Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

UMA ABORDAGEM ESTATÍSTICA SOBRE A ESTIMAÇÃO DE *REDSHIFTS* DE QUASARES USANDO DADOS DO S-PLUS

Gabriela Pereira Soares

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Uma abordagem estatística sobre a estimação de *redshifts* de quasares usando dados do S-PLUS

Gabriela Pereira Soares
Orientador: Rafael Izbicki
Coorientadora: Lilianne Nakazono (IAG - USP)

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Setembro de 2022

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

A statistical approach to estimating quasar redshifts using S-PLUS data

Gabriela Pereira Soares
Advisor: Rafael Izbicki
Co-advisor: Lilianne Nakazono (IAG - USP)

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
September 2022

Gabriela Pereira Soares

Uma abordagem estatística sobre a estimação de *redshifts* de quasares usando dados do S-PLUS

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Gabriela Pereira Soares e aprovado pela banca examinadora.

Aprovado em 8 de setembro de 2022

Banca Examinadora:

- Rafael Izbicki
- Laerte Sodré
- Rafael Stern

Resumo

Redshift é um indicador cósmico utilizado para medir distâncias de objetos astronômicos. O estudo desta quantidade é importante para o entendimento da expansão do Universo e o afastamento das galáxias, conforme a cosmologia atual. Neste trabalho, temos interesse em estimar distâncias de quasares, que são corpos celestes luminosos conhecidos por apresentarem altos *redshifts*, indicando que estes estão a grandes distâncias da Terra.

A estimação do *redshift* pode ser realizada via espectroscopia, porém essa técnica possui um alto custo e tempo de observação cósmica. Assim, levantamentos fotométricos têm sido altamente valiosos neste campo, visto que também trazem informações relevantes para aferição do *redshift*, apesar de possuírem baixa resolução e menor precisão.

O objetivo deste trabalho é, a partir de dados fotométricos de quasares do S-PLUS (*Southern Photometric Local Universe Survey*), construir modelos estatísticos baseados na estimação de densidade condicional partindo do algoritmo FlexCoDE, a fim de melhorar a estimação de *redshifts* fotométricos. Além disso, há o interesse em estudar a influência de filtros de banda estreita (*narrow bands*) no modelo, disponíveis atualmente apenas no S-PLUS, e comparar com os resultados de um modelo de redes neurais previamente desenvolvido, com a finalidade de confirmar a significância dessas bandas.

Constatamos a partir das análises que, de fato, as *narrow bands* melhoram significativamente as estimativas da densidade condicional do *redshift* fotométrico, embora essa melhoria não seja observada em estimadores pontuais do *redshift*. Entretanto, o diagnóstico detectou que os modelos testados, tanto do FlexCoDE quanto das redes neurais, podem estar incorretamente especificados, se desviando da verdadeira densidade condicional.

Palavras-chave: *densidade condicional, estimação, filtro de banda estreita, FlexCoDE, quasar, redshift, S-PLUS.*

Abstract

Redshift is a cosmic index used to measure distances to astronomical objects. The study of this quantity is important for the understanding of the expansion of the Universe and the current objective of the stars, according to cosmology. In this work, we are interested in estimating distances of quasars, which are luminous celestial objects known by its high *redshifts*, indicating that they are at great distances from Earth.

The estimation of *redshift* can be performed via spectroscopy, but this technique has a high cost and requires a large amount of time for cosmic observation. Thus, photometric surveys have been highly valuable in this field, as they also provide relevant information for measuring *redshift*, despite having low resolution and less precision.

The goal of this work is to improve the estimation of photometric *redshifts* for quasars from S-PLUS (*Southern Photometric Local Universe Survey*). In order to do that, we build statistical models based on the estimation of conditional densities using the FlexCoDE algorithm. In addition, we study the influence of narrowband filters (*narrow bands*) on the model, currently available only in S-PLUS, and compare it with the results of a previously developed neural network model, with the purpose of confirming the significance of these bands.

We found from the analysis that *narrow bands* significantly improve the estimates of the conditional density of the photometric *redshift*, although this improvement is not observed in point estimators for the *redshift*. However, the diagnosis detected that the tested models, both from FlexCoDE and from neural networks, may be improved.

Keywords: *conditional density, estimation, FlexCoDE, narrow band, quasar, redshift, S-PLUS.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Desde o início da década de 1920, o astrônomo norte-americano Edwin Hubble observou que as galáxias tendem a se afastar de nós, com velocidades proporcionais à distância, o que viria a ser considerado como evidência de que o universo está em expansão (Hubble, 1942). Ele mostrou que a luz de galáxias distantes estava se desviando para comprimentos de onda mais longos, isto é, para a extremidade mais vermelha do espectro de cores. Este conceito é chamado de *redshift* (desvio para o vermelho), que está diretamente relacionado à velocidade com a qual esses objetos se afastam de nós. Ele corresponde ao aumento no comprimento de onda devido ao afastamento da fonte emissora em relação à fonte receptora, ou seja, à medida que um objeto se afasta de nós, a luz deste é deslocada para o vermelho.

A partir da década de 1960, astrônomos avistaram objetos brilhantes conhecidos como quasares, também chamados de QSO (*quasi-stellar objects*), os quais apresentavam *redshifts* maiores do que todas as galáxias já vistas anteriormente (Britannica, 2021). Estes corpos celestes são núcleos ativos de galáxias formados por buracos negros supermassivos, e são caracterizados por emitirem luz em todo espectro eletromagnético devido à acreção de matéria ao núcleo. Os quasares também representam um dos objetos mais luminosos e distantes do Universo. Portanto, obter uma grande amostra de quasares com uma medida precisa de suas distâncias é de extrema importância na comunidade astronômica, dada a possibilidade de se estudar as chamadas estruturas em larga escala. Além disso, os quasares mais distantes permitem o entendimento de como era o Universo bilhões de anos atrás (Kellermann, 2013).

O *redshift* pode ser estimado com maior precisão utilizando dados provenientes da espectroscopia, na qual a luz do objeto é decomposta em comprimento de onda a partir de

um espectrógrafo, obtendo-se a densidade espectral de energia (ou simplesmente espectro). Entretanto, trata-se de um método altamente custoso devido ao tempo de observação. Por outro lado, a fotometria, que é um processo de contagem de fótons via imageamento, é uma técnica de medição mais rápida. Neste método, filtros são inseridos no caminho da luz, onde cada filtro permite a passagem de fótons em uma determinada banda de comprimento de onda, e através de um dispositivo de carga acoplada (CCD) é possível aferir a quantidade de luz observada. Quando a fotometria é realizada em várias bandas, obtém-se o que é chamado de espectro de baixa resolução. Para mais informações sobre espectroscopia e fotometria, consultar Oliveira Filho e Saraiva (2004).

Grandes mapeamentos fotométricos são altamente valiosos para o campo de pesquisa astronômica. Neste trabalho, será abordado o conjunto de dados fotométricos observados pelo S-PLUS (*Southern Photometric Local Universe Survey*), que irá cobrir cerca de 9300 graus^2 no céu do hemisfério sul (Mendes de Oliveira *et al.*, 2019). As observações são feitas por meio de 12 bandas fotométricas, sendo 5 bandas largas e 7 bandas estreitas. Embora existam diversos levantamentos de imagem em grandes áreas, o S-PLUS é a única pesquisa no hemisfério sul que utiliza em seu espectro óptico 7 filtros de banda estreita. Deste modo, busca-se utilizar métodos estatísticos e de aprendizado de máquinas para estimar *redshifts* fotométricos de quasares com maior precisão em decorrência das *narrow bands* do S-PLUS.

Na Figura 1.1, podemos observar os espectros de um quasar, uma galáxia, uma estrela M6 e uma anã branca, de cima para baixo. Os fluxos observados pelos filtros de banda larga e estreita do S-PLUS são denotados pelos quadrados e círculos, respectivamente. Logo, fica evidente a forma como perdemos resolução com dados fotométricos em comparação aos dados espectroscópicos.

Considerando a importância de determinar a distância de objetos astronômicos para o entendimento do Universo, e levando em conta os desafios provocados pela fotometria, o presente trabalho tem como objetivo investigar modelos estatísticos que determinem o *redshift* fotométrico de quasares com maior precisão. Para isto, serão analisados os dados do levantamento fotométrico S-PLUS, verificando quais aspectos contribuem mais para a acurácia da estimação, levando em consideração as particularidades de seu sistema óptico.

Este trabalho está organizado da seguinte maneira: no Capítulo 2 apresentamos a pesquisa astronômica S-PLUS, seu sistema fotométrico, a estrutura e análise descritiva do conjunto de dados; no Capítulo 3 especificamos a metodologia a ser aplicada nos dados;

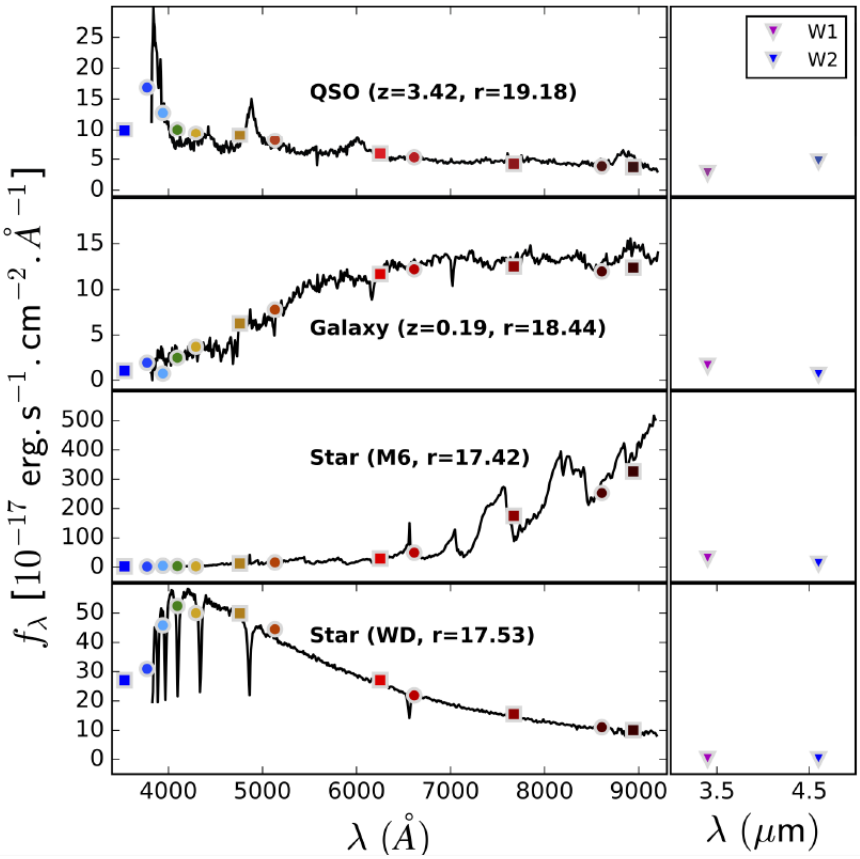

Figura 1.1: Gráfico do espectro (densidade de energia f_λ por comprimento de onda λ) observado pelo SDSS (curva preta) e o espectro fotométrico de baixa resolução do S-PLUS (quadrados e círculos coloridos) para um quasar, uma galáxia, uma estrela tipo espectral M6 e uma anã branca, de cima para baixo. Nos painéis da direita, encontram-se as medidas de densidade de energia em infravermelho obtidas pelo WISE (Nakazono *et al.*, 2021).

no Capítulo 4 descrevemos as métricas utilizadas para avaliar a performance dos modelos; no Capítulo 5 exibimos os resultados encontrados para os modelos em estudo; e por fim no Capítulo 6 consolidamos algumas conclusões e considerações finais sobre esta monografia.

Capítulo 2

A pesquisa astronômica S-PLUS

2.1 Apresentação

O *Southern Photometric Local Universe Survey*[1], também chamado de S-PLUS, é um levantamento fotométrico fundado com a colaboração de diversas instituições do Brasil, Chile e Espanha. Esta pesquisa observacional mapeará cerca de 9300 graus^2 no céu do hemisfério sul através do seu sistema óptico composto por 5 filtros de banda larga e 7 filtros de banda estreita, que serão descritos com mais detalhes na Seção 2.2. O telescópio utilizado está localizado no Observatório Interamericano Cerro Tololo (CTIO), no Chile, e seu espelho primário possui cerca de 0.8 metros de abertura.

O levantamento está dividido em 5 subcampos, de acordo com os interesses científicos da comunidade. A principal pesquisa é a *Main Survey* (MS), que cobre uma área de aproximadamente 8000 graus^2, e algumas de suas motivações estão na formação de uma base com indicadores ambientais, detecção de galáxias, classificação de objetos e estudo de populações estelares. Outra pesquisa é a *Ultra-Short Survey* (USS), que cobre a mesma área da MS, porém designada a explorar estrelas brilhantes de baixa metalicidade. A *Galactic Survey* (GS) abrange cerca de 1420 graus^2, cuja técnica de exposição permite a detecção de diferentes populações estelares. A pesquisa *Marble Field Survey* (MFS) possui objetos específicos de estudo, como a Pequena e Grande Nuvem de Magalhães e o aglomerado de Hydra. Por fim, a *Variability Fields Survey* (VFS) é um levantamento qualificado para detectar fontes variáveis. Para mais detalhes, consulte Mendes de Oliveira *et al.* (2019).

O catálogo de dados fotométricos que está sendo construído pelo S-PLUS é extrema-

[1] http://www.splus.iag.usp.br/

mente importante para o campo da astrofísica e astronomia, devido ao grande volume de informações obtidas em tempo razoável de observação cósmica, quando comparado a um estudo espectroscópico. Além disso, o *Southern Photometric Local Universe Survey* se destaca por ser o único levantamento com quantidade considerável de dados já observados em 7 bandas estreitas no hemisfério sul. Outras pesquisas como J-PLUS[2] (*Javalambre Photometric Local Universe Survey*) e J-PAS[3] (*Javalambre Physics of the Accelerating Universe Astrophysical Survey*) também fazem uso de filtros de banda estreita. No caso do J-PAS, a meta é cobrir ao menos 8500 graus^2 do céu utilizando 54 filtros de banda estreita. Já o J-PLUS tem como foco o estudo observacional do hemisfério norte, fazendo uso de 5 filtros de banda larga e 7 filtros de banda estreita, assim como o S-PLUS.

2.2 Sistema Fotométrico

Para entendermos o funcionamento de um sistema fotométrico, precisamos primeiramente definir alguns conceitos básicos de astrofísica. A luz é uma onda eletromagnética composta por partículas chamadas fótons, em que a distância com a qual a forma da onda se repete é denominada como comprimento de onda. Este, por sua vez, é denotado por λ e geralmente é medido em unidades de angstrom (Å), que equivalem a 0.1 nanômetros. O número de oscilações em que essa onda se repete a cada segundo é chamado de frequência, que é inversamente proporcional ao comprimento de onda.

Nos estudos de imageamento, usamos filtros de banda para restringir o intervalo de comprimento de onda permitido na passagem de luz. Neste processo, determinamos a intensidade do brilho de um objeto em uma banda específica a partir da magnitude, que é uma escala logarítmica do brilho e indiretamente proporcional a esta (Carroll e Ostlie, 2017).

Na astronomia, o conceito de cores é definido como a diferença de magnitude entre dois filtros. Ela é obtida subtraindo-se o valor da banda com comprimento de onda mais longo pelo valor da banda com comprimento de onda mais curto. Na Figura 2.1, podemos observar a escala de cores de acordo com suas respectivas frequências de onda. A construção das cores e suas relações com o *redshift* serão estudadas com maior aprofundamento nas Seções 2.3 e 5.

Nos estudos observacionais de astronomia, cada telescópio tem o seu próprio sistema de

[2] http://www.j-plus.es/
[3] http://www.j-pas.org/

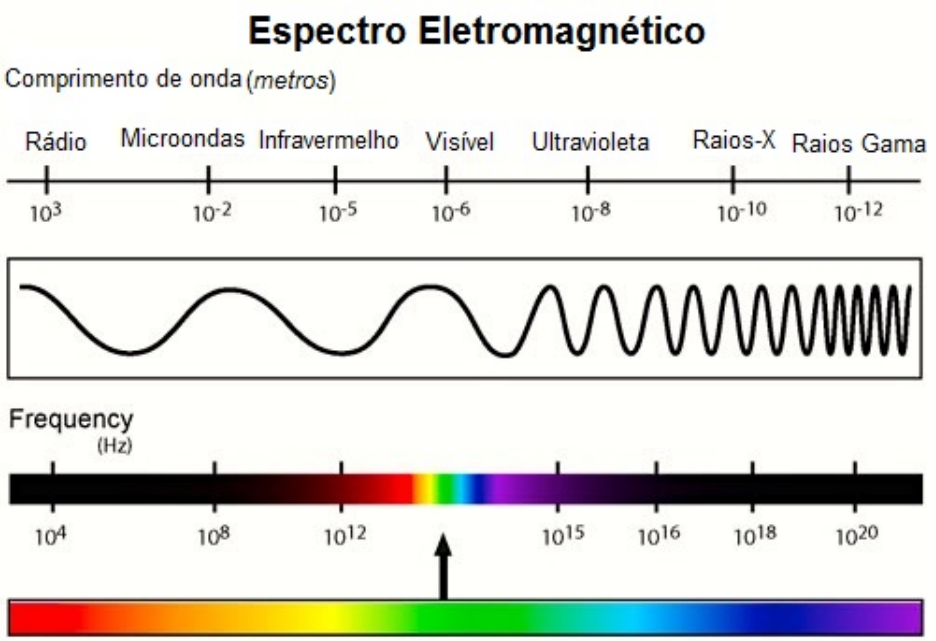

Figura 2.1: Cores e comprimentos de onda no espectro visível e não visível (Garabini, 2017).

filtros. Por exemplo, o telescópio do *Wide-field Infrared Survey Explorer*[4] (WISE) possui bandas de passagem no infravermelho, que é uma radiação não vista ao olho humano. Já o satélite *Galaxy Evolution Explorer*[5] (GALEX) utiliza filtros no espectro ultravioleta. Abordaremos com mais detalhes estas magnitudes na Seção 2.3.

O sistema óptico de filtros do S-PLUS é conhecido como *Javalambre System*, composto por 5 filtros de banda larga (*ugriz*), semelhantes aos do levantamento espectroscópico SDSS (*Sloan Digital Sky Survey*), e 7 filtros de banda estreita (prefixo "J0"), que capturam uma fração muito menor do espectro de luz. Essas 12 bandas definem o espectro fotométrico de baixa resolução, que mede a magnitude a partir do fluxo de fótons que passa por cada banda.

Na Figura 2.2, temos a curva de transmissão do sistema de filtros do S-PLUS, que representa a razão entre a quantidade de luz que atravessa a banda sobre a quantidade incidente (IUPAC, 2019), em que o eixo x corresponde ao comprimento de onda λ em unidades de angstrom (Å). As bandas estreitas estão centradas em *spectral features* conhecidas, que revelam propriedades químicas e físicas estelares, representadas pelas faixas verticais. Como quasares estão sob efeito da expansão do Universo e, portanto, têm velocidade de afastamento, o espectro como um todo se desloca para comprimentos de onda mais vermelhos.

[4] https://www.nasa.gov/mission_pages/WISE/mission/index.html
[5] http://www.galex.caltech.edu/about/overview.html

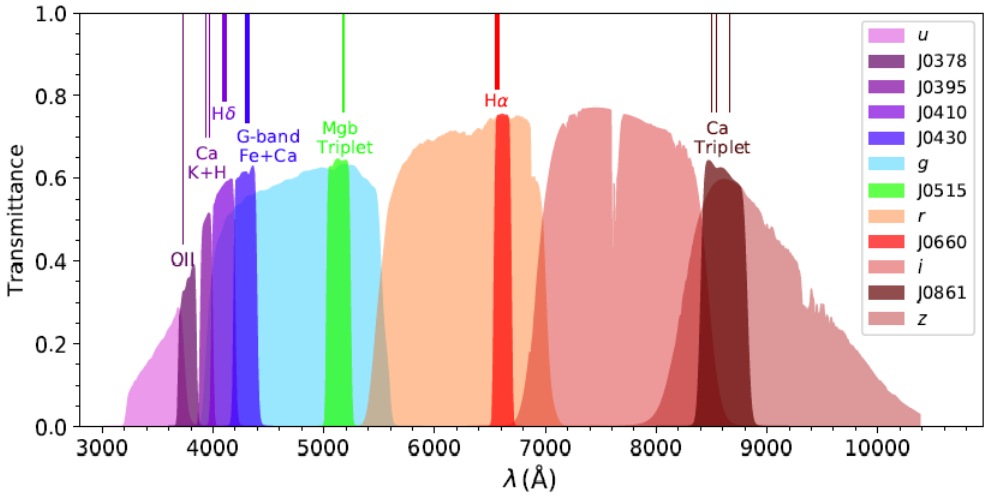

Figura 2.2: Curva de transmissão do sistema de bandas S-PLUS (Mendes de Oliveira *et al.*, 2019).

2.3 Conjunto de Dados

As observações do conjunto de dados que trabalharemos provêm do terceiro *data release* do S-PLUS (DR3), cobrindo uma área de 336 graus^2 de uma região conhecida como *Stripe 82* (Jiang *et al.*, 2014).

A base de dados disponibilizada para este estudo abrange no total cerca de 32 mil observações e 560 colunas referentes a informações de identificação do quasar, morfologia, fotometria, calibração, classificação e qualidade fotométrica de cada banda. As variáveis de morfologia dizem respeito aos parâmetros de detecção de imagem do *SExtractor* (Bertin e Arnouts, 1996), um programa de segmentação e catalogação de objetos astronômicos, e descrevem medidas de posição, área, forma e classificação a partir dos *pixels* da imagem.

As variáveis fotométricas mensuram a magnitude do objeto nas diferentes combinações entre as 12 bandas S-PLUS e os 6 tipos de aberturas, assim como seus respectivos erros de medida e razão sinal-ruído. Tais aberturas são:

- `AUTO`: abertura baseada no raio elíptico de Kron, possui baixa relação sinal-ruído;
- `PETRO`: abertura baseada no raio elíptico Petrosiano, deriva propriedades físicas;
- `APER_3`: abertura circular fixa com 3"de diâmetro;
- `APER_6`: abertura circular fixa com 6"de diâmetro;
- `ISO`: abertura isofotal, é a que melhor preserva a forma do objeto;
- `PStotal`: abertura circular fixa com correção de abertura para fontes pontuais.

As variáveis relativas às magnitudes do S-PLUS serão denotadas da forma {`banda`}_{`abertura`}.

Por exemplo, a variável `r_PStotal` diz respeito à magnitude medida na banda r com a abertura `PStotal`. Nesta monografia faremos as análises essencialmente com as variáveis do S-PLUS que foram medidas na abertura `PStotal`.

A amostra de quasares foi confirmada via espectroscopia a partir do conjunto de dados do SDSS realizando-se um *cross-matching*, isto é, um cruzamento entre as tabelas de quasares do S-PLUS e do SDSS considerando a posição desses objetos no céu. Segundo Lyke *et al.* (2020), a taxa de classificações erradas de quasares do SDSS é em torno de 0.3% a 1.3%.

A determinação dos *redshifts* no conjunto de dados também foi realizada via espectroscopia a partir do levantamento de dados do SDSS. O intuito deste trabalho é fixar o *redshift* espectroscópico z_{spec} como referencial para estimar o *redshift* fotométrico z_{phot} a partir de preditores fotométricos $\mathbf{x}$ com maior precisão, dando ênfase às *narrow bands*.

Além do *cross-matching* realizado para determinar o *redshift* espectroscópico, o catálogo do S-PLUS também inclui no conjunto de dados 2 filtros fotométricos no infravermelho do *Wide-field Infrared Survey Explorer* (WISE), chamados `W1_mag` e `W2_mag`, medidos no sistema Vega, e 2 filtros no ultravioleta do *Galaxy Evolution Explorer* (GALEX), chamados `FUVmag` e `NUVmag`, medidos no sistema AB assim como as bandas do S-PLUS. Essas magnitudes serão importantes na modelagem pois pesquisas realizadas por DiPompeo *et al.* (2015) e Yang *et al.* (2017) apontam que a inclusão destas melhoram a precisão do *redshift*.

A fim de gerar resultados comparáveis, os dados utilizados para análise descritiva e modelagem correspondem à amostra utilizada em Nakazono & Ruiz (em preparação), que contém 32687 registros de quasares. Neste artigo, os dados foram filtrados de maneira que teremos a variável `r_PStotal`≤ 22 e o *redshift* espectroscópico $0 < z_{spec} < 7$, possibilitando informações com boa qualidade fotométrica. A partir das magnitudes medidas nos filtros fotométricos, disponíveis no banco de dados, foram construídas as variáveis de cores tomando-se a banda `r_PStotal` como referencial para a subtração. Isto é, para cada filtro do S-PLUS, WISE e GALEX, fazemos a diferença entre a magnitude medida na banda com menor comprimento de onda e a magnitude na banda com maior comprimento de onda, nesta ordem, sendo a `r_PStotal` uma dessas bandas. A Tabela 2.1 ilustra a construção dessas variáveis. Vale ressaltar que utilizamos apenas as bandas do S-PLUS medidas na abertura `PStotal` e que as cores foram criadas somente após a tratativa dos valores faltantes nos filtros de banda, que será melhor detalhada na Seção 2.3.1.

Tabela 2.1: Cores construídas a partir dos filtros tomando-se como base seus respectivos comprimentos de onda. Adaptação de Nakazono *et al.* (2021).

Filtros	**Comprimento de onda efetivo (Å)**	**Cores Criadas**
FUV	1528	FUV - r
NUV	2310	NUV - r
u	3536	u - r
J0378	3770	J0378 - r
J0395	3940	J0395 - r
J0410	4094	J0410 - r
J0430	4292	J0430 - r
g	4751	g - r
J0515	5133	J0515 - r
r	6258	
J0660	6614	r - J0660
i	7690	r - i
J0861	8611	r - J0861
z	8831	r - z
W1	46000	r - W1
W2	34000	r - W2

2.3.1 Análise Descritiva

Nesta seção, iremos realizar algumas análises preliminares com o objetivo de compreendermos o comportamento das variáveis e sumarizar informações acerca dos dados. Como o foco principal do trabalho são as bandas fotométricas, analisaremos apenas essas bandas e outros atributos associados à elas.

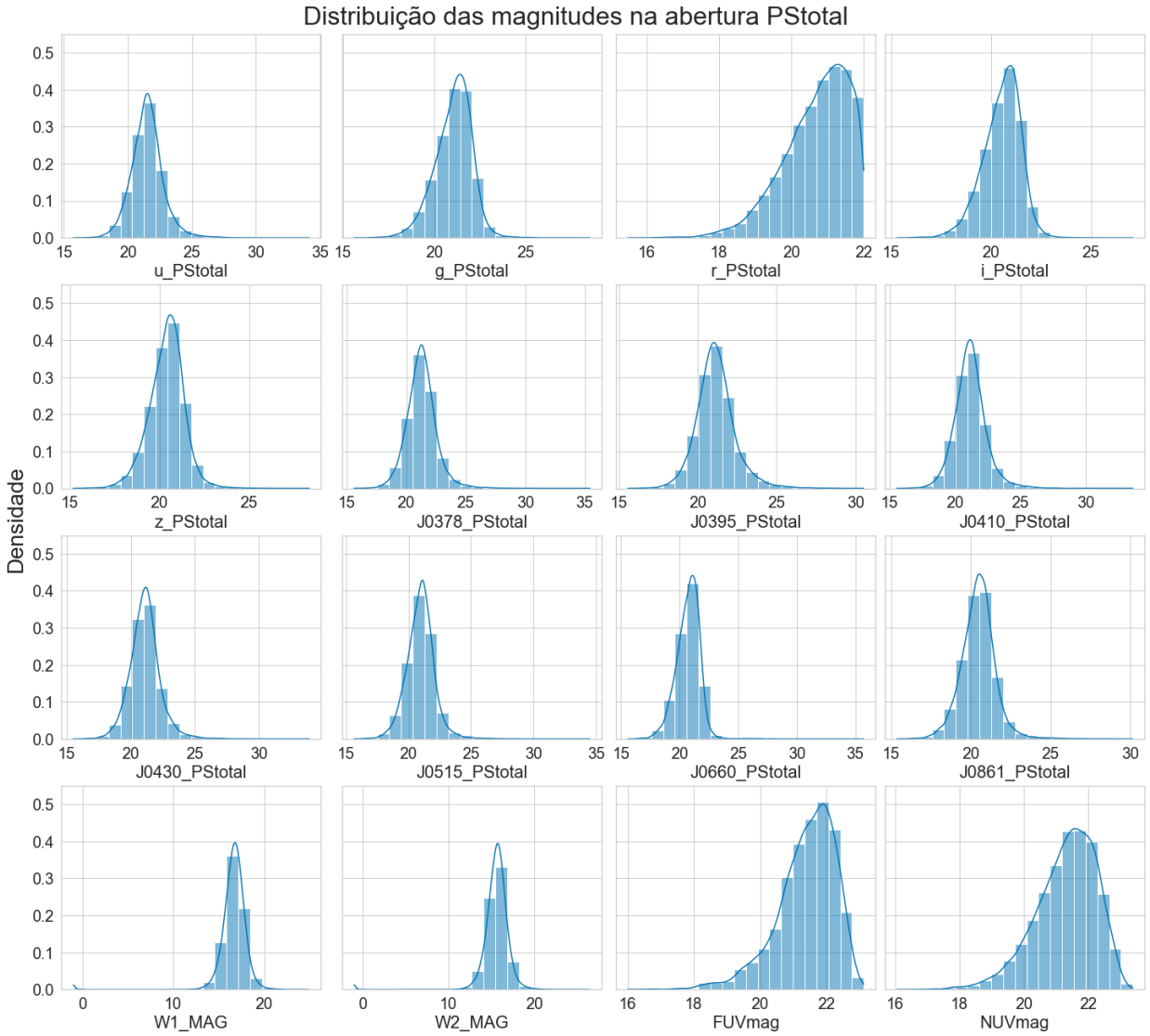

Figura 2.3: Histogramas da distribuição das magnitudes do S-PLUS, WISE e GALEX com as densidades de *kernel*.

Na Figura 2.3, podemos observar o histograma dos dados em cada uma das magnitudes do S-PLUS, WISE e GALEX. Nota-se que a maioria delas apresenta distribuições quase que simétricas, com exceção da `FUVmag`, `NUVmag` e `r_PStotal`. Vale relembrar também que foi realizada uma filtragem nos dados na qual foram removidos quasares com `r_PStotal`$>$ 22, justificando seu formato assimétrico à esquerda.

As variáveis de erro denotam a incerteza de medida das magnitudes nos diferentes

filtros. De maneira geral, quando a magnitude se encontra até aproximadamente o valor 20 temos um erro próximo de 0 nos filtros do S-PLUS, isto é, temos uma confiabilidade maior do valor calculado para a magnitude. Já acima desse valor, o erro tende a crescer quase que exponencialmente, como exemplificado na Figura 2.4 para uma das bandas.

Figura 2.4: Diagrama de dispersão entre a magnitude J0861 na abertura PStotal e seu respectivo erro de medida.

Tabela 2.2: Medidas de posição e dispersão dos erros relacionados às magnitudes

height	**Média**	**D.P.**	**Min.**	**25%**	**50%**	**75%**	**90%**	**Max.**
e_u_PStotal	2.98	152.52	0.00	0.13	0.27	0.55	1.70	26951.11
e_g_PStotal	0.23	2.71	0.00	0.06	0.12	0.20	0.31	449.14
e_r_PStotal	0.13	0.69	0.00	0.05	0.10	0.16	0.22	23.04
e_i_PStotal	0.16	0.60	0.00	0.06	0.11	0.18	0.28	45.36
e_z_PStotal	0.38	2.61	0.00	0.10	0.18	0.30	0.49	331.66
e_J0378_PStotal	5.63	541.02	0.00	0.17	0.33	0.69	2.36	97424.15
e_J0395_PStotal	3.28	16.97	0.01	0.25	0.47	1.08	15.21	1841.39
e_J0410_PStotal	5.52	279.56	0.01	0.21	0.40	0.83	3.37	39212.93
e_J0430_PStotal	4.18	228.79	0.00	0.18	0.35	0.68	2.03	39913.90
e_J0515_PStotal	3.50	372.87	0.00	0.14	0.26	0.46	0.91	67174.87
e_J0660_PStotal	2.59	406.63	0.00	0.06	0.11	0.18	0.28	73495.52
e_J0861_PStotal	0.92	11.14	0.00	0.13	0.23	0.40	0.73	1587.34
W1_MAG_ERR	0.16	1.73	-1.00	0.04	0.08	0.15	0.27	171.72
W2_MAG_ERR	0.33	14.29	-1.00	0.06	0.12	0.21	0.39	2521.75
e_FUVmag	0.31	0.11	0.02	0.23	0.32	0.40	0.46	0.54
e_NUVmag	0.26	0.12	0.01	0.17	0.26	0.36	0.43	0.50

Na Tabela 2.2, podemos notar uma assimetria extremamente grande nas variáveis relacionadas aos erros de magnitude, com exceção do `e_FUVmag` e `e_NUVmag`. Isto é notável

tanto pelo alto desvio padrão quanto pela discrepância do 3º quartil para o valor máximo, principalmente nos erros das bandas estreitas. Entretanto, vemos que no percentil 0.9 a maioria dos valores ainda se encontram relativamente baixos, indicando que pelo menos 90% dos dados possuem uma precisão e uma qualidade aceitável.

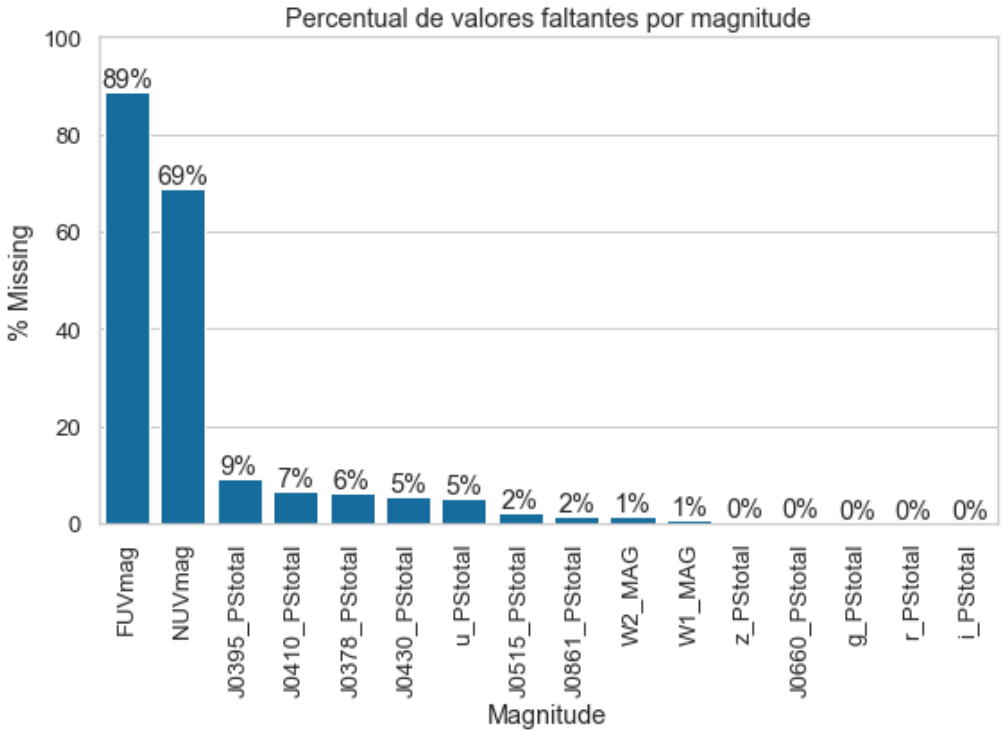

Figura 2.5: Percentual de valores faltantes nas magnitudes do S-PLUS, WISE e GALEX.

A Figura 2.5 exibe o percentual de valores *missing* nos filtros do S-PLUS, WISE e GALEX. Como podemos ver, as magnitudes `FUVmag`, `NUVmag` apresentam mais de 80% e 60% dos dados com valores faltantes, respectivamente. Dados *missing* nesses filtros podem indicar que os quasares não foram identificados no catálogo do GALEX. Isso pode ser explicado pelo fato de que a radiação correspondente ao comprimento de onda $\lambda < 912$Å é absorvida pelo gás neutro do ambiente e, portanto, objetos com alto *redshift* não seriam identificáveis nos filtros de menor comprimento de onda (Steidel *et al.*, 1999). Além dessas colunas, vemos também dados faltantes em algumas bandas do S-PLUS e também do WISE, porém com um percentual bem menor, que são os casos quando o fluxo está abaixo do limite de detecção da imagem.

Os dados faltantes serão substituídos pelo número 99, como proposto por Almeida-Fernandes *et al.* (2022). O valor 99 simboliza a alta magnitude não detectada pelo dispositivo e, como nosso modelo será baseado em árvores, as divisões nas covariáveis serão do tipo $x_j > k$, onde x_j corresponde à covariável j e k o ponto de corte, e não será afetada

por valores extremos. Portanto, o mais relevante para os dados *missing* não é a quantia imputada em si, e sim representá-los por um valor acima da maior magnitude observada.

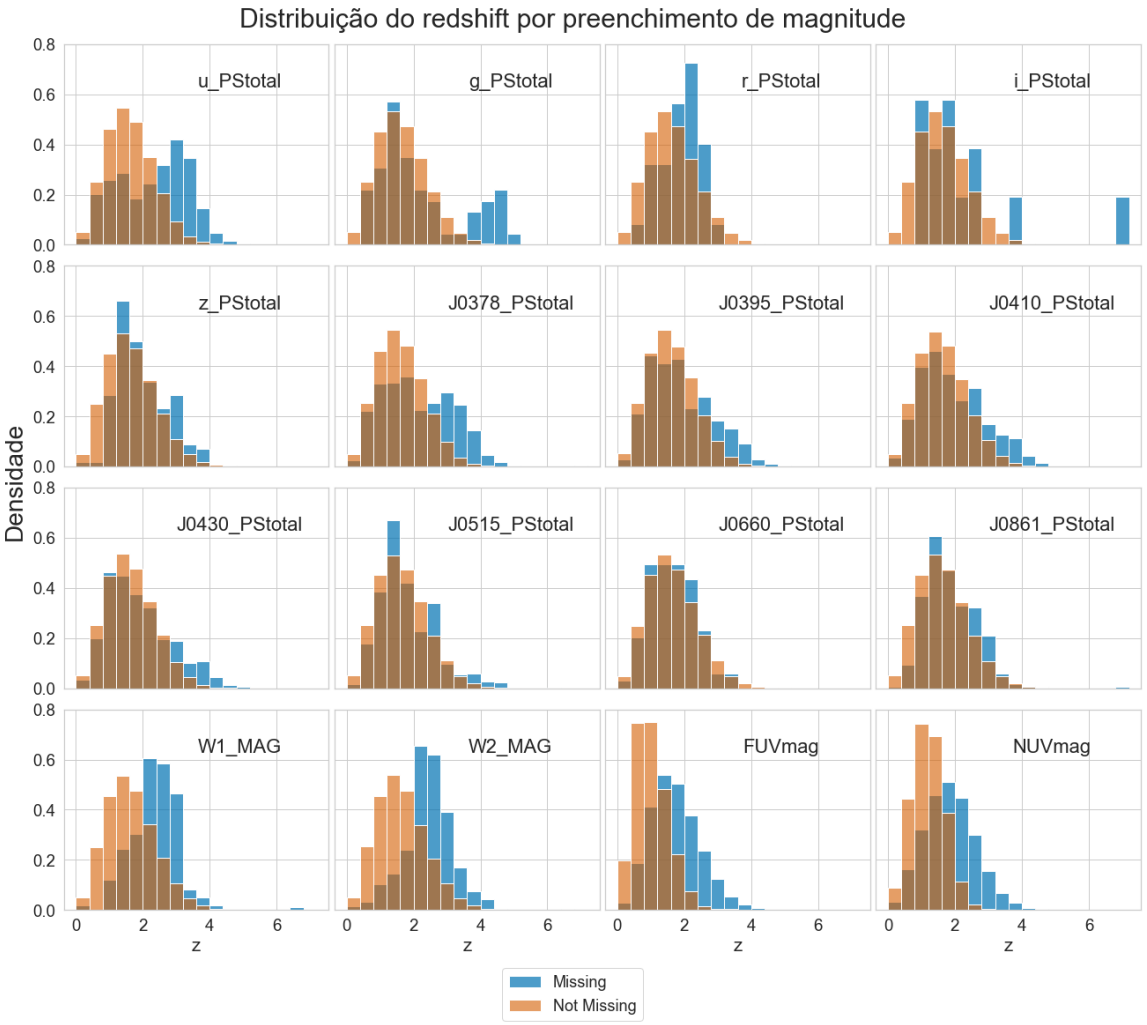

Figura 2.6: Histogramas da distribuição do *redshift* espectroscópico com e sem valores faltantes nas magnitudes do S-PLUS, WISE e GALEX.

Vemos na Figura 2.6 a distribuição do *redshift* espectroscópico considerando dados faltantes e completos em cada uma das bandas. Por exemplo, no primeiro gráfico, temos em laranja o histograma do *redshift* considerando apenas os dados completos na banda `u_PStotal`, e em azul apenas os dados faltantes nessa magnitude. Nota-se que, na maioria dos filtros, a distribuição do *redshift* tende a se deslocar para a direita quando consideramos apenas os valores *missing*, dando indícios de que existe relação direta entre a não-detecção nessas bandas com o desvio para o vermelho. Ou seja, a falta de informação nessas colunas também é uma informação para o *redshift*.

Como mencionado anteriormente, foram construídas variáveis de cores a partir das magnitudes do WISE, GALEX e S-PLUS na abertura `PStotal`, que servirão como pilares

dos modelos estatísticos mais adiante, em que os valores *missing* das bandas foram substituídos por 99 antes da criação das cores. A Figura 2.7 mostra o gráfico de correlação entre as cores e o *redshift* espectroscópico z_{spec}. Nesta imagem podemos concluir que no plano bidimensional não há nenhuma variável com forte associação ao *redshift* ou alta correlação entre as cores, exceto pelos poucos quadrados laranjas ou azuis um pouco mais intensos, como por exemplo entre `r-i` e `g-r`, e entre `NUV-r` e `FUV-r`, que apresentam correlações consideráveis.

Vale ressaltar que a medida de importância dessas variáveis na estimação de z_{phot} pode ser afetada pela multicolinearidade de algumas variáveis em métodos baseados em árvores, pois quando duas ou mais *features* correlacionadas aparecem no mesmo modelo, a importância de uma reduz o efeito da outra.

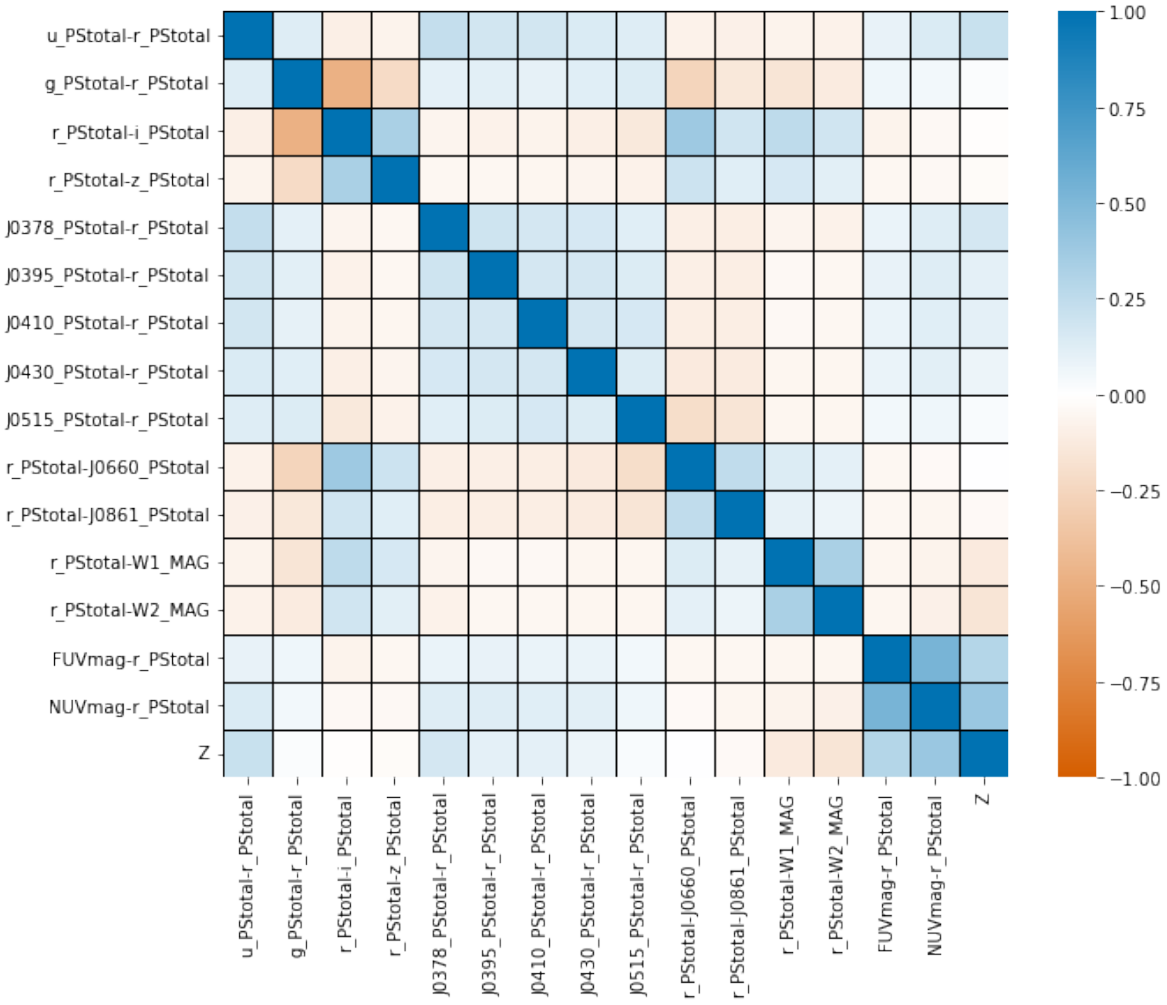

Figura 2.7: Mapa de calor da correlação entre as cores e o *redshift* espectroscópico.

Capítulo 3

Métodos de estimação

3.1 Estimação da esperança condicional

A estimação do *redshift* fotométrico tem sido tratada como um problema de regressão via modelos de *machine learning* na maior parte da literatura, principalmente devido às melhorias nas tecnologias de aprendizado e a disponibilidade de grandes levantamentos astronômicos. Tal abordagem tem sido amplamente utilizada em diversas pesquisas astronômicas, inclusive para os dados do S-PLUS, como em Lima (2020) e Lima *et al.* (2022), que utilizam particularmente redes neurais artificiais, processos gaussianos e inferência bayesiana.

O trabalho de Nakazono & Ruiz (em preparação) também trouxe resultados preliminares no que diz respeito à estimação pontual fotométrica do *redshift* no conjunto de dados do S-PLUS que utilizaremos neste trabalho. O modelo com florestas aleatórias apresentou boa performance, principalmente com os dados faltantes. Entretanto, não houve evidências de que os filtros de banda estreita do S-PLUS melhoram a precisão da estimativa do *redshift* fotométrico.

Os algoritmos estatísticos, de forma geral, buscam gerar estimativas pontuais da esperança condicional do *redshift* dadas as variáveis fotométricas, $\mathbb{E}[Z|\mathbf{x}]$. A esperança de uma variável aleatória contínua Y é definida como o valor médio ou esperado de Y, dada pela expressão

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} f_Y(y)dy, \tag{3.1}$$

em que $f_Y(y)$ é a função de densidade de probabilidade de Y, que será mais detalhada

em (3.4).

Assim, considerando duas variáveis aleatórias Y e U, a esperança condicional de Y dado $U = u$ é definida como o valor médio esperado da variável Y considerando o valor observado de $U = u$, dada pela equação a seguir:

$$\mathbb{E}[Y|U = u] = \int_{-\infty}^{\infty} y f_{Y|U}(y, u)dy, \tag{3.2}$$

na qual $f_{Y|U}(y, u)$ é a função de densidade de probabilidade condicional, que será melhor detalhada em (3.6).

Neste trabalho, denotaremos por $\mathbb{E}[Z|\mathbf{x}] := \mathbb{E}[Z|\mathbf{X} = \mathbf{x}]$ como sendo a esperança condicional do *redshift* fotométrico z_{phot} dado o vetor de covariáveis observadas $\mathbf{X} = (x_1, x_2, \ldots, x_p)$.

Além dos métodos tradicionais de aprendizado de máquinas (Izbicki e dos Santos, 2020), podemos estimar também a esperança condicional $\mathbb{E}[Z|\mathbf{x}]$ a partir da função de densidade condicional estimada $\hat{f}(z|\mathbf{x})$ por meio da expressão

$$\hat{\mathbb{E}}[Z|\mathbf{x}] = \int_{-\infty}^{\infty} z\hat{f}(z|\mathbf{x})dz. \tag{3.3}$$

3.2 Estimação da densidade condicional

Apesar de métodos de regressão baseados na estimação da esperança condicional do *redshift*, $\mathbb{E}[Z|\mathbf{x}]$, serem popularmente mais utilizados, Ball e Brunner (2010) mostraram em estudos recentes que estimar a função de densidade de probabilidade condicional (FDC) do *redshift*, $f(z|\mathbf{x})$, melhora significativamente os resultados; veja também Izbicki e Lee (2016); Izbicki *et al.* (2017); Freeman *et al.* (2017); Dalmasso *et al.* (2020); Izbicki *et al.* (2022). Isso porque a função de densidade traz mais informações do que uma simples média, principalmente no contexto da fotometria, na qual a distribuição de z geralmente é assimétrica, multimodal e com erros heterocedásticos.

A função de densidade de probabilidade (FDP) de uma variável aleatória contínua Y é uma função não-negativa $f : \mathcal{Y} \longrightarrow [0, \infty]$ tal que, para um intervalo (a, b) no domínio $\mathcal{Y}$, a probabilidade de Y pertencer ao intervalo é dada por

$$P(a < Y < b) = \int_a^b f_Y(y)dx. \tag{3.4}$$

Considere duas variáveis aleatórias Y e U possuindo uma função densidade de probabilidade conjunta $f_{YU}(y,u)$ que satisfaz, para uma região $A \subset \mathbb{R}^2$, a equação

$$P([Y,U] \in A) = \iint_A f_{YU}(y,u)dydu. \tag{3.5}$$

Então, definimos a função densidade de probabilidade condicional (FDC) de Y dado que $U = u$ como sendo

$$f_{Y|U}(y,u) = \frac{f_{YU}(y,u)}{f_U(u)}, \tag{3.6}$$

restrito a $f_U(u) > 0$.

Neste trabalho, denotaremos por $f(z|\mathbf{x}) := f_{Z|\mathbf{X}}(z,\mathbf{x})$ a função densidade de probabilidade condicional do *redshift* fotométrico z_{phot} dado o vetor de covariáveis observadas $\mathbf{X} = (x_1, x_2, \ldots, x_p)$.

3.2.1 O algoritmo FlexCoDE

Existem diversas metodologias designadas a estimar a densidade condicional $f(z|\mathbf{x})$ na literatura. Por exemplo, Rosenblatt (1969) propôs utilizar estimadores de densidade de *kernel* a fim de predizer a distribuição conjunta $f(z,\mathbf{x})$ e a distribuição marginal $f(\mathbf{x})$ para, então, a partir da Equação (3.6), estimar $f(z|\mathbf{x})$. Outras técnicas como regressão polinomial local (Fan *et al.*, 1996) e regressão quantílica (Takeuchi *et al.*, 2009) também trazem soluções alternativas para determinar a densidade condicional de uma variável. Entretanto, tais abordagens não produzem boa performance para conjuntos de dados em alta dimensão, devido ao custo computacional para escolha de parâmetros de ajuste.

FlexCoDE (*Flexible nonparametric conditional density estimation*) é um método não-paramétrico elaborado por Izbicki e Lee (2017) que visa a conversão de estimadores de esperança condicional em estimadores de densidade condicional. Sua implementação é eficiente com banco de dados de alta dimensão devido à flexibilidade na escolha de modelos de regressão para estimação dos coeficientes.

Formalmente, considere n observações independentes e identicamente distribuídas $(\mathbf{x}_1, z_1), (\mathbf{x}_2, z_2), \ldots, (\mathbf{x}_n, z_n)$ com $\mathbf{x} \in \mathbb{R}^p$. Seja também uma base ortonormal definida $(\phi_i)_{i\in\mathbb{N}}$. Então, para o método FlexCoDE, a função de densidade de probabilidade condi-

cional em $\mathbf{X} = \mathbf{x}$ pode ser escrita como

$$f(z|\mathbf{x}) = \sum_{i \in \mathbb{N}} \beta_i(\mathbf{x}) \phi_i(z), \tag{3.7}$$

em que $\beta_i(\mathbf{x})$ são chamados de coeficientes de expansão do modelo. Tais coeficientes são dados por

$$\beta_i(\mathbf{x}) = \langle f(\cdot|\mathbf{x}), \phi_i \rangle = \int_{\mathbb{R}} \phi_i(z) f(z|\mathbf{x}) dz = \mathbb{E}[\phi_i(Z)|\mathbf{x}]. \tag{3.8}$$

Desta forma, podemos estimar os coeficientes de expansão do modelo via métodos de regressão a partir dos dados observados da amostra. O algoritmo FlexCoDE permite uma flexibilidade na escolha do método de regressão para estimar os coeficientes de acordo com a estrutura específica dos dados. Assim, o estimador será dado por:

$$\hat{f}(z|\mathbf{x}) = \sum_{i=1}^{I} \hat{\beta}_i(\mathbf{x}) \phi_i(z), \tag{3.9}$$

sendo I um parâmetro de ajuste que determina a quantidade de coeficientes de expansão no modelo. Na prática, este *tuning* parâmetro é escolhido a partir do conjunto de validação, cujo valor aumenta à medida que a densidade apresenta mais protuberâncias.

Para selecionar o melhor número de coeficientes do modelo, inicialmente definimos a função de perda (*loss function*) como

$$\begin{aligned} L(\hat{f}, f) &= \int \int (\hat{f}(z|\mathbf{x}) - f(z|\mathbf{x}))^2 d\mathbb{P}(\mathbf{x}) dz \\ &= \int \int \hat{f}(z|\mathbf{x})^2 d\mathbb{P}(\mathbf{x}) dz - 2 \int \int \hat{f}(z|\mathbf{x}) f(z, \mathbf{x}) d\mathbf{x} dz + C, \end{aligned} \tag{3.10}$$

em que C é uma constante arbitrária de integração indefinida.

Deste modo, escolhemos o valor de I que minimiza a função de perda estimada, avaliada no conjunto de validação $(\mathbf{x}_1^*, z_1^*), (\mathbf{x}_2^*, z_2^*), \ldots, (\mathbf{x}_m^*, z_m^*)$, dada pela expressão

$$\hat{L}(\hat{f}, f) = \sum_{i=1}^{I} \frac{1}{m} \sum_{k=1}^{m} \hat{\beta}_i^2(\mathbf{x}_k^*) - 2 \frac{1}{m} \sum_{k=1}^{m} \hat{f}(z_k^*|\mathbf{x}_k^*)^2. \tag{3.11}$$

O algoritmo do FlexCoDE já está implementado nas linguagens *R* (Izbicki, 2019) e *Python* (Pospisil, 2019b), que serão usadas na aplicação dos dados fotométricos do S-

PLUS.

3.2.2 Bayesian Mixture Density Network

As Redes Bayesianas de Densidade de Mistura ou *Bayesian Mixture Density Network* (BMDN), utilizadas no artigo de Lima *et al.* (2022), é uma classe de modelos de aprendizado profundo que combina uma rede neural convencional (no caso, uma rede bayesiana) com um modelo de densidade de mistura. O modelo de mistura é um método que modela uma função de densidade de probabilidade com base em uma soma ponderada de outras distribuições mais simples.

O modelo de *deep learning* utilizado neste artigo é baseado no Multi-Layer Perceptron (MLP), em que todos os neurônios de uma camada estão conectados a todos os neurônios da camada anterior e posterior, denominada como camada densa. O treinamento da rede é realizado através da atualização dos pesos que conectam esses neurônios, com base nos erros cometidos nas predições em cada etapa de *feedforward* e *backpropagation*.

A arquitetura do modelo proposto em Nakazoro & Ruiz (em preparação) é similar àquela utilizada em Lima *et al.* (2022), composta basicamente por uma camada de *input*, que no caso são as covariáveis fotométricas $\mathbf{x}$. Em seguida tem-se blocos de camadas internas formados por dois componentes: a chamada `DenseVariational`, que é uma variação da camada densa; e a `BatchNormalization`, que aplica uma transformação nos valores. Após os blocos tem-se uma camada densa, e por fim a camada de *output* que gera funções gaussianas descritas por uma média μ_j, um desvio padrão σ_j e um peso w_j. A Figura 3.1 ilustra o sistema arquitetado para esta rede.

A função de densidade condicional estimada pode ser obtida combinando essa funções gaussianas a partir da seguinte equação:

$$\hat{f}(z|\mathbf{x}) = \sum_{j=1}^{m} w_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(\mathbf{x}-\mu_j)^2}{2\sigma_j^2}\right), \tag{3.12}$$

em que m é o número de funções gaussianas retornadas pelo modelo.

Esse modelo também foi usado no artigo de Nakazoro & Ruiz (em preparação) para o mesmo conjunto de dados deste trabalho e seus resultados serão comparados aos do FlexCoDE.

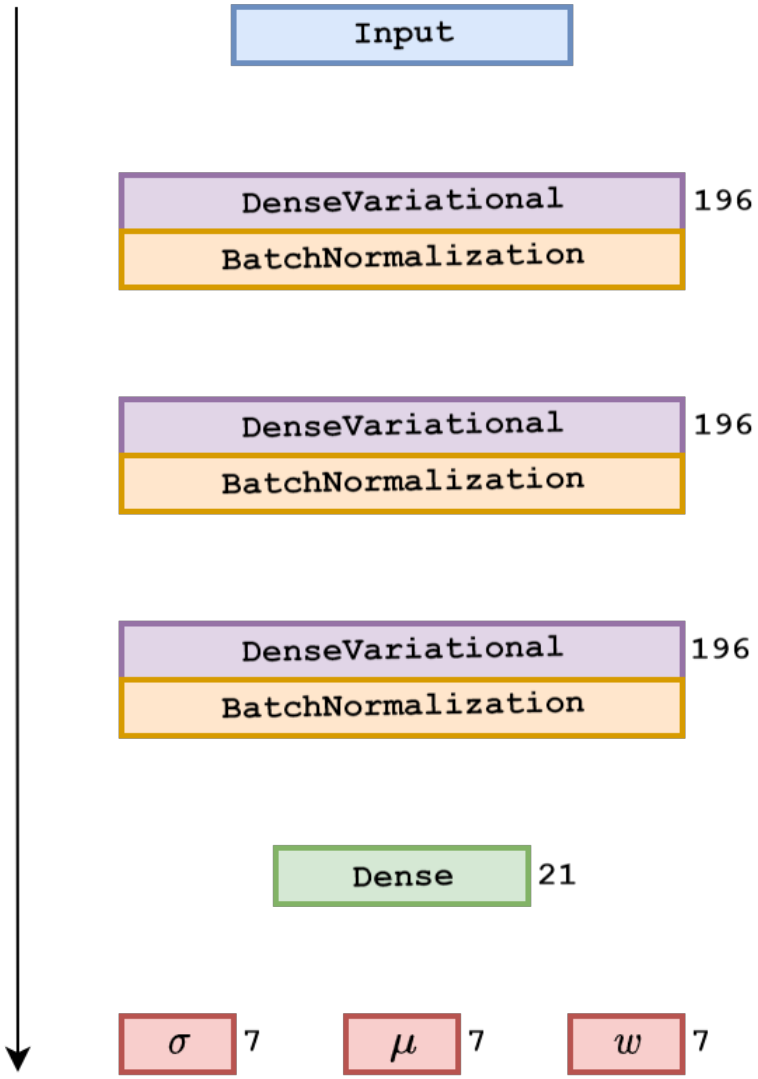

Figura 3.1: Arquitetura da Rede de Densidade de Mistura. A camada de entrada é representada em azul, seguida por blocos de camadas de `DenseVariational`, em roxo, e `BatchNormalization`, em laranja; a camada densa em verde e a camada de saída (uma `MixtureNormal`) em vermelho. Os números indicam a quantidade de neurônios em cada camada.

Capítulo 4

Medidas de qualidade do ajuste

4.1 Desvio Absoluto Mediano Normalizado

O desvio absoluto mediano normalizado, denotado por σ_{NMAD}, é uma medida robusta utilizada para avaliar a qualidade do ajuste. Essa estatística é adequada para situações em que os dados não apresentam distribuição normal e, por usar a mediana como medida de posição, não é sensível a *outliers*. Sua fórmula é dada por

$$\sigma_{NMAD} = 1.48 \cdot \text{mediana}\left(\frac{|\delta_z - \text{mediana}(\delta_z)|}{1 + z_{spec}}\right), \tag{4.1}$$

em que δ_z é o vetor de diferenças entre o *redshift* espectroscópico z_{spec} e o *redshift* fotométrico predito $\hat{z}_{phot}$ estimado pelas variáveis fotométricas. Esta medida será empregada restritamente às estimativas pontuais do conjunto de teste, ou seja, nas esperanças condicionais estimadas.

4.2 Raiz do Erro Quadrático Médio

O erro quadrático médio é outra medida de avaliação da precisão dos valores estimados em relação aos valores de referência. A vantagem da raiz do erro quadrático médio σ_{RMSE} em relação ao erro quadrático médio σ_{MSE} é que ele se mostra na mesma escala da variável. Assim, definimos como

$$\sigma_{RMSE} = \sqrt{\frac{1}{m^*}\sum_{i=1}^{m^*}(\delta_{z_i})^2}, \tag{4.2}$$

em que m^* é o tamanho da amostra de teste e δ_{z_i} é a diferença entre cada *redshift* espectroscópico $z_{(spec)i}$ e o *redshift* predito pelas variáveis fotométricas $\hat{z}_{(phot)i}$. Esta medida também será usada apenas nas estimativas das esperanças condicionais.

4.3 Fração de outliers

A fração de *outliers*, também chamada de *catastrophic failures* (falhas catastróficas), é uma medida que indica a fração de objetos cujos erros encontram-se acima de 0.15. Um *outlier* é definido por

$$\eta = \frac{|\delta_z|}{1 + z_{spec}} => 0.15, \tag{4.3}$$

em que δ_z é o vetor de diferenças entre o *redshift* espectroscópico z_{spec} e o *redshift* fotométrico predito $\hat{z}_{phot}$ estimado pelas variáveis fotométricas.

Usaremos essa métrica para avaliar as predições pontuais do *redshift*.

4.4 Risco Estimado e Erro Padrão

Como mencionado anteriormente, a função de perda para o problema da estimação da densidade condicional é dada pela Equação 3.10. Ela também pode ser escrita como:

$$\hat{L}(\hat{f}, f) = \mathbb{E}[\int \hat{f}(Z|\mathbf{x})dz] - 2\mathbb{E}[\hat{f}(Z|\mathbf{x})] + C_f, \tag{4.4}$$

em que C_f é uma constante arbitrária proveniente da integral.

Assim, avaliamos a performance do modelo através do risco estimado definido na Equação 3.11.

Também usaremos o erro padrão da função de perda estimada para avaliar a variabilidade das predições a partir da fórmula abaixo:

$$S_{\hat{L}} = \frac{1}{\sqrt{m^*}} \sqrt{\frac{\sum_{i=1}^{m^*} (\hat{L}_i(\hat{f}, f) - \bar{L}(\hat{f}, f))^2}{m^*}}, \tag{4.5}$$

na qual m^* é o tamanho amostral do conjunto de teste, $\hat{L}_i(\hat{f}, f)$ é a função de perda estimada avaliada na observação i e $\bar{L}(\hat{f}, f)$ é o valor médio da perda estimada. Neste caso, essas métricas serão aplicadas somente nas estimativas de densidade condicional.

4.5 Transformação Integral de Probabilidade

Para avaliarmos quão bem o modelo de densidade condicional se ajusta aos dados observados, podemos construir um gráfico de cobertura com base nas estimativas da densidade condicional, também chamado de valores PIT (*Probability Integral Transform*) (Schmidt *et al.*, 2020).

Seja $\hat{f}(z|\mathbf{x})$ a estimativa da densidade condicional de z dado $\mathbf{x}$. Calculamos a cobertura de $\hat{f}(z|\mathbf{x})$ da seguinte forma:

$$PIT(\hat{f}, z) = 1 - \hat{F}(z|\mathbf{x}), \tag{4.6}$$

em que $\hat{F}(z|\mathbf{x})$ é a função de distribuição acumulada estimada de z dado os valores observados $\mathbf{x}$, isto é,

$$\hat{F}(z|\mathbf{x}) = \int_0^z \hat{f}(z|\mathbf{x})dz, \tag{4.7}$$

que pode ser estimada a partir de uma aproximação numérica.

O cálculo dos valores de cobertura já está implementado tanto na linguagem *R* quanto na linguagem *Python*, e a documentação está disponível em Pospisil (2019a).

Se a densidade condicional estimada for similar à densidade condicional verdadeira, então esperamos que a distribuição dos valores PIT se assemelhem a uma distribuição uniforme padrão $U(0, 1)$. Assim, podemos construir um histograma para os valores de cobertura de todos os quasares, bem com um gráfico *P-P Plot* (ou gráfico de probabilidade-probabilidade), que traça pontos de probabilidade cumulativa dos valores observados *versus* dados gerados de uma uniforme padrão.

4.6 "Bookmaker" Odds

Considere um *threshold* definido Δz pequeno e z_{max} o valor de z que maximiza a função de densidade de probabilidade condicional estimada $\hat{f}(z|\mathbf{x})$. A medida de *"bookmaker" odds* (Benitez, 2000) descreve a razão de chances entre a probabilidade da variável Z pertencer ao intervalo $H_c = z_{max} \pm \Delta z$ e a probabilidade da variável Z não pertencer a

este mesmo intervalo. Em termos matemáticos, definimos

$$P(H_c|\mathbf{x}) = \int_{z_{max}-\Delta z}^{z_{max}+\Delta z} f(z|\mathbf{x})dz \tag{4.8}$$

como sendo a área sob a curva dentro da região $H_c = z_{max} \pm \Delta z$ e

$$P(\bar{H}_c|\mathbf{x}) = \int_{0}^{z_{max}-\Delta z} f(z|\mathbf{x})dz + \int_{z_{max}+\Delta z}^{\infty} f(z|\mathbf{x})dz \tag{4.9}$$

como sendo a área sob a curva fora da região $H_c = z_{max} \pm \Delta z$.

Então, a medida de *"bookmaker" odds* é definida como

$$O(H_c|\mathbf{x}) = \frac{P(H_c|\mathbf{x})}{P(\bar{H}_c|\mathbf{x})}, \tag{4.10}$$

em que as probabilidades podem ser calculadas a partir de uma aproximação numérica das integrais (por exemplo, Regra dos Trapézios) usando a função de probabilidade condicional estimada $\hat{f}(z|\mathbf{x})$ e um intervalo discreto para z.

Essa métrica será usada para quantificar a concentração das curvas de probabilidade condicional do *redshift* fotométrico para cada amostra. Basicamente, quanto mais concentrada for uma curva de densidade, maior será a probabilidade estimada $\hat{P}(H_c|\mathbf{x})$ e, consequentemente, maior será a *odds*.

Capítulo 5

Resultados

5.1 *Data Splitting*

Os dados disponibilizados foram previamente divididos aleatoriamente em 75% para treinamento do modelo e 25% para teste e avaliação do ajuste. Dentro do conjunto de treinamento, os dados foram particionados em 5 *folds*, de maneira estratificada em relação ao *redshift*. Usaremos 4 *folds* para treinar o modelo, ou seja, 80% do conjunto de treinamento, e o *fold* remanescente para selecionar os parâmetros de ajuste. Repetimos este procedimento alterando cada *fold* como dados de validação. A Figura 5.1 ilustra essa segmentação.

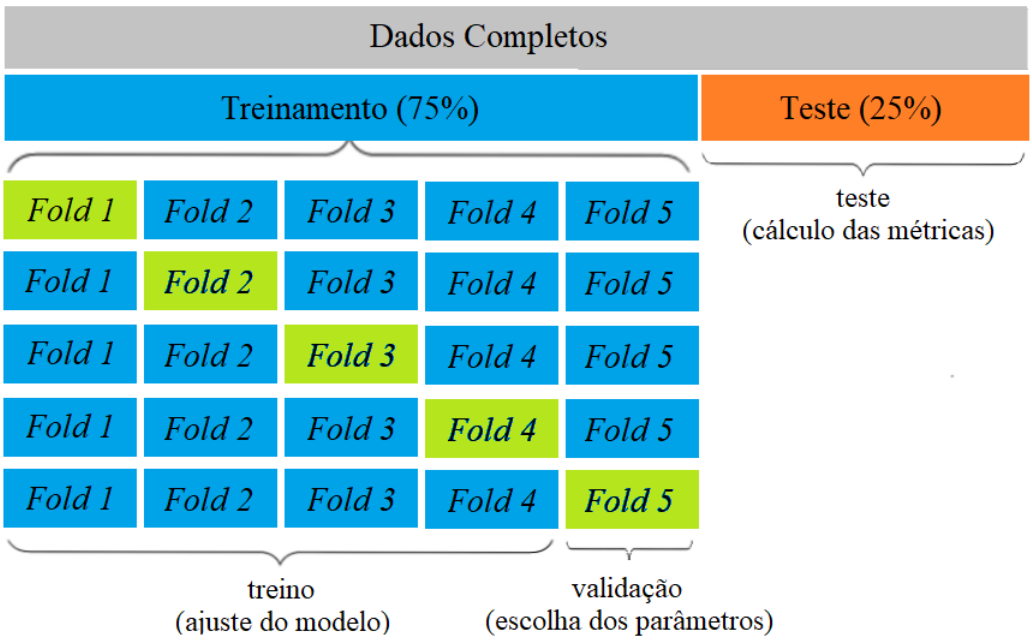

Figura 5.1: Esquematização da divisão da amostra.

Por fim, o conjunto de teste será empregado no cálculo das métricas definidas na Seção 4. A Figura 5.2 mostra o histograma do *redshift* espectroscópico para esses três conjuntos. Vemos que eles seguem a mesma distribuição, mantendo a quantidade de *redshifts* observados em diferentes intervalos de maneira proporcional.

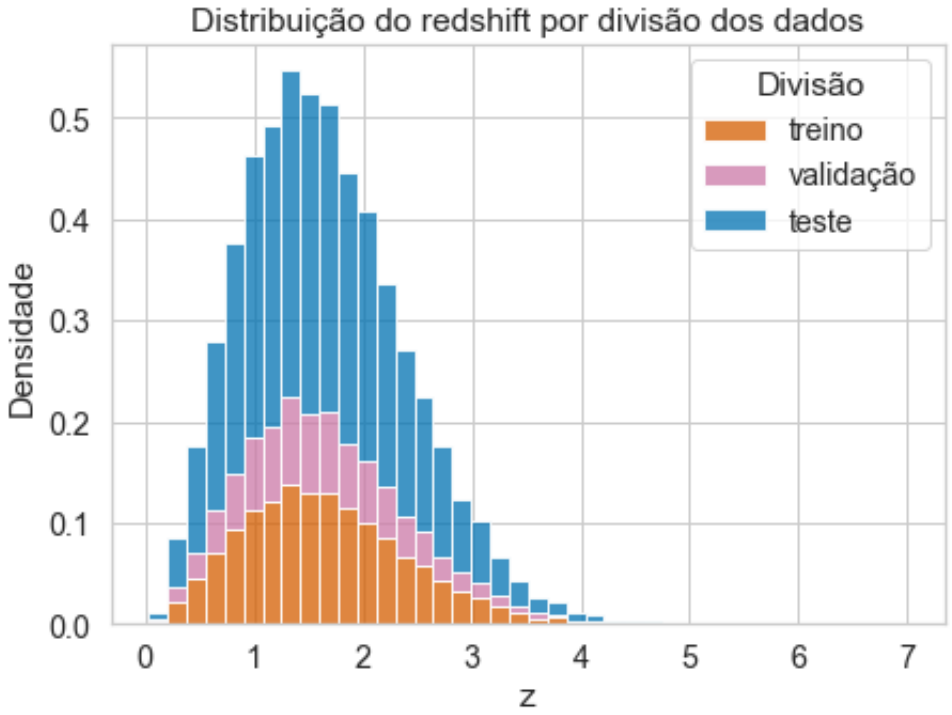

Figura 5.2: Histograma da distribuição do *redshift* espectroscópico nas amostras de treinamento, validação e teste.

5.2 Densidade condicional e a influência das *narrow bands*

5.2.1 *Pipeline* do experimento

Uma vez que o interesse científico é melhorar a predição do *redshift* a partir de variáveis fotométricas e, além disso, verificar se os filtros de banda estreita trazem informação relevante para estimação do mesmo, construiremos essencialmente dois modelos: um baseado nas 4 cores construídas a partir das bandas largas do S-PLUS, juntamente com as cores formadas a partir das 2 magnitudes do WISE e as 2 magnitudes do GALEX, totalizando 8 variáveis preditoras; e outro modelo adicionando-se também as 7 cores feitas com as bandas estreitas do S-PLUS. A Tabela 5.1 ilustra melhor a composição destes modelos.

Tabela 5.1: Composição de variáveis preditoras para cada base de modelos.

Cores Criadas	**Modelo sem *narrow bands***	**Modelo com *narrow bands***
FUV - r	✓	✓
NUV - r	✓	✓
u - r	✓	✓
J0378 - r	⊘	✓
J0395 - r	⊘	✓
J0410 - r	⊘	✓
J0430 - r	⊘	✓
g - r	✓	✓
J0515 - r	⊘	✓
r - J0660	⊘	✓
r - i	✓	✓
r - J0861	⊘	✓
r - z	✓	✓
r - W1	✓	✓
r - W2	✓	✓

Com o objetivo de obter maior representatividade e variabilidade dos dados, realizaremos esse procedimento para os 5 *folds* definidos no *data splitting*, alternando os papéis entre treino e validação para cada *fold*, resultando em 5 modelos sem as *narrow bands* e 5 modelos com essas variáveis. Computamos a densidade condicional estimada a partir da média das predições de $f(z|\mathbf{x})$ obtidas entre os 5 modelos usando o conjunto de teste.

Faremos também o cálculo da esperança condicional estimada a partir de (3.3), de forma a analisar a similaridade com o *redshift* espectroscópico e contrastar com o grau de informatividade que a densidade condicional traz.

Posteriormente, avaliaremos a performance dos ajustes a partir das métricas definidas na Seção 4, verificando a adequabilidade dos modelos aos dados observados. Assim, iremos comparar os resultados obtidos do ajuste com e sem as *narrow bands*, averiguando a influência destas na estimativa do *redshift*.

Ressalta-se que os dados utilizados para este procedimento foram tratados, conforme mencionado na Seção 2.3.1. Além disso, tanto os filtros de banda larga quanto os de banda estreita do S-PLUS foram empregados na abertura `PStotal`, dado que esta representa a abertura mais importante e com melhores resultados em pesquisas anteriores.

5.2.2 Escolha dos parâmetros do FlexCoDE

Para ajustar o modelo via FlexCoDE, partimos da função já implementada no programa *R*. Esta função possui os seguintes argumentos (parâmetros de ajuste):

- **nIMax**: o número máximo possível de componentes da expansão em série, isto é, o algoritmo encontrará o melhor I tal que $I \leq$ nIMax;
- **system**: escolha da base ortonormal para z;
- **regressionFunction**: o método de regressão a ser utilizado para estimar os coeficientes de expansão;
- **regressionFunction.extra**: parâmetros adicionais relacionados à função de regressão;
- **chooseDelta**: valor *booleano*, indica se o parâmetro δ, limite de área para remover saliências espúrias, deve ser escolhido;
- **chooseSharpen**: valor *booleano*, indica se o parâmetro α, parâmetro para aprimorar a estimativa final tal que $\hat{f}_{new}(z|\mathbf{x}) = (\hat{f}(z|\mathbf{x}))^{\alpha}$, deve ser escolhido.

Tanto o número ideal de componentes de expansão, I, quanto os parâmetros δ e α foram escolhidos tal que minimizem a função de perda estimada definida em (3.11) no conjunto de validação. Caso `chooseDelta = FALSE` teríamos $\delta = 0$, ou seja, não removeríamos as protuberâncias da densidade, assim como se `chooseSharpen = FALSE` teríamos $\alpha = 1$.

Para o nosso estudo, selecionamos como base ortonormal $\phi_i(z)$ a base de Fourier, definida por

$$\phi_1(z) = 1; \quad \phi_{2i+1}(z) = \sqrt{2}\sin(2\pi i z); \quad \phi_{2i}(z) = \sqrt{2}\cos(2\pi i z), \quad i \in \mathbb{N}, \tag{5.1}$$

que é o *default* da função do FlexCoDE e que, por meio de análises preliminares, foi a que apresentou melhor performance.

Também definiremos a quantidade máxima de coeficientes de expansão como `nIMax=45`, que demonstrou um custo computacional e complexidade razoáveis. A função de regressão a ser usada para estimar os parâmetros $\beta_i(\mathbf{x})$ será *Random Forest*, visto que já havia apresentado boa performance como modelo de regressão diretamente para a estimação de z em pesquisas anteriores e possui robustez a dados faltantes. O algoritmo *XGBoost* também foi testado previamente como regressor, porém obteve pior desempenho.

5.2.3 Avaliação e comparativa dos modelos de banda larga e estreita

Após ajustarmos os 10 modelos de densidade condicional via FlexCoDE conforme a definição destes mencionada no *pipeline*, obtivemos os seguintes resultados parciais:

Tabela 5.2: Hiperparâmetros estimados e métricas dos modelos com e sem os filtros de banda estreita.

	Modelo sem *narrow bands*					**Modelo com *narrow bands***				
Modelo	I	α	δ	$\hat{L}(\hat{f}, f)$	$S_{\hat{L}}$	I	α	δ	$\hat{L}(\hat{f}, f)$	$S_{\hat{L}}$
1	36	1.062	0.129	-1.308	0.017	45	2.639	0.225	-2.832	0.049
2	38	1.062	0.096	-1.313	0.017	45	2.639	0.225	-2.834	0.050
3	38	1.062	0.064	-1.312	0.016	45	2.639	0.225	-2.808	0.049
4	44	1.062	0.064	-1.307	0.017	45	2.639	0.193	-2.836	0.049
5	34	1.062	0.096	-1.309	0.016	45	2.639	0.225	-2.814	0.049

Podemos observar a partir da Tabela 5.2 à esquerda que a quantidade ideal de coeficientes de expansão variou bastante para cada modelo, de 34 a 44. O valor de α foi o mesmo para todos os modelos, resultando em $\alpha = 1.062$, um valor bem próximo de 1 inclusive, indicando que o aprimoramento para a estimativa final não foi tão grande. Já a estimativa para o valor de δ também variou de acordo com cada ajuste, resultando em valores de 0.129, 0.096 e 0.064. Apesar dos diferentes parâmetros encontrados, as estimativas para a função de perda e seu erro padrão não variaram tanto, com uma diferença apenas na segunda ou terceira casa decimal.

Já analisando a Tabela 5.2 à direita, referente aos resultados do modelo incluindo as cores de bandas estreitas, vemos que o valor de I atingiu o limiar definido para seleção do mesmo. Isso significa que, a partir do conjunto de validação, teríamos a quantidade ideal de coeficientes de expansão superior a 45. Foram feitas algumas tentativas incrementando o valor do argumento da função `nIMax`. Entretanto, a diferença em relação tanto às estimativas da densidade condicional quanto ao próprio risco estimado não foi tão significativa, o que implica que a inserção de muitos coeficientes e o aumento da complexidade do modelo não é compensada pelo pouco ganho de performance.

O que se pode verificar com notoriedade é que, de maneira geral, temos uma diminuição significativa na função de perda em relação aos modelos sem as *narrow bands*, com uma diferença de cerca de 1.57, apesar do aumento no seu erro padrão.

Realizando a média das predições no conjunto de teste e computando novamente a função de perda estimada em ambos os conjuntos de modelos, fica evidente na Tabela 5.3

a influência das bandas estreitas do S-PLUS na performance do ajuste.

Tabela 5.3: Riscos estimados e erros padrões dos modelos com e sem *narrow*.

Modelo	$\hat{L}(\hat{f}, f)$	$S_{\hat{L}}$
Sem *narrow bands*	-1.348	0.016
Com *narrow bands*	-2.921	0.047

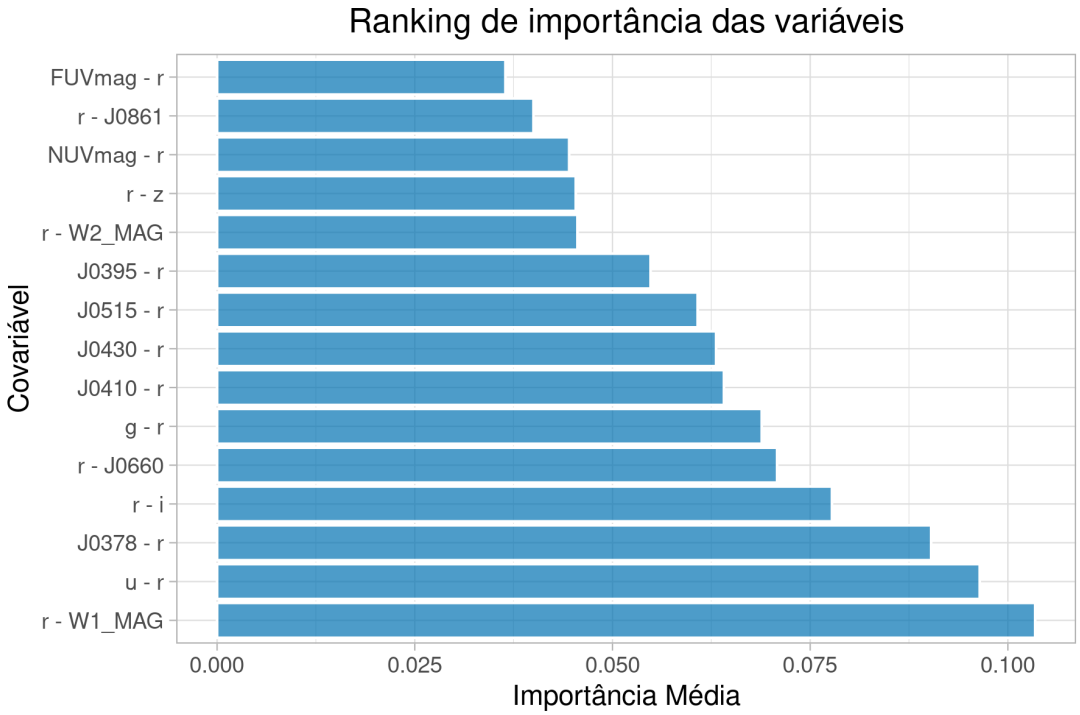

Figura 5.3: Gráfico de importância das variáveis no modelo 1 com as *narrow bands*.

A Figura 5.3 mostra o grau de importância das variáveis para a estimação dos coeficientes $\beta_i(\mathbf{x})$, especificamente do modelo 1, que reflete indiretamente na importância para a estimação do *redshift*. Podemos observar que a cor `r - W1_MAG` contribui fortemente para a predição, enquanto que as cores com ambas as magnitudes do GALEX, `FUVmag - r` e `NUVmag - r`, demonstram muito menos importância que as demais. Isso se deve possivelmente ao fato dessas bandas possuírem poucos valores conhecidos, resultando em pouca informação acerca do *redshift*.

Já as cores formadas pelos filtros de banda larga e estreita do S-PLUS aparecem se alternando nas posições do *ranking*, indicando que essas covariáveis possam estar dividindo suas contribuições no modelo. Entretanto, ainda temos um cenário de grande importância de algumas das *narrow bands*, como é o caso da cor `J0378 - r` que aparece na terceira posição.

Agora, para analisarmos a estimativa pontual dos modelos com e sem os filtros de banda estreita, fazemos o cálculo utilizando uma aproximação numérica de (3.3) levando em consideração as predições médias feitas anteriormente. Desta forma, obtemos as seguintes métricas:

Tabela 5.4: Métricas de avaliação para as estimativas pontuais.

Modelo	σ_{RMSE}	σ_{NMAD}	η
Sem narrow bands	0.442	0.099	0.228
Com narrow bands	0.434	0.058	0.189

Percebe-se que o modelo contendo as *narrow bands* resulta em valores quase iguais ou um pouco menores nas métricas de avaliação para as estimativas pontuais. Porém, é uma diferença menos significativa quando comparada à discrepância no risco estimado das densidades condicionais, indicando que não há grande influência das bandas estreitas na estimativa da esperança condicional do *redshift*. Contudo, vale ressaltar que um modelo de densidade condicional traz muito mais informações sobre o *redshift* do que uma simples média.

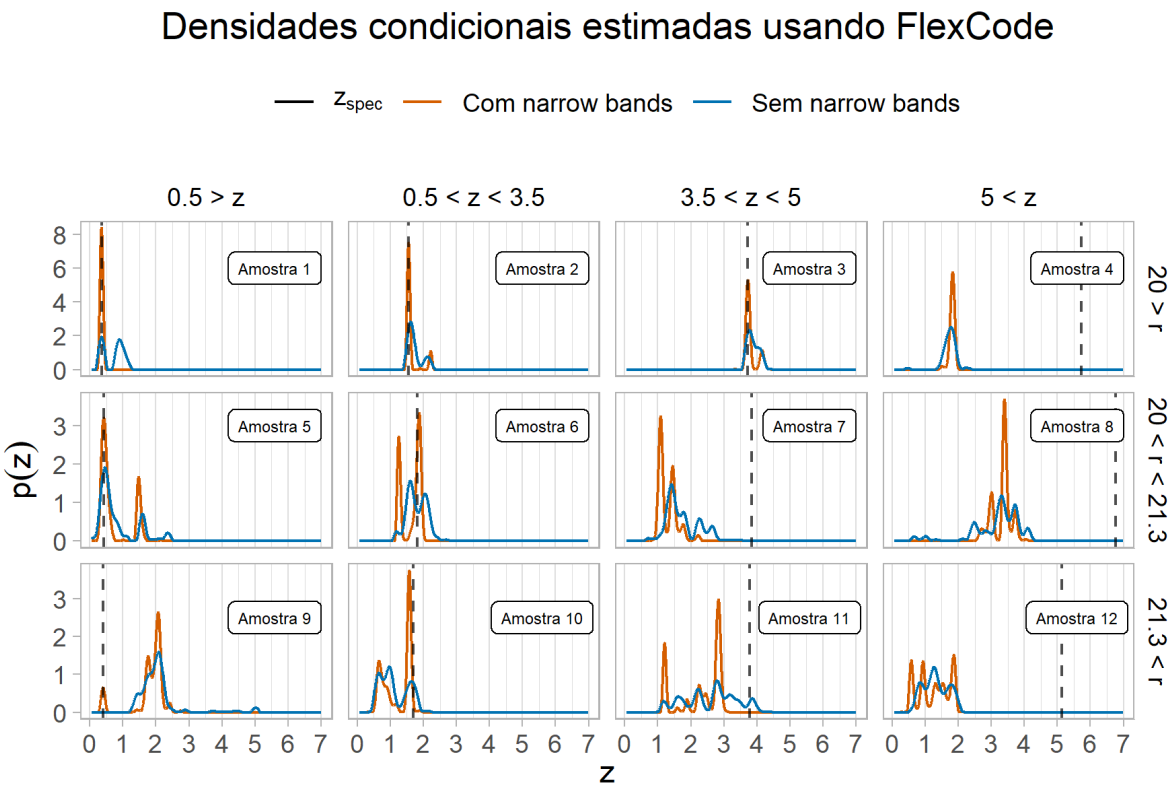

Figura 5.4: Gráfico das estimativas das densidades condicionais em 12 amostras de quasares. A curva laranja corresponde à densidade condicional estimada pelo modelo com as *narrow bands*, a curva azul se refere à densidade estimada pelo modelo sem essas bandas, e a linha vertical preta tracejada é o *redshift* espectroscópico.

Foram selecionadas 12 amostras de quasares em diferentes intervalos de z_{spec} e da magnitude `r_PStotal` com o propósito de comparar as densidades condicionais, como mostradas na Figura 5.4. Assim, fica evidente a diferença das densidades entre os modelos com e sem as *narrow bands*. As curvas de densidade do modelo com as bandas estreitas apresentam geralmente picos mais elevados e estreitos, em contraposição com as curvas do modelo sem essas bandas, que são mais achatadas e menores. Por outro lado, a densidade condicional do modelo com as bandas estreitas aparenta possuir maior multimodalidade, de forma mais intensa por exemplo no gráfico da última amostra.

Vemos também que o *redshift* espectroscópico, representado pela reta tracejada vertical na cor preta, se encontra bem próximo a uma das modas da curva em laranja nas amostras com $z < 3.5$, até mesmo no quasar da amostra 9 cuja reta vertical preta está centralizada no pico menor. Além disso, em alguns casos, o modelo consegue centralizar bem sua maior moda ao *redshift* espectroscópico e com alta densidade, conforme vemos nos gráficos das duas primeiras amostras.

Já quando as amostras possuem um $z_{spec} > 5$, nenhum dos dois modelos conseguem ajustar bem as densidades condicionais próximas ao *redshift* espectroscópico. Segundo Lyke *et al.* (2020), os objetos com $z > 5$ devem ser considerados suspeitos pelo fato de muitas vezes trazerem classificações espectroscópicas enganosas e medidas não confiáveis.

Tabela 5.5: Magnitude e *redshift* espectroscópico das amostras selecionadas com as estimativas pontuais de z_{phot} dos modelos com e sem *narrow bands*.

Amostra	**r_PStotal**	z_{spec}	z_{mean} **sem narrow**	z_{mean} **com narrow**
1	17.662	0.369	0.703	0.356
2	18.836	1.543	1.732	1.637
3	19.958	3.727	3.888	3.815
4	19.263	5.721	1.740	1.816
5	20.664	0.431	0.800	0.729
6	20.781	1.809	1.796	1.641
7	21.263	3.844	1.742	1.296
8	20.310	6.773	3.177	3.319
9	21.471	0.396	2.033	1.823
10	21.987	1.689	1.109	1.222
11	21.918	3.777	2.613	2.306
12	21.703	5.130	1.316	1.265

A Tabela 5.5 exibe um descritivo dessas amostras, e por ela podemos comparar também as estimativas pontuais obtidas pelos modelos com e sem filtros de banda estreitas. Vemos que ambos os modelos apresentam estimativas de z_{phot} bem próximas de maneira geral.

Porém, quando o valor do desvio para o vermelho aumenta, nenhum modelo consegue estimar adequadamente.

Para mensurar quão concentradas estão as curvas de densidade condicional dos modelos, calculamos a medida de *odds* para cada amostra através de (4.10) considerando um *threshold* de $\Delta z = 0.02$ e computamos a média e o erro padrão, descritas na Tabela 5.6.

Tabela 5.6: *Odds* média e erro padrão para os modelos com e sem filtros de banda estreita.

Modelo	$\bar{O}(H_c\|\mathbf{x})$	$\hat{S}_O$
Sem narrow bands	0.0913	0.0004
Com narrow bands	0.2703	0.0013

Nota-se que o valor médio da *odds* do modelo considerando as *narrow bands* é 3 vezes maior que a *odds* média do ajuste sem considerá-las, indicando que as densidades condicionais são em média bem mais concentradas quando adicionamos as cores com informações de banda estreita, apesar de um aumento no erro padrão também. Reitera-se, no entanto, que a medida de *odds* não leva em consideração o *redshift* espectroscópico, ou seja, embora tenhamos curvas mais densas, não necessariamente elas estão centralizadas ou próximas do referencial. A medida que nos informa este fato é a própria função de perda, conforme visto na Tabela 5.3.

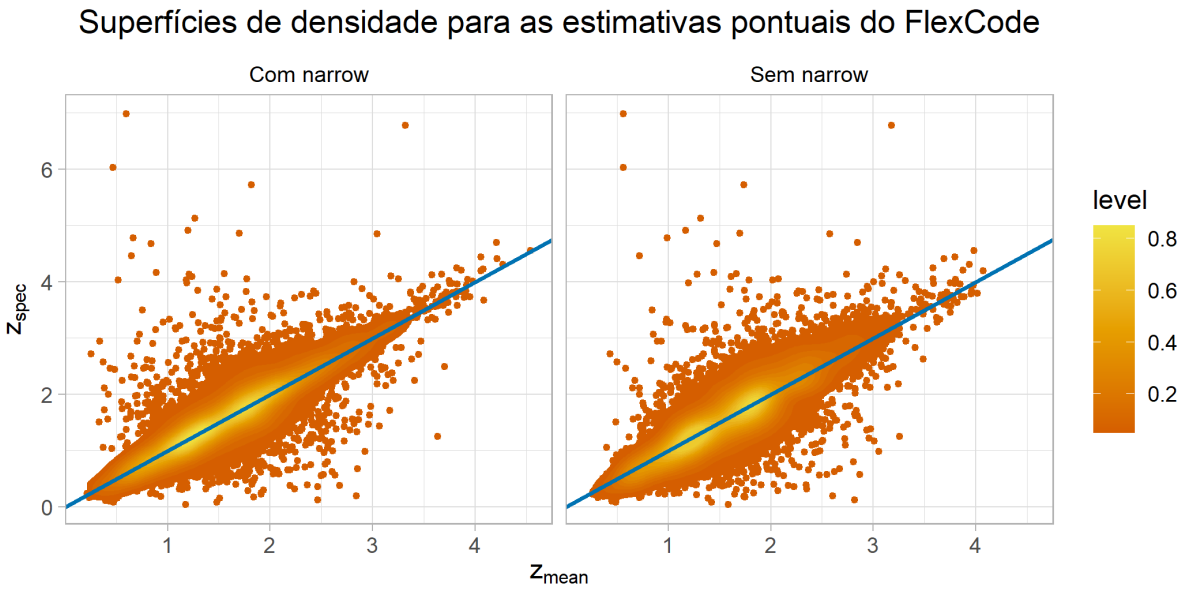

Figura 5.5: Mapa de densidade de pontos dos *redshifts* espectroscópicos versus valores médios preditos estimados. No lado esquerdo, temos os *redshifts* fotométricos estimados pelo modelo com as *narrow bands* e do lado esquerdo os valores obtidos com o modelo sem as *narrow bands*.

Apesar de destacarmos o ganho de informação das densidades condicionais com a

presença dos filtros de banda estreita, vemos que as estimativas pontuais do *redshift* desses modelos são bem próximas, de acordo com a Figura 5.5. Nota-se que a densidade de pontos são bem similares nos dois gráficos, com uma concentração um pouco mais acentuada próxima da reta de identidade com a informação das bandas estreitas, mas nada muito significativo. Além disso, vemos uma grande nuvem de pontos distantes da reta em azul, principalmente acima dela onde temos *redshifts* espectroscópicos mais elevados e erros maiores de predição.

Ou seja, a predição de z levando em consideração somente a média condicional nos levaria a uma conclusão limitada de que as *narrow bands* não contribuem significativamente para a estimação do *redshift*, fato que se opõe aos resultados observados até agora por meio da densidade condicional.

Usaremos também como ferramenta de diagnóstico dos modelos o gráfico de cobertura das estimativas de densidade condicional, dado pelas Figuras 5.6 e 5.7. Neste gráfico, temos o histograma dos *PIT values* e o intervalo de confiança para uma distribuição uniforme padrão.

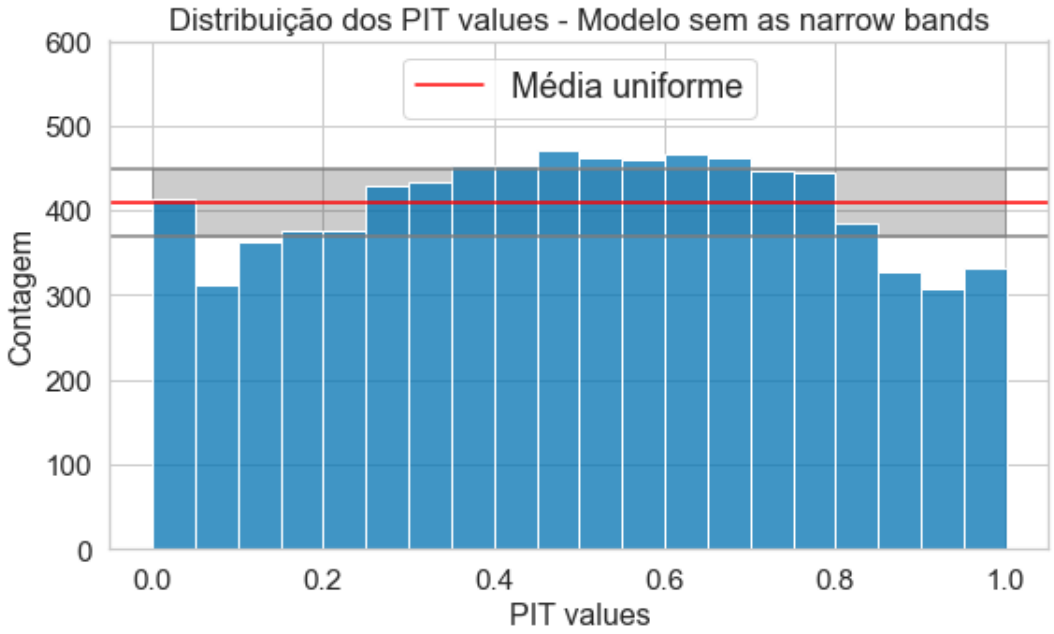

Figura 5.6: Histograma dos valores PIT baseados na densidade condicional estimada sem as *narrow bands*.

Pode-se notar que as barras de frequência do histograma na Figura 5.6 se desviam um pouco da banda de confiança da distribuição uniforme padrão, principalmente nos valores mais próximos das extremidades, indicando que os *PIT values* desse modelo possam não ser uniformemente distribuídos. Ou seja, a densidade condicional estimada de z_{phot} não

se assemelha tanto à densidade condicional real, dado as covariáveis deste modelo.

Já na Figura 5.7 vemos um cenário ainda mais crítico, uma vez que as barras laterais do histograma encontram-se bem acima dos limites do intervalo de confiança. Ademais, vemos um comportamento oscilatório nos dados, visto que as barras em torno de 0.1 e 0.9 passam a ficar abaixo do limite inferior e, no centro, dentro do intervalo.

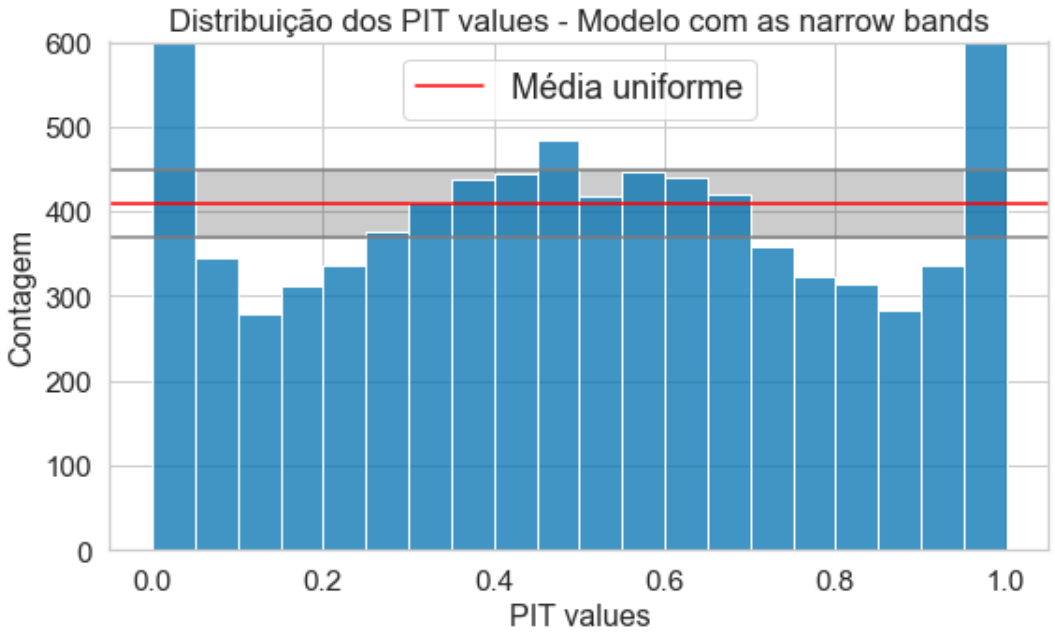

Figura 5.7: Histograma dos valores PIT baseados na densidade condicional estimada com as *narrow bands*.

Isso nos traz evidências de que o modelo possa não estar estimando bem a densidade condicional de z levando em conta as bandas estreitas, além das demais variáveis já utilizadas no modelo anterior. Tal fato pode ser justificado possivelmente por conta da curva de densidade do modelo com as bandas estreitas apresentar "picos" bem mais altos do que o modelo sem *narrow*, concentrando as probabilidades em um intervalo muito pequeno de z e, consequentemente, fazendo com que o *redshift* espectroscópico de diversas amostras de quasares se encontre em uma região com densidade próxima de zero, principalmente quando temos um z_{spec} mais alto, conforme já visto anteriormente.

A mesma verificação pode ser realizada a partir do gráfico *P-P Plot*, exibido nas Figuras 5.8 e 5.9, que compara os pontos de probabilidade cumulativa dos dados da amostra com a distribuição de probabilidade cumulativa específica de teste, que no nosso caso é a uniforme padrão.

Observa-se que na Figura 5.8 os pontos de probabilidades encontram-se relativamente próximos da reta de identidade, apesar de um desvio um pouco maior quando a dis-

tribuição cumulativa da uniforme está entre 0.7 e 0.9, indicando que a distribuição dos valores de cobertura sobre a densidade condicional estimada sem uso das *narrow bands* se assemelha um pouco à uma distribuição uniforme padrão.

Os pontos do gráfico da Figura 5.9 também não parecem se distanciar muito da reta de identidade, principalmente nos valores entre 0.25 e 0.75, apesar de possuírem caudas um pouco mais pesadas conforme já havíamos identificado no histograma da Figura 5.7, indicando uma menor similaridade com a distribuição uniforme padrão.

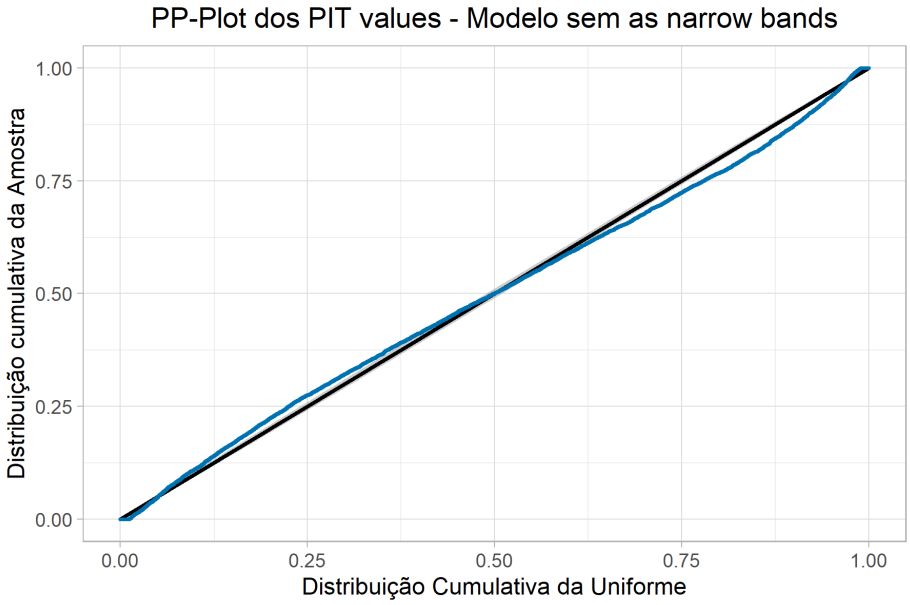

Figura 5.8: Gráfico de probabilidade-probabilidade dos *PIT values* do modelo sem *narrow*.

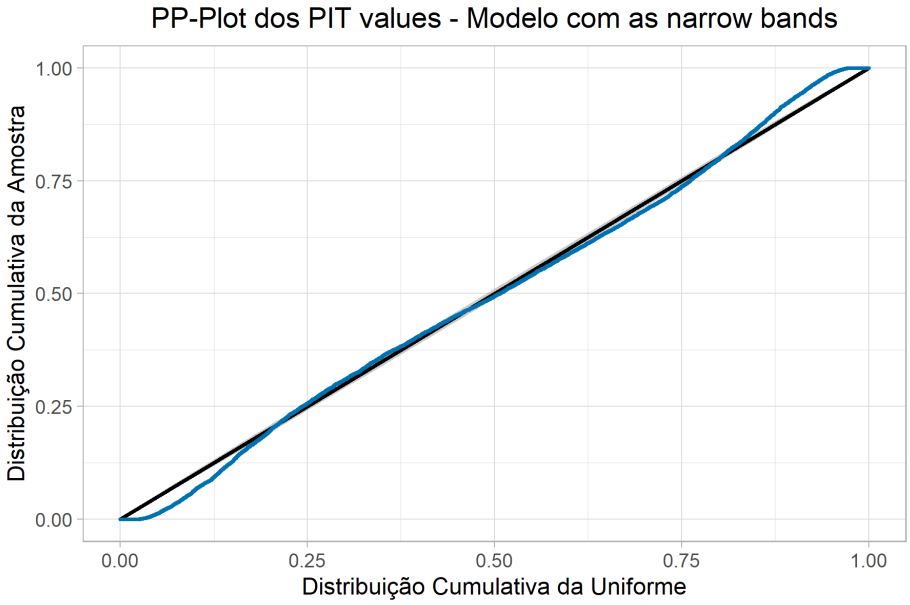

Figura 5.9: Gráfico de probabilidade-probabilidade dos *PIT values* do modelo com *narrow*.

Um teste de Kolmogorov-Smirnov foi realizado em ambos *PIT values* para verificar a similaridade entre a distribuição amostral e a distribuição empírica da uniforme padrão, em que hipótese nula é que não existe diferenças entre elas. O p-valor para a amostra do modelo com as *narrow bands* foi de 1.2×10^{-11} e para o modelo sem as *narrow bands* foi de 1.2×10^{-9}, ou seja, ambos os testes rejeitam a hipótese nula, nos dando indícios de que os valores de cobertura não seguem uma distribuição uniforme padrão. Isso implica que os os modelos ajustados não estão estimando tão bem suas respectivas funções de densidade a *posteriori*.

5.2.4 Comparação com os resultados das redes neurais

Uma vez que constatamos uma melhoria significativa na predição do *redshift* fotométrico utilizando-se cores com filtros de banda estreita por meio do algoritmo FlexCoDE, torna-se interessante verificar se estes resultados se repetem a partir de outro método de estimação, a fim de concluir de fato se o mérito de aprimoramento se dá pelas bandas estreitas ou se há um viés do próprio modelo com elas.

Deste modo, reaproveitaremos as saídas do modelo de Redes Bayesianas de Densidade de Mistura obtidas em Nakazono & Ruiz (em preparação), cuja base de dados para treinamento, validação e teste é exatamente a mesma deste trabalho, tornando a comparação mais justa e direta. Além disso, o ajuste foi realizado tanto com as cores formadas pelas *narrow bands* quanto sem elas.

Assim, foram calculadas as estimativas de densidade condicional nos dados de teste por meio da equação (3.12) utilizando-se os vetores de média, desvio padrão e pesos, em que cada vetor continha 7 valores. A Tabela 5.7 apresenta a função de perda estimada e seu respectivo erro padrão para os modelos.

Tabela 5.7: Riscos estimados e erros padrões dos modelos de redes neurais com e sem *narrow*.

Modelo	$\hat{L}(\hat{f}, f)$	$S_{\hat{L}}$
Sem *narrow bands*	-1.462	0.021
Com *narrow bands*	-2.852	0.051

Vemos que o modelo de redes neurais sem a presença das bandas estreitas apresentou uma perda estimada menor que o FlexCoDE sem essas bandas, enquanto que o modelo na presença das mesmas usando o algoritmo FlexCoDE obteve um desempenho um pouco melhor que a BMDN, comparando as Tabelas 5.7 e 5.3. Todavia, ambos os métodos

de estimação são concordantes na diminuição significativa da função de perda estimada quando inserimos a informação das *narrow bands* na predição do *redshift* fotométrico.

Já quando analisamos as métricas de performance das estimativas pontuais destes novos modelos, não vemos uma diferença tão significativa dos resultados obtidos anteriormente, conforme exibido na Tabela 5.8, apesar de uma melhoria um pouco maior no σ_{RMSE} do que observado no FlexCoDE. Assim, temos indícios de que não há influência das *narrow bands* na estimação da esperança condicional do *redshift* fotométrico, independentemente do modelo usado.

Tabela 5.8: Métricas de avaliação para as estimativas pontuais.

Modelo	σ_{RMSE}	σ_{NMAD}	η
Sem narrow bands	0.478	0.080	0.205
Com narrow bands	0.450	0.051	0.193

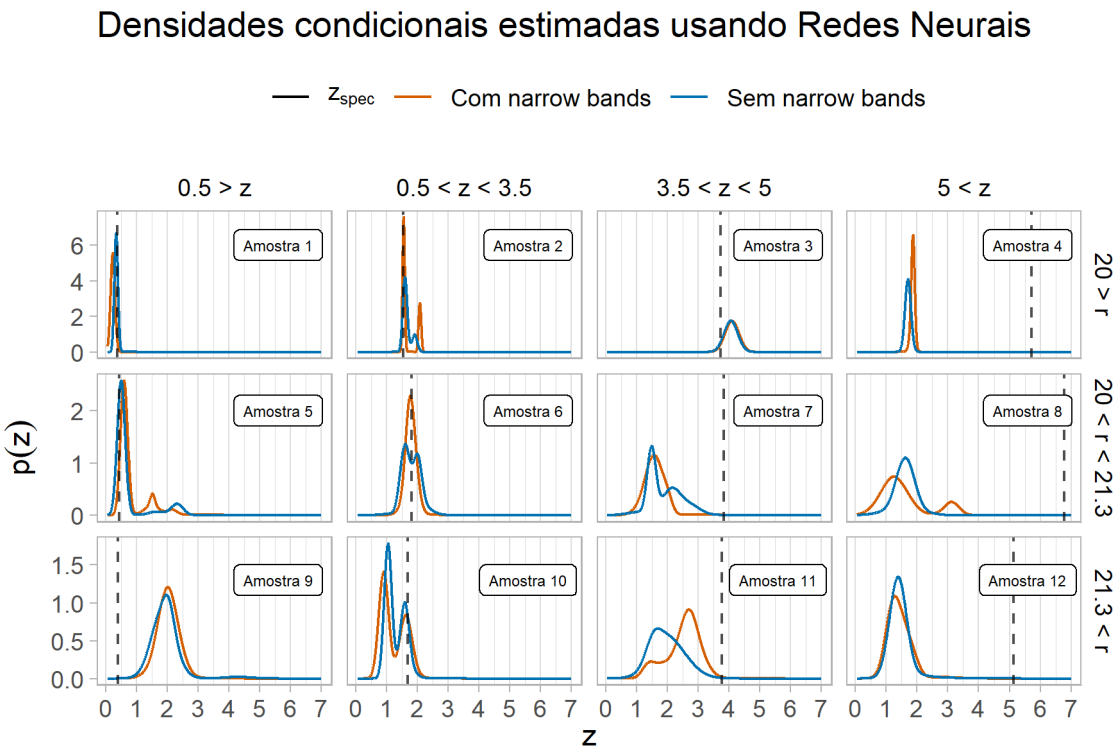

Figura 5.10: Gráfico das estimativas das densidades condicionais em 12 amostras de quasares. A curva laranja corresponde à densidade condicional estimada pela rede neural com as *narrow bands*, a curva azul se refere à densidade estimada pela rede neural sem essas bandas, e a linha vertical preta tracejada é o *redshift* espectroscópico.

Avaliamos também as curvas de densidade condicional para as mesmas amostras selecionadas anteriormente e plotamos nos gráficos exibidos na Figura 5.10. O primeiro ponto que podemos notar é que as curvas em laranja e azul parecem estar mais semelhan-

tes entre si quando comparado ao cenário visto no modelo de FlexCoDE, indicando que as estimativas de $f(z|\mathbf{x})$ não se diferem tanto com a presença das bandas estreitas neste caso, apesar da alteração na *loss function*.

Ou seja, enquanto que temos picos mais elevados e estreitos para as estimativas de densidade condicional incluindo as *narrow bands* no método FlexCoDE, nas redes neurais as curvas são quase equiparadas, e em alguns casos a curva sem os filtros de banda estreita (em azul) até supera a curva com as mesmas (em laranja), como vemos na última amostra da Figura 5.10.

Todavia, o comportamento no que diz respeito à eficiência na predição parece ser o mesmo do algoritmo FlexCoDE, no sentido que também temos uma maior assertividade quando o *redshift* espectroscópico encontra-se numa margem inferior à 3.5, e um maior desvio da moda da função ao z_{spec} quando este possui altos valores.

As estimativas de esperança condicional de z, bem como o *redshift* espectroscópico e magnitude observados para cada amostra encontram-se na Tabela 5.9. Podemos dizer de maneira geral que as conclusões são as mesmas retiradas da Tabela 5.5, no sentido de que a estimação pontual do *redshift* fotométrico não se difere muito com a informação das bandas estreitas para esses quasares.

Tabela 5.9: Magnitude e *redshift* espectroscópico das amostras selecionadas com as estimativas pontuais de z_{phot} das redes neurais com e sem *narrow bands*.

Amostra	**r_PStotal**	z_{spec}	z_{mean} **sem narrow**	z_{mean} **com narrow**
1	17.662	0.369	0.341	0.231
2	18.836	1.543	1.603	1.559
3	19.958	3.727	4.063	4.105
4	19.263	5.721	1.702	1.867
5	20.664	0.431	0.502	0.597
6	20.781	1.809	1.610	1.775
7	21.263	3.844	1.500	1.576
8	20.310	6.773	1.616	1.245
9	21.471	0.396	1.964	2.008
10	21.987	1.689	1.050	0.908
11	21.918	3.777	1.698	2.704
12	21.703	5.130	1.388	1.284

O valor médio da *odds* e seu erro padrão para os modelos de redes neurais são expostos na Tabela 5.10, e a partir dela podemos concluir que, apesar da Figura 5.10 causar a impressão de que as curvas de densidade do modelo com as bandas estreitas são semelhantes às curvas sem essas bandas, a função de densidade condicional estimada com as *narrow*

bands é, em média, bem mais concentrada do que a função estimada sem as *narrow bands*, dado que o valor de $\bar{O}(H_c|\mathbf{x})$ é 2.5 vezes maior.

Tabela 5.10: *Odds* média e erro padrão para as redes neurais com e sem filtros de banda estreita.

Modelo	$\bar{O}(H_c\|\mathbf{x})$	$\hat{S}_O$
Sem narrow bands	0.1081	0.0008
Com narrow bands	0.2542	0.0027

Embora não tenhamos visto diferenças significativas na estimação pontual dessas amostras, a Figura 5.11 apresenta uma pequena melhoria em relação aos erros de predição, observando-se uma diminuição de pontos distantes da reta em azul quando comparamos o segundo gráfico com o primeiro. Isso vai de encontro com a diminuição do σ_{RMSE} um pouco mais relevante do que a diminuição constatada no FlexCoDE (Tabelas 5.4 e 5.8).

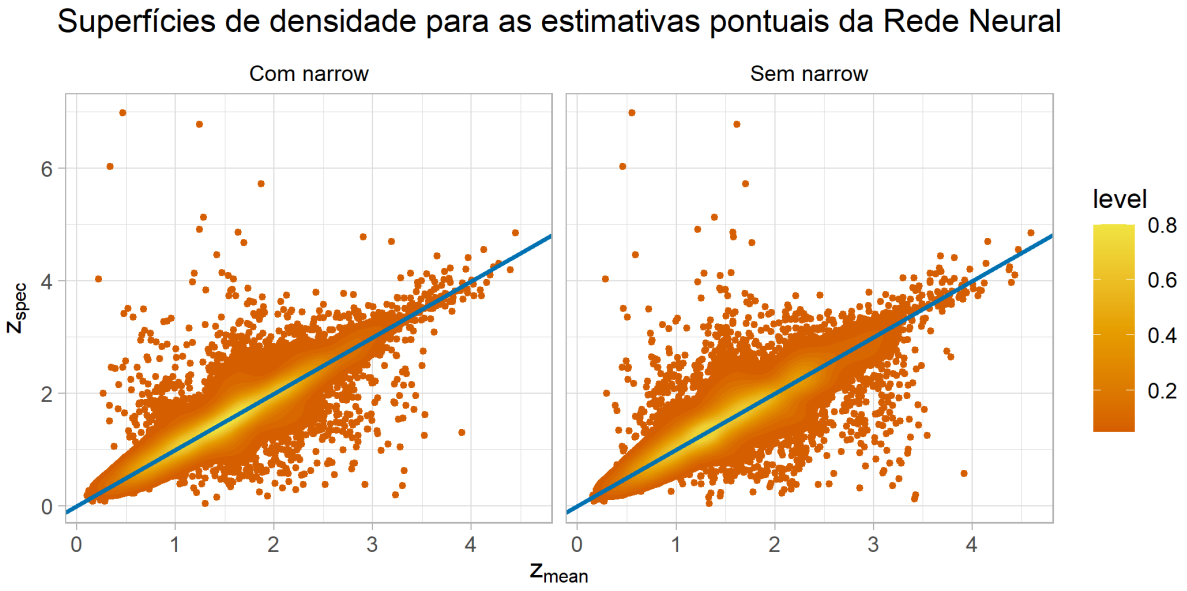

Figura 5.11: Mapa de densidade de pontos dos *redshifts* espectroscópicos versus valores médios preditos estimados. No lado esquerdo, temos os *redshifts* fotométricos estimados peloa rede neural com as *narrow bands* e do lado esquerdo os valores obtidos com a rede neural sem as *narrow bands*.

Pode-se notar que as barras de frequência do histograma na Figura 5.12 encontram-se um pouco mais próximas dos limites estabelecidos pela banda de confiança da distribuição uniforme padrão, porém ele também apresenta extremidades com frequências um pouco mais elevadas, dando pequenos indícios de desvio da uniformidade dos *PIT values* na ausência das bandas estreitas.

Quando analisamos a distribuição dos valores de cobertura na rede neural com as *narrow bands* o cenário se intensifica ainda mais, considerando as barras mais altas próximas

de 0 e 1 conforme a Figura 5.13. Ou seja, ambos os modelos de rede neural parecem não estarem conseguindo estimar bem a distribuição condicional do *redshift* fotométrico dadas suas covariáveis.

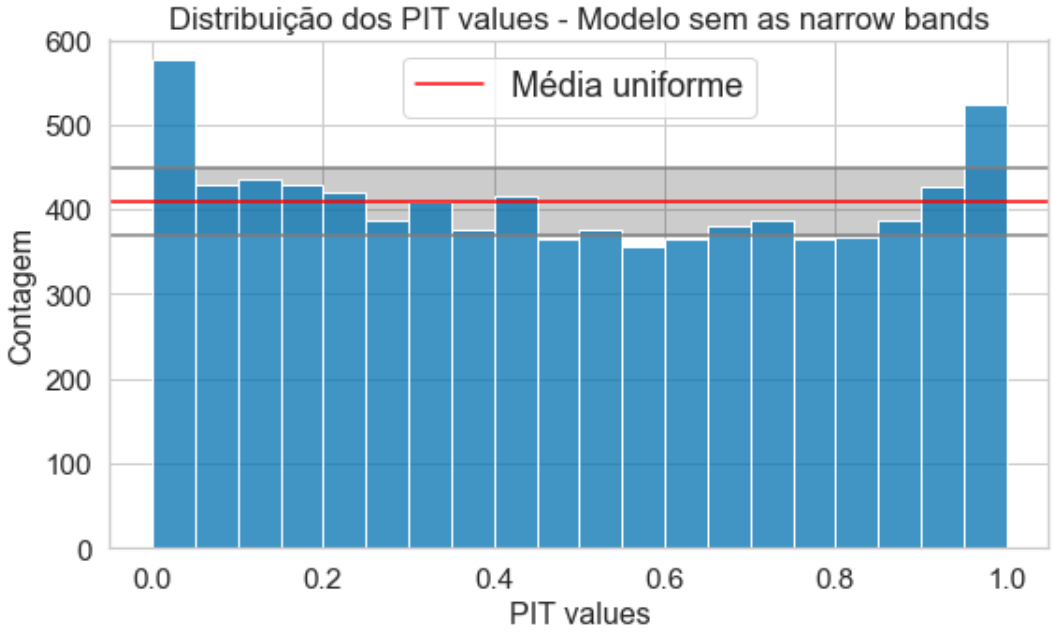

Figura 5.12: Histograma dos valores PIT baseados na densidade condicional estimada da rede neural sem as *narrow bands*.

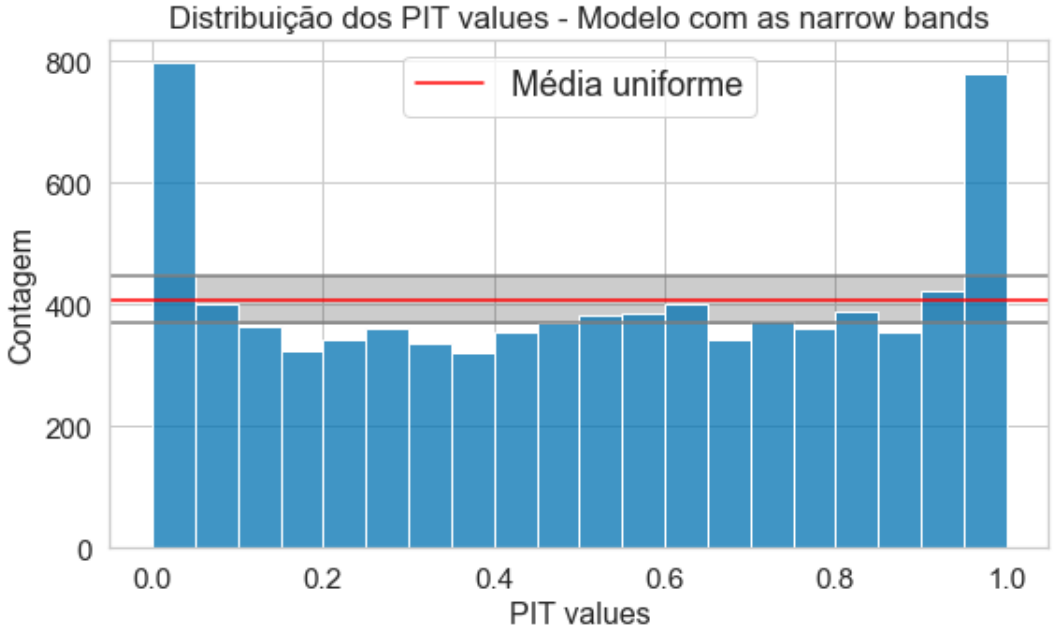

Figura 5.13: Histograma dos valores PIT baseados na densidade condicional estimada da rede neural com as *narrow bands*.

De maneira análoga, observa-se que as caudas do *P-P Plot* na Figura 5.15 tendem a ser mais pesadas na presença das bandas estreitas do que observado na Figura 5.14, em

que esta última não parece se distanciar tanto da reta de uniformidade.

Realizando-se o teste de Kolmogorov-Smirnov para verificação da hipótese de que os valores PIT seguem uma distribuição uniforme padrão, obtemos os p-valores de 2.2×10^{-16} e 2.356×10^{-7} para as redes neurais com e sem *narrow bands*, respectivamente. Ou seja, o diagnóstico do modelo BMDN também parece não se adequar para os dados, assim como constatado no FlexCoDE.

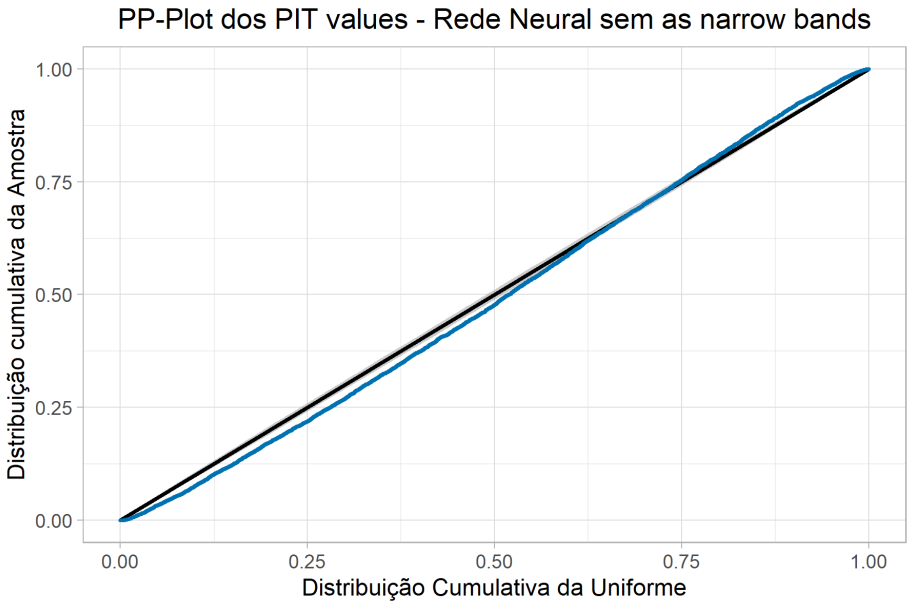

Figura 5.14: Gráfico de probabilidade-probabilidade dos *PIT values* da rede neural sem *narrow*.

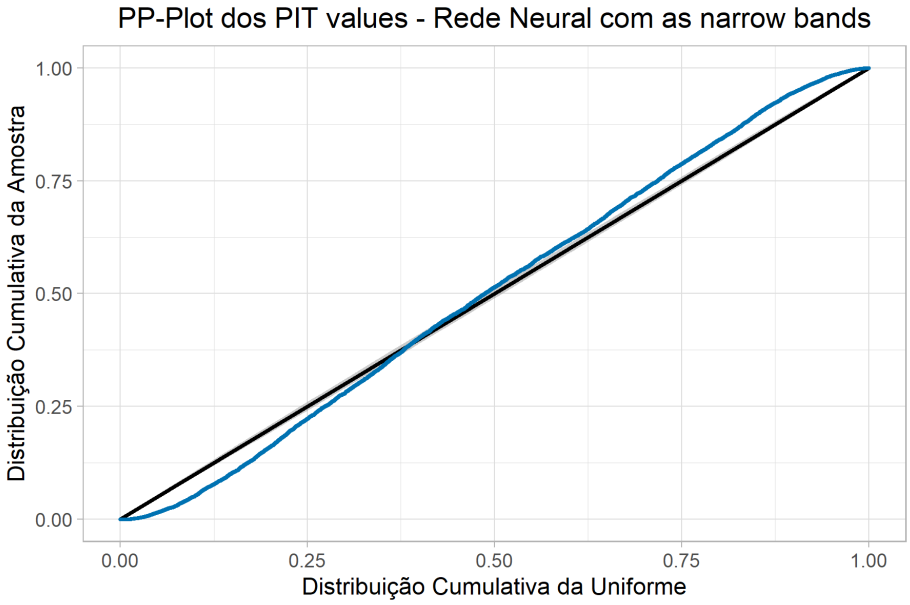

Figura 5.15: Gráfico de probabilidade-probabilidade dos *PIT values* da rede neural sem *narrow*.

Capítulo 6

Considerações finais

Este trabalho de conclusão de curso teve como finalidade aplicar métodos estatísticos para estimação da densidade condicional do *redshift* de quasares a partir de medidas fotométricas em dados observados pelo levantamento fotométrico *Southern Photometric Local Universe Survey*. Além disso, levando em conta a peculiaridade de seu sistema fotométrico composto por filtros de bandas estreitas, o estudo foi orientado à exploração e influência destas bandas na predição do *redshift*.

Considerando o ganho de informação acerca das características do *redshift* ao estimar sua função de densidade de probabilidade condicional em vez da esperança condicional, aplicamos o método não-paramétrico FlexCoDE para executar tal tarefa. O modelo foi empregado tanto na presença das cores criadas a partir das *narrow bands* quanto na ausência destas, com a finalidade de verificar a importância das bandas estreitas na estimação do *redshift* fotométrico.

A partir das diversas análises realizadas, constatamos que a informação das *narrow bands* trouxeram resultados promissores, diminuindo significativamente a função de perda estimada nos modelos do FlexCoDE. Observou-se que as curvas de densidade condicional estimadas do *redshift* incluindo as bandas estreitas apresentaram um comportamento mais concentrado e saliente do que as curvas sem essas bandas. Resultados análogos foram obtidos a partir das Redes Bayesianas de Densidade de Mistura, confirmando que a significativa melhoria da estimação do *redshift* fotométrico se dá principalmente pelas próprias *narrow bands*, e não apenas pelo método escolhido.

Entretanto, ambos os modelos ajustados não apresentaram um diagnóstico tão bom pela análise dos *PIT values*, dado que os valores de cobertura das densidades condicionais estimadas se desviaram da hipótese de uniformidade padrão. Embora essa diagnose

não tenha sido tão satisfatória, os frutos decorrentes deste trabalho ainda são válidos e extremamente relevantes, uma vez que tanto os resultados pelo algoritmo FlexCoDE quanto os resultados pelas Redes Bayesianas de Densidade de Mistura apontaram para a mesma conclusão acerca das *narrow bands*, mesmo sendo provenientes de fundamentações matemáticas tão distintas.

Assim, em futuras análises, podem ser investigadas outras metodologias de estimação de densidades condicionais com a finalidade de obter ajustes ainda mais precisos e com diagnósticos melhores. E, além disso, incentivar novas estratégias de pesquisa astronômica com a implementação de bandas estreitas, visto que elas trazem boas informações com respeito ao *redshift* fotométrico.

Referências Bibliográficas

Almeida-Fernandes, F., SamPedro, L., Herpich, F., Molino, A., Barbosa, C., Buzzo, M., Overzier, R., de Lima, E., Nakazono, L., Oliveira Schwarz, G. *et al.* (2022). Data release 2 of s-plus: Accurate template-fitting based photometry covering 1000 deg2 in 12 optical filters. *Monthly Notices of the Royal Astronomical Society*, **511**(3), 4590–4618.

Ball, N. M. e Brunner, R. J. (2010). Data mining and machine learning in astronomy. *International Journal of Modern Physics D*, **19**(07), 1049–1106.

Benitez, N. (2000). Bayesian photometric redshift estimation. *The Astrophysical Journal*, **536**(2), 571.

Bertin, E. e Arnouts, S. (1996). Sextractor: Software for source extraction. *Astronomy and astrophysics supplement series*, **117**(2), 393–404.

Britannica, E. (2021). "redshift". Disponível em: https://www.britannica.com/science/redshift. Acessado em: 12 jan. 2022.

Carroll, B. W. e Ostlie, D. A. (2017). *An introduction to modern astrophysics*. Cambridge University Press.

Dalmasso, N., Pospisil, T., Lee, A. B., Izbicki, R., Freeman, P. E. e Malz, A. I. (2020). Conditional density estimation tools in python and r with applications to photometric redshifts and likelihood-free cosmological inference. *Astronomy and Computing*, **30**, 100362.

DiPompeo, M., Bovy, J., Myers, A. e Lang, D. (2015). Quasar probabilities and redshifts from wise mid-ir through galex uv photometry. *Monthly Notices of the Royal Astronomical Society*, **452**(3), 3124–3138.

Fan, J., Yao, Q. e Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83**(1), 189–206.

Freeman, P. E., Izbicki, R. e Lee, A. B. (2017). A unified framework for constructing, tuning and assessing photometric redshift density estimates in a selection bias setting. *Monthly Notices of the Royal Astronomical Society*, **468**(4), 4556–4565.

Garabini, V. (2017). O uso de micro-ondas para o melhoramento de vias públicas. Disponível em: https://www.linkedin.com/pulse/o-uso-de-micro-ondas-para-melhoramento-vias-p%C3%BAblicas-veber-garabini?trk=public_profile_article_view. Acessado em: 11 ago. 2022.

Hubble, E. (1942). The problem of the expanding universe. *Science*, **95**(2461), 212–215.

IUPAC, C. o. C. T. (2019). "transmittance". Disponível em: https://goldbook.iupac.org/terms/view/T06484. Acessado em: 15 jan. 2022.

Izbicki, R. (2019). Flexcode. Disponível em: https://github.com/rizbicki/FlexCoDE. Acessado em: 12 fev. 2022.

Izbicki, R. e dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. ISBN 978-65-00-02410-4.

Izbicki, R. e Lee, A. B. (2016). Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, **25**(4), 1297–1316.

Izbicki, R. e Lee, A. B. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, **11**(2), 2800–2831.

Izbicki, R., Lee, A. B. e Freeman, P. E. (2017). Photo-z estimation: An example of nonparametric conditional density estimation under selection bias. *The Annals of Applied Statistics*, **11**(2), 698–724.

Izbicki, R., Shimizu, G. e Stern, R. B. (2022). Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, **23**(87), 1–32.

Jiang, L., Fan, X., Bian, F., McGreer, I. D., Strauss, M. A., Annis, J., Buck, Z., Green, R., Hodge, J. A., Myers, A. D. *et al.* (2014). The sloan digital sky survey stripe 82 imaging data: Depth-optimized co-adds over 300 deg2 in five filters. *The Astrophysical Journal Supplement Series*, **213**(1), 12.

Kellermann, K. I. (2013). The discovery of quasars and its aftermath. *arXiv: History and Philosophy of Physics*.

Lima, E., Sodré Jr, L., Bom, C., Teixeira, G., Nakazono, L., Buzzo, M., Queiroz, C., Herpich, F., Castellon, J. N., Dantas, M. *et al.* (2022). Photometric redshifts for the s-plus survey: Is machine learning up to the task? *Astronomy and Computing*, **38**, 100510.

Lima, E. V. R. d. (2020). *Photometric redshifts for S-PLUS using machine learning techniques*. Tese de doutorado, Universidade de São Paulo.

Lyke, B. W., Higley, A. N., McLane, J., Schurhammer, D. P., Myers, A. D., Ross, A. J., Dawson, K., Chabanier, S., Martini, P., Des Bourboux, H. D. M. *et al.* (2020). The sloan digital sky survey quasar catalog: Sixteenth data release. *The Astrophysical Journal Supplement Series*, **250**(1), 8.

Mendes de Oliveira, C., Ribeiro, T., Schoenell, W., Kanaan, A., Overzier, R., Molino, A., Sampedro, L., Coelho, P., Barbosa, C., Cortesi, A. *et al.* (2019). The southern photometric local universe survey (s-plus): improved seds, morphologies, and redshifts with 12 optical filters. *Monthly Notices of the Royal Astronomical Society*, **489**(1), 241–267.

Nakazono, L., Mendes de Oliveira, C., Hirata, N., Jeram, S., Queiroz, C., Eikenberry, S. S., Gonzalez, A., Abramo, R., Overzier, R., Espadoto, M. *et al.* (2021). On the discovery of stars, quasars, and galaxies in the southern hemisphere with s-plus dr2. *Monthly Notices of the Royal Astronomical Society*, **507**(4), 5847–5868.

Oliveira Filho, K. d. S. e Saraiva, M. d. F. O. (2004). Astronomia e astrofísica. *São Paulo: Editora Livraria da Física*, **780**.

Pospisil, T. (2019a). cdetools: Tools for conditional density estimates. Disponível em: https://github.com/tpospisi/cdetools. Acessado em: 30 mar. 2022.

Pospisil, T. (2019b). Flexcode. Disponível em: https://github.com/tpospisi/FlexCode. Acessado em: 12 fev. 2022.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. *Multivariate analysis II*, **25**, 31.

Schmidt, S., Malz, A., Soo, J., Almosallam, I., Brescia, M., Cavuoti, S., Cohen-Tanugi, J., Connolly, A., DeRose, J., Freeman, P. *et al.* (2020). Evaluation of probabilistic photometric redshift estimation approaches for the rubin observatory legacy survey of space and time (lsst). *Monthly Notices of the Royal Astronomical Society*, **499**(2), 1587–1606.

Steidel, C. C., Adelberger, K. L., Giavalisco, M., Dickinson, M. e Pettini, M. (1999). Lyman-break galaxies at $z > 4$ and the evolution of the ultraviolet luminosity density at high redshift. *The astrophysical journal*, **519**(1), 1.

Takeuchi, I., Nomura, K. e Kanamori, T. (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Computation*, **21**(2), 533–559.

Yang, Q., Wu, X.-B., Fan, X., Jiang, L., McGreer, I., Green, R., Yang, J., Schindler, J.-T., Wang, F., Zuo, W. *et al.* (2017). Quasar photometric redshifts and candidate selection: A new algorithm based on optical and mid-infrared photometric data. *The Astronomical Journal*, **154**(6), 269.

Apêndice A

Códigos de Programação

Os códigos de programação desenvolvidos para as análises deste Trabalho de Graduação estão disponíveis no repositório *GitHub* e podem ser acessados aqui ou pelo link https://github.com/GabrielaPereiraSoares/Projeto_TCC.