TESS Asteroseismic Predictions for Red Giants using Machine Learning

M. Schofield[1,2]⋆, G. R. Davies[1,2], W. J. Chaplin[1,2], M. F. Randrianandrasana

[1] *Department of Physics and Astronomy, the University of Birmingham, Birmingham B15 2TT, UK*
[2] *Stellar Astrophysics Centre (SAC), Department of Physics and Astronomy, Aarhus University, Ny Munkegade 120, DK-8000 Aarhus C, Denmark*

Last updated 2017 May 31; in original form 2017 May 31

ABSTRACT

Summary: This paper presents a method to predict Red Giant mode detectability with TESS, using the Machine Learning algorithm Classification. It requires only the global parameters $\Delta\nu$, $\nu_{\max}$, the stellar magnitude and length of observation.

Method: Lightcurves for *Kepler* stars with fitted radial mode frequencies were used to generate equivalent TESS lightcurves. The lightcurves were cut down, *Kepler* white noise was removed, the bandpass was adjusted, and TESS white noise was added. A detection test was run on the observed modes in these 'TESS-like' lightcurves. Classifiers were then used to predict mode detectability with TESS based upon the global asteroseismic parameters $\nu_{\max}$ and $\Delta\nu$, as well as the stellar magnitude.

Application: By changing only the length of dataset and instrumental noise level, this tool can make predictions for any future mission such as K2, PLATO or CoRoT. This is therefore an ideal tool for target selection.

1 INTRODUCTION

1.1 Introducing Machine Learning

Astronomy is increasingly becoming a field of big-data analysis (Kremer et al. 2017). One of the main tools for handling this larger amount of information is Machine Learning. In most situations, Machine Learning is used to solve problems in one of two ways; Supervised or Unsupervised Learning. Supervised Learning involves problems where there is a known result.

For example, Supervised Learning has been used to classify types of variable star using previously labelled data (Nun et al. (2014), Elorrieta et al. (2016)). This previously labelled data is known as training data: this is used to train the Machine Learning algorithm. In the problem of variable stars, lightcurves that had already been classified were used to train the algorithm (this is the training dataset). This algorithm was then used to classify the lightcurves of unidentified stars (this is known as the testing dataset).

In Unsupervised Learning, there are no known results (labels). The aim of Machine Learning in this case would be to find trends between variables. This could be used to identify similar stars by analysing their lightcurves without using previously classified data (for example, Valenzuela & Pichara (2018)).

The aim of this work is to use Machine Learning to make predictions about mode detection probability ($P_{\rm det}$). In this case, $P_{\rm det}$is the known label, so this is a Supervised Learning problem[1].

Within Supervised Learning, two common algorithms are that are used are Classification and Regression. In Regression, the relationship between variables is interpreted using a measure of uncertainty (such as using χ^2 values). Models are fitted using the independent data, and uncertainty is measured. The models are then improved by reducing this uncertainty. Note that regression is used when the label is continuous. For example, predicting the magnitude of a star is a problem suited to regression, as a star can have any magnitude.

Conversely, Classification algorithms work by assessing similarity[2]. In Classification, the training set is separated into groups based on the similarity of the data. The more information that was gained by splitting the data, the better. For example, if the problem were to separate Red Giant stars from Main-Sequence stars, a star could be classified as either a Red Giant (1), or not a Red Giant (0). In the case of identifying Red Giants, having a Luminosity above $\sim 10 L_{\odot}$ would be a strong indicator that the star was a Red Giant, so the data could be separated into groups here. This is not the only problem where Classification can be used on Red Giant stars (for example, see **?**).

1.2 Introducing Asteroseismic Target Selection

Satellites such as *Kepler* have allowed asteroseismology of solar-like and Red Giant stars to advance rapidly since the last century Chaplin & Miglio (2013). Power spectra can now be resolved to detect individual modes of oscillation in Red Giant stars (Davies & Miglio 2016).

Future space missions such as TESS (Ricker et al. 2014), K2 (Howell et al. 2014), CoRoT (Baglin et al. 2006) and PLATO (Rauer et al. 2014) will add to our understanding of stellar structure and evolution. In order to select asteroseismic targets for these missions, the potential detectability modes inside these stars needs to be estimated.

[1] https://machinelearningmastery.com

[2] http://www.simafore.com

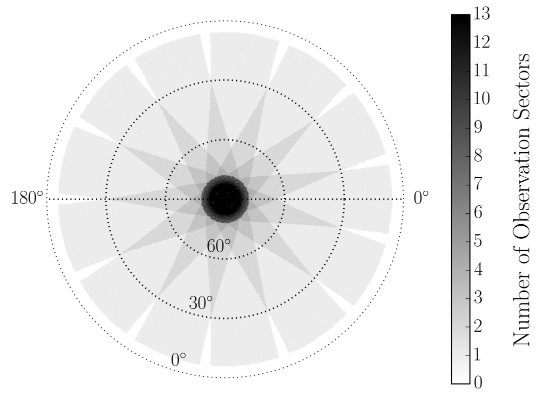

Figure 1. TESS field of view, centred around the ecliptic pole. Each strip will be observed for 27 days before the satellite rotates. Image taken from Campante et al. (2016).

In this work, individual fitted modes from Davies & Miglio (2016) were used to make asteroseismic predictions for TESS. By separating the targets into those with detected modes, and those without, Machine Learning was used to select a set of targets for future observation. Section 2 describes how the timeseries of every star was treated before transforming it to a power spectra. Section 3 then goes on to describe the detection test that was run of every fitted mode.

Lastly, Section 4 describes classification into stars with detected modes, and stars without. This was done by giving information on every target (the frequency of maximum solar-like amplitude $\nu_{\max}$, the large separation between modes of te same angular degree $\Delta\nu$, the magnitude, and which modes were detected inside the star) was given to a supervised Classifier. This data given to the Classifier is called the training set-it was used to train the algorithm. 30% of the sample was kept to test the algorithm-this data is known as the testing set. The Classifier recognised patterns between the variables in the training set, and made predictions on the testing set with a ?????????% accuracy.

2 TRANSFORMING THE LIGHTCURVES

Firstly, the timeseries data from *Kepler* observations needed to be adjusted for a different satellite and mission. This could either be done in the time or frequency domain. Both methods were tested and compared. The time domain method was chosen to transform the lightcurves, and was used in the rest of this work.

Several different adjustments needed to be made to the *Kepler* data. One difference between the missions is the length of observation. The *Kepler* mission observed for 4 years, while TESS' nominal 2 year mission will observe stars for between 27 days to 1 year, according to the star's ecliptic latitude (Figure 1). This distinction is made clear in the timeseries, see Figure 2.

As well as reducing the dataset length, the bandpass that TESS will observe in is much redder than that of *Kepler*, Figure 3. This has the effect of reducing the amplitude of stellar signals (i.e the signal due to stellar granulation and oscillation). Campante et al. (2016) found this bandpass correction to be 0.85.

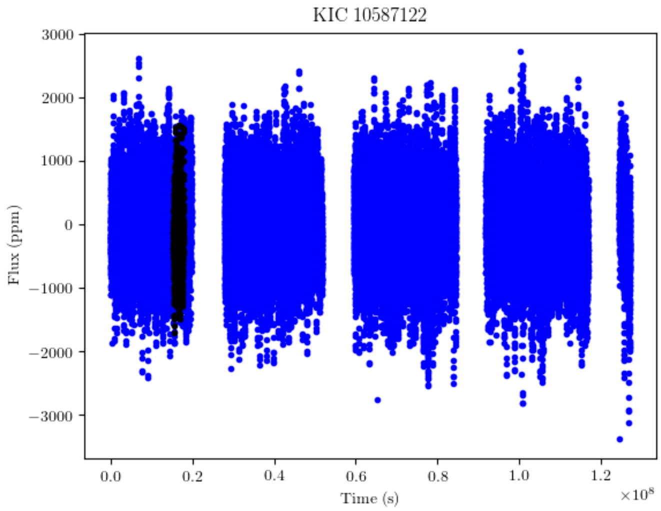

Figure 2. The 4-year long Power Spectra of KIC 10587122 is plotted in blue. Overplotted is the 27-day time segment with most coverage (the period with fewest gaps in the data). Reducing the length of observation this drastically will badly hamper mode detectability.

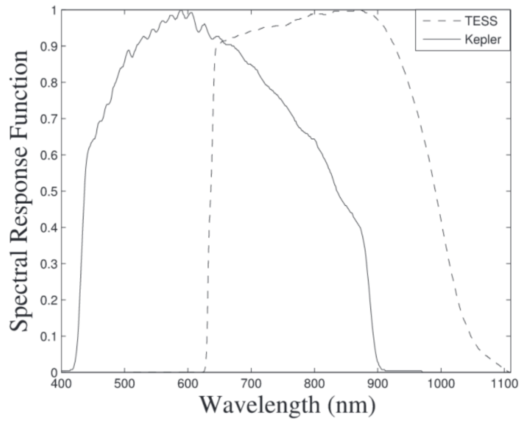

Figure 3. The bandpass of the *Kepler* and TESS missions. TESS will observe at longer (i.e redder) wavelengths than *Kepler*. This will reduce the amplitude of oscillations and granulation, whilst the white noise level will by unaffected.

Thirdly, the instrumental noise level in *Kepler* is different to the noise level in TESS. The noise level for the *Kepler* satellite depends on the *Kepler* magnitude of the star, K_p. The noise can be calculated using

$$\frac{10^6}{c} \times \sqrt{c + 9.5 \times 10^5 \left(\frac{14}{K_p}\right)^5}, \tag{1}$$

where

$$c = 1.28 \times 10^{0.4(12-K_p)+7}, \tag{2}$$

from Chaplin et al. (2011).

As well as subtracting this noise from the signal, instrumental noise from TESS needed to be added. This noise was estimated using the 'calc noise' IDL procedure (from William Chaplin, private communication), which depends on the I_c−band magnitude of the star and the number of pixels in the photometric aperture used when observing the star. This is given by

$$N_{\rm aper} = 10 \times (n + 10), \tag{3}$$

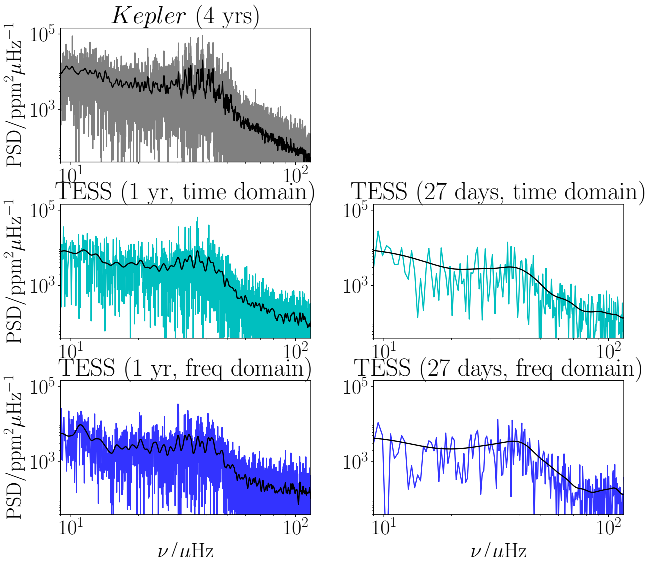

Figure 4. The Power spectra of KIC 10587122 is plotted five times. The original Power Spectra is plotted in grey. The power spectra after making the data look like TESS are plotted in blue. The data was transformed in the time domain (light blue) or frequency domain (dark blue). The left column shows 1 year of TESS observation (the maximum). The right column shows 27-days (the minimum). Based on this, the time series was chosen to transform the data in.

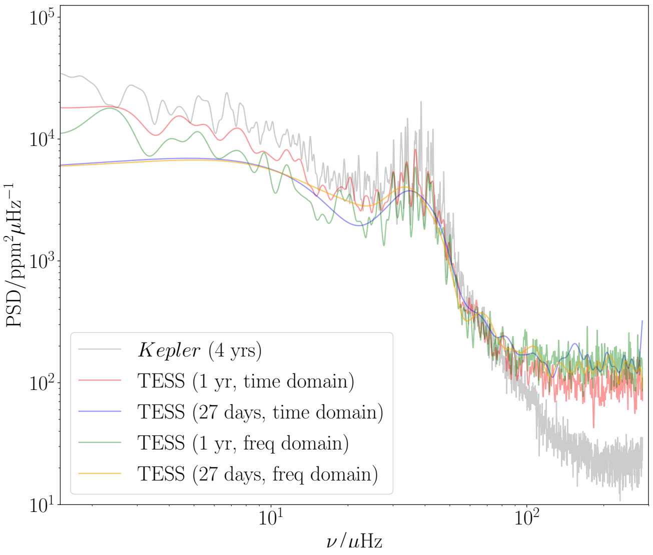

Figure 5. The Power Spectra of KIC 10587122. The original power spectra is in grey. The data was transformed into TESS observation and overplotted. The transformation was done in the time and frequency domains for comparison. Based on this, the time series was chosen to transform the data in.

where n is

$$n = 10^{-5.0} \times 10^{0.4\times(20-I_{\rm mag})}. \tag{4}$$

These three adjustments - the length of observation, the bandpass, and the noise level - were performed in the time and frequency domains for comparison. The methods used are outlined in Figures 6 and 7. The resulting Power Spectra are compared in Figures 4 and 5. From these results, it was decided to add noise in the time domain, before transforming to frequency.

After noise was added to the stars and power spectra were made, a detection test was run on every fitted mode of every star.

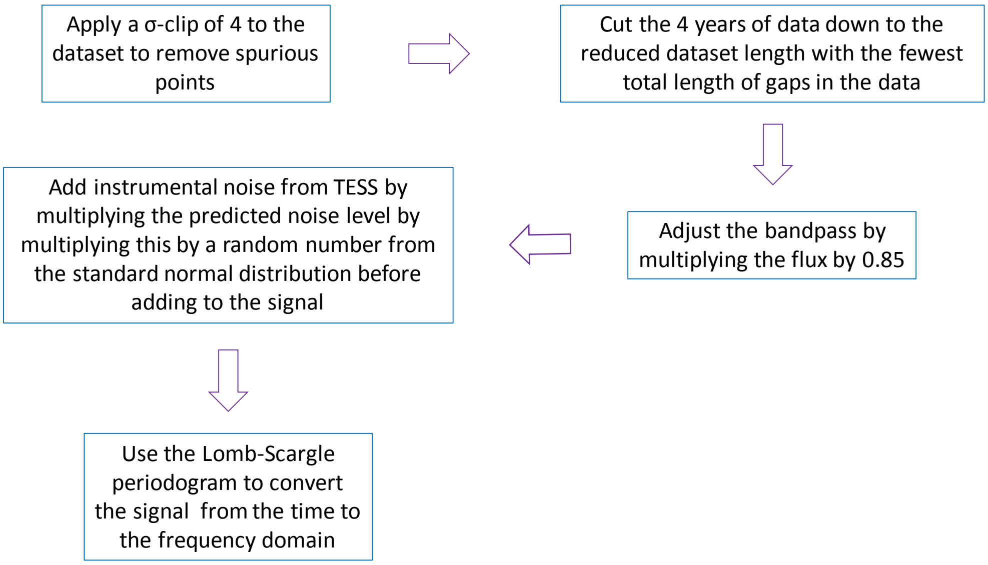
Apply a σ-clip of 4 to the dataset to remove spurious points
Cut the 4 years of data down to the reduced dataset length with the fewest total length of gaps in the data
Adjust the bandpass by multiplying the flux by 0.85
Add instrumental noise from TESS by multiplying the predicted noise level by multiplying this by a random number from the standard normal distribution before adding to the signal
Use the Lomb-Scargle periodogram to convert the signal from the time to the frequency domain

Figure 6. Flow chart of the method to convert the data from *Kepler* to TESS observations in the time domain.

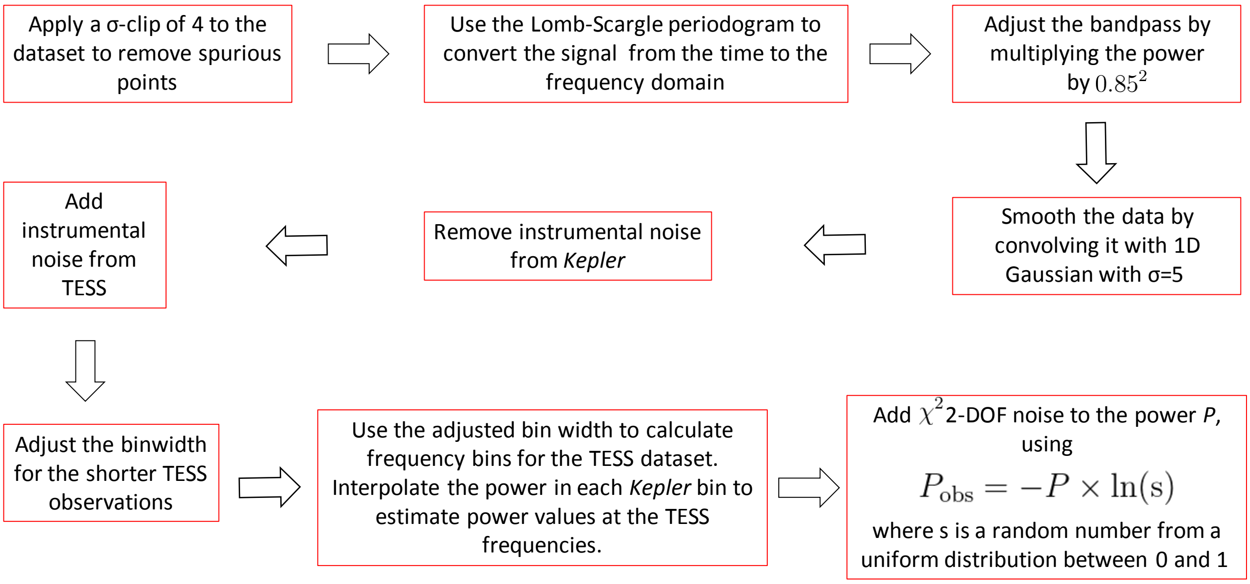
Apply a σ-clip of 4 to the dataset to remove spurious points
Use the Lomb-Scargle periodogram to convert the signal from the time to the frequency domain
Adjust the bandpass by multiplying the power by 0.85^2
Smooth the data by convolving it with 1D Gaussian with σ=5
Remove instrumental noise from *Kepler*
Add instrumental noise from TESS
Adjust the binwidth for the shorter TESS observations
Use the adjusted bin width to calculate frequency bins for the TESS dataset. Interpolate the power in each *Kepler* bin to estimate power values at the TESS frequencies.
Add χ^2 2-DOF noise to the power *P*, using
$P_{\rm obs} = -P \times \ln({\rm s})$
where s is a random number from a uniform distribution between 0 and 1

Figure 7. Flow chart of the method to convert the data from *Kepler* to TESS observations in the frequency domain.

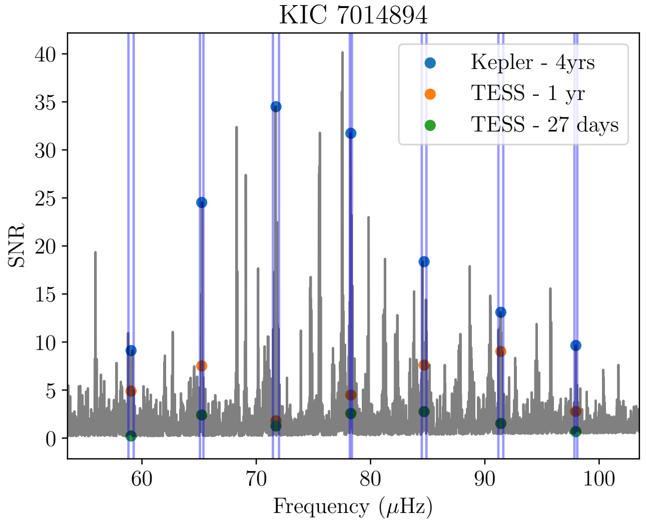

Figure 8. The power spectra of KIC 7014894 after background subtraction. The SNR value of every mode in the star was extracted from the SNR spectrum. The values for every mode in the original Kepler SNR are plotted in blue. The orange points are the SNR values after degrading the signal to 1 year of TESS observation. The green points are to 27 days.

3 DETECTION TEST

Section 2 described the method to transform the *Kepler* lightcurves into TESS observations. A detection test was then run on the stars, to determine which modes were still visible with observation by TESS, and which could not be recovered.

First, a moving median from Davies & Miglio (2016) was used. The width of the moving median is from the star's envelope width. This predicted envelope width was calculated as

$$\Gamma_{\rm env} = 0.66 * \nu_{\rm max}^{0.88}, \tag{5}$$

from Mosser et al. (2012). The moving median was used to interpolate between frequencies in the power spectrum. This moving median is a good estimate of the background of the star. It was divided out of the power spectrum to the get Signal-to-Noise spectrum,

$$\mathrm{SNR} = P/B. \tag{6}$$

An example of this is shown in Figure 8.

Once the SNR spectrum for the star was recovered, the SNR values of every mode were extracted. To ensure the correct SNR value was used, a window was fitted around each peak-bagged mode frequency. The size of the window was given as the linewidth of the mode. The highest value in the window was taken as the SNR of the mode. Figure 9 is an example of this process for one star.

Once all mode SNR values for the star were calculated, a detection test was run on every mode (Chaplin et al. 2011), (Campante et al. 2016).

The probability P that the SNR lies above some threshold $\mathrm{SNR_{thresh}}$ is

$$P(\mathrm{SNR} \geqslant \mathrm{SNR_{thresh}}) = p. \tag{7}$$

A false-alarm probability p of 1% was set; there is a 99% chance that the signal is due to solar-like oscillations, rather than noise. Equation 7 is solved for $\mathrm{SNR_{thresh}}$ by substituting P with

$$P = \int_x^\infty \frac{e^{-x}}{\Gamma(N)} x^{N-1} dx. \tag{8}$$

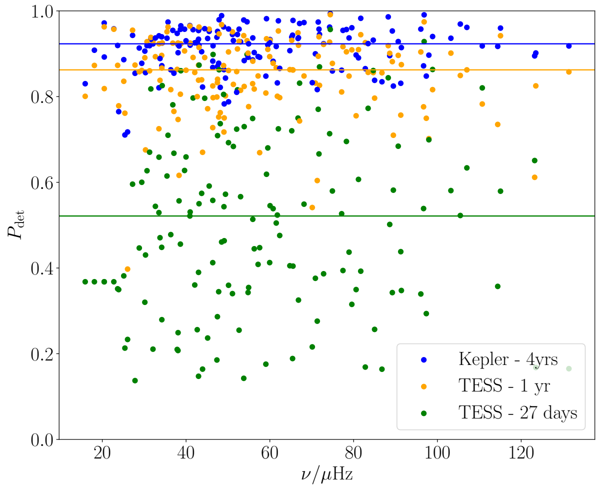

Figure 9. A plot showing the result of the detection test, after running on every mode in 20 stars. The results of the original power spectra are plotted in blue. The results of 1 year of TESS observation are in orange. 27 days of TESS observation is in green. At this short an observation, detecting individual modes will be extremely difficult.

$\Gamma(N)$ is the Gamma function. The lower bound of Equation 8 is set to $x = 1 + \mathrm{SNR_{thresh}}$. N is the number of frequency bins. The noise in this bin is assumed to follow $\chi 2$ $2n_{\rm bins}$ d.o.f statistics.

Once $\mathrm{SNR_{thresh}}$ is found, Equation 8 is solved again. This time it is solved for P by setting $x = (1 + \mathrm{SNR_{thresh}})/(1 + \mathrm{SNR})$. Here, SNR is the observed Signal-to-noise Ratio calculated from Equation 6. This calculates the probability that the frequency spike was due to stochastic excitation in the convective envelope of the star.

This detection test was applied to every mode of every star in the 1000-star sample. Classification was then applied to this dataset. This is described in Section 4.

4 CLASSIFICATION

Section 2 describes how the lightcurves of every star were treated before a detection test was run on the modes in Section 3. After this, Classification was applied to the stars to separate them into a suitable target list, and a list of stars that are not suitable for observation with TESS.

This could be done without using Machine Learning, although it takes much longer. In the age where datasets from missions such as MAST and Gaia (Gaia Collaboration et al. 2016) exist, tools need to be developed to make use of the huge amount of information we now have on stars from across the sky.

4.1 Preparing the data

Firstly, the detection probabilities of every mode were taken from Section 3. Each probability was changed to a 1 or 0 depending on if a detection was made. For each mode as observed by *Kepler*,

$$P_{\rm det, Kep} = \begin{cases} 1 & \text{if } P_{\rm det} \geqslant 0.9 \ , \\ 0 & \text{if } P_{\rm det} < 0.9 \ . \end{cases} \tag{9}$$

Modes will be more difficult to detect in TESS due to the larger white noise levels and shorter observation times. The threshold for determining a detection was therefore reduced;

$$P_{\mathrm{det,TESS}} = \begin{cases} 1 & \text{if } P_{\mathrm{det}} \geqslant 0.5 \quad , \\ 0 & \text{if } P_{\mathrm{det}} < 0.5 \quad . \end{cases} \tag{10}$$

Using equations 9 and 9, every mode was classified as detected (1) or undetected (0). The same three radial modes were used for every star: the mode closest to $\nu_{\mathrm{max}} l_n$, the radial mode one overtone below that l_{n-1}, and the radial mode above $\nu_{\mathrm{max}} l_{n+1}$. It was important to use the same information for every star so that the algorithm could learn the patterns between the variables.

A classifier is an algorithm that can learn a relationship between variables. The classifier will map from some initial information about the star (the X data), to some unknown information (the Y data). In this work, the X data were magnitude (K_p or I_{mag}), ν_{max}, $\Delta\nu$, T_{eff} and [Fe/H]. The Y data were the 3 radial modes for every star.

(Davies & Miglio 2016) had peak-bagged 1000 Red Giant stars. The more samples available, the better the classifier will be able to learn the relationship between variables. In order to increase the number of samples, the stellar magnitude was perturbed 100 times iteratively for every star. The noise level for the star was adjusted each iteration, and a detection probability was calculated for the modes. After removing gaps in the data, this left 60,000 samples.

The 60,000 samples were separated into a training dataset, and a testing set. 70% of the samples were used to train the classifier (46,410 stars); 30% was used to test the algorithm (19,890 stars). To train the Classifier, the X and Y data in the training set was given to the algorithm ($\mathrm{X_{train}}$ and $\mathrm{Y_{train}}$). Once the Classifier had been made, the X data from the testing set was given to it ($\mathrm{X_{test}}$). The Classifier then predicted a set of Y data for the testing set ($\mathrm{Y_{pred}}$). This was compared to the actual Y data for the testing set ($\mathrm{Y_{test}}$).

For the original *Kepler* sample, the Classifier produced a set of $\mathrm{Y_{pred}}$ data with a score of 0.93. The data was then modified for observation by TESS using the process described in Section 2, and the classifier was used again. It produced a set of $\mathrm{Y_{pred}}$ data with a score of ??????????

REFERENCES

Baglin A., et al., 2006. p. 3749, http://adsabs.harvard.edu/abs/2006cosp...36.3749B

Campante T. L., et al., 2016, preprint, 1608, arXiv:1608.01138

Chaplin W. J., Miglio A., 2013, Annual Review of Astronomy and Astrophysics, 51, 353

Chaplin W. J., et al., 2011, The Astrophysical Journal, 732, 54

Davies G. R., Miglio A., 2016, arXiv:1601.02802 [astro-ph]

Elorrieta F., et al., 2016, Astronomy and Astrophysics, 595, A82

Gaia Collaboration et al., 2016, Astronomy and Astrophysics, 595, A2

Howell S. B., et al., 2014, Publications of the Astronomical Society of the Pacific, 126, 398

Kremer J., Stensbo-Smidt K., Gieseke F., Steenstrup Pedersen K., Igel C., 2017, preprint, 1704, arXiv:1704.04650

Mosser B., et al., 2012, A&A, 537, A30

Nun I., Pichara K., Protopapas P., Kim D.-W., 2014, The Astrophysical Journal, 793, 23

Rauer H., et al., 2014, Experimental Astronomy, 38, 249

Ricker G. R., et al., 2014, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003

Valenzuela L., Pichara K., 2018, Monthly Notices of the Royal Astronomical Society, 474, 3259

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