# STA2453 Proposal

Yufei Liu

2025-01-22

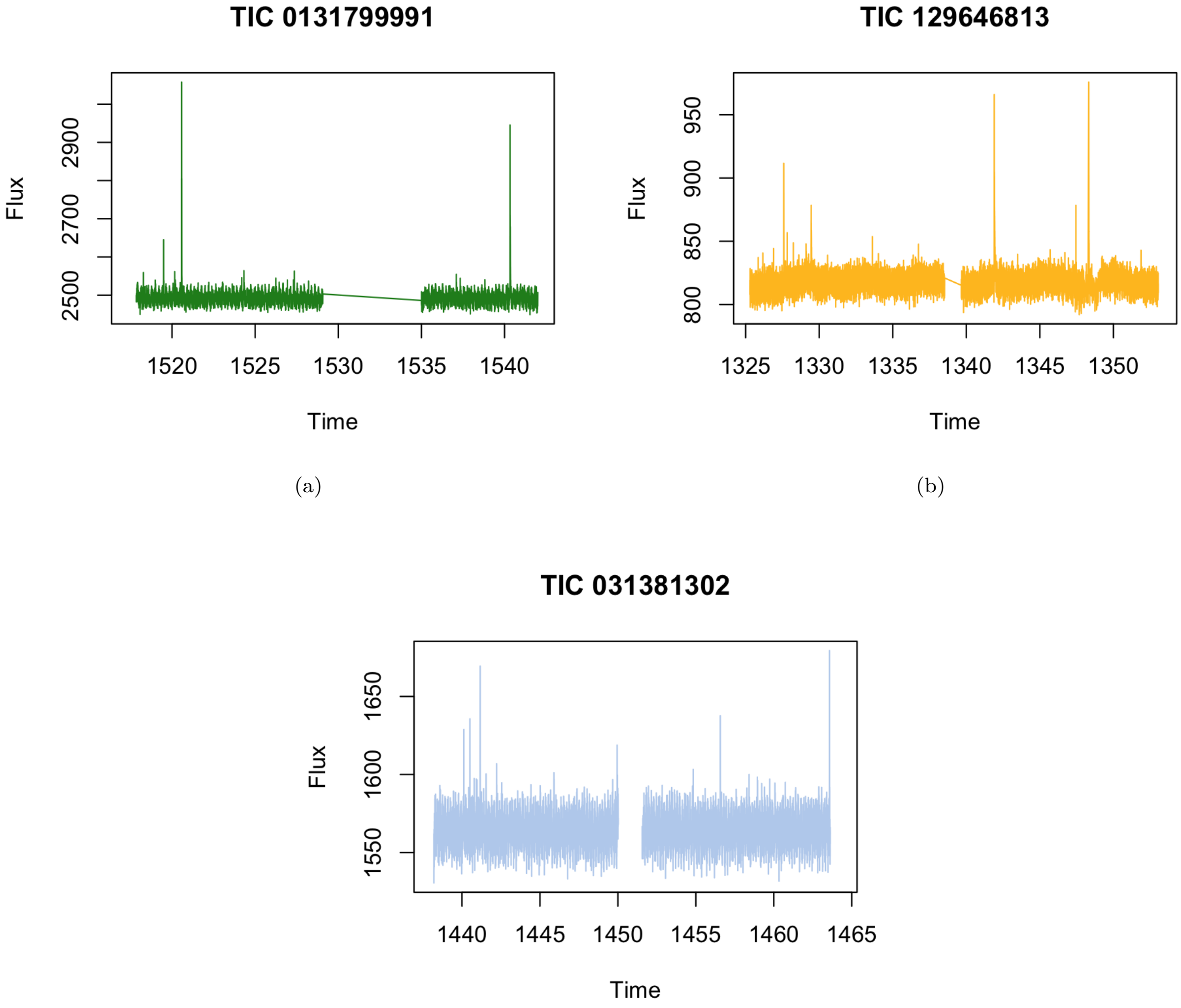
## 1 Proposal

For this project, we are interested in detecting potential stellar flares using Transiting Exoplanet Survey Satellite (TESS) data. Specifically, we are going to use the light curve, which is a plot of time versus flux as shown in Figure 1, to identify variability and possible flares for three stars, i.e. TIC 031381302, TIC 129646813, and TIC 0131799991. As it is a time series data, we will focus on two variables from the dataset, one is TIME in days and another is PDCSAP_FLUX, which indicates the variability. Time in days will be Barycentric Julian Date, which is the Julian Date been corrected for differences in the Earth's position with respect to the Solar System center of mass. PDCSAP flux refers to the Pre-search Data Conditioned Simple Aperture Photometry (SAP) flux, it is chosen instead of SAP flux since it is usually cleaner and may have fewer systematic trends because of detrending manipulations. The SAP flux was obtained by summing all pixel values in a pre-defined aperture as a function of time.

Since the dataset is unlabeled, i.e. we don't have predefined labels for flares and manually labelling process may lead to bias and costs a lot, we will consider unsupervised learning approached for the flare detection. As shown in Figure 1, for all 3 stars, the flux is stable for some time and there are very clear spikes, which refers to the sudden and sharp increase of the flux. So in this case, we will consider the unsupervised learning models for clustering and anomaly detection that could identify data points that are significantly different from the majority.

For the clustering approach, we plan to use the DBSCAN, which is a non-parametric algorithm that group data points by the density. It takes the advantages that it does not require specifying number of clusters, which is useful as we do not know the flare distributions, and also based on the time series plot we have, the most observations are flat without clear trends, which can be seen as similar density. It is also robust to the noise. Another model that could be considered is the isolation forest, which is a tree-based anomaly detection algorithm. It fits the data since it oriented to isolate anomaly, which aligns with our goal, in this case, flares. It also does not require specifying the distribution and robust to noise. To implement the methods, we plan to use Python to process lightcurve data, apply models, and validate. Since we have datasets for 3 stars, we may consider using one for training and assess the model performance on other. One challenge in the project would be missing values as we could see from the plot. To deal with irregular time series, we consider some smoothing techniques or other imputations.

For the next couple of week, we will start the EDA to find more data features and start to try different approaches. We may also need to consider the computational efficiency, which drives to choose the machine learning approaches instead of traditional time series models.



(a)

(b)

(c)

Figure 1: Time versus flux plots for the three stars.