



Publication Year	2020
Acceptance in OA	2022-01-24T10:43:45Z
Title	Identification of Single Spectral Lines through Supervised Machine Learning in a Large HST Survey (WISP): A Pilot Study for Euclid and WFIRST
Authors	BARONCHELLI, IVANO, Scarlata, C. M., Rodighiero, G., Rodríguez-Muñoz, L., BONATO, MATTEO, Bagley, M., Henry, A., Rafelski, M., Malkan, M., Colbert, J., Dai, Y. S., Dickinson, H., MANCINI, CHIARA, Mehta, V., Morselli, L., Teplitz, H. I.
Publisher's version (DOI)	10.3847/1538-4365/ab9a3a
Handle	http://hdl.handle.net/20.500.12386/31346
Journal	THE ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES
Volume	249

Identification of single spectral lines through supervised machine learning in a large HST survey (WISP): a pilot study for Euclid and WFIRST

I. BARONCHELLI,¹ C. M. SCARLATA,² G. RODIGHIERO,¹ L. RODRÍGUEZ-MUÑOZ,¹ M. BONATO,^{3,4} M. BAGLEY,⁵ A. HENRY,⁶ M. RAFELSKI,^{7,8} M. MALKAN,⁹ J. COLBERT,¹⁰ Y. S. DAI(戴昱),¹¹ H. DICKINSON,^{12,2} C. MANCINI,¹ V. MEHTA,² L. MORSELLI,¹ AND H. I. TEPLITZ¹⁰

¹*Dipartimento di Fisica e Astronomia, Università di Padova, vicolo Osservatorio, 3, 35122 Padova, Italy.*

²*MN Institute for Astrophysics, University of Minnesota, 116 Church St. SE, Minneapolis, MN 55455, USA.*

³*INAF—Istituto di Radioastronomia and Italian ALMA Regional Centre, Via Gobetti 101, I-40129, Bologna, Italy*

⁴*INAF—Osservatorio Astronomico di Padova, Vicolo dell’Osservatorio 5, I-35122, Padova, Italy*

⁵*College of Natural Sciences, The University of Texas at Austin, 2515 Speedway, Austin, TX 78712, USA*

⁶*Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD, 21218, USA*

⁷*Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA*

⁸*Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA*

⁹*Department of Physics and Astronomy, UCLA, Physics and Astronomy Bldg., 3-714, LA CA 90095-1547, USA*

¹⁰*IPAC, Mail Code 314-6, Caltech, 1200 E. California Blvd., Pasadena, CA 91125, USA.*

¹¹*Chinese Academy of Sciences South America Center for Astronomy (CASSACA)/NAOC, 20A Datun Road, Beijing 100101, China*

¹²*School of Physical Sciences, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK*

ABSTRACT

Future surveys focusing on understanding the nature of dark energy (e.g., Euclid and WFIRST) will cover large fractions of the extragalactic sky in near-IR slitless spectroscopy. These surveys will detect a large number of galaxies that will have only one emission line in the covered spectral range. In order to maximize the scientific return of these missions, it is imperative that single emission lines are correctly identified. Using a supervised machine-learning approach, we classified a sample of single emission lines extracted from the WFC3 IR Spectroscopic Parallel survey (WISP), one of the closest existing analogs to future slitless surveys. Our automatic software integrates a SED fitting strategy with additional independent sources of information. We calibrated it and tested it on a “gold” sample of securely identified objects with multiple lines detected. The algorithm correctly classifies real emission lines with an accuracy of 82.6%, whereas the accuracy of the SED fitting technique alone is low (~50%) due to the limited amount of photometric data available (≤ 6 bands). While not specifically designed for the Euclid and WFIRST surveys, the algorithm represents an important precursor of similar algorithms to be used in these future missions.

1. INTRODUCTION

A *spectroscopic* measurement provides the most precise estimate of the redshift of a given source. Even the most precise spectroscopic surveys, however, have a fraction of spectroscopic failures due to incorrectly identified spectral features (see, e.g., [Hasinger et al. 2018](#)).

The typical approach adopted in a spectroscopic survey exploits a flagging scheme used to characterize the reliability of a galaxy’s redshift measurement. This approach is well exemplified by the VIMOS VLT deep survey ([Le Fèvre et al. 2005](#)), where the quality of the emission lines’ classification is expressed through a four step scale (or flag). The probability that a classification is not correct can be estimated for each flag value using different methods. In the VIMOS sur-

vey, even the best quality sample shows a residual ~0.5-1.0% chance of lines being incorrectly identified. This fraction rises to up ~50% for the lowest-quality sample. The effect of a misidentified line identification results in what are commonly referred to as catastrophic redshift failures, or “*outliers*” in the real versus measured redshift plot.

The danger of line misidentification is that the Gaussian uncertainty associated with the λ position is directly translated to the redshift uncertainty. However, given the line misidentification problem, the actual redshift uncertainty is often not Gaussian and it is characterized by the presence of multiple peaks. The Gaussian assumption, therefore, can lead to a large underestimation of the final uncertainty associated with any physical quantity derived from line fluxes and redshifts.

In an alternative to the spectroscopic approach, many color-based criteria allow us to estimate redshifts, albeit with a lower precision than when spectra are available. Photometric redshifts have historically been used to build samples of

high- z galaxies by exploiting the shape and features of the UV/optical part of the spectra, in particular the presence of the *Balmer* and *Lyman breaks* (at $\lambda \sim 4000\text{\AA}$ and 1216\AA , respectively) as, for example, in the initial works in the Hubble deep field (e.g. Clements & Couch 1996; Madau et al. 1996; Dickinson 1998; Stevens & Lacy 2001). More refined techniques allow us to extrapolate information from photometric measurements by fitting the spectral energy distributions (SEDs) with theoretical and/or empirical models (for a recent review of this topic see, e.g., Salvato et al. 2018). Photometric analyses may also result in outliers, as a consequence, e.g., of the misidentification of the Balmer/Lyman break.

Various strategies can be adopted to reduce the number of catastrophic failures in redshift estimates, e.g., by combining spectroscopic and photometric analysis. First of all, instead of a simplistic Gaussian approximation for each source’s redshift, one can use the information in the full redshift probability distribution function (PDF). PDFs are typically created by most common software that performs fitting to spectral energy distributions (SEDs), such as, e.g., *Hyperz* (Bolzonella et al. 2000). Besides allowing for a more formally correct treatment of the uncertainty and its propagation, the redshift PDF¹ provides a way to compute the probability of a redshift being an outlier. Using the full redshift PDF allows for the selection of samples with different degrees of purity. In any case, the correct computation of PDFs from SED fitting strategies crucially depends on the availability of template galaxy models able to fit the photometric data.

The outlier problem can also be mitigated using photometric priors. These priors may include galaxy colors (as it is the case in photometric redshift estimates via SED fitting) or observed galaxy fluxes. To first approximation, brighter sources are more likely located at lower redshift, so this information can help to disentangle, for example, bright local objects with high metallicity (i.e., steeper optical continuum) from high redshift sources. In this context, some publicly available SED-fitting software provide the user with the option of applying optical priors (e.g. *EAZY*: Brammer et al. 2008). Additional empirical techniques can be exploited to improve the precision of photometric redshift estimates and to correct outliers in specific cases. For example, Baronchelli et al. (2018) demonstrate how additional optical priors, in combination with a far-IR detections, can help improve the accuracy.

The approaches described above can be used to identify single emission lines in galaxy spectra. However, while the photometric redshift estimate can be refined for every source using these techniques, the wavelength position of a spectral line remains the primary source of information in the case of a spectroscopic determination. In other words, these techniques can help in identifying an emission line, but they

do not affect the spectroscopic redshift of a source when the emission lines are already unequivocally identified.

In this paper, we address the problem of correctly identifying single emission lines detected in grism spectra of the HST-WISP survey (WFC3 Infra-red Spectroscopic Parallel survey, Atek et al. 2010, Baronchelli, I. et al. in preparation), by combining different sources of information.

Another goal of this paper is to provide a testing ground for the definition of similar algorithms to be used in the context of the future ESA’s Euclid (Laureijs et al. 2011) and NASA’s WFIRST (Green et al. 2011) missions, in order to maximize the scientific return of their near-IR spectroscopic surveys. It is worth noting that the spectroscopic coverage of the grisms and the photometric bands available in the WISP survey are both very similar to those that are planned to be employed by Euclid and WFIRST. These similarities, together with the wide sky area covered by WISP, make this survey one of the most important proxies for future space-based spectroscopic missions. The Euclid and WFIRST surveys will probably benefit from a large amount of ancillary data. In this sense, focussing on the WISP survey, our analyses represent a *pilot* study of these future missions. In any case, the modular structure of our algorithm is specifically designed to easily include and remove additional modules and sources of information.

The paper is organized as follows: in Section 2 we present the WISP survey and the samples we use to calibrate and test the algorithm. Section 3 describes the algorithm and its modular structure. In Section 4 we compute the precision of the algorithm, also in terms of completeness and contamination of differently selected samples. In Section 5 we report the new classification of WISP sources obtained using our software, while in Section 6 we discuss the implications of our work in the context of future dark energy missions. The main results and future perspectives are finally summarized in Section 7.

2. DATA

We tested our algorithm on the second data release of the WISP survey (WFC3 Infra-red Spectroscopic parallel survey: Atek et al. 2010, Baronchelli, I. et al. in preparation)².

2.1. The WISP survey

WISP is a pure-parallel, near-infrared slitless grism spectroscopic survey that efficiently collects WFC3 data while other HST instruments are in use (P.I. M. Malkan, Atek et al. 2010). The spectral coverage of the WFC3’s grisms, G102 (0.8-1.1 μm , $R \sim 210$) and G141 (1.07-1.7 μm , $R \sim 130$), enables the detection of the $H\alpha$ line from $z \sim 0.3$ to $z \sim 1.5$ and the [O III] emission lines from $z \sim 0.7$ to $z \sim 2.3$. To aid in extracting the 1D spectra from the dispersed images and to enable the wavelength calibration, the WISP fields were also observed in direct imaging mode with filters chosen to match the grisms’ spectral coverage: F110W for G102 and either

¹ Multiple peaks are commonly observed in a typical PDF. These peaks are due to the degeneracy existing among different SED models when fitted to the available photometric data.

² <https://archive.stsci.edu/prepds/wisp/>

F140W or F160W for G141. A relevant fraction of the WISP fields (\sim one third) are also observed with the WFC3 UVIS camera with a subset of the available filters (F475X, F600LP, F606W, and F814W). Finally, about half of the fields are also covered by *Spitzer*/IRAC observations in channel 1 and/or 2 ($3.6\mu\text{m}$ and $4.5\mu\text{m}$, respectively). To date, the WISP survey has observed 483 fields, collectively covering an area of more than 2000 arcmin² (the actual available data in the WISP spectroscopic catalog refer to a total area of 1520 arcmin²).

Being a parallel survey, the observing strategy of WISP depends on the details of the primary observations. This means that the depth of the coverage varies from field to field. Accordingly, we divided visit opportunities into two categories: “short” and “long” visits, corresponding to fewer than four, and four or more orbits, respectively. During the short opportunities, only the G141 grism and the F140W (or F160W) filter are used, while observations in the G102 grism and in the F110W filter are added during long opportunities. In the latter case, the relative integration time between the G102 and G141 grisms is chosen to balance the sensitivity reached by the two grisms. Given these premises, the median 5σ depth reached in both grisms is $5 \times 10^{-17} \text{ erg s}^{-1} \text{ cm}^{-2}$, with a factor of approximately two field-to-field variation.

For the line detection, we apply a wavelet convolution and a SExtractor-type threshold through a custom line-finding software. This approach dramatically reduces the number of false positive detections. For the detection, we require at least three contiguous pixels with $S/N \geq 2.3$. This translates into $S/N \geq 4$ integrated over the full emission line (i.e. the sum over the pixels involved). However, this value does not directly correspond to the S/N ratio reported in the final catalog. In that case, the flux and flux uncertainty are measured after the lines and the full spectrum are fit (meaning that the signal becomes the area under the Gaussian curve rather than the sum of the individual pixel values).

Once detected, each emission line is identified through a visual inspection streamlined with a Python-based interactive line fitting and measuring code. Every object is analyzed by at least two separate reviewers, and a total of 10 reviewers were involved in this process. During this phase, false identifications due to contamination by overlapping spectra or to specific noise features are excluded from the sample. All identified real emission lines are then fitted with a Gaussian profile + continuum, using a line fitting algorithm that required little input from the reviewer. When an emission line is detected in a spectrum, additional emission lines, initially undetected, can also be fit even if their flux was below the original pixel-based detection threshold.

The quantities measured for each line (total flux, equivalent width (EW), full-width-half-maximum FWHM) are then merged into a catalog that has a unique entry for each galaxy with all the spectroscopic information. Differences among reviewers’ classifications and line fits are used to obtain average solutions, when possible, and to define the quality flag associated with each spectrum. In particular, when no agreement is found between the identifications of two reviewers, this information is included in the quality flag. In this case, in

the final catalog, only the solution associated with the line fit that minimizes the χ^2 parameter is included. More details on the creation of the WISP emission line catalog are presented in Bagley, M. et al. (2020, in preparation).

2.2. Calibration and test samples

The algorithm is calibrated (Section 3.2) and tested (Section 4) on two not independent (almost completely overlapping) subsamples extracted from the WISP spectroscopic catalog of secure identifications. The choice of using two overlapping samples is justified in Section 4.1. For these samples, the standard method used to identify the spectral lines can be applied, i.e., two or more spectral lines are detected above a 2σ threshold. Additionally, we require that the reviewers agree on the identification of the lines. We only include sources detected in fields covered by both WFC3 grisms. This selection reduces the WISP area usable for calibration and testing to ~ 900 arcmin². Finally, we exclude all the emission lines located at the low sensitivity ends of the grisms’ wavelength range ($\lambda < 8500\text{\AA}$, $\lambda > 16700\text{\AA}$ and $11000\text{\AA} < \lambda < 11400\text{\AA}$)³. We call these two subsamples the *calibration* (2128 sources) and the *test* (2283 sources) “gold” sample. The only difference between the two samples is that in the *calibration* sample we consider only sources detected in the F110W photometric band, while the same requirement is relaxed when testing the algorithm (measuring accuracy, completeness, and contamination). This selection is due to the fact that the F110W magnitude is required to calibrate one of the modules of the algorithm (the magnitude prior described in Section 3.2.2), but the same measurement is not necessary when running the software on unidentified spectral lines. Consequently, the two samples almost fully overlap, with the entire *calibration* sample included in the *test* sample (only $\sim 7\%$ of the sources used for the test are not used for the calibration).

After calibrating and testing the algorithm, we run the software on a WISP subsample of sources covered by both grisms but with only one emission line detected above a 2σ threshold (Section 5). We highlight the fact that the algorithm is designed to identify the brightest line detected in a spectrum while it is blind to the possible presence of additional lines besides the strongest one. Consequently, their presence in a spectrum does not influence the calibration and test of the algorithm itself.

3. THE ALGORITHM

3.1. Rationale

The purpose of the algorithm is to provide a probabilistic identification of the strongest emission line observed in a spectrum, by combining various sources of information.

In principle, when only one line is detected in a spectrum, an effective method to identify the line is to compare the spectroscopic redshift expected from different species/ions

³ The high level of noise makes it difficult to clearly identify the strongest line in this spectral region.

(e.g., $z_{H\alpha}$, $z_{[OIII]}$, $z_{[OII]}$, etc.), with the independent redshift solution suggested by an SED fit of the available photometric data. The reliability of this approach, however, strongly depends on the number of available photometric measurements and their wavelength coverage. The precision of such a method rapidly declines when only a few photometric measurements are available. In this case, the photometric redshift PDF will not clearly distinguish between, e.g., the $H\alpha$ and the $[OIII]$ redshift solutions. The presence of multiple peaks in the PDF can further complicate the issue, generating photometric redshift outliers.

Figure 1 (top panel) shows the schematic representation of a WISP grism spectrum (G102 and G141) with only one line detected. The same observation can be equally well described as being due to $H\alpha$, $[OIII]$ or $[OII]$ (cases A, B, and C in the Figure), if the emission due to the other additional lines is below the S/N detection threshold. In the same Figure, the bottom panel illustrates the effects of a poorly constraining multi-peak PDF, on the redshift determination.

For a Gaussian PDF, the *width* of the curve (its value of σ) univocally indicates the uncertainty associated with the value (on the x-axes) corresponding to the peak of the PDF itself. In the more general case, when the shape of the PDF is not Gaussian (e.g., bottom panel of Figure 1), it is difficult to unequivocally describe the width of the probability distribution function with one number. In this case, the redshift range $\Delta z(p)$ corresponding to a given total probability (e.g., $p=95\%$) is a more correct way to represent the uncertainty associated with the most likely value. Alternatively, one could also consider the integral of PDF*, defined as:

$$\int \text{PDF}^*(z) dz = \int \frac{\text{PDF}(z)}{\max(\text{PDF}(z))} dz \quad (1)$$

Lower PDF* integrals or smaller $\Delta z(p)$, however, are not always associated with more precise solutions, such as when a PDF shows multiple peaks. This is true even if each of the peaks can singularly be described by a Gaussian function with a small value of σ . This case represents the typical outlier problem. While each Gaussian peak is characterized by a small σ (high precision), the real uncertainty could be badly underestimated (low accuracy).

On the other hand, some correlations such as the magnitude (or size) versus redshift relation are not particularly tight. Using these quantities as priors to determine the redshift itself thus generates wide PDFs (low precision). Because of this characteristic, however, these methods are less prone to the problem of outliers. In other words, while the uncertainty is high, the measure of the uncertainty is more accurate.

Thus, the most effective way to preserve the precision while maximizing the accuracy (limiting the number of outliers) is to exploit the broader PDFs to identify the most probable peaks in the narrower PDFs. This can be obtained by simply multiplying all the PDFs to each other. Besides allowing the PDFs to be computed, the same parameters can provide estimates of the expected flux ratios. These ratios

can be used by the algorithm to obtain an independent prediction of the most prominent line observed in a spectrum.

In this section, we analyze in detail all the different methods that we eventually combine in our algorithm. Every method is treated as an independent module of the algorithm. In particular, each module represents one single method to obtain information on one only type from one or more input parameters. This approach allows us to more clearly describe the different types of information that the same input parameter can provide, the relations existing between input parameters and outputs, and the limits of each method.

The output of the algorithm we describe below is, for the strongest emission line in each spectrum, a set of indices $P_{\text{transition}}$. These indices are directly related to the probability that the observed emission line is correctly identified to each of the transitions considered ($H\alpha$, $H\beta$, $H\gamma$, $[OIII]$, $[OII]$, $[SII]$)⁴. The transition with the highest value of $P_{\text{transition}}$ will provide the (probabilistically) best redshift solution. We note that, during this process, the algorithm always assumes that the observed line is a real spectral feature and not a spurious detection due to noise or contamination.

3.2. Structure of the algorithm

The modular structure of the algorithm is organized as shown in Figure 2. In brief, the modules can be divided in three main categories, or blocks, depending on the kind of information provided. Using PDFs, the algorithm can estimate the probability associated with each value of z (regression) in the redshift range considered ($0 < z < 3.3$). The flux ratios allow us to forecast which species/transition is most probably responsible for the strongest emission (classification). Finally, comparing the results with the input test sample, the algorithm can automatically fine-tune the final function, in order to increase the accuracy of the entire process (optimization). In a scheme:

- **REGRESSION BLOCK:** photo- z PDFs estimated from SED fitting (Section 3.2.1), J band apparent magnitude (Section 3.2.2), apparent size (Section 3.2.3), and equivalent width of the strongest line (Section 3.2.4);
- **CLASSIFICATION BLOCK:** a set of probability ratios estimated using existing relations between flux ratios and J band apparent magnitude, size, line EW, and J-H color index (Section 3.2.5);
- **OPTIMIZATION BLOCK:** a posteriori fine-tuning of the probability ratios based on the observed wavelength (Section 3.2.6).

⁴ These indices do not correspond to the actual probabilities for the following reason: first, the algorithm considers only the listed species, without taking into account different possibilities. Moreover, the algorithm does not compute the probability that a detected line is not a real emission, but the result of noise or contamination. For these reasons, the indices are normalized so that the highest probability index $P_{\text{transition}}$ is always equal to 1.

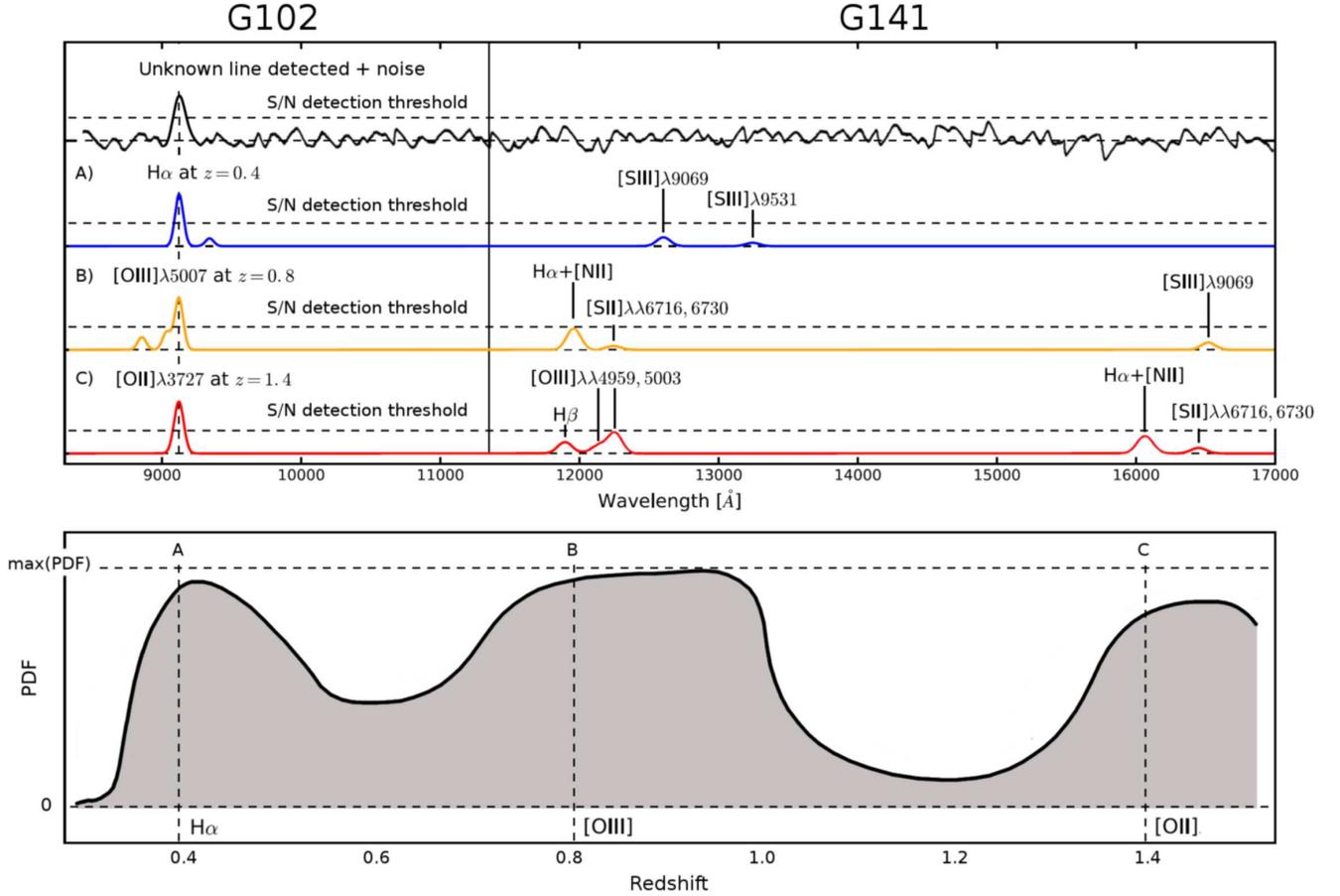


Figure 1. Top panel: In the upper of the four plots, the schematic representation of a WISP grism spectrum with only one line detected above the S/N threshold (uppermost plot). The lower three plots of the top panel represent three different possible solutions: H α (A), [O III] (B), and [O II] (C). **Bottom panel:** a poorly constrained PDF with multiple probability peaks. This kind of PDF would not be able to discriminate among the three possibilities shown in the top panel (A, B, C). This Figure is a modified version of Figure 3 of Bagley et al. (2017).

In this section we describe the single modules and how we calibrated them, while in Section 3.3, we detail on how the modules are combined with each other to output the final probability estimates.

3.2.1. Photo- z PDF from SED fitting

We compute photometric redshifts using the *Hyperz* software (Bolzonella et al. 2000). For each WISP field and for each source, we considered all the measurements in the available photometric bands among the WFC3/IR F110W, F140W, F160W, WFC3/UVIS F606W, F600LP, F814W, and IRAC 3.6 and 4.5 μm filters. Figure 3 shows the number of sources covered by every combination of photometric bands. Given the limited number of bands used in the photometric redshift calculation (larger than 2 only for approximately 55% of the sample), the resulting photometric redshift estimate will not be precise. However, we are not interested in the absolute value of the photometric redshift, typically assumed to correspond to the highest peak of the PDF, but rather in the full shape of the PDF itself. This PDF will be combined with additional probability functions as described in Section 3.3. Therefore, even a photo- z estimation derived

from two bands only (corresponding to one single color index) adds information on the redshift of a source. For example, the poorly constrained PDF shown in the bottom panel of Figure 1 does not allow for a reliable estimate of the photometric redshift. However, it does allow us to safely say that the source considered is *not* located below $z \sim 0.3$ or in the range $1.1 \lesssim z \lesssim 1.3$. This kind of information can be particularly relevant if combined with additional information from other independent sources⁵.

For the *Hyperz* run, we considered a combination of template models from Bruzual & Charlot (2003), characterized by an exponentially declining star formation rate (SFR) with $\text{SFR} \propto \exp(-t/\tau)$, where $\tau = 0.3, 1, 2, 3, 5, 10, 15,$ and 30 Gyr. We consider solar metallicity $Z=Z_{\odot}$ and an extinction law in the Calzetti et al. (2000) form, with A_V ranging from 0.0 to 3.0.

We computed photometric redshifts for all the sources in the sample, regardless of the number of photometric bands

⁵ For example, the upper panel of Figure 15 shows three poorly constrained PDFs. Their combination is shown in the bottom panel of the same Figure.

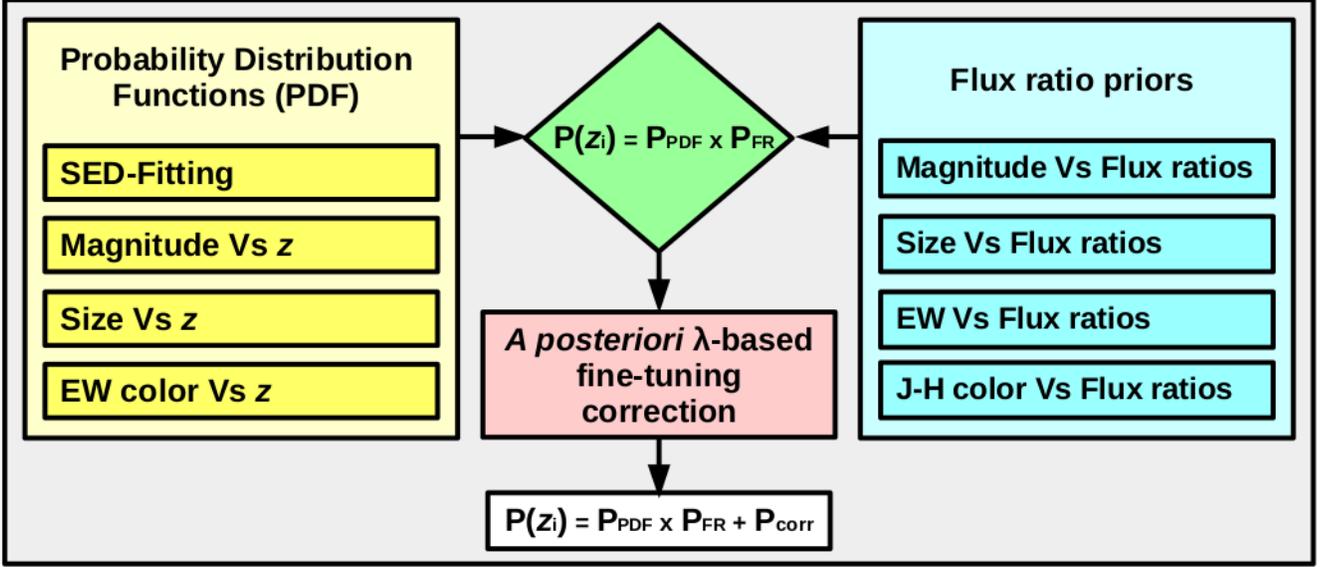


Figure 2. Organization of the modules in the algorithm. The modules can be divided into three categories, or blocks, depending on the kind of information supplied. For each spectrum (for each source), the probability distribution functions PDF(z) provide a continue probability estimation (regression), between $z = 0$ and $z = 3.3$. The line flux ratios can be used to predict which, among the considered species/transitions, is more likely responsible for the strongest line observed (classification). Finally, the probability ratios can be fine-tuned using a λ -dependent *a posteriori* correction (optimization).

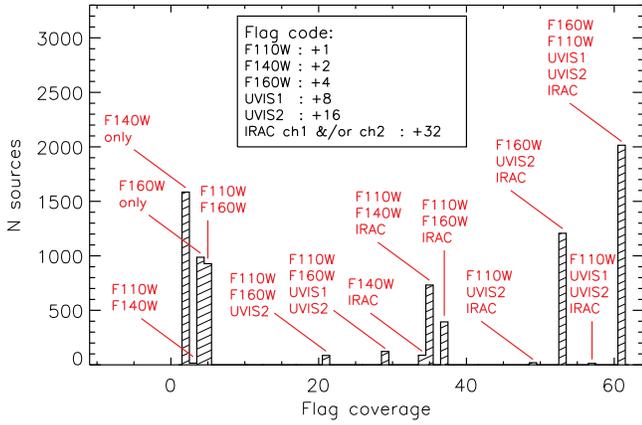


Figure 3. Number of sources versus coverage in the available photometric bands, for the original WISP spectroscopic catalog (at least one spectral line measured). No selection is applied. The coverage flag (x axes) is given by the sum of: +1 when the source is observed through the F110W filter, +2 for F140W, +4 for F160W, +8 for UVIS1, +16 for UVIS2, and +32 for IRAC channel 1 and/or channel 2 coverage. The naming convention “UVIS1” and “UVIS2” is used to represent the bluer and the redder UVIS filters used in a single observation, while a combination of either F475X or F606W and F600LP or F814W were actually used in the WISP survey.

available. The typical output PDF, rarely shows well defined and unique probability peaks. More commonly, multiple peaks are present. In some cases, their presence is due to the degeneracy of the input parameters, generating solutions at

different redshifts with similar probabilities⁶. In other cases, they are a consequence of the discretized grid of models used for the fit, especially when the maxima are close to each other (in redshift). To correct for the effect of model discretization, we smooth the PDFs using a Gaussian filter. Although a good compromise between precision and mitigation of the discretization effect can be obtained with a constant value of σ_z (~ 0.1), we decided to vary σ_z as a function of the integral of the PDF* (see Equation 1), for the reasons explained next.

As we previously discussed, narrow photometric redshift PDFs may still be centered on the wrong redshifts, with resulting uncertainties that are underestimated (resulting in outliers). In general, the overconfidence in the estimation of the z -PDF is a well known problem, especially for what concerns their low-probability tails (see, for example, Wittman et al. 2016). Because of the small (underestimated) uncertainties, the (possibly wrong) photometric redshift solution would prevail over any other redshift indicator. Thus, in these cases, also the emission line identification would also be compromised. To limit the effects of this problem, we smooth the z -PDFs, a similar solution to that adopted by, e.g., Rodríguez-Muñoz et al. (2019). Differently from that work, where a constant $\sigma = 0.2$ is considered, in our Gaussian smoothing, σ is proportional to the inverse of the PDF* integrals and asymptotically converging to $\sigma_z = 0.1$:

$$\sigma_z = 0.1 \times \left(1 + \frac{P_z^*}{\int_{0.0}^{3.3} PDF^*(z) dz} \right). \quad (2)$$

⁶ Often, few photometric optical data can equivalently well be fitted by both the spectrum of an obscured low redshift source or by the spectrum of an high redshift galaxy with low/normal extinction.

Both the best average smoothing factor ($\sigma_z = 0.1$) and the constant value $P_z^* = 0.4$ are empirically determined by maximizing the accuracy of the algorithm when run on the test “gold” sample. Figure 4, shows the value of the variable smoothing factor σ_z as a function of the original photo- z PDF* integral.

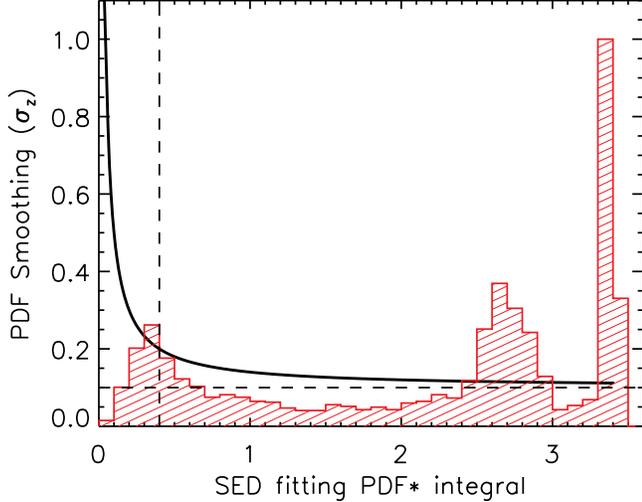


Figure 4. The black curve represents the width of the Gaussian filter used to smooth the photo- z PDF (σ_z in Equation 2), as a function of the PDF* integral $\int_{0.0}^{3.3} \text{PDF}^*(z) dz$. The red histogram shows the distribution of the PDF* integral for the WISP galaxies. Generally speaking, high values of the PDF* integral indicate poorly constrained photometric redshifts (a value of 3.4 corresponds to a PDF* equal to 1.0 everywhere between $z=0$ and $z=3.3$). However, given the possible presence of multiple probability peaks (outliers) and the intrinsic discretization of the grid of models used for the SED fit, the precision of the algorithm is not guaranteed by narrower PDFs. The horizontal dashed line represents the best average smoothing factor able to mitigate the effects of the use of a discretized grid of models without limiting the precision of the final PDFs. In Equation 2 this value corresponds to an asymptotic limit for σ_z ($\sigma_z \rightarrow 0.1$ for high values of the PDF* integral). The vertical dashed line represents the best empirically derived value of P_z^* , for which the smoothing factor σ_z is doubled with respect to its asymptotic value.

The SED fitting method relies on the combination of color indices in different photometric bands. As a consequence, if two different sources are characterized by very different apparent magnitudes or sizes, but identical color indices for all the bands considered, there will be no difference between their output redshifts. Thus, to improve the method, we include the magnitude and size priors described in Sections 3.2.2 and 3.2.3.

3.2.2. Photo- z PDF from F110W (J band) magnitude

Galaxies can not be arbitrarily bright (Schechter 1976). The consequence is that, when surveying a fixed solid angle

Ω , galaxies that *look* very bright are preferentially located at low redshifts, while faint sources more likely correspond to high redshift galaxies.

Figure 5 shows the observed F110W magnitude (hereafter J) as a function of the spectroscopic redshift for WISP sources in the calibration “gold” sample. The algorithm uses the linear fit to these data (solid line in the Figure) as photo- z prior. Specifically, given the J magnitude of a galaxy, the algorithm assumes a Gaussian shaped PDF(z), centered at the redshift indicated by the linear fit, and with a σ equivalent to the horizontal dispersion of the data in the plot of Figure 5 (right panel).

While we computed the linear fit using all the data, when applying the prior we limit the range of validity to: $J_{\min} < J < J_{\max}$, with $J_{\min}=20$ and $J_{\max}=24$. Sources with $J < J_{\min}$ or $J > J_{\max}$ are set to $J=J_{\min}$ and $J=J_{\max}$ respectively. This empirical approach allows us to limit the contamination of the numerically small sample of sources dominated by the [O II] emission, from sources leaking from the tails of the numerically more consistent [O III] and $H\alpha$ distributions.

As shown in Figure 3, many sources are not covered by observations in the F110W band. However, to first approximation, we can convert between the H (F140W or F160W) and the J band magnitudes, as shown in Figure 6. In this figure we plot the magnitude-dependent correction that can be applied to the measurements obtained using the F140W and F160W filters, to recover the magnitude expected in the F110W band. The same figure also shows that the magnitude correction does not depend on which emission line is the strongest in the spectrum.

3.2.3. Photo- z PDF from apparent size

The mass-size relation (e.g. Poggianti et al. 2013; van der Wel et al. 2014) and the shape of the stellar mass function (see e.g. Lapi et al. 2017, and Figure 4 therein) indicate that the same arguments used for the apparent magnitude can be applied to the apparent size as well: apparently larger sources are more likely located at lower rather than higher redshifts. Following the Mattig equation (Mattig 1958), the apparent size of a galaxy decreases with the distance, up to $z \sim 1.5$.

In Figure 7, we show the galaxies’ apparent size (A_IMAGE⁷) computed in the J band (F110W), as a function of the spectroscopic redshift. We include only sources in the WISP calibration “gold” sample. The linear fit to the size-magnitude relation is used as an additional redshift prior in our algorithm. In particular, given the apparent size of a source, we assume a Gaussian PDF in redshift, centered at the redshift indicated by the linear fit, and with a σ equivalent to the local horizontal dispersion of the data in the plot of Figure 7 (right panel).

Also in this case, we computed the linear fit using all the data, but when applying the prior we limit

⁷ The A_IMAGE parameter is an output of the SExtractor software corresponding to the RMS of the luminosity distribution of a galaxy, measured in pixels, along the semi-major axes.

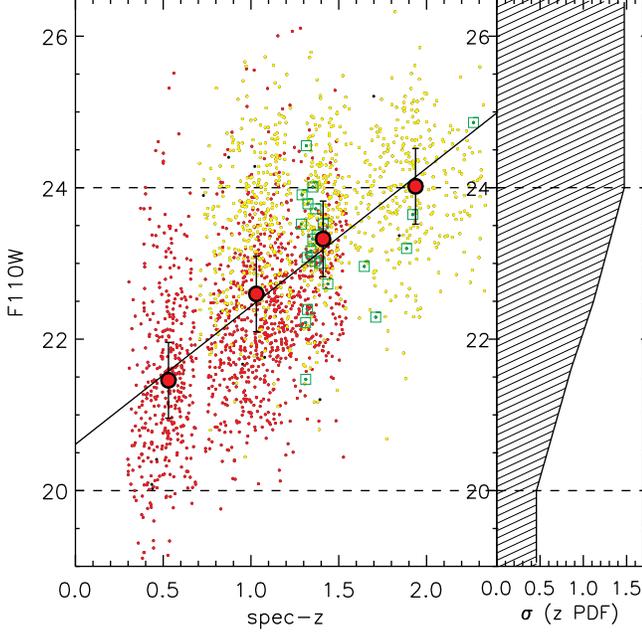


Figure 5. Left panel: Relation between AB magnitude measured in the F110W band (J) and spectroscopic redshift for the calibration “gold” sample. All the data are considered for the linear fit, that we use as a prior in the algorithm. However, the prior does not consider values of magnitude $J < J_{\min}$ or $J > J_{\max}$ (dashed lines): in these cases, the prior assumes $J = J_{\min} = 20$ and $J = J_{\max} = 24$ respectively. This method limits the contamination of the numerically small sample of sources dominated by the [O II] emission, from [O III] dominated sources. The actual nature of the strongest emission line is represented with yellow circles for [O III], red circles for $H\alpha$ and green circles surrounded by squares for [O II]. **Right panel:** given the J magnitude of a source, the PDF is assumed to be a Gaussian function, with σ equivalent to the horizontal dispersion of the data around that specific value of magnitude (± 0.5 mag).

the range to: $A_IMAGE_{\min} < A_IMAGE < A_IMAGE_{\max}$, with $A_IMAGE_{\min} = 2.5$ pixels and $A_IMAGE_{\max} = 10$ pixels. Sources with sizes larger or smaller than these limits are set to $A_IMAGE = A_IMAGE_{\max}$ and $A_IMAGE = A_IMAGE_{\min}$, respectively. When no images are available in the J band, the A_IMAGE parameter is computed using the H band image. Given the negligible differences in the J and H PSFs and sampled stellar populations, no corrections are considered in these cases.

3.2.4. Photo-z PDF from apparent equivalent width

Figure 8, shows the relation existing between the apparent equivalent width of the measured lines and the spectroscopic redshift of the sources. As we do for the apparent J magnitude and size priors (Sections 3.2.2 and 3.2.3 respectively), we compute a linear fit to the observed relation (solid line in the Figure). The best fitting relation is used as a redshift prior by our algorithm. In particular, given the EW of an observed emission line, we consider a Gaussian PDF centered at the

redshift indicated by the linear fit and with a value of σ given by the local horizontal dispersion (right panel of Figure 8).

Also in this case, we limit the range of validity to $\log(EW_{\min}) < \log(EW) < \log(EW_{\max})$, where $\log(EW_{\min}) = 1.75$ and $\log(EW_{\max}) = 2.75$. It is possible to observe (Figure 8) that these thresholds allows us to approximately separate sources dominated by the $H\alpha$ and [O II] emission from those dominated by [O III], helping to limit the contamination of the [O II] sample from the other more numerous samples.

3.2.5. Line flux ratio priors

Unless the z -PDF is particularly narrow, it is difficult to automatically classify an emission line using just the PDF itself. For example, while $\lambda_{H\alpha} \sim \lambda_{[SII]}$, it is very unlikely that [S II] can be the unique single line observed in a spectrum, because $H\alpha$ is commonly brighter than [S II]. In general, detecting the $H\alpha$ emission line in a randomly selected spectrum is more common than detecting [O III], and [O III] is more common than [O II] or other lines such as the fainter $H\beta$, $H\gamma$, or [S II]. Then, in order to obtain a proper classification, the expected flux ratios between different lines can be used as an additional source of information.

The average line flux ratios measured in the calibration “gold” sample allows us to rescale the detection probabilities of the different species/transitions considered. For example, let us assume the case of a spectrum with only one line detected, and a similar value of the photo- z PDF measured at $z_{H\alpha}$ and $z_{[SII]}$, with z -PDF ~ 0 for all the other observable emission lines. Since the $H\alpha$ emission is always stronger than that due to [S II], the line detected must be classified as $H\alpha$. Because the observed wavelength is known (λ_i^{obs}), identifying the nature of the (strongest) emission line measured ($i = H\alpha$, in this example) automatically provides a measure of the spectroscopic redshift: $z = z_{H\alpha} = (\lambda_{H\alpha}^{\text{obs}} / \lambda_{H\alpha}) - 1$.

For each of the species/transitions i considered, we computed the average observed flux ratio $F_i / F_{H\alpha}$ as a function of J magnitude, apparent size, equivalent width, and J-H color. This flux ratio is calculated for sources in the calibration “gold” sample and in the wavelength range of observability of the emission lines⁸. As shown in Figures 9 and 10, the line flux ratios show some correlation with all the parameters considered. Thus, if one or more of the parameters considered are available, it is possible to compute the expected relative strength of different lines. This information can be directly translated into the relative probability ($P_{FR}(i)$) that a given species/transition, i , corresponds to the strongest line measured in the spectrum. In particular, $P_{FR}(i)$ corresponds to the weighted average of the various $P_{FR}^j(i)$ obtained from

⁸ We do not exploit the ratio between [O III] and [O II] fluxes, computed in the calibration “gold” sample, above $z \sim 1.50$. This is due to a suspected bias in the original classification of strong [O II] emitters above this redshift. When $H\alpha$ is redder than the upper λ limit of the G141 grism, and if the [O III] and $H\beta$ lines are weaker than the [O II] emissions, the latter tends to be wrongly misidentified as $H\alpha$, especially when the S/N ratio is low.

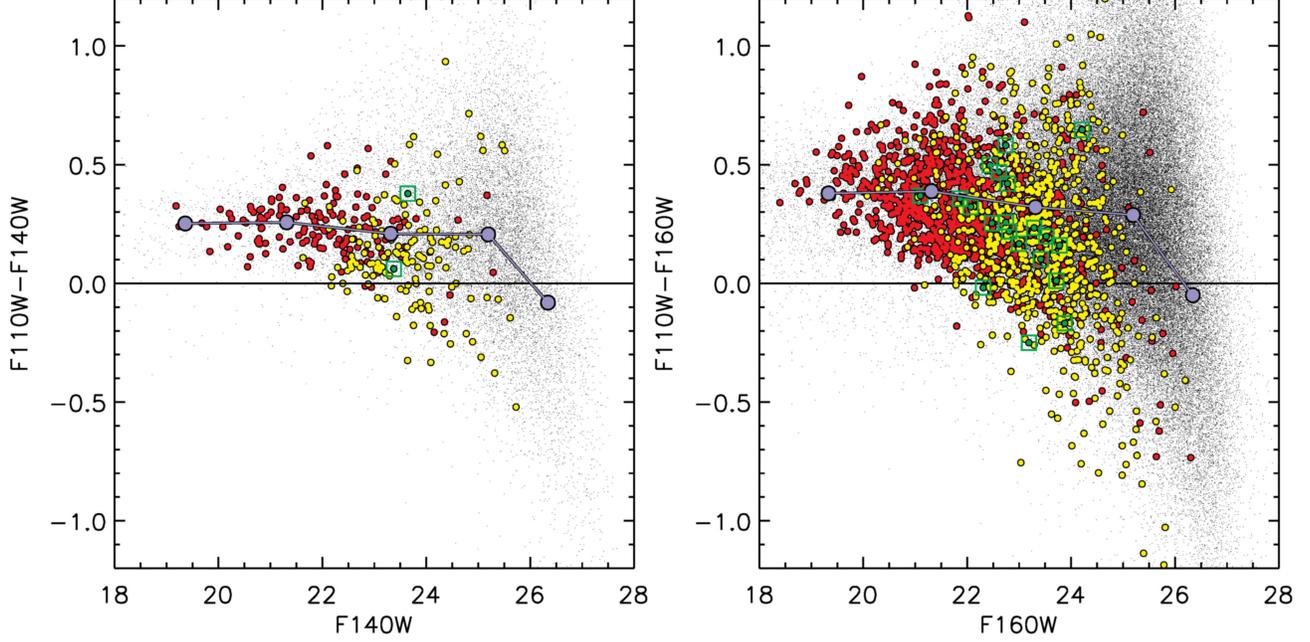


Figure 6. Left panel: J-H (F110W-F140W) color index as a function of the F140W magnitude. We use this color index to infer the expected F110W magnitude from the F140W (see Section 3.2.2). The correction is weakly dependent on magnitude, and does not depend on the identification of the strongest emission. Red circles show H α , yellow circles [O III], green circles surrounded by a square [O II], black dots for no lines detected. The average value of the color index is represented using light purple lines and filled circles, in 5 bins of magnitude. **Right panel:** J-H color index computed for the F110W-F160W combination, as a function of F160W.

each of the different indicators j considered (J magnitude, apparent size, equivalent width and J-H color):

$$P_{\text{FR}}(i) = [P_{\text{FR}}^{\text{J}}(i)W_{\text{FR}}^{\text{J}}(i) + P_{\text{FR}}^{\text{size}}(i)W_{\text{FR}}^{\text{size}}(i) + P_{\text{FR}}^{\text{EW}}(i)W_{\text{FR}}^{\text{EW}}(i) + P_{\text{FR}}^{\text{J-H}}(i)W_{\text{FR}}^{\text{J-H}}(i)] / \sum_j W_{\text{FR}}^j(i), \quad (3)$$

The weights $W_{\text{FR}}^j(i)$ are the square inverse of the dispersion of the relations, computed as a function of the reference parameters (see upper panels of Figures 9 and 10):

$$W_{\text{FR}}^j(i) = \sigma_j^{-2} \quad (4)$$

Given the observed wavelength of the strongest emission line λ_{obs} , the spectroscopic redshift solution depends on the unknown nature of the species/transition i responsible for such an emission as $z_i = (\lambda_{\text{obs}}/\lambda_i) - 1$, where λ_i is the rest frame wavelength of the species/transition i . Hence, the algorithm considers both the probability $P_{\text{PDF}}(z_i)$, associated with z_i being the correct redshift solution, and $P_{\text{FR}}(i)$, corresponding to the species/transition i being the responsible for the strongest emission:

$$P(z_i) = P_{\text{FR}}(i) \times P_{\text{PDF}}(z_i). \quad (5)$$

The previous equation does not provide the complete description of the algorithm, for which we refer to Section 3.3. In fact, the complete computation includes the *a posteriori* optimization (Section 3.2.6) and the limitations described in Section 3.2.7

3.2.6. Observed wavelength prior

The calibration of all the methods described is performed by using the high quality data of the calibration "gold" sample. However, since these data correspond to real observations, the phase space used to classify the emission lines can not be homogeneously calibrated. Additionally, we use a large but simplified (quasi-linear) set of functions describing the correlations between the observational parameters and the outputs. While this approach prevents us from *overfitting* the calibration data set (because the output does not depend on a discretized configuration of the inputs, but on their weighted average), the same method results in an uneven precision along the dimensions of the parameter space.

In particular, possible biases could be introduced at different λ_{obs} (i.e. at different redshifts). For example, while the average flux ratios are calibrated as a function of magnitude, color, size, and EW, no redshift evolution is taken into account for these relations, since there is not *a priori* knowledge of the redshift itself. Similar biases are visible in the combined photo- z PDF, showing systematic over- and under-estimations of the photometric redshift at different values of the spectroscopic reference.

Some of these redshift dependent biases can be corrected by using the wavelength of the observed lines as a prior. Combining all the methods described in the previous sections, the algorithm computes probability ratios between the different species (the probability associated with the most probable species is assumed equal to 1.0). We can exploit the wavelength position of the detected lines for *a posteriori*

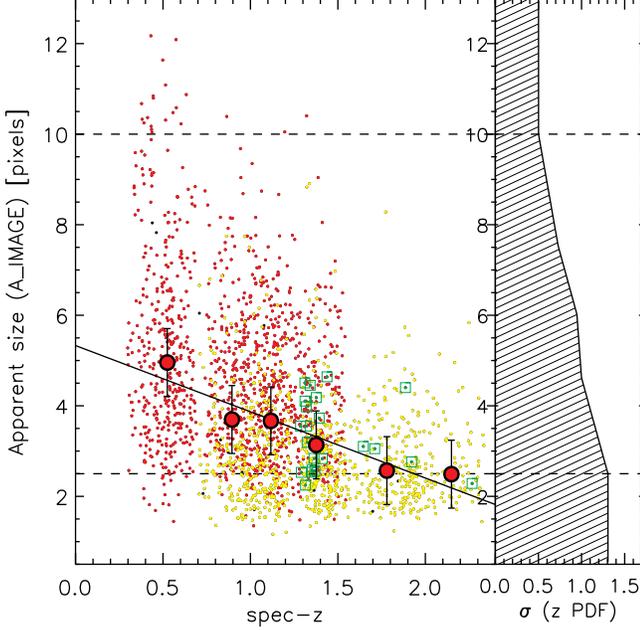


Figure 7. Left panel: Relation between apparent size ($S_{\text{Extractor}} A_{\text{IMAGE}}$ parameter) measured in the F110W band (J) and spectroscopic redshift, for the calibration “gold” sample. All the data are used to compute the linear fit, that we use as a prior in the algorithm. However, the prior does not consider values of A_{IMAGE} $A < A_{\text{min}} = 2.5$ pixels or $A > A_{\text{max}} = 10$ pixels (dashed lines): in these cases, the prior assumes $A = A_{\text{min}}$ and $A = A_{\text{max}}$ respectively. The actual nature of the strongest emission line is represented with yellow circles for [O III], red circles for $H\alpha$ and green circles surrounded by squares for [O II]. **Right panel:** given the value of A_{IMAGE} of a source, the PDF is assumed to be a Gaussian function, with σ equivalent to the horizontal dispersion of the data around that specific value of A_{IMAGE} (± 0.75).

fine-tuning of these probability ratios. This approach allows for the optimization of the algorithm.

In the WISP survey, the strongest emission lines measured are mostly either $H\alpha$ or [O III] (1293 and 949 objects, respectively, in the test “gold” sample). For an additional small fraction of sources, [O II] is the most prominent line (24 objects in the test “gold” sample). For this reason, to improve the overall accuracy of the algorithm, the most effective approach consists in the fine-tuning of the ratio between the probabilities computed for $H\alpha$ and [O III]. Secondly, another correction can be applied to the ratio between the [O III] and [O II] probabilities to further improve the performances of the algorithm.

In the left panel of Figure 11, we show the original $H\alpha$ over [O III] probability ratio obtained for all the sources of the test “gold” sample, before the fine-tuning correction is applied: $P_{H\alpha/[OIII]} = P(H\alpha)/P([O III])$. It is possible to see that the majority of the emission lines are already correctly identified ($P_{H\alpha/[OIII]} < 0$ for [O III] and $P_{H\alpha/[OIII]} > 0$ for $H\alpha$). However, these results can be improved by deriving, for each bin of observed wavelength, the value of $P_{H\alpha/[OIII]}$ that better

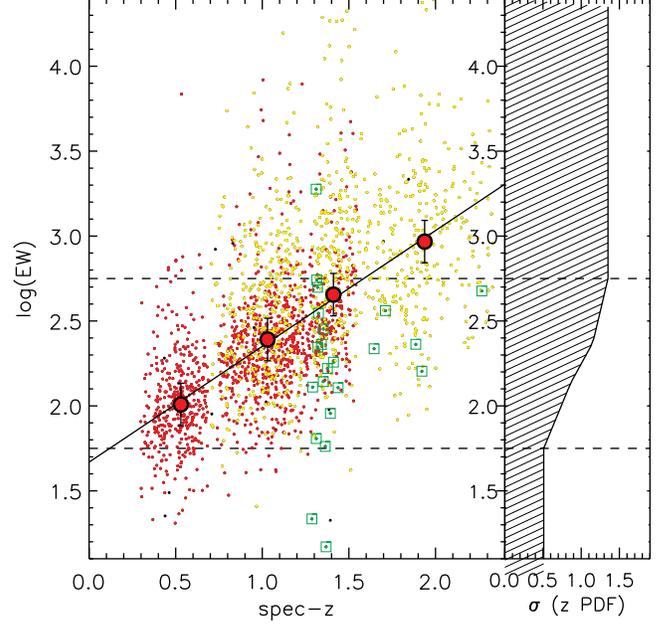


Figure 8. Left panel: Relation between apparent equivalent width and spectroscopic redshift, for the calibration “gold” sample. All the data are used to compute the linear fit, that we use as a prior in the algorithm. However, the prior does not consider values of equivalent width $\log(EW) < \log(EW_{\text{min}}) = 1.75$ or $\log(EW) > \log(EW_{\text{max}}) = 2.75$ (dashed lines): in these cases, the prior assumes $\log(EW) = \log(EW_{\text{min}})$ and $\log(EW) = \log(EW_{\text{max}})$ respectively. The nature of the strongest emission line is represented with yellow circles for [O III], red circles for $H\alpha$ and green circles surrounded by squares for [O II]. **Right panel:** given the value of EW, the PDF is assumed to be a Gaussian function, with σ equivalent to the horizontal dispersion of the data around that specific value of EW (± 0.125).

separates the two emission lines: $P_{H\alpha/[OIII]}^{\text{best}}(\lambda_{\text{obs}})$ (black line in the left panel of Figure 11). In Table 1, we report the tabulated values of $P_{H\alpha/[OIII]}^{\text{best}}(\lambda_{\text{obs}})$ that we computed in equally spaced wavelength bins ($\Delta\lambda = 500\text{\AA}$) ranging from 8500 to 16500 \AA (the left column in Table 1 represents the mean λ of the spectral lines in each bin). The right panel of Figure 11, shows the data distribution after the application of the fine-tuning correction $P_{H\alpha/[OIII]}^{\text{best}}(\lambda_{\text{obs}})$.

While we do not modify the value of $P_{H\alpha}$, the correction factor $P_{H\alpha/[OIII]}^{\text{best}}(\lambda_{\text{obs}})$ is applied to both $P_{[OIII]}$ and $P_{[OII]}$. This choice allows us to keep unaltered the probability ratio between [O III] and [O II] that is independently corrected.

Similarly to what was done for the $H\alpha$ and [O III] pair, we applied a correction to the [O II] over [O III] probability ratio. In the left panel of Figure 12, we show the original [O III] over [O II] probability ratio obtained for all the sources of the test “gold” sample, before the fine-tuning correction is applied: $P_{[OIII]/[OII]} = P([O III])/P([O II])$. In this case, given the lack of sources with prominent [O II] emission above $\lambda_{\text{obs}} \sim 9000\text{\AA}$, the probability ratio can not be

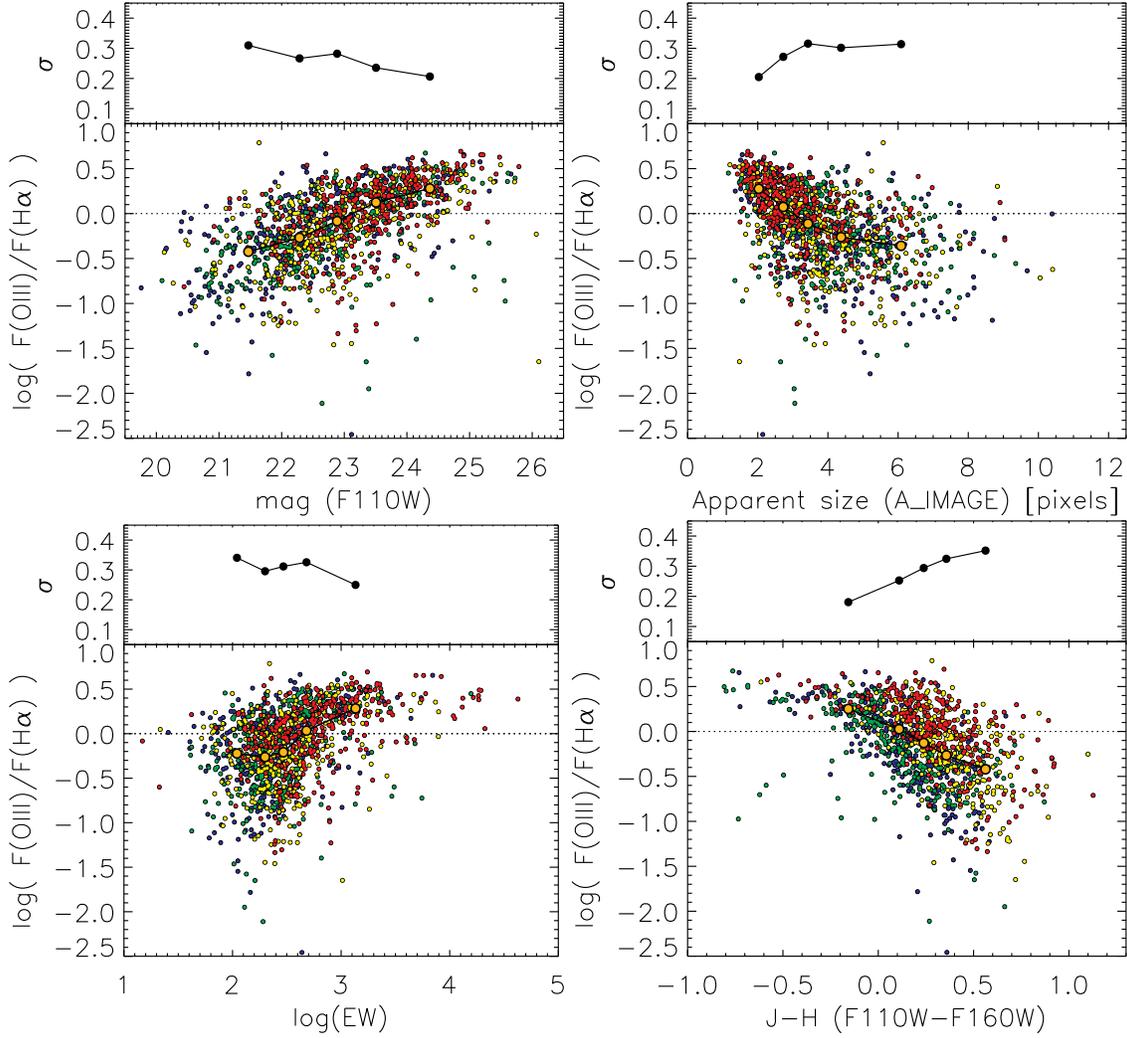


Figure 9. Ratio between $H\alpha$ and $[O\text{ III}]$ fluxes, measured as a function of apparent F110W magnitude (top left panel), size (top right), equivalent width (bottom left) and $J-H$ (F110W-F160W) color index (bottom right), for the sources in the calibration “gold” sample. A similar relation is computed using also the F110W-F140W color index, but it is not shown in these plots. For each relation, the dispersion (σ) is shown in the upper parts of each panel. The possible dependence of these relations on the redshift can not be directly taken into account, since the redshift is not known *a priori*. However, using the observed wavelength prior (Section 3.2.6), it is possible to indirectly correct for such a possible effect. The relations are shown for 4 different bins of redshifts, from $z \sim 0.87$ to $z \sim 1.42$. In order of increasing redshift, blue ($< z > \sim 0.87$), green ($< z > \sim 1.05$), yellow ($< z > \sim 1.22$), red ($< z > \sim 1.43$).

precisely fine-tuned. However, above $\lambda_{\text{obs}} \sim 11500\text{\AA}$, a systematic deviation of this ratio from $P_{[\text{OIII}]/[\text{OII}]}=0$ is immediately evident and easy to correct. In Table 2, we report the values of $P_{[\text{OIII}]/[\text{OII}]}$ that we applied to $P([\text{O II}])$ to fine tune the $P([\text{O III}])/P([\text{O II}])$ ratio.

3.2.7. Limitations

The wavelength range covered by WISP is limited. For this reason, the expected flux ratios can not always be used to identify the strongest *measured* line. For example, even if the flux ratios indicate that $H\alpha$ is the *expected* strongest line, $H\alpha$ may fall outside of the observable wavelength range. The brightest *measured* line may be due to other species/transitions, (e.g. $[O\text{ III}]$ or $[O\text{ II}]$).

In particular, $H\alpha$ falls outside the grism at $z > 1.6$, corresponding to $\lambda_{\text{obs}}(H\alpha) > 17000\text{\AA}$ and $\lambda_{\text{obs}}([O\text{ III}]) > 12900\text{\AA}$. Therefore, when 1) the observed wavelength of the strongest (single) observed emission line is $\lambda_{\text{obs}} > 12900\text{\AA}$, and 2) the expected $F_{H\alpha}/F_{[OIII]}$ flux ratio is > 1 , there is a non-negligible chance that $H\alpha$ is indeed stronger than $[O\text{ III}]$ (and $[O\text{ II}]$), but outside the range of observability. Given these circumstances, when the conditions 1) and 2) are verified at the same time, we set $P_{\text{FR}}(H\alpha)/P_{\text{FR}}([OIII]) = 1$, while keeping the values computed for $P_{\text{FR}}([OIII])$ and $P_{\text{FR}}([OII])$. This is equivalent to not considering the expected flux ratio between $H\alpha$ and $[O\text{ III}]$.

Similarly to the pair $H\alpha - [O\text{ III}]$, the same problem is expected to affect the $[O\text{ III}] - [O\text{ II}]$ pair when $[O\text{ III}]$ falls

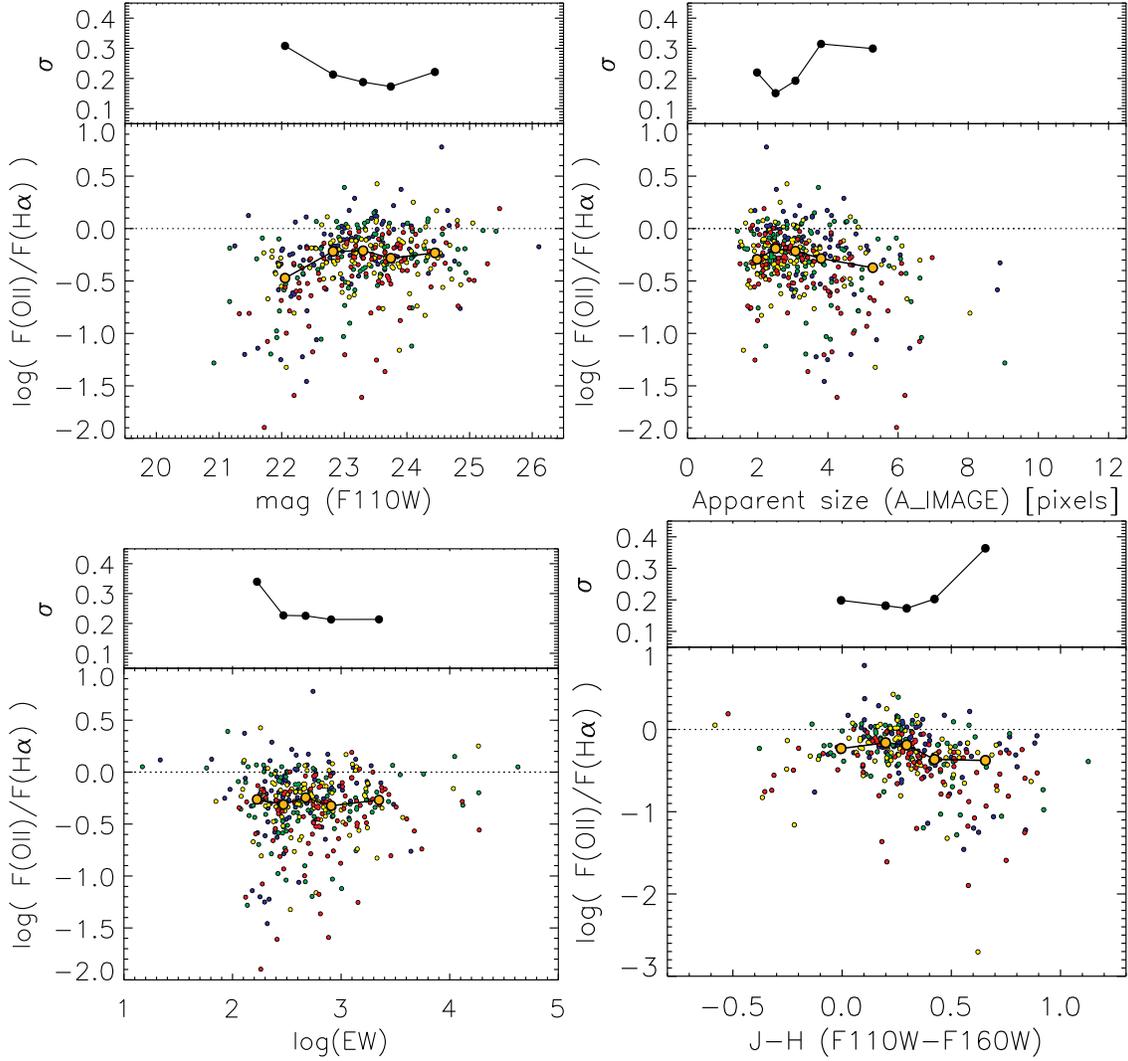


Figure 10. Ratio between $H\alpha$ and $[O II]$ fluxes, measured as a function of apparent F110W magnitude (top left panel), size (top right), equivalent width (bottom left) and J-H (F110W-F160W) color index (bottom right), for the sources in the calibration “gold” sample. A similar relation is computed using also the F110W-F140W color index, but it is not shown here. For each relation, the dispersion (σ) is shown in the upper parts of each panel. The possible dependence of these relations on the redshift can not be directly taken into account, since the redshift is not known *a priori*. However, using the observed wavelength prior (Section 3.2.6), it is possible to indirectly correct for such a possible effect. The relations are shown for 4 different bins of redshifts, from $z \sim 1.31$ to $z \sim 1.50$. In order of increasing redshift, blue ($\langle z \rangle \sim 1.32$), green ($\langle z \rangle \sim 1.37$), yellow ($\langle z \rangle \sim 1.43$), red ($\langle z \rangle \sim 1.50$).

outside the range of observability. However, this mistake is only theoretically possible, since the depth of the WISP survey make the observability of a $z \gtrsim 2.5$ source unlikely. For this reason, we do not apply any correction to the expected $P_{FR}([OIII])/P_{FR}([OII])$ ratio.

On the opposite side of the wavelength range covered by the grisms, the same problem can arise if $[O III]$ is the *intrinsic* strongest line but it is located below $\lambda \sim 8000\text{\AA}$, corresponding to $\lambda_{obs}(H\alpha) < 10536\text{\AA}$. While the flux ratios could correctly indicate $[O III]$ as the expected strongest line, the strongest *measured* line that the algorithm must indicate is $H\alpha$. We estimated the effect of this possibility by excluding the flux ratios when $\lambda < 10536\text{\AA}$ and $F_{H\alpha}/F_{[OIII]} < 1$. The results that we obtain confirm that the overall effect is

negligible on both the test sample (accuracy $\sim 81.7\%$) and the “single line” sample (for which the recovered distribution in z does not change significantly). For this reason we do consider the flux ratios indicators also when $\lambda < 10536\text{\AA}$.

We emphasize on the fact that the probability P_{FR} , obtained from the expected flux ratios and corrected as described above, is not used alone. In fact, the outputs of the *classification* block are combined with the outputs of the *regression* and *optimization* blocks, as we will better describe in detail in Section 3.3.

The test “gold” sample includes 632 sources for which the limitations described above apply, out of a total of 2283 galaxies. For this subsample, after applying the correction described, we measure an accuracy (i.e. fraction of spectral