Draft: Calibration of DESI readsfiht measurements using machine learning

by

Sergio David Lobo Bolaño

Undergraduate Monograph

Submitted to the Department of Physics
in fulfillment of the requirements for
the degree of

Bachelor of Science in Physics

Advisor

Jaime Ernesto Forero-Romero

Department of Physics
UNIVERSIDAD DE LOS ANDES
COLOMBIA
November 8, 2018

Contents

1 Introduction **1**

1.1 DESI 1

1.2 Machine Learning 1

2 Methodology **3**

2.1 Introduction 3

2.2 Model classes 3

2.3 Data sampling, train-validation-test 3

2.4 Data Preprocessing 3

2.5 Computational Resources analysis 3

2.6 Model training, evaluation and testing 3

3 Data **5**

3.1 Introduction 5

3.2 Data descripcion 5

3.2.1 Simulated expected results - truth file 5

3.2.2 Simulated Observations - target file 5

3.3 Overview of the dataset 7

3.3.1 Redshift relations 7

3.3.2 Spectral types 8

3.3.3 Galaxy-type objects 10

3.4 Relevant Features 10

4 Results **15**

4.1 HPC results 15

4.2 Model results 15

4.3 Comparison - Metric evaluation BIC 15

4.4 Selección de modelo 15

5 Conclusions 17

Chapter 1

Introduction

Se hace una breve explicación del problema, la motivación y la solución propuesta. Se enuncian los objetivos del proyecto.

1.1 DESI

1.2 Machine Learning

Chapter 2

Methodology

2.1 Introduction

2.2 Model classes

2.3 Data sampling, train-validation-test

2.4 Data Preprocessing

2.5 Computational Resources analysis

2.6 Model training, evaluation and testing

Chapter 3

Data

3.1 Introduction

3.2 Data descripcion

The dataset used consist of two FITS files corresponding to the simulated observations of the DESI instrument and the ”truth” data from cosmological simulations. Each data file has a key column called TARGETID, this variables is the same across all simulations and identifies a particular object in the sky, thus it is needed to relate both datasets.

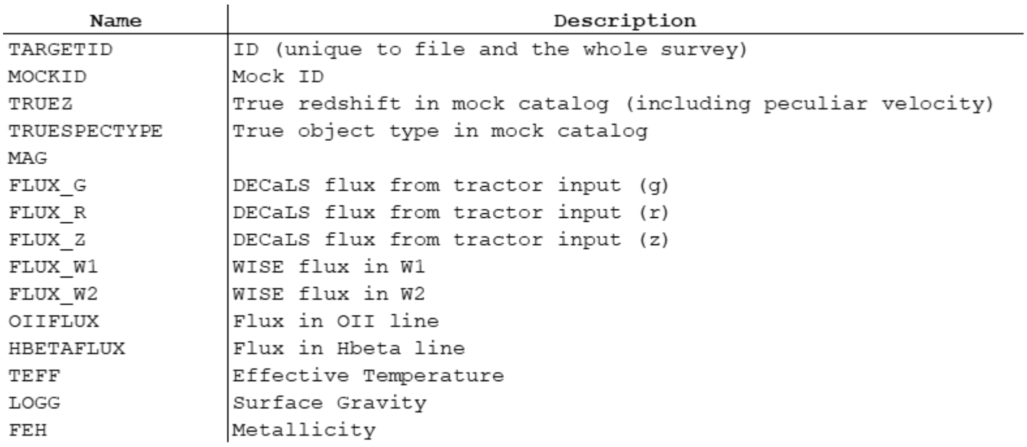
3.2.1 Simulated expected results - truth file

This file contains the data from the cosmological simulation of the target objects. Thus, this file contains the expected redshift that the instrument *should* measure. The complete list of columns in the file is shown in table 3.1[1]. Since we aim to correct the measurements of redshifts given by the instrument, the variable TRUEZ will be our target variable or ”*output*” to the machine learning models. The rest of the information will be used mostly for understanding the dataset but no as part of the ML model, since this information would not be available to the instrument in reality. This dataset contains the redshift of 24.851.543 objects.

3.2.2 Simulated Observations - target file

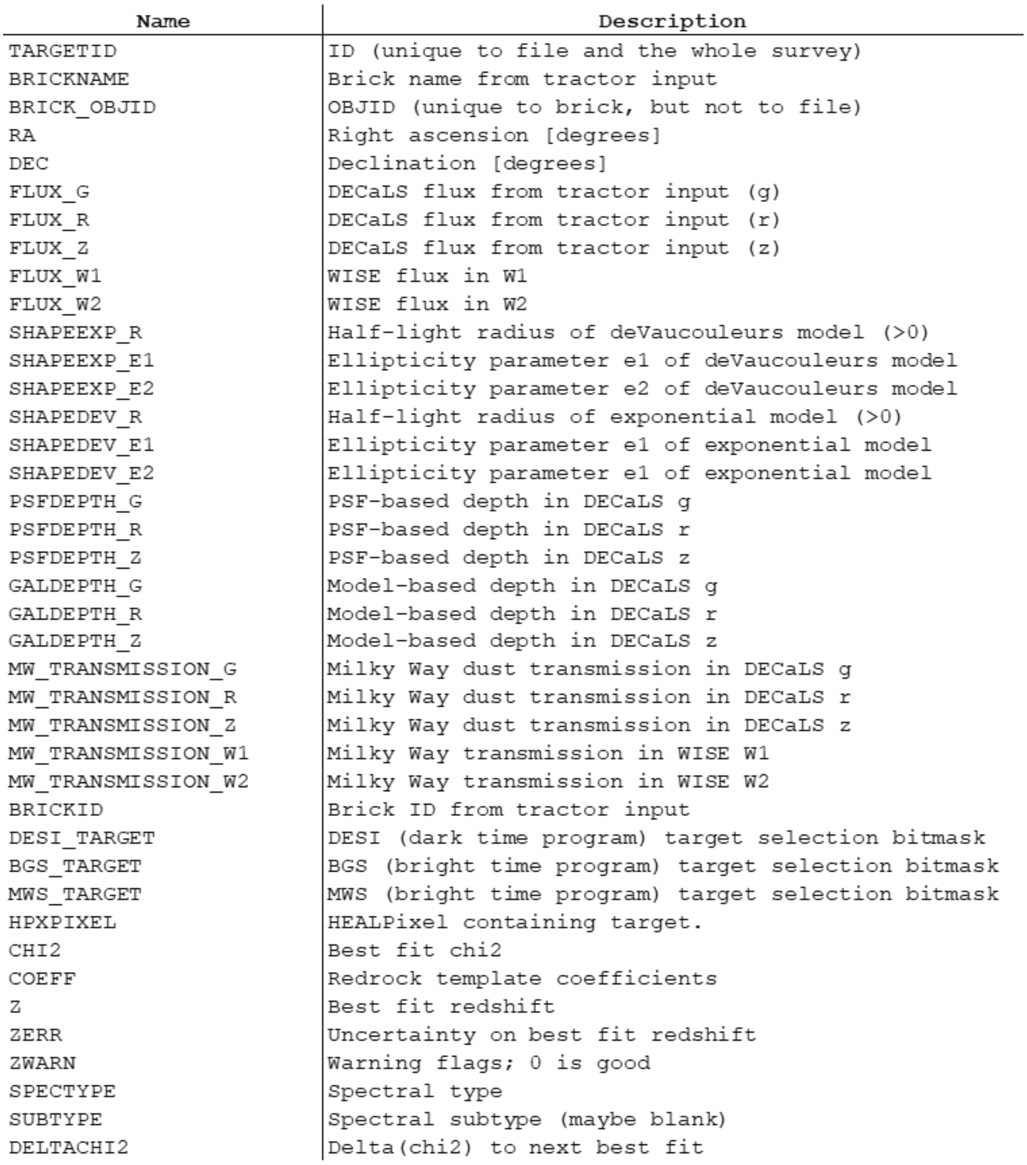
This file contains the redshift values of the targets as measured by the instrument in simulation. Apart from this, it also contains the columns shown in table 3.2. The different characteristics listed in Table 3.2 are the ones that we will use as input in our machine learning models, since is the data available from the instrument. This dataset contains the redshift of 2.131.896 objects. Since the files don´t have the same amount of points, they were cut to the one with the least (the target file) by linking the rows by its TARGETID.

[1] Additional information on `https://desidatamodel.readthedocs.io/en/stable/`



Name	Description
TARGETID	ID (unique to file and the whole survey)
MOCKID	Mock ID
TRUEZ	True redshift in mock catalog (including peculiar velocity)
TRUESPECTYPE	True object type in mock catalog
MAG	
FLUX_G	DECaLS flux from tractor input (g)
FLUX_R	DECaLS flux from tractor input (r)
FLUX_Z	DECaLS flux from tractor input (z)
FLUX_W1	WISE flux in W1
FLUX_W2	WISE flux in W2
OIIFLUX	Flux in OII line
HBETAFLUX	Flux in Hbeta line
TEFF	Effective Temperature
LOGG	Surface Gravity
FEH	Metallicity

Table 3.1: Columns in the cosmological simulation data file



Name	Description
TARGETID	ID (unique to file and the whole survey)
BRICKNAME	Brick name from tractor input
BRICK_OBJID	OBJID (unique to brick, but not to file)
RA	Right ascension [degrees]
DEC	Declination [degrees]
FLUX_G	DECaLS flux from tractor input (g)
FLUX_R	DECaLS flux from tractor input (r)
FLUX_Z	DECaLS flux from tractor input (z)
FLUX_W1	WISE flux in W1
FLUX_W2	WISE flux in W2
SHAPEEXP_R	Half-light radius of deVaucouleurs model (>0)
SHAPEEXP_E1	Ellipticity parameter e1 of deVaucouleurs model
SHAPEEXP_E2	Ellipticity parameter e2 of deVaucouleurs model
SHAPEDEV_R	Half-light radius of exponential model (>0)
SHAPEDEV_E1	Ellipticity parameter e1 of exponential model
SHAPEDEV_E2	Ellipticity parameter e1 of exponential model
PSFDEPTH_G	PSF-based depth in DECaLS g
PSFDEPTH_R	PSF-based depth in DECaLS r
PSFDEPTH_Z	PSF-based depth in DECaLS z
GALDEPTH_G	Model-based depth in DECaLS g
GALDEPTH_R	Model-based depth in DECaLS r
GALDEPTH_Z	Model-based depth in DECaLS z
MW_TRANSMISSION_G	Milky Way dust transmission in DECaLS g
MW_TRANSMISSION_R	Milky Way dust transmission in DECaLS r
MW_TRANSMISSION_Z	Milky Way dust transmission in DECaLS z
MW_TRANSMISSION_W1	Milky Way transmission in WISE W1
MW_TRANSMISSION_W2	Milky Way transmission in WISE W2
BRICKID	Brick ID from tractor input
DESI_TARGET	DESI (dark time program) target selection bitmask
BGS_TARGET	BGS (bright time program) target selection bitmask
MWS_TARGET	MWS (bright time program) target selection bitmask
HPXPIXEL	HEALPixel containing target.
CHI2	Best fit chi2
COEFF	Redrock template coefficients
Z	Best fit redshift
ZERR	Uncertainty on best fit redshift
ZWARN	Warning flags; 0 is good
SPECTYPE	Spectral type
SUBTYPE	Spectral subtype (maybe blank)
DELTACHI2	Delta(chi2) to next best fit

Table 3.2: Columns in the Simulated Observations data file

3.3 Overview of the dataset

3.3.1 Redshift relations

To understand the dataset, the first thing we need to do is see how the Z and TRUEZ variables behave. In Figure 3.1 we see three distinct regions: a 45-degree line that correspond to the redshifts measurements of DESI that are very close to the expected 'real' value, a square region where the data seems to scatter randomly except for some line groupings, and a third region of horizontal line near True Z = 0. The task is therefore to dissolve the square region and have all the points along the diagonal line.

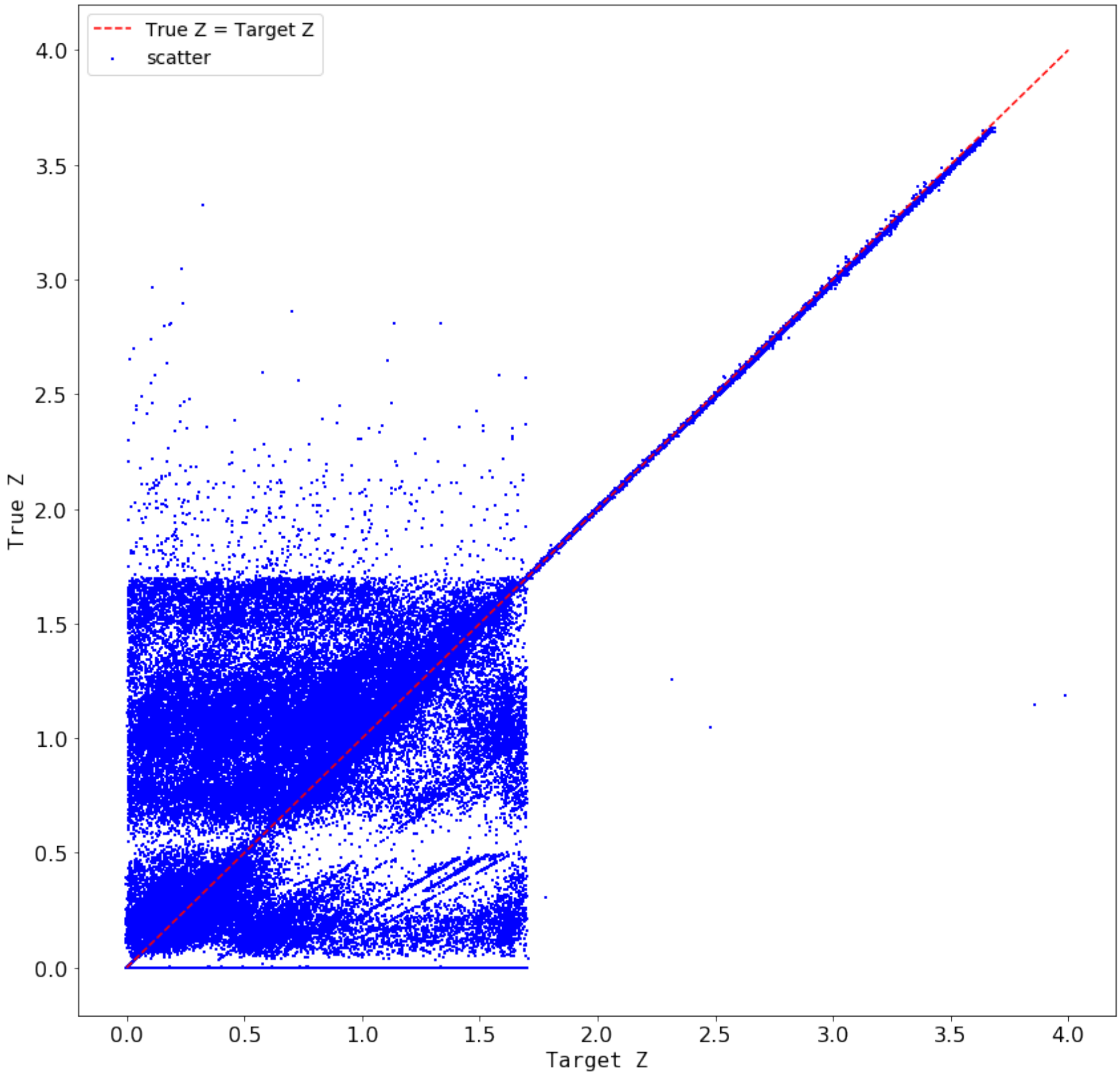
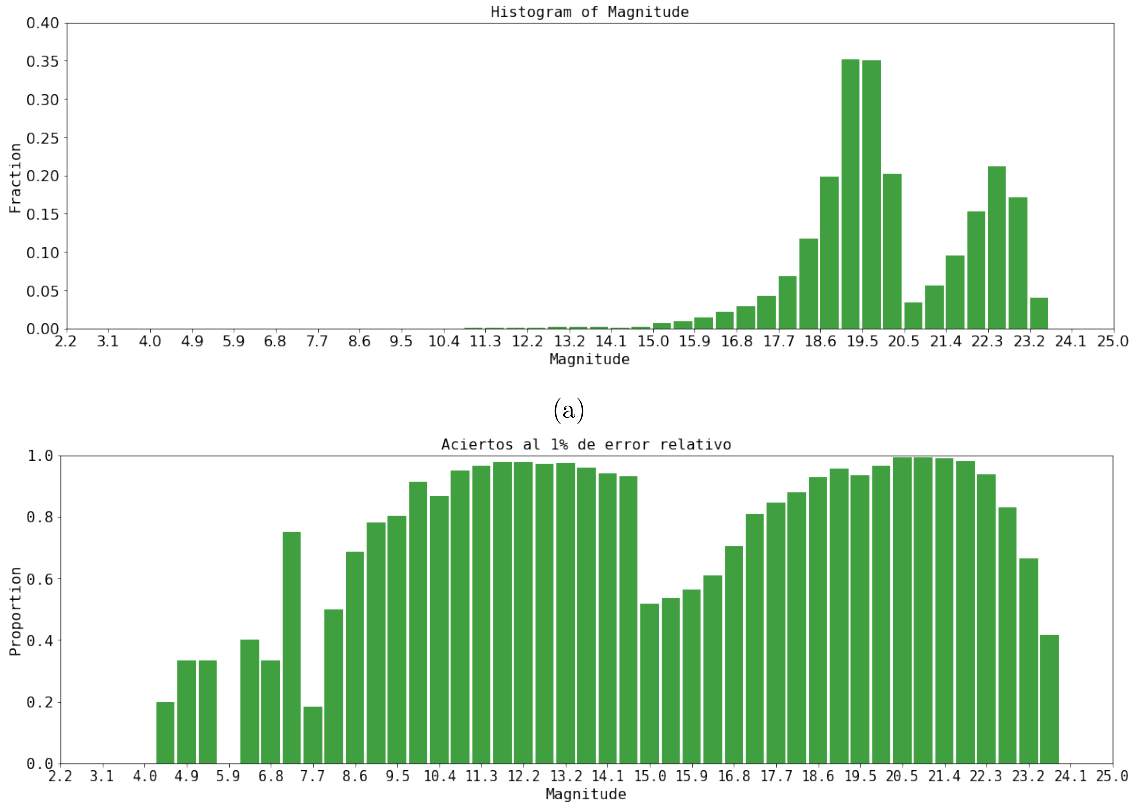


Figure 3.1: Relation between the 'observed' readshift - target Z, and the 'generated' redshift - True Z.

Now, cutting along the MAG (magnitude) variable, it is possible to see the distribution of magnitudes of the gathered data an see if there is any relation with the square region. Figure 3.2a shows this distribution and Figure 3.2b shows the fraction redshifts of each bin that is within a 1% error of the true value. We can see that the majority of the redshift is near its true value, however some valleys indicate that some regions (for example between



(a)

(b)

Figure 3.2: (a) Distribution of magnitudes in the dataset. (b) Proportion of the number of measured redshifts that are within the 1% error of their corresponding TRUEZ

15 and 17 in magnitude) are not that good. For simplicity, from now on we will refer to the **simulated expected redshift** as TRUEZ (the output to the ML models) and the **simulated observations of the targets** as Z (the input to the models) or TARZ.

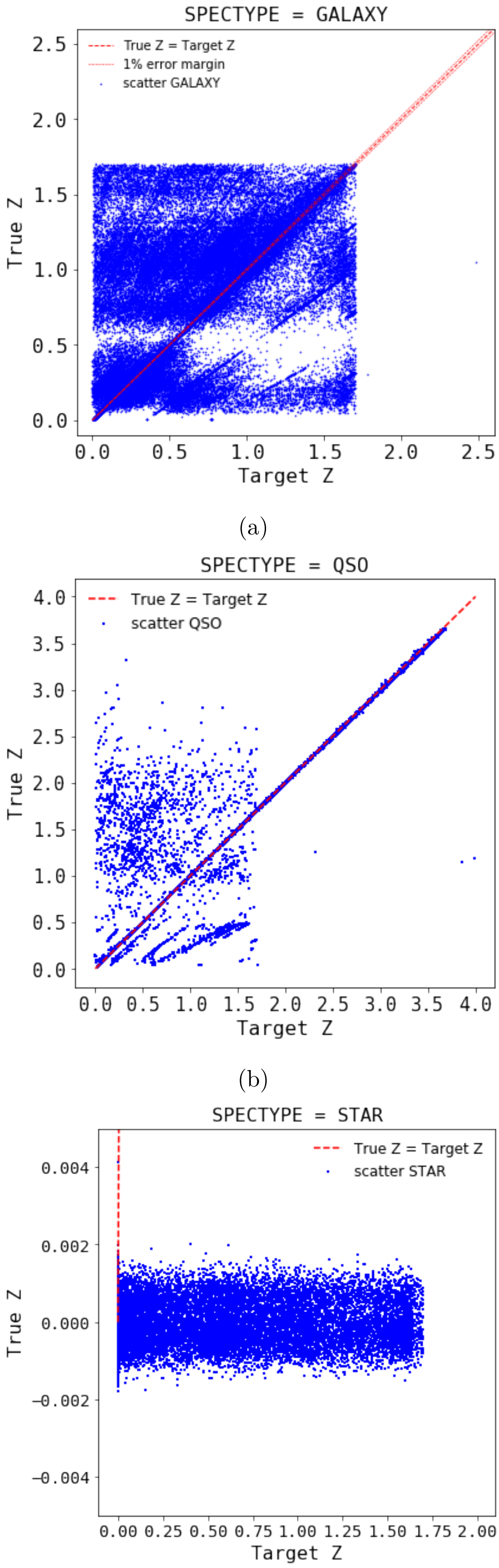
3.3.2 Spectral types

From figure 3.2 it is possible to infer that the different magnitudes of the targets can be related to the different regions on figure 3.1, and this difference in magnitude is also related to the SPECTYPE of each target. The spectral type of the data in the tar-file is distributed as shown in Table 3.3. Near 85% of the dataset is composed of Galaxy-type objects, for this reason, this is the most interesting class.

SPECTYPE	N samples	% of dataset
Galaxy	1796213	84.25
QSO	194319	9.11
Star	141364	6.63
Total	2131896	100

Table 3.3: Spectral type distribution of the tar file.

The redshift relations similar to Figure 3.1 discriminated by SPECTYPE are shown in



(a)

(b)

(c)

Figure 3.3: Redshift relation for (a) Galaxy-type objects, (b) QSO-type objects, and (c) Star-type objects

Figure 3.3, where the three regions in Figure 3.1 seems to be related to each spectral type. The GALAXY objects are distributed along a square and there are lines formed at angles different than 45, which means that there is a relation between the two variable but 'out of calibration'. It is worth mentioning that Figure 3.3a can be deceiving, because 96.34% of the Galaxy-type data is within the 1% margin line in the figure, which means that the square is formed only by the 3.66 % of the galaxy data, corresponding to approximately 65.742 measurements, still a lot.

The QSO-type objects in Figure 3.3b correspond to the 45-degree line in Figure 3.1 since the majority of points are along this line, although the same line patter and dispersion of the galaxy-type objects are present, but in least quantity. However, the Star-type in Figure 3.3c is randomly scattered over the TARZ range and correspond to the horizontal line in Figure 3.1. Once again, the most interesting SPECTYPE is GALAXY, because it has the majority of error in TARZ, and also presents different patterns, QSO are already fine and are not a priority while STAR is completely random a represent a small fraction of the whole dataset. From now on we will focus on Galaxy-type objects only.

3.3.3 Galaxy-type objects

Galaxy-type objects are clasified as Bright Galaxy Survey (BGS), Emission Line Galaxies (ELG) and Luminous Red Glaxies (LRG) distributing according to Table 3.4. In this case, the sub-types are more evenly distributed. The redshift relations of each sub-type is shown in Figure 3.4. BGS and ELG follow similar pattern, however ELG is more disperse while BGS is clustered along the TARZ = TRUEZ line, whereas LRG sub-type is already perfect.

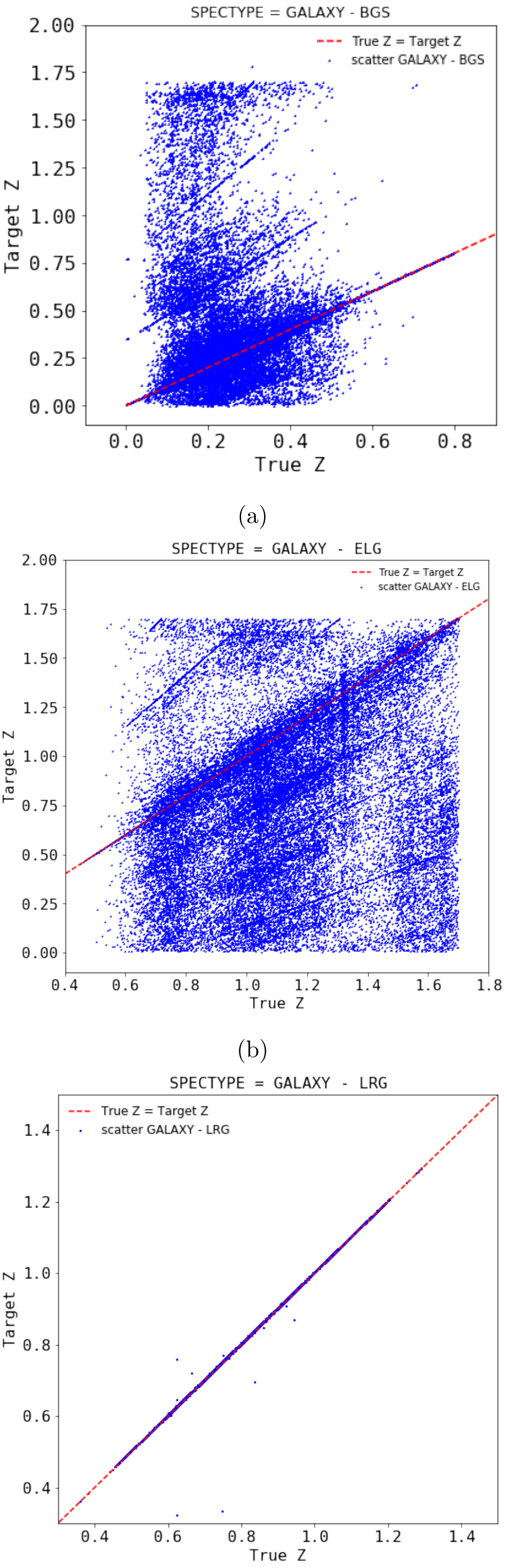
Galaxy subtype	N samples	% of dataset
BGS	889336	49.51
ELG	601847	33.51
LRG	305030	16.98
Total	1796213	100

Table 3.4: Distribution of Galaxy sub-types

Note that the line structures are more visible in Figure 3.4a than in Figure 3.4b, therefore, to extract the relevant features in the tar dataset that may be related to this structure, we will use the BGS subset and then see if the feature extracted are also useful for the ELG subset. Therefore, from now on we will use only the BGS data subset and proceed to find the relevant features (in Table 3.2)

3.4 Relevant Features

The ability of the instrument for measuring the correct redshift will probably depend on the quality of the fluxes that the fiber optics receive. The fluxes of the objects are related



(a)

(b)

(c)

Figure 3.4: Redshift relation for Galaxy (a) BGS-type objects, (b) ELG-type objects, and (c) LRG-type objects

to its magnitude, as we saw en Figure 3.2, magnitude is related to the correct prediction of the instrument's Z, but since the information of magnitude is known to instrument in form of fluxes, this variables will be explored next. The six fluxes variables are named in Table 3.2. The distributions of the fluxes in the BGS dataset are shown in Figure 3.5.

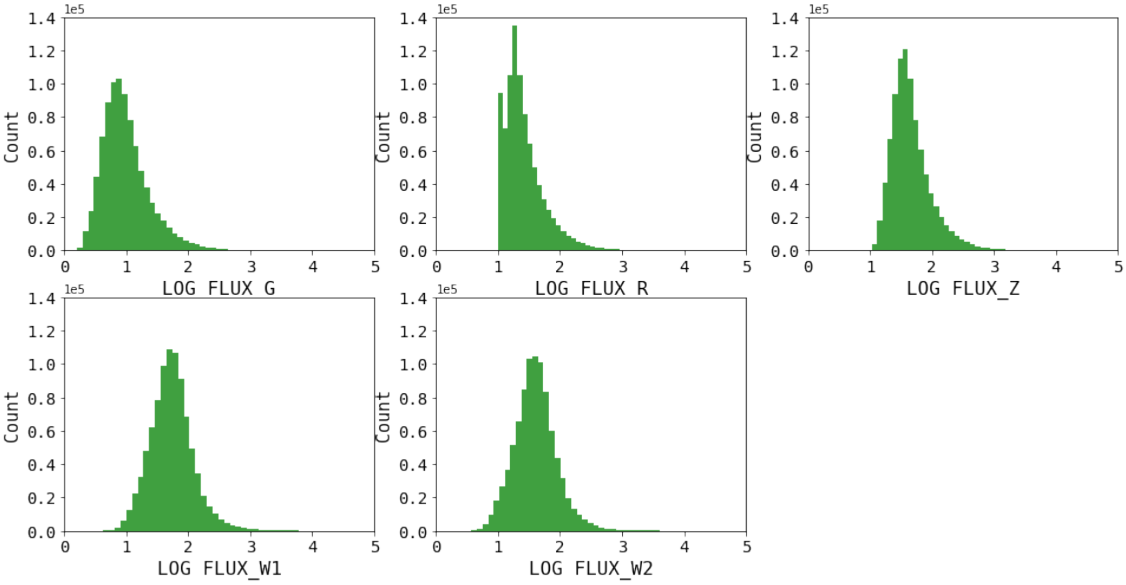


Figure 3.5: Distribution of the flux variable in the BGS subset.

In order to keep track of the relation between TARZ y TRUEZ and see the behavior of the flux variable with respect to the two, we define the following variable

$$\alpha = \frac{TRUEZ}{TARZ}, \tag{3.1}$$

therefore, alpha have a value near 1 when the redshifts are along de 45-degree line.

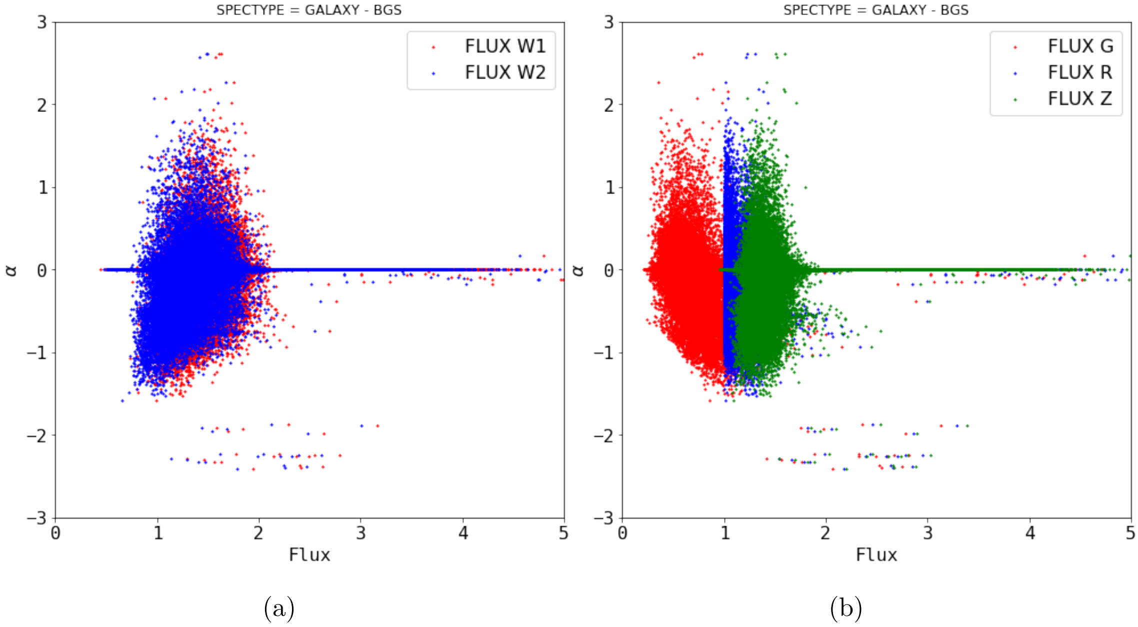


Figure 3.6: Relation between fluxes, TRUEZ and TARZ. (a) α as a function of W1 and W2 fluxes. (b)α as a function of G, R and Z fluxes. $\log_{10}$ values presented.

We see in Figure 3.6 that each flux has a similar behaviour, however, they are dispersed in different values and present different minimum values. As the flux magnitude increases, α tends to equal 1, meaning that as the fluxes received by the instrument increases, the redshift measurements are closer to the expected real values. Therefore, the dispersion regions may contain the information necessary for the ML models to learn no predict better redshifts. To see the influence of the dispersion at low fluxes over α, the cuts at different fluxes values in Figure 3.7 show the distribution of α at particular values of the fluxes.

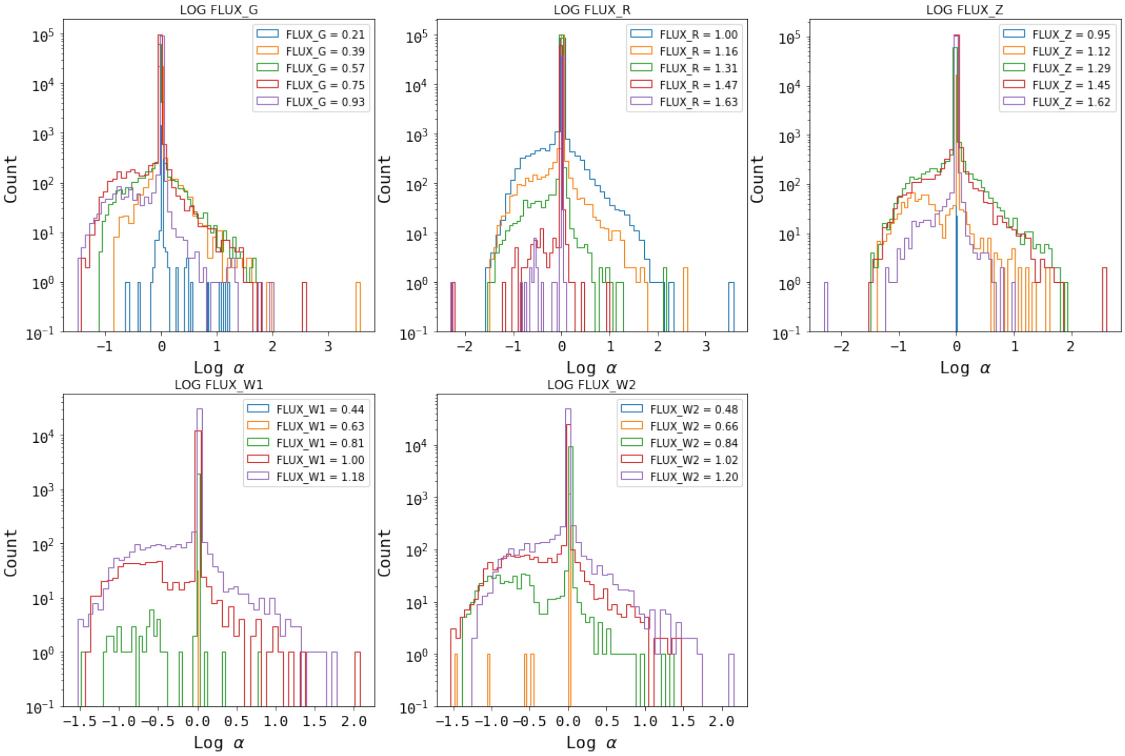


Figure 3.7: Distribution of α at different cuts in the flux variable.

The plots in Figure 3.7 are constructed by taking thin slices at fixed values of the fluxes in Figure 3.6 and constructing the histogram of the points that lie within the slice. The first approximation is to fit the distribution to a Gaussian distribution and estimate its mean and variance. The relation between mean and fluxes is shown in Figure 3.8, were the green points indicate the mean of α for the slice taken at the given flux.

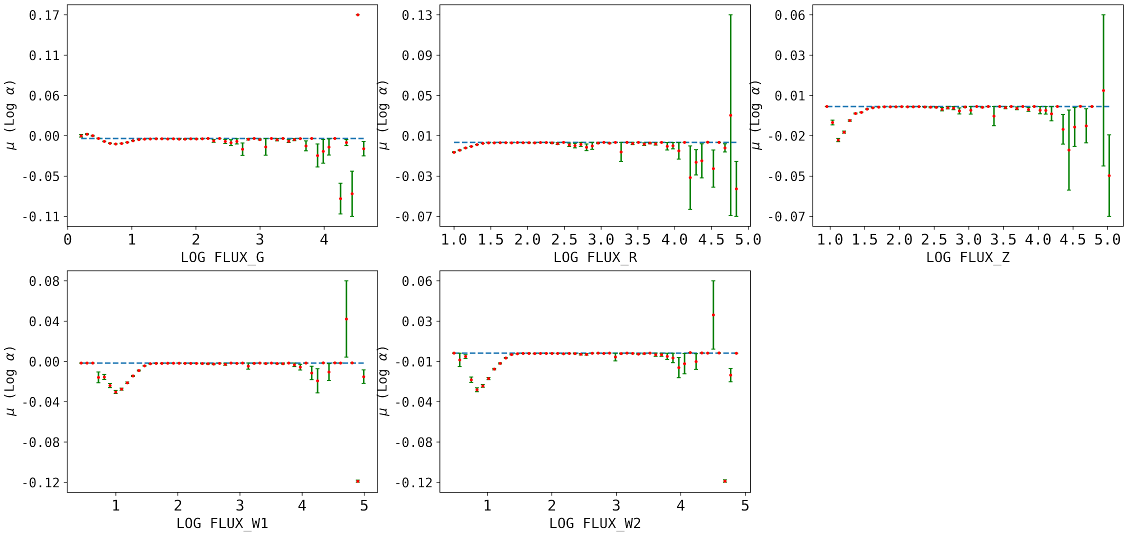


Figure 3.8: Mean and standard error α for every Flux interval (slice).

Chapter 4

Results

4.1 HPC results

4.2 Model results

4.3 Comparison - Metric evaluation BIC

4.4 Selección de modelo

Chapter 5

Conclusions

Bibliography

[1] N. Andersson and G. L. Comer. Relativistic fluid dynamics: Physics for many different scales. *Living Reviews in Relativity*, 10(1):1, 2007.

[2] G. K. Batchelor. *An introduction to fluid dynamics.* Cambridge university press, 2000.

[3] M. Blau. *Lecture notes on general relativity.* Albert Einstein Center for Fundamental Physics Bern Germany, 2011.

[4] J. Foster and J. D. Nightingale. *A short course in General Relativity.* Springer Science & Business Media, 2010.

[5] J. B. Hartle. *Gravity: An introduction to Einstein´s General Relativity. Mathematica Programs.*

[6] L. Herrera. Relativistic fluids and the physics of gravitational collapse. *arXiv preprint arXiv:0909.3474*, 2009.

[7] L. Herrera, A. Di Prisco, E. Fuenmayor, and O. Troconis. Dynamics of viscous dissipative gravitational collapse: a full causal approach. *International Journal of Modern Physics D*, 18(01):129–145, 2009.

[8] L. Herrera and N. Santos. Dynamics of dissipative gravitational collapse. *Physical Review D*, 70(8):084004, 2004.

[9] W. A. Hiscock and L. Lindblom. Stability and causality in dissipative relativistic fluids. *Annals of Physics*, 151(2):466–496, 1983.

[10] W. Israel. Nonstationary irreversible thermodynamics: a causal relativistic theory. *Annals of Physics*, 100(1-2):310–331, 1976.

[11] W. Israel and J. Stewart. Transient relativistic thermodynamics and kinetic theory. *Annals of Physics*, 118(2):341–372, 1979.

[12] P. S. Joshi. *Gravitational collapse and spacetime singularities.* Cambridge University Press, 2007.

[13] L. Landau and E. Lifshits. *Fluid Mechanics, by L.D. Landau and E.M. Lifshitz.* Teoreticheska fizika. Pergamon Press, 1959.

[14] B. Lautrup. *Physics of continuous matter: exotic and everyday phenomena in the macroscopic world.* CRC Press, 2004.

[15] R. Maartens. Causal thermodynamics in relativity. *arXiv preprint astro-ph/9609119*, 1996.

[16] M. M. May and R. H. White. Hydrodynamic calculations of general-relativistic collapse. *Physical Review*, 141(4):1232, 1966.

[17] C. W. Misner and D. H. Sharp. Relativistic equations for adiabatic, spherically symmetric gravitational collapse. *Physical Review*, 136(2B):B571, 1964.

[18] C. W. Misner, K. S. Thorne, and J. A. Wheeler. *Gravitation.* Macmillan, 1973.

[19] J. R. Oppenheimer and H. Snyder. On continued gravitational contraction. *Physical Review*, 56(5):455, 1939.

[20] T. Padmanabhan. *Gravitation: foundations and frontiers.* Cambridge University Press, 2010.

[21] R. Penrose. Gravitational collapse and space-time singularities. *Physical Review Letters*, 14(3):57, 1965.

[22] G. Pinheiro and R. Chan. Radiating gravitational collapse with shear viscosity revisited. *General Relativity and Gravitation*, 40(10):2149–2175, 2008.

[23] E. Poisson. A relativistic toolkit. *CUP, Cambridge*, page 85, 2004.

[24] A. Prasanna, J. Narlikar, C. Vishveshwara, and I. A. of Sciences. *Proceedings of the Workshop on Gravitation and Relativistic Astrophysics, Ahmedabad, 18-20 January 1982.* Published for the Indian Academy of Sciences, Bangalore, by World Scientific Pub. Co., 1984.

[25] L. Rezzolla and O. Zanotti. *Relativistic hydrodynamics.* Oxford University Press, 2013.

[26] B. Schutz. *A first course in general relativity.* Cambridge university press, 2009.

[27] P. Sharan. *Spacetime, geometry and gravitation*, volume 56. Springer Science & Business Media, 2009.

[28] R. C. Tolman. Effect of inhomogeneity on cosmological models. *Proceedings of the National Academy of Sciences*, 20(3):169–176, 1934.

[29] T. Tsumura, T. Kunihiro, and K. Ohnishi. Derivation of covariant dissipative fluid dynamics in the renormalization-group method. *Physics Letters B*, 646(2):134–140, 2007.

[30] P. Vaidya. Nature, 171, 260. *Google Scholar*, 1953.

[31] S. Weinberg. *Gravitation and cosmology: principles and applications of the general theory of relativity*, volume 67. Wiley New York, 1972.