CHANCE

ISSN: 0933-2480 (Print) 1867-2280 (Online) Journal homepage: https://amstat.tandfonline.com/loi/ucha20

ARIMA for the Stars: How Statistics Finds Exoplanets

Eric D. Feigelson

To cite this article: Eric D. Feigelson (2019) ARIMA for the Stars: How Statistics Finds Exoplanets, CHANCE, 32:3, 35-40, DOI: 10.1080/09332480.2019.1662699

To link to this article: https://doi.org/10.1080/09332480.2019.1662699

Published online: 06 Sep 2019.

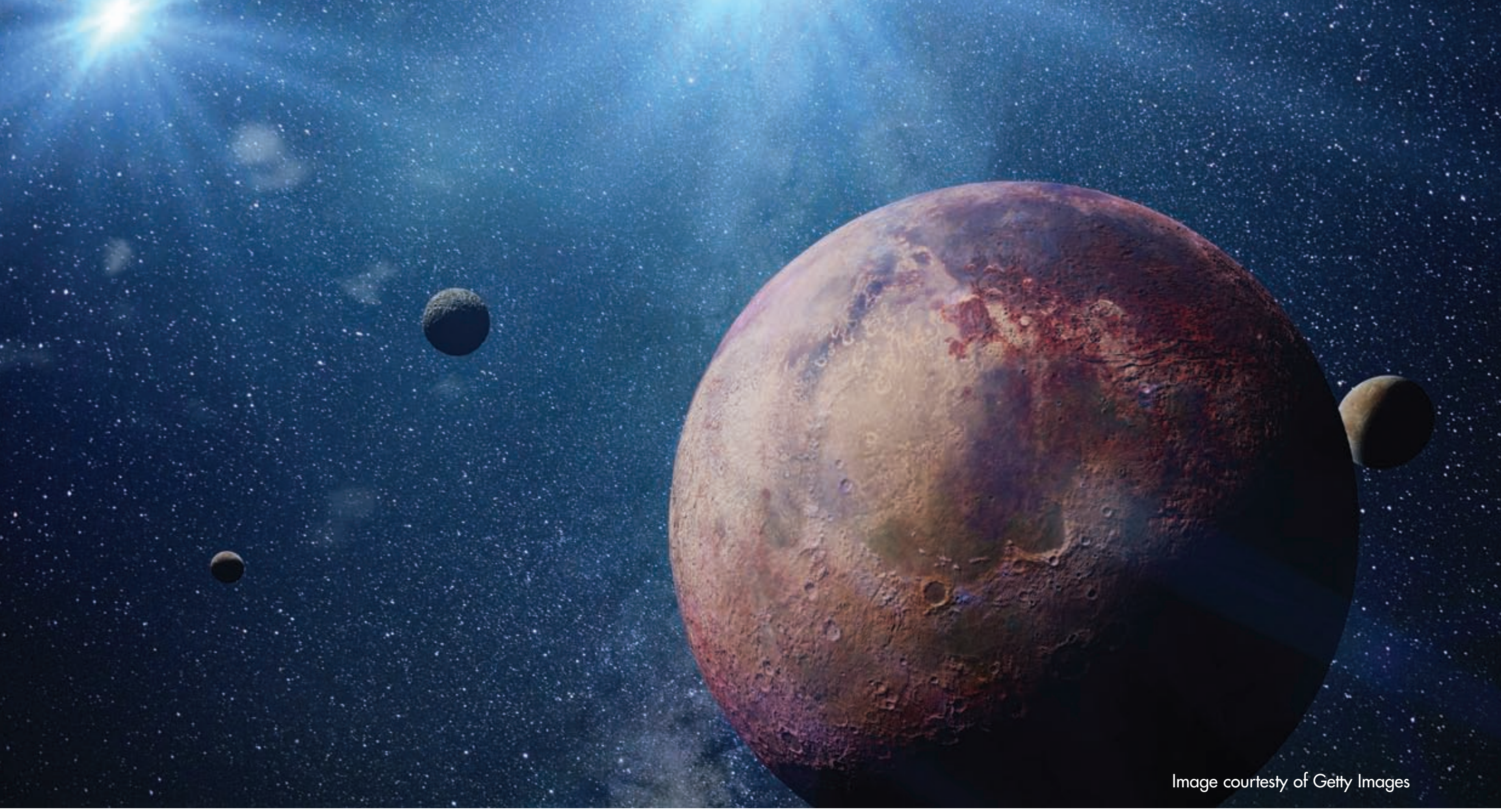

Image courtesty of Getty Images

ARIMA for the Stars: How Statistics Finds Exoplanets

Eric D. Feigelson

Planet, Planets Everywhere!

In the first century BC, the Roman poet Lucretius wrote, "you are bound to admit that in other parts of the universe there are other worlds inhabited by many different peoples and species of wild beasts." In the 17th century, Bernard de Fontenelle popularized the plurality of worlds with a best-seller among the nobility of Versailles. A generation ago, astronomer Carl Sagan entranced millions with lyrical phrases like "Think of how many stars, and planets, and kinds of life there may be in this vast and awesome universe."

From a scientific perspective, all of this was fantasy. While astronomers had established in the 19th century that the stars at night were luminous gaseous spheres like the sun, there was no evidence at all for planets orbiting the stars—until 1995. At that point, Swiss astronomers Michel Mayor and Didier Queloz had measured subtle Doppler shifts in the wavelengths of spectral lines from the starlight of 51 Pegasi, a nearby sun-like star. The observed periodic sinusoidal pattern in the star's motion indicated the star is being pulled back and forth by an invisible Jupiter-like planet orbiting every 4.2 days.

Their discovery was revolutionary, giving birth to the field of exoplanetary astronomy that is now a significant segment of all astronomical research effort.

Today, we know that most stars seen in the nighttime sky have planetary systems. Quantitative evaluations are still uncertain, but the best estimates are that stars typically have about five planets and perhaps 1% to 10% of them have Earth-like planets in Earth-like orbits where life could exist on the surface. (This fraction is known as $\eta_{\oplus}$, voiced as "eta-Earth.")

It can be inferred that there are hundreds of millions of habitable planets in the Milky Way galaxy,

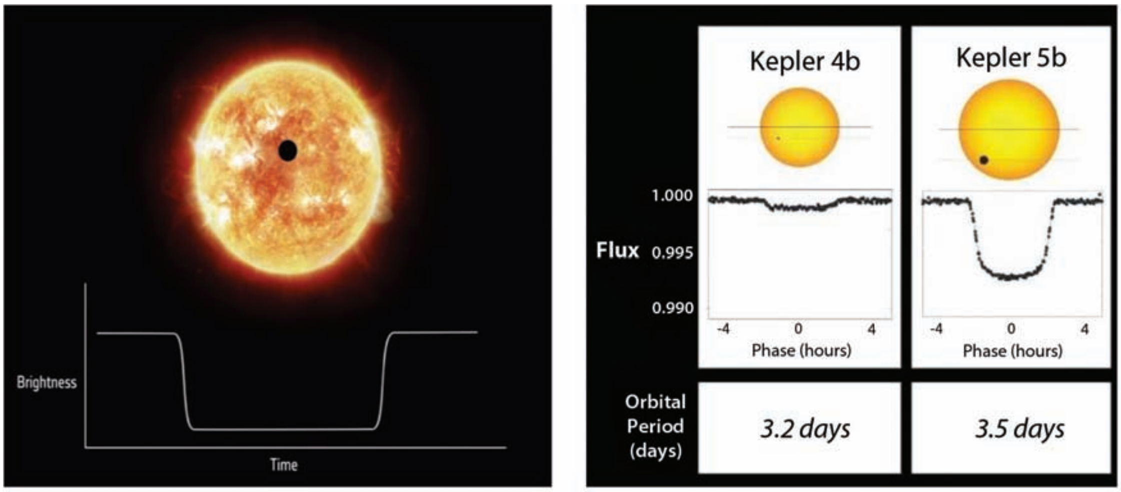

Figure 1. Diagram of an exoplanet transiting a star and the dip in brightness that occurs every orbit (left). Transit light curves from two planets found early in the Kepler mission, one a Neptune-size and the other Jupiter-size (right). Transits typically have amplitude of 0.01% to 1%, so other sources of brightness variations often overwhelm the planetary signal. (NASA/Kepler mission.)

the closest only a few lightyears away. No evidence for extraterrestrial life—Lucretius's wild beasts—has emerged, but the discovery of ubiquitous exoplanets motivates powerful research efforts directed toward this goal.

Astrostatistics and Exoplanets

Orbital models of multiplanet systems rely on sophisticated Bayesian nonlinear regression procedures where astrophysical processes constrain the prior distributions of model parameters. In part, the link between astronomy and statistics arises from the overabundance of the underlying populations. Of the billions of planets in our galaxy, we have so far sampled only a few thousand. Statistics is crucial here; a few years ago, two research groups analyzing the same data set arrived at $\eta_{\oplus}$estimates that differed by a factor of 7 due to different statistical analysis procedures.

Statistical inference is also critical for the detection and characterization of exoplanets because the signals are so weak. One effective method for discovering exoplanets detects changes in the velocity of the star toward and away from Earth as the planet orbits the star. Earth pulls the sun only 9 centimeters per second, so detecting its Doppler signature requires a spectrograph with better than 1-in-a-million precision measurements that is stable for years.

This is still beyond current engineering capabilities, but the latest astronomical spectrographs can detect Jupiter-mass planets out to Earth-like orbits, or Earth-like planets out to Mercury-like orbits.

A second effective method for discovering planets is somewhat less-demanding on our instruments. When an orbiting planet passes (or "transits") in front of a star, a small portion of the starlight is briefly eclipsed. For a Jupiter-size planet, the light is diminished by roughly 1%; for an Earth-size planet, the diminution is around 0.01%. Measuring star brightnesses (called "stellar photometry") to this precision is quite feasible today.

Figure 1 illustrates the dip in starlight produced by a planetary transit. In addition to various mountain-top telescopes, several satellite observatories are devoted to transit searches: Kepler and TESS missions launched by the U.S. National Aeronautics and Space Agency and the Cheops mission from the European Space Agency.

Since only a small fraction of orbits produce transits when viewed from any particular direction, many stars must be photometrically monitored for months or years. A variety of large-scale transit surveys are now underway, from both space-based satellite observatories and ground-based mountaintop observatories.

The Challenge of Planetary Transit Detection

A decade ago, NASA launched the Kepler mission with high confidence it would find many Earth-like planets, thanks to its instrumental precision (around 0.003%) and planned four-year

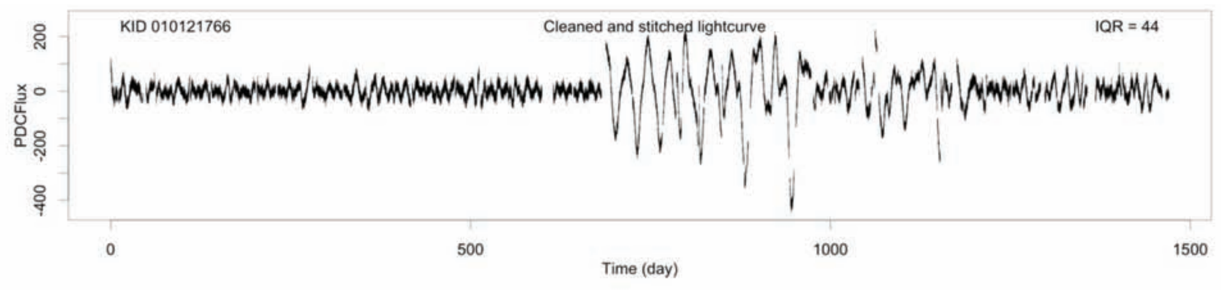

Figure 2. A four-year light curve from NASA's Kepler mission showing stellar brightness variations measured every half-hour during 2009–13. KID = Kepler identification number for this star, PDCFlux = star brightness after the "Pre-Search Data Conditioning" module is applied to reduce instrumental effects, and IQR = InterQuartile Range. In this star, the magnetic activity suddenly increased producing quasi-periodic variations from rotationally modulated starspots after about two years of observation. A low-dimensional ARFIMA model produced near-Gaussian white noise residuals but, as with most stars, no transiting planet was found here. (This plot and Figure 3 plots, obtained by Caceres, et al. 2019.)

photometric survey of ~200,000 stars. I did identify several thousand larger planets—most Kepler discoveries are super-Earth-size planets with orbital periods between two and 100 days—but hopes were foiled for the smaller planets.

It was the stellar variability that caused the problem: More stars exhibited brightness changes than expected. They mostly arise from magnetic activity and convection as seen on the sun and manifested as cool starspots, hot faculae, and flares.

While some stellar time series (called "light curves" by astronomers) can be quiet, others can show complex nonstationarity variations. The temporal behaviors are not simple: stochastic autocorrelated variations from superposed microflares, quasi-periodicities as starspots rotate in and out of view, and explosive white light flares. The phenomena cannot be modeled realistically astrophysically and are thus unpredictable. Figure 2 shows a typical nonstationary Kepler light curve.

The typical procedure has two stages: reducing the unwanted stellar variations in the time domain and searching for periodic transits in the frequency domain. The initial reduction of stellar variability in the time domain is generally performed with nonparametric procedures: wavelet analysis (as in the official Kepler pipeline) and high-dimensional local regression procedures like Gaussian Processes regression are most-often applied. A few researchers use advanced signal processing methods like Independent Component Analysis, correntropy, Empirical Mode decomposition, or Singular Spectrum Analysis.

Once the star brightness changes have been reduced, statistical procedures for finding periodic dips in brightness due to a transiting planet are well-established. Fourier transforms are not efficient because the transit shape is not sinusoidal; a "box least-squares" (BLS) matched filter for transit-shape dips is used instead. A BLS periodogram is constructed, light curves are "folded" (plotted modulo) to the most prominent spectral peak and are "vetted" by expert astronomers to make sure they exhibit properties expected of transiting planets.

This procedure is beset by four big problems:

1. It takes a time domain procedure that can treat a wide variety of unpredictable behaviors.
2. It requires a frequency domain procedure to find very faint transit-shaped periodic dips produced by small, Earth-sized planets. Periodograms have non-Gaussian power distributions that can easily give rise to spurious spectral peaks.
3. If the vetting stage is seen as a classification procedure, then the classes are badly imbalanced, with dozens of non-transiting cases for each transit. Even an occasional misidentification of a spectral peak with noise can flood the community with false alarms, wasting valuable telescope time for astronomical follow-up studies.

TERMS AND DEFINITIONS

ARIMA: In AutoRegressive Integrated Moving Average (ARIMA) time series models, the AR component treats dependence on recently past brightness values, MA component treats dependence on recently past brightness changes, and the I represents the differencing operation to reduce non stationarity. Fractional differencing in ARFIMA treats long-memory correlations corresponding to the astronomers' $1/f^{\alpha}$ "red noise." A periodic transit can be added as an exogeneous variable to the aperiodic ARIMA model using the ARIMAX formalism. We calculated model fits using the forecast CRAN package in the R statistical software environment developed by statistician Rob Hyndman and colleagues.

Box Least Squares and Transit Comb Filter: The BLS and TCF periodograms are frequency domain alternatives to the Fourier periodogram that are designed to detect periodic dip and spike shapes expected from planetary transits, rather than the periodic sinusoidal shape of Fourier analysis.

Random Forest: RF is an effective classifier using ensembles of decision trees based on "features" drawn from the data set and analysis by the scientist. Important advantages of RF for this problem include: a metric-free design permitting features with different scales and units; self-validation using out-of-bag sampling in each tree of the forest; and balanced treatment of imbalanced training sets.

Time series diagnostics: At each stage of the time series analysis, the analyst can examine how close the data are to white (uncorrelated) Gaussian noise. These include the Anderson-Darling test for normality, Ljung-Box portmanteau test for autocorrelation, and adjusted Dickey-Fuller test for stationarity.

4. Some stars have periodic variability for other reasons and disguise themselves as transiting planets. These include (quasi-)periodic variations from convection or pulsations in some stars, and eclipsing binary star systems that contaminate the target star light curve. Light curves with real, but non-planetary, periodicities are called astronomical false positives.

Penn State's Approach: AutoRegressive Planet Search (ARPS)

We have tried a different approach to stellar variability reduction based on a classic technique from time series analysis known as Box-Jenkins analysis. A low-dimensional, often-linear model is constructed where the brightness is not a function of time, but of recent past brightness levels and past changes (see sidebar). These are known as ARIMA models and are generally fit by maximum likelihood estimation, with model complexity determined by maximizing the Akaike Information Criterion.

However, the Kepler light curves are often nonstationary where the mean levels change quasi-periodically or in other peculiar ways (Figure 1). ARIMA models with a differencing operation treat many forms of non stationarity. If long-memory processes are present, ARFIMA models can be used.

The ARIMA residuals are then searched for periodic transits, but the differencing operator converts a box-like dip into ingress and egress spikes. Gabriel Caceres, while a graduate student at Penn State, developed a matched filter for a periodic sequence of double-spikes; he called this the Transit Comb Filter (TCF). The TCF periodogram is then examined for peaks representing repeated transit-like variations in the ARIMA residuals.

The ARPS procedure then faces a Big Data classification problem where the vast majority of Kepler light curves have no transit, and the small fraction with true planetary transits. Fortunately, NASA's Kepler Team recently released a "golden" list of transiting planet candidates, many confirmed with telescopic study by the broader astronomical community. These serve as a "planet training set" for a multivariate classifier, while random light curves (supplemented with sets of confirmed false positives like eclipsing binaries) serve as a "non-planet" training set.

We developed a list of several dozen "features" for a Random Forest multivariate classifier. The features are drawn from the original light curve, ARIMA residuals, TCF periodogram, folded light curve, and various time series diagnostics. Interestingly, the most-important feature emerged from an ARIMAX fit that gave a unique measure of the statistical significance of the transit depth.

ARIMAX analysis (see sidebar) is used most commonly in econometrics and had never before been applied to astronomical problems.

We optimized a Random Forest (RF) decision tree classifier using ROC curves. We carefully chose a threshold of RF probabilities on scientific grounds, balancing the recovery of true positives with the need to avoid the potentially overwhelming number of false alarms and false positives from the imbalanced classes.

Finally, we vetted the light curves satisfying the RF threshold for problems and identified promising cases. Several dozen

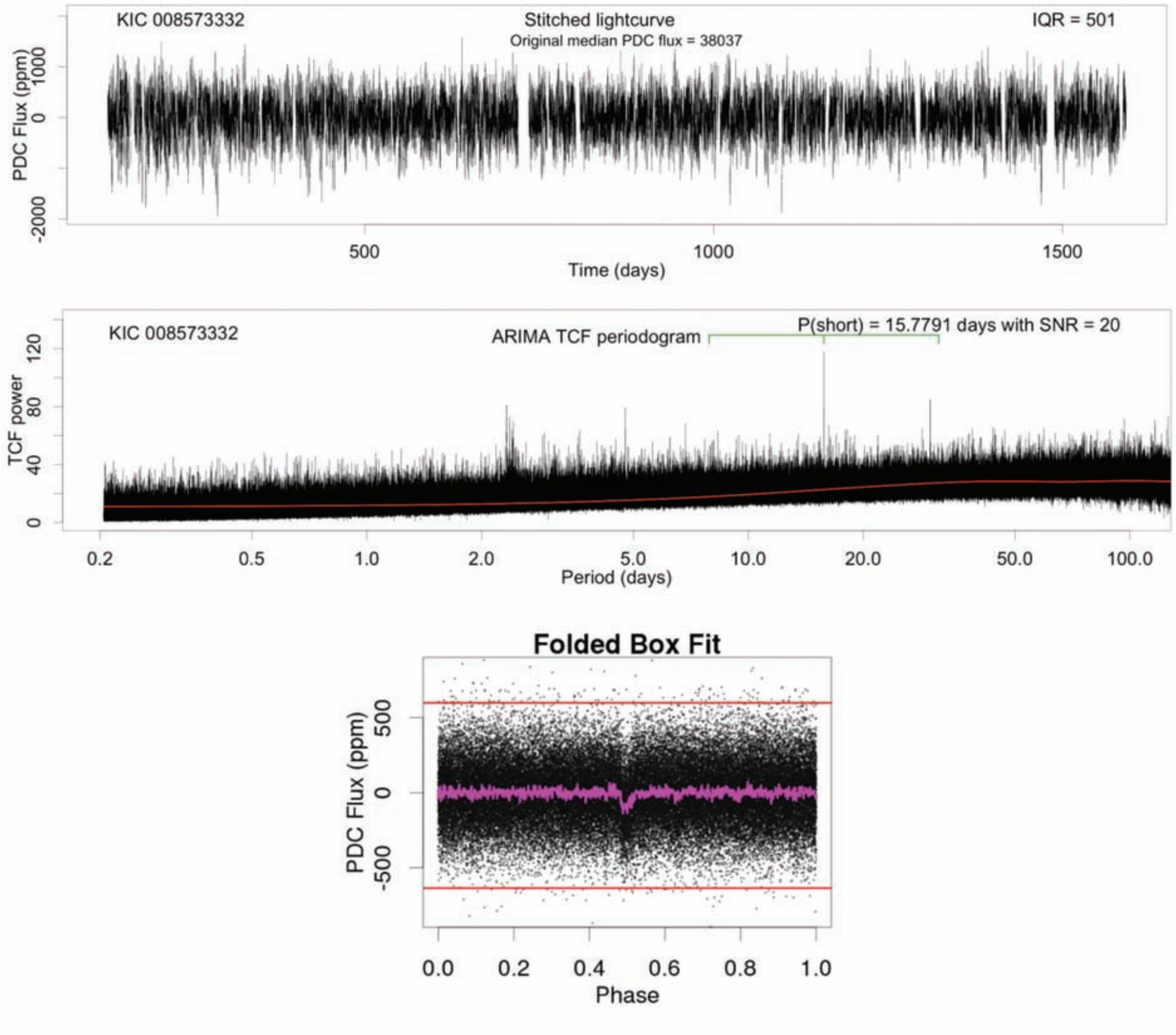

Figure 3. Light curve of a star with strong, choppy autocorrelation in units of brightness parts per million (ppm) (top). The TCF periodogram of ARIMA residuals shows a promising peak at period days (middle). The folded light curve, seen here as residuals of an ARIMAX model, shows the faint box-shaped dip expected from a planetary transit (bottom). The units of flux are ppm.

new planetary candidates emerge from this effort, most of them Earth-sized with very short orbital periods (0.2–10 days). These are planets so close to the host star that the rock surface itself is probably molten.

Figure 3 shows one of these promising discoveries. The middle panel shows a TCF spectral peak corresponding to an orbital period of 15.8 days, and the bottom panel shows a transit depth around 100 ppm corresponding to a planet with 1% the radius of the host star. If real, this Earth-sized planet has a very hot, possibly molten surface.

The Future of ARPS

From the perspective of a time series analyst, our procedure may not be particularly innovative: Box-Jenkins ARIMA-type modeling has been widely used since the 1970s, TCF can be viewed as a variant of Fourier analysis, and Random Forest classification has been successful in many fields since the 2000s. But in astronomy, this is quite an unusual approach; astronomers have weak education in applied statistical methodology and, in particular, are not familiar with ARIMA modeling.

The autoregressive planet search method outlined here can be developed further to improve the census of small planets, increasing sensitivity and reducing false alarms and false positives. Nonlinear ARFIMA and GARCH models give better fits to stellar variations in some cases.

Preprocessing by outlier removal can clean up some noisy periodograms. The TCF algorithm can be extended to treat gradual, rather than sudden, ingress and egress spikes. Newer machine learning classifiers like XGBoost can replace the Random Forest stage, and meta-classifiers might reduce the need for expert vetting.

Research groups in Israel and the U.S. are investigating the effectiveness of Deep Learning neural networks where the full database of light curves, rather than selected features, are entered into the classifier.

We have also investigated applying the ARPS approach to ground-based transit surveys where light curves of millions of stars are constructed. The problem here is the irregular spacing of the light curve time series data. Typically, a star can be observed only about eight hours per day at night, and is visible for only about six out of 12 months per year. The daily gaps in observation can be mitigated by placing telescopes on different continents or at the South Pole. We have simulated planet detection from ground-based surveys of this type and happily find that ARIMA and TCF methods work even with 70%–90% "missing data."

The Future of Statistics and Exoplanets

Exoplanetary research is in its third decade, but the work is still in its early phases. Nearly every aspect of the observational studies requires advanced statistical methodology:

- Algorithms for reducing stellar variations, both from the observational conditions and intrinsic to the star;
- Accurate subtraction stellar emission to reveal faint planetary emission in time series, astronomical spectra, and images;
- High-dimensional parametric modeling of multiplanetary orbits fitted to sparse data sets;
- Highly imbalanced classification problems from big surveys;
- Correcting observational selection biases to quantify underlying planetary populations; and
- Searching for biosignatures.

The field of astrostatistics must lead the way in future exoplanet studies, and this requires a combined effort of astronomers and statisticians. Astronomers traditionally have little training in statistics, and few statisticians are familiar with the scientific issues or are embedded in astronomical research teams. Astronomical projects fund "software" and "data analysis" computation, but not development of the methodology that underlies the computing effort, so cross-disciplinary collaboration is still rare.

The field of astrostatistics is constantly improving, with exponential growth in the use of machine learning and Bayesian approaches in the astronomical research literature. Many challenges are being effectively addressed, and many more will be faced in the future. If we ever find the extraterrestrial "wild beasts" predicted by Lucretius, it is likely that statistics will play a crucial role in the discovery.

Further Reading

Caceres, C.A., Feigelson, E.D., Babu, G.J., Bahamonde, N., Cristen, A., Meza, C., and Cure, M. 2019. AutoRegressive Planet Search, *Astronomical Journal*, in press arxiv:1901.05116, 1905.03766 and 1905.09852.

Feigelson, E.D., Babu, G.J., and Caceres, G.A. 2018. Autoregressive time series methods for time domain astronomy, *Frontiers of Physics* 6, #80. arxiv:1901.08003.

Haswell, C.A. 2010. Transiting Exoplanets. Cambridge, UK: Cambridge University Press.

Hyndman, R.J., and Athanasopoulos, G. 2018. *Forecasting: Principles and Practices,* 2nd edition. OTexts. *otexts.com/fpp2*.

About the Author

Eric Feigelson is professor of astronomy and astrophysics, and of statistics, at Penn State University. He has been working with G. Jogesh Babu and other statisticians for 35 years developing curriculum, offering summer schools, organizing meetings, and conducting research in astrostatistics.