Automated reliability assessment for redshift measurements

S. Jamal, V. Le Brun, O. Le Fèvre, D. Vibert, A. Schmitt,
C. Surace, P.-Y. Chabaud, M. Gray, F. Fauchier, M. Leurent

Aix Marseille University, CNRS, LAM (Laboratoire d'Astrophysique de Marseille)
Marseille, France

I. Introduction

Future large-scale surveys as Euclid will produce a large set of data *(1.2 10^9 observed sources)*

→ Need for fully automated data-processing pipelines.

Primary feature to measure : the redshift z.

Photometric redshifts:

$\mathbf{z}_{\text{phot, estimate}}$: *template fitting, artificial neural network, Bayesian inference* [1][2].

Spectroscopic redshifts:

$\mathbf{z}_{\text{spec, estimate}}$: *cross-correlation* [3][4], χ^2 *minimization* [5][6].

❖*References*

[1] *BPZ, N. Benitez, 1998.*
[2] *ZEBRA, R. Feldman, 2006.*
[3] *Darth Fader, D. Machado & al. , 2013.*
[4] *J. Tonry & M. Davis, 1979.*
[5] *EZ, B .Garilli. & al., 2010*
[6] *P. Schuecker, 1993*

II. **Redshift estimation** - *example* (1/3)

Observed spectrum

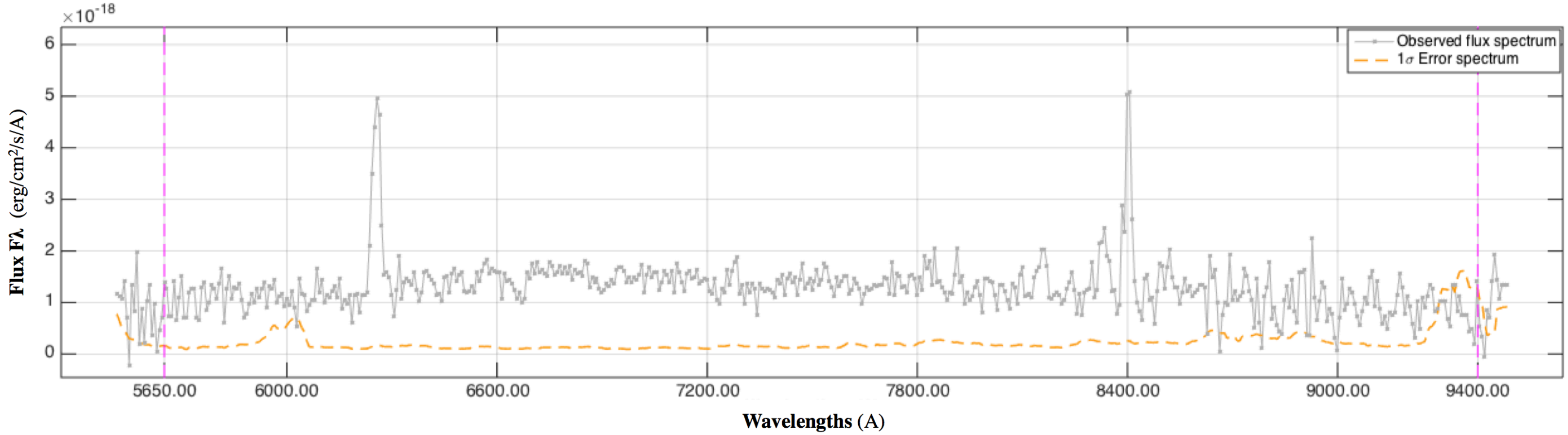
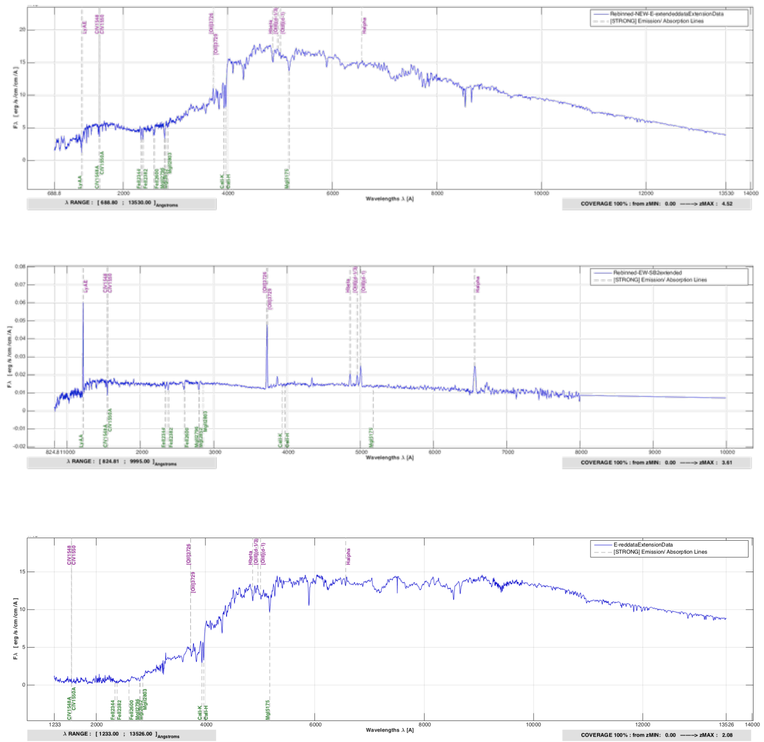


Figure 1 *– Galaxy spectrum from VVDS (Deep F02)*

Spectroscopic redshift ?

II. **Redshift estimation** - *example* (2/3)

[Reference] Spectroscopic templates



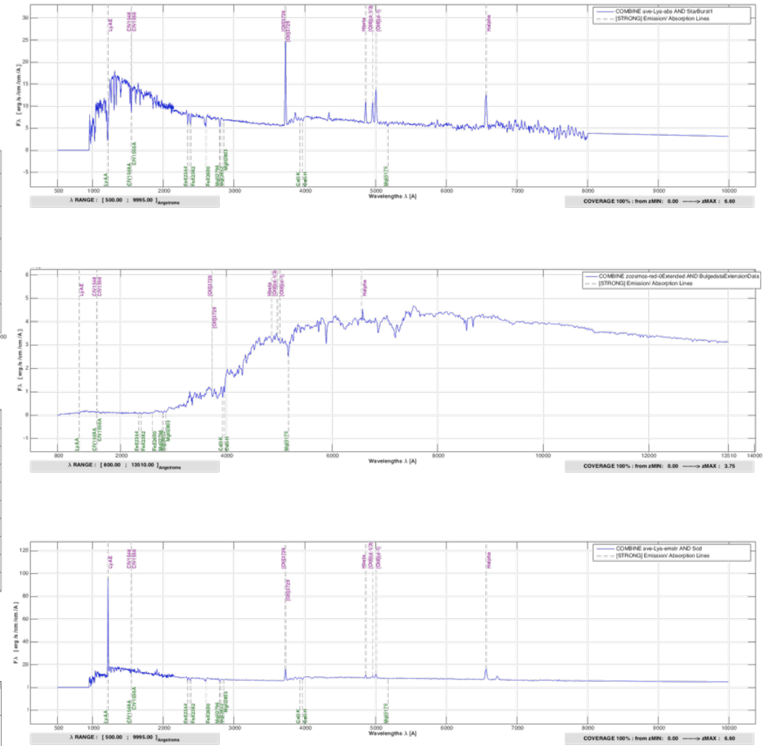


Figure 2 – Reference set of spectroscopic templates

II. **Redshift estimation** - *example* (3/3)

Best solution (z, Tpl)

Estimated redshift $z_{MAP} = \text{argmax}_z \; p\,(\,z \mid data\,)$

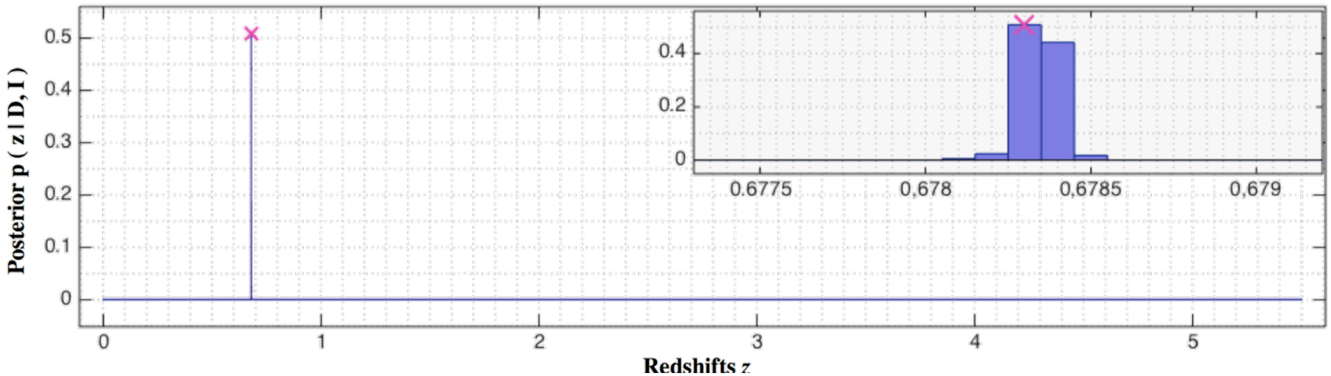


Figure 3
Display of the zspec probability density function

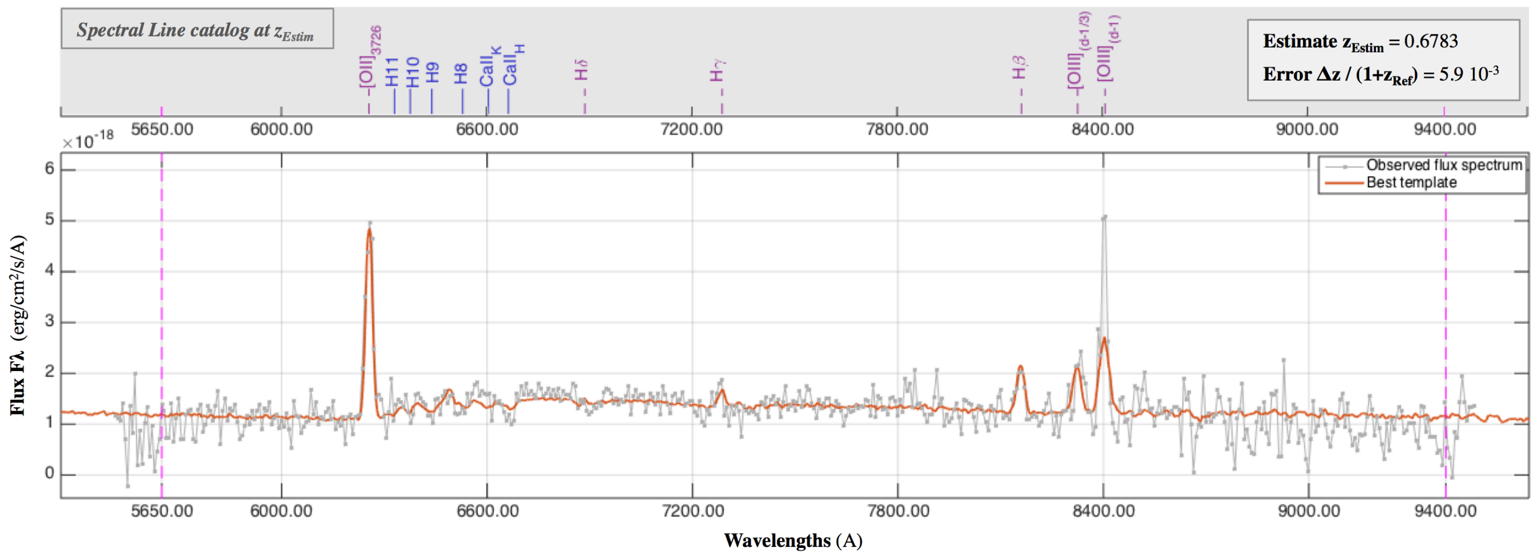


Figure 4
Best template a z_{MAP}

III. zSpec - Quality assessment (1/8)

'Best' redshift solution	• **Max**imization of the posterior PDF p (z ǀ data, priors) • **Min**imization of the chi-square operator

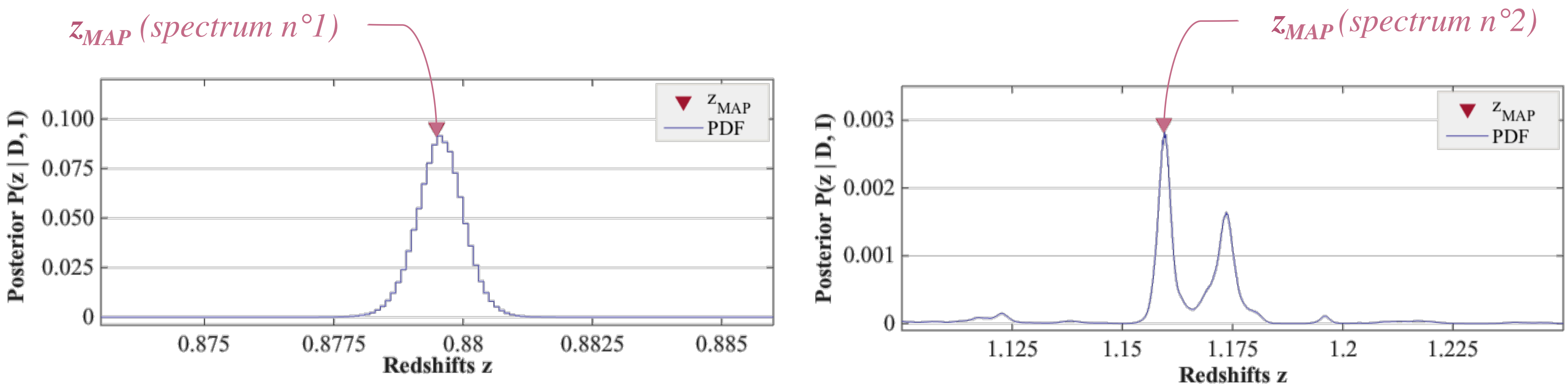


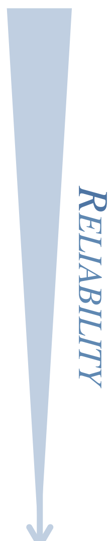
Figure 5 – *Display of two zspec PDFs, obtained for two different VVDS (Deep field) spectra*

Same *level of confidence between these 2 redshifts ?*

PDF: *Probability distribution function*
MAP: *Maximum-a-Posteriori estimate*

III. zSpec - Quality assessment (2/8)

Unimodal PDF *= 1 strong redshift candidate*



Multimodal PDF *= ++ redshift candidates* (!!)

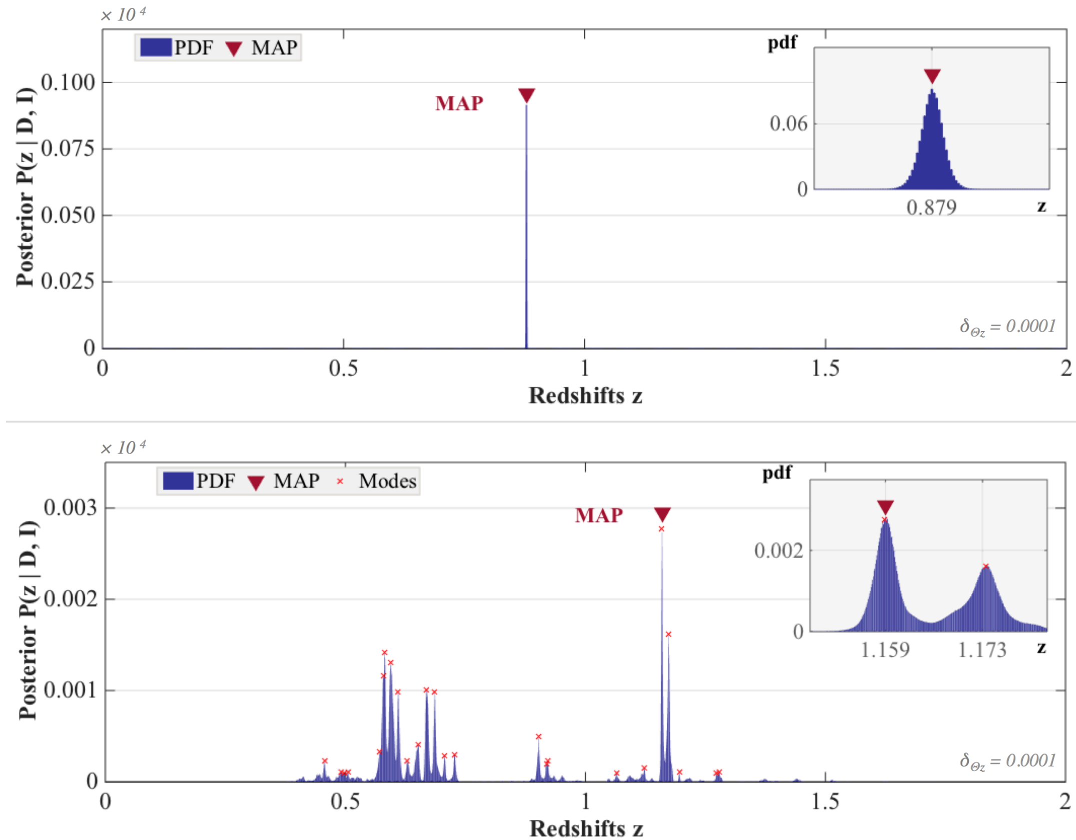


Figure 6 *– Display of unimodal and multimodal zspec PDFs, obtained for two VVDS (Deep field) spectra*

PDF: *Probability distribution function*
MAP: *Maximum-a-Posteriori estimate*

III. zSpec - Quality assessment (3/8)

Inputs *of an automated system ?*

Descriptors of the zPDF

- *Significant modes*
  ⋮
- *$P(z_{MAP})$*
- *CR characteristics*
- *Dispersion of the zPDF*

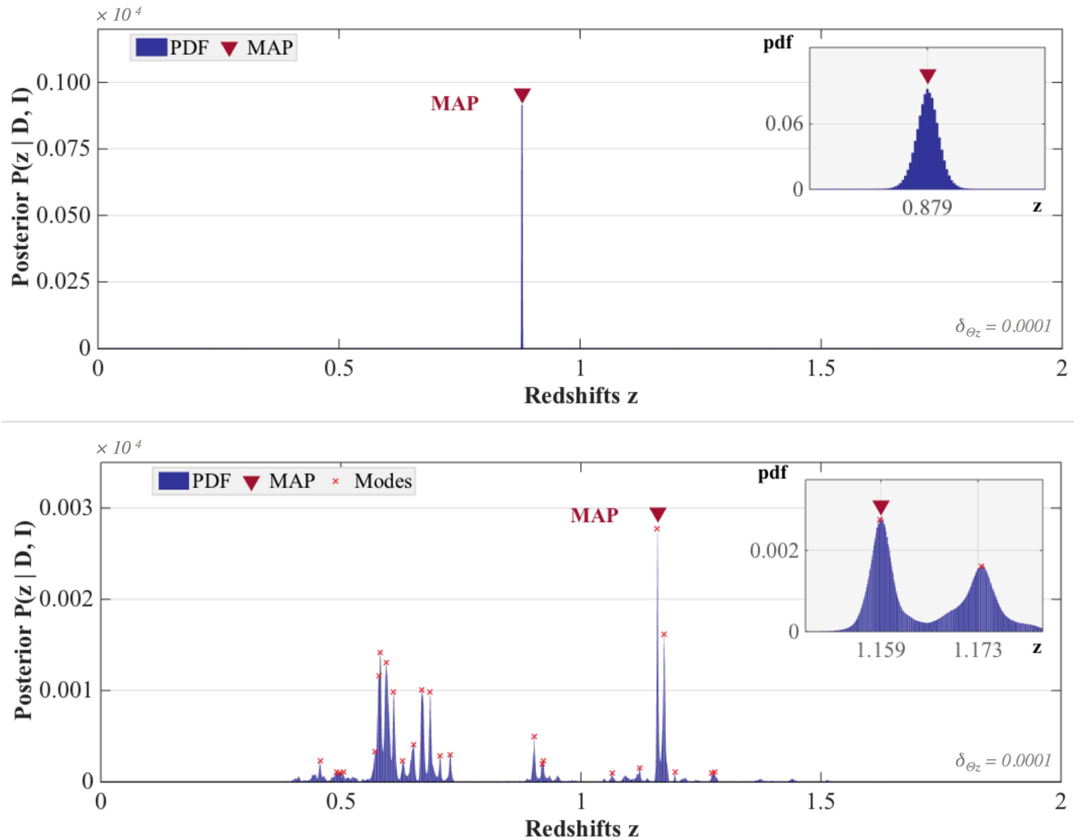


Figure 6 – *Display of unimodal and multimodal zspec PDFs, obtained for two VVDS (Deep field) spectra*

PDF *: Probability distribution function*
MAP: *Maximum-a-Posteriori estimate*
CR *: credibility region with 95% of probability*

III. **zSpec - Quality assessment** (4/8)

From the *zPDF* of a spectrum s_i, extract a list of L descriptors

⇒**Feature vector** $\boldsymbol{x_i} = (d_1 \quad \ldots \quad d_L)$

For a set of **N** spectra $(s_i)_{\mathrm{i} \in \{1 \ldots \mathrm{N}\}}$

Explanatory variables $\underline{\mathbf{X}} = (x_i)_{\text{Descriptors of zPDFs}}$

Response variables $\underline{\mathbf{Y}} = (y_i)_{\text{Quality Flags}}$

Relationship ?

Machine Learning

III. zSpec - Quality assessment (5/8)

SUPERVISED APPROACH ?

Label prediction using the generated mapping

Train a classifier → To produce a mapping between X and Y

Explanatory variables $\underline{\mathbf{X}} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}$

Response variables $\underline{\mathbf{Y}} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}_{\text{Quality Flags}}$

Use known labels in a database

III. zSpec - Quality assessment (6/8)

SUPERVISED APPROACH ?

~~*Label prediction using the generated mapping*~~

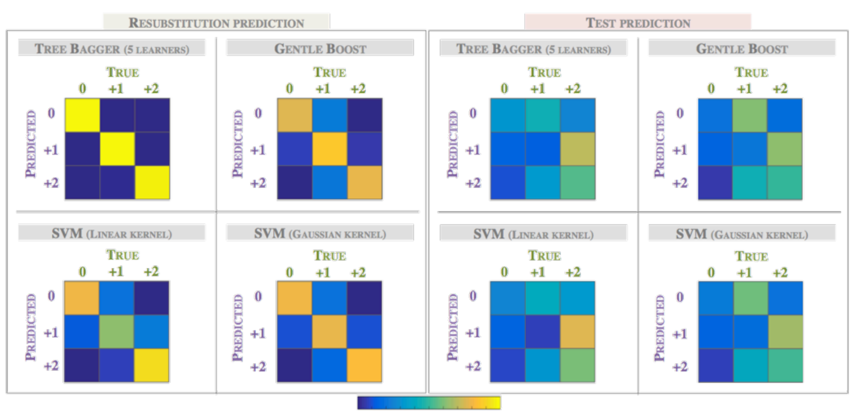
Train a classifier → ~~To produce a mapping between X and Y~~

No
Mismatch because of subjective labels

Explanatory variables $\underline{\mathbf{X}} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}$

Response variables $\underline{\mathbf{Y}} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$ Quality Flags

ECOC
(Error Correcting Output Codes)
for multiclass problems

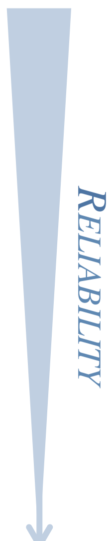


High off-diagonal element in confusion matrices (!!)

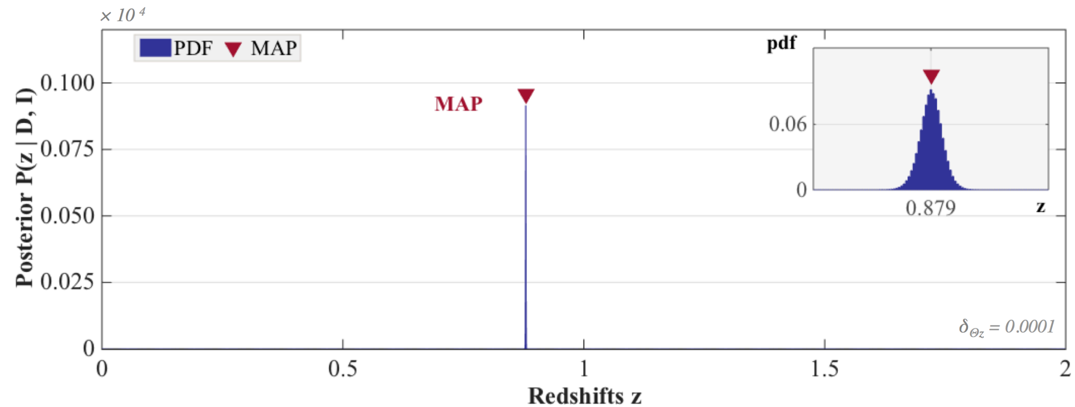
Figure 7
Confusion matrices obtained using existing redshift quality flags

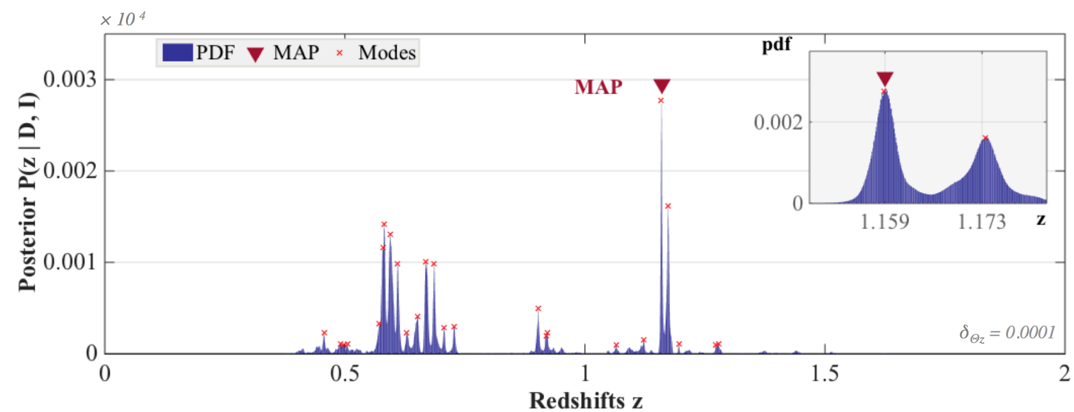
III. zSpec - Quality assessment (7/8)

Unimodal PDF



Multimodal PDF





III. zSpec - Quality assessment (8/8)

UNSUPERVISED APPROACH

Clustering algo → To partition the data into distinct clusters

Explanatory variables $\underline{\mathbf{X}} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}$

Response variables $\underline{\mathbf{Y}} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$ Quality Flags

VI. **ML tests** (1/6)

STEP (1) *Build a reference set*

Database of ***reliable*** *spectra*

zPDFs

Feature matrix $\boldsymbol{X}$

Quality labels $\boldsymbol{Y}$

$y_i \in \{\text{'C1'}, \ldots, \text{'C5'}\}$

controlled

TRAINING SET (X,Y)

VI. **ML tests** (2/6)

STEP (1) *Build a reference set*

Using the VVDS[1] database *(~24000 spectra)*

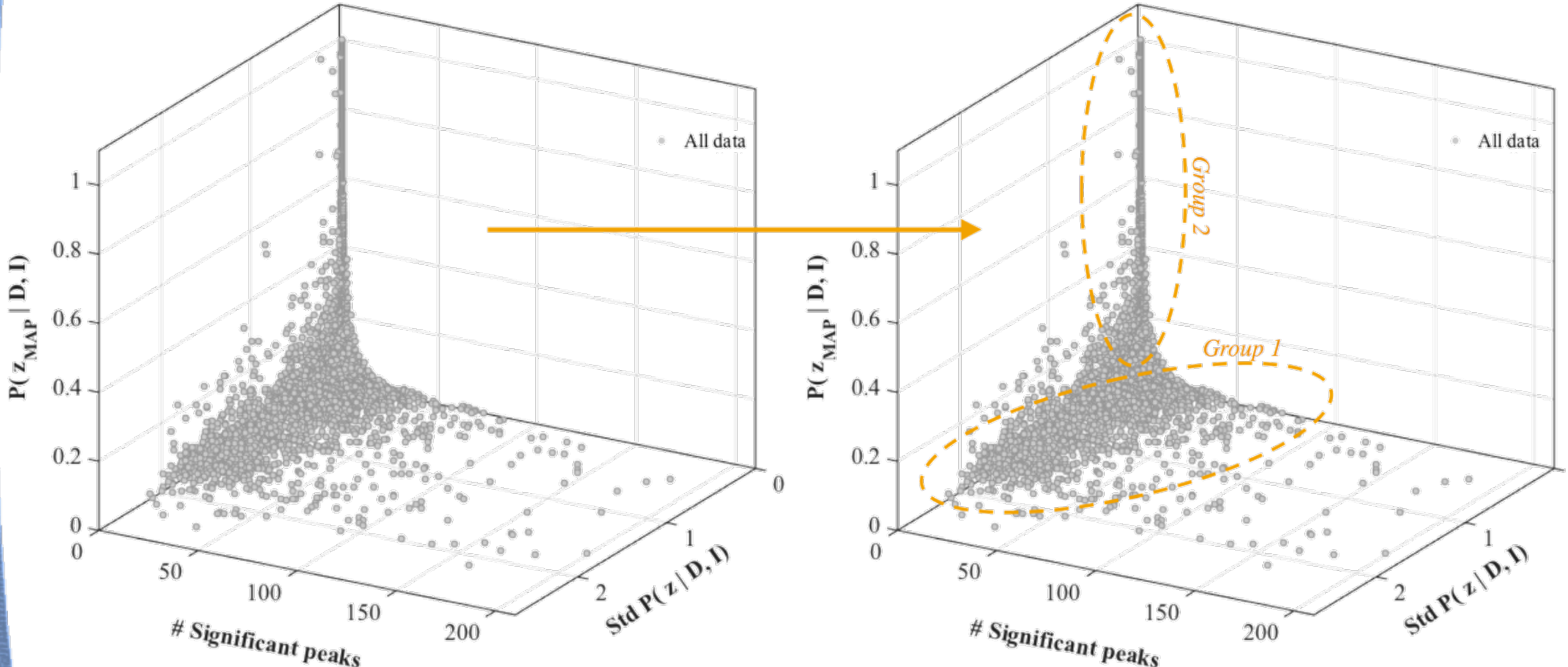


Figure 8 – *Data representation in a selected 3D space*

[1] *VIMOS VLT Deep Survey (VVDS)* http://cesam.lam.fr/vvds/

VI. **ML tests** (3/6)

STEP (1) *Build a reference set*

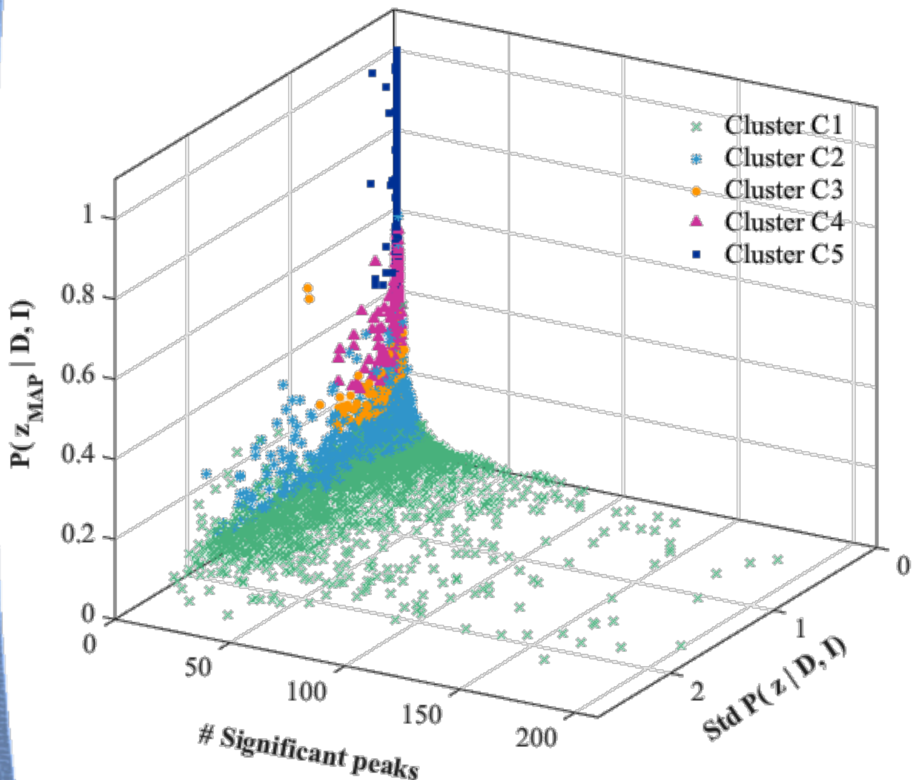
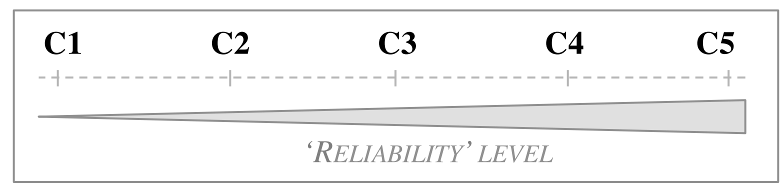


Figure 9 – Data representation in a selected 3D space after clustering

***FCM** algo*

(z_{Qual}) ***CLUSTERS***



VI. **ML tests** (4/6)

STEP (1) *Build a reference set*

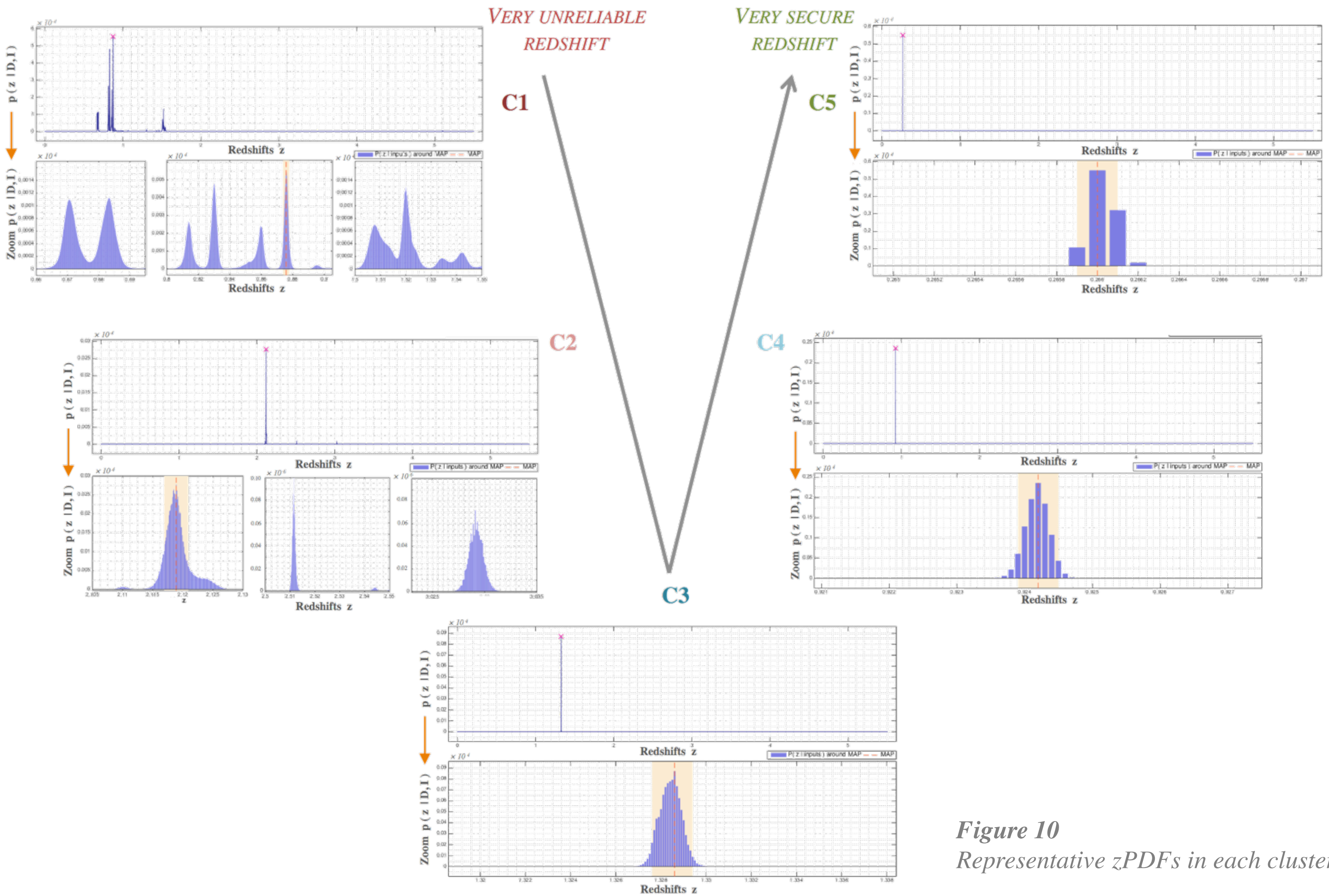


Figure 10
Representative zPDFs in each cluster

VI. ML tests (5/6)

STEP (1) *Build a reference set*

*Database of **reliable** spectra*

zPDFs

*Feature matrix **X***

*Quality labels **Y***

$y_i \in \{\text{'C1'}, \ldots, \text{'C5'}\}$

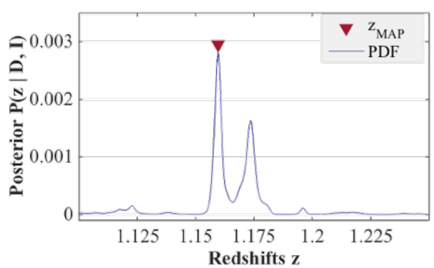
controlled

TRAINING SET *(X,Y)*

map

STEP (2) *Predict a reliability flag for unlabelled data*

New spectrum s_k *(unlabelled)* → *Compute the zPDF* → *Extract descriptors from the zPDF* → ***PREDICTED CLASS*** y_k *{'C1',...,'C5'}*



$$x_k = (d_1 \ldots d_L)$$

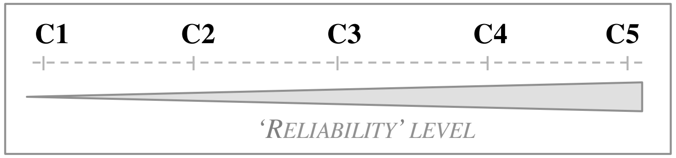
- *Significant modes*

⋮

- *Dispersion*

VI. **ML tests** (6/6)

STEP (2) *Predict a reliability flag for unlabelled data*



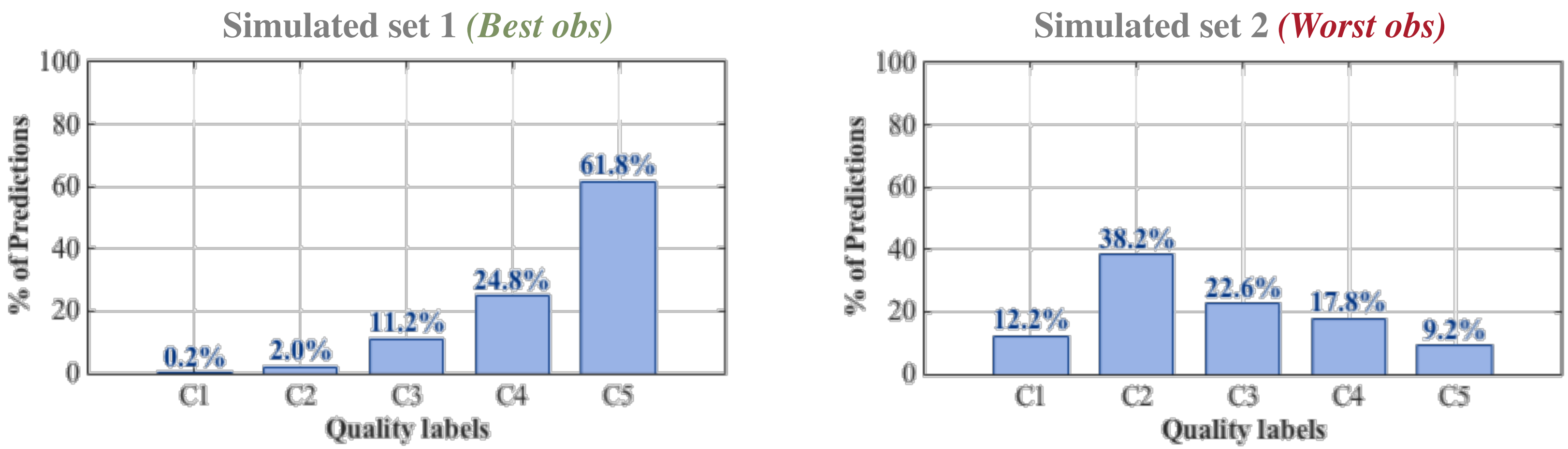


Figure 11 – *Redshift reliability predicted clusters of simulated datasets (S1-S2)*

V. Perspectives

Result An automated quality assessment of the estimated redshift z_{spec} via:

- Exploiting the redshift *zPDF* p (z | data, priors)
- Machine Learning (ML) algorithms

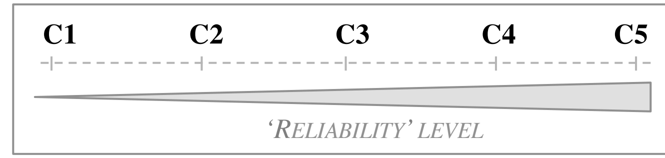
Next ?

- Fuzzy approach
- Performance evaluation on simulated data
- Advanced ML-algorithms (complex learning scheme? need? etc.)

Thank you for your attention

VI. **ML tests** (5/6)

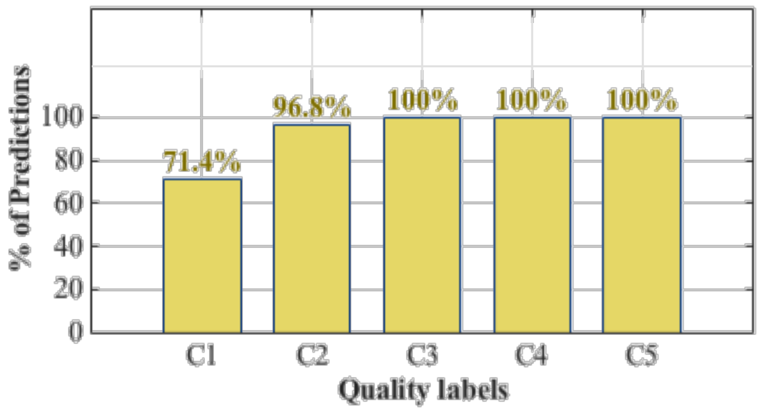
(z_{Qual}) CLUSTERS



Redshift error $\varepsilon_z = |z_{Ref} - z_{Estim}| \,/\, (1+z_{Ref})$

Criterion $\varepsilon_z \leq 0.001$ *for example*

Simulated set 1 ***(Best obs)***



Simulated set 2 ***(Worst obs)***

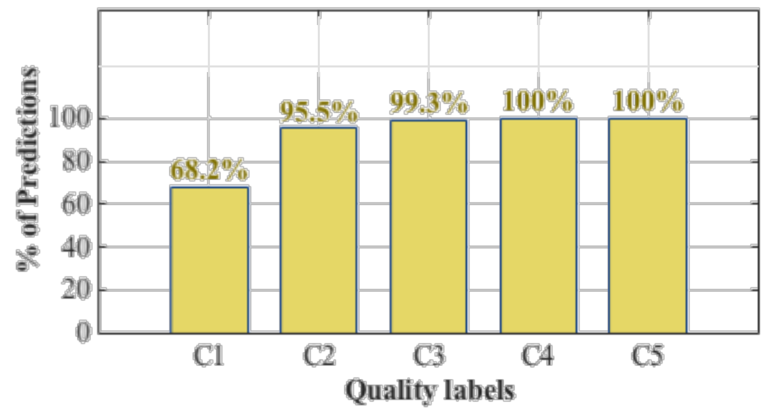


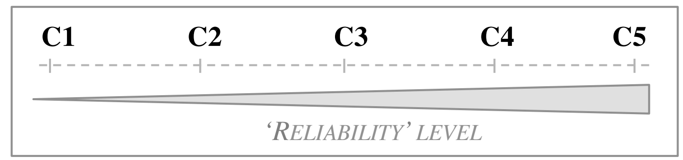
Figure 12 – *Fraction of redshift error* $\Delta z/(1+zRef) \leq 0.001$ *in each predicted partition*

VI. **ML tests** (6/6)

Redshift error $\varepsilon_z = |z_{Ref} - z_{Estim}| / (1+z_{Ref})$

Criterion $\varepsilon_z \leq 0.001$ *for example*

(z_{Qual}) CLUSTERS



Simulated set 1 *(Best obs)*

Simulated set 2 ***(Worst obs)***

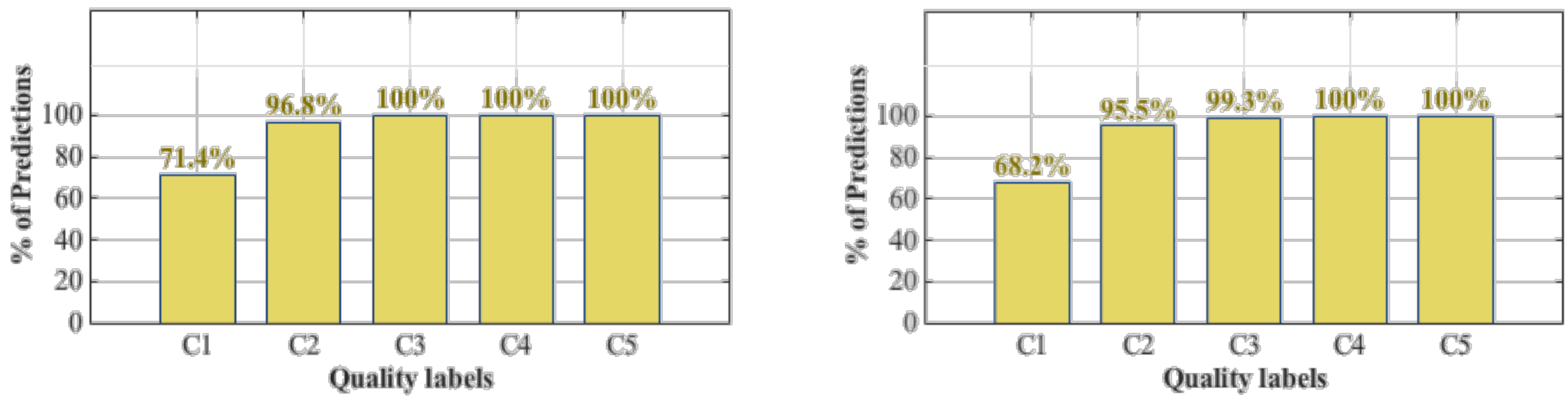


Figure 13 – *Fraction of redshift error* $\Delta z/(1+zRef) \leq 0.001$ *in each predicted partition*

↘ in C1/C2

Correlation [z_{Qual} ; ε_z] ?