STA2453 Proposal - Stellar Flares Detection

Isaac Baguisa

2025-01-29

Introduction

This project aims to detect stellar flares, and score the performance of the model to identify these events. Stellar flares are intense bursts of energy, emitted from a star that are thought to be caused by magnetic reconnection. They are usually indicated by a sudden increase in brightness, followed by a slower decay. Detecting and analyzing these flares is crucial for understanding stellar behaviour and its potential effects on its environments. The following questions are aimed to be answered in this project: Can non-parametric models effectively detect stellar flares from brightness time series data? How do these models score based on simulations? How can we validate the quality of simulations for testing the models?

Dataset and Variables of Interest

The dataset consists of time series data from the Transiting Exoplanet Survey Satellite (TESS) mission. Three stars (TIC 0131799991, TIC 031381302, and TIC 129646813) are used for analysis. The primary task involves detecting spikes in brightness, which indicate flares. Missing data occurs due to satellites turning off, and are not missing at random. To account for these gaps, interpolation methods will be used. The most important variables for this analysis are time, pdcsap_flux (Pre-Search Data Conditioning Simple Aperture Photometry - PDCSAP Flux), and pdcsap_flux_err, which represent the observation time, corrected photometric flux, and the associated uncertainty respectively. Additional variables, such as flux, sap_flux, and their errors, can be used for cross-validation.

Proposed Model Approaches

This project aims to detect stellar flares by implementing non-parametric approaches:

Kernel Density Estimation (KDE) is appropriate for unlabeled time series data because it estimates the underlying probability density function without assuming a specific distribution. This allows for the detection of flares as low-density regions in the flux data, indicating rare and significant deviations. By focusing on density anomalies, KDE can identify flare-like events in noisy datasets. KDE will estimate the probability density of the flux values, and flares will be identified as anomalies based on low-density regions in the data.

Gaussian Process Regression (GPR) is appropriate for time series data because it models the data as a continuous distribution over functions, providing both predictions and uncertainty estimates. This is useful for identifying flares as points with large deviations from the predicted trend, particularly when dealing with sparse observations. GPR's non-parametric properties ensures that it can adapt to the complex variability. GPR will model the brightness time series as a distribution over functions, providing both predictions and uncertainty estimates. Points with large deviations will highlight potential flares.

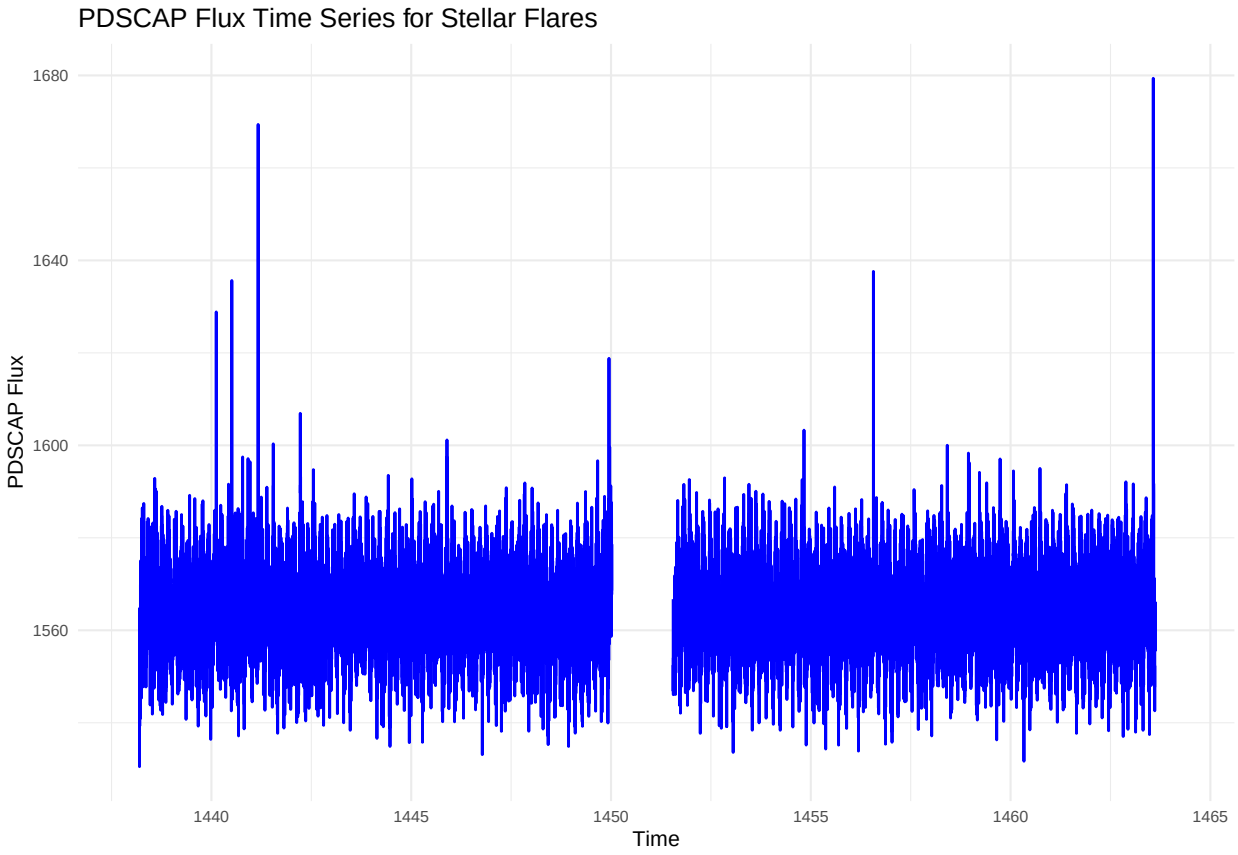
Loess (locally estimated scatterplot smoothing) and Residual Analysis is a non-parametric regression method that locally fits a curve to the data, capturing the underlying trend without assuming a global functional form. This is useful for time series with fluctuations and noise, as it highlights deviations from the smoothed curve. Residual analysis identifies flare events as points where observed flux significantly deviates from the smoothed trend. Loess will smooth the time series to extract the underlying trend. Residuals will highlight potential flares.

Proposed Scoring Based Method: Simulation-Based Validation

To address the challenge of working with unlabeled data, simulations will be used to validate the models. Simulated time series will replicate the statistical properties of the real data, including mean, variance, and noise characteristics. Synthetic flares will be inputted as sudden spikes with realistic intensity, duration, and decay patterns. Random noise will be added to simulate observational uncertainties. Flares in the synthetic data will be labeled 1 for flare, and 0 for non-flare, to be used as a benchmark for model performance. The quality of the simulations will be justified by comparing their statistical properties (e.g., distributions, autocorrelations) to the real data. Simulations generated using data from one or two stars (e.g., TIC 0131799991, TIC 031381302) will be used to test model performance on the third star (e.g., TIC 129646813). This approach evaluates the robustness of the models and of the simulations across different stars.

Additional simulations will be attempted by getting data of a star with no flares, and inputting synthetic flares, and testing the model based on these datasets. Classification scoring methods such as precision, recall, F1 score, and ROC-AUC will be used to assess the models' ability to detect synthetic flares.

Figures



Gantt Chart