AutoRegressive Planet Search

A new statistical approach to exoplanet transit detection

Eric Feigelson

with Gabriel Caceres, Andrew Stuhr,

G. Jogesh Babu & colleagues

Center for Astrostatistics

Penn State University

NARIT-EACOA Summer Workshop on Astrostatistics & Astroinformatics

August 2019

Outline

1. The problem: Stellar variability impedes planet detection

2. A solution: Parametric stochastic autoregressive models to reduce variability.

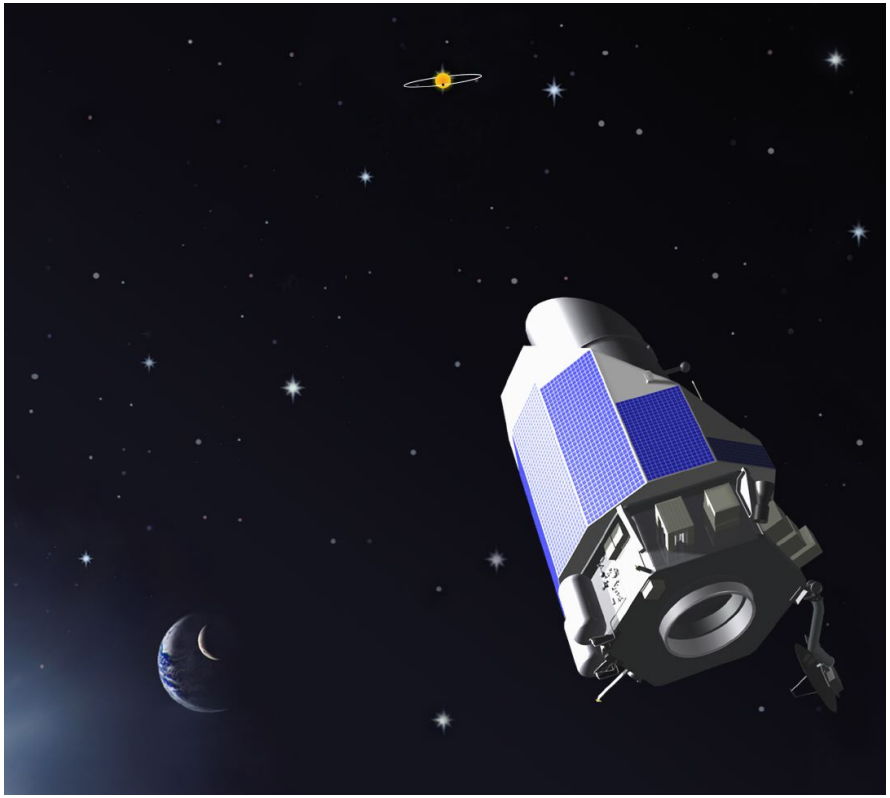
3. ARIMA and the ARPS methodology

4. ARPS applications to ~200,000 Kepler stars

Transiting planet observatories: Space-based and ground-based

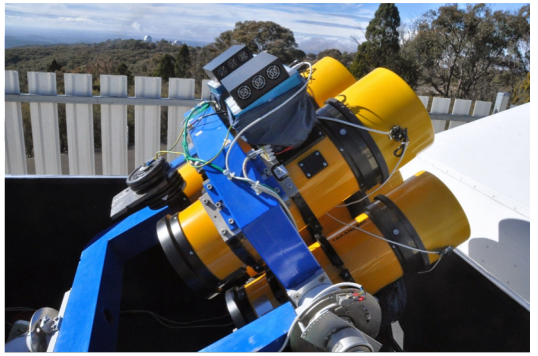
Kepler

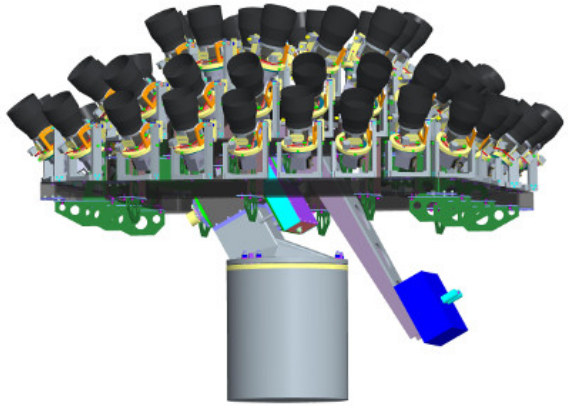
~2300 confirmed planets



also CoRot, TESS, Plato

HATNet
Hungarian Automated Telescope Network





Telescopes in Chile, Namibia, Australia

~100 confirmed planets

also WASP, OGLE, TrES, XO, Qatar, KELT, AST3

Stellar variability
A major problem for planet detection

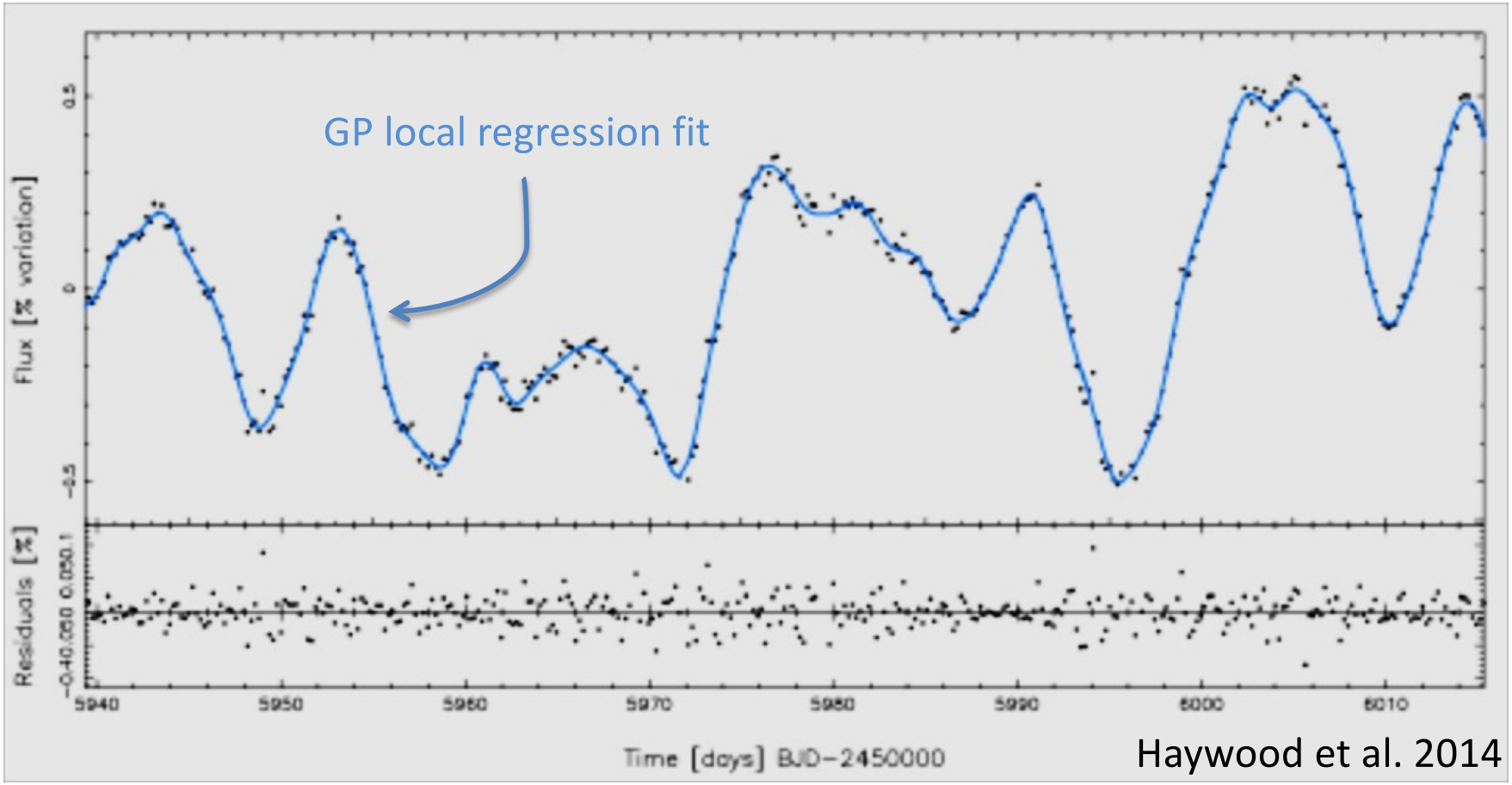
The discovery of planets in astronomical radial velocity or photometric time series (`light curves') involves:

- Time domain suppression of variations intrinsic to the star

- Frequency domain periodogram to reveal periodic signature of planetary orbit

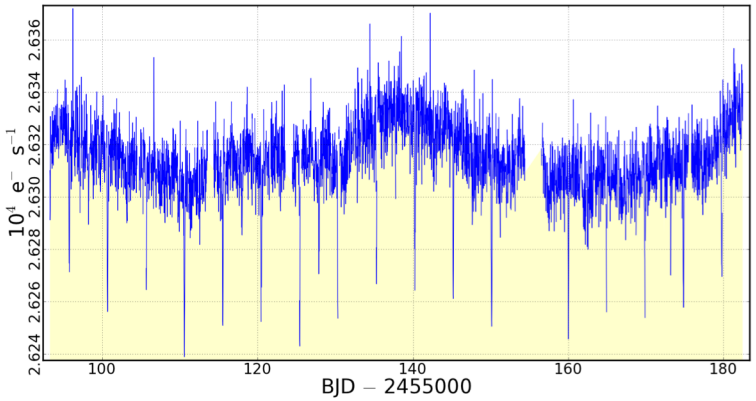
"[E]xoplanets may still be detected by exploiting differences in timescale, shape and wavelength dependence between the planetary and stellar signals. ... [Stellar] variability, combined with residual instrumental systematics, is still limiting the detection of habitable planets by Kepler." (Suzanne Aigrain IAU 2015)

For ***space-based observatories*** with ~0.01 mmag photometric precision, the variations are mostly from stellar magnetic activity

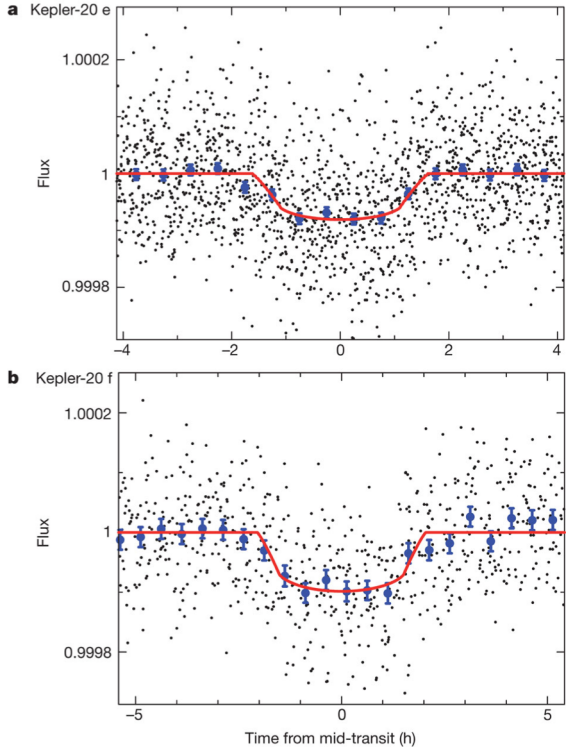


Magnetic activity is well-studied on our Sun and strongly magnetically active stars. But until the Kepler mission, it was not recognized how stellar magnetic activity is so pervasive in normal stars.

Sometimes it is easy to find the recurring transits ….



… but other times it is hard but important

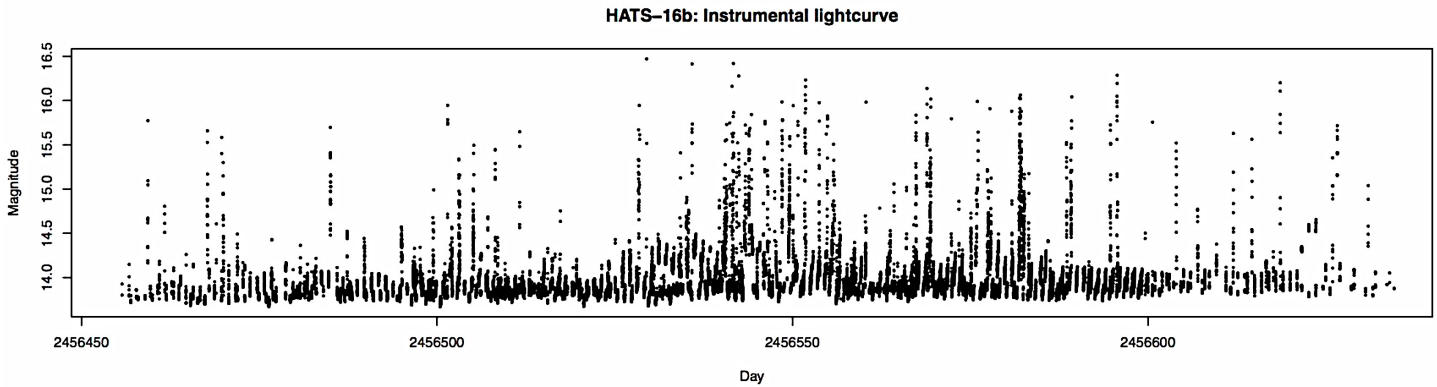


Two Earth-sized planets orbiting Kepler-20

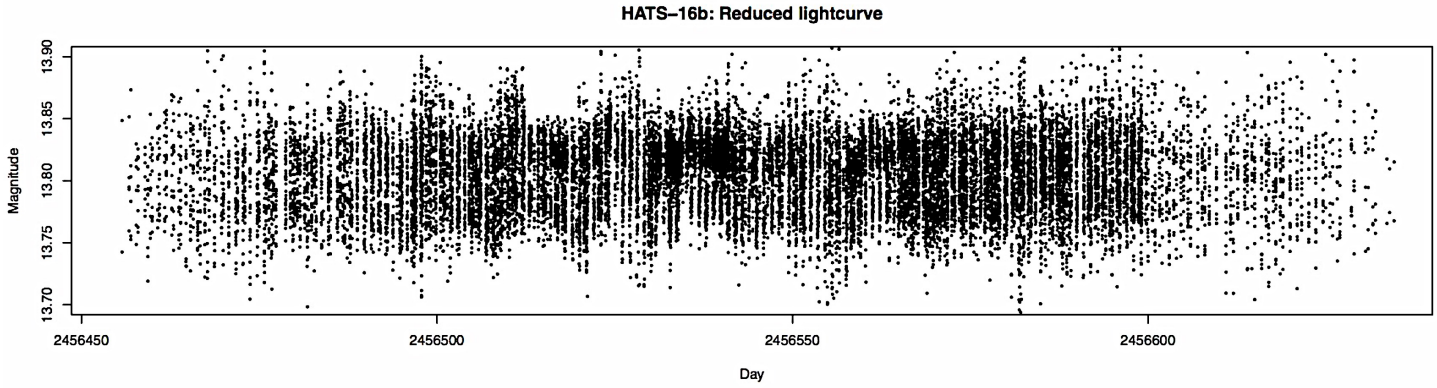
Fressin et al. 2012 Nature

Ground-based light curves are mostly dominated by atmospheric & instrumental variations

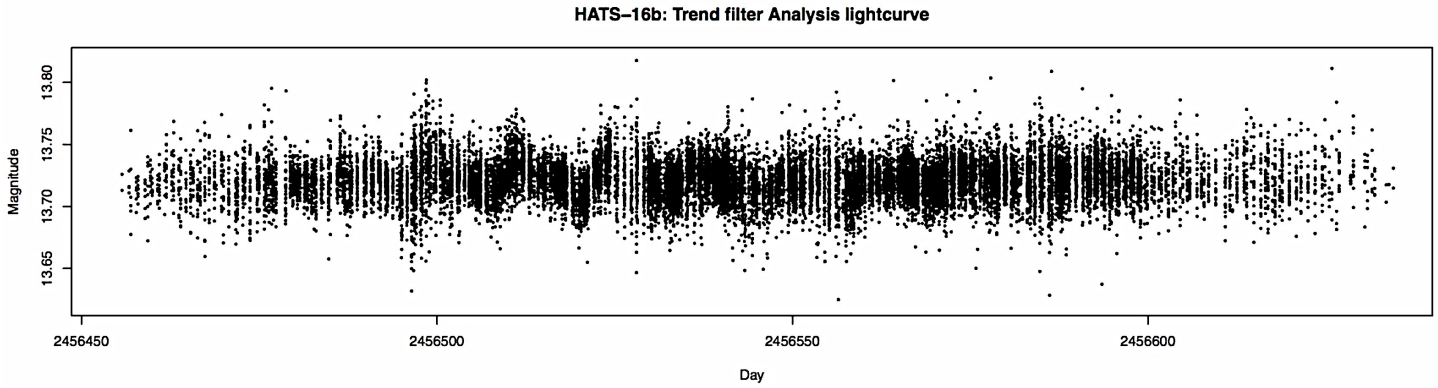
HAT-S raw lightcurve



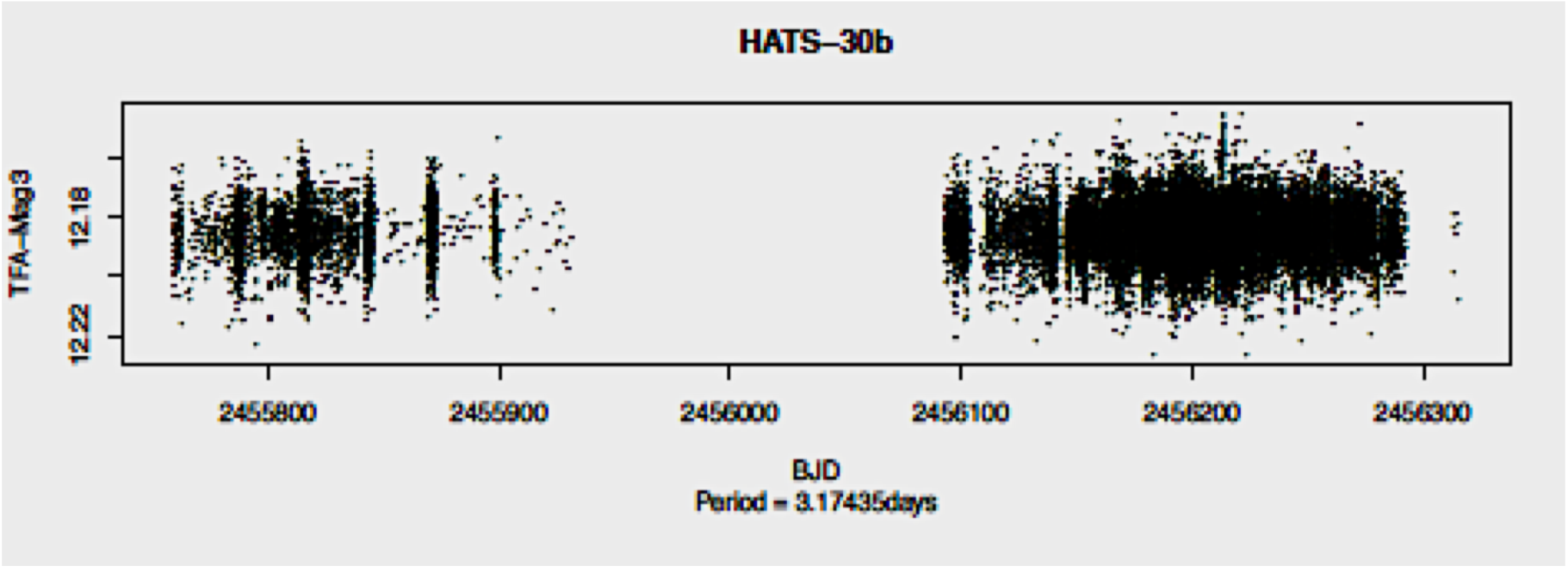
After instrumental effects are reduced



After atmospheric effects are reduced (TFA)



Complex observing cadences for HAT-S survey



Cadences for other ground-based surveys are typically sparser and more irregular … more difficult to find periodicities.

Stellar variability reduction methods

Nonparametric modeling

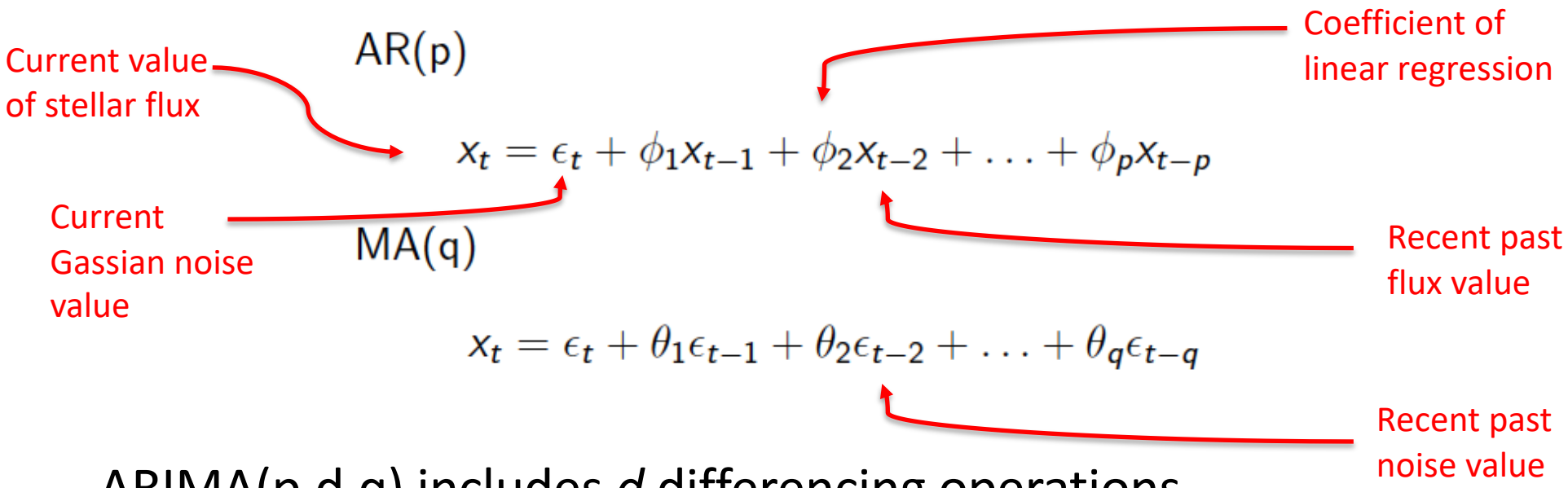
- *Wavelet analysis* (Jenkins 2002 & Kepler pipeline, Carter & Winn 2009)
- *Gaussian Processes regression* (Gibson 2014, Aigrain et al 2016, Luger et al 2016)
- Advanced signal processing methods: *Independent Component Analysis, correntropy, trend entropy, Empirical Mode Decomposition, Singular Spectrum Analysis, …* (Waldmann et al 2013, Huijse et al 2012, Roberts et al 2013, Greco et al 2016, Boufleur et al 2018, …)

Parametric modeling

Rarely used because stellar & atmospheric variations do not follow any obvious function: Flux ~ f(time). However, textbooks in time series analysis are dominated by ***stochastic autoregressive regression models***,
Flux ~ f(past fluxes, past changes). These are broad model families widely used in engineering signal processing, econometrics, and other fields (millions of Google hits).

We have found these models are also very effective for time domain astronomy!!

Autoregressive modeling for evenly spaced time series



ARIMA(p,d,q) includes *d* differencing operations

$$x'_t = x_t - x_{t-1}$$

ARFIMA(p,d,q) has *fractional differencing*

CARMA (continuous), VARIMA (vector), SARIMA (periodic), GARCH and dozens of other extensions to the ARMA family

Roughly speaking …

- AR & MA components model short-memory processes
- I differencing operator removes many forms of trend and non-stationarity
- F component models long-memory processes and is equivalent to the astronomers' $1/f^{\alpha}$-type red noise. The coefficient *d* is arithmetically related to α and the economists' Hurst parameter.

The ARIMA and ARFIMA models can remove an enormous variety of temporal variations seen in astronomical data

Methodological procedures

1. ARIMA-type models are fit by maximum likelihood estimation (MLE) giving a unique solution for a chosen order (p,d,q)
2. Order selection (model complexity) performed using the penalized likelihood measure, Akaike Information Criterion
3. Residual analysis to evaluate model adequacy: Is the best-fit model a good fit? Are the residuals consistent with Gaussian white noise? Tools: Autocorrelation function, Ljung-Box test, augmented Dickey-Fuller test, Anderson-Darling test, etc

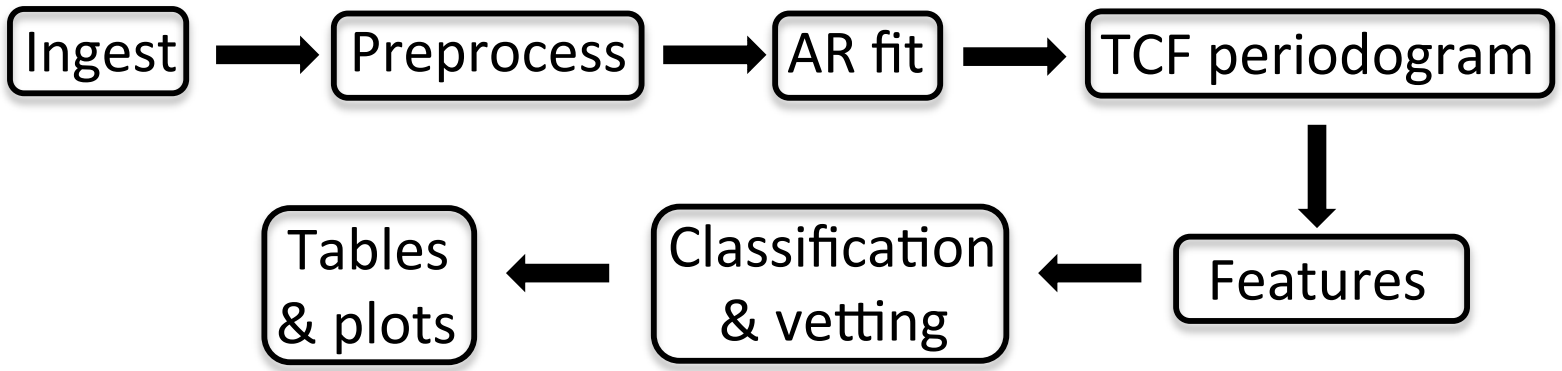
These MLE methods have no free parameters

e.g. choice of smoothing bandwidth or kernel

ARPS: Four steps to exoplanet transit detection

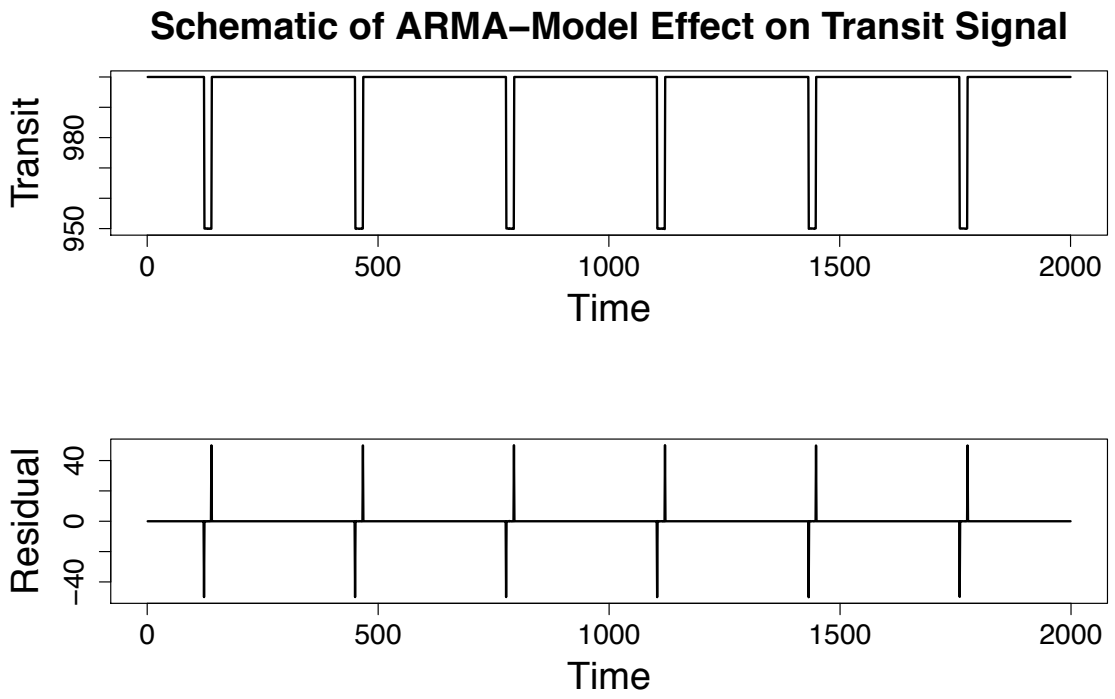
1. Pre-process the data
2. Reduce/remove uninteresting stellar variability but …

   "Don't eat the planet!" *David Jones, SAMSI*
3. Conduct periodicity search for recurring transits
4. Establish decision criteria to report *Planet Candidates*



Problem:

The differencing operator transforms box-shaped transit signal into a periodic double-spike signal



To find the planetary signal in the ARIMA/ARFIMA residuals, we convolve the residual time series with a matched filter for a periodic double-spike pattern we call the ***Transit Comb Filter*** (G. Caceres). Planetary signals should appear as peaks in a periodogram based on the TCF.

Caceres, Feigelson et al. 2019a

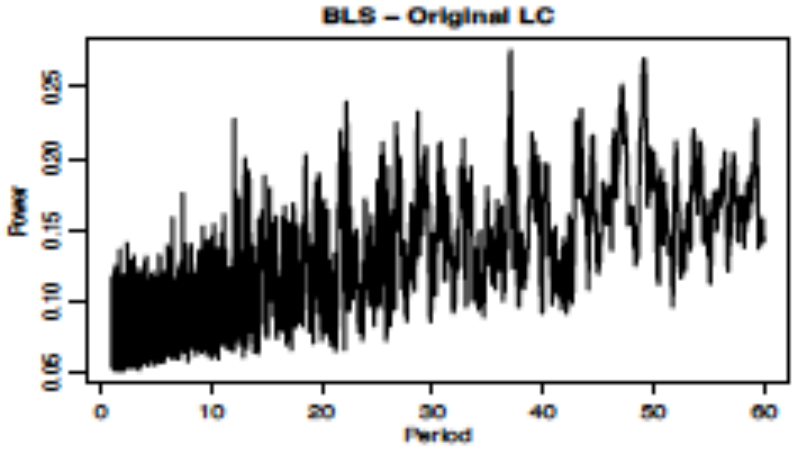
Comparison of TCF and BLS periodograms

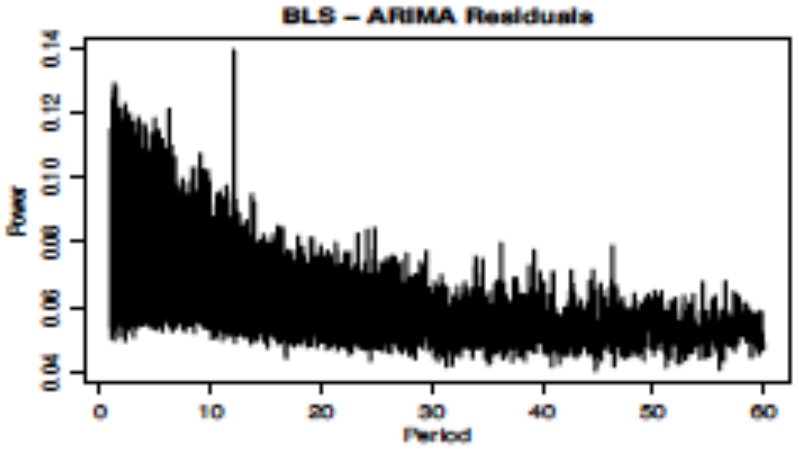
KIC 010024701

Here neither method detects periodicity in the original lightcurve

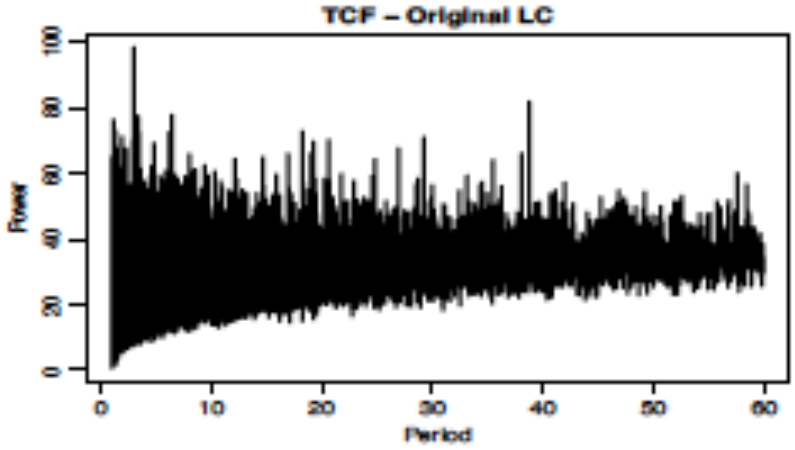
BLS detect periodicity in ARIMA residuals but with lower SNR than TCF

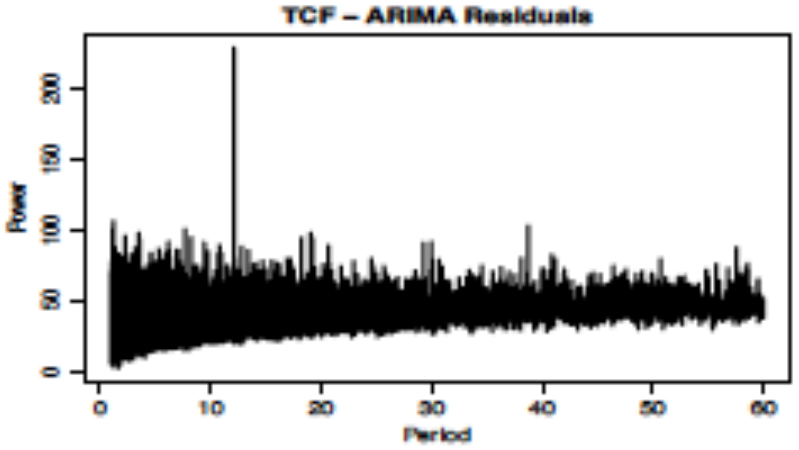
BLS





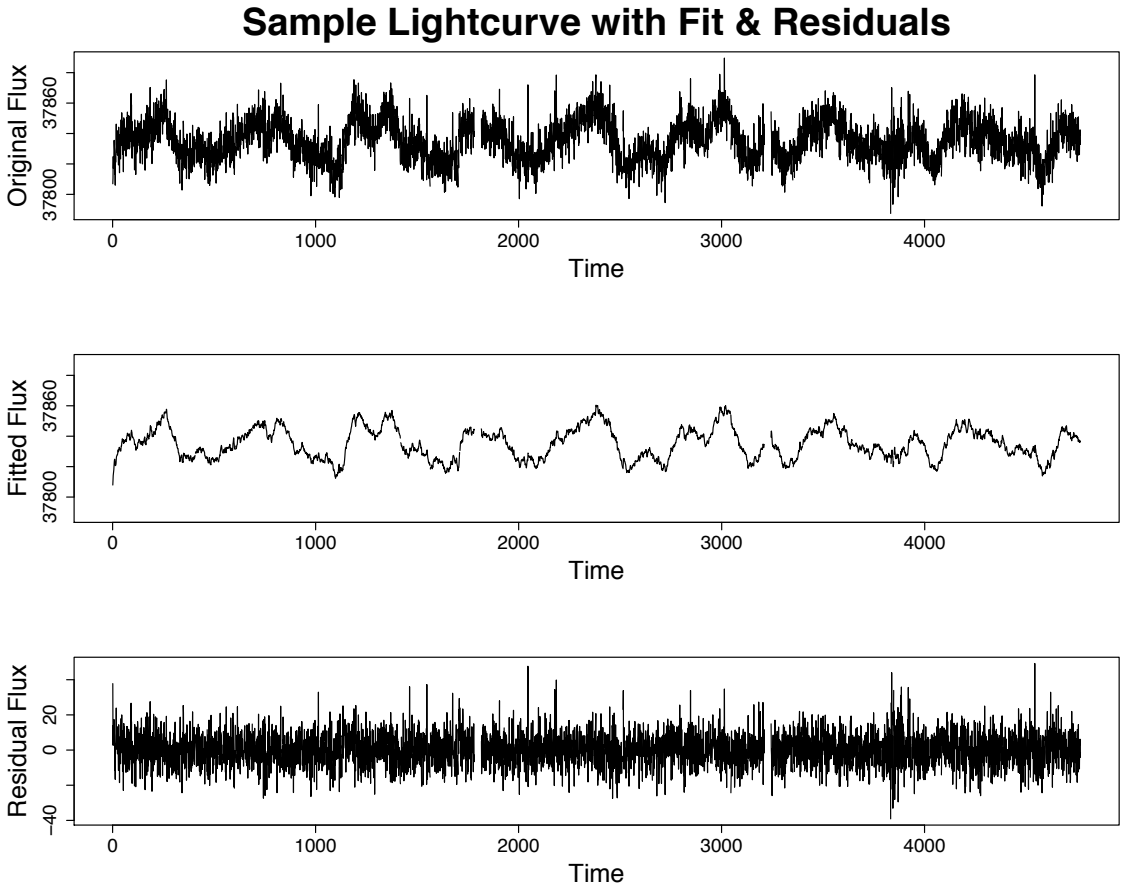
TCF





KARPS: Kepler AutoRegressive Planet Search

The ARPS procedure is applied to ~200,000 Kepler stars



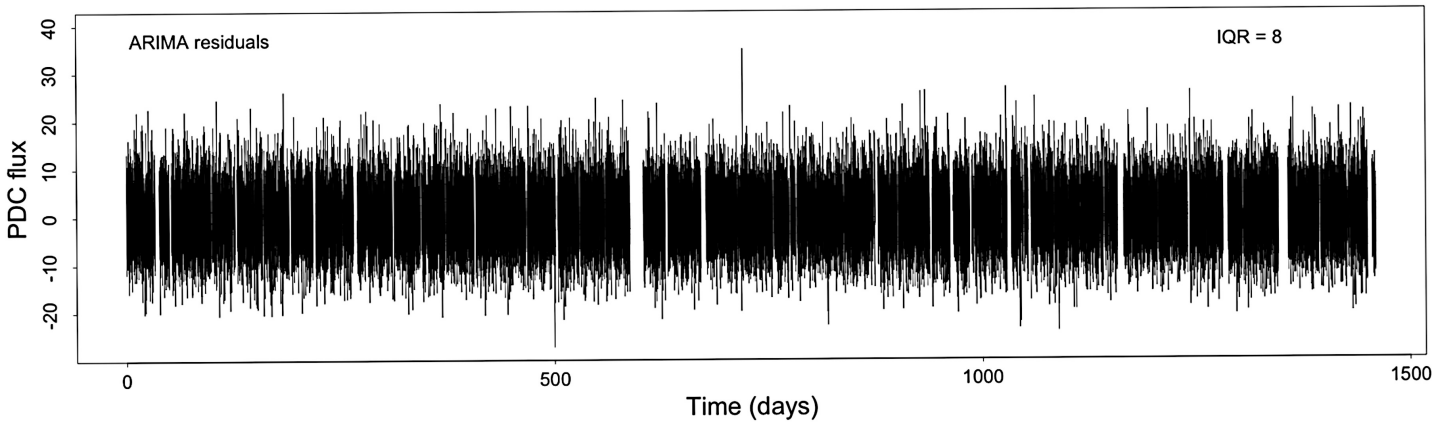
Typical 4 yr Kepler lightcurve

Maximum likelihood ARFIMA model

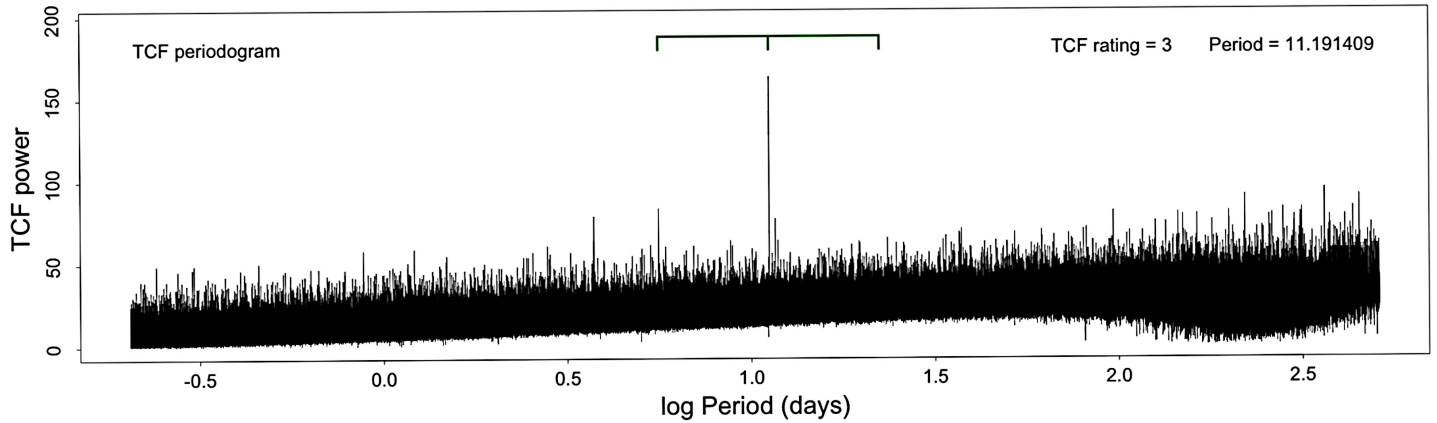
Residuals

Caceres, Feigelson, et al. 2019b

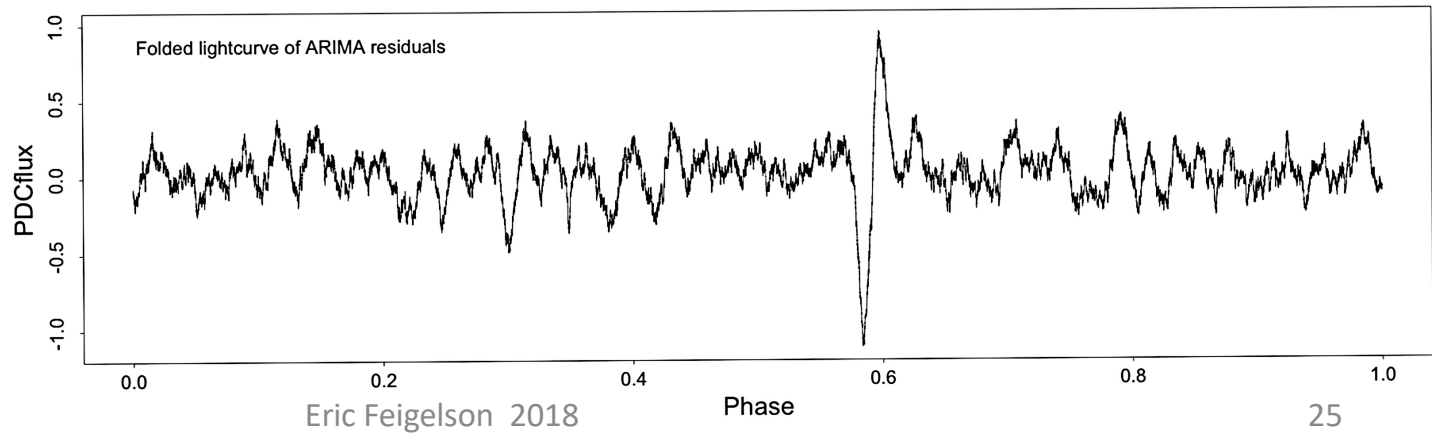
ARFIMA residuals



Transit Comb Filter periodogram



Folded light curve at best TCF period (double-spike)



Example #1

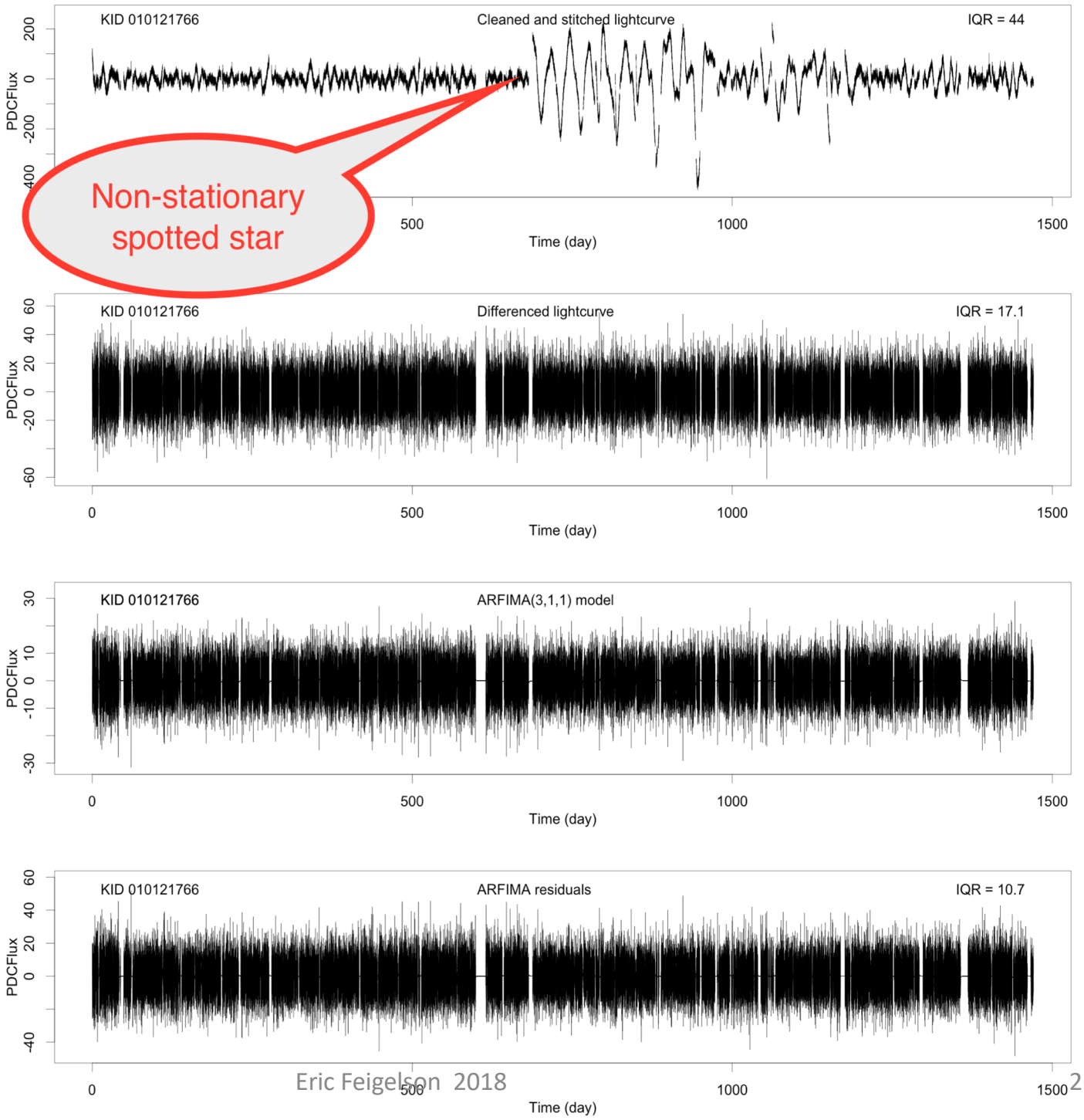
Observed 4 year light curve

After differencing

ARFIMA model

ARFIMA residuals

Noise reduced from IQR=44 to 10

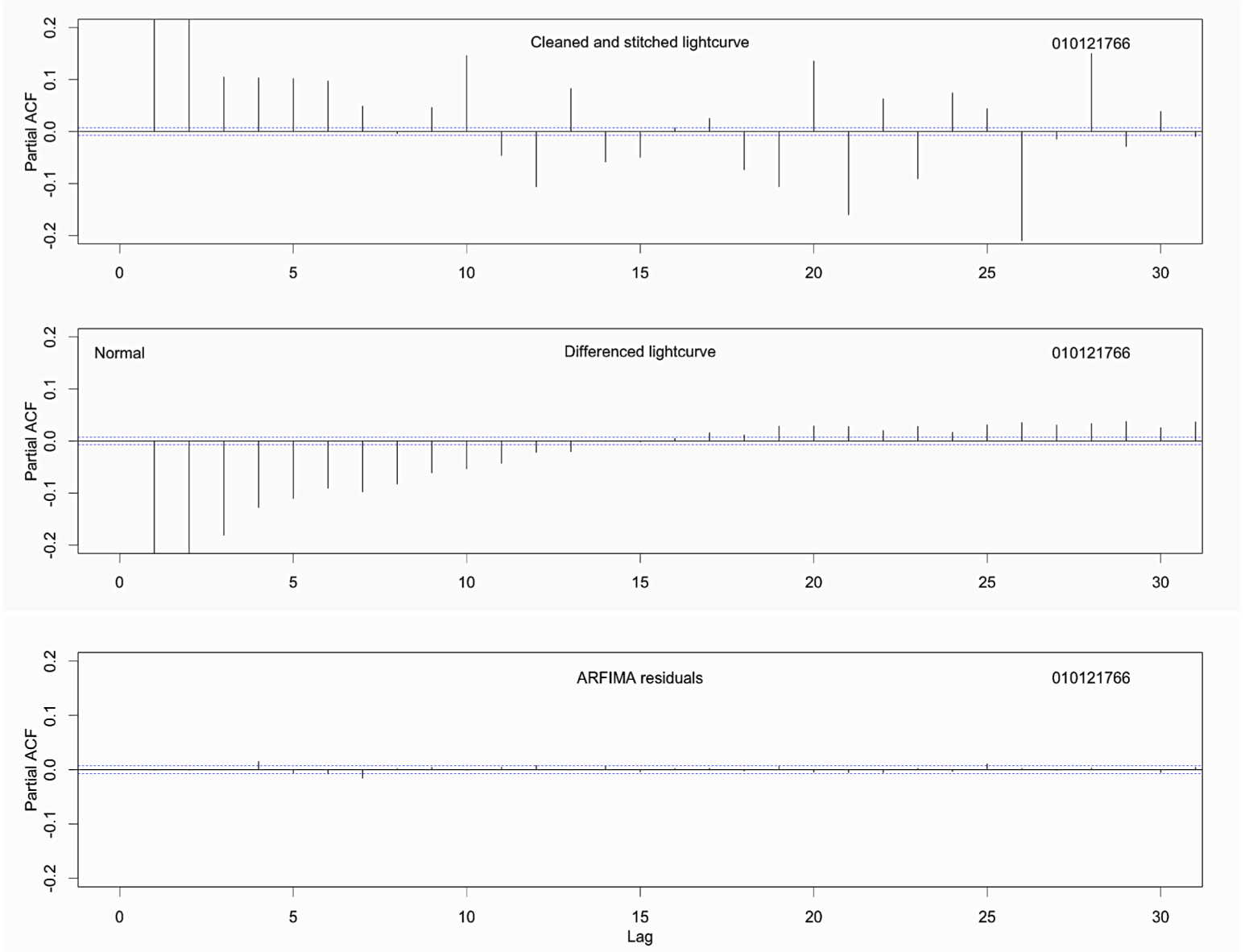


Partial autocorrelation function

Original lightcurve

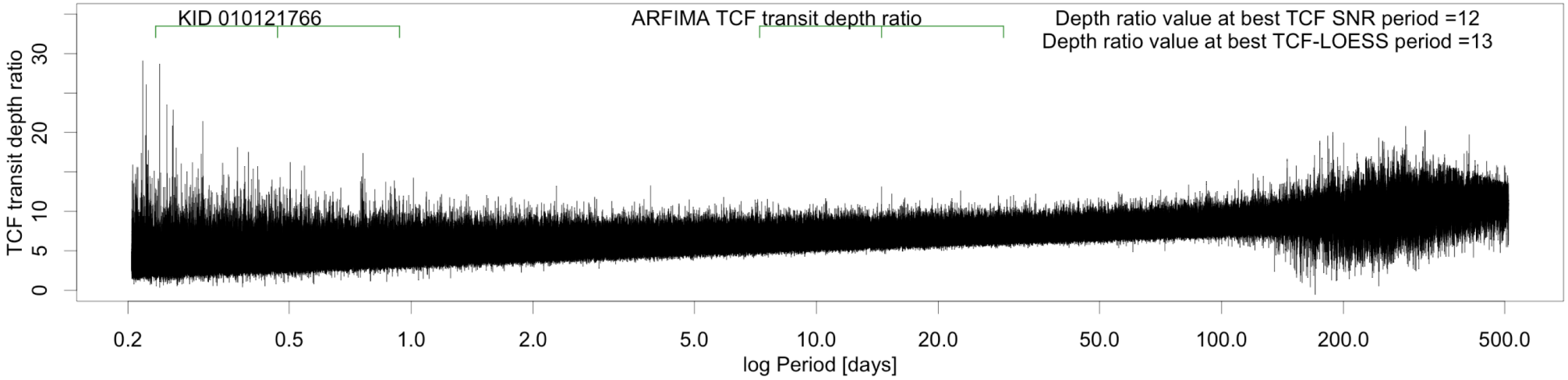
After differencing

ARFIMA residuals



Example #1 TCF periodogram ….. No periodicity seen !!

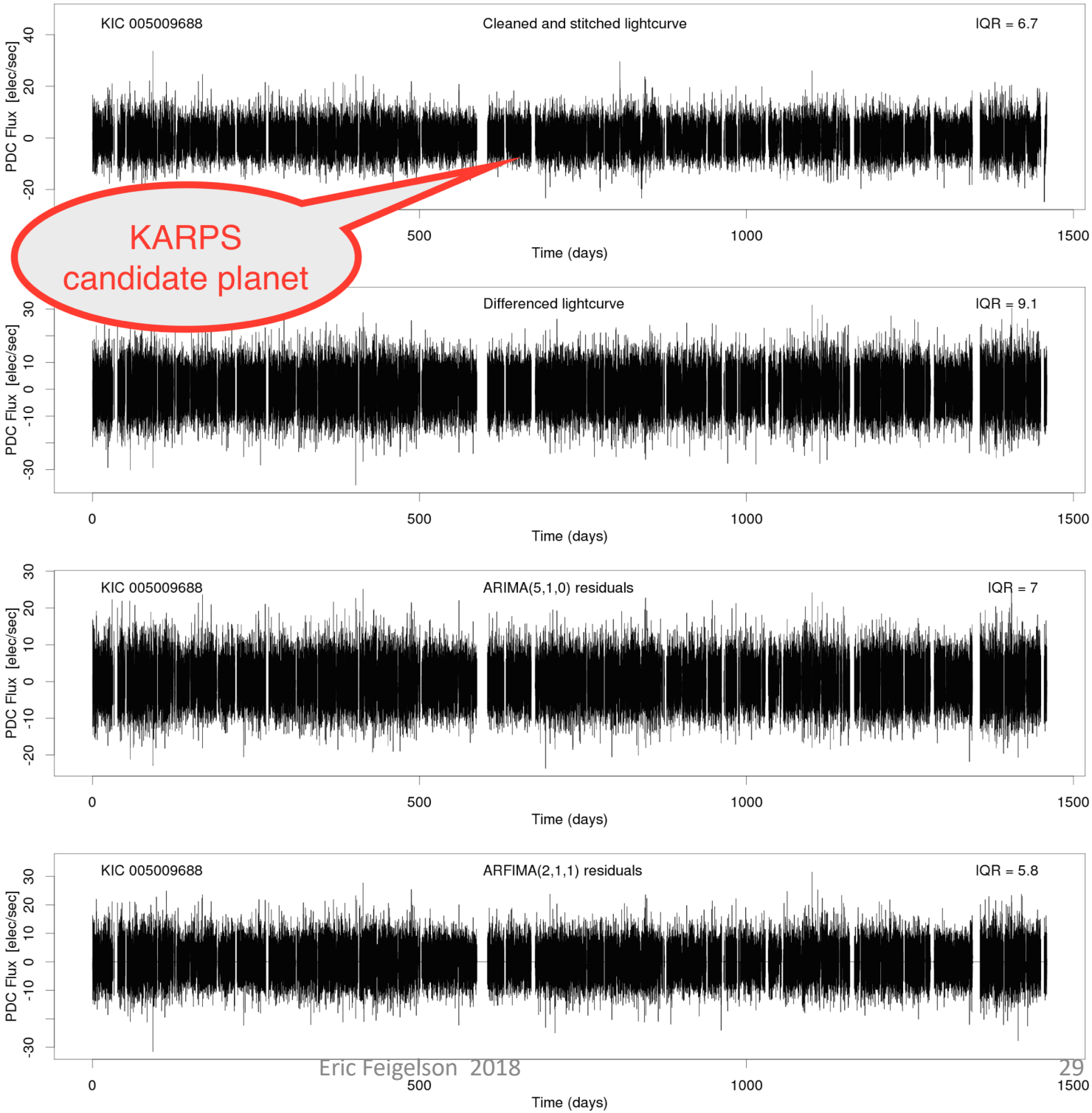
(This is the case for >95% of Kepler stars)



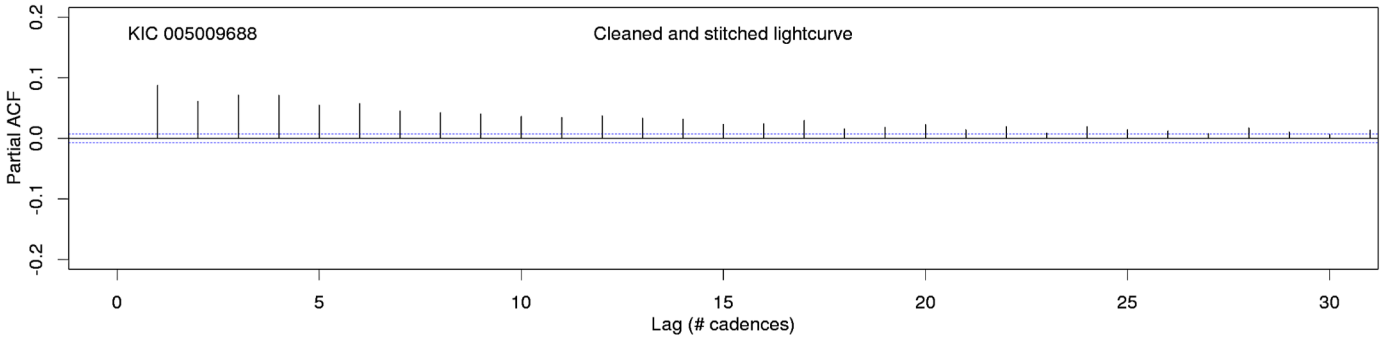
Example #2

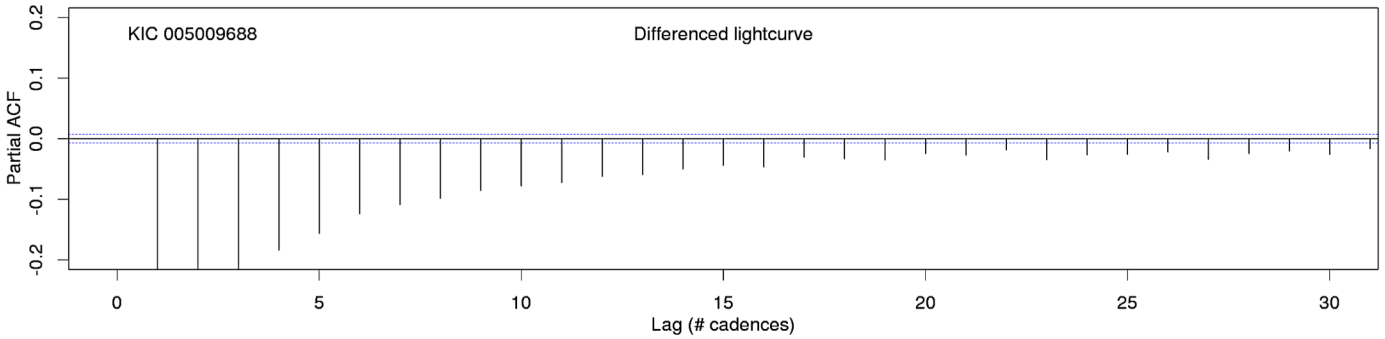
No apparent variability in light curve

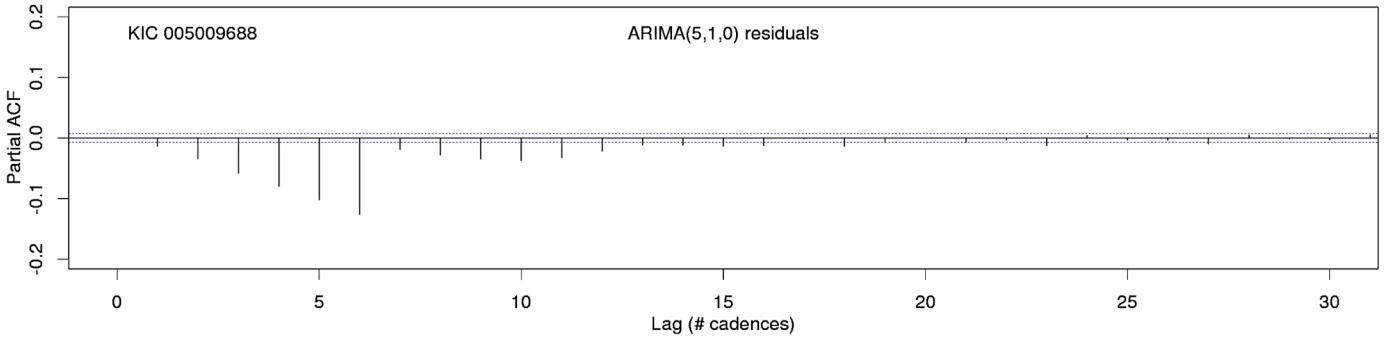
No improvement in noise with ARIMA models but …



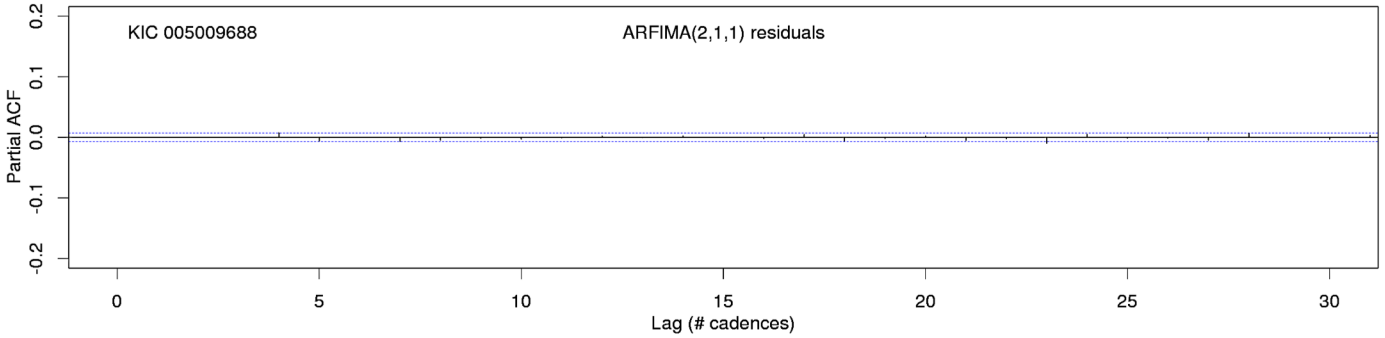
… autocorrelation is present and is …





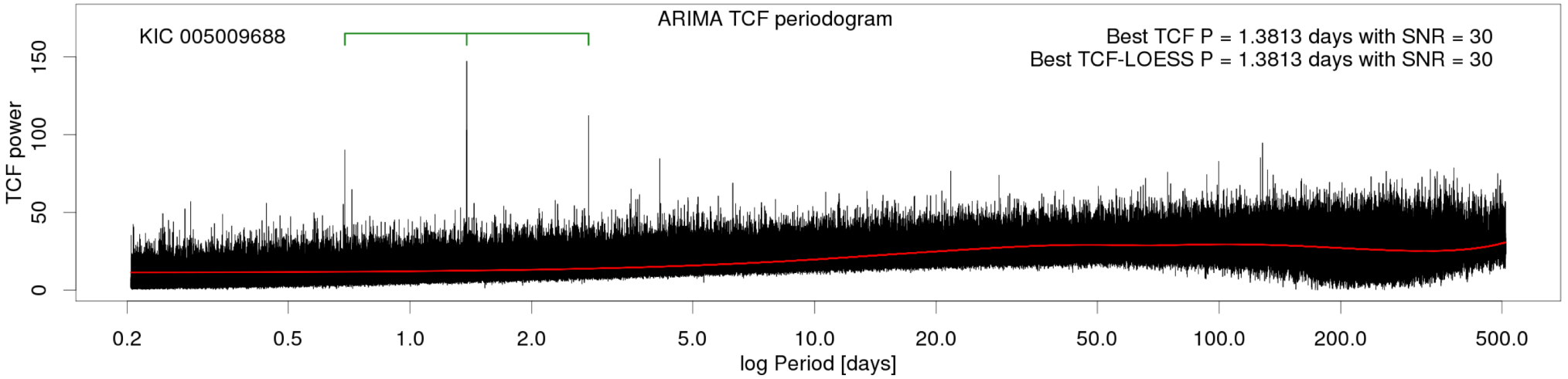


… removed with ARFIMA model



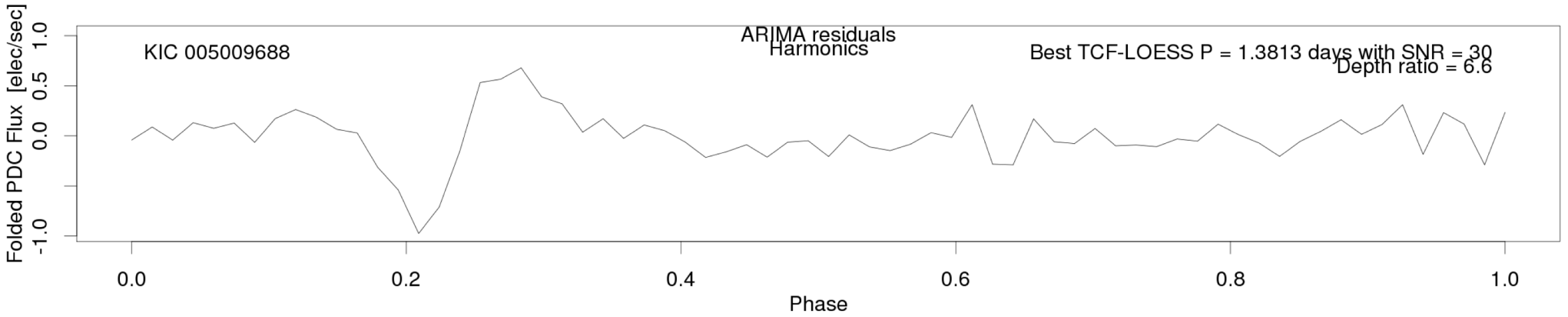
TCF periodogram show periodicity with harmonics

P=1.38 days with peak SNR = 30



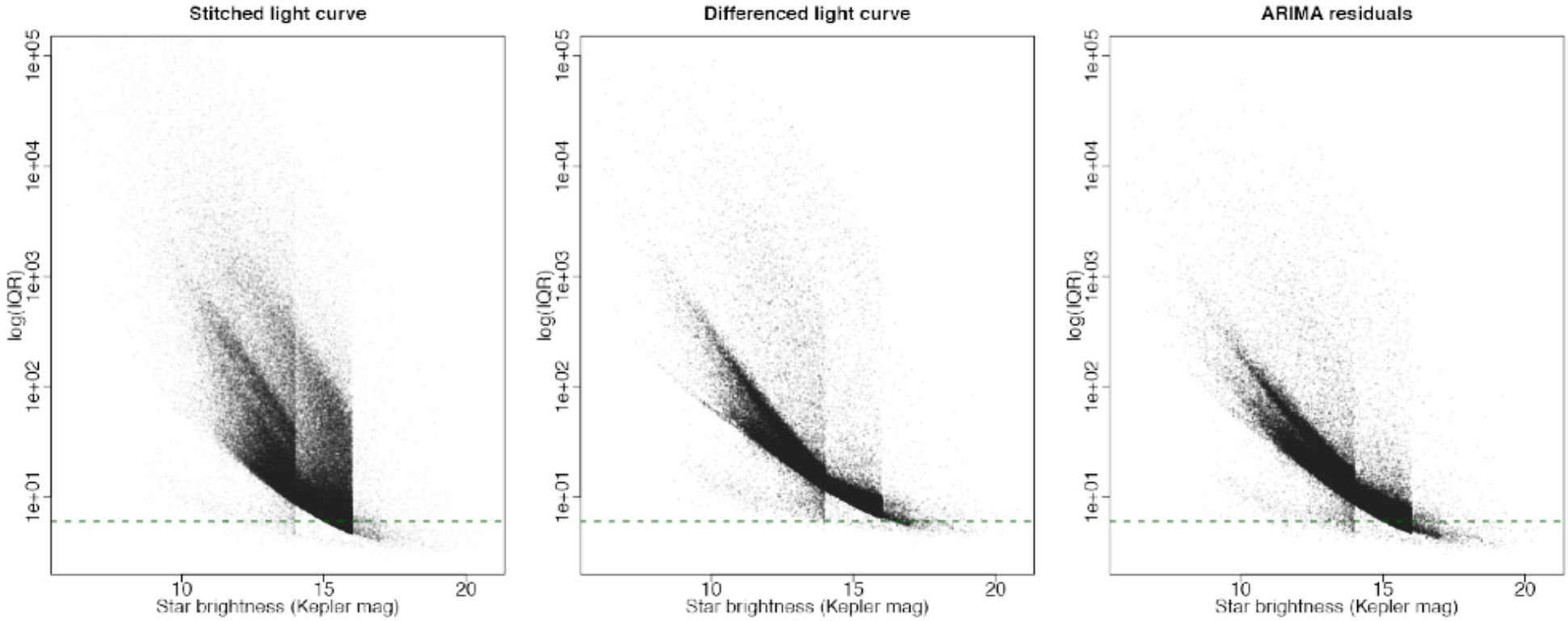
Folded lightcurve shows double-spike shape …

a new candidate planet !!



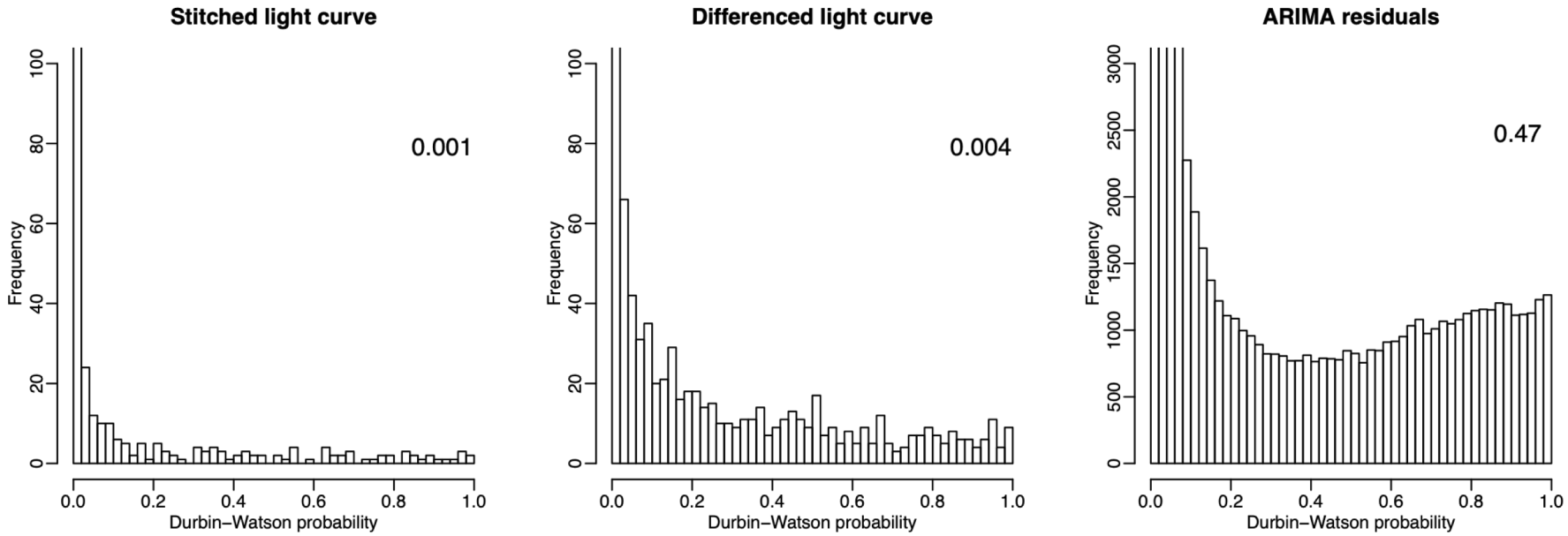
Results from full Kepler study

Improvements in lightcurve noise from ARIMA modeling

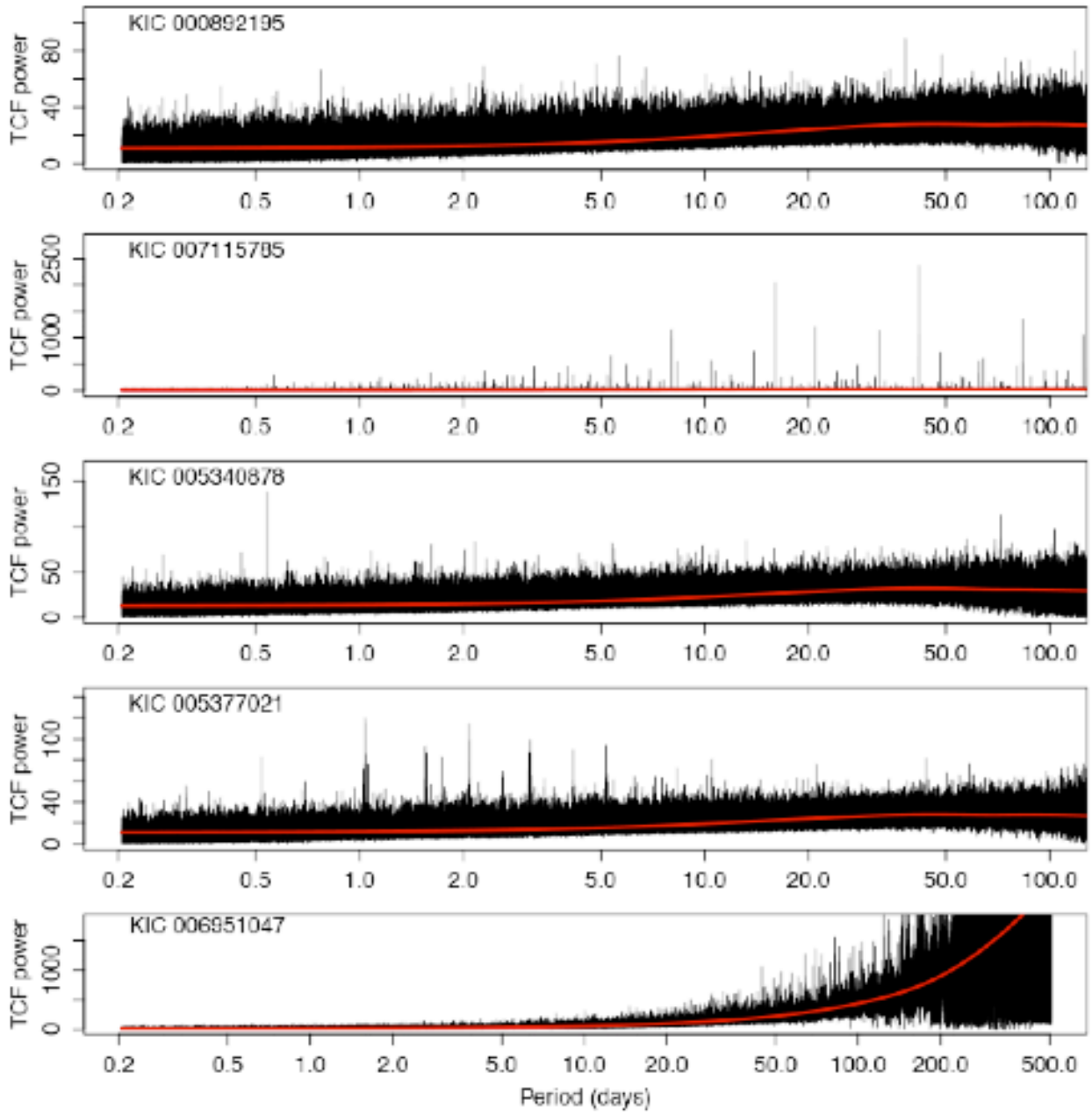


Caceres, Feigelson et al. 2019b

Improvements in lightcurve autocorrelation from ARIMA modeling



Examples of TCF periodograms



No peaks – very common

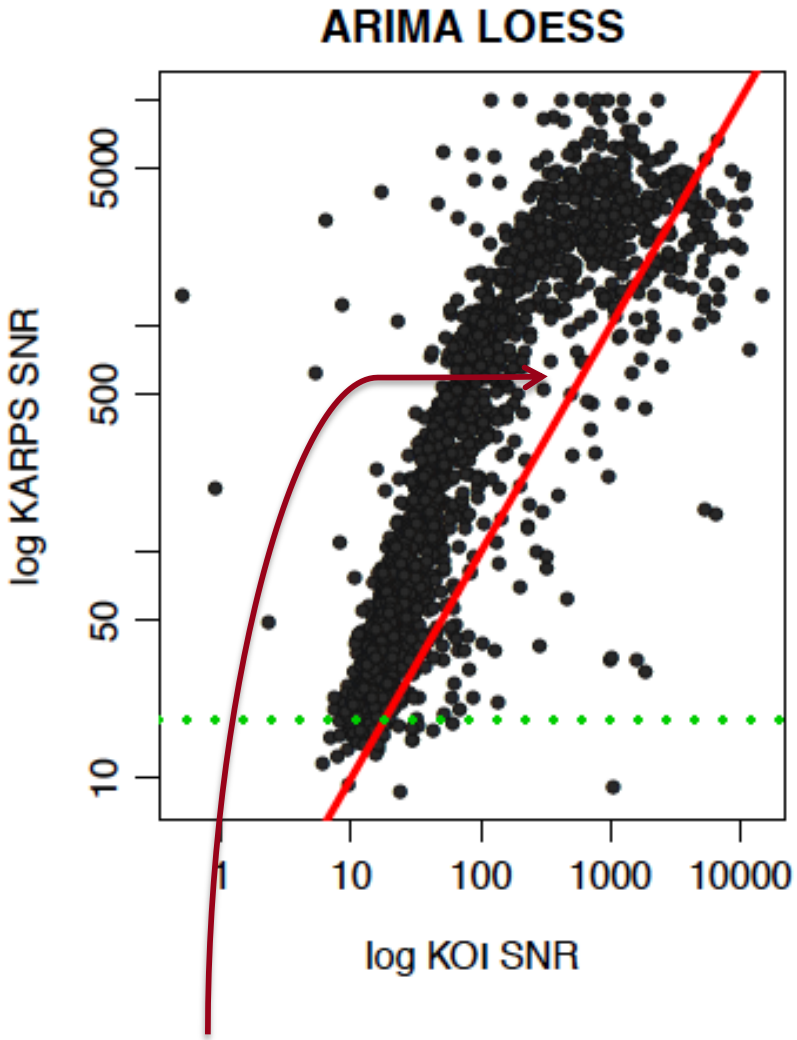
Strong periodic variable

New transit?

Red giant?

TCF corrupted by outliers

Comparison of KARPS TCF SNR and KOI ModelSNR



Offset suggests KARPS TCF is more sensitive than KOI modeling

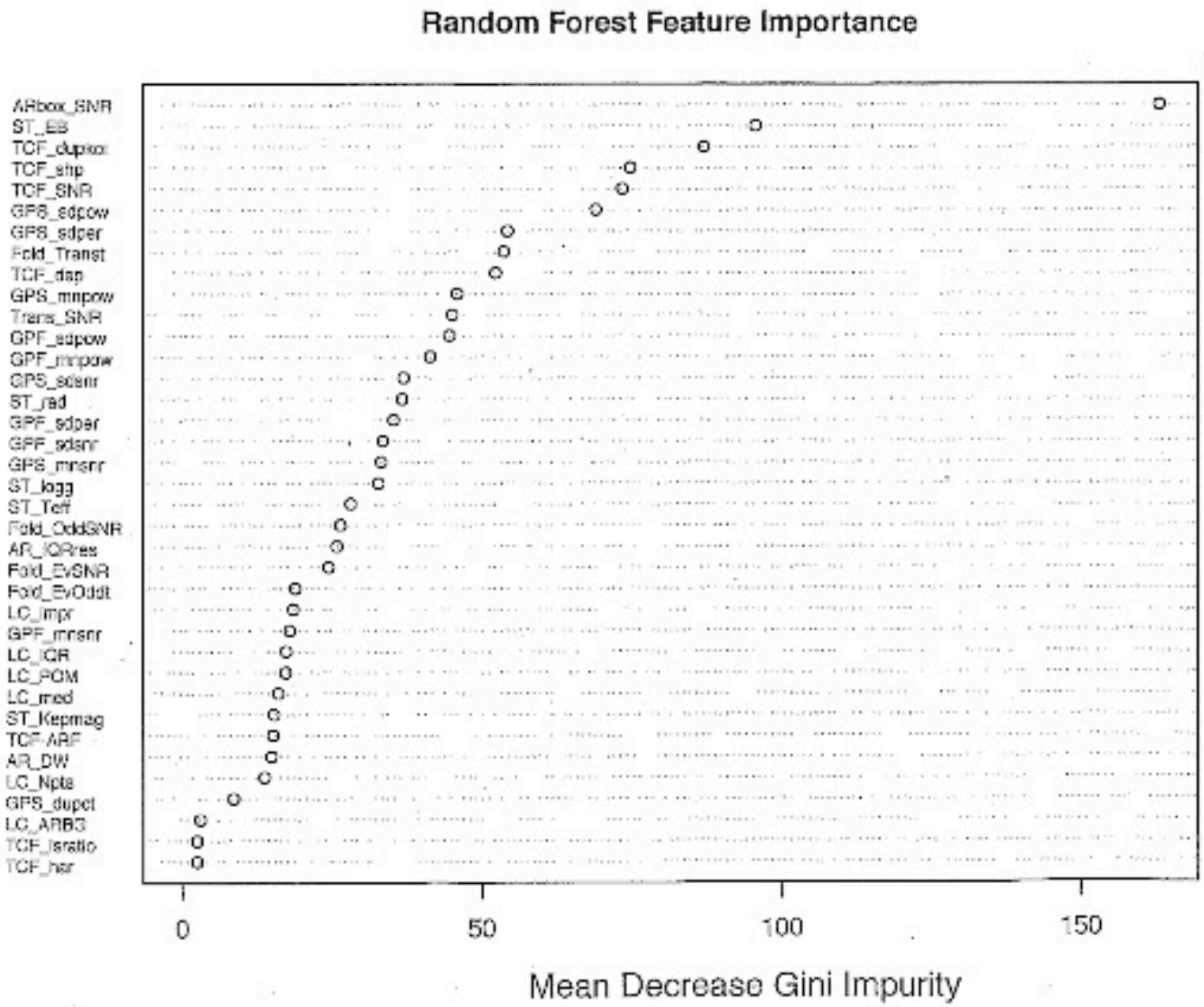
Caceres, Feigelson et al. 2019b

Random Forest classification and ROC curves for candidate planet selection

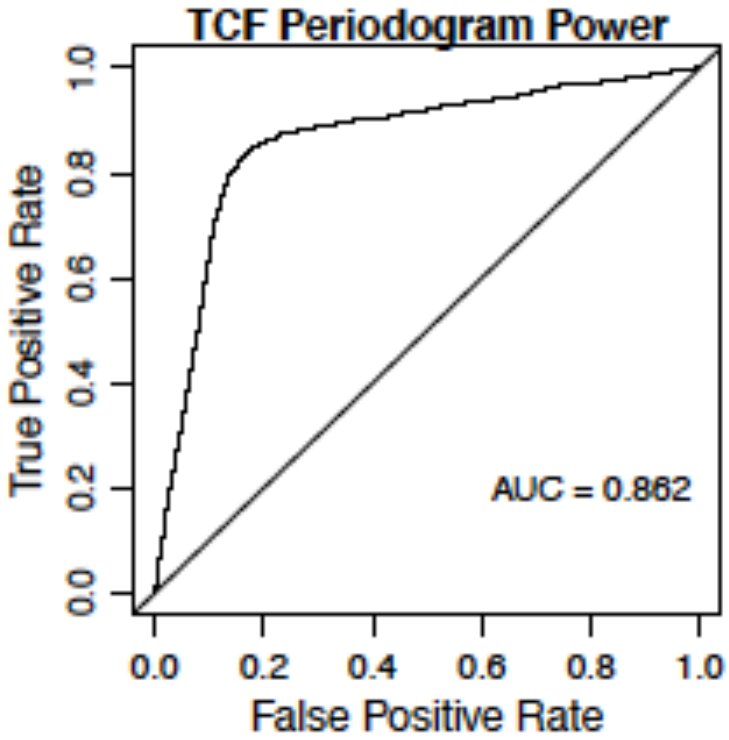
- The exoplanetary community places high level of trust in the Kepler Mission 'Confirmed Candidate' classification based on astronomical followup studies. These can be used as a training set for classification based on KARPS analysis.

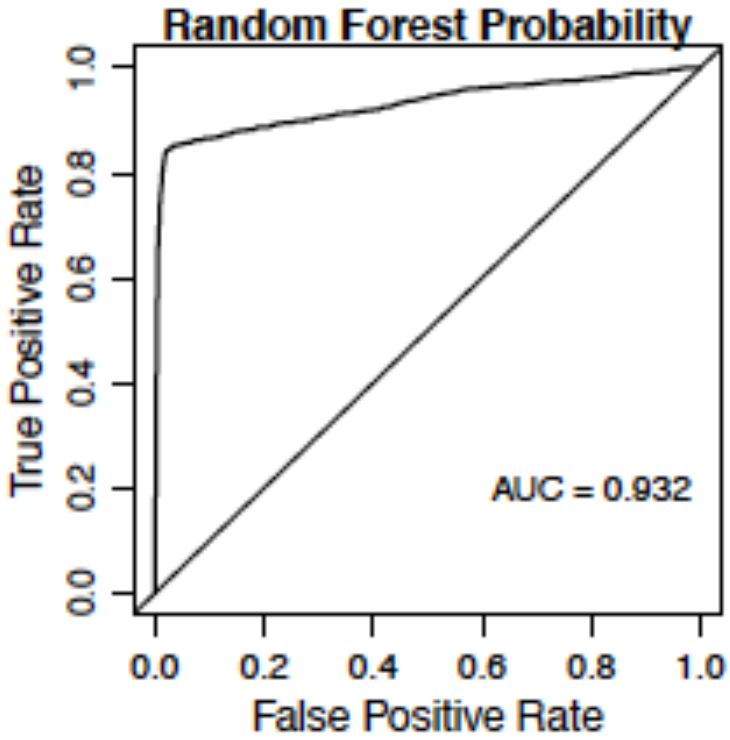
- Statisticians have extensively methodology to find classification criteria based on desired performance of True Positive & False Positive classifications. We use **Random Forest** decision trees with ~20 input features from lightcurves and TCF periodograms, with decision thresholds based on **receiver operating curves** (ROC).

Feature importance in Random Forest classifier

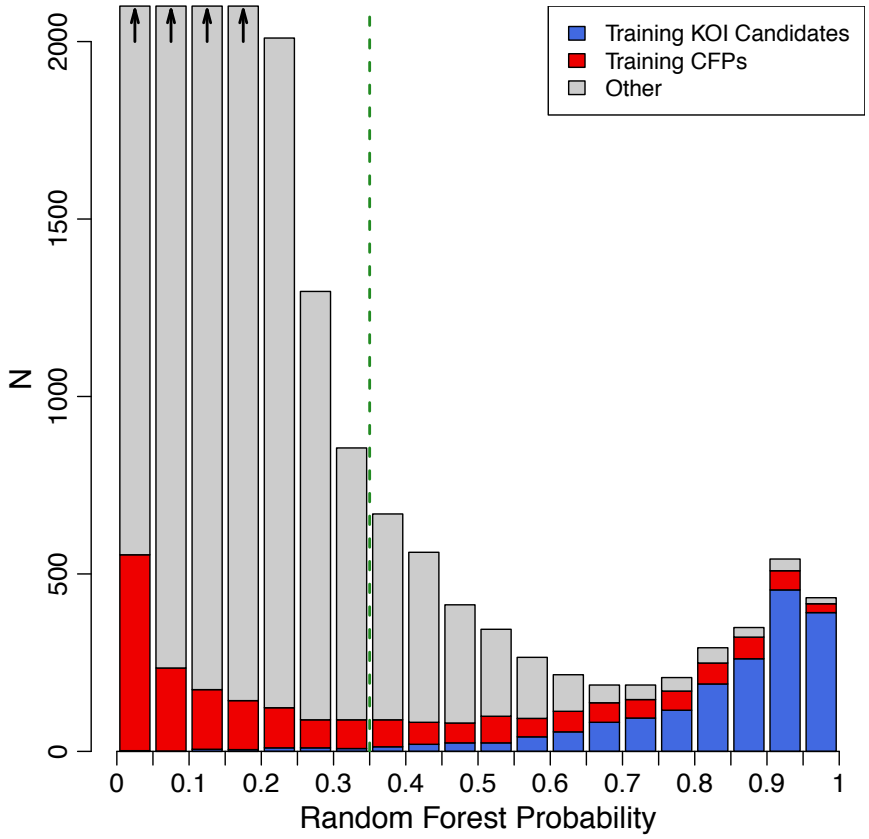


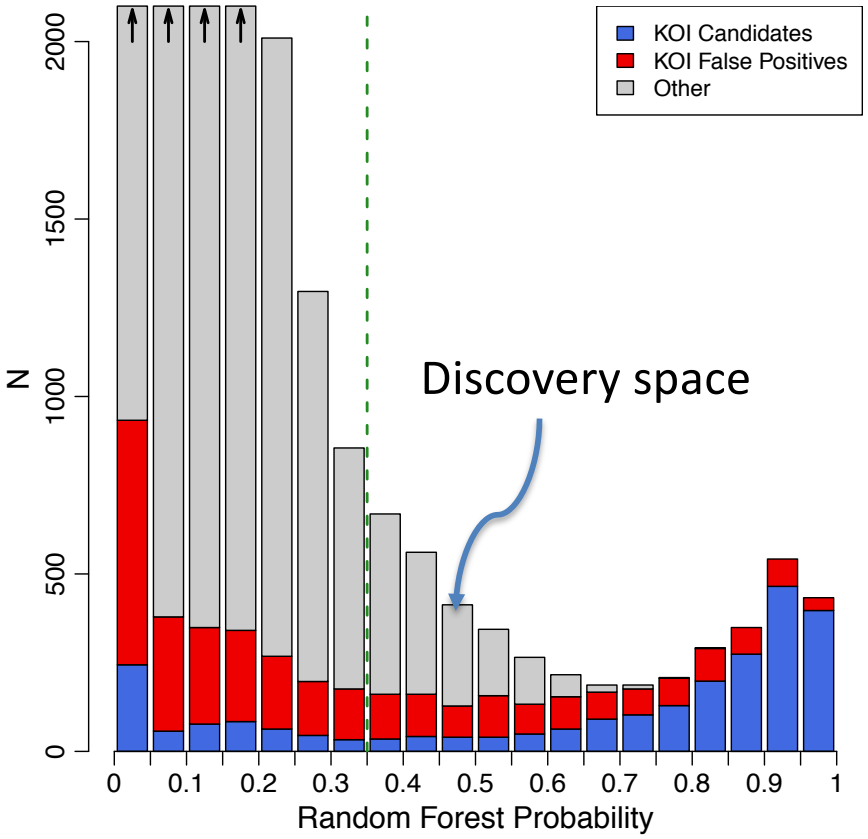
ROC curves for 1 vs. 37 features





Recovery & discovery with Random Forest classifier





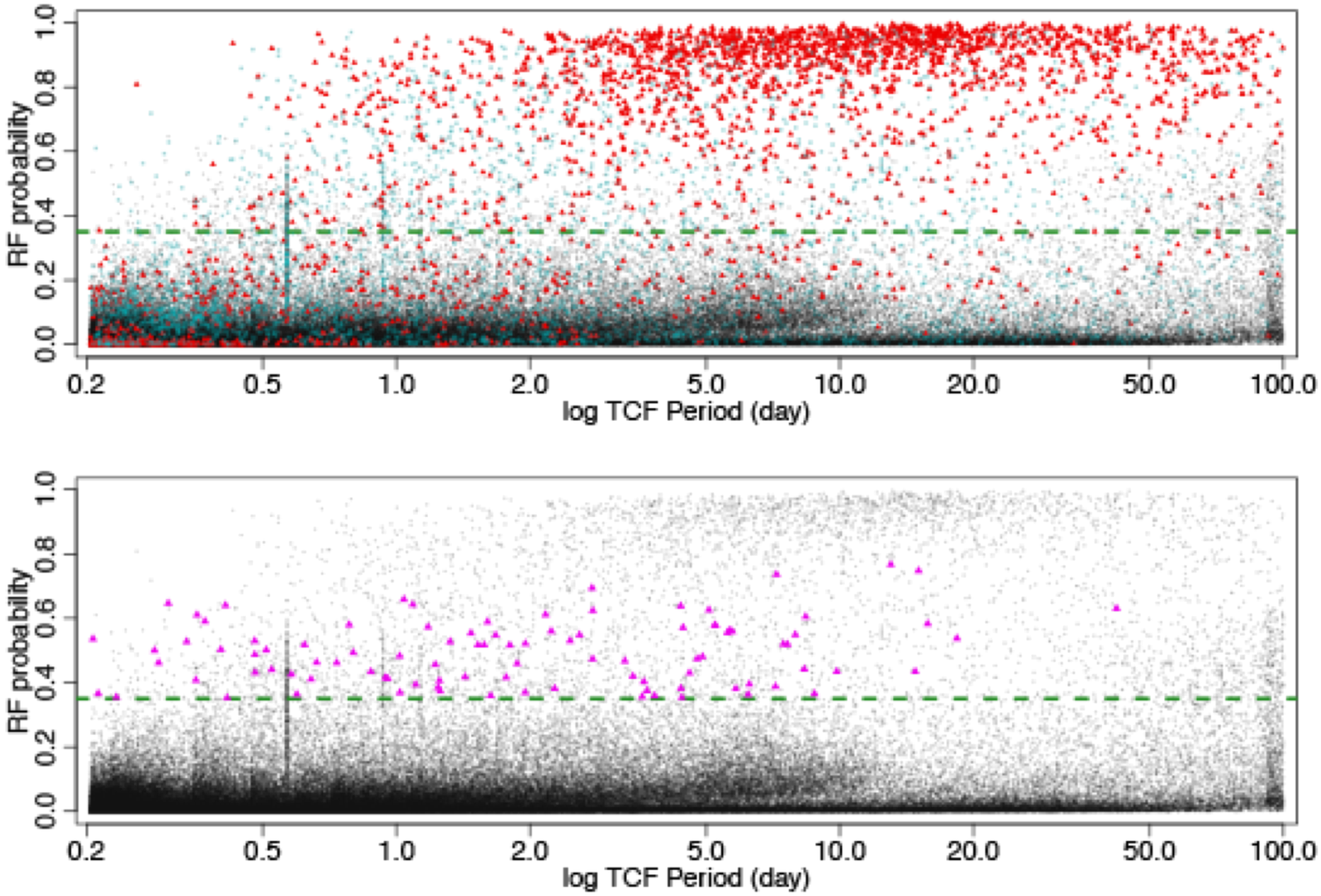
Blue = confirmed planets Red = confirmed False Positives Gray = random stars

Result for discovery of new Kepler planets

Red = previously known KOI planets

Magenta = new KARPS planets

Gray = stars without planets



Caceres, Feigelson et al. 2019b

ARPS Conclusion

Low-dimensional stochastic ARIMA-type models can be very effective in removing autocorrelated noise in stellar lightcurves, leaving periodic transits in the residuals.

TCF gives an effective periodogram for ARIMA residuals.

Random Forest is an effective classifier if strong training sets are available. False Positive rejection is tricky but feasible.

Coding is easy: extensive econometric software in R. Computational burden is reasonable.