Period Search
Software Requirement Specification

prepared by: Jan Cuypers
approved by: L. Eyer
reference: GAIA-C7-SP-ROB-JCU-002-4
issue: 4
revision: 0
date: 2012-07-09
status: Issued

Abstract

In this document, we present the Software Requirements Specifications for the period search. The result of the period search should be the most probable period(s) present in the time series of data with an indication of its (their) significance. The latter could also mean that no periodicity is present (anymore). It is part of the **Variability Characterisation** CU7 functional component.

Document History

Issue	Revision	Date	Author	Comment
4	0	2012-07-09	ILT	Issue for Livelink: references fixed, approval and status updated
3	4	2012-06-14	JCU	Only minor corrections
3	3	2012-05-11	JCU	References to weighted and other formulae - Mantis 0012068 and minor corrections
3	2	2011-08-16	JCU	Frequency error combination described - Mantis 0010293
3	1	2011-07-01	JCU	Final minor corrections
3	0	2011-06-27	LPG	Major release for cycle 11
2	3	2011-06-24	JCU	Typos corrected, merging section removed, minor changes
2	2	2011-01-04	ILT	Verification method of requirements set to AUT
2	1	2010-07-08	JCU	First changes for new issue, error on frequency
1	8	2010-03-12	ILT	Requirements identifiers updated (CU7 added)
1	7	2010-03-11	ILT	Version number added to requirements
1	6	2009-07-17	LPG	Corrections after review
1	5	2009-02-24	JCU	Final comments added
1	4	2009-02-17	LPG	Added new activity diagram and traceability
1	3	2009-02-17	JCU	Comments of review 2 incorporated
1	2	2009-02-16	LPG	Removed MDB direct references and proof-read
1	1	2009-02-13	JCU	First Issue with first ideas on significance and probabilities
D	1	2007-12-13	JCU	Fifth draft, merged with the SRS documents on the individual methods
D	1	2007-09-25	JCU	Fourth draft
D	1	2007-08-05	JCU	Third draft
D	1	2007-04-13	JCU	Second draft
D	1	2007-04-10	JCU	First draft

Contents

1 Introduction

1.1 Objectives

In this document, we present software specifications for the general period search. The goal is to detect possible periods with significance levels in series of Gaia measurements of either magnitudes, fluxes, or radial velocities.

The time sampling of Gaia measurements is uneven because of the scanning law of the spacecraft. Therefore, classical methods of period analysis of variable stars will be used to detect periodicities. These methods are based on different assumptions and some could be more efficient than others in finding periodicities in certain classes of variable stars (e.g. pulsational variations versus eclipse phenomena). The methods may also differ in the ability to select a true period amongst the aliases caused by the peculiar time coverage of the Gaia data. For both reasons, it is important to investigate and implement different methods for period search, as will be detailed further.

The software required here will not only have to select the most probable period(s) present in the time series or give an indication that no periodicity is present, but also the possibility should exist to compare the results of the different period search methods. This could be based on the probability associated with the most probable period.

1.2 Scope

This document covers the requirements to extract period information from time series.

1.3 Assumptions

This specification assumes that the input source time series (of magnitudes, fluxes, or radial velocity), at least calibrated on a relative scale, are readily available in the system

The period search will be performed for the sub-set of objects which have been detected as variable and some other TBD classes. There may be additional constraints applied (on brightness for example) to further reduce the size of the processed object set, depending on the method performance (see also Sect. 2.3.1).

1.4 Applicable Documents

AL-001	Special Variability Detection Software Requirements Specifiction
CA-002	Statistical Parameter Determination Software Requirement Specification
JCU-007	Variability Characterisation Software Requirement Specification
JCU-008	Period Search Significance calculation (Part 1)
JCU-011	On the computation of the Deeming, Lomb-Scargle, Harmonic Least Squares and related periodograms
LE-005	CU7 Software Development Plan
LPG-002	Time Series Modeling Software Requirement Specification
PD-006	CU7 Top-level Functional Analysis and Software Requirements Specification
TL-001	DPAC Product Assurance Plan
WOM-011	Software Engineering Guidelines for DPAC

1.5 Reference Documents

[JCU-011], Cuypers, J., 2012, *On the computation of the Deeming, Lomb-Scargle, Harmonic Least Squares and related periodograms*,
GAIA-C7-TN-ROB-JCU-011,
URL http://gaia.esac.esa.int/dpacsvn/DPAC/CU7/docs/characterisation/technotes/GAIA-C7-TN-ROB-JCU-011/GAIA-C7-TN-ROB-JCU-011.pdf

[JCU-007], Cuypers, J., Guy, L., 2011, *Variability Characterisation - Software Requirement Specification*,
GAIA-C7-SP-ROB-JCU-007,
URL http://www.rssd.esa.int/llink/livelink/open/2913936

[JCU-008], Cuypers, J., Rimoldini, L., 2012, *Period Search Significance calculation (Part 1*,
GAIA-C7-TN-ROB-JCU-008,
URL http://gaia.esac.esa.int/dpacsvn/DPAC/CU7/docs/characterisation/technotes/GAIA-C7-TN-ROB-JCU-008/GAIA-C7-TN-ROB-JCU-008.pdf

[CA-002], De Ridder, J., 2011, *Statistical Parameter Determination - Software Requirement Specification*,
GAIA-C7-SP-IVS-CA-002,
URL http://www.rssd.esa.int/llink/livelink/open/2913937

[PD-006], Dubath, P., Lecoeur-Taibi, I., Mowlavi, N., et al., 2011, *CU7 Top-Level Functional Analysis and Software Requirements Specification*,
GAIA-C7-SP-GEN-PD-006,
URL http://www.rssd.esa.int/llink/livelink/open/2786539

[LE-005], Eyer, L., Lecoeur, I., Beck, M., et al., 2011, *CU7 Software Development Plan*, GAIA-C7-PL-GEN-LE-005, URL http://www.rssd.esa.int/llink/livelink/open/2786584

[LPG-002], Guy, L., De Ridder, J., 2011, *Variability Characterization - Time Series Modelling SDD*, GAIA-C7-SP-GEN-LPG-002, URL http://www.rssd.esa.int/llink/livelink/open/3083180

[AL-001], Lanzafame, A., 2006, *Specific object studies - Solar-like variability (magnetic activity) and flare stars*, GAIA-C7-TN-UNCT-AL-001, URL http://www.rssd.esa.int/llink/livelink/open/2720497

[TL-001], Levoir, T., Damery, J., Hoar, J., et al., 2010, *DPAC Product Assurance Plan*, GAIA-C1-PL-CNES-TL-001, URL http://www.rssd.esa.int/llink/livelink/open/2439085

[WOM-011], O'Mullane, W., Hoar, J., Levoir, T., et al., 2011, *Software Engineering Guidelines for DPAC*, GAIA-C1-UG-ESAC-WOM-011, URL http://www.rssd.esa.int/llink/livelink/open/2760364

Baliunas et al., 1985, Ap. J. 294, 310
Burke et al., 1970, J. Roy. Astron. Soc. Canada 64, 353
Cuypers, J., 1987, Acad. Analecta, 49(3)
Deeming, T.J., 1975, Astrophys. Space Sci. 63, 137
Dworetsky, M.M., 1983, MNRAS 203, 917
Gilliland, R.L., Fischer, R., 1985, PASP 97, 285
Jurkevich, I., 1971, Astrophys. Space Sci. 13, 154
Kuiper, N.H., 1960, Proc. of the Koninkl. Nederl. Akad. van Wet. Ser. A. 63, 38
Kovacs, G., 1981, Astrophys. Space Sci. 78, 175
Lafler, J., Kinman, T.D., 1965, Ap. J. Suppl. 11, 216
Lomb, N.R., 1976, Astrophys. Space Sci 39, 447
Renson, P., 1978, A&A. 63, 125
Scargle, J.D., 1982, Ap. J. 263, 835
Stellingwerf, R.F., 1978, Ap. J. 224, 953

1.6 Definitions, acronyms, and abbreviations

The following is a list of acronyms used in this document.

Acronym	Description
CU	Coordination Unit (in DPAC)
DPAC	Data Processing and Analysis Consortium
DPC	Data Processing Centre
FAP	False Alarm Probability
HLSQ	Harmonic Least Squares
MDB	Main DataBase
PDM	Phase Dispersion Minimisation
SADT	Structured Analysis and Design Technique
SRS	Software Requirement Specification
TBD	To Be Defined (Determined)
UML	Universal Modeling Language
WP	Work Package

2 General description and requirements

2.1 Context

From the functional point of view, the Period Search is part of the **variability characterisation** component, which groups all tasks "characterizing" (i.e., "describing", "quantifying" and"modelling") the variability behaviour, as illustrated in Fig. 1.

- **Inputs:** time series of Gaia measurements, available from the CU7 variability database, that have been identified (and flagged) as variable. Magnitude will mainly be used, but it will also be possible to use other measurements from the spectrophotometry, or even radial velocities from CU6.
- **Outputs:** the most probable period(s) with their significance levels and error estimates.
- **Control parameters:** there are a number of parameters (minimum and maximum trial frequency, frequency steps and resolution etc.) in this method which depends mostly on the sampling of the time series and on the S/N ratio of the measurements. In principle, the best parameter values for the Gaia case can be derived from further tests and simulations. The goal is then to fix all parameters to optimum values and to end up with a method without any free parameter (see further).
- **Input producer and output consumer:** The CU7 variability database for the input time series. The output period will go to the (periodic) model fitting if considered significant. Output periods will further be used in various subsequent processing, such as in components of **Classification** and **Specific Object Studies** and stored in the CU7 variability database.

2.2 Decomposition

Several methods for period analysis have to be implemented. The choice of the optimal set of method(s) is an area of ongoing investigation and optimization. The candidate methods are described below as separate modules. Initially for each method a SRS document existed, but now these documents are merged.

Since several methods need the same initialisations (normalization, frequency start, end, step. . .) the first modules will supply these values.

The general scheme of the methods is to compute a periodogram (or a *frequencygram*) for a list of (N_f) periods/frequencies. These data are stored in a FrequencyGramme, if necessary. The

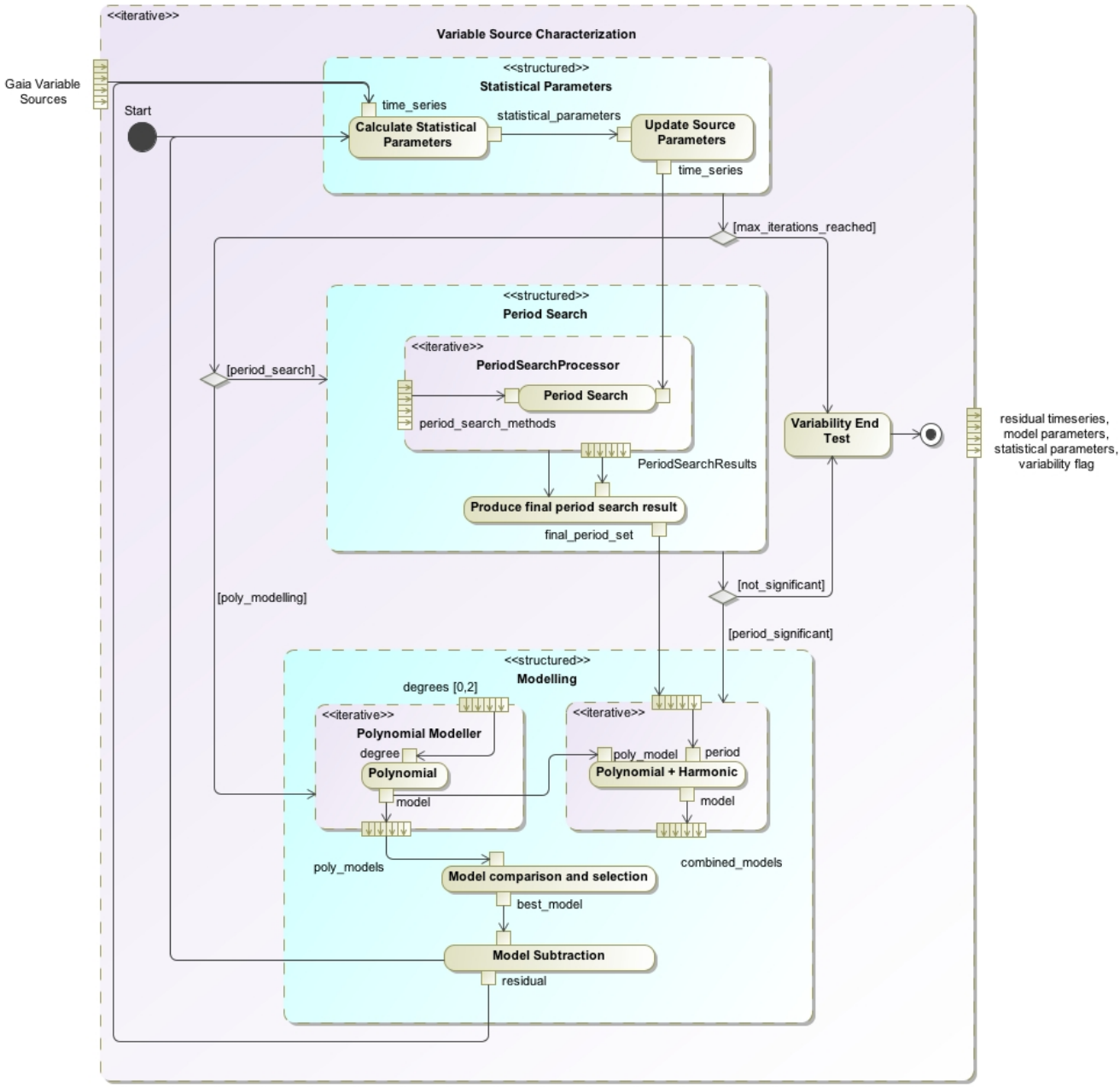

FIGURE 1: Context diagram

next step is to identify the highest/lowest (depending on the method) independent peak(s) in the FrequencyGramme and store them in a period set. The highest/lowest peak(s) provide the most probable period(s). If possible, each of the probable periods should have an associated significance level and an error estimate.

A first option is to apply several methods to all time series. The results of all the methods should then be merged to produce the final set of output periods. A strategy for the merging has to be searched for, if some extra efficiency is expected. Another option to be considered, is to apply only the most efficient (TBD) period search method on all time series. If a period resulting from this method, is highly significant (threshold TBD), the period search will stop and pass this period and its significance as output. If the period is less significant (value TBD), other period search methods could be applied to see if they can produce a (more) significant result. Tests will indicate if any of the methods are more suitable for certain classes of variable objects than the most efficient method(s) or if statistical or other parameters give an indication to use another method as well. First results of the tests gave no indications of methods being more efficient for some classes of variable stars only, but the conclusions of the tests are not finalized yet. The gain in computation time for the second option has to be evaluated in view of possible disadvantages, as finding the wrong period etc.

2.3 Requirements

2.3.1 Performance

The software performance is an important issue as we may want to apply this method to the $\approx 10^8$ variable Gaia objects. Great care will have to be taken to optimize all steps of this method, and it may be that we will have to apply further criteria to reduce the number of objects to be processed. For example, tests will be carried out to see whether we can define a magnitude threshold below which the S/N will be too low to get any meaningful periods.

2.3.2 Re-usability

Parts could be of use elsewhere (TBD).

2.3.3 Implementation Constraints

None

2.3.4 Other Non Functional Requirements

None

3 Modules

3.1 Normalising, folding and sorting of the time series

A time series is formed by a set of observation times t_i with associated measured quantities x_i, such as the magnitudes, the fluxes, colours or radial velocities. If we have the epochs of the observations as $t_i, i = 1, \ldots, n$ with n as the number of epochs, and if t_0 is taken as the reference epoch, the phases ϕ_{ij} for the period P_j or frequency f_j are:

$$\phi_{ij} = \left[\frac{t_i - t_0}{P_j}\right] = [(t_i - t_0)f_j]\,; i = 1, \ldots, n; j = 0, \ldots, N_f$$

where the square brackets indicate the decimal part. The phases are dimensionless since they are a ratio of times, and $0 \leq \phi_i < 1$. We will use the name folded time series for a time series with phases computed for period P_j or frequency f_j.

It is not necessary to use the expression given above for the calculation of the phase of every trial period/frequency. Indeed, because each trial frequency, in our case, can be written as $f_j = f_0 + j\Delta f (j = 0, \ldots, N_f)$ with f_0 the minimum trial frequency of the considered interval and Δf the frequency step, it is possible to rewrite the phase for trial frequency f_j as:

$$\phi_{ij} = \phi_{i0} + j\Delta\phi_i; i = 1, \ldots, n; j = 0, \ldots, N_f$$

with ϕ_{i0} the phase of t_i for f_0 and $\Delta\phi_i$ the phase for Δf. In this case, ϕ_{i0} and $\Delta\phi_i$ only need to be calculated once, to be able to calculate all other phases ϕ_{ij}. It should be checked whether this approach of calculating the phases results in a significant gain in computation time or not.

It might be preferable, for efficient calculation purposes only, that the time series are shifted so that $t_1 = t_0$. For some methods it might be necessary to sort the resulting folded time series according to increasing phase, rearranging the phase and all values (magnitude, fluxes, radial velocities) of the time series in the same way. It could also be of use (TBD for which or for all methods) that the x-values are 'normalized'. This could mean e.g. that the $\sum_{i=1}^{N} x_i^2 = 1$.

3.1.1 Requirements

CU7-WP711-02000-S-FUN-020	1.0	AUT	Issued
It shall be possible to fold a time series around a provided period and to sort the resulting folded time series according to increasing phase.			
Parent: CU7-WP711-00000-S-FUN-120			

3.1.2 Input / Output

Name	Description	Access Type
LightCurve		Input
FoldedLightCurve	A folded lightcurve for a given frequenncy	Output

3.2 Choice of the interval for the trial frequencies and the frequency step

The minimum and the maximum of the trial frequencies (or periods) can be derived from the input time series. At the end of the mission, it is expected that most of the time series will be almost of the same length and about the lifetime of the Gaia satellite. For simplicity we take this (optimistically) as $T = 2000$ days. This means that the frequency step has to be at least $1/5T \approx 0.0001d^{-1}$ in order to be sure that a peak in the frequencygram is well enough sampled.

The maximum trial period (or minimum frequency) should be no longer than half of the total observation time span. In our approximation this will about 1000 days or 0.001 d^{-1} in frequency.

The maximum frequency (or minimum period) is, at the moment, only limited by considerations of computation speed. If required, an upper limit can be established in function of expected variability type and data noise. Further work is required to derive a meaningful minimum period in the case of Gaia measurements. In view of the limited amount of variable stars with very short periods (high frequencies) and the lower period limit of some classes (e.g. δ Scuti stars), an upper limit of $30d^{-1}$ in frequency could be set. The Special Variability Detection can set a higher upper limit for a selection of stars (see AL-001).

For the final runs (on the complete datasets) the set of trial frequencies will come from the interval $[0.001, 30]d^{-1}$ with a step $\Delta f = 0.0001d^{-1}$ (from about $0.01\ to\ 300\ \mu Hz$ with a step of $0.001\ \mu Hz$). This will result in $N_f = 299\,990$ frequencies to scan in total.

Additional work is required to see if a different step for each time series has some advantages and to see if the upper frequency limit should be varied for each time series as well. If the answer to one of this questions is yes, a more dynamical approach to calculate the minimum and maximum frequency for the frequency search can be used. E.g. starting from $2/T$, but the upper end will still have to be set, since the unequidistant time sampling of the Gaia data will not allow to define a classical Nyquist frequency. If no, the appropriate values for the interval and the frequency step should be calculated once and kept fixed for all stars and methods. For partial results of the Gaia satellite, e.g. after one year ($T = 365d$), the step can be made appropriately larger if desired: $\Delta f = 1/5Td^{-1}$ or $\Delta f = 1/10Td^{-1}$

3.2.1 Requirements

CU7-WP711-02000-S-FUN-040	1.0	AUT	Issued
It shall be possible to derive a vector of appropriate trial frequencies from a time series.			
Parent: CU7-WP711-02000-S-FUN-020			

3.2.2 Input / Output

Name	Description	Access Type
LightCurve	Lightcurve	Input
FrequencySet	A set of frequencies to be used for period analysis	Output

3.3 Period Search Methods

3.3.1 Deeming method

The period analysis method as the one described by Deeming (1975) needs for each frequency of the list $f_j, j = 0, \ldots, N_f$ the quantity

$$P_N(f_j) = \frac{1}{N}\left\{\left(\sum_{i=1}^{N} x_i \sin 2\pi f_j t_i\right)^2 + \left(\sum_{i=1}^{N} x_i \cos 2\pi f_j t_i\right)^2\right\}.$$

The sine and cosine functions in this expression can be calculated in a recursive way if the frequency step $\Delta f = f_{j+1} - f_j$ used for the frequencygram is constant. In that case, a recursive scheme can be set up:

$$x_i \sin 2\pi f_{j+1} t_i = x_i \cos 2\pi \Delta f t_i \sin 2\pi f_j t_i + x_i \sin 2\pi \Delta f t_i \cos 2\pi f_j t_i$$

$$x_i \cos 2\pi f_{j+1} t_i = x_i \cos 2\pi \Delta f t_i \cos 2\pi f_j t_i - x_i \sin 2\pi \Delta f t_i \sin 2\pi f_j t_i$$

where the $x_i \cos 2\pi \Delta f t_i$ and $x_i \sin 2\pi \Delta f$ can be calculated before and stored.

The method can be optimized in several ways. Based on suggestions in several articles in literature, Cuypers (priv. communication) has developed a fast algorithm for this method. It has not been tested in all situations, but, if intermediate storage is not a problem, and the performance in Java is comparable to other computer languages, this algorithm is to be preferred. Details about the calculation and the version using weights can be found in the note JCU-011.

How to calculate the false alarm probability for this method is given in the paper JCU-008.

3.3.1.1 Requirements

CU7-WP711-02000-S-FUN-060	1.0	AUT	Issued
It shall be possible to compute the Deeming periodogram for a time series for a list of frequencies.			
Parent: CU7-WP711-00000-S-FUN-120, CU7-WP711-02000-S-FUN-040			

3.3.1.2 Input/Output

Name	Description	Access Type
LightCurve	Lightcurve	Input
FrequencySet	A set of frequencies to be used for period analysis	Input
FrequencyGramme	The Deeming periodogram	Output

3.3.2 Lomb-Scargle method

Lomb (1976) and Scargle (1982) described how the expression for the classical periodogram for an equidistantly sampled and uninterrupted time series could be applied to the typical non-equidistant and gapped time series of astronomical observations and made equal to linear least-squares fitting to the model

$$Asin(2\pi ft) + Bcos(2\pi ft).$$

The method can be considered as an extension of the Deeming method and, as a consequence, a similar optimized algorithm exists. Here, for each of frequency of the list $f_j, j = 0, \ldots, N_f$ the following quantity has to be computed:

$$P'_N(f_j) = \frac{1}{2}\left\{\frac{\{\sum_{i=1}^{N} x_i \sin 2\pi f_j(t_i-\tau)\}^2}{\sum_{i=1}^{N}\sin^2 2\pi f_j(t_i-\tau)} + \frac{\{\sum_{i=1}^{N} x_i \cos 2\pi f_j(t_i-\tau)\}^2}{\sum_{i=1}^{N}\cos^2 2\pi f_j(t_i-\tau)}\right\}$$

with

$$\tan 4\pi f_j\tau = \frac{\sum_{i=1}^{N}\sin 4\pi f_j t_i}{\sum_{i=1}^{N}\cos 4\pi f_j t_i}.$$

The offset τ makes $P'_N(f_j)$ independent of shifting all t_i by any constant. Therefore, any date t_0 can be used and complete equivalence by linear least-squares fitting to the model

$$Asin(2\pi ft) + Bcos(2\pi ft)$$

is assured.

It can be shown (Cuypers, private communication) that the quantity $P'_N(f_j)$ can be modified in such a way that no separate and a priori computation of τ is necessary. In general, this results in a considerable gain in computation speed. More about this and how to use weights in this method can be found in the note JCU-011.

The correct significance level (or false alarm probability) of the frequencies extracted by this method, has been the subject of several papers. The details on the false alarm probability and its calculation can be found in paper JCU-008.

3.3.2.1 Requirements

CU7-WP711-02000-S-FUN-080	1.0	AUT	Issued
It shall be possible to compute the Lomb-Scargle periodogram for a time series for a list of frequencies.			
Parent: CU7-WP711-00000-S-FUN-120, CU7-WP711-02000-S-FUN-040			

3.3.2.2 Input/Output

Name	Description	Access Type
LightCurve	Lightcurve	Input
FrequencySet	A set of frequencies to be used for period analysis	Input
FrequencyGramme	The Lomb-Scargle periodogram	Output

3.3.3 Harmonic Least Squares (HLSQ) Method

Here the quantity to be computed for each frequency of the list $f_j, j = 0, \ldots, N_f$ is directly related to the sum of the squared residuals after fitting of the model:

$$R(f) = \sum_{i=1}^{N} \{x_i - x_i^c(f_j)\}^2$$

where

$$x_i^c = \bar{a} \cos 2\pi f_j (t_i - \tau) + \bar{b} \sin 2\pi f_j (t_i - \tau) + \bar{c}.$$

This is equivalent to searching for the maximum of

$$\Delta R(f_j) = \sum_{i=1}^{N} x_i^2 - R(f_j)$$

or (after the indicated normalization):

$$\Delta R(f_j) = 1. - R(f_j).$$

Remark that this only differs from the Lomb-Scargle method by the constant in the fit. This means that the same recursive methods and optimized algorithms can be used here as well. With some extra computation higher harmonics (2f, 3f, ...) can be included, if necessary. How to optimize the calculations and the introduction of weights is described in JCU-011. The calculation of the false alarm probability for this method is in JCU-008.

3.3.3.1 Requirements

CU7-WP711-02000-S-FUN-100	1.0	AUT	Issued
It shall be possible to compute the HLSQ periodogram for a time series for a list of frequencies.			
Parent: CU7-WP711-00000-S-FUN-120, CU7-WP711-02000-S-FUN-040			

3.3.3.2 Input/Output

Name	Description	Access Type
LightCurve	Lightcurve	Input
FrequencySet	A set of frequencies to be used for period analysis	Input
FrequencyGramme	The HLSQ periodogram	Output

3.3.4 String Length methods

A simplified version of these methods was introduced by Lafler and Kinman (1965). They stated that, for the period present in the data, the quantity:

$$\theta_{LK} = \frac{\sum_{i=1}^{N}(x_{i+1} - x_i)^2}{\sum_{i=1}^{N}(x_{i+1} - \bar{x})^2}$$

where $x_{N+1} = x_1$ and $\bar{x}$ is the mean of the x_i, is minimal. Because

$$\theta_{LK} = \frac{\sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i x_{i+1}}{\sum_{i=1}^{N}(x_{i+1} - \bar{x})^2}$$

only $X = \sum_{i=1}^{N} x_i x_{i+1}$ has to be calculated and maximized.

Burke, Rolland and Boy (1970), and others later on, minimized the length of the connection line between two successive points in the phase diagram:

$$\theta_B = \sum_{i=1}^{N}\{(x_{i+1} - x_i)^2 + (\phi_{i+1} - \phi_i)^2\}^{1/2}$$

Dworetsky (1983) called this a string length method. The quantities x_i should be dimensionless, e.g. scaled with $x_{max} - x_{min}$ as unity in order to give measurements and phase differences the same weights, but other options for the scaling are possible.

Some variants of this methods were introduced by Renson (1978), e.g.

$$\theta_1 = \frac{\sum_{i=1}^{N}(x_{i+1} - x_i)^2}{(\phi_{i+1} - \phi_i) + \epsilon}$$

where $\phi_{N+1} = \phi_1 + 1$ and ϵ equals the possible phase difference of measurements within the observational error.

These methods have not been used very often, but can be considered as reliable. Some were also rather well performing on the periodic variables of the Hipparcos in the tests done.

Major drawbacks of these methods are the uncertainty of the choice of the scaling, the need for a estimate of ϵ and the slow computational time. There are not many ways to optimize this method in function of performance, since for every trial frequency the data have to be sorted according to phase and since phases and (normalized) data have to be added, no simplifications in the formulae are straightforward. Good estimates of false alarm probabilities are not available at the moment.

3.3.4.1 Requirements

CU7-WP711-02000-S-FUN-120	1.0	AUT	Issued
It shall be possible to compute the String Length periodogram for a time series for a list of frequencies.			
Parent: CU7-WP711-00000-S-FUN-120, CU7-WP711-02000-S-FUN-040			

3.3.4.2 Input/Output

Name	Description	Access Type
LightCurve	Lightcurve	Input
FrequencySet	A set of frequencies to be used for period analysis	Input
FrequencyGramme	The String length periodogram	Output

3.3.4.3 Input / Output

3.3.5 Jurkevich-Stellingwerf method

In this methods (Jurkevich, 1971; Stellingwerf, 1982) the phases constructed for each frequency of the list $f_j, j = 0, \ldots, N_f$ are divided into M parts, called bins. In general, the best results are found with an odd number of bins. Moreover, the higher the number of bins, the bigger the chance of the occurrence of empty bins in case of a relative low number of data points (as will be the case for the Gaia data sets). Therefore, M = 5 or 7 are expected to be the optimal choice.

The Jurkevich statistics θ_J is defined as the ratio of the sum of the dispersions calculated for the data in each bin separately to the global dispersion of the data. By definition, the lower θ_J, the higher the probability that the tested period is present in the data. In view of the re-usability of algorithms, it might be preferential to store the (1-θ_J) values instead of the θ_J values.

The Jurkevich statistic compares the sum of the individual dispersions of each phase bin to the global dispersion of the whole time series, and is defined as:

$$\theta_J = \frac{V_M^2}{N - M} \Big/ \frac{V^2}{N - 1}$$

where:

$$V_M^2 = \sum_{j=1}^{M}\sum_{i=1}^{N_k}(x_{ik} - \bar{x}_k)^2 \quad with \quad \bar{x}_k = \sum_{i=1}^{N_k} x_{ik}/N_k$$

$$V^2 = \sum_{k=1}^{M}\sum_{i=1}^{N_k}(x_{ik} - \bar{x})^2 \quad with \quad \bar{x} = \sum_{i=1}^{N} x_i/N$$

and N_k denotes the number of measured quantities x_{ik} within bin k ($k = 1, \ldots, M$).

If we additionally assume that the measured quantities are transformed such that $\bar{x} = 0$ and that they are 'normalized' such that $\sum_{i=1}^{N} x_i^2 = 1$, then the expressions can be simplified considerably resulting in a significant gain in computation time.

Remark that it is not necessary to sort the the Folded Time Series (the phases) according to increasing phase to be able to calculate the Jurkevich θ_J statistics.

This method has properties as described in ANOVA methods and it has been shown that the false alarm probability follows a β-distribution (see JCU-008).

3.3.5.1 Requirements

CU7-WP711-02000-S-FUN-140	1.0	AUT	Issued
It shall be possible to compute the Jurkevich-Stellingwerf periodogram for a time series for a list of frequencies.			
Parent: CU7-WP711-00000-S-FUN-120, CU7-WP711-02000-S-FUN-040			

3.3.5.2 Input/Output

Name	Description	Access Type
LightCurve	Lightcurve	Input
FrequencySet	A set of frequencies to be used for period analysis	Input
FrequencyGramme	The Jurkevich-Stellingwerf periodogram	Output

3.3.6 Kuiper method

Once a folded, and sorted time series is available, we can compute the Kuiper values (KV). Let us consider a folded time series with ϕ_i, and Flux$_i$, $i = 1, \ldots, n$ (usually called a folded light curve). We derive KV as follows.

1. We redefine the fluxes in cumulative flux (CF), that is for each measurement, we define

$$\text{CF}(i) = \text{CF}(i-1) + \frac{\text{Flux}_i}{\sum_{i=1}^{n} \text{Flux}_i}$$

   where $i = 1, \ldots, n$ and with CF(0)=0. The cumulative magnitude also satisfies CF(n) = 1 and CF(i) $\geq$ 0.

2. We compare the cumulative distribution of the observed magnitude with the distribution $(i/n, i = 1, \ldots, n)$ of a set of constant magnitude (equal to $\sum_{i=1}^{n}$ Magnitude$_i/n$). The difference Δ between the two distributions is

$$\Delta_i = \text{CF}(i) - i/n, i = 1, \ldots, n$$

   The value of the Kuiper test KV is then

$$KV = \Delta_i maximun - \Delta_i minimum$$

   Note that the maximum Δ_i is larger or equal to zero, and the minimum Δ_i is smaller or equal to zero.

A sketch of the distributions of the CF and of constant values are presented on Fig. 2, where the Δ_i maximum and minimum are also illustrated.

The Kuiper test was originally devised to determine if points are uniformly distributed on a circle. Tests still have to indicate how much gain there is in using this, rather slow, method.

3.3.6.1 Requirements

CU7-WP711-02000-S-FUN-160	1.0	AUT	Issued
It shall be possible to compute the Kuiper Value (KV) from a folded time series.			
Parent: CU7-WP711-00000-S-FUN-120, CU7-WP711-02000-S-FUN-040			

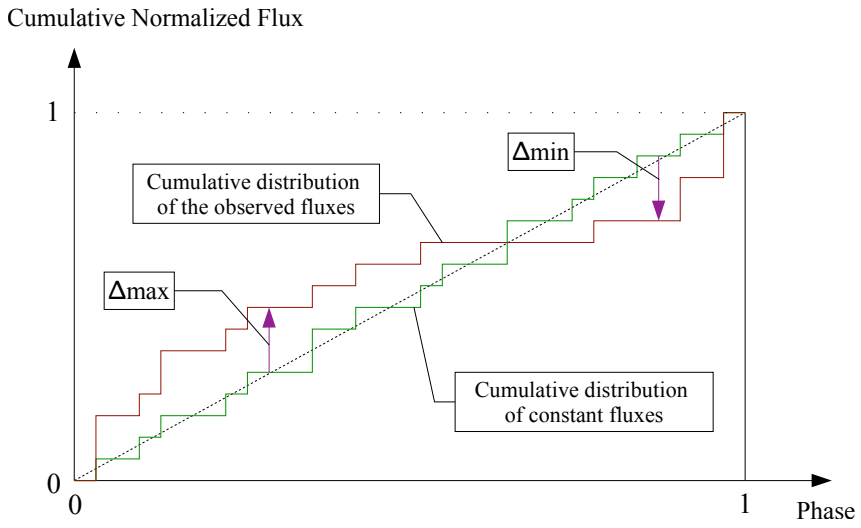

FIGURE 2: Cumulative distribution illustration

3.3.6.2 Input/Output

Name	Description	Access Type
FoldedLightCurve		Input
KuiperValue	result of Kuiper test	Output

3.4 False alarm probability (FAP)

The False Alarm Probability (FAP) of a period (frequency) is the probability that a value larger then the corresponding computed value in the periodogram (frequency-gram) of the time series occurs as the result of (random) noise only. The calculation should take into account that one has chosen the highest peak amongst a large number of peaks.

It should be possible to calculate for each method, either analytically, either with Monte Carlo simulations an estimate of the FAP for each result of any period search method. However, for some methods this is not straightforward. Theory and/or tests indicate that the FAP can derived from β (or β-like) distributions, although the parameters to be used are not always easy to calculate. Implementation of this FAP depends on the results of some tests. A technote on this subject has been prepared (JCU-008).

3.4.1 Requirements

CU7-WP711-02000-S-FUN-180	1.0	AUT	Issued
It shall be possible to compute a false alarm probability for every result for every method used.			
Parent: CU7-WP711-00000-S-FUN-60			

3.5 Producing the final period search result

3.5.1 Description and Objectives

Preliminary tests indicated that more than one period search method could be necessary to identify a maximal number of significant periods. If this is also computationally feasible, the software has to offer this possibility. Several strategies could be considered: one is to harmonize, compare and merge the results of the different methods to produce the final result. Another is to use methods successively. A workable version of the last strategy could be, to use one, preferably fast and efficient, method and only use other methods, if triggered by some statistical parameters or other indications. All options are under consideration and could change over the cycles.

3.5.2 Error on the frequency

For the error on the frequency, we will use, as a first estimate, the classical expression for equidistant observations with a sinusoidal variation (see e.g. Cuypers (1987) for details)

$$\frac{\sqrt{6}\sigma_n}{\pi A\sqrt{N}T},$$

with σ_n the expected standard deviation of the noise in the data, A the amplitude of the corresponding harmonic component of the model, N the number of data points and T the total time span of the observations. All this values are readily available and σ_n can be estimated from given error estimates or from the residuals after the modelling. To be complete, one should also take into account the error introduced because of the limited and discrete sampling. Kovacs (1981) showed that for a sinusoidal signal with frequency f without noise sampled on N equidistant times, the extremal value of the periodogram is displaced by about

$$\frac{0.11}{fT^2}$$

on average. Similar expressions were used by Baliunas et al. (1985) and Gilliland and Fisher (1985), but remain approximations for non-sinusoidal cases and time series with unequal sampling. This error is in general very small and only significant if the frequency is small (long periods) and the time base is short. If the signal-to-noise ratio is not high, this error will be negligible compared to the other errors.

The errors can be combined, assuming that they are independent:

$$\Delta f = \left[\frac{6\sigma_n^2}{\pi^2 A^2 NT^2} + \frac{0.0121}{f^2T^4}\right]^{1/2}.$$

3.5.3 Successive period search

If a method, as e.g. the Deeming method (3.3.1), associates a false alarm probability with a given result, it is easy to select those time series were this false alarm probability is extremely low (smaller than a value TBD) and pass those with the period search result to the modelling module for further processing. This will eliminate a lot of trivial cases where it is not necessary to use different methods. If the false alarm probability of the result of the first method is not very low (value TBD), other methods could be applied to produce a final result. It could also be a decision not to search with other methods if the FAP is very high (value TBD). In this way it is possible that a large fraction of the time series needs only one (fast) period search and this could result in a shorter computation time. Tests are necessary to quantify this.

3.5.4 Requirements

CU7-WP711-02000-S-FUN-200	1.0	AUT	Issued
It shall be possible to extract the most probable frequency or frequencies from the sets of frequencies.			
Parent: CU7-WP711-00000-S-FUN-060			

CU7-WP711-02000-S-FUN-220	1.0	AUT	Issued
It shall be possible to give and/or flag the significance of the most probable frequency(frequencies) from the sets of frequencies.			
Parent: CU7-WP711-00000-S-FUN-060			

CU7-WP711-02000-S-FUN-240	1.0	AUT	Issued
An estimate of the frequency error shall be provided for the most probable frequency or frequencies from the sets of frequencies .			
Parent: CU7-WP711-00000-S-FUN-060			

3.5.5 Input / Output

Name	Description	Access Type
FrequencySet	Set of frequencies from the different search routines	Input
BestFrequency	The best frequency estimate, its error a probability and/or a probability flag	Output

4 Traceability

Parent Requirement	Requirements in this document
CU7-WP711-00000-S-FUN-060	CU7-WP711-02000-S-FUN-200, CU7-WP711-02000-S-FUN-220, CU7-WP711-02000-S-FUN-240
CU7-WP711-00000-S-FUN-120	CU7-WP711-02000-S-FUN-20, CU7-WP711-02000-S-FUN-60, CU7-WP711-02000-S-FUN-80, CU7-WP711-02000-S-FUN-100, CU7-WP711-02000-S-FUN-120, CU7-WP711-02000-S-FUN-140, CU7-WP711-02000-S-FUN-160
CU7-WP711-00000-S-FUN-60	CU7-WP711-02000-S-FUN-180
CU7-WP711-02000-S-FUN-040	CU7-WP711-02000-S-FUN-60, CU7-WP711-02000-S-FUN-80, CU7-WP711-02000-S-FUN-100, CU7-WP711-02000-S-FUN-120, CU7-WP711-02000-S-FUN-140, CU7-WP711-02000-S-FUN-160
CU7-WP711-02000-S-FUN-020	CU7-WP711-02000-S-FUN-40