ML for photo-z’s

Amaro Valeria,	University Federico II of Naples
Brescia Massimo,	INAF OACN
Cavuoti Stefano,	INAF OACN
Longo Giuseppe,	University Federico II of Naples
Vellucci Civita,	University Federico II of Naples

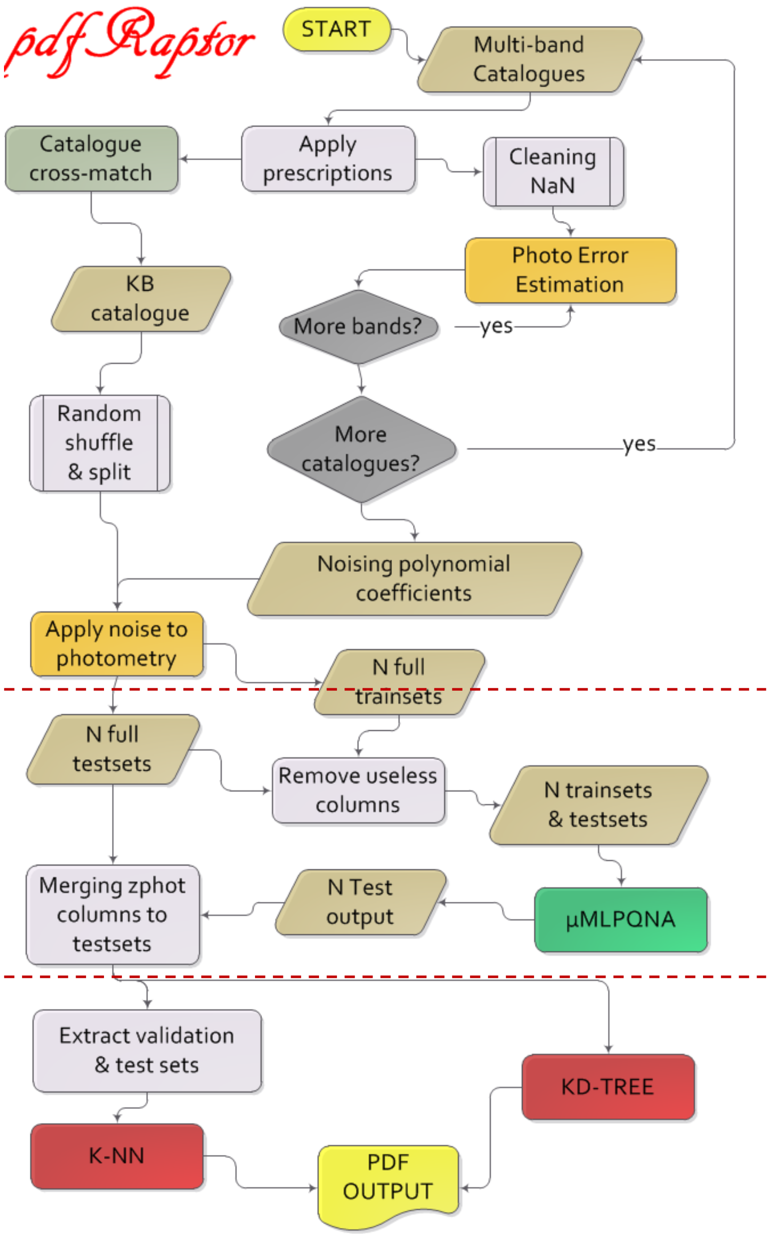
Error budget for regression with ML methods

1. Internal errors (initialization of weights, model topology, etc)
2. Propagation of photometric errors
3. Cross-talking between adiacent bins
4. Systematics
5. Etc.

Photo-z with MLPQNA

- □ **PHAT1 Contest** (*Cavuoti et al. 2012, A&A, 546, A13*)
- □ **GALEX+SDSS+UKIDSS+WISE QSOs** (*Brescia et al. 2013, ApJ, 772, 2, 140*)
- □ **CLASH-VLT** (*Biviano et al. 2013, A&A, 558, A1*)
- □ **EUCLID PHZ** (*Coupon et al. 2014, Challenge #1 internal report*)
- □ **SDSS DR9** (*Brescia et al. 2014, A&A, 568, A126*)
- □ **KiDS DR2** (*Cavuoti et al. 2015, MNRAS, accepted, in press*)
- □ **VST VOICE** (*Covone et al. 2015, in prep.*)
- □ **XMM** (*Vaccari et al. 2015, in prep.*)

pdfRaptor pipeline architecture

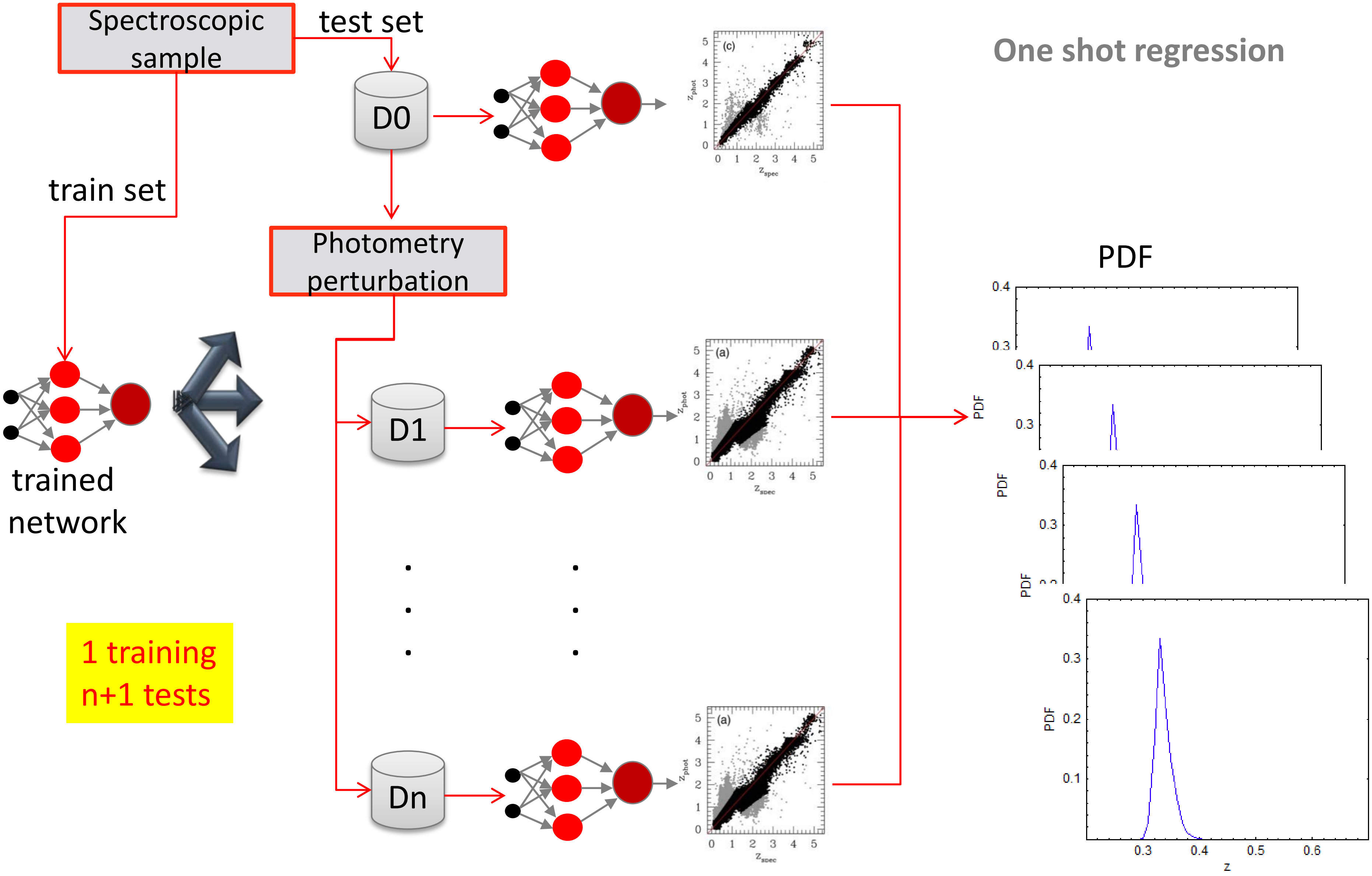


Data Pre-processing: photometric evaluation and error estimation of the multi-band catalogue used as KB of the photo-z experiment.

Photo-z calculation: training/test phase to be performed through the selected interpolative method (in this case µMLPQNA, which stands for multi-thread MLPQNA).

PDF calculation: methods designed and implemented to furnish a PDF evaluation for the photo-z produced.

base algorithm processing flow



Spectroscopic sample

test set

train set

D0

Photometry perturbation

D1

Dn

trained network

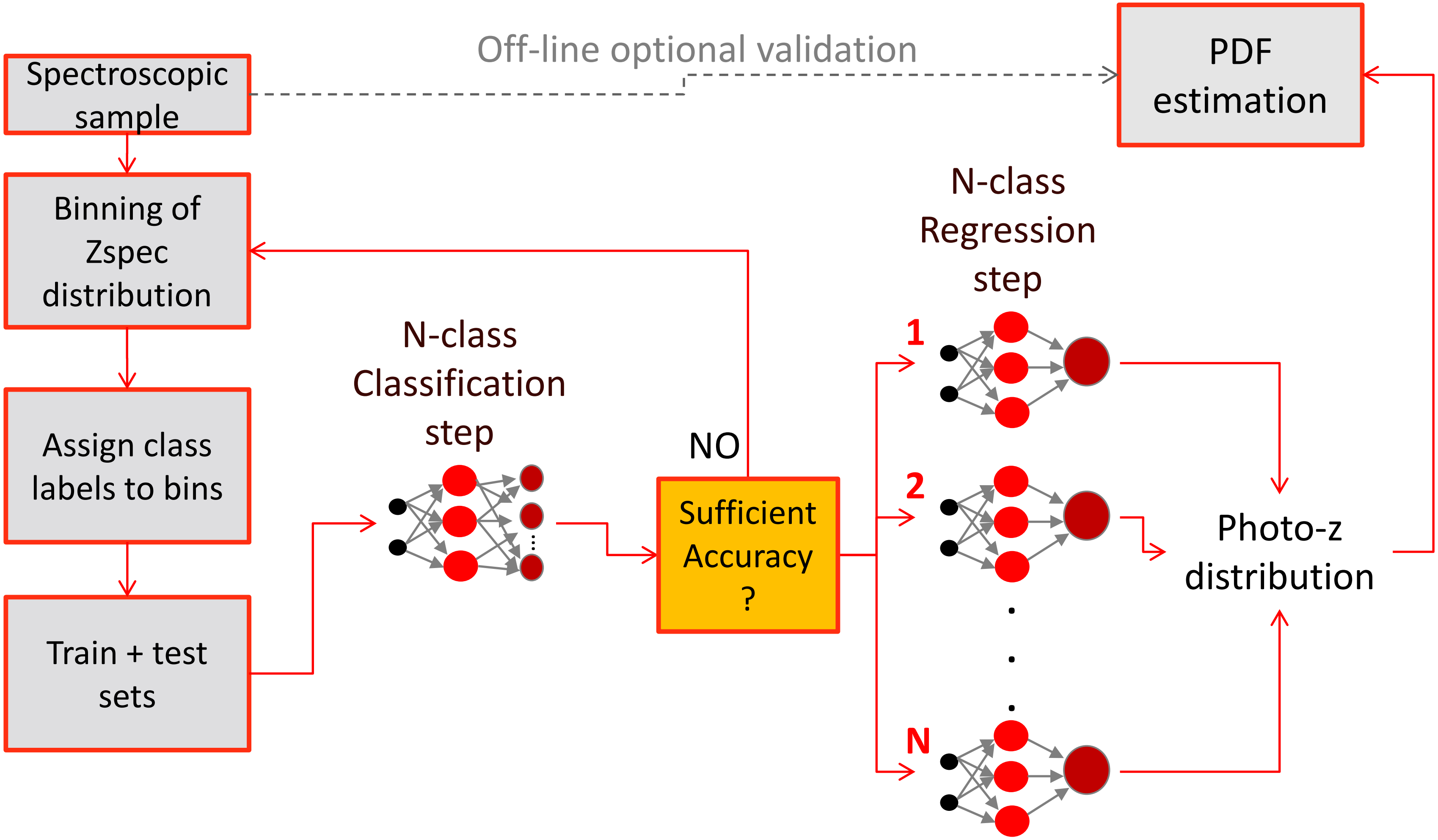
One shot regression

PDF

1 training
n+1 tests

PDF base algorithm processing flow

Hierarchical approach based on classification + multi-regression



Spectroscopic sample

Binning of Zspec distribution

Assign class labels to bins

Train + test sets

N-class Classification step

NO

Sufficient Accuracy ?

Off-line optional validation

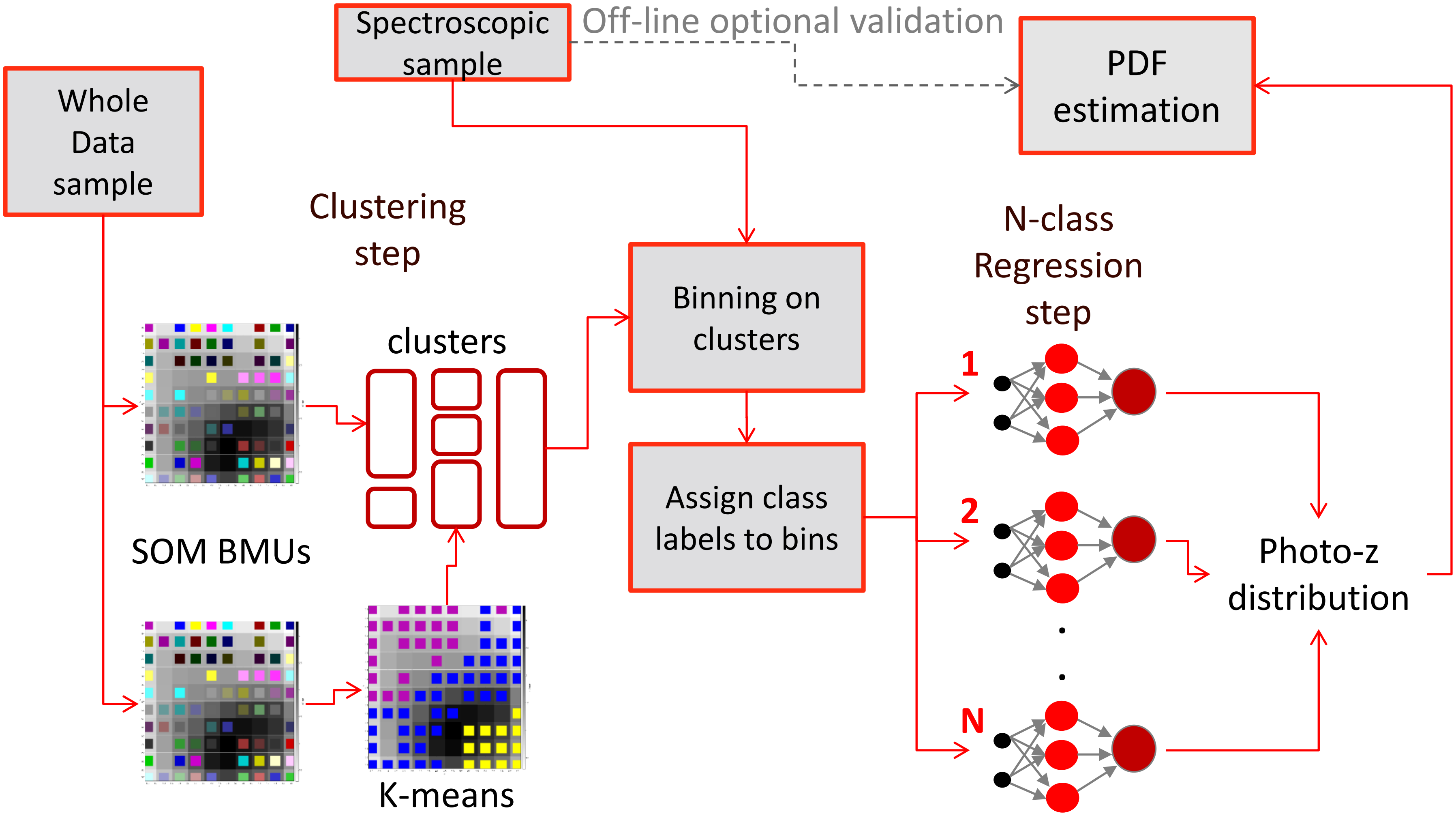
N-class Regression step

Photo-z distribution

PDF estimation

PDF base algorithm processing flow

Hierarchical approach based on clustering + multi-regression



Whole Data sample

Spectroscopic sample

Off-line optional validation

PDF estimation

Clustering step

clusters

SOM BMUs

K-means

Binning on clusters

Assign class labels to bins

N-class Regression step

Application to Data Challenge 2.

Pre-processing

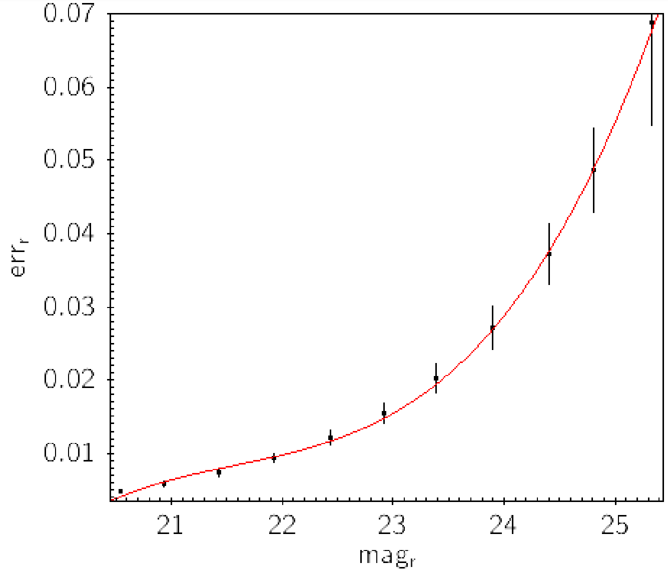
- cleaning from NaN entries ($22-29)
- application of the spectroscopic reliability flag(”reliable_S15”==1) according to the scheme by Salvato et al. 2015;
- the elimination of stars and the conservation of the AGN ($59) : this last column is defined by means of the columns $57 (star flag, if star, flag=1) and $58(AGN flag, if AGN, flag=1)(($57==1&&$58==0)?1:0)
- application of a prescription through the Sextractor Flags (“FLAG_DETECT”<4).
- Cuts in magnitudes (5 sigma): g: 24.95; r: 24.60; i: 23.72; ACS: 24.82; z: 23.21; Y: 24.57; J: 24.35 ; H: 23.89
- cut on mag error higher than 1.

- ***Errors on mags are used to derive the photometric errors rules. 100 perturbations for each data point.***
- ***Colours derived after perturbation of the mags***

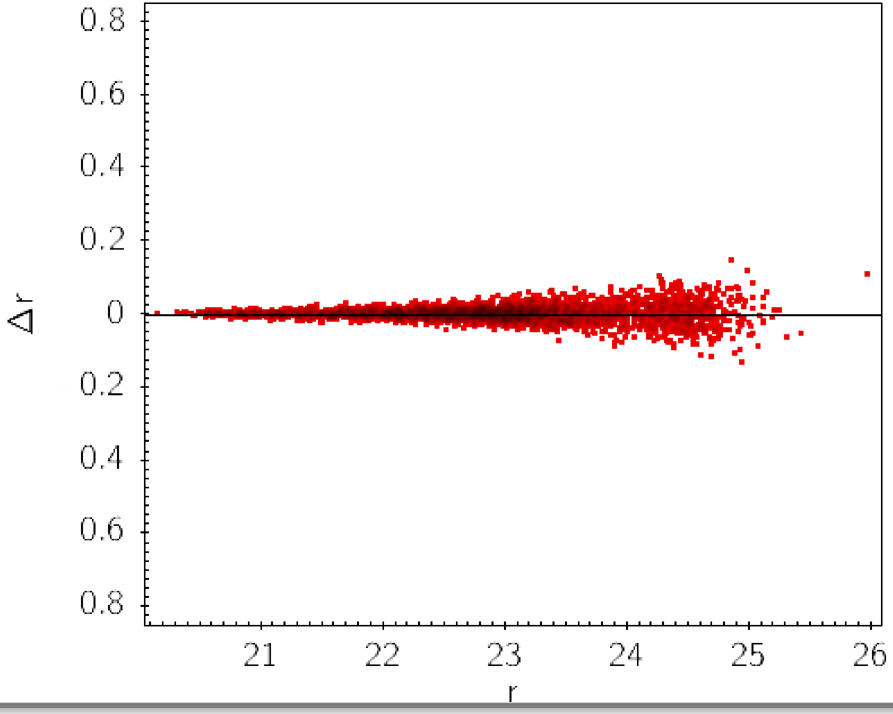
Photometry perturbation

Given a dataset A, a normal distribution on A, and

$N_{samples}$	number of objects in a given dataset A
$N_{perturb}$	number of perturbations to be done
N_{mags},	number of affected magnitudes
p_b	polynomial used to perturb mag of band b
$alpha_b$	perturbation constant for the band b
$mag_b(o_i)$	mag value of the band b for the object o_i



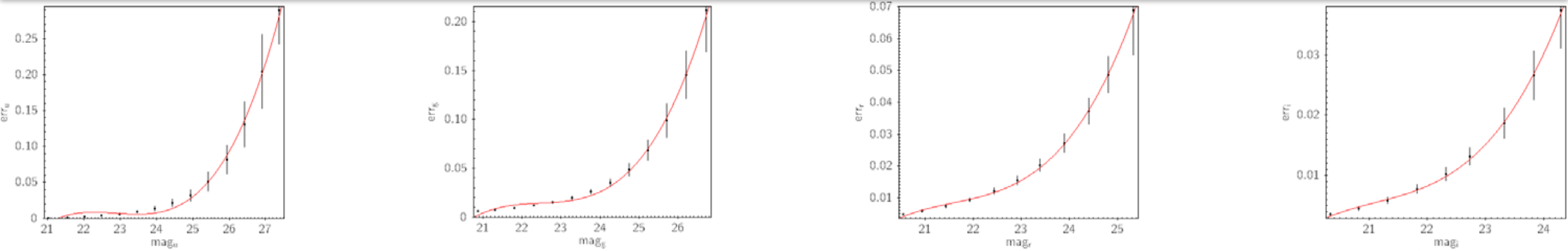
$$m_{ijperturbed}(o_i) = m_{ij} + alpha_b * p_b \circ (mag(o_i)) * N_A(0;1)$$



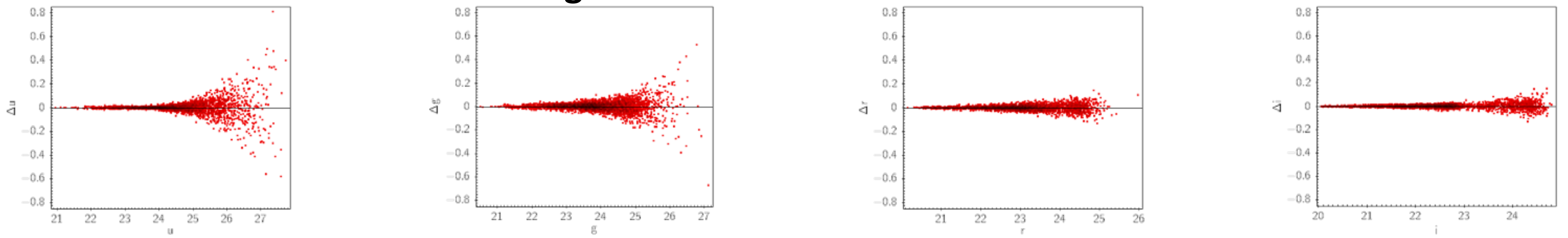
where the symbol "∘" stays for the scalar product,
$N_A(0;1)$is a normal distribution with the dimension of the dataset A to be perturbed, i.e. a distribution of a number $N_{samples}$ of values in the interval (-1,1).

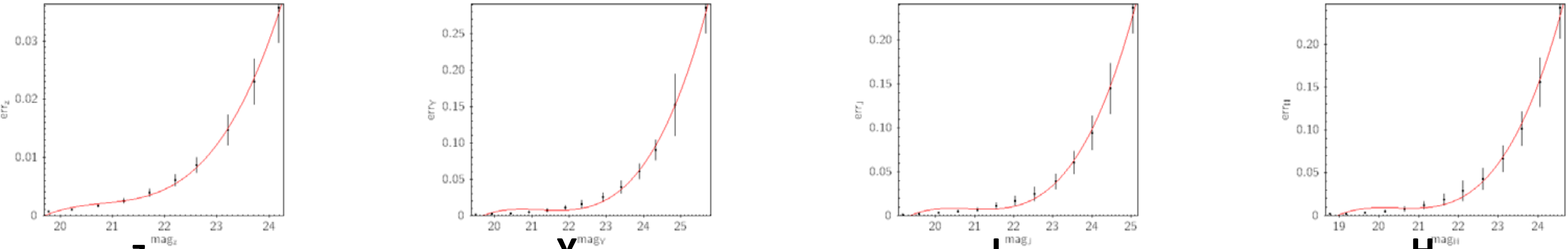
The variation of the percentage of noise is ensured by the randomly generated normal distribution at each step.

Photometry perturbation

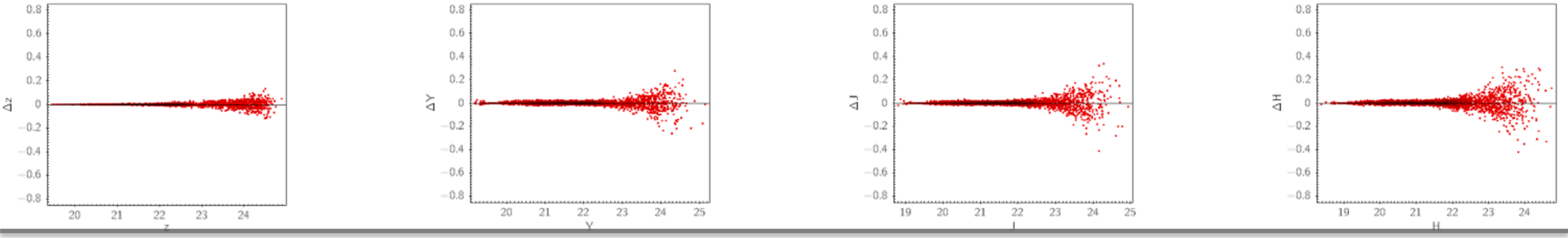


u g r i

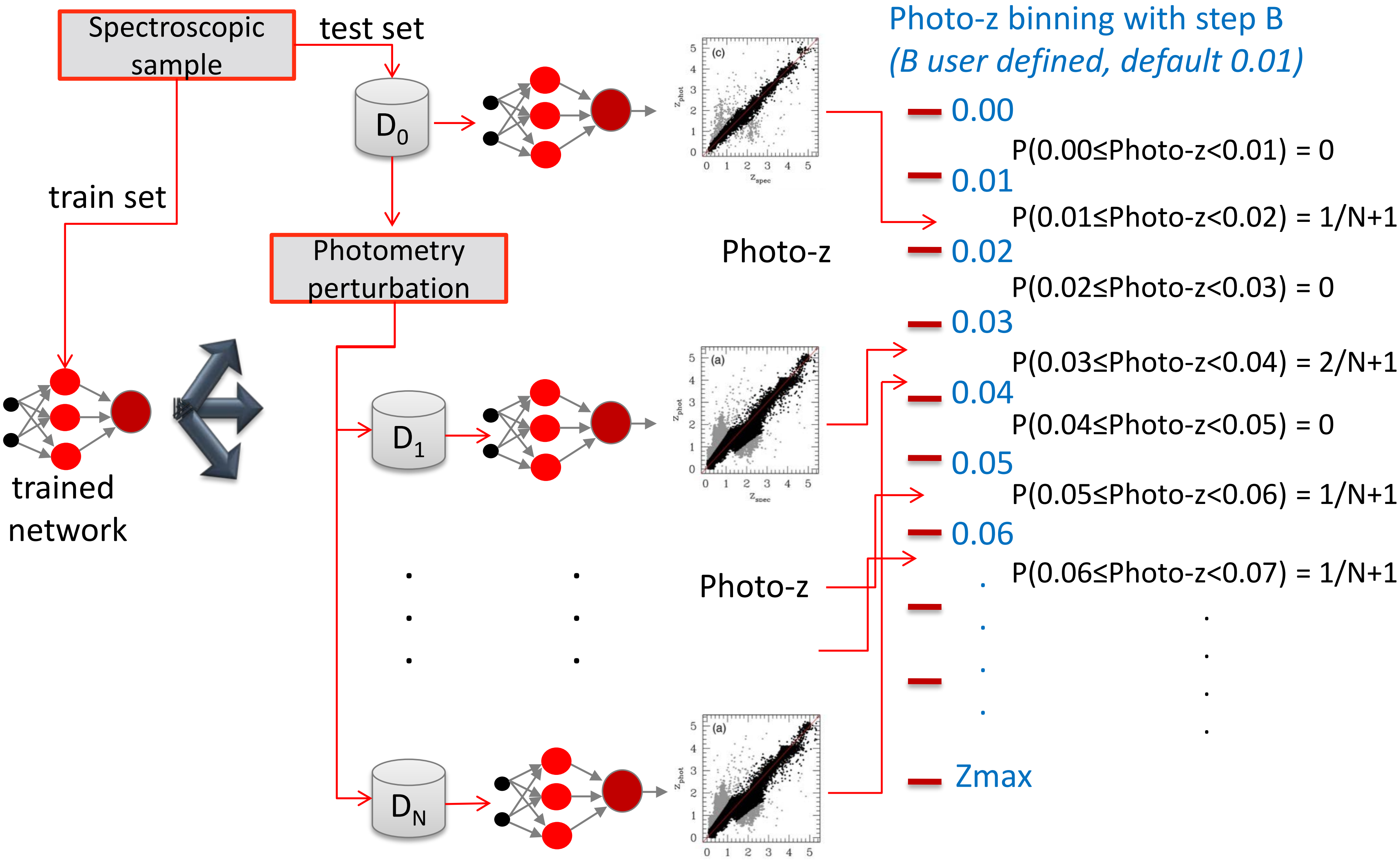




z Y J H



PDF base algorithm processing flow



Spectroscopic sample

test set

train set

D_0

Photometry perturbation

trained network

D_1

D_N

Photo-z

Photo-z

Photo-z binning with step B

(B user defined, default 0.01)

0.00

P(0.00≤Photo-z<0.01) = 0

0.01

P(0.01≤Photo-z<0.02) = 1/N+1

0.02

P(0.02≤Photo-z<0.03) = 0

0.03

P(0.03≤Photo-z<0.04) = 2/N+1

0.04

P(0.04≤Photo-z<0.05) = 0

0.05

P(0.05≤Photo-z<0.06) = 1/N+1

0.06

P(0.06≤Photo-z<0.07) = 1/N+1

Zmax

$$PDF(Photo\text{-}z) = \{P(Z_i \leq Photo\text{-}z < Z_{i+B}) = C_{B,i}/N+1\}_{[Zmin,\ Zmax]}$$

Features

Extensive set of experiments led to a PS with 17 features:

9 colours : g-r,r-i,i-z,z-Y,Y-J,J-H, VIS-Y, VIS-J, VIS-H
8 mags: g, r, i, VIS, z, Y, J, H

Training & Test

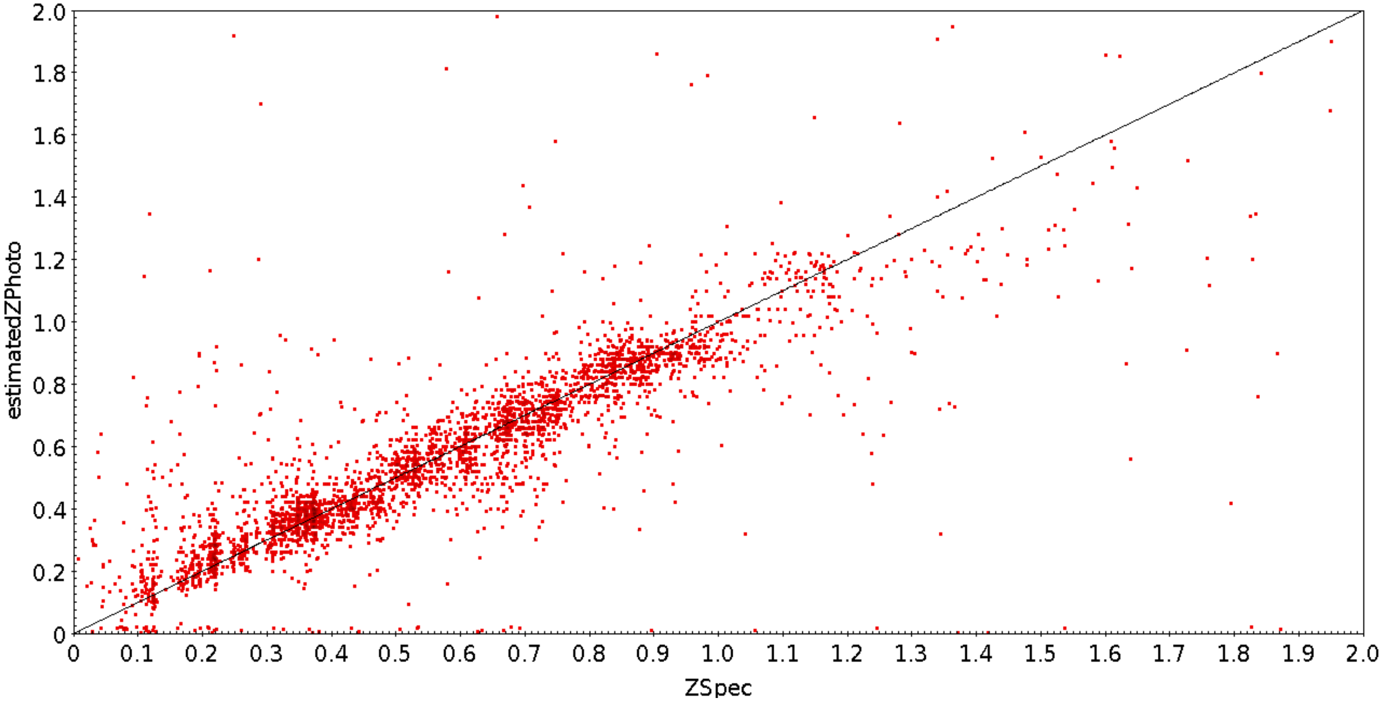
The training set has been randomly shuffled and split in a train set (with the 70% of the training set samples) and a test set (with a 30% of the training set objects)

101 runs of MLPQNA

Network topology and training parameters	
inputNeurons	17 (all the mags available+9 colours)
hiddenLayers	2
hiddenLayer1Neurons	35
hiddenLayer2Neurons	16
restarts	80
epochs	10000
threshold	0.001
decay	0,1

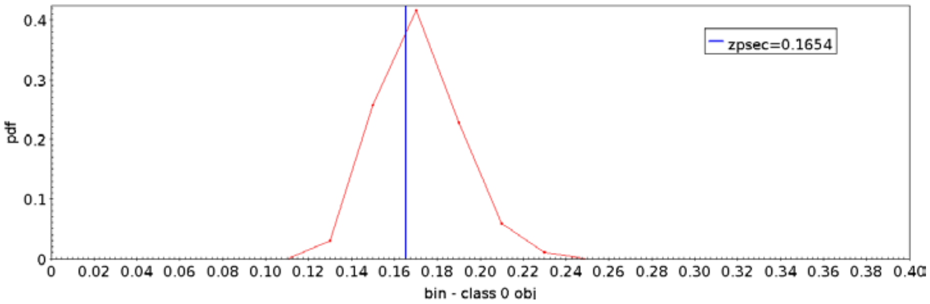
Overall performances for «best» experiment

parameter	network calibration catalogue (8 mags+9 colours)
|bias|norm	0.012
σ norm	0.145
Nmad_norm	0.044
σ68	0.049
σ95	0.220
%outliers>0.15	8,17
Train/test/total # objs	8218/3512/11730 (split 70-30 %)

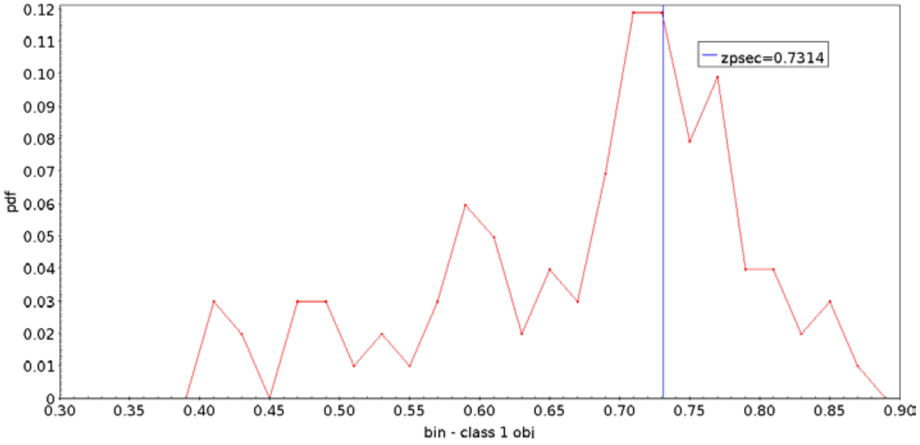


PDFs can be split into four groups:

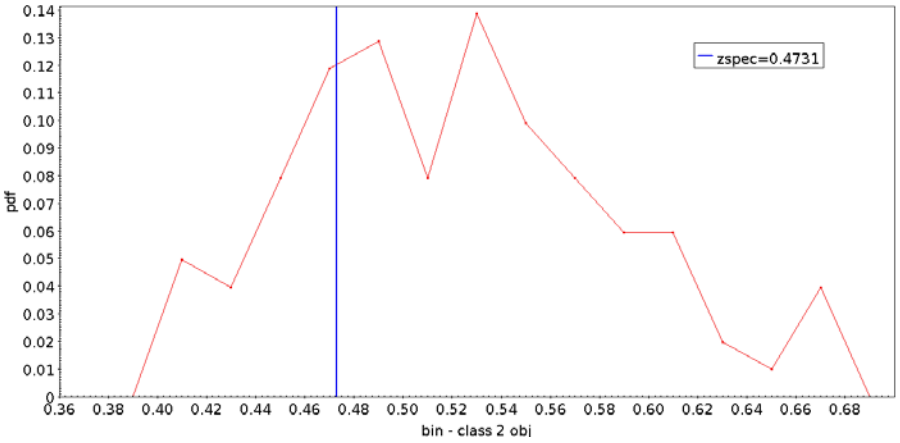
- pdfwidth: the amplitude of the whole pdf;
- pdfNbins : the total number of bins that compose the pdf;
- pdfPeakHeight: the amplitude of the peak of the pdf;
- pdfNearPeakwidth: the amplitude of the pdf near the peak, i.e the distance between the latest pdf bin where the pdf is ≠ 0 and higher than the peak bin, and the latest pdf bin where the pdf is ≠ 0 lower than the peak bin.



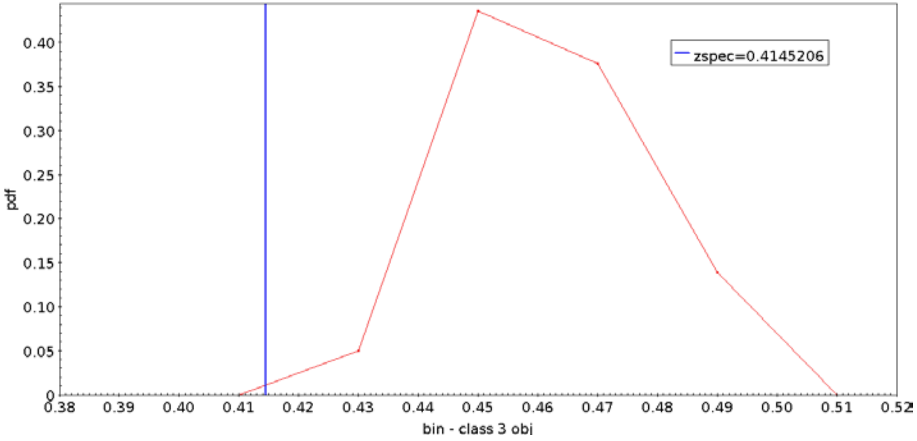
0 = Zspec falls within PDF peak



1 = Zspec falls within 1 bin from PDF peak



2 = Zspec falls within the PDF

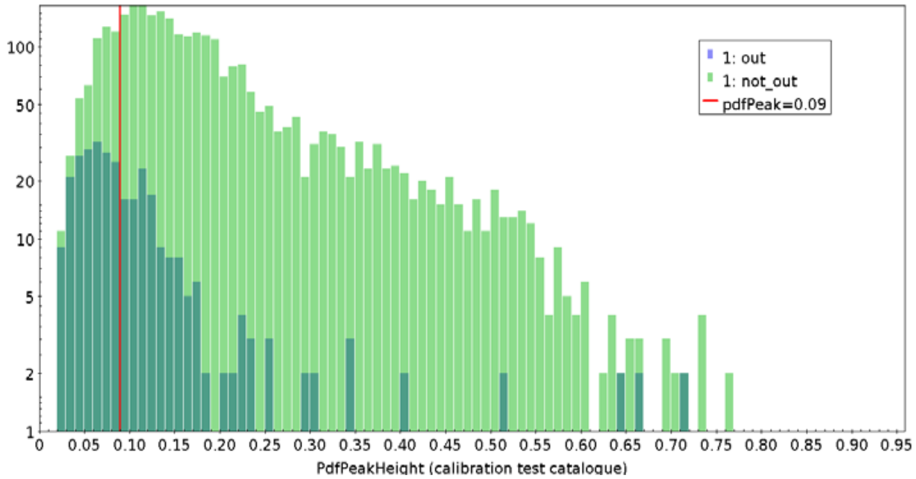


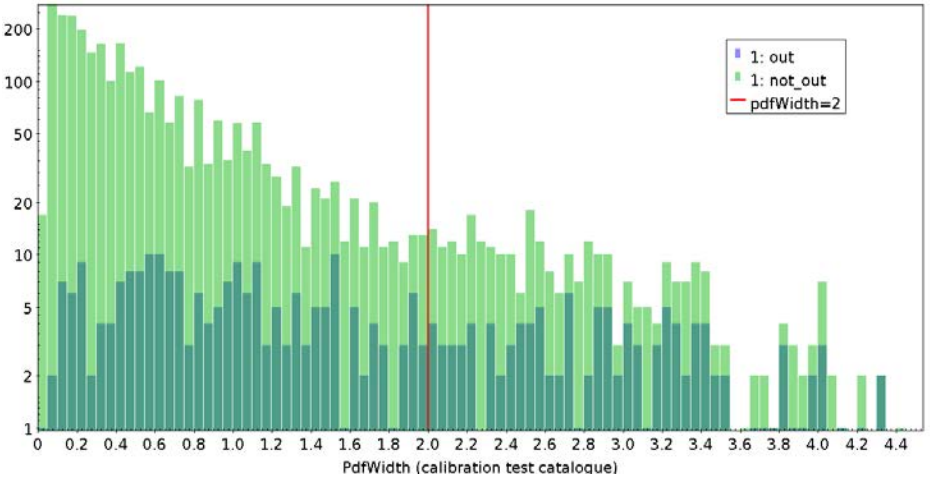
3 = Zspec falls outside the PDF

class	calibration test set (#3512 objs)
0 (zsspec within the pdf peak)	407 (12%)
1 (zspec within 1 bin from the peak)	761 (22%)
2 (zspec within the pdf)	1938 (54 %)
3 (zspec outside the pdf)	406 (12%)

Outliers are not distributed at random in the OPS

- |zspec-zphotestimated|/(1+zspec)>0.15 for outliers
- |zspec-zphotestimated|/(1+zspec)<0.15 for no-outliers





This cut reduces the test set samples of ~14% from 3512 to 3035. The recalculation of the statistics in correspondence of this cut,

- the σ68 is reduced to **0.040** (see table 2 for a comparison);
- the fraction of outliers is reduced to **4.41%** (see table 2 for a comparison).

Summary:

pdfbase algorithm has been applied to the validation catalogue
No NaN entries(#140,944/190462 of the original catalogue) and all the prescriptions applied to the calibration catalogue are “translated” in appropriate flags.

In particular, are flagged “0” in the “USE” flag column (see below), all the samples of the validation catalogue with:

- mag values deepest of the depth mag cut values within 5 sigma, applied to the calibration catalogue;
- all the samples with the FLAG_DETECT >4
- all the samples with mag errors >1
- all the samples with zphotestimated>2
- all the samples with Pdf Peak Height <0.09
- all samples with Pdf Width >2

Conclusions and future steps

Our «best experiment» falls within the requirement box but it can be largely improved

Some remarks on Data Challenge

1. We need clean training data for classification purposes (confusing labels STAR/AGN)
2. Once again , validation set was not blind

1. **Training sample is still small.**
   Corollary: blind tests should be truly blind. By removing blindness…..
   Results improve
2. **PDFs at the moment do not take into account the minimal contribution from inizialization of weights.**
3. **There is a correlation between PDF and position in the parameter space** (to be invstigated with MCS)
4. Need to obtain chi^2 from SED fitting to optimize evluation of errors (dependent on SED morphological type)
5. Need to investigate better combination of features (Data Driven Approach)
6. Optimize regression to subclasses