ASTRONOMICAL DATA ANALYSIS SOFTWARE AND SYSTEMS XIV
ASP Conference Series, Vol. 347, 2005
P. L. Shopbell, M. C. Britton, and R. Ebert, eds.

VOStat: A Distributed Statistical Toolkit for the Virtual Observatory

Matthew J. Graham, S. G. Djorgovski, A. A. Mahabal, and Roy D. Williams

California Institute of Technology, Pasadena, CA 91125, USA

G. Jogesh Babu, Eric D. Feigelson, and Daniel E. Vanden Berk

The Pennsylvania State University, University Park, PA 16802, USA

Robert Nichol

ICG, University of Portsmouth, PO1 2EG, UK

Larry Wasserman

Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract. The nature of astronomical data is changing: data volumes are following Moore's law with a doubling every 18 months and data sets consisting of a billion data vectors in a 100-dimensional parameter space are becoming commonplace. Sophisticated statistical techniques are crucial to fully and efficiently exploit these and maximize the scientific return. A long-standing limitation, however, on the range and capability of such analyses has been the paucity of non-proprietary software.

VOStat is the result of a cross-disciplinary collaboration between astronomers and statisticians to meet these challenges; it is a prototype knowledge-based statistical toolkit implemented within the VO paradigm for the entire astronomical community. VOStat consists of an easily extensible distributed web-based framework transparently accessed via a single science endpoint.

An exploratory science application is presented to demonstrate some of the functionality currently offered by VOStat.

1. Introduction

The use of statistics in astronomy is commonplace: of the 15000 astronomical studies carried out each year, 5% explicitly mention "statistics" in their abstract whilst 20% consider variable objects or multivariate datasets. However, the statistical methodologies that are predominantly employed in these studies predate the Second World War:

- Fourier transform (Fourier 1807)
- Least squares (Legendre 1805), Chi-squared (Pearson 1901)
- Kolmogorov-Smirnov (Kolmogorov 1933)
- Principal Component Analysis (Hotelling 1936)

As a whole, astronomy currently produces ~ 1 TB of data per night and within a decade, the LSST[1] alone will produce $\sim$ 13 TB per night. If the prospect of Petabyte-sized data archives were not daunting enough, each data point will occupy a position in a parameter space consisting of several hundred dimensions. Successfully data mining these data sets mandates new sophisticated statistical techniques that are easily accessible to the general astronomical community and implemented in a distributed fashion to take full advantage of the power of the Virtual Observatory and the Grid.

2. VOStat

VOStat[2] is a cross-disciplinary collaboration between astronomers, statisticians and VO scientists specifically to address these issues. It consists of both a pedagogical component to teach astronomers how to apply statistical methods properly, e.g. when is it appropriate to use the Kolmogorov-Smirnov test to determine goodness-of-fit and when is it more appropriate to use the Anderson-Darling test, and a software component to offer them easy access to such methods.

VOStat is implemented upon an extensible distributed web-based framework so that it is simple to expose new functionality, which could be legacy applications, whilst allowing software to run in its optimal environment (hardware and software) without unnecessary porting. By using the VOTable[3] standard as the default data format, data metadata can be easily passed around to permit efficient process initialisation prior to computation whilst the data itself only needs to be transferred when required.

3. Accessibility

The complexity of the distributed network is hidden via access through a single science gateway. Three interfaces to the gateway are being implemented:

- an interactive web form to allow users to play with the software and test data sets
- web services for developers who want to incorporate the software within their own applications in exactly the same way as one does with an external library function
- module/plug-in for popular VO data exploration tools, such as Mirage[4], VOPlot[5] and DataScope[6].

---

[1] `http://www.lssto.org`

[2] `http://www.vostat.org`

[3] `http://www.ivoa.net/Documents/latest/VOT.html`

[4] `http://skyservice.pha.jhu.edu/develop/vo/mirage/default.aspx`

[5] `http://vo.iucaa.ernet.in/~voi/voplot.htm`

[6] `http://heasarc.gsfc.nasa.gov/vo/`

4. Functionality

VOStat currently provides access to selected functionality from the open source statistics package R[7] and multi-resolutional k-dimensional trees. The types of activity that can be carried out are:

- descriptive statistics (e.g. boxplot)
- two- and k-sample tests (e.g. Wilcoxon rank-sum)
- density estimation (e.g. kernel smoothing)
- correlation and regression (e.g. P.C.A.)
- censored data (e.g. survival)
- multivariate classification (e.g. H clustering)
- outlier detection (e.g. k-d trees)

5. Example: Outliers in Color-Color Space

The detection of outliers in a high-dimension parameter space will certainly be a common data mining activity with very large data sets. To illustrate how this might be achieved with VOStat, we took a sample of 1000 randomly-selected objects from the Palomar-Quest synoptic sky survey (Graham et al. 2004), each with 9 colors.

Boxplot A boxplot (Figure 1) shows the relationships between the colors in terms of the mean, median, overlap and outliers.

K-means clustering Figure 2 shows the relationships between the colors having identified 5 clusters in the data using K-means clustering.

Probability density association Figure 3 shows the $B-R$ vs. $R-I$ distribution for the objects identified as having the lowest probability of being associated with a cluster in the parameter space, i.e. highest probability of being an outlier, using k-d trees.

Visual inspection Figure 4 shows BRI plots of the highlighted object in Figure 3. It can clearly be seen to be a B-band dropout.

Acknowledgments. This work is supported in part by the NSF grant DMS-0101360.

References

Graham, M. J., et al. 2004, in ASP Conf. Ser., Vol. 314, ADASS XIII, ed. F. Ochsenbein, M. Allen, & D. Egret (San Francisco: ASP), 14

[7] `http://www.r-project.org/`

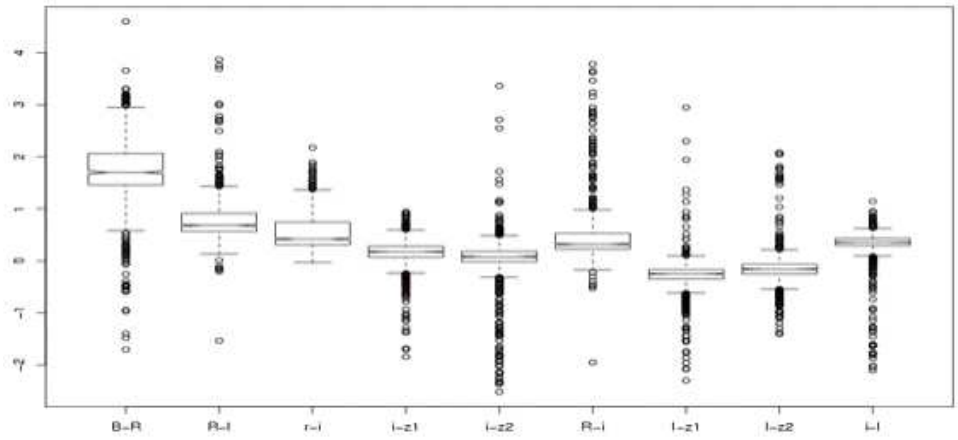

Figure 1. Boxplot of the data

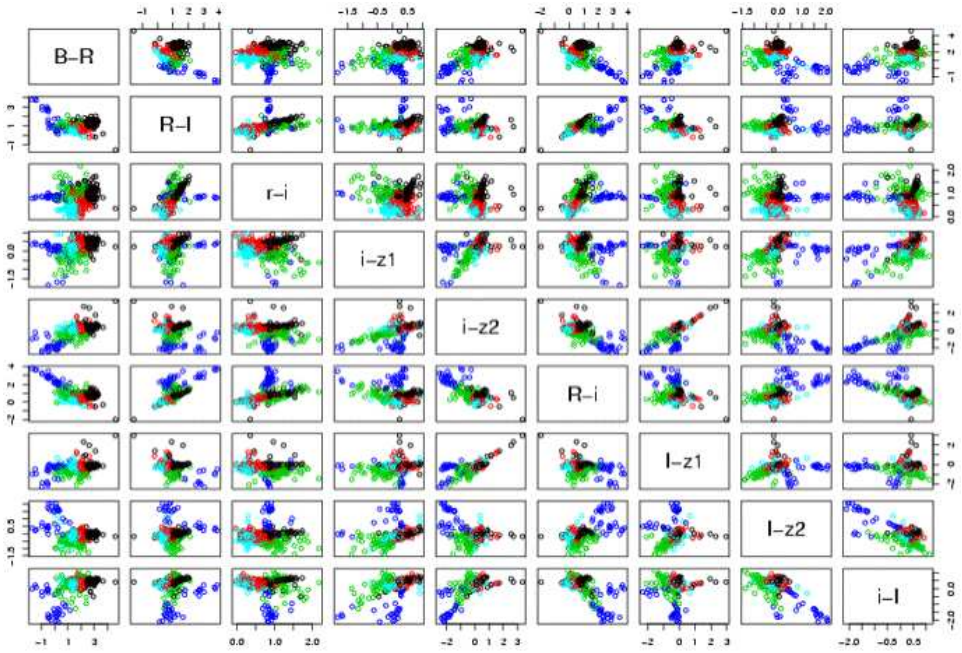

Figure 2. Plot of colors vs. colors having identified 5 clusters with K-means algorithm

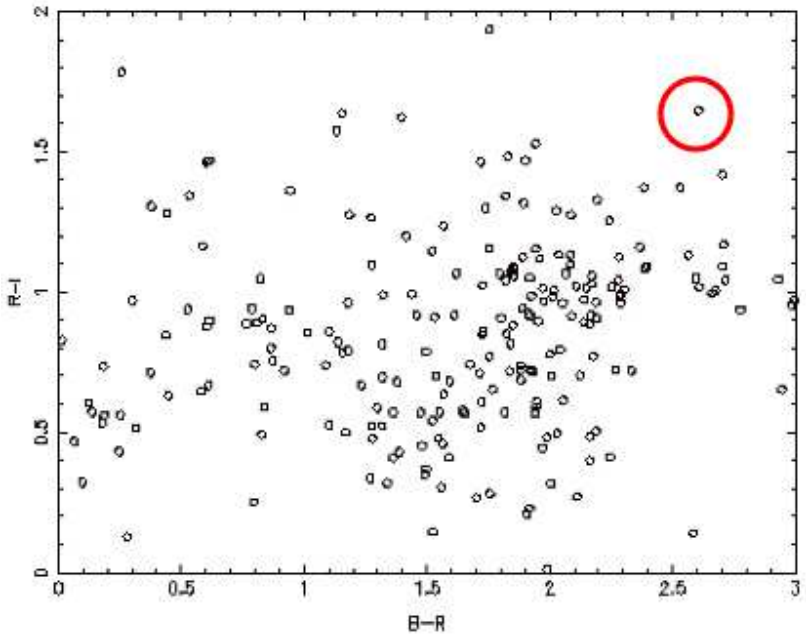

Figure 3. Color-color plot for lowest density association objects

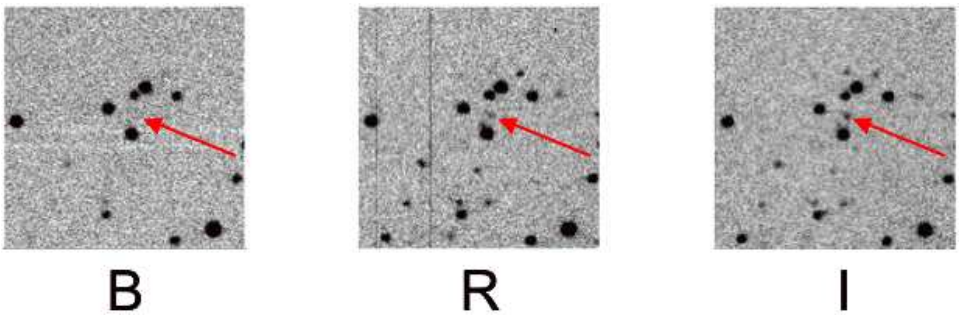

Figure 4. Visual plot of highlighted object