A&A 611, A53 (2018)
DOI: 10.1051/0004-6361/201731305
© ESO 2018

Astronomy & Astrophysics

Automated reliability assessment for spectroscopic redshift measurements*

S. Jamal[1], V. Le Brun[1], O. Le Fèvre[1], D. Vibert[1], A. Schmitt[1], C. Surace[1], Y. Copin[2],
B. Garilli[3], M. Moresco[4,5], and L. Pozzetti[5]

[1] Aix Marseille Univ., CNRS, LAM, Laboratoire d'Astrophysique de Marseille, 13013 Marseille, France
e-mail: [sara.jamal;vincent.lebrun]@lam.fr
[2] Université Lyon, Université Lyon 1, CNRS/IN2P3, Institut de Physique Nucléaire de Lyon, 69622 Villeurbanne cedex, France
[3] INAF – Istituto di Astrofisica Spaziale e Fisica Cosmica Milano, via Bassini 15, 20133 Milano, Italy
[4] Dipartimento di Fisica e Astronomia, Università di Bologna, via Gobetti 93/2, 40129 Bologna, Italy
[5] INAF–Osservatorio Astronomico di Bologna, via Gobetti 93/3, 40129 Bologna, Italy

Received 2 June 2017/ Accepted 9 September 2017

ABSTRACT

Context. Future large-scale surveys, such as the ESA *Euclid* mission, will produce a large set of galaxy redshifts ($\geq 10^6$) that will require fully automated data-processing pipelines to analyze the data, extract crucial information and ensure that all requirements are met. A fundamental element in these pipelines is to associate to each galaxy redshift measurement a quality, or reliability, estimate.

Aims. In this work, we introduce a new approach to automate the spectroscopic redshift reliability assessment based on machine learning (ML) and characteristics of the redshift probability density function.

Methods. We propose to rephrase the spectroscopic redshift estimation into a Bayesian framework, in order to incorporate all sources of information and uncertainties related to the redshift estimation process and produce a redshift posterior probability density function (PDF). To automate the assessment of a reliability flag, we exploit key features in the redshift posterior PDF and machine learning algorithms.

Results. As a working example, public data from the VIMOS VLT Deep Survey is exploited to present and test this new methodology. We first tried to reproduce the existing reliability flags using supervised classification in order to describe different types of redshift PDFs, but due to the subjective definition of these flags (classification accuracy ~58%), we soon opted for a new homogeneous partitioning of the data into distinct clusters via unsupervised classification. After assessing the accuracy of the new clusters via resubstitution and test predictions (classification accuracy ~98%), we projected unlabeled data from preliminary mock simulations for the *Euclid* space mission into this mapping to predict their redshift reliability labels.

Conclusions. Through the development of a methodology in which a system can build its own experience to assess the quality of a parameter, we are able to set a preliminary basis of an automated reliability assessment for spectroscopic redshift measurements. This newly-defined method is very promising for next-generation large spectroscopic surveys from the ground and in space, such as *Euclid* and WFIRST.

Key words. methods: data analysis – methods: statistical – techniques: spectroscopic – galaxies: distances and redshifts – surveys

1. Introduction

Next-generation experiments in Cosmology face the formidable challenge of understanding dark matter (DM) and dark energy (DE), two major components seemingly dominating the Universe content and evolution.

To improve our understanding of the Universe evolution history, the investigation of the distribution of galaxies over large volumes of the Universe at different cosmic times now constitutes a key requirement for future observational programs such as *Euclid* (Laureijs et al. 2011), WFIRST (Green et al. 2012), and LSST (Ivezic et al. 2008) that will exploit cosmological probes such as Weak Lensing (WL) and Galaxy Clustering (GC: baryon acoustic oscillations – BAO, redshift space distortions – RSD) to define the role of the dark components (Albrecht et al. 2006).

In GC, the detection of the BAOs at the sound horizon scale ($r_{\rm s} \approx 105\ h^{-1}$ Mpc) is used to investigate the role of DE in the evolution of the expansion through measurements of the Hubble parameter $H(z)$ and the comoving angular distances $D_{\rm A}(z)$ (Beutler et al. 2011), while the detection of the distorsions in the redshift space is used to probe the structures' growth and DE models by measuring the parameter combination $g_\theta = f(z)\sigma_8(z)$, where $f(z)$ and σ_8 refer to the growth rate and the rms amplitude (in a sphere of radius 8 h^{-1} Mpc) of the density fluctuations (Beutler et al. 2012), respectively. The WL is used to map the matter distribution (dark + visible) in the Universe and constrain the expansion history through precise measurements of shapes and distances of lensed galaxies (Huterer 2002; Linder & Jenkins 2003).

* A table of the reclassified VVDS redshifts and reliability is only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/qcat?J/A+A/611/A53

In Cosmology, the redshift z is a fundamental quantity, which links distances and cosmic time through the use of a cosmological model. Accurate redshift measurements are at the core of all modern experiments aiming at precision cosmology for a better understanding of the Universe content, focused on the dominant DM and DE components, as the cosmological probes GC and WL that require precise redshift measurements to build robust statistical models to constrain the DE equation-of-state and investigate the content of the dark Universe (Abdalla et al. 2008; Wang et al. 2010). In particular, 3D galaxy distribution maps from GC measurements entail precise measurements of spectroscopic redshifts, while cosmic shear measurements in WL require, along with high-quality imaging and photometry, the selection of sources using redshift measurements for two reasons: First, the galaxies in front of the lens are not affected by the gravitational lensing but they dilute the signal of the galaxy source in the background, and second, the galaxies at the same redshift as the lens contribute to the intrinsic alignment that disrupts the WL measurements.

As part of the future large-scale experiments in Cosmology designed to address the DE and DM origin, the *Euclid* mission is a M-Class ESA mission from the ESA Cosmic Vision program that aims to probe the expansion and the LSS growth histories in the Universe. Through the combination of cosmological probes (BAO, RSD, WL, clusters of galaxies, supernovae – SNe), *Euclid* will achieve an unprecedented level of accuracy and control of systematic effects to derive precise measurements of the Hubble parameter $H(z)$, the linear growth rate of structures γ, the DE equation-of-state parameters (ω_p, ω_a), the non-Gaussianity amplitude f_{NL} and the rms fluctuation of the matter over-density σ_8, among other cosmological parameters (Laureijs et al. 2011).

By covering a large fraction of the sky (Wide: ~15 000 deg^2, Deep: a total of 40 deg^2), the mission will perform a photometric survey in the visible and three near-infrared bands to measure the weak gravitational lensing by imaging approximately 1.5 billion galaxies with a photometric redshift accuracy of $\sigma_z/(1+z) \leq 0.05$, in addition to a spectroscopic slitless survey of approximately 25 million galaxies with a redshift accuracy of $\sigma_z/(1+z) \leq 0.001$ in order to derive precise measurements of the galaxy power spectrum (Laureijs et al. 2011). The wide-field *Euclid* survey will be particularly challenging because of the large-size sample of faint distant galaxies, for which the spectroscopic redshifts need to be automatically measured, and their corresponding reliability evaluated.

For large-scale surveys such as *Euclid*, the sheer amount of data requires the development of robust and fully automated data-processing pipelines to analyze the data, extract useful information (e.g., redshift) and ensure that all requirements are met.

Distinct approaches to estimate redshifts have been used in a broad range of galaxy surveys. Photometric redshifts z_{phot} are estimated using spectral energy distribution (SED) template fitting (e.g., Hyper-z Bolzonella et al. 2000; Le Phare Ilbert et al. 2006), classification with neural networks to produce a mapping between photometric observables and reference data (e.g., ANNz, Collister & Lahav 2004), or Bayesian inference to compute a posterior z_{phot} PDF with prior information from integrated flux in filters, colour or magnitude: BPZ (Benitez 1999), ZEBRA (Feldmann et al. 2006), EAZY (Brammer et al. 2008). On the other hand, spectroscopic redshifts z_{spec} are estimated from the direct application of cross-correlation or chi-square-fitting methods between the observed data and a reference set of spectroscopic templates (Tonry & Davis 1979; Simkin 1974; Schuecker 1993; Machado et al. 2013), or using spectral feature detection (emission/absorption lines and continuum features including spectral discontinuities in the UV-visible domain such as the Lyman break or the Balmer and D4000A breaks) that can be very powerful (Schuecker 1993). Some codes (EZ, Garilli et al. 2010) combine spectral lines detection with cross-correlation or chi-square fitting to inject prior knowledge about more plausible redshift solutions.

Despite their overall performances in redshift estimation, most algorithms in use today still suffer from numerous modeling and computational deficiencies, as the major recurrent issues with the z_{spec} estimation algorithms remain the strong correlation between reliable spectral feature detection and the quality of the observed spectrum, the difficulty to define a representative set of reference templates, and the use of a pre-generated redshift grid Θ_z that might be beneficial for rapid and parallel processing but could induce a "bias" regarding the redshift space to probe.

In galaxy surveys, a key issue often overlooked is the necessary evaluation of the quality of a redshift measurement because spectroscopic redshift measurement methods may be affected by a number of known or unknown observational biases that may produce some errors in the output redshift, ranging all the way to a catastrophic measurement far from the real galaxy redshift. Further, despite the general trend that consists in linking the reliability of a redshift measurement to the S/N of detected spectral features, the noise in the data usually presents a strongly non-linear dependency on the flux spectrum for various reasons (e.g., the wavelength-dependency of the background flux), which makes the definition of a precise redshift reliability criterion even more difficult.

A number of previous faint galaxy surveys have adopted redshift reliability assessments, either by using empirical thresholds applied to a single metric operator (Baldry et al. 2014; Cool et al. 2013), or by combining independent reliability assessments performed by more than two experienced astronomers in order to smooth-out the observer bias of each individual and produce a remarkably repetitive reliability assessment (Le Fèvre et al. 2013, 2015; Garilli et al. 2014; Guzzo et al. 2014). All methods imply subjective information, either by selecting "adequate" thresholds from a constructed sample or by involving a human operator within the (visual) verification process that becomes largely unfeasible for samples over 10^5 galaxies. For massive spectroscopic surveys such as *Euclid* or WFIRST, there is a critical need for a fully automated reliability flag definition that will adapt to the observed data and display a greater use of all available information.

In this paper, we propose to exploit a Bayesian framework for the spectroscopic redshift estimation to incorporate all sources of information and uncertainties of the estimation process (prior, data-model hypothesis), and produce a full z_{spec} posterior PDF, that will be the starting point of our automated reliability flag definition.

To test the proposed methodology of assessing the redshift reliability, we use a new redshift estimation software called algorithms for massive automatic Z evaluation and determination (AMAZED) developed as part of the Processing function (PF-SPE) in charge of the 1D spectroscopic data-processing pipeline of the *Euclid* space mission.

The paper is organized as follows. After introducing the subject, we present the data used in this study in Sect. 2, and in Sect. 3 we describe the Bayesian formalism of the spectroscopic redshift estimation. Section 4 is focused on the proposed automated reliability assessment method, where we first describe the principle, then present preliminary results of

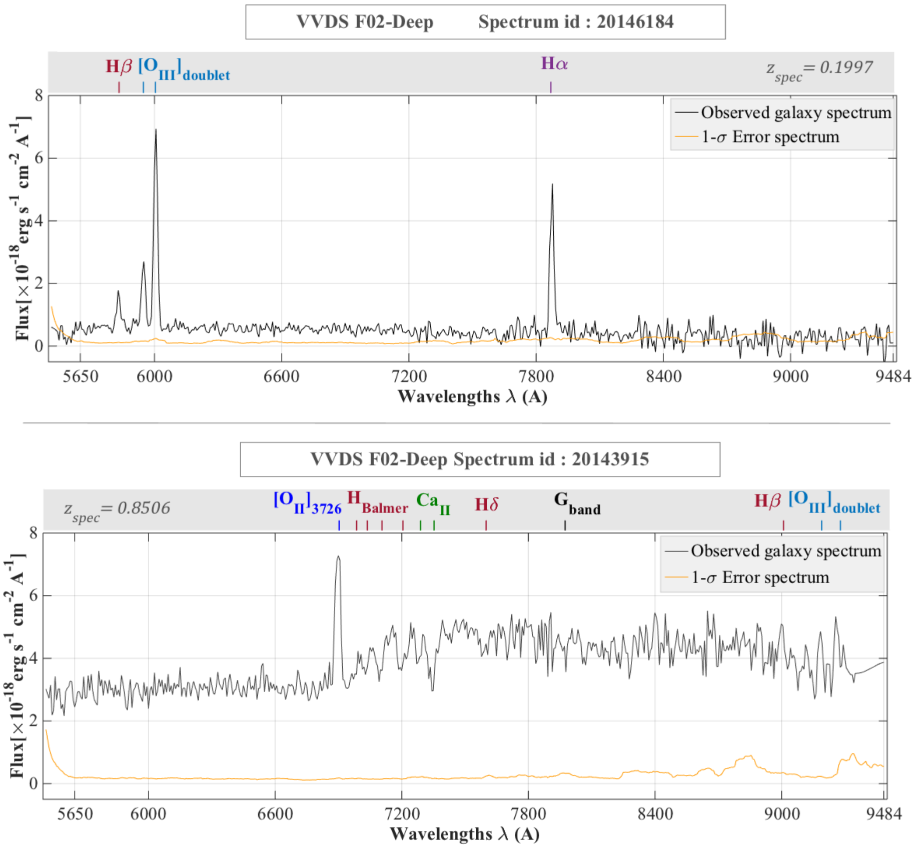

Fig. 1. VVDS-Deep galaxy spectra with identifiable spectral features at known redshifts: emission/absorption line, D4000A break.

supervised and unsupervised classification techniques using the public database of the VIMOS VLT Deep Survey (VVDS, Le Fèvre et al. 2013). In Sect. 5, we present our results of redshift reliability predictions using preliminary simulations of *Euclid* spectra covering a wavelength range [1.25−1.85] μm, and we finally conclude in Sect. 6.

2. Reference data

To test the proposed method of assessing a redshift reliability, we use public data from the VIMOS VLT Deep Survey[1] (VVDS) in this study. The large VIMOS VLT Deep Survey (Le Fèvre et al. 2013) is a combination of three i-band magnitude limited surveys: Wide ($17.5 \leq i_{AB} \leq 22.5$; 8.6 deg^2), Deep ($17.5 \leq i_{AB} \leq 24$; 0.6 deg^2) and Ultra-Deep ($23 \leq i_{AB} \leq 24.75$; 512 arcmin^2), that produced a total of 35 526 spectroscopic galaxy redshifts between 0 and 6.7 (22 434 in Wide, 12 051 in Deep and 1041 in UDeep) with a spectral resolution ($R \simeq 230$, dispersion 7.14A) approaching that of the upcoming *Euclid* mission ($R \geq 380$ for a 0.5″object, dispersion 13.4A) as illustrated in Fig. 1.

The VIMOS Interactive Pipeline and Graphical Interface (VIPGI) data-processing software included background subtraction, decontamination, filtering and extraction of 1D spectra from 2D spectral images using sophisticated packages (Scodeggio et al. 2005). The VIMOS 1D spectroscopic data was processed using the EZ software (Garilli et al. 2010) to compute spectroscopic redshift measurements and reliability flags by combining reliability assessments (visual checks) of at least two experienced astronomers (Le Fèvre et al. 2013). The VVDS project provides a reference sample with a range of redshifts and reliability flags well-suited for testing our methods in a broad parameter space.

To evaluate our automated redshift reliability assessment method (Sect. 4), we use the VVDS data in two stages. First we exploit the existing redshift reliability flags of the VVDS data as a reference to assess the performances of supervised classification algorithms in predicting a similar redshift reliability label. Then, after partitioning the VVDS data into distinct clusters of redshift reliability flags using unsupervised classification, we compare these results with the original VVDS redshift flags to evaluate the performances of the proposed methodology and unveil possible discrepancies.

3. Spectroscopic redshift estimation

3.1. Description

To derive a redshift, the widely used template-based algorithms rely on the hypothesis that “there exists a reference template spectrum that is a true (and sufficient) representation of the observed data”, implying that the observed spectrum can be described by at least one spectroscopic template of the reference library.

Using a set of rest-frame templates and a fixed grid of redshift candidates in Θ_z, for each pair (redshift z, template M_t) we compute the Least-Square metric:

$$\chi^2(z,t) = \sum_{i\in\Lambda} {\sigma_i}^{-2}(d_i - at_{i,z})^2, z \in \Theta_z, \tag{1}$$

or the cross-correlation:

$$xc(z,t) = \frac{1}{\sigma_d\,\sigma_{t,z}} \sum_{i\in\Lambda}(d_i - \mu_d)(t_{i,z} - \mu_{t,z}){\sigma_i}^{-2}, z \in \Theta_z, \tag{2}$$

where d_i and σ_i refer respectively to the observed flux and noise spectra at pixel i, $t_{i,z}$ is the redshifted template interpolated at

[1] http://cesam.lam.fr/vvds/

pixel i, and ($\mu_{t,z}$; $\sigma_{t,z}$) and (μ_d; σ_d) are the mean and standard-deviation of the redshifted template and the observed spectrum respectively. The wavelength range in use Λ contains n data-points, Θ_z refers to the redshift space to probe, and a is a scale factor referring to the amplitude of the redshifted template that is usually computed at each trial from (weighted) least-square estimation.

The estimated $z_{\rm spec}$ results from a joint-estimation of the pair (z, M_t) and is performed by optimizing a chosen metric: maximization of the cross-correlation function or minimization of the chi-square operator.

In general, the accuracy of the template-based methods is tied to the representativeness and wavelength coverage of the spectroscopic templates M_t in use.

3.2. Bayesian inference

Assuming a linear and Gaussian data model with i.i.d. (independent and identically distributed) residuals $\{N_i\}_{i\in\Lambda}$, the probability of observing the spectrum $\{D_i\}_{i\in\Lambda}$, at a redshift z given a template model M_t and any additional information I is described by the likelihood function $\mathcal{L}(z, M_t)$ (cf. Appendix A):

$$\begin{aligned}\mathcal{L}(z, M_t) &= p(D|z, M_t, I) = \prod_{i\in\Lambda} p(N_i|z, M_t, I)\\ &= \prod_{1\leq i\leq n} (\sqrt{2\pi}\sigma_i)^{-1} \exp^{-\frac{1}{2}\chi^2(z,t)}, \qquad (3)\\ \ell_{(z,M_t)} &= \log(\mathcal{L}(z, M_t))\\ &= -\frac{1}{2}\chi^2(z,t) - \frac{N}{2}\log(2\pi) - \sum_{i\in\Lambda}\log(\sigma_i). \qquad (4)\end{aligned}$$

Via the Bayes rule, the joint posterior distribution is:

$$p(z, M_t|D, I) = \frac{p(D|z, M_t, I) \times \pi(z,t)}{p(D|I)}, \qquad (5)$$

$$\begin{aligned}\log(p(z, M_t|D, I)) = &-\frac{1}{2}\chi^2(z,t) + \log(\pi(z,t))\\ &- \log\left(\iint_{z,M_t} \pi(z,t)\exp^{-\frac{1}{2}\chi^2(z,t)}\, \mathrm{d}z\, \mathrm{d}M_t\right) \qquad (6)\end{aligned}$$

where $\pi(z,t)$ is the joint-prior distribution of the pair (z, M_t).

The 1D posterior distribution is obtained by marginalizing over M_t:

$$p(z|D, I) = \int_{M_t} p(z, M_t|D, I)\, \mathrm{d}M_t. \qquad (7)$$

The “best” redshift $\widehat{z}_{spec}$ is the MAP (Maximum-A-Posteriori) estimate:

$$z_{\rm MAP} = \mathrm{argmax}_z\ p(z|D, I). \qquad (8)$$

This Bayesian formalism was not clearly stated for the spectroscopic redshift estimation. As for now, a posterior $z_{\rm spec}$ PDF can be computed and prior information, if available, can easily be integrated.

Furthermore if the hypothesis of the datamodel is readjusted, the equations can be rapidly and accurately revised in the likelihood expression (cf. Appendix A).

The template library used in this study includes a set of 9 continuum spectra of spiral, elliptical, starburst, and bulge galaxies, supplemented with 12 templates displaying different shapes and level for the continuum and the emission lines that were built by the VVDS team to take into account the diversity of galaxy spectra observed during the survey.

The spectroscopic templates that had only optical data were extended in the UV down to 912A by exploiting the closest templates with UV data, and below 912A by using nul flux spectra. In the infrared, a blackbody continuum was used to extrapolate the templates up to 20 000A.

This large wavelength coverage ensures that the intersection between the observed spectra and the templates is verified at each redshift trial.

3.3. Numerical computation

In the Bayesian inference, if our state of knowledge about a certain quantity θ is vague, a non-informative prior, such as the flat prior, is usually computed.

$$\int_{\Delta\theta} p(\theta|data)\, \mathrm{d}\theta = 1. \qquad (9)$$

Using a flat prior for redshift estimation implies that all redshifts and all templates are viewed as equiprobable solutions. The estimation algorithm will explore the full template library and the entire redshift grid and compute a (marginalized) posterior redshift PDF as displayed in Fig. 2.

If extra information about the pair (z, M_t) is available, the joint prior will be more informative as it will display a refined structure in the (z, M_t) space. For example, to estimate photometric redshifts, integrated flux in filters, colour, or magnitude can be used as priors to efficiently probe the redshift space. In Benitez (1999), the joint-prior $p(z, T|m_0)$ provides additional information about the most eligible spectral objects, T, with a magnitude, m_0, that could be observed at certain redshifts, z. However, for spectroscopic redshift estimation, there is no clear definition of a (data-independent) prior, a choice justified by the fact that spectroscopic data is more informative than photometry.

4. Reliability assessment

As the size of massive surveys in astronomy continues to expand, assessing redshifts’ reliability becomes increasingly challenging. The need for fully automated reliability assessment methods is now part of the requirements for future surveys, and is justified by the fact that automation provides predictable and consistent performances while the behavior of a human operator remains unpredictable and often inconsistent and therefore can require several independent observers to smooth out personal biases.

Moreover, the need for automation comes from the orders-of-magnitude increase in the total number of spectra that need to be processed. Visual examination of all spectra in a survey (2dF, DEEP2, VVDS, VIPERS, zCOSMOS, VUDS, PRIMUS, etc.) is extremely difficult for samples containing 10^5 objects or more, and will be completely impossible for next-generation spectroscopy surveys with more than 50×10^6 objects.

In general, existing approaches to automate the reliability assessment as well as the associated quality control in most engineering applications, such as the intrusion detection systems (IDS) that aim to evaluate the traffic quality by identifying any malicious activity or policy violation within a network, include:

1. Anomaly detection systems (ADS), where a component is labeled as an outlier if it deviates from an expected behavior using a set of thresholds or reference data (Chandola et al. 2009; Patcha & Park 2007). The ADS usually proceed by monitoring the system activity and detecting any sort of violation based on specific criteria or invariable standards.

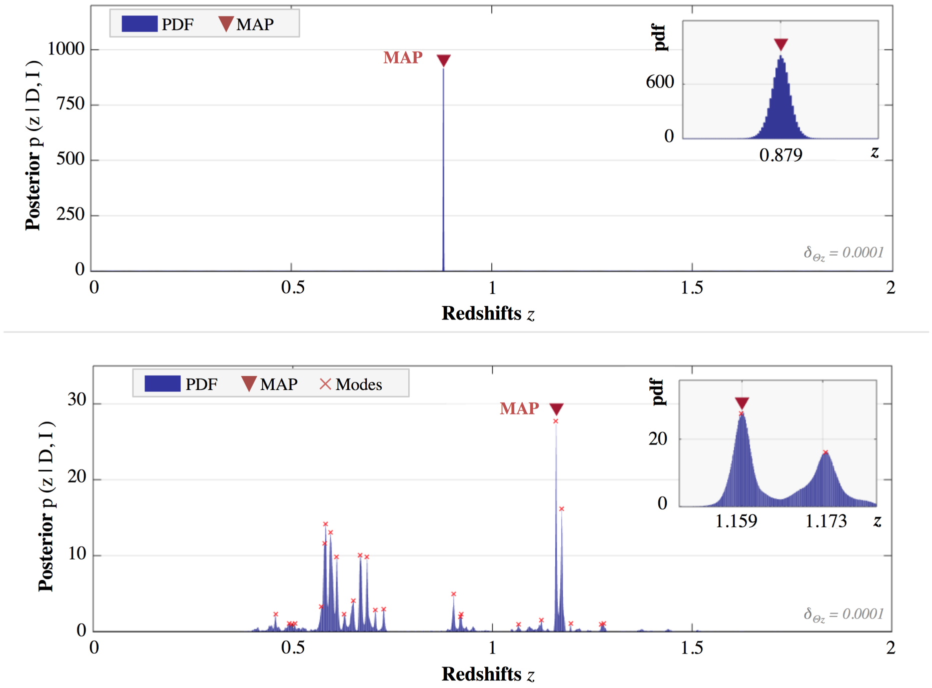

Fig. 2. Posterior redshift PDF of two VVDS-Deep spectra. *Top*: a unimodal zPDF characterizes a very reliable redshift measurement (a single peak at z=0.879). *Bottom*: a multimodal zPDF refers to multiple redshift solutions (multiple peaks) possibly with similar probabilities associated to a diminished confidence level of the MAP estimate at $z = 1.159$. The quantity δ_{Θ_z} refers to the fixed step of the redshift grid used to compute the zPDFs.

2. Supervised classification that exploits prior knowledge of a referenced training set to predict a label (Shahid et al. 2014).

Both methods deliver great performances in general, but still have some limitations: irrelevant thresholds to new data for the ADS, and poor representativity of the training set in classification, and so on.

To automate the redshift reliability assessment, reproducing the ADS reasoning scheme by setting empirical thresholds might not be the best option when dealing with massive surveys. However, the use of machine learning (ML) techniques can still be a viable option but first requires the search for a valid model and a coherent set of entries.

In this work, the method to automate the redshift reliability flag definition stems from an attempt to address questions about the meaning of a "reliable" redshift:

1. What guides an experienced astronomer to declare an estimated redshift as a plausible solution; apart from visual inspection of the data and its fitted template?
2. Is there some disregarded information within the z-estimation process that we can further exploit?
3. How can a system "perceive" the same information as a human does?

Spectroscopic redshift measurements are obtained from χ^2 minimization or maximization of the posterior probability $p(z|D, I)$ in Bayesian inference (cf. Sect. 3), and usually no further analysis of the computed functions is conducted afterwards. When computing the posterior redshift PDF, broadly two types of probability density function can be observed (cf. Fig. 2): a unimodal PDF versus a multimodal distribution. In both cases, a pipeline will provide a redshift estimation z_{MAP} but the estimated redshifts from these two different types of PDFs definitely do not show the same level of reliability. In fact, the multimodal PDF refers to numerous redshift candidates possibly with similar probabilities, while a strong unimodal PDF with a prominent peak and low dispersion depicts a more "reliable" redshift estimation of the data.

We exploit such characteristics of the posterior PDF to build a discretized descriptor space that will be the entry point for ML techniques to predict a reliability label. Our approach aims to build the "experience" of an automated system in order to assess the quality of a redshift measurement from the zPDF.

4.1. Description

In machine learning, the typical entries of the model are a response vector $\boldsymbol{Y}$ and a feature matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_M \end{pmatrix} = \begin{array}{cc} & \begin{array}{ccc} d_1 & & d_P \end{array} \\ \begin{array}{c} s_1 \\ \\ \\ s_M \end{array} & \begin{pmatrix} x_{11} & \cdots & x_{1P} \\ \vdots & \ddots & \vdots \\ x_{M1} & \cdots & x_{MP} \end{pmatrix} \end{array} ; \; \boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_M \end{pmatrix}, \tag{10}$$

where $\boldsymbol{x}_j = (x_{j,1} \ldots x_{j,P})$ is the P-dimensional feature vector of the jth observational data $(s_j)_{j\in\{1,\ldots M\}}$, and y_j is its response variable.

If the response vector $\boldsymbol{Y}$ of the model is unknown, the prediction of a label y_j using only the distribution of $\mathbf{X}$ in the feature space refers to clustering (unsupervised classification). Otherwise, we talk about supervised classification whose goal is to define a mapping between the observable entries $\mathbf{X}$ and their associated response variables $\boldsymbol{Y}$ through a dual training/test scheme.

In ML, the design of the entry model is decisive. What could be the optimal selection of informative and independent features to accurately describe the zPDF? Can a single operator, such as the integral under the redshift solution z_{MAP} or the difference in probability between the first two peaks (modes), be a unique and sufficient descriptor? No definite answers can be given, since this approach of "quantifying the spectroscopic redshift reliability" from the zPDF is new. Each set of selected features will define a different descriptor space that a classifier could separate differently.

In this study, our selected ML entries are redshift reliability flags ($\boldsymbol{Y}$) and descriptors of the zPDFs ($\mathbf{X}$), where the feature vector $\boldsymbol{x}_j$ associated to the observation $s_j = \{D\}$ consists of a list of eight tailored descriptors of the zPDF:

- The quantity $P(z_{\mathrm{MAP}}|D, I) \approx p(z_{\mathrm{MAP}}|D, I) \times \delta_{\Theta_z}$, where δ_{Θ_z} is the fixed step of the redshift grid.
- The number of significant modes in the PDF. The "significance" of a mode is determined by partitioning the set of detected peaks of the PDF into two categories (strong/weak) based on their prominence and height in order to avoid including the extremely-low density peaks (10^{-100} usually) that result from the conversion of logPDFs into a linear scale.
- The difference in probability of the first two best redshift solutions (z_{MAP}, z_2): $P(z_{\mathrm{MAP}}|D, I) - P(z_2|D, I)$.
- The dispersion $\sigma = [\int(z - \bar{z})^2 p(z)\mathrm{d}z]^{1/2}$, with $\bar{z} = \int z p(z)\mathrm{d}z$.
- The cumulative probability in the region R_2^*: $[z_{\mathrm{MAP}} \pm \delta]$ where the parameter δ is chosen equal to 0.001.
- The characteristics of the CR*(restricted version of the credibility region with 95% in probability): number of z candidates, width Δz, cumulative probability.
  In Bayesian Inference, the CR is analogous to the frequentist CI (confidence interval).
  For a $100(1 - \alpha)\%$ level of credibility, the CR is defined as: $\int_{CR} p(z|D, I)\mathrm{d}z = 1 - \alpha$.
  The restricted CR* used sets (optional) maximal bounds to the search region around z_{MAP} to accelerate the operation.

Displays of distinct zPDFs are presented in Figs. 3–6, where the descriptors listed above highlight interesting features of the zPDFs. For example, it is possible to obtain a similar dispersion for two zPDFs but a different number of significant redshift modes (Fig. 3), or the other way around: multimodal zPDFs with a comparable number of redshift modes with different amplitudes, and a different dispersion (cf. Fig. 4) or difference in probability between the first two best redshift solutions (cf. Fig. 5). Also, unimodal zPDFs can vary as they can display wider or narrower restricted CR (cf. Fig. 6) or different values of the dispersion σ.

Using the eight listed key descriptors, we estimate that the main features of the zPDF can be inferred. This design is not immutable. Supplementing the feature matrix with additional information about the observed spectra, s, or designing a different feature selection can also be explored.

4.2. Classification

4.2.1. Model

The ML entries in this study are obtained from a collection of zPDFs computed from M spectra of the VVDS to which a reliability label $(y_i)_{i\in\{1,\ldots M\}}$ is known to belong to one of the flags (Le Fèvre et al. 2005, 2013):

- flag 1, "Unreliable redshift";
- flag 2, "Reliable redshift";
- flag 9, "Reliable redshift, detection of a single emission line";
- flag 3, "Very reliable redshift with strong spectral features";
- flag 4, "Very reliable redshift with obvious spectral features".

The redshift reliability flags in the VVDS are determined by confronting independent redshift measurements performed by several observers on the same spectra.

By comparing the redshift measurements with internal duplicated observations or with published redshifts from different surveys, the VVDS spectroscopic redshift flags have been empirically paired with a probability for "a redshift to be correct": the VVDS redshift reliability flags $\{1, 2, 9, 3, 4\}$ are associated with probabilities of [50–75]%, [75–85]%, ~80%, [95–100]%, and 100% , respectively, that the measured redshifts are correct.

Using supervised classification, the objective is to predict similar redshift reliability flags for new unlabeled data. However, since the reproducibility of the VVDS redshift reliability flags is difficult because of their subjective definition and the confusion between "quality of a redshift" and "specific information about the data", we first decided to regroup the VVDS flags, as following:

- "Class 0", consisting of the "VVDS flags 1" to depict the uncertain redshifts.
- "Class +1", consisting of the "VVDS flags 2–9" to depict the reliable redshifts.
- "Class +2", consisting of the "VVDS flags 3–4" to depict the very reliable redshifts.

A three-class classification problem is then set. For multi-class problems, the error-correcting-output-codes (ECOC), as introduced in Dietterich & Bakiri (1995), are adapted for several learners, such as support vector machines (SVM), Tree templates, and Ensemble classifiers. A description of the ECOC is provided in Appendix B.

4.2.2. Preliminary tests

Classification tests are conducted using a VVDS subset of 24519 spectra with a constraint on the redshift accuracy $|z_{\mathrm{MAP}} - z_{\mathrm{ref}}|/(1 + z_{\mathrm{ref}}) \leq 10^{-3}$ for the VVDS flags $\{2, 9, 3, 4\}$. Our main objective is to build a descriptor space from a diverse set of zPDFs and evaluate the ability of the system to predict a redshift reliability label.

The dataset is decomposed into a "Training set" and a "Test set" (cf. Tables 1–3).

Different classifiers are tested in this study to carry out a careful analysis and avoid blindly trusting the results in cases of overfitting. We assess that different techniques should provide a different but not very disparate level of performance. Three classifiers are selected: the SVMs with linear and Gaussian kernels, an ensemble of bagging trees (referred to simply as TreeBagger) and a GentleBoost ensemble of decision trees. A general description of the classifiers and the multi-class measures is provided in Appendices C and D.

To evaluate the performance of a classifier, two tests are conducted:

- test 1: resubstitution;
- test 2: test prediction.

In the resubstitution, the "Training set" is reused as the "Test set" during the prediction phase. Extremely low prediction errors are expected ($\lesssim$1% classification error rate): if a bijective relation exists between the observables $\mathbf{X}_{\mathrm{train}}$ and the response vector $\boldsymbol{Y}_{\mathrm{train}}$, the generated mapping from the training phase is supposedly accurate. The predicted labels $\boldsymbol{Y}_{\mathrm{pred}}$ in resubstitution tests are therefore expected to resemble the true labels $\boldsymbol{Y}_{\mathrm{train}}$ with high accuracy, otherwise a clear mismatch between the features matrix $\mathbf{X}$ and the response vector $\boldsymbol{Y}$ of the ML model is reported. In such a case, the predictions of the second test ("Test prediction") would be baseless, since the mapping produced from

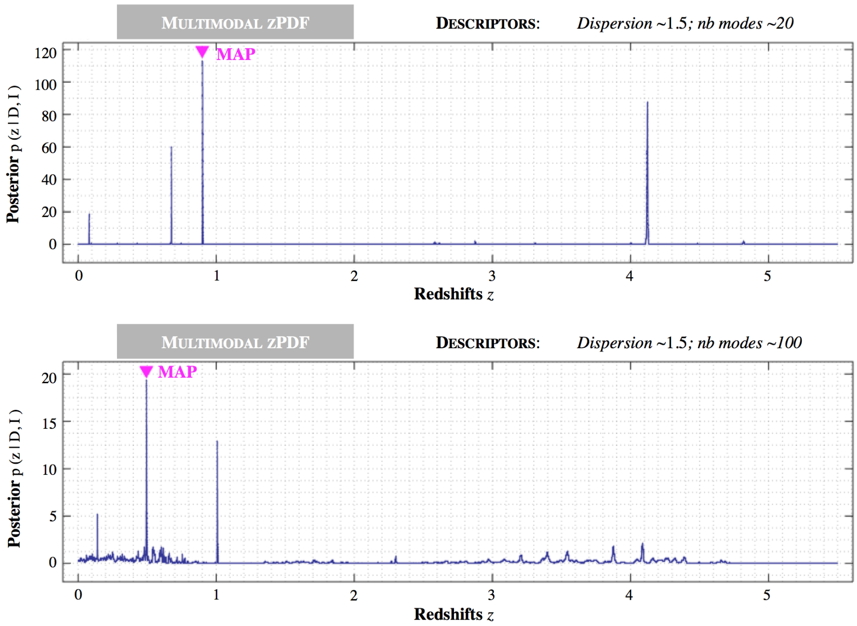

Fig. 3. Display of two zPDFs with multiples modes and a similar dispersion.

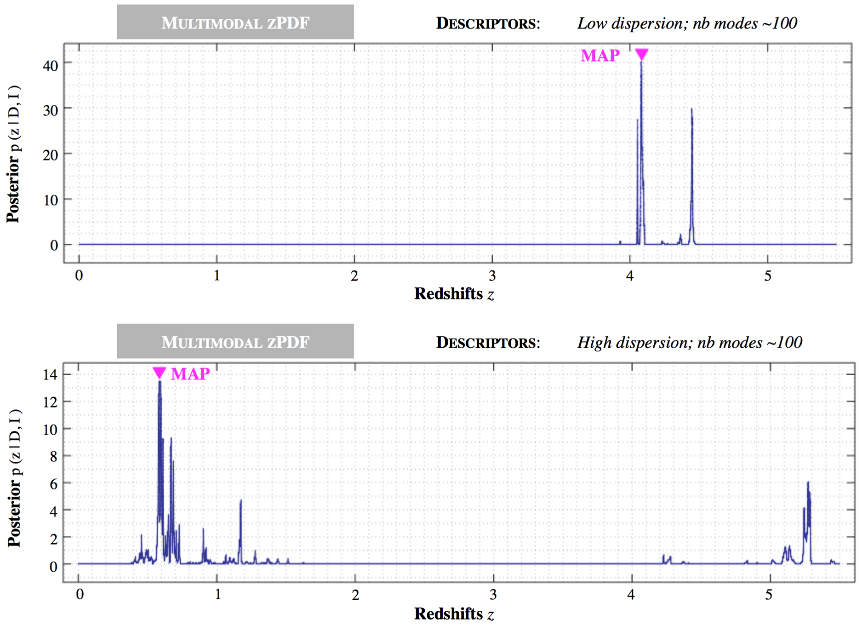

Fig. 4. Display of two zPDFs with multiples modes and a different dispersion.

the training phase is truly unusable. The overall performances reported in Tables F.9 and F.10 in addition to the confusion matrices (cf. Tables F.1 to F.8) representing the fraction of the predicted labels versus the true classes in $\boldsymbol{Y}_{\text{test}}$, support this conclusion. Most classifiers seem unable to predict the true labels in resubstitution: non-zero off-diagonal elements in the matrices and a high error-rate, implying that a correct mapping between the feature matrix and the existing VVDS redshift reliability flags cannot be produced.

We would like to point out the singular case of the TreeBagger that seems to generate a good mapping in resubstitution (error rate 0.08% on average) in comparison with the SVMs that are commonly-known as robust classifiers (error rate >10% in average). It seems reasonable to consider that the observed dissimilarity between the different classifiers in resubstitution is due to the sensitivity of the bagging trees to several parameters as the number of learners or the trees depth that can coerce the training into focusing on irregular patterns and establish an erroneous mapping). As anticipated from the resubstitution results (high error rate), we also find that the test predictions present a significant error rate (~40% on average).

To summarize, these first results of supervised classification show that trying to match the subjective VVDS flags with descriptors of the zPDF gives poor results.

The entries and hypotheses for ML have to be reexamined.

4.3. Clustering and fuzzy classification

From the previous results, doubts can be raised regarding the engineered zPDF feature space derived from a collection of 24519 VVDS spectra. However the selected set of descriptors seems to be a viable description that portrays an existing but hidden structure of the feature space.

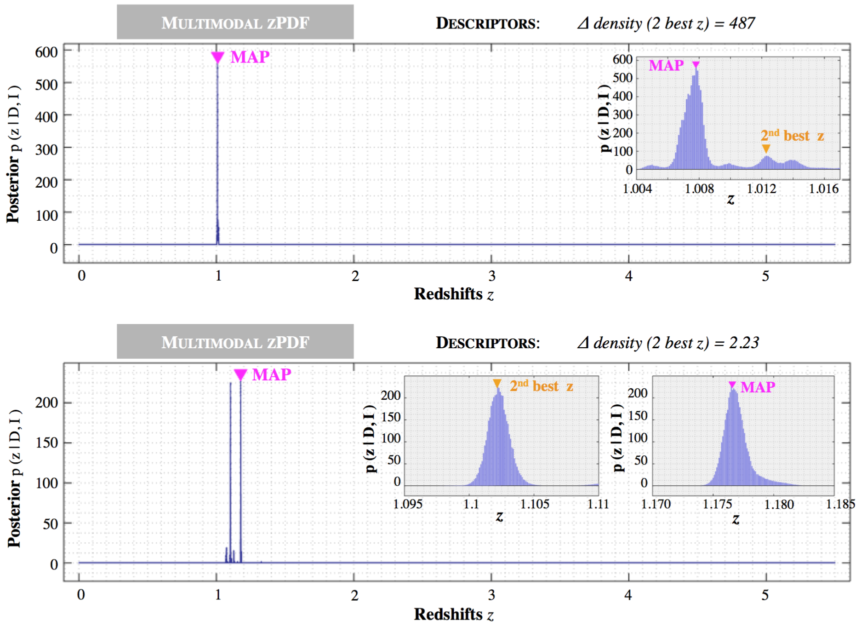

Fig. 5. Display of two zPDFs with multiples modes and different characteristics of the two best z solutions.

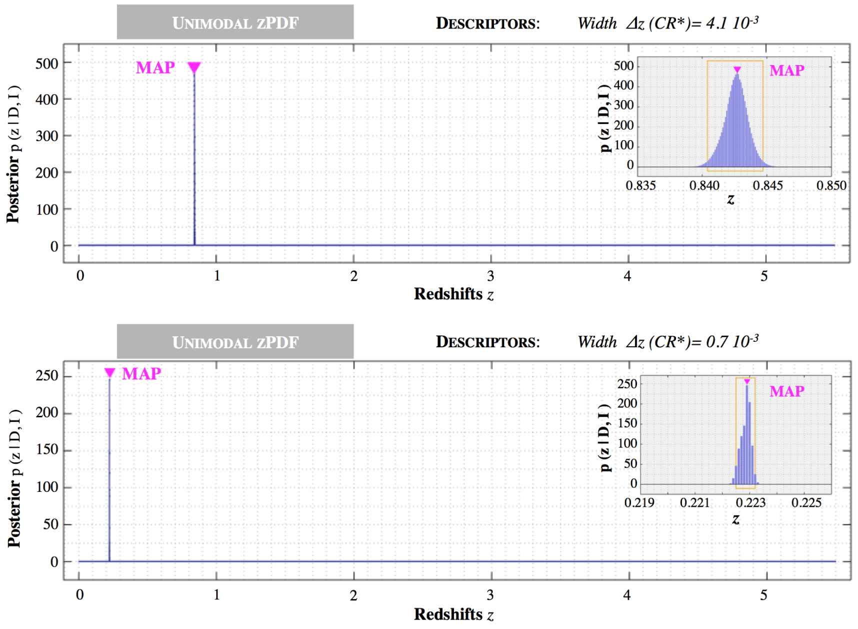

Fig. 6. Display of two zPDFs with a single mode and a different width of CR*.

Table 1. Description of the VVDS dataset used in this study.

Type	z reliability flags		Counts	z range
Primary objects	"Unreliable"	1	6768	0.0070–5.2280
	"Reliable"	9	632	0.0195–4.9285
	"Reliable"	2	4743	0.0017–4.4345
	"Very reliable"	3	6455	0.0266–4.5400
	"Very reliable"	4	5921	0.0213–3.8352

Clustering, known as unsupervised classification, is used in this Section to unveil the intricate structure and bring into light some properties of the data in the descriptor space.

4.3.1. Partitioning the descriptor space

In unsupervised classification, prior knowledge about class membership is unavailable. Partitioning the descriptor space into K manifolds is realized by applying separation rules only to the feature matrix **X**.
By representing the zPDFs feature matrix **X** in 3D (cf. Fig. 7), a simple bi-partitioning is introduced:

- Group 1: high dispersion and low $P(z_{\mathrm{MAP}}|D, I)$ referring to multimodal PDFs or platykurtic unimodal PDFs.
- Group 2: medium dispersion and high $P(z_{\mathrm{MAP}}|D, I)$ depicting strongly peaked unimodal PDFs.

In each category, we choose to reapply a bi-partitioning to decompose the data into a dichotomized pattern (cf. Fig. 8). This partitioning strategy, applied to the entire descriptor components and not only to the two descriptor components as in

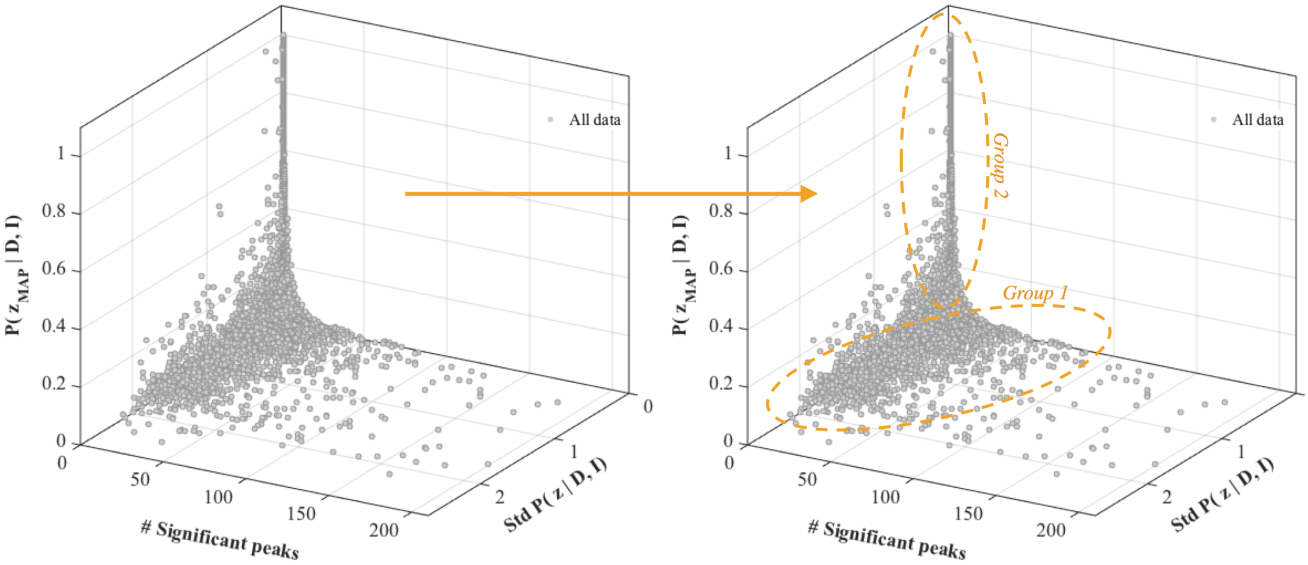

Fig. 7. 3D representation of the feature matrix **X**. Broadly, two categories are noticeable. The first group refers to the zPDFs with high dispersion, large $z_{\rm MAP}$ peak and low probability $P(z_{\rm MAP}|D, I)$ that can be assimilated to multimodal PDFs or platykurtic unimodal PDFs. The second group characterizes the zPDFs with medium-to-low dispersion, narrow $z_{\rm MAP}$ peak and high $P(z_{\rm MAP}|D, I)$ that depict strongly peaked unimodal PDFs.

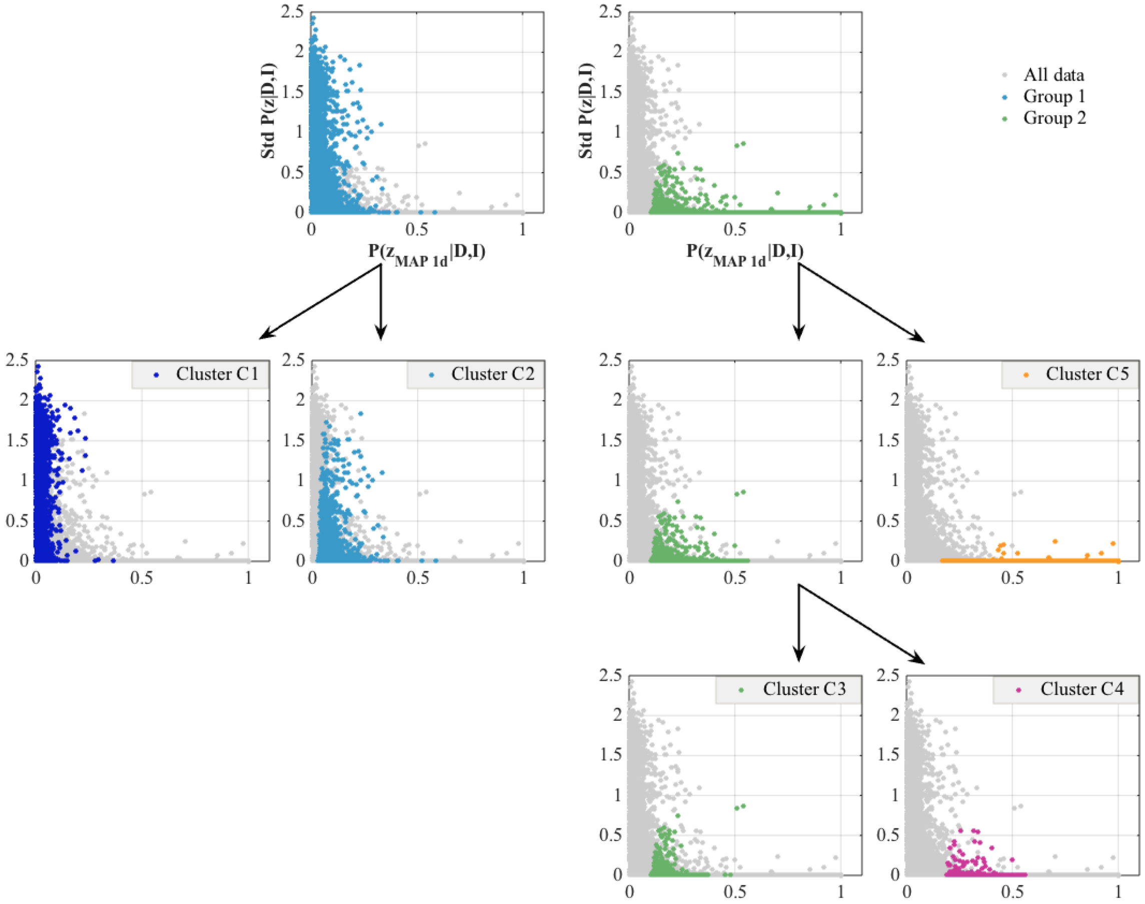

Fig. 8. Clustering the zPDFs features **X** in a dichotomized pattern. The clustering strategy exploits the classic FCM algorithm at each step to decompose the input data into two sub-classes using the entire set of descriptors. The final categories $\{C_k\}_{k\in\{1,2,3,4,5\}}$ are displayed in distinct colors.

the displays, alongside with the number of clusters, the feature selection and the ML algorithms tested in this work as a novelty to automate the redshift reliability, are not immutable and can be readjusted according to the data in hand. Further evaluations will be conducted on these aspects of ML to develop a robust and precise automated assessment of redshift reliability.

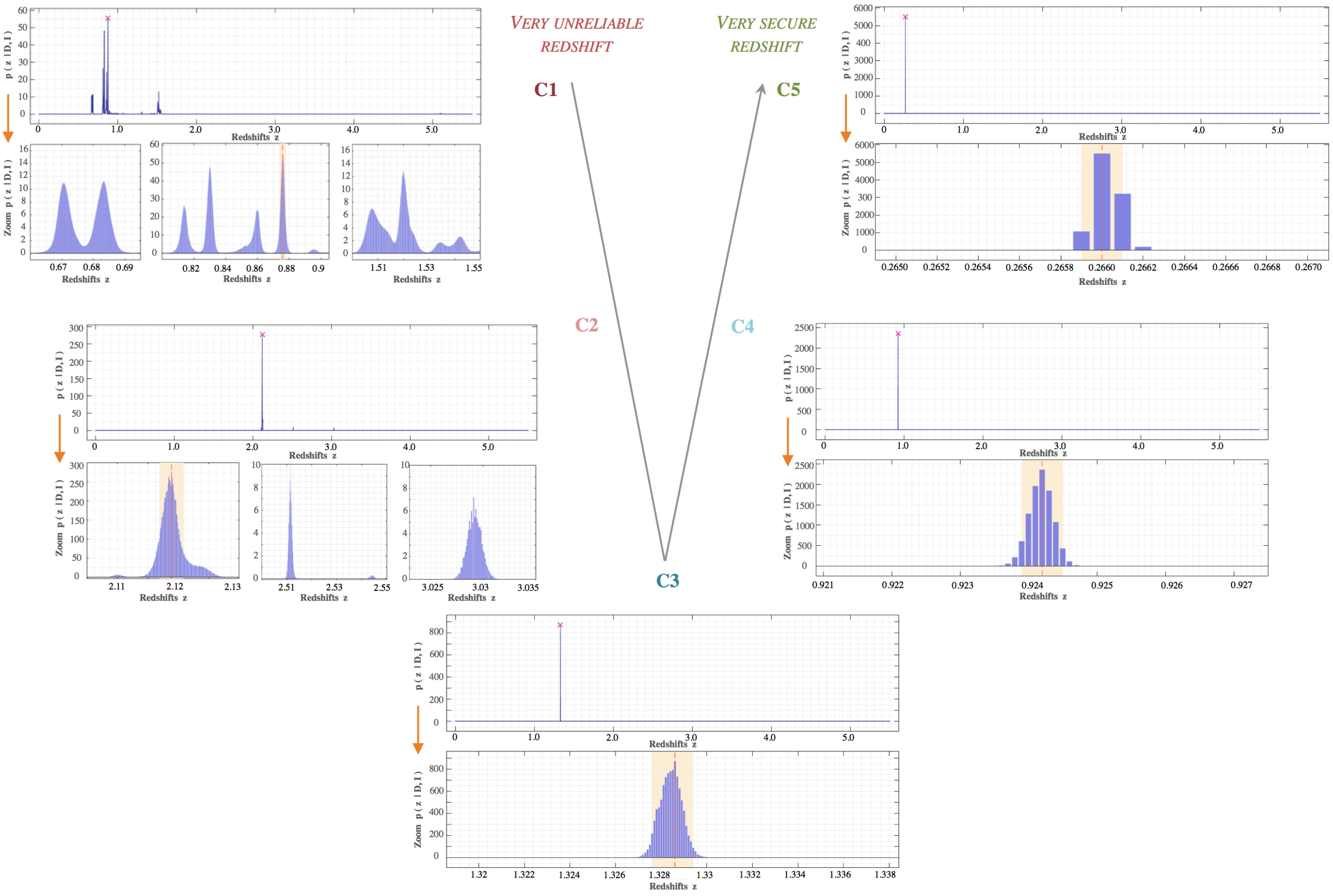

Fig. 9. Representative zPDFs for each clusters. Display of representative zPDFs in each cluster obtained from clustering. The shift in the confidence level in the $\{C_k\}_{k\in\{1,2,3,4,5\}}$ clusters is apparent in the type of the zPDF: from multimodal zPDFs to unimodal zPDFs with narrower z_{MAP} peaks, the confidence level ranges from "extremely unreliable redshift estimate" ($C1$) to "very certain redshift estimate" ($C5$).

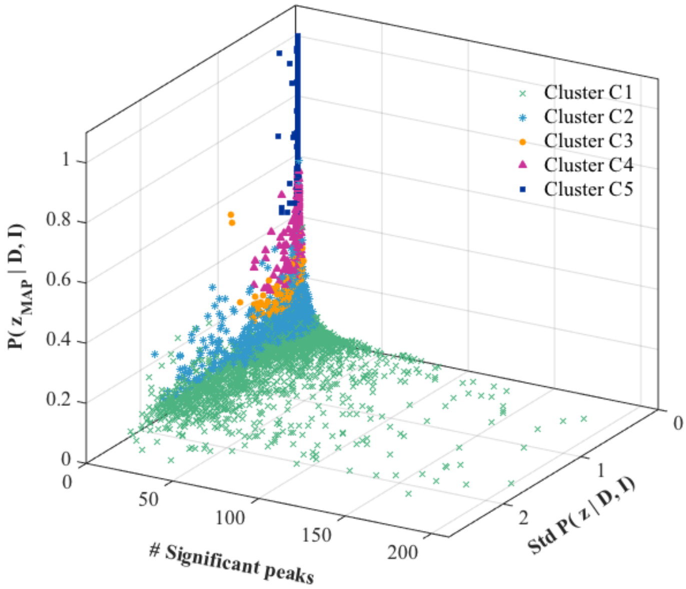

Fig. 10. Clusters repartition in a selected 3D space. The five zReliability clusters described in Sect. 4.3.1 are associated to different types of redshift PDFs, where the two extreme categories C_1 and C_5, respectively, describe highly dispersed multimodal zPDFs and peaked unimodal zPDFs.

Table 2. Training set (total of 16346 VVDS spectra).

	Label	Counts	%
Train set	"0"	4512	27.60
	"+1"	3583	21.92
	"+2"	8251	50.48

Table 3. Test set (total of 8173 VVDS spectra).

	Label	Counts	%
Test set	" 0 "	2256	27.60
	"+1"	1792	21.93
	"+2"	4125	50.47

Using the classic clustering algorithm FCM (Fuzzy C-Means) to minimize the intraclass variance (cf. Appendix E), the final groups identify distinct partitions in the feature space (cf. Figs. 9 to 11). In this study, the selection of the number of clusters is an empirical process based on the analysis of the intermediate partitions and testing different configurations. We assess that the final architecture is a viable solution amongst others.

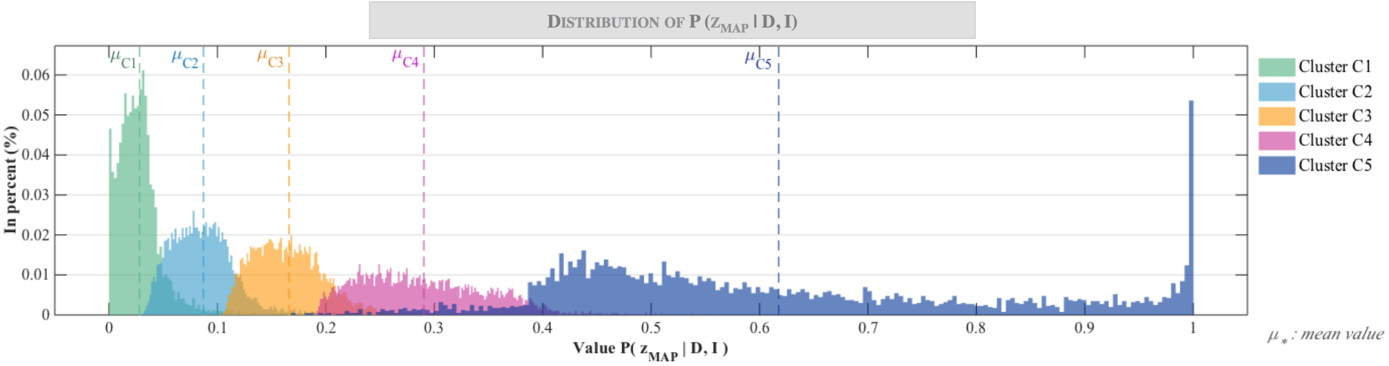

Fig. 11. Distribution of the probability value $P(z_{\mathrm{MAP}}|D, I)$ within the five partitions. The distribution of the probability values (component of the descriptor space) $P(z_{\mathrm{MAP}}|D, I)$ show distinct properties of the clusters. The five zReliability clusters (cf. Sect. 4.3.1) are associated with different categories of redshift PDFs.

- **"Cluster** $C1$**"**: highly dispersed PDFs with multiple equiprobable modes, $P(z_{\mathrm{MAP}}) \sim 0.028 \pm 0.023$.
- **"Cluster** $C2$**"**: less dispersed PDFs, with few modes and low probabilities $P(z_{\mathrm{MAP}}) \sim 0.087 \pm 0.033$.
- **"Cluster** $C3$**"**: low σ, intermediate probabilities $P(z_{\mathrm{MAP}}) \sim 0.166 \pm 0.035$.
- **"Cluster** $C4$**"**: unimodal PDFs with low dispersion, higher probabilities $\mathrm{P}(z_{\mathrm{MAP}}) \sim 0.290 \pm 0.059$.
- **"Cluster** $C5$**"**: strong unimodal PDFs with extremely low dispersion, better probabilities $P(z_{\mathrm{MAP}}) \sim 0.618 \pm 0.204$.

The coordinates of the clusters' centroids in the descriptor space are reported in Table 4.

$$\begin{cases} \text{Class } \{C_k\} \text{ centroid } \boldsymbol{g}_k = \frac{1}{M_k} \sum_{y_j \in C_k} \boldsymbol{x}_j \\ \text{Class } \{C_k\} \text{ variance } \boldsymbol{V}_k = \frac{1}{M_k} \sum_{y_j \in C_k} (\boldsymbol{x}_j - \boldsymbol{g}_k)^\top (\boldsymbol{x}_j - \boldsymbol{g}_k), \end{cases} \tag{11}$$

where M_k is the number of elements in cluster C_k.

Tables 5 and 6 report the intraclass dispersion $\sqrt{W}$ and the interclass dispersion $\sqrt{B}$ that characterize the newly defined clusters:

$$\begin{cases} \text{Interclass variance } \boldsymbol{B} = \frac{1}{M} \sum_k M_k (\boldsymbol{g}_k - \boldsymbol{g})^\top (\boldsymbol{g}_k - \boldsymbol{g}) \\ \text{Intraclass variance } \boldsymbol{W} = \frac{1}{M} \sum_k M_k \boldsymbol{V}_k, \end{cases} \tag{12}$$

with the total variance:

$$\boldsymbol{V} = \frac{1}{N} \sum_{j=1}^{M} (\boldsymbol{x}_j - \boldsymbol{g})^\top (\boldsymbol{x}_j - \boldsymbol{g}) = \boldsymbol{B} + \boldsymbol{W}, \tag{13}$$

where $\boldsymbol{g}$ is the global centroid and M is the full number of elements in the descriptor space.

The variance tables show that the intraclass variance, $\boldsymbol{W}$, is generally small in comparison to the interclass variance, $\boldsymbol{B}$, except for two descriptors (the dispersion and the number of modes). This results from the fact that the cluster $C1$ allows wider variations for these two components. Since the class $C1$ refers by definition to multimodal zPDFs associated to very unreliable redshift measurement, the results remain coherent.

Given the possibility that the clustering results might be unreliable due to inherent computational limitations or an incorrect modeling of the descriptor space, the full content of each partition $(C_k)_{k\in\{1,2,3,4,5\}}$ is investigated. We find that, overall, the zPDFs within each class, C_k , verify the properties listed above. The newly defined partitions genuinely describe a homogeneous representation of the data in the feature space.

4.3.2. Cluster analysis

In this section, we compare the initial VVDS redshift reliability flags and the new clusters in order to point out peculiar cases of misclassifications: unexplained discrepancies between the manually attributed flags in the VVDS database and those resulting from the unsupervised classification (cf. Sect. 4.3.1).

Two examples of misclassification are reported in Figs. 12 and 13:

1. A misclassification of a "VVDS Flag 1: unreliable redshift estimation" as $C5$ (unimodal zPDF and very reliable $\widehat{z}_{\mathrm{spec}}$) is presented in Fig. 12. The misclassification is due to the mismatch between the flux spectrum and its noise component, where the latter seems very inadequate when considering the good quality of the data. A problem regarding the generation of the 1D data (flux & noise components) from the 2D → 1D extraction can be noted.
2. A different type of misclassification illustrated in Fig. 13, where a "VVDS flag 9: secure redshift estimation with an identifiable strong EL" is identified as $C1$ (for very multimodal zPDFs and extremely unreliable $\widehat{z}_{\mathrm{spec}}$). This discrepancy between the VVDS flag and the new label from clustering could be ascribed to an imprecise computation of the zPDF due to a lack of representative templates at the given redshift, or a biased evaluation of a human operator.

To evaluate the misclassification rate for the entire VVDS dataset used in this study, Tables 7 and 8 summarize the repartition of the initial VVDS flags $\{1; 2; 9; 3; 4\}$ within the predicted reliability clusters.

We find that:

- The green cells represent the "expected" behavior: the cluster $C1$ is mainly composed of the unreliable redshift "VVDS flags 1" (∼86%), while the majority of the "VVDS flags 4" are in $C4$/$C5$ (∼81%) and the "VVDS flags 3" are in $C3$/$C4$ (∼68%).
- The gray cells represent a "gray area": the clustering provides homogeneous partitioning in comparison with the VVDS flags, as it properly incorporates the full information

Table 4. Coordinates of the clusters' centroids in the descriptor space.

Selected descriptors	Class $\{C_k\}$ centroid				
	$C1$	$C2$	$C3$	$C4$	$C5$
Dispersion σ	0.524	0.049	0.005	0.002	5e-4
$P(z_{\mathrm{MAP}}\vert D, I)$	0.028	0.087	0.166	0.290	0.618
Card($z \in$ CR*)	24.06	20.98	11.16	6.45	3.16
Width $\Delta z \in$ CR*	2.3e-3	2.0e-3	1.0e-3	5.5e-4	2.2e-4
$\sum_i P(z_i \in \mathrm{CR}^*)$	0.387	0.890	0.957	0.964	0.978
$\sum_i P(z_i \in R_2^*)$	0.364	0.884	0.998	1.0	1.0
Significant peaks	107.89	2.30	1.27	1.05	1.00
ΔP(two "best" z)	0.013	0.049	0.130	0.274	0.743
Nb elements M_k	3156	6720	5677	4966	4030

Table 5. Intraclass dispersion in the descriptor space.

Selected descriptors	Class $\{C_k\}$ dispersion				
	$C1$	$C2$	$C3$	$C4$	$C5$
Dispersion σ	0.585	0.167	0.039	0.023	0.008
$P(z_{\mathrm{MAP}}\vert D, I)$	0.023	0.033	0.035	0.059	0.204
Card($z \in$ CR*)	7.36	5.22	2.01	1.31	1.37
Width $\Delta z \in$ CR*	7.4e-4	5.2e-4	2.0e-4	1.3e-4	1.4e-4
$\sum_i P(z_i \in \mathrm{CR}^*)$	0.191	0.093	0.006	0.010	0.016
$\sum_i P(z_i \in R_2^*)$	0.175	0.107	0.006	0.003	0.001
Significant peaks	718.76	1.87	0.50	0.23	0.03
ΔP(two "best" z)	0.030	0.045	0.063	0.071	0.245

Table 6. Class dispersions in the descriptor space.

Selected descriptors	Variance $\boldsymbol{V} = \boldsymbol{B} + \boldsymbol{W}$		
	$\sqrt{\boldsymbol{V}}$	$\sqrt{\boldsymbol{B}}$	$\sqrt{\boldsymbol{W}}$
Dispersion σ	0.2851	0.1709	0.2282
$P(z_{\mathrm{MAP}}\vert D, I)$	0.2131	0.1929	0.0905
Card($z \in$ CR*)	8.64	7.65	4.01
Width $\Delta z \in$ CR*	8.6e-4	7.7e-4	4.0e-4
$\sum_i P(z_i \in \mathrm{CR}^*)$	0.207	0.189	0.085
$\sum_i P(z_i \in R_2^*)$	0.223	0.207	0.084
Significant peaks	260.28	35.63	257.83
ΔP(two "best" z)	0.271	0.247	0.112

Table 7. Repartition of the initial VVDS redshift reliability flags within the predicted labels (in absolute values).

		VVDS initial flags					
		$F1$	$F9$	$F2$	$F3$	$F4$	*Total*
Clusters	$C1$	2776	39	233	85	23	3156
	$C2$	3023	252	2055	1169	221	6720
	$C3$	657	212	1534	2345	899	5647
	$C4$	241	104	750	2019	1852	4966
	$C5$	71	25	171	837	2926	4030
	Total	6768	632	4743	6455	5921	24 519

Table 8. Repartition of the initial VVDS redshift reliability flags within the predicted labels (in percent).

		VVDS initial flags				
		$F1$	$F9$	$F2$	$F3$	$F4$
Clusters	$C1$	41.0%	6.2%	4.9%	1.3%	0.4%
	$C2$	44.7%	39.9%	43.3%	18.1%	3.7%
	$C3$	9.7%	33.5%	32.3%	36.3%	15.2%
	$C4$	3.6%	16.5%	15.8%	31.3%	31.3%
	$C5$	1.0%	4.0%	3.6%	13.0%	49.4%

from the input data (cf. observed flux and its associated noise component).

We find that the "VVDS flags 2–9" in $C4$/$C5$ (~20% each) are associated with extremely bright objects with easily identifiable spectral features that make the estimated redshifts very secure.

On the other hand, the "VVDS flags 4" predicted in $C3$ (~15%) are associated to noisier spectra with scarce spectral features in comparison with the "VVDS flags 4" in $C5$. The redshift reliability level for these spectra is thereby diminished.

Similarly, the prediction of "VVDS flag 2" in $C2$ (~43%) is due to the degradation of data quality in comparison with the "VVDS flags 2" located in $C3$.

The main reason behind these discrepancies lies in having different observers conducting the redshift-quality checks, as each person has their own understanding of a "redshift reliability" depending on their experience and knowledge of objectively assessing whether a redshift is deemed a secure estimation or not.

- The red cells are associated with peculiar cases of "abnormal" zPDFs resulting from incorrect noise spectra and/or human misclassification. In particular, the 71 cases listed of "VVDS flag 1" in $C5$ result from a mismatch between the flux and noise components; the noise component seems extremely low considering the reduced data quality. Having very low noise components contributes to reinforce that the flux information depicts a real observation even when it is not the case. We obtain, finally, extremely peaked zPDFs that are predicted as $C5$.

  For the 23 spectra of "VVDS flag 4" in $C1$, 13 cases are related to highly dispersed multimodal zPDFs where a confusion between the oxygen emission line [OII]$_{3726\mathrm{A}}$ and Lyα is reported: both emission lines are strong candidates which gives at least two significant modes detected in the zPDF. Also, the fact that the associated peaks are very distant in the redshift space results in a high dispersion value, σ, of the zPDF. The prediction in $C1$ is highly driven by these characteristics. We also report four cases within these 23 spectra that are associated to low S/N spectra: an important noise component annihilates the confidence in the flux vector and therefore produces highly multimodal zPDFs predicted in $C1$. For the remaining six cases of "VVDS flag 4" in $C1$, they result from an excessively-high noise component that produces very degenerate zPDFs, also predicted in $C1$.

Table 9. Training set (total of 16347 VVDS spectra).

	Label	Counts	%
TRAIN SET	*C*1	2104	12.87
	*C*2	4480	27.41
	*C*3	3765	23.03
	*C*4	3311	20.25
	*C*5	2687	16.44

Table 10. Test set (total of 8172 VVDS spectra).

	Label	Counts	%
TEST SET	*C*1	1052	12.87
	*C*2	2240	27.41
	*C*3	1882	23.03
	*C*4	1655	20.25
	*C*5	1343	16.43

The main result from the cluster analysis is that existing redshift reliability flags cannot be reproduced with a 100% accuracy due to their subjective definition, however a general trend can be retrieved as the majority of the VVDS initial redshift flags can be described by one or two of the redshift reliability clusters $\{C_k\}_{k\in\{1,2,3,4,5\}}$.

4.3.3. Re-using the clusters for redshift reliability label predictions

– Classification tests

The clustering results showed a great coherency between the automated definition of redshift reliability labels using the zPDF's features matrix and our understanding of "a redshift reliability". The idea presented in this Section consists in re-using the new labels of the 24519 VVDS spectra as the response vector $\mathbf{Y}_{\rm train}$ in supervised classification, to train a classifier to predict redshift reliability labels for new unlabeled data. For this purpose, classification tests are performed using the Training and Test sets in Tables 9 and 10. The resubstitution and test predictions are also used to verify once again the accuracy of the partitioning and objectively assess whether the FCM dichotomized strategy produced "random results" or a "a true description of the zPDFs" in the descriptor space.

Similar performances are observed for several classifiers in resubstitution, with extremely low off-diagonal elements in the confusion matrices and an average per-class error rate $\lesssim$1% (cf. Tables F.11 to F.14, and Table F.19) for all four classifiers, which is a clear contrast with the results in Sect. 4.2. By having low resubstitution errors, the mapping is deemed a reliable reproduction of the input data, and the prediction of $\mathbf{X}_{\rm test}$ can be examined. We find in test predictions that the confusion matrices for several classifiers offer a good predictive power (average per-class error rate <2%), with the Linear SVM scoring slightly lower results (cf. Tables F.15 to F.18, and Table F.20).

– Fuzzy approach

In ML, two main approaches exist: "hard" partitioning where an object is said to belong to a unique class (binary membership), and "soft/fuzzy" partitioning where the membership of an object to a class is expressed in terms of a probability between 0 and 1 (Wahba 1998, 2002).

In the classification tests, "soft" partitioning is used to compute the posterior class prediction probability in order to evaluate the classifier predictive power. The class posterior probabilities $p(Label\,|\,\{C_1, C_2, C_3, C_4, C_5\}\,;Descriptors)$ are obtained by minimizing the Kullback-Leibler divergence (Hastie & Tibshirani 1998).

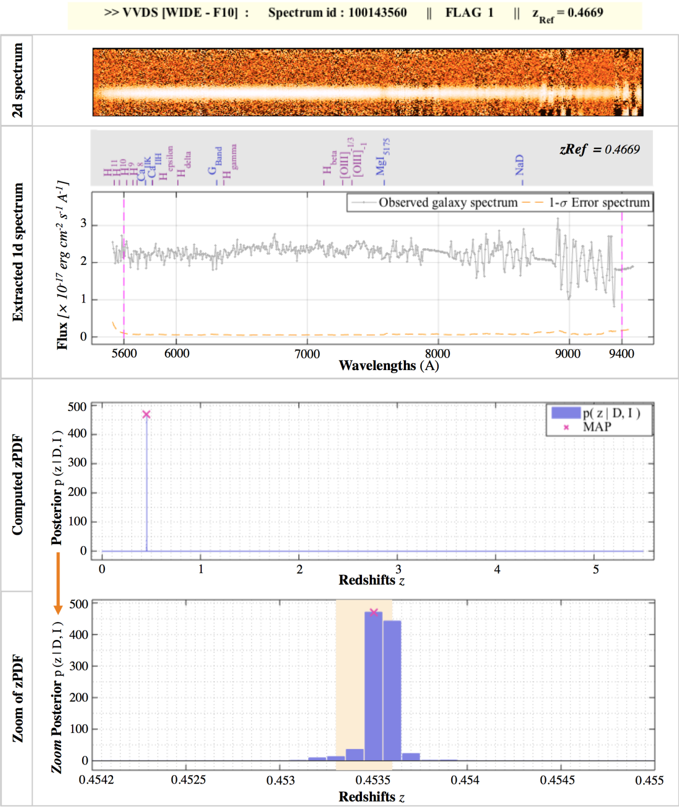

Fig. 12. Misclassification – case 1. A "VVDS flag 1: very unreliable redshift" is predicted by the classifier in the new category "*C*5" for very reliable redshifts. The spectrum displays very low noise components that reinforce the confidence in the measured flux pixels. However the extracted 1D spectrum appears distorted considering the initial 2D spectrum. The extraction 2D → 1D induced a bias in the estimation of the (falsely unimodal) zPDF.

In the test predictions on the evaluated VVDS dataset, we find that most class prediction probabilities fall between 0.7 and 1.0 (bright colors in Fig. 14). However, we estimate that it could be possible for new data to be assigned to a redshift reliability label with a lower probability, meaning that the classifier cannot project with certainty the unlabeled zPDF into the descriptor space as a result of an incorrect PDF (numerical limitations, degraded input spectra, etc.), or if it is located close to the margins of two or more clusters. For such cases, a new class of "Unidentified" objects has to be set apart from the labeled clusters $\{C_k\}_{k\in\{1,2,3,4,5\}}$. This particular point on the class prediction using soft partitioning is addressed further in the following section.

5. Tests on mock simulations for the *Euclid* space mission

An end-to-end simulation pipeline is currently under development for *Euclid* using catalogs of realistic input sources with spectro-photometric information and an instrumental model for the spectrophotometer NISP designed to perform slitless spectroscopy and imaging photometry in the near-infrared (NIR) wavelength domain. For *Euclid*, observations of the same field will be obtained from the combination of three or more different roll angles (referring to different orientations of the grisms) in

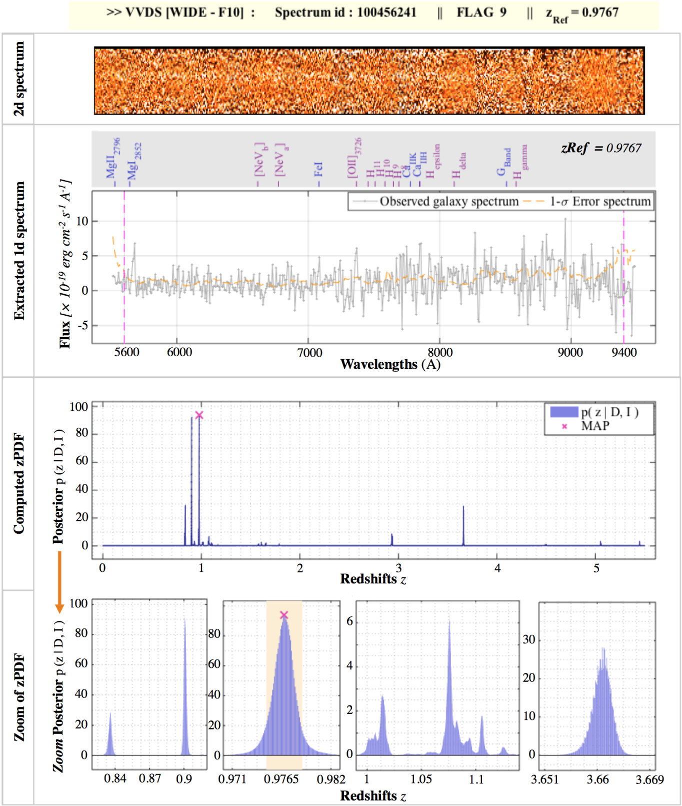

Fig. 13. Misclassification – case 2. A "VVDS flag 9: reliable redshift, detection of a single emission line" is predicted by the classifier in the new category "$C1$" for very unreliable redshifts. The spectrum displays a strong noise component that annihilates the confidence in the measured flux pixels (especially the spectral emission line $[\mathrm{OII}]_{3726}$ at 7365A). Several redshift solutions are declared as plausible solutions (a multimodal zPDF).

Table 11. Parameters of preliminary mock simulations for *Euclid*.

Properties	Redshift range	[0.95; 1.40]
	Magnitude J_{AB} range	[21.8; 24.5]
	Extinction $E(B-V)$	[0.00; 0.57]
	$\log(f\mathrm{H}\alpha)$ [erg s^{-1}cm^{-2}]	[–16.2; –14.1]
Simulator	Source size in arcsec, sigma	0.10 (set S1)
		0.50 (set S2)
	Sky background in e^{-}s^{-1}pix^{-1}	0.8 (set S1)
		2.0 (set S2)

order to alleviate the superposition of overlapping spectra due to the slitless mode.

Using the pixel simulator software TIPS (Zoubian et al. 2014), 1D spectra are obtained from 2D dispersed images after subtracting the sky background from the raw data and combining co-added image stamps of different roll angles. In these preliminary simulations for *Euclid*, a contamination model (zodiacal light, adjacent sources, etc.) is not included.

Table 11 reports the main characteristics of the simulated data of Hα EL galaxies at redshifts in the range $0.95 \leq z \leq 1.40$.

The *Euclid* simulations are not associated with a redshift reliability flag, and thereby are qualified as "unlabeled data" in this work. To test the performance of the redshift reliability assessment method, two sets of unlabeled spectra are used (S1, S2), with a total of 3169 spectra per set.

Table 12. zReliability predictions (in absolute values) of preliminary mock simulations for *Euclid*.

	Predictions in absolute values				
Set	"$C1$"	"$C2$"	"$C3$"	"$C4$"	"$C5$"
S1	3	61	313	835	1957
S2	383	1275	555	662	294

Table 13. zReliability predictions (in percent) of preliminary mock simulations for *Euclid*.

	Predictions in %				
Set	"$C1$"	"$C2$"	"$C3$"	"$C4$"	"$C5$"
S1	0.09	1.92	9.88	26.35	61.75
S2	12.09	40.23	17.51	20.89	9.28

By varying the source size and the sky background level, the difference in data quality between the two datasets is noticeable: Fig. 15 displays sample spectra for each dataset.

5.1. Reliability class predictions

The redshift PDFs of the *Euclid* simulated datasets are computed using a constant prior in Θ_z (cf. Figs. 16 and 17), and projected into the mapping (cf. Sect. 4.3) using soft partitioning to predict redshift reliability labels. Class prediction results are reported in Tables 12 and 13.

The system computes predominantly multimodal zPDFs with high dispersion when the useful information cannot be retrieved from the data because of low S/N: the estimated redshifts are deemed unreliable, which explains the high percentage of S2 spectra in the clusters $C1/C2$ (~52.3%). In contrast, for high S/N data, the system identifies the majority of redshifts as very reliable: high percentage of S1 spectra in the clusters $C4/C5$ (~88.1%).

Moreover, the highlighted cells (in magenta) within the result tables indicate two particular cases we anticipated to be null fractions when considering the data quality: We denote on one hand few spectra in the dataset S1 (high S/N) that are associated with unreliable redshift measurements (~2% predicted in $C1/C2$), and on the other hand a small fraction of spectra in S2 (low S/N) that is linked to very reliable redshifts (~9% predicted in $C5$). Such results can easily be understood by looking at the distribution in $[\log(f\ \mathrm{H}\alpha), J_{AB}]$ of the input spectra (cf. Fig. 18). We find that:

- Bright objects are mainly located in $C5$, while the majority of faint objects are predicted as $C1/C2$, in particular when the flux spectrum is embedded in a strong noise (S2). This distribution can be assimilated to a shift $C1 \rightarrow C5$ according to the intrinsic properties of the observed object.
- The difference in absolute values (cf. Table 12) between the results in S1 and S2 is due to the increased noise level from the sky background that injects a higher uncertainty in the observed flux spectrum. The redshift reliability is decreased in S2 in comparison with less noisy data (S1).
  The repartition in absolute values seems to describe a shift $C5 \rightarrow C1$ according to observational constraints (S/N).

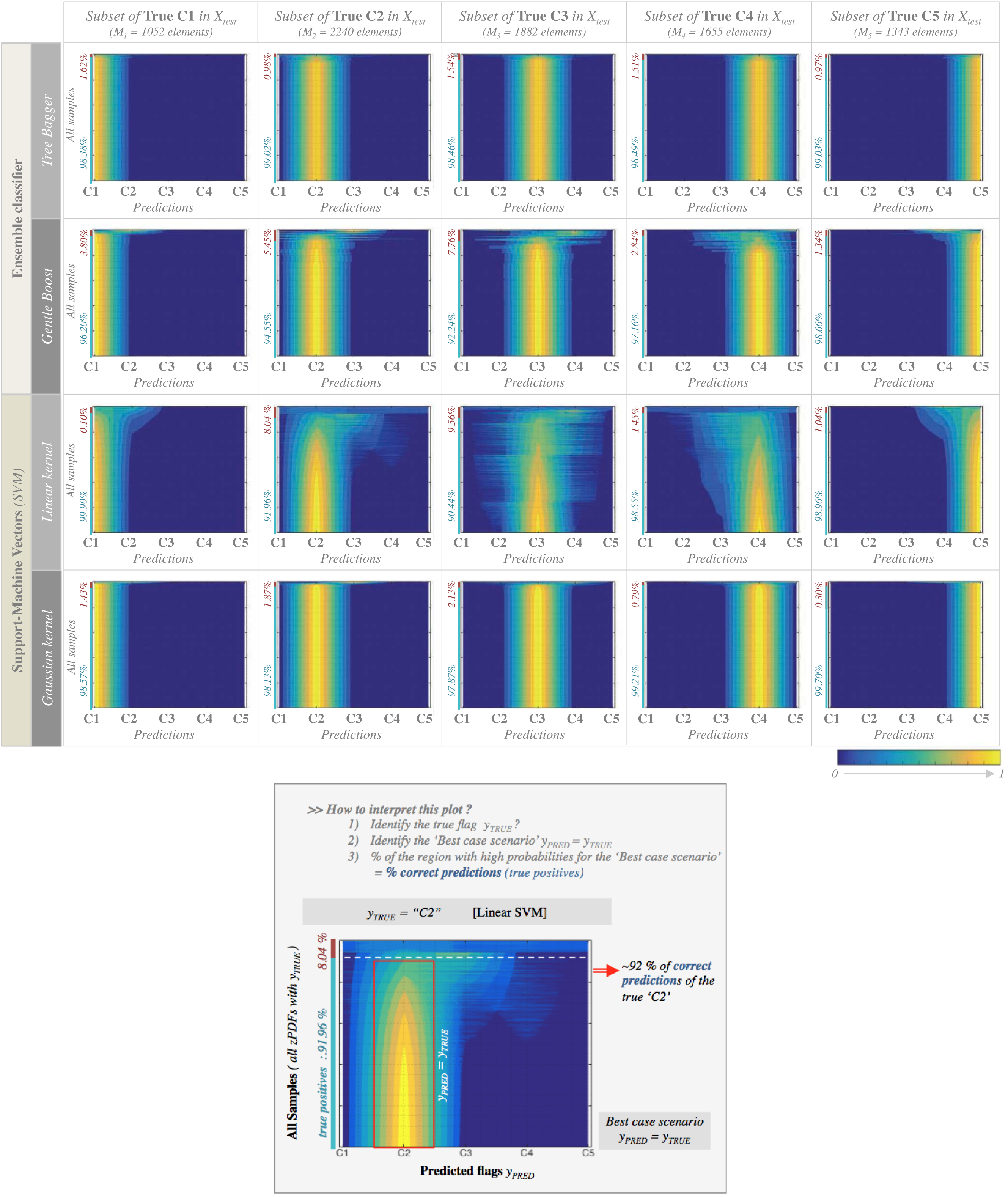

Fig. 14. Class posterior probabilities. The predictive power of several classifiers is displayed for each true class in $\mathbf{Y}_{\text{test}}$. Most prediction probabilities fall between ~[70–100]% (bright colors). For example: the Linear SVM correctly predicts ~92% (2060 elements) of the subset of "true $C2$" (around 2240 elements) in $\mathbf{Y}_{\text{test}}$ (cf. Table F.17) with class prediction probabilities between 0.7 and 1 (bright colors).

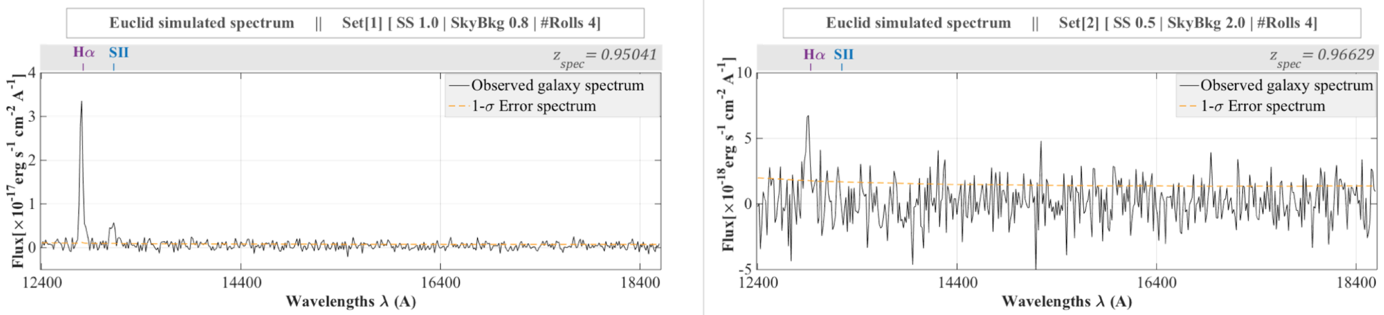

Fig. 15. Simulated *Euclid* spectra. *Left*: simulated galaxy spectrum (id: 53678850) in the dataset S1 with an identifiable Hα line at 12803A. *Right*: simulated galaxy spectrum (id: 56932048) in the dataset S2 with a Hα emission line at 12908A.

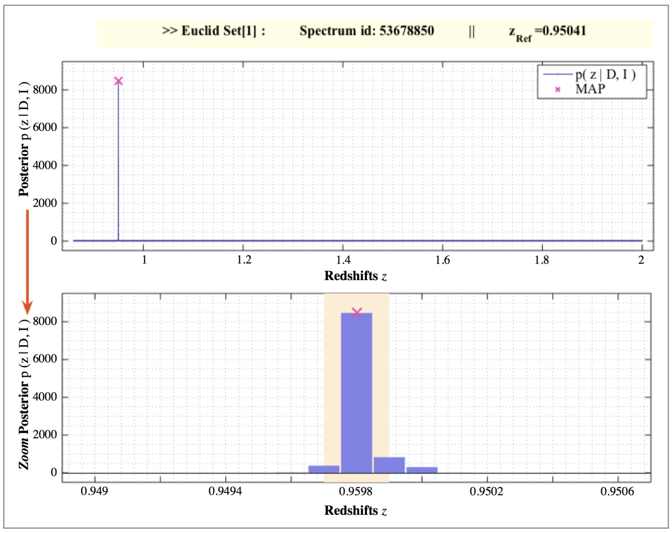

Fig. 16. Computed zPDF for a galaxy spectrum in the dataset S1. The redshift probability density function is computed using a constant prior over Θ_z.

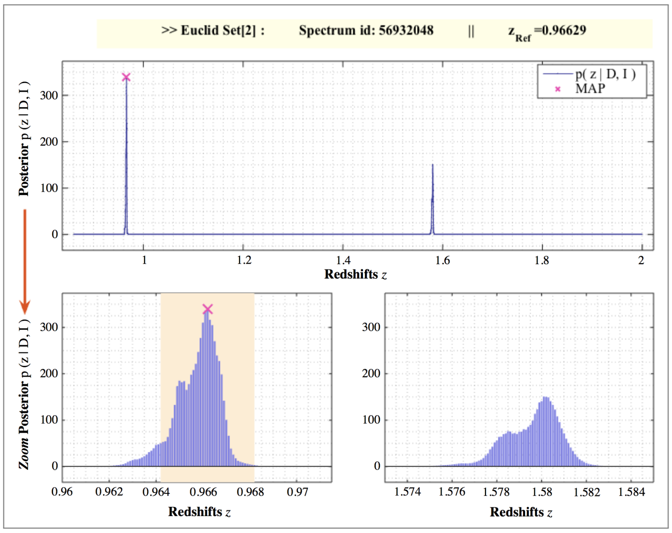

Fig. 17. Computed zPDF for a galaxy spectrum in the dataset S2. The redshift probability density function is computed using a constant prior over Θ_z.

5.2. Redshift error distribution

We further investigate the distribution of the redshift error $\varepsilon_z = |z_{\mathrm{MAP}} - z_{\mathrm{ref}}|/(1 + z_{\mathrm{ref}})$ within the predicted clusters (cf. Table 14).

We find that the majority of incorrect redshift estimations ($\varepsilon_z > 10^{-3}$) are located in the clusters $C1/C2$ for "unreliable redshifts" since low S/N data are more likely to be associated with inaccurate redshift measurements.

For the two datasets, the fraction of spectra associated with low redshift error ($\varepsilon_z \leq 10^{-3}$) is ~100%, ~99%, ~95%, and <70% in $C5/C4$, $C3$, $C2$, and $C1$, respectively.

From this particular result, one approach would be to identify a possible correlation between the redshift reliability clusters and a specific range of redshift errors in order to define a probability for "a redshift to be correct" within the $\{C_k\}$ clusters, in a similar way to that used for VVDS.

In this direction, the next step will be to conduct similar tests on a wide basis of *Euclid* simulated datasets (with a contamination model) to statistically constrain the correlation between redshift errors and the redshift reliability clusters.

5.3. Fuzzy approach for label prediction

As previously stated in Sect. 4.3.3, soft partitioning in ML provides extra information about the classifier predictive power that can be affected by several factors as possible outliers in the training set or numerical limitations associated to the zPDF computation.

In this study, we find that the majority of S1 and S2 spectra are associated with class probability predictions higher than 99%, as in the example of Table 15. However, peculiar cases, related to class predictions falling within the margins of two or more reliability clusters are detected, as in the example reported in Table 16 where the class posterior probabilities of the cluster $C4$ are quite close to the predicted class $C3$.

In this study, the predictions associated to lower-class probabilities are extremely few, with a confusion clearly stated between adjacent clusters ($C1$ with $C2$ or $C4$ with $C5$ for example).

A confusion entailing predicting a label $C4/C5$ as $C1$ (and vice-versa) could have been more problematic and can result from an erroneous computation of the zPDF or an incorrect spec-

Table 14. Redshift error distribution within the predicted redshift reliability classes for preliminary mock simulations for *Euclid*.

	Fraction of spectra with $\varepsilon_z \leq 10^{-3}$				
Set	"$C1$"	"$C2$"	"$C3$"	"$C4$"	"$C5$"
S1	1/3	59/61	313/313	835/835	1957/1957
S2	260/383	1221/1275	550/555	662/662	294/294

Notes. The initial predictions are indicated in gray.

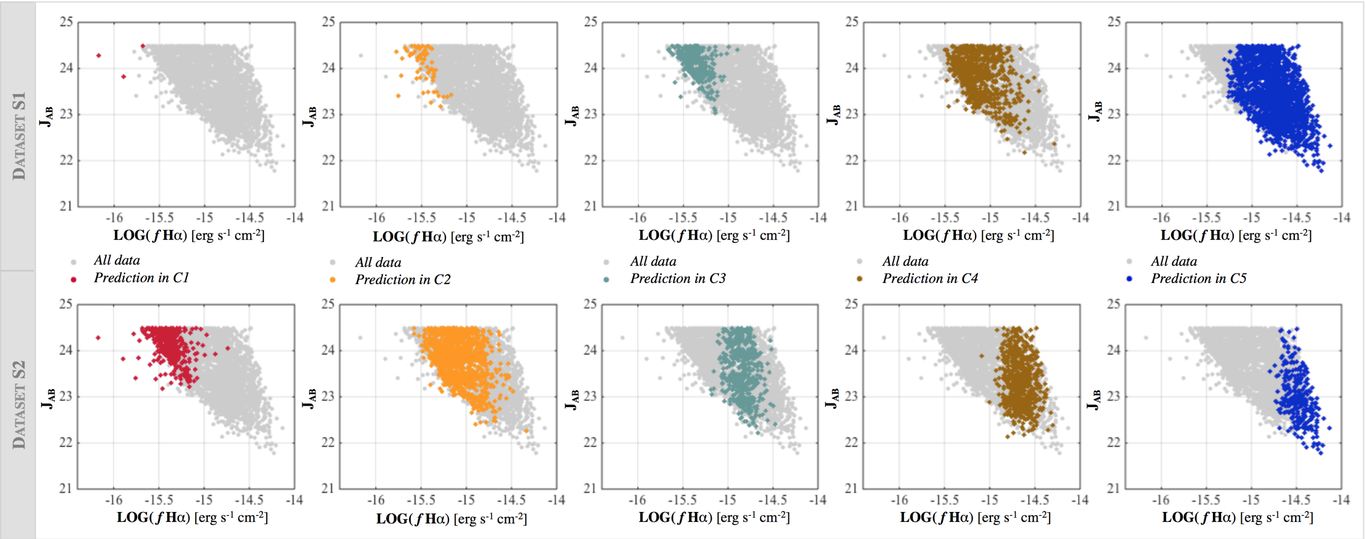

Fig. 18. log(f Hα), J_{AB} distribution of the reliability class predictions for unlabeled simulated galaxy spectra for *Euclid*. The number of faint objects predicted in $C1/C2$ increases when the noise component in the data is important ($S1 \rightarrow S2$). The increased sky background injects a strong noise component of the spectra that annihilates the confidence in a measured redshift, resulting in multimodal zPDFs with high dispersion ($C1/C2$). In contrast, extremely bright objects with an identifiable Hα line are located in $C4/C5$, because the redshift estimation is deemed very reliable when distinct spectral features are found.

Table 15. Class posterior probabilities of two simulated *Euclid* spectra.

		Class probability(in %)				
Set	Spectrum ID	$C1$	$C2$	$C3$	$C4$	$C5$
S1	53678850	0.00	0.17	0.02	0.01	99.81
S2	56932048	99.88	0.05	0.02	0.03	0.03

Notes. In green, the probability associated with the predicted class.

Table 16. Class posterior probabilities for a simulated spectrum in S2.

		Class probability(in %)				
Set	Spectrum ID	$C1$	$C2$	$C3$	$C4$	$C5$
S2	114440656	17.36	8.09	29.43	32.10	13.03

Notes. In green, the probability associated with the predicted class.

troscopic data (flux and noise components). We estimate that soft partitioning can be used to unveil such peculiar cases and improve the clustering by identifying possible outliers in the descriptor space that can be assigned to the “Unidentified” class independently from the $\{C_k\}_{k\in\{1,\ldots5\}}$ clusters.

5.4. Discussion

The results obtained using preliminary mock simulations for *Euclid* show that the new automated reliability redshift definition can be used to quantify the reliability level of spectroscopic redshift measurements. This method could be useful for cosmological studies that require accurate redshift measurements. By using 1D spectra of newly released *Euclid* simulations, upcoming studies will focus on the correlation between the distribution of redshift errors and the redshift reliability clusters to define the probability for “a redshift to be correct” in the $\{C_k\}_{k\in\{1,\ldots5\}}$ clusters in a similar approach as in VVDS.

6. Summary and conclusions

By mapping the posterior PDF $p(z|D, I)$ into a discretized feature space and exploiting ML algorithms, we are able to design a new automated method that correlates relevant characteristics of the posterior zPDF, such as the dispersion of the probability distribution and the number of significant modes, with a reliability assessment of the estimated redshift.

The proposed methodology consists of three steps:

1. Using a set of representative spectra, compute the redshift posterior PDFs $p(z|D, I)$ and extract a set of features to build the descriptor matrix $\mathbf{X}$.
2. Generate a reliable partitioning $\mathbf{Y}$ of the feature space using clustering techniques and prior knowledge, if available.
3. Use the partitioning to train a classifier that will predict a quality label for new unlabeled observations.

Using the zPDFs descriptors, we first tried to bypass the first two steps by exploiting existing reliability flags to train a classifier (supervised classification), but the results obtained (Sect. 4.2) justify the need for new homogeneous partitions [steps 1 and 2] of the feature space because the reproducibility of the existing quality flags cannot be achieved due to their subjective definition: the combination of several visual checks performed by different observers cannot derive homogeneous and objective criteria of redshift reliability for an automated system to learn from.

The results of unsupervised classification in Sect. 4.3 displayed great coherency in describing distinct categories of zPDFs: the multimodal zPDFs with equiprobable redshift solutions and high dispersion, versus the unimodal zPDFs with a narrower peak around the $z_{\rm MAP}$ solution, each depicting a different level of reliability for the measured redshift.

To predict a redshift reliability flag for unlabeled data (Sect. 5), our methodology consists in projecting the unlabeled zPDF [step 3] into the mapping generated from known zPDF descriptors $\mathbf{X}$ and their associated z reliability labels $\boldsymbol{Y}$ to predict the class membership.

A fuzzy approach can also be used to predict the class prediction probability and provide relevant information about the classifier performance and possible discrepancies in the input data.

To conclude, the proposed method to automate the redshift reliability assessment is simple and flexible; the only requirement being robust redshift estimation algorithms with representative templates and a good computational efficiency to produce accurate redshift PDFs. For the spectroscopic redshift estimation, the use of the Bayesian framework allows to incorporate multiple sources of information as a prior and any readjustment of the data/model hypotheses into the estimation process and produce a posterior zPDF.

In this work, we have demonstrated that by using a simple entry model and a few ML-algorithms that exploit descriptors of the redshift PDF, it is possible to capture an accurate description of the spectroscopic redshift reliability. This approach paves the way for fully automated processing pipelines of large spectroscopic samples as for next-generation large-scale galaxy surveys. We expect to further develop and test our method for the needs of the *Euclid* space mission when large simulations of realistic spectra become available. Advanced techniques in ML, such as neural networks and deep learning, will be explored to build a complex learning scheme.

Acknowledgements. We thank the referee for helpful comments and multiple suggestions that significantly improved this paper. This work is part of a thesis co-funded by the University Aix-Marseille and the Centre National d'Études Spatiales (CNES) in the context of the upcoming *Euclid* mission of the European Space Agency, as developed by the *Euclid* Consortium. The authors would like to thank the members of the AMAZED team who provided helpful feedback and insightful discussions on this topic. This work benefitted from access to databases managed by the CESAM (Astrophysical Data Center of Marseille) in LAM (Laboratoire d'Astrophysique de Marseille). We would also like to express our thanks to all the contributors whose fruitful comments and revision of this manuscript helped us expand our understanding and improve the paper.

References

Abdalla, F. B., Amara, A., Capak, P., et al. 2008, MNRAS, 387, 969

Albrecht, A., Bernstein, G., Cahn, R., et al. 2006, ArXiv e-prints [arXiv:astro-ph/0609591]

Baldry, I. K., Alpaslan, M., Bauer, A. E., et al. 2014, MNRAS, 441, 2440

Benitez, N. 1999, in ASP Conf. Ser., 536, 571

Beutler, F., Blake, C., Colless, M., et al. 2011, MNRAS, 416, 3017

Beutler, F., Blake, C., Colless, M., et al. 2012, MNRAS, 423, 3430

Bolzonella, M., Miralles, J.-M., & Pelló, R. 2000, A&A, 363, 476

Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJS, 686, 1503

Chandola, V., Banerjee, A., & Kumar, V. 2009, ACM Comput. Surv., 41, 15

Collister, A. A., & Lahav, O. 2004, PASP, 116, 345

Cool, R. J., Moustakas, J., Blanton, M. R., et al. 2013, ApJ, 767, 118

Cristianini, N., & Shawe-Taylor, J. 2000, An Introduction to Support Vector Machines: and Other Kernel-based Learning Methods (Cambridge University Press)

Dietterich, T. G. 2000, in Multiple Classifier Systems (Springer), 1

Dietterich, T. G., & Bakiri, G. 1995, J. Artificial Intelligence Res., 2, 263

Fawcett, T. 2006, Pattern Recognition Lett., 27, 861

Feldmann, R., Carollo, C. M., Porciani, C., et al. 2006, MNRAS, 372, 565

Garilli, B., Fumana, M., Franzetti, P., et al. 2010, PASP, 122, 827

Garilli, B., Guzzo, L., Scodeggio, M., et al. 2014, A&A, 562, A23

Green, J., Schechter, P., Baltay, C., et al. 2012, ArXiv e-prints [arXiv:1208.4012]

Guzzo, L., Scodeggio, M., Garilli, B., et al. 2014, 566, A108

Hastie, T., & Tibshirani, R. 1998, Annals Stat., 26, 451

Huterer, D. 2002, Phys. Rev. D, 65, 3001

Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841

Ivezic, Z., Tyson, J. A., Abel, B., et al. 2008, ArXiv e-prints [arXiv:0805.2366]

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]

Le Fèvre, O., Vettolani, G., Garilli, B., et al. 2005, A&A, 439, 845

Le Fèvre, O., Cassata, P., Cucciati, O., et al. 2013, A&A, 559, A14

Le Fèvre, O., Tasca, L., Cassata, P., et al. 2015, A&A, 576, A79

Linder, E. V., & Jenkins, A. 2003, MNRAS, 346, 573

Machado, D. P., Leonard, A., Starck, J.-L., Abdalla, F. B., & Jouvel, S. 2013, A&A, 560, A83

Patcha, A., & Park, J.-M. 2007, Computer Networks, 51, 3448

Schuecker, P. 1993, ApJS, 84, 39

Scodeggio, M., Franzetti, P., Garilli, B., et al. 2005, PASP, 117, 1284

Shahid, M., Rossholm, A., Lövström, B., & Zepernick, H.-J. 2014, EURASIP J. Image Video Processing, 2014, 40

Simkin, S. M. 1974, A&A, 31, 129

Tonry, J., & Davis, M. 1979, ApJ, 84, 1511

Vapnik, V. N. 2000, The Nature of Statistical Learning Theory (Springer)

Wahba, G. 1998, Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV (MIT Press)

Wahba, G. 2002, PNAS, 99, 16524

Wang, Y., Percival, W., Cimatti, A., et al. 2010, MNRAS, 409, 737

Zoubian, J., Kümmel, M., Kermiche, S., et al. 2014, in ASP Conf. Ser., 485, 509

Appendix A: Assigning probabilities

In the data, if the noise is assumed Gaussian, additive, and i.i.d., the data model for a single observation D_k is:

$$\text{datum } d_k = \text{true } x_k + \text{noise } n_k; \quad N_k \sim \mathcal{N}(0;\sigma_k). \tag{A.1}$$

The variables d_k, n_k , and x_k are realizations of the random variables (r.v.) D_k, N_k and X_k.

By marginalizing over the r.v. X_k and N_k:

$$p(D_k|z, M_t, I) = \int\int p(D_k, N_k, X_k|z, M_t, I)\mathrm{d}X_k\mathrm{d}N_k. \tag{A.2}$$

Assuming $X_k \perp\!\!\!\perp N_k$:

$$\begin{aligned} p(D_k|z, M_t, I) = \int\int & p(X_k|z, M_t, I) \times p(N_k|z, M_t, I) \\ & \times p(D_k|N_k, X_k, z, M_t, I)\mathrm{d}X_k\mathrm{d}N_k. \end{aligned} \tag{A.3}$$

From Eq. (A.1), the likelihood function is:

$$p(D_k|N_k, X_k, z, M_t, I) = \delta(d_k - x_k - n_k) = \delta_k. \tag{A.4}$$

The Kronecker function δ_k implies $n_k = d_k - x_k$ and allows to rewrite $p(D_k|z, M_t, I)$ as:

$$\begin{aligned} p(D_k|z, M_t, I) &= \int \mathrm{d}x_k f_X(x_k) \int \mathrm{d}n_k f_N(n_k)\delta_k \\ &= \int f_X(x_k) f_N(d_k - x_k)\mathrm{d}x_k, \end{aligned} \tag{A.5}$$

where $f_X(x_k)$ and $f_N(n_k)$ are the probability density functions of the random variables X_k and N_k.

If the true value X_k is considered as a deterministic variable:

$$p(D_k|z, M_t, I) = f_N(d_k - x_k) = f_N(n_k) = p(N_k|z, M_t, I). \tag{A.6}$$

Otherwise, the probabilistic model of x_k has to be integrated into the full expression of $p(D_k|z, M_t, I)$.

The likelihood $\mathcal{L}(z, M_t)$ describes the probability of observing the full set of independent observations $D = \{D_k\}_{k\in\Lambda}$ given a redshift z and a template model M_t and any additional information I.

Considering the aforementioned hypotheses on the data model, the likelihood is defined as following:

$$\begin{aligned} \mathcal{L}(z, M_t) &= p(D|z, M_t, I) = p(D_1, \ldots D_n|z, M_t, I) \\ &= p(N_1, \ldots N_n|z, M_t, I) = \textstyle\prod_{k\in\Lambda} p(N_k|z, M_t, I) \\ &= \textstyle\prod_{1\leq i\leq n} \left(\sqrt{2\pi}\sigma_i\right)^{-1} \exp\left(-\frac{1}{2}\chi^2(z, t)\right) \end{aligned} \tag{A.7}$$

$$\chi^2(z, t) = \textstyle\sum_{i=1}^{n} \sigma_i^{-2}[d_i - a_{\text{opt}}\, t_{i,z}]^2,$$

where Λ is the wavelength range in use (with n datapoints), d_i and σ_i are the observed flux and noise spectra at pixel i, respectively, $t_{i,z}$ is the redshifted template interpolated at pixel i, and a_{opt} is the optimal amplitude obtained from (weighted) Least-Square (LS) estimation.

We would like to point out that the estimation is in reality obtained from marginalizing over nuisance parameters θ, such as the amplitude A (r.v.) in the chi-square expression:

$$p(z, M_t|D, I) = \int p(z, M_t, \theta|D, I)\mathrm{d}\theta. \tag{A.8}$$

The joint-posterior PDF can be rewritten as:

$$\begin{aligned} p(z, M_t|D, I) &= \int \frac{p(\theta, z, M_t|I) \times p(D|z, M_t, \theta, I)}{p(D)}\mathrm{d}\theta \\ &= \int \frac{p(z, M_t|I) \times p(\theta|z, M_t, I) \times p(D|z, M_t, \theta, I)}{p(D)}\mathrm{d}\theta \\ &= \frac{p(z, M_t|I) \times p(D|z, M_t, \theta_{opt}, I)}{p(D)} \\ &\quad \times \int p(\theta|z, M_t, I)\frac{p(D|z, M_t, \theta, I)}{p(D|z, M_t, \theta_{opt}, I)}\mathrm{d}\theta, \end{aligned} \tag{A.9}$$

where the highlighted integral in blue is usually approximated by a constant, and the computed likelihood in redshift estimation englobes the optimal estimation a_{opt} of the amplitude parameter in Eq. (A.7).

The amplitude a_{opt} is estimated at each trial (z, t):

$$\begin{aligned} a_{opt} &= (t_z^\top w\; t_z)^{-1}\; t_z^\top w\; s \\ &= \Big(\sum_{i=1}^{n} s_i t_{i,z}\sigma_i^{-2}\Big)\Big/\Big(\sum_{i=1}^{n} t_{i,z}^2\sigma_i^{-2}\Big), \end{aligned} \tag{A.10}$$

where $w = \mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_n^{-2})$ is the weight matrix.

Appendix B: ECOC for multi-class problems

The principle of ECOC (Error-Correcting-Output-Codes) is based on the binary reduction of the multi-class problem using a coding matrix $\mathcal{M} \in \{-1; 0; +1\}^{K\times L}$ to design a codeword.

$$\mathcal{M} = \begin{array}{c} \\ c_1 \\ \\ c_K \end{array} \!\!\begin{array}{c} \begin{array}{ccc} l_1 & & l_L \end{array} \\ \begin{pmatrix} m_{11} & \cdots & m_{1L} \\ \vdots & \ddots & \vdots \\ m_{K1} & \cdots & m_{KL} \end{pmatrix} \end{array}, \tag{B.1}$$

where:

- L: number of learners;
- K: number of distinct classes.

The codewords $\boldsymbol{m}_k = (m_{k,1}, \ldots, m_{k,L})$ translate the membership information for each class c_k given a binary scheme:

- $m_{kj} = -1$: c_k is the negative class for learner l_j;

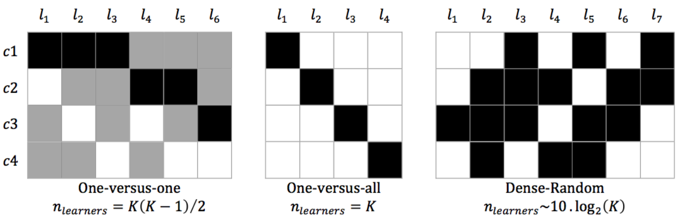
l_1 l_2 l_3 l_4 l_5 l_6
l_1 l_2 l_3 l_4
l_1 l_2 l_3 l_4 l_5 l_6 l_7
c1
c2
c3
c4
One-versus-one
$n_{learners} = K(K-1)/2$
One-versus-all
$n_{learners} = K$
Dense-Random
$n_{learners} \sim 10.\log_2(K)$

Fig. B.1. [Examples of ECOC coding design] . The black, white, and gray boxes refer respectively to $m_{k,j}$ = +1, –1 or 0.

- $m_{kj} = 0$: all observations associated with c_k are ignored by the learner l_j;
- $m_{kj} = +1$: c_k is the positive class for learner l_j.

Codewords are generated using existing coding strategies such as OVA (one-versus-all), OVO (one-versus-one) and dense random. Coding matrices for a example of a four-class problem are shown in Fig. B.1.

Each learner l_j is associated with two superclasses, $\{S_+; S_-\}$ referring to the positive and the negative classes, respectively, that are used to encode the response vector $\mathbf{Y}_{\text{train}}$ into a binary vector $[\mathbf{Y}_{\text{train}}]_{\text{j}}$.

Training the learner l_j with $\left\{\mathbf{X}_{\text{train}}; [\mathbf{Y}_{\text{train}}]_{\text{j}}\right\}$ is performed with the usual classifiers such as SVM, MLP, and so on.

A class prediction for an unlabeled spectra $\boldsymbol{x}_0$ in $\mathbf{X}_{\text{est}}$ is achieved in two steps:

- Step 1: each trained learner l_j provides a binary prediction: $(y_0)_j \in \{-1; +1\}$.
- Step 2: the bit vector $\boldsymbol{y}_0$ for all learners is decoded into the initial K-class by minimizing a distance metric Δ as the Euclidean distance $\Delta_k = \sum_{j=1}^{L} (m_{kj} - (y_0)_j)^{2^{1/2}}$ or a binary loss function $\Delta_k = \sum_{j=1}^{L} |m_{kj}|\; g(m_{kj}, s_j)$, where g is a binary loss function and s_j the score for learner j.
  The predicted class $\widehat{c_k}$ for $\boldsymbol{x}_0$ is associated with the index k for which the vector Δ is minimal.

Appendix C: Description of classification algorithms: SVM – Ensemble methods

C.1. Support-machine vectors

The SVM method classifies the data by finding the best hyperplane separating the datapoints of one class from those of another category.

Given a training set of M datapoints $(\boldsymbol{x}_i, y_i)_{i\in 1\ldots M}$, where $\boldsymbol{x}_i$ refers to the P-dimensional feature vector and y_i the associated label that indicates whether the datapoint belongs to the positive class ($y_i = +1$) or the negative class ($y_i = -1$), the objective of SVM is to separate the data into distinct classes using a separating rule in form of a parametrized function $f(\boldsymbol{x})$.

For linearly separable data, the equation of the hyperplane is:

$$f(\boldsymbol{x}) = f_{\boldsymbol{w},b}(\boldsymbol{x}) = \boldsymbol{w}.\boldsymbol{x} + b = 0, \tag{C.1}$$

where the scalar product $\boldsymbol{w}.\boldsymbol{x}$ is equivalent to $\boldsymbol{w}^\top \boldsymbol{x}$.

An infinity of hyperplanes verify Eq. (C.1), but only one hyperplane maximizing the margins between the observations and the hyperplane exists. This optimal hyperplane verify:

$$\begin{cases} (\boldsymbol{w}.\boldsymbol{x} + b) \geq +1 \text{ if } y_i = +1 \\ (\boldsymbol{w}.\boldsymbol{x} + b) \leq -1 \text{ if } y_i = -1 \end{cases} \Leftrightarrow y_i(\boldsymbol{w}.\boldsymbol{x} + b) \geq +1. \tag{C.2}$$

To find the “best” linear hyperplane minimizing the margins $2(\boldsymbol{w}.\boldsymbol{w}^\top)^{-1/2}$, the SVM algorithm consists in solving a quadratic problem:

$$\begin{aligned} &\underset{\boldsymbol{w},b}{\text{minimize}} \;\; \tfrac{1}{2}(\boldsymbol{w}\boldsymbol{w}^\top) \\ &\text{subject to: } y_i(\boldsymbol{w}.\boldsymbol{x}_i + b) \geq +1. \end{aligned} \tag{C.3}$$

For non-linearly separable data, the use of a kernel φ trick enables to map the distribution of the datapoints $\boldsymbol{x}$ into a projected space where $\varphi(\boldsymbol{x})$ can be linearly separable, and defines, in the same approach as in Eq. (C.2), a quadratic problem:

$$\begin{aligned} &\underset{\boldsymbol{W},B}{\text{minimize}} \;\; \tfrac{1}{2}(\boldsymbol{W}\boldsymbol{W}^\top) \\ &\text{subject to: } y_i(\boldsymbol{W}.\varphi(\boldsymbol{x}_i) + B) \geq +1, \end{aligned} \tag{C.4}$$

where $f(\boldsymbol{x}) = f_{\boldsymbol{W},B}(\boldsymbol{x}) = \boldsymbol{W}.\varphi(\boldsymbol{x}) + B = 0$.

The selection of an adequate kernel $\mathbf{K}$ is determined by a list of criteria. By definition, a kernel must be symmetric, definite positive, square integrable and satisfy:

$$\begin{cases} K_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i).\varphi(\boldsymbol{x}_j) \\ \exists(\lambda_1 \ldots \lambda_N) \in \mathbb{R}: \quad \sum_{i=0}^{N} \sum_{j=0}^{N} \lambda_i \lambda_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0. \end{cases} \tag{C.5}$$

Among commonly used kernels:

- linear: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i.\boldsymbol{x}_j$;
- power: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i.\boldsymbol{x}_j)^m$;
- Gaussian (rbf): $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{1}{2}|\boldsymbol{x}_i - \boldsymbol{x}_j|^2/\sigma^2)$.

Further details about the SVMs are available in Vapnik (2000) and Cristianini & Shawe-Taylor (2000).

C.2. Ensemble classifiers

The principle of ensemble methodology is to combine a set of predictions from different learners in order to improve the accuracy of a single learner.

Among the ensemble methods, two distinct approaches are identified:

1. averaging methods: Bagging/ Random forests...
2. boosting methods: AdaBoost/GentleBoost/RSBoost/...

– AdaBoost

AdaBoost, known also as “Adaptive Boosting” refers to a specific algorithm of boosted classifier defined as the sum of individual predictions from T weak learners. The algorithm aims to minimize the (weighted) classification error at each iteration:

$$\varepsilon_t = \sum_{i=1}^{N} d_i^{(t)} \mathbb{1}(y_i \neq h_t(\boldsymbol{x}_i)), \tag{C.6}$$

where:

- x_i is the feature vector of the ith observation;
- y_i is the true label of the ith observation;
- h_t is the prediction of learner t;
- $\mathbb{1}$ is the indicator function;
- $d_i^{(t)}$ is the weight of the ith observation at step t;
- t the iteration step from 1 to T.

At the first iteration, the weights $d_i^{(t)}$ are initialized (e.g., $d_i^{(t)} = 1/N$) and the weak learner h_t is obtained by minimizing the error ε_t. For the next iteration, the weights of the learner $(t + 1)$ are adjusted according to the performance of the previous one (t): whether increase $d_i^{(t+1)}$ for misclassified observations by learner t, or reduce the weights otherwise. The learner h_{t+1} is trained using the updated weights $d_i^{(t)}$ in the error ε_{t+1}.

After training, the prediction for a new data point, $\boldsymbol{x}$, is obtained by combining the individual predictions of all weak learners:

$$f(\boldsymbol{x}) = \sum_{t=1}^{\mathrm{T}} \alpha_t h_t(\boldsymbol{x}); \qquad \alpha_t = \frac{1}{2}\log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right). \tag{C.7}$$

The AdaBoost algorithm can also be viewed as a minimization of an exponential loss function:

$$\sum_{i=1}^{N} w_t \exp(-y_i f(\boldsymbol{x}_i)), \tag{C.8}$$

where w_t are normalized observational weights.

– LogitBoost

Following a similar approach to AdaBoost, the LogitBoost consists in training learners sequentially by minimizing an error function ε_t; the only difference being the minimization of the error function with respect to a fitted regression model $\widetilde{y}$ instead of y:

$$\varepsilon_t = \sum_{i=1}^{N} d_i^{(t)}(\widetilde{y}_i - h_t(\boldsymbol{x}_i))^2; \qquad \widetilde{y}_i = \frac{y_i^* - p_t(\boldsymbol{x}_i)}{p_t(\boldsymbol{x}_i)(1-p_t(\boldsymbol{x}_i))}, \tag{C.9}$$

where:

- y_i^* are modified labels: $y_i^* = 0$ if $y_i = -1$; and $y_i^* = 1$ otherwise;
- $p_t(\boldsymbol{x}_i)$ is the predicted class probability for the ith observation to be in the positive class “+1” given by the learner t.

– GentleBoost

Also called Gentle AdaBoost, this algorithm combines the methodology of AdaBoost and LogitBoost. An exponential loss function is minimized with a different optimization strategy to AdaBoost. Further, similarly to LogiBoost, weak learners fit a regression model $\widetilde{y}$ to the response variables y.

– Bagging

Bagging, referring to “bootstrap aggregation”, consists in generating m new training sets P_j, each of size N', by uniformly sampling with replacement from the initial training set $I = (\boldsymbol{x}_i, y_i)_{i\in 1...N}$.

The m models are trained separately and the class prediction of an unlabeled data $\boldsymbol{x}$ is obtained by combining the individual predictions of the m models: “averaging” if regression, or “voting” if classification.

Further details on the ensemble algorithms can be found in Dietterich (2000).

Appendix D: Measures for multi-class classification

For a binary classification, the confusion matrix represents the fraction of predicted labels versus the true classes. Four quantities are directly measured:

- TP: True Positives;
- TN: True Negatives;
- FP: False Positives;
- FN: False Negatives.

		True		
		pos	*neg*	*Total*
Predicted	*pos*	TP	FP	TP+FP
	neg	FN	TN	FN+TN
	Total	TP+FN	TN+FP	TP+FP+TN+FN

For multi-class classification, the approach consists in estimating these measures for each class. For example:

		True			
		“0”	“+1”	“+2”	*Total*
Predicted	“0”	4503	2	1	4506
	“+1”	9	3581	203	3793
	“+2”	0	0	8047	8047
	Total	4512	3583	8251	16346

Measures per class

	“0”	“+1”	“+2”
TP	4503	3581	8047
FP	3	212	0
FN	9	2	204
TN	11831	12551	8095

From the confusion matrix, the overall performances of the classification are quantified with the following measures:

multi-class classification $(C_k)_{k\in\{1...K\}}$

MEASURES PER-CLASS	ACCURACY(C_k)	$\frac{TP_k+TN_k}{TP_k+FN_k+TN_k+FP_k}$
	PRECISION(C_k)	$\frac{TP_k}{TP_k+FP_k}$
	SENSITIVITY(C_k)	$\frac{TP_k}{TP_k+FN_k}$
	F-SCORE(C_k)	$\frac{2*TP_k}{2*TP_k+FN_k+FP_k}$
	SPECIFICITY(C_k)	$\frac{TN_k}{FP_k+TN_k}$
AVERAGE PER-CLASS	ACCURACY	$\frac{1}{K}\sum_k Accuracy(C_k)$
	ERROR RATE	$\frac{1}{K}\sum_k Error(C_k)$
	PRECISION	$\frac{1}{K}\sum_k Precision(C_k)$
	SENSITIVITY	$\frac{1}{K}\sum_k Sensitivity(C_k)$
	F-SCORE	$\frac{1}{K}\sum_k Fscore(C_k)$

Further details are provided in Fawcett (2006).

Appendix E: Description of the FCM clustering algorithm

Similar to the k-means algorithm that aims to minimize the intraclass variance, the FCM (Fuzzy C-Means) algorithm exploits additional information about membership of the data to multiple clusters.

To partition a dataset $\boldsymbol{X} = (\boldsymbol{x}_1 \dots \boldsymbol{x}_M)^\top$ of P-dimensional vectors into K clusters, the algorithm aims to solve a quadratic problem in order to determine the optimal solution ($\mathbf{U}$, $\mathbf{G}$), where $\mathbf{G} = (\boldsymbol{g}_1 \dots \boldsymbol{g}_M)$ refers to the centroids of the final K clusters and $\mathbf{U} = (\mu_{ij})_{\left\{\substack{1\le i\le K\\ 1\le j\le M}\right.}$ is a coefficient matrix of class memberships for each element.

$$\begin{aligned} &\underset{\mathbf{U},\mathbf{G}}{\text{minimize}} \quad J(\boldsymbol{X};\mathbf{U},\mathbf{G}) \\ &\text{subject to: } \textstyle\sum_{i=1}^{K}\mu_{ij} = 1, \quad j = 1,\dots,M. \end{aligned} \tag{E.1}$$

The algorithm proceeds iteratively and converges when the estimated coefficient matrix at the iteration t is not very different from its previous estimation:

$$||\mathbf{U}^{(t)} - \mathbf{U}^{(t-1)}|| < \varepsilon_0 \text{ , where } \varepsilon_0 \text{ is fixed by the user.} \tag{E.2}$$

Further, the elements $(\boldsymbol{x}_j)_{j=1,\dots,M}$ are said to belong to the class $(c_i)_{i=1,\dots,K}$ for which the final coefficient μ_{ij} is maximal.

The matrix $\mathbf{U}$ can be further exploited to assess the membership level of the element $\boldsymbol{x}_j$ to its predicted class.

The cost function to minimize is:

$$J(\boldsymbol{X};\mathbf{U},\mathbf{G}) = \sum_{i=1}^{K}\sum_{j=1}^{M}\mu_{ij}^{m}D_{ijA}^{2}, \tag{E.3}$$

where the distance D_{ijA}, the coefficient μ_{ij} and the centroid $\boldsymbol{g}_i$ are defined as following:

$$D_{ijA}^2 = ||\boldsymbol{x}_j - \boldsymbol{g}_i||_A^2 = (\boldsymbol{x}_j - \boldsymbol{g}_i)^\top A(\boldsymbol{x}_j - \boldsymbol{g}_i), \tag{E.4}$$

$$\mu_{ij} = \left(\sum_{k=1}^{K} D_{ijA}^2 / D_{kjA}^2\right)^{-2/(m-1)}, \tag{E.5}$$

$$\boldsymbol{g}_i = \left(\sum_{j=1}^{M}\mu_{ij}^{m}\boldsymbol{x}_j\right)\Bigg/\left(\sum_{j=1}^{M}\mu_{ij}^{m}\right). \tag{E.6}$$

In the FCM, the fuzzifier parameter $m \geq 1$ is used to determine the level of fuzziness: if $m = 1$, the coefficient matrix is binary, which is equivalent to a hard partitioning. Usually, in the absence of prior information about the datamodel, the value $m = 2$ is used.

For the norm $||.||_A^2$ in Eq. (E.4), a common choice for the matrix A is the identity matrix, but it can be designed to incorporate individual variances of the data as $A = \text{diag}(\sigma_1^{-2} \dots \sigma_M^{-2})$ or the inverse of the covariance matrix.

Appendix F: Additional tables

Confusion matrices using modified VVDS flags

Table F.1. [Resubstitution prediction] – bagging trees.

		True			
		“0”	“+1”	“+2”	*Total*
Predicted	“0”	4503	2	1	4506
	“+1”	9	3581	203	3793
	“+2”	0	0	8047	8047
	Total	4512	3583	8251	16346

Table F.2. [Resubstitution prediction] – gentle boost.

		True			
		“0”	“+1”	“+2”	*Total*
Predicted	“0”	3520	271	0	3791
	“+1”	988	3070	1752	5810
	“+2”	4	242	6499	6745
	Total	4512	3583	8251	16346

Table F.3. [Resubstitution prediction] – SVM (linear kernel).

		True			
		“0”	“+1”	“+2”	*Total*
Predicted	“0”	3654	457	1	4112
	“+1”	851	2281	625	3757
	“+2”	7	845	7625	8477
	Total	4512	3583	8251	16346

Table F.4. [Resubstitution prediction] – SVM (Gaussian kernel).

		True			
		“0”	“+1”	“+2”	*Total*
Predicted	“0”	3646	372	2	4020
	“+1”	866	2823	1431	5120
	“+2”	0	388	6818	7206
	Total	4512	3583	8251	16346

Table F.5. [Test prediction] – bagging trees.

		True			
		“0”	“+1”	“+2”	*Total*
Predicted	“0”	722	271	442	1435
	“+1”	974	256	1424	2654
	“+2”	560	1265	2259	4084
	Total	2256	1792	4125	8173

Table F.6. [Test prediction] – gentle boost.

		True			
		“0”	“+1”	“+2”	*Total*
Predicted	“0”	447	272	253	972
	“+1”	1398	381	1799	3578
	“+2”	411	1139	2073	3623
	Total	2256	1792	4125	8173

Table F.7. [Test prediction] – SVM (linear kernel).

		True			
		“0”	“+1”	“+2”	*Total*
Predicted	“0”	559	272	330	1161
	“+1”	956	141	1317	2414
	“+2”	741	1379	2478	4598
	Total	2256	1792	4125	8173

Table F.8. [Test prediction] – SVM (Gaussian kernel).

		True			
		“0”	“+1”	“+2”	*Total*
Predicted	“0”	494	270	289	1053
	“+1”	1300	336	1638	3274
	“+2”	462	1186	2198	3846
	Total	2256	1792	4125	8173

General performances using modified VVDS flags

Table F.9. [Resubstitution prediction] – Measures from confusion matrices

Measures per class

		"0"	"+1"	"+2"
Tree bagger	Accuracy	99.93%	98.69%	98.75%
	Precision	99.93%	94.41%	100%
	Sensitivity	99.80%	99.94%	97.53%
	Specificity	99.97%	98.34%	100%
	F-score	99.87%	97.10%	98.75%
Gentle boost	Accuracy	92.27%	80.10%	87.78%
	Precision	92.85%	52.84%	96.35%
	Sensitivity	78.01%	85.68%	78.77%
	Specificity	97.71%	78.53%	96.96%
	F-score	84.79%	65.37%	86.68%
SVM (linear)	Accuracy	91.95%	83.01%	90.96%
	Precision	88.86%	60.71%	89.95%
	Sensitivity	80.98%	63.66%	92.41%
	Specificity	96.13%	88.44%	89.47%
	F-score	84.74%	62.15%	91.16%
SVM (rbf)	Accuracy	92.41%	81.30%	88.86%
	Precision	90.70%	55.14%	94.62%
	Sensitivity	80.81%	78.79%	82.63%
	Specificity	96.84%	82%	95.21%
	F-score	85.47%	64.87%	88.22%

Average per-class

Tree bagger	Accuracy	99.12%
	Error rate	0.88%
	Precision	98.11%
	Sensitivity	99.09%
	F-score	98.60%
Gentle boost	Accuracy	86.72%
	Error rate	13.28%
	Precision	80.68%
	Sensitivity	80.82%
	F-score	80.75%
SVM (linear)	Accuracy	88.64%
	Error rate	11.36%
	Precision	79.84%
	Sensitivity	79.02%
	F-score	79.43%
SVM (rbf)	Accuracy	87.52%
	Error rate	12.48%
	Precision	80.15%
	Sensitivity	80.74%
	F-score	80.45%

Table F.10. [Test prediction] – Measures from confusion matrices.

Measures per class

		"0"	"+1"	"+2"
Tree bagger	Accuracy	72.51%	51.87%	54.84%
	Precision	50.31%	9.65%	55.31%
	Sensitivity	32%	14.29%	54.76%
	Specificity	87.95%	62.42%	54.92%
	F-score	39.12%	11.52%	55.04%
Gentle boost	Accuracy	71.44%	43.62%	55.93%
	Precision	45.99%	10.65%	57.22%
	Sensitivity	19.81%	21.26%	50.25%
	Specificity	91.13%	49.90%	61.71%
	F-score	27.70%	14.19%	53.51%
SVM (linear)	Accuracy	71.87%	51.99%	53.91%
	Precision	48.15%	5.84%	53.89%
	Sensitivity	24.78%	7.87%	60.07%
	Specificity	89.83%	64.38%	47.63%
	F-score	32.72%	6.70%	56.82%
SVM (rbf)	Accuracy	71.60%	46.24%	56.26%
	Precision	46.91%	10.26%	57.15%
	Sensitivity	21.90%	18.75%	53.28%
	Specificity	90.55%	53.96%	59.29%
	F-score	29.86%	13.26%	55.15%

Average per-class

Tree bagger	Accuracy	59.74%
	Error rate	40.26%
	Precision	38.42%
	Sensitivity	33.68%
	F-score	35.90%
Gentle boost	Accuracy	57%
	Error rate	43%
	Precision	37.95%
	Sensitivity	30.44%
	F-score	33.79%
SVM (linear)	Accuracy	59.26%
	Error rate	40.74%
	Precision	35.96%
	Sensitivity	30.91%
	F-score	33.24%
SVM (rbf)	Accuracy	58.03%
	Error rate	41.97%
	Precision	38.11%
	Sensitivity	31.31%
	F-score	34.38%

Confusion matrices using partition labels

Table F.11. [Resubstitution prediction] – bagging trees.

		True					
		*C*1	*C*2	*C*3	*C*4	*C*5	*Total*
Predicted	*C*1	2104	3	0	0	0	2107
	*C*2	0	4476	4	0	0	4480
	*C*3	0	1	3758	2	0	3761
	*C*4	0	0	3	3309	0	3312
	*C*5	0	0	0	0	2687	2687
	Total	2104	4480	3765	3311	2687	16347

Table F.12. [Resubstitution prediction] – gentle boost.

		True					
		*C*1	*C*2	*C*3	*C*4	*C*5	*Total*
Predicted	*C*1	2067	37	0	0	0	2104
	*C*2	37	4346	24	0	0	4407
	*C*3	0	97	3582	14	0	3693
	*C*4	0	0	158	3263	20	3441
	*C*5	0	0	1	34	2667	2702
	Total	2104	4480	3765	3311	2687	16347

Table F.13. [Resubstitution prediction] – SVM (linear kernel).

		True					
		*C*1	*C*2	*C*3	*C*4	*C*5	*Total*
Predicted	*C*1	2101	8	0	0	0	2109
	*C*2	3	4335	48	0	0	4386
	*C*3	0	136	3545	37	0	3718
	*C*4	0	1	172	3261	18	3452
	*C*5	0	0	0	13	2669	2682
	Total	2104	4480	3765	3311	2687	16347

Table F.14. [Resubstitution prediction] – SVM (Gaussian kernel).

		True					
		*C*1	*C*2	*C*3	*C*4	*C*5	*Total*
Predicted	*C*1	2098	12	0	0	0	2110
	*C*2	6	4424	10	0	0	4440
	*C*3	0	44	3724	6	0	3774
	*C*4	0	0	31	3294	10	3335
	*C*5	0	0	0	11	2677	2688
	Total	2104	4480	3765	3311	2687	16347

Table F.15. [Test prediction] – bagging trees.

		True					
		*C*1	*C*2	*C*3	*C*4	*C*5	*Total*
Predicted	*C*1	1035	3	0	0	0	1038
	*C*2	17	2218	19	0	0	2254
	*C*3	0	19	1853	18	0	1890
	*C*4	0	0	10	1630	13	1653
	*C*5	0	0	0	7	1330	1337
	Total	1052	2240	1882	1655	1343	8172

Table F.16. [Test prediction] – gentle boost.

		True					
		*C*1	*C*2	*C*3	*C*4	*C*5	*Total*
Predicted	*C*1	1012	8	0	0	0	1020
	*C*2	40	2118	20	1	0	2179
	*C*3	0	114	1736	19	0	1869
	*C*4	0	0	125	1608	18	1751
	*C*5	0	0	1	27	1325	1353
	Total	1052	2240	1882	1655	1343	8172

Table F.17. [Test prediction] – SVM (linear kernel).

		True					
		*C*1	*C*2	*C*3	*C*4	*C*5	*Total*
Predicted	*C*1	1051	4	0	0	0	1055
	*C*2	1	2060	37	0	0	2098
	*C*3	0	176	1702	12	0	1890
	*C*4	0	0	143	1631	14	1788
	*C*5	0	0	0	12	1329	1341
	Total	1052	2240	1882	1655	1343	8172

Table F.18. [Test prediction] – SVM (Gaussian kernel).

		True					
		*C*1	*C*2	*C*3	*C*4	*C*5	*Total*
Predicted	*C*1	1037	4	3	1	1	1046
	*C*2	15	2198	17	0	0	2230
	*C*3	0	38	1842	2	0	1882
	*C*4	0	0	20	1642	3	1665
	*C*5	0	0	0	10	1339	1349
	Total	1052	2240	1882	1655	1343	8172

General performances using partition labels

Table F.19. [Resubstitution prediction]– Measures from confusion matrices.

Measures per class

		*C*1	*C*2	*C*3	*C*4	*C*5
Tree bagger	Accuracy	99.98%	99.95%	99.94%	99.97%	100%
	Precision	99.86%	99.91%	99.92%	99.91%	100%
	Sensitivity	100%	99.91%	99.81%	99.94%	100%
	Specificity	99.98%	99.97%	99.98%	99.98%	100%
	F-score	99.93%	99.91%	99.87%	99.92%	100%
Gentle boost	Accuracy	99.55%	98.81%	98.20%	98.62%	99.66%
	Precision	98.24%	98.62%	96.99%	94.83%	98.70%
	Sensitivity	98.24%	97.01%	95.14%	98.55%	99.26%
	Specificity	99.74%	99.49%	99.12%	98.63%	99.74%
	F-score	98.24%	97.81%	96.06%	96.65%	98.98%
SVM (linear)	Accuracy	99.93%	98.80%	97.60%	98.53%	99.81%
	Precision	99.62%	98.84%	95.35%	94.47%	99.52%
	Sensitivity	99.86%	96.76%	94.16%	98.49%	99.33%
	Specificity	99.94%	99.57%	98.63%	98.53%	99.90%
	F-score	99.74%	97.79%	94.75%	96.44%	99.42%
SVM (rbf)	Accuracy	99.89%	99.56%	99.44%	99.65%	99.87%
	Precision	99.43%	99.64%	98.68%	98.77%	99.59%
	Sensitivity	99.71%	98.75%	98.91%	99.49%	99.63%
	Specificity	99.92%	99.87%	99.60%	99.69%	99.92%
	F-score	99.57%	99.19%	98.79%	99.13%	99.61%

Average per-class

Tree bagger	Accuracy	99.97%
	Error rate	0.03%
	Precision	99.92%
	Sensitivity	99.93%
	F-score	99.93%
Gentle boost	Accuracy	98.97%
	Error rate	1.03%
	Precision	97.48%
	Sensitivity	97.64%
	F-score	97.56%
SVM (linear)	Accuracy	98.93%
	Error rate	1.07%
	Precision	97.56%
	Sensitivity	97.72%
	F-score	97.64%
SVM (rbf)	Accuracy	99.68%
	Error rate	0.32%
	Precision	99.22%
	Sensitivity	99.30%
	F-score	99.26%

Table F.20. [Test prediction] – Measures from confusion matrices.

Measures per class

		*C*1	*C*2	*C*3	*C*4	*C*5
Tree bagger	Accuracy	99.76%	99.29%	99.19%	99.41%	99.76%
	Precision	99.71%	98.40%	98.04%	98.61%	99.48%
	Sensitivity	98.38%	99.02%	98.46%	98.49%	99.03%
	Specificity	99.96%	99.39%	99.41%	99.65%	99.90%
	F-score	99.04%	98.71%	98.25%	98.55%	99.25%
Gentle boost	Accuracy	99.41%	97.76%	96.59%	97.67%	99.44%
	Precision	99.22%	97.20%	92.88%	91.83%	97.93%
	Sensitivity	96.20%	94.55%	92.24%	97.16%	98.66%
	Specificity	99.89%	98.97%	97.89%	97.81%	99.59%
	F-score	97.68%	95.86%	92.56%	94.42%	98.29%
SVM (linear)	Accuracy	99.94%	97.33%	95.50%	97.79%	99.68%
	Precision	99.62%	98.19%	90.05%	91.22%	99.11%
	Sensitivity	99.90%	91.96%	90.44%	98.55%	98.96%
	Specificity	99.94%	99.36%	97.01%	97.59%	99.82%
	F-score	99.76%	94.97%	90.24%	94.74%	99.03%
SVM (rbf)	Accuracy	99.71%	99.09%	99.02%	99.56%	99.83%
	Precision	99.14%	98.57%	97.87%	98.62%	99.26%
	Sensitivity	98.57%	98.13%	97.87%	99.21%	99.70%
	Specificity	99.87%	99.46%	99.36%	99.65%	99.85%
	F-score	98.86%	98.34%	97.87%	98.92%	99.48%

Average per-class

Tree bagger	Accuracy	99.48%
	Error rate	0.52%
	Precision	98.85%
	Sensitivity	98.68%
	F-score	98.76%
Gentle boost	Accuracy	98.17%
	Error rate	1.83%
	Precision	95.81%
	Sensitivity	95.76%
	F-score	95.79%
SVM (linear)	Accuracy	98.05%
	Error rate	1.95%
	Precision	95.64%
	Sensitivity	95.96%
	F-score	95.80%
SVM (rbf)	Accuracy	99.44%
	Error rate	0.56%
	Precision	98.69%
	Sensitivity	98.70%
	F-score	98.69%