The Cygnus Project- summary outline

Outline of the data pipeline.

1. Introduction

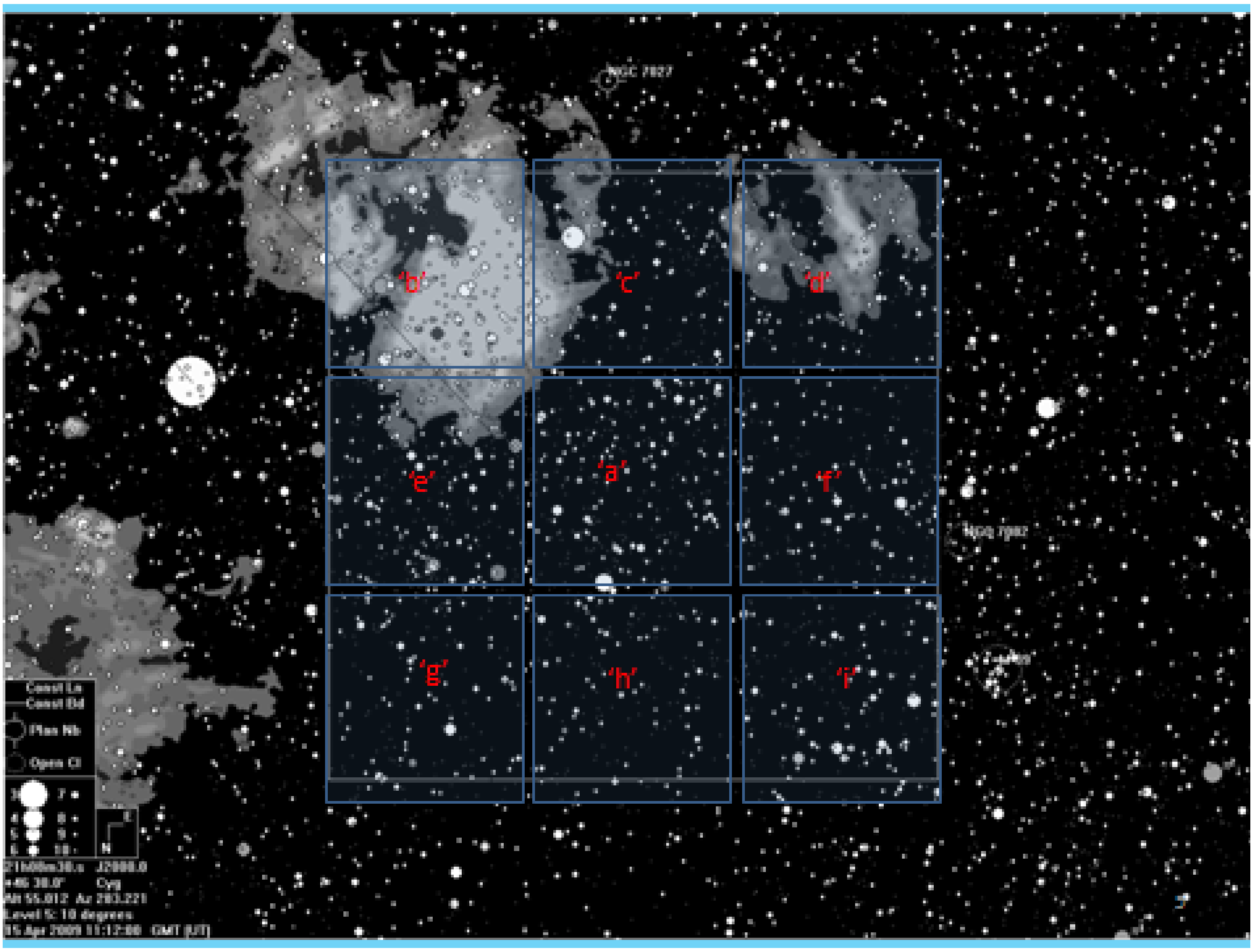
   This is a very brief summary of the key facts of this project which started life looking for extra-solar planets and was known as 'The Planet Project' but morphed into a mass new variable-star discovery project. A more detailed project description, including some technical detail will be published on this website in 2017 (The Cygnus Project – more details)

   Work started back in 2000 with a smallish camera but I really count the start as being from June 2003 when I bought the Apogee 16E camera and was able to have a large field of view.

2. Data Collection.

   Data was collected in Fits format and then converted to numbers and processed using programs all written in Dyalog APL. All the data I have collected is with the central y axis of the chip aligned along declination and central x along RA. I've changed telescopes over the years but the focal length has stayed very near to 750mm giving me a field of view close to 2.8 degrees square.

   Sky areas studied in Cygnus are as in the image below in - Deneb is the bright star to the left.



The areas in Cygnus are labelled ‘a’ to ‘i’, and by far the most data (over 73,000 images) was collected from area ‘a’.

I collected data from one area (labelled ‘p’ in some texts) in Auriga also, centred at radec: 05h 18m 0s, +41°50’ 0”. Nearly 21,000 images were collected from here.

3. Data Reduction

Many thousands of programs have been written (and continue to be) during this work. Their main features are the ability to handle many stars simultaneously, up to 10s of 1000s of them. The huge benefits of APL are that it is mathematics centred, super at handling large arrays and very concise.

The first step is to convert the Fits files to an image data only file and extract time and other info. The data is then converted to APL format. Then the images, in an apl workspace, are dark subtracted and divided by the flat field.

The next essential is to know where you are on the image in relation to the sky and this connection is formed by the picture coefficients.

3.1. Picture Coefficients

These are simply the mathematical relationship between the co-ordinates on the chip and the co-ordinates in the sky. I use a two dimensional (x and y) 4th order polynomial which needs a set of 15 coefficients. The programs find a set of 144 guide stars (one each in a 12 by 12 division of the image) and fit the coefficients to them.

> These 15 4th order coefficients are computed for each image and stored and then used in all subsequent calculations on that image. The ra and dec of any pixel (or any point on a pixel) on that image can be computed. Similarly, any ra and dec in the sky area covered by the image can be located on the image.
>
> So any event in the image can be located on the sky and position of any sky object, even if nothing is visible on the image, can be located in the image.
>
> That enables me, if necessary, to average the pixels of an invisible object thousands of times to, hopefully, make it visible!

So, any image can be probed at a specified ra/dec to retrieve data corresponding to that point in the sky for that 30 seconds of time. Next, we need to decide what objects to locate and study.

3.2 Starlists

The APL picture files are the initial basis for the data pipeline but they are not used throughout the process for obvious practical reasons: to study one star throughout the time period would otherwise need files retrievals on 70,000 images in area ‘a’, some 4.35 Tb. It was obvious early on that the data would need considerable pre-processing but it was essential not to lose the original data in the process. I’ve used two different routes to do this, these are the starlist based route and the ‘area’ route. The first which I have followed for most of the time since I started in 2003 either uses a starlist created from a chosen set of images, which I used exclusively from 2003 to 2011, or a starlist derived from a published one, such as the USNO ones. I originally rejected that approach because of the large number of errors in the USNO cats in Cygnus, but I’ve recently used them for other kinds of searches, more on that later. The starlist based strand uses either growth curves or ‘singlesets’.

The area method uses group sets, these are explained below.

The starlists are made by identifying the N brightest objects on a set of 30 good images. The ra and dec co-ordinates (hereinafter called radecs) of the objects in this list (obtained from the x,y and picture coefficients as above) is then compared to a published catalogue to get mag values and other IDs. When I started this work in 2003 the aim was to search for extra-solar planets so my catalogue was limited to the brightest 12,001 stars. I thought at the time that finding a 2% dip (as in HD29458) would not be possible for fainter stars. However, the 'noise' in the system, ie: variable stars- soon became of great interest* so I increased the catalogue to 17,000 and later again to 75,000 in area 'a'. However, I have so much data in area 'a' that, with averaging, at least 200,000 objects can be seen, but I've not studied all of them, yet.

*My interest was stimulated by noticing, for example, that star 12001 varied by over 4.4 magnitudes.

3.3 Using the Starlist

Having got a starlist and picture coefficients, the next step is to probe the reduced images (i.e.: images that have been dark subtracted, flat divided, converted to apl, multiplied by 10 and made integer) at the calculated positions. Again, there are several approaches to this. The most versatile sorting is into what I called 'singlesets' (ss).

3.3.1 Singlesets

This involves probing each image at the co-ordinates of each star in the starlist and extracting a square pixel set which varies in size depending on star brightness. So, after a longish night of, say, 500 images one has 500 picture sets each with 75,000 samples (one for each star making 37.5 million pixel sets). These are then rearranged into 75,000 sets of 500 samples, i.e.: 500 samples for each star. The ss approach is totally versatile in that further background tweaking can be done and, for complex objects, templates can be used to reduce interference by selecting a specific set of pixels, but it is storage and time heavy.

A later approach that I've used for the analysis of stars in areas 'b' to 'l' and in Auriga is to study the entire list via growth curves to identify stars that vary and that are affected by neighbours. Then a more select list of singlesets of those variable ones can be used to refine the data. That applies particularly to the very faint stars because they can then be studied pixel by pixel, tiny background errors removed and high quality averages made- i.e. based on averaging the pixel sets not averaging reduced data.

3.3.2 Growth Curves

Stars are particularly closely packed in Cygnus (which is why I chose it as a target) so the luxury of using a large collecting circle is not often available. In order to choose an optimum diameter (which may change with time during a night and seeing conditions) a choice of diameters for every star in every image gives great flexibility. Apart from choosing a diameter, the data to predict the loss as a function of diameter is also available so that if a small diameter has to be used the asymptotic value can be calculated. This leads to a lower noise set of data.

So, growth curves are essentially a set of light curves for a range of probe diameters around each star in the starlist. The project generated starlist for area 'a' is 75,000 long so each image is probed at that number of radec points with a set of collecting circles of different collecting areas. After some initial variations I have settled on 3 by 3 pixels square and 13 different circles increasing from 4 pixels diameter to 16 diameter. The small sizes are used for stars in close proximity to others. There is more detail on this in the main report, The Cygnus Project

3.3 Group Sets

The other method I use to find variable objects is the group-set approach.

Although it covers most of what one would want to do the starlist approach has two shortcomings:

a) Most of the pixels in each image is ignored so if something interesting happened between the catalogue stars it would not be noticed.

b) There may be an interesting star not in the list! In area 'a' there are in fact at least 150,000 measurable stars down to about magnitude 18. Not for me accurately measurable but some useful information can be obtained by averaging 20 plus images.

This approach is to divide every image into a number of small and equal areas and form sets of these sub-images. I used this method to search for transient happenings, in particular flares which may well happen in objects too faint to be in my starlist.

I've not in fact attempted to stack whole images but have found it easier to divide the images into squares, in fact 49 by 49 squares or 2501 of them, each 91 pixels on a side. Each square is centred on a precise ra/dec so that even over years (if I keep the orientation within one degree) the sub squares line up with a worst error of 2.4 arc-secs in a corner and that would very rarely happen. The sub squares are averaged in sets of 25 and the 91 pixels allows for a little overlap so the raw images are effectively reduced in storage requirement by about 20 times. Any of these star field averages can then be instantly compared with any other, from the same night to nights years apart. This system overcomes both problems, a and b. I've not taken full advantage of it yet because of other priorities but eventually all the variable stars in that 150,000 set will be listed and any events that happened also logged.

4 Finding Variability

Having got the data into a study-able form the interesting work starts, looking for variability. Knowing little about variable stars I classified them into very short period variables (VSPVs), SPVs, medium period variables (MPvs) and long period variables (LPVs) and eclipsing binaries. Of course there are sub categories, some stars produce spikes and dips, and some vary continuously. There are different techniques for each. The continuous VSPVs and SPVs are best found by Fourier and overlay[1] methods, the spiky or dippy ones overlay only. The analysis is done with varying degrees of averaging before analysis, from none up to about 30 samples. Great care is required when analysing averaged results or serious errors can result. The averaging must be by time interval of course. For MPVs and LPVs I just average each whole night's worth of data (for initial finding at least) and then compare the variability (i.e. the noise) of each set (with a variety of comparison stars for each) with the average for that brightness.

If the variable has a small range the 'fast' ones are much easier to find because one can do a periodogram using a week's data or a month or even a year which is a huge amount of averaging for a 0.1day period say. Of course, if the period changes, then a smaller window is necessary.

A big problem is aliases. I'm sure I have many errors due to that and there will be a separate report detailing several alias investigations.

4. Reporting

I only gradually developed a standard report format for each star so many of the reports in this website will be in different forms initially because the software helping to produce them evolved. Hopefully they will be standardised eventually. The standard form starts with a bitmap, or set of zoom bitmaps showing the environs of the star and listing those around it. Then there is basic data about the star and that is followed by comments, deductions and plots of light-curves, periodograms, phase plots etc. Also in the plots ‘ref’ mean comparison star of course.

note 1: Where you slice up the data into ‘period’ lengths, overlay them, then add and search for patterns, doing that for tens of thousands of periods to produce a periodogram.