Inject-retrieve:
A study on the detectability of transits and mono-transits in PLATO simulated data

Martin Binet

June 2023

Abstract

Context: PLATO is a European Space Agency (ESA) space telescope that is set to be launched in 2026. It will look at hundreds of thousands of stars, looking for planetary transits, in the hope to discover and characterize rocky Earth-like exoplanets. To prepare for the mission, PLATO scientists must get as much insight as possible on the telescope's abilities.

Aims: We aim to estimate the probability to detect a planet with PLATO, depending on its size, its period and the apparent magnitude V of its host star, among other parameters.

Methods: PLATO scientists have simulated hundreds of lightcurves for stars that the satellite will be looking at. Our framework ”inject-retrieve” will use these simulations, artificially inject planetary transit models to them, and check whether transit search programs can recover these transits. After doing this thousands of times, for different stars and orbit parameters, we will be able to interpret our recovering statistics as detection probabilities.

Results: To evaluate these expected performances, we compare them to those of Kepler, a disused telescope which allowed the discovery of thousands of planets. The two principal improvements with PLATO are its ability to look at brighter stars, and its much shorter observing cadence (25s, compared to 30min for Kepler). Our results do witness the importance of these improvements.

1 Introduction to exoplanets

1.1 What is an exoplanet?

An exoplanet is simply a planet outside the solar system. Precisely (see [3]), it is a body which:

1- Directly orbits a star;

2- Has a mass lower than both the minimum mass for deuterium fusion, currently estimated at 13 Jupiter masses; as well as 1/25 times its host star's mass, to ensure the L4 and L5 Lagrange points are stable;

3- Is massive enough to be in hydrostatic equilibrium, which gives it a round shape

The first confirmation of an exoplanet detection, named "51 Pegasi b" was in 1995, by Swiss astronomers Michel Mayor and Didier Queloz (see [11]). For this milestone, they were awarded the Nobel Prize in 2019. As of the 23rd of August 2023, there are now 5496 confirmed exoplanets.

The name of an exoplanet is always the star's name, followed by a lower case letter, starting from b for the star's closest planet, then c, d, e, etc... if there are multiple. So Michel Mayor's planet was orbiting around 51 Pegasi, a Sun-like star: this means that its size and temperature are close to the Sun's. If the star did not have a name before the planet's discovery, it is often name after the mission which discovered it, for instance "Kepler-62", around which 5 planets where discovered, including 2 "potentially habitable planets".

1.2 The transit discovery method

There are multiple ways to discover exoplanets. We will only mention the two main ones: transits and radial velocity.

The transit method is the easiest to understand: we observe the flux of light coming from a given star, at regular time intervals (the "cadence"). This flux, on average, only varies slowly over time. But if a planet comes in front of the star from our point of view, it will block part of the star's light, resulting in a temporary flux drop. The relative depth of this drop is approximately equal to the areas ratio, which is equal to $\frac{R_{planet}}{R_{star}}$. In other words, supposing a signal is confirmed to be a planet transit's, and we know the star's radius, we then have an estimate of the planet's radius. Also, it shows that the transit method has a selective effect in favor of larger planets.

But which proportion of planets actually transit their star when seen from Earth? To answer this, we can suppose that planetary systems come in all orientations with equal probability, which is a good assumption. In that case, for a given star of radius R_{star}, and planet of radius R_{planet} (we will keep these simple notations throughout the report) on a circular orbit of radius a, the transit probability is: $\frac{R_{star}+R_{planet}}{a} \approx \frac{R_{star}}{a}$ (see [4]). Thus, this generates another selection effect, in favor of planets closer to their host star.

Then, if we observe two transits of the same planet, we know that their time separation is the maximum orbital period. If we see no transit in between the two, it is then the actual period. Knowing the orbital period as well as the host star's mass also gives access to the semi-major axis a, through Kepler's third law (with the approximation $M_{planet} << M_{star}) : \frac{T^2}{a^3} = \frac{4\pi^2}{GM_{star}}$. a is approximately the average planet-star distance, or simply the distance for a near-circular orbit like Earth's. An example of a planetary transit signal is shown in 1.

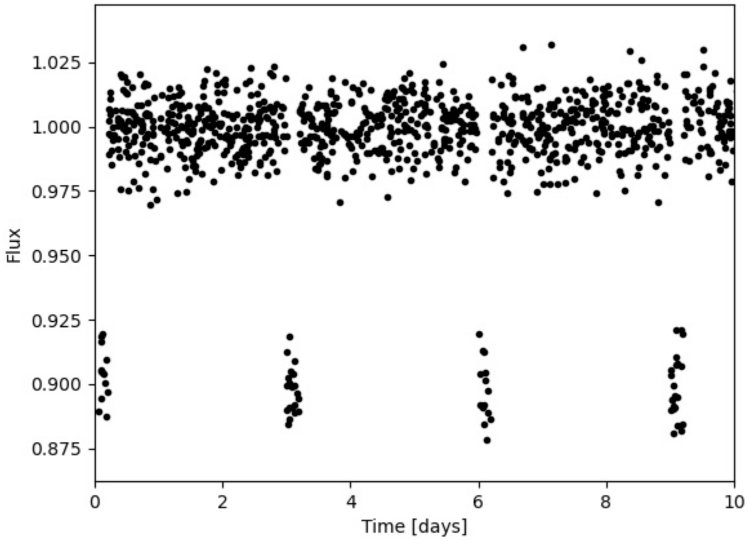


Figure 1: A typical lightcurve of a star with a transiting planet. Here the signal-to-noise ratio is very high, so it is easy to see the 4 transits, and deduce that the period is about 3 days. Also, the relative depth being 0.1, we know that the radius ratio (planet/star) is about 1/3. This is very big compared to the ratios in the solar system, with Jupiter being the largest planet, at only 1/10 of the Sun's radius. In fact, very few known exoplanets actually reach such a high ratio. 3 days is also a lot shorter than Mercury's 88 days, but actually many exoplanets have such short periods.

1.3 The radial velocity method

Apart from blocking some of its light in the case of a transit, a planet also has a an impact on a star's trajectory. Indeed, it is only approximately to true to say that a planet "orbits around its star". In reality, both objects rotate around their center of mass. Thus, the star usually being much more massive, this center of mass is often inside the star, but not in its own center, resulting in a small rotation for it too, the "reflex motion". For instance, the Sun has a reflex motion of a few meters per second (m/s), which is mostly induced by its heaviest planet, Jupiter.

Although it is theoretically possible to directly observe this motion for other stars, that process being called Astrometry, it has only been the first detection for two giant planets. But what is a lot easier to detect is the Doppler effect caused by this reflex motion. Indeed, unless it is perfectly perpendicular to us, it will have a radial velocity (RV) component, which will periodically be positive then negative. In order to measure it, it is not sufficient anymore to look at the photon flux for a given wavelength band (photometry), we must now look at the star's precise frequency distribution, using spectroscopy. By doing this, we see that the known spectral emission lines are periodically "red-shifted" (moved towards longer wavelengths) then blue-shifted (shorter wavelengths).

The radial velocity method then obviously has a strong selection effect in favour of more massive planets. The mean density of planets is variable, but this still means that we will rather find larger planets with RV method too. Furthermore, although we will not prove it here, the RV amplitude also decreases with semi-major axis, once again rather selecting planets close to their star.

We can see the result of both methods' selection effects on the current exoplanet population in 1: most known planets are very different than the solar system ones.

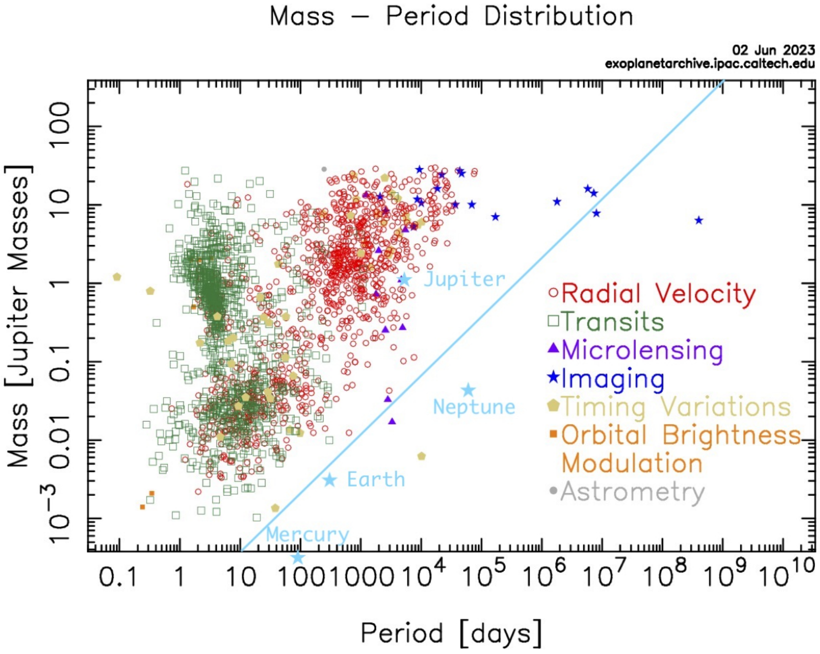


Figure 2: All known exoplanets as of 02/06/2023, plotted as a function of mass and orbital period (taken from [2]). Adding some of our solar planets, we clearly see that apart from Jupiter, they are far from the most frequent parameters. We have also drawn a straight line of slope 3/4 in log space, under which there are very few exoplanets have been detected, none of which with transit or RV methods.

1.4 Potentially habitable exoplanets

One of the most intriguing question of astronomy, is whether there is extraterrestrial life somewhere in the Universe. Because we have never detected a direct trace of such life, the best we can do is look for a planet which could potentially host life. For this, the two most basic properties are that it must be a rocky planet, and have the physical conditions which allow the presence of liquid water.

We know that most planets heavier than 10 Earth masses, or than larger than about 2.5 Earth radii, do not have solid surfaces, but rather thick layers of hydrogen, helium, and possibly ice. Although they can have different types of compositions, mostly dependant on their size and mass, they are all called "giant planets". This category includes Neptune, Uranus, Jupiter and Saturn, and can obviously be discarded in the search for habitable worlds.

Then, for the presence of liquid water, the key parameters are surface temperature and pressure. Indeed, water can only exist with a pressure higher than 0.01atm (1 atm being the sea-level pressure on Earth), as well as a temperature between 0°C and 100°C (the boiling temperature actually increases with pressure, so this bound can often be lower). A significant pressure implies the presence of an atmosphere, which this time eliminates most planets lighter than 0.2 masses, including Mars and Mercury. That is because their lower surface gravity allows most elements like Oxygen or Nitrogen to escape in space.

Because a planet's primary source of energy is its star, the temperature bounds come down to bounds on the amount of stellar light received. For a given star, because its light is equally emitted in all directions, this only depends on the planet's orbital distance (semi-major axis). This yields what is called a "habitable zone" (HZ). For instance, although its estimations vary a lot, the Sun's HZ goes approximately from 0.7 to 2 AU. But for fainter stars, the HZ can be a lot closer, like for Trappist-1. Indeed, this ultra cool star is known for hosting 7 exoplanets, including 4 in its habitable zone, even though these are located only 0.02 to 0.05 AU away from the star.

2 The PLATO mission

The PLATO telescope is set to be launched in the end of 2026, and its mission will last at least 4 years, and up to a maximum of 8.5 years (this upper limit comes from the amount of fuel it will carry).

2.1 The objectives

PLATO has two main scientific objectives: planetary science and asteroseismology (see [1] and [12]). The latter consists in studying oscillations in stellar lightcurves, which allows to probe stars' internal structures, as well as helping to determine their masses, radii and ages. All this information on a given star is then very useful to better characterize exoplanets found orbiting around it.

In this report, we will focus on planetary science. For this part, the mission's main goal is to find Earth-sized planets orbiting in the habitable zone of Sun-like stars. In the following, we will simply call these "Earth-like planets".

In the PLATO mission status report, Earth-sized is defined to be between 0.5 and 1.5 Earth radii ; and Sun-like is between spectral types F5 and K7, the Sun being of type G2.

There have been multiple attempts to predict how many such planets the mission will discover. The results are quite variable, mostly because of one very poorly known factor: the true planet occurrence. Indeed, because of the strong selection effects for planets of this type, it is very hard to estimate how many such planets actually exist on average per star. The estimations can vary from about 0.1 to 1.

But even with this high variability factor, the studies show that PLATO should find between about ten to a few tens of Earth-like planets. This may sound very little, but would actually be a lot compared to current numbers. Indeed, at the moment we only know 13 transiting planets within the above radius range and their star's habitable zone. Most of these were discovered by the Kepler telescope (early 2010s), which we will mention again later. If we add the Sun-like star condition (F5 to K7), this number drops down to 2: Kepler-62f and Kepler-442b.

It is also not surprising to note that both are really at the "easiest to detect" limit of our Earth-like definition: about 1.4 Earth radii, with stars smaller than 0.7 Sun radii. According to 1.2, the transit signal observed is then 4 times stronger than for the Earth around the Sun. In other words, PLATO will look for planets that previous telescopes would not have been able to detect.

2.2 The telescope

So what will allow PLATO to perform better than previous missions? To answer this, we first need to describe the telescope's main characteristics. It will carry 26 cameras, 2 of which are "fast cameras", with a cadence of 2.5s, which will look at very bright stars (magnitude 4 to 8), and work as guidance sensors.

The planetary science will come from the 24 "normal cameras", which have a cadence of 25 seconds. These 24 cameras will be split into 4 groups of 6. Within each group, all 6 cameras will be synchronised and will look at exactly the same field of view. But the different groups will be offsetted in time (6.25s between each), and most importantly in terms of sub-field of view (see 3).

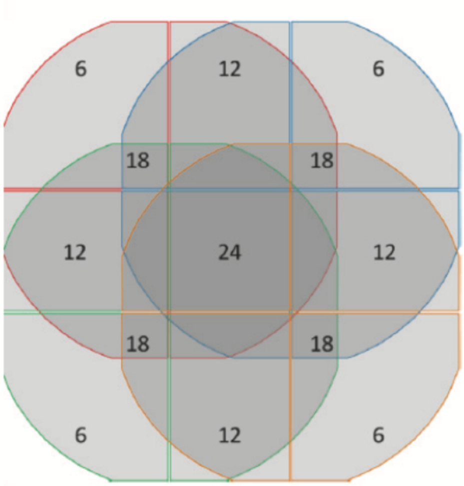


Figure 3: Schematic view of a complete PLATO field of view. The colored lines show the limits of each sub-field (for one group of 6 cameras). The shades of grey indicate the total number of cameras observing each zone of the field, with a minimum of 6 (1 group) in the corners, and a maximum of 24 (all 4 groups) in the center of the field.

Two fields of view, as represented in 3, have already been defined for PLATO: one in the Southern hemisphere, the other in the Northern. It is almost certain that the satellite will first look at the Southern field for 2 years. Then, nothing is decided yet, but a likely outcome is that it will then switch to the Northern field for the 2 remaining years of its primary mission duration.

The most natural comparison for PLATO is with the Kepler telescope. It has discovered more than 2500 transiting planets, which is about two thirds of all planets discovered with the transit method. For PLATO, the currently lowest estimation of the total number of planet discoveries is 4600 (see [1]). In 2023, the telescope discovering the most planets is TESS, but its technical abilities and mission design do not allow the detection of Earth-like planets.

Even when all 24 PLATO cameras are observing a star, the total light collecting area is only about the same as for the Kepler satellite. So that is not where PLATO will be better: the average noise in the data will not be any smaller.

The first main advance has already been mentioned: the 25s cadence. Indeed, the Kepler telescope had a much longer cadence of 30 minutes, which means PLATO will have 72 times as many data points per unit of time. This will allow the signal-to-noise ratio to be much higher, because it is proportional to the square root of the number of in-transit data points (see 5.5).

The second main difference is the brightness of the stars observed. Kepler did look at some stars with $V < 11$, but most of its input catalogue was above V = 11, because its detector would saturate with bright stars. For instance, the two Earth-like planets we mentioned earlier were around stars fainter than V = 14. Whereas PLATO's main input sample (”P1”) will all be stars between magnitudes 8 and 11. Once again, that will make the signal-to-noise ratio higher, simply because the signal will be stronger (more light collected), with an instrumental noise that will not be larger (it does not depend on star brightness).

2.3 Simulated lightcurves

In order to estimate the mission's performance, PLATO scientists have produced (and are constantly improving) simulated lightcurves. This means, for a given star in the input catalogue, the type of lightcurve that we expect to get, taking into account the telescope's precise technical specifications. When doing this, they first suppose that no planet is transiting the star.

The lightcurves we will use were produced from 2021 onward, with the software PLATOSim (see [7]). They include 1000 stars, which are supposed to be representative of the whole P1 sample they are taken from. That is why magnitudes that range from 8.5 to 11, and there are stars observed by 6, 12, 18 and 24 cameras.

They have only been simulated for a duration of 6 months, which corresponds to two ”quarters” of the mission (every 3 months the satellite must be re-calibrated, which corresponds to a quarter). The reason for this is computing power: with a cadence of 25s, these 6 months already amount to more than 600k data points!

3 Inject-retrieve: general information

Inject-retrieve is the name of the Python framework we have created. It is an implementation of the general "inject and recover" method (the name differs on purpose, in order to avoid confusion). All other Python frameworks we will mention are packages, which were not created for this project: we simply use them as they are.

The project is divided into two sub-projects, which will be talked about in detail in section 4 for the first, and section 5 for the second.

3.1 The inject and recover method

With the PLATO lightcurves in hand, we can then inject planetary transit models. In other words, produce a new lightcurve "supposing the star has a yet unknown transiting planet, of this specific size and period". Finally, we can check whether a transit search algorithm will recover this injected transit, to answer the question: "if there is actually a planet with such characteristics, will PLATO find it?". This type of study is called "inject and recover", and is exactly the purpose of this project.

Precisely, the steps of a single iteration of this method, in our inject-retrieve framework, are:

- Select a random star (see 3.3) and random orbital parameters (see 3.4);
- Load the selected star's raw lightcurve;
- Create a transit model with the selected star radius and orbital parameters, using the Batman package (see [8]);
- Inject the transit, by a simple Numpy multiplication: star lightcurve $*$ transit model (the latter is already normalized to 1, so the scale of the star lightcurve will not change);
- In the resulting lightcurve, remove the long-term trend (in other words, make it flat) with the .flatten method from the Lighkurve package (see [9]). This will also normalize the curve to 1. Without this step, the transit search which follows simply cannot work;
- Then remove outliers: data points that are further away from 1 than N standard deviations (we take N = 6). This step could be ignored, but the results would be worse;
- Perform a transit search (see 3.2);
- Write down in a new csv row all injected and recovered parameters for this iteration, the latter ones being the outputs of the transit search.

Then, the goal is to repeat the process thousands of times with different host stars and planet characteristics. As a result, the output is a csv file, with each row corresponding to an iteration. And when turned into a Pandas dataframe, each column then corresponds to either an injected parameter, or a recovered one.

From these csv files, we can extract recovery statistics, which we can interpret as detection probabilities. We will detail how this is done in 4 and 5 because it is different for each sub-project.

3.2 Transit search algorithms

In the project, we use two different transit search algorithms, depending on how many transits are injected in a given iteration of the program:

- If there are at least two transits, we call it a ”multi-transit” injection. In this case, we want to find the period of the signal. For this, we use the ”Box least squares” (BLS) algorithm implemented in the Lightkurve package. This algorithm takes as input a detrended (flattened) lightcurve, and looks for the best fit of a ”box signal” (see 4). BLS can take multiple arguments including the period range, but we do not use the latter, in order to mimic the situation where we are looking for a transit but do not have any constrain yet. The one argument we do use is the ”frequency factor”, which controls how precisely we want to look for the period. Detection rates are obviously better when it is more precise, but on the other hand the search becomes slower.

- If there is only one transit, it is a ”mono-transit”. This time, we use a brand new algorithm developed by PLATO scientists (not yet published), specifically for this purpose. It looks for a more precise fitting of the transit shape (which never looks like a simple box), and uses the GPU instead of the standard CPU. This allows the algorithm to be extremely fast, but still having very good detection rates. The authors have not given it a specific name yet, so we will simply call it ”GPU transit search” in the following.

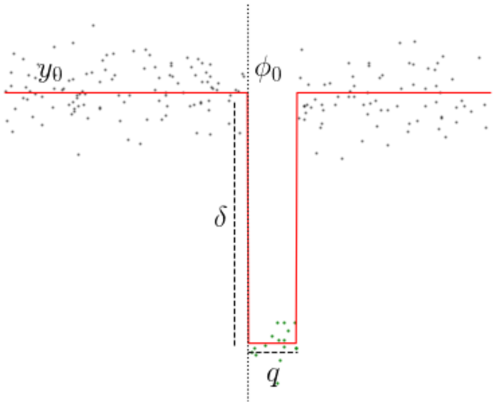


Figure 4: Example of a BLS fitting. In our case, y_0 is always 1 (the curves are normalized), δ is roughly the transit depth, q is the transit duration, and $\phi_0 + \frac{q}{2}$ is the time of mid-transit.

3.3 Star selection

For the purpose of this project, we will restrain the star selection a lot more than the F5-K7 spectral band. Indeed, as we have said earlier, this band includes stars as small as 0.7 Sun radii, whereas we will only keep those between 0.95 and 1.05 Sun radii, as well the same range for stellar mass. Although spectral type is induced by temperature, which is not directly related to mass and radius, there are scaling laws which ensure that this range is indeed well included in the F5-K7 band.

This yields only 33 Sun-like stars, out of the 1000 in the database. There is no filter on the number of cameras observing each star, meaning this selection is not necessarily representative of the whole P1 sample for that characteristic anymore. From these 33, there will then be a sub-selection specific the second part of the project (see 5).

3.4 Orbital parameters selection

We only consider circular orbits in this project. All orbital parameters are chosen randomly in a given range, specific to each sub-project:

- The orbital period (P), which directly constrains the semi-major axis;

- The planet radius (R_p), which will determine the relative depth of the transit signal (see 1.2);

- The time of mid-transit (t_0). With all lightcurves starting at t = 0, the first transit will occur when $t = t_0$, the second (if there is one) at $t = t_0 + P$... By "will occur", we mean the moment where the planet is closest to the centre of the star from our point of view, which is also halfway through the transit duration (hence the name "mid-transit").

- The impact parameter (b), which measures the geometric offset between the planet transit and the centre of the star. Depending on b, the transit (if there is one) can be full or partial (see 5); but in our analysis, we will only consider full transits.

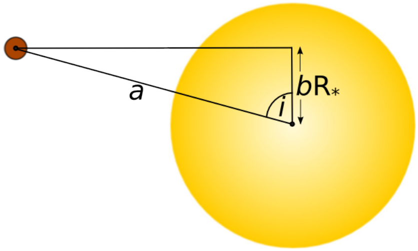


Figure 5: The definition of the impact parameter "b". The diagram (taken from [14]) is for when $t = t_0$, with the observer being located on the left. At this moment, the planet is offsetted by a distance bR_* from the star's centre. As a consequence, if we define $r = \frac{R_p}{R_*}$, the planet fully transits the star if $b + r \leq 1$; if $b - r < 1 < b + r$ only part of the planet will transit; and if $b - r \geq 1$ there will be no transit (which is by far the most likely case, see 1.2).

4 Studying multi-transits

As was said in 1.2, in order to truly confirm the presence of an exoplanet, and determine its orbital period, we need to observe multiple transits. So in the first part of the project, we injected transit models with periods short enough so that planets will transit at least twice in during the 6 months of data we have.

In fact, we took a very wide range of periods, to measure how much the detection rate is improved when there are more transits observed. We have also taken a wide range of planet radii, to see the importance of that variable too.

We will then compare the results with what we would get for true lightcurves from the Kepler mission.

4.1 Parameter selection laws

Here we detail the random selection laws for each iteration of the program in this sub-project.

- Host star lightcurve: uniformly chosen in the selection mentioned in 3.3. For the analysis, we will then split them into two groups depending on their apparent magnitude: bright stars between 8.5 and 10, and faint stars between 10 and 11;

- Period P: log-uniform law, from 0.5 to 1000 days. This yields multi-transits but also mono-transits, which allows to have a first insight on how harder it is to find a mono-transit;

- Planet radius R_p: log-uniform law from 0.005 to 0.15 R_{Sun} (approximately half-Earth to 1.5 Jupiters);

- Time of mid-transit t_0: uniform from 0 to P;

- Impact parameter: uniform from 0 to 1. This yields both full and partial transits, but as it was said in 3.4, only full transits are kept.

4.2 Binned lightcurves

As it was said in 2.3, the raw lightcurves produced by the PLATOSim software have more than 600k data points. Even if we only took one quarter of data (3 months, 300k points), the BLS transit search would still be way too slow to be able to compute thousands of samples in context of this project. Therefore, we need to bin the lightcurves. This means, reducing the number of data points by a factor N: for that, we split them into groups of N consecutive points, then take the mean flux in each group as a data point for the binned lightcurve.

In our case, we try two values of N: 4 and 12. They yield lightcurves of respective cadences 100 and 300 seconds (5min). We run our inject-retrieve code with both settings, in order to show the impact of reducing the cadence (which we expect to higher the detection rate, because binning is losing information).

4.3 Kepler lightcurves for comparison

For this sub-project, we found that it was interesting to compare our results for PLATO simulated lightcurves with the ones we would get for true Kepler lightcurves. To do this, we have fetched the said curves through the Python Lightkurve package, and used the same Sun-like criteria as for PLATO (between 0.95 and 1.05 Sun's mass and radius).

We have also filtered in terms of magnitude, to keep only the stars in the "faint" range we had for PLATO: $10 < V < 11$.

Using brighter stars might have been possible, but not very useful because they are so rare in the Kepler catalogue that they are not representative of the mission. In fact, Kepler has only discovered 14 planets (in 8 different systems) orbiting around stars with $V < 10$. This is only 0.6% of its 2500 planet discoveries. It is also very little compared to the predictions for PLATO discoveries in that range, which are at least a several hundreds with the currently planned observing strategy.

Finally, the last thing was to check that we only took stars for which we had not found a planet, nor even a false positive. That is to ensure a fair comparison with the PLATO simulations. All these filters combined yield 27 different stars, for which we also do a random uniform selection at each iteration.

For a side-to-side comparison, we use only 2 quarters of Kepler data, even though all 18 quarters (4.5 years, the duration of the first mission) are available online.

The observation we have made about the amount of data points in PLATO simulations (see 4.2) does not apply here, because of the 30 minutes cadence. Indeed, it leads to a total of only about 8000 data points in the 2 quarters. So there is no need to bin the lightcurves.

Also, for technical reasons we have not been able to run the GPU code with Kepler data, so we will only look at multi-transits.

As for orbital parameters selections, they are kept the same. Whenever an iteration selects parameters that lead to only a single transit (for instance, if the period is larger than 6 months), we ignore it and go straight to the next iteration.

4.4 Results

In this sub-project, we consider that the injected transit has been correctly retrieved if and only if:
- The absolute error on t_0 is smaller than 0.15 days;
- The relative error on the period is smaller than 1% (only applicable for multi-transits).

In that case, we label it "true positive".

Then, we group our injections into 11 radius ranges (the radii are expressed as fractions of R_{Sun}) and 11 number of transits ranges (or 12 for PLATO simulations, where we add the mono-transit column). As it is shown on the heatmaps made with the Seaborn package (see [15]), the ranges are defined to make a log scale. For each range of radius and number of transits, the corresponding number is the percentage of correct recoveries among all injected transits in that range, or the "True Positive Rate" (TPR, which we had called "detection rate" up to now). Each heatmap is made with about 50k samples in total, which means each TPR is a statistic from about $\frac{50{,}000}{11*11} \approx 400$ samples.

It is important to note that all other random parameters are hidden, including the host star. In the case of PLATO lightcurves, this also implies the number of cameras observing a star, which is not always 24 in our selection. We said in 2.2 that PLATO matched Kepler in terms of total collecting area with its 24 cameras combined. This means that on average, the collecting area is smaller, so the relative noise is larger.

In the following figures, we will show the results for:
- Kepler lightcurves (faint stars, 6)
- PLATO simulations, faint stars, 5min cadence (7a)
- Difference Kepler - PLATO 5min (7b)
- Difference PLATO 100s - PLATO 5min (still faint stars, 8a)
- Difference PLATO bright stars - faint (both 5min cadence, 8b)

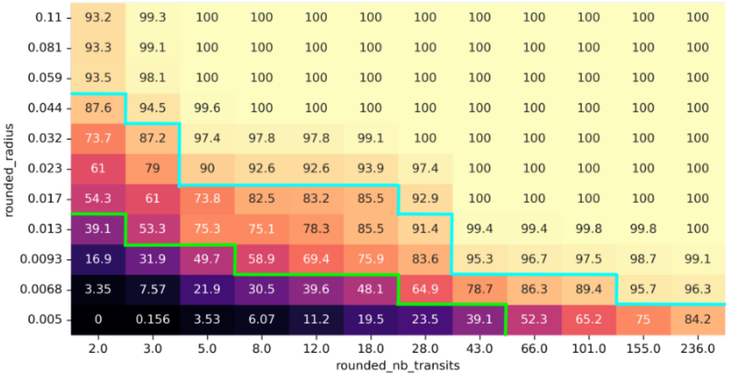
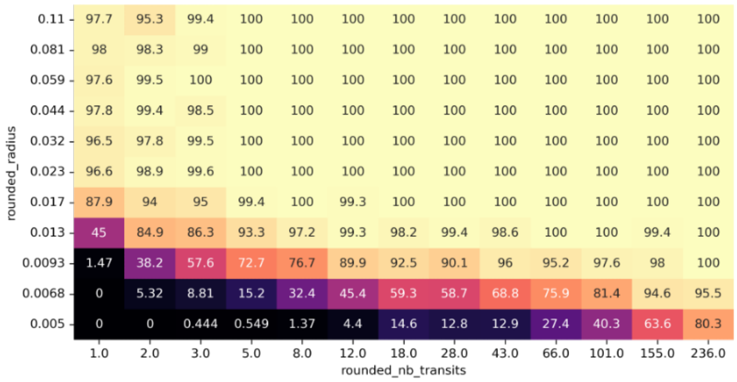
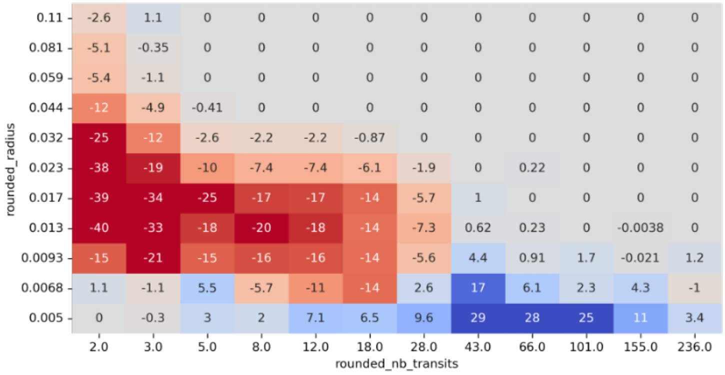


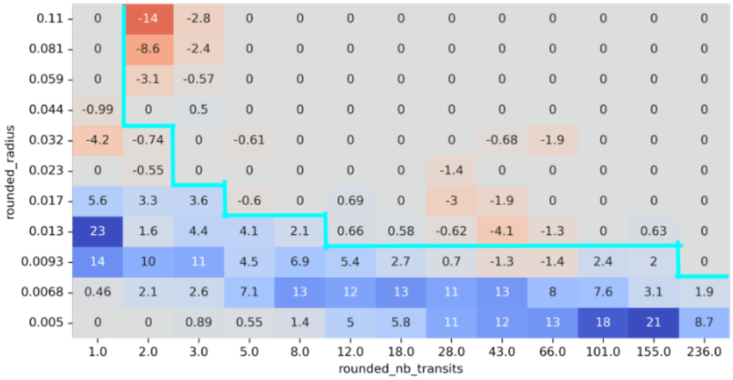
Figure 6: TPRs for Kepler lightcurves of Sun-like stars, with $10 < V < 11$, as a function of radius and number of transits. The values on the x and y axes are the minimum of the range. For instance, if we start from the bottom left, the 75.3% TPR shown on the fourth column and third row corresponds to injected signals with 5 to 7 transits, of planets with radii between 0.013 and 0.017 R_{Sun} (super-Earths). The green and blue lines correspond to the zones above which we have TPRs higher than 50% and 90% respectively. Two observations are comforting: first, the detection rate is an increasing function of both planet radius and number of transits. Secondly, for signals that are sufficiently "easy" (large radius and number of transits), we consistently get TPRs of 100%. With three significant digits, because there are about 400 injected transits per box, we know for sure that there were only true positives.



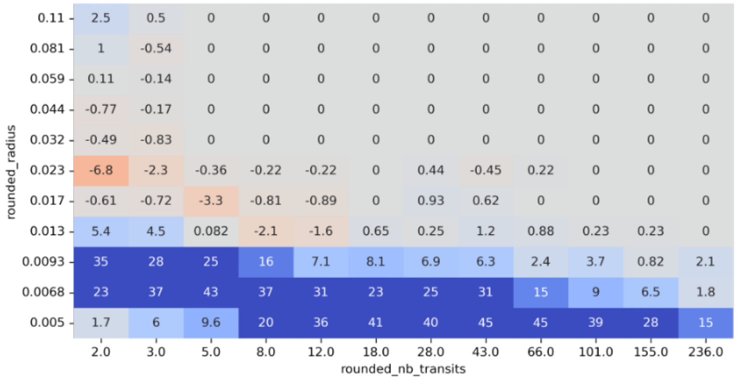
(a) TPRs for PLATO simulated lightcurves of faint ($10 < V < 11$) Sun-like stars, binned to a 5min cadence. The previous observation is still valid: TPRs increase with radius and number of transits. This holds even when comparing mono-transits to two-transit signals (first and second columns).



(b) TPR difference between the two previous heatmaps: Kepler - PLATO 5min cadence. We can clearly see a large parameter range where Kepler is outperformed by PLATO: from 2 to about 40 transits (periods between 5 and 100 days), and from 0.007 to 0.05 R_{Sun} (sub-Earth to super-Neptune). Out of the two technical improvements we have pointed out regarding PLATO, its ability to look at brighter stars does not play a role here because both heatmaps were made with results for stars with $10 < V < 11$. Therefore, we can assume the difference observed is thanks to the shorter cadence (5 minutes instead of 30). Then, we also see a range where Kepler perform better: sub-Earths (the Earth's radius is 0.0092 R_{Sun} with very short periods (more than 40 transits, so less than 5 days). This is not so surprising after the previous observation regarding the higher relative noise for PLATO: it struggles more to find small planets because the signal-to-noise ratio becomes very small.



(a) TPR difference induced by the shorter cadence: PLATO 100s - PLATO 5min. Both are for faint stars. It is really interesting to see that reducing the cadence does indeed improve transit detectability, especially in the intermediate parameter ranges, where the effiency is from about 10 to 50%. This means that in our previous comparison between Kepler and PLATO (7b), we have actually underestimated PLATO's abilities, by using 5min cadence instead of the true 25s (which we would expect to be even better than 100s). Also, we have shown again the blue line above which there is more than 90% TPRs. In those ranges, the differences observed (this time lightly in favor of the longer cadence) are most probably due to algorithm errors, like finding half or double the true period. But it does not really matter because these transits (giant planets and/or very short periods) are easy to see with the naked eye when looking at the lightcurves, so we could just consider the TPRs to be 100% in all cases.



(b) TPR difference between bright ($8.5 < V < 10$) and faint stars (($10 < V < 11$), both with 5min cadence. Once again, we get what we expected: a better detection rate for brighter stars. The difference is actually much larger than in the previous comparison. And we can say that figure 7b also underestimates PLATO's improvements because it compares TPRs for equal brightness, whereas on average, the new telescope will look at much brighter stars than Kepler.

5 Focus on mono-transits of Earth-like planets

Now that we have looked at the detectability of multi-transits in PLATO simulations, and compared it with the same study for Kepler data, we are going to focus on mono-transits for our second sub-project. Also, we will only look at the transits of Earth-like planets (see 2.1).

5.1 Why mono-transits?

In other words, the question is why would we study them, if they can hardly confirm and/or characterize exoplanets anyway (see 1.2)?

To answer this, we need to keep in mind two things. First, PLATO will most likely look at the same field without interruption (therefore the same stars) for a maximum of 2 years (see 2.2). Secondly, although it will represent a very small fraction of all discoveries, PLATO's main planetary science goal is to find Earth-like planets as defined in 2.1. Being located in the habitable zone implies their orbital periods will range from a few months for the closest planets orbiting the dimmest stars, to a few years for the furthest planets orbiting the lightest stars. As a consequence, on the one hand the former planets will be able to transit multiple times in a two-years window, but on the other hand, the latter will only have one transit, if any.

If we take the example of a planet in the Southern field, with a period of exactly two years, we know it should transit exactly once (supposing the orbit is aligned with Earth, see 1.2) in the first two years looking at the Southern field. But if PLATO switches to the Northern field for the following two, it will miss the next transit. What can be done though, is a "follow-up" by ground facilities. This means that if PLATO has detected single dip in a star's lightcurve, which seems to be due to a planetary transit, ground telescopes can then look at that star, and try to confirm (or deny) the planet's existence by looking for a second transit and/or a radial velocity signal (see 1.3).

In order to do this, we must have a good detection rate for mono-transits (as shown in 4.4, it is always lower than for multi-transits). But also, mono-transit search algorithms must not output too many false positives (see 5.5), because the ground facilities are limited, so they can only observe so many stars.

5.2 New choice of stars

This time, we will select host stars more thoroughly. We will choose 9 from our previous selection, each having a different combination of number of cameras and "rounded magnitude".

By rounded magnitude, we mean apparent magnitude rounded to the nearest integer, which is either 9, 10 or 11. For each of these values, we will select three stars, in order to have one observed by 6 cameras, another by 12, and the third by 24.

This will allow us to compute TPRs for each of these 9 settings, and compare them, to see how much of an impact each characteristic has. Also, using the statistical distribution of the whole PLATO stellar catalogue as reminded in René Heller's 2022 paper (see [5]), we will be able to give our own planet yield estimation, supposing no planet transits more than once (see 5.6). Because of this assumption in our computation, we expect to get a pessimistic prediction. In fact, we can then compare this prediction with the one in Heller's paper.

On a side note, we now only look at lightcurves in one of their two quarters (three months), because most stars in our previous selection are not observed by the same number of cameras in each quarter. Luckily, this allows to reduce the computing time per iteration, which is now mostly due to the flattening/detrending step, rather than the quick transit search.

5.3 Using raw lightcurves

As we said in 3.2, because we are looking at mono-transits, we are now using exclusively the GPU transit search algorithm, which is much faster. Luckily, this holds even for raw PLATO lightcurves, with 25s cadence (that is actually not so surprising, knowing the code was designed for PLATO).

After observing this, the first thing we did was a side-to-side comparison of the mono-transit TPRs that were shown in 4.4 for 100s cadence lightcurves, with the ones we would now get with 25s cadence. The results are in A, and they do witness again a new improvement, especially for Earth-size planets.

5.4 Parameter selection laws

This sub-section is the analog of 4.1, defining how we select each iteration's parameters.

- Host star lightcurve: see 5.2

- Period P: uniform law, from 340 to 360 days (like the Earth, with a little variability). Even if there is only one transit anyway, this parameter matters because the transit duration $\Delta t_{transit}$ directly depends on the period, with a scaling law in $P^{\frac{1}{3}}$. For these periods close to a year, transits last about 10 to 12 hours, whereas for short periods of a few days, they can be as quick as 1 to 2 hours. And naturally, for an equal number of transits, detection rates increase with transit duration;

- Planet radius R_p: uniform law from 0.5 to 1.5 R_{Earth} (the "Earth-like" boundaries);

- Time of mid-transit t_0: uniform from 0 to 90 days (anywhere in the chosen quarter).

- Impact parameter: uniform from 0 to 0.2. This only yields full transits.

5.5 Detection thresholds

In this new sub-project, we are more restrictive for a recovery to be considered a true positive. The problem with the previous choices was that without explicitly saying it, we considered all recoveries to be "positives". Then we compared with the injected parameters, to see whether it was a true or false positive.

But as we said in 5.1, false positives must be avoided as much as possible. For this reason, we define two other thresholds, under which a recovery is considered negative: signal-to-noise ratio (SNR) and signal detection efficiency (SDE).

The former is defined as $SNR = \frac{\delta}{\sigma} * \sqrt{n}$. δ is the recovered depth, σ is the lightcurve's standard deviation, and n is the number of in-transit points. Both σ and n are taken on the final lightcurve, meaning after it was flattened and cleared from outliers (see 3.1). For n, this changes very little from the raw curve, meaning we approximately have $n = N_{transits} * \frac{\Delta t_{transit}}{\Delta t_{cadence}}$ (in our case, $N_{transits} = 1$ and $\Delta t_{cadence} = 25s$).

The SDE, on the other hand, is a number directly outputted by the transit search program, which we save in each csv row. We will not explain in detail its computation, but it measures how well the transit model recovered fits the data. This figure is not directly comparable from one transit search algorithm to the other, for instance between this GPU code and Transit Least Squares (TLS, an improvement of BLS, see [6]) which is used in Heller's paper.

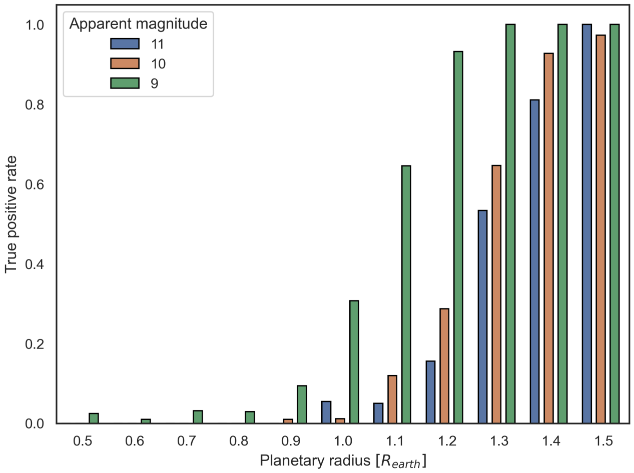
We have decided to take an SNR threshold of 6, instead of 7 in that paper, which we justify by the fact that we are looking for mono-transits instead of double/triple transits. Indeed, this divides our recovered SNR by a factor $\sqrt{2}$ or $\sqrt{3}$ respectively, meaning we would hardly have true positive of sub-earths if we kept 7. However, we set our SDE threshold to 16, which allows to have no false positive in our 10k iterations.

Finally, the time of mid-transit (t_0) error threshold has actually been loosened, it is now 0.25 days.

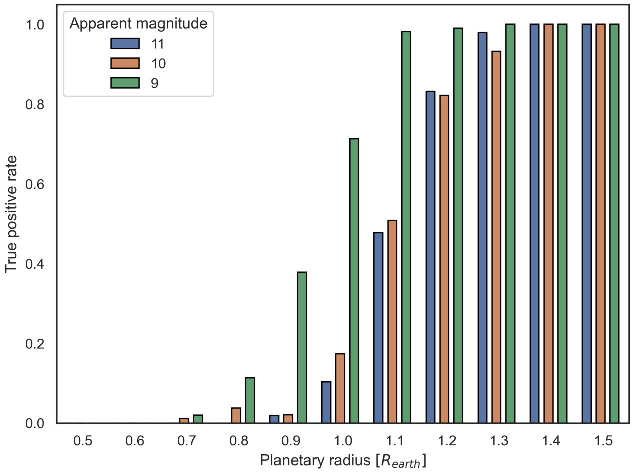
5.6 Results

As we said in 5.4, we are now using planets from 0.5 to 1.5 R_{Earth}. For the analysis, we will bin these radii to the closest tenth of R_{Earth} (0.5, 0.6... 1.5). The graphs 9a, 9b and 10a each correspond to a camera count, and there are different bars for each star brightness. Then, the fourth graph (10b) is a reproduction of the results in Heller's paper, for two transits of Earth-like planets, on stars observed by 24 cameras. It is not directly comparable because a different software was used to simulate PLATO lightcurves: PSLS (see [13]). Also, it was also made for a more restrictive range of radii (and under 0.8 R_{Earth} it finds no true positive, so TPRs are zero). Finally, the fifth figure (11) is the planet yield estimations we mentioned in 5.2.

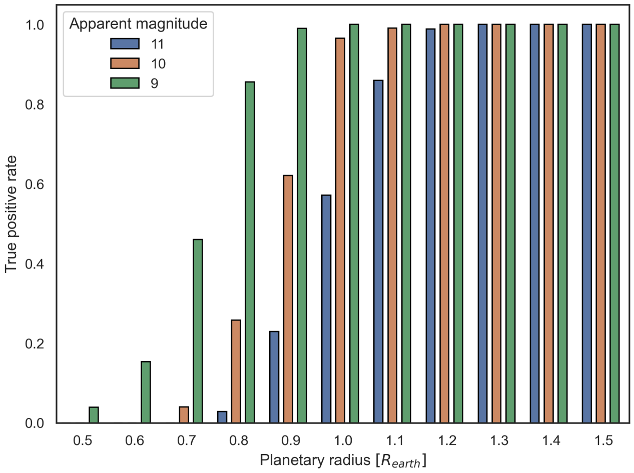
On a side note, this time the TPRs are not expressed as percentages but decimal numbers (1 corresponding to 100%...).



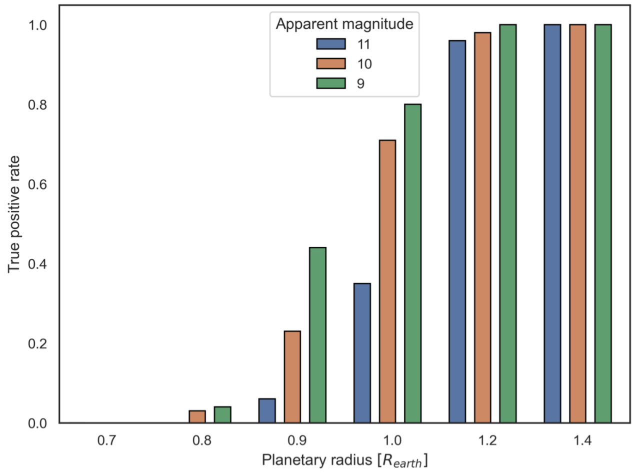
(a) TPRs of Earth-like mono-transits, for Sun-like stars observed with 6 cameras, depending on planet radius and star magnitude. It is comforting to see that TPRs increase both with star brightness (decreases with magnitude) and planet radius, apart from a few exceptions. In fact, the non-zero TPRs we get for very small planets (0.5 to 0.7 R_{Earth}) orbiting the magnitude 9 star, are most likely false positives due to dips in the raw PLATOSim lightcurves: in that case they are not caused by a planet. They might have been labeled as true positives if the injected transit happens to fall exactly at that place in the lightcurve.



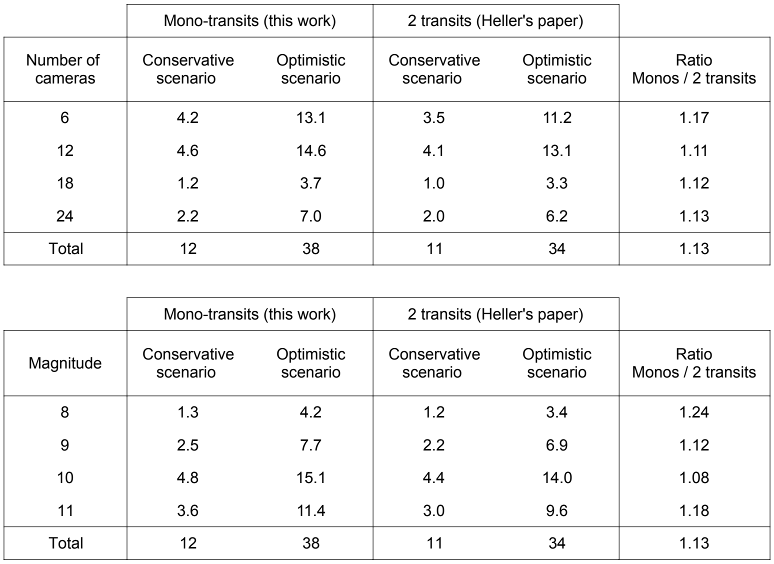
(b) TPRs for stars observed with 12 cameras, depending on planet radius and star rounded magnitude. They are still increasing with star brightness and planet size, but we do not get the surprising true positives for the smallest planets, that we have in the 6 cameras case. This not contradictory with the hypothesis we made, if we keep in mind that we are using different stars here, so there might just not be any ”planetary-like” dip in these lightcurves. On the other hand, as expected all TPRs starting at $0.8R_{Earth}$ are better here than in the 6 cameras case.



(a) TPRs for stars observed with 24 cameras, depending on planet radius and star rounded magnitude. Once again, they increase with star brightness and planet size, and are higher than in the 12 cameras case.



(b) TPRs for two transits of Earth-like planets, observed by 24 cameras, taken from Heller's paper [5]. Most parameters are very close to ours: the orbital period is fixed at 364 days, and the impact parameter is 0. The host stars are not the same but they are chosen the same way: G2 (Sun-like) stars with the same three magnitudes. Apart from the transit search algorithm and the simulation software, the other main difference lies in the lightcurves' time domain: here they are two years long instead of three months for us. This might be the best explanation for why the TPRs are lower than in our study (see fig 10a to compare at 24 cameras) although there are two transits. But it is still impressive for the mono-transit search to be performing so well.



	Mono-transits (this work)		2 transits (Heller's paper)		
Number of cameras	Conservative scenario	Optimistic scenario	Conservative scenario	Optimistic scenario	Ratio Monos / 2 transits
6	4.2	13.1	3.5	11.2	1.17
12	4.6	14.6	4.1	13.1	1.11
18	1.2	3.7	1.0	3.3	1.12
24	2.2	7.0	2.0	6.2	1.13
Total	12	38	11	34	1.13

	Mono-transits (this work)		2 transits (Heller's paper)		
Magnitude	Conservative scenario	Optimistic scenario	Conservative scenario	Optimistic scenario	Ratio Monos / 2 transits
8	1.3	4.2	1.2	3.4	1.24
9	2.5	7.7	2.2	6.9	1.12
10	4.8	15.1	4.4	14.0	1.08
11	3.6	11.4	3.0	9.6	1.18
Total	12	38	11	34	1.13

Figure 11: PLATO planet yield estimations. They are first divided into number of cameras, then into star magnitude. The figures for 2 transits are directly taken from Heller's paper. Our figures are computed with the help of our TPRs, with interpolation for the ones we had not determined directly (18 cameras and magnitude 8). Like we said previously, we use the same star distribution as Heller's paper, as well as the number of observed transits: 28 in the "pessimistic scenario" and 88 in the optimistic one.

6 Conclusions

The results of both sub-projects are truly interesting for our understanding of how the PLATO mission will be able to perform. First, our study of multi-transits (see 4) has allowed us to validate that the inject-retrieve framework was working properly. Indeed, on all the heatmaps we have shown, we get a detection rate that varies gradually from 0% for the transits that are most difficult to see (smallest planets, longest periods) to 100% for the easiest. This is true when we work with Kepler lightcurves, but also with PLATO simulations. For the latter, we also get more detections with brighter stars, which is another confirmation that the code works well.

In terms of comparison between Kepler and PLATO, it is very interesting to see that PLATO already seems to perform better when the lightcurves are binned to a 5min cadence, for equal star brightness. Since we have shown that increasing star brightness and reducing the cadence both contribute to better TPRs, we can really say that our study is promising for the PLATO mission.

To go into more depth, it could be interesting to do the same study using this time the Transit Least Squares algorithm instead of BLS, because it has been shown to have better performances (see [6]). Furthermore, it would be more precise to get PLATO TPRs separately for each camera count, as we did in our mono-transits study, instead of the mix we made here.

As for our second sub-project, there are two main takeaways: the performance assessment of the new GPU transit search algorithm, and the possibility to estimate planet yields. For the former, we have attempted to compare it with the TLS algorithm, using the results of [5]. It is really interesting to see that we obtained higher TPRs than the ones shown in that paper, with many parameters being the same (or close to the same) in both studies: planet sizes, camera counts, star magnitudes, periods and impact parameters. But unfortunately, there are also some significant differences in the lightcurves: the software used to simulate them, their duration (2 years vs 3 months), and the exact star selection (we have only taken stars with the same properties). So the next step would be to use TLS with the same PLATOSim lightcurves we used for the GPU code, which would directly remove all these differences. We would have to use shorter periods, in order to get two transits in three or six months. This would not correspond to Earth-like planets anymore, but that is not a problem if we are simply looking for a performance comparison (we would also have to run the code again with the shorter periods for the GPU code).

Regarding planet yields of Earth-like planets, our study could be extended by using stars of different spectral types from F5 to K7 (for each camera count and magnitude), and planets of different periods in the habitable zone range (which would be different for each spectral type). In fact, this has been done in [10]. In that paper, the true planet occurrence estimations are also different from [5], and are actually more precise in the sense that they do not consider the radius distribution to be uniform between 0.5 and 1.5 R_{Earth}. Finally, another simple improvement could be to consider a uniform distribution of impact parameters from 0 to 1.

With these changes, using the framework we have build, one could compute more solid planet yield estimations, which would be really interesting. But at the moment, we can already say that our study tends to confirm (even though the main variability is the true planet occurrence which we have not looked at) the predictions of about ten to a few tens of Earth-like planets discovered by PLATO.

References

[1] ESA Science & Technology - PLATO Definition Study Report (Red Book).

[2] NASA Exoplanet Archive.

[3] International Astronomical Union | IAU. 2006.

[4] Carole A. Haswell. *Transiting exoplanets*. Cambridge University Press, Cambridge, 2010. OCLC: ocn601110236.

[5] René Heller, Jan-Vincent Harre, and Réza Samadi. Transit least-squares survey. *Astronomy & Astrophysics*, 665:A11, sep 2022.

[6] Michael Hippke and René Heller. Optimized transit detection algorithm to search for periodic transits of small planets. , 623:A39, Mar 2019.

[7] N. Jansen et al. Platosim: An end-to-end plato camera simulator for modelling high-precision space-based photometry. *Astronomy & Astrophysics*, 2023.

[8] Laura Kreidberg. batman: BAsic Transit Model cAlculatioN in Python. *Publications of the Astronomical Society of the Pacific*, 127:1161, November 2015. ADS Bibcode: 2015PASP..127.1161K.

[9] Lightkurve Collaboration, J. V. d. M. Cardoso, C. Hedges, M. Gully-Santiago, N. Saunders, A. M. Cody, T. Barclay, O. Hall, S. Sagear, E. Turtelboom, J. Zhang, A. Tzanidakis, K. Mighell, J. Coughlin, K. Bell, Z. Berta-Thompson, P. Williams, J. Dotson, and G. Barentsen. Lightkurve: Kepler and TESS time series analysis in Python. Astrophysics Source Code Library, December 2018.

[10] F. Matuszewski, N. Nettelmann, J. Cabrera, A. Börner, and H. Rauer. Estimating the number of planets that PLATO can detect. *arXiv e-prints*, page arXiv:2307.12163, July 2023.

[11] Michel Mayor and Didier Queloz. A Jupiter-mass companion to a solar-type star. *Nature*, 378(6555):355–359, November 1995.

[12] Hike Rauer et al. The plato mission. In internal PLATO community review.

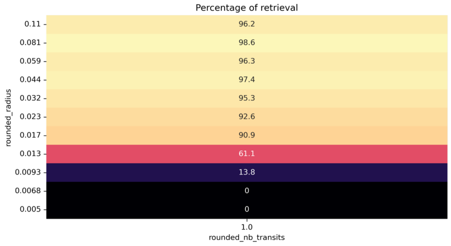
[13] R. Samadi, A. Deru, D. Reese, V. Marchiori, E. Grolleau, J. J. Green, M. Pertenais, Y. Lebreton, S. Deheuvels, B. Mosser, K. Belkacem, A. Börner, and A. M. S. Smith. The PLATO Solar-like Light-curve Simulator: A tool to generate realistic stellar light-curves with instrumental effects representative of the PLATO mission. *Astronomy & Astrophysics*, 624:A117, April 2019.

[14] Paul Strøm. The exoplanet transit method | PaulAnthonyWilson.com, August 2014.

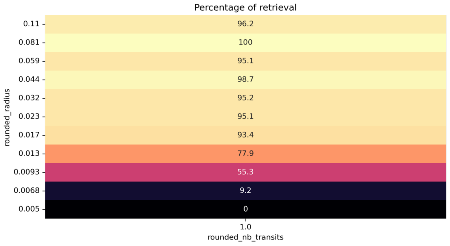
[15] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

A Cadence comparison for mono-transits

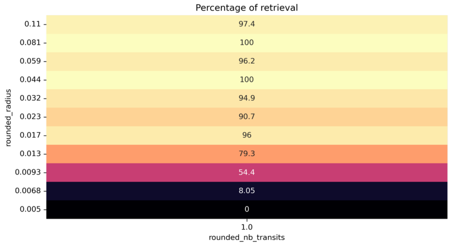
As mentioned in 5.3, here we show the impact of reducing the cadence from 100s to 25s on the detection of mono-transits in PLATO lightcurves. Although we are looking at mono-transits, we are still using the parameter selection of the first sub-project (see 4.1). That is why there are still planets of all sizes. As we can see on the figures, the TPRs increase by a significant margin, for both our faint and bright star selections.



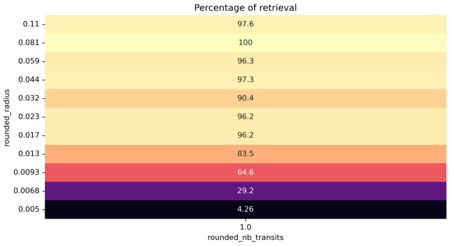
(a) Mono-transit TPRs for PLATO simulated lightcurves of faint ($10 < V < 11$) Sun-like stars, binned to a 100s cadence.



(b) Mono-transit TPRs for PLATO simulated lightcurves of faint ($10 < V < 11$) Sun-like stars, not binned (25s cadence).



(a) Mono-transit TPRs for PLATO simulated lightcurves of bright ($8.5 < V < 10$) Sun-like stars, binned to a 100s cadence.



(b) Mono-transit TPRs for PLATO simulated lightcurves of bright ($8.5 < V < 10$) Sun-like stars, not binned (25s cadence).