“PS1-STRM: Neural network source classification and photometric redshift catalogue for PS1 3π DR1”

Elias Kyritsis

Journal Club-30/1O/2020

Institute of Astrophysics, Crete

Pan-STARRS

[PanSTARRS dome, Created by Jeff Valenti]

PS1-STRM: Neural network source classification and photometric redshift catalogue for PS1 3π DR1

Róbert Beck[1,2]*, István Szapudi[1,2], Heather Flewelling[1], Conrad Holmberg[1,3], Eugene Magnier[1]

[1] *Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, HI, 96822, USA*

[2] *Department of Physics of Complex Systems, Eötvös Loránd University, Pf. 32, H-1518 Budapest, Hungary*

[3] *Platform Services, Stanford Health Care, 300 Pasteur Drive, Stanford, CA, 94305, USA*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The Pan-STARRS1 (PS1) 3π survey is a comprehensive optical imaging survey of three quarters of the sky in the *grizy* broad-band photometric filters. We present the methodology used in assembling the source classification and photometric redshift (photo-z) catalogue for PS1 3π Data Release 1, titled Pan-STARRS1 Source Types and Redshifts with Machine learning (PS1-STRM).

For both main data products, we use neural network architectures, trained on a compilation of public spectroscopic measurements that has been cross-matched with PS1 sources.

We quantify the parameter space coverage of our training data set, and flag extrapolation using self-organizing maps. We perform a Monte-Carlo sampling of the photometry to estimate photo-z uncertainty.

The final catalogue contains $2,902,054,648$ objects. On our validation data set, for non-extrapolated sources, we achieve an overall classification accuracy of 98.1% for galaxies, 97.8% for stars, and 96.6% for quasars.

Regarding the galaxy photo-z estimation, we attain an overall bias of $\langle \Delta z_{\mathrm{norm}} \rangle = 0.0005$, a standard deviation of $\sigma(\Delta z_{\mathrm{norm}}) = 0.0322$, a median absolute deviation of $\mathrm{MAD}(\Delta z_{\mathrm{norm}}) = 0.0161$, and an outlier fraction of $O = 1.89\%$.

The catalogue will be made available as a high-level science product via the Mikulski Archive for Space Telescopes at https://doi.org/10.17909//t9-rnk7-gr88.

Key words: catalogues – cosmology: large-scale structure of Universe – methods: data analysis – methods: numerical.

Introduction

Optical broad-band photometry ⟹ Very important for gathering information from the Universe

Numerous surveys are dedicated to obtain images and/or spectra !

E.g. SDSS

- Combination of spectroscopic + imaging measurements
- Coverage ≃ 14.000 deg^2 of the sky
- Results to ≃ 7.700 peer reviewed papers

Main usage of optical photometry ⟹ Distinguishing of astronomical objects to

Stars Galaxies QSOs

Not a trivial task if only info from broadband photometry is available !!!

Introduction

Traditional approaches for the source classification

- **Galaxy/ Non galaxy separation:** Galaxies are extended sources ⇨ A cut in their PSF aperture mag vs extended aperture mag. Problematic cases: Faint objects
- **QSO/Star separation:**
  - Cuts in colour - colour diagrams (Low redshift QSOs)
  - Optical & Infrared observations / Time-domain observations (High redshift OSOS)

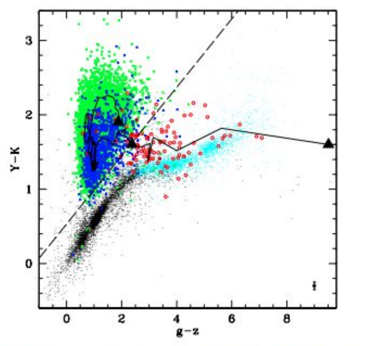


Figure 5. The $Y-K$ vs. $g-z$ colour-colour diagram of the SDSS-UKIDSS quasar and star samples. Green crosses, blue crosses and red circles denote quasars with $z \leqslant 2.2$, $2.2 < z \leqslant 4$ and $z > 4$, respectively. Black dots and cyan dots denote stars classified as 'STAR' and 'STAR_LATE' by SDSS spectroscopy. Dashed line indicates the separation criterion $Y-k > 0.46*(g-z)+0.53$. Solid curve is derived from the median colour-z relation (see section 3) of quasars. The solid triangles, from left to right, mark the colours of quasars with z=0.02, 4, and 5 respectively, derived from the colour-z relation. Typical error bars are shown in the lower-right corner.

[Wu & Jia, 2010]

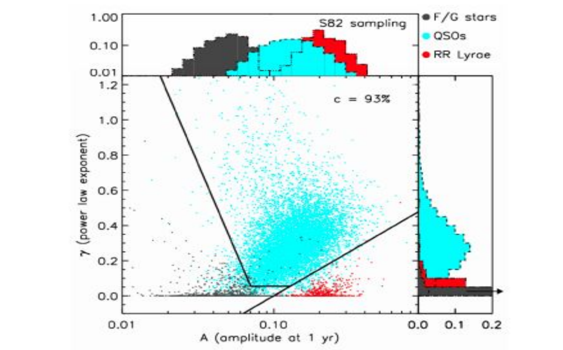


FIG. 5.— Distribution of the variability structure function parameters A and γ (Equation 3) for ~15,000 individual objects in Stripe 82. The spectroscopically confirmed quasars are shown as light blue points; confirmed RR Lyrae and color-selected F/G stars are shown in red and grey respectively. A separation of the quasars from the stellar locus contaminants is clearly seen. The three solid lines (Equations 7–9) define the region in which we estimate the quasar completeness c of our algorithm (which turns out to be 93% in this case, Table 5). Along the axes we show the projected A and γ distributions for the sub-samples.

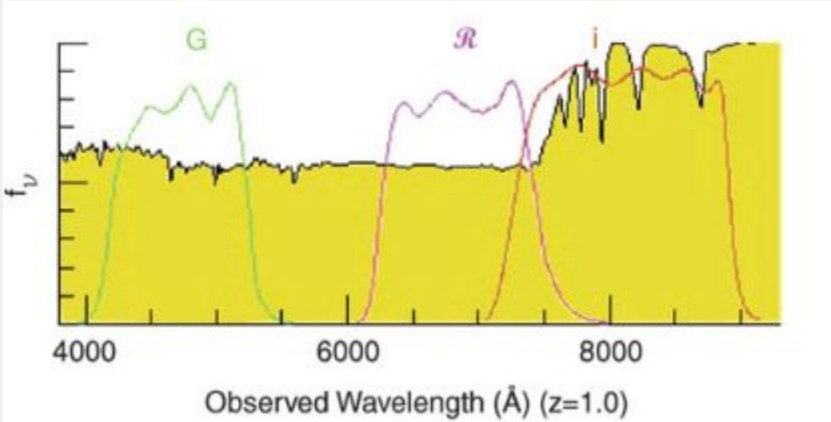
[Schmidt et al., 2010]

Introduction

Traditional approach for the photo-z estimation

Photo-z : Very important for the galaxies. Useful for distance measurements !

- Template fitting approaches



galaxies at other redshifts is also possible. The example shows a galaxy at $z = 1$ whose 4000 Å-break is located between the two redder filters. The 4000 Å-break occurs in stellar populations after several 10^7 yr (see Fig. 3.33) and is one of the most important features for the method of photometric redshift. Source: K.L. Adelberger 1999, *Star Formation and Structure Formation at Redshifts* $1 < z < 4$, astro-ph/9912153,

[Schneider P., Extragalactic Astronomy & Cosmology,2nd Ed, 2015]

Phot-z estimation with templates

1. A number of template-spectra (observations or populations synthesis models) is redshifted in λ .
2. For each template-spectrum and any z the expected galaxy colours are determined
3. This set of colours is compared with the observed galaxy colours.
4. The best match determines the galaxy’s z and type

Introduction

Problems with the traditional methods of source classification & photo-z estimation

- **Classification:** Boundary definitions, useful photometric bands, apertures ⟹ Based on a few low-dimensional projections of a complex high-dimensional space

- **Photo-z estimation:** Insufficient number of different filters, errors on magnitudes measurements

Solution

⇓

MACHINE LEARNING METHODS

- **Classification:** ML methods make automatically choices of photometric bands, aperture size, etc. based on a the entire multi-dimensional parameter space

- **Photo-z estimation:** ML methods seem to be more accurate than template fitting methods, when there is large number of spectro-z data available for the model's calibration

Introduction

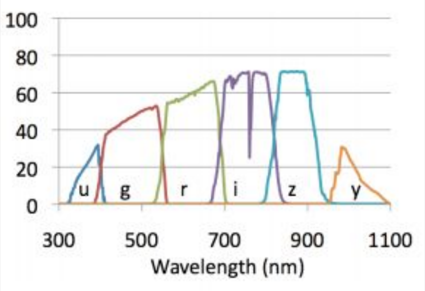
This work

- Based on Pan-STARRS1 3π survey Data Release 1
  - The currently largest imaging survey
  - Coverage ≈ 30.000 deg^2 of the sky
  - 10.7 billion unique objects - 3 billion sources confirmed in multiple bands

- Creation of ***Pan-STARRS1 Source Types and Redshifts with Machine Learning (PS1-STRM)*** by using machine learning for source classification & photo-z estimation

Data sets

Photometric data



[https://www.lsst.org/]

Broad-band photometric measurements : *g,r,i,z,y*

Photometry methodologies:

- Mean photometry: based on single-epoch detections
- Stack photometry: stacking all observations in a given(field and filter)
- Forced mean photometry: Objects detected in the stacks, not necessarily detected in single exposures

Which one is the best ?

For the purposes of a uniform classification and photo-z catalog ⟹ Forced Mean Photometry

Why ?

Deepest & accurate photometry available for all sources

Sufficient aperture magnitude

6 different photometries::

- PSF (describes well the flux from stars & QSOs)
- Kron (describes well the flux of extended sources)
- seeing-matched apertures
- 3.00” , 4.63”, 7.43”

6 different photometric values x 5 bands

⇓

30 measures of the multi-band flux of a light source

Data sets

Spectroscopic data

Table 1. The cross-matched source counts and used quality flags of the different surveys comprising our combined spectroscopic sample.

Survey	Total source count	Galaxies	Stars	Quasars	Quality flags
SDSS DR14	3, 616, 323	2, 310, 690	766, 251	539, 382	zWarning = 0x00, 0x10
DEEP2 DR4	18, 636	17, 143	631	862	ZQUALITY = 4
VIPERS PDR-2	53, 833	51, 523	2, 310	-	$\lfloor$zflg$\rfloor$ (mod 10) = 3, 4
WiggleZ	146, 686	146, 647	39	-	Q = 4, 5
zCOSMOS DR3	11, 867	11, 125	742	-	$\lfloor$CC$\rfloor$ (mod 10) = 3, 4
VVDS	6, 374	6, 374	-	-	ZFLAGS (mod 10) = 4
Combined	3, 853, 719	2, 543, 502	769, 973	540, 244	

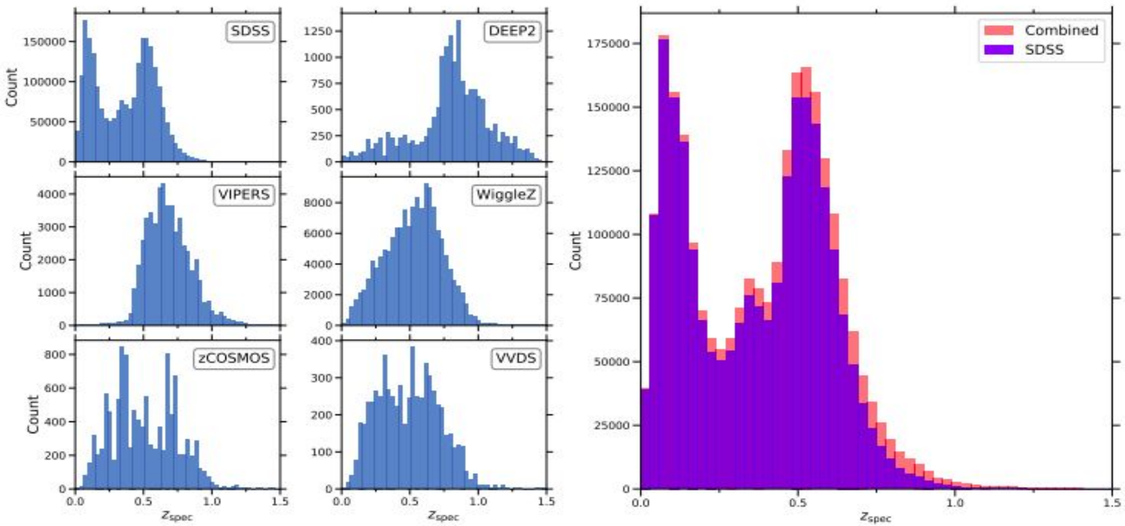


Figure 1. The redshift distribution of galaxies in the spectroscopic surveys that constitute our combined spectroscopic sample. Left panel: surveys are shown individually. Right panel: the combined sample is plotted, as well as the only-SDSS component.

Class & z : Determined by detailed analysis of high-resolution spectra

Cross-matching of spectroscopic sources with the PS1 3π DR1 and acceptance of only certain matches (closer than 1.5” ,Bayes factor > 10.000)

Keep only high quality spectra

Data sets

Dust maps

Photometry in PS1 3π DR1 in not corrected for extinction !

Data augmentation with 2 extra data sets:

1. PS1 observations of Galactic stars. Tracking the reddening until 4.5 kpc
2. Planck CMB (accounts for overall extinction)

Total Number of features for each object

6 different apertures x 5 bands

+

2 E(B-V) values from PS1 & Planck CMB

$\Downarrow$

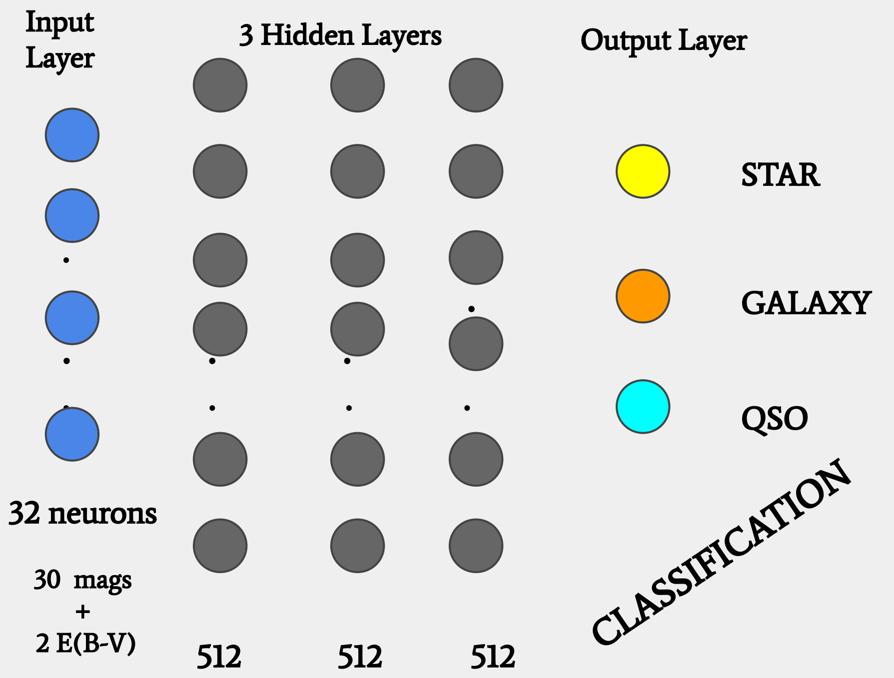
32 features

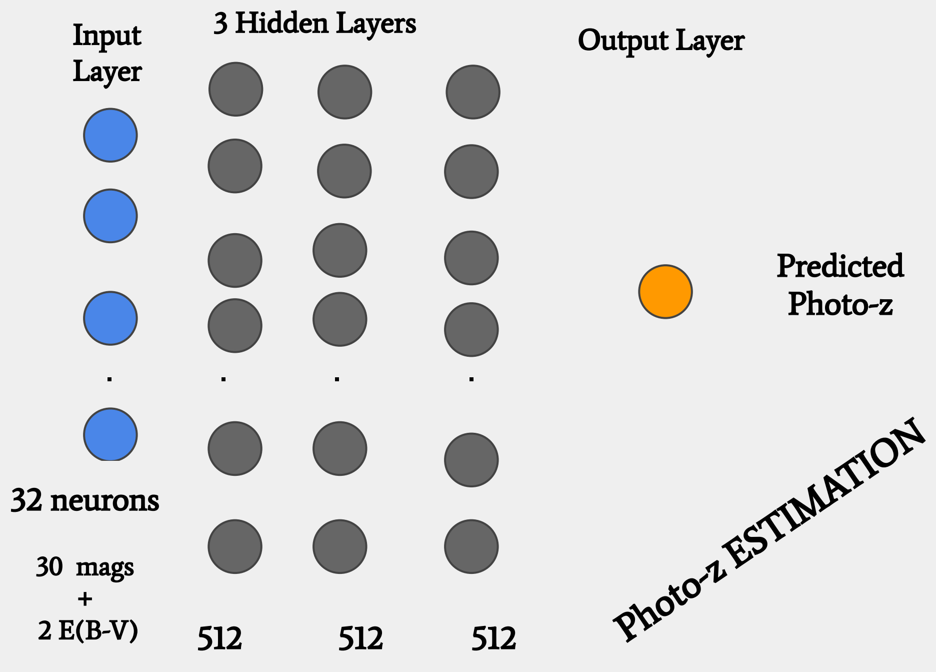
Methodology

Neural Network configuration

Why a NN ?

Flexible with non-linear models + Capable of recognizing useful patterns in multidimensional space





Methodology

Training Setup

- Training sample : 80 % of the ~3.8 million spectrosopic dataset
- Validation sample : 20 % of the ~3.8 million spectrosopic dataset

Classifier Model : Trained for 150 epochs

↓

Only objects that classify as galaxies used for the training of photo-z estimator
Mirroring the actual use case !

↓

Regression Model: Trained for 100 epochs

Methodology

Self Orginizing Maps (SOMs)

NN are not capable to extrapolate into regions of the input parameter space that are not covered from the training set

A way to quantify the training set coverage in the 30-d magnitude space is to use **SOMs**

SOMs: N.N model that identifies correlations in high dimensional data

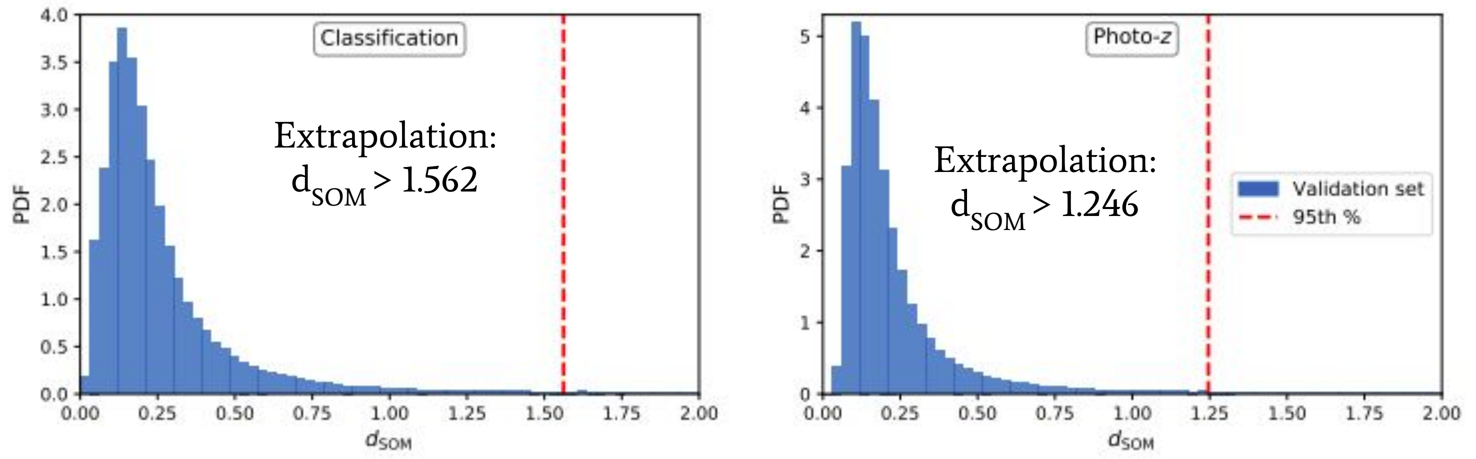


Figure 2. The distribution of d_{SOM}, the Euclidean distance (in normalized magnitude space) from the nearest cell centre in the SOM, for validation set objects. Vertical dashed lines represent the cut that defines whether an object is flagged as extrapolated. The left panel corresponds to the classification SOM and validation set, while the right panel shows the photo-z SOM and validation set.

Methodology

Catalogue processing

1st step: 2 E(B-V) values from dust maps based on l,b galactic coordinates

2nd step: 2 E(B-V) + 30 photometric fluxes ⟹ N.N. classification ⟹ p_{star},p_{gal},p_{QSO}

3rd step: **If** : p_{class} > 0.70 ⟹ Object is flagged with the corresponding class **else:** Object is flagged as *"Unsure"*

4th step: Based on SOM ⟹ d_{SOM} > 1.562 ⟹ Object is flagged as *"Extrapolated"*

5th step: Only objects flagged as *"Galaxies"* ⟹ N.N. photo-z estimation ⟹ $z_{phot,0}$

6th step: Monte Carlo sampling

100 multivariate Gaussian random samples with std= mag errors

100 realizations of the N.N. photo-z estimation

100 $z_{phot,0}$ values ⟹ Median = z_{phot}

7th step: Based on SOM ⟹ Galaxies with d_{SOM} > 1.244 ⟹ Object is flagged as *"Extrapolated"*

Validation Results

T1 : galaxy classified as galaxy
T0 : non-galaxy classified as non-galaxy

F1: non- galaxy classified as galaxy
F0: galaxy classified as non-galaxy

Metrics for the evaluation of the classification model

Purity = T1/(T1+F1), Completeness = T1/(T1+F0), Overall success = (T1+T0)/(T1+T0+F1+F0)

Table 2. Classification metrics for the galaxy, star and quasar classes, for different b decision boundary choices: P, the purity; C, the completeness; and S, the overall success rate. The fiducial decision boundary is $b = 0.7$. The metrics were evaluated on our validation data set. See the text for a detailed description of the metrics.

	Galaxy			Star			Quasar		
b	P_{gal}	C_{gal}	S_{gal}	P_{star}	C_{star}	S_{star}	P_{qso}	C_{qso}	S_{qso}
0.50	98.03%	98.86%	97.94%	94.61%	94.11%	97.75%	90.12%	85.87%	96.70%
0.60	98.30%	98.54%	97.91%	95.87%	92.62%	97.73%	92.36%	82.38%	96.57%
0.70	**98.56%**	**98.06%**	**97.77%**	**96.97%**	**90.68%**	**97.57%**	**94.17%**	**77.92%**	**96.23%**
0.80	98.82%	97.19%	97.38%	98.00%	87.88%	97.22%	95.99%	71.51%	95.59%
0.90	99.13%	95.03%	96.17%	98.92%	82.93%	96.41%	97.83%	60.45%	94.27%
0.95	99.34%	91.59%	94.04%	99.41%	77.64%	95.44%	98.62%	50.70%	92.99%
0.99	99.75%	64.20%	76.26%	99.79%	64.55%	92.89%	99.47%	29.09%	90.04%

Table 3. The same as Table 2, but the classification metrics were evaluated only on non-extrapolated sources within our validation data set.

	Galaxy			Star			Quasar		
b	P_{gal}	C_{gal}	S_{gal}	P_{star}	C_{star}	S_{star}	P_{qso}	C_{qso}	S_{qso}
0.50	98.36%	99.01%	98.25%	94.88%	95.01%	97.95%	90.85%	86.64%	97.01%
0.60	98.58%	98.73%	98.22%	96.04%	93.68%	97.94%	92.92%	83.49%	96.91%
0.70	**98.77%**	**98.36%**	**98.10%**	**97.04%**	**91.89%**	**97.79%**	**94.55%**	**79.44%**	**96.60%**
0.80	98.97%	97.73%	97.82%	98.04%	89.22%	97.46%	96.20%	73.75%	96.05%
0.90	99.22%	96.04%	96.88%	98.94%	84.36%	96.65%	97.89%	63.53%	94.88%
0.95	99.39%	93.18%	95.10%	99.43%	79.06%	95.67%	98.65%	53.86%	93.66%
0.99	99.77%	66.44%	77.67%	99.80%	65.83%	93.06%	99.49%	31.28%	90.68%

Validation Results

Photo-z

Table 4. Photo-z accuracy metrics computed on the base and Monte-Carlo sampled redshift estimates, for all validation set galaxies, and for non-extrapolated validation set galaxies. See the text for more details.

Data set	Estimate	$\langle \Delta z_{norm} \rangle$	$\sigma(\Delta z_{norm})$	$MAD(\Delta z_{norm})$	O
All validation	$z_{phot,0}$	0.0003	0.0342	0.0169	2.88%
All validation	z_{phot}	0.0010	0.0344	0.0170	2.99%
Non-extrapolated	$z_{phot,0}$	0.0005	0.0322	0.0161	1.89%
Non-extrapolated	z_{phot}	0.0013	0.0323	0.0163	2.00%

Standard literature metrics:

- $\Delta z_{norm} = (z_{phot} - z_{spec})/(1+z_{spec}) \equiv$ Redshift error
- $O \equiv |\Delta z_{norm}| > 0.15$: Outliers fraction
- $<\Delta z_{norm}> \equiv$ Average bias (only on non-outliers)
- $MAD(\Delta z_{norm}) \equiv$ Median Absolute Deviation

Added noise from MCS

⇩

Slightly lower performance

Suggested to use : $z_{phot,0}$

Validation Results

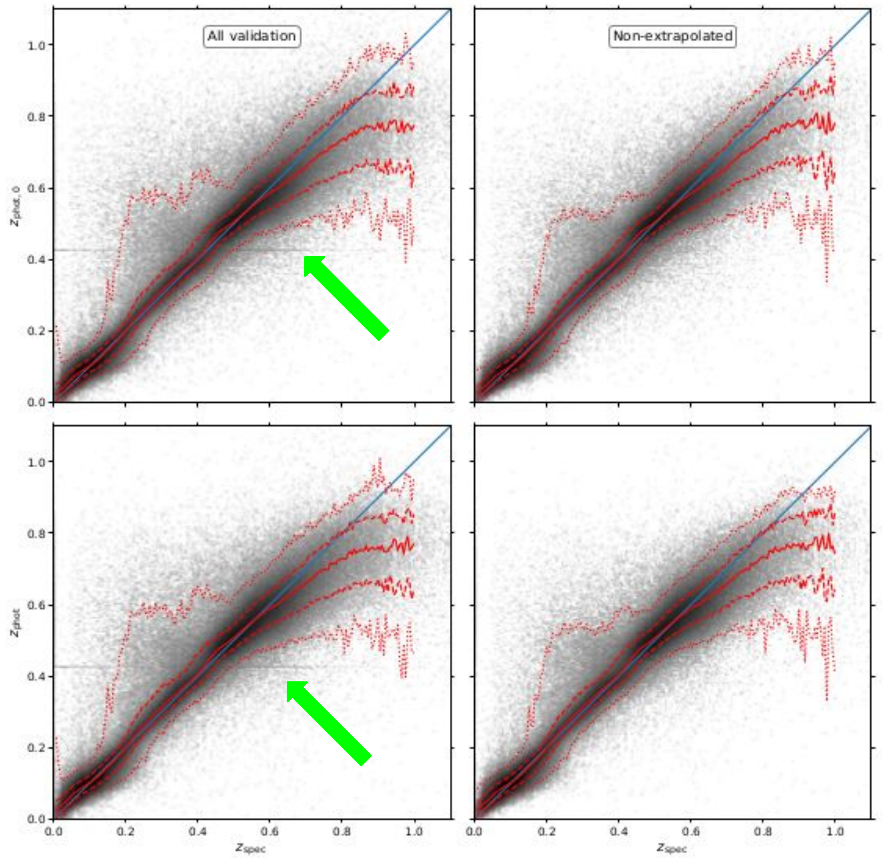


Figure 3. Photometric redshift estimation results, for the base estimate $z_{phot,0}$ and the Monte-Carlo sampled z_{phot}. The left column shows all validation set galaxies, while the right column shows only non-extrapolated validation set galaxies. In grayscale, we plot the logarithmic density of galaxies, so that even individual objects are visible. Solid, dashed and dotted lines show the sample median, 68% confidence interval, and 95% confidence interval, respectively. The main diagonal corresponds to the perfect estimation.

Extrapolated sources introduce unwanted features $z_{phot} \approxeq 0.43$ (object with missing photometry)

Suggestion for users :

1. Limit the analysis to Non-extrapolated sources !
2. Useful range: $z \in [0,0.6]$!

Take home message

- Creation of a new catalogue (PS1 3π DR1) including source classification and photo-z estimation by using machine learning methods
- Size: **2.902.054.648** objects

- Quantification of the parameter space of the training sample by using SOM. Definition of extrapolation boundaries

Non-extrapolated objects

-

	Galaxy			Star			Quasar		
b	P_{gal}	C_{gal}	S_{gal}	P_{star}	C_{star}	S_{star}	P_{qso}	C_{qso}	S_{qso}
0.70	98.77%	98.36%	98.10%	97.04%	91.89%	97.79%	94.55%	79.44%	96.60%

- Best photo-z estimation $z_{phot,0}$, since MCS add extra noise
- Future plans:
  - Optimization of the N.N hyperparameters
  - Inclusion of infrared observations

- Catalogue will be publicly available via **Mikulski Archive for Space Telescopes (MAST)**

BACK UP

$$B = \frac{L\,(\text{same source})}{L\,(\text{separate sources})} = \frac{2}{\sigma_1^2 + \sigma_2^2} exp\left\{ -\frac{\psi^2}{2\left(\sigma_1^2 + \sigma_2^2\right)} \right\}$$

Here σ_1 and σ_2 are the astrometric errors of two given galaxies, and ψ is the angular separation between them. We accepted matches with $B > 10,000$, thus ensuring that we only used rather certain matches.