0n the Extension of *CoRoT*

JC, SzCs, AE, HR

December 9, 2009

1 Summary

Triggered by the mail from Claire Moutou (see below, section 2), we investigate the impact of the length of the runs in the number of candidates found and their properties:

1. it seems that the **number of candidates found depends more on the nature of the stellar population and on the impact of the noise than on the length of the run;** although a dependence of the efficiency of the detection algorithms as a function of the period is suspected for short runs;

2. more **small candidates are found in long runs, mainly around bright targets** because the detection threshold approaches better the red noise limit;

3. **the accuracy of the physical parameters determined strongly depends on the length of the runs,** and up to one order of magnitude in the precision of the ephemeris can be gained going from a short run (20 days) to a long run (140 days);

To make several short runs instead of a long run probably entails the detection of more candidates, but not necessarily more interesting (i.e. smaller) or more accessible (i.e. better characterized or suitable for follow-up) objects. A more detailed study of the different factors that affect the detection efficiency will be done.

2 Claire's email 09.11.2009

I think we'd need some more work on the statistical analysis of previous runs, something like the "Juan's plot" shown in Jean's slides, but 1) with higher statistics, 2) for center and anticenter directions individually, 3) separately for stars brighter than 14.5 and for fainter stars. I'm sure that Corotlux users can also help to derive the useful range of a future run, given the actual threshold of corot. Juan, can you take the action to revisit your statistics and report on a wider context?
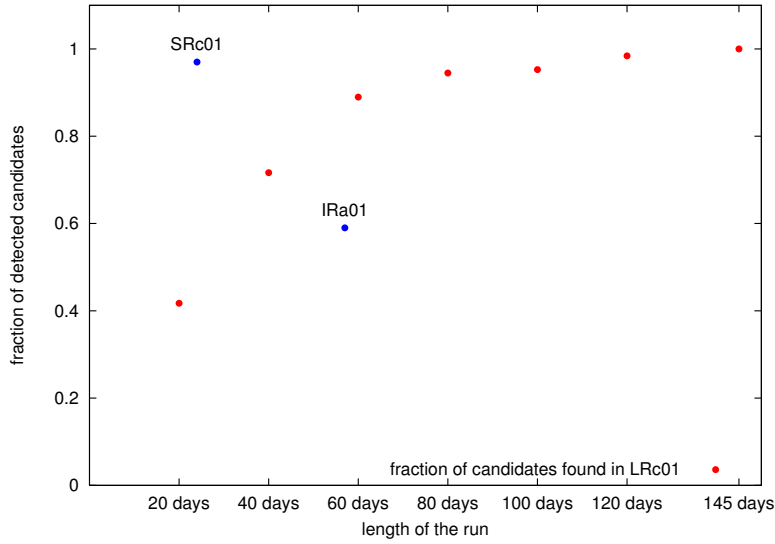
Figure 1: Yield plot: the expected fraction of candidates detected in one run as a function of the length of the run.

3 Motivation

The original yield plot (see appendix A) was built to test the performance of the detection algorithms as a function of the length of the run. The study was performed in the LRc01 run (145 days long in average) with only one filtering and one detection algorithm in the version 1.2 of the N2 data. The conclusions were, in short, that in an *intermediate* run of 80 days long 97% of the known candidates in the whole run were found. However, a warning was included about the impact of *intermediate* runs not on the number of candidates, but in the **quality** of the physical parameters which could be derived.

In this report we revisit the conclusions of the original statements analyzing and comparing the results of the runs which have been validated: IRa01, LRc01 and SRc01. In this context, validated means that each candidate, period and depth has been checked by human inspection to discard errors and false alarms. Therefore, the results should be homogeneous. The same analysis on LRa01 has been started and will be added soon.

4 The Yield Plot

The original yield plot, as defined in the appendix A, shows the expected fraction of candidates detected in one run as a function of the length of the run. We are going to compare the theoretical rates predicted by the yield
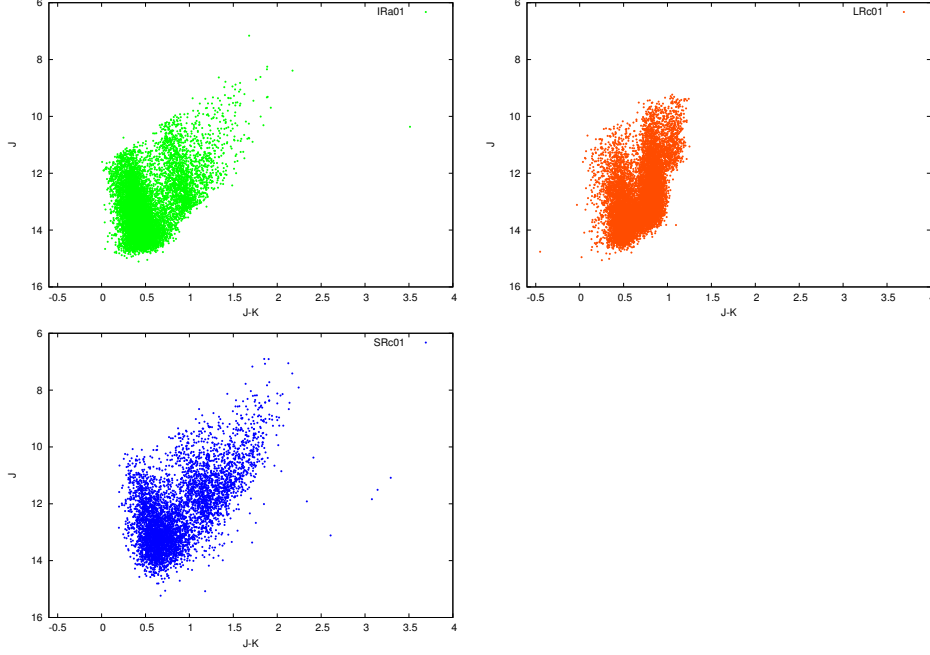
Figure 2: Color/magnitude diagram for IRa01 (upper left), LRc01 (upper right) and SRc01 (lower left).

Table 1: Details of the runs.

run	length (days)	stars	dwarf stars (%)	dwarf stars	candidates & binaries
IRa01	57	9 872	67%	7 898	195
SRc01	24	6 974	80%	4 637	188
LRc01	145	11 408	42%	4 791	200

plot with the validated results of IRa01 (SC09[1]), LRc01 (JC09[2]), and SRc01 (AE09[3]).

The fraction of candidates has been normalized in the following way: according to SA09[4], there is a fraction of 80% of dwarf stars in IRa01 compared to a fraction of only 42% of dwarf stars in LRc01. In AE09 a similar estimation (based on the J-H/J color-magnitude diagram) is done and a fraction of 67% of dwarf stars in SRc01 is found. This is the reason why we expect 97% ($188/200 \cdot 4791/4637 = 0.97$) and 59% ($195/200 \cdot 4791/7898 = 0.59$) of candidates respectively (see Table 1).

If the approach of the yield plot is right, a short run like SRc01 (that lasts

[1] Carpano et al. (2009) A&A, 506, 491-500.
[2] Cabrera et al. (2009) A&A, 506, 501-517.
[3] Erikson et al. (2009) in preparation.
[4] Aigrain et al. (2009) A&A, 506, 425-429.

24 days in average) we expect to find 50% of the candidates found in LRc01; and in IRa01 (that lasts 57 in average) we expect 85% of the candidates of LRc01. However, the figures found are 97% and 59% respectively (see Fig. 1). The positions of the runs IRa01 and SRc01 in the yield plot do not match their expected values. There are three possible explanations:

1. the yield plot is a good model, but the distribution function of candidates is not the same for IRa01, SRc01 and LRc01 (which could be the case if each run has a different stellar population);

2. the distribution function of candidates is the same in the 3 runs, but the yield plot is a bad model;

3. the yield plot is a bad model and the distribution function of candidates in the different runs is not the same.

We know from Table 1 and from the color/magnitude diagrams (see Fig. 2) that the stellar populations are not the same in the three runs, so we should not expect the distribution functions of candidates to be similar. But even though, we do not have enough information to prove whether the yield plot is a good model or not.

One can imagine a extreme case in which the number of candidates found is independent of the lenght of the run. This would explain why we find the same number of candidates in LRc01 and SRc01. The lack of candiates in IRa01 could be produced by a higher noise level which prevents the detection of small transits (see the discussion below). In this case, one have to provide an explanation for the dependence of the number of candidates as a function of the length of the run in the yield plot of LRc01. Is the noise level similar in the first 20 days of the run as in the following? Do we obtain the same results when analyzing shortened runs taking only the last 20 days, the last 40 days and so forth? Do we obtain the same dependence analyzing LRa01? This questions will be addressed in future work.

The dependence of the number of candidates on the length of the run is not fully understood. To analyze the problem further, we compare the results of the three runs in detail in the following.

5   The Distribution Function of Candidates

One can define the probability distribution of the expected number of candidates (both binary stars and planets) as a function of their period as:

$$N(p) = \int_0^p dp' \, N_{\mathrm{obs}} \, \eta(p') \, \epsilon(p') \, \Pi(p')$$

where N_{obs} is the observed number of stars; $\eta(p')$ is the fraction of stars with companions at a given period; $\epsilon(p')$ is the detection efficiency of the
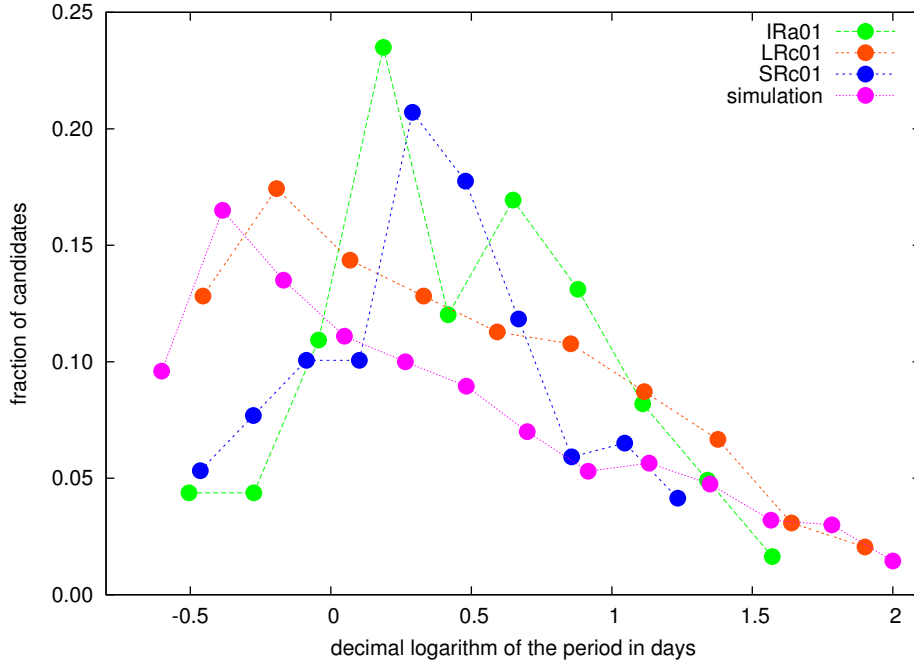
4

Figure 3: Histogram with the fraction of candidates found as a function of the decimal logarithm of their period in days.

algorithms and $\Pi(p')$ is the geometric probability of having a transit. One can naively set $\eta(p') = 1$; and $\epsilon(p') = 1$; and:

$$\Pi(p') \approx \frac{R_*}{a_{\mathrm{pl}}} = \frac{R_*(2\pi)^{2/3}}{[G(M_* + M_{\mathrm{pl}})]^{1/3}} p'^{-2/3}.$$

and therefore the fraction of candidates found with period between p and $p + \delta p$ is:

$$n(p; \delta p) = \int_p^{p+\delta p} dp' \Pi(p') = \int_p^{p+\delta p} dp'(\alpha + 1) \, K \, p'^\alpha = K(\delta p^{\alpha+1} - 1)p^{\alpha+1}$$

and in our case $\alpha+1 = 1/3$. At this point we realize that the assumptions are too simple because this result implies an exponential growth of the number of candidates with the period, which is unrealistic even for short runs. The error comes from the assumption $\eta(p') = 1$. If we take a more realistic distribution, like the log-normal distribution of DM91[5], we obtain the results shown in Fig. 3.

In Figure 3 we compare the histograms of the distributions of the fraction of candidates as a function of the decimal logarithm of their period (in days)

---

[5]Duquennoy & Mayor (1991) A&A, 248, 485-524

for the IRa01, LRc01, SRc01 and a simulated population of stars. For the simulation, we create a population of a certain number of stars with companions that follow the log-normal distribution of periods given by DM91. In this population we search for a certain number of eclipsing companions. For each companion, we define a random inclination; if the orientation of the orbit is favorable and the period of the companion is between 0.25 and 100 days (which are the maximum and minimum periods found in LRc01) we classify the star as a detected binary. Finally, we calculate the expected histogram of detections for a long run. The distribution of LRc01 and the simulated run follow quite the same distribution (the KS test value is 34%, although there are not enough values to have good statistics).

Comparing the histograms of Fig. 3 we notice that LRc01 follows the expected distribution if $\epsilon(p') = constant.$ (where ϵ is the factor that depends on the efficiency of the detection algorithms). The power-law decay for long periods is understood as the effect of the geometric factor. But IRa01 and SRc01 do not follow the same distribution. The reason why IRa01 and SRc01 have their peak in the distribution between 1 day and 3 days period instead of between 0.3 and 1 day as expected is not understood yet. Is this fact due to a different distribution of binaries or due to a selection effect of the detection algorithms? If the probability distribution function is right, it seems that the detection efficiency is constant (with the period) in LRc01 (and therefore its distribution is close to the theoretical one), but the detection efficiency seems to show some dependence on the period for SRc01 and IRa01 (and hence the peak of the distribution is displaced and the slope of the power law affected). Do the discontinuities produced by the passages through the South-Atlantic Anomaly have an impact in the detection of short-period candidates? This effect should act like the day/night discontinuities on ground-based surveys. However, as CoRoT orbits the Earth, the passages through the SAA change in phase and hence this effect is averaged out in longer runs.

6   The Magnitude Distribution

Figure 4 shows the histogram of the number of candidates as a function of R magnitude. If we normalize the fraction of candidates to the total number of targets in the run with the same magnitude, the fraction then looks constant (excluding the bright end, where we have fluctuations due to the low number of targets). It seems that the efficiency of detection does not depend strongly on the magnitude. But do all candidates in different magnitude bins have similar period and depth distributions?
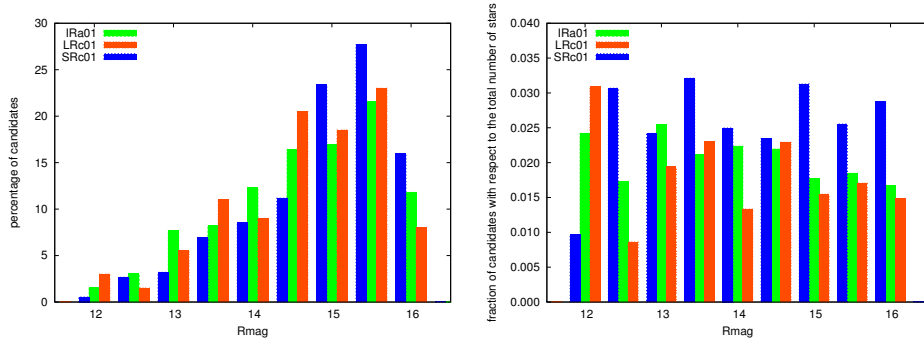
Figure 4: Histograms of the distribution of candidates as a function of R magnitude. In blue SRc01, in green IRa01 and in red LRc01. Left: the percentage of candidates with respect to the total number of candidates. Right: the fraction of candidates with respect to the number of targets in the same magnitude bin.

7   The Depth Distribution

Figure 5 shows, as a function of R magnitude, the measure of the transit signal as it was used in CM09[6] and JC09 based on the definition given by FP06[7]; in other words, this is the red-noise diagram. We expect a dependence of the transit depth as a function of the magnitude. It is not clear that IRa01 reaches the same limit as LRc01 and SRc01. This could be due to a source of systematic noise only present in this initial run, which was processed with a different pipeline than the succesive runs (see also the discussion in CM09). But this possibility has to be checked comparing the results of the analysis of the latest release of IRa01 (N2 v. 2.1), which so far have not produced a significative amount of new candidates.

Figure 6 shows, on the left, the distribution of the depth of the candidates for each run. We find more small candidates in LRc01, compared to SRc01 and specially IRa01. This is not caused by a bias in the magnitude (i.e. we detect more small candidates because we see more bright stars), but probably due to the fact that, for the same period, we are able to integrate more orbits in a longer run, achieving a better signal to noise ratio (the SNR grows as the square root of the number of points observed in transit), although a systematic additional source of noise in IRa01 is not discarded. We can see that in the bright end (when we keep only the candidates brighter than 14.5; see Fig. 6 right) the difference in the percentage of small candidates is less pronounced among the runs; so **we are actually finding smaller transits in longer runs.** This is an important difference when we are looking for planetary-like objects.

---

[6]Moutou et al. (2009) A&A, 506, 321-336.
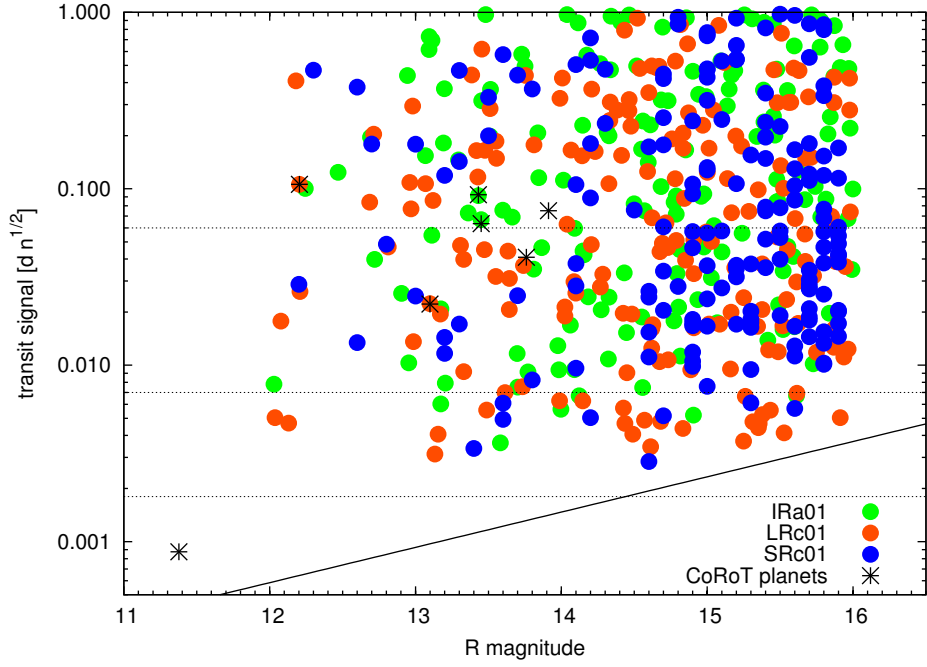[7]Pont et al. (2006) MNRAS, 373, 231-242.

Figure 5: Transit signal versus R magnitude for the candidates in the three runs The asterisks represent the position of the six planets, and the brown dwarf discovered by *CoRoT* (*CoRoT*-7b is in the bottom left corner). The horizontal dashed lines represent (from top to bottom) the expected signal produced by a Jupiter-size planet, a Neptune-size planet, and a 2 Earth-radii planet, respectively.

8 The Accuracy of the Period

The second big warning raised with the discussion of the yield plot was that although a comparable number of candidates could be obtained with a shorter run, the precision achieved in the determination of the physical parameters was not the same. This affects mainly the error in the estimation of the period, as it is shown in this section. A large error of the period makes the photometric follow-up impossible and therefore the confirmation process relies completely on the radial velocity follow-up, which becomes too risky because of the high amount of background contaminants. In Table 2 we show the median value of the published errors for the periods in the three runs[8]. We see a clear dependence of the median value with the length of the run: we gain an order of magnitude in LRc01 compared to IRa01.

Although strongly affected by low-number statistics, if we compare in

---

[8]in this case, the median is probably a better estimator than the mean.
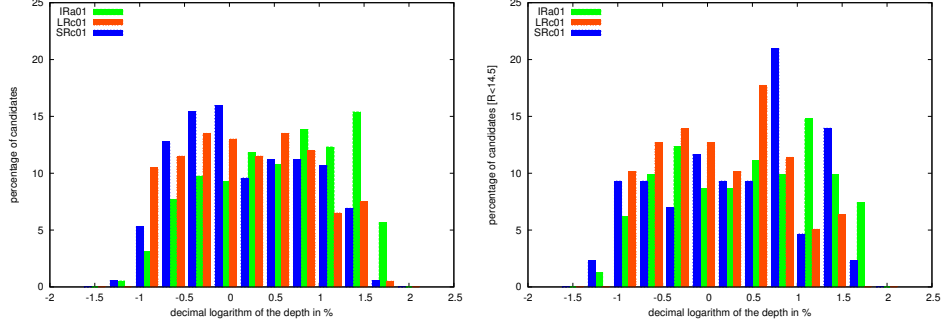
8

Figure 6: Histograms of the distribution of candidates as a function of their depth. In blue SRc01, in green IRa01 and in red LRc01. Left: the percentage of candidates with respect to the total number of candidates. Right: the percentage of candidates brighter than R=14.5 with respect to the total number of candidates brighter than R=14.5.

Table 2: Mean and median values for the error in the period

run	mean error (days)	median error (days)	accumulated error after 100 transits (hours)
SRc01	1.5e-3	3.8e-4	0.91
IRa01	3.7e-3	1.9e-4	0.45
LRc01	4.9e-4	1.5e-5	0.036

Table 3 the number of planets found in the different runs, we see that **no planet has ever been found in a short run.** A probable explanation might be that although even if small candidates around bright targets are found at reasonably the same rate in the different runs; shorter runs have intrinsic follow-up difficulties that make the planet detectability tougher. However, other possible explanations were given by CM09 and JC09.

Table 3: Planets found (in brackets planets not fully confirmed yet).

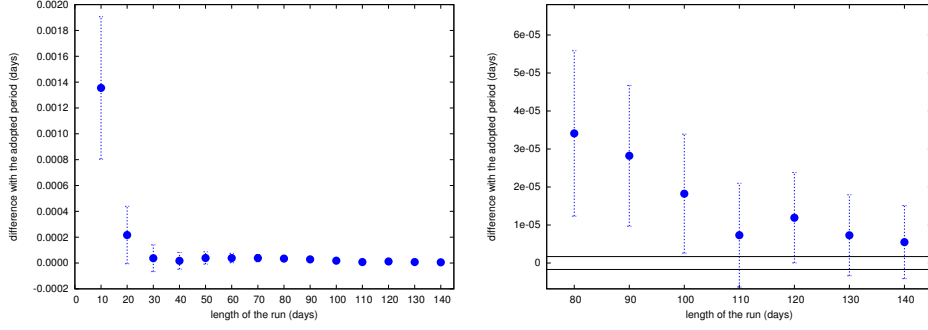| run	planets	run	planets	run	planets
SRc01	-	IRa01	1b, 4b	LRc01	2b, 3b, 8b, 10b
SRc02	-			LRc02	6b, 9b, 11b
SRa01	-			LRa01	5b, 7b, [13b]
SRa02	-			LRa02	
				LRc03	[12b]
total	0		2		11

9

Figure 7: Difference between the calculated period in a shortened run and the published period for *CoRoT*-2b. Right: zoom of 80-140 days.
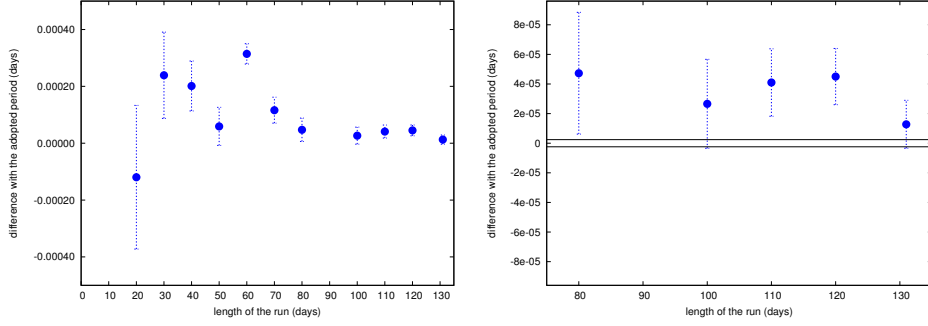

Figure 8: Difference between the calculated period in a shortened run and the published period for *CoRoT*-7b. Right: zoom of 80-140 days.
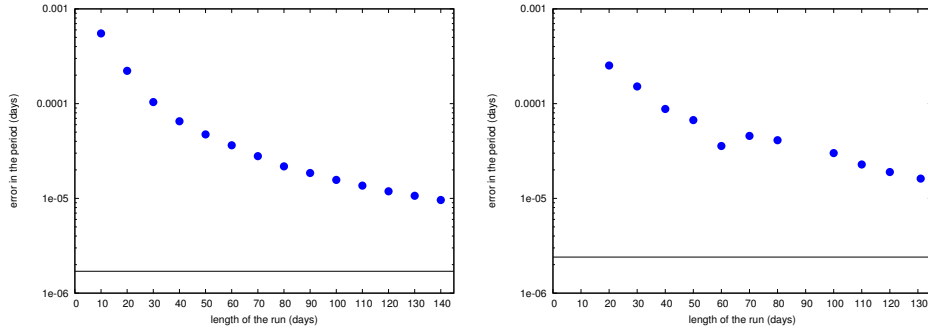

Figure 9: Dependence of the error in the determination of the period as a function of the length of the run compared with the published values for *CoRoT*-2b and *CoRoT*-7b.
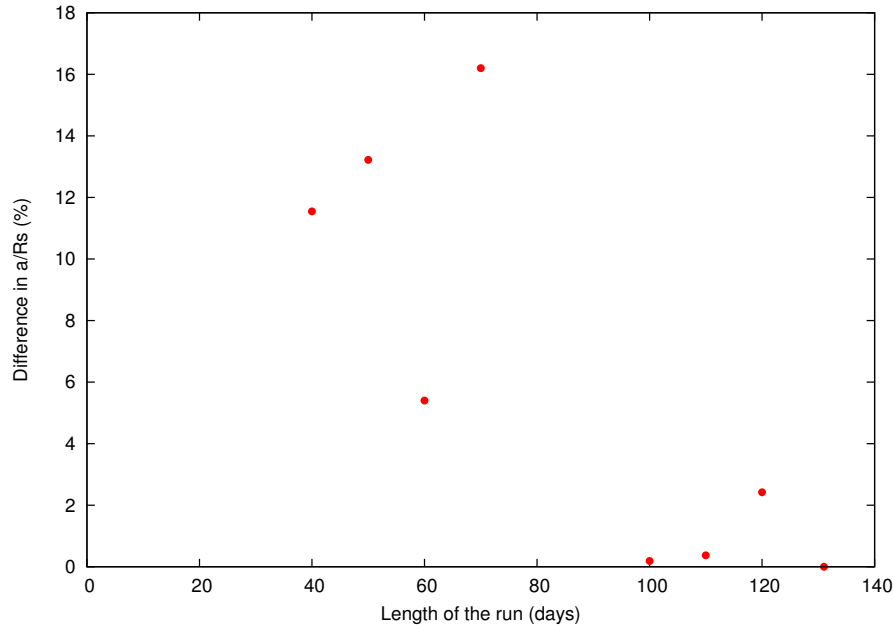
10

Figure 10: Deviations of the parameters a/R_s for each segment (see text for further explanation)

9 The Accuracy of the Physical Parameters: a test

To understand the dependence of the accuracy of the physical parameters the following test is designed: we study the value of the period and its error of the planets *CoRoT*-2b and *CoRoT*-7b as a function of the length of the run. The calculated periods and errors were compared to the published values, which are $1.742\,966\,4(17)$ for *CoRoT*-2b (RA08[9]) and $0.853\,585(24)$ for *CoRoT*-7b (LRS09[10]). Figures 7, 8 and 9 show the results: we can see a dependence in the error of the ephemeris on the length of the run; this evolution appears to stabilize after 100 days of observations. It is clear that the automatic methods used in this study do not reach the precision of the analysis performed for the papers, but the gain of one order of magnitude in the precision from a short run to a long run is consistent with the result of section 8.

9.1 The Accuracy of the Physical Parameters

We also investigated how the accuracies of the physical/geometrical parameters extracted from the light curve depend on the length of the run using the

---

[9]Alonso et al. (2008) A&A 482, L21.
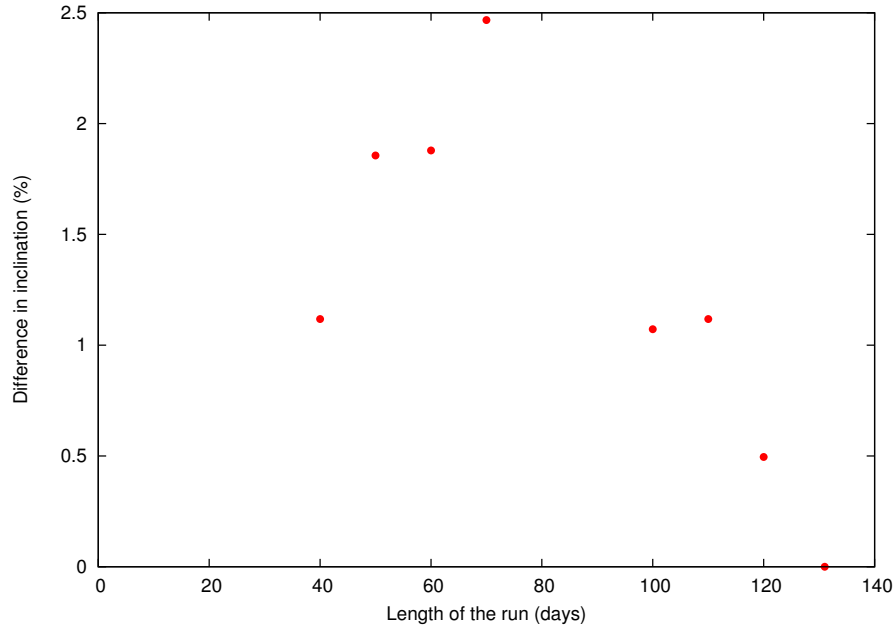[10]Léger, Rouan, Schneider et al. (2009) A&A, 506, 287.

Figure 11: Deviations of the parameters i for each segment (see text for further explanation)

same light curve segments as in the previous section (8). Each light curve segments was processed by the same procedure described in the *CoRoT*-6b paper: MF09[11]: Savitzky-Golay filtering and Harmony Search. As can be seen in that paper, this method yielded the same results (within the error bars) as the analysis performed by Roi Alonso with another set of tools for the same planetary system.

The light-curves were folded in 500 bins using for each segment the value of the period found in the previous subsection (8).

All these light curves were modeled in the same way as described in MF09. We used the MA02[12] formalism and the Harmony Search-algorithm for the optimization which worked quite well e.g. for the *CoRoT*-6b paper. The adjusted parameters were: a/R_s, R_p/R_s, inclination and two limb darkening coefficients for the quadratic limb darkening law. Figures 10 to 12 show the deviations of the parameters a/R_s, i and R_p/R_s for each segment (the deviations are calculated as the relative difference to the last - and longest: 131 days - segment).

We note that the typical relative errors in each segment were $\sim 5\%$ for a/R_s, $\sim 2\%$ for the inclination and $\sim 3\%$ for the radius ratio R_p/R_s. Taking this into account, one can conclude from Figs. 10 to 12 that **we need runs**

---

[11]Fridlund et al. (2009) A&A, submitted.
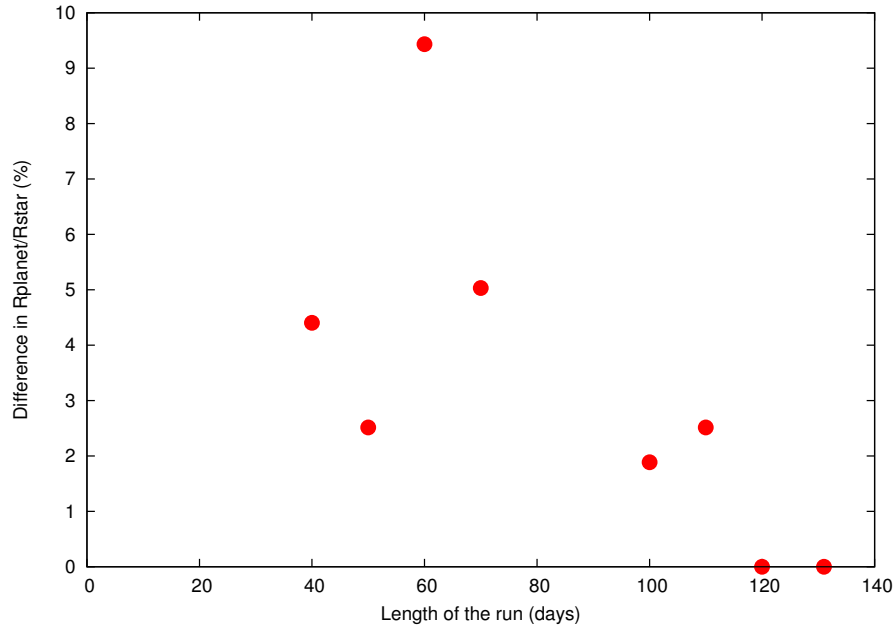[12]Mandel & Agol (2002) ApJ, L171.

Figure 12: Deviations of the parameters R_p/R_s for each segment (see text for further explanation)

longer than 100 days to determine with acceptable accuracy the physical parameters of *CoRoT*-7b.

One can formulate this result in an other way: since we have 117 transits of *CoRoT*-7b during 100 days, we need this number of transits for the average light curve to reach a good solution. If the next Earth-like planet has a longer period but the activity of the host star has the same level (since the amplitude will be more or less the same), we need proportionally longer runs! For a high-amplitude Jupiter-like planet $(1-2\%$ depth) it is no question that we can determine the physical parameters with appropriate accuracy (see e.g. the *CoRoT*-9b case).

10 Conclusions

We have investigated the impact of the length of the runs on the number of candidates found and their properties:

1. it seems that the **number of candidates found depends more on the nature of the stellar population and on the impact of the noise than on the length of the run;** although a dependence of the efficiency of the detection algorithms as a function of the period is suspected for short runs;

13

2. more **small candidates are found in long runs, mainly around bright targets** because the detection threshold approaches better the red noise limit;

3. **the accuracy of the physical parameters determined strongly depends on the length of the runs,** and up to one order of magnitude in the precision of the ephemeris can be gained going from a short run (20 days) to a long run (140 days);

To make several short runs instead of a long run probably entails the detection of more candidates, but not necessarily more interesting (i.e. smaller) or more accessible (i.e. better characterized or suitable for follow-up). Further analysis of the different open questions in this report will follow.

A Original Yield Report sent on 03.03.2009

On the performance of detection algorithms as a function of the length of the run

The SC is considering to split one of the future long runs of 150 days in two *intermediate* runs of 75 days. In order to better understand the advantages and limitations of this strategy, we decided to make a test on the performance of the detection algorithms in function of the length of the run.

The idea is to take the *CoRoT* data from the LRc01 and construct several *shortened runs*: the first run comprising just the first 20 days of all the light-curves in the LRc01; the second run comprising just the first 40 days and so on. We applied the standard pipeline to all these runs with the same filtering and detection threshold as was applied for the whole LRc01 and finally we counted the number of candidates found.

RESULTS

We can measure the performance of the detection algorithm as the fraction of candidates found in a shortened run compared to the total number of candidates known in the whole run; the results are given in the figure below. For example, analyzing only the first 80 days of the LRc01 we are able to find 94% of the candidates known in the whole run.
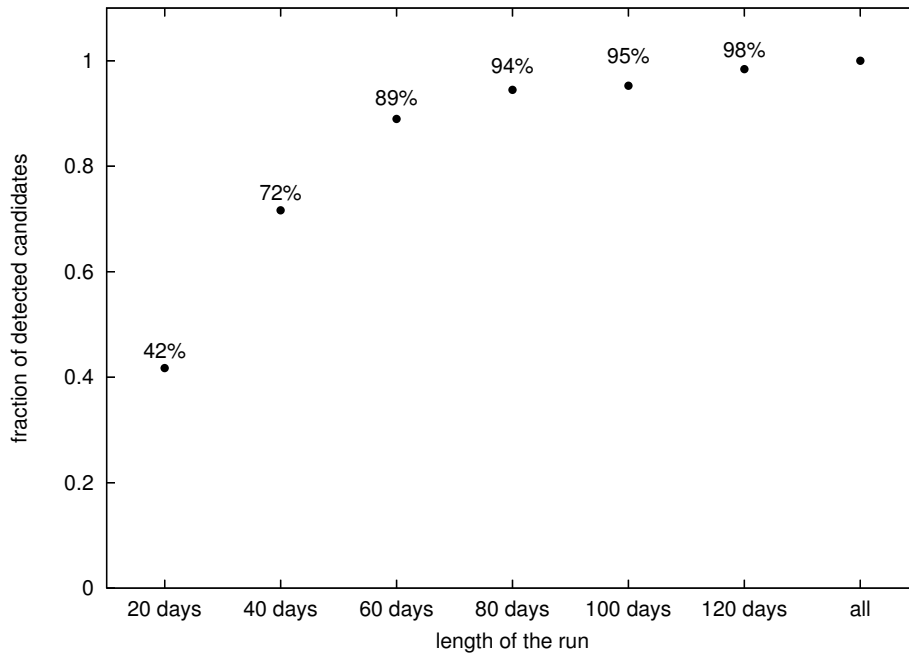
First important warning, candidates in this context are both binaries and planets. The detection algorithm finds a list of candidates and a posterior analysis, taking into account not only the significance of the signal, but also the information on the target star and the parameters of the candidate (period, duration, depth, colour), distinguish between potential planets and clear binaries.

Second important warning: this result is valid only with the automatic version of one detection algorithm. For example, in the 20 days shortened run exo2b and exo3b were not found by the automatic classification because they were slightly below the threshold. This threshold was relatively high, because we kept the values for the 153 days run (in a long run the signal to noise is better and we can put a higher threshold to avoid too many false positives). Looking at the individual light-curves, we would have found those candidates (11 transits of exo-2b and 4 transits of exo-3b were visible), so probably the values shown in the figure are lower limits. It would also be interesting to compare these results with those of different teams.

CONCLUSIONS

The main advantage of having two intermediate runs instead of one long run is the number of candidates: we should be able to almost double their number, as by observing 75 days we should be able to detect at least 90% of the potential candidates that would be found in a long run lasting 150 days.

However, it has to be understood that shortening the run we lose information: we lose precision in the characterization of the periods, so the follow-up is more difficult; we lose really long period candidates, such as LRc02_E1_0651 with a period of 95 days; we are almost losing the possibility of doing any transit timing variation analysis, because we need the longest time base line available; the oversampled fraction of the light curve for planet candidates will be shorter, so the characterization of the physical properties of the planet won't be as good as for a long run, etc.

Observing two intermediate runs we are close to double the number of candidates than if observing just a single long run; but the quality of the data will not be comparable in many aspects.

Juan Cabrera. March 2009

B Email Circulated on 06.03.2009

Dear all,

I join to this mail a cumulative plot with the candidates found as a function of the length of the run. The envelope lines are drawn just to guide your eyes, they should not be used to make any strong statement (they don't have any physical meaning).

Please note that with this kind of exercise we cannot talk about individual candidates, but just about trends: with a 20 days-long run we are likely to find just short period candidates (and some longer period candidates which are deep enough, but which will appear as monotransits); with a 80 days-long run we are likely to detect all the candidates detected in a 150 days-long run: we are just missing some dim candidates (at different periods) and some long period candidates. Take into account that in this graph there might be crowding effects, so not all the points might be visible.

My humble opinion is that the debate on the length of the run should not go in the direction of which is the smallest candidate which can be found for a given length of observation; but on the direction of the quality of the data.

Having 2 half runs will almost double the number of candidates and we won't be too much concerned by the sizes of the candidates because we are able to reach the instrumental noise limit most of the time; but the quality of the data won't be the same (i.e. exo-7b would have been found with just 20 days of observations, but I doubt that we could have done the analysis Roi and Szilard did on this planet with just 20 days of data).

Best regards

Juan