UNIVERSITY OF CRETE
SCHOOL OF SCIENCES AND ENGINEERING
DEPARTMENT OF PHYSICS

Observations of galaxies in the local universe with the space telescope *Gaia* ☾

Sophia Tsiatsiou

Supervisors: Prof. Vassilis Charmandaris, Dr. Ioannis Bellas-Velidis

*A thesis submitted in fulfilment of the requirements
for the degree of Masters of Science in the
Faculty of Physics*

HERAKLION 2018

Abstract

Observations of galaxies in the local universe with the space telescope *Gaia*

Gaia is a European Space Agency (ESA) cornerstone mission which was launched in 2013. While its main objective is to take a census of the stellar content of our Galaxy, it will also observe a large number of many other objects, among them over a million of unresolved galaxies. The objective of the Thesis is to extend, towards larger redshifts, the module Unresolved Galaxy Classifier (UGC) that is being developed for the *Gaia* ground-based pipeline by the Greek team, member of the Data Processing and Analysis Consortium (DPAC). The UGC is being developed to use low-resolution spectra of galaxies observed by *Gaia's* BP/RP spectrophotometer to classify the galaxies and to estimate specific astrophysical parameters. Our purpose is to develop an automatic procedure, based on supervised machine learning algorithm, through the Support Vector Machines (SVM), to predict the redshift of galaxies from BP/RP spectra up to $z = 0.6$. As a very first step we found which sources in the first *Gaia's* archive are galaxies by cross-matching the *Gaia* DR1 with the SDSS DR13 spectra table, which is composed of sources that have been observed spectroscopically. This led us to an estimate that more than 1.5 million of galaxies are observed by *Gaia*. In order to be able to create the redshift estimator, we selected a properly restricted subset of the matched galaxies (*Gaia* DR1 with SDSS DR13 BOSS) and downloaded their SDSS's photometric and spectroscopic parameters (among them the redshifts). The corresponding SDSS BOSS spectra have been modelled for the BP/RP spectrophotometer forming with the parameters an "empirical library" of *Gaia's* galaxies spectra. In a series of experiments with the SVM, trained and tested using the spectra and redshifts from the library, we defined an optimal SVM model to be used in the UGC. In this optimization the error in redshift prediction for galaxies with *Gaia's* magnitude G=17 within the range z=0.0-0.6 is 0.056, while in the middle part of the range it is as small as 0.029. Ways to improve this performance are discussed.

Περίληψη

Παρατηρήσεις γαλαξιών στο τοπικό Σύμπαν με το διαστημικό τηλεσκόπιο *Γαία*

Η Γαία αποτελεί μια ιδιαίτερα σημαντική αποστολή του Ευρωπαϊκού Οργανισμού Διαστήματος (ESA) που εκτοξεύτηκε το 2013. Αν και ο κύριος στόχος της είναι να κάνει μια απογραφή του αστρικού περιεχομένου του Γαλαξία μας, θα παρατηρήσει επίσης κι ένα μεγάλο αριθμό πολλών άλλων αντικειμένων, μεταξύ των οποίων και πάνω από 10^6 ανεξιχνίαστων γαλαξιών. Ο στόχος της εργασίας είναι να επεκτείνουμε, προς μεγαλύτερες μετατοπίσεις προς το ερυθρο (redshifts), το πρόγραμμα Unresolved Galaxy Classifier (UGC) που αναπτύσσεται για το επίγειο σύστημα επεξεργασίας δεδομένων της Γαίας από μια Ελληνική ομάδα, που είναι μέλος του Data Processing and Analysis Consortium (DPAC). Το UGC αναπτύσσεται για να χρησιμοποιεί φάσματα γαλαξιών χαμηλής-ανάλυσης που παρατηρούνται από το φασματοφωτόμετρο BP/RP της Γαίας για την ταξινόμηση των γαλαξιών και την εκτίμηση συγκεκριμένων αστροφυσικών παραμέτρων. Ο σκοπός μας είναι να αναπτύξουμε μια αυτόματη διαδικασία, βασισμένη σε έναν εποπτευόμενο αλγόριθμο μηχανικής μάθησης, μέσω του Support Vector Machines (SVM), ώστε να προβλέψουμε τα redshifts των γαλαξιών από τα BP/RP φάσματα μέχρι και για $z = 0,6$. Ως πρώτο βήμα βρήκαμε ποιες πηγές είναι γαλαξίες στο πρώτο αρχείο της Γαίας αντιστοιχίζοντας το *Gaia* DR1 με τον πίνακα φασμάτων του SDSS DR13, ο οποίος αποτελείται από πηγές οι ιδιότητες των οποίων έχουν υπολογιστεί φασματοσκοπικά. Αυτό μας οδήγησε στην εκτίμηση ότι η Γαία θα παρατηρήσει πάνω από 1,5 εκατομμύρια γαλαξίες. Για να μπορέσουμε να δημιουργήσουμε τον εκτιμητή του redshift, επιλέξαμε ένα κατάλληλο υποσύνολο από τους αντιστοιχισμένους γαλαξίες (*Gaia* DR1 με SDSS DR13 BOSS) και μελετήσαμε τις φωτομετρικές και φασματοσκοπικές παραμέτρους τους από το SDSS (ανάμεσα σε αυτές και τα redshifts). Τα αντίστοιχα φάσματα του SDSS BOSS μοντελοποιήθηκαν για το φασματοφωτόμετρο BP/RP διαμορφώνοντας έτσι μια "εμπειρική βιβλιοθήκη" των φασμάτων γαλαξιών της Γαίας. Σε μια σειρά από πειράματα με το SVM, εκπαιδεύοντας και δοκιμάζοντάς το χρησιμοποιώντας τα φάσματα και τα redshifts από τη βιβλιοθήκη, καθορίσαμε το βέλτιστο SVM μοντέλο που θα χρησιμοποιηθεί στο UGC. Στην βελτιστοποίηση το σφάλμα στο εκτιμόμενο redshift για τους γαλαξίες με μέγεθος G της Γαίας G=17 εντός της περιοχής z = 0,0-0,6 είναι 0,056, ενώ στο μεσαίο τμήμα της περιοχής είναι μόλις 0,029. Στο τέλος της εργασίας παραθέτουμε κάποιες σκέψεις για την βελτίωση της μεθόδου μας.

Acknowledgments

I wish to express my gratitude to my supervisors, Dr. Ioannis Bellas-Velidis and Prof. Vassilis Charmandaris, for their guidance, patience and support during the work for this thesis. Particularly, I am grateful for the collaboration with Dr. Ioannis Bellas-Velidis who helped me through our meaningful discussions to complete this work and for the time he devoted me. Also, I would like to thank Prof. Andreas Zezas for his help and guidance when I was an undergraduate student. Finally, special thanks should be awarded to my family for their support through the years in every possible way. Without them I would not be able to reach my goals. Last but not least, I would like to thank all the people who supported me. With Giannis Komis, Kostas Mouloudakis, Panagiotou Christos and Stauroula Karampatzaki we had a wonderful time in the University of Crete all these years.

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is `www.sdss.org`.

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astroísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

To my beloved family

Contents

Acronyms

AGN	Active Galactic Nucleus
AURA	Association of Universities for Research in Astronomy
AF	Astrometric Field
Apsis	Astrophysical Parameters Inference System
BOSS	Baryon Oscillation Spectroscopic Survey
BP/RP	Blue photometer/Red photometer
CAS	Catalogue Archive Server
CU	Coordination Unit
DPAC	Data Processing and Analysis Consortium
DPC	Data Processing Center
DR	Data Release
ESA	European Space Agency
GALAXY	Galaxy object type in SDSS
LSF	Line Spread Function
QSO	Quasar object type in SDSS
RVS	Radial-Velocity Spectrometer
SM	Sky Mapper
SDSS	Sloan Digital Sky Survey
SVM	Support Vector Machine
UGC	Unresolved Galaxy Classifier

List of Figures

List of Tables

1 Introduction

1.1 *Gaia* mission

1.1.1 The mission

Gaia is an European Space Agency (ESA) cornerstone mission[1] (Gaia Collaboration et al. 2016b). Its purpose is to construct the largest and most accurate 3D space catalogue ever made of about one billion stellar objects, mainly stars in our Galaxy. Moreover, it will provide the age, mass and chemical composition of 1% of the stars in the Milky Way. Observing the stellar content one can study the origin and evolution of our Galaxy. Outside of our Galaxy, *Gaia* will observe individual very bright stars in nearby galaxies, as well as the farthest and oldest known objects: quasars and primordial galaxies. *Gaia* can also provide information about the general characteristics of galaxies which are too remote to resolve their stars.

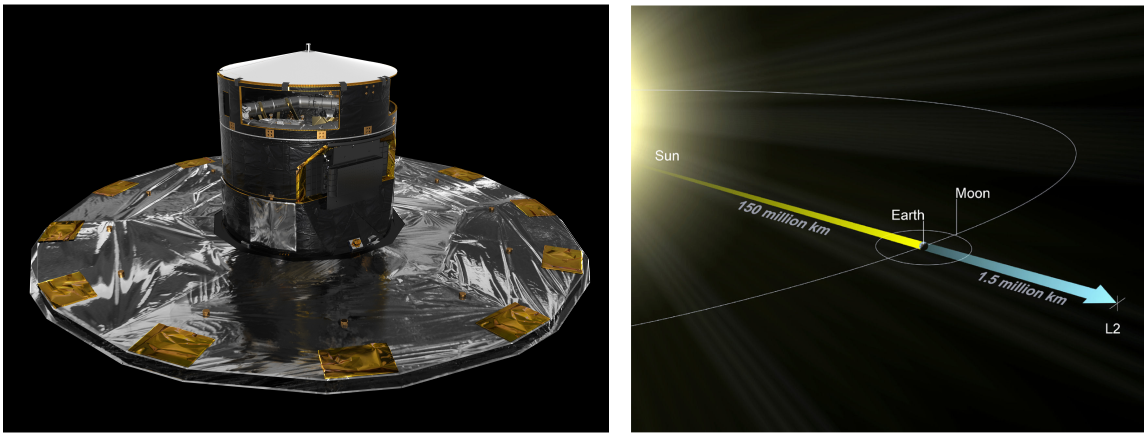


Figure 1: Left panel: The *Gaia* spacecraft. The 3m-diameter cylindrical structure holds the optical, the electronic, and the service systems, while the 11m in diameter flat structure is the payload's protection sunshield. Right panel: *Gaia* will operate in the vicinity of the L2, approximately 1.5 million km from the Earth, along the Sun-Earth line in the direction opposite to the Sun. The region around L2 is a gravitational saddle point where spacecraft can be maintained at roughly constant distance from the Earth for several years by small and energetically manoeuvres. (Credit: François Mignard, ESA/*Gaia*)

The observatory was launched on 19 December 2013 for its operating point. After about one month, an orbit insertion manoeuvre injected the spacecraft onto the stable manifold of the operational Lagrange point (L2) orbit. Its orbit has been chosen to be a controlled Lissajous orbit around the L2 (see Figure 1, right panel) of the Sun-Earth system in order to have a quiet environment for the payload in terms of thermo-mechanical stability. Interruptions in observations are avoided in the L2 point, since the Earth, Moon, and Sun lie inside *Gaia's* orbit. Complete check-out of the spacecraft and the instruments preceded the start of nominal operations (end of July 2014). The mission will be in operation over a period of five years. There is a potential extension for a period of one to four years, as the observatory has enough consumables to operate for approximately nine years, and its detectors are not degrading as fast as initially expected.

1.1.2 Telescopes, instruments, and observations

The optical system (Figure 2, left panel) of *Gaia* is composed of two identical telescopes separated by the basic angle of 106.5° and each has size 1.45×0.5 m^2 (1.7° by 0.6° on the sky)

[1] https://www.cosmos.esa.int/web/gaia

and 35m focal length providing high angular resolution. Three instruments, in a common focal plane (see Figure 2, right panel), detect the light collected by the telescopes: the astrometric, the photometric, and the spectroscopic instrument. *Gaia* will provide unprecedented accuracy in astrometry and photometry, as well as low-resolution two-band and limited-band high-resolution spectrophotometry. Astrometry and photometry will be performed for all the detected objects up to 20-21 magnitude. High-resolution spectra will be obtained only for the brighter sources up to 17th magnitude.

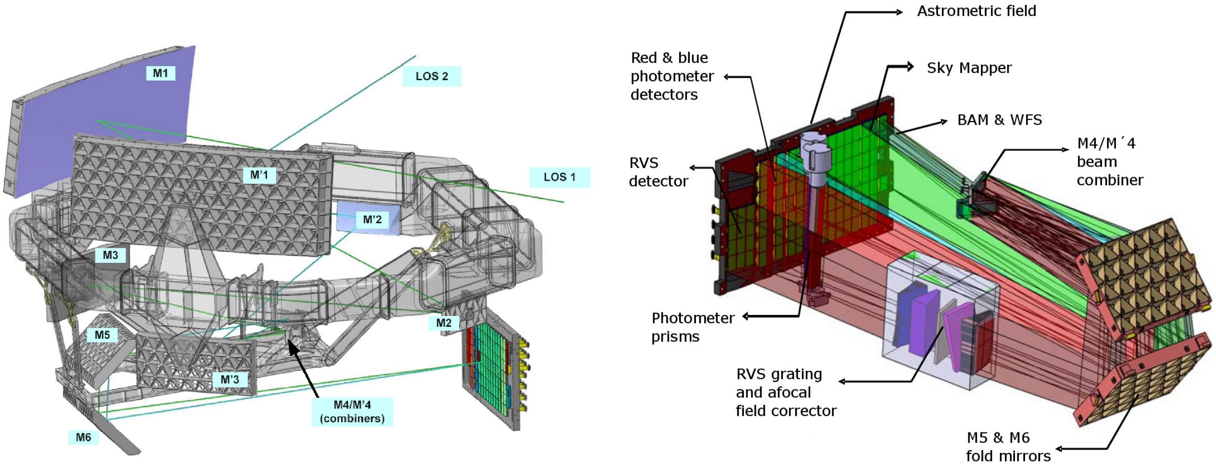


Figure 2: The *Gaia's* optical system with two main mirrors (Left) and the three instruments feeding the common focal plane (Right). (Credit: EADS Astrium)

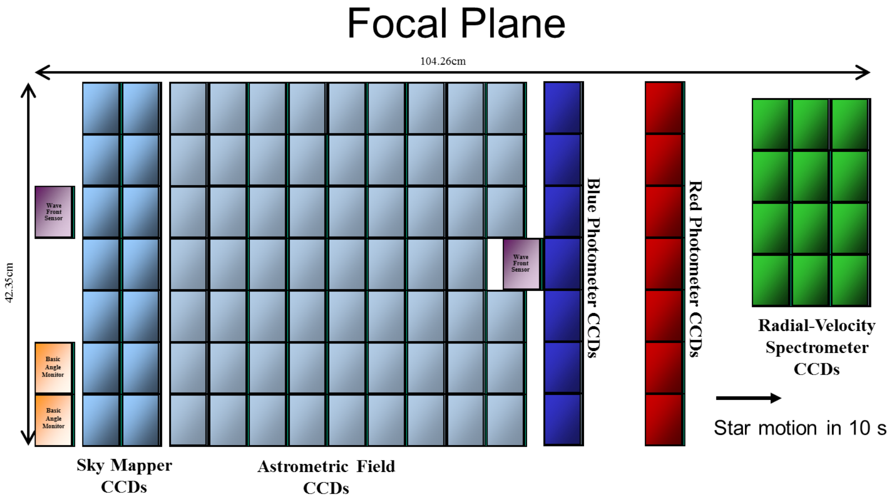


Figure 3: *Gaia's* focal plane. The images of the stars move from left to right. The 14 CCDs to the left correspond to the SM – blue, adjacent to the right are the 62 CCDs of the AF – light blue, followed by the CCDs of the BP, RP, and the RVS – green. Additionally two CCDs of the Wavefront Sensor – purple, and two for the Basic Angle Monitor are shown – orange. (Credit: A. Short & J. de Bruijne, ESA)

Inside the satellite the two telescopes are sharing a common focal plane. As the spacecraft slowly rotates, the light from the celestial object passes across the focal plane. It consists of 106 CCDs, a total of almost 1 billion pixels, which function as the detectors of the various instruments in the *Gaia* payload. Each CCD consists of 4500 × 1966 pixels (pixel size $10\mu m \times 30\mu m$). The focal plane serves the following five main functions:

- The Wavefront sensor and basic-angle monitor (purple and orange boxes respectively in Figure 3).

- The Sky Mapper (SM, blue boxes, in Figure 3), containing 14 CCDs (seven per telescope), which autonomously detects objects entering the fields of view and provides to the video-processing unit (VPU) details about the star transits through the subsequent CCDs. This is the only field where the images from the two mirrors are still separate (first telescope on the first column of SM CCDs, the second telescope on the second column).

- The main Astrometric Field (AF, light blue boxes, in Figure 3), covering 62 CCDs, devoted to angular-position measurements, providing astrometric parameters, such as positions, parallaxes, and proper motions of stars. The field provides also on unfiltered photometric measurements (broad-band white-light G-band fluxes).

- The Blue and Red Photometers (BP and RP, dark blue and red box respectively, in Figure 3), providing low-resolution, spectro-photometric measurements for each object, and contain seven CCDs each. The dispersion element consists of two prisms operating in 330-680 nm wavelength range for the BP and in 640-1050 nm for the RP. The spectral dispersion ranges from 3-27 nm per pixel for BP and 7-15 for RP, correspondingly. The spectral energy distribution measurements and the integrated BP-band and RP-band fluxes yield key astrophysical information for the observed sources.

- The Radial-Velocity Spectrometer (RVS, green boxes, in Figure 3), an integral-field spectrograph ($^{\lambda}/_{\Delta\lambda} \approx 11,500$ in the range of 845-872 nm around the Ca triplet) covering 12 CCDs collects high-resolution spectra of all objects brighter than 17th magnitude, allowing derivation of radial velocities and stellar atmospheric parameters.

Gaia was designed to map the whole celestial sphere and to repeatedly observe its objects. A “scanning law” has been prepared which dictates how the satellite’s spin axis evolves with time during the mission, as the satellite will continuously spin around its axis (see Figure 4, left panel). As a result, over a period of 6 hours, the two fields of view will scan across all objects located along the great circle “perpendicular” to the spin axis. Because of the basic angle separation, nearly the same part of the sky will transit the second field of view with a delay to the first field. *Gaia’s* spin axis does not point to a fixed direction in space but is carefully controlled so as to precess slowly on the sky. Therefore, the great circle changes slowly with time, allowing repeated full sky coverage over the mission lifetime. On average, each object on the sky is observed about 70 times (see Figure 4, right panel).

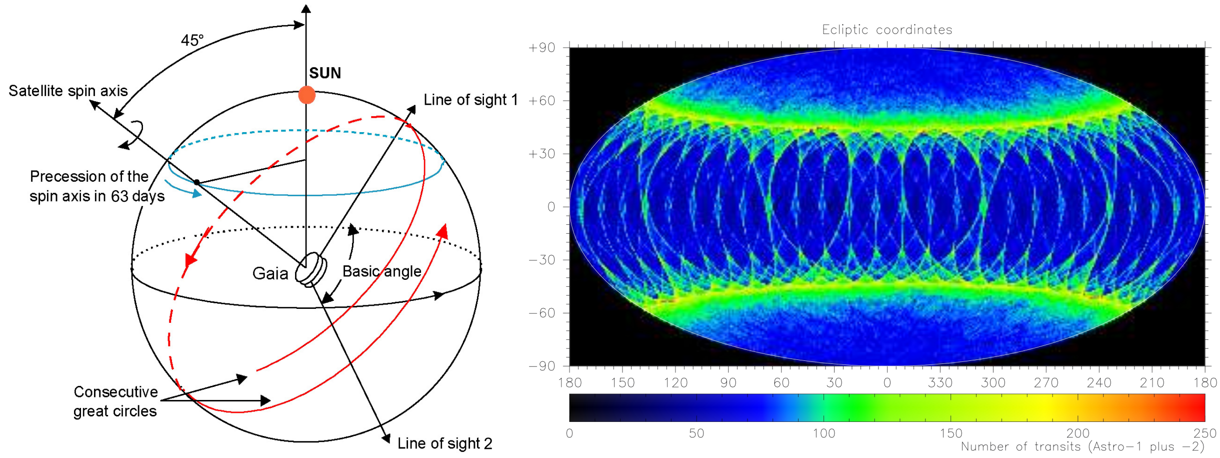


Figure 4: Left panel: During its 5-year operational lifetime, the satellite will continuously spin around its axis, with a constant speed of 60 arcsec s^{-1}. As a result, over a period of 6 hours, the two astrometric fields of view will scan across all objects located along the great circle "perpendicular" to the spin axis. *Gaia's* spin axis does not point to a fixed direction in space but is carefully controlled so as to precess slowly on the sky. As a result, the great circle that is mapped out by the two fields of view every 6 hours changes slowly with time, allowing repeated full sky coverage over the mission lifetime. Right panel: During its operational lifetime, *Gaia* will continuously scan the sky, roughly along great circles, according to a carefully selected pre-defined scanning law. The characteristics of this law, combined with the across-scan dimension of the astrometric fields of view, result in the above pattern for the distribution of the predicted number of transits on the sky in ecliptic coordinates. (Credit: J. de Bruijne, ESA)

1.1.3 Data processing and analysis

Although, there will be downloaded only small "windows" around the imaged sources, and not the whole CCD frames, the data that will be downloaded and the processing requirements for *Gaia* are among the most challenging. The combination of a large camera with the scanning law and a multi-year mission will result in an immense volume of data that will be collected. About 60 GB for nearly 600 million images are transmitted daily to three ground stations. The data that will be retrieved from the spacecraft during the total mission is estimated to be of 100 TB, so the total amount of processed and archived data will be approximately of 1 PB. The huge data volume generated during the mission requires a highly sophisticated system of software packages which run on computer clusters and supercomputers.

The responsibility of data processing is entrusted to a European consortium, the Data Processing and Analysis Consortium (DPAC) which consists of 400 expert scientists and engineers from around twenty countries, in order to produce the final objective of the *Gaia* catalogue. Coordinated by the DPAC Executive, the consortium is sub-divided into nine smaller, specialised units known as Coordination Units (CUs) with each unit being assigned a unique set of data processing tasks. The CUs draw their membership from multiple countries and are supported by the six Data Processing Centres (DPCs) (red dots in Figure 5, left panel). These are the centres at which the actual computer hardware for processing is available. The right panel in Figure 5 shows the flow of data within DPAC and the roles of the various CUs and DPCs.

In the course of the mission it is planned to produce some intermediate catalogues with the collected data. Those data require processing which is done by the DPAC team. So far, only one catalogue has been released, the Data Release 1 (DR1). The second data release is set for on April 2018, a third one is expected mid to late 2020, and the final data release in the end of 2022. Regarding the first data release, some limitations are expected. These limitations are related to the survey completeness, the astrometry, and the photometry. Some of them will be discussed in the Section 2.1.1.

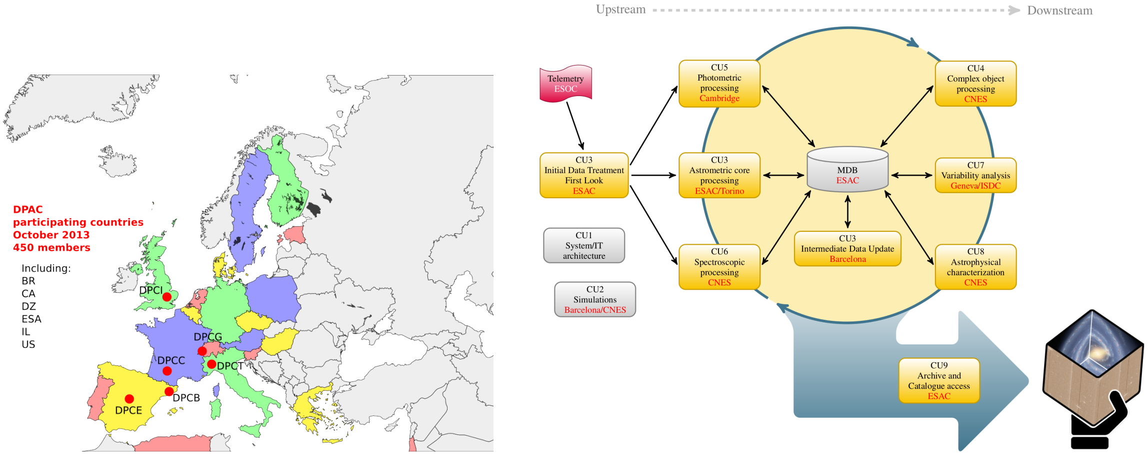


Figure 5: Left panel: DPAC membership map and Data Processing Centres (red dots). Right panel: Schematic overview of the way the data processing within DPAC is structured around the scientific units – CUs in charge of developing the algorithms for the data processing, and DPCs in charge of providing the hardware and the IT infrastructure to run the scientific algorithms. (Credit: ESA/*Gaia*)

A team of astronomers from Greece participates in DPAC since it was formed. Its main responsibility is the development of the Unresolved Galaxy Classifier (UGC) within the CU8 “Astrophysical Parameters”. UGC is a machine-learning based classifier and astrophysical parameters estimator for “unresolved” galaxies (too remote to be resolved in stars) observed by *Gaia’s* BP/RP spectrophotometer. The UGC is being implemented as a module of the ground-based pipeline “Apsis” (Bailer-Jones et al. 2013), operating in the DPC in Centre National d’Etudes Spatiales (CNES) in France. One of the front-end parameters that UGC estimates is the redshift of the galaxies. In addition to its scientific value in cosmological studies, it is important for the UGC itself because of its influence on the observed spectra. As a result, UGC uses the redshift estimator, internally, to simplify the classification and parameters regression of the spectra influenced by it. The results from UGC are planned to be included in the final *Gaia* archive.

1.2 Optical spectra of galaxies

The spectroscopy is a powerful diagnostic tool in astronomy. Analysing the emitted light from remote sources we can derive critical information about their physics and chemistry. The total energy and/or its integrated value in specific parts of the spectrum is used to characterize the sources (e.g. H-R diagram for stars). Particular features, spectral lines, either in absorption or emission, or bands, acts like “fingerprints” and their energy and profile provides more detailed information for the source, such as temperature, velocity, pressure, turbulent motions etc. On the other hand, the shifted position of these features with the respect to the laboratory, rest-frame spectrum, identifies the radial movement of the source toward or away from the observer.

In the cases of remote galaxies, their spectra reflect the content, the stellar population, the interstellar matter and dust, if any. One of the major factors influencing such a spectrum is the age and the star-forming history of the galaxy (Kennicutt 1992). Early-type galaxies (e.g. Elliptical), which already have been formed and contain mainly old population stars reveal spectra that resemble those of low mass main sequence stars, whereas late-type galaxies (e.g. Spiral or Irregular) undergo active star-forming demonstrated with characteristic intensive emission lines in their spectra, as seen in Figure 6.

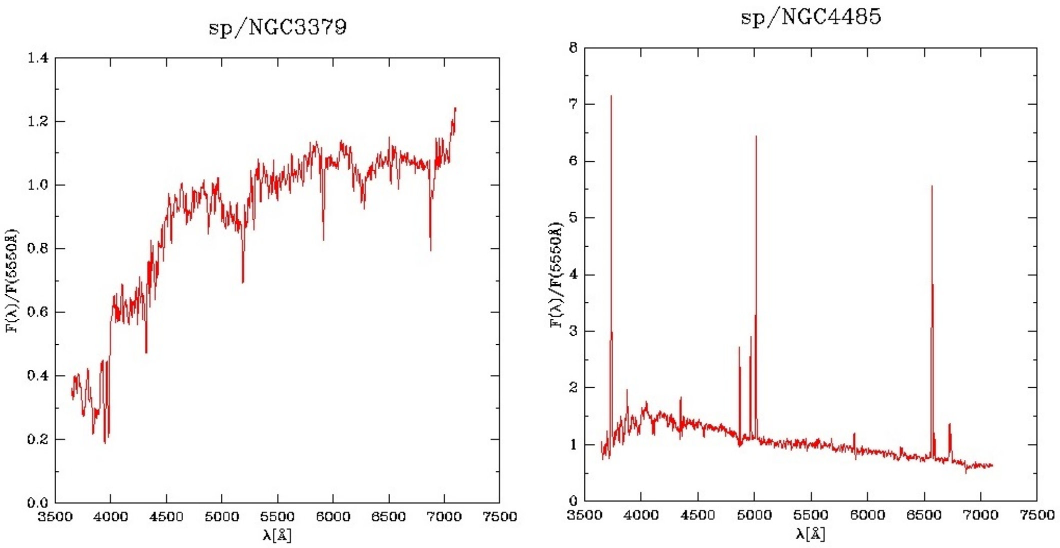


Figure 6: Optical spectra for NGC3379 galaxy (Left), which is an early-type galaxy (E1) containing almost old population stars and for NGC4485 (Right), which is a late-type galaxy (Irr) undergoing active star-formation. (Kennicutt 1992)

Another characteristic of the spectra of galaxies is the above mentioned displacement or "shift" concerning the wavelength (or frequency) scale. Because of the very large radial velocity of a distant galaxy its spectrum can differ significantly with the respect to the rest-frame one. The shifting because the Doppler effect "moves" normally not seen parts of the spectrum into the optical as well as "removes" other parts, causing a displacement of the spectral features (see Figure 7). This is well known as the observed "redshift" of distant objects due to the expansion of our Universe.

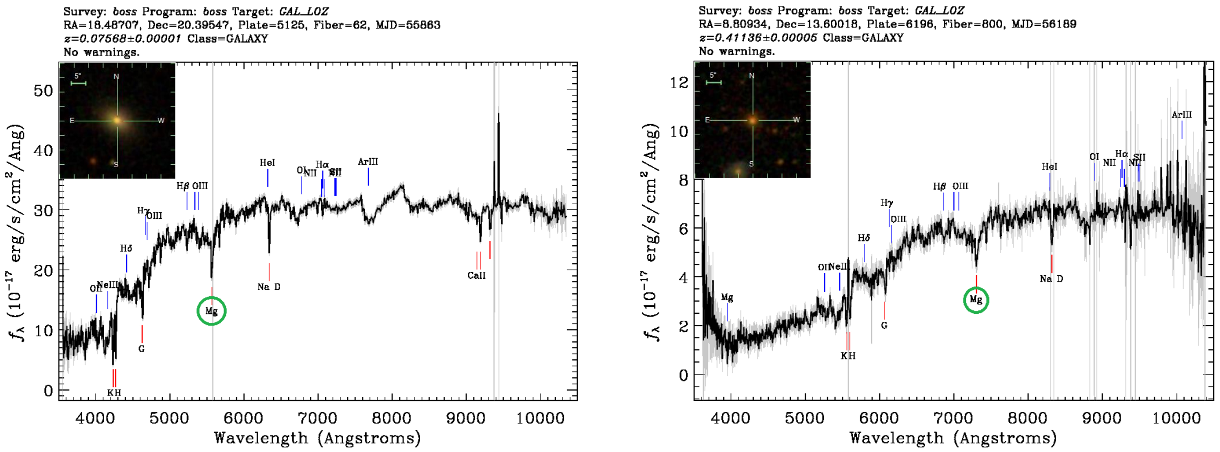


Figure 7: Spectra of galaxies from the SDSS's BOSS spectrograph in different redshifts. The first one (Left) has $z = 0.076$, while the second has $z = 0.411$, and it is evident the displacement of the absorption line Magnesium (Mg). Also, images of the corresponding galaxies are shown. The colour of a galaxy is change as we move to larger redshifts. (Credit: Sloan Digital Sky Survey)

1.2.1 Galaxies distance through redshift

As mentioned already, the shift of spectral lines from remote objects is known as redshift or blueshift. Most of the times astronomers observe redshifts, which implies that the object moves away from us, i.e. it occurs when the wavelength of an electromagnetic radiation, such as light, is shifted toward longer wavelengths. Redshifts are dimensionless quantities and they are defined as:

$$z \equiv \frac{\lambda_{obs} - \lambda_{th}}{\lambda_{th}} \tag{1}$$

where λ_{obs} is the observed wavelength and λ_{th} is the wavelength measured in the laboratory for a reference source at rest. The name of the z parameter, redshift, comes from the fact that the wavelength is shifted to the red end of the spectrum for all distant galaxies studied as early as in 1920s by the seminal observations of E. Hubble. The significance of this parameter is based on the fact that astronomers are able to detect or measure the distance and the relative velocity of remote objects.

In 1926 Edwin Hubble noted that almost all galaxies appeared to be moving away from us (Hubble 1926). So he came to the conclusion that the universe is expanding with all of the galaxies moving away from each other. This result had been achieved by studying the redshifts of galaxies's spectra. The velocity of a galaxy could be expressed mathematically as:

$$v = H * d \tag{2}$$

where v is the galaxy's radial outward velocity in km/s, d is the galaxy's distance from Earth in Mpc, and H is the constant of proportionality called the Hubble constant. The value of the constant is estimated by measuring the redshift of remote galaxies for which the distances are already known (e.g. by Cepheid variable, Supernovae). Uncertainties in the distances are causing varying estimates of the Hubble constant, which is currently estimated between 68 km/s/Mpc (Planck Collaboration et al. 2016) and 73 km/s/Mpc (Riess et al. 2018).

The redshifts appeared to be larger for fainter galaxies, i.e. more distant galaxies. Determining the redshift of galaxies for which their distance have been known, it has been found nearly linear dependence expressing numerically the above effect. The scaling parameter in this linear equation is the so-called Hubble constant and characterize the rate of the expansion of the Universe. So to determine an object's distance, we only need to know its velocity, which naturally follows from the redshift ($v = c * z$). Knowing the distance, taking into account the speed of the light and accepting the age of the Universe, we can measure how "old" is the deep-space object which image we are currently observing. A correspondence between the distance in time and the redshift is illustrated in Figure 8.

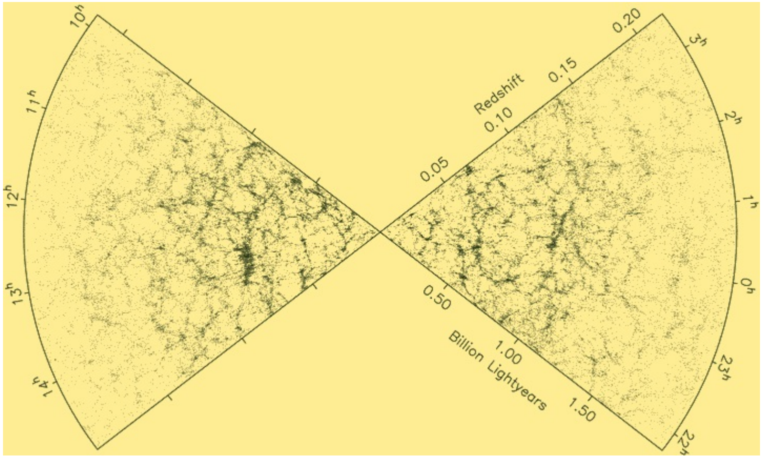


Figure 8: The redshift and the distance in the local Universe. (Credit: Two-degree-Field Galaxy Redshift Survey (2dFGRS))

1.3 Scope of the Thesis

The objective of our work is to prepare for the future survey of *Gaia* observed unresolved galaxies within the *Gaia* archive. The primary goal of this Thesis is to create an automated method for estimating the redshift of galaxies, based on low-dispersion *Gaia* BP/RP spectra, and to analyse its performance and applicability to be used as part of the UGC for the *Gaia's* ground-based pipeline. The low spectral resolution of *Gaia* does not allow us to resolve and identify well known spectral lines in order to derive the redshift of a galaxy. A very promising method is a machine-learning approach based on off-line training the estimator using spectra with known redshifts. Such an algorithm has already been implemented and tested in the prototype module. The challenge is that there are no available BP/RP observed spectra to be used for the purpose. A solution is to use either synthetic (modelled) spectra of galaxies (Tsalmantza et al. 2007) or real observed ones available in published surveys with already known redshifts. In both cases, the spectra shall be "converted" to *Gaia's* BP/RP using the appropriate instrument model. The synthetic spectra approach had been already implemented in the prototype, but it has been limited to small redshifts whereas it is now essential to expand the estimator to larger ones. In this work, we will use observed spectra of galaxies and to model them for BP/RP. The creation of such an "empirical library" of extragalactic sources of spectra will be an additional result towards the scope of the Thesis. Moreover, it will be used by another module (Discrete Source Classifier) of CU8 that is responsible to initially separate the different type of sources observed by *Gaia*.

In Section 2 we describe how we determined which of the billion of *Gaia's* observed sources are galaxies that are also provided in an applicable survey of optical spectra. Section 3 presents our methodology to combine the survey data and the list of galaxies observed by *Gaia* to create a library of BP/RP modelled spectra appropriate for this work. In Section 4 the development of the redshift estimator is presented, the training and testing, and an analysis of its performance. Finally, Section 5 summarizes the conclusions and we discuss possible future work.

2 Galaxies observed by *Gaia*

Gaia is a “billion star surveyor”, but because of its whole-sky observing nature, it will observe, as we already mentioned, extragalactic sources as well. As a first step of this work is to identify which sources observed by *Gaia* are galaxies. The currently available archive, *Gaia* DR1[2], is the base for this. This must be compared to another large-scale sky survey providing information, such as classification of the sources. The survey that has been used for this purpose is the Sloan Digital Sky Survey (SDSS), using the archive release SDSS DR13[3]. A simple cross-match using the coordinates of all *Gaia's* sources with galaxies in the SDSS is required to perform this task.

2.1 Archives used

2.1.1 *Gaia* DR1

Gaia's first data release (DR1) was published on 14 September 2016 (Gaia Collaboration et al. 2016a). DR1 contents positions and *Gaia* G magnitudes (`GaiaGmag`) for a billion stars and few millions of other sources that have been observed by *Gaia*. Additionally, the catalogue provides detailed astrometric information for about 2 million stars, light curves and characteristics for about 3000 variable stars, and positions and G magnitudes for more than 2000 quasars. The base catalogue in the archive contains 1,142,461,316 sources and its size is almost 170 GB, with more than five thousands files. This catalogue has been downloaded for the purpose of the cross-matching and parameters not essential for this task have been removed to reduce its size.

A remarkable note, about DR1, is that this release constituted by some limitations, as mentioned in Section 1.1.3. The first data release is not a complete survey, as described in the archive's site. First of all, the source list for the release is incomplete especially at the bright end and has an ill-defined faint magnitude limit, which depends on celestial position. Moreover, sources close to bright objects, as well extremely blue and red sources are missing. What concerns this thesis is that *Gaia* is optimized to observe point-like sources, so it does not observe galaxies that are extended. Also, the accuracy of the data obtained for the sources are influenced by the satellite scanning law – the number of transits on the sky depends on the position (see Figure 4). However, the astrometry and photometry is more than sufficient for the purposes of our cross-matching with the SDSS. Nevertheless, it should be noted that this release does not provide spectra neither from the BP/RP spectrophotometer nor from the RVS spectrograph. Spectra are expected in mid to late 2020, with the release of DR3.

2.1.2 SDSS DR13

SDSS is a major multi-spectral imaging and spectroscopic redshift survey using a dedicated 2.5 m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States, operating for over 15 years (Albareti et al. 2017). From 1998–2009 it observed in both imaging and spectroscopic mode, while later the telescope is used entirely in spectroscopic mode (Dawson et al. 2013). The spectrograph operates by feeding an individual optical fiber for each target through a hole drilled in an aluminium plate (Newman et al. 2004). Each hole is positioned specifically for a preselected target, so every field in which spectra are to be acquired requires a unique plate. In spectroscopic mode, the telescope tracks the sky in the standard way, keeping the objects focused on their corresponding fiber tips. Data collection began in 2000, and the final imaging data release covers over 35% of the sky, with photometric observations of around 500 million objects.

[2] gea.esac.esa.int/archive/

[3] www.sdss.org/dr13/

SDSS regularly publish the currently available data. In this work the DR13 is used. It is the first data release of the fourth phase, SDSS-IV (Blanton et al. 2017), and it was made available in 2016 July. DR13 includes the complete dataset of imaging and optical spectroscopy through July 2014 and July 2015 – more than four million sources from the SDSS's two optical spectrographs (legacy SDSS and BOSS). Access to the DR13 is granted through the Catalogue Archive Server (CAS) (Thakar 2008) via two primary modes: browser-based queries of the database are available through the SkyServer Web application in synchronous mode, and more advanced and extensive querying capabilities are available through the Catalogue Archive Server Jobs System or CasJobs[4] in asynchronous or batch mode that allows time-consuming queries to be run in the background (Li & Thakar 2008). The CAS is now part of the new SciServer collaborative science framework. Through CasJobs the necessary data for our work have been downloaded.

2.2 Cross-match and multiplicity

In order to find which sources in the *Gaia* DR1 are galaxies, we used a specific table of SDSS DR13 – SpecObj, which contains spectroscopic information for all objects with clean spectra. Using CasJobs (see Appendix A-1) only sources that have been classified as type "GALAXY" or "QSO" have been selected. 3,051,554 sources have been downloaded along with few parameters, such as positions. Their distribution by equatorial coordinates is presented in Figure 9, left panel. From now on we will mention those sources only as "galaxies". The *Gaia's* DR1 sources have been cross-matched with these galaxies. For this procedure a limit box of ±0.0005° (1.8″ which is close to fiber's size) in both right ascension and declination differences between *Gaia's* and SDSS's sources position has been used. A non-exclusive (all the matches counted) cross-match of SDSS galaxies against *Gaia* sources has been performed with a program of UGC and the results are presented in Table 1.

Table 1: Results of a non-exclusive cross-match of SDSS DR13 galaxies against *Gaia* DR1 sources.

Total sources *Gaia* DR1	SDSS DR13 galaxies (GALAXY & QSO)	*Gaia* matching SDSS	SDSS matched by *Gaia*
1,142,461,316	3,051,554	534,234	531,179

As it can be seen there are cases (∼ 1% of the total) where more than one *Gaia* source is matched to a single SDSS galaxy. Such a multiplicity in the matching has been analysed (see Table 2) and the most obvious reason seems to be the combination of the scanning law and the orientation of the CCD camera. For extended sources, like galaxies, and due to the CCD characteristics (orthogonal pixels, binning, etc.) and the scanning law, the processing system provides slightly different center coordinates for the source in different transits. To address this type of multiplicity, the brightest estimation of the *Gaia* source has been selected as the representative one.

Table 2: Multiplicity in *Gaia* sources matching SDSS galaxies.

Multimatch	*Gaia* sources	SDSS galaxies
single	525,383	525,383
double	5,894	2,947
triple	117	39
quadruple	4	1

[4] http://skyserver.sdss.org/casjob

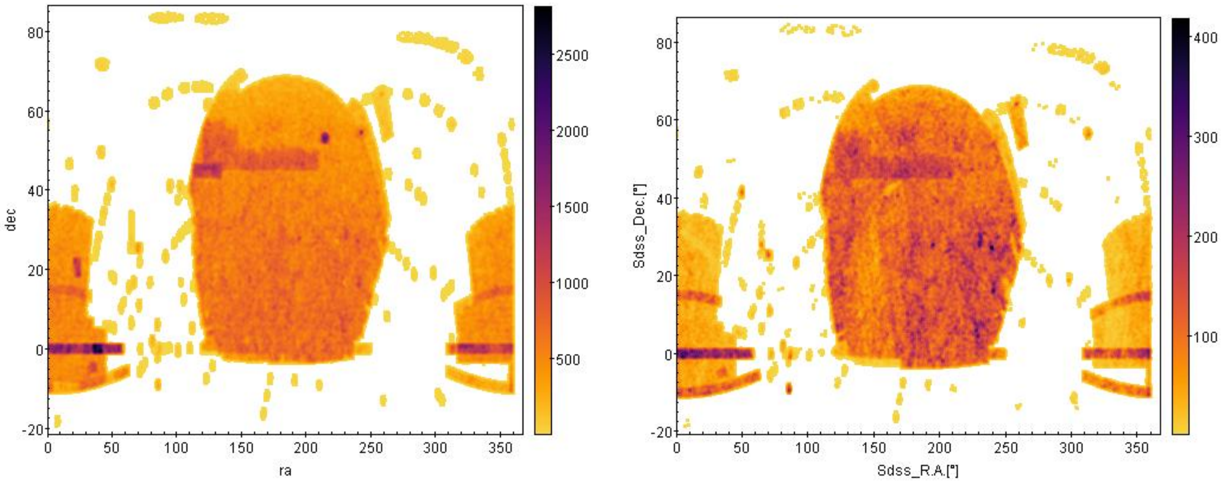


Figure 9: Left panel: Distribution on the sky of the complete SDSS source catalogue with available spectra (SpecObj table). Right panel: Fraction of the sources presented in the left panel that have been detected with *Gaia*.

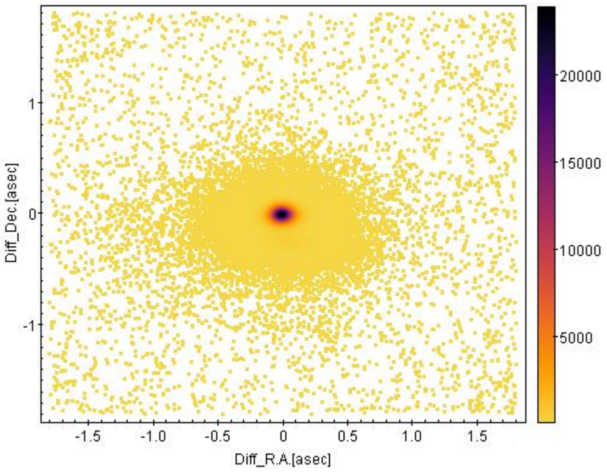


Figure 10: Distribution of the differences in right ascension and declination between the SDSS and *Gaia* for the matched galaxies.

The distribution in equatorial coordinates of the SDSS galaxies matched by *Gaia* is shown in the right panel of Figure 9. The comparison with the distribution of all the SDSS galaxies (left panel) shows no evident difference except, of course, the total number of sources. The difference in coordinates between the SDSS and *Gaia* matched sources is presented in Figure 10. With the exception of a relatively small number of sources, all differences are within a very small box of $\pm 0.1''$. So, if there are any errors in cross-matching, these are negligible, and the selected *Gaia* sources are definitely galaxies. Since SDSS covers about $^1\!/_3$ of the sky and *Gaia* observes the whole sky we can estimate that nearly 1.5 million of galaxies are observed by the satellite. Of course, the number is expected to be even larger as the base for matching was the table of SDSS galaxies being observed spectroscopically.

3 Spectral Library of *Gaia* galaxies

In order to create the redshift estimator (see Section 4) a set of *Gaia* BP/RP spectra of galaxies with known redshifts is necessary. As it has been pointed out, no such spectra are currently released, and they are expected to be published with the *Gaia* DR3 in about two years. A solution is to use spectra, from another catalogue, which can be simulated for the *Gaia's* spectrophotometer. We chose to use observed spectra of galaxies from the SDSS archive, the DR13. The choice was based on the fact that SDSS provides reliable redshifts and spectral classification of the extragalactic sources, which is very important since we want to deal only with the study of galaxies. The SDSS archive contains spectra obtained by two spectrographs, the Legacy which was used in SDSS I/II and a new one, the BOSS. The latter extended significantly the wavelength range recorded, bringing it quite close to the *Gaia's* BP/RP range (see Figure 11). It is natural thus to use for our purpose only the BOSS spectra.

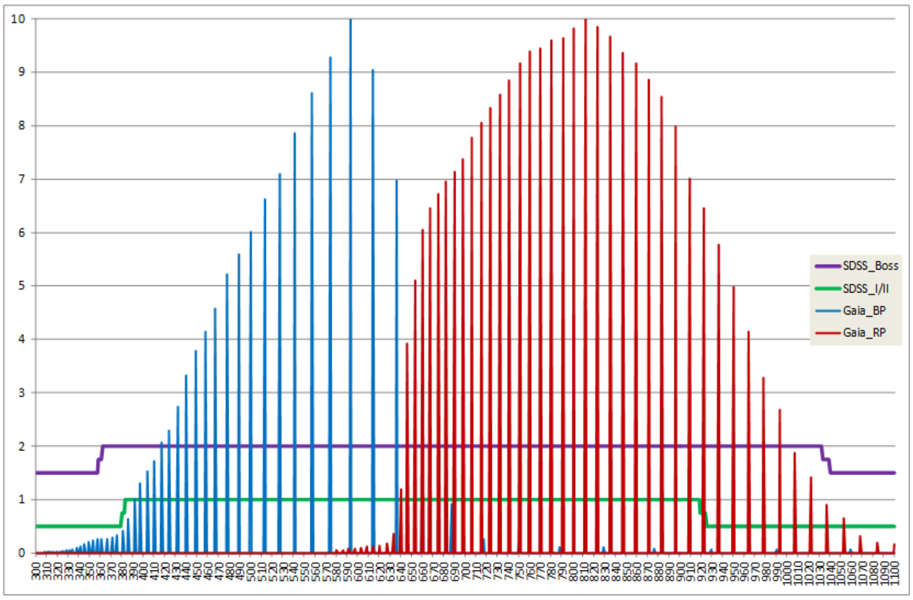


Figure 11: Wavelength coverage for *Gaia's* BP/RP spectrophotometer and the SDSS's spectrographs (BOSS and Legacy). The vertical blue and red lines indicate the exact locations (and throughput) of BP and RP wavelength sampling, while the purple line represents the BOSS's range, which covers almost the whole, the *Gaia's* range. Also, the green line is for Legacy spectrograph.

The Baryon Oscillation Spectroscopic Survey (BOSS) has been in operation since Fall 2009 and its first results were released in DR9. BOSS is an SDSS-III (2008-2014) spectroscopic survey mapping the clustering of galaxies and intergalactic gas in the distant universe, and measuring redshifts of galaxies and high redshift quasars. Its two identical spectrographs were rebuilt from the original SDSS's spectrographs. Each spectrograph has two cameras, one red and one blue, with a dichroic splitting the light at roughly 6,000 Å(the blue cameras cover 3,600-6,350 Å, whereas the red cameras cover 5,650-10,000 Å) and a resolution $R \sim 2,000$. Specifically, the blue channel of the spectrograph has resolution $R \sim 1,560 - 2,270$, while the red channel has $R \sim 1,850 - 2,650$. BOSS was the only extragalactic survey in SDSS-III that targeted both galaxies and quasars. An important upgrade from the original survey was the new optical fibers (1,000 rather than 640 per plate of Legacy) and a smaller aperture holes (2″ instead of 3″).

3.1 Source Library of *Gaia* observed galaxies

To use the BOSS spectra we have to select the galaxies observed also by *Gaia*. Similarly, as this has been done in Section 2.2, we downloaded the file with the necessary parameters for the SDSS DR13 BOSS data using CasJobs (see Appendix A-2.1). Then we cross-matched (non-exclusively) the tables from the two archives, the SDSS DR13 BOSS galaxies against the *Gaia* DR1 sources. From the first archive we used the spectroscopic parameters listed in the table SpecObj, which contains 1,858,658 galaxies. As for the data from the *Gaia* archive, the DR1 was used. Cross-matching between the two archives has been performed based on the coordinates exactly as before (limit box of ±0.0005°). The results are in Table 3, where we observe again that more than one *Gaia* source is matching a single SDSS galaxy. The results of multiplicity analysis are shown in Table 4. We found that in all 500 cases of double match there are indeed double sources, as shown for two cases in the Figure 12. Therefore, we have rejected all these double sources from further processing. The above decision leaves us with a table of 127,050 galaxies observed by *Gaia* for which we can access their BOSS spectra.

Table 3: Results of a non-exclusive cross-match of SDSS DR13 BOSS galaxies against *Gaia* DR1 sources.

Total sources *Gaia* DR1	SDSS DR13 BOSS galaxies (GALAXY & QSO)	*Gaia* matching SDSS	SDSS matched by *Gaia*
1,142,461,316	1,858,658	127,550	127,300

Table 4: Multiplicity in *Gaia* sources matching SDSS BOSS galaxies.

Multimatch	*Gaia* sources	SDSS BOSS galaxies
single	127,050	127,050
double	500	250

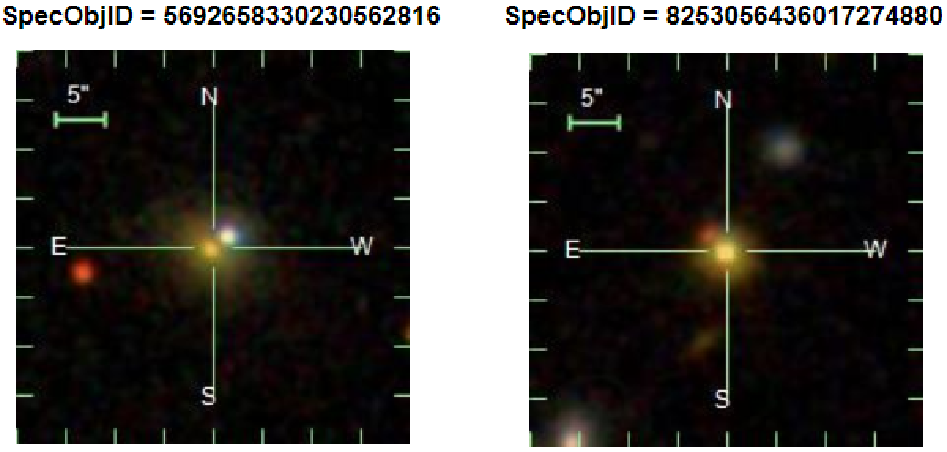

SpecObjID = 5692658330230562816
5"
N
E
W
S
SpecObjID = 8253056436017274880
5"
N
E
W
S

Figure 12: Examples of double sources observed by SDSS, which have been rejected.

Checking the table which had been created after the cross-matching and the multiplicity we concluded that we will need more parameters to the already existing ones from the SDSS archive. The new parameters will be spectroscopic and photometric from the SpecObj and the PhotoObj tables, respectively. The new query for the CasJobs is shown in Appendix A-2.2. Consequently, inspecting the resulting only-single-matches extended table we found erroneous data for a number of sources. Especially we looked for the parameters required for our task, which are listed in Table 5. There are about 700 cases (see Table 6) for which the corresponding source records have been removed.

Table 5: Necessary parameters from both catalogs (*Gaia*, SDSS) with their description.

Origin		Parameter	Description
SDSS	Photometric(1)	modelMag(2),(3)	The model magnitude, which uses the best of the De Vaucouleurs and Exponential model fits as a matched aperture to calculate the flux in all bands
		modelMagErr(2)	Error in better of De Vaucouleurs/Exponential magnitude fit
		petroR50(2),(3)	Radius containing 50% of Petrosian flux
		petroR50Err(2)	Error in radius with 50% of Petrosian flux error
		petroR90(2)	Radius containing 90% of Petrosian flux
		petroR90Err(2)	Error in radius with 90% of Petrosian flux error
	Spectroscopic	z(2)	Final Redshift
		zErr(2),(3)	Redshift error
		class	Spectroscopic class (GALAXY, QSO, or STAR)
		subClass	Spectroscopic subclass
		specObjID	Unique database ID
		plate	Plate number
		mjd	MJD of observation
		fiberID	Fiber ID
		snMedian(2),(3)	Median signal-to-noise over all good pixels
		ra	DR8 Right ascension of fiber, J2000
		dec	DR8 Declination of fiber, J2000
Gaia		GaiaGmag(3)	The *Gaia* magnitude from DR1

(1) The SDSS has a photometric system of five-bands – u,g,r,i,z-band – however, it is preferable to use mostly the r-band as it is closer to the *Gaia's* magnitude – GaiaGmag.
(2) Parameters checked for erroneous values.
(3) Parameters used for range limitation.

Table 6: Clearing sources with erroneous data. In these data no photometry have been performed.

Parameters	Wrong values	Remarks	Number of cases
modelMag_*, modelMagErr_*, petroR50_r, petroR50Err_r, petroR90_r, petroR90Err_r	null	no measurement	671
modelMag_*, modelMagErr_*, petroR50_r, petroR90_r	-9999	bad value code	12

Using the TopCat (Taylor 2005) tool, which is an interactive graphical viewer and editor for tabular data written in Java, we analysed the values of most of the parameters, listed in Table 5. The distribution of these parameters is presented in Figure 13. The histogram of modelMag_r shows that there are still sources with questionable faint magnitudes making these sources inappropriate. Moreover, in both, *Gaia* and SDSS magnitudes distributions there are quite bright sources that would yield overexposed BP/RP spectra. From the histogram of redshift it is evident that only quasars in SDSS are observed at larger redshifts, as expected. On the other hand, the petrosian radius (both `petroR50_r` and `petroR90_r`), distributions reveal persistence of quite extended galaxies, although the number of such cases is quite small. Finally, there are sources with unrealistic values of `snMedian`, which should also not be used.

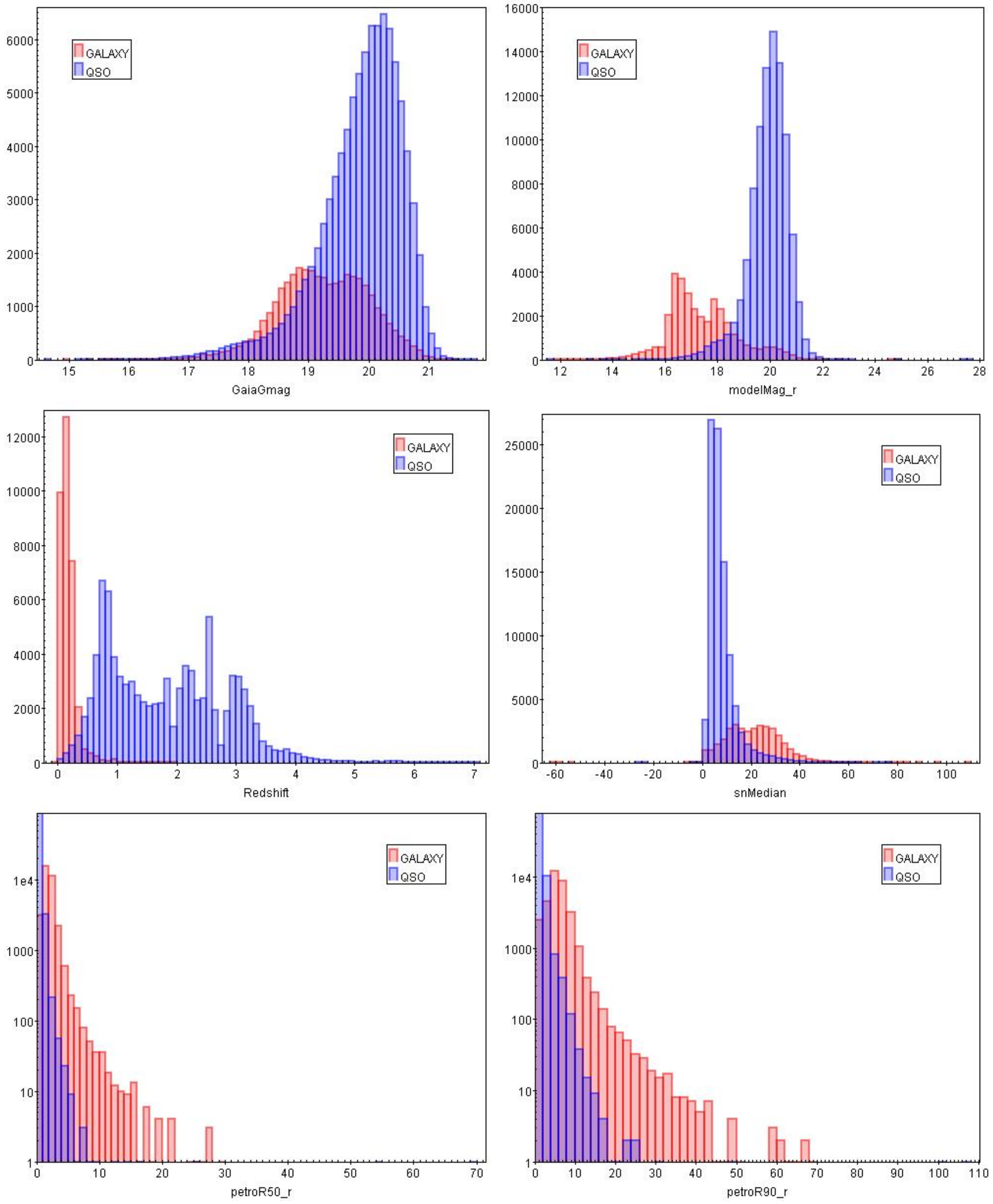


Figure 13: The diagrams show the distribution of the difference between the GALAXY (red bins) and QSO (blue bins) sources as a function of various parameters. Top left panel: *Gaia's* magnitude. Top right panel: Model magnitude in r-band of SDSS. Center left panel: Redshift. Center right panel: Signal-to-noise. Bottom left panel: Petrosian radius 50% of the flux. Bottom right panel: Petrosian radius 90% of the flux.

It is interesting to note the appearance of two clearly separated distributions in the modelMag_r corresponding to GALAXY and QSO sources which are not evident in the GaiaGmag. We investigated the difference between the SDSS and *Gaia* magnitudes which showed more clearly the separation (see Figure 14). The *Gaia* magnitude for GALAXY sources are systematically fainter than SDSS. This is a natural consequence of the windowing policy used in *Gaia* observations. A restricted size window is formed around a source within which the flux is measured. So, for the more extended galaxies a larger portion of its flux is lost and the magnitude is underestimated.

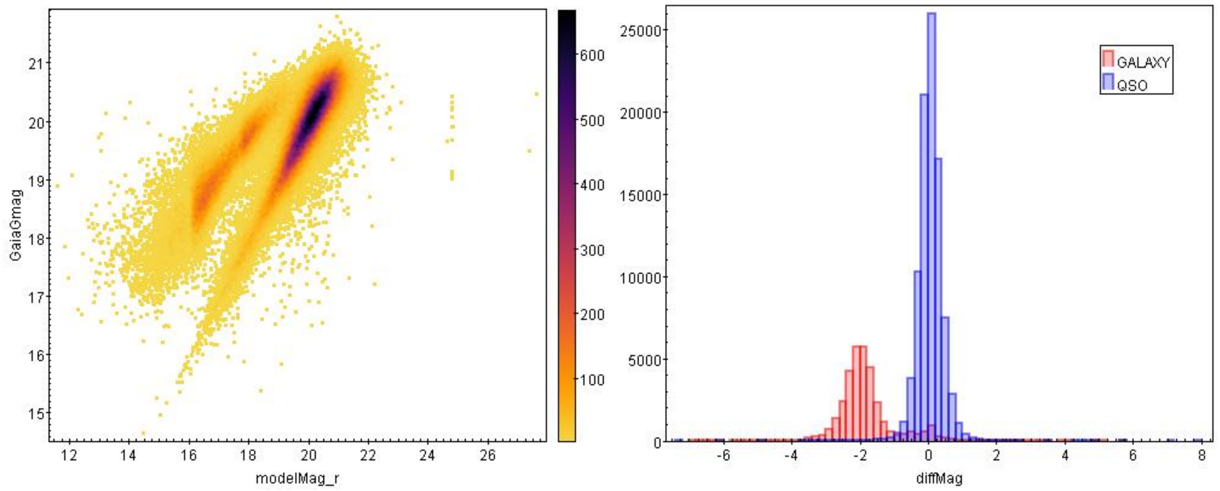


Figure 14: Left panel: Plot of the SDSS and *Gaia* magnitude, showing an interesting correlation. The one sequence represents the GALAXY sources, where the other represent the QSO sources. The GALAXY sources are fainter for *Gaia* due to the windowing policy of observations. Right panel: Distribution of the difference between the GALAXY and QSO sources for the difference of magnitudes between the model magnitude in the r-band with the *Gaia's* magnitude. In this distribution the difference between the mentioned types is more apparent.

The above analysis allowed us to set specific limitations on five of the parameters. The limits on the *Gaia* and SDSS magnitudes are necessary to avoid very bright and very faint sources in both archives. We adopted $2''$ limit for the petrosian radius (petroR50_r) in order to allow slightly extended but still unresolved galaxies. To avoid very noisy spectra we accepted a signal-to-noise ratio (snMedian_r) above 3. Finally, the redshift error must not be large since this parameter will be used later as target in the training. After applying these limits the created table, called from now on "Source Library", consists of 84,967 galaxies (18,245 GALAXY and 66,722 QSO type). The resulting new distributions with the respect to the previous, for two magnitudes and the redshift, are presented in Figure 15.

Table 7: Limitations on five of the parameters.

Limitations
$16 \leq \texttt{GaiaGmag} \leq 20.5$
$16 \leq \texttt{modelMag\_r} \leq 20.5$
$\texttt{petroR50\_r} \leq 2$
$\texttt{snMedian\_r} > 3$
$0 \leq \texttt{zErr} \leq 0.01$

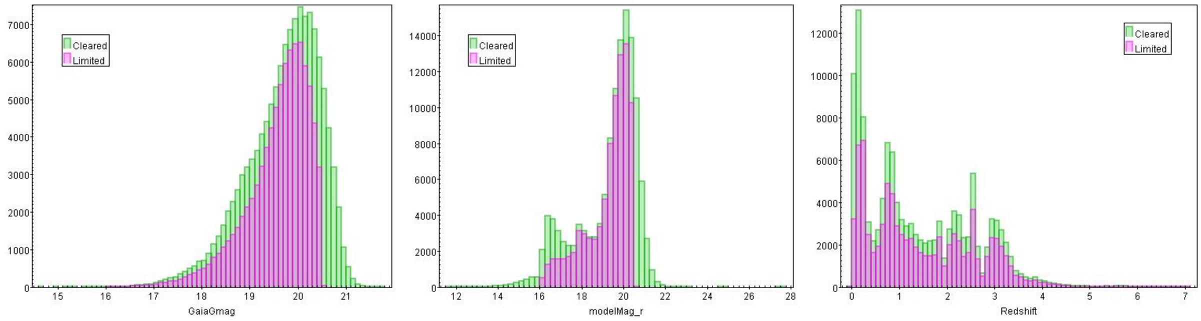


Figure 15: Distributions of the difference between the cleared (green bins) and limited (purple bins) data in terms of the various parameters. Left panel: *Gaia's* magnitude. Center panel: Model magnitude in r-band of SDSS. Right panel: Redshift.

3.2 SDSS spectra for *Gaia* galaxies

The Source Library is the base for the next step, the creation of “Spectral Library” of SDSS spectra for galaxies observed by *Gaia* modelled for its BP/RP instrument. Analysing the distribution of the sources by redshift in the library (see Table 8) it is obvious that above $z = 0.6$ the number of GALAXY sources rapidly drops whereas for the QSO the opposite is true. This can be clearly seen in the histogram of the redshift for the Source Library (see Figure 16) which presents the lower part of the histogram from Figure 13 in center left panel.

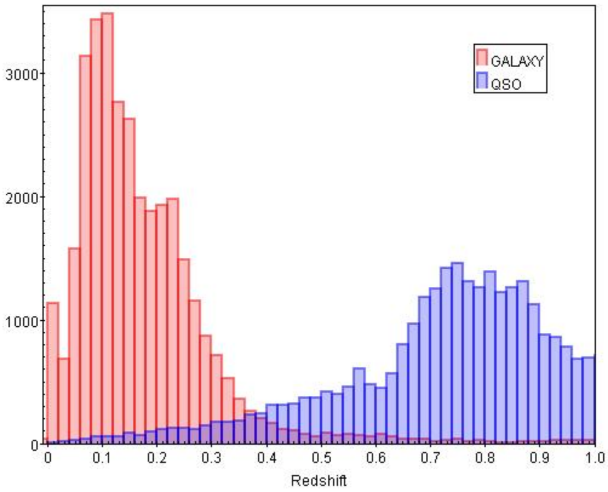


Figure 16: Distribution of the difference between the GALAXY (red bins) and QSO (blue bins) sources in terms of the redshift. Most of the normal galaxies lays in smaller redshifts in contrast with the quasars.

We have to note that the redshift estimator in the UGC prototype was designed for redshifts up to 0.2. Moreover, the UGC has been trained using synthetic models of normal galaxies (Karampelas et al. 2012). The Source Library lists observed by *Gaia* galaxies, for which SDSS spectra are available and can be modelled for BP/RP. Using such spectra, instead of synthetic ones, is more efficient to prepare the redshift estimator. Although UGC is dedicated to deal with normal galaxies, it is expected that it will encounter also quasars in the pipeline. So, it is essential to create a new redshift estimator for UGC to deal both with larger range of redshifts and different extragalactic sources. The Spectral Library, which will be discussed below, is the necessary foundation for this.

Table 8: Table from the Source Library. Illustrating their distribution with respect to redshift. In the Source Library there are also sources with negative redshifts. The number of sources we used as shown in column "Usable sources". According to the SDSS classification we can categorize the sources into subclasses (see below paragraph and Table 9). Also, each type contains sources that had been recognized as broadlines.

Type of	Number of			Redshift				Usable sources
sources	sources	0.0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0	≥ 1	0.0-0.6
GAL	16864	8666	7499	347	128	75	75	16512
G_SBR	303	121	155	25	0	0	0	301
G_SFR	562	405	136	21	0	0	0	562
G_AGN	470	340	112	18	0	0	0	470
Total GAL	18199	9532	7902	411	128	75	75	17845
QSO	65142	142	599	2496	7717	7205	7205	3237
Q_SBR	1412	123	733	556	0	0	0	1412
Q_SFR	47	10	16	21	0	0	0	47
Q_AGN	167	34	74	59	0	0	0	167
Total QSO	66768	309	1422	3132	7717	7205	7205	4863
Total	84967	9841	9324	3543	7845	7280	7280	22708

Limiting the Source Library to $0.0 \leq z \leq 0.6$ we have 22,708 galaxies (17,845 GALAXY and 4,863 QSO) (see Table 8). Although we do not use the extragalactic sources subclasses (STARFORMING, STARBURST, and AGN, see Table 9) provided by SDSS the table also presents the distribution by them. Moreover, if any galaxy or quasar has lines detected at the 10-sigma level with sigmas > 200 km/sec at the 5-sigma level, the indication "BROADLINE" is appended to their subclass. In Table 8 the types of sources include the sources with the corresponding broadline subclasses.

Table 9: The categorization of data into subclasses according to SDSS classification for galaxies and quasars.

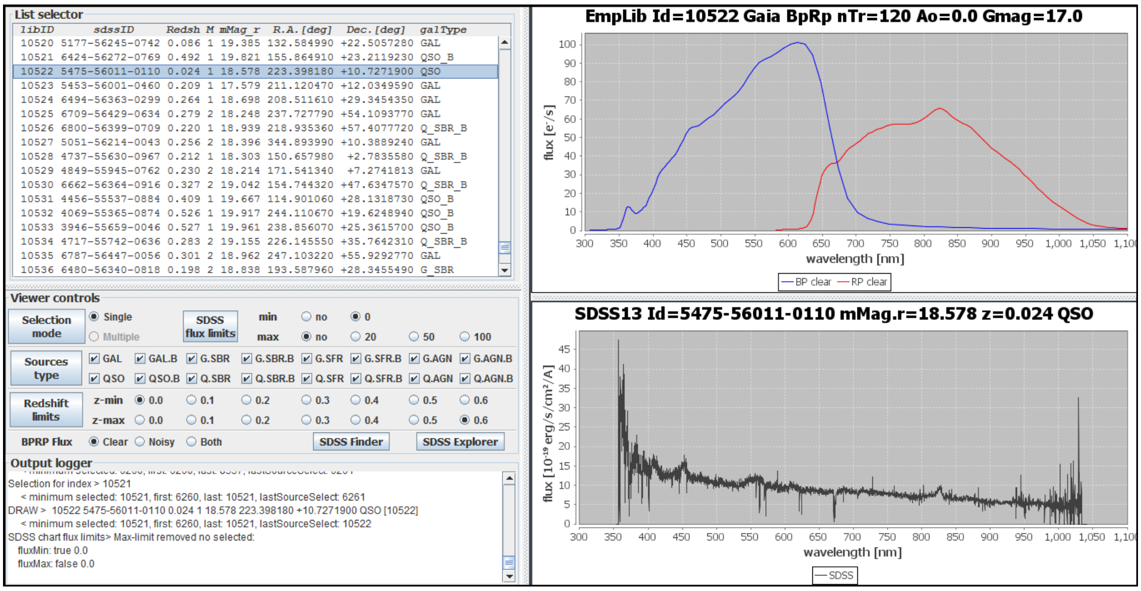
Subclasses	Description
STARFORMING	set based on whether the galaxy has detectable emission lines that are consistent with star-formation according to the criteria: $\log_{10}\left({}^{OIII}/_{H\alpha}\right) < 0.7 - 1.2\left(\log_{10}\left({}^{NII}/_{H\alpha}\right) + 0.4\right)$
STARBURST	set if the galaxy is star-forming but has an equivalent width of Hα greater than 50 Å
AGN	set based on whether the galaxy has detectable emission lines that are consistent with being a Seyfert or LINER: $\log_{10}\left({}^{OIII}/_{H\alpha}\right) > 0.7 - 1.2\left(\log_{10}\left({}^{NII}/_{H\alpha}\right) + 0.4\right)$

For the Spectral Library we selected almost half of the 22,708 sources, with respect to the photometry quality. In the selection we attempted to preserve the relative distribution of redshift and subclasses. We ended up with a subset of 11,102 sources for which we retrieved their SDSS BOSS spectra. The command to download is shown in Appendix A-3. The spectra wavelength units are in Å and the fluxes in $10^{-17} ergs\ cm^{-2} s^{-1} \mathring{A}^{-1}$. The logarithmic wavelength grid spacing is the same for all the spectra ($\log_{10}\lambda_{i+1} - \log_{10}\lambda_i = 0.0001$) but the starting wavelength differs from spectrum to spectrum. In our Spectral Library, to unify all the spectra, we expanded the wavelength logarithm scale from 3.5494 to 4.0173 (with zero fluxes) since it is necessary for each spectrum to cover a common wavelength range. This common grid includes 4,680 pixels. The spectra are accompanied by a table of corresponding parameters, subset of the Source Library.

3.3 Library of *Gaia* BP/RP simulated spectra of galaxies

Subsequently, we are going to model the selected spectra from the SDSS BOSS spectrograph for the *Gaia's* BP/RP spectrophotometer. A simulator, the Ulysses[5], has been used for this task. The Ulysses module is a simple BP/RP simulator written specifically to meet the needs of CU8 to quickly model end-of-mission spectra of sources as seen by *Gaia*. Ulysses simulates BP/RP spectra by convolving the input spectra from 300 to 1,100 nm with the Line Spread Function (LSF) of the *Gaia* optical instruments. The convolution is performed for every pixel in the BP/RP spectra. The transmissivity of the mirrors, filters, and prisms are also taken into account.

In our case, the input for the simulator are the SDSS's spectra, either as FITS or ASCII files, while the output will be the *Gaia's* BP/RP simulated spectra in *Gaia* compressed format, Gbin files. We provided values for the required parameters for simulation: the `GaiaGmag=17`, `numberofTransit=120`, and `oversampling=2` (which means that we have $2 * 60 = 120$ pix per BP/RP). Ulysses provides noiseless or noisy spectra and we used `numberofNoisySpectra=1`, which requires $\sigma = 1$ noise to be applied. The modelled spectrum will be in $photons\ cm^{-2}s^{-1}nm^{-1}$ and the wavelength scale is one and the same for all the spectra and gives the wavelength in nm for each sub-pixel. The resulting library which combines the input SDSS spectra the corresponding parameters and the modelled BP/RP spectra is the "Empirical Library" of the spectra of galaxies observed by *Gaia*.



List selector

```
libID     sdssID         Redsh M mMag_r  R.A.[deg]  Dec.[deg]   galType
10520 5177-56245-0742 0.086 1 19.385 132.584990 +22.5057280 GAL
10521 6424-56272-0769 0.492 1 19.821 155.864910 +23.2119230 QSO_B
10522 5475-56011-0110 0.024 1 18.578 223.398180 +10.7271900 QSO
10523 5453-56001-0460 0.209 1 17.579 211.120470 +12.0349590 GAL
10524 6494-56363-0299 0.264 1 18.698 208.511610 +29.3454350 GAL
10525 6709-56429-0634 0.279 2 18.248 237.727790 +54.1093770 GAL
10526 6800-56399-0709 0.220 1 18.939 218.935360 +57.4077720 Q_SBR_B
10527 5051-56214-0043 0.256 2 18.396 344.893990 +10.3889240 GAL
10528 4737-55630-0967 0.212 1 18.303 150.657980  +2.7835580 Q_SBR_B
10529 4849-55945-0762 0.230 2 18.214 171.541340  +7.2741813 GAL
10530 6662-56364-0916 0.327 2 19.042 154.744320 +47.6347570 Q_SBR_B
10531 4456-55537-0884 0.409 1 19.667 114.901060 +28.1318730 QSO_B
10532 4069-55365-0874 0.526 1 19.917 244.110670 +19.6248940 QSO_B
10533 3946-55659-0046 0.527 1 19.961 238.856070 +25.3615700 QSO_B
10534 4717-55742-0636 0.283 2 19.155 226.145550 +35.7642310 Q_SBR_B
10535 6787-56447-0056 0.301 2 18.962 247.103220 +55.9292770 GAL
10536 6480-56340-0818 0.198 2 18.838 193.587960 +28.3455490 G_SBR
```

Viewer controls

Selection mode: ● Single ○ Multiple

SDSS flux limits: min ○ no ● 0; max ● no ○ 20 ○ 50 ○ 100

Sources type: ☑ GAL ☑ GAL.B ☑ G.SBR ☑ G.SBR.B ☑ G.SFR ☑ G.SFR.B ☑ G.AGN ☑ G.AGN.B; ☑ QSO ☑ QSO.B ☑ Q.SBR ☑ Q.SBR.B ☑ Q.SFR ☑ Q.SFR.B ☑ Q.AGN ☑ Q.AGN.B

Redshift limits: z-min ● 0.0 ○ 0.1 ○ 0.2 ○ 0.3 ○ 0.4 ○ 0.5 ○ 0.6; z-max ○ 0.0 ○ 0.1 ○ 0.2 ○ 0.3 ○ 0.4 ○ 0.5 ● 0.6

BPRP Flux ● Clear ○ Noisy ○ Both

Output logger

```
Selection for index > 10521
   < minimum selected: 10521, first: 6260, last: 10521, lastSourceSelect: 6261
DRAW >  10522 5475-56011-0110 0.024 1 18.578 223.398180 +10.7271900 QSO [10522]
   < minimum selected: 10521, first: 6260, last: 10521, lastSourceSelect: 10522
SDSS chart flux limits> Max-limit removed no selected:
  fluxMin: true 0.0
  fluxMax: false 0.0
```

Figure 17: The visualization tool for the Empirical Library. On the left side is the control panel for selecting sources. On the right side the bottom panel presents the SDSS spectrum from the Spectral Library, while on the top is plotted the simulated BP/RP spectrum from the Empirical Library. Through the visualization tool we have the ability to determine the limit to the shown sources of the SDSS fluxes and the redshifts. Also, we can choose which types we want to see and whether the BP/RP spectra will be noisy or not.

[5] https://www.mpia.de/gaia/projects/ulysses

To visualize the Empirical Library a specific tool has been created in Java within the UGC (see Figure 17). It provides tools to view both the observed SDSS spectrum and the corresponding BP/RP simulated one, selecting the wished source from the displayed list of parameters. Optionally, the list can be reduced to specific ranges of redshift and/or different types of sources. Additionally, a link to the SDSS Explorer and Finder is provided to view the corresponding source in the original archive. Even though, the BP/RP spectra have low-resolution we can notice an emission line at $\sim$ 830nm in the BP/RP spectrum of Figure 17 which is most apparent in the spectrum of SDSS.

4 Redshift Estimator

The main purpose of the UGC module is to classify the observed by *Gaia* galaxies using the BP/RP spectra and to estimate the redshift and star-formation parameters. The redshift by itself, deforms the rest-frame spectrum causing ineffective tackling the problem directly. UGC solves this containing a set of classifiers and parameter estimators applicable to separate ranges of redshift. The module firstly estimates the redshift of a galaxy and then applies the corresponding components of the set to predict the galaxy type and the various star-formation parameters associated with the synthetic spectra used.

The redshift estimator of the UGC prototype has been prepared to work for a relatively small redshifts ($z \leq 0.2$). So far, it has been prepared using synthetic spectra of normal galaxies. As it has been mentioned in Section 3.2, it is necessary to extend significantly this redshift limit and to work with real observed spectra of extragalactic sources, including quasars as well. The Empirical Library (Section 3.3) is the base for the developing of the new redshift estimator. This estimator implements a supervised learning method, Support Vector Machines (SVM) (Chang & Lin 2011). A part of the library, it has been used to train the algorithm, while the whole library has been used to estimate its performance.

4.1 SVM approach

SVM are a set of related supervised learning methods used for classification and regression. A classification/regression task usually involves separating data into training and testing sets. Each instance in the training set contains one "target value" (i.e. the class labels for the case of classification, or the target parameter values for regression) and several "attributes" (i.e. the features or observed variables). The goal is to produce an SVM-model which optimally fits the target to the attributes. This model applied afterwards to data containing only attributes should successfully classify the target. In the model optimization procedure a test data set is used to monitor the performance. It includes both target and attributes but only the latter are submitted to the model which result is then compared to the corresponding target.

The redshift estimator applies the *epsilon-SVM* type of SVM. In our dataset the target is the redshift of a galaxy and the attributes are the pixel fluxes of the galaxy's BP/RP spectrum. The learning procedure is affected by few parameters common to SVM and its specific parameter with the same name (`epsilonSVR`). The selection of the kernel function is an important common step. The kernel transforms the input data space before learning optimizing the procedure. The non-linear transformation through Radial Basis Function (RBF) is commonly applied for regression problems and it is used in our estimator. However, it introduces a so-called `gamma` parameter which has to be optimized. On the other hand, the minimization of the SVM-model error introduces a penalty parameter of the error term, the `cost`, which has to be optimized as well. There are few other parameters that are fixed in our task.

In the UGC module there is an implementation of SVM using the *libsvm3.14*, a Java interface to the LIBSVM[6]. Additionally, UGC provides tools to select between few simple modes of data standardization. In machine learning, it is usually accepted to pre-process the target and/or the attributes rescaling the data. This is intended for better minimization the classification/regression error. UGC also includes a cross-validation method (SVM tuning) to determine optimal pair of `gamma` and `cost`.

4.2 SVM model for the Redshift estimator

Following the above mentioned approach we produced an SVM model for the redshift estimator finding the "best" from a series of experiments, using different values for the specific SVM para-

[6] https://www.csie.ntu.edu.tw/c̃jlin/libsvm/

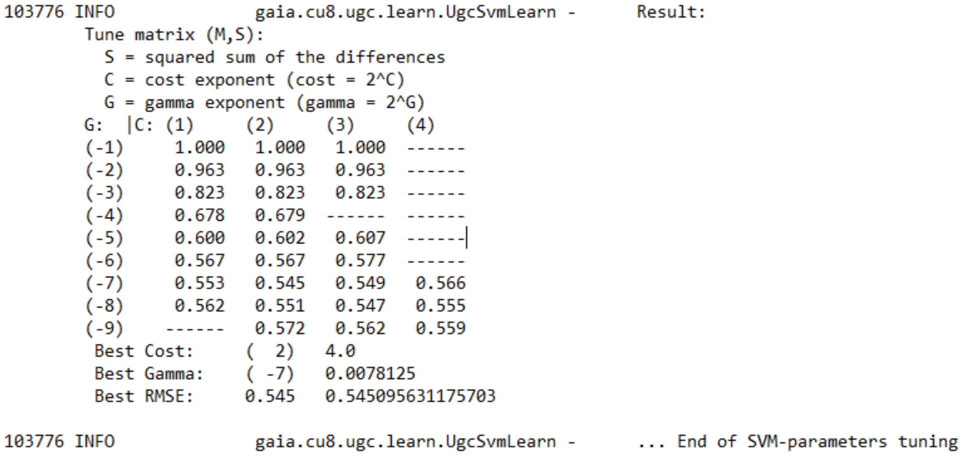
meters as mentioned above. The criterion for the SVM performance has been the error (standard deviation) of the difference for the known SDSS redshift between the predicted SVM redshift, `z_diff` = `z_sdss` − `z_pred`, for all the sources in the Empirical Library. Particularly we tested the provided by UGC normalization methods and different values `epsilonSVR` parameter.
The values used for the various experiments of the `epsilonSVR` parameter are:

(a) `epsilonSVR`=0.1
(b) `epsilonSVR`=0.01
(c) `epsilonSVR`=0.001

Table 10: The normalization methods for the SVM model.

Code number	Code name	Description
0	NO_NORM	No normalization
1	AVE_STD	Standardization for 1σ to [-1,1]
2	MIN_MAX	Rescaling the range of features to [0, 1]
3	AVE_3STD	Standardization for 3σ to [-1,1]
4	TOT_MAX	Rescaling to the total maximum to [0, 1]
5	AVE_5STD	Standardization for 5σ to [-1,1]

A standard procedure provided by UGC has been applied in all the experiments. It includes a sequence of three processes: TUNE, TRAIN, and TEST. An example image from the tuning procedure is seen in Figure 18, while a description of the tasks involved in the three processes follow:



```
103776 INFO            gaia.cu8.ugc.learn.UgcSvmLearn -     Result:
        Tune matrix (M,S):
          S = squared sum of the differences
          C = cost exponent (cost = 2^C)
          G = gamma exponent (gamma = 2^G)
        G:  |C: (1)     (2)     (3)     (4)
        (-1)     1.000   1.000   1.000  ------
        (-2)     0.963   0.963   0.963  ------
        (-3)     0.823   0.823   0.823  ------
        (-4)     0.678   0.679  ------  ------
        (-5)     0.600   0.602   0.607  ------|
        (-6)     0.567   0.567   0.577  ------
        (-7)     0.553   0.545   0.549   0.566
        (-8)     0.562   0.551   0.547   0.555
        (-9)    ------   0.572   0.562   0.559
         Best Cost:     (  2)   4.0
         Best Gamma:    ( -7)   0.0078125
         Best RMSE:     0.545   0.545095631175703

103776 INFO            gaia.cu8.ugc.learn.UgcSvmLearn -     ... End of SVM-parameters tuning
```

Figure 18: Tuning example for the error development. The local minimum value (0.545) is the best one. Then the `gamma` and `cost` parameters are calculated from their exponential form, i.e. `gamma` $= 2^G = 2^{-7}$, `cost` $= 2^C = 2^2$

- TUNE: SVM tuning in which by cross-validation we find the optimal values for the `gamma` and `cost`. The preferable values are those with the minimum error.
- TRAIN: SVM model creation using the optimal parameters and estimating an "internal" error of the model. Training results error in normalized and in real values redshift. It is also provides us with an output file with the estimated redshifts and difference from the given redshifts.
- TEST: SVM model applying to estimate its performance, i.e. estimating an"external" error. As the above processes, the testing also result error in normalized and in real values redshift and provides us with the identical output, but for the testing data set.

In TUNE and TRAIN a small part (no more than 10%) of the Empirical Library has been used as a training set. It has been created respecting the relative distribution of the spectra by redshift and by type (see Table 11). The training set contains 800 galaxies, namely 412 of type GALAXY and 388 of type QSO. The data set that has been used in TEST is the whole Empirical Library (see Table 12). The testing set contains 11,102 galaxies, namely 6,647 GALAXY and 4,455 QSO type.

Table 11: The distribution by redshift and classification of the 800 sources from the Empirical Library included in the training set.

Type of	Redshift range						Total range
sources	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	(0.0-0.6)
GAL	50	53	50	45	20	19	237
G_SBR	11	14	16	12	9	2	64
G_SFR	12	18	19	10	7	2	68
G_AGN	5	10	12	7	6	3	43
Total GAL	78	95	97	74	42	26	412
QSO	12	22	18	22	62	74	210
Q_SBR	2	14	20	37	36	9	118
Q_SFR	0	3	6	1	5	4	19
Q_AGN	1	4	7	10	10	9	41
Total QSO	15	43	51	70	113	96	388
Total	93	138	148	144	155	122	800

Table 12: The distribution by redshift and classification of the Empirical Library (11,102 sources) included in the testing set.

Type of	Redshift range						Total range
sources	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	(0.0-0.6)
GAL	224	1425	2978	1062	201	109	5999
G_SBR	24	51	76	57	19	3	230
G_SFR	46	95	70	29	15	5	260
G_AGN	11	59	53	19	13	3	158
Total GAL	305	1630	3177	1167	248	120	6647
QSO	26	78	174	371	813	1553	3015
Q_SBR	5	75	260	406	417	111	1274
Q_SFR	0	5	12	1	8	12	38
Q_AGN	1	12	25	33	30	27	128
Total QSO	32	170	471	811	1268	1703	4455
Total	337	1800	3648	1978	1516	1823	11102

Following the different runs (different values of `epsilonSVR` and normalization methods) we obtained the results shown in Table 13 and the diagrams that are shown in Appendix B. Table 13 lists the standard deviations values for each normalization and each `epsilonSVR` value. For each `epsilonSVR` we have 36 different cases, regarding the normalization in the redshifts (Target or T) of each spectrum (Source or S). Every possible normalization case is denoted by SnTn, where the number n represent the code number of the corresponding normalization method and the letters represent in what subject the normalization has been performed. The red values in the table are the smallest errors, corresponding to the best candidates.

Analysing the Table 13 and the diagrams in Appendix B we reach the conclusion that the best normalization case concerning the source is to leave the pixel fluxes of the galaxy's BP/RP spectrum without normalization. As for the target selection it is preferable to normalize it with applying the standardization for 1σ, as the error does not change for all three different values of `epsilonSVR`. On the other hand, the optimal value for the `epsilonSVR` parameters seems to be the `epsilonSVR` $= 0.01$, where in Table 13 of the testing set `epsilonSVR` shows a fixed error value. We have to note that we should focus on the testing and not on the training results. From the diagrams we come to the same conclusion. The performance of the method is visualized by the width of the histograms (the smaller the better) and the location of its peak which should be near zero. Similarly, the scatter plots of the difference between real and estimated redshift as a function of either the real redshift or the source brightness of SDSS should display a small scatter. Summarizing, we chose the SVM mobel to be applied in UGC with "`S0T1`" normalization case and the `epsilonSVR` $= 0.01$.

Table 13: Errors from the SVM runs with different normalization methods and `epsilonSVR`. In the normalization cases the letter "S" represents the sources from the SVM, i.e. in our case represents the pixel fluxes of the galaxy's BP/RP spectrum. As for the "T" letter represents the targets from the SVM, i.e. the redshifts of a galaxy. The numbers are the code names from the Table 10. Top: Errors from the training procedure. Bottom: Errors from the testing procedure.

`epsilonSVR`	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001
Normalization	S0			S1			S2			S3			S4			S5		
T0	0.074	0.047	**0.040**	0.081	0.053	0.054	0.080	0.068	0.069	0.082	0.056	0.057	0.086	0.079	0.079	0.078	0.064	0.045
T1	0.046	0.045	0.045	0.060	0.053	0.053	0.067	0.068	0.068	0.062	0.056	0.056	0.078	0.078	0.078	0.053	0.054	0.054
T2	0.048	0.046	0.046	0.061	0.054	0.054	0.068	0.069	0.069	0.063	0.057	0.057	0.078	0.078	0.078	0.065	0.055	0.055
T3	0.058	**0.041**	**0.041**	0.066	0.044	0.045	0.065	0.063	0.064	0.068	0.048	0.049	0.080	0.079	0.080	0.060	0.058	0.058
T4	0.055	0.045	0.046	0.065	0.061	0.061	0.064	0.060	0.061	0.067	0.044	0.044	0.084	0.078	0.078	0.061	0.054	0.055
T5	0.070	0.043	0.043	0.077	0.057	0.057	0.076	0.070	0.071	0.069	0.060	0.060	0.081	0.076	0.080	0.074	0.049	0.050

`epsilonSVR`	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001
Normalization	S0			S1			S2			S3			S4			S5		
T0	0.070	**0.056**	0.057	0.080	0.065	0.065	0.080	0.067	0.067	0.080	0.066	0.066	0.077	0.068	0.068	0.080	0.067	0.064
T1	**0.056**	**0.056**	**0.056**	0.066	0.064	0.064	0.066	0.066	0.066	0.067	0.064	0.064	0.067	0.067	0.068	0.065	0.064	0.064
T2	0.057	**0.056**	**0.056**	0.068	0.064	0.064	0.068	0.066	0.066	0.068	0.064	0.064	0.068	0.068	0.068	0.068	0.064	0.065
T3	0.060	**0.056**	0.057	0.070	0.063	0.063	0.069	0.065	0.065	0.070	0.063	0.063	0.071	0.068	0.068	0.069	0.065	0.065
T4	0.136	**0.056**	**0.056**	0.072	0.066	0.066	0.072	0.065	0.065	0.072	0.063	0.063	0.073	0.067	0.068	0.072	0.064	0.065
T5	0.065	**0.056**	**0.056**	0.077	0.065	0.065	0.076	0.067	0.068	0.076	0.066	0.066	0.073	0.067	0.068	0.076	0.064	0.064

4.3 Performance of the redshift estimator

The accepted SVM has been trained with the subset of the Empirical Library, which under the term "galaxies" combine two SDSS source types, GALAXY and QSO. So, in the training we included both types. In the following performance analysis of the redshift estimator we split the testing set (the whole Empirical Library) into two parts, by the source type, to examine whether there is a dependence of the SVM error on the source type. Despite the inclusion in the test set of the sources used for training, the performance results are not influenced by the small number (800) of these sources. In these tests we derived the difference between the SDSS-spectroscopic redshifts (known from the Empirical Library) and the SVM-estimated redshift. The parameter that quantifies the SVM performance is again the standard deviation, `std_z` of the differences.

The two tests of the SVM model where performed on 6,646 sources of GALAXY type and on 4,453 QSO, resulting in a corresponding redshift standard deviation of `std_z` = 0.053 and `std_z` = 0.059. The distribution of the differences in redshift (`z_diff`) is presented on Figure 19. The histogram has a nearly identical form for both GALAXY and QSO types.

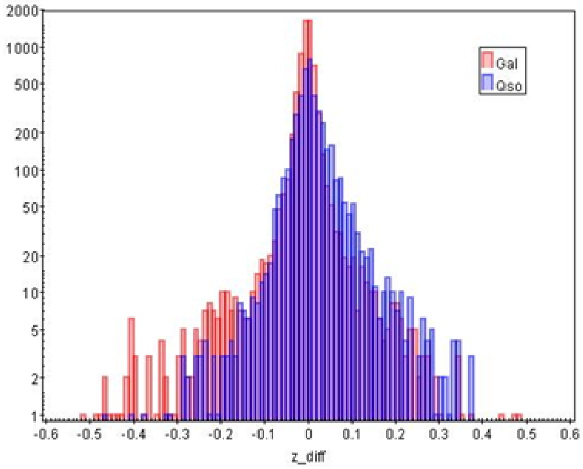


Figure 19: Histogram of the differences in redshift between the SDSS-spectroscopic measurement and the SVM estimated value. The normal galaxies are presented with red bins, while the quasars with blue bins. The distribution is almost the same for the both types.

However, the plot of the differences with the respect to the real redshifts (Figure 20, left panel) reveals some important details. It is evident that the QSO sources are distributed towards larger redshifts, which naturally follows their redshift distribution in the Empirical Library (Table 12). In the lower part of `z_sdss` there is a number of sources with overestimated redshift. Specifically, there are few tenths of sources with almost zero redshift, for which the SVM predicts large values. This behaviour is expected as the SVM is forced not to predict negative redshifts in this range. We shall avoid training SVM with such sources, even more because these are fairly nearby galaxies, which are typically extended. On the other extreme, at large redshifts, we have a number of underestimated z values. This is probably due to an "edge effect", because the SVM has not be trained for even larger redshifts. These two effects persist also in the histogram, although they are not clearly seen because of the relatively small number of cases. Furthermore, we also analysed whether these deviations in the redshift prediction depend on the sources used for the library. As it can be seen in the right panel of Figure 20, there is a strong increase of the scatter towards the faint SDSS sources. It is clear that the large errors are mainly caused by the inclusion of faint source spectra (lower signal-to-noise) in the library.

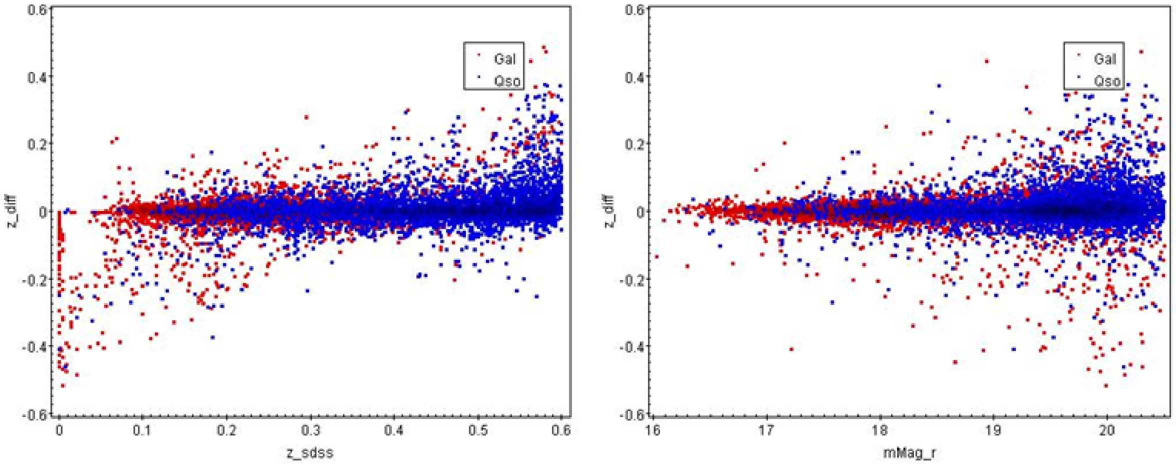


Figure 20: Plot of the difference between SDSS and *Gaia* estimated redshifts as a function of SDSS redshifts (Left) and SDSS magnitude (Right). GALAXY and QSO type sources are indicated as red and blue points, respectively.

These effects can also be seen by examining the SVM-error in particular ranges of redshift and magnitude. We observe the largest error in redshifts below `z_sdss` = 0.05 independently of the source type (see Figure 21). For half of the whole redshift range the standard deviation is below 0.050 and for its central part it is almost 0.025. As for the QSO type they show larger errors for `z_sdss` < 0.20, while for the GALAXY type the scatter increases above 0.40. On the other hand, the strong influence of the source magnitude on the standard deviation is demonstrated in Figure 22. It is evident that in the process of selecting SDSS spectra to create such a library we have to avoid sources of magnitude `modelMag_r` > 19.5.

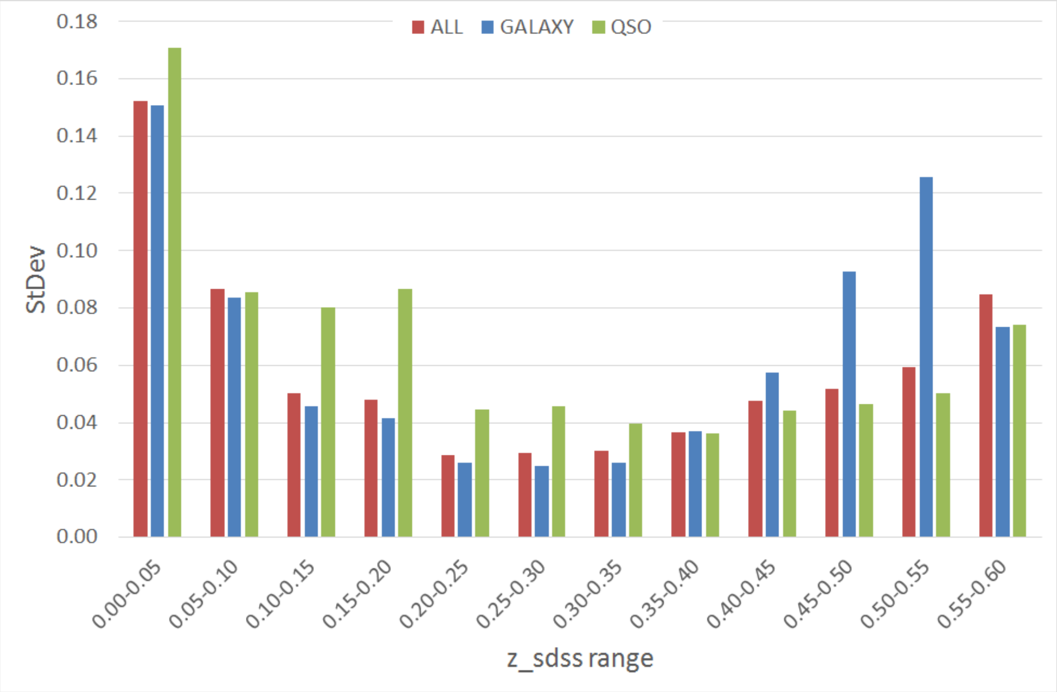


Figure 21: Distribution for particular ranges of the SDSS redshifts compared to the standard deviation of the difference of redshifts, for both GALAXY and QSO, as well as for each type separately. Both types are presented with red bins, while the GALAXY sources are presented with blue bins and the QSO with green bins.

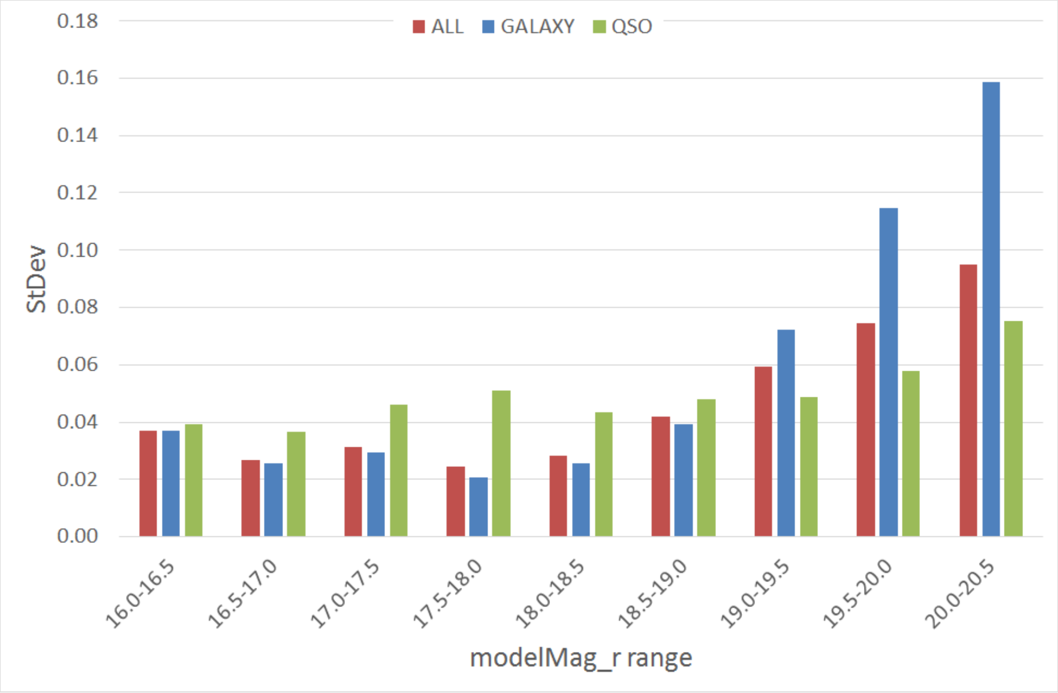


Figure 22: Distribution for particular ranges of the SDSS magnitude compared to the standard deviation of the difference of redshifts, for both GALAXY and QSO, as well as for each type separately. Both types are presented with red bins, while the GALAXY sources are presented with blue bins and the QSO with green bins.

Comparing the newly created SVM model to the original prototype UGC (Bellas-Velidis et al. 2012), its performance on the redshift prediction, appears worse (∼0.035 vs 0.056). However, we have to emphasize that the prototype had been developed for three times smaller redshift range, up to $z = 0.2$. Additionally, it was based on a set of synthetic spectra created on a restricted set of galaxy evolution models and then modelled for BP/RP simplifying the real output of the spectrophotometer. Moreover, the prototype optimize the solution implementing a number of SVM models working together. Furthermore, in our current library we also include observed SDSS spectra of quasars in addition to the ones of galaxies.

We can also compare our estimator with others, which are based on photometric data, (Beck et al. 2017), even though a such direct comparison is potentially not indicative, because of the different observational base used, colour indices instead of low-resolution spectra. Normalizing the errors of the SVM estimator of Beck et al. (2017), we derive nearly two times more accurate results. We have to note that their training and testing sets each includes quite large number of sources (almost 70,000).

Overall, the performance of our SVM model makes it appropriate for use as a front-end redshift estimator for the UGC module. The above analysis also indicates the ways to further improve its performance.

5 Discussion and Conclusions

In this thesis, we developed a redshift estimator for the Unresolved Galaxy Classifier (UGC), which improves the prototype, significantly expanding its redshift range by a factor of 3, making it applicable up to $z = 0.6$. Moreover, instead of synthetic spectra, we based our analysis on observed spectra of galaxies and quasars with known redshifts. The observed spectra were taken from the SDSS DR13 archive that were also observed by the *Gaia's* BP/RP spectrophotometer. We modelled the *Gaia's* spectra for our sample and created a new library. The modelled spectra were used in a number of experiments to optimize the SVM model and quantify its dependence on the redshift and magnitude ranges of the library spectra.

The performance of the prototype UGC in estimating the redshift, seems to be quite better than our method. However, as mentioned above, it is restricted in much narrower redshift range and it has been based on synthetic spectra simplifying the real situation. Moreover, the prototype combines a number of SVM models while in our work we merged those related to the redshift estimator into a single one.

The analysis show us ways to improve further the SVM model. Sources with very small redshift should be avoided, while the upper redshift limit of the training set must be extended by enriching the sample with more high z sources. In addition, spectra of faint sources (> 19.5 mag) should not be included, while the whole training sample can be easily doubled. Finally, a set of SVM models can be implemented in UGC, as in the prototype, optimizing its performance for different redshift ranges. Once we have in hand real spectra observed by *Gaia* BP/RP these instead of the modelled ones have to be used for the training.

The developed SVM model and its performance show us that the goal has been reached. The model is adequate to be implemented within UGC for the *Gaia's* ground-base pipeline. Its practical usefulness will be tested since *Gaia* is observing and will continue to observe more than 1.5 million of unresolved galaxies within this redshift range.

Appendix A

SDSS CasJob and spectra query commands

A-1 In order to perform the cross-match (Section 2.2) of *Gaia* DR1 sources with galaxies observed spectroscopically (in SDSS DR13), we used a query from CasJobs to download necessary basic parameters concerning the DR13. The downloaded file contained 2,401,646 GALAXY and 649,908 QSO sources. The parameters we used, are derived from the SpecObj table, which contains spectroscopic information for all objects with clean spectra. The query that we used is shown below.

```
SELECT s.specObjID, s.ra, s.dec, s.z, s.class, s.subClass
INTO MyDB.name
FROM SpecObj AS s
WHERE (s.class = 'GALAXY' or s.class = 'QSO')
```

A-2.1 Similarly, in order to perform the cross-match (in Section 3.1), with SDSS BOSS DR13, we used another CasJobs query (see below). The downloaded file contains 1,498,815 GALAXY and 359,843 QSO sources. The parameters we used, are derived from the SpecObj table. The BOSS spectra are identified by the parameter `s.class_noqso`. The two different source types are not distinguish in this parameter.

```
SELECT s.specObjID, s.ra, s.dec, s.z, s.class
INTO MyDB.name
FROM SpecObj AS s
WHERE (s.class_noqso = 'GALAXY')
```

A-2.2 After completing the cross-match and computing the multiplicity in Section 3.1, we performed another query (listed below) to extend the current table of parameters by joining information from the table of photometric objects. We created a table in order to join the existing sources with the new parameters. Expected the SpecObj, we used and the PhotoObj table, which contains attributes of each photometric object.

```
SELECT s.specObjID, s.ra, s.dec, s.z, s.zErr, s.z_noqso, s.zErr_noqso,
  s.class, s.subClass, s.plate, s.mjd, s.fiberID, s.snMedian,
  p.petroR50_r, p.petroR50Err_r, p.petroR90_r, p.petroR90Err_r,
  p.modelMag_u, p.modelMagErr_u, p.modelMag_g, p.modelMagErr_g, p.modelMag_r,
  p.modelMagErr_r, p.modelMag_z, p.modelMagErr_z, p.modelMag_i, p.modelMagErr_i
  p.dered_r, p.fiber2Mag_r, p.fiber2MagErr_r
INTO MyDB.name
FROM MyDB.name2 AS m
LEFT JOIN SpecObj AS s ON s.specObjID=m.specObjID
LEFT JOIN SpecPhoto AS t ON t.specObjID=m.specObjID
LEFT JOIN PhotoObj AS p ON p.ObjID=s.bestObjID
```

A-3 We downloaded the spectra of 11,102 sources from the Spectral Library and gathered them in a table. The current spectra are FITS files.

```
wget -nv -r -nH --cut-dirs=7
  -i selected_myfile.dat
  -B https://data.sdss.org/sas/dr13/eboss/spectro/redux/v5_9_0/spectra/lite/
```

Appendix B

SVM

The redshift estimator has been implemented through a supervised learning method, Support Vector Machines (SVM). The selection of the appropriate SVM has been based on a large number of experiments. These include various normalization methods and different values of `epsilonSVR` parameter. The spectra used for the whole process were taken from the Empirical Library. The histograms of the differences, z_diff = z_sdss - z_pred, between the real SDSS redshifts and the estimated from the SVM, as well as plots of these differences with respect to the real SDSS redshifts are presented below for each combination of the normalization and the parameter value.

In B-1 the diagrams from the TRAIN procedure are presented, where the "internal" errors are computed. In B-2 we include the diagrams from the TEST procedure, where the "external" errors are computed, based on the whole Empirical Library. Each normalization case is presented as SnTn, where the number n is the code number of the corresponding normalization method and the letters represent in what subject the normalization has been performed. The normalization in redshifts is denoted as T and for spectra is denoted as S.

B-1 TRAIN

1.1 Histograms for diff_z

1.1.1 epsilonSVR=0.1

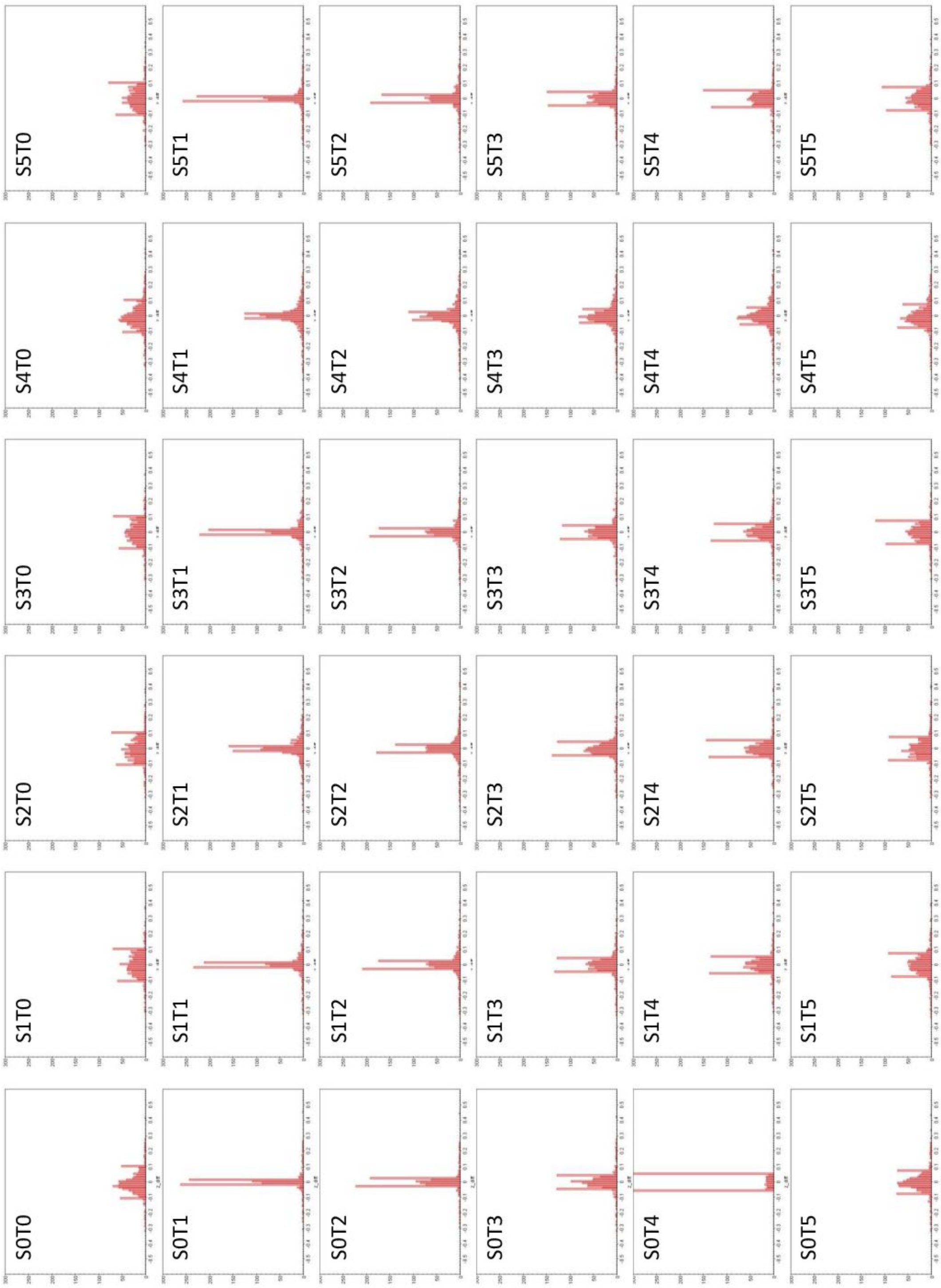


Figure 23: Histogram of z_diff from the train set for epsilonSVR=0.1. Each "window" shows the outcome of the training for all the possible normalization methods.

1.1.2 epsilonSVR=0.01

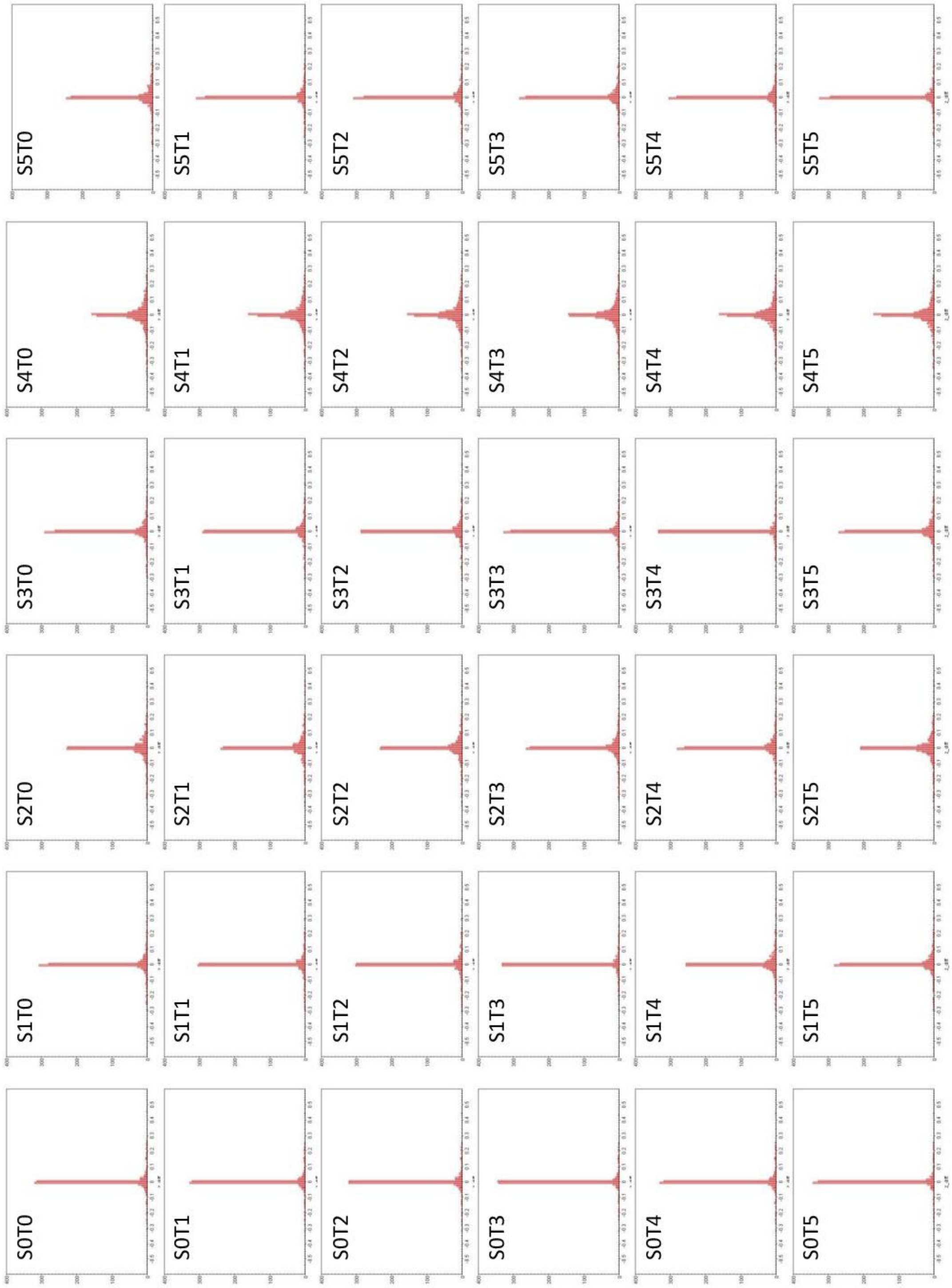
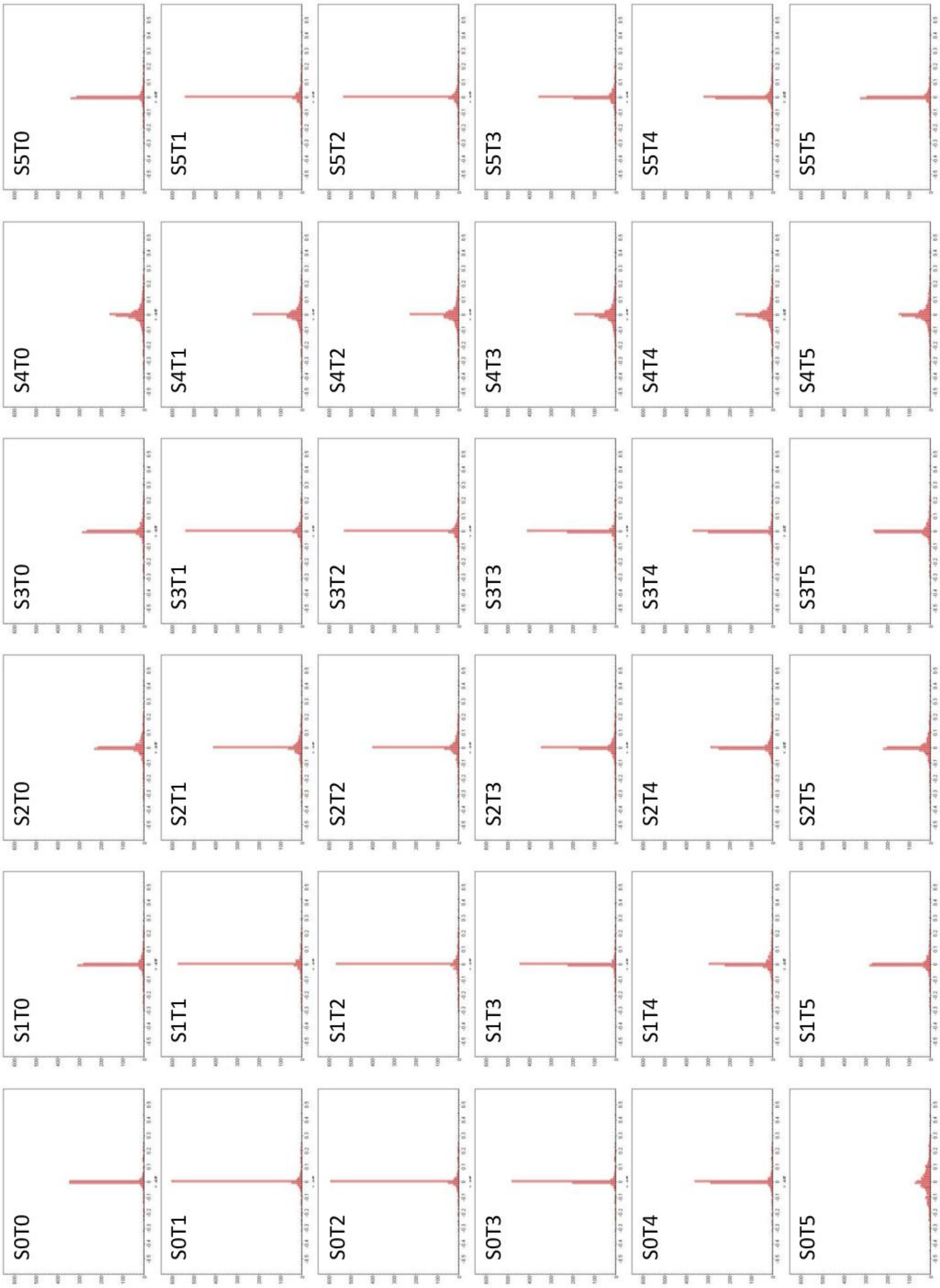


Figure 24: Histogram of z_diff from the train set for epsilonSVR=0.01. Each “window” shows the outcome of the training for all the possible normalization methods.

1.1.3 epsilonSVR=0.001



S0T0 S0T1 S0T2 S0T3 S0T4 S0T5

S1T0 S1T1 S1T2 S1T3 S1T4 S1T5

S2T0 S2T1 S2T2 S2T3 S2T4 S2T5

S3T0 S3T1 S3T2 S3T3 S3T4 S3T5

S4T0 S4T1 S4T2 S4T3 S4T4 S4T5

S5T0 S5T1 S5T2 S5T3 S5T4 S5T5

Figure 25: Histogram of z_diff from the train set for epsilonSVR=0.001. Each “window” shows the outcome of the training for all the possible normalization methods.

1.2 Plot for diff_z - sdss_z

1.2.1 epsilonSVR=0.1

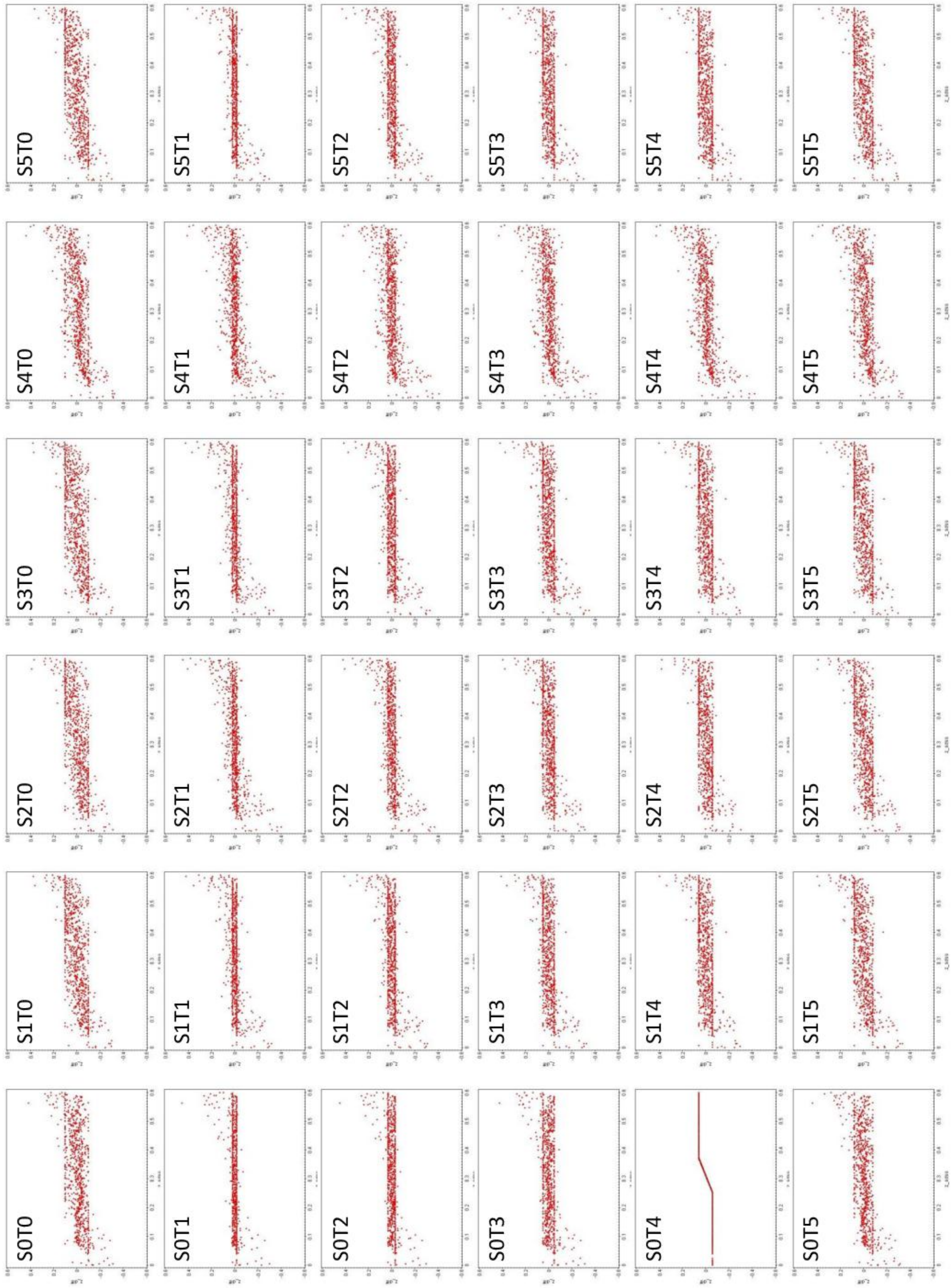


Figure 26: Plot of z_diff to z_sdss from the train set for epsilonSVR=0.1. Each "window" shows the outcome of the training for all the possible normalization methods.

1.2.2 epsilonSVR=0.01

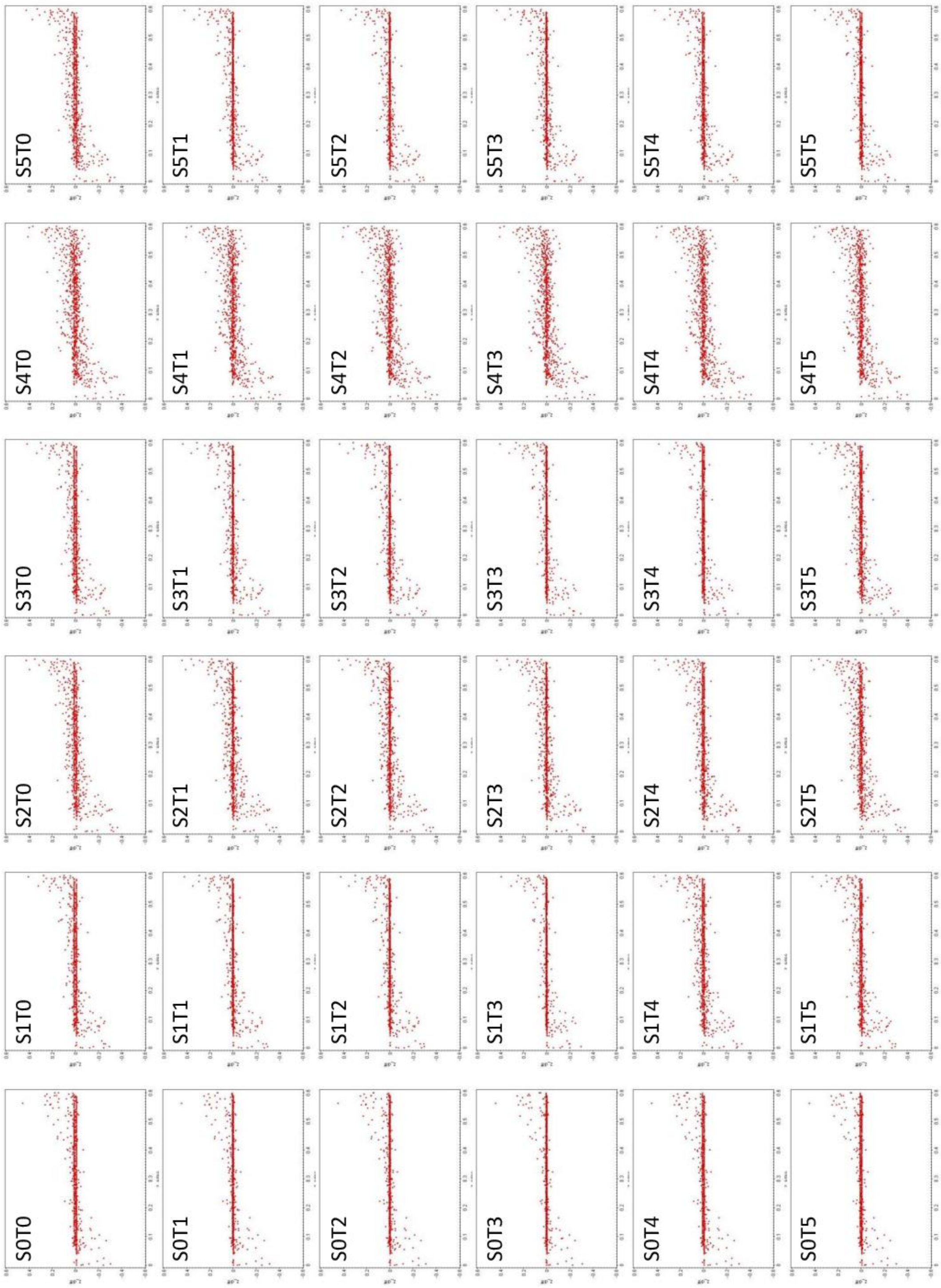


Figure 27: Plot of z_diff to z_sdss from the train set for epsilonSVR=0.01. Each "window" shows the outcome of the training for all the possible normalization methods.

1.2.3 epsilonSVR=0.001

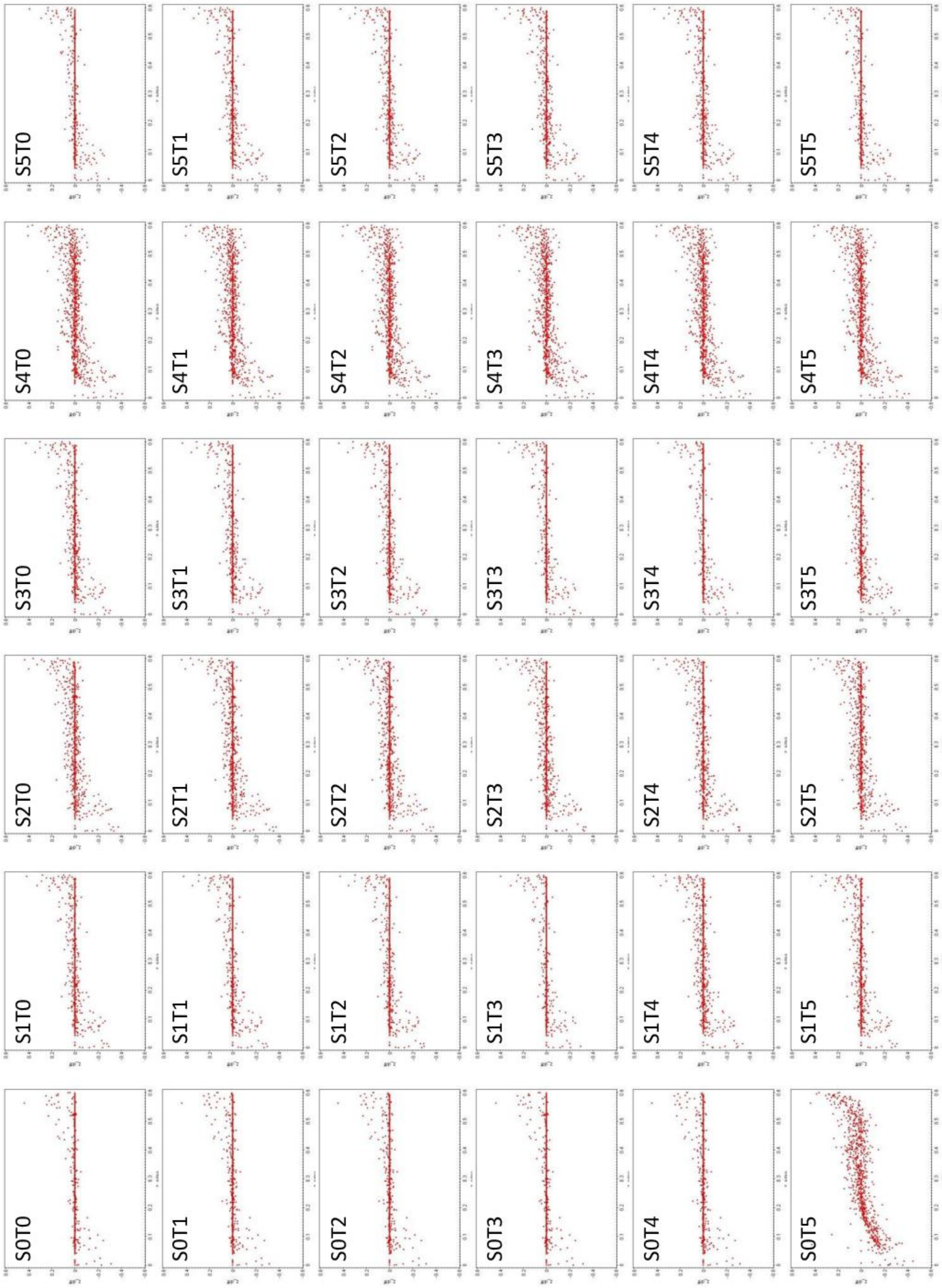


Figure 28: Plot of z_diff to z_sdss from the train set for epsilonSVR=0.001. Each "window" shows the outcome of the training for all the possible normalization methods.

B-2 TEST

2.1 Histograms for diff_z

2.1.1 epsilonSVR=0.1

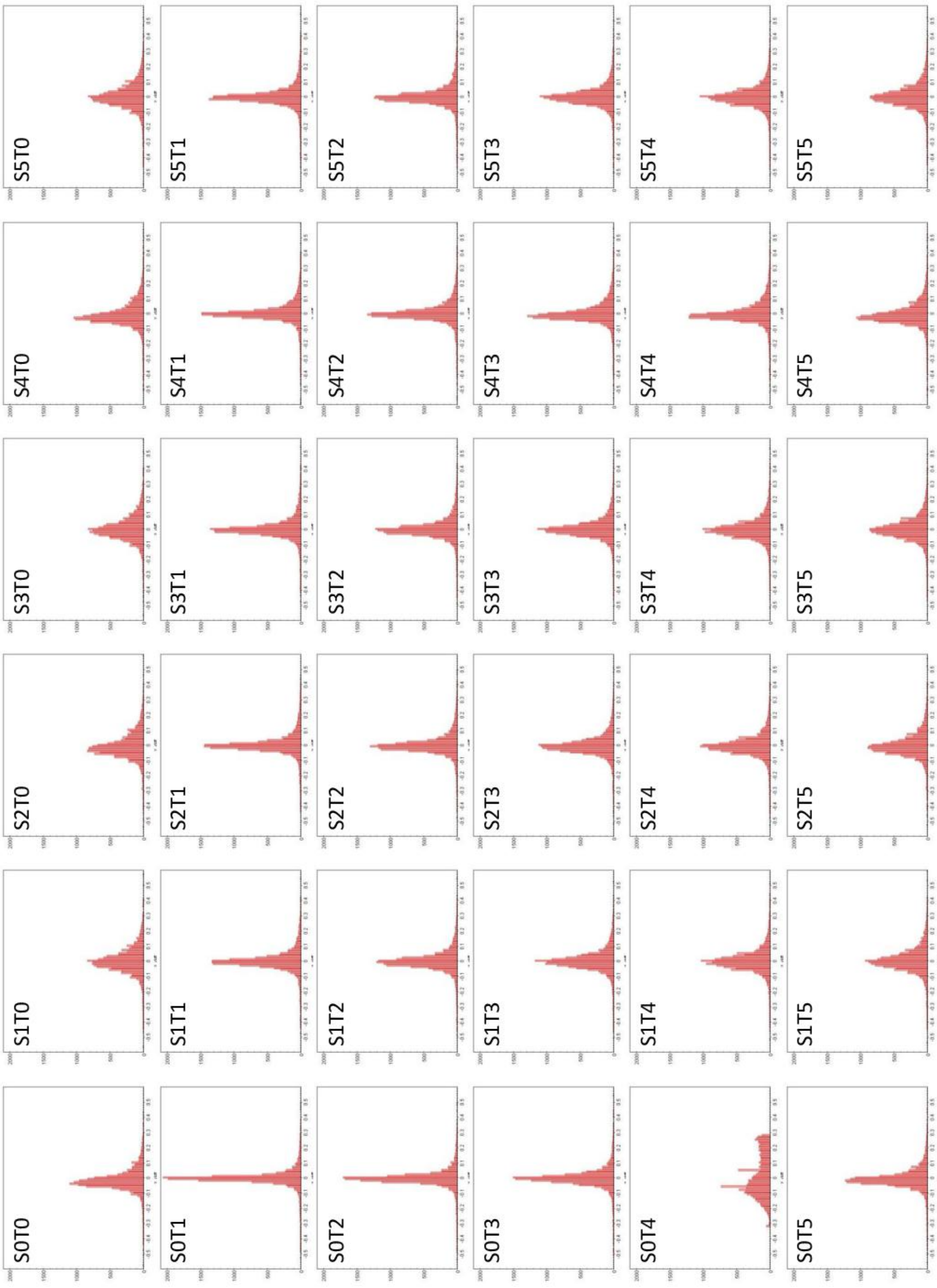


Figure 29: Histogram of z_diff from the test set for epsilonSVR=0.1. Each "window" shows the outcome of the testing for all the possible normalization methods.

2.1.2 epsilonSVR=0.01

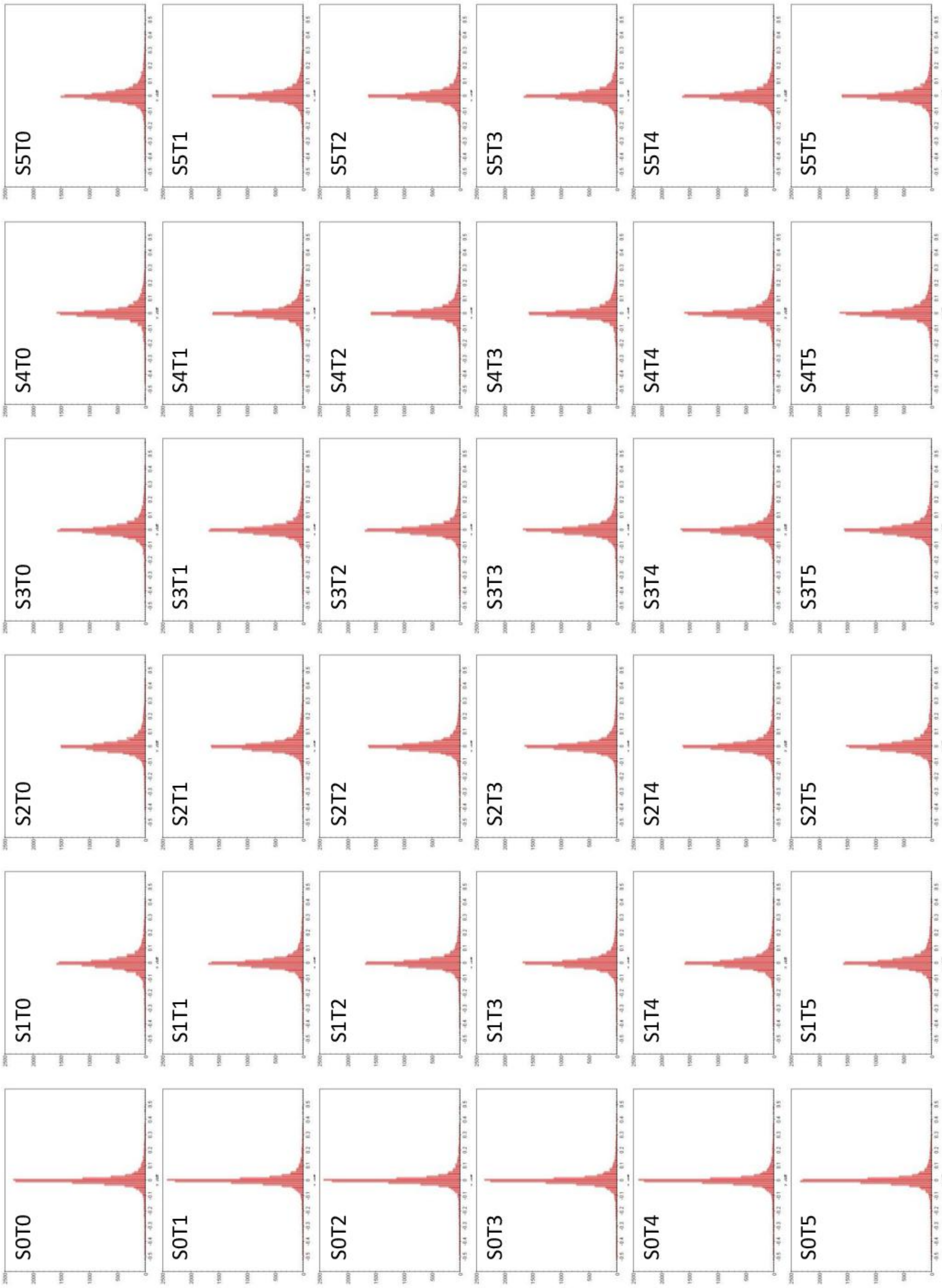


Figure 30: Histogram of z_diff from the test set for epsilonSVR=0.01. Each "window" shows the outcome of the testing for all the possible normalization methods.

2.1.3 epsilonSVR=0.001

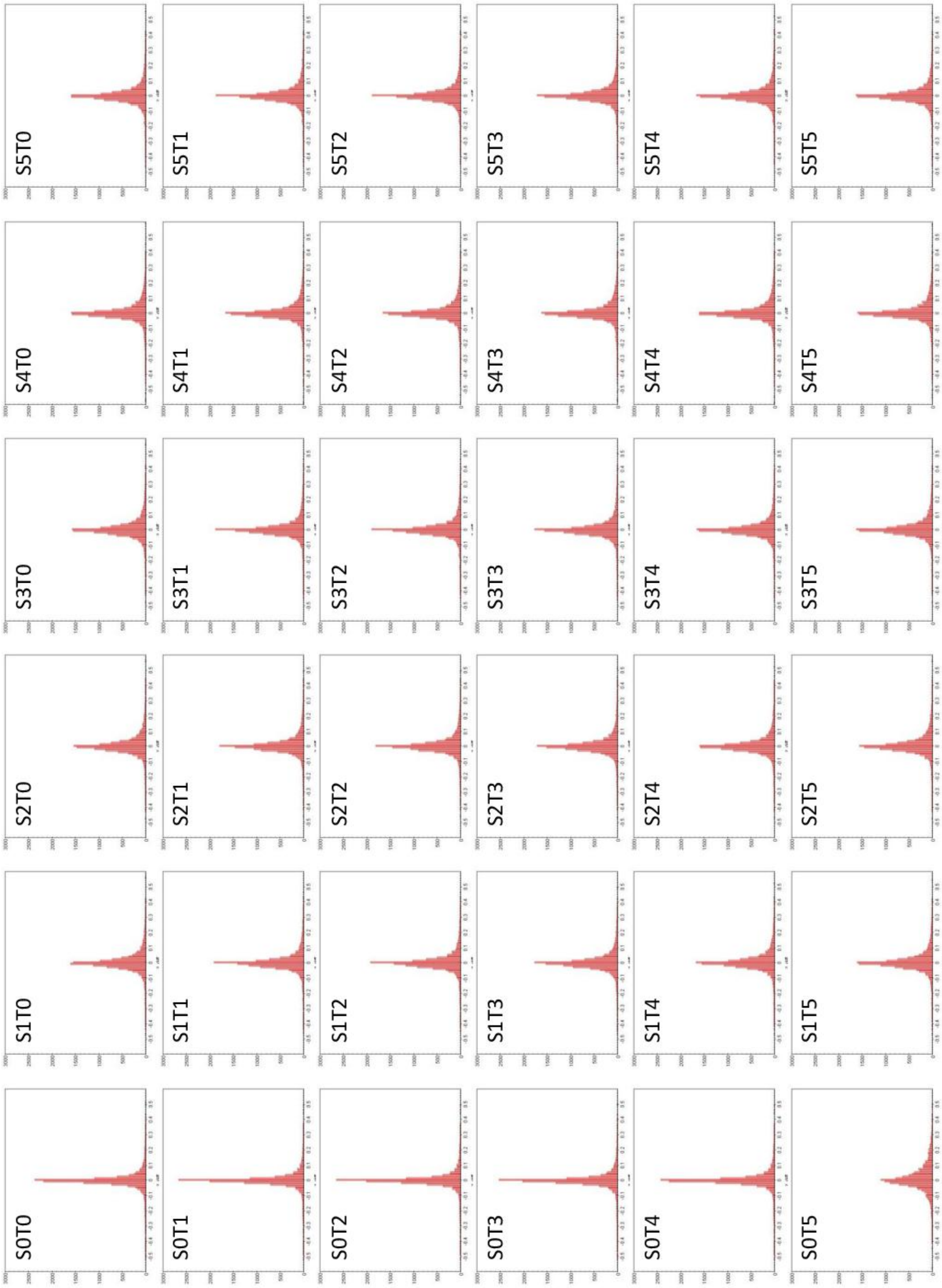


Figure 31: Histogram of z_diff from the test set for epsilonSVR=0.001. Each “window” shows the outcome of the testing for all the possible normalization methods.

2.2 Plot for diff_z - sdss_z

2.2.1 epsilonSVR=0.1

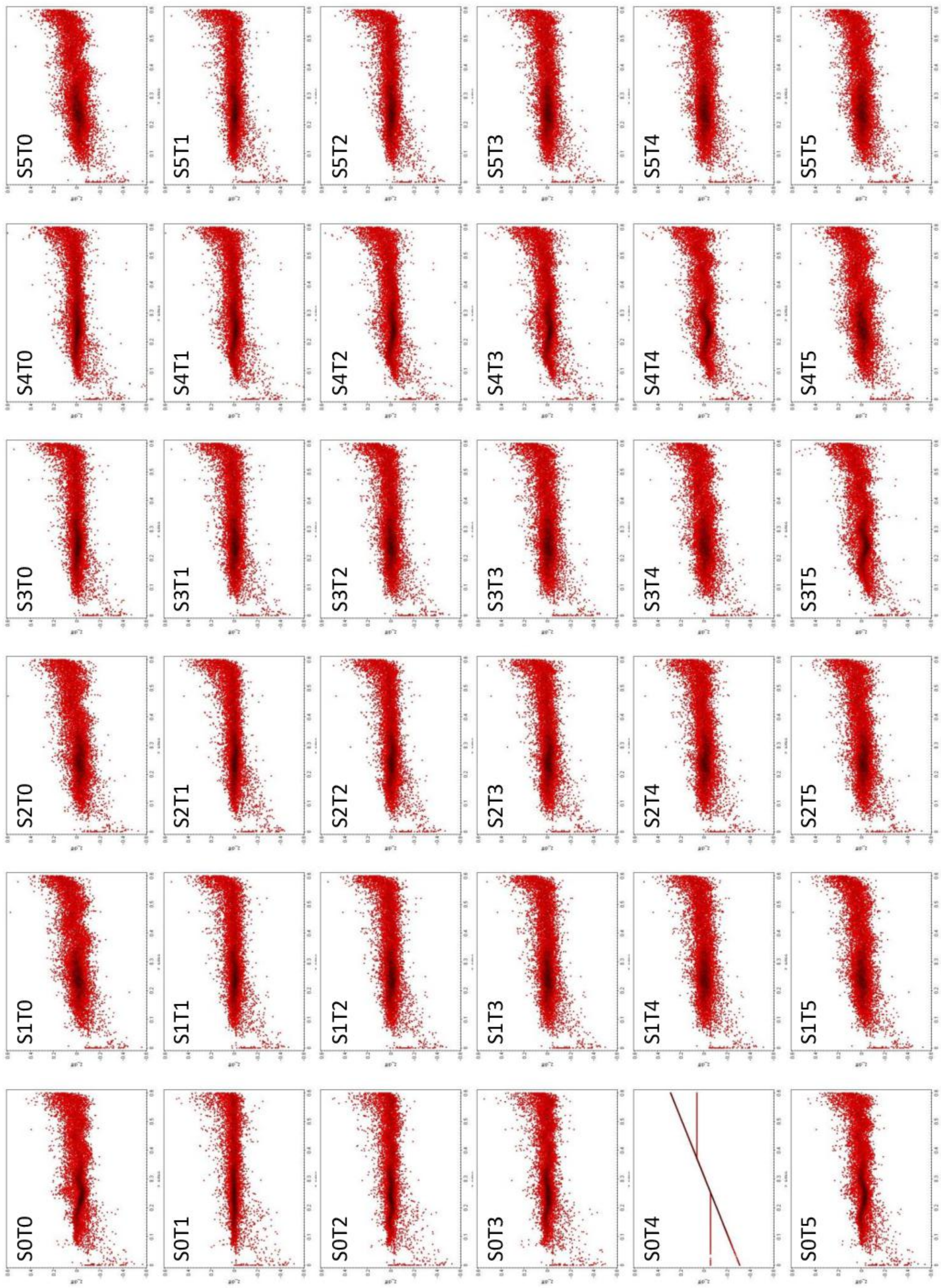


Figure 32: Plot of z_diff to z_sdss from the test set for epsilonSVR=0.1. Each "window" shows the outcome of the testing for all the possible normalization methods.

2.2.2 epsilonSVR=0.01

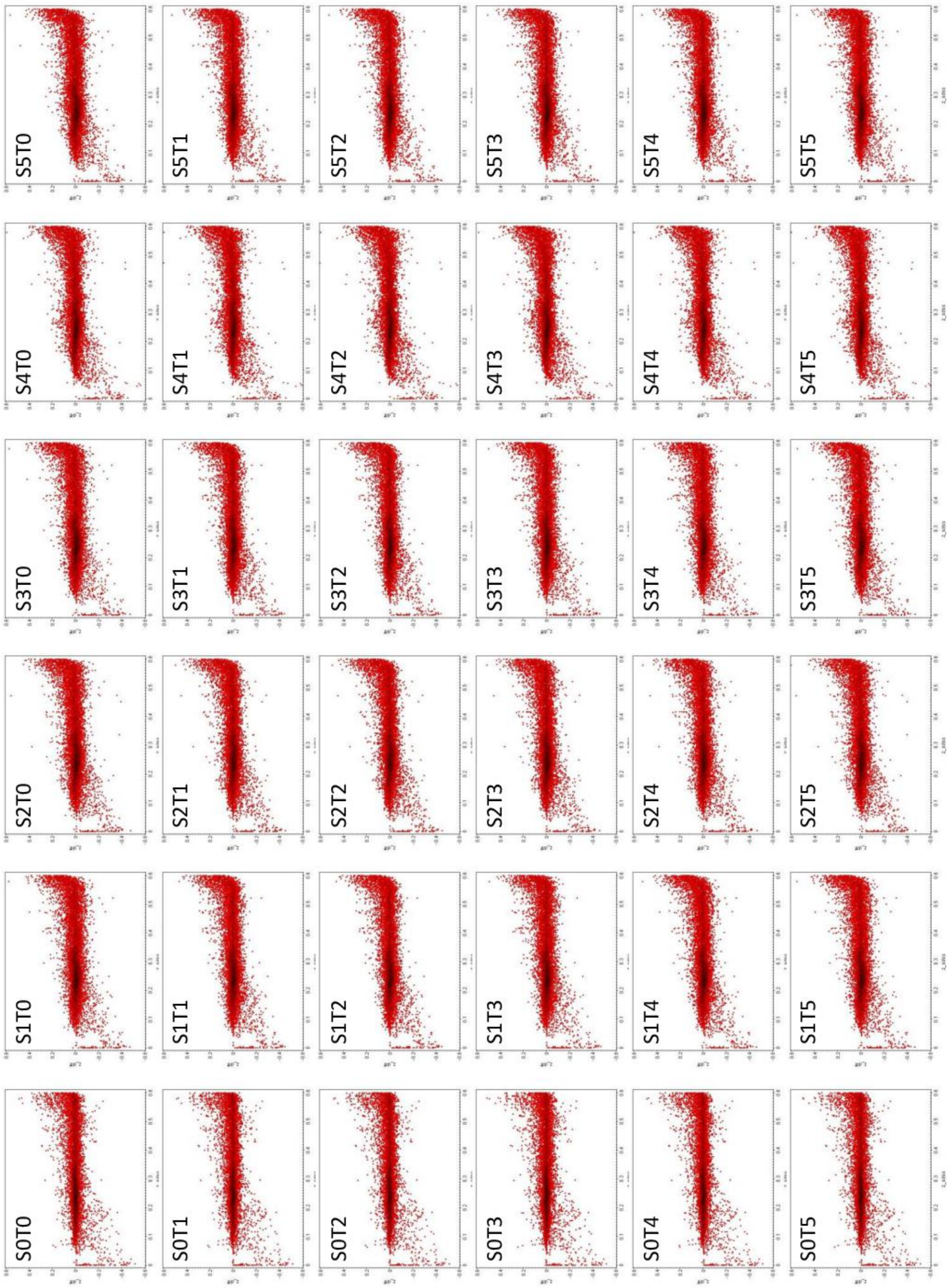


Figure 33: Plot of z_diff to z_sdss from the train set for epsilonSVR=0.01. Each “window” shows the outcome of the testing for all the possible normalization methods.

2.2.3 epsilonSVR=0.001

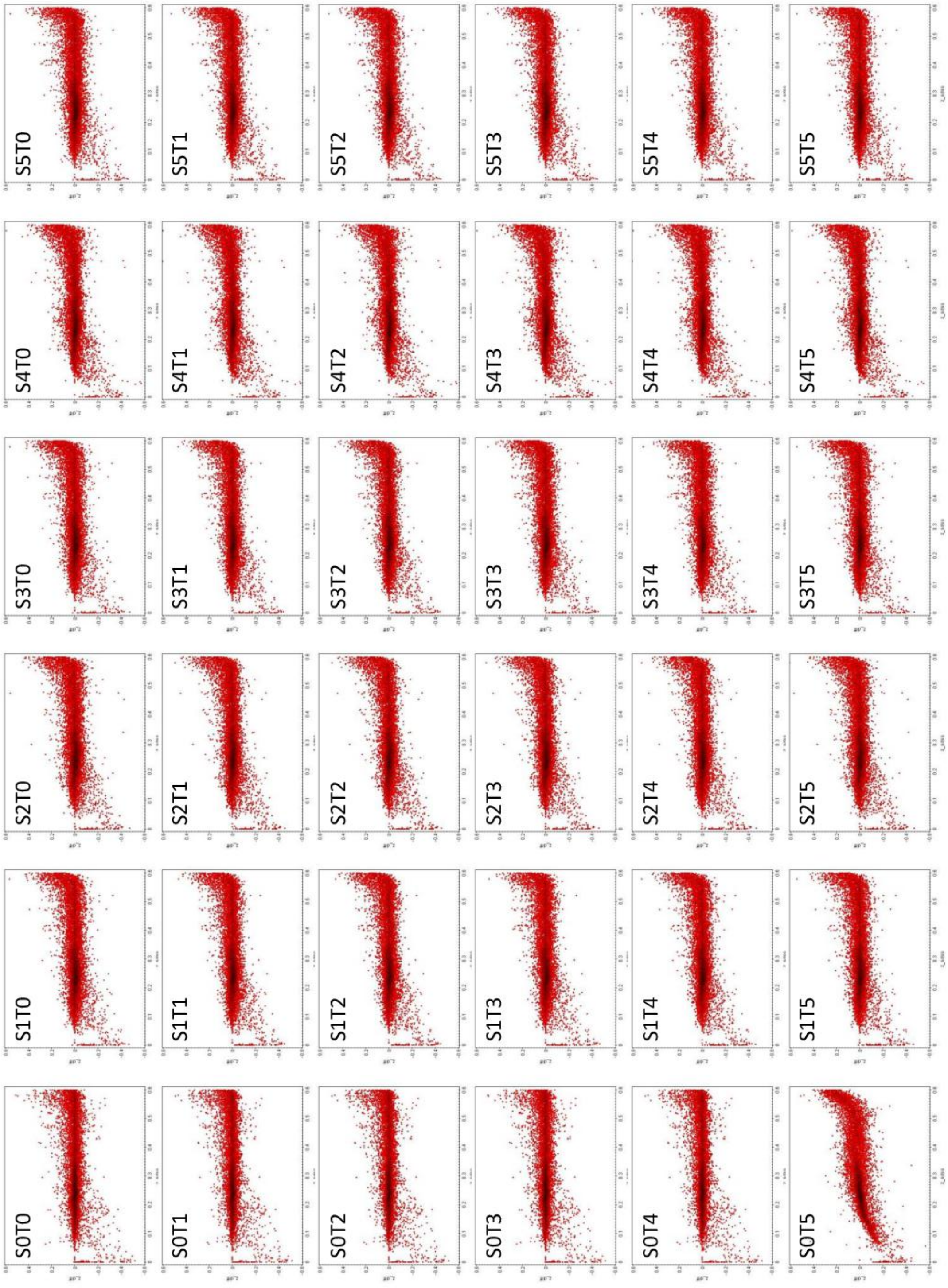


Figure 34: Plot of z_diff to z_sdss from the train set for epsilonSVR=0.001. Each "window" shows the outcome of the testing for all the possible normalization methods.

References

Albareti, F. D., Allende Prieto, C., Almeida, A., et al. 2017, ApJS, 233, 25

Bailer-Jones, C. A. L., Andrae, R., Arcay, B., et al. 2013, A& A, 559, A74

Beck, R., Lin, C.-A., Ishida, E. E. O., et al. 2017, MNRAS, 468, 4323

Bellas-Velidis, I., Kontizas, M., Dapergolas, A., et al. 2012, Bulgarian Astronomical Journal, 18, 3

Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28

Chang, C.-C. & Lin, C.-J. 2011, ACM Transactions on Intelligent Systems and Technology, 2, 27:1, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2016a, A& A, 595, A2

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016b, A& A, 595, A1

Hubble, E. P. 1926, ApJ, 64

Karampelas, A., Kontizas, M., Rocca-Volmerange, B., et al. 2012, A& A, 538, A38

Kennicutt, Jr., R. C. 1992, ApJS, 79, 255

Li, N. & Thakar, A. R. 2008, Computing in Science and Engineering, 10, 18

Newman, P. R., Long, D. C., Snedden, S. A., et al. 2004, in Proc. SPIE, Vol. 5492, Ground-based Instrumentation for Astronomy, ed. A. F. M. Moorwood & M. Iye, 533–544

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A& A, 594, A13

Riess, A. G., Rodney, S. A., Scolnic, D. M., et al. 2018, ApJ, 853, 126

Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29

Thakar, A. R. 2008, Computing in Science and Engineering, 10, 9

Tsalmantza, P., Kontizas, M., Bailer-Jones, C. A. L., et al. 2007, A& A, 470, 761