A&A 607, A105 (2017)
DOI: 10.1051/0004-6361/201730965
© ESO 2017

Astronomy & Astrophysics
Special issue

Gaia Data Release 1

Gaia Data Release 1

Cross-match with external catalogues. Algorithm and results

P. M. Marrese[1,2], S. Marinoni[1,2], M. Fabrizio[1,2], and G. Giuffrida[1,2]

[1] Space Science Data Center, ASI, via del Politecnico SNC, 00133 Roma, Italy
e-mail: paola.marrese@ssdc.asi.it
[2] INAF–Osservatorio Astronomico di Roma, via Frascati 33, 00040 Monte Porzio Catone (Roma), Italy

Received 10 April 2017 / Accepted 17 October 2017

ABSTRACT

Context. Although the *Gaia* catalogue on its own will be a very powerful tool, it is the combination of this highly accurate archive with other archives that will truly open up amazing possibilities for astronomical research. The advanced interoperation of archives is based on cross-matching, leaving the user with the feeling of working with one single data archive. The data retrieval should work not only across data archives, but also across wavelength domains. The first step for seamless data access is the computation of the cross-match between *Gaia* and external surveys.
Aims. The matching of astronomical catalogues is a complex and challenging problem both scientifically and technologically (especially when matching large surveys like *Gaia*). We describe the cross-match algorithm used to pre-compute the match of *Gaia* Data Release 1 (DR1) with a selected list of large publicly available optical and IR surveys.
Methods. The overall principles of the adopted cross-match algorithm are outlined. Details are given on the developed algorithm, including the methods used to account for position errors, proper motions, and environment; to define the neighbours; and to define the figure of merit used to select the most probable counterpart.
Results. Statistics on the results are also given. The results of the cross-match are part of the official *Gaia* DR1 catalogue.

Key words. astronomical databases: miscellaneous – catalogs – surveys – astrometry – proper motions

1. Introduction

The *Gaia* satellite will allow the positions, parallaxes, and proper motions to be determined with high accuracy for more than 1 billion sources reaching magnitude $G \sim 20.7$. All *Gaia* sources will also have multicolour photometry, while radial velocities will only be available for sources brighter than $G \sim 17$. The summary of the astrometric, photometric, and survey properties of *Gaia* Data Release 1 (DR1) are described in Gaia Collaboration (2016a), while the scientific goals of the mission are summarised in Gaia Collaboration (2016b). Combining the *Gaia* catalogue with other publicly available surveys obtained either from ground or from space more closely meets the modern astronomical research requirements. The main aim of adding a pre-computed cross-match to the official *Gaia* DR1 data is to complement *Gaia* with existing (and widely used by the scientific community) photometry and astrometry, thus allowing the full scientific exploitation of *Gaia*. The complexity and scientific issues related to cross-matching has become very popular now that the combined use of large data sets from different surveys and/or wavelength domains is more and more common. Arenou et al. (2017) shows how the comparison with external catalogues allows a deeper understanding of many of the parameters describing the performances of the *Gaia* catalogue. The results of the cross-match described here played an important role in the full sky tests utilised for the validation of *Gaia* DR1 data, constituting the first scientific exploitation of the cross-match results described in this paper.

In the following, detailed explanations of the general principles we followed and of the reasons behind our choices on each and all the scientific issues of the adopted cross-match algorithm are given. A detailed pinpointing of the caveats and of the failed cases is also available, allowing the scientists who use the cross-match results to be fully aware both of its quality and of its possible limitations.

The cross-matching (hereafter XM) of astronomical catalogues is a complex and challenging problem both scientifically and technologically, especially when matching large surveys which include several millions or billions of sources. In this paper we concentrate on the scientific issues, thus only a short description of the technological and computational implementation is given.

There are different approaches to the XM of astronomical catalogues, and XM algorithms can also be very different. It is important to correctly define both the scientific problem one is faced with and the objectives of the cross-match.

When a neighbour in the secondary catalogue is found close to a leading catalogue source, the first question to be answered is whether it should always be considered as the actual counterpart or not. In the second case the algorithm gets more complicated and some kind of a priori knowledge on the nature of the object that is being matched becomes necessary. The more the two catalogues being matched are different, the more caution should be used in considering the neighbours as counterparts. For example, one may want to match a large general purpose survey with

a survey of a particular class of objects, or the two catalogues could be largely not homogeneous because they were observed in different wavelength domains.

Depending on the scientific problem, an XM algorithm could require a one-to-one match or allow for a one-to-many or many-to-one matches. This will affect the possibility of using a symmetric algorithm.

An XM algorithm is always a trade-off between multiple requisites and a fraction of mismatched objects is always present. The scope of a given XM algorithm could be to minimise the absolute number of mismatches, or to minimise the number of mismatches among the rarest and most peculiar objects. For example, when matching general purpose surveys, it is important whether to use the magnitudes in the selection of the best match. The use of magnitudes and colours requires transformations between photometric systems that are usually based on synthetic photometry of normal stars. While using the magnitudes would help matching most of the objects in a given catalogue, it would probably worsen the matching of many relatively rare but very interesting objects such as variables, peculiar stars, and non-stellar objects. In addition, in many surveys not all objects have a colour (i.e. a fraction of objects may have been detected in one band only). One could aim for a simple algorithm which could be easily applied to many different catalogue pairs or, on the contrary, one could try to tailor the best possible algorithm for a given scientific problem. In some cases it would be more important for the same algorithm to be homogeneously applied to all the objects, in others a different definition of the figure of merit could be allowed (for example, in cases when some a priori information is available for only a fraction of the objects in a given catalogue).

The scientific details of a XM algorithm, in particular which object characteristics (available in the catalogues) to use in the definition of the best match, depend on the pair of catalogues being matched. The characteristics of each catalogue and how the two catalogues compare considering those characteristics are both important. A non-exhaustive list of the information that could be used in the definition of an XM algorithm includes the following: a) data available (positions, epochs, proper motions, parallaxes, photometry, binary, and/or variability characterisation); b) statistics on accuracy and precision of the data available; c) photometric depth (magnitude limit) and completeness; d) possible systematic errors on any of the data (astrometry and/or photometry) used in the cross-match; e) statistics on the availability of the information within a catalogue (for example how many objects have colour information); f) accuracy of photometric transformations between the two catalogues and their applicability limits; and g) angular resolution of each catalogue and the resolution difference between the two catalogues.

In Sect. 2 we outline the general principles that guided us in the definition of the XM algorithm, while in Sect. 3 we describe the details. In Sect. 4 a brief technical description of the XM implementation is given. Section 5 is dedicated to the description of the external surveys matched with *Gaia*. In Sect. 6 we discuss in general terms the XM computation results. Finally in Appendix A a validation test of the method applied to sources with unknown proper motion is discussed.

2. *Gaia* pre-computed cross-match: general principles

Among all the different approaches to cross-matching, we decided to define our algorithm according to the specific scientific problem we have.

The external catalogues to be matched with *Gaia* are all obtained in the optical/near-IR wavelength region (with the exception of allWISE, which extends in the medium-IR domain), are general surveys not restricted to a specific class of objects and have an angular resolution lower than *Gaia*. As such they are sufficiently homogeneous among themselves to allow the use of a single XM algorithm, which is adapted to each different catalogue using a small number of configurable parameters. Since the external catalogues are available together with *Gaia* DR1 data and their cross-match is part of the official *Gaia* DR1, consistency and homogeneity in the cross-match computations are an important requirement. We decided to match *Gaia* DR1 with each external survey separately and independently. A different approach, performing a simultaneous multicatalogue and multiwavelength cross-match (Pineau et al. 2017; Salvato et al. 2017), is less appropriate in our case as we concentrated our work on large optical/near-IR surveys.

The algorithm we defined to match *Gaia* data with publicly available astrometric/photometric surveys is not symmetric, and we always use *Gaia* as the leading catalogue. We assume that when a good neighbour[1] is found for a given *Gaia* object, then it is the counterpart. When more than one good neighbour is found, the best neighbour (i.e. the most probable counterpart according to the figure of merit we define, see Sect. 3) is chosen from good neighbours. The higher *Gaia* angular resolution with respect to the external catalogues requires a many-to-one algorithm; this is why the algorithm we used is not symmetric and why more than one *Gaia* object can have the same best neighbour in a given external catalogue. Two or more *Gaia* objects with the same best neighbour are denoted mates. True mates are objects resolved by *Gaia*, and are not resolved by the external survey.

An important requirement of the XM algorithm that we developed is completeness, and we thus defined the position errors to a 5σ level (see Sect. 3.3). In addition, when defining the XM algorithm, we decided to avoid features which would help the match of generic objects (i.e. normal well-behaved stars) at the cost of worsening the match of peculiar classes of objects. Since we computed the cross-match for several different surveys, we also valued consistency and homogeneity. We thus decided to avoid the use of a priori knowledge which in general surveys is not usually available for all (or the vast majority of) the objects or for all the external catalogues. We tried to avoid relying too much on assumptions while still using the scientific information present in the input catalogue data.

The chosen algorithm is positional and thus uses positions, position errors, their correlation if known, and proper motions. We used *Gaia* proper motions only, so the proper motion correction was applied only to the *Tycho-Gaia* Astrometric Solution (TGAS) subsample for *Gaia* DR1 (~2 million objects), while the vast majority of the *Gaia* DR1 stars do not have proper motions. While the figure of merit we used depends strongly on the angular distance between the *Gaia* target and the external catalogue counterpart candidate and on the position errors, it also depends on the local surface density of the external catalogue (environment).

We produced two separate XM outputs: a *BestNeighbour* table which lists each matched *Gaia* object with its best neighbour and a *Neighbourhood* table which includes all good neighbours for each matched *Gaia* object (see Sect. 6 for a detailed output description).

[1] A good neighbour for a given *Gaia* object is a nearby object in the external catalogue whose position is compatible within position errors with the target.

2.1. Use of magnitudes in the cross-match algorithm

When available and depending on the accuracy of the photometric conversion, photometric data can be considered when defining the XM figure of merit. In order to make use of the magnitudes in the evaluation of the best neighbour and in the process of scoring the neighbours, it is necessary to convert the external catalogue magnitudes to the *Gaia G* magnitude. While this is feasible in principle, one should bear in mind that in general the transformations between photometric systems show quite a large scatter, are not suited for peculiar objects, and have different accuracy for different catalogues. In addition colour information is not generally available for all the objects of a given external catalogue (due to different sensitivity in the different bands) and this in turn causes an inhomogeneous treatment of objects within a given catalogue. We thus avoided using the photometric information in the best neighbour selection.

2.2. Use of proper motions in the cross-match algorithm

Any XM algorithm between two astronomical catalogues is based on object positions and their errors. Proper motions should be taken into account when dealing with high proper motion stars or catalogues with very different epochs of observation or in the case of high confusion, high density regions. For the TGAS subsample we moved *Gaia* objects to the individual epoch of the possible matches (i.e. stars in the external catalogue which are within the search radius, see Sect. 3) and we propagated their position errors. This approach requires calculating *Gaia* object positions on the fly, rather than computing them to a median external catalogue epoch beforehand. We decided to discard the possibility of using the external catalogues' proper motions even when they are available as it introduces an inhomogeneity in the XM of the different catalogues. There is an additional problem when using the external catalogues' proper motions: while the catalogues usually contain positions at a reference epoch (J2000.0), the errors on the positions are given at a mean epoch. By definition, the mean epoch is the epoch which minimises the position errors in the proper motion fitting procedure and often the mean epoch is different for Right Ascension and Declination. It is of course possible to propagate errors from the mean to the reference epoch, but this implies an approximation as the coordinates at the mean epoch are not usually available. The algorithm we used to propagate an object position (and position errors) at a different epoch is described in the HIPPARCOS and *Tycho* Catalogues documentation (ESA 1997 first volume, in particular Sects. 1.2 and 1.5[2]). The adopted algorithm is based on a standard model of stellar motion, which assumes that stars move through space with a constant velocity vector. The rigorous treatment of the epoch transformation requires that the variation of all six parameters, α (Right Ascension), δ (Declination), π (parallax), $\mu_{\alpha*}$ (proper motion in $\alpha\cos\delta$), μ_δ (proper motion in δ), and V_R (Radial Velocity), must be considered. In our case we used only positions and proper motions; we did not use the parallax and radial velocity. As stated in ESA 1997 first volume, the simple formula[3] for transforming a celestial position from one epoch to a different one is not a good physical model for the star motion. The difference with respect to a rigorous model may become significant near the celestial poles or when propagating the position over a long time. Including proper motions does not address the astrometric binary problem as it would be necessary to include the binary orbits when available.

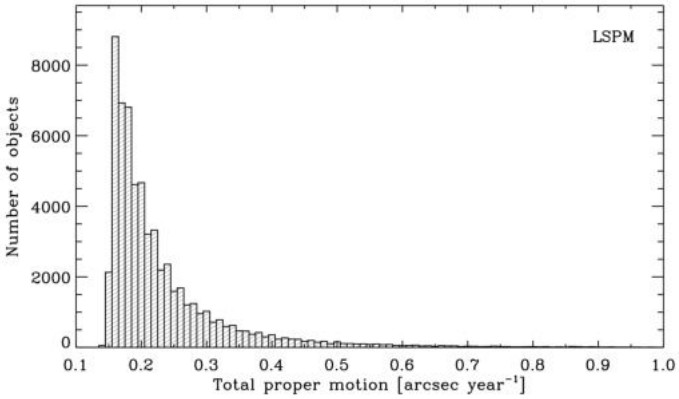


Fig. 1. LSPM proper motion distribution truncated at 1 arcsec yr^{-1}.

2.3. Accounting for epoch differences

The difference in coordinate epochs can be of decades in the most unfortunate cases. In order to account for the difference in coordinate epochs and for the sake of completeness, when *Gaia* proper motions are not available, we decided to increment the *Gaia* position errors in order to take into account not only the coordinates uncertainties, but also possible proper motions.

For high proper motion stars with unknown proper motion in *Gaia* DR1, we aim to obtain the completeness in the *Neighbourhood* output, if not the correct match in the *BestNeighbour* output: we may get the wrong counterpart (if another good neighbour is present), but the correct counterpart will probably be included in the *Neighbourhood* table and can be recovered. The effect of the unknown proper motions is potentially much larger than the position accuracy and depends strongly on the epoch difference between the *Gaia* and the external catalogues sources. We thus consider the effect of the unknown proper motions as a bias rather than a systematic uncertainty. In this context, we consider as *high* proper motion stars those objects whose motion in the sky, combined with the epoch difference between catalogues, can prevent a correct cross-match. The search radius is normally of the order of a few arcsecs to account for the random and systematic errors on positions in the *Gaia* and external catalogues. We thus consider as problematic high proper motion stars those objects that, by combining the proper motion with the epoch difference, can travel a distance comparable with the search radius.

We adopted what we consider to be a reasonable solution: define a proper motion threshold common to all external catalogues. This threshold together with the epoch difference is used to define the initial search radius and to broaden the astrometric errors in the leading (i.e. *Gaia*) catalogue (see Sect. 3). In order to define the proper motion threshold, we need to know the proper motion distribution of high proper motion stars. Figure 1 shows the proper motion distribution (truncated to 1 arcsec yr^{-1}) of the LSPM high proper motion star sample. According to Lépine & Shara (2005), the LSPM Catalog includes 61 977 stars with total proper motion higher than 150 mas yr^{-1} in the northern hemisphere and is complete to 99% for stars at high galactic latitude ($|b| > 15$) and 90% complete for stars at

[2] The original HIPPARCOS tool *pos_prop* implementing the algorithm is also made available (C and FORTRAN) at the following link: http://cdsarc.u-strasbg.fr/viz-bin/Cat?cat=I%2F239%2Fversion_cd%2Fsrc&target=http&

[3] $\alpha = \alpha_0 + (T - T_0)\mu_{\alpha*0}\sec\delta$; $\delta = \delta_0 + (T - T_0)\mu_{\delta 0}$.

low galactic latitude ($|b| \leq 15$) at $V = 19.0$. Most of the stars (~74%) in the LSPM Catalog have a proper motion smaller than 250 mas yr^{-1}, while ~52% of the stars have a proper motion lower than 200 mas yr^{-1}. While ideally the proper motion threshold should be the maximum known proper motion of a real object, we fixed the proper motion threshold at 200 mas yr^{-1} as a compromise between recovering large proper motion stars on the one hand and avoiding adding too many good neighbours and/or mismatches and preserving the performances of the XM calculations on the other.

Fixing a threshold for proper motions and using it to define the initial search radius (see Sect. 3) also influences the cross-match of high proper motion stars with a measured *Gaia* proper motion because the initial search area is defined around the *Gaia* coordinates of a given object before applying the proper motion correction which depends on the epoch of the counterpart candidate. Fortunately the high proper motion stars with a measured *Gaia* proper motion are a tiny fraction of the total: there are 6603 sources in TGAS with a total proper motion higher than 200 mas yr^{-1}. For GSC 2.3 we were able to match 6366 sources of the high proper motion subsample, while we recover 6182 of those sources in 2MASS XM output. This problem will be solved in the XM algorithm planned for *Gaia* DR2, and in the future high proper motion stars will be properly included in the XM output.

2.4. Environment

Given that the cross-match is both a source-to-source and a local problem, the definition of the best neighbour and the neighbour scoring should take into account the surroundings of the *Gaia* sources in the external catalogue. Therefore, the scientific information regarding the probability that a neighbour is a good neighbour resides in the angular distance, but also in the local surface density of the external catalogues. The local density of the external catalogues is calculated on-the-fly by counting the number of external catalogue sources within a fixed radius circle centred on each *Gaia* source (see Sects. 3 and 5 for details). This choice is not optimal, given the different densities between different catalogues and the large density variations with galactic coordinates within a given catalogue. It is, however, a trade-off between having a precise density (which requires a large number of stars and thus a large radius) and having a more accurate local density. By choosing a small radius (while keeping it much larger than the position errors with the exception of a few unfortunate cases), we obtain density estimates that are less precise, but more accurate. It is certainly true that a large part of the sky contains small numbers of stars in small areas. However, this is exactly why many of the objects in a survey are found in relatively dense areas. While a larger initial radius would be more appropriate for lower density regions, it would worsen the density determination when small-scale density variations occur, especially in dense regions like the Galactic Plane. Our choice is motivated by the fact that, even if we obtain a less precise density estimate in the easiest cases when the density is so low that there are very few candidates, we do obtain both precise and accurate (i.e. very local) densities for the difficult dense fields. Finally, an advantage of the local density calculated on-the-fly is that it is measured around the *Gaia* source position. The density is thus a characteristic of a given *Gaia* source and common to all good neighbours evaluated. As such, it does not affect the best neighbour selection.

3. *Gaia* pre-computed cross-match: details

The algorithm we prepared makes use of a plane-sweep technique which requires the catalogues to be sorted by declination, implies the definition of an active list of objects in the external catalogue for each *Gaia* object, and allows the input data to be read only once, thus making the XM computation faster (Devereux et al. 2005; Power & Devereux 2004; Abel et al. 2004; Devereux et al. 2004). In addition, we used a filter and refine technique: a first filter is defined by a large radius centred on a given *Gaia* object and is used to select neighbours and calculate the local surface density, while a second filter is used to select good neighbours among neighbours. Good neighbours are thus filtered on an object-by-object basis. The selection of the best neighbour among good neighbours is based on a figure of merit ("score" in the *Neighbourhood* output table).

Great circle distances between *Gaia* objects and counterpart candidates were evaluated using the haversine formula. We also made some tests using the special case for a sphere of the Vincenty formula obtaining identical results and very similar performances. A normal distribution for position errors is assumed and the position error ellipses are projected on the tangent plane. Even if the position errors are not truly Gaussian, the probability density function is expected to be peaked toward the mean within the error ellipse, and therefore a Gaussian is a reasonable approximation (see Sect. 3.2 for a discussion).

In the following the subscript G stands for *Gaia* and subscript E stands for the external catalogue.

3.1. Initial search radius (first filter)

The initial search radius is defined around each *Gaia* object as

$$R = \max(R_{\rm density}, R_{\rm epochDiff}), \tag{1}$$

where $R_{\rm density}$ is the radius used to calculate the local surface density (60″) and $R_{\rm epochDiff}$ is the radius needed to include in the XM output the stars with a proper motion up to the chosen threshold (200 mas yr^{-1}). The radius used to account for the epoch difference between catalogues is defined as

$$R_{\rm epochDiff} = H_\gamma \cdot PosErr_{\rm max} + \left(\frac{PM_{\rm ref} \cdot EpochDiff_{\rm max}}{1000}\right), \tag{2}$$

where $H_\gamma = 5$ corresponds to a confidence level $\gamma = 0.9999994267$; $PosErr_{\rm max}$ is the maximum of the combined position error; $PM_{\rm ref}$ is the proper motion threshold; and $EpochDiff_{\rm max}$ is the maximum reference epoch difference between *Gaia* and the external catalogue. The maximum combined position error is defined as

$$\begin{aligned} PosErr_{\rm max} = {} & \max[\max(RAerr_G), \max(DECerr_G)] \\ & + \max[\max(RAerr_E), \max(DECerr_E)], \end{aligned} \tag{3}$$

where $RAerr$ and $DECerr$ are respectively the uncertainties in Right Ascension and Declination. The maximum epoch difference between the two catalogues being matched is defined as

$$\begin{aligned} EpochDiff_{\rm max} = {} & \\ & \max\Big[|\max(refEpoch_G) - \min(refEpoch_E)|, \\ & |\min(refEpoch_G) - \max(refEpoch_E)|\Big] \end{aligned} \tag{4}$$

with R in arcsec, $PosErr_{\rm max}$ in arcsec, $PM_{\rm ref}$ in mas yr^{-1}, and $refEpoch$ in years.

The values of $EpochDiff_{\rm max}$ and $R_{\rm epochDiff}$ for all the external catalogues matched with *Gaia* are listed in Table 2.

3.2. Broadening of position errors

While the *Gaia* position errors for the TGAS subsample are propagated to the candidate counterpart epoch as described in Sect. 2.2, for the majority of *Gaia* sources we decided to account for the unknown proper motion systematic contribution. While how to define systematic uncertainties and estimate their magnitudes is open to debate, according to Sinervo (2003) the technique used should be consistent with how the statistical uncertainties are defined since systematic and statistical uncertainties are then combined when the results are compared with theoretical predictions. As stated in the cited paper, "a common technique for estimating the magnitude of systematic uncertainties is to determine the maximum variation in the measurement, Δ, associated with the given source of systematic uncertainty. Arguments are then made to transform that into a measure that corresponds to a one standard deviation measure that one would associate with a Gaussian statistic, with typical conversions being $\Delta/2$ and $\Delta/\sqrt{12}$, the former being argued as a deliberate overestimate, and the latter being motivated by the assumption that the actual bias arising from the systematic uncertainty could be anywhere within the interval Δ". For the sake of completeness as discussed in Sects. 2, 3.1, and 3.4, we fixed a very high confidence level for the initial search radius and the statistical uncertainties, namely the 2D equivalent of 5σ. We coherently decided to use the same proper motion threshold (200 mas yr^{-1}) used to define the initial search radius and the same high confidence level to define the unknown proper motion contribution. We increased the *Gaia* position errors using the following equations:

$$\begin{aligned}\sigma_{x_{G'}} &= \sigma_{x_G} + SysErr_x = \sigma_{x_G} + PM_{\rm ref} EpochDiff/5\\ \sigma_{y_{G'}} &= \sigma_{y_G} + SysErr_y = \sigma_{y_G} + PM_{\rm ref} EpochDiff/5.\end{aligned} \quad (5)$$

For each of the external catalogues matched with *Gaia*, Table 2 shows the maximum values ($SysErr_{\rm max}$) of the systematic contribution added. The actual size of the contribution varies with the exact epoch difference between a given *Gaia* source and the external catalogue counterpart candidate being evaluated. We recall here that, due to the presence of astrometric systematic uncertainties, position errors are not strictly Gaussian, even if the effect of unknown proper motions is not taken into account. Astrometric systematics are larger for the external catalogues than for *Gaia*. In the case of the external catalogues, they are due to the process of linking the observation to the ICRS reference frame, for example. The number, brightness, and colour distributions of reference stars influence the astrometric solution and introduce systematics both globally and locally. This systematic effect is usually smaller than the effect of proper motions and epoch differences.

3.3. Position error convolution ellipse

For the definition of the convolution ellipse, we followed the approach described in Pineau et al. (2011, see their Sect. 3 and Appendix A for details). The position error ellipses in equatorial coordinates are projected on the tangent plane centred on the *Gaia* object position, with the x-axis in the direction towards the external catalogue counterpart candidate. Position errors are respectively described as 2D Gaussians for *Gaia*[4] and external catalogue objects

$$\begin{aligned}&N_{G'}\left(x, y; \sigma^2_{x_{G'}}, \sigma^2_{y_{G'}}, \rho_{G'}\sigma_{x_{G'}}\sigma_{y_{G'}}\right)\\ &N_E\left(x-d, y; \sigma^2_{x_E}, \sigma^2_{y_E}, \rho_E\sigma_{x_E}\sigma_{y_E}\right),\end{aligned} \quad (6)$$

where d is the angular distance between the *Gaia* object and the external catalogue counterpart candidate. The density of probability that the two sources are at the same location is given by the convolution product of the two distributions

$$\begin{aligned}N_C\left(x, y; \sigma^2_{x_C}, \sigma^2_{y_C}, \rho_C\sigma_{x_C}\sigma_{y_C}\right) = f(x,y) = \frac{1}{2\pi\sigma_{x_C}\sigma_{y_C}\sqrt{1-\rho_C^2}}\\ \times \exp\left[-\frac{1}{2(1-\rho_C^2)}\left(\frac{x^2}{\sigma^2_{x_C}} + \frac{y^2}{\sigma^2_{y_C}} - \frac{2\rho_C xy}{\sigma_{x_C}\sigma_{y_C}}\right)\right],\end{aligned} \quad (7)$$

where

$$\begin{aligned}\sigma^2_{x_C} &= \sigma^2_{x_{G'}} + \sigma^2_{x_E}\\ \sigma^2_{y_C} &= \sigma^2_{y_{G'}} + \sigma^2_{y_E}\\ \rho_C\sigma_{x_C}\sigma_{y_C} &= \rho_{G'}\sigma_{x_{G'}}\sigma_{y_{G'}} + \rho_E\sigma_{x_E}\sigma_{y_E}.\end{aligned}$$

By using the eigendecomposition of the variance-covariance matrix of N_C, defining σ_M and σ_m as the semi-major and semi-minor axis in the eigenvector frame (x_1, y_1), changing to the polar coordinates (r,θ) and integrating over θ, the density of probability can be written as a Rayleigh distribution:

$$f(r) = r\ \exp\left(-\frac{1}{2}r^2\right) \quad (8)$$

where $r = \sqrt{\frac{x_1^2}{\sigma_M^2} + \frac{y_1^2}{\sigma_m^2}}$.

3.4. Good neighbours' selection (second filter)

The probability density function $f(x, y)$ defined in Eq. (7) depends on (x, y) only through the exponent component

$$\frac{1}{1-\rho_C^2}\left(\frac{x^2}{\sigma^2_{x_C}} + \frac{y^2}{\sigma^2_{y_C}} - \frac{2\rho_C xy}{\sigma_{x_C}\sigma_{y_C}}\right) = K_\gamma^2, \quad (9)$$

or equivalently

$$\begin{pmatrix}x\\ y\end{pmatrix}^t \Sigma^{-1}\begin{pmatrix}x\\ y\end{pmatrix} = K_\gamma^2, \quad (10)$$

where Σ is the covariance matrix (see also Pineau et al. 2011) and K_γ is known as the Mahalanobis distance. Equation (9) defines the lines of constant probability density; they are ellipses which define confidence regions (2D equivalent of confidence intervals), where K_γ^2 has a χ^2 distribution with 2 degrees of freedom. If we define K_γ equal to a critical value of the χ^2 distribution, the probability that (x, y) will fall within the ellipse is equal to the confidence level γ

$$P\left\{\frac{1}{1-\rho_C^2}\left(\frac{x^2}{\sigma^2_{x_C}} + \frac{y^2}{\sigma^2_{y_C}} - \frac{2\rho_C xy}{\sigma_{x_C}\sigma_{y_C}}\right) \le \chi^2_{2,\alpha}\right\} = 1-\alpha = \gamma, \quad (11)$$

where α is the probability that (x, y) will fall outside the ellipse.

[4] In Eq. (6), $\sigma_{x_{G'}}$ and $\sigma_{y_{G'}}$ stand for the broadened Gaia position errors defined in the previous section, with the exception of the TGAS subsample for which the errors were instead properly propagated using the known proper motion.

Good neighbours are defined as neighbours that fall within the ellipse defined by the confidence level γ:

$$\frac{1}{1-\rho_C^2}\left(\frac{x^2}{\sigma_{x_C}^2}+\frac{y^2}{\sigma_{y_C}^2}-\frac{2\rho_C xy}{\sigma_{x_C}\sigma_{y_C}}\right)\leq K_\gamma^2. \tag{12}$$

Considering that external catalogue sources have coordinates $x = d$ and $y = 0$, the above equation becomes

$$\frac{d}{\sigma_{x_C}\sqrt{1-\rho_C^2}}\leq K_\gamma, \tag{13}$$

where d is the angular distance, σ_{x_C} is the convolution ellipse error in the direction from *Gaia* object to the possible counterpart, ρ_C is the correlation between σ_{x_C} and σ_{y_C}, and K_γ depends on the confidence level γ:

- if $\gamma = 0.9973002038$, $K_\gamma = \sqrt{11.8290}$;
- if $\gamma = 0.9999366575$, $K_\gamma = \sqrt{19.3448}$;
- if $\gamma = 0.9999994267$, $K_\gamma = \sqrt{27.6310}$.

The adopted value of K_γ corresponds to a value of the confidence level γ of 0.9999994267, which in 1D is equivalent to 5σ. The high confidence level was chosen in order to improve the completeness of the cross-match. It should be noted that the *Neighbourhood* output table will contain only the good neighbours, which are not all the neighbours within a fixed radius, but all neighbours which are compatible within errors with the considered *Gaia* source.

3.5. Best neighbour selection: figure of merit

The definition of the figure of merit is inspired by de Ruiter et al. (1977), Wolstencroft et al. (1986), Sutherland & Saunders (1992), and Pineau et al. (2011). However, contrary to all of the above mentioned authors and consistently with the discussion in Sects. 1 and 2.1, we did not add any a priori knowledge on the counterpart candidate's magnitude either as the number of possible counterparts in the magnitude bin of the candidate being considered or the number of possible counterparts brighter than the candidate being considered. In the specific scientific case addressed in this paper, we have no expectations on the brightness of the correct match.

The figure of merit (FoM) we used evaluates the ratio between two opposite models/hypotheses, the counterpart candidate (i.e. the good neighbour) is a match or it is found by chance. The FoM depends on the angular distance and the position errors (both used in the definition of the dimensionless variable r), on the epoch difference, and on the local surface density of the external catalogue.

For each of the good neighbours, we compute the following FoM[5]:

$$FoM(r) = \frac{\mathrm{d}p(r|cp)}{\mathrm{d}p(r|spur)}\cdot \tag{14}$$

In Eq. (14), $\mathrm{d}p(r|cp)$ is the probability of finding a counterpart at a distance between r and $r + \mathrm{d}r$:

$$\mathrm{d}p(r|cp) = r\ \exp\left(-\frac{1}{2}r^2\right)\mathrm{d}r. \tag{15}$$

[5] The defined figure of merit is not a likelihood ratio because it is the ratio between two probabilities rather than between two likelihoods.

The probability of finding a spurious association is instead evaluated using the Poisson distribution, which is based on the assumption that celestial objects are locally randomly distributed and does not take into account the clustering of celestial objects. The Poisson probability of finding one or more objects by chance in an infinitesimal annulus area is

$$\begin{aligned}\mathrm{d}p(r|spur) &= \sum_{k=1}^{\infty}\frac{s^k}{k!}\exp(-s) \ = \ \sum_{k=0}^{\infty}\frac{s^k}{k!}\exp(-s) - Poi(0,s)\\ &= 1-\exp(-s)\approx s\end{aligned} \tag{16}$$

where

$$s = \rho\sigma_M\sigma_m\cdot \mathrm{d}A \approx \rho'\cdot 2\pi r\mathrm{d}r \tag{17}$$

is the number of sources within an infinitesimal annulus area $\mathrm{d}A$, while the factor $\sigma_M\sigma_m$ is needed to convert the measured density ρ into the polar coordinates (as was done in Sect. 3.3). The local surface density ρ is defined by counting the number of objects within the initial search radius. One of the reasons to prefer a more local surface density rather than a more precise one is that in order to increase precision, it is necessary to increase the size of the radius within which the density is evaluated making the assumption of a random distribution of sources less accurate.

The figure of merit is thus

$$FoM(r) = \frac{1}{2\pi\rho'}\exp\left(-\frac{1}{2}r^2\right)\cdot \tag{18}$$

In addition to making use of all the information available and not only the angular distance, the main advantage of the figure of merit with respect to a nearest neighbour is that it allows us to compare the goodness of best neighbours within a given catalogue. In the *Neighbourhood* output table we listed the asinh of the figure of merit defined above:

$$score = \mathrm{asinh}(FoM(r)) = \mathrm{asinh}\left[\frac{1}{2\pi\rho'}\exp\left(-\frac{1}{2}r^2\right)\right]\cdot \tag{19}$$

Since the FoM values cover a large range, applying an asinh make the output numbers more readable. It also has the advantage over the logarithm that it does not have a singularity at zero.

The best neighbour is defined as the good neighbour with the highest score value, while mates are flagged and counted after the best neighbours have been evaluated for all *Gaia* sources.

4. Technological implementation

The cross-match between catalogues which include hundreds of thousand of sources is a technological challenge, mainly because of the large number of angular distances required between pairs of sources in the two catalogues. Performance issues are even more critical when working within the framework of a large collaboration. The XM implementation we developed minimises the number of comparison by selecting a reasonable initial search radius (see Sect. 3.1), whose definition was a trade-off between completeness and performance, and calculating the angular distances for the second catalogue sources which would fall within the initial radius. In addition the *Gaia* catalogue was divided into several declination strips and the XM calculations for the different strips were run in parallel.

In our implementation, the XM calculations are performed in RAM, and the input data for both the first and the second catalogues are read only once: for each *Gaia* source an active list (which is a declination strip in the second catalogue) is defined

Table 1. XM computation performance: computation time (see Sect. 4).

Catalogue	Time (minutes)
UCAC4	39
GSC2.3	239
PPMXL	172
SDSS DR9	56
URAT-1	26
2MASS PSC	69
allWISE	450

and when passing to the following *Gaia* source the active list is updated, but not re-created. The input data are organised in MariaDB 10.1 (a mysql fork) MyISAM tables, since MyISAM is a light MySQL storage engine suitable to fast reading. The output is written in the Percona XtraDB engine for MariaDB, which is an enhanced version of the MySQL InnoDB engine. XtraDB is well suited for concurrent writing. A large effort was put in the detailed configuration of MariaBD (and its engines) in order to improve performance. While the input/output is supported by the MariaDB DBMS, the code is written in C language. All the calculations are performed in RAM by defining dedicated C data structures, which include the active lists and two different writing buffers for the *BestNeighbour* and the *Neighbourhood* outputs (see Sect. 6). The C data structures, as well as the number of *Gaia* strips and the size of the external catalogues active lists, have been optimised for performance on the server we used (256GB RAM, two processors with 8 cores at 2.0 GHz with hyper-threading for a total of 32 CPUs and two 1.2TB SAS disks at 10 K rpm). While the RAM was not an issue, the optimisation was a compromise between CPU usage and I/O limitations (actually the writing rather than the reading was a bottleneck). The performance depends on the characteristics of the external catalogues, mainly the stellar density. The best output performance was 200.000 inserts/s. Table 1 includes the time needed to perform the XM computations for the different external catalogues. It should be noted that the reported times do not include the time needed to ingest and prepare the catalogues nor the time needed to run consistency tests on the results.

5. External catalogue characteristics

The following is the list of external catalogues cross-matched with *Gaia* DR1 catalogue:

- UCAC4 (Zacharias et al. 2013);
- GSC 2.3 (Lasker et al. 2008);
- PPMXL (Röser et al. 2008; Roeser et al. 2010);
- SDSS DR9 primary objects (Ahn et al. 2012; Alam et al. 2015);
- URAT-1 (Zacharias et al. 2015);
- 2MASS PSC (Skrutskie et al. 2006);
- allWISE (Wright et al. 2010; Cutri & et al. 2013).

The main properties to consider when matching the external catalogues with *Gaia* are a) angular resolution; b) astrometric accuracy; c) how the catalogue is tied to the International Celestial Reference System (ICRS); d) coordinates epoch; e) the need to propagate astrometric errors when the catalogue proper motions are available[6]; and f) known issues and caveats. It is also important to take into account how the external catalogue properties compare to *Gaia* catalogue properties.

Table 2 lists the *Gaia* and external catalogue properties relevant to the cross-match when they are available. Figure 2 shows the sky coverage and the distribution of the surface density for *Gaia* and the external catalogues. The surface density was calculated by counting the number of sources in each pixel obtained using a Hierarchical Equal Area and isoLatitude Pixelization (HEALPix; Górski et al. 2005) tessellation with resolution $N_{\rm side} = 2^8$ which has 786 432 pixels with a constant area of $\Omega \sim 188.89$ arcmin^2.

The external catalogue quantities used by our XM calculations are positions, position errors, position error correlation, and epochs. Different surveys may have different definitions for some of these quantities and/or use different units. For example, UCAC4 uses the south pole distance instead of declination, some surveys report the epoch in Julian years and some in MJD, and position errors are defined in different ways. The external catalogue input quantities were thus homogenised in order to simplify the XM calculations.

In the following we list some caveats and known issues that are relevant when computing the cross-match. As stated in Zacharias et al. (2013), in UCAC4 if the computed position error (at mean epoch) of a star exceeds 255 mas, it is set to 255 mas. Similarly, according to the authors, the error in proper motion was truncated to 50 mas yr^{-1}, but respective stars were kept in UCAC4, if at least two observations from different CCD observations were matched or the star is either in the 2MASS, SPM, or NPM data files. Obviously, all large error objects need to be handled with caution, and some of these are simply non-existent. Since the publication of UCAC4 in August 2012, the authors[7] have suggested that the following corrections should be applied: identification of "streak objects" and removal and data correction of a small number of high proper motion stars.

There are some 3 350 256 objects in GSC 2.3 with RA and Dec errors equal to 0, while it is mandatory to have errors on coordinates in order to run the cross-match. We decided not to exclude these objects and to assign to them the largest position error found in the catalogue.

There is a small number of objects (23 945) in SDSS DR9 with large position error (greater than 10 arcsec and up to ~14.36 degrees either in RA or Dec). We decided to filter out these objects.

6. Results

The cross-match results are part of the official *Gaia* DR1 release and are available at the ESAC *Gaia* Archive[8] and at the SSDC *Gaia* Portal[9].

The XM output consists of two separate tables: *BestNeighbour* includes the best matches (selected as the good neighbour with the highest value of the score), while *Neighbourhood* includes all the good neighbours (selected using the second filter, see Eq. (13)). The XM output contents are described in Tables 3 and 4, respectively. The BestNeighbourMultiplicity field in the *BestNeighbour* output table addresses the binary stars and/or duplicates problem present in GSC 2.3, PPMXL, and 2MASS PSC. In these catalogues there is a fraction of source pairs with the same coordinates and position errors. Given that the astrometric

[6] Positions are given at epoch J2000.0, but errors on positions are given at mean epoch.

[7] http://www.usno.navy.mil/USNO/astrometry/optical-IR-prod/ucac

[8] https://gea.esac.esa.int/archive/

[9] http://gaiaportal.asdc.asi.it/

Table 2. Properties of *Gaia* and the external catalogues.

Catalogue	N Sources	$PosErr_{\max E}$[a] (arcsec)	Effective resolution (arcsec)	ICRS offset (mas)	$EpochDiff_{\max}$ (yr)	$SysErr_{\max}$[b] (arcsec)	$R_{EpochDiff}$[c] (arcsec)
Gaia DR1	1 142 679 769	0.1	0.1[d]	...	N/A	N/A	N/A
UCAC4	113 728 883	2.279[e]	~2	...	16.88	0.65	15.155
GSC 2.3	945 592 683	1.6	~4-5[f]	280	62.29	2.49	20.958
PPMXL	910 468 710	1.3421[e]	~4–5[f]	300	15.0	0.60	10.211
SDSS DR9	469 029 929	10.0	~1.4[f]	<100	16.29	0.65	53.758
URAT-1	228 276 482	0.429	~5[f]	...	2.689	0.11	3.183
2MASS PSC	470 992 970	1.21	5	15	17.57	0.70	10.064
allWISE	747 634 026	35.944	6.1, 6.4, 6.5,12.0[g]	...	5.0	0.20	181.220

Notes. [(a)] $PosErr_{\max E} = \max[\max(RAerr_E), \max(DECerr_E)]$, see Sect. 3.1. [(b)] $SysErr_{\max} = PM_{\rm ref} EpochDiff_{\max}/5$, see Sect. 3.2. [(c)] $R_{EpochDiff}$ is defined in Sect. 3.1, Eq. (2). [(d)] The effective angular resolution on the sky of *Gaia* DR1, in particular in dense areas, is not yet at this expected level (Arenou et al. 2017). [(e)] The maximum of position error refers to the propagated errors at J2000.0. [(f)] Effective resolution value is our best guess. [(g)] Angular resolution in the four $W1$–$W4$ bands, respectively.

Table 3. *BestNeighbour* output table content.

Field name	Short description
SourceId	*Gaia* source identifier
OriginalExtSourceId	Original external catalogue source identifier
AngularDistance	Haversine angular distance (arcsec)
NumberOfMates	Number of mates in *Gaia* catalogue
NumberOfNeighbours	Number of good neighbours in external catalogue
BestNeighbourMultiplicity	Number of neighbours with same probability as best neighbour
ProperMotionFlag	Use of *Gaia* proper motions (TGAS subsample)

Table 4. *Neighbourhood* output table content.

Field name	Short description
SourceId	*Gaia* source identifier
OriginalExtSourceId	Original external catalogue source identifier
AngularDistance	Haversine angular distance (arcsec)
Score	Figure of merit
ProperMotionFlag	Use of *Gaia* proper motions (TGAS subsample)

properties are the same, the calculated XM score is also identical: the XM algorithm picks one of them, with no possibility to distinguish between the two. Fortunately, these objects are quite rare, as shown in the last column of Table 5 where the number of sources with a bestNeighbourMultiplicity value greater than 1 is listed.

While it is not possible to discuss the correctness of the XM results on an object-by-object basis, it is possible to evaluate whether the general macroscopic results are as expected or if some features are present which could hint to a relevant fraction of mismatches.

Tables 5 and 6 respectively show some statistics of *BestNeighbour* and *Neighbourhood* output tables for the different external catalogues matched with *Gaia* DR1. The maximum value of angular distance and the fraction of matched pairs closer than 0″.5 depend on the astrometric precision and on the epoch difference, so that their correlation with astrometric accuracy and systematics is less obvious. Table 5 also shows that, even if in some cases the number of good neighbours for a given *Gaia* source is large, the vast majority of *Gaia* sources have a single neighbour in the external catalogue. The lower fraction of *Gaia* sources matched with PPMXL and GSC2.3 sources which have only one good neighbour is most probably due to a fraction of duplicated sources in those catalogues at photographic plates borders which are visible in Fig. 2. In addition Table 5 shows that the large majority of *Gaia* objects do not have mates; the *Gaia* DR1 effective angular resolution on the sky is not yet at the expected level (Arenou et al. 2017), which in turn is mainly due to heavy filtering in data reduction. The minimum and maximum score values listed in Table 6 show that even matches with very low values of the figure of merit are kept in the XM output. The selection of good neighbours is based on the criterion defined in Eq. (13), while no lower threshold for the figure of merit was fixed as it would be quite arbitrary. Matched pairs with a low score value should have a relatively large angular distance, large position errors, or be in fields with high stellar density. Table 7, Fig. 3 and the histograms shown in Fig. 6 address the issue of completeness by showing how many *Gaia* objects are matched and how many of the external catalogues sources are matched, and show respectively the sky and magnitude distribution of the matched sources. For example, the total number of sources in UCAC4 is around 10% of the number of sources in *Gaia* DR1, which is consistent with having ~10% of matched *Gaia* sources and ~95% of UCAC4 sources matched. This is also consistent with the flat sky distribution of *Gaia* versus UCAC4 best matches (shown in panels a and b of Fig. 3) and with the fact that matched and unmatched UCAC4 sources are evenly distributed in magnitude. On the contrary, GSC 2.3 and PPMXL reach fainter magnitudes outside the galactic plane but contain fewer objects closer to the galactic plane (probably due to their lower angular resolution compared to *Gaia*) as shown in Fig. 3, panels c, d and e, f, respectively. This is reflected in the Table 7 results: a significant fraction of *Gaia* sources are unmatched, as are a significant fraction of GSC 2.3 and PPMXL sources. The histogram in Fig. 6 shows that the GSC 2.3 and PPMXL sources with no *Gaia* match are mainly faint sources (Rf or $R2 > 19.0$). SDSS DR9 is much deeper in magnitude than *Gaia* and this is clearly visible in the XM results shown in Fig. 3. The histogram in panel c of Fig. 6 shows that the cross-match correctly selects the bright objects even if magnitudes are not directly used in the figure of merit definition. In the case of the *Gaia* versus allWISE cross-match, the results show that, as expected, the two surveys see quite a different sky.

The comparison of the surface density distribution in the sky of all sources (see Fig. 2) and of the matched sources (see Fig. 3), should help us understand whether some of the properties of the XM output are due to pairing failure or to features

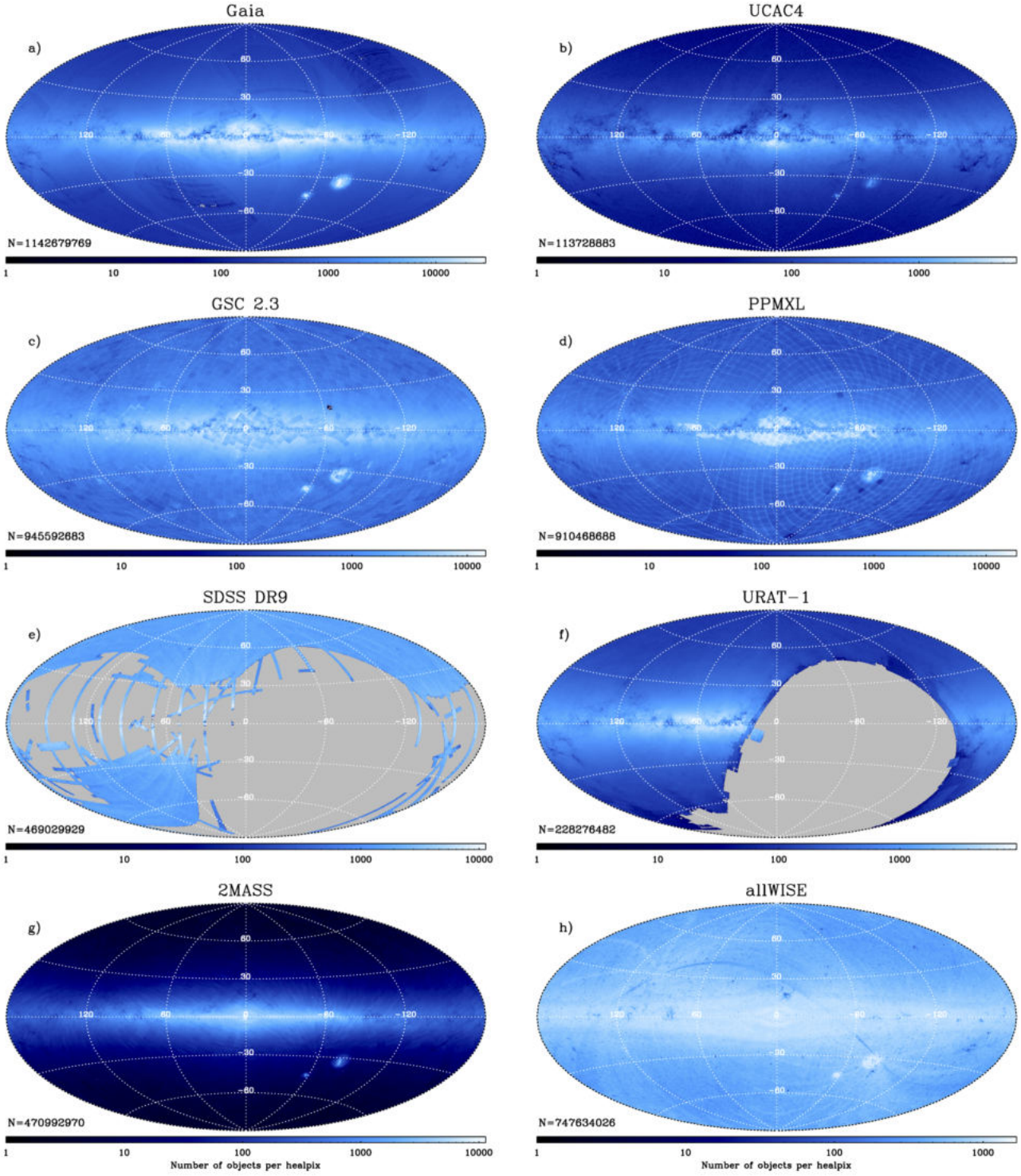


Fig. 2. Surface density distribution for *Gaia* and the external catalogues (see Sect. 5) obtained using a HEALPix (Górski et al. 2005) tessellation with resolution $N_{\rm side} = 2^8$. In grey are indicated the areas not covered by the survey.

already present in the catalogues. For example, the sky area around $(l, b) = (-120°, +40°)$, where the *Gaia* coverage is worse than average, can also be clearly distinguished in the matched sources sky distributions. *Gaia* sources with a large number of mates (see Cols. 6 and 7 in Table 5) are not numerous and are usually in very dense fields or are matched with a source with large position errors in the external catalogue. Fig. 4 shows a rather extreme example taken from the XM results for UCAC4. A single UCAC4 source (UCAC4 115-004819) is the best match for 13 different *Gaia* objects. The UCAC4 source is found in the very dense core of NGC 1718 star cluster (RA = $04^{\rm h}52^{\rm m}26\overset{\rm s}{.}44$, Dec = $-67°03'08\overset{''}{.}5$).

Figure 5 shows the angular distance distribution of the *Best-Neighbour* table for the different external catalogues. The top plots of each panel show the results of the published cross-match, while the bottom plots in each panel show the difference between the XM results calculated with and without the position error broadening. It is clear from the comparison between the top

Table 5. *BestNeighbour statistics*: Min/Max values of relevant output fields in *BestNeighbour* tables.

Catalogue	Angular distance (arcsec)	% with $d < 0\farcs5$[a]	Number of neighbours	% with single neighbour	Number of mates	% with no mates	BestNeighbour multiplicity	Sources with $m > 1$[b]
	max		max		max		max	
UCAC4	9.87	86.95	4	99.64	12	85.96	1	0
GSC 2.3	17.86	60.11	18	90.04	18	70.72	16	93 964
PPMXL	6.14	70.89	12	89.26	7	86.85	2	6
SDSS DR9	51.88	98.54	8	98.60	61	99.48	1	0
URAT-1	2.12	99.80	2	99.99	3	99.75	1	0
2MASS	6.82	85.32	5	98.92	6	88.03	2	8
allWISE	181.11	77.60	3	99.99	20	99.15	1	0

Notes. The fraction of *Gaia* matched stars closer than 0.5 arcsec, those without mates and with a single neighbour, and the number of *Gaia* matched sources with no multiplicity are also listed. [(a)] d = angular distance. [(b)] m = BestNeighbour multiplicity.

Table 6. *Neighbourhood* statistics: Min/Max values of relevant output fields in *Neighbourhood* tables.

Catalogue	Angular distance (arcsec)	Score	
	max	min	max
UCAC4	9.87	0.000000002625	17.553515364854
GSC2.3	18.54	0.000000000857	14.925308621133
PPMXL	6.15	0.000000002021	15.234638147405
SDSS DR9	52.40	0.000000000047	12.080360427928
URAT-1	2.12	0.000000044589	18.779691208067
2MASS	6.82	0.000000004135	14.712484824539
allWISE	181.11	0.000000000514	12.744128884004

Table 7. External catalogue XM results: the number of objects compared with the number of matched sources, the fraction of matched *Gaia* sources and the fraction of matched external catalogue sources.

Catalogue	Number of sources	Number of best matches[a]	% of *Gaia* sources matched[a]	% of external cat sources matched[a]	Number of neighbours
UCAC4	113 728 883	117 369 911	10.3	95.4	117 797 078
GSC2.3	945 592 683	844 343 562	73.9	74.8	937 463 454
PPMXL	910 468 710	714 367 484	62.5	73.1	801 968 492
SDSS DR9	469 029 929	97 018 148	8.5[b]	20.6	98 411 313
URAT-1	228 276 482	209 888 464	18.4[b]	91.8	209 888 621
2MASS	470 992 970	447 946 619	39.2	89.3	452 852 361
allWISE	747 634 026	311 033 691	27.2	41.4	311 035 922

Notes. The number of sources in the *Neighbourhood* tables is also listed. [(a)] "Number of best matches" includes the *mates*. This column and "% of *Gaia* sources matched" indicate distinct *Gaia* sources. "% of external cat sources matched" indicates the fraction of distinct external catalogue sources matched. [(b)] The percentage of matched *Gaia* sources in this case does not take into account the external catalogue limited sky coverage (see Fig. 2).

and bottom plots in each panel of Fig. 5 that without the position error broadening a large fraction of the matches are lost. On average, between ~12% and ~20% of the sources matched including the broadening are lost when original errors are used and epoch differences are ignored. The only exception is SDSS DR9 where a larger fraction of matches (~65%) are lost. In Appendix A a validation test of the position error broadening approach using the TGAS subsample is described for 2MASS PSC, UCAC4, and GSC 2.3. The test shows that the effect of unknown proper motions and epoch differences is not negligible and that broadening the position errors leads to much more accurate and complete results than those obtained when ignoring this effect.

It should also be noted that the algorithm we used is not able to distinguish between *true* mates, which are objects resolved by *Gaia* that are not resolved in the external survey, and *false*/*casual* mates. *False* mates are a pair (or small group) of *Gaia* objects with the same best neighbour in the external catalogue found at an angular distance which is larger than the effective resolution of the external catalogue. The discrimination between *true* and *false* mates thus depends on the angular resolution which is obviously much larger than position errors. While *false* mates are present because we chose a high confidence level (5σ) and we incremented the position errors in order to account for unknown proper motions, it is not correct to consider *false* all mates found

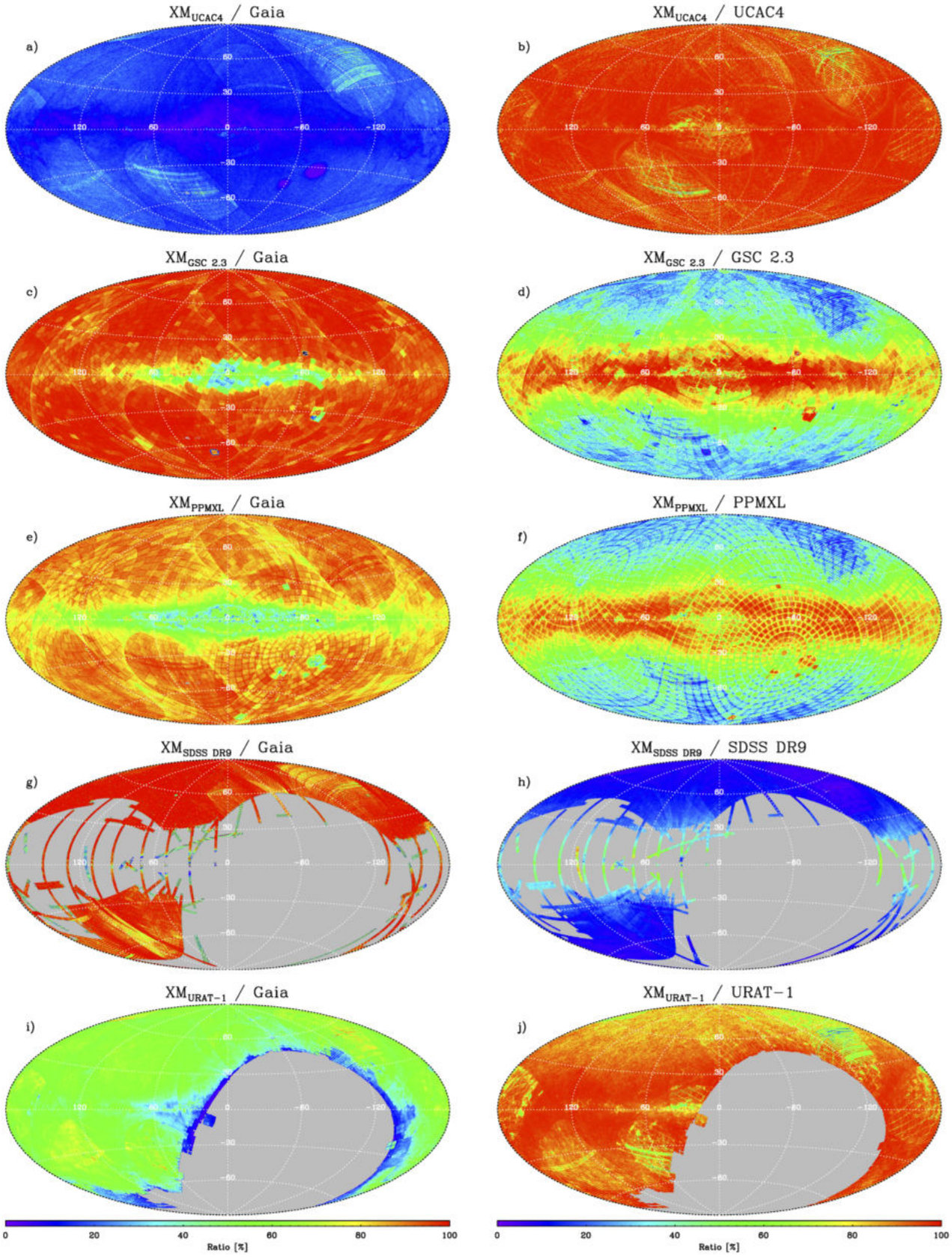


Fig. 3. Surface density map for matched sources obtained using a HEALPix (Górski et al. 2005) tessellation with resolution $N_{\rm side} = 2^8$. The *left column* figures show the fraction of *Gaia* sources matched with an external catalogue, while the *right column* figures show the fraction of distinct external catalogue sources matched with *Gaia*.

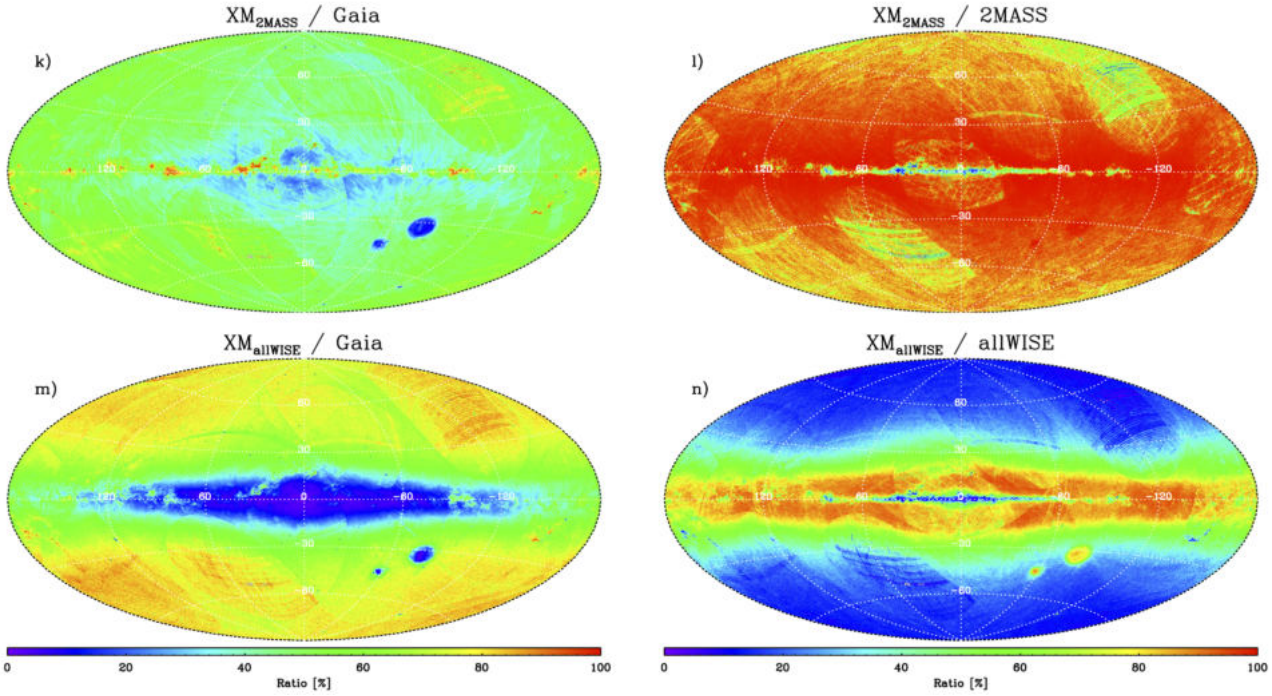


Fig. 3. continued.

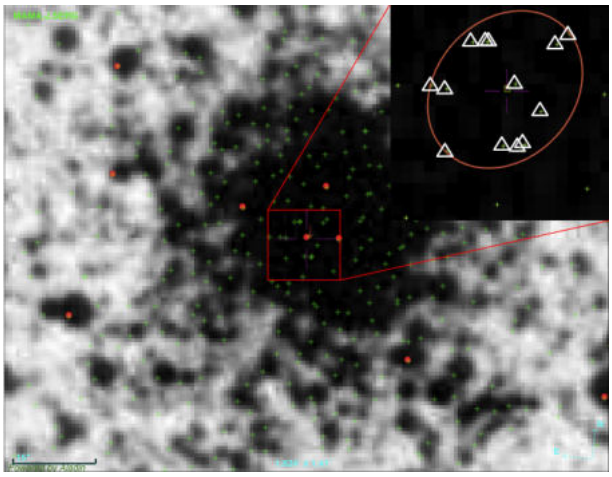


Fig. 4. Dense core of the NGC 1718 star cluster. Shown is the extreme case of UCAC4 115-004819, which is the best match for 13 different *Gaia* objects. UCAC4 objects are indicated by filled red dots, while *Gaia* sources by green crosses. This figure was obtained using the CDS Aladin tool (Bonnarel et al. 2000; Boch & Fernique 2014).

at distances which are large compared with the position errors, but not compared to the angular resolution. Matches which involve mates should be handled with particular care and a decision should be made about their being *true* or *false* on an object-by-object basis. In the case of *true* mates, when combining *Gaia* and an external catalogue photometry (for example in a colour magnitude diagram) the mates' fluxes should be added first.

The above analysis demonstrates that the XM results follow realistic expectations given the accuracy, precision, and diversity of input data sets, thus supporting the effectiveness of the adopted XM algorithm.

7. Conclusions

We developed a cross-match innovative in many respects and applied it in a consistent manner to a completely new survey of unprecedented astrometric accuracy such as *Gaia* DR1. The cross-match described in this paper is a large-scale XM which, quite uniquely, accounts for epoch differences and proper motions on an object-by-object basis. It uses an advanced algorithm based on a standard model of stellar motion when *Gaia* proper motions are available, and instead adds a systematic contribution (which depends on a proper motion threshold and on the epoch difference) to the position errors when *Gaia* proper motions are not available. In addition, the position errors are propagated to epoch J2000.0 for the surveys which list the coordinates at epoch J2000.0, but list position errors at a mean epoch (i.e. PPMXL and UCAC4). The adopted algorithm is also quite unique in that its definition of a many-to-one best match and the *mate* concept is new and accounts for the *Gaia* high angular resolution. The definition of the output itself is original and much different from what is generally done. We tried to supply scientists, in the output tables and in the analysis performed in this paper, with all the means to check the XM results and to understand whether this cross-match is appropriate for their scientific needs. They also have the completely new possibility of overriding the best match choice we made by using the *Neighbourhood* table and all the relevant quantities included there. For example, the angular distance and the figure of merit values could be complemented with a priori knowledge of counterpart magnitudes if a given scientific case benefits from it.

Since the XM algorithm described in this paper was developed for one very specific scientific case (i.e. matching *Gaia* data with large optical/IR surveys with an angular resolution lower than *Gaia's*), it is not appropriate for matching *Gaia* with the following:

a) Catalogues with a comparable angular resolution (HST data for example). In this case the mates should not be present and

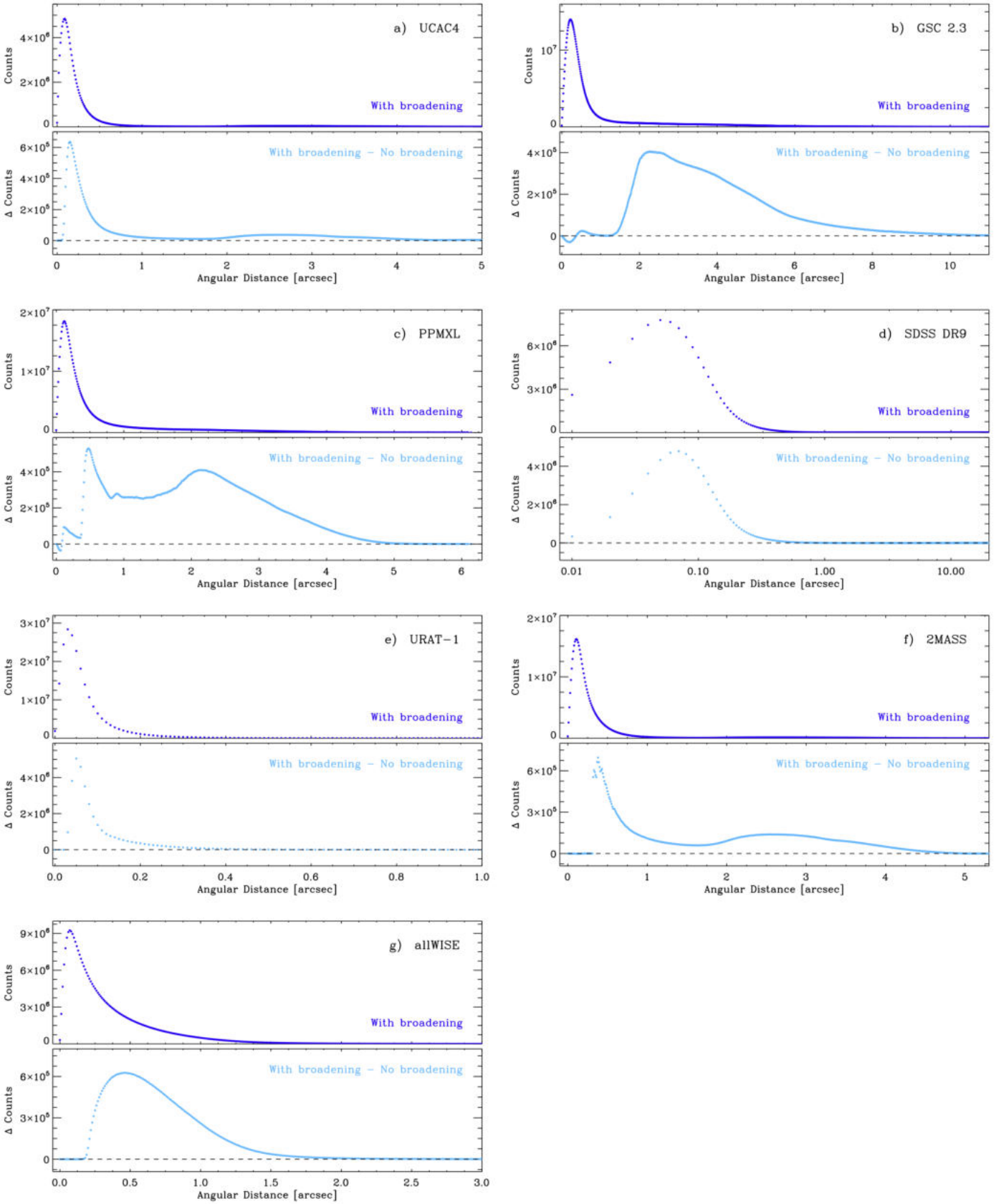


Fig. 5. Angular distance distribution of the matched pairs in the *BestNeighbour* table for the different external catalogues. For each catalogue in the *top plot* of each panel the results are shown for the algorithm used for *Gaia* DR1 cross-match (blue). The *bottom plots* of each panel show instead the difference between the XM results calculated with and without the position error broadening (light blue).

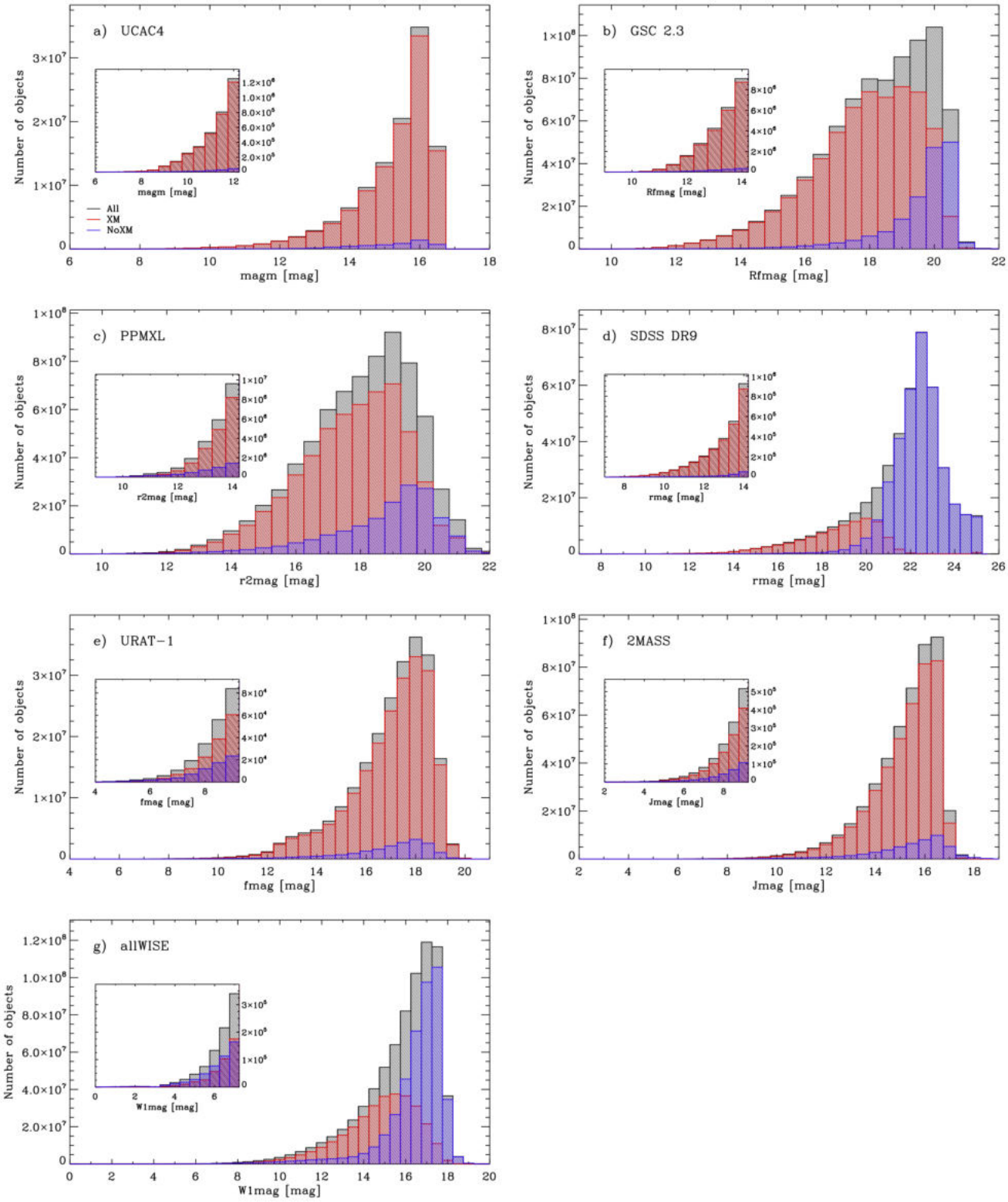


Fig. 6. Magnitude distribution, in the most populated band, for the sources in the external catalogues. In grey the catalogue distribution, in red the matched sources distribution, in blue the unmatched sources distribution.

a single *Gaia* counterpart should be chosen for each external catalogue source.

b) Catalogues obtained in wavelength regions different from optical/near-IR. In these cases the position accuracy and the density are very different from *Gaia*'s. It is probably better to use the external catalogue as the leading catalogue and there are probably good reasons to use magnitude/colours or other a priori information in the best match choice.

c) Lists of sparse objects. In these cases the completeness, the density, and the angular resolution of the sparse list are quite undefined. The best match is probably only one and should be chosen from the mates. The best match for a given *Gaia*

source could well be a close source which is not included in the list.

d) Catalogues of specific/peculiar objects rather than generic surveys. In these cases a priori information on the specific objects should be included in the best match selection criteria.

Acknowledgements. We would like to acknowledge the financial support of INAF (Istituto Nazionale di Astrofisica), Osservatorio Astronomico di Roma, ASI (Agenzia Spaziale Italiana) under contract to INAF: ASI 2014-049-R.0 dedicated to ASDC. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This research has made use of the Aladin Sky Atlas developed at CDS, Strasbourg Observatory, France. This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France. We would like to thank R. Smart, R.A. Power, D. Bastieri, G. Fanari, and G. Altavilla for very useful discussions, suggestions, help, and support.

References

Abel, D. J., Devereux, D., Power, R. A., & Lamb, P. R. 2004, An O(N log M) Algorithm for Catalogue Crossmatching, Tech. Rep. TR-04-1846, CSIRO ICT Centre
Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2012, ApJS, 203, 21
Alam, S., Albareti, F. D., Allende Prieto, C., et al. 2015, ApJS, 219, 12
Arenou, F., Luri, X., Babusiaux, C., et al. 2017, A&A, 599, A50
Boch, T., & Fernique, P. 2014, in Astronomical Data Analysis Software and Systems XXIII, eds. N. Manset, & P. Forshay, ASP Conf. Ser., 485, 277
Bonnarel, F., Fernique, P., Bienaymé, O., et al. 2000, A&AS, 143, 33
Cutri, R. M., et al. 2013, VizieR Online Data Catalog, II/328
de Ruiter, H. R., Willis, A. G., & Arp, H. C. 1977, A&AS, 28, 211
Devereux, D., Abel, D. J., Power, R. A., & Lamb, P. R. 2004, Notes on the implementation of Catalogue Cross Matching, Tech. Rep. TR-04-1847, CSIRO ICT Centre
Devereux, D., Abel, D. J., Power, R. A., & Lamb, P. R. 2005, in Astronomical Data Analysis Software and Systems XIV, eds. P. Shopbell, M. Britton, & R. Ebert, ASP Conf. Ser., 347, 346
ESA 1997, The HIPPARCOS and TYCHO catalogues. Astrometric and photometric star catalogues derived from the ESA Hipparcos Space Astrometry Mission, ESA SP, 1200
Gaia Collaboration (Brown, A. G. A., et al.) 2016a, A&A, 595, A2
Gaia Collaboration (Prusti, T., et al.) 2016b, A&A, 595, A1
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Høg, E., Fabricius, C., Makarov, V. V., et al. 2000, A&A, 355, L27
Lasker, B. M., Lattanzi, M. G., McLean, B. J., et al. 2008, AJ, 136, 735
Lépine, S., & Shara, M. M. 2005, AJ, 129, 1483
Myers, J., Sande, C., Miller, A., Warren, J., & Tracewell, D. 2002, SKY2000 Master Catalog, Version 4, Goddard Space Flight Center, Flight Dynamics Division
Pineau, F.-X., Motch, C., Carrera, F., et al. 2011, A&A, 527, A126
Pineau, F.-X., Derriere, S., Motch, C., et al. 2017, A&A, 597, A89
Power, R. A., & Devereux, D. 2004, Benchmarking Catalogue Cross Matching, Tech. Rep. TR-04-1848, CSIRO ICT Centre
Roeser, S., Demleitner, M., & Schilbach, E. 2010, AJ, 139, 2440
Röser, S., Schilbach, E., Schwan, H., et al. 2008, A&A, 488, 401
Salvato, M., Buchner, J., Budavari, T., et al. 2017, MNRAS, submitted [arXiv:1705.10711]
Sinervo, P. 2003, in Statistical Problems in Particle Physics, Astrophysics, and Cosmology, eds. L. Lyons, R. Mount, & R. Reitmeyer, 122
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Sutherland, W., & Saunders, W. 1992, MNRAS, 259, 413
Wolstencroft, R. D., Savage, A., Clowes, R. G., et al. 1986, MNRAS, 223, 279
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Zacharias, N., Finch, C. T., Girard, T. M., et al. 2013, AJ, 145, 44
Zacharias, N., Finch, C., Subasavage, J., et al. 2015, AJ, 150, 101

Appendix A: Validation of position error broadening approach with TGAS

We developed an XM algorithm which uses a new approach: the algorithm accounts for epoch differences and unknown proper motions rather than simply acknowledging the impossibility of taking into consideration their effects. With the aim of comparing the results of these two different approaches, we describe here a simple test using the *Gaia*-TGAS subsample. Three different XM algorithms were compared:

1. properly propagate for proper motion positions and their errors (Sect. 2.2);
2. ignore the proper motions and broaden the *Gaia* position errors (Sects. 2.3 and 3.2);
3. ignore the proper motions without applying any broadening to the *Gaia* position errors.

It is important to note that using TGAS proper motions implies not only moving the sources "closer" to their counterparts in the external catalogue, but also propagating (i.e. broadening) the position errors.

For the three described algorithms, we performed the XM calculations between the full *Gaia* catalogue and 2MASS PSC, UCAC4[10], and GSC 2.3. We then extracted the *Gaia*-TGAS subsample from the corresponding *BestNeighbour* tables.

A comparison of the results for the TGAS subsample is shown in Fig. A.1. The left panels show the Mahalanobis distance distributions in the three cases and compares them to the Rayleigh distribution. The plots show that in the first case (blue curves) the convolution of position errors is close to the theoretical expectations; in the second case (light blue) it is, as expected, definitely overestimated; and in the third case (yellow) it is clearly underestimated.

Assuming that the results of the first algorithm are correct, the fraction of correct matches for the second and third algorithms are defined as the number of sources with the same best neighbour as obtained with the first algorithm. The results (summarised in Table A.1) show that when a TGAS source is matched, all three algorithms produce, in the vast majority of cases, the same best match. However, the fraction of matches and the fraction of correct matches with respect to the first algorithm is always much larger when using the position error broadening (second algorithm) than when not broadening the position errors at all (third algorithm).

The right panels of Fig. A.1 show that the same sources are, as expected, matched at larger angular distances using the second algorithm (light blue), since source positions are not propagated to the external counterpart epoch. When the third algorithm is used (yellow), only closer counterparts are found. In the case of UCAC4 (which has much smaller position errors than 2MASS or GSC 2.3) the third method fails to match a good fraction of the TGAS sources.

The high proper motion stars are not over-represented in TGAS because of the large time interval between HIPPARCOS/*Tycho*2 and *Gaia* observations. The TGAS proper motion distribution is in fact similar to the UCAC4 and PPMXL corresponding distributions. However, given the smaller epoch difference between *Gaia* and URAT-1 or allWISE (see column $EpochDiff_{\max}$ in Table 2), the results of the second and third algorithms are expected to be less different in those cases (see also Fig. 5).

This test shows that for the TGAS subsample the effect of unknown proper motions and epoch differences is not negligible and that broadening the position errors leads to much more accurate and complete results than those obtained ignoring this effect.

We note that the TGAS subsample is not fully representative of the entire *Gaia* catalogue, given that the magnitude distribution is definitely different and bright sources are obviously more common in TGAS. Likewise, the TGAS counterparts in the external catalogues are not fully representative of the entire surveys. There are two main reasons for this: confusion and position error precision, which both depend on magnitude. Faint stars are difficult to detect around bright objects both for *Gaia* and the external catalogues, and thus the confusion around TGAS sources (and their counterparts in the external catalogues) is lower than average.

In addition, since bright (non-saturated) sources have normally more precise positions, this implies they are more difficult to match when astrometric systematics and proper motion effects are not taken into account (the third case described). This probably means that when considering the full catalogues rather than the bright TGAS subsample, the inadequacy of the third algorithm is less severe.

In order to assess to what extent the TGAS subsample is able to infer the validity of the position error broadening approach for the full catalogues, Table A.2 and Fig. A.2 were prepared.

For *Gaia*, 2MASS and UCAC4, Fig. A.2 shows in light blue the position error distribution for the bulk of the sources in each catalogue (normally they are the faint magnitude end). In yellow is shown the distribution of the position errors for TGAS (in the case of *Gaia*) or the TGAS counterparts (in the external catalogues) matched using the proper motions. In blue is shown the same distribution for all the sources in each catalogue in the same magnitude range as the TGAS (for *Gaia*) or TGAS counterparts (for the external catalogues). In Fig. A.2 the dotted red vertical line indicates the position error threshold within which are contained >99% of the TGAS or TGAS counterparts for each catalogue. Table A.2 summarises the fraction of sources with a position accuracy better than the threshold defined above. Both Fig. A.2 and Table A.2 indicate that the TGAS subsample is a fair approximation of the full catalogues in terms of position error distribution.

The GSC 2.3 results, while very good for the TGAS subsample itself, cannot be used to infer the correctness of the position error broadening approach for the full catalogue. Unfortunately, the TGAS subsample counterparts are almost always found (see the hatched region in Fig. A.2) among the bright stars which are saturated in the GSC 2.3 long exposure plates and were thus supplemented with data from *Tycho*-2 (Høg et al. 2000) and SKY2000 (Myers et al. 2002), as reported in Lasker et al. (2008).

The effect of the difference in confusion between a bright sample and a full survey is to reduce the fraction of correct matches; however, our claim that broadening the position errors allows us to recover the matches in the *Neighbourhood* table if not in the *BestNeighbour* table does hold. This is obviously not true when position errors are not broadened.

A more complete validation of the position broadening approach will be possible with *Gaia* DR2 data, when the vast majority of sources will have a published proper motion.

[10] For UCAC4 the position errors are propagated to epoch J2000.0 in all cases.

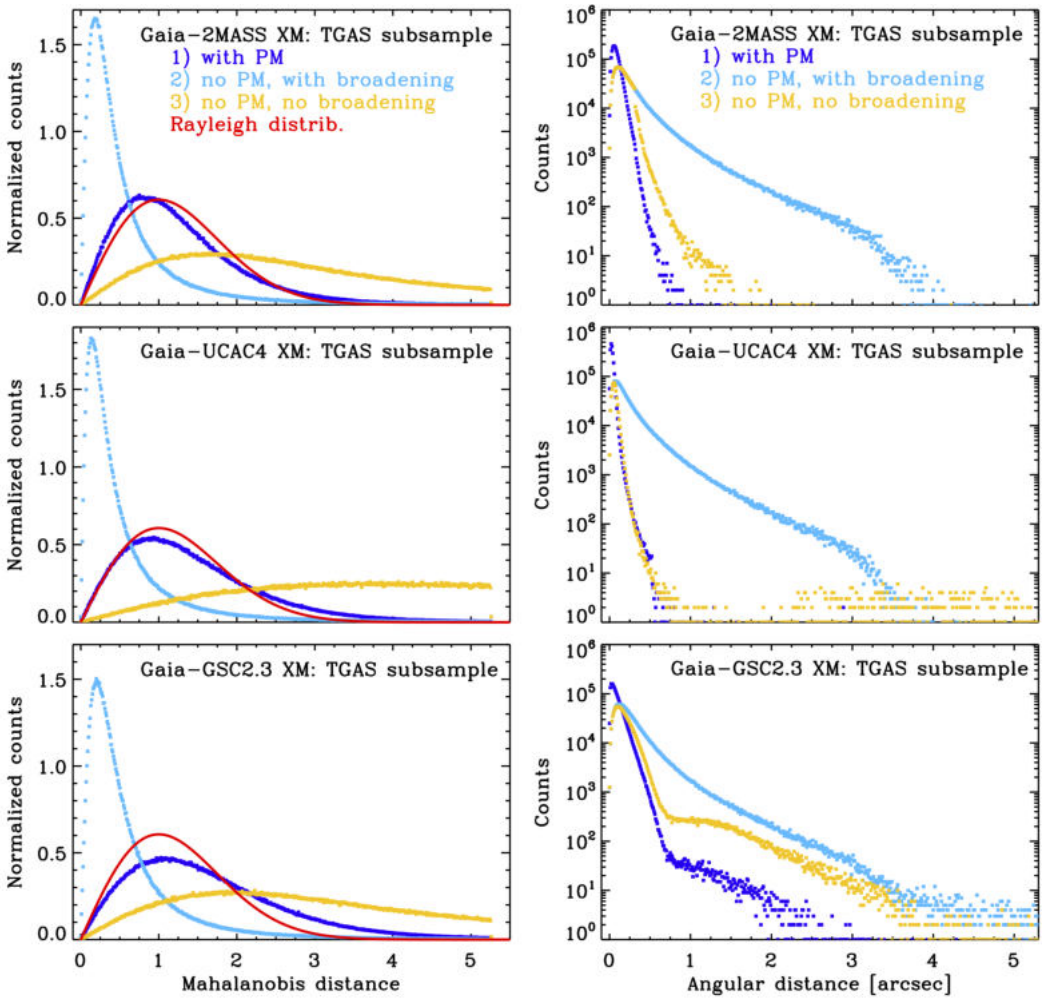


Fig. A.1. Mahalanobis (i.e. normalised) distance and angular distance distributions for three different algorithms (position propagation using proper motions in blue, position error broadening in light blue, and original position errors in yellow) for the *Gaia*-TGAS subsample matched with 2MASS PSC, UCAC4, and GSC 2.3.

Table A.1. Comparison of the results of the three different XM algorithms for TGAS. The number of sources in TGAS is 2 057 050.

Catalogue	N matches	Fraction of matches		Fraction of correct matches	
	algorithm 1	algorithm 2	algorithm 3	algorithm 2	algorithm 3
2MASS PSC	2 045 772	99.88	74.85	99.39	74.76
UCAC4	2 042 349	99.88	28.19	99.20	28.08
GSC 2.3[a]	2 026 532	100.72[b]	63.79	95.68	60.36

Notes. [(a)] See text for a discussion about GSC 2.3. [(b)] In the case of GSC 2.3 we find a few more matches with broadening than with proper motion propagation.

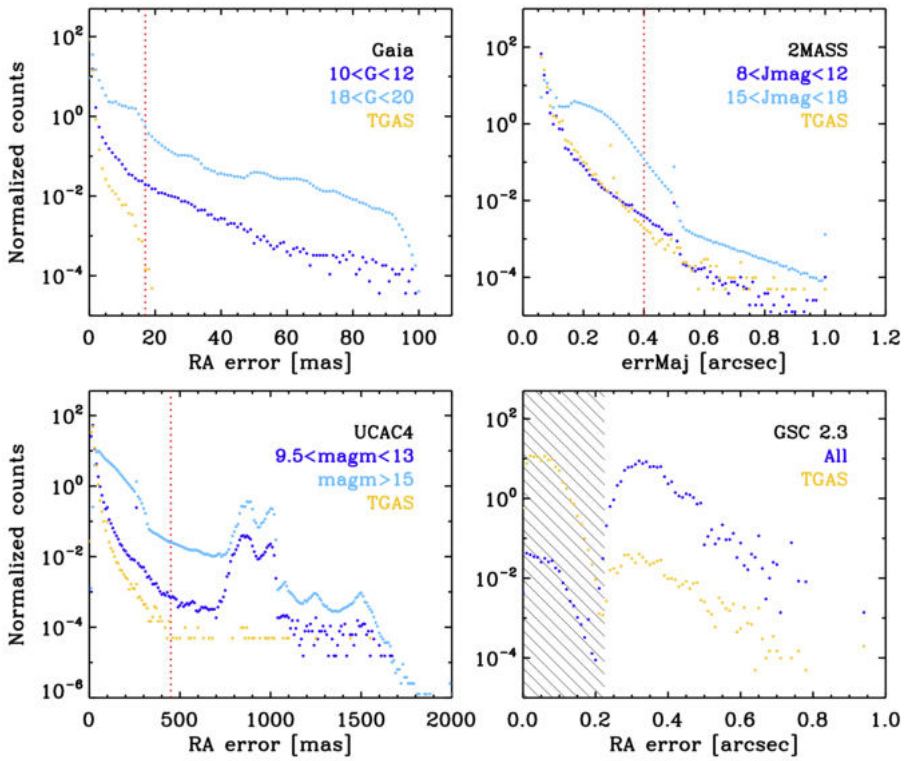


Fig. A.2. RA error distribution for *Gaia*, 2MASS PSC, UCAC4, and GSC 2.3. The position error distribution is shown in light blue for the bulk of the sources in each catalogue, and in blue for all the sources in each catalogue in the same magnitude range as TGAS (for *Gaia*) or TGAS counterparts (for the external catalogues). Each corresponding magnitude range is colour-coded and reported in the legend of each panel. In yellow is shown the distribution of the position errors for TGAS in the case of *Gaia* or the TGAS counterparts in the external catalogues matched using the proper motions. For GSC 2.3 the blue dots indicate the full catalogue error distribution. The distribution of Dec error shows the same behaviour.

Table A.2. Comparison between TGAS (or TGAS counterparts) and full catalogue error distributions.

		Gaia		
Position precision	TGAS sources	All	Bright	Bulk
RA error <= 17	~100%	94.82%	99.77 %	95.79%
		2MASS		
Position precision	TGAS counterparts	All	Bright	Bulk
errMaj <= 400	99.99%	99.66%	99.97%	99.49%
		UCAC4		
Position precision	TGAS counterparts	All	Bright	Bulk
RAerror <= 450	~100%	96.45%	99.47%	95.32%
		GSC23		
Position precision	TGAS counterparts	All		
RAerror <= 225	99.48%	0.35%		