Journal of Statistical Software

January 2007, Volume 18, Issue 11. *http://www.jstatsoft.org/*

Using R-based VOStat as a Low-Resolution Spectrum Analysis Tool

G. Jogesh Babu
The Pennsylvania State University

Ashish Mahabal
California Institute of Technology

Abstract

We describe here an online software suite **VOStat** written mainly for the Virtual Observatory, a novel structure in which astronomers share terabyte scale data. Written mostly in the public-domain statistical computing language and environment R, it can do a variety of statistical analysis on multidimensional, multi-epoch data with errors.

Included are techniques which allow astronomers to start with multi-color data in the form of low-resolution spectra and select special kinds of sources in a variety of ways including color outliers. Here we describe the tool and demonstrate it with an example from Palomar-QUEST, a synoptic sky survey.

Keywords: Virtual Observatory, color-color plot, clustering, **VOStat** web services.

1. Introduction

The nature of astronomical data is changing: data volumes are following Moore's law and doubling every 18 months, and datasets consisting of a billion data vectors in a 100-dimensional parameter space are becoming commonplace. 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 were developed more than 50 years ago: *Fourier transform* (Champeney 1973), *Least squares* (Laplace 1820; Stigler 1977), *Chi-squared* (Pearson 1901), *Kolmogorov-Smirnov* (Chakravarti, Laha, and Roy 1967), *Principal Component Analysis* (Hotelling 1936). Some of these procedures are incorrectly used. For example Kolmogorov-Smirnov does not work for multidimensional data.

Sophisticated statistical techniques are crucial to fully and efficiently exploit and maximize the scientific return. A long-standing limitation, however, on the range and capability of such

analysis has been the paucity of non-proprietary software.

2. Virtual Observatory

Astronomical surveys currently produce ≈ 1 TB of data per night and within a decade, the *Large Synoptic Survey Telescope* (http://www.lsst.org/) alone will produce ≈ 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 and will include elliptical error-bars.

As the data volume and complexity of astronomical findings have enormously increased in recent decades, a paradigm shift is underway in the very nature of observational astronomy. While in the past a single astronomer might observe a handful of objects, today data mining of large digital sky archives obtained at all wavelengths of light is becoming a major mode of study. The astronomical community thus faces a key task: to enable efficient and objective scientific exploitation of enormous multifaceted datasets. In recognition of this need, the Virtual Observatory (VO) initiative has emerged as a top priority, from the NAS Taylor/McKee Decadal Report on astronomy for 2000-2010, to federate numerous large digital sky archives and develop tools to explore and understand these vast volumes of data.

Major efforts are underway around the world, including both NASA and the NSF, to federate huge, uniform, multivariate and image databases collected by specialized observatories. Most VO efforts have focused on computational aspects of data access and mining from distributed, heterogeneous databases. Successfully data mining astronomical datasets 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 VO. The VO is making astronomical data easier to use through the creation and adoption of standards. A VO standard that has been adopted worldwide in the VO community is VOTable, a way to represent a table of data in XML with good meta-data about the semantic meaning of the data.

After the scientists have collected the sub-datasets of interest, powerful statistical techniques should be brought to bear to help them make astrophysical inferences. We have begun this effort with the creation of a prototype **VOStat** (http://vostat.org/) web service where scientists at point A can interactively request a statistical analysis using software at point B on data at point C and receive near-real-time answers. **VOStat** can also be accessed from the Center for Astrostatistics web site (http://astrostatistics.psu.edu/). **VOStat** was developed under a Focused Research Group grant funded by the NSF Division of Mathematical Sciences led by PI Babu and work done at PSU, Caltech and CMU. But only a fraction of the vast repertoire of astrostatistical needs have been implemented. The main engine that drives **VOStat** is R (R Development Core Team 2006).

VOStat 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. Help files provide statistical background. For example, the help file for **Shapiro-Wilks test for normality** reads:

> As some results in statistics rely on the assumption that the underlying popula-

tion under consideration is Gaussian, or the assumption of 'asymptotic normality', tests for normality are valuable. While any goodness-of-fit test (e.g., chi-squared) would suffice, several specialized statistical tests have been developed to measure the proximity of the dataset to the normal distribution including the Lilliefors test (special probabilities for the Kolmogorov-Smirnov 1-sample test), Anderson-Darling test, Shapiro-Wilks test, Shapiro-Francia test, D'Agostino's D, Spiegelhalter's T, and Martin-Iglewicz's I. The Shapiro-Wilks statistic W can be viewed as the correlation coefficient between the squared ordered values and normal scores. This R function requires a vector input with $3 < N < 5000$ points. Background: The test is derived by S. Shapiro and M. Wilks (Shapiro and Wilks 1965). The algorithm used by R was developed by P. Royston (Royston 1995).

A link to additional statistics background along with formulae are also provided in the help file.

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. **VOStat** accepts data in ASCII as well as VOTable format. By using the VOTable standard as the default format, data meta-data can be easily passed around to permit efficient process initialization prior to computation whilst the data itself only needs to be transferred when required. **VOStat** also returns the R script used in the computations along with the result. The web interface has links to help files written specifically for astronomers. A set of advanced level help files with links to further reading are being implemented.

The scientist uses astronomical databases and requests specific statistical calculations from the **VOStat** server (Figure 1). The computation is carried out either on the server or on distant VO computers, and results are communicated back to the scientist. At present, **VOStat** offers a limited suite of R procedures (some simple like a Kolmogorov-Smirnov test, others complex like the AGNES multivariate classifier) and specialized procedures developed specifically for very-large VO applications. The functionalities will be expanded with the integration of additional general-purpose R functionalities, and specialized modules for space science which are now stand-alone codes such as **ASURV** (LaValley, Isobe and Feigelson 1992).

The web interface is written in Perl. The complexity of the distributed network is hidden via access through a single science gateway. Interfaces to the gateway are implemented through an interactive web form to allow users to play with the software and test datasets. Module/plug-in for popular VO data exploration tools, such as VOPlot are being implemented in collaboration with Inter University Center for Astronomy and Astrophysics (IUCAA). This is in addition to native R plot capability.

VOStat currently provides access to selected functionality from the open source language and environment for statistical computing R. The types of activity that can be carried out include: 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., Linear regression), Principle Component Analysis, censored data (e.g., survival), multivariate classification (e.g., Hierarchical clustering).

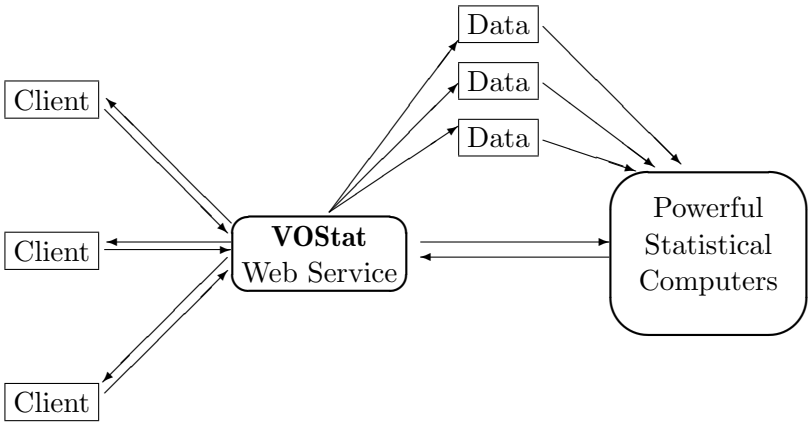

Figure 1: **VOStat** web services

3. Example: Looking for outliers

We use a dataset from the Palomar-QUEST sky survey to look for outliers (available at `http://astrostatistics.psu.edu/vostat/samples/questcols.vot`). The dataset consists of 1000 matched objects in a small region of the sky. The data are from two successive nights using Gunn *rizz* and Johnson *UBRI* filters (Figure 2). The different magnitudes act like a very low-resolution spectrum by providing information in broad bins over a few thousand Angstroms. We concentrate on the various colors obtained by taking ratios of the fluxes in the different bins (operationally we take differences of the magnitudes where a magnitude is defined as $m = -2.5 \cdot \log(\text{flux}) + c$, c being the zero point). Ordinary stars in our galaxy occupy a region in the color-color plot with a fairly well understood spread based on physical parameters like age, temperature etc. The color-color plot is referred to as the Hertzsprung-Russel diagram (HR diagram) and is often shown with some physical parameters forming a regular grid over the color-color plot. The aim here is to find objects that have atypical colors.

Any two filters can be used to form a color. Of all the colors possible in our dataset, we realize the nine where the two filters leading to the color either match or are close to each other in wavelength space in order to mimic a low-resolution spectrum. Besides these nine columns, the input file here has six others like ID, Date which do not pertain to the physical properties of the object. Choosing nine of the 15 columns for further processing is effected by a feature of the GUI (Figure 3) whereby we can choose a subset of columns from the input table. Since boxplot is very good at revealing the relationships between the different colors, including the mean, median, the overlap and the outliers for the set we start with using boxplot on the nine colors (Figure 4). That gives us an idea if certain colors can be fully ignored.

We then use hierarchical clustering to determine the clustering present in the data. The dendrogram (Figure 5) is used to determine where to make the cuts and then a list of possible outliers along with cluster centers and withinss is obtained using k-means (Figure 6) for those many clusters. k-means also allows for the visual inspection of the different clusters in terms of all the desired parameter pairs (Figure 7).

K-Density then associates a probability with each object about its being an outlier. We sorted

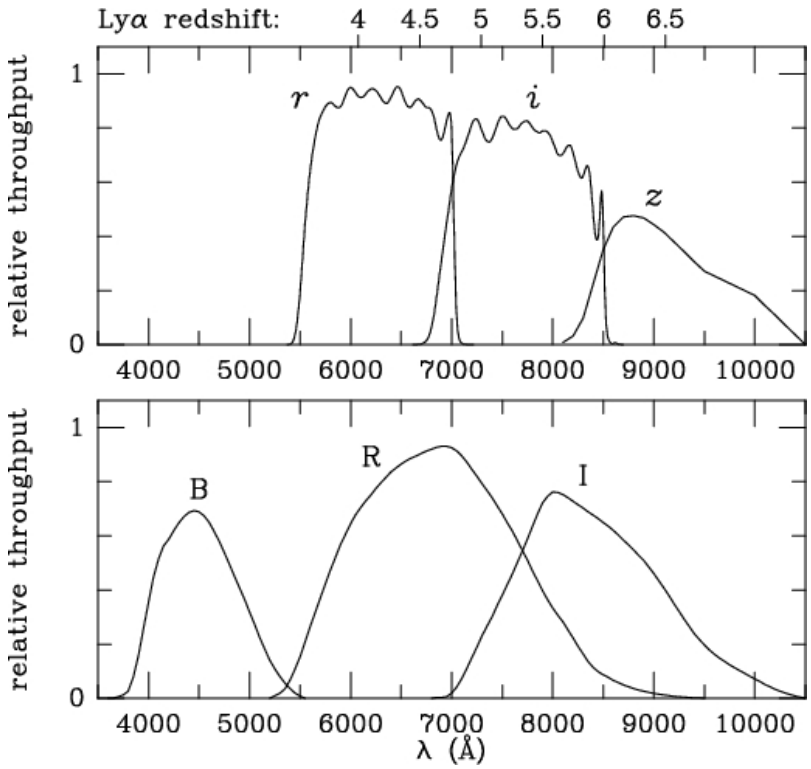

Figure 2: The wavelength range of Gunn r i z and Johnson BRI filters. The Johnson U has low efficiency and ignored here, and the two Gunn z are almost identical and their signal is combined.

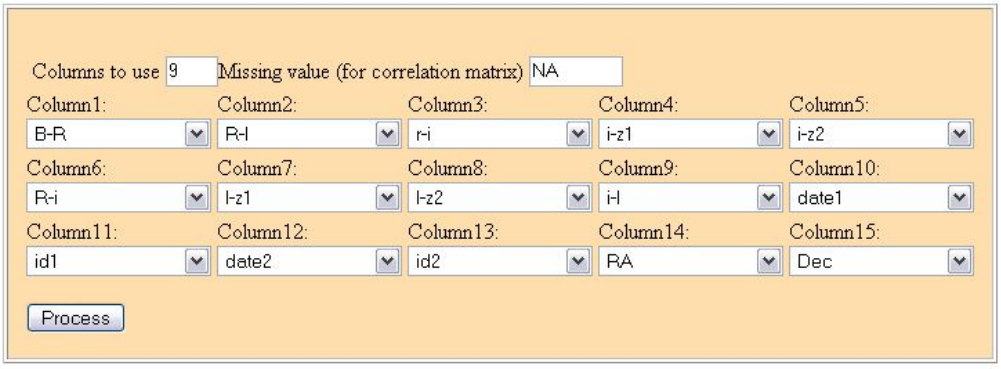

Figure 3: We choose the nine colors available from the 15 columns of the input file using a GUI feature.

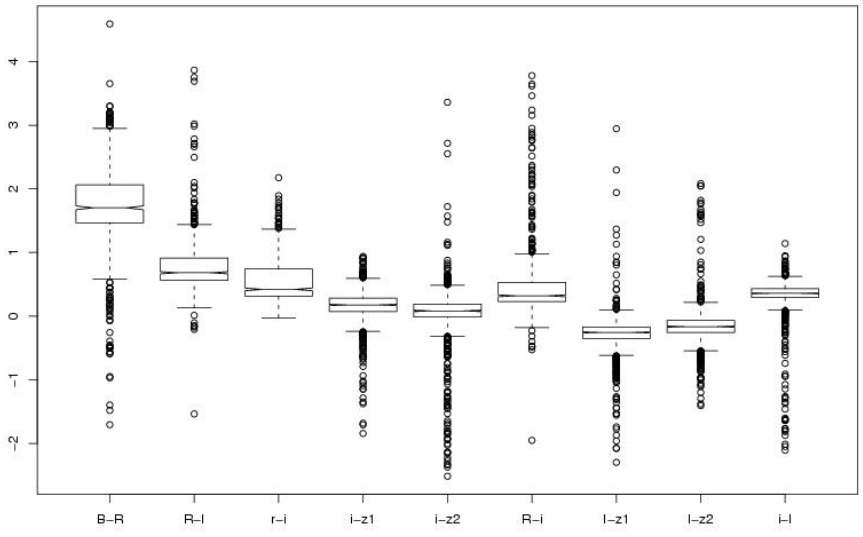

Figure 4: Boxplot with nine of the colors as input.

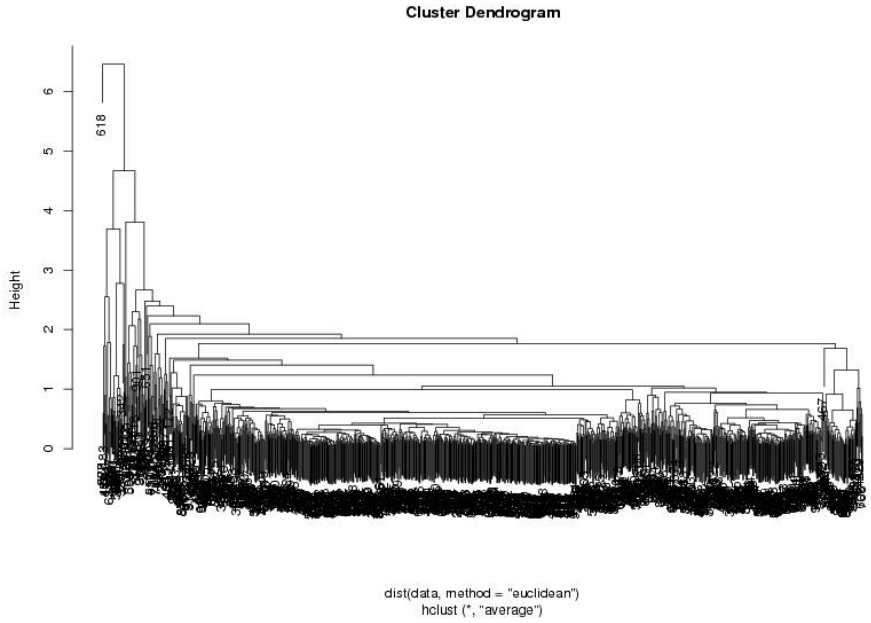

Figure 5: Dendrogram for the dataset using `hclust`. We decided to go with five clusters here.

```
> print(cl$size)
[1] 480  75  33 122 289
> print(cl$centers)
     X2.1553   X0.8833    X0.563     X0.2347     X0.1909   X0.5069    X.0.1855
1  1.5062098 0.5789365 0.3302465  0.12387542  0.05890729 0.2437696 -0.27625958
2  1.4074840 0.9671000 1.0442640 -0.29106000 -0.51805867 0.7100160 -0.77514267
3 -0.1626848 1.5104515 0.8905485 -0.08033333 -1.67197273 2.5554939 -0.62693030
4  2.6638713 1.1741434 0.9580811  0.40006393  0.39761967 0.7380221 -0.03850164
5  2.0228934 0.7910471 0.5455412  0.24857820  0.15998754 0.4003003 -0.23075917
     X.0.1417    X0.3764
1 -0.21129146  0.3351669
2 -0.54814400  0.2570840
3  0.96470909 -1.0450424
4 -0.03605738  0.4361213
5 -0.14216851  0.3907467
> print(cl$withinss)
[1]  95.40461 118.08204 142.91592 126.88979  87.19068
> print(cl$cluster)
  [1] 2 5 1 4 1 1 1 5 1 1 1 1 1 1 1 5 5 1 1 5 1 1 1 1 1 5 3 1 1 1 1 1 1 1 1 1 3
```

Figure 6: Centers and withinss of the five clusters using k-means.

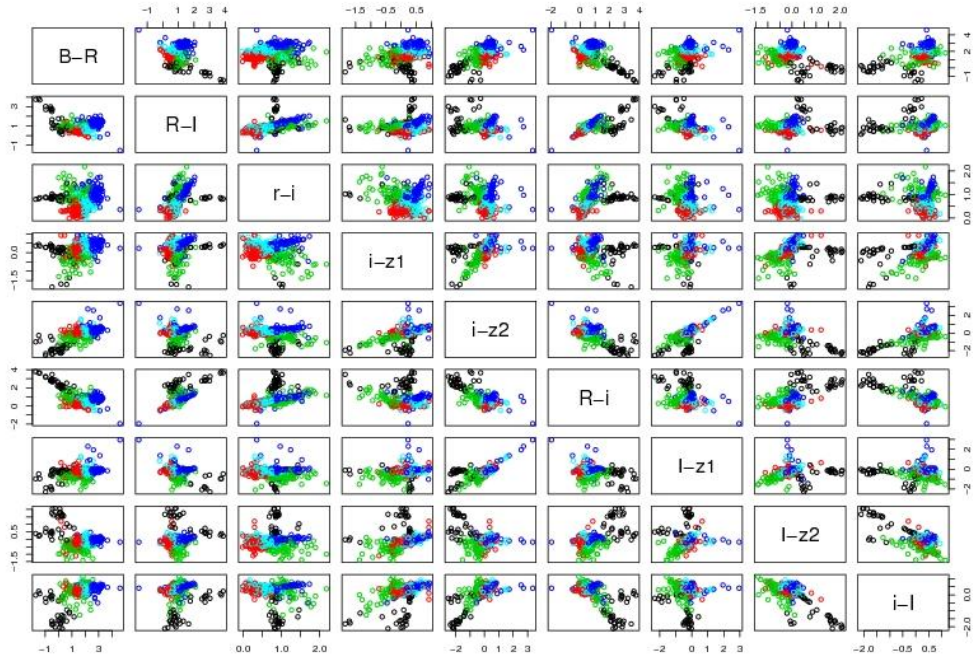

Figure 7: Pairwise plots for the nine parameters with members of the five clusters colored differently.

the probabilities dividing the set into 4 zones each having a quarter of the sorted points and then plot them as color-color plots using *BRI* (Figure 8) and r i z filters (Figure 9) for demonstration purposes. The scatter is clear even though we see here only two of the nine dimensions.

We then choose the outliers of interest and obtain image cutouts (using tools outside **VOStat**) for further science and follow-up in the traditional way. Example of a typical star (Figure 10) and an outlier (Figure 11) are shown here. Interesting objects can then be followed-up using various telescopes. The entire procedure can be easily automated for different datasets of a similar nature.

4. Future enhancements to VOStat

We are currently working on realizing the full potential of **VOStat**. Some planned modifications are cosmetic in nature and some others are fundamental changes that will make **VOStat** much more powerful. These include:

(a) The ability to fetch partial data (only certain rows and columns) from a data set. Statistical analysis can then be done on such large datasets available publicly,

(b) Implementing choices for data transformation (e.g., log, exp), prior to the analysis

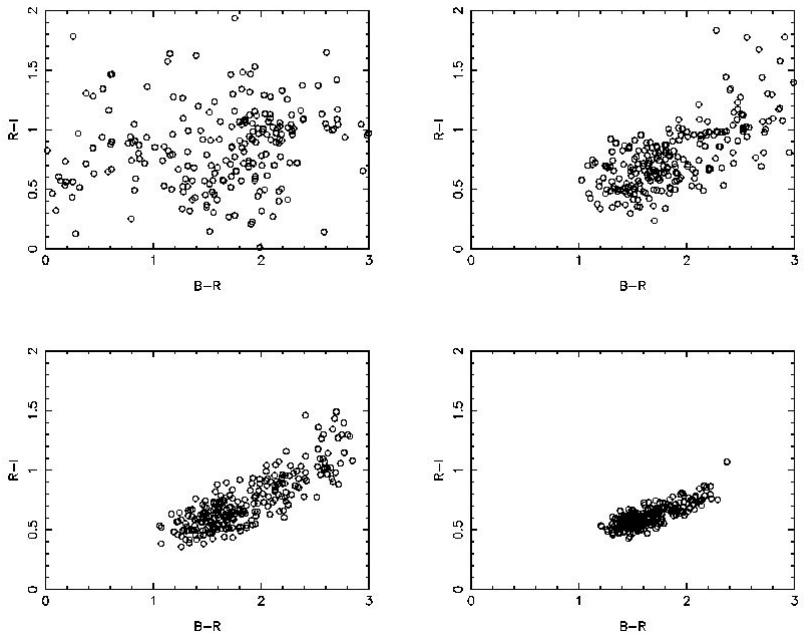

Figure 8: Color-color plot for the $B\ R\ I$ subset of magnitudes. In this and the next plot the x-y limits are chosen to highlight the scatter as well as the clustering in the different sub-figures.

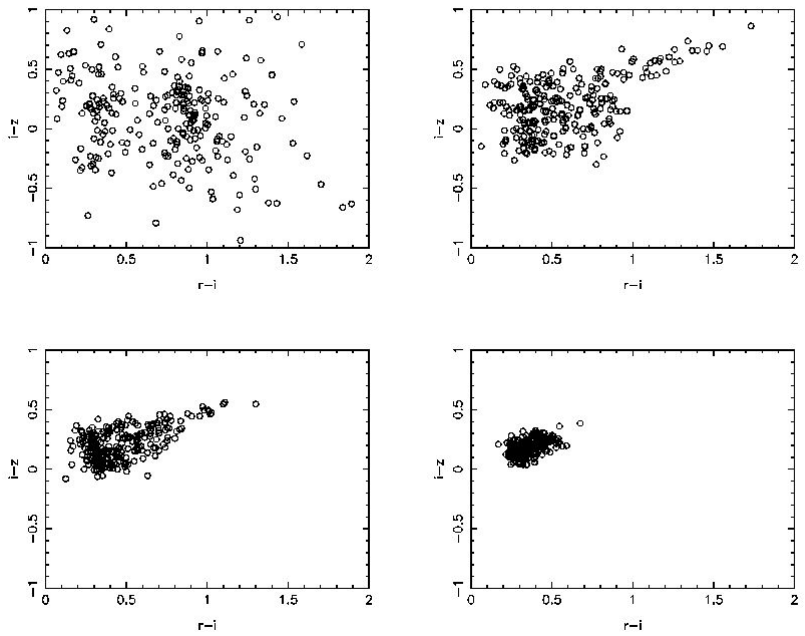

Figure 9: Color-color plot for the $r\ i\ z$ subset of magnitudes.

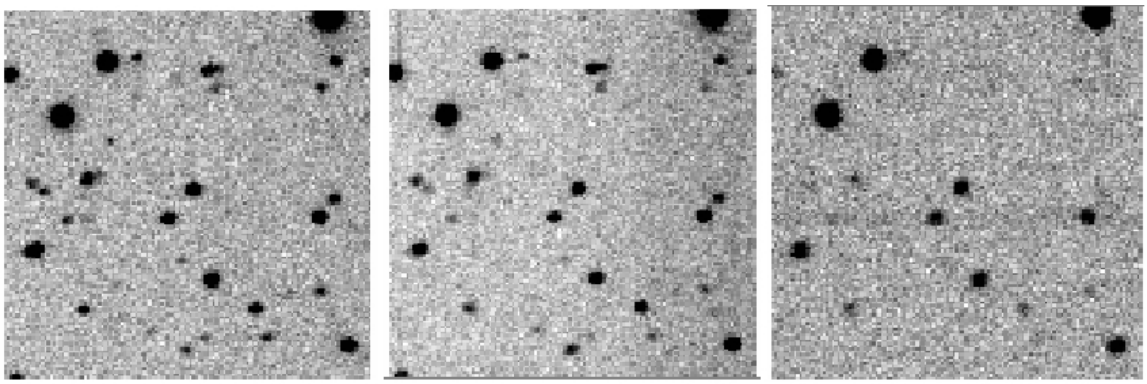

Figure 10: Example of a normal star in the sample. This object occupies one of the standard areas of the HR diagram and the fluxes in the different filters depicted indicate that it does not have extreme colors. This and the next cutout for the outlier are obtained with a tool not part of the **VOStat** suite. Further science for interesting objects could then be done using follow-up observations by telescopes.

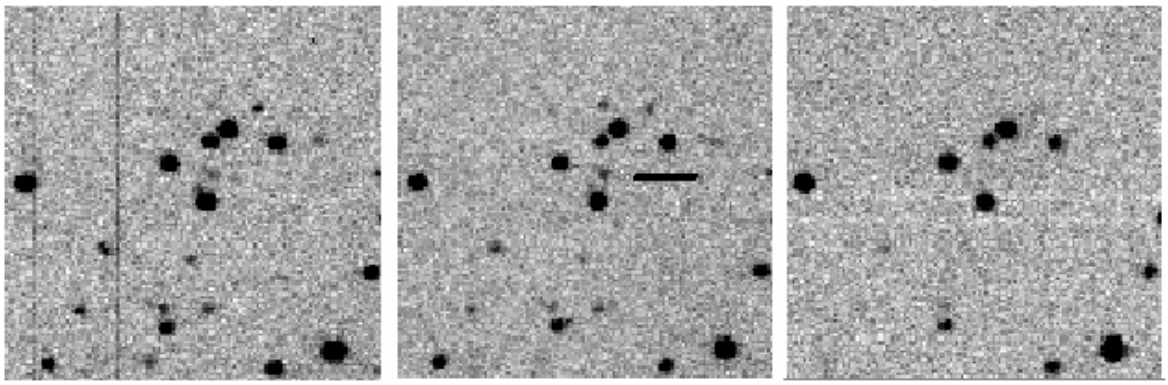

Figure 11: Example of an outlier (marked with a horizontal line). The fluxes indicate the extreme colors and this object occupies a non-standard area in the HR diagram and hence is atypical.

(c) Producing plots in alternative formats (e.g., jpeg, png, pdf),

(d) Allowing a series of operations rather than a single operation as is done now. One could then save scripts and automate complex operations on big sets.

References

Chakravarti I, Laha R, Roy J (1967). *Handbook of Methods of Applied Statistics*, volume 1. John Wiley and Sons.

Champeney D (1973). *Fourier Transforms and Their Physical Applications*. Academic Press, New York.

Hotelling H (1936). "Relations Between Two Sets of Variates." *Biometrika*, **28**, 321–377.

Laplace PS (1820). "Des méthodes analytiques du calcul des probabilités." In "Théorie analytique des probabilités," volume Livre 2. Courcier, Paris, 3 edition.

LaValley MP, Isobe T, Feigelson ED (1992). "Software Report: ASURV, Pennsylvania State University." *Bulletin of the American Astronomical Society*, **24**, 839–840.

Pearson K (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space." *Philosophical Magazine*, **2**(6), 559–572.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Royston P (1995). "A Remark on Algorithm AS 181: The W-Test for Normality." *Applied Statistics*, **44**, 547–551.

Shapiro SS, Wilks MB (1965). "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika*, **52**, 591–611.

Stigler SM (1977). "An Attack on Gauss, Published by Legendre in 1820." *Historia Mathematica*, **4**, 31–35.

Affiliation:

G. Jogesh Babu
Department of Statistics
319 Thomas Building
The Pennsylvania State University
University Park, PA 16802-2111, United States of America
E-mail: babu@stat.psu.edu

Ashish Mahabal
Astronomy Department
California Institute of Technology
MC 105-24
1200 East California Blvd
Pasadena, CA 91125, United States of America
E-mail: aam@astro.caltech.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 18, Issue 11 *Submitted:* 2006-10-16
January 2007 *Accepted:* 2007-01-10