Università degli Studi di Napoli *Federico II*

DOTTORATO DI RICERCA IN FISICA

Ciclo XXXII

Coordinatore: prof. Salvatore Capozziello

Improving the reliability of photometric redshift catalogues with Self-Organizing Maps

Dottorando
Oleksandra Razim

Tutori
Prof. Giuseppe Longo

Settore Scientifico Disciplinare FIS/05

Anni 2017/2021

Università degli Studi di Napoli *Federico II*

DOTTORATO DI RICERCA IN FISICA

Ciclo XXXII

Coordinatore: prof. Salvatore Capozziello

Improving the reliability of photometric redshift catalogues with Self-Organizing Maps

Dottorando
Oleksandra Razim

Tutori
Prof. Giuseppe Longo

Settore Scientifico Disciplinare FIS/05

Anni 2017/2021

Abstract

The already existing and upcoming massive surveys, such as KiDS, DES, Euclid and LSST, bring to reality many exiting possibilities in precision cosmology, as well as galaxy evolution and large-scale structure studies. However, to fully benefit from these surveys, we require redshifts for millions of galaxies. Currently, it is impossible to obtain these redshifts with spectroscopy only, since it would require immense observational time. For this reason, an alternative method, called photometric redshifts (photo-z) is used.

This thesis is dedicated to a new data cleaning methodology, that allows to significantly improve the quality of photo-z catalogues and to guarantee the reliability of their quality metrics (i.e., to perform photo-z calibration). This methodology is based on an unsupervised Machine Learning (ML) algorithm called Self-Organizing Maps (SOM). Different components of this methodology allow to tackle several important issues. Namely, in-cell SOM anomaly detection helps to alleviate contamination of a spectral redshift catalogue with unreliable measurements and reduce the percentage of catastrophic outliers in photo-z predictions. Another approach, SOM occupation map calibration, counters the deterioration of the reliability of photo-z catalogues caused by differences between the parameter space of the train and run datasets.

The methodology is tested on a deep 30-band photometric catalogue COSMOS2015. Photometric redshifts for this catalogue were obtained using a well-tested supervised ML algorithm MLPQNA. For additional comparison, SED (Spectral Energy Distribution) fitting photo-z were used. For both photo-z methods, the usage of the SOM-based data cleaning methodology reduces the percentage of catastrophic outliers by at least an order with corresponding improvements of other metrics. This result makes the SOM-based data cleaning a highly recommendable tool for preparing photo-z catalogues for the upcoming large surveys.

Contents

List of Figures

List of Tables

Abbreviations

ML	Machine Learning
AI	Artificial intelligence
KB	Knowledge base
NLP	Neural Language Processing
DNN	Deep Neural Network
CNN	Convolutional Neural Network
GPT	Generative Pre-trained Transformer
RF	Random Forest
SVM	Support Vector Machines

IR	Infrared
UV	Ultraviolet
RMSE	Root Mean Square Error
AGN	Active Galactic Nuclei
SNe	Supernovae
SFR	Star-Formation Rate

Surveys abbreviations

SDSS	Sloan Digital Sky Survey, Ahumada et al. [2020]
LSST	Vera Rubin Observatory Large Synoptic Survey Telescope, Ivezić et al. [2019]
LSST DESC	LSST Dark Energy Science Collaboration, The LSST Dark Energy Science Collaboration et al. [2018]
Euclid	Euclid, Laureijs et al. [2011]
COSMOS	Cosmic Evolution Survey, Scoville et al. [2007]
RST	Nancy Grace Roman Space Telescope, formerly Wide Field Infrared Survey Telescope (WFIRST), Spergel et al. [2015]
Gaia	Gaia, Gaia Collaboration et al. [2016]
SKA	Square Kilometre Array, Johnston et al. [2008]
DESI	Dark Energy Spectroscopic Instrument, DESI Collaboration et al. [2016b]
HDF	Hubble Deep Field, Williams et al. [1996]
AEGIS	All-wavelength Extended Groth Strip International Survey, Davis et al. [2007]
GOODS	Great Observatories Origins Deep Survey, Giavalisco et al. [2004]
DES	Dark Energy Survey, The Dark Energy Survey Collaboration [2005]
DES SV	Dark Energy Survey Science Verification data, Jarvis et al. [2016]
PAUS	Physics of the Accelerating Universe Survey, Padilla et al. [2019]
CFHTLenS	Canada-France-Hawaii Telescope Lensing Survey, Heymans et al. [2012]
CFHT/MegaCam	CFHT MegaCam, Boulade et al. [2003]
SPLASH	Spitzer Large Area Survey with Hyper-Suprime-Cam
IRAC	Infrared Array Camera, Eisenhardt et al. [2004]

HSC-SSP	Hyper Suprime-Cam Subaru Strategic Program, Aihara et al. [2018]
GALEX	Galaxy Evolution Explorer, Morrissey et al. [2007]

Chapter 1

Rather futurological introduction

1.1 Scientific revolutions of the 20th century

The advancement of computers, with the subsequent appearance of the Internet and ongoing progress of data science, is sometimes called the Digital or Third Industrial Revolution. Calling it a revolution might seem a bit preposterous, considering what the previous technological revolutions were and how they changed the fate of *Homo Sapiens Sapiens*.

The Neolithic or Agricultural Revolution took several millennia. Its early sign was the domestication of animals and plants, with dogs being the first around 20 000 − 40 000 years ago. By 10 000 BCE the Neolitic revolution gained momentum, and ended in Mesopotamia around 6 500 BCE [Diamond, 1998]. Writing, math, specialization, social hierarchy, money, calendar and astronomy have appeared there, at that point, and everything that we see now as a 'civilization' is a consequence of that transition.

The Industrial Revolution (also known as the First Industrial Revolution) took much less time, roughly from 1750 to 1850. It spread from Great Britain all across Europe and the United States. The most known change that happened in that period was the transition from handcrafting of goods to machine production. Looking back, we realize that it triggered an avalanche of social, political, scientific and technological events, from horrific, such as the two World Wars, to magnificent, such as the creation of the Universal Declaration of Human Rights and flying to space.

The Second Industrial Revolution was a logical continuation of the First one. It took approximately fifty years before the World War I, and its distinctive trait was the standardization and mass production of complex manufacturing machinery. These two processes made the building of large factories, the rapid expanse of modern transport and communication, and globalization as a whole, possible. During that period, people switched from solar time to railway time, which eventually led to the customary system of time zones. There was no need for this before the Second Industrial Revolution with its railways and telegraph, but after that we all started to live by the same clock. One might consider how it changed our perception of the world, when a concept as fundamental as time stopped being defined by the motion of the Sun and became governed by our transportation.

And then the 20th century arrived, with its numerous scientific breakthroughs. In a sense, they could be considered as revolutions of their own, if only we had enough time between them to fully grasp the scale of changes that they brought. It took one hundred years for the First Industrial Revolution to happen, around fifty for the Second, but the 20th century turned into a parade of game-changing innovations and massive social shifts. Even a quick recall shows that every generation had its own turning point.

First, the development of quantum theory and the theory of relativity had shattered our vision of the deterministic and common sense-based world. This change took around three decades. Then, powered by World War II and the rapid evolution of our understanding of the laws of nature, nuclear physics brought to reality the first potentially unlimited source of energy and the first weapon potent enough to cause mass extinction. Once again, it took only about thirty years, from 1911 to 1942, to walk the path from the Rutherford's publication on his planetary model of the atom, which still was two models away from the correct one, to the ignition of the first nuclear reactor, the Chicago Pile-1. Three years later, hundreds of thousands of people were killed with this new science during the bombing of Hiroshima and Nagasaki.

At the same period of time, two much quieter, yet life-saving transitions were happening: the first mass production of penicillin, also fuelled by World War II, and the first mass vaccination, the one against polio[1]. The middle of the 20th century is sometimes called the Atomic Age, while it should be called the Age of Antibiotics and Vaccines, considering that together with better sanitation they have mostly stopped the epidemics that were plaguing cities since humankind domesticated sheep and chickens.

Soon after that, with the launch of the Soviet satellite `Sputnik 1` in the 1957, the Space Age began. In only four years Yuri Gagarin reached space. In four more, the first spacewalk happened, and in another four, in 1969, Neil Armstrong made his small step on the surface of the Moon. One would have thought that humankind is just a few decades away from becoming a multi-planetary species, but something went wrong.

[1] Strictly speaking, mass smallpox vaccinations were organized since 18th century, but only on a small scale, and only by the mid 20th did it result in the eradication of the disease.

Unlike the new continents discovered by Columbus and Vasco da Gama, space could not offer anything of tangible, saleable value. There were no spices, no cheap gold, no lands to send there convicts for colonization, and therefore no fuel to keep the expansion going. The most breathtaking revolution of the 20th century, the one of dreamers and writers of the Golden Age of Science Fiction, soon came to a halt. Still, from the first satellite to the landing on the Moon, only twelve years passed.

Instead of driving us to Mars, the progress took a different turn. In the early 70's the epoch of personal computers has began. Several technological inventions and, importantly, design solutions, made it possible. The major factor was the progress of semiconductors, an essential ingredient to create small, universal and power-efficient processors; another one was putting semiconductors onto a small plate, e.g. the development of integral circuits. Perhaps, the third in importance was the development of periphery devices and Graphical User Interface (GUI), which made computers user-friendly. In 1974, the Altair 8800 became the first commercially successful personal computer. Soon after that the first Microsoft software - an interpreter for a programming language BASIC - was released. From that point, in less than ten years computers evolved from a bulky, expensive and complicated devices for big business and complex calculations into an expensive hobby, instrument for everyday work, and, eventually, a playground. From this playground the next technological breakthroughs sprung:

- The Internet. Its basis was laid in the 70-s with the APRANET, but its epoch has truly started in the late 80-s, when it became available to the private users and the technological stack for the World Wide Web was created.
- The social networks. Their prototypes existed since early 90-s but their real ascending has happened in early 2000-s. This controversial advancement is unique in many regards. For example, it is the first technological revolution that has virtually nothing to do with the production of goods. Even Internet has began first of all as an instrument for work, but social networks at the beginning were nothing more than entertainment. Another peculiarity is that it is the first technology that is often considered as a defining factor for the whole generation - generation Y, or millennials, - born during its development.
- And Machine Learning (ML) and Big data. They started to affect everyone's life in early 2010-s. We are now in the midst of this revolution, and likely on the cusp of the next one.

The last three technologies, together with a number of others, such as smartphones and Internet of Things (IoT), are considered to be the components of the Third Industrial Revolution.

This short list is by no means complete or objective, yet it shows that in the last century, new pivotal technologies were appearing *at least* every 20-30 years, and, obviously, they continue to evolve and shape humankind even when they are out of their "inflated expectations" phase of the hype curve [Dedehayir and Steinert, 2016]. As a result, it becomes impossible to trace the roots of the further changes[2]. The creation of the theory of relativity lead to the appearance of precise satellite navigation, and it lead to Google Maps in our smartphones [Ashby, 2003]; the horror tales of the Age of Atom (and, of course, a ghost of eugenics; a long echo of the World War II) lead to general wariness towards scientific advances, and this lead to pausing the research of gene therapy of embryos [Lander et al., 2019]; and the beginning of the cosmic age lead to monitoring satellites on the Earth's orbit, resulting in the confirmation of the global climate changes [Ablain et al., 2016, Guo et al., 2015]. But we can see these connections only well in a hindsight.

Obviously, predicting how the current Data revolution will affect the world is near impossible. Yet, everyone try; predictions are made by those who drive this transition as well as those who never wrote a single line of code, and their tone varies from alarmist to Utopian, bringing confusion and producing provocative headlines in newspapers. But to a large extent the field itself is to blame for this confusion; even its basic terminology is somewhat fuzzy. It is easy to determine what is a nuclear reactor, a spacecraft or a computer, but what should we consider as an artificial intelligence, which algorithms belong to ML and how big are the 'Big data'?

1.2 Artificial Intelligence, Machine Learning, Big data

The issue of terminology was raised for the first time by Alan Turing in his seminal paper "Computing Machinery and Intelligence" [Turing, 1950]. In there, Turing starts his considerations with a question: "Can machines think?". Right away he reflects that in order to answer it, we have first to define what it means 'to think', and an attempt to do this is likely to end up in a long and tedious argument. So he pragmatically proposes to use an 'imitation game' instead, now also known as the Turing test. The idea is that a researcher tries to discern between a 'thinking machine' and a person solely through dialogue with both, without visual or voice contact. Although such test equalizes 'thinking' to 'imitating *human* thinking process', in this way, at least, we would not depend on what edition of the dictionary we have at hand.

[2] So called Amara's law states: "We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run".

For several years the field continued developing a number of 'machine thinking' techniques with little need to delve into terminology, until in 1956, during the Dartmouth summer workshop, the term 'Artificial Intelligence' (AI) was coined. The organizers of the workshop picked it in order to avoid biasing towards one of the already existing terms, reserved for a specific research direction. In the funding proposal for the workshop, the term 'AI' is defined indirectly: "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines *use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves*" [McCarthy et al., 2006]. In other words, AI was defined not as just something that can fool a human interrogator, but as something that can operate as a human being, which is much harder (and more obscure) demand. Apparently, the participants of the workshop were content with this, because Marvin Minsky, one of the founders of this field and a participant of the workshop, in 1955 wrote in his report "Heuristic Aspects of the Artificial Intelligence Problem": "I do not feel that it would be at all useful to try to lay down an absolute definition of 'intelligence' or even 'intelligent behavior'... it seems wrong in spirit; we can often find very simple machines which, for certain tasks, exhibit performances which, if done by a man, we would have to call 'intelligent'." [Minsky, 1956]. In more than fifty years that have passed since the Dartmouth workshop, the definition of AI did not become any clearer. In fact, the current edition of the Merriam-Webster dictionary is even more general: "Artificial intelligence. 1: a branch of computer science dealing with the simulation of intelligent behavior in computers. 2: the capability of a machine to imitate intelligent human behavior".

The organizers of the Dartmouth workshop proposed "2-month, 10 man study of artificial intelligence", expressing their belief that "a significant advance can be made ... if a carefully selected group of scientists work on it together for a summer" [McCarthy et al., 2006]. Needless to say, two months were not enough, nor two decades. Interestingly, Turing in his paper in 1950 predicted that 'thinking machines' will appear only by the end of the century (and it was a fairly good estimation, considering the latest advancements in the field). It is an open question, whether the organizers of the Dartmouth workshop actually believed that some significant progress could be made in few months, or they wrote so in the proposal simply because it is much easier to obtain funding if one promises 'a significant advance' and not 'some curious ideas'.

After about a decade of active research it became clear that the possible pace of the AI progress was overestimated. The funding had drained out, and a period called 'an AI winter' has started. Yet, the research did not stop. Looking at Fig. 1.1, it is evident that

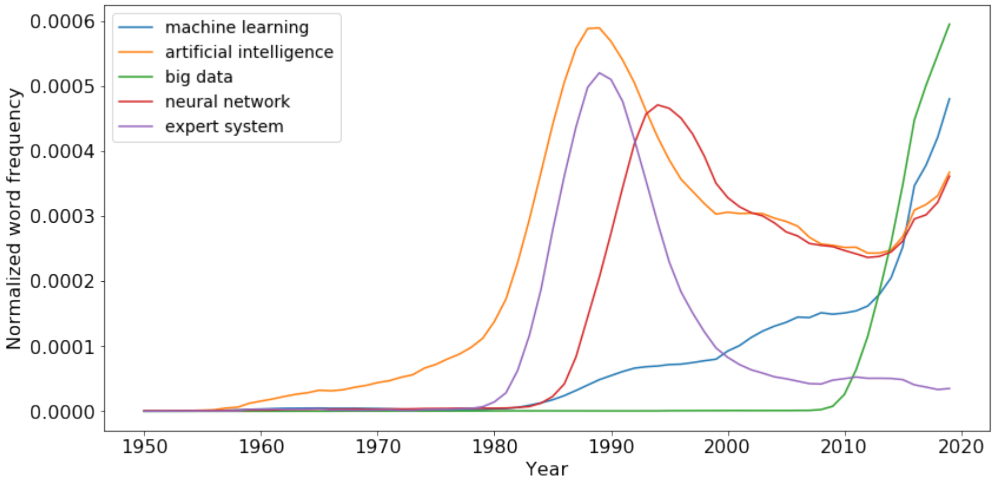


FIGURE 1.1: The frequency of the ML-related terms in literature in 1950-2020. The initial spike of interest towards AI was highly correlated with expert systems, but in a decade neural networks showed themselves as a more perspective technology. The 'AI winter' (late 80's - early 90's), the period of funding cuts of the research on the topic, was caused by the disparity between the original enthusiasm and real progress in the field. However, the studies continued, paving the way for the current advancements. The word frequency data obtained using Google Ngram service; the description of the methodology is given in Appendix A.1

it continued in a different direction and under different name. The AI winter became an ML spring.

Unlike AI, Machine Learning does not aim to imitate an abstract "intelligent human behavior", but merely "extract knowledge from data" [Müller and Guido, 2016], or even more specifically, "automatically extract meaningful patterns from raw data" [Goodfellow et al., 2016, Shalev-Shwartz and Ben-David, 2014]. Merriam-Webster dictionary specifies that Machine Learning is "the process by which a computer is able to improve its own performance (as in analyzing image files) by continuously incorporating new data into an existing statistical model". The common element here is *data*. The AI research at its early stages was mostly concentrated on developing pre-programmed expert systems that would be more or less universal and on incorporating knowledge disregarding its source [Langley, 2011]. Instead, the specific trait of ML is that it changes its own programming (by tuning model parameters) to create the best solution only for a given dataset. From a conceptual point of view, ML is a step backward in comparison with AI as it was understood in 1950s. From the practical point of view, this is what had to be done in order to achieve 'significant advance' in Natural Language Processing (NLP), Machine Vision and AI playing board games and fighting simulators.

Finally, in the early 2000 a concept of big data comes into play.

The term itself appeared in the late 1990s [Mashey, 1998] in the discussions on how to handle the growing volumes, variety and velocity of generation (so called 'three V', see Sagiroglu and Sinanc [2013]) of the data. The 'three V' are sometimes elaborated with other qualities, such as veracity of the data, its value, etc. Considering that the three 'V' have never been about some specific numbers, the big data is probably the most ill-defined term of the Digital revolution.

Speaking of volumes, the most important thing is that they are much larger than what can be handled with an ordinary software installed on a typical computer. Taking into account that 'ordinary software' can be very different for an owner of a small town grocery store and an astronomer working with the last data release of the Gaia survey, the most convenient dividing line is a volume that at least can be stored on a typical computer, which is nowadays of an order of Terabytes (Tb). In the paradigm of big data, it is usually assumed that this volume is distributed across a number of data storage nodes, and smooth work with the data is provided through usage of special technologies, such as MapReduce [Dean and Ghemawat, 2004] or Apache Spark [Salloum et al., 2016]. The key idea of these technologies is that every algorithm is being split to multiple elementary tasks which are performed independently on every node of the data storage cluster, and the results of their work are being combined to provide the final result to the user. This approach is very different from how the large datasets were used before the big data paradigms appeared. For example, scientific datasets, such as those generated by the CERN facilities or produced by astronomical observations, ordinarily were analyzed using supercomputers, and the software for such analysis was developed specifically for each task.

The variety of data assumes that the data can be stored in different formats, that the structure of the dataset is not defined in advance and that these data can be unstructured. It means that the big data algorithms have to be universal and easy to modify in case if some new tasks will appear and require additional knowledge extraction or data fusion, using the data that was not previously present.

Finally, the velocity means that the data is being generated in an increasing speed, and that the algorithms have to be handling it in real time or as close to it as possible. This aspect is relevant only for the fields where the infrastructure allows passing the large volumes of data to the final user; up to date this is one of the bottlenecks for scientific applications. E.g. astronomical facilities often storage multi-Tb datasets, but there is no built-in possibility to download large parts of them; the query interfaces usually have limits of several thousands of records. This difficulty spurred a discussion on the possibility of *bringing code to the data* instead of trying to *bring the data to code*, i.e. of providing computational resources for the users right at the data facility. This solution did not gain any serious popularity in the past; however, similar approach becomes common in computer science and industry. Large technological companies, e.g. Amazon and Google, offer cloud computing services at affordable prices, taking upon themselves the convoluted tasks of programming environment setup and management. Assuming that astronomical organizations will outsource the data storage to them, 'bringing code to the data' might become much more feasible.

Leaving AI, as the most obscure concept, out of consideration, we can notice that in astronomy, elements of ML and Big Data approaches have been in use long before these concepts became mainstream. Indeed, astronomy always had to deal with large volumes of heterogeneous, unlabelled data, and its discoveries often have been about finding a pattern, a correlation or a clustering of the data points in some non-obvious parameter space. So there is no surprise that astronomers have been amongst the first people to try the new technology. But before moving to the astronomical applications of ML, it is useful to consider the basic concepts and definitions that will be actively used in the next chapters.

1.3 Machine Learning basic concepts

As was defined above, ML is an automatic extraction of meaningful patterns from data. This extraction is done via learning process, when an ML algorithm adjusts its own parameters so that the final learned pattern was the best approximation of the dataset under investigation.

The dataset consists of records, or *instances* (usually they are the rows of the table), where one instance corresponds to one data point (e.g. a galaxy, a star, a transient event, etc.); the parameters of an instance (columns) are often called *features* (e.g. magnitudes in different bands or at different moments of time, chemical or morphological parameters, etc.). The features define a *parameter, or feature space* of the problem.

The ML algorithms usually have a number of user-defined *model hyperparameters*, which define the *model architecture*. Model architecture plays an important role in the ML performance, but there are no universal rules on selecting the architecture, so trying various hyperparameters is an essential part of creating an ML solution for the task at hand.

There are several types of ML algorithms:

- **Supervised learning.** For this algorithms the dataset has to contain *labels* (also called *targets*), which are the parameter of the data points that has to be learned. The target can be categorical (e.g. a class of star) or numerical (e.g. star-formation rate of a galaxy). The goal of supervised learning is to be able to predict target value with a maximum accuracy. In the case when the target values are categorical, the task is called *classification problem*; when the targets are numerical, the task is called *regression problem.* Many supervised ML algorithms can work in both regime, but with different setup and accuracy metrics.

  For the supervised problems it is typical to split the data in train, test, validation and run datasets. The train, validation and test datasets compose so-called *Knowledge Base (KB)*; for these datasets both features and targets are known. The train dataset is used for training the ML model; the validation dataset is used to evaluate its performance in each experiment, where various experiments are conveyed to determine the best architecture model. The test dataset is used at the stage of the final evaluation of the selected architecture. The train, validation and test datasets should not overlap, in order to avoid *overfitting* - a condition when the model have learned the 'right answers' only for the data it has seen, instead of learning the smooth n-dimensional correspondence between features and targets. Finally, the run dataset consists of only features; this is the dataset for which ML predictions are needed.

- **Unsupervised learning.** Unsupervised methods work on unlabelled data, looking for the patterns in the parameter space based on the topology, connectivity or local density of the data points of the dataset. The common classes of unsupervised ML are clustering, dimensionality reduction and anomaly detection algorithms. Without any targets to guide the learning process, in the most cases (apart from the simplest ones) there are multiple patterns that unsupervised ML can find in the same dataset. The interpretation of these patterns is a task left for the researcher. For this reason unsupervised ML is often used for data exploration and visualization.

- **Semi-supervised learning.** Semi-supervised algorithms take a large amount of unlabelled data and a small amount of labelled data, and analyze them simultaneously. Unlabelled data provides more information on the dataset topology, while labelled data serves as a guide towards the preferable solution, e.g. a version of clustering that maximizes the distance between two classes of objects.

There are multiple algorithms belonging to each category of ML; choosing the right one and then finding the optimal architecture is an important part of solving the problem, sometimes considered to be 'more art than science' due to its complexity. However, arguably, even more important part of the solution is a good data preprocessing: removing instances with incorrect measurements, filling absent values (or removing instances where some features are not present), normalization of the features, and feature selection.

In theory, it seems to be a good solution to give the ML algorithm all data that is present and to trusts it to find necessary information by itself. In practice, every additional feature brings not only meaningful signal, but also some amount of noise. What poses even more serious problem, though, is so-called *'curse of dimensionality'* [Bellman, 2010]. Its meaning is that the higher is the dimensionality of the parameter space, the larger is the distance between the data points and the harder it is for an ML algorithm to determine the topology of the dataset. The dependency is non-linear; it is often recommended that in order to compensate for the 'curse of dimensionality', the number of the training data points should be at least 5 times larger than the dimensionality of the parameter space [Koutroumbas and Theodoridis, 2008]. Given a fixed number of data points, obviously, it means that thoughtless increase of the number of the features leads to the deterioration of the model's performance.

For the datasets when there are only few features present, this is not an issue. However, ML tasks, including astronomical ones, sometimes deal with dozens and hundreds of features. In such a case, feature selection and sometimes feature engineering is an essential part of the data preprocessing.

Feature selection is a process of choosing which features are going to be used to train the model; the most relevant features can be chosen with automated methods or manually, based on the domain knowledge of the researcher. The aim of this process is to find features that correlate most strongly with the target value, but most weakly with other features chosen for the training (to minimize redundancy) and bring minimal amount of noise. The difficulty of this task is that ML models by construction are able to detect signal which is impossible to find with ordinary statistical methods. Besides, the main interest of using ML is to find some patterns that were not known beforehand, i.e. *not* incorporated in domain knowledge. As a result, brute force search across all combinations of features is the only method that is guaranteed to find the optimal feature set; unfortunately, usually it also so computationally expensive that providing it is absolutely impossible. Less direct feature selection methods, however, often provide satisfying results.

Finally, *feature engineering* is the process of creating new features by combining the old ones, e.g. by summation, division, etc. The usefulness of this process is not obvious; after all, no new information is going to appear if we sum two columns instead of using them separately, and many ML algorithms are doing the same various operations under the hood. Still, there are some benefits. First of all, combining several features might enhance the signal, while reducing the noise (similarly to averaging over several measurements). Secondly, it preserves the signal while reducing the dimensionality of the dataset. Finally, it is possible that the signal is contained in the relation between several features (think of how different types of galaxies are separated with colours, not magnitudes), and by suggesting this relation to the ML model directly we simplify its work. Obviously, the difficulties here are the same as with feature selection.

Now, after creating a rough map of ML concepts, we can look into how ML methods are used in astronomy.

1.4 Prehistory of Astroinformatics

The first mentions of AI application to astronomical problems begin to appear in the 1980-s, although nowadays we would hardly consider the used algorithms as AI ones. Back then, the most perspective direction for the AI research was expert systems; astronomers have developed several simple implementations of such (today we would call them decision trees based on the human-developed heuristic rules) and used them to classify sunspots [McIntosh, 1990, Miller, 1988, 1989b]. Several works have been dedicated to the application of the Bayesian approach: for example, to star/galaxy classification [Sebok, 1979, Slezak et al., 1988] and classification of stellar spectra [Goebel et al., 1989]. Neural networks, apart from a few attempts [see Adorf, 1989, and references therein], have remained unused until the 1990-s; then have started being investigated for classification of the stellar spectra [Weaver, 1990], correction of images obtained with adaptive optics [e.g. Angel et al., 1990, Lloyd-Hart et al., 1992, Sandler et al., 1991], star/galaxy [Odewahn et al., 1992], morphological galaxy [Storrie-Lombardi et al., 1992] and stellar [Hernandez-Pajares et al., 1992] classification, etc.

Interesting to note that apart from expert systems, there was a lot of hopes for Natural Language Processing (NLP). It was expected that advanced NLP software would help astronomers to manage proposal application and review, plan observations, search information in literature and transfer this information into a convenient knowledge base [Adorf, 1991, Albrecht, 1989, Rosenthal, 1988]; the Hubble Space Telescope software infrastructure included several of such systems [see Miller, 1989a]. In other words, astronomers were thinking about systems that would simplify the working routine, rather than classification and regression tasks. Although we have developed non-AI systems for these kinds of tasks, it is hard to argue that there is a lot to do in this direction. Currently, the best NLP instrument that we have is recommendation systems (e.g. such

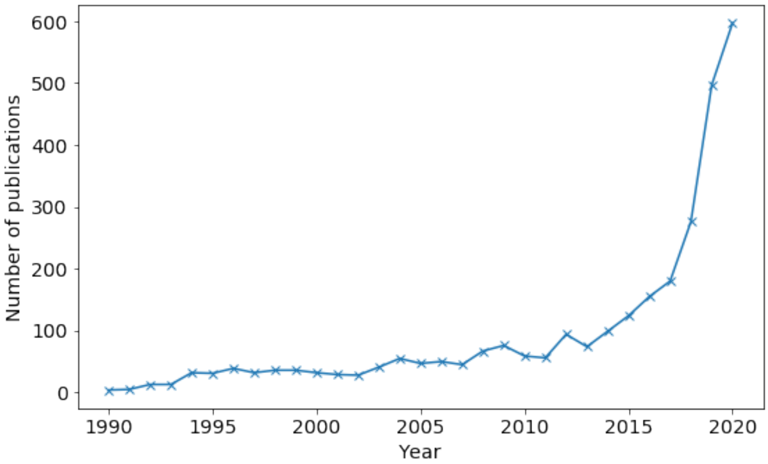


FIGURE 1.2: Number of refereed ML-related astronomical publications by years. The counts are taken using ADS API, the methodology is described in Appendix A.2

as implemented in the bibliography management service Mendeley), but it is evident that it is not enough to handle an information overload from thousands of papers. We may hope that the current advancements of the NLP algorithms will lead to the creation of robust and useful information extraction services, and that it will not take another three decades for such systems to migrate from commercial use to academia.

Starting from the late 1980-s, a number of initiatives was taken to handle the upcoming rise of the raw data. The most known and still ongoing is the Virtual Observatory [Arviset et al., 2012, Szalay and Brunner, 1999], but the growing number of large sky surveys and spread of programming skills amongst astronomers lead to appearance of custom data interfaces for almost every observational program. The abundance of heterogeneous and large astronomical datasets spurred a discussion on a new branch of astronomy called *astroinformatics* [e.g. Borne et al., 2009], by analogy with *bioinformatics*.

1.5 Current status

ML became a popular technology in early 2010-s; Fig. 1.1 shows that the turning point was passed around 2013. Most likely it is related to the progress of machine object recognition that allowed to achieve better-than-human scores in the visual recognition competition ImageNet in 2012. As it can be seen from Fig. 1.2, the astronomical community started to widely adopt these techniques around the same time, although in the last four years the growth sped up even more. A review of the most common ML techniques used in astronomy can be found in [Baron, 2019, Longo et al., 2019]. The problems for which they are used include star/galaxy classification [e.g. Bertin and Arnouts, 1996, Brescia et al., 2015, Cabayol et al., 2019, Jarrett et al., 2000, Kovács and Szapudi, 2015, Odewahn et al., 2004], morphological classification of galaxies [e.g. Barchi et al., 2020, de la Calleja and Fuentes, 2004, Huertas-Company et al., 2008, 2015, Moore et al., 2006, Simmons et al., 2017, Siudek et al., 2018, Vavilova et al., 2017], event detection [Brink et al., 2013, du Buisson et al., 2015, George et al., 2018, Goldstein et al., 2015, Masci et al., 2017, Mukund et al., 2017, Pearson et al., 2018, Schanche et al., 2019, Shallue and Vanderburg, 2018, Zevin et al., 2017] and search for peculiar objects (in a broad sense, e.g. gravitational lenses, galaxy mergers, galaxies at a specific phase of their evolution, fresh craters on the Moon and Mars surface, etc.; see e.g. Ackermann et al. [2018], Huertas-Company et al. [2018], Jacobs et al. [2017], Lanusse et al. [2018], Lee and Hogan [2020], Petrillo et al. [2017], Silburt et al. [2019]).

Calculation of photometric redshift, which is the main subject of this thesis, is sometimes considered a template use case of ML regression problem in astronomy [Brescia et al., 2018], due to the existence of large KBs and relative simplicity of the photometry-redshift correlation (see Connolly et al. [1995] on this subject). Shallow neural networks have been the first ML method applied to this task [Collister and Lahav, 2004, Firth et al., 2003, Tagliaferri et al., 2003, Vanzella et al., 2004] and remain quite competitive on the catalogues with small number of broad photometric bands and representative KBs (see e.g. Dahlen et al. [2013] for a comparison of the ANNz [Collister and Lahav, 2004, Sadeh et al., 2016] to other methods, Norris et al. [2019] for a comparison of the MLPQNA [Cavuoti et al., 2012] to other methods, and Euclid Collaboration et al. [2020] for a comparison of both). Other common ML methods, such as Random Forest [e.g. Carliles et al., 2010, Carrasco Kind and Brunner, 2013] and Support Vector Machines [e.g. Wadadekar, 2005] have also been used.

Photometric redshifts have been one of the few astronomical problems that have been extensively addressed with unsupervised machine learning, primarily with Self-Organizing Maps (SOM) algorithm. SOM have been used for photo-z prediction [Carrasco Kind and Brunner, 2014b, Geach, 2012], but in the last few years it became much more important as a tool for calibrating the photometric datasets to a spectroscopic ones [e.g. Hildebrandt et al., 2020b, Masters et al., 2015, Wright et al., 2020].

Since extragalactic astrophysics and especially cosmology requires reliable photo-z for the large samples of galaxies, including faint and high-redshift ones, there are many attempts to create better photo-z algorithms. Yet, there is not much space for improvements without including additional information. The latter, indeed, sometimes allows to obtain better photo-z predictions, and depending on the nature of this additional information some ML algorithms may be more preferable than other. For example, in order to include morphological information by using raw images as inputs, CNNs are required; Hoyle [2016] and Pasquet et al. [2019] have implemented such schemes and tested them on SDSS data, but the results did not show a convincing improvements over more traditional schemes. Another perspective direction is inclusion of medium and narrow photometric bands to the feature sets; Leistedt and Hogg [2017] have tried this with a hybrid SED fitting/ML methods and Heinis et al. [2016] used SVM (and genetic algorithm for feature selection). Neither approach brought any decisive improvement. Eriksen et al. [2020] achieved much better results with 40-band photometry and deep Mixture Density Network algorithm, but only after using synthetic dataset for pre-training their model. In more detail this subject is considered in Sec. 2.5.

This list of astronomical ML applications is far from a complete one, and new applications of ML to astronomical problems appear every day. There are two conditions, under which a problem can be solved using ML - and with some scientific gain: the problem has to be formalizable, and there should be datasets suitable for ML applications.

The requirement of formalization means that there should exist a clear performance metric: a well-defined classification system, or a value to predict. The simplicity of this requirement is often an illusion. For example, stars and galaxies are physically different, so provided a perfect telescope, we can confidently discern between them. Then, spiral, elliptical and irregular galaxies are *usually* different, but there are intermediate cases. These intermediate cases are physically motivated, since all types of galaxies are nothing more than conventionally chosen points in a multi-dimensional and not completely known continuous sequence of galaxy evolution. And as to the fine morphological galaxy classification, the system which it has to follow is mostly an arbitrary choice. Any attempt to pose this problem as a classification task is inevitably only an approximation of a regression task with an undetermined number of target values, and this approximation is not only rough, but also biased by the instrument, wavelength of observation and the researcher's focus of interests.

Evidently, this should not be a concern of ML *per se*. It is an intrinsic trait of astronomy, a field of science that studies exceptionally complex objects at an unimaginable distances with rather primitive (as our descendants in a few hundreds years will surely say) instruments and imperfect human brains. But this is where the second requirement for a problem comes into play: existence of the training datasets.

From the point of view of ML, astronomical datasets in general are far from ideal. There are several typical issues, that are not being accounted for by the ML developers from other fields, and therefore have to be tackled by the astroinformaticians themselves:

- It is a common case that a large percentage of data points in the dataset contains missing values for one or several features. These data points either have to be discarded from the analysis, or they should be subjected to a process of data imputation.

- Instead of simple Gaussian distribution, errors often have a complex dependency from observable parameters (e.g. magnitude errors may depend on the object's magnitude, angular size, its angular distance from the center of the field of view during the observation, weather conditions, etc.). ML algorithms have different sensitivity to this issue; more so, the majority of ML algorithms do not have common (e.g. included in popular programming language libraries) implementations that would take into account errors of physical measurements. As a consequence, astronomers have to develop such implementations themselves [e.g. Reis et al., 2019].

- Astronomical datasets are intrinsically incomplete, both due to Malmquist bias and due to the selection functions introduces by the observational strategy. ML models inherit these biases.

- Astronomical datasets are imbalanced, in a sense that different classes of objects appear with a different frequency. Sometimes such imbalance is of several orders, meaning that in a dataset with dozens of thousands of objects only a handful belong to the rarest class. ML algorithms tend to always predict more common classes and never 'notice' the rare ones. At the same time, the rare objects pose the biggest interest to science. To tackle this issue, either dataset resampling or model penalization are necessary. More so, we are usually interested in detecting objects that are not just *underrepresented*, but even previously unseen. In this case novelty detection algorithms have to be applied in addition to classification ones.

On the other hand, astronomy, unlike many other areas where ML is being applied, holds a treasure of multiple rich datasets, gathered, analyzed and labelled by experts over many decades. In theory, it would mean that astronomy and data science are bound to a fast and fruitful fusion. Unfortunately, the reality is not so bright.

1.6 Small problems with big data in Astronomy

The notion that astronomy is entering the age of Big data is a common one. A curious thing about it is that in some form, it is being announced for at least ten years now [e.g. Yasuda et al., 2004]. In 2004 - the year when Facebook and Gmail were launched and long before data science became one of the most attractive specializations - SDSS Data Release 3 (DR3) already contained observations of about 141 million of sources, and the preparations for the Petabyte-sized surveys were put into effect [Thakar et al., 2004]. Sixteen years later, big data analysis is a cheap service used by every other merchant, ML-powered algorithms retouch our faces on the fly during online conferences, and the Petabyte-sized survey LSST is on its brush-and-polish stage. Yet, the Catalog Archive System (CAS) and Astronomical Data Query Language (ADQL), developed in 2004 for the SDSS DR3, are still one of the most widely used and user-friendly interfaces for retrieving astronomical data. There are no tools that would, for example, automatically select all merging galaxies observed in HDF, or all craters on the Moon younger than a certain age. In fact, even the simplest tasks, such as crossmatching two catalogues or obtaining a few thousands of galaxy thumbnails, can take several hours (sometimes days, sometimes weeks) if these catalogues are not coming from one of the very few surveys with "big data adapted" interfaces. The world in general *has* entered the age of big data and Machine Learning long time ago; but astronomy is still entering it, and it seems like this transition is going to take a couple decades more.

The reasons to this have nothing to do with technological complications, but rather with the sociology of the astronomical community. First of all, it is small. Depending on the estimation, there are $\sim 10000 - 15000$ astronomers in the world; it is a population of a big village. Turning an innovation into a tool takes research, development, testing and maintenance, and there are not so many human resources in astronomy to invest them into anything non-vital.

The second issue is that astronomy does not bring short-term profit. There are few other kinds of human activity, even among fundamental sciences, that would be so unlucrative. Biology helps to develop drugs, Earth sciences are involved with politics through climate change research, physics provides new materials, but the only measurably valuable product that of fundamental astronomy are well-trained, creative and critical brains, and they are more likely to bring some short-term profit if being relocated elsewhere. All in all, the only regular investors in astronomy are governments, and with no noticeable dividends, the investments are rather small.

A historical observation: in Adorf [1990], an article in the "Artificial Intelligence Techniques for Astronomy" conference proceedings, we can read:

> "Problems in astronomy:
>
> - ...
> - computing in astronomy is lagging behind other scientific areas;
> - astronomical community is so small, no commercial interest, different from e.g. medicine or geology which receive attention in computer science research and from hardware vendors;

- advanced software engineering is too complex and distracting for astronomers, should be left to specialists (astronomers also don't build their own CCD-chips);
- software groups are often understaffed;
- strategic planning is often missing (notable exceptions: Space Telescope Science Institute and NRAO)."

Thirty years later, every single point in this list is still painfully valid.

There is another common notion, this time from data science: in any data analysis or ML project, 80% of the time is spent on finding, cleaning and organizing data, and only 20% on the actual analysis[3]. Searching for the right data can be a tedious task, especially when these data are not a generic dataset from a big observatory but a specific catalogue, prepared by a small team, uploaded only to an institute server, with counter-intuitive column names and mixed `NaN` values, unfindable unless one has read the fine print in the footnote of an application to some paper. Observational data, produced by large facilities, often pose another problem: in the absence of massive batch download and on-server crossmatch functions, gathering information about few hundreds or thousands of objects of interest requires downloading several data packages of tens or hundreds of gigabytes each. After that, selection and cross-matching procedures are performed on the user's computer. Rest assured, the whole process can take 80% of the working time and more.

Speaking from the point of technology, there is a straightforward way to shorten this part of the work: all that is needed is to create a system that would connect various astronomical data storages, guarantee the presence of complete and informative meta-data for the individual datasets, store the relations between individual datasets and support fast access to them. The system would help even in those cases when a simple direct cross-match is impossible, e.g. when it is necessary to find optical counterpart for an objects observed in another range. In such cases, well-organized data infrastructure would at least lessen the volume of the downloaded data by performing an on-server control that the sky areas, observed with two surveys, are overlapping. Essentially, it is a description of an idea of the Virtual Observatory, and similar systems are widely used in industry. However, in astronomy, the implementation of this straightforward idea struggles with incompatibility of various astronomical data storage systems (because they were developed at different times and no one has modernized them), incomplete and obscure meta-data (provided by the authors of the catalogues, not really enthusiastic

[3] Of course, it is just a variation of the well-known Pareto principle.

about uploading them in the first place), and general shortage of resources (especially human ones).

A reasonable question arises, if it is really that important. To answer it, we can make a back-of-the-envelope calculation. In the last decade, about 28000 refereed astronomical papers are being published annually[4]. Let's assume that 10000 of these publications are dedicated to some sort of data analysis[5]. If we manage to speed up the process of this analysis by 10% (by simplifying the process of *finding, cleaning and organizing data*), formally there would be $\sim$ 11000 publications instead[6]; one thousand more observations, discoveries or ideas.

But since astronomy is an extremely non-profitable area, there are very little stimuli for the governments to make investments into enhancements that would allow this to happen, because a few more things discovered about the Universe and a higher number (or higher quality) of scientific papers will not make a visible effect on the society, not in decades or even centuries. And astronomers, being few in numbers and mostly focused on their own projects with little time and efforts to spare, are not prone to fast and effective self-organization. As a result, astronomy was one of the first field to experiment with big data and ML, and will be one of the last to fully integrate these new technologies. Using the terminology introduced by sociologist Everett M. Rogers, astronomy is an early adopter and a laggard at the same time.
Nonetheless, changes do, inevitably, happen, at least, in the observational facilities under construction. One may hope that one day, when ML becomes an astronomer's everyday tool of the same mundanity as ADQL or basic statistic analysis, a modernization of the legacy archives will happen too.

1.7 ML-powered future of astronomy

Currently, the most prominent results in the ML-related fields are obtained with deep learning, e.g. with artificial NNs with large[7] number of hidden layers, and the progress is generally due to the increasing number of model's parameters [Hogarth, 2020]. The

[4] According to the ADS request `collection:astronomy property:refereed year:2010-2019`.

[5] To evaluate whether this assumption is close to reality we can make the following request: `(body:"dataset" or body:"data set" or body:"catalog") year:2019 collection:astronomy property:refereed`. It returns 10765 papers.

[6] Some would say that it would not be a positive change, considering that the information overload is already making it impossible to follow not only the events in the whole astronomical field, but even in several adjacent branches. But this is an issue that we have to solve, either with changing the publishing policies or with machine reading comprehension systems, not run away from it. And hopefully, in reality the number of paper would *lessen*, but they would become more concentrated on physical interpretations.

[7] 'Large' in general means 'more than one'. First DNNs were composed of less than ten layers, but now DNNs can be built of hundreds of layers and more.

number of parameters can go over a hundred of billions in the case of GPT-3 [Brown et al., 2020]; such algorithms require massive datasets and immense computational power, and the cost of training such a model goes over several million dollars. Although only big companies can allow to develop these algorithms, there is a positive side: such a model has to be trained only once, and after that it can be used for a multitude of applications. It means that in the near future an 'ML model rent service' will appear, and top notch AI models will become commonly used. Perhaps, the most appealing possibility for astronomy (and other sciences as well) is a long-desired automatic information extraction and summarization system for publications, which was the top line in the wish lists in the old astronomical AI reviews [Adorf, 1990, Rosenthal, 1988].

There are limits to increasing the number of model parameters, both technological and economical, so the AI community is looking for new approaches. For example, there is a growing interest to combining NLP algorithms with knowledge graphs obtained from various sources, e.g. from social network databases, connectivity of web pages or search query statistics. Unfortunately, even if these systems will be developed soon, astronomy is unlikely to benefit from it in the foreseeable future.

What can be expected is that as already existing algorithms become more and more user-friendly, up to the point where no knowledge of programming or ML theory is required, more and more astronomers will be including ML-based tools to their working routine for quick investigation of the datasets and selection of the objects of interest from the catalogues. Currently the bottleneck of this process is the data infrastructure; however, the new large surveys promise to be more ML-friendly. Another trend that arises for the new observational facilities is that they rely on ML algorithms for the preliminary data processing, e.g. for removing artifacts, classifying objects 'on the fly' and raising alerts for transient events.

One of the most rapidly growing areas of astronomical ML-applications is image processing with DNNs. It is explainable, since DNNs allow to extract task-specific information from the raw images, which previously was possible to do only manually, and only when the relevant image parameters were already known. Another reason for the rapid progress is that deep learning is commonly used in industry and there are numerous algorithms and libraries that can be used 'out of the box' with very little adjustments. However, in many regards astronomical datasets are different from those in industry and other sciences; astronomical data are sparse, often affected by variability, highly unbalanced, and suffer from non-homogeneous and non-Gaussian errors. As a result, astronomy would benefit from mastering less common algorithms that are better at handling these issues, e.g. one-shot/few-shots learning, Gaussian processes and recurrent neural networks.

The near future will bring us Petabyte-scale, multi-messenger, time-domain astronomy. An optimistic person can hope that in slightly more distant future, private space companies will speed up space exploration, and that it will result not only in communication satellites marring astronomical images [Hainaut and Williams, 2020, Tyson et al., 2020], but in some gains as well. Constellations of astronomical CubeSats observing the sky without breaks and in the wavelengths unavailable for the ground-based telescopes [Serjeant et al., 2020, Shkolnik, 2018] would bring the term of 'time-domain' to a new level, and launching complex orbital telescopes would be much less stressful if there was a possibility to convey some construction works already in the orbit. Both advancements would result in even larger growth of the volumes of astronomical datasets, fuelling further advancement of astroinformatics. At the same time, perhaps, one of the next state-of-the-arts DeepMind AIs will be assigned a task from some branch of astronomy - provided that astronomers will be able to formalize it and to prepare suitable datasets. After all, there are attempts to create an AI model that would be able to derive laws of physics from observations [e.g. de Silva et al., 2019, Iten et al., 2020, Wu and Tegmark, 2018], so in principle there is nothing *too fantastic* in imagining an AI algorithm trying to untangle the mystery of dark energy or dark matter by analyzing the astronomical and particle physics datasets simultaneously.

The most incomprehensible part of these changes is not technology though; the appearance of mobile phones and robots was hypothesized well before their actual creation. The most incomprehensible and unpredictable part of this equation are people. No one had predicted anything like WikiLeaks, fake news, Ice Bucket Challenge or video game streaming. The ungoing Data revolution promises - or threatens, depending on one's point of view - to bring immense changes, but truthfully, no one knows what they will be. The unpredictable human reactions plays the major role in this transition, making any prognosis a little more than random guessing. One of the common predictions is that a number of professions, mostly low-qualified ones, will become obsolete in the next twenty years[8], and that it will cause social tension and an epidemic of depression. But who knows, maybe the new generation, growing in an ever-changing world, taught not to trust social media and used to chatting with virtual assistants, will be completely comfortable with this new reality. Maybe they will be a generation of *hikikomori*, recluses living almost completely in the virtual world. Or maybe the opposite will happen, and this generation will dedicate itself to learning new things, arts, volunteering work in the countries that are still not completely over even with the vaccination revolution[9], or citizen science. If something like that will ever be true, astronomical community will finally

[8] The latest success of the AI systems hint that even high-qualified areas, like translating and programming, might become a machine's territory sooner than we think.

[9] The geopolitical consequences of such disparity are another rather scary and unpredictable subject.

have a chance to grow beyond the size of a village; after all, astronomy is not just one of the most unprofitable areas of human activity, but also one of the most inspiring ones.

1.8 About this thesis

The present work is done at the junction of astronomy and data science, fully withing the scope of astroinformatics. It is dedicated to photometric redshifts, a 'template case' for Machine Learning in astronomy.

Robust and reliable photo-z are critical for the new large-scale extragalactic surveys to succeed in their task of constraining cosmological parameters with a few-percent accuracy. However, different scientific problems are studied with different samples of galaxies for which various observables are present, and there is no algorithm that would be a 'silver bullet', a perfect solution for all of them. Additionally, precision cosmology requires not only very precise photo-z catalogues, but also the knowledge of their biases. A careful comparison of photo-z codes and calibration of their outcomes is necessary, and ML methods have been one of the key tools for this. Ch. 2 gives an elaborated review of the previous research done on photo-z, including the existing approaches to obtain them, auxiliary algorithms for photo-z catalogues calibration and validation, and the requirements posed by the upcoming cosmological surveys.

In this work, I use both supervised and unsupervised ML to obtain and calibrate photo-z for the COSMOS2015, a deep 30-band photometric catalogue which is often used as a model dataset for the upcoming surveys. Ch. 3 describes this work in detail, presenting a new ML-based methodology for improving the reliability of photo-z catalogues. For the first time the description of this new methodology was given in Razim et al. [2021] (submitted to MNRAS).

This new methodology is used for several purposes: a) to detect unreliable spec-z measurements from the catalogue used for training and evaluation, b) to estimate the reliability of photo-z predictions without comparing them to spec-z values and to remove those that are likely to be catastrophic outliers, and c) to calibrate the dataset without spec-z information (run dataset) to the one that was used for training. Combined together, different types of cleaning and calibration allow to reduce the percentage of catastrophic outliers by an order, with corresponding improvement of other metrics. As highlighted in Ch. 4, it makes the SOM-based data cleaning methodology a highly perspective tool for preparation of the massive photo-z datasets, e.g. based on the data observed by the LSST and Euclid, and for any survey that combines photometry and spectroscopy coming from different instruments.

Chapter 2

Photo-z

2.1 The Universe in 3D

In the 20th century a number of methods for determining astronomical distances was created, but the absolute majority of them share the same weakness: they work only for the specific types of objects and in a limited range of distances. For example, the method of stellar parallax works only for the closest stars; even now the most precise measurements, performed with GAIA, will allow us to map in 3D less than 1% of the stars in the Milky Way. Measuring distances to the stars on the outskirts of our galaxy requires other methods; local galaxies, more distant galaxies, quasars - every category of objects requires a new method, or rather several, in order for the measurements to be reliable.

As a result, astronomers now use a so called 'distance ladder'. The first rung in this ladder are the approaches that allow us to measure distances to the closest objects. They are used for calibrating the approaches with more far away range of applicability; those calibrated approaches, in turn, are used to verify the techniques for even larger distances, and so on. Finding the overlapping sample that can be used for reliable calibration of the more far-reaching method is often a non-trivial task, so the small sizes of these overlapping samples is the major factor that limits the precision of the overall distance scale.

But more critical issue is that in order to convey modern cosmological and astrophysical research it is not enough to determine distances only to some selected objects. We need them for millions and billions of galaxies, irrespective to their type and whether they contain some 'standard candle'. Among the variety of methods of determining the distances there is literally only one which is applicable to almost any galaxy in the Universe: the one based on the cosmological redshifts, measured either with spectroscopy or photometry.

There are three causes of redshift: a) the relative motion of the emitter and the observer along the line of sight (peculiar motion, which causes classical Doppler effect), b) the difference in the strength of the gravitational fields in the point of emission and the point of observation (gravitational redshift), and c) cosmological expansion of space (cosmological redshift). The measured redshift of any extragalactic object is a sum of these three components, and for the galaxies that are not gravitationally bound to the Milky Way the third component is the dominant one. The cosmological redshift is caused by the fact that when light travels from its galaxy of origin to the observer, it passes through the expanding space. The electromagnetic wave is being stretched together with the space, so the photon's wavelength is being increased, and this increase is proportional to the distance that light has traveled.

Mathematically it is equivalent to the emitting galaxy moving away from the observer [Bunn and Hogg, 2009], so it can be described through the Doppler–Fizeau equation:

$$\frac{\lambda_{obs}}{\lambda_{em}} = 1 + \frac{v}{c} \tag{2.1}$$

where λ_{em} is the emitted wavelength and λ_{obs} is the observed wavelength.

The redshift is measured as the relative change of the wavelength:

$$z = \frac{\lambda_{obs} - \lambda_{em}}{\lambda_{em}} \text{ or } \frac{\lambda_{obs}}{\lambda_{em}} = 1 + z \tag{2.2}$$

From 2.1 and 2.3:

$$1 + z = 1 + \frac{v}{c}, \text{ meaning that } v = cz \tag{2.3}$$

At the same time, the expansion of the Universe causes all galaxies to move away from each other, and the larger the distance between two galaxies, the larger is their relative velocity. According to the Hubble's law:

$$v = H_0 D \tag{2.4}$$

where v us the velocity of the galaxy, H_0 is the Hubble parameter and D is the distance to the galaxy.

As a result, from 2.3 and 2.4:

$$v = cz = H_0 D \tag{2.5}$$

This way, by measuring the redshift of some spectroscopic feature and assuming some value of H_0, we can directly obtain the distances to the galaxies.

This derivations assume flat Universe and require the knowledge of H_0, and these are not guaranteed, so in different branches of astronomy the redshifts are used differently. While in galaxy evolution studies they effectively substitute distances, in cosmology they are used to estimate the cosmological parameters, H_0 included, at different epochs, by comparing them with other distance measurements. This way, redshifts are the last rung of the distance ladder.

The first large spec-z surveys have been finished in the middle of the 20th century [de Vaucouleurs et al., 1976, 1964, Humason et al., 1956, Sandage, 1978]. Thanks to them astronomy started to map the large-scale structure of the universe, investigate galaxy evolution and derive cosmological parameters using large samples of galaxies. However, spectroscopy could not keep pace with photometry; for the massive wide surveys, only about one of one hundred of photometrically observed galaxies has measured spec-z[1], and for many observational programs this ratio is even worse. It is especially hard to obtain spec-z for the faint galaxies, meaning that spec-z catalogues have systematic bias towards bright massive galaxies and low redshifts affecting both astrophysical and cosmological research. For this reason, in the beginning of the 21th century a complementary method, photometric redshifts (or photo-z), came into play.

2.2 General idea of photo-z

The general idea of photo-z is based on the fact that both galaxy spectrum and its multiband photometry are the representations of the intrinsic spectral energy distribution (SED) of this galaxy. The essential difference between them is resolution. Spectrographs allow to distinguish the difference of the energy levels for the features that withstand from one another by angstroms. Photometry, on the other hand, allows to detect only the slope of the SED using the two points that correspond to the median wavelength values of the adjacent bands. These two points typically withstand from one another

[1] This is true for e.g. SDSS Data Release 16, based on the sizes of photometric and spectroscopic catalogues obtained via `http://skyserver.sdss.org/CasJobs/`

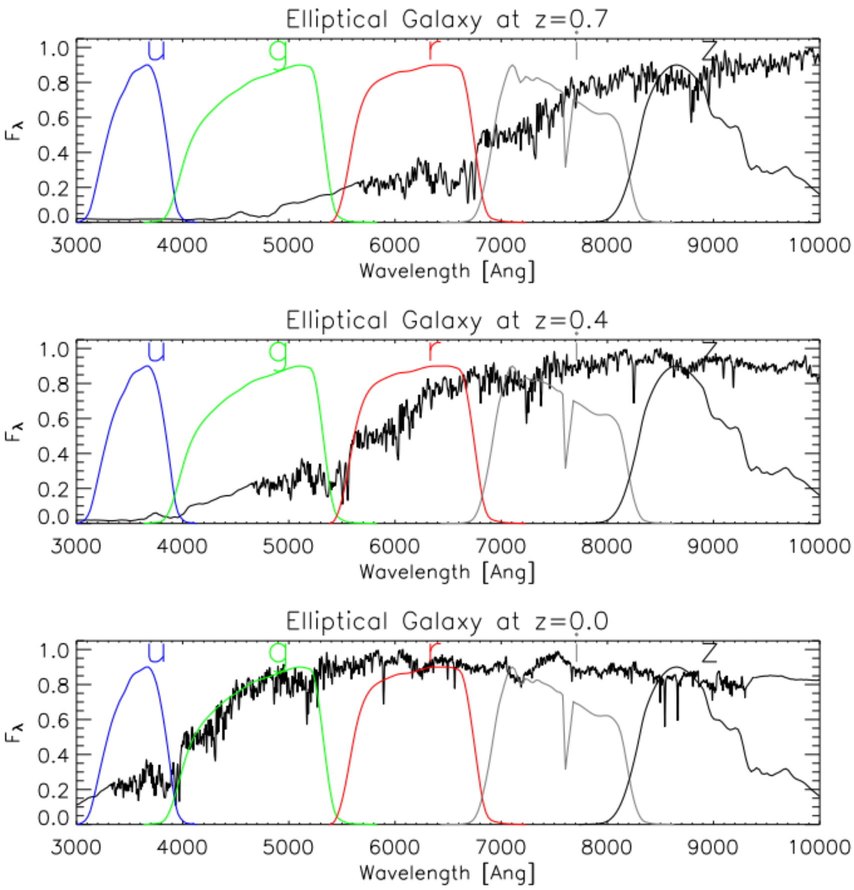


Figure 2.1: A SED of an elliptical galaxy [Bruzual and Charlot, 2003] with the SDSS filter bandpasses overlaid. The SED is shown at three redshifts, increasing from the bottom panel to the top one. With the increase of z, the Balmer break located at 3646Å in the restframe moves from `u`/`g` bands to the redder part of the spectrum. Image taken from Padmanabhan et al. [2007].

by 1 000 – 2 000 angstroms for optical broad bands with a significant increase of this scale in the IR part of the spectrum (e.g. Spitzer bands withstand one from another by $\sim$ 10 000 – 20 000Å); for the optical narrow bands this gap is $\sim$ 100 – 200Å. Galaxy spectrum allows us to measure redshift directly, by identifying a certain spectroscopic feature and comparing its observed wavelength with a wavelength of this feature in a rest-frame. Resolution of photometry, obviously, does not allow that. Instead, when a prominent spectroscopic feature, such as strong emission line or Lyman (912Å) or Balmer (3646Å) breaks, is redshifted, it moves from one photometric band to another, causing an increase or decrease of its flux (see Fig. 2.1). By these changes of the fluxes it is possible to estimate the wavelength of the feature and, consequently, derive the redshift.

There are a number of factors to take into account. Photometry in a certain band is an integrated flux of the galaxy in the band's wavelength range, but it is affected by the intergalactic absorption, Galactic interstellar absorption, atmosphere absorption (in case of ground-based observations) and the instrument's transmission curve in this band. Spectral features can be detected with photometric observations if they are prominent enough to cause a change of flux larger than photometric errors. Primarily such features are Lyman and Balmer breaks for ordinary galaxies and strong emission lines for quasars, although in the last ten years precise and narrow-band photometry also made it possible to 'detect' emission lines in the spectra of star-forming galaxies.

For the first time photo-z method was used by Baum [1962][2]. The motivation for this work was to probe the redshifts of the objects that were too faint for the spectroscopy of that time, with a final goal of distinguishing between the cosmological world models; essentially, this is the most important application of photo-z even now. Baum has measured photometry for 26 color-selected elliptical galaxies located in 8 clusters in the redshift range $z_{spec} < 0.44$, using 9 filters from UV to near-IR. Baum was one of the pioneers of using photoelectric detectors in astronomy, so the photometry was obtained with them instead of photographic plates. The 9 bands were enough to locate the Balmer break; after correcting the photometry for the atmosphere and instrument's transmission curves, Baum has compared the redshifts of the SED curves to the 'zero-redshift' SED of a single star. Despite the crudeness of such calibration, the photoelectric redshift for one of the most distant galaxies of the sample appeared to be in a good agreement with measured previously spec-z, and the subsequent spec-z measurements for most of the other galaxies were also close to the redshifts reported in the paper.

For more than a decade the method has been abandoned, but in the consequent few decades several works have confirmed its viability. In that period two main approaches to the problem have been outlined: a SED template fitting, that compares synthetic galaxy SEDs with the observed photometry, and empirical algorithms, that try to derive the correspondence between colours and redshifts from an observed sample of galaxies itself.

Simultaneously, some limitations of the photo-z methods were discovered. Firstly, it was demonstrated that in the case when the shape of the SED is known (i.e. in the case when the sample is homogeneous; the typical example of such are intrinsically luminous red elliptical galaxies), photo-z can be estimated with a precision $\sigma(\Delta z) < 0.03$ even from three-band photometry [Butchins, 1981, Koo, 1985]. From the simulated data it was shown that for the high signal-to-noise data and assuming no intrinsic variability caused by the evolution of the stellar population, precision of $\sigma(\Delta z) \sim 0.02$ is possible [Connolly

[2] In fact, it seems that the method has been developed earlier. In 1960 Baum and Minkowski already write: "Redshifts above 0.2 can be observed spectroscopically only for objects showing strong emission lines, but for all galaxies of normal color they are readily accessible to the method of multicolor photoelectric photometry. Using this method Baum has observed a redshift of 0.44 ± 0.03 for two average galaxies in the central part of the cluster." Baum and Minkowski [1960]

et al., 1995]. It appears that this limitation is a fundamental one, and obtaining better results is possible (but not guaranteed) only by introducing additional information.

Secondly, it became clear that for the medium-redshift galaxies ($z < 0.5$), 4000Å break (Balmer break) is the most important source of signal for the photo-z methods. This leads to a conclusion that the performance of the photo-z methods depends on the prominence of the Balmer break, which is different for different types of galaxies; for the bright red elliptical galaxies photo-z predictions are much more precise than for spirals and irregulars (the difference in metrics can be higher than one order). High signal-to-noise ratio of the photometry is essential to locate it precisely; additionally, it meant that for the high-redshift galaxies, for which Balmer break falls out of the optical part of the wavelength range, IR photometry is necessary.

Next, in the second half of the 20th century it became clear that the earlier assumption that galaxy type mix have been the same at all cosmological epochs, i.e. at all redshifts, is incorrect. It meant that the SED templates based on the observations of the low-redshift galaxies can be unsuitable for the higher redshifts. Synthetic templates can be created with any combination of parameters (e.g. SFR histories, metallicity, dust extinctions, etc.), but there is a risk of including physically unrealistic combinations, and using such templates may increase the percentage of catastrophically wrong fits [Firth et al., 2003].

Finally, when the observations of the high-redshift galaxies became possible, another issue was discovered: a degeneracy between high-z and low-z galaxies. The source of this degeneracy is the confusion between Lyman and Balmer break. It was evident that for breaking this degeneracy a NIR photometry is needed, and that photometry has to be multi-band in a sense that apart from the three bands to locate a continuum break at least two additional bands are required, bracketing the break region, to determine the continuum slope [Yee, 1998].

For some tasks, primarily cosmological, not only photo-z point estimations are used, but the Probability Distribution Functions[3] (PDFs) as well [Mandelbaum et al., 2008]. PDF describes the probability of a galaxy, given the observed photometry, to belong to a certain redshift as a function of redshift. The shape of the PDF, i.e. whether it is uni- or multi-modal and what is the amplitude and width of its maximums, helps to account for uncertainties introduced by both physical degeneration and photometric errors.

The progress of instrumentation in the end of the 20th century made possible large spec-z surveys. As stated in Giovanelli and Haynes [1991], "the redshift industry is among the most successful, as it can boast a sustained growth rate in excess of 10% per year over

[3] Sometimes they are called Probability Density Functions.

its all 80-years history, and has the potential to maintain its growth for the foreseeable future". For this reason, for a long time photo-z was not used in any cosmological or astrophysical research. The growth of the volume of the spec-z surveys continues up until now (e.g. the ongoing DESI is planned to map ~ 30 million of galaxies and quasars [DESI Collaboration et al., 2016a]), but the 'photometric industry' has been growing much faster than the spectroscopic one. Mosaic CCDs made it possible to observe large fields of sky with a high sensitivity, detecting millions of galaxies that are too faint to be observed spectroscopically. The turning point was the Hubble Deep Field observations, which provided multi-band photometric data for a large sample of faint galaxies; photo-z was the only method for obtaining redshifts for them [Yee, 1998]. In 1998, the first blind test competition of several photo-z methods was conveyed [Hogg et al., 1998]. A sample of 27 galaxies in the range $z_{spec} < 1.4$ was taken from the HDF observations, and it was demonstrated that various photo-z implementations show similar performance. In a few years, following the abundance of high-quality photometric measurements for millions of extragalactic objects, photo-z became a common substitute for spec-z.

Currently photo-z are used in multitude of studies that require good knowledge of the redshift distribution of some sample of objects, rather than high precision of the individual redshifts. These applications can be roughly separated into astrophysical, cosmological and large-scale structure investigations.

The examples of astrophysical questions that widely use photo-z are SFR evolution [e.g. Karim et al., 2011, Magnelli et al., 2009, Pérez-González et al., 2005], properties of dark matter halos [e.g. Hildebrandt et al., 2009, Mandelbaum et al., 2006, Parker et al., 2007, Wake et al., 2011], environmental effects on galaxy evolution, evolution of galaxy mergers, galaxy morphology and physical properties [e.g. Conselice et al., 2005, Guzzo et al., 2007, Ilbert et al., 2013, Lotz et al., 2008, McAlpine et al., 2013, Papovich et al., 2001], luminosity functions [e.g. Buchner et al., 2015, Finkelstein et al., 2015, Ilbert et al., 2010, Magnelli et al., 2011, McLure et al., 2013], evolution of AGNs and their role in galaxy evolution [e.g. Aird et al., 2010, Barger et al., 2005, Parsa et al., 2018, Tozzi et al., 2006, Wolf et al., 2003], etc.

Large-scale structure investigation is concentrating on identification of clusters [e.g. Adami et al., 2010, Arnouts et al., 1999, Hoekstra et al., 2013, Koester et al., 2007, Wen and Han, 2011] and general reconstruction of the cosmic web and its properties [e.g. Darvish et al., 2017, Jarrett, 2004, Kawaharada et al., 2010, Salimbeni et al., 2009].

Finally, photo-z are instrumental for constraining cosmological parameters; the specific applications include already mentioned identification and determination of parameters of galaxy clusters, studying Baryon Acoustic Oscillations [e.g. Abbott et al., 2019, Benítez et al., 2009a, Bernstein, 2006, Chaves-Montero et al., 2018, Zhan et al., 2008] and weak lensing [e.g. Bonnett et al., 2016, Hamana et al., 2020, Hikage et al., 2019, Hildebrandt et al., 2020a, Hoekstra et al., 2002, Hoyle et al., 2018, Schrabback et al., 2010]. A general review of cosmology with photo-z can be found in e.g. [Blake and Bridle, 2005, Weinberg et al., 2013].

Considering the importance of the photo-z for the cosmological research, the best determined requirements for the photo-z quality are coming from this area, especially from the weak lensing surveys. For this reason, this chapter often refers to the demands originating from the weak lensing in application to the determination of the cosmological constraints. The general logic of obtaining and calibrating photo-z remains the same for the other applications, with necessary corrections introduced by the specific traits of the galaxy sample under investigation.

2.3 Metrics and why they matter

The performance of the photo-z methods is usually evaluated by calculating residuals between spec-z and photo-z (eq. 2.6) for some sample of galaxies, and then investigating the distribution of these residuals.

$$\Delta z = \frac{z_{spec} - z_{phot}}{1 + z_{spec}} \tag{2.6}$$

Obviously, it requires a sample of galaxies for which the spec-zs are known, and in the ideal case this sample has to be representative. We cannot extrapolate the photo-z quality obtained for the bright low-redshift galaxies on the dataset containing also faint distant ones. Considering that photo-z method exists precisely because it is often impossible to obtain spec-z for faint distant galaxies, and taking into account that photo-z performance is generally lower for the star-forming spirals and irregulars, which were more common at high redshifts, any photo-z quality estimation is by default an *optimistic* one. Adding narrow band photometry has been shown to improve SED fitting photo-z quality for star-forming galaxies [Ilbert et al., 2009], however, narrow bands are usually shallower than broad bands, and for many surveys are not available at all.

At the initial stages, photo-z research almost completely relied on RMSE as the performance metric. But then it became clear that there are multiple reasons why photo-z predictions can be incorrect, and that RMSE alone is not enough to reveal which ones are the most relevant for a particular dataset. Currently there is a variety of metrics that are used to estimate the quality of the photo-z catalogues. They can be roughly divided in three groups:

1. Bias, traditionally determined as mean $\mu(\Delta z)$. It reflects the systematic error of the method; it is often related to the imbalance of the train dataset. For weak lensing analysis it is important that not only the overall bias, but also the bias in every redshift bin was low;

2. Scatter metrics:

   (a) Standard deviation $\sigma(\Delta z)$;

   (b) Standard deviation over the central part of the Δz distribution. Usually this is the part limited by 68% percentile: $\sigma_{68}(\Delta z)$;

   (c) Normalized Median Absolute Deviation (NMAD), defined as $\mathrm{NMAD}(\Delta z) = 1.48 \times median(|\Delta z|)$. This metric is also less sensitive to catastrophic outliers.

   Scatter errors take root in the intrinsic variability of the galaxies' properties and in photometric errors.

3. Percentage of catastrophic outliers with various definitions of outliers. The most common ones are:

   (a) $\eta_{0.15}$, defined as a number of sources with $|\Delta z| \geq 0.15$

   (b) η_{68}, defined as a number of sources with $|\Delta z|$ outside the central 68% area of the distribution of residuals;

   (c) $\eta_{3\sigma}$, defined as a number of sources with $|\Delta z|$ outside the 3σ range from the mean of the distribution of the residuals.

   Catastrophic outliers appear due to the colour-redshift degeneracy, blended galaxies, exotic sources, such as quasars and AGNs, mis-fit of the galaxy and SED template when SED fitting method used or non-representative KB in the case of ML photo-z methods (see § 2.8.1 for details). In the cosmological publications the definition and classification of outliers can be different and more detailed (see § 2.8).

In some cases additional parameters of the Δz distribution are determined, such as skewness and kurtosis. They are used to estimate whether the residuals follow the normal distribution.

Another matter is how to evaluate the quality of the photo-z PDFs. The PDF of a galaxy depends on its observed parameters, template library (in the case of SED fitting) or the KB (in the case of ML methods and SED fitting that uses observed spec-z for calibration), and on the implementation of the algorithm itself Schmidt et al. [2020]. The issue is that there is no 'perfect' PDF of a single observed galaxy with which the

PDF produced by an algorithm could have been compared. One possible solution to this is to rely on simulations, and another one is to use PDFs determined from a sample of observed galaxies. In any case, there are several metrics for the PDFs proposed, e.g. continuous ranked probability score (CRPS) and probability integral transform (PIT) [D'Isanto and Polsterer, 2018, Polsterer et al., 2016, Schmidt et al., 2020].

2.4 SED fitting methods

SED template fitting, the historically first photo-z approach, requires a library of predefined galaxy SED templates (essentially, synthetic galaxy spectra), obtained either from models or from observations. These templates are being redshifted with a small step and compared with the observed photometry of a galaxy. The best fitting template-redshift combination produces a rough classification of the galaxy and its photo-z.

Butchins [1981] essentially applied SED fitting photo-z method for the first time. He obtained photo-z for a sample of 141 galaxies, both elliptical and spiral, using only `B-V` and `B-R` colors. To do this he used 'simulated' photometry for four galaxy types, obtained through convolution of the observed model spectra with the transmission curves of the telescope. Then this simulated photometry was compared to the observed one, and the best-fit photo-z value was compared to the spec-zs derived from various catalogues. The RMS of the residuals between the photo-z and spec-z was ~ 0.03 in the range $z_{spec} < 0.5$. Based on the experiments with simulated photometric noise, Butchins also suggested that the photo-z scatter is mainly explained by the photometric errors. Then Puschell et al. [1982] obtained photo-z for a number of optically faint radio-emitting galaxies, that were assumed to be elliptical ones. In this paper IR photometry (`JHK`) was used together with the optical bands (`IR`) for the first time. It was also the first work where several SED templates for the same types of objects were tested in order to estimate how the evolution of the stellar populations affects photo-z performance.

Koo [1985] conveyed the first systematic investigation of how various factors affect the performance of the photo-z calculation. The study was performed using synthetic SED templates and considered the contribution of photometric errors, intrinsic variability of the type of the observed galaxy, its dust content and the epoch at which the galaxy was formed. The synthetic SEDs were then used to obtain photo-z for a heterogeneous sample of ~ 100 galaxies, all of which had $z_{spec} < 0.8$. The author used `UBVI` photometry, obtained with CCDs. A year later Loh and Spillar [1986] measured photometric redshifts for ~ 1000 galaxies; they used six bands covering $400 - 950$Å range, and the same model SEDs that were used in Koo [1985].

In the 90th it became clear that in order for SED fitting to be effective, the algorithm has to take into account a number of factors. In practice it means that apart from the mandatory processing - convolution of each template with the instrument's transmission curves and the redshifting itself - additional steps are needed. The most common ones:

- Accounting for the Galactic reddening, caused by the interstellar dust. This effect is coordinate-dependent; to counter it, the input photometry is corrected using Milky Way dust maps [see e.g. Galametz et al., 2017, and references therein].
- Accounting for the intergalactic absorption, caused by hydrogen clouds. This is particularly important for the high-redshift galaxies [see e.g. Furusawa et al., 2000, Massarotti et al., 2001b].
- Correcting the SED templates for the internal extinction, caused by the dust within the observed galaxy [e.g. Furusawa et al., 2000, Massarotti et al., 2001a, Mobasher and Mazzei, 2000].

Various modern implementations of SED fitting photo-z codes account for these and other factors differently; some apply corrections to the SED templates, and others model it as free parameters.

The advancement of the CCDs made possible a) large photometric datasets reaching far beyond the magnitude limits of spectroscopy, and b) large spectroscopic surveys that enabled a thorough investigation of the stellar populations, which resulted in a creation of reliable SED templates. As a result, in the late 90's the photo-z methodology gained its second breath. The rapid growth of the number of the publications can be traced back to the two seminal papers, Benítez [2000] and Bolzonella et al. [2000], both on SED fitting methods.

First SED fitting codes used maximum likelihood (or, equivalently, χ^2) criteria for the comparison of the SED template and the photometry. Benítez [2000] developed a `Bayesian photometric redshifts (BPZ)` code, which introduced a Bayesian approach instead. The benefit of the Bayesian approach is that it allows to utilize additional information obtained elsewhere, such as redshift-dependent distribution of galaxy types (and, consequently, SED templates), magnitude-dependent distributions of redshifts [e.g. Benítez, 2000, Brammer et al., 2008, Coe et al., 2006, Luo et al., 2010] or redshift-dependent distribution of stellar mass and star formation rates [e.g. Tanaka, 2015]. Additionally, the code presented in Benítez [2000] provided full PDFs and a number of reliability metrics, enabling an exhaustive analysis of uncertainties for the further applications. The `BPZ` was tested on HDF photometry with various combinations of photometric bands and SED templates.

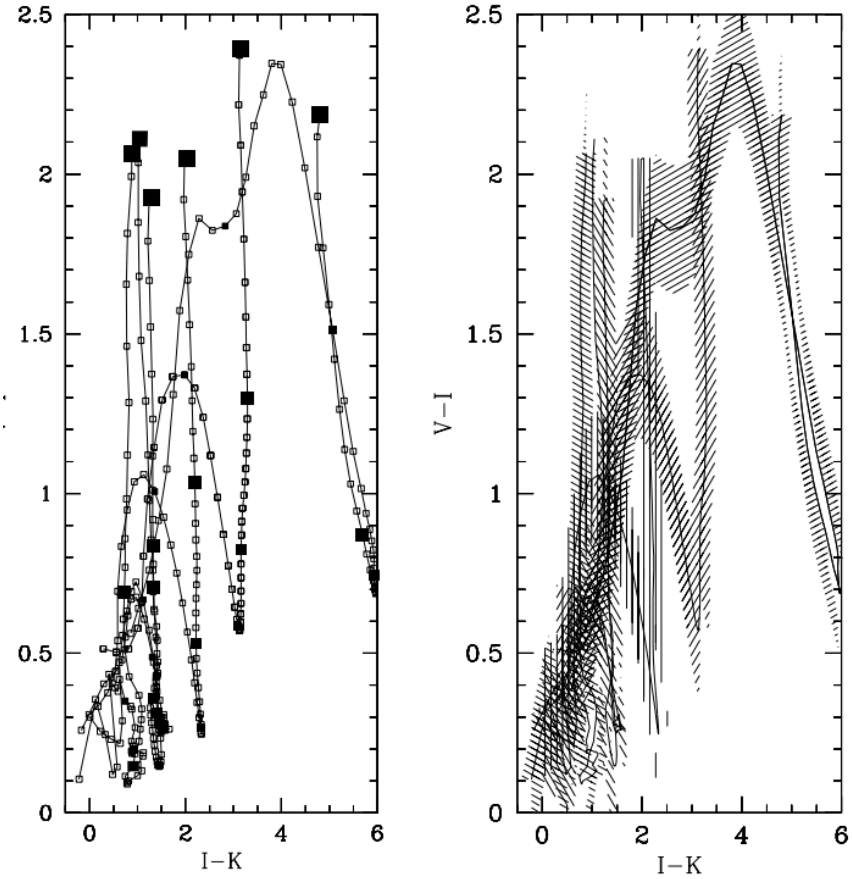


Figure 2.2: Photo-z color-redshift degeneracy in V-I and I-K colour space. The lines correspond to the different SED templates redshifted in the range $1 < z < 5$, the size of the filled squares corresponds to the redshift. Every crossing of the lines means color-redshift degeneracy, where an algorithm cannot discern between the two fits. The right panel shows an effect of photometric errors; it increases the probability of degeneracy. Adding more templates to the library is risky in a sense that they produce more degenerative fits. The solution to this is to use more colours, obtain better photometry and analyze full PDFs where degeneracy appears as several peaks. Image taken from Benítez [2000].

Bolzonella et al. [2000] used simulated and observed HDF data to test the photo-z quality against various template libraries, photometric errors, both random and systematic, dust reddening and physical parameters of the galaxies (e.g. metallicities, initial mass functions, presence of emission lines, etc.). The analysis was performed for narrow-field deep and wide-field shallow cases. The work was done with a new `hyperz` code; the authors opted for a standard χ^2 procedure instead of Bayesian one, motivating it by a notion that introducing some priors to the analysis makes it more dependant on the researcher's hypothesis.

Both `BPZ` and `hyperz` codes were made publictly available, became widely used in the next decades, served as a basis for several other photo-z implementations and remain benchmarks in the blind test competitions up until now. Later many other SED template fitting codes were developed, e.g. ZEBRA [Feldmann et al., 2006], EAZY [Brammer et al., 2008] and LePhare [Ilbert et al., 2006].

The main downside of the SED fitting methods is that they rely on predetermined galaxy SED templates, reddening maps, magnitude zero-points, point spread functions, etc.; in other words, SED photo-z are only as good as our knowledge of astrophysics in general. As was shown in e.g. Benítez [2000], it is especially important that the template library contained templates that are suitable for the galaxy sample under investigation. One might be tempted to add as much templates as possible; synthetic templates can be generated for any combinations of stellar populations and other physical parameters, and then one could trust that the photo-z code will choose the right one for every particular galaxy. But this strategy is tampered by the degeneracies (see Fig. 2.2).

Although template libraries become less reliable in the redshift range that is poorly covered by the spectroscopic surveys, in general SED fitting algorithms can produce predictions in the whole range of redshifts. It made SED fitting methods an important tool for cosmological surveys.

2.5 Empirical methods

Empirical methods aim to establish the correspondence between the colours and redshifts using an observed galaxy sample for which both photometry and spec-z are known, instead of fitting a pre-determined SED templates. First such methods were based on a simple linear or quadratic fit, with a first attempt taken by Connolly et al. [1995]. The authors used spec-z and UBRI photometry for a sample of ~ 300 galaxies, distributed in the range $z_{spec} < 0.4$, to calibrate color-redshift function and obtain photo-z for ~ 2000 galaxies. Importantly, in this paper the authors found out that the distribution of the galaxies in the UBR magnitude space is essentially two-dimensional. The location of a galaxy on this plane depends on its type, luminosity and redshift. The galaxies with a fixed redshift form a 'slab', whose intrinsic coordinates are the luminosity and colors; the 'slabs' corresponding to the different redshifts partially overlap in the color space, because blue distant galaxies, being redshifted, appear similar to the closer red ones. The ability to distinguish between the two overlapping 'redshift slabs' defines the accuracy of the photo-z methods, but the simplicity of the topology of this relation is the major factor that makes the photo-z method possible. Connoly et al. derived photometric redshifts using both linear and quadratic fit, obtained from the spectroscopic sample, with the quadratic fit producing better results (scatter of residuals between spec-z and

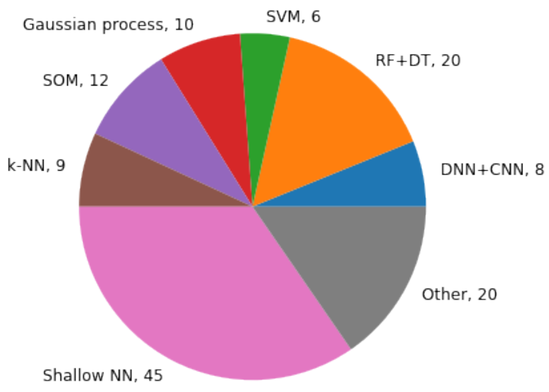


FIGURE 2.3: Popularity of the ML photo-z methods based on the refereed papers published between 2000 and 2020 and available in the ADS system. Number of the papers mentioning the respective algorithms is given after comma. For the methodology see Appendix A.3

photo-z 0.042); they also used simulated data to estimate what precision can be expected from photo-z predictions in an ideal case. Later few more works used polynomial fitting [e.g. Brunner et al., 1997, 2000, Budavári et al., 2005, Li and Yee, 2008, Wang et al., 1998], but starting from the early 2000-s this approach has been dominated by the ML methods.

Essentially, instead of using a finite number of templates obtained manually, ML methods derive an arbitrary number of "templates" right from the dataset under investigation. This change brings a number of benefits. Firstly, "learning templates" from the dataset means that the method does not rely on our knowledge of the galaxy astrophysics. This is important when we move to the higher redshifts, where the galaxy SED templates obtained from the studies of the nearby galaxies are not necessarily suitable. Secondly, the inner "templates" of the ML model come being already calibrated, corrected for the extinction and re-weighted according to the distributions of the galaxies withing this particular dataset. This explains why ML photo-z methods are mostly insensitive to the imprecise determination of the photometric zero-points and selection biases introduced by the observation strategies (provided that they are the same for the train and run datasets).

At the same time, the reliance of the ML methods on the data, and on the spec-z data in particular, is their greatest weakness. The majority of the supervised ML algorithms works well in interpolation regime, but is powerless in extrapolation, so it cannot make adequate predictions for the points that lie outside of the parameter space of the training dataset. For the photo-z case it means that ML models cannot produce meaningful predictions for the galaxies that are dissimilar in terms of their colour-redshift relation to those observed within the spectroscopic program.

Since the first works on ML photo-z in the early 2000-s, a number of algorithms have been adapted for this task (see Fig. 2.3). Briefly review the common ones.

2.5.1 Shallow neural networks

Artificial neural networks (ANN) or simply neural networks (NN) are a family of supervised ML algorithms, inspired by biological neural networks. Considering that in the next chapters I will use an NN-based algorithms, I give here a description of how they work.

ANN consists of nodes called artificial neurons, where each neuron works as a complex adding unit. Lets assume that a neuron takes x_m as feature vector; then the neuron's output will be calculated according to the equation:

$$y = f(\sum w_m x_m + b) \tag{2.7}$$

where f is an *activation function*, which can be linear, logistic, sigmoid or some other, w_m are the *weights* of the neuron, initialized with random values at the beginning of the training, and b is a linear bias.

The neurons of a NN are grouped in layers; the first (input) layer of a network takes the data, and the last (output) layer is composed of a number of neurons corresponding to the number of values (*targets* or *labels*) that have to be predicted. The intermediate layers are called hidden layers; typically the outputs of the neurons of the previous layer become inputs of the next one (Fig. 2.4), although this may depend on a particular algorithm. Types and number of layers and the number of neurons in each layer describe the *model architecture*; together with the neuron's activation function, learning rate and number of training epoch (or a threshold after which training stops) these are the main hyperparameters that have to be pre-defined by the user.

During the training data points are picked from the training set one by one; their feature vectors pass through the network, and the final outputs is being compared with the target values t_n, producing an error value. The function to produce error value is different for different tasks, but the simplest case is the mean square error (MSE):

$$E = \frac{1}{n}\sum_{i=1}^{n}(t_i - y_i)^2 \tag{2.8}$$

Then the weights of each neuron are being updated in such a way that the error would lessen. It is done by calculating the small addend $\Delta w_{i,j}$ for every weight value w_i of every neuron j. The simplest way of doing it is by using *gradient descent*, i.e. by calculating the slope of the error function in the current point and by moving in the direction of the local minimum:

$$\Delta w_{i,j} = \eta \frac{\partial E}{\partial w_{i,j}} \tag{2.9}$$

Since neurons of the 'later' layers take as inputs the values produced by the 'earlier' layers, updating weights goes in the backward direction: first the updates for the output neurons are calculated, then the updates for the last hidden layers and so on. This mechanism is called error *backpropagation*.

The goal of this process is to calculate such weight values that the NN was always producing the outputs that would be as close to the target values as possible; in other words, the complex non-linear function that describes the NN's outputs has to approximate the training data.

Shallow fully connected neural network is the simplest case of NN. Classical definition states that shallow neural network has only one or two hidden layers, but in this paragraph I will also include the cases of NNs with several more hidden layers, if these layers are all ordinary dense ones, without convolution, pooling or some other special layers. The reason for this deviation from the ordinary terminology is that in the last several years the term 'deep neural network' almost always assume large number of layers (more than ten) and complex architecture.

Shallow NNs have been the first ML algorithm applied to photo-z problem; the first papers were published in 2003 [Firth et al., 2003, Tagliaferri et al., 2003].

Tagliaferri et al. [2003] used a shallow NN with one hidden layer to obtain photo-z for SDSS Early Data Release [Stoughton et al., 2002]. The authors have tried various feature sets, combined of `ugriz` magnitudes in total and petrosian apertures, fluxes, etc., described the complications introduced by the imbalanced distribution of the target value, i.e. spec-z, and applied an unsupervised Self-Organizing Map algorithm to investigate the resulting photo-z catalogue's quality.

Firth et al. [2003] tried several model architectures, arriving to the conclusion that 3 hidden layers bring the best results and additional increase of the depth does not improve the photo-z. They also averaged the predictions of several NNs with the same architecture, but different initial weights, effectively introducing a simplified ensemble

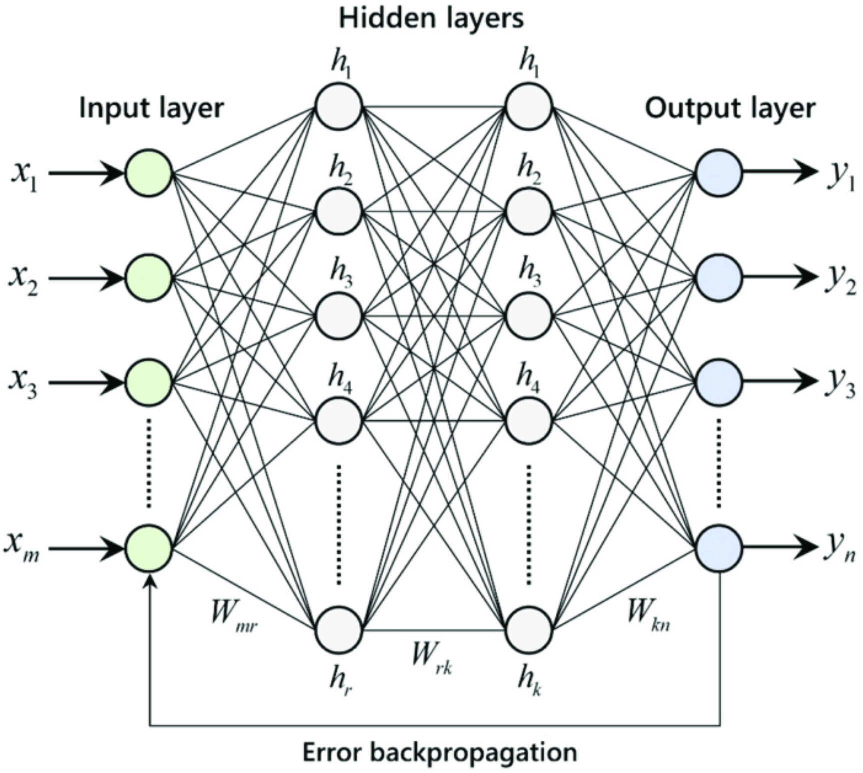


Figure 2.4: Fully connected shallow neural network. Image taken from Fernández-Cabán et al. [2018]

approach[4]. The investigation was conveyed for the 6-band mock photometry as well as for `ugriz` photometry for the galaxies from the SDSS-I limited by spec-z$<$0.5. The NN model showed comparable or sometimes better results than the SED fitting algorithm `hyperz` [Bolzonella et al., 2000]. The authors also investigated whether some morphological parameters, namely Petrosin radii and galaxy/star classification flag from the SDSS, improve the performance of their NN, and found that although in some redshift ranges it slightly improves the results, the overall statistics do not show reliable improvements. Authors explain it by the possibility that the benefit of adding morphological information is outweighed by the additional noise that comes with these features.

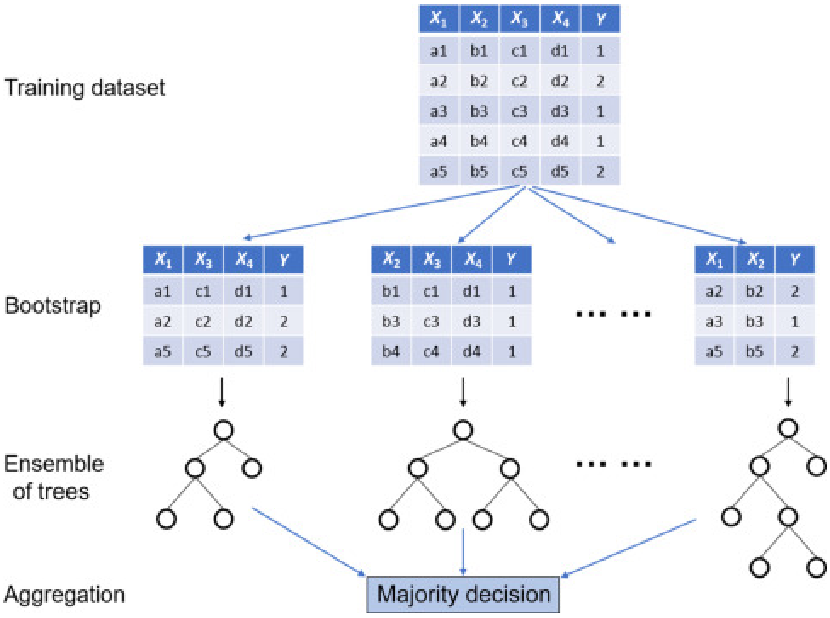
Later on a number of works have been investigating shallow NNs performance; the commonly used implementations are simple Multi-Layer Perceptron [Cavuoti et al., 2012, Tagliaferri et al., 2003, Vanzella et al., 2004] and ensemble-based ANNz/ANNz2 [Collister and Lahav, 2004, Sadeh et al., 2016].

[4] Ensemble approach in ML means using several ML algorithms, or the same algorithm but with different model architectures, to obtain predictions for the same data, and then averaging these predictions. It is based on a hypothesis that several weak predictors may compensate each other's errors.

2.5.2 Deep neural networks

Deep neural networks (DNNs) work on the same principle as shallow NNs, but the number of hidden layers is higher, and usually the architecture of the model is more complicated than for shallow NNS in a sense that there are various types of layers. DNNs are usually applied when the training data are images, data cubes, spectra, brightness curves or some other ’raw’ data, since their depth and flexibility of the architecture allows DNNs to extract useful patterns from such data. The ’useful patterns’ in astronomical data can be texture and asymmetry of galaxies, concentration index, distance between spectral lines in spectra, etc.; in other words, all the parameters that otherwise would have been extracted by some other software and often semi-manually. Elimination of this semi-manual or ’other software’ step minimizes biases and loss of information, enabling the DNN to fully utilize all the information contained in the data.

DNNs are used for photo-z predictions exactly with the hypothesis in mind that the raw images may provide more information than photometric and morphological parameters in a tabular form. Theoretically, morphological data may help to break colour-redshift degeneracy, which would reduce the percentage of outliers, especially for the datasets that occupy large redshift range. Improvement of the scatter of the residuals also possible, since DNN provided with the raw images of resolved galaxies should be able to perform internal ’galaxy classification’, which would compensate for the intrinsic variability of spirals and irregular galaxies. For this reason, several groups derived photo-z with DNNs. Hoyle [2016] has done it for ~ 60000 galaxies from SDSS Data Release 10, using differences between adjacent `griz` bands as three color channels. D’Isanto and Polsterer [2018] used Convolutional Neural Network (CNN) together with a Mixture Density Network (MDM) to derive photo-z PDFs for a dataset containing ~ 300000 quasars, ~ 200000 galaxies and ~ 200000 stars from SDSS Data Release 9; `ugriz` images as well as their differences were used. Pasquet et al. [2019] used only `ugriz` images without ’colour differences’ of ~ 500000 galaxies taken from SDSS DR12. In all three cases DNNs have shown the performance similar or better to those of reference ML models (RF, AdaBoost and k-NN), but not *significantly* better. It is also worth noticing that no DNN method used in any blind test competitions and their performance was not compared to the performance of the common photo-z codes. Considering that DNNs need significantly more computational resources (by few orders more in comparison to simpler ML methods) and large datasets of image cutouts (which is hard to prepare for the absolute majority of surveys due to the absence of data-sceince adapted interfaces; one may notice that all three works are performed for SDSS data), the gain does not seem that convincing. It is possible, though, that the situation will change once more



X_1	X_2	X_3	X_4	Y
a1	b1	c1	d1	1
a2	b2	c2	d2	2
a3	b3	c3	d3	1
a4	b4	c4	d4	1
a5	b5	c5	d5	2

X_1	X_3	X_4	Y
a1	c1	d1	1
a2	c2	d2	2
a5	c5	d5	2

X_2	X_3	X_4	Y
b1	c1	d1	1
b3	c3	d3	1
b4	c4	d4	1

X_1	X_2	Y
a2	b2	2
a3	b3	1
a5	b5	2

FIGURE 2.5: Random forest. The algorithm uses different sets of features X_i for every decision tree, and picks random data points for their training. This process is called bootstrapping. The resulting decision trees have different architecture, i.e. they use different sets of questions to make the classification. Then the predictions are being aggregated. Image taken from Misra and Li [2020].

surveys will develop better interfaces. Simulated cutouts also can help for pre-training the DNNs.

2.5.3 Random forest

Random forest (RF) is an ensemble method based on the decision tree algorithm.

Every decision tree passes the data through a sequence of questions with a fixed number of possible answers; the following questions depend on the answer for a previous question. For example, for the photo-z problem the first question can be 'is `mag` lesser than `median mag`?'; if the answer is positive, the next question asks whether the same `mag` is lesser than the median value of the new `mag` range, and so one until the range of the `mag` of the galaxy is not determined with a pre-defined precision. After doing so for every feature, the dataset ends up distributed in multiple *leaves* where each leaf corresponds to a small area of parameter space. This leaf is assigned median or mean spec-z value of its galaxies, and all galaxies from the run dataset, that will be classified as belonging to this leaf, will get this spec-z value as photo-z prediction.

RF creates multiple decision trees, that are being trained on different feature sets and on various random samples of data taken from the whole training set. Then the predictions of all decision trees are being combined together; multiple rough predictions are being averaged, producing a 'most likely' outcome, and outlying predictions, made by a few number of decision trees, are being down-weighted. As a result, RF is not prone to overfitting and not very sensitive to the noise in the training dataset. Another advantage of this algorithm is that it is computationally effective.

RF was used to obtain photo-z for large samples of galaxies taken from SDSS [Carliles et al., 2008, 2010] and DES [Sánchez et al., 2014]; implementations for obtaining PDFs were also created [Carrasco Kind and Brunner, 2013]. Due to its simplicity, robustness against outliers and modest computational requirements, RF often serves as an ML benchmark when testing other ML algorithms [e.g. Almosallam et al., 2016a, Carrasco Kind and Brunner, 2014b, Cavuoti et al., 2017b, D'Isanto and Polsterer, 2018, Pasquet-Itam and Pasquet, 2018].

2.5.4 Support Vector Machines

Support vector machines (SVM) [Burges, 1998, Cortes and Vapnik, 1995] is a family of algorithms that look for a hyperplane that would separate two samples of data points, belonging to the different classes. The learning process consists of finding such a hyperplane that would have a maximum margin between this hyperplane and the closest data points (Fig. 2.6). In the case when it is impossible to find a separating hyperplane in the original 'flat' parameter space, SVM attempts to map it to a feature space of higher dimensionality using a pre-defined kernel function, enabling in this way non-linear classification. A distinctive trait of these algorithms is that the kernel function and its properties are the only hyperparameters that have to be pre-determined. Also SVMs computational cost increases almost linearly with the increase of the parameter space dimensionality, so these algorithms are suitable for the data with large numbers of features.

The original version of SVM is a binary classification algorithm, but for photo-z task either a multi-label classification (where labels correspond to small ranges of redshift) or a regression version is used. The regression SVM attempts to find such a hyperplane that the data points were located within a certain small error margin from it. In Wadadekar [2005] SVMs were used on a set of $\sim$ 20000 bright galaxies from SDSS DR2, limited to $z_{spec} < 0.5$, and on a simulated dataset. The authors found that on this datasets adding 50% and 90% Petrosian flux radii to the `ugriz` photometry helps to lower scatter of predictions. Wang et al. [2008] used SVM for obtaining photo-z for $\sim$ 70000 quasars from SDSS DR5, and Han et al. [2016] used $\sim$ 106000 quasars from SDSS RD7 to show that a combination of SVM and KNN algorithms can reduce the percentage of catastrophic outliers in photo-z predictions by almost 10%. Finally, Jones and Singal [2017]

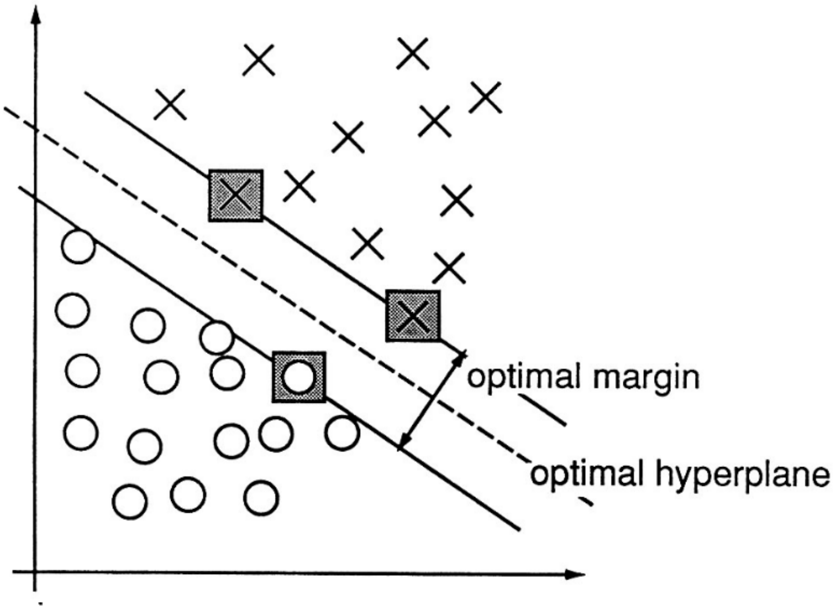


FIGURE 2.6: Support vector machines principle. The algorithm attempts to find a separating hyperplane with a maximum margin between it and the closest data points. Image taken from Cortes and Vapnik [1995].

developed a publicly available SVM code for photo-z calculation called `SPIDERz`; they tested it on several datasets, including PHAT-1 (18-band photometry from the GOODS survey), used for the blind competition in Hildebrandt et al. [2010], COSMOS2015 (30-band multi-source photometry) and AEGIS (5-band `ugriz` photometry). The authors also tried to add morphological parameters to the feature sets. In terms of the general performance the algorithm was competitive with other codes, both ML and SED fitting ones, but adding morphological parameters lead to the deterioration of the statistics. The authors concluded that noise of the morphological data outweights the benefit of additional information.

2.5.5 Gaussian processes

Gaussian processes (GP) [Rasmussen and Williams, 2006] is a family of Bayesian ML algorithms that aim to outline all the possible functions that can describe the correspondence between the data and target values (see Fig. 2.7). This approach allows to incorporate photometric errors in the analysis, and its outcome also includes uncertainties of the predictions. It is less common and more computationally expensive than the methods described above, although there are adapted for the photo-z GP models with a lower computational complexity [e.g. Almosallam et al., 2016b, Way et al., 2009].

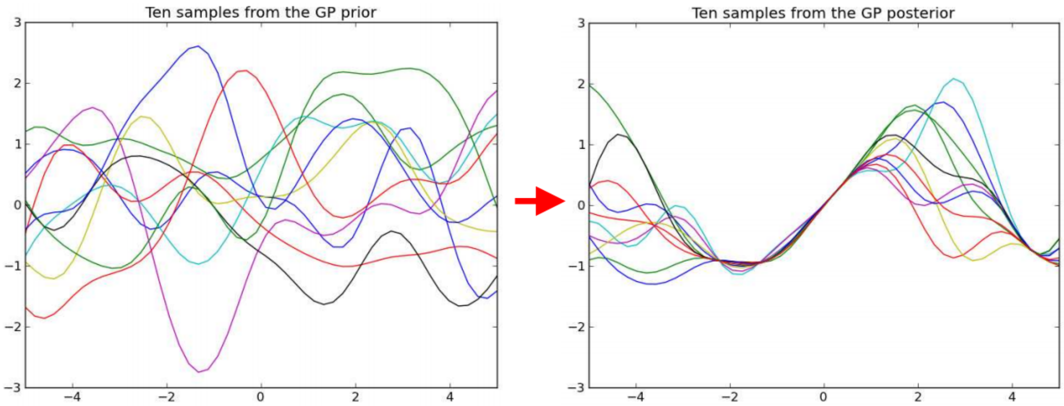


FIGURE 2.7: Gaussian process starts with a sample of possible functions defined by *a prior* probabilities, and then narrows it down, using the train data as conditions for calculation *a posterior* probability. In the ranges of input parameters where training data points are absent, the predictions will be given higher uncertainty. Image taken from de Freitas [2013].

Way and Srivastava [2006] applied GPs to the photo-z predictions for the first time; the authors compared it with linear and quadratic fits and shallow NNs and found that GPs perform better on the small datasets. Several feature sets were used, composed of SDSS, GALEX UV and 2MASS IR photometry and several morphological parameters. This work was continued in Way et al. [2009], where the authors used a modified, more computationally effective version of GPs; they also found out that adding morphological information to the feature sets leads to the deterioration of the photo-z statistics. Bonfield et al. [2010] also found that GPs are preferable over NNs on the small datasets and in the cases when there are undersampled ranges of redshifts (e.g. 'redshift desert' in the range $\sim 1.3 < z < 1.7$, where spectroscopic measurement in the optical wavelengths is complicated by the absence of the prominent spectral lines). Almosallam et al. [2016b] created a GPs code that takes into account uncertainties introduced by non-uniform data noise and non-uniform density of the data points. Finally, due to its good performance on sparse datasets, GPs were used by Duncan et al. [2018] to predict photo-z for IR, X-ray and optically selected AGNs. In some ranges (e.g. $1 < z < 2$), GPs showed significantly better performance than SED fitting method; however, for the galaxies that do not fit any AGN criteria, GPs showed worse scatter metrics than SED fitting.

2.5.6 Other algorithms

A number of other ML algorithms were used for prediction photo-z, although they are less common that the ones described above.

- **k-Nearest Neighbours (k-NN)**: k-NN assumes that the data points that are close in the parameter space will have similar target values. The proximity in the parameter space is usually measured with Euclidian distance. In the case of regression problem, such as photo-z, the predicted value is the mean of the target values of k nearest neighbours from the train set, where k is user-defined parameter. k-NN is commonly used for predicting photo-z for quasars [e.g. Ball et al., 2007, Curran, 2020, Richards et al., 2001, Zhang et al., 2013].

- **Self-Organizing Maps (SOM)**: SOM is an unsupervised method for data visualization and clusterization. It also groups similar data points together, but the grouping rule is more complicated than in the case of k-NN, and as a result SOM preserves intrinsic topology of the dataset. More detailed description of the algorithm is given in § 3.2.3. Nowadays the SOM is mostly used as an auxiliary algorithm for dataset calibration (see § 2.8), but Way and Klose [2012] used SOM to produce photo-z point estimations for the SDSS and PHAT0 galaxy samples and showed that they are of similar quality as those obtained with other photo-z algorithms. As was demonstrated in Carrasco Kind and Brunner [2014b], SOM also can be used for obtaining photo-z PDFs.

- **Boosted Decision Trees**: As was explained in § 2.5.3, decision tree is an algorithm that performs classification or regression through a series of 'yes or no' questions, that form a unique 'decision path' for every small range of the target value. Boosted Decision Trees is a modified version of this idea; after regression is finished for the first time, the algorithm repeats it, giving larger weight to the data points that were misclassified (or catastrophic outliers, speaking in the photo-z terms). Gerdes et al. [2010] successfully used this approach to derive photo-z estimations and their PDFs for SDSS galaxies as well as simulated SDSS and DES datasets.

2.5.7 Feature selection

Regardless of which algorithm we use, the quality of the photo-z predictions mostly depends on the data: on its quality, e.g. on the photometric errors, and on the presence of the redshift signal in the algorithm's input features. As was demonstrated in the early works on the topic [e.g. Butchins, 1981, Koo, 1985], even a few bands can be enough to derive photo-z of acceptable quality. An important nuance here is to take the right bands; e.g. for the low-redshift galaxies optical bands are enough, but for the galaxies with $z > 1$ IR photometry is needed to keep tracking the redshifted Balmer break. In other words, on the most basic level feature selection for photo-z calculation is done based on the general reasoning about the galaxies' astrophysics.

However, once we attempt to refine our predictions and utilize information contained in all the measured galaxy parameters, we cannot decide which features will be most useful from the physical considerations alone. Even if we have only magnitudes, it is not uncommon to have them measured in several apertures, so which ones should we use? Or maybe it is beneficial to derive ratios between the magnitudes in two different apertures, using it as a homemade concentration index? Should we use magnitude errors as inputs, giving an algorithm a hint on which magnitudes can be trusted, and which ones are likely to be imprecise?

In the case of SED fitting algorithm, there is not so much freedom in answering these questions. Different apertures can be tried one by one; magnitude errors can be included in the PDF production; and any additional information can be introduced only in a form of preliminary derived Bayesian priors. ML algorithms, on the other hand, in the general case can accept any kinds of features, and also their derivatives, in any combinations. It is possible to use all magnitudes in all apertures, errors, ratios, differences, square roots, physical parameters coming from any source, etc. And it is tempting to try these additional features; for example, we know that galaxy type mix is different at different redshifts, so why not to use some morphological features or star-formation rates for a better constrain of the redshift bin?

The question remains, though: how do we find the features that can bring the best result? Using all possible features is not a solution. First of all, the higher is the dimensionality of the data, the longer it will take for an ML model to train. But more importantly, due to the 'curse of dimensionality', expanding the dataset's feature space can worsen the performance of the model, so it is a good practice to keep the number of features as low as possible.

Some ML algorithms, e.g. CNNs, by themselves extract meaningful features from the raw images (although it is worth noting that this mechanism is imperfect and some 'feature engineering' in the form of e.g. brightness adjustment can be beneficial). In other cases, we have to determine the feature set before training the model. In the case when the number of features is small, we can do it by brute force search, i.e. by trying all possible feature combinations and comparing the model's performance for all of them. But the number of possible unique combinations of k out of total n features grows as $\frac{n!}{k!\cdot(n-k)!}$, meaning that once we have $n > 5 \sim 10$, brute force search becomes too computationally expensive.

There are multiple automatic methods of feature selection, starting from the simple ones, based on feature correlations or dimensionality reduction, and spanning to the complex ML-based solutions (see e.g. Guyon and Elisseeff [2003] for a general overview and Donalek et al. [2013] for the one on astronomical applications). However, as with any other data science task, there is no method that would always be better than others. More so, a feature set that was identified as optimal for one dataset is not guaranteed to be suitable for another dataset. It is fairly simple to explain using photo-z as an example: two surveys will have non-identical magnitude and redshift distributions, targeting strategies, instrumental biases, transmission curves, etc., so a band that holds a strong redshift signal in the first survey can be noisy and useless in the second one. Finally, it is possible that a feature can be of low informational value by itself, but useful in combination with some other feature.

Due to these complications, there are very few works investigating the benefits of feature selection for the ML photo-zs. Hoyle et al. [2015] used Boosted Decision Trees (BDT) to select most informative features for the SDSS. The authors investigated 85 features, including magnitudes in different apertures, colours, radii, ellipticity, etc., and discovered that adding several of the best BDT-determined 'best features' to the standard feature set of magnitudes improves the statistics of the photo-z obtained with NN-based code by $10-30\%$. Heinis et al. [2016] used Genetic Algorithms together with SVM to choose best features among magnitudes, colors, and also their logarithmic and exponential transformations (978 features in total) measured in the 25 broad, medium and narrow bands in the COSMOS field. The authors showed that the colours in narrow and medium bands bear maximum importance, but they do not compare their results, obtained with the 45 best features, with any 'standard feature set' of smaller size. Finally, Stensbo-Smidt et al. [2017] and D'Isanto et al. [2018] used forward feature selection. The idea of the forward feature selection is to start with only one feature and evaluate its predictive power; then take the best feature and use it together with another feature, going through all possible combinations; then take the best combination and add to it one more feature, and so on. Since this method is computationally demanding, the authors of both papers had to use simple and computationally inexpensive photo-z codes (Stensbo-Smidt et al. [2017] used k-NN and D'Isanto et al. [2018] used k-NN and Random Forest), and even then a massive parallelization was required. Both groups used SDSS data; Stensbo-Smidt et al. [2017] experimented only with magnitudes and colours in various apertures, while D'Isanto et al. [2018] included also magnitude errors, radii, ellipticity and extinctions, and a number of feature transformations, e.g. ratios and differences. Both studies showed that using 'best' feature sets improves the statistics in comparison with the standard ones by at least $10-20\%$, and that these 'best' features are not those that would be chosen from general reasoning.

From these works it is evident that automatic feature selection, indeed, can help to improve photo-z quality. However, their high computational cost and the necessity to perform this analysis for every dataset separately makes its wide spread unlikely.

2.6 Hybrid approaches

Both SED fitting and empirical methods have their pros and cons. A question arises, whether is is possible to construct a methodology that would combine both approaches in a way that would compensate the original methods' weaknesses. Several such attempts were made, using various approaches, e.g.:

- Empirical methods are used to derive SED templates from the dataset composed of observed spectra. Then these templates are used for SED fitting. The benefit of this method is that the templates are by definition physically realistic and do not rely on our knowledge. The downside of it is there is no guarantee that the templates derived from the low-redshift data will describe high-redshift galaxies well enough, and high-redshift data are often too scarce to produce good and representative empirical templates. Csabai et al. [2000] used this method on simulated and observed date in the HDF. The authors derived so-called *eigenspectra* which were used to produce SED templates, and showed that these templates provide better performance than those used in previous studies and obtained by traditional means.

  There is a variation of the previous approach, when the process starts with one pre-determined SED template, which is being adjusted for different subsets of the whole dataset. Such methodology was applied to quasars [Budavári et al., 2001] and SDSS galaxies [Csabai et al., 2003, Padmanabhan et al., 2005]. There is also a more recent implementation of this technique that uses SED template-guided Gaussian processes to derive new templates from the data [Leistedt and Hogg, 2017].

- SED template fitting is used for galaxy classification. Then photo-z ML algorithm is trained on the galaxy samples corresponding to each galaxy type separately. The benefit of this approach is that ML model is trained on a more homogeneous datasets, and it lowers the demand to the model's generalizing capability. It is especially important when there are some underrepresented types of galaxies, which would not be easy to learn for the model when mixed with other, more common types. The downside is that if the initial classification is mistaken, then every misclassified object will become a catastrophic outlier.

  Cavuoti et al. [2017b] used this approach with `ugri` KiDS photometry and compared its performance with the performance of the purely ML and SED fitting codes. The authors discovered that hybrid approach improves both scatter and outlier metrics.

A common variation of this approach is when the classification at the first step is done not with SED fitting, but with unsupervised ML instead [see e.g. Speagle and Eisenstein, 2017].

- SED templates are used to produce a mock dataset. The ML algorithm is pre-trained on this dataset, and then (if it is possible) additionally trained on the observed data. In other words, it is a transfer learning scheme. It was used in e.g. Eriksen et al. [2020] with a DNN photo-z model and multi-source broad- and narrow-band photometry in the COSMOS field.

Apart from that, there are various ways to combine photo-z predictions obtained separately by SED fitting and ML algorithms, e.g.:

- Brodwin et al. [2006] used a SED fitting algorithm for the ordinary galaxies from Spitzer/IRAC Shallow observations and NN-based code for the AGNs.
- Carrasco Kind and Brunner [2014a] used Bayesian approach to combine PDF predictions made by several ML and SED fitting methods.
- Finally, for the large cosmological surveys, which require highly reliable photo-zs, it is a standard practice to provide predictions made by a number of various algorithms [e.g. Sánchez et al., 2014]; obviously, the galaxies for which all photo-z predictions are in agreement have the most reliable photo-zs.

2.7 Performance comparison

In the last 20 years a huge variety of photo-z methods have appeared. A question arises, which one should be applied to a new survey, or, in the case of surveys like SDSS where multiple photo-z catalogues are available, which results should be used for further studies.

For a number of reasons, it is hard to compare performance of various photo-z codes. Firstly, the performance of any algorithm hugely depends on quality of the data. Color cuts, redshift limits, limiting by photometric errors and selection or omission of certain types of objects (e.g. AGNs) affect the statistics dramatically. More so, the spec-z catalogues used for calculating performance also can be different (e.g. authors may use different quality flag filtering). For this reason, without an independent crossmatch of the datasets, it is often impossible to directly compare the performance reported in publications, even when the origin of the datasets is the same.

Secondly, the metrics have various relevance for different astronomical problems, e.g. for weak lensing it is important that the bias was low in every photometric bin, while for luminosity functions of rare types of galaxies keeping the percentage of catastrophic outliers low can be more useful. The authors usually concentrate their efforts on improving those statistical indicators that are most important for their field of interest. Besides, even though in the last decade a conventional set of performance metrics has been shaped (see § 2.3), it is not common that all these metrics are reported, and comparing e.g. $\sigma_{\Delta z}$ with $NMAD_{\Delta z}$ is, obviously, impossible.

Finally, the performance of a photo-z code depends on multiple factors. For example, shallow NNs, the most common ML photo-z code, show excellent performance on datasets with abundant KBs, but struggle with interpolating in the redshift ranges where the training data are sparse; Gaussian processes, on the other hand, handle such situations much better. k-NN can be more robust on the sparse datasets, such as quasars, while for a dataset with resolved galaxies CNN can produce more accurate predictions than any feature-based algorithm. In other words, there is no universal solution; for every scientific task and every large dataset a number of approaches have to be tested. It is worth noting that although it is considered a good practice to compare the performance of a new photo-z code to some commonly used implementations, such comparisons have to be taken with a grain of salt. Every algorithm has its quirks and weaknesses, and comparing a *carefully adjusted* newly presented algorithm to a, say, RF or NN solution with *default settings*, is not very informative.

The necessity of having independent evaluations has resulted in the appearance of the so-called 'blind test competitions'. The competitors - scientific groups working on photo-z codes - receive a KB that can be used for training the ML models or calibrating the SED fitting algorithms. Another part of the dataset remains unpublished. At the end of the competition it is used to compare the performance of the developed solutions.

First such competition was presented in Hogg et al. [1998]; the KB was not provided, so the teams were allowed to use any data they could derive from public databases, and the test set was composed of only 27 galaxies observed in the HDF. The competing codes were all but one SED fitting ones; the only empirical implementation used third-order polynomial fitting. This competition resulted in a qualitative observation, that all solutions perform more or less similarly.

Later a number of such competitions were organized, using e.g. GOODS [Hildebrandt et al., 2010], CANDELS [Dahlen et al., 2013], COSMOS [Norris et al., 2019], LSST [Schmidt et al., 2020] and Euclid [Euclid Collaboration et al., 2020] data. Besides, a number of non-blind comparisons were performed for e.g. HDF-North [Cohen et al., 2000], SDSS EDR [Csabai et al., 2003], SDSS DR6 [Abdalla et al., 2011], HSC-SSP DR1 [Tanaka et al., 2018] and DES SV [Bonnett et al., 2016, Sánchez et al., 2014]. Salvato et al. [2019] also compares the performances of different codes on different datasets, highlighting the importance of the input data, namely number of photometric bands and volume of the spec-z catalogue. Despite the ever-growing demands for the photo-z quality, the absolute majority of these investigations confirmed the conclusion made in Hogg et al. [1998]: in general, various photo-z codes perform similarly. At the same time, a few details were discovered:

- Most photo-z codes underestimate uncertainties of their predictions [e.g. Dahlen et al., 2013, Hildebrandt et al., 2008, Schmidt et al., 2020];
- High uncertainties may indicate catastrophic outliers, but small uncertainties have almost no correlation with actual photo-z errors [e.g. Hildebrandt et al., 2008];
- ML methods usually outperform SED fitting in the low-redshift regime, where the training data are abundant [Dahlen et al., 2013, Euclid Collaboration et al., 2020, Hildebrandt et al., 2010, Norris et al., 2019, e.g.];
- SED fitting, predictably, outperforms ML in the very low-redshift and high-redshift regime, where ML methods do not have enough data for training [Abdalla et al., 2011, Euclid Collaboration et al., 2020, e.g.].

Overall, it is evident that the best strategy is, whenever possible, to obtain photo-z with several algorithms, preferably both SED fitting and ML ones, so that they compensated each other's weaknesses, with ML being the preferable approach for the datasets with abundant KBs, and SED fitting - for the scarce high-redshift datasets.

2.8 Calibration of photo-z

The estimation of the photo-z's quality made for some spec-z sample is an optimistic one. The photometric sample is usually deeper, fainter and more diverse than the spectroscopic one, so photo-z for the overall catalogue are by definition worse than for the KB. For the weak lensing research, one of the major and most demanding in terms of photo-z accuracy consumer of photo-z, it is important to know *how much* worse, since in this type of analysis photo-z are the main source of uncertainty. For example, for the LSST Dark Energy Science Collaboration (DESC) program, the redshifts after ten years of observations have to be known with a systematic uncertainty of $\mu(\Delta z) \leq 0.001$ and scatter accuracy of $\sigma(\Delta z) \leq 0.03$, which has to have a systematic uncertainty ≤ 0.001 [The LSST Dark Energy Science Collaboration et al., 2018]; and these requirements must hold not simply for the overall sample, but for the galaxies within several redshift bins. The percentage of outliers has to be below 10%, and Schaan et al. [2020] showed that without additional correction (namely, marginalization over outliers), 5% uncertainty of the percentage of outliers (i.e., assuming that the fiducial percentage of outliers is 10%, it is 0.005% of the total size of the dataset) can outweight all other uncertainties in the dark energy analysis. Even with careful accounting for these effects, they can significantly lessen the reliability of the obtained cosmological constraints. In fact, such situation appeared a few years ago, when the tomographic cosmic shear measurements performed on the KiDS and DES data revealed a mild tension between the results. The scientific groups of both surveys performed a re-analysis of photo-z, using the same data, but switching to each other's methodologies; after that the tension was alleviated (see Wright et al. [2020] and references therein).

For these reasons, in the last ten years a lot of efforts are dedicated not only to the development of better photo-z codes, but to the auxiliary methods for calibrating their outcome as well [see e.g. Bernstein and Huterer, 2010, Bordoloi et al., 2010, Cunha et al., 2012, 2014, Davis et al., 2017, Hartley et al., 2020, Hearin et al., 2010, Lima et al., 2008, Ma and Bernstein, 2008, Masters et al., 2015, Matthews and Newman, 2010, Newman, 2008, Newman et al., 2015, Schaan et al., 2020, Wright et al., 2020]. As Schaan et al. [2020] put it, "like in any physical experiment, obtaining a quality measurement requires the instrument to be precise (training of the photo-z algorithm), and this precision to be known (calibration of the photo-z algorithm)". Usually this calibration is done using spec-z sample, but selecting this spec-z sample can pose a problem as well; below we will discuss it a little more.

It is worth noting that the term 'calibration' is not well-defined in photo-z literature and bears a little different meaning in the cosmological publications, where redshift distribution is the main objective, not individual redshifts itself. For simplicity and for the purposes of the next chapters, here I outline a definition that utilizes ML terminology, but is applicable for photo-z obtained with any method. Let's assume that the KB is the sample of galaxies for which both photometry and spec-z are known, and the run dataset is the sample of galaxies for which only photometry is present and which will be used in the further astrophysical and cosmological analysis. Then, *photo-z calibration* is a set of methods that ensure that the photo-z statistics obtained for the KB are applicable to the whole run dataset with a level of uncertainty that will not compromise the reliability of the further studies. In practice, any photo-z calibration is composed of two steps: a) ensuring that the photo-z statistics obtained for the KB are robust and reliable, and b) ensuring that the run dataset and the KB are the same in terms of parameter space, or accounting for their differences using statistical methods. The second step is often called 'calibration' by itself, in a sense that the run dataset is being calibrated to the KB.

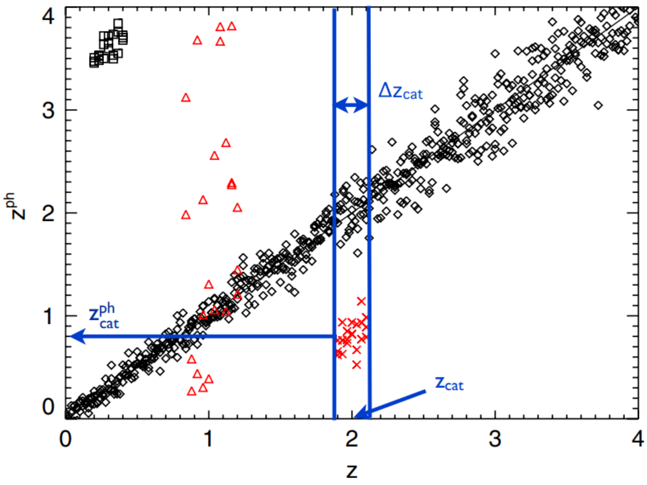


Figure 2.8: A schematic representation of types of photo-z outliers. Black diamonds correspond to the generally correct photo-z predictions. Black squares correspond to the localized catastrophic outliers caused by e.g. colour-redshift degeneracy that is observed on the spec-z sample and can be removed at the calibration stage. Red crosses appear to be similar to the black squares, but the nature of these outliers is different. They are localized catastrophic outliers that are underrepresented in the spec-z dataset; they cannot be removed from the photo-z catalogue and bias the following cosmological analysis. Finally, red triangles are uniform catastrophic outliers, which are the objects in a certain redshift range that are underrepresented in the spec-z train sample. Image taken from Hearin et al. [2010].

Before considering various calibration strategies, it is useful to look into the sources of photo-z errors first.

2.8.1 Types and sources of photo-z errors

Photo-z errors have various physical nature; they manifest differently and their contribution to the further analysis is also different. Using the terminology of Hearin et al. [2010], there are *localized* and *uniform* catastrophic outliers (see Fig. 2.8). The localized outliers are a population of objects which have a localized *true* redshift and localized *predicted* wrong photo-z; they cause systematic bias. The uniform outliers have 'random' photo-z residuals, and they mostly contribute to the scatter metrics.

The origins of photo-z errors are:

- **Intrinsic differences of the galaxies' SEDs.** Even in the ideal case, some baseline scatter of residuals is inevitable simply due to the intrinsic differences of the galaxies' SEDs and to the low wavelength resolution of photometry. The only way to lessen this scatter is by observing more galaxies of the types that have higher intrinsic differences and by obtaining photometry in medium- and narrow-bands.

- **Differences between spec-z and photometric samples in terms of colour/magnitude parameter space.** For the ML photo-z methods this factor is more obvious: ML algorithms by definition perform adequately only in the part of the parameter space which was occupied by the KB. SED fitting methods, in theory, can perform well even for the objects that are very dissimilar to those for which spec-z exist; however, it is not guaranteed, so for the precise cosmological measurements a good spec-z calibration sample is necessary.

  Failing to perform parameter space calibration results in higher scatter of predictions and in so-called *uniform catastrophic outliers* [Hearin et al., 2010], i.e. objects for which the predicted photo-z are incorrect and distributed randomly across the photo-z range, instead of being localized near some median value. There are two ways to handle this issue: to convey a follow-up spec-z survey, targeting the objects in the underrepresented parts of the parameter space, or to exclude these objects from the analysis. Both strategies have their pros and cons. Most often, the objects for which reliable spec-z are not available are the faint ones, for which obtaining spec-z may be impossible in principle. Rejecting all the objects from the underrepresented parts of the parameter space, on the other hand, significantly reduces the size of the galaxy sample and leads to the deterioration of reliability of the cosmological analysis. In a way, the demands of high completeness and accuracy of a spec-z survey confront each other; for a higher completeness, we have to observe fainter objects, but their accuracy is going to be lower [Cunha et al., 2014]. However, as Masters et al. [2015] have demonstrated, a careful planning of the follow-up observations can provide enough spec-z even for the faint parts of the colour space, with only a few percents of the photometric dataset left for rejection.

- **Colour-redshift degeneracy,** when a distant blue galaxy, being redshifted, becomes indistinguishable from a local intrinsically redder one. This outliers are usually *localized* ones [Hearin et al., 2010]; the degeneracy can be resolved with addition of the photometry in IR and UV bands that would help to better evaluate the shape of the SED [e.g. Benítez et al., 2009b], otherwise it can be taken into account using PDFs, where degenerate solutions appear as multimodality.

- **Blended sources**, which produce an apparent SED that is a sum of two different SEDs. The best way to handle this issue is to obtain photometry with an instrument with a higher spatial resolution, i.e. with from space [Schaan et al., 2020], however in principle photo-z codes should be able to assign low reliability to such objects.

- **Exotic sources, e.g. AGNs, quasars, SNe, Gamma-ray bursts**. These are a problematic class of objects for photo-z algorithms. Firstly, their SED is also essentially a combination of two components: one conditioned by the host galaxy and another appearing from the exotic or transient object itself. Secondly, their variability makes it complicated to obtain consistent photometry, especially when various bands are observed not at the same time [Salvato et al., 2019]. As a result, AGNs and other exotic objects often appear as catastrophic outliers when being treated together with the overall galaxy sample. To solve this issue it is necessary to use a suitable library of templates that accounts for the compound nature of these objects' SED or to train an ML model on the sample of these objects alone.

- **False outliers due to the contamination of the spec-z catalogue.** Some percentage of outliers (and some part of the scatter) is caused not by incorrect photo-z, but unreliable spec-z. Their origin is misidentification of spectral lines, unremoved atmospheric lines or spectroscopic analogue of color-redshift degeneracy, when in order to obtain spec-z a featureless spectra is being cross-correlated with a template spectra and Balmer and Lyman breaks are confused with each other [Cunha et al., 2014].

As we defined above, the process of photo-z calibration consists of two steps: making sure that the statistics obtained on the KB are reliable and robust, and calibrating the run dataset to the KB. The first step poses a number of requirements for the KB; spec-z catalogues have their own sources of inconsistencies, such as [Cunha et al., 2012]:

- Statistical shot noise, caused by too small size of the sample.
- Cosmic sample variance, caused by the fact that large-scale structure of the Universe introduces a coordinate-dependent distribution of the redshifts. On another patch of sky this distribution is going to be different, which affects photo-z priors, introduced to the SED fitting codes or learned from the dataset by the ML methods.
- Sample variance introduced by the observing conditions;
- Type incompleteness or imbalance, caused by the observational strategy or difficulties in observing certain types of objects, especially faint ones. Naturally, the objects which are the easiest for spec-z measurements - bright, with prominent spectroscopic features and well-defined SED shape - are also the objects for which photo-z are likely to have good accuracy. As a result, it is hard to determine the source of the statistics deterioration when photo-z code is being evaluated on the overall spec-z sample instead of the most reliable part of it. Some part of the deterioration is due to the photo-z failure, but some part is due to the 'false outliers' caused by the incorrect spec-z.

In § 3.4 I will apply a SOM-based data cleaning methodology to alleviate the contamination by some of these outliers.

The second step, the calibration of the datasets, has to ensure that the results obtained on the KB are applicable to the run dataset - and that all the unresolved uncertainties from the previous step are taken into account. To do this, several strategies exist.

2.8.2 Calibration strategies

The calibration of the run dataset aims to make its distribution in the parameter space the same as the distribution of the KB. The parameters, in the simplest case, are the colours, magnitudes and redshift; it is assumed that such calibration also ensures that the galaxy populations for the two datasets will be the same. The validity of this assumption depends on the number and coverage of photometric bands. In principle, nothing prevents using other parameters, e.g. morphological, for the calibration as well; it might be useful to break colour-redshift degeneracy in the deep surveys with absent IR and UV photometry.

There are three ways to do dataset calibration: culling the galaxies that are not represented in the KB, reweighting, or obtaining more spec-z for the underrepresented galaxies.

Culling in the simplest case is done via manual magnitude or colour cuts. It is often used at the stage of creation of the photo-z catalogues [see e.g. Cavuoti et al., 2015b, Eriksen et al., 2019, Wright et al., 2019], when it is not yet clear which sample of galaxies will be used in the further analysis. The culling criteria is more strict for the ML photo-z codes than for the SED fitting ones, since ML methods cannot make predictions outside of the KB's parameter space.

A finer way to do culling is to utilize an ML method to determine the boundaries of the KB's parameter space. In principle, many methods can be used for this, both supervised and unsupervised. In the case of a supervised ML the model has to be trained to determine whether an object belongs to the KB or the run dataset based on the photometry only. The objects from the run dataset for which this classifier fails are the objects that are well-represented by the KB, i.e. for which photo-z predictions can be trusted. Cunha et al. [2014] tested this approach with NN-based classifier. In the case of unsupervised ML, a clustering or dimensionality reduction algorithm has to be trained on the run dataset, and then post-labelled using the KB. This is the approach that I will use in Ch. 3.

Reweighting compensates the differences between the KB and run dataset by weighting the galaxies from the spec-z sample in such a way, that the resulting parameter distribution of the KB was similar to that of the run dataset. This approach was implemented for the first time by Lima et al. [2008] and tested on the mock DES and SDSS DR 6 datasets. The authors used nearest-neighbour algorithm in magnitude parameter space to establish the mapping between the spec-z and photometric samples and determine normalization weights. Later k-NN [Hildebrandt et al., 2017] and SOM-based [Buchs et al., 2019, Wright et al., 2020] implementations were created.

Newman [2008] proposed a completely different approach: to utilize spatial information instead of mapping the KB and the run dataset in the colour space. More precisely, the authors proposed to measure angular cross-correlation between the positions of the objects from the spectroscopic and photometric samples; taking into account that galaxies usually belong to groups and clusters, there is a high probability that the galaxies that are close to each other on the sky also have similar redshifts. Later this approach has been investigated in more detail by McQuinn and White [2013], and a publicly-available codes were developed [e.g. Morrison et al., 2017]. The advantage of these approaches is that they do not put much constraints on the spec-z sample; in principle, it can contain only selected type of objects for which spec-z are easy to obtain, e.g. luminous red galaxies.

The reweighting methods can be and often is used independently from the photo-z, since it allows to reconstruct the redshift distribution by itself - and this is exactly what is needed for many of the cosmological applications. Apart from the calibration, they can also be modified for obtaining priors for the Bayesian SED fitting photo-z codes.

At the same time, reweighting strategies have to be used with caution, and a careful propagation of the photo-z uncertainties through the subsequent analysis is needed. This is so because the influence of the redshift uncertainties on the cosmological models is redshift-dependent, so the same percentage of outliers in different redshift bins is harmful in various degree [Cunha et al., 2014, Hearin et al., 2010]. For example, let's consider two redshift bins, a low-redshift one and a medium-redshift one, with the same percentage of outliers, but with different density distributions of the KB and run datasets. The low-redshift bin has lesser number of galaxies in the run dataset, but almost all of them are observed spectroscopically. The normalizing weight is close to 1, and the statistics in general remain reliable. However, the medium-redshift bin has higher number of photometrically observed galaxies, and lower number of galaxies with reliable spec-z. As a result, spec-z sample is less representative for this bin than for the low-redhift bin, and the underrepresented objects are likely to have worse quality of spec-z. The reweighting algorithm has to take that into account. In principle, SOM-based solutions seem to be the most suitable tool for this task, thanks to the fact that their splitting of the parameter space into hypervolumes depends on the local density of the data points.

Obtaining more spec-z for the underrepresented galaxies is the most prolific method of dataset calibration, provided that the observational resources are available and the depth of the analysis is in principle within the reach of spectroscopy. Instead of removing some objects from consideration, which introduces selection biases to astrophysical research and reduces statistical reliability of cosmological studies, or propagating uncertainties originating from the incomplete spec-z sample and its difference from the photometric sample, this approach alleviates the uncertainties and preserves the variety of the galaxies. However, assuming a random sampling of the galaxies observed spectroscopically, the calibration of e.g. LSST would require a sample of ~ 100000 spectra, highly representative in terms of magnitudes and galaxy types, taken over several hundred square degrees, and for objects in the whole range of redshifts used for the analysis. Considering that the average success rate of a deep spectroscopic survey is $\sim 40 - 70\%$, that is essentially an impossible requirement [Newman et al., 2015].

In Masters et al. [2015] a much better approach was proposed, which is basically an inversion of the SOM-based reweighting algorithms. The idea is to train a SOM on the photometric dataset and then post-label it with already existing reliable spec-z. The cells which contain high number of galaxies from the run dataset but none or very few galaxies with reliable spec-z are the hypervolumes of the parameter space which require more information for calibration. The galaxies from these cells have to be targeted by the follow-up spec-z survey. In Masters et al. [2017] and Masters et al. [2019] this approach was used to organize spectroscopic observations of the calibration dataset in the COSMOS field, assuming Euclid-like photometric information available.

Similarly to the situation with photo-z codes, it is clear that combining multiple calibration methods, based on the task at hand, is the best possible strategy. It is also important to note that all the analysis mentioned above is cosmology-oriented, and mostly aim eliminating systematic biases in redshift distribution. The photo-z calibration for astrophysical purpose, which would be more concentrated on uniform catastrophic outliers and redshift distributions of specific types of galaxies, remains mostly unexamined.

2.9 Perspective directions

The next generation of large extragalactic surveys promises not only to constrain cosmological parameters with a precision of few percents, but also to provide us with an immense volume of information on galaxy parameters, variability and large-scale structure of the Universe. It means immense possibilities for astrophysics and galaxy evolution, and even imprecise photo-z distances will be instrumental to fully utilize them. Considering that there is no perfect photo-z or calibration code, but rather more and less effective combinations of data and methods, the volume and variability of the approaches is going to grow. However, there are seem to be two main trends.

The first trend is the appearance of combined pipelines, built using various ML, SED fitting and auxiliary approaches. Such pipelines commonly outperform standard methods, as it should be expected: ML allows to extract information directly from observations, SED fitting provides information known from previous observations, and various kinds of calibration remove biases, include spatial information and ensure the reliability of the metrics. It can be expected that with the advancement of the astronomical data infrastructure, even more complex algorithms will be created, e.g. such that will be automatically applying the most suitable code to the different populations of galaxies.

The second trend is to use not only broad band photometry, but other observables as well. The experiments on including morphological information into the analysis have been conveyed for at least twenty years, but in most cases they did not improve the results, apparently introducing more noise than redshift signal. However, this might change with a) DNN-based feature extractors, obtaining only relevant for the photo-z algorithms morphological parameters from the raw images directly, and b) creation of combined pipelines, that would use morphological parameters to narrow down the photo-z predictions and break colour-redshift degeneracy only for the objects from certain areas of the parameter space.

Unlike morphology, including medium and narrow bands have proved to be very useful, enabling the detection of bright emission lines in spectra of star-forming galaxies. At least, it is useful for the SED fitting algorithms; for now, ML-based codes do not show the similar degree of improvements. A possible explanation to this is that narrow bands contain a detectable redshift signal only for a small percentage of galaxies, so, perhaps, the existing ML photo-z codes are not adapted to learning this information. A further investigation of this matter is needed, e.g. in the direction of transfer learning with simulated datasets.

Other possible features that could improve the predictions are variability parameters (which will be provided abundantly by the LSST), SFRs, photometric errors, and imputed photometric data in the bands where non-detections are common. Obviously, in order to avoid the 'curse of dimensionality', preliminary (and possibly classification-aided) feature selection will be needed.

Chapter 3

COSMOS photo-z with Machine Learning methods

In this chapter, I obtain ML photo-z for the COSMOS2015 catalogue using shallow NN algorithm MLPQNA [Brescia et al., 2012, Cavuoti et al., 2012]. Although MLPQNA was used to obtain photo-z for multiple photometric catalogues before, this is the first time it is tested on a catalogue with narrow- and medium- band photometry. Additionally, here I test SOM-based data cleaning methodology for cleaning and calibrating source datasets. The aim of this step is to handle catastrophically wrong photo-z predictions of various nature (see § 2.8).

§ 3.1 describes the datasets and their basic preprocessing. § 3.2 gives information on the methods used. In § 3.3 I report the quality of the photo-z obtained with MLPQNA, investigate various feature sets and how the inclusion of the narrow bands affects the performance, and identify sources of photo-z errors. § 3.4 is dedicated to automatic detection of unreliable spec-z measurements, causing 'false outliers' (see § 2.8.1), and catastrophically wrong photo-z predictions. § 3.5 considers how we can use unsupervised ML to calibrate the parameter space of the run dataset to ensure the reliability of the photo-z quality estimations. In § 3.6 I describe how we can use our data-cleaning methodology to obtain reliable spec-z sample without quality flags. Finally, in § 3.7 I compare the quality of ML and SED fitting photo-z, discuss the nature of catastrophically wrong predictions, and consider how various data cleaning criteria can help to obtain spec-z and photo-z samples of various quality.

This chapter is based on the paper Razim et al. [2021].

3.1 Data

In this section I describe the three catalogues used in this work:

1. The COSMOS2015, which is our source of photometric measurements and SED fitting photo-z. Its detailed description can be found in Laigle et al. [2016].
2. A compilation of spectroscopic redshift available in literature for the same COSMOS field. This catalog is called the Main spec-z catalog.
3. The Deep Imaging Multi-Object Spectrograph (DEIMOS) spec-z catalogue [Hasinger et al., 2018], used to perform an additional, independent test of our methods.

3.1.1 COSMOS2015 photometric catalogue

The COSMOS2015 catalogue contains multi-wavelength broad-range (from mid-IR to near-UV) photometry for about half million objects, for which X-Ray and radio measurements, star formation rates and other additional information is available. Moreover, this catalogue includes SED fitting photo-zs, obtained with `LePHARE` software [Arnouts et al., 1999, Ilbert et al., 2006].

We downloaded the COSMOS2015 photometric data consisting of 34 bands:

- UV broad and medium bands: `NUVmag, FUVmag, u`;
- optical and near-IR broad bands: `B, V, ip, r, zp, zpp, Ks, Y, H, J, Hw, Ksw, yH`;
- mid-IR broad bands: `3_6mag, 4_5mag, 5_8mag, 8_0mag`;
- optical and near-IR medium bands: `IA484, IA527, IA624, IA679, IA738, IA767, IB427, IB464, IB505, IB574, IB709, IB827`;
- and optical and near-IR narrow bands: `NB711, NB816`.

For all of them $2''$, $3''$ *AUTO* and *ISO* apertures were available, except for Spitzer Large Area Survey with Hyper-Suprime-Cam (SPLASH; `3_6mag, 4_5mag, 5_8mag, 8_0mag`) and GALEX (`NUVmag, FUVmag`) bands.

The preprocessing of this dataset consisted of the following steps:

1. To ensure homogeneity of photometry and to enable a comparison with SED fitting photo-z, I follow the procedure described in Laigle et al. [2016]. I consider only those objects that lay within both UltraVISTA (this is done by using the condition `Area==0`) and COSMOS (`Cfl==1`) sky areas (see McCracken et al. [2012], Capak et al. [2007] and Scoville et al. [2007] for a detailed descriptions of these regions). As a result, $576,762$ objects remain out of the initially available $1,182,108$ samples.

2. I also exclude stars, X-ray and unclassified sources (`OType==0`). This leaves us with $551,538$ objects.

3. Finally, I remove saturated sources, by rejecting objects masked in optical broad bands (`Sat==0`). After this step, the final photometric catalogue consists of $518,404$ objects.

3.1.2 Main Spectroscopic catalogue

The Main spec-z catalog is extracted from the spectroscopic COSMOS master catalog maintained within the COSMOS collaboration. Our version of this catalogue includes only the publicly available redshifts prior Fall 2017. This catalogue contains 65 426 spectral redshifts, obtained with 27 different instruments in a spec-z range $0 < z_{\text{spec}} < 6.5$. The preprocessing consisted of the following steps:

1. To exclude stars, I remove sources with $z_{\text{spec}} < 0.01$; I also remove objects with $z_{\text{spec}} > 9$ to discard erroneous spec-z;

2. AGNs often pose a contamination problem for photo-z algorithms [Norris et al., 2019]. Therefore, I remove sources visible in X-Ray, using a catalogue of AGN sources detected by Chandra in the COSMOS field [Civano et al., 2012].

3. Then the resulting spec-z catalogue is cleaned from unreliable instances, using the available quality flags `Q_f`, described in Lilly et al. [2009]. Only robust spectroscopic redshifts are selected (i.e. with $\sim 99.6\%$ of spectroscopic verification), using the conditions `2<Q_f<5` and `22<Q_f<25`.

It is important to note that the main spec-z catalogue is a compilation of multiple catalogues. These data were obtained during different surveys with different targeting strategies and quality requirements during the last two decades, so the exact quality of the spectroscopic verification is impossible to estimate. As a result, the actual robustness of the final spec-z set may be lower, and this is one of the issues that I will address in the following sections. Another nuance is that for some objects the main spec-z catalogue

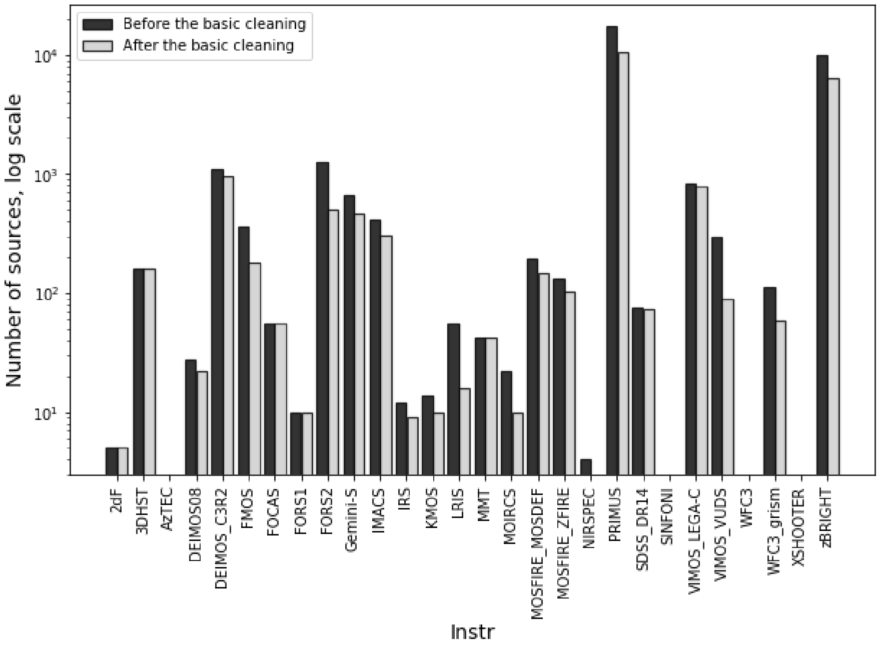


FIGURE 3.1: Number of objects in the main spec-z catalogue by instruments before and after the basic preprocessing steps. It can be seen that the preprocessing does not noticeably affect the instrument distribution.

contains multiple measurements made with different instruments, and in some cases these spec-z values have large residuals between each other (> 0.1). At this stage, I do not try to determine which measurements are correct, neither discard these objects. Instead, during the crossmatch, I simply use the spec-z measurements that are the best coordinate match to the COSMOS2015 objects. In § 3.3 I analyse these objects to clarify the nature of the photo-z outliers.

The resulting dataset is cross-matched with the COSMOS2015 photometric catalogue, obtaining $\sim 20\,000$ objects. The exact size of the dataset depends on the bands involved for limiting photometric errors, varying in the various experiments; see § 3.1.4. For the majority of this work, I used a dataset where I limit only photometric errors for broad and narrow visual, UV and IR bands in $3''$ aperture; it contains 19 893 objects. As shown in Fig. 3.1, the preprocessing does not noticeably affect the distribution of the sources by instrument of observation, and Fig. 3.2 shows that the spec-z distribution for $z_{\text{spec}} \leq 4$ is almost unaffected either. However, as it can be seen from this distribution, the number of objects for $z_{\text{spec}} > 1.2$ is approximately one order of magnitude lesser than the amount of closer objects. In absolute numbers, there are only ~ 700 galaxies in the redshift range $1.2 < z_{\text{spec}} < 4$. Such number of objects is not enough to effectively train the photo-z algorithm in this redshift range, and for this reason I limit further analysis and the resulting catalogue to $z_{\text{spec}} \leq 1.2$.

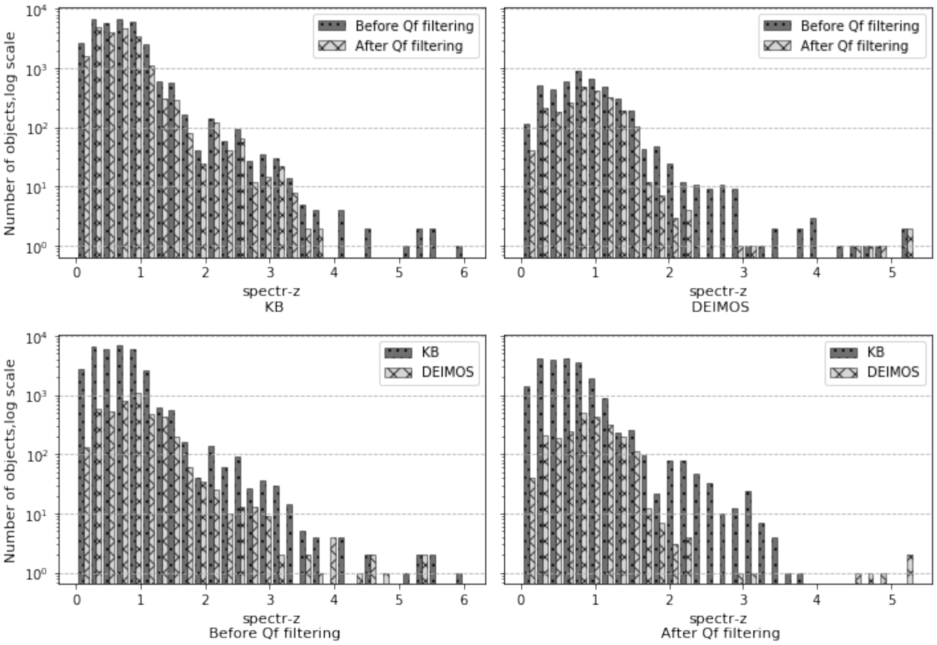


FIGURE 3.2: Spec-z distribution before and after quality flag (Qf) cleaning of the KB and DEIMOS datasets. The plots in the top row compare distributions dataset-wise, while the plots in the bottom row compare the KB and DEIMOS on the same cleaning stages.

3.1.3 DEIMOS spec-z catalogue

Since we want to test our methodology for selecting the subset of photometric data that is well covered by the KB, we need an independent spec-z catalogue different from the main spec-z catalogue selection function. For this purpose, I used the catalogue of spectroscopic redshift presented in Hasinger et al. [2018], acquired with DEIMOS, within different programs. This catalog provides redshifts for sources that are not included in the main spec-z catalog and that are somewhat fainter (see Fig. 3.3). For these differences, it represents an excellent benchmark data for this study.

In the preparation of the DEIMOS spec-z catalogue, I follow the same procedure described in § 3.1.2, aimed at discarding stars, AGNs and unreliable sources[1].

3.1.4 Final catalogues and SOM data cleaning prerequisites

After the basic preprocessing described in the previous sections, the following datasets are prepared (see Fig. 3.5 and Fig. 3.4 for the coordinate coverage):

[1] Note that the DEIMOS catalogue has two different quality flag columns. The one following the same scheme as in Lilly et al. [2009] is labeled "Qf".

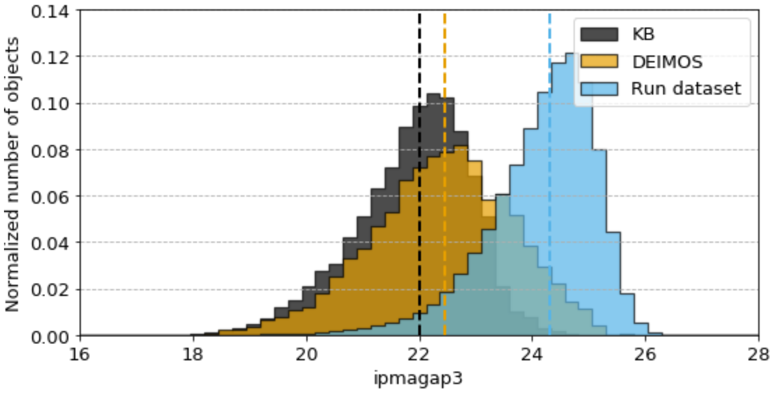


FIGURE 3.3: Magnitude distributions for the KB, DEIMOS and run datasets in the `ipmagap3` after standard cleaning but before any SOM filtering.

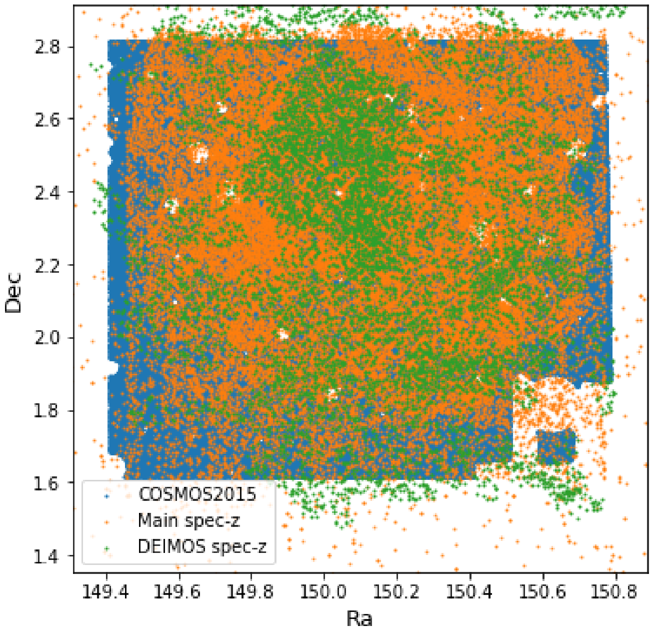


FIGURE 3.4: Coordinate coverage for the COSMOS2015, main spec-z and DEIMOS catalogues (all after the preprocessing).

1. KB (knowledge base), which is the intersection crossmatch between the COSMOS2015 and main spec-z catalogues. It contains both photometry and spec-z. For the ML photo-z experiments I randomly split this KB into train (70% of KB) and blind test (30%) datasets to provide reliable evaluation of the model performance. The photo-z algorithm also uses 10-fold cross validation during the training;

2. Run dataset, which is the COSMOS2015 catalogue after excluding the objects from the KB. It contains only photometric data;

3. DEIMOS dataset, which is an intersection of the run dataset and DEIMOS spec-z catalogue (meaning that it does not contain the objects from the KB). I use the DEIMOS as a control dataset to check how well our cleaning procedures work on

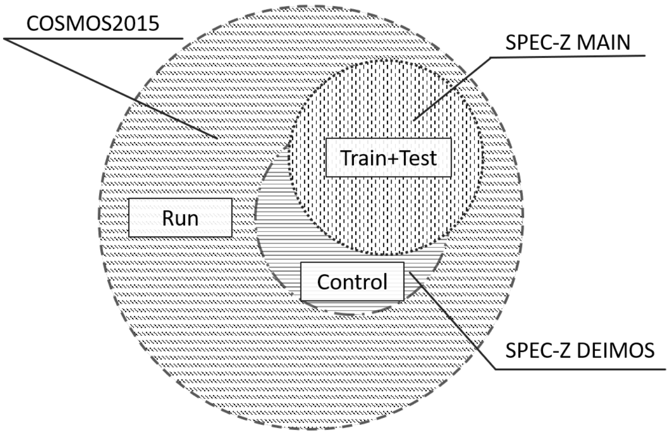


FIGURE 3.5: Venn diagram of the catalogues used in this work.

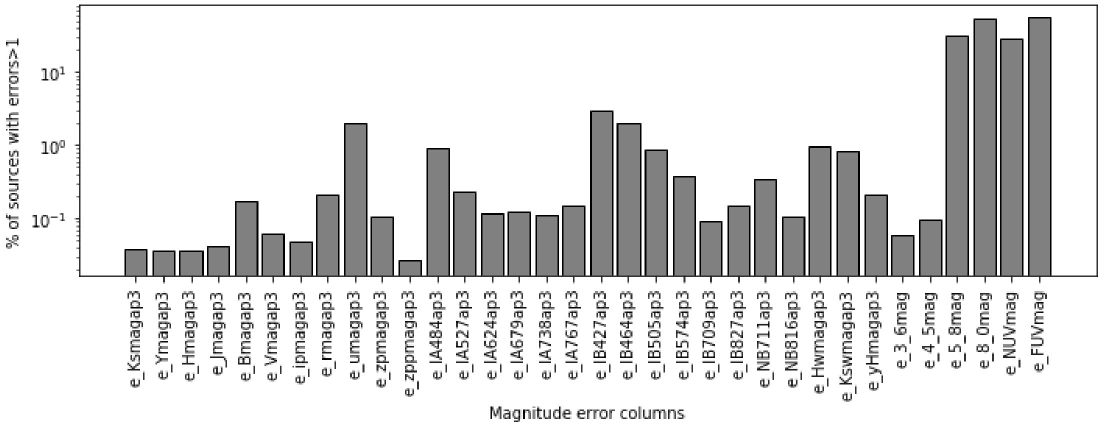


FIGURE 3.6: Number of objects in the KB with magnitude errors>1 by bands.

the data that come from an independent from the KB source and occupy a different hypervolume in the parameter space (see § 3.1.3).

In order to ensure the quality of the trained model, a reliable photometry is required, so I excluded all objects with high magnitude errors. Some bands have too many objects with unreliable photometry; if applied to these bands, this part of the preprocessing would reduce the size of the dataset by a factor of two or more, and narrow the area of parameter space where the photo-z algorithm would be applicable. To avoid this, I identified the bands affected by too many objects with large magnitude errors and excluded them from the experiments. As a rule of thumb this selection was operated by removing bands containing one order higher number of unreliable measurements than the others. As it can be seen from Fig. 3.6, these bands are `5_8mag, 8_0mag, NUVmag and FUVmag` and contain thousands of objects with large errors, while the other bands have up to hundreds of such galaxies.

Afterwards, I clean the photometry in the remaining bands, by limiting the magnitude error within the range $0 < e\_mag < 1$, where e_mag are magnitude errors for each band. The cleaning is performed for every feature set separately, since it allows us to preserve as many objects as possible (for example, in the experiments where I use only broad bands I do not remove the objects with high magnitude errors in narrow bands). Then in all ML experiments the software independently normalizes each band to the $[0, 1]$ range.

In order to take care of the mentioned low extrapolative power of ML models, we have to make sure that the run dataset is compliant with the KB in terms of parameter space coverage. Fig. 3.3 shows the magnitude distribution for the KB, DEIMOS and run datasets in `ipmagap3`. As it can be seen, our run dataset is noticeably deeper than the KB and DEIMOS datasets, as expected due to the spectroscopic bias induced by the targeting algorithms.

We then limit the magnitudes in the run dataset bands by their corresponding maximum values present within the KB. However, this is just a preliminary solution, since it does not guarantee a full similarity between the KB and run datasets, with respect to the parameter space. In § 3.2.5 I introduce a more accurate procedure to calibrate the run dataset to the parameter space of the KB.

3.2 Methods

In this work, I use two neural network algorithms. To calculate the photo-z point estimations, I use the supervised ML algorithm MLPQNA [Brescia et al., 2013, 2014]. Furthermore, in order to investigate and clean the datasets, I use an unsupervised ML approach, based on a modified version of the well-known dimensionality reduction algorithm SOM [Kohonen, 2013].

3.2.1 MLPQNA algorithm

MLPQNA is a neural network based on the classical Multi-layer Perceptron [Rosenblatt, 1963] with two hidden layers and hyperbolic tangent activation function of neurons. The general description of shallow neural networks was given in § 2.5.1, so here I outline only the distinctive traits of the MLPQNA.

MLPQNA uses Quasi Newton approximation of the Hessian error matrix as a learning rule [Nocedal, 2006]. MLPQNA was successfully used for photo-z estimation in a number of papers (such as, Biviano et al. [2013], Brescia et al. [2013, 2014], Cavuoti et al. [2012, 2015a, 2017a,b], Nicastro et al. [2018]). The model hyper-parameters were heuristically selected on the basis of our past experience and on an intensive test campaign. To

Parameter name	Parameter value
number of input neurons	N (number of features)
number of hidden layers	2
number of neurons in the hidden layer 1	$2N + 1$
number of neurons in the hidden layer 2	$N - 1$
number of output neurons	1 (for photo-z prediction)
weight decay (weight update multiplying coefficient)	0.001
restarts in every learning epoch	80
threshold (the difference between two consequent values of the learning metric after which the training stops)	0.01
epochs	20000

Table 3.1: MLPQNA model hyper-parameters settings (see Brescia et al. [2013] for details).

determine the number of neurons in each layer, I follow the rule of thumb described in Brescia et al. [2013]. This rule implies that the optimal number of neurons for the first hidden layer equals $2N + 1$ and the optimal number of neurons in the second hidden layer is $N - 1$, where N is number of features (i.e., photometric bands). The MLPQNA hyper-parameters are reported in Tab. 3.1.

3.2.2 Metrics

In order to evaluate the quality of photo-z estimations, I use standard metrics, described in § 2.3.

1. Standard deviation $\sigma(\Delta z)$;

2. Normalized median absolute deviation $\mathrm{NMAD}(\Delta z) = 1.48 \times median(|\Delta z|)$, which is less sensitive to catastrophic outliers than $\sigma(\Delta z)$;

3. Bias, defined as $mean(\Delta z)$;

4. Percentage of outliers $\eta_{0.15}$, defined as a number of sources with $|\Delta z| \geq 0.15$.

3.2.3 SOM

The Self Organizing Map (SOM) is a neural network algorithm first described in Kohonen [1982]. The idea behind SOM is the following: a parameter space of an arbitrary dimensionality is projected on a 2D map in such a way that neighbour instances in the original parameter space remain neighbours on the resulting map. For this purpose, the

SOM algorithm compares the feature vector of every object in the dataset (in our case the magnitude vector of every galaxy) with the weight vector of the same dimensionality associated with each cell on the 2D map. The object is then placed in the cell having the most similar weight vector (in terms of euclidean distance). Such cell is called Best-Matching Unit (BMU). The weights of the BMU and its neighbour cells are updated in such a way that they become more similar to the feature vector of the object. This mechanism ensures that, at the end of training loop, the map learns the representation of the parameter space of the entire training dataset in a self-organized way. That is why SOM is commonly used to perform an unsupervised data exploration. In photo-z estimation applications, several authors demonstrated that SOM can be used for different tasks. For example, Geach [2012] and Way and Klose [2012] used it to obtain photo-z, while Carrasco Kind and Brunner [2014b] applied it to estimate photo-z PDFs. Finally, Masters et al. [2015] adapted the SOM to check the coverage of the photometric parameter space by a given spectroscopic sample, thus indirectly creating a suitable method to optimize a spectroscopic follow-up strategy.

In this work I use the SOM for two different purposes: i) to detect potentially unreliable spec-z in the KB (see Sect. 3.2.4 and 3.4); ii) to ensure that the run dataset occupies the same part of the parameter space as the KB, i.e. to remove objects in the run dataset that are photometrically different from those in the KB (see Sect. 3.2.5 and 3.5). In our experiments I use a modified version of MiniSOM[2].

I use photometric bands as the input features for the SOM training. To analyse the datasets, I colour-label the resulting maps using either photo-z, spec-z, or the number of objects within each cell (also called the cell's occupation). The latter mapping is further referred as occupation maps, and it allows us to check how well the dataset, used for the labelling, covers the parameter space of the SOM training dataset.

Most of the SOM hyper-parameters (specifically, number of epochs, sigma, learning rate and neighbourhood function) are chosen via grid search. Their final values are listed in Tab. 3.2. The size of the map is chosen in such a way that, on average, each cell contains more than 30 galaxies from the training set. The choice of this size is based on the best compromise between the two competing goals: the reliability of statistics within each cell, and the need to capture the data topology with the maximum finesse. A SOM with a small number of cells provides us with a higher number of galaxies per cell, thus improving the reliability of the statistical indicators. On the other side, a larger SOM shows more details of the data distribution in the parameter space, but the statistics within some cells become unreliable. For this reason, SOM maps of different sizes are

[2] The original version of MiniSOM can be found in the repository `https://github.com/JustGlowing/minisom`. Our modified version is available at `https://github.com/ShrRa/minisom`.

Parameter name	Parameter value
width (low-resolution SOM)	25
height (low-resolution SOM)	28
width (high-resolution SOM)	67
height (high-resolution SOM)	64
num_features	10
epochs	6000
sigma	5
learning rate	0.5
neighborhood_function	bubble

TABLE 3.2: SOM settings used for all experiments in this study. For more details about low and high resolution maps, see § 3.2.3.

used. Specifically, I use small low-resolution maps to determine the anomalous sources within each cell, and large high-resolution maps to investigate the train dataset coverage of the parameter space defined by the run catalogue.

3.2.4 Spec-z in-cell outlier cleaning with SOM

The trained SOM places objects with similar feature vectors in the same or neighbour cells. If some object has a photometry-spec-z relationship that appears to be anomalous for its BMU, such object will be considered as an outlier in terms of spec-z distribution of this BMU. In order to avoid the potential confusion between these in-cell outliers with the traditional outliers of a photo-z distribution, hereafter I will refer the in-cell outliers as *anomalous sources* or *anomalies.*

We assume that these objects have a lower photo-z reliability, so I drop them out. In order to exclude such objects, for every galaxy the algorithm calculates the coefficient:

$$K_{\text{spec}} = \frac{\left\langle z_{\text{spec}}^{\text{BMU}} \right\rangle - z_{\text{spec}}^{\text{obj}}}{\sigma(z_{\text{spec}}^{\text{BMU}})} \tag{3.1}$$

where $\left\langle z_{\text{spec}}^{\text{BMU}} \right\rangle$ is the mean spec-z obtained by averaging over the objects falling in the same BMU, $z_{\text{spec}}^{\text{obj}}$ is the spec-z of the galaxy, and $\sigma(z_{\text{spec}}^{\text{BMU}})$ is the standard deviation of the spec-z distribution within the BMU cell. This coefficient has the same meaning as a standard score, or Z-score, often used in statistics. The prefix K instead of Z is chosen in order to avoid confusion with redshifts. Typically, objects are considered to be outliers if $|K_{spec}| > 3$. Yet, other criteria were also tried in order to see how it affects the photo-z quality (see § 3.4).

3.2.5 Parameter space calibration with SOM occupation map

As I pointed out earlier, neural networks cannot perform extrapolation. It means that, in order to obtain reliable results, the run dataset has to occupy the same area of the parameter space that is well sampled by the KB. Besides, as was explained in § 2.8, calibration of the datasets is needed even when SED fitting photo-z algorithms are used, otherwise the reliability of the photo-z performance is not guaranteed. Fig. 3.3 shows that, in terms of magnitude distribution, our train and run datasets are quite different. Therefore, in order to avoid biasing, we need to perform dataset calibration and estimate the statistics for each magnitude bin independently.

In the photo-z publications it is common to perform dataset calibration in the simplest form, by limiting the magnitudes of the run dataset to the maximum magnitudes of the KB, i.e. by cutting the faint-end "tail" of the magnitude distributions (e.g. Cavuoti et al. [2015b], Eriksen et al. [2019], Wright et al. [2019]). But this approach poses a major problem: if extreme cuts are adopted, the coverage of the parameter space is ensured but at the risk of loosing many objects with reliable photo-z. With a soft limiting value, on the other hand, the run dataset is affected by too many objects that lay outside of the KB parameter space.

Using SOM occupation maps, I implement a more accurate calibration, based on the approach first investigated by Masters et al. [2015]. The idea behind the method is simple: we train the SOM on the run dataset and then project the KB on the trained map. Since the entire SOM represents a projection of the parameter space of the run dataset, the galaxies in the KB will be clustered only in a subset of cells, with a fairly large number of cells either poorly occupied or completely empty. For these cells we do not have spectral information and hence photo-z predictions cannot be trusted. In order to capture in detail the coverage of the run dataset parameter space by the KB, I use a high-resolution SOM map of size $67X64$ (Tab. 3.2).

3.3 MLPQNA photo-z for COSMOS

3.3.1 Baseline photo-z experiments with MLPQNA

Here I calculate the photometric redshifts for the KB after the standard preprocessing described in §3.1.

As was explained in § 2.5.7, the choice of the feature set affects the results of ML process significantly, but selecting optimal features for the photo-z prediction task is not a trivial matter. Considering that in the case of COSMOS2015 we have four magnitudes in more than 30 bands, it is computationally impossible to convey MLPQNA experiments in the parameter space of the full dimensionality, and, consequently, some feature selection is needed. Performing a full brute force search is also too computationally expensive. For this reason, I tried manual feature selection based on domain knowledge, some brute force search on small subsamples of the total feature set, and several computationally inexpensive automated feature selection algorithms. As a result of this search, I performed ~ 400 experiments with different combinations of features, namely: with all photometric bands, only broad bands, broad bands plus one or more narrow bands, with five "SDSS-like" bands, etc. I also investigated whether there is any difference in photo-z performance when we use photometry in the same bands, but obtained with different instruments. To do this, I used same baseline broad-band feature sets, where one of the bands was replaced with the same band from a different source (i.e. VIRCAM/VISTA `Ks` and `H` bands were being replaced with WIRCAM/CFHT `Ksw` and `Hw`)

Tab. 3.3 reports the results of several basic experiments that I conveyed to obtain baseline results and narrow down the further search. The first five experiments use the same basic SDSS-like (based on the similarity of the central wavelengths of the bands) set of five broad bands, one in UV and four in the visual parts of the range, but different apertures. The experiments show comparable results, although $2''$ and $3''$ apertures demonstrate slightly lower bias and percentage of outliers than pseudo-total Kron magnitudes in `aper_auto` and isophotal magnitudes in `aper_ISO`. This is consistent with the results obtained for the SED fitting photo-z, reported in Laigle et al. [2016]. In the same article it is suggested that the lower quality of photo-z with `aper_auto` and `aper_ISO` can be explained in two ways: by the contamination by blended sources, or by the fact that these apertures are determined based on the images taken in the detection band, so it is possible that for the faint sources, these magnitudes in other bands are going to be noisier than fixed-aperture magnitudes.

The experiments with $2''$ and $3''$ apertures show very similar performance, and using them together does not bring noticeable improvement. For consistency of the comparison of the ML and SED fitting photo-z, in the further experiments I used magnitudes in $3''$ aperture, unless specified otherwise.

The next two rows in Tab. 3.3 (`exp006` and `exp007`) are excerpts from the series of experiments dedicated to the investigation of how adding broad UV and IR and narrow bands affects photo-z performance. In more details these experiments are reported in § 3.3.2 and § 3.3.3, and here they are presented for general comparison.

The `exp006` in Tab. 3.3 reports the results of the experiment with the feature set composed of all broad bands available, i.e. five visual[3], one UV and four IR. Compared to the results obtained in experiments with the SDSS-like feature sets, this experiment shows noticeable improvement of NMAD: from $\sim 0.025 - 0.028$ to 0.018. However, the improvement of $\sigma_{\Delta z}$ is weak, and the percentage of outliers does not drop at all. In § 3.3.6 I will investigate this discrepancy and show that the major source of outliers in our dataset is the contamination of the spec-z sample; the important consequence of this is that without additional data cleaning, percentage of outliers and $\sigma_{\Delta z}$ are less reliable metrics than NMAD. For this reason, in the next paragraphs, dedicated to the various aspects of feature selection, I will mostly concentrate on NMAD.

The `exp007` in Tab. 3.3 is an example of an experiment with a feature set composed of all broad and one narrow band. Unlike adding IR bands, this expansion does not bring a serious improvement; the same is fair for the `exp008`, where all available bands, both broad and narrow ones, were used. This absence of improvements contradicts the findings for the SED fitting algorithms, which were shown to benefit from the addition of the narrow bands [e.g. Salvato et al., 2019, and references therein]. I will return to this in § 3.3.3.

In order to make sure that we do not run into the 'curse of dimensionality', I performed a series of experiments to test how the quality of photo-z predictions depends on the size of the train sample. They were done with the setup of `exp006`, i.e. with 10 broad bands. Fig. 3.7 shows the results of these experiments. It demonstrates that after the training sample reaches $\sim 8\,000 - 10\,000$ objects, the improvement of statistics essentially stops.

3.3.2 Visual, UV and IR bands

The strongest redshift signal is encoded in the gradient between the magnitudes bracketing either Balmer or, for high-redshift galaxies, Lyman break. Being redshifted, these breaks change their wavelength, and, consequently, fall within different bands, meaning that the significance of each band for the photo-z calculation is redshift-dependent. We check how different combinations of the bands belonging to the visual, IR and UV parts of the spectra affect the quality of the ML photo-z. Tab. 3.4 reports the results of these experiments, and Fig. 3.8 demonstrates the residual-redshift distribution for each feature set.

[3] We included the NIR `zpp` band in the 'visual' feature set for two reasons: firstly, to have another 5-band feature set for a fair comparison with the SDSS-like, and secondly, to perform an experiment with the bands all coming from one instrument (in this case, Suprime-Cam/Subaru) to see whether there will be any degradation when we add photometry from other instruments.

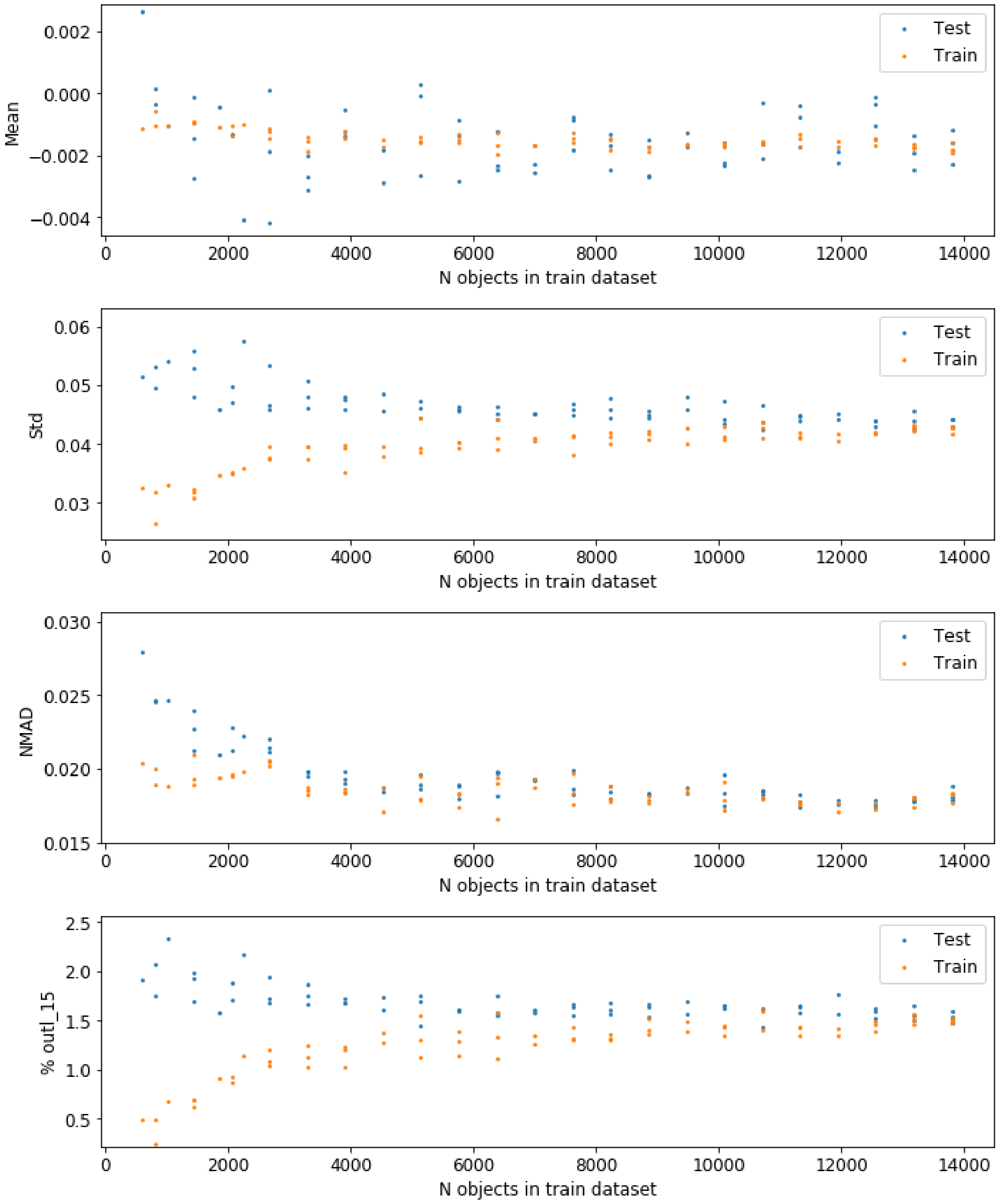


FIGURE 3.7: Photo-z quality dependency from the size of the train dataset. Three experiments were performed for each number of objects in the training sample. The setup of `exp006` from Tab. 3.3 was used.

Exp ID	Description	bands	Mean	$\sigma_{\Delta z}$	NMAD	$\eta_{0.15}$
exp001	SDSS-like, aper2	u, B, r, ip, zpp	0.000	0.053	0.025	1.999
exp002	SDSS-like, aper3	u, B, r, ip, zpp	-0.001	0.051	0.027	1.807
exp003	SDSS-like, aper_auto	u, B, r, ip, zpp	-0.003	0.055	0.027	2.364
exp004	SDSS-like, aper_ISO	u, B, r, ip, zpp	-0.004	0.059	0.028	2.346
exp005	broad bands, aper2 and aper3	u, B, r, ip, zpp	-0.001	0.051	0.025	1.756
exp006	broad bands, aper3	B, H, J, Ks, V, Y, ip, r, u, zpp	-0.002	0.048	0.018	2.138
exp007	broad bands + one narrow band, aper3	B, H, J, Ks, V, Y, ip, r, u, zpp, IB574	-0.002	0.048	0.019	1.642
exp008	all bands, aper3	Ks, Y, H, J, B, V, ip, r, u, zp, zpp, IA484, IA527, IA624, IA679, IA738, IA767, IB427, IB464, IB505, IB574, IB709, IB827, NB711, NB816, Hw, Ksw, yH	-0.002	0.049	0.017	1.877

Table 3.3: Results of the baseline experiments performed with the trained MLPQNA on the blind test set of the KB, after having just applied the standard KB cleaning procedure.

The first thing to notice is that the baseline experiment `exp002`, performed with SDSS-like features, has higher NMAD than all the experiments that include visual bands. This is easy to explain by the fact that our SDSS-like band set is chosen based on the central wavelengths of the bands only, without any consideration of their bandwidth. In the case of COSMOS2015, the adjacent `B` and `r` band do not overlap, and the gap between them degrades the quality of the photo-z predictions for galaxies in the range $\sim 0.4 < z_{spec} < 0.9$, which inevitably affects the overall statistics. This effect is enhanced by the fact that the majority of the galaxies lie in that range, so the worsening of NMAD for them has higher weight than better NMAD for the galaxies in the range $z_{spec} < 4$, which is also observed with SDSS-like features due to the presence of the `u` band. In any case, adding the `V` band improves the statistics in the middle of redshift distribution.

Exp ID	Description	N bands	Mean	$\sigma_{\Delta z}$	NMAD	$\eta_{0.15}$
exp002	Baseline SDSS-like (`u, B, r, ip, zpp`)	5	-0.001	0.051	0.027	1.807
r_exp001	Vis (`B, V, r, ip, zpp`)	5	-0.003	0.055	0.024	2.242
r_exp003	IR (`Y, J, H, Ks`)	4	-0.011	0.109	0.066	13.489
r_exp004	Vis, UV	6	-0.003	0.052	0.020	1.929
r_exp005	Vis, IR	9	-0.003	0.053	0.020	2.260
r_exp006	Vis, UV, IR	10	-0.002	0.052	0.018	2.068
r_exp007	Vis, UV, `Y`	7	-0.002	0.049	0.019	2.103
r_exp008	Vis, UV, `J`	7	-0.004	0.053	0.018	2.086
r_exp009	Vis, UV, `H`	7	-0.003	0.047	0.020	1.756
r_exp010	Vis, UV, `K`	7	-0.002	0.050	0.020	1.964

Table 3.4: The results with visual, UV and IR bands, obtained on the test set of the KB.

The IR bands alone (`r_exp003`) are predictably not enough to make good photo-z predictions, but adding them to visual bands (`r_exp005`) reduces both NMAD and mean of residuals. Adding UV band (`r_exp004`) has the same effect. Curiously, in both cases the improvement mostly comes from the galaxies in range $z_{spec} < 0.3$ (see Tab. 3.5). For the visual+UV feature set it was to be expected, since for the galaxies with $z_{spec} < 0.3$ Balmer break moves from `u` to `B` band. The improvement introduced by the IR bands is harder to explain; however, it is beneficial to add IR bands to the feature set even if `u` band is already included (`r_exp006`). A more careful analysis with adding each IR band to the feature set separately (experiments `r_exp007-r_exp010`) shows that the main improvement in terms of NMAD is introduced by the `J` band; at the same time, using only this band causes a slight increase of the mean bias.

It seems likely that IR bands in this case allow the MLPQNA to extract information contained in the overall shape of the SED, rather than detect some specific SED feature. Considering that the feature set composed of all broad bands produces the best results, in the majority of the further experiments, I use the feature set composed from all broad visual, IR and UV bands, unless specified otherwise.

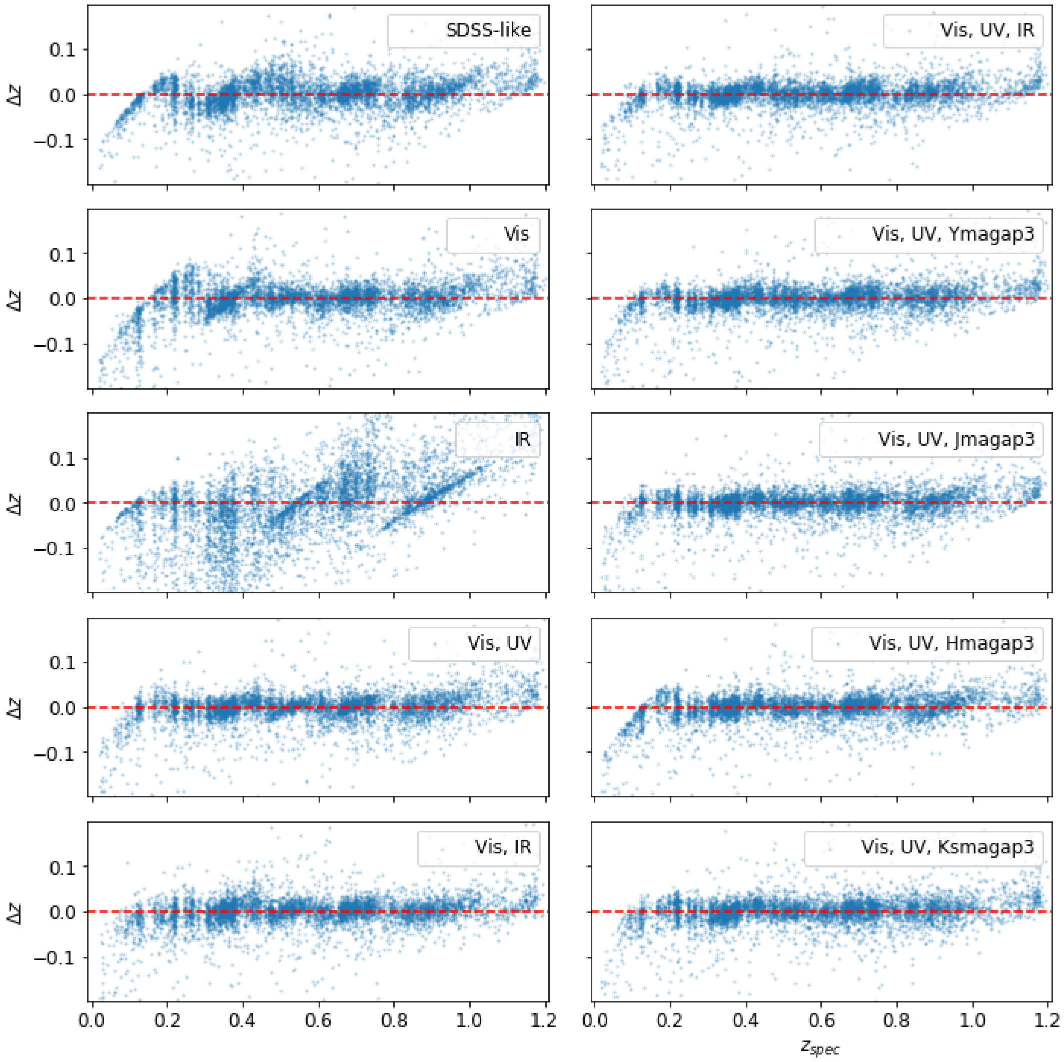


FIGURE 3.8: Photo-z quality with various feature sets

3.3.3 Narrow band influence

As was shown by e.g. Ilbert et al. [2009], SED fitting photo-z methods demonstrate better performance when SED templates take into account prominent emission lines, especially when narrow-band photometry is present. The improvement is most significant for star-forming galaxies and AGNs [e.g. Salvato et al., 2009]. However, it was never exhaustively investigated how adding narrow bands to the feature set affects the performance of the ML photo-z methods. Leaving such investigation for the future, I did try to perform a percursory check.

Tab. 3.6 reports the results of one of the series of these experiments. In there, I used a basic feature set composed of broad visual, UV and IR bands, complemented with all narrow bands one by one. In comparison with the baseline experiment (`exp006`), no experiment shows robust improvements for neither of metrics. Using all narrow bands in

z_{spec}	Vis	Vis, UV	Vis, IR	Vis-(Vis, UV)	Vis-(Vis,IR)	(Vis,UV)-(Vis,IR)
0.1	0.062	0.058	0.062	0.004	0.000	-0.004
0.2	0.045	0.023	0.024	0.022	0.021	-0.001
0.3	0.038	0.020	0.027	0.018	0.011	-0.007
0.4	0.019	0.016	0.018	0.003	0.001	-0.002
0.5	0.026	0.020	0.024	0.006	0.002	-0.004
0.6	0.015	0.013	0.017	0.002	-0.002	-0.004
0.7	0.016	0.016	0.016	0.000	0.000	0.000
0.8	0.019	0.019	0.016	0.000	0.003	0.003
0.9	0.023	0.021	0.018	0.002	0.005	0.003
1.0	0.023	0.022	0.018	0.001	0.005	0.004
1.1	0.030	0.035	0.026	-0.005	0.004	0.009
1.2	0.031	0.034	0.025	-0.003	0.006	0.009

Table 3.5: NMAD in z_{spec} bins for different feature sets. The first column gives the upper boundary of the redshift bin; the next three columns report the NMADs for the feature sets composed of broad visual, visual+UV and visual+IR bands correspondingly. The next two columns indicate the difference between the NMADs obtained with different feature sets; essentially, they show the improvement of NMAD that comes with adding UV or IR bands. This difference is largest for the two redshift bins in range $0.1 < z_{spec} < 0.3$. The final column reports the difference between the NMADs obtained with the features sets that include UV and IR. Adding UV band brings some improvement for $z_{spec} < 0.6$, adding IR bands improves redshifts for $z_{spec} > 0.6$. Such clear separation by the median of the redshift distribution suggests that it can be introduced by the ML model rather than some physical effect. In any case, the performance difference between these two feature sets is fairly small for all redshift bins.

addition to all broad bands lowers NMAD (see `exp008` in Tab. 3.3), but only just: the improvement is from 0.018 to 0.017.

There are several causes that could prevent the ML model from learning the information contained in the narrow bands, starting from issues with photometry homogenization and finishing with unsuitable ML model architecture. In any case, SED template fitting example shows that this information is possible to utilize, and finding a way to do it should be subject of further work.

Exp ID	Band	Mean	$\sigma_{\Delta z}$	NMAD	$\eta_{0.15}$
exp006	Baseline	-0.002	0.048	0.018	2.138
n_exp031	IA484	-0.002	0.051	0.018	2.086
n_exp032	IA527	-0.003	0.053	0.019	1.929
n_exp033	IA624	-0.003	0.050	0.019	1.721
n_exp034	IA679	-0.001	0.050	0.020	1.790
n_exp035	IA738	-0.002	0.048	0.019	1.964
n_exp036	IA767	-0.003	0.051	0.019	1.947
n_exp037	IB427	-0.001	0.046	0.020	1.843
n_exp038	IB464	-0.003	0.053	0.019	1.912
n_exp039	IB505	-0.003	0.053	0.018	1.860
n_exp040	IB574	-0.002	0.048	0.019	1.895
n_exp041	IB709	-0.001	0.050	0.019	2.068
n_exp042	IB827	-0.002	0.051	0.019	2.155
n_exp043	NB711	-0.002	0.051	0.019	1.982
n_exp044	NB816	-0.002	0.051	0.019	1.947

Table 3.6: The photo-z statistics obtained for the feature sets that include visual, UV and IR broad bands, complemented with an indicated narrow band.

3.3.4 Colours vs. magnitudes

The major part of the redshift information is contained in the slope of the SED rather than in magnitude values themselves. For this reason, photo-z codes commonly use colours instead or together with magnitudes. For ML photo-z algorithms, it is an open question whether using colours is any better or worse than using magnitudes. From one point of view, in ML tasks, suitable feature engineering simplifies the algorithms' job and sometimes allows to significantly improve the performance. From another point of view, ML algorithm should be able to determine the utility of such simple 'composite feature' as the magnitude differences by itself. At the same time, other magnitude relations, e.g. ratios or colours obtained between non-adjacent bands, can hold some redshift

Exp ID	Description	N features	Mean	$\sigma_{\Delta z}$	NMAD	$\eta_{0.15}$
c_exp023	colours: Vis	5	-0.002	0.054	0.021	2.155
c_exp024	colours: Vis, UV	6	-0.001	0.051	0.020	2.034
c_exp025	colours: Vis, IR	8	-0.004	0.052	0.020	2.051
c_exp026	colours: Vis, UV, IR	9	-0.001	0.051	0.020	2.016

Table 3.7: Photo-z performance with feature sets based on colours obtained for visual, UV and IR broad bands.

information as well. By using only simple adjacent-band colours, we might prevent the model from learning more complex dependencies.

We tested the performance of MLPQNA using colours instead of magnitudes for the four main cases of wavelength coverage, e.g. for the visual part of the spectrum, visual+UV, visual+IR and visual+UV+IR (see Tab. 3.7). The results are essentially the same as when magnitudes were used; however, in the case of 'full coverage' feature set (visual+UV+IR), magnitude-based feature set brings slightly lower NMAD than colours-based (0.018 for magnitudes-based, 0.020 for colours-based).

3.3.5 Automatic feature selection

Apart from testing physically motivated feature sets, I performed some experiments with automated feature selection algorithms. There are plethora of such algorithms, which are often split in three categories:

- Filter methods. These methods are the most general and computationally inexpensive; essentially, they measure the correlation between each of the feature and the target value, and the features with the highest correlation coefficient are considered most useful. I used the simplest Pearson's correlation matrix.
- Wrapper methods. These approaches iteratively add or remove features to/from feature sets, using performance of some ML algorithm as a criterion for whether the feature is important. To obtain best results possible, the feature relevance metric should be based on the same ML method that will be used in the the end for the target prediction. However, often it is impossible because the ML model is computationally expensive and training it at every iteration of the feature selection process would take too long. For example, this is true for MLPQNA. For this reason, I used a basic Random Forest (RF) regression model with default

sklearn settings (100 estimators with no limit for the maximum depth of the tree), combining it with the Recursive Feature Elimination (RFE) algorithm.

- Embedded methods. These methods are also iterative; at every iteration they evaluate the relevance of the features and remove those that are less useful than a predefined threshold. I used Parameter handling investigation LABoratory (ΦLAB) method, presented in Delli Veneri et al. [2019] and Brescia et al. [2019]. This method combines embedded and filter approaches; it uses so called *shadow features*, the noisy copies of the real ones [Kursa and Rudnicki, 2010], to determine feature importance threshold, and *Naive LASSO* (Least Absolute Shrinkage, Tibshirani [2012]) regularization to remove least relevant features.

The selection with Pearson's correlation matrix and RFE was performed using Python sklearn implementations and only on magnitudes and colours taken from the adjacent broad band magnitudes. Feature selection with ΦLAB also investigated all possible differences and ratios between all magnitudes, since their production is part of our implementation of the code. In every case, ten most relevant features were selected, in order to provide a fair comparison with the baseline feature set composed of ten broad bands.

Tab. 3.8 reports the results of the photo-z experiments with the feature sets derived by these automatic methods.

In comparison with the baseline feature set composed of all ten broad bands, the feature set selected with correlation matrix performs the worst. From the physical point of view it is not surprising; the selected features are mostly narrow bands, which might correlate with spec-z, but are not sufficient to detect Balmer break. Looking at the correlation matrices themselves (Fig. 3.9 and Fig. 3.10), we can also get some insight on why correlation-based feature selection methods are in general not the best choice for the photo-z task. Traditionally, apart from checking how various features correlate with the target value, it is useful to check how they correlate with each other and select such a feature set that its inner correlation was minimal. This additional step helps to reduce redundancy and minimize the dimensionality of the parameter space. However, in the case of photo-z, photometric bands (especially adjacent ones) are bound to correlate with each other, due to the fact that they are the points on a smooth and mostly monotone galaxy SED. Consequently, removing one of the features from the highly correlated pair would break the most useful feature pairs. For example, the correlation matrices for the broad and narrow bands (Fig. 3.9) show that these feature sets in general have more 'internal correlations' between their features than colour-based (Fig. 3.10), yet in the experiments with MLPQNA using colours does not bring any better results than

Photo-z method	Feature selection	Bands	Mean	$\sigma_{\Delta z}$	NMAD	$\eta_{0.15}$
MLPQNA	Baseline (exp006)	B, H, J, Ks, V, Y, ip, r, u, zpp	-0.002	0.048	0.018	2.138
MLPQNA	CorrMatrix	IB574, V, IA527, IA624, IB505, r, IA484, IA679, r-ip, IB464	-0.004	0.072	0.028	4.050
RF	Baseline (exp006)	B, H, J, Ks, V, Y, ip, r, u, zpp	-0.004	0.059	0.026	2.710
RF	RF-RFE	IB574, u-B, B-V, V-r, r-ip, ip-zpp, zpp-Y, Y-J, J-H, H-Ks	-0.002	0.055	0.017	2.360
MLPQNA	RF-RFE	IB574, u-B, B-V, V-r, r-ip, ip-zpp, zpp-Y, Y-J, J-H, H-Ks	-0.002	0.051	0.023	2.103
MLPQNA	ΦLab	u/B, r-ip, B/V, u-B, IA527/IA624, H-Ks, IA679-IA738, V-r, IB464/IB505, V/r	-0.003	0.051	0.024	1.959

Table 3.8: Photo-z quality for feature sets obtained with automatic feature selection.

using magnitudes. Similarly, all IR bands show high internal correlation, as well as high correlation with the IR zpp magnitude which I arbitrary included in the 'visual' feature set; still, adding the four IR bands J, H, Y, Ks improves the photo-z quality in comparison with visual-only feature set, and this improvement is higher than if we add only one IR band.

To test the next feature selection method, I first used RF in basic configuration with the baseline feature set (Tab. 3.8, third row), and then re-trained the same model on the RFE-selected features (Tab. 3.8, fourth row). With the baseline feature set, RF performed noticeably worse than MLPQNA (RF produced photo-z with NMAD = 0.026, NMAD of MLPQNA was 0.018). The RFE-selected features lowered the NMAD obtained with RF to 0.017. With MLPQNA, however, this feature set performed worse than baseline, producing NMAD = 0.023. Curiously, the RFE-selected features are mostly colours with an addition of one narrow band. The different behavior of RF and MLPQNA with the

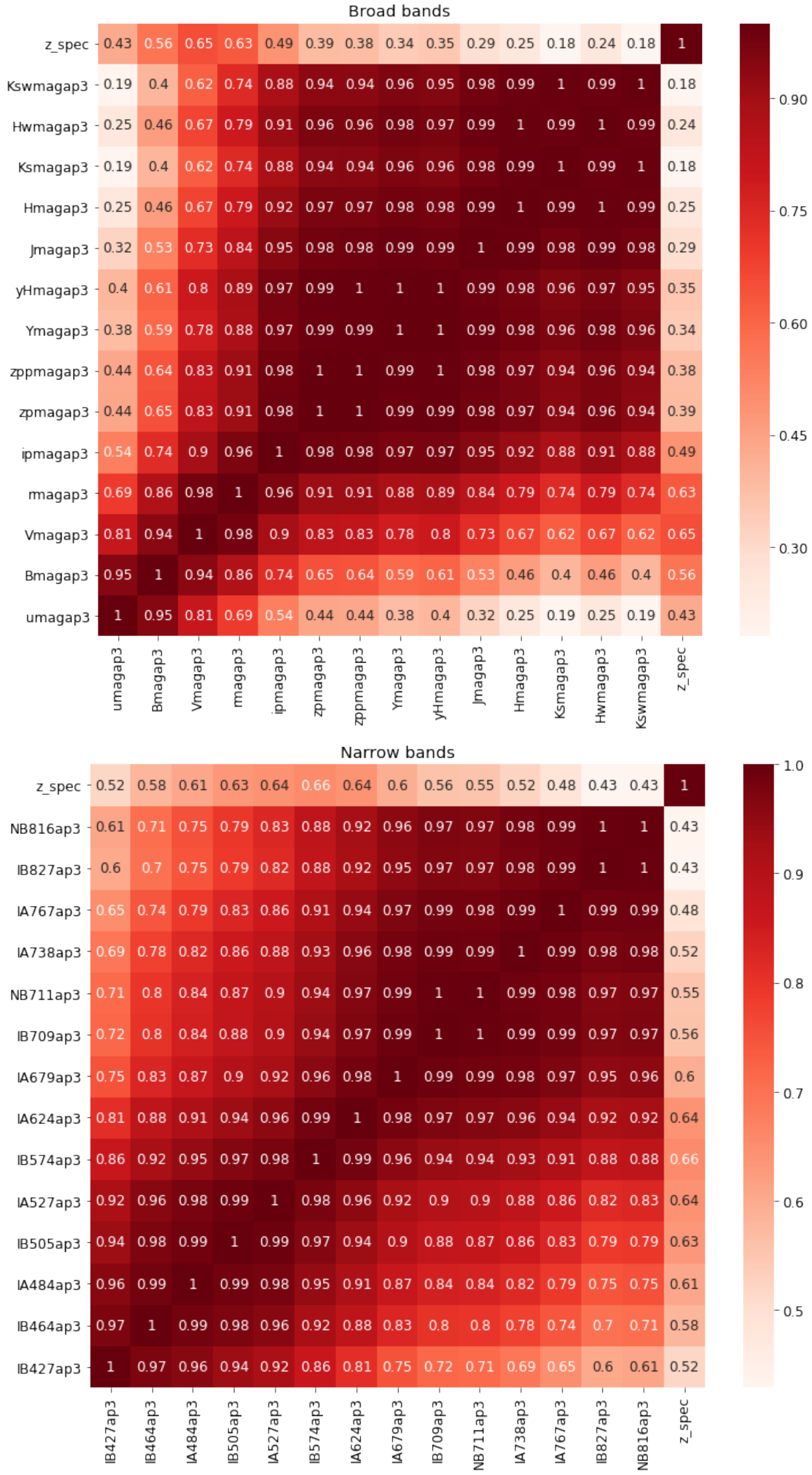
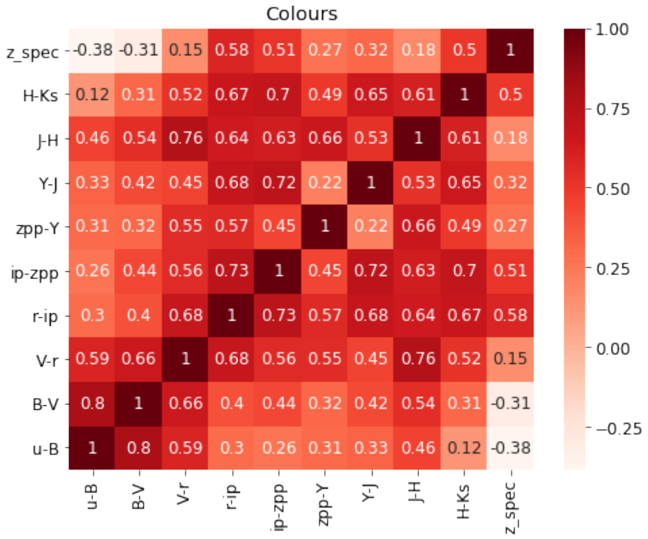


Figure 3.9: Feature correlation matrices for broad and narrow bands. During feature selection process all correlation matrices were analyzed together; the split for three matrices is done only for the sake of more clear visualization.



	u-B	B-V	V-r	r-ip	ip-zpp	zpp-Y	Y-J	J-H	H-Ks	z_spec
z_spec	-0.38	-0.31	0.15	0.58	0.51	0.27	0.32	0.18	0.5	1
H-Ks	0.12	0.31	0.52	0.67	0.7	0.49	0.65	0.61	1	0.5
J-H	0.46	0.54	0.76	0.64	0.63	0.66	0.53	1	0.61	0.18
Y-J	0.33	0.42	0.45	0.68	0.72	0.22	1	0.53	0.65	0.32
zpp-Y	0.31	0.32	0.55	0.57	0.45	1	0.22	0.66	0.49	0.27
ip-zpp	0.26	0.44	0.56	0.73	1	0.45	0.72	0.63	0.7	0.51
r-ip	0.3	0.4	0.68	1	0.73	0.57	0.68	0.64	0.67	0.58
V-r	0.59	0.66	1	0.68	0.56	0.55	0.45	0.76	0.52	0.15
B-V	0.8	1	0.66	0.4	0.44	0.32	0.42	0.54	0.31	-0.31
u-B	1	0.8	0.59	0.3	0.26	0.31	0.33	0.46	0.12	-0.38

FIGURE 3.10: Feature correlation matrix for colours. Unlike two previous correlation matrices, this one has two features with anti-correlation with spec-z: `u-B` and `B-V`; however, for our purposes only their absolute values are important.

two feature sets, magnitudes-based and colour-based, demonstrates the importance and model-dependence of correct feature selection.

Finally, the MLPQNA was trained with the features selected by ΦLAB. The most relevant features selected by this algorithm are all ratios and differences between adjacent broad bands or between narrow bands withstanding from each other by one narrow band. Similarly to the other feature selection methods, the results with this feature set is worse than with the ordinary baseline magnitudes.

Despite the generally negative results of the feature selection experiments, it brings some useful insights that might help with further, more elaborate study in this direction. First of all, it is clear that in the case of photo-z task, filtering methods are unlikely to be useful for feature selection. On the other hand, the serious improvement of RF performance with the RFE-selected features together with an absence of such improvement in the case of MLPQNA suggests that the development of a faster implementation of wrapper method for MLPQNA should be the preferred approach.

3.3.6 Outlier analysis

Keeping in mind that COSMOS2015 has exceptionally rich and well-calibrated photometry, one could expect to obtain photo-z catalogue of high precision. Our expectations were also based on the very precise results obtained with SED fitting in Laigle et al. [2016], and on the fact that in previous works (e.g. done for SDSS-DR9 [Cavuoti et al.,

		ML photo-z				SED photo-z			
spec-z bin	Num objects	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$
[Overall]	5967	0.048	0.019	-0.0023	1.64	0.094	0.011	-0.0041	2.23
[0.0; 0.2]	422	0.1	0.027	-0.0416	8.29	0.278	0.012	-0.0552	7.82
[0.2; 0.4]	1445	0.043	0.016	-0.0081	1.31	0.056	0.008	-0.0034	1.94
[0.4; 0.6]	1158	0.038	0.016	0.0003	1.04	0.067	0.01	-0.0002	2.07
[0.6; 0.8]	1489	0.036	0.018	0.0001	1.01	0.052	0.012	0.0002	1.48
[0.8; 1.0]	1085	0.032	0.019	0.0041	0.55	0.044	0.014	0.0057	1.2
[1.0; 1.2]	368	0.051	0.028	0.0292	2.99	0.094	0.027	-0.0068	3.53

Table 3.9: Statistics for the test set calculated in spec-z bins. ML photo-z used here were obtained during the exp007 from Tab. 3.3 using ten broad and one medium bands.

2017b] and KiDS [Brescia et al., 2014]), MLPQNA generally performed on a comparable or better level than other photo-z algorithms.

Yet, in all the experiments, reported in Tab. 3.3, the accuracy achieved in previous works was not reached. In particular, $\sigma(\Delta z_{\mathrm{ML}})$ and percentage of outliers are noticeably higher than those reported in the aforementioned papers for KiDS and SDSS-DR9. It is especially strange that the addition of the IR and UV bands (see Tab. 3.3 `exp002` and `exp006, exp007` and `exp008`) does not significantly reduce the percentage of outliers, despite the fact that it should have been able to break colour-redshift degeneracy, one of the most common sources of catastrophic outliers.

The COSMOS2015 is deeper than the catalogues used in those publications, and it combines photometry coming from various sources, so the deterioration of the overall statistics could be attributed to the issues of photometry calibration or the quality of spectroscopy for the high-z sources. But even for low-z galaxies (i.e. with $z_{\mathrm{spec}} < 0.5$), both the percentage of outliers and $\sigma(\Delta z_{\mathrm{ML}})$ turned out to be of lower quality (see Tab. 3.9).

These photo-z outliers can be degenerate solutions, exotic objects or misinterpreted spectra (e.g. false outliers), and thanks to the fact that some objects in our main spec-z catalogue contain more than one spec-z measurement, we can try to disentangle these contributions. In order to do this, I selected objects with more than one measurement of spec-z and calculate maximum difference between these measurements, calling it spec-z scatter. Then I calculated the percentage of outliers separately for objects with a single spec-z measurement, multiple measurements and small (< 0.1) scatter, and multiple measurements and large (≥ 0.1) scatter. Tab. 3.10 reports these calculations for ML and SED fitting photo-z.

From this table we can deduce several facts:

1. For the objects with multiple spec-z values the percentage of outliers for ML and SED fitting photo-z is essentially the same.
2. For the objects with small spec-z scatter this percentage is significantly lower ($\eta_{0.15} \sim 0.2\%$) than for the objects with large spec-z scatter ($\eta_{0.15} \sim 11\%$).
3. For the objects with single spec-z measurement the percentage of outliers for SED fitting is $\sim 50\%$ higher than for ML.

The most probable explanation here is that for the photo-z outliers with large spec-z scatter, the specific spec-z measurement used to estimate the photo-z residual is incorrect. Consequently, the majority of such outliers are likely to have correct photo-z predictions. It also implies that some percentage of the outliers with single spec-z measurements should be attributed to incorrect spec-z measurements. At the same time, we should not assume that the percentage of incorrect spec-z measurements in this group is the same as for the objects with multiple measurements, since there is no guarantee of similarity of the selection functions for these two categories of objects.

Instead, I used the SOM cleaning procedure, described in § 3.2.4, to select a set of objects with reliable spec-z even without multiple spec-z measurements.

3.4 Photo-z after in-cell outlier cleaning with SOM

3.4.1 Spec-z in-cell outliers

We train the SOM, using the KB with the set of broad bands that gave us the best results for the MLPQNA experiments (`B,H,J,Ks,V,Y,ip,r,u,zpp`, `exp006` from Tab. 3.3). Fig. 3.12 shows the resulting SOM map, colour-labelled with, respectively, the mean and standard deviation of spec-z and ML and SED fitting photo-z. In order to discard objects with anomalous spec-z values, I calculated the in-cell spec-z outlier coefficients K_{spec} for each source, as defined in Eq. (3.1); then I binned the whole KB according to the value of this coefficient and calculate the statistics of photo-z residuals for each bin. This allows

		ML photo-z		SED photo-z	
Case	Num objects	Num outliers	% outliers	Num outliers	% outliers
Total	5967	98	1.64	133	2.23
Single measurement	3745	63	1.68	98	2.62
Multiple measurements, spec-z scatter <0.1	1945	4	0.21	5	0.26
Multiple measurements, spec-z scatter ≥ 0.1	277	31	11.19	30	10.83

Table 3.10: Statistics for ML and SED outliers for the test set of the KB for objects with different number and scatter of spec-z measurements.

us to check how the quality of photo-z correlates with similarity between the spec-z of a given source and the mean spec-z of its BMU.

Fig. 3.11 shows that the majority of the objects have relatively small K_{spec} (second row of the figure), and the statistics are much better for them than for the objects with larger absolute values of K_{spec}.

In the majority of the bins ML photo-z's have lower standard deviations (third row from the top) and lower percentage of outliers (last row from the top) than SED photo-z's, but higher NMAD (fourth row). Predictably, mean residuals (second row from the bottom) have inverse correlation with K_{spec}, implying that photo-z predictions, for objects with spec-z lower than the median spec-z of their BMU, are biased towards higher values, and vice versa.

By limiting our dataset to galaxies with absolute value of K_{spec} smaller than 1, we reduce the percentage of outliers from 1.64 to 0.19 for ML photo-z and from 2.23 to 0.7 for SED fitting. The standard deviation of residuals also is reduced by a factor ~ 2 (see Tab. 3.12).

Remarkably, after this cleaning the statistics are effectively the same as for the objects with multiple spec-z measurements and low spec-z scatter. In other words, removing in-cell anomalous spec-z leaves us with a reliable set of spec-z with no need to use repeated spec-z measurements, and this set is twice as large as the one with multiple measurements.

Tab. 3.11 shows that while the percentage of outliers for the ML photo-z drops not only for the objects with large spec-z scatter, but for other categories as well, for the SED fitting the improvement for the objects with single spec-z measurement is weaker.

		ML photo-z		SED photo-z	
Case	Num objects	Num outliers	% outliers	Num outliers	% outliers
Total	4311	8	0.19	30	0.70
Single measurement	2683	5	0.19	24	0.89
Multiple measurements, spec-z scatter <0.1	1468	0	0.00	1	0.07
Multiple measurements, spec-z scatter ≥ 0.1	160	3	1.88	5	3.12

Table 3.11: Statistics for ML and SED outliers for the test set of the KB for objects with different number and scatter of spec-z measurements after removing in-cell anomalous spec-z ($|K_{\mathrm{spec}}| \leq 1$).

3.4.2 Photo-z cleaning with SOM

Taking into account that the run dataset does not contain spec-z, it appeared useful to check whether the quality of the dataset could be improved by using photo-z in-cell anomaly detection instead of spec-z. To do so, I calculated outlier coefficients for SED and ML photo-z (K_{SED} and K_{ML} correspondingly), and applied the same filtering using these coefficients instead of K_{spec}. It turns out that such filtering improves the overall statistics, even though less than the K_{spec} cleaning (upper half of Tab. 3.12). As expected, the relative improvements are stronger for the SED fitting photo-z than for the ML photo-z; in § 3.7.3 I will discuss this aspect in more detail.

3.5 Calibration of the datasets with SOM

3.5.1 Calibration of the DEIMOS dataset

Before applying the trained MLPQNA model to the run dataset, we have to discard galaxies that are not photometrically similar to the galaxies of the train dataset. To estimate the performance of such photometry cleaning, I used the DEIMOS dataset. Considering that the DEIMOS dataset is slightly deeper than the train dataset, we can expect that the cleaning process will mostly remove faint objects. Fig. 3.13 shows occupation maps for the train, test, DEIMOS and run datasets, respectively. For each object of the test, DEIMOS and run datasets I derive the occupation of its BMU by the galaxies of the train dataset. As expected, Fig. 3.18 demonstrates that the statistics tends to be better for cells with higher occupation by the train dataset. Based on the

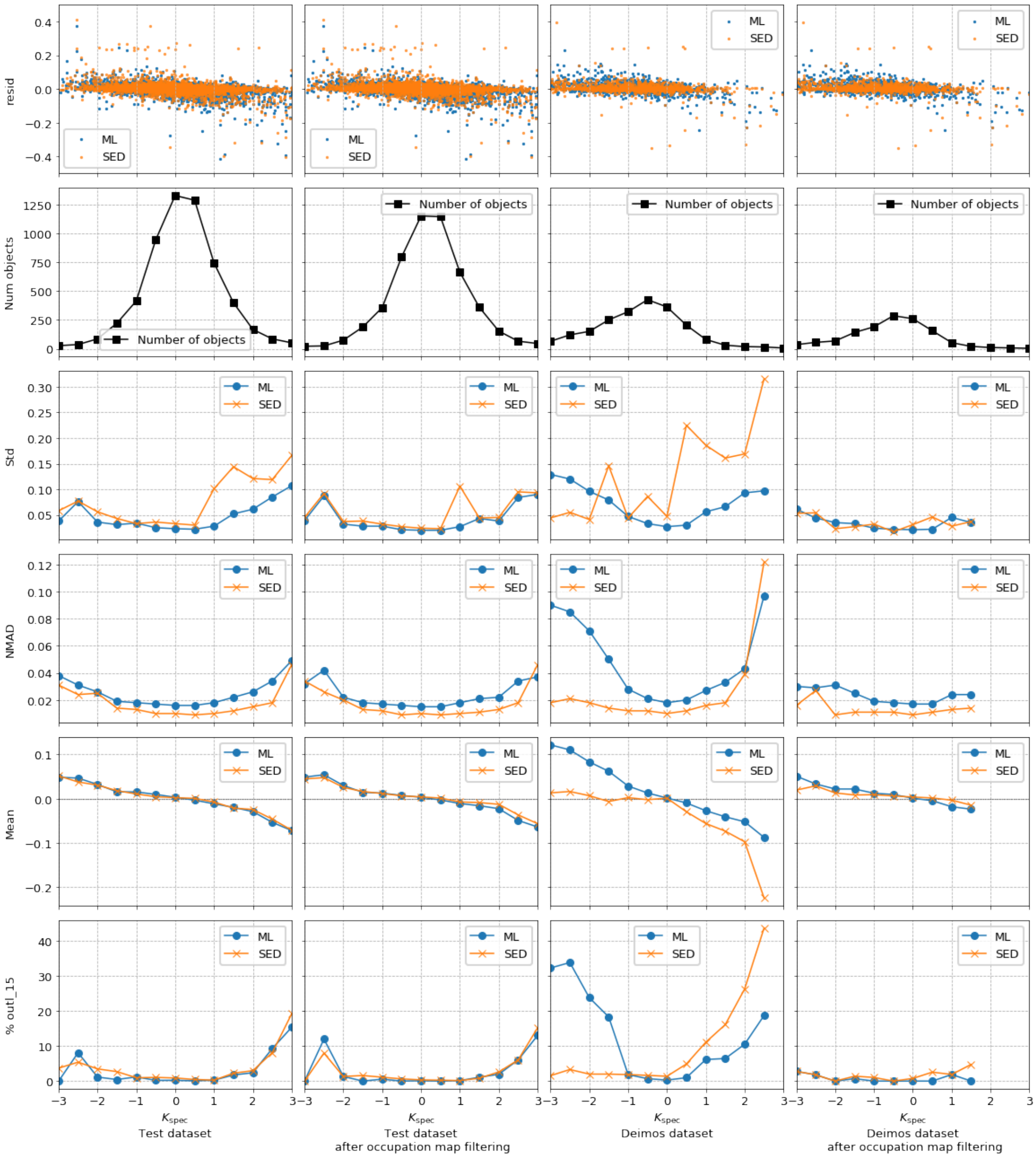


FIGURE 3.11: Statistics for the test and DEIMOS datasets in $K_{\rm spec}$ bins before and after occupation map filtering. Bins with number of objects < 15 are considered to be unreliable and excluded from these plots.

statistics for the DEIMOS dataset (right panel of Fig. 3.18), I chose to leave only the objects that belong to the cells with occupation > 5, since more strict criteria do not bring further improvements.

Tables 3.12 and 3.13 report the overall statistics for the test and DEIMOS datasets, respectively. The statistics are given for the different stages of cleaning, including the cases of combination of cleaning with occupation map and removal of the objects with anomalous spectroscopic or photometric redshifts. From these tables, we see that the effect of occupation filtering is much stronger for the DEIMOS than for the test dataset (since the test dataset by construction covers the same regions of the parameter space

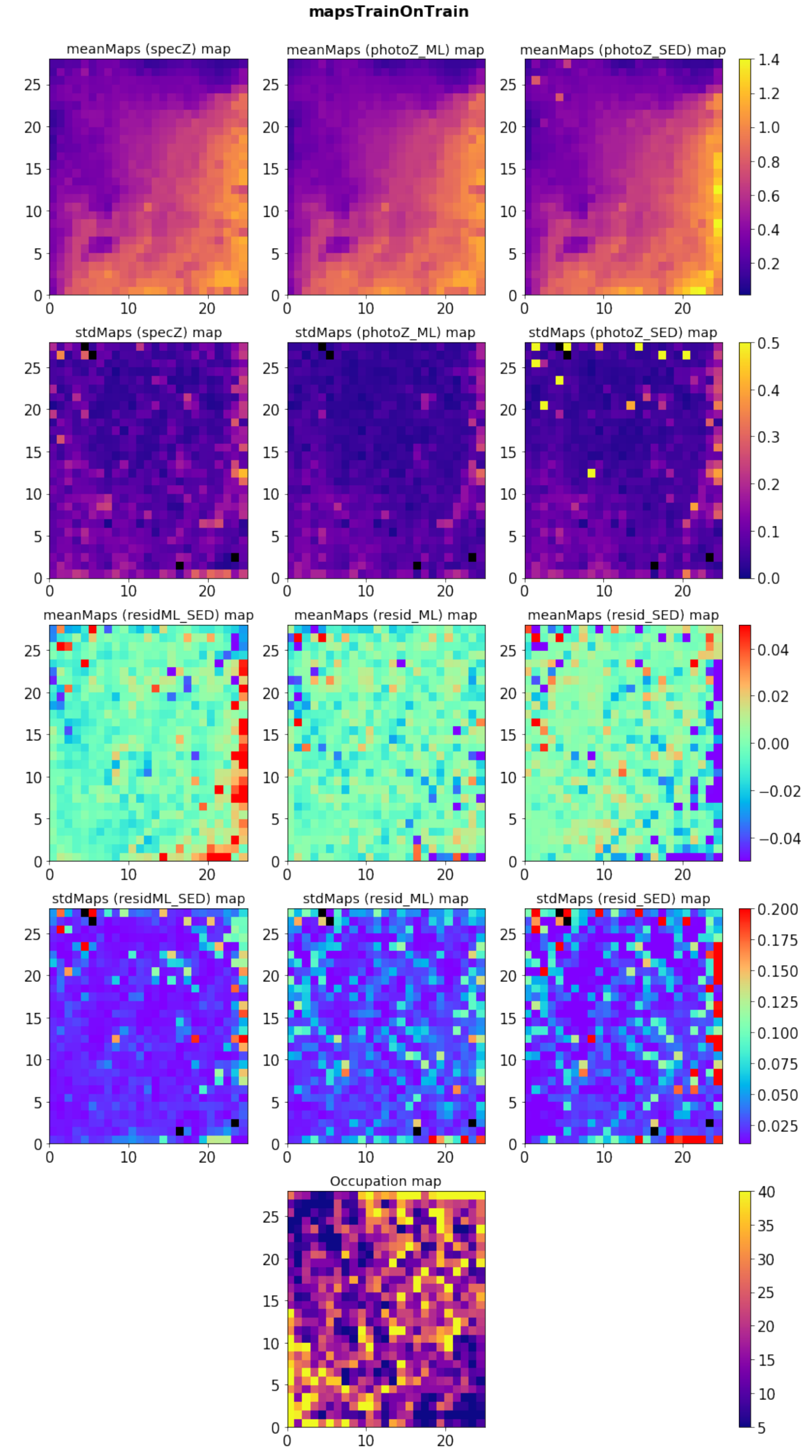


Figure 3.12: SOMs built and labelled with KB. The first two rows from the top illustrate mean and standard deviation of the redshifts (spectroscopic on the left, ML in the center and SED fitting on the right) within each cell. The fourth and fifth rows illustrate mean and standard deviation of residuals: on the left are the maps calculated for the residuals between ML and SED fitting photo-z, in the center and on the right are the residuals for ML and SED fitting photo-z. The last plot reflects how many objects are within each cell. Black cells on the occupation and mean maps imply that these cells are empty. On the maps of standard deviation black cells mean that occupation of the cell equals 1 and standard deviation cannot be calculated.

		ML photo-z				SED photo-z			
Filtering	Num objects	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$
No filtering	5967	0.048	0.019	-0.0023	1.64	0.094	0.011	-0.0041	2.23
$\lvert K_{\text{spec}}\rvert \leq 1$	4311	0.025	0.017	0.0002	0.19	0.052	0.01	0.0006	0.7
$\lvert K_{\text{ML}}\rvert \leq 1$	4071	0.045	0.018	-0.002	1.28	0.077	0.01	-0.0026	1.65
$\lvert K_{\text{SED}}\rvert \leq 1$	4133	0.043	0.018	-0.002	1.06	0.061	0.01	-0.0021	1.43
trainMapOccupation > 5	5167	0.041	0.018	-0.0021	1.18	0.058	0.01	0.0001	1.49
K_{spec} + trainMapOccupation	3761	0.022	0.016	-0.0002	0.05	0.05	0.009	0.0018	0.35
K_{ML} + trainMapOccupation	3587	0.039	0.017	-0.0015	1.06	0.062	0.01	-0.0003	1.17
K_{SED} + trainMapOccupation	3624	0.038	0.017	-0.0016	0.94	0.038	0.01	0.0002	0.99

Table 3.12: Statistics for ML and SED fitting photo-z calculated for the test dataset after different types of filtering. The upper part of the table presents the statistics calculated for the dataset without spec-z and photo-z outliers. The lower part reports the effects of photometric filtering with occupation map.

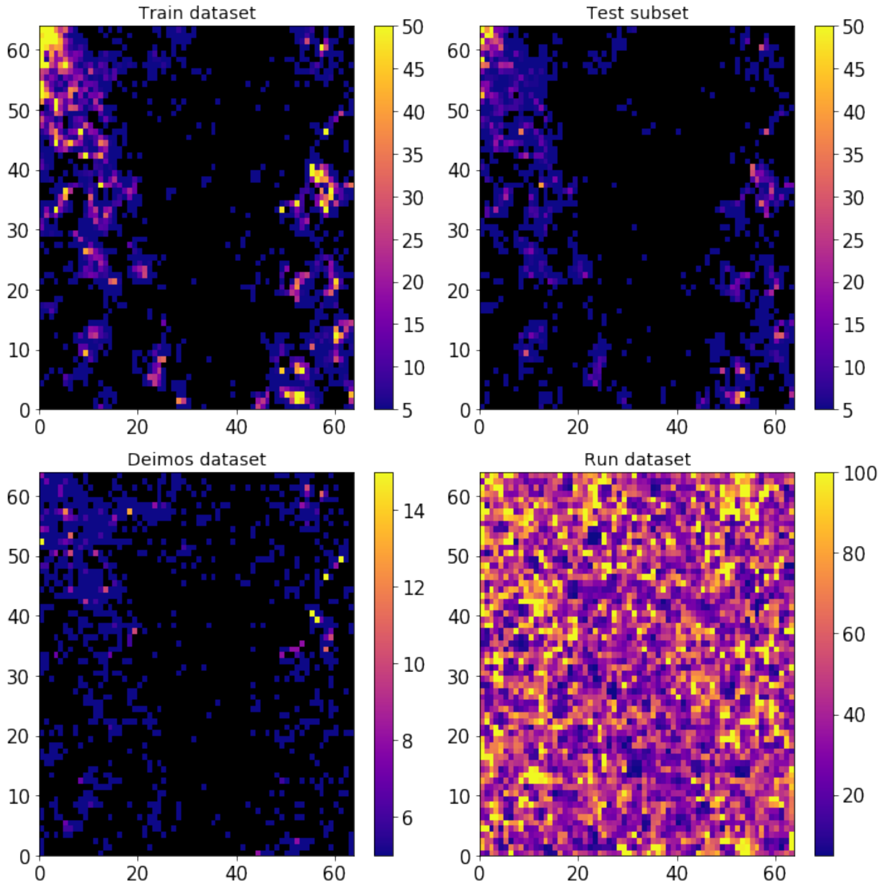


Figure 3.13: Occupation maps for all datasets projected on SOM trained with run dataset.

		ML photo-z				SED photo-z			
Filtering	Num objects	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$
No filtering	2255	0.099	0.032	0.0347	10.86	0.142	0.014	-0.0082	5.06
$\lvert K_{\mathrm{spec}}\rvert \leq 1$	1075	0.035	0.02	0.0018	1.02	0.127	0.011	-0.0103	2.88
$\lvert K_{\mathrm{ML}}\rvert \leq 1$	1209	0.095	0.033	0.0392	12.57	0.09	0.013	-0.0031	4.55
$\lvert K_{\mathrm{SED}}\rvert \leq 1$	1183	0.085	0.029	0.0245	8.96	0.078	0.013	0.0005	3.3
trainMapOccupation>5	1382	0.058	0.023	0.0127	2.1	0.059	0.012	0.0085	2.68
K_{spec} + trainMapOccupation	758	0.025	0.018	0.0017	0.13	0.031	0.01	0.0038	0.92
K_{ML} + trainMapOccupation	724	0.064	0.022	0.0136	1.93	0.06	0.011	0.008	2.62
K_{SED} + trainMapOccupation	741	0.046	0.02	0.0063	1.48	0.044	0.011	0.0075	1.89

TABLE 3.13: Statistics for ML and SED fitting photo-z for the DEIMOS dataset after different types of filtering. Upper part of the table describes statistics after spec-z and photo-z outlier removal, while the lower part reports the effects of photometry filtering with occupation map.

as the train dataset). Still, even on the test dataset the effect of the filtering with occupation map is comparable or better than the effect of the filtering of the photo-z anomalous sources. Also, occupation map filtering is more cost-effective in a sense that it discards much less objects than spec-z or photo-z anomalous source filtering.

From Tab. 3.13 we see that without any filtering applied to the DEIMOS dataset, our ML photo-z has worse indicators than SED photo-z. This can be easily understood by remembering that part of the DEIMOS dataset lays outside of the parameter space of the train dataset. But for objects that belong to the cells with good occupation, the percentage of outliers and standard deviations for both ML and SED fitting photo-z are very close.

After additional spec-z anomalous source filtering, the statistics improves even more and reaches approximately the same values as for the KB; the scatter plots in Fig. 3.14 clearly illustrate the improvement.

3.5.2 Calibration of the run dataset

The COSMOS2015 contains $\sim$ 500 000 galaxies. After the standard preprocessing described in § 3.1 and excluding objects that belong to the KB, the run catalogue consists of $\sim$ 190 000 galaxies. For all these galaxies, I calculated ML photo-z, and in order to determine the reliability of these redshifts, I calculated K_{SED}, K_{ML} and the occupation of their BMU by the objects in the train dataset, as described in the previous subsections.

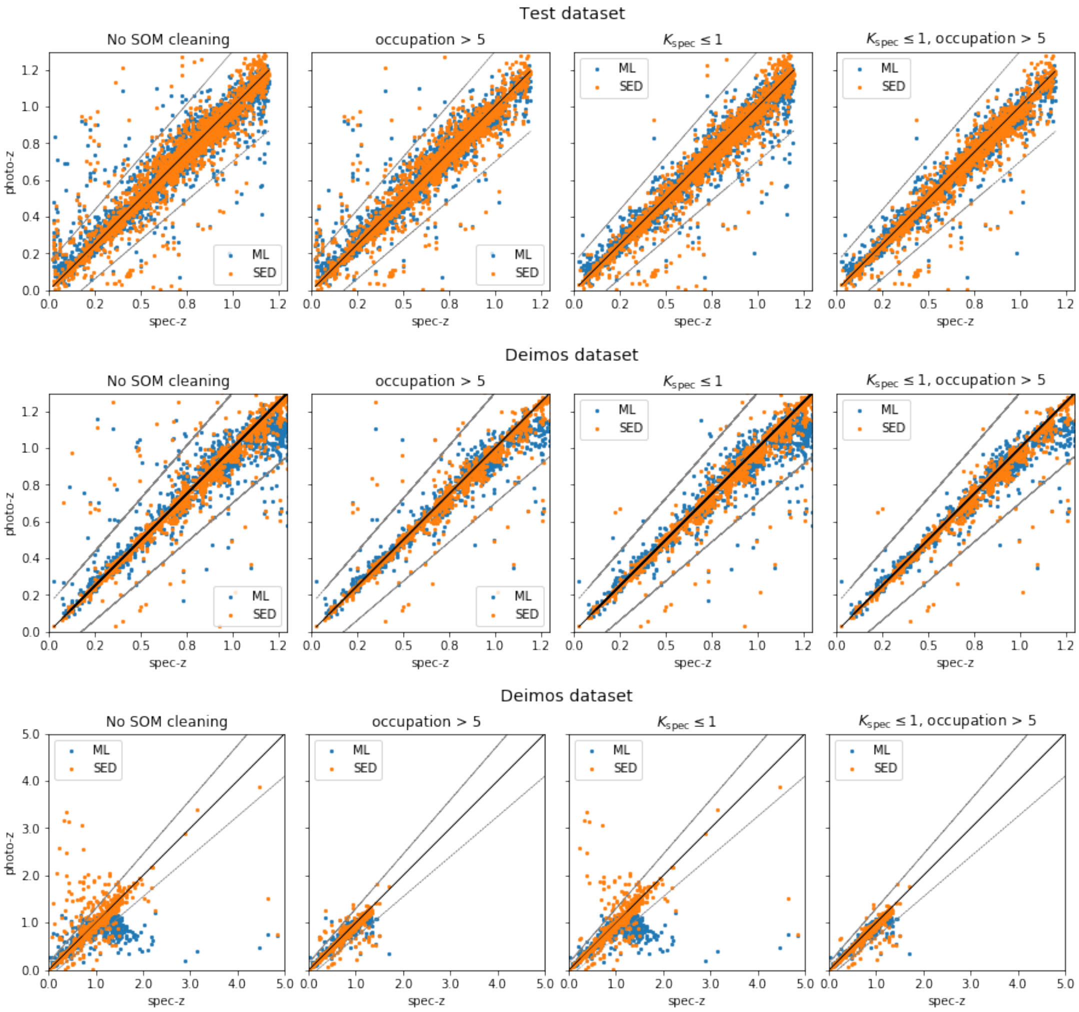


FIGURE 3.14: Scatter plots of ML and SED fitting $z_{\rm phot}$ against $z_{\rm spec}$. In the first column on the left are the datasets before SOM filtering procedures, in the second column are the results after only $z_{\rm spec}$ outlier filtering, in the third are the plots after only occupation map filtering, and in the last column are the datasets after the two cleaning together. The dotted lines show outlier boundaries defined as $z_{\rm photo} = z_{\rm spec} \pm 0.15 \cdot (1 + z_{\rm spec})$. The first two rows show the results for the test and DEIMOS datasets limited to $z_{\rm spec} < 1.2$ and $z_{\rm phot} < 1.2$. It can be seen that the two cleaning procedures remove different outliers. The last row demonstrates the predictions for the DEIMOS for the whole range of redshifts. It can be seen that both ML and SED fitting require occupation map filtering to select the objects with good predictions. Without it SED fitting produces a lot of catastrophic outliers in the whole range of $z_{\rm spec}$, and ML systematically fails for the objects with $z_{\rm spec} > 1.2$. Another observation to make is that the cleaning procedures are most effective against those outliers for which the predicted redshifts are higher than $z_{\rm spec}$, while the objects with underestimated $z_{\rm phot}$ are likely not to be removed.

	DEIMOS dataset					Run dataset				
Filtering	N obj	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$	N obj	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$
No filtering	2255	0.102	0.033	0.0355	11.35	194509	0.238	0.129	0.1373	40.23
$\lvert K_{\mathrm{spec}}\rvert \leq 1$	1075	0.07	0.023	0.0049	2.88	NA	NA	NA	NA	NA
$\lvert K_{\mathrm{ML}}\rvert \leq 1$	1209	0.088	0.034	0.0392	11.75	137152	0.225	0.11	0.1243	37.76
$\lvert K_{\mathrm{SED}}\rvert \leq 1$	1183	0.072	0.03	0.0224	7.44	146773	0.206	0.101	0.1217	34.93
trainMapOccupation> 5	1382	0.047	0.023	0.0031	1.52	43279	0.083	0.028	-0.0019	3.59
K_{spec} + occupation map	758	0.036	0.02	-0.0028	0.92	NA	NA	NA	NA	NA
K_{ML} + occupation map	724	0.043	0.022	0.0048	0.97	33296	0.08	0.028	-0.0029	3.48
K_{SED} + occupation map	741	0.03	0.022	-0.0015	0.27	35189	0.056	0.026	-0.0031	2.17

Table 3.14: Statistics for ML/SED residuals for the DEIMOS and run datasets after different types of filtering. Upper part of the table describes statistics after spec-z and photo-z outlier removal, lower part reports the effects of photometry filtering with occupation map. Taking into account that for the run dataset spectral information is absent, the rows corresponding to spec-z cleanings for run dataset are empty.

Since we do not have spec-z for the galaxies in the run dataset, we can only perform an accuracy test comparing SED and ML photo-z. To do so, I construct ML/SED residuals $\Delta z_{\mathrm{ML/SED}} = (z_{\mathrm{SED}} - z_{\mathrm{ML}})/(1+z_{\mathrm{SED}})$. Obviously, such a test suffers from biases introduced by both photo-z methods, so it can be used only for qualitative estimation of the photo-z robustness.

Tab. 3.14 compares the statistics for ML/SED photo-z residuals for the DEIMOS and run catalogues after applying different filters. The dataset calibration with occupation map appears to be the most important step, since it allows to reduce the percentage of outliers by almost an order of magnitude. Removal of SED fitting photo-z anomalous sources also improves the statistics. Maximum improvement for both datasets is achieved by a combination of occupation map filtering and removal of SED fitting photo-z anomalous sources. It reduces the percentage of outliers from 11% to 0.27% for the DEIMOS dataset, and from 40% to 2% for the run dataset. $\sigma(\Delta z)$ for the DEIMOS drops from 0.102 to 0.03, and from 0.238 to 0.056 for the run dataset. NMAD changes from 0.033 to 0.022 for the DEIMOS dataset and from 0.129 to 0.026 for the run dataset.

3.6 A purely data-driven selection of the reliable spec-zs

The successful removal of the unreliable spec-z measurements, described in § 3.4, suggests a possibility to create a data-driven complement for the quality flags system. This system would not be able to replace the standard quality checks, but rather be an additional check and a tool for choosing *likely* correct spec-z measurements from a sample of doubtful ones.

The idea is to apply the SOM in-cell anomaly filtering to the full uncleaned spec-z sample. By removing objects with high $|K_{spec}|$ from this sample, I created a catalogue of galaxies with spec-z values that are typical for their multi-dimensional colours. Obviously, there are two issues with this approach. The first one is that in-cell anomaly cleaning only removes galaxies with strongly mismatching redshift and colours, and does not detect objects with non-catastrophic spec-z uncertainties. For this reason, the sample of galaxies obtained with this methodology is expected to have higher scatter of residuals between spec-z and *true* redshifts. And the second is that this cleaning is likely to remove rare and anomalous objects, even if they have reliable spec-z. For some applications, such as predicting photo-z with ML methods or mapping large-scale structure, this can be acceptable, but for many astrophysical applications it will be not. However, in this cases it is possible to expand the `Q_f`-selected catalogue with galaxies that have unreliable `Q_f`, but low $|K_{spec}|$, which would provide a compromise between quality and completeness in terms of galaxy types.

In order to test this idea, I prepared a spec-z sample where only objects with $0.01 < z_{spec}$ and $z_{spec} > 9$, e.g. those that can be contaminated by stars and those without measured spec-z, are removed; the AGN and quality flags cleaning, described in § 3.1.2, was omitted. As a result, instead of 19 893 objects obtained after standard pre-processing, we have a dataset of $\sim$ 30 630 objects. Then I trained SOM algorithm on this dataset and remove in-cell spec-z outliers, using $|K_{spec}| \leq 1$ criteria. This leaved us with 24 058 objects, which spec-z distribution is essentially the same as for the spec-z sample after standard `Q_f` cleaning, with a slight overabundance in the high-redshift part of the distribution (Fig. 3.15). It is also worth to remember that the criteria $|K_{spec}| \leq 1$ is somewhat arbitrary; choosing other threshold will allow to preserve even larger sample at cost of lower reliability of the spec-z.

After that, I obtained photo-z for this sample, using the same MLPQNA setup as in § 3.3. For additional comparison, I also calculated photo-z for a random galaxy sample of the same size, taken from the non-cleaned spec-z dataset.

Tab. 3.15 reports the results of these experiments in comparison with a baseline experiment before and after in-cell spec-z anomaly cleaning. It shows that using only SOM

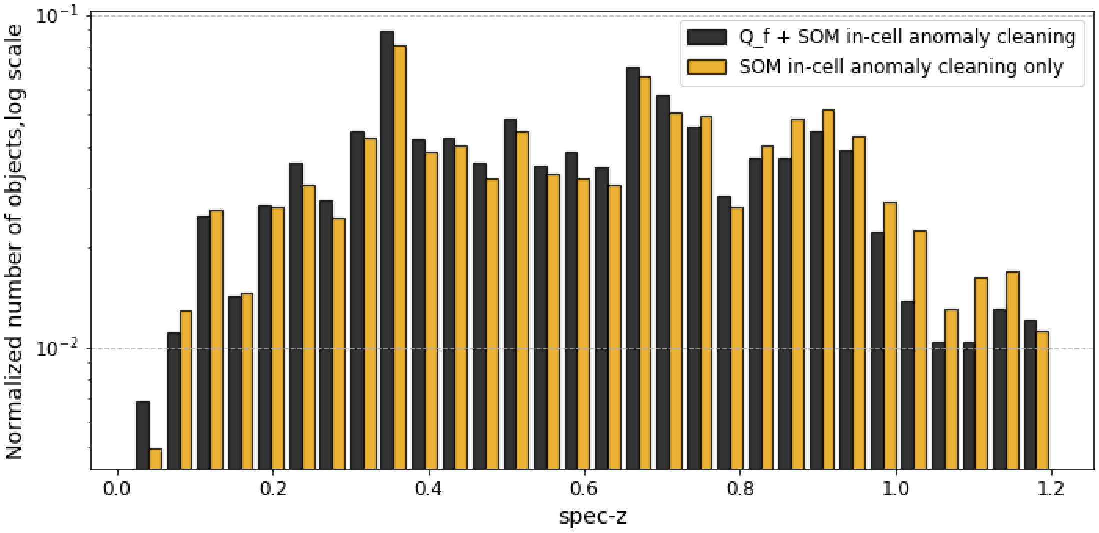


FIGURE 3.15: Spec-z distribution for the test dataset with and without standard `Q_f` cleaning

in-cell anomaly filtering, we are unable to obtain the same quality of photo-z as with a combination of standard quality flag pre-processing and SOM filtering. However, for the ML photo-z the difference is rather small; with the combined standard and SOM cleaning, we have NMAD $= 0.017$ and $\eta_{0.15} = 0.19$, while with SOM-only cleaning the statistics are NMAD $= 0.018$ and $\eta_{0.15} = 0.56$. This deterioration of statistics comes with a benefit of increasing the size of the spec-z dataset by $\sim 67\%$. Remarkably, unlike ML, SED fitting photo-z show noticeable degradation on this dataset; mean bias is worsened by two orders, from 0.0006 to -0.0134, and percentage of outliers increases from 0.7 to 1.93. The experiment with a random sample taken from the uncleaned spec-z catalogue demonstrates that without any form of data cleaning, photo-z quality is significantly worse: for ML photo-z, the percentage of outliers is $\sim 4.3\%$, NMAD equals 0.025, and mean bias is exceeding -0.005, and for SED fitting photo-z the percentage of outliers is $\sim 5.78\%$, NMAD equals 0.015, and mean bias is -0.0238.

Unsurprisingly, Tab. 3.16 shows that $|K_{spec}|$ criteria and `Q_f` system have some rough correlation. In the case of our spec-z sample, only five `Q_f` categories (`1, 2, 3, 4`, and `9`) have enough objects to provide reliable statistics. Two of them, `Q_f=3` and `Q_f=4`, correspond to secure redshifts, `Q_f=9` marks secure best-guess one-line redshifts, and `Q_f=1` and `Q_f=2` mean insecure and probable redshifts correspondingly [Lilly et al., 2009]. The percentage of objects with $|K_{spec}| \leq 1$ differs for these categories from $> 80\%$ for the reliable `Q_f=3, 4` to $\sim 65\%$ for `Q_f=1`. This result is unsurprising because the `Q_f` system is calibrated with photo-z, and every photo-z algorithm in its core performs some sort of analysis of the 'typicality' of a redshift for the given colour combination, just as SOM in-cell anomaly analysis does. The benefit of of SOM-cleaning is that it is both

			ML photo-z				SED photo-z			
Exp ID	Description	Num obj	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$
exp007	Baseline	5967	0.048	0.019	-0.0023	1.64	0.094	0.011	-0.0041	2.23
exp007-clean	Baseline, $K_{spec} \leq 1$	4311	0.025	0.017	0.0002	0.19	0.052	0.010	0.0006	0.7
d_exp027	'Dirty' spec-z, $\|K_{spec}\| \leq 1$	7218	0.033	0.018	-0.0009	0.57	0.135	0.012	-0.0134	1.93
d_exp028	'Dirty' spec-z, random sample	7218	0.072	0.025	-0.0055	4.32	0.178	0.015	-0.0238	5.78

Table 3.15: Photo-z performance for the test sample with and without standard `Q_f` cleaning. The first two rows describe baseline `exp007` experiment, which was done on the spec-z sample cleaned with standard `Q_f` criteria (§ 3.1.2), before and after SOM in-cell spec-z anomaly filtering. The third row describes an experiment done on the spec-z sample that was not cleaned with `Q_f`, but only with SOM in-cell filtering. The last row reports an experiment done on a random sample from the spec-z catalogue which was not cleaned neither with `Q_f` nor SOM in-cell filtering.

`Q_f`	N obj	% with $\|K_{spec}\| \leq 1$	`Q_f`	N obj	% with $\|K_{spec}\| \leq 1$
1	1085	65.3	12	1	100.0
2	8626	71.3	21	41	61.0
3	8378	81.5	22	53	84.9
4	11417	83.0	23	75	85.3
5	4	50.0	24	39	79.5
6	4	50.0	29	23	78.3
9	884	80.3			

Table 3.16: Percentage of objects with $|K_{spec}| \leq 1$ for different spec-z quality flags (`Q_f`).

data-driven, i.e. not limited by the SED template library and SED fitting setup, and easy to interpret and visualize, which is a rare quality for the ML photo-z algorithms.

3.7 Discussion

As I have shown in the previous section, there are several ways in which we can apply SOM to improve the quality of the photo-z catalogues. In this section I compare the effects of the SOM cleaning on ML and SED fitting photo-z (§ 3.7.1), consider the nature of spec-z and photo-z in-cell anomalies (§ 3.7.2 and § 3.7.3), and define general strategies for using the SOM cleaning methodology on other datasets (§ 3.7.5).

3.7.1 SED fitting vs. ML

In all the experiments performed on the KB, the ML photo-z distribution has a lower percentages of outliers and lower standard deviations than the SED fitting photo-z distribution, but higher NMAD (see Tab. 3.12). For the DEIMOS dataset, before SOM filtering, the situation is different: ML photo-z have a significantly higher percentage of outliers ($\sim$ 11% against 5% for SED fitting), due to the fact that DEIMOS contains many objects laying outside of the boundaries of the parameter space sampled by the KB. Dataset calibration with occupation map discards the majority of these outliers and the statistics of both ML and SED fitting photo-z become similar to those for the test dataset (see Tab. 3.12 and 3.13). In particular, for ML $\sigma_{\Delta z}$ decreases from 0.099 to 0.058 (for the KB it equals 0.041), and for SED fitting $\sigma_{\Delta z}$ from 0.142 to 0.059 (0.058 for the KB). The percentage of outliers also drops from 10.86% to 2.1% for ML photo-z (1.18% for the KB) and from 5.06% to 2.68% for SED fitting (1.49% for the KB).

The difference between the two methods in terms of NMAD can be explained by the fact that SED fitting methods benefit from the inclusion of the narrow-band photometry. It allows the SED fitting methods to detect emission lines passing through a certain wavelength range [Ilbert et al., 2009]. On the contrary, in the case of MLPQNA, the same inclusion of the additional narrow bands does not lead to significant improvements (Tab. 3.3). Earlier work by several authors (see e.g. Heinis et al. [2016] and Eriksen et al. [2020]) demonstrated that inclusion of the narrow-band photometry generally improves ML photo-z only after additional preparations, such as transfer learning or a extensive preliminary feature selection.

Even after occupation map filtering and removing in-cell anomalous spec-z, SED fitting produces a higher number of catastrophic outliers than ML. Also, the residuals for these outliers are larger than for ML outliers (Fig. 3.14: for the test dataset the mean absolute value of residuals for SED fitting catastrophic outliers is 0.35, while for ML outliers it's 0.26; for the DEIMOS the mean absolute residual for SED fitting is 0.29, and for ML it's 0.22). As can be seen from the second row of Fig. 3.12, the SOM map of the standard deviations of SED fitting photo-z has a number of cells with significantly higher values than appear on the $\sigma(z_{\rm spec})$ and $\sigma(z_{\rm ML})$ maps. These are the cells where the majority of the catastrophic outliers are located. From Fig. 3.16 it also can be seen that there for SED fitting photo-z there is a localized group of catastrophic outliers which do not disappear after dataset calibration with occupation map, but disappear after in-cell SED photo-z anomaly removal. It seems likely that these outliers appear due to the lack of a suitable SED template or SED fitting failure.

A somewhat unexpected result is that SED fitting seriously benefits from occupation map calibration. For the test dataset the percentage of outliers drops from 2.23% to 1.49%, while for DEIMOS it goes from 5.06% to 2.68%, which is even better than what is obtained after spec-z filtering. One possible explanation to this is that the objects in the areas of parameter space that are poorly covered by the spec-z catalogue are likely to be faint and at a high spec-z, thus having a less reliable photo-z.

The aforementioned results for ML photo-z are obtained using only 10 broad and 1 medium band. In the case of SED fitting photo-z more than thirty (broad, medium and narrow) bands were used; still, the statistics for SED fitting and ML are quite close. It implies some consequences for the observational strategies of the future surveys. Depending on the design of a survey, it can be more beneficial to invest resources into obtaining either a larger spectroscopic KB with ML photo-z algorithms in mind, or additional medium/narrow band photometry for SED fitting. Both decisions lead to similar quality of the photo-z catalogue with some differences in the NMAD and in the percentage of outliers. Obviously, obtaining both an extensive spectroscopic catalogue and medium/narrow band photometry will allow to use both techniques and to increase the reliability of the photo-z predictions (e.g. Cavuoti et al. [2017b], also see § 3.7.5).

3.7.2 Spec-z anomalous sources filtering

Removal of the spec-z anomalous sources drastically reduces the percentage of outliers for both photo-z methods. As was shown in § 3.3, the photo-z outliers mostly appear to be misinterpreted spectra, and removing them from the analysis makes the performance estimation more correct. Yet, there are much more objects with high values of K_{spec} than ML or SED fitting outliers. A question arises, what are all these objects.

Looking at Fig. 3.11, we can see that in general the statistics (including NMAD, which is less sensitive to hard outliers) deteriorate smoothly with the increase of the absolute value of K_{spec}. This smooth dependency is present not only for ML photo-z, but for SED fitting photo-z as well. It means that within each cell of SOM (which represents a small hyper-volume in the photometric parameter space), both ML and SED fitting perform better for "typical" spec-z than for atypical ones. This is true even for those regions of parameter space that are well-covered by the spec-z catalogue. In other words, K_{spec} works as an additional, finer indicator of whether a given galaxy is well-represented in the spectroscopic KB. However, it is not completely clear to what degree this 'under-representation' has some physical ground, and to what degree it is simply a consequence

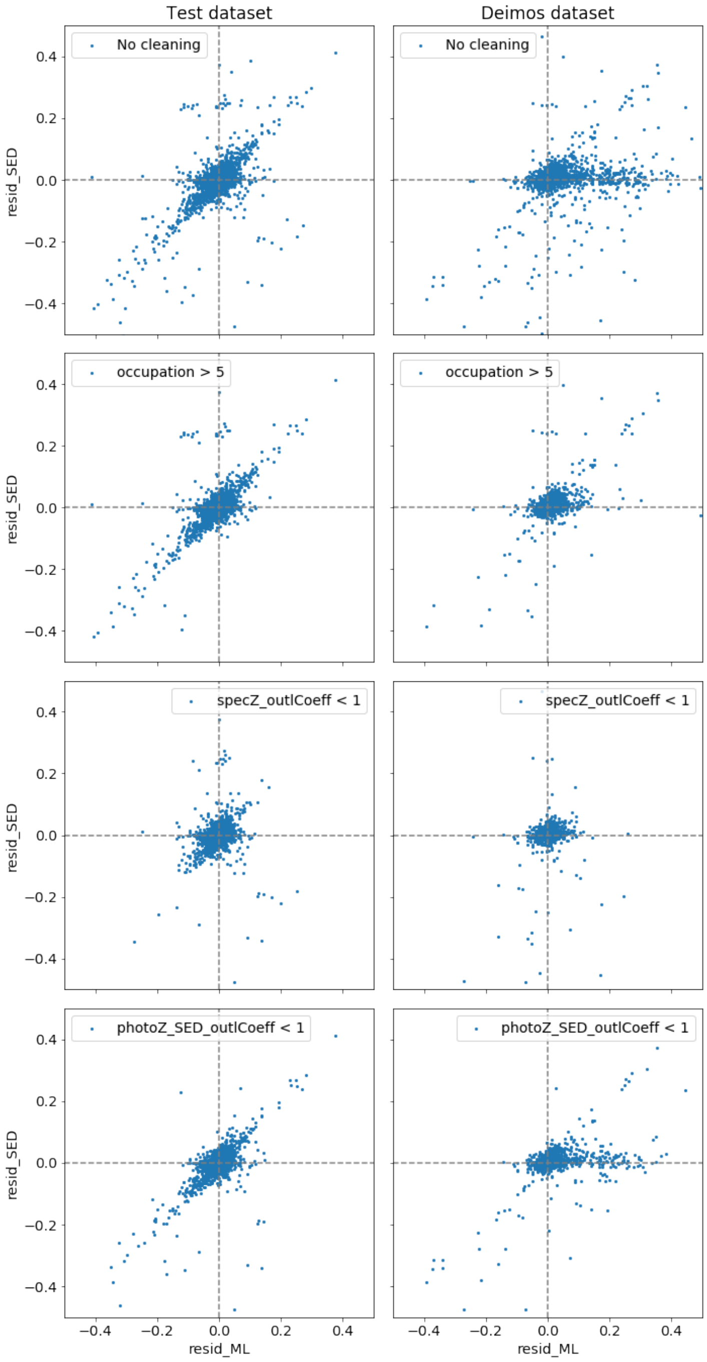


Figure 3.16: Correlation plots of ML and SED fitting photo-z residuals. Left panel: the test dataset, right panel: the DEIMOS. Several groups of object are present. On the left panel there is an evident diagonal stripe of objects for which ML and SED fitting predictions are very close, but different from spec-z values. This stripe disappears on the third row, where in-cell spec-z anomalies are removed. There is also a clearly distinguishable group of objects with `resid_ML` ~ 0 and `resid_SED` ~ 0.2; using the terminology from § 2.8.1, they appear to be localized outliers, some type of objects that are represented in the KB (and well-learned by the MLPQNA), but not represented in the SED library. This group is also present in the plots from the right panel. These objects disappear when SED in-cell anomaly are removed (bottom row). Finally, on the top plot in right panel there is a group of objects for which ML predictions are much worse than SED fitting ones; these objects disappear both after dataset calibration with occupation map (second row) and in-cell spec-z anomaly removal (third row).

of the fact that every photo-z code makes redshift predictions with a resolution which is rougher than the resolution of spec-z measurements.

In any case, removing objects with high K_{spec} allows us to create a reliable spec-z sample, useful for comparing the performance of different photo-z algorithms. In our case, for the test dataset this sample is more than two times larger than the set of the objects with multiple spec-z measurements that are in good agreement with each other. The usage of this method seems to be highly beneficial for the preparation of any future compound spec-z catalogues, and possibly for the verification of the new spec-z surveys against the old ones.

3.7.3 Photo-z anomalous sources and occupation map filtering

To clean photo-z values for the run dataset, the only possibility is to use K_{ML} or K_{SED} to discard either ML or SED fitting photo-z anomalies. However, the performance is different in the two cases. As it can be seen from Tab. 3.12 and 3.13, discarding SED fitting anomalies improves the statistics slightly more than removing ML anomalous sources. Obviously, this is related to the large percentage of outliers obtained in SED fitting photo-z and to their large residuals.

On the DEIMOS dataset, photo-z anomalous sources filtering, together with occupation map cleaning allows to reach $\sigma_{\Delta \mathrm{z}} \approx 0.046$ and percentage of outliers $\approx 1.48\%$ for ML photo-z, and $\sigma_{\Delta \mathrm{z}} \approx 0.044$ and percentage of outliers $\approx 1.89\%$ for SED fitting. At the same time, for ML/SED residuals on the DEIMOS the same cleaning procedure brings $\sigma_{\Delta \mathrm{z}} \approx 0.03$ and percentage of outliers $\approx 0.27\%$, which is quite close to the statistics for ML and SED photo-z after occupation map filtering together with removing spec-z anomalous sources.

Fig. 3.19 shows the values of different statistics in `ipmagap3` magnitude bins at different stages of cleaning of the DEIMOS dataset. As expected, with occupation map calibration the improvements are mostly achieved due to the filtering in the fainter part of magnitude distribution. Instead of completely loosing the faint objects, as it would have happened if we used the traditional cut-off procedure, we are preserving those which have fairly good quality of photo-z predictions. Fig. 3.20 shows the same picture for the run dataset. For what the standard deviation is concerned, the difference appears for objects fainter than `ipmagap3` ~ 21, but the most significant effect is observed for objects with `ipmagap3` > 23.

3.7.4 Potential biases introduced by SOM cleaning procedures

Every astrophysical or cosmological application has different requirements to the selection function of the redshift sample. For example, for weak lensing analysis, it is primarily the size of the sample in each redshift bin, which should be large enough to provide high statistical reliability, low mean bias and percentage of catastrophic outliers, which is needed to minimize systematic errors of cosmological parameters, and reliability of the redshift quality estimators, to correctly account for these systematics. For other tasks, such as studying luminosity functions of certain types of objects, the completeness of the sample becomes more important. In any case, in order to obtain meaningful scientific results, it is necessary to take into account the biases introduced by the data cleaning procedures.

As can be seen from Fig. 3.19, all cleaning procedures change magnitude distribution of the sample: fainter objects have much higher chances of being discarded. However, as can be seen from the upper panel of Fig. 3.17, the normalized distribution of spec-z of the test sample does not change much. In comparison to the redshift distribution before SOM cleaning, the distribution after such cleaning has a sharp diminishing of the number of objects with $z_{spec} < 0.02$ (considering the high percentage of outliers in the closest redshift bin, evident from e.g. Tab. 3.9, one might suspect some residual contamination by stars, perhaps introduced during crossmatch of photometric and spectroscopic catalogues), and much softer decline of the relative share of objects with $1 < z_{spec} < 1.2$, but the shape of the middle part of the distribution remains intact. The lower panel of Fig. 3.17 shows that in the case of the DEIMOS, the SOM cleaning procedures do exactly what they are supposed to do: remove objects with z_{spec} higher than 1.2, which was the limit for the KB. Similarly to the distribution for the test sample, some decline of the percentage of objects at the tails of the distribution is present, and the middle part of it remains essentially unaffected.

The cut of the total size of the spec-z sample is a serious downside of the method, especially for the cosmological applications. On the other hand, as I have shown in § 3.6, SOM in-cell anomaly filtering can be used to *increase* the size of the sample, if one is willing to sacrifice rare types of objects and accept potentially higher scatter of spec-z uncertainties.

3.7.5 Filtering strategies

The optimal strategy of the data cleaning with SOM depends on the nature of a dataset and on the task at hand:

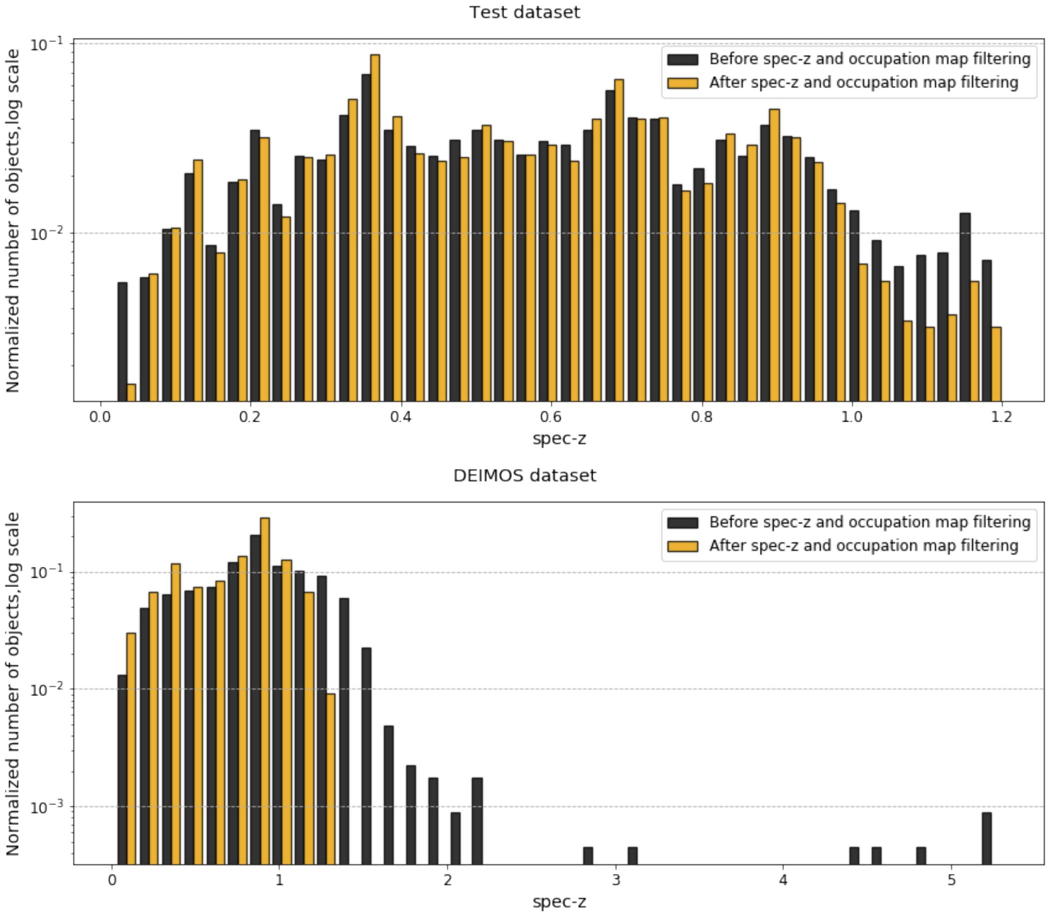


FIGURE 3.17: Spec-z distributions before and after SOM cleaning.

- Identification of spec-z anomalous sources is useful for finding unreliable spec-z and under-represented objects.
- Identification of photo-z anomalies allows us to improve the quality of a photo-z catalogue, especially for SED fitting photo-z. It can be applied to the run catalogue (i.e., for objects without spectral information), but this procedure is effective only for datasets that are well sampled by the KB.
- For the run datasets that are not well sampled by the KB, SOM filtering with an occupation map is the most important step. It leads to significant improvements not only in the case of ML based methods but also in the case of SED fitting.

In this work I used all three types of cleaning with the following thresholds:

1. spec-z anomalies filtering: $K_{\mathrm{spec}} \leq 1$.
2. Occupation map filtering: `trainMapOccupation` > 5.
3. photo-z anomalies filtering: $K_{\mathrm{SED}} \leq 1$.

	Test dataset					DEIMOS dataset				
Photo-z	Num objects	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$	Num objects	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$
ML	3508	0.021	0.016	-0.0001	0.03	725	0.024	0.018	0.002	0
SED	3508	0.049	0.009	-0.0001	0.06	725	0.027	0.009	0.0022	0.55
Mixed	3508	0.018	0.009	0.0012	0.03	725	0.017	0.01	0.0051	0

Table 3.17: Statistics for ML, SED fitting and mixed (selected based on the ML/SED residual value) residuals for the test and DEIMOS datasets. Both datasets were cleaned with the following conditions: trainMapOccupation > 5, $|K_{\text{spec}}| \leq 1$, $|K_{\text{SED}}| \leq 1$.

4. In cases when better NMAD is needed and percentage of outliers is less critical, SED fitting photo-z are preferable. For the tasks that demand the lowest percentage of outliers, ML photo-z show better results. Finally, it is possible to choose between the two values of photo-z: when photo-z predictions are similar (based on the grid search on the test and DEIMOS datasets, it is recommended to use the criteria of ML/SED residual < 0.5), it is better to select SED fitting value, and when the residual is > 0.5, ML photo-z are more likely to be correct. Tab. 3.17 reports the statistics for photo-z selected in this way for the test and DEIMOS datasets. With such selection we obtain both good NMAD and low percentage of outliers.

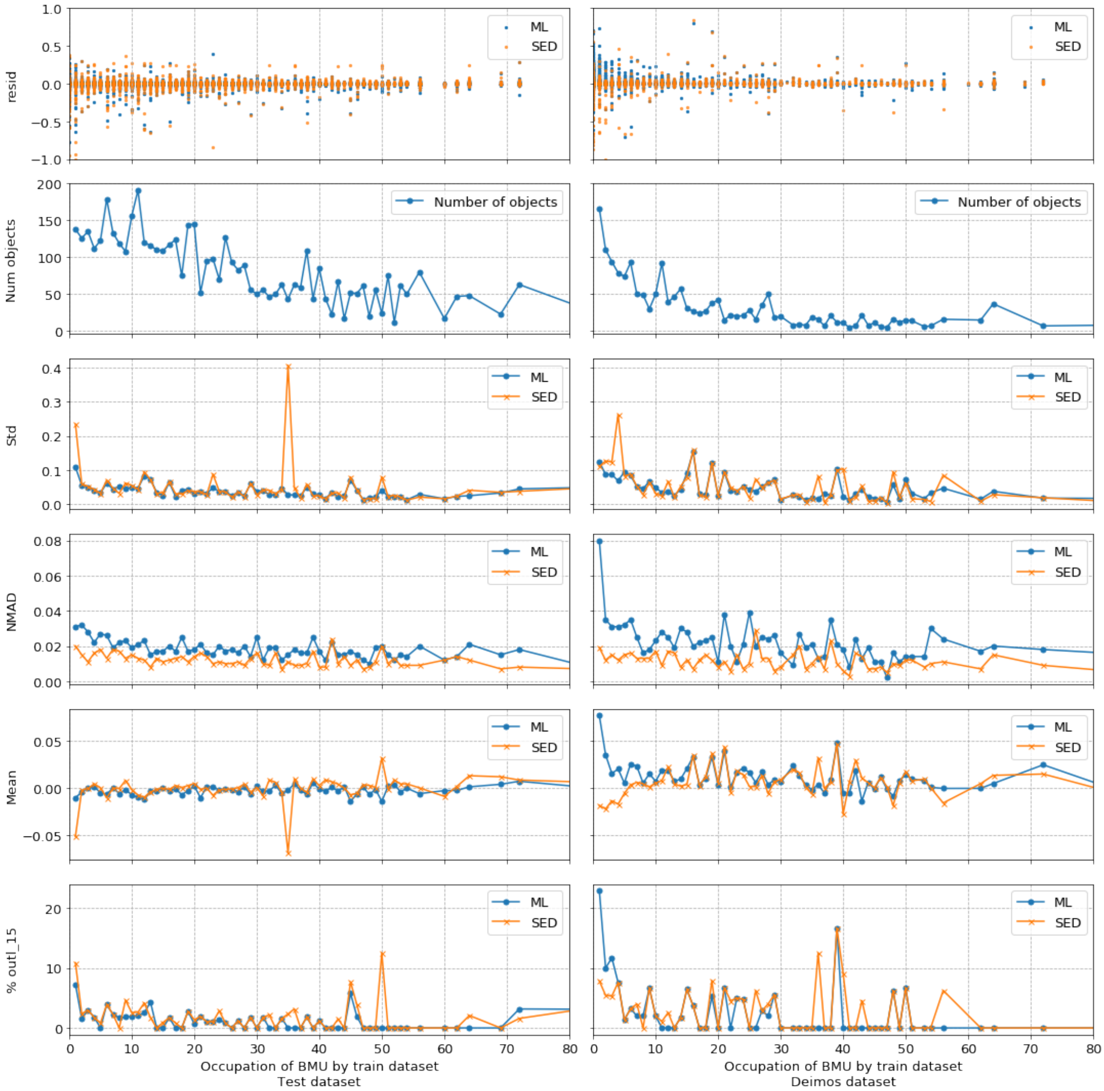


FIGURE 3.18: Dependency of the statistics for the test and DEIMOS datasets from the occupation of their BMU by objects from train dataset. The x-axis is limited to occupation ≤ 80 since there are not enough objects in the cells with bigger occupations to calculate reliable statistics.

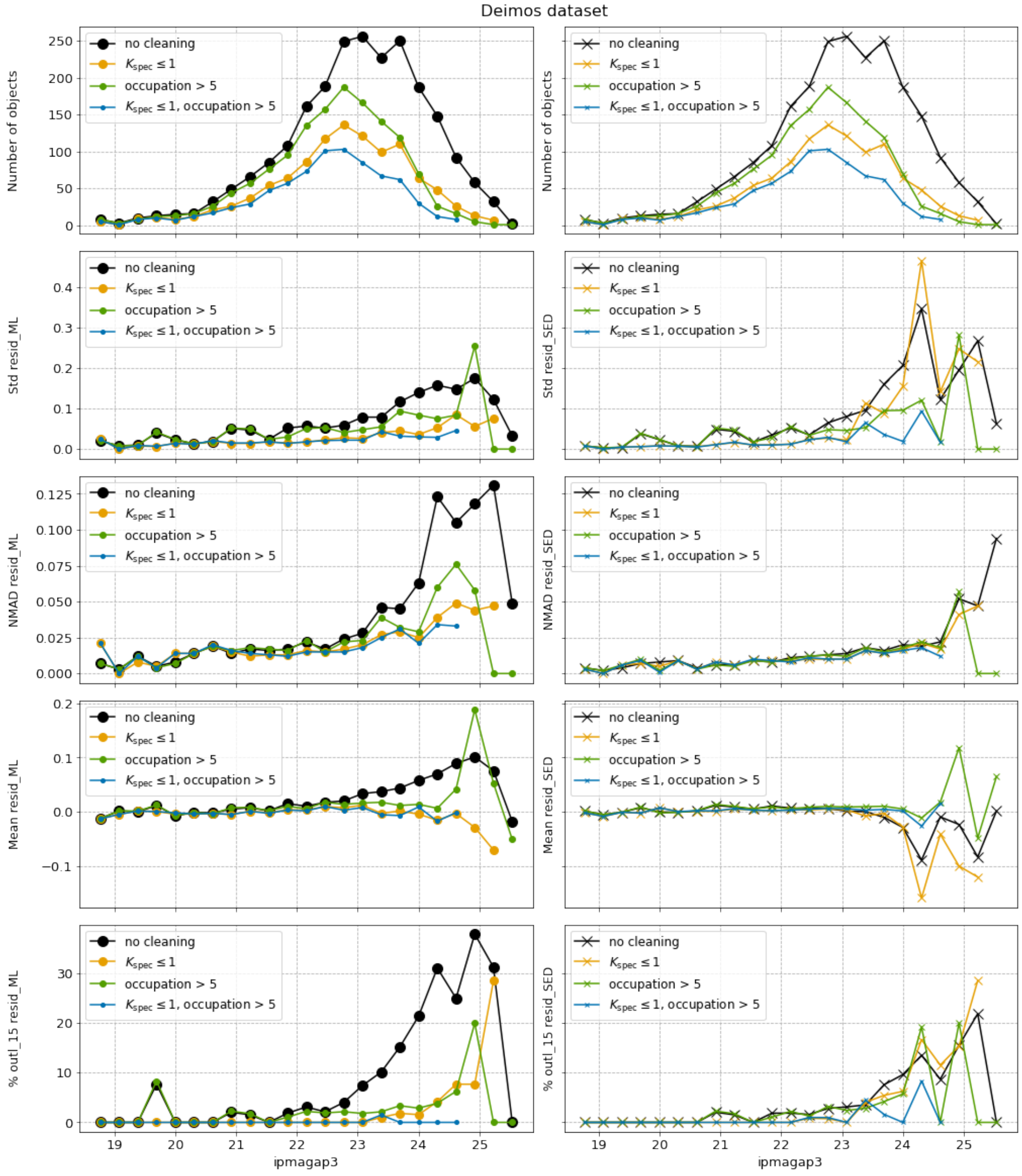


FIGURE 3.19: Statistics for the DEIMOS datasets in `ipmagap3` bins after applying different filters. Left panel: ML photo-z residuals, right panel: SED fitting photo-z residuals. Bins with number of objects < 15 are considered to be unreliable and excluded from these plots.

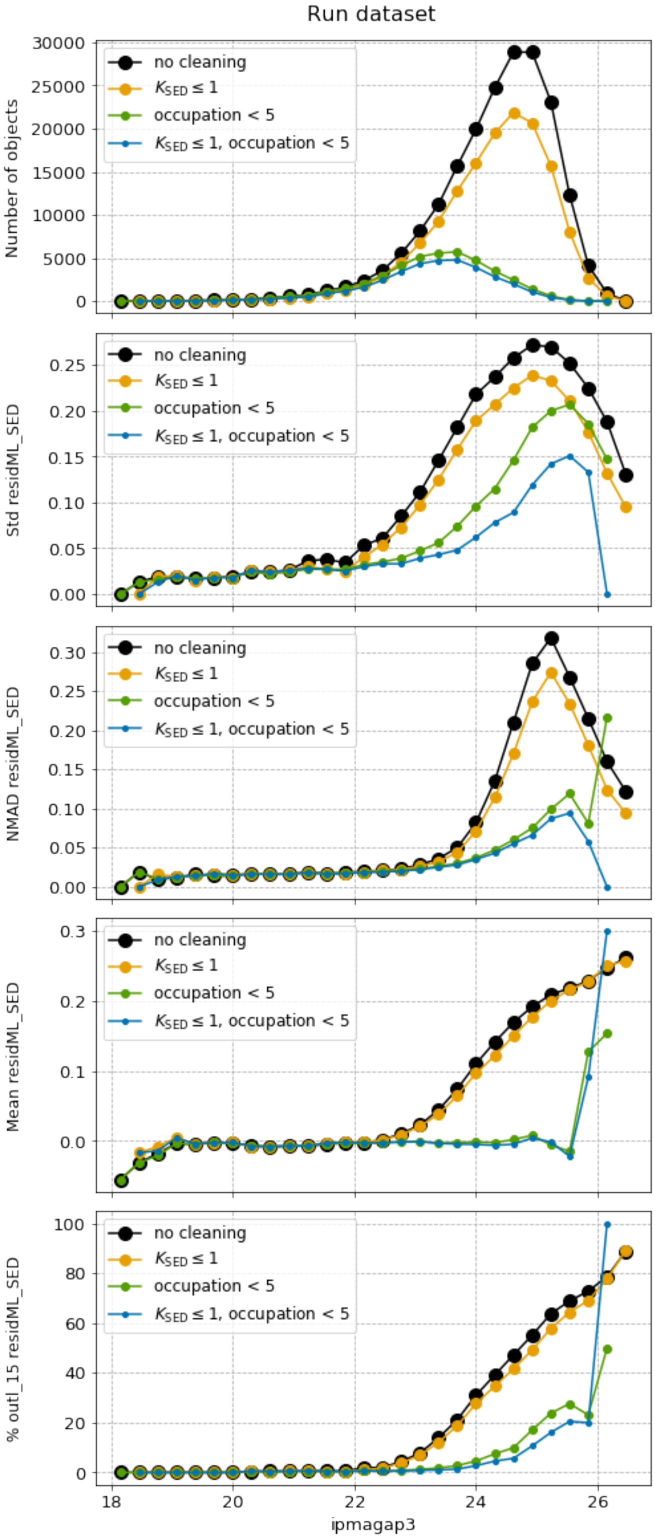


FIGURE 3.20: Statistics for ML/SED residuals for run dataset in `ipmagap3` bins after applying different filters. Bins with number of objects < 15 are considered to be unreliable and excluded from these plots.

Chapter 4

Conclusion

In this work, ML photo-z for the COSMOS2015 catalogue were calculated for the first time; it was done using MLPQNA algorithm. For the training and testing I used multi-instrument spectroscopic KB with $z_{\text{spec}} \leq 1.2$ and various sets of photometric bands, obtaining the best results with a feature set composed by 10 broad and 1 narrow bands. The comparison of the statistics for ML photo-z and SED fitting photo-z, calculated by Laigle et al. [2016] using the whole set of COSMOS2015 bands, showed that ML photo-z has lower percentage of outliers and $\sigma_{\Delta z}$, but higher NMAD. Particularly, for the test dataset without additional SOM cleaning MLPQNA produces photo-z with $\sigma_{\Delta z} = 0.048$, NMAD $= 0.019$ and $\eta_{0.15} = 1.64\%$, while SED fitting photo-z has $\sigma_{\Delta z} = 0.094$, NMAD $=$ 0.011 and $\eta_{0.15} = 2.23\%$ (Tab. 3.12).

Unlike SED fitting algorithms, the ML photo-z algorithm does not significantly benefit from the inclusion of the most medium and narrow photometric bands (Tab. 3.3 and § 3.3.3). To check this, I used various feature sets, chosen from general physical reasoning as well as from the results of the experiments with automatic feature selection algorithms (§ 3.3.5). Finding a way to exploit the information contained in those bands should be a subject of further work.

The experiments demonstrated that a significant percentage of outliers have similar values of ML and SED fitting photo-z. By analysing the objects with multiple spec-z measurements, I discovered that the majority of such outliers have unreliable spec-z values, which makes it likely that the photo-z prediction for these objects are correct and the actual percentage of outliers for both photo-z methods is significantly lower. On the subset of galaxies with multiple similar spec-z measurements, $\eta_{0.15} \sim 0.2\%$ for both photo-z methods.

I tested the possibility of using Self-Organizing Maps (SOM) for removing unreliable spec-z and creating a high-quality spec-z sample. To do this, I calculated a coefficient $K_{\rm spec}$ that quantifies how much a spec-z of a given galaxy differs from the mean spec-z of all the galaxies belonging to the same SOM cell, e.g. of the galaxies that are most photometrically similar. The resulting $K_{\rm spec}$ were used to remove objects with incorrect spec-z: as a consequence, the percentage of outliers for the test set of the KB dropped from 1.64 to 0.19 for ML photo-z and from 2.23 to 0.7 for SED fitting, and $\sigma_{\Delta z}$ for both ML and SED fitting photo-z improved almost by a factor of ~ 2 (see Tab. 3.12). At the same time, $K_{\rm spec}$ is sensitive to the intrinsic in-homogeneity of the galaxy population, caused by physical reasons. In this way, $K_{\rm spec}$ serves as a fine indicator of whether a given object is well-represented within the KB.

In § 3.6 I used $K_{\rm spec}$ cleaning to prepare a reliable spec-z dataset without using standard quality flags provided withing the catalogue. The performance of ML photo-z algorithm on this dataset is better than for the sample cleaned with quality flags but not cleaned with SOM, and only slightly worse than for the sample after both types of cleaning. Additionally, the sample obtained with only SOM cleaning is by 67% larger than the sample for which quality flag cleaning was also used. However, SED fitting shows serious degradation of statistics on this sample.

To ensure that our run dataset occupies the same area of the parameter space as the KB, I also used SOM occupation maps to calibrate the datasets. I found that after this cleaning, on the control DEIMOS spec-z dataset, that is slightly deeper than the KB spec-z catalogue, the percentage of outliers drops from 11% to 2% for ML photo-z and from 5% to 3% for SED fitting photo-z, with consequent improvements of other metrics. The details are reported in Tab. 3.13.

The statistics also improve after excluding the objects with SED fitting photo-z values that are anomalous for their SOM cells. Removing objects with anomalous ML photo-z does not improve the results significantly. Using both occupation map and SED fitting photo-z in-cell anomalies filtering, it is possible to bring the statistics for the DEIMOS dataset to the order of those for the KB. To be more precise, the standard deviation drops from $\sigma_{\Delta z} = 0.099$ to $\sigma_{\Delta z} = 0.046$ for ML and from $\sigma_{\Delta z} = 0.142$ to $\sigma_{\Delta z} = 0.044$ for SED fitting, and the percentage of outliers lessens from $\eta_{0.15} = 10.86$ to $\eta_{0.15} = 1.48$ for ML and from $\eta_{0.15} = 5.06$ to $\eta_{0.15} = 1.89$ for SED fitting. This result allows us to select the parts of photometric catalogues for which our photo-z predictions, obtained with any algorithm, can be trusted.

All in all, the SOM in-cell anomaly detection, presented in this work, proved to be a viable method for selecting reliable spec-z samples from a contaminated catalogue and a good tool for identifying SED fitting photo-z outliers. The SOM occupation map calibration is recommendable for ensuring the reliability of the future photo-z catalogues. Together, in-cell anomaly detection and occupation map calibration allow to create larger spec-z and photo-z catalogues which simultaneously have better quality and reliability; consequently, it is recommended to apply this methodology to the verification of the redshift catalogues of the upcoming Euclid and LSST surveys.

Appendix A

Methodology for bibliography analysis

A.1 The frequency of the ML-related terms in literature in 1950-2020

The data for the Fig. 1.1 are obtained using Google Books Ngram Viewer[1]. This engine provides information on the frequency with which the words (or combinations of words) appear in printed sources in a given time range in a chosen text corpora. The frequencies are normalized for the number of books published in a given year, so the rise of the number of publication does not affect the plot. For the Fig. 1.1 I used case-insensitive search across 'English (2019)' corpora, which includes books printed in any country and on any topic (as opposed to e.g. 'English Fiction' corpora). The code for reproducing the plot is available in the GitHub repository `https://github.com/ShrRa/google-ngrams-2020updated`.

A.2 Number of ML-related astronomical publications by years

The data for the Fig. 1.2 are obtained using the Astrophysics Data System (ADS) API[2]. The request used is `database:astronomy, property:refereed, abs:("machine learning" or "neural network" or "perceptron" or "neural-network" or "random forest" or "support vector machine" or "self-organizing map" or "k-nearest`

[1] `https://books.google.com/ngrams/`

[2] https://github.com/adsabs/adsabs-dev-api

`neighbor")`, e.g. it counts only refereed astronomical papers that contain one of the popular ML-related terms in title, abstract or keywords. The code for reproducing the plot is available in the GitHub repository `https://github.com/ShrRa/thesis_astroML`.

A.3 Popularity of the ML photo-z methods

The data for the Fig. 2.3 are obtained using the ADS API. The requests retrieve only refereed astronomical papers published between 2000 and 2020 (`year:2000-2020`) with the words 'photometric redshift' in the title (e.g. `database:astronomy, property:refereed, title:"photometric redshift"`). For each category of methods several related ML terms were grouped together and added to the main part of the request:

- **DNN+CNN**: `abs:("deep neural network" or "convolutional neural network")`;
- **RF+DT**: `abs:("random forest" or "decision tree" or "TPZ")`;
- **SVM**: `abs:("support vector machine")`;
- **Gaussian process**: `abs:("gaussian process" or "gaussian processes")`;
- **SOM**: `abs:("self-organizing map")`;
- **k-NN**: `abs:("k-nearest neighbor")`;
- **Shallow NN**: `abs:("neural network" or "perceptron" or "neural-network") -abs:("deep neural network" or "convolutional neural network")`;
- **Other**: `abs:"machine learning" -abs:("neural network" or "perceptron" or "TPZ" or "Mixture" or "Gaussian process" or "Gaussian processes" or "support vector machine" or "self-organizing map" or "decision tree" or "random forest")`;

The code for reproducing the plot is available in the GitHub repository `https://github.com/ShrRa/thesis_astroML`.

Appendix B

SOM repository

The source catalogues for this work can be obtained as described in Sect. 3.1. The final catalogue, containing MLPQNA photo-z and SOM-produced parameters that can be used for selecting objects with high-confidence predictions will be published via CDS Vizier facility. The code for reproducing this work is available in the GitHub repository https://github.com/ShrRa/COSMOS_SOM. The MLPQNA software is available within the PhotoRApToR[1] (PHOTOmetric Research APplication To Redshifts Cavuoti et al. [2015a]) package.

[1] http://dame.oacn.inaf.it/dame_photoz.html#photoraptor

Acknowledgements

It takes a village to raise a PhD. The full list of acknowledgements and gratitudes could make another chapter, so I have to settle for a brief and highly incomplete version of it.

First and foremost my gratitude goes to my family. My grandparents, prominent scientists and medical doctors, did not live to see me becoming a PhD, but I am sure they did not expect anything less of me. They did everything in their power to provide me with a good start, which was not always easy considering the part of space-time continuum where we were located. No less I am indebted to my babysitter (as she would be called now, although she has always been a family to me), for she was the one who made sure that I would not grow up a completely asocial bookworm. Finally, I have to thank my mother, a scientist, a doctor, an adventurer and a thinker. She has always been and still is my unconditional supporter and a brilliant partner in countless conversations on all kinds of scientific topics. I am fairly sure that my enchantment with the Universe in general and space specifically is at least partially an echo of her dreams of becoming an astronaut.

I have always been exceptionally lucky with teachers. I owe a noticeable part of my intelligence to Elena Arinkina, Marina Petrakova, Ilya Gelfgat and Igor Kolupaev; they are going to be a great source of inspiration to me, if I am ever to tutor pupils of myself.

Anastasia, Anastasia the Redhead and Bogdan remain my partners in conversations in my head no matter how long we have not spoken in real life. The same goes for Alyona, to whom, much to my bitterness and sorrow, I will never have a chance to say this.

There is a number of people from the Astronomy Division of the Department of Physics of the V.N. Karazin University who I should be thanking here. Among them is Andrei Gretskii, who had been my first supervisor there, albeit for a very brief time. Also, my special gratitude goes to Prof. Vadim Kaydash, my BSc and MSc supervisor, for his patience and good humor, and Prof. Elena Bannikova, who had never been my supervisor but a great teacher and a mentor of sorts nonetheless. Apart from my education, I am grateful to them for accepting my candidacy for a junior software engineer position in the Institute of Astronomy in the interim between my graduate and PhD years. Another person who I should thank is Dr. Oleksiy Golubov, whose energy and enthusiasm were a great support in the end of my university term. Also I would like to thank Prof. Massimo Capaccioli who helped me to find a position when there were very little chances for this.

My coming to Naples was not easy. Rosella, Valeria Amaro, Civita Velucci and Maurizio D'Addona were exceptionally kind and helpful in the first weeks. I owe even more to Rosella Raguza and her family, who took care of me when I needed it most. Their kindness towards me, a complete stranger, was one of the first and most important lessons I got in Naples.

Guido Celentano, whose official position is the secretary of the Department of Physics at the University of Naples, is unofficially a guardian angel of all the Department's students, and a sole messenger of order in this chaotic city. Apart from thanking him, I thank the fate for his existence.

For the other lessons, in science, in academic life and in the ways things are in general, I am deeply grateful to the group I have had a privilege to work in and to the members of the ITN SUNDIAL. Being part of these two communities brought me a great deal of experience, opportunities and insights. My immediate gratitude for collaboration and tutoring goes to Prof. Massimo Brescia, Dr. Stefano Cavuoti and Prof. Maurizio Paolillo. I am especially grateful to my supervisor Prof. Giuseppe Longo, who made sure that I survive the first days in Naples and provided me with both insightful discussions and a lot of research freedom in the next three years. I hope one day *astroinformatics*, or whatever it will evolve into, will help us to figure out how the Universe works.

The reviewers of this thesis, Prof. Michael Biehl and Dr. Mara Salvato, put a lot of efforts into making it better and provided numerous useful comments. All the remaining mistakes are mine, but their number has been noticeably reduced thanks to their work.

Also, I would like to thank my colleagues and friends, Dr. Angela Raj and soon-to-be Dr. Olena Torbaniuk, for science, travels, and some gossiping now and then.

Finally, I thank Dr. Kseniia Sysoliatina, my dearest friend, my partner in science, my patient listener, challenger, critic and supporter rolled into one talented, generous person. I would have dropped out of university sometime around third year if not for you, my dear, and I would have been regretting it for the rest of my life.

This project has received financial support from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 721463 to the SUNDIAL ITN network.

Bibliography

Abbott T. M. C., Abdalla F. B., Alarcon A., et al. Dark Energy Survey Year 1 Results: Measurement of the Baryon Acoustic Oscillation scale in the distribution of galaxies to redshift 1. *MNRAS*, 483(4):4866–4883, March 2019. doi: 10.1093/mnras/sty3351. ADS URL: https://ui.adsabs.harvard.edu/abs/2019MNRAS.483.4866A.

Abdalla F. B., Banerji M., Lahav O., and Rashkov V. A comparison of six photometric redshift methods applied to 1.5 million luminous red galaxies. *MNRAS*, 417(3):1891–1903, November 2011. doi: 10.1111/j.1365-2966.2011.19375.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2011MNRAS.417.1891A.

Ablain M., Legeais J. F., Prandi P., et al. Satellite altimetry-based sea level at global and regional scales. *Surveys in Geophysics*, 38(1):7–31, November 2016. doi: 10.1007/s10712-016-9389-8. URL: https://doi.org/10.1007/s10712-016-9389-8.

Ackermann Sandro, Schawinski Kevin, Zhang Ce, et al. Using transfer learning to detect galaxy mergers. *MNRAS*, 479(1):415–425, September 2018. doi: 10.1093/mnras/sty1398. ADS URL: https://ui.adsabs.harvard.edu/abs/2018MNRAS.479..415A.

Adami C., Durret F., Benoist C., et al. Galaxy structure searches by photometric redshifts in the CFHTLS. *A&A*, 509:A81, January 2010. doi: 10.1051/0004-6361/200913067. ADS URL: https://ui.adsabs.harvard.edu/abs/2010A&A...509A..81A.

Adorf H. M. *Connectionism and neural networks*, volume 329, pages 213–245. Springer, 1989. doi: 10.1007/3-540-51044-3_25. ADS URL: https://ui.adsabs.harvard.edu/abs/1989LNP...329..213A.

Adorf H. M. Artificial Intelligence in Astronomy - a Forecast. In Heck Andre, editor, *Artificial Intelligence Techniques for Astronomy*, page 1, January 1990. ADS URL: https://ui.adsabs.harvard.edu/abs/1990aita.proc....1A.

Adorf H. M. Artificial intelligence for astronomy. ESO course held in 1990. *The Messenger*, 63:69–72, March 1991. ADS URL: https://ui.adsabs.harvard.edu/abs/1991Msngr..63...69A.

Ahumada Romina, Allende Prieto Carlos, Almeida Andrés, et al. The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. *APJS*, 249(1):3, July 2020. doi: 10.3847/1538-4365/ab929e. ADS URL: https://ui.adsabs.harvard.edu/abs/2020ApJS..249....3A.

Aihara Hiroaki, Armstrong Robert, Bickerton Steven, et al. First data release of the Hyper Suprime-Cam Subaru Strategic Program. *PASJ*, 70:S8, January 2018. doi: 10.1093/pasj/psx081. ADS URL: https://ui.adsabs.harvard.edu/abs/2018PASJ...70S...8A.

Aird J., Nandra K., Laird E. S., et al. The evolution of the hard X-ray luminosity function of AGN. *MNRAS*, 401(4):2531–2551, February 2010. doi: 10.1111/j.1365-2966.2009.15829.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2010MNRAS.401.2531A.

Albrecht R. *Applications of Artificial Intelligence in Astronomy - a View Towards the Future*, volume 329, page 247. Springer-Verlag, 1989. doi: 10.1007/3-540-51044-3_26. ADS URL: https://ui.adsabs.harvard.edu/abs/1989LNP...329..247A.

Almosallam Ibrahim A., Jarvis Matt J., and Roberts Stephen J. GPZ: non-stationary sparse Gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts. *MNRAS*, 462(1):726–739, October 2016a. doi: 10.1093/mnras/stw1618. ADS URL: https://ui.adsabs.harvard.edu/abs/2016MNRAS.462..726A.

Almosallam Ibrahim A., Lindsay Sam N., Jarvis Matt J., and Roberts Stephen J. A sparse Gaussian process framework for photometric redshift estimation. *MNRAS*, 455(3):2387–2401, January 2016b. doi: 10.1093/mnras/stv2425. ADS URL: https://ui.adsabs.harvard.edu/abs/2016MNRAS.455.2387A.

Angel J. R. P., Wizinowich P., Lloyd-Hart M., and Sandler D. Adaptive optics for array telescopes using neural-network techniques. *Nature*, 348(6298):221–224, November 1990. doi: 10.1038/348221a0. ADS URL: https://ui.adsabs.harvard.edu/abs/1990Natur.348..221A.

Arnouts S., Cristiani S., Moscardini L., et al. Measuring and modelling the redshift evolution of clustering: the Hubble Deep Field North. *MNRAS*, 310(2):540–556, December 1999. doi: 10.1046/j.1365-8711.1999.02978.x. ADS URL: https://ui.adsabs.harvard.edu/abs/1999MNRAS.310..540A.

Arviset C., Gaudet S., and IVOA Technical Coordination Group . The IVOA Architecture. In *European Planetary Science Congress 2012*, pages EPSC2012–626, September 2012. ADS URL: https://ui.adsabs.harvard.edu/abs/2012epsc.conf..626A.

Ashby Neil. Relativity in the global positioning system. *Living Reviews in Relativity*, 6(1), January 2003. doi: 10.12942/lrr-2003-1. URL: https://doi.org/10.12942/lrr-2003-1.

Ball Nicholas M., Brunner Robert J., Myers Adam D., et al. Robust Machine Learning Applied to Astronomical Data Sets. II. Quantifying Photometric Redshifts for Quasars Using Instance-based Learning. *ApJ*, 663(2):774–780, July 2007. doi: 10.1086/518362. ADS URL: https://ui.adsabs.harvard.edu/abs/2007ApJ...663..774B.

Barchi P. H., de Carvalho R. R., Rosa R. R., et al. Machine and Deep Learning applied to galaxy morphology - A comparative study. *Astronomy and Computing*, 30:100334, January 2020. doi: 10.1016/j.ascom.2019.100334. ADS URL: https://ui.adsabs.harvard.edu/abs/2020A&C....3000334B.

Barger A. J., Cowie L. L., Mushotzky R. F., et al. The Cosmic Evolution of Hard X-Ray-selected Active Galactic Nuclei. *Aj*, 129(2):578–609, February 2005. doi: 10.1086/426915. ADS URL: https://ui.adsabs.harvard.edu/abs/2005AJ....129..578B.

Baron Dalya. Machine Learning in Astronomy: a practical overview. *arXiv e-prints*, art. arXiv:1904.07248, April 2019. ADS URL: https://ui.adsabs.harvard.edu/abs/2019arXiv190407248B.

Baum W. A. Photoelectric Magnitudes and Red-Shifts. In McVittie George Cunliffe, editor, *Problems of Extra-Galactic Research*, volume 15, page 390, January 1962. ADS URL: https://ui.adsabs.harvard.edu/abs/1962IAUS...15..390B.

Baum William A. and Minkowski R. Observations of a Large Redshift. *Aj*, 65:483, January 1960. doi: 10.1086/108083. ADS URL: https://ui.adsabs.harvard.edu/abs/1960AJ.....65Q.483B.

Bellman Richard. *Dynamic Programming*. Princeton University Press, USA, 2010. ISBN 0691146683.

Benítez N., Gaztañaga E., Miquel R., et al. Measuring Baryon Acoustic Oscillations Along the Line of Sight with Photometric Redshifts: The PAU Survey. *ApJ*, 691 (1):241–260, January 2009a. doi: 10.1088/0004-637X/691/1/241. ADS URL: https://ui.adsabs.harvard.edu/abs/2009ApJ...691..241B.

Benítez N., Moles M., Aguerri J. A. L., et al. Optimal Filter Systems for Photometric Redshift Estimation. *ApJ*, 692(1):L5–L8, February 2009b. doi: 10.1088/0004-637X/692/1/L5. ADS URL: https://ui.adsabs.harvard.edu/abs/2009ApJ...692L...5B.

Benítez Narciso. Bayesian Photometric Redshift Estimation. *ApJ*, 536(2):571–583, June 2000. doi: 10.1086/308947. ADS URL: https://ui.adsabs.harvard.edu/abs/2000ApJ...536..571B.

Bernstein G. Metric Tests for Curvature from Weak Lensing and Baryon Acoustic Oscillations. *ApJ*, 637(2):598–607, February 2006. doi: 10.1086/498079. ADS URL: https://ui.adsabs.harvard.edu/abs/2006ApJ...637..598B.

Bernstein Gary and Huterer Dragan. Catastrophic photometric redshift errors: weak-lensing survey requirements. *MNRAS*, 401(2):1399–1408, January 2010. doi: 10.1111/j.1365-2966.2009.15748.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2010MNRAS.401.1399B.

Bertin E. and Arnouts S. SExtractor: Software for source extraction. *AAPS*, 117:393–404, June 1996. doi: 10.1051/aas:1996164. ADS URL: https://ui.adsabs.harvard.edu/abs/1996A&AS..117..393B.

Biviano , Rosati, P. , Balestra, I. , et al. Clash-vlt: The mass, velocity-anisotropy, and pseudo-phase-space density profiles of the z = 0.44 galaxy cluster macs j1206.2-0847. *A&A*, 558:A1, 2013. doi: 10.1051/0004-6361/201321955. URL: https://doi.org/10.1051/0004-6361/201321955.

Blake Chris and Bridle Sarah. Cosmology with photometric redshift surveys. *MNRAS*, 363(4):1329–1348, November 2005. doi: 10.1111/j.1365-2966.2005.09526.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2005MNRAS.363.1329B.

Bolzonella M., Miralles J. M., and Pelló R. Photometric redshifts based on standard SED fitting procedures. *A&A*, 363:476–492, November 2000. ADS URL: https://ui.adsabs.harvard.edu/abs/2000A&A...363..476B.

Bonfield D. G., Sun Y., Davey N., et al. Photometric redshift estimation using Gaussian processes. *MNRAS*, 405(2):987–994, June 2010. doi: 10.1111/j.1365-2966.2010.16544.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2010MNRAS.405..987B.

Bonnett C., Troxel M. A., Hartley W., et al. Redshift distributions of galaxies in the Dark Energy Survey Science Verification shear catalogue and implications for weak lensing. *Phys. Rev. D*, 94(4):042005, August 2016. doi: 10.1103/PhysRevD.94.042005. ADS URL: https://ui.adsabs.harvard.edu/abs/2016PhRvD..94d2005B.

Bordoloi R., Lilly S. J., and Amara A. Photo-z performance for precision cosmology. *MNRAS*, 406(2):881–895, August 2010. doi: 10.1111/j.1365-2966.2010.16765.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2010MNRAS.406..881B.

Borne Kirk, Accomazzi Alberto, Bloom Joshua, et al. Astroinformatics: A 21st Century Approach to Astronomy. In *astro2010: The Astronomy and Astrophysics Decadal Survey*, volume 2010, page P6, January 2009. ADS URL: https://ui.adsabs.harvard.edu/abs/2009astro2010P...6B.

Boulade Olivier, Charlot Xavier, Abbon P., et al. MegaCam: the new Canada-France-Hawaii Telescope wide-field imaging camera. In Iye Masanori and Moorwood Alan F. M., editors, *Instrument Design and Performance for Optical/Infrared Ground-based Telescopes*, volume 4841 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 72–81, March 2003. doi: 10.1117/12.459890. ADS URL: https://ui.adsabs.harvard.edu/abs/2003SPIE.4841...72B.

Brammer Gabriel B., van Dokkum Pieter G., and Coppi Paolo. EAZY: A Fast, Public Photometric Redshift Code. *ApJ*, 686(2):1503–1513, October 2008. doi: 10.1086/591786. ADS URL: https://ui.adsabs.harvard.edu/abs/2008ApJ...686.1503B.

Brescia M., Cavuoti S., D'Abrusco R., et al. Photometric Redshifts for Quasars in Multi-band Surveys. *ApJ*, 772:140, August 2013. doi: 10.1088/0004-637X/772/2/140. ADS URL: https://ui.adsabs.harvard.edu/abs/2013ApJ...772..140B.

Brescia M., Cavuoti S., Longo G., and De Stefano V. A catalogue of photometric redshifts for the SDSS-DR9 galaxies. *A&A*, 568:A126, Aug 2014. doi: 10.1051/0004-6361/201424383. ADS URL: https://ui.adsabs.harvard.edu/abs/2014A&A...568A.126B.

Brescia M., Cavuoti S., and Longo G. Automated physical classification in the SDSS DR10. A catalogue of candidate quasars. *MNRAS*, 450(4):3893–3903, July 2015. doi: 10.1093/mnras/stv854. ADS URL: https://ui.adsabs.harvard.edu/abs/2015MNRAS.450.3893B.

Brescia M., Salvato M., Cavuoti S., et al. Photometric redshifts for X-ray-selected active galactic nuclei in the eROSITA era. *MNRAS*, 489(1):663–680, October 2019. doi: 10.1093/mnras/stz2159. ADS URL: https://ui.adsabs.harvard.edu/abs/2019MNRAS.489..663B.

Brescia Massimo, Cavuoti Stefano, Paolillo Maurizio, et al. The detection of globular clusters in galaxies as a data mining problem. *MNRAS*, 421(2):1155–1165, April 2012. doi: 10.1111/j.1365-2966.2011.20375.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2012MNRAS.421.1155B.

Brescia Massimo, Cavuoti Stefano, Amaro Valeria, et al. Data Deluge in Astrophysics: Photometric Redshifts as a Template Use Case. *arXiv e-prints*, art.

arXiv:1802.07683, February 2018. ADS URL: https://ui.adsabs.harvard.edu/abs/2018arXiv180207683B.

Brink Henrik, Richards Joseph W., Poznanski Dovi, et al. Using machine learning for discovery in synoptic survey imaging data. *MNRAS*, 435(2):1047–1060, October 2013. doi: 10.1093/mnras/stt1306. ADS URL: https://ui.adsabs.harvard.edu/abs/2013MNRAS.435.1047B.

Brodwin M., Brown M. J. I., Ashby M. L. N., et al. Photometric Redshifts in the IRAC Shallow Survey. *ApJ*, 651(2):791–803, November 2006. doi: 10.1086/507838. ADS URL: https://ui.adsabs.harvard.edu/abs/2006ApJ...651..791B.

Brown Tom B., Mann Benjamin, Ryder Nick, et al. Language Models are Few-Shot Learners. *arXiv e-prints*, art. arXiv:2005.14165, May 2020. ADS URL: https://ui.adsabs.harvard.edu/abs/2020arXiv200514165B.

Brunner Robert J., Connolly Andrew J., Szalay Alexander S., and Bershady Matthew A. Toward More Precise Photometric Redshifts: Calibration Via CCD Photometry. *ApJ*, 482(1):L21–L24, June 1997. doi: 10.1086/310674. ADS URL: https://ui.adsabs.harvard.edu/abs/1997ApJ...482L..21B.

Brunner Robert J., Szalay Alex S., and Connolly Andrew J. Evolution in the Clustering of Galaxies for Z<1.0. *ApJ*, 541(2):527–534, October 2000. doi: 10.1086/309488. ADS URL: https://ui.adsabs.harvard.edu/abs/2000ApJ...541..527B.

Bruzual G. and Charlot S. Stellar population synthesis at the resolution of 2003. *MNRAS*, 344(4):1000–1028, October 2003. doi: 10.1046/j.1365-8711.2003.06897.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2003MNRAS.344.1000B.

Buchner Johannes, Georgakakis Antonis, Nandra Kirpal, et al. Obscuration-dependent Evolution of Active Galactic Nuclei. *ApJ*, 802(2):89, April 2015. doi: 10.1088/0004-637X/802/2/89. ADS URL: https://ui.adsabs.harvard.edu/abs/2015ApJ...802...89B.

Buchs R., Davis C., Gruen D., et al. Phenotypic redshifts with self-organizing maps: A novel method to characterize redshift distributions of source galaxies for weak lensing. *MNRAS*, 489(1):820–841, October 2019. doi: 10.1093/mnras/stz2162. ADS URL: https://ui.adsabs.harvard.edu/abs/2019MNRAS.489..820B.

Budavári Tamás, Csabai István, Szalay Alexander S., et al. Photometric Redshifts from Reconstructed Quasar Templates. *Aj*, 122(3):1163–1171, September 2001. doi: 10.1086/322131. ADS URL: https://ui.adsabs.harvard.edu/abs/2001AJ....122.1163B.

Budavári Tamás, Szalay Alex S., Charlot Stéphane, et al. The Ultraviolet Luminosity Function of GALEX Galaxies at Photometric Redshifts between 0.07 and 0.25. *ApJ*, 619(1):L31–L34, January 2005. doi: 10.1086/423319. ADS URL: https://ui.adsabs.harvard.edu/abs/2005ApJ...619L..31B.

Bunn Emory F. and Hogg David W. The kinematic origin of the cosmological redshift. *American Journal of Physics*, 77(8):688–694, August 2009. doi: 10.1119/1.3129103. ADS URL: https://ui.adsabs.harvard.edu/abs/2009AmJPh..77..688B.

Burges Christopher J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, Jun 1998. ISSN 1573-756X. doi: 10.1023/A:1009715923555.

Butchins S. A. Predicted redshifts of galaxies by broadband photometry. *A&A*, 97(2): 407–409, April 1981. ADS URL: https://ui.adsabs.harvard.edu/abs/1981A&A....97..407B.

Cabayol L., Sevilla-Noarbe I., Fernández E., et al. The PAU survey: star-galaxy classification with multi narrow-band data. *MNRAS*, 483(1):529–539, February 2019. doi: 10.1093/mnras/sty3129. ADS URL: https://ui.adsabs.harvard.edu/abs/2019MNRAS.483..529C.

Capak P., Aussel H., Ajiki M., et al. The First Release COSMOS Optical and Near-IR Data and Catalog. *APJS*, 172(1):99–116, September 2007. doi: 10.1086/519081. ADS URL: https://ui.adsabs.harvard.edu/abs/2007ApJS..172...99C.

Carliles S., Budavári T., Heinis S., et al. Photometric Redshift Estimation on SDSS Data Using Random Forests. In Argyle R. W., Bunclark P. S., and Lewis J. R., editors, *Astronomical Data Analysis Software and Systems XVII*, volume 394 of *Astronomical Society of the Pacific Conference Series*, page 521, August 2008. ADS URL: https://ui.adsabs.harvard.edu/abs/2008ASPC..394..521C.

Carliles Samuel, Budavári Tamás, Heinis Sébastien, et al. Random Forests for Photometric Redshifts. *ApJ*, 712(1):511–515, March 2010. doi: 10.1088/0004-637X/712/1/511. ADS URL: https://ui.adsabs.harvard.edu/abs/2010ApJ...712..511C.

Carrasco Kind Matias and Brunner Robert J. TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests. *MNRAS*, 432(2): 1483–1501, June 2013. doi: 10.1093/mnras/stt574. ADS URL: https://ui.adsabs.harvard.edu/abs/2013MNRAS.432.1483C.

Carrasco Kind Matias and Brunner Robert J. Exhausting the information: novel Bayesian combination of photometric redshift PDFs. *MNRAS*, 442(4):3380–3399, August 2014a. doi: 10.1093/mnras/stu1098. ADS URL: https://ui.adsabs.harvard.edu/abs/2014MNRAS.442.3380C.

Carrasco Kind Matias and Brunner Robert J. SOMz: photometric redshift PDFs with self-organizing maps and random atlas. *MNRAS*, 438(4):3409–3421, Mar 2014b. doi: 10.1093/mnras/stt2456. ADS URL: https://ui.adsabs.harvard.edu/abs/2014MNRAS.438.3409C.

Cavuoti S., Brescia M., Longo G., and Mercurio A. Photometric redshifts with the quasi Newton algorithm (MLPQNA) Results in the PHAT1 contest. *A&A*, 546:A13, Oct 2012. doi: 10.1051/0004-6361/201219755. ADS URL: https://ui.adsabs.harvard.edu/abs/2012A&A...546A..13C.

Cavuoti S., Brescia M., De Stefano V., and Longo G. Photometric redshift estimation based on data mining with PhotoRApToR. *Experimental Astronomy*, 39(1):45–71, March 2015a. doi: 10.1007/s10686-015-9443-4. ADS URL: https://ui.adsabs.harvard.edu/abs/2015ExA....39...45C.

Cavuoti S., Brescia M., Tortora C., et al. Machine-learning-based photometric redshifts for galaxies of the ESO Kilo-Degree Survey data release 2. *MNRAS*, 452(3):3100–3105, September 2015b. doi: 10.1093/mnras/stv1496. ADS URL: https://ui.adsabs.harvard.edu/abs/2015MNRAS.452.3100C.

Cavuoti S., Amaro V., Brescia M., et al. METAPHOR: a machine-learning-based method for the probability density estimation of photometric redshifts. *MNRAS*, 465(2):1959–1973, Feb 2017a. doi: 10.1093/mnras/stw2930. ADS URL: https://ui.adsabs.harvard.edu/abs/2017MNRAS.465.1959C.

Cavuoti S., Tortora C., Brescia M., et al. A cooperative approach among methods for photometric redshifts estimation: an application to KiDS data. *MNRAS*, 466(2): 2039–2053, Apr 2017b. doi: 10.1093/mnras/stw3208. ADS URL: https://ui.adsabs.harvard.edu/abs/2017MNRAS.466.2039C.

Chaves-Montero Jonás, Angulo Raúl E., and Hernández-Monteagudo Carlos. The effect of photometric redshift uncertainties on galaxy clustering and baryonic acoustic oscillations. *MNRAS*, 477(3):3892–3909, July 2018. doi: 10.1093/mnras/sty924. ADS URL: https://ui.adsabs.harvard.edu/abs/2018MNRAS.477.3892C.

Civano F., Elvis M., Brusa M., et al. The Chandra COSMOS Survey. III. Optical and Infrared Identification of X-Ray Point Sources. *APJS*, 201:30, August 2012.

doi: 10.1088/0067-0049/201/2/30. ADS URL: https://ui.adsabs.harvard.edu/abs/2012ApJS..201...30C.

Coe Dan, Benítez Narciso, Sánchez Sebastián F., et al. Galaxies in the Hubble Ultra Deep Field. I. Detection, Multiband Photometry, Photometric Redshifts, and Morphology. *Aj*, 132(2):926–959, August 2006. doi: 10.1086/505530. ADS URL: https://ui.adsabs.harvard.edu/abs/2006AJ....132..926C.

Cohen Judith G., Hogg David W., Blandford Roger, et al. Caltech Faint Galaxy Redshift Survey. X. A Redshift Survey in the Region of the Hubble Deep Field North. *ApJ*, 538 (1):29–52, July 2000. doi: 10.1086/309096. ADS URL: https://ui.adsabs.harvard.edu/abs/2000ApJ...538...29C.

Collister Adrian A. and Lahav Ofer. ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. *PASP*, 116(818):345–351, April 2004. doi: 10.1086/383254. ADS URL: https://ui.adsabs.harvard.edu/abs/2004PASP..116..345C.

Connolly A. J., Csabai I., Szalay A. S., et al. Slicing Through Multicolor Space: Galaxy Redshifts from Broadband Photometry. *Aj*, 110:2655, December 1995. doi: 10.1086/117720. ADS URL: https://ui.adsabs.harvard.edu/abs/1995AJ....110.2655C.

Conselice Christopher J., Blackburne Jeffrey A., and Papovich Casey. The Luminosity, Stellar Mass, and Number Density Evolution of Field Galaxies of Known Morphology from z = 0.5 to 3. *ApJ*, 620(2):564–583, February 2005. doi: 10.1086/426102. ADS URL: https://ui.adsabs.harvard.edu/abs/2005ApJ...620..564C.

Cortes Corinna and Vapnik Vladimir. Support-vector networks. *Mach. Learn.*, 20(3): 273–297, Sep 1995. ISSN 1573-0565. doi: 10.1007/BF00994018.

Csabai I., Connolly A. J., Szalay A. S., and Budavári T. Reconstructing Galaxy Spectral Energy Distributions from Broadband Photometry. *Aj*, 119(1):69–78, January 2000. doi: 10.1086/301159. ADS URL: https://ui.adsabs.harvard.edu/abs/2000AJ....119...69C.

Csabai István, Budavári Tamás, Connolly Andrew J., et al. The Application of Photometric Redshifts to the SDSS Early Data Release. *Aj*, 125(2):580–592, February 2003. doi: 10.1086/345883. ADS URL: https://ui.adsabs.harvard.edu/abs/2003AJ....125..580C.

Cunha Carlos E., Huterer Dragan, Busha Michael T., and Wechsler Risa H. Sample variance in photometric redshift calibration: cosmological biases and survey requirements. *MNRAS*, 423(1):909–924, June 2012. doi: 10.1111/j.1365-2966.2012.20927.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2012MNRAS.423..909C.

Cunha Carlos E., Huterer Dragan, Lin Huan, et al. Spectroscopic failures in photometric redshift calibration: cosmological biases and survey requirements. *MNRAS*, 444(1):129–146, October 2014. doi: 10.1093/mnras/stu1424. ADS URL: https://ui.adsabs.harvard.edu/abs/2014MNRAS.444..129C.

Curran S. J. QSO photometric redshifts from SDSS, WISE, and GALEX colours. *MNRAS*, 493(1):L70–L75, March 2020. doi: 10.1093/mnrasl/slaa012. ADS URL: https://ui.adsabs.harvard.edu/abs/2020MNRAS.493L..70C.

Dahlen Tomas, Mobasher Bahram, Faber Sandra M., et al. A Critical Assessment of Photometric Redshift Methods: A CANDELS Investigation. *ApJ*, 775(2):93, October 2013. doi: 10.1088/0004-637X/775/2/93. ADS URL: https://ui.adsabs.harvard.edu/abs/2013ApJ...775...93D.

Darvish Behnam, Mobasher Bahram, Martin D. Christopher, et al. Cosmic Web of Galaxies in the COSMOS Field: Public Catalog and Different Quenching for Centrals and Satellites. *ApJ*, 837(1):16, March 2017. doi: 10.3847/1538-4357/837/1/16. ADS URL: https://ui.adsabs.harvard.edu/abs/2017ApJ...837...16D.

Davis C., Gatti M., Vielzeuf P., et al. Dark Energy Survey Year 1 Results: Cross-Correlation Redshifts in the DES – Calibration of the Weak Lensing Source Redshift Distributions. *arXiv e-prints*, art. arXiv:1710.02517, October 2017. ADS URL: https://ui.adsabs.harvard.edu/abs/2017arXiv171002517D.

Davis M., Guhathakurta P., Konidaris N. P., et al. The All-Wavelength Extended Groth Strip International Survey (AEGIS) Data Sets. *ApJ*, 660(1):L1–L6, May 2007. doi: 10.1086/517931. ADS URL: https://ui.adsabs.harvard.edu/abs/2007ApJ...660L...1D.

de Freitas Nando. Machine learning - Introduction to Gaussian processes, Feb 2013. URL: https://www.youtube.com/watch?v=4vGiHC35j9s&feature=youtu.be&ab_channel=NandodeFreitas. [Online; accessed 28. Dec. 2020].

de la Calleja Jorge and Fuentes Olac. Machine learning and image analysis for morphological galaxy classification. *MNRAS*, 349(1):87–93, March 2004. doi: 10.1111/j.1365-2966.2004.07442.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2004MNRAS.349...87D.

de Silva Brian M., Higdon David M., Brunton Steven L., and Kutz J. Nathan. Discovery of Physics from Data: Universal Laws and Discrepancies. *arXiv e-prints*, art. arXiv:1906.07906, June 2019. ADS URL: https://ui.adsabs.harvard.edu/abs/2019arXiv190607906D.

de Vaucouleurs G., de Vaucouleurs A., and Corwin J. R. Second reference catalogue of bright galaxies. *Second reference catalogue of bright galaxies*, 1976:0, January 1976. ADS URL: https://ui.adsabs.harvard.edu/abs/1976RC2...C......0D.

de Vaucouleurs Gerard Henri, de Vaucouleurs Antoinette, and Shapley Harlow. *Reference catalogue of bright galaxies*. University of Texas Press, 1964. ADS URL: https://ui.adsabs.harvard.edu/abs/1964rcbg.book.....D.

Dean Jeffrey and Ghemawat Sanjay. Mapreduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150, San Francisco, CA, 2004.

Dedehayir Ozgur and Steinert Martin. The hype cycle model: A review and future directions. *Technological Forecasting and Social Change*, 108, 04 2016. doi: 10.1016/j.techfore.2016.04.005.

Delli Veneri M., Cavuoti S., Brescia M., et al. Star formation rates for photometric samples of galaxies using machine learning methods. *MNRAS*, 486(1):1377–1391, June 2019. doi: 10.1093/mnras/stz856. ADS URL: https://ui.adsabs.harvard.edu/abs/2019MNRAS.486.1377D.

DESI Collaboration , Aghamousa Amir, Aguilar Jessica, et al. The DESI Experiment Part I: Science,Targeting, and Survey Design. *arXiv e-prints*, art. arXiv:1611.00036, October 2016a. ADS URL: https://ui.adsabs.harvard.edu/abs/2016arXiv161100036D.

DESI Collaboration , Aghamousa Amir, Aguilar Jessica, et al. The DESI Experiment Part I: Science,Targeting, and Survey Design. *arXiv e-prints*, art. arXiv:1611.00036, October 2016b. ADS URL: https://ui.adsabs.harvard.edu/abs/2016arXiv161100036D.

Diamond Jared M. *Guns, Germs, and Steel: the Fates of Human Societies*. W. W. Norton & Co., New York, 1998. ISBN 0393038912 9780393038910 0393317552 9780393317558.

D'Isanto A. and Polsterer K. L. Photometric redshift estimation via deep learning. Generalized and pre-classification-less, image based, fully probabilistic redshifts. *A&A*, 609:A111, January 2018. doi: 10.1051/0004-6361/201731326. ADS URL: https://ui.adsabs.harvard.edu/abs/2018A&A...609A.111D.

D'Isanto A., Cavuoti S., Gieseke F., and Polsterer K. L. Return of the features. Efficient feature selection and interpretation for photometric redshifts. *A&A*, 616:A97, August 2018. doi: 10.1051/0004-6361/201833103. ADS URL: https://ui.adsabs.harvard.edu/abs/2018A&A...616A..97D.

Donalek Ciro, Arun Kumar A., Djorgovski S. G., et al. Feature Selection Strategies for Classifying High Dimensional Astronomical Data Sets. *arXiv e-prints*, art. arXiv:1310.1976, October 2013. ADS URL: https://ui.adsabs.harvard.edu/abs/2013arXiv1310.1976D.

du Buisson L., Sivanandam N., Bassett Bruce A., and Smith M. Machine learning classification of SDSS transient survey images. *MNRAS*, 454(2):2026–2038, December 2015. doi: 10.1093/mnras/stv2041. ADS URL: https://ui.adsabs.harvard.edu/abs/2015MNRAS.454.2026D.

Duncan Kenneth J., Jarvis Matt J., Brown Michael J. I., and Röttgering Huub J. A. Photometric redshifts for the next generation of deep radio continuum surveys - II. Gaussian processes and hybrid estimates. *MNRAS*, 477(4):5177–5190, July 2018. doi: 10.1093/mnras/sty940. ADS URL: https://ui.adsabs.harvard.edu/abs/2018MNRAS.477.5177D.

Eisenhardt P. R., Stern D., Brodwin M., et al. The Infrared Array Camera (IRAC) Shallow Survey. *APJS*, 154(1):48–53, September 2004. doi: 10.1086/423180. ADS URL: https://ui.adsabs.harvard.edu/abs/2004ApJS..154...48E.

Eriksen M., Alarcon A., Gaztanaga E., et al. The PAU Survey: early demonstration of photometric redshift performance in the COSMOS field. *MNRAS*, 484(3):4200–4215, April 2019. doi: 10.1093/mnras/stz204. ADS URL: https://ui.adsabs.harvard.edu/abs/2019MNRAS.484.4200E.

Eriksen M., Alarcon A., Cabayol L., et al. The PAU Survey: Photometric redshifts using transfer learning from simulations. *MNRAS*, 497(4):4565–4579, August 2020. doi: 10.1093/mnras/staa2265. ADS URL: https://ui.adsabs.harvard.edu/abs/2020MNRAS.497.4565E.

Euclid Collaboration , Desprez G., Paltani S., et al. Euclid preparation. X. The Euclid photometric-redshift challenge. *A&A*, 644:A31, December 2020. doi: 10.1051/0004-6361/202039403. ADS URL: https://ui.adsabs.harvard.edu/abs/2020A&A...644A..31E.

Feldmann R., Carollo C. M., Porciani C., et al. The Zurich Extragalactic Bayesian Redshift Analyzer and its first application: COSMOS. *MNRAS*, 372(2):565–577, October 2006. doi: 10.1111/j.1365-2966.2006.10930.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2006MNRAS.372..565F.

Fernández-Cabán Pedro, Masters Forrest, and Phillips Brian. Predicting roof pressures on a low-rise structure from freestream turbulence using artificial neural networks. *Frontiers in Built Environment*, 4, 11 2018. doi: 10.3389/fbuil.2018.00068.

Finkelstein Steven L., Ryan Jr., Russell E., Papovich Casey, et al. The Evolution of the Galaxy Rest-frame Ultraviolet Luminosity Function over the First Two Billion Years. *ApJ*, 810(1):71, September 2015. doi: 10.1088/0004-637X/810/1/71. ADS URL: https://ui.adsabs.harvard.edu/abs/2015ApJ...810...71F.

Firth Andrew E., Lahav Ofer, and Somerville Rachel S. Estimating photometric redshifts with artificial neural networks. *MNRAS*, 339(4):1195–1202, March 2003. doi: 10.1046/j.1365-8711.2003.06271.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2003MNRAS.339.1195F.

Furusawa Hisanori, Shimasaku Kazuhiro, Doi Mamoru, and Okamura Sadanori. New Improved Photometric Redshifts of Galaxies in the Hubble Deep Field. *ApJ*, 534(2):624–635, May 2000. doi: 10.1086/308794. ADS URL: https://ui.adsabs.harvard.edu/abs/2000ApJ...534..624F.

Gaia Collaboration , Prusti T., de Bruijne J. H. J., et al. The Gaia mission. *A&A*, 595:A1, November 2016. doi: 10.1051/0004-6361/201629272. ADS URL: https://ui.adsabs.harvard.edu/abs/2016A&A...595A...1G.

Galametz Audrey, Saglia Roberto, Paltani Stéphane, et al. SED-dependent galactic extinction prescription for Euclid and future cosmological surveys. *A&A*, 598:A20, February 2017. doi: 10.1051/0004-6361/201629333. ADS URL: https://ui.adsabs.harvard.edu/abs/2017A&A...598A..20G.

Geach James E. Unsupervised self-organized mapping: a versatile empirical tool for object selection, classification and redshift estimation in large surveys. *MNRAS*, 419(3):2633–2645, Jan 2012. doi: 10.1111/j.1365-2966.2011.19913.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2012MNRAS.419.2633G.

George Daniel, Shen Hongyu, and Huerta E. A. Classification and unsupervised clustering of LIGO data with Deep Transfer Learning. *Phys. Rev. D*, 97(10):101501, May 2018. doi: 10.1103/PhysRevD.97.101501. ADS URL: https://ui.adsabs.harvard.edu/abs/2018PhRvD..97j1501G.

Gerdes David W., Sypniewski Adam J., McKay Timothy A., et al. ArborZ: Photometric Redshifts Using Boosted Decision Trees. *ApJ*, 715(2):823–832, June 2010. doi: 10.1088/0004-637X/715/2/823. ADS URL: https://ui.adsabs.harvard.edu/abs/2010ApJ...715..823G.

Giavalisco M., Ferguson H. C., Koekemoer A. M., et al. The Great Observatories Origins Deep Survey: Initial Results from Optical and Near-Infrared Imaging. *ApJ*, 600(2):L93–L98, January 2004. doi: 10.1086/379232. ADS URL: https://ui.adsabs.harvard.edu/abs/2004ApJ...600L..93G.

Giovanelli Riccardo and Haynes Martha P. Redshift surveys of galaxies. *ARA&A*, 29: 499–541, January 1991. doi: 10.1146/annurev.aa.29.090191.002435. ADS URL: https://ui.adsabs.harvard.edu/abs/1991ARA&A..29..499G.

Goebel J., Volk K., Walker H., et al. A Bayesian classification of the IRAS LRS atlas. *A&A*, 222:L5–L8, September 1989. ADS URL: https://ui.adsabs.harvard.edu/abs/1989A&A...222L...5G.

Goldstein D. A., D'Andrea C. B., Fischer J. A., et al. Automated Transient Identification in the Dark Energy Survey. *Aj*, 150(3):82, September 2015. doi: 10.1088/0004-6256/150/3/82. ADS URL: https://ui.adsabs.harvard.edu/abs/2015AJ....150...82G.

Goodfellow I., Bengio Y., and Courville A. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016. ISBN 9780262035613. URL: https://books.google.co.in/books?id=Np9SDQAAQBAJ.

Guo Hua-Dong, Zhang Li, and Zhu Lan-Wei. Earth observation big data for climate change research. *Advances in Climate Change Research*, 6(2):108–117, 2015. ISSN 1674-9278. doi: https://doi.org/10.1016/j.accre.2015.09.007. URL: http://www.sciencedirect.com/science/article/pii/S1674927815000519. Special issue on advances in Future Earth research.

Guyon Isabelle and Elisseeff André. An introduction of variable and feature selection. *J. Machine Learning Research Special Issue on Variable and Feature Selection*, 3:1157–1182, 01 2003. doi: 10.1162/153244303322753616.

Guzzo L., Cassata P., Finoguenov A., et al. The Cosmic Evolution Survey (COSMOS): A Large-Scale Structure at z=0.73 and the Relation of Galaxy Morphologies to Local Environment. *APJS*, 172(1):254–269, September 2007. doi: 10.1086/516588. ADS URL: https://ui.adsabs.harvard.edu/abs/2007ApJS..172..254G.

Hainaut Olivier R. and Williams Andrew P. Impact of satellite constellations on astronomical observations with ESO telescopes in the visible and infrared domains. *A&A*, 636:A121, April 2020. doi: 10.1051/0004-6361/202037501. ADS URL: https://ui.adsabs.harvard.edu/abs/2020A&A...636A.121H.

Hamana Takashi, Shirasaki Masato, Miyazaki Satoshi, et al. Cosmological constraints from cosmic shear two-point correlation functions with HSC survey first-year data. *PASJ*, 72(1):16, February 2020. doi: 10.1093/pasj/psz138. ADS URL: https://ui.adsabs.harvard.edu/abs/2020PASJ...72...16H.

Han Bo, Ding Hong-Peng, Zhang Yan-Xia, and Zhao Yong-Heng. Photometric redshift estimation for quasars by integration of KNN and SVM. *Research in Astronomy and*

Astrophysics, 16(5):74, May 2016. doi: 10.1088/1674-4527/16/5/074. ADS URL: https://ui.adsabs.harvard.edu/abs/2016RAA....16...74H.

Hartley W. G., Chang C., Samani S., et al. The impact of spectroscopic incompleteness in direct calibration of redshift distributions for weak lensing surveys. *MNRAS*, 496 (4):4769–4786, June 2020. doi: 10.1093/mnras/staa1812. ADS URL: https://ui.adsabs.harvard.edu/abs/2020MNRAS.496.4769H.

Hasinger G., Capak P., Salvato M., et al. The DEIMOS 10K Spectroscopic Survey Catalog of the COSMOS Field. *ApJ*, 858:77, May 2018. doi: 10.3847/1538-4357/aabacf. ADS URL: https://ui.adsabs.harvard.edu/abs/2018ApJ...858...77H.

Hearin Andrew P., Zentner Andrew R., Ma Zhaoming, and Huterer Dragan. A General Study of the Influence of Catastrophic Photometric Redshift Errors on Cosmology with Cosmic Shear Tomography. *ApJ*, 720(2):1351–1369, September 2010. doi: 10.1088/0004-637X/720/2/1351. ADS URL: https://ui.adsabs.harvard.edu/abs/2010ApJ...720.1351H.

Heinis S., Kumar S., Gezari S., et al. Of Genes and Machines: Application of a Combination of Machine Learning Tools to Astronomy Data Sets. *ApJ*, 821(2):86, April 2016. doi: 10.3847/0004-637X/821/2/86. ADS URL: https://ui.adsabs.harvard.edu/abs/2016ApJ...821...86H.

Hernandez-Pajares M., Comellas F., Monte E., and Floris J. Classifying Stars: A Comparison between Classical, Genetic and Neural Network Algorithms. In *European Southern Observatory Conference and Workshop Proceedings*, volume 43 of *European Southern Observatory Conference and Workshop Proceedings*, page 325, September 1992. ADS URL: https://ui.adsabs.harvard.edu/abs/1992ESOC...43..325H.

Heymans Catherine, Van Waerbeke Ludovic, Miller Lance, et al. CFHTLenS: the Canada-France-Hawaii Telescope Lensing Survey. *MNRAS*, 427(1):146–166, November 2012. doi: 10.1111/j.1365-2966.2012.21952.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2012MNRAS.427..146H.

Hikage Chiaki, Oguri Masamune, Hamana Takashi, et al. Cosmology from cosmic shear power spectra with Subaru Hyper Suprime-Cam first-year data. *PASJ*, 71(2):43, April 2019. doi: 10.1093/pasj/psz010. ADS URL: https://ui.adsabs.harvard.edu/abs/2019PASJ...71...43H.

Hildebrandt H., Wolf C., and Benítez N. A blind test of photometric redshifts on ground-based data. *A&A*, 480(3):703–714, March 2008. doi: 10.1051/0004-6361:20077107. ADS URL: https://ui.adsabs.harvard.edu/abs/2008A&A...480..703H.

Hildebrandt H., Pielorz J., Erben T., et al. CARS: the CFHTLS-Archive-Research Survey. II. Weighing dark matter halos of Lyman-break galaxies at z = 3-5. *A&A*, 498(3):725–736, May 2009. doi: 10.1051/0004-6361/200811042. ADS URL: https://ui.adsabs.harvard.edu/abs/2009A&A...498..725H.

Hildebrandt H., Arnouts S., Capak P., et al. PHAT: PHoto-z Accuracy Testing. *A&A*, 523:A31, November 2010. doi: 10.1051/0004-6361/201014885. ADS URL: https://ui.adsabs.harvard.edu/abs/2010A&A...523A..31H.

Hildebrandt H., Viola M., Heymans C., et al. KiDS-450: cosmological parameter constraints from tomographic weak gravitational lensing. *MNRAS*, 465(2):1454–1498, February 2017. doi: 10.1093/mnras/stw2805. ADS URL: https://ui.adsabs.harvard.edu/abs/2017MNRAS.465.1454H.

Hildebrandt H., Köhlinger F., van den Busch J. L., et al. KiDS+VIKING-450: Cosmic shear tomography with optical and infrared data. *A&A*, 633:A69, January 2020a. doi: 10.1051/0004-6361/201834878. ADS URL: https://ui.adsabs.harvard.edu/abs/2020A&A...633A..69H.

Hildebrandt H., van den Busch J. L., Wright A. H., et al. KiDS-1000 catalogue: Redshift distributions and their calibration. *arXiv e-prints*, art. arXiv:2007.15635, July 2020b. ADS URL: https://ui.adsabs.harvard.edu/abs/2020arXiv200715635H.

Hoekstra Henk, Yee Howard K. C., and Gladders Michael D. Constraints on Ω_m and σ_8 from Weak Lensing in Red-Sequence Cluster Survey Fields. *ApJ*, 577(2):595–603, October 2002. doi: 10.1086/342120. ADS URL: https://ui.adsabs.harvard.edu/abs/2002ApJ...577..595H.

Hoekstra Henk, Bartelmann Matthias, Dahle Håkon, et al. Masses of Galaxy Clusters from Gravitational Lensing. *Space Sci. Rev.*, 177(1-4):75–118, August 2013. doi: 10.1007/s11214-013-9978-5. ADS URL: https://ui.adsabs.harvard.edu/abs/2013SSRv..177...75H.

Hogarth Nathan Benaich and Ian. State of AI Report 2020, 2020. URL: https://www.stateof.ai/.

Hogg David W., Cohen Judith G., Blandford Roger, et al. A Blind Test of Photometric Redshift Prediction. *Aj*, 115(4):1418–1422, April 1998. doi: 10.1086/300277. ADS URL: https://ui.adsabs.harvard.edu/abs/1998AJ....115.1418H.

Hoyle B. Measuring photometric redshifts using galaxy images and Deep Neural Networks. *Astronomy and Computing*, 16:34–40, July 2016. doi: 10.1016/j.ascom.2016.03.006. ADS URL: https://ui.adsabs.harvard.edu/abs/2016A&C....16...34H.

Hoyle B., Gruen D., Bernstein G. M., et al. Dark Energy Survey Year 1 Results: redshift distributions of the weak-lensing source galaxies. *MNRAS*, 478(1):592–610, July 2018. doi: 10.1093/mnras/sty957. ADS URL: https://ui.adsabs.harvard.edu/abs/2018MNRAS.478..592H.

Hoyle Ben, Rau Markus Michael, Zitlau Roman, et al. Feature importance for machine learning redshifts applied to SDSS galaxies. *MNRAS*, 449(2):1275–1283, May 2015. doi: 10.1093/mnras/stv373. ADS URL: https://ui.adsabs.harvard.edu/abs/2015MNRAS.449.1275H.

Huertas-Company M., Rouan D., Tasca L., et al. A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images. I. Method description. *A&A*, 478(3):971–980, February 2008. doi: 10.1051/0004-6361:20078625. ADS URL: https://ui.adsabs.harvard.edu/abs/2008A&A...478..971H.

Huertas-Company M., Gravet R., Cabrera-Vives G., et al. A Catalog of Visual-like Morphologies in the 5 CANDELS Fields Using Deep Learning. *APJS*, 221(1):8, November 2015. doi: 10.1088/0067-0049/221/1/8. ADS URL: https://ui.adsabs.harvard.edu/abs/2015ApJS..221....8H.

Huertas-Company M., Primack J. R., Dekel A., et al. Deep Learning Identifies High-z Galaxies in a Central Blue Nugget Phase in a Characteristic Mass Range. *ApJ*, 858 (2):114, May 2018. doi: 10.3847/1538-4357/aabfed. ADS URL: https://ui.adsabs.harvard.edu/abs/2018ApJ...858..114H.

Humason M. L., Mayall N. U., and Sandage A. R. Redshifts and magnitudes of extragalactic nebulae. *Aj*, 61:97–162, January 1956. doi: 10.1086/107297. ADS URL: https://ui.adsabs.harvard.edu/abs/1956AJ.....61...97H.

Ilbert O., Arnouts S., McCracken H. J., et al. Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey. *A&A*, 457(3): 841–856, October 2006. doi: 10.1051/0004-6361:20065138. ADS URL: https://ui.adsabs.harvard.edu/abs/2006A&A...457..841I.

Ilbert O., Capak P., Salvato M., et al. Cosmos Photometric Redshifts with 30-Bands for 2-deg^2. *ApJ*, 690(2):1236–1249, January 2009. doi: 10.1088/0004-637X/690/2/1236. ADS URL: https://ui.adsabs.harvard.edu/abs/2009ApJ...690.1236I.

Ilbert O., Salvato M., Le Floc'h E., et al. Galaxy Stellar Mass Assembly Between 0.2 $<$ z $<$ 2 from the S-COSMOS Survey. *ApJ*, 709(2):644–663, February 2010. doi: 10.1088/0004-637X/709/2/644. ADS URL: https://ui.adsabs.harvard.edu/abs/2010ApJ...709..644I.

Ilbert O., McCracken H. J., Le Fèvre O., et al. Mass assembly in quiescent and star-forming galaxies since z = 4 from UltraVISTA. *A&A*, 556:A55, August 2013. doi: 10.1051/0004-6361/201321100. ADS URL: https://ui.adsabs.harvard.edu/abs/2013A&A...556A..55I.

Iten Raban, Metger Tony, Wilming Henrik, et al. Discovering Physical Concepts with Neural Networks. *Phys. Rev. D*, 124(1):010508, January 2020. doi: 10.1103/PhysRevLett.124.010508. ADS URL: https://ui.adsabs.harvard.edu/abs/2020PhRvL.124a0508I.

Ivezić Željko, Kahn Steven M., Tyson J. Anthony, et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *ApJ*, 873(2):111, March 2019. doi: 10.3847/1538-4357/ab042c. ADS URL: https://ui.adsabs.harvard.edu/abs/2019ApJ...873..111I.

Jacobs C., Glazebrook K., Collett T., et al. Finding strong lenses in CFHTLS using convolutional neural networks. *MNRAS*, 471(1):167–181, October 2017. doi: 10.1093/mnras/stx1492. ADS URL: https://ui.adsabs.harvard.edu/abs/2017MNRAS.471..167J.

Jarrett T. H., Chester T., Cutri R., et al. 2MASS Extended Source Catalog: Overview and Algorithms. *Aj*, 119(5):2498–2531, May 2000. doi: 10.1086/301330. ADS URL: https://ui.adsabs.harvard.edu/abs/2000AJ....119.2498J.

Jarrett Thomas. Large Scale Structure in the Local Universe - The 2MASS Galaxy Catalog. *PASA*, 21(4):396–403, January 2004. doi: 10.1071/AS04050. ADS URL: https://ui.adsabs.harvard.edu/abs/2004PASA...21..396J.

Jarvis M., Sheldon E., Zuntz J., et al. The DES Science Verification weak lensing shear catalogues. *MNRAS*, 460(2):2245–2281, August 2016. doi: 10.1093/mnras/stw990. ADS URL: https://ui.adsabs.harvard.edu/abs/2016MNRAS.460.2245J.

Johnston S., Taylor R., Bailes M., et al. Science with ASKAP. The Australian square-kilometre-array pathfinder. *Experimental Astronomy*, 22(3):151–273, December 2008. doi: 10.1007/s10686-008-9124-7. ADS URL: https://ui.adsabs.harvard.edu/abs/2008ExA....22..151J.

Jones E. and Singal J. Analysis of a custom support vector machine for photometric redshift estimation and the inclusion of galaxy shape information. *A&A*, 600:A113, April 2017. doi: 10.1051/0004-6361/201629558. ADS URL: https://ui.adsabs.harvard.edu/abs/2017A&A...600A.113J.

Karim A., Schinnerer E., Martínez-Sansigre A., et al. The Star Formation History of Mass-selected Galaxies in the COSMOS Field. *ApJ*, 730(2):61, April 2011. doi:

10.1088/0004-637X/730/2/61. ADS URL: https://ui.adsabs.harvard.edu/abs/2011ApJ...730...61K.

Kawaharada Madoka, Okabe Nobuhiro, Umetsu Keiichi, et al. Suzaku Observation of A1689: Anisotropic Temperature and Entropy Distributions Associated with the Large-scale Structure. *ApJ*, 714(1):423–441, May 2010. doi: 10.1088/0004-637X/714/1/423. ADS URL: https://ui.adsabs.harvard.edu/abs/2010ApJ...714..423K.

Koester B. P., McKay T. A., Annis J., et al. A MaxBCG Catalog of 13,823 Galaxy Clusters from the Sloan Digital Sky Survey. *ApJ*, 660(1):239–255, May 2007. doi: 10.1086/509599. ADS URL: https://ui.adsabs.harvard.edu/abs/2007ApJ...660..239K.

Kohonen Teuvo. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, Jan 1982. ISSN 1432-0770. doi: 10.1007/BF00337288. URL: https://doi.org/10.1007/BF00337288.

Kohonen Teuvo. Essentials of the self-organizing map. *Neural Netw.*, 37:52–65, January 2013. ISSN 0893-6080. doi: 10.1016/j.neunet.2012.09.018.

Koo D. C. Optical multicolors : a poor person's Z machine for galaxies. *Aj*, 90:418–440, March 1985. doi: 10.1086/113748. ADS URL: https://ui.adsabs.harvard.edu/abs/1985AJ.....90..418K.

Koutroumbas Konstantinos and Theodoridis Sergios. *Pattern Recognition*. Academic Press, Cambridge, MA, USA, Oct 2008. ISBN 978-1-59749272-0. URL: https://www.elsevier.com/books/pattern-recognition/koutroumbas/978-1-59749-272-0.

Kovács András and Szapudi István. Star-galaxy separation strategies for WISE-2MASS all-sky infrared galaxy catalogues. *MNRAS*, 448(2):1305–1313, April 2015. doi: 10.1093/mnras/stv063. ADS URL: https://ui.adsabs.harvard.edu/abs/2015MNRAS.448.1305K.

Kursa Miron and Rudnicki Witold. Feature selection with boruta package. *Journal of Statistical Software*, 36:1–13, 09 2010. doi: 10.18637/jss.v036.i11.

Laigle C., McCracken H. J., Ilbert O., et al. The COSMOS2015 Catalog: Exploring the 1 $<$ z $<$ 6 Universe with Half a Million Galaxies. *APJS*, 224:24, June 2016. doi: 10.3847/0067-0049/224/2/24. ADS URL: https://ui.adsabs.harvard.edu/abs/2016ApJS..224...24L.

Lander Eric S., Baylis Françoise, Zhang Feng, et al. Adopt a moratorium on heritable genome editing. *Nature*, 567(7747):165–168, March 2019. doi: 10.1038/d41586-019-00726-5. URL: https://doi.org/10.1038/d41586-019-00726-5.

Langley Pat. The changing science of machine learning. *Machine Learning*, 82(3):275–279, February 2011. doi: 10.1007/s10994-011-5242-y. URL: https://doi.org/10.1007/s10994-011-5242-y.

Lanusse François, Ma Quanbin, Li Nan, et al. CMU DeepLens: deep learning for automatic image-based galaxy-galaxy strong lens finding. *MNRAS*, 473(3):3895–3906, January 2018. doi: 10.1093/mnras/stx1665. ADS URL: https://ui.adsabs.harvard.edu/abs/2018MNRAS.473.3895L.

Laureijs R., Amiaux J., Arduini S., et al. Euclid Definition Study Report. *arXiv e-prints*, art. arXiv:1110.3193, October 2011. ADS URL: https://ui.adsabs.harvard.edu/abs/2011arXiv1110.3193L.

Lee Christopher and Hogan James. Automated crater detection with human level performance. *arXiv e-prints*, art. arXiv:2010.12520, October 2020. ADS URL: https://ui.adsabs.harvard.edu/abs/2020arXiv201012520L.

Leistedt Boris and Hogg David W. Data-driven, Interpretable Photometric Redshifts Trained on Heterogeneous and Unrepresentative Data. *ApJ*, 838(1):5, March 2017. doi: 10.3847/1538-4357/aa6332. ADS URL: https://ui.adsabs.harvard.edu/abs/2017ApJ...838....5L.

Li I. H. and Yee H. K. C. Finding Galaxy Groups in Photometric-Redshift Space: The Probability Friends-of-Friends Algorithm. *Aj*, 135(3):809–822, March 2008. doi: 10.1088/0004-6256/135/3/809. ADS URL: https://ui.adsabs.harvard.edu/abs/2008AJ....135..809L.

Lilly Simon, Le Brun Vincent, Maier Christian, et al. zcosmos – 10k-bright spectroscopic sample. *Astronomy & Astrophysics*, 523, 10 2009. doi: 10.1088/0067-0049/184/2/218.

Lima Marcos, Cunha Carlos E., Oyaizu Hiroaki, et al. Estimating the redshift distribution of photometric galaxy samples. *MNRAS*, 390(1):118–130, October 2008. doi: 10.1111/j.1365-2966.2008.13510.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2008MNRAS.390..118L.

Lloyd-Hart M., Wizinowich P., McLeod B., et al. First Results of an On-Line Adaptive Optics System with Atmospheric Wavefront Sensing by an Artificial Neural Network. *ApJ*, 390:L41, May 1992. doi: 10.1086/186367. ADS URL: https://ui.adsabs.harvard.edu/abs/1992ApJ...390L..41L.

Loh E. D. and Spillar E. J. Photometric Redshifts of Galaxies. *ApJ*, 303:154, April 1986. doi: 10.1086/164062. ADS URL: https://ui.adsabs.harvard.edu/abs/1986ApJ...303..154L.

Longo Giuseppe, Merényi Erzsébet, and Tiňo Peter. Foreword to the Focus Issue on Machine Intelligence in Astronomy and Astrophysics. *PASP*, 131(1004):100101, November 2019. doi: 10.1088/1538-3873/ab2743. ADS URL: https://ui.adsabs.harvard.edu/abs/2019PASP..131j0101L.

Lotz Jennifer M., Davis M., Faber S. M., et al. The Evolution of Galaxy Mergers and Morphology at z $<$ 1.2 in the Extended Groth Strip. *ApJ*, 672(1):177–197, January 2008. doi: 10.1086/523659. ADS URL: https://ui.adsabs.harvard.edu/abs/2008ApJ...672..177L.

Luo B., Brandt W. N., Xue Y. Q., et al. Identifications and Photometric Redshifts of the 2 Ms Chandra Deep Field-South Sources. *APJS*, 187(2):560–580, April 2010. doi: 10.1088/0067-0049/187/2/560. ADS URL: https://ui.adsabs.harvard.edu/abs/2010ApJS..187..560L.

Ma Zhaoming and Bernstein Gary. Size of Spectroscopic Calibration Samples for Cosmic Shear Photometric Redshifts. *ApJ*, 682(1):39–48, July 2008. doi: 10.1086/588214. ADS URL: https://ui.adsabs.harvard.edu/abs/2008ApJ...682...39M.

Magnelli B., Elbaz D., Chary R. R., et al. The 0.4 $<$ z $<$ 1.3 star formation history of the Universe as viewed in the far-infrared. *A&A*, 496(1):57–75, March 2009. doi: 10.1051/0004-6361:200811443. ADS URL: https://ui.adsabs.harvard.edu/abs/2009A&A...496...57M.

Magnelli B., Elbaz D., Chary R. R., et al. Evolution of the dusty infrared luminosity function from z = 0 to z = 2.3 using observations from Spitzer. *A&A*, 528:A35, April 2011. doi: 10.1051/0004-6361/200913941. ADS URL: https://ui.adsabs.harvard.edu/abs/2011A&A...528A..35M.

Mandelbaum R., Seljak U., Hirata C. M., et al. Precision photometric redshift calibration for galaxy-galaxy weak lensing. *MNRAS*, 386(2):781–806, May 2008. doi: 10.1111/j.1365-2966.2008.12947.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2008MNRAS.386..781M.

Mandelbaum Rachel, Hirata Christopher M., Broderick Tamara, et al. Ellipticity of dark matter haloes with galaxy-galaxy weak lensing. *MNRAS*, 370(2):1008–1024, August 2006. doi: 10.1111/j.1365-2966.2006.10539.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2006MNRAS.370.1008M.

Masci Frank J., Laher Russ R., Rebbapragada Umaa D., et al. The IPAC Image Subtraction and Discovery Pipeline for the Intermediate Palomar Transient Factory. *PASP*, 129(971):014002, January 2017. doi: 10.1088/1538-3873/129/971/014002. ADS URL: https://ui.adsabs.harvard.edu/abs/2017PASP..129a4002M.

Mashey John R. Big data... and the next wave of infrastress. slides, 1998. URL: https://static.usenix.org/event/usenix99/invited_talks/mashey.pdf.

Massarotti M., Iovino A., and Buzzoni A. A critical appraisal of the SED fitting method to estimate photometric redshifts. *A&A*, 368:74–85, March 2001a. doi: 10.1051/0004-6361:20000553. ADS URL: https://ui.adsabs.harvard.edu/abs/2001A&A...368...74M.

Massarotti M., Iovino A., Buzzoni A., and Valls-Gabaud D. New insights on the accuracy of photometric redshift measurements. *A&A*, 380:425–434, December 2001b. doi: 10.1051/0004-6361:20011409. ADS URL: https://ui.adsabs.harvard.edu/abs/2001A&A...380..425M.

Masters Daniel, Capak Peter, Stern Daniel, et al. Mapping the Galaxy Color-Redshift Relation: Optimal Photometric Redshift Calibration Strategies for Cosmology Surveys. *ApJ*, 813(1):53, Nov 2015. doi: 10.1088/0004-637X/813/1/53. ADS URL: https://ui.adsabs.harvard.edu/abs/2015ApJ...813...53M.

Masters Daniel C., Stern Daniel K., Cohen Judith G., et al. The Complete Calibration of the Color-Redshift Relation (C3R2) Survey: Survey Overview and Data Release 1. *ApJ*, 841(2):111, June 2017. doi: 10.3847/1538-4357/aa6f08. ADS URL: https://ui.adsabs.harvard.edu/abs/2017ApJ...841..111M.

Masters Daniel C., Stern Daniel K., Cohen Judith G., et al. The Complete Calibration of the Color-Redshift Relation (C3R2) Survey: Analysis and Data Release 2. *ApJ*, 877(2):81, June 2019. doi: 10.3847/1538-4357/ab184d. ADS URL: https://ui.adsabs.harvard.edu/abs/2019ApJ...877...81M.

Matthews Daniel J. and Newman Jeffrey A. Reconstructing Redshift Distributions with Cross-correlations: Tests and an Optimized Recipe. *ApJ*, 721(1):456–468, September 2010. doi: 10.1088/0004-637X/721/1/456. ADS URL: https://ui.adsabs.harvard.edu/abs/2010ApJ...721..456M.

McAlpine K., Jarvis M. J., and Bonfield D. G. Evolution of faint radio sources in the VIDEO-XMM3 field. *MNRAS*, 436(2):1084–1095, December 2013. doi: 10.1093/mnras/stt1638. ADS URL: https://ui.adsabs.harvard.edu/abs/2013MNRAS.436.1084M.

McCarthy John, Minsky Marvin L., Rochester Nathaniel, and Shannon Claude E. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, 27(4):12, Dec. 2006. doi: 10.1609/aimag.v27i4.1904. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/1904.

McCracken H. J., Milvang-Jensen B., Dunlop J., et al. UltraVISTA: a new ultra-deep near-infrared survey in COSMOS. *A&A*, 544:A156, August 2012. doi: 10.1051/0004-6361/201219507. ADS URL: https://ui.adsabs.harvard.edu/abs/2012A&A...544A.156M.

McIntosh Patrick S. The Classification of Sunspot Groups. *Sol. Phys.*, 125(2):251–267, September 1990. doi: 10.1007/BF00158405. ADS URL: https://ui.adsabs.harvard.edu/abs/1990SoPh..125..251M.

McLure R. J., Dunlop J. S., Bowler R. A. A., et al. A new multifield determination of the galaxy luminosity function at z = 7-9 incorporating the 2012 Hubble Ultra-Deep Field imaging. *MNRAS*, 432(4):2696–2716, July 2013. doi: 10.1093/mnras/stt627. ADS URL: https://ui.adsabs.harvard.edu/abs/2013MNRAS.432.2696M.

McQuinn Matthew and White Martin. On using angular cross-correlations to determine source redshift distributions. *MNRAS*, 433(4):2857–2883, August 2013. doi: 10.1093/mnras/stt914. ADS URL: https://ui.adsabs.harvard.edu/abs/2013MNRAS.433.2857M.

Miller Glenn. *Artificial intelligence applications for Hubble Space Telescope operations*, volume 329, pages 3–31. Lecture Notes in Physics, 1989a. doi: 10.1007/3-540-51044-3_14. ADS URL: https://ui.adsabs.harvard.edu/abs/1989LNP...329....3M.

Miller Richard W. WOLF - A computer expert system for sunspot classification and solar-flare prediction. *JRASC*, 82:191–203, August 1988. ADS URL: https://ui.adsabs.harvard.edu/abs/1988JRASC..82..191M.

Miller Richard W. *WOLF — A computer expert system for sunspot classification and solar flare prediction*, volume 329, page 107. Springer-Verlag, 1989b. doi: 10.1007/3-540-51044-3_20. ADS URL: https://ui.adsabs.harvard.edu/abs/1989LNP...329..107M.

Minsky Marvin. *Heuristic Aspects of the Artificial Intelligence Problem.* Ed. Services Technical Information agency, 1956. URL: https://books.google.it/books?id=fvWNo6_IZGUC&dq=%22artificial+intelligence%22&hl=en&source=gbs_navlinks_s.

Misra Siddharth and Li Hao. Chapter 9 - noninvasive fracture characterization based on the classification of sonic wave travel times. In Misra Siddharth, Li Hao, and He Jiabo, editors, *Machine Learning for Subsurface Characterization*, pages 243–287. Gulf Professional Publishing, 2020. ISBN 978-0-12-817736-5. doi: https://doi.org/10.1016/B978-0-12-817736-5.00009-0. URL: http://www.sciencedirect.com/science/article/pii/B9780128177365000090.

Mobasher B. and Mazzei P. The effect of dust on photometric redshift measurement: a self-consistent technique. *A&A*, 363:517–525, November 2000. ADS URL: https://ui.adsabs.harvard.edu/abs/2000A&A...363..517M.

Moore Jason A., Pimbblet Kevin A., and Drinkwater Michael J. Mathematical Morphology: Star/Galaxy Differentiation & Galaxy Morphology Classification. *PASA*, 23 (4):135–146, February 2006. doi: 10.1071/AS06010. ADS URL: https://ui.adsabs.harvard.edu/abs/2006PASA...23..135M.

Morrison C. B., Hildebrandt H., Schmidt S. J., et al. the-wizz: clustering redshift estimation for everyone. *MNRAS*, 467(3):3576–3589, May 2017. doi: 10.1093/mnras/stx342. ADS URL: https://ui.adsabs.harvard.edu/abs/2017MNRAS.467.3576M.

Morrissey Patrick, Conrow Tim, Barlow Tom A., et al. The Calibration and Data Products of GALEX. *APJS*, 173(2):682–697, December 2007. doi: 10.1086/520512. ADS URL: https://ui.adsabs.harvard.edu/abs/2007ApJS..173..682M.

Mukund N., Abraham S., Kandhasamy S., et al. Transient classification in LIGO data using difference boosting neural network. *Phys. Rev. D*, 95(10):104059, May 2017. doi: 10.1103/PhysRevD.95.104059. ADS URL: https://ui.adsabs.harvard.edu/abs/2017PhRvD..95j4059M.

Müller Andreas C. and Guido Sarah. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, Sebastopol, CA, USA, 2016. ISBN 978-1-44936989-7. URL: https://books.google.it/books/about/Introduction_to_Machine_Learning_with_Py.html?id=vbQlDQAAQBAJ&redir_esc=y.

Newman Jeffrey A. Calibrating Redshift Distributions beyond Spectroscopic Limits with Cross-Correlations. *ApJ*, 684(1):88–101, September 2008. doi: 10.1086/589982. ADS URL: https://ui.adsabs.harvard.edu/abs/2008ApJ...684...88N.

Newman Jeffrey A., Abate Alexandra, Abdalla Filipe B., et al. Spectroscopic needs for imaging dark energy experiments. *Astroparticle Physics*, 63:81–100, Mar 2015. doi: 10.1016/j.astropartphys.2014.06.007. ADS URL: https://ui.adsabs.harvard.edu/abs/2015APh....63...81N.

Nicastro F., Kaastra J., Krongold Y., et al. Observations of the missing baryons in the warm-hot intergalactic medium. *Nature*, 558(7710):406–409, June 2018. doi: 10.1038/s41586-018-0204-1. ADS URL: https://ui.adsabs.harvard.edu/abs/2018Natur.558..406N.

Nocedal . *Numerical optimization / Jorge Nocedal, Stephen J. Wright*. Springer Series in Operations Research and Financial Engineering. Springer, Berlin, 2nd edition edition, 2006. ISBN 978-0387-30303-1.

Norris Ray P., Salvato M., Longo G., et al. A Comparison of Photometric Redshift Techniques for Large Radio Surveys. *PASP*, 131(1004):108004, October 2019. doi: 10.1088/1538-3873/ab0f7b. ADS URL: https://ui.adsabs.harvard.edu/abs/2019PASP..131j8004N.

Odewahn S. C., Stockwell E. B., Pennington R. L., et al. Automated Star/Galaxy Discrimination With Neural Networks. *Aj*, 103:318, January 1992. doi: 10.1086/116063. ADS URL: https://ui.adsabs.harvard.edu/abs/1992AJ....103..318O.

Odewahn S. C., de Carvalho R. R., Gal R. R., et al. The Digitized Second Palomar Observatory Sky Survey (DPOSS). III. Star-Galaxy Separation. *Aj*, 128(6):3092–3107, December 2004. doi: 10.1086/425525. ADS URL: https://ui.adsabs.harvard.edu/abs/2004AJ....128.3092O.

Padilla Cristóbal, Castander Francisco J., Alarcón Alex, et al. The Physics of the Accelerating Universe Camera. *Aj*, 157(6):246, June 2019. doi: 10.3847/1538-3881/ab0412. ADS URL: https://ui.adsabs.harvard.edu/abs/2019AJ....157..246P.

Padmanabhan Nikhil, Budavári Tamás, Schlegel David J., et al. Calibrating photometric redshifts of luminous red galaxies. *MNRAS*, 359(1):237–250, May 2005. doi: 10.1111/j.1365-2966.2005.08915.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2005MNRAS.359..237P.

Padmanabhan Nikhil, Schlegel David J., Seljak Uroš, et al. The clustering of luminous red galaxies in the Sloan Digital Sky Survey imaging data. *MNRAS*, 378(3):852–872, July 2007. doi: 10.1111/j.1365-2966.2007.11593.x. ADS URL: https://ui.adsabs.harvard.edu/abs/2007MNRAS.378..852P.

Papovich Casey, Dickinson Mark, and Ferguson Henry C. The Stellar Populations and Evolution of Lyman Break Galaxies. *ApJ*, 559(2):620–653, October 2001. doi: 10.1086/322412. ADS URL: https://ui.adsabs.harvard.edu/abs/2001ApJ...559..620P.

Parker Laura C., Hoekstra Henk, Hudson Michael J., et al. The Masses and Shapes of Dark Matter Halos from Galaxy-Galaxy Lensing in the CFHT Legacy Survey. *ApJ*, 669(1):21–31, November 2007. doi: 10.1086/521541. ADS URL: https://ui.adsabs.harvard.edu/abs/2007ApJ...669...21P.

Parsa Shaghayegh, Dunlop James S., and McLure Ross J. No evidence for a significant AGN contribution to cosmic hydrogen reionization. *MNRAS*, 474(3):2904–2923, March 2018. doi: 10.1093/mnras/stx2887. ADS URL: https://ui.adsabs.harvard.edu/abs/2018MNRAS.474.2904P.

Pasquet Johanna, Bertin E., Treyer M., et al. Photometric redshifts from SDSS images using a convolutional neural network. *A&A*, 621:A26, January 2019. doi: 10.1051/

0004-6361/201833617. ADS URL: https://ui.adsabs.harvard.edu/abs/2019A&A...621A..26P.

Pasquet-Itam J. and Pasquet J. Deep learning approach for classifying, detecting and predicting photometric redshifts of quasars in the Sloan Digital Sky Survey stripe 82. *A&A*, 611:A97, April 2018. doi: 10.1051/0004-6361/201731106. ADS URL: https://ui.adsabs.harvard.edu/abs/2018A&A...611A..97P.

Pearson Kyle A., Palafox Leon, and Griffith Caitlin A. Searching for exoplanets using artificial intelligence. *MNRAS*, 474(1):478–491, February 2018. doi: 10.1093/mnras/stx2761. ADS URL: https://ui.adsabs.harvard.edu/abs/2018MNRAS.474..478P.

Pérez-González Pablo G., Rieke George H., Egami Eiichi, et al. Spitzer View on the Evolution of Star-forming Galaxies from z = 0 to z ~3. *ApJ*, 630(1):82–107, September 2005. doi: 10.1086/431894. ADS URL: https://ui.adsabs.harvard.edu/abs/2005ApJ...630...82P.

Petrillo C. E., Tortora C., Chatterjee S., et al. Finding strong gravitational lenses in the Kilo Degree Survey with Convolutional Neural Networks. *MNRAS*, 472(1):1129–1150, November 2017. doi: 10.1093/mnras/stx2052. ADS URL: https://ui.adsabs.harvard.edu/abs/2017MNRAS.472.1129P.

Polsterer Kai Lars, D'Isanto Antonio, and Gieseke Fabian. Uncertain Photometric Redshifts. *arXiv e-prints*, art. arXiv:1608.08016, August 2016. ADS URL: https://ui.adsabs.harvard.edu/abs/2016arXiv160808016P.

Puschell J. J., Owen F. N., and Laing R. A. Near-infrared photometry of distant radio galaxies - Spectral flux distributions and redshift estimates. *ApJ*, 257:L57–L61, June 1982. doi: 10.1086/183808. ADS URL: https://ui.adsabs.harvard.edu/abs/1982ApJ...257L..57P.

Rasmussen Carl Edward and Williams Christopher K. I. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. ADS URL: https://ui.adsabs.harvard.edu/abs/2006gpml.book.....R.

Razim Oleksandra, Cavuoti Stefano, Brescia Massimo, et al. Towards reliable photometric redshifts with machine learning methods. *MNRAS*, 2021.

Reis Itamar, Baron Dalya, and Shahaf Sahar. Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets. *Aj*, 157(1):16, January 2019. doi: 10.3847/1538-3881/aaf101. ADS URL: https://ui.adsabs.harvard.edu/abs/2019AJ....157...16R.

Richards Gordon T., Weinstein Michael A., Schneider Donald P., et al. Photometric Redshifts of Quasars. *Aj*, 122(3):1151–1162, September 2001. doi: 10.1086/322132. ADS URL: https://ui.adsabs.harvard.edu/abs/2001AJ....122.1151R.

Rosenblatt F. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. *American Journal of Psychology*, 76:705, 1963.

Rosenthal D. A. Applying artificial intelligence to astronomical databases - a surveyof applicable technology. In *European Southern Observatory Conference and Workshop Proceedings*, volume 28 of *European Southern Observatory Conference and Workshop Proceedings*, pages 245–259, January 1988. ADS URL: https://ui.adsabs.harvard.edu/abs/1988ESOC...28..245R.

Sadeh I., Abdalla F. B., and Lahav O. ANNz2: Photometric Redshift and Probability Distribution Function Estimation using Machine Learning. *PASP*, 128(968):104502, October 2016. doi: 10.1088/1538-3873/128/968/104502. ADS URL: https://ui.adsabs.harvard.edu/abs/2016PASP..128j4502S.

Sagiroglu Seref and Sinanc Duygu. Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, May 2013. doi: 10.1109/cts.2013.6567202. URL: https://doi.org/10.1109/cts.2013.6567202.

Salimbeni S., Castellano M., Pentericci L., et al. A comprehensive study of large-scale structures in the GOODS-SOUTH field up to z $\sim$ 2.5. *A&A*, 501(3):865–877, July 2009. doi: 10.1051/0004-6361/200811570. ADS URL: https://ui.adsabs.harvard.edu/abs/2009A&A...501..865S.

Salloum Salman, Dautov Ruslan, Chen Xiaojun, et al. Big data analytics on Apache Spark. *Int. J. Data Sci. Anal.*, 1(3):145–164, Nov 2016. ISSN 2364-4168. doi: 10.1007/s41060-016-0027-9.

Salvato M., Hasinger G., Ilbert O., et al. Photometric Redshift and Classification for the XMM-COSMOS Sources. *ApJ*, 690(2):1250–1263, January 2009. doi: 10.1088/0004-637X/690/2/1250. ADS URL: https://ui.adsabs.harvard.edu/abs/2009ApJ...690.1250S.

Salvato Mara, Ilbert Olivier, and Hoyle Ben. The many flavours of photometric redshifts. *Nature Astronomy*, 3:212–222, Jun 2019. doi: 10.1038/s41550-018-0478-0. ADS URL: https://ui.adsabs.harvard.edu/abs/2019NatAs...3..212S.

Sánchez C., Carrasco Kind M., Lin H., et al. Photometric redshift analysis in the Dark Energy Survey Science Verification data. *MNRAS*, 445(2):1482–1506, December 2014. doi: 10.1093/mnras/stu1836. ADS URL: https://ui.adsabs.harvard.edu/abs/2014MNRAS.445.1482S.

Sandage A. Optical redshifts for 719 bright galaxies. *Aj*, 83:904–937, August 1978. doi: 10.1086/112271. ADS URL: https://ui.adsabs.harvard.edu/abs/1978AJ.....83..904S.

Sandler D. G., Barrett T. K., Palmer D. A., et al. Use of a neural network to control an adaptive optics system for an astronomical telescope. *Nature*, 351(6324):300–302, May 1991. doi: 10.1038/351300a0. ADS URL: https://ui.adsabs.harvard.edu/abs/1991Natur.351..300S.

Schaan Emmanuel, Ferraro Simone, and Seljak Uros. Photo-z outlier self-calibration in weak lensing surveys. *JCAP*, 2020(12):001, December 2020. doi: 10.1088/1475-7516/2020/12/001. ADS URL: https://ui.adsabs.harvard.edu/abs/2020JCAP...12..001S.

Schanche N., Collier Cameron A., Hébrard G., et al. Machine-learning approaches to exoplanet transit detection and candidate validation in wide-field ground-based surveys. *MNRAS*, 483(4):5534–5547, March 2019. doi: 10.1093/mnras/sty3146. ADS URL: https://ui.adsabs.harvard.edu/abs/2019MNRAS.483.5534S.

Schmidt S. J., Malz A. I., Soo J. Y. H., et al. Evaluation of probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST). *MNRAS*, September 2020. doi: 10.1093/mnras/staa2799. ADS URL: https://ui.adsabs.harvard.edu/abs/2020MNRAS.tmp.2629S.

Schrabback T., Hartlap J., Joachimi B., et al. Evidence of the accelerated expansion of the Universe from weak lensing tomography with COSMOS. *A&A*, 516:A63, June 2010. doi: 10.1051/0004-6361/200913577. ADS URL: https://ui.adsabs.harvard.edu/abs/2010A&A...516A..63S.

Scoville N., Aussel H., Brusa M., et al. The Cosmic Evolution Survey (COSMOS): Overview. *APJS*, 172(1):1–8, September 2007. doi: 10.1086/516585. ADS URL: https://ui.adsabs.harvard.edu/abs/2007ApJS..172....1S.

Sebok W. L. Optimal classification of images into stars or galaxies - a Bayesian approach. *Aj*, 84:1526–1536, October 1979. doi: 10.1086/112570. ADS URL: https://ui.adsabs.harvard.edu/abs/1979AJ.....84.1526S.

Serjeant Stephen, Elvis Martin, and Tinetti Giovanna. The future of astronomy with small satellites. *Nature Astronomy*, 4:1031–1038, November 2020. doi: 10.1038/s41550-020-1201-5. ADS URL: https://ui.adsabs.harvard.edu/abs/2020NatAs...4.1031S.

Shalev-Shwartz Shai and Ben-David Shai. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 978-1-10-705713-5.

Shallue Christopher J. and Vanderburg Andrew. Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90. *Aj*, 155(2):94, February 2018. doi: 10.3847/1538-3881/aa9e09. ADS URL: https://ui.adsabs.harvard.edu/abs/2018AJ....155...94S.

Shkolnik Evgenya L. On the verge of an astronomy CubeSat revolution. *Nature Astronomy*, 2:374–378, May 2018. doi: 10.1038/s41550-018-0438-8. ADS URL: https://ui.adsabs.harvard.edu/abs/2018NatAs...2..374S.

Silburt Ari, Ali-Dib Mohamad, Zhu Chenchong, et al. Lunar crater identification via deep learning. *Icarus*, 317:27–38, January 2019. doi: 10.1016/j.icarus.2018.06.022. ADS URL: https://ui.adsabs.harvard.edu/abs/2019Icar..317...27S.

Simmons B. D., Lintott Chris, Willett Kyle W., et al. Galaxy Zoo: quantitative visual morphological classifications for 48 000 galaxies from CANDELS. *MNRAS*, 464(4): 4420–4447, February 2017. doi: 10.1093/mnras/stw2587. ADS URL: https://ui.adsabs.harvard.edu/abs/2017MNRAS.464.4420S.

Siudek M., Małek K., Pollo A., et al. The VIMOS Public Extragalactic Redshift Survey (VIPERS). The complexity of galaxy populations at 0.4 < z < 1.3 revealed with unsupervised machine-learning algorithms. *A&A*, 617:A70, September 2018. doi: 10.1051/0004-6361/201832784. ADS URL: https://ui.adsabs.harvard.edu/abs/2018A&A...617A..70S.

Slezak E., Bijaoui A., and Mars G. Galaxy counts in the Coma supercluster field. II. Automated image detection and classification. *A&A*, 201:9–20, July 1988. ADS URL: https://ui.adsabs.harvard.edu/abs/1988A&A...201....9S.

Speagle Joshua S. and Eisenstein Daniel J. Deriving photometric redshifts using fuzzy archetypes and self-organizing maps - I. Methodology. *MNRAS*, 469(1):1186–1204, July 2017. doi: 10.1093/mnras/stw1485. ADS URL: https://ui.adsabs.harvard.edu/abs/2017MNRAS.469.1186S.

Spergel D., Gehrels N., Baltay C., et al. Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report. *arXiv e-prints*, art. arXiv:1503.03757, March 2015. ADS URL: https://ui.adsabs.harvard.edu/abs/2015arXiv150303757S.

Stensbo-Smidt Kristoffer, Gieseke Fabian, Igel Christian, et al. Sacrificing information for the greater good: how to select photometric bands for optimal accuracy. *MNRAS*, 464(3):2577–2596, January 2017. doi: 10.1093/mnras/stw2476. ADS URL: https://ui.adsabs.harvard.edu/abs/2017MNRAS.464.2577S.

Storrie-Lombardi M. C., Lahav O., Sodre Jr., L., and Storrie-Lombardi L. J. Morphological Classification of Galaxies by Artificial Neural Networks. *MNRAS*, 259: 8P, November 1992. doi: 10.1093/mnras/259.1.8P. ADS URL: https://ui.adsabs.harvard.edu/abs/1992MNRAS.259P...8S.

Stoughton Chris, Lupton Robert H., Bernardi Mariangela, et al. Sloan Digital Sky Survey: Early Data Release. *Aj*, 123(1):485–548, January 2002. doi: 10.1086/324741. ADS URL: https://ui.adsabs.harvard.edu/abs/2002AJ....123..485S.

Szalay A. S. and Brunner R. J. Astronomical archives of the future: a Virtual Observatory. *Future Gener. Comput. Syst. (Netherlands*, 16(1):63–72, November 1999. ADS URL: https://ui.adsabs.harvard.edu/abs/1999FGST...16...63S.

Tagliaferri Roberto, Longo Guiseppe, Andreon Stefano, et al. *Neural Networks for Photometric Redshifts Evaluation*, volume 2859, pages 226–234. Springer Berlin, 2003. doi: 10.1007/978-3-540-45216-4_26. ADS URL: https://ui.adsabs.harvard.edu/abs/2003LNCS.2859..226T.

Tanaka Masayuki. Photometric Redshift with Bayesian Priors on Physical Properties of Galaxies. *ApJ*, 801(1):20, March 2015. doi: 10.1088/0004-637X/801/1/20. ADS URL: https://ui.adsabs.harvard.edu/abs/2015ApJ...801...20T.

Tanaka Masayuki, Coupon Jean, Hsieh Bau-Ching, et al. Photometric redshifts for Hyper Suprime-Cam Subaru Strategic Program Data Release 1. *PASJ*, 70:S9, January 2018. doi: 10.1093/pasj/psx077. ADS URL: https://ui.adsabs.harvard.edu/abs/2018PASJ...70S...9T.

Thakar A. R., Szalay A. S., O'Mullane W., et al. Brave New World: Data Intensive Science with SDSS and the VO. In *American Astronomical Society Meeting Abstracts*, volume 205 of *American Astronomical Society Meeting Abstracts*, page 113.01, December 2004. ADS URL: https://ui.adsabs.harvard.edu/abs/2004AAS...20511301T.

The Dark Energy Survey Collaboration . The Dark Energy Survey. *arXiv e-prints*, art. astro-ph/0510346, October 2005. ADS URL: https://ui.adsabs.harvard.edu/abs/2005astro.ph.10346T.

The LSST Dark Energy Science Collaboration , Mandelbaum Rachel, Eifler Tim, et al. The LSST Dark Energy Science Collaboration (DESC) Science Requirements Document. *arXiv e-prints*, art. arXiv:1809.01669, September 2018. ADS URL: https://ui.adsabs.harvard.edu/abs/2018arXiv180901669T.

Tibshirani R. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7: 1456–1490, 2012.

Tozzi P., Gilli R., Mainieri V., et al. X-ray spectral properties of active galactic nuclei in the Chandra Deep Field South. *A&A*, 451(2):457–474, May 2006. doi: 10.1051/0004-6361:20042592. ADS URL: https://ui.adsabs.harvard.edu/abs/2006A&A...451..457T.

Turing A. M. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX (236):433–460, October 1950. doi: 10.1093/mind/lix.236.433. URL: https://doi.org/10.1093/mind/lix.236.433.

Tyson J. Anthony, Ivezić Željko, Bradshaw Andrew, et al. Mitigation of LEO Satellite Brightness and Trail Effects on the Rubin Observatory LSST. *Aj*, 160(5):226, November 2020. doi: 10.3847/1538-3881/abba3e. ADS URL: https://ui.adsabs.harvard.edu/abs/2020AJ....160..226T.

Vanzella E., Cristiani S., Fontana A., et al. Photometric redshifts with the Multilayer Perceptron Neural Network: Application to the HDF-S and SDSS. *A&A*, 423:761–776, August 2004. doi: 10.1051/0004-6361:20040176. ADS URL: https://ui.adsabs.harvard.edu/abs/2004A&A...423..761V.

Vavilova I. B., Dobrycheva D. V., Vasylenko M. Yu., et al. Machine learning technique for morphological classification of galaxies at z<0.1 from the SDSS. *arXiv e-prints*, art. arXiv:1712.08955, December 2017. ADS URL: https://ui.adsabs.harvard.edu/abs/2017arXiv171208955V.

Wadadekar Yogesh. Estimating Photometric Redshifts Using Support Vector Machines. *PASP*, 117(827):79–85, January 2005. doi: 10.1086/427710. ADS URL: https://ui.adsabs.harvard.edu/abs/2005PASP..117...79W.

Wake David A., Whitaker Katherine E., Labbé Ivo, et al. Galaxy Clustering in the NEWFIRM Medium Band Survey: The Relationship Between Stellar Mass and Dark Matter Halo Mass at 1 < z < 2. *ApJ*, 728(1):46, February 2011. doi: 10.1088/0004-637X/728/1/46. ADS URL: https://ui.adsabs.harvard.edu/abs/2011ApJ...728...46W.

Wang D., Zhang Y., and Zhao Y. Estimating Photometric Redshifts of Quasars Using Support Vector Machines. In Argyle R. W., Bunclark P. S., and Lewis J. R., editors, *Astronomical Data Analysis Software and Systems XVII*, volume 394 of *Astronomical Society of the Pacific Conference Series*, page 509, August 2008. ADS URL: https://ui.adsabs.harvard.edu/abs/2008ASPC..394..509W.

Wang Yun, Bahcall Neta, and Turner Edwin L. A Catalog of Color-based Redshift Estimates for Z <~4 Galaxies in the Hubble Deep Field. *Aj*, 116(5):2081–2085, November 1998. doi: 10.1086/300592. ADS URL: https://ui.adsabs.harvard.edu/abs/1998AJ....116.2081W.

Way M. J. and Klose C. D. Can Self-Organizing Maps Accurately Predict Photometric Redshifts? *PASP*, 124(913):274, Mar 2012. doi: 10.1086/664796. ADS URL: https://ui.adsabs.harvard.edu/abs/2012PASP..124..274W.

Way M. J. and Srivastava A. N. Novel Methods for Predicting Photometric Redshifts from Broadband Photometry Using Virtual Sensors. *ApJ*, 647(1):102–115, August 2006. doi: 10.1086/505293. ADS URL: https://ui.adsabs.harvard.edu/abs/2006ApJ...647..102W.

Way M. J., Foster L. V., Gazis P. R., and Srivastava A. N. New Approaches to Photometric Redshift Prediction Via Gaussian Process Regression in the Sloan Digital Sky Survey. *ApJ*, 706(1):623–636, November 2009. doi: 10.1088/0004-637X/706/1/623. ADS URL: https://ui.adsabs.harvard.edu/abs/2009ApJ...706..623W.

Weaver W. B. Automatic Classification of WN Spectra. In *Bulletin of the American Astronomical Society*, volume 22, page 848, March 1990. ADS URL: https://ui.adsabs.harvard.edu/abs/1990BAAS...22..848W.

Weinberg David H., Mortonson Michael J., Eisenstein Daniel J., et al. Observational probes of cosmic acceleration. *Phys. Rep.*, 530(2):87–255, September 2013. doi: 10.1016/j.physrep.2013.05.001. ADS URL: https://ui.adsabs.harvard.edu/abs/2013PhR...530...87W.

Wen Z. L. and Han J. L. Galaxy Clusters at High Redshift and Evolution of Brightest Cluster Galaxies. *ApJ*, 734(1):68, June 2011. doi: 10.1088/0004-637X/734/1/68. ADS URL: https://ui.adsabs.harvard.edu/abs/2011ApJ...734...68W.

Williams Robert E., Blacker Brett, Dickinson Mark, et al. The Hubble Deep Field: Observations, Data Reduction, and Galaxy Photometry. *Aj*, 112:1335, October 1996. doi: 10.1086/118105. ADS URL: https://ui.adsabs.harvard.edu/abs/1996AJ....112.1335W.

Wolf C., Wisotzki L., Borch A., et al. The evolution of faint AGN between z =~1 and z =~5 from the COMBO-17 survey. *A&A*, 408:499–514, September 2003. doi: 10.1051/0004-6361:20030990. ADS URL: https://ui.adsabs.harvard.edu/abs/2003A&A...408..499W.

Wright Angus H., Hildebrandt Hendrik, Kuijken Konrad, et al. KiDS+VIKING-450: A new combined optical and near-infrared dataset for cosmology and astrophysics. *A&A*,

632:A34, December 2019. doi: 10.1051/0004-6361/201834879. ADS URL: https://ui.adsabs.harvard.edu/abs/2019A&A...632A..34W.

Wright Angus H., Hildebrandt Hendrik, van den Busch Jan Luca, and Heymans Catherine. Photometric redshift calibration with self-organising maps. *A&A*, 637:A100, May 2020. doi: 10.1051/0004-6361/201936782. ADS URL: https://ui.adsabs.harvard.edu/abs/2020A&A...637A.100W.

Wu Tailin and Tegmark Max. Toward an AI Physicist for Unsupervised Learning. *arXiv e-prints*, art. arXiv:1810.10525, October 2018. ADS URL: https://ui.adsabs.harvard.edu/abs/2018arXiv181010525W.

Yasuda N., Mizumoto Y., Ohishi M., et al. Astronomical Data Query Language: Simple Query Protocol for the Virtual Observatory. In Ochsenbein Francois, Allen Mark G., and Egret Daniel, editors, *Astronomical Data Analysis Software and Systems (ADASS) XIII*, volume 314 of *Astronomical Society of the Pacific Conference Series*, page 293, July 2004. ADS URL: https://ui.adsabs.harvard.edu/abs/2004ASPC..314..293Y.

Yee H. K. C. Photometric Redshift Techniques: Reliability and Applications. *arXiv e-prints*, art. astro-ph/9809347, September 1998. ADS URL: https://ui.adsabs.harvard.edu/abs/1998astro.ph..9347Y.

Zevin M., Coughlin S., Bahaadini S., et al. Gravity Spy: integrating advanced LIGO detector characterization, machine learning, and citizen science. *Classical and Quantum Gravity*, 34(6):064003, March 2017. doi: 10.1088/1361-6382/aa5cea. ADS URL: https://ui.adsabs.harvard.edu/abs/2017CQGra..34f4003Z.

Zhan Hu, Wang Lifan, Pinto Philip, and Tyson J. Anthony. Measuring Baryon Acoustic Oscillations with Millions of Supernovae. *ApJ*, 675(1):L1, March 2008. doi: 10.1086/529546. ADS URL: https://ui.adsabs.harvard.edu/abs/2008ApJ...675L...1Z.

Zhang Yanxia, Ma He, Peng Nanbo, et al. Estimating Photometric Redshifts of Quasars via the k-nearest Neighbor Approach Based on Large Survey Databases. *Aj*, 146(2):22, August 2013. doi: 10.1088/0004-6256/146/2/22. ADS URL: https://ui.adsabs.harvard.edu/abs/2013AJ....146...22Z.