GALACTIC DUST MAP VIA RR LYRAE PERIOD-COLOR RELATION

Abhishek Kattuparambil[1]

[1]*Department of Astronomy, University of California, Berkeley, CA, USA 94720*

Keywords: RR Lyrae — Light Curve — Fourier Analysis — Period-Luminosity Relation — Period-Color Relation — Extinction Map — Metropolis-Hastings Monte Carlo Markov Chain — Galactic Dust Map

1. INTRODUCTION

Gaia is a space observatory launched by the European Space Agency in 2013, which has recorded astrometry for roughly two billion sources with remarkable precision. The Gaia mission set out to document 1% of the Milky Way, focusing on bright stars falling within an extended visual photometric band. In this paper, we will focus on `gaiadr3.vari_rrlyrae`, a catalog with information on RR Lyrae stars. These measurements allow for the estimation of apparent/absolute magnitudes, distance, color, periods, extinction, and therefore, the creation of a galactic dust map for RR Lyrae variable stars. After the Gaia Data Release 3 (GAIA DR3) earlier this year, a plethora of new measurements have been made public. In this paper, we will verify and analyze the periodic pulsations of RR Lyrae stars and derive a linear estimator for the period-magnitude relation by using this data in conjunction with distance data from `external.gaiaedr3_distance`. Furthermore, we will also derive a linear estimator for the period-color relation, which we will use alongside photometric data to determine extinction values and create a galactic dust map. Finally, we will compare our extinction map to the SFD Map and analyze any discrepancies.

2. GAIA ARCHIVE

During its conception, GAIA was an acronym, standing for Global Astrometric Interferometer for Astrophysics. Gaia's interferometry technique has changed since then, and it has lost its lengthy moniker, yet it continues to record astrometric and photometric data with extreme precision. The European Space Agency provides well-documented public access to all of Gaia's data in the Gaia Archive[1], which can be queried directly from the website, or through API requests abstracted by the `astroquery.gaia` package. All queries must be written in Astronomical Data Query Language (ADQL), which is syntactically very similar to SQL, but contains added functionality to simplify complex queries for stellar data.

[1] https://gea.esac.esa.int/archive/

3. RR LYRAE

RR Lyrae variable stars are post-Main Sequence stars that lie on the instability strip, a region on the HRD where stars pulsate and exhibit a periodic brightness. The pulsation is caused by the κ-mechanism, a phenomenon in which a star undergoes a continual cycle of ionizing hydrogen and increasing temperature until gas becomes opaque and the star begins cooling again. This pulsation allows us to directly observe a period, and we will create a linear estimator for both the period-luminosity and period-color relations, where period is measured in log days.

4. LIGHT CURVES

By plotting the G-band magnitude (m_G) against time, we can generate light curves for each RR Lyrae source, and derive a functional form for $m_G(t)$. However, the GAIA data is sampled at irregular discrete intervals, such that the magnitude measurements are sparse over time. Fortunately, RR Lyrae are known for having observable periods. To determine the period of our measurements, we will run the data through a Lomb-Scargle Periodogram, which detects periodic signals in unevenly spaced measurements. See Figure 1. We will fold the time interval into a single period (t % P) to produce the light curve in Figure 2. Additionally by folding this data, we can reduce the Fourier domain to the interval [0,1] to simplify future computations.

4.1. *Fourier Analysis*

Since the light curves exhibit periodic fluctuations, we will fit a Fourier series to the measured data. A Fourier series is a linear combination of k sine and cosine waves, oscillating to fit a function $m_G(t)$. Since we are calculating $m_G(t)$ in phase-space, its domain is the interval

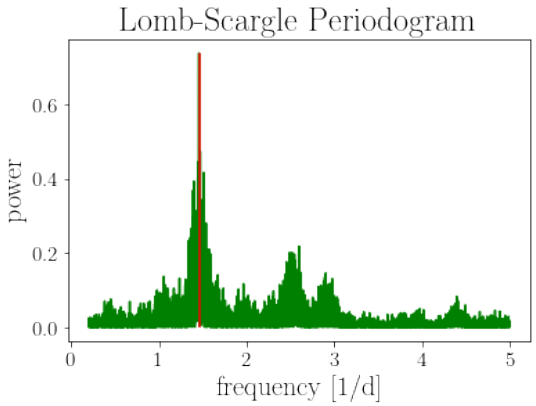

Figure 1. Lomb-Scargle periodogram for flux measurements of source 5817567360327589632 over time. There is a large spike at a frequency of 1.46 $\frac{1}{d}$, which has been marked in red. This corresponds to a period (P) of 0.686 days.

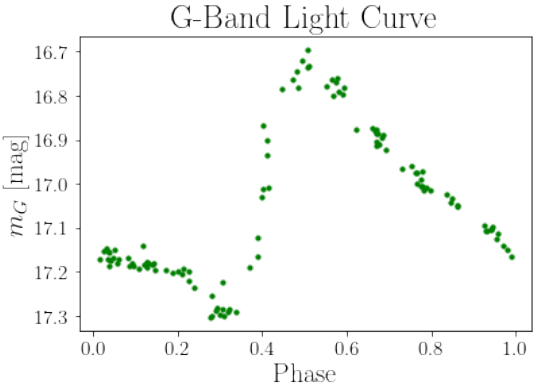

Figure 2. G-band light curve for source 5817567360327589632 folded into one period. We can see a clear correlation between m_G and phase, demonstrating the star's periodic brightness and the existence of a functional form for $m_G(t)$.

$[0, 1]$, and the corresponding Fourier parameter $L = \frac{1}{2}$.

$$m_G(x) = \frac{a_0}{2} + \sum_{k=1}^{K} a_k sin(2\pi Pkt) + b_k cos(2\pi Pkt) \quad (1)$$

By deriving the period ω from the Lomb-Scargle periodogram, we can solve for the Fourier coefficients by solving the following linear equation.

$$\vec{y} = X\vec{\beta} \quad (2)$$

where $\vec{y}$ is the vector of all the recorded fluxes and $\vec{\beta}$ is the vector of all the Fourier coefficients in the order $a_0, b_0, a_1, b_1, ...$ Therefore, X is the matrix of the trignometric functions to me multiplied through the coeffi-

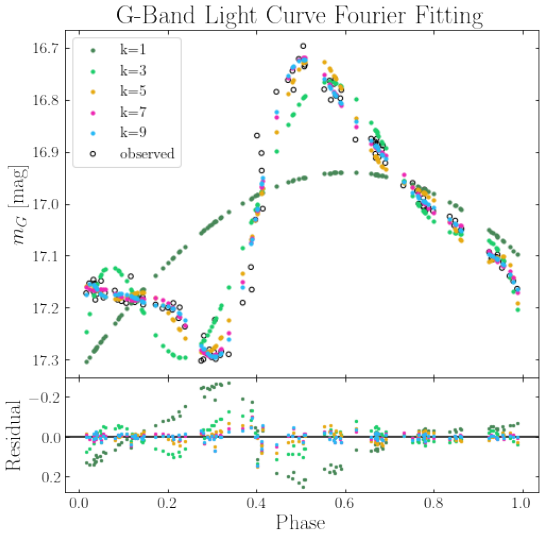

Figure 3. Fourier models $m_G(t)$ for $k = 1, 3, 5, 7, 9$. Fourier series for small values of k have fewer terms, and are not expressive enough to fit complex curves. The structure of this particular light curve is apparent only in the Fourier models for $k > 7$.

cients. Our final expression resembles

$$\begin{bmatrix} m_{G0} \\ m_{G1} \\ \vdots \\ m_{Gk} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & sin(\gamma t_0) & cos(\gamma t_0) & \dots & cos(k\gamma t_0) \\ \frac{1}{2} & sin(\gamma t_1) & cos(\gamma t_1) & \dots & cos(k\gamma t_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2} & sin(\gamma t_k) & cos(\gamma t_k) & \dots & cos(k\gamma t_k) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ b_1 \\ \vdots \\ a_k \\ b_k \end{bmatrix} \quad (3)$$

where $\gamma = 2P\pi$. We are unable to perfectly fit the data, so we cannot directly solve this linear system. Instead, we will compute the least-squares solution via `numpy.linalg.lstsq`. This returns our coefficient vector $\vec{\beta}$, whose elements we plug into Equation (1) to recover the Fourier series.

If we underestimate the complexity of our light curve ($K = 1$), we will *underfit* the data, such that our model is not representative of the data's structure. If we overestimate (ex: $K = 50$), we will force the series to fit the exact data points, which will *overfit* the intrinsic variance and possible outliers. To avoid either modeling error, we employ cross-validation on the data set, splitting the original table into training and validation sets. By training the model on the training set, and testing it on the validation set, we can determine the effectiveness of the model on new data using the χ^2/N metric, which will decrease for the training data and eventually increase for the validation data as we increase k. We choose a value for k which ensures that both χ^2/N val-

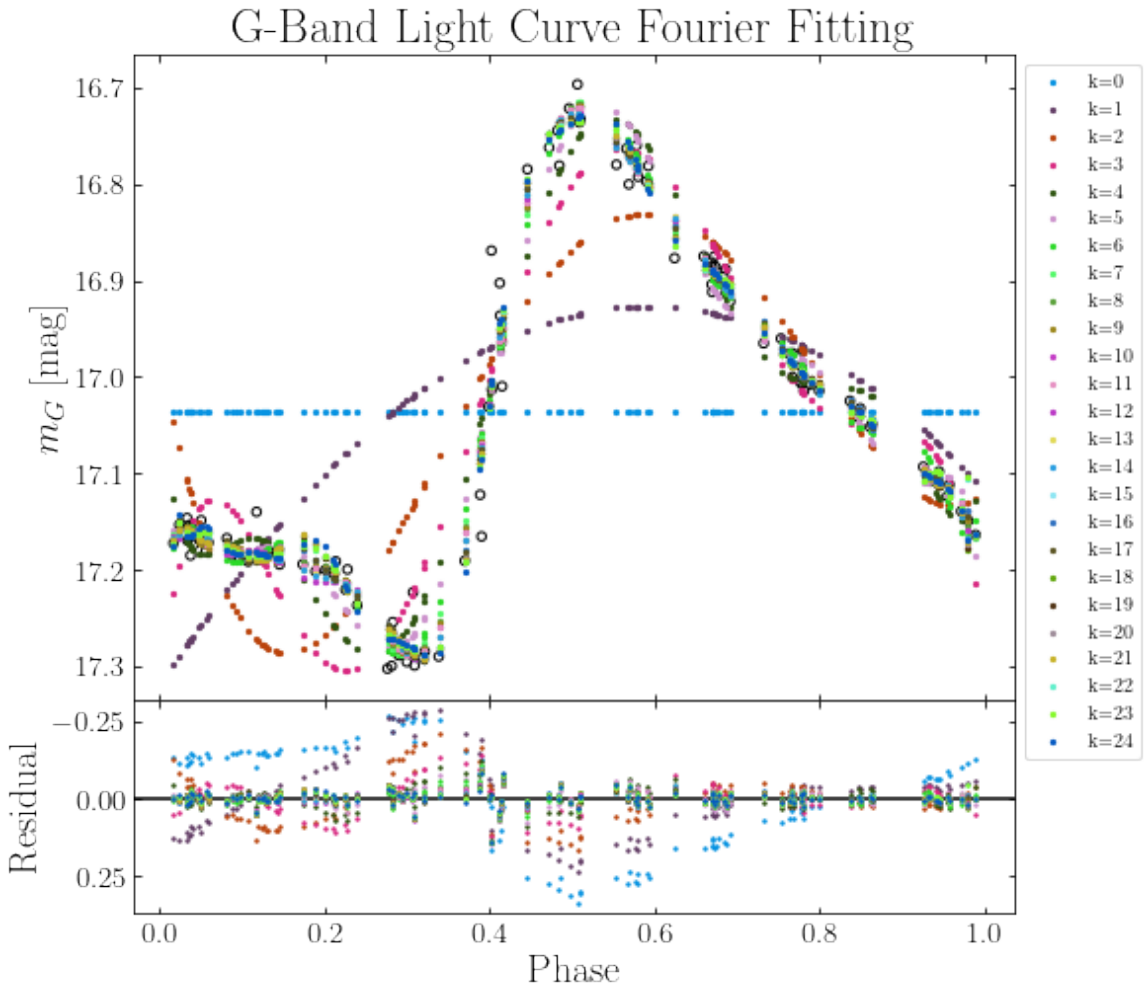

Figure 4. Fourier models $m_G(t)$ for $k \in [1 \,..\, 25]$ and the corresponding residuals. After applying cross-validation to randomly selected training and validation datasets, we determined that any k within the range $[8 \,..\, 15]$ performed similarly on the validation set. However since the splits were randomly decided, we averaged the results over 10 runs and determined $k = 8$ to be the optimal value.

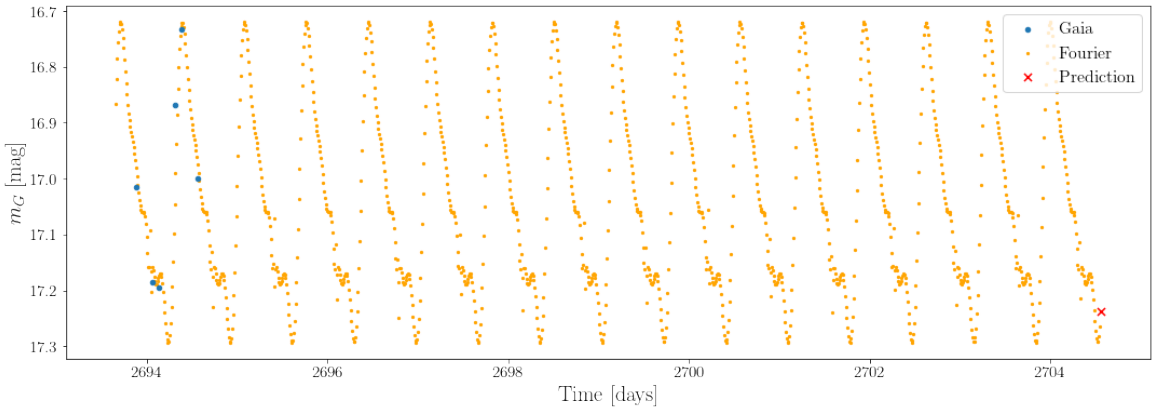

Figure 5. Light curve extrapolation 10 days from last GAIA measurement for DR3 source 5817567360327589632. Blue points are the final few measurements from the Gaia DR3 photometry data. Orange points are sampled from a Fourier series built with $k = 8$, and our final prediction for time $t = 2704.56$ is $m_G = 17.26$

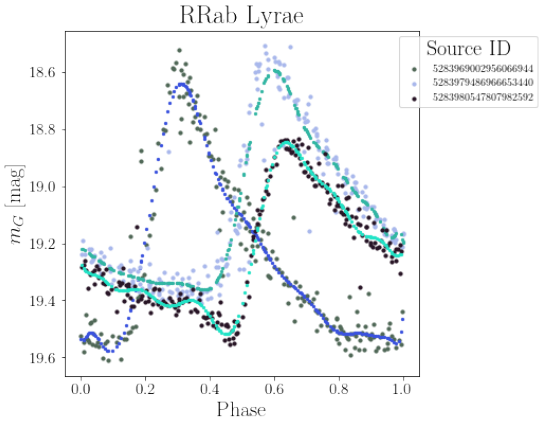

Figure 6. Fourier models for three RRab stars. Source IDs are listed in the legend, and the Fourier models are overplotted on the same x-space as the measurements. The models are apparent and follow a shape close to the mean of the measurements. These light curves are steep and complex, typical of RRab stars.

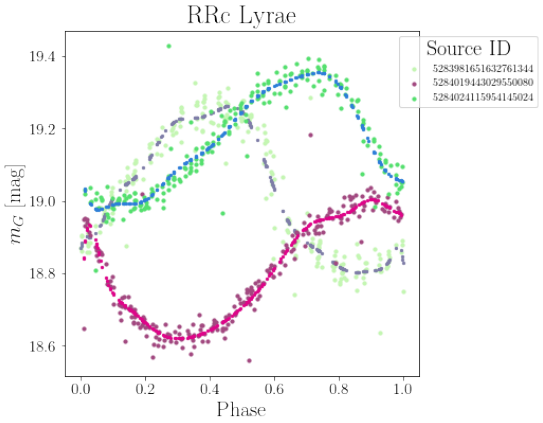

Figure 7. Fourier models for three RRc stars. Source IDs are listed in the legend, and the Fourier models are once again overplotted. Notice that the light curves are shorter and rounder, indicative of RRc stars.

ues are as small as possible.

Since we have only been querying stars with periods recorded in the column `pf`, we have been limited to the RR Lyrae variability class "RRab". RRab starts are the most common (∼90%) class of RR Lyrae, and are identifiable by their steep light curves. See Figure 6. On the other hand, RRc stars have short, rounded periods which are much less complex than examples from other variability classes. See Figure 7.

5. PERIOD-LUMINOSITY RELATION

RR Lyrae variable stars have a known period-luminosity relation, allowing observers to derive the absolute magnitude of a source once recording the period.

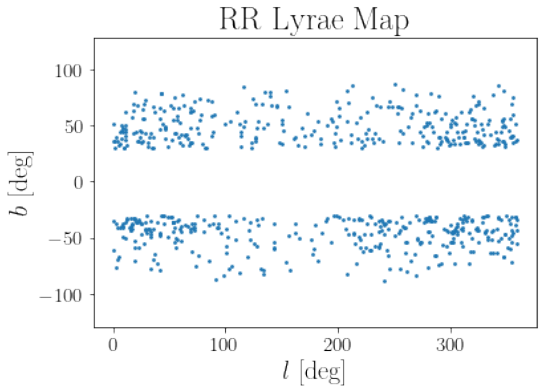

Figure 8. Map of RR Lyrae variable stars in Galactic coordinates, where l and b are the galactic longitude and latitude, respectively. Since the population density of the stars increases as b approaches 0 from either side, we can confirm that the stars are concentrated in the Milky Way disk.

The distance to a RR Lyrae star can then be derived using Equation 4.

$$m - M = 5\log(d) – 5 \tag{4}$$

where m is the apparent magnitude, M is the absolute magnitude, and d is the distance to the source in parsecs. This is a typical procedure where the RR Lyrae star is used as a standard candle, a term for a source of known brightness. In this section, we will derive the period-luminosity relation for RR Lyrae variable stars, once again utilizing Gaia Archive data. To query RR Lyrae stars along with their photometric data, we will join the `gaiadr3.vari_rrlyrae` and `gaiadr3.gaia_source` tables on the DR3 `source_id`. However, there is a large amount of dust concentrated in the Milky Way, which reduces and reddens incoming light, leading to inflated apparent magnitudes. Therefore, we will only query for stars above or below the Milky Way disk, with Galactic latitude $|b| > 30$. See Figure 8. Additionally, to ensure the measurements are precise, we filter out stars with parallax errors larger than 20% and stars further than 4 kpc away.

5.1. *Bailer-Jones Catalog*

To compare our distances to those calculated with astrometric parameters, we will query `external.gaiaedr3_distance`, also known as the Bailer-Jones catalog (Bailer-Jones 2018). These distance values are calculated with Metropolis-Hastings Markov Chain Monte Carlo Sampling with a gaussian prior on parallax uncertainties. This algorithm will be explained in Section 5.2. Nonetheless, Bailer-Jones distances are more precise than naive distance calculations using parallax, as depicted in Figure 10. The

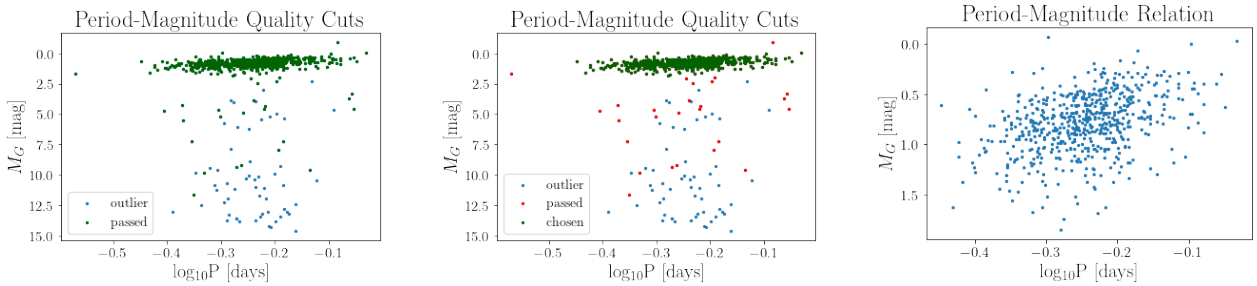

Figure 9. Period-absolute magnitude relation with entire query, cuts from Lindegren (2018), and additional visual cuts, respectively. The plot furthest on the left shows a definite relation, with many outliers lying off the median. The plot in the middle keeps most of the points along the main relation, yet retains a significant amount of outliers. Our final visual cuts narrows down our relation such that our axes change, yet all these points lie across a narrow relation.

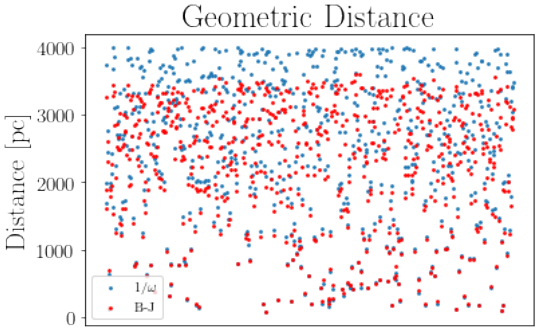

Figure 10. Bailer-Jones distances `r_med_geo` plotted in red and naive distance calculations $\frac{1}{\varpi}$ plotted in blue. As distances increase, the difference between the two grows larger, with the naive method consistently overestimating the Bailer-Jones distances.

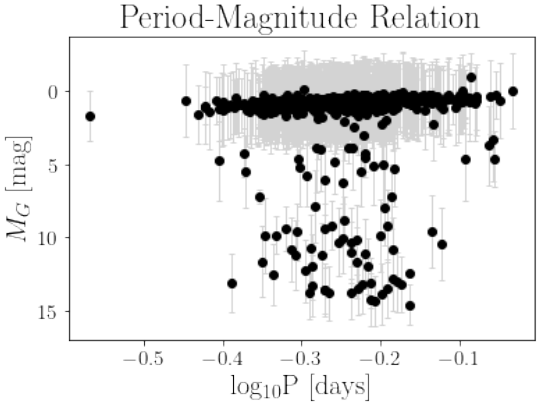

Figure 11. Period-absolute magnitude relation for our entire query, with errorbars calculated from `r_lo_geo` and `r_hi_geo` quantiles from the `external.gaiaedr3_distance` dataset. Notice that our errorbars are significant, and that our sources are not constrained to the median relation.

naive method is simple; when distance d is measured in parsecs, and parallax ϖ is measured in arcseconds, they exhibit an inverse relationship

$$d = 1/\varpi \tag{5}$$

Now that we have reliable distance data, we can use Equation 4 to determine the absolute magnitudes of our RR Lyrae variables. Additionally, we can use error measurements from the data to derive errorbars for each calculation. See Figure 11. Many of these stars lie off the main relation, due to incorrectly measured parallaxes. Directed by the equations C.1 and C.2 listed in Lindegren (2018), we make quality cuts upon the data to remove invalid measurements such that our data forms the tightest relation possible. Furthermore, we will make visual cuts on $\log(P)$ and M_G, using the standard deviation of the parameters to guide our bounds.

5.2. *Markov Chain Monte Carlo (MCMC) Sampling*

MCMC Sampling is a stochastic sampling method which constructs a Markov Chain to have an equilibrium distribution modeling an existing probability distribution, such that we can sample the distribution by recording states from the chain. This algorithm is ideal for modeling data with intrinsic scatter, which will define the width of the final equilibrium distribution. After passing in parameters to model the mean and standard deviation of our final distribution, the MCMC Sampling algorithm will attempt to optimize each value such that the resulting probability distribution is as tight as possible. In the case of the Bailer-Jones catalog, the distances were calculated using a Metropolis-Hastings (M-H) MCMC, which takes steps by sampling from a proposal distribution. If the new point is more preferable (according to the log-likelihood), we will always accept it. However, similar to the simulated annealing optimization method, even if the new point is analytically worse, we may accept it with a probability r, such that we do not get stuck in local minima. In this section, we will be using a No U-Turn Sampler (NUTS) MCMC to estimate the period-luminosity relation. The NUTS MCMC constructs its path to avoid backtracking by avoiding random walks and taking informed steps based on a variety of gradient and differential information.

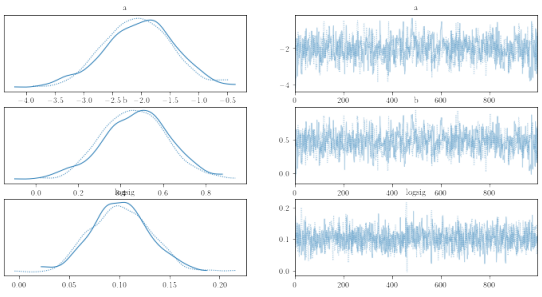

Figure 12. Trace plots for the a, b, and $\sigma_{scatter}$ parameters we will be using to create our linear estimator $M_G = a\log[P/day]+b$. The distributions are densities of the parameters, such that the peaks represent successful points in parameter space which the algorithm hovered around during its search for the optima.

We will describe a linear estimator $M_G = a\log[P/day]+b$ to fit our data, and will derive the according parameters using a NUTS MCMC. For our initial values, we will create three Gaussian variables $a, b, \sigma_{scatter}$ with means $\mu_a = -2, \mu_b = -1$, and $\mu_\sigma = -1$ respectively. Each variable was given a standard deviation $\sigma = 1$, since the initial guesses were fairly accurate, cross-references with results from a linear least squares estimation. See Figures 12, 13, and 14.

5.3. *Wide-field Infrared Survey Explorer*

Wide-field Infrared Survey Explorer (WISE) is a space telescope that has performed an all-sky astronomical survey in the near-infrared band, and is being repurposed to identify asteroids on Earth-bound trajectories. Regardless, we need the photometry data obtained in the infrared survey, so we again turn to the Gaia Archive, this time using the `gaiadr3.allwise_neighbourhood` and `gaiadr1.allwise_original_valid` tables to cross-reference RR Lyrae variable stars with the corresponding WISE data. Since these stars are very bright and found in clusters, we will use the `w2mpro` magnitude, which is calculated using profile-fitting photometry. Once again, we will plug this apparent magnitude and our Bailer-Jones distances into Equation 4 to calculate absolute magnitudes. As evidenced by Figures 14 and 15, the W2 period-magnitude relation is steeper. In fact, we can use the results of our MCMC to claim with high confidence that the near-infrared relation is roughly 1.5x steeper than its counterpart in the visual band Klein (2014) Beaton (2018).

6. PERIOD-COLOR RELATION

Period color relations are imperative to photometry, as they allow the derivation of period from color measurements. The period allows for the calculation of a

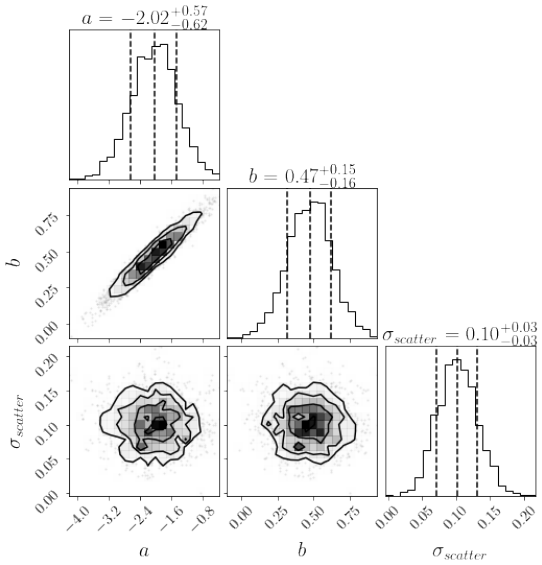

Figure 13. Corner plots for the a, b, and $\sigma_{scatter}$ parameters we will be using to create our linear estimator. These plots depict the posterior probability distributions of each parameter, and their covariances. As we can see, each parameter's individual posterior distribution was essentially normal. There is an obvious correspondence between a and b since changing either would move the entire line, and force the other to recalibrate to fit the data. Additionally, the one-dimensional posterior probabilities allow us to make confidence interval estimates for the values of our parameters. The mean and standard deviation of the resulting parameters are listed above.

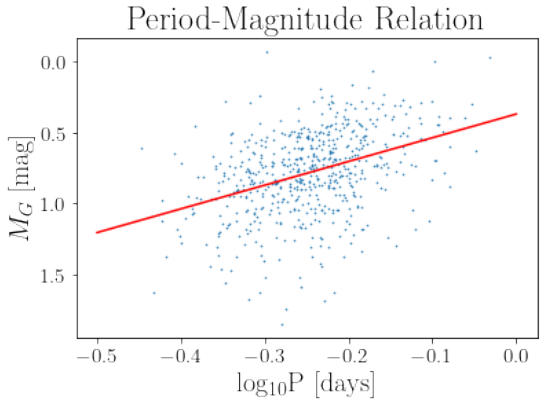

Figure 14. Period-Absolute magnitude relation for the cleaned dataset, and the overplotted line determined by the resulting parameters of our NUTS MCMC algorithm. The final parameters were determined to be slope $a = -1.67$,y-intercept $b = 0.37$,and instrinsic scatter $\sigma_{scatter} = -1.30$. Each parameter has a defined errorbar, yet we only overplot the line described by the mean parameters for simplicity.

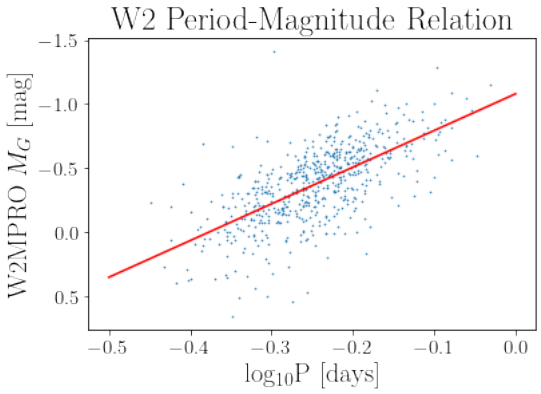

Figure 15. Period-Absolute magnitude relation for our WISE query, and the overplotted line determined by our NUTS MCMC. Initial values were $a = -3, b = -1$, and $\sigma_{scatter} = -1$. Once again, each paramater had a standard deviation $\sigma = 1$. Once the Markov Chain converged, the mean parameters were determined to be $a = -2.86, b = -1.08$, and $\sigma_{scatter} = -1.48$.

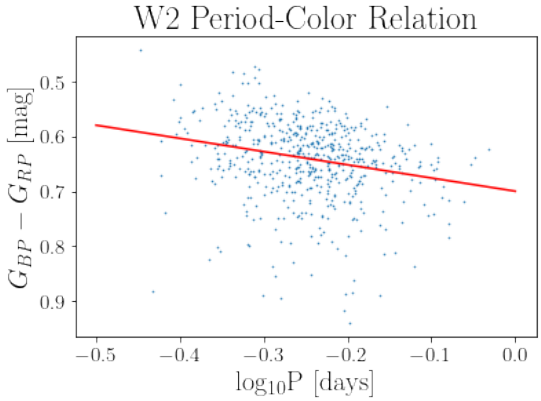

Figure 16. Period-Color relation for our WISE query, and the overplotted line determined by our NUTS MCMC. Initial values were $a = 1, b = 0.5$, and $\sigma_{scatter} = -2$. Each paramater distribution began with a standard deviation $\sigma = 1$. Once the Markov Chain converged, the mean parameters were determined to be $a = 0.24, b = 0.70$, and $\sigma_{scatter} = -2.66$.

multitude of other astrometric and photometric variables. We will fit a linear estimator to the relation between log(P) and `bp_rp`. See Figure 16. Now that we have derived a period-color relation for RR Lyrae variable stars, we can extrapolate the results to the entire `gaiadr3.vari_rrlyrae` catalog to create a galactic dust map. We begin by calculating the color excess for the W2 sources using

$$E(G_{BP}-G_{RP}) = (G_{BP}-G_{RP})_{\text{observed}} - (G_{BP}-G_{RP})_{\text{intrinsic}} \quad (6)$$

where $(G_{BP} - G_{RP})_{obs}$ is `bp_rp` and $(G_{BP} - G_{RP})_{intr}$ is the prediction from our MCMC linear estimator. We

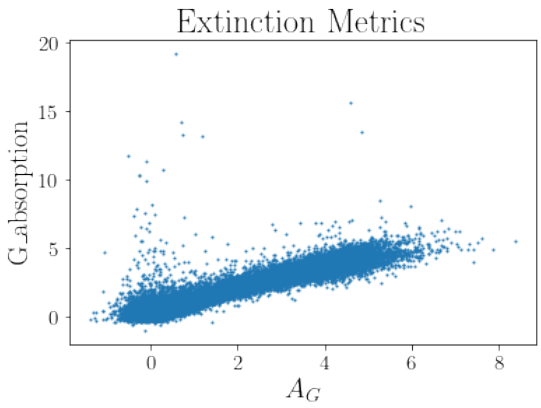

Figure 17. Comparison between calculated A_G extinction values and `gaiadr3.vari_rrlyrae.G_absorption`. There is a linear relationship between these variables, which confirms that the values are directly correlated, and the dust maps should resemble each other, albeit over a different interval for absorption values.

then use these color excesses to derive the G-band extinction, which we assume is of the form

$$A_G = 2E(G_{BP} - G_{RP}) \quad (7)$$

. We compare these extinction values (Figure 17) to the provided `G_absorption` column, which represents interstellar absorption in the G band.

7. GALACTIC DUST MAP

In this section, we will create a Galactic Dust Map to examine patterns between interstellar extinction in the galactic plane. We can plot our entire dataset in Galactic coordinates using an Mollweide projection, and color the points using the extinction values derived in the previous section. We will apply the same C.1 and C.2 quality cuts from Lindegren (2018) to clean the raw data. See Figure 18.

We expect the large-scale structure of the SFD to match our predictions, however the different small-scale structure is expected. In the SFD map, the authors elected to model the temperature and optical depth of actual dust particles to draw conclusions for their attenuation. Additionally, their measurements were completed in a (B-V) band, while our measurements are recorded in the (B-R) band. Additionally, the SFD map only accounts for the presence of dust, so it is an attenuation map, rather than the extinction map we derived in this section.

We have examined the periodic, pulsating nature of RR Lyrae variable stars, and have derived a period-color relation using sources in areas of the galactic plane with low extinction ($|b| > 30$). Since the relation was built

Table 5.2. MCMC Linear Estimator Parameters

Relation	a	b	$\sigma_{scatter}$
Period-Absolute M_G	-1.67	0.47	-1.30
W2 Period-Absolute M_G	-2.86	-1.08	-1.48
Period-Color	0.24	0.7	-2.66

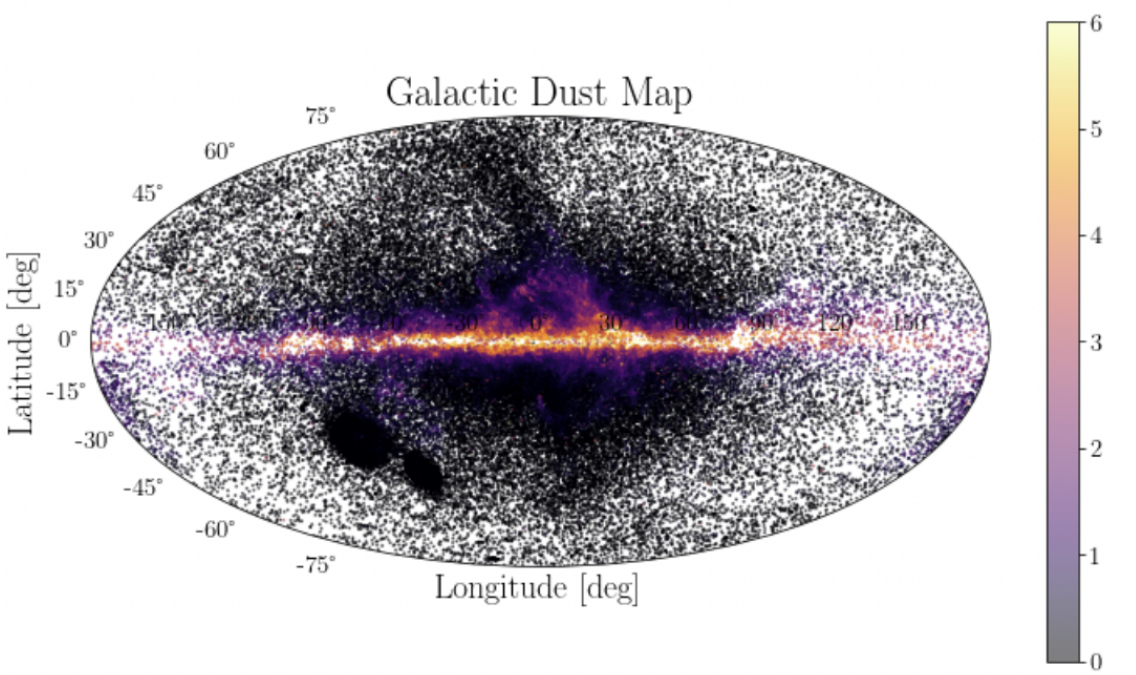

Figure 18. Galactic dust map colored with A_G values from the previous section. As expected, there is a concentration of dust in the Milky Way disk, and is begins to lessen as we go further out, with the exception of occasional clusters contributing large amounts of extinction to the galactic plane.

using relatively undampened measurements, we can extrapolate these to the entire RR Lyrae catalog to create a galactic dust map, where extinction values are determined by deviation from the derived period-color relation.

REFERENCES

Bailer-Jones, C. A. 2018, Estimating distances from parallaxes. V: Geometric and photogeometric distances to 1.47 billion stars in Gaia Early Data Release 3. https://arxiv.org/abs/2012.05220

Beaton, R. L. 2018, Old-Aged Stellar Population Distance Indicators. https://arxiv.org/abs/1808.09191

Klein, C. R. 2014, Towards precision distances and 3D dust maps using broadband Period–Magnitude relations of RR Lyrae stars. https://arxiv.org/abs/1404.4870

Lindegren, L. 2018, Gaia Data Release 2 – The astrometric solution. https://arxiv.org/abs/1804.09366

Schlegel, D. J. Finkbeiner, D. P. D. M. 1998, Maps of Dust IR Emission for Use in Estimation of Reddening and CMBR Foregrounds. https://arxiv.org/abs/9710327

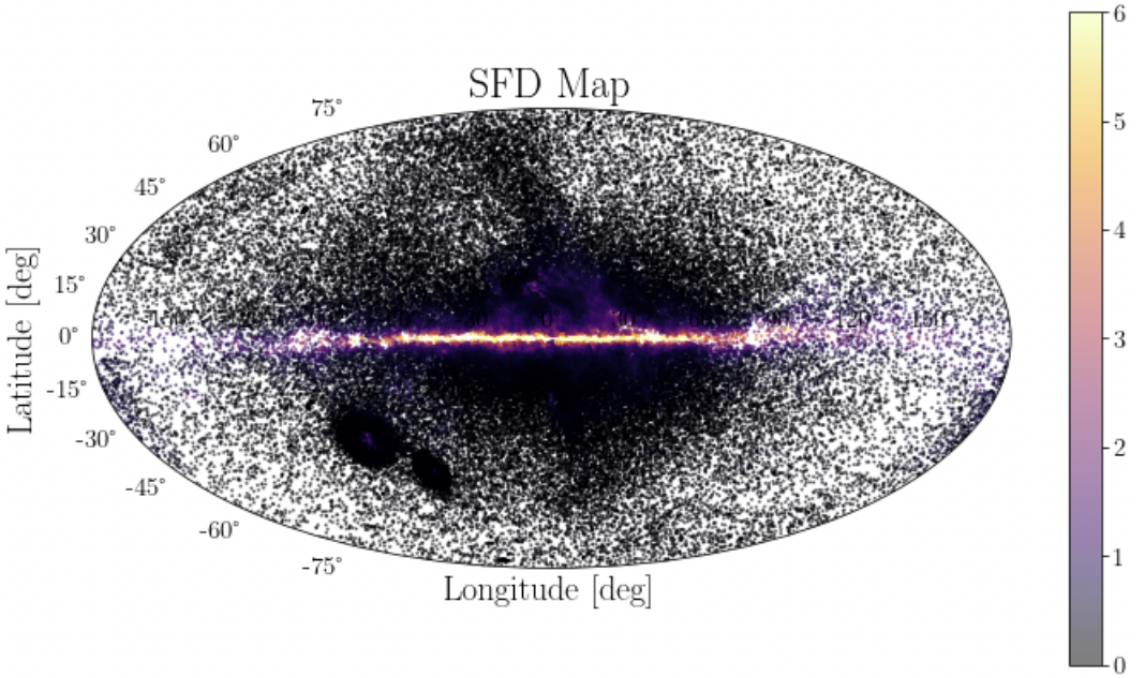

Figure 19. The most widely used Galactic dust map: the SFD map Schlegel (1998). The large-scale structure is similar to the generated dust map; the dust concentration in the Milky Way disk is apparent, and the general extinction caused by fringe galaxies matches the predictions. However, the small-scale details of the map do not match. The generated map underestimates dust in the fringe galaxies above the Milky Way disk, and also does overestimates dust for a few galaxies just below the disk.