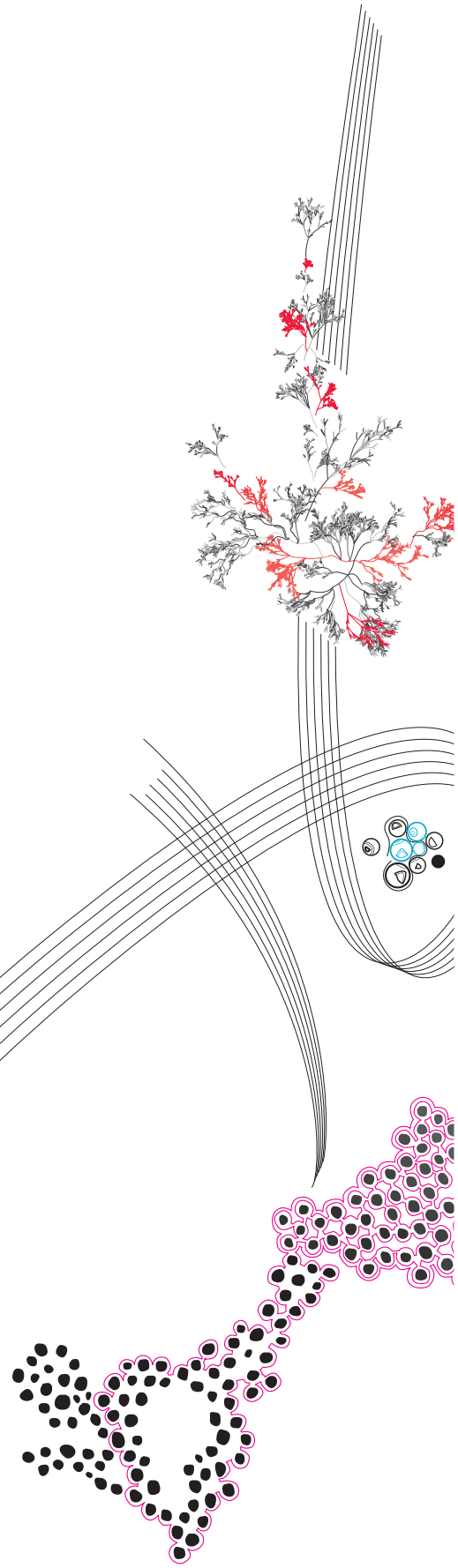


BSc Thesis Applied Mathematics

Detection of exoplanets by means of the Bump Detection Method

M. O. Karlashchuk

Supervisor: dr. K. Proksch

January, 2020

Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science

UNIVERSITY OF TWENTE.

Preface

This paper is the conclusion of the research that I have performed for the completion of my Bachelor Applied Mathematics. I hope the paper will enlighten you more about the world of Mathematics and its application in astronomy, specifically about the topic 'exoplanets'.

First of all, I would like to thank Katharina for her guidance during the past months. It has been a pleasure to work with her and I am happy that she believed in me and encouraged me during difficult moments of my research. Second, I want to thank my parents for their patience and love during the past five and a half years. Last but not least, I would like to thank my close friends for all their support.

Detection of exoplanets by means of the Bump Detection Method

Karlashchuk. O. Mariya*

January, 2020

Abstract

The detection of exoplanets can be accomplished by using the Bump Detection Method, which uses many statistical hypotheses tests simultaneously in order to determine whether or not a potential exoplanet is present, visible as a bump in a data set. In order to draw reliable conclusions from the hypotheses tests it is important that the Type I error is controlled. This can be done by determining a suitable significance level for the hypothesis tests. Simulation tests have shown that it is best to choose a significance level between 0.005 and 0.01, such that reliable results follow concerning the detection of exoplanets. These values have been used for tests with data from the NASA Kepler mission, which - after comparison with physical results - lead to the conclusion that it is possible to detect exoplanets using the Bump Detection Method.

Keywords: bump detection, hypothesis testing, Kepler mission, statistical methods, matched filter

1 Introduction

For a very long time various scientists have put much effort in searching for habitable planets outside of our Solar System, which are referred to as exoplanets. The task of finding exoplanets is rather challenging and requires very sensitive and modern equipment. In 1992 the first planets orbiting a pulsar [6] - a certain type of a neutron star - have been found in our universe by Aleksander Wolszczan and Dale Frail, 1000 light years away from the Sun in our galaxy [12]. Three years later another exoplanet orbiting a sun-like star was discovered 50.45 light years away by Michel Mayor and Didier Queloz [10]. In the subsequent years the field of exoplanet search gained more and more interest. In 2009 NASA launched the Kepler space telescope which observed a part of the Milky Way galaxy in search of exoplanets [10]. This mission has continued for approximately nine years, where 530,506 stars and 2,662 planets have been monitored [11]. From the Kepler mission data at various stages of preprocessing are publicly available.

There are multiple methods that are used for the detection of exoplanets. Some exoplanets can be found through direct imaging by telescopes, but the majority must be detected through indirect methods. The most common ways to detect exoplanets are by means of the Radial Velocity Method and the Transit Method [2]. The former method is based on measuring the Doppler velocity of the star that the exoplanet orbits, while the

*Email: m.o.karlashchuk@student.utwente.nl

latter is based on the detection of dips in the brightness of a star. This paper covers the latter method, therefore only the Transit Method will be elaborated, or - as it will be called in this paper - the Bump Detection Method, which is used for the statistical analysis of first simulated data and later on Kepler data.

The scope of this paper is to determine how a transit of an exoplanet can be detected by means of the Bump Detection Method, given the light curve of its host star. First of all, a statistical model will be provided, whereafter the Bump Detection Method will be applied based on statistical significance testing. Second, performance simulations will be investigated, before applying the Bump Detection Method on the Kepler data. To enhance the detection of exoplanets a linear filter will be applied to the data. Finally, the paper concludes with an analysis of the results retrieved from the application of the Bump Detection method in the Kepler data and recommendations for further research.

2 Statistical model for the Bump Detection method

The Bump Detection Method (BDM) [5] relies on monitoring the brightness of a star. During the observation it is possible that a planet transits in front of the observed star. The brightness of the star drops in that case. If this event occurs periodically and the reversed 'bump' has a certain size, one could speak of the presence of an exoplanet. Usually the data that is retrieved from monitoring the brightness is inverted - i.e. multiplied by -1 -, such that one would see a bump in the data instead of a drop, hence the name of this detection method. Inverting the data is possible due to identical outcomes of statistical tests.

An example of a drop in a light curve can be found in Fig. 1, which is similar to data detected by the inversed BDM.

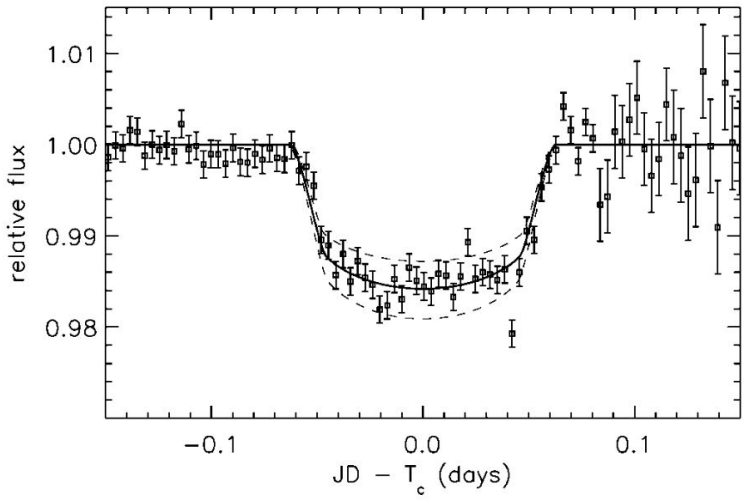


Figure 1: Example of a photometric time series of a transit curve; the series is plotted as a function of time [4].

Before the application of the BDM to the simulated data will be explained, it is necessary to state formal definitions first.

2.1 Statistical hypothesis test

To define statistical hypothesis tests properly, the definitions according to Hogg et al. [7] and Shao [13] are used:

Definition 1. Statistical hypothesis; A statistical hypothesis is an assertion about the distribution of one or more random variables. If the statistical hypothesis completely specifies the distribution, it is called a simple statistical hypothesis; if it does not, it is called a composite statistical hypothesis.

Assume a random variable X has a density function $f(x;\theta)$ with $\theta \in \Theta$, where Θ is a family of populations. Based on the observed X one tests the hypotheses $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 are disjoint sets such that $\Theta_0 \cup \Theta_1 = \Theta$. The hypotheses are usually referred to as the null-hypothesis H_0 and the alternative hypothesis H_1. The formal definition of a statistical hypothesis test is the following.

Definition 2. Statistical test; A test of a statistical hypothesis is a rule which, when experimental sample values have been obtained, leads to a decision to accept or to reject the null-hypothesis under consideration.

Whenever a hypothesis test is performed H_0 is tested against H_1. The outcome of this test is either accepting or rejecting H_0 in favour of H_1. One must keep in mind that accepting H_0 does not imply that H_0 is true; it simply means that based on the data at hand it is likely that H_0 is not wrong.

While performing a statistical hypothesis test, a certain statistic $\mathrm{T} = T(X_1, \ldots, X_n)$ is compared to a critical value c Here $X_1, \ldots, X_n$ are random variables, with n as sample size.

Definition 3. Critical value; A critical value is a point distribution that is compared to T to determine whether or not the null-hypothesis can be rejected.

Usually the critical value is determined using the significance level α, which determines the size of the critical region. Example 1 demonstrates how the critical value can be computed in a specific example.

Example 1. Assume the random variables $X_1, \ldots, X_N$ are independent and identically distributed with $X_i \sim N(\theta, \sigma^2)$ where $\theta > 0$ and $\sigma^2 > 0$ are known. Suppose the following hypotheses are determined:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

such that $\theta_1 > \theta_0$. Assume the significance level is set at $\alpha = 0.05$. Then H_0 is rejected if the test statistic $\mathrm{T} = T(X_1, \ldots, X_N) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (X_i - \theta_0) \geq c$, where c is the critical value. It is clear that $\mathrm{T} \sim N(0, \sigma^2)$. Then performing the test gives:

$$\begin{aligned} \mathbf{P}(\mathrm{T} \geq c \mid H_0) &= \alpha \\ \mathbf{P}\Big(\frac{\mathrm{T} - 0}{\sigma} \geq \frac{c - 0}{\sigma}\Big) &= \alpha \\ \mathbf{P}(Z \geq \tilde{c}) &= \alpha \\ 1 - \Phi(\tilde{c}) &= \alpha \end{aligned}$$

with $\tilde{c} = \frac{c}{\sigma}$ and Φ is the cumulative distribution function of the standard normal distribution. Using the normal distribution table [7] the value of $\tilde{c}$ can be determined, which will be $\tilde{c} = 1.645$. This test is also called the Z-test. $\triangle$

It is possible that the null-hypothesis is rejected while being true. In that case one is dealing with a Type I error - also known as a false positive. When H_0 is not rejected while H_1 is true, one could speak about a Type II error, or a false negative. The focus of this paper is mainly on the Type I error, which will be explained in the next subsection.

In general it is preferred to minimize the probability of making an error of Type I or II given a predefined significance level α. This means that one wants to increase the power of a hypothesis test while fixing α, such that it is known whether the best rejection region is chosen. The power of a test can be defined as the correct rejection of H_0, thus $\beta(\theta) = \mathbf{P}(\mathbf{T} \geq c \mid \theta_0)$. Using the Neyman-Pearson lemma it can be determined whether a test is the most powerful test at a significance level α. The Neyman-Pearson lemma is stated as follows.

Theorem 1. Fundamental lemma of Neyman-Pearson; Consider the hypothesis test $H_0 : \theta \in \Theta_0 = \{\theta_0\}$ versus $H_1 : \theta \in \Theta_1 = \{\theta_1\}$ with critical value c at significance level α. Define the likelihood-ratio as $\Lambda(X_1, \ldots, X_n) = \frac{L(\theta_0|x_1,\ldots,x_n)}{L(\theta_1|x_1,\ldots,x_n)}$, where the likelihood function is $L(\theta \mid x_1, \ldots, x_n) = \Pi_{i=1}^{n} f(x_i, \theta)$ for $\theta \in \Theta$, if the observations are independent and identically distributed. The test is the most powerful test at significance level α if $\mathbf{P}(\Lambda(X_1, \ldots, X_n) \leq c \mid H_0) = \alpha$ holds. This test with power $\beta(\theta)$ is a uniformly most powerful (UMP) test if $\beta(\theta) \geq \beta^*(\theta)$ for every $\beta^*(\theta)$ that is a power function.

The test in Example 1 is an example of a UMP test.

2.2 Model of detected data

For the data analysis further on in the paper a time series of observed data points will be used. Therefore the following model for the collected data is assumed:

$$Y_i = \mu_i + \epsilon_i, \text{ with } i = 1, \ldots, n \tag{1}$$

where μ_i is the integrated brightness of a star, such that

$$\mu_i = S \cdot I_n = \begin{cases} S & \text{if } I_n = \{P_0, \ldots, P_0 + L - 1\} \\ 0 & \text{otherwise} \end{cases}$$

for a bump with signal strength $S > 0$, signal length L on an interval $I_i \subset \{1, \ldots, n\}$ [5], start position of the signal P_0 and ϵ_i as the noise or observation error, such that $\epsilon_i \sim N(0, 1)$. To determine whether an exoplanet transits the star, each data point Y_i is compared to a set threshold, as an increase in the signal might be visible as an increase in the value of the data points. Whenever a bump is detected, i.e. a data point Y_i is larger than the threshold and the bump appears periodically above this threshold, one is able to confirm the presence of an exoplanet.

Suppose the data points $Y_1, \ldots, Y_n$ are defined as in (1) and are independent with a normal distribution[1]. The probability density function (pdf) $f(y_i, \theta_i)$ of the data points depends

[1]This assumption is made to simplify any further calculations. Assuming that Y_i has a different distribution complicates the statistical tests remarkably, which does not fit in the scope of this paper.

on a parameter $\theta_i \in \Theta$. For the model in (1) it is defined that $\Theta = \{0, k\}$. Resulting from this, the following point hypotheses can be formed:

$$H_0 : T(Y_i) = Y_i \in \Theta_0 = \{0\} \tag{2}$$
$$H_1 : T(Y_i) = Y_i \in \Theta_1 = \{k\} \tag{3}$$

In this specific case, consider that 0 is the base line of the signal. Whenever the test rejects the null-hypothesis it can be concluded that a bump, and thus an exoplanet is detected. Technically, the hypotheses that are tested against each other are *'there is no exoplanet'* (null-hypothesis) versus *'there is an exoplanet'* (alternative hypothesis).

It is a rather difficult task to determine the threshold c in such way that every exoplanet will be detected faultlessly. It may occur that a bump is detected, while this may only be due to strong noise. In this case a Type I error (or false-positive error) occurs: an exoplanet is detected, while not being there. Because it is desirable to make as many true detections as possible, the probability of a Type I error occurring needs to be reduced. To reduce this probability, the threshold (critical value) must be changed. The level of the threshold depends on various parameters that are related to the detected data, e.g. the length of the bump, the variance of the data or the data size.

One of the aims of this paper is to determine which parameters have the most influence on the definition of the threshold. Next to this, it is desirable to find a suitable filter for the data to reduce the noise ϵ_i and increase the signal strength such that the probability of a type I error occurring is low. This way a suitable value of α can be chosen, when the BDM is applied to the Kepler data, which is analysed later on. On that account simulations will be run in order to determine which value of α is fit.

3 Variation of significance level in simulated data

During the detection of exoplanets one must find a compromise between reasonable choices of the critical value c and the significance level α. If it is preferred to choose a low significance level: this implies that the critical value - in the context of exoplanet detection, the threshold - increases. This results in detection of less peaks, meaning the probability that the wanted signal is detected is low. However, choosing the critical value too high could result in detection of no peaks at all. Choosing a high significance level could lead to the detection of peaks in the data that are not the wanted signal, leading to a high probability of the occurrence of the Type I error. The correlation between α and c can be clearly seen in the standard normal distribution table [7], namely $c = \Phi^{-1}(1-\alpha)$, where Φ^{-1} is strictly increasing.

To determine what values of α are suited for practical situations, a simulation study is performed with produced synthetic data sets according to the model in (1). The code for these tests can be found in Appendix D. In this study several values for a certain parameter have been tested against different values of α to see for which significance level the most bumps could be detected. To obtain reliable results, multiple simulations have been performed. For each simulation a new data set has been created. In total 150 simulations[2] have been run. The signal was hidden at the same location within the time series.

[2]This number is based on a small test that has been performed. Running 150 tests on different data sets appeared to have less false rejections than running 50 tests, e.g. For large numbers the results did not differ significantly. Therefore this number appears to be the best choice.

The parameters that were changed were the sample size n, the signal length L, the signal strength S and the standard deviation σ. For each test run, one of these parameters was changed while the α was changed as well, the rest was kept fixed. The hypothesis test that has been performed is the same as described in Section 2.2.

3.1 Results for unfiltered random data set

The obtained results are the average number of (false) rejections for the distinct parameters per each different value of α. While one parameter is changed, the other are fixed. The parameters are kept fixed at $n = 250,\ L = 5,\ S = 2,\ \sigma = 1$.

	α			
n	0.0001	0.005	0.01	0.05
50	0	0.2667	0.4133	2.1
100	0.0067	0.4333	0.9067	5.06
150	0.0067	0.6	1.3	7.0467
200	0.02	1.1	1.8133	10.0867
250	0.02	1.3467	2.4	12.38
300	0.0133	1.4133	3.0733	13.9267

(a) Average number of false rejections

	α			
n	0.0001	0.005	0.01	0.05
50	0.2067	1.6933	2.2333	5.28
100	0.22	1.8267	2.9267	8.1867
150	0.18	2.1067	2.9867	10.1467
200	0.2867	2.3933	3.66	13.28
250	0.1867	2.7533	4.14	15.5733
300	0.2133	3.02	5.0533	17.0933

(b) Average number of rejections

Table 1: Average number of false rejections (a) and all rejections (b) for varying sample size n of 150 simulations.

Table 1 shows the average number of false rejections and all rejections, respectively. As the sample size n gets larger, the average number of false rejections also increases. For smaller values of α this number is low, whereas the average number of false rejection increases to a great extent for larger values of α. For example, for $n = 300$ the average number of false rejections for $\alpha = 0.0001$ is 0.0133, compared to 13.9267 for $\alpha = 0.05$, which is almost 1000 times larger. Concurrently, it can be observed that low values of α give much more true rejections than higher values of α. The true rejections can be determined by substracting the number of false rejections from the number of all rejections. These two observations lead to the conclusion that it is better to keep α low in order to prevent the occurrence of a Type I error, especially if the data set is large.

	α			
L	0.0001	0.005	0.01	0.05
3	0.0067	1.28	2.7067	11.8267
5	0.0533	1.1067	2.6	12.36
7	0.0067	1.1133	2.5467	12.26
9	0.0133	1	2.2533	11.8867
11	0.0133	1.16	2.6133	11.4467

(a) Average number of false rejections

	α			
L	0.0001	0.005	0.01	0.05
3	0.12	2.1133	3.8667	13.8667
5	0.24	2.5667	4.4267	15.42
7	0.32	3.2467	5.1333	16.68
9	0.34	3.4267	5.4	17.64
11	0.46	4.26	6.6667	18.4267

(b) Average number of rejections

Table 2: Average number of false rejections (a) and all rejections (b) for varying signal length L of 150 simulations.

In general it could be said that for an increasing signal length L the average number of rejections also increases, which can be observed in Table 2b. On the other hand, the number of false rejections in table 2a does not vary much, the outcomes tend to fluctuate moderately. Even so, the conclusion that can be drawn is that, with the ascent of the signal

length, the number of false rejections is declining which results in a smaller probability of a Type I error occurring. The last remark that can be made is that the number of (false) rejections in both tables scale proportionally to α, which is consistent with the construction of the test and the independence of the data points.

	α			
S	0.0001	0.005	0.01	0.05
1	0.02	1.2133	2.52	12.3867
2	0.0267	1.3067	2.56	12.74
3	0.0133	1.2	2.4533	12.6933
4	0.0267	1.22	2.4533	12.2067
5	0.02	1.18	2.5333	12.1667

(a) Average number of false rejections

	α			
S	0.0001	0.005	0.01	0.05
1	0.0333	1.42	2.96	13.7
2	0.2667	2.78	4.4867	15.98
3	1.1333	4.5733	6.1333	17.2333
4	3.24	5.84	7.1733	17.1733
5	4.52	6.1333	7.5267	17.1667

(b) Average number of rejections

Table 3: Average number of false rejections (a) and all rejections (b) for varying signal strength S of 150 simulations.

The results for the average number of (false) rejections for a varying signal strength may be interpreted almost the same as for the varying signal length. The only pronounced difference seen in Table 3 is that the average number of rejections increases much more than for the previous results for a varying signal length. This leads to less false rejections in total each time the signal strength increases. Similarly, the average number of false rejections also scales proportionally with the increasing α. With this in mind, one could decide for small values of α in case of a strong signal, thus a high S.

	α			
σ	0.0001	0.005	0.01	0.05
1	0.0267	1.1267	2.2933	12.1933
1.25	0.4067	4.72	7.5	23.2333
1.5	1.62	10.4467	14.3467	32.7933
1.75	3.8333	16.9533	22.32	43.0333
2	7.7267	23.3333	30.1933	50.3467

(a) Average number of false rejections

	α			
σ	0.0001	0.005	0.01	0.05
1	0.1667	2.6067	3.9533	15.2
1.25	0.74	6.24	9.5	26.36
1.5	2.2067	12	16.52	35.9867
1.75	4.7733	18.7667	24.72	46.0133
2	8.8067	25.1467	32.42	53.4

(b) Average number of rejections

Table 4: Average number of false rejections (a) and all rejections (b) for varying standard deviation σ of 150 simulations.

As for the increasing σ the changes within the values are much more extreme, compared to the outcomes for the other parameters. With every increment of σ the average number of (false) rejections almost doubles. What is remarkable is that, in contrast with all previous outcomes, the number of true rejections is overall larger for larger values of α than for smaller values, although the numbers of false rejections are fairly high. This should be kept in mind when choosing a suitable significance level for the data.

For each varied parameter the results can be summarized as follows:

- For an increasing sample size it is better to keep α low
- For any signal length it is best to keep α below 0.01
- For a large signal strength it is best to keep α low as well
- For an increasing standard deviation it is acceptable to choose larger values of α

To conclude, in general it seems better to keep α as low as possible, which gives better outcomes for the number of false rejections. However, the outcomes could be improved such that higher values of α will give a low number of false rejections as well. This is elaborated on in the next section.

4 Filtering simulated data

As may be concluded from Section 3, the average number of false rejections is fairly high for some parameters. When certain parameters are increased, the average number of false rejections becomes very large. To assure that this number decreases, thus ensuring the correct detection of exoplanets, the signal-to-noise ratio (SNR) of the data must be maximised [3]. Applying a matched filter to a data set ensures this maximization since this is the optimal linear filter in presence of additive stochastic noise [9][14].

4.1 Introduction to matched filter

Simply said, the matched filter enhances the signal that is present within the data set. It is assumed that the shape of the signal is known and the filter equals this specific physical shape. In case of exoplanet detection it is assumed the filter is a time series containing a bump, as shown in Figure 2. Because of the shape of the signal, the filter represents the pure signal without noise. Using the matched filter, it is possible to find a correlation

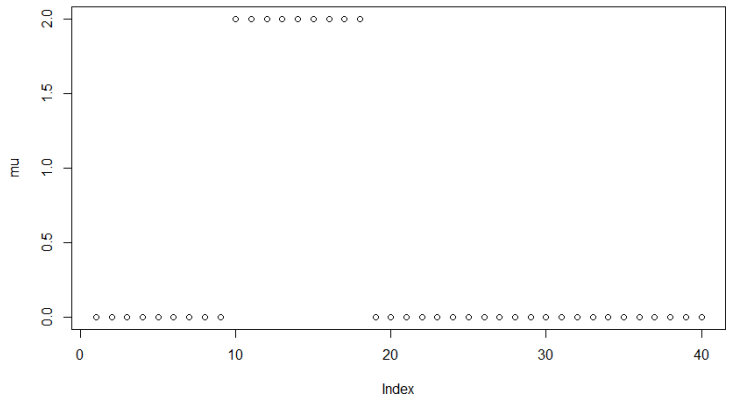


Figure 2: An example of a matched filter with $S = 2$ and $L = 9$.

between the signal in the data set and the filter itself. Consider $f_i = \frac{1}{\sigma\sqrt{L}} I_{\{P_0,\ldots,P_0+L-1\}}$ to be the matched filter and Y_i the unfiltered data as in (1). Then, the result of the filtered output is the convolution of f_i and Y_i:

$$X_i = \sum_{k=-\infty}^{\infty} f_{k-i} \cdot Y_k, \quad i \in \{1, \ldots, n - L + 1\} \tag{4}$$

such that $Y_k = 0$ if $k \leq 0$ or $k > n$. Applying the matched filter to the data will result in a peak in the filtered data, which appears when the filter encounters the signal. The maximum of this peak is when the filter is aligned with the data. Figure 3 is an example of the application of a matched filter to a random data set. The black points represent the unfiltered data, where the red dots are the filtered data points. The location of the signal

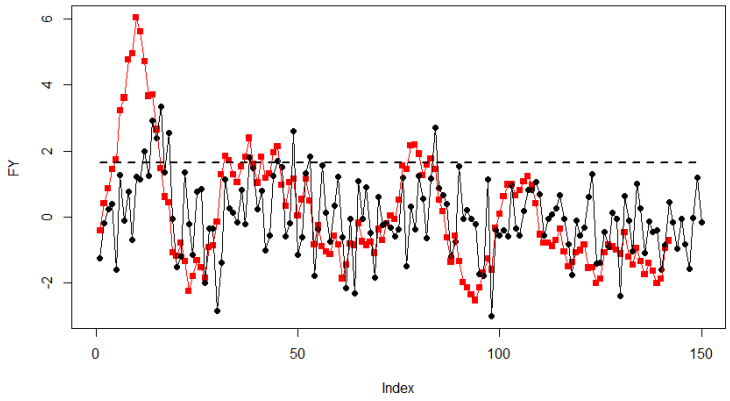


Figure 3: Unfiltered random data set (black) and random data set where matched filter is applied (red). $n = 150$, $L = 9$, $S = 2$, $\sigma = 1$. The dotted line is the threshold that is set ($c = 1.96$). The peak in the filtered data appears to be more prominent that in the unfiltered data.

in the filtered data easier to detect than in the unfiltered data. To conclude, the filtered signal is not present at the last $L - 1$ points. This is evident, because the last part of the time series where the filter is matched is located precisely there.

4.2 Increasing detection power by filtering data

By applying the matched filter to the simulated data it can be shown using the Neyman-Pearson lemma that the detection power is increased, which means the probability of rightfully rejecting H_0 is increased. First, it is assumed that the position P_0 of the signal of length L in the simulated data set is known. Then the next hypotheses follow:

$$H_0 : \overrightarrow{\mu} \equiv 0 \tag{5}$$
$$H_1 : \overrightarrow{\mu} = S \cdot I_n \tag{6}$$

with $\overrightarrow{\mu} = (\mu_1, \ldots, \mu_n)$ and $I_n = \{P_0, \ldots, P_0 + L - 1\}$. In this situation - after filtering - it is assumed that $\Theta = \mathbf{R}^n$, because vectors are constantly compared with each other. Hence, this results in the disjoint sets $\Theta_0 = \{(0, \ldots, 0) \in \mathbf{R}^n\}$ and $\Theta_1 = \{(0, \ldots, 0, S, \ldots, S, 0, \ldots, 0) \in \mathbf{R}^n\}$. Because this is a point hypothesis test, as in Example 1 it is possible to apply the Neyman-Pearson lemma to the test to see whether this is the most powerful test.

Lemma 1. The statistical test with null-hypothesis (5) and alternative hypothesis (6) is the most powerful test.

Proof. According to Theorem 1 a statistical test is most powerful at significance level α if $\mathbf{P}(\Lambda(Y_1, \ldots, Y_n) \leq c \mid H_0) = \alpha$ holds. Assume that $Y_1, \ldots, Y_n$ are indentically distributed such that $f(y_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(y_i - \theta_i)^2}{2\sigma}$ is the pdf of Y_i.

Calculating the likelihood ratio results in

$$\begin{aligned}
\Lambda(Y_1,\ldots,Y_n) &= \frac{L(\theta_0|y_1,\ldots,y_n)}{L(\theta_1|y_1,\ldots,y_n)} \\
&= \frac{\Pi_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-y_i^2}{2\sigma}\right)}{\Pi_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i-\theta_i)^2}{2\sigma}\right)} \\
&= \frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(\sum_{i=1}^n \frac{-y_i^2}{2\sigma}\right)}{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(\sum_{i=1}^n \frac{-(y_i-\theta_i)^2}{2\sigma}\right)} \\
&= \exp\Big(-\sum_{i=1}^n \frac{y_i^2}{2\sigma} + \sum_{i=1}^n \frac{(y_i-\theta_i)^2}{2\sigma}\Big) \\
&\text{[Split the second sum in three different parts]} \\
&= \exp\Big(-\sum_{i=1}^n \frac{y_i^2}{2\sigma} + \sum_{i=1}^{P_0-1} \frac{y_i^2}{2\sigma} + \sum_{i=P_0}^{P_0+L-1} \frac{(y_i-S)^2}{2\sigma} + \sum_{i=P_0+L}^n \frac{y_i^2}{2\sigma}\Big) \\
&= \exp\Big(\sum_{i=P_0}^{P_0+L-1} \frac{-2y_iS+S^2}{2\sigma}\Big).
\end{aligned}$$

This results leads to the following probability that must hold for the most powerful test:

$$\mathbf{P}\Big(\exp\Big(\sum_{i=P_0}^{P_0+L-1} \frac{-2y_iS+S^2}{2\sigma}\Big) \le c \mid H_0\Big) = \alpha.$$

Applying the natural logarithm function to both sides of the inequality will result in

$$\mathbf{P}\Big(\sum_{i=P_0}^{P_0+L-1} \frac{-2y_iS+S^2}{2\sigma} \le \ln c \mid H_0\Big) = \alpha.$$

Splitting the sum gives

$$\begin{aligned}
&\mathbf{P}\Big(\frac{S^2}{2\sigma}\sum_{i=P_0}^{P_0+L-1} 1 - \frac{2S}{2\sigma}\sum_{i=P_0}^{P_0+L-1} y_i \le \ln c \mid H_0\Big) = \alpha \\
&\mathbf{P}\Big(\frac{S^2}{2\sigma}(L-1) - \frac{S}{\sigma}\sum_{i=P_0}^{P_0+L-1} y_i \le \ln c \mid H_0\Big) = \alpha
\end{aligned}$$

Now the constant on the left-hand side of the inequality can be moved to the right-hand side. After this, multiplying both sides of the inequality with $-\frac{1}{S\sqrt{L}}$ yields

$$\begin{aligned}
&\mathbf{P}\Big(\frac{1}{\sigma\sqrt{L}}\sum_{i=P_0}^{P_0+L-1} y_i \ge \frac{\ln c}{S\sqrt{L}} + \frac{S(L-1)}{2\sigma\sqrt{L}} \mid H_0\Big) = \alpha \\
&\mathbf{P}\Big(\frac{1}{\sigma\sqrt{L}}\sum_{i=P_0}^{P_0+L-1} y_i \ge \tilde{c}\Big) = \alpha
\end{aligned}$$

□

and this gives indeed a test with the best power, meaning this is a UMP test.

5 Variation of significance level in filtered simulated data

In this section the experiments of Section 3 have been repeated, only now there will be dealt with filtered data. The goal is to reduce the average number of false rejections for the varying parameters, leading to recommendations on choosing a suitable α when dealing with the Kepler data.

5.1 Results for filtered random data set

	α			
n	0.0001	0.005	0.01	0.05
50	0	0.28	0.3467	1.633
100	0	0.3867	1.0467	4.4333
150	0.0267	0.5133	1.1133	5.9267
200	0.0133	1.08	2.12	9.3067
250	0.0333	1.06	2.2933	12.06
300	0.0133	1.1467	3.1667	14.4133

(a) Average number of false rejections

	α			
n	0.0001	0.005	0.01	0.05
50	2.2467	5.4	6.5667	9.2867
100	2.1333	5.52	7.3933	12.1
150	1.9733	5.9533	7	13.6867
200	2.0733	6.42	8.38	17.04
250	2.3133	6.5	8.7533	19.9667
300	2.12	6.34	9.3933	21.9667

(b) Average number of rejections

Table 5: Average number of false rejections (a) and all rejections (b) for varying sample size n of 150 simulations of filtered data.

First of all, the difference between the average number of false rejections between the unfiltered and filtered data in Table 5 is not significant. For higher values of α the results for the filtered data are slightly better, but not significantly. However, if one takes a look at the average number of rejections it can be established that this number has increased. Both of these observations imply the increase of the average number of true rejections. Last but not least, because more rejections are made for the filtered data, the detection of peaks is more precise - peaks that should have been rejected in the unfiltered data are now rejected, hence the correct detection of the actual signal is ensured.

	α			
L	0.0001	0.005	0.01	0.05
3	0.0133	1.0333	2.2267	11.34
5	0.0267	1.18	2.4	11.2533
7	0.0467	0.98	2.0067	11.6867
9	0.0067	1.0533	2.0333	10.8
11	0.04	1	2.6067	11.3267

(a) Average number of false rejections

	α			
L	0.0001	0.005	0.01	0.05
3	0.34	2.4667	3.5933	14.3467
5	2.1667	6.4733	8.5933	19.12
7	4.8467	8.98	11.0933	22.4067
9	7.12	11.36	13.18	23.4
11	8.8667	12.9067	15.5667	26.1

(b) Average number of rejections

Table 6: Average number of false rejections (a) and all rejections (b) for varying signal length L of 150 simulations of filtered data.

After filtering, the average number of false rejections for the lowest α is slightly higher than before filtering for the varying L. For the rest of the outcomes the average number is slightly higher compared to the results in Table 6a. A prominent change is the increase of the average number of rejections. It can be seen in table 6b that for $L = 3$ this number is somewhat lower than before filtering; when L starts to increase, the average number of rejections is increasing to a great extent. Just as for n this implies that for a longer signal the number of true rejections that is made is large. The larger α becomes, the more true rejections are made.

S	α			
	0.0001	0.005	0.01	0.05
1	0	0.9933	2.6533	11.92
2	0.0333	1.1933	2.4467	11.22
3	0.0067	1.0733	2.7	11.7
4	0.0067	0.9	2.22	11.14
5	0.0067	1.4267	2.5067	10.6133

(a) Average number of false rejections

S	α			
	0.0001	0.005	0.01	0.05
1	0.12	2.1133	4.1867	15.58
2	1.8067	6.5467	8.48	18.66
3	5.8267	8.8267	10.7933	20.4733
4	7.76	9.3867	10.9067	20.4
5	8.2933	10.3267	11.5533	20.02

(b) Average number of rejections

Table 7: Average number of false rejections (a) and all rejections (b) for varying signal strength S of 150 simulations for filtered data.

The average number of (false) rejections for filtered data, as is shown in Table 7, are slightly less than for the unfiltered data. The only big difference is for $\alpha = 0.0001$; the average number of false rejections is almost 0, where the average number of rejections is almost twice as high, meaning that a lot of true rejections have been made. In conclusion, if the S becomes larger, it is better to choose a higher value of α, due to the larger number of true rejections.

σ	α			
	0.0001	0.005	0.01	0.05
1	0.0133	1.1733	2.5	12.3333
1.25	0.0467	1.1533	2.1933	12.3533
1.5	0	1.2933	2.4733	11.2467
1.75	0.0133	0.98	2.1867	11.58
2	0.0133	0.92	2.48	11.97633

(a) Average number of false rejections

σ	α			
	0.0001	0.005	0.01	0.05
1	2.1667	7.0333	8.7333	19.9467
1.25	1.1867	4.48	6.5667	18.5533
1.5	0.2867	3.92	5.5133	16.8733
1.75	0.24	2.2333	4.28	15.7867
2	0.08	2.3533	4.2333	15.2533

(b) Average number of rejections

Table 8: Average number of false rejections (a) and all rejections (b) for varying standard deviation σ of 150 simulations for filtered data.

Filtering the data set has made the most impact on the results for changing σ. While comparing Tables 4a and 8a, one can observe that the average number of false rejections has decreased a lot, especially for larger values of α. The same conclusion can be drawn for the average number of rejections. As already mentioned in Section 4, due to the higher SNR the fluctuation of the data points is much less after filtering. The results of the hypothesis tests are a good illustration of this phenomenon. Though the numbers in Table 4 are lower than in Table 8, the number of true rejections is approximately the same. Even so, it is better to reduce the number of (false) rejections in general to rectify the detection of a signal in the given data set.

5.2 Recommendations for the Kepler data set

After performing statistical tests on simulated data sets, it is possible to give recommendations about choices for α for statistical tests that will be performed on actual Kepler data. All in all, the results above have shown that in any case it is desirable to filter any given data set using the matched filter. Furthermore, it is important to consider what kind of data set one is dealing with. Due to the large size of the Kepler data set it is best to keep the significance level low ($\alpha \leq 0.01$). For a longer signal length it is best to let α be relatively large, namely $\alpha \geq 0.005$. These two points lead to the conclusion that it is best to choose α between 0.005 and 0.01. The fact that the results for the unfiltered and

filtered data set for varying S and σ are not significantly unalike, justifies the choice for $0.005 \leq \alpha \leq 0.01$.

6 Kepler data

From the Kepler data three objects (time series data sets) - KIC10910878, KIC4563268 and KIC11954842 - have been retrieved. The plots of these data sets can be found in Appendix B. Each of these objects has a certain number of candidates (potential exoplanets) hidden in the data. For each of these objects a test has been run for $\alpha \in \{0.005, 0.0075, 0.01\}$ to see how much peaks would be detected. All the parameters that have been tested in Section 3 and 5 are unknown, except for the sample size - each object consists of 1640 observation points. For each test run the significance level and the signal length have been varied. The results that are obtained from the tests are the number of rejections. This does not give information about how many peaks there are present, but just that there is a possibility that peaks could be present.

6.1 Results for Kepler data after application of the BDM

In Table 9 the number of rejections for filtered and unfiltered data is denoted. The first thing that must be noted is that for each data set the number of rejections for unfiltered data increases per increasing α. Though, this number remains the same for varying L, because only the filtered data depends on this parameter (which is used for the matched filter).

	α		
L	0.005	0.0075	0.01
6	24 (22)	27 (27)	30 (31)
10	28 (22)	30 (27)	31 (31)
14	34 (22)	36 (27)	37 (31)
18	38 (22)	39 (27)	41 (31)

(a)

	α		
L	0.005	0.0075	0.01
6	30 (13)	37 (18)	40 (22)
10	41 (13)	45 (18)	49 (22)
14	31 (13)	41 (18)	48 (22)
18	12 (13)	18 (18)	22 (22)

(b)

	α		
L	0.005	0.0075	0.01
6	13 (12)	14 (15)	14 (18)
10	18 (12)	18 (15)	22 (18)
14	12 (12)	12 (15)	14 (18)
18	0 (12)	2 (15)	6 (18)

(c)

Table 9: Number of rejections for the Kepler data KIC10910878 (a), KIC4563268 (b) and KIC11954842 (c) with a varying α and L. The number between brackets denotes the number of rejections for the unfiltered data, the other is for the filtered data.

For the first object KIC10910878 it can be clearly seen that when L is increasing, this is also the case for the number of rejections. The shorter the signal, the larger these peaks are for the filtered data. This is made visible in Figure 4. The large number of rejections implies that more candidate planets may be present. Figure 4 shows two large peaks, meaning it is more than likely that at least one candidate is present. If one takes a closer

look, several periodical peaks appear just above the set threshold. This could imply the presence of another candidate.

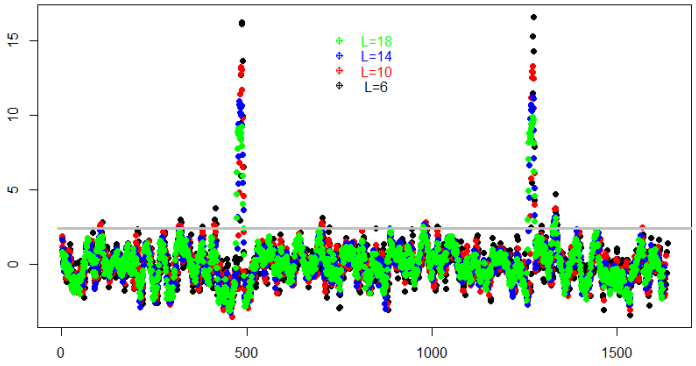


Figure 4: Filtered data points of object KIC10910878 with varying signal length L. The value of α is set at 0.001.

The outcomes of the hypothesis test on the second object KIC4563268, denoted in Table 9b, appear to be different compared to the outcomes for KIC10910878, which can be viewed in Table 9a. Here it can be seen that for $L = 10$ the number of rejections is the highest. However, in Figure 11 the peaks of a candidate appear to be the highest for $L = 6$. In some cases, the peaks are the largest for $L = 10$. In any case, at least one candidate can be detected for each value of L. The presence of more candidates remains unclear. Table

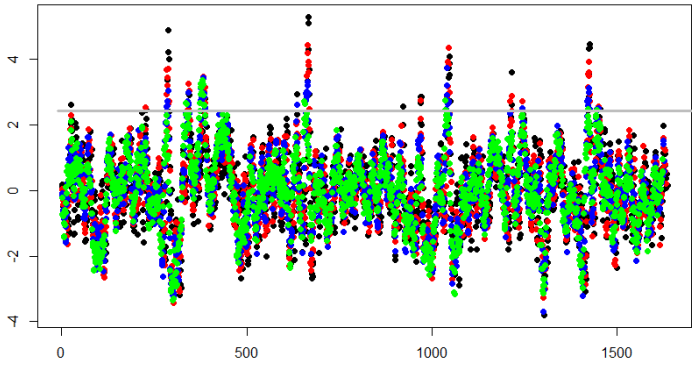


Figure 5: Filtered data points of object KIC4563268 with varying signal length L. The value of α is set at 0.001. The colour code for the varying L is identical to the colour code as in Figure 4.

9c denotes the results for the last Kepler object KIC11954842. Again, it can be seen that for $L = 10$ the most rejections are made. What is interesting is that for smaller numbers of L the number of rejections is almost the same for each α, here it starts to differ for $L = 18$. This may be due to extremely fluctuating data. Looking at Figure 6 one can deduce that it is very likely that a candidate is present, though the signal length of this

candidate remains unclear. The figure shows that some peaks are detected best for $L = 6$, while other peaks appear above the threshold for $L = 10$. This could be a peak of another candidate.

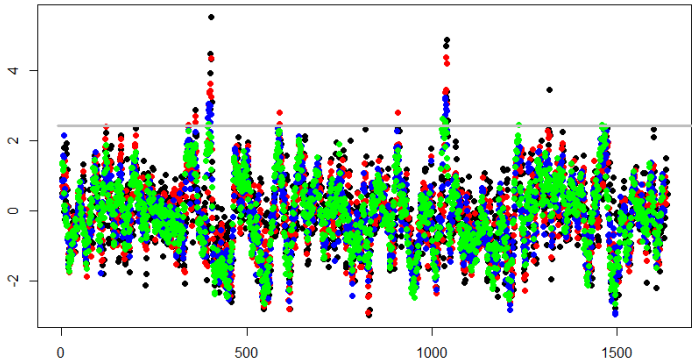


Figure 6: Filtered data points of object KIC11954842 with varying signal length L. The value of α is set at 0.001. The colour code for the varying L is identical to the colour code as in Figure 4.

6.2 Reevaluation of the results for Kepler data

Since the objects have already been observed and studied by physicists, it is possible to compare the results obtained in Section 6.1 with the results from real-life observations. The page summaries [1] can be found in Appendix C. Each new observation was made approximately half an hour after the previous observation. Thus the period between two data points is 0.49 hours.

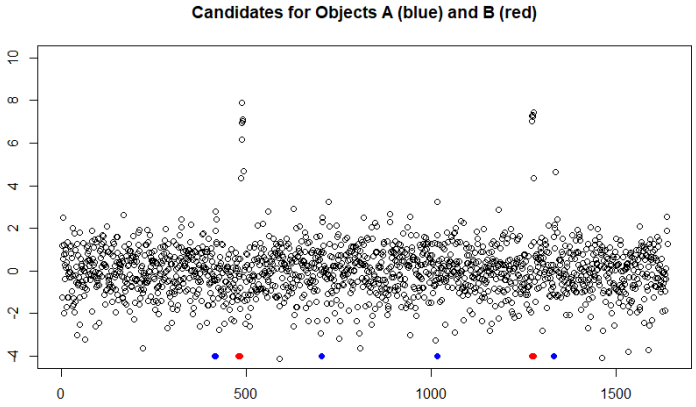


Figure 7: Raw data of Kepler object KIC10910878 (black transparent dots) with positions of candidates depicted below the data. The red points are the positions of Candidate 1 and the blue points are the positions of Candidate 3.

Figure 7 shows that the largest peaks are located at $i = 487$ and $i = 1273$. This would mean that the period between both peaks in days is approximately $\frac{(1273-487)\cdot 0.49}{24} \approx 16.05$

days. Comparing this with the pages summaries leads to the conclusion that the large peaks indicate the presence of Candidate 1. In Appendix C.1 is can be seen that the verified period is approximately 16.07 days, which is close to the result of this paper.

The second candidate has more positions that appear in Figure 7. Based on data points $i = 415$ and $i = 704$ it can be calculated that the period of this candidate is $\frac{(704-415)\cdot 0.49}{24} \approx 5.90$ days. This corresponds with Candidate 3 in Appendix C.3, which has a period of approximately 6.26 days.

According to the page summaries a third candidate should be present as well, namely Candidate 2. However, this candidate has a period of approximately 41.20 days. The analysis of the object of this paper has a period of approximately 34.17 days. This means that even if Candidate 2 would be detected, this would emerge in a single peak. Hence, no conclusion can be drawn about the presence of a potential exoplanet, since this peak could also be noise.

The peaks of the second object could not be identified using the provided script, even though the peaks are visible in the filtered data. Therefore no conclusion can be drawn regarding the detected objects. According to the page summaries two candidates should be present in the object. Figure 11 supports this, due to the clearly visible peaks in the filtered data.

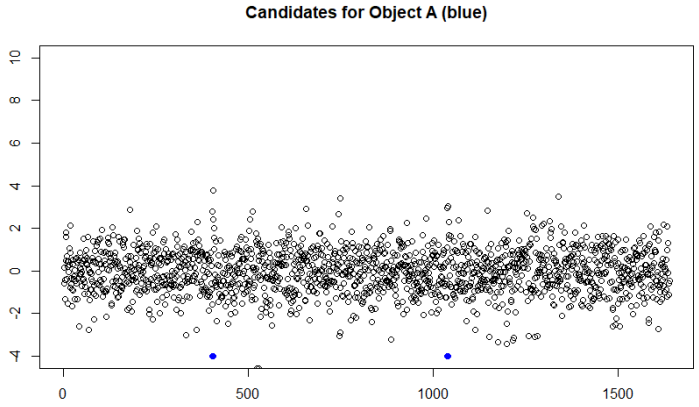


Figure 8: Raw data of Kepler object KIC11954842 (black transparent dots) with positions of candidates depicted below the data. The blue points are the positions of Candidate 1.

In Figure 8 the positions of a potential candidate are depicted. The locations of these points are $i = 403$ and $i = 1039$. Calculating the period of this candidate will result in a period of $\frac{(1039-403)\cdot 0.49}{24} \approx 12.99$ days. According to Appendix C.6 the period of this candidate is also approximately 12.99 days, which means Candidate 1 in Object 3 is located successfully.

7 Conclusion and Discussion

The goal of this paper was to apply statistical hypotheses tests to data sets in order to determine whether or not it is possible to detect exoplanets. First of all, a model was defined for the data points of the simulated data. Next to this it has been shown that the BDM could be applied to this type of data, which contains the most powerful statistical point hypothesis test. This has been done using the Neyman-Pierson lemma.

Hereafter, hypothesis tests - as formulated in (2) and (3) - have been performed on the simulated unfiltered data. For these hypotheses n, L, S and σ were varied, after which the (false) rejections were calculated. These outcomes were compared to each other for various values of α to observe how this would influence the detection of a wanted signal. For future research, one could refine on the different values of α to see whether the change of the number of (false) rejections clearly. Next to this, it would be worthwhile to inspect σ specifically, since the transitions within the number of (false) rejections for this parameter were the largest.

Filtering the simulated data sets that have been tested has proven to be beneficial. It could be clearly seen that the number of false rejections has been reduced, where this was not necessarily the case for the total number of rejections. This leads to the conclusion that more correct rejections have been made, hence more precise detections have been achieved. What could improve the detection of a wanted signal even more is applying the Holm-Bonferroni method [8] on the simulated data. Briefly explained, this method is used for multiple-hypothesis testing to control the family-wise error rate - meaning the probability that one or more errors of Type I will occur. Since this paper was dealing with multiple-hypothesis testing for simulated data, and further on the Kepler data, this method could have been useful.

At last the BDM has been applied to three Kepler data sets, or objects, to determine whether or not potential exoplanets could be detected by means of this method. Due to the absence of knowledge about the position P_0 and the precise length L of the wanted signal it was not possible to determine the number of false rejections, meaning it was not possible to determine specifically whether or not a signal could be identified. Nevertheless, using the results of the total number of rejections and the plots it was possible to deduce if there could be a potential exoplanet present or not. This was certainly the case for each of the Kepler data sets.

Revising the outcomes of the application of the BDM to the objects and comparing these to the physical outcomes has lead to the conclusion that for Object KIC10910878 and KIC11954842 two and one candidates could be found, respectively. To conclude this research, it can be said that the BDM has been proven a reliable method for exoplanet detection. For future work one could look more into the detection of the location of peaks. This has now been done by taking the highest and second highest point of a candidate in the data and using these for the calculation of the period.

References

[1] Kepler Science Operations Center Pipeline at NASA Ames Research Center. Kepler and k2 data processing pipeline. `https://keplerscience.arc.nasa.gov/pipeline.html`, last accessed on 2020-01-14.

[2] G.J. Babu. *A review of exoplanets detection methods*, volume 244, pages 79–87. Springer New York LLC, 1 2018.

[3] J.B. Carlin and L.W. Doyle. Basic concepts of statistical reasoning: Hypothesis tests and the t-test. *Journal of paediatrics and child health*, 37(1):72–77, 2001.

[4] D. Charbonneau, T.M. Brown, D.W. Latham, and M. Mayor. Detection of planetary transits across a sun-like star. *The Astrophysical Journal Letters*, 529(1):L45, 1999.

[5] F. Enikeeva, A. Munk, F. Werner, et al. Bump detection in heterogeneous gaussian regression. *Bernoulli*, 24(2):1266–1306, 2018.

[6] N.K. Glendenning. *Compact stars: Nuclear physics, particle physics and general relativity*. Springer Science & Business Media, 2012.

[7] R.V. Hogg, J.W. McKean, and A.T. Craig. *Introduction to Mathematical Statistics*. Pearson education international. Pearson Education, 2005.

[8] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[9] Y. I. Ingster and I.A. Suslina. Detection of a signal of known shape in a multichannel system. *Journal of Mathematical Sciences*, 127(1):1723–1736, 2005.

[10] M. Mayor and D. Queloz. A jupiter-mass companion to a solar-type star. *Nature*, 378(6555):355, 1995.

[11] D. Overbye. Kepler,the little nasa spacecraft that could, no longer can. *The New York Times*.

[12] NASA Exoplanet Exploration Program. Search. `{https://exoplanets.nasa.gov/alien-worlds/historic-timeline/#first-exoplanets-discovered}, Lastaccessedon2019-11-01`,.

[13] J. Shao. *Mathematical statistics: exercises and solutions*. Springer Science & Business Media, 2006.

[14] G. Turin. An introduction to matched filters. *IRE transactions on Information theory*, 6(3):311–329, 1960.

A Table of symbols

Symbol	Description
n	Sample size of a data set
L	Length of the signal (bump)
S	Strength of the signal (bump)
σ	Standard deviation of a data point Y_i in a data set
α	Significance level set to perform a statistical test
c	Critical value of a statistical test
Z	Standard normal random variable
Φ	Cumulative distribution function of the standard normal distribution

Table 10: Table of symbols

B Plots of unfiltered Kepler objects

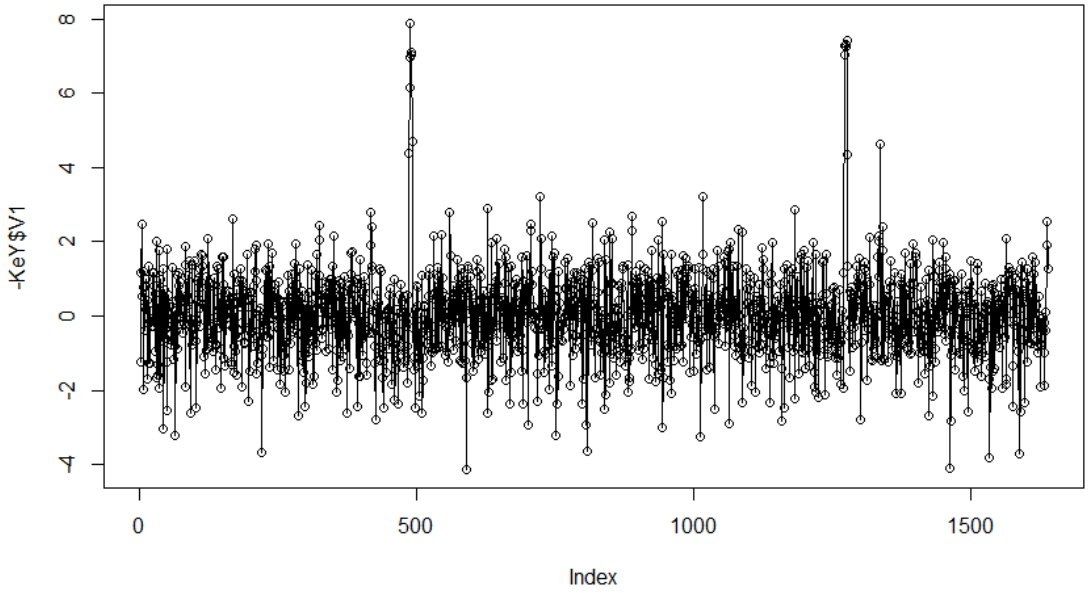


Figure 9: Data points of object KIC4563268.

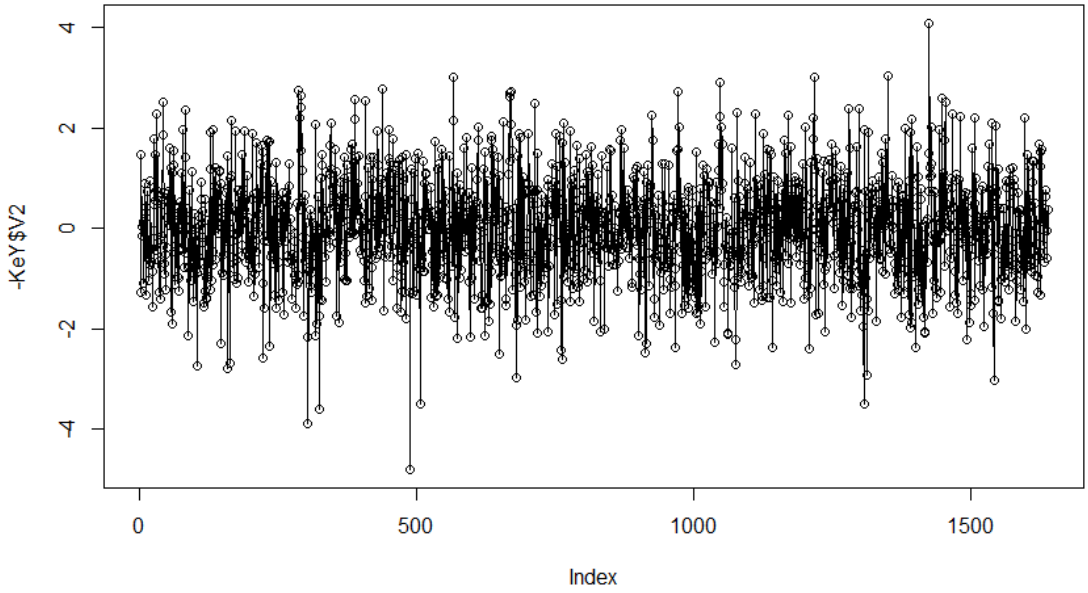


Figure 10: Data points of object KIC4563268.

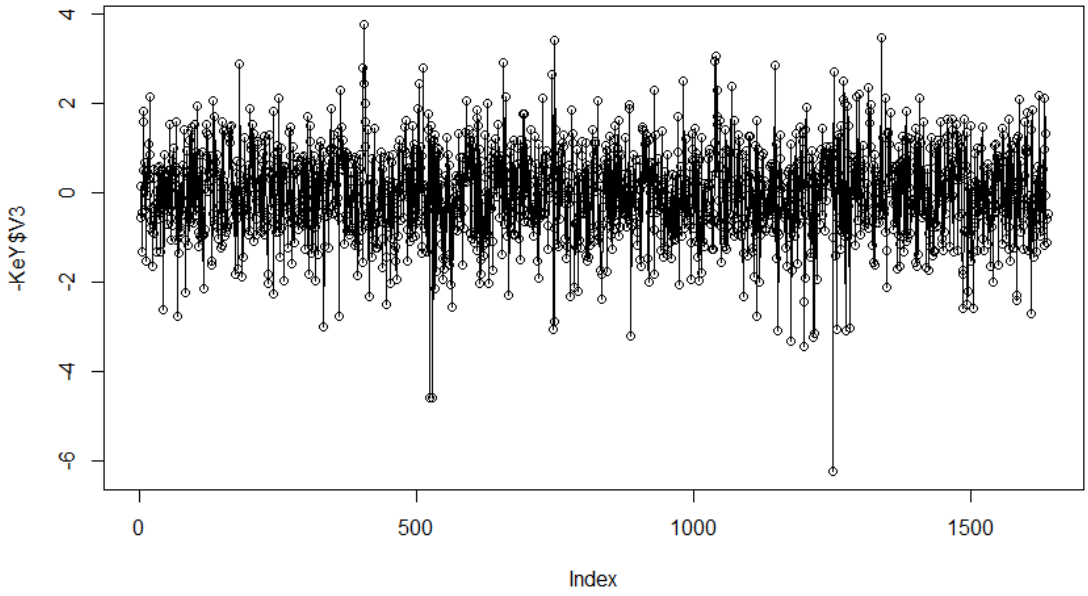
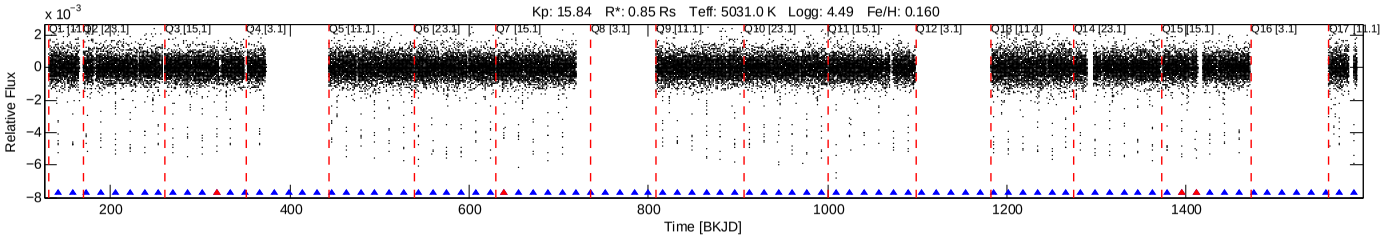


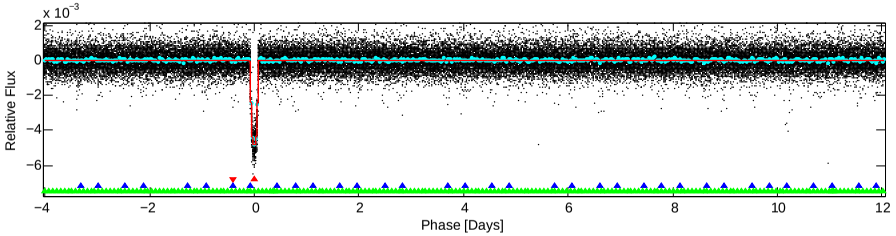
Figure 11: Data points of object KIC4563268.

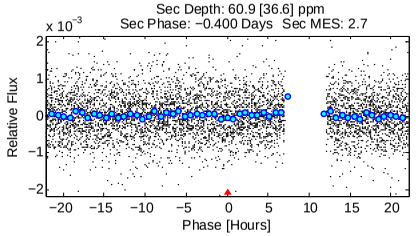
C Summaries of Kepler data objects

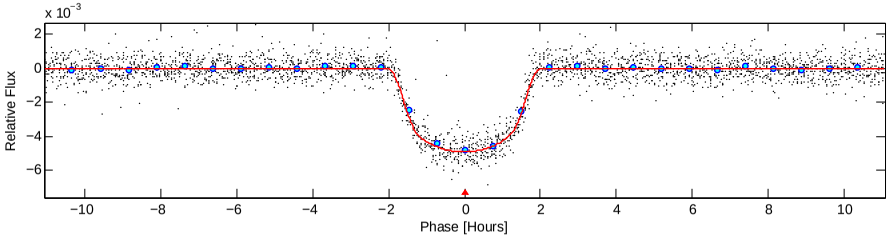
C.1 Object 1 - candidate 1

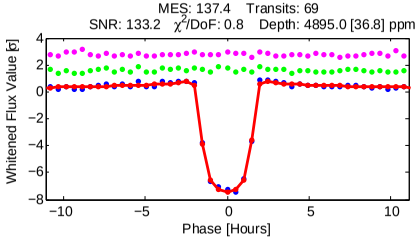
KIC: 10910878 Candidate: 1 of 3 Period: 16.069 d
KOI: K00757.01 Name: Kepler−229c Corr: 0.980

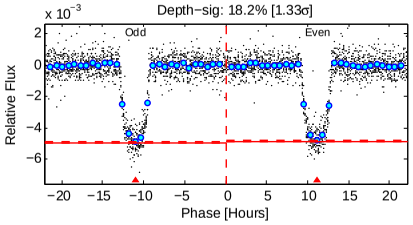


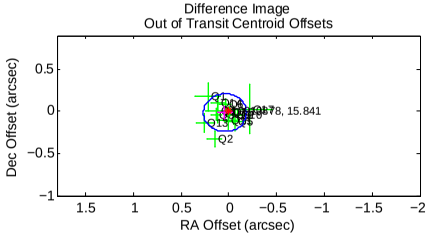












DV Fit Results:

Period = 16.06864 [0.00001] d
Epoch = 141.4851 [0.0006] BKJD
Rp/R* = 0.0671 [0.0026]
a/R* = 28.30 [3.56]
b = 0.64 [0.12]

Seff = 30.79 [4.75]
Teq = 601 [23] K
Rp = 6.21 [0.58] Re
a = 0.1158 [0.0099] AU

Ag = 11.65 [7.23] [1.47σ]
Teffp = 1716 [261] K [4.26σ]

DV Diagnostic Results:

ShortPeriod−sig: 100.0% [51.53σ]
LongPeriod−sig: 100.0% [88.82σ]
ModelChiSquare2−sig: 94.9%
ModelChiSquareGof−sig: 100.0%
Bootstrap−pfa: 0.00e+00
RollingBand−fgt: 0.94 [60/64]
GhostDiagnostic−chr: 3.289

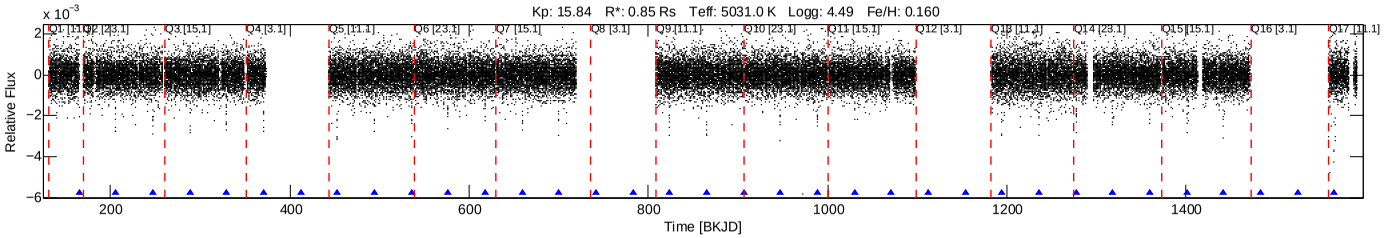
Centroid−sig: 6.3%
Centroid−so: 0.055 arcsec [0.62σ]
OotOffset−rm: 0.044 arcsec [0.58σ]
KicOffset−rm: 0.031 arcsec [0.38σ]
OotOffset−st: 4/4/1/5 [14]
KicOffset−st: 4/4/1/5 [14]
DiffImageQuality−fgm: 1.00 [14/14]
DiffImageOverlap−fno: 0.93 [13/14]

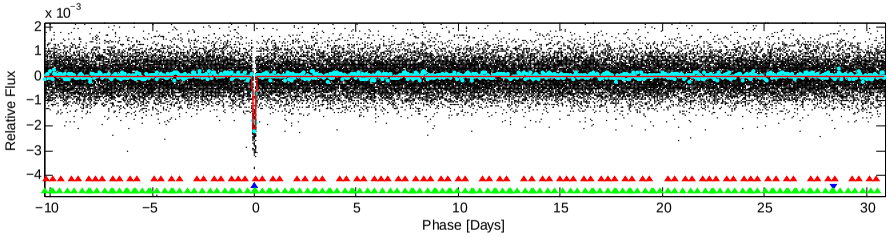
Software Revision: svn+ssh://murzim/repo/soc/tags/release/9.3.42@60958 −− Date Generated: 29−Jan−2016 03:41:18 Z

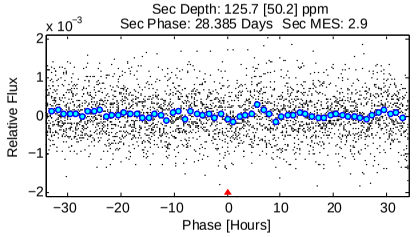
This Data Validation Report Summary was produced in the Kepler Science Operations Center Pipeline at NASA Ames Research Center

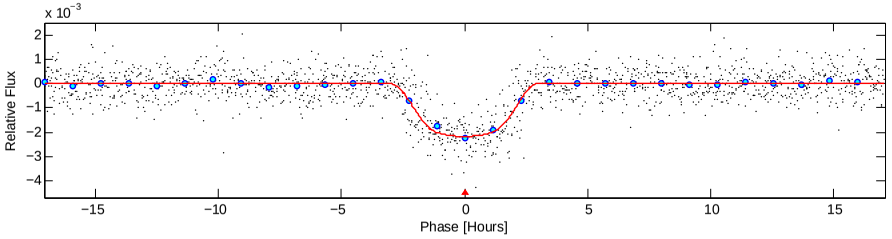
C.2 Object 1 - candidate 2

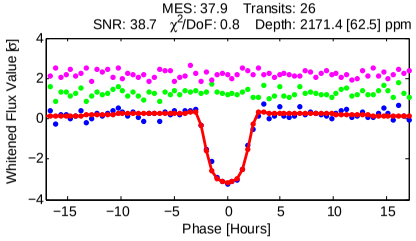
KIC: 10910878 Candidate: 2 of 3 Period: 41.197 d
KOI: K00757.02 Name: Kepler−229d Corr: 0.894

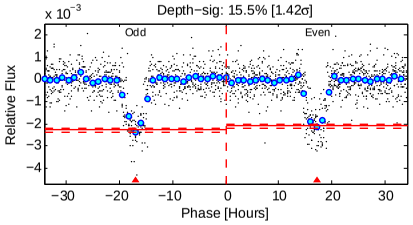


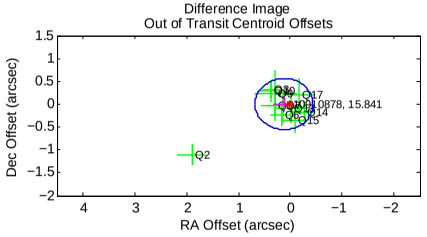












DV Fit Results:

Period = 41.19698 [0.00015] d
Epoch = 165.0073 [0.0030] BKJD
Rp/R* = 0.0535 [0.0014]
a/R* = 28.57 [1.87]
b = 0.92 [0.01]

Seff = 8.78 [1.35]
Teq = 439 [17] K
Rp = 4.95 [0.44] Re
a = 0.2169 [0.0185] AU

Ag = 132.91 [56.71] [2.33σ]
Teffp = 2304 [235] K [7.90σ]

DV Diagnostic Results:

ShortPeriod−sig: 100.0% [88.82σ]
LongPeriod−sig: N/A
ModelChiSquare2−sig: 52.3%
ModelChiSquareGof−sig: 100.0%
Bootstrap−pfa: 2.11e−278
RollingBand−fgt: 1.00 [23/23]
GhostDiagnostic−chr: 5.864

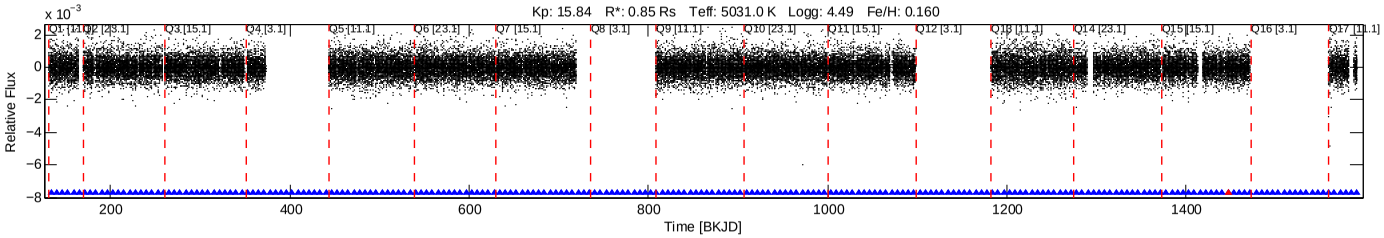
Centroid−sig: 12.3%
Centroid−so: 0.426 arcsec [1.41σ]
OotOffset−rm: 0.124 arcsec [0.66σ]
KicOffset−rm: 0.079 arcsec [0.50σ]
OotOffset−st: 4/3/0/4 [11]
KicOffset−st: 4/3/0/4 [11]
DiffImageQuality−fgm: 1.00 [11/11]
DiffImageOverlap−fno: 0.91 [10/11]

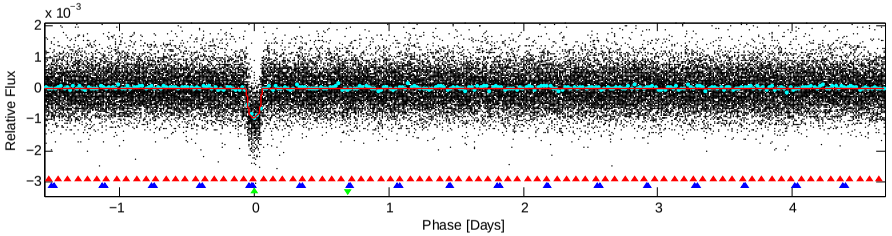
Software Revision: svn+ssh://murzim/repo/soc/tags/release/9.3.42@60958 −− Date Generated: 29−Jan−2016 03:41:26 Z

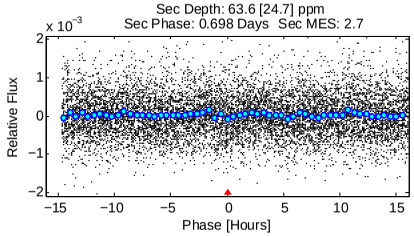
This Data Validation Report Summary was produced in the Kepler Science Operations Center Pipeline at NASA Ames Research Center

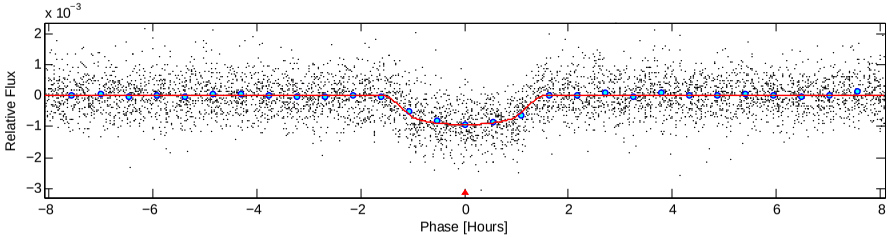
C.3 Object 1 - candidate 3

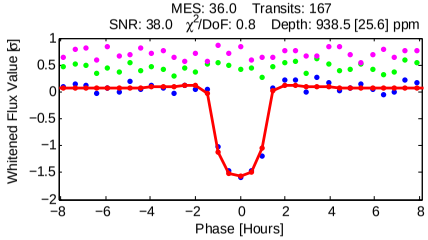
KIC: 10910878 Candidate: 3 of 3 Period: 6.253 d
KOI: K00757.03 Name: Kepler−229b Corr: 0.981

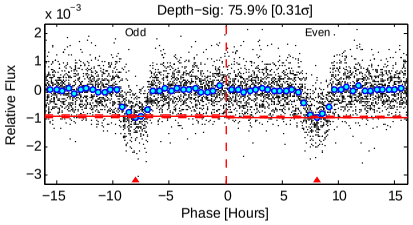


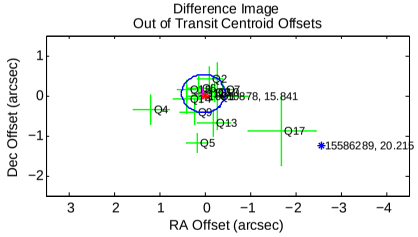












DV Fit Results:

Period = 6.25296 [0.00001] d
Epoch = 133.7825 [0.0013] BKJD
Rp/R* = 0.0310 [0.0089]
a/R* = 12.19 [12.09]
b = 0.77 [0.54]

Seff = 108.39 [16.72]
Teq = 823 [32] K
Rp = 2.87 [0.85] Re
a = 0.0617 [0.0053] AU

Ag = 16.24 [11.45] [1.33σ]
Teffp = 2554 [443] K [3.89σ]

DV Diagnostic Results:

ShortPeriod−sig: N/A
LongPeriod−sig: 100.0% [51.53σ]
ModelChiSquare2−sig: 100.0%
ModelChiSquareGof−sig: 100.0%
Bootstrap−pfa: 1.01e−271
RollingBand−fgt: 0.99 [153/154]
GhostDiagnostic−chr: 3.911

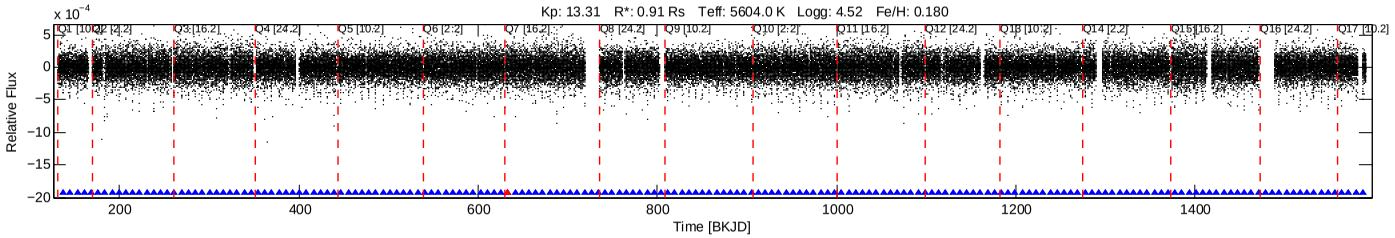
Centroid−sig: 23.6%
Centroid−so: 0.395 arcsec [1.16σ]
OotOffset−rm: 0.059 arcsec [0.37σ]
KicOffset−rm: 0.054 arcsec [0.42σ]
OotOffset−st: 4/4/1/5 [14]
KicOffset−st: 4/4/1/5 [14]
DiffImageQuality−fgm: 1.00 [14/14]
DiffImageOverlap−fno: 1.00 [14/14]

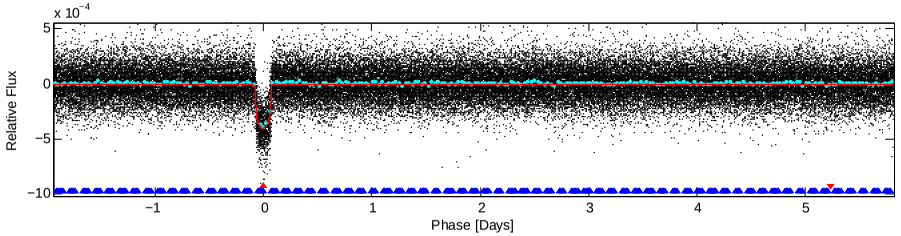
Software Revision: svn+ssh://murzim/repo/soc/tags/release/9.3.42@60958 −− Date Generated: 29−Jan−2016 03:41:32 Z

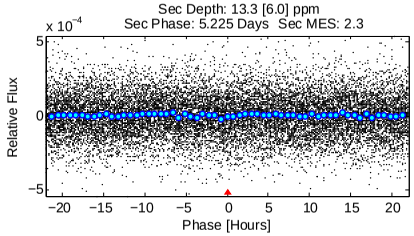
This Data Validation Report Summary was produced in the Kepler Science Operations Center Pipeline at NASA Ames Research Center

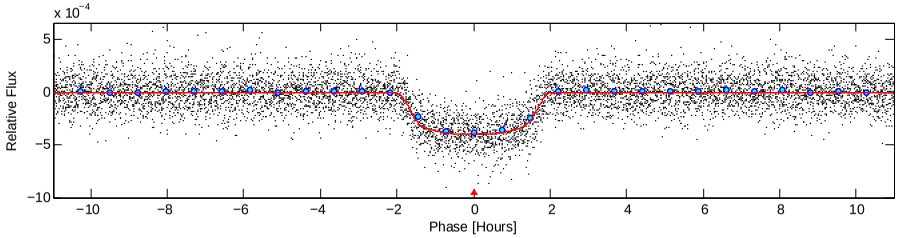
C.4 Object 2 - candidate 1

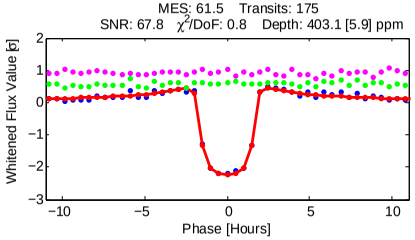
KIC: 4563268 Candidate: 1 of 2 Period: 7.752 d
KOI: K00627.01 Corr: 0.976

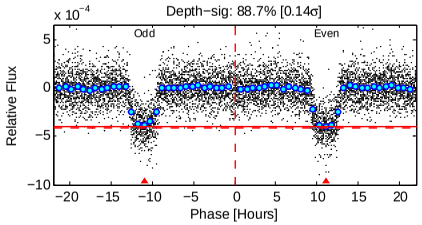


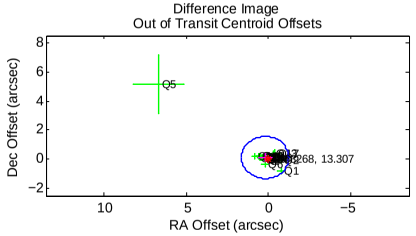












DV Fit Results:

Period = 7.75191 [0.00001] d
Epoch = 137.4145 [0.0009] BKJD
Rp/R* = 0.0210 [0.0018]
a/R* = 9.33 [3.40]
b = 0.84 [0.13]

Seff = 124.64 [26.74]
Teq = 852 [46] K
Rp = 2.09 [0.34] Re
a = 0.0768 [0.0097] AU

Ag = 9.82 [5.13] [1.72σ]
Teffp = 2332 [287] K [5.09σ]

DV Diagnostic Results:

ShortPeriod−sig: 100.0% [17.36σ]
LongPeriod−sig: N/A
ModelChiSquare2−sig: 99.9%
ModelChiSquareGof−sig: 100.0%
Bootstrap−pfa: 0.00e+00
RollingBand−fgt: 0.99 [166/167]
GhostDiagnostic−chr: 3.999

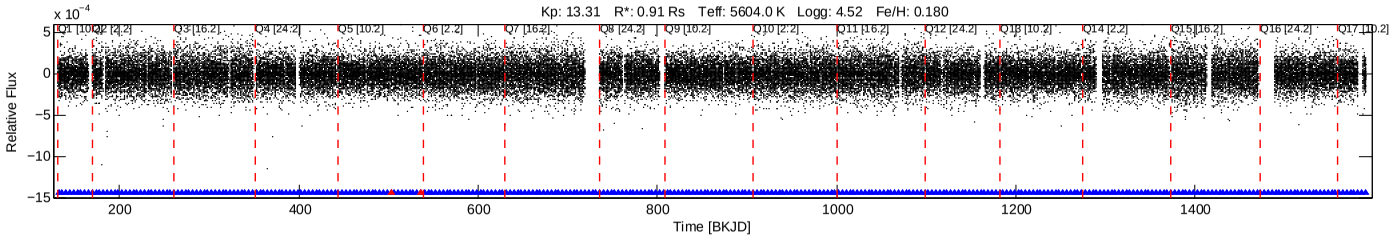
Centroid−sig: 0.0%
Centroid−so: 0.648 arcsec [2.24σ]
OotOffset−rm: 0.250 arcsec [0.52σ]
KicOffset−rm: 0.080 arcsec [0.26σ]
OotOffset−st: 4/4/4/5 [17]
KicOffset−st: 4/4/4/5 [17]
DiffImageQuality−fgm: 0.88 [15/17]
DiffImageOverlap−fno: 1.00 [17/17]

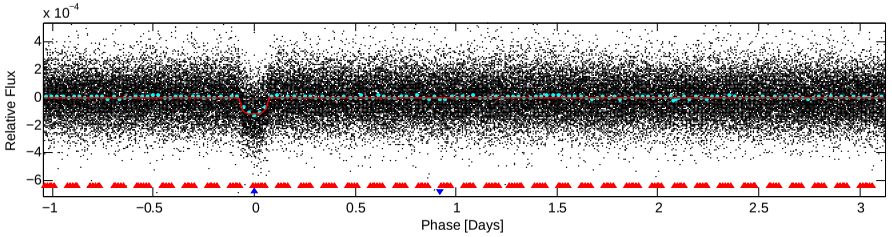
Software Revision: svn+ssh://murzim/repo/soc/tags/release/9.3.42@60958 −− Date Generated: 31−Jan−2016 03:55:58 Z

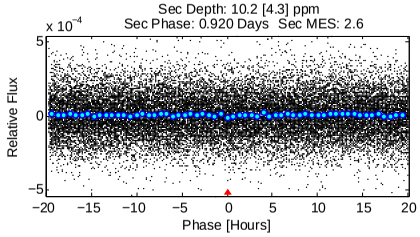
This Data Validation Report Summary was produced in the Kepler Science Operations Center Pipeline at NASA Ames Research Center

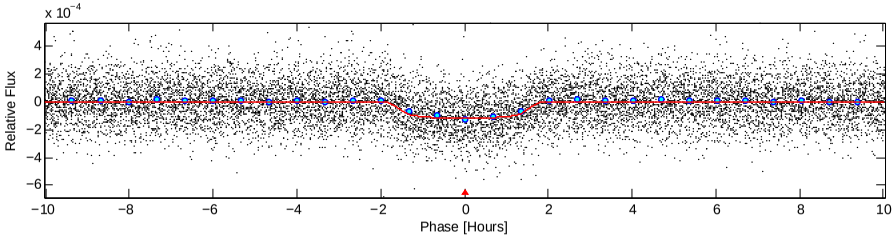
C.5 Object 2 - candidate 2

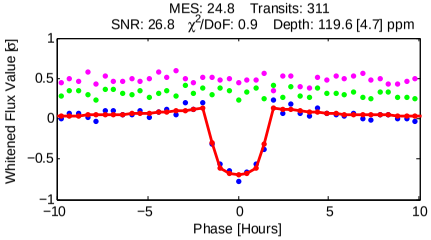
KIC: 4563268 Candidate: 2 of 2 Period: 4.165 d
KOI: K00627.02 Corr: 0.985

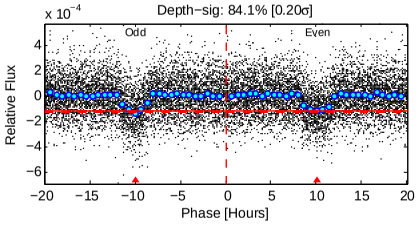


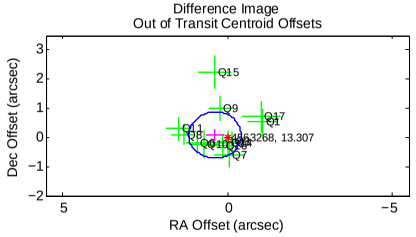












DV Fit Results:

Period = 4.16538 [0.00001] d
Epoch = 132.0379 [0.0018] BKJD
Rp/R* = 0.0118 [0.0028]
a/R* = 4.83 [4.83]
b = 0.88 [0.27]

Seff = 285.33 [61.20]
Teq = 1048 [56] K
Rp = 1.17 [0.32] Re
a = 0.0508 [0.0064] AU

Ag = 10.50 [6.97] [1.36σ]
Teffp = 2917 [467] K [3.97σ]

DV Diagnostic Results:

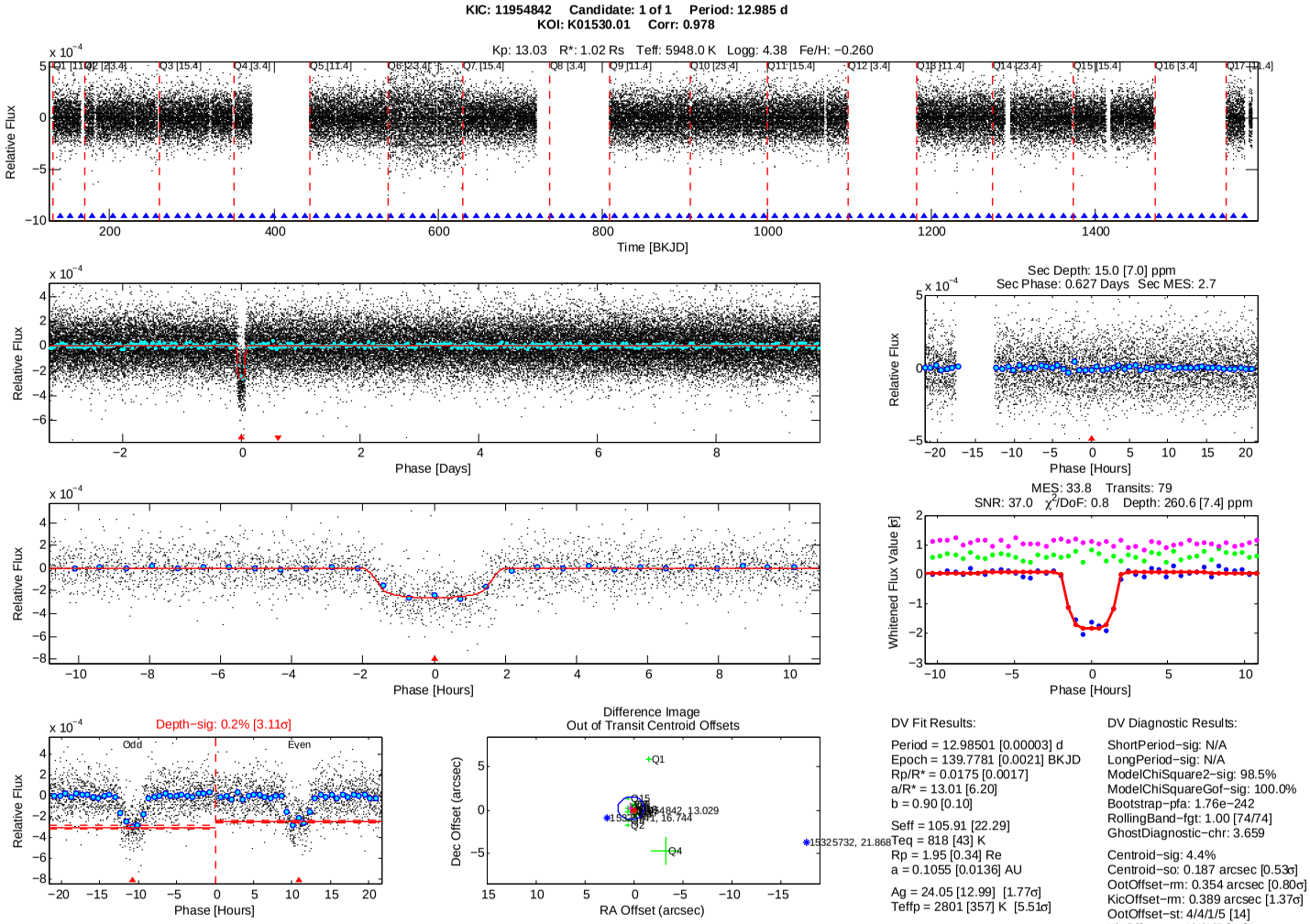
ShortPeriod−sig: N/A
LongPeriod−sig: 100.0% [17.36σ]
ModelChiSquare2−sig: N/A
ModelChiSquareGof−sig: N/A
Bootstrap−pfa: 8.59e−130
RollingBand−fgt: 0.99 [295/297]
GhostDiagnostic−chr: 5.671

Centroid−sig: 0.0%
Centroid−so: 0.380 arcsec [0.78σ]
OotOffset−rm: 0.388 arcsec [1.46σ]
KicOffset−rm: 0.068 arcsec [0.30σ]
OotOffset−st: 3/4/2/3 [12]
KicOffset−st: 3/4/2/3 [12]
DiffImageQuality−fgm: 0.75 [9/12]
DiffImageOverlap−fno: 1.00 [17/17]

Software Revision: svn+ssh://murzim/repo/soc/tags/release/9.3.42@60958 −− Date Generated: 31−Jan−2016 03:56:05 Z

This Data Validation Report Summary was produced in the Kepler Science Operations Center Pipeline at NASA Ames Research Center

C.6 Object 3 - candidate 1



DV Fit Results:

Period = 12.98501 [0.00003] d
Epoch = 139.7781 [0.0021] BKJD
Rp/R* = 0.0175 [0.0017]
a/R* = 13.01 [6.20]
b = 0.90 [0.10]

Seff = 105.91 [22.29]
Teq = 818 [43] K
Rp = 1.95 [0.34] Re
a = 0.1055 [0.0136] AU

Ag = 24.05 [12.99] [1.77σ]
Teffp = 2801 [357] K [5.51σ]

DV Diagnostic Results:

ShortPeriod−sig: N/A
LongPeriod−sig: N/A
ModelChiSquare2−sig: 98.5%
ModelChiSquareGof−sig: 100.0%
Bootstrap−pfa: 1.76e−242
RollingBand−fgt: 1.00 [74/74]
GhostDiagnostic−chr: 3.659

Centroid−sig: 4.4%
Centroid−so: 0.187 arcsec [0.53σ]
OotOffset−rm: 0.354 arcsec [0.80σ]
KicOffset−rm: 0.389 arcsec [1.37σ]
OotOffset−st: 4/4/1/5 [14]
KicOffset−st: 4/4/1/5 [14]
DiffImageQuality−fgm: 0.86 [12/14]
DiffImageOverlap−fno: 1.00 [14/14]

Software Revision: svn+ssh://murzim/repo/soc/tags/release/9.3.42@60958 −− Date Generated: 01−Feb−2016 00:53:04 Z

This Data Validation Report Summary was produced in the Kepler Science Operations Center Pipeline at NASA Ames Research Center

D R code: hypothesis testing on random data set

```
# Number of iterations
k <- 150

#Packages needed:
require(zoo)

# Function to calculate number of (false) rejections:
# variation_alpha(number of iterations, sample size, lengths,
# strengths, sigmas, significance level)
variation_alpha <- function(k,n,L,S,sigma,ct){
  #Create empty vectors
  udr <- vector("integer", k)
  udfr <- vector("integer", k)
  fdr <- vector("integer", k)
  fdfr <- vector("integer", k)

  #for-loop for calculation of mean (falsely) rejected data
  for (i in 1:k){
    #Producing data:
    mu<-c(rep(0,L),rep(S,L),rep(0,(n-2*L)))#vector of mean values
    Y<- mu+rnorm(n,0,sigma) #Observation vector;
                            #correspnds to Vectors Vi of data files
    c<- qnorm(1-ct) # Determining threshold
    #rnorm command to simulate random draws from normal distribution;
    #also avalable for other distributions, e.g. rbinom, runif,....

    #Filter the data:
    FY<-1/(sqrt(F)*sigma)*rollapply(Y,F,sum) # Filtered (matched
                                             # filter) and normalized
                                             # data
    Fmu<-1/(sqrt(F)*sigma)*rollapply(mu,F,sum) # Filtered mu-vector

    ####Evaluation
    sum(Y>c)# number of rejections, unfiltered data
    sum(Y[mu==0]>c) # number of FALSE REJECTIONS, unfiltered data
    sum(FY>c)# number of rejections, filtered data
    sum(FY[Fmu==0]>c) # number of FALSE REJECTIONS, unfiltered data

    udr[i]<-sum(Y>c)
    udfr[i]<-sum(Y[mu==0]>c)
    fdr[i]<-sum(FY>c)
    fdfr[i]<-sum(FY[Fmu==0]>c)
    i <- i+1
  }
  udr_total <- sum(udr)/length(udr)
  udfr_total <- sum(udfr)/length(udfr)
  fdr_total <- sum(fdr)/length(fdr)
```

```
  fdfr_total <- sum(fdfr)/length(fdfr)

  print(udr_total)
  print(udfr_total)
  print(fdr_total)
  print(fdfr_total)
}

#Standard settings for parameters
n1 <- 250
L1 <- 5
S1 <- 2
sigma1 <- 1
c_ch1 <- 0.05

# Paramaters that change
n_ch <- c(50,100,150,200,250,300) #different sample sizes
L_ch <- c(3,5,7,9,11) #different signal lengths
S_ch <- c(1,2,3,4,5) #different signal strengths
sigma_ch <- c(1, 1.25, 1.5, 1.75, 2) #different sigmas
c_ch <- c(0.0001, 0.005, 0.01, 0.05) #different significance levels

# Displaying results of the hypothesis test that is performed per
# variable
for (i in c_ch){
  for (l in n_ch){
    print("#(false) rejections for sample size")
    print(l)
    variation_alpha(k,l, L1, S1, sigma1, i)
  }
}

for (i in c_ch){
  for (m in L_ch){
    print("#(false) rejections for signal length")
    print(m)
    variation_alpha(k,n1, m, S1, sigma1, i)
  }
}

for (i in c_ch){
  for (s in S_ch){
    print("#(false) rejections for signal strength")
    print(s)
    variation_alpha(k,n1, L1, s, sigma1, i)
  }
}

for (i in c_ch){
```

```
  for (x in sigma_ch){
    print("#(false) rejections for sigma")
    print(x)
    variation_alpha(k,n1, L1, S1, x, i)
  }
}
```

E R code: hypothesis testing on Kepler data set

```
    #Packages needed:
require(zoo)

KeY <- read.delim('Kepler.txt', header = TRUE, dec = ".", sep=" ")
#Observation vector; correspnds to Vectors Vi of data files

# Parameters which you can vary. (If you set S<-0 you get the
# situation with no signal at all.):
n<-length(KeY\$V1) # Sample size: Should be at least 3*L
L<-5 # Length of signal
F<-L # Length of window in filter, here set equal to L,
     # but could also be different
S<-9 # Strength of the signal
sigma<-sd(KeY\$V1) # Standard deviation of the normally distributed
                   # random variables
c<- qnorm(1-0.0001) # Critical value used

#Producing data:
mu<-c(rep(0,L),rep(S,L),rep(0,(n-2*L))) #vector of mean values
Yinv <- KeY\$V1
Y <- Yinv * -1

#rnorm command to simulate random draws from normal distribution;
#also avalable for other distributions, e.g. rbinom, runif,....
#Filter the data:
FY<-1/(sqrt(F)*sigma)*rollapply(Y,F,sum) # Filtered (matched filter)
                                        # and normalized data
Fmu<-1/(sqrt(F)*sigma)*rollapply(mu,F,sum) # Filtered mu-vector

#Plots:
plot(FY, ylim=c(min(c(Y,FY)), max(c(Y,FY))), main="",pch=15,
xlim=c(0,n),col="red")
points(Y,pch=16)
points(rep(c,n),type="l",lty=2,lwd=2)

lines(FY,col="red")
lines(Y)

####Evaluation
sum(Y>c)# number of rejections, unfiltered data
sum(Y[mu==0]>c) # number of FALSE REJECTIONS, unfiltered data
sum(FY>c)# number of rejections, filtered data
sum(FY[Fmu==0]>c) # number of FALSE REJECTIONS, filtered data
```