Time Series in the AAVSO Database: Identification, Characterization and Utility

Brian Hill, Department of Physics & Astronomy, Saint Mary's College of California, Moraga, CA

George Silvis, Observer Code SGEO, 194 Clipper Road, Bourne, MA

Prepared for AAVSO / SAS Joint Session; June 15-17, 2017; Ontario, CA

Abstract

The AAVSO International Database (AID) holds over 30 million variable star observations and is currently growing at a rate of over 1 million observations per year. The nature and purpose of the observations is diverse and changing. We identify a subset of these observations as "candidate time series," characterize these in terms of their length (number of observations) and their duration (time from first observation to last observation). The characterization reveals a bimodality which we use to define classification as time series. We illustrate the utility of our classification by making some queries of interest more discriminating.

Context

The AAVSO comprises education, amateur/professional collaboration and scientific data collection. This last is most evident in the AAVSO International Database (AID) to which members submit data. We wish to explore what the AID can tell us about membership activities and how that is changing over time. What quickly becomes evident is that modes of observation strongly impact the analysis, and that an analysis based simply on number of observations submitted will be biased by observing modes that result in large numbers of records: e.g., time series for eclipsing binaries or exoplanets which often result in hundreds of observations in a single observing session. This preliminary study is to see if we can make a compelling classification of time series in order to better report descriptive statistics of membership activities and trends in the data.

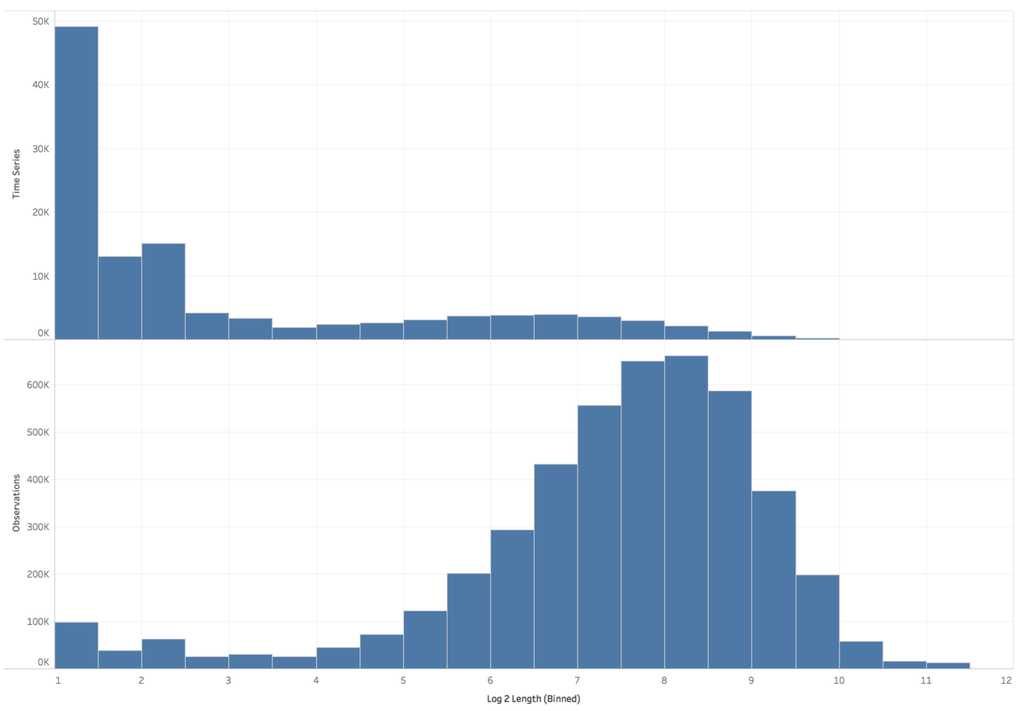
Identification

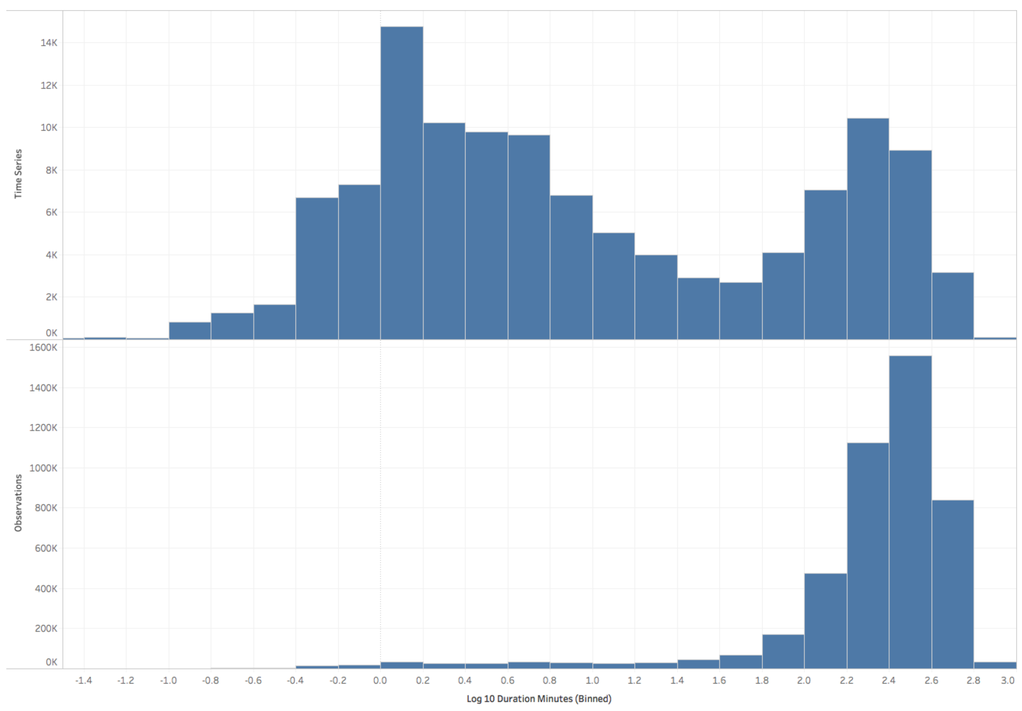
To enable a preliminary identification and analysis of time series, the AAVSO exported 4,820,120 observations ranging from Julian Date 2457022.5985 (2014-12-31 02:21:50.4 UT) to Julian Date 2457815.6840 (2017-03-03 04:24:57.6 UT). The export was limited to CCD, PEP and DSLR observation types. Each observation can have up to 41 attributes. Among the most important attributes for identification and analysis are the Julian date, the observer and the target. In the database, the Julian date is captured in the database column, "JD," the observer is captured in the column, "obscode," and the target or star name is captured in the column "name." All observations carry a unique id, captured in a four-byte integer column, "unique_id."

Candidates observations for inclusion in a candidate time series are required to be separated by a JD interval of less than 0.5 (12 hours), and have the same observer and the same name. A Python script using these criteria was created to run over all observations in the database and populate a table of candidate time series. Candidate time series must also have non-zero duration and at least two observations. 117,829 candidate time series were found. These time series include an overwhelming majority (4,576,283 out of 4,820,120) of the observations in the database export (94.9%). Candidate time series were ordered ascending by Julian date and all observations in the candidate time series were assigned a "timeseries_id" which is equal to the unique id of the earliest observation in the time series. In addition to its times series id, each time series has a length (the number of observations in the series), and a duration (the time elapsed from the first observation to the last observation in the series).

Characterization

As noted above, we characterize candidate time series by their length and their duration. Because time series with long lengths count for a disproportionate number of observations in the database, our visualizations are presented both in terms of the number of time series and in terms of the length-weighted number. Because both lengths and durations are positive quantities spanning multiple decades, our histograms bins are logarithmic:

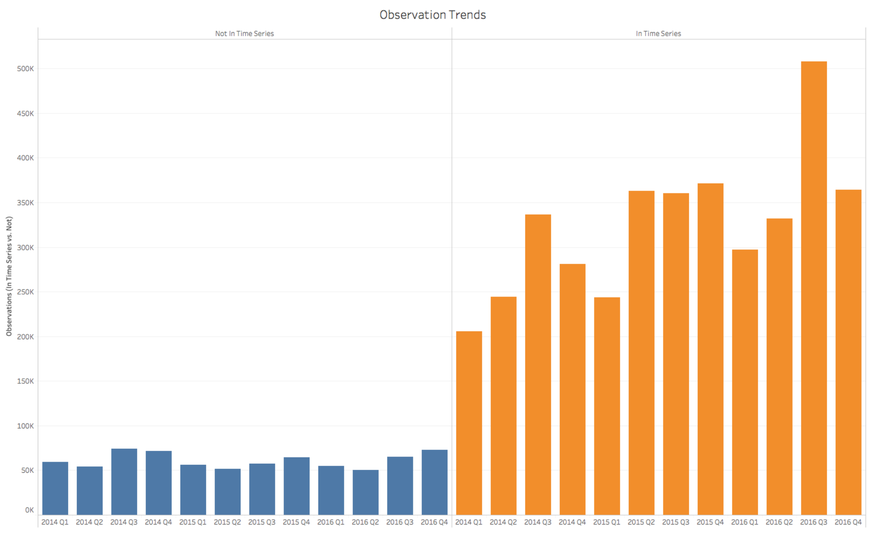


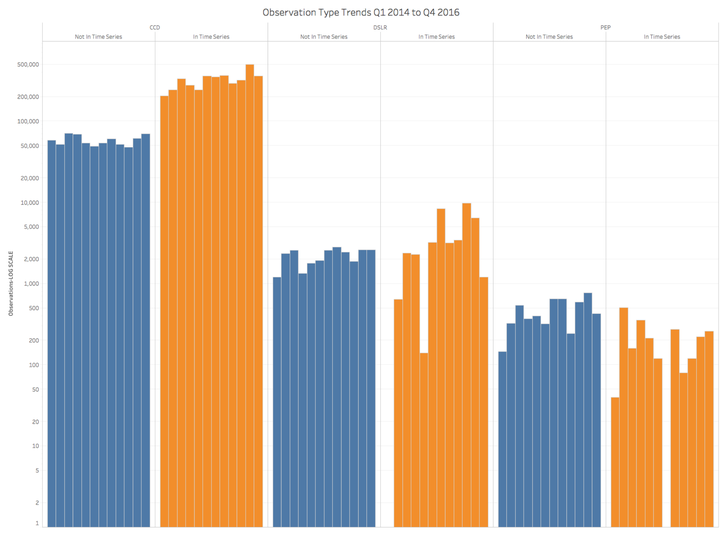


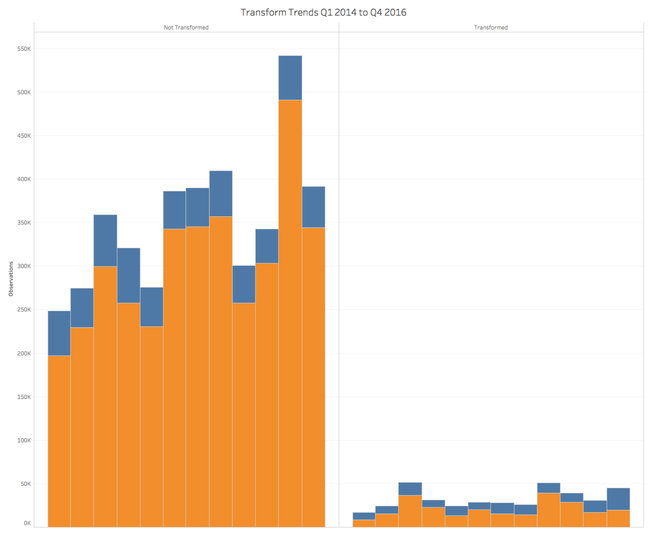
In the pair of plots on the left, we observe a bimodality in length with a minimum between the two populations occurring between 2^3 (8) and 2^4 (16) observations. Based on this, we choose to place a cut at 10 observations. In the pair of plots on the right, we observe a bimodality in duration with a minimum between the two populations occurring at about $10^{1.4}$ (25.1) and $10^{1.8}$ (63.1) minutes. We therefore choose to place an additional cut at 30 minutes. Additionally, a minimum cadence cut of 6 minutes was applied. To summarize, a time series is at least 10 observations, spanning at least 30 minutes, occurring at a cadence never less than 10 observations per hour. We will simply refer to such time series as "time series" rather than as "candidate time series." There are 35,667 such time series containing the great majority (4,060,381 out of 4,820,120) of the observations in the database export (84.2%).

Utility

To illustrate the utility of the above characterization of time series, we make a few commonly performed queries more discriminating by performing them separately based on time series membership. In all three plots below, observations that are part of a time series are in orange. Observations that are not part of any time series are in blue:







Among the things that can be observed in the preceding three plots are that non-time-series observations are roughly flat in the 12 quarters shown. Time series observations may have grown by as much as a factor of two in this time. Time series dominates CCD data collection, but not DSLR or PEP data collection. Note that scales were necessitated by the vast difference in number between CCD, DSLR an PEP data collection. Finally, transformed data is a growing fraction of non-time-series observations, but a flat or possibly declining fraction of time-series observations. It would be extremely interesting to extend the trend graphs back in time, and as new data comes in to continue adding additional quarters.

Future

Provided that the AAVSO membership finds the above criteria for time series to be sensible and interesting, we propose to expand the analysis to the entire AAVSO database, and to record the Python script results in a new table so that it is readily available for further database queries. Email the authors, Brian Hill at brh3@stmarys-ca.edu and George Silvis at george@gasilvis.net, with any comments or suggestions on this work and future directions.