A statistical analysis of BRITE data

M. Kondrak[1], K. Zwintz[1] and R. Kuschnig[2]

1. Institut für Astro- und Teilchenphysik, Universität Innsbruck
Technikerstr. 25/8, 6020 Innsbruck, Austria
2. Graz University of Technology, Institute of Communication Networks and Satellite Communications, Infeldgasse 12, 8010 Graz, Austria

Reducing raw data received by satellites or telescopes is an important and challenging step in order to deduce scientific results. Measurements do not only depend on properties of the scientific target, but also on intrinsic instrumental parameters. To remove the undesired instrumental influence on the measured signal, one needs to determine the correlation between both. This is a non-trivial process, since instrumental parameters may correlate with each other and can be time-dependent.

Removing outliers may look simple, but an overcorrection can decrease the quality of the result and might eliminate significant scientific signals. If using a time-series analysis, which is a proper method to obtain information about periodic stellar variability, it is important not to cut off data which may contribute to a periodic pattern, although they might look like outliers.

Within this work a software tool has been developed to assist the user with the reduction of BRIght Target Explorer (BRITE) Constellation data. A uniform reduction tool speeds up and simplifies the data analysis and allows to prioritize the scientific data analysis over the data handling.

1 Introduction

Automation of data processing has many advantages. Routines can be adapted to similar input data by adjusting their respective configuration. Modifications and user-input can be written automatically to logfiles, allowing bug tracking and reproduction of results. This method is faster and more user-friendly than documenting every step by hand. Results can be confirmed by others using the same routines together with the same input documented in the logfile. Using standardized routines within a working group allows easy and fast comparison of results. Sharing these routines with others may improve the overall result by implementing new ideas. Also people, who are new to the specific scientific field of research, can be given an introduction using a tutorial. Visualization of the output can help inexperienced users to understand the effects of varying different input parameters. Therefore, basics of data reduction can be learned faster and more efficiently. Besides these advantages, there are also disadvantages writing such routines.

A lot of resources have to be spent before any scientific outcome is produced. The development of a suitable reduction routine requires a lot of research. Various ideas of data reduction have to be considered. The routines should be designed for easy installation, be user-friendly and well documented so that further development is possible. During this design process no scientific output is produced. Furthermore, a lot of testing using various datasets is required in order to eliminate bugs occurring within specific inputs or generated by unpredictable user-input. Testing on different

Satellite name	Short	Country	Launch date	Filter	Period [min]
BRITE-Austria	BAb	Austria	25.02.2013	blue	100.36
UniBRITE	UBr	Austria	25.02.2013	red	100.37
Lem	BLb	Poland	21.11.2013	blue	99.57
BRITE-Toronto	BTr	Canada	19.06.2014	red	98.24
Heweliusz	BHr	Poland	19.08.2014	red	97.10

Table 1: List of all BRITE Constellation satellites.

operating systems and with other versions of libraries should be done to ensure the stability and robustness of the routines. Feedback from testers should help to eliminate questions, increase the user-friendliness and include required features. Another big disadvantage is the unpredictability of the input data. Since every measurement and dataset is unique and no further information on the stellar behaviour may be available, the routines might not handle the data correctly. Also datasets used for testing might not include such unpredicted behaviour. In general, data reduction routines only provide a viable output if the user is experienced with the theory behind the data reduction itself and the usage of such routines.

A data reduction routine is needed to process a lot of input data requiring similar treatment. Therefore, it was decided to design a software tool which assists the user in the data reduction of BRIght Target Explorer (BRITE) Constellation data. Designing such a routine requires statistical analysis of the input data, theoretical backgrounds of data reduction and a schematic pattern of the necessary reduction steps. The data reduction process in this work includes removal of outliers, decorrelation between measurement and instrumental parameters, orbital clipping and visualization of these reduction steps.

2 Data acquisition and format

BRITE Constellation consists of five nanosatellites. These 7 kg cube-shaped satellites orbit the Earth at an altitude of about 800 km in a 100-minute sun-synchronous polar orbit, taking typically images every 20 seconds for about 15–35 minutes per orbit. A five-lens telescope with a 3 cm aperture together with a 4008 × 2672 pixel CCD detector and either a blue (390–460 nm) or red (550–700 nm) filter allows them to do precise photometric time-series measurements of bright stars (down to $V = 4$ mag at 1 mmag precision of each BRITE orbit) within their 24° field-of-view. Different filters help to identify different modes with very similar frequencies in non-radially pulsating stars (Weiss et al., 2014). A list of the BRITE satellites is given in Table 1.

The CCD images taken by the BRITE satellites are converted to FITS-files by measuring the Point-Spread-Function (PSF), pre-processed and then delivered as ASCII-files (Popowicz, 2016). Each file contains a header with parameters of the observational setup used followed by columns that contain the observational data. Currently four different versions of data-formats exist. The first and second versions contain data taken in normal mode. In later versions the so-called chopping mode (Pablo et al., 2016) was introduced to increase image quality by removing hot pixels via nodding the satellite. The two versions in normal or chopping mode differ by just an additional column with a quality value added.

3 Statistical data analysis

The measured photon flux signal ($f(t)$) of a star at time t does not only depend on the stellar luminosity ($l(t)$) which may be related to stellar variability, but also on external perturbations like stray light ($s(t)$), intrinsic instrumental parameters like the temperature of the CCD ($T(t)$) or the x- and y-center position of the measured PSF ($x(t), y(t)$) and other parameters ($z(t)$).

$$f(t) = F(l(t), s(t), T(t), x(t), y(t), z(t)) \tag{1}$$

The challenge of the data reduction is to retrieve the a priori unknown correlation function F to deduce the time-dependency of the stellar luminosity out of the measurements. In this simple case the instrumental parameters only depend on time. In reality, they can also depend on each other and on the external perturbations, so Eq. (1) extends to:

$$f(t) = F(l(t), s(t, \phi, \theta, z), T(t, \phi, \theta, x, y, z), x(t, \phi, \theta, z), y(t, \phi, \theta, z), z(t, \phi, \theta)) \tag{2}$$

E.g.: the influence of the Moon's stray light depends on the position of the Moon, the position (ϕ and θ), alignment and stability of the satellite. The stability of the satellite may depend further on Earth's magnetic field, the performance of the star tracker and so on. Determining and treating such multidimensional correlations requires a lot of resources. A Bayesian approach using untreated raw data would be necessary, which is beyond the scope of this work.

3.1 Theoretical background and basic assumptions

In order to perform basic decorrelation a few assumptions are made. The instrumental parameters do not depend on each other, leading to Eq. (1). The variations of the instrumental parameters are small, allowing them to be described by a low order polynomial.

A polynomial of second order was tested for decorrelation between flux and temperature using datapoints within one orbit. The correlation between flux and temperature showed a minimum at about 24°C, leading to an increased flux above and below that temperature. The initial assumptions do not allow such behaviour, therefore a linear fit is further used for decorrelation. This might lead to an inaccurate description of the correlation process.

In order to describe the probability of a linear relation between the measured signal and a single instrumental parameter, the Pearson product-moment correlation coefficient or Pearson's r is used:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

r ranges from -1 (negative), 0 (no) to $+1$ (positive linear correlation) for two datasets $\{x_1, ..., x_n\}$ and $\{y_1, ..., y_n\}$ containing n values. r^2 is another useful quantity to describe correlation. $0 < r^2 < 1$ is the ratio of the explained variation to the total variation and denotes the strength of a linear relation between x and y. E.g. $r = 0.922$ or $r^2 = 0.850$ denotes that 85% of the total variation in y can be explained by the linear relationship between x and y. The other 15% of the total variation in y remains unexplained.

3.2 Data analysis of the Cygnus field

A statistical analysis of about 230 datasets taken by four of the five BRITE satellites was performed using data in chopping mode. Reduced data were used to test against the correlation of instrumental parameters. Only a linear relationship between the measured signal and one instrumental parameter was considered. Multidimensional correlations were neglected.

To constrain the linear correlation, a best linear fit was performed between the stellar magnitude and the corresponding instrumental parameter. The slope of this linear fit then accounts for the strength of a given correlation. To account for the variable quality of the unprocessed raw data, different weights are used. To determine the probability of a linear correlation the Pearson product-moment correlation coefficient was calculated using an implemented command within the *numpy* library of Python. To account for greater significance of larger datasets, a weight using the number of datapoints was also used. The analysis was also performed with the product of those two weights.

The results are listed in Table 2. Values for the y-center correlation are not given, but they are similar to the values of the x-center. It can be seen that the temperature correlation is stronger than the x-center correlation. Also the chopping mode at UBr and BTr leads to very small correlations. UBr needs further investigations, because of its opposite signed temperature correlation.

4 BRITE-data-reduction-tool

The first step in the design process of a new software tool is the selection of a proper programming or scripting language. Since programming languages like C or C++ are preferable for large project handling complex algorithms and data structures, where execution speed is critical, scripts are preferable (Ousterhout, 1998) and used in this work to achieve the task of data reduction.

Python scripts using the *matplotlib* library perform the visualisation and handle graphical user-input and selections. A shell script guides the user through the reduction process, coordinates and handles the communication between the user and the Python scripts, formats the data, performs calculations and writes a logfile. The reduction process was inspired by Pigulski (2015).

The reduction tool is build modularly, allowing to rearrange existing subroutines, implementing them within other scripts, using them for bug tracking, and extending the existing reduction procedure with new modules. Each module, except for the decorrelation, writes the output to the standard output of the terminal, preventing files from being overwritten and thus ideal for independent usage. Only the decorrelation module writes its result to the standard error output in the terminal. The main disadvantage is that bug and error tracking is more complicated. The output of each script, including the main shell script, is given in comma separated values (CSV) format. The modularity allows treating data of various versions and input format, supposing proper modules are available and the main shell script is modified accordingly.

After an overview of the lightcurve, the user can set the size of an aperture and exclude datapoints whose x- or y-center coordinates lie outside a given range. Datapoints above or below a certain threshold, i.e. outliers, can be removed automatically in the next step. Furthermore, the data can be decorrelated using a linear approach,

Sat	*now.*	*corr.w.*	*linew.*	corr. × line w.	# datasets
Temperature correlation					
BLb	−25.458	−58.067		−64.904	25
BAb	−1.711	−3.890		−3.485	72
UBr	9.354	19.975		13.533	36
BTr	−0.393	1.165		−2.539	96
All	−2.003	−4.362		−2.454	230
X-center correlation					
BLb	−1.084	−2.467	−0.953	−2.370	25
BAb	−0.924	−0.852	−0.482	−0.925	72
UBr	0.089	−0.811	−0.118	−0.531	36
BTr	0.097	−0.467	−0.073	−0.821	96
All	−0.353	−0.575	−0.114	−0.818	230
X1-center correlation					
BLb	−9.600	−26.251	−11.260	−35.289	25
BAb	15.604	31.647	7.790	24.732	72
UBr	−7.325	−28.985	−7.503	−20.160	36
BTr	2.315	0.614	2.155	0.766	96
All	3.661	4.221	0.211	−6.208	230
X2-center correlation					
BLb	11.636	37.271	11.996	39.959	25
BAb	−2.649	−8.037	−1.895	−8.404	72
UBr	2.914	8.864	3.093	8.354	36
BTr	−3.451	−16.262	−3.433	−14.542	96
All	−0.549	−1.064	−1.787	−9.317	230

Table 2: Results of the linear correlation between measured data and instrumental parameters. Columns are satellite (Sat), no weight (no w.), weight determined from the correlation (corr. w.), weight from the number of data points (line w.), weight from correlation multiplied by line weight (corr. × line w.) and number of data sets (# datasets). The slope of the linear correlation between stellar magnitude and various instrumental parameters is multiplied by a factor of 1000.

since the methodology described in Section 3 provided good results. Afterwards, sigma clipping can remove points with a large scattering within an orbit. As a last step, an experienced user can manually remove datapoints, which originally were not detected as outliers by the reduction tool. In addition, the tool automatically rejects a whole orbit if more datapoints than a configurable threshold are removed.

To get information about the stellar variability, a time-series analysis has to be performed. No tool for performing a time-series analysis is provided within this work. The best approach for getting frequency information of the star has to be selected and performed by the user. The scripts developed within this work are licensed under the GNU General Public Licence (GPL). They are tested and functional under Linux. OS X support is planed if resources are available. The scripts are publicly available at: 'https://github.com/Ashoka42/BRITE-data-reduction-tool'

5 Conclusion

The data reduction process, especially the decorrelation is a non-trivial process. In order to get better estimates of the correlation between the observational data and instrumental properties, the unreduced original FITS-files should be used for an analysis to extract further information. In addition, a statistical analysis should be done on these data to point out the influence of ageing effects by cosmic ray hits, increased stabilization of the satellite pointing by software improvements, increased image quality due to the introduced chopping mode and other long term trends. Together with huge volumes of data, a Bayesian approach seems to be necessary and more suitable than linearisation to determine the intrinsic behaviour of the correlation.

Using a software tool not only provides a unification of the data reduction process and thus reproducibility throughout various working groups and comparability of results using different reduction parameters. They also are a rather easy way to introduce other people to the field or test different reduction approaches. Routines speed up the reduction process, giving scientists more time interpreting the results than handling the data.

Acknowledgements. Based on data collected by the BRITE-Constellation satellite mission, designed, built, launched, operated and supported by the Austrian Research Promotion Agency (FFG), the Austrian Aeronautics and Space Agency and the University of Vienna, the Technical University of Graz, the Canadian Space Agency (CSA) and the Foundation for Polish Science & Technology (FNiTP MNiSW) and National Centre for Science (NCN). MK acknowledges support from the project "Asteroseismology with BRITE-Constellation" (Nachwuchsförderung der Universität Innsbruck 2015, PI: K. Zwintz). MK also acknowledges support by the Austrian Stience Fund FWF within the DK-ALM (W1259-N27). KZ is supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (FWF, project V431-NBL). MK also thanks Bram Buysschaert for sharing his ideas and own scripts of another data reduction. MK thanks Ela Zoclonska for discussions and access to the original FITS-files. MK thanks Stefan Gössl for testing and feedback of the script.

References

Ousterhout, J. K., *Scripting: Higher-Level Programming for the 21st Century*, *Computer* **31**, 3, 23 (1998), URL `http://dx.doi.org/10.1109/2.660187`

Pablo, H., et al., *The BRITE Constellation Nanosatellite Mission: Testing, Commissioning, and Operations*, PASP **128**, 12, 125001 (2016), `1608.00282`

Pigulski, A., *Analysis of BRITE data - a cookbook*, in BRITE Photometry Wiki (2015)

Popowicz, A., *Image processing in the BRITE nano-satellite mission*, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Proc. SPIE, volume 9904, 99041R (2016)

Weiss, W. W., et al., *BRITE-Constellation: Nanosatellites for Precision Photometry of Bright Stars*, PASP **126**, 573 (2014), `1406.3778`