Highlights on Spanish Astrophysics X, Proceedings of the XIII Scientific Meeting of the Spanish Astronomical Society held on July 16 – 20, 2018, in Salamanca, Spain. B. Montesinos, A. Asensio Ramos, F. Buitrago, R. Schödel, E. Villaver, S. Pérez-Hoyos, I. Ordóñez-Etxeberria (eds.), 2019

Proper motion and other challenges in cross-matching *Gaia* observations.

F. Torra[1], M. Clotet[1], J. J. González-Vidal[1], C. Fabricius[1] and J. Castañeda[1]

[1] Institut de Ciències del Cosmos, Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, E08028 Barcelona, Spain.

Abstract

The cross-matching (XM) in *Gaia* is a sophisticated process that provides a consistent match between observations and sources in the working catalogue for subsequent data reduction processes

Although the fraction of high proper motion stars that *Gaia* observes is small, their absolute number is not, and therefore the proper motion as well as other parameters have to be taken into account in the cross-matching of *Gaia* observations. In consequence, we describe the improvements and the identification of new proper motion sources thanks to a generalized algorithm based on clustering analysis, and a post-processing algorithm which identifies variable stars.

These improvements with respect the *Gaia* DR2 catalogue will imply a better identification of the observations of these kinds of stars and more precise astrometric and photometric parameters for subsequent data releases.

1 Introduction

Gaia [5] is a mission of the European Space Agency (ESA) which aims to measure the positions, motions and distances of more than 1 billion stars producing the most precise three-dimensional map of our Galaxy. Before the astrometric and photometric reductions, a pre-processing and source list creation (presented in [3]) is necessary to determine the parameters of each of the celestial objects (sources) that *Gaia* observes and, more specifically, the observation-to-source cross-matching (XM) of *Gaia* objects which provides the link between the *Gaia* detections and the sources.

The XM process has two preparatory stages in order to isolate groups of observations and sources of a specific sky region which are matched in the final XM resolution stage (see [2]). Furthermore, the resolution stage is divided into two substages, a first clustering stage and a final conflict resolution stage to solve all conflict scenarios, as described previously in [2].

We describe a generalization of the clustering algorithm used for *Gaia* DR2 [4] and we focus on interesting cases such as the high proper motion sources not found in DR2 and the variable stars, providing significant improvements of the algorithms for *Gaia* DR3.

2 Cluster analysis

Cluster analysis aims to divide data into groups (the so-called clusters), where the objects in each cluster are similar between them and different from objects within other clusters.

The model pretends to be independent from other catalogues, so the input only consists of a set of observations and, therefore, the positions and motions have to be determined as the number of observations in the cluster increases. Following a proposal by Lindegren [6], we consider here a hierarchical agglomerative algorithm because it presupposes very little in the way of data characteristics (i.e. in our case it does not require previous knowledge of the number of clusters to be created).

In order to decide which clusters should be agglomerated a measure of dissimilarity between sets of observations is required, which is a positive semi-definite symmetric mapping of pairs of observations and/or clusters of observations onto the reals (i.e. $\Delta(C_i, C_j) \geq 0$ and $\Delta(C_i.C_j) = \Delta(C_j, C_i)$ for clusters $C_i,\ C_j$). Note that, the triangular inequality is not necessarily satisfied for our type of problem.

Moreover, we have to consider an efficient algorithm to agglomerate the observations according to the corresponding definition of the dissimilarity.

2.1 The minimum variance criterion

The dissimilarity measure chosen is the Ward's dissimilarity which is defined as the increase of the sum of squared residuals when using a common coordinates compared to the value obtained when the two terms are separately minimized,

$$\Delta(C_i, C_j) = R(C_i \cup C_j) - R(C_i) - R(C_j), \tag{1}$$

where $C_i,\ C_j$ are two disjoint clusters and $R(C)$ is the sum of squared residuals in the corresponding cluster C.

If we consider the simple model where the coordinates of the observations do not depend on time, the dissimilarity between the clusters C_i and C_j can be written as

$$\Delta(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \left\| \boldsymbol{x}(C_i) - \boldsymbol{x}(C_j) \right\|^2, \tag{2}$$

where n_i (n_j) is the number of observations in the cluster C_i (C_j), and the vector $\boldsymbol{x}(C) = \frac{1}{n} \sum_{O \in C} \boldsymbol{x}(O)$ is the cluster center given by the center of gravity of the observations in the cluster.

This dissimilarity allows to agglomerate with the minimum increase in information loss and can be generalized to a linear model such as the inclusion the proper motion.

Note that the norm used in (2) may contain weight factors if required to include more parameters such as the magnitude.

2.2 Nearest Neighbor Chain

An efficient algorithm for hierarchical clustering is the nearest neighbor chain [7] based on the construction of nearest neighbor chains and reciprocal nearest neighbors. More specifically, the algorithm builds a chain of nearest neighbors, starting from an arbitrary (agglomerable) cluster, until a pair of mutual nearest neighbours has been found and agglomerated.

In this algorithm the agglomeration is carried all the way to the point where all observations are in a single cluster but for the XM process this makes little sense.

Therefore we consider that the agglomeration only makes sense while the dispersion of residuals within the clusters is below a given limit. This dispersion is measured by the variance $\sigma^2(C) = R(C)/n$ and the limit depends on *Gaia* observation error and the model error caused by not including the parallax.

3 Inclusion of proper motion

In Section 2.1 we have supposed that the coordinates do not depend on time, but for the inclusion of the proper motion we have to consider a linear model for each direction u,

$$u(t) = u_0 + u_1 t \tag{3}$$

where u_0 is the mean position and u_1 is the proper motion.

The linear system in matricial form is

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{u} + \boldsymbol{e} \tag{4}$$

where $\boldsymbol{u} = (u_0, u_1)$, $\boldsymbol{b}$ is a $n-$vector of observations, $\boldsymbol{e}$ is a $n-$vector of observation errors, and $\boldsymbol{A}$ is a $2 \times n-$matrix with the time functions.

Therefore, applying the definition of (1) and using some equations, the dissimilarity in u-direction can be expressed

$$\Delta_u(C_i, C_j) = (\widehat{\boldsymbol{u}}_i - \widehat{\boldsymbol{u}}_j)^T \boldsymbol{N}_i(\boldsymbol{N}_i + \boldsymbol{N}_j)^{-1}\boldsymbol{N}_j(\widehat{\boldsymbol{u}}_i - \widehat{\boldsymbol{u}}_j), \tag{5}$$

where $\boldsymbol{N}_i = \boldsymbol{A}^T\boldsymbol{A}$ is the normal matrix and $\widehat{\boldsymbol{u}} = \boldsymbol{N}^{-1}(\boldsymbol{A}^{-1}\boldsymbol{b})$ minimizes the sum of squared residuals.

Note that the above equation reduces to (2) when the normal matrices are of dimension 1×1, i.e., without applying the linear model.

In the example shown in Figure 1, the XM created more than one source in DR2 because the proper motion model was not used in the clustering stage. Moreover, the observations which are processed for first time are unmatched in the input of the final stage because they are so far to be matched in the preparatory stages. However, the generalized XM algorithm including the proper motion merges the sources of the previous cycle (and therefore calculate the proper motion) thanks to the new implementation and the major number of observations.

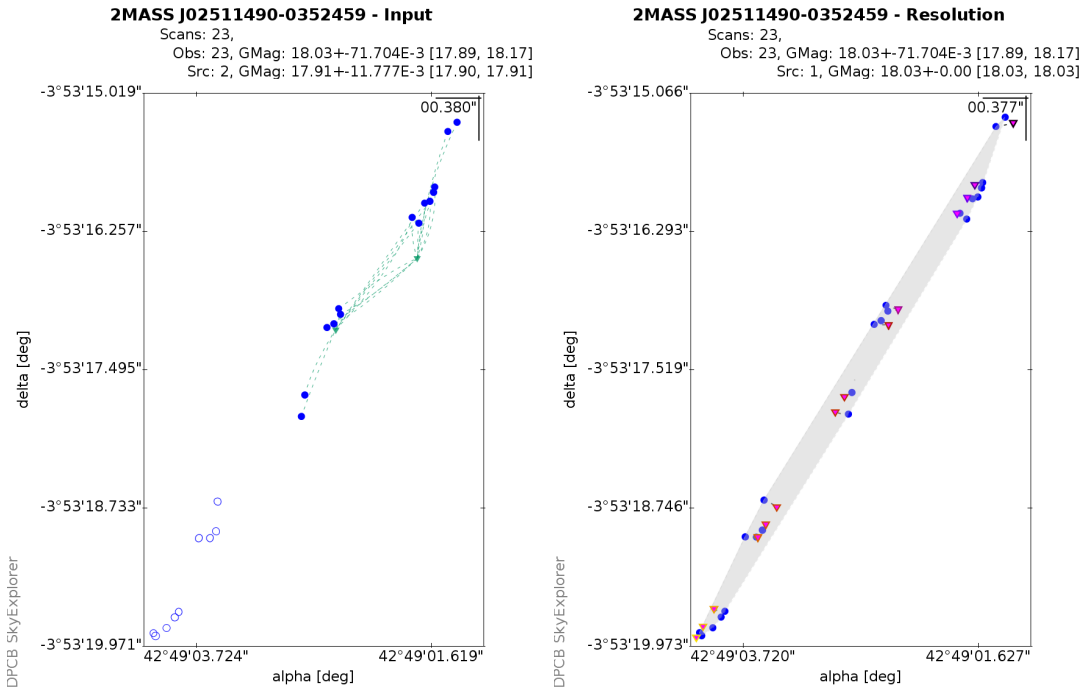

Figure 1: XM resolution around 2MASS J02511490-0352459. Left: XM resolver input including observations (blue dots, empty for unmatches in the input of final resolver), input sources (green triangles) and input resolver links (dashed green lines); right: resolution including the observations, the New Source propagated to the observation epoch (triangles) and the grey area is the cluster region.

4 Post-processing analysis

As mentioned in Section 2.1, a magnitude criterion can be included in the clustering algorithm with a scale factor which makes a magnitude error comparable with an error in position. This criterion leads to separate valid and spurious detections into different clusters, and improves the resolution in crowded areas significantly. However, it creates problems for variable stars creating several clusters at the same position (and different scans) but with different magnitude.

In this post-processing procedure, we detect these clusters with very close centres (about 120 mas) and without any common scan (i.e. compatible in time). After that, they are agglomerated into a single one without using any magnitude criterion. Therefore, this post-process avoids the creation of several sources corresponding to a variable source.

Gaia17aru is a confirmed cataclysmic variable star detected by the Gaia Science Alerts system (http://gsaweb.ast.cam.ac.uk/alerts) which runs at the Cambridge Institute of Astronomy to look for sources that suddenly change dramatically in brightness. This Alert is split into 2 clusters without the post-processing analysis, the brighter transits (outbursts) being grouped in one cluster, and the fainter with the other, whereas applying the post-

processing all the transits are grouped together with one of the input sources (see Figure 2).

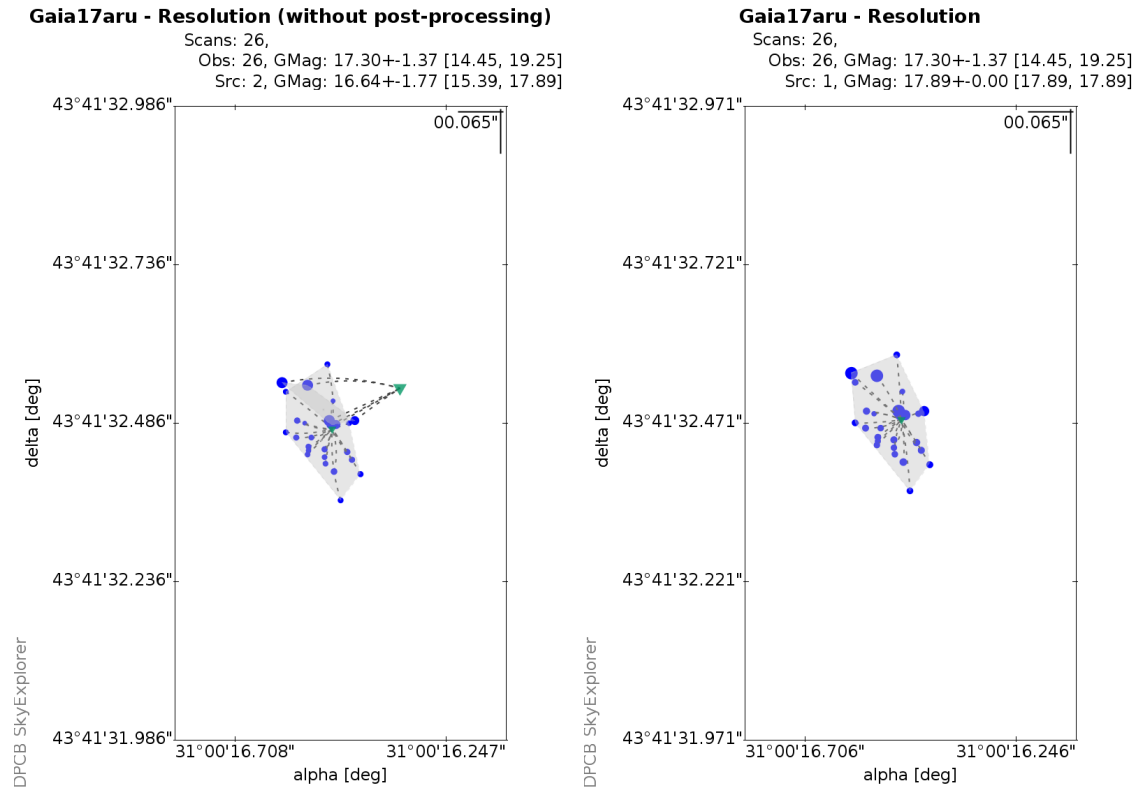

Figure 2: XM resolution around Gaia17aru. Left: XM resolver without post-processing including observations (blue dots), input sources (green triangles) and resolver links (dashed black lines); right: resolution with post-processing including the observations and the persisting source (green triangle). Grey areas are the cluster regions.

5 Conclusion

We have shown that the cross-matching algorithm for *Gaia* can include parameters such as the proper motion and the magnitude which are necessary to identify some kinds of stars and provide precise parameters of them, applying a generalization of the Ward's dissimilarity and defining suitable post-processing algorithms. Therefore, the number of high proper motion stars and variable stars will increase in *Gaia* DR3 respect to *Gaia* DR2 and in addition their parameters will be more precise because of the creation of new sources and the increasing number of matched observations.

These improvements on the clustering algorithm may imply an update of the *Gaia* DR2 identification and the astrometric and photometric parameters because some of the observations will be matched to other sources (see details on the source evolution in [1]).

Acknowledgments

This work was supported by the MINECO (Spanish Ministry of Economy) through grant ESP2016-80079-C2-1-R (MINECO/FEDER, UE) and MDM-2014-0369 of ICCUB (Unidad de Excelencia 'María de Maeztu').

References

[1] Castañeda, J., Torra, F., Clotet, M., et al. 2018, Highlights of Spanish Astrophysics X (this volume).

[2] Clotet, M., González-Vidal, J. J., Castañeda, J., et al. 2017, Highlights on Spanish Astrophysics IX, pp.634-639.

[3] Fabricius, C., Bastian, U., Portell, J., et al. 2016, A&A, 595, A3.

[4] Gaia Collaboration (Brown et al.) 2018, A&A, 616, A1.

[5] Gaia Collaboration (Prusti et al.) 2016, A&A, 595, A1.

[6] Lindegren, L. 2005, Gaia Data Processing and Analysis Consortium (DPAC) technical note GAIA-LL-060.

[7] Murtagh, F. 1983, The Computer Journal, 26(4), pp.354-359.