Analysis of Irregularly Spaced Time Series

Eric D. Feigelson
Penn State University
LSST/TVS Workshop June 2018

EDF roles:
Core member, LSST/ISSC
Assoc Dir, Penn State Center for Astrostat
Lead author, astrostatistics textbook
President, IAU Commission Astroinfo/stat
AAS Journals, Statistics Editor

Outline

- **The challenges of astronomical time series analysis**
- **Thumbnail review of standard TSA (regular cadence)**
- **Irregular time series: Use standard TSA methods**
  - Examples: nonparametric local regression, ARIMA for Kepler → HAT-South
- **Irregular time series: Specialized methods**
  - Examples: Edelson-Krolik DCF, Stellingwerf's PDM, Lomb-Scargle periodogram, CARFIMA
- **LSST challenge: Classification of billions of sparse, irregular time series**

Conclusion: It's not easy!! Lots of methodology to learn Use R or Matlab

Challenges

Wide range of temporal behaviors

- **Periodic phenomena:** binary orbits (stars, exoplanets), stellar rotation (pulsars, spots), pulsations (asteroseismology, Cepheids)
- **Stochastic phenomena:** accretion (cataclysmic variables, X-ray binaries, Seyfert galaxies, QSOs
- **Explosive phenomena**: thermonuclear (novae, X-ray bursters), magnetic reconnection (solar/stellar flares), star death (supernovae, gamma-ray bursts)

Difficulties in astronomical time series

- **Irregular data streams:** Diurnal & annual cycles, telescope allocation/cadence committees
- **Heteorscedastic measurement errors:** S/N differs from point to point

Standard TSA

Fundamental concepts

- **Cadence & noise** Nearly all standard methodology assumes: (a) regularly spaced observations, although missing data (NAs) are often allowed; and (b) homoscedastic Gaussian white noise, $\varepsilon=N(0,\sigma^2)$, plus signal of interest

- **Stationarity** Statistical properties unchanged by shifts in time. Variations can be deterministic or stochastic. Nonstationarity includes: trends, heteroscedasticity, and change points.

- **Periodicity** Measured levels repeat with fixed periods. The signal is concentrated in a narrow range of frequencies (spectral, Fourier, harmonic analysis)

Standard TSA

- **Autocorrelation** Measured levels depend on previous values. The nonparametric autocorrelation function *ACF(k)* gives fraction of variance attributed to correlation at lag *k*:

$$ACF(k) = \Sigma(x_t - \langle x \rangle)(x_{t-k} - \langle x \rangle) / \Sigma (x_t - \langle x \rangle)^2$$

- **Stochastic process** Dynamic evolution with random ingredients. Short-memory behaviors are concentrated in a few lags of the ACF. Long-memory behaviors include $1/f^{\alpha}$ `red' noise.

- **(Non)parametric methods** Nonparametric procedures make no assumption regarding the global behavior. Parametric procedures assume a mathematical form to the global behavior; regression is used to find the best-fit parameter values.

Standard TSA

Nonparametric time domain methods

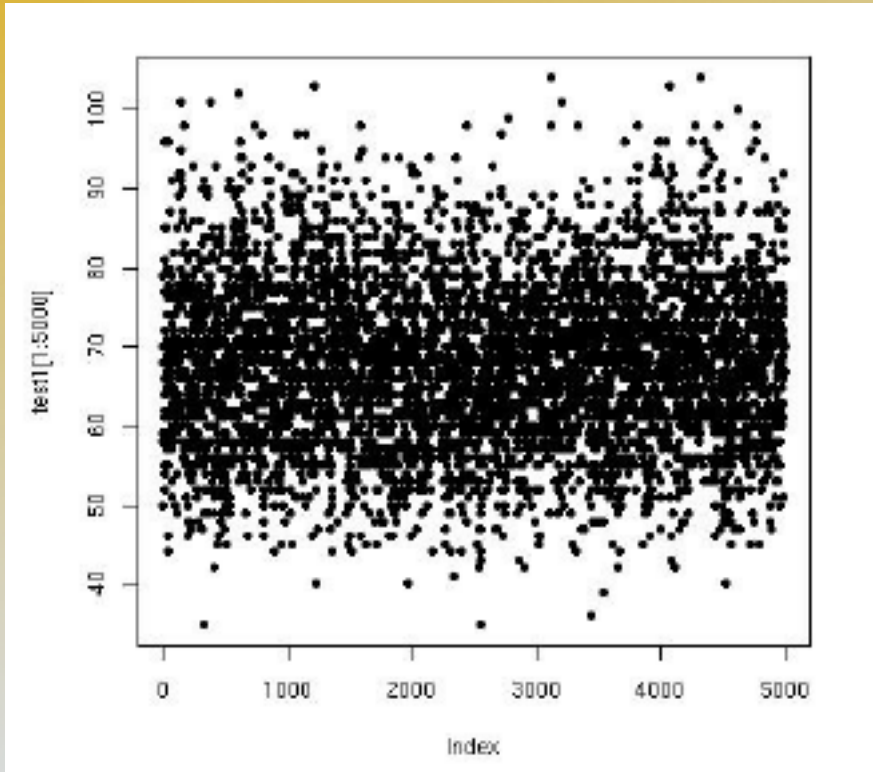
- **(Partial) autocorrelation function** For zero-mean Gaussian white noise, the ACF is asymptotically normal (i.e. probabilities can be inferred): *Var(ACF) = (n-k)/n(n+2).* *PACF(k)* gives effect at lag k with smaller lags removed.

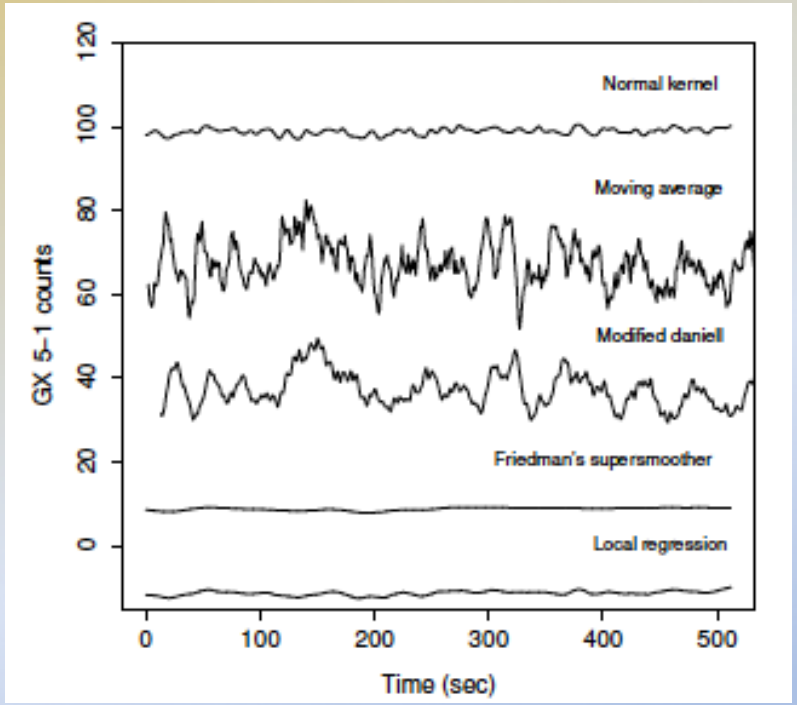
- **Density estimation = smoothing** Kernel density estimation, local polynomial regressions (splines, LOESS, …), Gaussian Processes regression. Preferred if deterministic global functional form is not known.

- **Differencing operator** Replacing x_t with $x_t - x_{t-1}$ removes many types of trend. Reversed with integrating operator.

Standard TSA

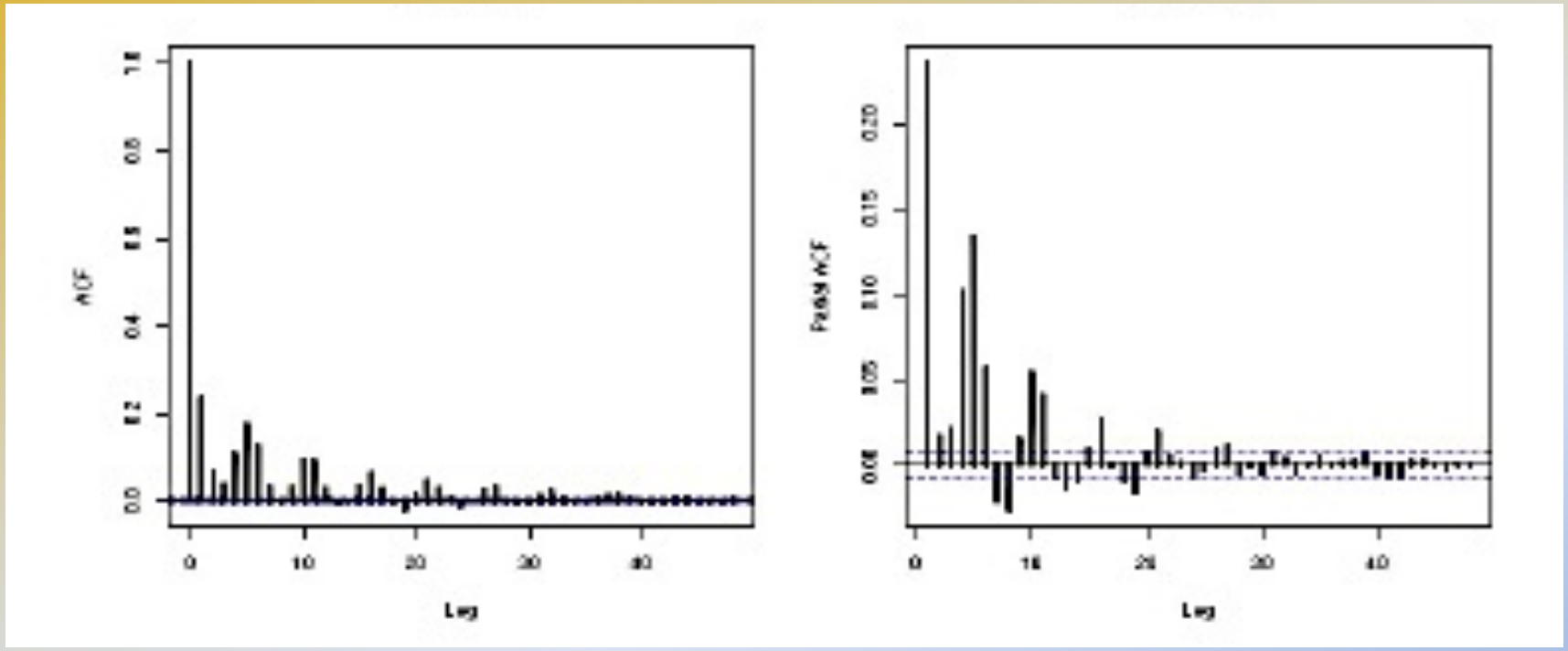
X-ray counts from GX 5-1





Feigelson & Babu (2012)

Standard TSA



Standard TSA

Parametric time domain methods

- **Deterministic trend** Variation in time has known form, $x_t = f(t)$, such as polynomial or exponential (not common in astronomy)

- **Stochastic trend** Many stochastic behaviors can be successfully fit with linear autoregressive models: ARMA (short-memory stationary), ARIMA (with nonstationary trend), ARFIMA (with long-memory $1/f^{\alpha}$), GARCH (with volatility), etc.

$$\mathbf{AR(p)\ model}\quad x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + w_t$$

$$\mathbf{MA(q)\ model}\quad x_t = w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}$$

ARMA-type models are fit by maximum likelihood with order (p,q) chosen with the Akaike Information Criterion. Lots of methodology: parameter errors, goodness-of-fit tests, hierarchical models, Bayesian inference, Kalman filter updates, etc

Standard TSA

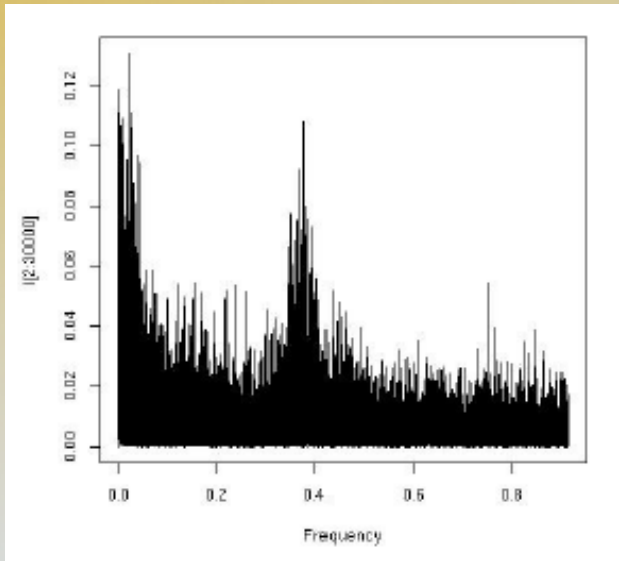
Nonparametric frequency domain method

- **Fourier analysis** Fourier transform (1807), power spectrum (Schuster periodogram) concentrates strictly periodic signal into sharp peak. Theorems are highly restrictive: evenly spaced data of infinite duration, Gaussian white noise with single sinusoidal-shaped signal at fixed period. For realistic data, difficult to establish significance of periodogram peaks.

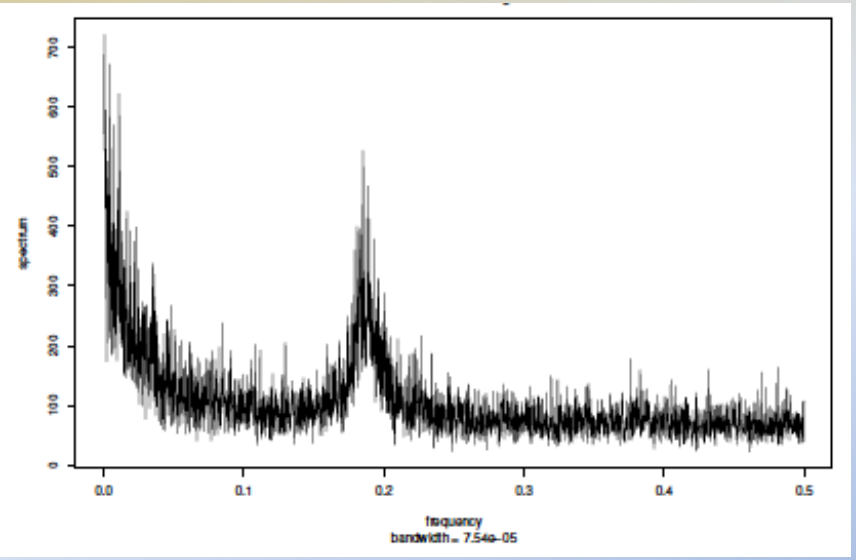
- **Modern Fourier analysis** improves the power spectrum with smoothing (reduced variance), tapering (reduced bias), detrending (reduced zero-frequency power), and pre-whitening (removing strong periodicity). Multitapering is recommended.

Standard TSA

Raw Fourier periodogram
of GX 5+1 time series



… after smoothing and tapering



Standard TSA

Advanced signal processing methods are being introduced into astronomy by leading astrostatistics groups:

- *Independent Component Analysis* (Waldmann et al. 2012)
- *Correntropy* (Huijse et al. 2012)
- *Hilbert-Huang transform* (Kolotkov et al. 2015)
- *Empirical Mode Decomposition* (Roberts et al. 2013)
- *Singular Spectrum Analysis* (Boufleur et al. 2018)
- Dynamic time warping ()
- … …
- *Ensemble methods to identify astronomical instrumental & atmospheric effects: SysRem, TFA, EPD, PDC-MAP, ARC* (Tamuz et al. 2005, Kovac et al. 2005, Bakos et al. 2007, Stumpe et al. 2012, Roberts et al. 2013)

Standard TSA

Strategy for time series analysis

The ACF, ARMA maximum likelihood fit, wavelet transform, and the Fourier transform all contain the same information as the original time series if an infinite number of coefficients are maintained.

Each specializes in highlighting different characteristics of the dataset.

Choice of method depends on properties of the variability and the scientific question being addressed. There is no problem using different methods for different purposes.

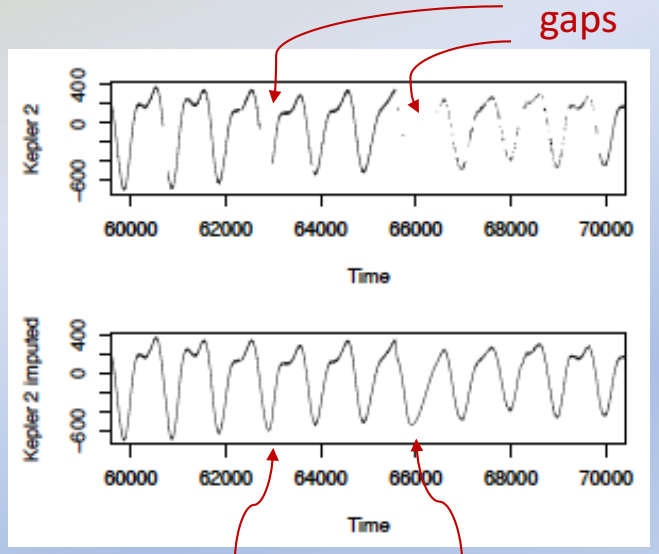
Irregular TS

Irregular time series: Try standard methods

- Nonparametric local regression (= smoothing) directly works for irreg TS: splines, LOWESS, GP regression, etc. Use methods that give heteroscedastic confidence intervals (local bootstrap for lots of data, Gaussian assumption for sparse TS).

- Drizzle data onto a fixed time grid, average values where coverage is dense, impute (interpolate) across gaps, or leave gaps as NA values. Sophisticated TS imputation methods in R/CRAN packages. Then use standard methods.

- Penn State methodology studies show reasonable performance of ARIMA for lightcurves with 15-90% NAs (Kepler → HAT-South cadences).



Irregular TS

Irregular time series: Specialized methods from statistics, signal processing & econometrics

Almost no methodology is available!

One approach …

Continuous-time autoregressive parametric models: CAR(p), CARMA(p,q), CARFIMA(p,d,q)

Astronomers have used low-dimensional CAR and CARMA models to characterize quasar variability (Kelly et al. 2009, 2014). But these models assume stationarity, no red noise, and simple behaviors.

Good news: Statisticians have developed MLE for the richer CARFIMA family, and a R/CRAN package emerged following SAMSI Astrostatistics Program (Tak & Tsai 2017).

Irregular TS

Irregular time series: Specialized methods from astronomy

- Edelson-Krolik Discrete Correlation Function is variant of ACF with NAs. Probably standard method.
- Structure function. Not used by statisticians. No theorems on confidence intervals.
- Red noise modeling from linear regression on log-Fourier power spectrum. Same as econometricians GPH method … not recommended. Obtaining α in $1/f^{\alpha}$ noise is not straightforward; try ARFIMA.

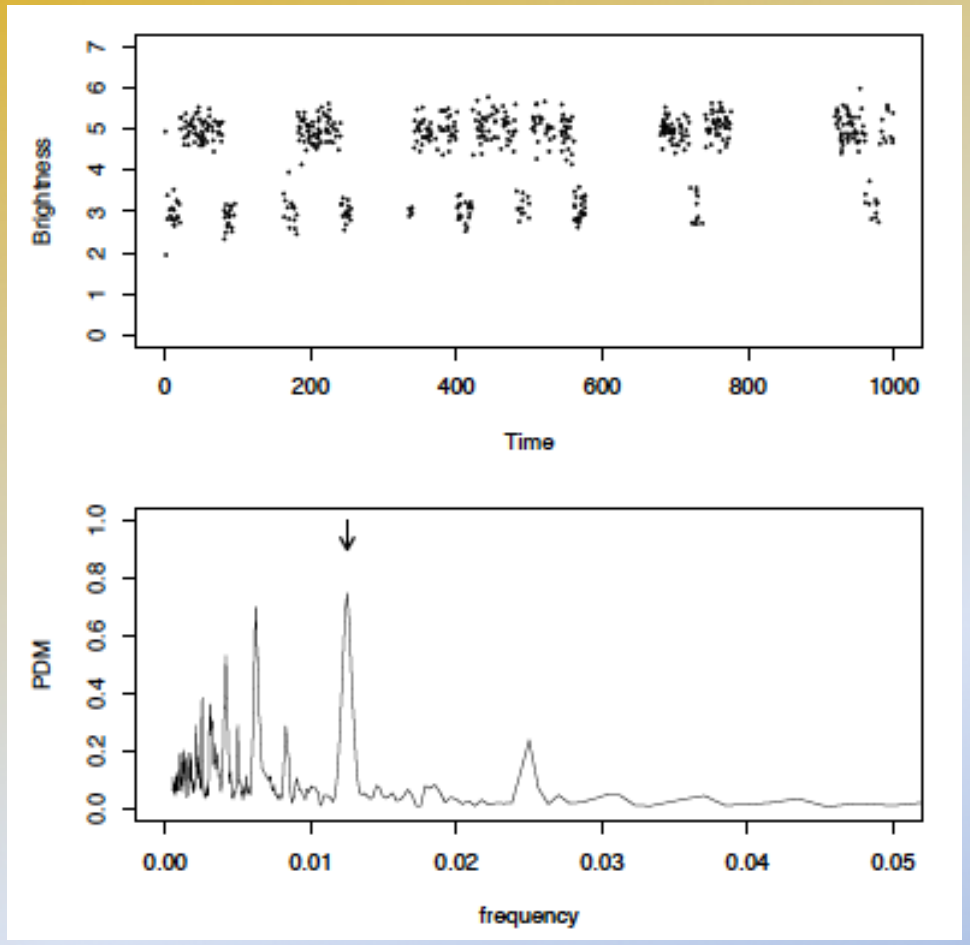
Irregular TS

Astronomical methods for periodicity search in irregular time series

- Several nonparametric statistics of folded lightcurves give periodograms: Lafler-Kinman (1965 = Durbin-Watson statistic), Stellingwerf phase dispersion minimization (1978) = ANOVA statistic (Schwarzenberg-Czerny 1989), Dworetsky minimum string length (1983), Gregory-Loredo Bayesian model (1992). These periodograms treat arbitrary irregularity and shape (e.g. V-shaped eclipse, asymmetry from elliptical orbit). Box Least Squares (Kovacs et al. 2002) and Transit Comb Filter (Caceres et al. 2018) give specialized periodograms for exoplanet transits. No statistical evaluation of any of these methods, hence no probabilities for periodogram peaks.

- Lomb-Scargle (1982) generalization of Fourier transform for irregular spacing. Heritage in least-squares fitting of sinusoids in 1960-70s. Clear Bayesian interpretation (Bretthorst 2003). Difficult to evaluate significance of periodogram peaks; see review by vanderPlas (2018).

Irregular TS



Simulated transit-like periodicity in a heavily gapped time series

Stellingwerf PDM periodogram showing strong aliases of simulated period

[R/CRAN package RobPer]

Irregular TS

Change Point Analysis

Wide variety of nonparametric and parametric methods to find sudden change in level or variability behaviors. ***Relevance to astronomy has not been investigated.***

- CUSUM (cumulative sum) tests used extensively in manufacturing quality control allow NAs in evenly spaced data.
- Many financial time series methods for irregularly spaced data to treat high frequency stock trading.
- Many extensions Wald's Sequential Probability Ratio Test for sudden change in mean, variance, etc.
- Many tests of sudden change of ARMA coefficients.

Irregular TS

LSST challenge: Classification of billions of sparse, irregular time series

Gather scalar `features' from many (non)parametric time series methods and feeding them into a machine learning classifier like Random Forests (e.g. J.Richards et al. 2011). These can be highly effective at classifying new variable objects.

Two main challenges (neither involves methodology):

- must generate large, high-quality training sets: full feature lists for many members of each class of variable objects.
- must have sufficiently rich time series on test data to discriminate between classes. This may not work effectively until several years of LSST cadences are obtained.

Final comments

- Time series analysis is huge field of methodology and there is a lot for us to learn. I suggest that everyone be familiar at the level of Feigelson/Babu (2012) and Chatfield (6th ed 2003).
- Characterization of time series with weird LSST-type cadences is not in any textbook. Non-trivial challenges arise. Nonetheless, standard methods can often be effective or built upon.
- Temporal characterization can be effective for variability timescales longer than cadence spacings. Otherwise, characterization is very primitive.
- Beware of probabilities or False Alarm rates from irregularly spaced time series. Careful Monte Carlo analysis may be needed.
- Beware False Alarm rates from periodicity searches for two reasons: mathematically (very) uncertain; aperiodic processes often produce temporary quasi-periodicities.
- A vast array of statistics codes are already written, most in R/CRAN and Matlab (Signal Processing & Econometrics toolboxes). Can wrap R/CRAN from Python.