Learning algorithms at the service of WISE survey

Katarzyna Małek[1,2,3], T. Krakowski[1], M. Bilicki[4,3], A. Pollo[1,5,3], A. Solarz[2,3], M. Krupa[5,3], A. Kurcz[5,3], W. Hellwing[6,3], J. Peacock[7], T. Jarrett[4]

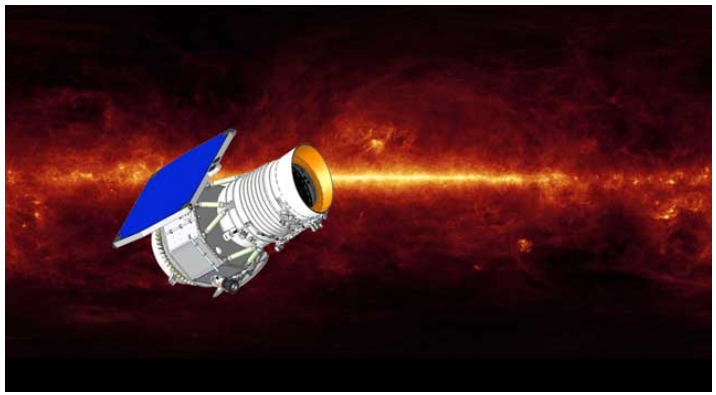
[1] National Centre for Nuclear Research, ul. Hoża 69, 00-681 Warszawa, Poland
[2]Nagoya University, Furo-cho, Chikusa-ku, 464-8602 Nagoya, Japan
[3]Kepler Institute of Astronomy, University of Zielona Góra, ul. Lubuska 2, 65-265 Zielona Góra, Poland
[4]Department of Astronomy, University of Cape Town, Rondebosch, South Africa
[5]The Astronomical Observatory of the Jagiellonian University, ul. Orla 171, 30-244 Kraków, Poland
[6]Institute for Computational Cosmology, Department of Physics, Durham University, Durham DH1 3LE, U.K
[7]Institute of Astronomy, University of Edinburgh, Royal Observatory, Edinburgh EH9 3HJ, UK

WISE at 5, 2015

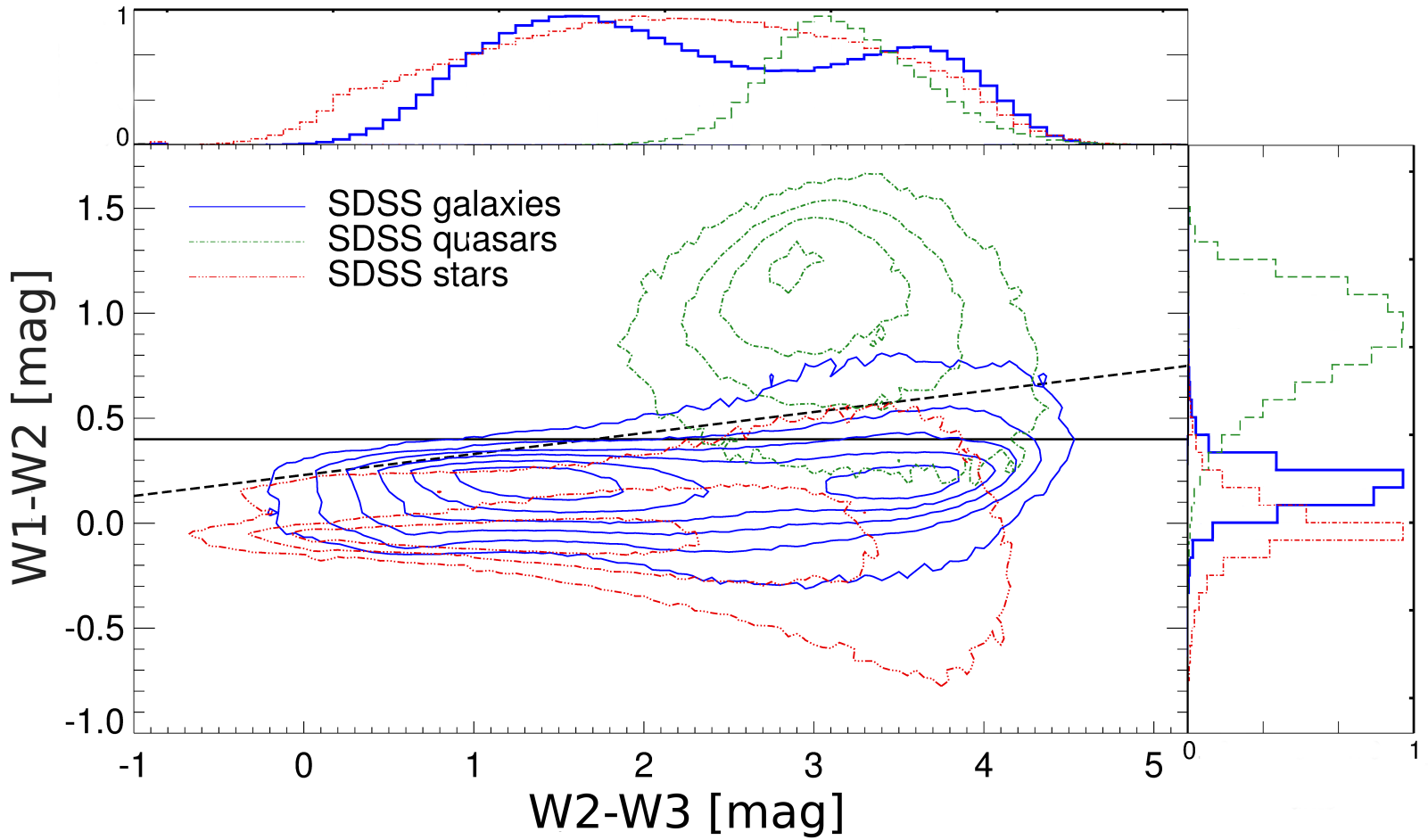
How to perform the ultimate classification of the WISE data?

- **W**ide-field **I**nfrared **S**urvey **E**xplorer - the biggest Infrared all sky survey,
- over 1 million images covering the whole sky in 4 infrared wavelengths,
- map of the sky, PSC
- good resolution of the data,
- great opportunity to study different types of galaxies, quasars; huge impact for cosmology.

No sufficient auxiliary data to classify all these sources.



PROBLEM TO SOLVE



WISExSDSS DR10 $\sim$ 1 700 000 objects

AIM

- to develop an automatic object classification in WISE into galaxy/quasar/star categories, with high completeness and purity, based on photometric properties of the sources;
- we focused on obtaining clean catalogs of stars, galaxies, and quasars,
- keeping at the same time **the biggest possible sample** of WISE sources to take advantage of the all-sky information

IDEA

1. match two all-sky catalogs: AllWISE and SuperCOSMOS;

IDEA

1. match two all-sky catalogs: AllWISE and SuperCOSMOS;
2. use spectroscopic data (SDSS DR 10) to create calibration sets of three source types (galaxies, stars, and quasars);

IDEA

1. match two all-sky catalogs: AllWISE and SuperCOSMOS;
2. use spectroscopic data (SDSS DR 10) to create calibration sets of three source types (galaxies, stars, and quasars);
3. use the calibration data as training sets for a machine learning algorithm to perform the final classification of the WISE x SuperCosmos data;

IDEA

1. match two all-sky catalogs: AllWISE and SuperCOSMOS;
2. use spectroscopic data (SDSS DR 10) to create calibration sets of three source types (galaxies, stars, and quasars);
3. use the calibration data as training sets for a machine learning algorithm to perform the final classification of the WISE x SuperCosmos data;
4. as a result we can obtain classification of 170 milion of WISE sources, with known purity and contamination distribution.

DATA

The second all-WISE dataset (Cutri et al. 2013) availabe to download from `http://irsa.ipac.caltech.edu/frontpage/`, combining data from the cryo-genic and post-cryogenic survey phases.

- almost 750 million sources with S/N ≥ 5 in at least one of the bands,
- averaged 95% completeness in unconfused areas is W1: 17.1, W2: 15.7, W3: 11.5 and W4: 7.7 in Vega magnitudes,

The SuperCOSMOS Sky Survey (hereafter **SCOS**, Hambly et al.2001 a,b,c):

- digitized photographs in three bands (B, R, I), obtained via automated scanning of source plates from
  - South: the United Kingdom Schmidt Telescope (UKST), and
  - North: the Palomar Observatory Sky Survey-II (POSS-II).
- calibrated using:
  - SDSS photometry in the relevant areas,
  - 2MASS J band over the rest of the sky.
- publicly available from the SuperCOSMOS Science Archive (`http://surveys.roe.ac.uk/ssa/`),
- contains **1.9 bilion of sources** (photometry, morphology and quality)

SAMPLE SELECTION

WISE:

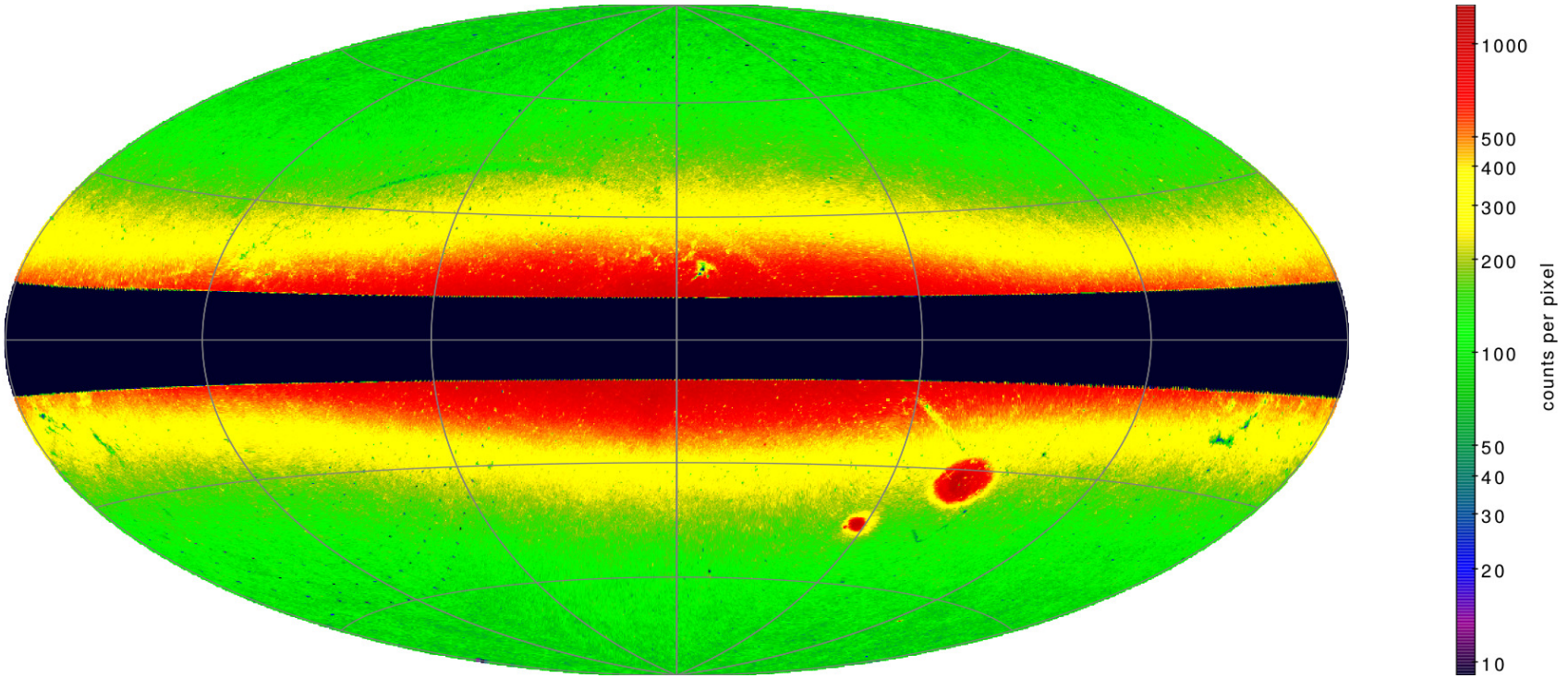
- S/N ratios larger than 2 in the W1 and W2 bands,
- removal of obvious artefacts (cc_flags[1,2] $\neq$ `DPHO`) $\Longrightarrow$ 603 million detection over the whole sky,
- rejection of the Galactic Plane ($|b| < 10^{\circ}$) $\Longrightarrow$ 460 million of objects (83% of the sky available at $|b| > 10^{\circ}$),
- $W1 < 17$ magnitude threshold for all-sky uniformity $\Longrightarrow$ **343 million sources** at $|b| > 10^{\circ}$.

SCOS:

1. required detections in both the B and R bands,
2. flux cuts for uniformity: B $<$ 21 mag and R $<$ 19.5 mag,
3. galactic latitude cut ($|b| > 10^{\circ}$), where blending and high extinction make the SCOS photometry unreliable.

FINAL SAMPLE: WISExSuperCOSMOS catalog

170 milion of sources



MACHINE LEARNING ALGORITHMS

The machine learning algorithms

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

1. **Unsupervised** algorithms identify groups in the data *a priori*;
   - clustering algorithms;
2. **Supervised** algorithms are trained to recognize the pattern.
   - Support vector Machines (SVM)

The machine learning algorithms

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

1. **Unsupervised** algorithms identify groups in the data *a priori*;
   - clustering algorithms;
2. **Supervised** algorithms are trained to recognize the pattern.
   - **Support vector Machines (SVM)**

SVM - the main concept

- to calculate **decision planes** between a set of objects having different class memberships, which are defined by the **Training Sample**,

SVM - the main concept

- to calculate **decision planes** between a set of objects having different class memberships, which are defined by the **Training Sample** $\Rightarrow$ **quantities that describe the properties of each class of objects**,

SVM - the 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,

SVM - the 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.

SVM

- to search for a hyperplane, SVM uses kernel function[2], and
- a soft-boundary SVM method called `C-SVM`:
  - C - trade-off parameter between large margin of different classes of objects and mis-classifications.
  - γ parameter determines the topology of the decision surface.

Both parameters, C and γ, need to be **tuned** based on the **Training Sample**.

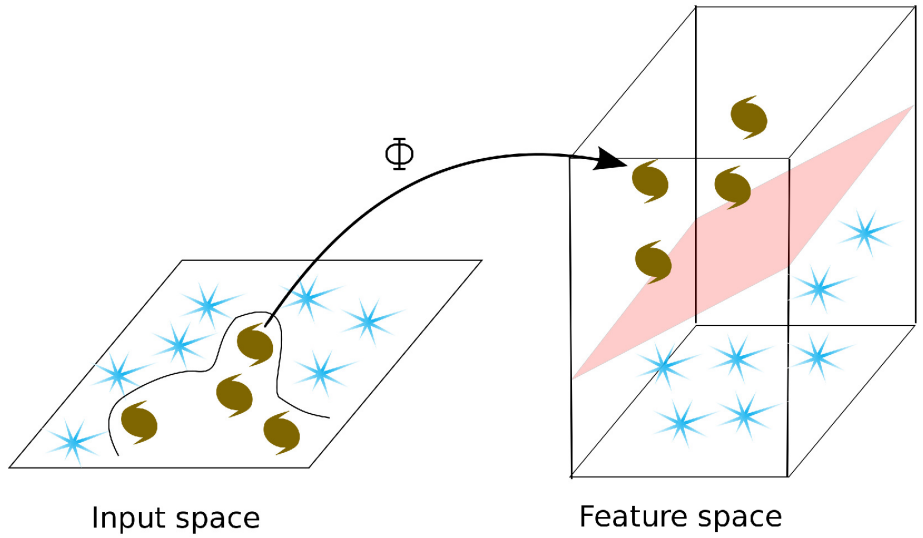
[2] Gaussian radial basis kernel (RBK) function for this work.

SVM: practical point of view

1. manually classify the Training Sample,

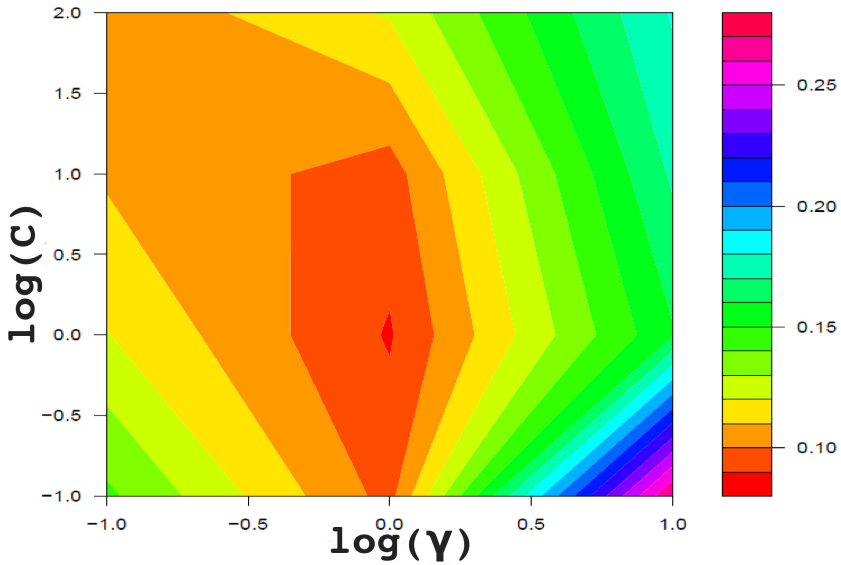
SVM: practical point of view

1. manually classify the Training Sample,
2. for each object in this subset define a feature vector,



SVM: practical point of view

1. manually classify the Training Sample,
2. for each object in this subset define a feature vector,
3. Train algorithm and optimize C and γ.



Training Sample

Training Sample - the heart of the method

- The WISExSCOS catalog includes z$<$ 0.5 galaxies (Bilicki et al. 2014; Bilicki et al. 2015), as well as stars and higher-redshift quasars,

Training Sample - the heart of the method

- The WISExSCOS catalog includes z$<$ 0.5 galaxies (Bilicki et al. 2014; Bilicki et al. 2015), as well as stars and higher-redshift quasars,
- $\Rightarrow$ the training sets require good-quality and high-reliability pre-classification using spectroscopic measurements,

Training Sample - the heart of the method

- The WISExSCOS catalog includes z$<$ 0.5 galaxies (Bilicki et al. 2014; Bilicki et al. 2015), as well as stars and higher-redshift quasars,
- $\Rightarrow$ the training sets require good-quality and high-reliability pre-classification using spectroscopic measurements,
- $\Rightarrow$ the best choice: spectroscopic sample from the **Sloan Digital Sky Survey DR10** (SDDS DR10, Ahn et al. 2014) cross-matched with WISExSCOS catalog.

Training sample: WISExSCOSxSDSS

SDSS DR10 $\Rightarrow \sim$ 3.4 million spectroscopic sources: 59% - galaxies, 26% - stars, and 15% - quasars.

Pairing up these sources with our WISExSCOS flux-limited catalogue within 1' to obtain the **Trainig Sample** gave over **1.3 million common objects**, of which:

- 68% were galaxies,
- 25% stars, and
- 7% of quasars.

Training sample: WISExSCOSxSDSS

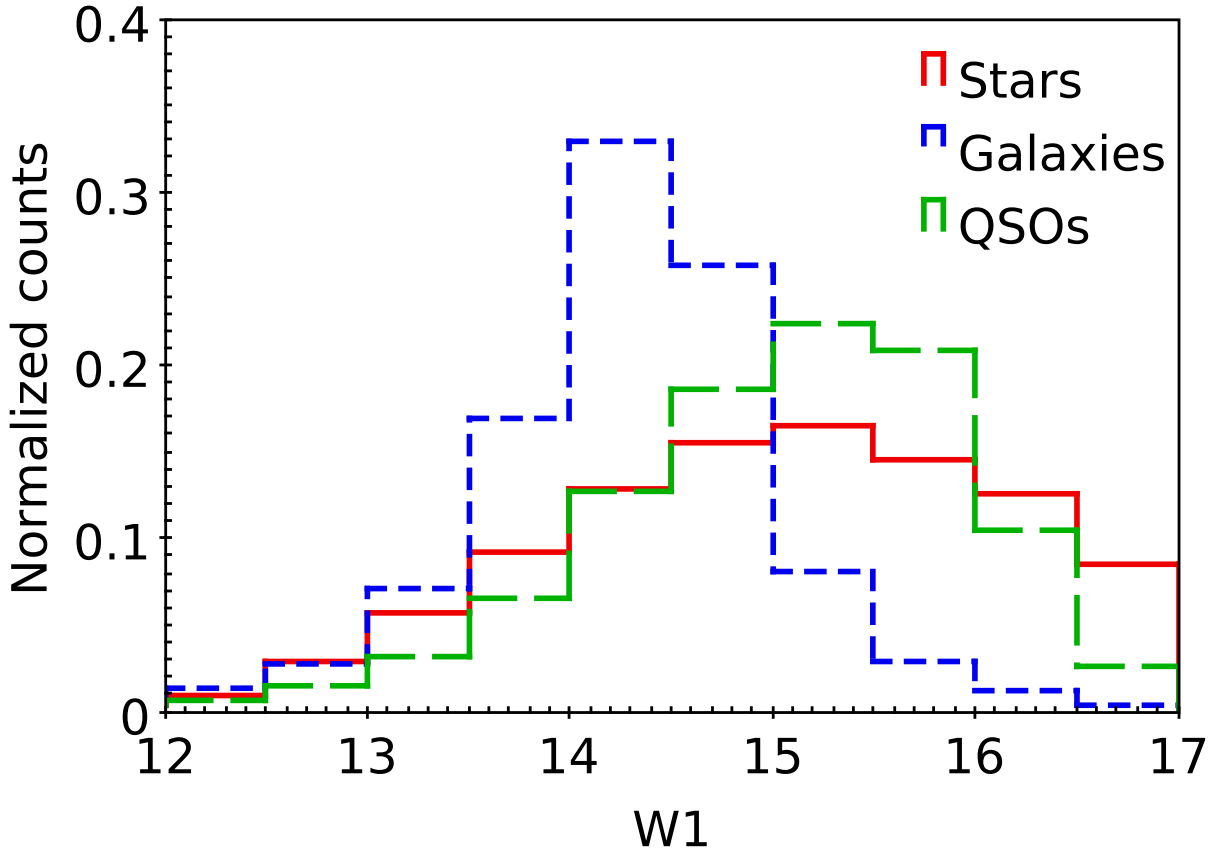


Figure 1 : Original SDSSxWISExSCOS cross-matched data.

Training sample: WISExSCOSxSDSS

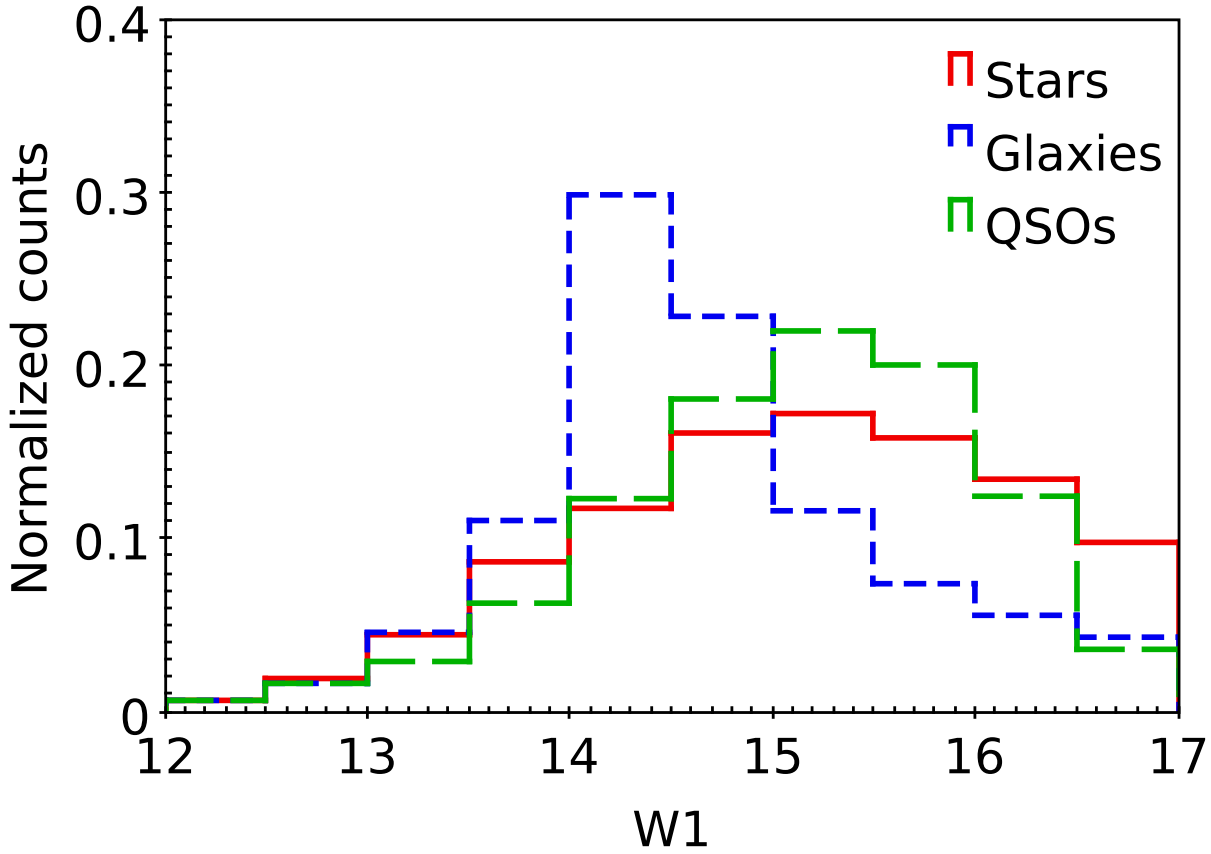


Figure 1 : Training set after oversampling.

Training sample: WISExSCOSxSDSS

We divided the training sample into 5 bins:

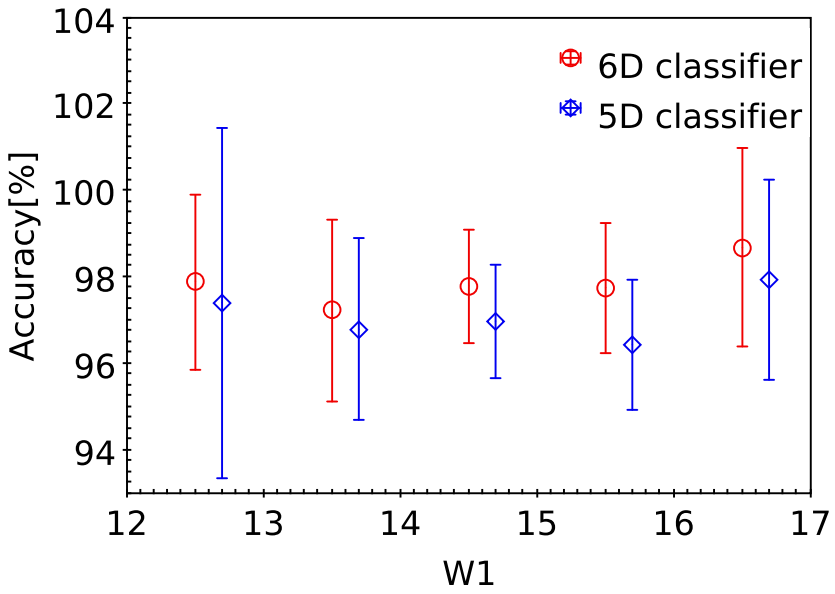
- W1 $<$ 13 mag,
- 13 mag $\leq$ W1 $<$ 14 mag,
- 14 mag $\leq$ W1 $<$ 15 mag,
- 15 mag $\leq$ W1 $<$ 16 mag, and
- 16 mag $\leq$ W1 $<$ 17 mag.

For the training set we select 5000 galaxies, 5000 quasars, and 5000 stars. The rest of the WISExSCOSxSDSS catalog was used as independent test sets to check the stability of the clasisfier,

As the parameter space we choose 5 parameters:

`W1` magnitude, `W1-W2` colour, `R-W1` colour, `B-R` colour and the `w1mag13` differential aperture of the W1 magnitude.

digression: 5D vs 6D classifier (+ W3)



The accuracy, completeness, and purity for both classifiers are very high, and the contamination levels hardly exceed 5%. The differences between the 5D and 6D classifiers are not significant. We decided not to use the $W3$ passband, as the most of our sources have only $W3$ flux upper limits estimated.

Parameters which can help us to understand the performance of the classifier

Parameters which can help us to understand the performance of the classifier

Total Accuracy:

$$TA = \frac{1}{10}\sum_{i=1}^{10} A_i \tag{1}$$

The accuracy for a given iteration is defined as

$$A_i = \frac{\mathrm{TG} + \mathrm{TQ} + \mathrm{TS}}{\mathrm{TG} + \mathrm{TQ} + \mathrm{TS} + \mathrm{FG} + \mathrm{FQ} + \mathrm{FS}}. \tag{2}$$

where:
TG - true galaxies, TQ - quasars and Ts - stars from the training sample, properly classified as galaxies, quasars, and stars, respectively; and FG - false galaxies, being real quasars or stars misclassified as galaxies, with false quasars (FQ) and false stars (FS) defined in a similar manner.

Parameters which can help us to understand the performance of the classifier

completeness (c):

$$c_g = \frac{\mathrm{TG}}{\mathrm{TG} + \mathrm{FGS} + \mathrm{FGQ}}, \tag{3}$$

purity (p, 1-contamination):

$$p_g = 1 - \frac{\mathrm{FSG} + \mathrm{FQG}}{\mathrm{TG} + \mathrm{FSG} + \mathrm{FQG}}, \tag{4}$$

where:
FGS and FGQ determine galaxies mis-classified respectively as stars and quasars, and FSG, FQG define stars and quasars mis-classified as galaxies. Definitions for stars and quasars follow in an analogous way.

RESULTS

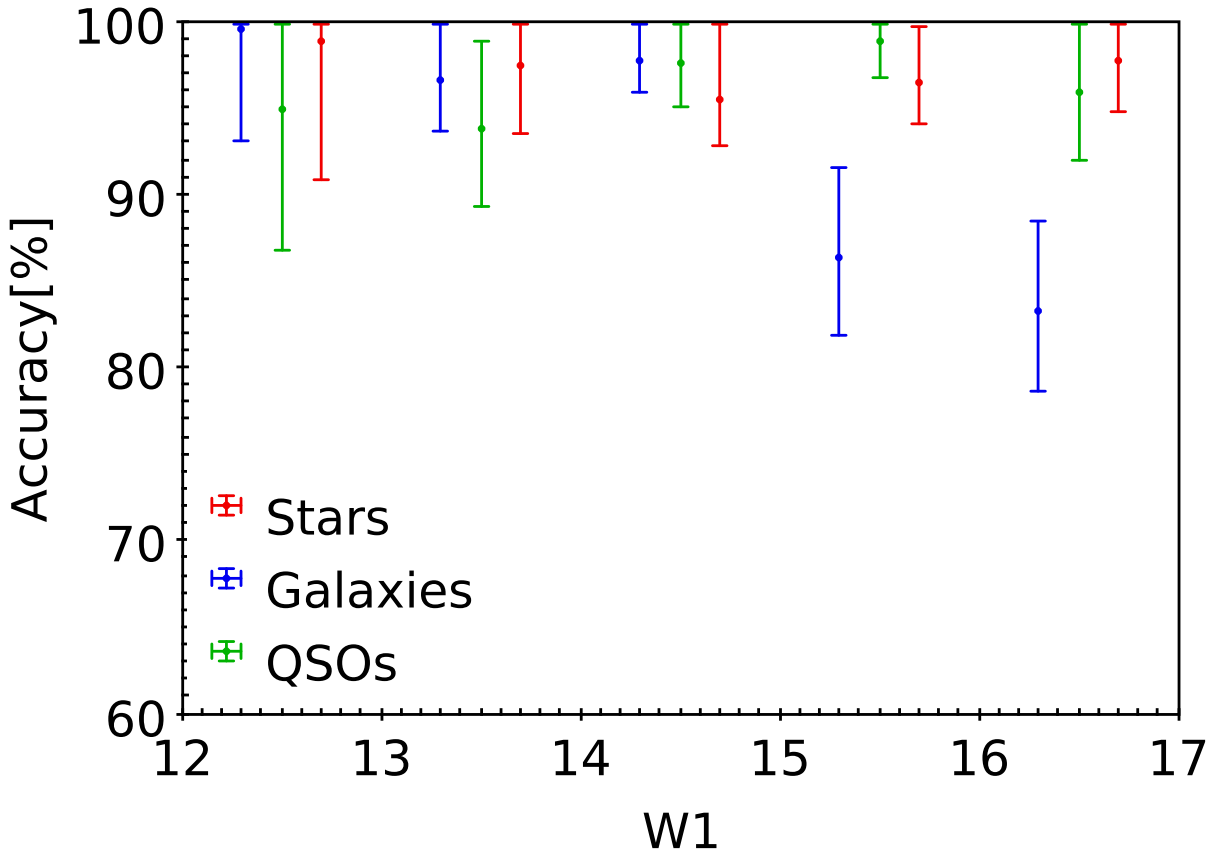


Figure 2 : Accuracy of the classifier vs W1

▸ TS distribution

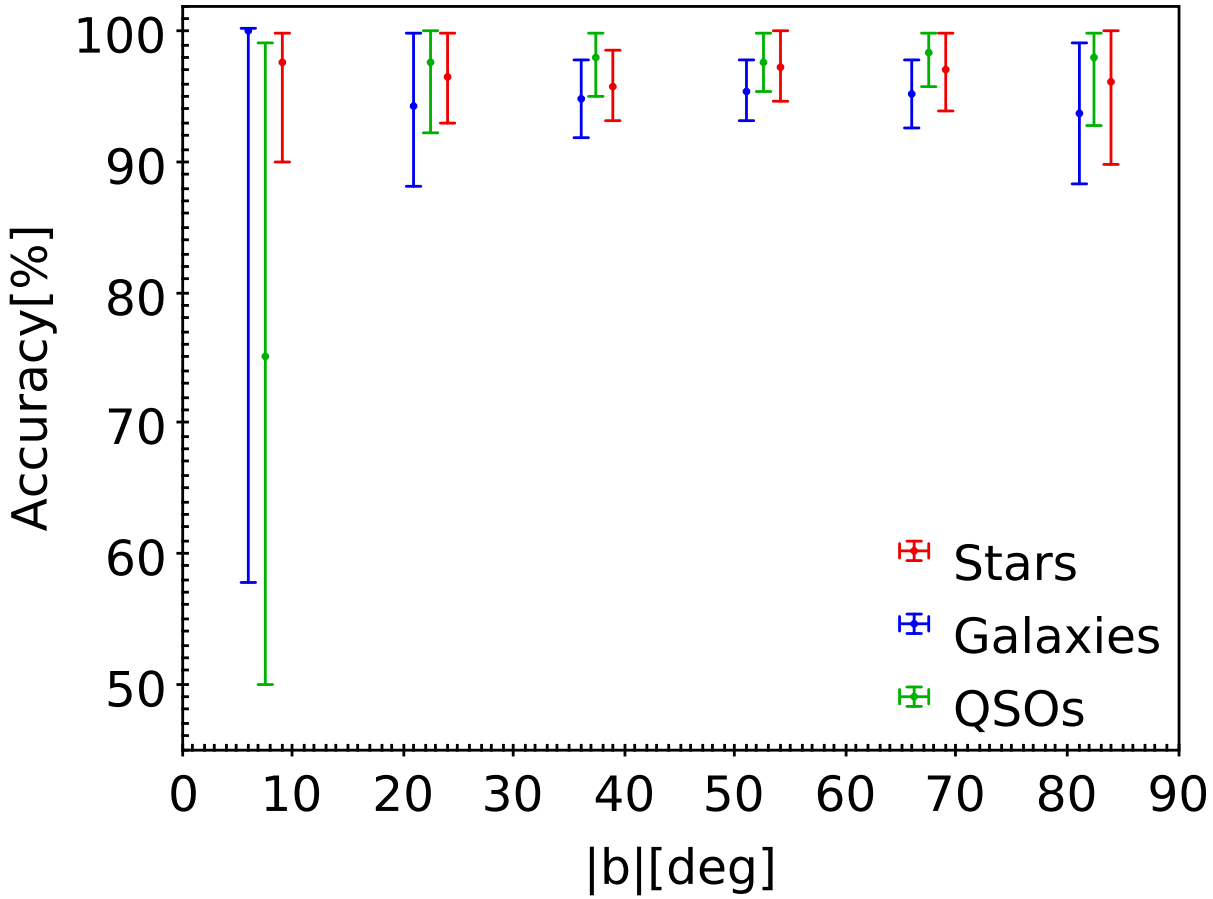


Figure 2 : Accuracy of the classifier vs $|b|$

▸ TS distribution

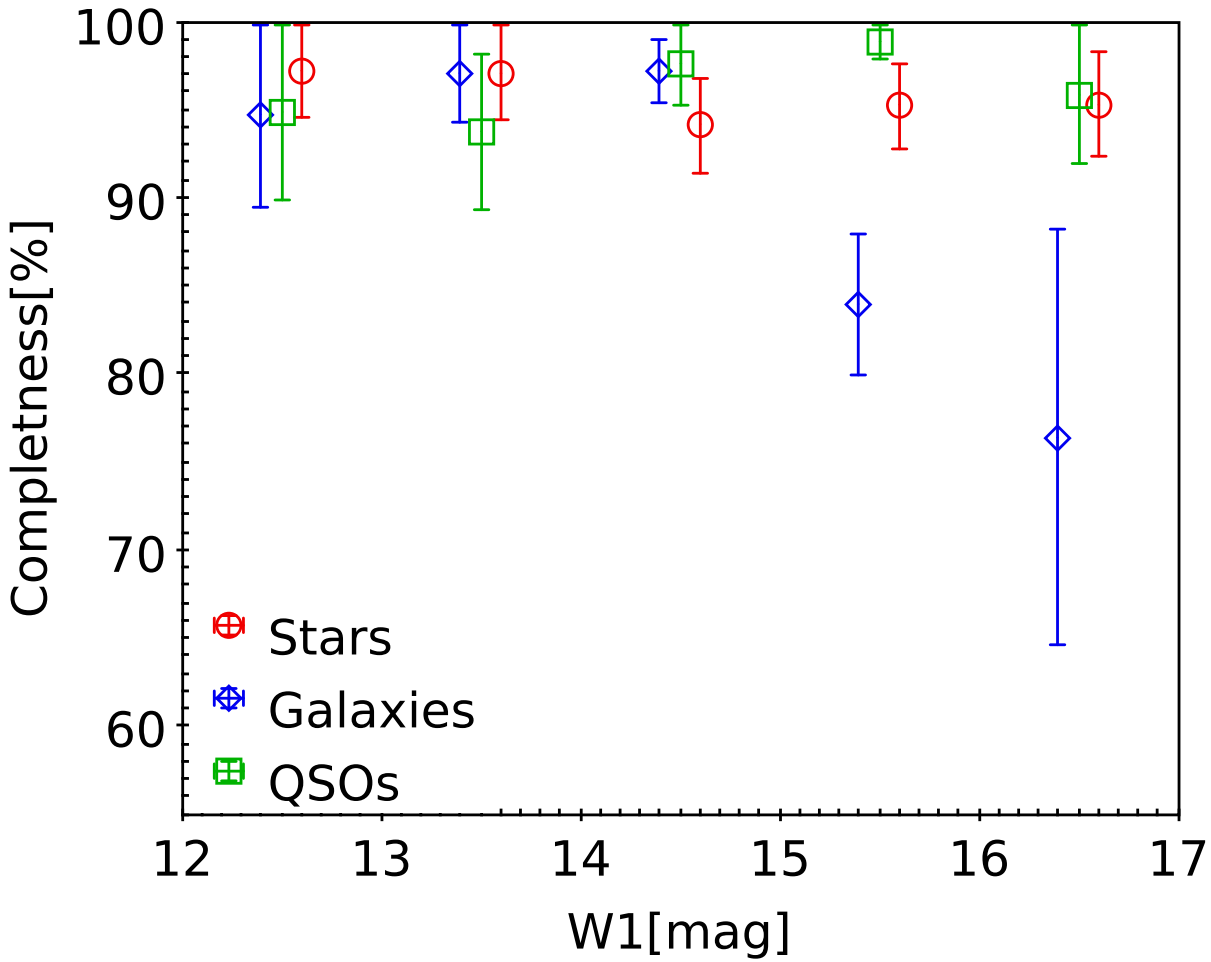


Figure 3 : Completness of the WISExSCOSxSDSS sample vs W1 mag

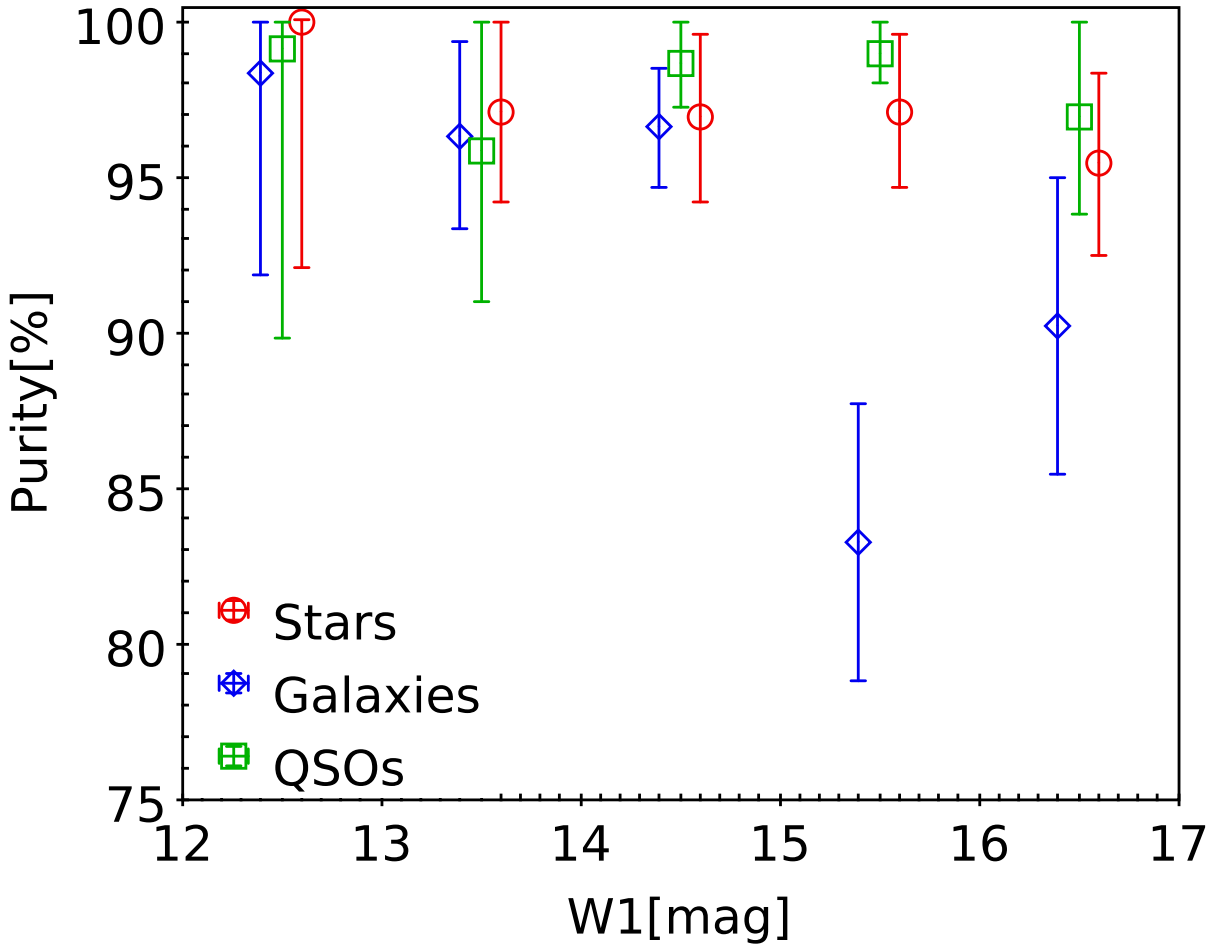


Figure 3 : Purity of the WISExSCOSxSDSS sample vs W1 mag

PRELIMINARY RESULTS: FOUR WISExSuperCOSMOS SUBSAMPLES

PRELIMINARY RESULTS FOR THE FINAL CATALOG

North Galactic Pole

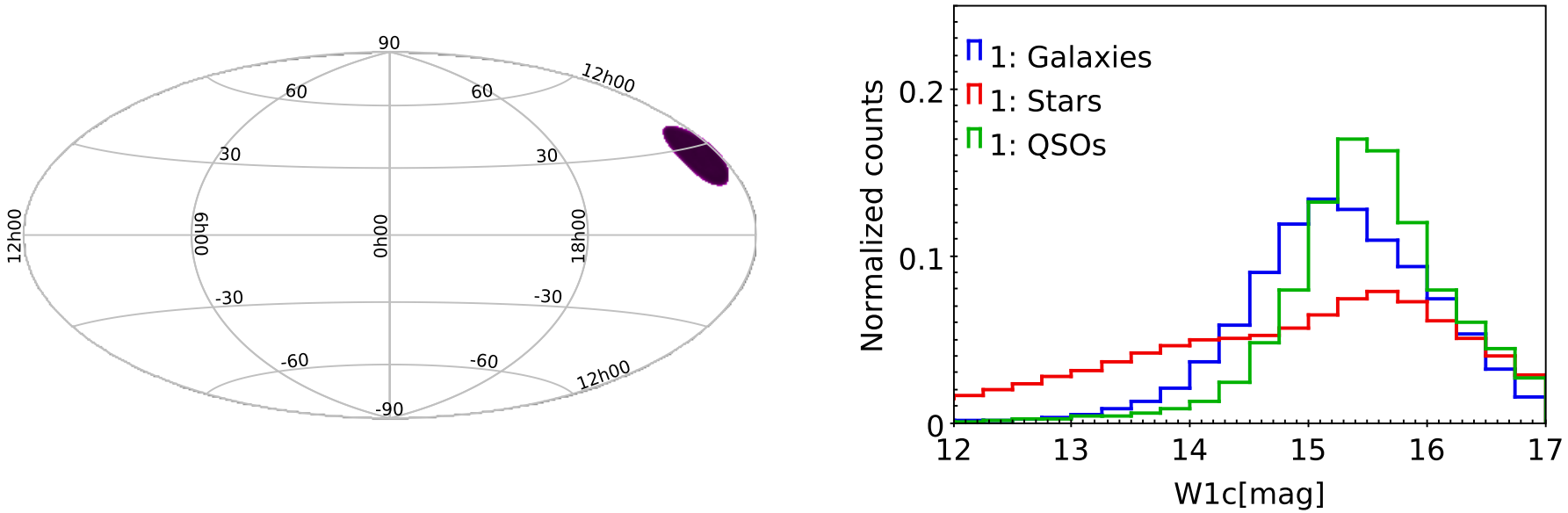
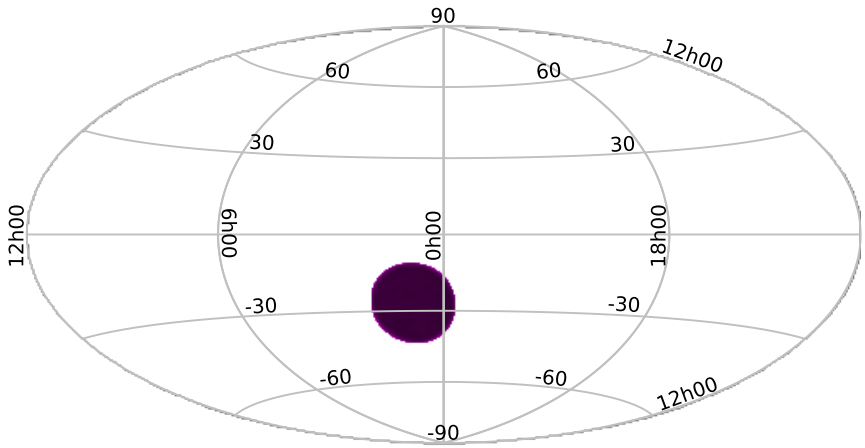


Figure 4 : **504 144 objects:** stars: 286 537 (56.84%), galaxies: 185 622 (36.82%), quasars: 31 985 (6.34%)

PRELIMINARY RESULTS FOR THE FINAL CATALOG

South Galactic Pole



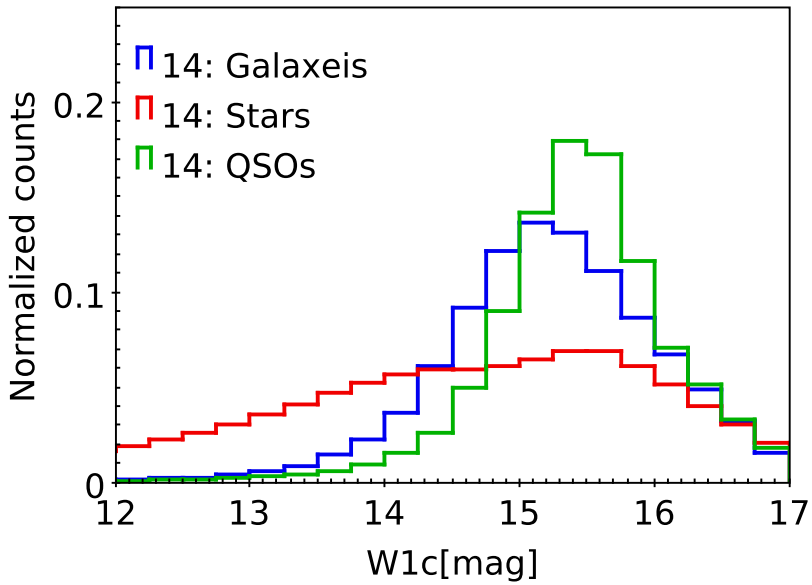
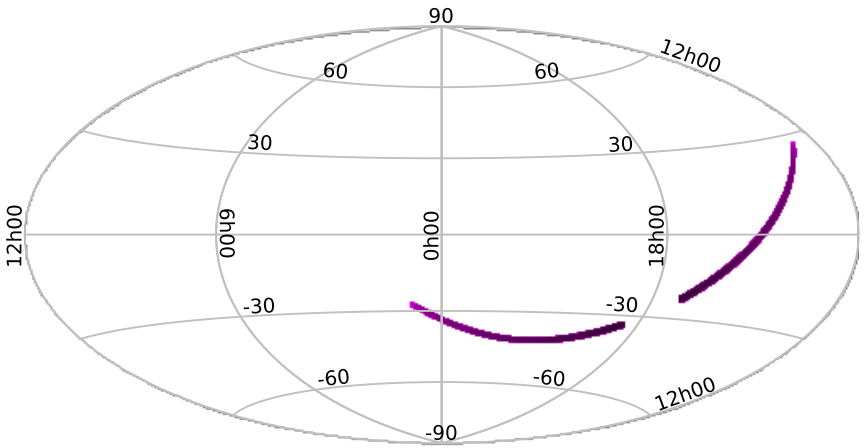


Figure 5 : **1 146 531 objects:** stars: 644 341 (56.20%), galaxies: 433 132 (37.78%), quasars: 69 058 (6.02%)

PRELIMINARY RESULTS FOR THE FINAL CATALOG

crossing the Bulge ($0° \leqslant 1 \leqslant 1°$)



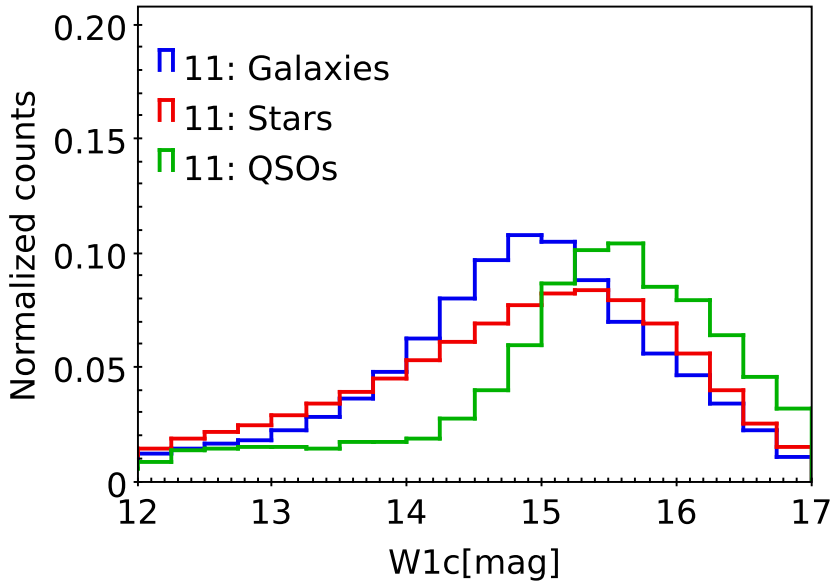
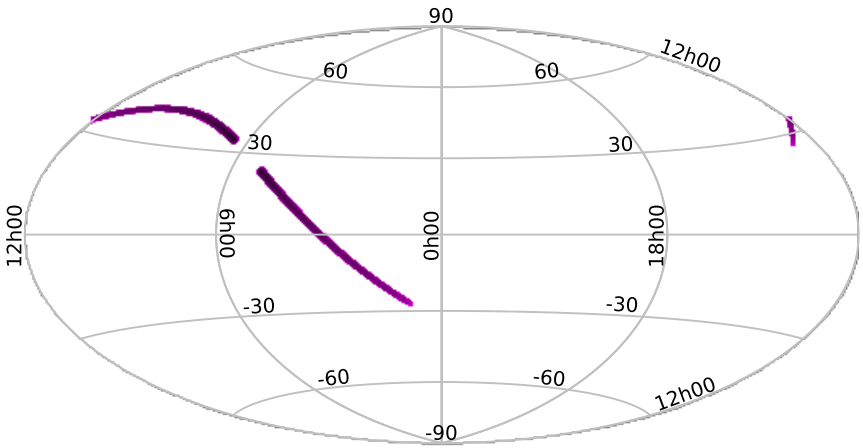


Figure 6 : **661 402 objects:** stars: 553 358 (83.66%), galaxies: 94 759 (14.33%), quasars: 13 285 (2.00%)

PRELIMINARY RESULTS FOR THE FINAL CATALOG

Galactic anti-center ($179^\circ \leqslant l \leqslant 180^\circ$)



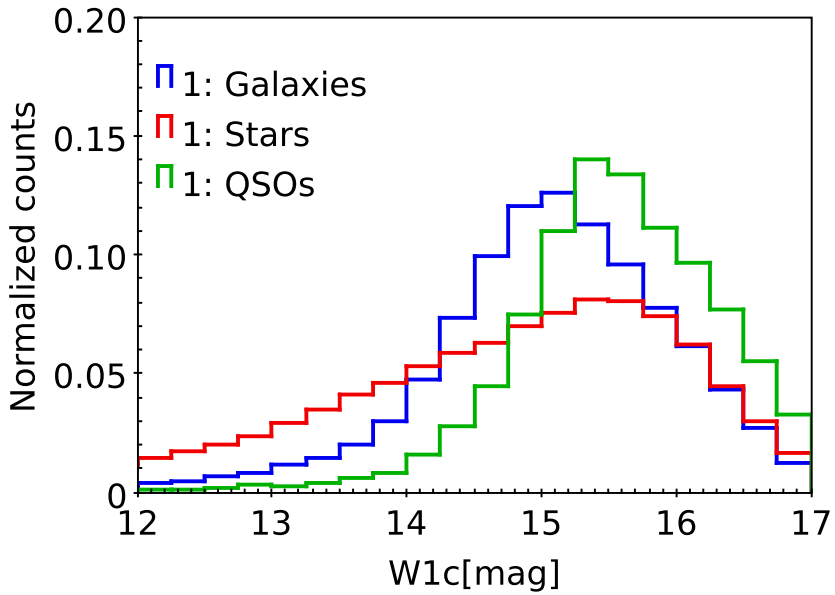


Figure 7 : **327 367 objects:** stars: 259 457 (79.25%), galaxies: 58 029 (17.23%), quasars: 9 881 (3.02%)

CONCLUSIONS

- WISExSuperCOSMOS catalog (170 milion of sources),

- WISExSuperCOSMOS catalog (170 milion of sources),
- Training Sample: WISExSuperCOSMOSxSDSS catalog of galaxies, stars, and quasars with spectroscopic classification (1.3 milion of sources),

- WISExSuperCOSMOS catalog (170 milion of sources),
- Training Sample: WISExSuperCOSMOSxSDSS catalog of galaxies, stars, and quasars with spectroscopic classification (1.3 milion of sources),
- C and γ tuning in 5 W1 bins,

- WISExSuperCOSMOS catalog (170 milion of sources),
- Training Sample: WISExSuperCOSMOSxSDSS catalog of galaxies, stars, and quasars with spectroscopic classification (1.3 milion of sources),
- C and γ tuning in 5 W1 bins,
- computation of the accuracy, purity, contaminations in function of W1 and b,

- WISExSuperCOSMOS catalog (170 milion of sources),
- Training Sample: WISExSuperCOSMOSxSDSS catalog of galaxies, stars, and quasars with spectroscopic classification (1.3 milion of sources),
- C and γ tuning in 5 W1 bins,
- computation of the accuracy, purity, contaminations in function of W1 and b,
- run the final 5D classifier (`W1`, `W1-W2`, `R-W1`, `B-R`,`w1mag13`) against the 4 test subsamples of the WISExSCOS catalog,

- WISExSuperCOSMOS catalog (170 milion of sources),
- Training Sample: WISExSuperCOSMOSxSDSS catalog of galaxies, stars, and quasars with spectroscopic classification (1.3 milion of sources),
- C and γ tuning in 5 W1 bins,
- computation of the accuracy, purity, contaminations in function of W1 and b,
- run the final 5D classifier (W1, W1−W2, R−W1, B−R,w1mag13) against the 4 test subsamples of the WISExSCOS catalog,
- run the final classifier against the whole 170 milion of WISExSuperCOSMOS catalog (the results will be publish very soon).

- WISExSuperCOSMOS catalog (170 milion of sources),
- Training Sample: WISExSuperCOSMOSxSDSS catalog of galaxies, stars, and quasars with spectroscopic classification (1.3 milion of sources),
- C and γ tuning in 5 W1 bins,
- computation of the accuracy, purity, contaminations in function of W1 and b,
- run the final 5D classifier (`W1`, `W1-W2`, `R-W1`, `B-R`,`w1mag13`) against the 4 test subsamples of the WISExSCOS catalog,
- run the final classifier against the whole 170 milion of WISExSuperCOSMOS catalog (the results will be publish very soon).

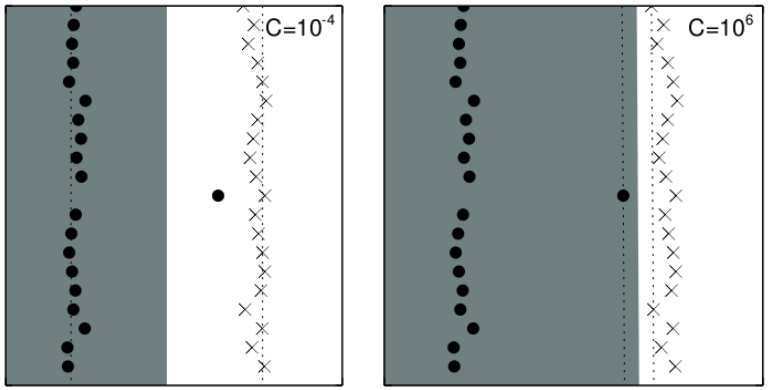
The aim is reached

THANK YOU FOR YOUR ATTENTION

final catalog:
almost ready, need additional 2 weeks of computation,
more about our results:
Krakowiak et al., in preparation, Kurcz et al., in preparation

Non-linearly separable data

- C - trade-off parameter between large margin M and poor classifications ξ_i

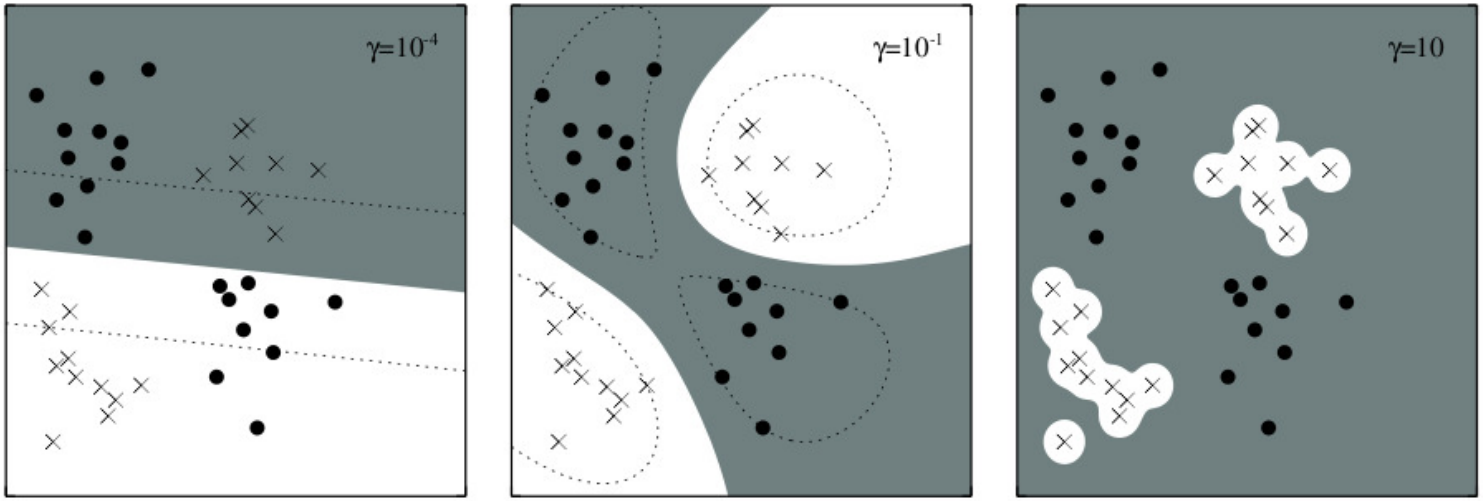


- large C: single outlier on the wrong side results in small margin;
- reducing C: individual ξ_i penalize classification less
- classifier mis-classifies small number of objects but has larger margin;

Non-linearly separable data

To determine the topology of the decision surface: kernel functions with adjustable parameter γ.

- radial basis: $k(x, x') = exp(-\gamma||x - x'||^2)$



- low values - stiff boundaries;
- too high - over-fitting;

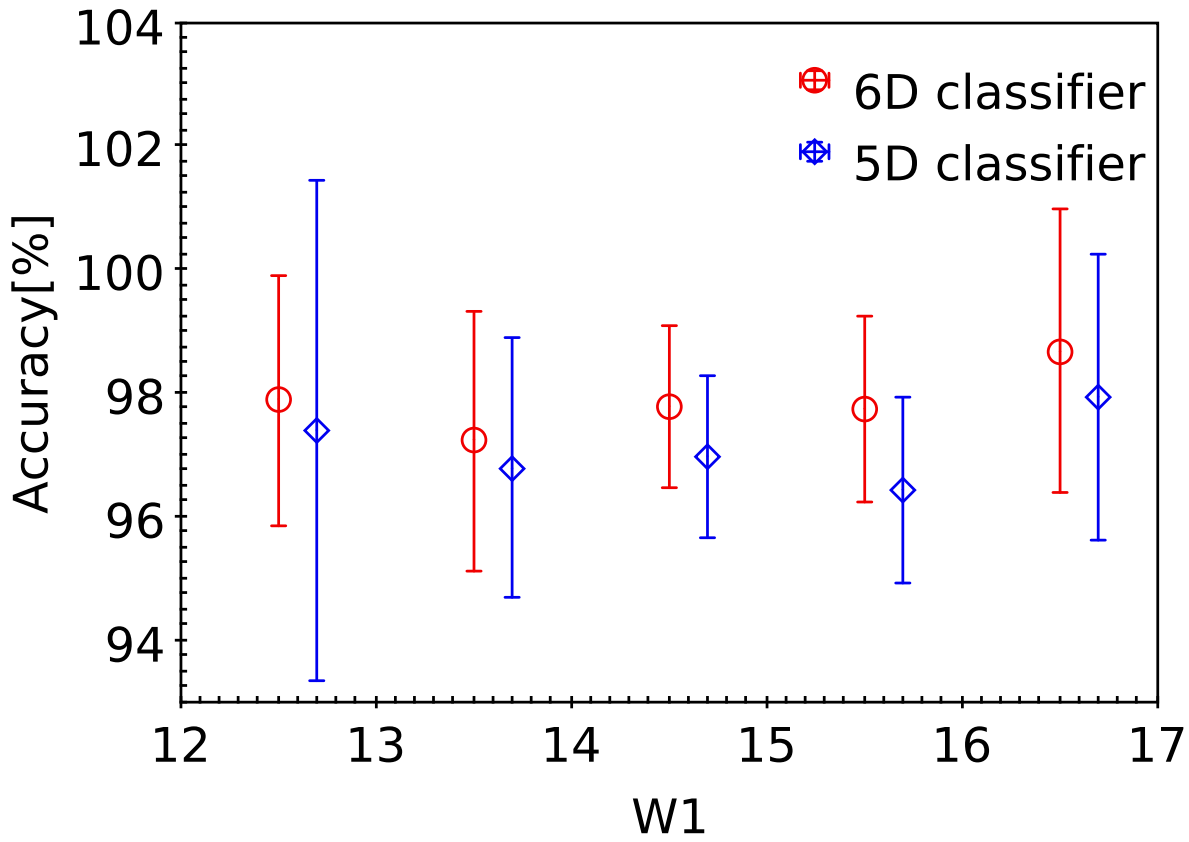
C and γ tuning

We have performed a grid-search using a 10-fold cross-validation technique:

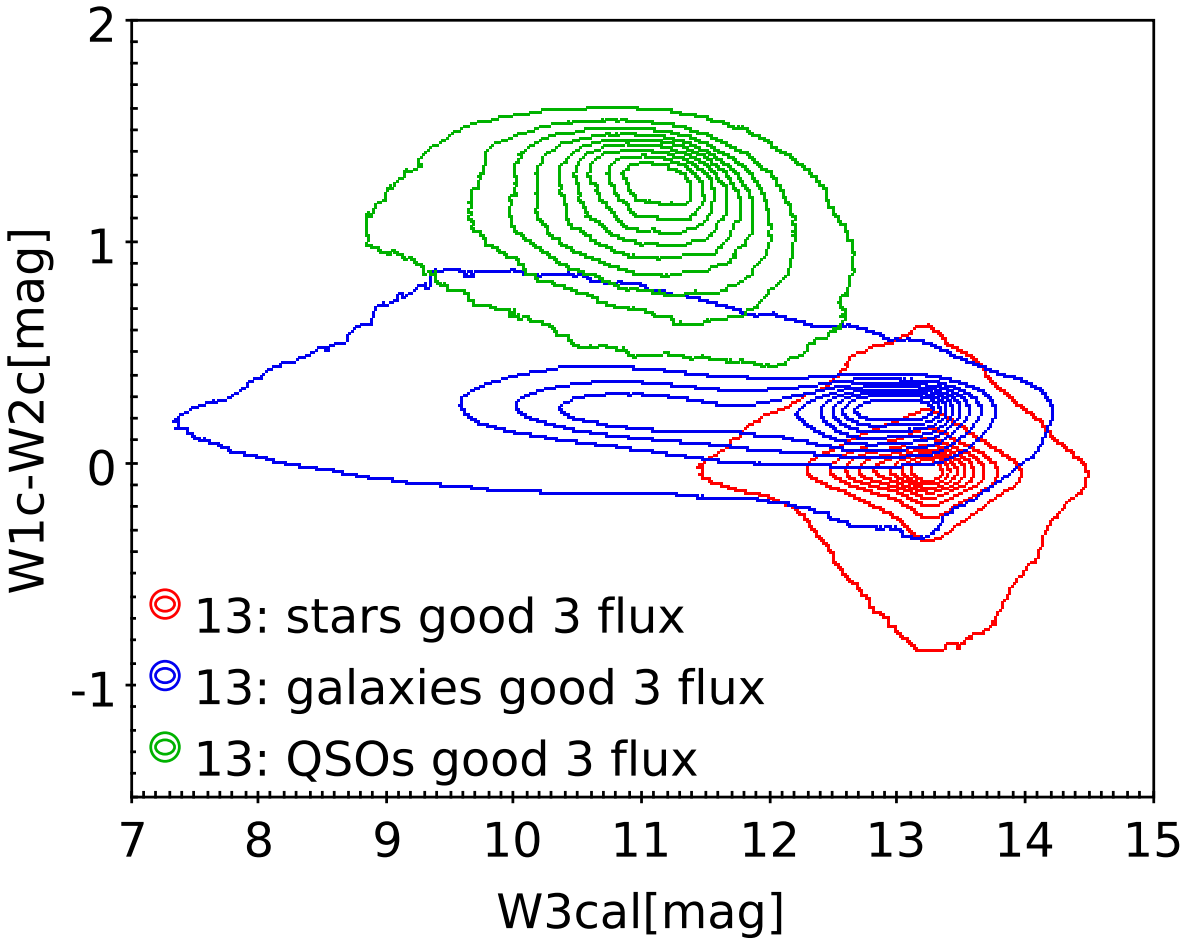
- we divided the full Training Sample into 10 subsets of equal size, and we selected 9 subsets to train the classification model and test it against the remaining subset (so called `self-check`).

It is the iterative process, and its result is a hyperspace between different classes of objects with tuned parameters controlling the size of margins between classes and ratio of mis-classifications.

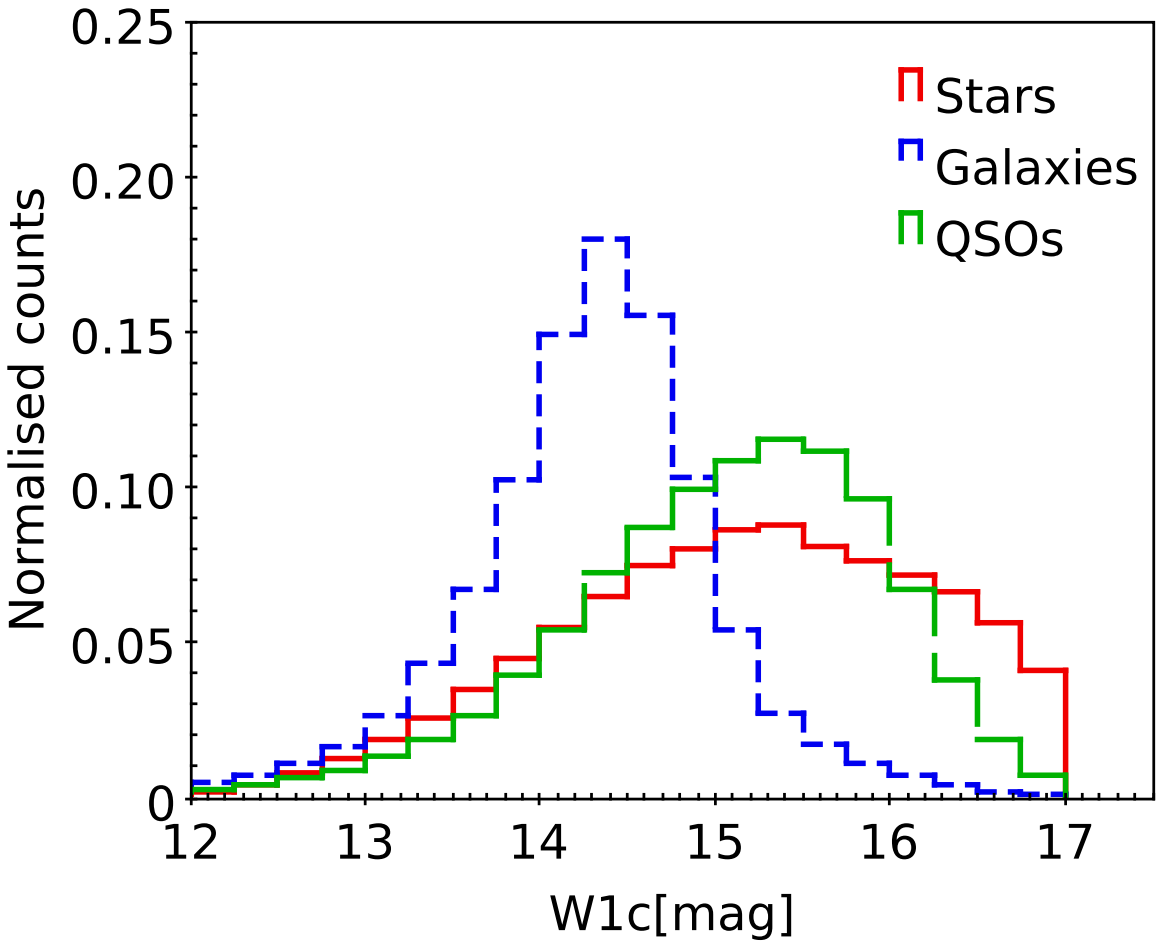
5D vs 6D classifier



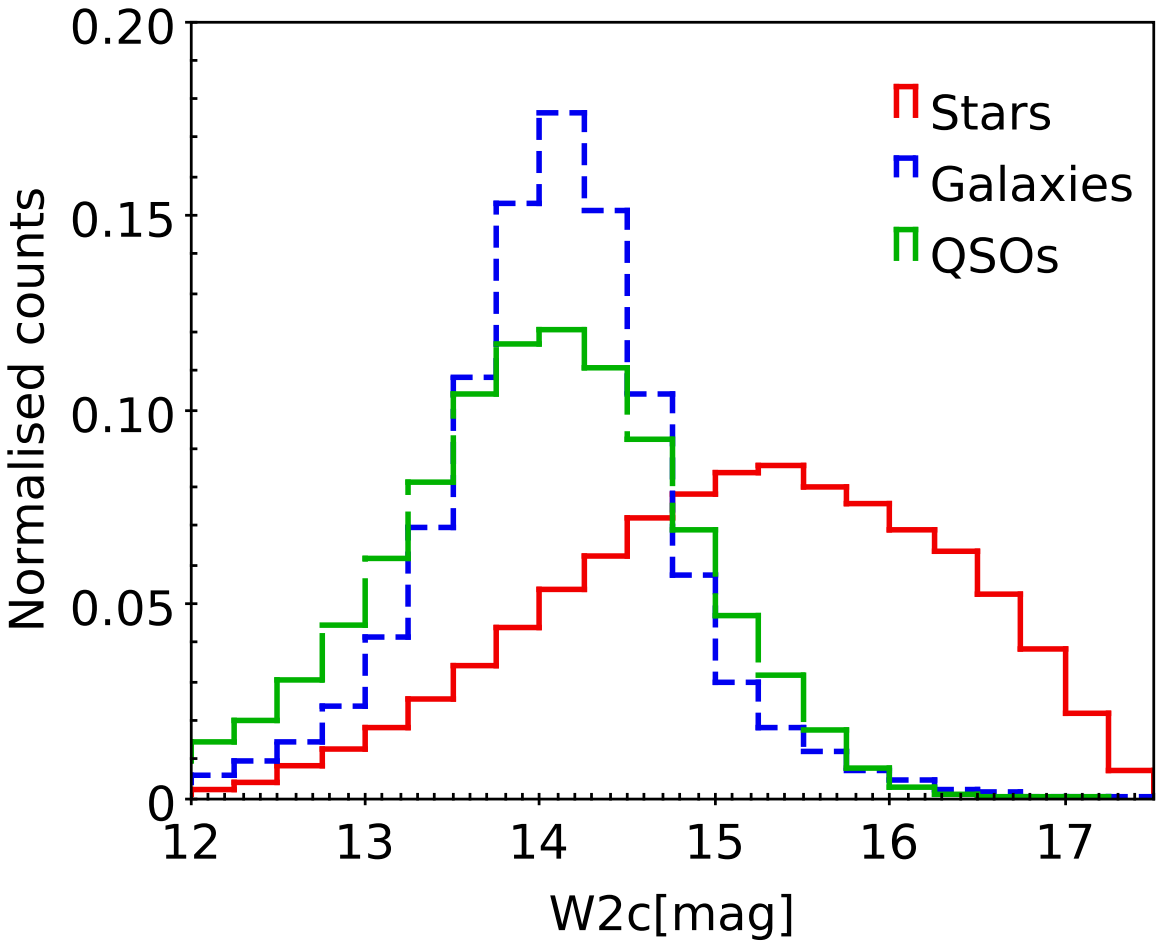
W3 vs W1-W2


W1c-W2c[mag]
W3cal[mag]
2
1
0
-1
7
8
9
10
11
12
13
14
15
13: stars good 3 flux
13: galaxies good 3 flux
13: QSOs good 3 flux

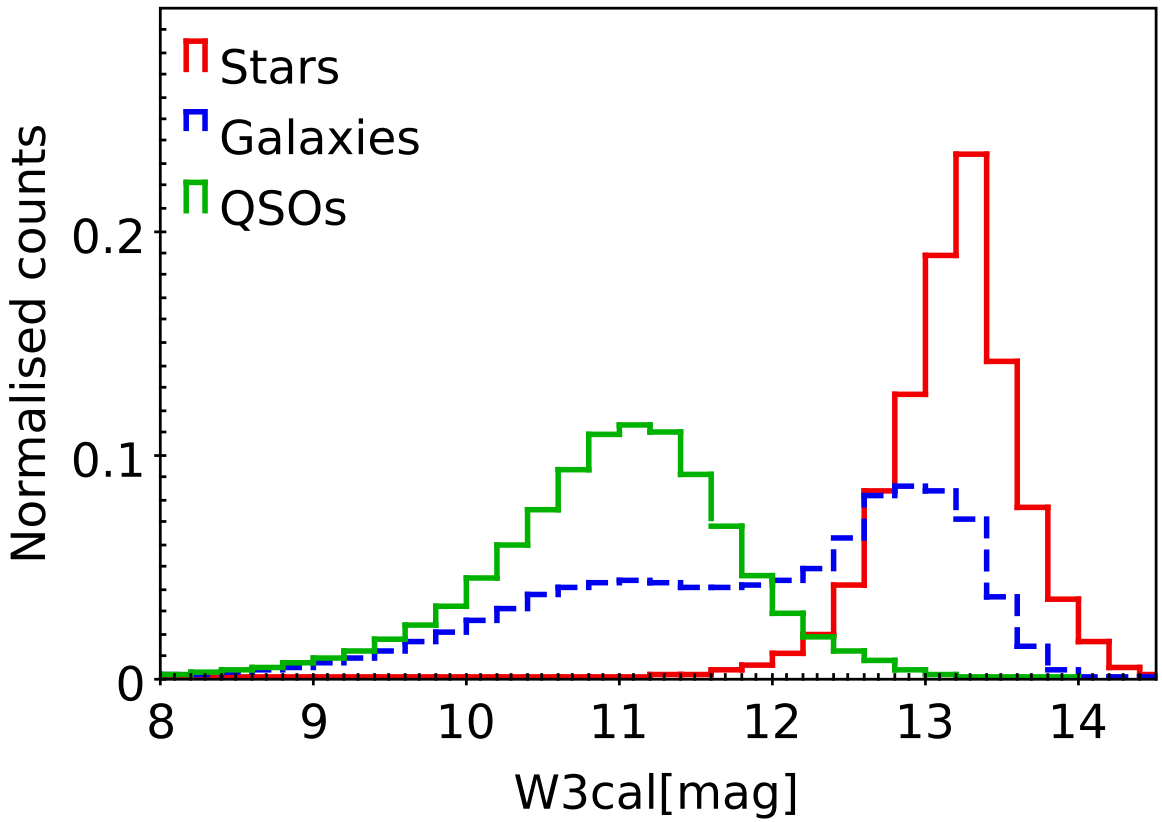
WISExSCOSxSDSS W1



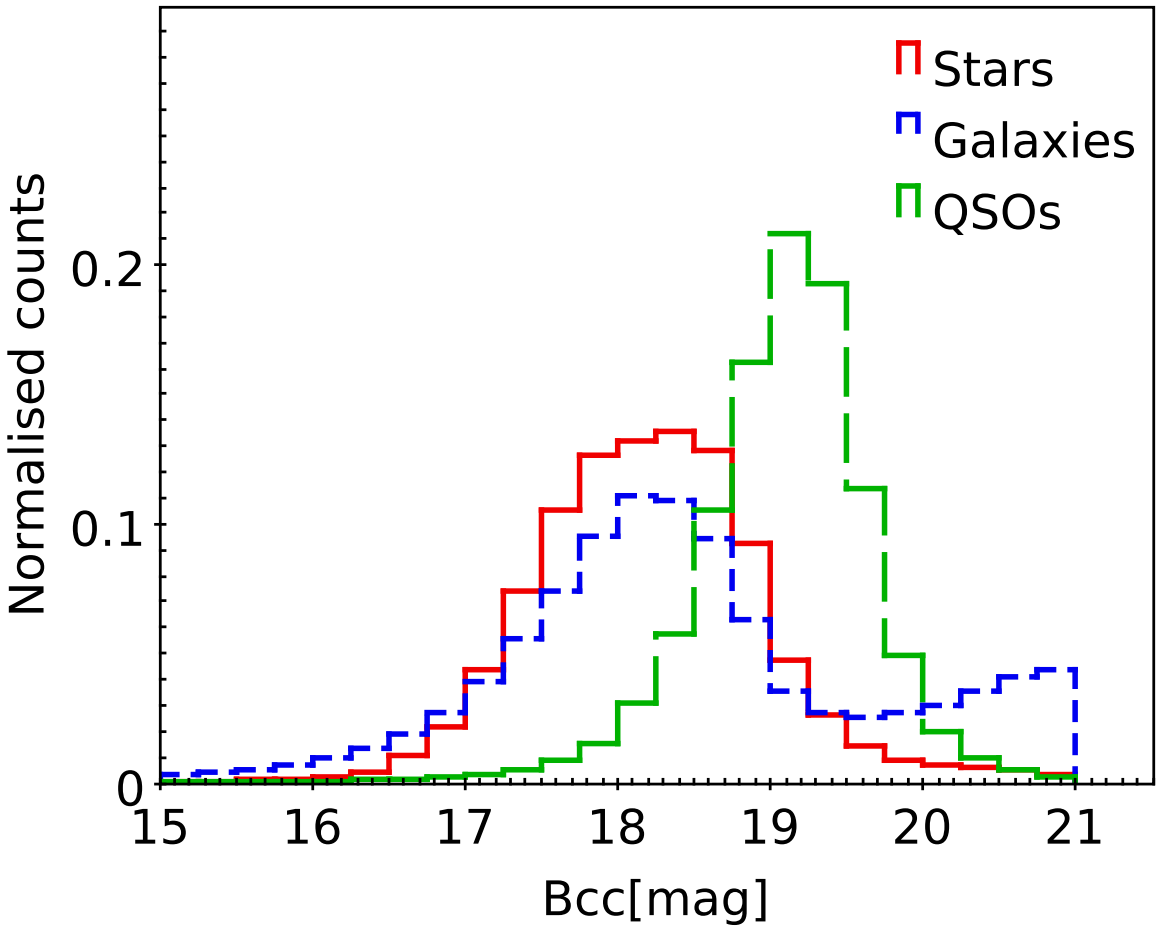
WISExSCOSxSDSS W2



WISExSCOSxSDSS W3



WISExSCOSxSDSS B



WISExSCOSxSDSS R