Gaia DU831-QSOC

Porto 06-09 / 06 / 2011

Collaborators : J. Surdej, J. Poels, J-F Claeskens, A. Smette and P. Geurts

QSOC (Quasar Classifier)

- APs derivation (redsift, quasar type, …) using BP/RP spectra:
  - BP / RP spectrum = 180 pixels each = function { quasarType, redshift, spectral index (the power law continuum slope), the total equivalent width, the dust extinction}

- Preparation of the training data for QSOC and for DSC:
  - DSC = Discrete Source Classifier is the heart of CU8. It classifies sources using BP/RP, RVS, Astrometry and Gmag.
  - DSC : very good QSO classification (~94% true positives)
  - DSC responsible : Kester Smith from MPIA

Training data preparation

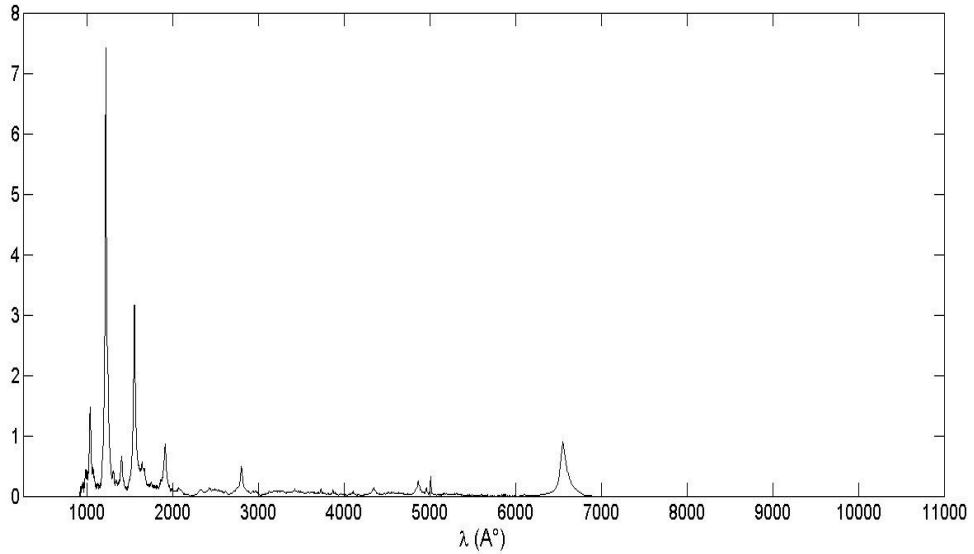
- 2 families of data :
  - Purely synthetic.
  - Semi synthetic : extrapolation of SDSS spectra from [380,920]nm to the BP/RP range [300,1100]nm

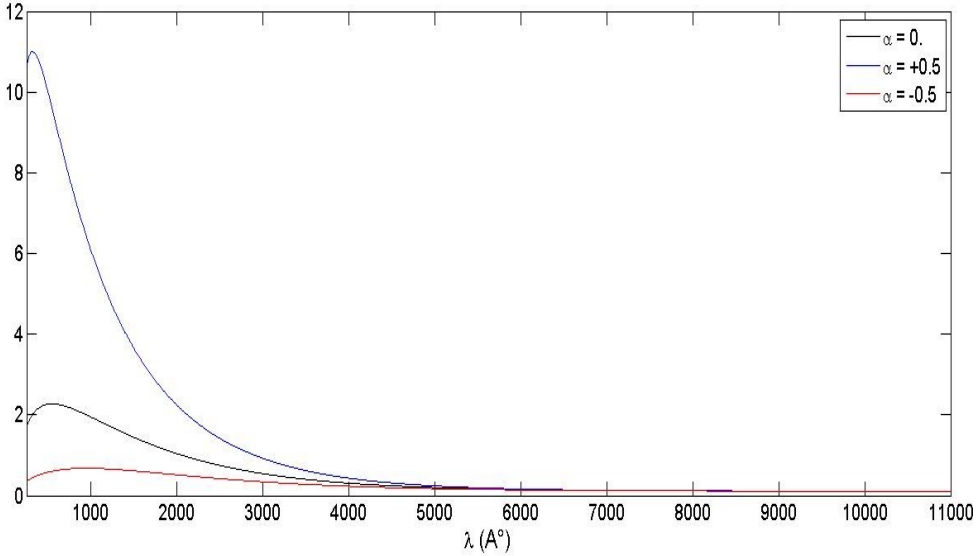
- Quasar types : should we use templates, Principal components ?
  - BAL QSOs (10% of QSOs) : no general template
  - Type1 and Type2 (20% of QSOs) : no defined class border.

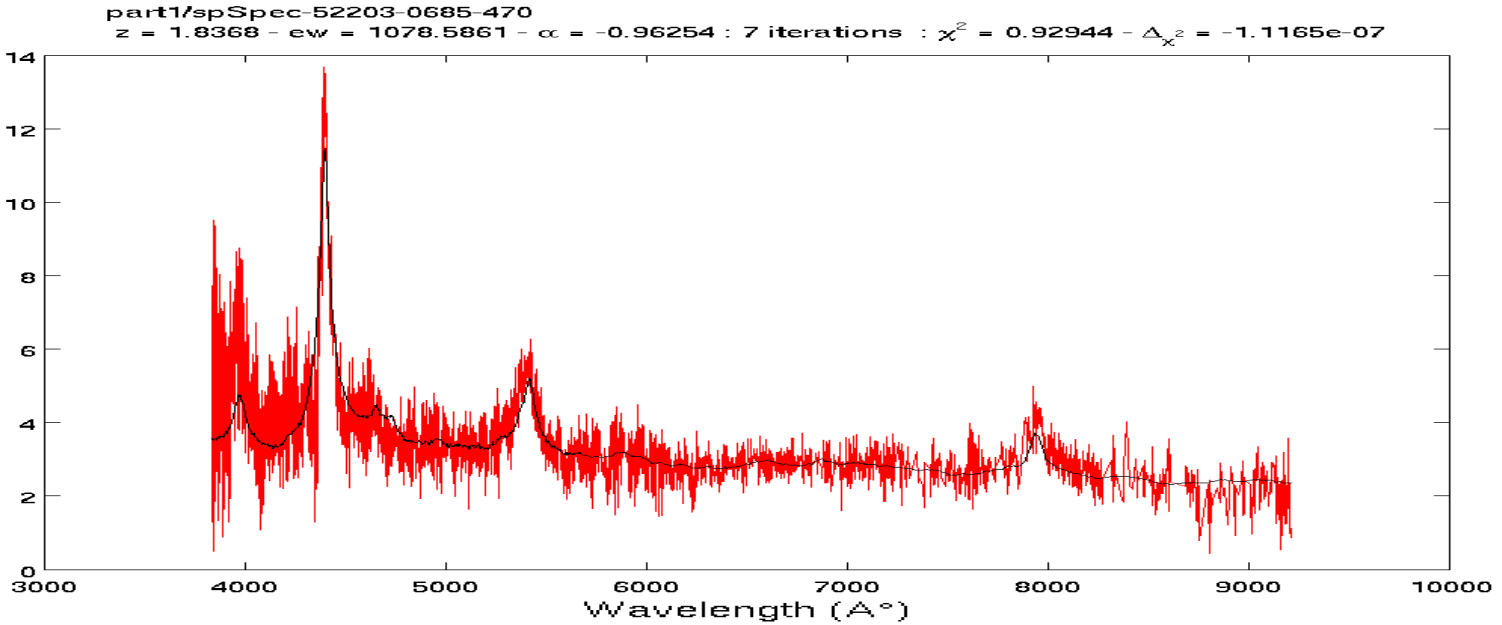
- Continuum modeling : 1 or 2 combined power laws.

- Dust extinction : at which resdhift it happens ?

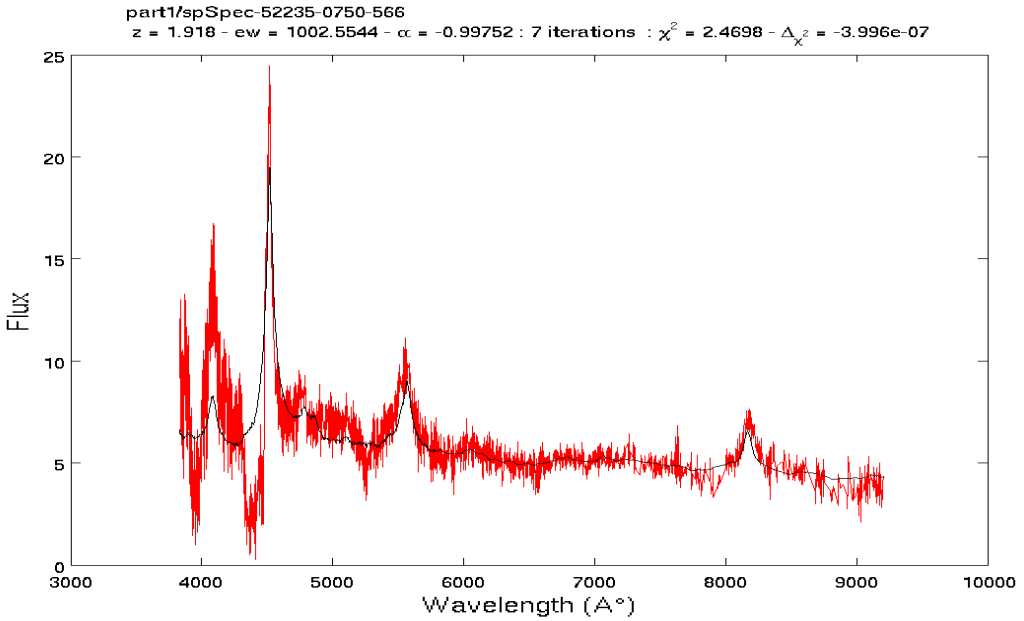
Training data preparation : Template + continuum



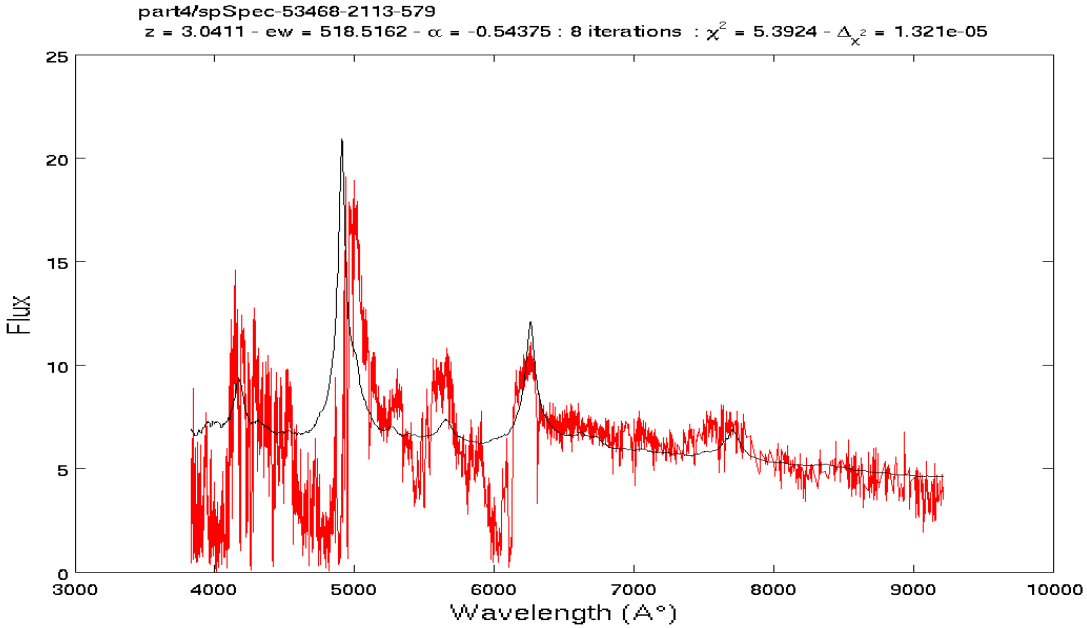




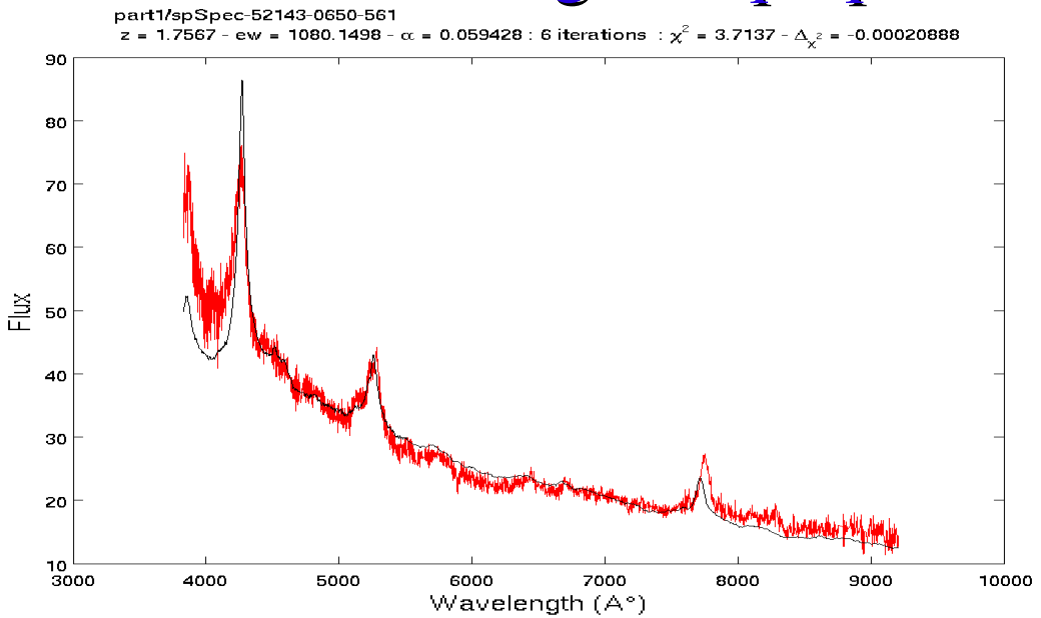
Training data preparation : Template + continuum



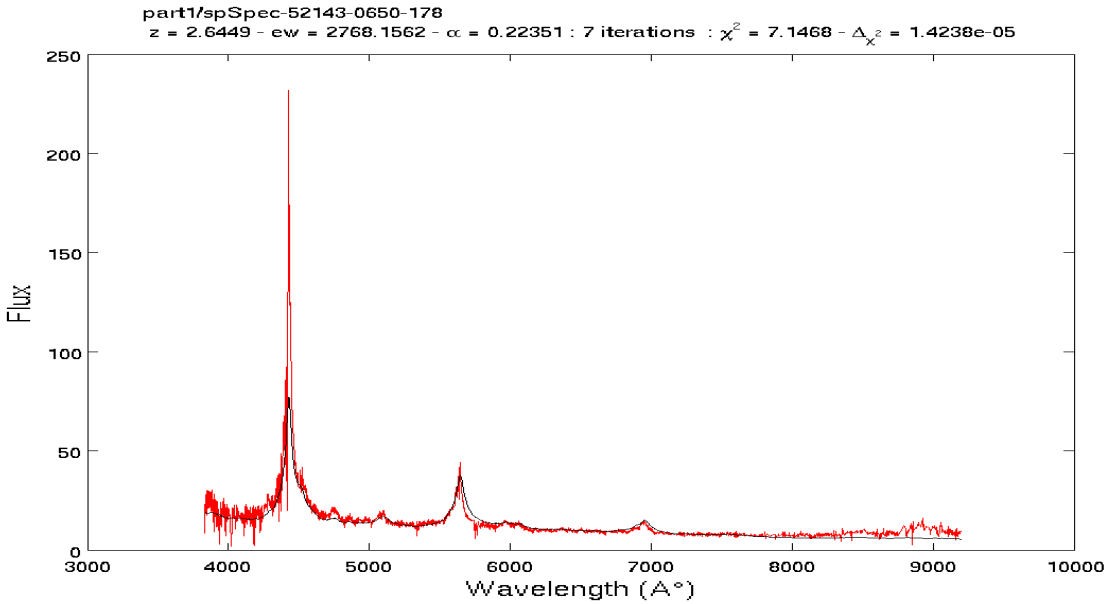
BALs



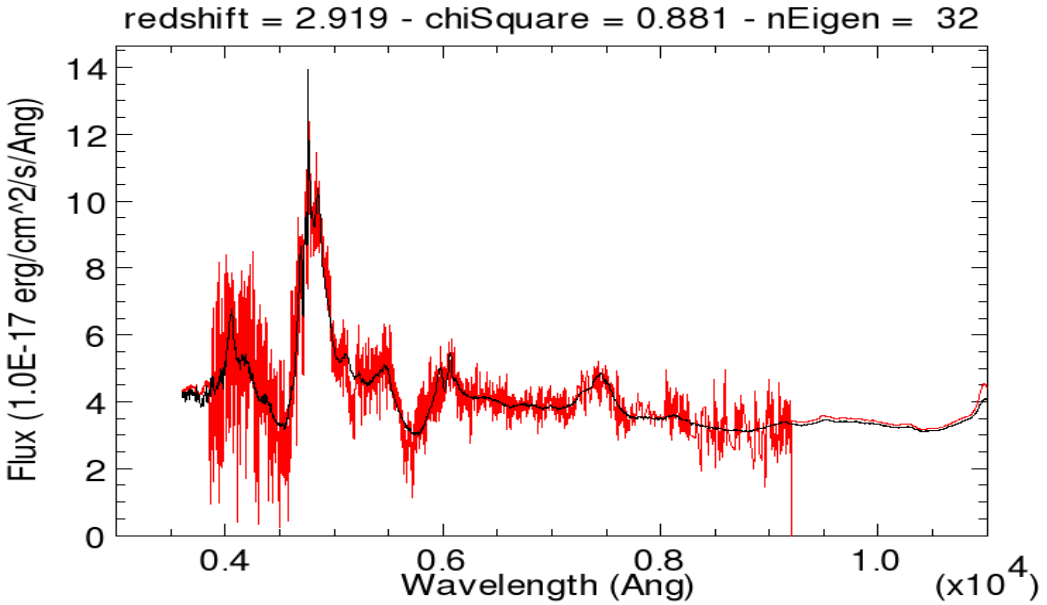
Training data preparation : Template + continuum



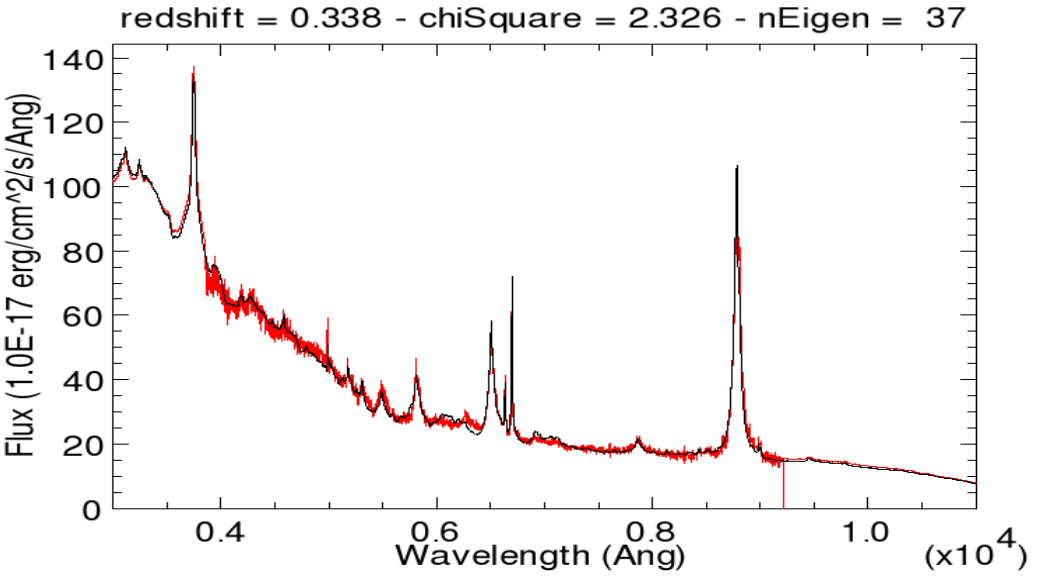
- Differential RedShift
- ≠ Line (Amplitude – Width)



Training data preparation : PCA

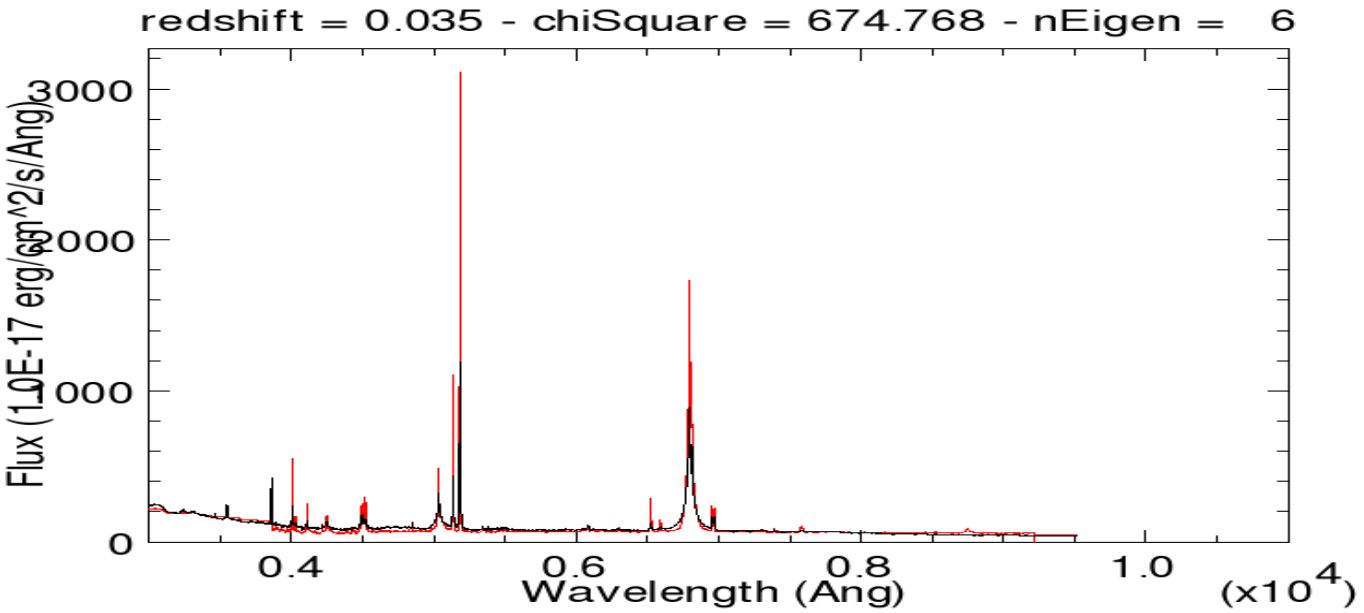


Yip et al. 2004, AJ, 128, pp 2603-2630



Training data preparation : PCA

- Better performances comparing to « Template + Continuum » method.
- Numerical difficulties : how to ensure positive fluxes in the extrapolated ranges.
- Can be improved by separating quasar types : PCs for each type (under progress)



QSOC : Ideal organogram.

- Foreach AP {redshift, continuum slope, total equivalent width, extinction} : finds and approached value by combining 2 (or more) pattern recognition algorithms (ERT, Annz, SVMc).

- Finds the quasar type using classification pattern recognition algorithm (SVMc or ERT).

- Uses inference to improve the approached APs using the extinction parameters from CU8/TotalGalacticExtinction package.

- Using a general minimisation process to refine the aproached APs using noise-free reference library => variance matrix and GoF.

Extremely Randomized Trees

Pierre Geurts, Damien Ernst and Louis Wehenkel, Machine Learning, 2006, Volume 63, Number 1, Pages 3-42

- Tree algorithm : regression and classification modes.
- Very easy to implement and to optimise : 2 algorithm parameters.
- Very easy to interpret (if (input > threshold) then ... otherwise ...).
- Very fast learning and predicting processes.
- Provides the AP variance at each tree leaf which can be used as an error estimation.
- Leaf output is a linear combination of inputs => CAN NOT extrapolate + BAD interpolation of data gaps !
- Depends highly on random numbers : not a deterministic process.

Neural network Annz

Photometric redshifts using Artificial Neural Networks (Collister & Lahav 2004)
http://www.homepages.ucl.ac.uk/~ucapola/annz.html

- Very light data model (comparing to trees) : a simple matrix of weights.
- Can learn using many starting values for the weights (changing the random seed) and combine the prediction with all the models => avoid local minima.
- User friendly code (C++).
- Slow learning but not huge memory consuming (comparing to Matlab nnet toolbox).
- Provides the AP variance.
- May suffer from imbalanced data.

Application : Photometric redshift using SDSS photometry

REGRESSION							
	Method parameters	Run 01 reddened 5 filters	Run 02 reddened 5 filters + Au	Run 03 reddened 4 colors + 1 filter	Run 04 reddened 4 colors + 1 filter + Au	Run 05 dereddened 5 filters	Run 06 dereddened 4 colors + 1 filter
Knn	RMS	0.415	0.469	0.392	0.414	0.408	0.387
	Number of nearest neighbors	17	10	25	18	14	27
ERT	RMS	0.416	0.416	0.380	0.378	0.409	0.375
	Number of trees	200	200	200	200	200	200
	Number of instances per node	5	6	12	9	6	11
	extraTreesK	4	6	5	5	5	5
RF	RMS	0.418	0.419	0.382	0.380	0.412	0.377
	Number of trees	200	200	200	200	200	200
	Number of instances per node	5	4	10	8	3	14
	extraTreesK	3	4	2	3	3	3
Mart	RMS	0.429	0.428	0.386	0.383	0.421	0.383
	Number of trees	200	200	200	200	200	200
	Maximum number of splits	860	843	310	262	750	194
	Mart mutation rate	0.0225	0.0250	0.0225	0.0245	0.0265	0.0280
SVR	RMS	0.430	0.467	0.408	0.417	0.426	0.400
	Cost = 2 ^	-1.2500	-0.8750	-0.5000	-1.1250	0.2500	-1.0000
	Gamma = 2 ^	2.0000	0.8750	-1.5000	-1.1250	2.0000	-1.0000
	Epsilon-SVR	0.0088	0.0075	0.0030	0.0053	0.0056	0.0063
Ann(Annz)	RMS	0.378	0.374	0.379	0.374	0.372	0.372
	Hidden layer nodes	{20,20}	{5,100}	{19,10}	{20,10}	{35,5}	{24,6}
Ann (matlab)	RMS	0.390	0.387	0.390	0.382	0.388	0.388
	Hidden layer nodes	{20,12}	{7,42}	{50,20}	{8,85}	{20,20}	{12,83}

QSOC : purely synthetic data

- Simulated data :
  - Based on purely sybthetic spectra : power law continuum + emission line template (Type 1) !
  - Random (RAN) libraries of End of mission combined BP/RP spectra => Mag0 (noise free, 20x10^3), {Mag15,Mag18.5, Mag20.0} (1 noise realisation, 20x10^3).

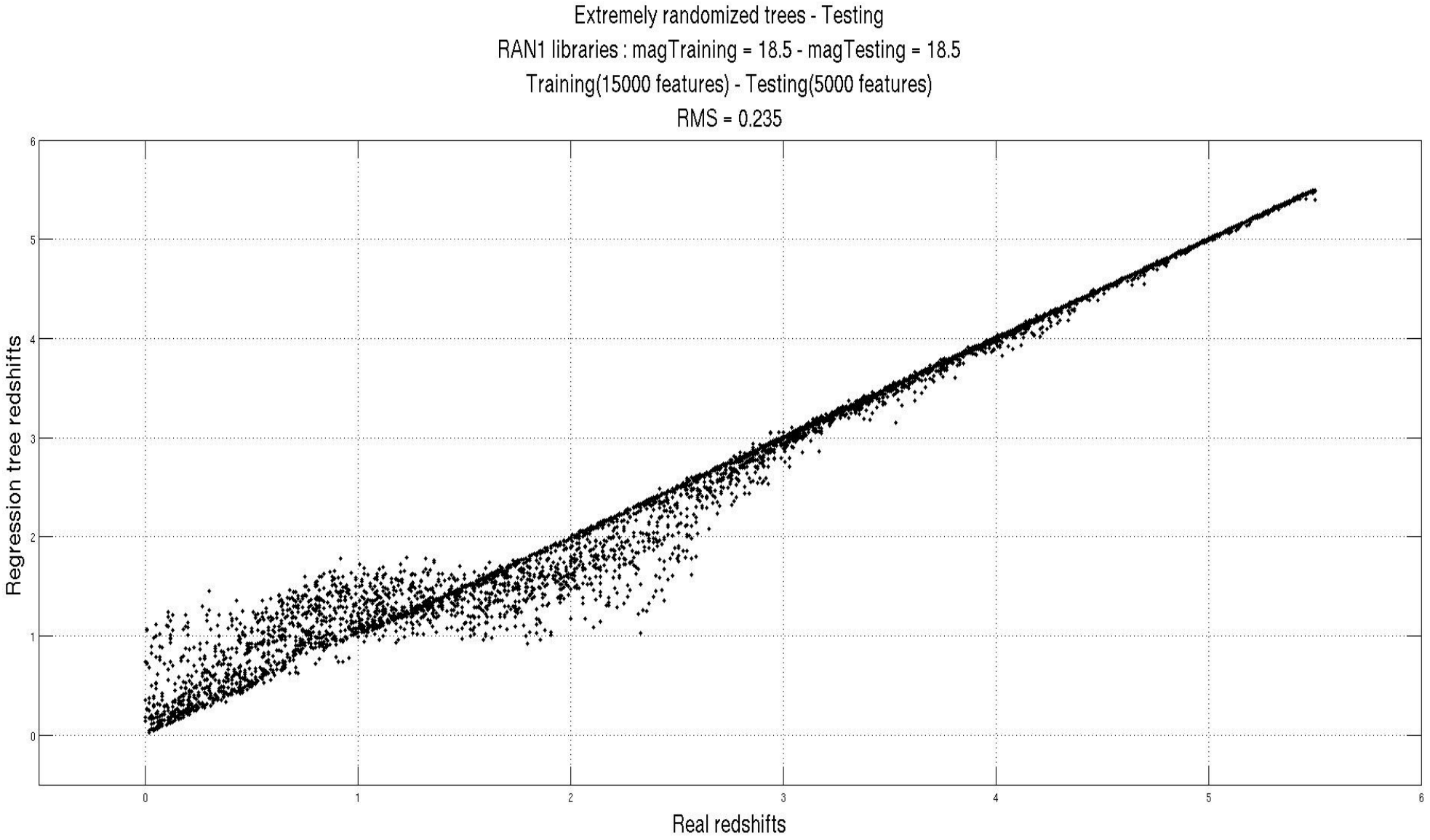
- 3 runs : each library is splited to 4 x ¼ : cross-validation using 3 x ¼ and test using ¾ – ¼
  - Run 01 : RAN library without extinction
  - Run 02 : RAN library with unknown extinction
  - Run 03 : RAN library with known extinction (A_0 is an input)

- Remove pixels for which the amplitude (over the learning library) < 3 * sigma

- Regression algorithms : Knn, ERT(1), ERT(all), Annz(1),Annz(all), RF(1) and RF(all)

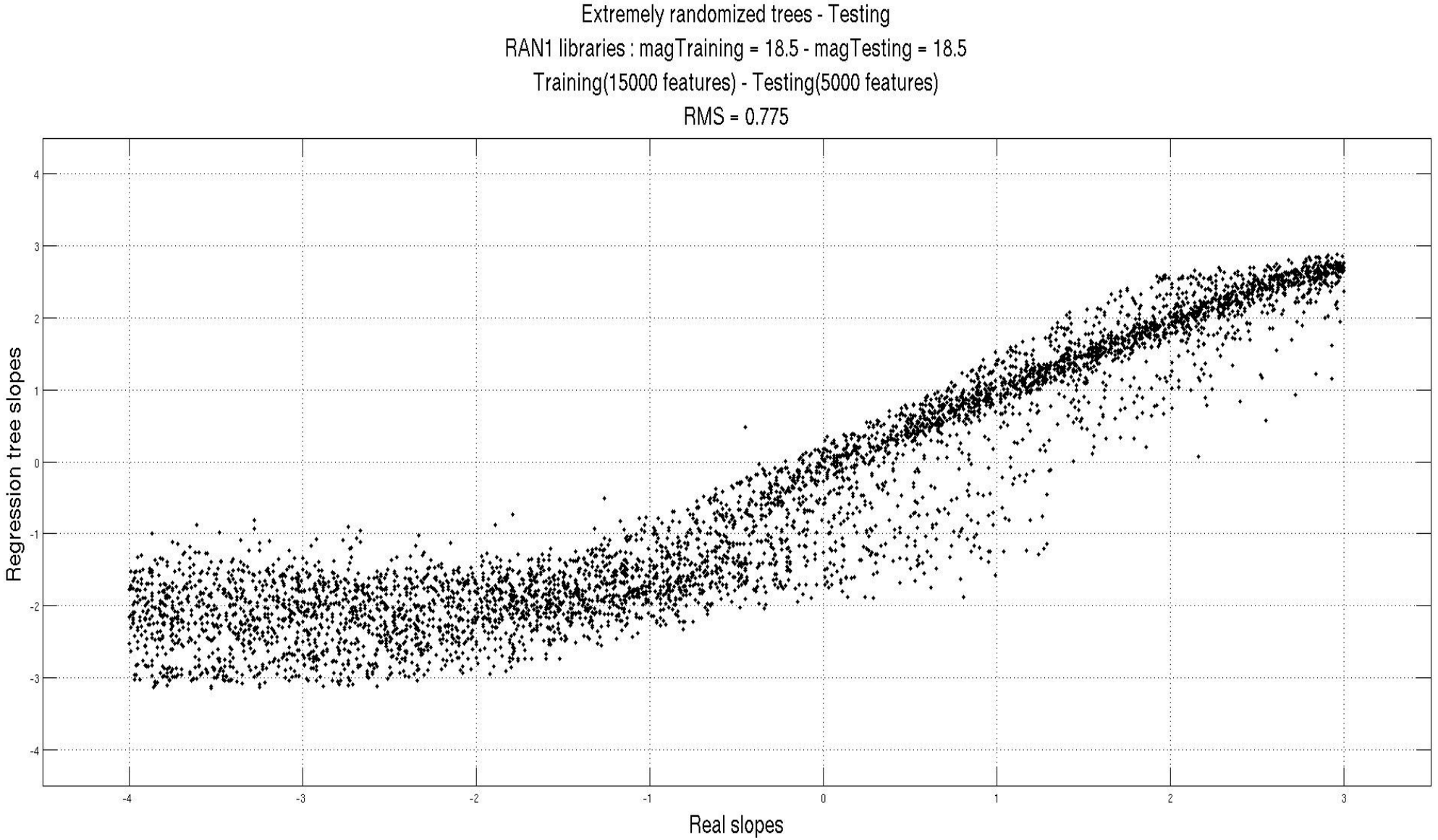
QSOC : purely synthetic data

1) End of mission (72 transits) - No extinction



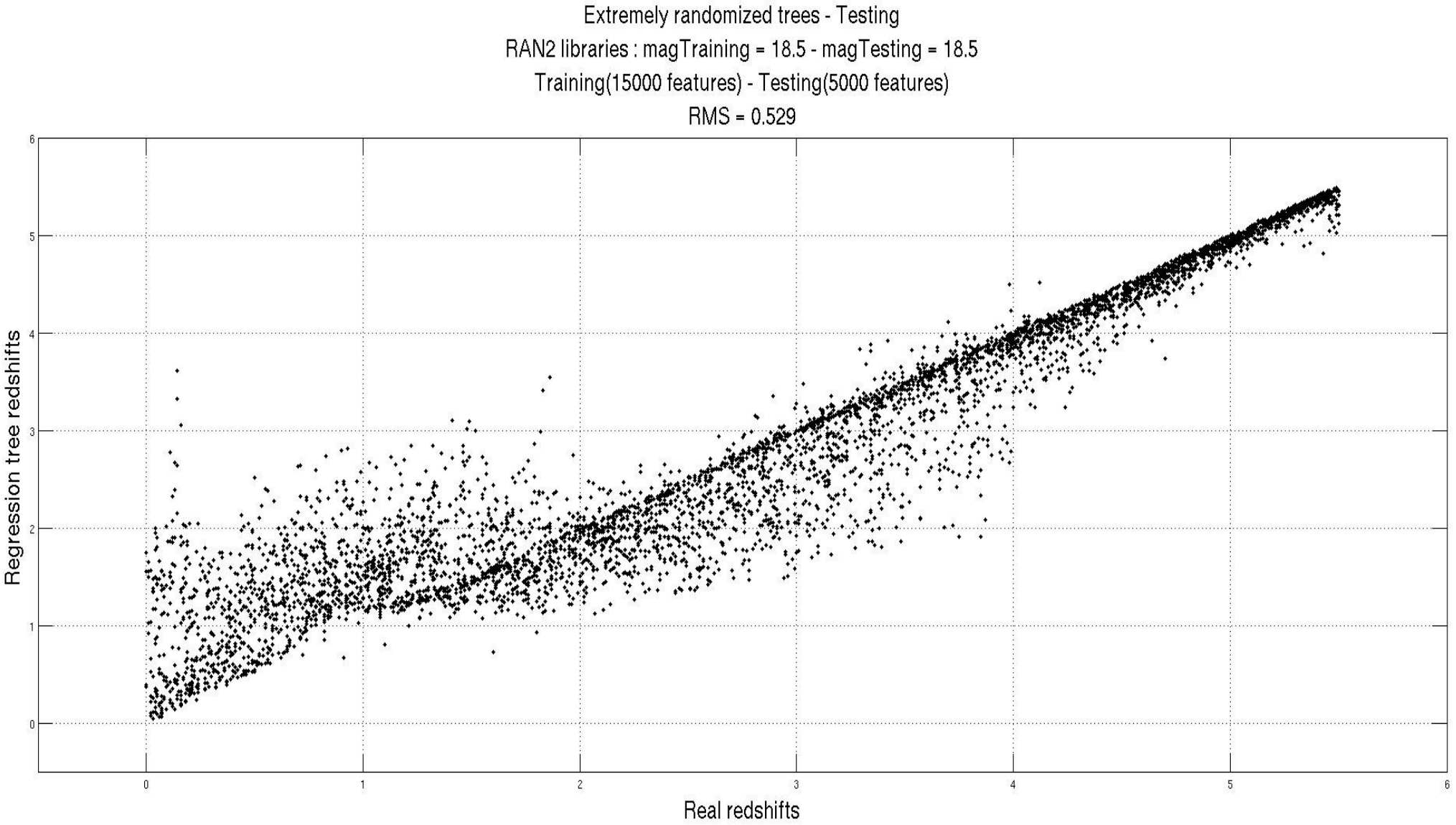
QSOC : purely synthetic data

1) End of mission - No extinction



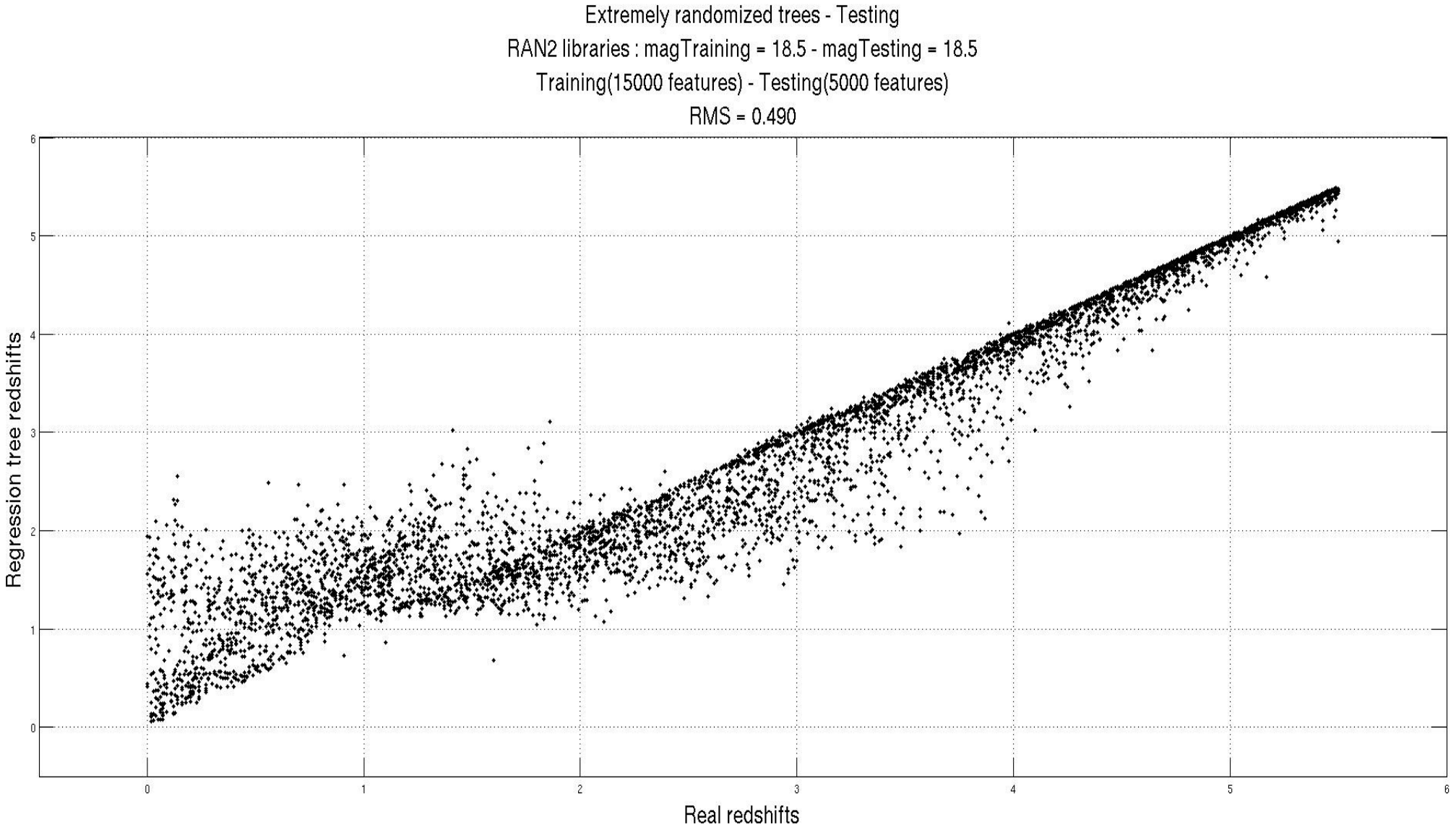
QSOC : purely synthetic data

2) End of mission (72 transits) - Unknown extinction



QSOC : purely synthetic data

3) End of mission (72 transits) - Corrected extinction



QSOC : purely synthetic data

4) Epoch spectrum - Corrected extinction