PHOTOMETRIC REDSHIFTS IN KIDS WITH MLPQNA: COMPLETENESS VS ACCURACY.

Massimo Brescia
INAF – Capodimonte Astronomical Observatory – Napoli
Stefano Cavuoti
INAF – Capodimonte Astronomical Observatory – Napoli
Giuseppe Longo
University of Naples Federico II – Napoli

Public Service Announcement: What We Do

DAME Program is a joint effort between University Federico II, Caltech and INAF-OACN, aimed at implementing (as web 2.0 apps and services) a scientific gateway for data exploration on top of a virtualized distributed computing environment.

Multi-purpose data mining with machine learning Web App REsource

Extensions
- DAME-KNIME
- ML Model plugin

Specialized services:
- text mining (VOGCLUSTERS)
- Transient classification (STraDiWA)
- EUCLID Mission Data Quality

Other Services:
- CLASH-VLT Data Archive
- PhotoRaptor
- GPU-based models

http://dame.dsf.unina.it/
Science and management
Documents
Science cases
Newsletters

http://www.youtube.com/user/DAMEmedia
DAMEWARE Web Application media channel

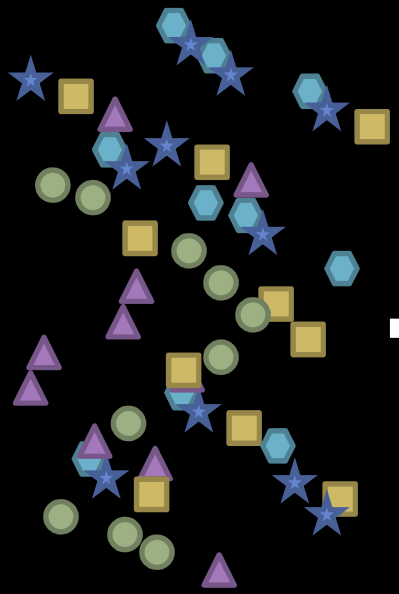
Machine Learning: Supervised

A Supervised Method tries to reproduce a bias, extending a preexisting knowledge on new patterns…

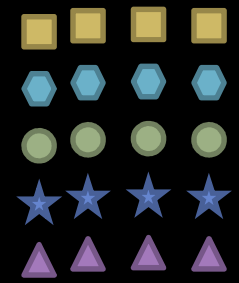
- Good for interpolation of data, bad for extrapolations;
- They need extensive bases of knowledge (KB, i.e. uniformly sampling the parameter space) which are difficult to obtain;
- Errors are easy to evaluate;
- Relatively easy to use;
- They reproduce all biases and preconceived ideas present in the KB.

Supervised Methods are subdivided into: Classification and Regression algorithms

Machine Learning: Supervised

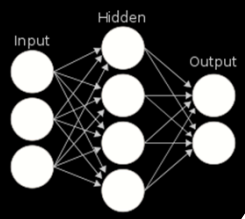


Knowledge Base



Train Set

Train

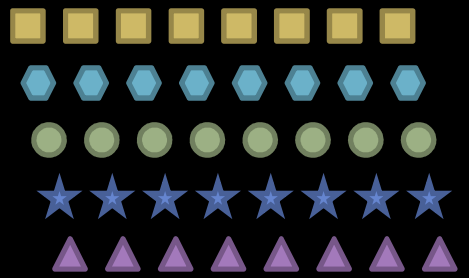


Trained Network

Blind Test Set

Test

The World



New Knowledge

Photometric Redshifts: the Data Mining Approach

Photometric redshifts are treated as a regression problem (i.e. function approximation) , hence a DM problem:

$\mathbf{X} \equiv \{x_1, x_2, x_3, \ldots x_N\}$ **input vectors**

$\mathbf{Y} \equiv \{x_1, x_2, x_3, \ldots x_M\}$ **target vectors** $M << N$

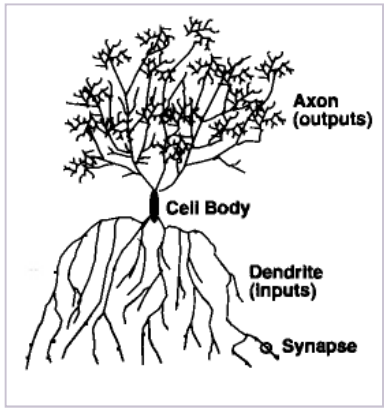
find $\hat{f}$: $\hat{\mathbf{Y}} = \hat{f}(\mathbf{X})$ **is a good approximation of** $\mathbf{Y}$

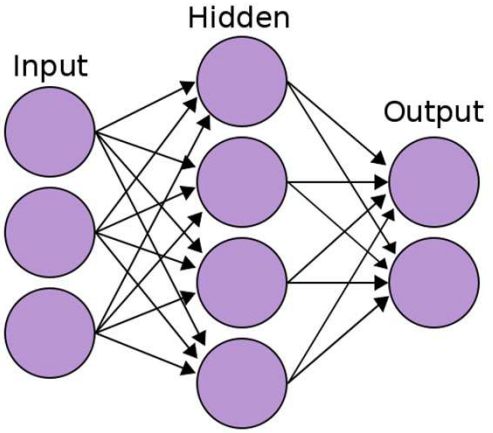
KB = Knowledge Base

KB(from VO) (set of templates) → Mapping function → Knowledge (phot-z's)

Multi Layer Perceptron

A Multi Layer Perceptron is a mathematical operator that mimics the brain behavior:





Neurons are connected by «activation functions» we have different kind of MLP changing the way with they found the best solution

Training rules:

- **Quasi Newton**
- **Back Propagation**
- **Genetic Algorithm**
- **Levenberg Marquardt**

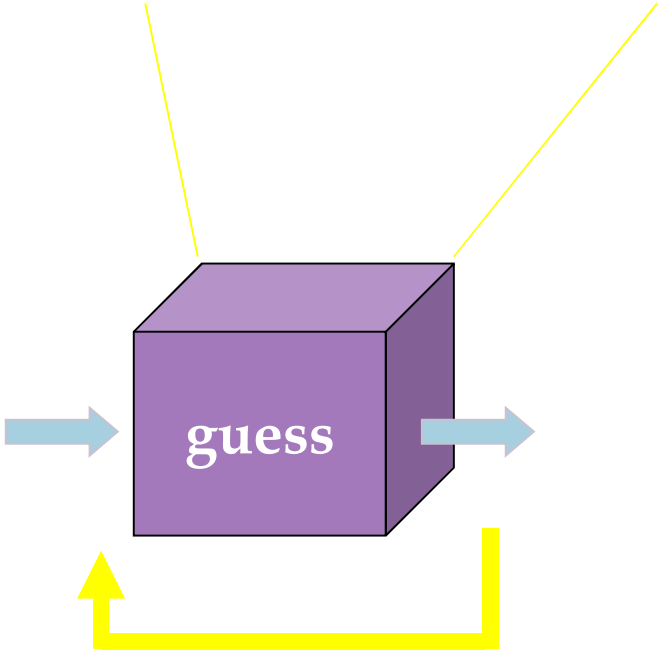


Photo-z: what we do

- Brescia, M.; Cavuoti, S.; D'Abrusco, R.; Longo, G.; Mercurio, A.; 2013, Photometric redshifts for Quasars in multi band Surveys, ApJ 772, 2, 140
- Cavuoti, S.; Brescia, M.; Longo, G.; Mercurio, A.; 2012, Photometric Redshifts with Quasi Newton Algorithm (MLPQNA). Results in the PHAT1 Contest, A&A, Vol. 546, A13, pp. 1-8
- D'Abrusco, R.; Staiano, A.; Longo, G.; Brescia, M.; De Filippis, E.; Paolillo, M.; Tagliaferri, R.; 2007, Mining the SDSS data. I. Photometric redshifts for the nearby universe, ApJ., 663, 752-764
- Cavuoti, S.; Brescia, M.; D'Abrusco, R.; Longo, G.; Paolillo, M.; 2014, Photometric classification of emission line galaxies with Machine Learning methods, Monthly Notices of the Royal Astronomical Society, Volume 437, Issue 1, p.968-975
- Brescia, M.; Cavuoti, S.; Paolillo, M.; Longo, G.; Puzia, T.; 2012, The detection of Globular Clusters in galaxies as a data mining problem, Monthly Notices of the Royal Astronomical Society, Volume 421, Issue 2, pp. 1155-1165, available at arXiv:1110.2144v1
- Brescia M., Cavuoti, S., Djorgovski, G.S., ,Donalek, C., Longo, G.,,Paolillo, M., 2011, Extracting knowledge from massive astronomical data sets, arXiv:1109.2840v1, to appear in Astrostatistics and data mining in large astronomical databases, L.M. Barrosaro et al. eds, Springer Series on Astrostatistics, 15 pages
- …

Currently:

- Euclid OU-Photo z (Galaxies and AGNs);
- VOICE/SUDARE
- SDSS DR10 (Galaxies and Quasars)
- CLASH-VLT
- KIDS… (of course)

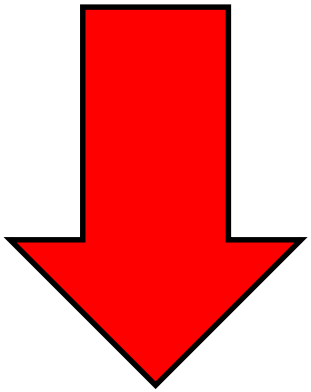
KIDS experiments

About 50 experiments were performed in order to find the best configurations, in particular in order to identify the most effective features; the results can be further improved.

Spectroscopic Base of Knowledge:

- RA/DEC;
- Optical aperture magnitudes (UGRI) at 2 arcsec (magap4) and 4 arcsec (magap6);
- NIR aperture magnitudes (HJKZY) at 2 arcsec (magap3), 2.8 arcsec (magap4) and 5.7 arcsec (magap6);
- GAMA Z Heliocentric spectroscopic redshift;
- SDSS Heliocentric spectroscopic redshift;
- NQ normalized redshift quality flag, related to GAMA (High Quality for NQ > 2);

Best Results (for the moment)



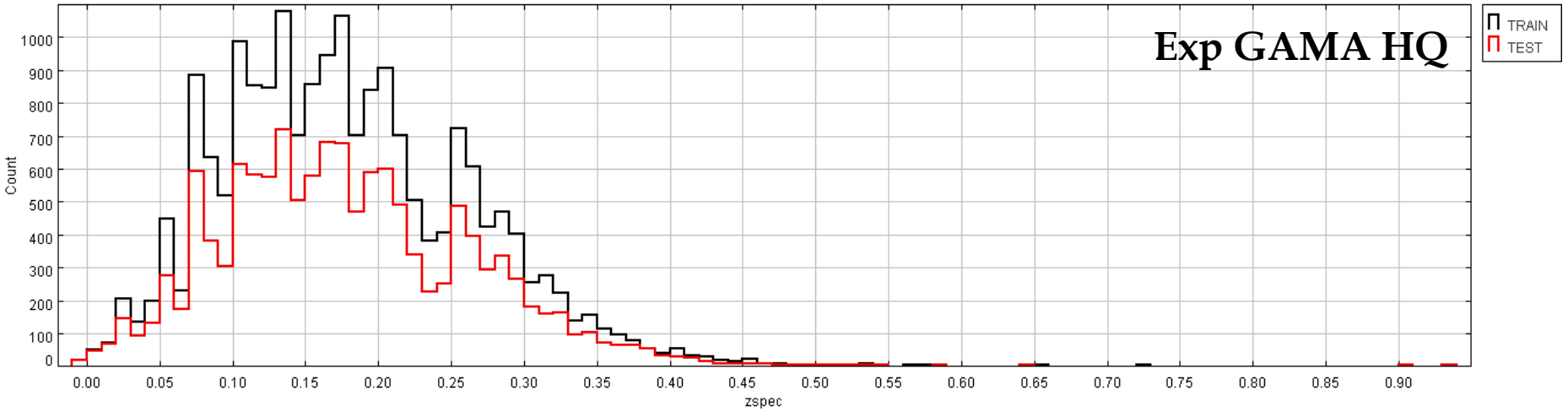
Experiments

Experiment LABEL	OPT-G		OPT-GS	OPTNIR-GS			
Experiment ID	1	2	3	4	5	6	7
input features	8	8	8	13	13	13	23
OPT magap_4	UGRI	UGRI	UGRI	UGRI	UGRI	UGRI	UGRI
OPT magap_6	UGRI	UGRI	UGRI	UGRI	UGRI	UGRI	UGRI
NIR magap_3	-	-	-	HJKZY	-	-	HJKZY
NIR magap_4	-	-	-	-	HJKZY	-	HJKZY
NIR magap_6	-	-	-	-	-	HJKZY	HJKZY

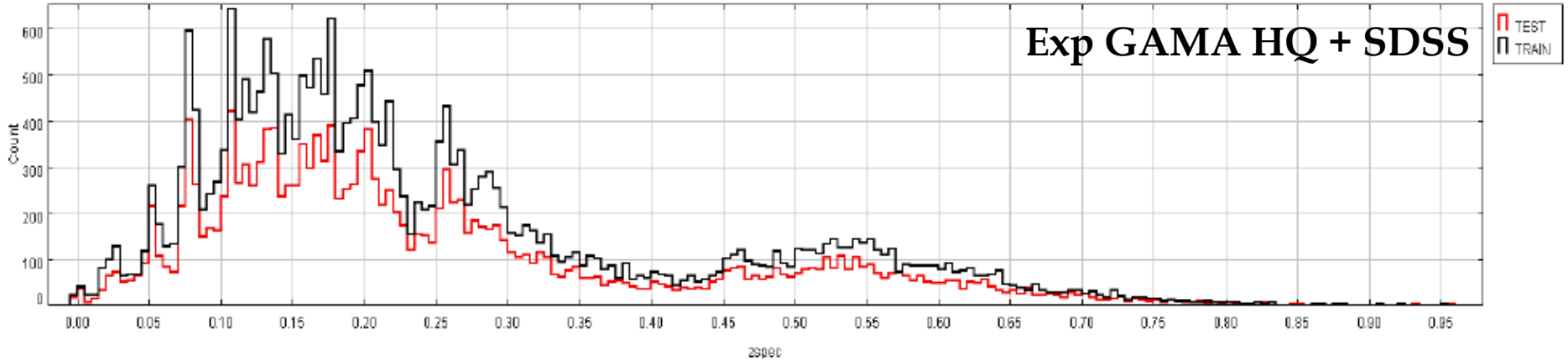
Experiments Results

BLIND TEST SET STATISTICS on Δz = (zspec-zphot)/(1+zspec)						
Experiment ID	|bias|	σ	NMAD	Outliers % |Δz| > 0.15	Outliers % |Δz| > σ	Outliers % |Δz| > 2σ
1	0.00094	0.0278	0.0201	0.31	20.0	3.22
2-HQ	0.00068	0.0267	0.0196	0.28	20.9	3.67
3	**0.00086**	**0.0305**	**0.0206**	**0.42**	**18.23**	**3.06**
3-HQ	**0.00099**	**0.0267**	**0.0198**	**0.30**	**21.2**	**3.34**
4-HQ	0.00072	0.0273	0.0204	0.27	21.9	3.33
5	**0.00071**	**0.0266**	**0.0155**	**0.41**	**14.88**	**3.01**
5-HQ	**0.00017**	**0.0220**	**0.0139**	**0.25**	**15.23**	**2.62**
6-HQ	0.00001	0.0231	0.0149	0.27	16.46	2.96
7-HQ	0.00007	0.0218	0.0139	0.27	16.26	2.96

Experiments Optical Zspec Distribution

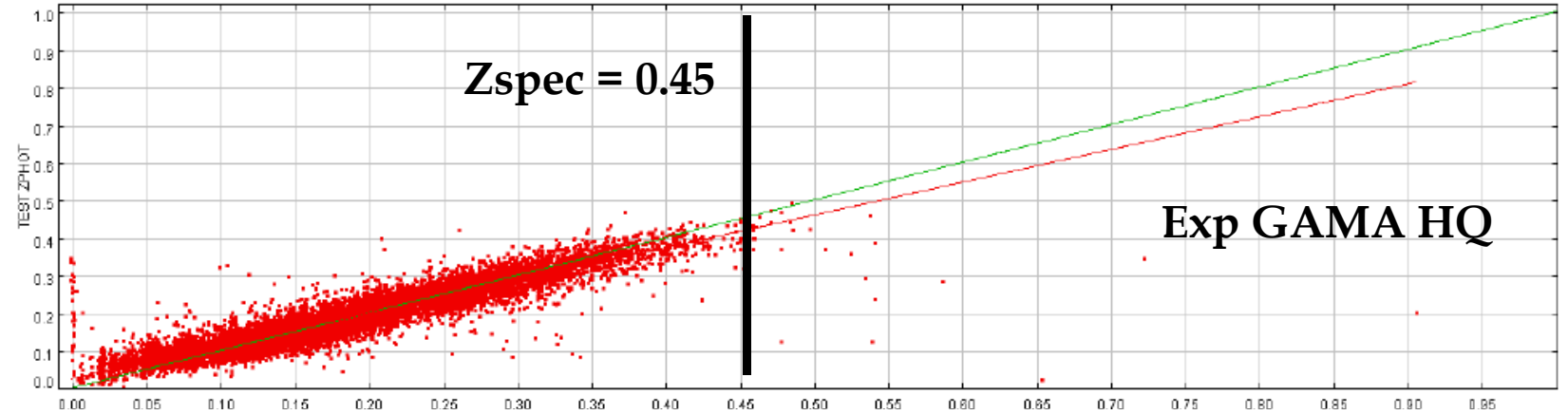


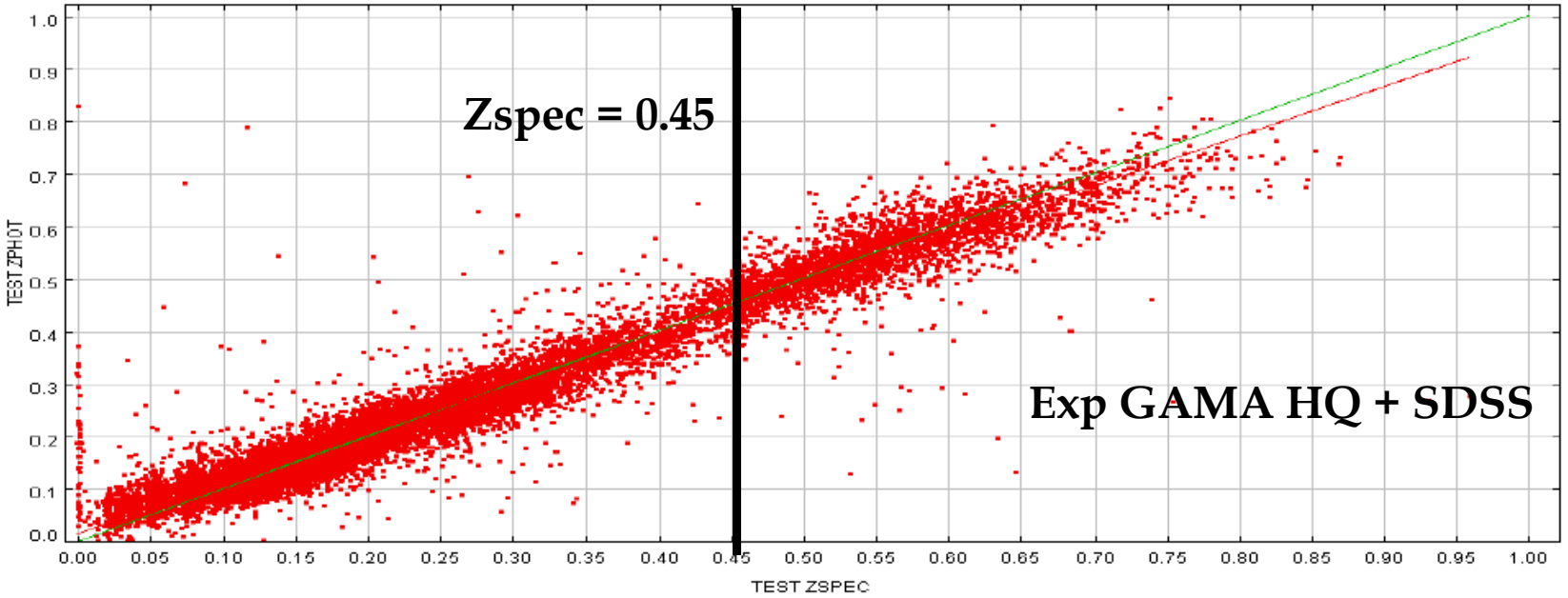
zspec distribution of train (black) and test (red) sets, based on OPTICAL photometry and GAMA spectroscopic information (HQ redshifts)



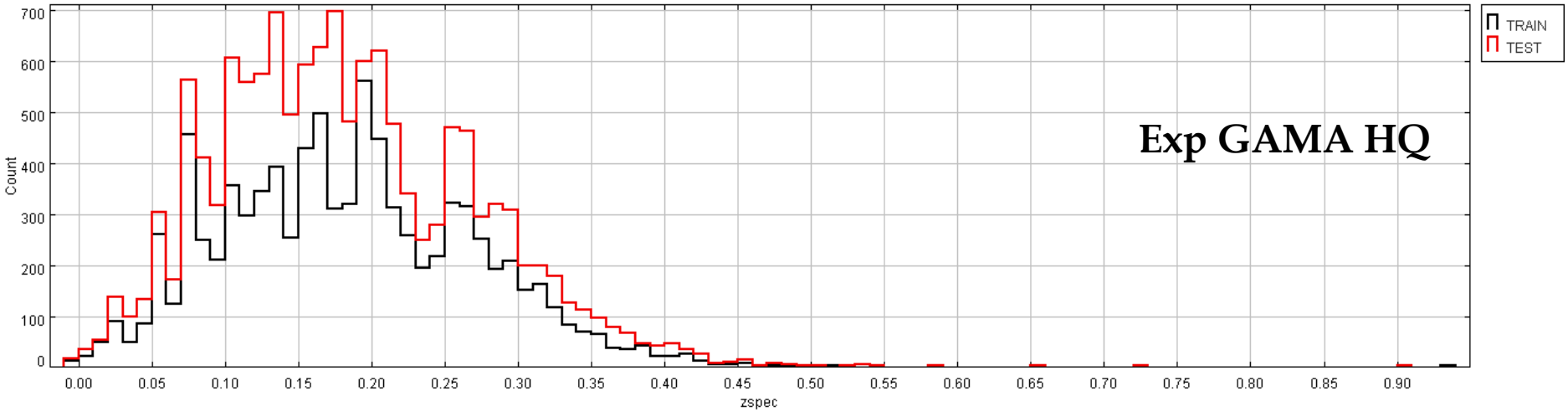
zspec distribution of train (black) and test (red) sets, based on OPTICAL photometry and GAMA+SDSS spectroscopic information

Experiment ID	| bias |	σ	NMAD	Outliers % | Δz | > 0.15	Outliers % | Δz | > σ	Outliers % | Δz | > 2σ
GAMA HQ + SDSS	0.00086	0.0305	0.0206	0.42	18.23	3.06
GAMA HQ	0.00099	0.0267	0.0198	0.30	21.2	3.34

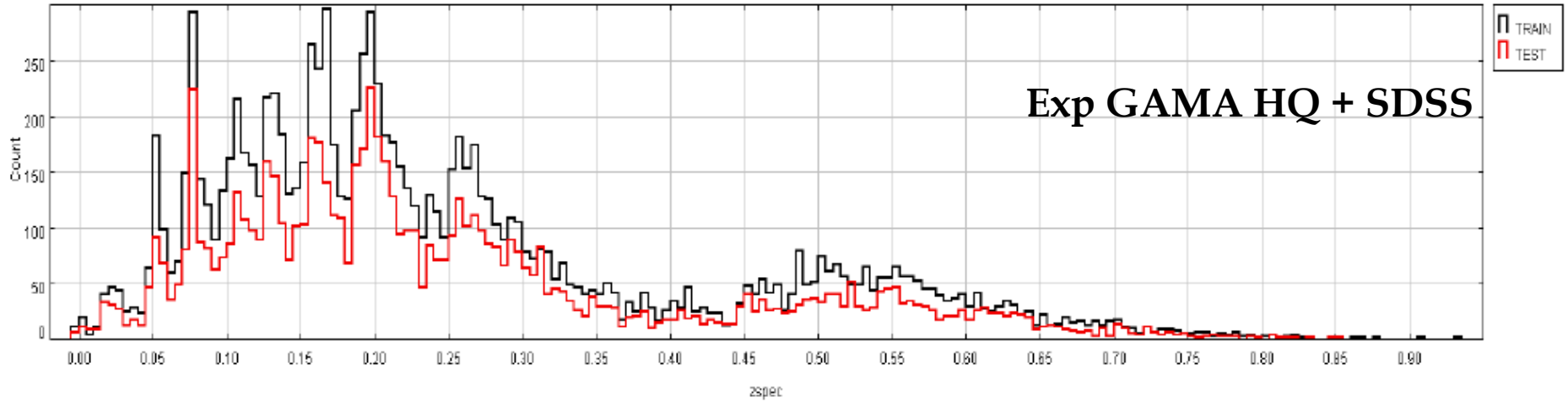




Experiments Optical + NIR Zspec Distribution

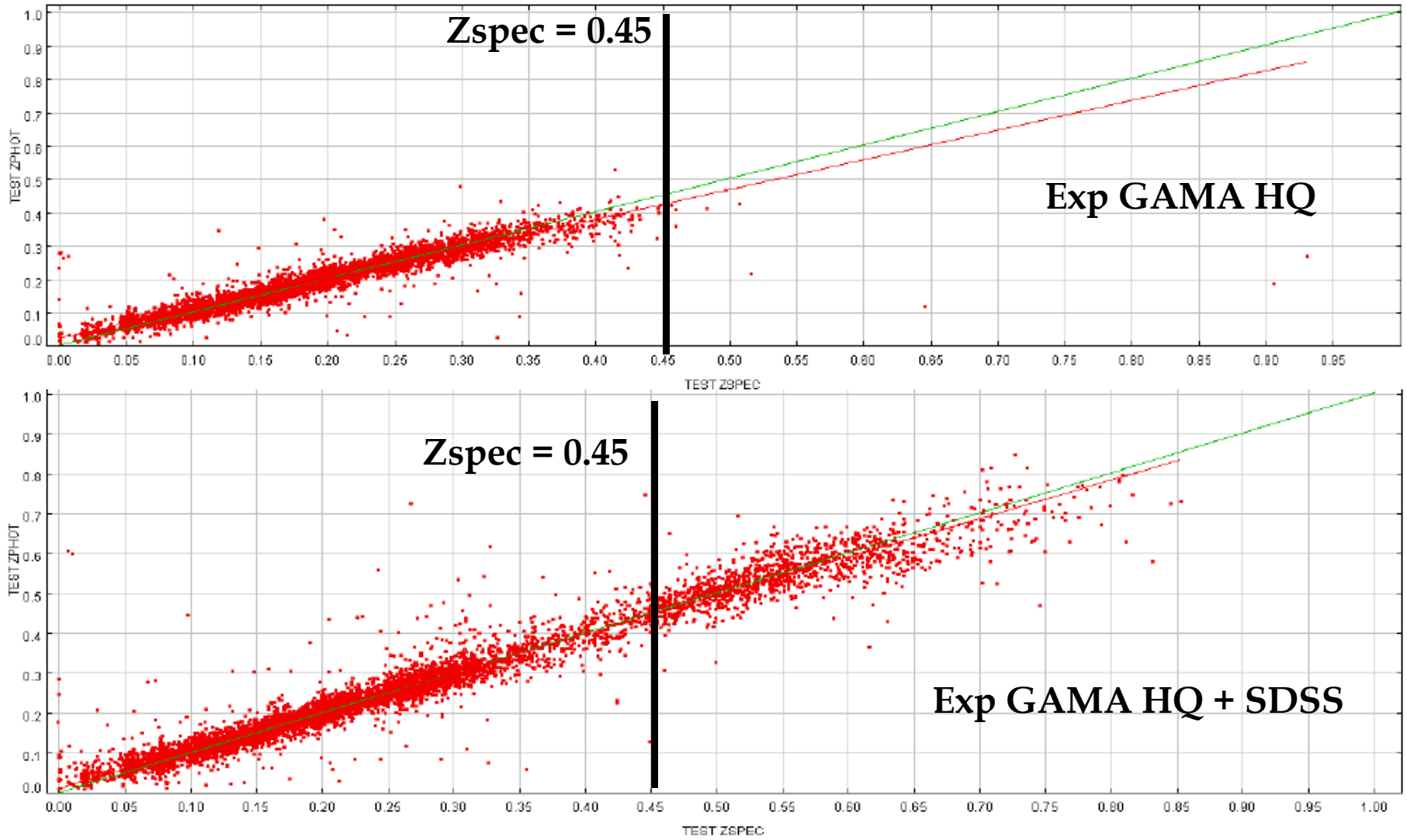


zspec distribution of train (black) and test (red) sets, based on OPTICAL+NIR photometry and GAMA spectroscopic information (HQ redshifts)



zspec distribution of train (black) and test (red) sets, based on OPTICAL+NIR photometry and GAMA+SDSS spectroscopic information

Experiment ID	|bias|	σ	NMAD	Outliers % |Δz| > 0.15	Outliers % |Δz| > σ	Outliers % |Δz| > 2σ
GAMA HQ+SDSS	0.00071	0.0266	0.0155	0.41	14.88	3.01
GAMA - HQ	0.00017	0.0220	0.0139	0.25	15.23	2.62



Conclusions and Future Developments. I

- The use of High Quality (HQ) data (e.g. NQ > 2) only, leads to a significant gain in the photo-z. This is confirmed by all experiments;
- The mixture of GAMA + SDSS spectroscopic data slightly extends the spectroscopic base of knowledge;
- By comparing the experiments OPTICAL+NIR with a single NIR band, the best performances are obtained with NIR magnitude aperture 2.8 arcsec;
- As expected, NIR band improves z-phot. It is confirmed that the additional NIR contribution is more relevant than the reduction of knowledge base (approximately halved);

- The Knowledge Base available at higher redshift (e.g. Z>0.45) is still very limited to give the possibility to evaluate the performances at higher redshift;
- The presence of some objects scattered around the minimum zspec (around zero) seems to indicate a residual presence of stars within the sample.

AND…

Conclusions and Future Developments. II

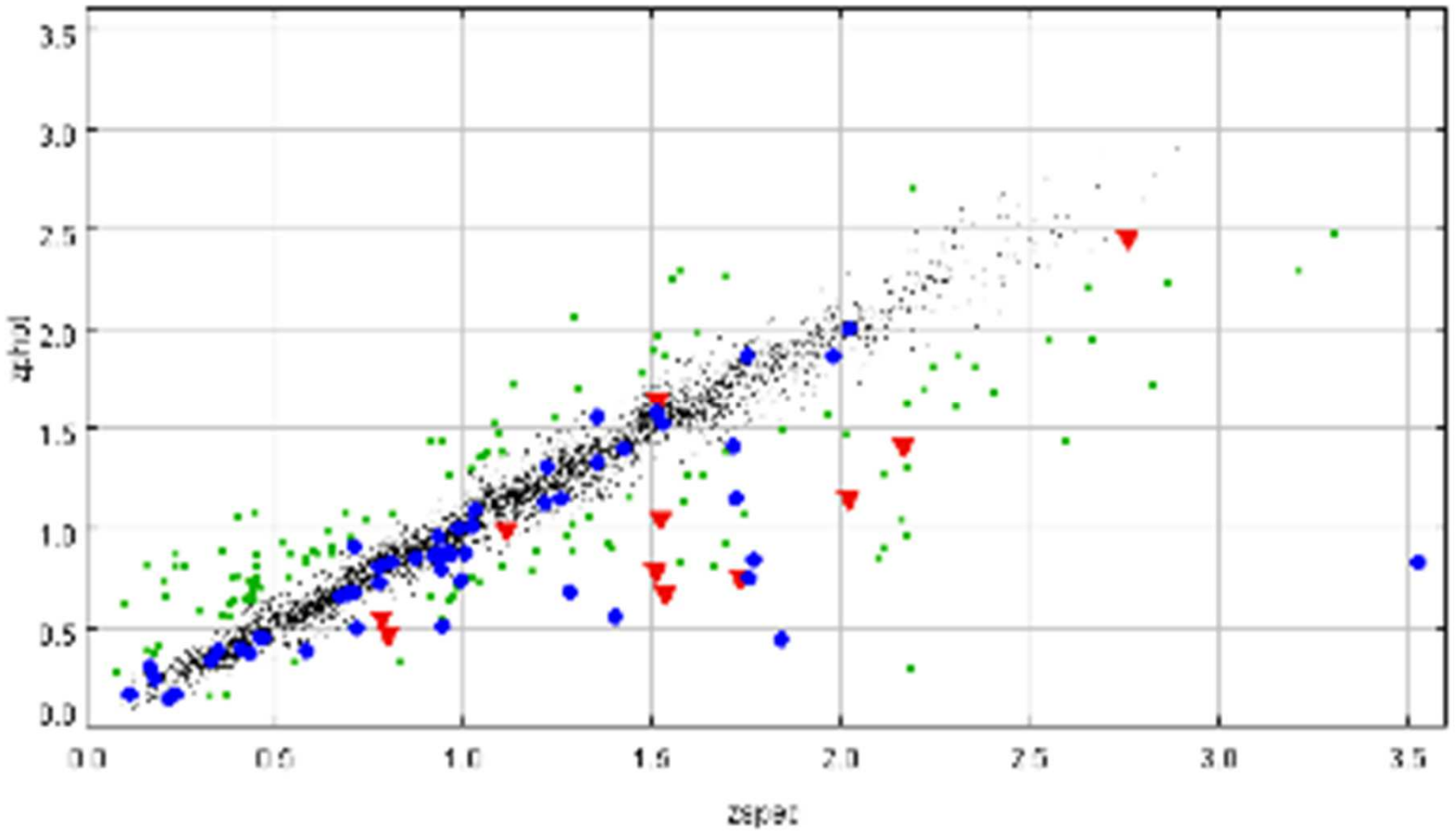
Improvements may come:

- Improved Knowledge Base (better coverage of photometric/redshift parameter space);
- Better photometry (especially in the NIR);
- Larger photometric coverage (other bands)
- **Better treatment of catastrophic outliers (by sacrificing a little completeness against accuracy);**

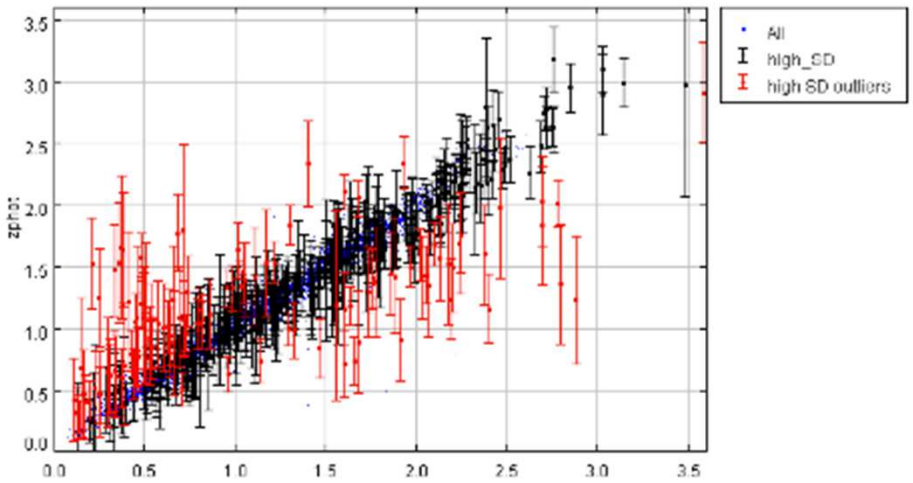
Petrillo, Longo, Brescia, Cavuoti, 2014,
in preparation (characterization of catastrophic outliers in the QSO and galaxy SDSS samples)

• 10 independent trainings lead to 10 independent Photo-z estimates for objects in the test set…
• This allows to derive individual st. dev for each object in the photometric sample

Results for the SDSS QSO sample



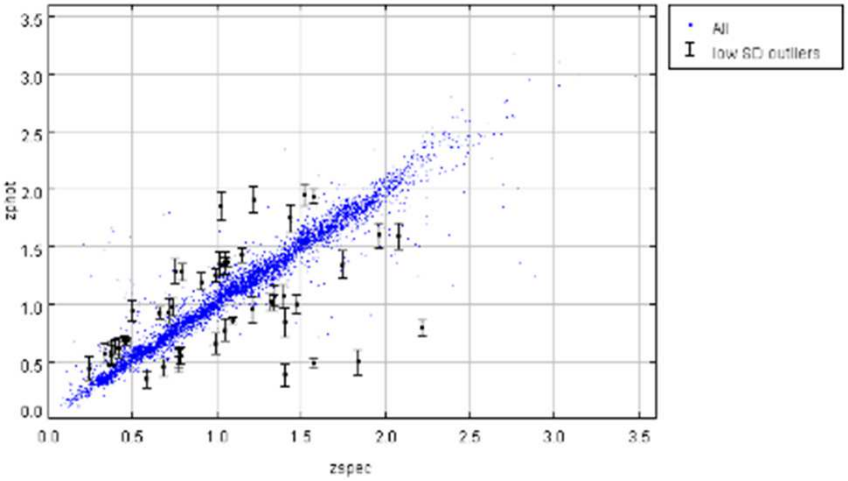
Resulting scatter plot (zspec vs zphot) for the experiment with all blazars and gravitationally lensed quasars placed in the test-set. Black dots are all test-set sources; blue circles are blazars; red triangles are lensed quasars; green dots are the outliers.



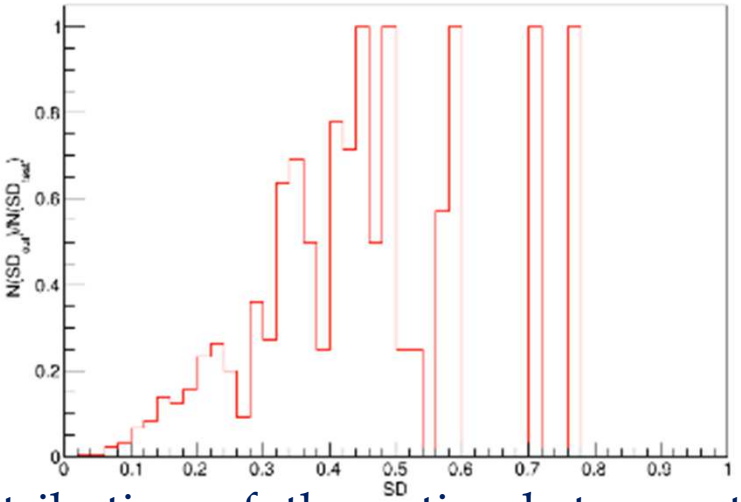
Scatter plot (zspec vs zphot) for the mean experiment. Sources with a standard deviation greater than 0.125 are marked in black or red (if are outliers). Blue dots indicate the entire test-set

Two types of behaviour: low standard deviation and high standard deviation.

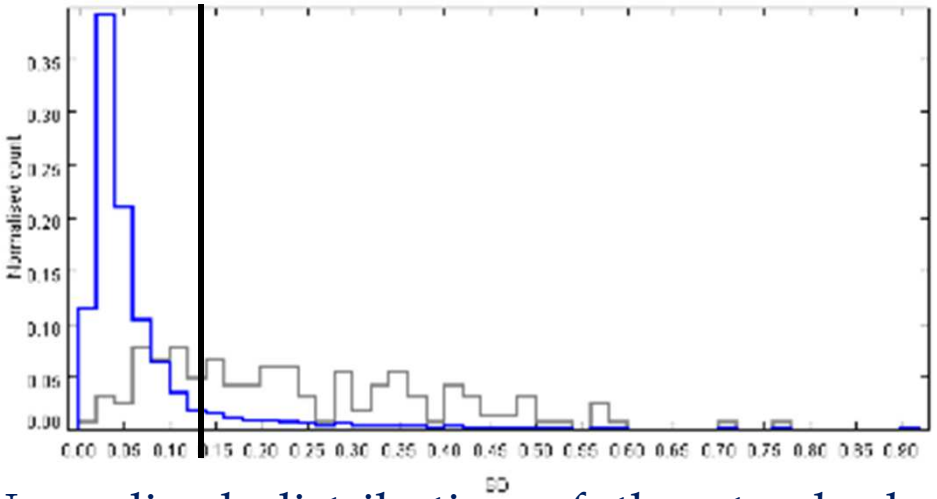
THE HIGH STANDARD DEVIATION component is dominated by catastrophic outliers



Scatter plot (zspec vs zphot) for the mean experiment. Outliers with a standard deviation less than 0.125 are marked in black



Distribution of the ratios between the value of the standard deviation of the outliers and the standard deviation of the entire test-set



Normalized distribution of the standard deviation of the entire test-set (Blue line) and of the outliers only (Grey line)

Lost of completeness < 20%

	Homogenous	*Average*	*Average low SD*
Dataset	14284	14284	(14284 - 367)
BIAS(Δz)	0.002	0.0001	0.0007
$\sigma(\Delta z)$	0.14	0.12	0.077
MAD(Δz)	0.043	0.036	0.034
RMS(Δz)	0.14	0.12	0.077
NMAD(Δz)	0.063	0.054	0.050
$> 2\sigma(\Delta z)$	2.94%	3.17%	3.67%
$> 4\sigma(\Delta z)$	1.14%	0.10%	0.40%
BIAS(Δz_{norm})	0.003	0.003	0.0005
$\sigma(\Delta z_{norm})$	0.70	0.059	0.037
MAD(Δz_{norm})	0.021	0.018	0.017
RMS(Δz_{norm})	0.070	0.060	0.037
NMAD(Δz_{norm})	0.031	0.027	0.025
$> 2\sigma(\Delta z_{norm})$	2.66%	2.98%	3.87%
$> 4\sigma(\Delta z_{norm})$	0.84%	0.89%	0.57%

The number of sources in the dataset, the statistical indicators and percentages of catastrophic outliers for the average experiment (second column) and the homogeneous experiment for comparison (first column). The recomputed quantities after the removal of 367 objects with an high standard deviation (third column)