0.0.1 Decision Trees

The use of a pre-classified star catalogue as a training set for a machine learning algorithm is not new. The General Catalogue of Variable Stars (GCVS, (Samus' et al., 2017)) exemplifies such a catalogue, having compiled variable objects since 1946. This approach has been applied to data from VVV (Molnar et al., 2022), ASAS-SN (Jayasinghe et al., 2018, 2019), Gaia DR2 (Rimoldini et al., 2019) and DR3 (Rimoldini et al., 2023), EROS-II light curves (Kim et al., 2014), and in the identification of microlensing events (Husseiniova et al., 2021).

It follows that we can use this approach to classify the variable stars in the PRIMVS catalogue. This method is not without caveats however, most notably the biases inherited from the training set.

The cross match of PRIMVS with the Gaia DR3 all-sky classification catalogue (Rimoldini et al., 2023) yielded 118,172 sources with an FAP < 0.3 and a 'best class score' > 0.7. This selection forms the training set which was then used with a gradient boosted decision tree classifier 'XGBoost' (Chen and Guestrin, 2016). Figure 1 shows the distribution of classes found in the cross-matched data. This distribution is not even across all classes and reflects the selection bias of both the Gaia DR3 all-sky classification catalogue and the VVV PRIMVS catalogue. This is a notable caveat, especially as the differences in the data (e.g. optical vs near-IR) likely leads to different selection biases.

Figure 2 shows the confusion matrix achieved from using the Gaia DR3 all-sky classification catalogue to form our training set for classifying PRIMVS. It can be seen that the majority of classes are correctly identified with a high completeness. All of the White Dwarf and RCB variables are misclassified as EBs and LPVs respectively. RCB variables are hydrogen-poor, carbon/helium-rich, high-luminosity stars. Their variability is characterised by high amplitude (1-9mag) aperiodic changes on the order of hundreds of days. This is superimposed by periodic pulsations up to several tenths of a magnitude on the time scale of tens of days (Clayton, 1996). The light curves of RCBs are therefore complex and likely span a large range in any feature space. Considering this with their scarcity, it is not surprising we misclassify all of them as LPVs, a much more common class with similar features. Given this, RCBs and White Dwarfs were removed from the data.

We can calculate how confident the model is with the highest probable class. Figure 3 shows the 'Entropy' versus the 'Confidence metric' for each class with a probability> 0.5 for its most likely class. where 'Entropy' is the entropy across the classes, (i.e. $S = \sum P_{class} \ln(P_{class})$). Therefore, a lower Entropy suggests that the model's predictions are more certain because the probability distribution across classes is less uniform – One class has a much higher probability compared to others, indicating a strong preference by the model for that class. The 'Confidence metric' is the difference between the most likely and next most likely class.

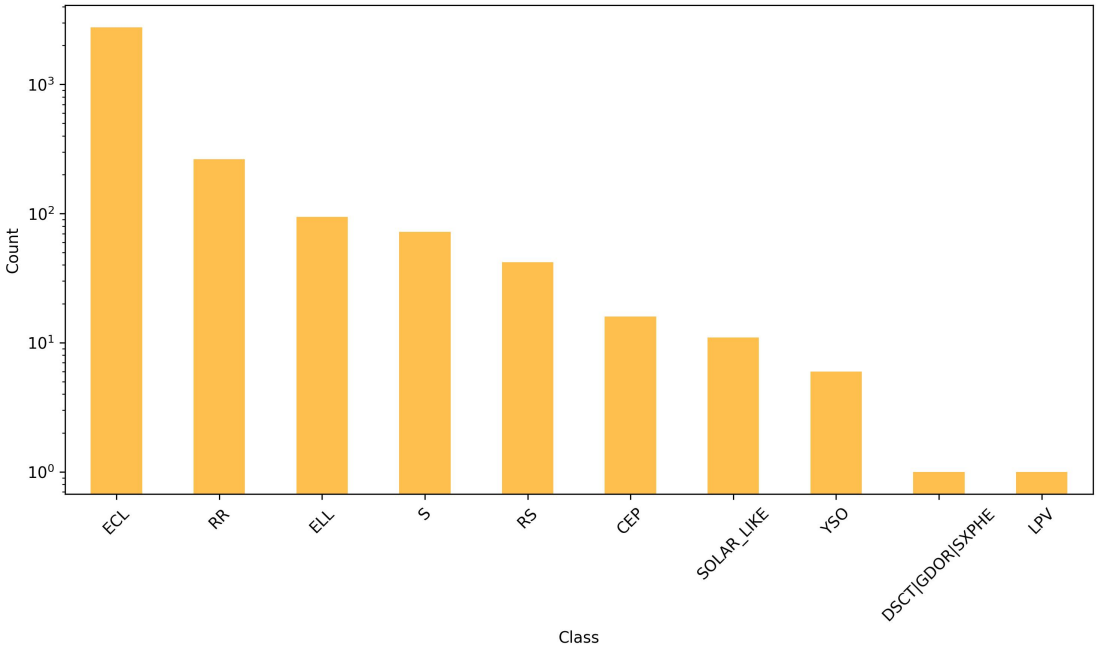

FIGURE 1: Training set of cross-matched VVV-Gaia data. Where; 'ECL' is eclipsing binaries, 'RR' is RR Lyraes, 'ELL' is Ellipsoidals, 'S' is short time-scale objects, 'RS' is RS Canum Venaticorum variables, 'CEP', is Cepheids, 'SOLAR_LIKE' is for Solar-like objects, 'YSO' is young stellar objects, 'DSCT||GDOR||SXPHE' is for delta-scuti like objects, and 'LPV' is for long period variables.

A Bailey diagram is an excellent tool for versifying how sensible our classifications are. Figure 4 shows the Bailey diagram constructed from the highest probability sources for each class. For both figure 4 and figure 5 we select only sources with a probability > 0.7, entropy < 0.2 and confidence metric > 0.9. The same colours and markers are also used to represent each class throughout figures 3,4, and 5 The Cepheid population is clearly visible and takes the expected form on the plot. The expected bimodal distribution of Cepheids can be seen as a loose 'V' shape at 10 days (Bono et al., 2000). We also see good agreement with Kains et al. (2019) in terms of our LPV, Delta Scuti and RR Lyrae placements.

Figure 5 shows the stellar classifications across the VVV survey region in relation to their positions within the Milky Way. This plot provides insights into the typical locations of different stellar populations. Cepheids, marked as red dots, are young, luminous stars commonly found in the thin disk throughout the galaxy Skowron et al. (2019). Figure 5 shows our sample of Cepheids throughout the disk mostly within $|l| < 1.5$, with an increased density at $|b| < 6$. Long-period Variables, such as Miras and semi-regulars are typically older, evolved stars and thus are more prevalent in the galactic bulge and halo, where older stellar populations dominate (Wood and Bessell, 1983). Figure 5 shows these objects homogeneously spaced throughout the disk with significantly higher densities towards the inner bulge. RR Lyrae stars, yellow plus signs, are old, metal-poor stars found mainly in the galactic bulge and halo, highlighting regions

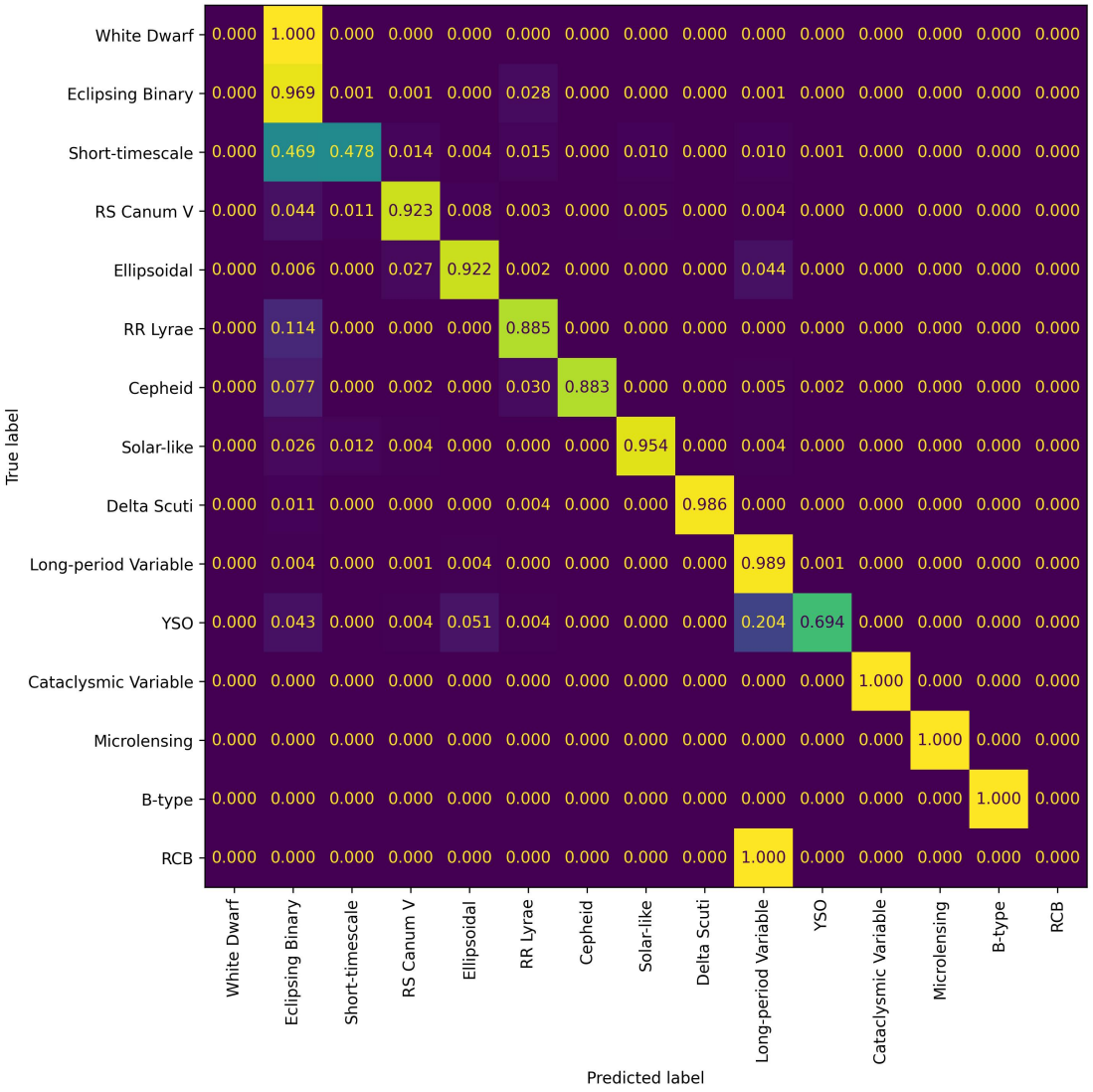

FIGURE 2: Confusion matrix of grouped classes from the Gaia Data Release 3 All-sky classification (Rimoldini et al., 2023). 'RCB' is R Coronae Borealis variables and YSO is Young Stellar Objects.

with ancient star populations ,(Cabrera Garcia et al., 2023; Ramos et al., 2018). This seems to be in agreement with figure 5.

Similar to the autoencoder, interactive plots and data visualisations that are not suitable for the pdf format can be found at: https://nialljmiller.com/projects/PRIMVS/PRIMVS.html.

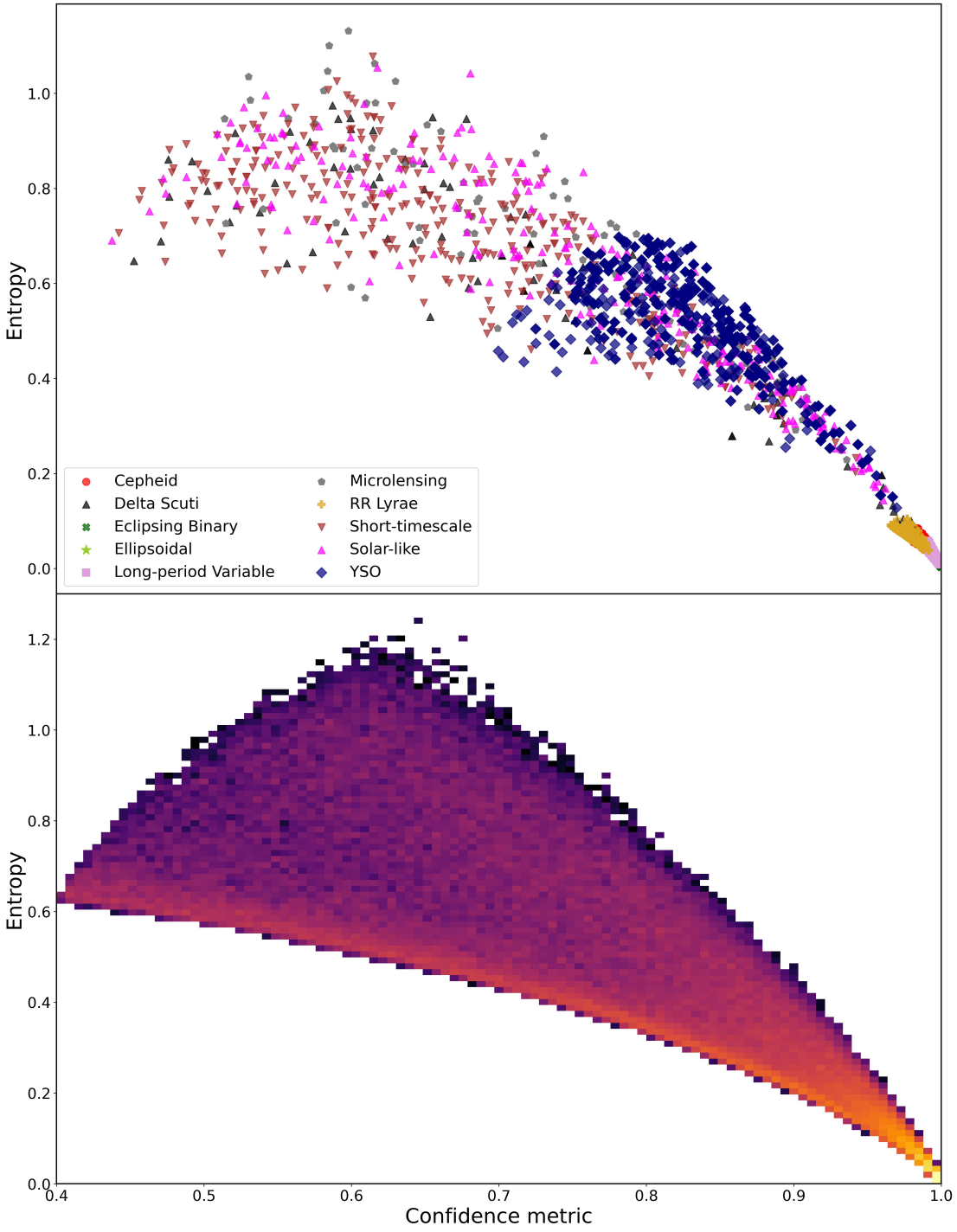

FIGURE 3: Top: A scatter plot of 'Entropy' versus the 'Confidence metric' for each each class with a probability$>$ 0.5. Bottom: A 2D histogram of 'Entropy' versus the 'Confidence metric' for the same data

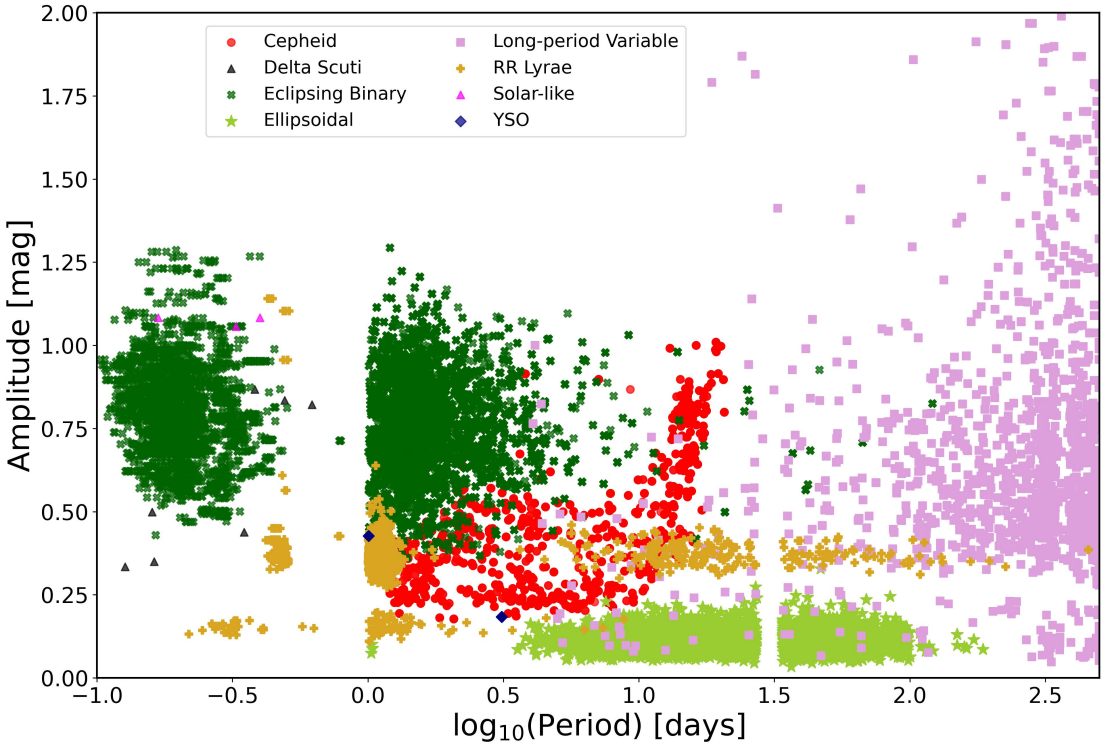

FIGURE 4: A plot of $\log_{10}(Period)$ verses Amplitude for the most confident predictions (top 10%) of each class from our decision tree which was trained using the Gaia DR3 all-sky classification catalogue.

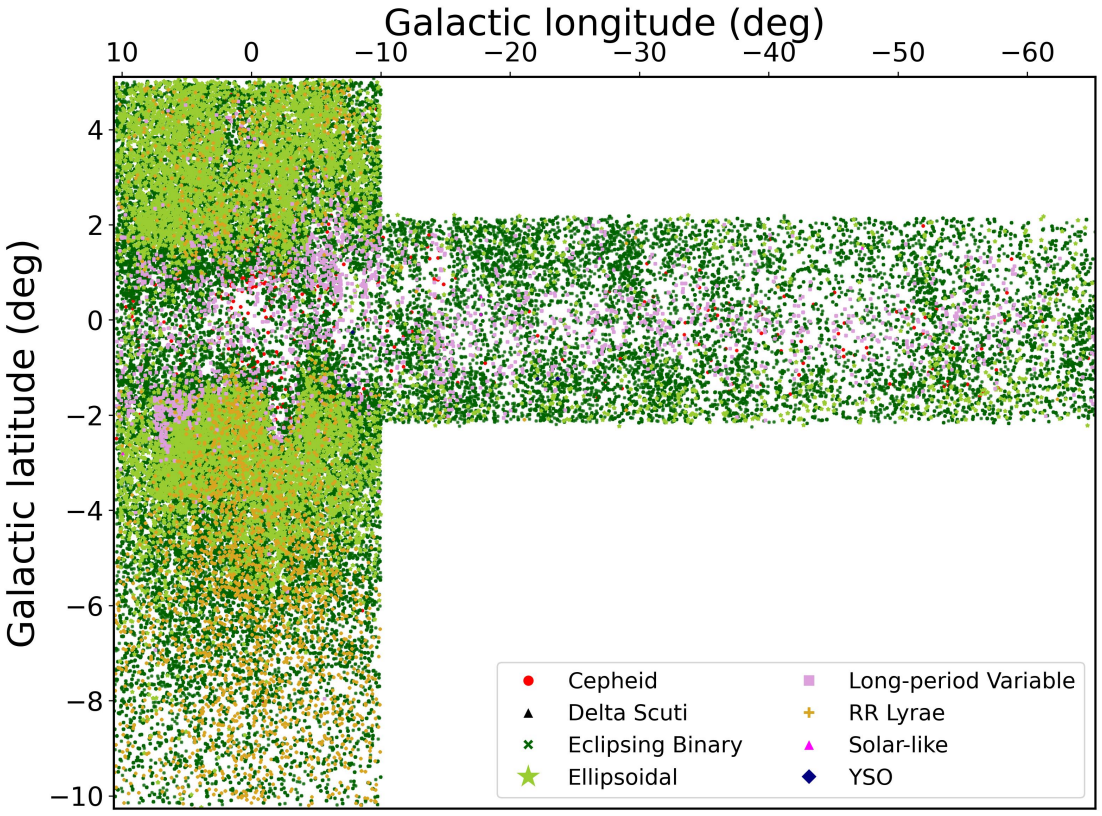

FIGURE 5: Spatial distribution of stellar classes across the VVV survey region in the context of the Milky Way. The decision tree based classification uses the Gaia DR3 all-sky classification catalogue as its training set.

Bibliography

Bono, G., Marconi, M., and Stellingwerf, R.F., 2000. Classical Cepheid pulsation models — VI. The Hertzsprung progression. *A&A*, 360:245.

Cabrera Garcia, J., Beers, T.C., Huang, Y., et al., 2023. Probing the Galactic halo with RR Lyrae stars – V. Chemistry, kinematics, and dynamically tagged groups. *Monthly Notices of the Royal Astronomical Society*, 527(3):8973.

Chen, T. and Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *arXiv e-prints*, arXiv:1603.02754.

Clayton, G.C., 1996. The R Coronae Borealis Stars. *PASP*, 108:225.

Husseiniova, A., McGill, P., Smith, L.C., et al., 2021. A microlensing search of 700 million VVV light curves. *MNRAS*, 506(2):2482.

Jayasinghe, T., Kochanek, C.S., Stanek, K.Z., et al., 2018. The ASAS-SN catalogue of variable stars I: The Serendipitous Survey. *MNRAS*, 477(3):3145.

Jayasinghe, T., Stanek, K.Z., Kochanek, C.S., et al., 2019. The ASAS-SN catalogue of variable stars - II. Uniform classification of 412 000 known variables. *MNRAS*, 486(2):1907.

Kains, N., Calamida, A., Rejkuba, M., et al., 2019. New variable stars towards the Galactic Bulge - I. The bright regime. *MNRAS*, 482(3):3058.

Kim, D.W., Protopapas, P., Bailer-Jones, C.A.L., et al., 2014. The EPOCH Project. I. Periodic variable stars in the EROS-2 LMC database. *A&A*, 566:A43.

Molnar, T.A., Sanders, J.L., Smith, L.C., et al., 2022. Variable star classification across the Galactic bulge and disc with the VISTA Variables in the Vía Láctea survey. *MNRAS*, 509(2):2566.

Ramos, R.C., Minniti, D., Gran, F., et al., 2018. The vvv survey rr lyrae population in the galactic center region*. *The Astrophysical Journal*, 863(1):79.

Rimoldini, L., Holl, B., Audard, M., et al., 2019. Gaia Data Release 2. All-sky classification of high-amplitude pulsating stars. *A&A*, 625:A97.

Rimoldini, L., Holl, B., Gavras, P., et al., 2023. Gaia Data Release 3. All-sky classification of 12.4 million variable sources into 25 classes. *A&A*, 674:A14.

Samus', N.N., Kazarovets, E.V., Durlevich, O.V., et al., 2017. General catalogue of variable stars: Version GCVS 5.1. *Astronomy Reports*, 61(1):80.

Skowron, D.M., Skowron, J., Mróz, P., et al., 2019. A three-dimensional map of the Milky Way using classical Cepheid variable stars. *Science*, 365(6452):478.

Wood, P.R. and Bessell, M.S., 1983. Long-period variables in the galactic bulge : evidence for a young super-metal-rich population. *ApJ*, 265:748.