Report on the article, “pyUPMASK: an improved unsupervised clustering algorithm” by M. S. Pera, G. I. Perren, A. Moitinho, H. D. Navone, and R. A. Vazquez. (A&A)

General assessment:

The article is presented as a new method of “field decontamination” or “cluster membership analysis” for a collection of multi-dimensional data. The method, called pyUPMASK, is put forward as an improvement on one previously published by one of the co-authors, called UPMASK. In the introduction, they state that UPMASK was designed to select cluster members using the positions on the sky and photometry. In this study, the type of variables used is generalized and increased, although the validation of the method is only based on two sets of variables: PHOT (positions and photometry (reduced to 4D)) and PM (positions and proper motions); see section 3.3.

The authors include several innovations. In the analysis of the positions (first loop), they improve the speed and reliability in detecting possible groupings compatible with a uniform distribution. They also include various methods of clustering analysis, which can be chosen by the user depending on their scientific objective, on the variables available and on the computing capabilities they have access to.

An interesting set of figures of merit is also included to evaluate the goodness of the classification in each case, as well as the estimated computing time for the pyUPMASK and UPMASK codes.

The aims of the article are appropriate and suitable, one can appreciate the large amount of computational work undertaken and value the merit of a new contribution to the pre-existing set of algorithms and mathematical tools for the detection and analysis of stellar clusters. It is clear that the data releases from Gaia are giving a strong push to the development of new analysis tools.

Nevertheless, in its current state, the paper does not attain the standards of A&A for publication. It needs a thorough revision.

General revisions required:

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.

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.
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.
d) The writing of the method and the organization of the article need to be revised. There are some confusing aspects, some of which I mention in the section below on specific issues.

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?
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.

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.

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.

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.

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.

Summary:
The analysis carried out for the PHOT and the PM subset should be separated, comparing these different methods.

PHOT	**PM**	**PM & PHOT**
pyUPMASK - UPMASK	pyUPMASK(PM) – UPMASK(PHOT)	
pyUPMASK - ASTECA	pyUPMASK - ASTECA	pyUPMASK - ASTECA

I introduce ASTECA into the comparative analysis because, given that (a) pyUPMASK is developed by a team that has members who participated in the design of both UPMASK and ASTECA, (b) both codes have the same scientific objective, and (c) ASTECA includes a broad set of input variables, like pyUPMASK, it would therefore not be difficult to perform this comparison and elucidate the performances of each of these packets for working with different datasets.

Aside from this, the introduction needs rewriting with consideration of the contributions that are key to the development of the methods analyzed, as well as mentioning the other approaches to the problem that have been developed in recent years.

Lastly, the paper as a whole is in need of an edit to remove confusing or contradictory sentences and to help the reader attain a better understanding of the method.

Appendix: A list of publications that have dealt with the development of methods of membership analysis since 1971. Among others:

\bibitem[Cantat-Gaudin \& Anders(2020)]{2020A&A...633A..99C} Cantat-Gaudin, T. \& Anders, F.\ 2020, \aap, 633, A99.
doi:10.1051/0004-6361/201936691

\bibitem[Balaguer-N{\'u}{\~n}ez et al.(2020)]{2020MNRAS.492.5811B} Balaguer-N{\'u}{\~n}ez, L., L{\'o}pez del Fresno, M., Solano, E., et al.\ 2020, \mnras, 492, 5811. doi:10.1093/mnras/stz3610

\bibitem[Xu et al.(2019)]{2019ChA&A..43..225X} Xu, S.-. kun ., Wang, C., Zhuang, L.-. hua ., et al.\ 2019, \caa, 43, 225. doi:10.1016/j.chinastron.2019.04.001

\bibitem[Gao(2018)]{2018PASJ...70...68G} Gao, X.-H.\ 2018, \pasj, 70, 68. doi:10.1093/pasj/psy059

\bibitem[Sampedro \& Alfaro(2016)]{2016MNRAS.457.3949S} Sampedro, L. \& Alfaro, E.~J.\ 2016, \mnras, 457, 3949. doi:10.1093/mnras/stw243

\bibitem[Priyatikanto \& Arifyanto(2015)]{2015PKAS...30..271P} Priyatikanto, R. \& Arifyanto, M.~I.\ 2015, Publication of Korean Astronomical Society, 30, 271. doi:10.5303/PKAS.2015.30.2.271

\bibitem[Gao et al.(2014)]{2014ChA&A..38..257G} Gao, X.-. hua ., Chen, L., \& Hou, Z.-. jie .\ 2014, \caa, 38, 257. doi:10.1016/j.chinastron.2014.07.004

\bibitem[Javakhishvili et al.(2006)]{2006A&A...447..915J} Javakhishvili, G., Kukhianidze, V., Todua, M., et al.\ 2006, \aap, 447, 915. doi:10.1051/0004-6361:20040297

\bibitem[Balaguer-N{\'u}{\~n}ez et al.(2004)]{2004A&A...426..819B} Balaguer-N{\'u}{\~n}ez, L., Jordi, C., Galad{\'\i}-Enr{\'\i}quez, D., et al.\ 2004, \aap, 426, 819. doi:10.1051/0004-6361:20041332

\bibitem[Platais et al.(1998)]{1998AJ....116.2423P} Platais, I., Kozhurina-Platais, V., \& van Leeuwen, F.\ 1998, \aj, 116, 2423. doi:10.1086/300606

\bibitem[Galadi-Enriquez et al.(1998)]{1998A&A...337..125G} Galadi-Enriquez, D., Jordi, C., \& Trullols, E.\ 1998, \aap, 337, 125

\bibitem[Chen et al.(1997)]{1997A&A...318...29C} Chen, B., Asiain, R., Figueras, F., et al.\ 1997, \aap, 318, 29

\bibitem[Cabrera-Cano \& Alfaro(1990)]{1990A&A...235...94C} Cabrera-Cano, J. \& Alfaro, E.~J.\ 1990, \aap, 235, 94

\bibitem[Cabrera-Cano \& Alfaro(1985)]{1985A&A...150..298C} Cabrera-Cano, J. \& Alfaro, E.~J.\ 1985, \aap, 150, 298

\bibitem[Zhao et al.(1982)]{1982ChA&A...6..293Z} Zhao, J.-. liang ., Tian, K.-. ping ., Xu, Z.-. hai ., et al.\ 1982, \caa, 6, 293. doi:10.1016/0275-1062(82)90004-2

\bibitem[Slovak(1977)]{1977AJ.....82..818S} Slovak, M.~H.\ 1977, \aj, 82, 818. doi:10.1086/112132

\bibitem[McNamara \& Sanders(1976)]{1976A&A....52...53M} McNamara, B.~J. \& Sanders, W.~L.\ 1976, \aap, 52, 53