MNRAS **000**, 1–8 (2019) Preprint 30 September 2019 Compiled using MNRAS LATEX style file v3.0

Unsupervised Machine learning for Transient Discovery and Anomoly Detection from Fast Cadenced Optical Light Curves.

Sara Webb[1,2]★, Michelle Lochner[3], Daniel Muthukrishna[4], Jeff Cooke[1,2,5], Simon Goode[1,2] , Chris Flynn[1,2]

[1]*Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Mail Number H29, PO Box 218, 31122, Hawthorn, VIC, Australia*
[2]*ARC Centre of Excellence for Gravitational Wave Discovery (OzGrav), Australia*

Last updated 2019 September 22

ABSTRACT

This is a guide for preparing papers for *Monthly Notices of the Royal Astronomical Society* using the `mnras` LATEX package. It provides instructions for using the additional features in the document class. This is not a general guide on how to use LATEX, and nor does it replace the journal's instructions to authors. See `mnras_template.tex` for a simple template.

Key words: editorials, notices – miscellaneous

1 INTRODUCTION

In the era of large current and upcoming time-domain surveys, classification and discovery of optical transient sources will begin to become reliant on machine classification to handle the large amounts data being collected. Current ground based surveys such as the Zwicky Transient Facility (ZTF), Dark Energy Survey (DES) and the All Sky Automated Survey for Supernovae (ASAS-SN) are able to scan thousands of square degrees continuously amounting in terabytes of data annually, and recently the Panoramic Survey Telescope and Rapid Response System Survey (Pan-STARRS) delivered the first petabyte scale optical data release ((Bellm et al. 2018; Collaboration: et al. 2016; Shappee et al. 2014; Stubbs et al. 2010; Chambers et al. 2016). While space based time-domain missions have provided unprecedented photometry, light curves and proper motions for galactic sources, with Kepler & K2 targeting ~400,000+ individual stars, TESS is expected to target at least 200,000 targets of the ~9.5 million targets in TESS input catalog, and Gaia is already releasing almost 2 billion sources (Borucki et al. 2010; Howell et al. 2014; Stassun et al. 2018). The importance of being able to mine these increasing amounts of data to not only identify known transients but make discoveries of new or anomalous sources is paramount to the success of future transient astronomy.

Supervised machine learning has been utilised already by several surveys and teams in astronomy for source identification over large data sets. A large majority of the work to date has been in the identification of variable stars and quasi-stellar objects from light curves via multivariate Gaussian mixture models, random forest classifiers, support vector machines, or Bayesian networks (Debosscher et al. 2007; Richards et al. 2011; Pichara et al. 2012; Bloom et al. 2012; Kim & Bailer-Jones 2016; Kim et al. 2011; Mackenzie et al. 2016; Pichara & Protopapas 2013). The work aforementioned all successfully shows the power in machine classification of sources via trained algorithms, and explores the most successful feature properties for both folded and unfolded light curves with the most common features being compiled into the python FATS package by Nun et al. (2015). Features are able to express given time-series data as a set of normalised values, each representing a measurable property or characteristic of light curve. Classification of non-folded light curves of extragalactic transient sources has also been explored, moving away from traditional template fitting to computationally more robust supervised and semi-supervised techniques also requiring feature extraction (Richards et al. 2011; Karpenka et al. 2012; Lochner et al. 2016; Narayan et al. 2018; A.Möller et al. 2016).

What is limiting by the current supervised techniques is the need for full light curves (non real-time) and prior knowledge of transient classes for network training. Work towards real-time classification of supernovae by Muthukrishna et al. (2019) and Möller & de Boissière (2019) have shown the power in deep Recurrent Neutral Networks (RNNs), run on Graphics Processing Units (GPUs), and their ability to provide fast real-time classifier that don't rely on extracting computationally expansive features of the input data, only requiring features of training data. Even with the advances of machine in astronomy, mining data for unknown or anomalous events is relatively unexplored, as the majority of current networks work off pre-trained algorithms. Work by Mackenzie et al. (2016) developed an unsupervised feature learning algorithm which takes subsections of variable light curves to cluster and use as features to train a linear support vector machine. This work is the first of its kind to eliminate the need of traditional feature extraction, limiting the computational burden and biases of feature selection. Only limited work into actual transient classification or anomaly detection via unsupervised means has been preformed. Unsupervised clustering

★ Contact e-mail: webb.sara.a@gmail.com

of variable star light curves was preformed by Valenzuela & Pichara (2018) by creating variability trees using k-medoids clustering algorithm of fragmented light curves. They show this method was able to be used to create an unsupervised variability tree, backtrace known light curves to determine their individual tree structure and then preform similarity searches of unknown light curves. This method offers a novel and computationally fast approach to data exploration but is again limited by the need of known light curve examples for similarity searches. Giles & Walkowicz (2019) again utilised clustering of light curves in full, using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to cluster Kepler light curves to identify outliers for visual inspection. They showed the successful extraction of the known anomalous Boyajian's star through their method, however the limitations of DBSCAN assuming clusters to have constant density was identified Giles & Walkowicz (2019). It should be noted that currently all work preformed on light curve classification in astronomy have used long cadenced light curves, spanning from 30 minute cadence to several day cadence with folded light curves.

Currently, the majority of wide field surveys explore a limited region of the luminosity-timescale phase space, with an average cadence of hours to days between visits to fields, with only few programs exploring the phase space shorter then 1 hour cadence (see, Roykoff et al. (2005); Rau et al. (2009); Lipunov et al. (2004, 2007); Berger et al. (2014)). What is largely unexplored by these surveys is the phase space of transient events occurring on seconds to minutes time scales, events which are often referred to as 'foreground fog' in traditional surveys. There are several events expected to occur over seconds to minutes and understanding the transient Universe on these timescales is crucial for understanding the transient Universe as a whole. Take the upcoming Large Synoptic Survey Telescope (LSST), each night upwards of 1 million alerts are predicted to be generated, it will be invaluable to be able to meaningfully quantify the expected volume of short timescale events will assist in follow-up priorities (LSST Science Collaboration et al. 2009). Using fast cadenced light curves from the Deeper, Wider, Faster program we will be exploring this phase space

2 THE DEEPER, WIDER, FASTER PROGRAM

Several new and interesting astronomical fast transient events have been discovered in recent decades and the progenitors and physical mechanisms behind many of them are still relatively poorly known (eg. Fast Radio Bursts (FRBs), supernova shock breakouts and other rapidly evolving extragalatic events (see Lorimer et al. (2007); Garnavich et al. (2016); Prentice et al. (2018); Perley et al. (2018) respectively). What has limited our ability to detect and understand these events is the capability to gather data in short, regular time intervals before, during and after the events as well as over a range of wavelengths. The Deeper, Wider, Faster program (DWF Andreoni & Cooke (2018)) has been designed with these challenges specifically in mind, constructing a multi-wavelength and simultaneous observational program, of over 30 facilities to date [1]. DWF takes a 'proactive' approach to transient astronomy, with multi-wavelength observations of the target fields taken continuously over 1-3 hour periods, capturing before, during and after many of the transient events. DWF unitises facilities with large field of views, targeting three square degrees at once using the Dark Energy Camera (DECcam) as our primary imager, taking continuous 20 second exposures. Using DECam, DWF has an image cadence 63 times higher then the Deep Lens Survey from the early 2000s (Wittman et al. 2002) to a similar magnitude limit, and cadence at least 45 times higher then the on-going Zwicky Transient Facility and surveying 6 magnitudes deeper (Kulkarni & Rau 2006). From our real-time processing, we are able to rapidly identify candidates and coordinate rapid-response and long term follow-up of candidates. DWF was begun in 2014 and since its inception has had two commissioning runs and six operational runs (see **?**, Cooke et al., in prep).

The unique design of DWF allows exploration of transients on the second-to-minute timescales, providing further understanding into the classes of already observed fast transient events as well as exploring events theorised to occur on these timescales. The optical component of DWF is able to explore a region of parameter space not yet reached by previous transient surveys, by taking continuous, high-cadenced 20 second exposures, imaging with the wide-field sensitive Dark Energy Camera (DECam) on the 4m Blanco telescope in Chile or 30 second exposures using the Hyper-SurprimeCam (HSC) imager on the 8m Subaru telescope in Hawaii. Note. This work will only focus on the data gathered by DECam. Work by Andreoni et al. (2019) utilised the unique DWF data and the *'Mary'*, our transient difference image discovery pipeline to constrain extragalatic transients on the minute timescales. In this work, we will be examining light curves generated purely from science images for all sources in our chosen fields, and exploring the ability to identify known and unknown transient and variable sources through the use of unsupervised machine learning. By examining every source light curve through an unsupervised network, we aim to not only distinguish clear separations of sources in feature space but identify and classify unknown and outlier sources to comprehensively explore transient event and source variability on the minutes-to-hours timescale.

3 DATA

This work will explore unsupervised methods applied to light curves generated from the fast cadenced data collected during DWF runs using DECam. As aforementioned previously, we collect 20 second continuous imaging of targeted fields, acquired in a single band, the *'g'* filter. We choose the continuous use of the *'g'* filter to maximize depth with DECam, reaching ~0.5 deeper in comparison to the other filters. With an average seeing on 1.0 arcseconds and airmass of 1.5 (relatively high airmass due to the field constraints of simultaneous multi-facility observations), the expected limiting magnitude in *'g'* band is m(AB) ~ 23. The DECam images are post-processed through the NOAO High-Performance Pipeline System (Valdes & Swaters 2007; Swaters & Valdes 2007; Scott et al. 2007) and then transferred to the OzStar supercomputer at Swinburne University of Technology for our data analysis. The DECam 62 CCD mosaic is separated into individual fits files for each extension. Each CCD is processed separately for source extraction using SExtractor and all source magnitudes are corrected for exposure time and magnitude offsets against the SkyMapper Data Release 2 catalogue Bertin & Arnouts (2010); Onken et al. (2019). A master source list is compiled by cross matching all extracted sources from each CCD, over all exposures within 0.5 arcsecond radius into one catalogue of source positions. This master catalogue is used to to create light curves for each source, with any non-detections in single exposure

[1] http://bit.do/DWF

having the CCD exposure detection upper limit represented in the light curve.

In totality DWF has targeted 15 separate fields, accumulating over 1 million source detections and 6 million nightly light curves. In this work we will be analysing light curves gathered over 1.5 hours of continuous observation on the Antlia field, as proof of concept of unsupervised clustering for transient and anomaly detection. The data was collected on February 6^{th} 2017, using a field centre of RA: 10:30:00.0 DEC:-35:20:00.0 on the Antlia cluster of galaxies.

4 METHODOLOGY

In this section we will outline our methodology across several stages, outlined in figure X below and explained in the following sections. Note that all stages are preformed on nightly light curves with an average cadence of 68 seconds between light curve points, accounting for both exposure and readout. We utilize python for all stages, using the following packages scikit-learn, hdbscan, FATS, astropy, numpy, pandas and matplotlib (Pedregosa et al. 2011; McInnes et al. 2017; Nun et al. 2015; Price-Whelan et al. 2018; Oliphant 2015; McKinney et al. 2010; Hunter 2007). In this paper we will present this method applied to DWF light curves of sources from the Antlia field on the night of February 5th 2017.

4.1 Features

To represent our unique fast-cadenced data, we use a mixture of features developed and used primarily for the identification of variable stars and quasi-stellar objects. We extract 25 unique features from each light curve using both the FATS python package and manual calculation (see Appendix A for the full feature list). The majority of our features were first presented in the work by Richards et al. (2011) in classifying variable stars from sparse and noisy time-series data, We chose the features from this work that were not specifically restricted to folded light curves or periodic sources, these features being amplitudes, Beyond 1 standard deviation, linear trend, maximum slope, median absolute deviation, median buffer range parentage, pair slope trend, range of cumulative sum, skew, small kurtosis, standard deviation, see Richards et al. (2011) figure 8 for more detail. The other features we have taken from stellar variability detection are focused around Fourier decomposition, giving in H_1 (amplitudes), R_{21} (ratio of 2nd to 1st amplitude) and R_{31} (ratio of 3rd to 1st amplitude). The remaining features were taken from work in quasi-stellar object selection, these being auto correlation length, consecutive points, variability index and Stetson K_{AC} as used by Kim et al. (2011) and mean, σ and τ taken form an continuous auto regressive model fitted to our data from Pichara et al. (2012).

4.2 HDBSCAN

We utilise the newly released python library[2] by McInnes et al. (2017) to implement Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), a method first proposed by Campello et al. (2013). HDBSCAN takes the approach of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and converts it into a hierarchical clustering algorithm by varying the value of epsilon to identify clusters of varying densities, see McInnes et al. (2017). To better understand how HDBSCAN works we will first outline the original DBSCAN algorithm by Ester et al. (1996). DBSCAN is able to take a set of given points perform nearest neighbour searches in a given feature space to determine clusters of over densities, points closely related in distance, and identify outlier points that exist in low density regions as noise. DBSCAN requires two parameters, epsilon (ϵ), which represents the radius of the neighbourhood search, and minimum number of points ($minPts$), which must exist in a neighbourhood to constitute a dense region. What has limited the use of DBSCAN in the past is the inability to vary ϵ in a given data set, requiring clusters to have similar densities, however, HDBSCAN is able to implement changing values of ϵ to successfully explore clusters of varying densities. We opt to limit the restrictions on cluster decision making when using HDBSCAN using parameters as shown in table X. We aim to create as many distinct clusters in our feature space as the algorithm will allow to maximize our classification of the quality and events present in the data and limit the outliers to only data points in very low density regions.

Parameter	Setting
Algorithm	Best
Allow single cluster	False
Alpha	1.0
Cluster selection method	eom
Core distance n jobs	4
Leaf size	40
Metric	Euclidean
Min cluster size	5

Table 1. Parameter settings for HDBSCAN algorithm for our analysis

4.3 t-SNE

To help visualise the clustering of objects in our high dimensional feature space we use the T-distributed Stochastic Neighbor Embedding (t-SNE) algorithm developed by van der Maaten & Hinton (2008). The t-SNE algorithm uses the same Euclidean distance metric to measure the proximity of all features in higher dimensional space and converts these distances to probabilities using a normal distribution. A similarity matrix of the probabilities is stored for the higher dimensional space, and the feature space is then randomly collapsed down to either 2 or 3 dimensions where the Euclidean distance is calculated once again using a t-distribution to assign probabilities and saved as a second similarity matrix. The two matrices are then minimized using the sum of Kullback-Leibler divergence of all data points using a gradient descent method to return a 2 or 3 dimensional representation of the distance of data in our feature space.

4.4 Astronomaly

To explore our sub clusters for the most anomalous light curves we use the python package Astronomaly which is comprised of a python back end and JavaScript front end to easily explore the data (Lochner et al in prep 2019). We run Astronomaly on each cluster, and unitize the inbuilt isolation forest algorithm, using our already calculated features, to then visually inspect the highest ranking anomalous light curves, as well as the interactive t-SNE plot to explore the lower dimensional cluster space.

[2] https://hdbscan.readthedocs.io/en/latest/

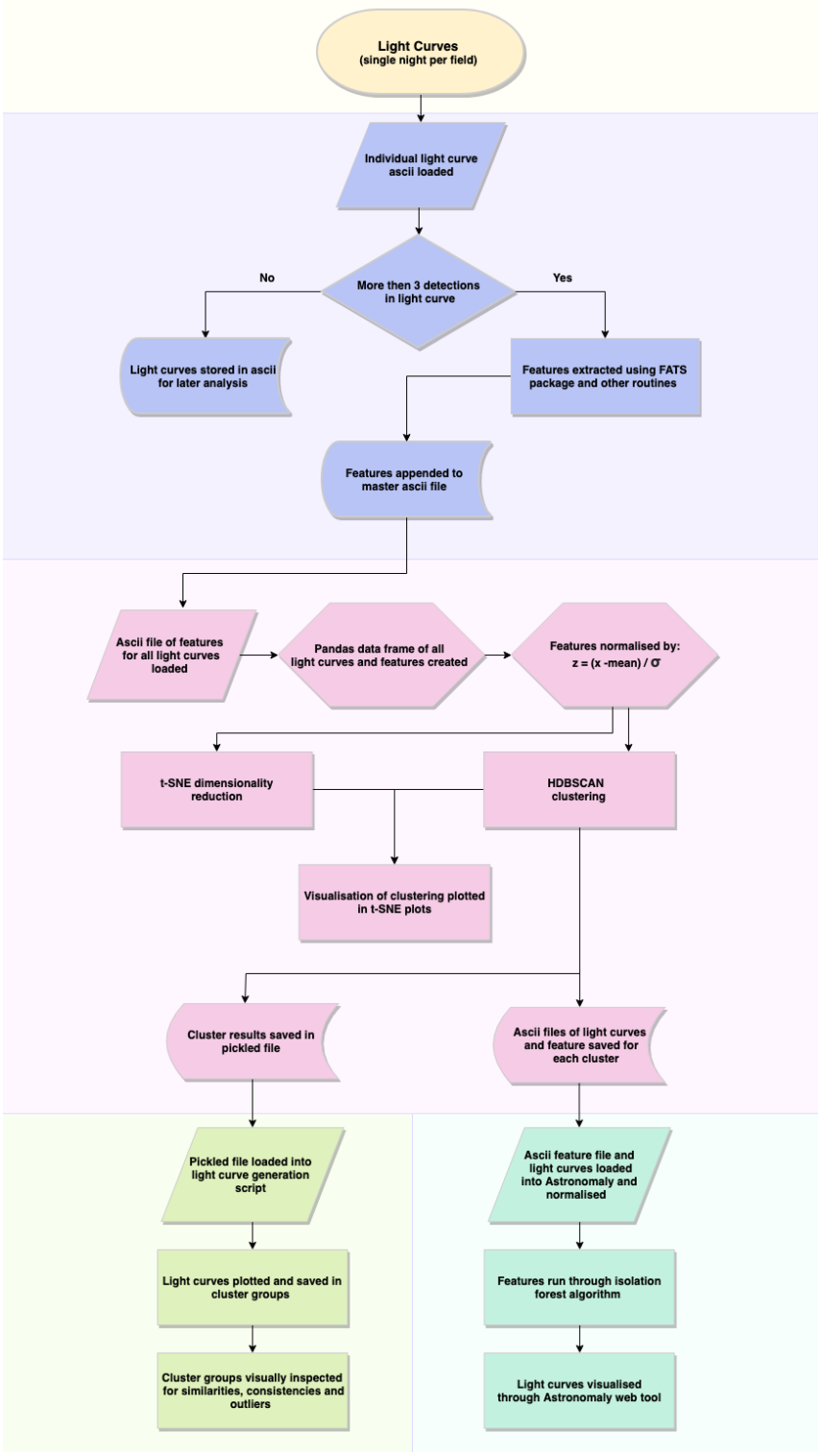

Figure 1. Workflow of light curve analysis, broken in to four stages. The first stage, indicated by purple, is feature extraction. Each light curve is looped through this process to write a master ascii file of all light curve IDs and features. The next stage, indicated by pink, is the unsupervised clustering and visualisation using t-SNE dimensionality reduction and HDBSCAN clustering. The clusters are saved into two separate outputs, the first being a pickled data frame of all light curve IDs and cluster assignments ascii feature files for each cluster, and the second an ascii feature files for each cluster. The final stages are light curve visualisation, the first, shown in blue, is outlier identification in each cluster using Astroanomaly. Each cluster set is loaded and an isolated forest is computed across all features to show the most anomalous light curves of the clusters via the interactive web GUI. The second stage of visualisation, shown in green is the generation of all light curve plots. These plots are able to be scanned through rapidly to validate the clustering success as well as identify outlier light curves from the identified noise.

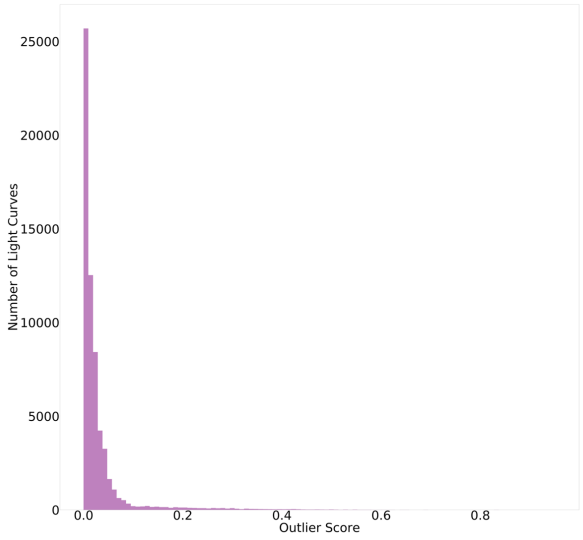

Figure 2. Histogram of outlier scores for 170206 Antlia light curves, note the majority of light curves have very low outlier scores below 0.1, indicating that the majority of light curves exist in similar feature space.

5 RESULTS

We started with a total of 70350 sources identified in the Antlia field from the five night master source list complied, of these, 61844 light curves met the criterion of having $N_{det} > 3$. The chosen 25 features were extracted from each of the 61844 light curves. The light curves contain photometric information of sources from 80 exposures over a 1.5 hour period. A total of 29 clusters were Identified through the HDBSCAN clustering algorithm, as well as a grouping of noise, light curves that did not satisfy the distance requirements to join the identified clusters. See Appendix B for individual cluster information. Using HDBSCAN the outlier scores for each light curve using the GOSH outlier algorithm, and from these results we are able to firstly verify that the vast majority of our data have lower outlier scores, see figure 2, with ~90% of light curves below a score of 0.1.

5.1 Cluster Sub Grouping

The 29 clusters can be broken down into 12 sub groups shown in table 2. The majority of clusters fall into the sub groups of photometic anomalies caused by telescope dithering, ccd artifacts/cosmic rays or photometric correct issues. However two very clear sub groupings, A and B, separated out variable sources discussion of which will be followed in section 5.2. Representation of the clusters in feature space can been seen in figures 3 and 4, where the feature space has been reduced into 2 and 3 dimensions respectively using t-SNE. The t-SNE plots clearly show the feature space is dominated by one main cluster (number 28), which is unsurprising, as we expect the majority of sources in the field to be unchanging over the minutes to hours time scales. Figure 3 further shows the grouping of clusters with related light curves by highlighting the sub groups containing more then one cluster. From the sub grouping of clusters, we are able to meaningfully quantify the our light curves for this field, finding that 91% are grouped into one cluster, of seemingly unchanging sources, while 1.6% of light curves were affected by telescope dithering and/or the use of the hexapod[3] on the DECam instrument, and only 0.56% of light curves had photometric correction issues over the first 5 exposures (of the 80) due to the initial 5 point dither pattern. We further identify 24 light curves which have been contaminated by cosmic rays/hot pixel spots during only one exposure throughout the night, and ~ 5% of light curves and 12 light curves of faint sources below our detection threshold for majority of exposures.

5.2 Sub groups identifying Variable Sources

Two sub groups of light curves showing variability were identified in groups A and B, with A showing 7 clear variable sources, and B showing 9 noisier light curves with possible astrophysical variability as well as possible variability caused by photometric correction . Each identified source was cross matched to the International Variable Star Index (VSX) catalogue (Watson et al. 2006). From group A we found that 4/7 identified sources were cataloged as either eclipsing binaries, RR lyrae or spotted stars with periods ranging between 0.27 - 0.45 (days). Of the remaining three source, two appear to be steadily rising between 0.25-0.50 magnitudes over the 1.5 hours of exposure, not unlike the other identified variables. The remaining one uncatalogued variable source appears to be varying with a period of ~ 1.5hours, fluctuating ~ 0.15 magnitudes from peak to trough.

5.3 Un-clustered Noise

3270 light curves which were identified as noise and not assigned to a specific cluster. See Appendix C for individual cluster details.

WORKING ON Using Astronomaly to visually inspect the light curves classified as noise, additional variable sources were identified. See figure five for a subset of identified varying sources over a range of magnitudes. It's interesting to note the variable sources that were grouped as noise have larger variance in their magnitude change then those grouped in cluster 1.

We chose to recover the 5 RR lyrae stars in our field of view, as identified in Gaia data release 2, to determined the periodic timescales of variable sources which are able to be identified and discovered using this method.

6 CONCLUSION

WORKING ON

REFERENCES

A.Möller et al., 2016, Journal of Cosmology and Astroparticle Physics, 2016, 008

Andreoni I., Cooke J., 2018.

Andreoni I., et al., 2019, arXiv e-prints, p. arXiv:1903.11083

Bellm E. C., et al., 2018, Publications of the Astronomical Society of the Pacific, 131, 018002

Berger E., et al., 2014, ApJ, 779

Bertin E., Arnouts S., 2010, SExtractor: Source Extractor (ascl:1010.064)

Bloom J. S., et al., 2012, Publications of the Astronomical Society of the Pacific, 124, 1175

[3] The Hexapod mechanism is a set of six pneumatically driven pistons that actuate to precisely align the optical elements between exposures.

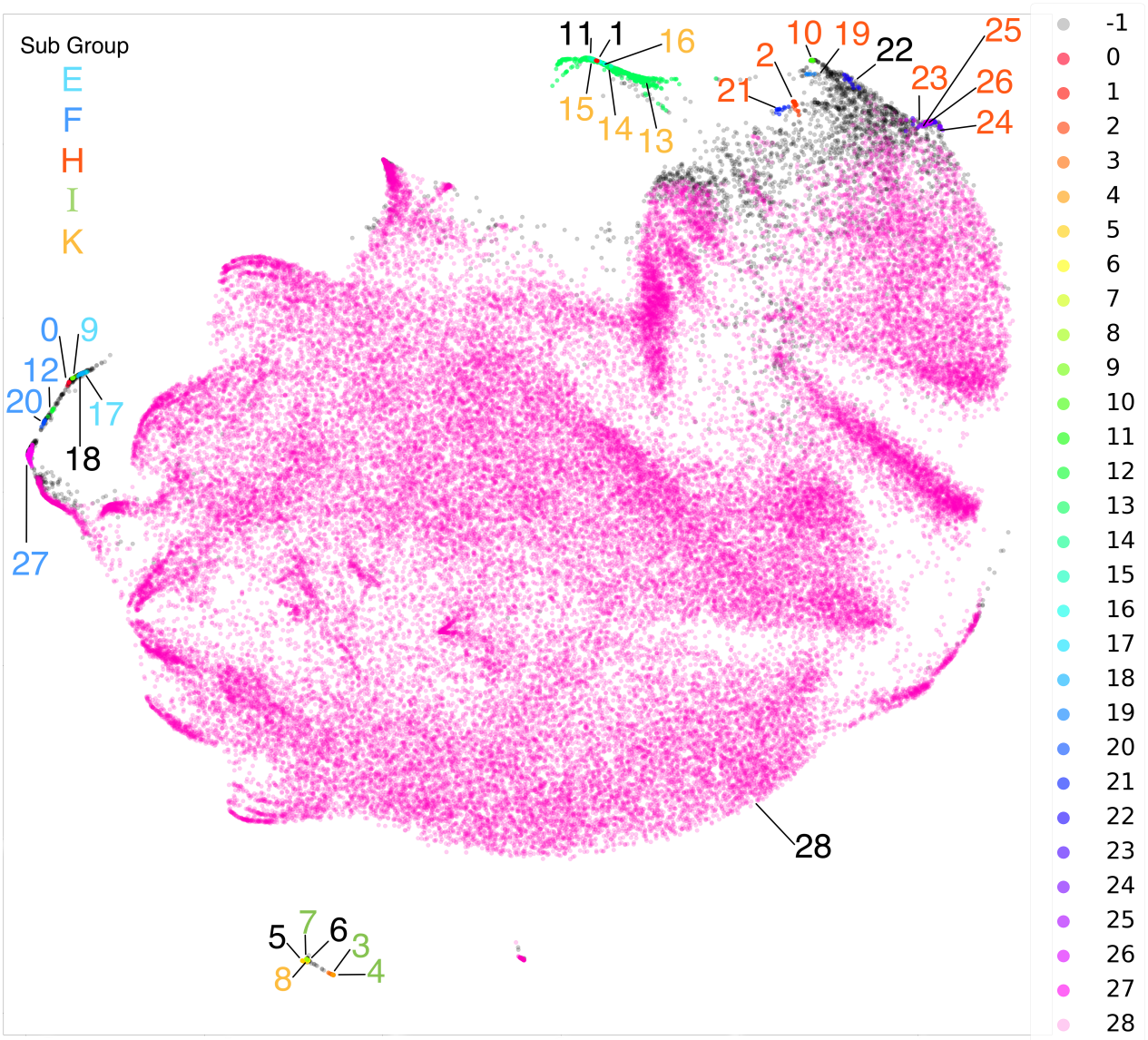

Figure 3. Feature space of light curves collapsed down to 2 dimensions using t-SNE with each cluster identified by HDBSCAN coloured individually. The sub groups containing more than one cluster have their labels colored with the corresponding sub group. Sub group E is light blue, F is blue, H is red, I is green and K is orange. Those clusters indicated by black numbers are belonging to their own sub group.

Borucki W. J., et al., 2010, Science, 327, 977

Campello R. J. G. B., Moulavi D., Sander J., 2013, in Pei J., Tseng V. S., Cao L., Motoda H., Xu G., eds, Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 160–172

Chambers K. C., et al., 2016, arXiv e-prints, p. arXiv:1612.05560

Collaboration: D. E. S., et al., 2016, Monthly Notices of the Royal Astronomical Society, 460, 1270

Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, A&A, 475, 1159

Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. KDD'96. AAAI Press, pp 226–231, http://dl.acm.org/citation.cfm?id=3001460.3001507

Garnavich P. M., et al., 2016, ApJ, 820

Giles D., Walkowicz L., 2019, MNRAS, 484, 834

Howell S. B., et al., 2014, PASP, 126, 398

Hunter J. D., 2007, Computing in Science & Engineering, 9, 90

Karpenka N. V., Feroz F., Hobson M. P., 2012, Monthly Notices of the Royal Astronomical Society, 429, 1278

Kim D.-W., Bailer-Jones C. A. L., 2016, A&A, 587, A18

Kim D.-W., Protopapas P., Byun Y.-I., Alcock C., Khardon R., Trichas M., 2011, The Astrophysical Journal, 735, 68

Kim D.-W., Protopapas P., Bailer-Jones C. A. L., Byun Y.-I., Chang S.-W., Marquette J.-B., Shin M.-S., 2014, A&A, 566, A43

Kulkarni S., Rau A., 2006, ApJ Letters, 644

LSST Science Collaboration et al., 2009, arXiv e-prints, p. arXiv:0912.0201

Lipunov V. M., et al., 2004, ApJ, 611

Lipunov V. M., et al., 2007, Astronomy Reports, 51

Lochner M., McEwen J. D., Peiris H. V., Lahav O., Winter M. K., 2016, ApJS, 225, 31

Lorimer D., et al., 2007, Science, 318

Mackenzie C., Pichara K., Protopapas P., 2016, ApJ, 820, 138

McInnes L., Healy J., Astels S., 2017, The Journal of Open Source Software, 2

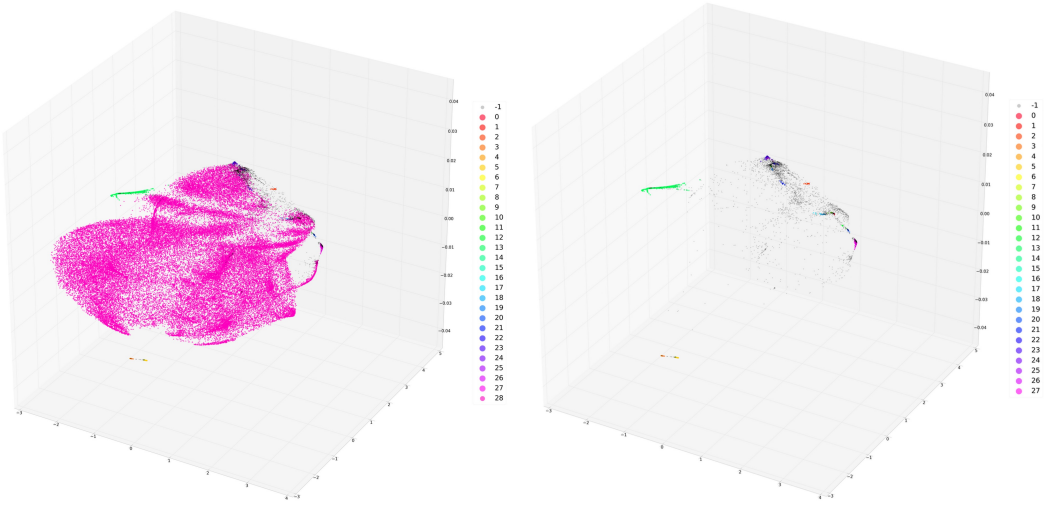

Figure 4. Left: Feature space of light curves collapsed down to 3 dimensions using t-SNE with each clusters identified by HDBSCAN coloured individually. Its apparent the most predominate cluster is number 28, shown in fuchsia, which occupies the majority of space. Right: The same plot as to the left with the main cluster removed, highlighting the noise, in black, and identified clusters in color. The orange to red smaller clusters, although the smallest, have the largest apparent distance to the main cluster.

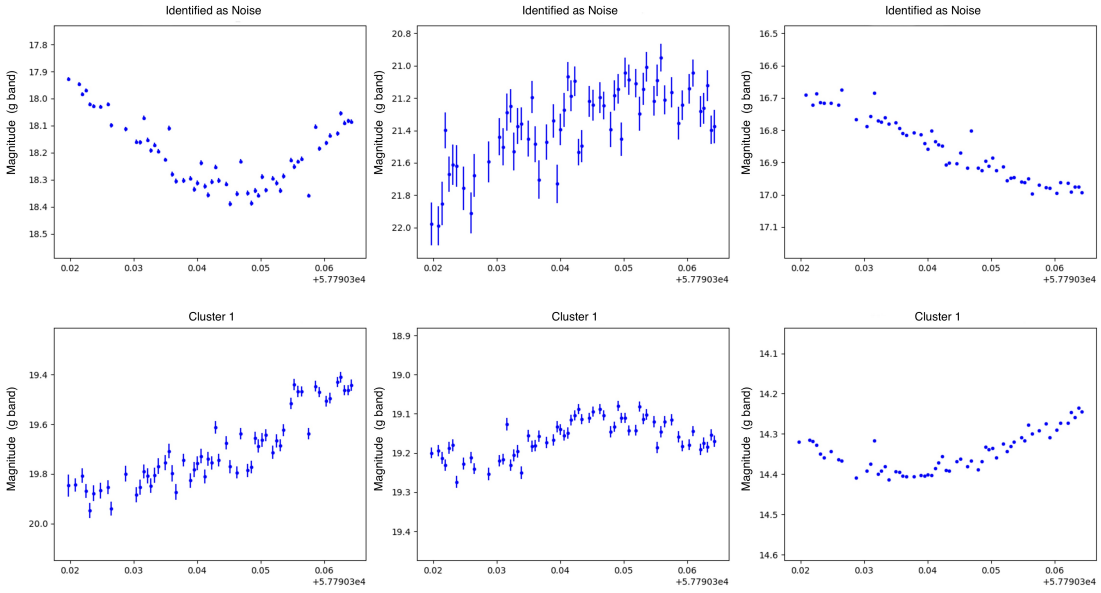

Figure 5. Identified variable sources from Antlia observation from 170206

McKinney W., et al., 2010, in Proceedings of the 9th Python in Science Conference. pp 51–56

Möller A., de Boissière T., 2019, arXiv e-prints, p. arXiv:1901.06384

Muthukrishna D., Narayan G., Mandel K. S., Biswas R., Hložek R., 2019, arXiv e-prints, p. arXiv:1904.00014

Narayan G., et al., 2018, The Astrophysical Journal Supplement Series, 236, 9

Nun I., Protopapas P., Sim B., Zhu M., Dave R., Castro N., Pichara K., 2015, arXiv e-prints, p. arXiv:1506.00010

Oliphant T. E., 2015, Guide to NumPy, 2nd edn. CreateSpace Independent Publishing Platform, USA

Onken C. A., et al., 2019, Publ. Astron. Soc. Australia, 36, e033

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Perley D. A., et al., 2018, Monthly Notices of the Royal Astronomical Society, 484, 1031

Pichara K., Protopapas P., 2013, The Astrophysical Journal, 777, 83

Pichara K., Protopapas P., Kim D. W., Marquette J. B., Tisserand P., 2012, MNRAS, 427, 1284

Prentice S. J., et al., 2018, ApJ, 865, L3

Price-Whelan A. M., et al., 2018, AJ, 156, 123

Protopapas P., Huijse P., Estévez P. A., Zegers P., Príncipe J. C., Marquette J.-B., 2015, The Astrophysical Journal Supplement Series, 216, 25

Rau A., et al., 2009, 121

Richards J. W., et al., 2011, ApJ, 733, 10

Roykoff E. S., et al., 2005, ApJ, 631

Sub Group	Clusters	# Light Curves	Properties of Light Curves
A	1	7	Varying sources.
B	11	9	Possible varying sources.
C	18	10	Faint defuse sources (galaxies) or sources amongst defuse galaxies.
D	22	12	Faint source below detection threshold for majority of exposures.
E	9, 17	14	Sources appear in ccd chip gaps 50% of observations and only partially on ccds for remaining exposures.
F	0, 12, 20, 27	24	Sources unchangingone bright detection caused by ccd anomaly.
G	6	42	Noisy light curves with images showing possible extinction from clouds.
H	2, 10, 19 21, 23, 24 25, 26	157	Sources only detected during 5 point dithers during observations.
I	3, 4, 7	191	Sources appearing 0.2 magnitudes brighter in first four light curve points.
J	5	310	Faint and noisy sources bouncing around 0.4 magnitudes over . observations.
K	8, 13, 14, 15, 16	1,003	Sources near edge of ccd resulting in in dimming and brightening as the source moves ccd position during observations.
L	28	56693	Flat light curves of unchanging sources.

Table 2. Sub grouping of like clusters and their properties.

Scott D., Pierfederici F., Swaters R. A., Thomas B., Valdes F. G., 2007, in Shaw R. A., Hill F., Bell D. J., eds, Astronomical Society of the Pacific Conference Series Vol. 376, Astronomical Data Analysis Software and Systems XVI. p. 265

Shappee B. J., et al., 2014, The Astrophysical Journal, 788, 48

Stassun K. G., et al., 2018, AJ, 156, 102

Stetson P. B., 1996, PASP, 108, 851

Stubbs C. W., Doherty P., Cramer C., Narayan G., Brown Y. J., Lykke K. R., Woodward J. T., Tonry J. L., 2010, The Astrophysical Journal Supplement Series, 191, 376

Swaters R. A., Valdes F. G., 2007, in Shaw R. A., Hill F., Bell D. J., eds, Astronomical Society of the Pacific Conference Series Vol. 376, Astronomical Data Analysis Software and Systems XVI. p. 269

Valdes F. G., Swaters R. A., 2007, in Shaw R. A., Hill F., Bell D. J., eds, Astronomical Society of the Pacific Conference Series Vol. 376, Astronomical Data Analysis Software and Systems XVI. p. 273

Valenzuela L., Pichara K., 2018, MNRAS, 474, 3259

Watson C. L., Henden A. A., Price A., 2006, Society for Astronomical Sciences Annual Symposium, 25, 47

Wittman M., et al., 2002, Proceedings of the SPIE

van der Maaten L., Hinton G., 2008, Journal of Machine Learning Research, 9, 2579

APPENDIX A: FEATURES

APPENDIX B: LIGHT CURVE TRAITS

This paper has been typeset from a TEX/LATEX file prepared by the author.

ID (Gaia DR2)	**Type** (Days)	**Period**	**Cluster**	**Observations**
5445663255931713664	RRab	0.68	13	Located very nearby in feature space to new variables uncovered in cluster 1. See Appendix B.
5445874740121318656	RRab	0.508	-1 (noise)	Only partial detections due to dithering where source is partly on ccd for majority of exposures.
5447173538231827584	RRc	0.33	28	Seemingly flat on 1.5 hour time scales.
5444099161984715776	RRc	0.31	28	Below limiting magnitude for majority of exposures and seemingly unchanging in any detections.
5447200613705402752	RRab	0.57	28	Seemingly flat on 1.5 hour time scales.

Table 3. Caption

Feature	**Description**	**Inputs**	**Refs**
Amplitudes	Half the difference between the median of the maximum 5% and the median of the minimum 5% Magnitude.	Magnitude	Richards et al. (2011)
Auto correlation length	Length of linear dependence of a signal with itself at two points in time	magnitudes	Kim et al. (2011)
Beyond1Std	Percentage of points beyond one standard deviation from the weighted mean	Magnitude & error	Richards et al. (2011)
CAR_{mean}	The mean of a continuous time auto regressive model using a stochastic differential equation	Magnitude & time & error	Pichara et al. (2012)
CAR_{σ}	The variability of the time series on time scales shorter than τ	Magnitude & time & error	Pichara et al. (2012)
CAR_{τ}	The variability amplitude of the time series	Magnitude & time & error	Pichara et al. (2012)
H_1	Amplitude derived using the Fourier decomposition	Magnitude	Kim & Bailer-Jones (2016)
Con	The number of three consecutive data points that are brighter or fainter then 2σ and normalized by N -2	Magnitude	Kim et al. (2011)
Linear Trend	Slope of a linear fit to the light curve	Magnitude & time	Richards et al. (2011)
MaxSlope	Maximum absolute magnitude slope between two consecutive observations	Magnitude & time	Richards et al. (2011)
Mean	The mean magnitude	Magnitude	Kim et al. (2014)
Mean Variance	the ratio of the standard deviation to the mean magnitude	Magnitude	Kim et al. (2011)
Median Absolute Deviation	The median discrepancy of the data from the median data	Magnitude	Richards et al. (2011)
Median Buffer Range Percentage	Fraction of photometric points with amplitude/10 of the median magnitude	Magnitude	Richards et al. (2011)
Pair Slope Trend	The fraction of increasing first differences minus the fraction of decreasing first differences	Magnitude	Richards et al. (2011)
Q31	The difference between the 3rd and 1st quarterlies	Magnitude	Kim et al. (2014)
R_{21}	2^{nd} to 1^{st} amplitude ratio derived using the Fourier decomposition	Magnitude	Kim & Bailer-Jones (2016)
R_{31}	3^{rd} to 1^{st} amplitude ratio derived using the Fourier decomposition	Magnitude	Kim & Bailer-Jones (2016)
Rcs	Range of cumulative sum	Magnitude	Richards et al. (2011)
Skew	The skewness of the sample	Magnitude	Richards et al. (2011)
Slotted Auto Correlation Function Length	Slotted auto correlation length	Magnitude & time	Protopapas et al. (2015)
Small Kurtosis	Small sample kurtosis of magnitudes	Magnitude	Richards et al. (2011)
Standard Deviation	Standard deviation of the magnitudes	Magnitude	Richards et al. (2011)
Stetson K_{AC}	Stetson K applied to the slotted auto correlation function of the light curve	Magnitude	Stetson (1996); Kim et al. (2011)
Variability Index	Ratio of the mean of the square of successive differences to the variance of data points	Magntiude	Kim et al. (2011)

Table A1. Caption

Cluster	Number of Light	Notes
Noise	3270	light curves with majority non-detections as well as possible variable sources and photomery affected by telescope dithering.
0	8	Sources with majority non detections and one bright detection caused by ccd anomaly.
1	7	Variable Sources.
2	36	Sources only detected during 5 point dither at beginning of observations and once during end of night dithers otherwise in chip gap.
3	23	Sources appearing 0.2 mags brighter in first 5 light curve points, a result of photometer corrections using different standard stars for ccd corrections.
4	109	Similar photometric effects to cluster 3.
5	310	Faint noisy source bouncing around 0.4 magnitude differences
6	42	Very noisy light curves with cutouts showing possible extinction from weather in exposures
7	59	Similar photometric effects to cluster 3.
8	37	Sources near edge of cdd resulting in deeming and brightening as the source is slightly moved nearer and further from the ccd edge during observations.
9	7	Dither affects on photometery.
10	40	Similar photometric effects to cluster 2.
11	9	Noisy light curves, possible varying sources.
12	6	Similar photometric effects to cluster 0.
13	920	Similar photometric effects to cluster 8.
14	21	Similar photometric effects to cluster 8
15	19	Similar photometric effects to cluster 8.
16	6	Similar photmetric effects to cluster 8.
17	7	Similar photmetric effects to cluster 9.
18	10	Faint defuse sources (galaxies) orsources amongst defuse galaxies
19	5	Similar photometric effects to cluster 2.
20	10	Similar photometric effects to cluster 0.
21	12	Similar photometric effects to cluster 2 with more detections (source not at close to edge as cluster 2).
22	12	Majority non-detections, 4 detections of very faint sources throughout observations.
23	26	Similar to cluster 2, with only detections at beginning of observations.
24	8	Similar to cluster 2 with only detections at beginning of observations.
25	14	Similar to cluster 2, more variance in magnitudes between detections.
26	16	Similar to cluster 2 with only detections at beginning of observations.
27	102	Flat light curves with one brighter point caused by ccd anomaly.
28	56693	Seemingly flat light curves within photometric errors.

Table B1. Clusters Identified from Antlia field light curves using HDBSCAN.