The VIMOS Public Extragalactic Redshift Survey (VIPERS)

A support vector machine classification at the service of the VIMOS Public Extragalactic Redshift Survey (VIPERS)

Małek Katarzyna[1], Solarz A.[1], Pollo A.[1,2], Fritz A.[3], Garilli B.[3], Scodeggio M.[3], Iovino A.[3], Granett B.[4], and the VIPERS Team

1st Roman Juszkiewicz Symposium, Warsaw, 2015

Małek et al. A&A, 2013, 557, A16
Małek et al. A&A, 2015 in prep.

[1] NCBJ, ul. Hoża 69, 00-681 Warszawa, POLAND,
[2] Jagiellonian University, ul. Orla 171, 30-244 Kraków, POLAND,
[3] INAF – IASF, via Bassini 15, 20133 Milano, ITALY,
[4] INAF – OA di Brera, via Brera 28, 20122 Milano, via E. Bianchi 46, 23807 Merate, ITALY

VIPERS

1. MOTIVATION
2. METHOD - machine learning algorithms,
3. DATA
   a. training sample,
4. RESULTS
   a. 3 & 5D classifier of galaxies/stars/AGNs,
   b. 6D classifier for narrow & broad line AGNs

Katarzyna.Malek@ncbj.gov.pl

1. MOTIVATION

- the amount of astronomical data collected by satellites/ground-based surveys is steadily **increasing**,

1. MOTIVATION

- the amount of astronomical data collected by satellites/ground-based surveys is steadily **increasing**,
- the zoo of collected data (photometry, redshifts, spectral lines, morphology) is constantly expanding - how **to recognise** and **extract** the most discriminating **patterns** and allow full systematisation of the data?

Katarzyna.Malek@ncbj.gov.pl

1. MOTIVATION

- the amount of astronomical data collected by satellites/ground-based surveys is steadily **increasing**,
- the zoo of collected data (photometry, redshifts, spectral lines, morphology) is constantly expanding - how **to recognise** and **extract** the most discriminating **patterns** and allow full systematisation of the data?

The CLASSIFICATION of sources is one of the basic and crucial tasks to perform before moving on to ANY scientific analysis.

Katarzyna.Malek@ncbj.gov.pl

MOTIVATION:

HOW TO PERFORM THE ULTIMATE CLASSIFICATION IN THE NEW SURVEYS?

Katarzyna.Malek@ncbj.gov.pl

1. MOTIVATION

bright sources: the distinction stars/galaxies can be made based upon **morphological measurements**; point sources → stars, while extended sources → galaxies,

fainter sources: **colour-colour diagrams** (different types of objects will appear in different colour regions in such diagrams due to the shape of the SED),

AGNs: how to classify them based on the photometric data only?

1. MOTIVATION

- AUTOMATIC CLASSIFICATION → crucial for upcoming surveys (huge amount of data),
- intelligent MACHINE LEARNING ALGORITHMS (based on multi-dimensional parameter spaces) can help us not to get lost in the zoo of data, and PRESELECT sources for more sophisticated scientific analysis.

Katarzyna.Malek@ncbj.gov.pl

METHOD:

MACHINE LEARNING ALGORITHMS

Katarzyna.Malek@ncbj.gov.pl

2. METHOD

The machine learning algorithms are designed to infer patterns in the data:

1. **unsupervised algorithms** identify groups in the data a priori,
2. **supervised algorithms** are trained to recognize the pattern (e.g. Support Vector Machines *SVM*).

Support Vector Machine (SVM)

Katarzyna.Malek@ncbj.gov.pl

2. METHOD/SVM-main concept

- to **calculate decision planes** between a set of objects having different class memberships, which are defined by the TRAINING SAMPLE → quantities that describe the properties of each class of objects,

Katarzyna.Malek@ncbj.gov.pl

2. METHOD/SVM-main concept

- to **calculate decision planes** between a set of objects having different class memberships, which are defined by the TRAINING SAMPLE,
- SVM searches for the optimal separating **hyperplane** between the different classes of objects by maximizing the margin between the classes' closest points; to find boundaries one needs to choose the bunduary method (in our case **C-SVM**, with C and gamma parameters), and a **kernel function** (Gaussian radial in this work).

2. METHOD/SVM-main concept

- to **calculate decision planes** between a set of objects having different class memberships, which are defined by the TRAINING SAMPLE,
- SVM searches for the optimal separating **hyperplane** between the different classes of objects,
- the objects are classified based on their relative position in the **N-dimensional parameter space** to the separation boundary.

Katarzyna.Malek@ncbj.gov.pl

2. METHOD/SVM-practical point of view

- manually classify the TRAINING SAMPLE,

AGNs

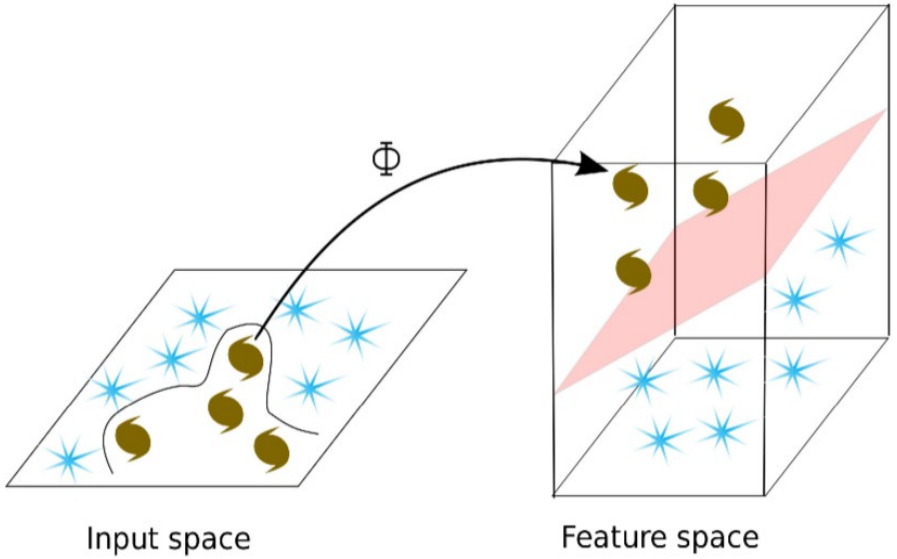
Stars

Galaxies

Katarzyna.Malek@ncbj.gov.pl

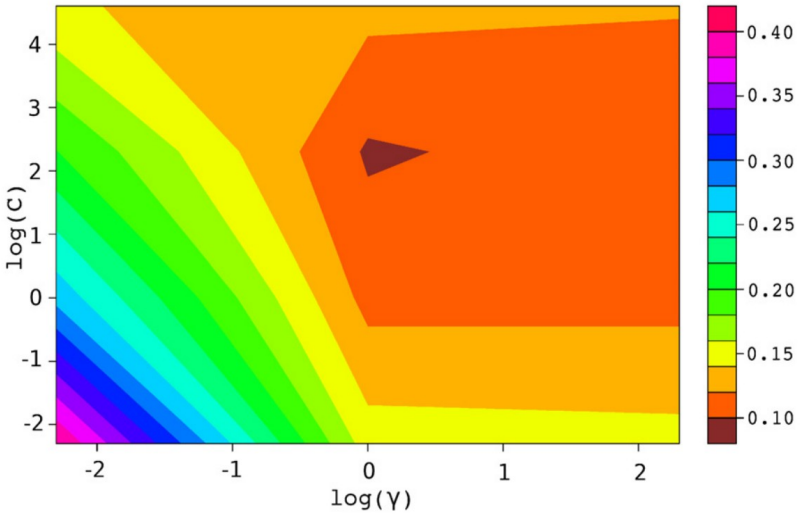
2. METHOD/SVM-practical point of view

- manually classify the TRAINING SAMPLE,
- for each object in this subset define a feature vector (in or case photometric data),



2. METHOD/SVM-practical point of view

- manually classify the TRAINING SAMPLE,
- for each object in this subset define a feature vector (in or case photometric data),
- train algorithm and optimize all parameters.



Katarzyna.Malek@ncbj.gov.pl

2. METHOD/SVM-practical point of view

- manually classify the TRAINING SAMPLE,
- for each object in this subset define a feature vector (in or case photometric data),
- train algorithm and optimize all paramaters,
- automatically classify the data

FINAL CLASSIFICATION

DATA:

PHOTOMETRIC DATA + SPECTROSCOPIC CLASSES (USED TO BUILD OUR TRAINING SAMPLE)

Katarzyna.Malek@ncbj.gov.pl

3. DATA

3.1. Photometric data

a) **CFHTLS (u, g, r, i, z)** photometry (a joint Canadian-French programme), a subsample of CFHTLS T0005 catalogue with spectroscopic redshift measured by VIPERS.

b) **WIRCam** data: near infrared **Ks** measurements taken from Wide-field InfraRed Camera, coming from the dedicated follow-up observations for the VIPERS project (Arnouts et al., in prep.),

Magnitudes used to develop classifier are in the AB photometry system and were corrected for foreground Galactic extinction,

3. DATA

3.1. Photometric data

a) CFHTLS (u, g, r, i, z) ,

b) WIRCam (Ks),

3.2. Spectroscopic data

c) **PDR 1 VIPERS survey:** 55 358 redshifts (galaxies, stars, AGNs)

d) **VIMOS-VLT Deep Survey (VVDS):** VVDS-Deep (F02 field) and VVDS-Wide (F22 field) surveys for a training sample of AGNs (objects classified as AGNs by Gavignaud et al. 2007)*.

*homogeneous photometric information (u,g,r,i,z,Ks), similar sample selection, similar flags

Katarzyna.Malek@ncbj.gov.pl

3. DATA

The VIMOS Public Extragalactic Redshift Survey (VIPERS)

- an ongoing ESO Large Programme aimed at measuring spectroscopic redshifts for ~ 10^5 galaxies and covering in total ~ 24 deg^2 on the sky,
- redshift range $0.5 < z < 1.2$,
- the galaxy target sample selected from optical photometric catalogs CFHTLS to the limit of $i_{AB} < 22.5$

  a simple and robust color pre-selection in (g – r) vs (r – i) applied to remove stars (but not all), and galaxies at $z < 0.5$.

Katarzyna.Malek@ncbj.gov.pl

TRAINING SAMPLE:

a set of objects with confirmed classes which will serve as a template for distinguishing the sources whose class we want to determine.

4. TRAINING SAMPLE

- **GALAXIES** with the highest confidence level of redshift measurements (more than 16 000),
- **AGNs** given the small number of AGNs detected in the VIPERS fields with the highest confidence of redshift measurements (398 objects) we have also merged the VIPERS sample with objects classified as broad-line AGNs in the VVDS survey (100 objects),
- **STARS** stars observed by VIPERS are interlopers within the galaxy and AGN samples and are thus not representative of the stellar class; to build an unbiased star training sample we added stars from the VVDS Wide F22 overlap with the VIPERS W4 field (2 232 stars).

Katarzyna.Malek@ncbj.gov.pl

4. TRAINING SAMPLE

Table 2. Number (N) of galaxies, AGNs, and stars in our training sample after using the oversampling method.

	$19 \leqslant i' < 20$	$20 \leqslant i' < 21$	$21 \leqslant i' < 22$	$22 \leqslant i' < 22.5$
N galaxies	1884	5483	6778	2126
N AGNs	1520	4440	5440	1760
N stars	2232	4440	5 440	2232

Katarzyna.Malek@ncbj.gov.pl

4. TRAINING SAMPLE

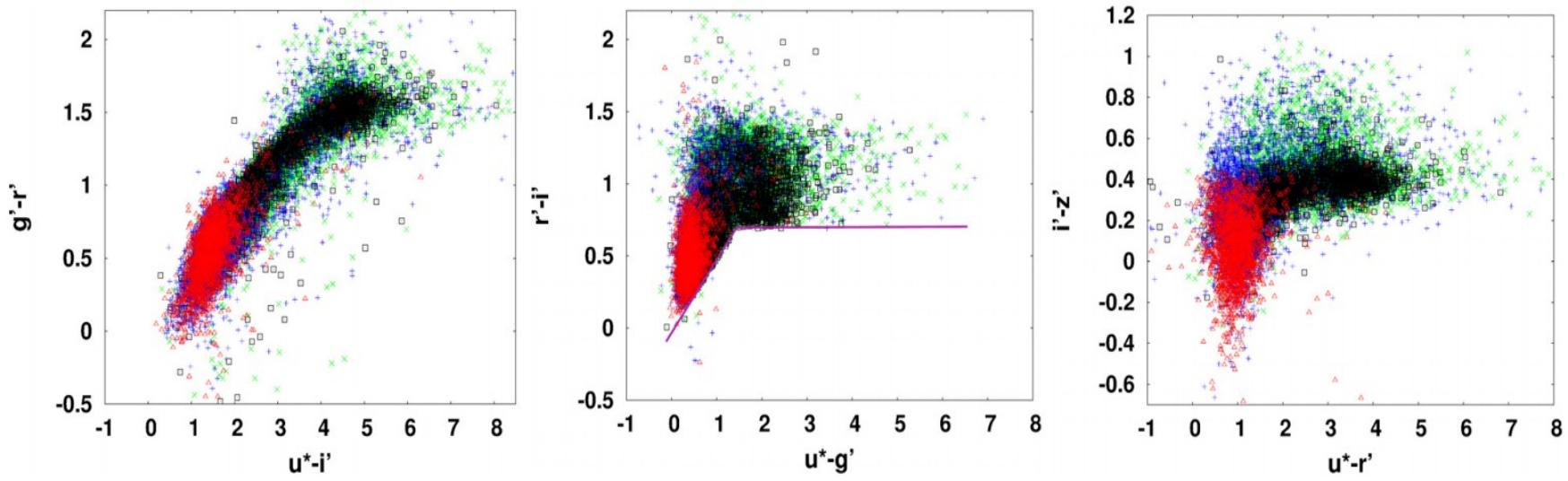


Fig. 3. Representative colour–colour plots for the galaxy training sample. Open black squares represent objects with i'-apparent magnitude between 19 and 20 mag; green X-s – galaxies with i' magnitude between 20 and 21 mag; objects with i' apparent magnitude between $21 \leqslant i < 22$, and $22 \leqslant i < 22.5$ mag are marked as blue +-s and open red triangles, respectively; in the *middle panel* of colour–colour plots, the boundaries of VIPERS selection are marked as magenta lines.

4. TRAINING SAMPLE

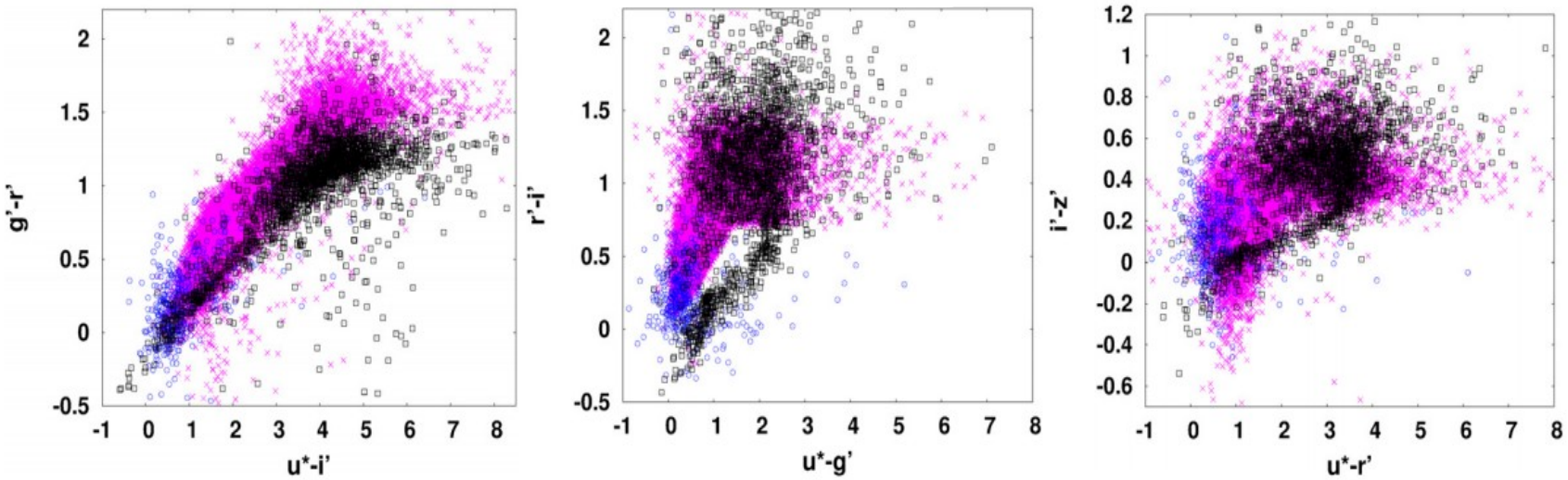


Fig. 7. Representative colour–colour plots for all objects used for the training sample. Pink x-s represent galaxies. Open blue circles correspond to the AGN sample, and open black squares to the stellar sample.

5. RESULTS

TWO **GALAXY/AGN/STAR** CLASSIFIERS:

- 3D (pure OPTICAL) **(u-g), (g-r) & (r-i)**

Katarzyna.Malek@ncbj.gov.pl

5. RESULTS

TWO **GALAXY/AGN/STAR** CLASSIFIERS:

- 3D (pure OPTICAL) **(u-g), (g-r) & (r-i)**
- 5D (OPTICAL+NIR) **(u-g), (g-r), (r-i),** (i-z) & (z-Ks)

Katarzyna.Malek@ncbj.gov.pl

5. RESULTS

TWO **GALAXY/AGN/STAR** CLASSIFIERS:

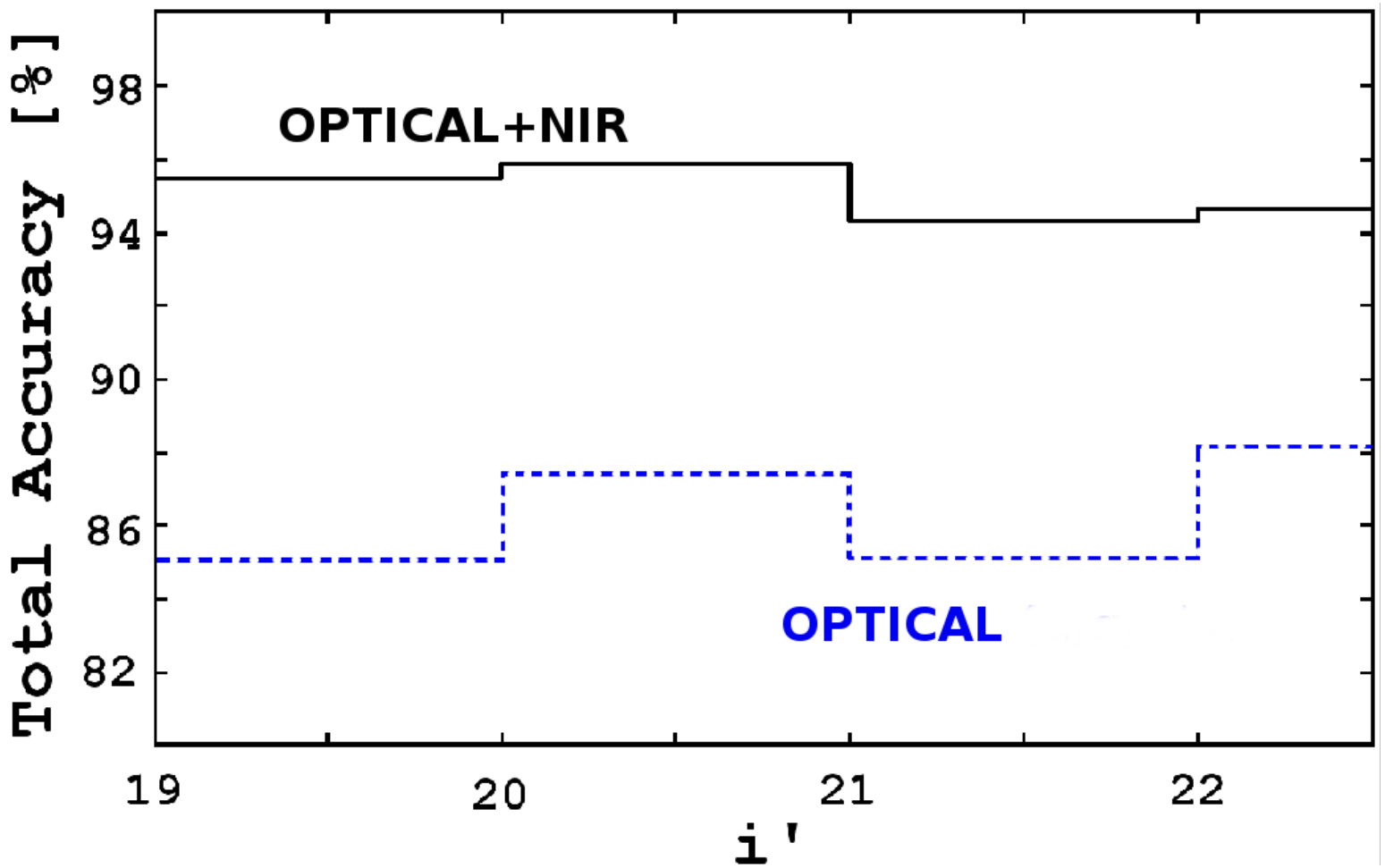
- 3D (pure OPTICAL) (u-g), (g-r) & (r-i)
- 5D (OPTICAL+NIR) (u-g), (g-r), (r-i), (i-z) & (z-Ks)

$$Accuracy = \frac{TG + TAGN + TS}{TG + TAGN + TS + FG + FAGN + FS}$$

$$Total\ Accuracy = \frac{\sum_{N}^{i=1} Accuracy_i}{N}$$

where N = 10 is the number of validation iterations

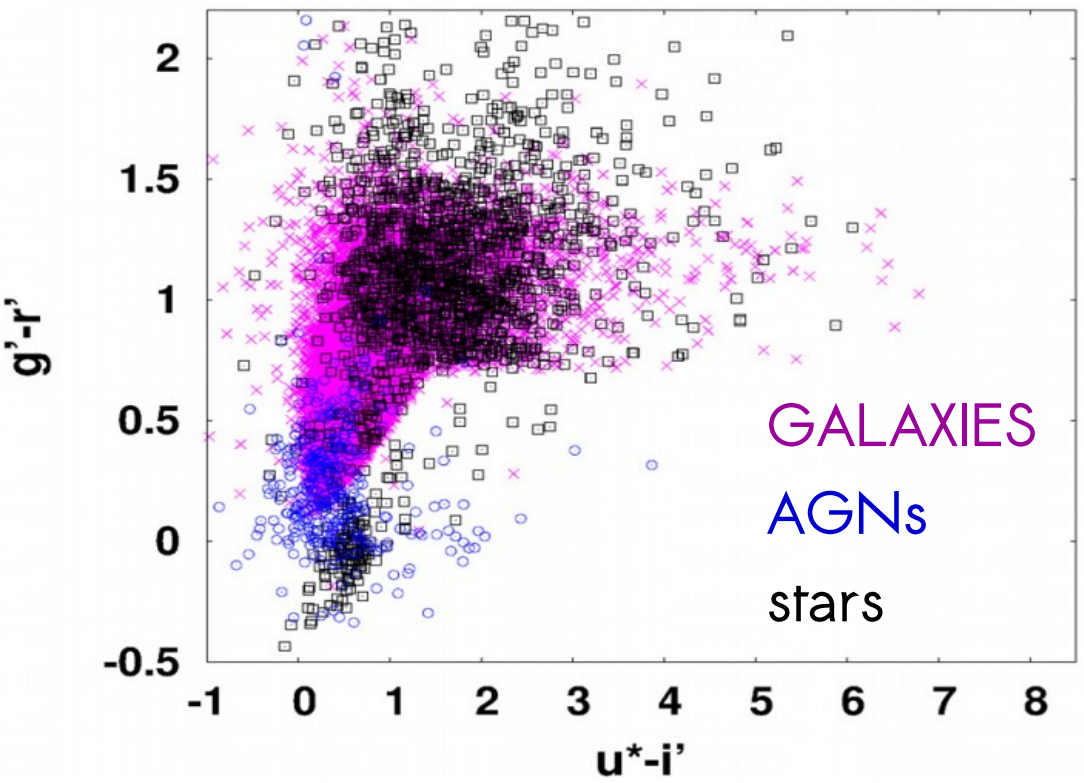
5. RESULTS



Katarzyna.Malek@ncbj.gov.pl

5. RESULTS

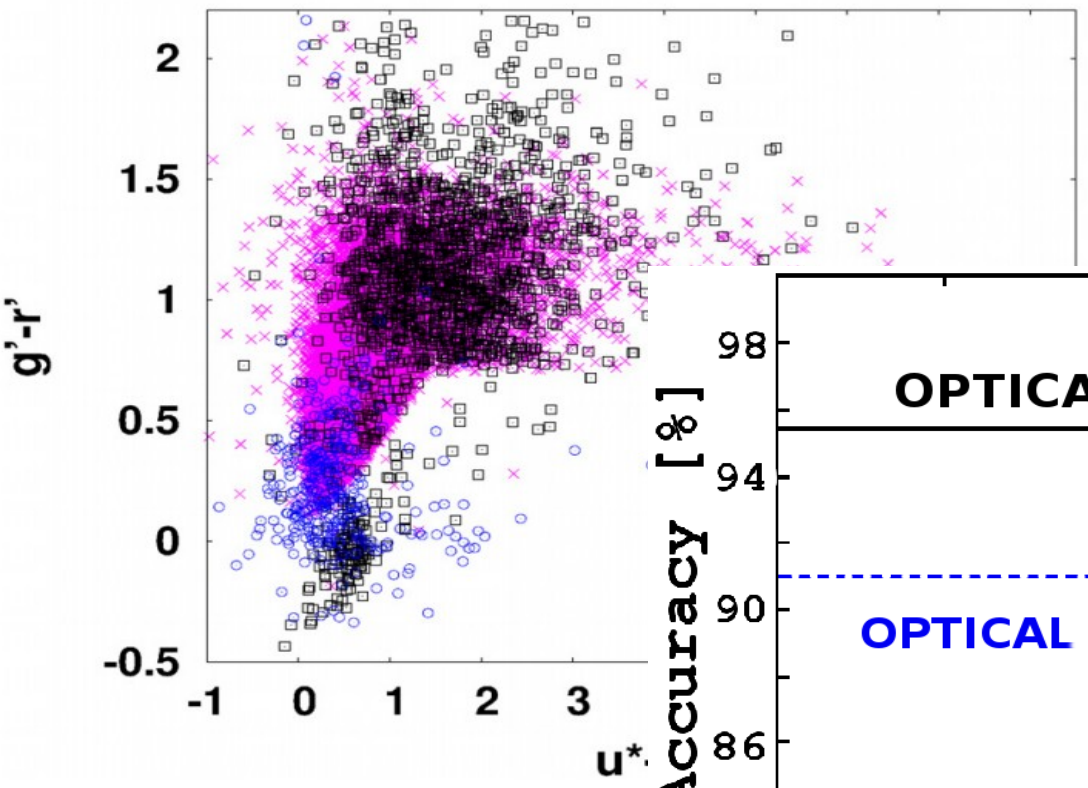
Objects with high redshift confidence not used as a training sample,



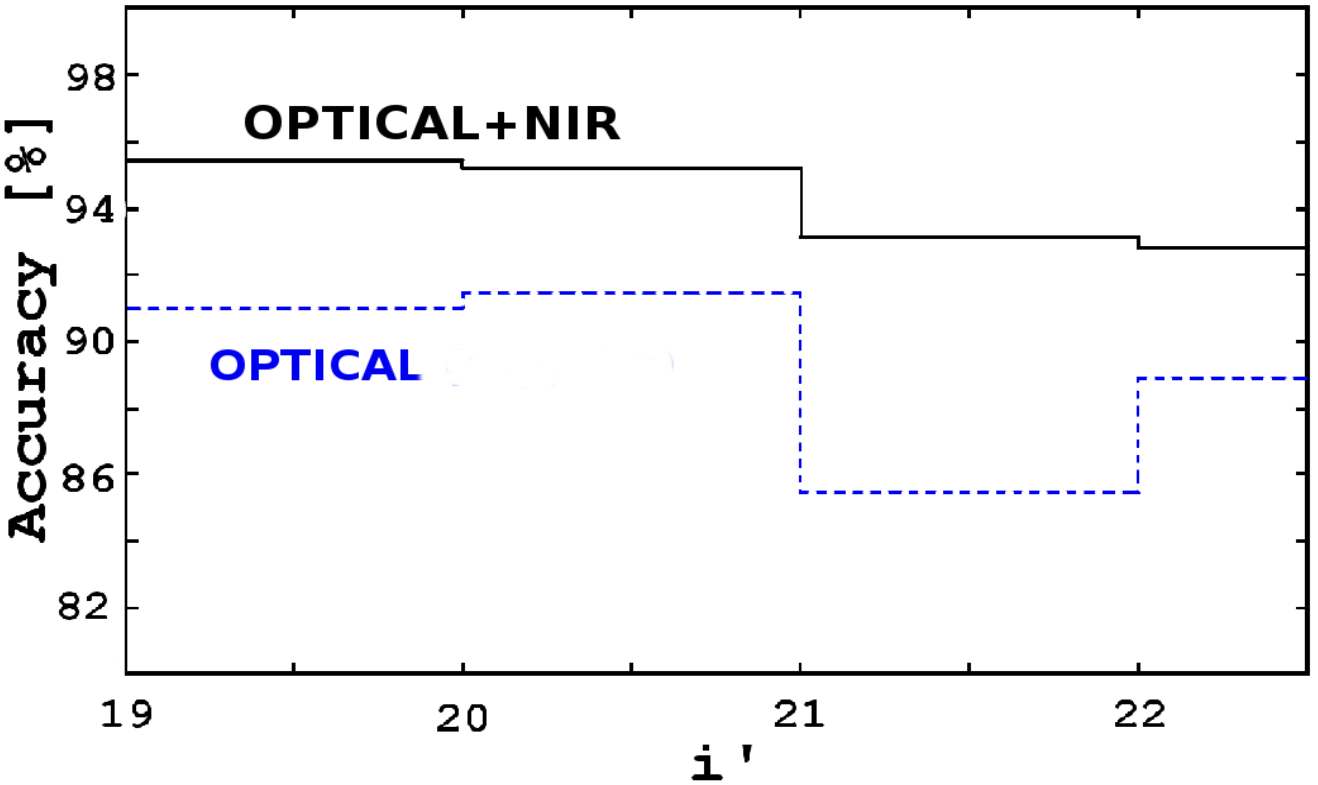
Katarzyna.Malek@ncbj.gov.pl

5. RESULTS

Objects with high redshift confidence not used as a training sample,



GALAXIES ONLY



Katarzyna.Malek@ncbj.gov.pl

5. RESULTS

Table 7. Results of the test of the optical+NIR classifier for GAL_3, and AGNs and stars with redshifts measurements on a confirmation level ≥ to 99%.

	$19 \leqslant i' < 20$			$20 \leqslant i' < 21$			$21 \leqslant i' < 22$			$22 \leqslant i' < 22.5$		
SVM/true	Galaxy	AGN	Star	Galaxy	AGN	Star	Galaxy	AGN	Star	Galaxy	AGN	Star
Number of sources	445	69	337	3 271	1340	428	7 667	127	701	2 156	37	263
Galaxy	**95.38**	12.52	4.17	**95.17**	7.37	3.27	**93.09**	10.46	3.42	**92.72**	14.90	9.09
AGN	2.42	**77.34**	3.70	2.72	**82.08**	3.27	4.30	**82.75**	1.43	5.29	**75.54**	6.44
Star	2.20	10.14	**92.13**	2.11	10.55	**93.46**	2.61	6.79	**95.15**	1.99	9.56	**84.47**

Notes. Values marked in bold correspond to the correctly classified objects (galaxies, AGNs, and stars) in i'-based apparent magnitude bins. The ratio of the classified objects is given in percentage.

galaxies are misclassified as AGNs in the faintest magnitude bins

1) the decrease in the quality of the photometry for the less luminous sources (lower SN ratio),

2) some of them hosting faint AGNs (not recognised during the visual verification/validation), since with the decreasing luminosity the host galaxy becomes dimmer and the AGN component becomes more significant.

5. RESULTS

5D classifier: very good agreement with the VIPERS spectroscopic sample with confidence level of z measurements equal to 95%:

classification accuracy:

- 94% for  ,
- 93% for ★,
- 82% for AGNs.

By classifying the sources with low-quality spectra, we can improve the classification and enlarge the samples that may be used for analysis. Using the optical+NIR classifier, we confirmed the class of **4900** objects with low flags.

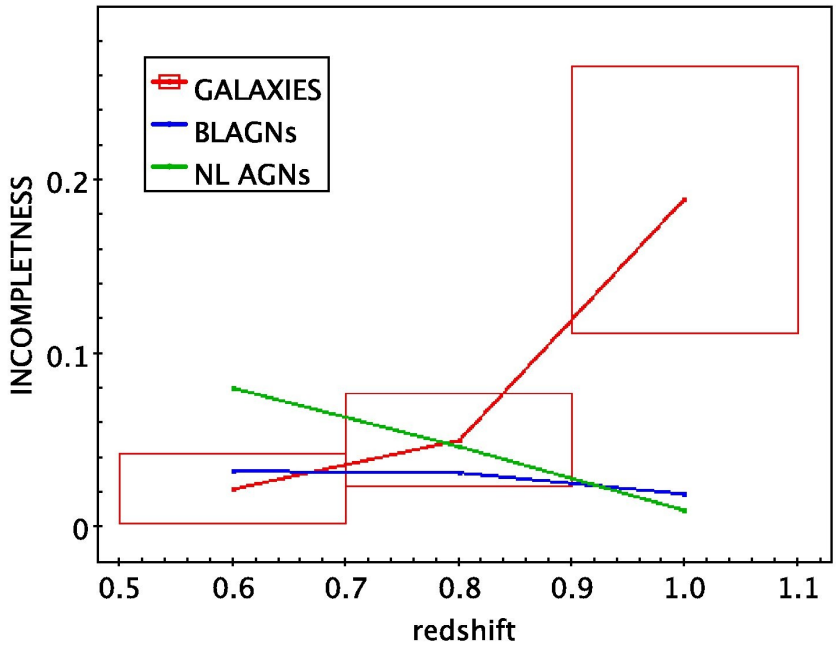
Katarzyna.Malek@ncbj.gov.pl

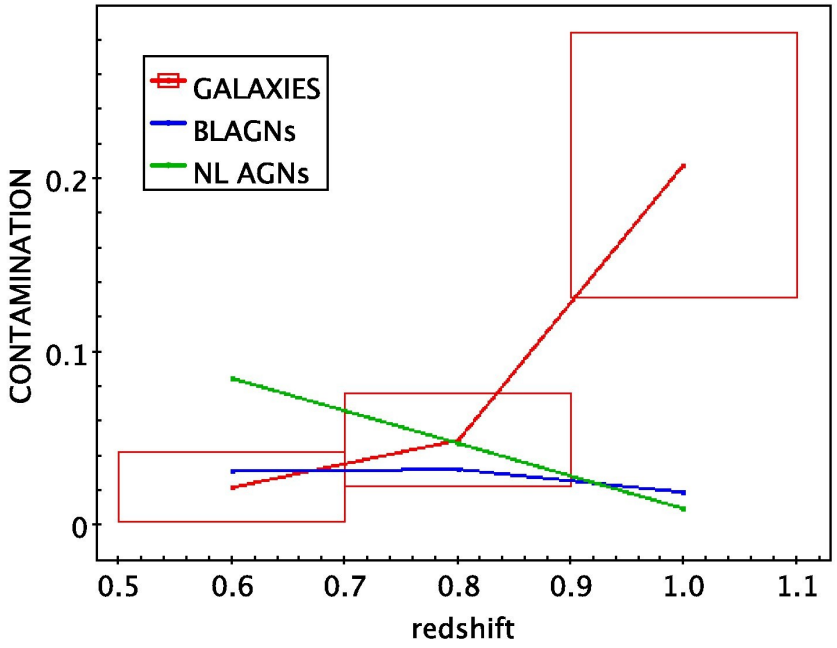
5. RESULTS

- new classfier to classify **narrow line AGNs** which are not visible at VIPERS survey (main emission line are out of the range for VIPERS sample),
- new Training Sample (galaxies, BLAGNs ← from VIPERS + NLAGNs from VVDS and zCOSMOS,
- new parameter space: 6D classifier, photometric information only: **(u-g), (g-r), (r-i), (i-z), (z-Ks) & Sersic parameter**

Katarzyna.Malek@ncbj.gov.pl

5. RESULTS





- total efficiencies for the classifier are very high: more than **96%** for 20-22.5 bins, **92%** for the most luminous bin,
- it means that we can use **purely photometric sample** what is very promising for the **future surveys**.

Katarzyna.Malek@ncbj.gov.pl

CONCLUSIONS:

- **an excellent classification** for galaxies, AGNs and stars, and the very promising results of the galaxy-NLAGN-BLAGN classifier,
- great tool for a new surveys (designe & selection, classifications, ..),
- help for a final validation of the data.

Katarzyna.Malek@ncbj.gov.pl

Thank you for your attention

Katarzyna.Malek@ncbj.gov.pl