Fully Automated Approaches to Analyze Large-Scale Astronomy Survey Data

A. Prša

Villanova University & University of Ljubljana

`andrej.prsa@villanova.edu`   (610) 519-4887

with

E. F. Guinan[1], E. J. Devinney[1], S. G. Engle[1], M. DeGeorge[1],
G. P. McCook[1], P. A. Maurone[1], J. Pepper[2], D. James[3],
D. H. Bradstreet[4], C. R. Alcock[5], J. Devor[5], R. Seaman[6]
T. Zwitter[7], K. Long[8], R. E. Wilson[9], I. Ribas[10], A. Gimenez[11]

April 1, 2009

Abstract

Observational astronomy has changed drastically in the last decade: manually driven target-by-target instruments have been replaced by fully automated robotic telescopes. Data acquisition methods have advanced to the point that terabytes of data are flowing in and being stored on a daily basis. At the same time, the vast majority of analysis tools in stellar astrophysics still rely on manual expert interaction. To bridge this gap, we foresee that the next decade will witness a fundamental shift in the approaches to data analysis: case-by-case methods will be replaced by fully automated pipelines that will process the data from their reduction stage, through analysis, to storage. While major effort has been invested in data reduction automation, automated data analysis has mostly been neglected despite the urgent need. Scientific data mining will face serious challenges to identify, understand and eliminate the sources of systematic errors that will arise from this automation. As a special case, we present an artificial intelligence (AI) driven pipeline that is prototyped in the domain of stellar astrophysics (eclipsing binaries in particular), current results and the challenges still ahead.

[1]Villanova University; [2]Vanderbilt University; [3]University of Hawaii; [4]Eastern University; [5]Harvard-Smithsonian CfA; [6]NOAO; [7]University of Ljubljana; [8]STSCI; [9]University of Florida; [10]University of Barcelona; [11]INTA/CSIC.

1

1 Introduction

One of the most important changes in observational astronomy of the 21$^\mathrm{st}$ Century is a rapid shift from classical object-by-object observations to extensive automated surveys. As CCD detectors improve in sensitivity and their costs decrease, more and more small and medium-size observatories are refocusing their attention[1] to the investigation of stellar variability through systematic wide-field sky-scanning missions. This trend is additionally powered by the success of pioneering surveys such as EROS (Palanque-Delabrouille et al. 1998), MACHO (Cook et al. 1995), OGLE (Udalski et al. 1997), ASAS (Pojmanski 2002), their space counterpart Hipparcos (Perryman & ESA 1997) and others. Such surveys produce massive amounts of data that pose a significant challenge to reduction and analysis. Surveys and missions currently commissioned (i.e. Kepler (Borucki et al. 2007), LSST (Tyson 2002), Pan-STARRS (Kaiser 2004) and Gaia (Perryman et al. 2001)) will produce petabytes of data daily; spectroscopic surveys such as RAVE (Steinmetz et al. 2006), SEGUE (Newberg & Sloan Digital Sky Survey Collaboration 2003) and Hermes (Raskin & Van Winckel 2008) will open the doors for complementary spectroscopy for millions of sources. Yet currently-available tools fall short of the needs of proper analysis.

In this white paper we limit ourselves to stellar astrophysics (in particular, eclipsing binary stars – EBs hereafter), but the points raised are readily applicable to other areas of astronomy, such as the study of pulsating variable stars, astroseismology, stellar rotation, population theory, etc. To date, about a thousand papers have been published on EBs with physical and geometrical parameters determined to better than 3% accuracy. For an eclipsing binary expert it takes 1-2 weeks to reduce and analyze a single eclipsing binary light-curve the old-fashioned way. There are currently about 10,000 photometric/RV data-sets that in principle allow modeling to a 3% accuracy. According to Hipparcos results, about 0.8% of the overall stellar population are EBs (917 out of 118,218 stars, Perryman & ESA 1997). Projecting these statistics to other large surveys allows estimating the number of EBs expected in survey databases: $\sim$136,000 in ASAS, $\sim$56,000 in the OGLE LMC field, $\sim$16,000 in OGLE SMC field, $\sim$80,000 in TASS (Droege et al. 2006), etc. Gaia will make a revolution in these numbers since the aimed census of the overall stellar population is $\sim$1 billion down to $V = +20\,\mathrm{mag}$ (Perryman et al. 2001), yielding millions of EBs and tens of millions of variable stars. Finally, with LSST essentially complete to $V = 24.5$, the yield of EBs will reach the tens of millions. Even if all observational facilities collapsed at that point so that no further data got collected, it would still take 500 years for all 12,500 members of the IAU to analyze these data! Given the unique capability of EBs to yield accurate stellar masses, radii, temperatures and distances, and realizing that many of these are accessible by small-size ground instruments, EBs should definitely hold one of the top positions on observational candidates list.

---

[1]A comprehensive list of more than a hundred such facilities may be found at http://www.astro.physik.uni-goettingen.de/~hessman/MONET/links.html.
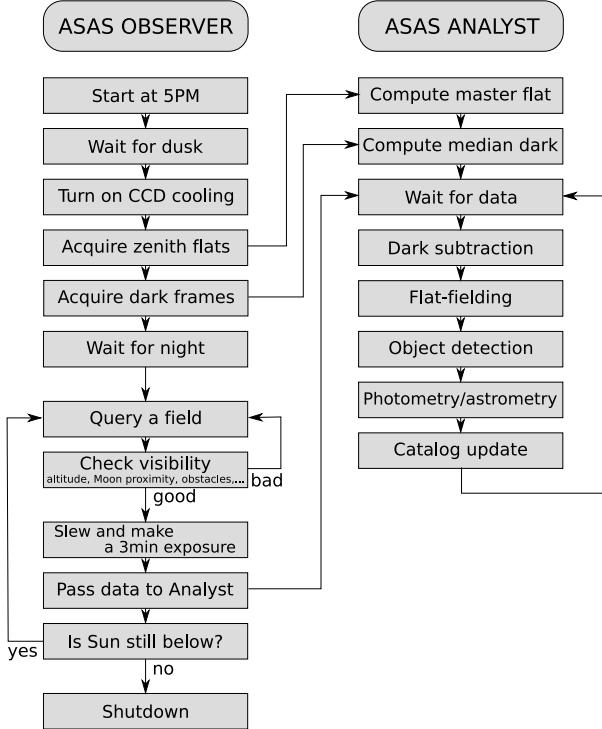
Figure 1: ASAS project's automated data acquisition pipeline (Pojmanski 1997).

2 Brief review of process automation

Data acquisition is the most automated aspect of the pipeline. An example of a fully automatic data acquisition and analysis pipeline is that of the All-Sky Automated Survey (ASAS, Pojmanski 1997), depicted in Figure 1. The level of sophistication is already such that it assures accurate and reliable data from both ground-based and space surveys.

Variability classification, however, has proved to be much more involved than perhaps initially expected. Fundamentally different objects (i.e. radial pulsators and ellipsoidal variables) produce essentially indistinguishable light curves and follow-up spectroscopy is paramount for identifying their true nature. A series of systematic analyses were conducted by Rucinski (1997b,a, 1998) and later Maceroni & Rucinski (1999) and Rucinski & Maceroni (2001) that highlighted the importance of the Fourier decomposition technique (FDT) for classification of variable stars. The technique itself – fitting a 4th order Fourier series to phased data curves and mapping different types of variables in Fourier coefficient space – was originally proposed for EBs by Rucinski (1973) and has been used ever since, most notably for classifying ASAS data (Pojmanski 2002; Paczyński et al. 2006). Alcock et al. (1997), analyzing 611 bright EBS from the MACHO database (Cook et al. 1995), proposed a new decimal classification scheme for categorizing EB types. Wyrzykowski et al. (2003, 2004) identified 2580 EBs in the LMC and 1351 EBs in the SMC. They employed a novel classification approach using Artificial neural networks (ANN) as an image recognition algorithm, based on phased data curves that have been converted to low-resolution images as
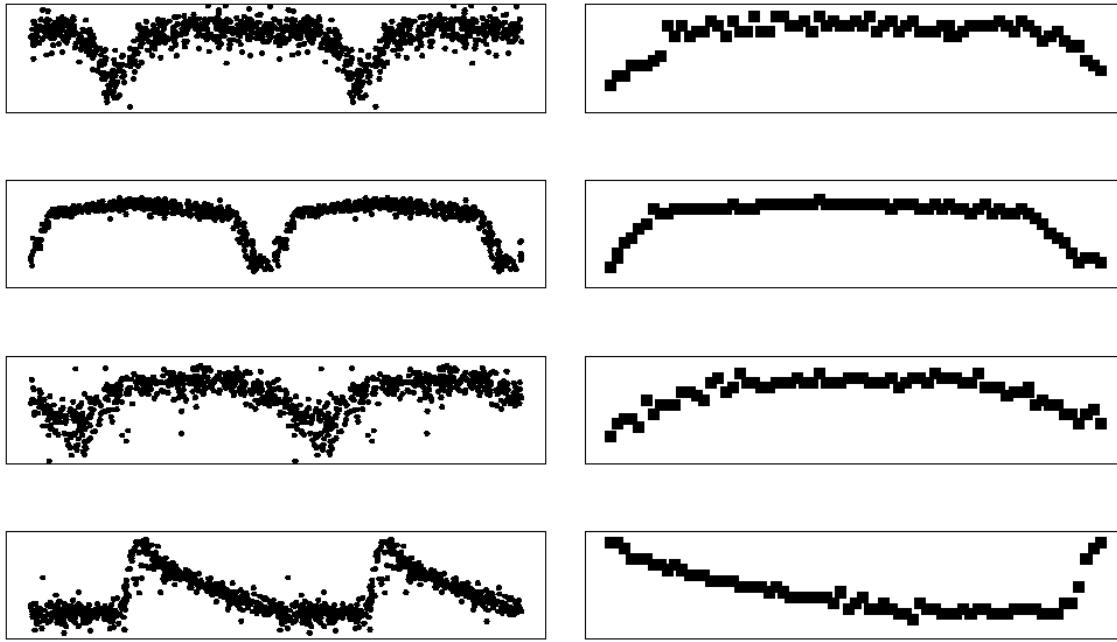
Figure 2: Conversion of phased light curves (left) to 70×15 pix images (right), which are fed to the neural network image recognition algorithm. Taken from Wyrzykowski et al. (2003).

depicted on Figure 2. Their classification pipeline was backed up by visual examinations of results.

Approaches to automating light curve solutions have taken various forms to date. Wyithe & Wilson (2001, 2002), in their work to establish the best distance indicators among detached and semi-detached binaries in the Small Magellanic Cloud, obtained starting parameters for the rigorous WD model by comparing each candidate light curve with a set of template model light curves, sending the best match to an automated version of the differential corrector program (DC). The latter could be computationally prohibitive to apply to the expected large future data-sets. Employing less rigorous physical models, of course, is one approach to computational efficiency. Thus, Tamuz et al. (2006) employ the EBOP ellipsoidal model (Popper & Etzel 1981). Using this engine, they arrive at an initial solution after a combination of grid search, gradient descent and geometrical analysis of the LC. Devor (2005) illustrates an automated pipeline employing a simple model of spherical stars without tidal or reflection physics, whose starting values are similarly obtained. Prša et al. (2008) have devised a neural network based engine EBAI (Eclipsing Binaries via Artificial Intelligence; http://www.eclipsingbinaries.org) that is capable of processing thousands of light curves in just a few seconds; it yields principal parameters of the analyzed variable.

3   Artificial Intelligence

Advances in Artificial Intelligence (AI) and the continued operation of Moore's law that predicts exponential growth of the processing power have created the opportunity for significant progress in solving the types of problems that are limited by the lack of human capital. A new approach, the *Intelligent Data Pipeline* (IDP), has been prototyped in the domain of EBs which uses AI techniques to operate autonomously on large observational data-sets to produce results of astrophysical value. The IDP is designed to handle the complete process of variable discovery, classification of variability and management of the solution process for the discovered EBs (Devinney et al. 2005, 2006; cf. Fig. 3). The IDP employs ANNs in the processing modules, while the supervisory knowledge, now implicit in humans, is encoded in control modules as rules appropriate for each processing module. The supervisory modules have the task of keeping the process on track and providing physically meaningful results through each phase of the processing pipeline.

ANNs are very simple algorithms that involve little beyond summation and multiplication, while having the capability of being *trained* on a physical content. While some may find the opaqueness of Artificial Neural Networks (ANN) problematic, their success in many areas, including classification, real-time robotic control and others is a powerful answer.
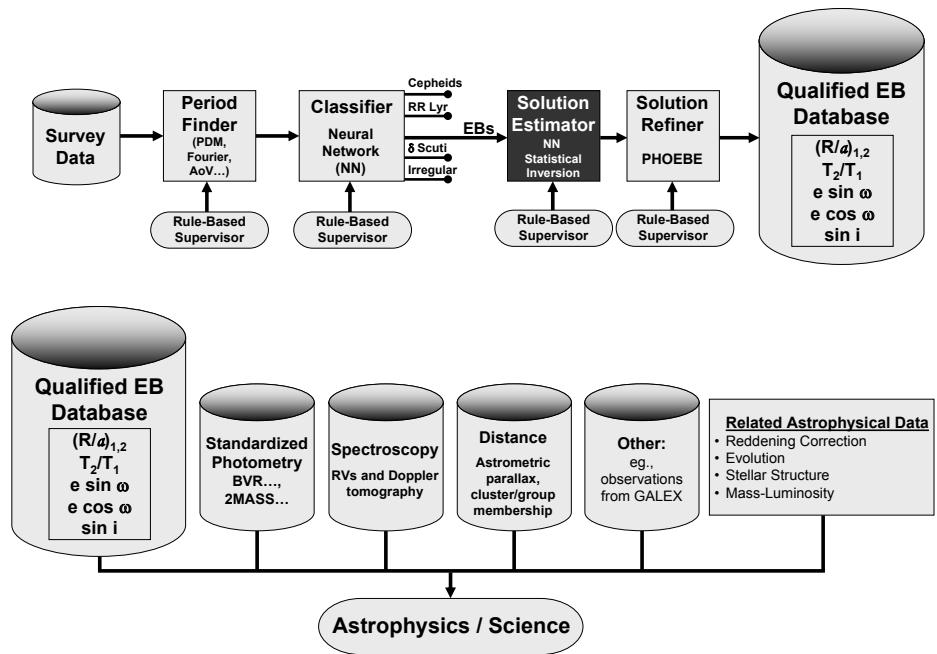

Figure 3: Intelligent Data Pipeline (IDP). Complete survey data is piped through a period finder algorithm that is controlled by a rule-based system. All variable sources are then passed to the ANN-based classifier. Light curves consistent with EB signatures are passed to the Solution Estimator block.

In their basic form, ANNs are systems of multiple layers (Fig. 4). Each layer consists of a given number of independent units. Each unit holds a single value. These values are propagated from each unit on the current layer to all units on the subsequent layer by weighted connections. Propagation is a simple linear combination $y_i = \sum_j w_{ij} x_j$, where x_j are the values on the current layer, w_{ij} are weighted connections, and y_i are the values that enter the subsequent layer. Before they are stored in their respective units, y_i are first passed through a (non-linear) *activation function* A_f. This function, typically a sigmoid function – $A_f(y_i) = 1/[1 + \exp(-(y_j - \mu)/\tau)]$ – introduces non-linear properties to the network. Coefficients μ and τ are selected so that $A_f(y_i)$ fall in the $(-1, 1)$ interval. It is this value that is stored in the i-th unit on the subsequent layer. Layers in the three-layer network are usually denoted *input*, *hidden*, and *output* layer. ANN is thus a non-linear *mapping* from the input layer to the output layer. In our implementation in the domain of EBs, the ANN maps the input light curves to the output set of principal physical parameters.

Training the network implies determining the weights w_{ij} on weighted connections. The
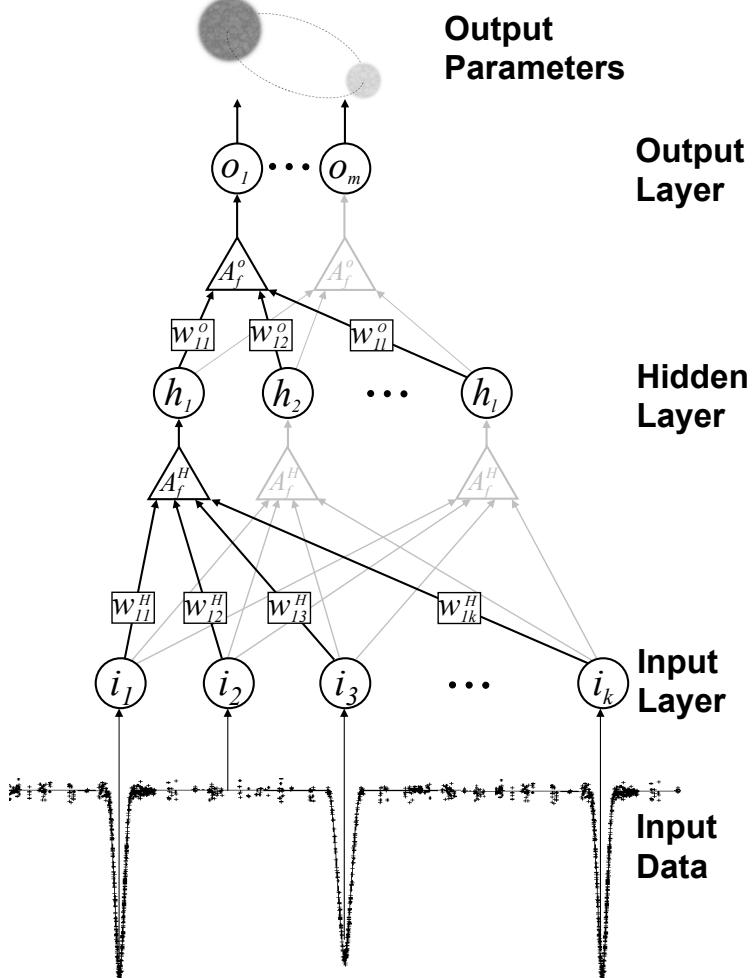

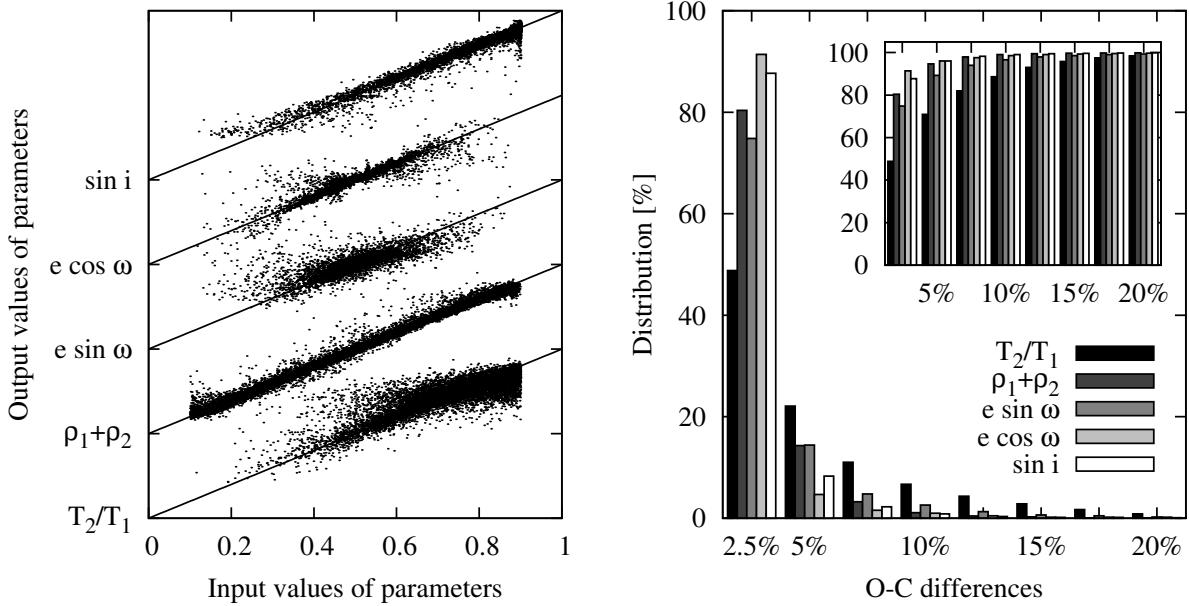Figure 4: Schematic view of a three-layered, fully connected ANN.

Figure 5: ANN performance on a sample of 10,000 LCs. Left: comparison between the input parameters (known from generating a sample) and output parameters provided by the network. $T_{1,2}$ are effective temperatures of EB components, $\rho_{1,2}$ are their fractional radii, e is eccentricity, ω is argument of periastron, and i denotes inclination. Parameters are offset by 0.5 for clarity and a guideline is provided for easier comparison. Right: distribution of residuals (main graph) and their cumulative distribution (inset). The bars depict the fraction of EBs with errors between 0% and 2.5% (first bin), 2.5% and 5% (second bin), etc. 90% of all LCs have errors smaller than 10% in *all* parameters.

back-propagation algorithm relies on a sample of LCs (the training set) with known physical parameters; these are called *exemplars*. All LCs are propagated through the network and their outputs are compared to the known values. The weights are then modified so that the discrepancy between the two sets is minimized. This is an iterative process that needs to be done only once. After training, the network is ready to process any input LC extremely quickly.

To evaluate ANN performance, we created a set of 10,000 synthetic light curves for eclipsing binary stars and passed it through a trained ANN. To each LC we added variable amounts of white noise, simulating different S/N ratios. Fig. 5 depicts the results that show clear statistical viability: 90% of the sample resulted in parameters with errors less than 10%. The success rate of recognition is comparable to that of the learning sample, and the underlying distribution of errors for both data-sets is indistinguishable. This demonstrates the capability of the ANN to successfully recognize data it has never seen before.

4 Discussion

The importance of results that will be achieved by developing novel fully automated approaches can hardly be overstated. In the domain of EBs, their analysis yields:

- calibration-free physical properties of stars (i.e. masses, radii, surface temperatures, luminosities);

- accurate stellar distances;

- precise stellar ages;

- stringent tests of stellar evolution models.

The products of state-of-the-art EB modeling are seminal to many areas of astrophysics:

- calibrating the cosmic distance scale;

- mapping of clusters and other stellar populations (e.g. star-forming regions, streams, tidal tails, etc) in the Milky Way;

- determining initial mass functions and studying stellar population theory;

- understanding stellar energy transfer mechanisms (including activity) as a function of temperature, metallicity and evolutionary stage;

- calibrating stellar color-temperature transformations, mass-radius-luminosity relationships, and other relations basic to a broad array of stellar astrophysics;

- studying stellar dynamics, tidal interactions, mass transfer, accretion, chromospheric activity, etc.

In addition, spectroscopic surveys such as RAVE, SEGUE and Hermes will provide observations of thousands of spectroscopic binaries that will allow the determination of metallicity, leading to chemical tagging, galactic stratigraphy and population memberships.

The enormous inflow of data that marked the previous decade exposed the deficiencies of the analysis tools in this decade. Manual analysis will have to be limited to the astrophysically most interesting cases; all other sources will need to be processed in a fully automated fashion. In this white paper we presented one possible approach to automation – artificial neural networks – that has a unique capability of processing hundreds of thousands of LCs in a matter of minutes. Suitable training data-sets will have to be created that would allow for a wide range of light curve types to be automatically processed. The community will need to invest significant effort to further develop automation methods and update the current tools to cope with this challenge. In addition, greater attention needs to be paid to intelligent components, such as expert systems, to insure appropriate flow down the data pipeline.

5 Recommendations

Our recommendations to the Decadal Survey 2010 regarding actions that need to be taken in order to address the challenges pointed out above are:

- form a dedicated center (such as MAST or HEASARC) that would address the issues of data analysis automation; such a center would employ 3-5 FTE in software engineering and theoretical scientific modeling;

- form a narrowly focused IAU commission that would steer community efforts – i.e. the National Virtual Observatory (NVO) interface, the choice of a programming language (i.e. python), specifications for application deployment, a well-defined suite of test cases, etc;

- organize regular workshops and splinter meetings at the AAS and IAU symposia to address the application of Artificial Intelligence and other fully automated methods in astronomy data mining;

- secure adequate funding through NSF/NASA for technology research and implementation through specialized calls for proposals.

References

Alcock, C., Allsman, R. A., Alves, D., et al. 1997, AJ, 114, 326

Borucki, W. J., Koch, D. G., Lissauer, J., et al. 2007, in Astronomical Society of the Pacific Conference Series, Vol. 366, Transiting Extrapolar Planets Workshop, ed. C. Afonso, D. Weldrake, & T. Henning, 309–+

Cook, K. H., Alcock, C., Allsman, H. A., et al. 1995, in ASP Conf. Ser. 83: IAU Colloq. 155: Astrophysical Applications of Stellar Pulsation, ed. R. S. Stobie & P. A. Whitelock, 221–+

Devinney, E., Guinan, E., Bradstreet, D., et al. 2005, in Bulletin of the American Astronomical Society, 1212–+

Devinney, E. J., Guinan, E., DeGeorge, M., Bradstreet, D. H., & Giammarco, J. M. 2006, in Bulletin of the American Astronomical Society, Vol. 38, Bulletin of the American Astronomical Society, 119–+

Devor, J. 2005, ApJ, 628, 411

Droege, T. F., Richmond, M. W., Sallman, M. P., & Creager, R. P. 2006, ArXiv Astrophysics e-prints

Kaiser, N. 2004, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. J. M. Oschmann, Jr., Vol. 5489, 11–22

Maceroni, C. & Rucinski, S. M. 1999, AJ, 118, 1819

Newberg, H. J. & Sloan Digital Sky Survey Collaboration. 2003, in Bulletin of the American Astronomical Society, Vol. 35, Bulletin of the American Astronomical Society, 1385–+

Paczyński, B., Szczygieł, D. M., Pilecki, B., & Pojmański, G. 2006, MNRAS, 368, 1311

Palanque-Delabrouille, N., Afonso, C., Albert, J. N., et al. 1998, A&A, 332, 1

Perryman, M. A. C., de Boer, K. S., Gilmore, G., et al. 2001, A&A, 369, 339

Perryman, M. A. C. & ESA. 1997, The HIPPARCOS and TYCHO catalogues. (Noordwijk, Netherlands: ESA Publications Division, 1997, Series: ESA SP Series vol no: 1200, ISBN: 9290923997 (set))

Pojmanski, G. 1997, Acta Astronomica, 47, 467

Pojmanski, G. 2002, Acta Astronomica, 52, 397

Popper, D. M. & Etzel, P. B. 1981, AJ, 86, 102

Prša, A., Guinan, E. F., Devinney, E. J., et al. 2008, ApJ, 687, 542

Raskin, G. & Van Winckel, H. 2008, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7014

Rucinski, S. M. 1973, Acta Astronomica, 23, 79

Rucinski, S. M. 1997a, AJ, 113, 1112

Rucinski, S. M. 1997b, AJ, 113, 407

Rucinski, S. M. 1998, AJ, 115, 1135

Rucinski, S. M. & Maceroni, C. 2001, AJ, 121, 254

Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, AJ, 132, 1645

Tamuz, O., Mazeh, T., & North, P. 2006, MNRAS, 367, 1521

Tyson, J. A. 2002, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. J. A. Tyson & S. Wolff, Vol. 4836, 10–20

Udalski, A., Kubiak, M., & Szymanski, M. 1997, Acta Astronomica, 47, 319

Wyithe, J. S. B. & Wilson, R. E. 2001, ApJ, 559, 260

Wyithe, J. S. B. & Wilson, R. E. 2002, ApJ, 571, 293

Wyrzykowski, L., Udalski, A., Kubiak, M., et al. 2003, Acta Astronomica, 53, 1

Wyrzykowski, L., Udalski, A., Kubiak, M., et al. 2004, Acta Astronomica, 54, 1