Universidade de São Paulo
Instituto de Astronomia, Geofísica e Ciências Atmosféricas
Departamento de Astronomia

Vitor Martins Cernic

Síntese de populações de galáxias com técnicas de aprendizado de máquina

São Paulo
2020

Vitor Martins Cernic

Síntese de populações de galáxias com técnicas de aprendizado de máquina

Trabalho de Conclusão de Curso apresentado ao Instituto de Astronomia, Geofisica e Ciências Atmosféricas da Universidade de São Paulo como requisito parcial para a obtenção do título de Bacharel em Astronomia.

Vertente: Computação
Orientador: Prof. Dr. Laerte Sodré Júnior

São Paulo
2020

À todos aqueles que compartilharam essa jornada comigo, tais cujos nomes e experiências convividas estarão sempre presentes em mim.

Agradecimentos

À minha família;

À Milena, por sempre estar ao meu lado;

Aos meus colegas de infância, que sempre me apoiaram;

Ao meu orientador Laerte, por ser quem eu mais pude contar com suporte dentro da Universidade;

À todos do meu círculo de pesquisa, em especial ao Ariel e ao Erik, por serem algumas das pessoas mais pacientes e de bom coração que me inspiraram durante a graduação;

Aos colegas Tajan e Fabrícia, por me ajudarem nos momentos difíceis da graduação;

À FAPESP, pelo apoio financeiro, sob o projeto n^o: 2019/05355-8.

Esta tese/dissertação foi escrita em LaTeX com a classe IAGTESE, para teses e dissertações do IAG.

"*The cake is a lie.*"

GLaDOS

Resumo

Este projeto é dividido em duas partes, com cada uma utilizando diferentes bases de dados para estudos e aplicações.

Na primeira parte geramos uma conjunto de treinamento de tamanho 231643 objetos $\times$ 100 *redshifts* $\times$ (12 bandas fotométricas + 1 valor de *redshift*), com cada um desses objetos relacionado a um parâmetro de metalicidade ou idade média, que, futuramente, deve treinar uma rede neural de *Deep Learning* (DL). Para isso utilizamos os resultados de uma nova aplicação do *software* STARLIGHT a uma amostra de mais de 200 mil galáxias realizada por Werle (2019). Estes resultados inovam pois a síntese espectral é feita levando em conta a emissão no UV das galáxias, o que melhora a estimativa das propriedades de populações estelares. Por conta da natureza desses dados, adicionamos as linhas espectrais faltantes aos espectros utilizando o *software* DOBBY, Flórido (2018). Estes espectros passam por um processo de *data augmentation* onde colocamos cada objeto em $N = 100$ *redshifts* diferentes entre 0.01 e 1. Calculamos finalmente a fotometria desses objetos artificiais com base nos 12 filtros do levantamento (*survey*) fotométrico S-PLUS.

Para a segunda parte do projeto, treinamos uma rede neural utilizando como base de dados as galáxias contempladas pelo S-PLUS que também fazem parte da aplicação do STARLIGHT à uma amostra do *Sloan Digital Sky Survey* (SDSS) de Cid Fernandes et al. (2005). Utilizando uma validação *holdout*, a rede apresenta resultados promissores para estimativa do `t_flux` (a idade média da populacão estelar ponderada em fluxo) ao treinar por 100 épocas. Depois de 100 épocas a rede começa a mostrar sinais de *overfitting*. Treinamos também a rede utilizando um método de validação conhecido por *k-fold* que, apesar de mais robusto, obteve resultados piores, retornando um valor de erro médio absoluto (MAE) de 0.320 após 100 épocas.

Abstract

This project is divided into two parts, with each one using different databases for our studies.

In the first part, we created a training set of shape 231643 objects $\times$ 100 redshifts $\times$ (12 photometric filters + 1 redshift value), with each of those objects related to a stellar population's parameters like metallicity or average stellar age. This dataset will be used, in the future, to train a Deep Learning (DL) neural network. To accomplish this, we utilized the results from a new application of the STARLIGHT software to a sample of more than 200 thousand galaxies by Werle (2019). These results are innovative because the spectral synthesis was done by taking into account the UV emission of the galaxies, which further improves the estimate of the stellar population's parameters. Because of the nature of this data, we added the missing spectral lines using the DOBBY software, Flórido (2018). The spectra go through a data augmentation process, creating artificial spectrums for $N = 100$ redshifts between 0.01 and 1. Finally, we took the photometry of each artificial object based on the 12 S-PLUS filters.

For the second part of the project, we trained a neural network with a database consisting of all galaxies observed by the S-PLUS that were also part of a STARLIGHT application to a sample from the Sloan Digital Sky Survey (SDSS) by Cid Fernandes et al. (2005). The algorithm shows promising results using holdout validation when estimating `t_flux` (the stellar population's average age weighted by flux) for a 100 epochs training. After that period, the network starts showing signs of overfitting. We also trained the neural network applying a validation method known as k-fold cross-validation that, while normally more robust, obtained worse results, providing a mean average error (MAE) of 0.320 after 100 epochs.

Lista de Figuras

1.1 Uma esquematização geral de como funciona um algoritmo de *Machine Learning*. 21

2.1 Resultados da análise de Werle (2019) para três objetos diferentes, um em cada linha, mostrando o fluxo normalizado $F_\lambda/F_{\lambda 0}$ por comprimento de onda λ. Em azul está o espectro sintetizado combinando-se a espectroscopia óptica do SDSS com a fotometria do Galex. Em vermelho está a extrapolação para o ultravioleta do ajuste obtido usando-se apenas o espectro óptico. Os fluxos das bandas do Galex são representados por círculos azuis. Os círculos laranja marcam o fluxo esperado no UV para este modelo. Créditos: Werle (2019) 24

2.2 Na figura de cima temos o espectro sem linhas de emissão (somente espectro estelar) da amostra de Werle (2019). Na figura de baixo vemos o espectro com as linhas de emissão adicionadas (espectro estelar mais espectro nebular) do ajuste do DOBBY. 26

2.3 Gráfico de fluxo normalizado por comprimento de onda (*wavelength*) de um espectro. Em azul temos um espectro direto da amostra de Werle (2019) em um redshift no referencial de repouso ($z = 0$) enquanto em vermelho temos esse mesmo espectro em $z = 0.15$ por meio da conta da Equação 2.1. . . . 27

2.4 Os 12 filtros do S-PLUS em gráfico de comprimento de onda λ por eficiência $E(\lambda)$. Cada filtro é colorido de acordo com a legenda da direita. Créditos: Mendes de Oliveira et al. (2019). 28

2.5 Cálculos de fluxo normalizado F_λ por comprimento de onda λ. As linhas vermelhas representam o espectro enquanto os pontos azuis representam a fotometria calculada para cada filtro. A posição das abscissas dos pontos é dada pelo λ_{eff}. Na figura de cima, o espectro está em redshift de repouso ($z = 0$) enquanto na figura de baixo temos o mesmo espectro em $z = 0.15$. 29

2.6 Representação visual do cubo de dados gerado. Do lado esquerdo temos os dados de entrada (*inputs*), onde cada um dos 231643 objetos têm seu espectro posto em 100 *redshifts* diferentes e para cada um desses novos espectros é calculado sua fotometria com base nos 12 filtros do S-PLUS junto com um valor de *redshift*. Do lado direito estão os dados alvos (*targets*), onde cada espectro está relacionado a um parâmetro de metalicidade ou idade média. 30

3.1 Arquitetura da rede, construída com uma camada de entrada à esquerda de tamanho 13 (12 magnitudes fotométricas + 1 *redshift*), seguida por duas camadas escondidas (*hidden layers*) de 128 neurônios cada uma. Por fim, uma camada de saída de tamanho 1 relacionada a um parâmetro de população estelar. 32

3.2 Gráfico de erro médio absoluto (MAE) ao aplicar à rede o conjunto de validação para cada uma das 1500 épocas (figura de cima). A figura de baixo representa o mesmo conjunto de dados porém vizualizando até 100 épocas. O método de validação utilizado foi o *holdout*. 33

3.3 Gráfico de erro médio absoluto (MAE) ao aplicar à rede o conjunto de validação para cada uma das 200 épocas. O método de validação utilizado foi o *K-Fold Cross-Validation*. 34

Lista de Tabelas

1.1 Nome do filtro, comprimento de onda efetivo λ_{eff} e intervalo de comprimento de onda $\Delta\lambda$ para cada um dos 12 filtros do S-PLUS. A coluna "Comment" traz informações adicionais sobre a relação dos filtros com outros *surveys* ou quais linhas espectrais (em baixos *redshifts*) as bandas estreitas pretendem observar. 20

Sumário

1. Introdução 19
1.1 Motivação 19
1.2 *Machine Learning* e a importância da base de dados 21

2. A amostra de dados de Werle (2019) 23
2.1 Estudo dirigido 25
2.2 Adição de linhas espectrais 25
2.3 Espectro em diferentes *redshifts* 26
2.4 Cálculo da fotometria de cada espectro 27
2.5 O cubo de dados 30

3. Dados de Cid Fernandes et al. (2005) 31
3.1 A arquitetura da rede neural 31
3.2 Treinamento da rede 32

4. Conclusões 35

Referências 37

Apêndice 39

A. Descrição dos levantamentos fotométricos 41

Capítulo 1

Introdução

1.1 Motivação

Com a globalização e o avanço da tecnologia nos últimos anos, o tempo onde a humanidade era capaz de estudar todo o conteúdo produzido em uma única área do conhecimento há muito se fora. A área de *data science* cresce como nenhuma outra, e instituições ao redor do mundo buscam maneiras eficientes de tratar seus grandes conjuntos de dados. Vivemos hoje no período do *Big Data*, uma época onde o avanço tecnológico e científico da sociedade estão atrelados à eficiência dos modelos que analisam as informações de forma massiva, ao invés de estudar os dados separadamente. Mas não se deve deixar enganar pensando que esse saturamento de informações é algo ruim para a humanidade. Grandes conjuntos de dados, a partir de análises estatísticas robustas, revelam informações sobre eventos que a mente humana nunca foi capaz de compreender completamente, e relações que antes pareciam abstratas começam a tomar formas mais concretas. Com *data science* conseguimos explorar as relações desde tarefas que parecem mundanas, como diferenciar um gato de um cachorro, até tarefas que pareciam nunca antes possíveis, como obter o *redshift* dada uma fotometria para milhares/milhões de galáxias.

É inevitável que a Astronomia, como a maioria das outras áreas da ciência, seja afetada pelo fenômeno da avalanche de dados. Conforme o avanço na tecnologia de telescópios aumenta, o volume de dados gerados cresce de maneira significativa, fazendo com que a quantidade de informações que eram produzidas por levantamentos inteiros (*surveys*) há uma década sejam fabricadas hoje em uma noite. Esse gigantesco fluxo atual é discutido por Kremer et al. (2017), onde se nota que a tendência para o futuro é que esse volume só aumente, devido aos levantamentos em andamento, como o *Javalambre Physics of the*

Tabela 1.1 - Nome do filtro, comprimento de onda efetivo λ_{eff} e intervalo de comprimento de onda $\Delta\lambda$ para cada um dos 12 filtros do S-PLUS. A coluna "Comment" traz informações adicionais sobre a relação dos filtros com outros *surveys* ou quais linhas espectrais (em baixos *redshifts*) as bandas estreitas pretendem observar.

Filter name	λ_{eff} [Å]	$\Delta\lambda$ [Å]	Comment
uJava	3574	330	Javalambre *u*
J0378	3771	151	[O II]
J0395	3941	103	Ca H+K
J0410	4094	201	Hδ
J0430	4292	200	G-band
gSDSS	4756	1536	SDSS-like g
J0515	5133	207	Mgb Triplet
rSDSS	6260	1462	SDSS-like r
J0660	6614	147	Hα
iSDSS	7692	1504	SDSS-like i
J0861	8611	408	Ca Triplet
zSDSS	8783	1072	SDSS-like z

Accelerating Universe Astrophysical Survey (J-PAS), *Javalambre-Photometric Local Universe Survey* (J-PLUS) e *Southern Photometric Local Universe Survey* (S-PLUS), ou aos esperados para um futuro próximo, como o *Large Synoptic Survey Telescope* (LSST) e o *Euclid*. Com isso, é necessário a criação de novas ferramentas para que esses dados consigam ser compreendidos pelos astrônomos, já que esses levantamentos são as bases para abordar as grandes perguntas da astronomia atual: "do que é feito o Universo?"; "o que é a misteriosa matéria escura?"; "como a Via Láctea se formou?"; entre muitas outras.

Neste projeto iniciamos a criação de uma ferramenta baseada em aprendizado de máquina (*machine learning*, ML) para extrair alguns parâmetros de populações estelares como metalicidade e idade média a partir das fotometrias e *redshift* de alguns dos levantamentos fotométricos mais importantes atualmente: o J-PAS, J-PLUS e S-PLUS.

Os procedimentos que iremos desenvolver serão aplicados inicialmente aos objetos do Universo local ($z < 0.1$) observados pelo levantamento S-PLUS. Diferente dos levantamentos fotométricos mais comuns de 5 filtros, o S-PLUS e J-PLUS são *surveys* fotométricos com 12 filtros (7 bandas estreitas e 5 bandas largas *u, g, r, i, z*) descritos na Tabela 1.1. Enquanto as bandas largas contêm informações sobre o contínuo de um espectro, as bandas estreitas representam melhor regiões de linhas de emissão, e essas informações adicionais

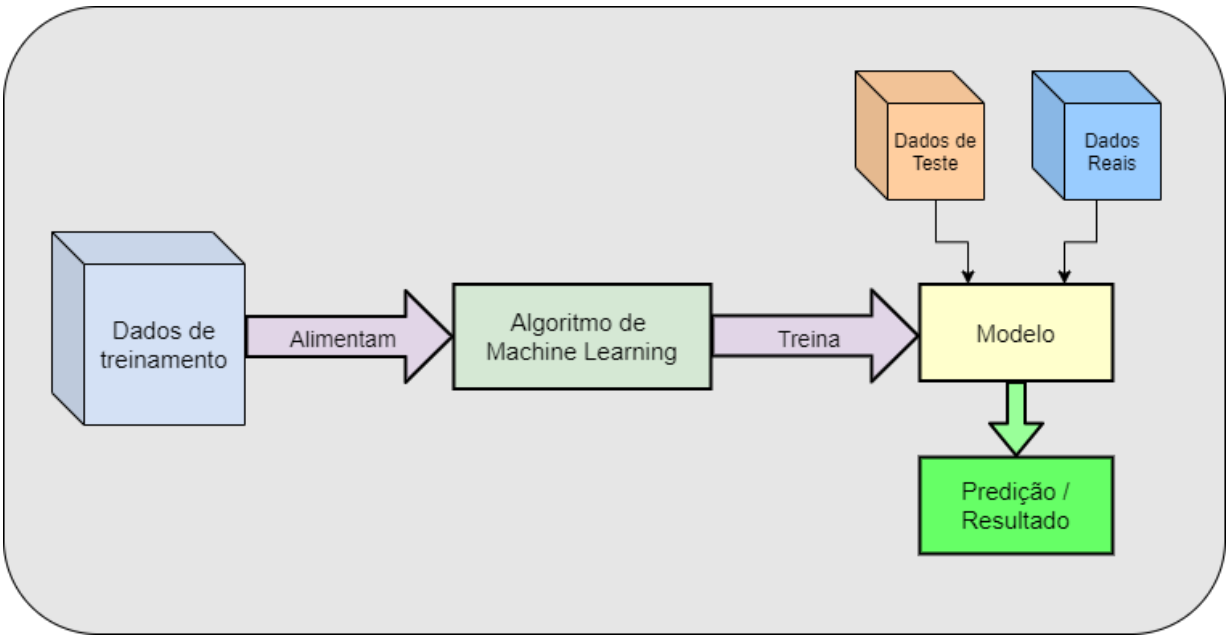

Figura 1.1: Uma esquematização geral de como funciona um algoritmo de *Machine Learning*.

sobre o espectro devem ajudar o algoritmo a fazer predições melhores. No Apêndice A apresentamos uma descrição mais detalhada desses três levantamentos.

1.2 *Machine Learning* e a importância da base de dados

A maior parte das aplicações de ML podem ser esquematizadas de uma forma geral como na Figura 1.1. Essas aplicações partem de uma extensa base de dados de treinamento que, ao ser fornecida para algum algoritmo de *machine learning*, treina um certo modelo. Esse modelo passa por diversos testes para estimar sua eficácia utilizando uma base de dados de validação, e quando atinge um resultado satisfatório podemos aplicar este modelo a dados reais.

Em nosso caso, como queremos achar a relação entre fotometria e parâmetros estelares, os dados (conjunto) de treinamento devem conter, de um lado, fotometrias, e do outro lado, propriedades de populações estelares. É possível perceber pela Figura 1.1 que essa base de dados é essencial para o funcionamento do algoritmo. Se os dados de treinamento não estiverem representando corretamente a relação que queremos estudar, ao aplicar os dados reais ao modelo não obteremos resultados satisfatórios. Isso significa que é preciso ter certeza de que as fotometrias e parâmetros obtidos são confiáveis.

Aplicamos nesse projeto uma subárea do *Machine Learning* chamada de *Deep Learning*

(DL). Fazendo o uso de camadas densas (com muitos neurônios) o DL consegue, na maioria dos casos, encontrar resultados melhores e mais acurados do que algoritmos genéricos de ML. Por conta de tecnologias como Keras por Chollet et al. (2015), TensorFlow, entre outras, a implementação de redes densas é trivial para cientistas que querem fazer o uso dessas ferramentas.

Dividimos esse projeto em duas partes, cada uma com bases de dados diferentes. A primeira parte a ser estudada, na Seção 2, é a síntese espectral da tese de doutorado de Ariel Werle, Werle (2019), cujos resultados utilizamos para criar um conjunto de treinamento que futuramente será utilizado para treinar uma rede neural. Para a segunda parte do projeto, estudamos os dados vindos da aplicação do *software* STARLIGHT por Cid Fernandes et al. (2005) às galáxias da região contemplada pelo *survey* S-PLUS. Nesta última parte, desenvolvida na Seção 3, utilizamos novos dados para treinar um algoritmo de *Deep Learning*.

Capítulo 2

A amostra de dados de Werle (2019)

Para a primeira parte desse projeto, desejamos criar um conjunto de treinamento com dados de ponta que será utilizado, futuramente, para treinar uma rede neural que consiga predizer, com uma certa precisão, os parâmetros das populações estelares de uma galáxia dada sua fotometria e seu *redshift*. Tomamos como ponto de partida as sínteses espectrais criadas por uma nova aplicação do *software* STARLIGHT a uma amostra de 231643 galáxias realizadas por Werle et al. (2018). O STARLIGHT é um programa que extrai propriedades físicas de galáxias ajustando seus espectros com uma combinação linear de espectros de populações estelares simples, publicado por Cid Fernandes et al. (2005). Nesta parte, serão utilizados os resultados da síntese espectral feita por Ariel Werle em sua tese de doutorado, Werle (2019). Werle usou uma versão modificada do STARLIGHT para gerar sínteses espectrais de galáxias, e sua implementação traz várias inovações em relação às sínteses originais. Em particular, combina os espectros do SDSS com dados fotométricos do Galaxy Evolution Explorer (GALEX) para uma amostra de 231643 galáxias, usando modelos do estado da arte para populações estelares. Estes resultados inovam pois a síntese é feita levando em conta a emissão no ultravioleta (UV) das galáxias, o que melhora a estimativa das propriedades das populações estelares.

Na Figura 2.1 vemos a comparação entre os dados obtidos por Werle (2019) com os espectros do SDSS e do GALEX para três objetos diferentes. Estes resultados ilustram o enorme ganho na modelagem de espectros com o STARLIGHT quando se inclui na análise, além do espectro óptico, a fotometria UV. A mudança significativa no ajuste no UV leva a uma melhor determinação das populações estelares, principalmente as associadas a estrelas jovens.

Como o objetivo do trabalho de Werle (2019) é a determinação das populações estelares

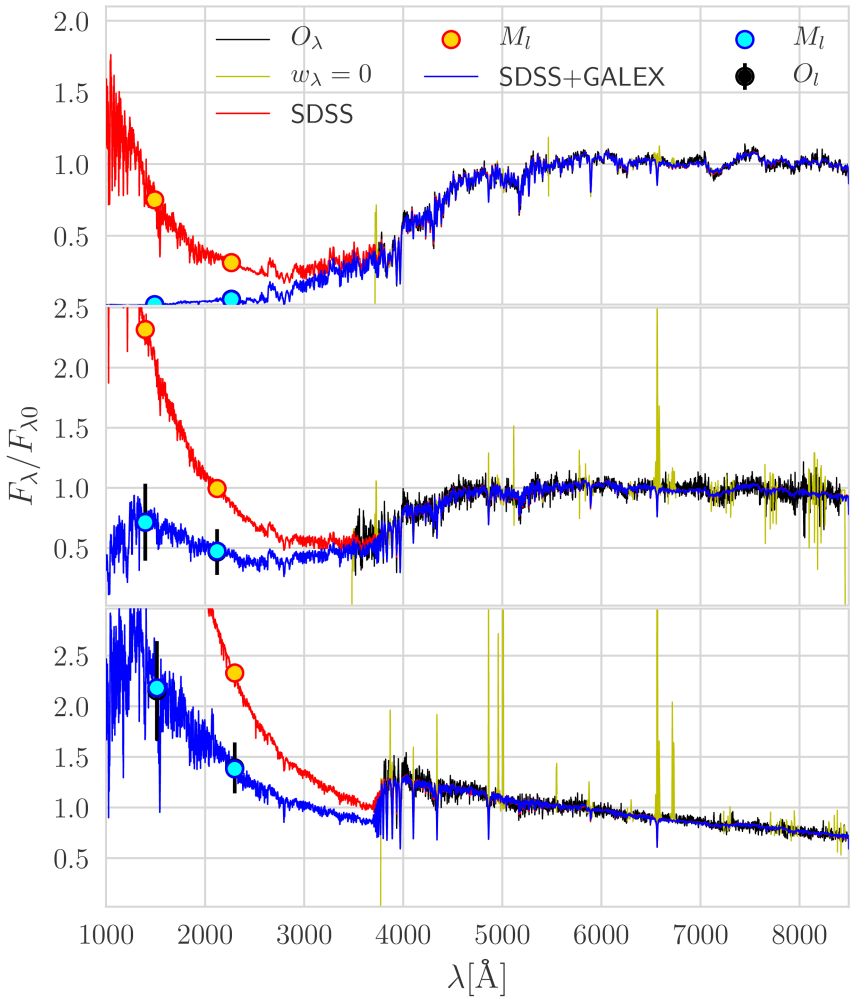

Figura 2.1: Resultados da análise de Werle (2019) para três objetos diferentes, um em cada linha, mostrando o fluxo normalizado $F_\lambda/F_{\lambda 0}$ por comprimento de onda λ. Em azul está o espectro sintetizado combinando-se a espectroscopia óptica do SDSS com a fotometria do Galex. Em vermelho está a extrapolação para o ultravioleta do ajuste obtido usando-se apenas o espectro óptico. Os fluxos das bandas do Galex são representados por círculos azuis. Os círculos laranja marcam o fluxo esperado no UV para este modelo. Créditos: Werle (2019)

das galáxias, apenas o contínuo dos espectros é ajustado, ou seja, as linhas de emissão são excluídas das análises pois são produzidas pelo gás ionizado, não pelas estrelas. O problema disso é que a inclusão das linhas espectrais é essencial para fotometria (principalmente no caso de bandas estreitas) e devemos considerá-las para criar modelos realísticos para o S-PLUS e outros levantamentos de dados.

Além de adicionar as linhas de emissão à amostra de Werle (2019), é necessário calcular a fotometria sintética dos espectros dada a natureza dos levantamentos fotométricos que queremos aplicar o algoritmo. Para isso, nos baseamos nos 12 filtros do S-PLUS definidos em Mendes de Oliveira et al. (2019), mostrados na Tabela 1.1. O cálculo dessa fotometria é discutido na Seção 2.4.

2.1 Estudo dirigido

Para iniciar o desenvolvimento desse conjunto de treinamento, o primeiro passo foi fazer um estudo sobre a os dados provenientes da pesquisa de Werle (2019). Auxiliado tanto pelo orientador Laerte Sodré quando pelo próprio Ariel Werle, isso significou estudar extensamente tanto a base de dados do S-PLUS quanto as sínteses espectrais de Werle (2019).

Foi criado também pelo professor Laerte Sodré Júnior um grupo de estudos semanal voltado para aprender os conceitos de inteligência artificial abordados no livro Deep Learning with Python, Chollet (2018).

Tomar esse tempo para entender os dados e ferramentas que estamos trabalhando foi fundamental para conseguir fazer as análises de tratamento dos dados das seções seguintes.

2.2 Adição de linhas espectrais

Para adicionar as linhas de emissão ao espectro de Werle (2019), utilizamos o *software* DOBBY, Flórido (2018). Este é um código que ajusta linhas de emissão assumindo que elas apresentam um perfil gaussiano. Entramos em contato com o grupo responsável pelo DOBBY para nos auxiliarem no projeto. Essa comunicação, facilitada pela familiaridade do Ariel Werle com este grupo, nos obteve, para cada um dos 231643 objetos da amostra, um novo espectro contendo as linhas de emissão. Um desses objetos está representado na Figura 2.2.

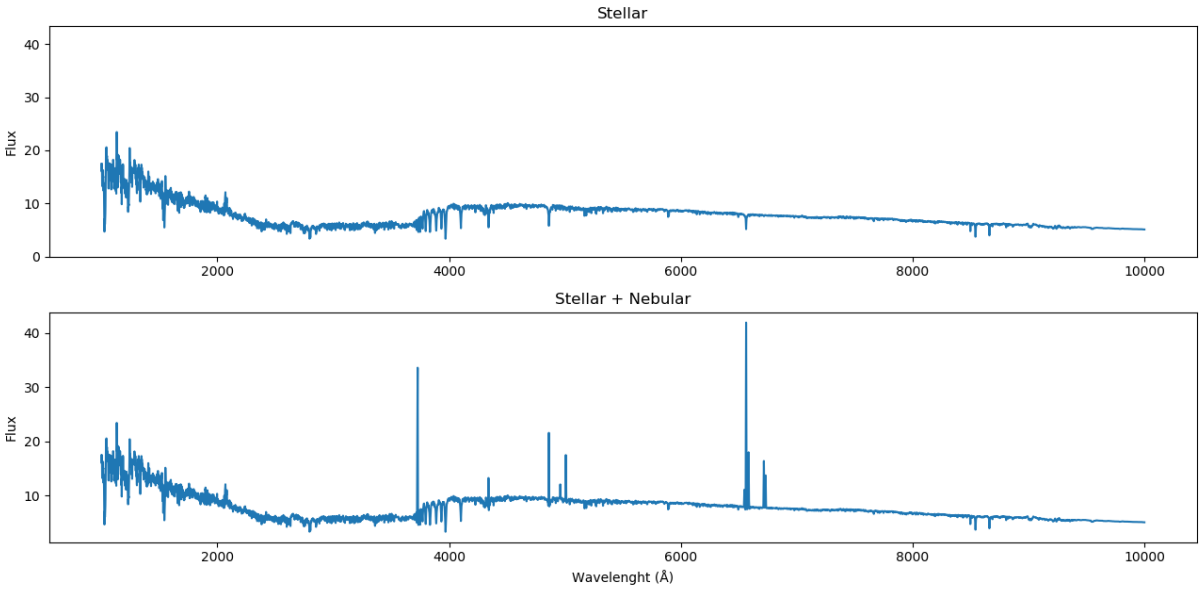

Figura 2.2: Na figura de cima temos o espectro sem linhas de emissão (somente espectro estelar) da amostra de Werle (2019). Na figura de baixo vemos o espectro com as linhas de emissão adicionadas (espectro estelar mais espectro nebular) do ajuste do DOBBY.

2.3 *Espectro em diferentes redshifts*

Dado um certo fóton com comprimento de onda λ_0 emitido por um objeto em referencial de repouso, o comprimento de onda λ do fóton observado se relaciona com λ_0 por um fator de $(1+z)$, onde z é o chamado *redshift*. Em termos matemáticos, definimos a Equação 2.1

$$\lambda = \lambda_0(1+z). \tag{2.1}$$

Uma das vantagems de utilizar a amostra de Werle (2019) é que os espectros são ajustados em *redshift* $z = 0$, ou seja, no referencial de repouso. Isso significa que como conhecemos λ_0, podemos utilizar a Equação 2.1 para aplicar a esse espectro um valor de z e calcular qual seria o espectro desse objeto em diferentes redshifts. Na Figura 2.3 vemos um dos objetos da amostra representado tanto em $z = 0$ quanto em $z = 0.15$.

A ideia é que aplicando diferentes valores de z para um espectro, podemos criar diversos objetos artificiais. Essa técnica é conhecida por *data augmentation*, e é muito utilizada em aplicações de *Machine Learning* para aumentar o conjunto de treinamento do algoritmo. Dessa forma, colocando cada espectro em N *redshifts* diferentes, podemos criar um conjunto de treinamento com $231643 \times N$ objetos artificiais.

Decidimos definir $N = 100$, escolhendo z a partir de uma distribuição uniforme entre

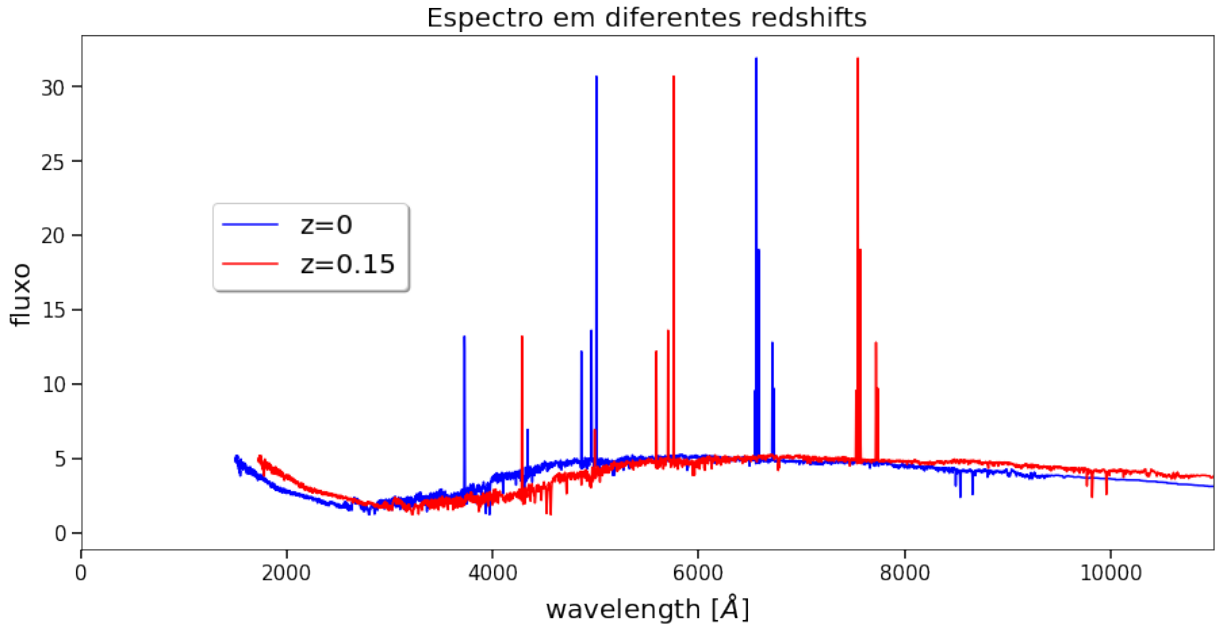

Figura 2.3: Gráfico de fluxo normalizado por comprimento de onda (*wavelength*) de um espectro. Em azul temos um espectro direto da amostra de Werle (2019) em um redshift no referencial de repouso ($z = 0$) enquanto em vermelho temos esse mesmo espectro em $z = 0.15$ por meio da conta da Equação 2.1.

0.01 e 1. A escolha do intervalo genérico de *redshifts* vêm do fato de ser onde existe a maior concentração de objetos no S-PLUS. Foi discutido a possibilidade de tentar preservar a distribuição de redshifts do S-PLUS no conjunto de treinamento, fazendo um estudo mais complexo sobre a distribuição dos objetos do levantamento. Essa possibilidade poderá ser considerada futuramente, dependendo dos resultados dos algoritmos que utilizarem essa base de dados.

2.4 Cálculo da fotometria de cada espectro

A fotometria dos espectros depende de quais filtros são escolhidos para serem aplicados. Ela é calculada a partir do fluxo $F(\lambda)$ do espectro, que depende do desenho do filtro e fatores físicos do telescópio. Além disso, é necessário conhecer também a função de eficiência $E(\lambda)$ de cada filtro. Essa eficiência vem da multiplicação entre a média da curva de transmissão do filtro, da transmissão atmosférica, da eficiência do CCD, e das curvas de reflexividade do espelho primário. As funções de eficiência de cada filtro representados pela Figura 2.4 são fornecidas diretamente pelos *websites* de cada um dos levantamentos. Neste caso, mostramos as curvas de eficiência para o S-PLUS.

A magnitude m_{filtro} de um certo filtro pode ser calculada a partir da Equação 2.2

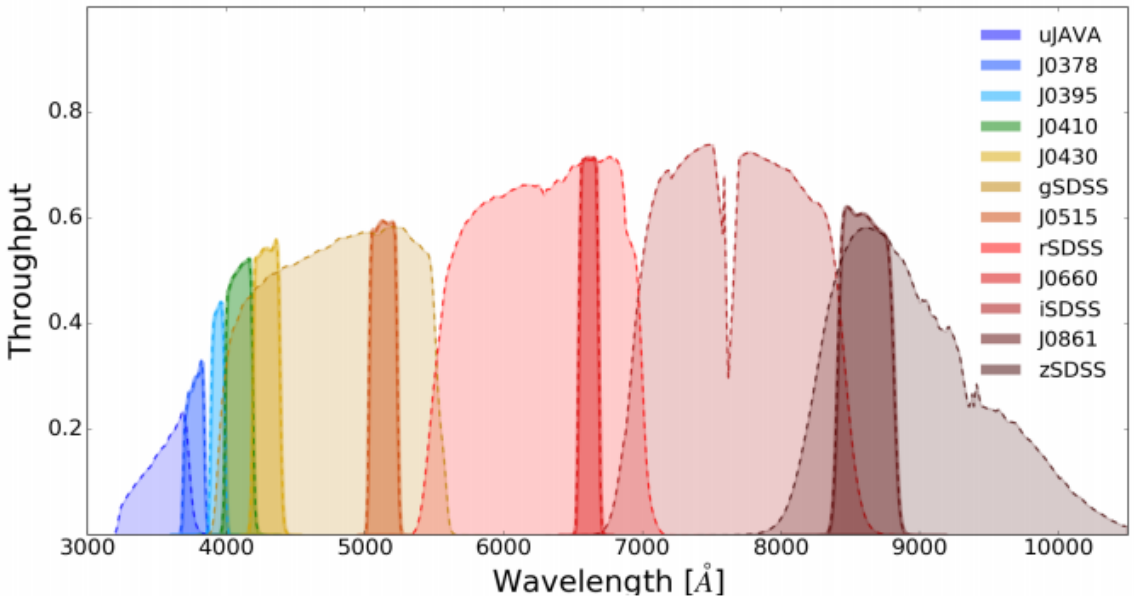

Figura 2.4: Os 12 filtros do S-PLUS em gráfico de comprimento de onda λ por eficiência $E(\lambda)$. Cada filtro é colorido de acordo com a legenda da direita. Créditos: Mendes de Oliveira et al. (2019).

$$m_{filtro} = \frac{\int \lambda E(\lambda) F(\lambda) d\lambda}{\int \lambda E(\lambda) d\lambda}. \tag{2.2}$$

Podemos também calcular o comprimento de onda efetivo λ_{eff} (ou comprimento de onda central)

$$\lambda_{eff} = \sqrt{\frac{\int \lambda E(\lambda) d\lambda}{\int E(\lambda)/\lambda d\lambda}}. \tag{2.3}$$

Mais detalhes dessas definições encontram-se no manual do Synphot, STScI development Team (2018).

Computamos as Equações 2.2 e 2.3 para calcular a fotometria de cada espectro sintético com base nas 12 fotometrias do S-PLUS. Utilizando integrais numéricas em python pela função `quad` da biblioteca `scipy.integrate`, obtivemos como resultado desse código a Figura 2.5.

Com o código criado, calculamos a fotometria para todos os espectros gerados e finalizamos as análises com esta base de dados do projeto.

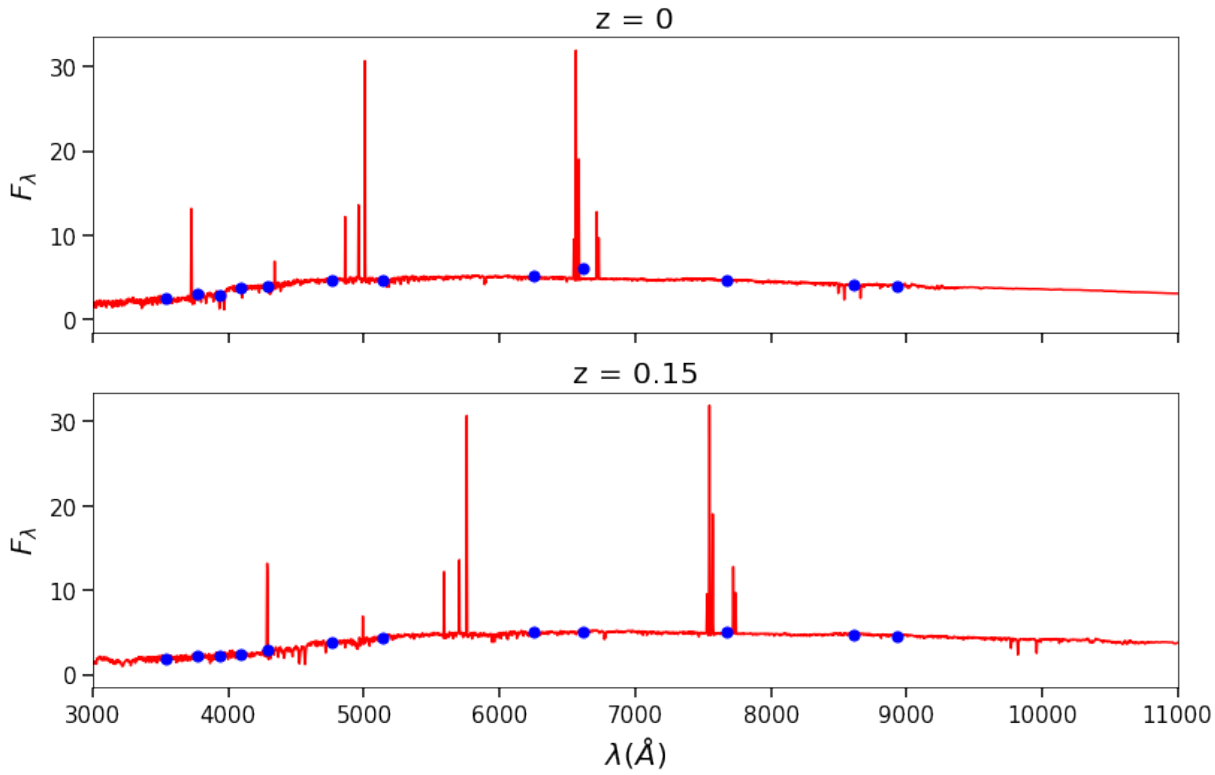

Figura 2.5: Cálculos de fluxo normalizado F_λ por comprimento de onda λ. As linhas vermelhas representam o espectro enquanto os pontos azuis representam a fotometria calculada para cada filtro. A posição das abscissas dos pontos é dada pelo λ_{eff}. Na figura de cima, o espectro está em redshift de repouso ($z = 0$) enquanto na figura de baixo temos o mesmo espectro em $z = 0.15$.

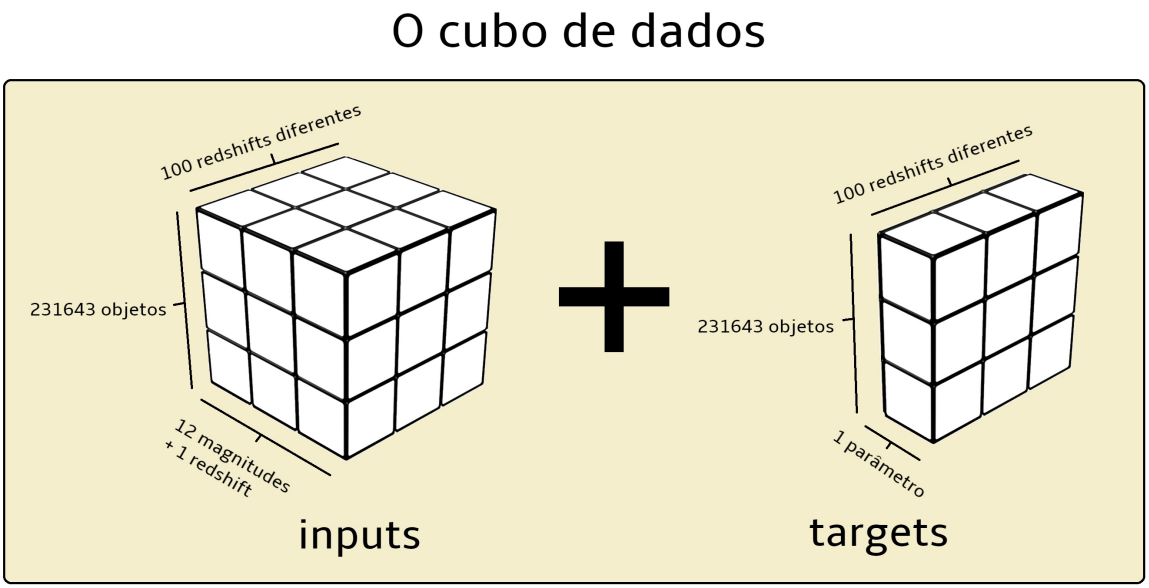

Figura 2.6: Representação visual do cubo de dados gerado. Do lado esquerdo temos os dados de entrada (*inputs*), onde cada um dos 231643 objetos têm seu espectro posto em 100 *redshifts* diferentes e para cada um desses novos espectros é calculado sua fotometria com base nos 12 filtros do S-PLUS junto com um valor de *redshift*. Do lado direito estão os dados alvos (*targets*), onde cada espectro está relacionado a um parâmetro de metalicidade ou idade média.

2.5 O cubo de dados

O conjunto de dados criado consiste em um "cubo" de tamanho 231643 objetos $\times$ $N = 100$ redshifts $\times$ (12 bandas do S-PLUS + 1 *redshift*), representado pela Figura 2.6, com cada um desses objetos relacionado a um parâmetro de população estelar, tal como metalicidade média, idade média, entre outros.

Vale notar aqui que o *redshift* será utilizado como um dos valores de entrada, já que é um parâmetro que é facilmente obtido e deve ajudar no aprendizado da rede. No caso do cubo de dados criado, conhecemos muito bem tal valor de *redshift* por conta da Seção 2.3. Quando trabalharmos com dados reais, pretendemos utilizar algum *redshift* fotométrico.

A próxima parte do projeto, descrita no Capítulo 3, mostra nosso desenvolvimento em criar uma rede que treina em dados fotométricos para estimar parâmetros de populações estelares, porém utilizando outra base de dados. Esperamos que futuramente esses resultados aqui criados (o cubo de dados) sirvam, da mesma forma, como um conjunto de treinamento para uma rede semelhante.

Capítulo 3

Dados de Cid Fernandes et al. (2005)

Para a segunda parte do projeto, treinamos uma rede neural com os dados de Cid Fernandes et al. (2005). Esses dados são uma aplicação do STARLIGHT às galáxias do SDSS, nos fornecendo valores de parâmetros como metalicidade e idade média para todos os objetos desse levantamento. Realizando um *cross-match* entre as galáxias contempladas pelo S-PLUS e a as galáxias utilizadas na aplicação no Cid Fernandes et al. (2005), obtemos um conjunto de treinamento com 55803 objetos, com suas fotometrias oferecidas pelo S-PLUS, com seus *redshifts* espectroscópicos vindos do SDSS e seus parâmetros estelares fornecidos pelo STARLIGHT.

Notamos que o conjunto de treinamento utilizado nessa etapa do projeto não é o mesmo que o criado no Capítulo 2. Os resultados mostrados aqui são preliminares e ainda estão em desenvolvimento. Pretendemos, futuramente, aplicar os mesmos passos mostrados nas próximas seções à base de dados criada no Capítulo 2.

3.1 A arquitetura da rede neural

Dado a fotometria nas 12 bandas do S-PLUS de um objeto e seu *redshift*, esperamos que a rede forneça um único valor correspondente ao parâmetro que queira ser estudado. Decidimos então criar um algoritmo de regressão por se encaixar nessas condições impostas. Escolhemos por utilizar *Deep Learning* dados seus resultados recentes terem sido mais robustos que outros algoritmos de *Machine Learning*. Para as análises apresentadas, vamos estimar o valor do parâmetro `t_flux` (a idade média da populacão estelar ponderada em fluxo).

O pacote Keras, Chollet et al. (2015), foi escolhido como biblioteca utlizada para esse

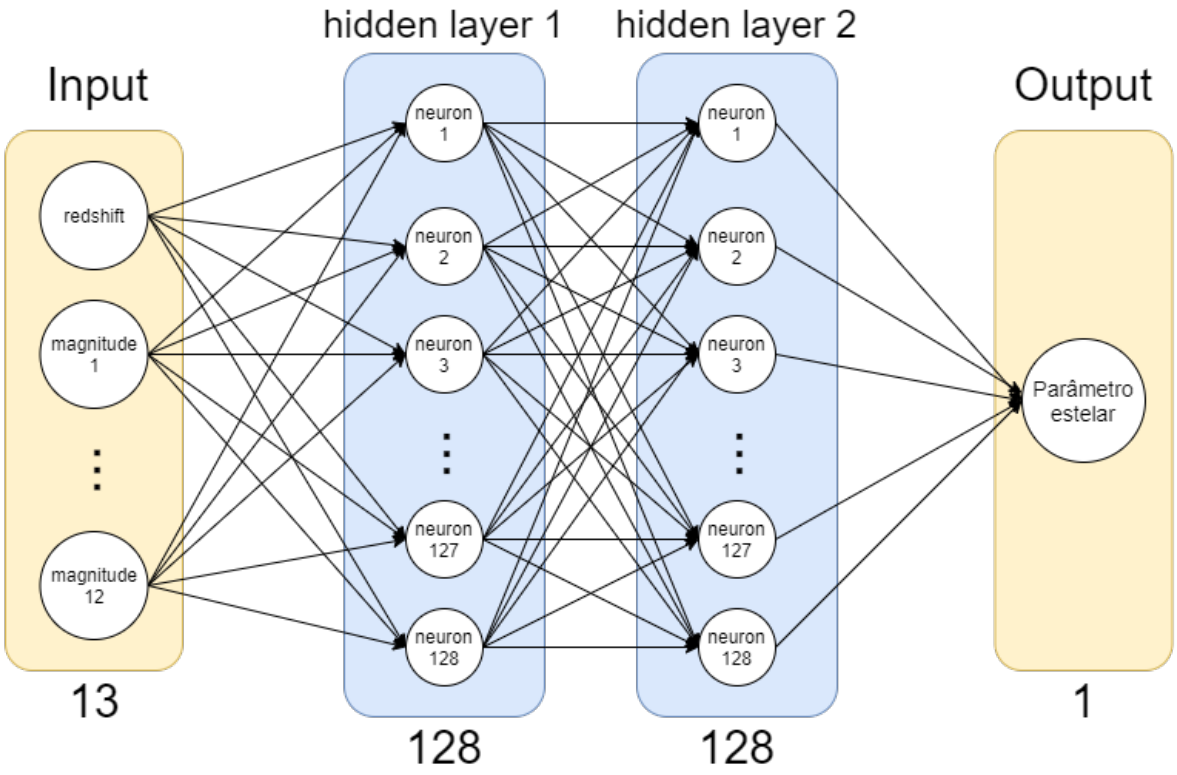

Figura 3.1: Arquitetura da rede, construída com uma camada de entrada à esquerda de tamanho 13 (12 magnitudes fotométricas + 1 *redshift*), seguida por duas camadas escondidas (*hidden layers*) de 128 neurônios cada uma. Por fim, uma camada de saída de tamanho 1 relacionada a um parâmetro de população estelar.

projeto. A arquitetura da rede foi construída com uma camada de entrada de tamanho 13 (as 12 fotometrias + 1 valor de *redshift*) seguida por duas camadas densas com 128 neurônios cada uma, com uma função de ativação linear retificada (ReLU). Por fim, a rede têm uma última camada de saída de tamanho 1. Utilizamos uma otimização `rmsprop` e uma *loss* computada pelo erro médio absoluto (MAE). A Figura 3.1 mostra uma representação visual da arquitetura.

Como explicado na Seção 2.5, utilizamos um valor de *redshift* como um dos parâmetros de entrada já que esse é facilmente obtido em dados reais e deve ser um fator importante para ajudar a rede a treinar. Aqui utilizamos um valor de *redshift* espectroscópico (z_{spec}), fornecido pelo SDSS, ao invés dos *redshifts* fotométricos (z_{phot}) fornecidos pelo S-PLUS.

3.2 Treinamento da rede

Aplicamos à base de dados um processo de *scaling*, onde restringimos tanto os valores de entrada (12 magnitudes fotométricas e 1 *redshift*) quanto os valores de saídas (parâmetros

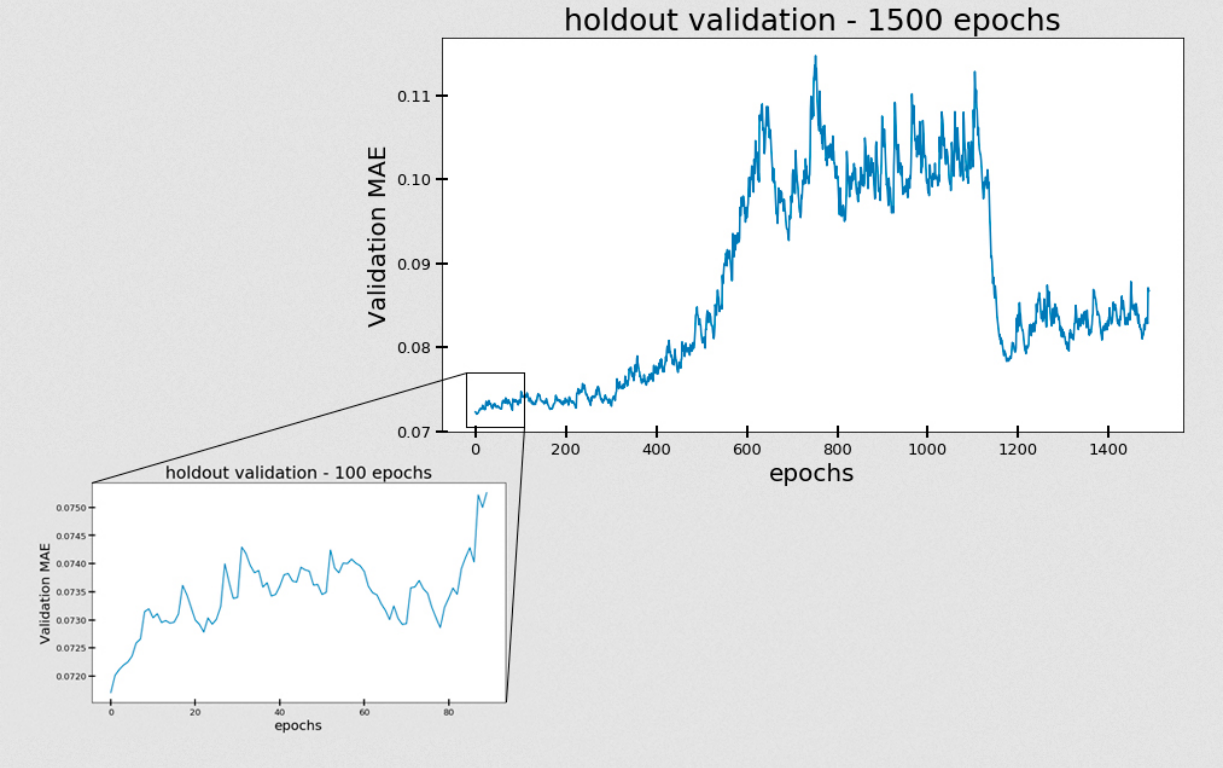

Figura 3.2: Gráfico de erro médio absoluto (MAE) ao aplicar à rede o conjunto de validação para cada uma das 1500 épocas (figura de cima). A figura de baixo representa o mesmo conjunto de dados porém vizualizando até 100 épocas. O método de validação utilizado foi o *holdout*.

das populações estelares) entre 0 e 1. Isso é feito como um pré processamento para ajudar a rede a aprender mais facilmente.

Inicialmente testamos um método de validação chamado de *holdout*, onde dividimos a base de dados aleatoriamente entre 75% dos dados para treinamento e 25% para validação. Treinamos a rede por 1500 épocas para estudar o *overfitting*, obtendo o resultado da Figura 3.2. Notamos que a partir de 200 épocas o *overfitting* começa a se tornar relevante dado o aumento constante do erro médio absoluto (MAE) para os dados de validação.

Experimentamos também um método de validação chamado *K-fold Cross-Validation*. Esse método consiste em separar a base de dados em K conjuntos de mesmo tamanho e treinar a rede K vezes, cada vez utilizando um conjunto diferente de validação. A vantagem é que todos os dados são usados tanto para treinamento quanto para teste. Para isso, divimos os dados em 4 conjuntos diferentes, sempre separando 75% para treino e 25% para validação. O resultado desse estudo para 200 épocas, mostrado na Figura 3.3, é um pouco pior do que o obtido para as 100 primeiras épocas por *holdout*. Isso é um pouco inesperado, já que geralmente o *k-fold* é um método mais robusto. Com esses resultados,

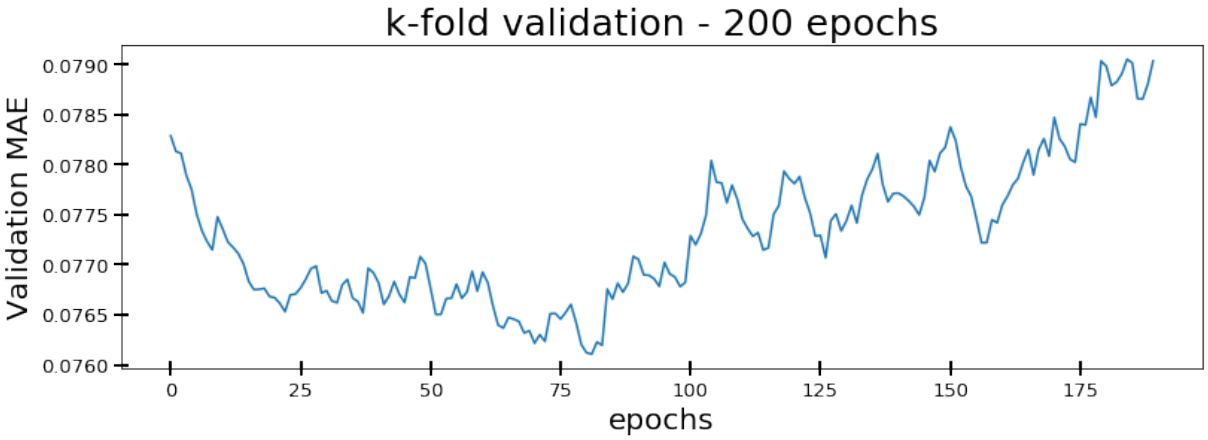

Figura 3.3: Gráfico de erro médio absoluto (MAE) ao aplicar à rede o conjunto de validação para cada uma das 200 épocas. O método de validação utilizado foi o *K-Fold Cross-Validation*.

acabamos treinando uma rede final com *holdout* por 100 épocas.

Por meio das funções do Keras, podemos estudar algumas métricas que dizem sobre a precisão do modelo. No caso de regressões, calculamos o erro médio absoluto do conjunto de validação após treinar a rede final por 100 épocas. Vale notar que para esse cálculo não aplicamos *scaling* aos dados já que queremos uma métrica que representa informações sobre os dados originais. Com isso, utilizando a classe `evaluate` do Keras, encontramos um valor (*score*) de 0.320 para o MAE do conjunto de validação da rede final.

Capítulo 4

Conclusões

A partir das sínteses espectrais de Werle (2019) fomos capazes de adicionar as linhas espectrais faltantes por meio do *software* DOBBY, gerar espectros artificiais em diferentes *redshifts* e calculamos a fotometria desses espetros. Obtivemos, assim, um cubo de dados de tamanho 231643 objetos $\times$ 100 *redshifts* $\times$ (12 bandas fotométricas + 1 valor de *redshift*).

Durante a segunda parte do projeto, treinamos uma rede neural de *Deep Learning* com os dados de uma aplicação do STARLIGHT por Cid Fernandes et al. (2005) obtendo bons resultados para um número de épocas não muito grande.

O treinamento da rede ainda é algo que está em desenvolvimento. Apesar de conseguirmos resultados interessantes para 100 épocas utilizando *redshifts* espectroscópicos, precisamos estudar como a acurácia do modelo muda ao utilizar *redshifts* fotométricos (fornecidos pelo S-PLUS) ao invés dos espectroscópicos, já que estes são somente obtidos por análises de *surveys* espectrais como SDSS. Assim como treinamos os dados de Cid Fernandes et al. (2005), pretendemos treinar em um futuro próximo uma rede com a base de dados vinda de Werle (2019) que foi construída nesse projeto. Conseguindo isso, teremos uma rede que poderá estimar parâmetros de metalicidade e idade média para *surveys* fotométricos com uma boa precisão. É importante notar que essa síntese por Cid Fernandes et al. (2005) foi feita analisando-se espectros, porém nosso método pretende obter os mesmos parâmetros utilizando a fotometria. Pretendemos ser capazes de predizer, também, outros parâmetros não incluídos no estudo de Cid Fernandes et al. (2005).

Com um resultado satisfatório para o treinamento da rede, pretendemos disponibilizar na internet os códigos utilizados para que outros astrônomos possam usufruir. Da mesma forma que transformamos os espectros de Werle (2019) em um banco de dados para treinar uma rede baseada no S-PLUS, podemos gerar dados similares para outros *surveys*

fotométricos, como o J-PLUS, J-PAS, etc.

A astronomia depende cada vez mais de *data science*. O volume de dados que os astrônomos têm acesso hoje é maior do que se poderia imaginar décadas atrás. O *machine learning*, junto com *deep learning*, ganha cada vez mais espaço na área da astronomia, e sua existência já passa de um auxílio aos métodos tradicionais para uma dependência necessária.

Como próximos passos, pretendemos melhorar o algoritmo para a obtenção da idade média em massa e luz, além de extender os resultados para outros parâmetros do S-PLUS. Em breve novos dados do S-PLUS serão disponibilizados para a comunidade, e com a implementação deles esperamos obter resultados cada vez melhores. Dado todos esses fatores, durante os próximos meses as análises decorrentes desse projeto devem ser aprimoradas.

Muitos resultados promissores no campo da astronomia fotométrica estão prestes a serem disponibilizados em um curto prazo. Novas fotometrias do S-PLUS, por exemplo, prometem ser muito mais acuradas e devem a ir a público em pouco tempo. Como o estudo de *redshifts* fotométricos é de grande interesse, o avanço nessa área é rápido e novas publicações devem aparecer em breve, melhorando a rede. Somando ambos esses fatores, assim como evoluções no âmbito geral da astronomia, geram-se grandes expectativas para esse projeto nos próximos meses, e é provável que os frutos aqui colhidos sirvam como base para outros astrônomos.

Referências Bibliográficas

Chollet, F. (2018). *Deep Learning with Python*. Shelter Island, NY: Manning Publications

Chollet F., et al. 2015. *Keras*. `https://keras.io`

Cid Fernandes R., Mateus A., Sodré L., Stanińka G., Gomes J. M. (2005). MNRAS, 358, 363

Flórido, T. 2018. *Análise de linhas de emissão em galáxias: o gás difuso ionizado nas galáxias do MaNGA*. Universidade Federal de Santa Catarina, Santa Catarina, Brazil

Kremer J., Stensbo-Smidt K., Gieseke F., Pedersen K., Igel C. (2017). IEEE Intelligent Systems, 32, 16-22

Mendes de Oliveira C., et al. (2019). MNRAS, 489, 241-267

STScI development Team. 2018. ASCL, 1811.001

Werle, A. 2019. *Analisys of SDSS spectra and GALEX photometry with STARLIGHT: stellar populations and dust attenuation in local galaxies*. Universidade Federal de Santa Catarina, Santa Catarina, Brazil

Werle A., Cid Fernandes R., Vale Asari N., et al. (2018). MNRAS, 483, 2382-2397

Apêndice

Apêndice A

Descrição dos levantamentos fotométricos

O J-PAS, *Javalambre Physics of the Accelerating Universe Astrophysical Survey 1* é uma colaboração hispano-brasileira que deve conduzir um levantamento fotométrico de ∼8500 graus quadrados do céu com um sistema de 59 filtros (54 estreitos, com ∼110 ° A de largura cada, e 5 de banda larga), a partir de um observatório dedicado, o *Javalambre Astrophysical Observatory* (JAO), em Teruel, Espanha, operado pelo CEFCA (Centro de Estudios de Física del Cosmos de Aragón, Teruel). O JAO conta com dois telescópios, um de 80cm- o T80-N- para calibração, e um de 2.5m- o T250- que conduzirá o levantamento propriamente dito. Este último estará equipado com uma câmera CCD de 1.2 Gpixel, com um campo de visão de 3 graus quadrados (a segunda maior do mundo em número de pixels, atrás apenas da usada no projeto Pan-Starrs). Diversos pesquisadores do Depto. de Astronomia do IAG participam do J-PAS. O *white paper* do projeto pode ser lido em Benítez et al. (2014, arXiv1403.5237).

O levantamento J-PLUS, acrônimo de *Javalambre Photometric Local Universe Survey2*, com o T80-N, está em andamento desde o final de 2015. Em 29 de setembro de 2017 o CEFCA anunciou o acesso aberto ao *Early Data Release* (EDR) deste *survey*, que consiste em 18 apontamentos em diferentes regiões do céu, cobrindo 36 graus quadrados. Os catálogos disponibilizados juntos com as imagens contêm mais de 400 mil objetos astronômicos observados nas 12 bandas do projeto, sendo 150 mil estrelas e 101 mil galáxias até a magnitude $r < 21$. Os primeiros artigos deste survey foram publicados há pouco tempo (ver Cenarro et al. 2018, arXiv:1804.02667) e vários outros estão em processo de revisão. Temos já vários artigos submetidos e outros em preparação.

O projeto S-PLUS, *Southern Photometric Local Universe Survey 3*, equivalente ao J-PLUS e conduzido com um telescópio semelhante ao T80-N, mas a partir do hemisfério

sul, em Cerro Tololo, no Chile, onde o telescópio T80-S está instalado, começou a coletar dados em janeiro de 2017. Em 18/12/2017 foi anunciado o *First Internal Data Release of Stripe82 fields*, consistindo de 88 apontamentos que cobrem grande parte da Stripe 82, previamente observada pelo SDSS e por outros *surveys*, o que permite aferir a qualidade das observações, além, claro, de se fazer ciência.