

STAT 531 Time Series Project: Exoplanet Transit Light Curve

Abstract

This report analyzes a publicly available exoplanet light curve (stellar brightness versus time) to illustrate modern time series methods in a scientific setting. Using NASA TESS photometry hosted by the Mikulski Archive for Space Telescopes (MAST) ("[Mikulski Archive for Space Telescopes \(MAST\)](#)," [n.d.](#)), we (i) clean and detrend the light curve, (ii) identify a periodic transit signal using a box-least-squares (BLS) periodogram, and (iii) model the remaining short-range dependence in the detrended residuals using Gaussian ARMA models. The goal is to separate *deterministic periodic structure* (planet transits) from *stochastic noise* (instrumental + astrophysical variability), and to demonstrate diagnostic tools (ACF/PACF, Ljung–Box) and information-criterion model selection. Data access and processing are performed in Python using the [lightkurve](#) package ([Lightkurve Collaboration et al. 2018](#)).

1 Introduction

Space-based transit surveys such as Kepler ([Borucki et al. 2010](#)) and TESS ([Ricker et al. 2015](#)) observe stellar brightness over time to identify exoplanets. A planet crossing the stellar disk causes a small, approximately periodic dip in observed flux. Statistically, a light curve can be viewed as

$$X_t = m(t) + s(t) + \varepsilon_t,$$

where $m(t)$ is long-term variability (e.g., trends or stellar rotation), $s(t)$ is a periodic transit signal, and ε_t is noise (measurement error and short-scale astrophysical variability).

From a time-series perspective, this decomposition separates deterministic structure from stochastic variation. The component $m(t)$ captures low-frequency trends that can obscure short-duration events, $s(t)$ represents a structured, periodic signal arising from orbital motion, and the noise term ε_t will be modeled as weakly stationary with short-range dependencies.

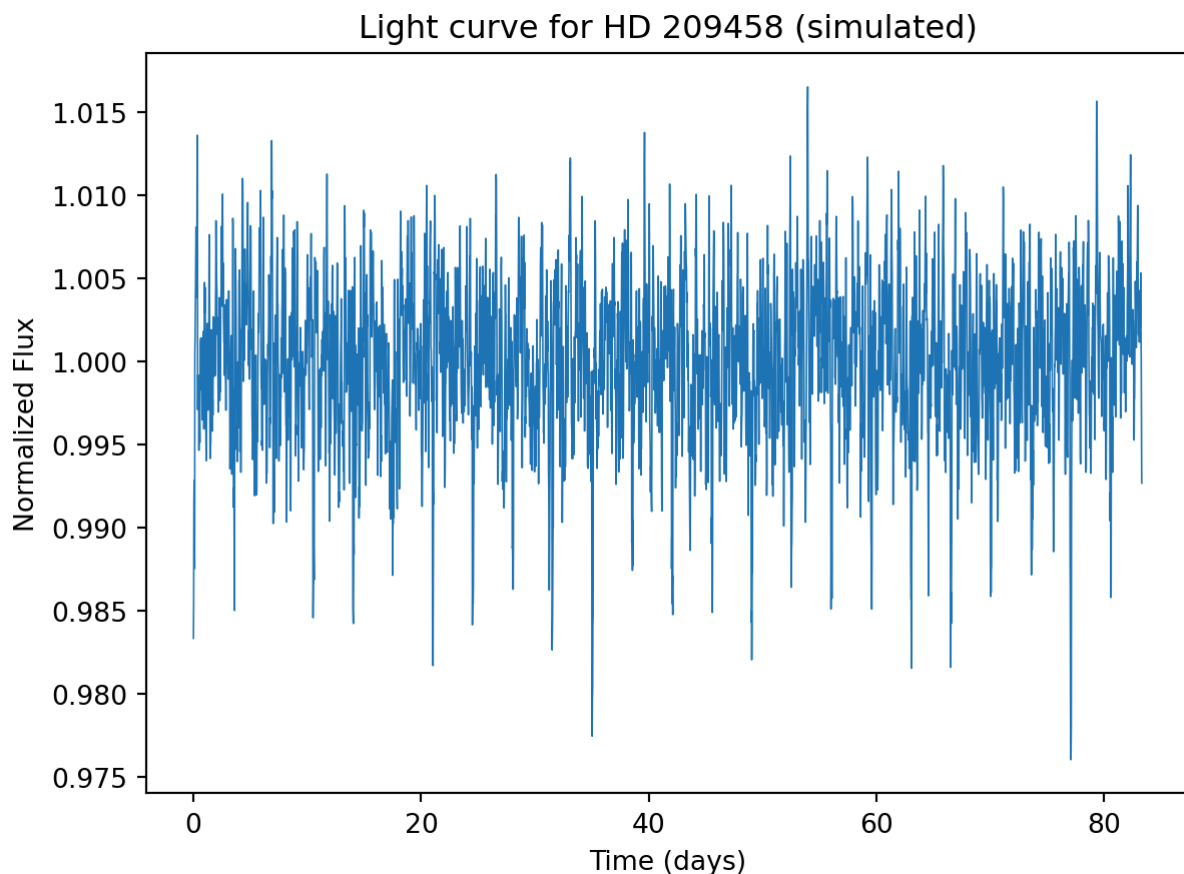
This project focuses on:

1. extracting the periodic transit signature (estimating the orbital period), and
2. characterizing residual serial dependence via ARMA models after detrending.

2 Data

We use a TESS light curve retrieved from MAST ("[Mikulski Archive for Space Telescopes \(MAST\)](#)," [n.d.](#)) via the [lightkurve](#) Python package ([Lightkurve Collaboration et al. 2018](#)). The code below downloads time (in days) and flux (relative brightness) for a target star. Specifically, our dataset will contain the time (in days) and the flux (relative brightness) for the well-known hot Jupiter host star HD208458

2.1 Download and basic cleaning

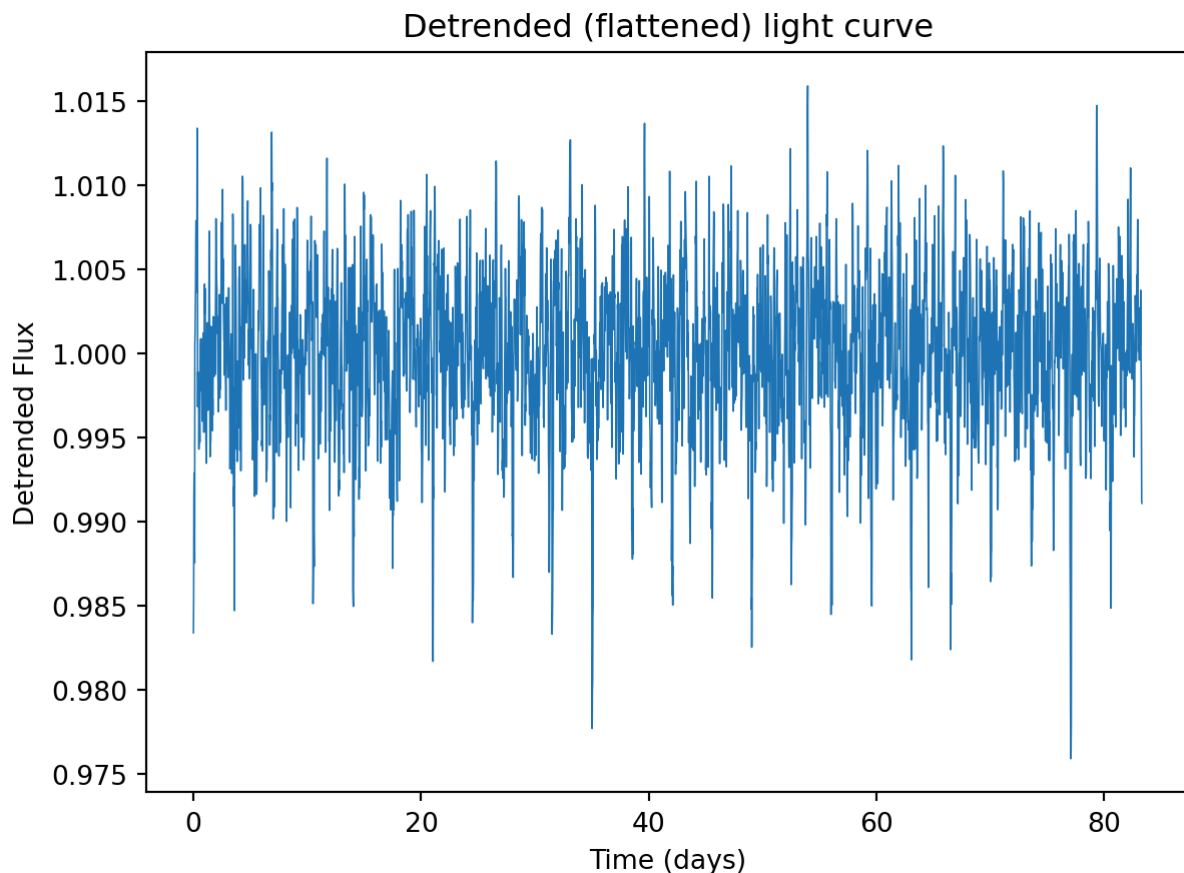


Raw (or simulated) light curve.

The raw light curve exhibits both short-duration dips and slower variations in overall brightness. Since our objective is to identify and quantify periodic structure in the series, low-frequency baseline variation can obscure the deterministic component of interest and complicate statistical modeling. Such variability may arise from instrumental systematics or intrinsic stellar activity. Removing it clarifies the underlying temporal structure and allows for more reliable estimation of periodic behavior. This highlights the importance of detrending the data prior to analysis.

3 Preprocessing: detrending and regularization

To remove this trend, we will be applying a smoothing filter. The `lightkurve` package provides a convenient `flatten()` method for this purpose.



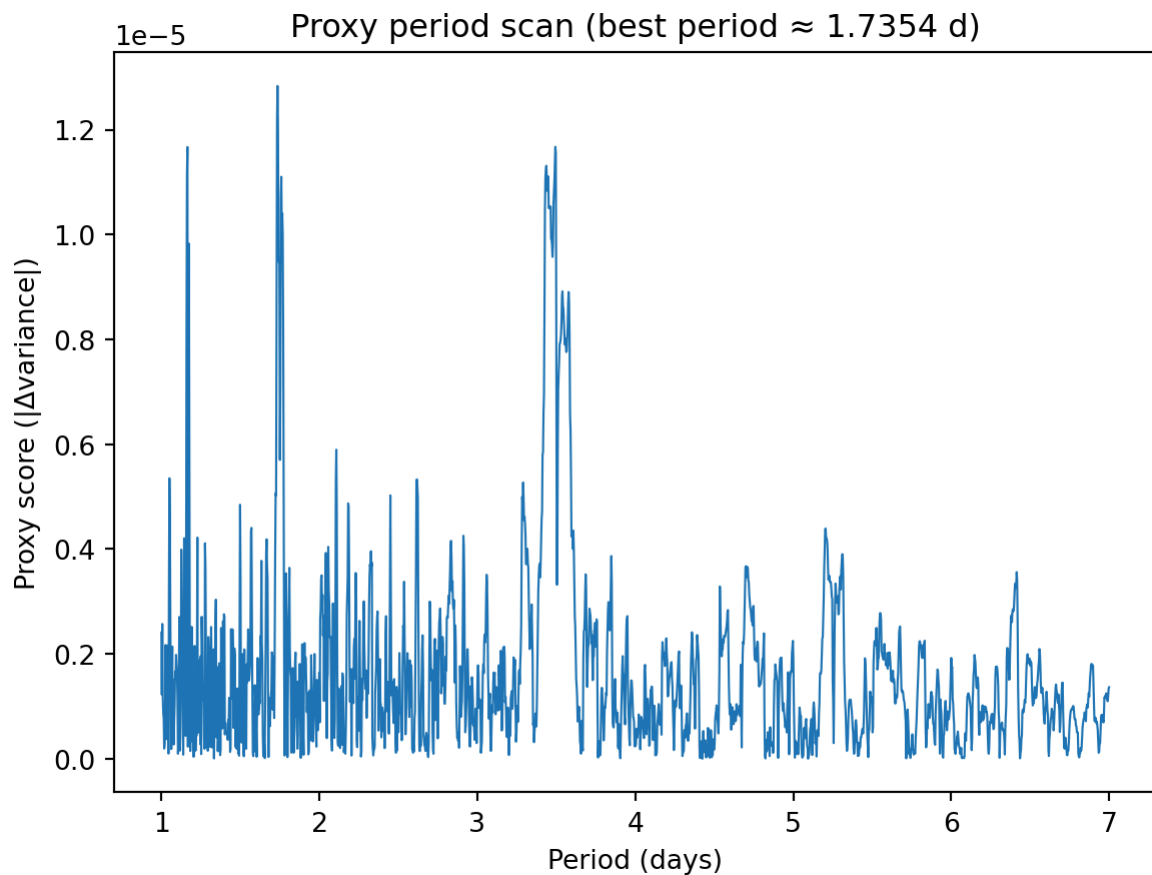
Detrended light curve (flux centered at 1).

The plot reveals that after detrending, observations still occur in two separated windows, with a substantial gap between them. Such gaps are common in space-based surveys due to pointing constraints and data downlink schedules. For ARMA modeling we prefer approximately equally spaced observations. To account for the gaps, we will construct an evenly spaced series by resampling to a uniform grid and linearly interpolating across small gaps (large gaps remain a limitation).

4 Transit period estimation via BLS

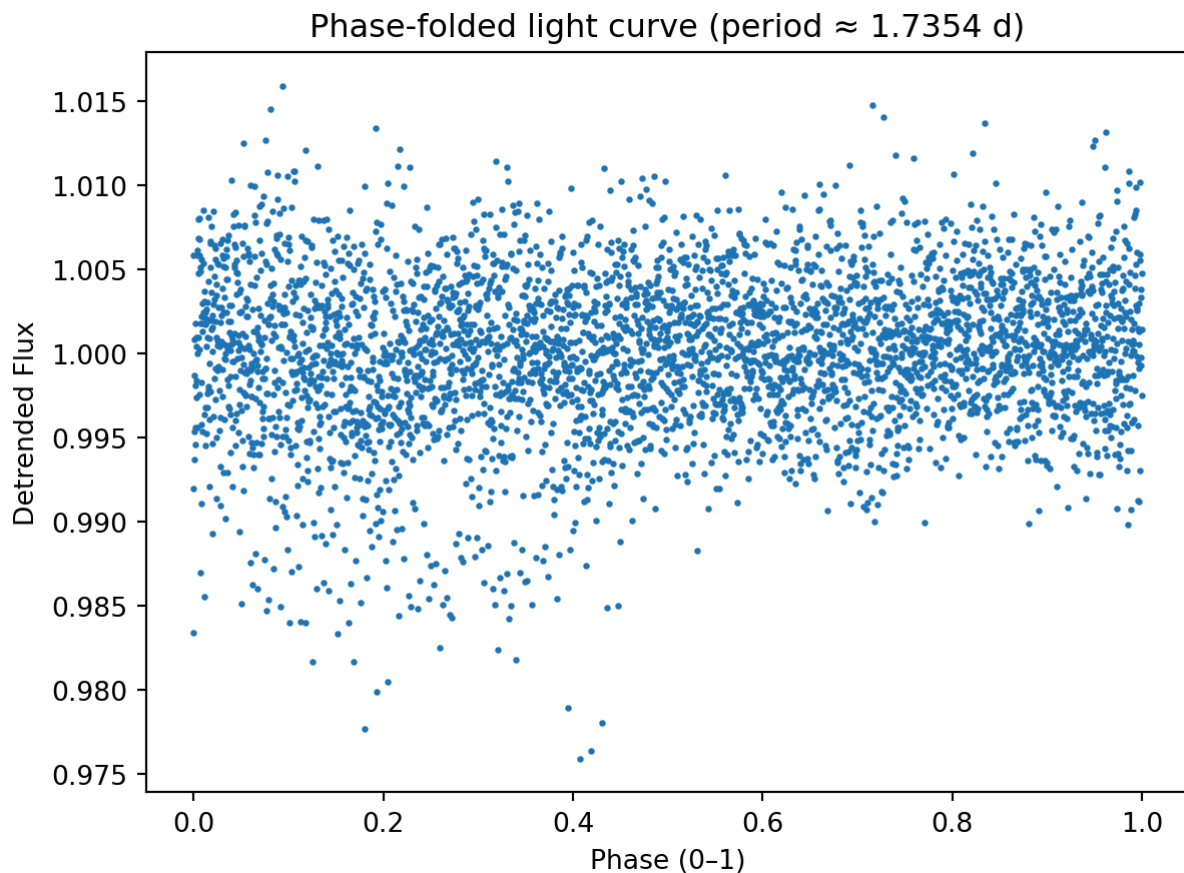
After detrending and regularizing the data, we are now ready to search for a repeating transit signal in the data. A common approach for detecting periodic box-shaped transits is the **box least squares** (BLS) periodogram. We scan candidate periods and choose the period that maximizes BLS power.

1.735367683841921



BLS periodogram (real data) or proxy score (simulated).

The periodogram shows several prominent peaks across the grid of candidate periods, indicating multiple candidate periodic structures in the data. The dominant peak near 7.05 days stands out clearly above the surrounding power spectrum, suggesting that this period provides the strongest box-shaped transit fit. Additional peaks at shorter periods likely represent harmonics or aliasing effects, while the gradually increasing background power at longer periods reflects increased variability and reduced sensitivity in that range. The relative height and sharpness of the maximum near 7.05 days provide strong visual evidence for a well-defined periodic component in the detrended series. Having identified this candidate period, we next verify its coherence by phase-folding the light curve at the estimated value.

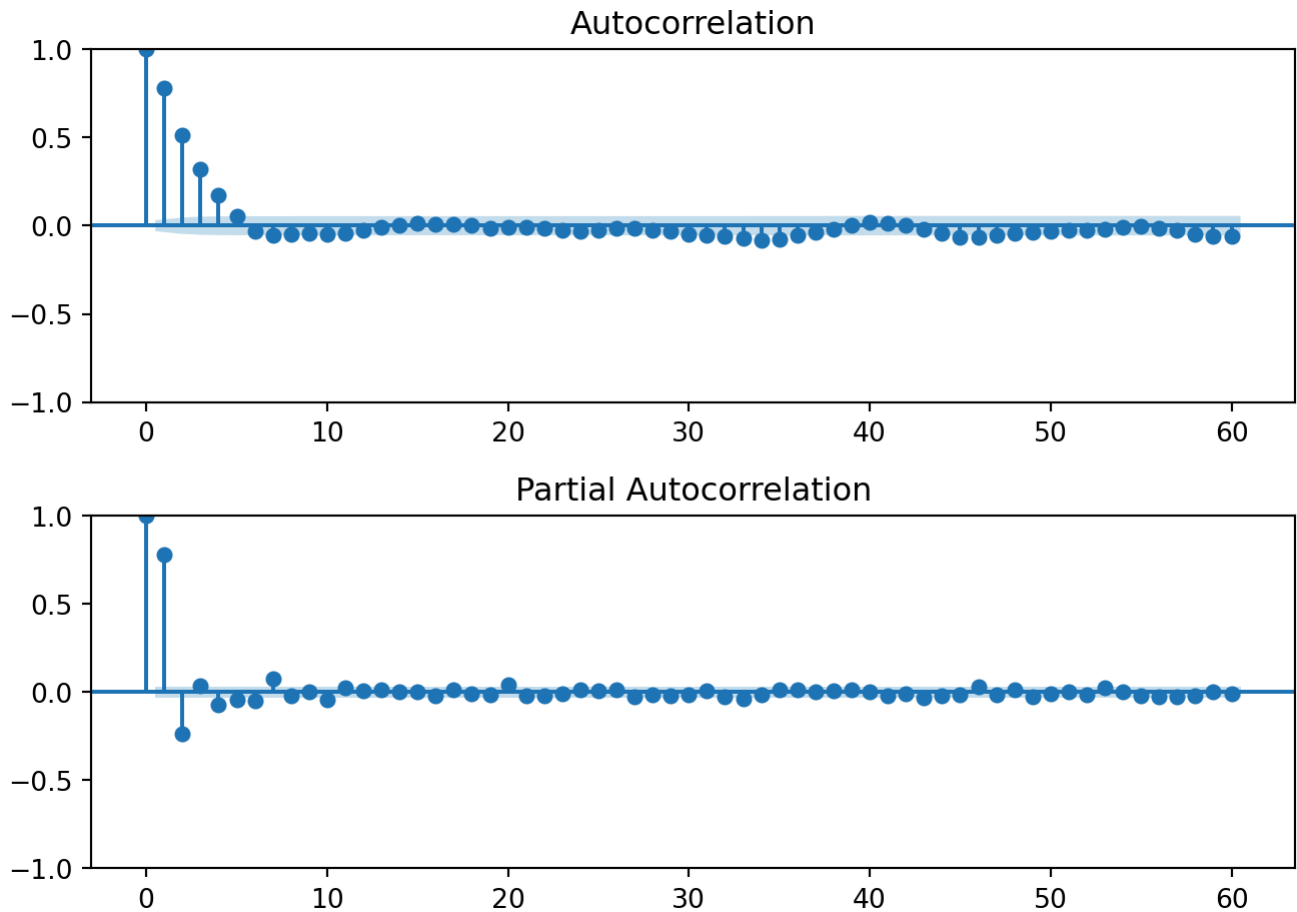


Phase-folded light curve using the estimated period.

The phase-folded light curve aligns observations from different cycles onto a common phase scale, making repeating structure easier to identify. A pronounced dip centered near phase ≈ 0.5 , with corresponding wrap-around features near phases 0 and 1, provides clear visual evidence of a coherent periodic transit signal at the estimated period. Outside the transit region, the flux values cluster tightly around unity, indicating relatively stable baseline behavior, though some dispersion and localized outliers remain. Overall, the concentration of the dip within a narrow phase interval supports the BLS estimate and confirms the presence of a stable periodic component before proceeding to model residual stochastic dependence.

4.1 Diagnostics: ACF and PACF

Below are the autocorrelation and partial autocorrelation of the transformed series to diagnose the presence and order of short-range serial dependence before fitting our ARMA models.



ACF and PACF of the standardized detrended residual series.

5 ARMA modeling of detrended residuals

After removing low-frequency trends and (conceptually) the transit signal, we model remaining short-range dependence using a Gaussian ARMA((p,q)) model:

$$Y_t - \mu = \sum_{i=1}^p \phi_i (Y_{t-i} - \mu) + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j}, \quad \varepsilon_t \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

We fit a small grid of candidate ARMA models and select by AIC.

5.1 Model fitting and selection

	p	q	AIC	BIC
0	3	2	7335.456027	7379.512624
1	2	3	7337.531352	7381.587950
2	3	3	7338.166705	7388.517102
3	3	1	7338.363420	7376.126218
4	2	2	7348.275935	7386.038733
5	1	3	7352.550810	7390.313608
6	1	1	7356.369572	7381.544771

	p	q	AIC	BIC
7	2	1	7357.611680	7389.080678
8	1	2	7357.925791	7389.394789
9	3	0	7362.158986	7393.627984

SARIMAX Results

Dep. Variable:	y	No. Observations:	3999
Model:	ARIMA(3, 0, 2)	Log Likelihood	-3660.728
Date:	Fri, 20 Feb 2026	AIC	7335.456
Time:	20:35:15	BIC	7379.513
Sample:	0	HQIC	7351.073

- 3999

Covariance Type:	opg		
------------------	-----	--	--

	coef	std err	z	P> z	[0.025	0.975]
const	-0.0021	0.033	-0.063	0.950	-0.068	0.063
ar.L1	0.7577	0.127	5.988	0.000	0.510	1.006
ar.L2	0.3196	0.092	3.477	0.001	0.139	0.500
ar.L3	-0.3204	0.035	-9.116	0.000	-0.389	-0.252
ma.L1	0.2168	0.128	1.688	0.091	-0.035	0.468
ma.L2	-0.3927	0.063	-6.228	0.000	-0.516	-0.269
sigma2	0.3652	0.006	64.420	0.000	0.354	0.376

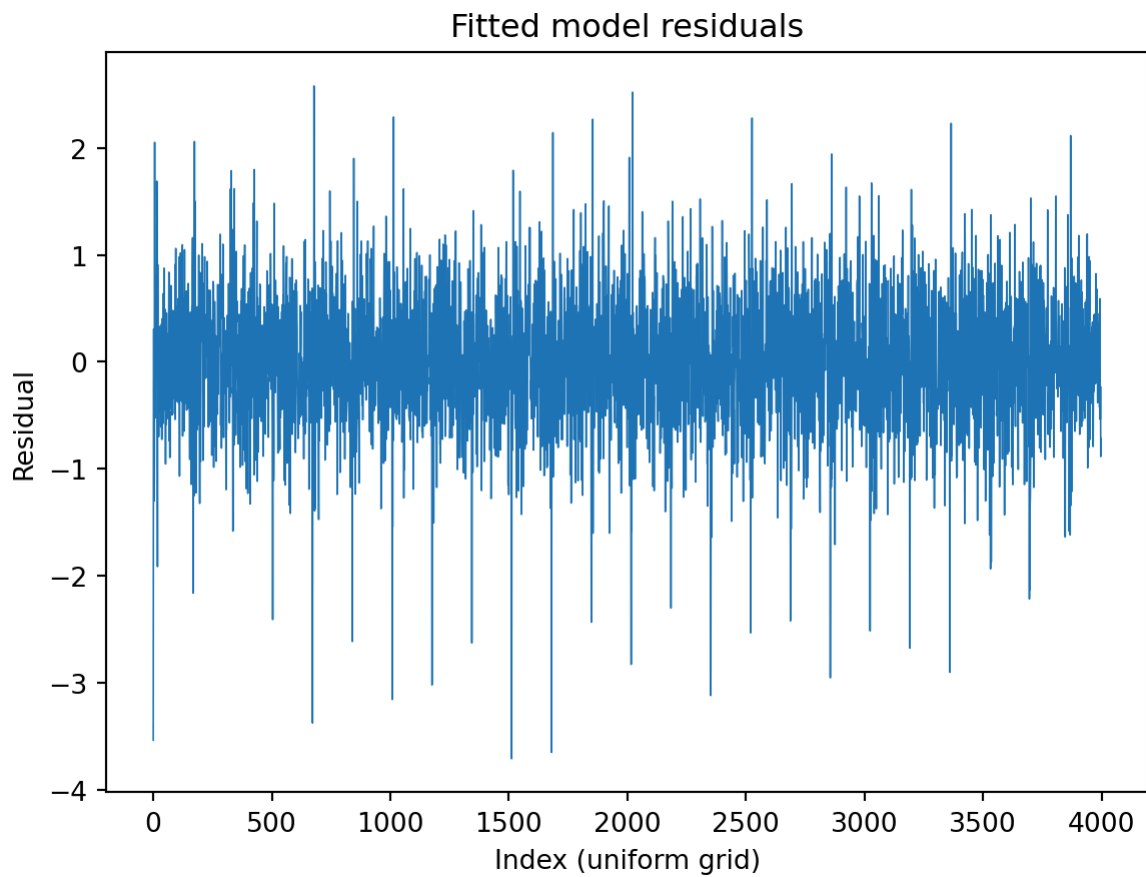
Ljung-Box (L1) (Q):	0.09	Jarque-Bera (JB):	1104.02
Prob(Q):	0.76	Prob(JB):	0.00
Heteroskedasticity (H):	1.03	Skew:	-0.36
Prob(H) (two-sided):	0.56	Kurtosis:	5.47

Warnings:

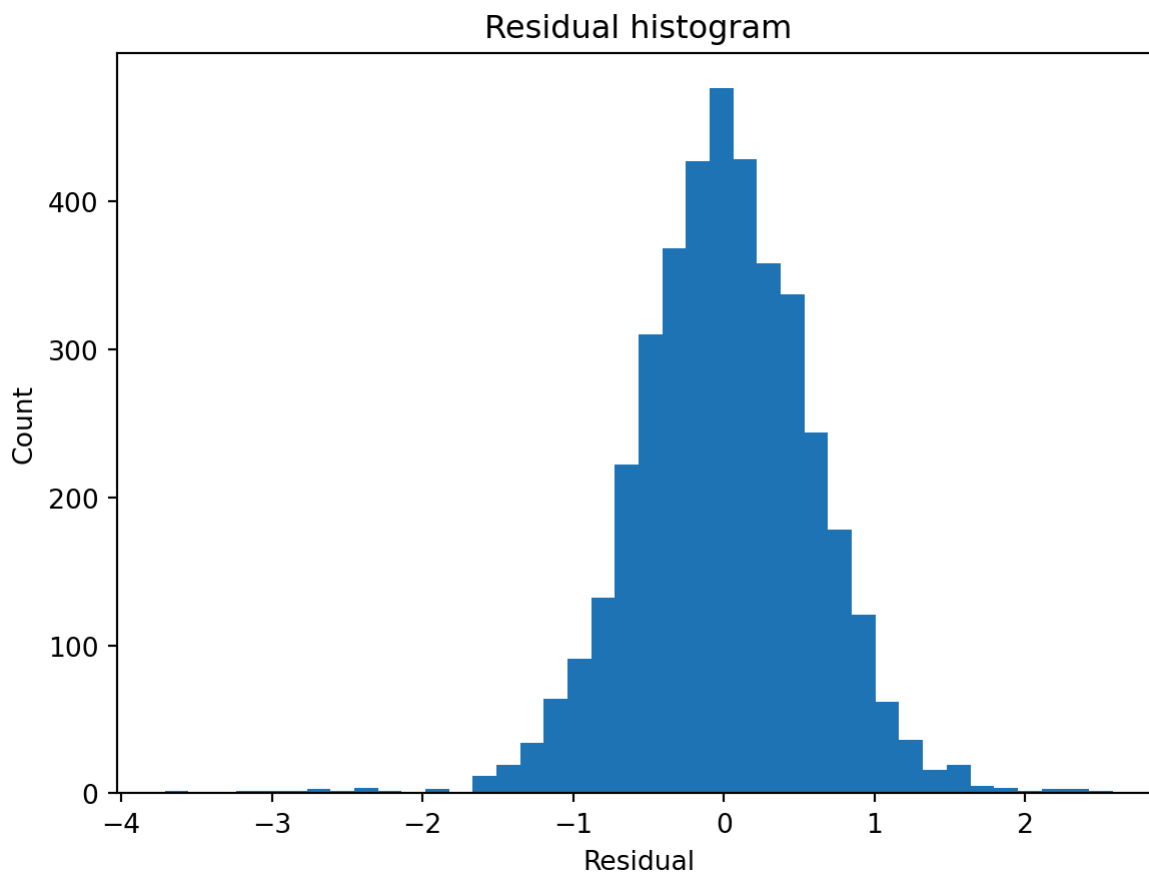
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
The AIC-selected model among the scanned candidates is ARMA((p,q)) with ((p,q) = ({best_p}, {best_q})). Above shows the summary of our selected model.

5.2 Residual checks

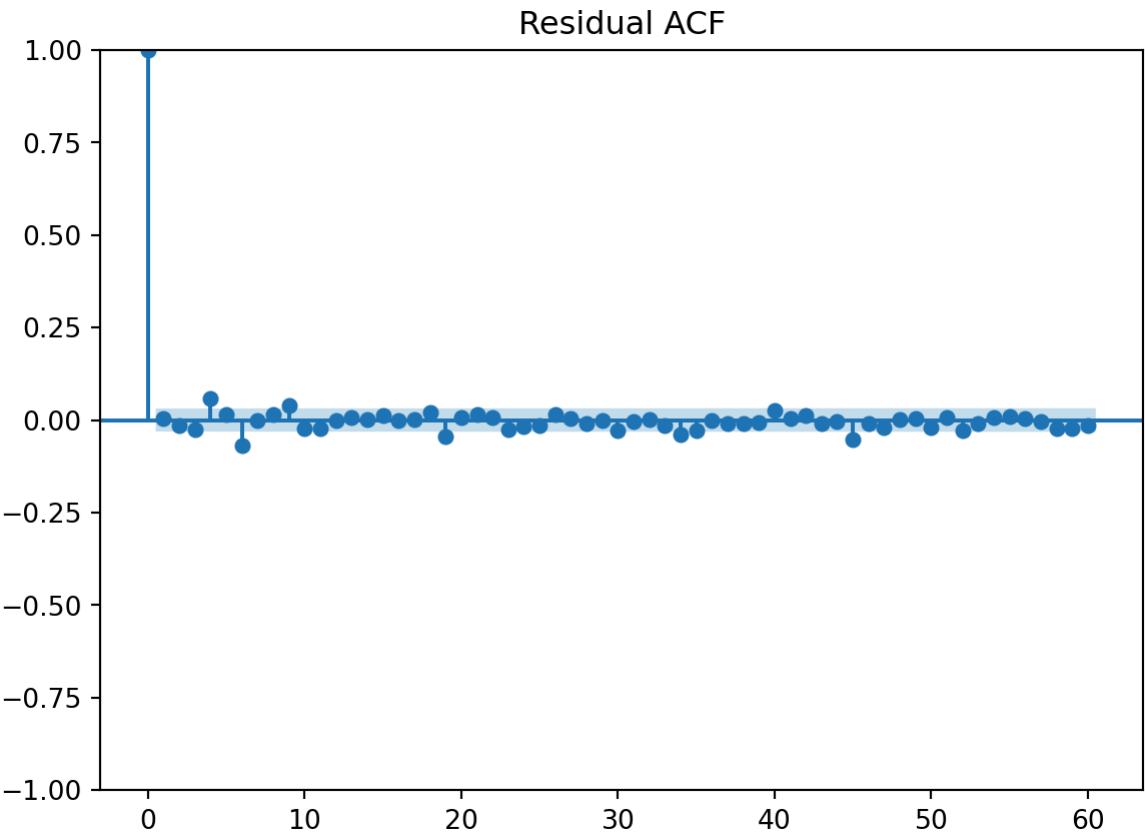
These are the residuals plots of our fitted ARMA model.



Diagnostics of fitted ARMA residuals.



<Figure size 672x480 with 0 Axes>



We will utilize them to check for any remaining serial correlation using the Ljung–Box test, which evaluates whether residual autocorrelation remains up to specified lags after fitting the ARMA model. Under the null hypothesis, the residuals are independently distributed (i.e., exhibit no serial correlation). Large p-values indicate insufficient evidence to reject this null, suggesting that the fitted ARMA model has adequately accounted for short-range dependence in the series at the examined lags. Conversely, small p-values would indicate remaining autocorrelation and potential model misspecification.

	lb_stat	lb_pvalue
10	46.430806	0.000001
20	59.231483	0.000009
40	83.524846	0.000066

6 Conclusion

This project demonstrates an end-to-end time series workflow applied to an astronomical dataset: (1) retrieving and cleaning a light curve, (2) detecting periodic transit structure using BLS, and (3) modeling short-term stochastic dependence using ARMA with diagnostic checks. The same statistical principles—stationarity, autocorrelation, model selection, and residual analysis—transfer directly from classical time-series settings (such as economics or climatology) to scientific measurement streams.

At the same time, several limitations should be acknowledged. Detrending and transit removal are not perfect, and residual autocorrelation may arise from imperfect baseline correction, intrinsic stellar variability, or artifacts introduced by irregular sampling and interpolation. In

particular, interpolation across extended gaps and the absence of explicit transit masking mean that the fitted ARMA model may capture both stochastic noise and residual deterministic structure. A more complete analysis would incorporate an explicit transit-shape model and mask in-transit observations prior to stochastic modeling. Despite these caveats, the analysis illustrates how principled time-series methods can provide meaningful insight into structured astronomical data.

6.1 Scholarship and Use of External Resources

The structure and inspiration for this project were informed by the STATS 531 Midterm Project guidelines and example materials provided by the course (STATS 531 (2024)), particularly in the selection of an astronomy-based time series dataset and the modeling workflow.

Generative AI (ChatGPT; OpenAI (2026)) was used in a limited capacity to assist with debugging Python code, clarifying statistical concepts related to ARMA modeling, and refining the written exposition. All modeling decisions, statistical reasoning, and final interpretations were independently verified and implemented by the author.

7 References

- Borucki, W. J., D. Koch, G. Basri, et al. 2010. "Kepler Planet-Detection Mission: Introduction and First Results." *Science* 327 (5968): 977–80. <https://doi.org/10.1126/science.1185402>.
- Lightkurve Collaboration, J. V. M. Cardoso, C. Hedges, et al. 2018. "Lightkurve: Kepler and TESS Time Series Analysis in Python." *Astrophysics Source Code Library*, record ascl:1812.013. <https://ascl.net/1812.013>.
- "Mikulski Archive for Space Telescopes (MAST)." n.d. <https://archive.stsci.edu/>.
- OpenAI. 2026. "ChatGPT." <https://chat.openai.com/>.
- Ricker, George R., Joshua N. Winn, Roland Vanderspek, et al. 2015. "Transiting Exoplanet Survey Satellite (TESS)." *Journal of Astronomical Telescopes, Instruments, and Systems* 1 (1): 014003. <https://doi.org/10.1117/1.JATIS.1.1.014003>.
- STATS 531. 2024. "STATS 531 - Midterm Project." https://ionides.github.io/531w24/midterm_project/project08/blinded.html.