Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **508,** 5734–5756 (2021) https://doi.org/10.1093/mnras/stab2588
Advance Access publication 2021 September 11

# A method for finding anomalous astronomical light curves and their analogues

J. Rafael Martínez-Galarza [1]★ Federica B. Bianco,[2,3,4,5] Dennis Crake,[1,6,7] Kushal Tirumala,[8] Ashish A. Mahabal [8] Matthew J. Graham [8] and Daniel Giles [9]

[1]*Center for Astrophysics | Harvard & Smithsonian, 60 Garden St, Cambridge, MA 02138, USA*
[2]*Department of Physics and Astronomy, University of Delaware, 217 Sharp Lab, Newark DE 19716 USA*
[3]*Joseph R. Biden, Jr. School of Public Policy and Administration, University of Delaware, 184 Academy St, Newark DE 19716 USA*
[4]*University of Delaware Data Science Institute, 100 Discovery Blvd, Newark, DE 19713, USA*
[5]*Center for Urban Science and Progress, New York University, 370 Jay St, Brooklyn NY 11201, USA*
[6]*School of Physics and Astronomy, University of Southampton, Hampshire SO17 1BJ, UK*
[7]*Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK*
[8]*Division of Physics, Mathematics and Astronomy, California Institute of Technology, Pasadena CA 91125, USA*
[9]*The SETI Institute. 189 Bernardo Ave, Suite 200, Mountain View CA 94043, USA*

Accepted 2021 September 8. Received 2021 September 8; in original form 2020 September 15

## ABSTRACT

Our understanding of the Universe has profited from deliberate targeted studies of known phenomena, as well as from serendipitous unexpected discoveries, such as the discovery of a complex variability pattern in the direction of KIC 8462852 (Boyajian's star). Upcoming surveys such as the Vera C. Rubin Observatory Legacy Survey of Space and Time will explore the parameter space of astrophysical transients at all time-scales, and offer the opportunity to discover even more extreme examples of unexpected phenomena. We investigate strategies to identify novel objects and to contextualize them within large time-series data sets in order to facilitate the discovery of new classes of objects as well as the physical interpretation of their anomalous nature. We develop a method that combines tree-based and manifold-learning algorithms for anomaly detection in order to perform two tasks: 1) identify and rank anomalous objects in a time-domain data set; and 2) group those anomalies according to their similarity in order to identify analogues. We achieve the latter by combining an anomaly score from a tree-based method with a dimensionality manifold-learning reduction strategy. Clustering in the reduced space allows for the successful identification of anomalies and analogues. We also assess the impact of pre-processing and feature engineering schemes and investigate the astrophysical nature of the objects that our models identify as anomalous by augmenting the Kepler data with Gaia colour and luminosity information. We find that multiple models, used in combination, are a promising strategy to identify novel light curves and light curve families.

**Key words:** methods: data analysis – methods: statistical – stars: flare – stars: peculiar.

## 1 INTRODUCTION

Scientific discovery in astronomy can be thought of as happening via two complementary approaches. The *question-driven* approach attempts to provide answers to questions that have already been conceived based on our present knowledge of existing theories, models, or observed phenomena. Cosmological supernovae (SNe type Ia) are a good example. They are relatively well-known objects to us in terms of their energetic output, redshift distribution, and spectral properties, and we design surveys with observational parameters fine-tuned to find them on the basis of those known properties to improve our understanding of cosmology as well as stellar physics. The *exploration-driven* approach, on the other hand, attempts to enable us with the capability to find objects that are unknown, unexpected, or extremely rare, by either expanding the space of observational parameters (for example by increasing the spatial resolution available to us with a larger telescope) or by employing novel ways to dissect the increasingly complex data sets that are becoming available to us (Li et al. 2021). These unexpected discoveries often require additions or modifications to our theoretical apparatus, and, on occasion, force us to formulate new hypotheses.

This approach has led to serendipitous discoveries, which are prevalent in astronomy and which have produced significant breakthroughs. Recent examples include the discovery of peculiar light curves in *Kepler* data such as KIC 8462852 commonly known as Boyajian's star (Boyajian et al. 2016), the discovery of the interstellar object 1I/'Oumuamua (Meech et al. 2017), and the detection of quasi-periodic oscillations in the X-ray light curve of galaxy G 159 (Miniutti et al. 2019), among many others. The question then becomes, how do we make serendipity 'systematic' (Giles & Walkowicz 2019) in order to increase the chance of discovery in the era of large astronomical data sets? A related question that is at the core of this paper is whether it is possible to find analogues of a particular anomaly of interest.

★ E-mail: jmartine@cfa.harvard.edu

Anomalies are traditionally identified as data points that lie beyond some threshold distance from the bulk of the population of other data points in some representative space (see for example the comparative study in Goldstein & Uchida 2016). This distance is usually determined in terms of the population scatter, often measured by the standard deviation. For example, assuming a Gaussian distribution, one can flag one in 370 objects ($3\sigma$) or one in 1.7 million objects ($5\sigma$) as 'outliers'. However, there is no reason to think that an arbitrary data set in an arbitrary data space should be described by a Gaussian distribution. Furthermore, Aggarwal & Yu (2001) have shown that as the number of dimensions increases, the proximity-related measures of similarity between objects become less meaningful, since in a high-dimensional space objects are more sparse, and more likely to show up as proximity-based outliers. In addition, anomaly detectors can also be affected by false alarms, either false positives or false negatives.

Time domain-based studies of astrophysical phenomena date back centuries, but in the context of modern astrophysical surveys an example can be found in Eyer et al. (2019), where the authors also emphasize the pivotal role of time-domain surveys in the future of astrophysics. Reviews of anomaly detection methods in time-series data can be found in Goldstein & Uchida (2016) and Blázquez-García et al. (2020). Several unsupervised and semisupervised algorithms for astrophysical anomaly detection have been used, including approaches that use Euclidean proximity or clustering information to isolate anomalies (Dutta et al. 2007; Henrion et al. 2013; Giles & Walkowicz 2019), and approaches that use more complex representations of the data that do not involve their projection into an Euclidean space, such as neural networks, ensemble methods, and active learning (Baron & Poznanski 2017b; Druetto et al. 2019; Margalef-Bentabol et al. 2020; Škoda, Podsztavek & Tvrdík 2020), as well as Gaussian processes (Chen et al. 2018). Significant effort has been put into the problem of anomaly detection in supernova surveys, in particular by the Supernova Anomaly Detection (SNAD) group, which has used isolation forest and active learning to boost the discovery of unusual objects (Pruzhinskaya et al. 2019; Aleo et al. 2020; Ishida et al. 2021). There are also significant differences in the feature engineering aspects of those algorithms. While some of them require a feature extraction step that produces a set of synthetic features to represent the data, others work directly on the data points, either light curves, spectra, or images. While each of these methods can perform well for specific data sets, one could ask whether different methods applied to the same data set would find the same anomalies.

One potential caveat of anomaly detection algorithms based on a similarity score calculated using a certain set of features is that the selected features might not fully characterize the anomalous nature of the objects. That is, the objects can be anomalous in a certain space, but not in all possible representations. This makes it hard to efficiently identify objects that are 'like' a particular object of interest when dealing with a large data set, because the anomaly score based on those features is usually a one-dimensional quantity. Our hypothesis in the present work is that by combining a manifold-learning proximity method with an independently derived anomaly score from a tree-based method, one can effectively break the degeneracies in the one-dimensional space of the anomaly score, and make useful inferences regarding which anomalies are similar to each other, simplifying the discovery of analogues and new classes. This approach also offers the advantage that a different set of features can be used for the anomaly scoring and for the clustering, offering additional flexibility.

Previous work has made significant contributions towards the goal of finding analogues to Boyajian's star and other anomalies. For example Giles & Walkowicz (2019, 2020) apply a clustering method to a set of synthetic features derived for Kepler light curves and demonstrate that their method is capable of identifying anomalies such as Boyajian's star, as well as cataclysmic variables. Schmidt (2019) use a photometric selection method to isolate analogues of KIC 8462852 by looking for light curve dips in All Sky Automated Survey for Supernovae (ASAS-SN; Kochanek et al. 2017) data, and they find about 20 similar objects that deserve follow-up studies (see also Lochner & Bassett 2021). Yet, to the best of our knowledge, no reproducible methods have been proposed with the specific goal of finding analogues to a light curve of interest, be it Boyajian's star or other type of anomalous light curve. In this paper we aim to provide a recipe not only for finding the most compelling objects in a time domain survey, but also for finding any analogues (*i.e.* objects with similar light curves) of those objects. The method combines:

(i) A tree-based anomaly detection algorithm, the unsupervised random forest (URF; Shi & Horvath 2006) that operates on the joint space of light curve points and power spectrum.

(ii) Two manifold-learning algorithms: *t*-SNE (Maaten & Hinton 2008) and UMAP (McInnes, Healy & Melville 2018), which operate on an image representation of the light curves and finds low-dimensional embedded representations of these images.

We apply this combined method to the full set of Quarter 16 light curves from the *Kepler* Space Observatory, and test it against a set of previously identified anomalies in the *Kepler* data set (Giles & Walkowicz 2020). We rank the anomalies according to their URF-based similarity score and compare them with the scores for the general population in order to assess the ability of the method to identify *bona-fide* anomalous objects. In order to test our ability to find analogues, we investigate the location and clustering properties of these anomalies in the space of embedded features derived from the manifold methods, and whether other similar and previously unknown anomalies are identified. By using *Gaia* observations of the sources to construct their Hertzprung–Russell diagram, we find that the method is able to identify anomalies that share astrophysical properties, either intrinsic or extrinsic.

The Kepler telescope generated evenly sampled time-series. This is possible with a space mission, but impossible with ground-based surveys, due to weather, moon phase, visibility, and, generally, optimization of survey strategies (Bianco et al. 2021). Surveys like the Catalina Realtime-Transient Survey (Drake et al. 2012) and the Zwicky Transient Facility (Bellm et al. 2019) are an invaluable reservoir of transients, all measured with unevenly sampled light curves. Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST; Ivezić et al. 2019) will generate an unprecedentedly large data set including tens of billions of stars measured with over $\sim$800 unevenly sampled data points. To test if our results generalize to unevenly sampled data sets such as the LSST data, we generate a second data set from the Kepler data by sub-sampling the original time-series, and repeat our analysis on these uneven, sparse light curves.

This paper is organized as follows. In Section 2 we describe the data set used to test the performance of our method, and discuss the impact of different pre-processing in the process of finding anomalies. In Section 3 we describe the methods used in the paper for anomaly detection, dimensionality reduction, and clustering. In Section 4 we present our set of *bona-fide* anomalies, show that they can be identified with our method, and identify additional anomalies in the data set. We also show how the manifold methods can be used to group similar anomalies together, and how those similarities relate to shared astrophysical properties. Finally, in Section 5, we discuss and summarize our findings.

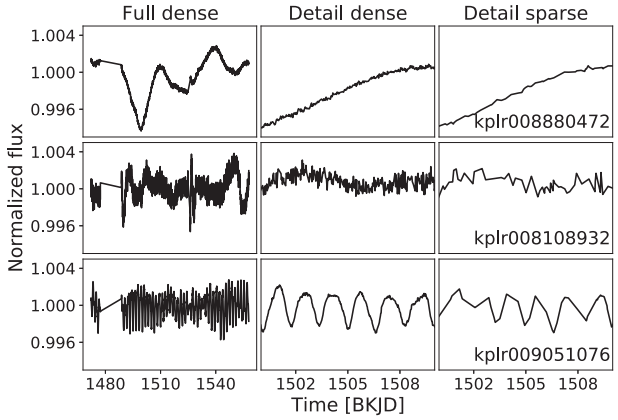


**Figure 1.** Examples of normalized Kepler light curves used in this paper. The left-hand panels show the full 85 d long Q16 dense light curves for three different Kepler targets. The middle panels show a detail of the same light curves plotted over a period of 10 Julian days. The right-hand panels show the same detail but this time for the corresponding sparse version of the light curves. The Kepler source IDs are shown. Note the data gap early on in each light curve due to missing observations.

## 2 DATA SET AND FEATURE EXTRACTION

In this section, we describe the data set used in the subsequent analysis as well as the feature extraction approaches that we employed to create the input to the anomaly detection and dimensionality reduction methods.

### 2.1 Kepler light curves

The data set considered here is comprised of 147 036 long cadence light curves from the *Kepler* telescope archive. These light curves were obtained by searching the Mikulski Archive for Space Telescopes (MAST) for all light curves in Quarter 16, and may be obtained from the MAST archive at https://dx.doi.org/10.17909/T9SW2G. From the total of over 160 000 light curves, we then excluded those that we were not able to pre-process with the methods of Section 2.3. The downloaded light curves are de-trended from spacecraft effects and have a uniform cadence, with one photometry point obtained every 30 min. Our pre-processed version on the light curves (see Section 2.3) covers a total span of approximately 85 d, starting at Barycentric Kepler Julian Date (BKJD) 1472 and ending at BKJD 1558, for a total of 3520 measurements per light curve. We used the most up-to-date pipeline processed data (PDCsap flux) according to the Kepler Data Release Notes and the Kepler Data Processing Handobook (Jenkins 2017). Fluxes are given in relative flux units as provided in the standard Kepler de-trended light curves delivered by the MAST archive. We normalized the fluxes to have a mean value of 1, *i.e.* we divided them by their mean. These normalized fluxes are not formally standardized, *i.e.* their standard deviations are those resulting from this normalization process, without any further adjustments. We chose this approach because we do not want the anomaly detection being dominated by the noise level. In what follows, however, we use standardized and normalized indistinctly. In Fig. 1 we show a small sample of the light curves studied here, in order to illustrate the wide range of variability properties in the Kepler targets.

In order to assess the effect that uneven sparse sampling has on our ability to find anomalies, we have also produced a separate data set by uniform random sub-sampling of the full light curves, selecting only ∼10 per cent of the original data points time-stamps (the same time-stamps for all light curves). We refer to this second set of light curves as the *sparse* set, as opposed to the *dense* set of the original light curves. Three sparse light curves are shown in Fig. 1, right-hand panel.

### 2.2 Feature engineering

Selecting and engineering features from astronomical light curves (and more generally from time-series and other one-dimensional data) is a crucial step in setting up a successful anomaly detection algorithm, and, in general, any machine learning model, either supervised or unsupervised (Fulcher 2017). The decision regarding which features to use should be based upon the predictive power of these features and on how efficiently they can be used to split the data set into multiple classes, or as is the case here, on how well this set of features represent the different types of variability of the original data set.

Feature selection is crucial not only to maximize the efficiency of our methods, but especially if we intend to use the features to physically characterize the objects. In the case of supervised tree methods, on which some of the models we use are based, since at each step the model deals with only a single feature and features are never combined in a mathematical sense, the predictions are generally robust against covariance (Breiman 2001; Biau 2012). Consider, however, the case of two features that are completely co-linear: the trees in the forest will use one or the other based on an initial random selection. When encountering the second feature the trees will disregard it since the information carried by the feature has already been used in splits based on the other co-linear feature. However, this affects the feature importance evaluation. Trees assess the importance of features based on how much each feature contributes to splitting. In the covariant example, the second co-linear feature does not ultimately contribute and therefore its covariance with other features has suppressed its importance.

Feature engineering is particularly challenging when the sampling of a time-series is sparse and uneven, especially if the sampled epochs are different for different light curves. In that case, phase information may be lost, and the individual photometric measurements are rendered much less useful for the analysis (Che et al. 2018). Yet, due to technical and astronomical constraints, this is necessarily the case for ground-based astronomical time domain surveys (*e.g.* SDSS; York et al. 2000; ASASSN, Kochanek et al. 2017, etc.), and will be the case of upcoming surveys such as the LSST (Ivezić et al. 2019). Additionally, depending on the number of light curves to analyse and the specific features to be extracted, the process of feature engineering can be computationally expensive. It is therefore desirable to use feature extraction methods that do not require heavy processing.

It is common to use statistical parameters of the time-series as the set of features (Dubath et al. 2011; Richards et al. 2011; Nun et al. 2016; Johnston & Oluseyi 2017). A number of feature extraction packages are available for time-series. One such code is the Feature Analysis for Time-series (FATS) code (Nun et al. 2015), which is able to evaluate over 40 features in time-series with the epochs, magnitudes, magnitude errors, and filters as an input, and extracts statistical features such as means, standard deviations, linear trends, variability index, skewness, kurtosis, etc. (see also Malanchev et al. 2021). One issue with this approach of feature extraction is that not all of the features included are properly defined for all light curves, since there might be missing data, different number of elements, etc. As a result, these features may be undefined or unreliable for possibly a significant fraction of the light curves.

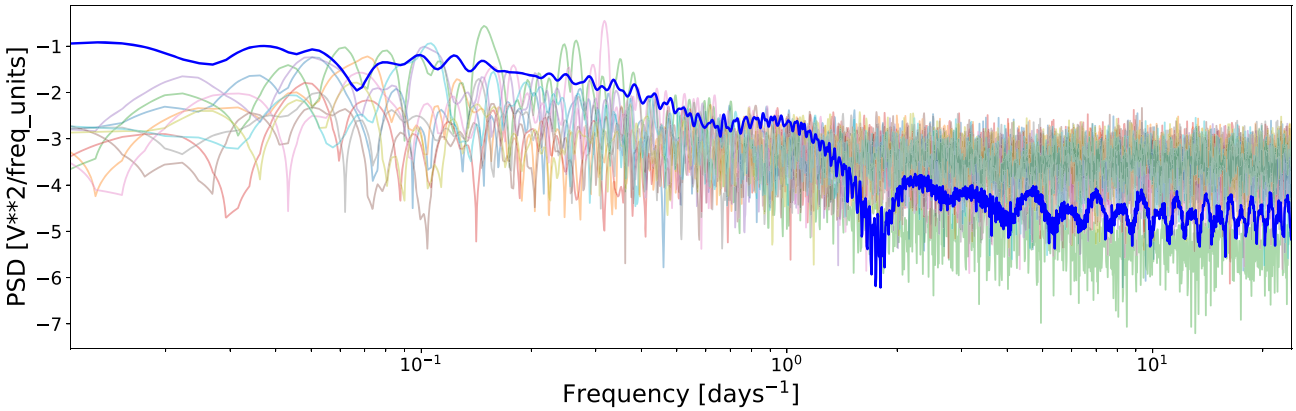


**Figure 2.** Example periodograms for 10 Kepler Q16 light curves, with Boyajian's star indicated in thick blue. Note the particular structure of Boyajian's periodogram, indicating the pronounced dips in the light curve. The logarithmic frequency range covers periods from 1 hr to about 90 d. All periodograms are obtained over the same array of frequencies.

Rather than relying on second-order features, it seems appropriate to remain as close as possible to the data themselves, therefore avoiding biases introduced by the extraction process. In this paper, we have chosen to use a combination of the light curve points themselves and their power spectrum as the input features for the anomaly detection, and an image representation of the light curves as the input for the dimensionality reduction algorithms. The latter method falls within the representation learning approach (Bengio, Courville & Vincent 2013), in which a different representation of the data (as opposed to engineered features) is used in order to extract useful information for classification and/or anomaly detection tasks. Representation learning has been used for astronomical analysis, for example in Jamal & Bloom (2020) and Szklenár et al. (2020) for classification purposes, but not as extensively for anomaly detection (but see Storey-Fisher et al. 2021, for the use of generative models for anomaly detection).

Using the power spectrum (or rather the periodogram since the data are unevenly sampled due to the presence of a gap in the data) is desirable especially for light curves that do not share a common phase reference. Periodogram analysis can also be used to find the best period for periodic variables. However, period-finding itself is a significant challenge (Graham et al. 2013; VanderPlas 2018), and therefore we do not use estimated periods, but the periodogram values themselves as the features.

### 2.2.1 *Periodograms*

The periodogram of a light curve is based on its Fourier transform and it measures the signal's power as a function of angular frequency. Lomb (1976) and Scargle (1982) generalized the concept of periodogram for the case of unevenly sampled data. Their normalized periodogram can be written as:

$$P_N(\omega) = \frac{1}{2\sigma_y}\left[\frac{\left[\sum_k (y_k - \hat{y})\cos\omega(t_k - \tau)\right]^2}{\sum_k \cos^2\omega(t_k - \tau)} + \frac{\left[\sum_k (y_k - \hat{y})\sin\omega(t_k - \tau)\right]^2}{\sum_k \sin^2\omega(t_k - \tau)}\right], \quad (1)$$

where $\sigma_y$ is the variance of the photometry $y_k$ and $\tau$ is a time offset that orthogonalizes the model and makes the expression independent on a time translation. As demonstrated in Lomb (1976), this expression is fundamentally equivalent to estimating the harmonic content given a least-squares fit to a sinusoidal model consisting of a single component. It can therefore easily be computed using a fast Fourier transform.

Given a light curve, we can evaluate the periodogram in equation (1) for a discrete number of frequencies, and use the values of the periodogram evaluated at each of these frequencies as features for our algorithm. This is conceptually similar to using the individual pixels in a spectrum as the features, except that using the same array of frequencies for all light curves, we can compare them using the same absolute reference. This approach can also be used for multiband non-simultaneous light curves, since a different periodogram can be calculated for each filter, and a final array of features can be obtained by concatenating the single-band periodograms.

For the dense light curves, we have extracted periodograms with 3000 single points covering a logarithmic range of frequencies corresponding to periods between 1 hr (twice the Kepler cadence) and 90 d (the approximate duration of the observations). This range does not only cover the periods sampled by the observations, but also the typical time-scales of different stellar variability phenomena. For example, an *HST* survey of the variability properties of luminous ($M_I < -5$) stars in M51 (Conroy et al. 2018) finds that the variability fraction for these is $\sim 50$ per cent, with many stars showing typical time-scales between 1 and 100 d. More in general, the most common pulsating variable stars have characteristic time-scales that range from a few minutes to a couple of years (Eyer & Mowlavi 2008). For the sparse light curves we construct periodograms of 300 points, covering a range of frequencies between 4 hr and 90 d. In Fig. 2 we show examples of the dense light curve periodograms computed for a sub-sample of relatively normal light curves in the Kepler data set, together with the periodogram of Boyajian's star, one of our bona-fide anomalies. Note that, with respect to the periodogram of Boyajian's star, 'normal' light curves have a more evenly distributed spectral power as a function of frequency.

In terms of transferability, we note that the approach adopted here requires the generation of the power spectrum prior to the anomaly detection analysis. The properties of this power spectrum will be affected by the cadence and total number of points in the light curve, which implies that, in principle, it would not be possible to transfer a model trained on *Kepler* data and apply it to a different survey, such as LSST. However, as long as all the cadence has been the same for all the objects belonging to a survey, the Lomb–Scargle algorithm provides a method to obtain a periodogram that is consistent for all objects in that survey. Our investigations on simualted LSST-like data sets such as the one resulting from the PLAsTiCC data challenge

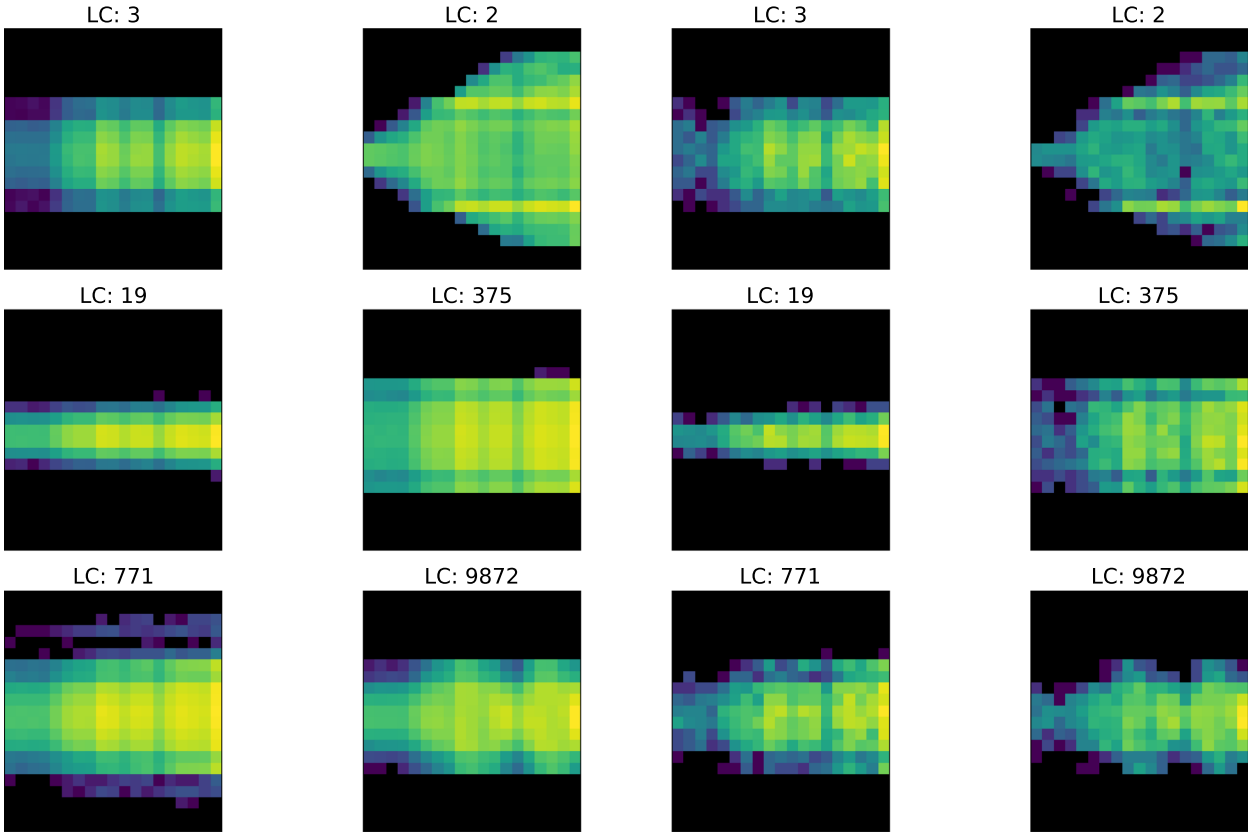


**Figure 3.** A plot of raw *DMDT* representations of dense (six left-hand panels) and sparse (six right-hand panels) light curves, for a random selection of six objects from our data set.

(Kessler et al. 2019) have shown that a similar approach as the one adopted here can be used to study the much more sparse and irregular time-series of ground-based surveys.

For anomaly detection with the tree-based method, we have constructed the feature vector for each light curve as the concatenation of the light curve points and the periodogram points. This yields a total of 6520 features for the dense light curves, and 632 features for the sparse light curves.

### 2.2.2 *DMDT maps*

One additional feature extraction method, the *DMDT* approach proposed in Mahabal et al. (2017), is adopted to obtain the input features for the manifold learning-based dimensionality reduction, and is described below.

Light curves are sequences of stellar brightness as a function of time, and the *DMDT* map is a light curve image representation that simply considers differences in magnitude — *DM* — against differences in time — *DT*. The values of magnitude differences and time differences are then binned, resulting in a 2D image representation of the light curve (we refer the reader to Mahabal et al. 2017, for more information on the specifics on the *DMDT* representation of light curves). The output feature space is a $\mathbb{R}^{21\times19}$ matrix (its dimension relates to the desired dimensions of the *DMDT* mapping) representing the pixel matrix of the *DMDT* images. We show examples of *DMDT* representations for dense and sparse light curves in Fig. 3.

After producing the *DMDT* map for each light curve, we then flatten the matrices into 399-long arrays, each array representing a light curve. We consider this our starting high dimensional space for the dimensionality reduction algorithms. While keeping the 2D information of the matrix would be ideal to reduce the amount of information lost, to the best of our knowledge, there are currently no codes available that are robust, tested, and scalable that can handle distances between 2D data sets. One possible approach is proposed in Johnston et al. (2019), where the authors represent light curves of variable stars as matrices of features, and then define a metric that defines distances directly between matrices. The authors acknowledge that the method requires some improvements before it can be fully exploited. We have therefore opted for flattening the matrices, as is common practice for the treatment of 2D images. Distances can be directly estimated in the multidimensional space of the vector elements.

## 2.3 The influence of pre-processing choices

Anomaly detection methodologies identify sources whose properties stand out with respect to all other objects in a data set. These sources might represent examples of unknown types of objects, in which case we may require the formulation of new physical hypotheses to explain their properties, or instances of rare evolutionary stages of known types. In both cases, they represent an expansion of our discovery space. But anomaly detection can also reveal instrumental or processing artefacts.

In the specific case of light curves, such artefacts can include spurious trends in the light-curve baselines or bad pixels values that result in artificial spikes or dips in the time-series. Therefore, in order to characterize the effect of data artefacts in our analysis, we explore

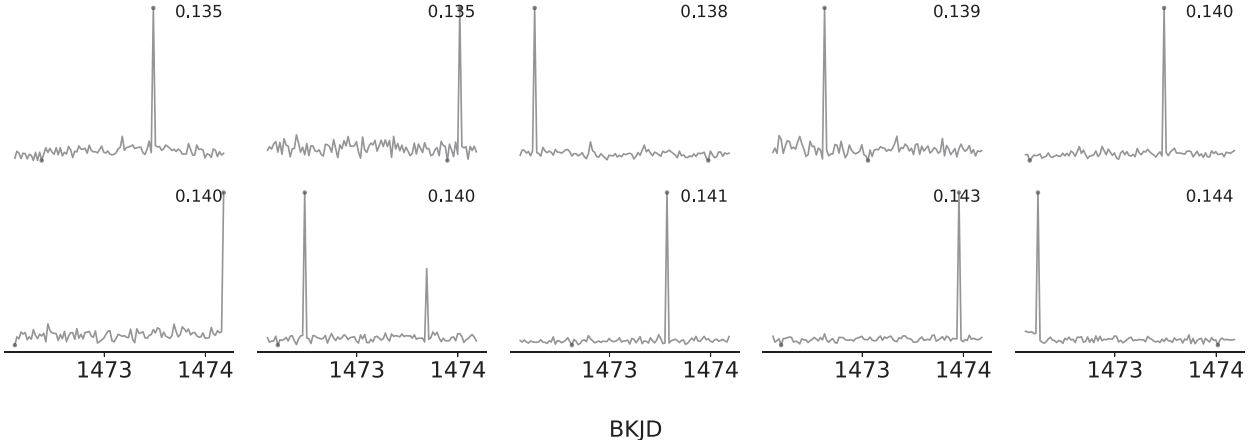


**Figure 4.** Normalized dense partial (5-d) Kepler Q16 light curves with corresponding normality scores derived from the IF method for 10 light curves with low anomaly score. The anomaly score is dominated by single-point events which may be due to artefacts or extremely high energy and short duration events, such as flares, which are rare, and therefore arguably should be contributing to the anomaly estimate. The brightest and dimmest point in each light curve are indicated by a blue point.

how different pre-processing approaches affect our ability to find anomalies.

We consider the results of applying an anomaly detection algorithm to our set of *normalized* light curves both before and after further pre-processing. For demonstration purposes, we use here the isolation forest (IF) method, which will be explained in detail in the next section (Section 3). For now, it is sufficient to say that the IF generates a score such that the lower the score value, the more unlike-the-others the object is, and, formally, having decided what percentage of objects are expected to be anomalous through the hyperparmeter `contamination`, anomalies are associated with negative scores. We note that the IF leads to the identification of similar anomalies as the URF methods (Section 3.1), which we extensively use in our analysis. By comparing the anomaly scores of the pre-processed light curves with those of the original light curves, we expect to learn to what extent pre-processing (or the lack of pre-processing) influences our ability to detect anomalies.

In Fig. 4 we show the 10 least anomalous light curves, according to a similarity score derived from the IF method. The fact that prominent spikes consisting of a single bright data point are present in all of them deserves some attention. These spikes can be due to events unrelated to the source (*e.g.* cosmic rays), or to instrumental artefacts, such as hot pixels. They could also be very high-energy astrophysical events such as star flares, but we note that these spikes are unlikely to be high-energy flares simply due to how commonly they appear in our data compared to flare statistics (see for example Paudel et al. 2019). These large features dominate the information-content in the normalized light curves, and therefore the similarity score of objects, potentially hiding less prominent, anomalous features. Here we are interested in finding unique anomalous light curves, so we cannot let either data artefacts or flares to dominate the anomaly score when the methods are applied to the normalized light curves.

We investigate the sensitivity of the anomaly score to pre-processing by removing these spikes using two different approaches and then comparing the anomaly scores obtained for the normalized light curves, and both sets of pre-processed light curves. The first spike-removal approach involves low-pass filtering: we take the light curve's rolling average over a window of 10 points, henceforth suppressing any high-frequency variations. We call these the low-pass filtered light curves set *LPlcv*'s. The second approach involves directly removing the outlying data points by replacing every point outside of a $3\sigma$ deviation range from the light curve mean with the corresponding value of a *LPlcv*'s generated with a window of size 15 data points (7.5 hr). We refer to this last set of pre-processed time-series as *Cleanlcv*'s.

The anomaly scores of the normalized light curves do not strongly correlate with the scores for either of the two pre-processing methods (left-hand and center panels, Fig. 5). The Pearson's $r$ correlation coefficients between each of the two sets of pre-processed light curves and the unprocessed version are $r \sim -0.10$ and $r \sim -0.16$: there is a weak *inverse* proportionality, weak but statistically significant at a $3\sigma$ level, for both *Cleanlcv* and *LPlcv* ($p$-value $< 0.01$). On the other hand, the two pre-processed light curve sets result in very similar anomaly scores (right-hand panel Fig. 5). In this case, the Pearson's $r$ value is 0.96. Although the IF reported 'anomaly' threshold, where the anomaly score becomes negative, is somewhat arbitrary it is worth noting that all scores are positives for the unprocessed light curve set, while the pre-processed light curves find anomalies, consistently with the implicit request set up in the choice of parameters (`contamination = 0.1`).

Altogether, this demonstrates that the single spikes dominate the model's decision when assigning an anomaly score, and may hide important anomalous behaviour of lesser flux amplitude by biasing the scores. Truly anomalous light curves (or at least not those with the most obvious data artefacts) are only found by these methods when such artefacts have been effectively removed and such features should be removed to reveal more subtle anomalies.

We further investigate the scores in order to understand why the presence of spikes is associated with a low anomaly score in the original light curve set. We look at the correlation of the anomaly scores for each pre-processing scheme with three simple light curve statistics: standard deviation ($\sigma_t$) of the original light curve, standard deviation of the low pass filter version $\sigma_{tLP}$, and normalized flux range over the standard deviation $R_{SNR} = \frac{max(\vec{t})-min(\vec{t})}{\sigma_t}$. This last statistics is significantly impacted by the presence of spikes. We find a strong correlation between the score obtained on the normalized time-series and $R_{SNR}$ (Pearsons $r \sim 0.5$, $p$-value $< 0.001$); see Fig. 6. Meanwhile, individually, neither the data range nor the standard deviation of the original time-series is a good predictor of the normality score ($r \sim 10^{-3}$ and 0.04, respectively). Conversely, we find a strong anticorrelation of the score of the pre-processed time-series, both *LPlcv* and *Cleanlcv*, with the standard deviation of the original light

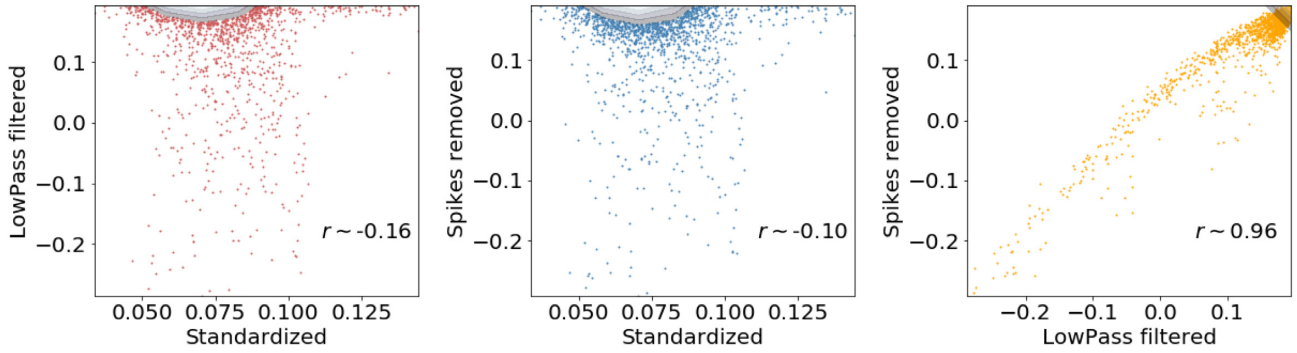


**Figure 5.** Comparison of anomaly scores produced by the IF method for light curves pre-processed differently: standardized, low pass filters, and cleaned by replacing $3\sigma$ outliers with a local mean value. The correlation in the scores generated for standardized light curves and light curves where sharp features are removed is low and inverse (Pearson's $r$ is reported in each panel) regardless of whether sharp features are low pass filtered or cleaned by replacing the values with a local mean. Conversely, the specific choice of pre-processing (by low-pass filtering or by replacement) has little influence on the anomaly score. (Regions of high point density are visualized as contours, instead of scatter points).

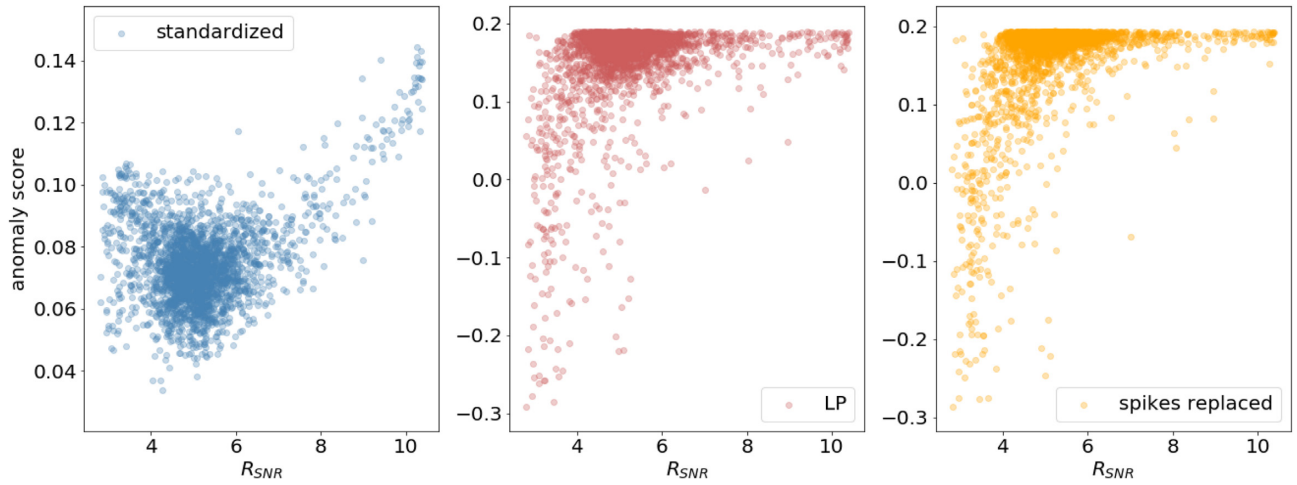


**Figure 6.** The SNR of the light curves, measured as $R_{SNR} = \frac{max(\vec{t})-min(\vec{t})}{\sigma_t}$, plotted against the anomaly score measured by the IF model for standardized light curves (left), light curves that have been low-pass filtered (*LPlcv*, center), and light curves were spikes are replaced by the value of the local mean (*Cleanlcv*, right). The dependency of the score on the $R_{SNR}$ is weak for the standardized light curve, and significant for the low-pass filtered and cleaned-by-replacement versions.

curve, and of the pre-processed versions (all $p$-values $< 0.001$), as well as with the $R_{SNR}$ metric; but for the latter, the distribution is far more evenly distributed, still with a significant lack of inliers at high $R_{SNR}$.

In the presence of light curves with significant spikes, therefore less anomalous objects are largely dominated by these artefacts, as they are not easily isolated in the multidimensional space of all light curve points. Once we remove them, however, the pattern of the least anomalous objects changes to contain those light curves with smaller flux variances and a white noise spectrum. More severe perturbations of the light curves are then associated with anomalies. Pre-processing of light curves in order to remove non-astrophysical artefacts is therefore a necessary step in anomaly detection.

For the remainder of this paper, we use light curves pre-processed according to the *Cleanlcv* method. As illustrated in Fig. 7, this method removes the undesired spike artefacts without significantly affecting the frequency structure of the original data in other segments of the light curve. Only a very small fraction of true (non-spike) light curve points are affected by the $3\sigma$ threshold that we have imposed, and when they are affected, the effect is not dramatic, because most of the astrophysical dips and flares last for a few hours at least. The main effect of this pre-processing approach is that dips that deviate significantly from the mean value of the light curve are slightly less pronounced compared to the original light curve, as a result of the smoothing. We note that a similar normalization and spike-removal approach has been used in previous work as a pre-processing step for anomaly detection (Rebbapragada et al. 2009).

The results in this section are reproducible and the code can be found in the `Github` repository that we have provided.

## 3 METHODS

In this section we introduce the anomaly detection and dimensionality reduction methods employed in our analysis. These methods are used with different purposes along the paper. The unsupervised RF method is used to estimate anomaly scores for the entire set of objects. The IF method is used (see previous section) to support our analysis on pre-processing. Finally, the manifold methods ($t$-SNE and UMAP) are used to reduce the dimensionality of the light curves and to group similar anomalies. For the latter, the addition of an Euclidean metric in the low-dimensionality space

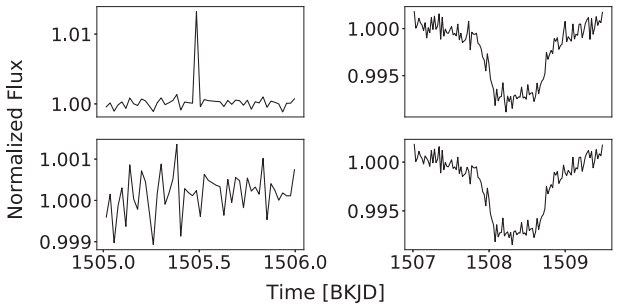


**Figure 7.** Two light curve segments showing the effect of the *Cleanlcv* pre-processing approach. The top panels show the original light curve points, whereas the bottom panels show the processed light curves. We note that the spike, potentially an artefact, seen in the top left-hand panel is removed by our method, while the light curve remains unaffected for real astrophysical features, such as the transit-like dip seen in the right-hand panels.

can be used as an additional anomaly measure, as explored in Section 4.3.2.

### 3.1 Unsupervised random forest

In order to understand how *similarity* is defined in random forest (RF) methods, we ought to briefly describe the basics of their functionality as classifiers. An RF is an ensemble of decision trees. Each of these trees is a model in which the final prediction is based on a series of comparisons of the values of the predictors (features) against threshold values. Every tree therefore corresponds to a partition of the feature space by axis-aligned boundaries, where the class of each final partition is given by the class of the majority of objects in that partition. These methods are inherently 'greedy' and cannot exhaustively explore the parameter space (consider for example that a single continuous variable offers an infinite number of binary splits). As a result, decision trees have high variance (different trees give a different results) and are subject to overfitting. Ensemble approaches reduce the variance of these methods: each tree in an RF uses a random subset of objects from the training set and a subset of features. The forest then predicts the class of an object as the majority-vote prediction of all the trees in the ensemble.

The URF method for anomaly detection was first introduced by Shi & Horvath (2006) in the context of tumor discovery using data sets comprising tumor marker expressions and has been adapted for the first time in astrophysics to look for anomalous objects in a large (∼2 million sources) data set of SDSS spectra (Baron & Poznanski 2017a,b; Reis et al. 2018). URF is an adaptation of RF to unsupervised learning. The method works in two stages: the training of a supervised classifier, and the generation of an anomaly metric obtained by propagating the data set of interest through the trained classifier. The classifier is trained to distinguish between members of the original data set, and members of a *synthetic* data set that is generated by sampling the marginal distributions from that original data set for each of the features independently. That is, the synthetic data set is identical to the original data set in its marginal distributions, but it lacks the correlations between features that are present in the original data. The labels for the classifier are thus either *original* or *synthetic*, and the training set is the entire original data set augmented with the synthetic data set. The training is performed using all objects in the data set for which an anomaly score wants to be determined.

The anomaly score is then computed by propagating the original data set through the trained classification model. The similarity score between two objects is measured as a normalized count of individual trees in which two given objects ended up in the same leaf node and with the same class. The *weirdness* or anomaly score of an object can then be thought of as the average of the pair-wise dissimilarity between that particular object and all the other objects in the data set. Note that this dissimilarity measure is not necessarily associated to distance in an Euclidean space, because objects can be isolated in only a few of the many possible dimensions of the features space, and also because of the random nature of the forest.

Formally, the similarity between the $i$-th and $j$-th objects in the data set can be calculated as:

$$D_{ij} = 1 - N_{\rm leaf}/N_{\rm tree}, \tag{2}$$

where $N_{\rm leaf}$ is the number of trees for which the $ij$ pair were both classified as *real* in the same leaf node, and $N_{\rm tree}$ is the total number of trees. The *weirdness* of object $i$ is then the average value of $D_{ij}$ for all possible pairings of $i$ with all other $j$ objects in the data set.[1]

In this work, we implement the URF using the *scikit-learn* (Pedregosa et al. 2011) RF architecture, following the general recipe described in Baron & Poznanski (2017b). The URF has several important hyper-parameters that require fine-tuning in order to avoid overfitting (and which of course depend on the number of $n_{\rm feat}$ features of each ligth curve; see Section 2.2). The most important hyperparameters are the number of trees in the forest ($N_{\rm tree}$); the maximum depth of the tree (the length of the longest path from the root of the tree to a leaf), `max_depth`; and the maximum numbers of features that are considered when looking for the best split, `max_features`. We performed the tuning of these parameters through a validation process that involved a gridsearch and an 80–20 training-test split. For the light curves and features described in Section 2, we found that achieving an optimal non-overfitting validation accuracy for the classifier also results in a well-defined peak of anomalous scores in the distribution of the URF scores. That is, if the classifier is less accurate, the peak of anomalies defined by the distribution of URF scores is less distinct from the less anomalous objects. The desired levels of accuracy are achieved with $N_{\rm tree} = 700$, `max_depth` $= 100$, and `max_features` $= \log_2(n_{\rm feat})$ in the case of the full dense light curves curves, and $N_{\rm tree} = 150$, `max_depth = default`, and `max_features` $= \sqrt{n_{\rm feat}}$ in the case of the sparse light curves.

#### 3.1.1 Isolation Forest

A related method is the IF that we have used in Section 2.3 to evaluate the effect of pre-processing of the data. IFs are ensemble methods based on Isolation Trees (Liu, Ting & Zhou 2012). In this method, which comprises a forest of isolation trees, each tree is partitioning a data set based on randomly selected features and randomly selected splitting points for each feature. The anomaly score is proportional to the number of random splits required to isolate an object (smaller score indicates more anomalous objects, negative scores indicate 'outliers'), averaged over all trees in the forest. Anomalies require fewer partitions, *i.e.* they are easy to isolate. The average number of partitions over a large number of random trees can therefore be considered as a measure of similarity to the bulk of the data. For the pre-processing analysis of Section 2.3, we use the Pedregosa

[1]For a graphical demonstration of how weirdness is defined in this algorithm, we refer the reader to the `GitHub` repository associated to Baron & Poznanski (2017b): https://github.com/dalya/WeirdestGalaxies/blob/master/outlier_detection_RF_demo.ipynb

et al. (2011) implementation of this method (see also Buitinck et al. 2013) and we embrace the default parameters: the number of samples used by each tree is set to `max_samples` $= 2^8$ and the number of trees to 100. The 'contamination' parameter sets the expectation for the fraction of objects in the set that are outliers, and it is set to `contamination = 0.1`. These parameters were demonstrated in Liu et al. (2012) to be effective under a large range of circumstances. The IF anomaly score is intuitively interpretable. For this, we chose it to guide our pre-processing choices. Yet IF is an extremely powerful methods for anomaly detection in spite of its simplicity (see for example the benchmarking study Emmott et al. 2013).

### 3.2 Manifold learning methods

We now describe two related manifold learning methods: *t*-SNE and UMAP. The methods are designed to visualize high-dimensional data in a lower typically two-dimensional space where reciprocal distances in the high dimensional space are (optimally) preserved.

The resulting embeddings can also be used for anomaly detection. As we detail in Section 4.3, an anomaly score can be defined by looking at the distribution of distances to the nearest neighbour in the converged distribution for both *t*-SNE and UMAP methods. This Euclidean anomaly score, which selects a very specific type of anomalies, can be used complementary to the score derived from the URF algorithm.

#### *3.2.1* t*-distributed Stochastic Neighbour Embedding*

The *t*-distributed Stochastic Neighbour Embedding (SNE) method, or *t*-SNE, was introduced in Maaten & Hinton (2008), improving upon the well known non-linear dimensionality reduction algorithm SNE (Hinton & Roweis 2003). SNE works by embedding multidimensional Euclidean distances with conditional probabilities, which is what represents the similarities between data points. In other words, suppose we have a data point $x_i$ in the high dimensional space. Then consider a normal distribution of distances from $x_i$, wherein points near $x_i$ have a higher probability density under the distribution and further points have a lower probability density under the distribution. Then the similarity between $x_i$ and another data point $x_{i'}$ is the conditional probability $P_{x_{i'}|x_i}$ that $x_i$ will choose $x_{i'}$ as a neighbour under the normal distribution just described.

Then we replicate the process for the lower dimensional space, for which we get another set of conditional probabilities. SNE then attempts to minimize the Kullback–Leibler (KL) divergence (or relative entropy; Kullback & Leibler 1951) between the two probability distributions using gradient descent. However, SNE is computationally very expensive, largely because of the asymmetry imparted by the use of KL divergence as the distance metric; *t*-SNE attempts to resolve this issue by looking at a 'symmetric' SNE, specifically a symmetric version of the cost function with similarly simple gradients. *t*-SNE also redefines the lower dimensional distribution using a student *t*-distribution in place of the Gaussian distribution to solve the crowding problem, which stems from the fact that there is not enough area in a two-dimensional plot to accurately embed distances between points that are close, which leads to loss of information.

Here we use the Python `scikitlearn` implementation of *t*-SNE with hyperparameters (`n_components = 2`, `perplexity = 200`, `learning_rate = 50.0`, `early_exaggeration = 5.0`) for dense light curves, and similar hyperparameters for sparse light curves (with the only difference being `early_exaggeration = 20.0`); `n_components` represent the dimension of the space we want to map into. `Perplexity` is related to the number of neighbours to be considered when considering a certain data point (described in McInnes et al. 2018), defining a notion of similarity; `learning_rate` affects the gradient descent portion of the tSNE algorithm, with too fast a learning rates resulting in ball like clusters where neighbours are equidistant, and too slow a learning rate creating dense cloud clusters; `early_exaggeration` relates to how tightly points are clumped in the embedding space so that we can control the visualization of the high dimensional data (this affects all points similarly, so it mostly impacts visualization, and not the overall similarity results from the algorithm). We ranged the values of the three hyperparameters over different experiments, recording the KL divergence of each model after training on the full data set. We then set the hyperparameters based on the model with the lowest KL divergence.

#### *3.2.2 Uniform Manifold Approximation and Projection*

Uniform Manifold Approximation and Projection (UMAP) is a non-linear dimensionality reduction technique introduced very recently in McInnes et al. (2018). Most dimensionality reduction techniques have a very similar structure: they aim to find some low dimensional representation of data that minimizes information loss between the same representation applied to the high dimensional data set. UMAP works as follows: consider data points $x_1, \ldots, x_n$. We then create a $k$-neighbour weighted graph by considering $k$-neighbours of each $x_i$, and adding an edge in the graph with a defined weight $w$ that depends on the diameter of the $k$-neighbourhood of $x_i$, and the distance between $x_i$ and the closest neighbour. Note that the weight, as defined in McInnes et al. (2018), is not symmetric. We handle this by, given $a = w(x_i, x_j)$, $b = w(x_j, x_i)$, defining a new weight $w'(x_i, x_j) = a + b - ab$. Then the same process is repeated in the lower dimensional space, resulting in a new weight function for the lower dimensional space. Then UMAP minimizes the cross-entropy between the two weight functions as specified by the cost function so that the lower dimensional weights encapsulate (as closely as possible) the information from the higher dimensional weights. Like in *t*-SNE, this optimization is done via stochastic gradient descent.

Here we use the UMAP implementation provided in McInnes et al. (2018), with hyperparameters (`n_neighbours = 200`, `min_dist = 0.4`, `learning_rate = 0.25`) for dense light curves. We used hyperparameters (`n_neighbours = 200`, `min_dist` $= 0.1$, `learning_rate` $= 0.8$) for sparse light curves. `n_neighbours` is similar to `perplexity` in the *t*-SNE algorithm, controlling the number of neighbours to be considered when defining the weights between data points in the weighted graph. `min_dist` is similar to the `early_exaggeration` parameter in the *t*-SNE algorithm, and controls how tightly points are clumped in the embedding space (to create more interpretable visualizations). Similar to the *t*-SNE method, we vary the hyperparameters for the model, and pick the model that gives the most stable embeddings.

### 3.3 Processing requirements and scalability

We now describe the computational performance of our algorithms in order to provide some insight as to how our method scales to even larger data sets. Overall, processing times for the number of objects and the number of features involved in this project are not

computationally prohibitive. The URF anomaly detection method took a few hours to process the entire set of 147 036 dense light curves, each containing about 6300 features, using a 2.4 GHz 8-Core Intel Core i9 Processor, typical of a Mac laptop. One advantage of ensemble methods such as the URF is that trees can be trained independently, and therefore parallelization is possible. The URF computing time is linear in the number of features and in the number of trees. The generation of the similarity matrix for the calculation of the anomaly score requires a total of $n_{\rm obects} \times n_{\rm obects\text{-}1}$ simple arithmetic operations, and it therefore scales roughly as $n^2_{\rm objects}$.

The machine that ran the *t*-SNE/UMAP outlier pipeline had a GNU/Linux OS, x86-64 architecture (Linux kernel version 3.10.0-957.el7.x86-64) with 31 GB of free memory. DMDT generation for dense light curves took around 36.76 hr, and DMDT generation for sparse light curves took around 0.23 hr. Generating TSNE embeddings for dense light curves took about 1.04 hr and UMAP embeddings for dense light curves took around 2.36 hr. Generating *t*-SNE/UMAP embeddings for sparse light curves took around 2 hr. Finding outliers given a specific embedding took around 1 hr. In terms of storage, the dense and the sparse DMDT data sets occupy ∼450 MB each. Storing the *t*-SNE/UMAP embeddings for dense and sparse light curves took around 2MB each. No batch processing was therefore required for our largest data set.

## 4 RESULTS

This section summarizes the results of implementing our method for the identification of anomalies in the Kepler data set, and for the identification of analogues to those light curves. We describe the list of *bona-fide* anomalies that we have used as a ground truth to evaluate the performance of the method. We first apply the URF method to define a list of anomalies, and several diagnostics of their anomalous behaviour. We then investigate how the 2D embeddings from the DMDT maps can be used to identify objects that are analogous to those identified with URF as anomalies. In the last subsection we examine some astrophysical implications of our findings.

### 4.1 *Bona-fide* anomalies

To ensure our methods are effective, we use a control set of known anomalous objects as a first test of our method: rare objects should have a high anomalous score, as measured by our algorithm. In order to construct this set, we have relied on the work of (Debosscher et al. 2007), who performed a comprehensive census of 35 variability classes and identified members of each class via an exhaustive literature search that they later used to inform different supervised classifiers. Among the uncommon behaviours identified, they find:

(i) Ellipsoidal variables
(ii) $\gamma$-Doradus stars
(iii) Slowly pulsating B stars
(iv) RR Lyrae stars
(v) RV-Tauri stars
(vi) Classical cepheids

We constructed a list of 150 objects containing members of each of these classes. While these classes are not necessarily rare in the overall Galactic field population, stellar variability is in fact rare in the *Kepler* Input Catalogue (KIC). Variability classes such as classical Cepheids, ellipsoidal variables, $\gamma$-Doradus stars, and pulsating B stars are have less than 1000 identified members in the KIC, out of a total of about 13 million targets. RR Lyrae stars and RV-Tauri stars are even rarer, with probably less than 100 identified objects. Even eclipsing binaries are below the 10 000 limit. For the purpose of this paper, thus, stellar variability is 'anomalous', and our algorithm should also serve as a variability detector.

To the group of *bona-fide* variables, we have added Boyajian's star, which shows a truly unique behaviour, unlike any other objects in the aforementioned classes. Some members of these classes, and Boyajian's star itself, have been successfully identified by existing anomaly detection methods, (notably in Giles & Walkowicz 2020). However, no established mechanism exists to find analogues to those objects in a given data set.

We present the full list of *bona-fide* anomalies in online Table A1.

### 4.2 Finding anomalies with URF.

To find anomalous light curves, we apply the URF algorithm with the hyper-parameters described in Section 3.1 to the set of features composed of the array of the pre-processed (normalized, spike-removed) light curve points concatenated with the periodogram points for each of the 147 036 light curves in our data set. The URF results are reproducible and presented at the `GitHub` repository that we have provided. The total number of features used for the dense light curves was 6520, whereas for the sparse light curves it was 632. The hyper-parameters of the URF were tuned to maximize accuracy during the classification step as described in Section 3. We compute the URF weirdness scores for each light curve, and for both the dense and sparse data sets. In this section we present the resulting anomaly scores, highlight the important features for anomaly detection, and specifically search for the *bona-fide* anomalies in the ranked list of weirdness scores, to evaluate the performance of our method in finding light curves of interest. We also provide a list of new anomalies identified with our method.

#### *4.2.1 The identification of anomalous objects*

In Fig. 8 (top panel), we show the histogram of URF weirdness scores for all the 147 036 dense light curves. Note that the absolute range of URF score is not *per se* informative of the differences between objects, as it depends on the total number of trees and depth of the RF, and we have therefore re-scaled the URF scores to the range (0, 1). The relative differences between objects are informative. We identify at least three groups of objects based solely on the values of the URF scores. There is a main core of 'normal' objects, with normalized anomaly score centred at around 0.25, which includes about 60 per cent of all objects. Then there is a clear peak of objects that represent 18 per cent of the total, and that all have scores larger than about 0.85. We will take these to be anomalies. Finally, there is an excess of objects with scores centred at around 0.55, near the center of the range. This excess of objects, as we will see later in Section 5, mostly corresponds to members of the 'red clump', a group of red giants that are slightly hotter than other red-giant-branch stars of the same luminosity, and have a degenerate helium core (Girardi 1999).

Based on the distribution of the anomaly scores, we select a weirdness score of 0.85 as the URF anomaly threshold hereafter. At this score the distribution in the top panel of Fig. 8 departs from a decreasing behaviour and starts increasing, signaling an inflexion point beyond which there is an excess of objects with high anomaly scores. We note that the selection of this threshold is informed by the distribution, but it is to some extent arbitrary, because the transition between normal and anomalous objects is a continuous, rather than a discontinuity.

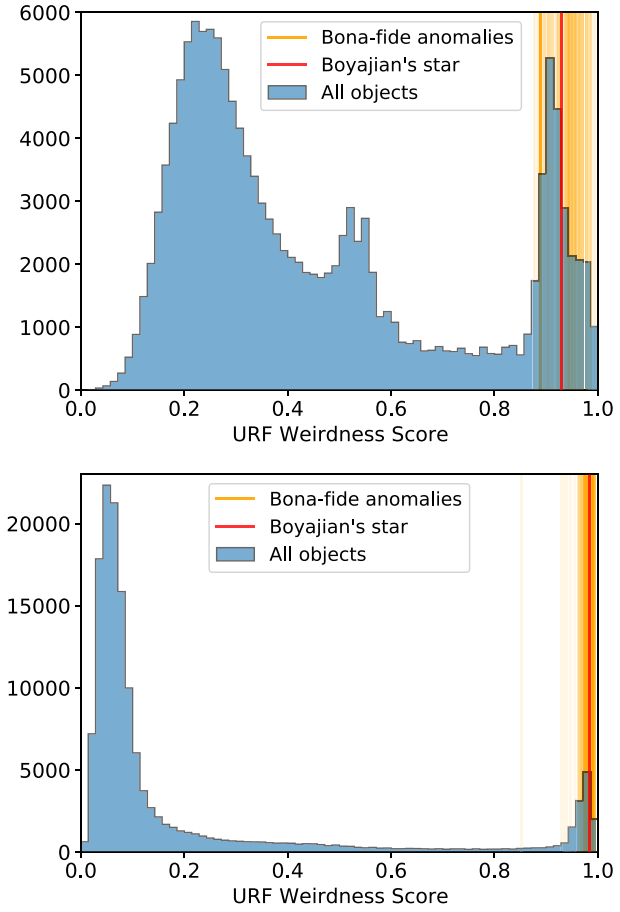


**Figure 8.** Histograms of URF weirdness scores measured for dense (top) and sparse (bottom) Kepler Q16 dense light curves. Indicated in orange and red are the scores of the *bona-fide* anomalies and Boyajian's star, respectively.

We also indicate in Fig. 8 the URF weirdness scores for all the *bona-fide* anomalies of online Table A1, shown as vertical orange lines. We observe that all of the anomalous objects of interest fall within the anomalous peak, *i.e.* they are all in the top 18 per cent of the weirdness scores (Boyajian's star is in the top 7 per cent). We determined the uncertainty in the anomaly score by running the full algorithm 10 times for a subset of about 20 000 objects, and measuring the standard deviation of the anomaly score for each object across these 10 runs. We find that the typical uncertainty depends on the mean value of the score, with high anomaly scores having smaller standard deviations. Specifically, the URF score is uncertain at the level of a few per cent (2 per cent-3 per cent) for objects with mean URF score above 0.85, and at the level of about 10 per cent for objects in the low end of the URF score distribution. The extreme low-score end of the distribution is populated by objects whose features are consistent with being sampled from the distribution that describes the bulk of the objects, specifically near its mean. In this respect, they are unremarkable as anomalies. One example of such objects is KIC 8211660 (URF score 0.0015), which is shown as a black line in Figs 9–11.

As discussed above, the *bona-fide* anomalious objects that we just demonstrated to be able to recover represent a few different astrophysical phenomena with different photometric properties. The anomaly scores are loosely related with 'anomaly type', but there is a significant overlap between classes. We can quantify this by looking at the average scores of anomalies for which an independent classification is available in the SIMBAD database. For example, among the most anomalous objects are $\delta$-Scuti stars, for which the mode of the URF score is 0.99 (top 0.6 per cent). Eruptive stars have a similar mode, but have a larger fraction of members with much lower anomaly scores. Eclipsing binaries have a score mode of 0.975 (top 1.6 per cent). $\gamma$-Dor stars have a score mode of 0.96 (top 4 per cent). Rotating variables have a score mode of 0.91 (top 8 per cent), and represent the vast majority of the high anomaly score peak. Long period variables have a score mode of 0.88 (top 17 per cent), and so on. Other types of object are also found (in smaller numbers) to have high score, including RRLyrae, cataclysmic variables, Cepheids, White Dwarfs, ellipsoidal variables, Mira variables, as well as active galactic nuclei, and quasars. Remarkably, a fraction as large as $\sim$ 4 per cent of all the Q16 targets, or about 5000 objects, are both unclassified and within the anomalous peak.

We show examples of light curves found by our method to have a high anomaly score in Figs 9–11. Fig. 9 shows known variability types (we remind the reader that these types are rare in the KIC), such as eclipsing binaries, long period variables, and $\delta$-Scuti stars. They all show significant amplitude variations with respect to the light curves of objects with low anomaly score (we have plotted the light curve of one of these normal objects, KIC 8211660, as a dark line for reference). They also have periodic or semiperiodic behaviours. Fig. 10 shows light curves of objects of known type that achieve a high URF anomaly score, but that belong to classes that are even rarer in our data set. These include RR Lyrae stars (only a handful in the entire data set), ellipsoidal variables, and Mira stars. Just as in the first group, these anomalies have wide amplitude variations and a tendency for periodic behaviour. Finally, in Fig. 11 we show objects found by our method to be anomalous and that are not classified in the SIMBAD database. Some of them appear in the literature as candidate dwarf novae, extrasolar planetary transits, and variable RGB stars, but the majority remain unclassified. Boyajian's star appears in this group, and it is shown in the center panel.

The rich set of different anomaly types that we are able to identify indicates that the URF algorithm applied to the set of features constructed from the light curve points and their periodograms is an effective approach to identifying time-domain anomalies of varying astrophysical nature, including paradigm-changing objects such as Boyajian's star, but also a wide range of variability types that depart from the behaviour of the majority of Kepler's sources. The complexity of the light curves result in significant dispersion and considerable overlaps between different groups in the distribution of URF scores. This implies that, given an unclassified anomaly, we cannot associate it with a specific class, or tell whether it belongs to a potentially novel class of objects, based on the URF score alone. In Section 4.3 we propose a method to identify analogues.

We provide the full list of unclassified anomalies found with our method in the electronic version of the paper, together with their URF anomaly scores, low-dimensional manifold embedding features, and where available, also *Gaia* absolute magnitudes and colours.

### 4.2.2 *Sparse light curves*

Can we still identify anomalous objects when the amount of information is reduced? Time-domain surveys do not generally result in regular and well sampled light curves, like Kepler's. In fact,

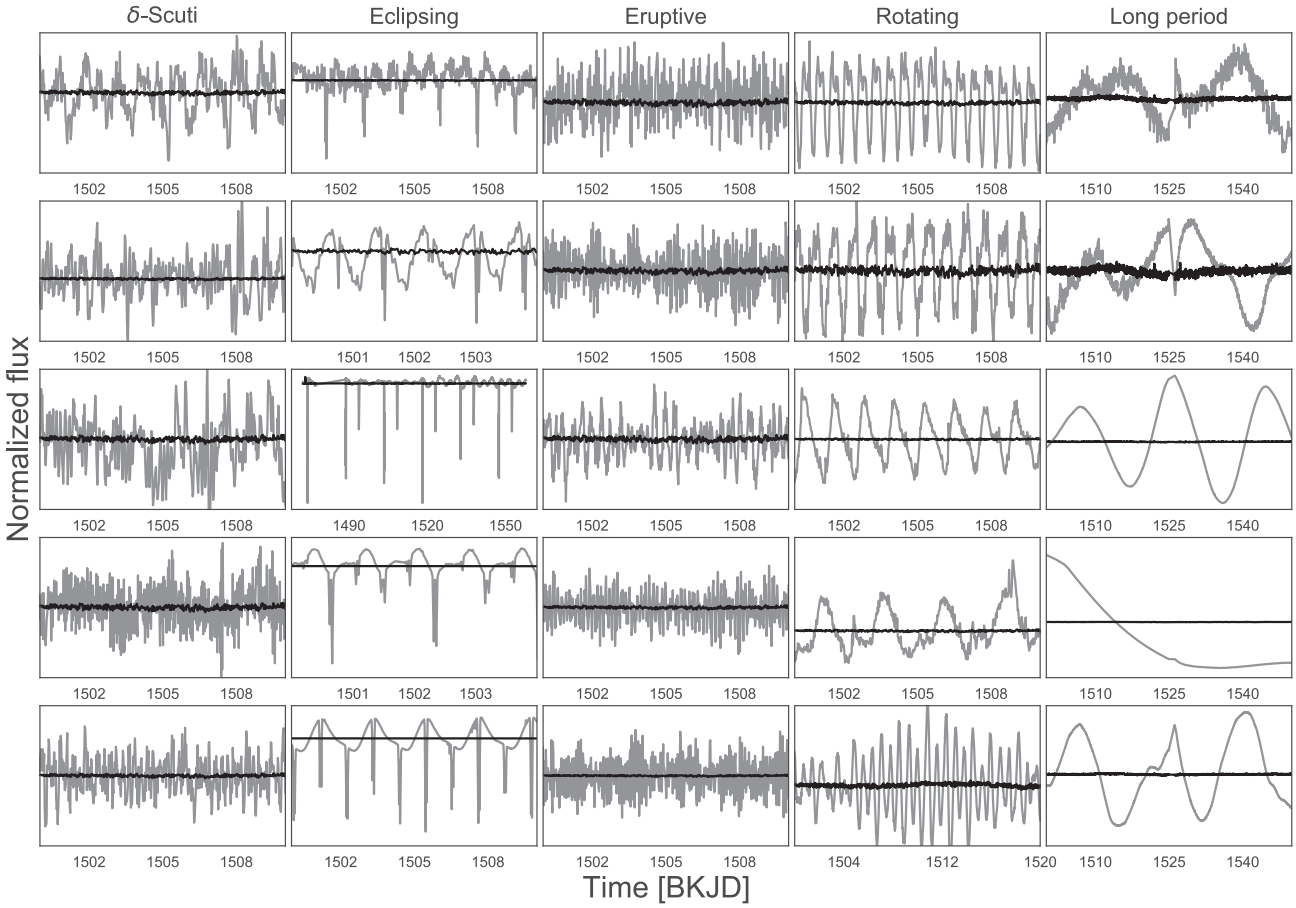


**Figure 9.** A selection of URF anomalous light curves with an associated common class. Note that the ranges in both axes change from light curve to light curve. For comparison, each light curve is shown together with the light curve of low-URF score object KIC 8211660 (black line).

ground-based surveys, including the upcoming LSST, will have significantly more irregular and sparser cadences when compared to the data set studied here so far.

We have applied our method to the set of sparse light curves generated by randomly sampling 10 per cent of the original light curve points. Just as with the original data set, we have generated the periodograms by applying the Lomb-Scargle method to the sparse light curves, and concatenated them to the data arrays to form our feature vector. We have then adjusted the hyper-parameters of the URF algorithm to maximize accuracy in the classification step as described in Section 3.1.

The resulting histogram of URF weirdness scores is shown in the bottom panel of Fig. 8. The sparser light curves have resulted in a much less structured distribution of the scores, but the general result holds: most of the *bona-fide* anomalies, Boyajian's star included, remain within the peak of anomalous object, which is now much more separated from all the rest of the objects and only includes 8 per cent of all objects (the anomaly peak in the case of the full-sampled light curves contained 18 per cent). This suggests that information about the anomalous nature of objects is mostly contained at low frequencies. However we do not identify as anomalous all the original anomalies.

What information are we missing in the sparse light curves for the purposes of anomaly identification? One would be tempted to infer that the information lost is related to high-frequency oscillatory modes, to which we have lost sensitivity when under-sampling the original data set. But the difference in the high frequency cuts of the periodograms for dense and sparse light curves is not very significant (1 hr versus 4 hr), and most of the *bona-fide* anomalies have characteristics time-scales larger than a few hours. In fact, what we are losing by only considering the sparse set is the ability to discern significant large-scale differences in the distribution of spectral power between high frequencies and low frequencies.

To illustrate this, in Fig. 12 we show the light curves (both dense and sparse) and the corresponding periodograms for two rotating variable stars with high anomaly scores. One of the two (KIC 9096191) is identified as an anomaly in both the dense and sparse data sets, whereas the other (KIC 6266324) is only identified in the dense data set. We argue that the reason for this difference is that any relative low-frequency power excess in the power spectrum is damped by the loss of data points (*i.e.*the overall shape of the power spectrum is flatter for the sparse light curves).[2] This implies that anomalies selected mainly on the basis of an uneven spectral power distribution will only be spotted when the dense light curves are used. Some, of course, might still be selected as anomalies in the sparse light curve if, for example, the amplitude of the variability is unusually large, as is the case for KIC 9096191.

A more general question is: how sparse can the cadence of a light curve get before one loses the ability to detect anomalies entirely? In Fig. 13 we show the fraction of the 100 most anomalous

[2]In noisy light curves of fainter or farther away objects, this excess is less pronounced because noise adds power primarily in high frequencies.

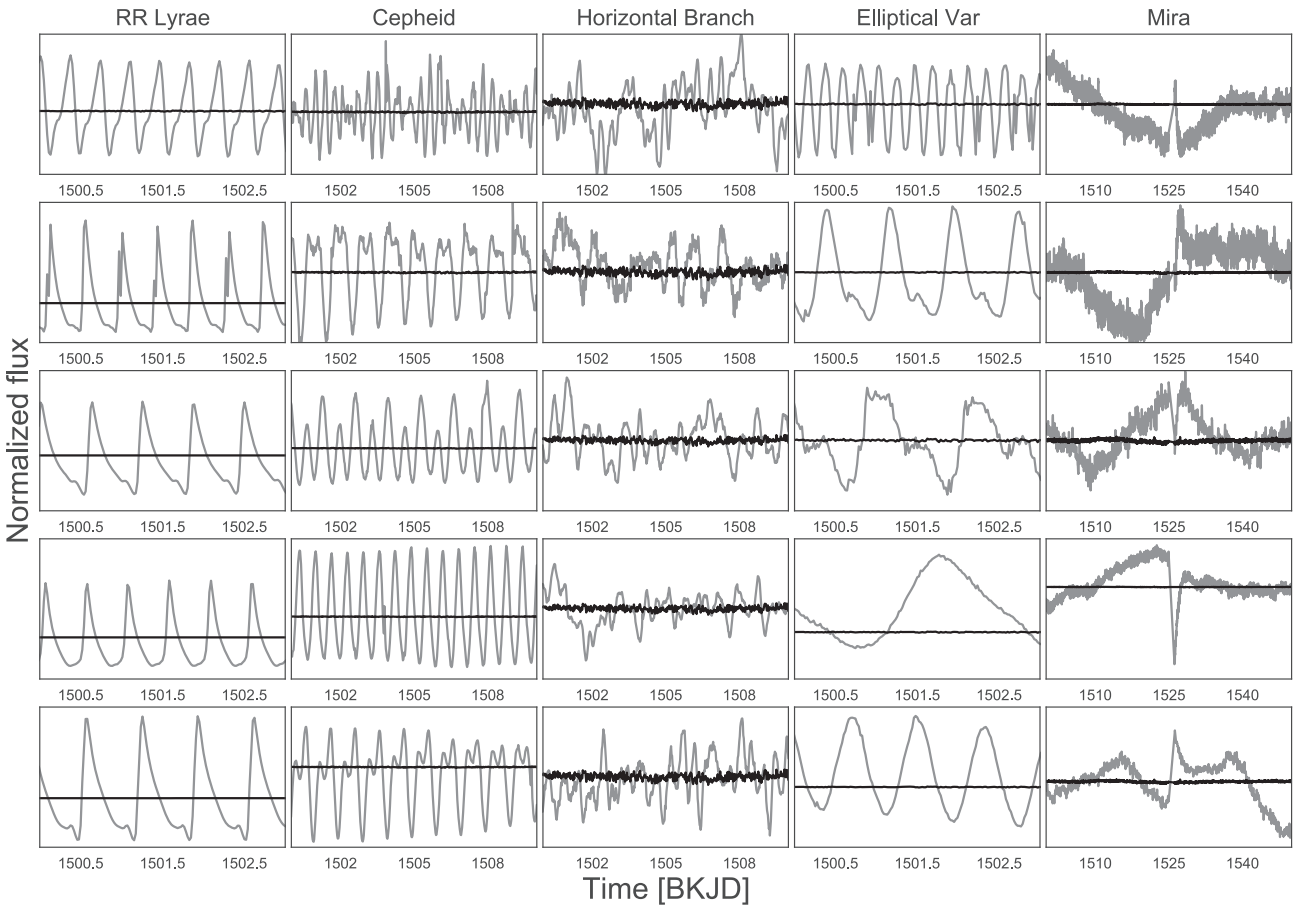


**Figure 10.** A selection of URF anomalous light curves associated with less common know classes. Note that the ranges in both axes change from light curve to light curve. For comparison, each light curve is shown together with the light curve of low-URF score object KIC 8211660 (black line).

objects found using the dense light curves that are still identified as anomalous in sparse light curves, for increasing separation between light curve points, for sub-sample of 2500 Q16 light curves. The average separation between points needs to increase by a factor of about 5 before the recovery rate drops below 80 per cent. This means that we could decrease the *Kepler* Q16 cadence from 30 min to 2.5 hr and we would still detect 80 per cent of the anomalies that we have found using the full light curves.

### *4.2.3 What drives the anomalous nature of light curves?*

We now turn to the question: what makes a Kepler light curve anomalous? In this section we provide an answer in terms of the algorithm used and the features that were fed to it in computing the anomaly scores.

In Fig. 14 we show the distribution of feature importances derived during training of the URF classifier step. To estimate the importance of each feature, we have used the `feature_importances_` attribute of the random forest classifier in `scikit-learn`, which assigns importances to features during training according to the accumulated decrease in impurity every time a feature is used in splitting. Important features are therefore those that, when used for splitting, result in a maximal isolation (less impurity) of anomalies. Leftwards of the vertical lines are the features associated with light curve points, rightwards of the line are the features associated with spectral power at different frequencies. There is relatively little variation in importance among features that represent light curve points, although there is a small increase of importance in the later half of the light curve. In fact, this upward trend follows quite closely the behaviour of the *average* light curve for the entire Kepler quarter, and might be the result of a residual in the de-trending process. The dip near the mid-point of the light curve features is related to a discontinuity due to lack of data around BKJD 1525. Something similar happens between BKJD 1477 and BKJD 1487. In general, light curve points are important in the classification step when they represent significant deviations from the mean flux.

In the features associated with the power spectrum (rightward of the vertical line in Fig. 14) a more clear trend emerges. Spectral power values that contribute the most to the classification step, and therefore to the anomaly score, are either shorter than a few hours, or longer than about 10 d. The classification and anomaly detection are thus mostly affected by characteristic variability time-scales of hours or weeks, whereas days-long characteristic time-scales are less relevant for distinguishing between objects.

We argue that features that are important in classification step of the URF (which, as a reminder, is the step in which the model learns to discriminate between true and synthetic data, see Section 3) also dominate the anomaly detection. For example, anomalous light curves should have more points that deviate from the mean with respect to more common light curves. Similarly, anomalous light curves should have significant differences in the spectral power at those frequencies that were relevant to perform the classification during the training phase. This would explain why pulsating stars with short

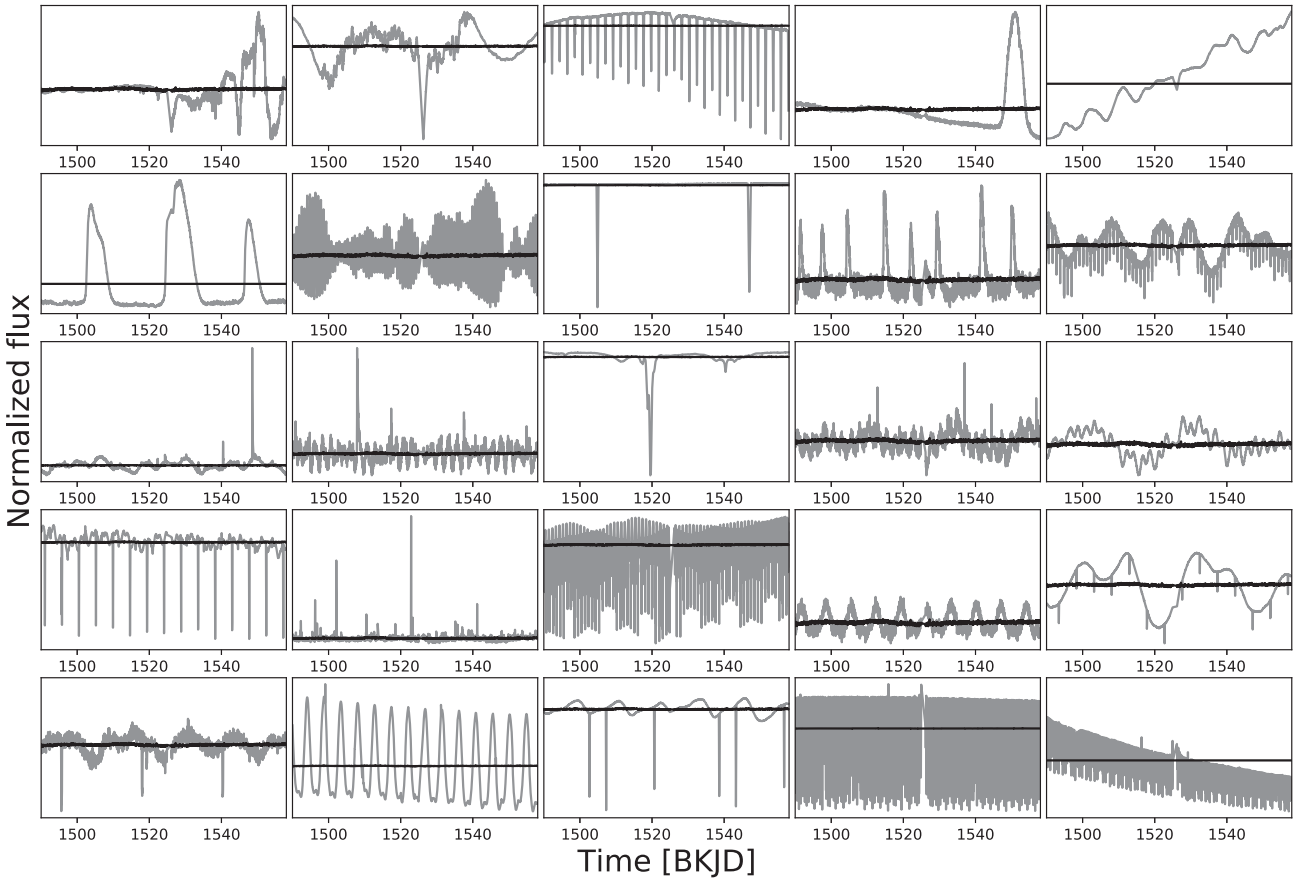


**Figure 11.** A selection of URF anomalous light curves without a current classification in the SIMBAD database. Boyajian's star is shown in the center panel. Note that the y-axis range in changes from light curve to light curve, whereas the time axis is kept the same for all. For comparison, each light curve is shown together with the light curve of low-URF score object KIC 8211660 (black line).

periods, such as $\delta$-Scuti stars, or long period variables are selected by the method, especially when they also have high amplitude variations. In particular, the combination of long (weeks/months) modulations with prominent dips and flares, as is the case of Boyajian's star (or certain types of eclipsing binaries), result in an anomalous behaviour recognized by our method.

In order to test our hypothesis, we look at how the most important features are distributed for both normal and anomalous light curves. We show the marginal distributions of a few pairs of features in Fig. 15. We selected the plotted features by ranking all features by importance, from more to less important, and then, for each group of high frequency power features, low frequency power features, and time-ordered light curve points, selecting from the top 5. The left-hand panel shows the distribution of spectral power density (PSD) at two frequencies roughly corresponding to time-scales of 1.5 hr (PSD-1) and 3 hr (PSD-2), for a representative group of normal objects (low URF score) and a representative group of anomalous objects (high URF score). For the same objects, the middle panel shows the distributions for low frequencies (15 d in the PSD-3 axis and 60 d in the PSD-4 axis), and the right-hand panel shows the normalized flux at BKJD 1536 and the spectral power at 1.5 hr (PSD-1).

We note that anomalies distribute quite differently in the joint space of features compared to normal objects. Of relevance is the fact that anomalous objects have less relative spectral power at high frequencies, but show a larger degree of correlation between the spectral power values at those frequencies. The relative variability amplitudes of anomalies are significantly enhanced with respect to the normal objects.

From this perspective, Boyajian's star is remarkable in more than one way. It shows pronounced dips seen in the light curve that deviate more than 20 per cent from the mean flux, but additionally it also shows shallower (0.5 per cent) modulations with characteristic time-scales of several weeks. These modulations result in a complex and unique power spectrum dominated by the large time-scales, some power bumps in short time-scales, and a noticeable decrease in spectral power at frequencies corresponding to time-scales of half a day, as well as higher frequency oscillations with similar characteristic periods. This time-scale coincides with the typical duration of the main flux dips (see Fig. 2).

### 4.3 Finding analogues of light curves of interest

We have demonstrated that the URF score is a good measure of the anomalous nature of a light curve, and we have investigated how anomalous light curves differ from normal light curves in the space of features defined by the light curve points and the power spectrum. However, we have also pointed out that the URF score alone is insufficient for finding light curves that are similar between them, which is an essential task if we are to find analogues of specific light curves or truly unique objects. In this section we investigate whether manifold-learning methods that reduce the dimensionality of the

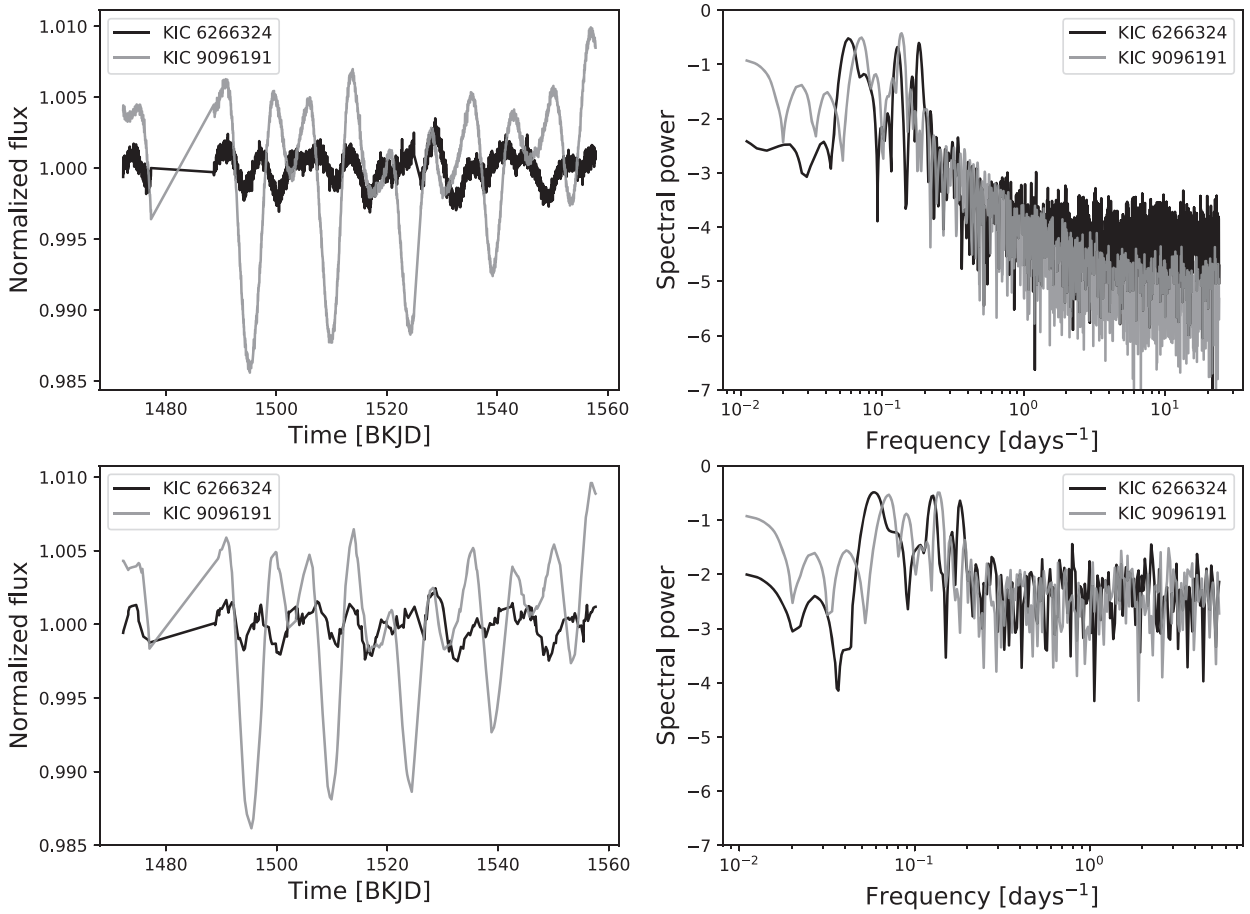


**Figure 12.** Effect of information loss on anomaly detection. Shown are the light curves (left-hand panels) and periodograms (right-hand panels) for two rotation-variable objects, and for both the dense (top panels) and sparse (bottom panels) versions. KIC 9096191 (grey line) is identified as an anomaly in both the dense and sparse data sets. KIC 6266324 (black line) is identified as an anomaly only in the dense data set. Note that the power spectrum is more uniform across frequencies for the sparse light curves.

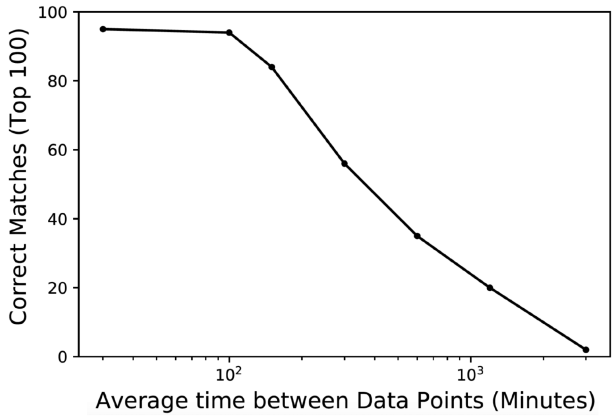


**Figure 13.** The recovery rate of anomalies as a function of the average separation between points, for a subset of 2500 light curves.

light curves are suitable for the identification of groups containing anomalous light curves sharing similar properties. We show the results of this dimensionality reduction and explore its own value as an anomaly detection method, before we move on to combining the results with the URF scores in order to find groups of similar anomalous objects.

#### *4.3.1 Reduced dimensionality for finding analogues*

*t*-SNE and UMAP (as described in Sections 3.2.1 and 3.2.2 respectively) are separately applied to a representative feature data set, namely the *DMDT* image representation of light curves introduced in Mahabal et al. (2017), and described in Section 2.2.2. The hyperparameters used are summarized in Section 3.2. Each of the two algorithms reduces the dimensionality of the data to only two dimensions, while trying to preserve similarity between light curves. We show the resulting *t*-SNE and UMAP embeddings for the dense light curves in Fig. 16, where each point represents a light curve, and has been colour-coded by its independently determined URF score.

Despite the loss of information associated with the dimensionality reduction, the embedding maps show a complex structure that relates to the broad range of variability types present in the *Kepler* data set. Both maps show distint, differenciated regions. The *t*-SNE embedding map is somewhat more complex, fragmented into different 'islands' and showing a less connected representation, whereas the UMAP embedding map shows a more continuous distribution and no fragmentation. Our analysis indicates that this distinction is due to the fact that *t*-SNE is more sensitive to baseline trends, *i.e.* similar objects might end up in different *t*-SNE islands if they look alike but have a different upward or downward trend. But both maps show a population of objects represented by dense filaments that span a considerable range of feature values.

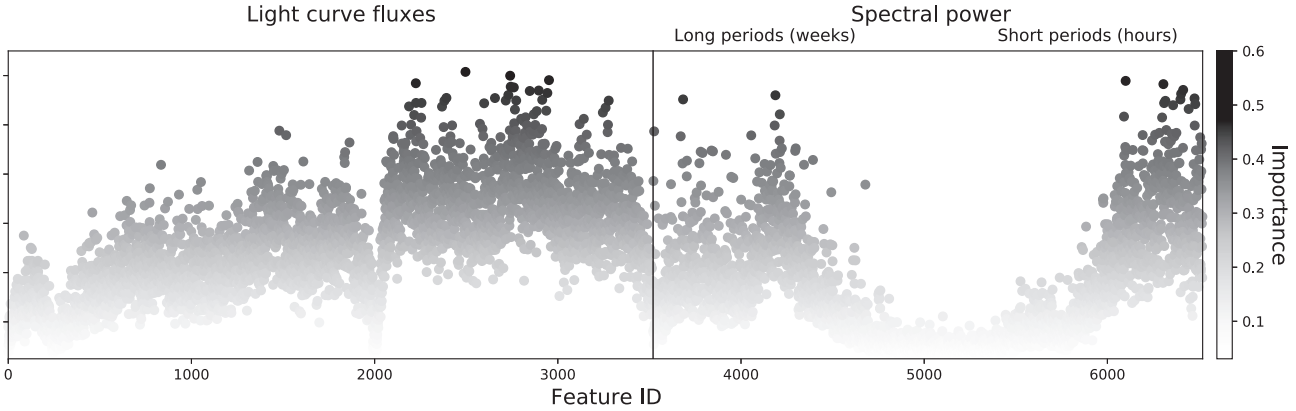


**Figure 14.** Feature importance distribution for the URF. The vertical line marks the boundary between normalized fluxes (left) and frequencies (right). Each dot represents a feature, and they are colour-coded by normalized importance.

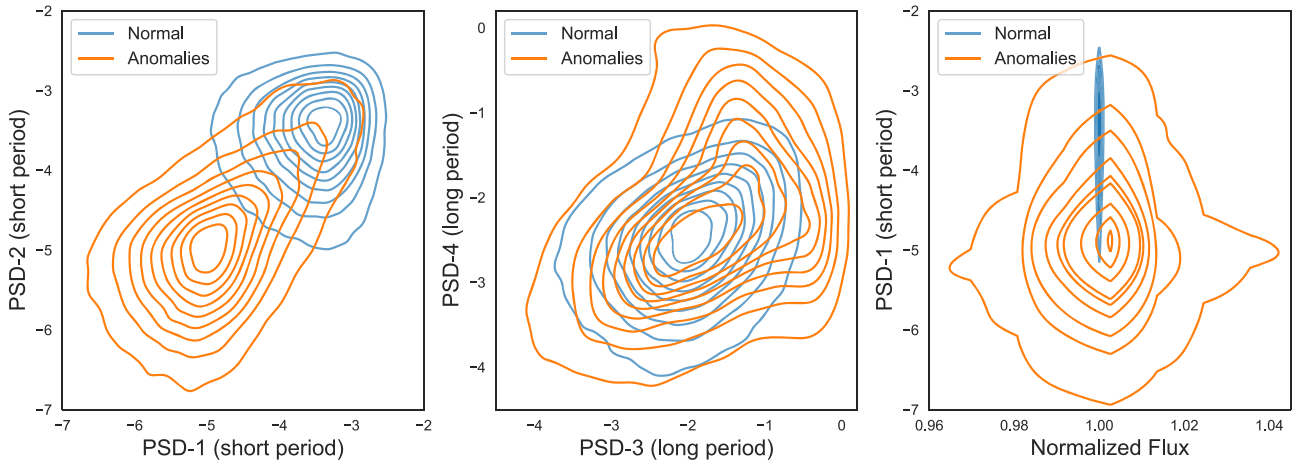


**Figure 15.** Two-dimensional density distributions of several important features within the URF model for normal (blue) and anomalous (orange) light curves. Shown are for different power spectral densities (PSDs), and one normalized flux value.

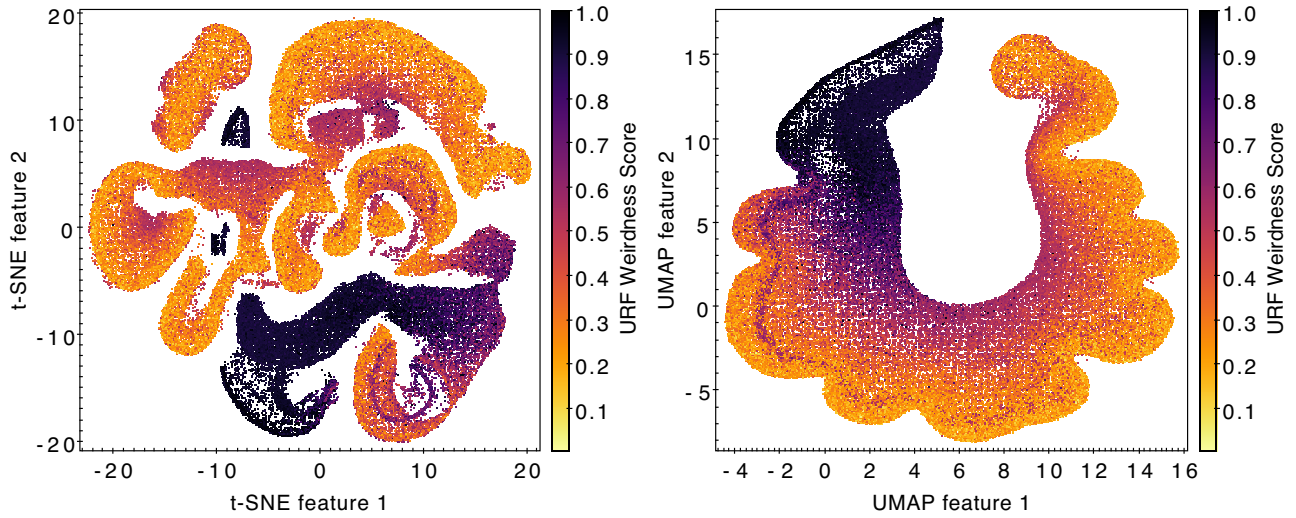


**Figure 16.** *Left*: The 2D embedding of the dense light curves using the *t*-SNE algorithm, colour coded by URF weirdness score. *Right*: The 2D embedding of the dense light curves using the UMAP algorithm, colour coded by URF weirdness score.

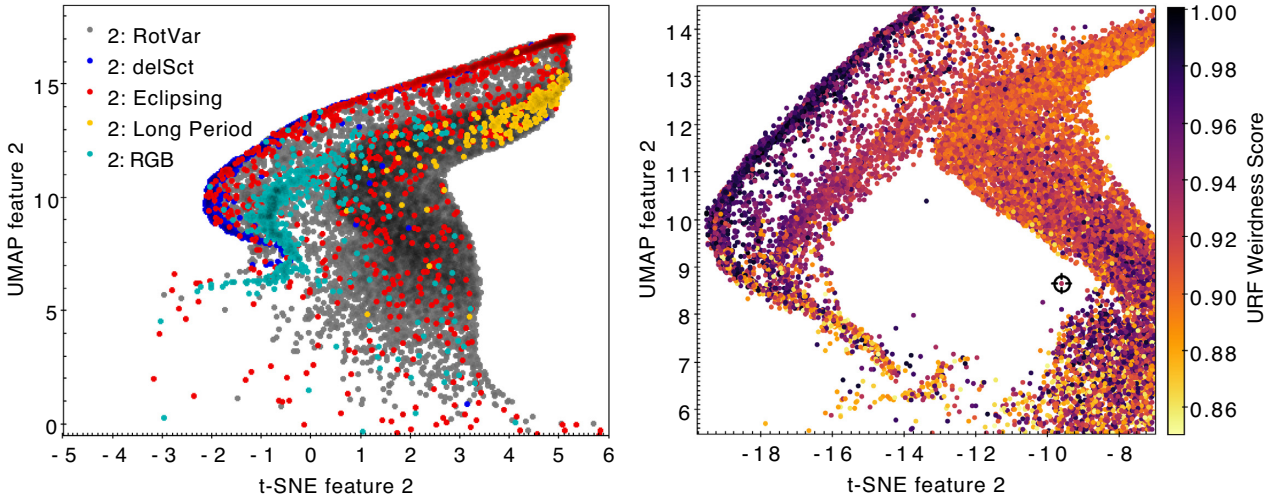


**Figure 17.** *Left*: The anomalous tip of the UMAP embedding map, with different types of anomalous objects (rotating variables, $\delta$-Scuti stars, eclipsing binaries, long period variables and oscillating RGB stars) indicated, from a crossmatch of our object list with the SIMBAD database. *Right*: A particular projection of the 4-dimensional space of the four combined embeddings, in the region close to the location of Boyajian's star, which is indicated by the cross mark. Shown are only points corresponding to anomalies, and the colour coding has been scaled to include only values in the anomalous peak of the distribution.

Remarkably, both maps show a striking correlation between the location of a light curve in the 2D space of the map, and the independently determined URF score for the light curve. This suggests that the properties that make a light curve anomalous, *i.e.* increased amplitude variability, having certain characteristic variability time-scales, or having truly unique light curve features, are to some extent *embedded* in these maps. More specifically, members of the anomalous peak of the URF distribution occupy a well defined area of these two maps, roughly corresponding to the black areas in the figures. In the two-dimensional (four-dimensional if both maps are considered) space of the manifold embeddings, the degeneracies between different types of light curves having similar URF scores are now broken, allowing us to associate objects with similar characteristics. This implies that anomalies that are locally close to each other in the 4-dimensional space of these embeddings are similar between them, *i.e.* they have analog behaviours.

To illustrate the last point, in the left-hand panel of Fig. 17 we show the anomalous tip of the UMAP embedding map, this time explicitly indicating different types of known anomalies that we have identified by cross-matching our list of objects with the SIMBAD database. These classes closely match the types included in our list of *bona-fide* anomalies. We note that objects of the same class tend to cluster together along well differentiated regions, or filaments of the map. The clustering is not exclusive, and there are relatively small overlaps between types, but considering the existing ambiguity in the classification of the sources this can be expected. Where, in this reduced space, do we find very unique anomalous objects, such as Boyajian's star? In both the *t*-SNE and UMAP embeddings of Fig. 16, Boyajian's star falls within the group of anomalous objects dominated by rotating variables, which in the *t*-SNE mapping corresponds to the elongated shape located at $y \sim -10$ and spanning the range $-8 < x < -5$. In each individual map it is difficult to distinguish it from other, less remarkable anomalies. Its truly unique nature comes to light, however, if we move to the 4-dimensional space of the combined *t*-SNE and UMAP features.

As illustrated in the right-hand panel of Fig. 17, Boyajian's star sits alone in an isolated region of the map, easily distinguishable from other anomalies, that in this region are mostly rotating variables. The two isolated anomalies closest to Boyajian's star in this region are KIC 5385778, which just like Boyajian's star show a combination of dips and long-term variability (although the dips are regular and far less pronounced), and KIC 3544657, which displays long-term irregular variability and shallow, irregular dips. They are, however, too separated from each other to be considered true analogues. They are all unique in their own merits. Using a similar approach, anomalous objects that belong to a particular rare class (*e.g.* RRLyrae stars[3]) and that have not yet been classified, can be found.

In online Appendix A, online Tables A2–A5, we provide small samples of the selected anomalies, that we have grouped by their probable class. In order to associate anomalies of unknown class to a specific labelled group, we first identified regions in the UMAP embedding with high density and high purity of a given class (*e.g.* Long Period Variables occupy a very specific region in the space of UMAP embeddings in Fig. 17). We then manually selected objects in this region of the 2D map that had an anomaly score above 0.85, and that also did not have a previous classification according to the SIMBAD database. The labels we have assigned in the tables should be taken only as indicative, as no formal classification has been performed on these anomalies.

### 4.3.2 *Distinguishing Euclidean anomalies in* t*-SNE and UMAP*

The fact that unique anomalies such as Boyajian's star appear isolated in the four-dimensional space of the *t*-SNE and UMAP embeddings suggests that the Euclidean distance in this space could be used as an anomaly score, at least for some types of light curves. We now investigate what types of anomalies can be identified based on a distance metric in the embedding maps, and whether those anomalies are similar to those found with the URF method.

Since the *t*-SNE and UMAP methods work by closely simulating the probability distribution $q_{i|j}$ of point $i$ picking point $j$ as a neighbour

[3]RRLyrae stars sit at the very tip of the anomalous region in the UMAP embedding map. Whereas in the general context of all variable stars they are not rare, they are extremely rare in the data set considered here.

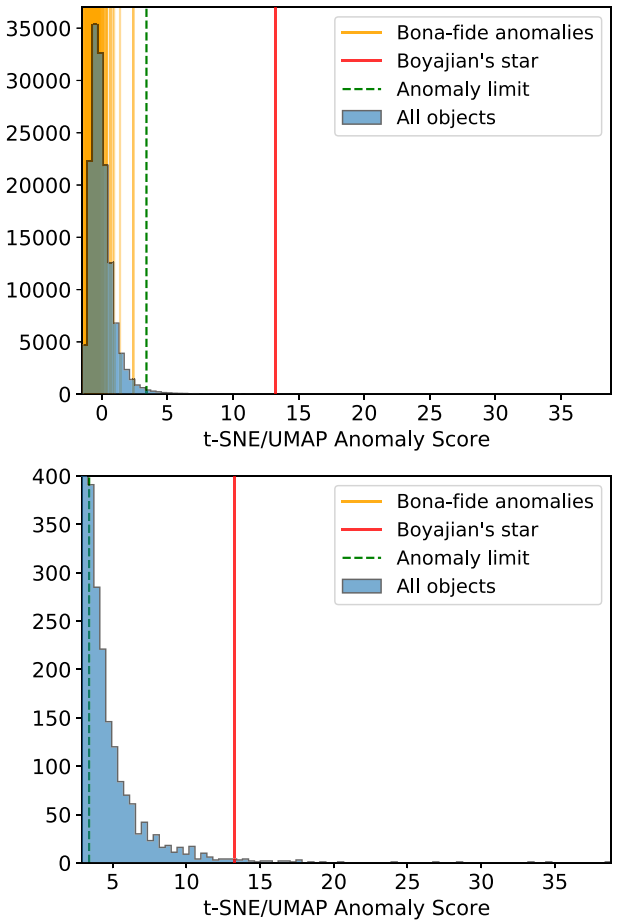


**Figure 18.** Histograms of manifold anomaly scores for the dense Kepler Q16 light curves. Indicated in orange and red are the scores corresponding to *bona-fide* anomalies, including Boyajian's star, as well as the anomaly limit (in green). The top panel shows all the objects, whereas the bottom panel shows the high anomaly score region of the distribution.

in the high dimensional space, we can build a distribution based on the pairwise Euclidean distances between points in the 4-dimensional space of the combined *t*-SNE and UMAP embeddings. Specifically, for each point we consider its distance to the nearest neighbour. We then consider the distribution formed by normalizing these distances and take all points that are at least $3.4\sigma$'s from the mean in order to set a threshold for anomalous objects. The specific value of $3.4\sigma$ was chosen heuristically by inspecting the number of anomalies as a function of the threshold, and choosing the value at which the second derivative of this curve was closest to zero for both the *t*-SNE and UMAP. The points selected are those that are furthest from the cluster and, by the Euclidean metric, they are anomalies in the manifold-based method.

In Fig. 18 we show the distribution of manifold anomaly scores computed as described for the dense light curves. In a similar manner as we did in Fig. 8, we indicate the scores of Boyajian's star and the rest of the *bona-fide* anomalies with vertical lines. We note that this time the majority of our *bona-fide* anomalies (in fact, all of them except for Boyajian's star) fall outside of the anomaly area, with many of them having fairly low scores. It is not too surprising that the methods may find different anomalies: *t*-SNE and UMAP use the *DMDT* image pixels as the features, which contains different information compared to the light curves-periodogram combinations. Second, as pointed out in the introduction, Euclidean distances become less meaningful as the dimension of the data grows, and thus when a high dimensional space is reduced to only a few dimensions with manifold approaches like the ones used here. Yet, Boyajian's star is clearly identified using this approach: Euclidean distances are still meaningful locally, and in fact we preferentially find certain types of isolated anomalous objects using distances between manifold features (see Fig. 19).

However, in the more general case, we are unable to know how many anomalies we are missing when we select them based on the Euclidean metric alone. We can understand this in terms of individual, or *one-off* anomalies versus groups anomalies. The Euclidean distance identifies *one-off* anomalies, isolated sources of which only one example exists, as well as anomalies that have only a few outlying features in an otherwise regular light curve (as in the case of rapid exoplanet transits), but it fails to identify groups of anomalies, such as RR Lyrae stars in the present Kepler data set, whose anomalous behaviour is more related to the overall variation of the light curve, taking into account many of the features at once. The URF method, on the other hand, successfully identifies both group anomalies with overall different feature distributions, and *one-off* anomalies. A similar conclusion applies for the *t*-SNE and UMAP embeddings when the sparse light curves are considered. The correlation between the location of a light curve in the embedding map and its URF score is still present for these sparse light curves. But the general idea about using Euclidean metrics for identifying *one-off* anomalies still holds.

We find a total of 1552 anomalous dense light curves using the 4-dimensional Euclidean metric and the diagnostic described above. Online Table A6 list the first 20 when they are ordered by the Euclidean score. The manifold anomaly scores for those object do not correlate with the URF scores, with some of the manifold anomalies having low URF scores (and vice-versa). But the most remarkable manifold anomalies are also selected in the URF method. Boyajian's star, as expected, ranks high in the manifold Euclidean score, at position 33 from the top most anomalous object. In Fig. 19 we show some representative examples of anomalies being selected by the manifold method, but not selected in the URF method. They are all in the top 150 of Euclidean scores. We note that, unlike those identified with the URF method, these anomalies are not as clearly dominated by large amplitude variations. Deviations from the mean normalized flux are not as pronounced as in the URF anomalies of Figs 9–11. Instead, they appear to be dominated by complex periodic or semiperiodic behaviour, either in the form of fast dips, such as those happening in transits with a typical duration comparable to or less than the Kepler cadence, or of periodic oscillations with more than one mode and characteristic time-scales of a few days, which incidentally is the time-scale that the URF method deems as less important (see Fig. 14). This ∼days-long periodicity, together with the fast transits, account for the vast majority of the manifold anomalies. The fact that these anomalies are being selected is probably due to the presence of features in the DMDT maps associated with these periodic behaviours, although it is difficult to interpret how those patterns reveal themselves in the 2D embeddings to which the images are eventually reduced. As a matter of fact, the *DMDT* maps for these anomalies do not look very different from the randomly selected maps of Fig. 3. This suggests that the 4-dimensional Euclidean anomaly metric selects anomalies based on subtle *DMDT* map differences between a given anomaly and the average map of the objects that are most like it. Sudden rapid dips in a light curve that is otherwise ordinary are a good example of this type of anomaly.

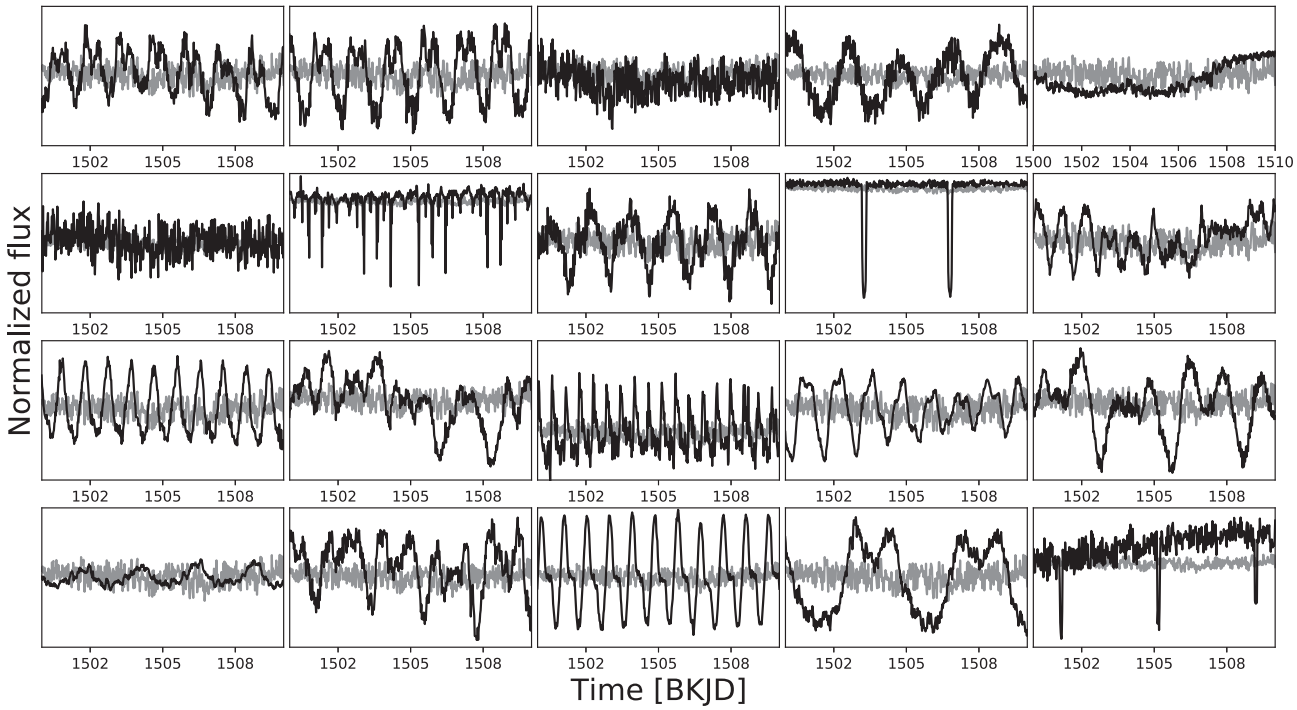


**Figure 19.** Examples of anomalous light curves identified using the Euclidean distance in the manifold embeddings of the *DMDT* maps. The anomalies are shown as black lines, whereas the grey line represents the reference object KIC 8211660.

In contrast, the *DMDT* maps of URF anomalies, of which we show a few in Fig. 20 show a more complex structure: they are more widely populated in the vertical DM axis (which is explained by the fact that they are amplitude-dominated), and show vertical and horizontal stripes representing either periodicities or magnitude dips. The manifold dimensionality reduction using either *t*-SNE or UMAP is very successful at translating those features into groups of similar, or analog light curves (which is why we see a correlation between the URF anomaly score and the 2D embeddings), but not very successful at identifying those complex features as anomalies. The Euclidean metric, we argue, is a good indicator of *local* anomalies (*i.e.* those with small deviations from an otherwise normal light curve). The additional information that we gain about light curve behaviour from the spectral power features in the URF method appears to be fundamental in the identification of a broader range of anomaly types, including those selected *a priori* as our *bona-fide* anomalies (RR Lyrae stars, Cepheids, etc.), that we know to be rare in the KIC. We therefore conclude that the URF score that we have derived in Section 4.2 for our set of objects is better suited for finding all types of anomalies, both *one-off* cases with no counterparts (either because they show a few outlying features, or because they distribute differently in the joint space of all features), and groups of objects on known type that are rare in a given data set.

#### *4.3.3 The full algorithm*

We now present an integrated view of our algorithm. Depending on the specific interest of the user, the method proposed here can be used in two modules in order to explore time-domain data. The user can choose to only find anomalies using the URF method applied to the light curve and periodogram points. But the user can in addition find analogues to anomalous light curves of interest by performing dimensionality reduction to the DMDT maps generated from the light curves, and then look for neighbours of the light curve of interest in the space of low dimensionality. In Fig. 21 we show a flow chart that summarizes the method.

We could have adopted the perfectly viable option of combining all the features that we have derived, including light curve points, periodograms, and DMDT maps, and run the URF algorithm on the resulting embeddings. This would have resulted in different scores, and potentially different anomalies. However, we have demonstrated in Fig. 17 that the URF scores and the *t*-SNE/UMAP embeddings are correlated, and therefore contain similar information. The advantage of splitting the analysis in two parts (anomaly identification with URF using the periodogram and light curve points, plus analog identification with the *t*-SNE/UMAP embeddings using the DMDT maps) is that we avoid redundancy (given the correlation already mentioned) while allowing for modular analysis. For example, a user can decide to perform only the anomaly detection part using URF, for which calculating the periodograms alone is less computationally expensive than calculating both the periodograms and the DMDT maps. Similarly, a user can opt for getting the DMDT maps, run them through t-SNE/UMAP, and look for light curves that are similar based on their proximity in these embeddings.

### 4.4 Astrophysical implications

We now turn to the question of whether there is a correlation between anomalous variability of an object and its overall physical properties. Our results indicate that anomalous light curves can be associated to known astrophysical processes, such as intrinsic stellar pulsations of different types (*e.g.* Arras, Townsley & Bildsten 2006), flares (Davenport et al. 2014; Paudel et al. 2018), or extrinsic binary eclipsing phenomena (Prša & Eclipsing Binary Working Group 2013; Prša et al. 2016), but also to phenomena that are a challenge to current models, such as the case of Boyajian's star. Therefore, you would

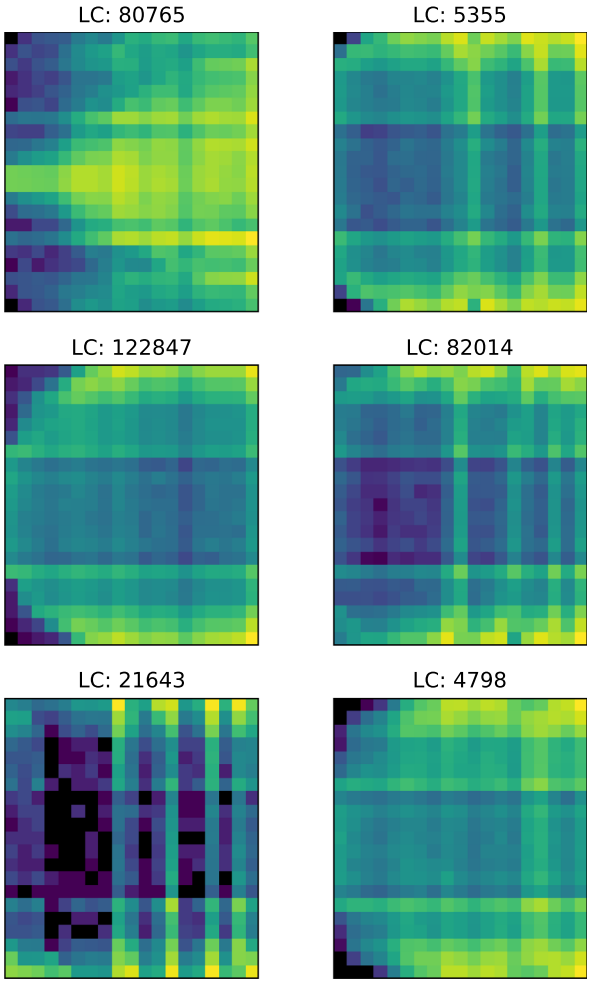


**Figure 20.** DMDT maps for a sample of our *bona-fide* anomalies. Shown clockwise from upper left are Boyajian's star, a classical Cepheid, a $\gamma$-Dor variable, a slow pulsating B star, an RRLyrae star,and an ellipsoidal binary.

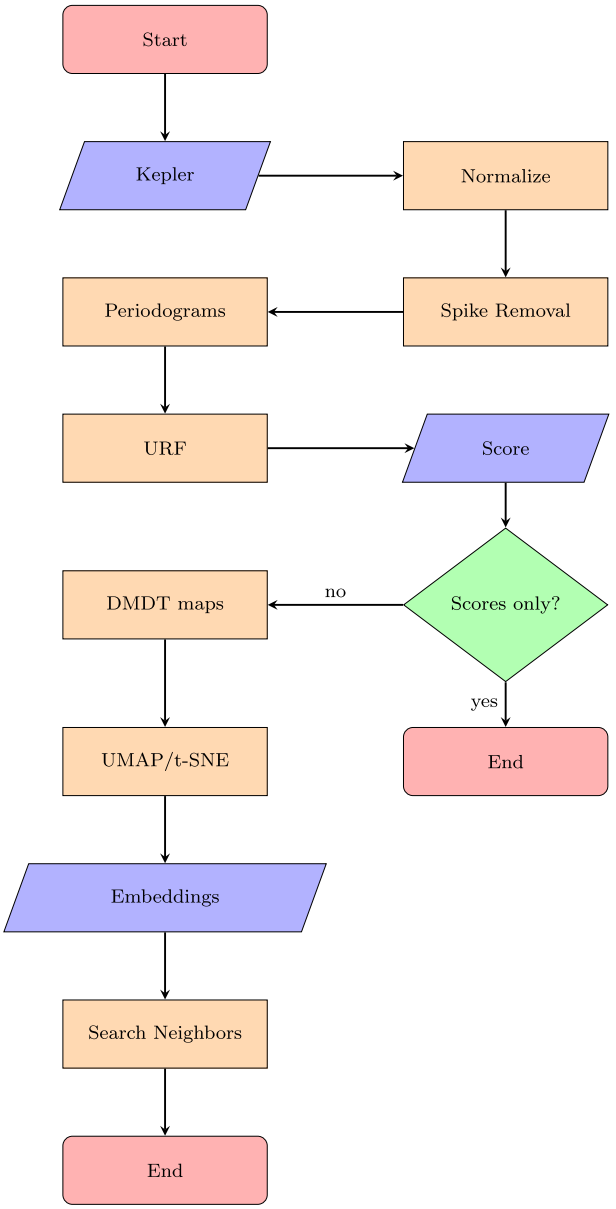


**Figure 21.** A flow chart showing how our proposed algorithm works, both for finding anomaly scores only, and for finding analogues of light curves of interest.

expect that there is some correlation between the URF score and the luminosity and temperatures of these objects.

In order to investigate this, we have obtained *Gaia* DR2 colours and derived G band magnitudes (Gaia Collaboration et al. 2016, 2018) for all the objects in our sample with available *Gaia* measurements, and produced a Hertzprung-Russell diagram that we have colour-coded according to the URF anomaly score. We show this diagram in Fig. 22. The left-hand panel shows the entire set of objects analysed in this paper, whereas the right-hand panel shows those objects that we consider anomalous, with a URF score larger than 0.85.

A clear correlation is revealed between astrophysical properties and anomalous variability. The majority of anomalies lie along the main sequence (MS), but those with the highest URF score are not preferentially located along the MS. Instead, pulsators in the instability strip such as $\delta$-Scuti stars, $\gamma$-Doradus stars, and RR Lyrae stars take the first place among the most anomalous objects, followed by eruptive stars and red dwarfs with high amplitude stochastic variability. White dwarfs and hot sub-dwarfs also belong to the group of the most anomalous objects, also due mostly to their high amplitude variations. Then come most of eclipsing binaries and rotational variables, with the latter being the group most represented along the main sequence. The last place among the anomalous is taken by irregular pulsating giants, long period variables, and Mira stars, in that order. The URF score gradient for the anomalies is clearly seen in the right-hand panel of Fig. 22.

The point density of objects is different between the two diagrams in Fig. 22, which indicates that anomalies distribute differently in their physical properties with respect to the general population of KIC objects. With respect to the latter, there is an overdensity of anomalies in the instability strip, where pulsations due to instabilities in stars with ionized He atmospheres are common. In addition, anomalies cluster more strongly compared to the KIC at lower luminosities along the main sequence. This might be related to the larger incidence of eclipsing binaries and flaring young stars at late spectral types.

Perhaps the most unexpected anomalies are those that lie along the MS, and for which no particular strong pulsation modes or high amplitude variations are expected, and that are not rotational variables or eclipsing binaries. These objects, of which Boyajian's star is an example, are local anomalies that can be identified using an

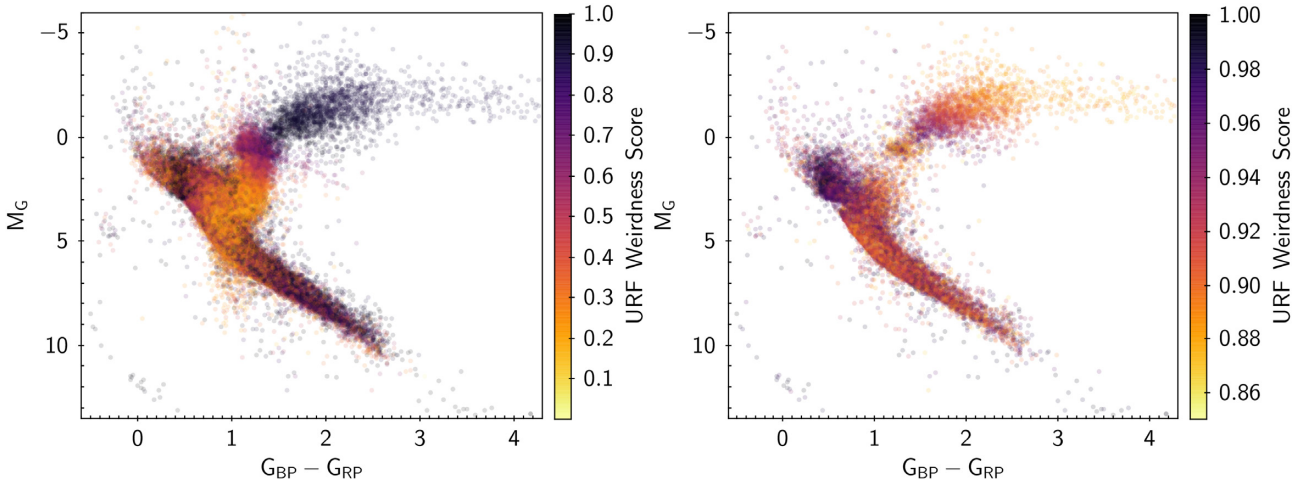


**Figure 22.** *Left*: The Hertzprung-Russell diagram for those stars in our set with measured *Gaia* magnitudes and parallaxes, colour-coded by the URF score for the dense light curves. *Right*: The diagram for the anomalous objects, as determined using the dense light curves (URF score larger than 0.85). The colourbar has been renormalized to contain only values larger than 0.85.

Euclidean metric in the space of manifold embeddings, as we have discussed. The correlation between URF anomaly score, embedded low-dimensionality features, and astrophysical properties indicates a possible path toward new discoveries. In particular, the application of a similar analysis to TESS light curves will allow us to identify scientifically compelling light curves for further analysis. We are currently performing this analysis on TESS data (Crake et al. in prep.). Note that this type of analysis is not limited to light curves with regular cadences. Our experimental setup allows us to adjust for differences in light curve lengths, cadences, and gaps between light curve points. Both the spectral decomposition and the *DMDT* images can be produced for light curves of any length and cadence.

Red clump stars deserve a special mention. They are seen in the left-hand panel of Fig. 22, as the purple clump of objects at higher temperatures and lower luminosities compared to the anomalous giants. We have seen that they also stick out in the distribution of anomaly scores, having URF scores of about 0.55. These are RGB stars supported by helium fusion on their core. They tend to clump in the HR diagram because they end up having the same luminosity at their red giant stage regardless of their initial age or composition (Girardi 2016). In terms of variability, they have amplitude variations that are less pronounced compared to other RGB stars (but still clearly seen in Kepler light curves), and these amplitude variations are also less regular compared to the RGB pulsations. In the UMAP embedding of the right-hand panel in Fig. 16, they live in the purple filament seen to the left, just below the anomalous filament of the instability strip pulsators.

## 5 DISCUSSION AND CONCLUSIONS

Anomaly detection in time domain astronomy enables the discovery of unique light curves that could lead to the formulation of new hypotheses. Such is the case of the anomalous object Boyajian's star, whose *Kepler* light curve has become the gold standard in the search for anomalous light curves. The search for anomalies is even more relevant in the era of *TESS* and the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST), which will deliver thousands of light curves per night over a period of 10 yr. Yet, it would seem a very difficult task to define an anomalous light curve, as the anomalous nature of an astronomical time-series depends on the details of the methods employed, as well as on the features that are used to represent the light curve ensemble. Here we have proposed an approach for the detection of time-domain anomalies and of their analogues, and we have applied it to *Kepler* data. We have accomplished the following:

(i) By comparing our candidates for anomalous light curves with a list of *bona-fide* objects that we knew in advance to be rare in the context of the Kepler Input Catalog (KIC), we have shown that we can isolate rare and unique light curves, including Boyajian's star and a broad range of variable stars that are underrepresented in the KIC.

(ii) We provide an empirical definition: an anomaly is an object whose most relevant features (*i.e.* those that most efficiently reduce impurity during the random forest classification), distribute differently compared to the corresponding features of non-anomalous objects. (Figs 14 and 15).

(iii) Starting from a simple set of features, namely the light curve points and their power spectra, we provide a specific measure for anomalous behaviour, the URF score, whose distribution over the entire data set defines a clear set of anomalies, including at least 5000 anomalies that have not been previously classified in the KIC (Fig. 8).

(iv) Using an image representation of the light curves and manifold methods for dimensionality reduction, we break the degeneracy in the one-dimensional URF score, and provide a nearest-neighbours method to group objects with similar characteristics, facilitating the identification of analogues to interesting anomalies (Figs 16 and 17).

(v) Using Euclidean distances in the four-dimensional space created by assembling the *t*-SNE and UMAP projections, group anomalies (such as rare known stellar types) are not identified, but *one-off* anomalies are clearly identified. The latter can be truly unique stars that, like Boyajian's star, that have no counter part in the sample, or otherwise regular objects with a few extremely outlying features.

Along the way, we have also demonstrated the impact of pre-processing on anomaly detection, as observational or instrumental artefacts such as spurious spikes or bad pixels can have a significant impact on the anomaly score (Figs 5 and 6). We have also shown

that our method can be applied to sparser, unevenly sampled light curves by deploying it on a version of the light curves with only 10 per cent of the points and demonstrated it is still possible to find some of the relevant anomalous objects, while missing anomalies that are selected based on large-scale differences in their power spectra (Fig. 12). This is relevant to extend our work to ground-based surveys.

Our work is complementary to previous methods that have attempted to identify time-domain anomalies in *Kepler* data (*e.g.* Giles & Walkowicz 2019) and incorporates new features that allow for the identification of analogues.

The remarkable correlation between the location of a light curve in the space of *t*-SNE and UMAP embedded 2-dimensional features, and the independently determined URF weirdness score, indicates that the apparent complexity of the Kepler light curves, with a broad range of frequencies, amplitudes, transient behaviour, and periodicity, can be represented by a handful of numbers, with anomalies occupying specific regions in these representations. In particular, we have shown that 'group' anomalies such as RR Lyrae stars, instability strip pulsators, long period variables, live together in specific areas of this lower-dimensional space, whereas 'isolated' anomalies, such as Boyajian's star, stick out locally as objects without neighbours in the vicinity of the embedded maps.

Finally, we have demonstrated that there is a linkage between anomalous behaviour and astrophysical properties (Fig. 22). We have shown that, although *Kepler* anomalies live in many different region of the HR diagram (most of them along the main sequence), the distribution of their astrophysical properties is different compared to the bulk of the objects. We have also shown that 'group' anomalies live in physically meaningful regions of the HR diagram, such as the instability strip and the red giant branch, but that the most fertile ground for discovery lies within the realm of individually isolated anomalous light curves of objects that live in otherwise uneventful regions of the HR diagram.

## ACKNOWLEDGEMENTS

We thank the referee for a very detailed report that made this article significantly better.

We thank the organizers and participants of the *Detecting the Unexpected* workshop that took place at STScI in 2017. The ideas for this work came from a hack during that workshop and have produced also other papers. In particular, we thank Lucianne Walkovicz for a continuous exchange of ideas and for proposing the original hack. We also thank Dalya Baron for useful insight about the use of the URF method. We thank the original hackers' team which included Kelle Cruz, and Umaa Rebbapragada

The authors acknowledge the support of the Vera C. Rubin Observatory Legacy Survey of Space and Time Transient and Variable Stars Science Collaboration (TVS SC), of which most of the authors are member and that provided opportunities for collaboration and exchange of ideas and knowledge.

This paper includes data collected by the Kepler mission and obtained from the MAST data archive at the Space Telescope Science Institute (STScI). Funding for the Kepler mission is provided by the NASA Science Mission Directorate. STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 5-26555. This work has made use of data from the European Space Agency (ESA) *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

This work made use of several `Python` modules including:

(i) `numpy` (Harris et al. 2020)
(ii) `maplotlib` (Hunter 2007)
(iii) `scikit-learn` (Pedregosa et al. 2011)
(iv) `seaborn` (Waskom et al. 2017)

## DATA AVAILABILITY

The data underlying this article were accessed from the Mikulski Archive for Space Telescopes (MAST), at https://mast.stsci.edu/portal/Mashup/Clients/Mast/Portal.html. The derived data generated in this research can be accessed from the `GitHub` repository https://github.com/kushaltirumala/WaldoInSky.

## REFERENCES

Aggarwal C. C., Yu P. S., 2001, in Proceedings of the 2001 ACM SIGMOD international conference on Management of data. Association for Computing Machinery, New York, p. 37
Aleo P. D. et al., 2020, Res. Notes Am. Astron. Soc., 4, 112
Arras P., Townsley D. M., Bildsten L., 2006, ApJ, 643, L119
Baron D., Poznanski D., 2017a, Astrophysics Source Code Library, record ascl:1611.07526
Baron D., Poznanski D., 2017b, MNRAS, 465, 4530
Bellm E. C. et al., 2019, PASP, 131, 018002
Bengio Y., Courville A., Vincent P., 2013, Proc. IEEE, 35, 1798
Bianco F. et al., 2021, ApJS, (arXiv:2108.01683)
Biau G., 2012, J. Mach. Learn. Res., 13, 1063
Blázquez-García A., Conde A., Mori U., Lozano J. A., 2020, preprint (arXiv:2002.04236)
Boyajian T. S. et al., 2016, MNRAS, 457, 3988
Breiman L., 2001, Mach. Learn., 45, 5
Buitinck L. et al., 2013, preprint (arXiv:1309.0238)
Che Z., Purushotham S., Cho K., Sontag D., Liu Y., 2018, Scientific Reports, 8, 6085
Chen H., Diethe T., Twomey N., Flach P., 2018, Anomaly Detection in Star Light Curves using Hierarchical Gaussian Processes. ESANN
Conroy C. et al., 2018, ApJ, 864, 111
Davenport J. R. A. et al., 2014, ApJ, 797, 122
Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, A&A, 475, 1159
Drake A. J. et al., 2012, in Griffin E., Hanisch R., Seaman R., eds, Proc. IAU Symp. 285, New Horizons in Time Domain Astronomy. Kluwer, Dordrecht, p. 306
Druetto A., Roberti M., Cancelliere R., Cavagnino D., Gai M., 2019, Rojas I., Joya G., Catala A., eds, Lecture Notes in Computer Science, Advances in Computational Intelligence, Vol. 11507, Springer Nature, Switzerland, p. 390
Dubath P. et al., 2011, MNRAS, 414, 2602
Dutta H., Giannella C., Borne K., Kargupta H., 2007, Proceedings of the 2007 SIAM International Conference on Data Mining, SIAM, Philadelphia
Emmott A. F., Das S., Dietterich T., Fern A., Wong W.-K., 2013, in Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description. ODD '13. Association for Computing Machinery, New York, p. 16
Eyer L., Mowlavi N., 2008, J. Phys. Conf. Ser., 118, 012010
Eyer L., Süveges M., De Ridder J., Regibo S., Mowlavi N., Holl B., Rimoldini L., Bouchy F., 2019, PASP, 131, 088001
Fulcher B., 2017, preprint (arXiv:1709.08055)
Gaia Collaboration et al., 2016, A&A, 595, A1
Gaia Collaboration et al., 2018, A&A, 616, A1
Giles D., Walkowicz L., 2019, MNRAS, 484, 834
Giles D. K., Walkowicz L., 2020, MNRAS, 499, 524

Girardi L., 1999, MNRAS, 308, 818
Girardi L., 2016, ARA&A, 54, 95
Goldstein M., Uchida S., 2016, PLoS One, 11, 4
Graham M. J., Drake A. J., Djorgovski S. G., Mahabal A. A., Donalek C., Duan V., Maker A., 2013, MNRAS, 434, 3423
Harris C. R. et al., 2020, Nature, 585, 357
Henrion M., Hand D. J., Gandy A., Mortlock D. J., 2013, Statistical Analysis and Data Mining. Wiley, New York, p. 53
Hinton G. E., Roweis S. T., 2003, Advances in Neural Information Processing Systems, Vol. 15, MIT Press, Cambridge, p. 833
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ishida E. E. O. et al., 2021, A&A, 650, A195
Ivezić Ž. et al., 2019, ApJ, 873, 111
Jamal S., Bloom J. S., 2020, ApJS, 250, 30
Jenkins J. M., 2017, Kepler Data Processing Handbook: Philosophy and Scope. Kepler Science Document KSCI-19081-002
Johnston K. B., Oluseyi H. M., 2017, New Astron., 52, 35
Johnston K. B., Caballero-Nieves S. M., Peter A. M., Petit V., Haber R., 2019, in Teuben P. J., Pound M. W., Thomas B. A., Warner E. M., eds, ASP Conf. Ser. Vol. 523, Astronomical Data Analysis Software and Systems XXVII. Astron. Soc. Pac., San Francisco, p. 83
Kessler R. et al., 2019, PASP, 131, 094501
Kochanek C. S. et al., 2017, PASP, 129, 104502
Kullback S., Leibler R. A., 1951, Annu. Math. Stat., 22, 79
Li X., Ragosta F., Clarkson W. I., Bianco F. B., 2021, preprint (arXiv:2107.10281)
Liu F. T., Ting K. M., Zhou Z.-H., 2012, ACM Transactions on Knowledge Discovery from Data, 6, 1
Lochner M., Bassett B. A., 2021, Astron. Comput., 36, 100481
Lomb N. R., 1976, Ap&SS, 39, 447
McInnes L., Healy J., Melville J., 2018, preprint (arXiv:1802.03426)
Maaten L. v. d., Hinton G., 2008, J. Mach. Learn. Res., 9, 2579
Mahabal A., Sheth K., Gieseke F., Pai A., Djorgovski S. G., Drake A. J., Graham M. J., 2017, IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, p. 1
Malanchev K. L. et al., 2021, MNRAS, 502, 5147
Margalef-Bentabol B., Huertas-Company M., Charnock T., Margalef-Bentabol C., Bernardi M., Dubois Y., Storey-Fisher K., Zanis L., 2020, MNRAS, 496, 2346
Meech K. J. et al., 2017, Nature, 552, 378
Miniutti G. et al., 2019, Nature, 573, 381
Nun I., Protopapas P., Sim B., Zhu M., Dave R., Castro N., Pichara K., 2015, preprint (arXiv:1506.00010)
Nun I., Protopapas P., Sim B., Chen W., 2016, AJ, 152, 71
Paudel R. R., Gizis J. E., Mullan D. J., Schmidt S. J., Burgasser A. J., Williams P. K. G., Berger E., 2018, ApJ, 861, 76
Paudel R. R., Gizis J. E., Mullan D., Schmidt S. J., Burgasser A. J., Williams P. K., Youngblood A., Stassun K., 2019, MNRAS, 486, 1438
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Prša A. et al., 2016, ApJS, 227, 29
Prša A., Eclipsing Binary Working Group, 2013, Giants of Eclipse, Vol. 45, BAAS, p. 40102
Pruzhinskaya M. V., Malanchev K. L., Kornilov M. V., Ishida E. E. O., Mondon F., Volnova A. A., Korolev V. S., 2019, MNRAS, 489, 3591
Rebbapragada U., Protopapas P., Brodley C. E., Alcock C., 2009, Mach. Learn., 74, 281
Reis I., Poznanski D., Baron D., Zasowski G., Shahaf S., 2018, MNRAS, 476, 2117
Richards J. W. et al., 2011, ApJ, 733, 10
Scargle J. D., 1982, ApJ, 263, 835
Schmidt M., 2019, preprint (arXiv:1907.02574)
Shi T., Horvath S., 2006, J. Comput. Graph. Stat., 15, 118
Škoda P., Podsztavek O., Tvrdík P., 2020, A&A, 643, A122
Storey-Fisher K., Huertas-Company M., Ramachandra N., Lanusse F., Leauthaud A., Luo Y., Huang S., Prochaska J. X., 2021, MNRAS, 508, 2946
Szklenár T., Bódi A., Tarczay-Nehéz D., Vida K., Marton G., Mező G., Forró A., Szabó R., 2020, ApJ, 897, L12
VanderPlas J. T., 2018, ApJS, 236, 16
Waskom M. et al., 2017, mwaskom/seaborn: v0.8.1. Zenodo. Available at: https://doi.org/10.5281/zenodo.883859
York D. G. et al., 2000, AJ, 120, 1579

## SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

The code used in this work is available at https://github.com/kushaltirumala/WaldoInSky

**unclassified_anomalies.csv**
**appendix.pdf**

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TEX/LATEX file prepared by the author.