



Kent Academic Repository

Ryan, Benjamin W., Stokes-Geddes, Holly and Froebrich, Dirk (2025) *A survey for variable young stars with small telescopes – X. Comparing stochastic YSO light curves*. Monthly Notices of the Royal Astronomical Society, 543 (2). pp. 1133-1145. ISSN 1365-2966.

Downloaded from

<https://kar.kent.ac.uk/111435/> The University of Kent's Academic Repository KAR

The version of record is available from

<https://doi.org/10.1093/mnras/staf1495>

This document version

Publisher pdf

DOI for this version

Licence for this version

CC BY (Attribution)

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in **Title of Journal**, Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our [Take Down policy](https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies) (available from <https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies>).

A survey for variable young stars with small telescopes – X. Comparing stochastic YSO light curves

Benjamin W. Ryan,[★] Holly Stokes-Geddes[★] and Dirk Froebrich^{★}

Centre for Astrophysics and Planetary Science, School of Physics and Astronomy, University of Kent, Canterbury CT2 7NH, UK

Accepted 2025 August 26. Received 2025 August 22; in original form 2025 March 31

ABSTRACT

Light curves of young stars exhibit photometric variability over hours to decades and across a wide range of amplitudes. On time-scales beyond a few rotation periods, these light curves are typically stochastic. The variability arises from a combination of accretion rate changes, line-of-sight extinction variations, and evolving spotted stellar surfaces. We aim to develop a methodology to quantitatively compare the full variability statistics of these inhomogeneously sampled light curves with model calculations. To achieve this, we converted the light curves into variability fingerprints. They map the probability of variation by a given amount over a given time-scale. Applying principal component analysis to these fingerprints produces a stable distribution of the first two principal components. We show that this distribution is a continuum without clusters. Adding a model-generated fingerprint to an observational sample does not significantly alter the distribution of the sample, allowing a robust comparison between the model and observed light curves to assess statistical realism. We show that photometric uncertainties, timing, and observing cadence have a minimal impact on model placement within the observational distribution. The main source of variance among highly variable light curves of young stars is the time-scale of the onset of significant variability (above 0.3 mag), with time-scales of one to three months being the most critical. The secondary cause of variance is long-term (above 1.5 yr) dimming or rising trends.

Key words: stars: formation – stars: pre-main-sequence – stars: star spots – stars: variables: T Tauri – stars: variables: Herbig Ae/Be.

1 INTRODUCTION

Photometric variability has long been known as a defining feature of young stars (Joy 1945). This prevalence of variability in young stellar objects (YSOs) is caused by a variety of physical reasons such as rotation and surface spots, variations in mass accretion rates, and changes in the line of sight extinction due to orbiting disc material. Typically, light curves of YSOs are influenced by a combination of all of these effects simultaneously, creating stochastic variability in particular at time-scales above a few rotation periods, which can make it difficult to easily disentangle the effects of the different physical causes. Furthermore, the variations in brightness can reach up to several magnitudes and occur on time-scales from days to tens of years. Thus, numerous monitoring campaigns have been devoted to variability studies of YSOs in the past. See, for example, the *Protostars and Planets* reviews by Bouvier et al. (2007), Herbst et al. (2007), Audard et al. (2014), Bouvier et al. (2014), and Fischer et al. (2023) for an overview of some aspects of this vast field.

Hence, to study the underlying causes of YSO variability using photometry, we must cover all the relevant time-scales and observe a statistically significant number of objects. Furthermore, observations in multiple filters are paramount to potentially distinguish variations

caused by different physical mechanisms. There are multiple studies available that cover some of these aspects (high cadence, long-term observations, multifilter coverage, large sample; e.g. Grankin et al. 2007; Cody et al. 2014; Kochanek et al. 2017), but not all simultaneously. Thus, we started the Hunting Outbursting Young Stars (HOYS) project in 2014 (Froebrich et al. 2018) with the aim of monitoring a large number of YSOs in multiple optical filters with a daily cadence over several decades.

Some YSO light curves can (in part) show periodic modulation due to rotation – see reviews by Herbst et al. (2007) and Bouvier et al. (2014) – or orbiting warped inner disc material (e.g. Bouvier et al. 1999). While these regular variations are common at shorter time-scales (e.g. Cody, Hillenbrand & Rebull 2022), at longer time-scales the light curves are typically stochastic. Thus, the characterization of these is non-trivial. Multiple attempts have been made in the past to characterize light-curve properties with simple metrics such as variability indices (e.g. Stetson 1996; Sokolovsky et al. 2017), or quasi-periodicity (Q) and asymmetry (M) indices (e.g. Cody et al. 2014; Lakeland & Naylor 2022). While these are easy to calculate and compare (with each other or with models) for non-homogeneously sampled light curves, they essentially reduce the vast information of the entire light curve to a single number/dimension. Thus, they cannot capture the full extent of the variability information available for each YSO. Furthermore, it is not clear whether they represent the best way to distinguish YSO light curves from each other.

* E-mail: bwr2@kent.ac.uk (BWR); hs442@kent.ac.uk (HS-G); d.froebrich@kent.ac.uk (DF)

There have been attempts to remedy this situation (e.g. Scholz & Eislöffel 2004; Findeisen, Cody & Hillenbrand 2015; Rigon et al. 2017), upon which we have built our variability fingerprinting of light curves (Evitts et al. 2020; Froebrich et al. 2022). These fingerprints convert the light curves into a two-dimensional map showing the probability that a star varies by a given magnitude on a given time-scale. They can hence capture the entire range of variability of each light curve. To some extent, the fingerprints are similar to the analyses of structure functions presented, for example, by Sergison et al. (2020) and Lakeland & Naylor (2022). However, they distinguish between dimming and brightening more comprehensively and will be applied to data taken over much longer time-scales (10 yr) with nearly continuous coverage in our work.

These variability fingerprints can also be created from simulated light curves based on numerical models of the evolving young stars and their surrounding discs. To determine whether the variability in the simulated light curves is statistically similar to the observed light curves of YSOs, we need to develop a method to quantitatively compare samples of these fingerprints. Unsupervised machine learning techniques, such as dimension reduction and clustering with DBSCAN or k-means, might provide a method to achieve this. However, tests need to be performed to verify if these methods result in individual clusters of objects or if they create a continuum distribution. Furthermore, we need to examine how the various methods perform if samples of real data are investigated together with model fingerprints. Ideally, we are seeking to develop a methodology that creates clusters from a sample of fingerprints, and that is stable and robust. In other words, adding individual fingerprints based on model light curves into the sample to show where amongst the observed objects the model falls should not significantly change the assignments of the fingerprints to their clusters, or their relative positions within a continuum distribution.

In this paper, in Section 2 we describe the data calibration and verification as well as the selection of our highly variable YSO sample. In Section 3 we lay out the determination of the variability fingerprints and discuss their accuracy. Finally, in Section 4 we discuss the methodology and results of fingerprint analysis, as well as its stability, when comparing model and observed light curves.

2 PHOTOMETRY DATA AND VARIABLE STAR SELECTION

In this section, we describe the photometry data and their calibration, as well as the selection of variable stars for our analysis.

2.1 HOYS data and calibration

All photometry data used are provided by the HOYS citizen science project (Froebich et al. 2018). It aims to provide long-term, high cadence, multi-filter, optical photometry of 25 young (age < 10 Myr), nearby (distance < 1 kpc) star-forming regions. The data are collected by a mixture of ~130 professional, university, and amateur observatories. However, data taking for most of the light curves analysed is dominated by three 40-cm class observatories: (i) the University of Kent's Beacon observatory (Froebich et al. 2018); (ii) the AstroLAB IRIS Observatory (Evitts et al. 2020); (iii) the Remote Observatory Atacama Desert (ROAD; Hambsch 2012; Froebich et al. 2022). Collectively, they currently account for two-thirds of all images and just over half of the total exposure time.

All images for the project are bias-, dark-, and flat-corrected, uploaded to our publicly available web server,¹ and preliminarily calibrated. The astrometry in the images is solved using the `As-trometry.net` software (Hogg et al. 2008). Aperture photometry is performed using the Source Extractor software (Bertin & Arnouts 1996). Deep images of all regions obtained under photometric conditions in u , B , V , R_c , and I_c filters are used as reference for relative photometry. In this work, we only investigate data taken in the filters with most data, which are referred to as V , R , and I , hereafter. This follows the steps outlined in Froebich et al. (2018). The parameters of a photo-function (Moffat 1969; Bacher, Kimeswenger & Teutsch 2005) and fourth-order polynomial are fit and the relative magnitude $m_{\lambda,c}^i$ for each star i is determined from the instrumental magnitudes $m_{\lambda,in}^i$ following equation (1), where λ indicates the filter used:

$$m_{\lambda,c}^i = \mathcal{F}(m_{\lambda,in}^i) = A \log \left[10^{B(m_{\lambda,in}^i - C)} + 1 \right] + \mathcal{P}_4(m_{\lambda,in}^i). \quad (1)$$

Because a significant fraction of the amateur data are taken using slightly non-standard filters, or under non-photometric conditions, a second additional calibration step to correct colour terms needs to be applied to all data, the details of which are described in Evitts et al. (2020). We first identify non-variable stars in all fields, by selecting objects with a low Stetson index \mathcal{J} (Stetson 1996), and we measure their magnitudes and colours. These stars are then used to fit the parameters of a second-order polynomial in magnitude and colour (without cross-terms), including a common offset, p_0 , to correct the colour terms. The final calibrated magnitudes $m_{\lambda,f}^i$ for each star are then determined following

$$m_{\lambda,f}^i = \mathcal{W}(m_{\lambda,c}^i, V^i - I^i) = p_0 + \mathcal{P}^2(m_{\lambda,c}^i) + \mathcal{P}^2(V^i - I^i). \quad (2)$$

For the purpose of this work, we use the $V^i - I^i$ colour of the stars, but any other colour can be chosen. This colour is determined for each star from the median of the $m_{V,c}^i$ and $m_{I,c}^i$ values, that are available within 5 d of the image that is calibrated.

This colour correction usually significantly improves the root-mean-square (rms) scatter of the magnitude offsets of all calibration stars. We use this scatter of stars that have the same magnitude (within ± 0.1 mag) as the star in question, to determine the photometric uncertainty of the calibration (Evitts et al. 2020). All measurements with an uncertainty above 0.3 mag are excluded from the light curves analysed in this work.

2.2 Removing photometry outliers

The above detailed calibration ensures that the systematic photometry uncertainties are minimized. The nature of our HOYS data requires further data quality control to remove potentially erroneous data. Here, we detail the additional selection criteria applied to our data before they are used for analysis.

(i) Photometry closer than 5 arcmin to very bright stars ($G_{mag} < 6$ mag) has been removed for observatories where we have identified problems with point spread function wings influencing the photometry (including the University of Kent's Beacon observatory).

(ii) Some stars apparently have two brightness measurements on the same image. This is caused either by good seeing and the object being a binary (apparent or real), or by tracking issues in the image. All such photometry has been removed.

¹<https://astro.kent.ac.uk/HOYS-CAPS/>

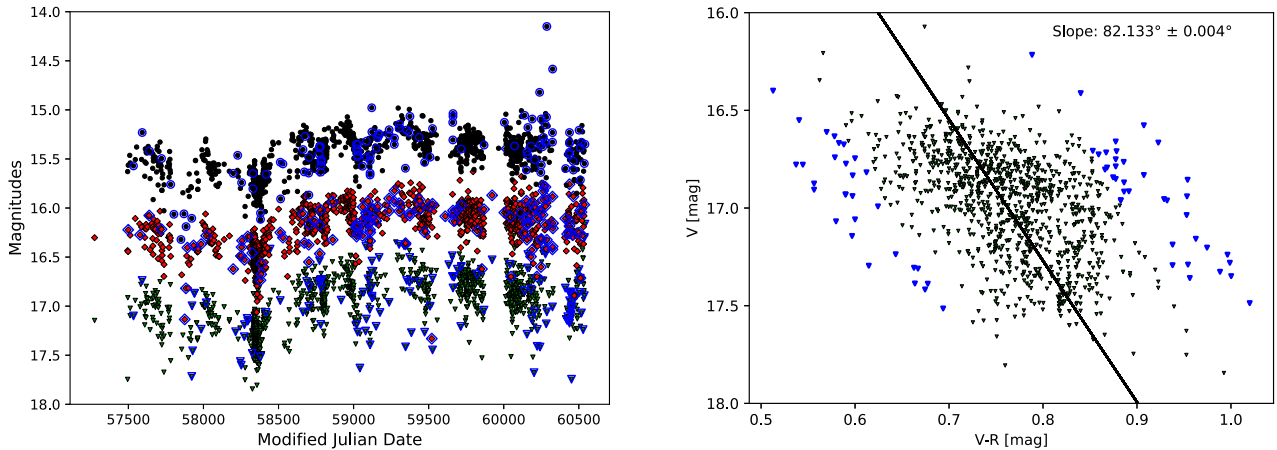


Figure 1. Left: example light curve of the variable source FHK 116. We show the V - (green triangles), R - (red diamonds), and I -band (black circles) light curves. The points highlighted in blue have been flagged up as potentially erroneous and have been removed from any subsequent analysis. Right: example $V - R$ vs V CMD for the same source. The best-fitting slope is overplotted and the 3σ outliers are indicated as blue triangles. Note the very different scales on the x - and y -axis.

(iii) We are not interested in very rare and short-term events such as flares. Thus, for each light curve and filter, we determine the mean and rms of all magnitude measurements. Any points that deviate by more than four times that value from the mean are removed.

(iv) For each light curve, a V vs $R - I$ colour-magnitude diagram (CMD) is created. For each V magnitude measurement, the $R - I$ colour is determined as the difference of the average of all R and I measurements within 2 d of the V data being taken. A linear perpendicular distance fit to the CMD has been obtained, and the rms scatter perpendicular to the line of best fit has been determined iteratively. All points further away from the line of best fit than three times that rms value are removed from the light curve. This is repeated for all possible combinations of the V , R , and I filters in the CMD (i.e. nine times).

Fig. 1 shows how the procedure performs for a typical variable source – in this case, the object FHK 116 (Hillenbrand et al. 2022), which is also known as 2MASS J20513340+4434543 or Gaia DR3 2163149219097677568. The left panel shows the V - (green triangles), R - (red diamonds), and I -band (black circles) light curves. We highlight in blue all data points that have been flagged up as potentially erroneous by our procedure and that are excluded from analysis. In the right panel of Fig. 1, an example $V - R$ vs V CMD for the same source is displayed. The best linear fit and the outliers that are removed are indicated.

2.3 Identification of variables

The above data extraction (performed on 2024 June 18), calibration and ‘cleaning’ have been applied to the just over 3000 light curves of potential HOYS cluster members identified in Froebrich et al. (2024) based on their *Gaia* Data Release 3 (DR3) astrometric properties (parallax and proper motion). This includes all clusters, even if they were not analysed in detail in that paper. No selection based on any other object properties (such as disk bearing or not) have been made. Here we only require a sample of clearly variable light curves to test our methodology (see Section 4). We removed all objects that had less than 100 data points in V , R , and I remaining after the cleaning.

For each light curve, we determined the Welch–Stetson index \mathcal{I} (Stetson 1996). This metric measures the correlation of the variations

between two different filters, and is defined for the combination of V and I filters in the following equation:

$$\mathcal{I}_{V,I} = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^N \left(\frac{V_i - \bar{V}}{\sigma_{V,i}} \right) \left(\frac{I_i - \bar{I}}{\sigma_{I,i}} \right)}. \quad (3)$$

It relies on having N contemporary pairs of brightness measurements and their uncertainties in the two filters. These are determined in the same way as for the colours in the CMDs (see Section 2.2), over a time period of 2 d.

In this work, we focus only on the data in the V , R , and I filters. Thus, we have three possible combinations for the Welch–Stetson index: $\mathcal{I}_{V,I}$, $\mathcal{I}_{V,R}$, and $\mathcal{I}_{R,I}$. Fig. 2 shows how these depend on the apparent magnitude of the sources. The majority of stars show only low-level variability, but the \mathcal{I} values go up to values of 1323. In our analysis, we found that 8, 5, and 6 per cent of the \mathcal{I} values were greater than 5 for $\mathcal{I}_{V,R}$, $\mathcal{I}_{V,I}$, and $\mathcal{I}_{R,I}$, respectively. These are not shown in Fig. 2 for clarity. For our subsequent analysis of highly variable sources, we select all stars that have an \mathcal{I} value above 2 for all three filter combinations. Such values usually indicate significant variability above the level of observational noise (e.g. Carpenter, Hillenbrand & Skrutskie 2001; Cody et al. 2014). This creates a sample of 240 highly variable sources. Note that the specific choices of time window lengths used to define contemporaneous brightness pairs for the Welch–Stetson index, as well as the threshold for selecting highly variable sources, do not affect our subsequent results.

3 VARIABILITY FINGERPRINTS

In this section, we describe how our YSO light curves are converted into variability fingerprints. We discuss their range and resolution, and how their uncertainties can be estimated.

3.1 Fingerprint determination

The light curves of all investigated variable YSOs have individual cadences. This is caused by the nature of HOYS, with its distributed observing. Thus, almost no two light curves can be directly compared, and we need to create a measure that allows us to capture the statistical

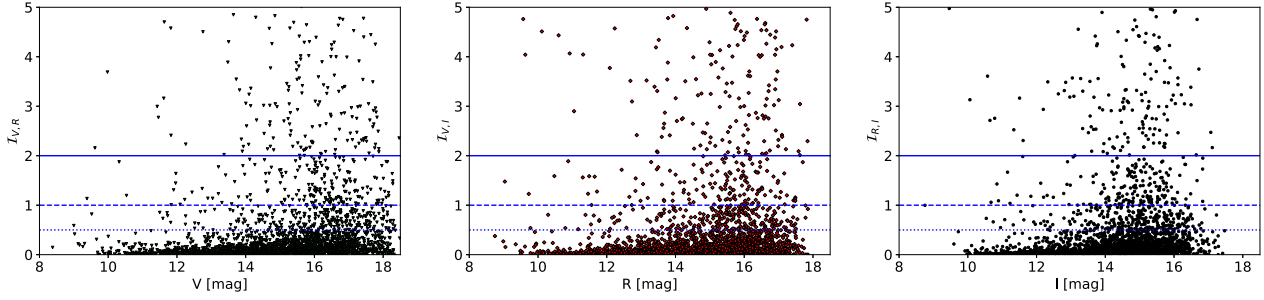


Figure 2. Welch–Stetson index of our YSO sample determined from all three filter combinations, $\mathcal{I}_{V,R}$, $\mathcal{I}_{V,I}$, and $\mathcal{I}_{R,I}$ from left to right. The cut-off for highly variable objects used in this study is marked at $\mathcal{I} = 2$ with a solid line. We also indicate $\mathcal{I} = 1$ and 0.5 with dashed and dotted lines, respectively.

properties of a star’s variability and to compare this between objects. These are the variability fingerprints.

We have introduced these fingerprints generated for the HOYS data in Evitts et al. (2020). They are based on earlier works by, for example, Scholz & Eislöffel (2004), Findeisen et al. (2015), and Rigón et al. (2017). The purpose of these variability fingerprints is to determine the probability $P(\Delta t, \Delta m)$ that a star varies by an amount Δm when observed at a time difference Δt apart. This can be done for light curves obtained in any filter. It assumes that the total length of the light curve and the observing cadence is sufficient so that all time-scales and magnitude changes have been covered. This is obviously not the case for rare events, such as major bursts or long dimming events. Thus, events that only occasionally occur in a fraction of light curves can only be captured statistically, by analysing a large sample of light curves and their fingerprints.

The determination of the fingerprints for a light curve with N data points ($t_i, m_i, i = 1, \dots, N$) is done using the following procedure.

(i) We determine all $N(N - 1)/2$ pairs of positive time differences $\Delta t = t_j - t_i$ (with $t_j > t_i$) in the light curve, as well as their corresponding magnitude difference $\Delta m = m_j - m_i$.

(ii) The distribution of the $(\Delta t, \Delta m)$ values is converted into a two-dimensional histogram. There are hence \mathcal{T} columns for the bins in Δt and \mathcal{M} rows for bins in Δm . The bin sizes chosen are discussed in Section 3.2. Each of the cells (pixels) in this histogram contains the number of $(\Delta t, \Delta m)$ pairs, i.e. $N(x, y)$ with $x = 1, \dots, \mathcal{T}$ and $y = 1, \dots, \mathcal{M}$.

(iii) To obtain the probabilities $P(\Delta t, \Delta m)$, the counts in each pixel are normalized by the total number of counts in the same column (range of Δt values) they are in, as shown in the following equation:

$$P(x, y) = \frac{N(x, y)}{\sum_{i=1}^{\mathcal{M}} N(x, i)}. \quad (4)$$

3.2 Fingerprint range and resolution

For our analyses, we need to choose the minimum and maximum ranges as well as the bin sizes and scaling in the variability fingerprints along both axes. In particular, the bin size determines the number of counts in each bin and thus the accuracy by which the probabilities can be determined. These are discussed in Section 3.3. Furthermore, the bin size sets the range of Δt and Δm over which the individual probabilities $P(\Delta t, \Delta m)$ are determined.

Along the Δt axis, the maximum range is set by our survey duration, which is up to 10 yr for some of the HOYS fields. Based

on the design of the survey and our science goals, we set a minimum range of 1 d. It is not practical to use linearly scaled bins along the Δt axis due to the wide range of scales. Thus, we use a log scaling (base 10) for the bin size in Δt , between the borders of $0.0 \leq \log[\Delta t \text{ (d)}] \leq 3.5$. This corresponds to time differences from 1 d to 8.67 yr. Equally sized bins in this log scaling then correspond to the same ratio of time differences. For example, a bin size of $\log[\Delta t \text{ (d)}] = 1$ means that the bin contains time differences that are a factor of 10 apart from smallest to largest, and bins of size $\log[\Delta t \text{ (d)}] = 0.3$ would correspond to time differences that are a factor of 2 apart from smallest to largest.

Along the Δm axis, the range is set by the typical maximum variability that the investigated objects have. We find that for most of our variable objects, this reaches about ± 1 – 2 mag. However, a few sources vary by more. Using linearly scaled bin sizes along the Δm axis is a valid choice. The bin size should then not be below the typical photometric uncertainty of the brightness measurements, which is of the order of a few per cent (e.g. Froebrich et al. 2018; Evitts et al. 2020). Thus, we use a bin size of 0.05 mag.

However, for larger variability amplitudes, the behaviour of a source is not expected to change significantly from one such bin to the next. In other words, the probability that a star varies by $+1.00$ mag on a given time-scale should be very close to the probability that it varies by $+1.05$ mag on the same time-scale. Hence, we are using *adaptive* bin sizes along the Δm axis, where each row of pixels away from the $\Delta m = 0$ mag axis has the range Δm increased by a value of \mathcal{X} times the range of the first pixel. A value of $\mathcal{X} = 0$ corresponds to linearly scaled pixels discussed above. For all values of \mathcal{X} between zero and a few, the results obtained do not significantly influence the results. We hence use $\mathcal{X} = 1$ as an example throughout the remainder of this paper. This means that the first pixels range from 0.00 to ± 0.05 mag. Each subsequent pixel is wider by the amount of 0.05 mag. Thus, the second pixels range from ± 0.05 to ± 0.15 mag, the third pixels from ± 0.15 to ± 0.30 mag, etc. This allows us to cover the variability range from -1.8 to $+1.8$ mag with only 16 of these adaptive pixels. The range and pixel number can be adjusted by choosing a different value for \mathcal{X} .

Fig. 3 shows several V-band fingerprints with different resolutions (including one with adaptive pixels) for the object FHK 116, the same source for which the light curves are plotted in Fig. 1. Up to time-scales of about 1 yr, the source does not vary considerably. On larger time-scales, the object is more likely to brighten than to become fainter, which is caused by the short dimming event that the object underwent near MJD = 58 400 d. The higher the resolution, the more detail is obviously visible. However, the accuracy of the map also decreases with higher resolution. We note that, other than the number

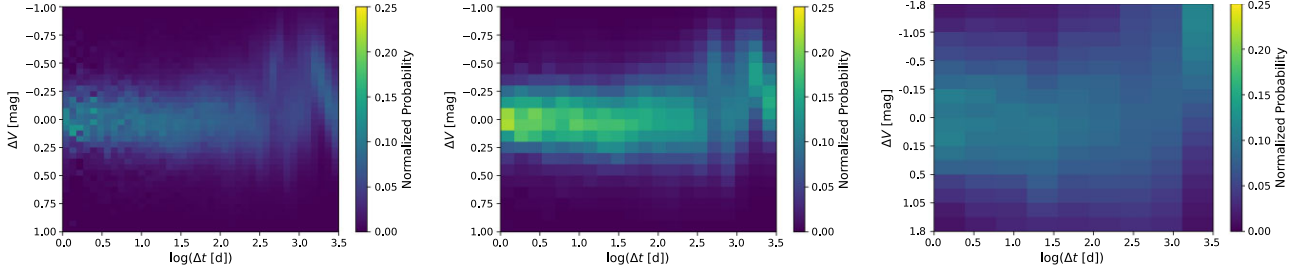


Figure 3. Example V -band variability fingerprints of the source FHK 116. From left to right we show the resolutions 40×40 , 20×20 , and 9×16 adaptive pixels. The range along the time axis is identical in all panels, but the adaptive pixel map on the right has a larger magnitude range than the other two maps.

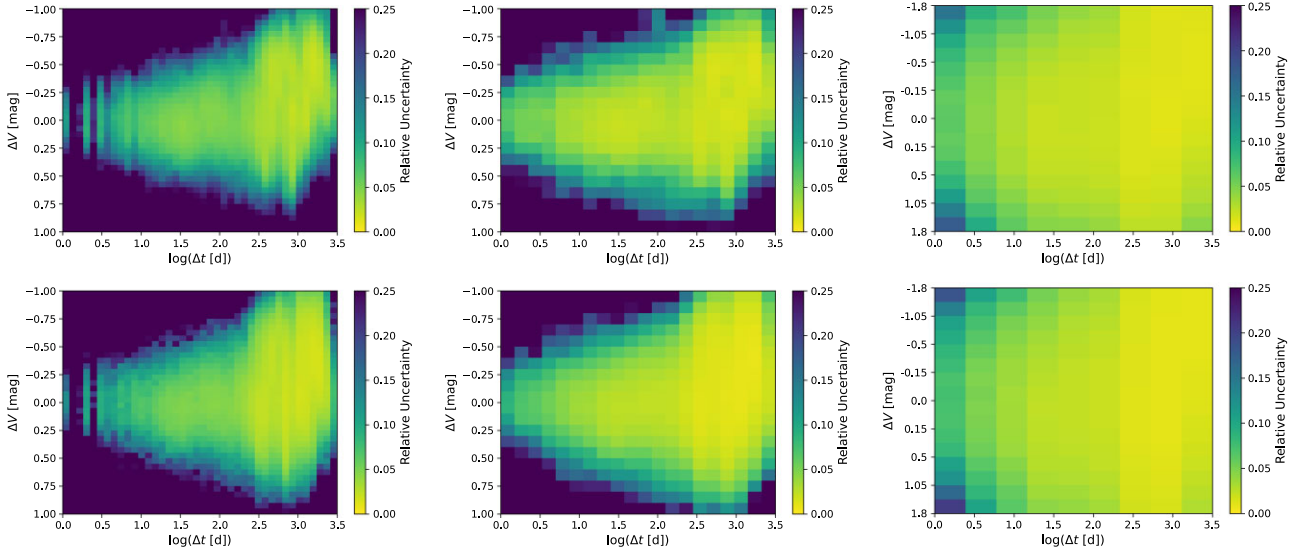


Figure 4. Example V -band variability fingerprint error maps of the source FHK 116. From left to right we show the resolutions 40×40 , 20×20 , and 9×16 adaptive pixels. The top row shows the maps determined from bootstrapping the light curve, and the bottom row when using the Poisson counting statistics.

of pixels without data in the fingerprints or low signal-to-noise (S/N) probabilities (see Section 3.3), the exact choice of bin size and range along both axes has no qualitative influence on the results of the analysis performed in Section 4. Only minor quantitative differences will occur if any of these values are changed, which have no influence on the results and conclusions of our work.

3.3 Fingerprint accuracy

The probabilities $P(\Delta t, \Delta m)$ in the fingerprints have varying uncertainties, depending on the number of counts $N(\Delta t, \Delta m)$ in each pixel and how close the individual $(\Delta t, \Delta m)$ pairs are to the pixel boundaries. Ideally, we only want to include high-significance $P(\Delta t, \Delta m)$ values in any subsequent analysis. Thus, we estimate the relative uncertainties of the fingerprints by two different methods: light-curve bootstrapping and Poisson counting statistics.

3.3.1 Light-curve bootstrapping

Each data point (m_i) in our light curve has an associated photometric uncertainty $\sigma(m_i)$ (see Section 2.1). We assume that there are no uncertainties for the times of the observation t_i as they are in any case of the order of the exposure times (typically minutes), which is at least three orders of magnitude below the smallest time differences included in the fingerprints.

To evaluate the uncertainties of the fingerprints, we repeat the determination of the fingerprint for bootstrapped light curves. In other words, each light-curve point is replaced by a random magnitude value. This value is chosen from a normal distribution with a mean of m_i and a variance of $\sigma(m_i)$. This is repeated a number of times. We then use the average pixel value from all these fingerprints as the $P(\Delta t, \Delta m)$ value, and the standard deviation as its uncertainty $\Delta P(\Delta t, \Delta m)$. We have run extensive tests to evaluate how many of these bootstrapping repeats are required to obtain accurate estimates for the fingerprint uncertainties. We find that typically after 30 000 repeats, the uncertainties will not vary by more than 10^{-4} from their value. Thus, this number was chosen for all bootstrapping uncertainty calculations.

In the top row of Fig. 4, we show some example V -band fingerprint relative uncertainty maps of object FHK 116 for several resolutions. It is evident that the accuracy decreases with increasing resolution. In particular, for the highest-resolution map (left panel), pixels that do not cover any integer day time difference have very high uncertainties due to our typical cadence of 1 d. Note that for the adaptive pixel maps (right panel), there are almost no low S/N pixels, even for the shortest time-scales.

3.3.2 Poisson counting statistics

The above described light-curve bootstrapping is very computationally intensive, as 30 000 fingerprints need to be calculated for each

light curve to obtain the uncertainties. We have thus investigated whether a much simpler (faster) method can lead to the same results. This is based on the assumption that the number of counts $N(x, y)$ in a pixel of the fingerprints follows the Poisson counting statistics. This would mean that their uncertainty is equal to the square root of their value. If this assumption is correct, then the uncertainties for each pixel can be estimated by simple error propagation of equation (4). This means that the relative uncertainties of the fingerprints can be determined as shown in the following equation:

$$\frac{\Delta P(x, y)}{P(x, y)} = \sqrt{\frac{1}{N(x, y)} + \left[\sum_{i=1}^M N(x, i) \right]^{-1}}. \quad (5)$$

In the bottom panel of Fig. 4, we show the relative uncertainty maps determined for the *V*-band fingerprints with the Poisson counting statistics, for the same maps as in the top row. We find that they are visually very similar, both qualitatively and quantitatively.

3.3.3 Comparison of fingerprint uncertainties

Here we briefly investigate the quantitative differences between the two uncertainty calculation methods. We have determined the uncertainty maps for both methods for all variable sources in our sample and for various resolutions. For each map, a linear regression between the Poisson and bootstrapping uncertainty values for all pixels has been performed. We have excluded all pixels in the maps where the relative uncertainties are worse than 0.33, that is, where the *S/N* of the $P(\Delta t, \Delta m)$ values is below 3. Note that such pixels are not very common in our maps (see, e.g. Fig. 4). In total, only about 10 per cent of pixels have such a low *S/N*, and these are exclusively confined to short time-scales and high amplitudes.

We find that the uncertainties based on the Poisson counting statistics are about 1.15 ± 0.10 times larger than the ones estimated from bootstrapping the light curves. This factor does not systematically depend on the resolution of the fingerprints or the filter in which the light curve has been observed. Thus, we can conclude that the fingerprint relative uncertainties can be estimated using the much faster method that relies on the assumption of Poisson counting statistics. However, they should be scaled down by a factor of 1.15 to represent the correct relative uncertainties based on the light-curve bootstrapping. We note that if fingerprint pixel sizes are used that are significantly different to the ones in this work, the scaling factor should be re-evaluated.

4 COMPARING VARIABILITY FINGERPRINTS

As outlined above, we aim to find a methodology that allows us compare the YSO variability fingerprints and to potentially identify clusters amongst them. In future, we plan to compare the observed fingerprints with fingerprints obtained from artificial observations of model discs. Thus, our goal is to establish a procedure that creates a preferably two-dimensional representation of the distribution of fingerprints in a sample, which is also robust. In other words, the addition of an individual object to the sample does not significantly change the distribution. Hereafter, we refer to this kind of distribution as the fingerprint landscape.

4.1 Pre-analysis

The variability fingerprints are high-dimensional data sets. Each pixel represents one dimension. For example, in the 9×16 adaptive pixel configuration, we have 144 dimensions. Thus, to

establish a two-dimensional landscape (which can also be used for clustering) a dimension reduction tool needs to be applied. For the purpose of this paper, we have tested principal component analysis (PCA) and *t*-distributed stochastic neighbour embedding (*t*-SNE). We have also tested two clustering algorithms, DBSCAN and *k*-means. Both are iterative but have different approaches to selecting cluster membership. Centroids are used in *k*-means, and core/border points in DBSCAN. The latter also uses *t*-SNE as part of its workflow.

We applied PCA using the standard scaler, which normalizes the value in each dimension (fingerprint pixel) to a mean of zero and a standard deviation of one (Hastie, Tibshirani & Friedman 2009). This is sufficient as all pixel values in the fingerprints represent probabilities (i.e. are already limited to between zero and one). The dimension reduction was done for several configurations (9×16 , 20×20 , etc.) of our 240 fingerprints (obtained from *V*-, *R*-, or *I*-band data) to investigate how much of the variance can be captured in just two dimensions. Generally, the configuration leading to the maximum value of variance captured should lead to the best clustering results.

We find that for all configurations, the first two components capture just below 50 per cent of the total variance. As there is no significant difference between these values, all subsequent analysis is only presented for the 9×16 adaptive pixel configuration and *V*-band data. Note that the results for all other fingerprint resolutions and filters do not significantly differ qualitatively and quantitatively. Generally, much higher values than 50 per cent for the total variance in the first two components are expected for distinct clusters (e.g. Murty & Devi 2016). Our low values thus indicate that there are most likely no distinct clusters identifiable in the fingerprint landscape, but that instead they form a continuum or single cluster.

For DBSCAN, three parameters must be determined. These are: epsilon (ϵ), which is directly a part of DBSCAN; perplexity (\mathcal{PP}), which is a parameter for *t*-SNE; and the minimum number (\mathcal{N}) of objects in a cluster. The values for these parameters can be determined by calculating the Davies–Bouldin index (DBI) for their values. We used ranges of $1 \leq \epsilon \leq 20$, $5 \leq \mathcal{PP} \leq 50$, and $5 \leq \mathcal{N} \leq 50$ with the condition that the number of clusters is greater than one. These ranges do encompass typical accepted values (\mathcal{PP} ; van der Maaten & Hinton 2008), or are sensible given our data (ϵ, \mathcal{N} ; Davies & Bouldin 1979; Ester et al. 1996). The combination that leads to the minimum DBI is the optimal value for the parameters. The minimum DBI obtained was approximately 0.45. The same value was obtained for a number of combinations of parameters, one of which was $\epsilon = 16$, $\mathcal{PP} = 7$, and $\mathcal{N} = 35$. The DBSCAN clustering result for this example is shown in the left-hand panel of Fig. 5.

A DBI of under 0.40 suggests highly compact clusters that are well separated (Davies & Bouldin 1979). Our index does not meet this criterion but is close to it. However, there are a large number of outliers evident in Fig. 5, roughly 40 per cent. This adds some further insight into the clustering outcome. The DBI is a metric of the clusters formed only. The large number of outliers spread evenly throughout the left panel of the figure shows that there is no real separation between the clusters, and the data as a whole are a continuum. For some of the combinations of parameters, the DBI index was greater than 2, indicating clusters that either are not compact or are poorly separated, or both. Some analysis was also carried out on the parameter space using a silhouette score. A maximum silhouette score of 0.4 was found for $\epsilon = 18$, $\mathcal{PP} = 5$, and $\mathcal{N} = 20$. Generally, a silhouette score above 0.5 is required for separated distinct clusters (Murty & Devi 2016). The low silhouette score could derive from the fingerprints forming a continuum, or

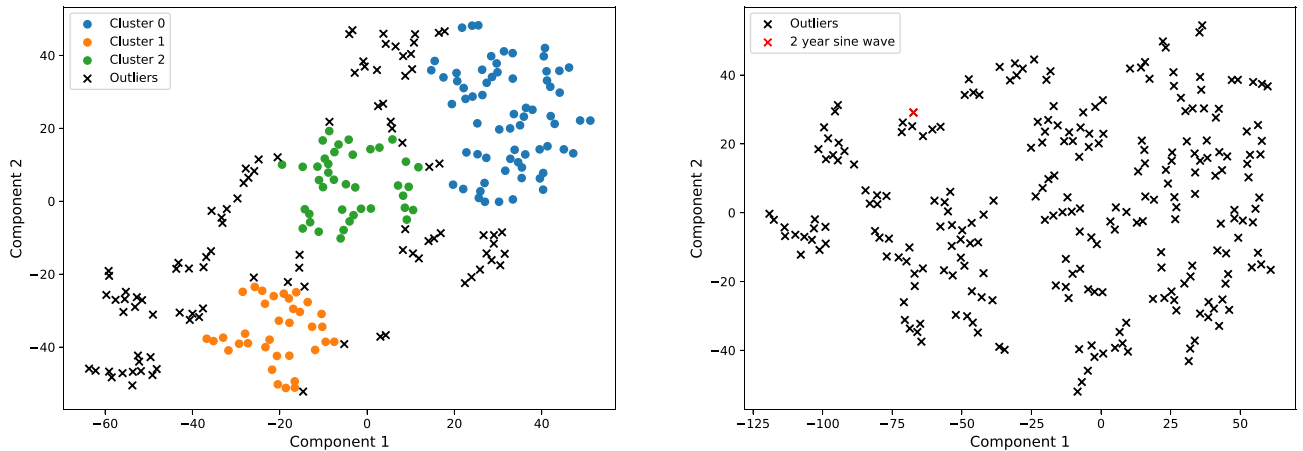


Figure 5. Left: DBSCAN clustering results for the 9×16 V -band adaptive pixel fingerprints of 240 variable YSOs. Right: the same data and process as in the left panel but with one model fingerprint added, which corresponds to a sine-wave-like light curve with an amplitude of 1 mag and a period of 2 yr.

because it is not suitable as a metric for concave clusters such as those formed by DBSCAN (Rousseeuw 1987).

Hence, all the indications show that the variability fingerprints in our sample do not form distinct clusters. Rather they form a continuum distribution in the landscape. Our subsequent analysis thus focuses on the stability and robustness of this landscape. Hence, the clustering algorithms applied to the data in the subsequent sections are for illustrative purposes only. The above analysis clearly shows that no algorithm will identify real clusters in our data.

4.2 Analysis with PCA, t -SNE, and DBSCAN

4.2.1 Clustering with DBSCAN

DBSCAN was chosen to illustrate clustering in our data set as t -SNE can be made part of its workflow. In addition, it is not sensitive to noise and will still form clusters. If there are outliers present, they will be identified as such and not given cluster membership based on an arbitrary condition, such as the number of clusters set manually (as in k -means; see Section 4.3.1).

We used the 240 object strong 9×16 V -band fingerprint sample and applied a dimension reduction using PCA from 144 down to 10 components. The output was passed to t -SNE and the dimension reduced further to two. Note that changing the value from which t -SNE reduces the dimension to two has no influence on the results, as long as t -SNE is used at all. DBSCAN was used with $\epsilon = 16$, $\mathcal{PP} = 7$, $\mathcal{N} = 35$, and 4000 iterations (see Section 4.1). The results of this clustering are shown in the left panel of Fig. 5. One can see that the fingerprints are grouped in three distinct clusters, but there are also a large number (102 – or 43 per cent) that have been labelled as outliers. We have visually inspected a number of randomly chosen pairs of light curves that have been assigned to the same cluster. They appear to show similar behaviour (see also Fig. A1 in the Appendix).

4.2.2 Stability of PCA and t -SNE

As indicated above, we require stability of the fingerprint landscape. To analyse this, we randomly selected one of the YSO light curves. The observing time (cadence) was maintained but the magnitude values were replaced by a sine wave with a 2-yr period and 1-mag amplitude. This was fingerprinted in the same way as all light curves and the fingerprint added into the sample. PCA and t -SNE were

used for dimension reduction. DBSCAN was used to illustrate the clustering with exactly the same parameters as in the previous section, just with this one artificial object added. The results are shown in the right-hand panel of Fig. 5.

There are significant differences in the landscape between the two versions. Some of the substructures are recognizable in both versions, but generally the entire structure has changed. However, points in close proximity do roughly remain in close proximity to each other. Notably, after adding just this one additional source, no clusters are formed at all, and all objects are classed as outliers. Thus, these significant changes in the landscape, as seen in both panels of Fig. 5, render the method not useful for our purpose. We note that for other values of ϵ , \mathcal{PP} , and \mathcal{N} the changes can be slightly less severe; however, there is no combination of these parameters that creates a stable landscape when one object is added.

We have also done extensive tests using different randomly selected light curves to create the artificial fingerprint, different fingerprint resolutions, filters of the photometry (V , R , I), and DBSCAN parameters (ϵ , \mathcal{PP} , \mathcal{N}). In all cases, we obtain the same qualitative and quantitative results. This method does not generate a stable landscape of points. This is caused by the non-linearity of the t -SNE process and the fact that our data are not clustered but form a continuum, as indicated by the DBI scores obtained (see Section 4.1). Thus, while the combination of PCA, t -SNE, and DBSCAN has some utility in grouping light curves with similar properties into clusters, the instabilities in the generated landscape caused by t -SNE make it unsuitable for our aim to insert an artificial light curve for comparison with the observed YSO sample.

4.3 Analysis with PCA and k -means

4.3.1 Clustering with k -means

The k -means clustering is generally thought of as a less sophisticated algorithm than DBSCAN. It converges quickly, often within a few iterations, and has only one parameter (the number of clusters) that needs to be set. This determines the number of centroids around which the clusters are formed. The process is sensitive to outliers and noisy data. Here we illustrate the k -means analysis of our data for the 9×16 adaptive pixel fingerprints made from the V -band light curves of 240 variable YSOs – the same as used in the previous section. We used four clusters as the input for k -means, but note that

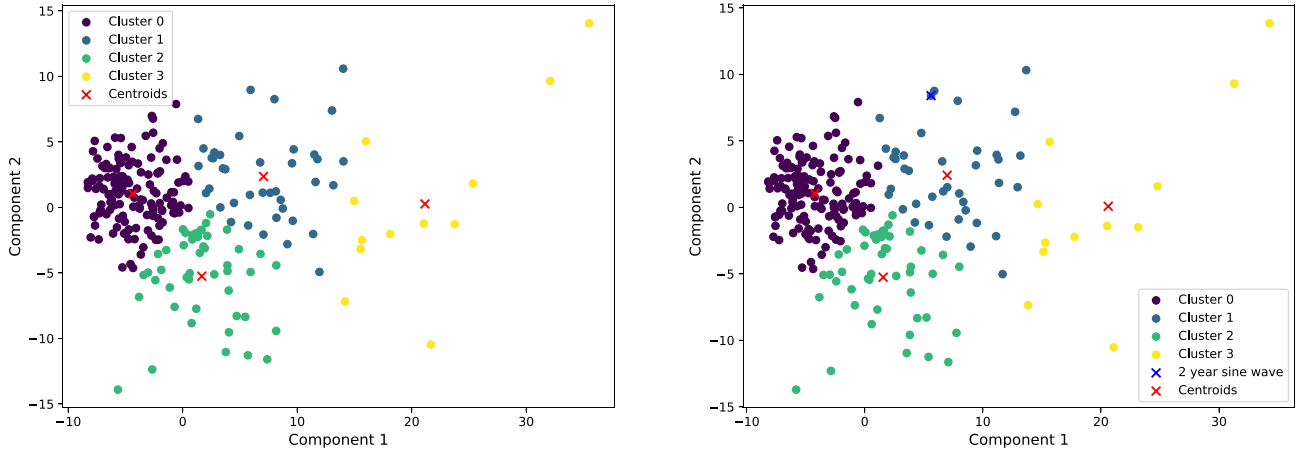


Figure 6. Left: clustering results with PCA and k-means using the 240 sized sample, the 9×16 V-band adaptive pixel fingerprints and four clusters. Right: the same process as the left panel with one artificial fingerprint added, which corresponds to a sine-wave-like light curve with an amplitude of 1 mag and a period of 2 yr.

specific results obtained do not depend on the exact value chosen. Scaling was applied using the standard scaler, and PCA was applied to reduce the dimensions from 144 to two.

The visual representation of the results of using PCA and k-means to illustrate the clustering can be seen in the left-hand panel of Fig. 6. Unlike the DBSCAN results in Fig. 5, the clusters formed by k-means have less separation. The landscape created by PCA is clearly a continuum instead of distinct clusters. A visual inspection of light curves assigned to different clusters but close to each other in the landscape shows that they subjectively look more similar than light curves of objects in the same cluster but far away from each other in the landscape. This is further illustrated by Fig. A1 in the Appendix and also holds for artificial light curves (see later in Section 4.4).

The differing outcomes observed when applying DBSCAN versus k-means are not a result of inherent structures in the data (the fingerprints represent a continuum distribution; see Section 4.1), but rather stem from the properties of the dimensionality reduction and clustering methods themselves. The k-means algorithm, typically applied after linear transformations such as PCA, is limited to identifying convex clusters centred around means. In contrast, DBSCAN, when used following a non-linear transformation such as t -SNE, can detect arbitrarily shaped clusters due to t -SNE's tendency to exaggerate local structure (Wang et al. 2021). However, this enhanced local separation does not indicate true clustering, especially in the absence of any real discrete groups. Instead, the apparent clusters produced by DBSCAN after t -SNE, especially with low perplexity and appropriate density parameters, are artefacts of these methods imposing structure on to fundamentally unclustered data. This highlights how combinations of dimensionality reduction and clustering techniques can generate misleading results when applied to continuous data sets.

4.3.2 Stability of PCA

We have established that the fingerprint landscape does not exhibit distinct clusters, and that the combination of PCA and t -SNE does not yield a stable configuration. We now assess whether using PCA alone produces a stable landscape. As in the earlier DBSCAN test, one light curve was randomly selected, and its original cadence retained while replacing its magnitudes with a synthetic sine wave (period, 2 yr;

amplitude, 1 mag). A fingerprint was generated for this artificial light curve and added to the data set before applying PCA. The resulting landscape is shown in the right-hand panel of Fig. 6, with k-means applied solely to illustrate the spatial distribution of fingerprints.

A comparison of the two panels in Fig. 6 shows that the positions of the original 240 fingerprints do not change by more than a small fraction of the typical separation between points (this is discussed in Section 4.4). Thus, the association of fingerprints with the four illustrative clusters has remained almost unchanged. Only one object shifted from cluster 0 to cluster 1. The synthetic sine wave was associated with cluster 1. Crucially, the structure of the landscape itself remains stable after the inclusion of the additional object.

To further quantify this stability, we performed several additional tests. We only display the landscape as determined by PCA in the subsequent figures, as we have established in Section 4.1 that there is no evidence of intrinsic clustering in our data. We utilize our 240 fingerprint sample and select randomly one of the objects. All individual photometry data points in that light curve are then bootstrapped (i.e. varied according to their photometric uncertainty). This means each magnitude is replaced by a number randomly drawn from a normal distribution with a mean value corresponding to the measured magnitude and a variance corresponding to its uncertainty. We created 1000 bootstrapped light curves (and fingerprints) for this object and determined the landscape with PCA for each of them separately with the other 239 objects. The results of the stability test for a typical source are shown in Fig. 7. Note that the size of the uncertainty ellipse is governed by photometric noise of the object and, at most, objects have twice the variation of the one shown.

Fig. 7 shows the average position in the landscape of the 1000 repeats for the unchanged objects as blue dots. The mean position for the bootstrapped light curve is shown in red. We also overplot as the diameter of the light-blue and light-red shaded ellipses the 1σ deviations from the mean for all objects. Note that these deviations are scaled up by a factor of 20 for better visibility. However, the changes in the position within the landscape for the unaltered objects are still smaller than the symbol size used for the mean positions. Considering the scaling in Fig. 7, we can conclude that the photometric uncertainty in the HOYS light curves does not significantly change the fingerprint landscape. At most, the uncertainties in the photometry move positions in the landscape by less than the typical separation of nearby neighbours.

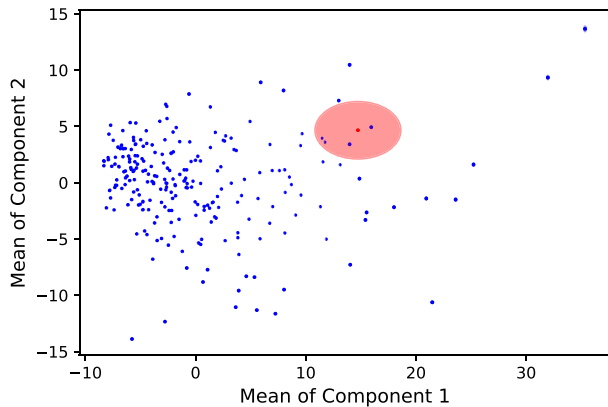


Figure 7. Stability test using PCA for dimension reduction. A randomly selected object (highlighted in red) has its light curve bootstrapped 1000 times. The mean positions (points) and standard deviations (shaded ellipses) in the fingerprint landscape are shown where the standard deviations are scaled up by a factor of 20 for better visibility.

4.4 PCA with artificial light curves

We have established above that the fingerprints of light curves of variable young stars form a continuum distribution. However, in Section 4.3.2 we have shown that the fingerprint landscape is stable when applying PCA for dimension reduction. Furthermore, the photometric uncertainties in the HOYS data have no significant influence on the position of an object in the landscape. As outlined in Section 1, our long-term aim is to use artificial light curves from numerical models to investigate whether they are consistent with the observed fingerprints. Thus, we need to consider what influence the timing and observing cadence have on the position of a particular underlying light curve in the fingerprint landscape. We thus slightly change the stability analysis done above using two sets of artificial light curves.

4.4.1 Sine wave light curves

We insert 2-yr period and 1-mag amplitude sine waves as artificial light curves (fingerprints) into the sample. This was done 960 times, and each of the 240 cadences of the original sample was used four times. In each case, a random phase shift was applied to the sine wave. This test shows how the observing cadence and the timing of the observations (phase shift for the sine-wave-like light curves) influence the position of the associated fingerprint in the landscape. We show the results in Fig. 8. The mean positions for the 240 variable objects are shown as blue points. Their standard deviation is again too small to be plotted. The black point indicates the mean position of all 960 artificial sine waves with a 2-yr period and 1-mag amplitude. The diameter of the grey-shaded ellipse represents the standard deviation from the mean (i.e. the typical scatter of the position of this type of light curve in the fingerprint landscape).

In contrast to the initial stability test, the standard deviations are much larger. This indicates that the timing of the observations and the observing cadence have an influence on the position of an artificial light curve (fingerprint) in the landscape. However, this range of positions is significantly smaller than the overall spread of points. This will allow us to evaluate qualitatively and quantitatively how much overlap there is in the variability characteristic of an artificial/model light curve and those of an observed sample of variable stars. Qualitatively, we can investigate where in relation

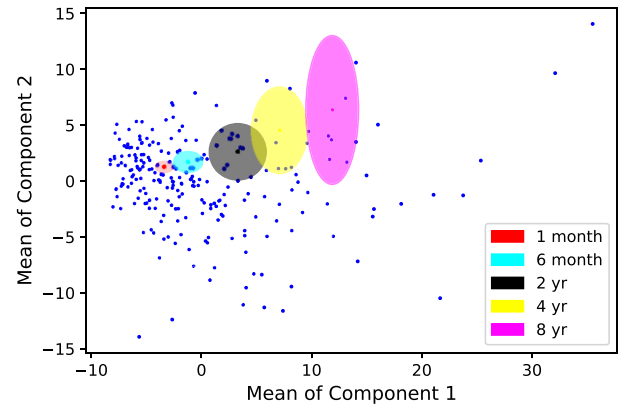


Figure 8. Stability test using PCA for our sample of 240 variable objects plus light curves representing sine waves. The approximate periods (see text for details) are indicated in the legend. The amplitudes of the sine waves are 1 mag in all cases. For each period, 960 repeats with bootstrapped photometry, random phase shifts, and varying cadences are used to determine the mean positions (points) and their standard deviations (shaded ellipses).

Table 1. Approximate periods (listed in Fig. 8) and the exact irrational periods used in the simulated sine wave light curves.

Approx. periods	One month	Six months	2 yr	4 yr	8 yr
Exact periods	$\frac{1}{3.8\pi}$ yr	$\frac{1}{0.66\pi}$ yr	$\frac{1}{0.15\pi}$ yr	$\frac{1}{0.08\pi}$ yr	$\frac{1}{0.04\pi}$ yr

to the observed fingerprints the model light curves are placed. If they are separated, then the model leads to variability characteristics that are not present in the observed sample. If there is overlap, then we can investigate whether there are characteristics in the light curve or YSO properties that distinguish objects inside and outside the overlap region in the landscape. A detailed discussion of this – including how the overlap of two samples can be quantified, how much overlap there is between YSO samples of different properties (e.g. evolutionary stages, inner disc excess emission, etc.) or model light curves – will be the subject of a follow-up investigation and is outside the scope of this current work. Here, we solely focus on establishing and testing the methodology to create a stable fingerprint landscape based on observed light curves of variable YSOs.

We have performed additional tests identical to the one discussed above, with different sine-wave-like model light curves. We used five different periods for sine waves with amplitudes of 0.5 and 1.0 mag. The periods correspond roughly to one month, six months, and 2, 4, and 8 yr. The exact values chosen (listed in Table 1) ensure that the periods are not exact integer days to avoid sampling issues with our typical observing cadences. Obviously, typical YSO light curves are not represented by sine waves. The exceptions are spotted stars, but these generally have small amplitudes and short periods (e.g. Froebrich et al. 2021; Herbert, Froebrich & Scholz 2023). Thus, one would expect only a small amount of overlap between these simple model light curves and the observed sample.

The results of this test for the 1.0-mag amplitude sine waves are summarized in Fig. 8. We can see that, regardless of the period and amplitude, the range in the landscape covered by the model fingerprints is small, compared with the spread of the entire YSO sample. The scatter in the landscape increases systematically with the period. This is understandable as our fingerprints only cover time-scales of up to 8.5 yr. Thus, the shorter periods fit into all the light curves multiple times. Hence, their variability characteristics are fully sampled in each light curve. However, for longer periods, each light

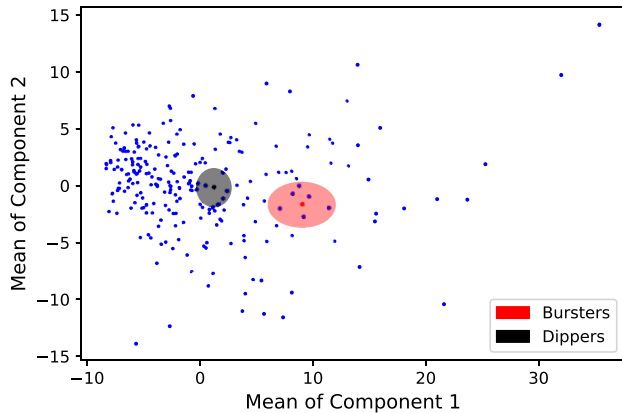


Figure 9. Placement test using PCA for our sample of 240 variable objects plus model light curves representing bursters and dipper. See text for details on the model parameters. The mean positions (points) and their standard deviations (shaded ellipses) for the models and data are determined in the same way as for Fig. 8.

curve only samples a part of the variations, thus causing the increased scatter. Furthermore, there are systematic changes of the average positions of the models with respect to the landscape. In particular, the placement is ordered by period; that is, the closer the periods are to each other, the closer their placement in the landscape. Longer-period light curves are placed systematically further away from the majority of observed fingerprints, to the right (positive values) along principal component 1 (PC 1). They are also systematically shifted towards slightly higher (positive) values along principal component 2 (PC 2). The same qualitative and quantitative behaviour is evident if we use a 0.5-mag amplitude in the sine waves. In Section 4.5, we investigate which light-curve properties are responsible for the placement in the fingerprint landscape, and we discuss how this explains the systematic behaviour seen in the simulated sine-wave-like light curves.

4.4.2 Dipper and burster light curves

In addition to the simple sine waves, we briefly investigate quantitatively the placement of burster- or dipper-like light curves in the landscape. A more detailed, qualitative investigation of this will be included in the above-mentioned follow-up investigation. We performed the same simulations (in terms of cadence, phase shifts, and repeats) as above, but changing the shape of the light curve into periodic dippers or bursters. The dippers are represented by symmetric dimming events, 0.4-mag deep. They are placed into the light curves every 150 d, with a duration of two months. The middle 60 per cent of the duration is spent in the dim state, with the remainder in ingress and egress. Bursts are placed into the light curve as one magnitude brightness increases every two years with a duration of six months. The rise is short within 10 d, then there is a 2.5 month plateau before the brightness decreases towards the baseline.

Fig. 9 shows the placement of these models in the landscape. We can see that their placement in the landscape and the scatter follow the same principles as for the sine waves. Models with more frequent events (in this case the dippers) are placed at lower PC 1 values, and vice versa. The shape of the variation seems to have more of a subtle influence on the placement along PC 2. We note that this is just a proof of concept example. The exact locations and scatter of the model light curves will systematically depend on their parameters. Furthermore, these simple model dipper/burster light curves do not

represent the full complexity of real YSO light curves. It should also be stressed that real objects coinciding with the locations of these models have the same variability statistics (fingerprints) but not necessarily light curves with the same dimming/brightening features as in these simplified models.

4.5 Placement in PCA-generated landscape

As discussed in detail above, the distribution of objects within the fingerprint landscape forms a continuum rather than distinct clusters. Nevertheless, certain regions of the landscape are more densely populated, while some objects appear as clear outliers. This suggests that specific light-curve characteristics are more prevalent than others. Furthermore, the artificial sine-wave-like light curves show a systematic trend in their placement, which is governed by the period. Thus, we investigate here which light-curve properties determine the placement of objects in the fingerprint landscape.

This is best done by visualizing the loadings matrices for the two components generated by the PCA. These are shown in Fig. 10. The colour code shows the weighting for each pixel in the fingerprints to the placement along each of the two components. In other words, if fingerprints have large pixel values in areas of the loading matrices that are red coloured, they will be placed towards larger values along the respective component (and vice versa).

For PC 1 (left panel in Fig. 10), we find that all fingerprint pixels that characterize light-curve variations with an amplitude of less than 0.3 mag will push an object towards large values of PC 1. In contrast, all variability in the light curve above an amplitude of 0.3 mag pushes the object towards smaller PC 1 values. Thus, a theoretically non-variable light curve would appear at very large PC 1 values. A high-amplitude light curve with stochastic variations on very short time-scales will be placed at very low PC 1 values. Light curves that start to vary on a given time-scale will be placed in between the two extremes.

We find that PC 1, derived from our sample of highly variable YSO light curves, therefore primarily reflects the time-scale for the onset of significant variability (exceeding 0.3 mag). This indicates that the timing of the onset of such variability serves as the most effective distinguishing feature among these light curves. The largest contributions to PC 1 (i.e. the highest absolute values in the loadings matrix) occur at time-scales between one and three months. This range is consistent with the findings of Rigon et al. (2017), who reported that most Class II YSOs exhibit low-amplitude (below 0.3 mag) optical variability driven by rotational modulation on time-scales of a few weeks. In contrast, stronger, longer-term variability associated with the inner disc and accretion processes affects roughly one-fifth of the population. Our results thus confirm that the onset of this additional, larger-amplitude variability typically occurs on time-scales of one to three months.

The loadings matrix for PC 2 (right panel in Fig. 10) is different. It is composed of pixel values close to zero for all time-scales below about 1.5 yr. Thus, for the placement of an object along PC 2, only the long-term behaviour of the light curve matters. In particular, light curves that show long-term brightening (i.e. negative ΔV values) will be placed at lower PC 2 values. Light curves that show long-term dimming behaviour have higher values of PC 2.

The loadings matrices derived from the sample of YSO light curves remain qualitatively and quantitatively consistent across varying conditions. Specifically, the choice of filter, the resolution of the fingerprints, and the threshold applied to the variability indices have no impact on the interpretation of the loadings matrices. As a result, replacing a single object in the sample produces only negligible

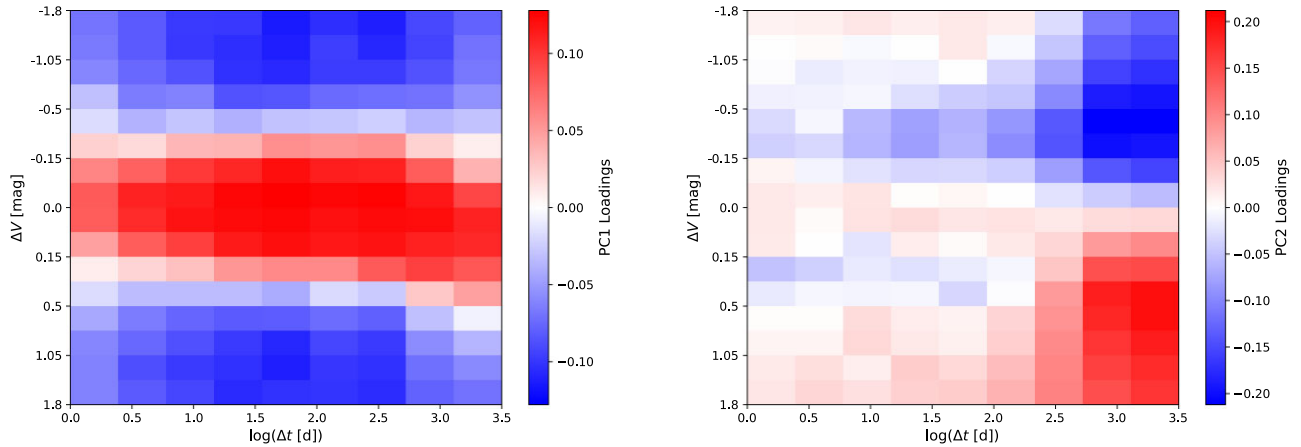


Figure 10. Loadings matrices of the 9×16 adaptive pixel fingerprints for principal component 1 (left) and principal component 2 (right). The colour scale shows how much each pixel value contributes to the positioning in the fingerprint landscape.

changes to the loadings matrices, accounting for the landscape stability described in previous sections. However, introducing a sample with substantially different characteristics, such as non-variable sources or a dominance of pure sine wave light curves, might alter the loadings matrices, potentially affecting the interpretation of PC 1/2.

The loading matrices for PC 1/2 can now also help us understand the systematic placement of sine-wave-like light curves in the landscape, as seen in Fig. 8 and discussed in Section 4.4. All these individual model light curves have been placed into the sample one by one. Thus, the loading matrices for each case are almost identical to those shown in Fig. 10. The placement of the sine waves along PC 1 is hence easy to understand. The variability in these light curves starts at time-scales of the order of a quarter of the period. Thus, the longer the period, the more non-variable does the light curve appear at shorter time-scales, and the further it is moved towards higher PC 1 values. If we create the same plot as in Fig. 8 but use 0.5-mag amplitude sine waves, then the models are systematically placed towards larger PC 1 values compared with the 1-mag amplitudes. This again is understandable because more of the detectable variations will be at amplitudes below 0.3 mag, which cause the objects to move towards larger PC 1 values. Along PC 2, the sine wave models systematically move towards larger values with increasing period. The long-term behaviour of the sine waves should be symmetric. However, there is a subtle feature in the loadings matrix for PC 2. For all time-scales below about 1.5 yr, variations of low amplitude cause a shift towards positive PC 2 values. The longer the period in the sine wave, the more likely we find low-amplitude variations at shorter time-scales, explaining the systematic shift. This feature is more evident in the higher-resolution maps.

5 CONCLUSION

We have refined the treatment of photometry in our inhomogeneous HOYS data set. Building on the standard calibration established in Froebrich et al. (2018) and Evitts et al. (2020), we exclude unreliable photometry near bright stars and from images with tracking issues. Additionally, potential photometric outliers in colour–magnitude space are identified and removed from the analysis. Using our previous astrometric selection of ≈ 3000 members in the monitored

young clusters (Froebrieh et al. 2024) and long-term V -, R -, and I -band photometry, we identify a sample of 240 highly variable YSOs.

We construct variability fingerprints from the light curves of these sources, mapping the probability that an object varies by a given amount over a given time-scale. This enables a quantitative comparison of the variability statistics of stochastically varying sources with randomly sampled light curves. Our data allow us to probe variability from ± 0.05 to ± 2.0 mag on time-scales from 1 d to 8.6 yr. Over 90 per cent of the fingerprints achieve an S/N ratio above three, with low-S/N regions confined to short time-scales and large amplitudes. Fingerprint uncertainties closely follow Poisson counting statistics.

Applying various dimension reduction and clustering techniques, we assess whether fingerprints form distinct groups and whether individual objects maintain stable relative positions when modifying the sample. Our analyses show that fingerprints create a continuous distribution rather than discrete clusters, consistent with a sample exhibiting variability from multiple physical processes, including accretion rate changes, line-of-sight extinction variations, and spotted stellar surfaces. However, objects close together in the landscape have similar light-curve morphologies.

We find that dimension reduction via t -SNE does not yield a stable landscape due to its non-linear nature. Individual object positions shift significantly when modifying the sample, though the large-scale structure remains recognizable. Conversely, PCA produces a highly stable landscape where adding or altering a single object results in only marginal shifts of the remaining points. This stability allows us to incorporate model-generated fingerprints and assess their placement relative to observations.

To test this, we simulate simple sinusoidal and burster/dipper light curves with varying HOYS cadences and randomized phase shifts. These models occupy a limited region of the observed landscape, indicating that observing time, cadence, and photometric errors do not significantly affect an object’s placement.

Analysis of the PCA loadings matrices shows that the primary source of variance among fingerprints derived from highly variable YSO light curves is the time-scale at which significant variability (exceeding 0.3 mag) begins, with time-scales of one to three months being the most prominent. The second most important contributor to the variance is long-term behaviour, specifically gradual fading or brightening trends occurring over time-scales longer than 1.5 yr.

These results demonstrate that PCA of variability fingerprints provides a robust, quantitative method to compare the variability statistics of observed YSO light curves with artificial light curves derived from models of the underlying physical processes.

ACKNOWLEDGEMENTS

We would like to thank all contributors of observational data for their efforts towards the success of the Hunting Outbursting Young Stars project. HS is supported by the Science and Technology Facilities Council under grant number STFC Kent ST/Y509267/1.

DATA AVAILABILITY STATEMENT

The data underlying this article are available in the HOYS data base at <https://astro.kent.ac.uk/HOYS-CAPS/>.

REFERENCES

- Audard M. et al., 2014, in Henning T. K. et al. (eds), *Protostars and Planets VI*. University of Arizona, Tuscon, AZ, p. 387
- Bacher A., Kimeswenger S., Teutsch P., 2005, *MNRAS*, 362, 542
- Bertin E., Arnouts S., 1996, *A&AS*, 117, 393
- Bouvier J. et al., 1999, *A&A*, 349, 619
- Bouvier J., Alencar S. H. P., Harries T. J., Johns-Krull C. M., Romanova M., 2007, in Reipurth B., Jewitt D., Keil K. eds, *Protostars and Planets V*. University of Arizona Press, Tuscon, AZ, p. 479 ([arXiv:astro-ph/0603498](https://arxiv.org/abs/astro-ph/0603498))
- Bouvier J., Matt S. P., Mohanty S., Scholz A., Stassun K. G., Zanni C., 2014, in Henning T. K. et al. (eds), *Protostars and Planets VI*. University of Arizona Press, Tuscon, AZ, p. 433
- Carpenter J. M., Hillenbrand L. A., Skrutskie M. F., 2001, *AJ*, 121, 3160
- Cody A. M. et al., 2014, *AJ*, 147, 82
- Cody A. M., Hillenbrand L. A., Rebull L. M., 2022, *AJ*, 163, 212
- Davies D. L., Bouldin D. W., 1979, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, 224
- Ester M., Kriegl H.-P., Sander J., Xu X., 1996, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, Cambridge, MA, p. 226, <https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf>
- Evitts J. J. et al., 2020, *MNRAS*, 493, 184
- Findeisen K., Cody A. M., Hillenbrand L., 2015, *ApJ*, 798, 89
- Fischer W. J., Hillenbrand L. A., Herczeg G. J., Johnstone D., Kospal A., Dunham M. M., 2023, in Inutsuka S., Aikawa Y., Muto T., Tomida K., Tamura M. eds, *ASP Conf. Ser. Vol. 534, Protostars and Planets VII*. Astron. Soc. Pac., San Francisco, CA, p. 355 ([arXiv:2203.11257](https://arxiv.org/abs/2203.11257))

- Froebich D. et al., 2018, *MNRAS*, 478, 5091
- Froebich D. et al., 2021, *MNRAS*, 506, 5989
- Froebich D., Eislöffel J., Stecklum B., Herbert C., Hamsch F.-J., 2022, *MNRAS*, 510, 2883
- Froebich D. et al., 2024, *MNRAS*, 529, 1283
- Grankin K. N., Melnikov S. Y., Bouvier J., Herbst W., Shevchenko V. S., 2007, *A&A*, 461, 183
- Hamsch F. J., 2012, *Journal of the American Association of Variable Star Observers*, 40, 1003
- Hastie T., Tibshirani R., Friedman J., 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, Berlin
- Herbert C., Froebich D., Scholz A., 2023, *MNRAS*, 520, 5433
- Herbst W., Eislöffel J., Mundt R., Scholz A., 2007, in Reipurth B., Jewitt D., Keil K. eds, *Protostars and Planets V*. University of Arizona Press, Tuscon, AZ, p. 297 ([arXiv:astro-ph/0603673](https://arxiv.org/abs/astro-ph/0603673))
- Hillenbrand L. A., Kiker T. J., Gee M., Lester O., Braunfeld N. L., Rebull L. M., Kuhn M. A., 2022, *AJ*, 163, 263
- Hogg D. W., Blanton M., Lang D., Mierle K., Roweis S., 2008, in Argyle R. W., Bunclark P. S., Lewis J. R. eds, *ASP Conf. Ser. Vol. 394, Astronomical Data Analysis Software and Systems XVII*. Astron. Soc. Pac., San Francisco, CA, p. 27
- Joy A. H., 1945, *ApJ*, 102, 168
- Kochanek C. S. et al., 2017, *PASP*, 129, 104502
- Lakeland B. S., Naylor T., 2022, *MNRAS*, 514, 2736
- van der Maaten L., Hinton G., 2008, *Journal of Machine Learning Research*, 9, 2579
- Moffat A. F. J., 1969, *A&A*, 3, 455
- Murty M., Devi V., 2016, *Pattern Recognition: An Algorithmic Approach*. Springer, Berlin
- Rigon L., Scholz A., Anderson D., West R., 2017, *MNRAS*, 465, 3889
- Rousseuw P. J., 1987, *Journal of Computational and Applied Mathematics*, 20, 53
- Scholz A., Eislöffel J., 2004, *A&A*, 419, 249
- Sergison D. J., Naylor T., Littlefair S. P., Bell C. P. M., Williams C. D. H., 2020, *MNRAS*, 491, 5035
- Sokolovsky K. V. et al., 2017, *MNRAS*, 464, 274
- Stetson P. B., 1996, *PASP*, 108, 851
- Wang Y., Huang H., Rudin C., Shaposhnik Y., 2021, *Journal of Machine Learning Research*, 22, 1

APPENDIX A: EXAMPLE LIGHT-CURVE PAIRS

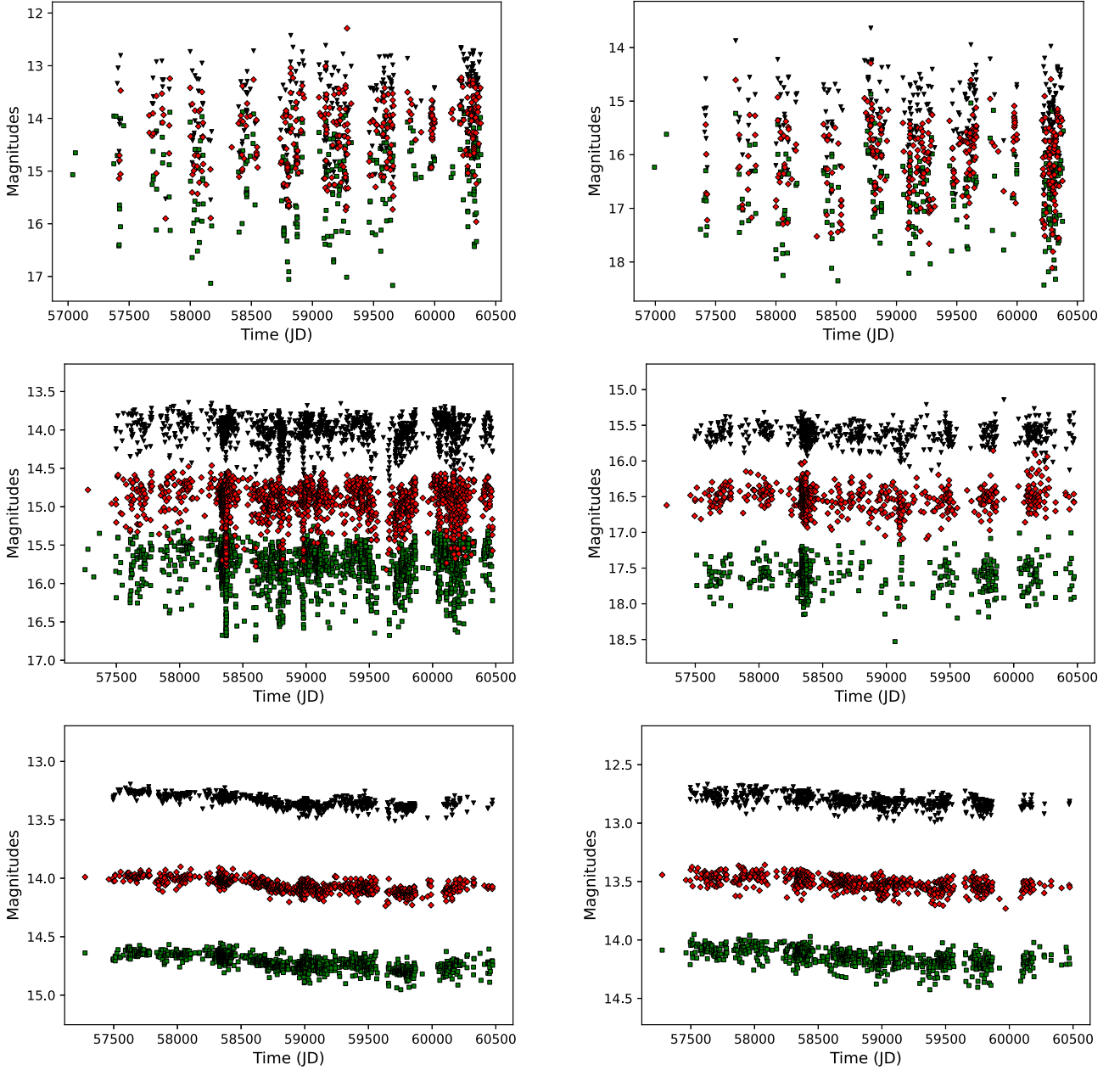


Figure A1. Example HOYS light curves. Each row contains two examples of light curves that are in close proximity to each other in the fingerprint landscape. In the top row, we have two sources with low PC 1 values, the middle row contains objects with intermediate PC 1 values, and the bottom row shows light curves with high values for PC 1.

This paper has been typeset from a \LaTeX file prepared by the author.