Identification of ultracool dwarfs in J-PLUS DR2 using Virtual Observatory tools and Machine Learning techniques

Pedro Mas Buitrago

Supervisores

Enrique Solano Márquez (CAB)

Ana González Marcos (Universidad de La Rioja)

UNIVERSIDAD COMPLUTENSE MADRID

MINISTERIO DE CIENCIA E INNOVACIÓN

AGENCIA ESTATAL DE INVESTIGACIÓN

CENTRO DE ASTROBIOLOGÍA CSIC - INTA

EXCELENCIA MARÍA DE MAEZTU

1. Scientific context
2. VO methodology
3. Machine Learning methodology
4. Conclusiones and future work

Scientific context: Ultracool dwarfs

- UCDs comprise the lowest mass members of the stellar population and brown dwarfs
- From the M7 V to the extended L, T and Y spectral types
- Effective temperatures of $T_{eff} \lesssim 2900\ K$
- Relevant role in the search for Earth-like exoplanets, in the study of Galactic kinematics and in the understanding of the boundary between stellar and substellar objects

Scientific context: J-PLUS

- Javalambre Photometric Local Universe Survey, conducted from the Observatorio Astrofísico de Javalambre (OAJ)
- Multi-filter survey with 12 optical bands
- Large coverage of the electromagnetic spectrum, which allows a more accurate determination of physical parameter
- The J-PLUS Data Release 2 covers 2 176 deg^2

UNIVERSIDAD COMPLUTENSE MADRID

EXCELENCIA MARÍA DE MAEZTU

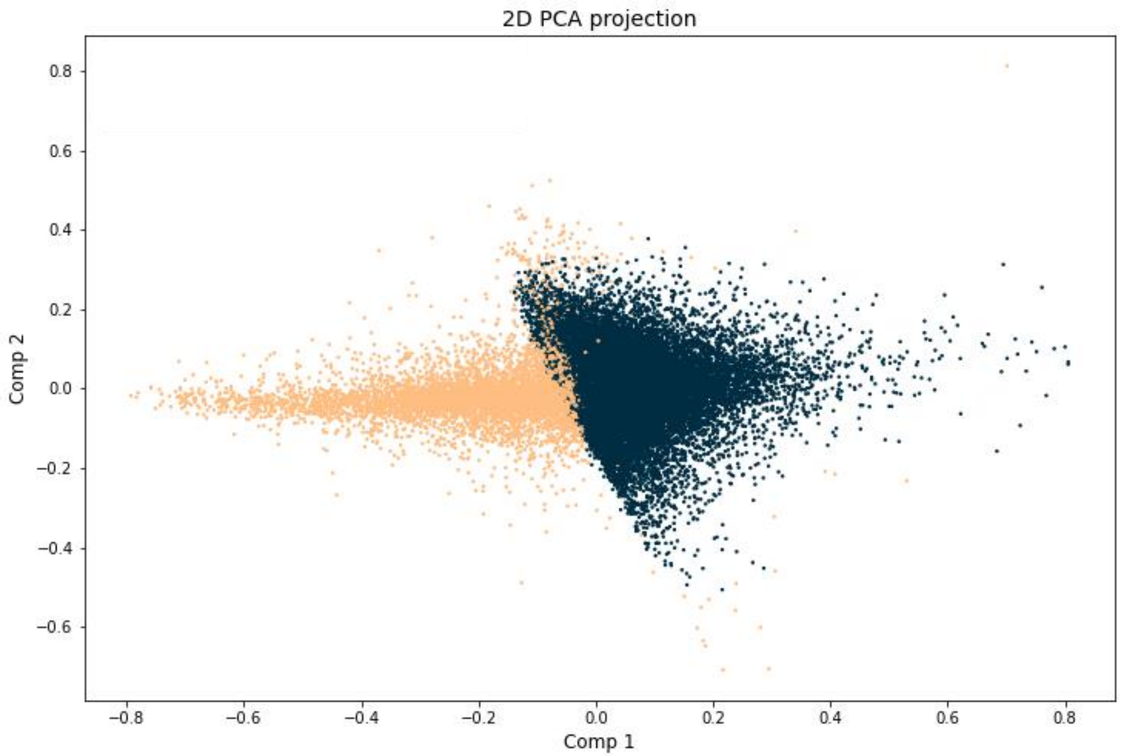

SDSS9
J-PLUS DR2

Scientific context: SVO

- The Virtual Observatory (IVOA) is an international initiative to provide seamless access to the data available from astronomical archives and services as well as state-of-the-art tools
- The vision that astronomical datasets and other resources should work as a whole
- This is made possible by standardization of data and metadata, by standardization of data exchange methods, and by the use of a registry, which lists available services and what can be done with them
- The Spanish Virtual Observatory is part of IVOA

Scientific context: PCA step

- Reduces a lower data set to a lower dimension by identifying the axes that account for the largest amount of variance
- The expectation behind is that the entire data set can be well characterized along a small number of dimensions
- Deterministic nature, i.e. different runs of PCA on a given dataset will always produce the same results

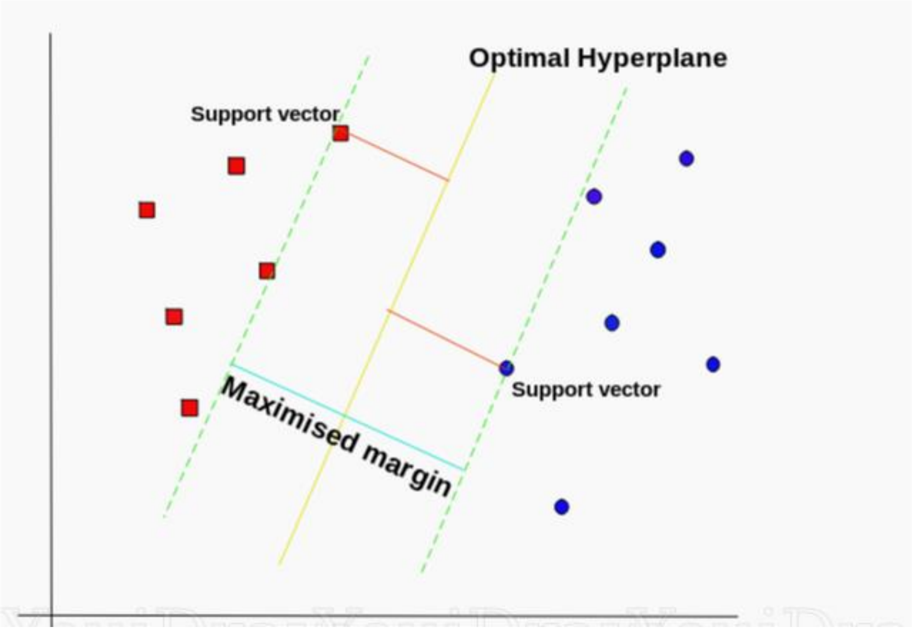


UNIVERSIDAD COMPLUTENSE MADRID

EXCELENCIA MARÍA DE MAEZTU

Scientific context: SVM algorithm

- Supervised ML algorithm
- The idea behind is to find a hyperplane that separates data into two classes while maximizing a marging
- Linear classifier: we can gain linear separation by mapping the data to a higher dimensional space with different kernels (polynomial, Radial Basis Function…)



https://gdcoder.com/support-vector-machine-vs-logistic-regression/

UNIVERSIDAD COMPLUTENSE MADRID

EXCELENCIA MARÍA DE MAEZTU

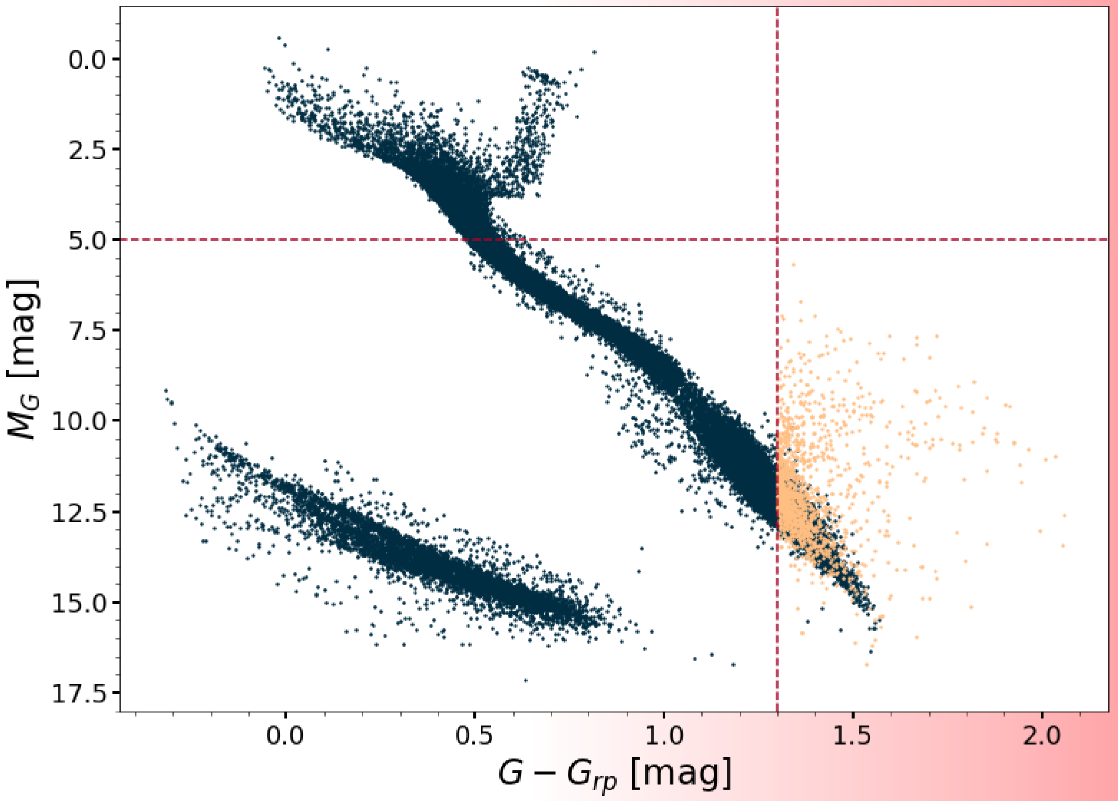
VO Methodology: Pre-screening process

Two astrometric approaches

- Parallax-based:
  - Relative error of less than 20% in parallax
  - $G - G_{RP} > 1.3$ - Pecaut & Mamajek (2013)
- Proper motion-based:
  - Relative error of less than 20% in both pm components
  - Only sources with non-zero pm
  - $G - G_{RP} > 1.3$ - Pecaut & Mamajek (2013)

One photometric approach

- Colour cut of $r - z > 2.2$ using J-PLUS photometry



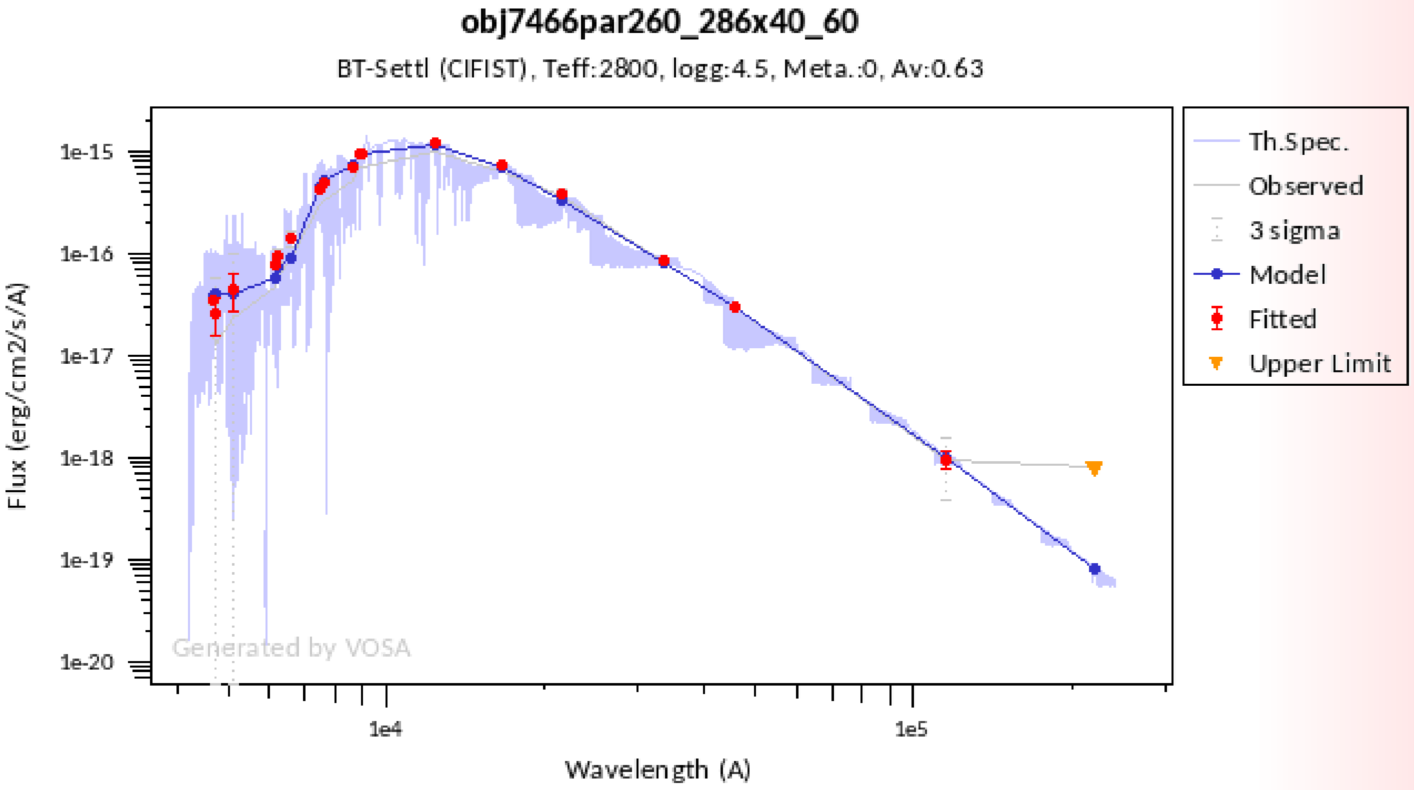
UNIVERSIDAD COMPLUTENSE MADRID

CENTRO DE ASTROBIOLOGÍA CSIC - INTA

EXCELENCIA MARÍA DE MAEZTU

VO Methodology: Final candidate UCDs

- We used VOSA to complement J-PLUS photometry with optical and infrared VO catalogues
- VOSA fit: effective temperature cut of $T_{eff} \leq 2900\ K$



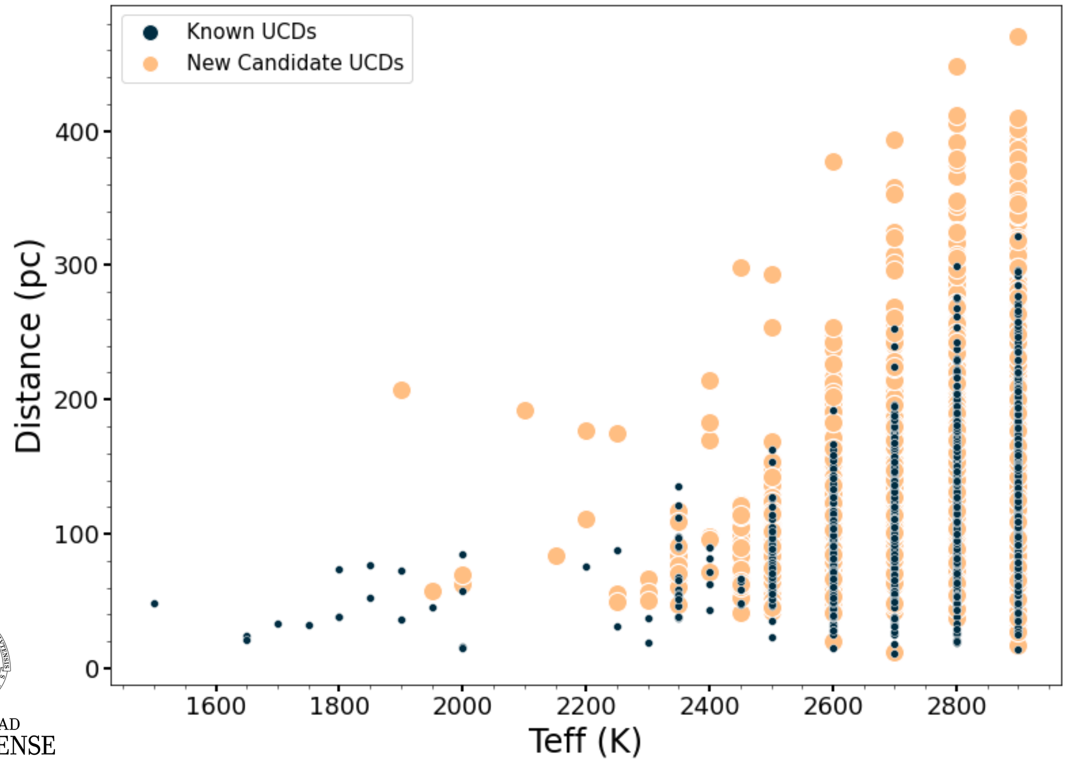
UNIVERSIDAD COMPLUTENSE MADRID

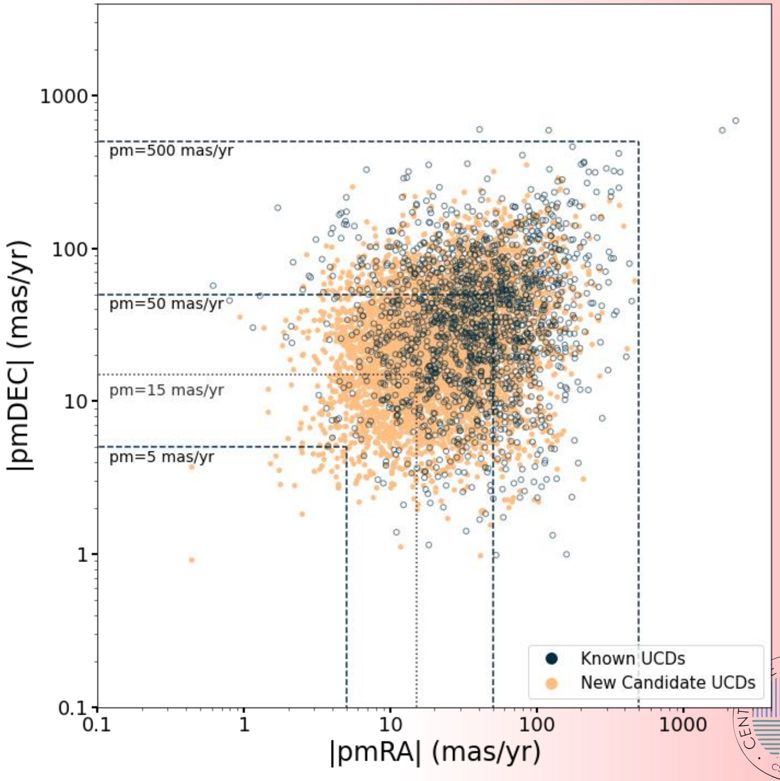
CENTRO DE ASTROBIOLOGÍA CSIC - INTA

EXCELENCIA MARÍA DE MAEZTU

VO Methodology: Results

- We identified a total of 9 810 candidate UCDs, of which only 1 981 were previously reported (increase of ~135% in J-PLUS DR2 sky coverage)
- In-Depth kinematic and binarity analysis of the candidate UCDs
- Our methodology allows us to go to further distances and smaller proper motions in the search for UCDs



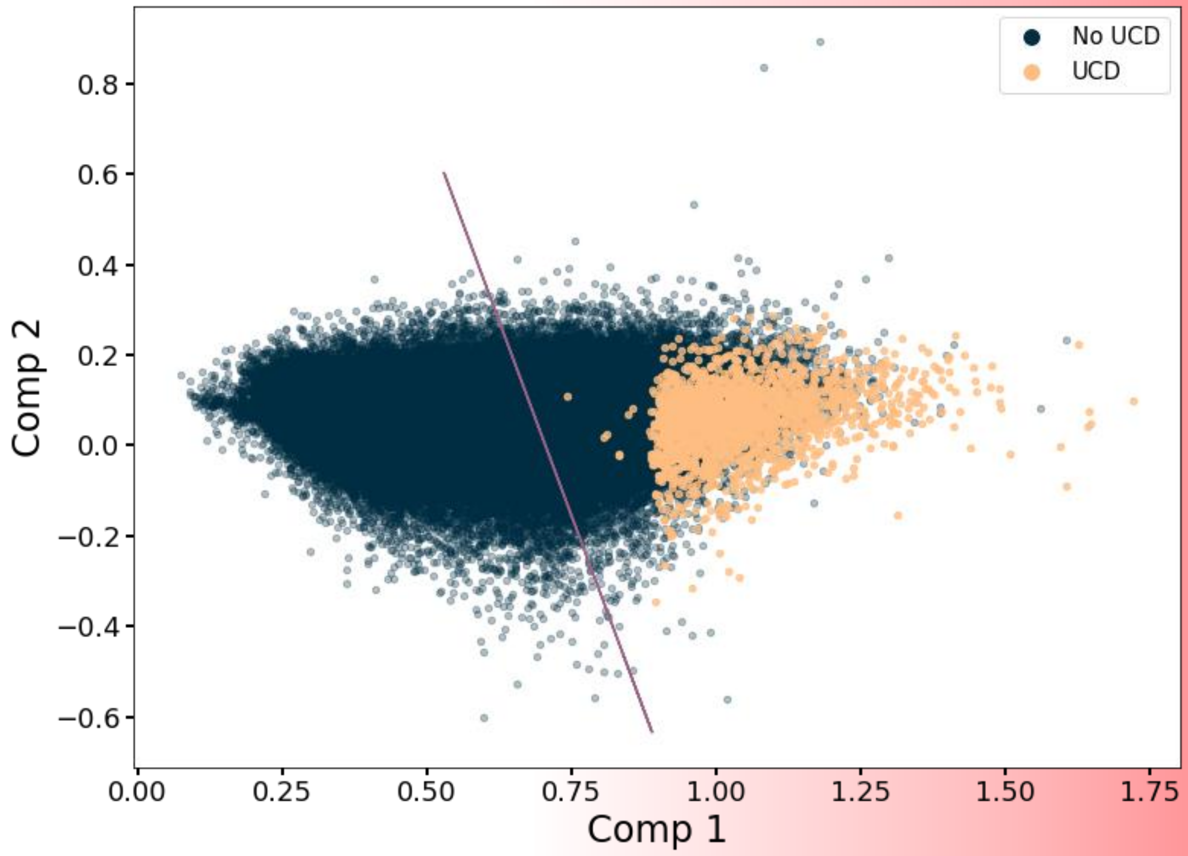


UNIVERSIDAD COMPLUTENSE MADRID

EXCELENCIA MARÍA DE MAEZTU

ML Methodology: PCA step

1. We relied only J-PLUS photometry, using as features eight different colours ($i-z, r-i, J0861-i$ ...)
2. We labeled the instances as positive or negative class using the candidate UCDs obtained with the VO methodology
3. Stratified sampling for training and test sets
4. PCA model: 94% of the variance along the two first Principal components
5. First cut in the identification of UCDs to reduce the imbalance of the data



UNIVERSIDAD COMPLUTENSE MADRID

EXCELENCIA MARÍA DE MAEZTU

ML Methodology: SVM step

1. The SVM model is developed using the reduced sample obtained with the PCA filtering
2. Exhaustive search for the optimal SVM hyperparameters using *GridSearchCV* class from the Python package SCIKIT-LEARN
3. Best recall score with an RBF kernel and hyperparameters $C = 1000, \gamma = 0.001$
4. Recall of 98% and 96% in the test set and in the blind test, respectively

UNIVERSIDAD
COMPLUTENSE
MADRID

EXCELENCIA
MARÍA
DE MAEZTU

ML Methodology: Results

- The PCA filter is able to remove the hottest objects ($T_{eff} \geq 4\ 100\ K$)
- We recover nearly all the UCD objects, but we still need to apply VOSA to a significant number of objects
- Restrictive methodology in terms of photometric quality

Conclusions and future work

- Paper under review by the J-PLUS consortium
- We consolidated and further developed a search methodology, introduced in Solano et al. (2019), to be used for deeper and larger surveys like J-PAS and Euclid
- The ML methodology is more efficient in the sense that it allows a greater number of non-UCD objects to be discarded prior to analysis with VOSA
- Real turning point: ML methodology that more significantly filters the number of objects we need to analyze with VOSA
    - Independent Component Analysis (ICA)
    - Ensemble learning

Thanks for your attention

SVO
Spanish Virtual Observatory

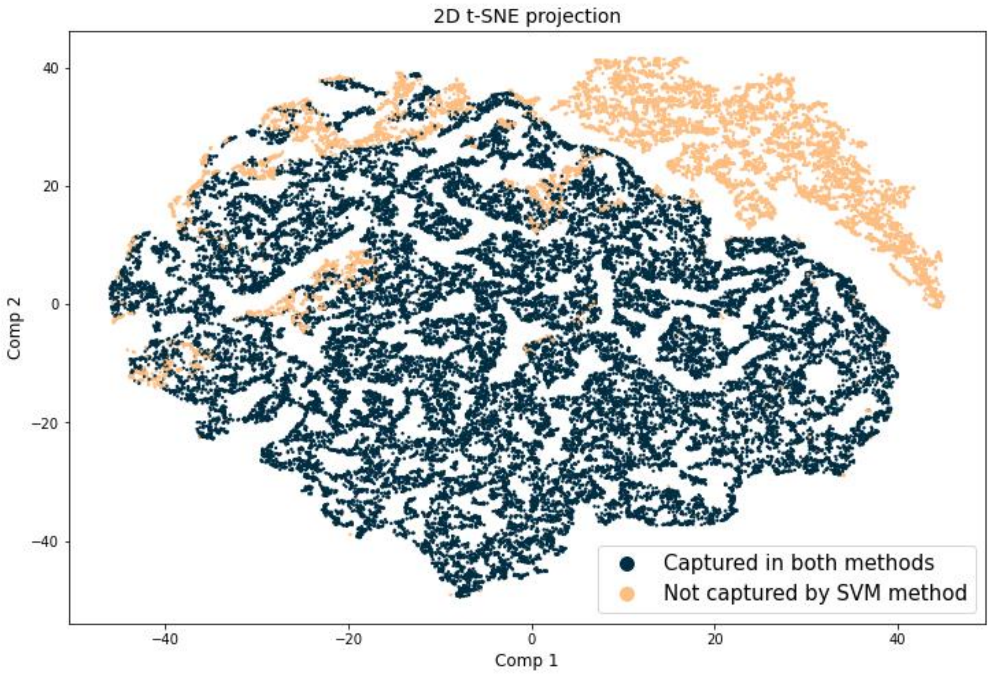
UNIVERSIDAD
COMPLUTENSE
MADRID

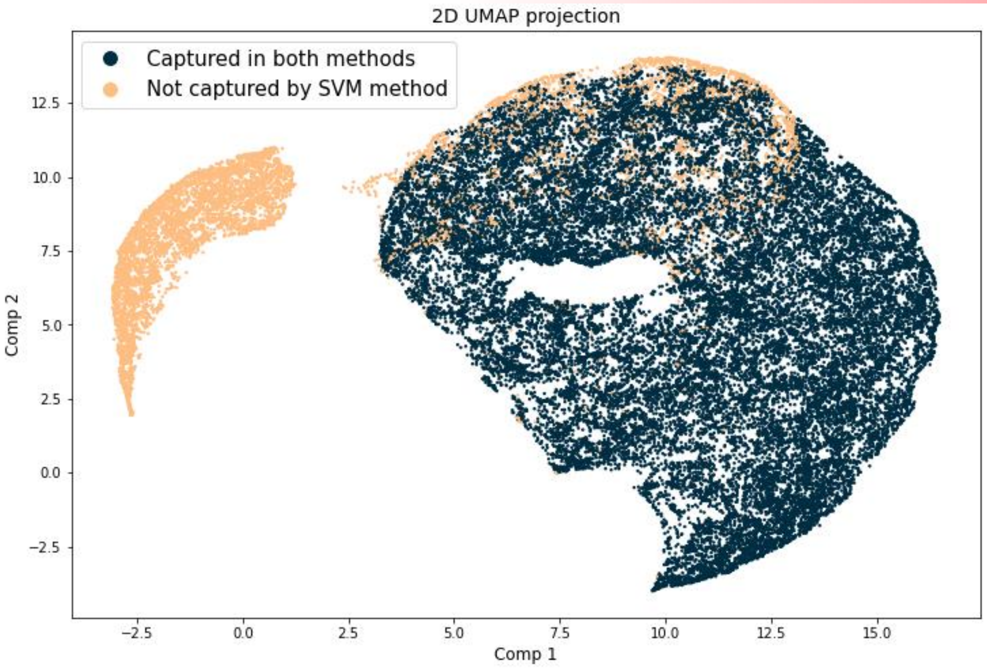
MINISTERIO
DE CIENCIA
E INNOVACIÓN

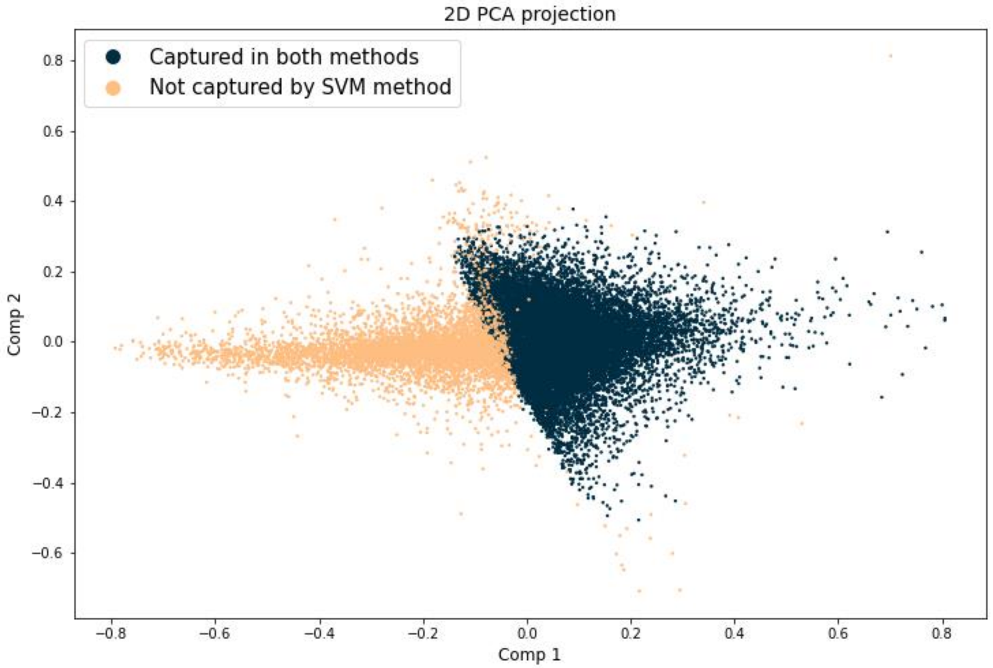
AGENCIA
ESTATAL DE
INVESTIGACIÓN

CENTRO DE ASTROBIOLOGÍA
CSIC - INTA

EXCELENCIA
MARÍA
DE MAEZTU







UNIVERSIDAD
COMPLUTENSE
MADRID

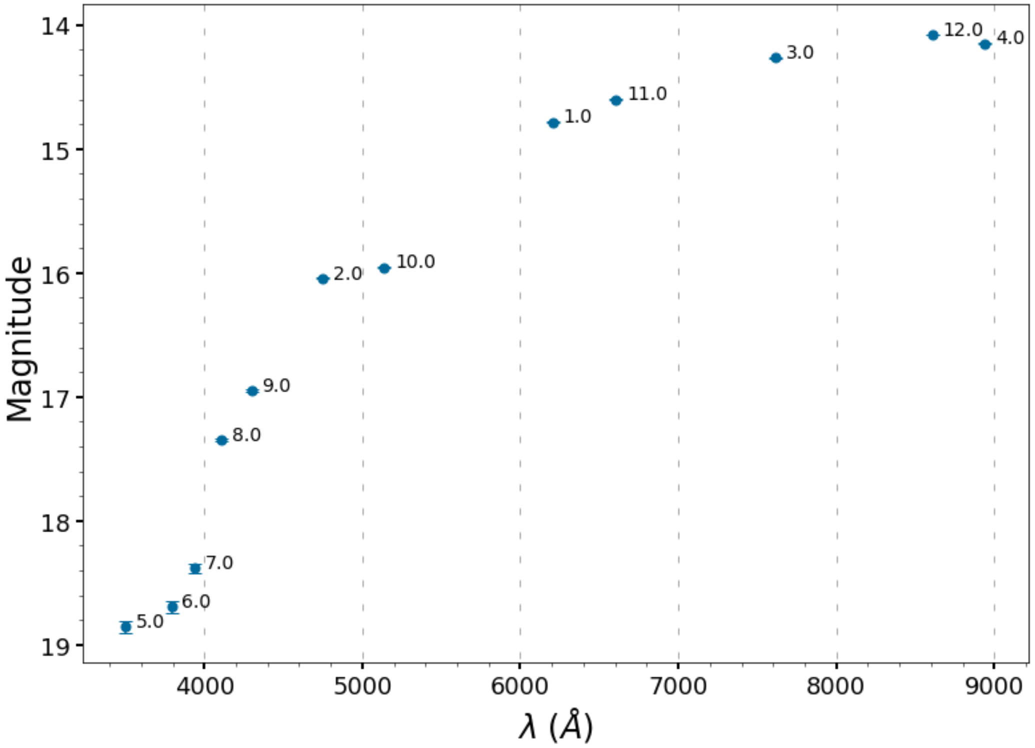
EXCELENCIA
MARÍA
DE MAEZTU

Tools

- VO tools: Aladin, Topcat, VOSA, STILTS…
- Astronomical Data Query Language (ADQL) and VO TAP protocol
- Several Python packages (Pandas, Seaborn, Scikit-Learn, MocPy…)
- Machine Learning algorithms:
    - Principal Component Analysis (PCA)
    - Support Vector Machine (SVM)

Data sources

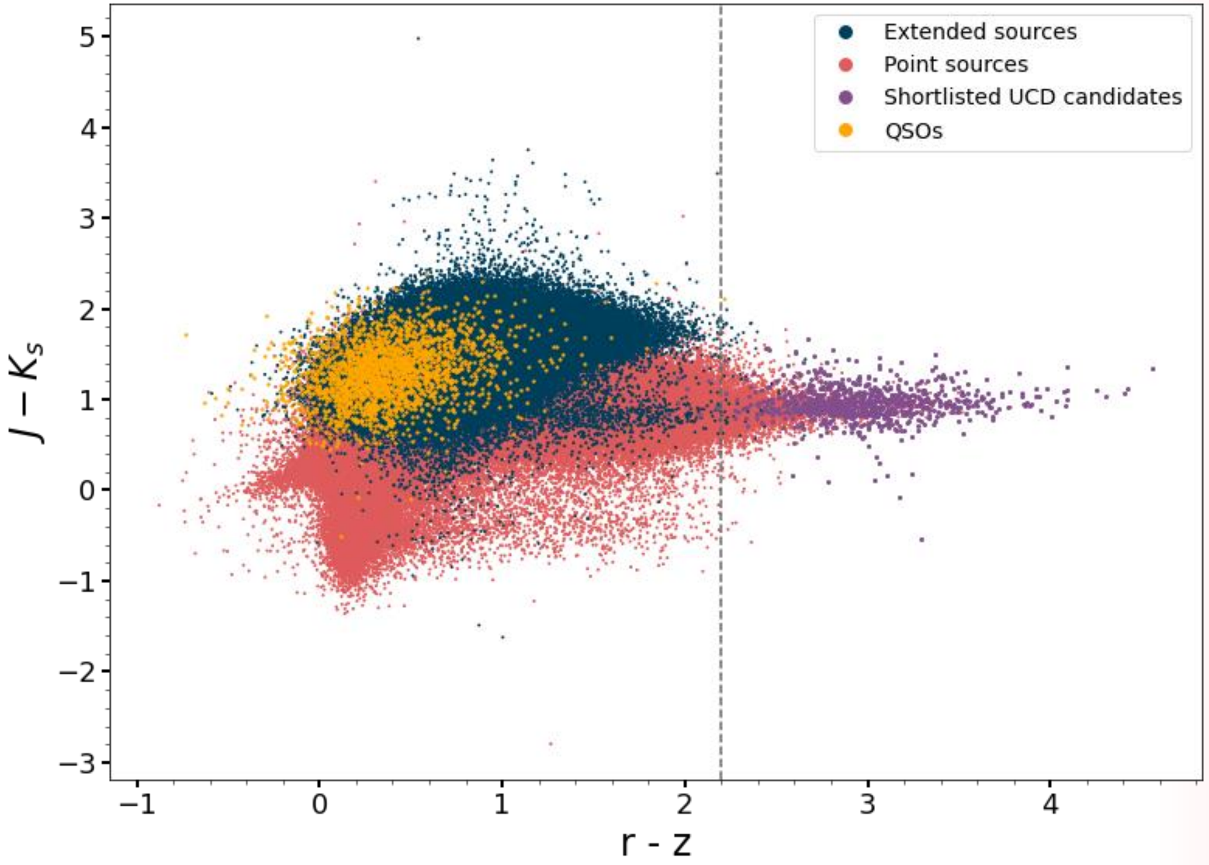
- J-PLUS DR2 photometry
- Parallaxes and proper motions from Gaia EDR3
- Complementary optical and infrared photometry from different VO catalogues available in VOSA



VO Methodology: Pre-screening process

One photometric approach

- Colour cut of $r - z > 2.2$ using J-PLUS photometry



UNIVERSIDAD COMPLUTENSE MADRID

EXCELENCIA MARÍA DE MAEZTU

VO Methodology: Obtaining the data

1. We divided the sky coverage of J-PLUS DR2 in 37 regions of 20x20 deg^2
2. We tessellated each region into smaller circular sub-regions of 1 deg radius
3. We built a Python code that to query the J-PLUS database over all the 20x20 deg^2 regions iteratively
4. We cross-matched the queried sources with *Gaia* EDR3 to obtain the astrometric information

```
SELECT filter_id, alpha_j2000, delta_j2000,
    mag_aper_6_0, mag_err_aper_6_0
FROM jplus.MagABSingleObj
WHERE alpha_j2000 between 2 and 5
AND delta_j2000 between 2 and 3
AND flags=0
AND filter_id between 1 and 4
AND class_star>0.1
```