Universidade Estadual de Campinas

Instituto de Física *Gleb Wataghin*

F 896 Monografia

Estimativas de *redshifts* fotométricos para o Dark Energy Survey

Aluno: Pedro Rangel Caetano

E-mail: p.r.caetano@gmail.com

Orientadora: Profa. Dra. Flávia Sobreira Sanchez

E-mail: sobreira@ifi.unicamp.br

Departamento de Raios Cósmicos e Cronologia

Instituto de Física Gleb Wataghin

Universidade Estadual de Campinas

Campinas-SP

Dezembro 2017

Resumo

O Dark Energy Survey (DES) é um levantamento fotométrico em 5 filtros no visível e infravermelho próximo criado com o objetivo de permitir medidas precisas de parâmetros cosmológicos, em especial da equação de estado da energia escura. Para este objetivo, é importante obter estimativas dos *redshifts* dos objetos observados, o que só pode ser feito neste caso a partir da fotometria multibanda. Estudamos neste trabalho estimativas de *redshifts* fotométricos nos dados do ano 3 do DES. Utilizamos o algoritmo Bayesian Photometric *Redshifts* (BPZ), que teve seu desempenho caracterizado quanto a dispersão, vieses, fração de outliers e fidelidade de reprodução da distribuição de galáxias em *redshift* utilizando uma amostra de validação preparada a partir do casamento da fotometria do DES com outros levantamentos espectroscópicos.

Abstract

The Dark Energy Survey (DES) is a photometric survey in 5 optical and near-infrared bands designed to make precision measurements of cosmological parameters and constraint the dark energy equation of state. To this end, it is necessary to obtain *redshifts* for the observed sources, and in DES case this must be done using multiband photometry. Here we study photometric *redshift* estimation on DES year 3 data. We use the algorithm Bayesian Photometric *Redshifts* (BPZ), which has its performance characterized on bias, dispersion, outlier fraction and fidelity of reprodution of *redshift* distribution with a validation sample prepared by matching DES data with spectroscopic surveys.

Biografia do Autor

Pedro Rangel Caetano nasceu em 1995 em Sorocaba, interior de São Paulo. Ingressou no bacharelado em Física na Unicamp em 2013 e, após uma breve excursão pela matemática, dedica-se agora à Cosmologia Observacional.

A meus pais, por tudo.

Agradecimentos

Agradeço a meus pais por todo o amor e apoio durante minha graduação e antes dela, por ter me ensinado a explorar o mundo um pouquinho além da superfície, a reconhecer o caráter coletivo das ideias e do trabalho humano e, sobretudo, por mostrar no dia-a-dia como lutar para sobreviver e sonhar com um mundo mais justo onde tantas vidas e sonhos não sejam tolhidas pelo egoísmo e pela mesquinharia.

À professora Flávia Sobreira, pela orientação e revisão do texto.

Aos meus amigos, por todas as discussões, aprendizados, nonsense e por tornar a vida interessante e possível, dividindo comigo o caminho tantas vezes árido e sombrio pelos corredores do IFGW.

Agradeço também ao Laboratório Interinstitucional de e-Astronomia (LIneA) e seu pessoal pelo fornecimento da infraestrutura sem a qual este trabalho seria impossível, e em especial à Júlia Gschwend, pela ajuda com os pipelines do photo-z e sugestões ao texto. Também agradeço a Ricardo Ogando, Christophe Benoist e Paulo Pellegrini pelas sugestões e correções incorporadas nesta versão final.

Por fim, agradeço a todos os trabalhadores da Unicamp e do mundo, por possibilitar uma oportunidade infelizmente tão rara. Todo o conhecimento é de vocês também, embora seja tantas vezes roubado descaradamente.

Sumário

Resumo **ii**

Abstract **iii**

Biografia do Autor **iv**

1 Introdução **1**

2 Medidas fotométricas de *redshift* **4**

2.1 Desvios para o vermelho 5

2.2 Técnicas espectroscópicas e fotométricas 5

2.3 Métricas de caracterização de *redshifts* fotométricos 11

3 Ferramentas computacionais **13**

3.1 *Redshifts* fotométricos como um problema de inferência 14

3.2 Portal Científico do DES 18

4 Resultados **21**

4.1 Dados utilizados 21

4.2 Amostra fotométrica 22

4.3 Amostra de validação 25
4.4 Rodadas realizadas 25

5 Conclusão **31**

Bibliografia **32**

Apêndices **39**

A Aplicações do *redshift* à cosmologia **40**
A.1 Distâncias e *redshifts* 42
A.2 A Lei de Hubble 44

B Inferência bayesiana **46**

1

Introdução

A cosmologia observacional é uma área bem estabelecida e com influência crescente nas últimas décadas. Um longo caminho foi trilhado desde os trabalhos pioneiros no começo do século XX, e hoje se trata de uma área rica em dados e técnicas, capaz de fazer testes de precisão dos modelos cosmológicos propostos.

De fato, a descoberta da expansão acelerada do universo ([37], [36]) no fim da década de 1990, que rendeu aos líderes dos grupos envolvidos o prêmio Nobel de 2011, é provavelmente uma das mais importantes descobertas do século passado, se tratando de uma das principais anomalias que aponta para além dos modelos padrão da cosmologia e física de partículas.

Dentro do arcabouço do modelo cosmológico padrão, conhecido como ΛCDM[1], a taxa de expansão do Universo é relacionada à sua composição em diversos componentes, como, e.g., matéria bariônica e radiação eletromagnética. Cada componente é modelada como um fluído perfeito, cuja evolução é descrita por sua equação de estado $P = w\rho$, relacionando sua pressão P à densidade ρ. Neste modelo, a expansão acelerada é explicada com a adição de uma constante cosmológica Λ, que pode ser interpretada como uma nova componente do Universo com pressão negativa e cuja densidade se mantêm constante conforme o Universo expande (i.e., sua equação

[1] CDM são as iniciais de Cold Dark Matter - matéria escura fria - e Λ representa a constante cosmológica

de estado w é constante e igual a -1), chamada energia escura. Pelos valores mais aceitos, este modelo implica que, hoje, 73 % do Universo (em termos de densidade de energia) corresponde a energia escura, 23 % a matéria escura e apenas 4 % a matéria bariônica.

Não se sabe ao certo porém se a energia escura é de fato uma constante cosmológica, ou seja, se sua equação de estado não muda com a evolução do universo. Há modelos alternativos, como a quintessência, que incluem esta possibilidade. É também, inclusive, possível que a expansão cósmica seja uma indicação de falha da Relatividade Geral[3].

Neste cenário, se faz essencial caracterizar com mais precisão o comportamento da energia escura, de forma a obter evidências experimentais que facilitem a resolução deste mistério. A Dark Energy Task Force [3] propôs um cronograma, numa escala de tempo de décadas, de experimentos com este fim. O Dark Energy Survey(DES) é um dos experimentos propostos dentro da 3a. geração deste cronograma, almejando melhorar cerca de 4.6 vezes as medidas da equação de estado da energia escura e da sua taxa de variação[2]. Para isto o DES utilizará 4 diferentes sondas cosmológicas: aglomerados de galáxias, oscilações acústicas bariônicas(BAO), lenteamento fraco e supernovas do tipo IA. Cada uma destas sondas é independente das demais, e sujeita a diferentes tipos de desvios sistemáticos. A combinação entre elas, além de aumentar a precisão das medidas, as torna mais confiáveis, pois fornece um teste de consistência.

O Dark Energy Survey é um levantamento astronômico medindo posições, cores e formatos de galáxias numa área de 5000 graus quadrados (aproximadamente 1/8 do céu). Adicionalmente, uma área menor de 30 graus quadrados é visitada frequentemente em busca de supernovas. Espera-se que sejam medidos cerca de 300 milhões de galáxias até *redshift* 1.2, assim como 100.000 aglomerados de galáxias e 3000 supernovas Ia. As medidas são realizadas no telescópio Victor M. Blanco, de 4 m de diâmetro, que fica no Cerro Tololo Inter-American Observatory (CTIO), com a Dark Energy Camera (DECam), uma câmera com sensor CCD de aproximadamente 500 Megapixels e 7 filtros que selecionam regiões específicas do espectro visí-

[2]Parametrizando a evolução da equação de estado com o fator de escala como $w(a) = w_0 + (1-a)w_a$, a melhora por um fator de 4.6 refere-se à figura de mérito no espaço de parâmetros w_0 e w_a, que por sua vez corresponde ao inverso da área cobrindo a região de 95 % de nível de confiança.

vel e infravermelho próximo, cobrindo um campo de visão de 2.2 graus quadrados[28]. O DES é uma colaboração internacional liderada pelo Fermilab e incluindo centenas de pesquisadores. A participação brasileira ocorre por meio do Laboratório Interinstitucional de e-Astronomia (LIneA) e inclui a Unicamp.

Para a maior parte das medidas realizadas pelo DES é importante conhecer o *redshift* das galáxias detectadas, provocado pela expansão do Universo e associado à distância das galáxias. O DES, porém, não é um levantamento espectroscópico, portanto esta medida não é direta. Os *redshifts* são inferidos a partir da fotometria multibanda nas bandas g, r, i, z e Y, sendo denominados *redshifts* fotométricos.

Este trabalho se dedica a explorar um algoritmo de *redshifts* fotométricos bastante utilizado, denominado Bayesian Photometric *Redshifts* (BPZ)[6], descrevendo seu funcionamento e caracterizando sua performance nos dados do ano 3 do DES. Também discutiremos a física dos *redshifts*, bem como seu papel e relevância na cosmologia de forma geral.

2

Medidas fotométricas de redshift

O *redshift*, ou desvio para o vermelho, é uma das quantidades mais importantes em estudos de cosmologia e astrofísica extragalática. A partir dele diversas outras grandezas são derivadas, como a distância comóvel da fonte, a idade do universo quando a luz captada foi emitida e a luminosidade intrínseca do objeto sendo observado. Destas quantidades, por sua vez, são obtidas medidas essenciais para vincular parâmetros cosmológicos e estudar a estrutura em grande escala do Universo (como a função de correlação), para entender os processos de formação e evolução de galáxias (como a função de luminosidade) ou para obter a distribuição de matéria escura do Universo (como em técnicas de lenteamento fraco). Em suma, a obtenção de *redshifts* muitas vezes é o que permite que as observações astronômicas obtenham profundidade e evolução temporal.

No DES a maioria das medidas realizadas depende da obtenção de *redshifts* para as galáxias observadas. As técnicas utilizadas para este fim, porém, possuem algumas sutilezas pois o DES é um levantamento fotométrico e a inferência de *redshifts* fotométricos pode não ser tão precisa, e produzir grandes desvios, conhecidos como "erros catastróficos". Este capítulo é dedicado à discussão das técnicas fotométricas de determinação de *redshifts*, introduzindo os conceitos básicos necessários, bem como as técnicas utilizadas para avaliar e garantir a confiabilidade dos

resultados.

2.1 Desvios para o vermelho

O desvio para o vermelho, ou *redshift*, de uma fonte astronômica é o nome dado ao deslocamento do espectro observado na direção de maiores comprimentos de onda, com relação ao espectro observado no referencial de repouso da fonte. O *redshift* costuma ser denotado pela letra z, e quantitativamente expresso como

$$z = \frac{\Delta\lambda}{\lambda} = \frac{\lambda' - \lambda}{\lambda}, \tag{2.1}$$

onde λ e λ' são os comprimentos de onda de um fóton medidos no referencial da fonte e do observador, respectivamente. É possível também escrevê-lo em termos das frequências emitida f e observada f',

$$z = \frac{f}{f'} - 1. \tag{2.2}$$

O *redshift* pode ser determinado identificando um ou mais artefatos de comprimento de onda conhecido (correspondentes a transições espectrais conhecidas, por exemplo) num espectro medido.

As causas possíveis para o *redshift* de uma fonte podem ser diversas. No Apêndice A, discutimos como o efeito Doppler, a expansão do Universo e potenciais gravitacionais causam este fenômeno, bem como que tipo de informações podemos inferir em cada um dos casos.

2.2 Técnicas espectroscópicas e fotométricas

A maneira mais óbvia de se determinar o *redshift* de uma fonte é obter o seu espectro e, após encontrar marcadores com comprimento de onda conhecido (uma linha devida à uma transição de um elemento conhecido, por exemplo), determinar a diferença dos comprimentos de onda medidos e de referência para o marcador, obtendo assim uma medida direta de $\Delta\lambda$, e portanto de z.

Esta técnica costuma ser bastante precisa, mas é onerosa: o tempo de integração necessário para obter o espectro de uma galáxia é bastante longo e galáxias distantes, e mais tênues, são de difícil observação. Além disso, o número de medições simultâneas realizadas costuma ser limitado, pois tipicamente a luz de cada galáxia é guiada até o espectroscópio por fibras ópticas que devem ser precisamente posicionadas no plano focal do telescópio na posição da galáxia, sendo desafiador realizar este posicionamento quando há muitas galáxias no campo. De fato, o levantamento Sloan Digital Sky Survey (SDSS) coletou, entre 2000 e 2016, espectros de cerca de 5 milhões de objetos e o DESI, proposto para começar a coleta de dados em 2018, espera coletar da ordem de dezenas de milhões de espectros, inclusive utilizando um sistema controlado por computador para posicionamento rápido das fibras ópticas, enquanto o DES observará cerca de 300 milhões de galáxias em seus cinco anos de coleta de dados. Estas dificuldades enfrentadas por levantamentos espectroscópicos também podem gerar vieses de seleção na amostra medida (por exemplo, a colisão de fibras pode gerar subamostragem de tipos de galáxias mais aglomerados), o que pode ser um problema para algumas aplicações.

Levantamentos espectroscópios de grande escala são portanto de difícil realização. É possível, porém, obter *redshifts* por fotometria, isto é, utilizando imagens dos objetos com diferentes filtros, que permitem a passagem de luz apenas dentro de uma faixa de comprimentos de onda. Medindo a quantidade de luz em cada filtro, obtemos essencialmente uma versão de baixa resolução do espectro da fonte[1]. Como espectros de galáxias frequentemente apresentam artefatos que são detectáveis mesmo nestas baixas resoluções, como a descontinuidade no espectro em 4000 Ângstron, caracterizada por uma brusca queda na densidade espectral de fluxo para comprimentos de onda abaixo de 4000 Ângstron, é possível obter estimativas do *redshift*, que neste caso é chamado *redshift* fotométrico (ou photo-z, de forma abreviada).

Uma das principais vantagens desta técnica é que é possível obter com relativa facilidade uma grande quantidade de medições simultaneamente. A DECam por exemplo, que é a câmera

[1]Na verdade, é possível obter fotoespectros de boa resolução utilizando muitos filtros de banda estreita, e alguns levantamentos utilizam ou utilizarão esta técnica, como ALHAMBRA (20 filtros, FWHM = 310 Angstrom) e J-PAS (56 filtros).

utilizada no DES, registra em cada exposição milhares de galáxias. Este grande volume de dados balanceia a baixa precisão das medições individuais, sendo possível obter boas medidas de grandezas estatísticas, como a distribuição espacial das galáxias ou a distribuição em luminosidade de galáxias num certo intervalo de *redshift*.

A grande quantidade de dados, por outro lado, pode tornar estas medições desafiadoras. Há um processo de redução desde a obtenção das imagens até a elaboração de um catálogo de objetos contendo a posição de cada objeto no céu, seu *redshift* fotométrico e sua fotometria. Cada uma destas etapas envolve um processo de calibração minucioso e pode ser computacionalmente cara. No DES o processo de redução dos dados é realizado no National Center for Supercomputing Aplications (NCSA), e a participação brasileira também utiliza um cluster localizado no Laboratório Nacional de Computação Científica (LNCC).

Nosso foco neste trabalho é a última etapa do processo descrito acima, a determinação dos *redshifts* de posse da fotometria dos objetos (em particular, das galáxias). Há uma miríade de algoritmos dedicados a este problema, mas veremos que todos se situam em uma de duas principais abordagens: a técnicas de ajuste de modelos e a de aprendizado de máquina. O algoritmo aqui estudado, BPZ, é um dos mais utilizados da primeira categoria.

Conceitos básicos de fotometria

Considere uma fonte emitindo radiação eletromagnética que eventualmente é detectada com algum dispositivo. Podemos medir a intensidade da radiação obtendo a quantidade de energia depositada pela fonte a cada intervalo de tempo numa área determinada do detector, sendo esta grandeza denominada *fluxo*. Outra grandeza relevante é a *densidade espectral de fluxo*, ou seja, o fluxo devido a uma fonte num intervalo $d\nu$ ao redor de uma frequência dada ν.

O fluxo medido na Terra para diferentes objetos varia enormemente: o fluxo devido ao Sol é da ordem de 10^{12} vezes maior que o fluxo das estrelas mais tênues observadas a olho nu. Além disso, a percepção humana de brilhos é logarítmica. É portanto conveniente introduzir uma escala logarítmica de fluxos. A *magnitude aparente* de um objeto com fluxo f é definida como

$$m = -2.5 \log_{10} f + C,$$

onde C é uma constante que determina o ponto 0 da escala. Note que diferenças de magnitudes representam razões de fluxos, pois

$$m_1 - m_2 = -2.5 \log_{10} \frac{f_1}{f_2}.$$

A cor de um objeto está ligada à razão dos fluxos obtidos em diferentes regiões do espectro visível: isto motiva a definição dos *índices de cor* como a diferença das magnitudes medidas em diferentes filtros. No DES, por exemplo, são usados os filtros g, r, i, z e Y. Podemos então calcular os indíces de cor $g - r$, $r - i$, $i - z$ e assim por diante.

Métodos de ajuste de modelos

As técnicas de estimativa de *redshift* fotométrico baseadas em ajuste de modelos usam modelos previamente selecionados dos tipos de espectros que esperamos encontrar. Estes modelos são chamados *Spectral Energy Distributions*, ou SEDs. Uma biblioteca de SEDs pode ser montada tanto a partir de espectros reais obtidos de uma amostra de galáxias como a partir de modelos de evolução galáctica.

Um algoritmo de ajuste de modelos típico recebe como entradas uma biblioteca de SEDs, a resposta espectral de cada filtro do sistema fotométrico utilizado, as magnitudes medidas para cada um destes filtros e o intervalo de busca para o *redshift*. O intervalo de busca é então particionado, originando um conjunto discreto de *redshifts*. Para cada par (SED, *redshift*) o programa calcula o espectro esperado caso observássemos uma galáxia deste tipo neste *redshift* e, realizando uma convolução da SED com a resposta dos filtros são encontradas as magnitudes esperadas nesta situação. Comparando as magnitudes esperadas com as observadas é possível quantificar a qualidade do ajuste. Repetindo este processo percorrendo todo o intervalo de busca e toda a biblioteca de SEDs, o valor atribuído para o *redshift* fotométrico do objeto é aquele correspondente ao melhor ajuste.

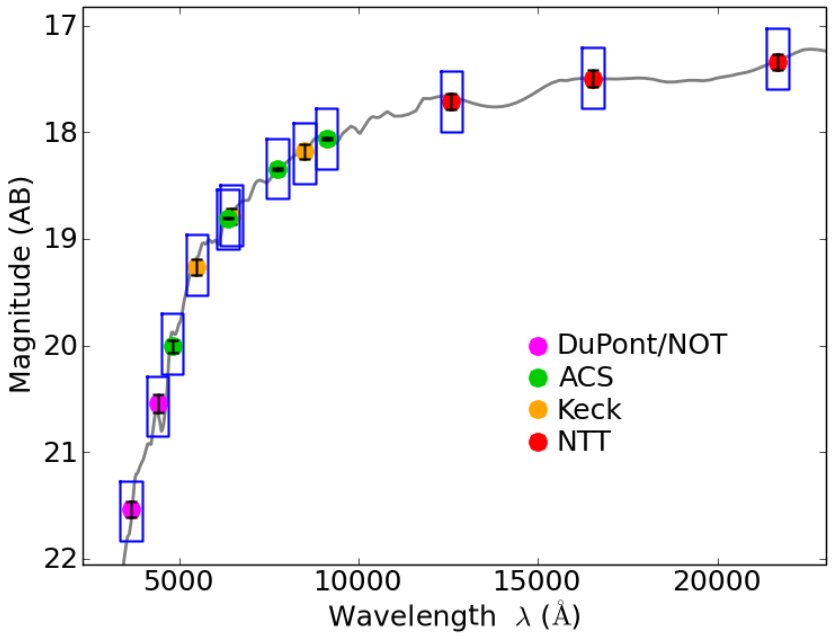

Figura 2.1: Exemplo de ajuste de modelo, retirado de [10]. A linha cinza corresponde à SED que melhor ajusta as magnitudes observadas, já deslocada por um *redshift* $z = 0.18$ (correspondente ao valor inferido para o *redshift* fotométrico). As caixas azuis correspondem às magnitudes obtidas pela convolução da SED com as respostas espectrais dos filtros, e suas respectivas incertezas. Os pontos coloridos correspondem às magnitudes observadas.

Na verdade, é possível obter com este processo a distribuição de probabilidade de *redshift* como função das SEDs e dos *redshifts*. Isto é importante pois é possível que haja degenerescências na relação entre as magnitudes nos filtros observados e o conjunto de modelos utilizado, isto é, é possível que diferentes pares (SED, *redshift*) forneçam as mesmas magnitudes nas bandas observadas, sendo assim impossível determinar qual é o *redshift* real do objeto. Tipicamente nestes casos o programa nos forneceria uma distribuição de probabilidade com 2 ou mais picos, sendo portanto capaz de perceber a degenerescência e nos dar alguma informação sobre ela. Adicionalmente, tendo a distribuição de probabilidade completa temos uma melhor caracterização da precisão da medida obtida. Algumas análises científicas podem também se beneficiar

desta informação adicional.

Por outro lado, métodos de ajuste a modelos são dependentes de boas bibliotecas de modelos, e podem ser bastante sensíveis às escolhas particulares de modelos realizadas.

O algoritmo estudado neste trabalho, Bayesian Photometric *redshifts* (BPZ), realiza ajuste de modelos dentro de uma abordagem bayesiana, incluindo modelagem da informação a priori disponível. Estas informações são importantes pois permitem a resolução de algumas das degenerescências mencionadas anteriormente.

Métodos empíricos

A outra abordagem comumente utilizada no cálculo de *redshifts* fotométricos é a de aprendizado de máquina. Neste caso, o problema de determinação de *redshifts* fotométricos é reformulado como um problema de aprendizado supervisionado, ou seja, é fornecido ao programa um conjunto de medições para os quais o *redshift* é conhecido (normalmente foi obtido por técnicas espectroscópicas), a partir do qual o programa aprenderá a inferir *redshifts* desconhecidos.

A entrada do algoritmo consiste portanto num conjunto $\{({m_i}^j, z^j)\}$, onde j indexa as medições (e i os filtros utilizados). Este conjunto é fracionado em duas partes: os conjuntos de treinamento e validação. Numa primeira etapa, o programa tenta inferir os *redshifts* do conjunto de treinamento a partir das magnitudes de entrada, se adaptando até que obtenha um bom desempenho. Posteriormente, o programa tenta inferir os *redshifts* do conjunto de validação a partir das magnitudes apenas, sendo assim obtida uma caracterização da resposta do algoritmo. Se o desempenho for bom no conjunto de validação o programa pode ser utilizado para extrapolar *redshifts* desconhecidos a partir apenas das magnitudes, como desejado.

Para garantir que os resultados obtidos sejam confiáveis é importante que as amostras de treinamento e validação sejam semelhantes à amostra fotométrica para a qual deseja-se inferir o *redshift*. Isto significa que idealmente as distribuições de magnitudes e cores destas amostras devem ser semelhantes.

Algoritmos baseados em aprendizado de máquina são tipicamente rápidos e possuem bom desempenho, além de não dependerem de uma biblioteca de modelos. A caracterização da pre-

cisão dos resultados é porém mais difícil. Embora seja possível, muitos destes algoritmos não retornam uma distribuição de probabilidades para o *redshift*, sendo assim mais difícil caracterizar a precisão dos resultados. Além disso, ao contrário de algoritmos de ajuste de modelos, que tendem a avisar quando não foi possível obter uma resposta confiável, eles tendem a falhar silenciosamente.

Há uma infinidade de métodos de aprendizado de máquina supervisionado, como redes neurais, árvores de decisão e vizinhos próximos. Estas estratégias são implementadas por uma grande variedade de algoritmos.

2.3 Métricas de caracterização de *redshifts* fotométricos

Para garantir a confiabilidade dos *redshifts* inferidos, é importante caracterizar o desempenho do algoritmo, verificando seu comportamento em uma subamostra para a qual conhecemos o *redshift* "verdadeiro" (espectroscópico). Há quatro aspectos importantes a serem caracterizados: os vieses (medem se os *redshifts* fotométricos são sistematicamente superiores ou inferiores aos verdadeiros), a dispersão (mede quão dispersos ao redor do *redshift* verdadeiro são distribuídos os *redshifts* fotométricos) os erros catastróficos (mede quão frequentemente o *redshift* fotométrico possui grande discordância com o *redshift* real) e a fidelidade de reprodução da distribuição de *redshifts*. Seguindo [38], denotaremos por z_{phot} e z_{spec} os *redshifts* fotométricos e espectroscópicos, ϵ_{phot} as incertezas de z_{phot}, $N(z_{phot})$, $N(z_{spec})$ as distribuições do número de galáxias por *redshift*, $\Delta z = z_{photo} - z_{spec}$ a distribuição dos desvios e $\Delta z' = \Delta z/\epsilon_{phot}$ a distribuição dos desvios normalizados pelas incertezas. Há diversas métricas capazes de mensurar o desempenho dos *redshifts* fotométricos nos 4 aspectos descritos acima, descritas na Tabela 2.3. É importante obtê-las não somente para a amostra completa, mas também determinar a evolução das mesmas com *redshift*, o que pode ser obtido particionando o conjunto de validação em vários intervalos de *redshift* e computando as métricas para cada partição.

Métrica	Descrição	Tipo
$\overline{\Delta z}$	média da distribuição Δz	Viés
Δz_{50}	mediana da distribuição Δz	Viés
$\overline{\Delta z'}$	média da distribuição de z'	Viés
$\sigma_{\Delta z}$	desvio padrão da distribuição Δz	Dispersão
σ_{68}	metade do intervalo ao redor da mediana Δz_{50} contendo 68 % da amostra	Dispersão
$\sigma_{\Delta z'}$	desvio padrão da distribuição $\Delta z'$	Dispersão
$\text{out}_{2\sigma}$	fração da amostra para a qual $\lvert\Delta z - \overline{\Delta z}\rvert > 2\sigma_{\Delta z}$	Erros catastróficos
$\text{out}_{3\sigma}$	fração da amostra para a qual $\lvert\Delta z - \overline{\Delta z}\rvert > 3\sigma_{\Delta z}$	Erros catastróficos
$N_{poisson}$	r.m.s. da distribuição $N(z_{phot}) - N(z_{spec})$, normalizado por flutuações poissonianas $\sqrt{N(z_{spec})}$	Reprodução de $N(z)$

Tabela 2.1: Métricas utilizadas para caracterização de vieses, dispersão, erros catastróficos e reprodução da distribuição real de *redshifts* dos *redshifts* fotométricos. Seguimos as convenções de [38], mas há outras utilizadas na literatura.

3

Ferramentas computacionais

Neste capítulo discutiremos com maior detalhe o funcionamento do algoritmo que usamos para determinar os *redshifts* fotométricos neste trabalho: o Bayesian Photometric *Redshifts* (BPZ) [6]. O BPZ é um algoritmo de ajuste de modelos que utiliza uma abordagem bayesiana para derivar uma distribuição de propabilidade de encontrarmos uma galáxia num *redshift* z para a qual foram medidas magnitudes $\{m_i \pm \Delta m_i\}_{i \in D}$ em um certo conjunto de filtros D. A partir desta distribuição podemos inferir um valor de *redshift* fotométrico tomando a moda (o pico) ou a mediana desta distribuição, bem como quantificar a incerteza desta medida a partir da largura da mesma, que pode ser determinada com o desvio padrão, por exemplo. A distribuição em si, porém, contêm mais informação, e de fato várias técnicas modernas aproveitam este fato para melhorar análises em cosmologia.

3.1 *Redshifts* fotométricos como um problema de inferência

A distribuição de probabilidade dos *redshifts*

A introdução das técnicas de inferência bayesiana, descritas no Apêndice B, no problema de estimativa de *redshifts* fotométricos é bastante natural. Primeiramente vamos introduzir algumas convenções. Para cada observação, denotaremos por m_0 a magnitude observada numa banda de referência pré-definida. Se $F = \{f_1, f_2, ..., f_n\}$ é o conjunto de filtros utilizado, e $m = \{m_1, ..., m_n\}$ as magnitudes de uma galáxia em cada um dos filtros, chamaremos de C o conjunto de cores observado $\{m_2 - m_1, m_3 - m_2, ..., m_n - m_{n-1}\}$. T denotará um modelo qualquer na biblioteca de modelos utilizada.

A primeira observação a ser feita é que a informação que desejamos obter é a distribuição de probabilidades $P(z)$ dos *redshifts* para uma galáxia. Dada a magnitude de referência m_0 já possuímos uma distribuição $P(z|m_0)$, correspondente à informação que já possuímos a priori, tanto devido a estudos empíricos anteriores como a restrições físicas. Por exemplo, embora uma galáxia com magnitude m_0 possa estar a um *redshift* arbitrário sabemos previamente que a luminosidade de galáxias está limitada a um certo intervalo. Isto implica que *redshifts* muito baixos ou muito altos são improváveis, pois nestes casos a galáxia possuiria uma luminosidade muito baixa ou muito alta para a magnitude observada.

Sabemos entretanto que a galáxia foi observada com cores C. Se conseguirmos modelar a função de verossimilhança $P(C|z, m_0)$ aplicando o teorema de Bayes podemos encontrar a distribuição de probabilidade dos *redshifts* para a galáxia após a observação

$$P(z|C, m_0) \propto P(C|z, m_0)P(z|m_0).$$

Marginalizando sob os modelos

Na verdade, para encontrarmos $P(z, C|m_0)$ precisamos levar em conta os vários modelos disponíveis em nossa biblioteca. Aqui, o modelo associado à observação é uma variável acessória, sob a qual podemos marginalizar

$$P(z|C, m_0) \propto \sum_T P(C|z, T, m_0)P(z, T|m_0).$$

A função de verossimilhança $P(C|z, T, m_0)$ pode ser obtida agora verificando a qualidade do ajuste do modelo T num *redshift* z às cores observadas. Isto pode ser feito considerando o χ^2 do ajuste: se f_α são os fluxos observados em cada um dos filtros $\alpha \in F$, com incertezas associadas $\sigma_{f\alpha}$, $f_{T\alpha}$ são os fluxos previstos para o modelo T num *redshift* z e a é uma constante de normalização para o espectro do modelo temos

$$\chi^2(z, T, a) = \sum_\alpha \frac{(f_\alpha - a f_{T\alpha})^2}{\sigma_{f\alpha}^2}, \tag{3.1}$$

e a probabilidade de obtermos as cores C para o *redshift* z e modelo T é

$$P(C|z, T, a, m_0) \propto exp\left(-\frac{\chi^2(z, T, a)}{2}\right).$$

O parâmetro a deve ser agora introduzido como parâmetro acessório, por isso marginalizamos novamente,

$$\begin{aligned}
P(z|C, m_0) &\propto \int da \sum_T P(z, T, a|C, m_0) \\
&\propto \sum_T \int da P(C|z, T, a, m_0)P(z, T, a|m_0) \\
&\propto \sum_T \int da P(C|z, T, a)P(z, T|m_0)p(a|m_0) \\
&\propto \sum_T p(z, T|m_0) \int da p(a|m_0)e^{-\chi^2(z,T,a)/2}
\end{aligned}$$

onde assumimos que $P(C|z,T,a,m_0)$ é independente de m_0 e que z e T são independentes de a, logo $p(z,T,a|m_0) = p(z,T|m_0)p(a|m_0)$. Como não possuímos nenhum modelo para $p(a|m_0)$ assumiremos que esta probabilidade é constante. Neste caso, sendo a_m o valor para o qual $\chi^2(z,T,a)$ é máximo, obtemos,

$$P(z|C,m_0) \propto \sum_T p(z,T|m_0) F_{TT}(z)^{-1/2} \exp\left[-\frac{\chi^2(z,T,a_m)}{2}\right] \tag{3.2}$$

sendo o máximo a_m de $\chi^2(z,T,a_m)$ dado por

$$a_m = \frac{F_{OT}}{F_{TT}} \tag{3.3}$$

e os valores F_{OT} e F_{TT} por

$$F_{OO} = \sum_\alpha \frac{f_\alpha^2}{\sigma_{f_\alpha}^2}, \quad F_{TT} = \sum_\alpha \frac{f_{T_\alpha}^2}{\sigma_{f_\alpha}^2}, \quad F_{OT} = \sum_\alpha \frac{f_\alpha f_{T_\alpha}}{\sigma_{f_\alpha}^2}. \tag{3.4}$$

Probabilidades à priori

Com a equação (3.2) já essencialmente concluímos a modelagem do problema: se soubermos calcular todos os termos ali presentes já seremos capazes de determinar o *redshift* fotométrico de uma galáxia dada. As equações (3.1), (3.3), (3.4) já nos possibilitam calcular a parte referente à verossimilhança, portanto só falta-nos obter as probabilidades à priori $P(z,T|m_0)$. Podemos usando a regra do produto escrever as probabilidades à priori como

$$P(z,T|m_0) = p(z|T,m_0)p(T|m_0), \tag{3.5}$$

efetivamente a decompondo em probabilidades a priori para modelos e *redshifts*. Estas probabilidades são modeladas pelo BPZ em termos de parâmetros α_t, z_{0t}, k_{mt} e k_t , onde os subíndices t denotam o tipo da galáxia à qual os modelos se referem: $t = 1$ corresponde a galáxias *early-type*[1], $t = 2$ a galáxias espirais e $t = 3$ a galáxias irregulares. A probabilidade à priori dos

[1]Galáxias *early-type* compreendem simultaneamente as galáxias elípticas e lenticulares, sendo o nome devido à uma interpretação errônea do diagrama de classificação morfológica proposto por Hubble.

Tipo espectral	α_t	z_{0t}	k_{mt}	k_t
Early types	2.46	0.542	0.112	0.296
Espirais	1.836	0.399	0.101	0.143
Irregulares	1.180	0.134	0.143	–

Tabela 3.1: Parâmetros dos priors utilizados, do levantamento COSMOS. Note que k_t não se aplica às galáxias irregulares pois $P(T|m_0)$ é neste caso escolhido para que a soma seja 1.

modelos é assumida do tipo decaimento exponencial

$$p(T|m_0) = f_t \exp\left(-k_t(m_0 - 20)\right), \tag{3.6}$$

enquanto a probabilidade dos *redshifts* é modelada como

$$p(z|T, m_0) \propto z^{\alpha_t} \exp\left[-\left(\frac{z}{z_{mt}(m_0)}\right)^{\alpha_t}\right], \tag{3.7}$$

com $z_{mt}(m_0)$ dado por

$$z_{mt}(m_0) = z_{0t} + k_{mt}(m_0 - 20) \tag{3.8}$$

O conjunto dos onze parâmetros envolvidos nos priors pode ser determinado ajustando os priors à uma população de galáxias representativa com *redshifts* conhecidos. [6] descreve como o processo de ajuste pode ser modelado como um problema de inferência bayesiana muito semelhante ao que fizemos até aqui, e determina os parâmetros utilizando a amostra do Hubble Deep Field - North (HDF-N). O prior que utilizaremos aqui foi porém obtido utilizando uma amostra do levantamento COSMOS ([26]), cujos parâmetros são indicados na Tabela 3.1.

BPZ

O algoritmo Bayesian Photometric *Redshifts*(BPZ)[6, 7, 11] é uma implementação da estratégia delineada nas seções anteriores, escrita em Python. As entradas do algoritmo consistem em um catálogo, fornecido na forma de Tab Separated Values (tsv) contendo as magnitudes

em cada uma das bandas e seus erros associados. Também é necessário indicar ao algoritmo os filtros correspondentes a cada banda e a biblioteca de modelos a ser utilizada. O BPZ também é capaz de interpolar entre os modelos fornecidos.

Além de calcular a distribuição de probabilidades em *redshift*, o BPZ retorna um catálogo contendo a moda da distribuição z_B (isto é, o valor de *redshift* mais provável), o intervalo ao redor deste valor contendo 95 % da probabilidade total, um parâmetro de chance, correspondente à integral da distribuição numa região $\Delta z = 0.2(1 + z_B)$ ao redor do máximo, o valor do modelo correspondente ao melhor ajuste e o valor de χ^2 do ajuste obtido. O parâmetro de chance é especialmente útil para avaliar a qualidade das medidas: valores próximos a 1 indicam que a distribuição de probabilidade é um pico estreito, e portanto o valor obtido é confiável. Valores baixos do parâmetro de chance indicam distribuições largas ou com vários picos, portanto é possível reduzir a quantidade de outliers realizando cortes utilizando este parâmetro (é entretanto necessário cuidado, pois estes cortes, além de reduzir o tamanho da amostra, podem introduzir vieses inaceitáveis para certas análises).

Por fim, o BPZ também oferece a possibilidade de corrigir pequenos vieses na calibração da fotometria. Isto é possível rodando-o numa amostra com *redshift* espectroscópico conhecido, onde o algoritmo tentará inferir desvios em cada banda de forma a melhorar o desempenho. Este processo funciona como uma espécie de treinamento do algoritmo, embora se trate de um algoritmo de ajuste de modelos.

3.2 Portal Científico do DES

Um dos aspectos desafiadores de um projeto como o DES, que envolve massiva quantidade de dados e uma longa cadeia de processamento, é realizar uma análise otimizada ao mesmo tempo em que são registradas todas as versões dos dados e software utilizados, para garantir a reproducibilidade dos resultados obtidos.

No DES, a resposta a este problema passa pelo portal científico do DES. Este portal consiste numa interface, acessível por navegador web, onde estão implementadas pipelines de proces-

samento de dados realizando as várias tarefas envolvidas na redução de dados e nas análises científicas do DES. Este sistema automaticamente registra as configurações utilizadas em cada rodada. O portal científico foi desenvolvido pelo Linea como parte do investimento para garantir a participação brasileira no DES.

A estrutura do portal[24, 22] é organizada de forma que há 4 fases de processamento de dados. As três primeiras, Data installation, Data Preparation e Catalog Creation, coletivamente chamadas End to End (E2E), envolvem os passos necessários para a criação de catálogos prontos para as análises científicas propriamente ditas, chamados Valued Added Catalogs (VACs). A última fase consiste nas análises científicas propriamente ditas, chamada Science Pipelines. A Figura 3.2 ilustra esquematicamente a estrutura do portal.

O cálculo dos *redshifts* fotométricos ocorre na fase de Data Preparation, envolvendo 4 pipelines, descritas abaixo:

- Spectroscopic Sample: Nesta fase é criada a amostra espectroscópica. Este processo envolve a seleção dos levantamentos espectroscópicos a serem utilizados, resolução de duplicatas, isto é, casos em que uma mesma galáxia foi medida por vários levantamentos, e controle de qualidade.

- Training Set Maker: Nesta fase o catálogo a ser utilizado na validação e treinamento dos algoritmos de photo-z é criado. Isto envolve fazer o pareamento da amostra espectroscópica obtida no passo anterior com a fotometria do DES.

- Photo-z Training: Nesta etapa os algoritmos de photo-z são treinados e validados. É possível realizar diversos tipos de seleção e cortes de qualidade nos dados, e comparar métricas entre diversos algoritmos. Também são automaticamente produzidos diversos gráficos relevantes para a avaliação do algoritmo.

- Photo-z Compute: Por fim, nesta etapa são calculados redshits fotométricos para toda a amostra fotométrica utilizando as configurações selecionadas no passo anterior.

Há dezenas de algoritmos de *redshifts* fotométricos instalados no portal, incluindo a versão mais recente do BPZ (1.99.0+6). A maior parte dos resultados obtidos aqui são provenientes de análises realizadas no portal.

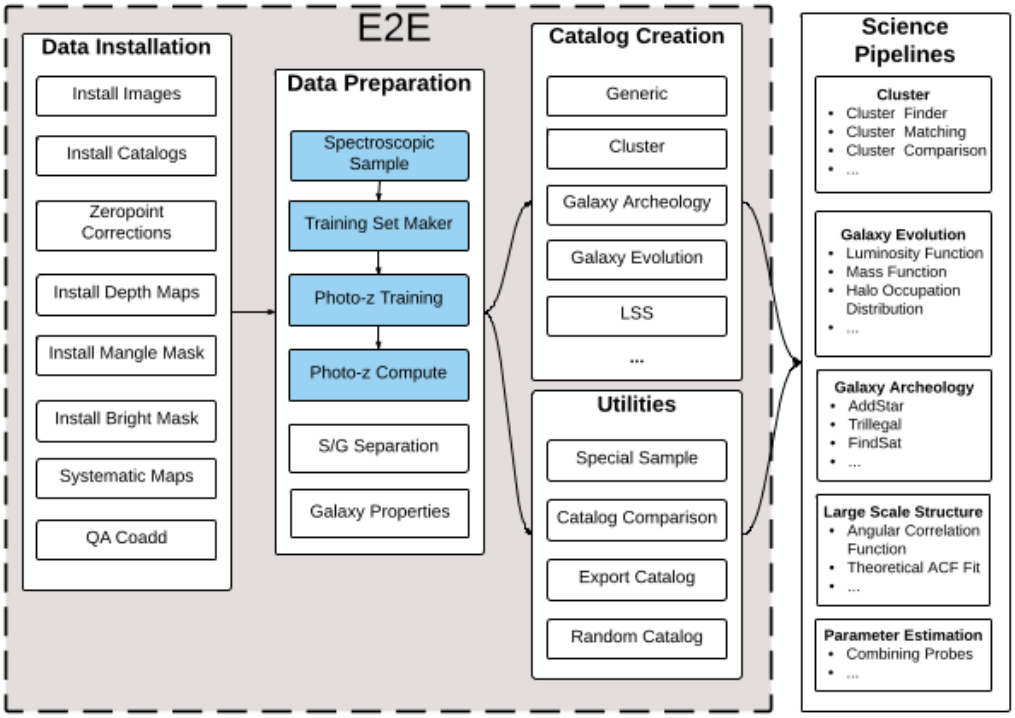

Figura 3.1: Estrutura do portal. Figura retirada de [24] - submetido.

4

Resultados

Neste capítulo discutiremos as rodadas realizadas com o BPZ nos dados do ano 3 do DES. Foi realizada uma rodada de validação numa amostra preparada a partir do casamento de uma amostra espectroscópica, obtida de levantamentos espectroscópicos diversos, e da amostra fotométrica do *release* Y3A2 do DES. Também inferimos os *redshifts* fotométricos para uma subamostra contendo cerca de 1 % das galáxias no catálogo `Y3A2_COADD`, a partir dos quais obtivemos a distribuição em *redshifts* das galáxias medidas. Esta mesma distribuição também foi inferida utilizando as distribuições de probabilidade completas fornecidas pelo BPZ.

4.1 Dados utilizados

Amostra espectroscópica

A amostra espectroscópica foi formada pela compilação de 17 levantamentos, totalizando 578.444 medições. Apenas objetos da melhor faixa de qualidade especificada pelos levantamentos foram utilizados. Duplicatas foram resolvidas considerando que fontes cuja separação angular é inferior a 1 segundo de arco pertencem ao mesmo objeto, sendo mantido o *redshift* medido com menor erro (ou do levantamento mais recente, em caso de empate). As várias amostras compiladas nesta amostra espectroscópica unificada estão descritas na Tabela 4.1.

Levantamento	**Número de objetos**	**Porcentagem**	***redshift* médio**
SDSS_DR7[1]	90332	23.5 %	0.17
WIGGLEZ[35]	80893	21.1%	0.59
DR12_CMASS[2]	51728	13.5 %	0.57
VIPERS[23]	50903	13.3 %	0.69
DEEP2[16, 17]	35314	9.2 %	0.87
EBOSS_DES_ELG [13]	19874	5.2%	0.92
DES_AAOMEGA [42]	14437	3.8%	0.6
ZCOSMOS[27]	12813	3.3 %	0.54
SDSS_OZDES[43]	9055	2.4 %	0.53
GAMA[20]	7997	2.1%	0.22
ACES[14]	4746	1.2%	0.59
SPT_GMOS[5]	2195	0.6%	0.54
UDS[9, 30]	1417	0.4%	1.14
ATLAS[29]	805	0.2%	0.38
CDB[39]	515	0.1%	0.60
SPARCS[33]	403	0.1%	0.92
VUDS[40]	283	0.1%	2.21

Tabela 4.1: Levantamentos espectroscópicos utilizados na compilação da amostra espectroscópica utilizada.

4.2 Amostra fotométrica

A amostra fotométrica utilizada provêm do *release* Y3A2 do DES. Esta amostra consiste nos dados já reduzidos e coadicionados, isto é, as exposições de cada filtro em várias épocas já passou pelo processo de redução, que é realizado pelo DES Data Management (DESDM)[31] e envolve diversas correções compreendendo efeitos instrumentais da DECam (como não-linearidade, ruído térmico e bad pixels, para citar alguns), calibração da fotometria, correção de efeitos

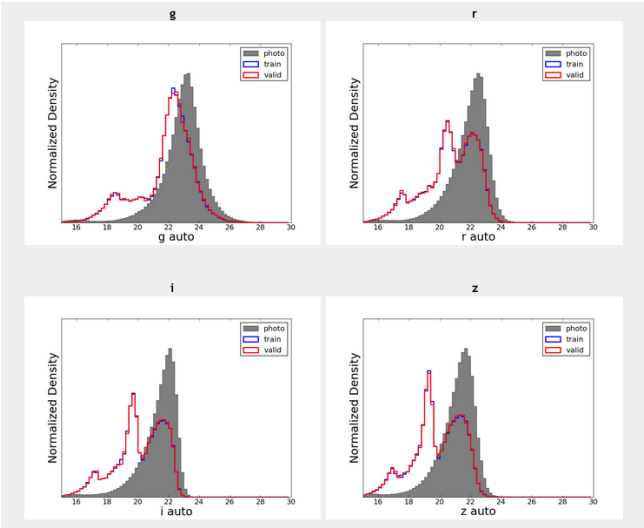

Figura 4.1: Histogramas das distribuições nas bandas g, r, i e z dos objetos na amostra fotométrica (cinza) e nas subamostras de treinamento e validação (azul e vermelho).

atmosféricos e várias outras. Como parte deste processo, fontes (como galáxias e estrelas) são detectadas e medidas com o programa `Sextractor` [8]. Além disso, as várias exposições de época única (i.e., individuais), da mesma região do céu são empilhadas originando uma imagem mais profunda e com relação sinal/ruído melhor. Este processo é descrito detalhadamente em [21, 31].

Em suma, ao fim da redução obtêm-se um catálogo de objetos com suas posições e várias medidas de magnitude e morfologia nas 5 bandas espectrais utilizadas pelo DES. A amostra completa contêm 399.263.026 objetos. A distribuição dos objetos em cada um dos filtros utilizados pode ser observada na Figura 4.1.

Como galáxias são corpos extensos a medição de magnitude possui certa ambiguidade. Por isto há vários tipos de magnitudes medidos para cada banda, a depender da maneira como se define a região onde a luz será medida. Utilizaremos aqui as magnitudes do tipo `MAG_AUTO`, onde se define uma abertura elíptica englobando a galáxia, com tamanho inferido numa banda de referência (no nosso caso, a banda i). Esta medida pode estar sujeita a inclusão de luz de outros objetos quando o brilho superficial é baixo. Para mitigar a contaminação de vizinhos,

foram mantidos apenas objetos para os quais `Sextractor` reportou uma detecção limpa.

Separação Estrela/Galáxia e filtragem de outliers

Outro ponto importante é que nem todos os objetos detectados são galáxias: há também estrelas e outliers como satélites ou mesmo artefatos do sistema de detecção. Para separar estrelas de galáxias, foi utilizado o classificador `Y1_MODEST_CLASS` baseado na morfologia do objeto, desenvolvido para as análises da liberação de dados do ano 1. Para isto, é necessário corrigir efeitos atmosféricos, que costumam ser modelados por uma função gaussiana que "borra" as imagens dos objetos, isto é, a distribuição espacial de fluxo observada corresponde à distribuição espacial de fluxo oriunda do objeto convolucionada com esta Point Spread Function (PSF). Corrigindo efeitos da PSF, estrelas são fontes pontuais e galáxias fontes extensas. `Y1_MODEST_CLASS` aproveita-se deste fato, e de quantidades medidas pelo `Sextractor` que quantificam a área ocupada pelo objeto para separar estrelas de galáxias. Já que alguns objetos, como galáxias distantes e quasares, são quase pontuais, o classificador também utiliza a informação da magnitude na banda i. Uma vez aplicado, restam 310.470.517 galáxias na amostra utilizada.

Outros outliers são retirados realizando cortes de cor. São descartados todos os objetos cujas cores $g - r$, $r - i$ e $i - z$ não estejam no intervalo $[-2, 4]$.

Subamostra fotométrica

A fim de demonstrar todo o processo de obtenção de *redshifts* fotométricos, selecionamos uma subamostra dos dados do catálogo `Y3A2_COADD` a partir de uma requisição SQL tal que cada galáxia possui uma probabilidade de 1 % de ser selecionada. Esta subamostra possui 3.993.267 objetos, e devido ao processo de seleção aleatório possui as mesmas características da amostra fotométrica total, como cores e distribuições de magnitudes em cada banda. A distribuição espacial desta amostra pode ser vista na Figura 4.2 (à esquerda).

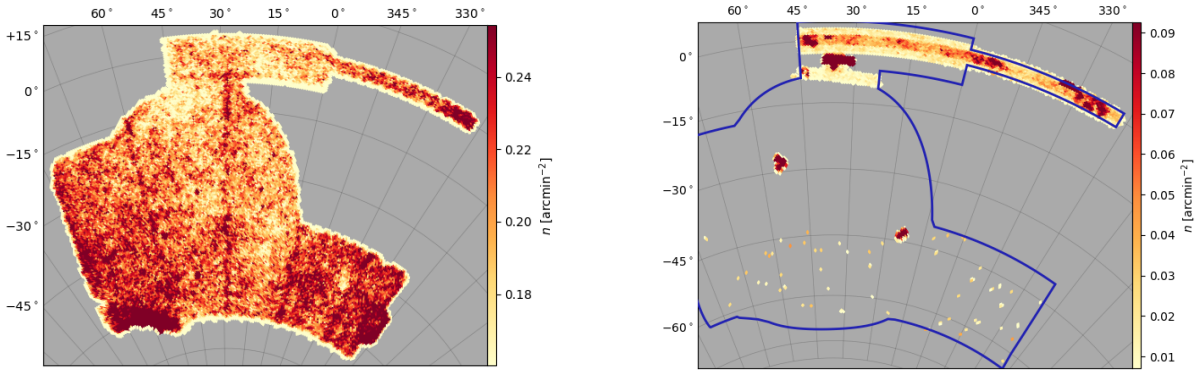

Figura 4.2: Mapas celestes da densidade de objetos para a subamostra fotométrica utilizada (esquerda) e para a amostra de validação (direita). O eixo horizontal corresponde à ascensão reta e o vertical à declinação.

4.3 Amostra de validação

A amostra de validação foi criada a partir do casamento das amostras espectroscópicas e fotométricas previamente mencionadas, determinando que a fotometria e a espectroscopia pertencem ao mesmo objeto caso as duas observações distem menos de 1 segundo de arco. Esta amostra foi subdividida em duas metades, um conjunto de treinamento e um conjunto de validação. O conjunto de treinamento foi utilizado para determinar os *shifts* de correção da fotometria enquanto o conjunto de validação para avaliar o desempenho do algoritmo. Também foi realizado um corte, retirando todos os objetos cuja relação sinal/ruído da detecção é menor que 5, após o qual cada uma das subamostras passou a conter 68.136 galáxias. As distribuições de magnitude de cada uma das subamostras em cada filtro pode ser observada na Figura 4.1.

4.4 Rodadas realizadas

Em todas as rodadas, utilizamos a versão 1.99.0 do BPZ e a biblioteca de SEDs CWW+SB [12, 25]. Os *priors* utilizados são aqueles descritos na seção 3.1, calibrados no levantamento COSMOS.

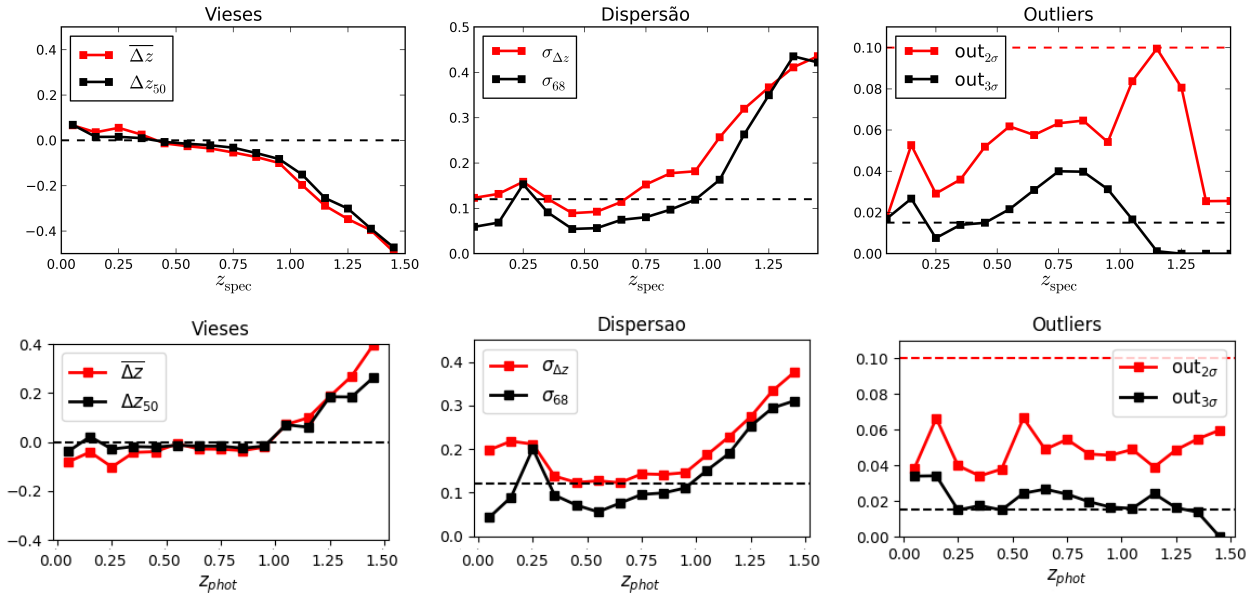

Figura 4.3: Evolução das métricas em *redshift* (espectroscópico e fotométricos na primeira e segunda linhas, respectivamente). As linhas tracejadas indicam os valores máximos aceitáveis para análises científicas do DES.

Rodada de validação

Rodamos o BPZ na subamostra de treinamento utilizando a flag `ONLY_TYPE`. Neste modo o BPZ calcula as correções fotométricas em cada uma das bandas de forma a ajustar os *redshifts* fotométricos inferidos aos *redshifts* espectroscópicos da amostra. Foram obtidos os valores [-0.01, 0.013, -0.002, -0.001] para os *shifts* nas bandas g, r, i e z, respectivamente. A banda Y não foi utilizada.

Utilizando os *shifts* descobertos, executamos o BPZ na subamostra de validação. As métricas de caracterização que discutimos anteriormente foram calculadas (cf. Tabela 4.2), bem como sua evolução com *redshift* (cf. Figura 4.3).

Métrica	$\overline{\Delta z}$	Δz_{50}	$\sigma_{\Delta z}$	σ_{68}	out$_{2\sigma}$	out$_{3\sigma}$	$\overline{\Delta z}'$	$\sigma_{\Delta z'}$	$N_{poisson}$
Valor	-0.0311	-0.0148	0.1613	0.0869	0.0515	0.0239	-0.1044	0.7485	16.1596

Tabela 4.2: Métricas calculadas para a rodada de validação do BPZ.

Observamos que os valores globais das métricas de dispersão estão aceitáveis: σ_{68} está dentro dos requisitos científicos do DES (que estipulam um valor máximo de 0.12). Note ainda que a distribuição $\Delta z'$ é bem comportada: o valor de $\sigma_{\Delta z'}$ é próximo a 1, indicando que as incertezas estimadas para os *redshifts* são confiáveis. Erros catastróficos e vieses não se comportam muito bem, entretanto: embora $\text{out}_{2\sigma}$ obedeça à restrição requerida (máximo de 0.1), $\text{out}_{3\sigma}$ é consideravelmente superior ao limite aceitado de 0.015, e $\overline{\Delta z}$ e Δz_{50} ambos mostram considerável desvio sistemático.

Comparando nossas métricas com as obtidas pela colaboração nos dados do período de verificação científica [38] notamos que o desempenho em dispersão é bom: os melhores algoritmos naquele trabalho atingem um σ_{68} próximo a 0.08, comparável ao obtido aqui. Este valor também se aproxima do observado no levantamento CFHTLS[15, 38] ($\sigma_{68} \approx 0.085$), que possui características semelhantes ao DES. O desempenho observado é consideravelmente melhor que o do BPZ em [38], que atinge cerca de 0.1 na mesma métrica, o que pode ser atribuído ao treinamento que realizamos (não realizado por [38]), ou à diferente escolha de priors utilizada.

Os resultados obtidos para os vieses e erros catastróficos também são típicos. Na análise de [38] observa-se que códigos de ajuste de modelos costumam apresentar vieses maiores, com $|\overline{\Delta z}|$ de 0.03 ou mais. O resultado obtido aqui se aproxima bastante do obtido pelo BPZ naquele trabalho, de cerca de -0.03. Isto pode indicar que haja algum desvio sistemático no conjunto de modelos utilizado. Nossos resultados para erros catastróficos são um pouco piores, mas próximos aos observados no período de verificação científica: a maior parte dos algoritmos apresentou $\text{out}_{3\sigma}$ entre 0.015 e 0.022. Há espaço para melhora, entretanto: os requisitos científicos do DES permitem que sejam descartados até 10 % da amostra com as piores incertezas na determinação do *redshift* fotométrico, o que não fizemos aqui. Realizando este corte, ou utilizando o parâmetro `ODDS` do BPZ descrito anteriormente é possível que estes valores melhorem consideravelmente.

Na análise da evolução das métricas em *redshift*, a primeira constatação é que a dispersão é consideravelmente ruim para *redshifts* pequenos, o que é um comportamento estranho, já que estes estão mais próximos e possuem uma fotometria melhor. Este fenômeno também é

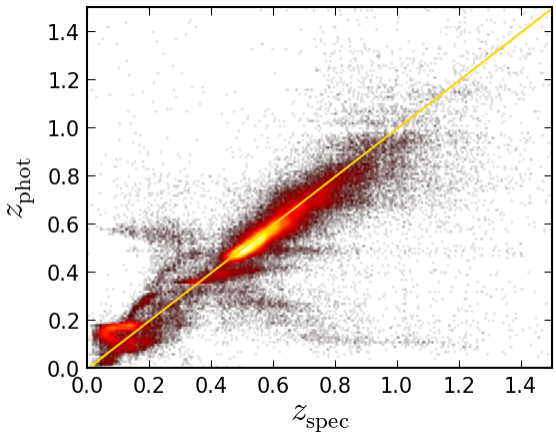

Figura 4.4: Gráfico do *redshift* fotométrico z_{photo} vs. *redshift* espectroscópico z_{spec} para a amostra de validação. A linha amarela corresponde a $z_{photo} = z_{spec}$, e a cor é proporcional à densidade de pontos.

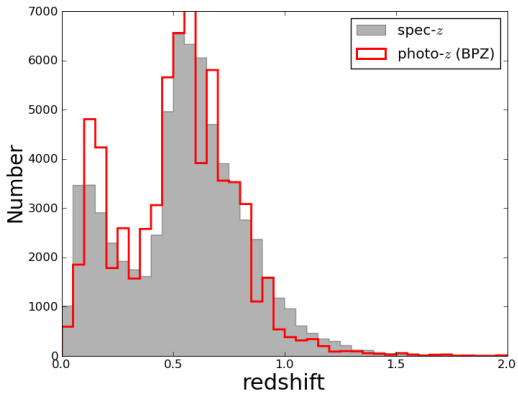

Figura 4.5: Histograma dos *redshifts* espectroscópicos e fotométricos para a amostra de validação. Observe que o BPZ é razoavelmente capaz de reproduzir a distribuição verdadeira de *redshifts*.

observado no gráfico de z_{photo} vs. z_{spec} (Figura 4.4). Trata-se na verdade de um fenômeno bem conhecido, causado pela degenerescência entre as quebras de Lyman e de 4000 Angstrom, dois artefatos importantes, correspondentes a uma brusca queda no fluxo para comprimentos de onda menores que 912 Angstrom (o menor comprimento possível para a série de Lyman) e 4000 Angstrom, respectivamente. Apenas com as informações disponíveis nos filtros g, r, i e z pode ser impossível distinguir a quebra de 4000 Angstrom de uma galáxia próxima da quebra de Lyman de uma galáxia mais distante e avermelhada. Isto confunde o algoritmo, daí o aumento na dispersão. Um remédio possível é adicionar um filtro no ultravioleta próximo, que ajuda a distinguir as duas situações, entretanto isto não está no planejamento do DES. Outros algoritmos lidam melhor com a situação, o que indica que talvez seja possível mitigar este problema introduzindo novos modelos bem escolhidos, como realizado por [32] para o levantamento CLASH, onde com a introdução de um novo modelo de galáxia elíptica mais avermelhada corrigiu-se um problema semelhante em baixos *redshifts*.

À parte disto, notamos que as métricas mantêm-se relativamente constantes para *redshifts* entre 0 e 1. A partir daí tanto vieses quanto a dispersão crescem fortemente. Neste regime as galáxias são mais fracas, e portanto sua fotometria é mais ruidosa e este comportamento é esperado, embora seja possível que cortes cuidadosos dos objetos com maiores erros, ou escolhas mais adequadas para a fotometria (restringindo o tamanho da região utilizada para obter as magnitudes `MAG_AUTO` com o `Sextractor`, por exemplo) sejam capazes de melhorar os resultados.

Rodada na subamostra fotométrica

Após a validação, a fim de demonstrar o processo completo de inferência de *redshifts* fotométricos, executamos o BPZ na subamostra contendo 1 % dos objetos de `Y3A2_COADD`, utilizando os shifts inferidos durante a validação. Com este catálogo podemos estimar a distribuição em *redshift* das galáxias medidas. Há na verdade dois métodos possíveis: um deles consiste em realizar um histograma dos valores z_B dos *redshifts* fotométrico fornecidos pelo BPZ, que correspondem aos picos da densidade de probabilidade $p(z)$. Outra abordagem consiste em armazenar toda a distribuição de probabilidade para todos os objetos. A distribuição em *redshift* das galáxias pode ser então estimada somando as distribuição de probabilidade individuais. A Figura 4.6 ilustra os resultados destas duas abordagens.

A obtenção de um catálogo como este é o ponto de partida de uma gama de análises científicas em cosmologia e astrofísica extragaláctica. Exemplos incluem a investigação da estrutura do Universo em grande escala usando a função de correlação de 2 pontos (onde podem ser medidas as oscilações acústicas bariônicas), o estudo dos processos de evolução galáctica, explorando a função de luminosidade e sua evolução com *redshift* e a investigação do conteúdo de matéria, principalmente escura, presentes nas galáxias, aglomerados e nos vazios do espaço extragaláctico, utilizando os tênues efeitos de lenteamento gravitacional detectáveis por correlações entre os achatamentos de galáxias.

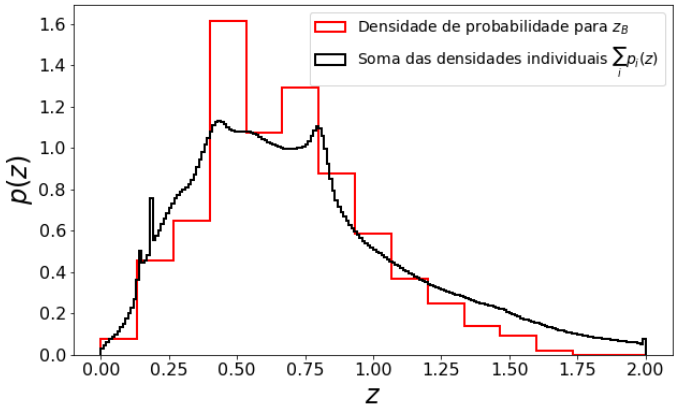

Figura 4.6: Densidade de probabilidade associada à distribuição em *redshift* da subamostra fotométrica estudada.

5

Conclusão

Com os avanços recentes da cosmologia experimental, que se tornou uma disciplina rica em dados, é evidente que serão crescentes os desafios na análise destes. *Redshifts* fotométricos são críticos neste processo, pois são essenciais para a realização de levantamentos de grande escala, e é necessária uma análise cuidadosa para garantir que erros nesta etapa não se propagem por toda a análise. Neste contexto, e tendo em vista os experimentos e resultados reportados aqui, torna-se clara a importância do estabelecimento de uma ferramenta como o Portal Científico do DES, que fornece flexibilidade ao mesmo tempo em que garante a reprodutibilidade das análises.

Além disto, concluímos que o código BPZ é capaz de produzir estimativas de *redshift* fotométrico com a qualidade necessária para análises científicas. Entretanto também é claro que há desafios e espaços para melhoras em muitas direções, que foram pouco exploradas neste trabalho. Observamos que com o aumento do *redshift* a qualidade das estimativas de photo-z cai, o que é um padrão conhecido. Nas próximas gerações de levantamentos, a magnitude limite das observações tende a crescer, e será importante obter estimativas confiáveis de photo-z para objetos em altos *redshifts*, o que é desafiador, pois é difícil medir objetos tão tênues com espectroscopia, o que faz com que amostras de treinamento para algoritmos de aprendizado de

máquina sejam limitadas e enviesadas nesta região. Por outro lado, algoritmos de ajuste de modelos são dependentes de boas bibliotecas de modelos, e é difícil obtê-las neste regime, já que estamos olhando para objetos muito antigos sobre os quais o conhecimento é limitado e podem existir novas classes de galáxias desconhecidas, tornando nossa bibliotecas incompletas.

Para enfrentar este problemas análises mais complexas vem sendo desenvolvidas, desde as mais simples envolvendo a utilização de pesos para que as amostras de treinamento e validação tornem-se mais representativas da fotometria utilizada, quanto técnicas mais sofisticadas envolvendo estudos de correlação cruzada[34] ou simulações. Algumas outras possibilidades a serem exploradas futuramente incluem o treinamento de algoritmos empíricos em dados simulados e a utilização de SEDs teóricas para galáxias jovens, obtidas a partir de modelos de síntese de populações estelares.

Além da fronteira em profundidade, outra grande desafio futuro reside na precisão. As medidas do DES já possuem precisão numa escala de poucos pontos percentuais, e neste cenário erros sistemáticos no processo de redução de dados tornam-se muito importantes. O crescimento da fotometria de banda média e curta, onde as observações não são necessariamente mais profundas, mas envolvem um número maior de filtros, como é o caso dos levantamentos J-PLUS (16 filtros), ALHAMBRA (20 filtros), COSMOS (dezenas de filtros em várias faixas do espectro eletromagnético) e J-PAS (56 filtros) também será importante. A obtenção de *redshifts* fotométricos neste caso é levada a outra escala de precisão, criando naturalmente novos desafios. A cobertura espectral também possui um impacto importante, e a análise conjunta de levantamentos em diferentes faixas espectrais pode aumentar consideravelmente a precisão. No caso do DES, a inclusão da fotometria no infravermelho próximo do levantamento VISTA pode diminuir a dispersão em até 30 % [4].

Em suma, *redshifts* são quantidades essenciais para uma miríade de aplicações, e técnicas fotométricas de inferência são uma ferramenta importante, quando não essencial, para estudos contemporâneos.

Bibliografia

[1] K. N. Abazajian et al. “The Seventh Data Release of the Sloan Digital Sky Survey”. Em: ApJS 182, 543-558 (jun. de 2009), pp. 543–558. DOI: 10.1088/0067-0049/182/2/543. arXiv: 0812.0649.

[2] S. Alam et al. “The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample”. Em: MNRAS 470 (set. de 2017), pp. 2617–2652. DOI: 10.1093/mnras/stx721. arXiv: 1607.03155.

[3] A. Albrecht et al. “Report of the Dark Energy Task Force”. Em: *ArXiv Astrophysics e-prints* (set. de 2006). eprint: astro-ph/0609591.

[4] M. Banerji et al. “Photometric redshifts for the Dark Energy Survey and VISTA and implications for large-scale structure”. Em: MNRAS 386 (maio de 2008), pp. 1219–1233. DOI: 10.1111/j.1365-2966.2008.13095.x. arXiv: 0711.1059.

[5] M. B. Bayliss et al. “SPT-GMOS: A Gemini/GMOS-South Spectroscopic Survey of Galaxy Clusters in the SPT-SZ Survey”. Em: ApJS 227, 3 (nov. de 2016), p. 3. DOI: 10.3847/0067-0049/227/1/3. arXiv: 1609.05211.

[6] N. Benítez. “Bayesian Photometric Redshift Estimation”. Em: ApJ 536 (jun. de 2000), pp. 571–583. DOI: 10.1086/308947. eprint: astro-ph/9811189.

[7] N. Benítez et al. “Faint Galaxies in Deep Advanced Camera for Surveys Observations”. Em: ApJS 150 (jan. de 2004), pp. 1–18. DOI: 10.1086/380120. eprint: astro-ph/0309077.

[8] E. Bertin e S. Arnouts. “SExtractor: Software for source extraction.” Em: A&AS 117 (jun. de 1996), pp. 393–404. DOI: 10.1051/aas:1996164.

[9] E. J. Bradshaw et al. “High-velocity outflows from young star-forming galaxies in the UKIDSS Ultra-Deep Survey”. Em: MNRAS 433 (jul. de 2013), pp. 194–208. DOI: 10.1093/mnras/stt715. arXiv: 1304.7276.

[10] D. Coe et al. “A High-resolution Mass Map of Galaxy Cluster Substructure: LensPerfect Analysis of A1689”. Em: ApJ 723 (nov. de 2010), pp. 1678–1702. DOI: 10.1088/0004-637X/723/2/1678. arXiv: 1005.0398 [astro-ph.CO].

[11] D. Coe et al. “Galaxies in the Hubble Ultra Deep Field. I. Detection, Multiband Photometry, Photometric Redshifts, and Morphology”. Em: AJ 132 (ago. de 2006), pp. 926–959. DOI: 10.1086/505530. eprint: astro-ph/0605262.

[12] G. D. Coleman, C.-C. Wu e D. W. Weedman. “Colors and magnitudes predicted for high redshift galaxies”. Em: ApJS 43 (jul. de 1980), pp. 393–416. DOI: 10.1086/190674.

[13] J. Comparat et al. “SDSS-IV eBOSS emission-line galaxy pilot survey”. Em: A&A 592, A121 (ago. de 2016), A121. DOI: 10.1051/0004-6361/201527377. arXiv: 1509.05045.

[14] M. C. Cooper et al. “The Arizona CDFS Environment Survey (ACES): A Magellan/IMACS Spectroscopic Survey of the Chandra Deep Field-South”. Em: MNRAS 425 (set. de 2012), pp. 2116–2127. DOI: 10.1111/j.1365-2966.2012.21524.x. arXiv: 1112.0312.

[15] J. Coupon et al. “Photometric redshifts for the CFHTLS T0004 deep and wide fields”. Em: A&A 500 (jun. de 2009), pp. 981–998. DOI: 10.1051/0004-6361/200811413. arXiv: 0811.3326.

[16] M. Davis et al. “Science Objectives and Early Results of the DEEP2 Redshift Survey”. Em: *Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II*. Ed. por P. Guhathakurta. Vol. 4834. Proc. SPIE. Fev. de 2003, pp. 161–172. DOI: 10.1117/12.457897. eprint: astro-ph/0209419.

[17] M. Davis et al. “The All-Wavelength Extended Groth Strip International Survey (AEGIS) Data Sets”. Em: ApJ 660 (maio de 2007), pp. L1–L6. DOI: 10.1086/517931. eprint: astro-ph/0607355.

[18] T. M. Davis e C. H. Lineweaver. “Expanding Confusion: Common Misconceptions of Cosmological Horizons and the Superluminal Expansion of the Universe”. Em: PASA 21 (2004), pp. 97–109. DOI: 10.1071/AS03040. eprint: astro-ph/0310808.

[19] T. M. Davis et al. “The Effect of Peculiar Velocities on Supernova Cosmology”. Em: ApJ 741, 67 (nov. de 2011), p. 67. DOI: 10.1088/0004-637X/741/1/67. arXiv: 1012.2912.

[20] S. P. Driver et al. “Galaxy and Mass Assembly (GAMA): survey diagnostics and core data release”. Em: MNRAS 413 (maio de 2011), pp. 971–995. DOI: 10.1111/j.1365-2966.2010.18188.x. arXiv: 1009.0614.

[21] A. Drlica-Wagner et al. “Dark Energy Survey Year 1 Results: Photometric Data Set for Cosmology”. Em: *ArXiv e-prints* (ago. de 2017). arXiv: 1708.01531.

[22] A. Fausti Neto et al. “DES Science Portal: II- Creating Science-Ready Catalogs”. Em: *ArXiv e-prints* (ago. de 2017). arXiv: 1708.05642 [astro-ph.IM].

[23] B. Garilli et al. “The VIMOS Public Extragalactic Survey (VIPERS). First Data Release of 57 204 spectroscopic measurements”. Em: A&A 562, A23 (fev. de 2014), A23. DOI: 10.1051/0004-6361/201322790. arXiv: 1310.1008.

[24] J. Gschwend et al. “DES Science Portal: I - Computing Photometric Redshifts”. Em: *ArXiv e-prints* (ago. de 2017). arXiv: 1708.05643.

[25] A. L. Kinney et al. "Template Ultraviolet to Near-Infrared Spectra of Star-forming Galaxies and Their Application to K-Corrections". Em: ApJ 467 (ago. de 1996), p. 38. DOI: 10.1086/177583.

[26] C. Laigle et al. "The COSMOS2015 Catalog: Exploring the 1 < z < 6 Universe with Half a Million Galaxies". Em: ApJS 224, 24 (jun. de 2016), p. 24. DOI: 10.3847/0067-0049/224/2/24. arXiv: 1604.02350.

[27] S. J. Lilly et al. "The zCOSMOS 10k-Bright Spectroscopic Sample". Em: ApJS 184 (out. de 2009), pp. 218–229. DOI: 10.1088/0067-0049/184/2/218.

[28] H. Lin, B. Flaugher e Dark Energy Survey Collaboration. "The Dark Energy Survey". Em: *American Astronomical Society Meeting Abstracts #214*. Vol. 214. American Astronomical Society Meeting Abstracts. Maio de 2009, p. 669.

[29] M. Y. Mao et al. "The Australia Telescope Large Area Survey: spectroscopic catalogue and radio luminosity functions". Em: MNRAS 426 (nov. de 2012), pp. 3334–3348. DOI: 10.1111/j.1365-2966.2012.21913.x. arXiv: 1208.2722.

[30] R. J. McLure et al. "The sizes, masses and specific star formation rates of massive galaxies at 1.3 < z < 1.5: strong evidence in favour of evolution via minor mergers". Em: MNRAS 428 (jan. de 2013), pp. 1088–1106. DOI: 10.1093/mnras/sts092. arXiv: 1205.4058.

[31] J. J. Mohr et al. "The Dark Energy Survey data processing and calibration system". Em: *Software and Cyberinfrastructure for Astronomy II*. Vol. 8451. Proc. SPIE. Set. de 2012, p. 84510D. DOI: 10.1117/12.926785. arXiv: 1207.3189 [astro-ph.IM].

[32] A. Molino et al. "CLASH: accurate photometric redshifts with 14 HST bands in massive galaxy cluster cores". Em: MNRAS 470 (set. de 2017), pp. 95–113. DOI: 10.1093/mnras/stx1243. arXiv: 1705.02265.

[33] A. Muzzin et al. "The Gemini Cluster Astrophysics Spectroscopic Survey (GCLASS): The Role of Environment and Self-regulation in Galaxy Evolution at z ˜ 1". Em: ApJ 746, 188 (fev. de 2012), p. 188. DOI: 10.1088/0004-637X/746/2/188. arXiv: 1112.3655.

[34] J. A. Newman. "Calibrating Redshift Distributions beyond Spectroscopic Limits with Cross-Correlations". Em: ApJ 684, 88-101 (set. de 2008), pp. 88–101. DOI: 10.1086/589982. arXiv: 0805.1409.

[35] D. Parkinson et al. "The WiggleZ Dark Energy Survey: Final data release and cosmological results". Em: Phys. Rev. D 86.10, 103518 (nov. de 2012), p. 103518. DOI: 10.1103/PhysRevD.86.103518. arXiv: 1210.2130.

[36] S. Perlmutter et al. "Measurements of Ω and Λ from 42 High-Redshift Supernovae". Em: ApJ 517 (jun. de 1999), pp. 565–586. DOI: 10.1086/307221. eprint: astro-ph/9812133.

[37] A. G. Riess et al. "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant". Em: AJ 116 (set. de 1998), pp. 1009–1038. DOI: 10.1086/300499. eprint: astro-ph/9805201.

[38] C. Sánchez et al. "Photometric redshift analysis in the Dark Energy Survey Science Verification data". Em: MNRAS 445 (dez. de 2014), pp. 1482–1506. DOI: 10.1093/mnras/stu1836. arXiv: 1406.4407 [astro-ph.IM].

[39] M. Sullivan et al. "VizieR Online Data Catalog: Type Ia supernovae luminosities (Sullivan+, 2010)". Em: *VizieR Online Data Catalog* 740 (jan. de 2011).

[40] L. A. M. Tasca et al. "The VIMOS Ultra Deep Survey first data release: Spectra and spectroscopic redshifts of 698 objects up to z_{spec} 6 in CANDELS". Em: A&A 600, A110 (abr. de 2017), A110. DOI: 10.1051/0004-6361/201527963. arXiv: 1602.01842.

[41] R. Wojtak, T. M. Davis e J. Wiis. "Local gravitational redshifts can bias cosmological measurements". Em: J. Cosmology Astropart. Phys. 7, 025 (jul. de 2015), p. 025. DOI: 10.1088/1475-7516/2015/07/025. arXiv: 1504.00718.

[42] F. Yuan et al. "OzDES multifibre spectroscopy for the Dark Energy Survey: first-year operation and results". Em: MNRAS 452 (set. de 2015), pp. 3047–3063. DOI: 10.1093/mnras/stv1507. arXiv: 1504.03039.

[43] F. Yuan et al. “OzDES multifibre spectroscopy for the Dark Energy Survey: first-year operation and results”. Em: MNRAS 452 (set. de 2015), pp. 3047–3063. DOI: 10.1093/mnras/stv1507. arXiv: 1504.03039.

Apêndices

A

Aplicações do redshift à cosmologia

O *redshift* de uma fonte astronômica é uma informação extremamente importante em cosmologia. Discutiremos aqui brevemente as razões, começando por um panorama dos tipos de fenômenos físicos responsáveis.

Efeito Doppler

Suponha que uma fonte de radiação eletromagnética se afaste radialmente de um observador com velocidade v. Da relatividade especial, sabemos que os comprimentos e intervalos de tempos medidos no referencial de repouso da fonte e do observador serão diferentes. De fato, utilizando as transformações de Lorentz mostra-se que o observador medira um *redshift* z dado pela equação

$$1 + z = \sqrt{\frac{1 + \frac{v}{c}}{1 - \frac{v}{c}}}. \tag{A.1}$$

Este é o denominado efeito Doppler relativístico. No limite não-relativístico $v/c \ll 1$ recuperamos a relação

$$z = \frac{\Delta\lambda}{\lambda} = \frac{v}{c}. \tag{A.2}$$

Redshift cosmológico

Num universo em expansão ou contração, o comprimento de onda de radiações eletromagnéticas varia juntamente com o espaço. De fato, se a expansão do Universo for descrita por um fator de escala $a(t)$ que varia com a idade do universo t e vale $a_0 = 1$ atualmente, objetos a distâncias d_0 hoje distam $d(t) = a(t)d_0$ quando o universo possui idade t. Assumindo que a luz emitida por uma galáxia distante expanda juntamente com o Universo em sua viagem até nossos olhos ou CCDs, concluímos que o *redshift* observado z estará relacionado com o fator de escala $a(t)$ no instante em que a luz foi emitida por

$$1 + z = \frac{\lambda'}{\lambda} = \frac{a_0}{a(t)} = \frac{1}{a(t)}, \tag{A.3}$$

pois as razões entre o comprimento de onda observado λ' e o emitido λ e entre o fator de escala na atualidade e no instante em que a luz foi emitida devem ser iguais.

Este tipo de *redshift* é denominado *redshift* cosmológico. É possível obtê-lo de forma mais rigorosa estudando as geodésicas tipo-luz na métrica de Friedmann-Robertson-Walker, que é tipicamente utilizada para modelar o espaço-tempo em escalas grandes o bastante.

Redshift gravitacional

A relatividade geral prevê que a passagem do tempo é afetada pela distribuição de matéria e energia. De fato, no limite de campos fracos da relatividade geral é possível mostrar que quando uma fonte de radiação está em um potencial gravitacional ϕ_{fonte}, enquando o observador está em um potencial ϕ_{obs}, o observador detecta um *redshift* da radiação dado por

$$z = \frac{\phi_{obs} - \phi_{fonte}}{c^2}. \tag{A.4}$$

Este tipo de *redshift* é conhecido como *redshift* gravitacional.

Redshift total

No cenário mais geral possível, uma fonte astronômica está sujeita aos 3 fenômenos descritos nas subseções anteriores. Devido à expansão do Universo, observamos um *redshift* z_{cosmo}. Além disso, a fonte pode possuir uma velocidade peculiar, isto é, uma velocidade no seu referencial local, já descontando os efeitos da expansão do Universo, que por sua vez causa um *redshift* Doppler z_{pec}. Por fim, a presença de flutuações de densidade no caminho entre a fonte e o observador pode gerar um *redshift* gravitacional z_{grav}. O *redshift* total observado, z, é a combinação dos anteriores,

$$1 + z = (1 + z_{grav})(1 + z_{pec})(1 + z_{cosmo}). \tag{A.5}$$

Segundo [19], as velocidades peculiares geram *redshifts* da ordem de 10^{-3}, enquanto [41] argumenta que *redshifts* gravitacionais são da ordem de 10^{-5} a 10^{-4}. Os *redshifts* devidos à velocidades peculiares e flutuações de densidade são entretanto distribuídos aleatoriamente. Considerando uma grande amostra de fontes, na média seus efeitos se cancelam, e o *redshift* medido corresponde essencialmente ao *redshift* cosmológico.[1]

A.1 Distâncias e *redshifts*

Uma das utilidades mais importantes dos *redshifts* é a determinação das distâncias comóveis (i.e., das distâncias num sistema de coordenadas que expande juntamente com o Universo). Considere um universo plano, homogêneo e isotrópico em expansão. Este universo é bem descrito por uma métrica de Friedmann-Robertson-Walker (FRW)

$$ds^2 = -c^2dt^2 + a^2(t)\left(d\chi^2 + \chi^2 d\Omega^2\right), \tag{A.6}$$

[1] Um efeito que pode, porém, enviesar medidas de *redshift* cosmológico é a presença de uma flutuação local de densidade, que adicionaria um *redshift* gravitacional a todas as medidas. De fato parece haver indícios de que estamos numa região com densidade abaixo da média, e isto pode afetar medidas precisas de parâmetros cosmológicos[41].

onde $a(t)$ é o fator de escala, χ são as coordenadas comóveis e $d\Omega = \sin\theta d\theta d\phi$ é um diferencial de ângulo sólido.

A distância comóvel de um ponto à origem no instante t é obtida a partir da integração de $d\chi$ na trajetória de um fóton emitido radialmente da fonte em t_{em} e detectado agora, em t_0. Trajetórias de fótons são geodésicas nulas, portanto $ds = 0$ e obtemos na métrica FRW $d\chi = c/a(t)dt$, portanto

$$\chi = \int_{t_{em}}^{t_0} \frac{cdt}{a(t)}$$

Lembrando da equação (A.3), porém, se definirmos o fator de escala a de forma que hoje $a(t_0) = a_0 = 1$ temos que o *redshift* da fonte e o fator de escala no instante da emissão do fóton se relacionam por $1 + z = \frac{1}{a(t_{em})}$. Derivando esta relação podemos mudar variáveis na integral acima

$$\chi = \int_0^z \frac{cdz'}{H(z')}, \tag{A.7}$$

onde $H(z) = \frac{\dot{a}(z)}{a(z)}$ é o parâmetro de Hubble para *redshift* z. A equação anterior nos diz portanto que há uma relação entre *redshift* cosmológico e distância comóvel, biúnivoca desde que o parâmetro de Hubble seja sempre positivo. A relação exata, porém, é dependente do modelo cosmológico utilizado, já que a evolução de $H(z)$ depende de parâmetros cosmológicos.

A partir da distância comóvel, várias outras medidas de distância podem ser obtidas. A distância física no momento t (ou no *redshift* z) entre dois pontos separados por uma distância comóvel χ é, por exemplo,

$$d = a(t)\chi = \frac{\chi}{1+z}. \tag{A.8}$$

Além do *redshift*, normalmente para medir distâncias no universo utilizam-se dois métodos: velas padrão ou réguas padrão. Velas padrão são fontes com luminosidade conhecida L. Sabendo que o fluxo F observado cai com o quadrado da distância, utilizando velas padrão obtemos uma distância até a fonte chamada distância de luminosidade, d_{lum}. Devido à expansão do

universo tanto a energia dos fótons quanto a taxa com que chegam ao observador diminuem proporcionalmente a $(1+z)$ (cf. equação (2.2) e definição do fator de escala), logo a distância de luminosidade é dada por

$$d_{lum} = (1+z)d_0 = (1+z)\chi, \tag{A.9}$$

sendo d_0 a distância física atual (que é igual à distância comóvel, já que adotamos $a_0 = 1$). Por outro lado, podemos determinar a distância até uma régua padrão de tamanho l notando que o ângulo θ subentendido é igual à razão l/d_{ang}, onde d_{ang} é a distância do tamanho angular. Como esta distância na verdade é a distância da fonte no instante em que luz foi emitida, a distância atual d_0 é $a = 1/(1+z)$ vezes maior,

$$d_{ang} = \frac{d_0}{1+z} = \frac{\chi}{1+z} \tag{A.10}$$

A.2 A Lei de Hubble

Em 1929, Edwin Hubble mediu a distância de algumas dezenas de galáxias com *redshift* conhecido. Já se sabia que a maior parte das galáxias possuia *redshifts*, e não blueshifts. As medidas de Hubble, entretanto, revelaram que além disso o *redshift* era proporcional à distância. Hubble interpretou este *redshift* como sendo causado pelo efeito Doppler, e concluiu que a velocidade v de afastamento das galáxias é proporcional à sua distância d,

$$v = Hd, \tag{A.11}$$

sendo H a constante de Hubble. As distâncias medidas por Hubble se mostraram superestimadas, e portanto o valor de H obtido por ele, de cerca de 500 km s^{-1} Mpc^{-1} é consideravelmente maior que o valor aceito hoje, de 68 ± 2 km s^{-1} Mpc^{-1}.

Este resultado é naturalmente entendido em um Universo em expansão. Sendo a distância de uma fonte função do fator de escala $d(t) = a(t)d_0$, é claro que a velocidade $v(t) = \dot{d}$ vale

$$v(t) = \dot{d}(t) = \frac{\dot{a}(t)}{a(t)} a(t) d_0 = H(t) d(t), \tag{A.12}$$

onde $H(t) = \frac{\dot{a}(t)}{a(t)}$ é o parâmetro de Hubble. Para fontes próximas, $H(t) \approx H_0$ e recuperamos a lei de Hubble.

Observe, porém, que a interpretação dada por Hubble possui sutilizesas[18]. Para objetos suficientemente distantes o *redshift* pode ser consideravelmente maior que 1 (a radiação cósmica de fundo, por exemplo, possui um *redshift* $z_{rec} = 1100$). A lei de Hubble, portanto, indica que a velocidade de afastamento nestes casos é superior à velocidade da luz. Neste regime é claro que a aproximação $H(t) \approx H_0$ falha, e o cálculo da velocidade de afastamento $v = \dot{a}\chi$ é na verdade dependente de parâmetros cosmológicos (pois é sensível à evolução de H com o tempo), entretanto ainda é verdade que observamos objetos cuja velocidade de afastamento é maior que c. Isto não é um problema, pois não há um referencial inercial onde esta velocidade é de fato observada.

A lei de Hubble pode também ser entendida interpretando o *redshift* observado como um *redshift* cosmológico. De fato, lembrando da equação (A.7) para a distância comóvel em termos do *redshift* cosmológico, e considerando que para objetos próximos $H(z) \approx H_0$ temos

$$d_0 = \chi \approx \frac{cz}{H_0},$$

e igualando z ao *redshift* Doppler não-relativístico obtemos imediatamente a lei de Hubble.

B

Inferência bayesiana

Em estatística, **inferência** é um processo onde, observando um conjunto de dados supostamente obtidos de uma distribuição de probabilidade subjacente, são inferidas propriedades desta distribuição. Exemplos cotidianos deste processo incluem a estimativa da média e desvio padrão de uma distribuição de alguma grandeza física a partir de um conjunto de medições, ou de parâmetros de algum modelo físico a partir de ajuste a dados experimentais.

Há duas abordagens distintas deste mesmo problema, conhecidas por **bayesiana** e **frequentista**, que, embora tipicamente sejam consistentes entre si, supõem interpretações diferentes para o conceito de probabilidade, além de utilizarem-se de ferramentas distintas.

O paradigma frequentista interpreta a probabilidade de ocorrência de um evento A num experimento descrito por um espaço amostral Ω como a frequência relativa deste evento no limite em que o número de realizações do experimento vai ao infinito. Isto é, se em n repetições do experimento observa-se A n_A vezes, temos

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n}. \tag{B.1}$$

Esta é a definição clássica de probabilidade, e associadas a ela encontram-se diversas ferramentas, como os ensembles tipicamente utilizados na mecânica estatística, ou os valores p e as

diversas estatísticas utilizadas em testes de hipóteses.

O paradigma bayesiano por sua vez interpreta probabilidades como uma representação do grau de confiança atribuído a uma proposição que pode ser verdadeira ou falsa. Estas probabilidades gozam de duas propriedades: se $\overline{X}$ é a negação da proposição X então

$$P(X) + P(\overline{X}) = 1 \tag{B.2}$$

$$P(X, Y) = P(X|Y)P(Y) \tag{B.3}$$

onde $P(X, Y)$ corresponde à probabilidade de X e Y serem verdadeiras e $P(X|Y)$ corresponde à probabilidade de X ser verdadeira caso se saiba que Y é verdadeira. Estas duas propriedades fazem com que seja possível atualizar a probabilidade que se atribui a uma proposição quando novas evidências são obtidas. Duas ferramentas são especialmente úteis. A primeira é o teorema de Bayes, que é uma consequência direta da regra do produto (eq. (B.3)),

$$P(H|D, I) = \frac{P(D|H, I)P(H|I)}{P(D|I)}. \tag{B.4}$$

Aqui I denota a informação à priori que já se possui, isto é, o conjunto de proposições que já se sabe (ou se supõe) verdadeiras, e costuma ser ignorado quando não há risco de confusão. No contexto do teorema de Bayes $P(D|H, I)$ costuma ser chamada verossimilhança (do inglês, *likelihood*)), enquanto $P(H|I)$ e $P(H|D, I)$ são nomeadas probabilidades à priori e à posteriori, respectivamente. O teorema de Bayes permite que se atualize a probabilidade de uma hipótese H dado um conjunto de observações D quando se conhecem as probabilidades das observações assumindo a hipótese como verdadeira (o que tipicamente é conhecido, já que normalmente a hipótese corresponde a algum modelo, que portanto fornece receitas para o cálculo da probabilidade de várias observações).

Adicionalmente, como consequência da regra da soma (eq. (B.2)) obtemos o processo de marginalização. Dadas proposições Y_i, onde $i \in \Lambda$ é um índice num conjunto discreto, dizemos que Y_i são mutuamente exclusivas quando a proposição Y_i e Y_j é falsa, ou seja, $P(Y_i, Y_j) = 0$,

para quaisquer $i \neq j$. Além disso, Y_i são exaustivas quando $\cup_{i\in\Lambda} Y_i$[1] é verdadeira, ou seja, $P\left(\cup_{i\in\Lambda} Y_i\right) = 1$. Neste caso, dado um conjunto de proposições mutuamente exclusivas e exaustivas temos

$$\sum_{i\in\Lambda} P(X, Y_i|I) = P(X|I). \tag{B.5}$$

Este processo também pode ser aplicado quando o conjunto Λ é contínuo, sendo neste caso as somas substituídas por integrais[2],

$$\int_{\Lambda} P(X, Y|I)dY = P(X). \tag{B.6}$$

A combinação da regra de marginalização com o teorema de Bayes pode ser usada para atualizar a probabilidade de hipóteses quando as verossimilhanças dependem de variáveis que não nos interessam (por exemplo, quando o modelo associado à hipótese depende de parâmetros), isto é, se N_i são um conjunto de parâmetros acessórios mutuamente excludente e exaustivo, o teorema de Bayes marginalizado em N_i fornece

$$P(H|D, I) = \sum_i P(H, N_i|D, I) = \sum_i \frac{P(D|H, N_i, I)P(H, N_i|I)}{P(D|I)}. \tag{B.7}$$

Por fim observamos que muitas vezes não se conhece a probabilidade a priori dos dados, $P(D|I)$. Isto entretanto não é um problema, já que ele age principalmente como um fator de normalização no teorema de Bayes, e podemos obter toda a informação sobre a forma da distribuição desconsiderando-o.

[1] A união corresponde ao ou lógico.

[2] Tecnicamente é necessário que Λ seja um espaço de medida, sendo a integral entendida no sentido de Lebesgue.