Answers to referee’s report

We would like to thank the referee for his/her report. We have addressed all the corrections/comments/suggestions in the report below, and marked the changes in the article with boldface .

General revisions

a) *In the introduction, when they review the historical approaches to the problem of stellar cluster membership selection, they go directly from Sanders (1971) to UPMASK (Krone-Martins & Moitinho 2014). A jump of 43 years without there being any contribution worthy of mention until the arrival of UPMASK! I have appended a list of articles devoted to this purpose and published up until 2020. In particular, the first method that used KDE for determining membership probabilities in a 4-dimensional space (positions and proper motions) was published in 1990 by Cabrera–Cano & Alfaro in A&A, and was the basis for the further development of non-parametric methods designed to this end. This concerns the methodology, but if we look at the variables used, two co authors of this study participated in the design of a method to select cluster members using an N-dimensional space including the photometric subspace (ASTECA; Perren et al. 2015). I believe that, since UPMASK was developed with the objective of obtaining cluster members from photometric data and that pyUPMASK is an improvement and updating of UPMASK, a historical review of the different methods that have introduced photometry into the analysis and what the contribution of each one is, should be included in the introduction.*

Both the articles for UPMASK (Krone-Martins & Moitinho 2014; Sect. 1. ‘*Introduction’*) and ASteCA (Perren et al. 2015; Sect. 2.8 ‘*Field star decontamination’*) contain large sections commenting on the history of membership estimation in star clusters with lots of references. We have added to the Introduction a few of the articles mentioned by the referee, and directed the reader to those two references for a more detailed recount.

b) *In the comparison of pyUPMASK and UPMASK, they use two sets of variables: PHOT (positions and photometry (reduced to 4D)), and PM (positions and proper motions). Figure 6 summarizes, on the basis of different figures of merit, the cases in which each code provides the best results. This analysis requires the results to be distinguished by the set of variables used. Since UPMASK only uses positions and photometry, it would* *be of great interest to know whether pyUPMASK improves the performance only with the PHOT data and not mixing them with the PM. That is to say, when only using positions and photometry, does pyUPMASK give a better classification than UPMASK or does it merely speed up the process? If, however, pyUPMASK, working only with PM, provides better figures of merit than UPMASK (working only with photometry), the conclusion could be that the proper motions contain more or better information than the photometry concerning their membership of the stellar system, but that the new algorithm does not significantly improve the classification when we use PHOT.*

- "*Since UPMASK only uses positions and photometry*" this is not correct. Both methods were tested using the same sets of synthetic data, i.e. PHOT and PM. UPMASK is not limited to photometric data, even though that was the only data employed in the Krone-Martins & Moitinho (2014) article. All the performance analysis values shown (including of course Fig 6) are comparable between both methods, because both methods analyzed the exact same set of synthetic clusters (PHOT + PM). We have made this point more clear in Sect 3.1

This fact notwithstanding we have added an appendix showing Fig 6 but segregated between results for the PHOT and PM datasets, for clarity. This is mentioned in Sect. 4.

c) *It seems surprising that there is no comparison of pyUPMASK with an analysis based on the ASTECA toolkit. As is mentioned in the article, UPMASK has had more than 50 citations since 2014/01 (52 in NASA-ADS to date), but ASTECA has had 49 since 2015/04. Both are very widespread and both use photometric data and are in an open code. Furthermore, pyUPMASK and ASTECA are participated on by two co-authors in common. It is important to know which code gives us the best information, including the photometric data. Gaia not only provides us with astrometry but also an excellent collection of homogeneous photometric data.*

ASteCA was purposely left out of the article because the method employed by this code to assess membership probabilities is not directly comparable with either UPMASK or pyUPMASK. Unlike these two methods which are **unsupervised**, ASteCA requires that the cluster region and field region are a priori defined. This is done by estimating the center and radius of the cluster (and a surrounding field region) before the membership algorithm can be applied. This means that ASteCA uses a **supervised** method of membership estimation, as one of the classes (field stars) must be clearly identified and the other (cluster stars) approximately identified (as it is contaminated by field stars).

Being non-comparable methods, we do not believe that adding ASteCA to the analysis is reasonable. We have explained this point clearly in Sect 2.

Specific issues

a) *The following sentences appear on page 6 of the referee copy: "The last three methods (AGG, KNN, VOR) have a characteristic in common: no stochastic process or approximation is employed by either of them. In other words, they are deterministic.* ***This means that, for the same input data, the exact same result (i.e., clustering) will be obtained for different runs****." This last sentence can lead to confusion: assuming that I use the same input data and parameters, different runs will lead to one single result. Is this correct?*

Yes, that is correct. We have edited this sentence with a wording closely resembling the one used by the referee for clarity.

b) *Equation (6), which represents a spatial model for cluster and field together, with the field defined by a uniform distribution, has already been used previously in combination with another model for the proper motions. The first article where this combination was put forward was "Astrometric Criteria for Selecting Physical Members of Open Clusters with Low Astrometric Precision - Application to NGC559", De Graeve, E.; Publications of the Vatican Observatory V.1:16, P. 1, 1979. Another later citation, which includes an exponential model rather than a Gaussian one for the spatial distribution of the cluster is: \bibitem[Jones \& Walker(1988)]{1988AJ.....95.1755J} Jones, B.~F. \& Walker, M.~F.\ 1988, \aj, 95, 1755. doi:10.1086/114773.*

We do not have access to the 1979 reference, but we have included a mention to the 1988 reference in Sect. 2.2.3.

c) *The following sentences (at the beginning of section 2.2.4) are also confusing: "Once a run of the inner loop is finished, each star in the observed field is classified to be either a cluster member or a field star. This is a hard binary classification, meaning that only probability values of 0 and 1 are assigned. The KDE block takes these binary probabilities and turns them into continuous probabilities in the range [0, 1]". As is described in Section 2.2.3 and Fig. 2, a membership probability has already been estimated through its spatial distribution that is not only limited to 0 or 1. It seems to me that this paragraph needs clarifying or rewriting. As it stands, I understand that although a continuous spatial probability is used to carry out a first classification between member and non-member, the information that moves on to the next segment is only the 0 or 1 classification, but not the probability in the subspace of the positions.*

- "*a membership probability has already been estimated through its spatial distribution that is not only limited to 0 or 1*", this is correct.
- "*although a continuous spatial probability is used to carry out a first classification between member and non-member, the information that moves on to the next segment is only the 0 or 1 classification, but not the probability in the subspace of the positions.*", this is also correct.

We have modified Sect 2.2.4 to make this more clear.

d) *"...the contamination index (CI), defined as the number of field stars to cluster members in the frame. The maximum CI in our set of synthetic clusters is 200." The quoted sentences define CI as nf/ncl. However, in ASTECA the CI was set as nf/(nf+ncl). Evidently, both express the field star contamination. Yet it seems strange that "CI" (same notation) is defined in two different ways in two packets (and their corresponding papers) that share two authors. Perhaps simply changing the notation of CI to CIpy, or something similar, could make things a little clearer.*

The referee is correct in noting that both definitions do not match. The CI definition used here (nf/ncl) was selected to match the "contamination rate" used in the UPMASK article (Krone-Martins & Moitinho 2014). We have made this clear in Sect 3.1

e) *There are two other sentences that seem contradictory and might lead to confusion. In subsection 3.1 (Synthetic datasets), p. 11, it reads: "The first sub-set is equivalent to that used in the original UPMASK article (KMM14), as it is composed of clusters with synthetic photometry generated with the same process as that used in KMM14.* ***We will refer to this sub-set as PHOT hereinafter.*** *The* ***second sub-set contains 280 clusters and, although it also contains synthetic photometry, it was generated adding synthetic proper motions to all the stars in the frame."*** *From this reading, one can infer that the second dataset contains PMs and photometry. However, in subsection 3.3 (Input Parameters Selection), p.17, it states: "The PHOT set was processed using all the available photometry as input (V; B-V;U-B; V-I; J-H; H K) but selecting only the four principal dimensions after the principal component analysis dimensionality reduction.* ***For the PM set we used only the proper motions, and no photometry.*** *" The variables that form the PHOT subset and the PM subset need to be made clear, since the comparisons between the performances of different methods should be done with the same sets of variables. This links directly with my last point in "general considerations", and with the next point.*

- "*one can infer that the second dataset contains PMs and photometry.",* that is correct it does.
- "***For the PM set we used only the proper motions, and no photometry"*** this means that although the PM set contains photometry in addition to proper motions, we did not use this photometric data in the analysis.
- "*the comparisons between the performances of different methods should be done with the same sets of variables",* as explained in **'General revisions' b)** this is precisely what we did.

Sect 3.1 was modified to make this more clear, as explained in '**General revisions**' **b)**. We also removed the mention to the photometry of the PM set in Sect 3.3, to avoid confusion.

f) *Also on p.17, it says: "Proper motions are generally regarded as better cluster members discriminators than photometry. We were able to confirm this by checking that the results (with either UPMASK or pyUPMASK)* ***degraded if photometry was added to the proper motions as input data for the PM set****." This is a very interesting conclusion, which leads us to think that the comparative analysis shown in Fig. 6 does not provide the appropriate information, and that the results of the performances of the two methods, when UPMASK only works with PHOT and pyUPMASK with PHOT or PM, should not be mixed.*

- "*when UPMASK only works with PHOT* " this is not correct, as mentioned in '**General revisions**' **b)**.