MNRAS **000**, 1–13 (2017) Preprint 27 March 2017 Compiled using MNRAS LATEX style file v3.0

Preliminary Data Access Center : User Report

Krzysztof Suberlak, [1]★ Željko Ivezić, [1] and the PDAC team
[1] *Department of Astronomy, University of Washington, Seattle, WA, United States*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
A report on user experience of the Preliminary Data Access Center (PDAC). We test the quality and ease of access to the data. PDAC will pave the way to the Science User Interface and Tools (SUIT). We employ both in-detail study of individual objects, and a statistical study of an ensemble of objects. We evaluate user-friendliness of the current interface, and make recommendations for its future improvements.

1 INTRODUCTION

This is a document to report on the user experience testing of the Preliminary Data Access Center. The Large Scale Synoptic Telescope (LSST) will produce a big volume of data. Such unprecedented data stream poses new challenges to provide an easy access for users, in such a way that they can quickly find what they need, and thus be able to focus on the science goal that they would like to achieve. The detail description of such online user-interface called Science User Interface and Tools is outlined in documents LDM-130 (SUIT requirements) and LDM-492 (SUIT Vision). An idea of having an interface to the data is not new : there exists Aladin, SDSS CAS jobs, IPAC IRSA, Mikulsky NASA Archive, NED, and many other archives. These allow a user to query for data (either via SQL query, or interface), returning the data table. Some user interfaces (eg. IRSA) have some rudimentary plotting capabilities. There have been ideas of a new interface, that would not only eg. plot the lightcurve and display the spectrum, but also allow the user to run some machine learning algorithms, or simple models that can help narrow down the query, or obtain science results in the browser. Namely, Victor Pankratius, from MIT, in his talk "Computer-Aided Discovery: Towards Scientific Insight Generation with Machine Support" outlined the idea of an ipython notebook - access to data, which lives in the cloud, is allocated some CPU share and memory, and allows one to upload / download the data and run the model in real time, which is especially helpful to geoscientists doing fieldwork, where new data acquisition conditions their next step.

These requirements and the vision for SUIT have been further described on confluence pages[1]. Some technical notes about current implementation of SUIT by PDAC are also available via confluence pages[2].

[1] https://confluence.lsstcorp.org/display/DM/Science+User+Interface+and+Tools
[2] https://confluence.lsstcorp.org/display/DM/Guide+to+PDAC+version+1

2 OVERVIEW OF PERFORMED TESTS

We test a variety of aspects of PDAC : the user interface, infrastructure, and database ingestion. The user interface is similar to IRSA, which aids the ease of access. In Section 3 we describe the functionality available through user interface. It is a work in progress, hence any deficiency outlined may become updated in real-time, whereas some recommendations, if met with approval, may have a longer implementation timescale. In Section 4 we describe the structure of available data : both data that is available directly from NCSA (internal catalogs) , and data that is available from IRSA (external catalogs). In that section we also provide an overview of query and analysis methods available directly through the User Interface, as well as through SQL. Finally, in Section 5 we consider the quality of database ingestion, answering the question of how well was a given dataset loaded into PDAC. In particular we compare the S82 forced photometry dataset, an outcome of the Summer 2013 reprocessing, to the same data stored locally at the University of Washington.

3 USER INTERFACE: WHAT WE SEE

In order to access PDAC we follow the directions [2] that include logging to NCSA via VPN https://vpn.ncsa.illinois.edu/ using Cisco AnyConnect Secure Mobile Client, and opening in the web browser http://lsst-sui-proxy01.ncsa.illinois.edu/suit. This opens the main interface screen, which allows to select the database, and perform the desired query.

Currently, PDAC v1, in the upper-left corner of the interface, under tab 'LSST Data' (see Fig. 1) includes the Summer 2013 DM-stack reprocessed SDSS Stripe 82 data (database name `sdss_stripe82_00`) , hosted at the NCSA on the LSST prototype ("integration cluster") hardware, in Qserv [Gregory Dubois-Felsmann, priv.comm. 02-20-2017, slack]. The only other locally stored database (as of March 2017), is WISE catalog, that is not yet accessible via the graphical user interface (it can be queried as Data Base `wise_00`, with catalogs 'Object' containing objects (like

DeepSource in S82 above), and 'ForcedSource' containing forced photometry (like DeepForcedSource in S82)).

The upper-left corner of the interface also leads to 'External Images' and 'External Catalogs'. The Catalogs are all NASA/IPAC[3] Infrared Science Archive(IRSA) publicly accessible catalogs, including GAIA, WISE, 2MASS, SPITZER, etc. (see Fig. 2).

4 INFRACTRUCTURE : WHAT IS AVAILABLE AND HOW TO GET IT

As we described in Section 3, the main user interface allows access to the internally stored (at NCSA) SDSS Stripe 82 data reprocessed during the Summer 2013 [4] as part of Data Challange with the continuously developed LSST Stack [5].

The reprocessing included:

- coadding the data from all epochs in each of the ugriz SDSS filters. Measurements on coadds (per object) are available as `RunDeepSource` table, accessible via Catalogs –> 'DeepSource' . The single-band coadded images with MariaDB metadata are available as `DeepCoadd` table, accessibla via Images –> 'DeepCoadd' .
- using i-band detections to seed forced photometry on all epochs in all bands. The results of photometry are available as `RunDeepForcedSource` table, accessible via Catalogs –> 'Deep Forced Source' .
- For reference , the individual calibrated single epoch images are available as `Science_Ccd_Exposure` table, accessible via Images –> 'Science CCD Exposure'

Additional details of the schema are also outlined in the LSST Data Challenge Report [Shaw, Juric, Becker, Krughoff et al. 2013], and the LSST Database Schema Browser [6].

5 DATABASE INGESTION : IS WHAT WE GET WHAT WE EXPECTED TO GET?

6 METHODS

We perform single-object tests and statistical tests on an ensemble of objects .

7 POSITIONAL QUERY

First, we study in detail a particular source type - we consider examples of variable objects, confirmed by previous studies, such as RR Lyrae stars. Sesar et al. (2010) performed lightcurve template fits to 483 RR Lyrae lightcurves from SDSS (see Fig.3). Both fit parameters and lightcurves are publicly accessible in the online version of the journal. We apply positional query to PDAC against these objects (see Fig. 8 for object positions), download lightcurves for objectIds within the search radius (2 arcec by default), run Lomb-Scargle periodogram to find period, and plot the PDAC-pulled data phased on the best-fit period.

Comparing the S82 data stored at PDAC to the data from Sesar et al. (2010), we want to treat the latter as 'ground truth', but as a sanity check we perform Lomb Scargle periodogram testing to confirm the more detailed analysis of Sesar et al. (2010). Using *astroML* python module (Vanderplas et al. 2012), we sample the uniformly spaced frequency grid with N=5000 samples span between the smallest and the largest frequency reported in Table 1 of Sesar et al. (2010) ±10%, i.e. $\omega_{min} = 0.9(2\pi/P_{min})$, $\omega_{max} = 1.1(2\pi/P_{min})$. We use the default *astroML* Lomb Scargle periodogram settings, namely generalized LS (see Eq.20 in Zechmeister & Kürster (2009), and Section 10.3.2 in Ivezić et al. (2014)).

Using the same frequency grid for all 483 RR Lyrae, we compute Lomb-Scargle periodograms, and determine the best-fit period from the highest frequency peak (see eg. Fig. 5). We find that for about half of the SDSS lightcurves from Sesar et al. (2010), the Lomb-Scargle periodogram fitting single-term Fourier Series (LS) is sufficient to find the 'true' period, (see Fig. 7). We illustrate examples of RR Lyrae falling into each group : where with the naive LS we find the same period (Fig. 4), a smaller period (Fig. 5), , or a bigger period (Fig. 6) than the ground truth. For the same objects we also show PDAC lightcurves for which we also computed LS periodogram - see Figs. 10, 11 and 12, respectively.

Using the RA, Dec for the RR Lyrae from Sesar et al. (2010) we positionally query the PDAC `RunDeepForcedSource` database (Cone Search), to find objects within 2 arcsec radius. As shown on Fig. 8, not all objects have a match. For the 343 stars with a PDAC match, we obtain calibrated g-magnitude lightcurves querying the `RunDeepForcedSource` and `Science_Ccd_Exposure` for the zero point magnitudes per exposure. For these PDAC lightcurves we also calculate Lomb-Scargle periodogram and find the frequency with most-significant power.

We tested the periodogram results for few RR Lyrae using the NASA Exoplanet Archive Periodogram[7]. The calculation takes on average 15 seconds (illustrated on Fig. 14, same for for RR Lyr ID=4099 and 470994). See Table 1 for a summary of results.

7.1 Area query

Second, we query the S82 database against a small subset of a given S82 patch (few degrees), downloading lightcurves for ~ 100000 objects in that area of the sky. We plot color-color diagrams, as in Sesar et al. (2007), Fig.3 ,4, and color - magnitude diagrams to show the morphology of the Sgr dSph tidal stream (Sesar et al. 2010).

[3] Infrared Processing and Analysis Center, http://www.ipac.caltech.edu/project/lsst

[4] https://confluence.lsstcorp.org/display/DM/Properties+of+the+2013+SDSS+Stripe+82+reprocessing

[5] https://pipelines.lsst.io/index.html

[6] https://lsst-web.ncsa.illinois.edu/schema/index.php?t=DeepForcedSource&sVer=S12_lsstsim

[7] http://exoplanetarchive.ipac.caltech.edu/cgi-bin/Pgram/nph-pgram

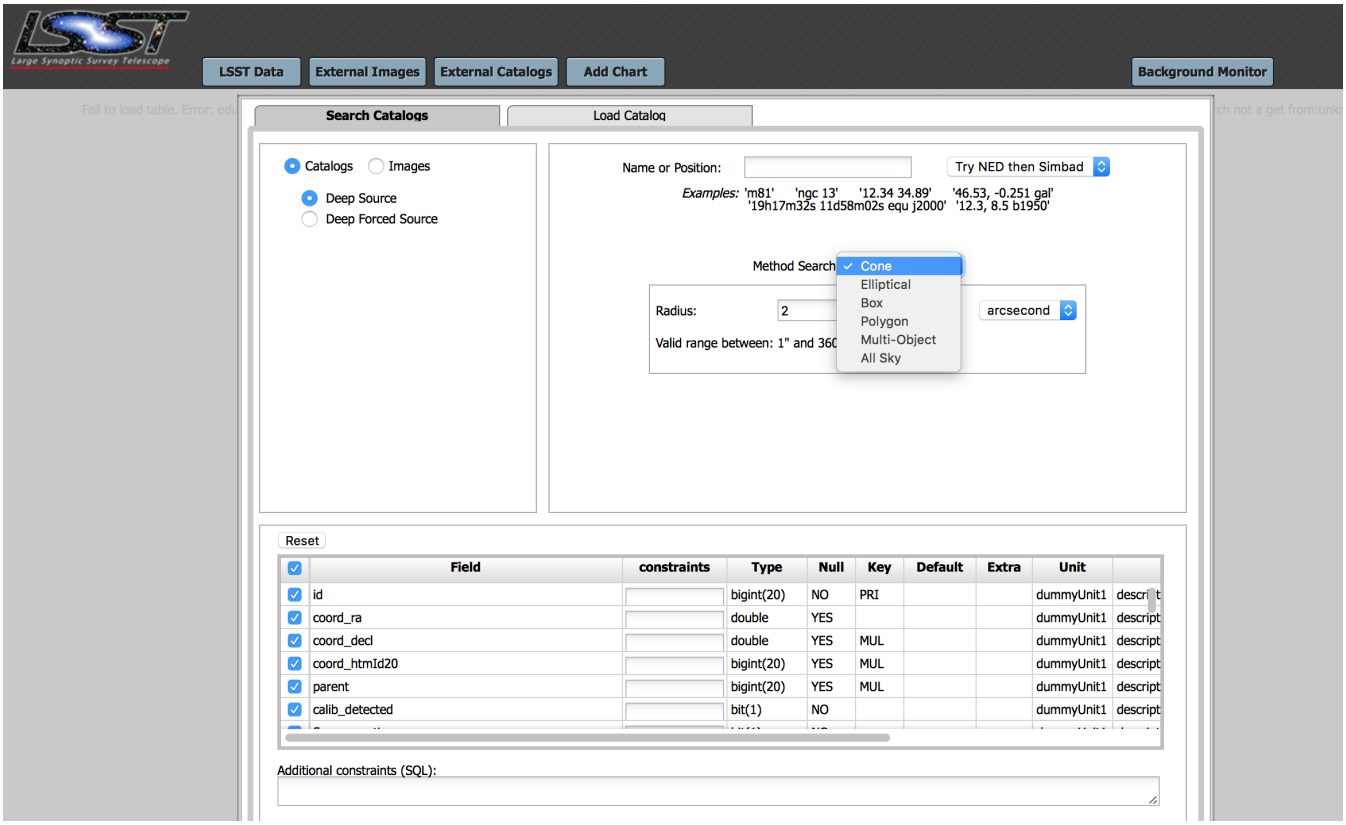

Figure 1. The main user interface of PDAC ver. 1

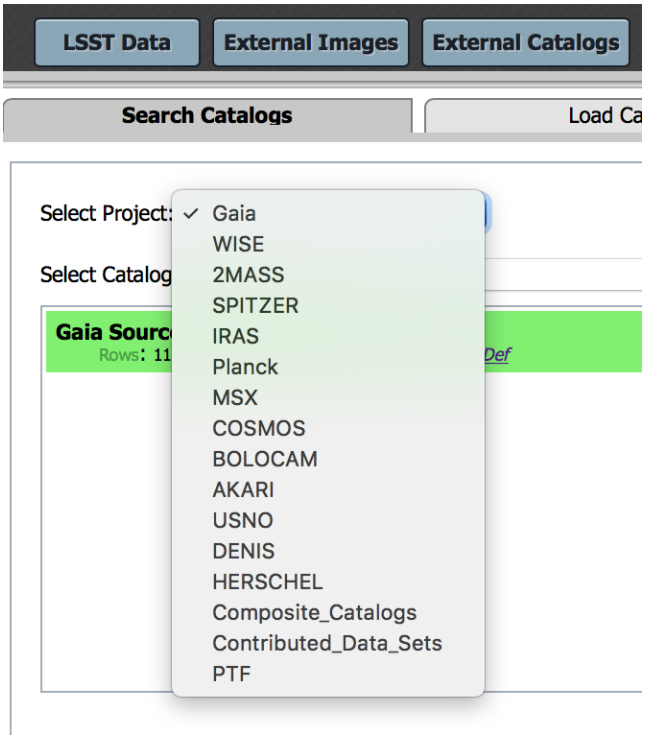

Figure 2. IPAC- hosted catalogs , accessible via IRSA.

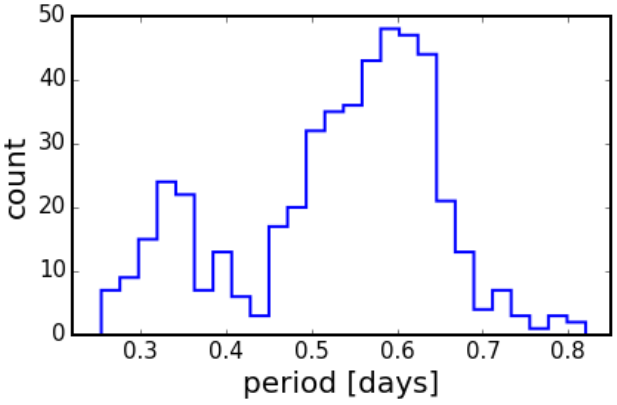

Figure 3. Distribution of RR Lyrae periods for 483 objects in (Sesar et al. 2010). Note the bimodal distribution, reflecting two main RR Lyrae types : 309 RRab (right) and 104 RRc (left) (see also Fig.16 in (Sesar et al. 2010)).

8 RESULTS

9 CONCLUSIONS

ACKNOWLEDGEMENTS

Thank you !

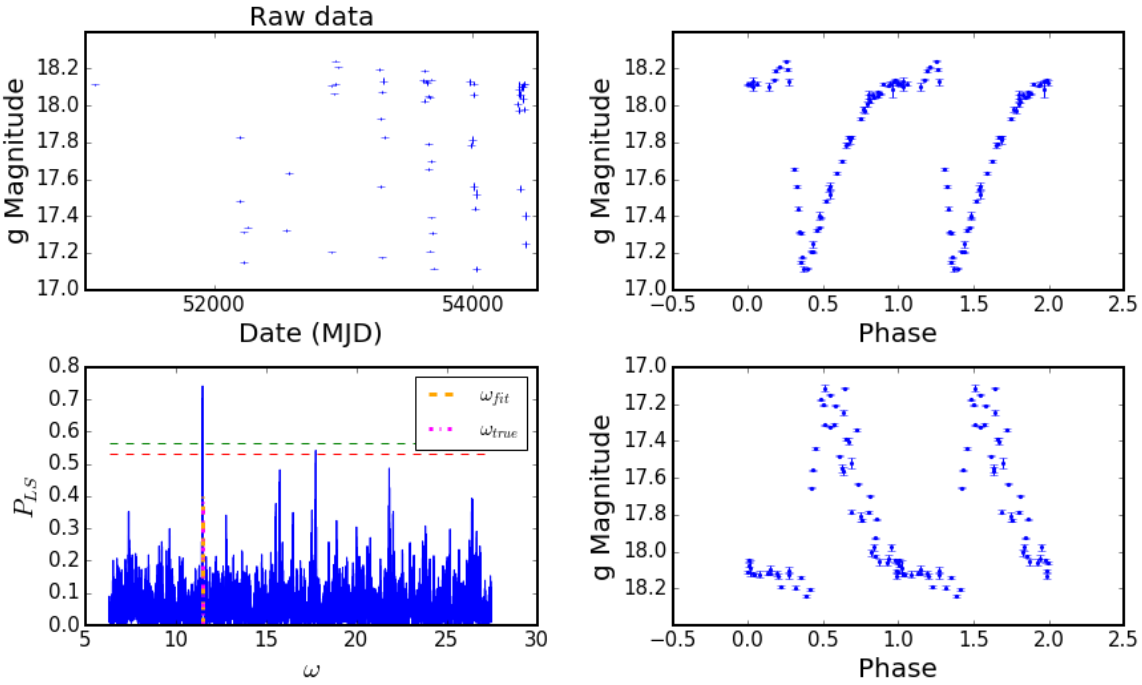

Figure 4. An example of the *astroML* Lomb Scargle periodogram performance, calculated for RR Lyr ID=13350 in SDSS g band (following Table 2 in (Sesar et al. 2010)), using the SDSS data from (Sesar et al. 2010). It took 18.6 miliseconds on a laptop to calculate this periodogram. The upper left panel depicts the raw SDSS lightcurve data. The upper right panel shows the phased lightcurve constructed with a cited period of 0.547987 days ('P_{true}'). The lower left panel shows the Lomb Scargle periodogram on a uniform frequency grid (5000 bins), where the orange and magenta vertical lines mark the location of the highest periodogram peak, and the frequency based on the reported period ($\omega_{true} = 2\pi/P_{true}$). The lower right panel shows the phased lightcurve constructed with the Lomb-Scargle Periodogram period of 0.547161 days, corresponding to the highest peak, $P_{fit} = 2\pi/\omega_{fit}$. The horizontal red and green lines mark the 5% and 1% significance levels for the highest peak, as found from 500 bootstrap resamplings (See `http://www.astroml.org/book_figures/chapter10/index.html`). The same object, but pulling the data from PDAC, is shown on Fig. 10

Table 1. Comparison of RR Lyrae periods obtained with different methods. First, P(S) is the 'ground truth' - period resulting from detailed template fitting by (Sesar et al. 2010). Second, P(LS) is the period corresponding to the most prominent frequency in the Lomb-Scargle periodogram (LS) computed on the SDSS lightcurve for a given object pulled from online journal data in (Sesar et al. 2010). Third, P(EXO) uses the same SDSS data from (Sesar et al. 2010) in g-band, to find the best period with the NASA Exoplanet Archive Periodogram service. Fourth, P(PDAC) uses the data pulled from PDAC, for which we find the best period using LS periodogram (same method as P(LS)).

ID	P(S)	P(LS)	P(EXO)	P(PDAC)
4099	0.641754	0.280827	0.64175	0.280827
13350	0.547987	0.547161	0.35365	0.547969
470994	0.346794	0.531667	0.34679	0.531667

REFERENCES

Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., & Gray, A. 2014, Statistics, Data Mining, and Machine Learning in Astronomy
Sesar, B., et al. 2010, ApJ, 708, 717
Sesar, B., et al. 2007, The Astronomical Journal, 134, 2236
Vanderplas, J., Connolly, A., Ivezić, Ž., & Gray, A. 2012, in Conference on Intelligent Data Understanding (CIDU), 47
Zechmeister, M., & Kürster, M. 2009, A&A, 496, 577

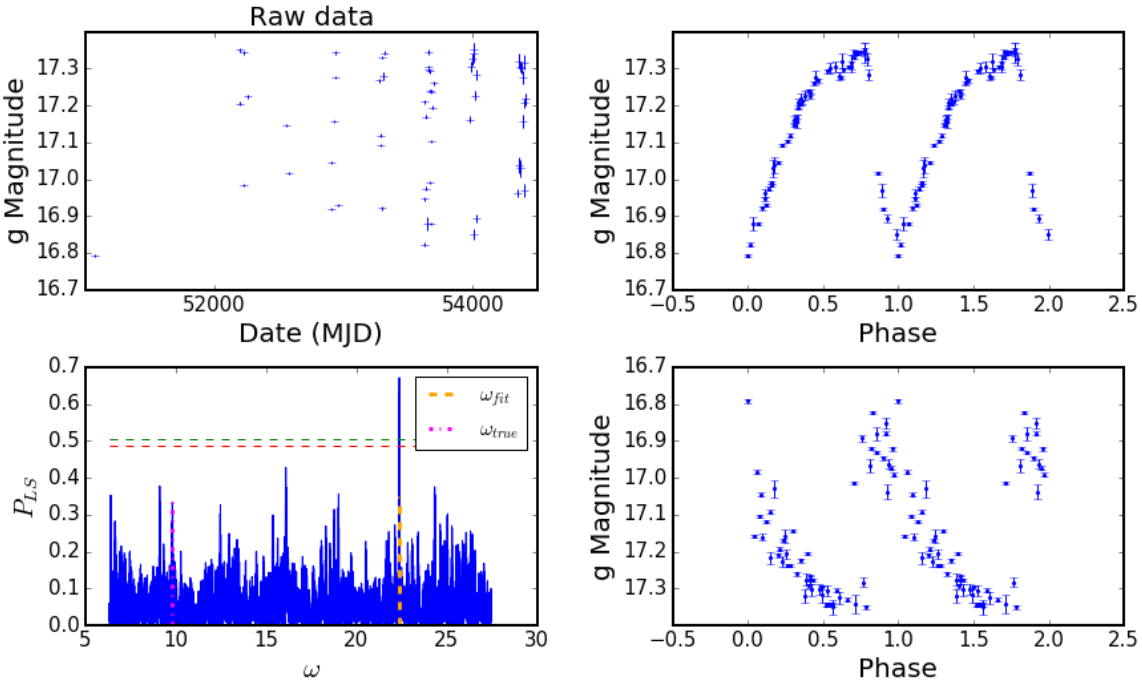

Figure 5. A periodogram, raw and folded lightcurve using SDSS data from (Sesar et al. 2010) for RR Lyr ID =4099 (the same object with PDAC data is shown on Fig. 11). It is an example of a failure of naive single Lomb Scargle periodogram performance - the ratio of $\omega_{true}/\omega_{fit}$ = 0.437. The 'true' period from (Sesar et al. 2010) is 0.641754 days, whereas the naive Lomb-Scargle periodogram approach yields the 'fit' period of 0.280827 days. Note that here ω_{fit} and ω_{true} significantly differ for this RR Lyr, and the 'true' frequency, backed-up by the full lightcurve fitting of (Sesar et al. 2010), appears as only one of insignificant periodogram peaks. Everything else as on Fig. 4.

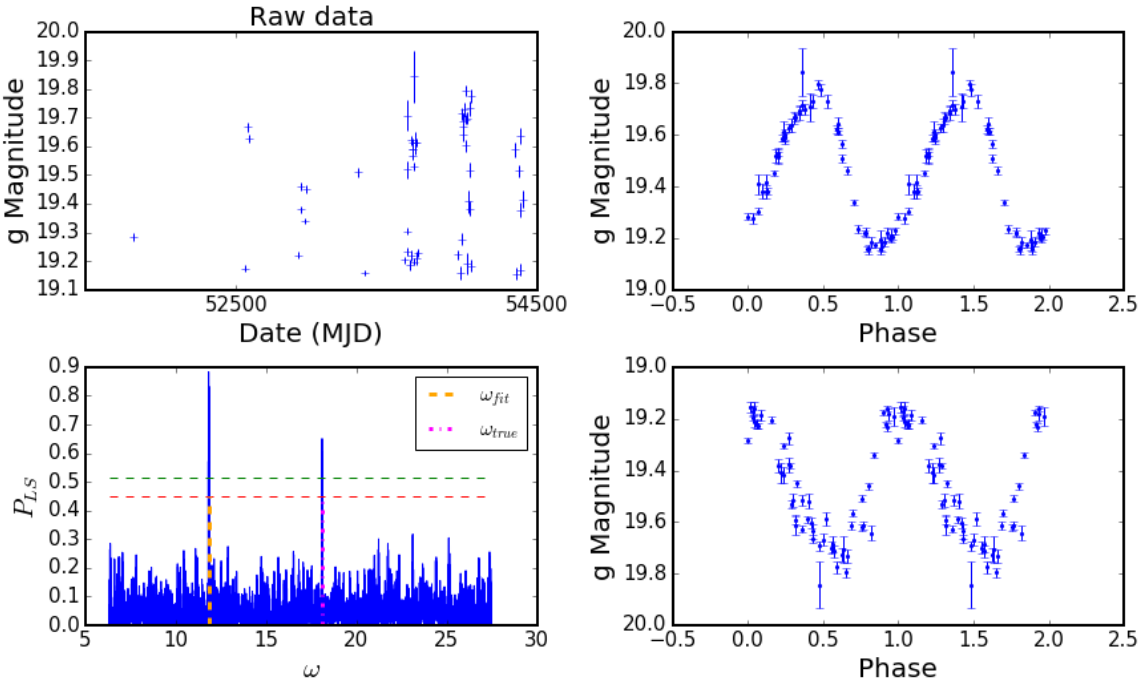

Figure 6. Same as Fig. 5, using the SDSS data pulled from (Sesar et al. 2010), with $\omega_{true}/\omega_{fit}$ = 1.53. Here RR Lyr ID=470994 has a cited period of 0.346794 days ('P_{true}'), whereas period derived from the Lomb-Scargle periodogram is 0.531667. It may be a good example of aliasing.

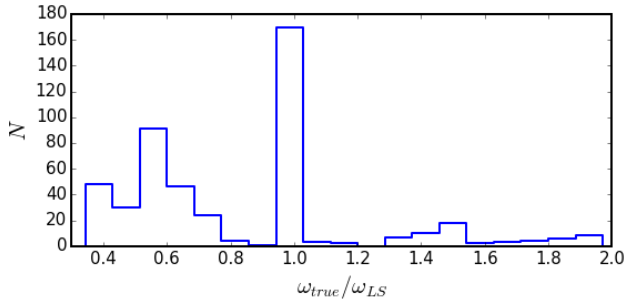

Figure 7. The distribution of the ratio of ω_{true} to ω_{fit}, where ω_{true} is inferred directly from the 'ground truth' - period cited in Table 2 of (Sesar et al. 2010). We take the same SDSS data from the paper (Table 1 in (Sesar et al. 2010)), and calculate the Lomb-Scargle single-term generalized periodogram. The frequency corresponding to the highest peak is ω_{fit}. Thus, wherever this ratio is approximately equal to 1, this means that the naive LS approach is able to recover the 'true' period. However, where the highest frequency peak is not the same as ω_{true}, the ratio will be smaller or bigger from 1. This may be caused by the inherent simplicity of the simple single-term Fourier Series fitting. Indeed, some RR Lyrae lightcurves may have shapes that are insufficiently described by a single sinusoid (as on Fig.10.18 in (Ivezić et al. 2014)).

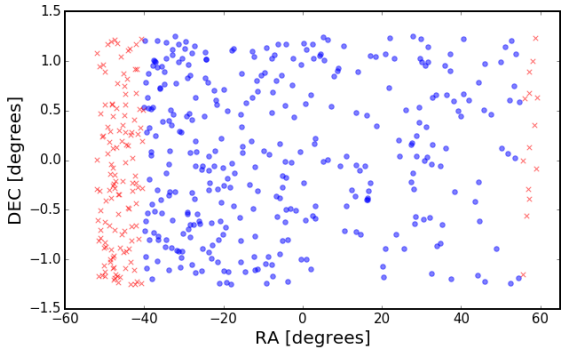

Figure 8. Results of positional query against 483 RR Lyrae stars from (Sesar et al. 2010), using their RA, Dec. Blue dots are 343 stars that have a match in the PDAC S82 dataset within 2 arcsec, and red crosses are 140 stars that did not. Increasing the search radius to 3 arcsec does not alter this result.

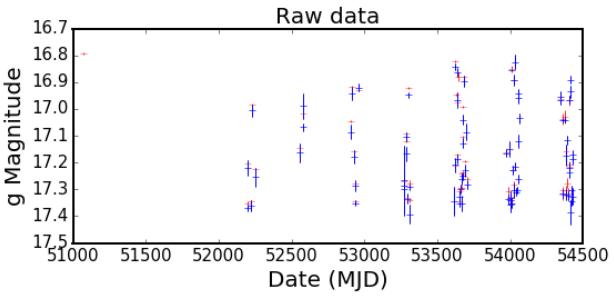

Figure 9. Comparison of RR Lyr ID=4099 from (Sesar et al. 2010) (red crosses), and PDAC (blue crosses). The two lightcurves have different length : 59 vs 162 points, respectively.

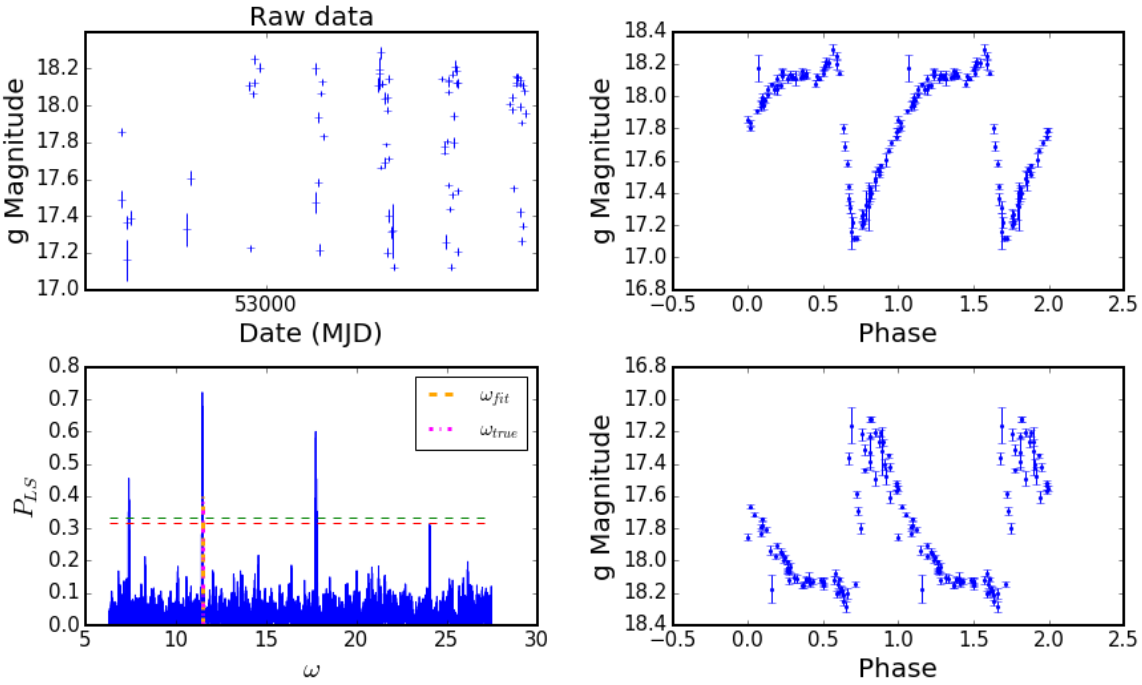

Figure 10. The same object as Fig. 4, but using data downloaded using PDAC. Using PDAC data, the RR Lyr ID=13350 has a best-fit period of 0.547969 days, almost identical to true period of 0.547987 from (Sesar et al. 2010). Panels the same as on Fig. 5

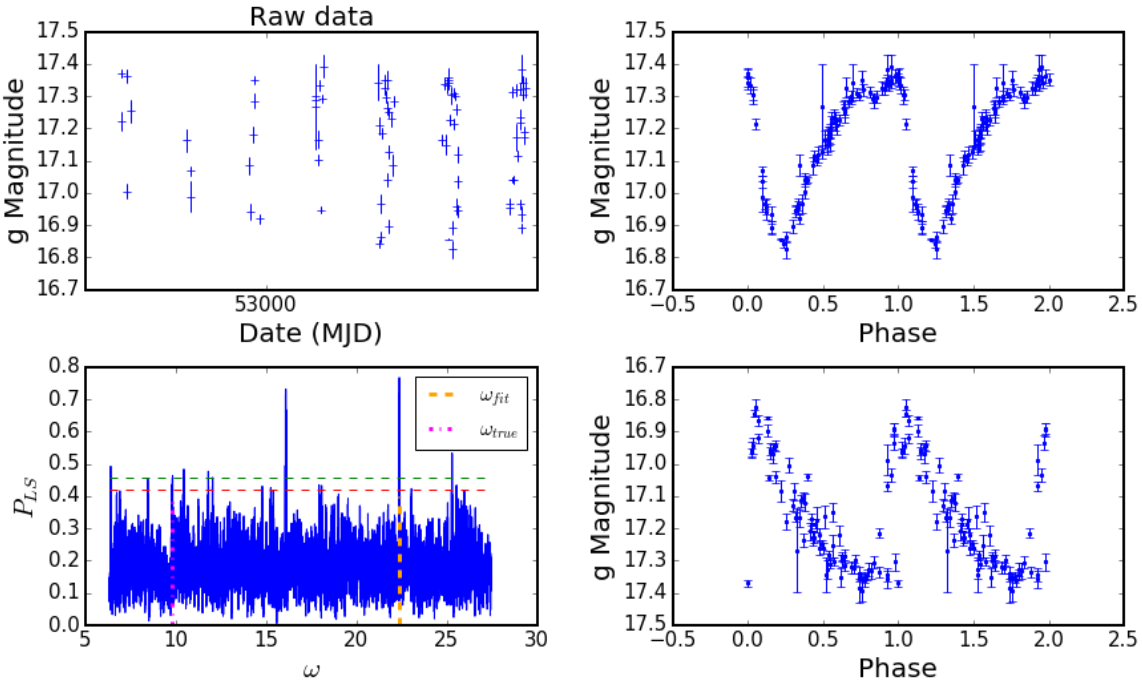

Figure 11. The same object as Fig. 5, but using data downloaded using PDAC. Calculating a naive LS periodogram using PDAC data for RR Lyr ID=4099 we find the best-fit period (frequency with highest power) of 0.280827 days, almost identical to the period found using LS periodogram on the SDSS (Sesar et al. 2010) data of 0.280827 days. Both are discrepant with respect to the 'true' period of 0.641754 days from (Sesar et al. 2010). Panels the same as on Fig. 4

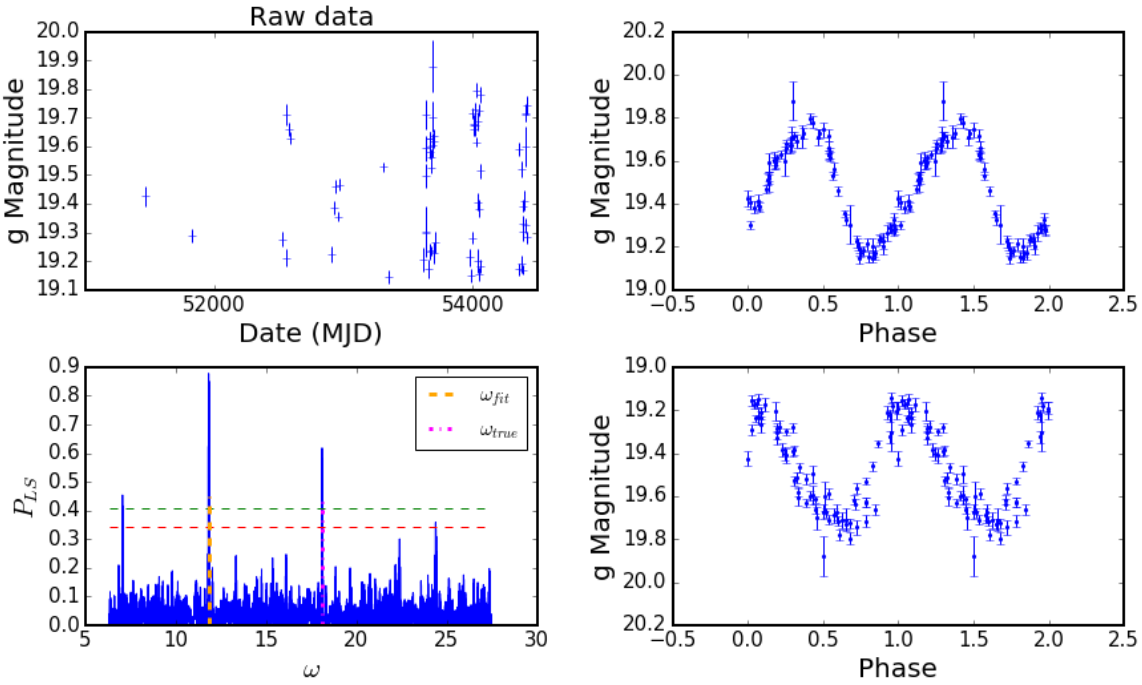

Figure 12. The same object as Fig. 6, but using data downloaded from PDAC. Calculating a naive LS periodogram using PDAC data for RR Lyr ID=470994 we find the best-fit period (frequency with highest power) of 0.531667 days, almost twice as high as the 'true' period of 0.346794 days from (Sesar et al. 2010). For this star we get an identical period if we use LS periodogram on SDSS data from (Sesar et al. 2010) as opposed to PDAC. Panels the same as on Fig. 4

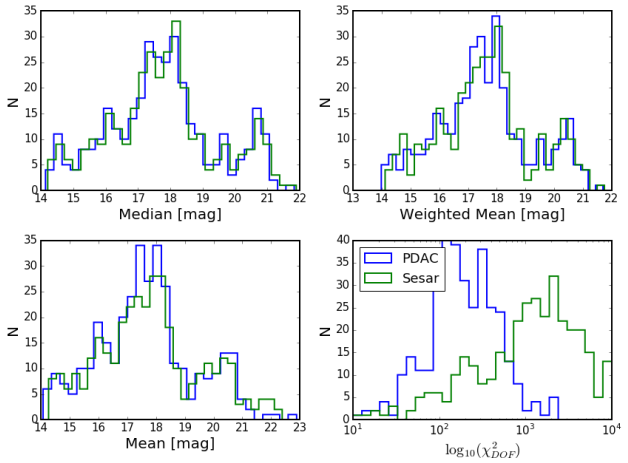

Figure 13. Comparison of the original (Sesar et al. 2010) lightcurves (green) against data for the same objects pulled from PDAC (blue). For each of the 383 lightcurves in SDSS *g*-band, without any pre-processing or clipping, we calculated the median, weighted mean, mean, and chi^2_{DOF}.

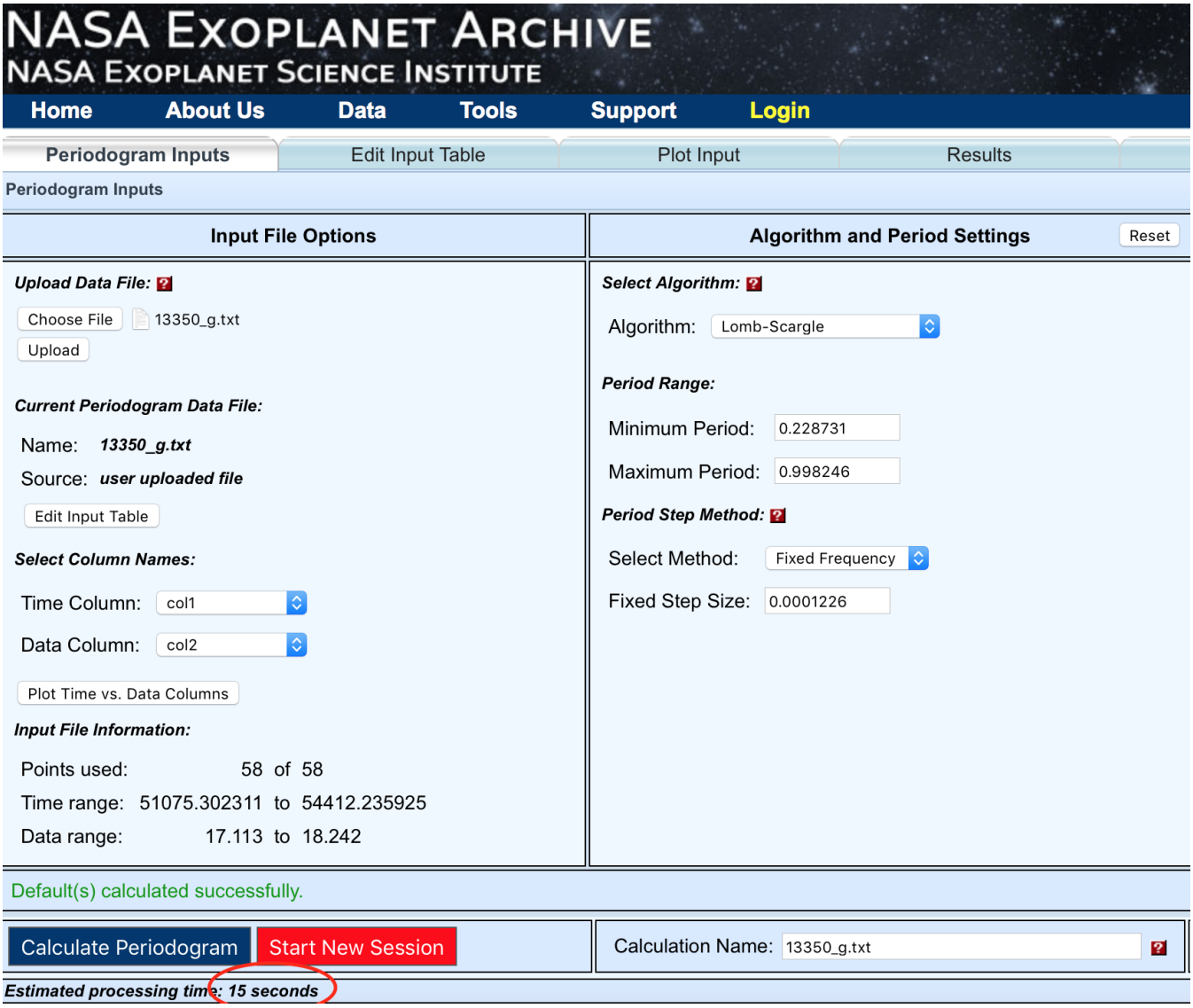

Figure 14. The same object as Fig. 4, and Fig 10, using the SDSS data from (Sesar et al. 2010). The highest significance frequency peak (power 21.58) corresponds to a period of 0.35365194 days. Only the second in significance peak (power 20.62) corresponds to the 'true' period of 0.547969 (Sesar et al. 2010). Note the bottom-left corner : the calculation took 15 secs for one lightcurve (compare to few miliseconds of AstroML code naive single-sinusoid approach that gave the same result for this particular object)

APPENDIX A: SQL QUERIES

A1 Cone query around given RA, DEC

To query `RunDeepForcedSource` table, for an object with RA = 0.935679, and DEC=1.115859, search radius 0.00055555 (all in degrees), saving the output of the query to `catalogCone1.json` file:

```
curl -o catalogCone1.json -d 'query=SELECT+objectId+FROM+RunDeepForcedSource+WHERE+\
                qserv_areaspec_circle(0.935679,1.115859,0.00055555);' \
                http://lsst-qserv-dax01.ncsa.illinois.edu:5000/db/v0/tap/sync
```

I perform such queries for each of the 483 objects in Sesar et al. (2010), using search radius of 2 arcsec.

A2 ID query to return lightcurves

We find that between 1 to 6 unique detection ID's are found within a search radius of 2 arsec for the 483 RR Lyrae. Given the IDs corresponding to the sought RA, DEC, location, to obtain the lightcurve in calibrated magnitudes, we need to query `RunDeepForcedSource` Table to provide information about individual exposures (mjd time of observation: `exposure_time_mid`, flux in a given filter `flux_psf`, flux error `flux_psf_err`, filter code `exposure_filter_id`, with `0,1,2,3,4` corresponding to `u,g,r,i,z`, respectively). We query `scienceCcdExposureId` table to obtain the flux for photometric zero point with assosiated error (`fluxMag0`, `fluxMag0Sigma`) with `exposure_id` in `RunDeepForcedSource` corresponding to `scienceCcdExposureId` in `Science_Ccd_Exposure`. We perform a `join` between these two tables, and store the result as `catalogForced1.json` file. Assuming that IDs that we query against are 3562429887808969, 3562429887812663, the query to obtain `g` magnitudes would be (see https://confluence.lsstcorp.org/display/DM/PDAC+sample+queries+and+test+cases#) :

```
curl -o catalogForced1.json -d 'query=\
SELECT \
  objectId, id, fsrc.exposure_id, fsrc.exposure_time_mid, exp.run, \
  scisql_dnToAbMag(fsrc.flux_psf,exp.fluxMag0) AS g, \
  scisql_dnToAbMagSigma(fsrc.flux_psf, fsrc.flux_psf_err, \
                        exp.fluxMag0, exp.fluxMag0Sigma) AS gErr \
FROM \
  RunDeepForcedSource AS fsrc, \
  Science_Ccd_Exposure AS exp \
WHERE \
  exp.scienceCcdExposureId = fsrc.exposure_id \
  AND fsrc.exposure_filter_id=1 \
  AND objectId IN (3562429887808969, 3562429887812663)\
ORDER BY exposure_time_mid' http://lsst-qserv-dax01.ncsa.illinois.edu:5000/db/v0/tap/sync
```

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