Astronomy & Astrophysics manuscript no. Ninsight_paperII_v01
November 24, 2016
©ESO 2016

arXiv:1611.07838v1 [astro-ph.IM] 23 Nov 2016

New Insights into Time Series Analysis

II - No Correlated Observations

C. E. Ferreira Lopes[1,2] and N. J. G. Cross[1]

[1] SUPA (Scottish Universities Physics Alliance) Wide-Field Astronomy Unit, Institute for Astronomy, School of Physics and Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
e-mail: ferreiralopes1011@gmail.com
[2] Departamento de Física, Universidade Federal do Rio Grande do Norte, Natal, RN, 59072-970 Brazil

Received November, 2016; accepted xxx, 2016

ABSTRACT

Context. Statistical parameters are used in finance, weather, industrial, science, among other vast number of different fields to draw conclusions. They are also used to identify variability patterns on photometric data in order to select non-stochastic variations, indicative of astrophysical effects. New more efficient selection methods are mandatory to analyses the huge amount of astronomical data.
Aims. Our aims are to improve the current methods used to select non-stochastic variations on non-correlated data.
Methods. The standard and new data-mining parameters to analyses non-correlated data are used to set the best way to discriminate stochastic and non-stochastic variations. A new approach including a modified Strateva function are used to select non-stochastic variations. Monte-Carlo simulation and public time-domain data are used to estimate its accuracy and performance.
Results. We introduce 15 modified statistical parameters covering different features of statistical distribution, like; average, dispersion, and shape parameters. Many of dispersion and shape parameters are unbound parameters, i.e. equations which do not require the calculation of the average. Unbound parameters are computed using single loop and so decreasing running time. Moreover, the majority of them have lower error than previous ones that is mainly observed for distributions having few measurements. From estimation of uncertainties together with a noise model we introduce a non-correlated variability index. It reduces the number of misselections by about 770%. We also improve the correlated indices by reducing $E_{\rm tot}$ by 18% using the even-mean.
Conclusions. Then non-correlated indices improve the selection criteria of non-stochastic variations. The even-averages provide a better estimation of mean and median and so this work also improves the correlated variability indices proposed in the first paper of this series. We consider that the first step of this project, where we set new techniques and methods that provide a huge improve on the efficiency of selection of variable stars, is now complete. Many of this techniques may be useful for vast number of different fields. Next, we will commence a new step of this project on the analysis of period search methods.

Key words. catalog – variable stars – infrared

1. Introduction

Statistical analysis is a vital concept in our lives because it is used to understand what's going on and thereby make a decision. They are also used to assess theoretical models by experiments that are limited by experimental factors, leading to uncertainty. Measurements are usually performed many times to increase the confidence level. The results are summarized by statistical parameters in order to communicate the largest amount of information as simply as possible. Statisticians commonly describe the observations by averages (such as the arithmetic mean, median, mode, and interquartile mean), dispersion (e.g. standard deviation, variance, range, interquartile range, absolute deviation), shape of the distribution (such as a skewness and kurtosis), and a measure of statistical dependence (like Spearman's rank correlation coefficient). These parameters are used in finance, weather, industry, experiments, science and in several other areas to characterize probability distributions. New insights on this topic should be valuable in many of the natural sciences, technology, economy, and quantitative social science research.

Improvements on data analysis methods are mandatory to analyze the huge amount of data collected in recent years. Large volumes of data having potential scientific results are left unexplored or delayed due to current inventory tools that are unable to produce clear samples. In fact, we risk underusing a large part of these data despite efforts having been undertaken. The current techniques of data processing can be improved considerably. For instance, the flux independent index, proposed by us in a previous paper, reduces the misselection of variable sources by about 250% (Ferreira Lopes & Cross 2016). A reliable selection on astronomical databases allows us put forward faster scientific results such as those enclosed in many current surveys (e.g. Kaiser et al. 2002; Udalski 2003; Pollacco et al. 2006; Baglin et al. 2007; Hoffman et al. 2009; Borucki et al. 2010; Bailer-Jones et al. 2013; Minniti et al. 2010). The reduction of misclassification at the selection step is crucial to follow up the development of the instruments themselves.

Statistical parameters like standard deviation and kurtosis as function of magnitude have being used as main way to select variable stars (e.g. Cross et al. 2009). This method assumes that for the same magnitude stochastic and non-stochastic variation have different statistical properties. To compute all current dispersion and almost all shape parameters the averages must also be calculated and thus we increase uncertainties as well as

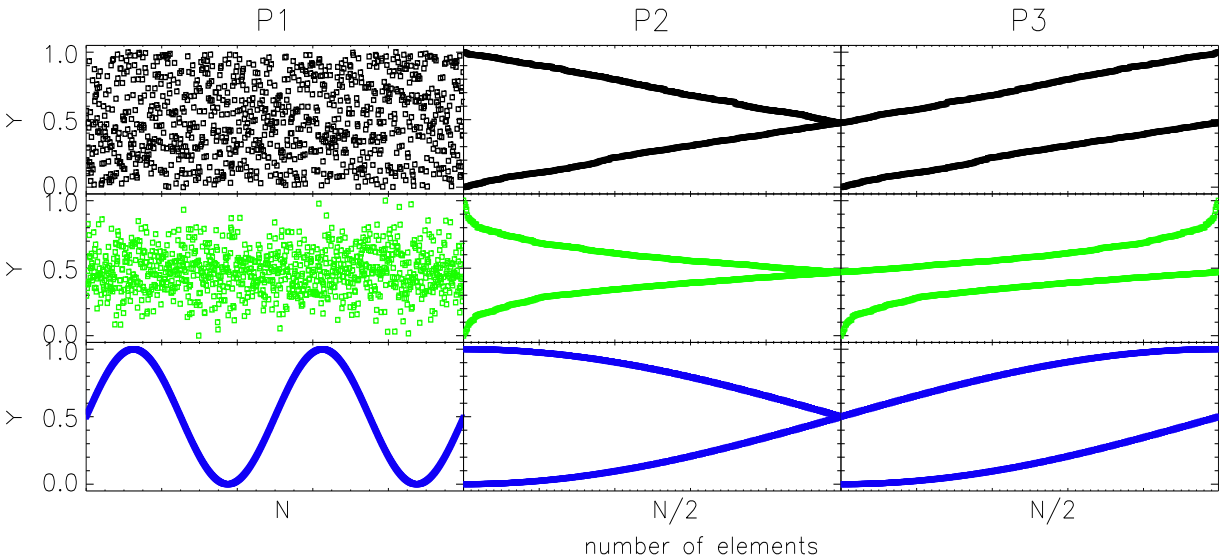


Fig. 1: Uniform (top panels), normal (middle panels), and sinusoidal (bottom panels) distributions having 1000 measurements as a function of the number of elements. The same distributions are displayed using different arrangements (ABCD panels - see Sect. 3 for a better description). The dashed grey lines mark the mean value and median values, respectively.

the processing time. Indeed, statistical properties still exist even where averages are unknown. In this fashion, Brys et al. (2004) proposed a robust measure of skewness, called the 'medcouple', from the comparison of quartiles and pairs of measurements that allow us to compute it without use averages. However, it has a long running time since the number of possible combination increases by factorial of the number of measurements. However, we can use a similar idea to propose new averages, dispersions, and shape parameters which have a smaller running time.

This work is the second in a series about new insights into time series analysis. In the first paper we assess the discrimination of variable stars from noise for correlated data using variability indices (Ferreira Lopes & Cross 2016). Now, new statistical parameters and their accuracy, in comparison with previous ones, are analyzed to increase the capability to discriminate stochastic and non-stochastic distributions. We also look into their dependence with the number of epochs to determine statistical weights to improve the selection criteria. Lastly we use a noise model to propose a new non-correlated variability index. Forthcoming papers are going to combine the selection criteria and study how to use the full current inventory of period finding methods to clean the sample selected by variability indices.

The notation used is described in Sect. 2, and next we suggest a new sets of statistic in Sect. 3. In Sections 4) and 5 the new parameters are tested and a new approach to model the noise and and select variable stars is proposed. Next, the selection criteria are tested on real data in Sect. 6. Finally, we summarise and make our conclusions in Sect. 8.

2. Notation

Let $Y' := y'_1 \leq y'_2 \leq \ldots \leq y'_c \leq \ldots \leq y'_{N'}$ from where the kernel function is defined by

$$Y := \begin{cases} y'_i \in Y' \quad \forall \quad y'_i = y'_i & if\ N' \ even \\ y'_i \in Y' \quad \forall \quad y'_i \neq y'_{Int(N/2)+1} & if\ N' \ odd \end{cases} \tag{1}$$

where $Int(N/2)$ means the integer part (floor) of half the number of measurements. The lower contribution to compute statistical parameters for symmetric distributions is given by $y'_{Int(N/2)+1}$ since it is the nearest measurement to the average. Y is a sample having an even number of measurements (N) that also can be discriminated into the sub-samples Y^- and Y^+ composed of measurements having $y_i \leq y_{N/2}$ and $y_i > y_{N/2}$, respectively. Different arrangements can be taken into consideration, such as;

$P1$ - unsystematic y_i values;
$P2$ - increasing order of Y^- and Y^+;
$P3$ - decreasing order of Y^- and increasing order of Y^+;

where the measurement of Y^- and Y^+ for $P2$ and $P3$ assume the same position on the x axis only to provide a better display. [Perhaps we should show a less symmetric distribution for completeness] Figure 1 shows the $P1-3$ arrangements for uniform (black), normal (green), and sinusoidal (blue) distributions. The symmetry found in the $P2$ and $P3$ yield unique information about the shape and dispersion. Therefore, we combine the measurements of Y^- and Y^+ to propose a new set of statistical parameters (see Table 1).

3. Even Statistic (E)

Non-parametric statistics are not based on probability distributions whose interpretation does not depend on the fitting of parametrized distributions. The typical parameters used for descriptive and inferential purposes are the mean, median, standard deviation, skewness, kurtosis, among others to name a few. They are defined to be a function of a sample that has no dependency on a parameter, i.e. its values are the same for any type of arrangement, as for instance $P1-3$ (see Fig. 1). All dispersion and almost all shape parameters are dependent of some kind of average, i.e. they describe distributions around average values. Indeed, dispersion and shape still exist even where the averages are unknown. For instance, the standard deviation and absolute deviation provide an estimate of dispersion about the mean value. On the other hand, statistical parameters only based on the measurements provide unbound estimates.

Consider a symmetric distribution where the mean value can be computed as $\sum_{i=0}^{Int(N/2)} y'_i/N + \sum_{i=Int(N/2)+1}^{N} y'_i/N$. Where

the first and second term are the weights of left and right side of the distributions having the same number of elements. On the other hand, for odd numbers of measurements, the weight for both sides of the distribution are not equivalent. The single measurement that can be withdrawn to correct such variation is $y'_{Int(N/2)+1}$ since if we withdraw any of the other y'_i measurements we would increase the difference. The kernel given by Eq. 1 provides samples having even numbers of measurements and so the same weight for both sides. The parameters proposed using Eq. 1 are named as 'even parameters'. Moreover, the even number of measurements allows us to compare single pairs of measurements among Y^- and Y^+ and so to estimate dispersion and shape values without taking account of the average. The new statistical parameters to compute averages (see Sect. 3.1), dispersions (see Sect. 3.2), and shapes (see Sect. 3.3) are described below.

3.1. Averages (A)

Considering the kernel given by Eq. 1(see Sect. 2) we propose new average statistics, given by

$$EA_\mu = \frac{1}{N}\sum_{i=0}^{N} y_i, \tag{2}$$

and

$$EA_m = \frac{y_{\frac{N}{2}} + y_{\frac{N}{2}+1}}{2} \tag{3}$$

where EA_μ and EA_m are named as even-mean and even-median. These expressions mimic the mean (A_μ) and median (A_M). Moreover, $EA_\mu = A_\mu$ and $EA_m = A_m$ when the number of measurements (N') is an even number. A comparison between them is performed in Sect. 4.1.

3.2. Dispersion Parameters (D)

Statistical dispersion is used to measure the amount of sample variance and it is computed using the absolute or square value of the distance between the measurements and the average. Improving the estimation of averages can provide a better accuracy of dispersion parameters, such as: mean standard deviation (D_{σ_μ}), median standard deviation (D_{σ_m}), mean absolute standard deviation (D_μ), and median absolute standard deviation (D_M). Therefore we propose the even-dispersion parameters that are computed using the even-averages (see Table 1 - the even-standard-deviation (ED_{σ_μ}), the even-median-standard-deviation (ED_{σ_m}), the even-mean-absolute-standard-deviation (ED_μ), and the even-median-absolute-standard-deviation (ED_m)). The accuracy of these parameters is assessed in Sect. 4.1. A note of caution, these parameters will return the same values as the previous ones for even numbers of measurements.

Moreover, using the single combination between measurements Y^- with Y^+ we also can estimate the amount of variation or dispersion of a sample. From the kernel given by Eq. 1 we propose the following parameters, written as;

$$ED = \frac{1}{N}\sum_{i=1}^{N/2}(y_{N-i} - y_i) = \frac{1}{N}\sum_{i=1}^{N/2}\left(y_{\frac{N}{2}+i} - y_i\right). \tag{4}$$

The two parts are the same since $\sum_{i=1}^{N/2} y_{N-i} = \sum_{i=1}^{N/2} y_{\frac{N}{2}+i}$. ED means even-absolute deviation because such a sum is always positive, i.e. $(y_{N-i} - y_i) \geq 0$ and $\left(y_{\frac{N}{2}+i} - y_i\right) \geq 0$. Moreover, a simple identity is found for distributions having an even number of measurements,

$$\begin{aligned} ED &= \frac{1}{N}\sum_{i=1}^{N/2}(y_{N-i} - y_i + EA_m - EA_m) \\ &= \frac{1}{N}\sum_{i=1}^{N/2}(y_{N-i} - EA_m) - (y_i - EA_m) \\ &= ED_m \end{aligned} \tag{5}$$

once $(y_{N-i} - EA_m) \geq 0$ and $(y_i - EA_m) \leq 0$. Indeed, we also can mimic the standard deviation by proposing two new even-dispersion parameters, given by,

$$ED_{(1)} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N/2}(y_{N-i} - y_i)^2} \tag{6}$$

and,

$$ED_{(2)} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N/2}\left(y_{\frac{N}{2}+i} - y_i\right)^2}. \tag{7}$$

The even-dispersion parameters are unbound, i.e. they are not dependent on the average. They allow us speak about the dispersion of a distribution instead of the dispersion about an average. Moreover, a strict relationship between $ED_{(1)}$ and $ED_{(2)}$ with ED_{σ_m} is found when we have even numbers of measurements:

$$\begin{aligned} ED^2_{(1)} &= \frac{1}{N-1}\sum_{i=1}^{N/2}(y_{N-i} - y_i + EA_m - EA_m)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N/2}[(y_{N-i} - EA_m) - (y_i - EA_m)]^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N/2}\Big[(y_{N-i} - EA_m)^2 + (y_i - EA_m)^2 \\ &\quad -2\,(y_{N-i} - EA_m)\times(y_i - EA_m)\Big] \\ &= ED^2_{\sigma_m} - 2\times Cov(y_{N-i}, y_i) \end{aligned} \tag{8}$$

while for $ED_{(2)}$,

$$\begin{aligned} ED^2_{(2)} &= \frac{1}{N-1}\sum_{i=1}^{N/2}\left(y_{\frac{N}{2}+i} - y_i + EA_m - EA_m\right)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N/2}\left[\left(y_{\frac{N}{2}+i} - EA_m\right) - (y_i - EA_m)\right]^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N/2}\Big[\left(y_{\frac{N}{2}+i} - EA_m\right)^2 + (y_i - EA_m)^2 \\ &\quad -2\left(y_{\frac{N}{2}+i} - EA_m\right)\times(y_i - EA_m)\Big] \\ &= ED^2_{\sigma_m} - 2\times Cov(y_{N-i}, y_i) \end{aligned} \tag{9}$$

Table 1: Variability statistical analyses in the present work.

N	Statistic	Definition	Reference
1	Even-mean	$EA_\mu = \frac{1}{N}\sum_{i=0}^{N} y_i$	Average
2	Even-median	$EA_m = \frac{y_{\frac{N}{2}} + y_{\frac{N}{2}+1}}{2}$	Average
3	Even-mean-standard-deviation	$ED_{\sigma_\mu} = \sqrt{\frac{1}{(N'-1)}\sum_{i=1}^{N'}\left(y'_i - EA_\mu\right)^2}$	Dispersion
4	Even-median-standard-deviation	$ED_{\sigma_m} = \sqrt{\frac{1}{(N'-1)}\sum_{i=1}^{N'}\left(y'_i - EA_M\right)^2}$	Dispersion
5	Even-mean-absolute-deviation	$ED_\mu = \frac{1}{N'}\sum_{i=1}^{N'}\left\lvert y'_i - EA_\mu\right\rvert$	Dispersion
6	Even-median-absolute-deviation	$ED_m = \frac{1}{N'}\sum_{i=1}^{N'}\left\lvert y'_i - EA_M\right\rvert$	Dispersion
7	Even-absolute-deviation	$ED = \frac{2}{N}\sum_{i=1}^{N/2}(y_{N-i} - y_i)$	Dispersion
8	Even-deviation (1)	$ED_{(1)} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N/2}(y_{N-i} - y_i)^2}$	Dispersion
9	Even-deviation (2)	$ED_{(2)} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N/2}\left(y_{\frac{N}{2}+i} - y_i\right)^2}$	Dispersion
10	Even-skewness	$ES = \frac{\frac{1}{N'}\sum_{i=1}^{N'}\left(y'_i - EA_\mu\right)^3}{ED_\mu^3}$	Shape
11	Even-kurtosis	$EK = \frac{\frac{1}{N'}\sum_{i=1}^{N'}\left(y'_i - EA_\mu\right)^4}{ED_\mu^4} - 3$	Shape
12	Even-shape (1)	$ES_{(1)} = \frac{\frac{1}{N}\sum_{i=1}^{N/2}(y_{N-i} - y_i)^3}{ED_{(1)}^3} - 2.26$	Shape
13	Even-shape (2)	$ES_{(2)} = \frac{\frac{1}{N}\sum_{i=1}^{N/2}\left(y_{\frac{N}{2}+i} - y_i\right)^3}{ED_{(2)}^3} - 1.52$	Shape
14	Even-shape (3)	$EK_{(1)} = \frac{\frac{1}{N}\sum_{i=1}^{N/2}(y_{N-i} - y_i)^4}{ED_{(1)}^4} - 6$	Shape
15	Even-shape (4)	$EK_{(2)} = \frac{\frac{1}{N}\sum_{i=1}^{N/2}\left(y_{\frac{N}{2}+i} - y_i\right)^4}{ED_{(2)}^4} - 2.46$	Shape

where Cov denotes covariance. Indeed, the second term in these equations will be additive since the covariance among Y^- and Y^+ is negative.

The identity give by Eqs. 5, 8, and 9 also is valid for odd number of measurements if we have a large number of measurements since $ED_m \simeq D_m$ and $ED_{\sigma_m} \simeq D_{\sigma_m}$. Similar relationships can be found using A_μ, A_m, EA_μ, and EA_m.

The dispersion of a distribution given by Eqs. 6 and 7 is the standard deviation about the averages minus two times the covariance among Y^- and Y^+. Moreover, for symmetric distributions where $y_{N-i} - EA_M = -(y_i - EA_m)$ as well as $EA_m = EA_\mu$, we can write the following identity,

$$\begin{aligned} ED_{(1)} &= \sqrt{ED_{\sigma_m}^2 - \frac{2}{N-1}\sum_{i=1}^{N/2}(y_{N-i} - EA_m)\times(y_i - EA_m)} \\ &= \sqrt{ED_{\sigma_m}^2 + \frac{2}{N-1}\sum_{i=1}^{N/2}(y_i - EA_m)^2} \\ &= \sqrt{2}\times ED_{\sigma_m} \end{aligned} \tag{10}$$

The ratio of $ED_{(1)}$ by ED_{σ_m} can be used to estimate if the measurements are symmetrically distributed.

3.3. Shape Parameters (S)

In a similar fashion to the dispersion parameters, we also can improve the accuracy of skewness (S_S) and kurtosis (S_K) using the even-averages. Therefore, we propose the even-skewness (ES) and even-kurtosis (EK) to estimate the distribution shape (see Table 1 10-11). Moreover, we also propose the higher moments of $ED_{(1)}$ and $ED_{(2)}$ as new even-shape-parameters, given by:

$$ES_{(1)} = \frac{\frac{1}{N}\sum_{i=1}^{N/2}(y_{N-i} - y_i)^3}{ED_{(1)}^3}. \tag{11}$$

and,

$$ES_{(2)} = \frac{\frac{1}{N}\sum_{i=1}^{N/2}\left(y_{\frac{N}{2}+i} - y_i\right)^3}{ED_{(2)}^3} \tag{12}$$

and,

$$EK_{(1)} = \frac{\frac{1}{N}\sum_{i=1}^{N/2}(y_{N-i} - y_i)^4}{ED^4_{(1)}}. \quad (13)$$

and,

$$EK_{(2)} = \frac{\frac{1}{N}\sum_{i=1}^{N/2}\left(y_{\frac{N}{2}+i} - y_i\right)^4}{ED^4_{(2)}} \quad (14)$$

$ES_{(1-2)}$ and $EK_{(1-2)}$ are unbound parameters, i.e. they are independent of the average. Where $ES_{(1-2)}$ mimics the skewness while $EK_{(1-2)}$ mimics kurtosis. A strict relationship between $ES_{(1-2)}$ and ES as well as $EK_{(1-2)}$ and EK is quite complicated once such definitions use distinct dispersion parameters. Indeed, we can use other dispersion parameters to broaden the list of even-shape parameters.

4. Simulating distributionsons

We performed 10^7 Monte Carlo simulations using the uniform, normal and sinusoidal distributions with a range of the number of measurements (number of epochs) varying from 10 to 100. Such intervals typify light curves from surveys such as Pan-STARRS, VVV, and Gaia. It means about 10^5 simulations for a given number of measurements. These simulations were performed for uniform, normal and sinusoidal distributions (see Figs. 1) where the first two mimic noise while the last variable stars. The statistical parameters have a higher statistical significance for distributions having a large number of measurements where the addition of measurements only implies small fluctuations. Therefore, we adopt as a "true parameter" value (P_{true}) those computed using 10^5 measurements. This value is used as a reference to analyse the error given by,

$$e_P = \frac{|P - P_{true}|}{P_{true}} \quad (15)$$

where P means the statistical parameter. This expression provides the mean error for P. In order to avoid singularities we shift the skewness and kurtosis values for $P'_{true} = P_{true} + 1$ as well as $P = P' + 1$ since they have $P_{true} \sim 0$.

4.1. Accuracy of the statistical parameters

Figure 2 shows e_P (see Eq.15) for the even-parameters (see Tab. 1 1-6 and 10-11) and its comparison with previous parameters (mean, median, mean standard deviation, median standard deviation, mean absolute deviation, median absolute deviation, skewness, and kurtosis) as a function of the number of measurements in the left panels and right panels, respectively. The left panels have the results of simulations for the whole range of measurements while the right panels only have the results for odd numbers of measurements because for even numbers of measurements the current and even-statistical parameters have the same values. Therefore, we only use the results for odd numbers of measurements, i.e. $11, 13, 15, \cdots$. Where $e_P/e_{P'} < 1$ means a higher accuracy for the new parameters compared to current parameters while $e_P/e_{P'} > 1$ means no improvement with the new parameters. The simulations were performed as described in the Sect. 4, from which we can observe that:

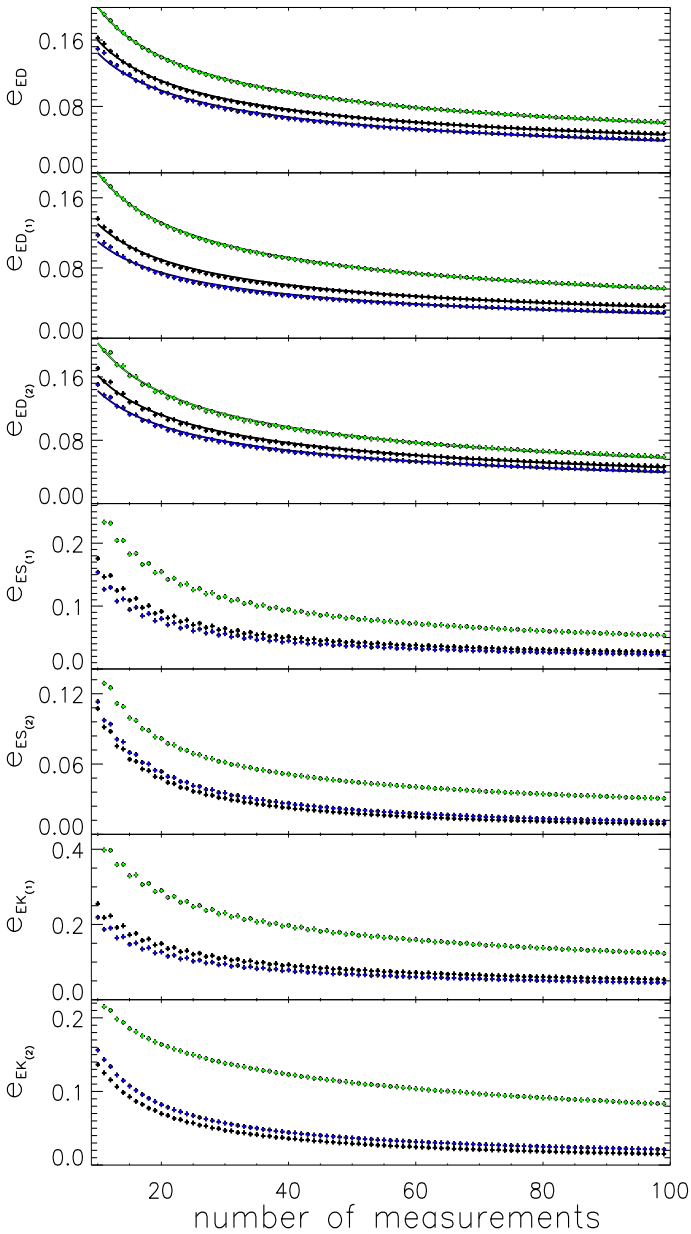


Fig. 3: e_P as function of number of measurements for the free even-parameters (see Tabs. 1). The colors are the same as Fig. 2.

- EA_μ shows lower errors for the normal distribution than found for uniform and sinusoidal distributions. Indeed, $e_{EA_\mu}/e_{A_\mu} \simeq 1$ for normal distributions. The error for 10 measurements is about twice that found for 100 measurements. The even-mean parameters are more accurate than the mean. For instance, the even-mean presents improved accuracy of about $\sim$ 4% and $\sim$ 8% for uniform and sinusoidal distributions over that found by the mean. On the other hand, $e_{EA_\mu} \simeq e_{A_\mu}$ for normal distributions.
- EA_m is more accurate than A_m for the whole range of measurements and distributions analyzed. They are about $\sim$ 12%, $\sim$ 9%, and $\sim$ 6% more accurate for sinusoidal, uniform, and normal distributions respectively.
- D_{σ_μ} is more accurate than ED_{σ_μ} for the whole range of measurements and distributions analyzed despite the improvement on the estimation of the mean. Such a difference is less than $\sim$ 0.1%. On the other hand, ED_{σ_m} is more accurate than

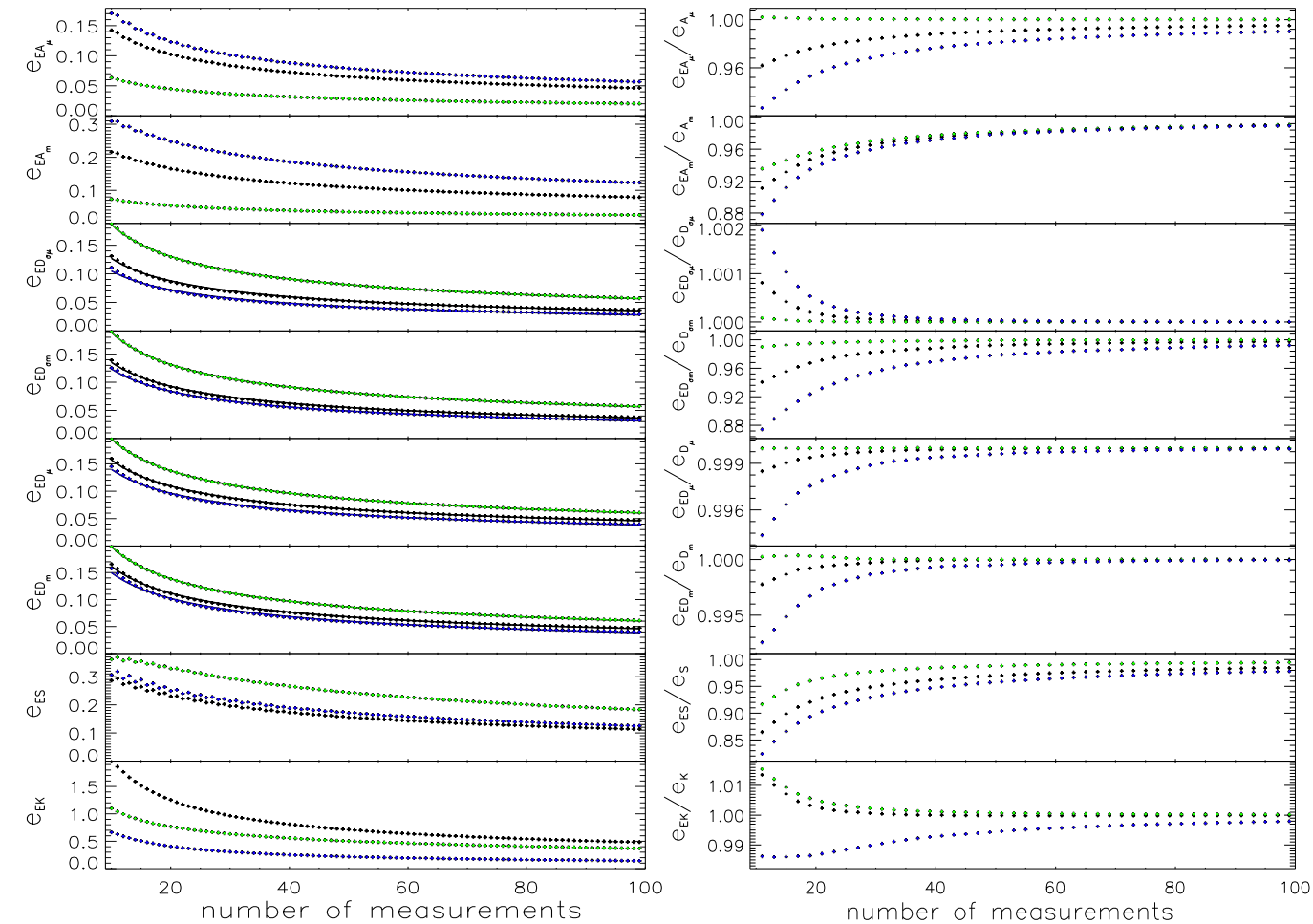


Fig. 2: e_P (left panels) and its comparison with previous statistical parameters (right panels) as a function of the number of measurements (see Tab. 1) for uniform (black points), normal (blue points), and sinusoidal (green points) distributions. The solid lines in the even dispersion diagrams set the models described in Sect. 4.2.

D_{σ_m} for the whole range of measurements and distributions analyzed. They are about ∼ 5%, ∼ 1%, and ∼ 13% more accurate for uniform, normal, and sinusoidal distributions respectively having 10 measurements.

– ED_μ and ED_m present a similar relative error and they are more accurate than ED_{σ_μ} or ED_{σ_m}. Moreover, they are more accurate than D_μ and D_m.
– ES and EK have an uncertainty higher than ∼ 10% up to 50 measurements. We find an increase of accuracy using ES than S. On the other hand, EK is about ∼ 1% less accurate than K for normal and uniform distributions compared with sinusoidal distributions.

To summarise, the accuracy of statistical parameters has a strong dependence of number of measurements and distribution type. The improvements in estimation of averages by even statistics allows us to improve the estimation of dispersion and shape parameters for all distributions analyzed. It is mainly observed for distributions where the probability to find measurements near to P_{true} is smaller. As result, $e_{P_{even}}/e_P \simeq 1$ for the normal distribution.

On the other hand, the unbound even statistical parameters display similar properties to those found for previous parameters (see Fig. 3). Indeed, the even-shape parameters present smaller errors than found by skewness and kurtosis. Even-parameters (see Table 1) have a similar or better accuracy to previous statistical parameters and so they can be used to characterize statistical distributions in a similar fashion to previous ones.

Table 2: Coefficients of Eq.16.

	Uniform		Normal		Sinusoidal	
P	$b_1^{(P)}$	$b_2^{(P)}$	$b_1^{(P)}$	$b_2^{(P)}$	$b_1^{(P)}$	$b_2^{(P)}$
ED_{σ_μ}	-7.3×10^{-3}	0.18	-4.0×10^{-3}	0.36	-7.9×10^{-3}	0.13
$ED_{\sigma m}$	-9.7×10^{-3}	0.21	-4.9×10^{-3}	0.37	-11.6×10^{-3}	0.18
ED_μ	-5.6×10^{-3}	0.26	-2.4×10^{-3}	0.39	-8.7×10^{-3}	0.22
ED_m	-7.7×10^{-3}	0.29	-3.5×10^{-3}	0.40	-12.9×10^{-3}	0.26
ED	-6.9×10^{-3}	0.28	-4.7×10^{-3}	0.42	-10.5×10^{-3}	0.24
$ED_{(1)}$	-8.8×10^{-3}	0.19	-5.2×10^{-3}	0.37	-10.1×10^{-3}	0.14
$ED_{(2)}$	-8.3×10^{-3}	0.29	-9.6×10^{-3}	0.45	-8.1×10^{-3}	0.23

4.2. The coefficients adjusted for sample size

We create adjusted coefficients for sample size because samples having few measurements have a large fluctuations in the estimated parameters. For instance, the Fisher-Pearson coefficient (given by $\sqrt{n \times (n-1)}/(n-1)$) for a sample having 10 and 100 measurements is 1.054 and 1.005, respectively. As result, for instance, this correction increases the value if the skewness is positive, and makes the value more negative if the skewness is negative. It cannot be used for parameters that only assume positive

values like standard deviation. However other adjusted coefficients have been proposed in a similar fashion, and indeed these coefficients increase the dispersion in a population since they enlarge the range of values. Moreover, we cannot assume the same adjustment coefficient for all statistical parameters since they have different accuracies (see Sect. 4).

We use the simulations described in the Sect. 4 to determine a model for different statistical parameters, given by

$$w_{(P)} = 1 - \left(b_1^{(P)} + \sqrt{\frac{b_2^{(P)}}{N}} \right) \tag{16}$$

where $b_1^{(P)}$ and $b_2^{(P)}$ are a real number constant (see Table 2).

5. Modelling the noise

Cross et al. (2009) used the Strateva function (see Strateva et al. 2001; Sesar et al. 2007, for more details) to fit the standard deviation as a function of magnitude to estimate a noise model (ζ). This method assumes that the majority of the sample are point sources, where the variability measurements are dominated by noise, rather than astrophysical variations. It provides a suitable model for photometric surveys at optical wavelengths if they have a single component of noise that increases in relative magnitude from bright to faint stars. However, the brightest stars can show much greater variation which comes from saturation and non-linearity of the detectors providing a source of variation that cannot be fit by these models. Such a situation is rare at optical wavebands but is quite frequently present for NIR data (see Fig. 5). Since the sky foreground emitted by the atmosphere is highly variable in the NIR, it causes a highly time-varying saturation limit, which can affect large parts of otherwise highly accurate time-series data for bright stars with substantial outliers having very small formal error estimates (Ferreira Lopes et al. 2015). These outliers will probably lead to a spurious impact upon the statistical parameters. Therefore, we propose a modification to the Strateva function that allows us to model such variations, the increase in the standard deviation for bright (saturated stars) and faint (photon noise) stars, given by;

$$\zeta_{(P)}(m) = c_0 + c_1 10^{0.4\times m} + c_2 10^{0.8\times m} + c_3 10^{-0.8\times m} \tag{17}$$

where all the coefficients are real numbers. Indeed, the three first terms were proposed previously (e.g Strateva et al. 2001; Sesar et al. 2007). Of course, the coefficients of the fit given by Eq. 17 will be different for distinct dispersion parameters and $c_3 \sim 0$ for optical wavebands, for instance.

The procedure adopted to model the standard deviation as a function of magnitude is similar to that used by (Cross et al. 2009). We compute EA_M and ED_M for bins having a width of 0.1 magnitude or having at least 100 objects. For this step, we only consider those stars having more than 20 measurements. Next, we compute $\zeta_{P'}(m)$ from a non-linear least-squares minimization using the Levenberg-Marquardt method Levenberg (1944); Marquardt (1963). This procedure is iterated once more on those stars having values below $2 \times ED_M$ to calculate the noise model. As a result, we obtain a better noise model since we withdraw probable variable stars.

5.1. Non-correlated and combined indices

We can use one or combine all even-dispersion parameters (see Table 1) to select non-stochastic variations. In order to combine them

$$I_{(P)} = \frac{w_{(P)} \times P}{\zeta_{w_{(P)}P}} \tag{18}$$

where $w_{(P)}$ and $\zeta_{w_{(P)}P}$ are given by Eqs. 16 and 17, respectively. From which, we propose to combine the results from different even-dispersion parameters into the following index:

$$X_f = \frac{\sum_{j=1}^{v} \omega_{P_j} I_{P_j}}{\sum_{j=1}^{v} \omega_{P_j}} \tag{19}$$

where f is the waveband used, ω_{P_j} is a weight related with each even-dispersion parameter, v is the number of dispersion parameters used, and I_{P_j} is given by Eq. 18. Indeed, I_{P_j} provides a normalized index allowing us to combine distinct dispersion parameters as well as the results from different wavebands. For this work, we consider $P = [ED_{\sigma\mu}, ED_{\sigma m}, ED_{\mu}, ED_m, ED, ED_{(1)}, ED_{(2)}]$ and ω_{P_j} as the inverse of $ED_{\sigma\mu}$ for each P_j.

The correlated indices are more efficient than non-correlated indices and so should be adopted when they are available (see Sect. 6) otherwise we can use X_f indices (see Fig. 4). This is likely to be the case in most situations where correlated indices can be calculated sensibly, but this still needs to be tested, especially in systems where the correlation order and number of permutations are very low. A note of caution, the flux independent indices are weakly dependent on magnitude but are strongly dependent of the time interval among correlated measurements, so should be used when the observations have a natural correlation interval that is shorter than the typical epoch interval (for more details see Sec 4.3 Paper I).

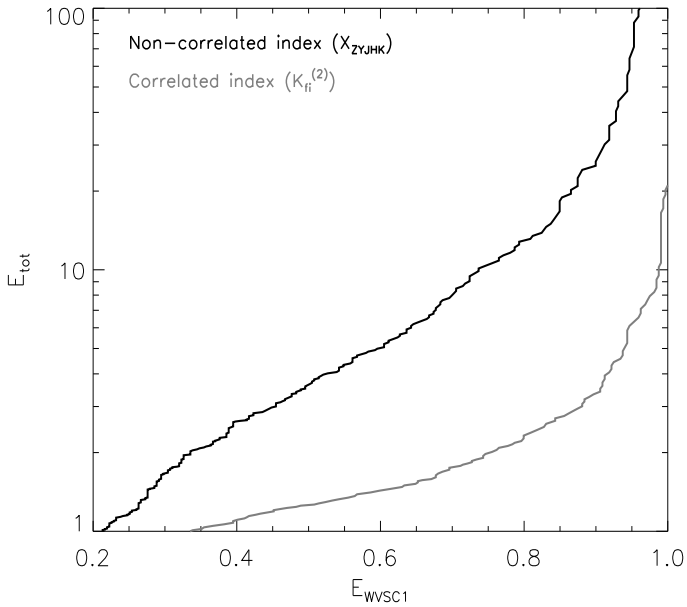


Fig. 4: E_{tot}as a function of E_{WVSC1} for correlated (grey line) and non-correlated indices (black line).

5.2. Broadband selection

Variable stars candidates for non-correlated data are usually selected from the noise model (see Sect. 5). Stars having values above $n \times D_{\sigma\mu}$ are selected for further analyses. This approach assumes that a few percent of entire sample are variable stars and have statistical values above the noise. The noise samples present distributions like uniform, normal, or distributions in between, while variable stars are more similar to sinusoidal distributions. Therefore, the dispersion parameters assume a different range of values for variable and non-variable stars that is mainly highlighted for samples having a high number of measurements (typically higher than 50). Indeed, such a difference must increase for higher amplitudes than that found for the noise. For few measurements (typically less than 20) stochastic and non-stochastic variations have large uncertainties increasing the mis-selection (see Sect. 4.1). We find a similar behaviour for correlated indices. For instance, Ferreira Lopes et al. (2015) uses cut-off surfaces linking magnitude, number of epochs, and variability indices to improve the selection criteria of variable stars while Ferreira Lopes & Cross (2016) use flux independent indices to propose an empirical relation between cut-off values and number of measurements without taking into account magnitude.

The adjusted coefficients for sample size, as presented in Sect. 4.2, reduce the population dispersion. Meanwhile, uncertainties about the range of values assumed by stochastic and non-stochastic variations also vary with the number of measurements. For non-stochastic variations having a good signal-to-noise and a large number of measurements such a range is different to that produced by stochastic variations. On the other hand, for distributions having just a few measurements, the range of values can significantly overlap. In the same fashion as the empirical selection criteria proposed by Ferreira Lopes & Cross 2016 (see Eq. 16), we propose the follow criteria,

$$f(\alpha,\beta) = \alpha + \sqrt{\frac{\beta}{N}} \tag{20}$$

where α and β are real positive values and N is the number of measurements. Where α is bigger than 1, we may find stochastic variations. Higher values of β provide a higher cutoff for small numbers of measurements or correlations. For instance, $f(1,4)$ for N equal to 10, 30, and 50 are 1.63, 1.36, and 1.28, respectively. Indeed lower values of α provide a more complete selection while higher values provide a more reliable selection.

6. Real data

We use the WFCAM Calibration 08B release (WFCAMCAL08B - Hodgkin et al. 2009; Cross et al. 2009) as a test database like we do in the first paper of this series. To summarise, this programme contains panchromatic data for 58 different pointings distributed over the full range in right ascension and spread over declinations of $+59\overset{\circ}{.}62$ and $-24\overset{\circ}{.}73$. These data have been used to calibrate the UKIDSS surveys Lawrence et al. 2007. During each visit the fields were usually observed with a sequence of filters, either JHK or $ZYJHK$ within a few minutes. This led to an irregular sampling with fields reobserved roughly on a daily basis, although longer time gaps are common, and of course large seasonal gaps are also present in the data set. For more information about design, the details of the data curation procedures, the layout, and about variability analysis on this database are described in detail in Hambly et al. 2008, Cross et al. 2009, and Ferreira Lopes et al. 2015.

The multi-waveband data were well fitted to test the statistical parameters using different wavebands ($ZYJHK$). Moreover, Ferreira Lopes et al. 2015 and Ferreira Lopes & Cross 2016 performed a comprehensive stellar variability analysis of the WFCAMCAL08B characterizing the photometric data and identifying 319 stars (WVSC1), of which 275 are classified as periodic variable stars and 44 objects as suspected variables or apparently aperiodic variables. In this paper we analyze the same sample from Ferreira Lopes et al. 2015 and Ferreira Lopes & Cross 2016. First, we selected all sources classified as a star or probable star having at least ten unflagged epochs in any of the five filters. This selection was performed from an initial database of 216, 722 stars. Next we test the efficiency of selection of variable stars using the statistical parameters presented in Sect. 3.

We compute all statistical parameters displayed in the Table 1 by the following algorithm: the photometry measured by the best aperture was selected; next the measurements having flags (`ppErrBits`) higher than 256 were removed; finally, we only consider the measurements having values $2 \times ED_{\sigma\mu}$ about EA_M. The last step was adopted to reduce the dispersion and improve the estimation of the statistical shape parameters. On the other hand it also mis-selects binary stars having few measurements at the eclipse, for instance (see Sect. 6.1 for more details). Such a step decreases the number of stochastic variations a lot but it also means that we can miss some variable stars (see table 3). Figure 5 shows even-statistical parameters and the standard deviation as a function of the K band magnitude. The noise model (Strateva) functions as well as WVSC1 are set by lines and black dots, respectively. The main results can be summarized as:

- The dispersion even parameters have a similar range than that found for the standard deviation. Moreover, the majority of WVSC1 stars have values above the stochastic variations and so these parameters can be used in the same fashion as the standard deviation to discriminate variable stars form noise. As expected, the diagram of ED is equal to ED_M as well as being similar to ED_μ.
- The Strateva and the modified-Strateva functions show similar values for almost all ranges of magnitude. The difference is a slope at lower magnitudes (bright stars) found for the modified-Strateva function. This allows us to reduce the mis-selection but we also remove some variable bright stars having small amplitude variations. A note of caution, Strateva and the modified-Strateva functions can present an incorrect model for fainter magnitudes since a small decrease in the dispersion is found. In these cases a magnitude limit can be adopted (Cross et al. 2009).
- The shape-even parameters give a good discrimination for many variable stars particularly for bright stars (see Fig. 5). However, almost all faint stars (magnitudes greater than $\sim$ 16) have values near to that found for stochastic variations. In this sense, the dispersion-parameters are better than shape-parameters at discriminating non-stochastic variations since we can see a clearer separation among them for all ranges of magnitude. The shape parameters may be useful to discriminate different kinds of light curve signatures and this will be addressed in a future paper in this series.

In summary, the even statistical parameters can be used in the same fashion as previous ones. The main goal of this paper is to study the criteria of selection of variable stars from noise and meanwhile these parameters may be useful for many other purposes in different branches of science and technology.

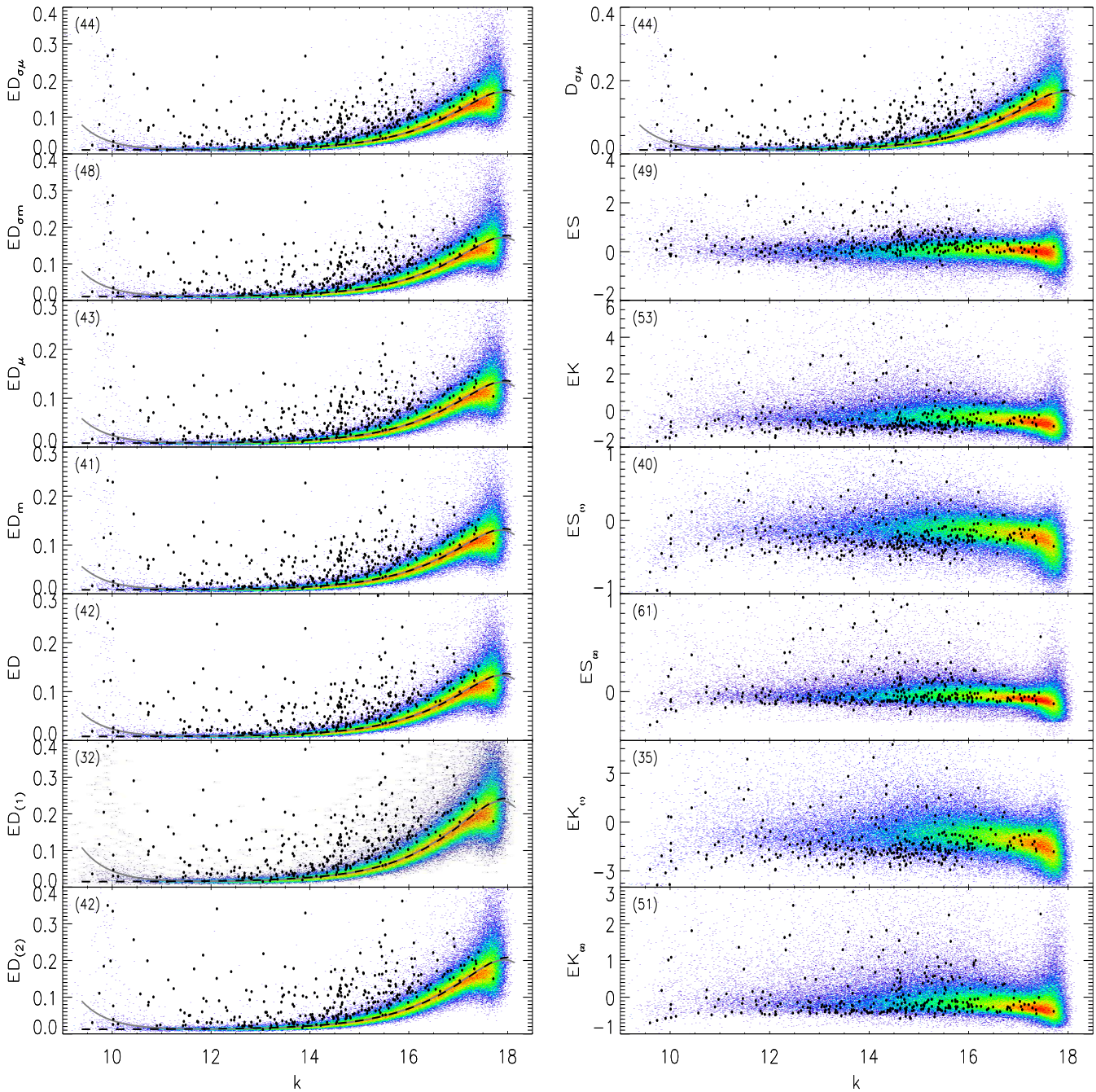


Fig. 5: Dispersion and shape parameters as function of magnitude where the black dots indicate the WVSC1 stars. The grey and dashed black lines mark the modified-Strateva and Strateva functions, respectively. The maximum number of sources per pixel is shown in brackets in each panel.

6.1. Testing photometric apertures

In order to test the dependence of the photometric aperture and extreme measures on the selection criteria the WFCAM analyzes were performed in seven different ways;

- **A1-5** Photometric measurements using a standard photometric aperture from 1 to 5 (0.5″, $\sqrt{(}0.5)''$, 1″, $\sqrt{(}2)''$, and 2″ radius, respectively);
- **BA** Photometric measurements using the best aperture (see Cross et al. 2009);
- **BAS** all measurements enclosed in $2 \times ED_{\sigma\mu}$ about EA_M of BA photometry are used;

where the measurements having flags higher than 256 were removed. The 3rd aperture (A3), corresponds to the default 1″ aperture, where the radius is slightly larger than the typical seeing FWHM, so an aperture centred on a point-source should contain > 95% of the light (in the ideal gaussian case - in reality a lot more is in lower surface brightness wings). Increasing the aperture size will increase the amount of signal, but at the expense of increasing the amount of sky too, such that the signal-to-noise decreases. Decreasing the aperture reduces the signal too much, also reducing the signal-to-noise ratio. Usually A3 gives the optimal signal-to-noise, but sometimes, nearby stars can affect the

Table 3: The efficiency metric E_{tot} (ratio of the number of selected sources by the total number of WVSC1 variable stars) and E_{WVSC1} (the ratio of the number of WVSC1 stars selected by the total number of WVSC1 stars), and α values computed from X_f for each waveband as well as using all $ZYJHK$ wavebands using $\beta = 4$. It was performed for different WFCAMCAL photometric apertures.

		Z		Y		J		H		K		YZJHK	
	E_{WVSC1}	α	E_{tot}	α	E_{tot}	α	E_{tot}	α	E_{tot}	α	E_{tot}	α	E_{tot}
	0.60	1.68	15.05	1.92	14.41	1.68	15.55	1.65	17.09	1.44	20.32	1.57	13.34
	0.65	1.55	19.40	1.74	19.30	1.59	18.46	1.50	22.35	1.35	25.51	1.46	16.68
	0.70	1.41	26.50	1.64	22.71	1.50	22.42	1.38	28.31	1.26	33.07	1.38	20.13
A1	0.75	1.26	38.66	1.51	28.65	1.37	30.78	1.30	34.25	1.17	45.64	1.27	26.32
	0.80	1.17	50.09	1.42	34.66	1.30	37.46	1.22	43.20	1.08	67.19	1.20	32.14
	0.85	1.11	60.61	1.29	47.59	1.23	46.24	1.11	64.89	0.98	110.34	1.09	46.92
	0.90	0.99	94.34	1.12	83.03	1.13	65.90	1.01	103.68	0.88	184.26	0.99	73.60
	0.60	1.75	14.20	2.06	7.84	1.80	7.94	1.76	8.73	1.57	10.59	1.67	6.63
	0.65	1.62	17.75	1.88	10.36	1.67	10.50	1.62	11.55	1.43	15.28	1.58	8.15
	0.70	1.46	24.14	1.74	13.25	1.56	13.87	1.47	16.22	1.32	21.76	1.44	11.59
A2	0.75	1.31	33.55	1.64	16.26	1.48	17.43	1.38	20.70	1.25	27.89	1.34	15.41
	0.80	1.23	41.27	1.54	20.30	1.39	22.59	1.29	27.24	1.15	42.38	1.27	19.35
	0.85	1.18	47.77	1.41	27.97	1.31	29.27	1.20	37.58	1.07	62.71	1.19	25.81
	0.90	1.03	78.98	1.20	54.83	1.19	45.32	1.12	53.23	0.94	130.12	1.11	36.35
	0.60	1.87	16.30	2.46	5.27	2.07	4.98	1.95	4.89	1.64	8.08	1.86	4.51
	0.65	1.76	19.05	2.20	6.96	1.93	6.47	1.80	6.32	1.49	12.35	1.74	5.59
	0.70	1.63	23.17	2.00	8.95	1.80	8.49	1.67	8.18	1.41	15.96	1.60	7.49
A3	0.75	1.47	30.48	1.92	10.17	1.70	10.67	1.59	9.82	1.32	22.12	1.50	9.64
	0.80	1.41	34.21	1.78	12.73	1.58	14.24	1.44	14.66	1.25	29.43	1.41	12.38
	0.85	1.29	44.22	1.67	15.94	1.45	20.68	1.33	21.48	1.16	44.15	1.33	16.03
	0.90	1.09	74.03	1.45	27.35	1.32	31.60	1.20	36.69	1.04	81.88	1.19	27.40
	0.60	1.87	19.36	2.38	8.25	2.05	7.34	1.90	7.61	1.53	15.62	1.78	7.27
	0.65	1.76	22.82	2.22	9.84	1.92	9.41	1.72	10.54	1.45	19.99	1.66	9.24
	0.70	1.54	32.59	2.08	11.87	1.79	12.20	1.62	13.34	1.34	28.88	1.56	11.75
A4	0.75	1.44	39.17	1.82	17.68	1.60	18.84	1.51	17.55	1.29	34.52	1.47	15.01
	0.80	1.30	51.76	1.67	23.82	1.49	25.31	1.41	23.47	1.20	49.01	1.31	24.61
	0.85	1.14	75.44	1.59	28.31	1.36	37.35	1.30	34.53	1.10	74.73	1.20	36.73
	0.90	1.00	111.85	1.39	47.61	1.26	52.13	1.15	64.66	1.01	111.72	1.11	54.13
	0.60	1.69	30.28	2.17	13.08	1.82	15.62	1.76	14.04	1.41	26.23	1.62	15.85
	0.65	1.54	38.71	2.06	15.31	1.75	18.01	1.60	19.47	1.30	36.15	1.52	20.42
	0.70	1.42	48.16	1.87	20.72	1.60	24.93	1.47	26.92	1.24	43.90	1.43	26.19
A5	0.75	1.30	60.91	1.69	28.68	1.46	35.14	1.35	37.96	1.20	50.10	1.31	37.77
	0.80	1.18	78.96	1.50	43.50	1.37	45.16	1.28	47.42	1.06	83.05	1.15	65.30
	0.85	1.04	111.12	1.35	63.83	1.21	73.63	1.20	62.46	1.00	103.14	1.08	86.08
	0.90	0.92	155.17	1.25	84.10	1.12	98.49	1.06	103.26	0.90	145.29	0.95	148.01
	0.60	1.95	16.54	2.58	6.41	2.16	5.86	2.15	5.86	1.77	7.34	1.99	4.97
	0.65	1.83	19.44	2.37	7.75	2.06	6.87	1.98	7.20	1.62	10.37	1.84	6.20
	0.70	1.74	22.06	2.15	9.90	1.92	8.60	1.82	9.23	1.52	13.31	1.69	7.93
BA	0.75	1.56	29.10	2.02	11.65	1.80	10.96	1.70	11.23	1.40	19.09	1.57	10.34
	0.80	1.44	35.92	1.90	13.70	1.63	15.95	1.60	13.75	1.30	26.97	1.48	12.79
	0.85	1.27	50.29	1.74	17.66	1.53	20.72	1.42	21.57	1.22	36.24	1.35	18.20
	0.90	1.14	69.10	1.53	27.58	1.41	29.10	1.29	32.88	1.09	65.17	1.25	25.85
	0.60	2.08	8.32	2.60	4.05	2.22	3.44	2.13	3.29	1.78	3.29	2.00	3.64
	0.65	1.94	9.72	2.38	4.70	2.11	3.79	1.97	3.93	1.68	4.16	1.88	4.18
	0.70	1.85	10.76	2.18	5.53	1.96	4.57	1.85	4.61	1.56	5.70	1.77	4.82
BAS	0.75	1.70	13.08	2.08	6.10	1.83	5.45	1.76	5.20	1.48	7.45	1.66	5.67
	0.80	1.59	15.27	1.95	7.04	1.72	6.73	1.64	6.45	1.36	11.85	1.55	6.74
	0.85	1.36	23.43	1.79	8.69	1.56	9.69	1.46	10.26	1.27	18.13	1.44	8.40
	0.90	1.16	41.06	1.56	13.16	1.43	14.57	1.30	18.83	1.15	35.23	1.28	13.47

measurements by adding in additional noise component from deblending images which relies on some imperfect modelling, and selecting a smaller aperture which will include less signal from the neighbour gives better results, which is why a variable aperture was selected by Cross et al. (2009).

We compute E_{tot} (ratio of the total number of sources selected to the total number of variable stars of WVSC1 catalog) for different values of E_{WVSC1} (ratio of number of selected variables stars in WVSC1 to the total number of variable stars in WVSC1). They were computed for each waveband as well as considering all wavebands ($ZYJHK$). Table 3 display α and its respective efficiency coefficient values. Such parameters allow us to analyze the efficiency of selection of variable stars from noise in the WFCAMCAL database using the WVSC1 set of variable stars as comparison stars.

Indeed, the X indices using all bands are more efficient than those found using single wavebands (see Fig. 7). The best result was found for the J waveband rather that for the Z and K wavebands. The efficiency decrease found for K waveband is related with the decrease of signal-to-noise while for Z waveband we find that the c_0 in Eqn 17 is significantly higher ($\sim$ 0.023 c.f. $\sim$ 0.014 for Y, J, H, K), which suggests greater across detector variations, since simple offsets in the zeropoint would be corrected by recalibration done by Cross et al. (2009). Calibrating the Z and Y bands was trickier than J, H, K because the calibration is extrapolated from 2MASS J, H, Ks (see Hodgkin et al. 2009), and more susceptible to extinction, particularly in the Z-band, which can vary on small scales in star forming regions.

We also perform the selection using the previous normal procedure to select variable stars using non-correlated data, i.e. select all sources with an magnitude RMS above n times sigma above the noise model function. We compute the standard deviation and the X index for K waveband using BA photometry. Considering one sigma above the Strateva function $\sim$ 81% of WVSC1 stars are selected but at the expense of an $E_{tot} \simeq 103$. Such an E_{tot} value is 7.7 larger than that found using our approach. This means that the modified-Strateva function joined with our empirical approach (see Eqs. 17 and 20) and statistical weights (see Sect. 4) increases the selection efficiency by about $\sim$ 770%.

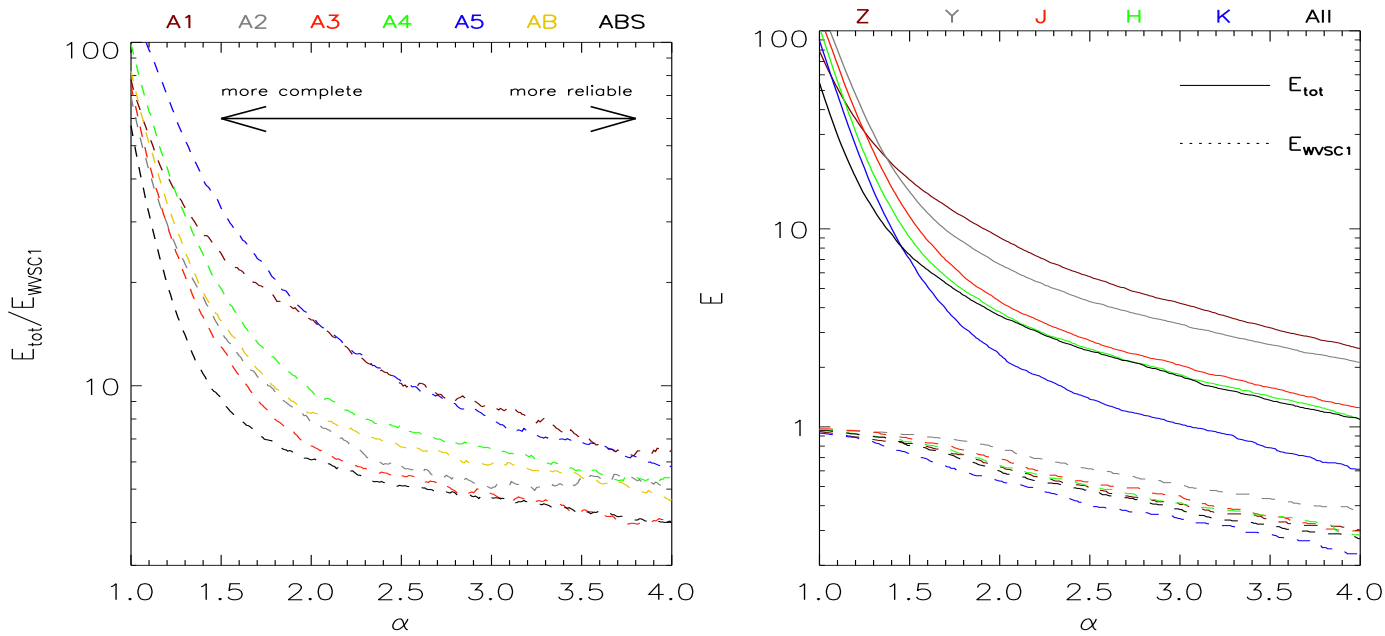


Fig. 6: The right panel shows E_{tot}/E_{WVSC1} for all apertures analyzed while the left panel shows the E_{tot} and E_{WVSC1} values for each waveband as well as their combinations ($ZYJHK$ wavebands) using BAS. Here the result for each photometric aperture and waveband are shown by different colours. Higher α values mean more reliable selection (lesser misclassification) against for more complete selection.

Table 4: The efficiency metric E_{tot}, E_{WVSC1}, and α_{cor} values computed from analyse of BA photometry for $K^{(2)}_{(fi)}$ and $K^{(3)}_{(fi)}$. α_{cor} is the values regarding to the Eq. 16 of Ferreira Lopes & Cross (2016).

	s = 2		s = 3	
E_{WVSC1}	α_{cor}	E_{tot}	α_{cor}	E_{tot}
0.60	0.21	1.44	0.30	0.86
0.65	0.22	1.54	0.33	0.97
0.70	0.24	1.77	0.35	1.11
0.75	0.25	1.97	0.38	1.31
0.80	0.27	2.27	0.40	1.55
0.85	0.28	2.74	0.43	1.92
0.90	0.30	3.38	0.46	2.65

7. Improvements on correlated indices and combined indices

The flux independent variability indices ($I^{(s)}_{fi}$) proposed by us (for more details see Ferreira Lopes et al. 2015; Ferreira Lopes & Cross 2016) are not dependent on the amplitude signal since they only use the correlation signal. However, they are dependent on the mean value. Therefore, the correlation values computed using even-averages are more accurate than those computed using mean value since the even-mean gives a value closer to the true center (see Sect. 3.1). As a result, the E_{tot} values presented in the Table 4 are reduced by about ~ 18% compared to those found in the Table 2 of paper I.

We also test the correlated indices for BAS, i.e. all measurements enclosed in $2 \times ED_{\sigma\mu}$ about EA_M of BA photometry. The results are not so different from those found for $E_{WVSC1} < 0.85$ (see Table 4) while for $E_{WVSC1} > 0.85$ we found an E_{tot} about 40% higher. The measurements related to eclipsing binary stars are removed when we use BAS. The correlated and non-correlated indices can fail for low signal-to-noise variations and for non-contact binaries having few measurements at the eclipses.

The WFCAMCAL database allows us to compute correlated indices having number of correlations higher than $N_2^{(min)}$ for about ~ 94% of data. Variable stars having fewer correlations or not previously detected will be explored in the next paper of this series, where we are going to propose a new periodicity search method as well as studying selection criteria to produce a clearer sample.

8. Conclusions

Statistical parameters were analyzed as a tool to discriminate variable stars from noise. We propose 15 new statistical parameters based on even-statistics, where seven of them are unbound parameters, i.e. they are independent of the average. However, they keep a strong relationship with previous ones. The even-mean and even-median are more accurate than the mean and median. It allows us to improve the efficiency level of correlated indices by about 14%. Moreover, we can also improve the accuracy of dispersion and shape parameters previous defined by using the even averages.

We use Monte Carlo simulations to test the new statistical parameters and propose weights to take account of the variations of statistical parameters as a function of the number of measurements. Moreover, we propose non-correlated indices that combine all dispersion parameters. It allows us to improve the estimation of dispersion as well as combine different wavebands and so reduce the misselection rate.

Finally, we also test the new statistical parameters as well as variability indices on the WFCAMCAL database. Using these data, we verify that the even parameters work in the same way as the previous ones. And so, we increase our inventory of tools to assess statistical distributions. Moreover, the approach proposed in this paper returns a misselection rate 770% less than previous similar statistical methods. Moreover, from a combination

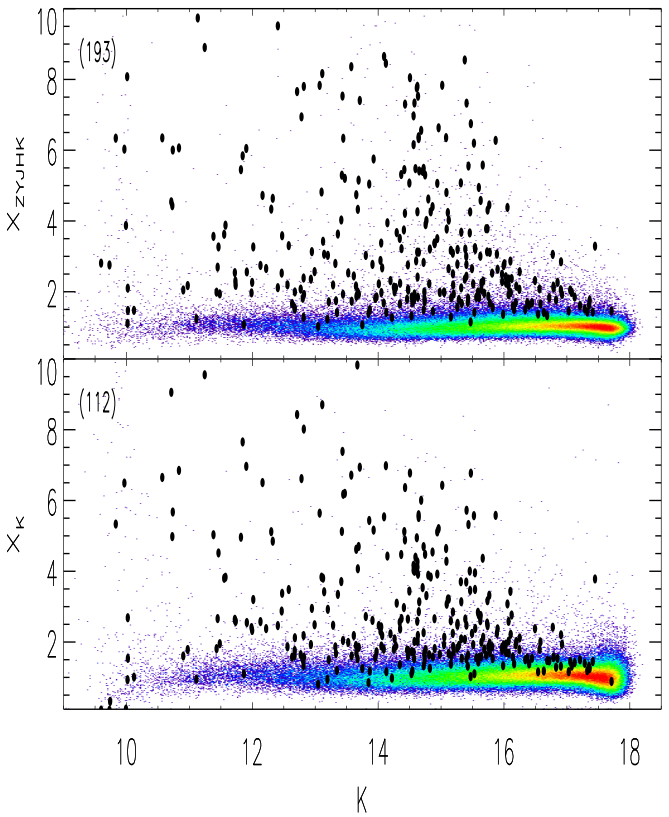
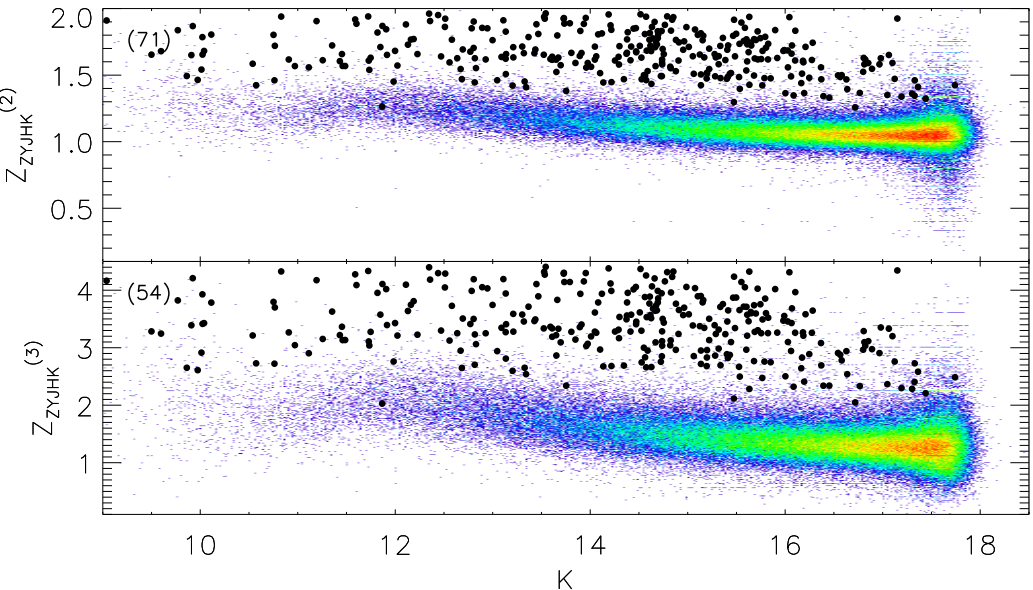


Fig. 7: X_f indices for K waveband (X_K bottom panel) and for all wavebands (X_{ZYJHK} top panel) as a function of magnitude. The maximum number of sources per pixel is shown in brackets in each panel.

of correlated and non-correlated indices we can reduce the mis-selection for 1 in every 3 sources selected. We have concluded that the first step of this project which was to study and improve the selection criteria for variable stars in photometric surveys. Next, we will start the second step of this project which is the study of periodicity methods and criteria to create clear samples.

References

Baglin, A., Auvergne, M., Barge, P., et al. 2007, in American Institute of Physics Conference Series, Vol. 895, Fifty Years of Romanian Astrophysics, ed. C. Dumitrache, N. A. Popescu, M. D. Suran, & V. Mioc, 201–209
Bailer-Jones, C. A. L., Andrae, R., Arcay, B., et al. 2013, A&A, 559, A74
Borucki, W. J., Koch, D., Basri, G., et al. 2010, Science, 327, 977
Brys, G., Hubert, M., & Struyf, A. 2004, Journal of Computational and Graphical Statistics, 13, 996
Cross, N. J. G., Collins, R. S., Hambly, N. C., et al. 2009, MNRAS, 399, 1730
Ferreira Lopes, C. E. & Cross, N. J. G. 2016, A&A, 586, A36
Ferreira Lopes, C. E., Dékány, I., Catelan, M., et al. 2015, A&A, 573, A100
Hambly, N. C., Collins, R. S., Cross, N. J. G., et al. 2008, MNRAS, 384, 637
Hodgkin, S. T., Irwin, M. J., Hewett, P. C., & Warren, S. J. 2009, MNRAS, 394, 675
Hoffman, D. I., Harrison, T. E., & McNamara, B. J. 2009, AJ, 138, 466
Kaiser, N., Aussel, H., Burke, B. E., et al. 2002, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4836, Survey and Other Telescope Technologies and Discoveries, ed. J. A. Tyson & S. Wolff, 154–164
Lawrence, A., Warren, S. J., Almaini, O., et al. 2007, MNRAS, 379, 1599
Levenberg, K. 1944, Quart. Appl. Math., 2, 164
Marquardt, D. 1963, SIAM J. Appl. Math., 11, 431
Minniti, D., Lucas, P. W., Emerson, J. P., et al. 2010, New A, 15, 433
Pollacco, D. L., Skillen, I., Collier Cameron, A., et al. 2006, PASP, 118, 1407
Sesar, B., Ivezić, Ž., Lupton, R. H., et al. 2007, AJ, 134, 2236
Strateva, I., Ivezić, Ž., Knapp, G. R., et al. 2001, AJ, 122, 1861
Udalski, A. 2003, Acta Astron., 53, 291


(71)
(54)
$Z_{ZYJHK}^{(2)}$
$Z_{ZYJHK}^{(3)}$
K
2.0
1.5
1.0
0.5
4
3
2
1
10
12
14
16
18