Appendix

Detail and explanation of formulae used in calculations

(Adapted from R. Raikar (2019), ref. 5.)

1. Procedures used for Data Cleaning

Normalization

Initially the data from the SPLASH and PHAT datasets had high signal to noise (S/N) ratios thus making them unreliable due to inconsistency. This caused a large difference in flux between the stars. It was important to have comparable sets of data, so the spectra between wavelengths 7796-8257.5 Å (range of analysis) was trimmed and then all stars were divided by the medium flux value $f_m(\lambda)$ (Kamath et al., 2017). This made the median of all the normalized spectra equivalent to 1.

When comparing the spectra, it was important to know the uncertainty in the measurements of the star's spectra. Therefore noise (variance) was defined as the uncertainty in the measurements of flux in the spectra as a function of wavelength. The variance is given by $variance_\lambda = \sigma^2$ where σ represents the noise and it is squared to obtain absolute, non-negative values. The uncertainty of measurements were obtained and calculated using inverse variance ($ivar$) with the formula,

$ivar_\lambda = \frac{1}{\sigma^2}$. The noise had to be normalized as well to match the normalized spectra flux values. Normalized inverse variance $ivar_n$ is calculated using

$$ivar_n = \frac{1}{(\frac{\sigma}{f_m})^2} = ivar_i \cdot (f_m)^2$$

where $ivar_i$ is the initial unnormalized inverse variance ($ivar_\lambda$) that was mentioned earlier.

Although the formulas by Kamath et al. were correct, these formulas were using datasets that were very noisy as they did not take into account the fact that there were many velocity measurements that were not reliable in terms of spectral analysis. Previous work used all ZQUAL values from the datasets including the very noisy spectra that would affect the medium values. ZQUAL is the quality of velocity measurements with Z standing for the redshift of a star. To quickly summarize, the Doppler effect, which measures the increase or decrease of frequency, in terms of redshifts and blueshifts. As a star moves away, the wavelength gets longer and its light is shifted to the red end of the spectrum. But if a star is moving closer, the wavelength gets shorter and the light shifts to the blue end of the spectrum. So the redshift Z measures how fast the star was moving away with ZQUAL categorizing these measurements based on reliability of measurements and labeling them with numbers -2 to 4. Each number represented the quality of measurement. Using all the stars included stars that had bad ZQUAL measurements, so the analysis had to be restricted to the analysis of the good ZQUAL values (1,3, and 4). This greatly improved our future analysis by eliminating highly noisy datasets.

Coadditions and Smoothing

To further reduce noise and magnify similar features in the populations, coadditions were used. Each star's normalized spectra was weighted and added. This value was then weighted with the normalized $ivar_n$ values. The final flux values of the coadded spectra were calculated using the equation (Hamren et al. 2015):

$$f = \frac{\sum(f_\lambda \cdot ivar_\lambda)}{\sum ivar} = \frac{1}{\sum ivar} \cdot \sum(f_\lambda \cdot ivar_\lambda)$$

After the coadditions, the spectra needed to be trimmed to avoid any abnormally high or low flux values of outliers that may affect the comparisons of populations. A sigma clipping function was defined to clip or remove any outliers in the coadding spectra. The standard deviations of the spectral flux at every wavelength was calculated and any outliers that were 3.5 standard deviations away from the medium flux, f_m, were clipped or removed from the coadded normalized spectra and $ivar_n$ datasets. The value of σ=*3.5* was arbitrarily chosen, but greatly reduced outliers that affected analysis. The now newly trimmed datasets were once again coadded for a better average.

An extra step was a Gaussian 1D kernel smoothing function which was applied to the coadded spectra to improve the quality of the spectrum's graph by further reducing noise. This function was mainly applied to be used to smooth any graphs at a 2 pixels level, a number previously determined by Kamath et al., to portray the best spectra for analysis with minimized noise but still representing the most true data.

Dividing the Normal Stars

The first few steps had been to trim the data based on the spectral quality (measure of signal to noise ratio) of the stars. The goal was to limit the amount of noise in the spectra and prevent discrepancies to provide a clearer analysis of the data. Before the trimming, the weak CN and carbon stars had a high signal to noise ratio compared to the normal stars.

To account for this discrepancy, the normal star population was trimmed into two samples whose cumulative sum distributions were based on and matched the weak CN and carbon populations respectively.

2. Distance weighting for CMD classification

During classification of stars based on their positions on color-magnitude diagrams (CMDs), the following formula was used to weight the size of the points on the CMDs based on their distance from a point chosen arbitrarily as the head of the long trail of normal stars in the undiluted template scores plot:

$$(P_{max} - P_{min}) * e^{(-D/D_0)} + P_{min}$$

where P_{max} and P_{min} are the maximum and minimum dot sizes for the scatter plot in matplotlib, D is the distance from the comet head, and the free parameter D_0 is used to control the distribution of point sizes. D_0 value showing best results in our analysis was 0.075.