

Ministério de Ciência, Tecnologia e Inovação  
Programa de Pós Graduação em Astronomia  
Observatório Nacional

Implantação e testes de algoritmos de *redshift*  
fotométrico no Portal do *Dark Energy Survey* com  
aplicação à Função de Luminosidade de Galáxias

**Julia de Figueiredo Gschwend**

2015



Observatório  
Nacional

IMPLANTAÇÃO E TESTES DE ALGORITMOS DE REDSHIFT FOTOMÉTRICO NO  
PORTAL DO DARK ENERGY SURVEY COM APLICAÇÃO À FUNÇÃO DE  
LUMINOSIDADE DE GALÁXIAS

Julia de Figueiredo Gschwend

Dissertação de Mestrado apresentada à  
Divisão de Programas de Pós-graduação,  
do Observatório Nacional, como parte dos  
requisitos necessários à obtenção do título de  
Mestre em Astronomia.

Orientador: Marcio Antonio Geimba Maia

Rio de Janeiro  
Maio de 2015

IMPLANTAÇÃO E TESTES DE ALGORITMOS DE REDSHIFT FOTOMÉTRICO NO  
PORTAL DO DARK ENERGY SURVEY COM APLICAÇÃO À FUNÇÃO DE  
LUMINOSIDADE DE GALÁXIAS

Julia de Figueiredo Gschwend

DISSERTAÇÃO SUBMETIDA À DIVISÃO DE PROGRAMAS DE PÓS GRADUAÇÃO  
(DPPG) DO OBSERVATÓRIO NACIONAL COMO PARTE DOS REQUISITOS  
NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM ASTRONOMIA.

Examinada por:

---

Prof. Dr. Marcio Antonio Geimba Maia,

---

Prof. Dr. Roderik Overzier,

---

Prof. Dr. Paulo Afrânio Lopes,

RIO DE JANEIRO, RJ – BRASIL  
MAIO DE 2015

de Figueiredo Gschwend, Julia

Implantação e testes de algoritmos de redshift fotométrico no Portal do Dark Energy Survey com aplicação à Função de Luminosidade de Galáxias/Julia de Figueiredo Gschwend. – Rio de Janeiro: DPPG/ON, 2015.

XXII, Ap. - 155 - p.: il.; 29, 7cm.

Orientador: Marcio Antonio Geimba Maia

Dissertação (mestrado) – MCTI/ON/Programa de Astronomia, 2015.

Referências Bibliográficas: p. 91 – 97.

1. *Redshifts* Fotométricos. 2. Função de Luminosidade de Galáxias. 3. Evolução de Galáxias. I. Geimba Maia, Marcio Antonio. II. Observatório Nacional, ON, Programa de Astronomia. III. Mestrado

*À minha mãe.*

# Agradecimentos

Gostaria de agradecer ao meu orientador, Dr. Marcio Maia, por todo o trabalho desenvolvido nos últimos dois anos, pelas muitas horas de dedicação à minha formação e principalmente por descortinar um mundo de possibilidades e descobertas, com a oportunidade de trabalhar na fronteira do conhecimento. Ao Dr. Paulo Pellegrini pela imensa generosidade e empenho em me ajudar a compreender toda a ciência envolvida neste trabalho. Aos Drs. Ricardo Ogando e Aurelio Carnero pela grande ajuda tanto nos temas científicos quanto nas questões computacionais que surgiram ao longo desta jornada. Ao Dr. Luiz da Costa, pela liderança inspiradora do grupo de pesquisa ao qual pertencço e pela muitas vezes que me confiou responsabilidades com as quais eu não acreditava ser capaz de lidar, acelerando meu amadurecimento profissional. Aos rapazes do time de TI do LIneA, por resolver todas as demandas computacionais geradas ao longo deste trabalho, sempre muito prestativos e competentes.

Agradeço à CAPES pelo financiamento desta pós-graduação, apesar da bolsa de estudos não ter sido suficiente para garantir a minha subsistência, mesmo tendo me afastado compulsoriamente de todas as outras atividades profissionais... o que me leva a agradecer ao meu marido Victor, pelo patrocínio da realização deste sonho e pela paciência e tolerância com minhas ausências, por me motivar quando eu desanimava e por me dar a força que eu precisei realizar todo esse trabalho, sem nenhuma expectativa de algo em troca, apenas satisfeito por me ver feliz.

Agradeço aos meus ex-professores, em especial à prof. Maria de Fátima, prof. James Skea, prof. Alfredo Sotto, prof. Pedro Rocha e prof. Marcelo França. E também aos meus ex-alunos, especialmente aos do CAP-UERJ e PENSI Icaraí, que demonstraram ter por mim a mesma admiração que eu tenho pelos meus mestres, o que me enche de alegria e motivação para continuar fazendo parte deste círculo *virtuoso*.

Agradeço à toda a minha família (as duas), especialmente à minha mãe, pela tolerância e incentivo, pela educação, pelos valores, e por todo o afeto que não poderia ser maior, ao meu pai, pelo apoio e por me transmitir a paixão pelas ciências exatas e à minha avó, exemplo máximo de ser humano. Agradeço aos meus amigos (em ordem alfabética para não gerar ciúmes) Codorna, Eduardo, Gabi, Livia (BFF), Luíza, Manteiga, Renata, Taysa, Thaís, Tol, Tumati, Will (que também entrou na categoria família), Yves e Zé, que foram pacientes com o meu abandono e que participaram dos momentos de relaxamento e descontração, também muito importantes no período deste trabalho.

Por fim, agradeço a todos os servidores do ON, professores, pesquisadores, pós-docs, alunos, pessoal do trabalho burocrático e da limpeza, por fazer desta instituição, que hoje é a minha segunda casa, um centro de pesquisa e pós-graduação de excelência, mesmo em um país onde muitas vezes a educação, a ciência e a tecnologia ficam em segundo plano.

*“ Prefiro morrer do que perder a vida.”*

***Roberto Gómez Bolaños***

# Resumo

Apresentamos nesta dissertação resultados preliminares em estudo sobre evolução de galáxias. Este estudo pretende examinar aspectos como o da evolução em densidade e luminosidade como uma função da idade do Universo em *redshifts*  $\lesssim 1$ .

Para esta dissertação o objetivo central foi a implementação de uma metodologia completa que culmina na estimativa da Função de Luminosidade de Galáxias (LF), desde a obtenção de dados fotométricos. Tal metodologia será posteriormente aplicada em dados liberados anualmente pelo levantamento *Dark Energy Survey* (DES).

Em uma primeira fase, foi implantado um pipeline para estimativa da LF no âmbito do Portal Científico do levantamento DES. Dado o grande volume de dados, ferramentas para manter a proveniência dos dados, distribuir catálogos e compartilhar resultados com os demais membros da colaboração se torna um expediente, mais do que útil - necessário.

Como o levantamento DES produz essencialmente dados de natureza fotométrica, temos a necessidade de estimar distâncias de galáxias através do uso de *redshifts* (desvios para o vermelho) fotométricos. Um grande esforço foi feito para a criação de *workflows* científicos que fazem o treinamento e validação de amostras para vários destes algoritmos. Além disso é preciso a aplicação dos algoritmos a amostras fotométricas – para as quais se quer determinar *redshifts* fotométricos (*photo-zs*).

Apresentamos uma análise do desempenho destes códigos para dados simulados e da fase de Verificação Científica do levantamento DES, verificando o impacto destes resultados na determinação de Função de Luminosidade. Examinamos o impacto de distintos critérios de seleção de amostras, tais como os limites na relação sinal-ruído dos objetos, no valor da magnitude limite da amostra, no intervalo de *redshifts* usado para o cálculo de *photo-zs*. Também testamos o desempenho dos algoritmos sem o uso da banda *u*, que não faz parte do levantamento DES.

Todos estes testes servem para avaliar o impacto na determinação da Função de Luminosidade para distintos intervalos de *redshift*, podendo assim examinarmos o comportamento da densidade e magnitude características como uma função da idade do Universo, um dos objetivos de nosso estudo sobre a evolução de galáxias.

Aprendemos com este trabalho, que a qualidade dos dados é tão importante quando a metodologia adotada, e que a limpeza da amostra, baseada na razão-sinal ruído se mostrou um procedimento importante, tendo seu impacto caracterizado pela melhora das métricas de análise da performance do *photo-z*.

A combinação entre a melhor acurácia obtida com códigos empíricos e a possibilidade de obtenção de outros parâmetros, além do *z* em métodos de ajuste a modelos mostrou-se a abordagem mais interessante para o estudo de evolução de galáxias através da Função de Luminosidade.

# Abstract

We present in this work preliminar results on galaxy evolution studies. These studies aim to examine issues such as the evolution in density and luminosity as a function of the age of the Universe in redshifts up to  $\approx 1$ .

The main goal of this work was the implementation of a full methodology, since the treatment of the photometric data, leading to the estimation of the Luminosity Function of Galaxies (LF). In the near future, this methodology will be applied on data from Dark Energy Survey (DES) annual releases.

Initially, a pipeline to estimate the galaxy Luminosity Function under the DES Scientific Portal scope was implemented. Given the large volume of data, tools to keep the source of the data, distribute catalogs and share results with other members of the collaboration becomes an expedient, more than helpful - necessary.

Since DES essentially produces photometric data, we need to estimate distances to galaxies using photometric redshifts. A great effort was made for the creation of scientific workflows that make the training and validation samples for several of these algorithms. Furthermore it is necessary to apply the algorithms to the samples for which one desires to determine the photo-z's - called photometric sample.

We present a performance analysis of these codes for simulated and Science Verification phase of DES by checking the impact of these results in the determination of Luminosity Function. We examined the impact of different criteria for sample selection, such as limits on the signal-to-noise ratio and the magnitude limit of the sample in the range of redshifts used, for the calculation of photo-z's. We also tested the performance of the algorithms without  $u$ -band, which is not present in DES.

All these tests are used to assess the impact in determining the Luminosity Function for different redshift bins, so one can examine the behavior of the density and magnitude characteristics of the Universe as a function of age, one of the goals of our study on the evolution of galaxies.

We learned from this work that the quality of the data is as important as the methodology adopted, and the cleaning of the sample, based on signal to noise ratio has been an important procedure, having its impact characterized by improvement on photo-z metrics.

The combination of the best accuracy obtained with empirical codes and the possibility of obtaining other results, in addition to photo-zs, provided by template-fitting methods seems to be a valuable approach to the study of evolution of galaxies through the Luminosity Function.

# Sumário

Lista de Figuras	xiii
Lista de Tabelas	xviii
<b>1 Introdução</b>	<b>1</b>
<b>2 Metodologia</b>	<b>6</b>
2.1 Magnitudes Aparentes . . . . .	7
2.2 <i>Redshift</i> . . . . .	7
2.2.1 <i>Redshifts</i> Espectroscópicos ( <i>spec-z</i> ) . . . . .	7
2.2.2 <i>Redshifts</i> Fotométricos ( <i>photo-z</i> ) . . . . .	8
2.3 Determinação da Distância das Galáxias . . . . .	11
2.4 Cálculo da Magnitude Absoluta . . . . .	12
2.5 A Função de Luminosidade . . . . .	13
2.5.1 O Método $1/V_{\text{máx}}$ . . . . .	16
2.5.2 O Ajuste à Função de Schechter . . . . .	16
<b>3 Ferramentas Computacionais</b>	<b>18</b>
3.1 Algoritmos para estimativa de <i>photo-z</i> . . . . .	18
3.1.1 Métodos Empíricos . . . . .	18
3.1.1.1 ANNz - <i>Artificial Neural Networks for Redshift</i> . . . . .	18
3.1.1.2 ArborZ - <i>Photometric Redshifts Using Boosted Decision Trees</i> . . . . .	20
3.1.1.3 TPZ - <i>Trees for Photo-Z</i> . . . . .	20
3.1.1.4 SkyNet . . . . .	21
3.1.1.5 <i>Weights</i> . . . . .	22
3.1.1.6 Pofz . . . . .	22
3.1.1.7 DESDM . . . . .	23
3.1.1.8 RVMz . . . . .	23
3.1.1.9 NIP-kNNz . . . . .	24
3.1.1.10 ANNz2 . . . . .	24
3.1.2 Método de Ajuste a Modelos . . . . .	25
3.1.2.1 LePHARE - <i>Photometric Analysis for Redshift Estimate</i> . . . . .	25
3.1.2.2 BPZ - <i>Bayesian Photometric Redshifts</i> . . . . .	26
3.1.2.3 EAZY - <i>Easy and Accurate Redshifts from Yale</i> . . . . .	26

3.1.2.4	ZEBRA - <i>Zurich Extra-galactic Bayesian Redshift Analyser</i> . . . .	27
3.1.2.5	PHOTOZ . . . . .	27
3.2	O Portal Científico do DES (PCDES) . . . . .	28
3.2.1	Visão Geral do PCDES . . . . .	28
3.2.2	O E2E para a criação de VACs . . . . .	30
3.2.2.1	O estágio <i>Data Installation</i> . . . . .	30
3.2.2.2	O estágio <i>Data Preparation</i> . . . . .	31
3.2.2.3	O estágio <i>VAC Creation</i> . . . . .	31
3.2.3	Os <i>Workflows</i> Científicos . . . . .	32
3.3	O <i>Pipeline Luminosity Function</i> . . . . .	33
<b>4</b>	<b>Descrição dos Dados</b>	<b>36</b>
4.1	Dados de Observação - SVA1 . . . . .	38
4.1.1	Conjunto Fotométrico . . . . .	38
4.1.2	Conjuntos de Treinamento e Validação . . . . .	39
4.1.3	Mapas de magnitude limite . . . . .	42
4.1.4	Separação Estrela-Galáxia . . . . .	42
4.1.4.1	MODEST_CLASS . . . . .	42
4.1.4.2	TPZ_SG_CLASS . . . . .	43
4.2	Dados de Simulação - BCC Aardvark v1.0 . . . . .	45
4.2.1	Conjuntos de Treinamento e Validação . . . . .	46
4.2.2	Conjunto Fotométrico . . . . .	46
<b>5</b>	<b>Testes e comparações de algoritmos de <i>Photo-z</i></b>	<b>48</b>
5.1	Comparação entre os códigos de <i>Photo-z</i> . . . . .	48
5.2	A Importância da Qualidade das Amostras . . . . .	54
<b>6</b>	<b>Resultados para a amostra BCC</b>	<b>59</b>
6.1	Definição de critérios de seleção da amostra . . . . .	59
6.2	<i>Photo-z</i> . . . . .	60
6.2.1	Validação do <i>Photo-z</i> . . . . .	60
6.2.2	<i>Photo-z</i> da amostra fotométrica . . . . .	64
6.3	Magnitudes Absolutas . . . . .	65
6.4	Função de Luminosidade . . . . .	66
<b>7</b>	<b>Resultados para a amostra SVA1</b>	<b>70</b>
7.1	Critérios de definição da amostra . . . . .	70
7.2	<i>Photo-z</i> . . . . .	70
7.2.1	Validação do <i>Photo-z</i> . . . . .	70
7.2.2	<i>Photo-z</i> da amostra fotométrica . . . . .	74
7.3	Magnitudes Absolutas . . . . .	76
7.4	Função de Luminosidade . . . . .	77
7.4.1	Tentativa de limpeza da contaminação remanescente . . . . .	79
7.4.1.1	Candidatos a objetos contaminantes no diagrama Cor-Cor . . . .	79

7.4.1.2	Exemplos de objetos problemáticos . . . . .	82
7.4.1.3	Comparação de resultados . . . . .	84
<b>8</b>	<b>Conclusões e Perspectivas</b>	<b>88</b>
	<b>Referências Bibliográficas</b>	<b>91</b>
<b>A</b>	<b>O Portal Científico do DES</b>	<b>Ap. - 98 -</b>
A.1	O <i>Pipeline</i> <code>training_set_maker</code> . . . . .	Ap. - 101 -
A.2	O <i>Pipeline</i> <code>photoz_training</code> . . . . .	Ap. - 102 -
A.2.1	Etapa pré-treinamento: componente <code>subsets_separation</code> . . . . .	Ap. - 103 -
A.2.2	Etapa principal: componente <code>photoz_train</code> . . . . .	Ap. - 104 -
A.2.3	Etapa pós-treinamento: componente <code>photo-z_valid</code> . . . . .	Ap. - 105 -
<b>B</b>	<b>Configurações das Rodadas</b>	<b>Ap. - 107 -</b>
<b>C</b>	<b>Descrição dos Dados: SVA1</b>	<b>Ap. - 108 -</b>
C.1	Magnitudes e erros . . . . .	Ap. - 108 -
C.2	Cores . . . . .	Ap. - 110 -
C.3	<i>Redshift</i> e Distribuição Espacial . . . . .	Ap. - 111 -
C.3.1	Profundidade das observações . . . . .	Ap. - 111 -
<b>D</b>	<b>Separação Estrela/Galáxia</b>	<b>Ap. - 112 -</b>
D.1	Magnitudes e erros . . . . .	Ap. - 115 -
<b>E</b>	<b>Descrição dos Dados: BCC</b>	<b>Ap. - 117 -</b>
E.1	Magnitudes e erros . . . . .	Ap. - 117 -
E.2	Cores . . . . .	Ap. - 119 -
E.3	<i>Redshift</i> e Distribuição Espacial . . . . .	Ap. - 120 -
<b>F</b>	<b>Resultados com a Amostra BCC</b>	<b>Ap. - 121 -</b>
F.1	Validação do <i>photo-z</i> . . . . .	Ap. - 121 -
F.1.1	Conjunto de Treinamento Representativo . . . . .	Ap. - 121 -
F.1.2	Conjunto de Treinamento OzDES-like (não representativo) . . . . .	Ap. - 124 -
F.2	<i>Photo-z</i> do conjunto fotométrico . . . . .	Ap. - 127 -
F.2.1	Subconjunto com 100 mil objetos . . . . .	Ap. - 127 -
F.2.1.1	Conjunto de Treinamento Representativo . . . . .	Ap. - 127 -
F.2.1.2	Conjunto de Treinamento OzDES-like (não representativo) . . . . .	Ap. - 127 -
F.2.2	Amostra fotométrica completa . . . . .	Ap. - 128 -
F.3	Produtos do LePhare . . . . .	Ap. - 128 -
F.3.1	Magnitudes Absolutas . . . . .	Ap. - 128 -
F.3.2	Classificação (preliminar) . . . . .	Ap. - 129 -
F.4	Função de Luminosidade . . . . .	Ap. - 132 -
F.4.1	LF por tipos espectrais (preliminar) . . . . .	Ap. - 135 -

## G Resultados com a Amostra SVA1

Ap. - 138 -

G.1	Validação do <i>photo-z</i> . . . . .	Ap. - 138 -
G.1.1	Conjunto de Treinamento OzDES . . . . .	Ap. - 138 -
G.1.2	Conjunto de Treinamento OzDES ponderado (representativo) . . . . .	Ap. - 141 -
G.2	<i>Photo-z</i> do conjunto fotométrico . . . . .	Ap. - 143 -
G.2.1	Conjunto de Treinamento OzDES . . . . .	Ap. - 143 -
G.2.2	Conjunto de Treinamento OzDES ponderado (representativo) . . . . .	Ap. - 144 -
G.2.3	<i>Photo-z</i> da amostra fotométrica completa . . . . .	Ap. - 144 -
G.3	Produtos do LePhare . . . . .	Ap. - 145 -
G.3.1	Magnitudes Absolutas . . . . .	Ap. - 145 -
G.3.2	Classificação (preliminar) . . . . .	Ap. - 145 -
G.4	Função de Luminosidade . . . . .	Ap. - 147 -
G.4.1	Amostra contaminada com estrelas . . . . .	Ap. - 150 -
G.5	LF por tipos espectrais (preliminar) . . . . .	Ap. - 154 -

# Lista de Figuras

2.1	Etapas para a determinação da Função de Luminosidade . . . . .	6
2.2	Exemplo de espectro obtido pelo projeto OZDES . . . . .	8
2.3	2.3a:Gráfico original do artigo : Schechter 1976 [69], onde o autor propõe a parametrização da LF com a expressão da equação 2.7. 2.3b: LF dependente do tipo morfológico (classificação de Hubble para galáxias do Universo local) e do ambiente. Gráfico retirado do artigo de revisão de Binggeli, Sandage & Tammann, 1988 [13]. . . . .	14
2.4	2.4a: representação da diferença nos formatos da LF de galáxias e da função de massa de halos prevista pelo formalismo de Press-Schechter. 2.4b: exemplo com dados observados (Benson et al. 2003 [8]) . . . . .	15
3.1	Esquema de uma rede neural artificial . . . . .	19
3.2	Esquema de árvore de decisões. . . . .	21
3.3	Exemplos de modelos de SEDs. 3.3a: exemplos representativos dos diferentes tipos morfológicos de galáxias contemplados no conjunto de onde foram selecionadas as SEDs utilizadas neste trabalho. 3.3b: figura original retirada de Coleman, Wu & Weedman (1980) [22] . . . . .	26
3.4	Captura de Tela - <i>Workflows</i> Científicos . . . . .	29
3.5	Área selecionada . . . . .	29
3.6	E2E - Cadeia de produção de VACs . . . . .	30
3.7	Captura de Tela - configuração do <i>pipeline Luminosity Function</i> . . . . .	32
3.8	<i>Pipeline Luminosity Function</i> . . . . .	34
3.9	Fluxograma do módulo <i>Luminosity Function</i> . . . . .	35
4.1	Distribuição Espacial dos dados do DES . . . . .	38
4.2	Distribuição Espacial da amostra SVA1 . . . . .	39
4.3	Fabricação dos conjuntos de treinamento . . . . .	40
4.4	Características fotométricas da amostra SVA1 na banda $i$ . . . . .	41
4.5	Mapas de magnitude limite nas observações . . . . .	42
4.6	Classificadores estrela/galáxia . . . . .	44
4.7	SPREAD_MODEL_I vs magnitude na banda $i$ . . . . .	44
4.8	Amostras de simulação . . . . .	46
4.9	Características fotométricas da amostra BCC na banda $i$ . . . . .	47
5.1	Distribuições de <i>spec-z</i> e magnitudes da amostra SV . . . . .	50

5.2	Métricas . . . . .	52
5.3	$z_{phot}$ versus $z_{spec}$ . . . . .	53
5.4	$\epsilon_{phot}$ versus $\overline{\Delta z}$ para métodos empíricos. . . . .	57
5.5	$\epsilon_{phot}$ versus $\overline{\Delta z}$ para métodos de ajuste a modelos . . . . .	58
5.6	$\epsilon_{phot}$ versus $\overline{\Delta z}$ para métodos que fornecem a $P(z)$ . . . . .	58
6.1	Gráfico de dispersão de $photo-z$ versus $spec-z$ . . . . .	62
6.2	Gráfico de densidades de $photo-z$ versus $spec-z$ . . . . .	63
6.3	$N(z)$ para o conjunto de validação . . . . .	64
6.4	Resultado do TPZ para todas as galáxias que serão utilizadas na determinação da função de luminosidade (seleção de objetos com $S/N > 5$ ). Árvores produzidas com o conjunto representativo. Em azul, o resultado com o método TPZ mode. Em vermelho, resultados para TPZ mean. Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1). . . . .	65
6.5	Distribuição de magnitudes absolutas geradas pelo LePhare para o conjunto fotométrico com seleção em $S/N > 5$ . À esquerda: magnitudes absolutas geradas a partir dos <i>redshifts</i> verdadeiros da simulação. À direita: magnitudes absolutas geradas a partir dos <i>photo-z's</i> calculados com o TPZ. Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1). . . . .	66
6.6	Comparação <i>true - z</i> vs. <i>photo - z</i> . . . . .	67
6.7	Comparação da evolução dos parâmetros de Schechter . . . . .	69
7.1	Gráfico de dispersão de $photo-z$ versus $spec-z$ . . . . .	72
7.2	Gráfico de densidades de $photo-z$ versus $spec-z$ . . . . .	73
7.3	$N(z)$ para o conjunto de validação . . . . .	74
7.4	Distribuições de <i>photo-z's</i> , para todas as galáxias que serão utilizadas na determinação da função de luminosidade. Em azul, o resultado com o método TPZ mode. Em vermelho, resultados para TPZ mean. Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1). . . . .	75
7.5	Comparação entre as distribuições de <i>photo-z's</i> , determinados pelo método TPZ mean (com árvores produzidas com o conjunto OzDES ponderado), para todas as galáxias que serão utilizadas na determinação da função de luminosidade (corte em $S/N > 5$ ). À esquerda, o histograma comum e à direita o histograma normalizado. . . . .	76
7.6	Distribuição de magnitudes absolutas geradas pelo LePhare para o conjunto fotométrico com corte em $S/N > 5$ . À esquerda, amostra após a remoção dos objetos classificados como estrelas. À direita, antes da remoção. Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1). . . . .	77
7.7	LF: Comparação entre amostra limpa e contaminada . . . . .	78
7.8	Diagramas Cor-Cor . . . . .	82
7.9	Exemplo de objetos duplos detectados como um único objeto. . . . .	83
7.10	Exemplo de objetos contaminados pelo fluxo de outros. . . . .	83
7.11	Exemplo de objetos contaminados pelo fluxo gerado pela detecção de satélites nas proximidades. . . . .	83

7.12	Exemplo de <i>photo-z</i> mal estimado. . . . .	84
7.13	Exemplo de <i>photo-z</i> mal estimado. . . . .	84
7.14	Exemplo de <i>photo-z</i> mal estimado. . . . .	84
7.15	Comparação entre LFs após remoção de contaminação . . . . .	85
7.16	Comparação da evolução dos parâmetros de Schechter. . . . .	86
A.1	O Portal Científico - Visão Global . . . . .	Ap. - 99 -
A.2	E2E - Cadeia de produção de VACs . . . . .	Ap. - 100 -
A.3	Esquema de funcionamento do <i>pipeline training_set_maker</i> . . . . .	Ap. - 101 -
A.4	Captura de Tela - <i>Product Log</i> do <i>pipeline training_set_maker</i> . . . . .	Ap. - 102 -
A.5	Gráficos gerados pelo <i>training_set_maker</i> . . . . .	Ap. - 102 -
A.6	<i>Workflow</i> do <i>pipeline photoz_training</i> . . . . .	Ap. - 103 -
A.7	Captura de Tela - <i>Product Log</i> do componente <i>subsets_separation</i> . . . . .	Ap. - 104 -
A.8	Distribuições de Magnitudes geradas no componente <i>subsets_separation</i> . . . . .	Ap. - 104 -
A.9	Captura de Tela - configurações do <i>pipeline photoz_training</i> . . . . .	Ap. - 105 -
A.10	Captura de Tela - <i>Product Log</i> do componente <i>photoz_valid</i> . . . . .	Ap. - 106 -
C.1	Magnitudes Aparentes . . . . .	Ap. - 108 -
C.2	Erros nas Magnitudes Aparentes . . . . .	Ap. - 109 -
C.3	Magnitudes Aparentes vs. Erros . . . . .	Ap. - 109 -
C.4	Diagramas Cor-Magnitude e Cor-Cor . . . . .	Ap. - 110 -
C.5	Cores . . . . .	Ap. - 111 -
C.6	Redshifts e Distribuição Espacial . . . . .	Ap. - 111 -
C.7	Mapas de profundidade . . . . .	Ap. - 111 -
D.1	Classificadores estrela/galáxia . . . . .	Ap. - 112 -
D.2	SPREAD_MODEL_I vs magnitude na banda <i>i</i> . . . . .	Ap. - 113 -
D.3	SPREAD_MODEL_I vs magnitude na banda <i>i</i> . . . . .	Ap. - 114 -
D.4	Banda <i>i</i> após separação estrela/galáxia . . . . .	Ap. - 115 -
D.5	Banda <i>i</i> para estrelas . . . . .	Ap. - 116 -
E.1	Magnitudes Aparentes . . . . .	Ap. - 117 -
E.2	Erros nas Magnitudes Aparentes . . . . .	Ap. - 118 -
E.3	Magnitudes Aparentes vs. Erros . . . . .	Ap. - 118 -
E.4	Diagramas Cor-Magnitude e Cor-Cor . . . . .	Ap. - 119 -
E.5	Cores . . . . .	Ap. - 120 -
E.6	Redshifts e Distribuição Espacial . . . . .	Ap. - 120 -
F.1	Gráfico de dispersão de <i>photo-z</i> versus <i>spec-z</i> - conjunto representativo . . . . .	Ap. - 122 -
F.2	Gráfico de densidades de <i>photo-z</i> versus <i>spec-z</i> - conjunto representativo . . . . .	Ap. - 123 -
F.3	N( <i>z</i> )- conjunto representativo . . . . .	Ap. - 123 -
F.4	Gráfico de dispersão de <i>photo-z</i> versus <i>spec-z</i> - conjunto OzDES-like . . . . .	Ap. - 125 -
F.5	Gráfico de densidades de <i>photo-z</i> versus <i>spec-z</i> - conjunto OzDES-like . . . . .	Ap. - 126 -
F.6	N( <i>z</i> )- conjunto OzDES-like . . . . .	Ap. - 126 -

F.7	Resultado do TPZ em um conjunto de 100 mil galáxias aleatoriamente selecionadas do conjunto fotométrico. Árvores produzidas com o conjunto representativo. . . . .	Ap. - 127 -
F.8	Resultado do TPZ em um conjunto de 100 mil galáxias aleatoriamente selecionadas do conjunto fotométrico. Árvores produzidas com o conjunto OzDES-like (não representativo). . . . .	Ap. - 127 -
F.9	Resultado do TPZ para todas as galáxias que serão utilizadas na determinação da função de luminosidade (remoção de objetos com $S/N < 5$ ). Árvores produzidas com o conjunto representativo. . . . .	Ap. - 128 -
F.10	Distribuição de magnitudes absolutas geradas pelo LePhare para o conjunto fotométrico (com remoção de objetos com $S/N < 5$ ). À esquerda: magnitudes absolutas geradas a partir dos <i>redshifts</i> verdadeiros da simulação. À direita: magnitudes absolutas geradas a partir dos <i>photo-z's</i> calculados com o TPZ. . . . .	Ap. - 128 -
F.11	Distribuição SEDs . . . . .	Ap. - 129 -
F.12	Distribuição SEDs em faixas de <i>redshift</i> . . . . .	Ap. - 130 -
F.13	Evolução da quantidade de modelos ajustados em cada grupo de tipos espectrais. À esquerda: SEDs ajustadas a partir dos <i>redshifts</i> verdadeiros da simulação. À direita: SEDs ajustadas a partir dos <i>photo-z's</i> calculados com o TPZ. . . . .	Ap. - 131 -
F.14	LF: BCC com <i>true - z</i> . . . . .	Ap. - 132 -
F.15	LF: BCC com <i>photo - z</i> . . . . .	Ap. - 133 -
F.16	LF: BCC com <i>true - z</i> com contagens por magnitude absoluta . . . . .	Ap. - 134 -
F.17	LF: BCC com <i>photo - z</i> com contagens por magnitude absoluta . . . . .	Ap. - 134 -
F.18	Evolução dos parâmetros de Schechter . . . . .	Ap. - 135 -
F.19	Evolução do ajuste de Schechter . . . . .	Ap. - 135 -
F.20	LF por tipos com <i>true - z</i> . . . . .	Ap. - 136 -
F.21	Evolução dos parâmetros de Schechter para diferentes tipos espectrais com <i>true - z</i> . . . . .	Ap. - 136 -
F.22	LF por tipos com <i>photo - z</i> . . . . .	Ap. - 137 -
F.23	Evolução dos parâmetros de Schechter para diferentes tipos espectrais com <i>photo - z</i> . . . . .	Ap. - 137 -
G.1	Gráfico de dispersão de <i>photo-z</i> versus <i>spec-z</i> - conjunto OzDES . . . . .	Ap. - 139 -
G.2	Gráfico de densidades de <i>photo-z</i> versus <i>spec-z</i> - conjunto OzDES . . . . .	Ap. - 140 -
G.3	$N(z)$ - conjunto OzDES . . . . .	Ap. - 140 -
G.4	Gráfico de dispersão de <i>photo-z</i> versus <i>spec-z</i> - conjunto OzDES ponderado . . . . .	Ap. - 142 -
G.5	Gráfico de densidades de <i>photo-z</i> versus <i>spec-z</i> - conjunto OzDES-like . . . . .	Ap. - 142 -
G.6	$N(z)$ - conjunto OzDES-like . . . . .	Ap. - 143 -
G.7	Resultado do TPZ em um conjunto de 100 mil galáxias aleatoriamente selecionadas do conjunto fotométrico. Árvores produzidas com o conjunto OzDES. . . . .	Ap. - 143 -
G.8	Resultado do TPZ em um conjunto de 100 mil galáxias aleatoriamente selecionadas do conjunto fotométrico. Árvores produzidas com o conjunto OzDES ponderado (representativo). . . . .	Ap. - 144 -
G.9	Resultado do TPZ para todas as galáxias que serão utilizadas na determinação da função de luminosidade (seleção de objetos com $S/N > 5$ ). Árvores produzidas com o conjunto OzDES. . . . .	Ap. - 144 -
G.10	Distribuição de magnitudes absolutas geradas pelo LePhare para o conjunto fotométrico com seleção de objetos com $S/N > 5$ , a partir dos <i>photo-z's</i> calculados com o TPZ. . . . .	Ap. - 145 -
G.11	Distribuição SEDs . . . . .	Ap. - 145 -

G.12 Distribuição SEDs em faixas de <i>redshift</i> . . . . .	Ap. - 146 -
G.13 Evolução da quantidade de modelos ajustados em cada grupo de tipos espectrais. SEDs ajustadas a partir dos <i>photo-z's</i> calculados com o TPZ. . . . .	Ap. - 146 -
G.14 LF: sva1 com <i>photo - z</i> . . . . .	Ap. - 147 -
G.15 LF: sva1 com <i>photo - z</i> com contagens por magnitude absoluta . . . . .	Ap. - 148 -
G.16 Evolução dos parâmetros de Schechter . . . . .	Ap. - 149 -
G.17 Comparação entre as distribuições de <i>photo-z's</i> e de magnitudes absolutas, para todas as galáxias que serão utilizadas na determinação da função de luminosidade (seleção em $S/N >$ 5). À esquerda, amostra após a remoção dos objetos classificados como estrelas. À direita, antes da remoção. *: a remoção dos objetos não garante 100% de pureza na amostra. . . .	Ap. - 150 -
G.18 Distribuição SEDs . . . . .	Ap. - 150 -
G.19 LF: sva1 com <i>photo - z</i> . . . . .	Ap. - 151 -
G.20 LF: sva1 com <i>photo - z</i> com contagens por magnitude absoluta . . . . .	Ap. - 152 -
G.21 Evolução dos parâmetros de Schechter . . . . .	Ap. - 153 -
G.22 LF por tipos com <i>photo - z</i> . . . . .	Ap. - 154 -
G.23 Evolução dos parâmetros de Schechter para diferentes tipos espectrais com <i>photo - z</i>	Ap. - 155 -

# Lista de Tabelas

2.1	Resumo das Métricas . . . . .	11
4.1	Características dos filtros utilizados na DECam. ( $\lambda$ em nm) . . . . .	36
5.1	Lista de métodos utilizados para estimar os <i>photo-z</i> 's. . . . .	49
5.2	Métricas relacionadas ao viés e ao espalhamento . . . . .	51
5.3	Frações de <i>redshifts</i> discrepantes . . . . .	51
5.4	Reprodutibilidade de $N(z)$ . . . . .	51
5.5	$\overline{\Delta z}$ e $\sigma_{68}$ após diferentes cortes . . . . .	55
5.6	$out_{2\sigma}$ e $out_{3\sigma}$ após diferentes cortes . . . . .	56
6.1	Comparação baseada na métrica $\sigma_{68}$ , entre as performances do TPZ, após diferentes seleções baseadas em razão sinal-ruído. . . . .	59
6.2	Métricas obtidas após cálculo do photo-z no conjunto de validação. . . . .	61
6.3	Limites definidos iterativamente para o ajuste à função de Schechter . . . . .	67
6.4	Parâmetros de Schechter . . . . .	68
7.1	Comparação baseada na métrica $\sigma_{68}$ , entre as performances do TPZ, após diferentes critérios de seleção baseado na razão sinal-ruído. . . . .	70
7.2	Métricas obtidas após cálculo do <i>photo-z</i> no conjunto de validação. . . . .	71
7.3	Limites definidos iterativamente para o ajuste à função de Schechter . . . . .	77
B.1	TPZ . . . . .	Ap. - 107 -
B.2	LePHARE . . . . .	Ap. - 107 -
F.1	Métricas obtidas após o cálculo do photo-z no conjunto de validação. Treinamento realizado com o conjunto <b>representativo</b> . Método: <b>TPZ mode</b> ( <i>photo-z</i> = $z$ correspondente ao pico da $P(z)$ ). . . . .	Ap. - 121 -
F.2	Métricas obtidas após cálculo do photo-z no conjunto de validação. Treinamento realizado com o conjunto <b>representativo</b> . Método: <b>TPZ mean</b> ( <i>photo-z</i> = média da $PDF(z)$ ) . . . . .	Ap. - 122 -
F.3	Métricas obtidas após cálculo do photo-z no conjunto de validação. Treinamento realizado com o conjunto <b>OzDES-like</b> (não representativo). Método: <b>TPZ mode</b> ( <i>photo-z</i> = $z$ correspondente ao pico da $PDF(z)$ ) . . . . .	Ap. - 124 -

F.4	Métricas obtidas após cálculo do photo-z no conjunto de validação. Treinamento realizado com o conjunto <b>OzDES-like</b> (não representativo). Método: <b>TPZ mean</b> ( <i>photo-z</i> = média da PDF( <i>z</i> )) . . . . .	Ap. - 124 -
F.5	Comparação baseada na métrica $\sigma_{68}$ , entre as performances do TPZ, após diferentes seleções em razão sinal-ruído. . . . .	Ap. - 127 -
F.6	Parâmetros de Schechter - BCC com <i>true - z</i> . . . . .	Ap. - 132 -
F.7	Parâmetros de Schechter - BCC com <i>photo - z</i> . . . . .	Ap. - 133 -
G.1	Conjunto de treinamento: OzDES. Método: TPZ mode ( <i>photo-z</i> = pico da PDF( <i>z</i> )) . . .	Ap. - 138 -
G.2	Conjunto de treinamento: OzDES. Método: TPZ mean ( <i>photo-z</i> = média da PDF( <i>z</i> )) . .	Ap. - 139 -
G.3	Conjunto de treinamento: OzDES ponderado. Método: TPZ mode ( <i>photo-z</i> = pico da PDF( <i>z</i> )) . . . . .	Ap. - 141 -
G.4	Conjunto de treinamento: OzDES ponderado. Método: TPZ mean ( <i>photo-z</i> = média da PDF( <i>z</i> )) . . . . .	Ap. - 141 -
G.5	Comparação baseada na métrica $\sigma_{68}$ , entre as performances do TPZ, após diferentes seleções em razão sinal-ruído. . . . .	Ap. - 143 -

# Lista de Acrônimos

- 2dFGRS** - 2dF Galaxy Redshift Survey
- 2MASS** - Two Micron All Sky Survey
- $\Lambda$ CDM** - Lambda-Cold Dark Matter (Lambda-Matéria escura fria)
- BCC** - Blind Cosmology Challenge
- BCG** - Brightest Cluster Galaxy
- CfA** - Center for Astrophysics
- CFHTLS** - Canada-France Hawaii Telescope Legacy Survey
- CFRS** - Canada-France Redshift Survey
- CTIO** - Cerro Tololo Interamerican Observatory
- DECam** - Dark Energy Camera
- DES** - Dark Energy Survey
- DESDM** - Dark Energy Survey Data Managment
- LF** - Luminosity Function (Função de Luminosidade)
- LSST** - Large Synoptic Survey Array
- MF** - Mass Function (Função de Massa)
- NCSA** - National Center for Supercomputing Applications
- OzDES** - Optical Redshifts for Dark Energy Survey
- PDF** - *Probability Density Function* (Função de densidade de probabilidade)
- PCDES** - Portal científico do DES
- photo-z*** - Redshift fotométrico
- SED** - Spectral Energy Distribution (Distribuição Espectral de Energia)
- SDSS** - Sloan Digital Sky Survey
- spec-z*** - Redshift espectroscópico

**SPT** - South Pole Telescope

**SVA1** - Science Verification Annual Release 1

**VHS** - VISTA Hemisphere Survey

**VVDS** - Vimos-VLT Deep Survey

# Capítulo 1

## Introdução

O estudo da evolução de galáxias é um campo altamente empírico da astrofísica extragaláctica, alimentado tanto por estudos de caracterização do Universo local, quanto por estudos comparativos da evolução das mais variadas propriedades em diferentes épocas do Universo. Tais estudos visam chegar a um desenho de uma história evolutiva consistente com a evolução observada, através da utilização de modelos e simulações em confronto com dados observacionais.

Embora não seja possível acompanhar a evolução de uma única galáxia devido à grande disparidade entre as escalas de tempo da vida humana e da “vida” das galáxias, podemos formular hipóteses sobre a sua evolução, a partir da observação de diferentes populações de galáxias em diferentes épocas.

Para termos um melhor entendimento da evolução de galáxias, precisamos lançar mão de duas abordagens complementares: (i) Estudar grandes amostras estatisticamente representativas, medindo a evolução com o tempo de várias quantidades observacionais, hoje possível graças aos grandes levantamentos extragalácticos. (ii) Estudar detalhadamente pequenas amostras, ou galáxias individuais, para compreender os processos físicos que levam à evolução, tais como a formação estelar, fluxos de matéria e energia, captura de gás, jatos de núcleos ativos e fusão de galáxias.

As diversas propriedades das galáxias, como idade da população estelar, histórico e taxa de formação estelar, metalicidade, extinção por poeira, luminosidade, cores, tamanhos e massa estelar, são funções do desvio para o vermelho (*redshift* -  $z$ ) e do ambiente. Apesar de muitos avanços terem sido feitos na compreensão do Universo local ( $z \approx 0.2$ ) através de levantamentos como o *Sloan Digital Sky Survey* (SDSS – York et al. 2000 [78]), são os grandes levantamentos fotométricos que se iniciam agora que irão apresentar resultados para a fronteira de  $z \approx 1.0$ , época em que o Universo tinha metade de sua idade atual. Estas observações são cruciais para criar uma ligação entre as fases de formação das galáxias em *redshifts* maiores, obtidas através de registros fósseis das suas populações estelares, e as das galáxias no Universo local.

O cenário mais aceito, atualmente, para explicar a formação das galáxias é o modelo de aglomeração hierárquica, no qual objetos maiores se formaram através da fusão de objetos menores. Trabalhos publicados na década de 70, como os de Press e Schechter (1974) [64] e White e Rees (1978) [74], foram os primeiros a propor um modelo de formação de galáxias baseado na formação hierárquica de halos de matéria escura. O subsequente esfriamento e condensação de bárions nos centros desses halos daria origem às protogaláxias que sofreriam fusões para originar as galáxias

observadas atualmente.

Neste cenário, esses objetos deveriam ser raros, ou sequer existir, em altos *redshifts*. Entretanto, estudos das populações estelares no Universo local mostram que as galáxias massivas de hoje apresentam populações estelares velhas (e.g., Papovich et al 2006 [60]), o que sugere a formação desses sistemas numa fase primordial do Universo ( $z \approx 5$ ) e uma evolução passiva desde então, como um dos possíveis caminhos evolutivos (Eggen et al. 1962 [33]).

Outra característica que vincula os modelos de evolução é a contagem de objetos. Na lógica do modelo hierárquico, se os objetos menores vão sofrendo fusões ao longo da sua evolução, espera-se que o número de objetos diminua com o passar do tempo. Para estudar esta característica, as ferramentas utilizadas neste estudo são a função de luminosidade (LF) e de massa (MF) em função do *redshift* e do ambiente. Outra forma é a busca por galáxias de grande massa a grandes *redshifts* através de critérios de seleção por cores.

Os novos levantamentos, que pretendem ir mais fundo e cobrindo grandes áreas, ajudarão a superar algumas das deficiências dos levantamentos antecessores: a falta de volume ou profundidades, suficientes para não serem afetados pela variância cósmica e a baixa densidade numérica no caso das galáxias de grande massa.

Apesar de sermos capazes de caracterizar teoricamente a formação de galáxias no contexto do modelo  $\Lambda$ CDM ( $\Lambda$  matéria escura fria), os processos que levam às propriedades físicas observadas hoje (como mecanismos de desencadeamento e extinção de formação estelar) ainda não estão claros. Para determinar como se dá a evolução das galáxias a partir de sua formação, dividem-se os objetos em subgrupos de acordo com densidade ambiental, tipo morfológico, cor, etc, para os quais se determinam LFs independentes. A LF, que denota a densidade numérica comóvel de galáxias em função de suas luminosidades, é uma poderosa ferramenta no estudo da distribuição desses objetos e da matéria visível presente nesses corpos ao longo da história do Universo.

Do levantamento *Center for Astrophysics* (CfA - Huchra et al. 1983 [42]), que confirmou a aglomeração de galáxias não só em projeção, mas também em 3D, passando pelo levantamento de *Las Campanas* (Lin et al. 1996 [50]) entre outros, até chegar aos levantamentos mais atuais – *2 degree Field Galaxy Redshift Survey* (2dFGRS - Colless 1998 [23]), *Two Micron All Sky Survey* (2MASS - Skrutskie et al. 2006 [71]) e SDSS – houve um grande avanço na determinação da LF local (até  $z \approx 0.2$ ). Já a faixa que se estende até  $z \approx 1.0$  foi explorada por levantamentos como o *Canada-France Redshift Survey* (CFRS - Lilly et al. 1995 [47]), o COMBO-17 (Wolf et al. 2003 [76]), *Vimos VLT Deep Survey* (VVDS - e.g. Ilbert et al. 2005 [43]) e o *zCOSMOS* (Lilly et al. 2007 [48]; Zucca et al. 2009 [81]), que estudaram a LF em função da cor, analisando os processos evolutivos tanto de galáxias *early-type* quanto *late-type* em *redshifts* intermediários. A evolução dos estudos nos próximos anos ocorrerá com levantamentos que irão abranger áreas maiores, faixas de comprimentos de onda mais extensas, e irão explorar galáxias mais fracas e mais distantes, como o *Dark Energy Survey* (DES - Abbott et al. 2005 [1]; Flaugher et al. 2005 [36]), o *Large Synoptic Survey Telescope* (LSST, Ivezić et al. 2006 [45]).

A MF é definida como a densidade numérica de galáxias por intervalo logarítmico de massa. Uma determinação empírica de como a massa estelar evoluiu desde a formação das galáxias é um importante vínculo a ser imposto sobre o modelo  $\Lambda$ CDM. Determinando a densidade de massa estelar e a evolução da MF das galáxias, é possível explorar a eficiência da formação estelar e o

*feedback* em diferentes escalas de massa. Além disso, a massa das galáxias está correlacionada a outras propriedades desses objetos e o estudo das relações de escala em função do *redshift* ajuda a compreender a formação estelar e a formação de galáxias por fusões.

Diversos levantamentos mediram a MF tanto para o Universo local quanto em *redshifts* maiores. No primeiro caso, podemos citar os levantamentos 2dF (Cole et al. 2001, [21]), dados conjuntos obtidos pelo 2MASS e o SDSS (Bell et al. 2003 [5]), COSMOS (Drory et al. 2009 [32]), entre outros. Já no regime de  $z > 1$ , há MFs calculadas utilizando dados do levantamento GOODS (Mortlock et al. 2011 [57]), do UKIDSS (Caputi et al. 2011 [18]), entre outros.

É observado um crescimento de um fator 10 na massa estelar entre  $z \approx 2 - 3$  e o Universo local (Dickinson et al. 2003 [31]; Elsner et al. 2008 [34]) com crescimento relativo menor para galáxias de grande massa ( $M > 10^{11} M_{\odot}$ ) a partir de  $z \approx 1 - 2$  (Wake et al. 2006 [73]; Pozzetti et al. 2010 [63]; Caputi et al. 2011 [18]).

Apesar dos esforços, a evolução da MF ainda é um campo repleto de incertezas, especialmente quando se leva em conta as dificuldades em se fazer uma estimativa acurada das massas estelares (Maraston et al. 2006 [53]; Conroy, Gunn & White 2009 [25]).

As características dos levantamentos de grande área, porém rasos, ou de campos pequenos e muito profundos realizados até hoje, levaram a um viés na caracterização de objetos com grandes *redshifts* que acabam sendo limitados, respectivamente, ou por fontes muito brilhantes (geralmente hospedeiras de AGNs ou surtos de raios gama) ou por objetos fracos mas de grande densidade numérica. Sendo assim, o estudo de galáxias de grande massa em grandes *redshifts*, que se acredita serem mais raras, pode ter sido prejudicado. A próxima geração de grandes levantamentos que aumentam o limite de magnitude observada (como o DES) em conjunção com a realização de missões de *follow-up* espectroscópico dos candidatos selecionados através de critérios de cor irá confirmar ou não a existência dessas galáxias de grande massa a grande  $z$ .

O DES é um levantamento fotométrico que se estenderá do óptico ao infravermelho próximo utilizando as bandas  $g, r, i, z$  e  $Y$ . A proposta do levantamento é de, através do uso de quatro experimentos distintos e independentes (abundância e distribuição espacial de aglomerados de galáxias, lenteamento fraco, oscilações acústicas de bárions e supernovas tipo Ia) (Annis et al. 2005 [2]), impor vínculos à equação de estado da energia escura. O DES cobrirá 5000 graus quadrados da calota galáctica Sul (Abbott et al. 2005 [1]) e em adição aos filtros mencionados o levantamento *VISTA Hemisphere Survey* (VHS, McMahon et al. 2013 [55]) irá prover dados nas bandas  $J, H$  e  $K$ . Além disso, aproximadamente 4000 graus quadrados da região observada irão se sobrepor ao levantamento *South Pole Telescope* (SPT, Ruhl et al. 2004 [67]).

O equipamento usado pelo DES é uma câmera CCD, a DECam ([37]), especialmente desenhada para o levantamento que cobre 3 graus quadrados de céu. Ela está montada no telescópio *Blanco* de 4 metros localizado no *Cerro Tololo Inter-American Observatory* (CTIO). De acordo com o planejamento, o levantamento terá 525 noites de observações distribuídas por cinco anos nos períodos de setembro a fevereiro (Abbott et al. 2005 [1]). Os limites de fotometria esperados são:  $g = 25.4$ ,  $r = 24.9$ ,  $i = 24.6$ ,  $z = 23.8$  e  $Y = 21.7$ , no sistema AB de magnitudes, uma vez finalizado o levantamento.

A primeira luz ocorreu no início de setembro de 2012 e a fase de comissionamento no início de 2013. Em fevereiro de 2015 encerrou-se o segundo ano de observações. Apesar de ser voltado para

o estudo da energia escura, a grande profundidade e área cobertas pelo DES irão proporcionar uma imensa amostra estatística que será de grande utilidade no campo de evolução de galáxias. O levantamento tornará possível a determinação de vínculos sobre a parte fraca ( $M \gtrsim M^*$ ) da LF em pequenos *redshifts*, além de estudar o comportamento da parte brilhante da LF em  $z \approx 4 - 6$ , com a seleção de *Lyman Break Galaxies* nesses regimes de  $z$ .

O avanço na determinação da LF em *redshifts* mais distantes é desafiado pela necessidade de amostras homogêneas e extensas, que não são possíveis de se obter utilizando *redshifts* espectroscópicos (*spec-zs*). Essa técnica de determinação de *redshifts*, apesar de mais precisa, requer integrações longas para atingir razões sinal/ruído suficientes. Além disso, grandes espectrógrafos, como o do SDSS só conseguem obter espectros para  $\sim 600$  fontes por vez. Já o DES irá apresentar espectroscopia apenas para uma fração pequena de seus objetos, frutos de levantamentos de *follow-up* e da compilação de dados publicados.

A solução está na utilização de *redshifts* fotométricos (*photo-zs*). Apesar de menos precisos, os métodos de cálculo de *photo-zs* se baseiam nas cores de cada objeto para determinar de forma mais rápida e incluindo objetos mais fracos, os *redshifts* de uma amostra. Os métodos se dividem em ajustes a modelos e ajustes empíricos, cada um com vantagens e desvantagens específicas. Enquanto os métodos de ajuste de modelos tendem a apresentar maior dispersão em seus resultados quando se compara *photo-zs* a *spec-zs*, métodos de ajuste empírico, como redes neurais, que dependem de uma amostra de treinamento que contenha *spec-zs*, tendem a apresentar maiores vieses (*bias*), principalmente nas regiões mais próximas aos limites inferior e superior de *redshift* da amostra de treinamento, apresentando, porém menor dispersão.

Apesar de haver dificuldade em se caracterizar o ambiente de galáxias em levantamentos fotométricos (e.g., Cooper et al. 2005 [26]), é possível separar a amostra fotométrica em populações de aglomerados (alta densidade) e de campo (baixa densidade) o que, unido ao poder estatístico da enorme amostra a ser produzida pelo DES, irá permitir avanços no estudo da evolução galáctica em diferentes ambientes. O trabalho de Ramos et al. (2011) [66] mostra que é possível determinar a LF utilizando amostras fotométricas. Quando comparado aos resultados obtidos para dados do VVDS com *spec-zs*, a utilização de *photo-zs* reproduziu a evolução dos parâmetros característicos da LF obtidos no primeiro caso. Ainda, testes com a amostra do *Canada-France-Hawaii Telescope Legacy Survey* (CFHTLS, Ilbert et al. 2006 [44] ; Coupon et al. 2009 [27] ) reproduziram a evolução dos parâmetros obtidas por Ilbert et al. (2005) [43] para a amostra espectroscópica do VVDS. A falta da banda *u* nas observações do DES terá impacto na determinação dos *photo-zs* de objetos com  $z < 0.4$  e  $z > 3$  (Ilbert et al. 2006 [44]).

Estudos recentes mostram um pequeno número de galáxias de massa superior a  $2.5 \times 10^{11} M_{\odot}$  na faixa de  $z \approx 3 - 4$  (e.g., Papovich et al. 2006 [60]; Marchesini et al. 2010 [54]) e algumas evidências, ainda não conclusivas, da existência desses objetos em  $z > 5$  (e.g., Yan et al. 2006 [77]; Mancini et al. 2009 [52]). Acredita-se que o DES tenha a combinação perfeita de área e profundidade para realizar a busca por essas galáxias (Davies et al. 2013 [29]). A ideia é utilizar uma técnica de seleção por cor similar àquelas aplicadas na seleção de galáxias *Lyman-break*.

A identificação de galáxias de grande massa em altos *redshifts* sugere que os objetos de grande massa do Universo local possam ter sido formados no Universo jovem, corroborando com a hipótese de uma possível evolução passiva.

Testes com galáxias simuladas de grande massa mostram que os principais contaminantes no espaço de cor desses objetos são (Davies et al. 2014 [29]):

- para  $z \approx 3$ , galáxias em baixos *redshifts* atenuadas por extinção;
- para  $z \approx 5$  e  $z \approx 6$ , estrelas anãs L e T.
- QSOs em todas as faixas de  $z$ .

Para resolver este problema, será necessário um *follow-up* espectroscópico nos candidatos que ocupam as áreas de menor contaminação no espaço de cores.

Neste trabalho apresentamos a definição de uma metodologia e a preparação de infraestrutura para uso dos dados do levantamento DES iniciando com a implantação e testes de ferramental para estimativas de *photo-z*, com aplicação a determinação de LF em simulações e dados da fase de verificação científica do levantamento.

Nos capítulos 2 e 3 e no apêndice A são apresentados os detalhes da metodologia utilizada em todas as etapas da determinação da LF, bem como as ferramentas computacionais necessárias. A descrição das amostras utilizadas é feita no capítulo 4, complementada pelos apêndices C, D e E. Os resultados e discussões são apresentados nos capítulos 5, 6 e 7, complementados pelos apêndices F e G. Por fim, as conclusões e perspectivas compõem o capítulo 8.

Parte dos resultados apresentados ainda não foram validados inteiramente pela colaboração DES, e portanto, devem ser considerados como preliminares.

# Capítulo 2

## Metodologia

Este trabalho utiliza a função de luminosidade (LF) como um exemplo de ferramenta para o estudo da evolução de galáxias. Os ingredientes necessários para a determinação da LF são: luminosidade (ou equivalentemente magnitude absoluta) e *redshift*. As magnitudes absolutas são derivadas das magnitudes aparentes e do *redshift*, logo, a matéria prima para a determinação da LF se resume à: dados fotométricos e *redshift*. A figura 2.1 mostra as etapas serem cumpridas para chegar à função.

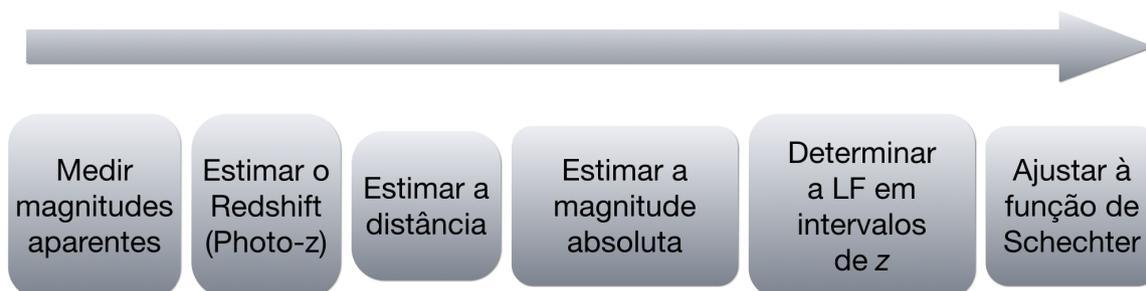


Figura 2.1: Esquema resumindo o passo-a-passo necessário para a obtenção de resultados de Função de Luminosidade a partir de dados fotométricos.

Na era dos grandes levantamentos, não apenas a LF, mas a maioria das análises são baseadas em dados exclusivamente fotométricos. É inviável do ponto de vista técnico-operacional obter o *redshift* através da técnica de espectroscopia. Desta demanda surge a importância do cálculo do *redshift* fotométrico (*photo-z*). Embora o *redshift* espectroscópico (*spec-z*) seja mais acurado e mais preciso, o poder estatístico das grandes amostras possibilitam o uso do *photo-z* em análises globais (Hildebrandt et al. 2010 [41]; Carrasco Kind & Brunner 2013 [19]).

Além da luminosidade (ou magnitude absoluta), a LF depende do volume que contém a amostra observada, que por sua vez depende da distância a que estão os objetos. Esta distância é estimada a partir do *redshift* fotométrico, utilizando as magnitudes aparentes e cores obtidas com as observações. A distância dos objetos também depende do modelo cosmológico adotado. Neste cálculo, nos baseamos em um modelo de Universo plano, preenchido por matéria escura e com constante cosmológica, definido pelos parâmetros:  $\Omega_k = 0$ ,  $\Omega_m = 0.3$ ,  $\Omega_\Lambda = 0.7$ , e  $H_0 = 70 \text{ kms}^{-1}\text{Mpc}^{-1}$ .

## 2.1 Magnitudes Aparentes

O principal observável em um levantamento fotométrico é o fluxo da luz proveniente das galáxias observadas. O DES possui um sistema integrado de calibração e redução das imagens, que converte as medidas de fluxo em magnitudes aparentes. Sendo assim, a matéria prima deste estudo é um catálogo de magnitudes aparentes, a partir do qual todas as outras etapas se desenvolverão. Em paralelo, fizemos um estudo semelhante com um catálogo simulado, onde as magnitudes aparentes foram artificialmente criadas para a realização de testes em condições controladas. No capítulo 4, abordamos com detalhes as características dos dados, observados e simulados, utilizados neste trabalho.

## 2.2 *Redshift*

A luz que recebemos das galáxias sofre um desvio  $\Delta\lambda$  em seus comprimentos de onda  $\lambda$ , como consequência do movimento relativo na componente da linha de visada. Tal desvio é quantificado através da grandeza *redshift* ( $z$ ), definida como:

$$z \equiv \frac{\Delta\lambda}{\lambda} = \frac{\lambda_{observado} - \lambda_{fonte}}{\lambda_{fonte}}. \quad (2.1)$$

Uma parte deste desvio é devido ao efeito Doppler, consequência do movimento peculiar relativo na componente radial, em relação a um mesmo referencial inercial ( $z_{pec}$ ). Este movimento pode ser de afastamento ou aproximação, podendo causar, respectivamente, um desvio para o vermelho ( $\Delta\lambda > 0$ ) ou para o azul ( $\Delta\lambda < 0$ ). O restante é devido ao *redshift* cosmológico ( $z_{cosmo}$ ), provocado pela expansão do Universo. Neste caso não importa o movimento peculiar de um corpo através de um sistema de coordenadas, e sim o afastamento das próprias coordenadas, podendo ser, inclusive, com velocidades superiores à velocidade da luz no vácuo. O *redshift* observado é então a composição dos desvios causados pelos dois fenômenos (Peacock, 1999 [61]):

$$1 + z_{observ} = (1 + z_{cosmo})(1 + z_{pec}).$$

Para a maioria dos objetos extragalácticos, o desvio é dominado pelo efeito cosmológico – fazendo com que o resultado seja essencialmente um desvio para o vermelho – e o efeito da velocidade peculiar pode ser desprezado.

### 2.2.1 *Redshifts* Espectroscópicos (*spec-z*)

O *redshift* de uma galáxia pode ser obtido diretamente da medida do seu espectro. No projeto OzDES (que será apresentado no capítulo 4, a medida do *spec-z* é feita de forma semi-automática pelo *software runz*, desenvolvido dentro do projeto, com a supervisão de uma equipe de astrônomos. O método utilizado pelo *runz* consiste em identificar linhas de emissão e/ou de absorção no espectro. O software também possui um conjunto de modelos de espectros de estrelas, galáxias e quasares, fornecendo como resultado adicional uma classificação quanto à natureza do objeto. Na figura 2.2 temos um exemplo de espectro analisado pelo programa *runz* do projeto OzDES.

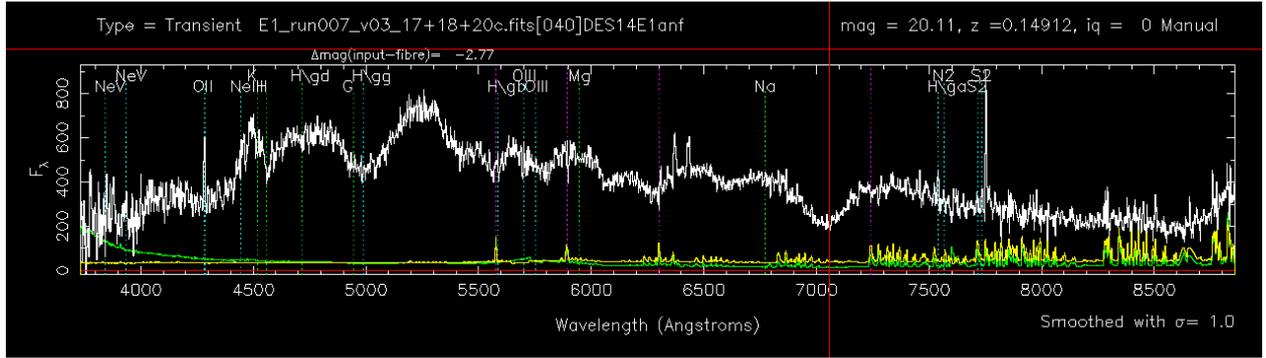


Figura 2.2: Exemplo de um espectro obtido no telescópio Anglo-Australiano durante as observações do projeto OzDES. Este objeto foi identificado como uma supernova do tipo IA, em uma galáxia a um *redshift*  $z \sim 0.15$  (Yuan et al. 2014 [79]).

No contexto dos grandes levantamentos, não é viável medir os *redshifts* de todos os objetos observados usando a técnica da espectroscopia. Entretanto, devido a sua maior acurácia, os *spec-zs* são utilizados para treinamento ou calibração de algoritmos de *photo-z*.

### 2.2.2 *Redshifts* Fotométricos (*photo-z*)

O *photo-z* consiste em uma medida de  $z$  onde os dados iniciais são apenas de natureza fotométrica – as magnitudes. Com esta técnica, é possível obter o  $z$  para um número muito maior de galáxias, utilizando menos tempo de telescópio (Hildebrandt et al. 2010 [41]), além de incluir objetos menos brilhantes em comparação com o *spec-z*. O grande volume de dados tratado estatisticamente compensa a menor precisão no *redshift*, tornando esta técnica bastante competitiva nos grandes levantamentos que visam fazer medidas na escala cosmológica (Carrasco Kind & Brunner 2013 [19]).

Há uma grande diversidade de métodos computacionais para realizar a estimativa do *photo-z* disponíveis na literatura. Eles podem ser divididos em dois grupos: os métodos empíricos e os métodos de ajuste a modelos. Detalhes dos dois tipos de métodos, bem como exemplos de algoritmos e seu funcionamento, serão mostrados no próximo capítulo.

Para quantificar a performance do cálculo do *photo-z*, um conjunto de métricas é calculado a partir da comparação entre os valores de *photo-z* e os de *spec-z* conhecido, tomado como verdadeiro (*true-z*). Com o objetivo de garantir a representatividade dos *redshifts* verdadeiros pelas medidas espectroscópicas, utilizamos apenas as medidas com bons parâmetros de qualidade (*flags*). Por exemplo, para o levantamento VVDS, uma das fontes de *redshifts* utilizadas neste trabalho, as *flags* 3 e 4 representam, respectivamente, 95 e 100% de certeza na identificação correta das linhas nos espectros.

O DES estabelece requisitos científicos voltados para as necessidades dos estudos centrais do levantamento visando a caracterização da energia escura. Alguns desses requisitos se referem à performance do *photo-z*, estabelecendo valores limites para algumas métricas, descritas a seguir.

Como as métricas são obtidas a partir de um conjunto de validação, elas podem não refletir exatamente a performance que o método teria na amostra fotométrica. Para contornar esta limitação, as métricas podem ser ponderadas por um conjunto de pesos. Para cada amostra de  $N$

galáxias é definido um vetor de pesos, com um peso  $w$  para objeto  $i$ , sendo usado na ponderação das métricas globais, onde:

$$\sum_{i=1}^N w_i = 1.$$

No caso deste trabalho, os pesos são calculados pelo algoritmo *weights*, que será descrito no próximo capítulo (mas o método poderia ser aplicado a qualquer distribuição de pesos). Eles refletem a importância de cada elemento do conjunto de validação (conjunto para o qual as métricas serão calculadas), na comparação com o conjunto fotométrico. Caso nenhum peso seja utilizado, é atribuído  $w_i = 1/N$  para todas as galáxias.

Para quantificar a acurácia do *photo-z*, o "viés"  $\Delta z_i$  é definido como:

$$\Delta z_i = z_{phot,i} - z_{spec,i}.$$

e o viés normalizado como:

$$\Delta z_i = \frac{(z_{phot,i} - z_{spec,i})}{\epsilon_{phot,i}},$$

onde  $\epsilon_{z_{phot}}$  é o erro do *photo-z*. Para quantificar o viés global, são utilizadas a média ( $\overline{\Delta z}$ ):

$$\overline{\Delta z} = \frac{\sum w_i \Delta z_i}{\sum w_i}$$

e a mediana ( $\Delta z_{50}$ ) da distribuição de  $\Delta z$ , que satisfaz a probabilidade:

$$P_{50} = P(\Delta z \leq \Delta z_{50}) = \int_0^{\Delta z_{50}} w(\Delta z) d(\Delta z) = \frac{1}{2}.$$

Ambas calculadas para a amostra de validação (ou o equivalente conjunto de teste). Este parâmetro reflete as diferenças sistemáticas entre o *photo-z* e o valor esperado (*spec-z*).

O espalhamento dos resultados é estimado a partir do desvio padrão de  $\Delta z$ :

$$\sigma_{\Delta z} = \left( \frac{\sum w_i (\Delta z_i - \overline{\Delta z})^2}{\sum w_i} \right)^{1/2}$$

e o percentil 68 ( $\sigma_{68}$ ), que corresponde a metade do intervalo em torno de  $\Delta z_{50}$  que contém 68% dos dados:

$$\sigma_{68} = \frac{1}{2} (P_{84} - P_{16}).$$

O  $\sigma_{68}$  é uma medida mais robusta da dispersão do que o  $\sigma_{\Delta z}$  que é influenciado por medidas com erros catastróficos. Os requisitos científicos DES definem que  $\sigma_{68}$  deve ser  $< 0.12$  para, pelo menos, 90% da amostra.

A fração de objetos com *redshifts* discrepantes (com  $|\Delta z|$  grande) é determinada pelas métricas  $out_{2\sigma}$  e  $out_{3\sigma}$ , que medem, respectivamente a fração de objetos que estão fora do intervalo  $\overline{\Delta z} \pm 2\sigma_{\Delta z}$  e  $\overline{\Delta z} \pm 3\sigma_{\Delta z}$ . Os requisitos científicos do DES estabelecem  $out_{2\sigma} < 0.1$  e  $out_{3\sigma} < 0.015$ .

Para avaliar os erros reportados pelos códigos de *photo-z* ( $\epsilon_{z_{phot}}$ ), é calculado o viés global normalizado:

$$\overline{\Delta z'} = \frac{\sum w_i \Delta z'_i}{\sum w_i},$$

onde a dispersão  $\sigma_{\Delta z'}$ , é o desvio padrão de  $\Delta z'$ . Idealmente obter-se-ia  $\overline{\Delta z'} = 0$  e  $\sigma_{\Delta z'} = 1$  para uma distribuição gaussiana de média zero e variância igual a 1.

A capacidade de reproduzir a distribuição verdadeira de *redshifts*  $N(z_{spec})$  é analisada através de dois parâmetros estatísticos. O primeiro, o  $N_{Poisson}$ , é uma estimativa do quão próximas são as duas distribuições  $N(z_{phot})$  e  $N(z_{spec})$ . Para cada intervalo  $j$  de *photo-z*, de largura 0.1 é calculada a diferença entre  $N(z_{phot})$  e  $N(z_{spec})$ , normalizada por flutuações poissônicas:

$$n_{poisson,j} = \frac{\left( \sum_{z_{phot,i} \in j} w_i N - \sum_{z_{spec,i} \in j} w_i N \right)}{\sqrt{\sum_{z_{spec,i} \in j} w_i N}}.$$

A métrica  $N_{poisson}$  é então calculada como a média quadrática da quantidade anterior:

$$N_{poisson} = \left( \frac{1}{n_{bins}} \sum_{j=1}^{n_{bins}} n_{poisson,j}^2 \right)^{1/2}.$$

O segundo parâmetro estatístico é o resultado do teste estatístico *Kolmogorov-Smirnov* (KS), que infere se  $N(z_{phot})$  e  $N(z_{spec})$  são consistentes entre si, como se fossem provenientes de uma mesma população parente, independente da divisão dos intervalos. KS é definido como a distância máxima entre as distribuições empíricas acumulativas, calculadas como:

$$F_{spec}(z) = \frac{\sum_{i=1}^N \Omega_{z_{spec,i} < z}}{\sum w_i} \quad e \quad F_{phot}(z) = \frac{\sum_{i=1}^N \Omega_{z_{phot,i} < z}}{\sum w_i},$$

onde,

$$\Omega_{z_{spec,i} < z} = \begin{cases} w_i, & \text{se } z_{spec,i} < z \\ 0, & \text{se } z_{spec,i} \geq z \end{cases} \quad e \quad \Omega_{z_{phot,i} < z} = \begin{cases} w_i, & \text{se } z_{phot,i} < z \\ 0, & \text{se } z_{phot,i} \geq z \end{cases}.$$

A estatística KS é então calculada como:

$$KS = \max_z (|F_{phot}(z) - F_{spec}(z)|).$$

Quanto menor o valor de KS, mais similares são as duas distribuições, indicando uma melhor performance do código em reproduzir a distribuição  $N(z)$ . A tabela 2.1 traz o resumo das definições das métricas, feitas nos parágrafos anteriores.

Tabela 2.1: Resumo das métricas usadas para a avaliação dos desempenhos dos códigos de *photo-z*. (Sánchez et al. 2014 [68])

Métrica	Descrição	Requisito DES
$\overline{\Delta z}$	média da distribuição de $\Delta z$	-
$\sigma_{\Delta z}$	desvio padrão da distribuição de $\Delta z$	-
$\Delta z_{50}$	mediana da distribuição de $\Delta z$	-
$\sigma_{68}$	metade do intervalo em torno de $\Delta z_{50}$ (68% dos dados)	$< 0.12$
$out_{2\sigma}$	fração de galáxias com: $ \Delta z - \overline{\Delta z}  > 2\sigma_{\Delta z}$	$< 0.1$
$out_{3\sigma}$	fração de galáxias com: $ \Delta z - \overline{\Delta z}  > 3\sigma_{\Delta z}$	$< 0.015$
$\overline{\Delta z'}$	Média da distribuição de $\Delta z' = \Delta z / \epsilon_{z_{phot}}$	-
$\sigma_{\Delta z'}$	Desvio padrão da distribuição de $\Delta z'$	-
$N_{Poisson}$	RMS de $\Delta N(z)$ , normalizado por flutuações poissônicas	-
$KS$	Teste Kolmogorov–Smirnov a para $N(z)_{photo-z}$ , $N(z)_{spec-z}$	-

## 2.3 Determinação da Distância das Galáxias

Para calcular a magnitude absoluta, é necessária a determinação da distância da galáxia, estimada a partir do *redshift*. Como a separação entre dois pontos do sistema de coordenadas varia com o tempo, o cálculo da distância comóvel deve considerar esta variação durante a trajetória do fóton, sendo descrita pelo fator de escala  $R(t)$ :

$$d_c = \int_0^r R(t) dr = \int_0^z \frac{c}{H(z)} dz, \quad (2.2)$$

onde o  $r$  é a coordenada radial da posição onde o objeto observado se encontra (tomando o observador na origem).  $R(t)$  é o fator de conversão para  $r$  em um determinado instante  $t$  (todas as quantidades com índices 0 se referem ao tempo presente):

$$r(t) = R(t)r(t_0)$$

$H(z)$  é o parâmetro de Hubble, dependente do tempo, e que é escrito em termos do *redshift* pela expressão 2.3, que é obtida da equação de Friedmann, considerando um modelo de Universo dominado por matéria e com constante cosmológica.

$$H = H_0 \left[ \Omega_m (1+z)^3 + \Omega_k (1+z)^2 + \Omega_\Lambda \right]^{1/2}. \quad (2.3)$$

Neste trabalho consideramos  $H_0 = 70 \text{ kms}^{-1}\text{Mpc}^{-1}$  e os parâmetros cosmológicos:  $\Omega_k = 0$ ,  $\Omega_m = 0.3$  e  $\Omega_\Lambda = 0.7$ .

## 2.4 Cálculo da Magnitude Absoluta

A magnitude absoluta  $M$  é definida como a magnitude aparente  $m$  que um objeto teria caso estivesse a uma distância de 10 pc, sendo obtida pela expressão:

$$m - M = 5 \log(d_L) - 5,$$

onde  $m$  é a magnitude aparente observada e  $d_L$  (distância de luminosidade) é a distância que provocaria uma redução no brilho aparente, levando ao fluxo observado  $F$ , seguindo a lei do inverso do quadrado da distância:

$$F = \frac{L}{4\pi d_L^2},$$

que corresponde à distância comóvel  $d_c$ , obtida das equações 2.2 e 2.3, corrigida por um fator de  $(1 + z)$ :

$$M = m - 5 \log(d_c(1 + z)) + 5.$$

Para objetos extragalácticos, a unidade mais conveniente para a distância é o Mpc ( $= 10^6$ pc). Sendo assim, a expressão para a magnitude absoluta torna-se:

$$M = m - 5 \log(d_c(1 + z)) - 25,$$

ainda sem levar em conta que a luz sofre extinção no caminho percorrido. Além disso, ela é válida para o caso onde a fonte e o observador estão em referenciais comoveis. Para corrigir a magnitude desses efeitos, dois termos são incorporados à equação acima:  $A$  para extinção e  $K(z)$  para o efeito provocado pelo movimento relativo entre fonte e observador:

$$M = m - 5 \log(d_c(1 + z)) - 25 - A - K(z). \quad (2.4)$$

A extinção  $A$ , da luz vinda de uma galáxia distante ocorre na própria galáxia e na Via Láctea. O efeito do meio intergaláctico, neste caso, pode ser desprezado. O gás e a poeira presentes na Via Láctea absorvem e espalham parte da luz vinda de objetos distantes. A extinção também é dependente do comprimento de onda (as bandas mais azuis sofrem mais extinção que as mais vermelhas).

A profundidade óptica da Via Láctea depende fortemente da latitude galáctica,  $b$ . O gás e a poeira, responsáveis pela extinção, estão mais concentrados no disco galáctico. Mapas disponibilizados em Schlegel, Finkbeiner & Davis (1998) [70] foram utilizados para obter os valores de  $A_{MW}$ , dependendo da posição no céu.

A correção  $K$  desfaz a diferença na magnitude aparente produzida pelo *redshift*. Quando uma distribuição espectral de energia (SED) é deslocada, com respeito uma banda em um sistema fotométrico em repouso. A expressão que determina o valor da correção  $K$  é composta por dois termos independentes. O primeiro termo corrige o efeito devido à diferença entre os comprimentos de onda da luz nos dois referenciais (do objeto e do observador). Sendo  $F(\lambda)$  a função que descreve o fluxo de uma galáxia em função do comprimento de onda (essencialmente a SED), o fluxo no

$\lambda$  medido no referencial de repouso de uma galáxia a um *redshift*  $z$ ,  $\lambda_z^{ef}$ , transformado para o referencial do detector como  $\lambda_{obs}^{ef}/(1+z)$  é dado por:

$$F(\lambda_{obs}^{ef}) \propto \frac{\int_0^\infty F(\lambda_{obs}^{ef}) S_i(\lambda) d\lambda}{\int_0^\infty F\left[\frac{\lambda_{obs}^{ef}}{(1+z)}\right] S_i(\lambda) d\lambda} F(\lambda_z^{ef}). \quad (2.5)$$

onde  $S_i(\lambda)$  é a sensibilidade do filtro na banda  $i$ , que depende de  $\lambda$ . O segundo termo corrige o efeito devido à diferença na largura (faixa de  $\lambda$ ) do sistema de detecção nos dois referenciais. A largura efetiva da banda  $i$  do detector vai ser diferente nos dois referenciais, por um fator  $(1+z)$ , fazendo com que o fluxo medido seja também alterado por este fator:

$$F(\lambda_{obs}^{ef}) \propto (1+z) F(\lambda_z^{ef}). \quad (2.6)$$

O primeiro efeito (eq.: 2.5) é seletivo, depende de  $\lambda$ , sendo nulo caso  $F(\lambda) = c^{te}$ . O segundo efeito (eq.: 2.6) independe da forma de  $F(\lambda)$ . Juntando as duas contribuições, e aplicando o logaritmo para converter o fluxo em magnitude, temos a expressão para o cálculo da correção K (Pence 1976 [62]; Oke & Sandage 1968 [58]):

$$K_i = 2.5 \log(1+z) + 2.5 \log \left\{ \frac{\int_0^\infty F(\lambda_{obs}^{ef}) S_i(\lambda) d\lambda}{\int_0^\infty F\left[\frac{\lambda_{obs}^{ef}}{(1+z)}\right] S_i(\lambda) d\lambda} \right\}.$$

## 2.5 A Função de Luminosidade

A LF ( $\phi(L)$ ) é definida como a quantidade de galáxias com luminosidade intrínseca no intervalo entre  $L$  e  $L + dL$ , por unidade de volume. É comum denotarmos a LF em termos da magnitude absoluta ( $\phi(M)$ ), o que é conceitualmente equivalente à abordagem em função da luminosidade ( $\phi(L)dL = \phi(M)dM$ ). A integral da LF nos fornece a quantidade total de luminosidade presente na amostra.

Em 1976, Paul Schechter propôs uma expressão analítica para a parametrização da LF global (para todos os tipos morfológicos ou espectrais) que se ajustou tão bem aos dados observacionais, que continua sendo a abordagem mais utilizada até os dias de hoje. A função de Schechter (eq.:2.7) possui dois regimes: uma lei de potências para galáxias de baixa luminosidade (muito numerosas) e um regime exponencial para galáxias de alta luminosidade (muito raras) como mostrada no painel 2.3a.

$$\phi(L)dL = \phi^* \left(\frac{L}{L^*}\right)^\alpha \exp\left(-\frac{L}{L^*}\right) d\left(\frac{L}{L^*}\right). \quad (2.7)$$

A inclinação da parte fraca é dada pelo parâmetro  $\alpha$  e o ponto que define a transição de um regime para o outro é definido pelos parâmetros  $L^*$  (ou  $M^*$ ) e  $\phi^*$ , sendo estes a luminosidade característica e a normalização dada em  $L^*$ , respectivamente. A maior fração da luminosidade do Universo está em galáxias com luminosidade da ordem de  $L^*$ . Uma aproximação ingênua, porém razoável, seria a consideração de um Universo povoado apenas com galáxias de  $L = L^*$  e com densidade igual a  $\phi^*$ .

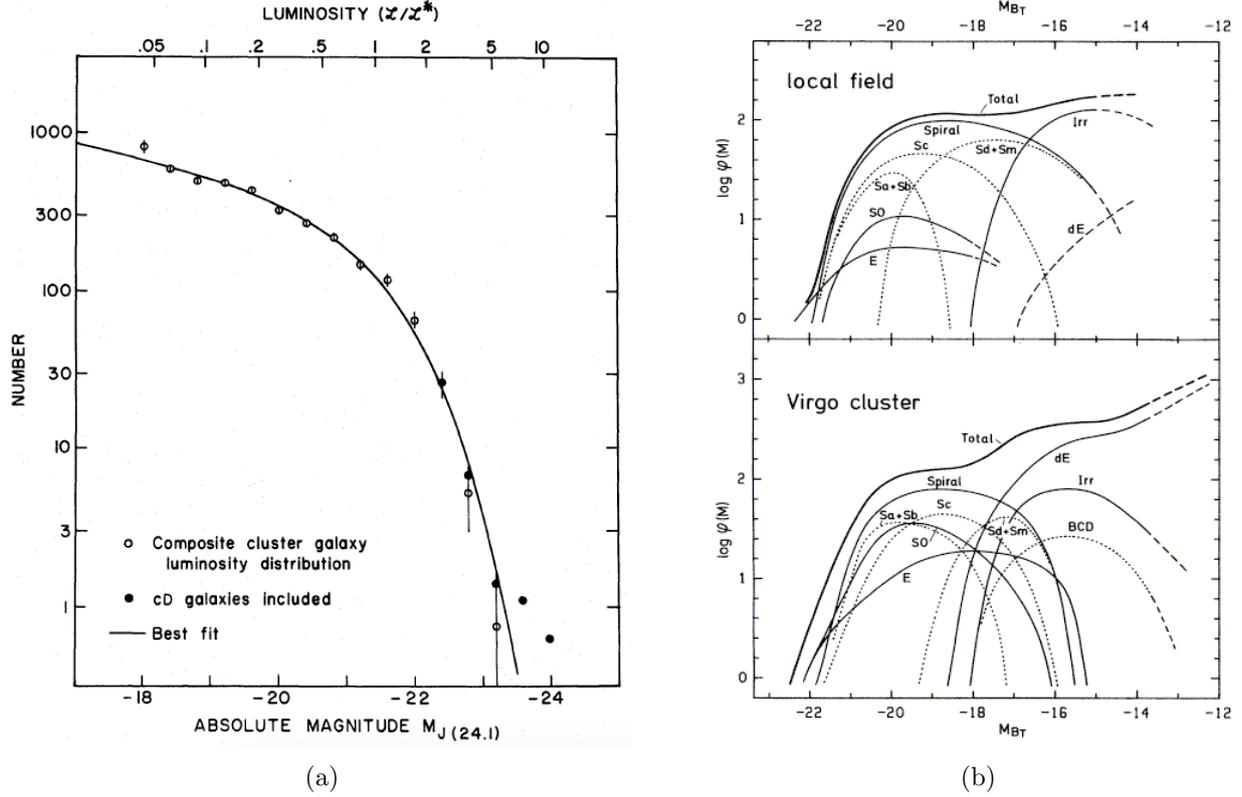


Figura 2.3: 2.3a:Gráfico original do artigo : Schechter 1976 [69], onde o autor propõe a parametrização da LF com a expressão da equação 2.7. 2.3b: LF dependente do tipo morfológico (classificação de Hubble para galáxias do Universo local) e do ambiente. Gráfico retirado do artigo de revisão de Binggeli, Sandage & Tammann, 1988 [13].

Estudos mais recentes (Wolf et al. 2003 [76]; Lilly et al. 2007 [48]; Zucca et al. 2009 [81]) mostram que a LF depende fortemente do tipo espectral e do ambiente, neste caso reforçando a ideia de dependência entre o ambiente e a predominância de determinados tipos espectrais. A LF global é o resultado da combinação das LF's para os diferentes tipos espectrais. Ela também reflete a soma das diferentes proporções de tipos encontrados em cada ambiente (no campo ou em aglomerados).

Por exemplo, para o Universo local, a LF das galáxias espirais e S0's possuem um formato aproximadamente gaussiano; para as elípticas, a LF se comporta como uma gaussiana enviesada; já para as galáxias espirais e irregulares, a LF segue fielmente o comportamento da função de Schechter (painel 2.3b).

A expressão da função de Schechter não é fruto de um processo físico específico, mas também não é uma conjectura 100% empírica, encontrada ao acaso. A formação de galáxias deriva de dois ingredientes fundamentais: a aglutinação de halos de matéria escura e a queda de bárions para o núcleo destes halos, necessários para a formação estelar. O formalismo Press-Schechter (ver eq.:2.8), descrito por William H. Press e Paul Schechter (1974) [64], é um modelo matemático que prevê o número de halos de matéria escura, em um dado instante, dependendo do número inicial de flutuações de densidade acima de um certo valor mínimo no Universo primordial.

$$N(M)dM = \frac{1}{\sqrt{\pi}} \left(1 + \frac{n}{3}\right) \frac{\bar{\rho}}{M^2} \left(\frac{M}{M^*}\right)^{(3+n)/6} \exp\left[-\left(\frac{M}{M^*}\right)^{(3+n)/3}\right] dM, \quad (2.8)$$

onde:  $\bar{\rho}$  é a densidade média de matéria no Universo (escura e bariônica);  $n$  é o índice do espectro de potências das flutuações do Universo primordial ( $P(k) \propto k^n$ ); e  $M^*$  é a massa crítica, acima da qual as estruturas vão se formar devido ao colapso gravitacional.

No contexto do modelo cosmológico  $\Lambda$ CDM, flutuações na densidade de energia foram impressas sobre o Universo jovem. Com o passar do tempo, estas flutuações tornam-se perturbações em massa, que crescem de forma linear. Só muito tempo depois, começando com escalas de massa pequenas e avançando ao longo do tempo para escalas maiores de massa, as perturbações entram em colapso gravitacional, para formar galáxias ou aglomerados de galáxias, no chamado modelo hierárquico de formação de estruturas.

O comportamento previsto pelo formalismo é que a distribuição de massa siga uma lei de potência para as pequenas massas, com um corte exponencial acima de uma massa característica que aumenta com o tempo, semelhante ao comportamento da LF global. Todavia, embora a luminosidade seja relacionada com a massa bariônica, que por sua vez se relaciona com a massa dos halos de matéria escura, o formato da distribuição de massas nos halos prevista pelo formalismo de Press-Schechter não é idêntico ao formato da LF global observada.

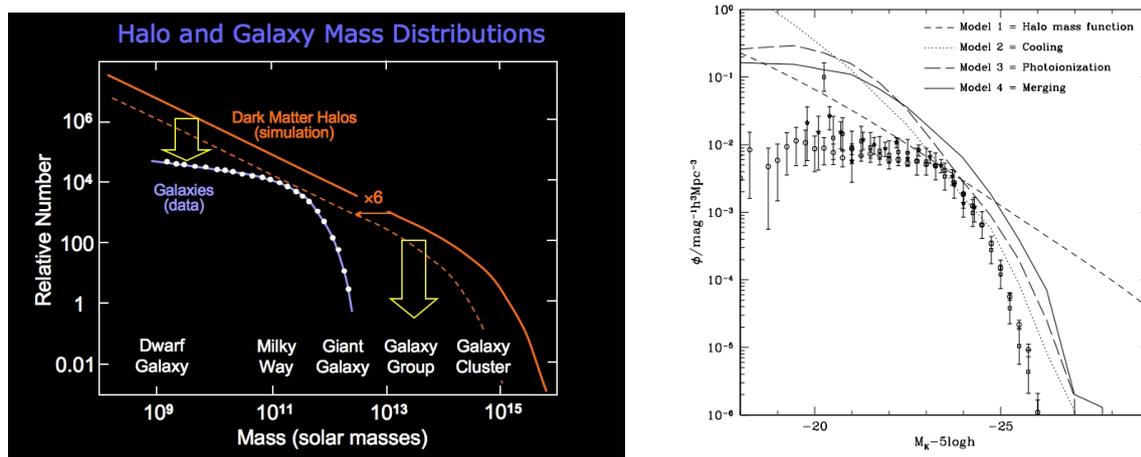


Figura 2.4: 2.4a: representação da diferença nos formatos da LF de galáxias e da função de massa de halos prevista pelo formalismo de Press-Schechter. 2.4b: exemplo com dados observados (Benson et al. 2003 [8])

Na comparação das duas curvas, há um déficit na LF global, nas duas extremidades da função (painéis 2.4a<sup>1</sup> e 2.4b). A explicação pode estar em diferentes processos físicos, nas diferentes escalas de massa ou estrutura. Nas menores escalas, para halos de massa da ordem da massa de galáxias anãs, a reionização do Universo teria aumentado a temperatura e a pressão para valores que dificultariam a formação dos pequenos halos. Para halos um pouco maiores, da ordem da massa da Via Láctea, surtos de formação estelar e ventos de supernovas poderiam ser os responsáveis por expulsar uma fração considerável dos bárions, tornando a massa total da estrutura menor que a prevista pelo formalismo. Para halos ainda maiores, da ordem da massa de grupos de galáxias, jatos de núcleos ativos poderiam ter reaquecido o gás que precisaria ser resfriado para poder formar halos menores (porém maciços). Em grande escala, para os halos da ordem dos aglomerados de

<sup>1</sup>[https://www.astro.virginia.edu/class/whittle/astr553/Topic04/t4\\_LF\\_origin\\_B.html](https://www.astro.virginia.edu/class/whittle/astr553/Topic04/t4_LF_origin_B.html)

galáxias, o potencial gravitacional no interior dos aglomerados seria grande o suficiente para que os bárions que caem no poço de potencial se tornem quentes demais para conseguir colapsar para formar estruturas (Benson et al. 2003 [8]).

### 2.5.1 O Método $1/V_{\text{máx}}$

O método  $1/V_{\text{máx}}$  é adotado para corrigir os efeitos de se ter uma amostra limitada em fluxo. O volume observado que contém as galáxias da amostra depende de suas luminosidades. As galáxias de alto brilho intrínseco são detectadas a grandes distâncias, enquanto as galáxias de baixo brilho, são vista apenas em distâncias mais próximas.

Sendo assim, o volume considerado no cálculo de  $\phi(L)$  (ou  $\phi(M)$ ) deve ser maior ao contabilizar as galáxias mais brilhantes e menor para as mais fracas.

O método  $1/V_{\text{máx}}$  consiste em ponderar a contagem das galáxias justamente pelo volume em que cada uma pertenceria à amostra,  $V_{\text{máx}}$  (considerando a sua magnitude absoluta e a magnitude aparente limite da amostra,  $m_{\text{lim}}$ ), assim, a correção é feita para todas as galáxias, uma a uma.

Considerando uma amostra com  $N$  galáxias cobrindo um ângulo sólido  $\Omega$ , com magnitudes aparentes até um dado limite  $m_{\text{lim}}$  e num intervalo de *redshift* entre  $z_{\text{min}}$  e  $z_{\text{max}}$ , em cada intervalo de magnitude absoluta de largura  $\Delta M = M_2 - M_1$ , a função  $\phi(M)dM$  deve ser tal que:

$$N = \int_{M_1}^{M_2} \phi(M)dM V_{\text{máx}} = \Phi(M)V_{\text{máx}},$$

onde  $V_{\text{máx}}$  é definido pela distância de cada objeto. Para pequenos valores de  $\Delta M$ , temos que:

$$\phi(M) = \sum \frac{1}{V_{\text{máx}}}.$$

Uma limitação deste método é a de que ele assume que as galáxias estão uniformemente distribuídas no céu, sem considerar as heterogeneidades da amostra. Portanto ela não é aconselhada para amostras coletadas de áreas pequenas, que sofrem efeitos da variância cósmica.

### 2.5.2 O Ajuste à Função de Schechter

Após calcular o valor de  $\phi(M)$  em cada intervalo de magnitude absoluta, o próximo passo é ajustar a curva da função de Schechter aos pontos obtidos, encontrando assim os parâmetros  $\phi^*$ ,  $M^*$  e  $\alpha$ . A função de Schechter em termos das magnitudes absolutas é dada pela expressão:

$$\phi(M)dM = (-0.4 \ln 10)\phi^*10^{0.4(\alpha+1)(M^*-M)} \exp -10^{0.4(M^*-M)}dM,$$

que deriva da equação 2.7 considerando a relação:

$$\log \frac{L}{L^*} = -0.4(M - M^*).$$

O ajuste é feito utilizando o método dos mínimos quadrados, com os três parâmetros livres para os 4 primeiros intervalos de  $z$ . Nos dois últimos, devido à dificuldade de caracterização do parâmetro  $\alpha$  em altos *redshifts*, adotamos o mesmo procedimento utilizado em Ramos et al (2011)

[66], fixando o valor de  $\alpha$  como sendo o obtido para o intervalo anterior (o quarto intervalo, que corresponde a  $0.6 < z < 0.8$ ).

# Capítulo 3

## Ferramentas Computacionais

Neste capítulo apresentamos as ferramentas usadas para o estudo de GE, que envolvem tanto software como hardware. Na parte de software temos códigos de estimativa de *photo-z*, e de cálculo da função de luminosidade. Na parte de hardware temos um *cluster* de processamento com mil *cores*, e uma unidade de grande capacidade de armazenamento. Estes recursos estão integrados através de um Portal Científico que pode ser acessado por navegador de internet. Nas seções abaixo descrevemos estas ferramentas.

### 3.1 Algoritmos para estimativa de *photo-z*

Como mencionado em capítulos anteriores, uma forma de se estudar a evolução de galáxias pode ser feita através da Função de Luminosidade. Pelo fato de usarmos dados do levantamento DES e este só provê o imageamento para as galáxias, temos que fazer a determinação de *redshifts* através da informação fotométrica. Para isso foi desenvolvido no Portal Científico do DES uma suíte de programas disponibilizados por diversos autores para o cálculo de *photo-z*.

A natureza dos programas de *photo-z* pode ser dividida em duas grandes categorias, levando-se em conta sua metodologia - métodos empíricos e os de ajuste a modelos, que descrevemos a seguir. A efetiva aplicação destes códigos aos dados de simulações e de observações do DES serão apresentados em capítulo subsequente.

#### 3.1.1 Métodos Empíricos

##### 3.1.1.1 ANNz - *Artificial Neural Networks for Redshift*

O ANNz (Collister & Lahav 2004 [24]) é um algoritmo empírico de *photo-z*, baseado em redes neurais artificiais. A partir de um conjunto de treinamento, onde o *redshift* espectroscópico (*spec-z*) é conhecido, o código utiliza um conjunto de redes neurais para encontrar uma relação funcional entre os observáveis que podem ser, por exemplo, magnitudes, cores, parâmetros de forma, entre outros e a quantidade desejada, neste caso, o *photo-z*.

A rede neural possui um certo número de camadas e cada camada possui um certo número de nós. Estas quantidades podem ser escolhidas pelo usuário no início do processamento, quando se define a arquitetura da rede. A primeira camada recebe a informação dos dados de entrada,

neste caso as magnitudes. A última camada fornece o valor do *photo-z* estimado (ver esquema na Fig. 3.1). A incerteza associada ao resultado é calculada através da propagação dos erros dos observáveis, segundo a relação funcional encontrada pelo algoritmo.

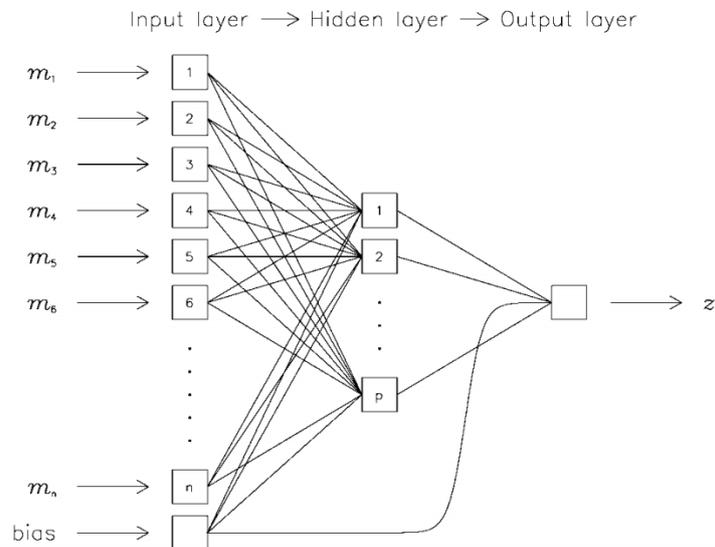


Figura 3.1: Esquema de rede neural artificial, mostrando as diferentes camadas: a de *inputs*, com vários parâmetros de entrada, a(s) camada(s) oculta(s), onde o algoritmo realiza todos os processos de treinamento e aplicação de pesos e, por fim, a camada de *output* com um único parâmetro, que é o *redshift* determinado, acompanhado de uma estimativa para o ser erro (figura retirada de Collister & Lahav (2004) [24])

Os nós são interconectados e cada conexão carrega um peso - um parâmetro livre na parametrização. Durante o treinamento, todos os pesos são determinados através da minimização de uma função de custo  $\mathcal{E}$ . Esses pesos, então, são gravados num arquivo que servirá de dado de entrada na próxima etapa. Para aumentar a precisão do *photo-z* estimado, o processo de treinamento pode ser feito repetidas vezes, criando assim um conjunto de redes definido por um conjunto de arquivos de saída. Uma vez gerados, eles podem ser aplicados a diferentes conjuntos de teste, sem a necessidade da repetição do treinamento para cada teste.

O algoritmo usa um método iterativo para fazer a minimização de  $\mathcal{E}$ , que é desencadeado a partir de uma semente dada pelo usuário. A cada iteração, a função  $\mathcal{E}$  é calculada para o conjunto de treinamento e também para um outro conjunto, de validação, onde o  $z$  também é conhecido. Ao atingir o número máximo de iterações definido pelo usuário, os pesos são determinados a partir da iteração onde a função  $\mathcal{E}$  atingiu o valor mínimo no conjunto de validação. Esta abordagem é útil para evitar o ajuste excessivo ao conjunto de treinamento, quando este é relativamente pequeno.

Assim como nos outros algoritmos baseados em métodos empíricos, a qualidade dos resultados gerados pelo ANNz é fortemente dependente do conjunto de treinamento. O código é competitivo apenas quando o conjunto de treinamento é suficientemente grande e representativo do conjunto onde o *photo-z* será medido. Segundo os autores, uma possível solução para lidar com a falta de dados de treinamento em algum domínio de cor ou *redshift*, nos casos em que a obtenção do dado espectroscópico é difícil, seria a utilização de dados simulados (Collister & Lahav 2004 [24]).

### 3.1.1.2 ArborZ - *Photometric Redshifts Using Boosted Decision Trees*

ArborZ (Gerdes et al. 2010 [38]) é um algoritmo que usa o método de árvores de decisão indutiva (*Boosted Decision Trees* - BDTs) para determinar o mapeamento empírico de observáveis para *redshift*. Assim como a maioria dos algoritmos de aprendizado de máquina, ArborZ determina esse mapeamento durante uma etapa de treinamento que antecede a fase de estimativa dos *photo-z*'s. Nesta etapa, o conjunto de treinamento é dividido em faixas de *redshift*, onde a largura de cada faixa é aproximadamente a metade da resolução esperada para o *photo-z*.

Para cada faixa de *redshift*, o código percorre todo o conjunto de treinamento, dando uma identificação positiva para as  $N$  galáxias cujo *redshift* cai naquela faixa, formando o conjunto “*signal*”.  $5N$  galáxias são aleatoriamente escolhidas entre as que possuem *redshift* a uma distância maior do que  $3\eta$  do centro da faixa (onde  $\eta$  é a resolução do *photo-z* escolhida pelo usuário). Essas galáxias compõem o chamado conjunto “*background*”.

Uma árvore de decisão individual tem pouco poder de classificação, devido a aleatoriedade das amostragens e da incerteza introduzida pelo tamanho de  $\eta$ . Entretanto, a “floresta” criada durante o processo de treinamento é a combinação ponderada de várias árvores, onde as que possuem menor índice de erro na classificação ganham um peso maior, tornando o método poderoso. O processo de separação do conjunto entre *signal* e *background* é então feito repetidamente, e assim o algoritmo vai aprendendo quais variáveis, dentre os observáveis, melhor separam os dois conjuntos.

Na etapa de teste, cada galáxia tem seus observáveis confrontados com as árvores de classificação em cada faixa de *redshift*, recebendo uma probabilidade pertencer a ela. Desse modo, o código fornece a função “densidade de probabilidade” (PDF) de *redshift* para cada objeto. A estimativa do *photo-z* vem da média da PDF e erro é estimado a partir da largura em torno do pico da PDF que contem 68% da área sob a curva.

### 3.1.1.3 TPZ - *Trees for Photo-Z*

TPZ (Carrasco Kind & Brunner 2013 [19]) é um método empírico baseado em florestas de árvores de decisão. O código é rápido, robusto, funciona de forma paralelizada e possui um sistema eficiente de armazenar as PDFs com economia de espaço.

As árvores de decisão são criadas a partir de uma sequência de perguntas que repartem recursivamente o multi-espaço de observáveis em dois ramos. Cada ramificação da árvore de decisão divide o conjunto de treinamento em regiões menores (Figura 3.2). Ao atingirem algum critério pré-definido para interromper as divisões (por exemplo um tamanho mínimo do último subconjunto, ou um limiar mínimo na variância), as sucessivas ramificações terminam em “folhas”, que são pequenas regiões do multi-espaço, com apenas poucos elementos de propriedades semelhantes. Cada subconjunto é um classificador à parte. A acurácia do método vem da combinação de várias árvores em cada estimativa, formando a “floresta” de árvores de decisão. Neste caso o resultado é a média dos resultados sugeridos por cada árvore individual.

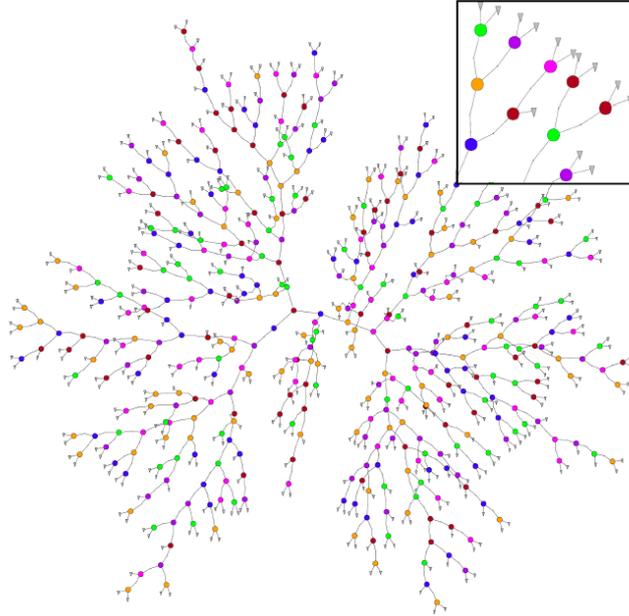


Figura 3.2: Esquema simplificado de uma árvore de decisões binárias, desenhadas radialmente. Cada cor representa um observável diferente. Cada extremidade (“folha”) contém uma predição específica, baseada na informação contida no nó terminal. (figura retirada de Carrasco Kind & Brunner (2013) [19])

Como resultado, além de um valor para o *redshift* e uma distribuição  $P(z)$  para cada galáxia, o código também fornece informações extras sobre a amostra, como um *ranking* de importância das variáveis, identificação de *outliers* e de áreas no multi-espaco de observáveis onde a estimativa do *photo-z* é menos confiável (áreas menos representadas pelo conjunto de treinamento). Os erros são estimados a partir da perturbação dos parâmetros das galáxias, da ordem de suas incertezas. Esse processo é repetido, simulando várias “novas observações” de cada galáxia. O erro é então estimado com base na variância do conjunto de resultados pós-perturbação.

### 3.1.1.4 SkyNet

SkyNet (Graff et al. 2014 [39] e Bonnett 2015 [14]) também é um método que utiliza redes neurais artificiais para encontrar um ou mais valores estimados a partir de um conjunto de observáveis. Ele é uma ferramenta ampla, capaz de realizar procedimentos tanto de maneira supervisionada, quanto não-supervisionada de aprendizagem de máquina para realizar diversas tarefas úteis para astronomia, como regressão, classificação, estimativa de densidade espacial e de aglomeração além de redução dimensional (Graff et al. 2014 [39]).

O SkyNet também realiza uma etapa de treinamento, onde o algoritmo mapeia a relação entre as entradas e as saídas através de conexões entre os nós da rede neural de forma ponderada. Aplicado ao cálculo de *redshift*, SkyNet possui um método de classificação onde associa uma probabilidade de que o objeto estudado pertença a uma dada classe. Definindo então as classes como intervalos de *redshift*, o conjunto de probabilidades em todos os intervalos forma a PDF de cada objeto.

### 3.1.1.5 *Weights*

Diferente dos outros métodos, o código *Weights* (Lima et al. 2008 [49]) foi idealizado para estimar a distribuição de objetos no espaço de *redshifts*,  $N(z)$ , ao invés de atribuir individualmente um valor de *photo-z* para cada galáxia. Segundo os autores, o  $N(z)$  tradicionalmente calculado por outros métodos, a partir do conjunto de estimativas de  $z$ , é enviesado pelos erros do *photo-z*. Em alguns casos (como o estudo de lentes gravitacionais fracas ou o estudo das oscilações acústicas dos bárions, por exemplo), é requerida apenas a distribuição  $N(z)$ , não sendo necessária a determinação dos valores individuais, o código foi desenvolvido de forma a contornar o viés introduzido no cálculo do *photo-z*.

O método consiste em associar pesos às galáxias do conjunto de treinamento, de modo que as distribuições de seus observáveis, quando ponderadas por estes pesos, sejam correspondentes às distribuições dos mesmos, no conjunto fotométrico. Os pesos são determinados a partir da técnica de identificação dos vizinhos mais próximos no espaço multidimensional de cores e magnitudes, o que garante a estabilidade do método em regiões pouco povoadas. Nesta técnica, são comparadas as densidades nos dois conjuntos, espectroscópico (de treinamento) e fotométrico, para as mesmas regiões do multi-espaço de observáveis. O peso atribuído a cada galáxia é a razão entre a densidade de galáxias vizinhas em um volume definido em torno dela no espaço de magnitudes do conjunto fotométrico e a densidade de vizinhos em um mesmo volume do conjunto de treinamento, normalizada pelo número total de galáxias do conjunto fotométrico (Cunha et al. 2009 [28]).

A vantagem do código é a possibilidade de lidar com um conjunto de treinamento não representativo do conjunto a ser estudado, com a ressalva de que este deve cobrir toda a faixa de observáveis por onde se estende o conjunto fotométrico, mesmo que contenha regiões pouco povoadas. A distribuição  $N(z)$  global da amostra é construída simplesmente através da soma dos pesos derivados em intervalos de *redshift*.

Este algoritmo é utilizado como etapa de treinamento do próximo código a ser descrito (Pofz) e pode também ser usado para selecionar ou rearranjar objetos em um conjunto de treinamento qualquer, tornando-o mais representativo do conjunto fotométrico que se quer estudar.

### 3.1.1.6 *Pofz*

Pofz (Cunha et al. 2009 [28]) é um código simples, que aproveita os pesos gerados na saída do código descrito anteriormente (*Weights*) para montar a PDF de cada galáxia. Segundo os autores, os estimadores tradicionais de *photo-z* são limitados por assumirem a premissa de que existe uma relação funcional entre os observáveis e o *redshift*. Entretanto, galáxias que ocupam uma pequena região no espaço de observáveis podem ter *redshifts* bem variados.

É possível associar esta região a uma PDF em função do *redshift* e dos observáveis. O formato da função seria determinada pela escolha dos observáveis, tamanho da região, erros fotométricos e a diversidade de SEDs das galáxias pertencentes a amostra. Caso a curva de probabilidade possua um pico estreito, a estimativa do *photo-z* seria precisa e acurada. No entanto se o pico for muito largo ou assimétrico, ou ainda, se a distribuição possuir múltiplos picos, a estimativa do *photo-z* terá grandes incertezas, o que potencialmente leva a erros catastróficos.

As tendências de superestimar o *photo-z* das galáxias em baixo *redshift* de subestimar o *photo-z*

das galáxias em alto *redshift* são consequências deste pressuposto fundamental – quando o pico principal da PDF ocorre em um valor de  $z$  baixo, se houver múltiplos picos, a probabilidade é maior de que eles sejam situados a *redshifts* mais altos que o do pico principal, levando a uma superestimativa do *redshift*. O mesmo raciocínio vale para quando o pico principal ocorre em  $z$  muito alto, e os demais picos em menores  $z$ s trazem o resultado para um valor subestimado.

Além disso, objetos com diferentes valores de *redshift* podem ocupar a mesma região no espaço de observáveis, por isso os estimadores lhes atribuirão o mesmo valor para o *photo-z*. Neste caso, para evitar o viés intrínseco à estimativa, é preferível usar a PDF das galáxias ( $P(z) \equiv P(z|\text{observáveis})$ ). Posteriormente, o valor individual do *photo-z* pode ser extraído da distribuição de probabilidades através de sua média, mediana, ou moda.

O método de pesos descrito anteriormente, leva diretamente a uma estimativa do  $P(z)$  de cada galáxia do conjunto fotométrico, que é dada a partir da distribuição ponderada de *redshifts* dos seus vizinhos mais próximos (no espaço de observáveis pertencentes ao conjunto de treinamento. Adicionalmente, a distribuição  $N(z)$  global pode ser obtida através do somatório das funções  $P(z)$  de todas as galáxias do conjunto fotométrico.

### 3.1.1.7 DESDM

O programa DESDM calcula os *photo-z*'s usando a mesma rede neural artificial utilizada na liberação de dados DR6 do levantamento *Sloan Digital Sky Survey*, descrita com detalhes em Oyaizu (2008) [59]. O programa leva o mesmo nome do *software* de gerenciamento de dados do DES pois ele é integrado ao *pipeline* de redução das imagens.

O processo de minimização para determinar os pesos da rede neural é feita sobre os dados do conjunto de treinamento, escolhendo o conjunto de pesos que fornecem a menor dispersão no *photo-z* quando aplicado ao conjunto de validação, após 300 iterações. Os pesos são gerados de forma iterativa, partindo de uma semente (um número aleatório qualquer) dada pelo usuário. O método permite a repetição do procedimento, partindo de diferentes sementes, para reduzir a sensibilidade às condições iniciais no procedimento de minimização. O *photo-z* é então tomado como sendo a média dos *photo-z*s calculadas a partir dos pesos otimizados para cada uma das 10 minimizações de rede.

A estimativa do erro no *photo-z* é feita utilizando o método empírico *nearest neighbor error* (NNE), descrito em Oyaizu (2008) [59]. O NNE estima o erro baseado na diferença entre o *photo-z* e o *spec-z* das 100 galáxias vizinhas mais próximas no multi-espaço de observáveis. O erro é definido como a largura correspondente a 68% de vizinhos próximos na distribuição de  $|z_{phot} - z_{spec}|$ .

### 3.1.1.8 RVMz

RVMz é um código empírico com base no algoritmo de máquina de vetores de relevância (Tipping, 2001 [72]), um método bayesiano de kernel para a regressão. A máquina de vetor de relevância (RVM) tem características semelhantes à máquina de vetor de suporte, mas inclui um tratamento bayesiano para a determinação dos pesos do modelo. Isto tem a vantagem de que os parâmetros que governam a complexidade do modelo e da variância do ruído são encontrados na própria execução, e, por conseguinte, a RVM não requerem validação cruzada para otimizar estes

parâmetros. A PDF é reconstruída pela combinação das incertezas nos conjuntos de dados e do modelo. O treinamento consiste em compartimentar os dados em  $k$  grupos, então  $k - 1$  desses grupos são usados para formar o modelo que é então avaliado pelo grupo faltante. Este processo é então repetido para todas as escolhas possíveis de grupos e o erro quadrático médio de predição resultante para *photo-z* é avaliada. Nesta fase, obtém-se o erro de modelo como o RMS da PDF previsto. Os detalhes do método estão sendo descritos em Rau et al. (em prep.).

### 3.1.1.9 NIP-kNNz

NIP-kNNz (Juan De Vicente et al. (2015), em preparação) calcula o *photo-z* a partir da busca pelo vizinho mais próximo com base no produto interno normalizado (NIP)(eq.: 3.1). Enquanto a distância euclidiana garante para que galáxias próximas no espaço de magnitudes tenha sido atribuído o mesmo *redshift*, o algoritmo não considera como vizinhos galáxias com a mesma cor, mas separados em magnitude global. Métricas sobre o NIP corrigem isto considerando duas galáxias como vizinhas quando elas têm cores semelhantes (ao invés de magnitudes), e, portanto, com *redshift* próximo. A métrica baseia-se na definição do produto interno:

$$\text{NIP} = \cos \alpha = \frac{\mathbf{M}_t \cdot \mathbf{M}_p}{M_t M_p}, \quad (3.1)$$

onde  $\mathbf{M}_t$  e  $\mathbf{M}_p$  são vetores no multi-espaço de magnitudes dos conjuntos de treinamento e fotométrico, respectivamente e  $\alpha$  é o “ângulo” entre os dois vetores (cujo cosseno define o produto interno).

Para estimar o erro do *photo-z*, o programa utiliza uma fórmula empírica que foi derivada para contabilizar três contribuições diferentes: (i) o primeiro termo considera a precisão do *redshift* espectroscópico; (ii) a segunda contribuição vem da caracterização dos erros do *photo-z* na amostra espectroscópica. NIP-kNN é rodado para todas as galáxias da amostra espectroscópica para obter seu *photo-z*. O erro no *photo-z* é estimado como a metade da diferença entre a *photo-z* e *spec-z* galáxia. Quando NIP-kNN é aplicado a uma galáxia na amostra fotométrica, ele herda não só o  $z$  das galáxias vizinhas, mas também o seu erro no *photo-z*; (iii) o terceiro termo é a distância métrica  $\sin(\alpha)$ , que considera a vizinhança de objetos semelhantes, multiplicada por uma constante determinada empiricamente.

### 3.1.1.10 ANNz2

ANNz2 é uma nova versão do *software* público ANNz (Collister & Lahav 2004 [24]). O novo código incorpora vários métodos de aprendizado de máquina, tais como redes neurais artificiais, árvores de decisão impulsionadas e busca por vizinhos mais próximos. Os diferentes algoritmos são utilizados em conjunto, a fim de otimizar o desempenho da reconstrução da distribuição  $N(z)$  e estimar as incertezas das soluções fotométricas. Isto é feito através da geração de uma grande variedade de soluções de aprendizado de máquina com, por exemplo, diferentes arquiteturas de rede, inicializadas por diferentes sementes aleatórias. A otimização é efetuada por ranking de soluções diferentes, de acordo com o seu desempenho, o qual é determinado pela respectiva dispersão do *photo-z* de cada solução. A solução com o melhor desempenho, ou seja, com menor dispersão no *photo-z*, é escolhida como o melhor estimador.

### 3.1.2 Método de Ajuste a Modelos

A seguir são descritos métodos que se utilizam de informação sobre a SED de objetos, que podem ser tanto de galáxias como de quasares.

#### 3.1.2.1 LePHARE - *Photometric Analysis for Redshift Estimate*

LePhare (Arnouts et al. 1999 [3] ; Ilbert et al. 2006 [44]) é um algoritmo essencialmente dedicado a estimar *photo-z*'s, mas que também pode ser usado para estimar outros parâmetros físicos como, por exemplo, a atribuição de uma classe espectral. Os modelos de SEDs em referencial de repouso, são sistematicamente desviados para o vermelho – seguindo limites inferior e superior, além de um passo, definidos pelo usuário – e convoluídos com as curvas de transmissão dos filtros (já considerando a eficiência do instrumento). Desta forma é possível calcular valores de fluxos nos diversos filtros como uma função do *redshift*. A opacidade do meio interestelar (Madau et al. 1996 [51]) também é considerada.

O ajuste dos fluxos estimados às SEDs é feito através da minimização da função de mérito  $\chi^2$ , pela condição:

$$\frac{\partial \chi^2}{\partial s} = 0 \quad \text{onde} \quad \chi^2 = \sum_I \left[ \frac{F_{obs,i} - sF_{mod,i}}{\sigma_i} \right]^2,$$

calculada a partir do fluxo observado,  $F_{obs}$  com sua incerteza  $\sigma$ , do fluxo estimado  $F_{mod}$  dado pelo modelo de uma SED, e do fator de normalização  $s$ , levando em conta os  $i$  filtros disponíveis (Ilbert et al. 2006 [44]).

$$s = \sum_i \left[ \frac{F_{obs,i} F_{mod,i}}{\sigma_i^2} \right] / \sum_i \left[ \frac{F_{mod,i}^2}{\sigma_i^2} \right].$$

Com isso já se associa, ao mesmo tempo, um *redshift* e um modelo de SED à cada galáxia, e consequentemente toda a informação que a determinada SED carrega, ou que dela se pode derivar, como por exemplo: tipo espectral, correção K, magnitude absoluta, massa, extinção, etc.

O código possui um método adaptativo de corrigir a diferença entre a calibração das magnitudes observadas e teóricas (dadas pelos modelos de SEDs). A partir de uma amostra com *redshifts* conhecidos, a calibração é obtida comparando as magnitudes teóricas com as observadas de forma iterativa, até que se atinja uma convergência nos valores dos “pontos zeros”. Esses valores são fornecidos junto com os resultados, como pequenos desvios que serão aplicados aos fluxos das SEDs antes do cálculo do *photo-z*. Tal procedimento pode ser feito separadamente para um conjunto de treinamento e os desvios obtidos posteriormente aplicados ao conjunto fotométrico. Os resultados podem ser aplicado na estimativa do *photo-z* no próprio conjunto de treinamento, de forma recursiva, o que é útil para medir a performance deste procedimento adicional. Segundo Hildebrandt et al. (2010) [41], a técnica ajuda a minimizar os efeitos sistemáticos das medidas de magnitude.

Além desta abordagem, há também a possibilidade de se utilizar informações de distribuições de *redshift* previamente conhecidas como mecanismo para minimizar os erros catastróficos. A distribuição disponível na biblioteca oferecida pelos autores do código se refere ao levantamento VVDS.

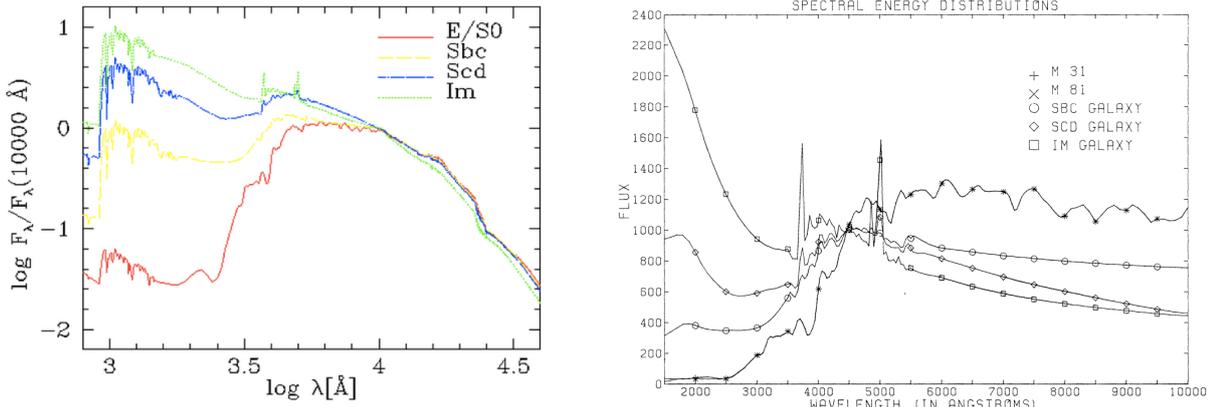


Figura 3.3: Exemplos de modelos de SEDs. 3.3a: exemplos representativos dos diferentes tipos morfológicos de galáxias contemplados no conjunto de onde foram selecionadas as SEDs utilizadas neste trabalho. 3.3b: figura original retirada de Coleman, Wu & Weedman (1980) [22]

Os modelos utilizados por nós perfazem um total de 29 SEDs, selecionadas do conjunto de 66 SEDs que foram utilizadas em [44] (Coleman, Wu & Weedman 1980 [22]) (12 elípticas, 16 espirais e 1 irregular). Esta seleção com um número menor modelos de galáxias elípticas minimiza a fração de galáxias que recebem classificação equivocada devido à degenerescência entre as SEDs de galáxias elípticas e de espirais corrigidas de extinção, uma consequência marcante da falta da banda  $u$  nas observações do DES. Estas conclusões foram herdadas de testes anteriores, utilizando amostras do levantamento VVDS através da comparação de resultados obtidos com e sem a utilização da banda  $u$ . Os painéis 3.3a e 3.3b mostram alguns exemplos de SEDs que representam os diferentes tipos espectrais previstos no conjunto utilizado.

### 3.1.2.2 BPZ - *Bayesian Photometric Redshifts*

O BPZ (Benítez 2000 [7]; Coe et al. 2006 [20]) é um método baseado em ajuste a modelos que retorna a distribuição de densidade de probabilidade  $P(z|m_{band})$  de uma dada galáxia possuir um redshift  $z$  quando suas magnitudes em cada banda são  $m_{band}$ . O programa também fornece um valor único de *photo-z*, calculado como o  $z$  correspondente ao máximo de  $P(z|m_{band})$ .

Um diferencial deste método é o uso de probabilidades previamente conhecidas, tais como a forma esperada das distribuições redshift e as frações de tipo galáxias, que podem ser facilmente obtidos a partir de estudos existentes na literatura. Se esta informação anterior é inexistente ou insuficiente, por exemplo, por causa da profundidade sem precedentes de um dado levantamento, as distribuições prévias podem ser calibradas usando a amostra de dados para os quais os *redshifts* fotométricos serão estimados.

### 3.1.2.3 EAZY - *Easy and Accurate Redshifts from Yale*

O programa EAZY (Brammer, van Dokkum & Coppi 2008 [15]) é um método de ajuste de modelos que foi otimizado especificamente para o uso quando amostras com *redshifts* espectroscópicos representativas da amostra fotométrica (no espaço de observáveis) não estão disponíveis para formar um conjunto de treinamento ideal. Dentre os recursos que distinguem-no de outros códigos

destaca-se o uso de um algoritmo de fatoração de matrizes para construir um conjunto mínimo de modelos que são combinações lineares de modelos semi-analíticos. Esses modelos (SEDs) tendem a ser mais representativos das galáxias, particularmente em altos redshifts, em comparação com os comumente usados, como os de Coleman, Wu & Weedman (1980) [22] ou Bruzual & Charlot (1993) [16]. O diferencial deste método é que ele não depende obrigatoriamente da existência de um conjunto de treinamento para fazer a estimativa do *photo-z*.

#### 3.1.2.4 ZEBRA - *Zurich Extra-galactic Bayesian Redshift Analyser*

O programa ZEBRA (Feldmann et al. 2006 [35]) é um estimador de *redshift* fotométrico flexível e poderoso, baseado em ajuste a modelos. O código produz uma distribuição de probabilidades para cada galáxia no espaço de *redshifts* e modelos ( $T$ ),  $P(z, T)$ , assim como distribuições marginalizadas para  $P(z)$  e  $p(T)$ . Os modelos são selecionados a partir de um conjunto prévio de SEDs, buscando os mais frequentes no ajuste do conjunto de treinamento, tomando  $z = z_{spec}$ .

O conjunto prévio de modelos de SEDs modelo foi produzido por um módulo de correção, através de interpolações dos modelos de Coleman, Wu & Weedman (1980) [22] e Kinney et al. (1996) [46]. O módulo permite ao usuário definir os intervalos de *redshift* dentro do qual os modelos são modificados para melhor atender a fotometria de entrada. Após selecionado o conjunto de SEDs apropriado, o modelo é atribuído pelo método estatístico da máxima verossimilhança e a informação a priori obtida do conjunto de treinamento é utilizada com base no método bayesiano para corrigir as PDFs de cada galáxia.

#### 3.1.2.5 PHOTOFIT

PHOTOFIT (Bender et al. 2001 [6]) é um algoritmo bayesiano de ajuste a modelos que fornece a probabilidade de uma galáxia estar em um dado *redshift*. A estimativa é dada através da multiplicação das probabilidades do melhor ajuste a um modelo (utilizando o método do  $\chi^2$ ) por probabilidades previamente conhecidas que relacionam o *redshift* com o fluxo ou magnitudes. A probabilidade total de um modelo é expressa por:

$$P(\vec{\mu}|m) \propto \mathcal{L}(m|\vec{\mu}) \cdot P(\vec{\mu}), \quad (3.2)$$

onde  $m$  corresponde a magnitudes ou fluxos observados e  $\vec{\mu}$  são os os parâmetros do modelo, como *redshift*  $z$  e magnitude absoluta.

O código possui um recurso que permite otimizar o seu desempenho através da calibração de pontos zeros. Esta calibração é feita de forma iterativa, a partir do deslocamento das magnitudes previstas pelos modelos após a comparação com os dados, levando em conta o *spec-z* da amostra de treinamento.

## 3.2 O Portal Científico do DES (PCDES)

Os grandes levantamentos da atualidade geram uma quantidade enorme de dados que exige uma infraestrutura computacional capaz de armazenar, distribuir e propiciar análises de uma forma eficiente. No caso do DES, os grupos de trabalho científicos se beneficiam da infraestrutura associada ao PCDES para o exame de qualidade dos dados obtidos, geração de catálogos e execução de análises, tudo de uma forma integrada, permitindo o compartilhamento de resultados entre os membros da colaboração, mantendo a proveniência e capacidade de reproduzir estes resultados.

### 3.2.1 Visão Geral do PCDES

O PCDES é acessado via navegador de internet é desenvolvido e mantido pelo LIneA<sup>1</sup>, oferecendo uma série de facilidades, entre elas: (a) a manutenção de um repositório de códigos centralizado que usa o sistema *git* para controle de versões; (b) a manutenção de bibliotecas científicas e pacotes de software; (c) a instalação e verificação automatizada da qualidade de dados obtidos pela DECam; (d) ferramentas para inspeção visual de imagens e catálogos associados (o *Tile Viewer*); (e) um banco de dados contendo catálogos astronômicos incluindo os de outros levantamentos; (f) um *cluster* de computadores e uma unidade de armazenamento de grande capacidade dedicados ao processamento via o Portal; (g) um sistema de ponta a ponta, o chamado “*end-to-end*” (E2E) para produção de catálogos com a possibilidade de combinar dados de outros levantamentos; (h) a integração de algoritmos científicos na forma de *workflows* (encadeamento de tarefas modularizadas de componentes e *pipelines*); (i) interfaces para acesso aos dados, metadados e resultados (da Costa et al., em preparação).

A estrutura do portal garante o total controle de proveniência e a reprodutibilidade dos resultados, duas demandas essenciais em projetos científicos de grande porte e com um grande número de participantes como o DES. A Figura 3.4 mostra a tela inicial do portal. Na parte superior, encontra-se um menu com as subdivisões em que o portal é organizado. Duas delas merecem destaque: *Data Server* e *Science Pipelines*.

O *Data Server* é um conjunto de ferramentas computacionais para facilitar o exame do status do levantamento, examinar imagens e catálogos, além de outros produtos científicos que podem ser disponibilizados por participantes da colaboração DES. Este conjunto é formado pelos aplicativos:

- **Observation History:** resumo sobre as observações do levantamento DES e diagnósticos objetivos da qualidade das imagens;
- **Data Releases:** lista de dados liberados e dados associados;
- **Footprint:** Mapa do céu em coordenadas equatoriais mostrando a cobertura atual do levantamento DES e as sobreposições com catálogos externos disponíveis, com *links* para os dados e as imagens (ver Figura 3.5);
- **Tile Viewer:** ferramenta para inspeção visual de imagens e catálogos.

---

<sup>1</sup>Laboratório Interinstitucional de e-Astronomia ([www.linea.gov.br](http://www.linea.gov.br))

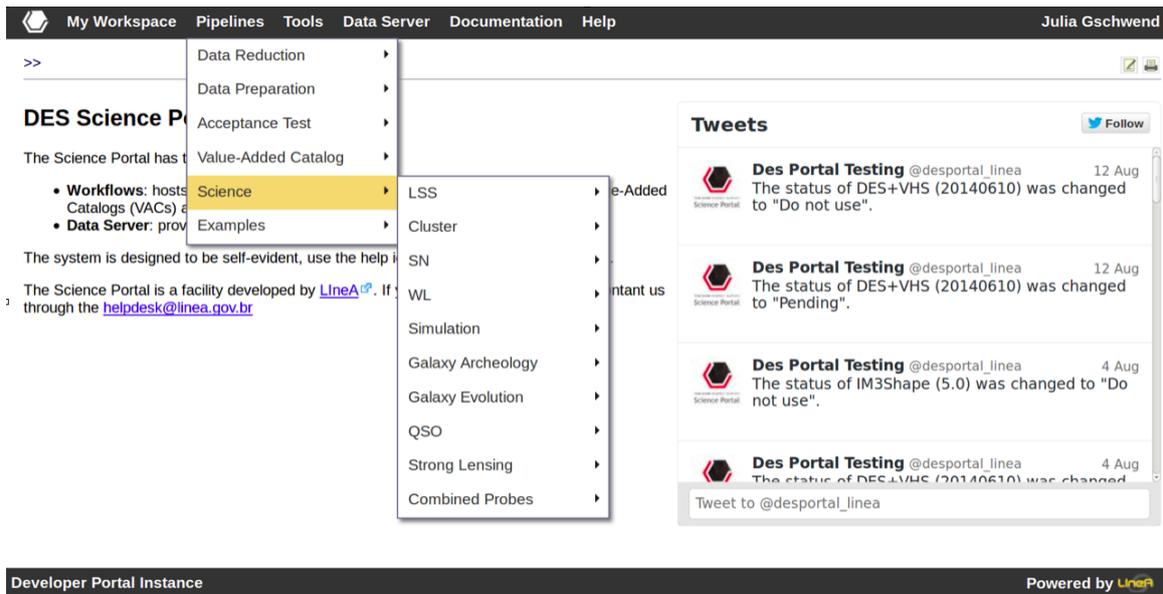


Figura 3.4: Captura de tela da interface gráfica do portal científico. No exemplo acima: menu de *workflows* científicos.

- **Catalog Server:** acesso aos catálogos com valores agregados, os *Value Added Catalogs* (VACs) – que são catálogos produzidos pelo portal, catálogos externos de referência, e simulações.
- **Science Products:** acesso aos resultados produzidos pelos *workflows* científicos ou carregados por outros autores.

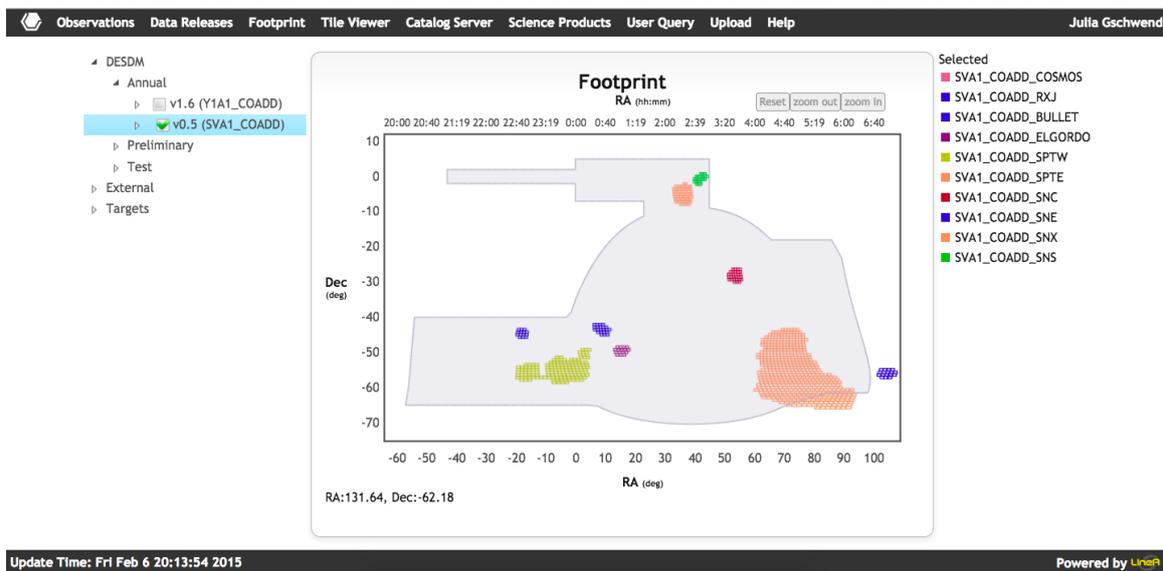


Figura 3.5: Área selecionada.

Os *Science Pipelines* consiste de um conjunto de *pipelines* científicos que recebem os VACs como entrada e geram resultados para as mais variadas linhas de pesquisa, como mostrado no menu da Figura 3.4.

### 3.2.2 O E2E para a criação de VACs

Uma das incumbências dos pesquisadores do consórcio DES-Brazil é a de prover catálogos com os dados processados, prontos para a análises científicas - os VACs.

Os VACs são gerados ao final da cadeia de produção E2E, ao fim de um processo completo, que conecta todas as etapas, desde a instalação dos dados observacionais, até a consolidação em um catálogo único, passando pela preparação e limpeza dos dados, determinação de parâmetros adicionais, como *photo-z*, propriedades espectrais, etc. Haverá um ou mais VACs para cada *pipeline* científico.

A Figura 3.6 mostra, de forma resumida, todas as etapas do E2E. Uma breve descrição será feita neste capítulo (mais detalhes serão encontrados no Apêndice A).

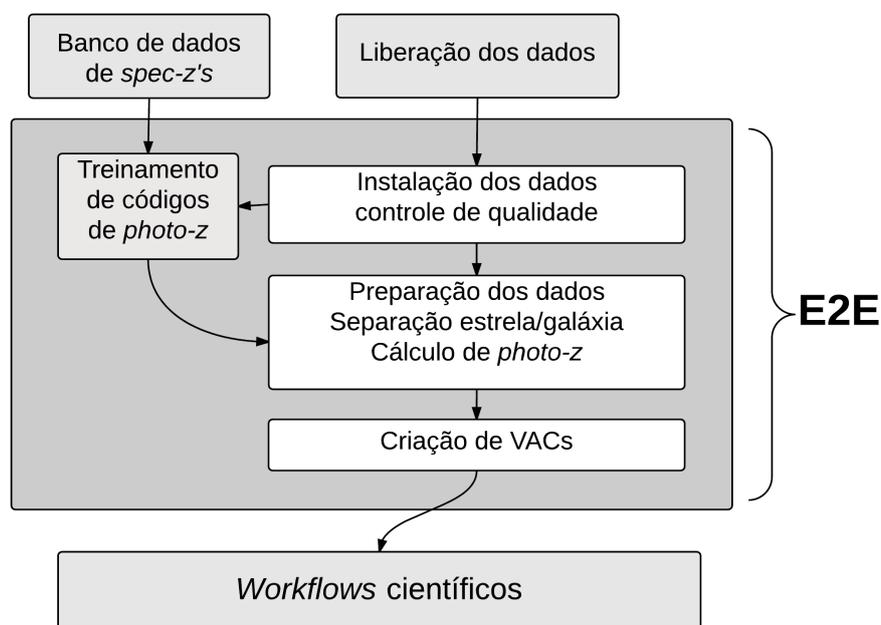


Figura 3.6: Esquema geral do E2E para a criação de VACs.

#### 3.2.2.1 O estágio *Data Installation*

No primeiro estágio do processo de criação de VACs são realizadas, de forma automatizada, as seguintes tarefas:

- A detecção automática de novas liberações de dados do DESDM<sup>1</sup>
- A transferência e ingestão dos produtos do DESDM para o banco de dados do LIneA.
- A identificação de objetos presentes em outros levantamentos (por exemplo: SDSS, 2MASS, APASS, ...) através da correspondência com a posição dos objetos nos catálogos.

<sup>1</sup>DES Data Management, a cadeia de produção de catálogos e imagens do DES.

- A criação de produtos adicionais tais como: correções de ponto zero de magnitudes (baseadas na técnica de regressão ao *locus* estelar), mapas de efeitos sistemáticos e conjuntos de treinamento para o cálculo *redshift* fotométrico (em fase final de implementação)
- Os testes de controle de qualidade de catálogos, baseados nos requisitos científicos estabelecidos nas diretrizes do levantamento.

### 3.2.2.2 O estágio *Data Preparation*

O segundo estágio do E2E é composto de quatro etapas:

- **Limpeza dos catálogos:** A limpeza dos catálogos é opcional e consiste em eliminar objetos do catálogo original, baseado em critérios objetivos que meçam a qualidade da observação. Este critério pode ser por exemplo: um corte em razão sinal-ruído ou em magnitude limite (em uma ou mais bandas), em parâmetros de qualidade provenientes do SExtractor<sup>2</sup>.
- **Separação Estrela/Galáxia:** A classificação dos objetos para futura separação de estrelas e galáxias pode ser feita através de diferentes métodos. Como exemplos, temos: a utilização de parâmetros de forma, que essencialmente vão medir se o objeto é uma fonte pontual ou extensa e métodos empíricos que levam em conta outros parâmetros além dos de forma, como brilho superficial, cor, etc.
- **Photo-z:** Nesta etapa, o *redshift* fotométrico é calculado para todos os objetos classificados como galáxias, utilizando diversos algoritmos. Todos os resultados são encaminhados para o estágio seguinte, o de criação de VACs. Os VACs podem conter um ou mais resultados de *photo-z*.
- **Propriedades espectrais:** A determinação de propriedades espectrais como tipo espectral, correção K, extinção e magnitude absoluta, é feita através do algoritmo de ajuste de modelos LePhare. Neste caso, há duas possibilidades: o algoritmo pode considerar o seu próprio resultado de *photo-z* para atribuir um modelo de SED e a partir dele derivar as demais propriedades, ou então, pode se apropriar do *photo-z* calculado por outro código, usar esse valor como dado de entrada. Todas as propriedades obtidas nesta etapa são entregues ao *pipeline* de criação de VACs.

Mais detalhes sobre o estágio *Data Preparation* são encontrados no Apêndice A.

### 3.2.2.3 O estágio *VAC Creation*

Uma vez que todos os ingredientes já foram produzidos nas etapas anteriores, com total controle de proveniência, versões e configurações, o *pipeline* responsável pela criação dos VACs simplesmente recolhe, dentre os produtos gerados, aqueles que são pertinentes a cada VAC. Por exemplo, o “VAC GE”, voltado para o estudo de evolução de galáxias, vai ter colunas diferentes do “VAC GA”, que por sua vez alimenta os *pipelines* de estudo da Via Láctea.

---

<sup>2</sup>Software que identifica e classifica objetos nas imagens digitais obtidas pela DECAM, mede o fluxo e parâmetros de forma, e determina posições em coordenadas celestes ([www.astromatic.net/software/sextractor](http://www.astromatic.net/software/sextractor)) (Bertin & Arnouts 1996 [11])

### 3.2.3 Os *Workflows* Científicos

*Workflows* científicos são o final de todo o encadeamento de ações dentro do portal científico. São eles que realmente tornam possível uma interpretação dos dados do levantamento com resultados científicos. Devido as grandes proporções do levantamento, não é viável aos cientistas do DES a manipulação de dados e a geração de resultados em seus computadores pessoais.

Além de toda a infraestrutura montada para atender à comunidade DES com catálogos prontos para a análise, o portal também oferece as ferramentas para esta análise, onde os próprios cientistas são criadores e desenvolvedores.

Os *Workflows* científicos atualmente em funcionamento no portal são: *i)* os para estudos de Estrutura em grande escala; *ii)* os para Aglomerados de galáxias; *iii)* os para Evolução de galáxias; *iv)* os para Lentes Gravitacionais; e *v)* Arqueologia da Via Láctea. Cada um deles contem um apanhado de programas e ferramentas voltadas para as necessidades específicas de cada área da astronomia relacionada. Todos estão em constante aprimoramento, acompanhando as demandas de cada grupo de cientistas, possibilitando o total controle de versões, histórico e documentação.

Um exemplo de *pipeline* que compõe o *workflow* científico voltado para o estudo da evolução de galáxias é o *Luminosity Function*. A partir de um VAC GE, ou seja, um catálogo de galáxias contendo os dados fotométricos e resultados anteriores (como *photo-z*, propriedades espectrais, etc), o *pipeline* calcula a função de luminosidade em fatias de redshift e ajusta a função de Schechter ao resultado obtido, encontrando os parâmetros  $\phi^*$  e  $M^*$  do ajuste. O *pipeline Luminosity Function* está em fase de implementação no portal, tendo seus componentes sido exaustivamente testados e utilizados separadamente para a produção de resultados mostrados neste trabalho. A Figura 3.7 mostra a tela inicial de escolha de parâmetros de configuração, como um exemplo da interface visível para o usuário.

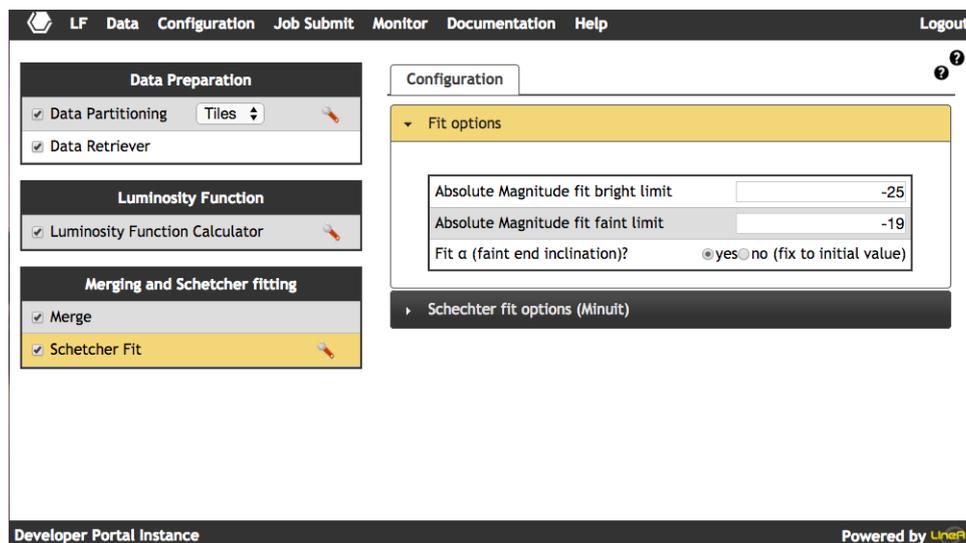


Figura 3.7: Tela de início do *pipeline Luminosity Function*. Em cada seção o usuário do portal pode escolher os parâmetros de configuração desejados.

### 3.3 O Pipeline Luminosity Function

O *pipeline Luminosity Function* (ainda em fase de implementação no PCDES) concatena uma série de módulos (ver figura 3.8), que foram utilizados na determinação da LF no presente trabalho.

O primeiro módulo realiza a tarefa de preparar os dados, deixando-os no formato esperado para os dados de entrada do módulo principal. O módulo principal, detalhado na figura 3.9, calcula as funções de luminosidade global e por tipos, utilizando o método  $1/V_{max}$ . A separação por tipos é baseada nos modelos ajustados pelo LePhare, sendo divididos em duas categorias: “*early-types*” e “*late-types*”.

Apesar do significado morfológico que estes nomes carregam, consideramos esta uma classificação espectral, pois ela é derivada da escolha das SEDs que melhor se ajustam aos dados fotométricos. O módulo produz como resultado uma tabela com os valores de  $\phi$ , com seus respectivos erros e as contagens de galáxias em cada intervalo de magnitude absoluta e em cada intervalo de *redshift*.

O módulo seguinte faz o ajuste dos valores de  $\phi$  obtidos à função de Schechter, através do método dos mínimos quadrados, fornecendo os parâmetros de Schechter:

$$\phi(M)dM = (-0.4 \ln 10)\phi^* 10^{0.4(\alpha+1)(M^*-M)} \exp -10^{0.4(M^*-M)} dM$$

,  
com seus respectivos erros, para cada intervalo de *redshift*, além de produzir gráficos para a interpretação dos resultados.

## Pipeline Função de Luminosidade

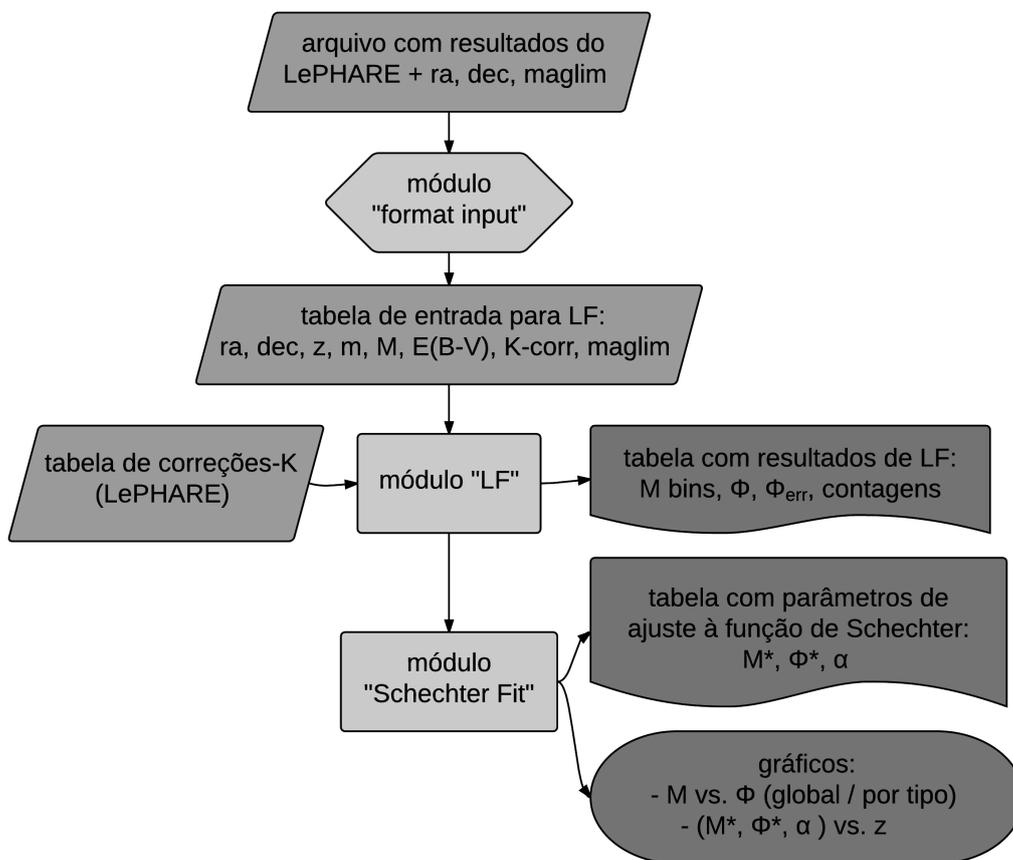


Figura 3.8: Fluxograma do *pipeline Luminosity Function*

# Módulo Função de Luminosidade

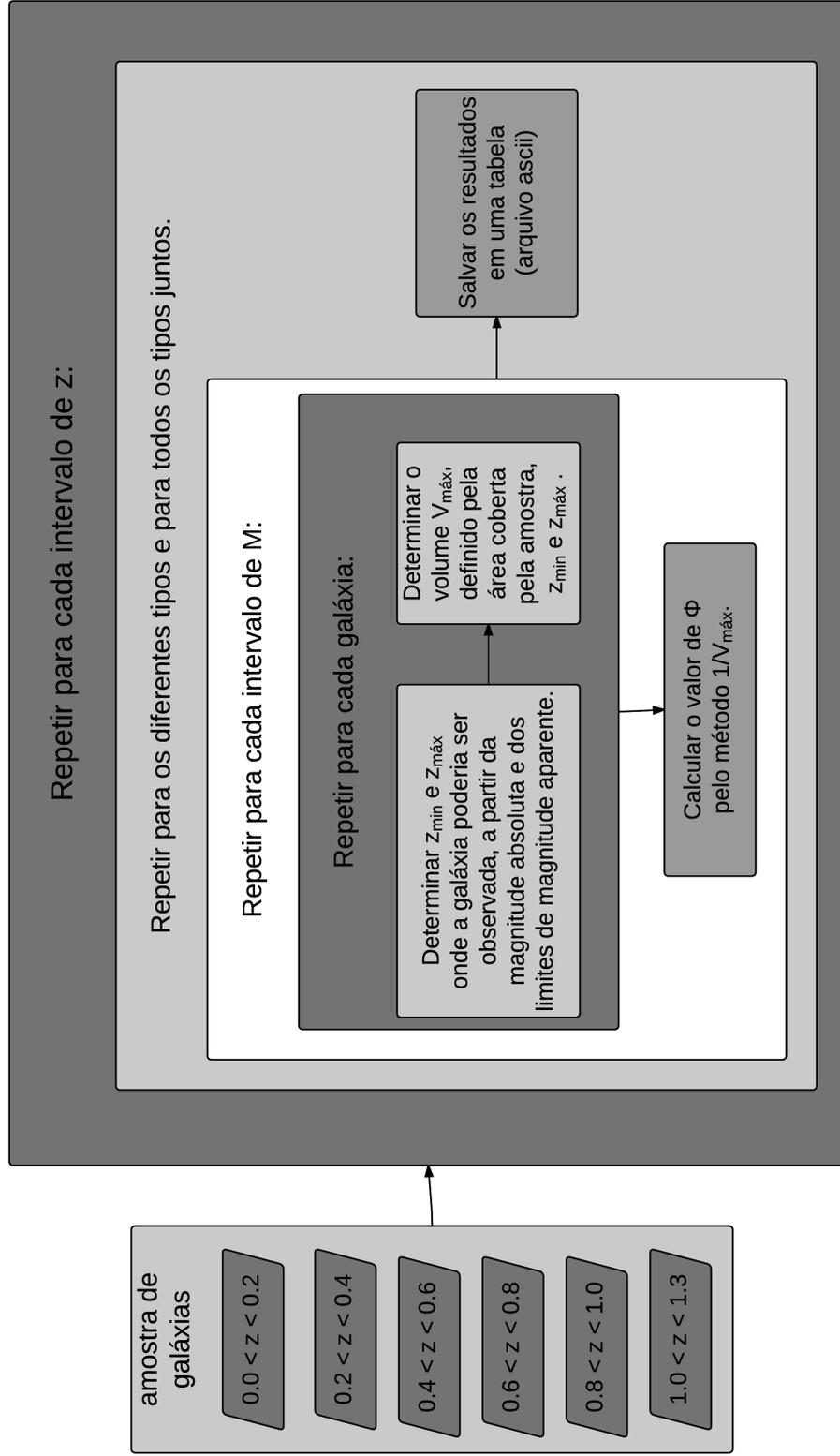


Figura 3.9: Fluxograma do módulo *Luminosity Function*

# Capítulo 4

## Descrição dos Dados

O Dark Energy Survey<sup>1</sup> é um levantamento fotométrico em 5 bandas (*grizY* – detalhes na tabela 4.1), com limites esperados  $g = 25.4$ ,  $r = 24.9$ ,  $i = 24.6$ ,  $z = 23.8$  e  $Y = 21.7$ , desenvolvido para mapear 5000 graus quadrados do céu no hemisfério sul, em 5 anos de observações, iniciadas em agosto de 2013. O objetivo central do levantamento é proporcionar o entendimento da expansão acelerada do universo, caracterizando a natureza da energia escura através de 4 grandes experimentos, cujos observáveis são: lentes gravitacionais fracas, aglomerados de galáxias, oscilação acústica de bárions e supernovas tipo Ia.

A estratégia adotada no levantamento permitirá a determinação de *redshifts* fotométricos de galáxias com uma precisão média de  $\sigma_z \sim 0.07$  para  $z < 1$ , chegando a  $\sigma_z \sim 0.02$  para galáxias em aglomerados, além de medidas de parâmetros de forma para aproximadamente 200 milhões de galáxias (Abbott et al. 2005 [1]).

Tabela 4.1: Características dos filtros utilizados na DECam. ( $\lambda$  em nm)

Filtro	$\lambda$ central	$\lambda$ mínimo	$\lambda$ máximo	FWHM	Máxima eficiência (%)
g	473	398	548	150	91-92
r	642	568	716	148	90-91
i	784	710	857	147	96-97
z	926	850	1002	152	97-98
Y	1009	953	1065	112	98-99

Para atender ao DES, foi construída a DECam, uma câmera de 570 megapixels, composta por 62 CCDs que cobrem 3 graus quadrados no céu a cada exposição. A DECam está instalada no telescópio *Blanco* de 4 metros, situado no CTIO, Chile.

As imagens produzidas pela DECam são reduzidas pelo sistema de gerenciamento de dados *Dark Energy Survey Data Management* (DESDM, Desai et al. 2012 [30]; Mohr et al. 2012 [56]), projeto desenvolvido no *National Center for Supercomputing Applications* (NCSA), na Universidade de Illinois em Urbana-Champaign nos Estados Unidos. O sistema DESDM utiliza os *softwares* SCAMP (*Software for Calibrating AstroMetry and Photometry* - Bertin 2006 [9]) para a leitura das imagens e calibrações astrométricas, SWarp (Bertin et

<sup>1</sup>[www.darkenergysurvey.org](http://www.darkenergysurvey.org)

al. 2002 [12]; Greisen & Calabretta 2000 [40]) para remapeamento e coadição das várias exposições, PSFEx (*PSF Extractor* - Bertin 2011 [10]) para modelagem da função PSF e SExtractor<sup>2</sup> (Bertin & Arnouts 1996 [11]) para detecção de objetos e medidas de fotometria.

Após serem reduzidos no NCSA, os dados seguem para o Fermilab, nos Estados Unidos e para o LIneA, no Brasil, onde passam por um controle de qualidade e em seguida alimentam a cadeia de produção de VACs.

Além dos dados fotométricos obtidos pela DECam, o DES também dispõe de dados espectroscópicos, utilizados para calibração e/ou treinamento de códigos de *photo-z*. Catálogos de *spec-zs* de levantamentos que possuem seus dados públicos, juntamente com medidas de *spec-zs*, realizadas dentro do projeto *Optical Redshifts for Dark Energy Survey* (OzDES, Yuan et al. 2015 [80]), são armazenados em um banco de dados para a criação de novos conjuntos de treinamento a cada liberação de dados do DES.

A área coberta pelo DES foi selecionada visando minimizar efeitos da extinção galáctica e a contaminação por estrelas. A figura 4.1 mostra a região selecionada para a realização do levantamento.

A figura 4.1 mostra a área que será coberta pelo levantamento ao final dos 5 anos e a área que foi coberta na fase de verificação científica. Além disso, mostra a sobreposição com a área do levantamento VHS, cujos dados no infravermelho serão complementares aos dados no óptico obtidos com o DES.

As “regiões de supernovas” foram projetadas para obter curvas de luz bem definidas de aproximadamente 3 mil supernovas do tipo Ia até  $z \approx 1,2$ , com erros estatísticos e sistemáticos pequenos o suficiente para cumprir com os objetivos do levantamento (dentre eles a utilização de dados de supernovas para restringir a determinação dos parâmetros da equação de estado da energia escura). A utilização de espectroscopia para determinação de *redshifts* nessas regiões as tornam importantes para o presente trabalho. Tais medidas alimentam o banco de dados de *spec-z's* utilizados na construção de conjuntos de treinamento.

Neste trabalho, serão analisados dois subprodutos do levantamento DES: simulações e dados da fase de verificação científica. Os dados de simulação servem como um cenário de observação ideal, mas que tentem mimetizar os vieses provocados pelo instrumental e pelas condições do céu. A vantagem do catálogo simulado está em se saber os fluxos e distâncias dos objetos, além da ausência de contaminação por estrelas. A aplicação de nossas rotinas de análise aos dados simulados, permitirá avaliar o desempenho e limitações das mesmas.

Os dados observados nos darão apenas um vislumbre sobre os resultados a serem obtidos futuramente pelo levantamento, pois os mesmos são dados de verificação científica, obtidos em uma fase de ajuste fino do instrumental e dos programas de redução dos dados.

---

<sup>2</sup>[www.astromatic.net/software/sextractor](http://www.astromatic.net/software/sextractor)

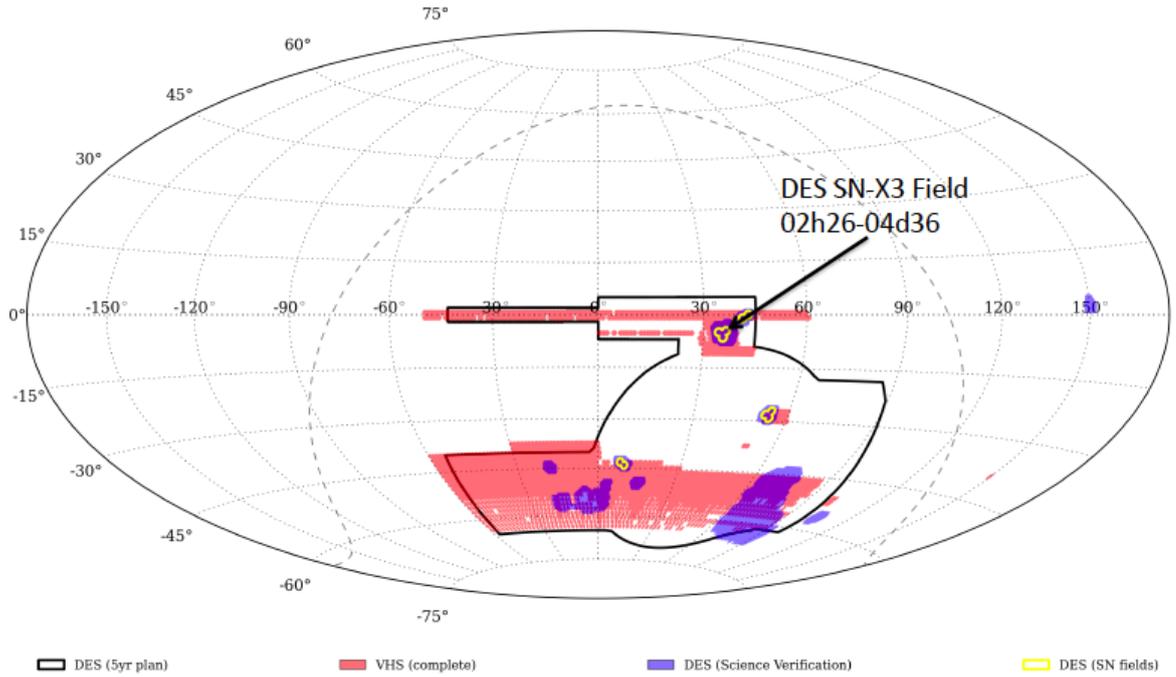


Figura 4.1: Área coberta pelo levantamento DES. O mapa em coordenadas equatoriais mostra o contorno em preto que representa a área planejada para ser coberta ao final dos 5 anos de observação. Em salmão é destacada a região coberta pelo levantamento VHS. Em roxo estão as regiões cobertas pelas observações na fase de verificação científica e em amarelo, as "regiões de supernovas". Figura retirada do artigo Banerji et al. (2015) [4]

## 4.1 Dados de Observação - SVA1

Em todos os conjuntos definidos abaixo, foram utilizadas as medidas de magnitude do tipo MAGAUTO, fornecidas pelo *software* SExtractor. As magnitudes MAGAUTO, são obtidas a partir da medida do fluxo (contagens) contido dentro de uma elipse, automaticamente definida pelo *software* para cada objeto, visando gerar uma estimativa da magnitude total da fonte. A descrição detalhada encontra-se em Bertin & Arnouts (1996) [11].

### 4.1.1 Conjunto Fotométrico

Como o levantamento DES ainda está em fase de liberação dos dados do primeiro ano de observações, para este trabalho utilizamos os obtidos na fase de verificação científica. A amostra denominada *Science Verification Annual Release 1* (SVA1) é composta por observações feitas em regiões estratégicas, que buscaram sobreposições com outros levantamentos (espectroscópicos) ou com os "campos de supernovas" (para os quais o DES vem realizando medidas espectroscópicas através do projeto OzDES). Dentro desta amostra, uma área de aproximadamente 100 graus quadrados (figura 4.2) foi selecionada para compor o conjunto fotométrico, contendo  $\sim 10$  milhões de galáxias, com magnitude limite na banda  $i$  23.9, para  $S/N \sim 5$  (figura: 4.4) .

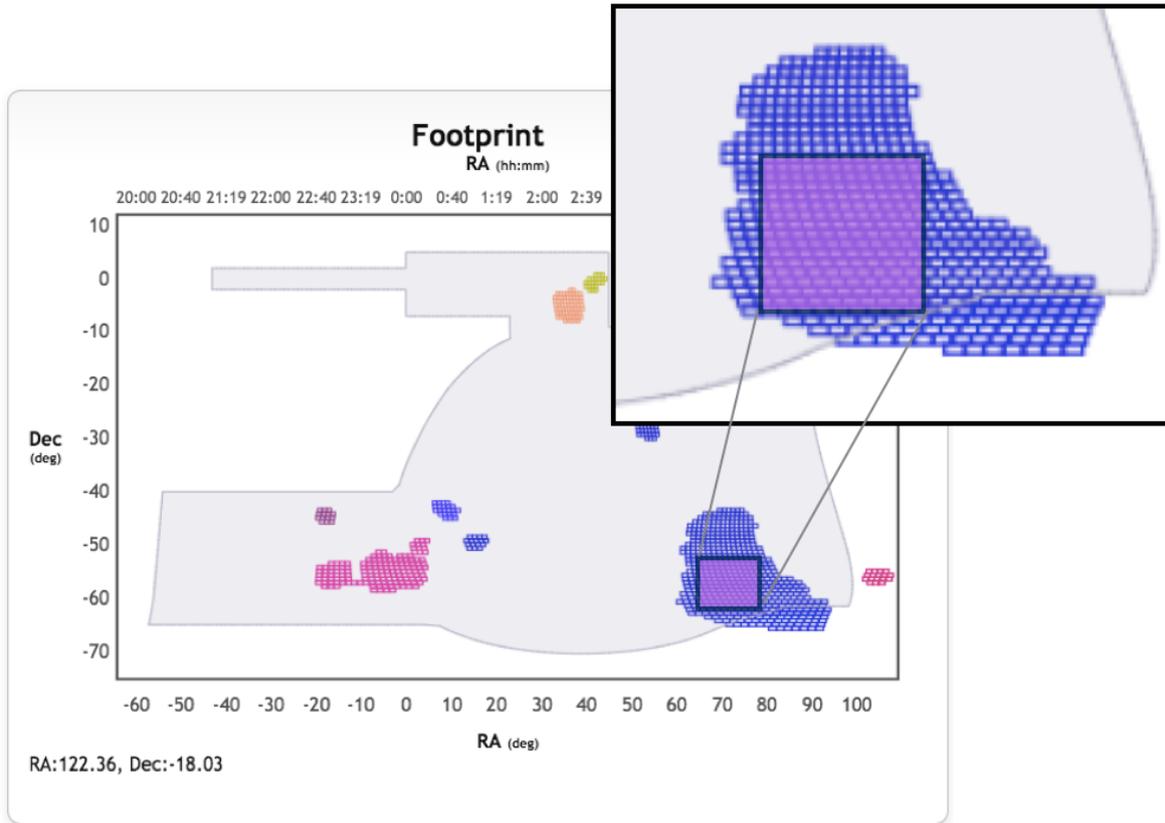


Figura 4.2: Distribuição espacial das regiões observadas na fase de verificação científica do DES. A região ampliada mostra a área selecionada para compor a amostra SVA1.

### 4.1.2 Conjuntos de Treinamento e Validação

O OzDES é um projeto associado ao DES que realiza medidas espectroscópicas com o objetivo de identificar e medir o espectro de candidatas a supernovas tipo Ia. Também são selecionadas galáxias para obtenção de *spec-z* e quasares cujos espectros são usados em mapas de reverberação. Este catálogo é acrescido de *redshifts* provenientes de outros levantamentos.

A partir da versão de novembro de 2014 do catálogo, um conjunto de treinamento foi criado através da correspondência com objetos observados pelo DES. Deste conjunto, foram selecionadas apenas as galáxias cujas “*flags*” de qualidade são consideradas confiáveis (que representam no mínimo 95% de certeza na identificação das linhas nos espectros), totalizando 40.125 galáxias, cobrindo uma faixa de  $0 < z < 1.6$  aproximadamente, e que daqui em diante denominamos “conjunto OzDES”.

A partir do conjunto OzDES e com a utilização do programa *Weights* (Lima et al. 2008 [49]), um segundo conjunto de treinamento foi gerado, com o intuito de simular um conjunto de treinamento ideal, representativo do conjunto fotométrico (no espaço de magnitudes).

O programa gera pesos que atribuem às galáxias do conjunto de treinamento a importância que elas teriam dentro do conjunto fotométrico. Esta importância é quantificada através da quantidade de vizinhos dentro de um certo raio no hiperespaço de magnitudes

que a galáxia do conjunto de treinamento possui no conjunto fotométrico. Tais pesos foram utilizados para adicionar as galáxias ao novo conjunto, um número inteiro de vezes, proporcional ao seu peso, formando o conjunto de treinamento OzDES ponderado, que é representativo do conjunto fotométrico.

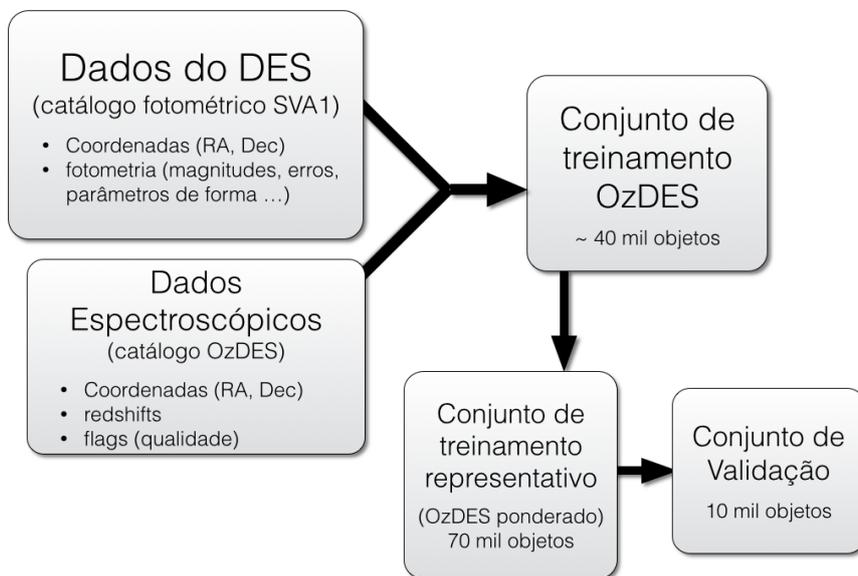


Figura 4.3: Combinação entre dados fotométricos do DES e dados de *redshifts* espectroscópicos do OzDES, junto com dados públicos de outros levantamentos para a construção de conjuntos de treinamento para códigos de *photo-z*.

Com o objetivo de testar a performance do cálculo do *photo-z* e medir o impacto da qualidade da amostra, 10 mil objetos foram aleatoriamente selecionados do conjunto de treinamento representativo, para compor o conjunto de validação. Neste conjunto serão calculadas as métricas descritas no capítulo 2.

A figura 4.3 mostra o esquema da construção do conjunto de treinamento OzDES a partir da correspondência entre as coordenadas das galáxias do catálogo espectroscópico OzDES e do conjunto fotométrico SVA1, bem como a construção dos conjuntos de treinamento ponderado e de validação, a partir do conjunto de treinamento OzDES.

A figura 4.4 mostra as características das amostras mencionadas, na banda  $i$  (banda onde foi determinada a função de luminosidade). O painel (a) mostra as distribuições de magnitudes dos conjuntos de treinamento e do conjunto fotométrico. Em cinza, temos a distribuição de magnitudes do conjunto fotométrico da SVA1. Em azul, vemos que o conjunto de treinamento gerado diretamente a partir da correspondência dos dados fotométricos com as galáxias que possuem *spec-z* disponível no catálogo OzDES apresenta uma distribuição de magnitudes diferente do conjunto fotométrico SVA1. Em vermelho temos a distribuição do conjunto de treinamento representativo, gerado a partir do conjunto OzDES, após a ponderação com a utilização do método de pesos, descrito no capítulo 3. Neste gráfico é possível notar a melhor representatividade das características fotométricas da amostra SVA1 pelo conjunto de treinamento OzDES ponderado, em comparação com o conjunto OzDES

primário.

O painel (b) mostra as distribuições de  $z$  dos dois conjuntos de treinamento, mostrando que a mudança na distribuição de magnitudes do conjunto OzDES, ao se construir o conjunto OzDES ponderado influencia diretamente a distribuição de  $z$ , reforçando o vínculo de dependência entre as duas grandezas. Os painéis (c) e (d) trazem informações sobre os erros nas magnitudes, na banda  $i$  (mais detalhes sobre as amostras e informações sobre as outras bandas estão presentes no apêndice C). No painel (d) vemos um comportamento bimodal na relação entre magnitudes e erros para os conjuntos de treinamento, consequência da predominância de dados de dois levantamentos espectrocópicos de profundidade diferentes na amostra, SDSS e VVDS. Esta assinatura também é percebida na distribuição de  $z$  (painel (b)), com dois picos na distribuição do conjunto OzDES (não ponderado).

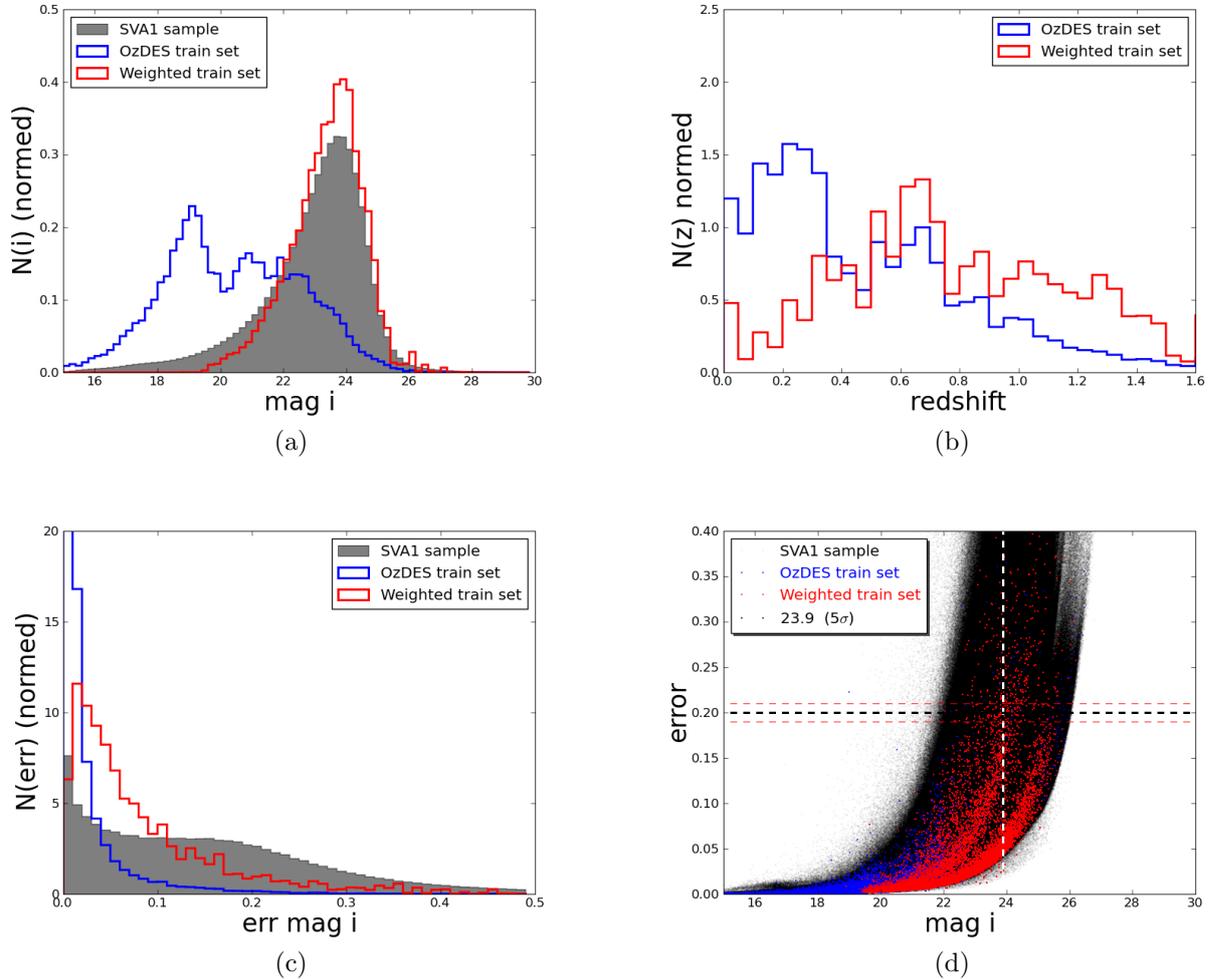


Figura 4.4: 4.4a: Distribuição de magnitudes na banda  $i$ ; 4.4b: distribuição de  $redshift$  dos conjuntos relativos à amostra SVA1; 4.4c: Distribuição de erros nas magnitudes na banda  $i$ ; Para os três histogramas, temos em cinza a amostra SVA1 completa, em azul o conjunto de treinamento OzDES e em vermelho o conjunto de treinamento representativo (OzDES ponderado). 4.4d: Relação entre as magnitudes aparentes na banda  $i$  e seus respectivos erros.

### 4.1.3 Mapas de magnitude limite

Juntamente com os dados observados, foi disponibilizado por colaboradores dentro do levantamento um catálogo adicional contendo a magnitude limite a  $S/N \sim 10$  da região a qual cada galáxia pertence e uma prescrição para a conversão destes valores para a magnitude limite a  $S/N \sim 5$ .

Os limites são dados para cada unidade de área, que corresponde a pequenas regiões de formato hexagonal, denominadas *healpix*. Como as imagens pertencem a uma fase de comissionamento, onde os equipamentos e *softwares* de redução ainda estavam sofrendo ajustes, observamos diferenças significativas na profundidade de diferentes regiões.

A figura 4.5 mostra a distribuição espacial das magnitudes limites dadas pelos mapas mencionados. O histograma à direita mostra que o limite mínimo onde temos a mesma profundidade (em magnitude) alcançada em todas as regiões é de  $\approx 23$  na banda  $i$ . Esta informação será útil para impor um limite na determinação do volume no cálculo da função de luminosidade.

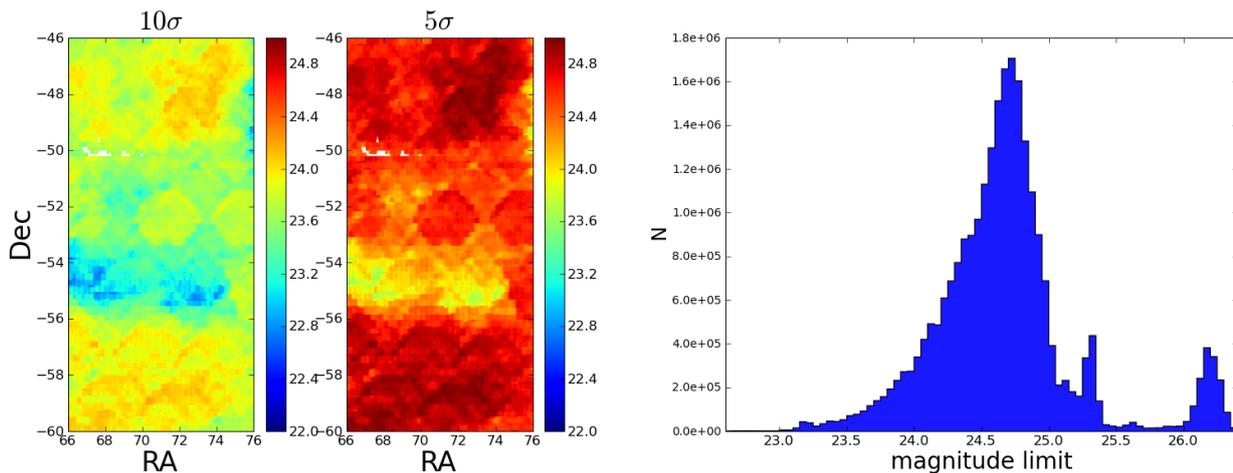


Figura 4.5: Mapas de magnitude limite alcançada para as razões sinal-ruído 10 e 5, respectivamente, e a distribuição de magnitudes limites para  $S/N \sim 5$ .

### 4.1.4 Separação Estrela-Galáxia

Pelo fato da amostra SVA1 ser uma amostra observacional, o conjunto fotométrico possui contaminação por estrelas. Dois métodos de separação das estrelas foram considerados e os descrevemos abaixo.

#### 4.1.4.1 MODEST\_CLASS

O primeiro método analisado é uma classificação baseada em um conjunto de condições mutuamente exclusivas a serem satisfeitas, baseadas em parâmetros de forma e na magnitude da banda  $i$  (todos os parâmetros são obtidos do *software* SExtractor). Neste caso a classificação é binária (painel: 4.7a).

## Galáxias:

$$\begin{aligned} & \text{FLAGS}_I \leq 3 \text{ e não } \{[(\text{CLASS\_STAR}_I > 0.3) \text{ e } (\text{MAG\_AUTO}_I < 18.0)] \\ & \text{ou } [(\text{SPREAD\_MODEL}_I + 3 \cdot \text{SPREADERR\_MODEL}_I) < 0.003] \\ & \text{ou } [(\text{MAG\_PSF}_I > 30.0) \text{ e } (\text{MAG\_AUTO}_I < 21.0)]\} \end{aligned}$$

## Estrelas:

$$\begin{aligned} & (\text{FLAGS}_I \leq 3) \text{ e } \{[(\text{CLASS}_s\text{STAR}_I > 0.3) \\ & \text{e } (\text{MAG\_AUTO}_I < 18.0)] \text{ e } (\text{MAG\_PSF}_I < 30.0) \\ & \text{ou } [(\text{SPREAD\_MODEL}_I + 3 \cdot \text{SPREADERR\_MODEL}_I) < 0.003] \\ & \text{e } ((\text{SPREAD\_MODEL}_I + 3 * \text{SPREADERR\_MODEL}_I) > -0.003)]\} \end{aligned}$$

FLAG\_I é um parâmetro de qualidade da fotometria, neste caso para banda  $i$ , dado pelo *software* SExtractor. A FLAG é uma quantidade acumulativa, correspondente à soma de números inteiros. Cada número representa um problema ocorrido na detecção, como por exemplo contaminação pelo fluxo de estrelas brilhantes próximas. FLAG = 0 significa que o objeto foi detectado em condições ideais.

CLASS\_STAR\_I e SPREAD\_MODEL\_I são parâmetros de forma (SPREADERR\_MODEL\_I é o erro do SPREAD\_MODEL\_I). Estas medidas fornecem uma informação sobre o espalhamento dos fótons em torno do pixel central na imagem detectada no CCD, permitindo classificá-lo como pontual ou extenso. Para um objeto extenso, a classificação como galáxia é direta. Já para objetos pontuais, existe uma probabilidade de que ele seja extragaláctico. Quasares e galáxias muito distantes podem ocupar uma área muito pequena na imagem de detecção. Por isso o MODEST\_CLASS é um parâmetro composto, que também considera as magnitudes. MAG\_AUTO e MAG\_PSF são duas medidas distintas de magnitude.

### 4.1.4.2 TPZ\_SG\_CLASS

O algoritmo TPZ, utilizado no cálculo do *photo-z*, é um método de classificação que pode ser usado para outros fins, além da estimativa do *redshift*. A partir de um conjunto de treinamento com estrelas e galáxias previamente conhecidas, o algoritmo pode ser trei-

nado para classificar objetos como estrelas ou galáxias, dependendo de suas características fotométricas. A partir das magnitudes, cores e de parâmetros de forma, o TPZ atribui uma probabilidade  $P$  de o objeto ser estrela.  $P = 0$  significa certeza na classificação do objeto como galáxia.  $P = 1$  significa certeza na classificação do objeto como estrela. Entretanto, ao invés de um resultado binário, apenas com 0 para as galáxias e 1 para as estrelas, o programa fornece uma distribuição de probabilidades, conforme mostra o painel 4.7b. A partir de uma recomendação do autor do programa, em trabalho colaborativo, neste trabalho consideramos galáxias os objetos cuja probabilidade de ser estrela é inferior a 14%.

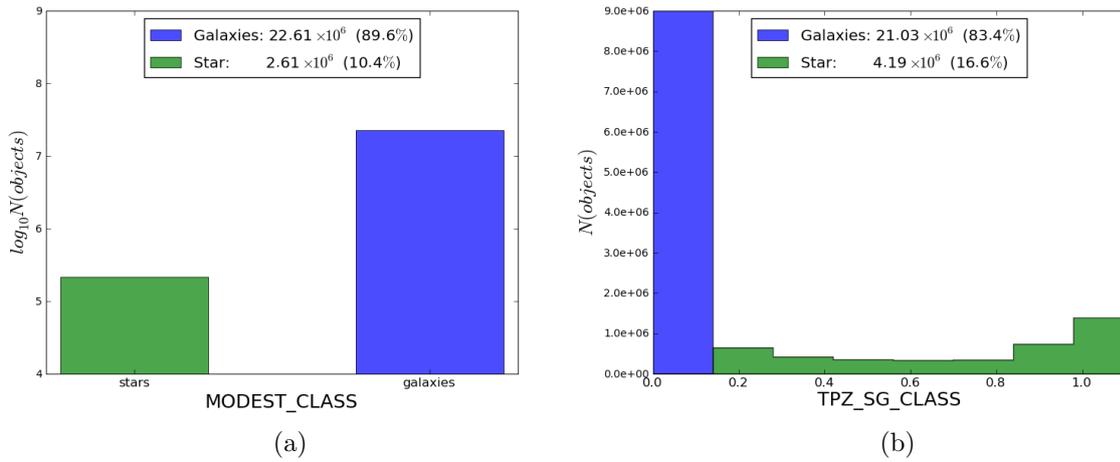


Figura 4.6: Distribuição dos parâmetros `MODEST_CLASS` e `TPZ_SG_CLASS`, utilizados para classificar os objetos quanto a sua natureza (estrela ou galáxia).

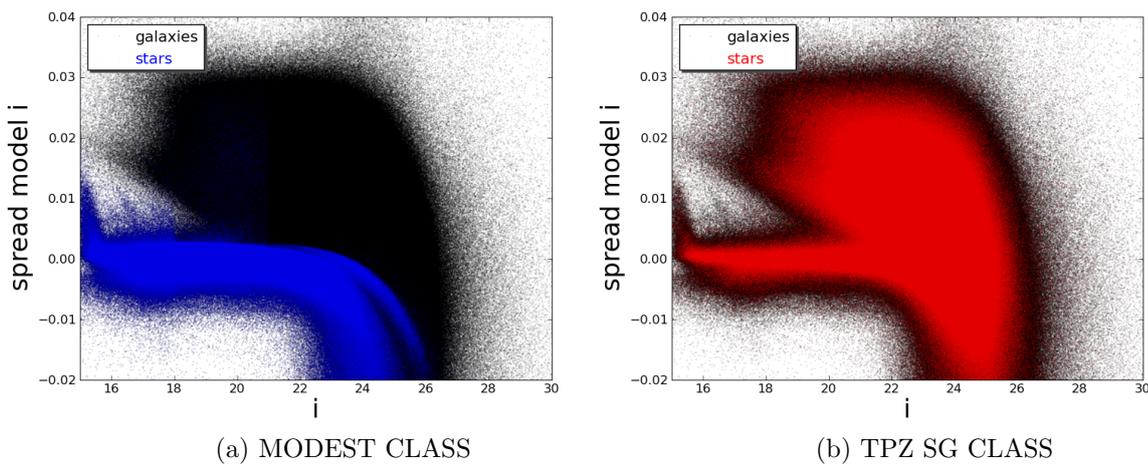


Figura 4.7: `SPREAD_MODEL_I` versus magnitude na banda  $i$ . Em preto os objetos classificados como galáxia e em azul ou vermelho, os objetos classificados como estrelas pelos métodos `MODEST_CLASS` e `TPZ_SG_CLASS`, respectivamente.

Ambos os métodos falham em algum grau, introduzindo estrelas na amostra de galáxias, ou retirando galáxias da amostra por engano. Estes equívocos impactam tanto na pureza, quanto na completeza do conjunto fotométrico. Com uma análise grosseira do gráfico apresentado na figura 4.7, nota-se facilmente que o método *TPZ\_SG\_CLASS* classifica muitos objetos extensos (valores altos de *SPREAD\_MODEL*) como estrelas, causando grande impacto na completeza da amostra de galáxias. Por este motivo, adotaremos o método *MODEST\_CLASS* como separador na amostra SVA1 (Os gráficos da figura 4.4 foram gerados com a amostra contemplando apenas objetos classificados como galáxias pelo classificador *MODEST\_CLASS*).

## 4.2 Dados de Simulação - BCC Aardvark v1.0

*Blind Cosmology Challenge* (BCC) é uma força-tarefa formada por um grupo de cientistas do DES com o objetivo de testar as ferramentas computacionais para, em última instância, extrair resultados cosmológicos. Para isso, foi criado um catálogo de galáxias simuladas, considerando as expectativas do levantamento DES no óptico e do VHS no infravermelho cobrindo assim uma área equivalente a um quarto do céu, até *redshift*  $z \sim 2$ . O catálogo contém galáxias com  $S/N > 5$  em pelo menos uma das bandas do DES, nos limites para as bandas  $grizY = [25.5, 25.0, 24.4, 23.9, 22.0]$  que foram derivados das magnitudes limites para  $S/N > 10$  nas bandas  $grizY = [24.9, 24.5, 23.7, 23.2, 21.5]$  (Busha et al. 2013 [17]). A partir de simulações de n-corpos, galáxias sintéticas foram adicionadas às distribuições de matéria escura através de um algoritmo empírico, utilizando um modelo semi-analítico. Tal algoritmo permite ajustar a concordância com informações observacionais sobre a evolução de populações de galáxias (como contagens e aglomeração espacial, por exemplo) no espaço multidimensional de luminosidades, cores e densidade.

O catálogo, que inclui  $\sim 320$  milhões de galáxias, com cores realistas, formas e erros fotométricos, foi adequado para apoiar estudos de aglomerados de galáxias e lentes gravitacionais, podendo também ser usado para estudos de evolução de galáxias e para testes de algoritmos de *photo-z*.

Para esta dissertação, foi tomado um subconjunto do BCC (daqui em diante chamado de "amostra BCC"), suficientemente grande para minimizar os efeitos da variância cósmica, de  $\sim 100$  graus quadrados de área, contendo  $\sim 13$  milhões de galáxias e magnitude limite de 24.4 na banda-*i*. Desta amostra, foram definidos quatro conjuntos: dois conjuntos para treinamento dos códigos de *photo-z*, um para a validação dos mesmos e um para a determinação da função de luminosidade (conjunto fotométrico, que é a própria amostra BCC completa), conforme mostra o esquema na figura 4.8.



Figura 4.8: Esquema da construção de sub-amostras a partir de amostra de dados simulados do BCC.

### 4.2.1 Conjuntos de Treinamento e Validação

Da amostra BCC, foram selecionados aleatoriamente 40.000 para formar o conjunto de treinamento ideal (ou representativo). Pela natureza aleatória da seleção (dada a grande quantidade de dados), ele preserva as mesmas características do conjunto de origem, tais como as distribuições de magnitudes e de *redshift*.

Para estudar uma situação mais próxima das condições reais, onde a amostra de galáxias que possui dados espectroscópicos disponíveis é, em geral, mais brilhante que a amostra fotométrica (devido à limitações observacionais das técnicas de espectroscopia), um segundo conjunto de treinamento, daqui em diante referido como “OzDES-like”, foi criado, buscando distribuições de cores, magnitudes e *redshift*, semelhantes ao conjunto OzDES.

Para construir esse catálogo foi utilizado um algoritmo que busca na amostra BCC o vizinho mais próximo a cada galáxia do catálogo OzDES no hiperespaço de magnitudes, usando distância euclidiana. Esta galáxia, que possui características fotométricas semelhantes à galáxia do conjunto OzDES é então adicionada ao conjunto OzDES-like.

Uma amostra menor, com outros 10.000 objetos, também aleatoriamente selecionados, foi construída para ser utilizada como amostra de validação. Com este conjunto, o método para o cálculo de *photo-z* foi aplicado, com o objetivo de estudar o seu desempenho.

### 4.2.2 Conjunto Fotométrico

O conjunto fotométrico simula o resultado das observações de um levantamento como o DES. Neste caso particular, tratando-se de uma amostra simulada, temos a informação sobre o *redshift* de antemão. Entretanto, como o objetivo é estudar como seria a LF no caso real, ela será calculada a partir de magnitudes absolutas derivadas do *photo-z*, assim como na amostra observada pelo DES.

O conjunto fotométrico de dados simulados é então definido como a amostra BCC com-

pleta, a partir da qual será determinada a LF.

A figura 4.9 mostra as características das amostras mencionadas, na banda  $i$  (banda onde foi determinada a função de luminosidade). No painel (a) é possível notar a boa representatividade das características fotométricas da amostra BCC, tanto para o conjunto de validação, quanto para o conjunto de treinamento representativo. Cabe também o destaque para a capacidade do conjunto de treinamento não representativo em mimetizar a distribuição de magnitudes do conjunto observado OzDES. Os painéis (c) e (d) trazem informações sobre os erros nas magnitudes, na banda  $i$ . Na comparação com a amostra SVA1, podemos perceber que para os dados simulados os erros são menores e menos espalhados. Acreditamos que os erros do catálogo BCC sejam subestimados, na tentativa de simular erros em observações reais (mais detalhes sobre as amostras e informações sobre as outras bandas estão presentes no apêndice E).

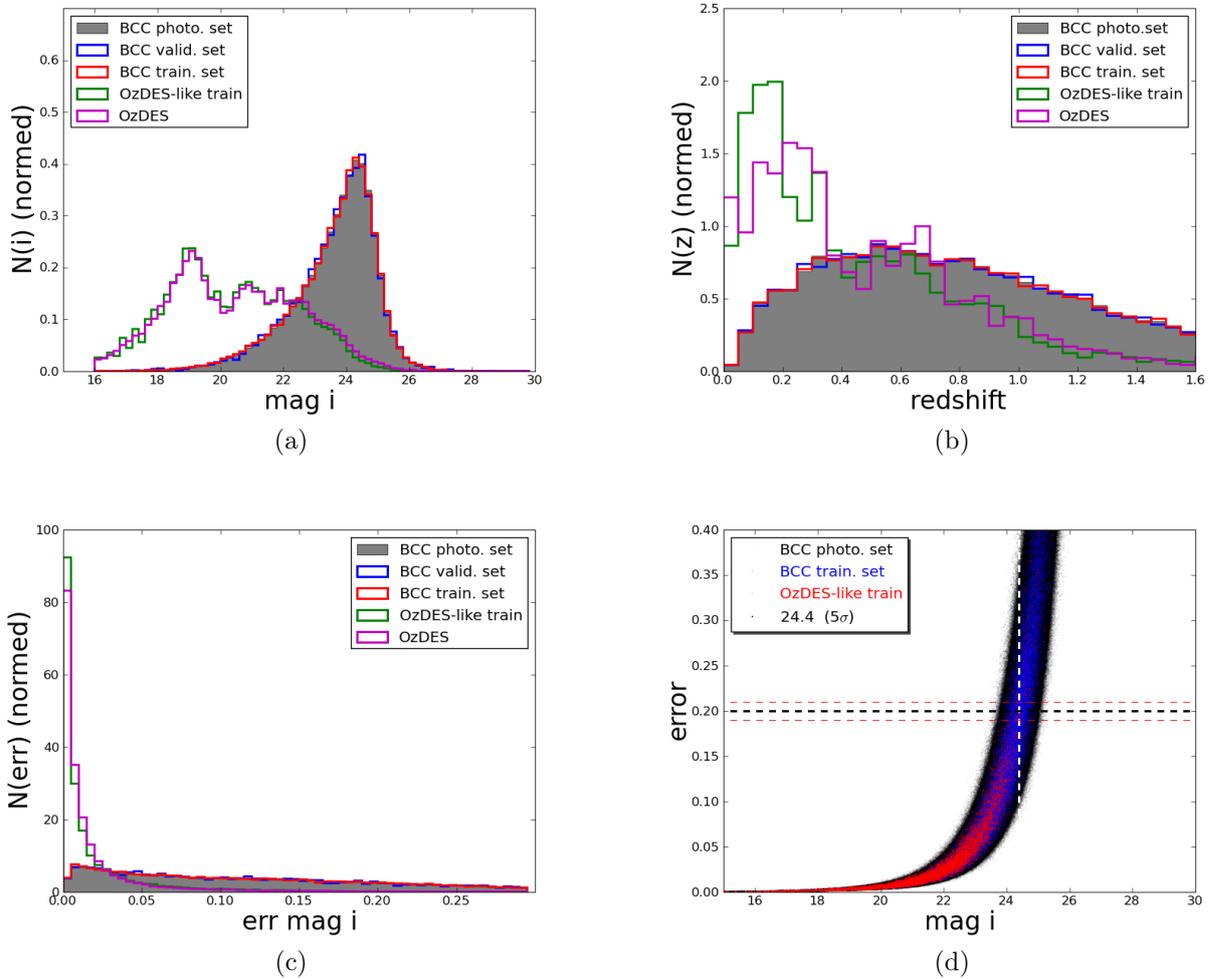


Figura 4.9: painel (a): Distribuição de magnitudes na banda  $i$ ; painel (b): distribuição de  $\text{redshift}$  dos conjuntos relativos à amostra BCC; painel (c): Distribuição de erros nas magnitudes na banda  $i$ ; Para os três histogramas, temos em cinza a amostra BCC completa, em azul o conjunto de validação, em vermelho, o conjunto de treinamento representativo, em magenta, o conjunto de referência de dados observados OzDES e em verde, o conjunto de treinamento não representativo (“OzDES-like”). Painel (d): Relação entre as magnitudes aparentes na banda  $i$  e seus respectivos erros.

# Capítulo 5

## Testes e comparações de algoritmos de *Photo-z*

Neste capítulo abordaremos tanto o objetivo final deste trabalho, que é a determinação da função de luminosidade, quanto os testes que a antecederam, como por exemplo, os que guiaram a escolha dos métodos para estimativas de *photo-z* e a definição das amostras usadas.

### 5.1 Comparação entre os códigos de *Photo-z*

Ao longo dos anos de 2013-14, foi realizado um estudo com o objetivo de testar o desempenho na determinação de *redshifts* fotométricos com os dados do DES, resultando no artigo Sánchez et al. (2014) [68]. Como subproduto deste trabalho foi feita uma avaliação do desempenho entre 20 resultados<sup>1</sup> para 13 algoritmos *photo-z* distintos. Na Tabela 5.1 indicamos os algoritmos usados.

Os dados fotométricos foram obtidos pela DECam na fase inicial de verificação científica (período de observação do fim de 2012 ao início de 2013), cobrindo uma área de  $\sim 200$  graus quadrados nas bandas *grizY*, atingindo profundidades, nos "campos de supernovas", próximas às aquelas esperadas para o levantamento, ao final dos 5 anos de observações.

---

<sup>1</sup>Os resultados apresentados são apenas os melhores obtidos para cada código.

Tabela 5.1: Lista de métodos utilizados para estimar os *photo-z*'s.

Código	Tipo	Referência
DESDM	Empírico	Oyaizu et al. 2008 [59]
ANNz	Empírico	Collister & Lahav 2004 [24]
TPZ	Empírico	Carrasco Kind & Brunner 2013 [19]
RVMz	Empírico	Tipping 2001 [72]
NIP-kNNz	Empírico	de Vicente et al.(em prep.)
ANNz2	Empírico	Sadeh et al. (em prep.)
ArborZ	Empírico	Gerdes et al. 2010 [38]
SkyNet	Empírico	Graff et al. 2008 [39]
BPZ	Ajuste a modelos	Benítez 2000 [7]
EAZY	Ajuste a modelos	Brammer et al. 2008 [15]
LePhare	Ajuste a modelos	Arnouts et al. 1999 [3] e Ilbert et al. 2006 [44]
ZEBRA	Ajuste a modelos	Feldmann et al. 2006 [35]
PHOTOZ	Ajuste a modelos	Bender et al. [6]

Na ocasião, para se fazer o treinamento e determinação de métricas de desempenho para os códigos de *photo-z*, foi utilizado um conjunto de  $\sim 15.000$  galáxias com *redshifts* espectroscópicos disponíveis para definição de duas amostras: a “*main*” e “*deep*”, cuja diferença está na profundidade atingida na fotometria. A figura 5.1 mostra as distribuições *redshift* e de magnitudes nas bandas *g* e *i* para as duas amostras.

Neste trabalho, abordaremos apenas os resultados obtidos com a amostra *main* (identificados como “test 1” no artigo), por semelhança com as amostras aqui utilizadas. Os dados da amostra *main* foram divididos em duas partes, uma para o treinamento e outra para a validação dos códigos.

O conjunto de métricas definido no capítulo 2 foi utilizado para quantificar e comparar a performance dos métodos, tendo sido calculadas para 90% da amostra com os menores erros no *photo-z*, tanto para a amostra completa (métricas globais), quanto para as definidas em intervalos de *redshift*. O método de pesos (Lima et al. 2008 [49]) foi utilizado como fator de ponderação, a fim de melhorar a representatividade do conjunto de treinamento em comparação ao conjunto de dados fotométricos.

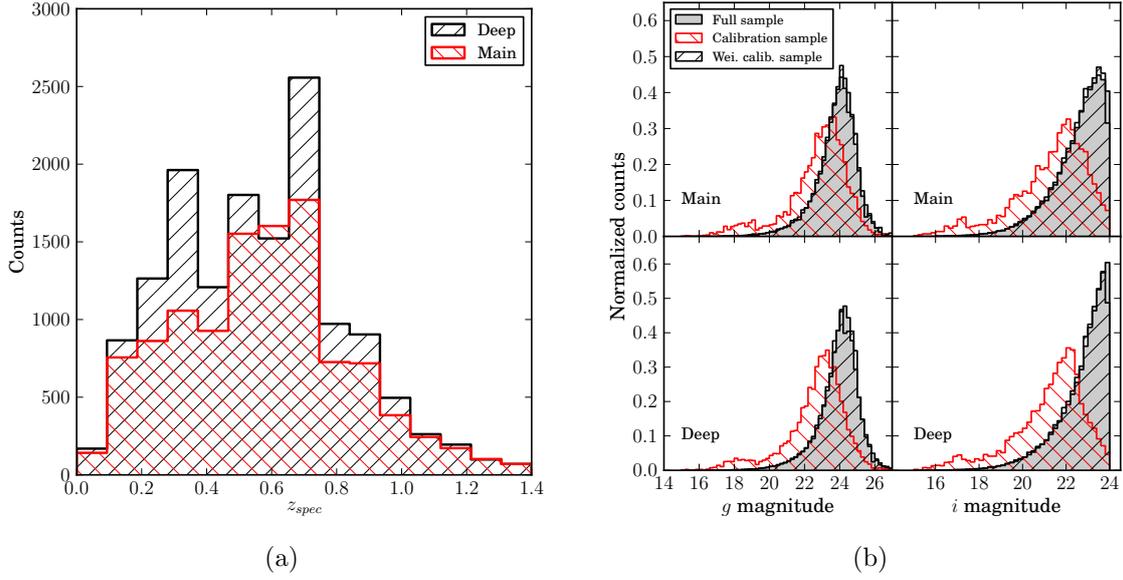


Figura 5.1: 5.1a: Distribuição de *redshifts* espectroscópicas dos conjuntos de treinamento. 5.1b: Distribuição de magnitudes – em cinza os conjuntos fotométricos, em preto os conjuntos de treinamento e em vermelho a distribuição ponderada pelos pesos. (figuras retiradas de Sánchez et al. [68])

Os erros das métricas foram calculados a partir de 100 subamostragens dos resultados, utilizando a técnica *bootstrap*, onde cada subamostra continha 90% dos dados, aleatoriamente escolhidos, com a possibilidade de repetição. O erro em cada métrica foi estimado a partir do desvio padrão do conjunto de 100 resultados. Os resultados são então confrontados com os requisitos científicos do DES que estipulam as condições necessárias para atender, principalmente, as demandas científicas do levantamento.

Tabela 5.2: Métricas relacionadas ao viés e ao espalhamento

Algoritmo	$\overline{\Delta z}$	$\Delta z_{50}$	$\sigma_{68}$	$\sigma_{\Delta z}$	$\overline{\Delta z'}$	$\sigma_{\Delta z'}$
DESDM	$-0.005 \pm 0.003$	$-0.003 \pm 0.002$	$0.094 \pm 0.002$	$0.135 \pm 0.005$	$-0.047 \pm 0.032$	$1.479 \pm 0.052$
ANNz	$0.002 \pm 0.003$	$-0.001 \pm 0.003$	$0.086 \pm 0.002$	$0.118 \pm 0.004$	$0.096 \pm 0.046$	$3.341 \pm 0.134$
TPZ	$-0.001 \pm 0.003$	$0.004 \pm 0.002$	$0.078 \pm 0.002$	$0.122 \pm 0.006$	$0.019 \pm 0.032$	$1.529 \pm 0.063$
RVMz	$-0.011 \pm 0.005$	$-0.004 \pm 0.002$	$0.116 \pm 0.004$	$0.180 \pm 0.008$	$-0.084 \pm 0.041$	$1.371 \pm 0.098$
NIP-kNNz	$-0.030 \pm 0.005$	$-0.011 \pm 0.002$	$0.120 \pm 0.004$	$0.197 \pm 0.009$	$-0.186 \pm 0.030$	$1.116 \pm 0.117$
ANNz2	$-0.002 \pm 0.003$	$-0.003 \pm 0.002$	$0.089 \pm 0.003$	$0.151 \pm 0.009$	$0.063 \pm 0.071$	$2.280 \pm 0.255$
BPZ	$-0.022 \pm 0.003$	$-0.021 \pm 0.002$	$0.097 \pm 0.003$	$0.137 \pm 0.006$	$-0.194 \pm 0.032$	$1.750 \pm 0.075$
EAZY	$-0.061 \pm 0.004$	$-0.063 \pm 0.003$	$0.109 \pm 0.003$	$0.153 \pm 0.010$	$-0.331 \pm 0.074$	$3.982 \pm 0.804$
LePHARE	$0.002 \pm 0.004$	$-0.007 \pm 0.003$	$0.111 \pm 0.003$	$0.171 \pm 0.008$	$1.177 \pm 0.379$	$42.883 \pm 6.603$
PhotoZ	$-0.029 \pm 0.003$	$-0.029 \pm 0.002$	$0.097 \pm 0.003$	$0.142 \pm 0.006$	$-0.268 \pm 0.020$	$1.003 \pm 0.030$
TPZ P(z)	$0.006 \pm 0.003$	$0.011 \pm 0.002$	$0.078 \pm 0.002$	$0.119 \pm 0.006$	$0.125 \pm 0.030$	$1.484 \pm 0.059$
ArborZ P(z)	$-0.008 \pm 0.004$	$0.001 \pm 0.004$	$0.128 \pm 0.003$	$0.153 \pm 0.005$	$-0.028 \pm 0.024$	$0.962 \pm 0.025$
ANNz2 P(z)	$-0.002 \pm 0.002$	$-0.002 \pm 0.002$	$0.085 \pm 0.002$	$0.118 \pm 0.004$	$-0.004 \pm 0.024$	$1.315 \pm 0.042$
SkyNet P(z)	$-0.001 \pm 0.002$	$0.001 \pm 0.002$	$0.077 \pm 0.002$	$0.104 \pm 0.003$	$-0.006 \pm 0.014$	$0.829 \pm 0.027$
BPZ P(z)	$-0.025 \pm 0.003$	$-0.025 \pm 0.003$	$0.101 \pm 0.002$	$0.132 \pm 0.006$	$-0.224 \pm 0.033$	$1.750 \pm 0.080$
ZEBRA P(z)	$-0.050 \pm 0.005$	$-0.030 \pm 0.002$	$0.109 \pm 0.004$	$0.177 \pm 0.012$	$-0.383 \pm 0.047$	$1.906 \pm 0.166$

Tabela 5.3: Frações de *redshifts* discrepantesTabela 5.4: Reprodutibilidade de  $N(z)$ 

Algoritmo	$out_{2\sigma}$	$out_{3\sigma}$	Algoritmo	$N_{poisson}$	$KS$
DESDM	$0.053 \pm 0.005$	$0.018 \pm 0.003$	DESDM	$7.035 \pm 0.486$	$0.056 \pm 0.004$
ANNz	$0.049 \pm 0.004$	$0.015 \pm 0.002$	ANNz	$6.355 \pm 0.480$	$0.052 \pm 0.005$
TPZ	$0.046 \pm 0.004$	$0.019 \pm 0.002$	TPZ	$4.122 \pm 0.505$	$0.044 \pm 0.005$
RVMz	$0.060 \pm 0.005$	$0.023 \pm 0.003$	RVMz	$8.382 \pm 0.548$	$0.083 \pm 0.007$
NIP-kNNz	$0.058 \pm 0.005$	$0.018 \pm 0.003$	NIP-kNNz	$3.633 \pm 0.482$	$0.054 \pm 0.007$
ANNz2	$0.042 \pm 0.003$	$0.021 \pm 0.002$	ANNz2	$5.381 \pm 0.458$	$0.052 \pm 0.004$
BPZ	$0.049 \pm 0.003$	$0.018 \pm 0.002$	BPZ	$7.919 \pm 0.622$	$0.099 \pm 0.008$
EAZY	$0.035 \pm 0.005$	$0.015 \pm 0.002$	EAZY	$11.148 \pm 0.868$	$0.195 \pm 0.009$
LePHARE	$0.047 \pm 0.002$	$0.024 \pm 0.002$	LePHARE	$6.632 \pm 0.444$	$0.087 \pm 0.006$
PhotoZ	$0.058 \pm 0.004$	$0.017 \pm 0.002$	PhotoZ	$4.565 \pm 0.464$	$0.080 \pm 0.006$
TPZ P(z)	$0.049 \pm 0.005$	$0.018 \pm 0.003$	TPZ P(z)	$2.607 \pm 0.449$	$0.044 \pm 0.006$
ArborZ P(z)	$0.056 \pm 0.005$	$0.016 \pm 0.003$	ArborZ P(z)	$4.774 \pm 0.449$	$0.064 \pm 0.009$
ANNz2 P(z)	$0.051 \pm 0.004$	$0.016 \pm 0.003$	ANNz2 P(z)	$5.010 \pm 0.404$	$0.045 \pm 0.004$
SkyNet P(z)	$0.072 \pm 0.006$	$0.015 \pm 0.002$	SkyNet P(z)	$4.091 \pm 0.404$	$0.052 \pm 0.005$
BPZ P(z)	$0.046 \pm 0.004$	$0.014 \pm 0.002$	BPZ P(z)	$5.776 \pm 0.948$	$0.098 \pm 0.007$
ZEBRA P(z)	$0.043 \pm 0.006$	$0.018 \pm 0.002$	ZEBRA P(z)	$6.692 \pm 0.550$	$0.112 \pm 0.008$

As tabelas 5.2 a 5.4 mostram os resultados das métricas para o *photo-z* de todos os códigos comparados. Para gerar estes resultados, foi utilizado um vetor com pesos para ponderação, tornando o conjunto de validação com características fotométricas mais próximas das do conjunto fotométrico ao qual ele deveria representar. Além disso, foram descartados 10% dos objetos com os maiores erros na medida do *photo-z* em cada código.

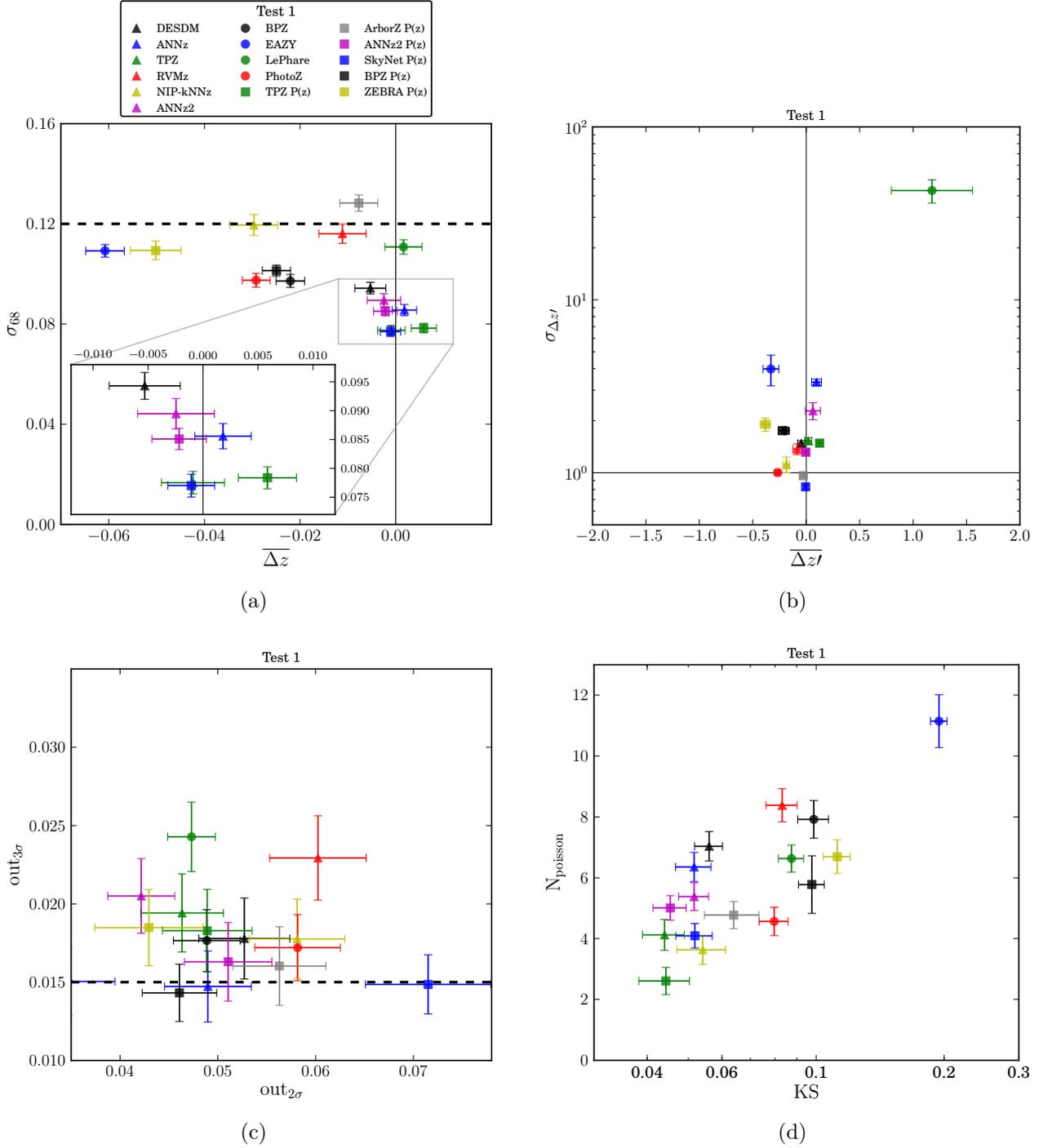
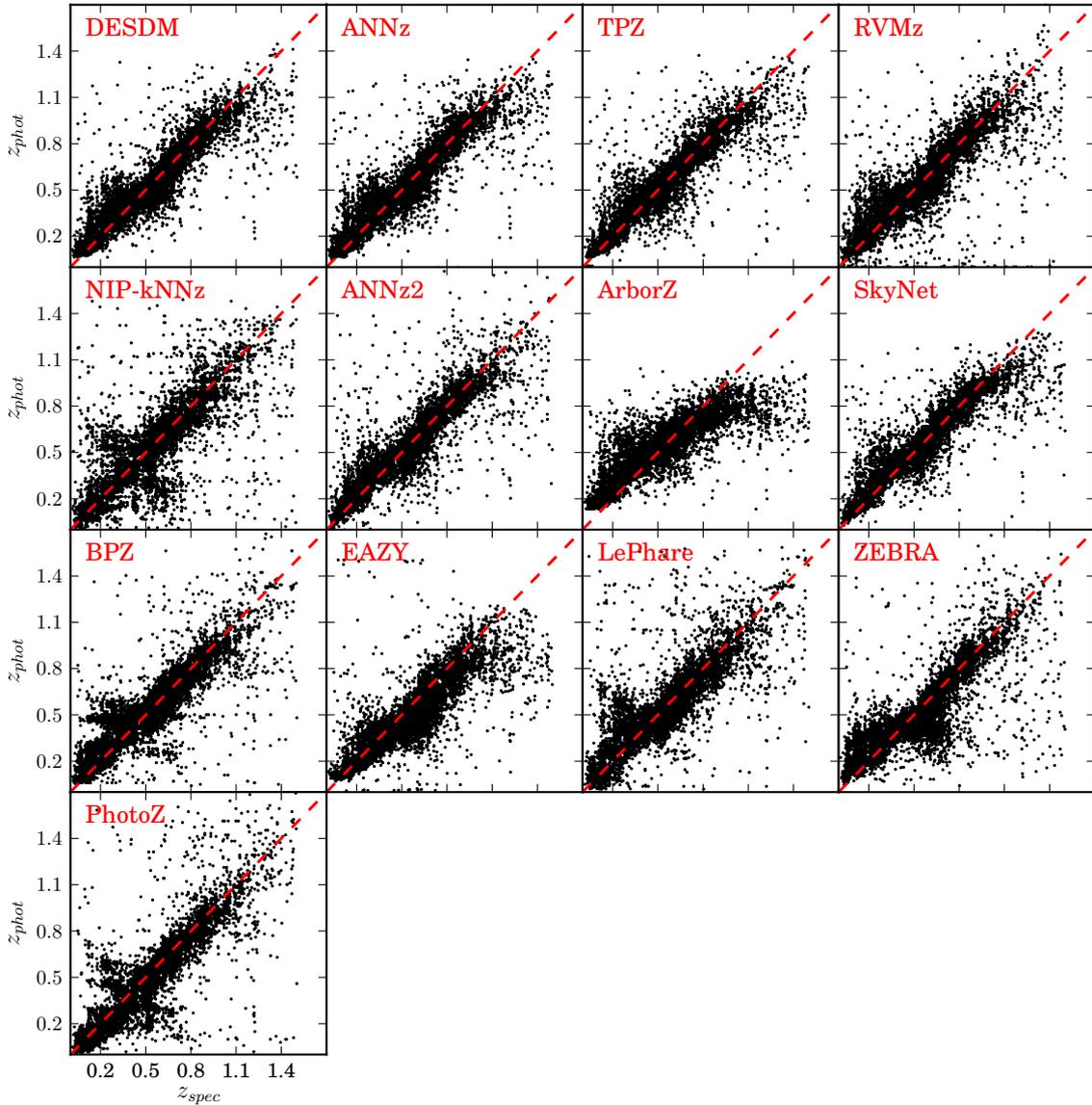


Figura 5.2: Métricas obtidas para os códigos de *photo-z*. As linhas tracejadas representam os requisitos científicos do DES. Triângulos representam os códigos baseados em métodos empíricos, círculos os códigos baseados em ajustes de modelos e quadrados os resultados obtidos a partir da  $P(z)$ . 5.2a:  $\sigma_{68}$  vs.  $\overline{\Delta z}$ . 5.2b:  $\sigma_{\Delta z'}$  vs.  $\overline{\Delta z'}$ . 5.2c:  $out_{3\sigma}$  vs.  $out_{2\sigma}$ . 5.2d:  $N_{poisson}$  vs.  $KS$ . (figura retirada de Sánchez et al. [68])

Figura 5.3:  $z_{phot}$  versus  $z_{spec}$  para todos os algoritmos avaliados. (figura retirada de Sánchez et al. [68])



## Sumário

- O painel 5.2a mostra  $\sigma_{68}$  vs.  $\overline{\Delta z}$ . A grande maioria dos códigos satisfaz o requisito científico DES com facilidade. Os códigos que ocupam a região de alta performance no gráfico (região ampliada) são todos baseados em métodos empíricos (com estimativa de valor único para o *photo-z* ou produzindo  $P(z)$ ).
- O painel 5.2b mostra  $\sigma_{\Delta z'}$  vs.  $\overline{\Delta z'}$ . Grande parte dos códigos apresentam valores altos de  $\overline{\Delta z'}$  (cujo valor esperado seria 1), mostrando que todos eles subestimam seus erros de *photo-z*. Entretanto uma boa quantidade de códigos mostra uma distribuição de erros próxima da gaussiana com média 0 e desvio padrão 1. Não há casos muito problemáticos para estas métricas. Os dois tipos de código povoam tanto regiões de alta quanto de baixa performance.

- O painel 5.2c mostra as frações de *redshifts* discrepantes  $out_{3\sigma}$  vs.  $out_{2\sigma}$ . O requisito científico DES para  $out_{2\sigma} (< 0.1)$  está fora dos limites do gráfico, mostrando que todos os códigos satisfazem-no. Por outro lado, o requisito para  $out_{3\sigma}$  só é satisfeito para uma pequena fração dos códigos (de ambos os tipos de métodos).
- O painel 5.2d mostra  $N_{poisson}$  vs.  $KS$ . Conforme o esperado, as duas métricas são fortemente correlacionadas. O gráfico mostra que uma melhor reconstrução da distribuição  $N(z)$  é obtida com os códigos que fornecem  $P(z)$ , quando comparada à utilização das estimativas de valor único para o *photo-z* dos mesmos códigos.

Em linhas gerais, os códigos baseados em métodos empíricos apresentam melhor desempenho tanto para a qualidade da estimativa de um valor único de *photo-z* quanto para a reconstrução da distribuição  $N(z)$ . Este resultado pode ser um indício de falta de acurácia dos modelos de SEDs utilizados pelos códigos de ajustes. Todavia, quando comparamos as frações de *redshifts* discrepantes, não há evidências de vantagem em um tipo de método sobre o outro.

Estes resultados podem variar de acordo com a faixa de *redshift* estudada. Tradicionalmente, códigos de ajustes a modelos funcionam melhor em regiões de alto *redshift*. Códigos empíricos são muito dependentes da boa representatividade da amostra pelo conjunto de treinamento, que é dificilmente satisfeita nessa faixa de  $z$ .

Baseado na métrica  $\sigma_{68}$ , o melhor resultado na estimativa de um valor único para o *photo-z* é o código empírico TPZ. O resultado deste teste justifica a escolha da metodologia adotada para a determinação da função de luminosidade. O *photo-z* da amostra fotométrica, para a qual será determinada a LF, será calculado pelo código TPZ.

## 5.2 A Importância da Qualidade das Amostras

Após uma avaliação dos resultados dos testes anteriores, as métricas foram recalculadas após dois cortes distintos que eliminam aproximadamente a mesma quantidade de objetos. O primeiro corte consiste em calcular o *photo-z* para todos os objetos e posteriormente eliminar os 10% com maiores erros no *photo-z* ( $\epsilon_{phot}$ ). O segundo consiste em eliminar, a priori, objetos que possuem  $S/N < 10$  em 3 ou mais bandas, o que também corresponde a  $\sim 10\%$  da amostra.

As tabelas 5.5 e 5.6 mostram os resultados das métricas recalculadas. Diferentemente da seção anterior, neste teste vamos apresentar todos os resultados obtidos, inclusive com repetição de código. Cada resultado é identificado por uma sigla que representa o nome da pessoa ou grupo responsável e o nome do código utilizado. Cada tabela está dividida em 3 blocos, organizando os códigos pelo método em que se baseiam.

No primeiro bloco estão os códigos empíricos, no segundo, os de ajuste de modelos, e no

terceiro, os que fornecem a  $P(z)^2$  para cada objeto (podendo ser de qualquer método). Em cada bloco, os códigos estão arbitrariamente em ordem crescente de  $\sigma_{68}$  após corte em S/N.

Tabela 5.5:  $\overline{\Delta z}$  e  $\sigma_{68}$  após diferentes cortes

Métrica	$\overline{\Delta z}$			$\sigma_{68}$		
	sem cortes	$\epsilon_{z_{phot}}$	S/N	sem cortes	$\epsilon_{z_{phot}}$	S/N
DesBrazil ANNz	-0.0016	0.0000	-0.0002	0.0722	0.0820	0.0668
MCK TPZ	-0.0013	0.0010	0.0013	0.0773	0.0850	0.0695
CSanchez ANNz	-0.0022	0.0020	-0.0004	0.0781	0.0850	0.0709
HLin DESDM	-0.0025	-0.0050	-0.0006	0.0789	0.0940	0.0725
ISadeh ANNz2	-0.0024	0.0050	0.0006	0.0793	0.0870	0.0731
JVicente kNNz	-0.0140	-0.0150	-0.0103	0.1099	0.1300	0.0989
NGreisel PhotoZ	-0.0247	-0.0290	-0.0228	0.0845	0.0970	0.0782
SJouvel LePHARE	0.0127	0.0150	0.0150	0.0982	0.1000	0.0940
PMarti BPZ	0.0017	-0.0220	0.0132	0.1023	0.0970	0.0962
DesBrazil LePHARE	0.0113	0.0020	0.0130	0.1052	0.1110	0.0994
MBanerji EAZY	-0.0487	-0.0530	-0.0483	0.1166	0.1300	0.1099
DesBrazil Zebra	-0.0211	-0.0340	-0.0102	0.1463	0.1850	0.1344
DesBrazil HyperZ	-0.0646	-0.0540	-0.0630	0.1543	0.1770	0.1494
MCK TPZ $P(z)$	0.0066	0.0090	0.0094	0.1018	0.0850	0.0694
DesBrazil LePHARE $P(z)$	0.0081	0.0030	0.0090	0.1049	0.1100	0.0985
PMarti BPZ $P(z)$	0.0005	-0.0250	0.0112	0.1097	0.1010	0.1028
Michigan ArborZ $P(z)$	-0.0104	-0.0310	-0.0050	0.1236	0.1470	0.1139
DesBrazil $P(z)$	-0.0582	-0.0070	-0.0347	0.1402	0.1170	0.1232
SJouvel LePHARE $P(z)$	0.2092	-0.0220	0.2119	0.3078	0.1090	0.3072

<sup>2</sup>Os códigos que aparecem em dois blocos são os que fornecem a  $P(z)$  além de uma estimativa de valor único para o *photo-z* de cada objeto.

Tabela 5.6:  $out_{2\sigma}$  e  $out_{3\sigma}$  após diferentes cortes

Métrica Algoritmo	$out_{2\sigma}$			$out_{3\sigma}$		
	sem cortes	$\epsilon_{z_{phot}}$	S/N	sem cortes	$\epsilon_{z_{phot}}$	S/N
DesBrazil ANNz	0.0497	0.0440	0.0459	0.0202	0.0150	0.0180
MCK TPZ	0.0543	0.0520	0.0467	0.0237	0.0310	0.0207
CSanchez ANNz	0.0498	0.0500	0.0479	0.0177	0.0140	0.0155
HLin DESDM	0.0511	0.0530	0.0482	0.0188	0.0180	0.0159
ISadeh ANNz2	0.0497	0.0470	0.0455	0.0212	0.0140	0.0161
JVicente kNNz	0.0559	0.0460	0.0571	0.0230	0.0200	0.0203
NGreisel PhotoZ	0.0564	0.0590	0.0562	0.0228	0.0160	0.0190
SJouvel LePHARE	0.0705	0.0680	0.0697	0.0213	0.0200	0.0223
PMarti BPZ	0.0545	0.0490	0.0531	0.0199	0.0170	0.0177
DesBrazil LePHARE	0.0569	0.0460	0.0531	0.0233	0.0240	0.0231
MBanerji EAZY	0.0524	0.0490	0.0466	0.0204	0.0170	0.0173
DesBrazil Zebra	0.0546	0.0700	0.0509	0.0245	0.0310	0.0209
DesBrazil HyperZ	0.0691	0.0640	0.0728	0.0140	0.0140	0.0120
MCK TPZ P(z)	0.0528	0.0590	0.0466	0.0213	0.0150	0.0207
DesBrazil LePHARE P(z)	0.0494	0.0500	0.0430	0.0237	0.0210	0.0204
PMarti BPZ P(z)	0.0480	0.0460	0.0447	0.0163	0.0140	0.0145
Michigan ArborZ P(z)	0.0600	0.0650	0.0576	0.0183	0.0150	0.0170
DesBrazil P(z)	0.0624	0.0650	0.0535	0.0268	0.0150	0.0258
SJouvel LePHARE P(z)	0.0555	0.0560	0.0565	0.0065	0.0190	0.0077

### Sumário:

- O corte em  $\epsilon_{phot}$  aumenta  $\overline{\Delta z}$  para a maioria dos códigos avaliados (11 de 19).
- O corte em S/N reduz  $\overline{\Delta z}$  para a maioria dos códigos avaliados (12 de 19).
- O corte em  $\epsilon_{phot}$  aumenta  $\sigma_{68}$  para a maioria dos códigos avaliados (14 de 19).
- O corte em S/N diminui  $\sigma_{68}$  em todos os códigos avaliados.
- A fração de objetos discrepantes é reduzida com os dois cortes, para a maioria dos códigos avaliados.
  - Corte em  $\epsilon_{phot}$ : 10 de 19 para  $out_{2\sigma}$  e 15 de 19 para  $out_{3\sigma}$ .
  - Corte em S/N: 16 de 19 para  $out_{2\sigma}$  e 17 de 19 para  $out_{3\sigma}$ .

O corte baseado nos erros, em geral, provoca aumento na dispersão do  $photo-z$  quando comparado ao  $spec-z$  (quantificada pela métrica  $\sigma_{68}$ ). Este corte também aumenta, em

geral, o viés médio introduzido por erros sistemáticos (quantificado por  $\overline{\Delta z}$ ). O requisito científico DES para a qualidade da fotometria prevê a possibilidade de que 90% da amostra atinjam os patamares estabelecidos. Apesar de sabermos que a maioria dos códigos de *photo-z* estimam o erro ( $\epsilon_{phot}$ ) através da propagação dos erros dos observáveis, os resultados dos testes realizados mostraram que remover os objetos com piores  $\epsilon_{phot}$ , a posteriori, não melhora a qualidade do resultado segundo as métricas globais. Uma possível explicação para este resultado seria o fato de que, conforme mostram as figuras 5.4 a 5.6, para a maioria dos códigos, não há correlação entre  $\epsilon_{phot}$  e o  $\overline{\delta z}$ . Além disso, os objetos removidos com este critério variam de um código para outro, tirando a imparcialidade da comparação entre eles.

Por outro lado, o corte em razão sinal-ruído, feito a priori, realmente elimina os dados de má qualidade, como galáxias com incertezas muito grandes em suas magnitudes e flutuações no ruído do céu que acidentalmente foram detectadas como galáxias. Esta limpeza nos dados provoca melhora na qualidade do resultado global do *photo-z*, o que pode ser quantificado pelas métricas  $\overline{\Delta z}$  e  $\sigma_{68}$ .

## Métodos Empíricos

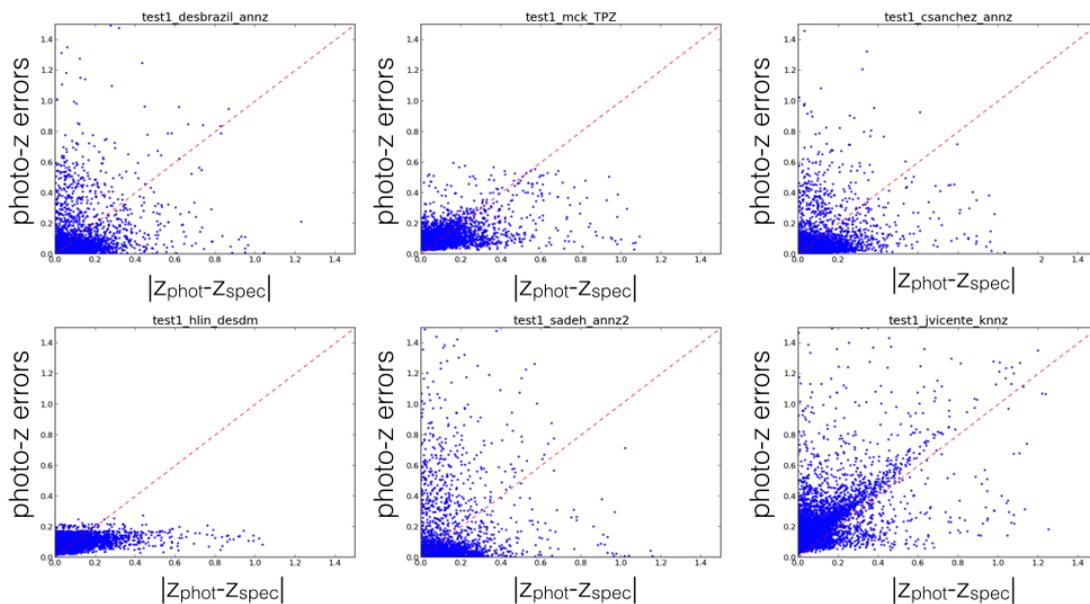


Figura 5.4:  $\epsilon_{phot}$  versus  $\overline{\Delta z}$  para métodos empíricos.

## Métodos de Ajuste a Modelos

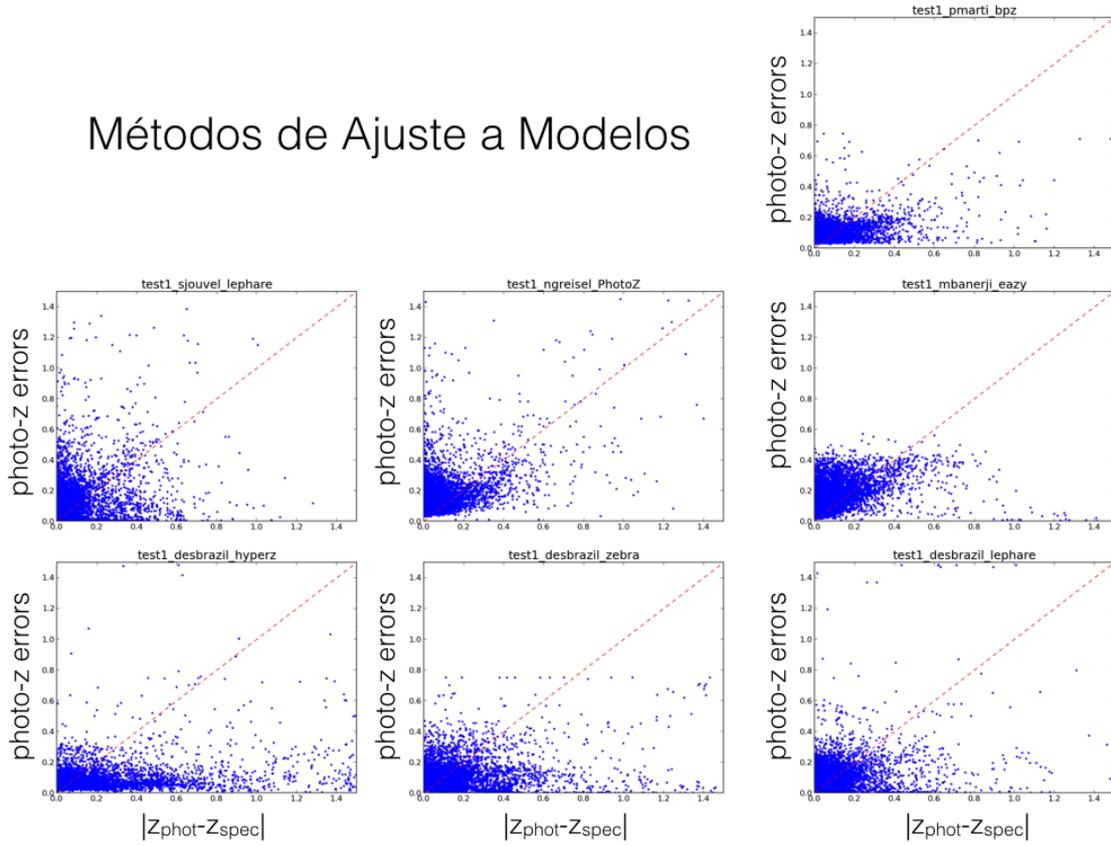


Figura 5.5:  $\epsilon_{phot}$  versus  $\overline{\Delta z}$  para métodos de ajuste a modelos

## Métodos que fornecem PDF(z)

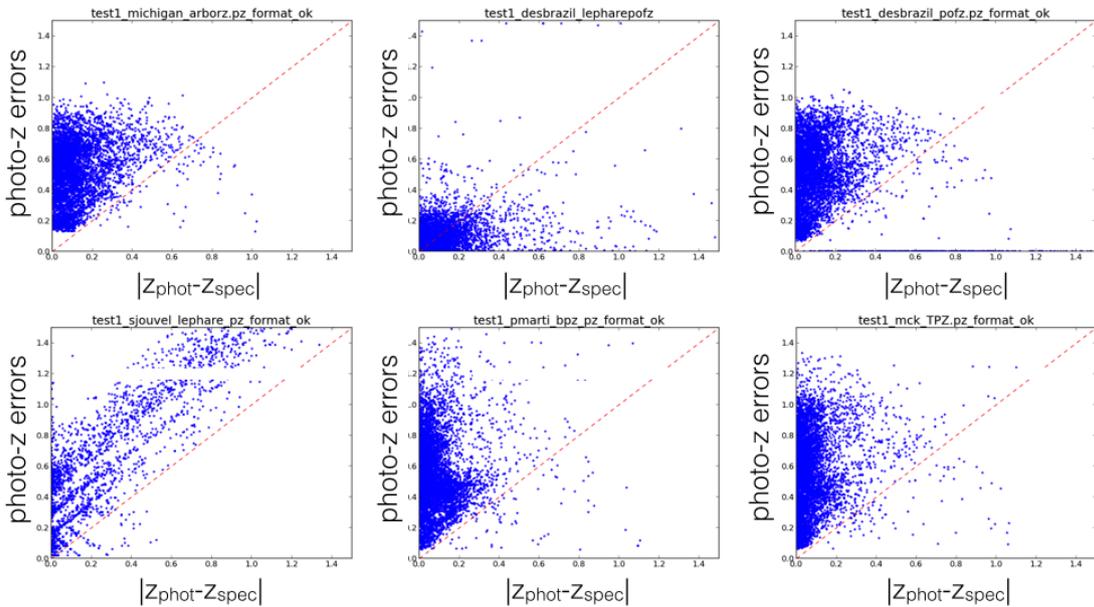


Figura 5.6:  $\epsilon_{phot}$  versus  $\overline{\Delta z}$  para métodos que fornecem a  $P(z)$ .

# Capítulo 6

## Resultados para a amostra BCC

### 6.1 Definição de critérios de seleção da amostra

A partir de resultados dos testes anteriores, percebemos que a qualidade dos dados fotométricos influencia diretamente a estimativa do *photo-z*. Por conseguinte, a performance da estimativa do *photo-z*'s influenciará na qualidade da determinação da LF. Sendo assim, antes de cumprirmos todas as etapas para a determinação da LF, precisamos remover os dados de má qualidade da amostra.

Uma abordagem simples é a remoção de objetos com baixa razão sinal-ruído ( $S/N$ ), cuja detecção e mensuração são duvidosas. Porém, é preciso definir quantitativamente quais objetos são considerados com qualidade suficiente para fazer parte da amostra. É importante ter em conta que um corte muito drástico pode remover objetos de interesse, prejudicando o cálculo da LF através do aumento da incompletude da amostra.

Para chegarmos objetivamente a um valor de  $S/N$  que represente um limite mínimo aceitável, utilizamos as métricas de avaliação da qualidade do *photo-z*, calculadas na amostra de validação para medir o impacto do corte em  $S/N$ . A tabela 6.1 mostra os resultados para a métrica  $\sigma_{68}$ , levando em conta os diferentes cortes em  $S/N$ , nos dois conjuntos de treinamento disponíveis e para os dois métodos de estimativa de *photo-z* disponibilizados no TPZ.

Tabela 6.1: Comparação baseada na métrica  $\sigma_{68}$ , entre as performances do TPZ, após diferentes seleções baseadas em razão sinal-ruído.

Conjunto	metodo	amostra completa	$S/N > 3$	$S/N > 5$	$S/N > 10$
Representativo	TPZ mode	0.1521±0.0007	0.1344±0.0010	0.1080±0.0006	0.0465±0.0003
Representativo	TPZ mean	0.1643±0.0008	0.1487±0.0009	0.1190±0.0007	0.0633±0.0004
OzDES-like	TPZ mode	0.1586±0.0009	0.1399±0.0007	0.1140±0.0007	0.0752±0.0004
OzDES-like	TPZ mean	0.1586±0.0010	0.1498±0.0008	0.1204±0.0007	0.0821±0.0005

Conforme o esperado, nota-se que quanto maior o limite de  $S/N$  mínimo permitido na amostra, menor o valor do  $\sigma_{68}$ , o que se reflete em uma menor dispersão dos *photo-z*'s em

torno do valor esperado ( $true-z^1$ ), portanto uma melhor performance na determinação do cálculo do *redshift*.

A escolha da  $S/N$  mínima a ser aceita seguirá as recomendações dos requisitos científicos do levantamento DES. Para a métrica  $\sigma_{68}$ , o requisito para o cumprimento das demandas dos estudos da energia escura é de  $\sigma_{68} < 0.12$ . A partir de  $S/N > 5$ , todos os resultados cumprem o requisito (dentro da margem de erro). Portanto, este será o corte adotado neste trabalho. Deste ponto em diante, todos os resultados apresentados se referem à amostra após serem removidos<sup>2</sup> os objetos com  $S/N < 5$ .

## 6.2 Photo-z

### 6.2.1 Validação do *Photo-z*

Para quantificar a qualidade do *photo-z* calculado pelo TPZ, usaremos as métricas definidas no capítulo 2. Os resultados obtidos para a amostra de validação após o treinamento ter sido realizado com o conjunto representativo são os que mostram a qualidade do *photo-z* utilizado na determinação da LF. O conjunto de treinamento OzDES-like foi criado para medir o impacto, na qualidade do *photo-z*, da utilização de um conjunto de treinamento criado a partir de dados espectroscópicos reais.

Devido a questões observacionais, os perfis das distribuições de magnitudes das amostras espectroscópicas são diferentes daqueles referentes à amostra fotométrica, causando impacto na determinação do *redshift*. Como a espectroscopia dispersa a luz em seu leque de comprimentos de onda, ela exige uma magnitude aparente mínima (de maior valor) mais brilhante que o exigido para um objeto ser detectado pela fotometria. Isso acontece porque o fluxo vindo de um objeto vai ser distribuído em vários *pixels* do detector, e cada um deles necessita de uma razão sinal-ruído mínima para realizar a detecção. Com isso, as galáxias que compõem o conjunto de treinamento são, em média, mais brilhantes (em magnitude aparente) que as galáxias que compõem o conjunto fotométrico.

A tabela 6.2 mostra o conjunto completo de métricas calculadas para o conjunto de validação, após eliminação dos objetos com  $S/N$  inferior a 5 (resultados para os demais cortes e para a amostra completa encontram-se no apêndice F).

---

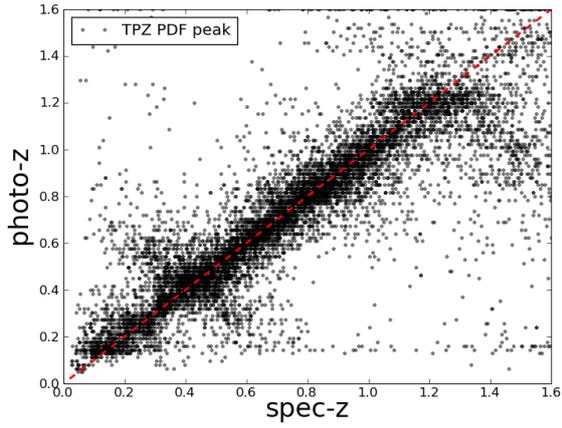
<sup>1</sup>Em alguns momentos, nos referiremos ao *true-z* como *spec-z*. Apesar de o *redshift* da simulação não ser o resultado de uma medida espectroscópica, ele simula o papel deste tipo de dados no cálculo do *photo-z*.

<sup>2</sup>Para evitar impactos nas distribuições de cores, o corte é feito apenas na considerando a banda  $i$ , banda de interesse para o cálculo da LF.

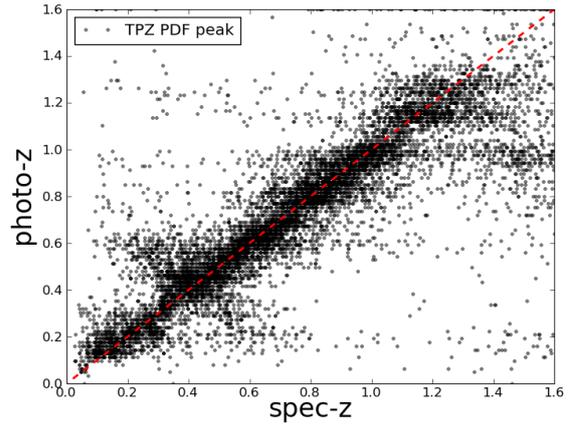
Tabela 6.2: Métricas obtidas após cálculo do photo-z no conjunto de validação.

Métrica	Conjunto Representativo		Conjunto OzDES-like	
	TPZ mode	TPZ mean	TPZ mode	TPZ mean
$\overline{\Delta z}$	$-0.0245 \pm 0.0008$	$0.0050 \pm 0.0007$	$-0.0267 \pm 0.0007$	$-0.0095 \pm 0.0007$
$\Delta z_{50}$	$-0.0031 \pm 0.0003$	$0.0043 \pm 0.0004$	$-0.0057 \pm 0.0004$	$-0.0009 \pm 0.0004$
$\sigma_{\Delta z}$	$0.2039 \pm 0.0013$	$0.1962 \pm 0.0010$	$0.2111 \pm 0.0012$	$0.1983 \pm 0.0011$
$\sigma_{68}$	$0.1080 \pm 0.0006$	$0.1190 \pm 0.0007$	$0.1140 \pm 0.0007$	$0.1204 \pm 0.0007$
$out_{2\sigma}$	$0.0590 \pm 0.0006$	$0.0595 \pm 0.0006$	$0.0603 \pm 0.0006$	$0.0603 \pm 0.0005$
$out_{3\sigma}$	$0.0216 \pm 0.0004$	$0.0216 \pm 0.0003$	$0.0227 \pm 0.0003$	$0.0232 \pm 0.0003$
$\overline{\Delta z'}$	$-0.1165 \pm 0.0049$	$-0.0384 \pm 0.0044$	$-0.1479 \pm 0.0047$	$-0.1040 \pm 0.0047$
$\sigma_{\Delta z'}$	$1.3763 \pm 0.0104$	$1.2879 \pm 0.0085$	$1.3802 \pm 0.0127$	$1.3241 \pm 0.0126$
$N_{poisson}$	$6.1084 \pm 0.0887$	$8.4844 \pm 0.0936$	$8.0476 \pm 0.0730$	$8.2502 \pm 0.1055$
$KS$	$0.0517 \pm 0.0011$	$0.0477 \pm 0.0009$	$0.0625 \pm 0.0010$	$0.0598 \pm 0.0010$

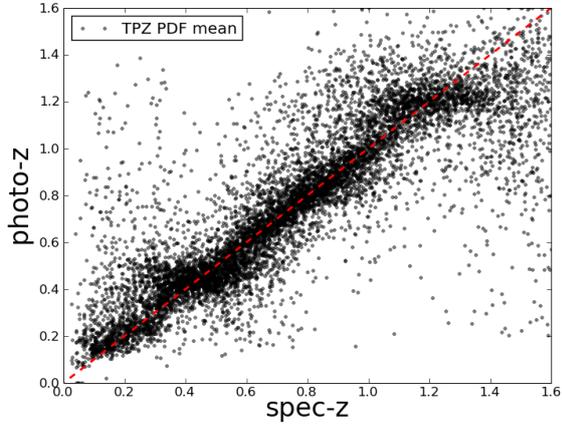
Segundo a métrica  $\overline{\Delta z}$ , o método TPZ mode provoca  $\sim 5$  vezes mais viés quando treinado pelo conjunto representativo e  $\sim 3$  vezes mais viés quando treinado pelo conjunto OzDES-like, em comparação ao método TPZ mean. Por outro lado, segundo a métrica  $\sigma_{\Delta z}$  que é mais robusta, para o conjunto representativo, o resultado com o método TPZ mode é  $\sim 30\%$  melhor que com o TPZ mean. Já para o conjunto OzDES like, o método TPZ mean possui  $\sigma_{\Delta z} \sim 6$  vezes menor. Segundo as métricas de dispersão  $\sigma_{\Delta z}$  e  $\sigma_{68}$ , os resultados para os métodos TPZ mean e TPZ mode concordam razoavelmente, dentro das margens de erro (com pequenas discrepâncias menores do que 2%). A fração de objetos com erros catastróficos é praticamente a mesma para os dois métodos, TPZ mode e TPZ mean, sendo ligeiramente menor no conjunto representativo (menos de 6%, de diferença).



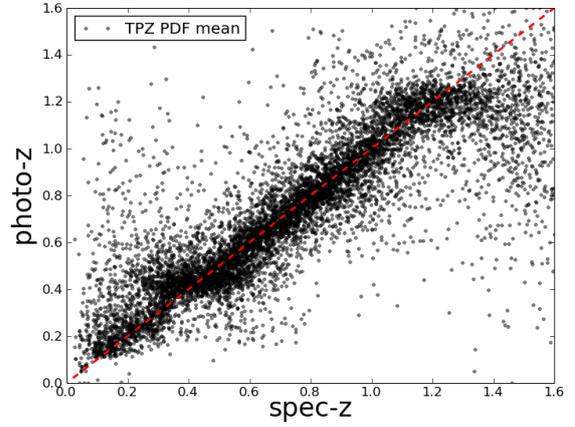
(a) TPZ mode - conj. Representativo



(b) TPZ mode - conj. OzDES-like



(c) TPZ mean - conj. Representativo



(d) TPZ mean - conj. OzDES-like

Figura 6.1: Gráfico de dispersão da relação entre  $photo-z$  e  $spec-z$ , para resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto representativo (à esquerda) e OzDES-like (à direita).

As figuras 6.1 e 6.2 mostram a dispersão dos resultados de  $photo-z$  em torno do valor esperado ( $spec-z$ , correspondente à reta a  $45^\circ$ ). Enquanto os gráficos do painel 6.1 evidenciam os objetos muito discrepantes, os gráficos do painel 6.2 nos permitem perceber com mais clareza as regiões de mais concentração de pontos ao longo da linha diagonal. Nestes gráficos fica bastante evidente uma característica peculiar ao método TPZ mode, que é a estratificação dos resultados, apresentando o gráfico com aspecto de listras. Esta característica pode ter sido introduzida como consequência da escolha inadequada dos parâmetros de configuração. Como neste caso o  $redshift$  é estimado como o valor central do intervalo onde ocorre o valor máximo da PDF, o conjunto de possíveis valores de  $photo-z$  é um conjunto discreto. Este efeito poderia ser minimizado definindo intervalos menores para a construção da PDF, o que, em contra-partida, aumentaria o tempo de processamento. Devido a essa característica, somada ao desempenho obtido das métricas, decidimos então prosseguir para

as demais etapas adotando o método TPZ mean.

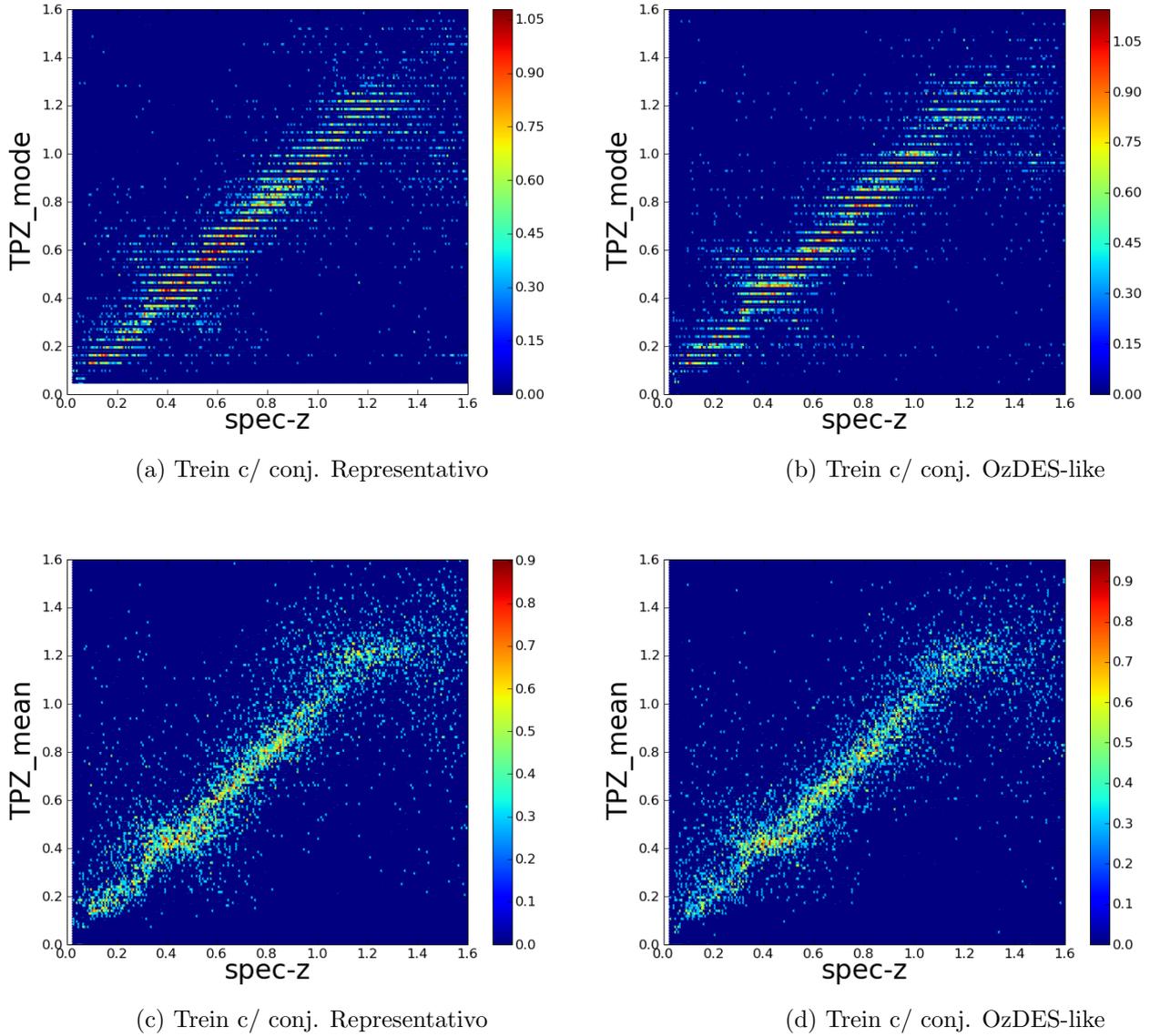


Figura 6.2: Gráfico de densidades da relação entre  $photo-z$  e  $spec-z$ , para resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto representativo (à esquerda) e OzDES-like (à direita). A escala de cores representa a densidade de pontos em cada região do gráfico, em escala logarítmica.

De acordo com as os painéis 6.3a, 6.3b e a figura 6.4, o método TPZ mean fornece uma reconstrução de  $N(z)$  menos ruidosa que o método TPZ mode, apesar de apresentar alguns picos indesejáveis.

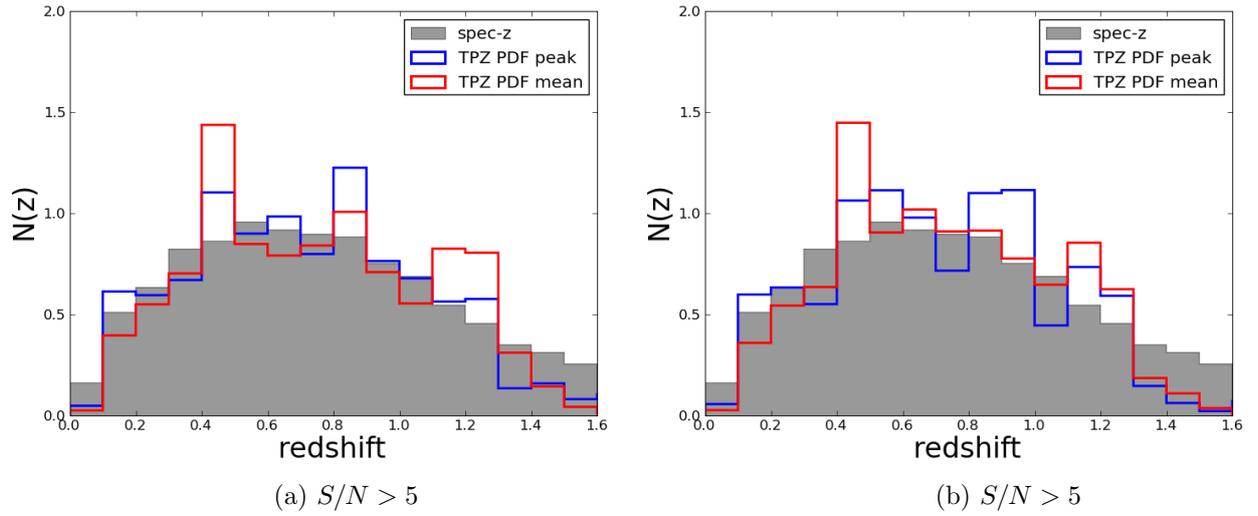


Figura 6.3: Distribuição de *redshifts* obtidos na amostra de validação, após o treinamento ter sido realizado com os conjuntos representativo (6.3a) e OzDES-like (6.3b). Em azul, o resultado com o método TPZ mode. Em vermelho, resultados para TPZ mean. Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1).

### 6.2.2 *Photo-z* da amostra fotométrica

O *photo-z* para a amostra fotométrica foi calculado utilizando os métodos TPZ mode e TPZ mean, selecionando apenas objetos com razão  $S/N > 5$ . As árvores de decisão produzidas com o conjunto de treinamento representativo na etapa anterior (de validação) foram reaproveitadas.

Segundo o gráfico da figura 6.4, os dois métodos falham, criando picos indesejáveis na distribuição  $N(z)$ . Ambos provocam um acúmulo artificial de objetos em três pequenos intervalos de  $z$ , devido a estimativa incorreta do mesmo. Como o TPZ mode provoca um pico muito mais alto próximo a  $0.8 < z < 0.9$ , optamos por utilizar o resultado gerado pelo método TPZ mean no cálculo das magnitudes absolutas e na determinação LF.

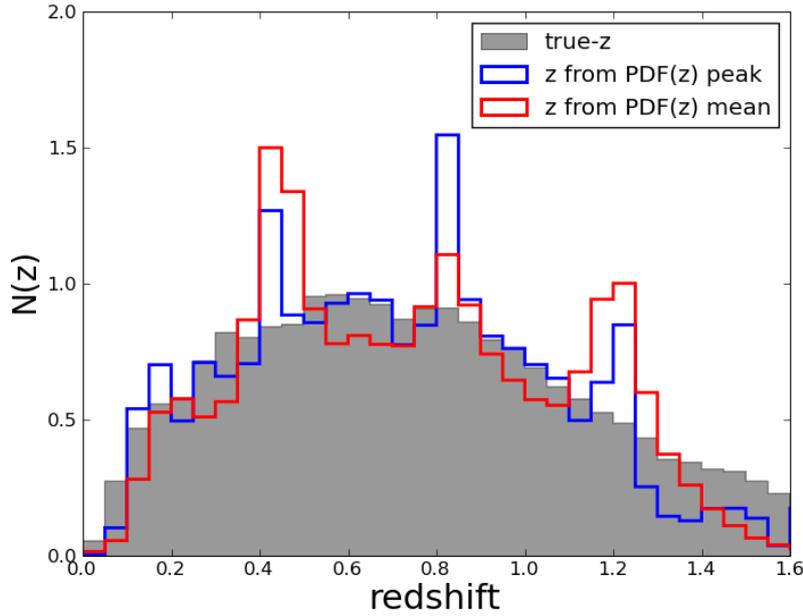


Figura 6.4: Resultado do TPZ para todas as galáxias que serão utilizadas na determinação da função de luminosidade (seleção de objetos com  $S/N > 5$ ). Árvores produzidas com o conjunto representativo. Em azul, o resultado com o método TPZ mode. Em vermelho, resultados para TPZ mean. Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1).

### 6.3 Magnitudes Absolutas

O cálculo das magnitudes absolutas foi efetuado usando-se o algoritmo LePhare, a partir do ajuste das magnitudes aparentes a um conjunto de 29 modelos de SEDs (Coleman, Wu & Weedman 1980 [22]), tendo como condição inicial os *redshifts* verdadeiros da simulação (6.5a) e o *photo-z* calculado pelo método TPZ mean (6.5b).

Na prática, o algoritmo LePhare é utilizado para obter uma estimativa do tipo espectral, através do ajuste das magnitudes a uma determinada SED, dado um *redshift* previamente conhecido. Com isso, obtém-se a extinção e a correção K correspondentes à SED escolhida, ingredientes necessários para o cálculo da magnitude absoluta. A figura 6.5 mostra a distribuição de magnitudes absolutas calculadas.

Nota-se que os erros introduzidos pelo *redshift* fotométrico causam deformações no perfil da distribuição de magnitudes absolutas. Essas deformações, tais como o excesso de galáxias nas faixas de  $-21$  a  $-22$  e de  $-17$  a  $-19$ , bem como a falta de galáxias na faixa de  $-16$  a  $-15$ .

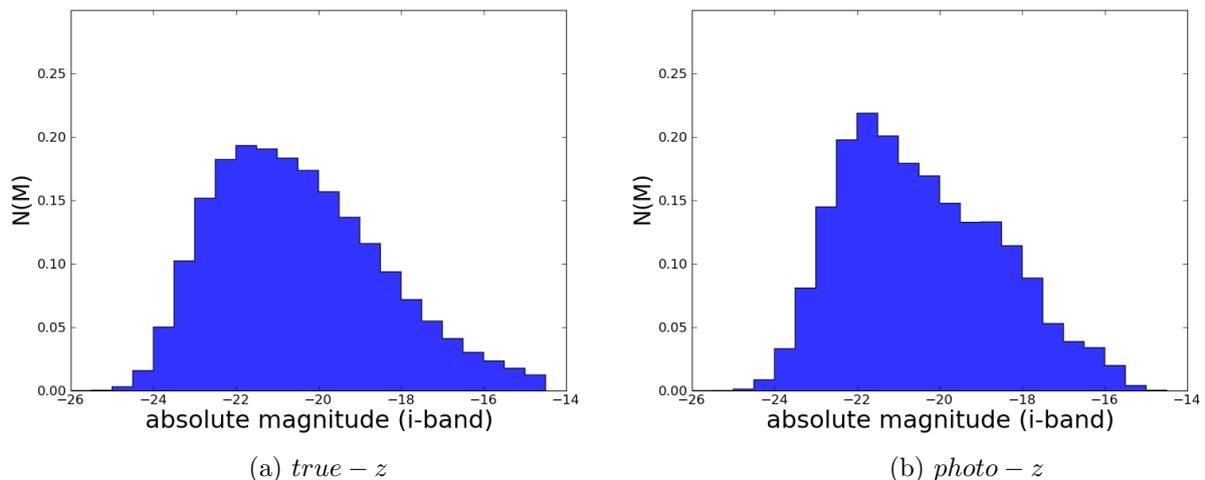


Figura 6.5: Distribuição de magnitudes absolutas geradas pelo LePhare para o conjunto fotométrico com seleção em  $S/N > 5$ . À esquerda: magnitudes absolutas geradas a partir dos *redshifts* verdadeiros da simulação. À direita: magnitudes absolutas geradas a partir dos *photo-z*'s calculados com o TPZ. Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1).

A distribuição de magnitudes absolutas derivadas do *true-z* também possui suas incertezas, fruto da dependência do ajuste às SEDs, um procedimento que pode falhar. Por exemplo, uma galáxia *late-type* cuja luz sofreu extinção pode ser erradamente associada a uma SED de galáxia do tipo *early-type*. A falta de informação na região do ultravioleta pode levar a degenerescências entre SEDs.

## 6.4 Função de Luminosidade

Os resultados apresentados a seguir foram obtidos a partir do método  $1/V_{max}$ , utilizando os *redshifts* originais da simulação e o *photo-z* calculado pelo TPZ. A magnitude limite da amostra foi estimada como 24.4 na banda *i*, a partir da magnitude mediana das galáxias que possuem razão  $S/N = 5 \pm 0.3$ . A LF foi ajustada à função de Schechter pelo método dos mínimos quadrados, definida nos intervalos indicados na tabela 6.3. Apesar de a amostra BCC possuir objetos até  $z = 2$ , mostraremos apenas a LF até  $z = 1.3$ , devido a grande incompletude na amostra para altos valores de  $z$ .

Os ajustes globais da LF obtidos para intervalos de *redshift* adotados são apresentados na figura 6.6. Representados em preto estão as densidades obtidas com os *redshifts* verdadeiros, e em vermelho com os *photo-zs* estimados pelo algoritmo TPZ. Nos intervalos de maior *redshift* ( $z > 0.8$ ), o parâmetro  $\alpha$  foi mantido fixo. De uma forma geral, os pontos que representam intervalos onde é possível se perceber uma perda substancial na completude (na parte fraca da função) ou um excesso de objetos (na parte brilhante) foram excluídos do ajuste (símbolos vazios). Em particular, este último é mais crítico para os *photo-zs*, causado, em parte, pela falta da banda *u*.

Tabela 6.3: Limites definidos iterativamente para o ajuste à função de Schechter

Redshift	<i>true</i> - <i>z</i>		<i>photo</i> - <i>z</i>	
	M mínima	M máxima	M mínima	M máxima
0.0 a 0.2	-26.0	-14.0	-26.0	-15.7
0.2 a 0.4	-26.0	-15.7	-26.0	-15.7
0.4 a 0.6	-25.4	-17.7	-25.1	-17.7
0.6 a 0.8	-25.1	-18.3	-25.1	-18.5
0.8 a 1.0	-25.6	-19.7	-25.6	-19.7
1.0 a 1.4	-25.6	-20.3	-25.1	-22.0

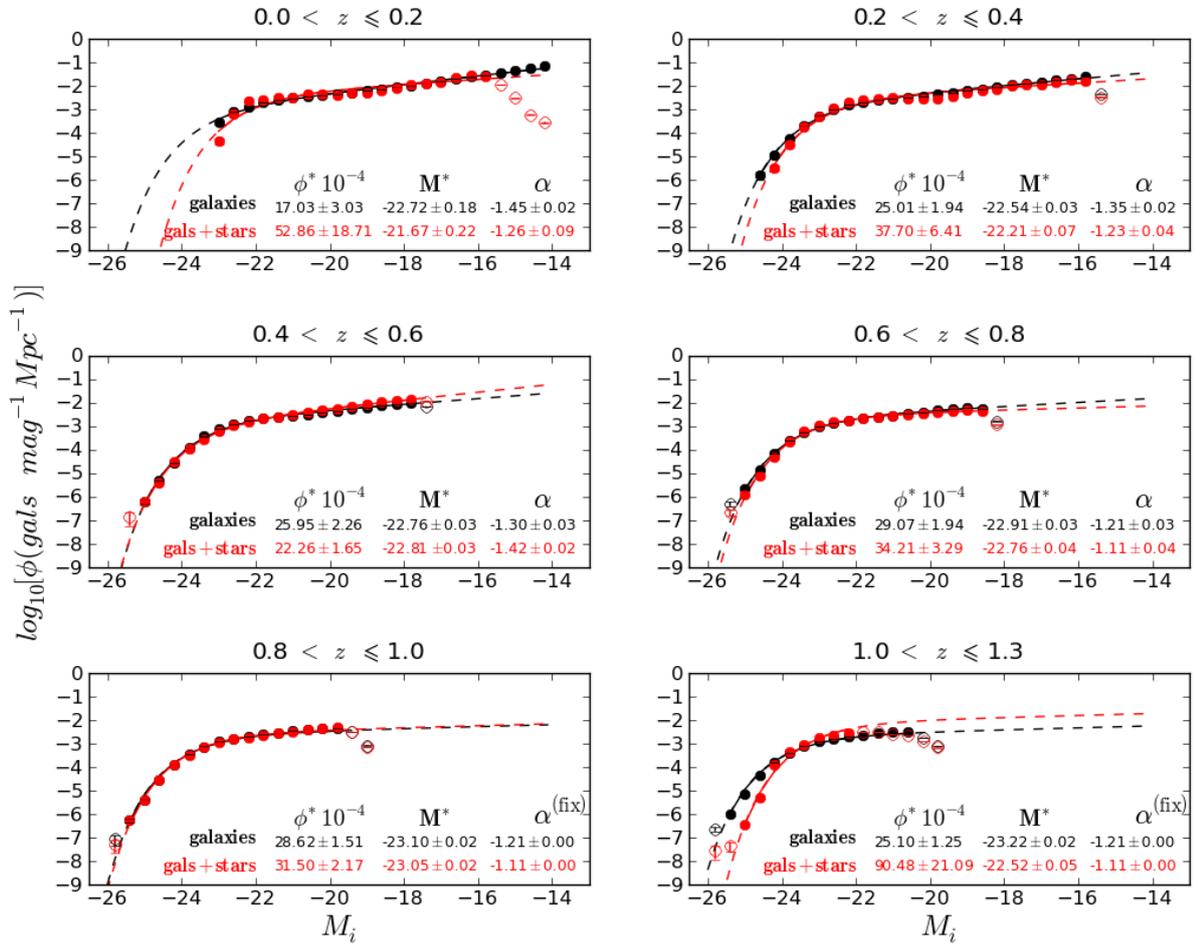


Figura 6.6: Função de luminosidade global para diversos intervalos de *redshift*. Comparação entre os resultados da função de luminosidade utilizando o *redshift* verdadeiro (em preto), proveniente da simulação e o *photo-z* calculado pelo TPZ (em vermelho). A parte tracejada representa extrapolação do ajuste. Os pontos vazios não foram considerados no ajuste.

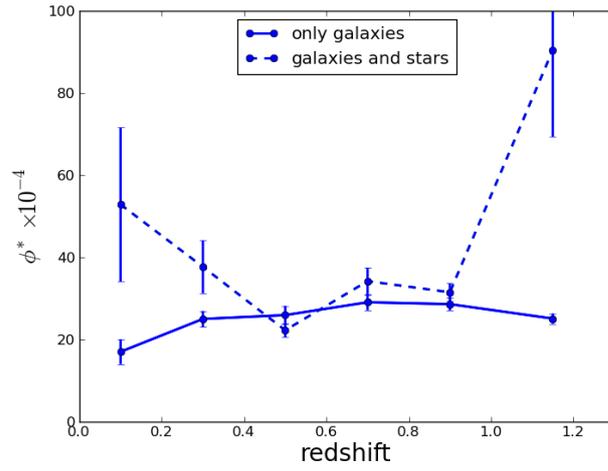
Tabela 6.4: Parametros de Schechter

Redshift	$\phi^* \times 10^{-4}$		$M^*$		$\alpha$	
	<i>true</i> - $z$	<i>photo</i> - $z$	<i>true</i> - $z$	<i>photo</i> - $z$	<i>true</i> - $z$	<i>photo</i> - $z$
0.0 a 0.2	17.03 $\pm$ 03.03	52.86 $\pm$ 18.71	-22.72 $\pm$ 0.18	-21.67 $\pm$ 0.22	-1.45 $\pm$ 0.02	-1.26 $\pm$ 0.09
0.2 a 0.4	25.01 $\pm$ 01.94	37.70 $\pm$ 06.41	-22.54 $\pm$ 0.03	-22.21 $\pm$ 0.07	-1.35 $\pm$ 0.02	-1.23 $\pm$ 0.04
0.4 a 0.6	25.95 $\pm$ 02.26	22.26 $\pm$ 01.65	-22.76 $\pm$ 0.03	-22.81 $\pm$ 0.03	-1.30 $\pm$ 0.03	-1.42 $\pm$ 0.02
0.6 a 0.8	29.07 $\pm$ 01.94	34.21 $\pm$ 03.29	-22.91 $\pm$ 0.03	-22.76 $\pm$ 0.04	-1.21 $\pm$ 0.03	-1.11 $\pm$ 0.04
0.8 a 1.0	28.62 $\pm$ 01.51	31.50 $\pm$ 02.17	-23.10 $\pm$ 0.02	-23.05 $\pm$ 0.02	-1.21* $\pm$ 0.00	-1.11 $\pm$ 0.00
1.0 a 1.3	25.10 $\pm$ 01.25	90.48 $\pm$ 21.09	-23.22 $\pm$ 0.02	-22.52 $\pm$ 0.05	-1.21* $\pm$ 0.00	-1.11 $\pm$ 0.00

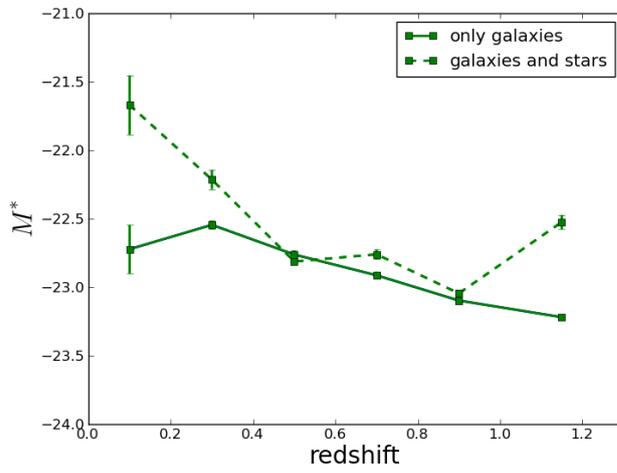
Para examinarmos a evoluao das galaxias, apresentamos a dependencia dos parametros determinados para a funao Schechter como uma funao do *redshift* (ver figura 6.7). Lembramos que nos dois intervalos de maior *redshift*, o  $\alpha$  foi mantido fixo, repetindo o valor obtido no intervalo anterior ( $0.6 < z < 0.8$ ).

Os dois resultados obtidos a respeito da evoluao de  $M^*$  e  $\phi^*$ , o primeiro a partir do *redshift* verdadeiro da simulacao e o segundo a partir do *photo-z* estimado pelo TPZ, concordam razoavelmente entre si para os quatro intervalos centrais de *redshift*. Discrepancias significativas sao observadas nos intervalos extremos (o primeiro,  $0.0 < z < 0.2$  e o ultimo,  $1.0 < z < 1.3$ ) (ver figura 6.7).

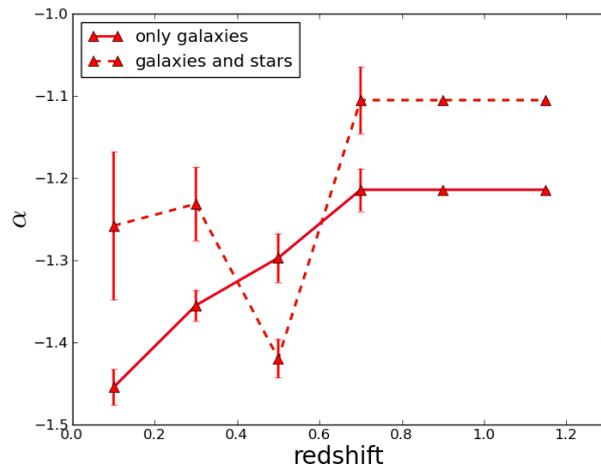
Comparando com resultados anteriores obtidos da literatura, com a ressalva de que os dados possuem calibraao diferente – e de que em alguns casos nem se referem a mesma banda – voltamos nossa atenao para o comportamento geral da evoluao de  $M^*$  e  $\phi^*$  com o *redshift*. Nos dois resultados, o comportamento da evoluao dos parametros concorda com resultados encontrados em Ilbert et al. (2005) [43] (na figura 7 do artigo) e Willmer et al. (2006) [75] (na figura 9 do artigo), que mostram um decrescimo suave no brilho intrinseco e uma constancia (com algum ruido) na densidade media de galaxias para todos os tipos espectrais juntos, com o passar do tempo cosmico (conforme o *redshift* decresce), no intervalo de  $0.2 < z < 1.0$ .



(a)  $\phi^*$



(b)  $M^*$



(c)  $\alpha$

Figura 6.7: Evolução dos parâmetros de Schechter. Linha cheia: LF obtida a partir dos *redshifts* verdadeiros da simulação. Linha tracejada: LF obtida a partir dos *photo-z's* calculados com o TPZ.

# Capítulo 7

## Resultados para a amostra SVA1

### 7.1 Critérios de definição da amostra

Para manter a coerência com a metodologia utilizada na amostra BCC e permitir uma comparação entre os resultados, adotamos na amostra SVA1 o mesmo corte baseado na razão sinal-ruído ( $S/N > 5$ ) utilizado com a amostra BCC. As tabelas 7.1 e 7.3 e os gráficos 7.1 a 7.3 mostram que a performance do TPZ na amostra observada é superior, comparada à performance na amostra simulada.

Tabela 7.1: Comparação baseada na métrica  $\sigma_{68}$ , entre as performances do TPZ, após diferentes critérios de seleção baseado na razão sinal-ruído.

Conjunto	método	sem cortes	$S/N > 3$	$S/N > 5$	$S/N > 10$
OzDES	TPZ mode	$0.0359 \pm 0.0003$	$0.0355 \pm 0.0002$	$0.0339 \pm 0.0002$	$0.0279 \pm 0.0001$
OzDES	TPZ mean	$0.0718 \pm 0.0004$	$0.0666 \pm 0.0004$	$0.0550 \pm 0.0003$	$0.0382 \pm 0.0002$
OzDES ponderado	TPZ mode	$0.0171 \pm 0.0001$	$0.0167 \pm 0.0001$	$0.0164 \pm 0.0001$	$0.0160 \pm 0.0001$
OzDES ponderado	TPZ mean	$0.0183 \pm 0.0001$	$0.0177 \pm 0.0000$	$0.0169 \pm 0.0001$	$0.0161 \pm 0.0001$

### 7.2 Photo-z

#### 7.2.1 Validação do *Photo-z*

Assim como para a amostra BCC, para quantificar a qualidade do *photo-z* calculado pelo algoritmo TPZ, usamos para a amostra SVA1 as mesmas métricas, definidas no capítulo 2.

O conjunto de treinamento derivado dos dados espectroscópicos é o conjunto OzDES. Outro conjunto, o “OzDES ponderado”, foi construído a partir do conjunto OzDES, com o objetivo de treinar o TPZ com uma amostra com distribuição magnitudes mais representativa dos dados da amostra fotométrica. É com este conjunto que serão criadas as árvores que alimentarão o cálculo do *photo-z*.

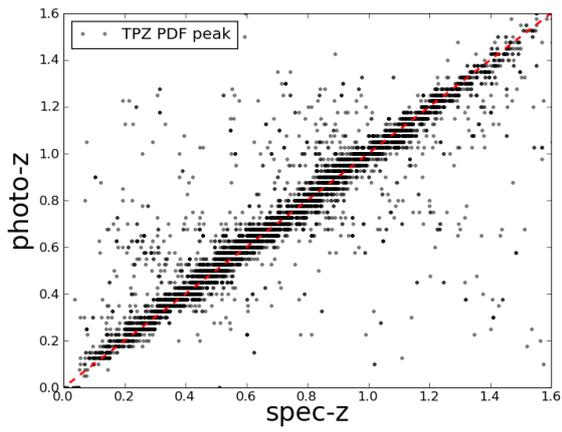
O conjunto completo de métricas calculadas para o conjunto de validação, após eliminação dos objetos com  $S/N < 5$  é mostrado na tabela 7.2. Os resultados para os demais

cortes e para a amostra completas encontram-se no apêndice G.

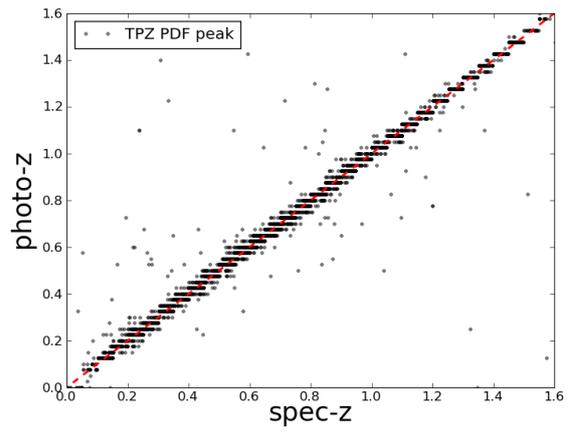
Tabela 7.2: Métricas obtidas após cálculo do *photo-z* no conjunto de validação.

Métrica	Conjunto OzDES		Conjunto OzDES ponderado	
	TPZ mode	TPZ mean	TPZ mode	TPZ mean
$\overline{\Delta z}$	0.0007±0.0005	0.0037±0.0005	0.0006±0.0002	0.0023±0.0002
$\Delta z_{50}$	0.0025±0.0001	0.0044±0.0001	0.0007±0.0001	0.0029±0.0001
$\sigma_{\Delta z}$	0.1404±0.0015	0.1262±0.0010	0.0557±0.0018	0.0520±0.0011
$\sigma_{68}$	0.0339±0.0002	0.0550±0.0003	0.0164±0.0001	0.0169±0.0001
$out_{2\sigma}$	0.0457±0.0005	0.0499±0.0006	0.0092±0.0003	0.0225±0.0006
$out_{3\sigma}$	0.0248±0.0004	0.0251±0.0004	0.0073±0.0003	0.0136±0.0005
$\overline{\Delta z'}$	0.0331±0.0023	0.0507±0.0020	0.0159±0.0009	0.0366±0.0009
$\sigma_{\Delta z'}$	0.6378±0.0076	0.5690±0.0037	0.2641±0.0033	0.2664±0.0021
$N_{poisson}$	3.5907±0.0567	4.3745±0.0798	0.6718±0.0287	0.7946±0.0301
$KS$	0.0502±0.0010	0.0483±0.0008	0.0413±0.0006	0.0372±0.0008

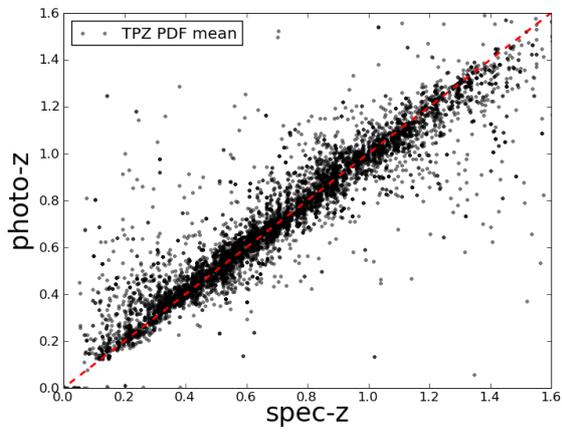
Conforme os resultados apresentados na tabela 7.2 e nas figuras 7.1 a 7.3, a transformação do conjunto de treinamento OzDES em um conjunto ponderado através da utilização dos pesos provoca melhora em todas as métricas que caracterizam viés e dispersão. O método TPZ mean apresenta viés e dispersão ligeiramente maiores que o método TPZ mode, porém suas métricas satisfazem os requisitos científicos.



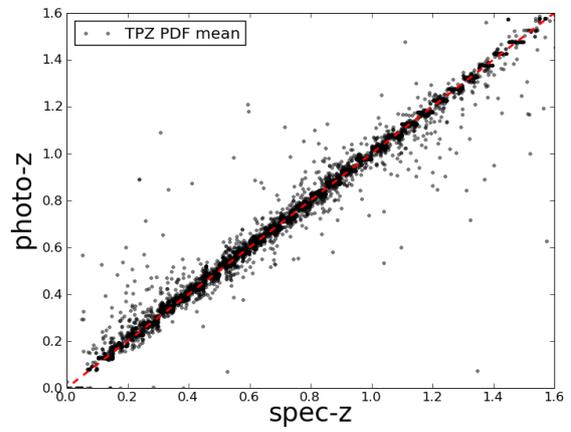
(a) TPZ mode - conj. Representativo



(b) TPZ mode - conj. OzDES-like

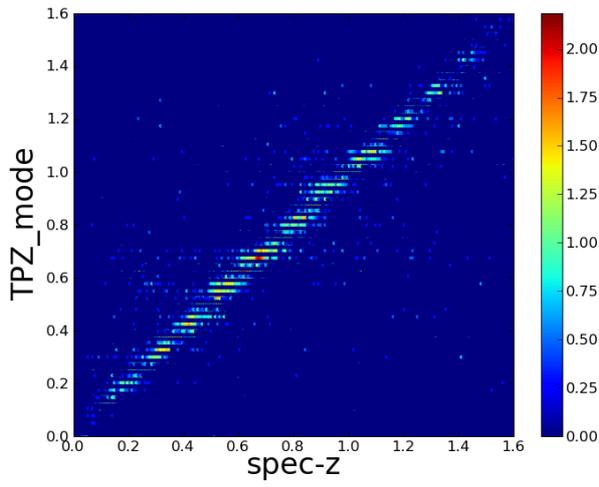


(c) TPZ mean - conj. Representativo

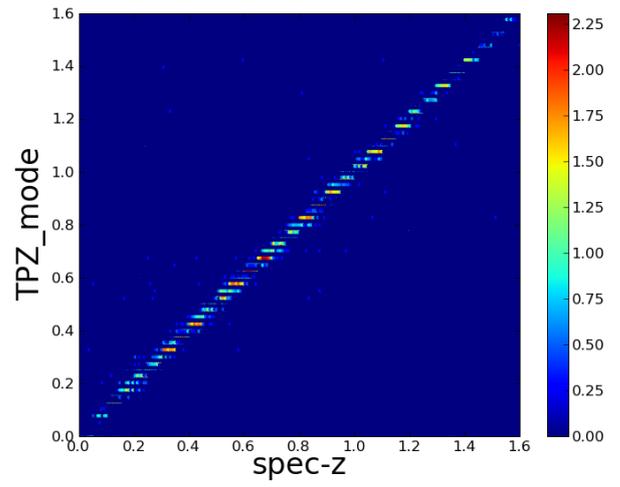


(d) TPZ mean - conj. OzDES-like

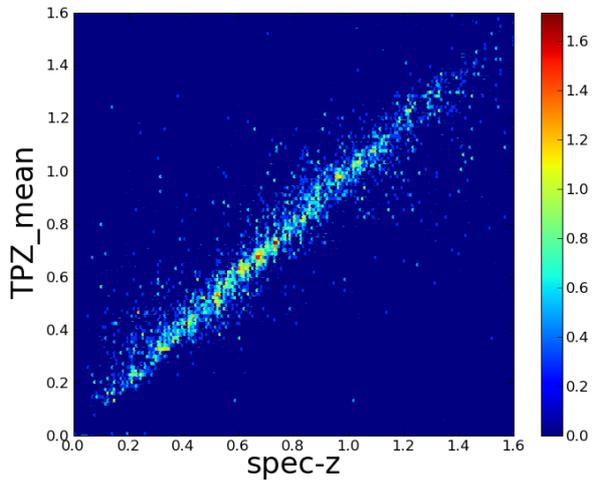
Figura 7.1: Gráfico de dispersão da relação entre  $photo-z$  e  $spec-z$ , para resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto **OzDES** (à esquerda) e **OzDES ponderado** (à direita).



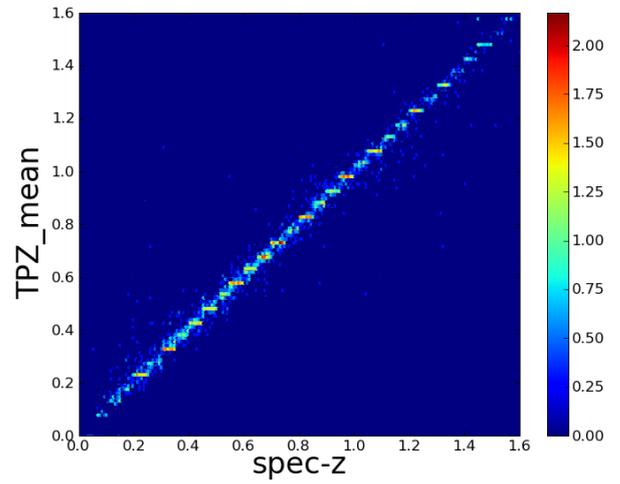
(a) Trein  $c/$  conj. Representativo



(b) Trein  $c/$  conj. OzDES-like



(c) Trein  $c/$  conj. Representativo



(d) Trein  $c/$  conj. OzDES-like

Figura 7.2: Gráfico de densidades da relação entre  $photo-z$  e  $spec-z$ , para resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto **OzDES** (à esquerda) e **OzDES ponderado** (à direita). A escala de cores representa a densidade de pontos em cada região do gráfico, em escala logarítmica.

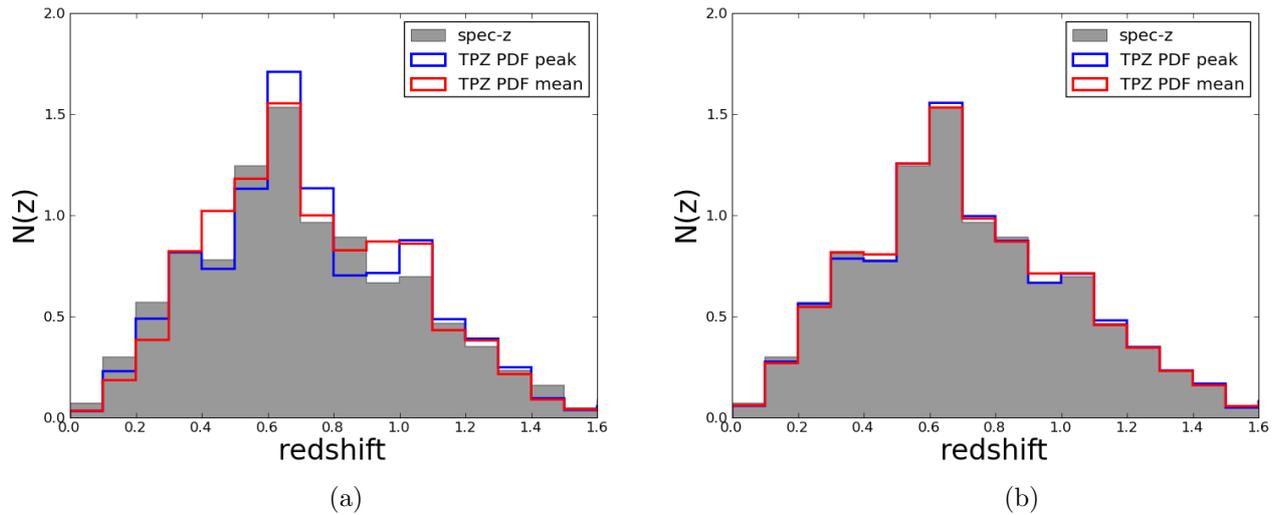


Figura 7.3: Distribuição de *redshifts* obtidos na amostra de validação, após o treinamento ter sido realizado com os conjuntos representativo (7.3a) e OzDES-like (7.3b). Em azul, o resultado com o método TPZ mode. Em vermelho, resultados para TPZ mean. Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1).

## 7.2.2 *Photo-z* da amostra fotométrica

O *photo-z* da amostra fotométrica foi calculado utilizando os métodos TPZ mode e TPZ mean, com a amostra selecionada apenas com objetos cuja razão  $S/N$  é superior a 5. As árvores de decisão produzidas com o conjunto de treinamento representativo na etapa anterior (de validação) foram reaproveitadas.

Assim como na Amostra BCC, a distribuição  $N(z)$  obtida com o TPZ mode é mais ruidosa que aquela obtida com o TPZ mean. Manteremos então a escolha do método TPZ mean para gerar os *photo-z*'s que alimentarão o LePhare no cálculo da magnitude absoluta, e consequentemente na determinação da LF. A figura 7.4 mostra as distribuições obtidas com os dois métodos para a amostra fotométrica.

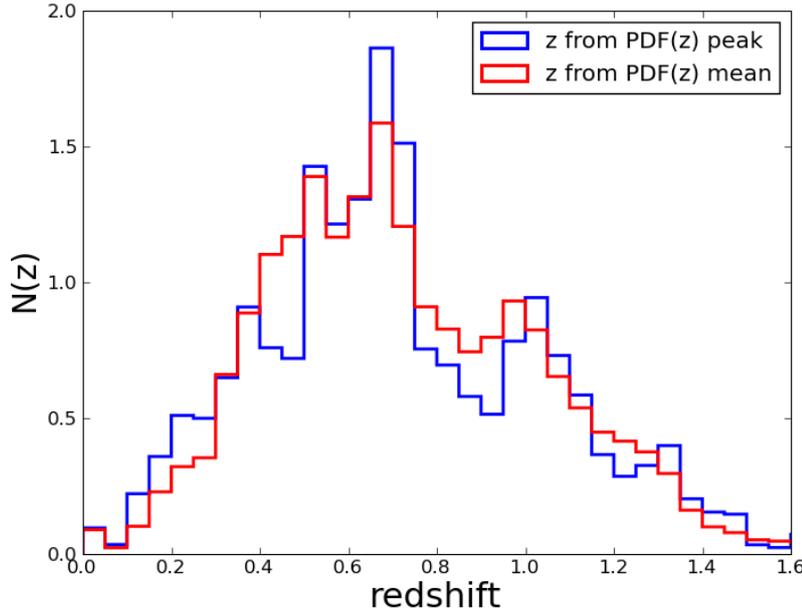


Figura 7.4: Distribuições de *photo-z*'s, para todas as galáxias que serão utilizadas na determinação da função de luminosidade. Em azul, o resultado com o método TPZ mode. Em vermelho, resultados para TPZ mean. Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1).

Ao contrário da amostra simulada BCC, com dados observados não temos nenhuma informação sobre os *redshifts* verdadeiros para compararmos *photo-z* calculado pelo TPZ. Entretanto, uma comparação útil pode ser feita com os resultados que são obtidos ao utilizar a amostra contaminada por estrelas. Ao observar as tendências provocadas pela presença de estrelas nos resultados, podemos especular sobre as incertezas provocadas pelas estrelas remanescentes, já que o método de separação estrela/galáxia não é 100% eficaz.

A figura 7.5 mostra a comparação entre as distribuições de *redshift* obtidas antes e após a remoção de estrelas utilizando o classificador "MODEST\_CLASS", descrito no capítulo 4. No painel (a) temos a real distribuição, em valores absolutos de objetos em cada intervalo de *redshift* e no painel (b) temos uma distribuição normalizada (área total sob as curvas = 1).

Ao deixar a amostra propositalmente contaminada por estrelas, observamos que o TPZ atribui *photo-zs* para essas estrelas em quase todo o intervalo estudado neste trabalho ( $0.0 < z < 1.3$ ). Apenas para  $z \gtrsim 1.2$  os efeitos da contaminação são desprezíveis. O maior impacto é observado na região em torno de  $z \approx 0.6$ . Como o método de separação pode falhar, acreditamos que, mesmo após a separação, alguns objetos que foram contabilizados no cálculo da LF possam ser estrelas, sobretudo no intervalo  $0.6 < z < 0.8$ . Veremos nas seções a seguir o impacto desta contaminação na LF.

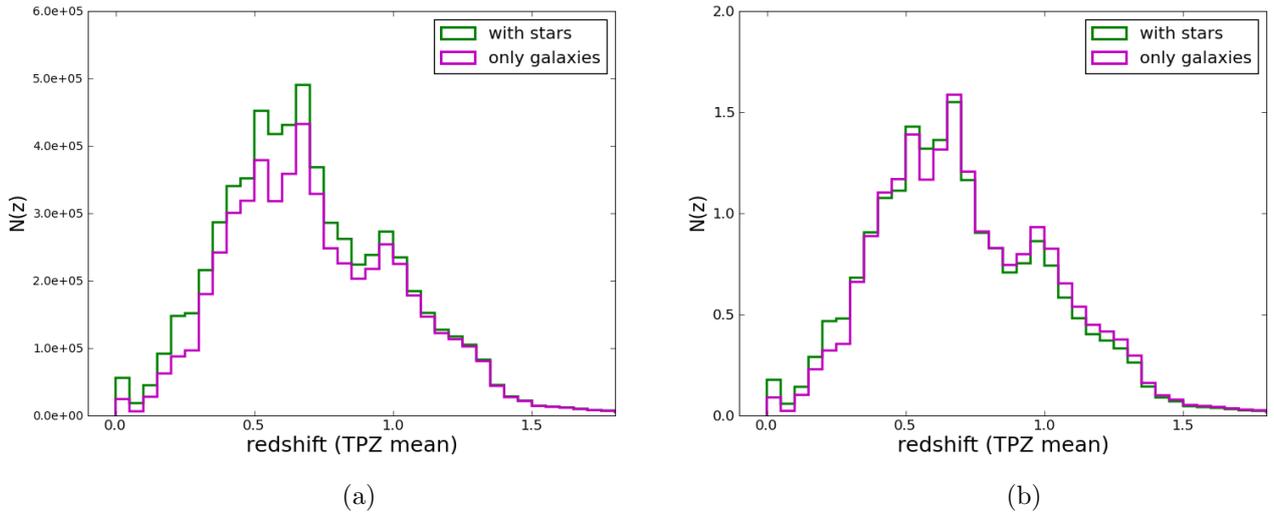


Figura 7.5: Comparação entre as distribuições de  $photo-z$ 's, determinados pelo método TPZ mean (com árvores produzidas com o conjunto OzDES ponderado), para todas as galáxias que serão utilizadas na determinação da função de luminosidade (corte em  $S/N > 5$ ). À esquerda, o histograma comum e à direita o histograma normalizado.

### 7.3 Magnitudes Absolutas

As magnitudes absolutas foram calculadas utilizando o código LePhare, a partir de  $photo-z$ s produzidos pelo algoritmo TPZ, com ajuste das magnitudes aparentes a um conjunto de 29 modelos de SEDs (Coleman, Wu & Weedman 1980 [22]).

Assim como foi mostrado no capítulo anterior para a amostra BCC, para a amostra SVA1 também utilizamos o algoritmo LePhare para obter uma estimativa do tipo espectral, através do ajuste das magnitudes a uma determinada SED, dado um  $redshift$  previamente conhecido. Com isso, obtém-se a extinção e a correção K correspondentes à SED escolhida, ingredientes necessários para o cálculo da magnitude absoluta.

Neste caso, como não temos o  $redshift$  verdadeiro, fornecemos ao LePhare apenas o  $photo-z$  obtido com o TPZ, e comparamos com os resultados obtidos com a mesma amostra, porém mantendo propositalmente a contaminação de estrelas. A figura 7.6 mostra as distribuições de magnitudes absolutas obtidas nos dois casos. Nota-se que a inclusão de estrelas causa um "alargamento" na distribuição de magnitudes, adicionando magnitudes extras em toda a faixa estudada, com destaque para as magnitudes mais brilhantes – contagens quase nulas no lado esquerdo do painel (a), aumentadas no painel (b).

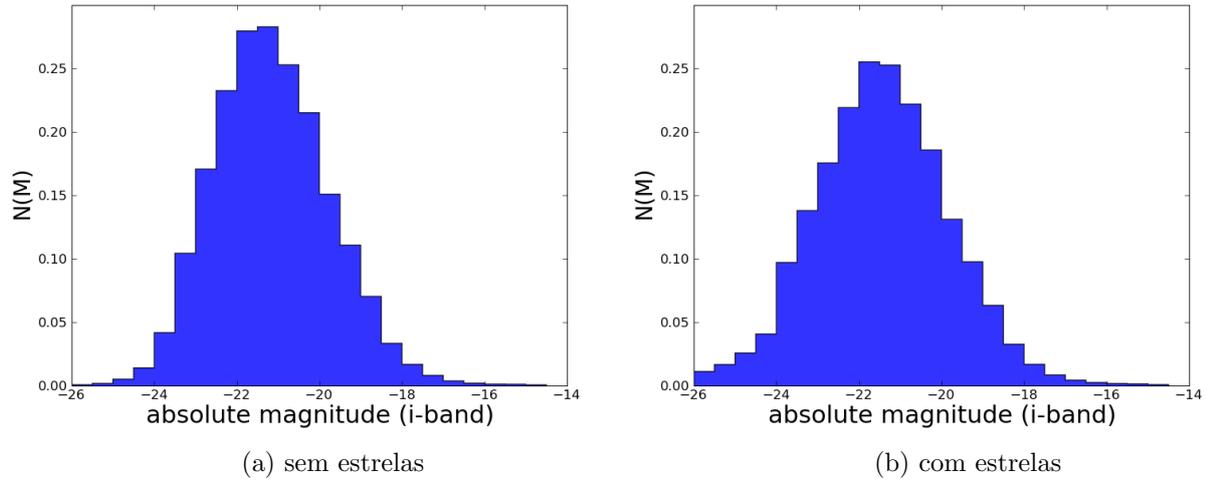


Figura 7.6: Distribuição de magnitudes absolutas geradas pelo LePhare para o conjunto fotométrico com corte em  $S/N > 5$ . À esquerda, amostra após a remoção dos objetos classificados como estrelas. À direita, antes da remoção. Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1).

## 7.4 Função de Luminosidade

Os resultados apresentados a seguir foram obtidos a partir do método  $1/V_{max}$ , utilizando o  $photo-z$  calculado pelo TPZ. Para evitar as consequências da heterogeneidade na profundidade das observações, o que prejudicaria o cálculo do volume na determinação da LF, adotamos um corte em magnitude limite  $m_i < 23.0$ . Chegamos a esse valor a partir da profundidade mínima atingida em todas as regiões, segundo o mapa de profundidades.

A LF foi ajustada à função de Schechter pelo método dos mínimos quadrados, definida nos intervalos indicados na tabela 7.3. Estes intervalos estão explícitos na figura 7.7 e correspondem aos pontos representados por círculo cheio. Os círculos vazados representam pontos que foram excluídos do ajuste à curva.

Tabela 7.3: Limites definidos iterativamente para o ajuste à função de Schechter

Redshift	M min	M max
0.0 a 0.2	-26.0	-14.0
0.2 a 0.4	-25.3	-16.3
0.4 a 0.6	-25.3	-18.8
0.6 a 0.8	-24.7	-19.7
0.8 a 1.0	-25.5	-21.0
1.0 a 1.4	-25.7	-21.3

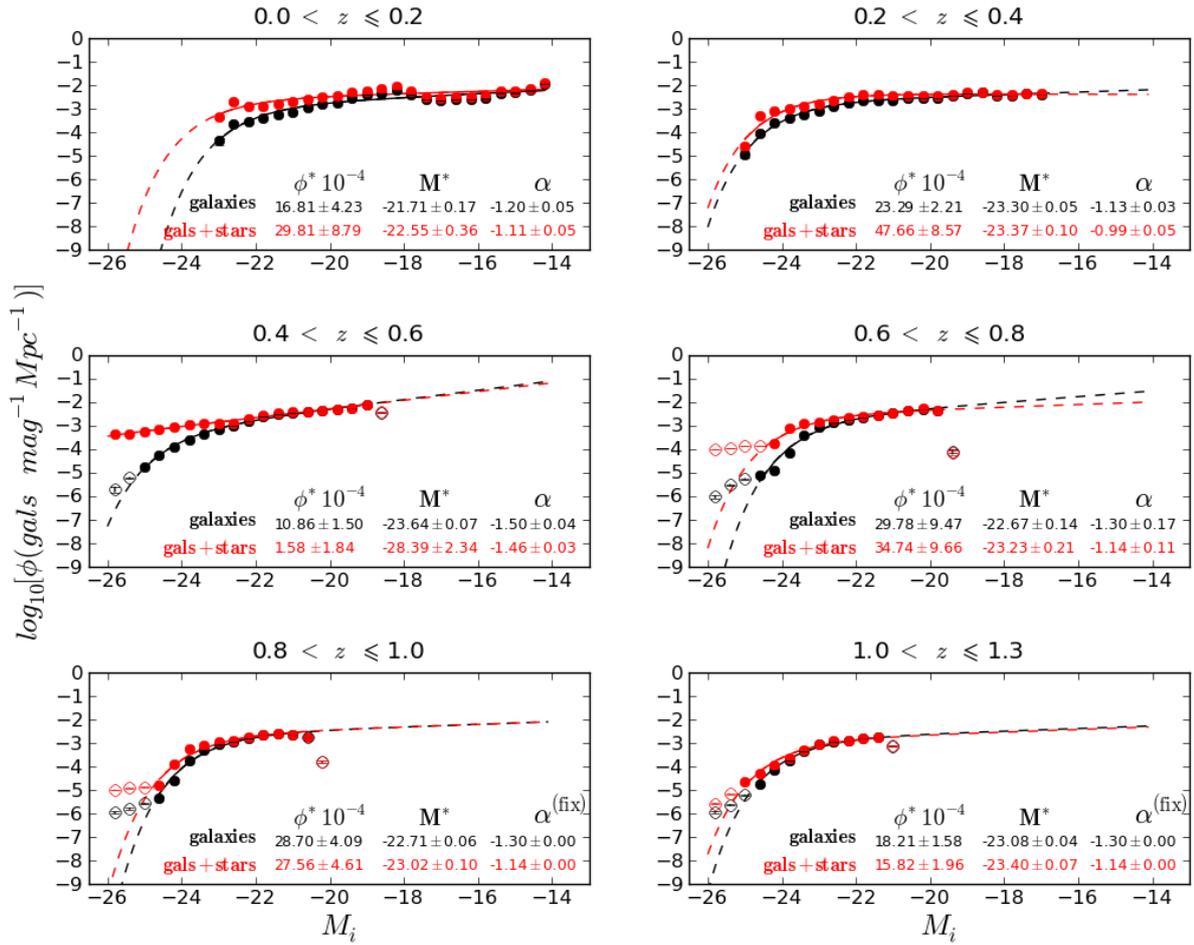


Figura 7.7: Função de luminosidade global em intervalos de *redshift*. Comparação entre os resultados da função de luminosidade utilizando o *photo-z* calculado pelo TPZ, com a amostra antes e após a remoção de estrelas.

A figura 7.7 mostra a função nos intervalos de *redshift* estudados. Em preto temos os pontos correspondente à função para o conjunto fotométrico SVA1, apenas para objetos considerados como galáxias. Em vermelho, temos os pontos gerados pelo cálculo da função na amostra propositalmente contaminada por estrelas. No intervalo de  $0.4 < z < 0.6$  podemos notar a grande presença de estrelas, com *photo-z* erradamente atribuído neste intervalo, e com magnitudes absolutas na parte brilhante da função (mais brilhantes que  $M^*$ ). Nota-se também que que o classificador *MODEST\_CLASS* resolve bem o problema da separação dos objetos neste intervalo. Após a remoção das estrelas, os pontos pretos caem satisfatoriamente sobre a linha da curva da função de Schechter, no regime mais brilhante de magnitudes absolutas.

Por outro lado, nos intervalos de maior *redshift* (onde  $z > 0.6$ ), ainda é observado um excesso de objetos contribuindo para a função na região de magnitudes absolutas mais brilhantes, mesmo após a separação. Este resultado pode ser interpretado como um indício

de que a amostra de galáxias ainda possa ter uma contaminação remanescente, fruto da falha do método separador. A seção a seguinte mostra uma tentativa de driblar este problema.

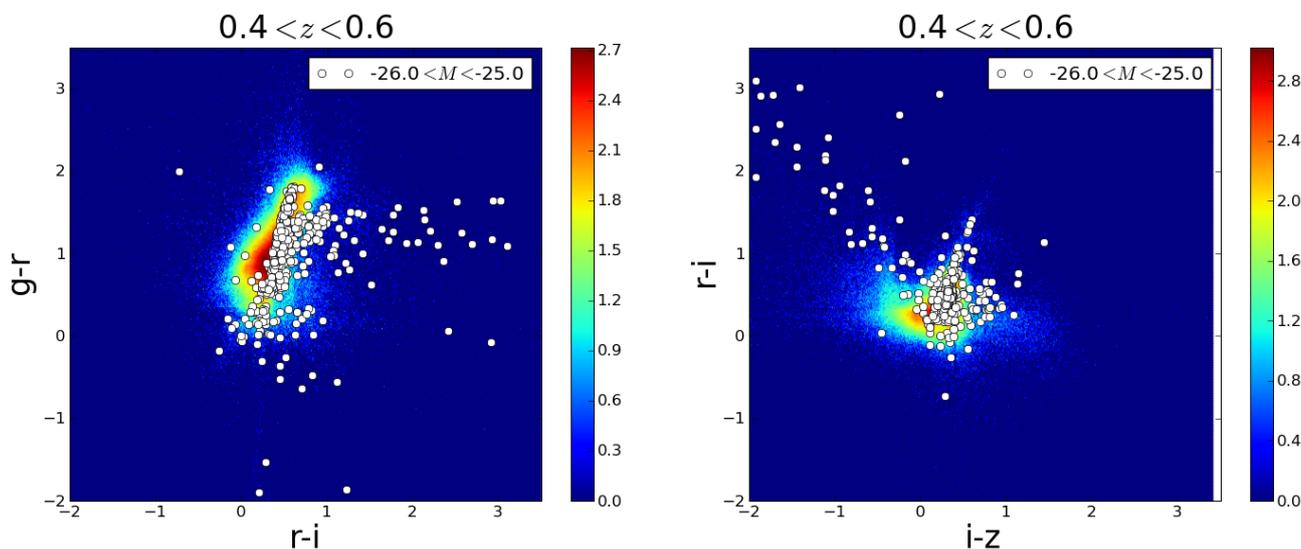
Outra possível explicação para o excesso observado na LF, em intervalos de magnitude absoluta mais brilhante, além da contaminação remanescente por estrelas, é a contaminação por quasares e galáxias com núcleo ativo e problemas nas observações, como mostraremos a seguir.

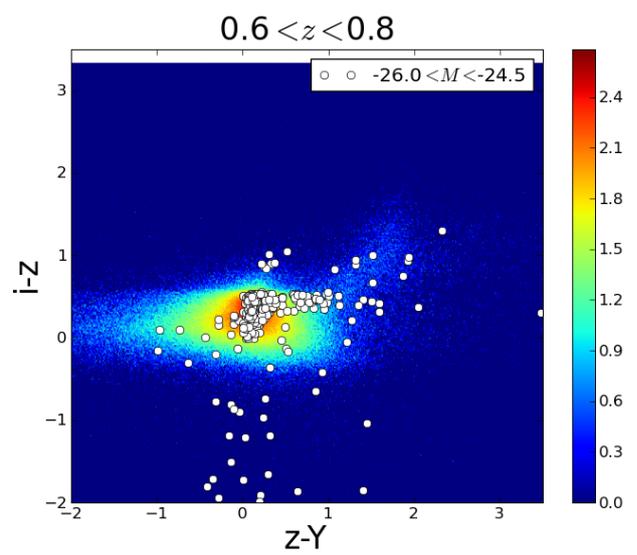
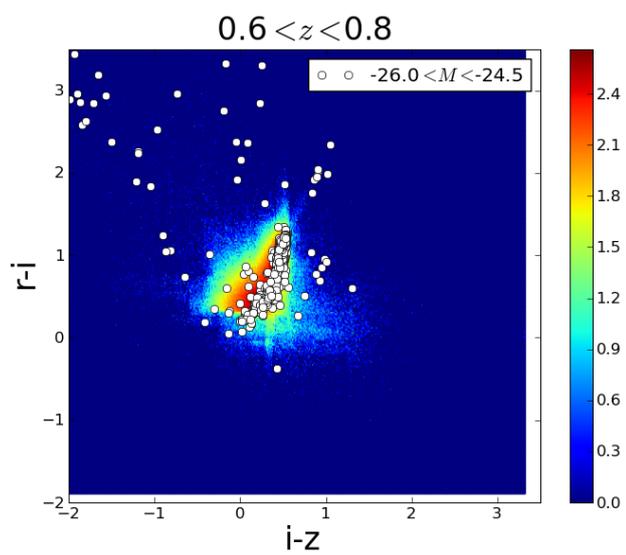
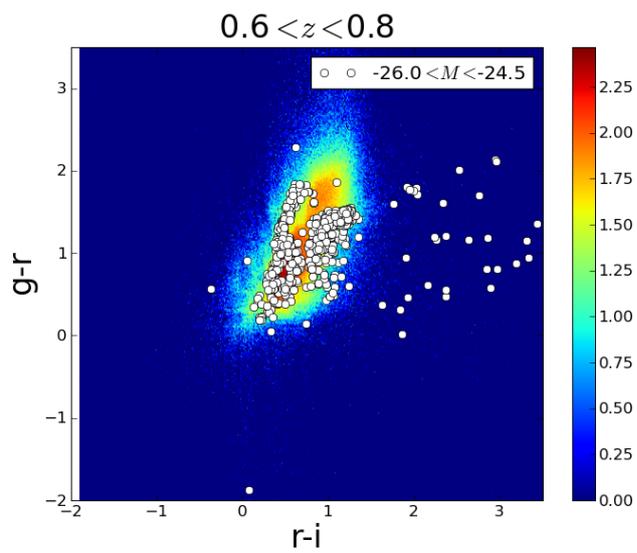
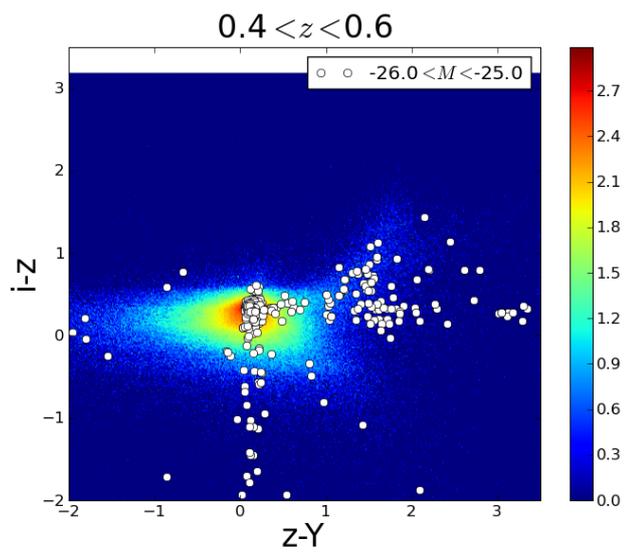
## 7.4.1 Tentativa de limpeza da contaminação remanescente

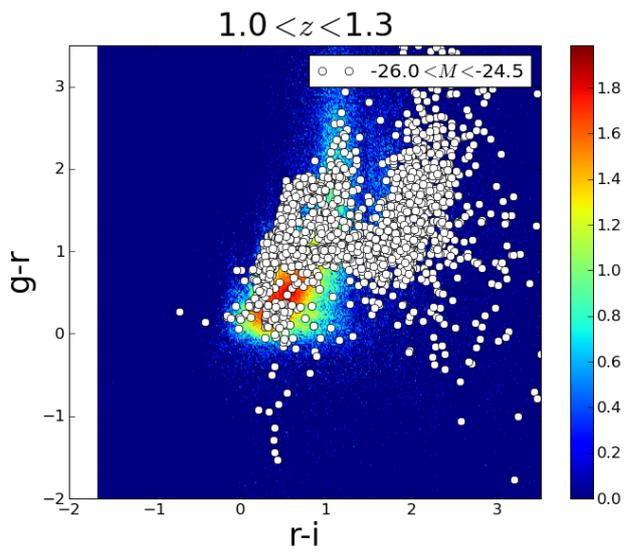
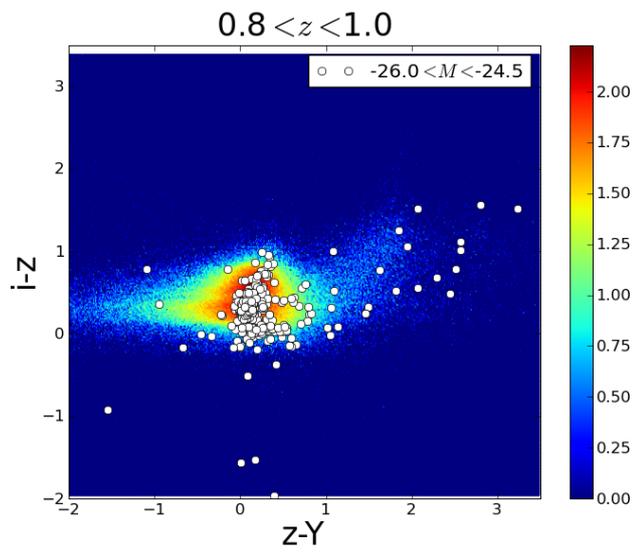
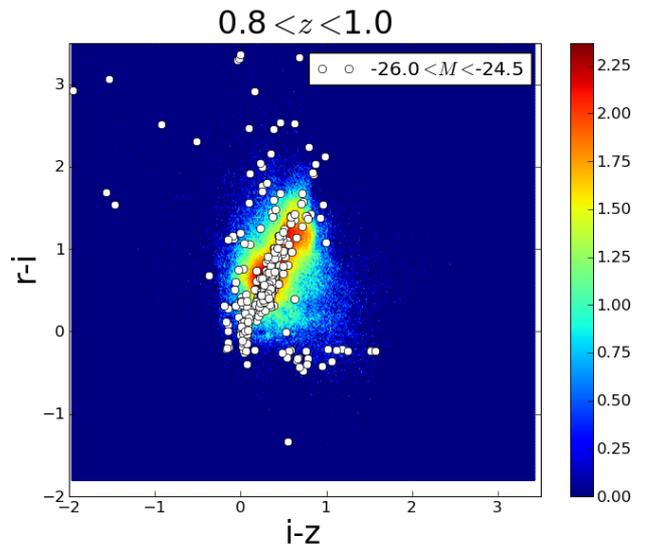
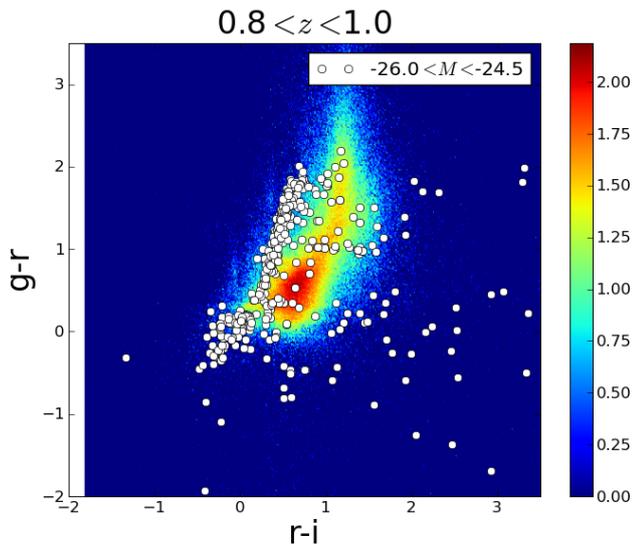
Ao encontrarmos um excesso nos valores de  $\phi(M)$ , em intervalos de magnitude absoluta mais brilhante, iniciamos uma busca pelos motivos pelos quais valores muito baixos de  $M$  tenham sido atribuídos a tantas galáxias dentro da amostra.

### 7.4.1.1 Candidatos a objetos contaminantes no diagrama Cor-Cor

Ao analisar os diagramas cor-cor (7.8), notamos que alguns objetos de magnitude absoluta muito brilhantes (por exemplo na faixa de  $-26.0$  a  $-25.0$ ) ocupam posições fora do *locus* ocupado pelas demais galáxias do mesmo intervalo de  $z$ . Fizemos então uma lista de alvos, e utilizamos a ferramenta *Tile Viewer* do PCDES para investigar a natureza desses objetos.







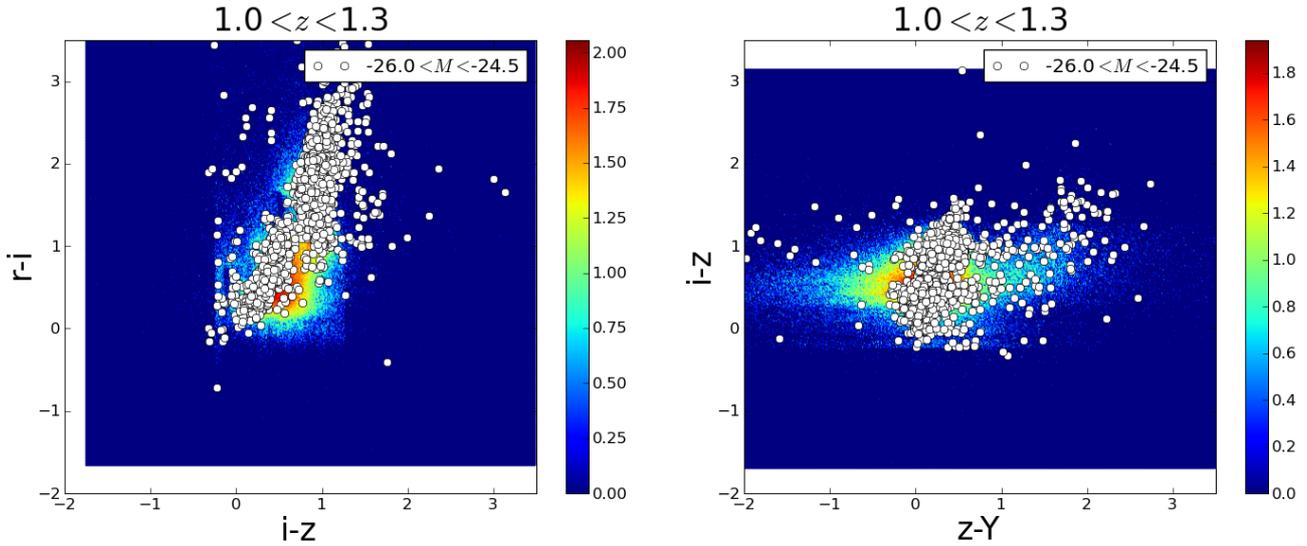


Figura 7.8: Diagramas cor-magnitude e cor-cor para a amostra SVA1 em intervalos de *redshift*. Os objetos com magnitude absoluta dentro dos intervalos onde a LF possui excesso, com respeito ao ajuste de Schechter, estão destacados em branco. As cores representam a densidade de pontos em escala logarítmica.

#### 7.4.1.2 Exemplos de objetos problemáticos

As figuras abaixo mostram os diversos tipos de problemas encontrados na inspeção visual das imagens. Encontramos casos de objetos duplos que estavam tendo os dois fluxos atribuídos a um único objeto, tornando-o mais brilhante; casos de objetos contaminados pelo fluxo de estrelas na vizinhança ou pelo fluxo gerado pelos nossos satélites, que passavam no campo de observação justamente no momento da obtenção das imagens na banda  $i$ ; encontramos também diversos objetos visivelmente com *photo-z* superestimado, levando a uma superestimativa da magnitude absoluta. Na figura 7.14, temos uma galáxia que poderia ser uma BCG (*Brightest Cluster Galaxy*), galáxias mais brilhantes de um aglomerado, que podem provocar excesso na LF, em comparação à curva parametrizada.

A quantidade de objetos com estes tipos de problemas é ínfima em comparação à quantidade de dados da amostra fotométrica (menos de 0.1% do total de galáxias). Porém, como os intervalos de magnitude absoluta muito brilhante são pouco povoados, a LF acaba sendo sensível à presença de tais objetos.

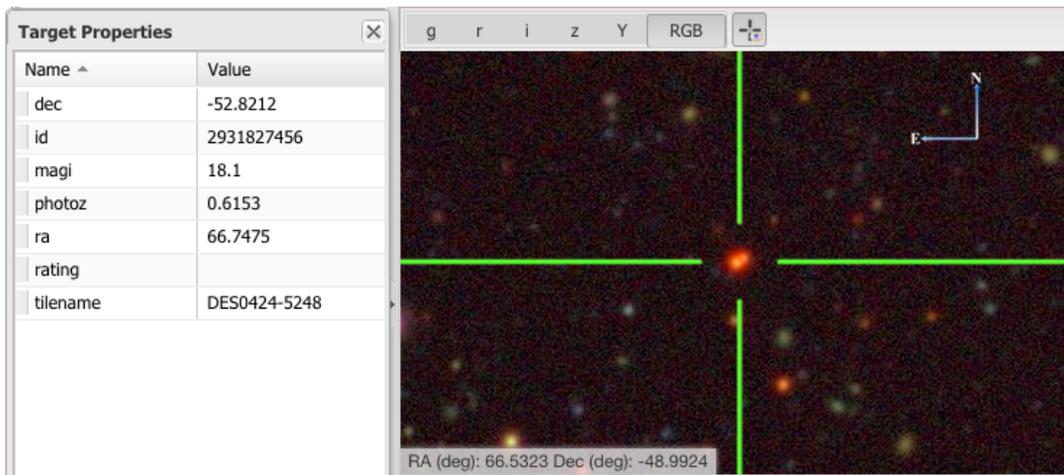


Figura 7.9: Exemplo de objetos duplos detectados como um único objeto.

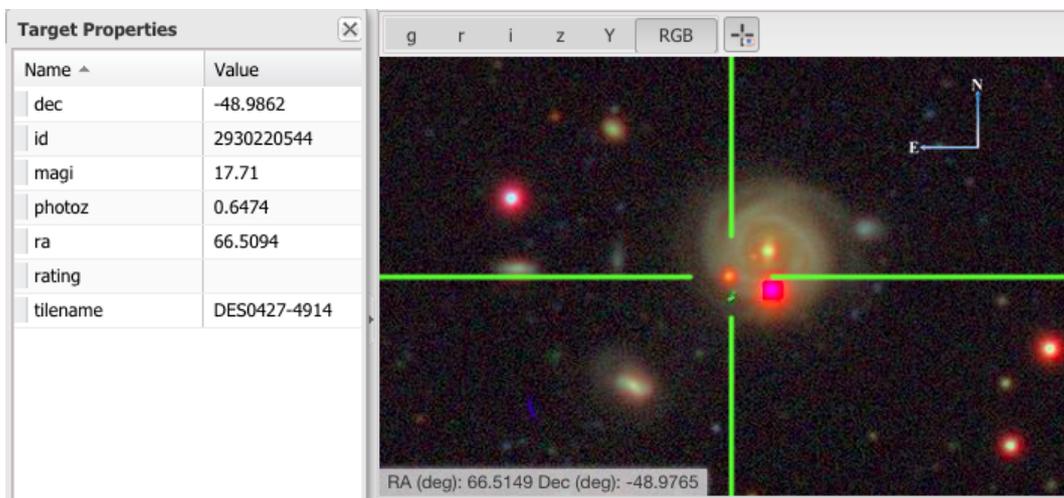


Figura 7.10: Exemplo de objetos contaminados pelo fluxo de outros.

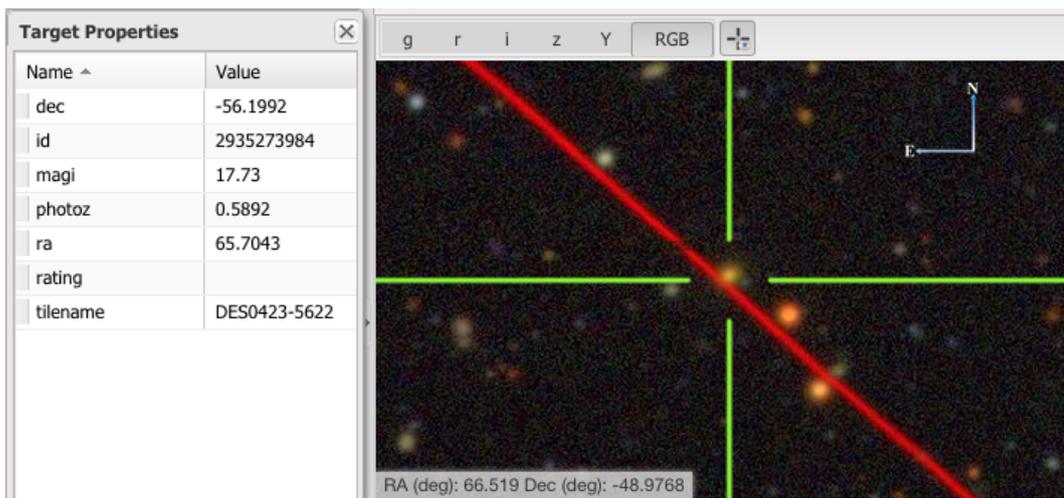


Figura 7.11: Exemplo de objetos contaminados pelo fluxo gerado pela detecção de satélites nas proximidades.

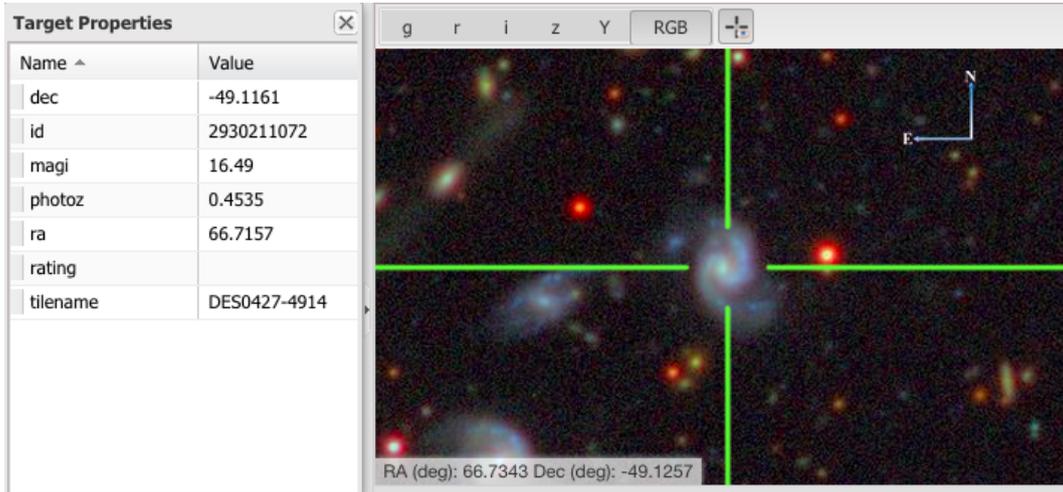


Figura 7.12: Exemplo de *photo-z* mal estimado.

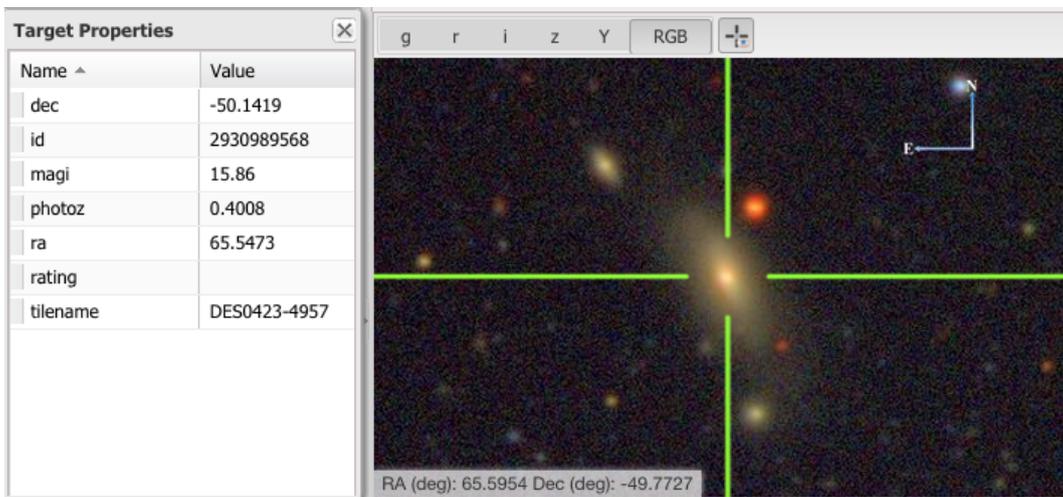


Figura 7.13: Exemplo de *photo-z* mal estimado.

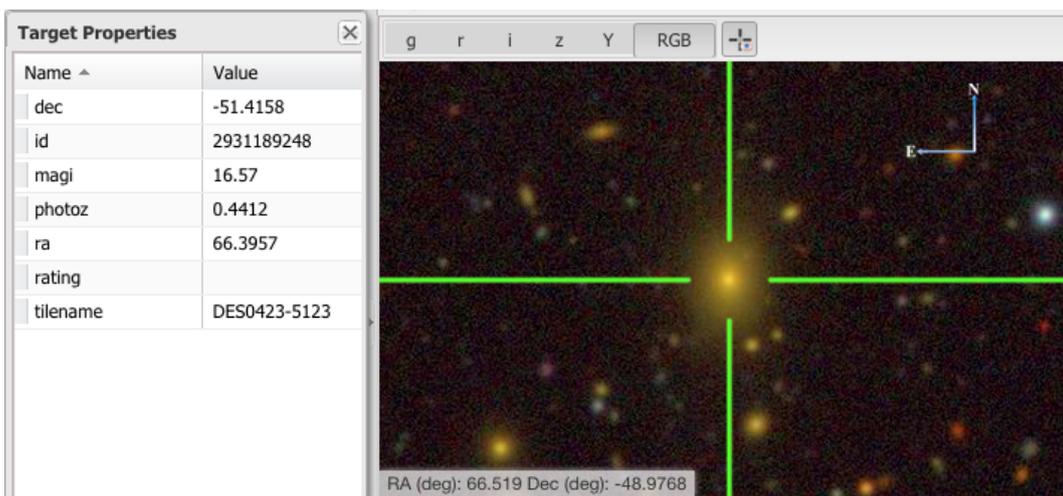


Figura 7.14: Exemplo de *photo-z* mal estimado.

### 7.4.1.3 Comparação de resultados

Para estimar, mesmo que grosseiramente, o impacto destes objetos na LF, fizemos um teste eliminando os objetos “*outliers*” no diagrama cor-cor. Notamos uma melhora signifi-

cativa nos dois intervalos de maior *redshift* ( $z > 0.8$ ), e uma melhora mais modesta nos intervalos intermediários ( $0.4 < z < 0.8$  – ver figura 7.15). O mesmo se observa nos gráficos de evolução dos parâmetros do ajuste (figura 7.16).

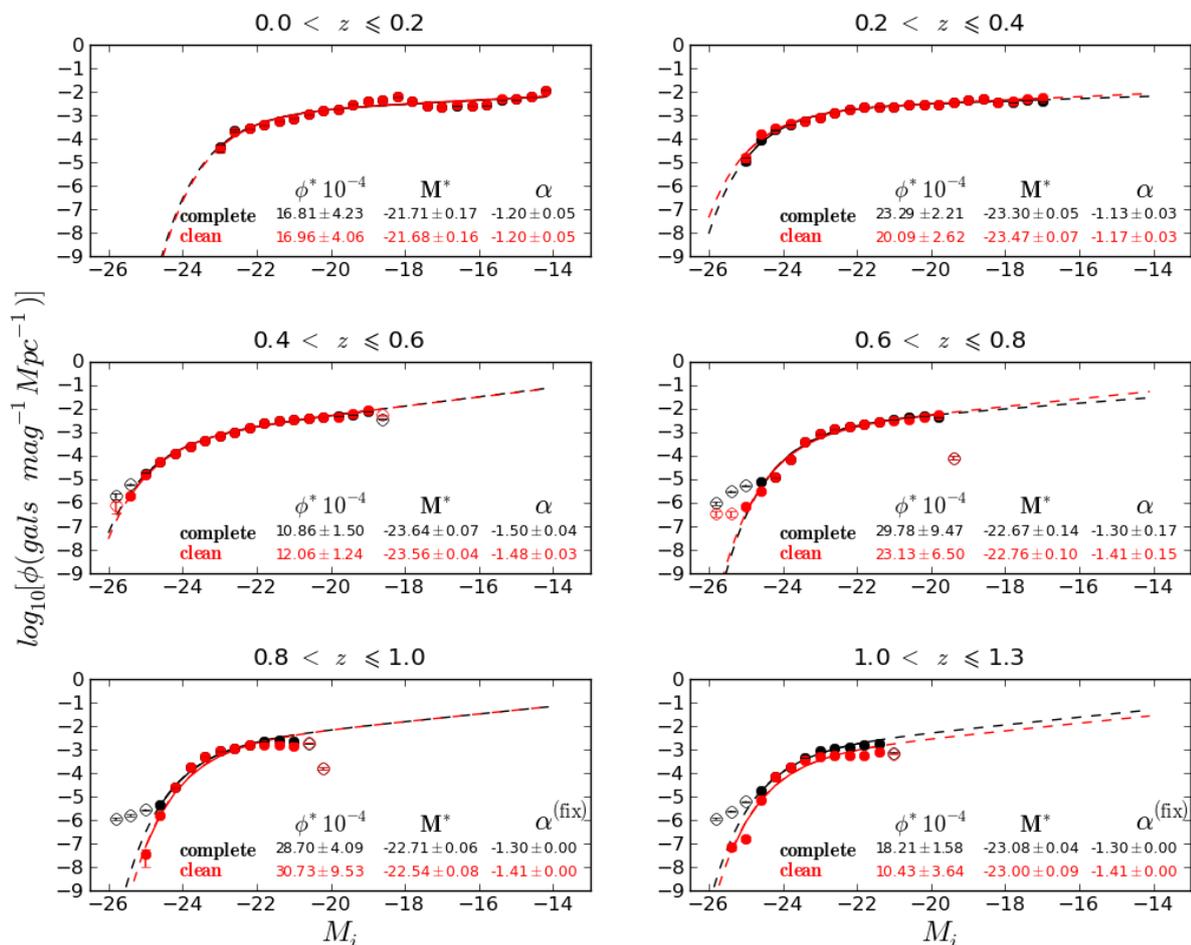
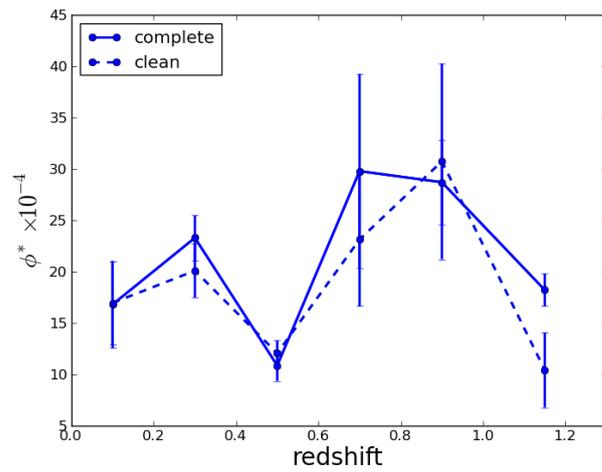
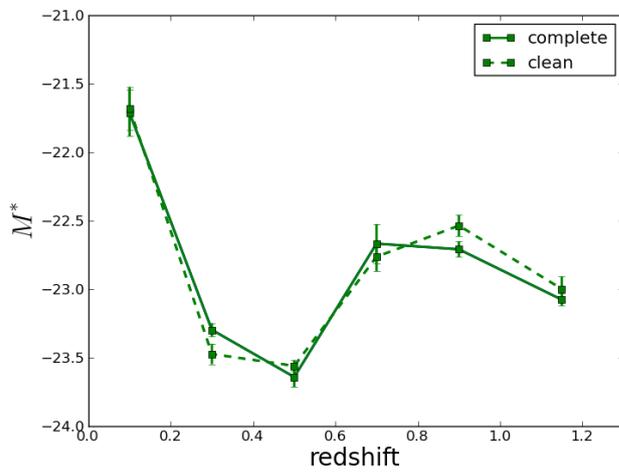


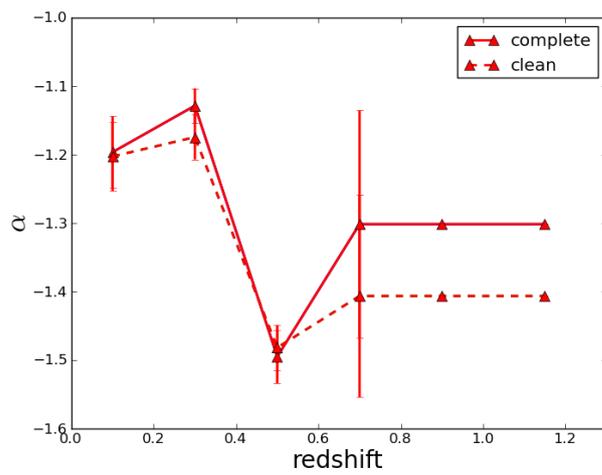
Figura 7.15: Função de luminosidade global em intervalos de *redshift*. Comparação entre os resultados da função de luminosidade antes e após a remoção dos objetos candidatos a contaminantes.



(a)  $\phi^*$



(b)  $M^*$



(c)  $\alpha$

Figura 7.16: Comparação da evolução dos parâmetros de Schechter.

A figura 7.15 mostra que a remoção de objetos com base no diagrama cor-cor melhora o ajuste, mas não resolve nos intervalos centrais, especialmente em  $0.6 < z < 0.8$ .

Comparando com resultados anteriores obtidos da literatura, assim como fizemos para a amostra BCC, reforçamos a ressalva de que os dados possuem calibração diferente – e de que em alguns casos nem se referem à mesma banda – voltamos então nossa atenção para o comportamento geral da evolução de  $M^*$  e  $\phi^*$  com o *redshift*.

Os resultados encontrados em Ilbert et al. (2005) [43] (na figura 7 do artigo) e Willmer et al. (2006) [75] (na figura 9 do artigo), mostram um decréscimo suave no brilho intrínseco e uma constância (com algum ruído) na densidade média de galáxias para todos os tipos espectrais juntos, com o passar do tempo cósmico (conforme o *redshift* decresce). O comportamento da evolução dos parâmetros que obtivemos mostra maior discrepância com os resultados citados justamente no intervalo de *redshift* onde não obtivemos êxito na eliminação da contaminação.

# Capítulo 8

## Conclusões e Perspectivas

Em preparação para o uso dos dados do levantamento DES para estudos sobre a evolução das galáxias realizamos implementação e testes com códigos de determinação de *redshift* fotométrico, e de estimativa de Função de Luminosidade. Para tal, usamos dados simulados (amostra BCC) e observacionais da fase de verificação científica do DES (amostra SVA1).

Com respeito aos testes de estimativas de *redshifts* fotométricos podemos concluir o seguinte:

- O uso do Portal Científico do DES torna-se uma vantagem, pois permite o encadeamento de *workflows* e *pipelines*, desde a identificação de novos dados liberados pelo DESDM, passando pela ingestão dos catálogos no *database* do LIneA, pela criação de amostras de treinamento e validação para os diversos códigos de *redshifts* fotométricos, suas execuções em catálogos fotométricos do levantamento, até a criação de VACs para uso em pipelines científicos.
- Com base em um conjunto de métricas e em requisitos científicos estabelecidos pela colaboração DES, os 13 códigos examinados, de uma forma geral, satisfazem as condições exigidas.
- Os códigos baseados em métodos empíricos apresentam melhor performance tanto na qualidade da estimativa de um valor único de *photo-z* quanto para a reconstrução da distribuição de  $N(z)$ .
- Códigos de ajustes a modelos tem melhor desempenho em regiões de alto *redshift*. Códigos empíricos são muito dependentes da boa representatividade da amostra pelo conjunto de treinamento, que é dificilmente satisfeita nessa faixa de  $z$ .
- O corte em  $\epsilon_{phot}$  aumenta  $\overline{\Delta z}$  para a maioria dos códigos avaliados.
- O corte em S/N reduz  $\overline{\Delta z}$  para a maioria dos códigos avaliados.
- O corte em  $\epsilon_{phot}$  aumenta  $\sigma_{68}$  para a maioria dos códigos avaliados.

- O corte em S/N diminui  $\sigma_{68}$  em todos os códigos avaliados. Uma relação sinal-ruído  $> 5$  é um bom compromisso.
- Baseado na métrica  $\sigma_{68}$ , o algoritmo com a melhor performance na estimativa de um valor único para o photo-z é o código empírico TPZ.
- A combinação entre a melhor acurácia obtida com códigos empíricos e a possibilidade de obtenção de outros resultados, além do  $z$  (como tipo espectral, extinção e correção-K), em métodos de ajuste a modelos mostrou-se a abordagem mais interessante para o estudo de GE através da LF.

Com respeito aos resultados obtidos para a função de luminosidade, reconhecemos que os mesmos são ainda preliminares para servirem de base para uma análise científica completa. Apesar de termos encontrado, para todas as amostras utilizadas, uma LF global com o formato próximo ao esperado e que se ajusta bem à parametrização da função de Schechter, a evolução dos parâmetros de ajuste ainda não é perfeitamente compatível com os resultados encontrados na literatura.

Este trabalho teve como o objetivo a definição de uma metodologia que viabilize o estudo de evolução de galáxias com os dados do DES nos próximos anos. Apesar de termos utilizado a LF como um exemplo de ferramenta para o estudo de GE, e de termos cumprido com todas as etapas, desde o tratamento dos dados fotométricos, até culminar na determinação da função, entendemos que ainda há muitos desafios a serem enfrentados na continuação deste trabalho no futuro. Acreditamos que o maior deles reside na qualidade do *photo-z* e é por isso que focamos nossos estudos justamente neste tema. Pouco adiantaria se concentrássemos esforços no aperfeiçoamento do cálculo da LF, por exemplo buscando métodos mais sofisticados que o método  $1/V_{max}$ , se tivéssemos que lidar com uma amostra com *redshifts* fotométricos de baixa qualidade.

O ferramental sendo desenvolvido, já permitiu algumas análises sobre os dados do levantamento DES. No entanto, ainda podemos ampliar e melhor definir os procedimentos e os códigos de análise. Assim, como perspectivas temos:

- Finalizar a implementação no Portal Científico do DES de alguns dos algoritmos de *photo-z* faltantes.
- Realizar novos testes de desempenho dos algoritmos de *photo-z* em amostras maiores e com mais dados espectroscópicos para treinamento e validação.
- Efetuar estimativas da LF na amostra do primeiro ano de dados do levantamento (Y1A1).
- Estimar LF para amostras separadas por classes espectrais distintas.
- Implementar algoritmo para estimativa da Função de Massa.

- Desenvolver algoritmo para busca de candidatas a galáxias maciças.

# Referências Bibliográficas

- [1] Abbott, T., Aldering, Annis, et al., 2005, “The Dark Energy Survey”, *arXiv*, (out.).
- [2] Annis, J., Bridle, S., Castander, F. J., et al., 2005, “Constraining Dark Energy with the Dark Energy Survey: Theoretical Challenges”, *ArXiv Astrophysics e-prints*, (out.).
- [3] Arnouts, S., Cristiani, S., Moscardini, L., et al., 1999, “Measuring and modelling the redshift evolution of clustering: the Hubble Deep Field North”, *MNRAS*, v. 310 (dez.), pp. 540–556. doi: 10.1046/j.1365-8711.1999.02978.x.
- [4] Banerji, M., Jouvel, S., Lin, H., et al., 2015, “Combining Dark Energy Survey Science Verification data with near-infrared data from the ESO VISTA Hemisphere Survey”, *MNRAS*, v. 446 (jan.), pp. 2523–2539. doi: 10.1093/mnras/stu2261.
- [5] Bell, E. F., McIntosh, D. H., Katz, N., et al., 2003, “The Optical and Near-Infrared Properties of Galaxies. I. Luminosity and Stellar Mass Functions”, *ApJS*, v. 149 (dez.), pp. 289–312. doi: 10.1086/378847.
- [6] Bender, R., Appenzeller, I., Böhm, A., et al., 2001, “The FORS Deep Field: Photometric Data and Photometric Redshifts”. In: Cristiani, S., Renzini, A., Williams, R. E. (Eds.), *Deep Fields*, p. 96. doi: 10.1007/10854354\_18.
- [7] Benítez, N., 2000, “Bayesian Photometric Redshift Estimation”, *ApJ*, v. 536 (jun.), pp. 571–583. doi: 10.1086/308947.
- [8] Benson, A. J., Bower, R. G., Frenk, C. S., et al., 2003, “What Shapes the Luminosity Function of Galaxies?” *ApJ*, v. 599 (dez.), pp. 38–49. doi: 10.1086/379160.
- [9] Bertin, E., 2006, “Automatic Astrometric and Photometric Calibration with SCAMP”. In: Gabriel, C., Arviset, C., Ponz, D., et al. (Eds.), *Astronomical Data Analysis Software and Systems XV*, v. 351, *Astronomical Society of the Pacific Conference Series*, p. 112, jul.
- [10] Bertin, E., 2011, “Automated Morphometry with SExtractor and PSFEx”. In: Evans, I. N., Accomazzi, A., Mink, D. J., et al. (Eds.), *Astronomical Data Analysis*

*Software and Systems XX*, v. 442, *Astronomical Society of the Pacific Conference Series*, p. 435, jul.

- [11] Bertin, E., Arnouts, S., 1996, “SExtractor: Software for source extraction.” *A&ASS*, v. 117 (jun.), pp. 393–404.
- [12] Bertin, E., Mellier, Y., Radovich, M., et al., 2002, “The TERAPIX Pipeline”. In: Bohlender, D. A., Durand, D., Handley, T. H. (Eds.), *Astronomical Data Analysis Software and Systems XI*, v. 281, *Astronomical Society of the Pacific Conference Series*, p. 228.
- [13] Binggeli, B., Sandage, A., Tammann, G. A., 1988, “The luminosity function of galaxies”, *ARAA*, v. 26, pp. 509–560. doi: 10.1146/annurev.aa.26.090188.002453.
- [14] Bonnett, C., 2015, “Using neural networks to estimate redshift distributions. An application to CFHTLenS”, *MNRAS*, v. 449 (maio), pp. 1043–1056. doi: 10.1093/mnras/stv230.
- [15] Brammer, G. B., van Dokkum, P. G., Coppi, P., 2008, “EAZY: A Fast, Public Photometric Redshift Code”, *ApJ*, v. 686 (out.), pp. 1503–1513. doi: 10.1086/591786.
- [16] Bruzual A., G., Charlot, S., 1993, “Spectral evolution of stellar populations using isochrone synthesis”, *ApJ*, v. 405 (mar.), pp. 538–553. doi: 10.1086/172385.
- [17] Busha, M. T., Wechsler, R. H., Becker, M. R., et al., 2013, “Catalog Production for the DES Blind Cosmology Challenge”. In: *American Astronomical Society Meeting Abstracts 221*, v. 221, *American Astronomical Society Meeting Abstracts*, p. 341.07, jan.
- [18] Caputi, K. I., Cirasuolo, M., Dunlop, J. S., et al., 2011, “The stellar mass function of the most-massive galaxies at  $3 < z < 5$  in the UKIDSS Ultra Deep Survey”, *MNRAS*, v. 413 (maio), pp. 162–176. doi: 10.1111/j.1365-2966.2010.18118.x.
- [19] Carrasco Kind, M., Brunner, R. J., 2013, “TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests”, *MNRAS*, v. 432 (jun.), pp. 1483–1501. doi: 10.1093/mnras/stt574.
- [20] Coe, D., Benítez, N., Sánchez, S. F., et al., 2006, “Galaxies in the Hubble Ultra Deep Field. I. Detection, Multiband Photometry, Photometric Redshifts, and Morphology”, *AJ*, v. 132 (ago.), pp. 926–959. doi: 10.1086/505530.
- [21] Cole, S., Norberg, P., Baugh, C. M., et al., 2001, “The 2dF galaxy redshift survey: near-infrared galaxy luminosity functions”, *MNRAS*, v. 326 (set.), pp. 255–273. doi: 10.1046/j.1365-8711.2001.04591.x.

- [22] Coleman, G. D., Wu, C.-C., Weedman, D. W., 1980, “Colors and magnitudes predicted for high redshift galaxies”, *ApJS*, v. 43 (jul.), pp. 393–416. doi: 10.1086/190674.
- [23] Colless, M., 1998, “Early Results from the 2dF Galaxy Redshift Survey”. In: Colombi, S., Mellier, Y., Raban, B. (Eds.), *Wide Field Surveys in Cosmology*, p. 77.
- [24] Collister, A. A., Lahav, O., 2004, “ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks”, *PASP*, v. 116 (abr.), pp. 345–351. doi: 10.1086/383254.
- [25] Conroy, C., Gunn, J. E., White, M., 2009, “The Propagation of Uncertainties in Stellar Population Synthesis Modeling. I. The Relevance of Uncertain Aspects of Stellar Evolution and the Initial Mass Function to the Derived Physical Properties of Galaxies”, *ApJ*, v. 699 (jul.), pp. 486–506. doi: 10.1088/0004-637X/699/1/486.
- [26] Cooper, M. C., Newman, J. A., Madgwick, D. S., et al., 2005, “Measuring Galaxy Environments with Deep Redshift Surveys”, *ApJ*, v. 634 (dez.), pp. 833–848. doi: 10.1086/432868.
- [27] Coupon, J., Ilbert, O., Kilbinger, M., et al., 2009, “Photometric redshifts for the CFHTLS T0004 deep and wide fields”, *A&A*, v. 500 (jun.), pp. 981–998. doi: 10.1051/0004-6361/200811413.
- [28] Cunha, C. E., Lima, M., Oyaizu, H., et al., 2009, “Estimating the redshift distribution of photometric galaxy samples - II. Applications and tests of a new method”, *MNRAS*, v. 396 (jul.), pp. 2379–2398. doi: 10.1111/j.1365-2966.2009.14908.x.
- [29] Davies, L. J. M., Maraston, C., Thomas, D., et al., 2013, “Detecting massive galaxies at high redshift using the Dark Energy Survey”, *MNRAS*, v. 434 (set.), pp. 296–312. doi: 10.1093/mnras/stt1018.
- [30] Desai, S., Armstrong, R., Mohr, J. J., et al., 2012, “The Blanco Cosmology Survey: Data Acquisition, Processing, Calibration, Quality Diagnostics, and Data Release”, *ApJ*, 757:83. doi: 10.1088/0004-637X/757/1/83.
- [31] Dickinson, M., Papovich, C., Ferguson, H. C., et al., 2003, “The Evolution of the Global Stellar Mass Density at  $0 < z < 3$ ”, *ApJ*, v. 587 (abr.), pp. 25–40. doi: 10.1086/368111.
- [32] Drory, N., Bundy, K., Leauthaud, A., et al., 2009, “The Bimodal Galaxy Stellar Mass Function in the COSMOS Survey to  $z \sim 1$ : A Steep Faint End and a New Galaxy Dichotomy”, *ApJ*, v. 707 (dez.), pp. 1595–1609. doi: 10.1088/0004-637X/707/2/1595.
- [33] Eggen, O. J., Lynden-Bell, D., Sandage, A. R., 1962, “Evidence from the motions of old stars that the Galaxy collapsed.” *ApJ*, v. 136 (nov.), pp. 748. doi: 10.1086/147433.

- [34] Elsner, F., Feulner, G., Hopp, U., 2008, “The impact of Spitzer infrared data on stellar mass estimates - and a revised galaxy stellar mass function at  $0 < z < 5$ ”, *A&A*, v. 477 (jan.), pp. 503–512. doi: 10.1051/0004-6361:20078343.
- [35] Feldmann, R., Carollo, C. M., Porciani, C., et al., 2006, “The Zurich Extragalactic Bayesian Redshift Analyzer and its first application: COSMOS”, *MNRAS*, v. 372 (out.), pp. 565–577. doi: 10.1111/j.1365-2966.2006.10930.x.
- [36] Flaugher, B., 2005, “The Dark Energy Survey”, *International Journal of Modern Physics A*, v. 20, pp. 3121–3123.
- [37] Flaugher, B., Diehl, H. T., Honscheid, K., et al., 2015, “The Dark Energy Camera”, *ArXiv e-prints*, (abr.).
- [38] Gerdes, D. W., Sypniewski, A. J., McKay, T. A., et al., 2010, “ArborZ: Photometric Redshifts Using Boosted Decision Trees”, *ApJ*, v. 715 (jun.), pp. 823–832. doi: 10.1088/0004-637X/715/2/823.
- [39] Graff, P., Feroz, F., Hobson, M. P., et al., 2014, “SKYNET: an efficient and robust neural network training tool for machine learning in astronomy”, *MNRAS*, v. 441 (jun.), pp. 1741–1759. doi: 10.1093/mnras/stu642.
- [40] Greisen, E. W., 2000, “Data Acquisition by Switching”. In: Manset, N., Veillet, C., Crabtree, D. (Eds.), *Astronomical Data Analysis Software and Systems IX*, v. 216, *Astronomical Society of the Pacific Conference Series*, p. 514.
- [41] Hildebrandt, H., Arnouts, S., Capak, P., et al., 2010, “PHAT: PHoto-z Accuracy Testing”, *A&A*, 523:A31. doi: 10.1051/0004-6361/201014885.
- [42] Huchra, J., Davis, M., Latham, D., et al., 1983, “A survey of galaxy redshifts. IV - The data”, *ApJs*, v. 52 (jun.), pp. 89–119. doi: 10.1086/190860.
- [43] Ilbert, O., Tresse, L., Zucca, E., et al., 2005, “The VIMOS-VLT deep survey. Evolution of the galaxy luminosity function up to  $z = 2$  in first epoch data”, *A&A*, v. 439 (set.), pp. 863–876. doi: 10.1051/0004-6361:20041961.
- [44] Ilbert, O., Arnouts, S., McCracken, H. J., et al., 2006, “Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey”, *A&A*, v. 457 (out.), pp. 841–856. doi: 10.1051/0004-6361:20065138.
- [45] Ivezić, Z., Tyson, A. J., Strauss, M. A., et al., 2006, “LSST Survey Strategy”. In: *American Astronomical Society Meeting Abstracts*, v. 38, *Bulletin of the American Astronomical Society*, p. 1017, dez.

- [46] Kinney, A. L., Calzetti, D., Bohlin, R. C., et al., 1996, “Template Ultraviolet to Near-Infrared Spectra of Star-forming Galaxies and Their Application to K-Corrections”, *ApJ*, v. 467 (ago.), pp. 38. doi: 10.1086/177583.
- [47] Lilly, S. J., Tresse, L., Hammer, F., et al., 1995, “The Canada-France Redshift Survey. VI. Evolution of the Galaxy Luminosity Function to  $Z$  approximately 1”, *ApJ*, v. 455 (dez.), pp. 108. doi: 10.1086/176560.
- [48] Lilly, S. J., Le Fèvre, O., Renzini, A., et al., 2007, “zCOSMOS: A Large VLT/VIMOS Redshift Survey Covering  $0 < z < 3$  in the COSMOS Field”, *ApJs*, v. 172 (set.), pp. 70–85. doi: 10.1086/516589.
- [49] Lima, M., Cunha, C. E., Oyaizu, H., et al., 2008, “Estimating the redshift distribution of photometric galaxy samples”, *MNRAS*, v. 390 (out.), pp. 118–130. doi: 10.1111/j.1365-2966.2008.13510.x.
- [50] Lin, H., Kirshner, R. P., Shectman, S. A., et al., 1996, “The Luminosity Function of Galaxies in the Las Campanas Redshift Survey”, *ApJ*, v. 464 (jun.), pp. 60. doi: 10.1086/177300.
- [51] Madau, P., Ferguson, H. C., Dickinson, M. E., et al., 1996, “High-redshift galaxies in the Hubble Deep Field: colour selection and star formation history to  $z \sim 4$ ”, *MNRAS*, v. 283 (dez.), pp. 1388–1404.
- [52] Mancini, C., Matute, I., Cimatti, A., et al., 2009, “Searching for massive galaxies at  $z \sim 3.5$  in GOODS-North”, *A&A*, v. 500 (jun.), pp. 705–723. doi: 10.1051/0004-6361/200810630.
- [53] Maraston, C., Daddi, E., Renzini, A., et al., 2006, “Evidence for TP-AGB Stars in High-Redshift Galaxies, and Their Effect on Deriving Stellar Population Parameters”, *ApJ*, v. 652 (nov.), pp. 85–96. doi: 10.1086/508143.
- [54] Marchesini, D., Whitaker, K. E., Brammer, G., et al., 2010, “The Most Massive Galaxies at  $3.0 < z < 4.0$  in the Newfirm Medium-band Survey: Properties and Improved Constraints on the Stellar Mass Function”, *ApJ*, v. 725 (dez.), pp. 1277–1295. doi: 10.1088/0004-637X/725/1/1277.
- [55] McMahon, R. G., Banerji, M., Gonzalez, E., et al., 2013, “First Scientific Results from the VISTA Hemisphere Survey (VHS)”, *The Messenger*, v. 154 (dez.), pp. 35–37.
- [56] Mohr, J. J., Armstrong, R., Bertin, E., et al., 2012, “The Dark Energy Survey data processing and calibration system”. In: *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, v. 8451, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 0, set. doi: 10.1117/12.926785.

- [57] Mortlock, A., Conselice, C. J., Bluck, A. F. L., et al., 2011, “A deep probe of the galaxy stellar mass functions at  $z$  1–3 with the GOODS NICMOS Survey”, *MNRAS*, v. 413 (jun.), pp. 2845–2859. doi: 10.1111/j.1365-2966.2011.18357.x.
- [58] Oke, J. B., Sandage, A., 1968, “Energy Distributions, K Corrections, and the Stebbins-Whitford Effect for Giant Elliptical Galaxies”, *ApJ*, v. 154 (out.), pp. 21. doi: 10.1086/149737.
- [59] Oyaizu, H., Lima, M., Cunha, C. E., et al., 2008, “Photometric Redshift Error Estimators”, *ApJ*, v. 689 (dez.), pp. 709–720. doi: 10.1086/592591.
- [60] Papovich, C., Moustakas, L. A., Dickinson, M., et al., 2006, “Spitzer Observations of Massive, Red Galaxies at High Redshift”, *ApJ*, v. 640 (mar.), pp. 92–113. doi: 10.1086/499915.
- [61] Peacock, J., 1999, “Book Review: Cosmological physics / Cambridge U Press, 1999”, *The Observatory*, v. 119 (out.), pp. 296.
- [62] Pence, W., 1976, “K-corrections for galaxies of different morphological types”, *ApJ*, v. 203 (jan.), pp. 39–51.
- [63] Pozzetti, L., Bolzonella, M., Zucca, E., et al., 2010, “zCOSMOS - 10k-bright spectroscopic sample. The bimodality in the galaxy stellar mass function: exploring its evolution with redshift”, *A&A*, 523:A13. doi: 10.1051/0004-6361/200913020.
- [64] Press, W. H., Schechter, P., 1974, “Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation”, *ApJ*, v. 187 (fev.), pp. 425–438. doi: 10.1086/152650.
- [65] Prevot, M. L., Lequeux, J., Prevot, L., et al., 1984, “The typical interstellar extinction in the Small Magellanic Cloud”, *A&A*, v. 132 (mar.), pp. 389–392.
- [66] Ramos, B. H. F., Pellegrini, P. S., Benoist, C., et al., 2011, “Evolution of Galaxy Luminosity Function Using Photometric Redshifts”, *AJ*, 142:41. doi: 10.1088/0004-6256/142/2/41.
- [67] Ruhl, J., Ade, P. A. R., Carlstrom, J. E., et al., 2004, “The South Pole Telescope”. In: Bradford, C. M., Ade, P. A. R., Aguirre, J. E., et al. (Eds.), *Z-Spec: a broadband millimeter-wave grating spectrometer: design, construction, and first cryogenic measurements*, v. 5498, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pp. 11–29, out. doi: 10.1117/12.552473.
- [68] Sánchez, C., Carrasco Kind, M., Lin, H., et al., 2014, “Photometric redshift analysis in the Dark Energy Survey Science Verification data”, *MNRAS*, v. 445 (dez.), pp. 1482–1506. doi: 10.1093/mnras/stu1836.

- [69] Schechter, P., 1976, “An analytic expression for the luminosity function for galaxies.” *ApJ*, v. 203 (jan.), pp. 297–306. doi: 10.1086/154079.
- [70] Schlegel, D. J., Finkbeiner, D. P., Davis, M., 1998, “Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds”, *ApJ*, v. 500 (jun.), pp. 525–553. doi: 10.1086/305772.
- [71] Skrutskie, M. F., Cutri, R. M., Stiening, R., et al., 2006, “The Two Micron All Sky Survey (2MASS)”, *AJ*, v. 131 (feb.), pp. 1163–1183. doi: 10.1086/498708.
- [72] Tipping, m. E., 2001, “Sparse Bayesian Learning and the Relevance Vector Machine”, *MLR*, v. 1 (jun.), pp. 211–214.
- [73] Wake, D. A., Nichol, R. C., Eisenstein, D. J., et al., 2006, “The 2df SDSS LRG and QSO survey: evolution of the luminosity function of luminous red galaxies to  $z = 0.6$ ”, *MNRAS*, v. 372 (out.), pp. 537–550. doi: 10.1111/j.1365-2966.2006.10831.x.
- [74] White, S. D. M., Rees, M. J., 1978, “Core condensation in heavy halos - A two-stage theory for galaxy formation and clustering”, *MNRAS*, v. 183 (maio), pp. 341–358.
- [75] Willmer, C. N. A., Faber, S. M., Koo, D. C., et al., 2006, “The Deep Evolutionary Exploratory Probe 2 Galaxy Redshift Survey: The Galaxy Luminosity Function to  $z \sim 1$ ”, *ApJ*, v. 647 (ago.), pp. 853–873. doi: 10.1086/505455.
- [76] Wolf, C., Meisenheimer, K., Rix, H.-W., et al., 2003, “The COMBO-17 survey: Evolution of the galaxy luminosity function from 25 000 galaxies with  $0.2 < z < 1.2$ ”, *A&A*, v. 401 (abr.), pp. 73–98. doi: 10.1051/0004-6361:20021513.
- [77] Yan, H., Dickinson, M., Stern, D., et al., 2006, “GOODS spitzer IRAC observations of high-  $z$  galaxies - Implications for reionization”, *NAR*, v. 50 (mar.), pp. 127–133. doi: 10.1016/j.newar.2005.11.026.
- [78] York, D. G., Adelman, J., Anderson, Jr., J. E., et al., 2000, “The Sloan Digital Sky Survey: Technical Summary”, *AJ*, v. 120 (set.), pp. 1579–1587. doi: 10.1086/301513.
- [79] Yuan, F., Lidman, C., Davis, T. M., et al., 2014, “OzDES Spectroscopic Classification of 5 Supernovae From DES”, *The Astronomer’s Telegram*, v. 6520 (out.), pp. 1.
- [80] Yuan, F., Lidman, C., Davis, T. M., et al., 2015, “OzDES multi-fibre spectroscopy for the Dark Energy Survey: first-year operation and results”, *ArXiv e-prints*, (abr.).
- [81] Zucca, E., Bardelli, S., Bolzonella, M., et al., 2009, “The zCOSMOS survey: the role of the environment in the evolution of the luminosity function of different galaxy types”, *A&A*, v. 508 (dez.), pp. 1217–1234. doi: 10.1051/0004-6361/200912665.

# Apêndice A

## O Portal Científico do DES

Neste apêndice abordaremos com mais detalhes as questões técnicas que envolvem os *pipelines* do PCDES mencionados ao longo do texto. A figura A.1<sup>1</sup> nos dá uma visão global da estrutura de funcionamento do portal, na época da escrita desta dissertação. Cabe ressaltar que esta não é uma estrutura estática. Ela está em constante desenvolvimento, sofrendo melhorias e adaptações conforme as demandas do levantamento. Conforme descrito no capítulo 3, o E2E para a criação de VACs é dividido em uma série de etapas (figura A.2), sendo aquelas relacionadas à produção dos *photo-z*'s, as mais relevantes para este trabalho.

Métodos empíricos e semi-empíricos de cálculo de *photo-z* realizam uma etapa inicial onde um conjunto de galáxias com *redshift* conhecido é utilizado para treinamento. Este procedimento gera arquivos, que podem variar significativamente em formato e conteúdo de um código para outro, e que carregam a informação do treinamento para o passo seguinte. Uma vez que esses arquivos podem ser reutilizados, a repetição da primeira etapa se torna desnecessária ela pode ser feita separadamente. A criação do conjunto de treinamento também é um procedimento que não necessita ser reproduzido a cada rodada de um código de *photo-z*. Por isso, dois *pipelines* adicionais foram criados para funcionar em paralelo, provendo os chamados *training sets* e os *training files*, arquivos resultantes do processo de treinamento.

---

<sup>1</sup>Fonte: <http://www.linea.gov.br/2014/08/portal-cientifico-desenvolvido-pelo-linea/>

# Science Portal Grand View

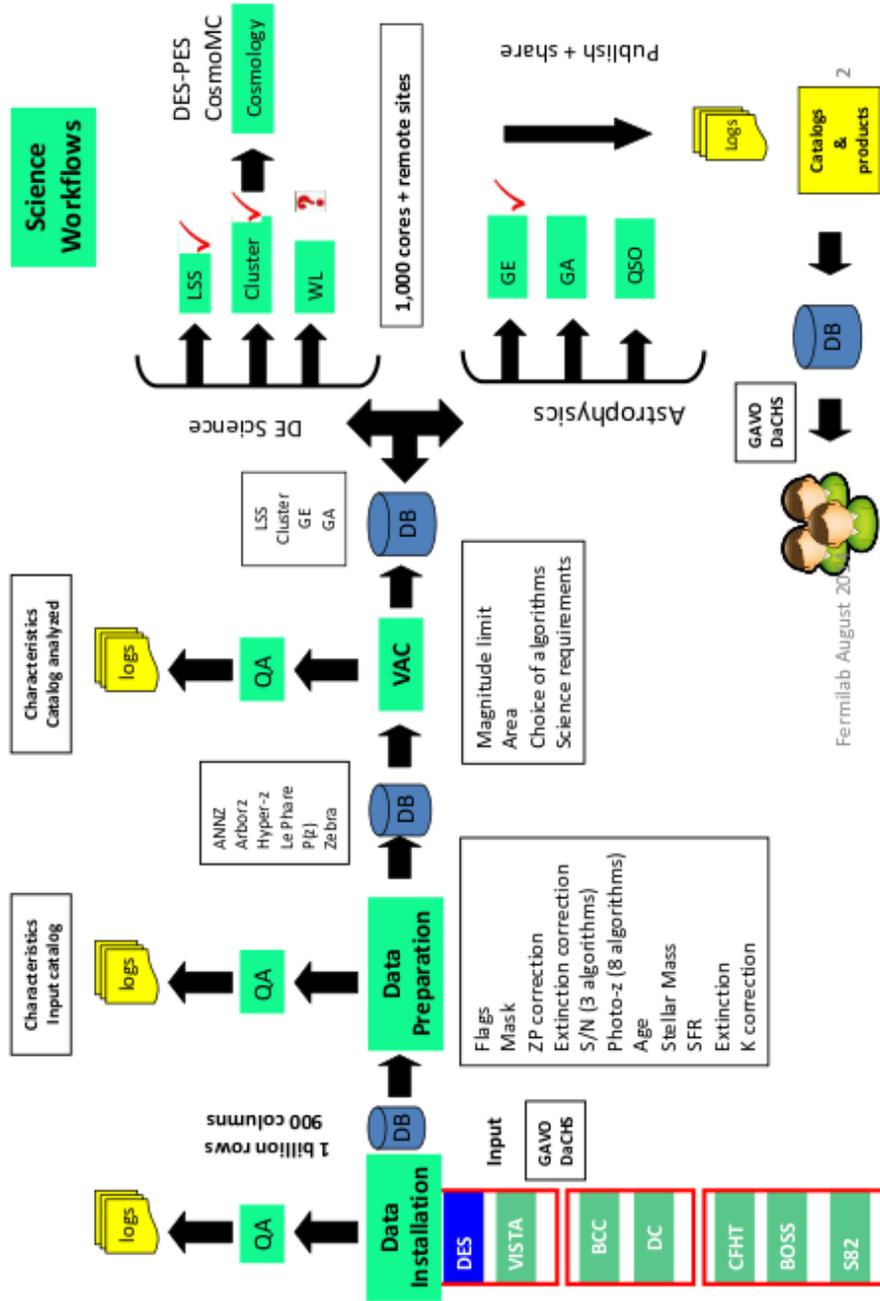


Figura A.1: Visão geral do funcionamento do portal científico.

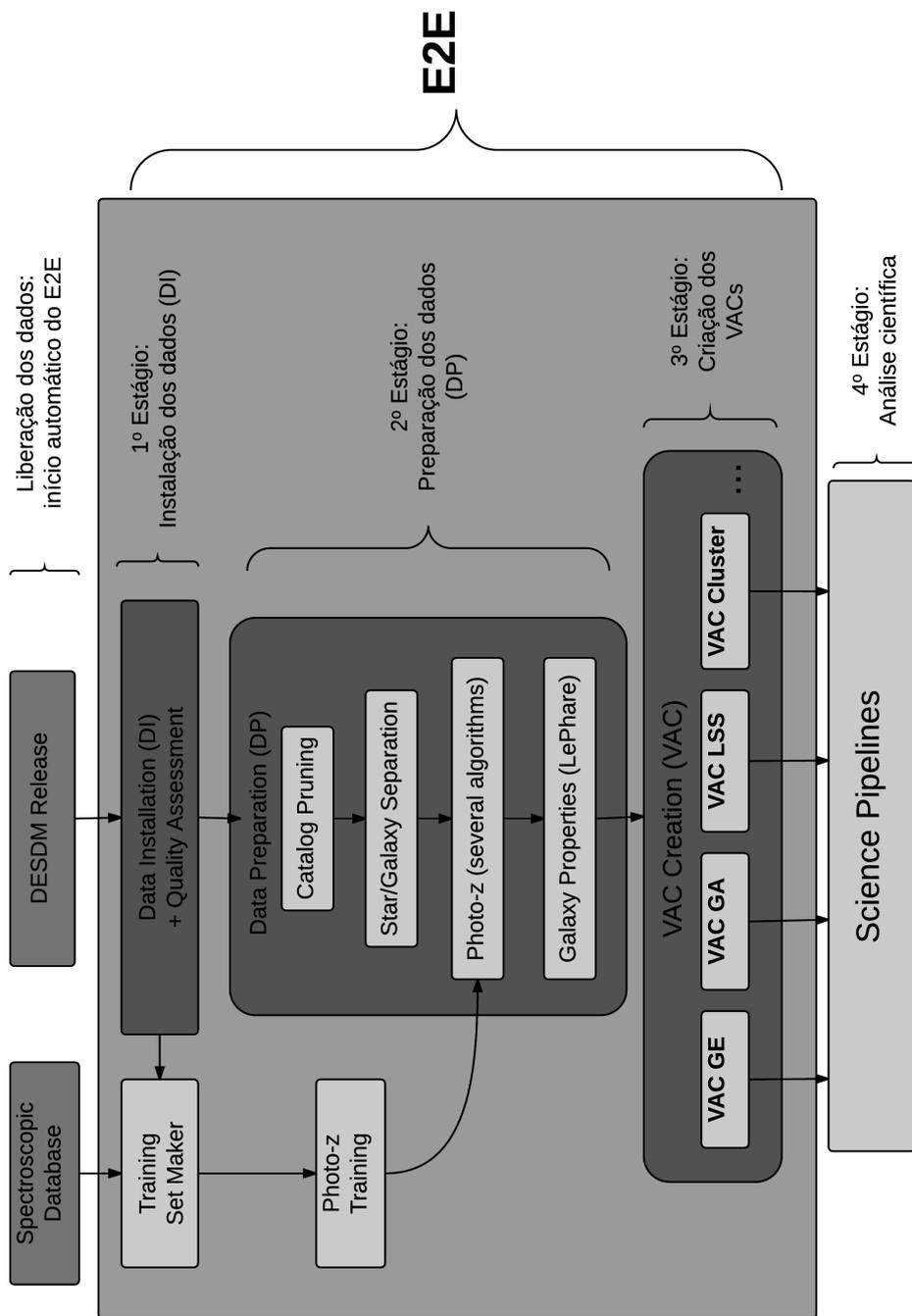


Figura A.2: Esquema geral do E2E para a criação de VACs.

## A.1 O *Pipeline training\_set\_maker*

A partir da combinação entre as medidas de posição dos dados do DES e de outros levantamentos (já previamente armazenados em um banco de dados) o *pipeline training\_set\_maker* cria conjuntos de treinamento que são entregues ao *pipeline* seguinte (detalhes na Figura A.3). Estes conjuntos são arquivos no formato FITS, que representam subconjuntos de uma determinada liberação oficial de dados do DES, adicionado aos *redshifts* disponíveis para a região do céu em questão e seus erros associados e/ou parâmetros de qualidade (quando disponíveis). O *pipeline* é rodado sempre que há uma nova liberação de dados, pois nesta ocasião a área coberta pela fotometria e a profundidade (devido à coadição de imagens) tendem a aumentar.

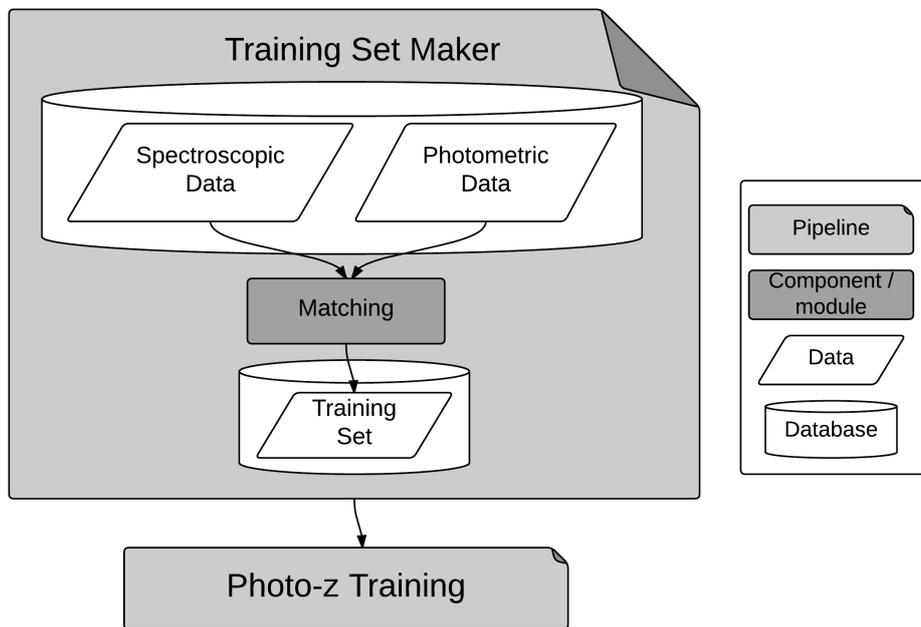


Figura A.3: *Pipeline training\_set\_maker*: Combinação entre dados fotométricos do DES e dados de *redshifts* espectroscópicos públicos de outros levantamentos para a construção de conjuntos de treinamento para códigos de *photo-z*

Além da associação entre a dados oriundos da fotometria e da espectroscopia, o *pipeline* também fornece uma "etiqueta" de identificação de qualidade a cada galáxia que codifica, não somente em quantas, mas em quais bandas a fotometria pode ser considerada confiável<sup>2</sup>.

O resultado do *pipeline* mostra as fontes dos *redshifts* espectroscópicos e a quantidade (absoluta e em porcentagem) de objetos obtidos de cada uma, além de um gráfico que mostra a posição dos objetos no céu e um histograma da distribuição de *redshifts* obtida na rodada.

<sup>2</sup>Razão sinal-ruído acima de 10 – requisito científico estabelecido pelo levantamento

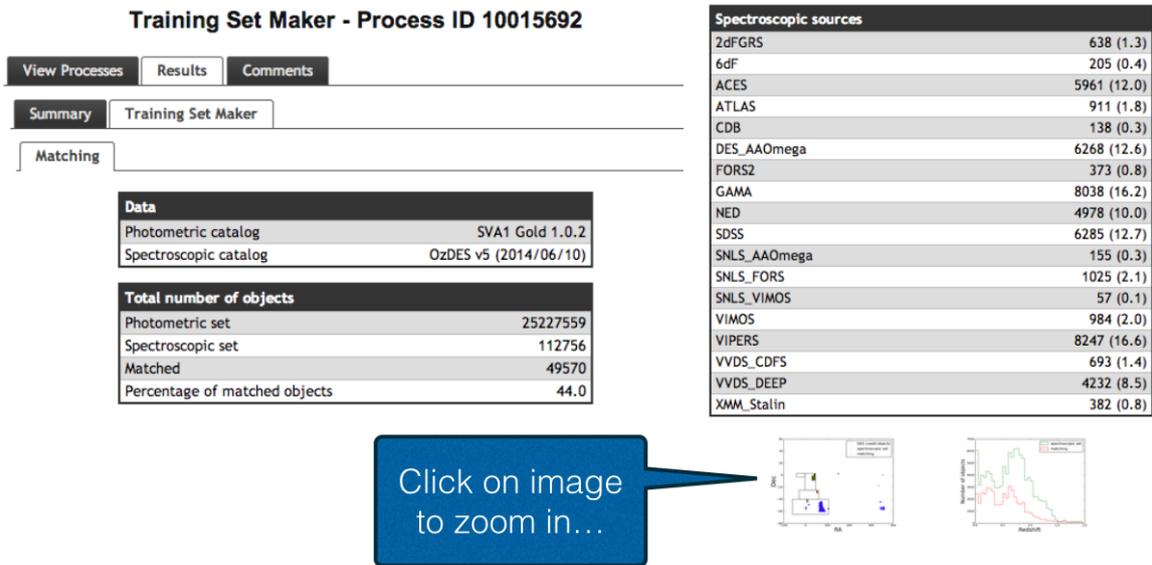


Figura A.4: Captura de tela do resultado de uma rodada do *pipeline training\_set\_maker*. A tabela à direita mostra a quantidade de objetos obtidos em cada fonte dados espectroscópicos (valor em porcentagem entre parênteses)

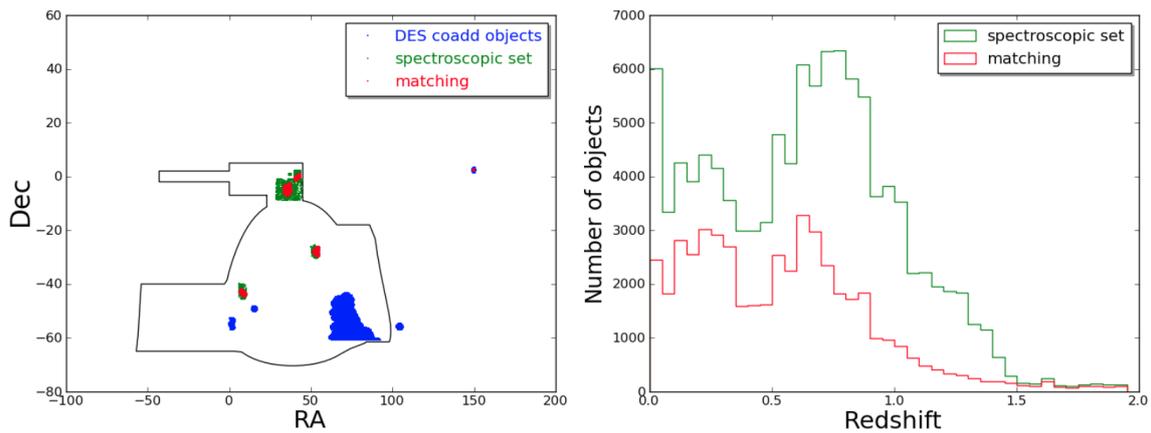


Figura A.5: À esquerda: Posições em graus (coordenadas equatoriais) dos objetos dos conjuntos espectroscópico, fotométrico e de treinamento gerado, em coordenadas equatoriais. À direita: distribuição de *redshifts* do conjunto espectroscópico e do conjunto de treinamento (parte do primeiro que teve combinação com a fotometria bem sucedida).

## A.2 O *Pipeline photoz\_training*

O *pipeline photoz\_training* tem como principal objetivo a realização da etapa de treinamento dos métodos empíricos de *photo-z* e a entrega deste resultado para o estágio *Data Preparation*, onde os *photo-z's* de todas as galáxias serão calculados.

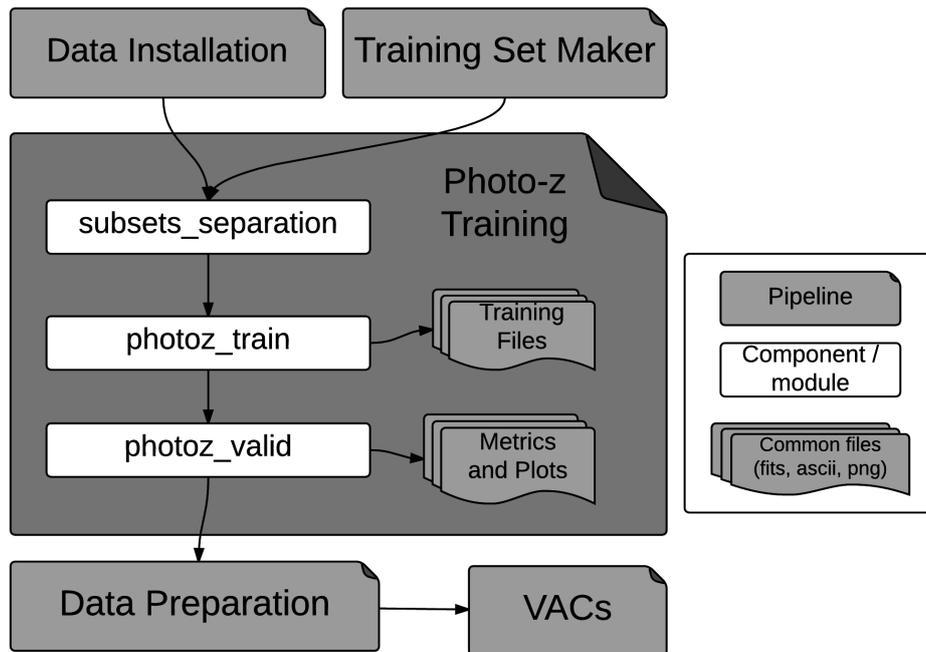


Figura A.6: Esquema do *workflow* do *pipeline* `photoz_training`. Os dados de entrada são oriundos do *pipeline* *Training Set Maker* (o conjunto de treinamento) e do *pipeline* *Data Installation* (o conjunto fotométrico de referência, para comparação e utilização, opcional, de pesos). Os produtos gerados são os arquivos com os resultados do treinamento, métricas e gráficos de diagnóstico.

### A.2.1 Etapa pré-treinamento: componente `subsets_separation`

- Separação dos dados em dois subconjuntos, um para o ao treinamento propriamente dito e outro para a validação do métodos e dos conjuntos de treinamento.
- Caracterização dos dados através de tabelas e gráficos de diagnóstico, mostrando características como as distribuições de magnitudes, erros, cores, redshift, relação entre razão sinal-ruído e os outros observáveis, etc.
- Utilização (opcional) do método de pesos para refazer os conjuntos de treinamento e validação imitando as características fotométricas do conjunto fotométrico de referência (os dados que serão usados no *pipeline* seguinte).

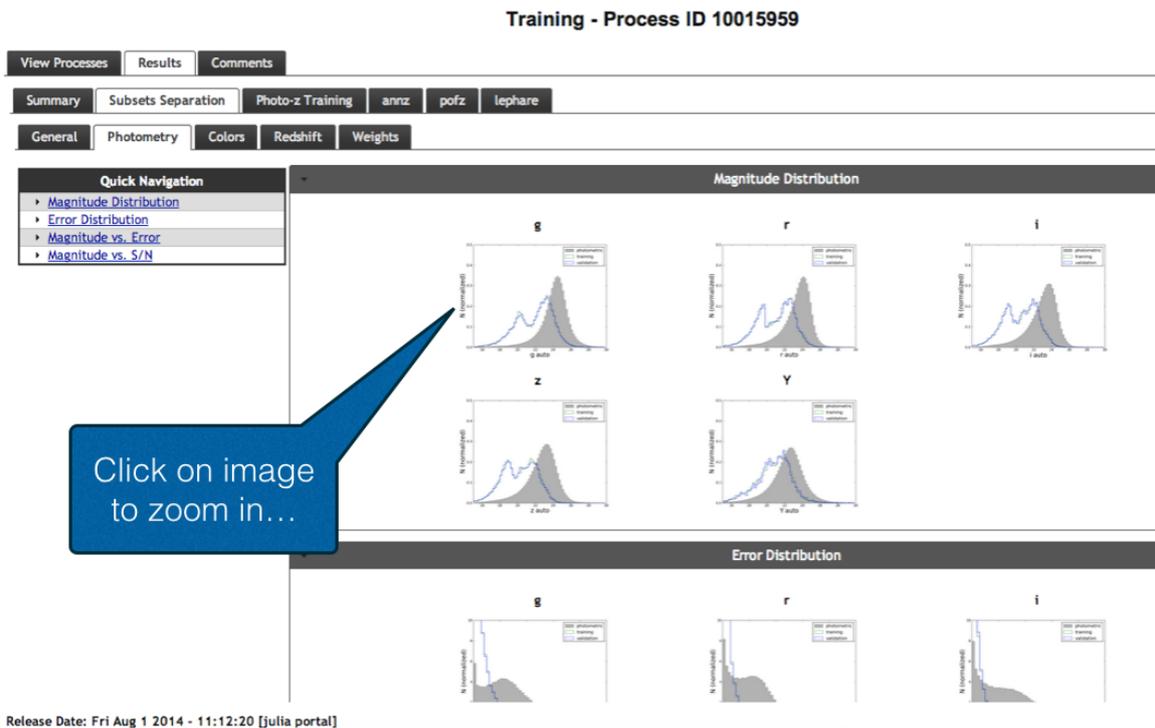


Figura A.7: Captura de Tela - *Product Log* do componente `subsets_separation`

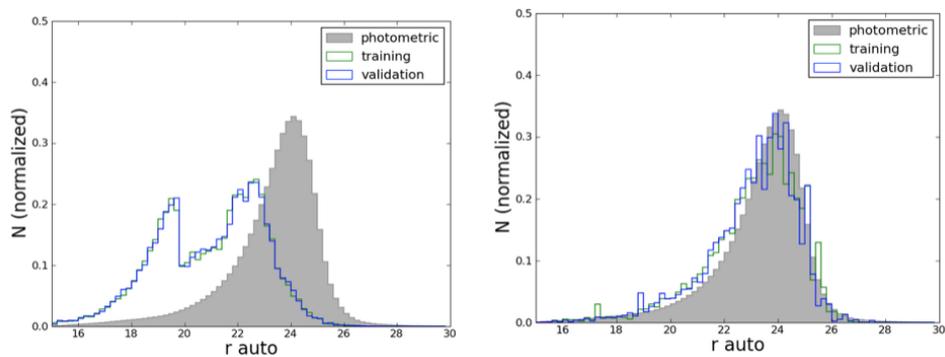


Figura A.8: Exemplo de gráfico de diagnóstico gerado pelo componente `subsets_separation`– Distribuições de magnitude na banda  $r$ , para dois casos: à esquerda sem e à direita com a utilização do método de pesos para tornar o conjunto de treinamento mais representativo do conjunto fotométrico a ser estudado.

## A.2.2 Etapa principal: componente `photoz_train`

- Realização do treinamento para os códigos empíricos e do procedimento de calibração das magnitudes teóricas do código LePHARE.
- Disponibilização dos arquivos com os resultados do treinamento para a próxima etapa e para o próximo *pipeline*.

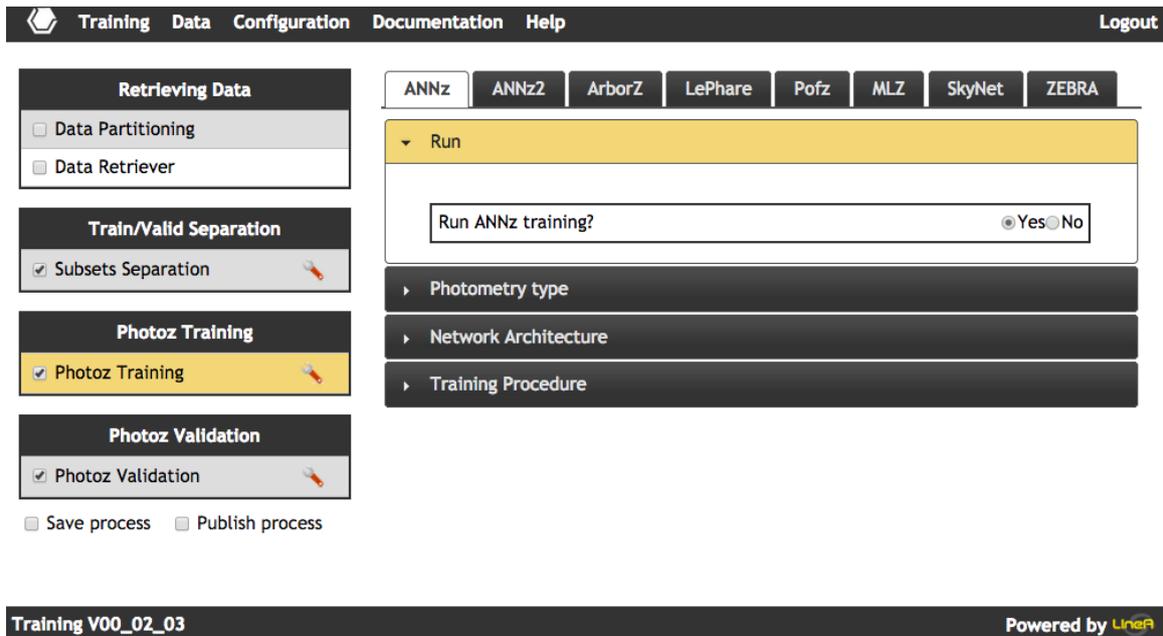


Figura A.9: Tela de início do *pipeline* `photoz_training`. Em cada aba o usuário do portal pode escolher os parâmetros de configuração desejados.

### A.2.3 Etapa pós-treinamento: componente `photo-z_valid`

- Cálculo do *redshift* fotométrico dos objetos do subconjunto de validação.
- Cálculo das métricas que definem a qualidade do *photo-z*, estabelecidas nas diretrizes do levantamento DES.
- Gráficos de comparação dos *redshifts* espectroscópico (valor esperado) e fotométrico (valos obtido), distribuição de erros e comportamento das métricas em intervalos de *redshift*.

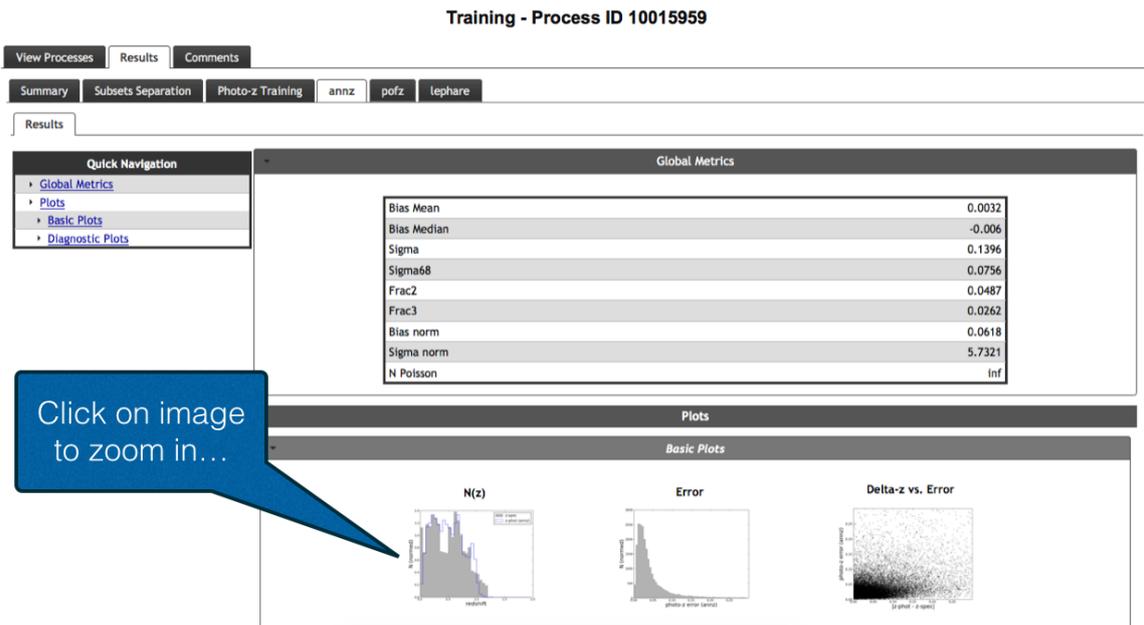


Figura A.10: Captura de Tela - *Product Log* do componente `photoz_valid`

Sendo assim, o *pipeline* `photoz_training` possui uma utilidade que vai além do elo na cadeia de produção dos VACs. A etapa de validação possibilita a avaliação da performance e, conseqüentemente, a comparação de métodos e códigos, gerando resultados como os que foram amplamente discutidos em Sánchez et al. (2014) [68]. Os códigos disponíveis atualmente são: ANNz, ANNz2, ArborZ, LePhare, Pofz, TPZ (dentro do pacote MLZ), SkyNet e DNF. Código em fase de implementação: ZEBRA.

# Apêndice B

## Configurações das Rodadas

Tabela B.1: TPZ

---

Inputs	<i>spec-z</i> , magnitudes e erros <i>grizY</i> , cores <i>g-r</i> , <i>r-i</i> , <i>i-z</i> , <i>z-Y</i>
Outputs	média e pico da PDF( <i>z</i> ), erro em cada estimativa
Configurações	Método: regressão Número de árvores: 20 Número de intervalos de <i>redshift</i> : 50 $z_{min} = 0.001$ ; $z_{max} = 2.0$
Tempo gasto	$\sim 8h/100$ mil objetos

---

Tabela B.2: LePHARE

---

Inputs	<i>photo-z</i> (resultado do TPZ), magnitudes <i>grizY</i> e os respectivos erros
Outputs	magnitude absoluta, correção-K, extinção, modelo do melhor ajuste
Configurações	Z_STEP: 0.02,2.,0.1 (bin, máx, bin para $z > 6$ , se houver) COSMOLOGY: 70,0.3,0.7 ( $H_0, \Omega_m, \Omega_\Lambda$ ) MOD_EXTINC: 15,29 (intervalo de SEDs para aplicar extinção) EXTINC_LAW: SMC_prevot.dat (Prevot et al. 1984 [65]) EB_V: 0.0,0.05,0.1 (valores de E(B-V) introduzidos nas SEDs) SEDs: 12 elípticas, 16 espirais e 1 irregular (Coleman, Wu & Weedmann 1980 [22])
Tempo gasto	$\sim 1h/100$ mil objetos

---

# Apêndice C

## Descrição dos Dados: SVA1

### C.1 Magnitudes e erros

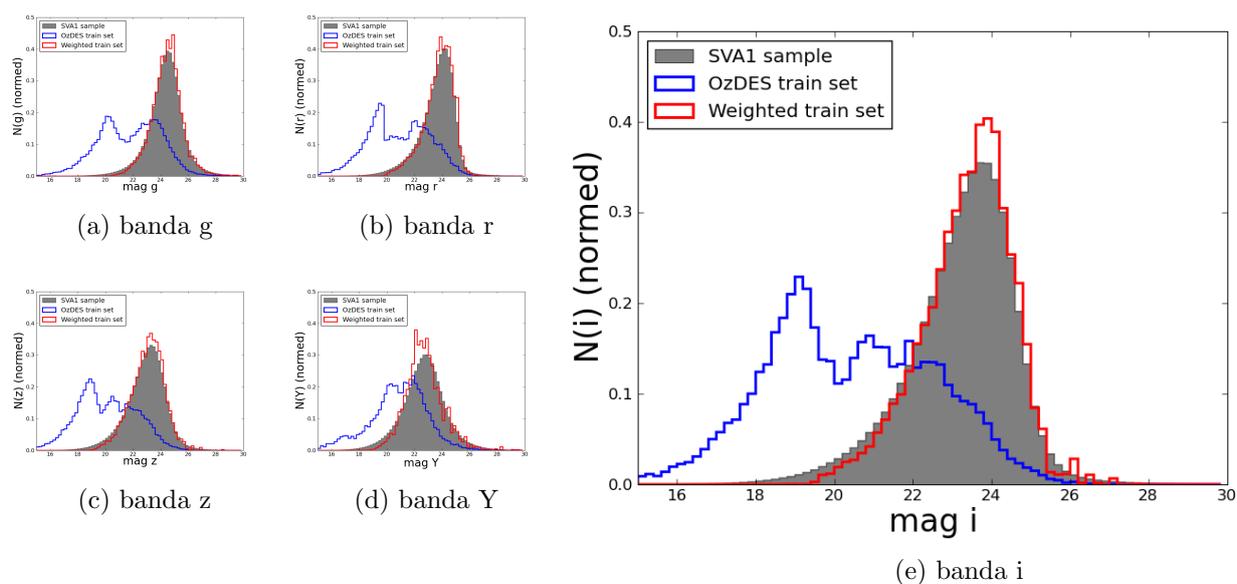


Figura C.1: Distribuição de magnitudes: Em cinza, a amostra SVA1 completa. Em azul, o conjunto de treinamento OzDES. Em vermelho, o conjunto de treinamento representativo (OzDES ponderado). Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1).

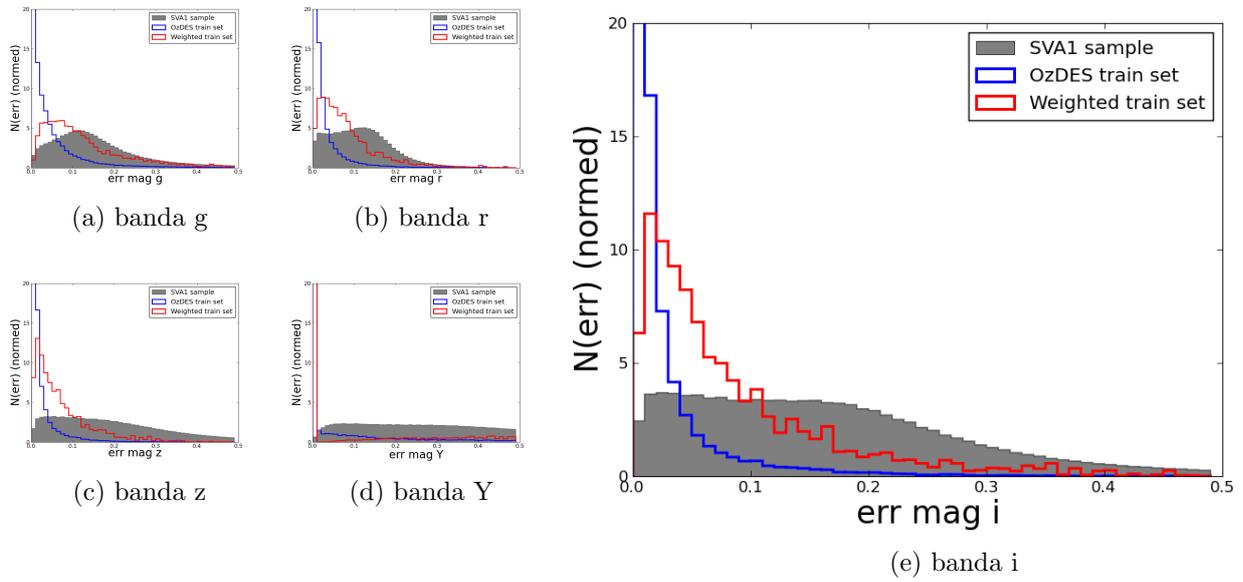


Figura C.2: Distribuição de erros nas magnitudes: Em cinza, a amostra SVA1 completa. Em azul, o conjunto de treinamento OzDES. Em vermelho, o conjunto de treinamento representativo (OzDES ponderado). Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1).

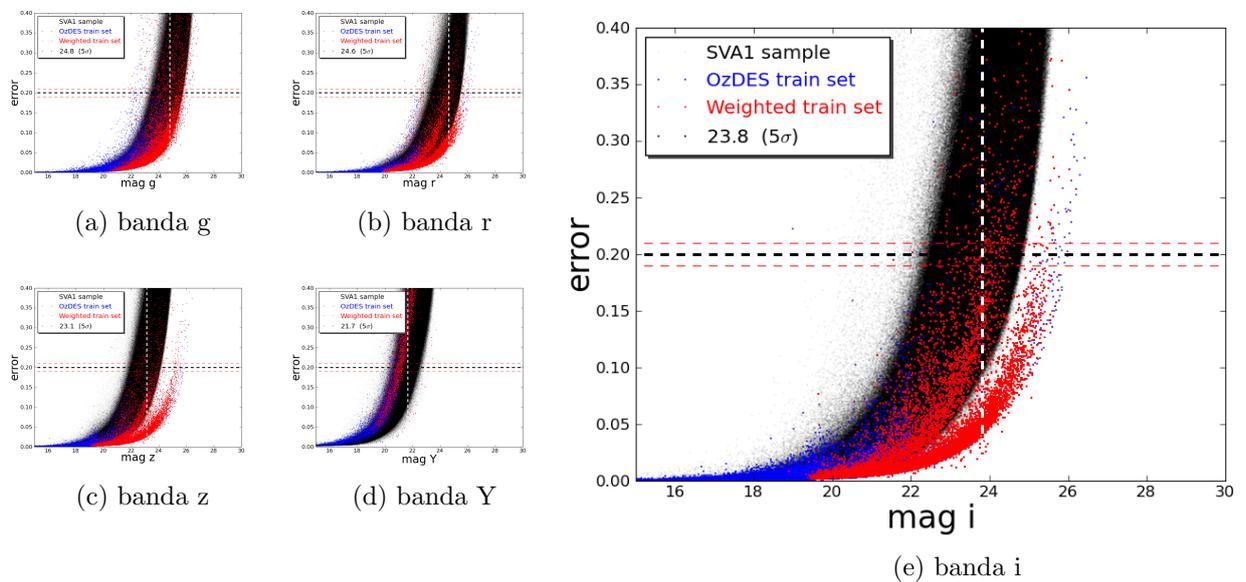


Figura C.3: Relação entre magnitudes aparentes e seus erros: Em preto, a amostra SVA1 completa. Em azul, o conjunto de treinamento OzDES. Em vermelho, o conjunto de treinamento representativo (OzDES ponderado). O número na legenda representa a magnitude média das galáxias com erros nas magnitudes de  $0.2 \pm 0.01$ .

## C.2 Cores

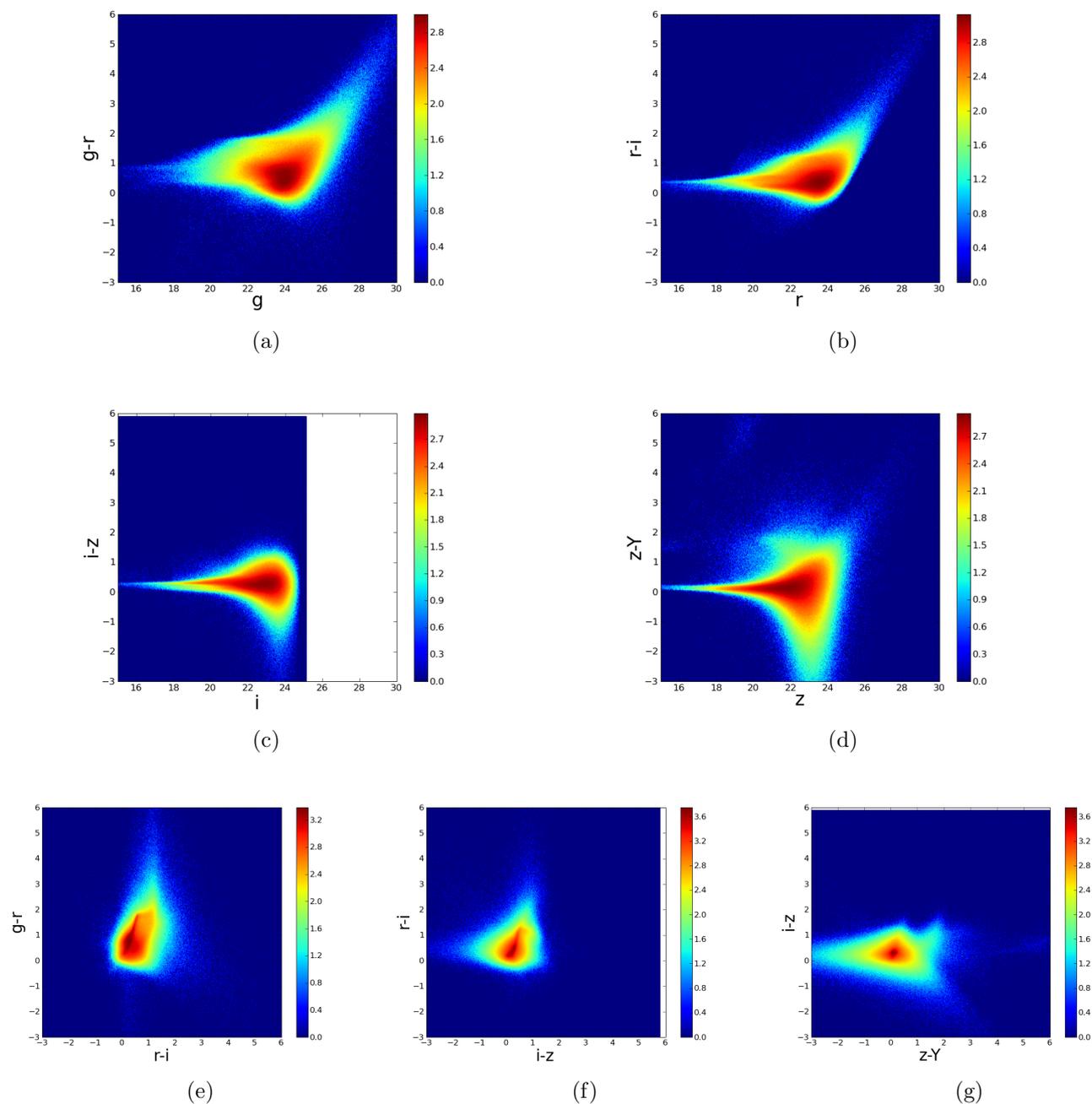


Figura C.4: Diagramas cor-magnitude e cor-cor para a amostra sva1 completa. As cores representam a densidade de pontos em escala logarítmica.

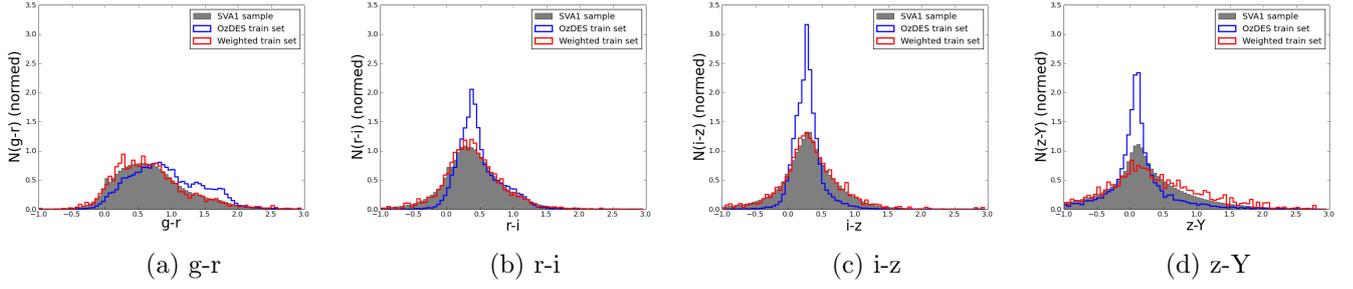


Figura C.5: Distribuição de cores: Em cinza, a amostra SVA1 completa. Em azul, o conjunto de treinamento OzDES. Em vermelho, o conjunto de treinamento representativo (OzDES ponderado). Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1).

### C.3 Redshift e Distribuição Espacial

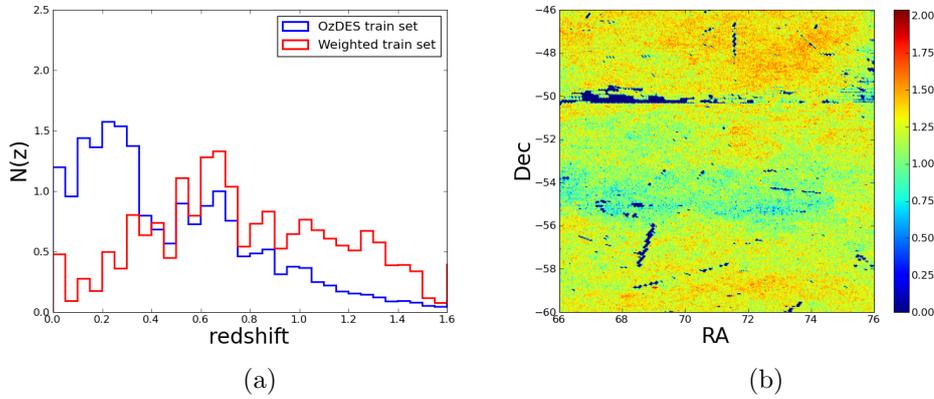


Figura C.6: C.6a: Redshifts das amostras simuladas e do conjunto observado OzDES: Em azul, o conjunto de treinamento OzDES. Em vermelho, o conjunto de treinamento representativo (OzDES ponderado). Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1). C.6b: Distribuição espacial projetada em duas dimensões das galáxias na amostra SVA1 completa. As cores representam a densidade de pontos em escala logarítmica.

#### C.3.1 Profundidade das observações

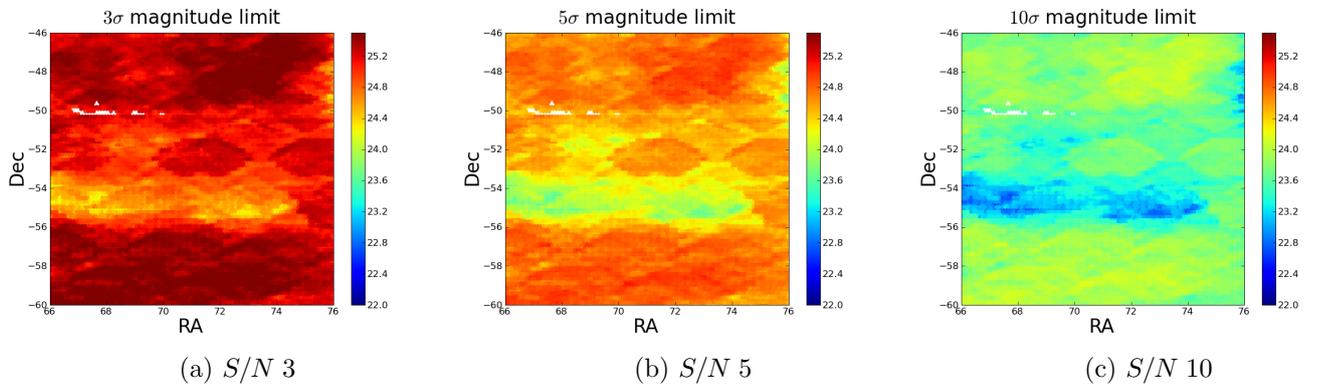


Figura C.7: Mapas de magnitude limite: as cores representam a magnitude limite à razão sinal-ruído 3, 5, e, 10, respectivamente.

# Apêndice D

## Separação Estrela/Galáxia

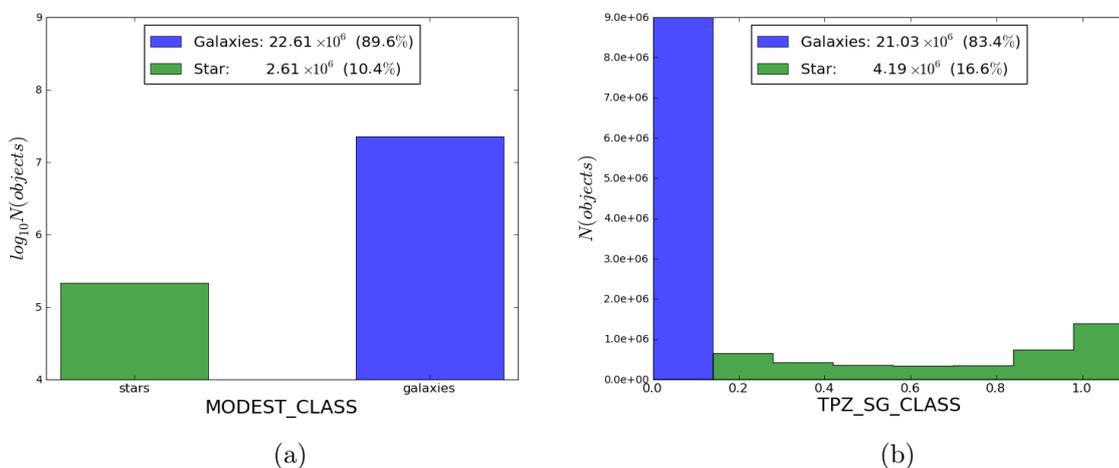
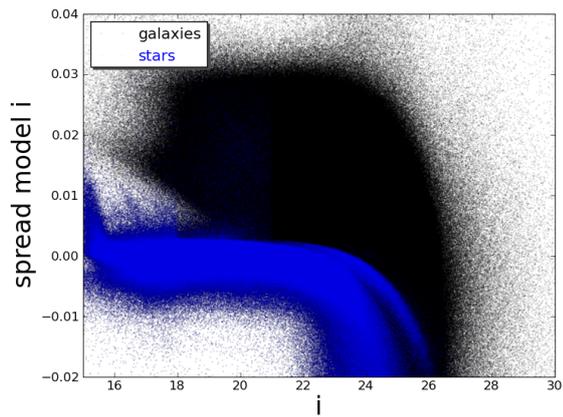
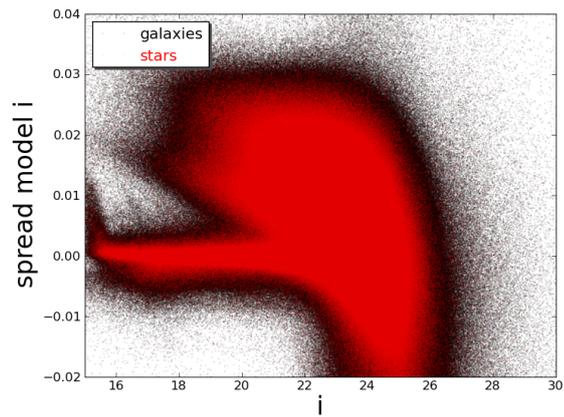


Figura D.1: Distribuição dos parâmetros MODEST\_CLASS e TPZ\_SG\_CLASS, utilizados para classificar os objetos quanto a sua natureza (estrela ou galáxia).

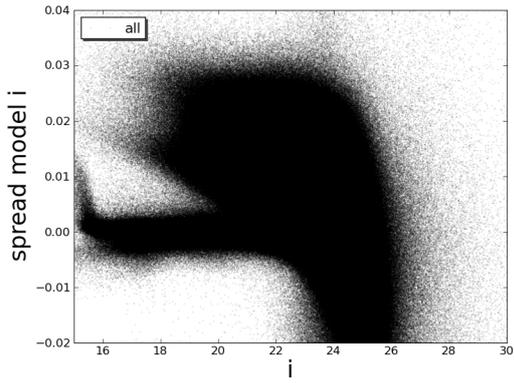


(a) MODEST CLASS

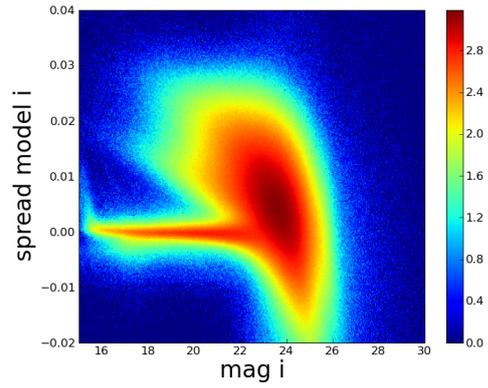


(b) TPZ SG CLASS

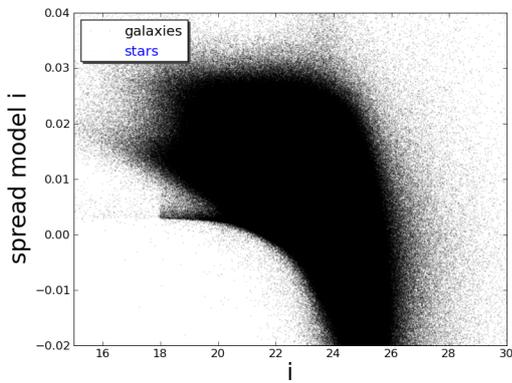
Figura D.2: SPREAD\_MODEL\_I versus magnitude na banda  $i$ . Em preto os objetos classificados como galáxia e em azul ou vermelho, os objetos classificados como estrelas pelos métodos MODEST\_CLASS e TPZ\_SG\_CLASS, respectivamente.



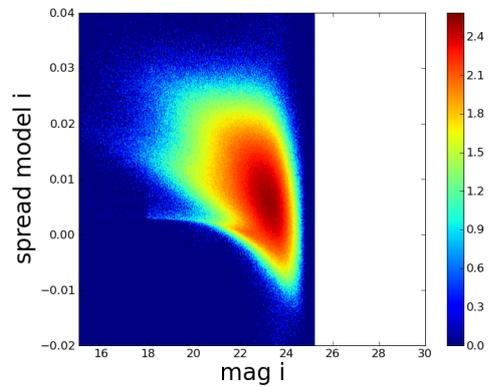
(a) c/ estrelas



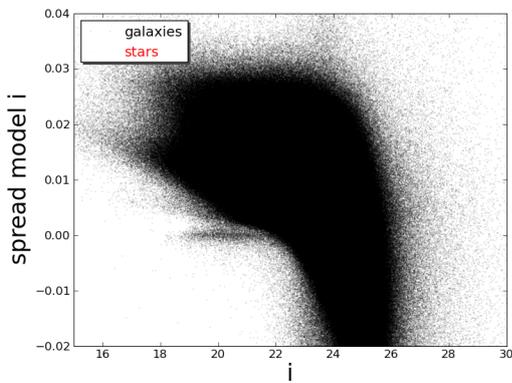
(b) c/ estrelas



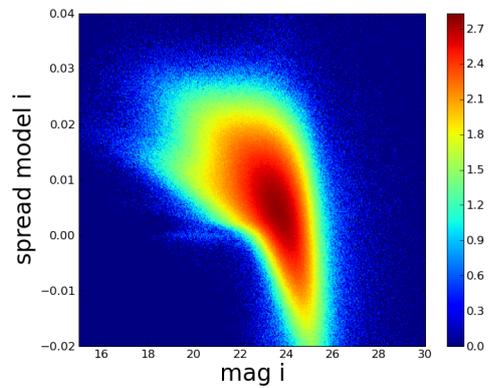
(c) MODEST CLASS



(d) MODEST CLASS



(e) TPZ SG CLASS



(f) TPZ SG CLASS

Figura D.3: SPREAD\_MODEL\_I versus magnitude na banda  $i$  antes e após separação estrela/galáxia por diferentes métodos. Nos gráficos de densidade, as cores representam a densidade de pontos em escala logarítmica.

## D.1 Magnitudes e erros

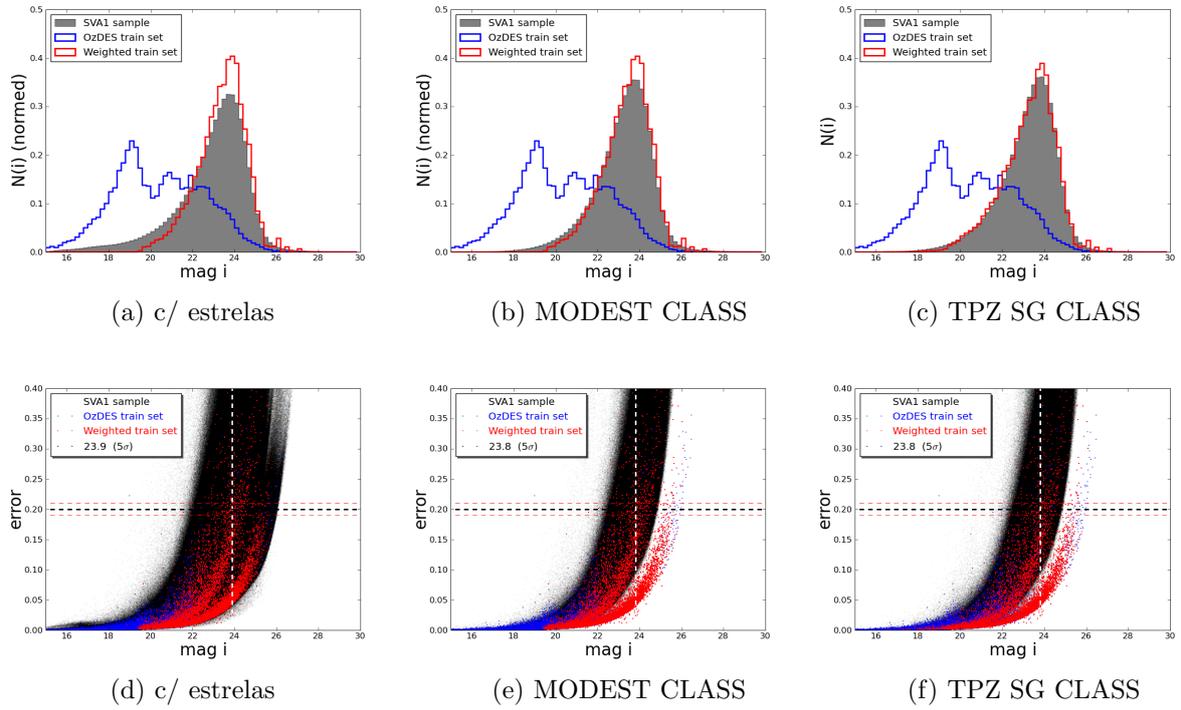
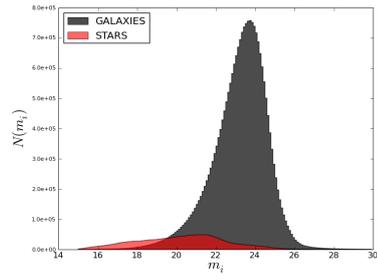
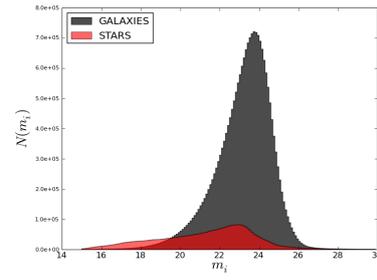


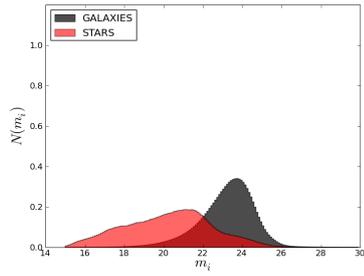
Figura D.4: Distribuição de magnitudes na banda  $i$  após separação estrela/galáxia: Em cinza, a amostra SVA1 completa. Em azul, o conjunto de treinamento OzDES. Em vermelho, o conjunto de treinamento OzDES ponderado. Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1)



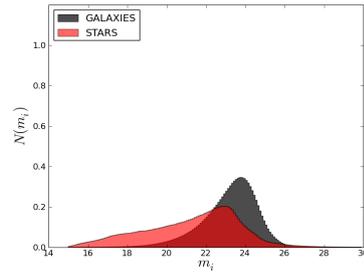
(a) MODEST CLASS



(b) TPZ SG CLASS



(c) MODEST CLASS



(d) TPZ SG CLASS

Figura D.5: Distribuição de magnitudes na banda  $i$  após separação estrela/galáxia, para estrelas e galáxias separadamente. Linha inferior: histogramas normalizados. Linha inferior: histogramas com quantidades reais.

# Apêndice E

## Descrição dos Dados: BCC

### E.1 Magnitudes e erros

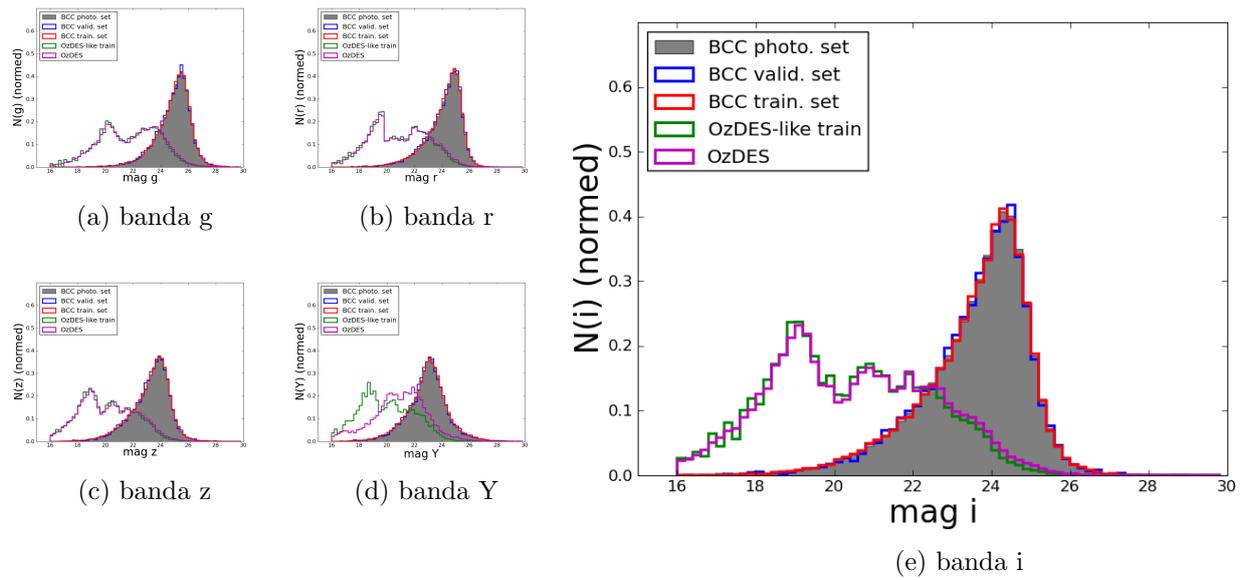


Figura E.1: Distribuição de magnitudes: Em cinza, a amostra BCC completa. Em azul, o conjunto de validação. Em vermelho, o conjunto de treinamento representativo. Em magenta, o conjunto de referência de dados observados OzDES. Em verde, o conjunto de treinamento "OzDES-like". Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1)

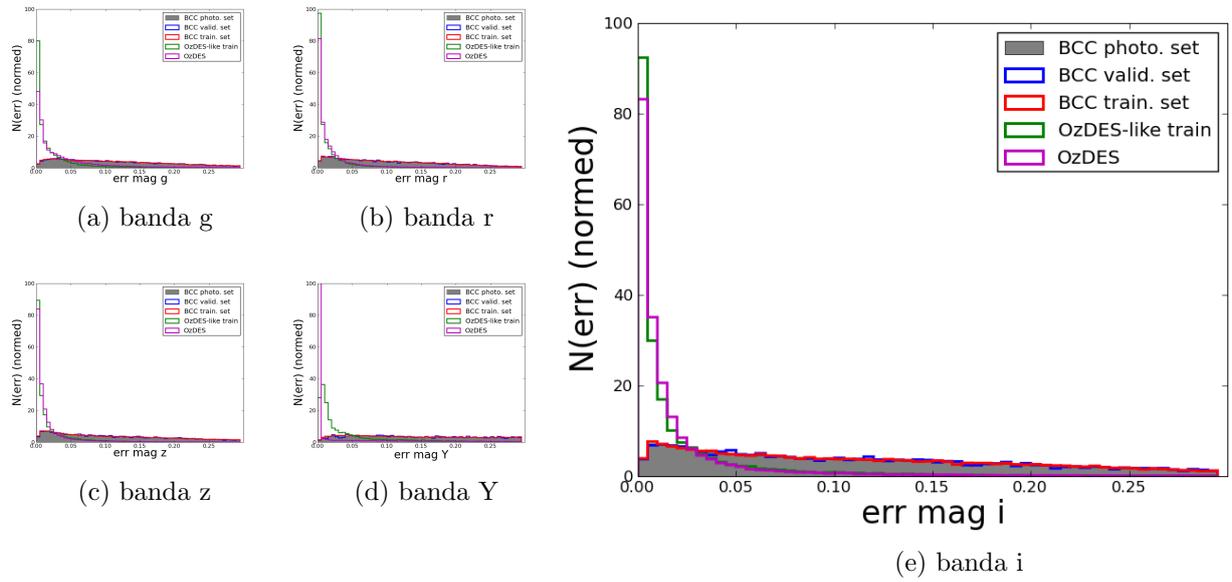


Figura E.2: Distribuição de erros nas magnitudes: Em cinza, a amostra BCC completa. Em azul, o conjunto de validação. Em vermelho, o conjunto de treinamento representativo. Em magenta, o conjunto de referência de dados observados OzDES. Em verde, o conjunto de treinamento "OzDES-like". Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1)

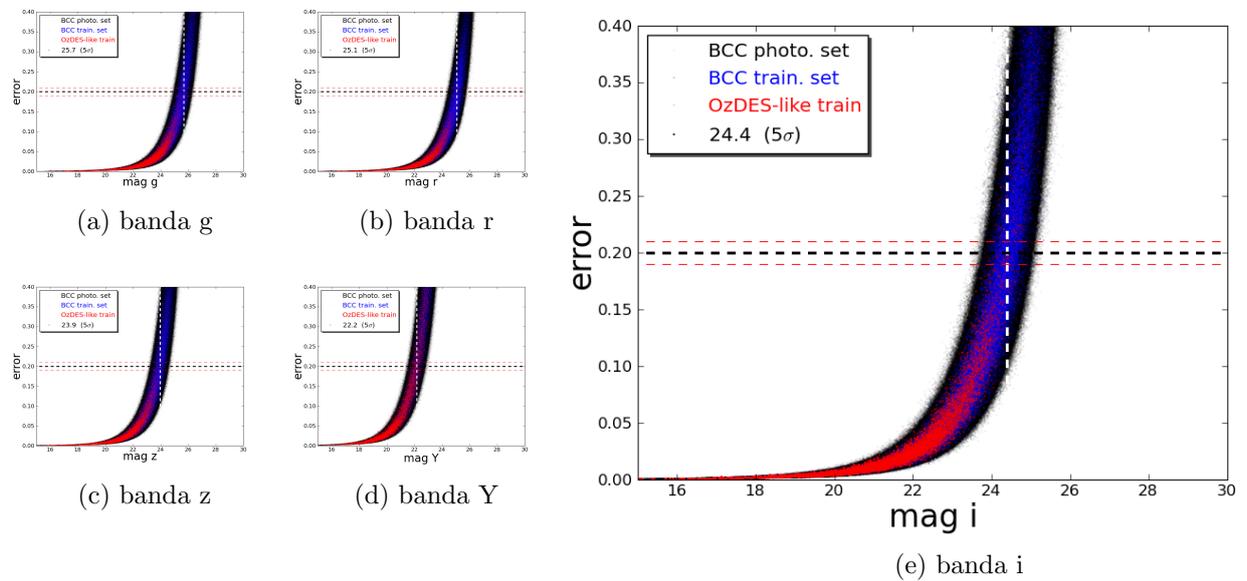


Figura E.3: Relação entre magnitudes aparentes e seus erros: Em preto, a amostra BCC completa. Em azul, o conjunto de treinamento representativo. Em vermelho, o conjunto de treinamento OzDES-like. O número na legenda representa a magnitude média das galáxias com erros nas magnitudes de  $0.2 \pm 0.01$ .

## E.2 Cores

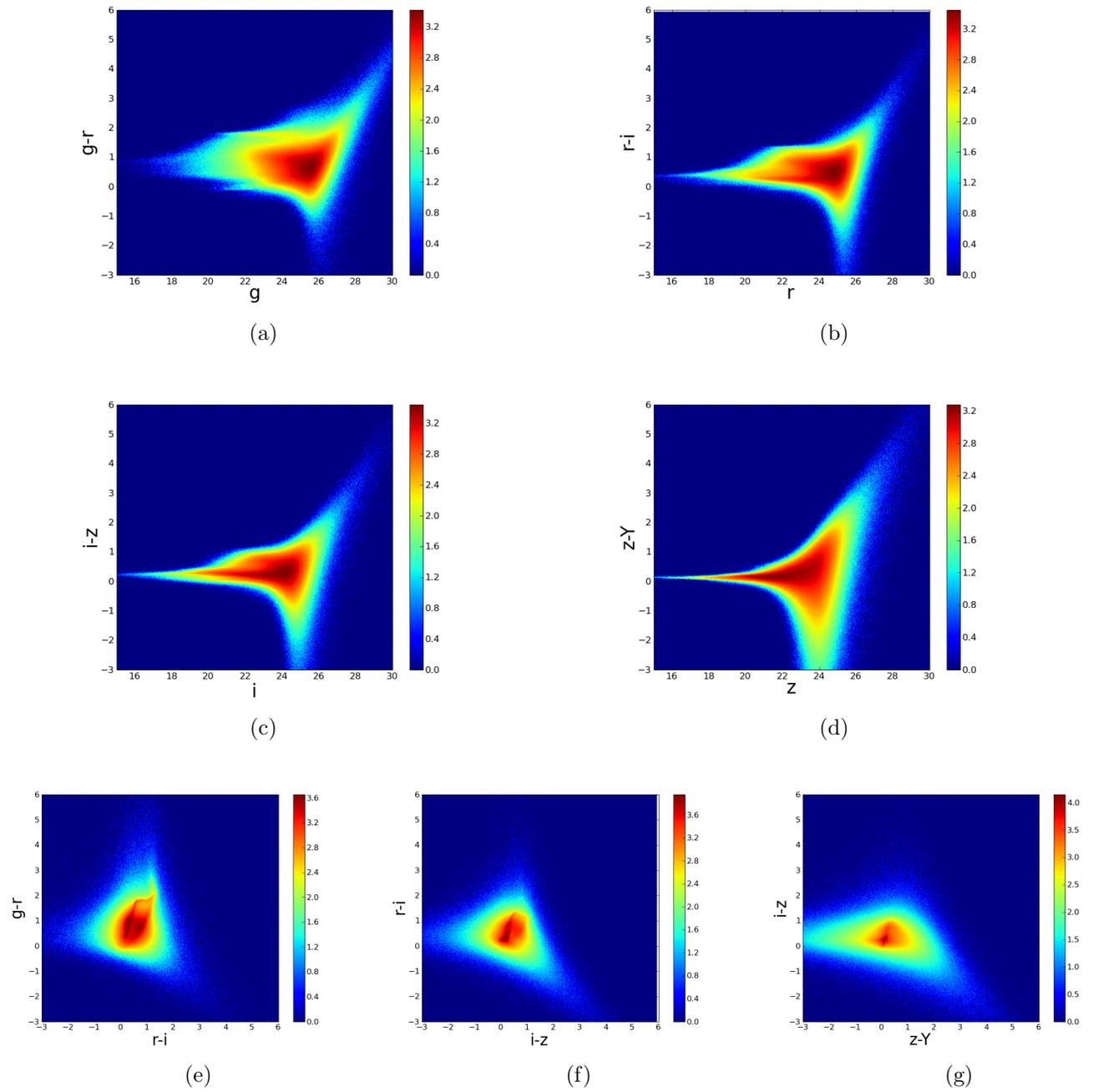


Figura E.4: Diagramas cor-magnitude e cor-cor para a amostra BCC completa. As cores representam a densidade de pontos em escala logarítmica.

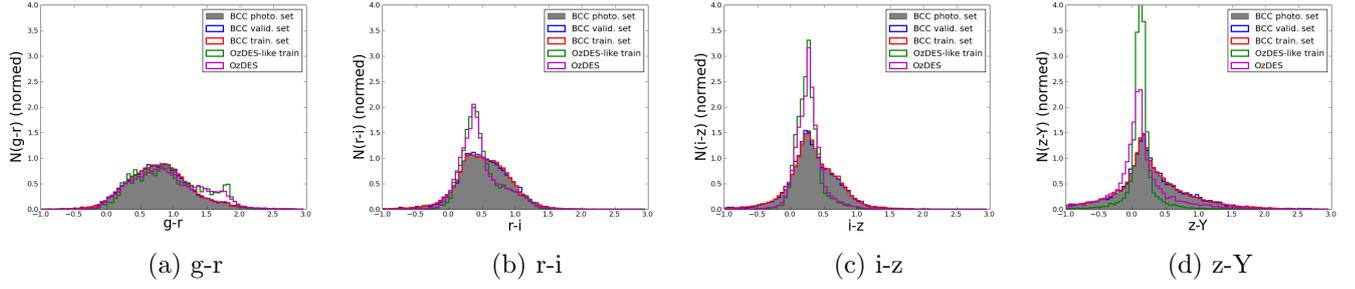


Figura E.5: Distribuição de cores: Em cinza, a amostra BCC completa. Em azul, o conjunto de validação. Em vermelho, o conjunto de treinamento representativo. Em magenta, o conjunto de dados observados OzDES. Em verde, o conjunto de treinamento "OzDES-like". Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1).

### E.3 Redshift e Distribuição Espacial

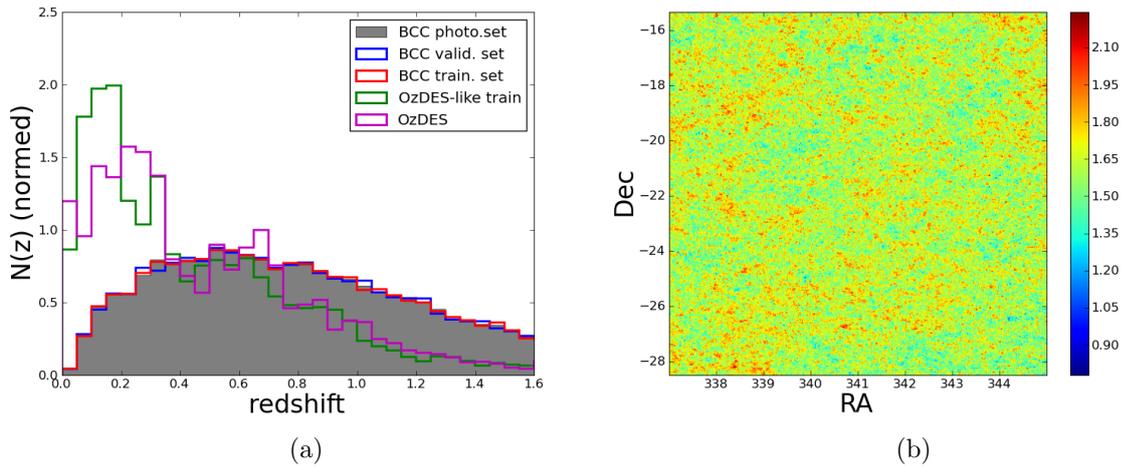


Figura E.6: E.6a: Redshifts das amostras simuladas e do conjunto observado OzDES: Em cinza, a amostra BCC completa. Em azul, o conjunto de validação. Em vermelho, o conjunto de treinamento representativo. Em magenta, o conjunto de dados observados OzDES. Em verde, o conjunto de treinamento "OzDES-like". Os histogramas são normalizados para facilitar a comparação (área total das figuras = 1). E.6b: Distribuição espacial projetada em duas dimensões das galáxias na amostra BCC completa, mostrando que a amostra não sofre efeitos de variância cósmica significativos. As cores representam a densidade de pontos em escala logarítmica.

# Apêndice F

## Resultados com a Amostra BCC

### F.1 Validação do *photo-z*

#### F.1.1 Conjunto de Treinamento Representativo

Tabela F.1: Métricas obtidas após o cálculo do *photo-z* no conjunto de validação. Treinamento realizado com o conjunto **representativo**. Método: **TPZ mode** (*photo-z* = *z* correspondente ao pico da  $P(z)$ ).

Métrica	amostra completa	$S/N > 3$	$S/N > 5$	$S/N > 10$
$\overline{\Delta z}$	$-0.0406 \pm 0.0011$	$-0.0345 \pm 0.0008$	$-0.0245 \pm 0.0008$	$-0.0072 \pm 0.0004$
$\Delta z_{50}$	$-0.0062 \pm 0.0006$	$-0.0055 \pm 0.0004$	$-0.0031 \pm 0.0003$	$0.0014 \pm 0.0002$
$\sigma_{\Delta z}$	$0.2779 \pm 0.0014$	$0.2418 \pm 0.0012$	$0.2039 \pm 0.0013$	$0.1207 \pm 0.0010$
$\sigma_{68}$	$0.1521 \pm 0.0007$	$0.1344 \pm 0.0010$	$0.1080 \pm 0.0006$	$0.0465 \pm 0.0003$
$out_{2\sigma}$	$0.0644 \pm 0.0005$	$0.0652 \pm 0.0005$	$0.0590 \pm 0.0006$	$0.0582 \pm 0.0006$
$out_{3\sigma}$	$0.0270 \pm 0.0003$	$0.0227 \pm 0.0004$	$0.0216 \pm 0.0004$	$0.0291 \pm 0.0004$
$\overline{\Delta z'}$	$-0.1556 \pm 0.0055$	$-0.1395 \pm 0.0040$	$-0.1165 \pm 0.0049$	$-0.0159 \pm 0.0033$
$\sigma_{\Delta z'}$	$1.4316 \pm 0.0100$	$1.3511 \pm 0.0086$	$1.3763 \pm 0.0104$	$0.9520 \pm 0.0065$
$N_{poisson}$	$6.3630 \pm 0.0989$	$6.1544 \pm 0.0923$	$6.1084 \pm 0.0887$	$4.9592 \pm 0.0770$
$KS$	$0.0531 \pm 0.0012$	$0.0581 \pm 0.0010$	$0.0517 \pm 0.0011$	$0.0286 \pm 0.0008$

Tabela F.2: Métricas obtidas após cálculo do photo-z no conjunto de validação. Treinamento realizado com o conjunto **representativo**. Metodo: **TPZ mean** ( $photo-z = \text{m\u00e9dia da PDF}(z)$ )

M\u00e9trica	amostra completa	$S/N > 3$	$S/N > 5$	$S/N > 10$
$\overline{\Delta z}$	$0.0040 \pm 0.0010$	$0.0054 \pm 0.0007$	$0.0050 \pm 0.0007$	$0.0015 \pm 0.0004$
$\Delta z_{50}$	$0.0052 \pm 0.0005$	$0.0057 \pm 0.0004$	$0.0043 \pm 0.0004$	$0.0027 \pm 0.0002$
$\sigma_{\Delta z}$	$0.2559 \pm 0.0011$	$0.2329 \pm 0.0011$	$0.1962 \pm 0.0010$	$0.1111 \pm 0.0008$
$\sigma_{68}$	$0.1643 \pm 0.0008$	$0.1487 \pm 0.0009$	$0.1190 \pm 0.0007$	$0.0633 \pm 0.0004$
$out_{2\sigma}$	$0.0663 \pm 0.0005$	$0.0639 \pm 0.0005$	$0.0595 \pm 0.0006$	$0.0545 \pm 0.0006$
$out_{3\sigma}$	$0.0228 \pm 0.0003$	$0.0239 \pm 0.0003$	$0.0216 \pm 0.0003$	$0.0248 \pm 0.0004$
$\overline{\Delta z'}$	$-0.0573 \pm 0.0052$	$-0.0456 \pm 0.0039$	$-0.0384 \pm 0.0044$	$0.0209 \pm 0.0031$
$\sigma_{\Delta z'}$	$1.3610 \pm 0.0100$	$1.2950 \pm 0.0083$	$1.2879 \pm 0.0085$	$0.8659 \pm 0.0052$
$N_{poisson}$	$8.5829 \pm 0.1049$	$8.0185 \pm 0.1113$	$8.4844 \pm 0.0936$	$7.7611 \pm 0.1004$
$KS$	$0.0700 \pm 0.0013$	$0.0647 \pm 0.0011$	$0.0477 \pm 0.0009$	$0.0405 \pm 0.0009$

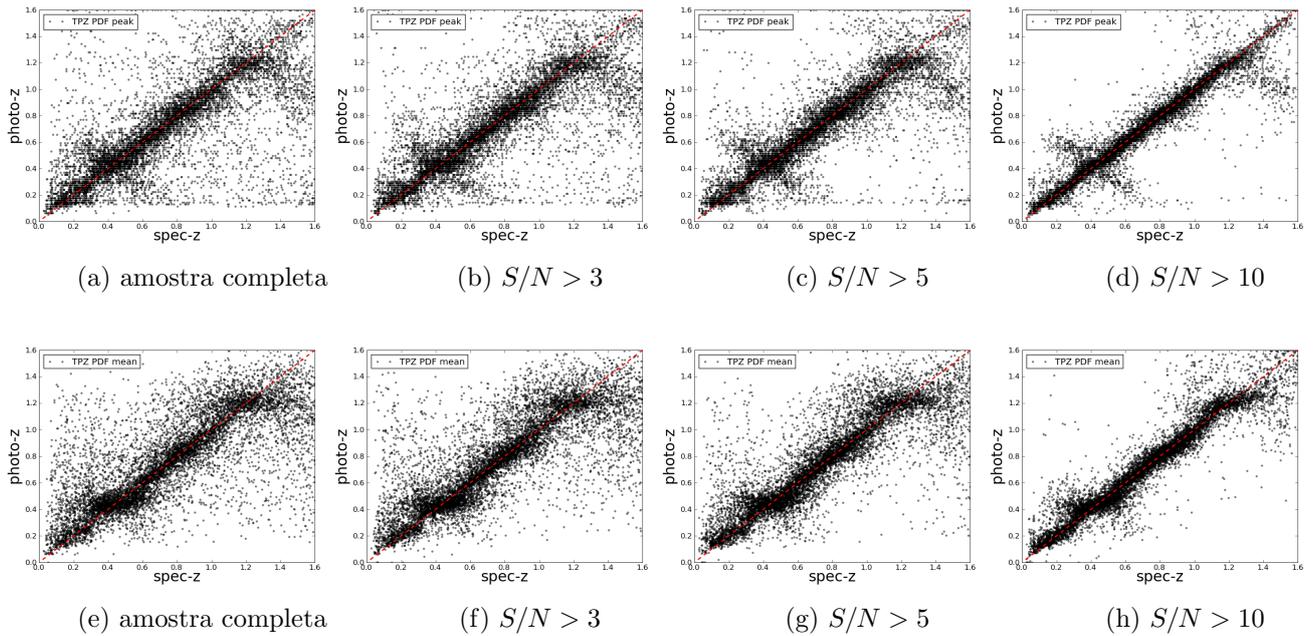


Figura F.1: Gr\u00e1fico de dispers\u00e3o da rela\u00e7\u00e3o entre  $photo-z$  e  $spec-z$ , para os diferentes sele\u00e7\u00f5es na amostra, baseadas na raz\u00e3o sinal-ru\u00eddo. Resultados obtidos para o conjunto de valida\u00e7\u00e3o, ap\u00f3s treinamento ter sido realizado com o conjunto **representativo**. Linha superior: **TPZ mode**; linha inferior: **TPZ mean**.

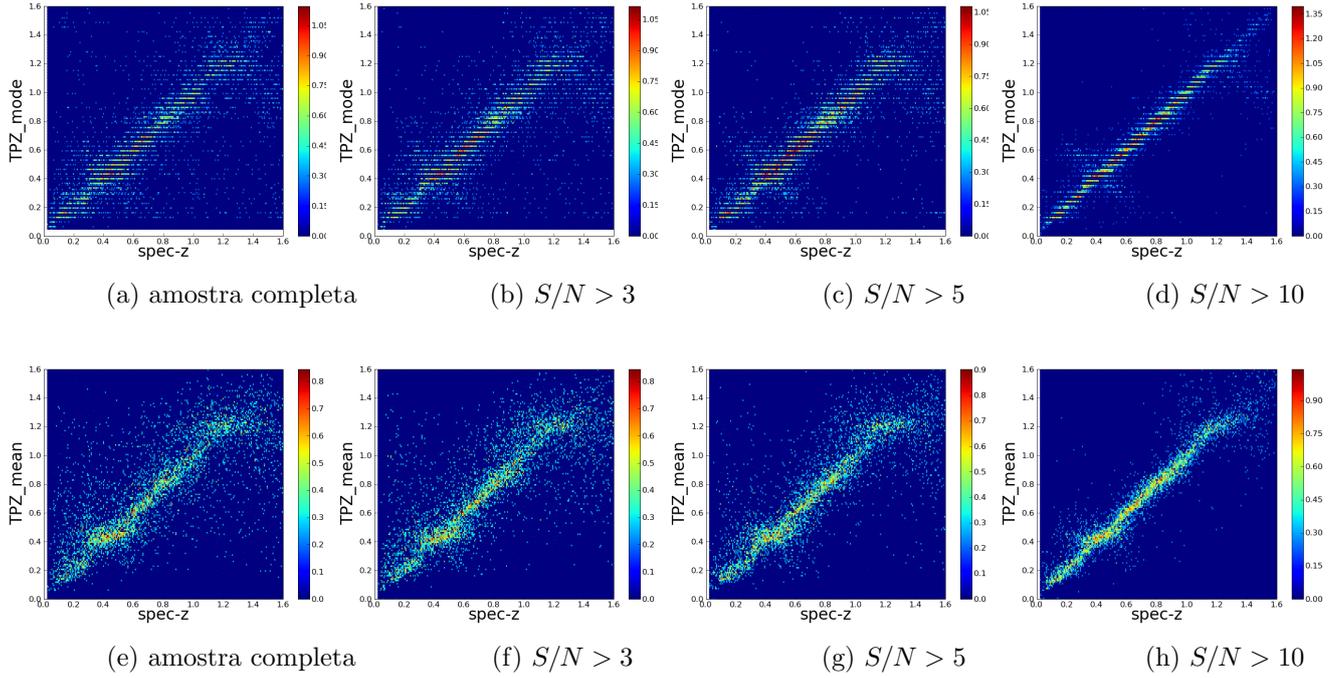


Figura F.2: Gráfico de densidades da dispersão da relação entre *photo-z* e *spec-z*, para os diferentes seleções na amostra, baseadas na razão sinal-ruído. Resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto **representativo**. A escala de cores representa a densidade de pontos em cada região do gráfico, em escala logarítmica. Linha superior: **TPZ mode**; linha inferior: **TPZ mean**.

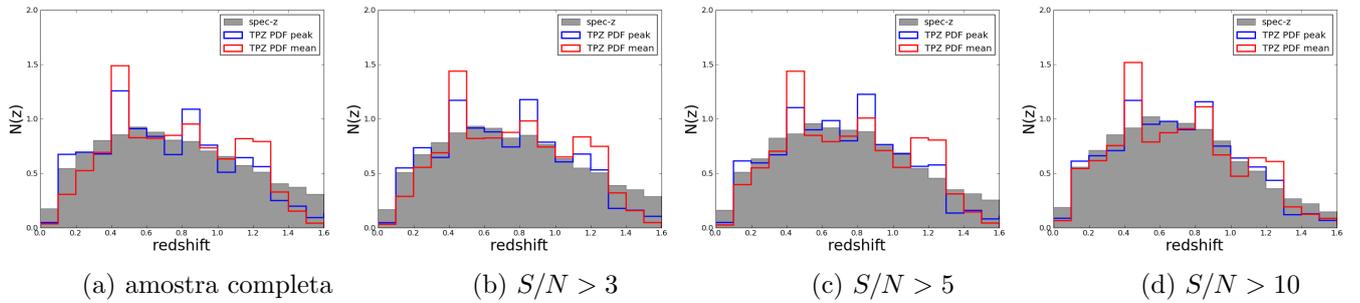


Figura F.3: Gráfico da distribuição de *redshifts* obtidos com o TPZ, em comparação com o *true-z* (*redshift* original da simulação), para os diferentes seleções na amostra, baseadas na razão sinal-ruído. Resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto **representativo**. Em cinza: *true-z*; em azul: **TPZ mode**; em vermelho: **TPZ mean**.

## F.1.2 Conjunto de Treinamento OzDES-like (nao representativo)

Tabela F.3: Metricas obtidas apos calculo do photo-z no conjunto de validaao. Treinamento realizado com o conjunto **OzDES-like** (nao representativo). Metodo: **TPZ mode** ( $photo-z = z$  correspondente ao pico da PDF( $z$ ))

Metrica	amostra completa	$S/N > 3$	$S/N > 5$	$S/N > 10$
$\overline{\Delta z}$	$-0.0272 \pm 0.0010$	$-0.0188 \pm 0.0008$	$-0.0267 \pm 0.0007$	$-0.0131 \pm 0.0005$
$\Delta z_{50}$	$-0.0058 \pm 0.0005$	$-0.0045 \pm 0.0003$	$-0.0057 \pm 0.0004$	$-0.0012 \pm 0.0002$
$\sigma_{\Delta z}$	$1.2803 \pm 0.0014$	$0.2450 \pm 0.0010$	$0.2111 \pm 0.0012$	$0.1426 \pm 0.0010$
$\sigma_{68}$	$0.1586 \pm 0.0009$	$0.1399 \pm 0.0007$	$0.1140 \pm 0.0007$	$0.0752 \pm 0.0004$
$out_{2\sigma}$	$0.0659 \pm 0.0004$	$0.0657 \pm 0.0006$	$0.0603 \pm 0.0006$	$0.0609 \pm 0.0007$
$out_{3\sigma}$	$0.0272 \pm 0.0003$	$0.0246 \pm 0.0004$	$0.0227 \pm 0.0003$	$0.0256 \pm 0.0004$
$\overline{\Delta z'}$	$-0.0878 \pm 0.0051$	$-0.1036 \pm 0.0047$	$-0.1479 \pm 0.0047$	$-0.0776 \pm 0.0044$
$\sigma_{\Delta z'}$	$1.3606 \pm 0.0105$	$1.4033 \pm 0.0081$	$1.3802 \pm 0.0127$	$1.2265 \pm 0.0089$
$N_{poisson}$	$7.6987 \pm 0.0881$	$9.3324 \pm 0.0906$	$8.0476 \pm 0.0730$	$5.7068 \pm 0.0795$
$KS$	$0.0591 \pm 0.0011$	$0.0636 \pm 0.0010$	$0.0625 \pm 0.0010$	$0.0432 \pm 0.0009$

Tabela F.4: Metricas obtidas apos calculo do photo-z no conjunto de validaao. Treinamento realizado com o conjunto **OzDES-like** (nao representativo). Metodo: **TPZ mean** ( $photo-z =$  media da PDF( $z$ ))

Metrica	amostra completa	$S/N > 3$	$S/N > 5$	$S/N > 10$
$\overline{\Delta z}$	$-0.0272 \pm 0.0011$	$0.0044 \pm 0.0008$	$-0.0095 \pm 0.0007$	$-0.0038 \pm 0.0005$
$\Delta z_{50}$	$-0.0058 \pm 0.0006$	$0.0032 \pm 0.0004$	$-0.0009 \pm 0.0004$	$0.0015 \pm 0.0003$
$\sigma_{\Delta z}$	$0.2803 \pm 0.0011$	$0.2350 \pm 0.0009$	$0.1983 \pm 0.0011$	$0.1378 \pm 0.0009$
$\sigma_{68}$	$0.1586 \pm 0.0010$	$0.1498 \pm 0.0008$	$0.1204 \pm 0.0007$	$0.0821 \pm 0.0005$
$out_{2\sigma}$	$0.0659 \pm 0.0005$	$0.0650 \pm 0.0006$	$0.0603 \pm 0.0005$	$0.0580 \pm 0.0006$
$out_{3\sigma}$	$0.0272 \pm 0.0004$	$0.0240 \pm 0.0003$	$0.0232 \pm 0.0003$	$0.0222 \pm 0.0003$
$\overline{\Delta z'}$	$-0.0878 \pm 0.0052$	$-0.0361 \pm 0.0046$	$-0.1040 \pm 0.0047$	$-0.0527 \pm 0.0042$
$\sigma_{\Delta z'}$	$1.3606 \pm 0.0098$	$1.3565 \pm 0.0078$	$1.3241 \pm 0.0126$	$1.1554 \pm 0.0077$
$N_{poisson}$	$7.6987 \pm 0.0969$	$8.1378 \pm 0.0876$	$8.2502 \pm 0.1055$	$8.5093 \pm 0.0954$
$KS$	$0.0591 \pm 0.0014$	$0.0720 \pm 0.0011$	$0.0598 \pm 0.0010$	$0.0547 \pm 0.0010$

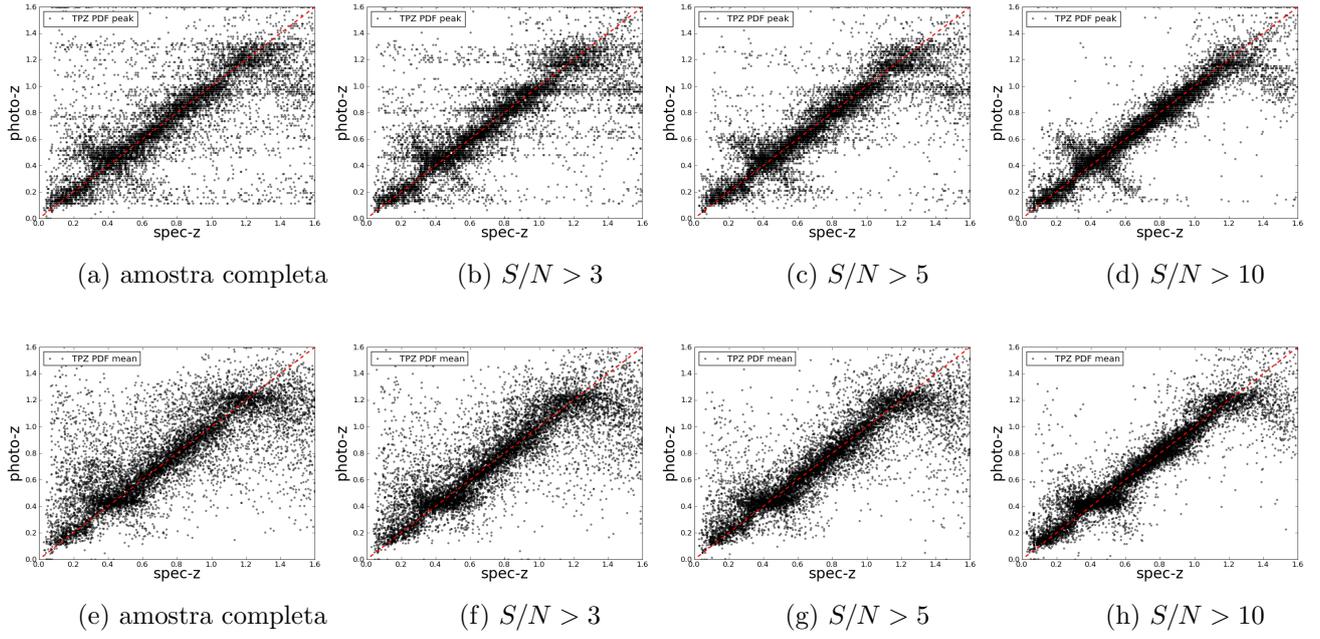


Figura F.4: Gráfico de dispersão da relação entre  $photo-z$  e  $spec-z$ , para os diferentes seleções na amostra, baseadas na razão sinal-ruído. Resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto **OzDES-like** (não representativo). Linha superior: **TPZ mode**; linha inferior: **TPZ mean**.

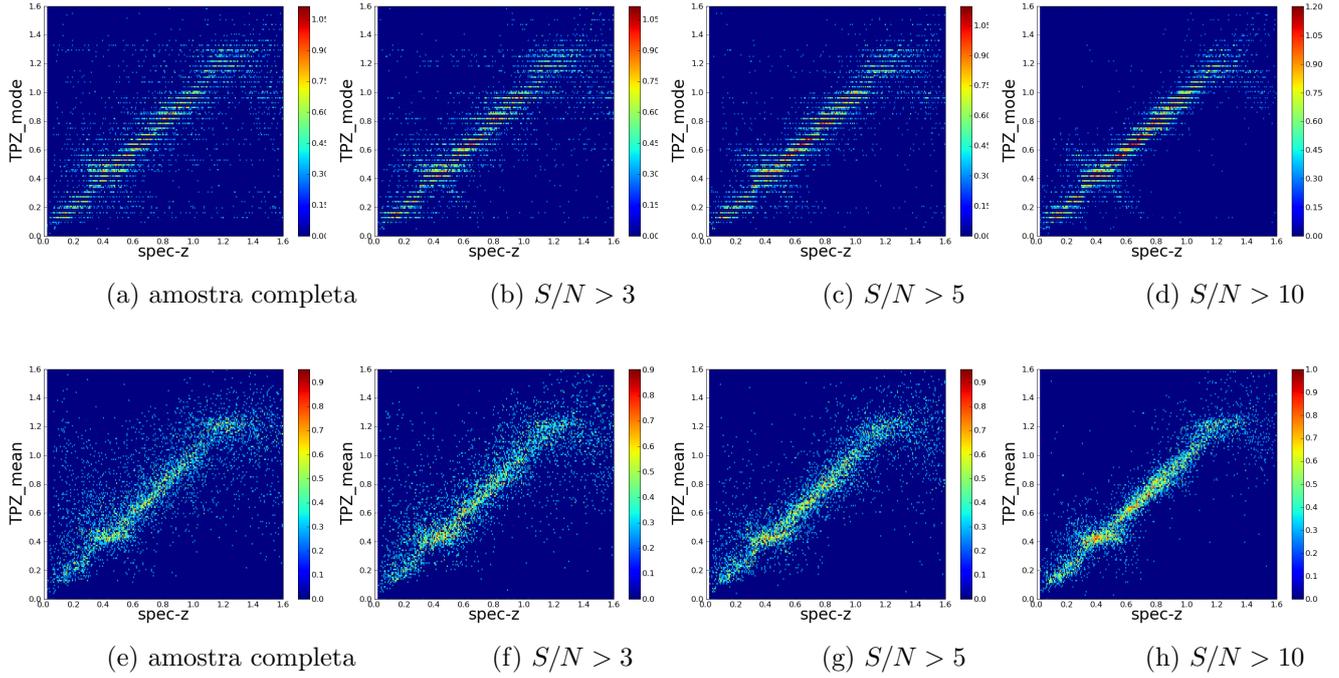


Figura F.5: Gráfico de densidades da dispersão da relação entre *photo-z* e *spec-z*, para os diferentes seleções na amostra, baseadas na razão sinal-ruído. Resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto **OzDES-like** (não representativo). A escala de cores representa a densidade de pontos em cada região do gráfico, em escala logarítmica. Linha superior: **TPZ mode**; linha inferior: **TPZ mean**.

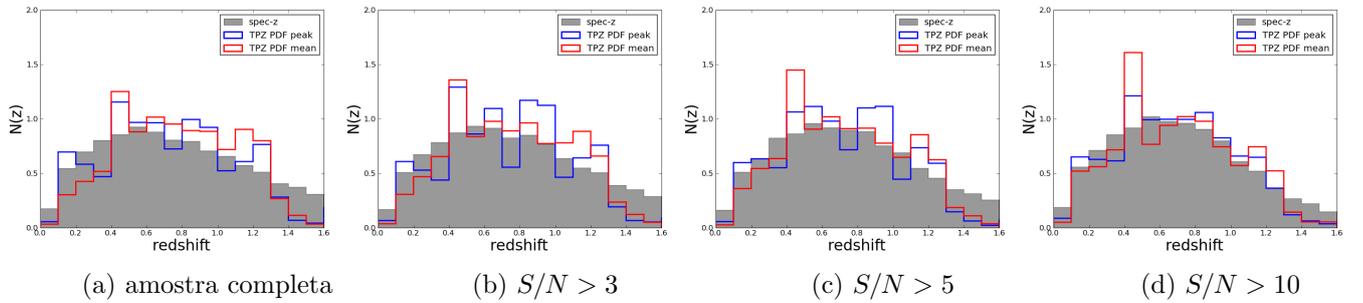


Figura F.6: Gráfico da distribuição de *redshifts* obtidos com o TPZ, em comparação com o *true-z* (*redshift* original da simulação), para os diferentes seleções na amostra, baseadas na razão sinal-ruído. Resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com conjunto **OzDES-like** (não representativo). Em cinza: *true-z*; em azul: **TPZ mode**; em vermelho: **TPZ mean**.

Tabela F.5: Comparação baseada na métrica  $\sigma_{68}$ , entre as performances do TPZ, após diferentes seleções em razão sinal-ruído.

Conjunto	método	amostra completa	$S/N > 3$	$S/N > 5$	$S/N > 10$
Representativo	mode	$0.1521 \pm 0.0007$	$0.1344 \pm 0.0010$	$0.1080 \pm 0.0006$	$0.0465 \pm 0.0003$
Representativo	mean	$0.1643 \pm 0.0008$	$0.1487 \pm 0.0009$	$0.1190 \pm 0.0007$	$0.0633 \pm 0.0004$
OzDES-like	mode	$0.1586 \pm 0.0009$	$0.1399 \pm 0.0007$	$0.1140 \pm 0.0007$	$0.0752 \pm 0.0004$
OzDES-like	mean	$0.1586 \pm 0.0010$	$0.1498 \pm 0.0008$	$0.1204 \pm 0.0007$	$0.0821 \pm 0.0005$

## F.2 *Photo-z* do conjunto fotométrico

### F.2.1 Subconjunto com 100 mil objetos

#### F.2.1.1 Conjunto de Treinamento Representativo

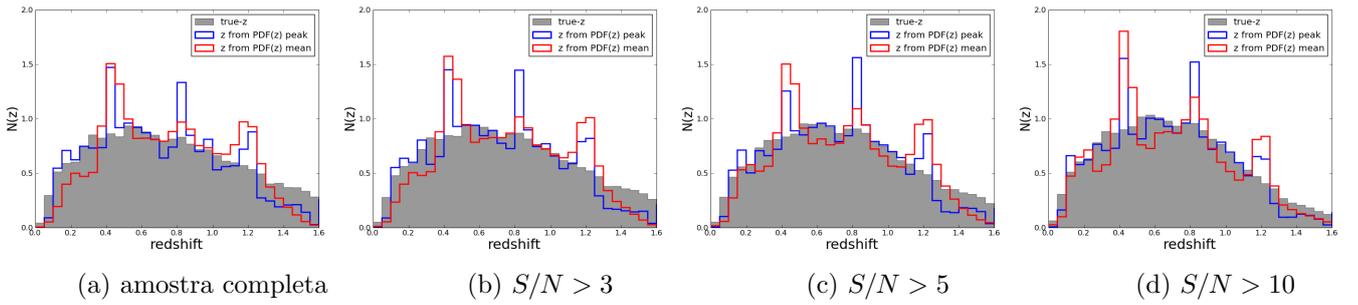


Figura F.7: Resultado do TPZ em um conjunto de 100 mil galáxias aleatoriamente selecionadas do conjunto fotométrico. Árvores produzidas com o conjunto representativo.

#### F.2.1.2 Conjunto de Treinamento OzDES-like (não representativo)

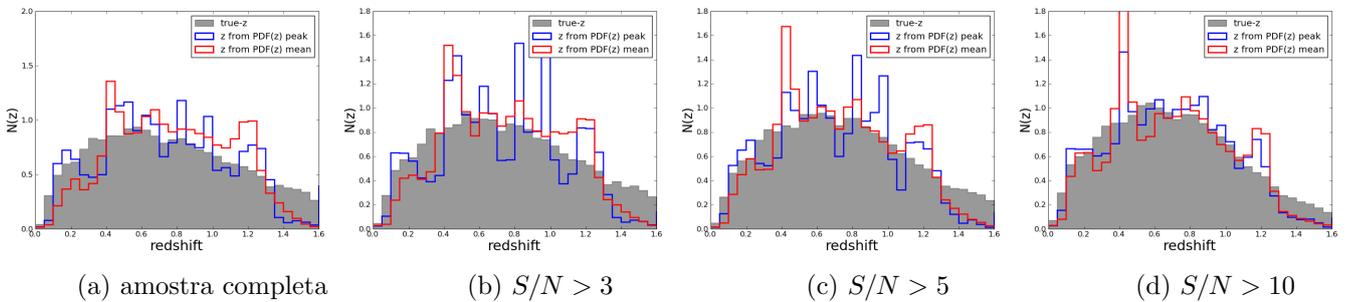


Figura F.8: Resultado do TPZ em um conjunto de 100 mil galáxias aleatoriamente selecionadas do conjunto fotométrico. Árvores produzidas com o conjunto OzDES-like (não representativo).

## F.2.2 Amostra fotométrica completa

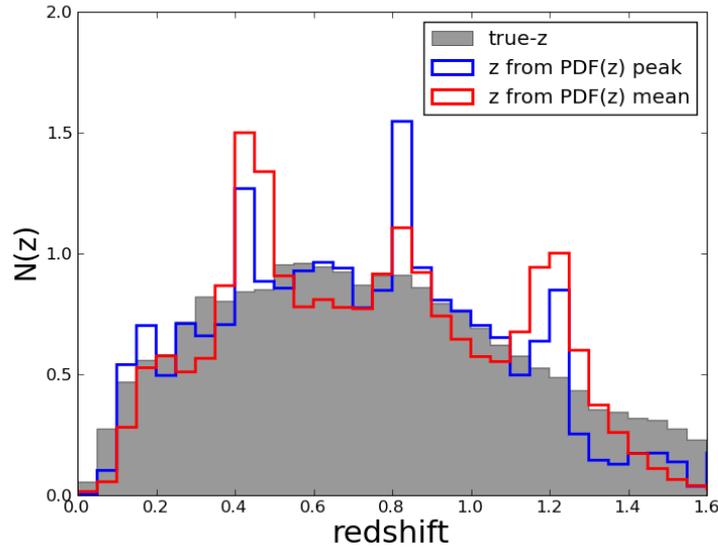


Figura F.9: Resultado do TPZ para todas as galáxias que serão utilizadas na determinação da função de luminosidade (remoção de objetos com  $S/N < 5$ ). Árvores produzidas com o conjunto representativo.

## F.3 Produtos do LePhare

### F.3.1 Magnitudes Absolutas

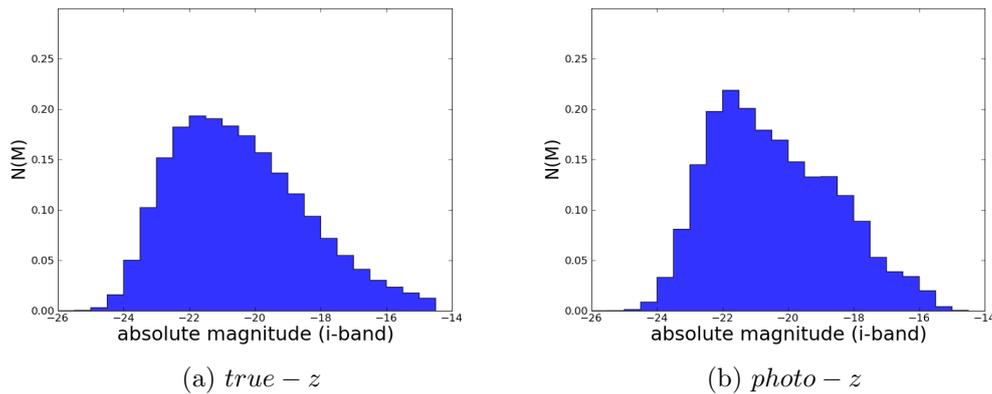


Figura F.10: Distribuição de magnitudes absolutas geradas pelo LePhare para o conjunto fotométrico (com remoção de objetos com  $S/N < 5$ ). À esquerda: magnitudes absolutas geradas a partir dos *redshifts* verdadeiros da simulação. À direita: magnitudes absolutas geradas a partir dos *photo-z*'s calculados com o TPZ.

### F.3.2 Classificação (preliminar)

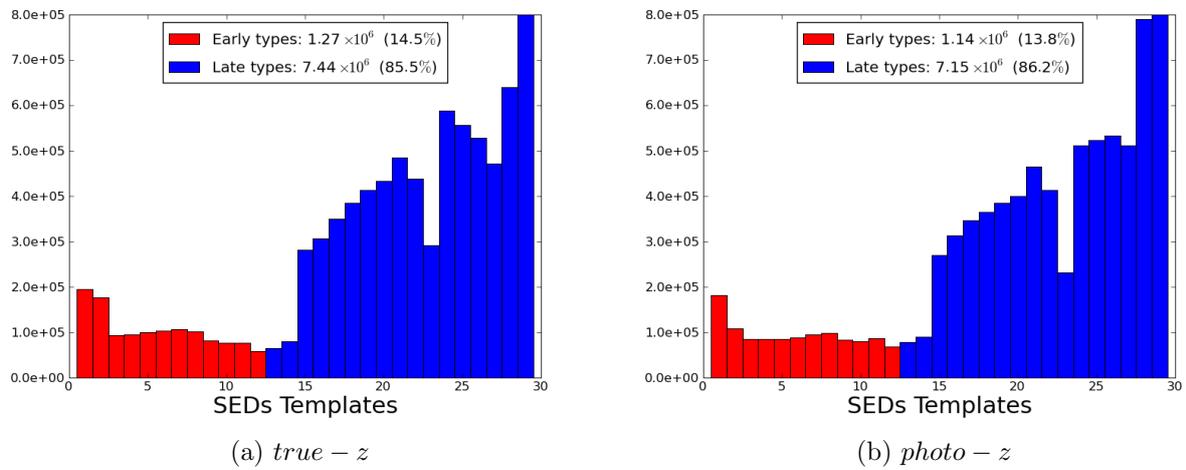


Figura F.11: Distribuição modelos de SEDs ajustados pelo pelo código LePhare. À esquerda: SEDs ajustadas a partir dos *redshifts* verdadeiros da simulação. À direita: SEDs ajustadas a partir dos *photo-z's* calculados com o TPZ.

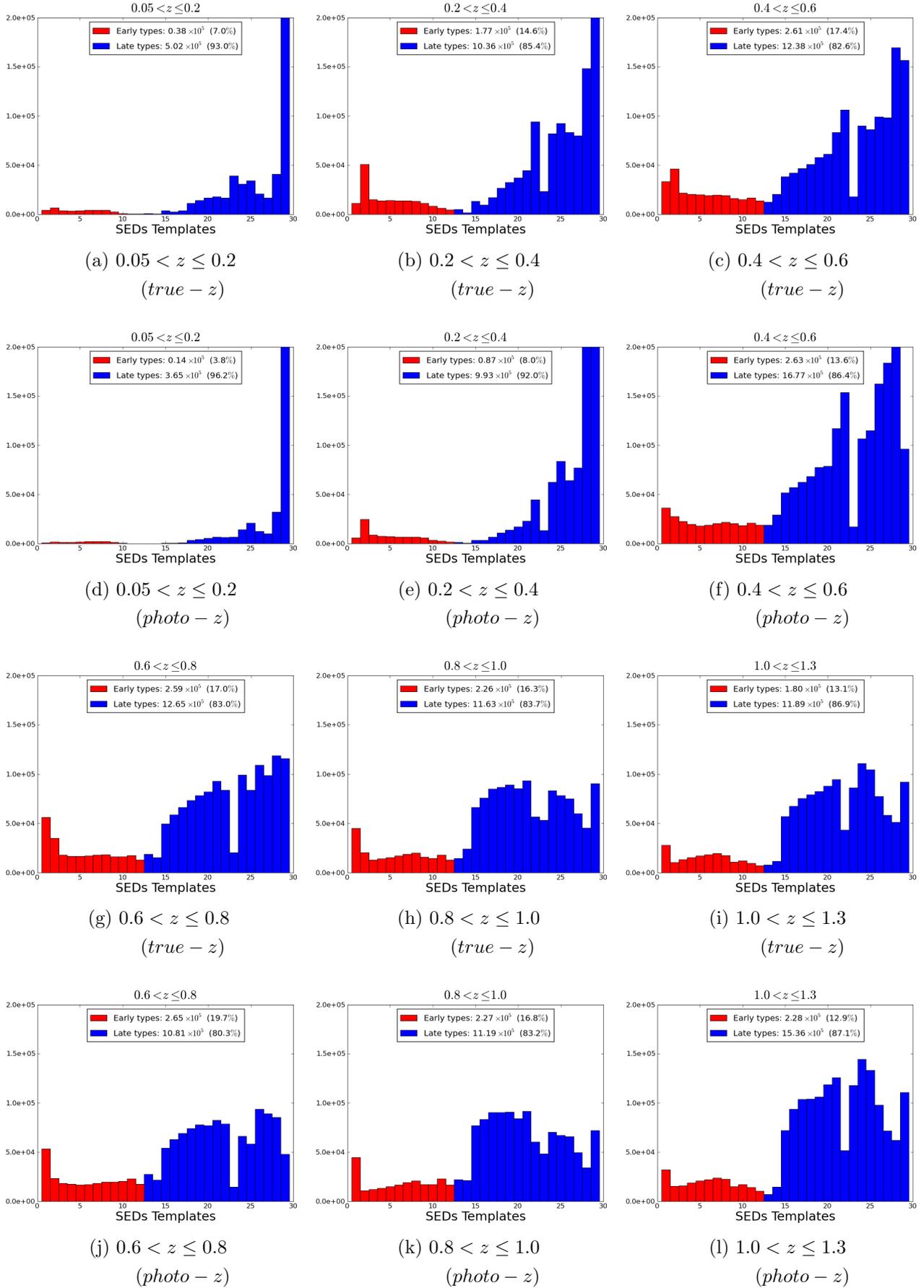
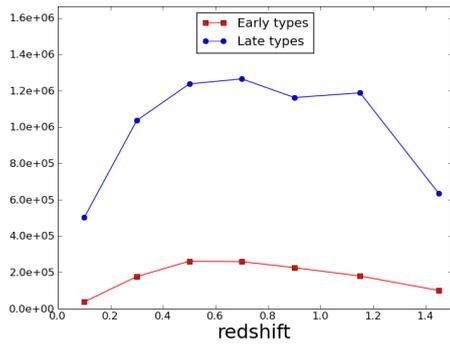
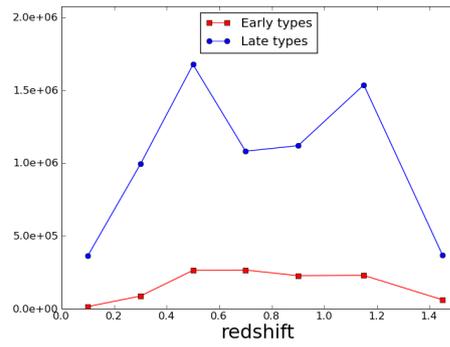


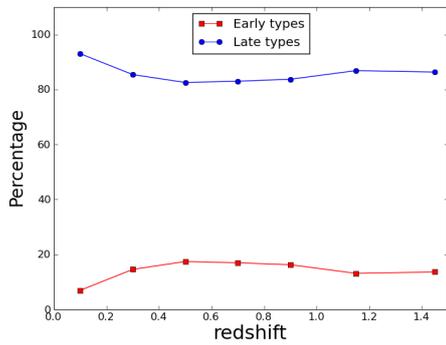
Figura F.12: Distribuição modelos de SEDs ajustados pelo código LePhare, em faixas de *redshift*. Linhas ímpares: SEDs ajustadas a partir dos *redshifts* verdadeiros da simulação. Linhas pares: SEDs ajustadas a partir dos *photo-z's* calculados com o TPZ.



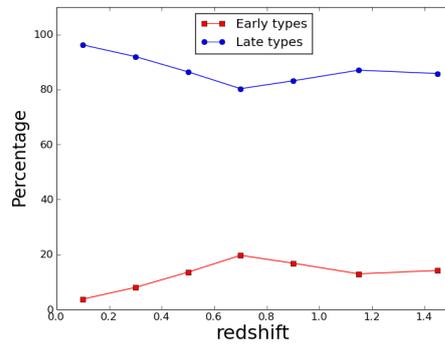
(a) Número (com *true - z*)



(b) Número (com *photo - z*)



(c) Porcentagem (com *true - z*)



(d) Porcentagem (com *photo - z*)

Figura F.13: Evolução da quantidade de modelos ajustados em cada grupo de tipos espectrais. À esquerda: SEDs ajustadas a partir dos *redshifts* verdadeiros da simulação. À direita: SEDs ajustadas a partir dos *photo-z's* calculados com o TPZ.

## F.4 Função de Luminosidade

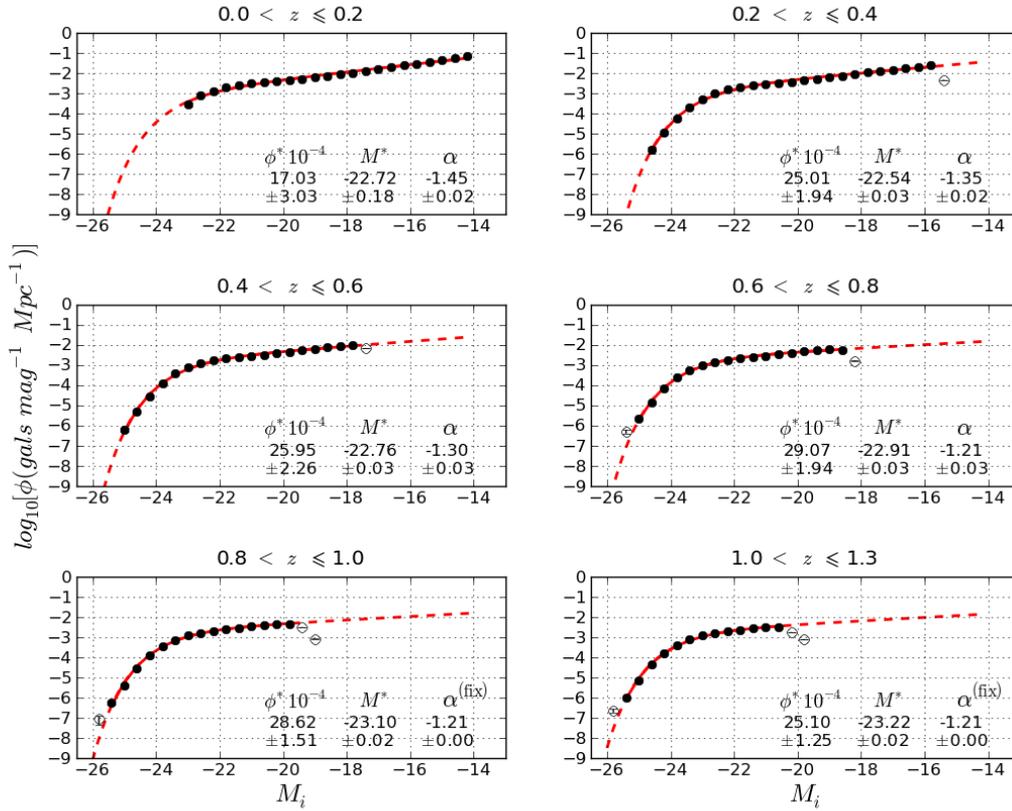


Figura F.14: Função de luminosidade global em intervalos de *redshift*, sendo este o valor verdadeiro, proveniente da simulação.

Tabela F.6: Parametros de Schechter - BCC com *true* -  $z$

Redshift	$\phi^*$	$M^*$	$\alpha$
0.0 a 0.2	$0.001703 \pm 0.000303$	$-22.72 \pm 0.18$	$-1.45 \pm 0.02$
0.2 a 0.4	$0.002501 \pm 0.000194$	$-22.54 \pm 0.03$	$-1.35 \pm 0.02$
0.4 a 0.6	$0.002595 \pm 0.000226$	$-22.76 \pm 0.03$	$-1.30 \pm 0.03$
0.6 a 0.8	$0.002907 \pm 0.000194$	$-22.91 \pm 0.03$	$-1.21 \pm 0.03$
0.8 a 1.0	$0.002862 \pm 0.000151$	$-23.10 \pm 0.02$	$-1.21 \pm 0.00$
1.0 a 1.3	$0.002510 \pm 0.000125$	$-23.22 \pm 0.02$	$-1.21 \pm 0.00$

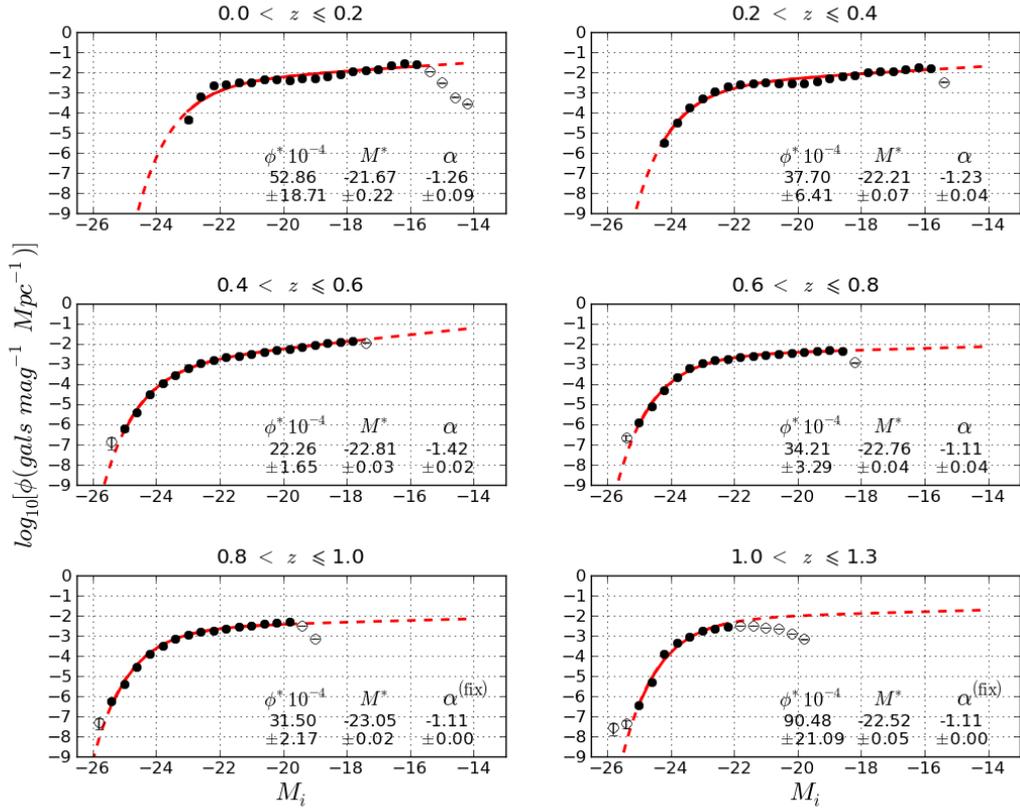


Figura F.15: Função de luminosidade global em intervalos de *redshift*, sendo este o valor do *photo-z* calculado com o TPZ.

Tabela F.7: Parametros de Schechter - BCC com *photo - z*

Redshift	$\phi^*$	$M^*$	$\alpha$
0.0 a 0.2	$0.005286 \pm 0.001871$	$-21.67 \pm 0.22$	$-1.26 \pm 0.09$
0.2 a 0.4	$0.003770 \pm 0.000641$	$-22.21 \pm 0.07$	$-1.23 \pm 0.04$
0.4 a 0.6	$0.002226 \pm 0.000165$	$-22.81 \pm 0.03$	$-1.42 \pm 0.02$
0.6 a 0.8	$0.003421 \pm 0.000329$	$-22.76 \pm 0.04$	$-1.11 \pm 0.04$
0.8 a 1.0	$0.003150 \pm 0.000217$	$-23.05 \pm 0.02$	$-1.11 \pm 0.00$
1.0 a 1.3	$0.009048 \pm 0.002109$	$-22.52 \pm 0.05$	$-1.11 \pm 0.00$

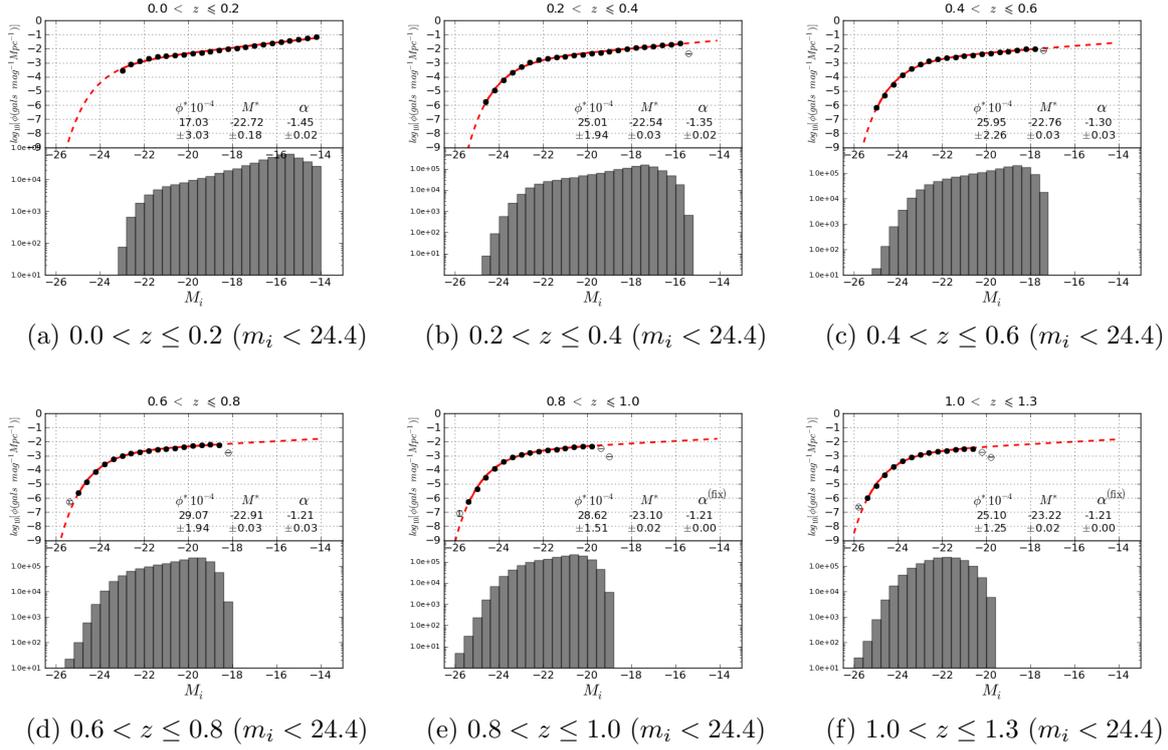


Figura F.16: Função de luminosidade global em intervalos de *redshift*, sendo este o valor verdadeiro, proveniente da simulação. Abaixo do gráfico da função, o histograma de contagens de galáxias por intervalos de magnitude absoluta, em escala logarítmica.

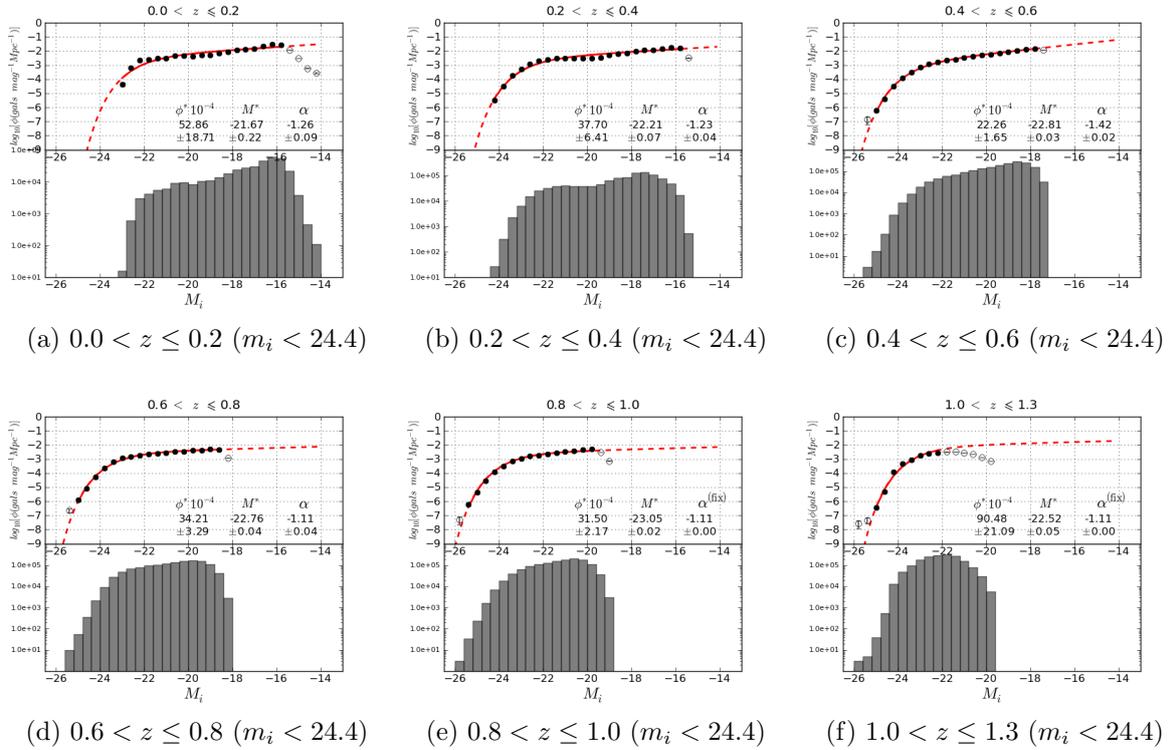


Figura F.17: Função de luminosidade global em intervalos de *redshift*, sendo este o valor do *photo-z* calculado com o TPZ. Abaixo do gráfico da função, o histograma de contagens de galáxias por intervalos de magnitude absoluta, em escala logarítmica.

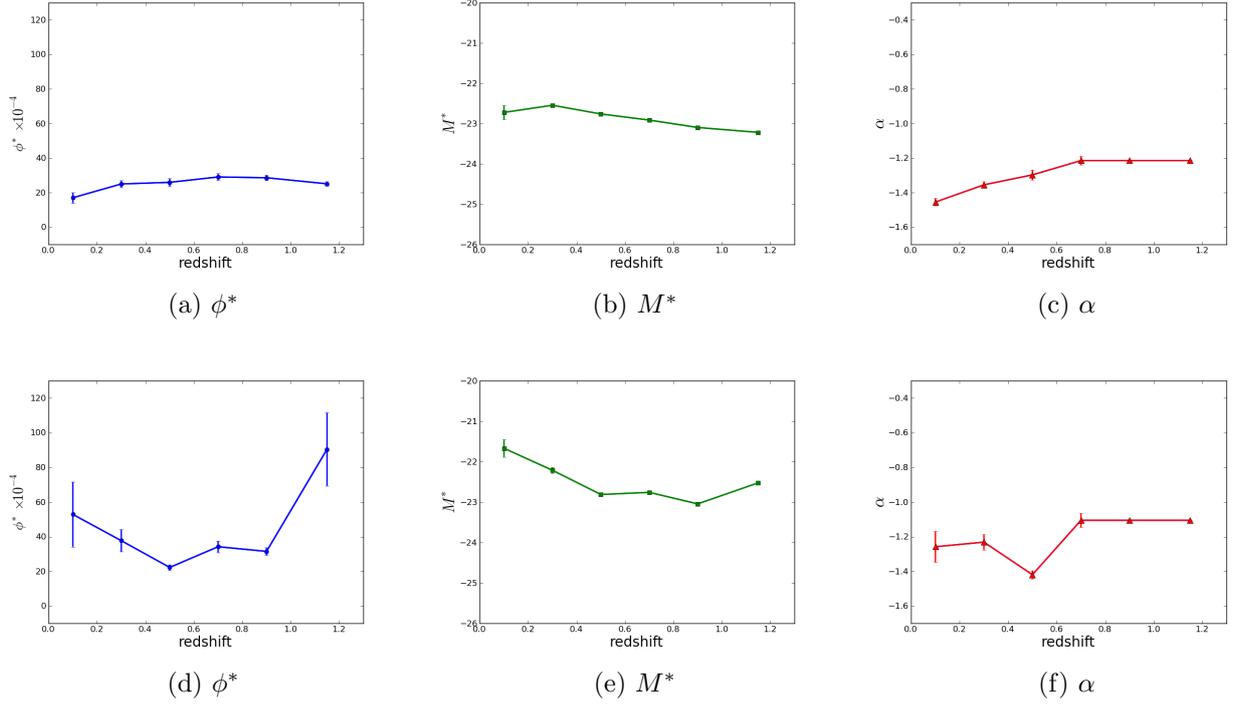


Figura F.18: Evolução dos parâmetros de Schechter. Linha de cima: LF obtida a partir dos *redshifts* verdadeiros da simulação. Linha de baixo: LF obtida a partir dos *photo-z's* calculados com o TPZ.

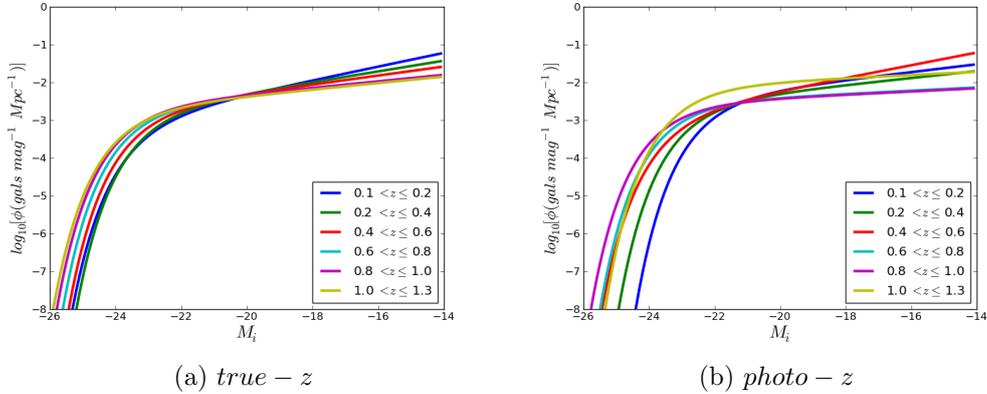


Figura F.19: Evolução da curva da LF ajustada à função de Schechter. À esquerda: LF obtida a partir dos *redshifts* verdadeiros da simulação. À direita: LF obtida a partir dos *photo-z's* calculados com o TPZ.

#### F.4.1 LF por tipos espectrais (preliminar)

Para estudar a dependência do comportamento da LF com o tipo espectral, utilizamos os modelos da biblioteca de SEDs do LePhare, descrita no capítulo 3, para classificar as galáxias em dois grupos: *early-types* (modelos baseados em SEDs de elípticas) e *late-types* (modelos baseados em SEDs de espirais e irregulares), para então calcular as LFs independentes. Como ainda estamos trabalhando nos ajustes da configuração do código e na escolha das SEDs que compõem a biblioteca utilizada, consideramos estes resultados preliminares.

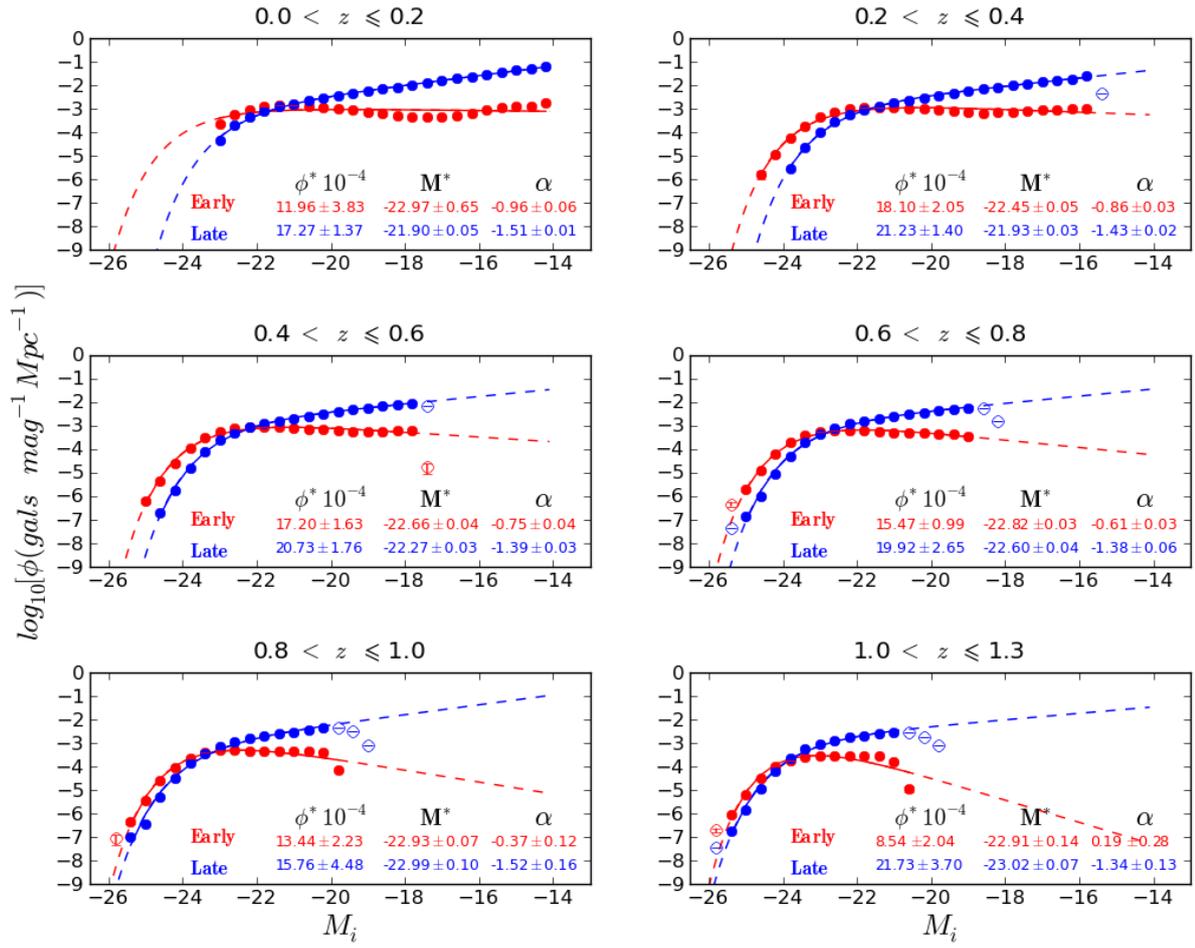


Figura F.20: Função de luminosidade após a divisão da amostra em dois tipos espectrais utilizando o *redshift* verdadeiro, proveniente da simulação.

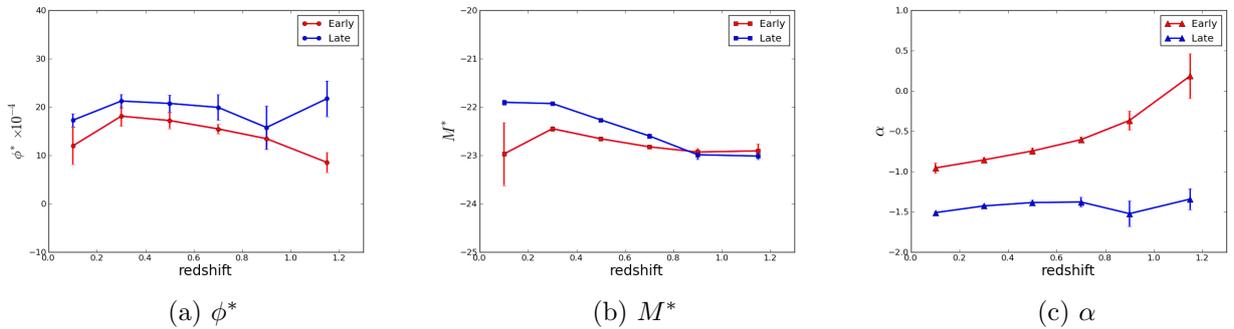


Figura F.21: Evolução dos parâmetros de Schechter da LF obtida a partir dos *redshifts* verdadeiros da simulação

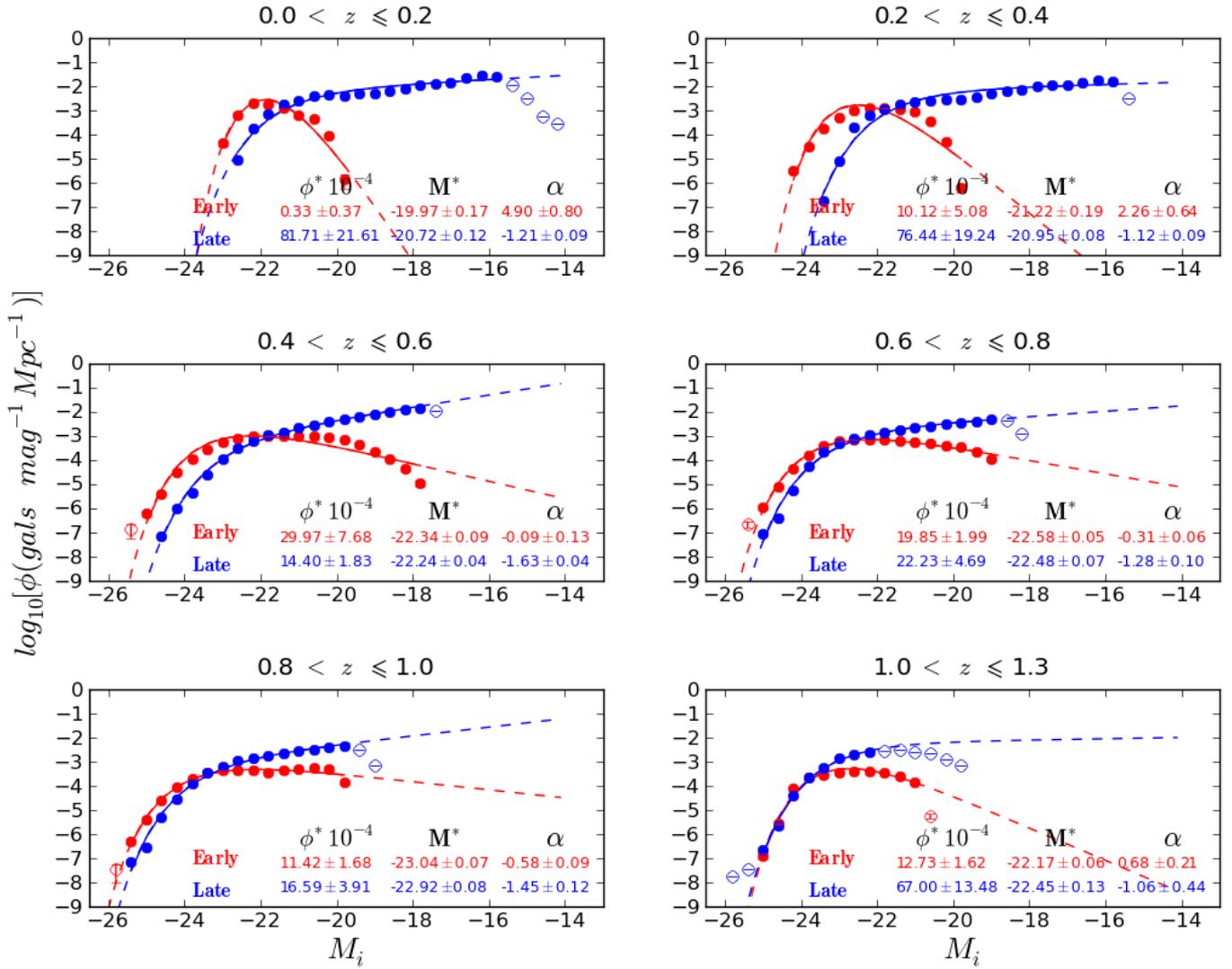


Figura F.22: Função de luminosidade após a divisão da amostra em dois tipos espectrais utilizando o *photo-z* calculado pelo TPZ.

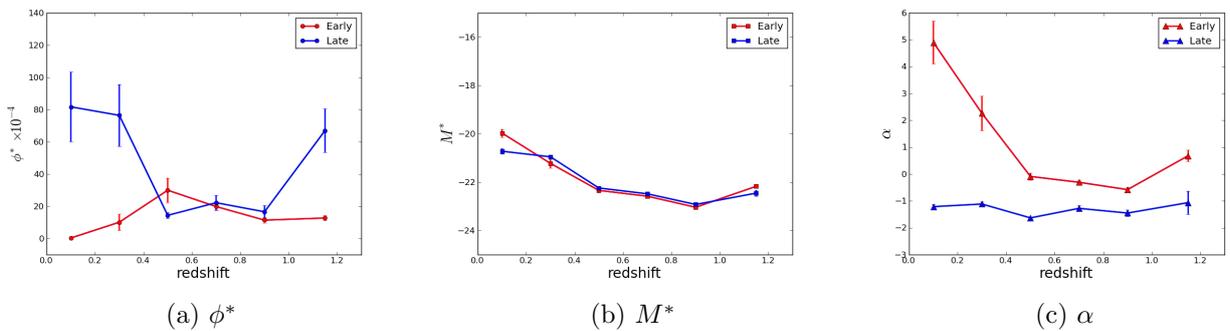


Figura F.23: Evolução dos parâmetros de Schechter da LF obtida a partir dos *photo-z*'s calculados com o TPZ.

# Apêndice G

## Resultados com a Amostra SVA1

### G.1 Validação do *photo-z*

#### G.1.1 Conjunto de Treinamento OzDES

Tabela G.1: Conjunto de treinamento: OzDES. Método: TPZ mode (*photo-z* = pico da PDF(*z*))

Métrica	amostra completa	$S/N > 3$	$S/N > 5$	$S/N > 10$
$\overline{\Delta z}$	$0.0003 \pm 0.0007$	$0.0001 \pm 0.0006$	$0.0007 \pm 0.0005$	$-0.0008 \pm 0.0004$
$\Delta z_{50}$	$0.0027 \pm 0.0001$	$0.0036 \pm 0.0001$	$0.0025 \pm 0.0001$	$0.0021 \pm 0.0002$
$\sigma_{\Delta z}$	$0.1676 \pm 0.0016$	$0.1560 \pm 0.0016$	$0.1404 \pm 0.0015$	$0.1269 \pm 0.0018$
$\sigma_{68}$	$0.0359 \pm 0.0003$	$0.0355 \pm 0.0002$	$0.0339 \pm 0.0002$	$0.0279 \pm 0.0001$
$out_{2\sigma}$	$0.0539 \pm 0.0007$	$0.0509 \pm 0.0008$	$0.0457 \pm 0.0005$	$0.0354 \pm 0.0006$
$out_{3\sigma}$	$0.0279 \pm 0.0005$	$0.0304 \pm 0.0005$	$0.0248 \pm 0.0004$	$0.0221 \pm 0.0005$
$\overline{\Delta z'}$	$0.0296 \pm 0.0026$	$0.0296 \pm 0.0024$	$0.0331 \pm 0.0023$	$0.0238 \pm 0.0024$
$\sigma_{\Delta z'}$	$0.6671 \pm 0.0068$	$0.6598 \pm 0.0079$	$0.6378 \pm 0.0076$	$0.6217 \pm 0.0096$
$N_{poisson}$	$3.7521 \pm 0.0732$	$4.2557 \pm 0.0647$	$3.5907 \pm 0.0567$	$2.6808 \pm 0.0456$
$KS$	$0.0458 \pm 0.0011$	$0.0558 \pm 0.0011$	$0.0502 \pm 0.0010$	$0.0413 \pm 0.0010$

Tabela G.2: Conjunto de treinamento: OzDES. Método: TPZ mean ( $photo-z = \text{media da PDF}(z)$ )

Métrica	amostra completa	$S/N > 3$	$S/N > 5$	$S/N > 10$
$\overline{\Delta z}$	$0.0123 \pm 0.0006$	$0.0082 \pm 0.0006$	$0.0037 \pm 0.0005$	$0.0004 \pm 0.0004$
$\Delta z_{50}$	$0.0064 \pm 0.0002$	$0.0051 \pm 0.0002$	$0.0044 \pm 0.0001$	$0.0037 \pm 0.0001$
$\sigma_{\Delta z}$	$0.1535 \pm 0.0009$	$0.1543 \pm 0.0013$	$0.1262 \pm 0.0010$	$0.1057 \pm 0.0012$
$\sigma_{68}$	$0.0718 \pm 0.0004$	$0.0666 \pm 0.0004$	$0.0550 \pm 0.0003$	$0.0382 \pm 0.0002$
$out_{2\sigma}$	$0.0590 \pm 0.0005$	$0.0514 \pm 0.0011$	$0.0499 \pm 0.0006$	$0.0399 \pm 0.0006$
$out_{3\sigma}$	$0.0301 \pm 0.0005$	$0.0258 \pm 0.0004$	$0.0251 \pm 0.0004$	$0.0187 \pm 0.0004$
$\overline{\Delta z'}$	$0.0714 \pm 0.0023$	$0.0526 \pm 0.0019$	$0.0507 \pm 0.0020$	$0.0360 \pm 0.0018$
$\sigma_{\Delta z'}$	$0.5933 \pm 0.0024$	$0.5828 \pm 0.0030$	$0.5690 \pm 0.0037$	$0.5057 \pm 0.0040$
$N_{poisson}$	$4.8761 \pm 0.0734$	$3.9674 \pm 0.0702$	$4.3745 \pm 0.0798$	$4.0086 \pm 0.0640$
$KS$	$0.0507 \pm 0.0009$	$0.0447 \pm 0.0007$	$0.0483 \pm 0.0008$	$0.0435 \pm 0.0008$

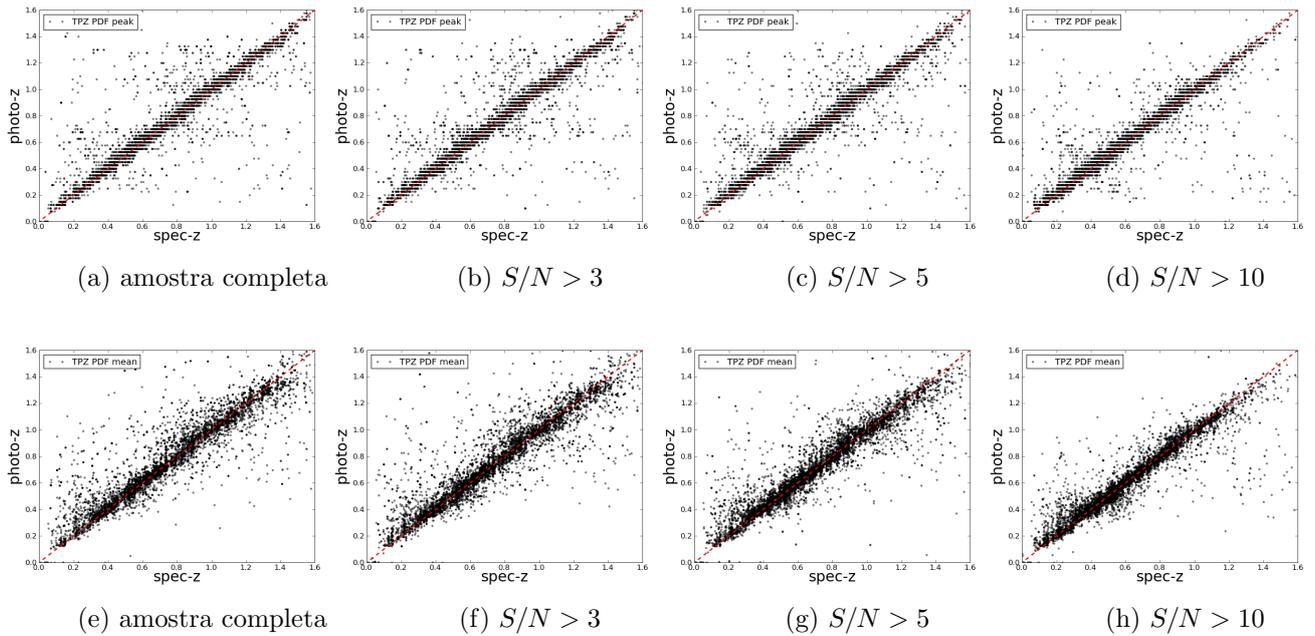


Figura G.1: Gráfico de dispersão da relação entre  $photo-z$  e  $spec-z$ , para os diferentes seleções na amostra, baseadas na razão sinal-ruído. Resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto **OzDES**. Linha superior: **TPZ mode**; linha inferior: **TPZ mean**.

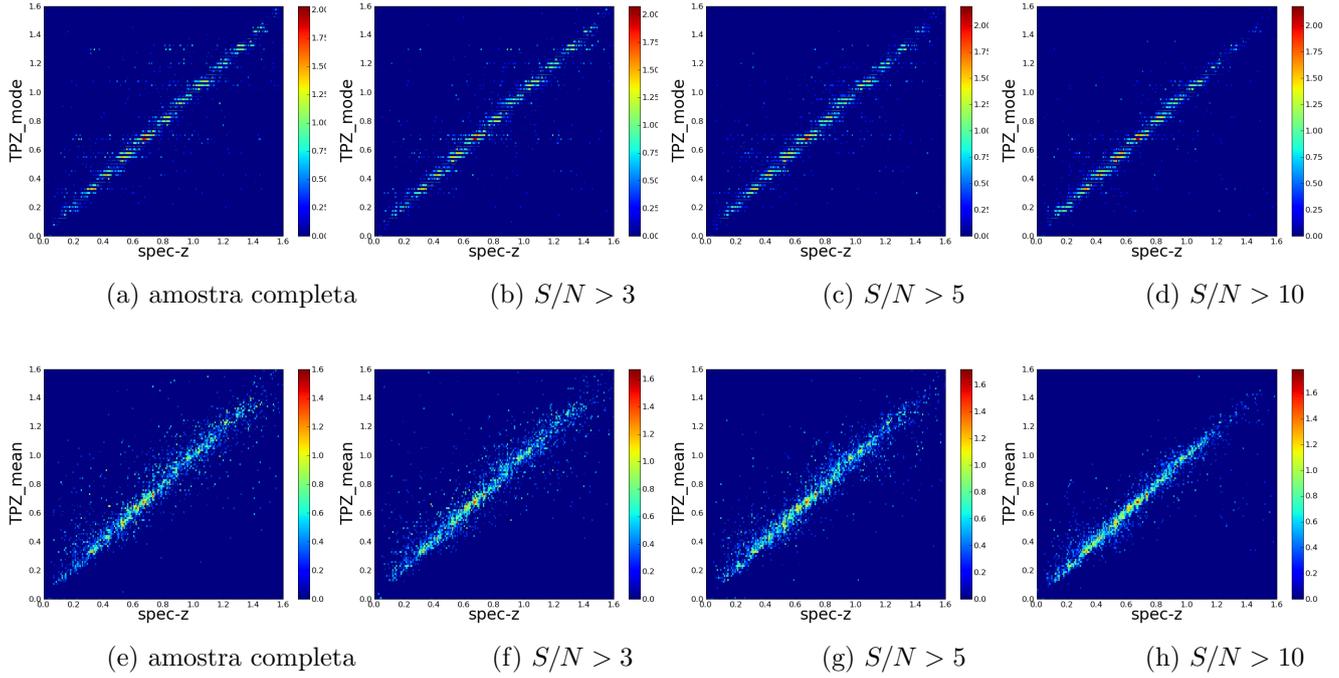


Figura G.2: Gráfico de densidades da dispersão da relação entre *photo-z* e *spec-z*, para os diferentes seleções na amostra, baseadas na razão sinal-ruído. Resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto **OzDES**. A escala de cores representa a densidade de pontos em cada região do gráfico, em escala logarítmica. Linha superior: **TPZ mode**; linha inferior: **TPZ mean**.

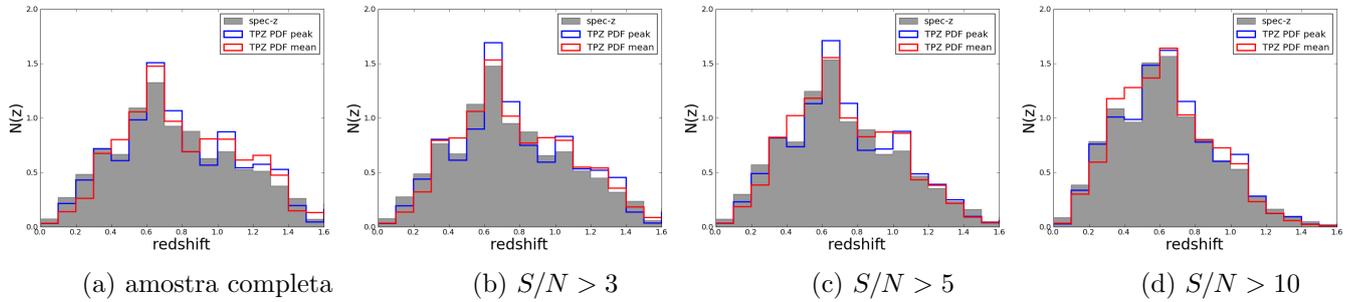


Figura G.3: Gráfico da distribuição de *redshifts* obtidos com o TPZ, em comparação com o *spec-z*, para os diferentes seleções na amostra, baseadas na razão sinal-ruído. Resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto **OzDES**. Em cinza: *true-z*; em azul: **TPZ mode**; em vermelho: **TPZ mean**.

## G.1.2 Conjunto de Treinamento OzDES ponderado (representativo)

Tabela G.3: Conjunto de treinamento: OzDES ponderado. Método: TPZ mode ( $photo-z$  = pico da PDF( $z$ ))

Métrica	amostra completa	$S/N > 3$	$S/N > 5$	$S/N > 10$
$\overline{\Delta z}$	0.0002±0.0002	0.0002±0.0002	0.0006±0.0002	0.0006±0.0002
$\Delta z_{50}$	0.0003±0.0001	0.0004±0.0001	0.0007±0.0001	0.0008±0.0001
$\sigma_{\Delta z}$	0.0642±0.0020	0.0495±0.0017	0.0557±0.0018	0.0431±0.0015
$\sigma_{68}$	0.0171±0.0001	0.0167±0.0001	0.0164±0.0001	0.0160±0.0001
$out_{2\sigma}$	0.0139±0.0005	0.0106±0.0004	0.0092±0.0003	0.0073±0.0003
$out_{3\sigma}$	0.0112±0.0004	0.0091±0.0003	0.0073±0.0003	0.0058±0.0003
$\overline{\Delta z'}$	0.0167±0.0012	0.0105±0.0012	0.0159±0.0009	0.0142±0.0009
$\sigma_{\Delta z'}$	0.3356±0.0086	0.3138±0.0088	0.2641±0.0033	0.2546±0.0041
$N_{poisson}$	0.9259±0.0389	0.6140±0.0360	0.6718±0.0287	0.7659±0.0282
$KS$	0.0396±0.0008	0.0433±0.0008	0.0413±0.0006	0.0464±0.0009

Tabela G.4: Conjunto de treinamento: OzDES ponderado. Método: TPZ mean ( $photo-z$  = média da PDF( $z$ ))

Métrica	amostra completa	$S/N > 3$	$S/N > 5$	$S/N > 10$
$\overline{\Delta z}$	0.0029±0.0002	0.0023±0.0002	0.0023±0.0002	0.0023±0.0001
$\Delta z_{50}$	0.0024±0.0001	0.0025±0.0001	0.0029±0.0001	0.0027±0.0001
$\sigma_{\Delta z}$	0.0636±0.0011	0.0537±0.0011	0.0520±0.0011	0.0421±0.0008
$\sigma_{68}$	0.0183±0.0001	0.0177±0.0000	0.0169±0.0001	0.0161±0.0001
$out_{2\sigma}$	0.0281±0.0009	0.0251±0.0006	0.0225±0.0006	0.0221±0.0005
$out_{3\sigma}$	0.0148±0.0004	0.0157±0.0004	0.0136±0.0005	0.0138±0.0003
$\overline{\Delta z'}$	0.0434±0.0012	0.0365±0.0010	0.0366±0.0009	0.0384±0.0009
$\sigma_{\Delta z'}$	0.3218±0.0039	0.2938±0.0034	0.2664±0.0021	0.2463±0.0015
$N_{poisson}$	1.5059±0.0475	0.8380±0.0354	0.7946±0.0301	1.0369±0.0342
$KS$	0.0340±0.0008	0.0359±0.0008	0.0372±0.0008	0.0418±0.0009

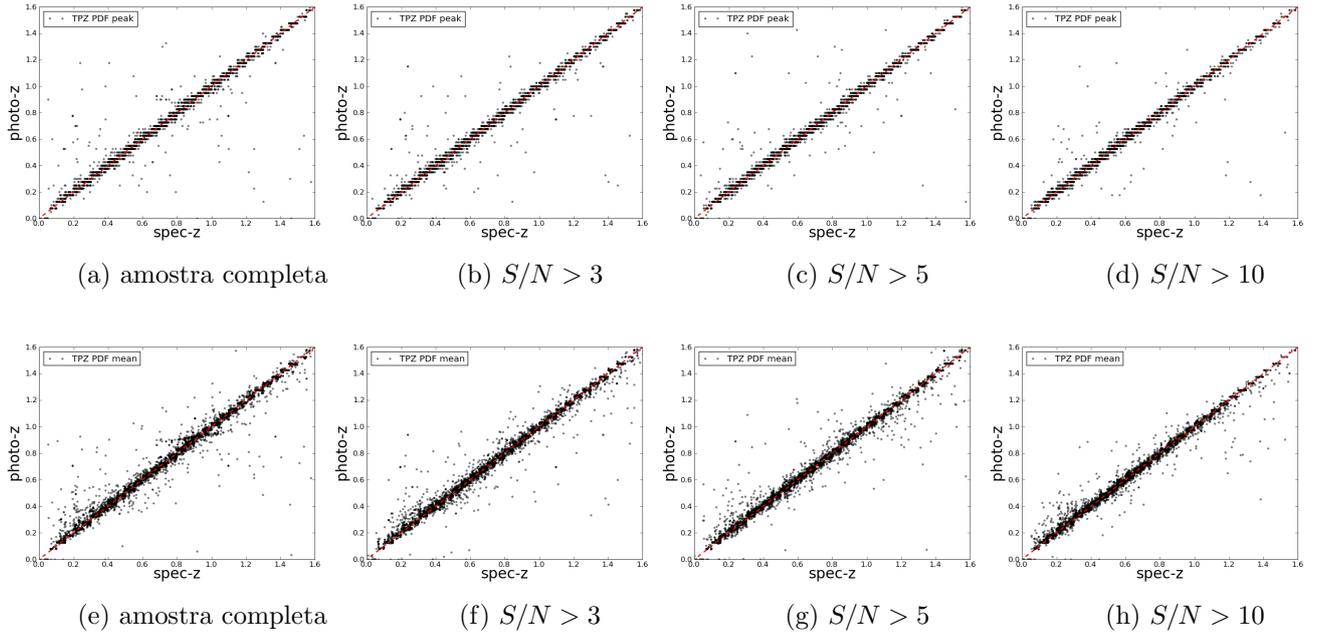


Figura G.4: Gráfico de dispersão da relação entre  $photo-z$  e  $spec-z$ , para os diferentes seleções na amostra, baseadas na razão sinal-ruído. Resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto **OzDES ponderado** (representativo). Linha superior: **TPZ mode**; linha inferior: **TPZ mean**.

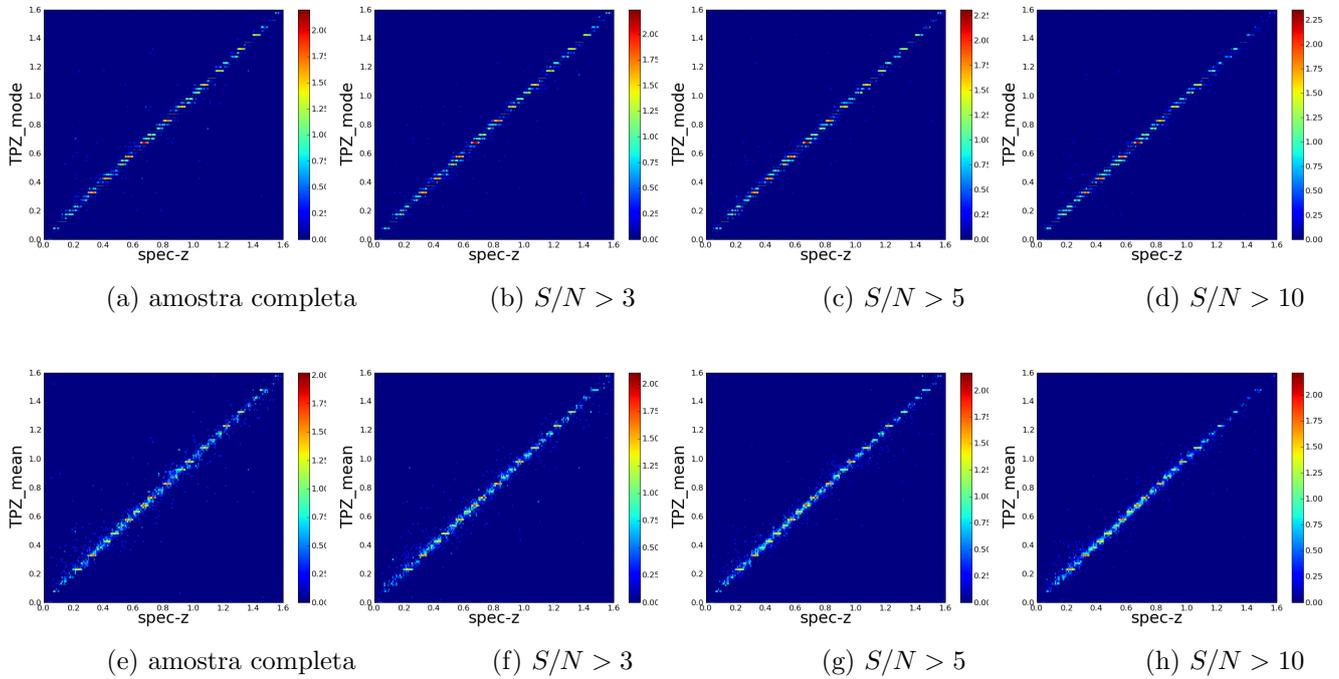


Figura G.5: Gráfico de densidades da dispersão da relação entre  $photo-z$  e  $spec-z$ , para os diferentes seleções na amostra, baseadas na razão sinal-ruído. Resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com o conjunto **OzDES ponderado** (representativo). A escala de cores representa a densidade de pontos em cada região do gráfico, em escala logarítmica. Linha superior: **TPZ mode**; linha inferior: **TPZ mean**.

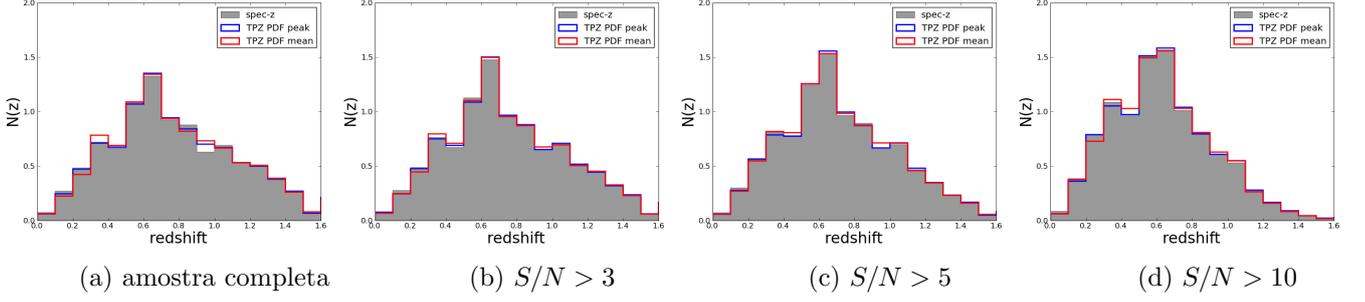


Figura G.6: Gráfico da distribuição de *redshifts* obtidos com o TPZ, em comparação com o *true-z* (*redshift* original da simulação), para os diferentes seleções na amostra, baseadas na razão sinal-ruído. Resultados obtidos para o conjunto de validação, após treinamento ter sido realizado com conjunto **OzDES ponderado** (representativo). Em cinza: *true-z*; em azul: **TPZ mode**; em vermelho: **TPZ mean**.

Tabela G.5: Comparação baseada na métrica  $\sigma_{68}$ , entre as performances do TPZ, após diferentes seleções em razão sinal-ruído.

Conjunto	metodo	amostra completa	$S/N > 3$	$S/N > 5$	$S/N > 10$
OzDES	mode	$0.0359 \pm 0.0003$	$0.0355 \pm 0.0002$	$0.0339 \pm 0.0002$	$0.0279 \pm 0.0001$
OzDES	mean	$0.0718 \pm 0.0004$	$0.0666 \pm 0.0004$	$0.0550 \pm 0.0003$	$0.0382 \pm 0.0002$
ponderado	mode	$0.0171 \pm 0.0001$	$0.0167 \pm 0.0001$	$0.0164 \pm 0.0001$	$0.0160 \pm 0.0001$
ponderado	mean	$0.0183 \pm 0.0001$	$0.0177 \pm 0.0000$	$0.0169 \pm 0.0001$	$0.0161 \pm 0.0001$

## G.2 *Photo-z* do conjunto fotométrico

### G.2.1 Conjunto de Treinamento OzDES

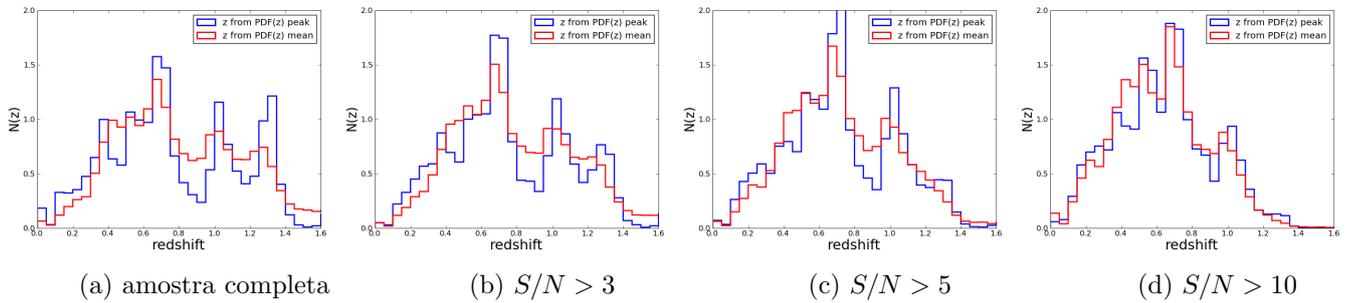


Figura G.7: Resultado do TPZ em um conjunto de 100 mil galáxias aleatoriamente selecionadas do conjunto fotométrico. Árvores produzidas com o conjunto OzDES.

## G.2.2 Conjunto de Treinamento OzDES ponderado (representativo)

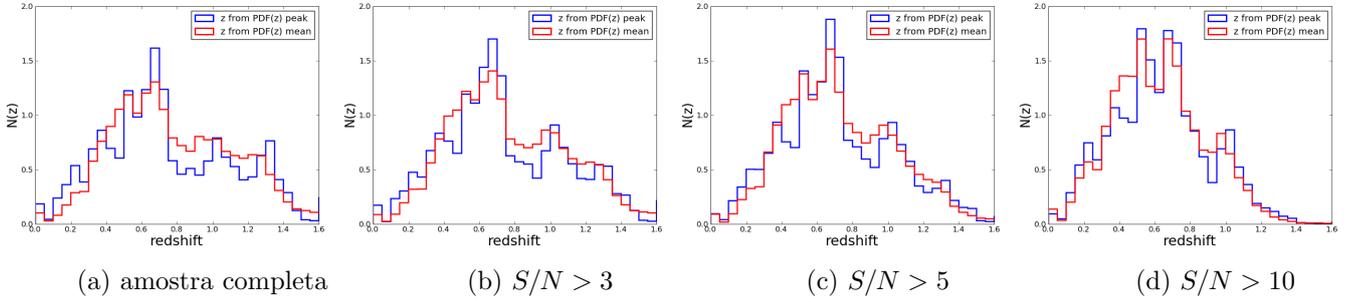


Figura G.8: Resultado do TPZ em um conjunto de 100 mil galáxias aleatoriamente selecionadas do conjunto fotométrico. Árvores produzidas com o conjunto OzDES ponderado (representativo).

## G.2.3 *Photo-z* da amostra fotométrica completa

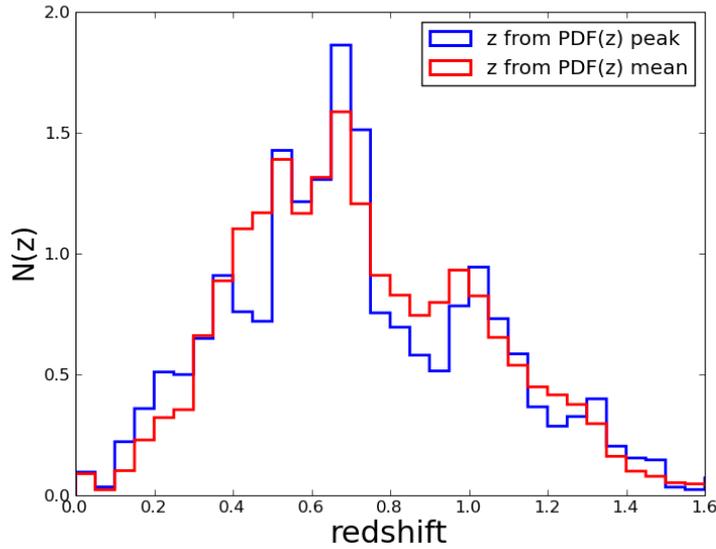
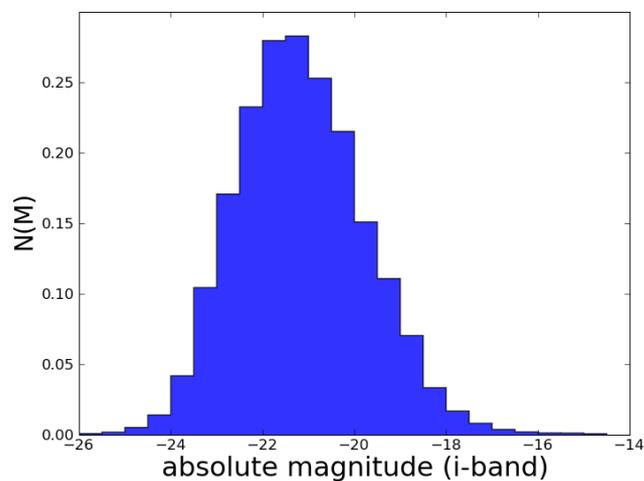


Figura G.9: Resultado do TPZ para todas as galáxias que serão utilizadas na determinação da função de luminosidade (seleção de objetos com  $S/N > 5$ ). Árvores produzidas com o conjunto OzDES.

## G.3 Produtos do LePhare

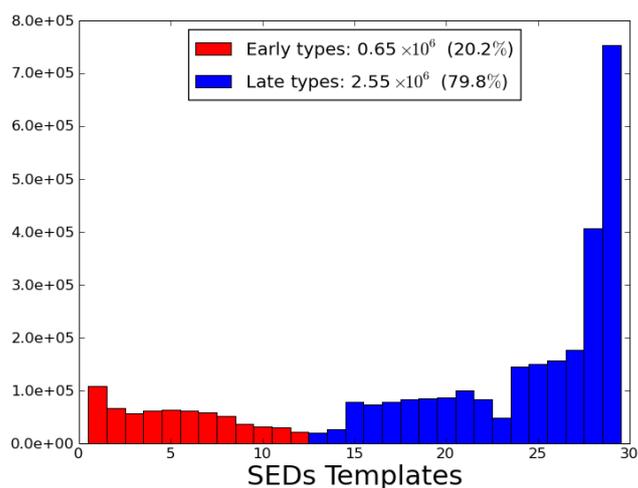
### G.3.1 Magnitudes Absolutas



(a) *photo - z*

Figura G.10: Distribuição de magnitudes absolutas geradas pelo LePhare para o conjunto fotométrico com seleção de objetos com  $S/N > 5$ , a partir dos *photo-z*'s calculados com o TPZ.

### G.3.2 Classificação (preliminar)



(a) *photo - z*

Figura G.11: Distribuição modelos de SEDs ajustados pelo código LePhare. SEDs ajustadas a partir dos *photo-z*'s calculados com o TPZ.

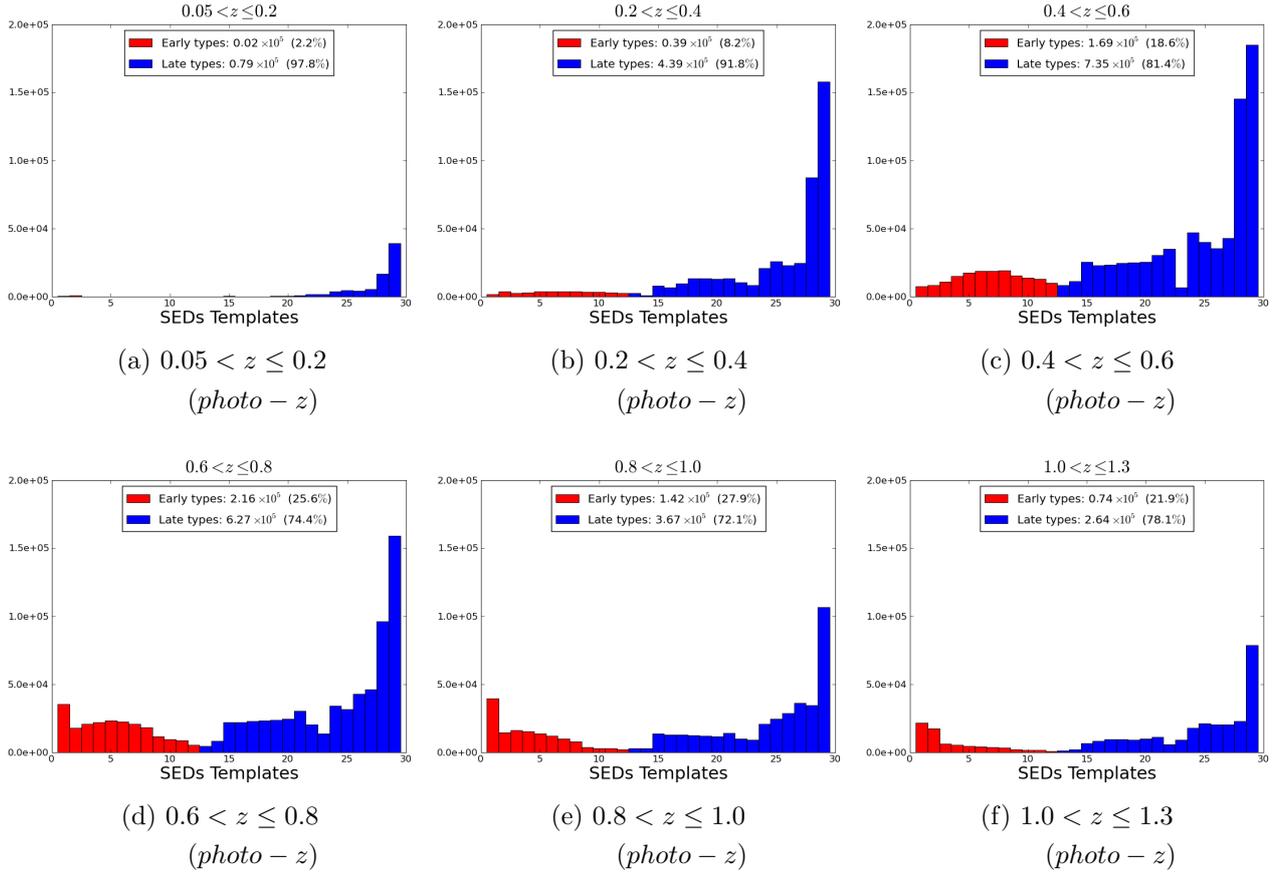


Figura G.12: Distribuição modelos de SEDs ajustados pelo código LePhare, em faixas de *redshift*. SEDs ajustadas a partir dos *photo-z's* calculados com o TPZ.

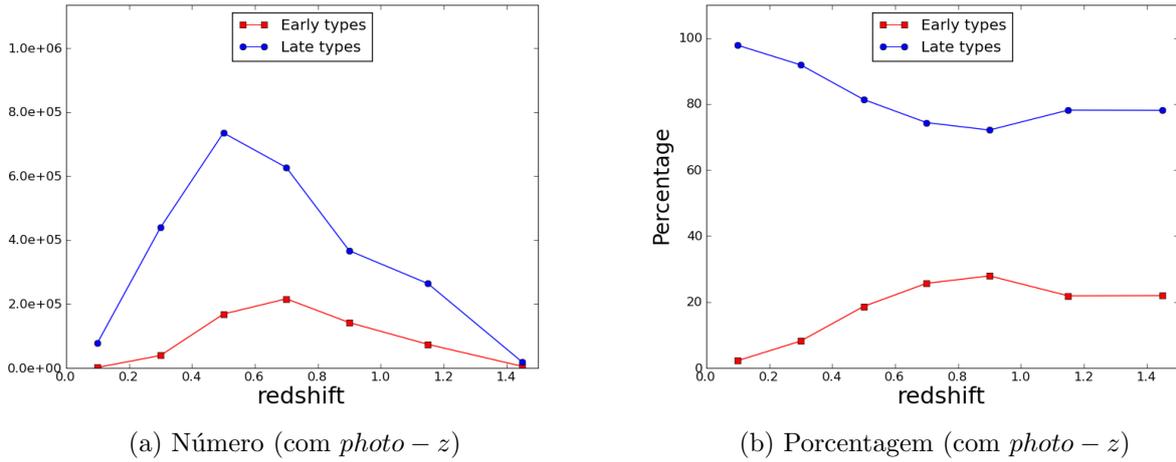


Figura G.13: Evolução da quantidade de modelos ajustados em cada grupo de tipos espectrais. SEDs ajustadas a partir dos *photo-z's* calculados com o TPZ.

## G.4 Função de Luminosidade

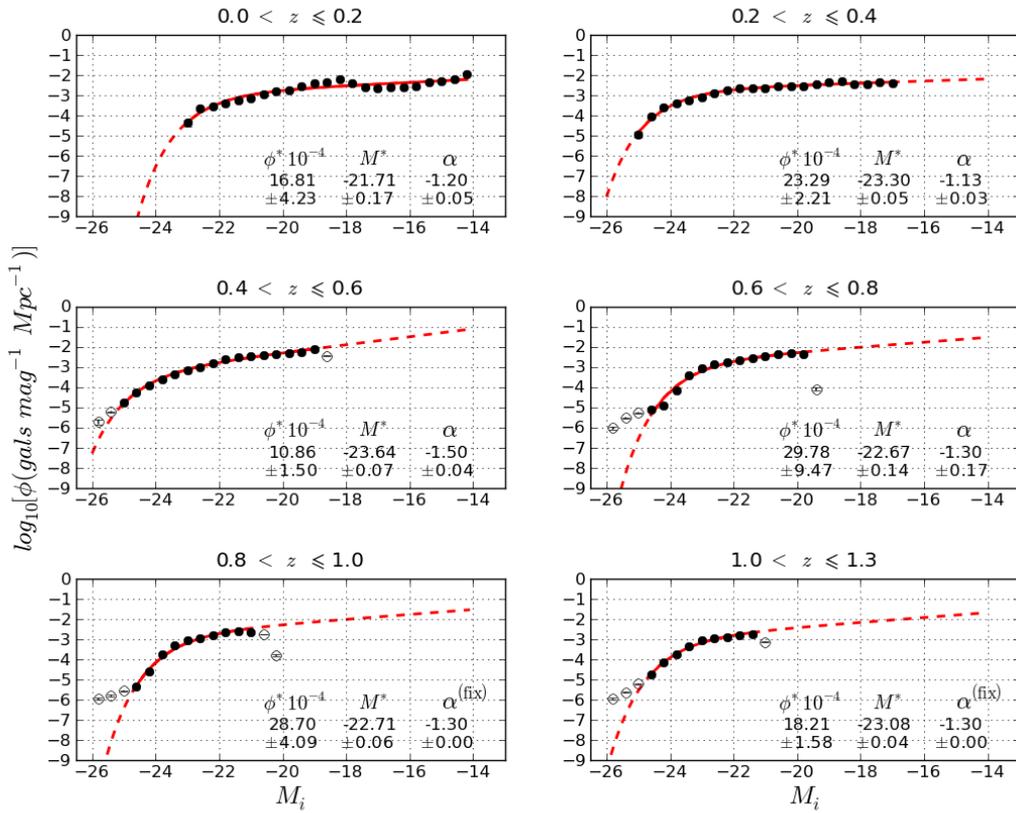
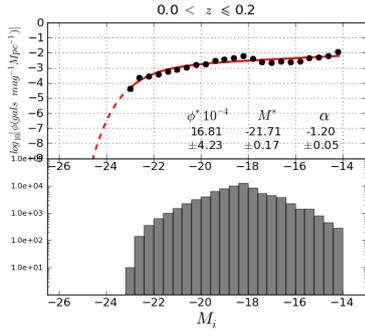
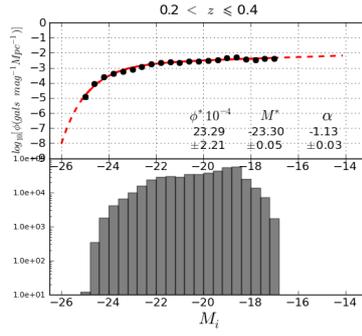


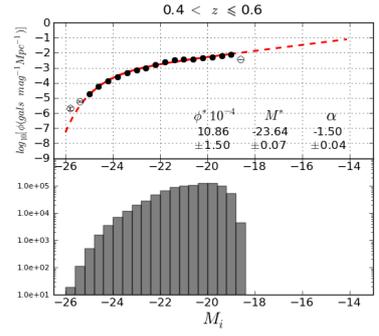
Figura G.14: Função de luminosidade global em intervalos de *redshift*, sendo este o valor do *photo-z* calculado com o TPZ.



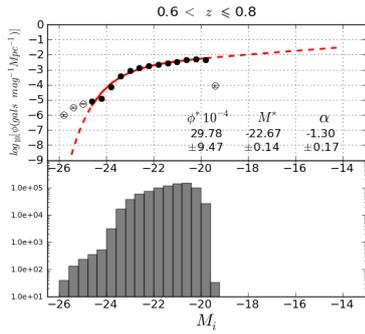
(a)  $0.0 < z \leq 0.2$  ( $m_i < 24.4$ )



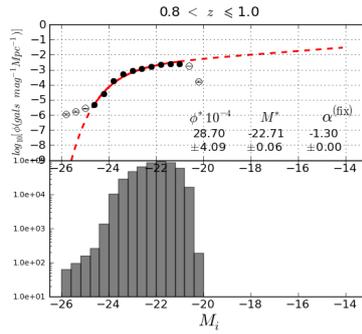
(b)  $0.2 < z \leq 0.4$  ( $m_i < 24.4$ )



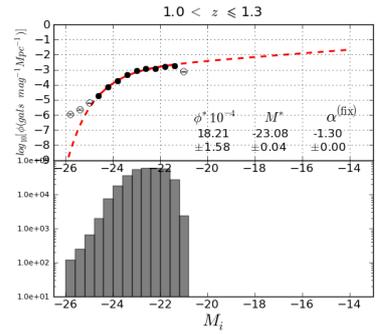
(c)  $0.4 < z \leq 0.6$  ( $m_i < 24.4$ )



(d)  $0.6 < z \leq 0.8$  ( $m_i < 24.4$ )



(e)  $0.8 < z \leq 1.0$  ( $m_i < 24.4$ )



(f)  $1.0 < z \leq 1.3$  ( $m_i < 24.4$ )

Figura G.15: Função de luminosidade global em intervalos de *redshift*, sendo este o valor do *photo-z* calculado com o TPZ e histograma com contagem de galáxias em intervalos de magnitude absoluta, em escala logarítmica.

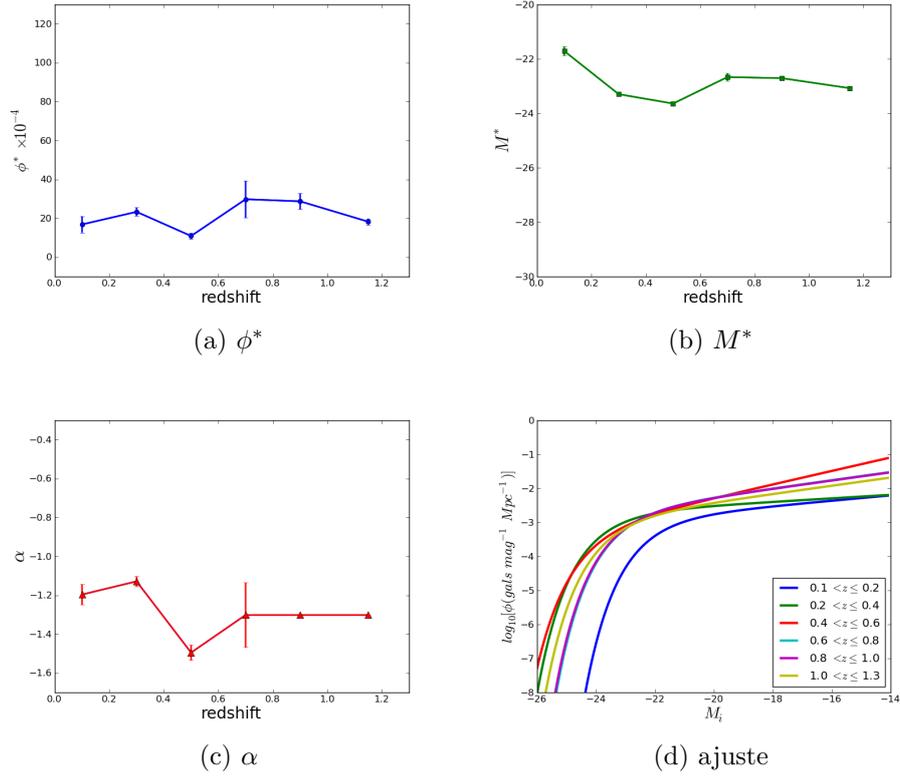


Figura G.16: Evolução dos parâmetros e da curva ajustada à função de Schechter. LF obtida a partir dos *photo-z's* calculados com o TPZ.

## G.4.1 Amostra contaminada com estrelas

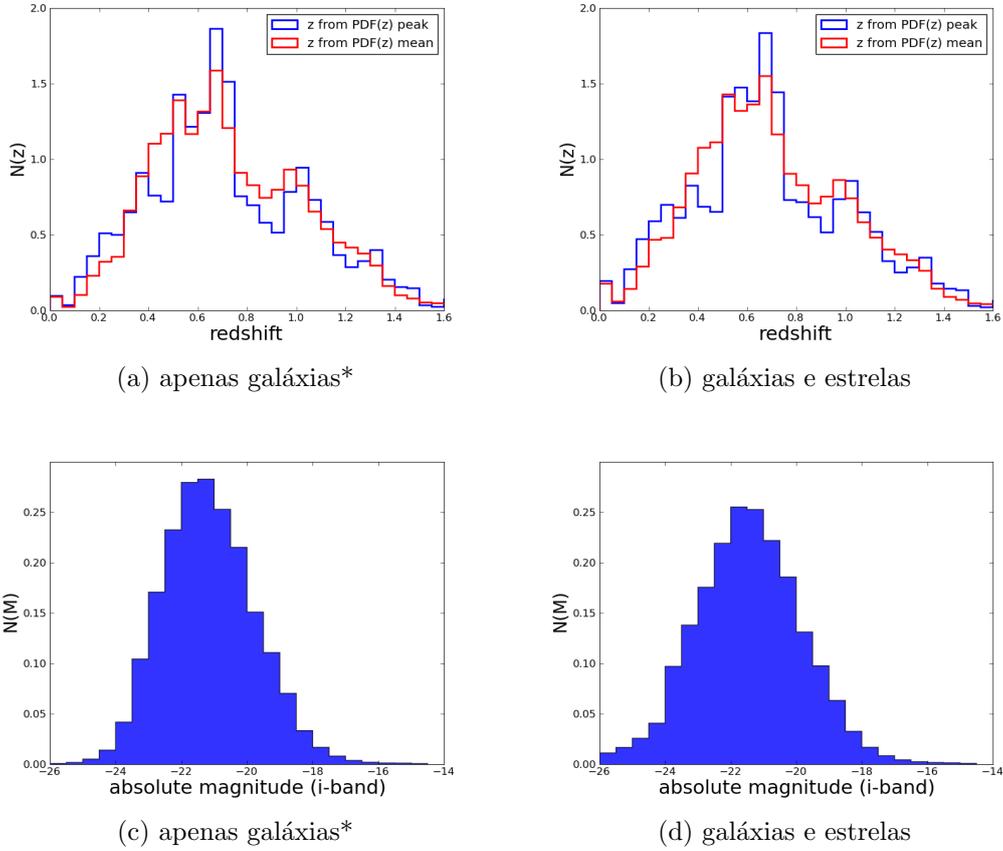


Figura G.17: Comparação entre as distribuições de *photo-z*'s e de magnitudes absolutas, para todas as galáxias que serão utilizadas na determinação da função de luminosidade (seleção em  $S/N > 5$ ). À esquerda, amostra após a remoção dos objetos classificados como estrelas. À direita, antes da remoção. \*: a remoção dos objetos não garante 100% de pureza na amostra.

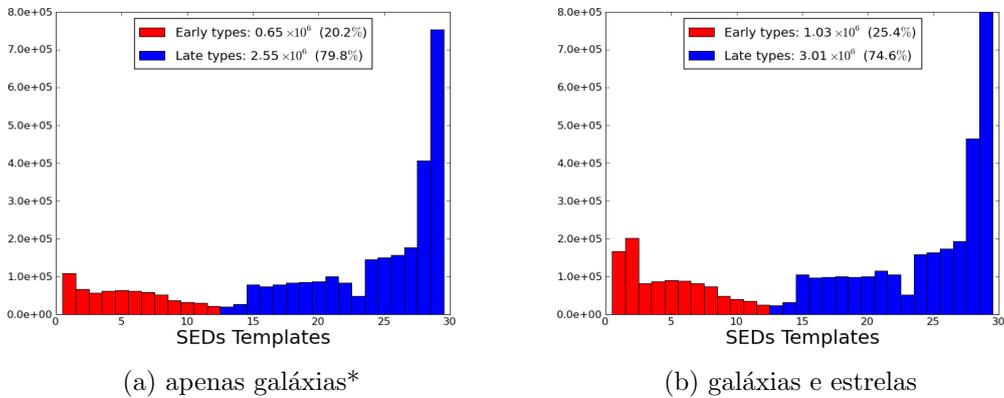


Figura G.18: Distribuição modelos de SEDs ajustados pelo código LePhare. SEDs ajustadas a partir dos *photo-z*'s calculados com o TPZ. À esquerda, amostra após a remoção dos objetos classificados como estrelas. À direita, antes da remoção. \*: a remoção dos objetos não garante 100% de pureza na amostra.

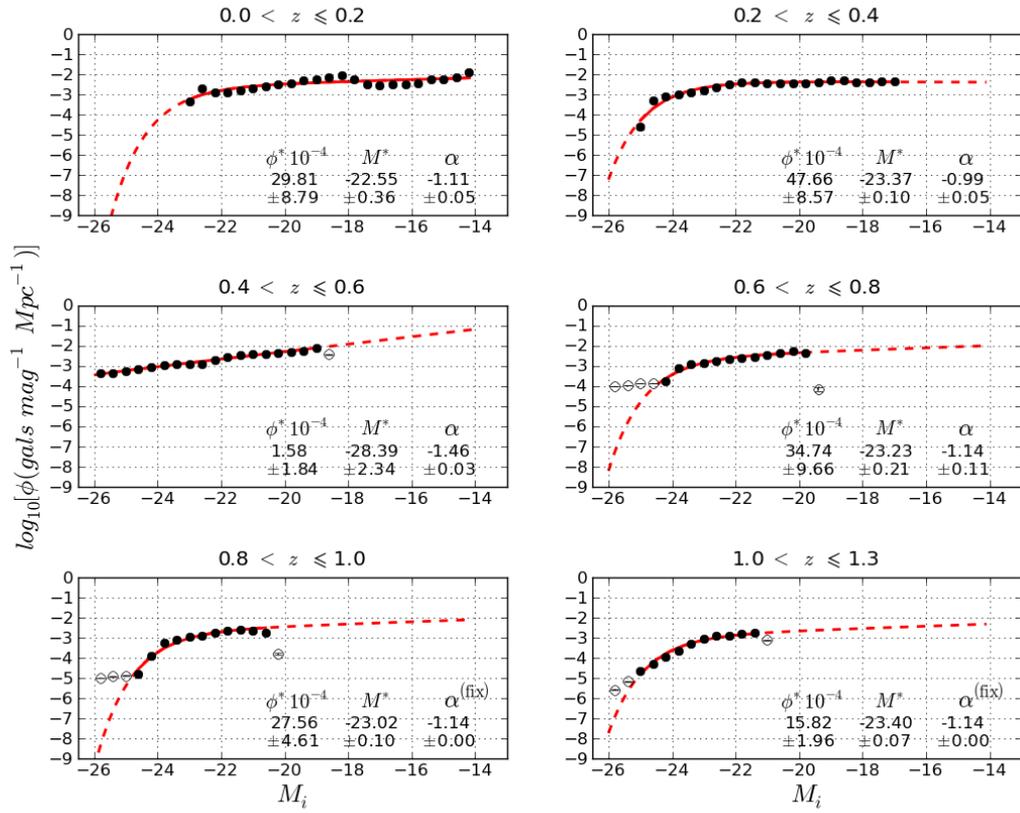
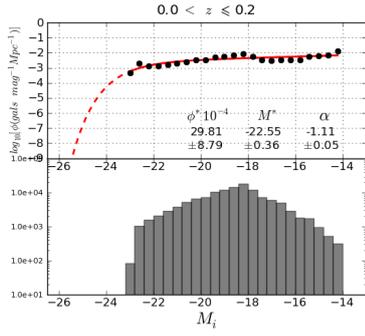
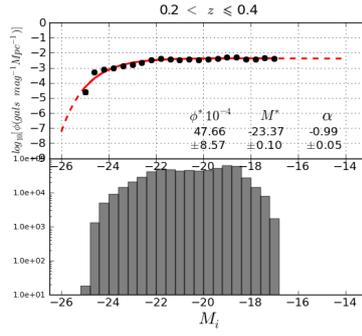


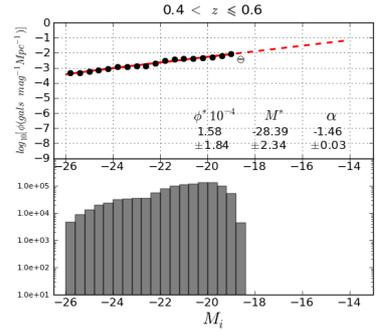
Figura G.19: Função de luminosidade global em intervalos de *redshift*, sendo este o valor do *photo-z* calculado com o TPZ.



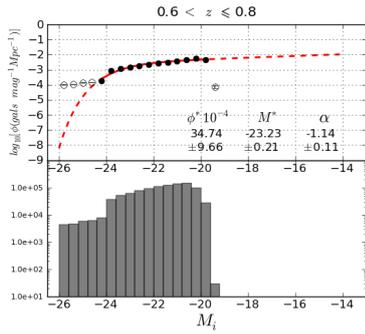
(a)  $0.0 < z \leq 0.2$  ( $m_i < 24.4$ )



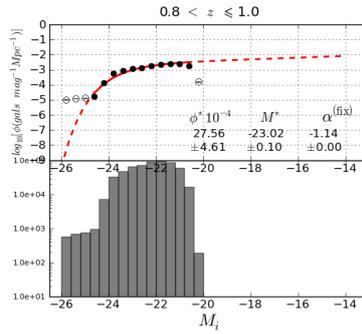
(b)  $0.2 < z \leq 0.4$  ( $m_i < 24.4$ )



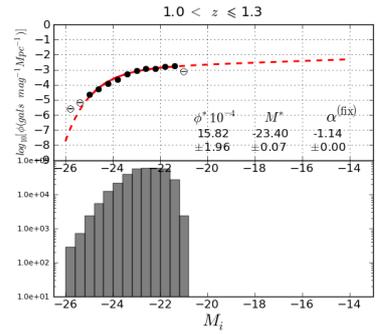
(c)  $0.4 < z \leq 0.6$  ( $m_i < 24.4$ )



(d)  $0.6 < z \leq 0.8$  ( $m_i < 24.4$ )



(e)  $0.8 < z \leq 1.0$  ( $m_i < 24.4$ )



(f)  $1.0 < z \leq 1.3$  ( $m_i < 24.4$ )

Figura G.20: Função de luminosidade global em intervalos de *redshift*, sendo este o valor do *photo-z* calculado com o TPZ e histograma com contagem de galáxias em intervalos de magnitude absoluta, em escala logarítmica.

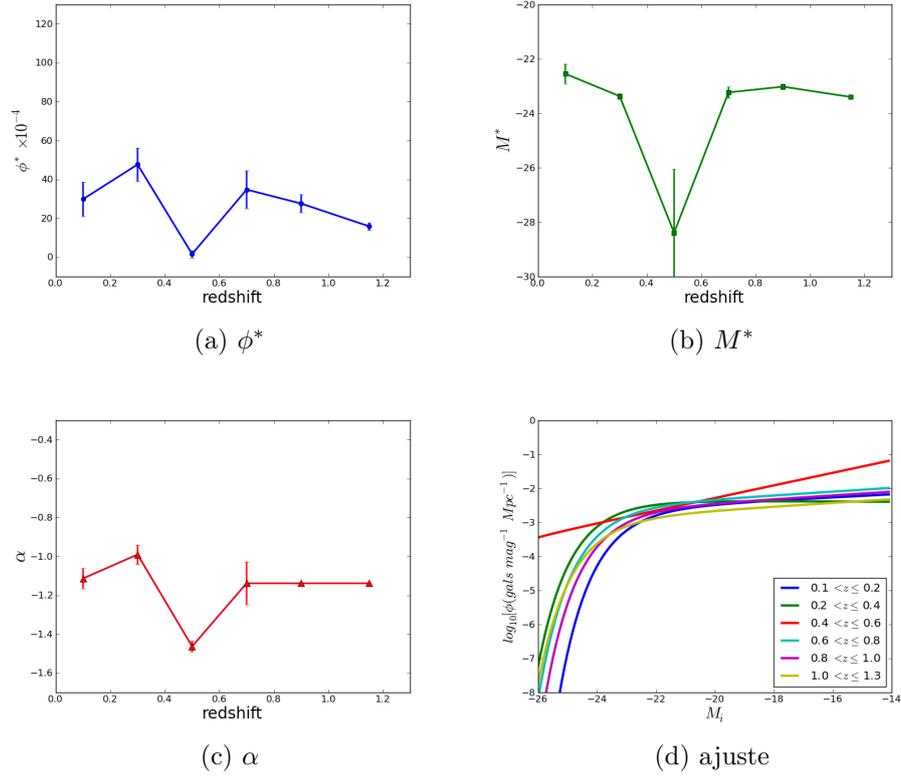


Figura G.21: Evolução dos parâmetros e da curva ajustada à função de Schechter. LF obtida a partir dos *photo-z's* calculados com o TPZ.

## G.5 LF por tipos espectrais (preliminar)

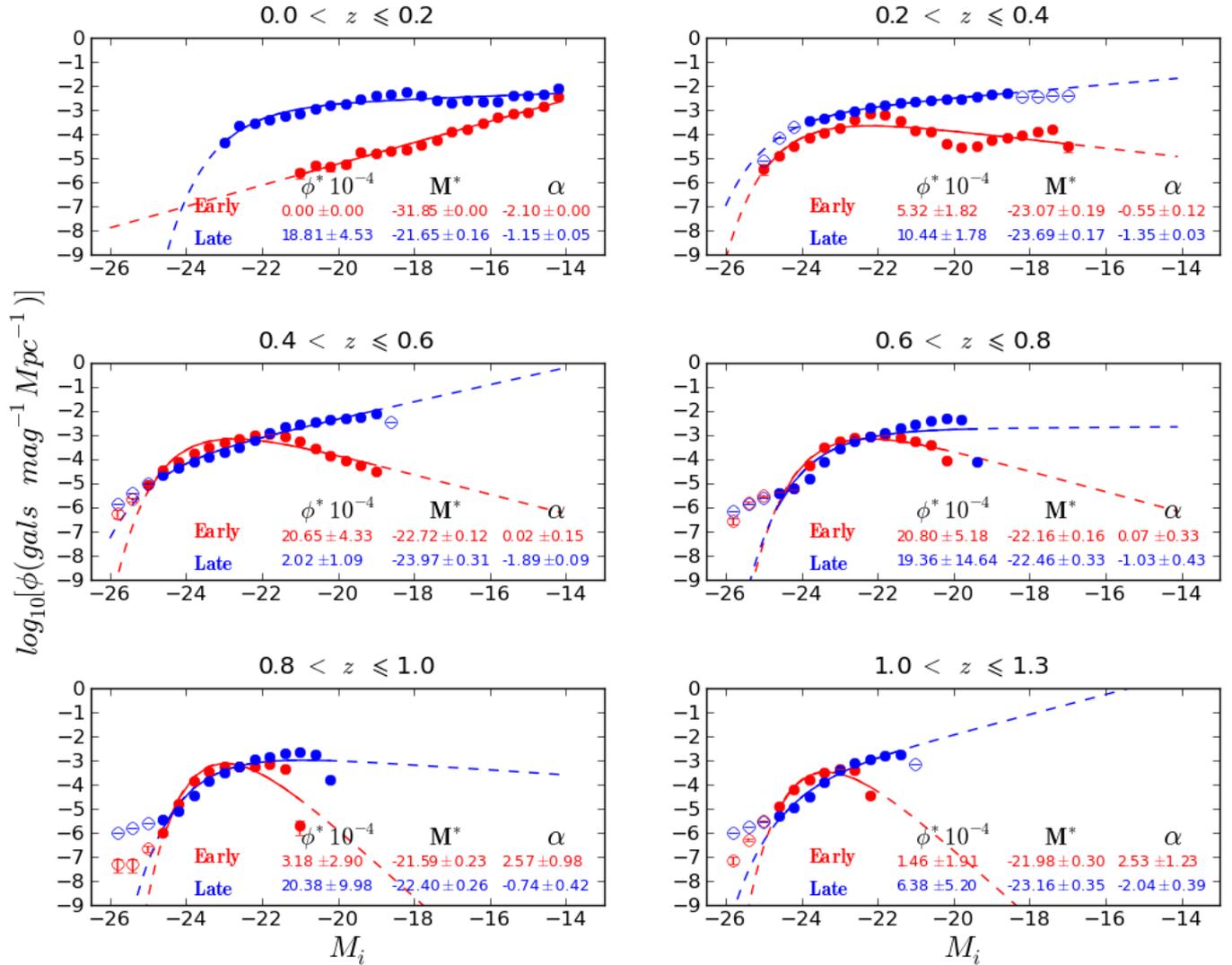
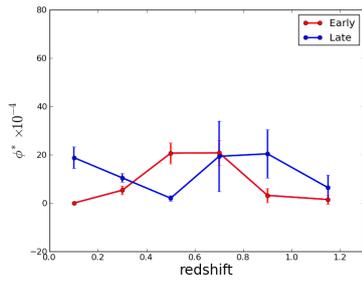
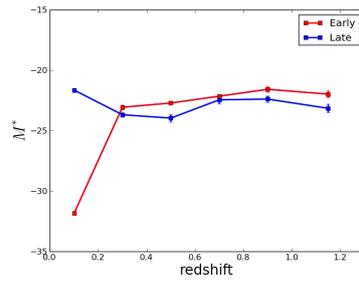


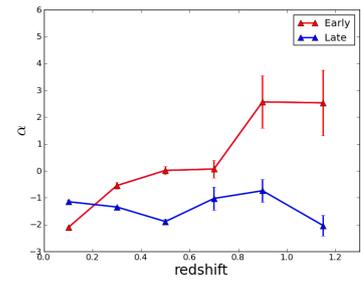
Figura G.22: Função de luminosidade após a divisão da amostra em dois tipos espectrais utilizando o  $photo-z$  calculado pelo TPZ.



(a)  $\phi^*$



(b)  $M^*$



(c)  $\alpha$

Figura G.23: Evolução dos parâmetros de Schechter da LF obtida a partir dos *photo-z*'s calculados com o TPZ.