Astronomy and Computing 11 (2015) 126–137

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

AstroStat—A VO tool for statistical analysis

A.K. Kembhavi [a], A.A. Mahabal [b], T. Kale [a], S. Jagade [a], A. Vibhute [a], P. Garg [a], K. Vaghmare [a,*], S. Navelkar [a], T. Agrawal [c,1], A. Chattopadhyay [d], D. Nandrekar [e,2], M. Shaikh [f,2]

[a] Inter-University Centre for Astronomy and Astrophysics, Pune, India
[b] California Institute of Technology (Caltech), Pasadena, USA
[c] Symantec, Pune, India
[d] Department of Statistics, Calcutta University, Kolkata, India
[e] John Hopkins University, Baltimore, MD, USA
[f] Synygy, Pune, India

ARTICLE INFO

Article history:
Received 5 May 2014
Received in revised form
31 October 2014
Accepted 17 February 2015
Available online 7 March 2015

Keywords:
Virtual observatory tools
Methods: statistical

ABSTRACT

AstroStat is an easy-to-use tool for performing statistical analysis on data. It has been designed to be compatible with Virtual Observatory (VO) standards thus enabling it to become an integral part of the currently available collection of VO tools. A user can load data in a variety of formats into AstroStat and perform various statistical tests using a menu driven interface. Behind the scenes, all analyses are done using the public domain statistical software—R and the output returned is presented in a neatly formatted form to the user. The analyses performable include exploratory tests, visualizations, distribution fitting, correlation & causation, hypothesis testing, multivariate analysis and clustering. The tool is available in two versions with identical interface and features—as a web service that can be run using any standard browser and as an offline application. AstroStat will provide an easy-to-use interface which can allow for both fetching data and performing power statistical analysis on them.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

AstroStat[3] is a powerful VO compatible tool, developed by the Virtual Observatory—India (VOI) project, for statistical analysis of data. It provides a number of statistical tests, ranging from the simple to the more complex and sophisticated, which are performed using a very simple to use graphical interface. The analysis is carried out using the highly developed statistical package R, which is available in the public domain. AstroStat uses in-built graphics for easy visualization of the data as well as the results of the tests performed. It incorporates various VO standards, so that it can easily be linked to a wide range of VO tools like the plotting and visualization tools VOPlot and TOPCAT and can use the Astronomical Data Query Language to obtain data from VO compatible services for statistical analysis.

AstroStat has evolved from the statistical analysis tool VOStat, which was first developed through a collaboration between groups from Caltech and Pennsylvania State University and later through collaboration between these two groups and VOI. VOStat is available as a web-service from the Centre for Astrostatistics at Penn State.[4] AstroStat has been developed as an independent tool by VOI, in collaboration with a group from Caltech, with important inputs from various astronomers, statisticians and software engineers.

The AstroStat code is made of two parts—the main backbone code written in Java and the R snippets which are made available to the user when a test is run. Both these codes are being made available to the community under GNU GPL license agreement.[5]

* Corresponding author.
E-mail addresses: akk@iucaa.ernet.in (A.K. Kembhavi), kaustubh@iucaa.ernet.in (K. Vaghmare).

[1] Author was affiliated with Persistent Systems Ltd., during the development of AstroStat.

[2] Author was affiliated with the Inter-University Centre for Astronomy and Astrophysics during the development of AstroStat.

[3] http://voi.iucaa.ernet.in:8080/astrostat.

[4] http://astrostatistics.psu.edu:8080/vostat/.

[5] The source code can be obtained by mailing a request to voindia@iucaa.ernet.in.

http://dx.doi.org/10.1016/j.ascom.2015.02.004
2213-1337/© 2015 Elsevier B.V. All rights reserved.

The present article is organized as follows. In Section 2, we provide an overview of the tool and in Section 3, the details of R as a statistical backend are discussed. In Sections 4 and 5, we cover the inner implementation details of AstroStat including descriptions of various VO standards. In Section 6 we provide an illustrative application of AstroStat and in Section 7 briefly discuss future directions.

2. An overview of AstroStat

AstroStat comes in two flavors—an offline version[6] bundled in the form of an executable Java Archive (.jar) and a web version which can be run in any standard browser. The interface, which has been designed with ease-of-use in mind, has been kept the same in both the versions. The primary interface comprises of three ever-present sections—(i) which enables the user to load data, (ii) a collection of tests categorized into Exploratory, Advanced and Expert, and (iii) a help section which presents a description of the currently selected test with examples and any extra notes. Section 4 appears on selecting a test and this provides options to select and transform columns, supply necessary parameters to the test (e.g. type of correlation when computing a correlation matrix), choose the nature of output etc. 2.

The typical workflow, from the end user's perspective, is shown in Fig. 1. The user first loads data into the application, in the form of a file either on the local hard drive or on a web server. Data can also be loaded using the Table Access Protocol (TAP) (Dowler et al., 2011) or through Simple Access Message Protocol (SAMP) (Taylor et al., 2011), as described in detail in Section 5. It is possible to load more than one file at a time and a list of all loaded files is available in the form of a drop-down menu. As a next step, the user selects one of the three categories of tests and a test within it. A complete list of all tests available can be found in Appendix A. The Help section updates itself to reflect the currently selected test and offers a quick overview of what the test does, possible examples and special notes, if any. When a test is selected, Section 4 appears where a user inputs parameters required by the test. Once done, the user clicks *Run Test* and AstroStat performs the analysis and displays the output in a tabular form with tooltips to aid interpretation.

Since all the four sections described above are always visible, the user can easily run another test or the same test with modified input parameters, or refer to the help section for a quick reminder of say, what exactly the output means, etc. The output is in a friendly and neatly formatted form and can be easily saved. The plots can be saved into a single ZIP file while the tables and other output data can be stored in a plain ASCII text format.

In addition to these features, AstroStat also offers other functional features like:

- A quick-look summary statistics pop-up for the currently loaded data.
- Ability to view both the tabular version and the original file. This allows the user to ensure that the data have been loaded correctly.
- The user can define new columns by performing common operations on existing columns. (e.g. sum of two columns, square of a column, etc.)
- One click access to the VOPlot service (Kale et al., 2004) for interactive plotting and data visualization.

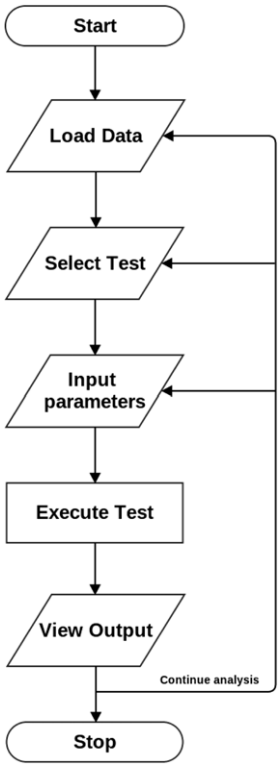

Fig. 1. A flow chart illustrating the user perspective of the workflow in AstroStat.

- Ability to view the R code used in the actual analysis so that a user may build upon this code for further work. If the user wishes to modify the R code provided to perform further analysis, this will have to be done outside of AstroStat in an R shell. The R code is provided under the GNU GPL license. At the time of writing this article, the R code provided by the web version to the user includes a lot of code which is especially needed for a seamless interaction between AstroStat and R. In a future release, we will clean the code being served to the user so that it can become easier for the user to modify it.

In the subsequent sections, we describe the detailed implementation and features of the tool.

3. Statistical backend

The R language (Ihaka and Gentleman, 1996) came into existence as a free counterpart of the S statistical language from Bell Labs. Like S, R (R Core Team, 2013) has all the common tools needed for advanced statistics: linear and non-linear modeling, various statistical tests, time series analysis, classification, clustering etc. Ross Ihaka and Robert Gentleman developed R with user participation in mind which has resulted in a very large number of contributions from the users. The Comprehensive R Archive Network (CRAN)[7] hosts the user packages and has easy interfaces to download and install any of the packages from geographically distributed mirror sites. In early 2014 the count has crossed 5000 packages. As it is arguably the most versatile open-source system for statistics we decided to use it as the backend for the AstroStat service. The original collaboration for developing such a service was between Caltech and Penn State with the coding to be done at Caltech (Mahabal et al., 2002; Graham et al., 2005). Part of the undertaking was to provide users with a set of tools as well as broad

[6] IMPORTANT: The AstroStat stand-alone or offline version is still in development. While the application can still be downloaded from http://voi.iucaa.ernet.in/~voi/AstroStat.html, it is not yet ready for the end user.

[7] http://cran.r-project.org/.

Download English Version:

https://daneshyari.com/en/article/10349343

Download Persian Version:

https://daneshyari.com/article/10349343

Daneshyari.com