Bias in galaxy cluster concentration and velocity anisotropy? **aap**, v. 520, p. A30, set. 2010.

Merrall, T. E. C.; Henriksen, R. N. Relaxation of a Collisionless System and the Transition to a New Equilibrium Velocity Distribution. **apj**, v. 595, p. 43–58, set. 2003.

MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada: uma abordagem aplicada.** Belo Horizonte: Editora UFMG, 2005.

Ogorodnikov, K. F. Statistical Mechanics of the Simplest Types of Galaxies. **sovast**, v. 1, p. 748, out. 1957.

Padmanabhan, T. Books-Received - Structure Formation in the Universe. **Journal of the British Astronomical Association**, v. 103, p. 193, ago. 1993.

Padmanabhan, T. **Theoretical Astrophysics, Volume III: Galaxies and Cosmology.** Northwestern University: American Astronomical Society, 2002.

Pinkney, J.; Roettiger, K.; Burns, J. O.; Bird, C. M. Evaluation of Statistical Tests for Substructure in Clusters of Galaxies. **apjs**, v. 104, p. 1, maio 1996.

Press, W. H.; Schechter, P. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. **apj**, v. 187, p. 425–438, fev. 1974.

Ribeiro, A. L. B.; de Carvalho, R. R.; Trevisan, M.; Capelato, H. V.; La Barbera, F.; Lopes, P. A. A.; Schilling, A. C. SPIDER - IX. Classifying galaxy groups according to their velocity distribution. **mnras**, v. 434, p. 784–795, set. 2013.

Ribeiro, A. L. B.; Lopes, P. A. A.; Trevisan, M. Non-Gaussian velocity distributions - the effect on virial mass estimates of galaxy groups. **mnras**, v. 413, p. L81–L85, maio 2011.

Ruckdeschel, P.; Kohl, M.; Stable, T.; Camphausen, F. R package distrmod: S4 classes and methods for probability models. **R News**, v. 35, p. 1–27, 2006.

Ryden, B. **Introduction to Cosmology.** Northwestern University: American Astronomical Society, 2016.

Schneider, M.; Knox, L.; Zhan, H.; Connolly, A. Using Galaxy Two-Point Correlation Functions to Determine the Redshift Distributions of Galaxies Binned by Photometric Redshift. **apj**, v. 651, p. 14–23, nov. 2006.

Serra, P.; Oser, L.; Krajnović, D.; Naab, T.; Oosterloo, T.; Morganti, R.; Cappellari, M.; Emsellem, E.; Young, L. M.; Blitz, L.; Davis, T. A.; Duc, P.-A.; Hirschmann, M.; Weijmans, A.-M.; Alatalo, K.; Bayet, E.; Bois, M.; Bournaud, F.; Bureau, M.; Crocker, A. F.; Davies, R. L.; de Zeeuw, P. T.; Khochfar, S.; Kuntschner, H.; Lablanche, P.-Y.; McDermid, R. M.; Sarzi, M.; Scott, N. The ATLAS^{3D} project - XXVI. H I discs in real and simulated fast and slow rotators. **mnras**, v. 444, p. 3388–3407, nov. 2014.

Shakouri, S.; Johnston-Hollitt, M.; Dehghan, S. An optical analysis of the merging cluster Abell 3888. **mnras**, v. 458, p. 3083–3098, maio 2016.

Svensmark, J.; Wojtak, R.; Hansen, S. H. Effect of asphericity in caustic mass estimates of galaxy clusters. **mnras**, v. 448, p. 1644–1659, abr. 2015.

Wing, J. D.; Blanton, E. L. An examination of the optical substructure of galaxy clusters hosting radio sources. **The Astrophysical Journal**, v. 767, n. 2, p. 102, 2013. Disponível em: <http://stacks.iop.org/0004-637X/767/i=2/a=102>.

Wojtak, R.; Łokas, E. L.; Mamon, G. A.; Gottlöber, S.; Prada, F.; Moles, M. Interloper treatment in dynamical modelling of galaxy clusters. **aap**, v. 466, p. 437–449, maio 2007.

Yahil, A.; Vidal, N. V. The Velocity Distribution of Galaxies in Clusters. **apj**, v. 214, p. 347–350, jun. 1977.

York, D. G.; Adelman, J.; Anderson JR., J. E.; Anderson, S. F.; Annis, J.; Bahcall, N. A.; Bakken, J. A.; Barkhouser, R.; Bastian, S.; Berman, E.; Boroski, W. N.; Bracker, S.; Briegel, C.; Briggs, J. W.; Brinkmann, J.; Brunner, R.; Burles, S.; Carey, L.; Carr, M. A.; Castander, F. J.; Chen, B.; Colestock, P. L.; Connolly, A. J.; Crocker, J. H.; Csabai, I.; Czarapata, P. C.; Davis, J. E.; Doi, M.; Dombeck, T.; Eisenstein, D.; Ellman, N.; Elms, B. R.; Evans, M. L.; Fan, X.; Federwitz, G. R.; Fiscelli, L.; Friedman, S.; Frieman, J. A.; Fukugita, M.; Gillespie, B.; Gunn, J. E.; Gurbani, V. K.; de Haas, E.; Haldeman, M.; Harris, F. H.; Hayes, J.; Heckman, T. M.; Hennessy, G. S.; Hindsley, R. B.; Holm, S.; Holmgren, D. J.; Huang, C.-h.; Hull, C.; Husby, D.; Ichikawa, S.-I.; Ichikawa, T.; Ivezić, Ž.; Kent, S.; Kim, R. S. J.; Kinney, E.; Klaene, M.; Kleinman, A. N.; Kleinman, S.; Knapp, G. R.; Korienek, J.; Kron, R. G.; Kunszt, P. Z.; Lamb, D. Q.; Lee, B.; Leger, R. F.; Limmongkol, S.; Lindenmeyer, C.; Long, D. C.; Loomis, C.; Loveday, J.; Lucinio, R.; Lupton, R. H.; MacKinnon, B.; Mannery, E. J.; Mantsch, P. M.; Margon, B.; McGehee, P.; McKay, T. A.; Meiksin, A.; Merelli, A.; Monet, D. G.; Munn, J. A.; Narayanan, V. K.; Nash, T.; Neilsen, E.; Neswold, R.; Newberg, H. J.; Nichol, R. C.; Nicinski, T.; Nonino, M.; Okada, N.; Okamura, S.; Ostriker, J. P.; Owen, R.; Pauls, A. G.; Peoples, J.; Peterson, R. L.; Petravick, D.; Pier, J. R.; Pope, A.; Pordes, R.; Prosapio, A.; Rechenmacher, R.; Quinn, T. R.; Richards, G. T.; Richmond, M. W.; Rivetta, C. H.; Rockosi, C. M.; Ruthmansdorfer, K.; Sandford, D.; Schlegel, D. J.; Schneider, D. P.; Sekiguchi, M.; Sergey, G.; Shimasaku, K.; Siegmund, W. A.; Smee, S.; Smith, J. A.; Snedden, S.; Stone, R.; Stoughton, C.; Strauss, M. A.; Stubbs, C.; SubbaRao, M.; Szalay, A. S.; Szapudi, I.; Szokoly, G. P.; Thakar, A. R.; Tremonti, C.; Tucker, D. L.; Uomoto, A.; Vanden Berk, D.; Vogeley, M. S.; Waddell, P.; Wang, S.-i.; Watanabe, M.; Weinberg, D. H.; Yanny, B.; Yasuda, N.; SDSS Collaboration. The Sloan Digital Sky Survey: Technical Summary. **aj**, v. 120, p. 1579–1587, set. 2000.

Zabludoff, A. I.; Huchra, J. P.; Geller, M. J. The kinematics of Abell clusters. **apjs**, v. 74, p. 1–36, set. 1990.

Apêndices

APÊNDICE A – Teorema do Virial

O chamado Teorema do Virial é um dos teoremas mais importantes e utilizados em astrofísica. Sua origem data da metade do século XIX, do estudo da teoria cinética de gases. Claussius, em 1870, define a grandeza

$$V_c = \frac{1}{2}\sum_i \vec{F_i}.\vec{r_i}, \tag{43}$$

como sendo o virial do sistema de partículas de coordenadas r_i, massa m_i e cada uma sujeita a uma força total F_i, isto é, $F_i = \sum_{j\neq i} f_{ij}$, onde f_{ij} é a força entre as partículas i e j

a derivação clássica do teorema do virial pode ser feita da seguinte maneira. Consideremos um sistema de N-corpos de massa m_i e definimos

$$G_c = \sum_i \vec{p_i}.\vec{r_i}, \tag{44}$$

onde p_i é a quantidade de movimento da partícula i. A equação 44 pode ser reescrita como

$$G_c = \frac{1}{2}\frac{d}{dt}(\sum_i m_i r_i^2) = \frac{1}{2}\frac{dI}{dt}, \tag{45}$$

utilizando a relação $(d\vec{r}/dt).\vec{r} = \frac{1}{2}d(\vec{r}.\vec{r})/dt$, e $I = \sum_i m_i r_i^2$ é o momento de inércia do sistema. Derivando em relação ao tempo a equação 44 e utilizando a definição 43 obtemos

$$\frac{dG_c}{dt} = \sum_i \vec{r_i}.\vec{p_i} + \sum_i \vec{r_i}.\vec{p_i} \tag{46}$$

$$= \sum_i m_i v_i^2 + \sum_i \vec{r_i}.\vec{p_i} \tag{47}$$

$$= 2T + 2V_c \tag{48}$$

lembrando que $2V_c = \sum_i \vec{F_i}.\vec{r_i} = \sum_i \vec{p_i}.\vec{r_i}$.

Retornando ao virial de Claussius, V_c, supomos que as forças que agem sobre as partículas são deriváveis de um potencial (Hipótese I),

$$\vec{f_{ij}} = -\vec{\bigtriangledown}_i \phi(r_{ij}). \tag{49}$$

e que o potencial é em lei de potência, $\phi(r) \propto r^k$ (Hipótese II). Assim obtemos:

$$\vec{f_{ij}} = -\vec{\bigtriangledown} c_{ij} r_{ij}^k = -k c_{ij} r_{ij}^{k-2} (\vec{r_i} - \vec{r_j}), \tag{50}$$

onde c_{ij} é a constante de proporcionalidade do potencial para as partículas i e j. Utilizando a definição do virial de Claussius, equação 43, obtemos:

$$2V_c = -k \sum_i \sum_{j>i} c_{ij} r_{ij}^{k-2} [(\vec{r_i} - \vec{r_j}).\vec{r_i} + (\vec{r_j} - \vec{r_i}).\vec{r_j}] \tag{51}$$

$$2V_c = -k \sum_i \sum_{j>i} c_{ij} r_{ij}^k \tag{52}$$

$$2V_c = -k \sum_i \sum_{j>i} \phi(r_{ij}), \tag{53}$$

onde utilizamos $\vec{F_i} = \sum_j \vec{f_{ij}}$ e $\sum_i \vec{F_i}.\vec{r_i} = \sum_i \sum_{j>i} \vec{f_{ij}}.\vec{r_i} + \vec{f_{ji}}.\vec{r_j}$. finalmente podemos escrever,

$$\frac{dG_c}{dt} = \frac{1}{2}\frac{d^2 I}{dt^2} = 2T - kU, \tag{54}$$

onde T e U são as energias cinética e potencial totais do sistema. Para o caso específico do campo gravitacional, $k = -1$ e obtemos a chamada identidade de Lagrange:

$$\frac{1}{2}\frac{d^2 I}{dt^2} = 2T + U. \tag{55}$$

O teorema do virial é obtido tomando-se a média da equação 55 durante um intervalo de tempo t_o,

$$\frac{1}{t_0}[G_c(t_0) - G_c(0)] = 2\bar{T} + \bar{U}; \bar{x}(t_0) \equiv \frac{1}{t_0}\int_0^{t_0} x(t)dt. \tag{56}$$

Se a média for tomada por um intervalo de tempo suficientemente longo e se o sistema estiver próximo de um estado de equilíbrio quase-estacionário, obtemos o teorema do virial na sua forma usual:

$$2\bar{T} + \bar{U} = 0. \tag{57}$$

A equação acima traduz o fato de que, para um sistema em equilíbrio, o momento de inércia não se altera com o tempo e, portanto, $d^2I/dt^2 = 0$.

Se, além disto o sistema for ergódico, podemos escrever simplesmente

$$2T + U = 0. \tag{58}$$

isto é, uma medida instantânea das energias cinética e potencial bastam para satisfazer o teorema do virial.