MNRAS **000**, 1–3 (2017) Preprint 26 February 2017 Compiled using MNRAS LATEX style file v3.0

Preliminary Data Access Center : User Report

Krzysztof Suberlak [1]★ Željko Ivezić, [1]
[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). Employing the SDSS and GAIA datasets 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.

Indeed, astronomers may find that quick look into the data, finding eg. all stars that exhibit RR Lyr variability and have been observed in a certain region of the sky, is very helpful.

Here we outline the user experience of PDAC (see PDAC technical description on [1]

Currently, PDAC v1, under tab 'LSST Data' in the upper-left corner of the interface (see Fig. 1) includes the Summer 2013 DM-stack reprocessed SDSS Stripe 82 data, hosted at the NCSA on the LSST prototype ("integration cluster") hardware, in Qserv [Gregory Dubois-Felsmann, priv.comm. 02-20-2017, slack]. 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'

Details of the S82 LSST reprocessing can be found in the PDAC document https://confluence.lsstcorp.org/display/DM/Properties+of+the+2013+SDSS+Stripe+82+reprocessing. 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 [2].

PDAC v1 under tab 'External Catalogs' also provides access to all NASA/IPAC Infrared Science Archive(IRSA) publicly accessible catalogs, including GAIA, WISE, etc. (see Fig. 2). These are stored at Infrared Processing and Analysis Center (IPAC) http://www.ipac.caltech.edu/project/lsst.

[1] https://confluence.lsstcorp.org/display/DM/Guide+to+PDAC+version+1

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

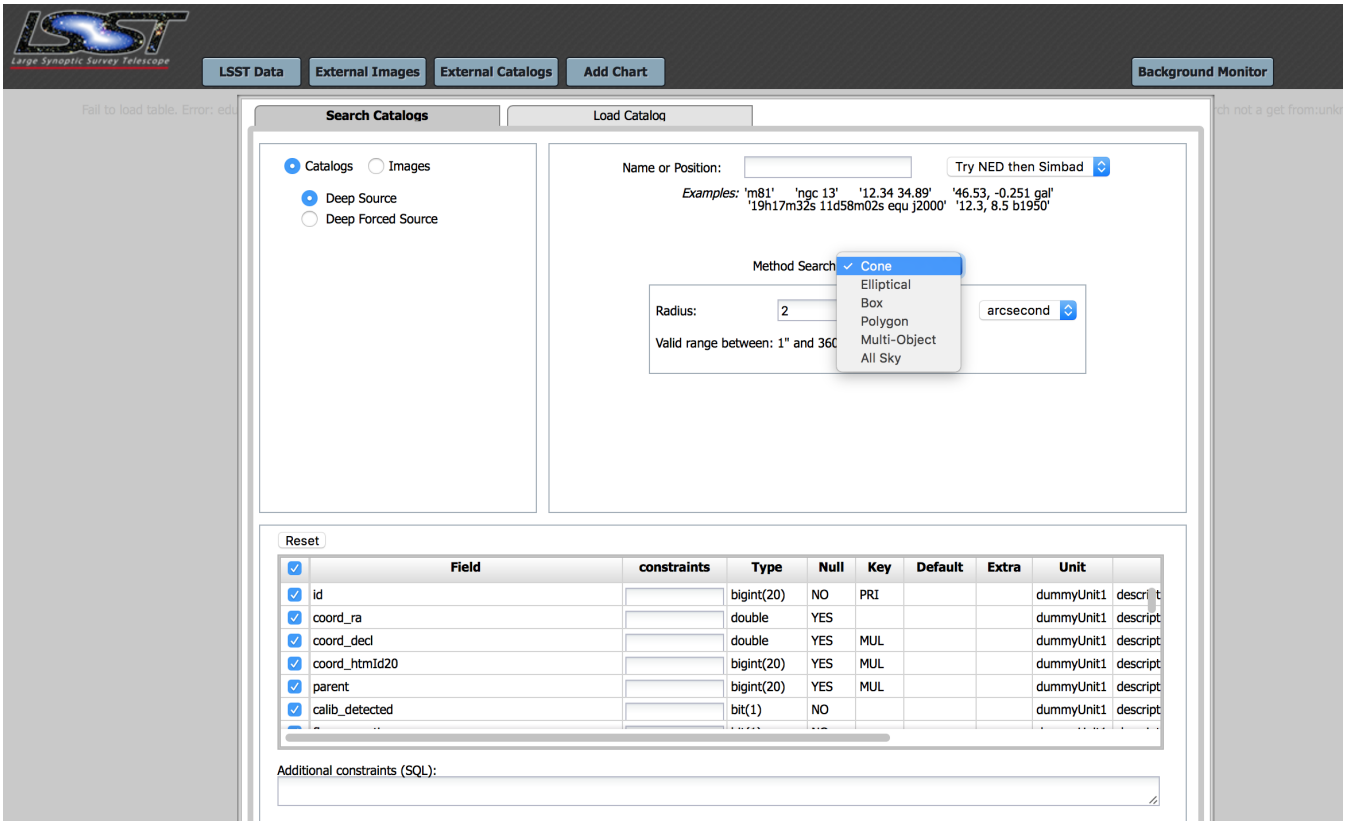

Figure 1. The main user interface of PDAC ver. 1

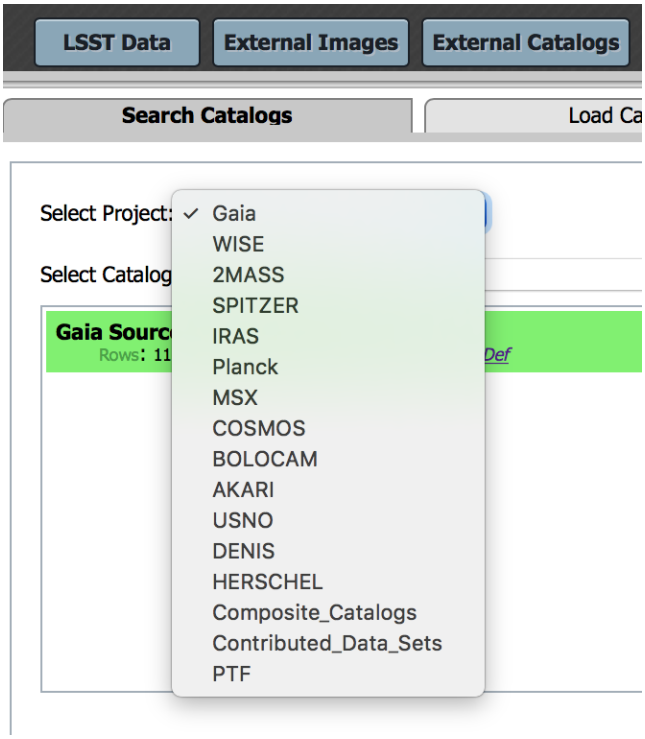

Figure 2. IPAC- hosted catalogs , accessible via IRSA.

2 METHODS

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

First, we study in detail a particular source - we consider examples of variable objects, confirmed by previous studies (eg. RR Lyrae from Sesar+2010, Table 1). We download these from the S82 dataset on PDAC, run Lomb-Scargle periodogram to find period, and plot the phased lightcurve. 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

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 Fig. 5). We find that for about half of the lightcurves the Lomb-

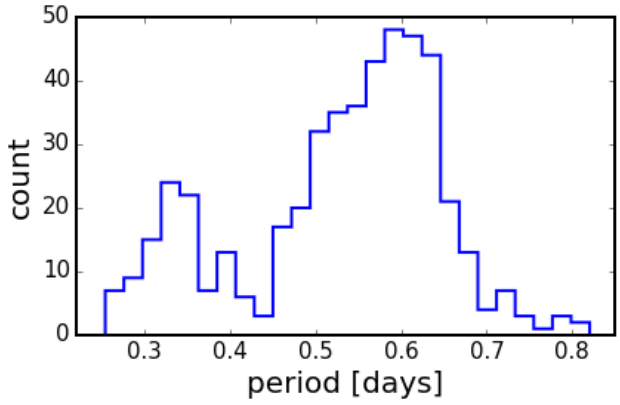

Figure 3. Distribution of RR Lyrae periods for 483 objects in (Sesar et al. 2010). Note the bimodal distribution, reflecting two main RR Lyrae types : 309 RRab (right) and 104 RRc (left) (see also Fig.16 in (Sesar et al. 2010)).

Scargle periodogram fitting single-term Fourier Series is sufficient to find the right period (note middle group centered on 1 on Fig. 5, and an example on Fig. 4). However, there are many cases where the naive single sinusoid is insufficient to correctly fit the period (groups outside of 1 on Fig. 4)

Using the Ra, dec for the RR Lyrae we positionally query the PDAC `RunDeepForcedSource` database to find objects within 2 arcsec radius. For these, we obtain calibrated g-magnitude lightcurves querying the `RunDeepForcedSource` and `Science_Ccd_Exposure` for the zero point magnitudes per exposure. Exactly as for Sesar et al. (2010) SDSS lightcurves considered before, for PDAC S82 lightcurves we also calculate Lomb-Scargle periodogram and find the most-significant frequency (fit the best period).

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 RESULTS

4 CONCLUSIONS

ACKNOWLEDGEMENTS

Thank you !

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

APPENDIX A: SOME EXTRA MATERIAL

If you want to present additional material which would interrupt the flow of the main paper, it can be placed in an Appendix which appears after the list of references.

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

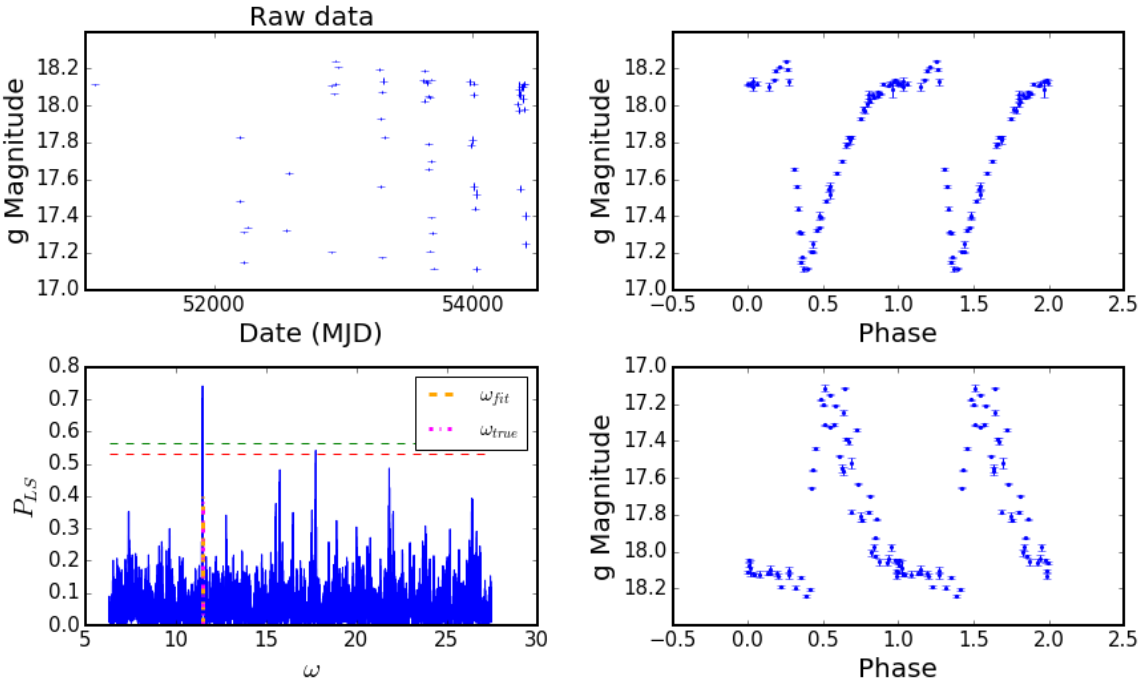

Figure 4. An example of the *astroML* Lomb Scargle periodogram performance, calculated for RR Lyr ID=13350 in SDSS g band (following Table 2 in (Sesar et al. 2010)). It took 18.6 miliseconds on a laptop to calculate this periodogram. The upper left panel depicts the raw SDSS lightcurve data. The upper 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, 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`)

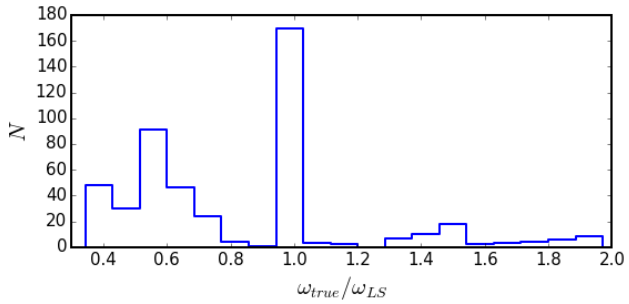

Figure 5. The distribution of the ratio of ω_{true} to ω_{fit}. Where the Lomb Scargle highest peak frequency is not the same as that reported in (Sesar et al. 2010), it may be because the naive single-term sinusoid is insufficient to describe the variability pattern (as in Fig.10.18, (Ivezić et al. 2014)).

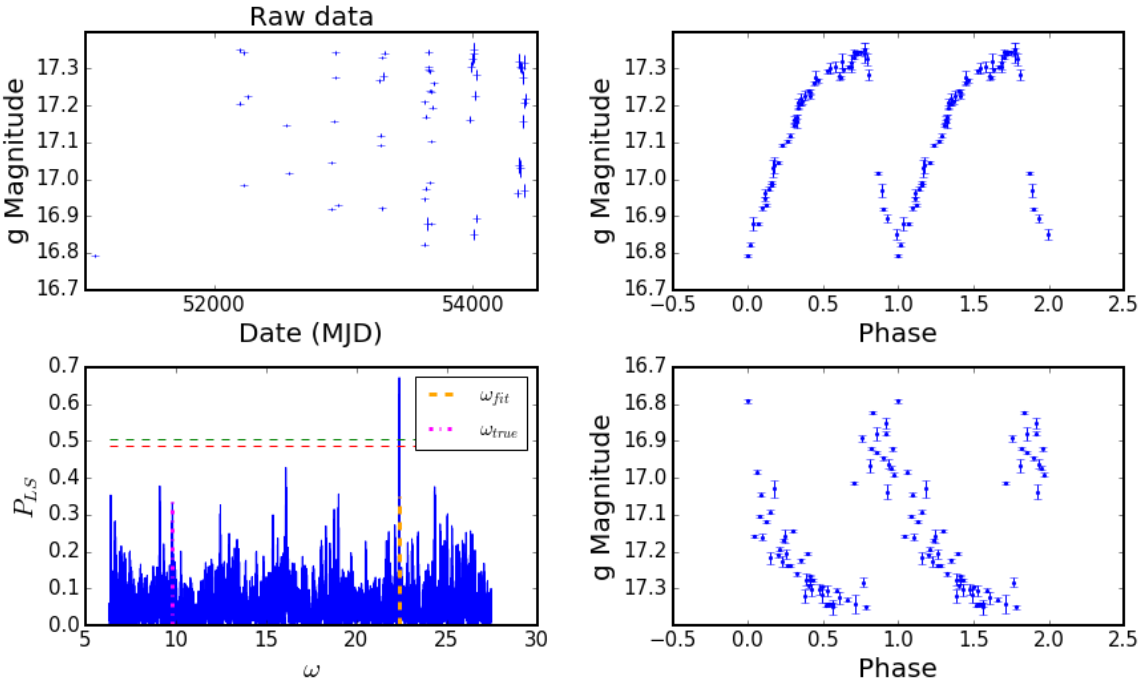

Figure 6. A failure of naive single Lomb Scargle periodogram performance - here the ratio of ω_{true} to ω_{fit} equals 0.437. Top-left : the raw SDSS lightcurve data for RR Lyr ID=4099 in g band. Top-right : the phased lightcurve constructed with a cited period of 0.641754 days ('P_{true}'). Bottom-left: the Lomb Scargle periodogram on a uniform frequency grid (5000 bins), with the orange and magenta vertical lines marking the location of the highest periodogram peak, and the frequency based on the reported period ($\omega_{true} = 2\pi/P_{true}$). Note that 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. As on Fig. 4, the horizontal red and green lines mark the 5% and 1% significance levels for the highest peak, as found from 500 bootstrap resamplings. Bottom-right : the phased constructed with the P_{fit} of 0.280827 days.

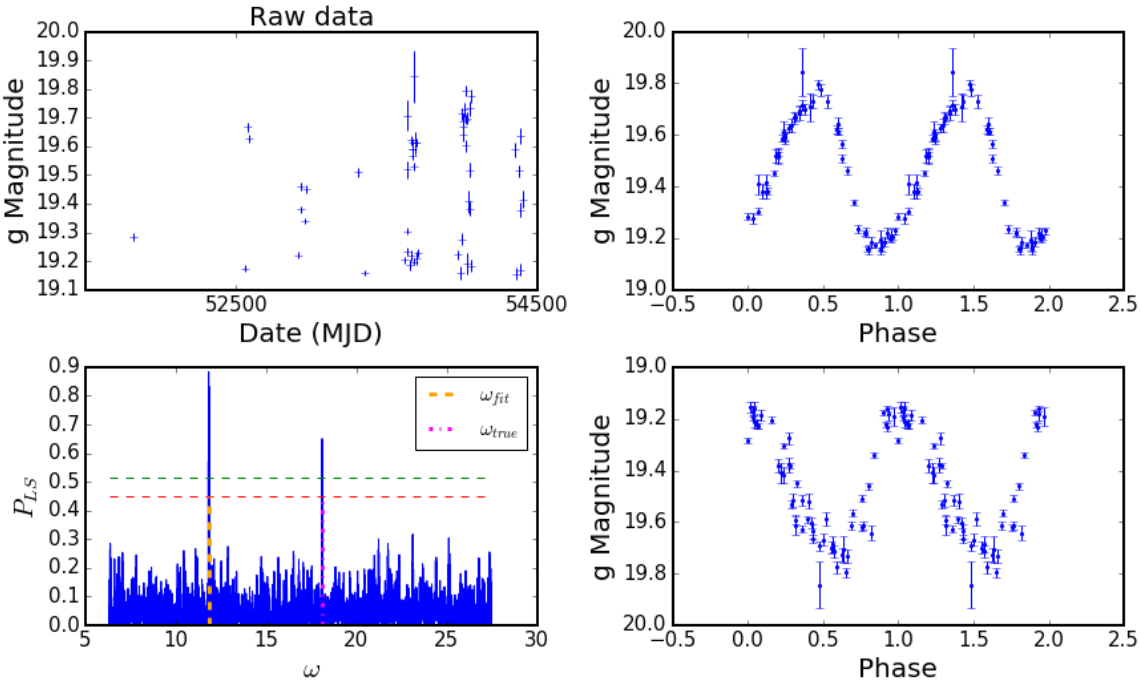

Figure 7. Same as Fig. 6, with $\omega_{true}/\omega_{fit}$ = 1.53. Here RR Lyr ID=470994 has a cited period of 0.346794 days ('P_{true}'), whereas period derived from the Lomb-Scargle periodogram is 0.531667.

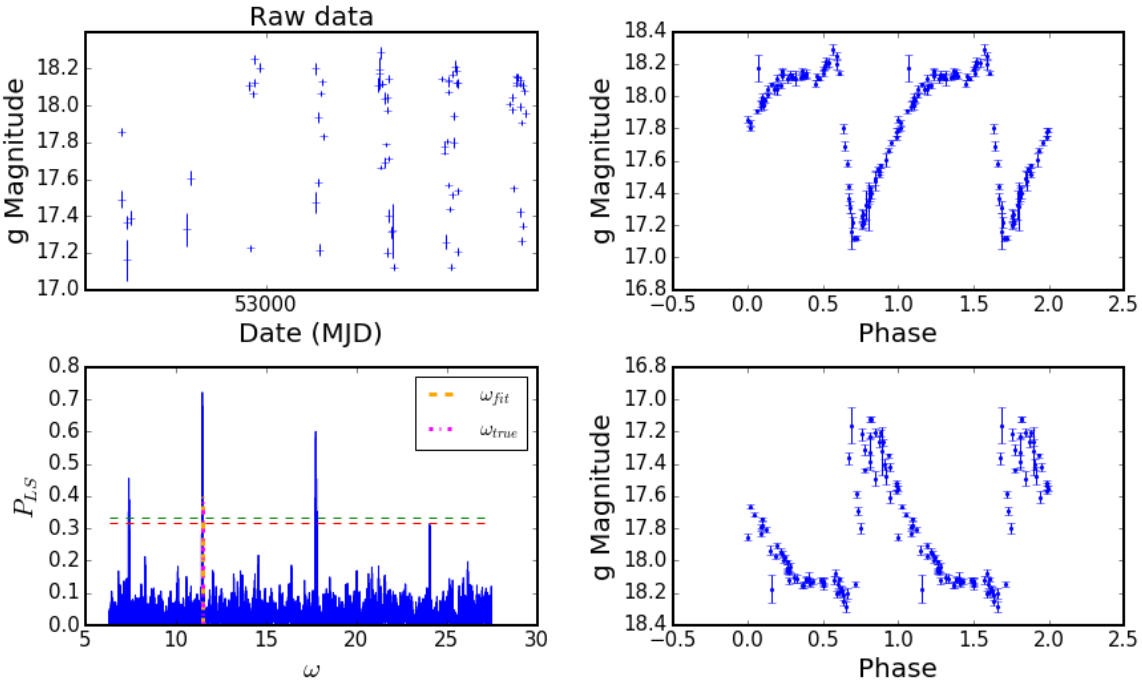

Figure 8. The same object as Fig. **??**, but using data downloaded using PDAC. Using PDAC data, the RR Lyr ID=13350 has a best-fit period of 0.547969 days, almost identical to the period found by (Sesar et al. 2010) of 0.547969 days. Top-right panel shows the phased PDAC data lightcurve folded on the 'true' period, and top-right : on the 'best-fit' period. Everything else as on Fig. **??**