The SARS algorithm: detrending CoRoT light curves with Sysrem using simultaneous external parameters

A. Ofir, R. Alonso, A.S. Bonomo, L. Carone, S. Carpano, B. Samuel, J. Weingrill, S. Aigrain, M. Auvergne, A. Baglin, et al.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY LETTERS

Mon. Not. R. Astron. Soc. **404,** L99–L103 (2010) doi:10.1111/j.1745-3933.2010.00843.x

The SARS algorithm: detrending *CoRoT* light curves with Sysrem using simultaneous external parameters

Aviv Ofir,[1⋆] Roi Alonso,[2,3] Aldo Stefano Bonomo,[2] Ludmila Carone,[4] Stefania Carpano,[5] Benjamin Samuel,[6] Jörg Weingrill,[7] Suzanne Aigrain,[8,9] Michel Auvergne,[10] Annie Baglin,[10] Pierre Barge,[2] Pascal Borde,[6] Francois Bouchy,[11,12] Hans J. Deeg,[13,14] Magali Deleuil,[2] Rudolf Dvorak,[15] Anders Erikson,[16] Sylvio Ferraz Mello,[17] Malcolm Fridlund,[18] Michel Gillon,[3,19] Tristan Guillot,[20] Artie Hatzes,[21] Laurent Jorda,[2] Helmut Lammer,[7] Alain Leger,[6] Antoine Llebaria,[2] Claire Moutou,[2] Marc Ollivier,[6] Martin Päetzold,[4] Didier Queloz,[3] Heike Rauer,[16,22] Daniel Rouan,[2] Jean Schneider[23] and Guenther Wuchterl[21]

[1]*School of Physics and Astronomy, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel*
[2]*LAM, UMR 6110, CNRS/Univ. de Provence, 38 rue F. Joliot-Curie, 13388 Marseille, France*
[3]*Observatoire de Genève, Université de Genève, 51 chemin des Maillettes, 1290 Sauverny, Switzerland*
[4]*Rheinisches Institut für Umweltforschung an der Universität zu Köln, Aachener Strasse 209, 50931, Köln, Germany*
[5]*RSSD, ESTEC/ESA, PO Box 299, 2200 AG Noordwijk, the Netherlands*
[6]*Institut d'Astrophysique Spatiale, Universite Paris XI, F-91405 Orsay, France*
[7]*Space Research Institute, Austrian Academy of Science, Schmiedlstr. 6, A-8042 Graz, Austria*
[8]*School of Physics, University of Exeter, Exeter EX4 4QL*
[9]*Oxford Astrophysics, University of Oxford, Keble Road, Oxford OX1 3RH*
[10]*LESIA, Observatoire de Paris-Meudon, 5 place Jules Janssen, 92195 Meudon, France*
[11]*Institut d'Astrophysique de Paris, UMR7095 CNRS, Université Pierre & Marie Curie, 98 bis boulevard Arago, 75014 Paris, France*
[12]*22 Observatoire de Haute-Provence, CNRS/OAMP, 04870 St Michel l Observatoire, France*
[13]*Instituto de Astrofísica de Canarias, E-38205 La Laguna, Tenerife, Spain*
[14]*Dept. de Astrofísica, Universidad de La Laguna, Tenerife, Spain*
[15]*Institute for Astronomy, University of Vienna, Türkenschanzstrasse 17, 1180 Vienna, Austria*
[16]*Institute of Planetary Research, DLR, 12489 Berlin, Germany*
[17]*IAG Universidade de Sao Paulo, Sao Paulo, Brazil*
[18]*Research and Scientific Support Department, European Space Agency, Keplerlaan 1, NL-2200AG, Noordwijk, the Netherlands*
[19]*IAG Université du Liège, Allée du 6 août 17, Liège 1, Belgium*
[20]*Universit de Nice Sophia Antipolis, CNRS, Observatoire de la Cte d'Azur, BP 4229, 06304 Nice, France*
[21]*Thüringer Landessternwarte, 07778 Tautenburg, Germany*
[22]*TU Berlin, Zentrum für Astronomie und Astrophysik, Hardenbergstr. 36, 10623 Berlin, Germany*
[23]*LUTH, Observatoire de Paris-Meudon, 5 place Jules Janssen, 92195 Meudon, France*

Accepted 2010 March 1. Received 2010 March 1; in original form 2010 February 9

ABSTRACT

Surveys for exoplanetary transits are usually limited not by photon noise but rather by the amount of red noise in their data. In particular, although the *CoRoT* space-based survey data are being carefully scrutinized, significant new sources of systematic noises are still being discovered. Recently, a magnitude-dependant systematic effect was discovered in the *CoRoT* data by Mazeh et al. and a phenomenological correction was proposed. Here we tie the observed effect to a particular type of effect, and in the process generalize the popular Sysrem algorithm to include external parameters in a simultaneous solution with the unknown effects. We show that a post-processing scheme based on this algorithm performs well and indeed allows for the detection of new transit-like signals that were not previously detected.

Key words: methods: data analysis – techniques: photometric – planetary systems.

⋆E-mail: avivofir@wise.tau.ac.il

1 INTRODUCTION

The limiting factor for most planetary transit surveys is not the theoretical photon noise but rather the practically achieved red noise from non-astrophysical sources (Pont, Zucker & Queloz 2006). The most capable transit surveys are the space-based surveys *CoRoT*[1] and *Kepler* because their stable environments allow for minimal red noise (and other benefits – such as continuous observations). Still, no instrument is perfect and the *CoRoT* light curves are known to show a number of significant effects that hinder transit detection, such as discontinuities of arbitrary magnitude due to high energetic proton flux near the South Atlantic Anomaly (SAA), residuals at the *CoRoT* orbital period, spacecraft jitter, CCD long-term aging and more. Many of these effects are correctable to satisfactory level at post-processing, but not for all stars and at all times.

On top of these effects, Mazeh et al. (2009) (hereafter M09) recently discovered that there are significant magnitude-dependant systematic effects in the *CoRoT* light curves, and they developed a phenomenological algorithm to correct for them. In this Letter we tie the above observation to a particular type of effect: added/subtracted linear flux, and are thus able to improve on their correction. In the process we generalize the Sysrem algorithm (Tamuz, Mazeh & Zucker 2005) to include arbitrary external parameters and show the benefits of using this modified version. In the following, we formulate the Sysrem generalization in Section 2, which is part of a complete post-processing scheme presented in Section 3, and then conclude.

2 THE SARS CORE

2.1 Algorithm

M09 first noted that there are magnitude-dependent systematic effects in the *CoRoT* data. They proposed to correct for the effects by fitting a parabola to the residuals of each exposure – but this correction is a purely phenomenological correction since there is no identified cause for the effects, and thus no explanation as to why a parabola is the best functional form. We hypothesize that the underlying physical mechanism M09 were trying to correct for is a constant flux that is either added or subtracted from all the light curves due to calibration errors, scattered light or other causes. Such an additive effect will create a large magnitude difference on faint stars, and small magnitude difference on bright stars, as M09 had originally observed. Indeed, the original authors had also considered this option (Tsevi Mazeh, personal communication) but they chose to use a more phenomenological correction rather than to tie the correction to this proposed physical mechanism. Since detrending algorithms cannot a priori disentangle additive from relative effects, we choose to simultaneously correct both types of effects, and so developed 'Simultaneous Additive and Relative Sysrem' – or the SARS algorithm – described below.

Suppose a matrix of photometric measurements of N stars ($i = 1, \ldots, N$) on M measurements ($j = 1, \ldots, M$) is given, so that the magnitude value of the ith star on the jth frame is m_{ij} and its associated error is $\sigma_{i,j}$. After removing from each stellar light curve (hereafter LC) its mean or median $\bar{m}_i$ we are left with the matrix of residuals r_{ij}. In the original Sysrem, the residuals of intrinsically constant stars are modelled using two contributions:

$$r_{ij} = A_j C_i + \text{noise}, \tag{1}$$

where A_j is the effect in each exposure and C_i is the effect's coefficient for each star. We note that since the data are in the magnitude system the effects found in this manner are relative in flux. In order to test out hypothesis that the magnitude-dependent effects stem from something that is additive in flux, we introduce the SARS model:

$$r_{ij} = A_j x_{ij} C_{A,i} + R_j C_{R,i} + \text{noise}. \tag{2}$$

Here the second term is exactly the usual Sysrem effect were we simply change the letter from A_j to R_j to designate that it is a relative effect. The first term stands for the additive effect by introducing $x_{ij} = 10^{0.4 m_{ij}}$ which makes sure that the effect is stronger for faint stars and weaker for bright stars – and in the correct functional form expected from additive flux effects. In practice, we use $x_{ij} = 10^{0.4(m_{ij} - m_{\text{rel}})}$, where m_{rel} is a constant number (e.g. the median of all the stars on all the exposures) to avoid overly large or small values for x_{ij}. As in Sysrem, minimizing the sum of squared residuals S,

$$S = \sum \left(\frac{r_{ij} - \text{model}}{\sigma_{ij}} \right)^2. \tag{3}$$

Given the best-fitting effects $\boldsymbol{R}$ and $\boldsymbol{A}$, and the corresponding coefficients $\boldsymbol{C_R}$ and $\boldsymbol{C_A}$:

$$A_j = \frac{\sum_i \frac{r_{ij} x_{ij} C_{A,i}}{\sigma_{ij}^2} - R_j \sum_i \frac{C_{R,i} x_{ij} C_{A,i}}{\sigma_{ij}^2}}{\sum_i \frac{x_{ij}^2 C_{A,i}^2}{\sigma_{ij}^2}}, \tag{4}$$

$$R_j = \frac{\sum_i \frac{r_{ij} C_{R,i}}{\sigma_{ij}^2} - A_j \sum_i \frac{C_{R,i} x_{ij} C_{A,i}}{\sigma_{ij}^2}}{\sum_i \frac{C_{R,i}^2}{\sigma_{ij}^2}}, \tag{5}$$

$$C_{A,i} = \frac{\sum_j \frac{r_{ij} x_{ij} A_j}{\sigma_{ij}^2} - C_{R,i} \sum_j \frac{A_j x_{ij} R_j}{\sigma_{ij}^2}}{\sum_j \frac{x_{ij}^2 A_j^2}{\sigma_{ij}^2}}, \tag{6}$$

$$C_{R,i} = \frac{\sum_j \frac{r_{ij} R_j}{\sigma_{ij}^2} - C_{A,i} \sum_j \frac{A_j x_{ij} R_j}{\sigma_{ij}^2}}{\sum_j \frac{R_j^2}{\sigma_{ij}^2}}. \tag{7}$$

As in Sysrem, the values of A_j, R_j, $C_{A,i}$, $C_{R,i}$ are iteratively refined until convergence is achieved. We found it important that in each iteration the new values of the effects are used to calculate the coefficients, and not the values of the previous iteration.

2.2 Further generalization

The formulae (4)–(7) do not 'know' that x_{ij} is meant to scale magnitude data to create flux-based correction – they only know that x_{ij} depends on external information not present in the original matrix of residuals. In this, SARS presents a significant departure from the original Sysrem by allowing for the detrending against any explicitly known external parameters as long as their effect can be encapsulated in some x_{ij}. For example, these can be distance from the centre of the CCD or pixel phase (or otherwise location based), CCD temperature (or otherwise weather related) or Moon phase (or

[1] The *CoRoT* space mission, launched on 2006 December 27, has been developed and is operated by CNES, with the contribution of Austria, Belgium, Brazil, ESA, Germany and Spain. *CoRoT* data become publicly available one year after release to the Co-Is of the mission from the *CoRoT* archive: http://idoc-corot.ias.u-psud.fr/

otherwise temporal effects), etc. It is thus easy to include multiple external effects in the detrending model (e.g. the SARS model of equation 2), and by minimizing the sum of squared residuals S to establish their effect on the data simultaneously with the effects of unknown sources.

2.3 Suggested good practices

Below we describe what we think are good practices when using SARS:

(i) Starting point: we note that already the Sysrem 'search space' was very large: as many parameters to adjust as there are stars + exposures. By simultaneously fitting more than one effect in SARS – we further enlarge this search space greatly. In the original Sysrem the starting point was deemed to be unimportant since Tamuz et al. (2005) claimed that in their simulations no matter what initial values were used, the same effect and coefficients were obtained. We believe that these simulations were somewhat lacking in that they used white noise only, with no red noise component, which allowed them to always find the (unique) global χ^2 minimum with no local minima to be avoided. We have also performed a similar test – but on real data, rather than simulated data, and found that sometimes (a few per cent of the runs) the global minimum was indeed missed. Fearing that the enlarged SARS search space would worsen the problem, we choose to start the iterations from a deterministic point. Assuming that the median of photometric measurements is rather robust, we start by finding a proxy to the relative and additive effects by $R_j = \mathrm{median}(r_{ij})$ and $A_j = \mathrm{median}(r_{ij}x_{ij})$, respectively. We set all $\boldsymbol{C_R}$ and $\boldsymbol{C_A}$ to unity since we wish to find effects that affect many (if not all) stars.

(ii) Convergence criteria: the convergence criteria for the above iterations was unspecified in Tamuz et al. (2005). We define it as the iteration when the maximal absolute value of total correction $\mathrm{abs}(A_j x_{ij} C_{A,i} + R_j C_{R,i})$ is smaller than some fraction f of the standard deviation of that particular object. We used $f = 0.5$ in our processing.

(iii) Once either additive or relative effects show no further correlation, one can use the regular Sysrem to look for additional effects of the other type since one may have different number of relative and additive effects – until no effects of either type are identified.

(iv) Bright stars both make planetary transits easier to detect, and are more susceptible to relative effects [which are later corrected by Sysrem/TFA (Kovács, Bakos & Noyes 2005)/other]. For these reasons some of the transit surveys intentionally monitor only the relatively bright stars in their field of view. On the other hand, fainter stars more readily show additive effects. We therefore propose to add more faint stars to the data of such surveys when using SARS as they might hold the key to better correct all stars.

3 THE SARS POST-PROCESSING SCHEME

3.1 Post-processing steps

The above SARS core is just one element of the SARS post-processing scheme. We were able to achieve complete automation with no human input from *CoRoT* N2 FITS files to clean LCs. The post-processing global structure was similar to the one used by M09.

(1) Resample to 512s: resampling is done for each *CoRoT* colour separately, if available.

(2) Divide to ∼10 d blocks, process blocks individually.

(3) Subtract a running median with a window the size of three *CoRoT* orbital periods, and reject outliers.

(4) Choose a 'learning set' to calculate the effects with.

(5) Apply the effects to all stars (we used three pairs of effects).

(6) Reset errors and reject bright outliers.

However, in order to achieve full automation we elaborate below on steps that were either unspecified or human dependant on M09.

(7) Outliers rejection is done in three tiers:

(i) Removal of solitary outliers that are far from a small-window (five-point) median filter.

(ii) Further outliers must meet two criteria: (1) that frame has anomalously high median absolute deviation (usually SAA-affected frames) and (2) the data point is far from a three-orbit median.

(iii) Before SARS-core application, frames must have a minimum number of valid learning-set stars (we used at least 100).

(8) Automatic choosing of the learning set aims to isolate the intrinsically and instrumentally constant stars. These stars are assumed to be numerous and similarly variable in the raw data. An initial learning set is chosen by multiple criteria:

(i) the alarm statistic of the LC (Tamuz, Mazeh & North 2006) must be part of the bulk of alarms;

(ii) the alarm statistic of the residuals must be part of the bulk of alarms;

(iii) the locus of constant stars on the log(rms) versus magnitude plot is along a straight line. Learning-set stars must not be far from that locus.

Next, the learning set is refined by a procedure inspired by techniques originally developed for photometric follow-up of transiting planets (Holman et al. 2006) and is aimed at delivering the best comparison signal (lowest relative noise).

(i) Given a set of N stars, calculate relative error on the total flux for all N subsets of $(N - 1)$ stars.

(ii) Compare the best subset (having the lowest relative flux error) with the relative flux error of the sum of N stars.

(a) If error reduced: repeat from step (i) with $(N - 1)$ stars.

(b) If error increased: optimal set reached.

This procedure guarantees that a local minimum in relative error is reached. We opt not to search for the global optimum since this is deemed too difficult (testing all subgroups of N stars require testing $N!$ configurations, where N is in the order several thousands). We note that the resultant learning set preferentially includes faint stars (see Fig. 1) which at least partially is because any variability is easier to spot on brighter stars. Interestingly, Fig. 1 and panel 2 of Fig. 2 show that despite the fact that faint stars are preferentially selected as learning-set stars – bright stars are better corrected, showing that indeed something was learned from the fainter objects and was well applied to the brighter stars.

(9) We use the SARS core three times.

(i) Use on the learning set only – used to recalibrate the errors only (see below).

(ii) Use on the learning set only (now with calibrates errors) – to calculate the effects.

(iii) Use on the whole data set – apply the already calculated effect to all the LCs.

(10) Errors resetting is done by

$$\mathrm{err}_{ij} = \mathrm{StarErr}(i)\frac{\mathrm{ExpErr}(j)}{\mathrm{median}(\mathrm{ExpErr})}, \quad (8)$$

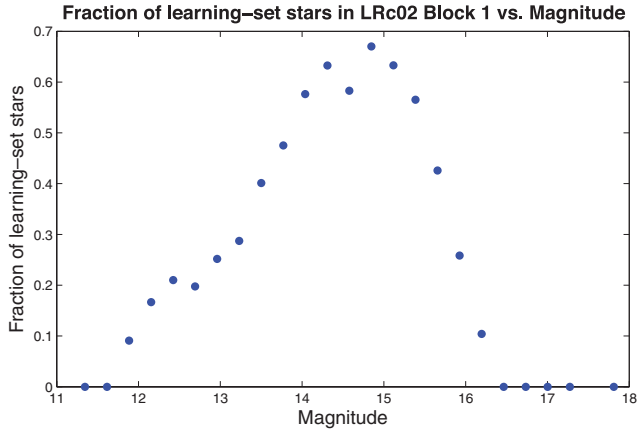

Figure 1. A representative plot of the fraction of learning-set stars in several magnitude bins (data here are from the example presented in Section 3.2, for the learning set of the first effects-pair in the first block of LRc02).

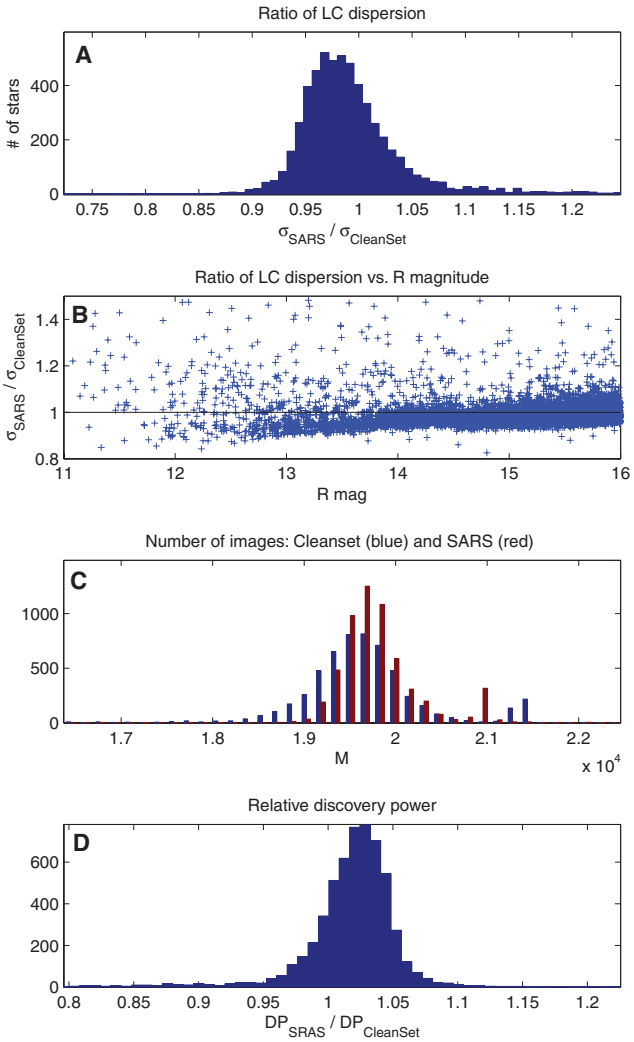

Figure 2. Comparison between SARS and CleanSet for one random field (CCD E1 of LRc02). Panel A: histogram of the ratio of LC dispersion in SARS and CleanSet. Note $\sigma_{\rm SARS}/\sigma_{\rm CleanSet} < 1$ for most stars. Panel B: the above ratio of LC dispersion versus R magnitude: the observed trend is explained in the text. Panel C: histogram of the number of remaining data points after outlier rejection in SARS and CleanSet. Panel D: histogram of the ratio of detection power (DP) between SARS and CleanSet.

where StarErr is estimated from the star's LC and ExpErr is estimated from the distribution of magnitude residuals of each exposure.

3.2 Results

A comparison of the performance of SARS-cleaned and M09-cleand LCs (the later sometimes dubbed 'CleanSet') of one random field (LRc02) is shown in Fig. 2: the SARS post-processing delivers lower LC dispersion than in M09's CleanSet for about 65 per cent of the stars, while keeping at least the same number of valid data points M (CCD E2) if not more (CCD E1). If we compare the detection power, which is defined as DP $\sim \sqrt{M}/\sigma$, we find that it is higher in SARS than in CleanSet for up to 80 per cent of the stars. We note that there is a small trend in the relative performance: the brighter the stars are the better SARS is relative to CleanSet. This is the expected result of the approximated functional form of the M09 correction: since there are many more faint stars in the data than bright stars, the parabolic least-squares correction of M09 tends to better suite the numerous faint ones, and so less fitting (due to the

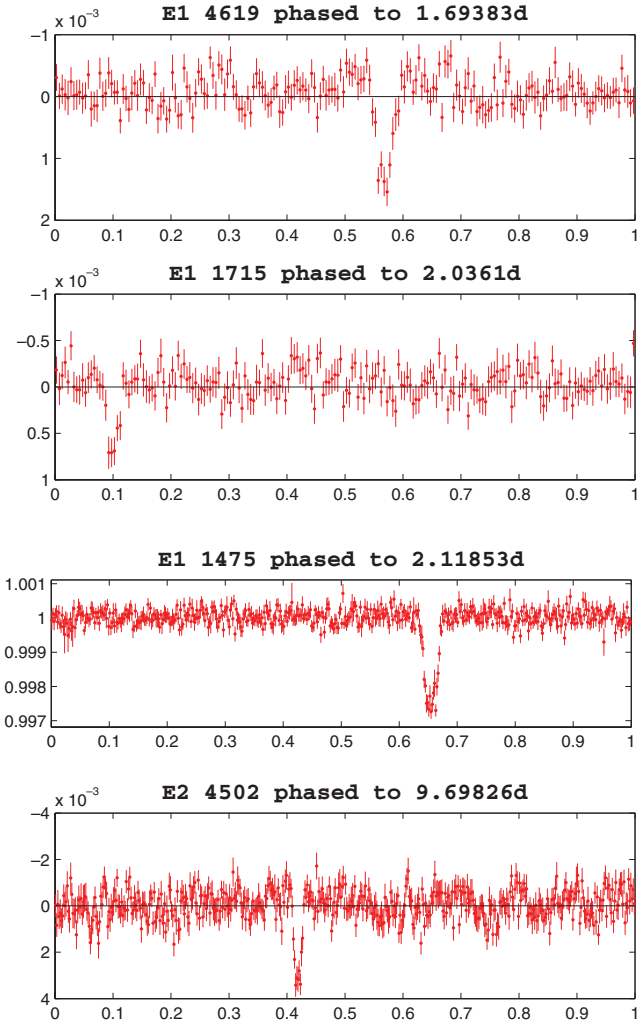

Figure 3. Example of four *CoRoT* LRa02 candidates that were not detected by any of the detection teams before the application of the SARS post-processing, and which are currently being followed up. The light curves are folded and binned to aid visibility.

approximated functional form) to the bright stars. Thus our initial hypothesis that the effects are additive seems even more robust.

This global statistics is also translated to specific detections: so far we have SARS-analysed three long runs (LRc02, LRa02 and LRa01) and found about 10 new transit-like signals that were not detected before in each field. For example in Fig. 3 we show four such LRa02 new transit candidates that were also chosen for follow-up.

4 DISCUSSION

At the start of the *CoRoT* mission the *CoRoT* Exoplanets Science Team (CEST) made the strategic choice of having multiple team analysing the exact same input data. By cross-checking each detection with different tools and cleaning techniques (e.g. Cabrera et al. 2009; Carpano et al. 2009) the CEST hoped to achieve the best possible transit candidates list for follow-up observations by the limited ground-based telescope resources. Here we present yet another step in the journey to clean photometric data sets in general and *CoRoT* data in particular: we generalize the popular and efficient Sysrem algorithm to include external parameters in a simultaneous solution, together with the unknown effects. This allows us to show that data from *CoRoT* is probably contaminated with additive, rather than relative, systematic effects – and that these effects are the probable cause behind the phenomenological observation of M09. The size of the additive correction abs($A_j x_{ij} C_{A,i}$) is comparable to that of the relative correction abs($R_j C_{R,i}$), with a median additive-to-relative ratio of about 0.5, but with a large scatter – making the additive correction larger than the relative correction for $\sim$1/3 of the data points. Additive effects can arise from scattered light or erroneous bias or background subtraction, and we believe that the additive effects can be used just as the regular Sysrem effects to help to trace down the origin of the systematics and thus to avoid them in the first place.

We believe that the main advantage of SARS is not in a dramatic change in the standard deviation σ of the LCs, but rather in the whiter colour of their noise, which in turn allows for lower background of spurious signals in the BLS spectra (Kovács, Zucker & Mazeh 2002) and thus the detection of shallower signals.

We were able to achieve good performance and complete automation which allows us to now process the entire mission data, and to look for – and find – ever shallower transit-like signals. For example on LRc02 target *CoRoT* ID 0105842933 we were able to clearly detect a very shallow signal, only 10^{-4} mag deep, in a $P = 1.08\,085$ d period. Not only that, but we were also able to show that this is an eclipsing binary since at the double period the odd and even eclipses have different depths, with the secondary eclipse still visible (on a binned LC) while having a depth below the 84 ppm depth of an exo-Earth around a Sun-like star.

We will make the SARS-cleaned light curves available for the *CoRoT* community, and it is our intention that when the proprietary period is over to make data generally available (upon request). We note that we have also allowed for the application of SARS to the residuals of astrophysically variable stars (pulsators, eclipsing binaries, etc.) which will allow to better clean them too – as part of a parallel CEST effort to look for transiting circumbinary planets (Ofir 2008; Ofir, Deeg & Lacy 2009).

REFERENCES

Cabrera J. et al., 2009, A&A, 506, 501
Carpano S. et al., 2009, A&A, 506, 491
Holman M. J. et al., 2006, ApJ, 652, 1715
Kovács G., Zucker S., Mazeh T., 2002, A&A, 391, 369
Kovács G., Bakos G., Noyes R. W., 2005, MNRAS, 356, 557
Mazeh T. et al., 2009, A&A, 506, 431 (M09)
Ofir A., 2008, MNRAS, 387, 1597
Ofir A., Deeg H. J., Lacy C. H. S., 2009, A&A, 506, 445
Pont F., Zucker S., Queloz D., 2006, MNRAS, 373, 23
Tamuz O., Mazeh T., Zucker S., 2005, MNRAS, 356, 1466
Tamuz O., Mazeh T., North P., 2006, MNRAS, 367, 1521

This paper has been typeset from a TEX/LATEX file prepared by the author.