MOTIVATION FOR MAPPING

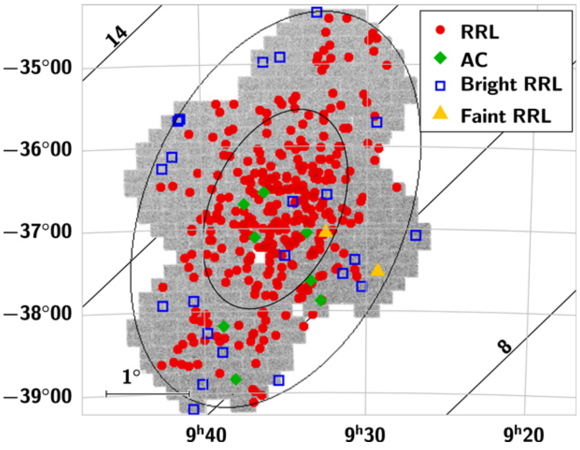
Pulsating variable stars that are periodic are frequently used in determining distances because a relatively easily and reliably measured characteristic (the period of pulsation) can lead us to the luminosity of the stars. So with a measurement of the apparent brightness of the star, we can figure out the distance. This is a STANDARD CANDLE method, which uses the inverse square law for light:

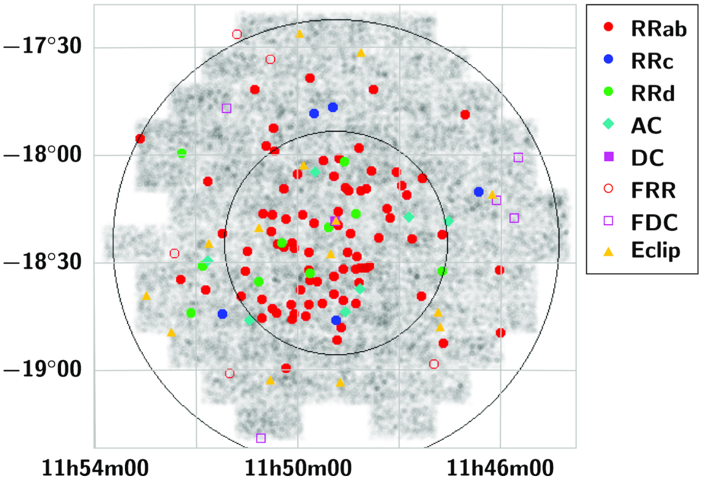
$$f = L/4\pi d^2$$

(where f is flux (or brightness), L is luminosity (or power), and d is distance).

The distance allows us to figure out the position of the star in 3-D space, and if we find groupings of stars in 3-D space (especially if they are also *moving* in the same direction), it can be a sign that a larger object (like a star cluster or galaxy) is there or was there.

Some examples are dwarf galaxy satellites of the Milky Way: the Antlia 2 dwarf (Vivas+ 2022), and the Crater 2 dwarf (Vivas+ 2020). In both cases concentrations of variable stars are seen in the sky.



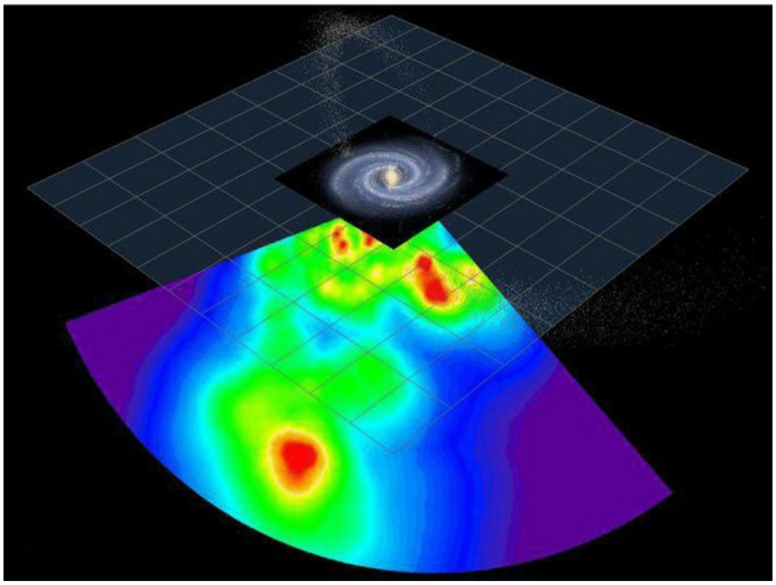


Sesar+ (2010) shows an example (below) of mapping in 3-D space using RR Lyrae stars from the earlier Sloan Digital Sky Survey (SDSS) in the northern hemisphere.

WHY DOES THIS WORK?

It helps a little to know the engine behind a star's pulsation in order to understand why there is a PL relationship... it comes about because pulsations involve FORCE IMBALANCE, where gravity is one of the forces. This sets a TIMESCALE for the variation. Here's a short argument:

- gravitational acceleration for gas near the star's surface: $g \approx \frac{GM}{R^2}$
- if no force opposes gravity, distance travelled is $d = \frac{1}{2}gt^2$



- inward fall covers approximately half a pulsation period ($t \sim \frac{1}{2}P$), and some fraction f of the star's radius ($d \sim fR$), so

$$\begin{aligned} f \cdot R &\sim \frac{1}{2}\left(\frac{GM}{R^2}\right)\left(\frac{1}{2}P\right)^2 \\ \frac{GM}{R^3} \cdot P^2 &\sim 8f \\ \frac{M}{\frac{4\pi}{3}R^3} \cdot P^2 = \bar{\rho} \cdot P^2 &\simeq \frac{6f}{\pi G} \approx \text{constant} \end{aligned}$$

While this isn't exact, it leads us to expect that pulsation period and MEAN DENSITY of a star should be related. LESS DENSE STARS TAKE LONGER TO PULSATE.

The connection to luminosity comes in via the important relation

$$L = 4\pi R^2 \sigma T_{\text{eff}}^4$$

decribing the power output of a hot spherical object (a "blackbody"). When put in magnitude form,

$$M_{\text{bol}} = -5 \log_{10} R - 10 \log_{10} T_{\text{eff}} + C_{onst}$$

We can eliminate the radius from this using the $\bar{\rho} \cdot P^2$ relation (although a mass dependence comes in), and the temperature can be converted to a color to get a period-luminosity-color or PLC relation:

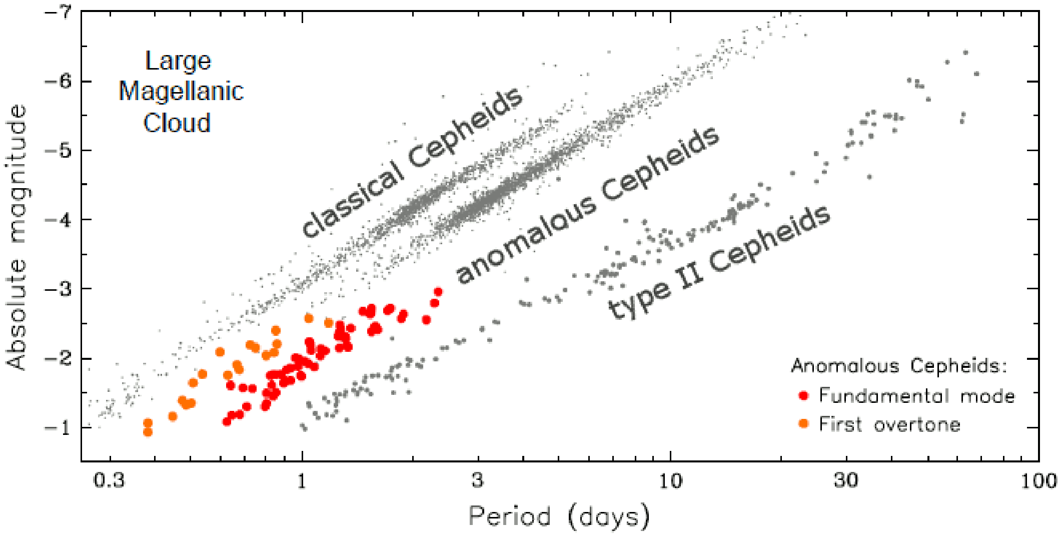
$$\langle M_V \rangle = \alpha \log_{10} P + \beta (B-V)_0 + \gamma_V$$

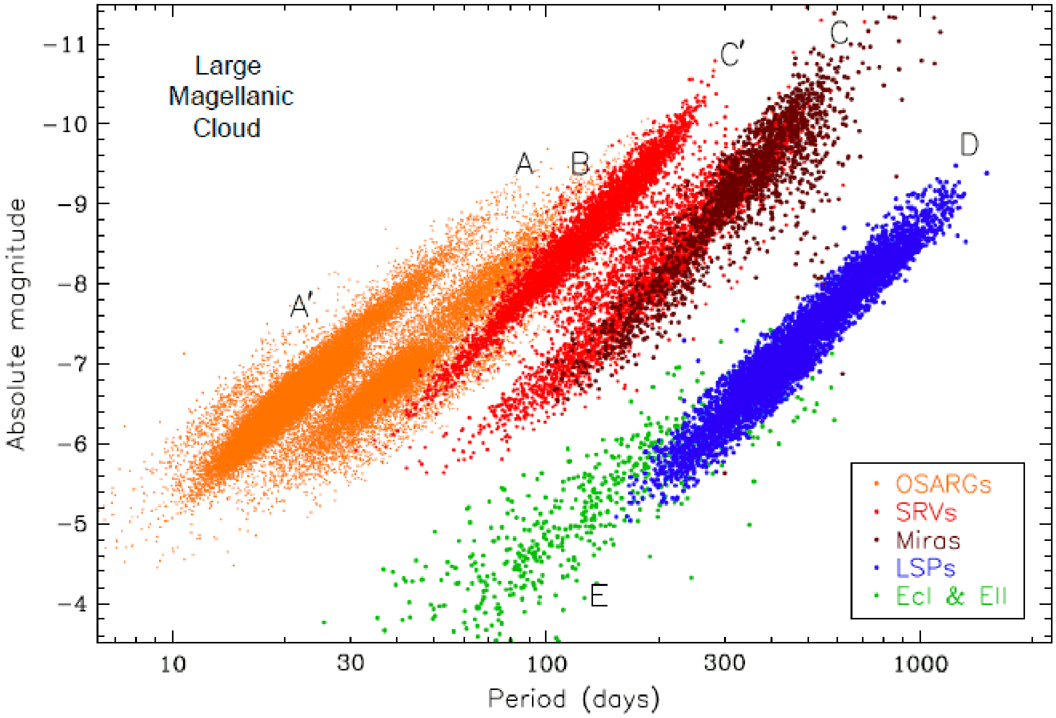
This is sometimes known as a "Leavitt Law", after the discoverer, Henrietta Swan Leavitt. Some examples of these linear relations are shown in the fiures below for different types of variable stars.

Based on these arguments, we expect the coefficients of the PLC relation to behave like so:

- $\alpha < 0$: increased period $\longrightarrow$ lower density $\longrightarrow$ higher luminosity
- $\beta > 0$: decreased temperature (which means increased color) at constant radius (to keep P constant) requires lower luminosity

These expectations are borne out in studies comparing pulsating stars of the same type in the same mode of pulsation. The type and mode of pulsation can often be identified from a combination of the period of the variation and the shape of the light curve.





PULSATING STARS

Before getting started, remember that we are focusing on *periodic* variable stars that vary in a consistent way, cycle after cycle. For these variables, we can make a *phased* light curve, that maps observations from every cycle onto just one representative cycle. The phase ϕ is something that only varies from 0 to 1, and represents the fraction of a cycle the star has gone through at a particular time. This can be calculated mathematically using the modulus ("mod") function:

$$\phi_i = \mathrm{mod}\left(\frac{t_i - t_0}{P}\right)$$

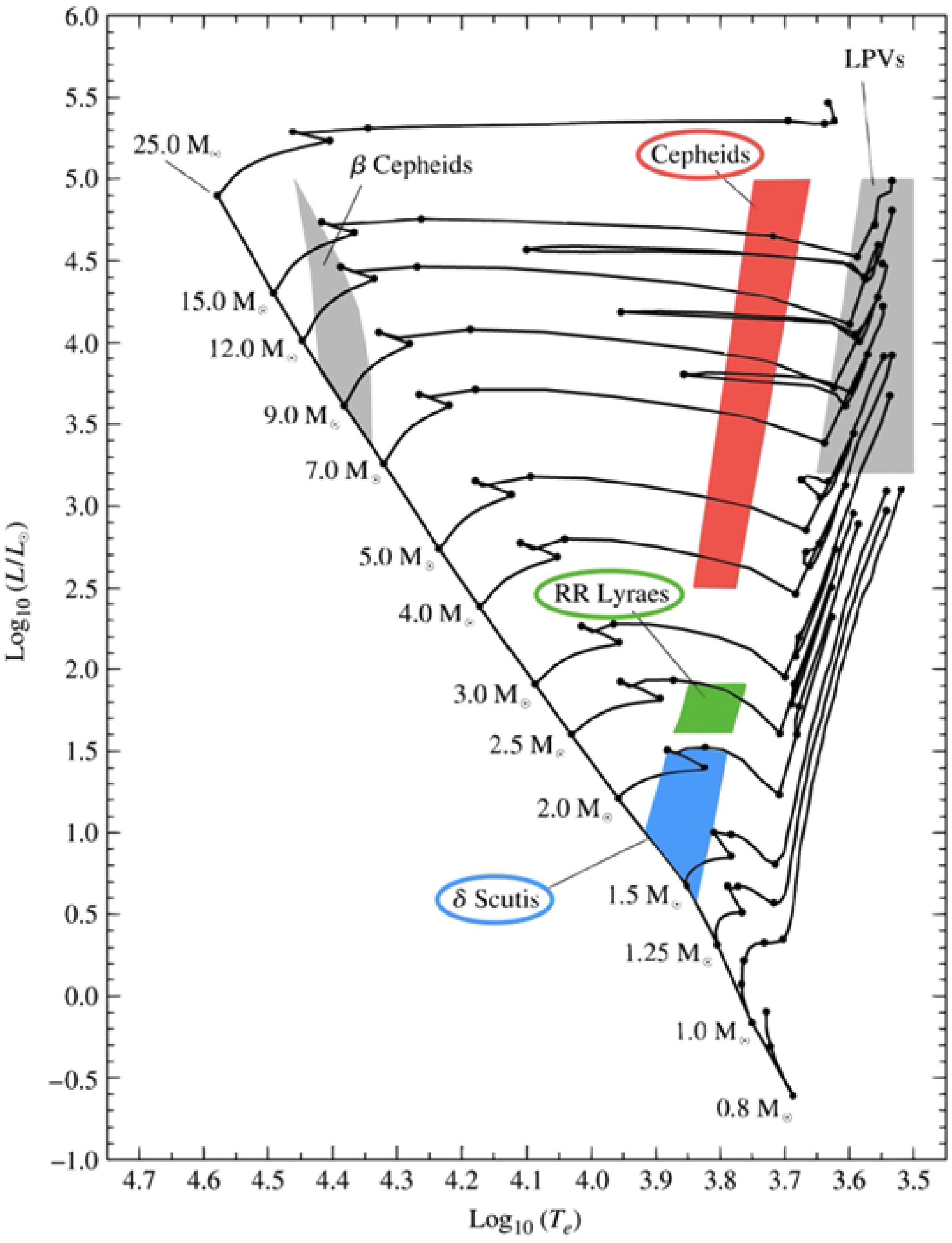
where t_i is the time of an observation, t_0 is the known start of one of the cycles, and P is the period. The mod function takes the decimal *remainder* of the division. (The integer part is the number of cycles that have passed since t_0.) Using this, observations at any time can be used to fill out a picture of the brightness variation.

We will focus on three types of pulsating variable stars:

- RR Lyraes: core He burning stars (with an H burning shell)
- Cepheids: supergiants that result from more massive stars
- Miras and long-period variables (LPVs): low-mass asymptotic giant stars (with H and He burning shells)

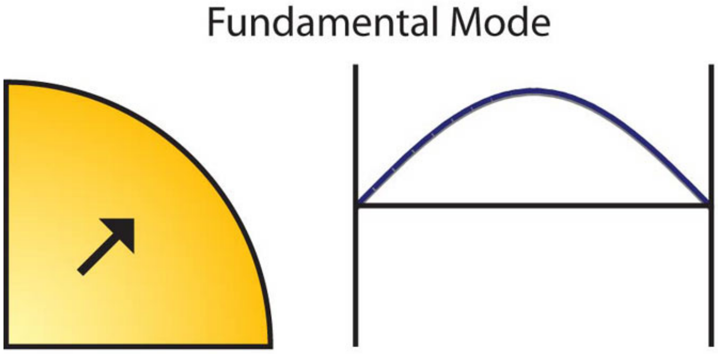
One common characteristic is that all of these are much higher in luminosity than the Sun, and are cooler and redder. The HR Diagram below shows the characteristics the different types generally have.

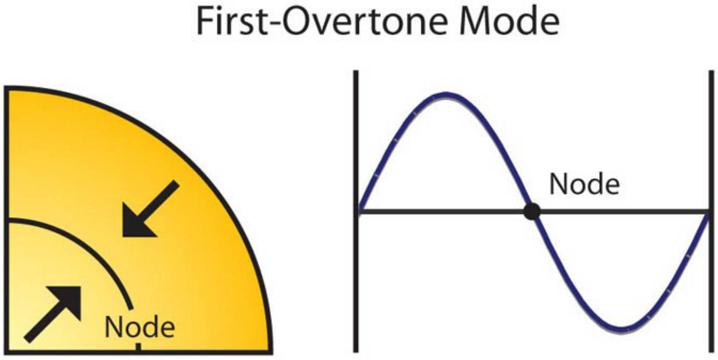
Type:	RR Lyr	Ceph	LPV
P range (d)	$0.3 - 1$	$1 - 70$	$30 - 1000$
L range ($L_\odot$)	$30 - 40$	$600 - 4 \times 10^4$	$10 - 10^4$



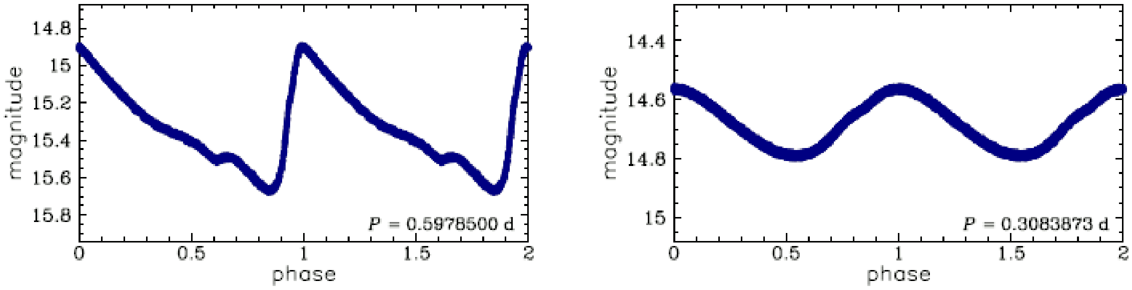
RR LYRAES

RR Lyraes, and the other types of pulsators we are focusing on, are radial pulsators, where the stellar layers move radially inward and outward. But different stars can have different (or multiple!) modes of oscillation. The most common are the fundamental mode (where the layers all move inward or outward together; called RRab or RR0 variables), and the first overtone (where different layers can be moving in opposite directions, with a "node" or stationary layer between; called RRc or RR1 variables). An illustration is below. The fundamental mode is generally longer period (lower frequency) than the first overtone, similar to the situation for vibrating strings or organ pipes. Some stars can pulsate in both modes simultaneously (called RRd variables), just as musical instruments can have multiple harmonics.





The fundamental mode generally has a larger swing in brightness over a cycle (the "amplitude"), and a sharper rise in brightness than the first overtone does. Some example light curves from the OGLE Survey are shown below.

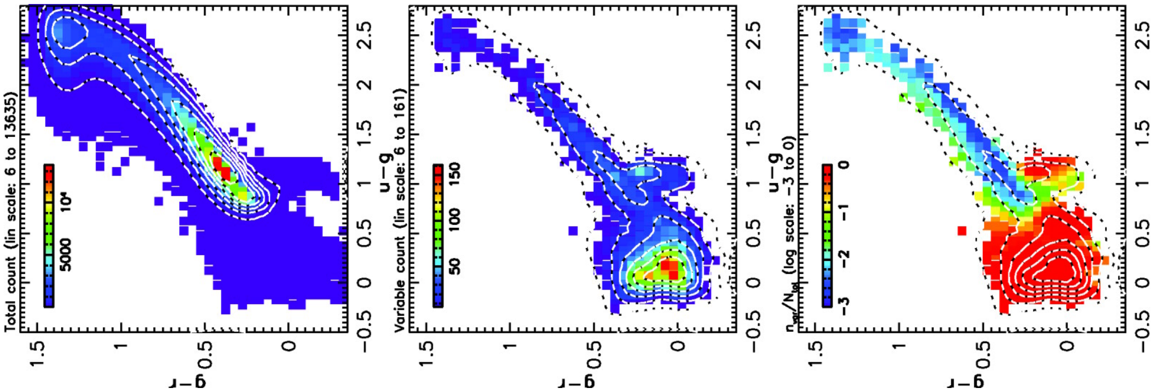


PROS and CONS of RR Lyr Variables:

- Pro: much less variation in luminosity from star to star than the other types because they arise in stars are in a very specific mass range and specific stage of life: low-mass (around $0.7 M_\odot$), burning He in their cores (following the main sequence stage when they were burning H). As a result, they can provide very good distances.

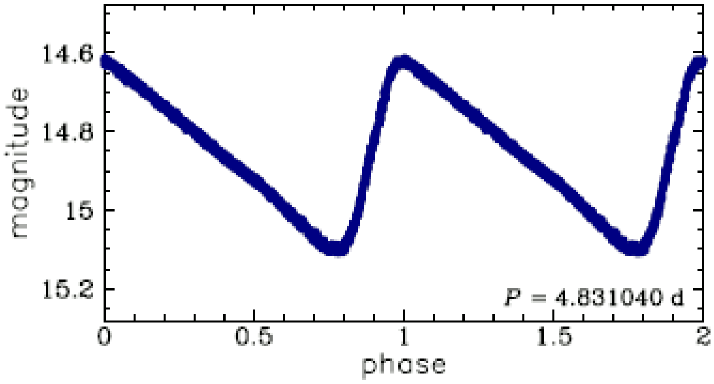
- Pro: Cover a limited range of color, and can be selected pretty efficiently that way. See the figure below. RR Lyraes stand out with $u - g \sim 1.1$ and $g - r \sim 0.2$. (Left is a plot for all stars, middle is a plot for detected variables, right is a plot for the frequency of variables.)

- Pro: Present in fairly large numbers in old, low metal content populations of stars.
- Con: Less luminous than other types, and so aren't as easily detectable at large distance.



CEPHEIDS

Cepheids cover a much wider range of luminosities, and they arise from much more massive stars that are crossing the HR Diagram after they have completed the main sequence. An example light curve from the OGLE survey is shown below.

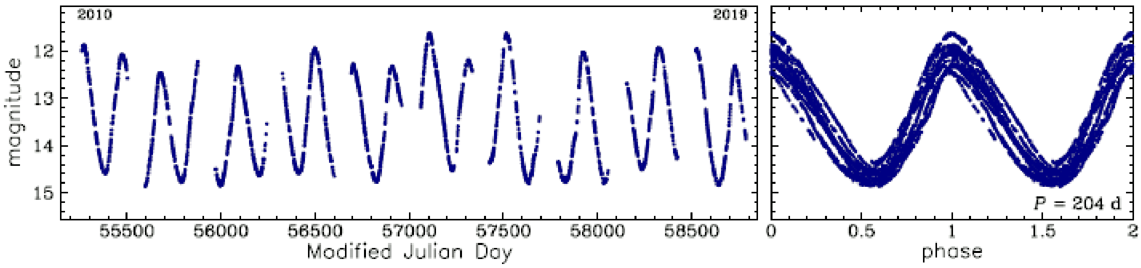


PROS and CONS of Cepheid Variables:

- Pro: Much more luminous than RR Lyrae, and have been used to measure distances to other galaxies as a result.
- Con: Not usually present in old (and metal-deficient) populations of stars because of their masses. Exceptions are Anomalous Cepheids and Type II Cepheids (including from lowest luminosity to highest luminosity, BL Her, W Vir, and RV Tau types), although these are generally lower in luminosity than the Type I Cepheids.

LONG-PERIOD VARIABLES:

These variables also cover a pretty wide range in luminosity, but also have much longer periods. Frequently their pulsation is irregular, with more than one mode occuring at the same time. An example of the light curve and a phased version for a Mira variable from the OGLE Survey are shown below.



PROS and CONS of LPVs:

- Pro: Very luminous and red, which makes them detectable at large distance and through a lot of gas and dust.
- Pro: Very large luminosity variations, making the variation easy to detect.
- Con: Long period requires much longer effort to characterize them
- Con: Multiple pulsation modes can make the proper PL relation difficult to identify.

MAGNITUDES AND COLORS

A Quick Primer

Fundamentally, what we are trying to measure in astronomy is "flux" (or brightness), related to how many photons or how much energy is being received per second per unit area. Fluxes in astronomy cover many powers of 10 in amount, but can't be negative. To make this easier to carry around, we use a logarithmic scale that very roughly corresponds to how your eye and brain experience light. This is the "**magnitude**" scale.

The fundamental equation for magnitudes is

$$m_1 - m_2 = -2.5 \log_{10} \left(\frac{f_1}{f_2} \right)$$

A *ratio* of fluxes (f_1/f_2) is on the right because we are always going to be comparing the brightness of different objects. In the logarithmic magnitude scale, this ratio is turned into a difference. Some consequences:

- Fainter objects have *larger* magnitudes... it was originally a kind of ranking system (the brightest stars in the sky were "first magnitude")
- A rule of thumb is that if two fluxes differ by a factor of 100, they will differ by 5 magnitudes.
- A 1% difference in flux translates into a 0.01 mag difference. (There is a similar translation for other percentage differences, as long as the percentage is small.)

Astronomical colors are a magnitude-based way of describing how bright one object is in different wavelengths. It probably isn't too hard to grasp that the ratio f_B/f_R (where f_B is flux in the blue part of the spectrum, and f_R is flux in the red part of the spectrum) tells you about color balance: a large ratio means the object would look bluer, and a small ratio means the object looks redder. To turn this into an astronomical color, we use an equation like

$$(B - R) = -2.5 \log_{10} \left(\frac{f_B}{f_R} \right) + C_{BR}$$

This means:

- Bluer objects have smaller color values, and redder objects have larger color values.
- Colors aren't affected by distance, unlike the individual magnitudes. A more distant star will have both of its fluxes decrease in the same way.
- Dust in space produces *reddening* and *attenuation* of starlight. In other words, it affects shorter (bluer) wavelengths more than longer (redder) wavelengths.

Importantly for astronomy, colors are related to surface temperature for a star. This is one of the most easily observed characteristics of a star, and pretty useful for grouping stars.

Magnitudes and Errors

Magnitudes in astronomy are *comparisons* of one sort or another — they can be comparisons of fluxes (photons per unit area per second) measured at different times for the same star, or fluxes for different stars. For the Difference Imaging Analysis (DIA) in Rubin data, there is a reference flux (determined from some template images) that our measurements are being compared to. If f_0 is the reference flux and Δf_i is a difference flux (measured on individual images with the template subtracted), then we can define the magnitude as

$$m_i = 25. - 2.5 \log_{10}(f_o - \Delta f_i)$$

The quantity in parentheses is the flux that would be measured on image i in ideal conditions ($f_i = f_0 - \Delta f_i$), and the 25 is our definition of the magnitude for a star that result in a flux of 1 count on an image (making the logarithm 0).

The uncertainty in a magnitude is related to the uncertainty in the flux, and how it compares to the size of the flux. The uncertainty can be written

$$\sigma_m = 2.5 \log_{10} \left(1 + \frac{\sigma_f}{f_0 - \Delta f_i} \right)$$

The larger the size of the uncertainty (in comparison to the flux), the larger the magnitude error. But because of the way magnitudes work (see above), a 1% uncertainty in flux translates into a 0.01 magnitude uncertainty.

VARIABILITY

The Weighted Average

Before determining whether an object is variable, we have to have a good idea of what its typical magnitude is. This is usually determined from within a set of data. Rather than simply average the magnitude measurements, we want to put more confidence in measurements that have less uncertainty, so we can weight each measurement by $1/\sigma_m^2$:

$$\bar{m} = \sum_i \frac{m_i}{\sigma_i^2} / \sum_i \frac{1}{\sigma_i^2}$$

Measurements with larger uncertainties will contribute less to the top sum because of the division by the uncertainty squared.

The χ^2 Test

We can get some idea of whether an individual measurement deviates significantly from the mean by calculating the quantity

$$\delta m_i = \frac{m_i - \bar{m}}{\sigma_i}$$

This compares the deviation to the uncertainty, and larger positive or negative values are more likely to be significant deviations. (If the uncertainties are calculated properly, approximately 67% of measurements should be within $1\sigma_i$ of the mean.)

But we do have more than one measurement. So we can calculate a statistic that will give us a feeling for how likely it is that a particular star is significantly variable. The χ^2 statistic can do that:

$$\chi^2 = \sum_i \left(\frac{m_i - \bar{m}}{\sigma_i} \right)^2$$

We square the quantity in parentheses because either positive (faintward) or negative (brightward) deviations from the mean are potentially interesting. By summing these, we get an idea of how the *whole group* of measurements deviates from the mean.

The one unfortunate aspect of χ^2 is that the value will generally increase as the number of measurements increases, so there isn't a nice well-defined range of values where we can say that χ^2 is telling us that the object is variable. This is solved with the *reduced* χ^2:

$$\chi^2_\nu = \frac{\chi^2}{N-1} = \frac{1}{N-1} \sum_i \left(\frac{m_i - \bar{m}}{\sigma_i} \right)^2$$

where N is the number of measurements. As a rough guide, most non-varying stars will have $0 < \chi^2_\nu < 2$, with 1 being the most typical value. A value of 0 would be very unusual: none of the measurements deviate from the mean! Noise will cause some natural variation.

Another nice aspect of the reduced χ^2 is that the most common value for a non-varying object doesn't change with magnitude - it should still be near 1. However, χ^2_ν does lose sensitivity as you look at fainter and fainter stars — it is harder to identify truly varying stars because the uncertainties are going up.

The Welch-Stetson Statistic I_{WS} and the Stetson J Statistic

Another (often more sensitive) statistic is I_{WS}. It uses the same quantities

$$\delta m_i = \frac{m_i - \bar{m}}{\sigma_i}$$

to get an idea of how deviant individual measurements are. However, it uses these deviations in a different way. The idea is to look for *correlated* deviations. If a measurement deviates significantly from the mean, it is pretty likely that another measurement near it (in time) will deviate in the same direction (both measurements brighter than the mean or both measurements fainter than the mean). So the statistic is

$$I_{WS} = \sqrt{\frac{1}{n(n-1)}} \sum_{\text{pairs}} \delta m_i \, \delta m_j$$

This is a bit more difficult to calculate. To compute it, you look at successive *pairs* of measurements (labeled i and j) that are close together in time (generally no more than a few minutes apart), and multiply the deviations together. If both measurements are brighter (or fainter) than the mean, the product will be positive, and this will make I_{WS} larger. (These are correlated measurements.) If one measurement is brighter than the mean and the other is fainter than the mean, the product will be negative, and it will make I_{WS} smaller.

The number of pairs n is used in normalizing I_{WS} so that its typical value is 1 for a non-varying star. (For noisy data, there will typically be a mixture of correlated and anti-correlated measurements, and I_{WS} should be near zero. Negative values for I_{WS} are possible in principle, but that would imply that most pairs of measurements are anti-correlated, or in other words, the measurements bounce back and forth brighter than and fainter than the mean. That would be weird.)

Some slight changes to the I_{WS} statistic make it a bit more robust. Instead of δm_i, it accounts for the fact that the observation was used in the determination of the mean for that filter:

$$\delta m_i = \sqrt{\frac{n}{n-1}}\frac{m_i - \bar{m}}{\sigma_i}$$

Single observations (without others taken near in time) can also be included with a different formulation $P_k = (\delta m_i)^2 - 1$, while pairs are still handled like $P_k = \delta m_i \cdot \delta m_j$. The Stetson J statistic is then

$$J = \frac{\sum_k w_k \, \mathrm{sgn}(P_k)\sqrt{|P_k|}}{\sum_k w_k}$$

where the w_k are weights and the "sgn" function just takes the sign of P_k.

Using the Statistics

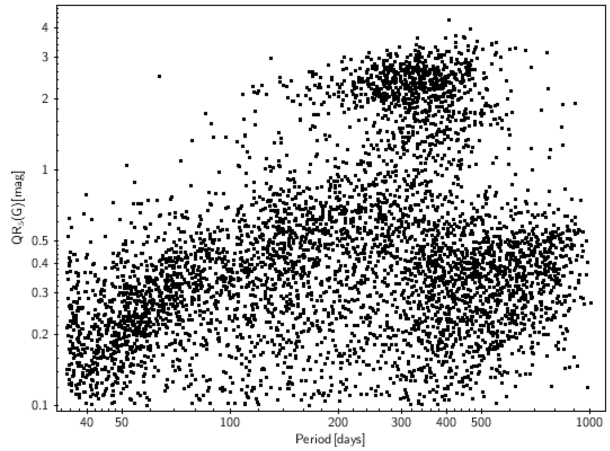
As with χ^2_ν, larger values of I_{WS} imply a likelihood of significant variations. However, there could also be systematic effects (like image processing errors) that can introduce fake variation into the light curves. In the end, you need to look at the light curves. However, the statistics help guide you in using your time most effectively — you can ignore the majority of stars that don't give large values for the variability statistics.

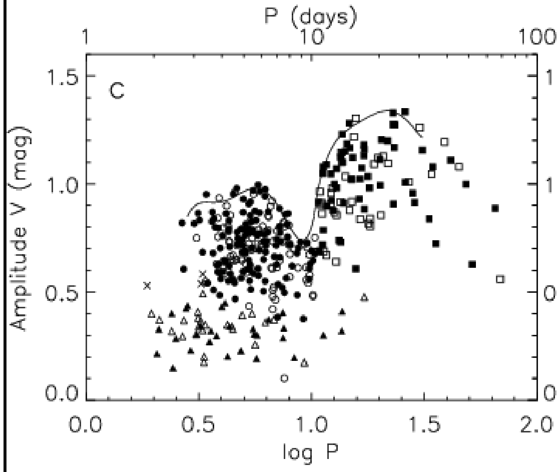
Period-Amplitude Relations

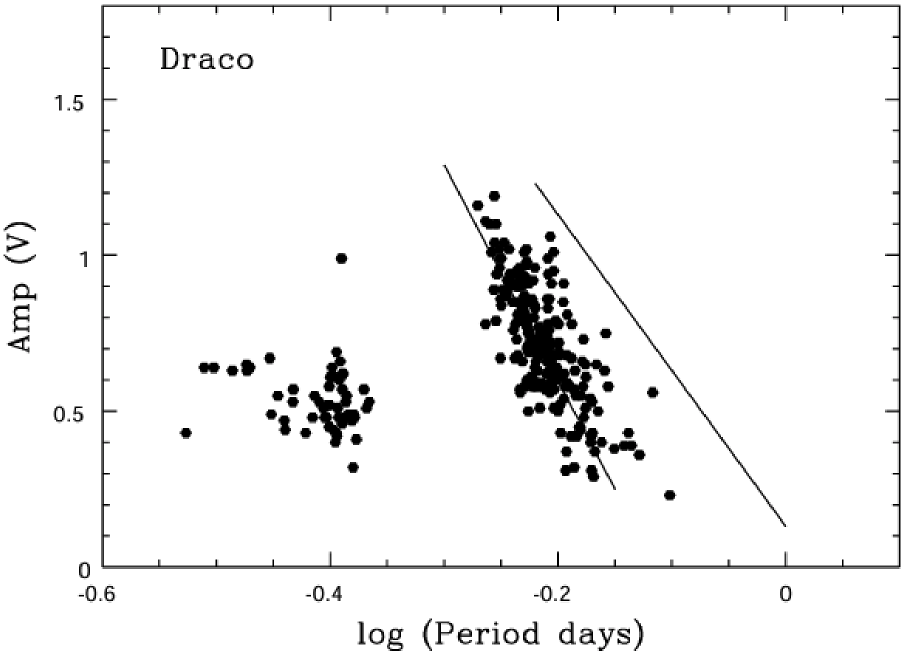
A common characteristic of pulsating variables is that there are often connections between the pulsation period and amplitude. This may make some sense in the context of musical instruments: longer periods (lower frequencies) often allow for larger vibrations in strings or in the bodies of the instruments.

Here are some observed period-amplitude relations. On the left is a relation for nearby LPVs (using the 5-95% interquantile range $QR_5(G)$ instead of an amplitude). On the whole, the amplitudes are pretty large. The largest amplitude group are mostly "Mira" variables, and they can mostly be identified with just their amplitudes. On the right is a period-amplitude relationship for Cepheid variables. Circles and squares are fundamental mode pulsators, and triangles are first overtone. There is an apparent distinction between Cepheids above and below a period of 10.5 d.

On the bottom is a similar diagram for RR Lyrae stars in a dwarf spheroidal galaxy called Draco. the larger period stars are fundamental mode pulsators, and the lower period (and usually amplitude) ones are first overtone. There are found to be different trends for different star clusters or galaxies (possibly due to chemical composition differences).







EXTINCTION

Along many sightlines, especially ones near the disk, there will be significant gas and dust extinction that will affect distance measurements by making the variable star seem to be less luminous. There are ways of minimizing this effect though.

REDDENING is frequently easier to measure than extinction, and it is defined using colors. For example,

$$E(B-V) = (B-V)_{\rm obs} - (B-V)_0$$

where $(B-V)_0$ is the unreddened color of the star. There is usually a proportional relationship with extinction, like

$$R_V = \frac{A_V}{E(B-V)} \approx 3.1$$

An "extinction-free" magnitude would have the extinction subtracted, but we can't generally measure it:

$$V - A_V = V - R_V \cdot E(B-V)$$

We can get close by subtracting a quantity related to the color. This is a WESENHEIT MAGNITUDE, and it only involves things we can measure:

$$W = V - R_V \cdot (B-V)$$

The observed color already has some (unknown) reddening in it, so we will correct for it (and more) by subtracting it.

W has the benefit of being the same, regardless of whether there is or isn't any extinction. For a star with no extinction,

$$W_0 = V_0 - R_V \cdot (B-V)_0$$

For the same star with extinction (and reddening),

$$\begin{aligned} W &= V - R_V \cdot (B-V) = (V_0 + A_V) - R_V \cdot [(B-V)_0 + E(B-V)] \\ &= V_0 - R_V \cdot (B-V)_0 + (A_V - R_V \cdot E(B-V)) \\ &= W_0 \end{aligned}$$

because the last term in parentheses is zero as long as the extinction and reddening have a constant ratio!

RUBIN OBSERVATORY AND ITS DATA

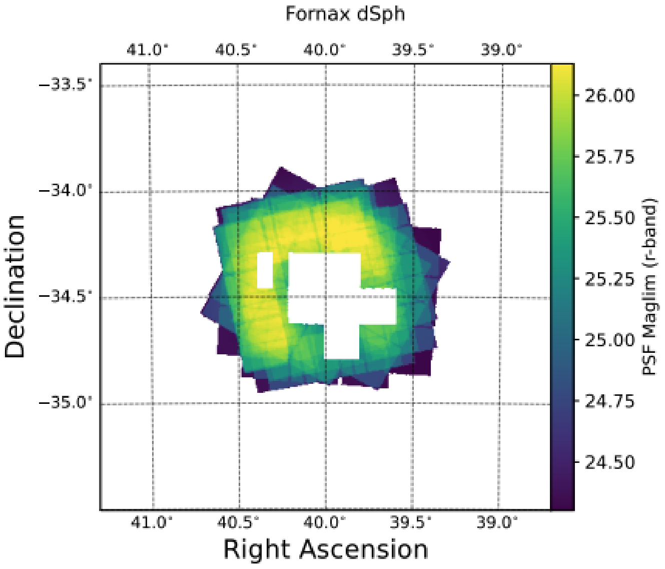
Because of the size of the datasets that will be generated by the Rubin Observatory, the data are going to be hosted on the cloud at the Rubin Science Platform. You can access the data there via:

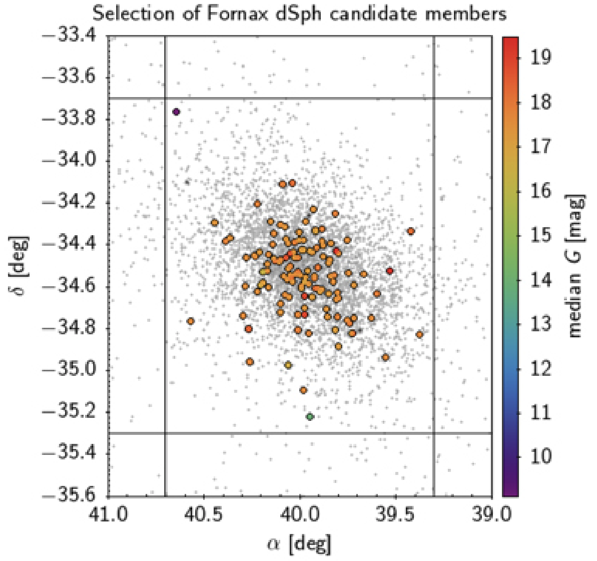
- Portal aspect: a web-interface to look at the different catalogs
- Notebook aspect: a Jupyter Lab-based platform that allows you to create and modify notebooks, as well as do analysis of data on the remote computer

The platform has many tutorials for both aspects that will be good to go through to get familiar with the interfaces.

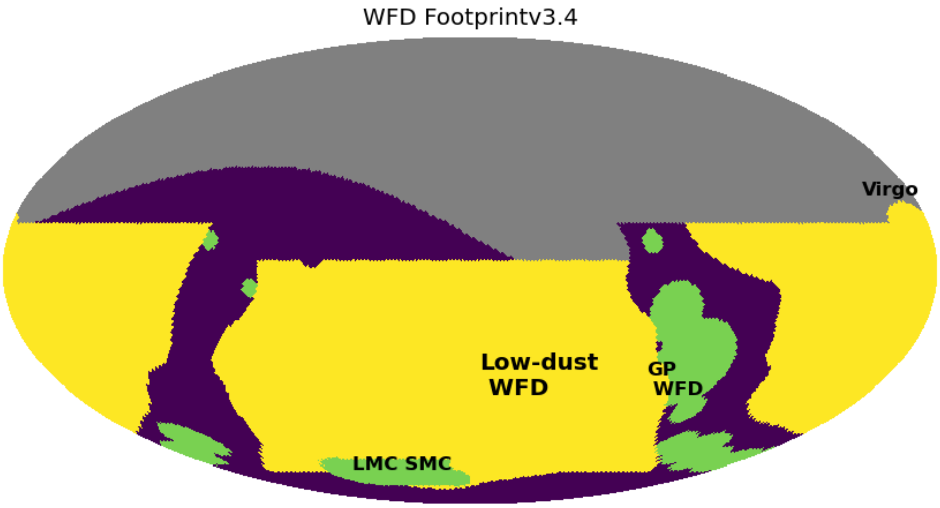
Some important details about the dataset:

- Before the Rubin Observatory started taking data, they developed a practice dataset called Data Preview 0 DP0.2, which involves simulated images of galactic and extra-galactic objects, and the photometry catalogs to go along with them. This dataset covers 300 deg^2 and 5 simulated years of observation. We may try do some testing on this.
- This summer, they will release a small amount of *real* data in Data Preview 1 (DP1). A couple of the fields that will be observed have known variable stars in them, with the Fornax dwarf spheroidal galaxy being the most relevant for this project. Candidate LPVs have been identified by the *Gaia* spacecraft (Lebzelter+ 2023), as shown below.





- For perspective, the full survey is shown in the image below (where the S celestial pole is at bottom, and celestial equator would stretch across the middle of the map). Most of the southern sky (about 19600 square degrees) will be surveyed in the "Wide Fast Deep" (or WFD) survey. This part is expected to have relatively low gas and dust extinction because it is looking out of the Galactic plane.



- Cadence: For the most part, Rubin is expected to visit each part of the visible sky (in the southern hemisphere) about every 3 days. At each visit, a couple of images in different filters will be taken.

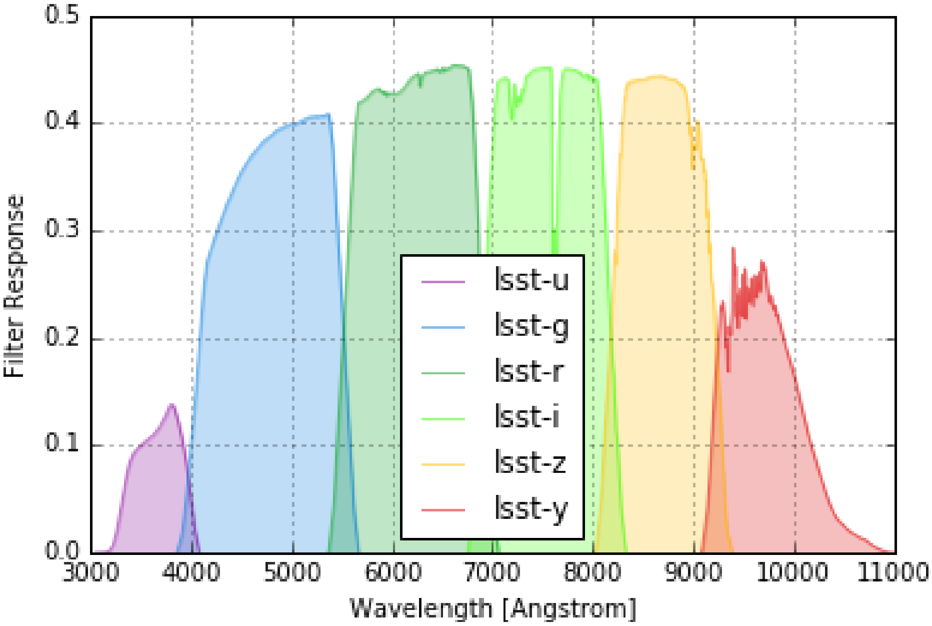
- Filters: Six filters will be used, labelled *ugrizy* from shortest wavelength (u in the ultraviolet) to longest wavelength (y in the infrared). This is to get useful color information for almost any star. See below for their transmission curves.

For the catalogs, the DP0.2 Schema Browser is a quick reference for the meaning of data columns in different catalogs. Probably the most important catalog for us will be the "DiaObject" catalog that contains information on *how* stars are varying. This is information we will want to use to *categorize or classify* objects.

Some "BIG" goals for the summer:

- Develop code for identifying likely variable stars in the catalogs that the Rubin Observatory will be keeping, and attempt to classify them using the information in the catalogs (like colors, and variability statistics). We can start by testing on the Fornax dwarf galaxy field that will be released this summer.

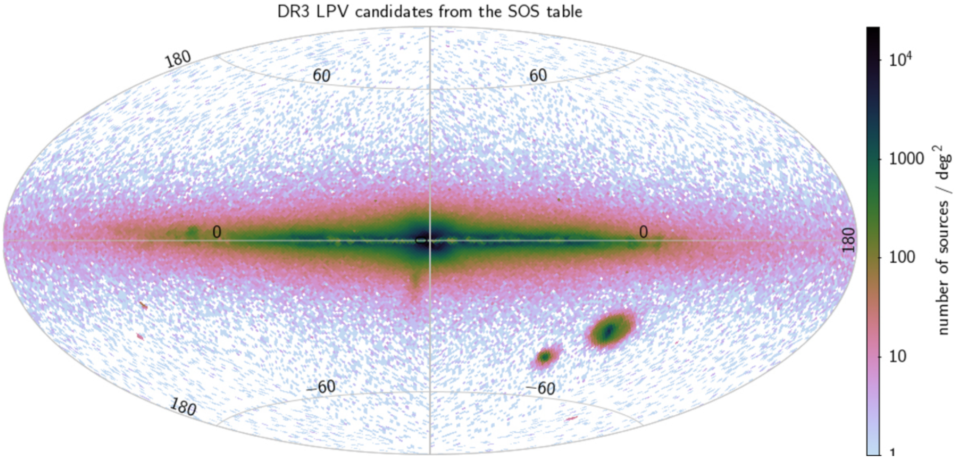
- With pulsating variables identified, determine a pulsation period from a time-series analysis. This will probably involve a periodogram to identify the pulsating frequencies with the most power, and examining phased light curves.



- Compare lists of identified variable stars with other catalogs to test our effectiveness in detecting variable stars in the Rubin data.
- Use a period-luminosity relationship to determine distance, and examine the positions in 3-D, using the distance, and galactic latitude and longitude. Can we identify the structure of a known satellite galaxy of the Milky Way? Plotting variable stars in galactic coordinates can help start to visualize likely satellites. The galactic center is in the direction of galactic longitude $\ell = 0°$ and galactic latitude $b = 0°$ and 8 kiloparsecs away. `astropy.coordinates` should allow you to convert your RA and DEC to galactic coordinates. Views with the stars projected onto the galactic plane ($Z = 0$) and projected onto the plane through the galactic center ($Y = 0$) can be nice for 2-D plots, but a 3-D plot you can move around maybe?

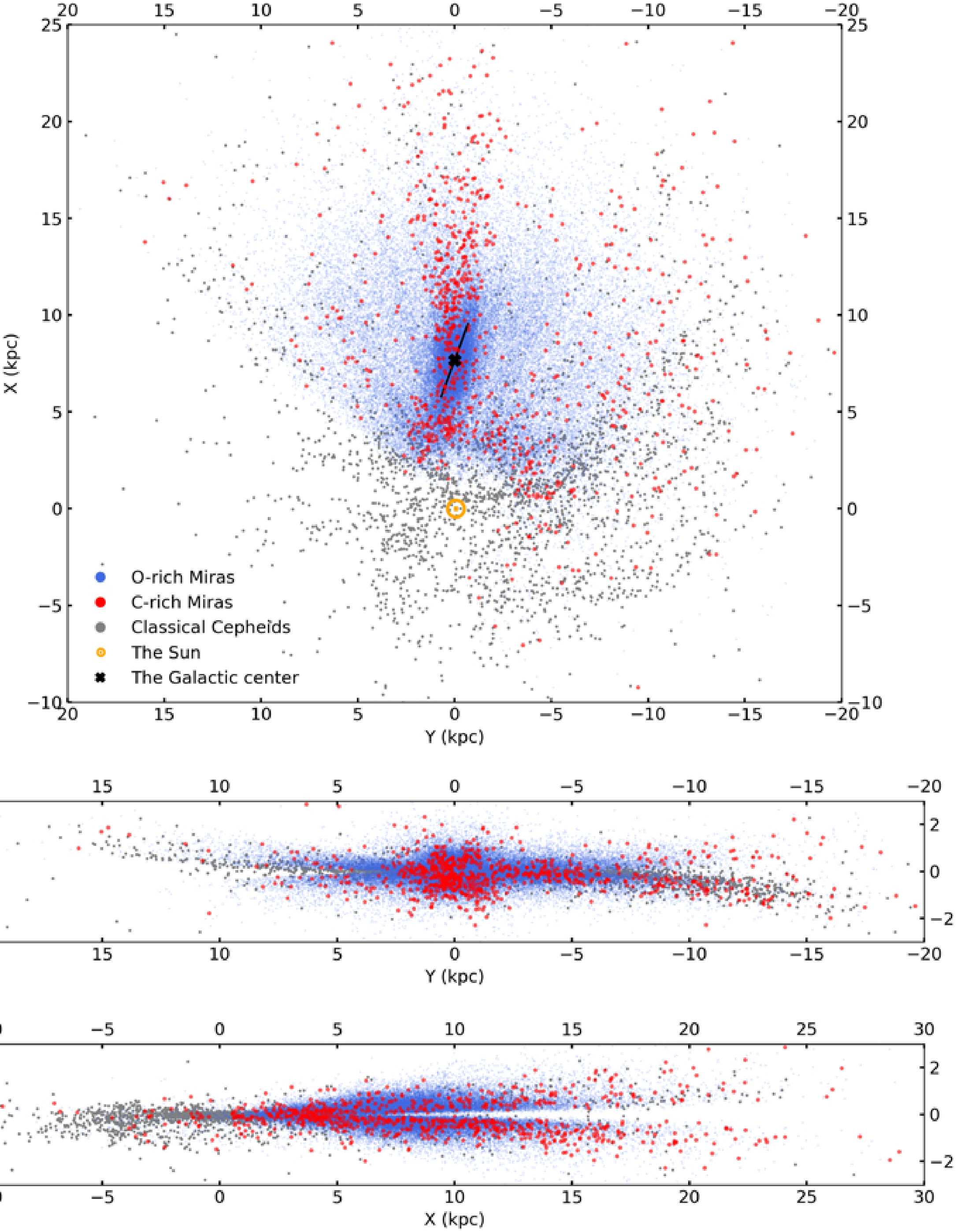
Additional References:

- Huang (2024): This is a review of Mira variables that covers a lot of the important topics, like the different types and $P - L$ relations.
- Lebzelter+ (2023): This article contains 1.7 million LPV candidates from the *Gaia* survey, covering the entire sky. An all-sky map is included below. The Fornax dwarf galaxy is barely visible below the "-60" curve in the lower right.
- Iwanek+ (2023): This article used nearly 66,000 Mira variables from the OGLE survey to look at the structure of the Milky Way's bar and part of the disk. Here is one of their maps:
- Konchady+ (2024): This study applied machine-learning techniques to identifying LPVs in a nearby galaxy called M33, and may be a good reference for seeing how


DR3 LPV candidates from the SOS table
180
60
60
0
0
180
-60
-60
180
10^4
1000
100
10
1
number of sources / deg^2

machine learning can be applied in this case. In addition, they used *gri* filters similar to Rubin, and their catalog is publicly available on GitHub.

- Bersier & Wood (2002): This pair have a catalog of variable stars in a subset of the Rubin field, but identify more than 500 RR Lyrae stars, 23 Cepheid-type stars, and 85 long-period variable candidates. Their tables are publicly available.
- Braga+ (2022): This study discusses a subset of RR Lyrae stars in Fornax and gives a catalog of those, but mentions more than 2000 detected RR Lyraes total, and more than 20 Cepheids. It might be possible to identify the variables from data here.



Some Possibly Useful Columns for Variable Identification and Classification: In the notes below “x” stands for one of the Rubin filters ($ugrizy$).

Detection:

- xPSFluxChi2; xPSFluxStetsonJ: These statistics can be good absolute indicators of variability even before we are able to figure out what the light curve looks like.

- xPSFluxMAD: This can also be an indicator of variability, but it has the drawback that the variations are not being compared to the measurement errors. So it might be misleading for faint (more noisy) variable stars.

- xPSFluxMean: We are ultimately going to need a good measure of the average flux in getting the distance, and also in getting colors (from ratios of average fluxes in different filter bands).
- nDiaSources; xPSFluxNdata: We'll generally want to make sure we have enough measurements to start to tell if there is variability. (nDiaSources counts all measurements identified for an object.)

Classification:

- xPSFluxLinearSlope: This might help identify LPVs early on... most will be varying in a linear way before they finish one cycle.
- xPSFluxPercentile05; xPSFluxPercentile95: These can provide a more robust indication of the amplitude of variation by eliminating outliers (in the lowest or highest 5% of the measurements). The amplitude in magnitudes would be related to $-2.5log_{10}$ of the ratio of these flux percentiles.
- xPSFMaxSlope: might help in separating stars with different light curve shapes (although it probably won't be helpful for the shortest period variable stars).
- xPSFluxSkew: This can be a good indication that the light curve is asymmetric (more time spent at lower fluxes than high fluxes, for example).