Searching for extragalactic variable stars using Machine Learning algorithms

Javier Alejandro Acevedo Barroso

Advisor: Alejandro García

Departamento de Física
Universidad de los Andes

February 9, 2021

Outline

Main scope of the project

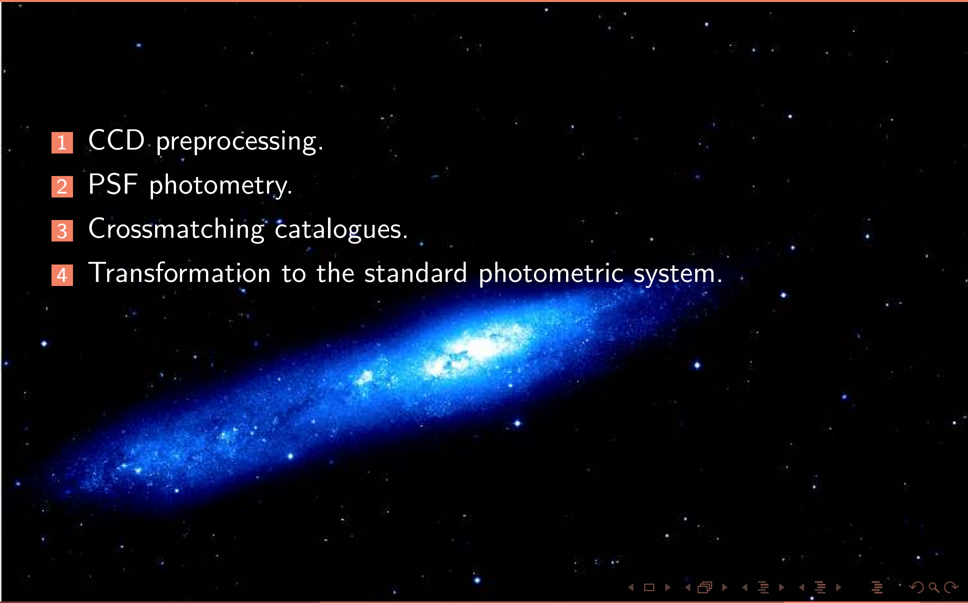
- To look for variable stars in the galaxy NGC 55 using public data from the European Southern Observatory archive to generate *light curves*, and then supervised machine learning techniques to classify them.

The two big objectives

- To analyse never-published wide field images for NGC 55 in order to generate light curves.
- To implement a variable star classifier using machine learning algorithms and OGLE Catalog of Variable Stars as training sample.

Generating light curves

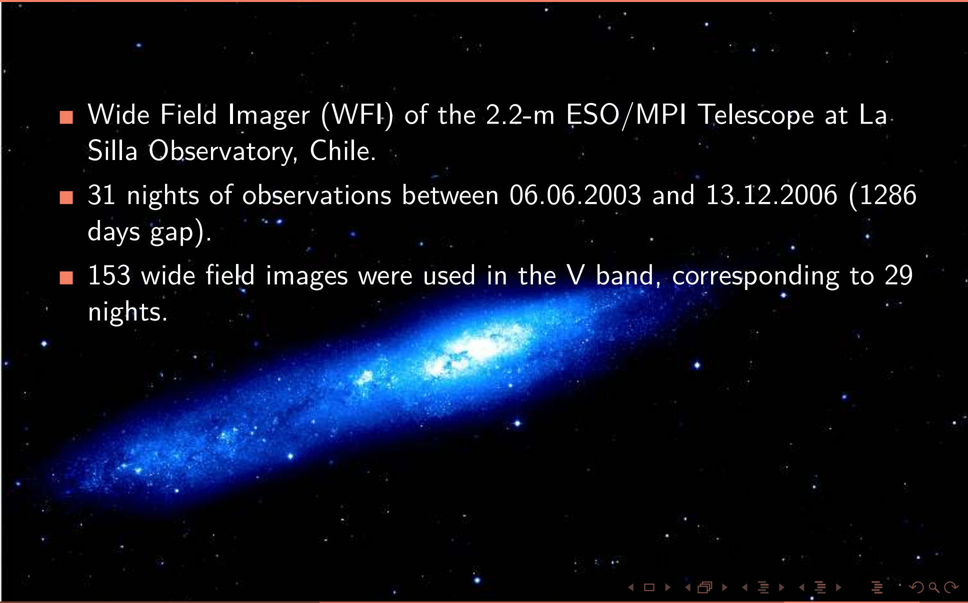
Methodology



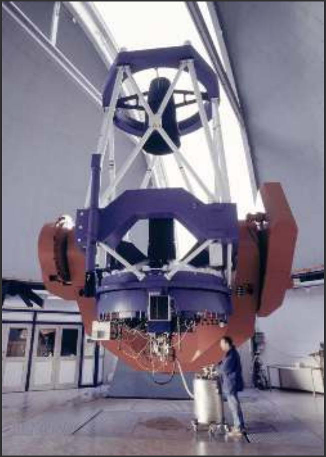
1. CCD preprocessing.
2. PSF photometry.
3. Crossmatching catalogues.
4. Transformation to the standard photometric system.

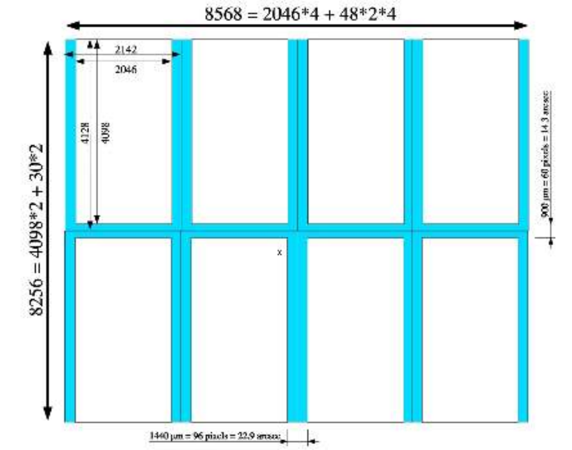
The photometric survey

- Wide Field Imager (WFI) of the 2.2-m ESO/MPI Telescope at La Silla Observatory, Chile.
- 31 nights of observations between 06.06.2003 and 13.12.2006 (1286 days gap).
- 153 wide field images were used in the V band, corresponding to 29 nights.



Instrumentation





Figures taken from *The Wide Field Imager Handbook* SciOps 2005.

Summary of survey parameters

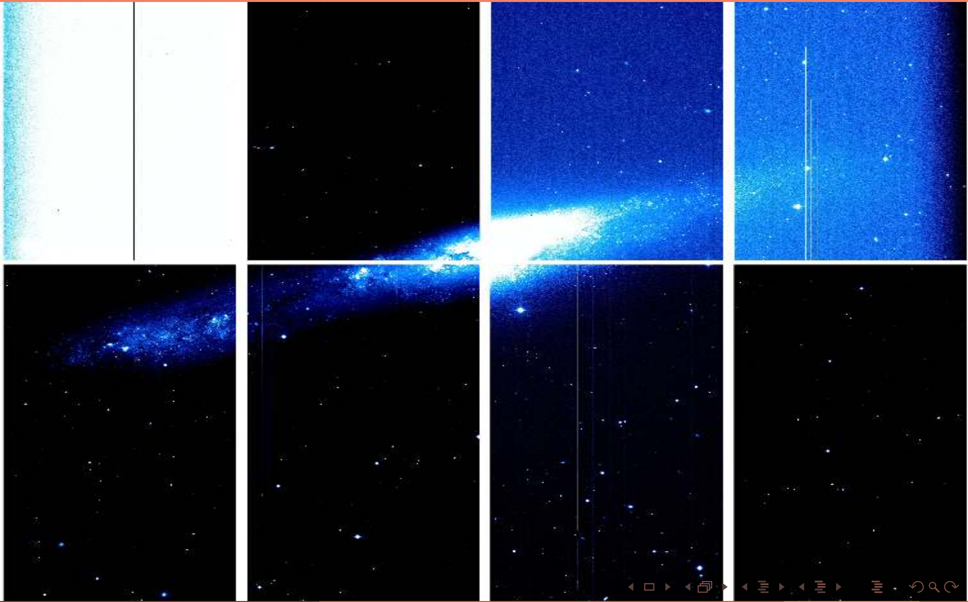
Total			Seeing [$''$]		
Nights	Images	Time (V band) [s]	Range	Average	Air masses
29	153	299.917	[0.5, 1.7]	1.1	1.02 - 1.52

CCD preprocessing

- Overscan removal.
- Average bias correction.
- Cosmic ray removal.
- Flat fielding.
- Bad pixel masking.
- Stack of dithered sequences (possible only in 23 of the 29 nights).

All the preprocessing was done using IRAF (Tody 1986) and its packages: CCDPROC (Valdes 1988), MSCRED and ESOWFI (Valdes 1998).

Raw image

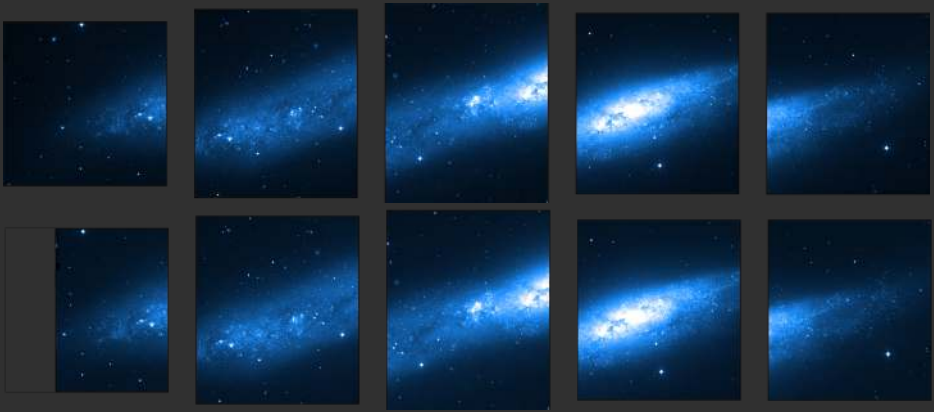


CCD corrected image

Stacked image

1. Fixing World Coordinate System on every image (and some bugs on IRAF).
2. Resampling the mosaic images in a single plane.
3. Matching photometric scale and zero point (mscimatch vs philmatch).
4. Creating the final stack using pixel-wide averages.

Photometry cuts - 20% overlap

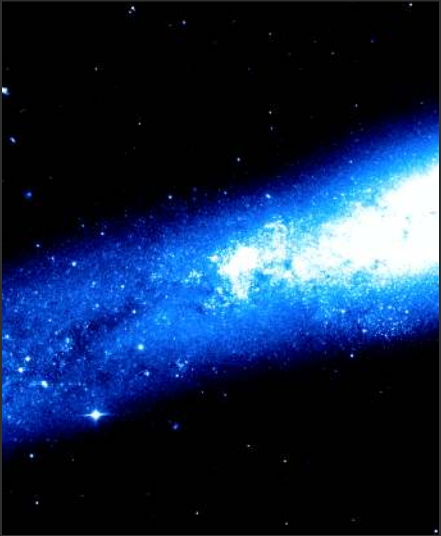


PSF Photometry - Methodology

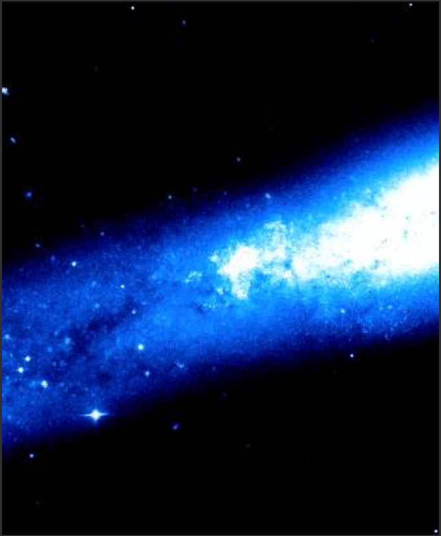
1. Measuring FWHM of previously selected stars, and obtaining average.
2. Finding sources with peaks higher than $\mu + 3\sigma$, and FWHM similar to the average from step 1.
3. Creating a list of the best 150 candidates for the PSF model, and manually filtering it.
4. Creating a PSF model with linear variations, a residue table and the function that best fits (Gaussian, Lorentzian, or Moffat).
5. Running PSF photometry on the entire cut.
6. Repeating steps 2 and 5 on the subtracted image, now using $\mu + 5\sigma$ as threshold.

This was done using IRAF's version of DAOPHOT (Stetson 1987).

PSF Photometry - Example



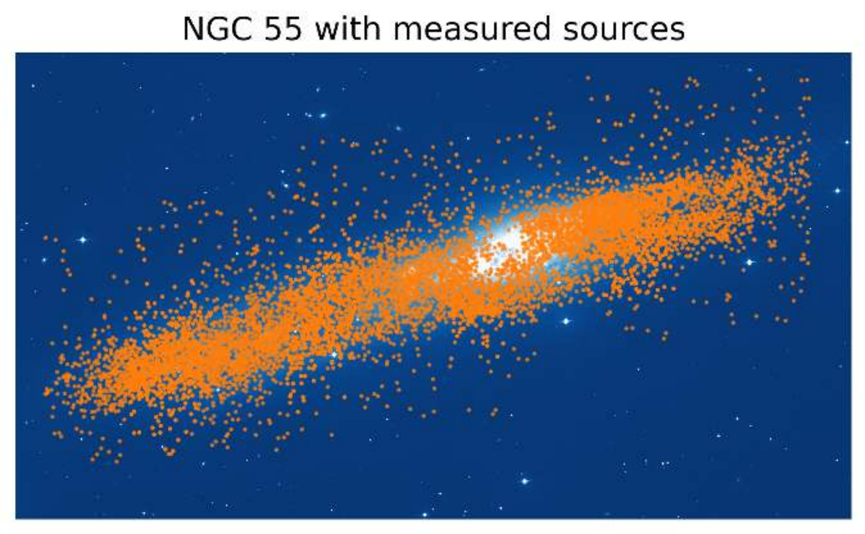
PSF Photometry - Example



DAOMATCH & DAOMASTER (Stetson 1993)

- When a stack was not possible, we matched all the measurements for the night, and obtained a final, more precise photometry.
- Then, five matches were made, one per cut.
- For the 6 nights without a stack we used the CCD chips instead of cuts.
- We kept objects that appeared in at least 20 frames, and obeyed $\sigma < 1.0$.
- Coordinate transformation included translation and rotation (6 parameters).

8756 light curves generated

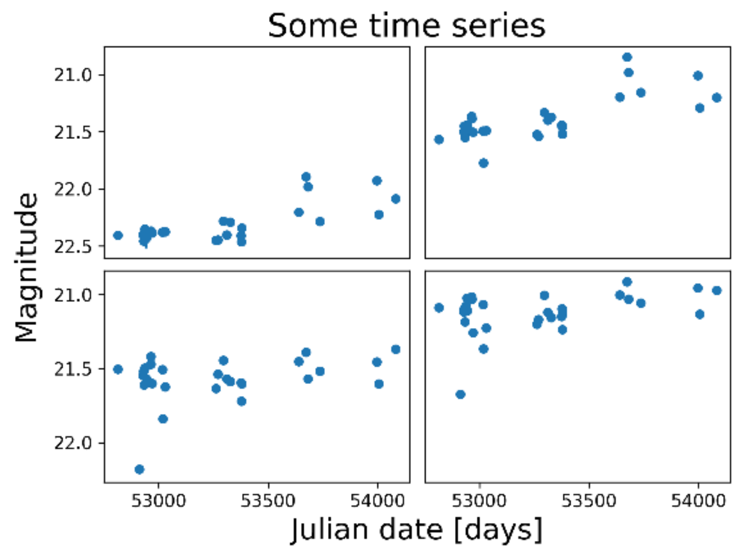


Summary of results

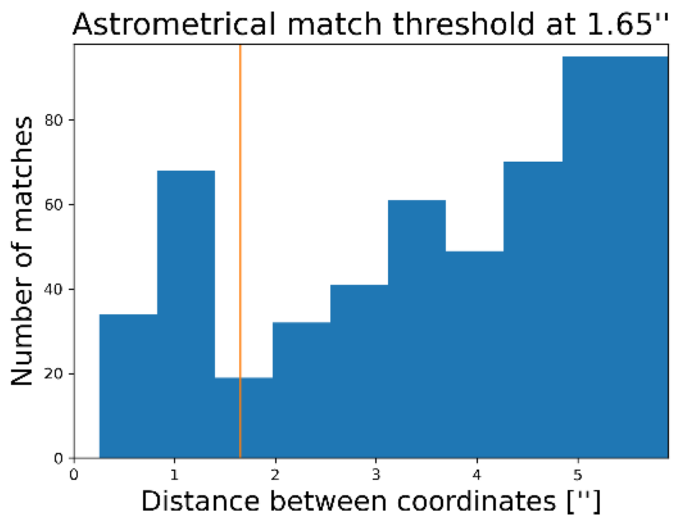
Cut	Min. detections	Max detections	No. of light curves
1	1331	7636	949
2	2371	9688	2129
3	2688	11171	1902
4	2460	9322	2092
5	2071	7508	1684
Total			8756

Table: Minimum and maximum number of detected stars on each cut, along with number of light curves generated.

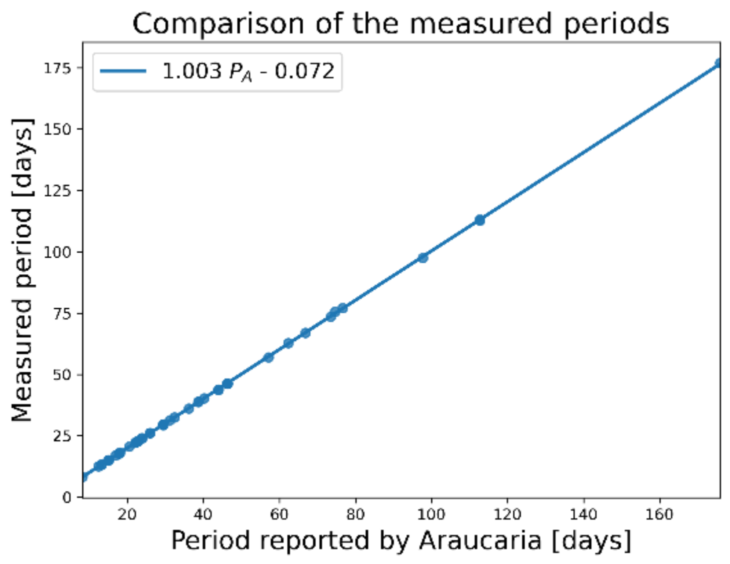
Final time series



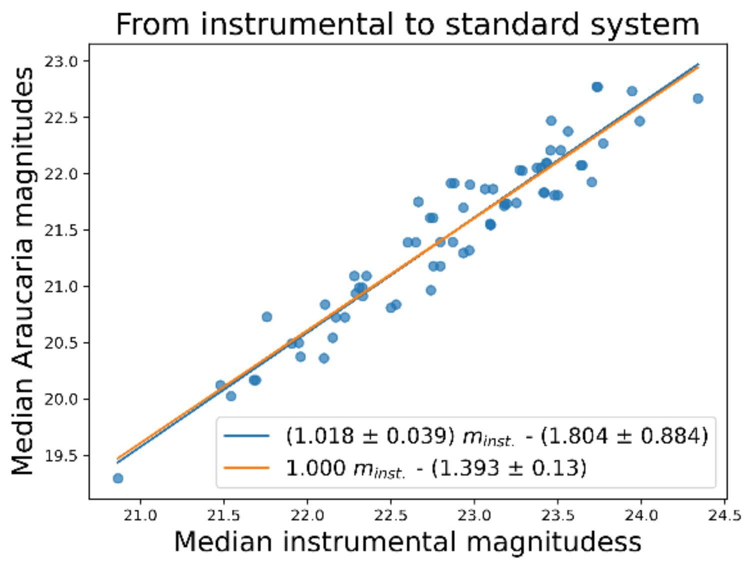
Match with Araucaria Project's previous findings (Pietrzyński *et al.* 2006).



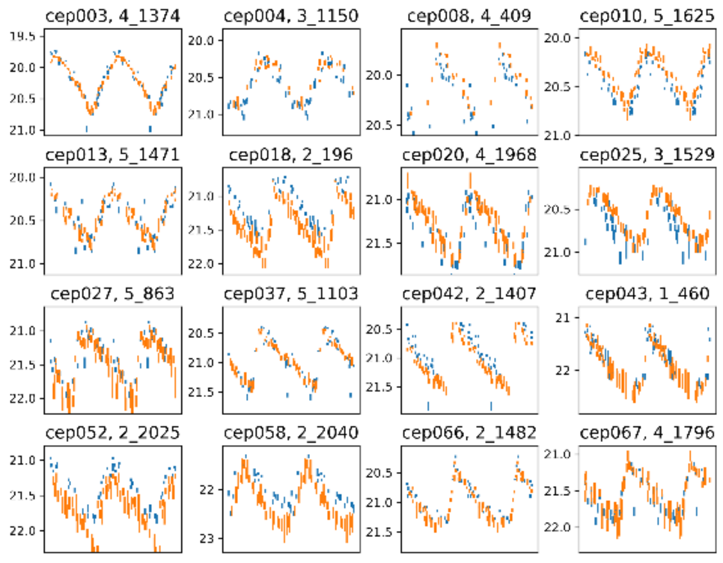
In total, 47 confirmed matches were obtained



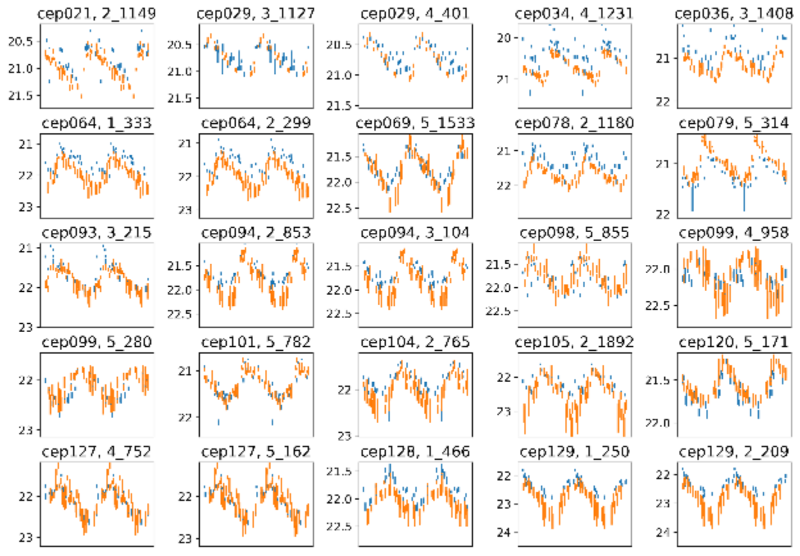
The transformation



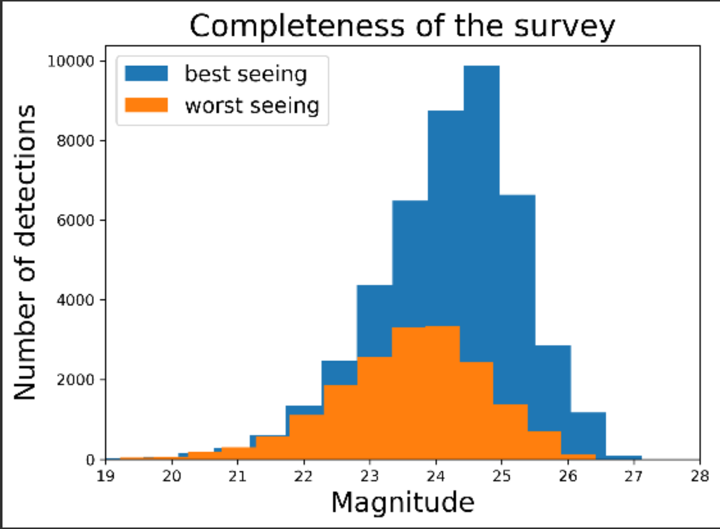
Light curves used in the calibration



Validation light curves (no colour term)



Completeness



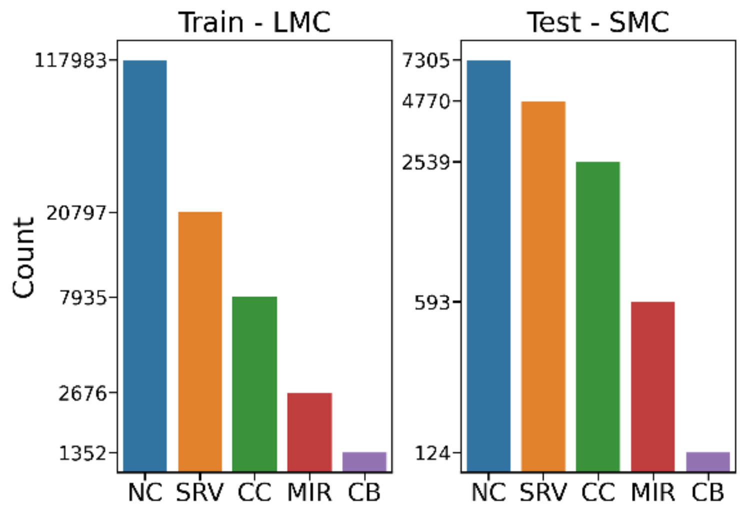
Classifying light curves

Finding periodic variables.

1. Compiling and preprocessing the training data.
2. Feature engineering.
3. Choosing a classifier algorithm.
4. Optimising hyperparameters.
5. Use the classifier to generate lists of candidates.
6. Visual inspection of the candidates and period determination.

Magellanic system as seen by OGLE

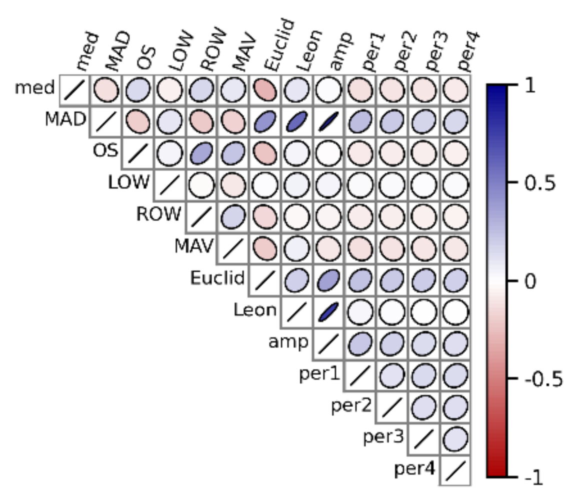
Heavy class imbalance



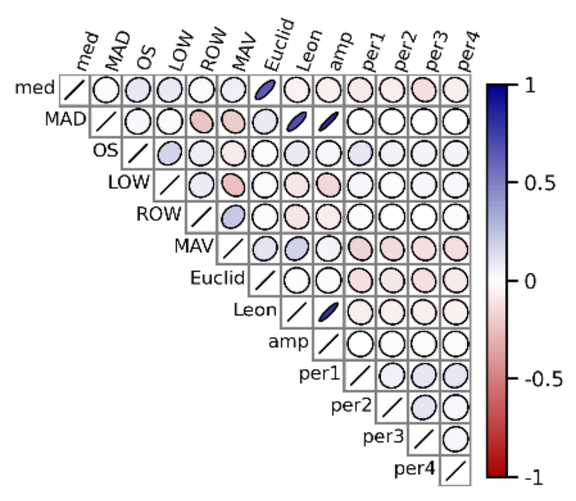
Representative features

1. Median.
2. Median absolute deviation (MAD).
3. Octile skewness.
4. Robust amplitude.
5. Left octile weight.
6. Right octile weight.
7. Modified Abbe value.
8. Average slope of successive observations.
9. Average Euclidean distance between successive observations.
10. Average of residuals after linear interpolation with adjacent observations.
11. The four most prominent Lomb-Scargle periods.

Feature correlation on OGLE data (LMC)



Feature correlation on NGC 55



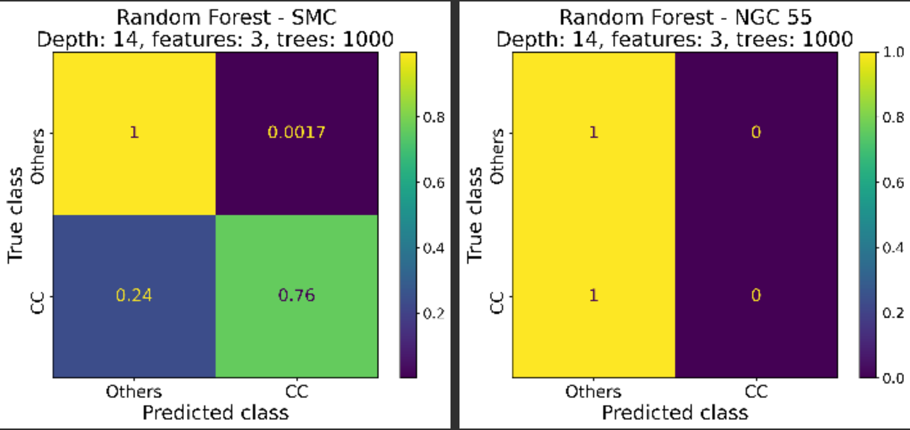
Algorithms tried

Algorithm	Implementation	Reference
Random Forest	`scikit-learn`	Breiman 2001
Bagged SVM	`scikit-learn`	Cortes *et al.* 1995
Balanced bagged SVM	`imbalanced-learn`	Cortes *et al.* 1995
Balanced Random Forest	`imbalanced-learn`	Chen *et al.* 2004
EasyEnsemble	`imbalanced-learn`	Liu *et al.* 2009
RUSBoost	`imbalanced-learn`	Seiffert *et al.* 2010
LightGBM	**Microsoft**	**Ke *et al.* 2017**

All the classifiers were written using Python 3.8

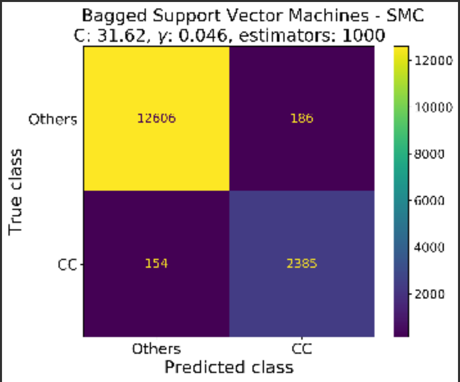
First generation: RF and Bagged SVM

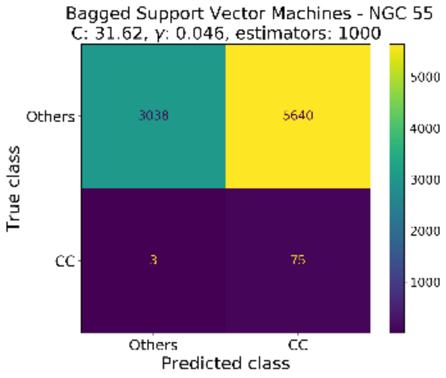
Utter failure at generalising.



Second generation: balanced bagged SVM, balanced Random Forest, EasyEnsemble, RUSBoost, and LightGBM

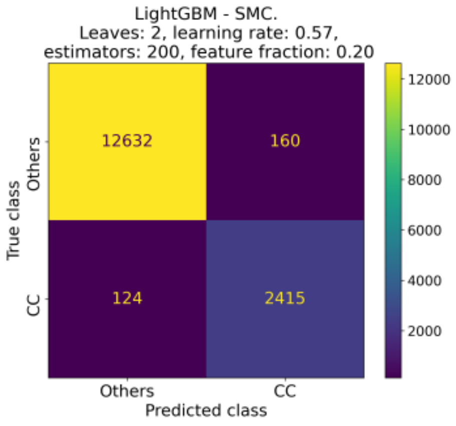
Example: Balanced bagged SVM

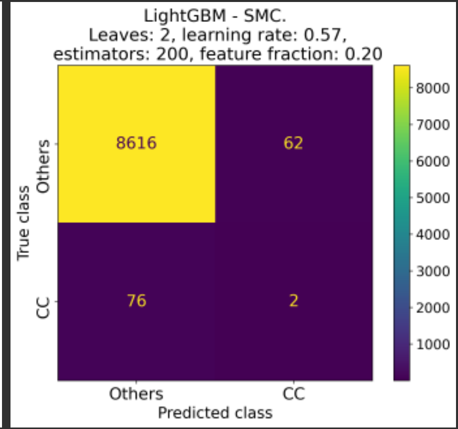




Second generation: balanced bagged SVM, balanced Random Forest, EasyEnsemble, RUSBoost, and LightGBM

Example: LightGBM





Third generation: LightGBM and Optuna

Hyperparameter search space

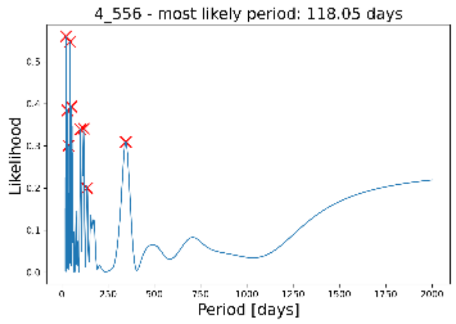
Hyperparameter	Initial distribution	Domain
Feature fraction	Uniform	[0.1, 1]
Bagging fraction	Uniform	[0.2, 1]
Number of leaves	Discrete uniform	[2, 128]
Bagging frequency	Discrete uniform	[1, 7]
Minimum child samples	Discrete uniform	[5, 100]
Number of trees	Discrete uniform	[1, 2000]
Early stopping rounds	Discrete uniform	[50, 500]
Learning rate	Logarithmic uniform	$[10^{-6}, 2]$
λ_1	Logarithmic uniform	$[10^{-6}, 2]$
λ_2	Logarithmic uniform	$[10^{-6}, 2]$

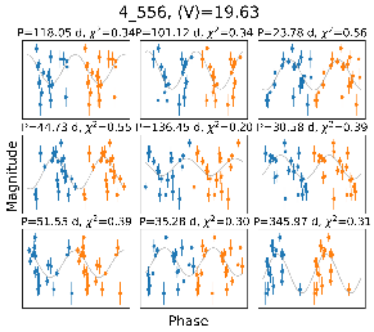
Classifier results

- 10000 trials were run, and the 10 best selected.
- We also discarded trials that suggested lists of more than 200 candidates.
- In the end, a list of 222 candidates was generated.

Manual exploration: step 1

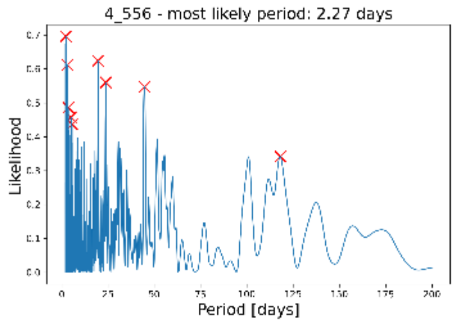
Periodogram for periods between 20 and 2000 days.

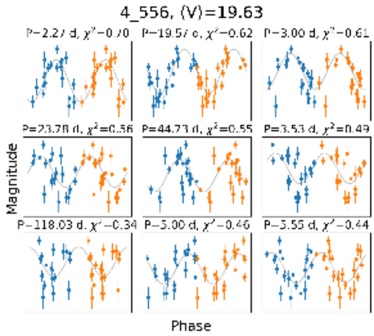




Manual exploration: step 2

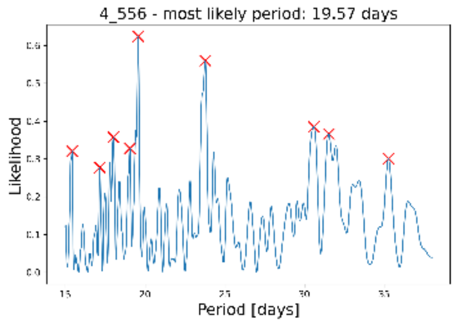
Periodogram for periods between 2 and 200 days.

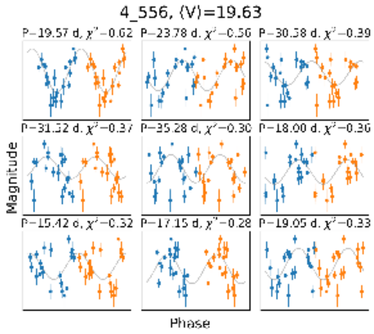




Manual exploration: step 3

Periodogram in the vicinity of the most interesting periods from the last one.



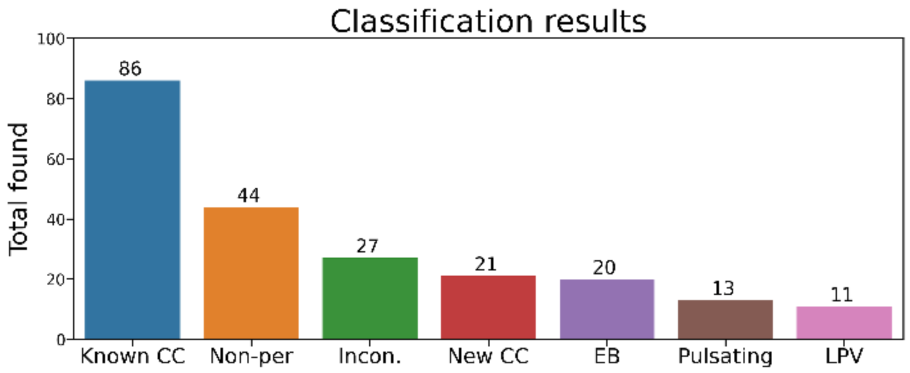


Results

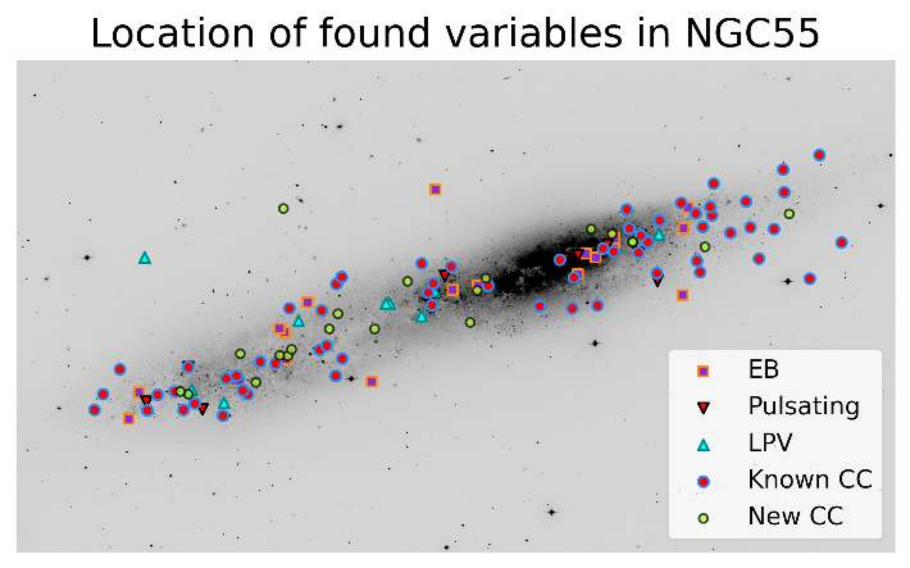
Three main results

- 8756 time series.
- 151 variable stars.
- Period-luminosity relation and distance to NGC 55.

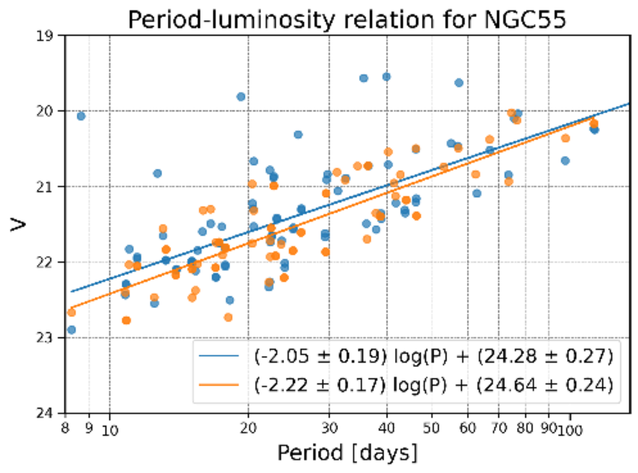
Class distribution



Sky distribution



Period-luminosity relation



Blue: This work's photometry, 2.2-m ESO/MPG telescope. **Orange**: Araucaria project, 1.3-m Warsaw telescope.

Distance determination

- Distance is determined using:
  $(V - M_V)_{\mathsf{NGC55}} = (V - M_V)_{\mathsf{LMC}} + Z_{\mathsf{NGC55}} - Z_{\mathsf{LMC}} - A_V$.
- Period-Luminosity law for LMC (Udalski 2000):
  $V_{\mathsf{LMC}} = (-2.775 \pm 0.031) \log \mathsf{P} + (17.066 \pm 0.021)$.
- Period-luminosity law for NGC 55 forcing the slope of LMC:
  $\mathsf{V}_{\mathsf{NGC55}} = (-2.775) \log \mathsf{P} + (25.12 \pm 0.11)$.

Putting it all together

$$(V - M_V)_{\text{NGC55}} = (V - M_V)_{\text{LMC}} + Z_{\text{NGC55}} - Z_{\text{LMC}} - A_V.$$

Quantity	Value [mag]	Paper
$(V - M_V)_{\text{LMC}}$	18.50	Freedman *et al.* 2001
Z_{LMC}	17.066	Udalski 2000
Z_{NGC55}	25.12	This work
A_V	0.04536	Schlegel *et al.* 1998
$(V - M_V)_{\text{NGC55}}$	26.69	This work

Method	Distance modulus	Distance [Mpc]	Paper
Planetary nebula luminosity function	26.81 ± 0.33	2.30 ± 0.35	van de Steene *et al.* 2006
Cepheid populations	26.40 ± 0.14	1.91 ± 0.13	Pietrzyński *et al.* 2006
Flux-weighted gravity-luminosity	26.85 ± 0.10	2.34 ± 0.11	Kudritzki *et al.* 2016
Cepheid populations	26.69 ± 0.11	2.16 ± 0.11	This work

Conclusions and future work

Conclusions

- 153 images taken over 29 nights were processed and 8756 light curves were generated.
- 86 out of 144 previously known Cepheids where recovered.
- 150 Variable stars were found using supervised machine learning.
- Success rate of the method was 68% (150/222).
- It is essential that the training data of the algorithms is as similar as possible to the new data (same camera, telescope and cadence).
- The methodology used works as an initial exploratory analysis, but fails to scale up.
- LightGBM outperforms all the other algorithms in classification.

Future work

- Exploring what is the minimum number of observations per light curve required to obtain a trustworthy classification, and how does it vary with the classes involved.
- Classification without explicitly defining features (it has been done before, but never with less than 100 nights per light curve).
- Generating light curves from the public data of NGC 247, NGC 300 and NGC 7793; all taken with the same instrumentation, and available at the ESO archive.

The end
Thank you

References I

- Tody, D. *The IRAF Data Reduction and Analysis System.* in *Instrumentation in astronomy VI* (ed Crawford, D. L.) **627** (Jan. 1986), 733.
- Stetson, P. B. DAOPHOT: A Computer Program for Crowded-Field Stellar Photometry. PASP **99,** 191 (Mar. 1987).
- Valdes, F. *The IRAF CCD Reduction Package - Ccdred.* in *Instrumentation for Ground-Based Optical Astronomy* (Jan. 1988), 417.
- Stetson, P. B. *Further Progress in CCD Photometry.* in *IAU Colloq. 136: Stellar Photometry - Current Techniques and Future Developments* (eds Butler, C. J. & Elliott, I.) (Jan. 1993), 291.
- Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20,** 273–297 (1995).

References II

- Schlegel, D. J., Finkbeiner, D. P. & Davis, M. Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds. ApJ **500,** 525–553. arXiv: astro-ph/9710327 [astro-ph] (June 1998).
- Valdes, F. G. *The IRAF Mosaic Data Reduction Package.* in *Astronomical Data Analysis Software and Systems VII* (eds Albrecht, R., Hook, R. N. & Bushouse, H. A.) **145** (Jan. 1998), 53.
- Udalski, A. The Optical Gravitational Lensing Experiment. Stellar Distance Indicators in the Magellanic Clouds and Constraints on the Magellanic Cloud Distance Scale. Acta Astron. **50,** 279–306. arXiv: astro-ph/0010151 [astro-ph] (Sept. 2000).
- Breiman, L. Random forests. *Machine Learning* **55,** 5–32 (Jan. 2001).

References III

- Freedman, W. L. *et al.* Final Results from the Hubble Space Telescope Key Project to Measure the Hubble Constant. ApJ **553,** 47–72. arXiv: astro-ph/0012376 [astro-ph] (May 2001).
- Chen, C., Liaw, A. & Breiman, L. Using Random Forest to Learn Imbalanced Data. *University of California, Berkeley,* 1–12 (Jan. 2004).
- SciOps, L. S. *The Wide-field Imager Handbook.* 2nd ed. (European Southern Observatory, 2005).
- Pietrzyński, G. *et al.* The Araucaria Project: The Distance to the Sculptor Group Galaxy NGC 55 from a Newly Discovered Abundant Cepheid Population. AJ **132,** 2556–2565. arXiv: astro-ph/0610595 [astro-ph] (Dec. 2006).

References IV

- van de Steene, G. C., Jacoby, G. H., Praet, C., Ciardullo, R. & Dejonghe. Distance determination to NGC 55 from the planetary nebula luminosity function. A&A **455,** 891–896 (Sept. 2006).
- Liu, X.-Y., Wu, J. & Zhou, Z.-H. Exploratory Undersampling for Class-Imbalance Learning. *Trans. Sys. Man Cyber. Part B* **39,** 539–550. ISSN: 1083-4419. https://doi.org/10.1109/TSMCB.2008.2007853 (Apr. 2009).
- Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J. & Napolitano, A. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *Trans. Sys. Man Cyber. Part A* **40,** 185–197. ISSN: 1083-4427. https://doi.org/10.1109/TSMCA.2009.2029559 (Jan. 2010).
- Kudritzki, R. P. *et al.* A Spectroscopic Study of Blue Supergiant Stars in the Sculptor Galaxy NGC 55: Chemical Evolution and Distance. ApJ **829,** 70. arXiv: 1607.04325 [astro-ph.GA] (Oct. 2016).

References V

Ke, G. *et al.* in *Advances in Neural Information Processing Systems 30* (eds Guyon, I. *et al.*) 3146–3154 (Curran Associates, Inc., 2017). http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.

Stacked image