XMM-LSS
INTERNAL
REPORT n°14-Mi

March 25, 2014

Steps towards identification of XXL X-ray sources

aka "the Geneva exercise 0.0"

L.Chiappetti[1]

INAF, IASF Milano, via Bassini 15, I-20133 Milano, Italy

Abstract. I report on the current preliminary exercise towards a multiwave XXL catalogue based on the material distributed by Sotiria Fotopoulou on 7 February 2014 to a restricted mailing list. A preliminary draft version of this report was uploaded on the XXL wiki on March 11, describing the steps I performed so far and the problems encountered and was promptly released to a restricted circle. *The current version supersedes such a draft.* My idea remains that the exercise should be repeated on a revised dataset, be it called 1.0 or 2.0.

Key words: LSS; XXL

1. Introduction

Since inception of the XXL project, the task of mantaining the XXL database has been assigned (*obtorto collo* in lack of better volunteers !) to IASF-Milano, building on the experience of the predecessor project, XMM-LSS (Chiappetti et al., 2013).

In the year before September 2013 the XXL database in Milano has reached a fairly stable state for what concerns the X-ray pointings. An overlap-free X-ray catalogue (actually divided in two parts, `XXLN` and `XXLS`) was released to the XXL consortium in such date, together with a relatively large number of correlated non-X-ray tables. Situation is described in Report XII (Chiappetti , 2013), available on the XXL wiki.

A *work version* multiwave catalogue (`XXLNOPT` and `XXLSOPT`) was created using the available tables, and announced to the database user and AGN mailing lists (on 9 October 2013).
Additional non-X-ray tables were added (and advertised to the database users) in November 2013 (updated SSDF V9 and DECam) and at end January 2014 (SSDS, WISE and IRAC North "V0"), in which date they were also advertised to the entire consortium.

Possible future updates (not yet applied) concerns the OM-SUSS table, the DECam table and the Marseille spectroscopy table (an update to each of such catalogues has been recently announced). Additionally the replacement of the XAMIN *3.3 X-ray tables with new ones based on* XAMIN *3.4 is under consideration.*

All the non-X-ray tables referred above are *subset tables* i.e. they include only the objects around (typically) $10''$ from X-ray source positions. The complete catalogues from which they are extracted are usually available in the public domain (or in some cases made available to the consortium by specific consortium members).

A mail by Sotiria Fotopoulou was circulated on 7 February 2014, with directions for retrieval of a large set of *complete* catalogues (and originally with an unrealistic deadline of one week for concurrent identification attempts). This dataset will be hereafter referred as *Sotiria's distribution* or *Sotiria's dataset* (for short, SD).

This report describes how the above catalogues were used in Milan concurrently with the previous tables used for `XXLNOPT` and `XXLSOPT` both for a comparison and prevalidation of the new datasets, and for a prelinary tuning of the traditional XMM-LSS-like procedure.

The plan of this report is as follows: section 2 describes the current dataset supplied by Sotiria (SD) and the potential problems and possible improvements; section 3 describes tne XMM-LSS-like procedure used in the past for identifications (3.1), the ingestion of the new data (3.2), the preliminaries for probability computation (3.3) and source association (3.4), and the new ranking performed only on newer data (3.5); section 4 first compares the application of the association procedure to the tables in SD to the preexisting tables in Milan DB (from 4.1 to 4.3), and then, from 4.4 onwards, describes the results of the application of the final steps of the procedure (probability ranking) to the new dataset only; while finally the conclusive section 5 indicates ideas for future "production" reprocessing as well as for possible tuning of the procedure.

Note that in the PDF version of this report links like this are active URLs giving access to the original web pages via the browser of your choice.

2. The input tables

The material distributed by Sotiria (SD) consists of 14 northern and 6 southern non-X-ray catalogues (plus a copy of `XXLN` and `XXLS`) in FITS format with a uniform layout packed in two `tar.gz` archives.
Two of the northern catalogues (IRAC1 and IRAC2) were later replaced by a revised version to cope with a problem with identifiers.

These catalogues are complete, i.e. include *all objects*, not just those in the surrounding of X-ray sources as it has been customary for the Milan DB. *(Note that the current choice of having a 10′ subset is incidental, and could be modified after motivated request. I actually often solicited whether such sample should be modified, for instance using a larger radius around extended X-ray sources).* The rationale behind the choice for the Milan DB were on one hand efficiency reasons (interactive queries on a relatively small table is faster), on the other hand competence reasons (complete databases are *in general* not useful for the X-ray-centric XXL work, also in most cases entire tables are publicly available on the *official* website of the specific survey, while in other cases their public usage is instead *interdicted* by data rights and Memorandums of Understanding).

When I say "*in general*", it means that I am aware of specific cases in which one might need the usage of all sources, e.g. to study the environment, to calibrate photometric redshifts, etc., or trivially, as occurred to me, to compute the density $n(brighter\ than\ m)$ necessary for chance probability computations (my page `http://sax.iasf-milano.inaf.it/~lucio/XXL/WebAux/probability.html` gives details on the various ways I made use of complete catalogues, either on the native official websites or ingesting them in temporary tables in the Milan DB).

Now Sotiria's distribution in principle validly raises a difference which makes worth reconsidering the above data policy. This difference is the fact that tables in SD are "standardized", and this does not refer only to the uniform format and layout (which is in itself a good thing, see section 2.4), but particularly to the fact a standard set of magnitudes is chosen, in a standard uniform system, and with standard corrections applied, at variance with what is distributed in the official sites.

While for the current release of SD I have been requested **not** to make the material available to the consortium via the Milan DB, the request by the PI that everybody in the consortium shall use an unique "endorsed" version of the non-X-ray tables suggests the following compromise:

- I will continue to ingest and release to the consortium a *subset* of a future version of each of Sotiria's tables (could be a 10″ subset or whatever requested). These will be *queryable* tables, accessed via queries using the interface of Milan DB (of course query results can then be exported).
- I might make available *entire* catalogues as *data products*. They will not be queryable, but can be retrieved in toto. The related files could reside in Milan or (preferably) elsewhere, but will anyhow be linked by the Milan DB providing the *data product URL*.

In all cases this will apply to a *next release* of SD, both because on one hand I understand that Sotiria herself does not consider current version as final, and because on the other hand I am convinced the current distribution requires upgrades in layout and content as suggested in the next subsections.

2.1. Documentation issues

Tables present in Milan DB stick to a reasonable documentation detail, which is usually presented in the so-called *table help* (the web page accessed clicking on a table name), in e-mail announcements (usually stored also on the XXL wiki) and often in detailed reports like this one (also stored or linked via the XXL wiki).

As a minimum, such documentation indicates the web site from where data have been retrieved, and the version or release of the data, plus any modification done with respect to the original/official data. I feel that such details are missing in current SD.

In particular since the version is not indicated, I am often confused in comparing data in SD tables with data in previously available Milan DB tables, and I do not know whether they refer to same or different versions, and which is referring to the latest version (which should be the choice). Compare in particular sections 4.2 and 4.3 amd figures from Fig. 3 to 8.

I therefore recommend that a future distribution includes such documentation details.

Another minor documentation item concerns column description, useful for ingestion and release via Milan DB, and are recalled later in sections 2.3 and 2.5.

2.2. Table name issues

In SD, catalogues are distributed as FITS files (see 2.4), named in the `tar.gz` archive with pathnames like e.g. `North/IRAC2_for_assoc.fits`, i.e. *emisphere*`/`*tablename*`_for_assoc.fits`.

A minor problem was that for revised IRAC1 and IRAC2 northern catalogues the directory name was `input` instead of `North`, which breaks the possibility of automated handling.

More relevant problems (in what is otherwise a perfectly sensible *file naming convention* occur during the ingestion of such FITS files into database tables.

- The Milan DB is articulated as a *single database*, i.e. it is not articulated in subdirectories, and given the

architecture of the DART interface *it is now too late to change*
- Therefore two tables in the northern and southern area cannot have the same name (compare e.g. GALEX, IRAC1, IRAC2, WISE). This was usually fixed renaming the southern table as e.g. `sGALEX`.
- Moreover `mysql` is not case sensitive. A table name can be written in upper, lower or mixed case, but they all refer to the same table. This might cause name clashes with existing tables like with `wircam` or `wise`. This was also fixed adding a suffix like e.g. `nWIRcam`.

Additionally, SD table names have no information on the release version, contrary to the convention used in Milan DB for most tables (there is e.g. an `ukidssdr10` or `galexgr6` table because they could later be replaced by a DR11 or GR7 issue, and actually have replaced earlier versions in the past, but for instance there is a `wise` table because it was taken from what should be the ultimate release).

I recommend that for the future table names are agreed in advance in such a way to avoid name clashes.

I present in table 1 a list of all *sensu lato* DB tables in three columns: SD tables *as ingested* (see 3.2) but not released, released tables used as member tables in `XXLNOPT` and `XXLSOPT` catalogues, released tables which became available after September 2013.

2.3. Column name issues

So far the Milan DB has tried to adopt a consistent naming convention, with some variations. One variation was between the *physical tables* and the *catalogues*. For the majority of the individual (physical) tables there was historically an attempt to name the same thing in the same way (for instance coordinates are called `ra, decl`, magnitudes are `mag`*band*, fluxes are `flux`*band* etc.). For catalogues (views with many member tables) a different convention was used for historical reasons (e.g. `fluxb` is renamed `Bflux`).

However the administration of the Milan DB has been found to be rather time consuming. One has for instance to insert column names, units, labels and description in an administrative database. This manual operation takes time. Therefore for recent ingestions I tried to automatize this loading the information from ASCII or FITS file headers. As a corollary, column names mantain their original name (for instance one data source may use `alpha_j2000` instead of `ra`).

Sotiria-supplied files appear to use an uniform column naming and content (see also 2.5). Each file has an *identifier* column, a pair of *ra,dec* columns and a number of *magnitude and error* pairs, one for each band.

The first three columns have names like *tablename*`_id`, *tablename*`_ra`, *tablename*`_dec`, i.e. prefixed by the table name, while the remaining columns have plain names like *band*`_MAG_`*method* and *band*`_MAGERR_`*method* where *method* is usually `AUTO`.

The prefix is somewhat redundant, since names like *tablename.columnname* are automatically generated by `mysql`, which will result in too verbose references like e.g. `nWISE.WISE_id`, somewhat complicated by the table renaming described above in 2.2. I suggest that the prefixes are removed in the final version.

A minor annoyance was caused for tables `BCS` and `DEcam` by the fact the sky coordinate columns were named in upper case (`RA` and `DEC`). I suggest that names of similar items always use consistently the same case, lower, upper or mixed. In the ingestion procedure (see 3.2) I forced such renaming manually.

Concerning declination, if the prefix is removed, I urge to use a name like `decl` not `dec` since "dec" is a reserved word in `mysql`.

I remind also that a *full fledged* database ingestion (which is **not** what was done in the current exercise, see 3.2) will require, besided the data ingestion, the filling of an *administrative database*, which, for each column shall provide: (a) a categorization for priority of listing (order); (b) the measurement units of the column; (c) a short caption or label; (d) a longer description; (e) optional VO UCD information. Filling those administrative information manually is very time consuming, therefore it would help if some of the administrative information (namely units and description) are provided "automatically" in the file header (see also 2.4).

2.4. File format issues

All files in SD are FITS files, and all of them (but the two X-ray tables produced directly from the Milan DB) have been produced by `topcat` (which uses some non-standard conventions but is however ok). The usage of an uniform format is definitely a bonus for database ingestion (although SQL requires FITS files to be converted to ASCII for ingestion) with respect to the heterogeneity of files coming from various archives and websites which usually required to adapt and debug an adhoc script instead of a single standard one (see 3.2).

A slightly minor problem is related to *identifier format* (see also 2.5). Some of the identifiers are *long numeric strings* (by long I mean it won't fit in a 32-bit integer). In SD such identifiers were coded with `TFORMn='K'` (64-bit integers), which is a somewhat unfortunate choice (for instance a viewer like `fv` seems not to handle it, `topcat` itself supports it in a ragged way [1], luckily my own FITS-ASCII converters support K-format decoding since 2010. Then for what concerns `mysql`,I harcoded handling of K integers as `char(20)` strings, despite the slight index inefficiency

[1] for instance one cannot search for an entry with a given value like `TABLE_id=value` but has to use `equals(toString(TABLE_id),"value")`

SD	XXL?OPT members	newer tables	Notes
North			
GALEX	galexgr6		
	omsuss		OM SUSS not considered by Sotiria
CFHTLS_W1	w1t7		tile not like '_'
CFHTLS_WA	w1t7		tile='A'
CFHTLS_WB	w1t7		tile='B'
CFHTLS_WC	w1t7		tile='C'
CFHTLS_D1	not loaded		
	zt7		CFHTLS zphot
	cfhtlens		CFHTLS alternate
SDSS		sdssdr10	
UDS			unknown to me
UKIDSS	ukidssdr10		
VIDEO			unknown to me
nWIRcam	wircam		Sotiria's table renamed
nWISE		wise	Sotiria's table renamed
IRAC1		not loaded	3.6μm-selected
IRAC2		irac2v0	4.5μm-selected
	swiredr6		SWIRE not considered by Sotiria
	marseillespec		
	ned		
	simbad		
	usno		
South			
sGALEX	galexgr6		Sotiria's table renamed
sBCS	bcsru		Sotiria's table renamed
	bcslmu		BCS alternate
sDECam		decam	Sotiria's table renamed
sWISE		wise	Sotiria's table renamed
sIRAC1	available not used	available not used	Sotiria's table renamed, 3.6μm-selected
sIRAC2	ssdf2v8	ssdf2v9	Sotiria's table renamed, 4.5μm-selected
	marseillespec		
	ned		
	simbad		
	usno		

Table 1. Database tables used in the present report in approximate wavelength order. The first column gives Sotiria's tables as ingested (not advertised). The second column gives Milan DB advertised tables used as members in XXLNOPT or XXLSOPT. The third column gives Milan DB advertised tables ingested after Sep 2013. Note that standard Milan DB non-X-ray tables may contain both northern and southern objects.

in using strings instead of numbers. I would suggest typing of identifiers is explicitly agreed in advance.

As anticipated in 2.3, I would recommend that, besides the column name (TTYPEn) and data format (TFORMn), one provides for each columns the measurement units except for adimensional quantities in TUNITn and a long column description in TCOMMn. The short label might default to the column name (therefore one should choose meaningful names).

2.5. Content issues

One of the first thing I made with the FITS files in SD was to inspect them not only for the format (see 2.4) but also for a quick check of the content compared (by sample) with the corresponding tables in the Milan DB (see correspondence in Table 1). Doing this I noticed three things:

- *identifiers* might match or not match
- *magnitude values* usually do not match
- Sotiria's files have a standard limited list of columns (see 2.3) while tables in Milan DB, although they have usually much less columns than the dataset available in the public archives, usually have more columns, including possibly useful "*service columns*"

Now the magnitude mismatch, although the most scientifically valuable, is possibly the least important and the most easily explained. Some catalogues were reprocessed, and most (all ?) magnitudes were normalized to a common system and standard corrections. All operations valuable

per se, which just need to be properly documented (see 2.1).

The other two items, identifiers and additional columns, are instead worth a detailed per-table examination.

2.5.1. Identifiers

Identifiers, from the point of view of the database, shall be *unique indices*. As such they allow for instance to update a table at different epochs avoiding the insertion of redundant entries (this feature was regularly exploited in the Milan DB).

Not all catalogues provide on their original sites an unique identifier in the form of a single column (but in `mysql` an index may refer to a combination of columns).

Identifiers are useful to make a reference to the original entry in the input (complete) catalogue, as well as for linking counterparts in member tables in a multiwave counterpart set. *However in the Milan DB, since I ingest subsets instead of complete catalogues, I prefer to use record sequence numbers (aka* `seq`*) for the latter purpose* and so I did also in the current exercise. Original identifiers can be accessed with a single level of indirection as any other column.

The following criticalities concerning individual tables were encountered

- For CFHTLS_W1 the column `CFHTLS_W1_id` appears to have nothing to do with the CFHT identifier (which is unique inside the tile) although there is a reasonable match in position. *Clarify ?*
- For SDSS the column `SDSS_id` appears to match *objid* in `sdssdr10`.
- For UKIDSS the column `UKIDSS_id` appears to match *id* in `ukidssdr10`.
- For WIRcam the column `WIRcam_id` appears to match `wircam.seq` (there was no other identifier in the files supplied by J.Willis).
- For IRAC north, the column `IRAC?_id` has been revised after the wrong column (not unique) was chosen by mistake. I presume now is some sort of sequence identifier in the complete dataset, since Bristol's distribution *did not include a proper id*. Therefore there is no obvious match with my `irac2v0` table (which has a local `seq` and uses the combination (α, δ) as identifier).
- For IRAC south (aka SSDF) I do not know the origin of the column `IRAC?_id`. I have no id (just a local `seq`) in all my SSDF tables (`ssdf2v8` and `ssdf2v9` etc.) since the original distribution by A.S.Stanford had no id (I am using the combination of tile id and X,Y coordinates for such purpose). Therefore there is no obvious match and I am uncertain *whether SD contains version 8 or version 9*.
- For WISE (N and S) the column `WISE_id` has the format of an IAU catalogue identifier, as it has my column `wise.designation`. However for associated objects they do not match (in some cases they do, in other they have little differences like e.g. `J231351.37-541359.5` and `J231351.39-541359.7`. This has to be clarified: *I suspect different versions were used*: Milan DB `wise` is based on AllWISE, the most recent (Nov 2013) release.
- For GALEX (N and S) the column `GALEX_id` appears to usually match `id` in `galexgr6`.
- For BCS I do not know the origin of the column `BCS_id`, which looks some sort of sequential identifiers. *It has nothing to do* with the `id` in my table `bcsru` (which is not unique and should be used in conjunction with the `tile` id).
- For DECam I do not know the origin of the column `DECam_id`, which looks some sort of sequential identifiers. *It has nothing to do* with the `id` in my table `decam` (which is not unique and should be used in conjunction with the `tile` id).

2.5.2. Additional columns

The following criticalities concerning individual tables were encountered

- For CFHTLS W1, my table `w1t7` had among others more columns like flags indicating whether the source is extended, masked or saturated, plus the *tile identification* (needed to associate images as data products) and the `ugriz` flag. It should be useful to reintroduce them.
- For the WA, WB and WC datasets it would be useful to have them in the same table with the other CFHT data (similar to what `w1t7` did), and to handle the *overlaps* between the ABC fields themselves and with the W1 tiles.
- Why are additional (my table `zt7`) or alternate my table `cfhtlens` CFHTLS datasets (including also "official" or semiofficial photometric redshifts) not considered ?
- For SDSS my table `sdssdr10` has many more columns. While I could not doubt that some of the different magnitude systems are redundant, and welcome a simplyfying choice by Sotiria, I think that some of the additional columns (particularly those coming from the SDSS spectroscopy, at least *redshifts*) should be recovered.
- For UKIDSS my table `ukidssdr10` included some flags (`class` and `multi`) from the original dataset, a flag telling whether the `survey` was DXS or UDS, and (on request by O.Melnyk) alternate "Hall" magnitudes (however not always present in the input data). While I welcome a simplyfying choice by Sotiria, I wonder if some *additional columns* should be recovered.
- For WIRcam my table `wircam` had additional aperture magnitudes as well as repeated *ugriz* ones. However,

with the possible exception of flags, I would be ready to accept Sotiria's choice (keep only K_s).
- For IRAC in the north, the main difference in the approach was that I ingested only the 4.5μm-selected catalogue (`irac2v0`). SD includes both the 3.6μm-selected (IRAC1) and the 4.5μm-selected (IRAC2). This has been justified by the fact the two surveys do not fully overlap at the edges. However there could be sources present only in IRAC1 (anyhow with measurements in both 3.6μm and 4.5μm bands), sources present only in IRAC2 (also with measurements in both bands), and sources present in both IRAC1 and IRAC2, with possibly different positions and two different sets of magnitudes (2 in each of the 2 bands), which is confusing and should be *somehow merged*. In addition my table has additional columns, privileging fluxes to magnitudes (which was an unfortunate choice since in this preliminary versions fluxes are in unusual and perhaps uncalibrated units), providing Bristol's CFHT association, and some other. While I would be ready to accept Sotiria's choices about magnitudes, additional columns shall be evaluated. Also one should consider that Bristol's distribution was a "version 0.1" likely to be soon superseded by a new one.
- For IRAC in the south, there is the same different approach about the two 3.6μm-selected and 4.5μm-selected caalogues. This survey corresponds to SSDF. I ingested and advertised, as directed by A.S.Stanford, the 4.5μm-selected both for public version 8 `ssdf2v8` and for the more recent private version 9 `ssdf2v9`. Actually I ingested but did not advertise also the other selection. Some form of *merging* as discussed above shall be considered. In addition my table has additional columns, privileging fluxes to magnitudes (in this case real fluxes) and some other like tile identification. While I would be ready to accept Sotiria's choices about magnitudes, additional columns shall be evaluated, specially *tile identifiers*.
- For WISE my table `wise` (common to north and south) had additional columns (flags etc.) somehow proposed by Geneva (N.Fourmanoit). However I would be ready to accept Sotiria's choice.
- Why is SWIRE not considered ?
- For GALEX my table `galexgr6` (common to north and south) had a few additional columns, i.e. privileging fluxes to magnitudes, and including tile identification and a tiling flag (homegrown). While I would be ready to accept Sotiria's choices about magnitudes, additional columns shall be evaluated and in particular it is mandatory to ensure the new dataset is free from *tiling overlaps*.
- For BCS there is the *vexata quaestio* of the two analyses (Rutgers and LMU) which appear to diverge already at image and sky coverage level. In Milan DB I stored independently results from both (`bcsru` and `bcslmu`) and with additional columns (tile identification and photometric redshift). While a full reanalysis which I presume was done by Sotiria would render preexisting zphot useless, one should consider to preserve *tile identifiers*.
- For DECam my table `decam` included additional per-band flags and information about *tiles and tiling overlaps*, as well as a ***griz** flag* telling which bands are covered by the specific tile. Possibly part of this information shall be recovered.

3. The LSS-like procedure

3.1. Previous work

The creation of the work version `XXLNOPT` and `XXLSOPT` multiwave catalogues was done with a procedure mimicked on the one described in Chiappetti et al. (2013). Essentially one has the following steps:

- I create a GCT (generalized or "glorified" correlation table), i.e. one having as many columns as member tables, and named *as* the corresponding member table.
- Initially I populate the GCT with the content of the pre-existing GCT underneath the X-ray catalogue (e.g. for `XXLNOPT` I copy the entries in `XXLN` having as members `north33`, `north33b` and `north33cd`)
- Then I add "placeholder entries" for "duplicated X-ray sources" (this step is unnecessary for identifications, it is just useful to see the position of nearby X-ray object in overlapping fields as overlay on images; this step will not be considered in the current exercise).
- Then I run a "`ptr`" script which handles the first optical table. It considers all optical objects within 6″ of each X-ray source, and may result in editing existing records or adding new records, namely
  - an X-ray source `x:b:cd` with a single counterpart in table `opt` will result in an entry `x:b:cd:opt=i`
  - if there are more counterparts, additional entries are added, e.g. `x:b:cd:opt=j` besides `x:b:cd:opt=i`
  - if the X-ray source has no counterparts the pointer remains null `x:b:cd:opt=null`
- Another "`ptr`" script which handles the next member table. It also considers possible counterparts within 6″ of each X-ray source, but compares their position within a predefined smaller radius (from 0.5″ to 2″ depending of the tables).
  - if an object in the next table is associated to an X-ray source and is within the smaller radius of a counterpart in the previous optical table, the record is updated as `x:b:cd:opt:nxt=k`
  - if the object in the next table is associated with the X-ray source but not to any object in the previous optical table, an entry is added as `x:b:cd:opt=null:nxt=k`

  - if the X-ray source and its previous counterparts have no counterparts in the next table its pointer remains null `x:b:cd:opt:nxt=null`
- A similar step is repeated for each further member table, with a further adhoc "`ptr`" script which compares it to all previous members within predefined radii
- At the end one should somehow perform a ranking procedure of the obtained counterpart sets, e.g. using the probabilities described in `http://sax.iasf-milano.inaf.it/~lucio/XXL/WebAux/probability.html` however this step was not performed for `XXLNOPT` and `XXLSOPT`

The above procedure is rather cumbersome, as all "`ptr`" scripts have to be customised by hand (adding new members, hardcoding the correlation radii, handling table-specific peculiarities like the name of the RA,Dec columns and other).

3.2. Ingestion

The pretty standard format of FITS files in SD allowed to write a general script to handle "quick" database ingestion. By "quick" ingestion I mean (a subset of) the content of the files is ingested in a database table, but the *administration information necessary for advertising to users is* **not** *filled*.

The script is invoked with a calling sequence like e.g. `tempingest.csh North IRAC1 ch1_MAG_AUTO ra dec` i.e. I have to pass as arguments also the names of the "preferred magnitude" column and RA,Dec columns.

The need to pass the RA,Dec column names is due to the fact two tables used different (case) names as described in 2.3. These names were later normalized to lower case.

The script suffered some adjustments on the fly to cope with renaming of tables as described in 2.2 and Table 1.

The choice of the "preferred magnitude" column is explained describing what the script does:

- FITS file is untarred to staging area
- FITS file is converted to ASCII into staging area
- `TFORM` and `TTYP` are extracted from FITS header and used to generate the list of columns in the `create table` SQL statements
- the ASCII file is loaded into the created temporary database table (e.g. `tempIRAC1`)
- the 10″ subset around X-ray position is extracted into the definitive database table (e.g. `IRAC1`)
- another semitemporary table (e.g. `densIRAC1`) is preserved. It contains for *all* objects just RA,Dec truncated to 2 decimal figures (0.01 deg) and the "preferred magnitude". This will be used in the step described below in 3.3

So the idea is to keep only the 10″ subset as permanent table. The resulting "compression factor" is reported in Table 2.

Table	preferred mag	ingested	total
CFHTLS_W1	i_old_MAG_AUTO	120132	5506442
CFHTLS_D1	i_old_MAG_AUTO	17150	554396
CFHTLS_WA	r_MAG_AUTO	2799	116958
CFHTLS_WB	r_MAG_AUTO	4057	112602
CFHTLS_WC	g_MAG_AUTO	1076	48239
GALEX	NUV_MAG_AUTO	36218	1145789
IRAC1	ch1_MAG_AUTO	37938	1182681
IRAC2	ch2_MAG_AUTO	34159	1031289
SDSS	i_MAG_modelmag	15217	436518
UDS	i_MAG_AUTO	23296	776787
UKIDSS	K_MAG_AUTO	24263	661735
VIDEO	K_MAG_AUTO	13399	399535
nWIRcam	K_MAG_auto	50885	2407202
nWISE	W1_MAG_AUTO	13057	301389
sBCS	i_MAG_AUTO	76848	3284726
sDECam	i_MAG_AUTO	123887	6072801
sGALEX	NUV_MAG_AUTO	10989	310133
sIRAC1	ch1_MAG_AUTO	77972	4147550
sIRAC2	ch2_MAG_AUTO	82626	4364890
sWISE	W1_MAG_AUTO	10186	238291

Table 2. For each table it reports the chosen "preferred magnitude", the number of objects ingested within the 10″ sample, and the total number of objects in the original files.

Note that I assumed that `i_old` corresponded to what CFHTLS calls *i'* and `i_new` to the so-called *y* magnitude but it is possible that such assumption is incorrect. *See note at end of 4.3 !!*

3.2.1. Correlation tables

I have also adapted the current general-purpose fast script which creates the (6″) correlation tables between any generic table and an X-ray table, into a special script which does the same, but does not register the correlation table in the administrative database. I have run it for all tables. In production version the original script will have to be used.

3.3. Probability computation

The customary (XMM-LSS style) *probability ranking* (see Chiappetti et al. (2013) and references therein) presupposes the computation of a probability

$$probability = 1 - exp(-\pi\ n(brighter\ than\ m)\ r^2)$$

which depends on the distance between the X-ray source and the counterpart, on the brightess of the counterpart (magnitude or flux), and on the density of sources $n(brighter\ than\ m)$. This density is usually coarsely modelled with a linear fit in log-log space.

Although I haven't computed such probabilities for XXL so far, I systematically collected material for the modelling of the density for most tables ingested

(and advertised) in the Milan DB. All details, including data source, methods, coefficients and plots, are reported in page `http://sax.iasf-milano.inaf.it/~lucio/XXL/WebAux/probability.html` and won't be repeated here.

As a propedeutic exercise to probability ranking for the new tables in SD, I computed density and coefficients as described below.

3.3.1. Area

The estimate of the area covered by a particular catalogue/survey is approximated in the following way. I use the semi-temporary tables described in section 3.2. Since they contain the coordinate of all sources truncated to 0.01 deg, and assuming sources are so thick that there is at least one in each 0.01×0.01 deg box, the count `select count(distinct r1,d1)/10000 from dens`*table* is the covered area in square degrees.

The computed areas are reported in Table 3 and can be compared with the areas reported in my web page for the tables advertised in the Milan DB.

I am not sure whether the results for `densCFHTLS_W1` make sense, since they were computed using `i_old` magnitude which I assumed to be i' but it looks like to be y instead ! *See note at end of 4.3 !!*

The areas for WA, WB and particularly WC looks unexpectedly low (I'd expect them to be about 1 square degree each).

The areas for GALEX and SDSS are also low than what I computed (26 or 22 sq deg instead of 55). However I computed them for a rectangular area enclosing XXL while I do not know the selection area for Sotiria's tables (see lack of documentation in section 2.1). *I suspect they were tailored to the XXL FOVs.* Similar argument for WISE (19 instead of 43), BCS (41 instead of 49), DECam (42 instead of 52), southern GALEX (19 instead of 67), southern WISE (19 instead of 42) and SSDF (IRAC, 43 instead of 52).

3.3.2. Number density

The next step is to compute the number counts $N(< m)$ (the actual densities are obtained as $n(< m) = N(< m)/area$. This is done in `mysql` running a simple script which is edited each time to replace the database table name and magnitude column name, saving results in a temporary table and storing that in an ASCII file with the number count in 0.5 magnitude bins.

The script benefits of the standard format of Sotiria's files. The results are not always comparable with what reported in my pages, since in some cases the database tables privileged flux instead of magnitude, or used different magnitudes.

Anyhow the results are plotted in the next section.

3.3.3. Coefficients

Densities are fitted approximating them with a linear relation in log-log space $n(< m) = 10^{a+bm}$ using an IDL procedure. The fit is performed in a magnitude range chosen empirically case by case. The same IDL procedure is also used to produce the plots in Fig. 1 and 2.

The coefficients a and b, together with the fitted magnitude range, are reported in Table 3.

3.4. Identification procedure

I have used standard scripts to create two GCTs `testxon` and `testxos` with all chosen member tables (see below) and initialize them with the X-ray data (first two steps in 3.1, the third step, placeholder creation, was omitted).

I made up a parametric script `ptr-first.newsql` to run the first step, i.e. handle the first optical table. It is called passing such table as argument like
`wrapper4.csh ptr-first.newsql glorxxln testxon north33 CFHTLS_W1` or
`wrapper4.csh ptr-first.newsql glorxxls testxos south33 sBCS`.

After such step I have in the north, for 14134 X-ray sources, that 11837 have 31073 counterpars (31040 distinct), and 2297 have none. In the south, for 11863 X-ray sources, 10176 have 21328 counterparts (21289 distinct) and 1687 have none.

The next steps are driven by a configuration file. The configuration file lists member tables in order of processing. This order is not the same as the physical order in which member tables are listed in the GCT (which is controlled by the "traditional" script for the creation step).

The configuration file does not just list the member tables *but also their relevant RA,Dec columns*, because such name may vary for Sotiria's tables because of her prefix-convention, but also for Milan DB tables because of the different naming convention used for older or newer tables, as described in 2.3. E.g.

```
CFHTLS_W1 CFHTLS_W1_ra CFHTLS_W1_dec
nWIRcam WIRcam_ra WIRcam_dec
w1t7 ra decl
cfhtlens ALPHA_J2000 DELTA_J2000
```

The choice of member tables for the current exercise has been the following:

- in the GCT I place first the X-ray tables
- then the Milan DB members used for `XXLNOPT` and `XXLSOPT` i.e. those in the second column of Table 1
- then the "newer" Milan DB members (third column of Table 1)
- finally Sotiria's tables (first column of Table 1)
- Concerning the processing order instead, Sotiria's tables are processed first (with the idea their successors

Table	Area (sq deg)	magnitude range			a	b
CFHTLS_W1	24.2736	i (old ?)	15	25	-9.97803	0.289021
CFHTLS_D1	1.0198	i (old ?)	15	25	-9.27190	0.291301
CFHTLS_WA	0.7765	r	15	25	-9.60764	0.301016
CFHTLS_WB	0.7761	r	15	25	-9.41931	0.291846
CFHTLS_WC	0.2194	g	15	25	-9.88457	0.303393
GALEX	26.1076	NUV	15	22	-11.0099	0.340601
IRAC1	23.1013	3.6μm	13	21	-10.5944	0.372975
IRAC2	22.7215	4.5μm	13	21	-10.9313	0.385549
SDSS	22.3410	i	16	20	-9.18764	0.288256
UDS	1.2524	i	15	25	-10.2747	0.334115
UKIDSS	6.1069	K	12	20	-9.97558	0.344430
VIDEO	1.8084	K	13	22	-9.76473	0.332281
nWIRcam	16.7183	K	16	23	-11.3564	0.400782
nWISE	19.2658	W1	12	18	-9.71328	0.325762
sBCS	41.4699	i	13	22	-16.3704	0.634184
		i	13	18	-18.2582	0.756058
		i	18	22	-10.3099	0.334430
sDECam	42.1223	i	15	24	-11.6731	0.370729
sGALEX	19.4623	NUV	15	22	-11.4070	0.349856
sIRAC1	42.9423	3.6μm	13	21	-11.4481	0.409375
sIRAC2	43.0632	4.5μm	13	21	-11.6774	0.415645
sWISE	19.0751	W1	12	18	-9.72923	0.328101

Table 3. Density fit coefficients computed for each table in the given magnitude range and converting number counts to density using the indicated area. A single power law fit is particularly poor for BCS, in which case a double power law fit is also indicated. For CFHTLS W1 I doubt that the magnitude chosen is what should be intended.

will be the only one processed in the production version)
- the Milan DB members used for `XXLNOPT` and `XXLSOPT` are processed next (but not the `usno` table)
- the "newer" Milan DB members are processed next
- the `usno` table is processed last (the "last" table, however defined in the "superwrapper" script, requires some special "final" processing)

An awk script onto the configuration files is used to generate a list of invocation of the "superwrapper" script in the wished processing order, e.g.

```
superwrap.csh CFHTLS_D1 testxon glorxxln north33
superwrap.csh CFHTLS_WA testxon glorxxln north33
...
superwrap.csh usno testxon glorxxln north33
```

The superwrapper script generates a "`ptr`" script for the appropriate table which makes reference to all *preceding* members, and submits them to `wrapper4.csh`.

There is one important detail which is not handled in the configuration script, but it is asked interactively, and that is the correlation radii between the current member and all its preceding members. These values are logged to a file, and this logfiles have been used to generate Tables 4 and 5.

The procedure was tested emulating `XXLNOPT`, i.e. writing a configuration file for its GCT, and running the superwrapper script in dry-run mode (generating the "`ptr`" scripts without executing), and comparing the generated scripts with the ad-hoc "`ptr`" scripts used originally.

The new procedure is general and automatic enough, however in the specific execution described in this report, it requires a *manual patch at least* for table `galexgr6`. In fact such tables contains (specially in the northern area) what are likely to be several repeated observations of the same object. Since they are tentatively flagged with a `tiling` flag, the patch is such that only one object is used for such cases. I do not know whether Sotiria's `GALEX` tables are affected by the same problem but I did not apply any patch for them.

Note that `XXLNOPT` included a member table `zt7`, but since this shares identifiers with (a subset of) `w1t7` entries, it was correlated on identifier and not on distance. Since this is of no interest for the present exercise, `zt7` was defined as member but not processed (and its pointers left empty).

The preliminary draft ended here with the description of the procedure, and resumed with subsections 4.1 to 4.3 (and the prologue) of section 4.

3.5. The refined ranking procedure

The ranking procedure used for the published `2XLSSOPTd` multiwave catalogue (Chiappetti et al., 2013), generically referred as "XMM-LSS-like procedure", involved the following logical steps:

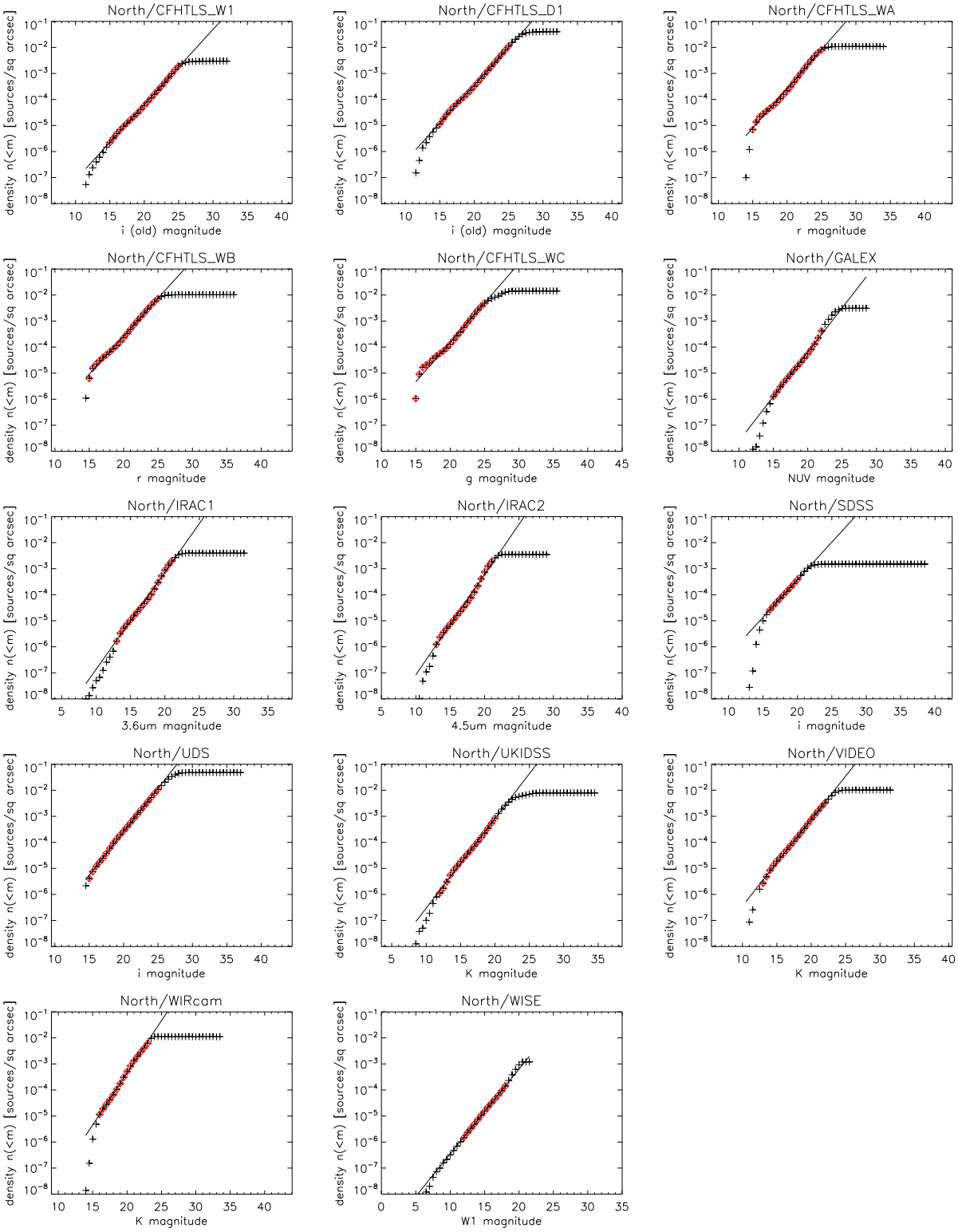

Fig. 1. Density fits in the northern area. Red diamonds indicates the points used in the fit. Coefficients resulting from the fit are reported in Table 3

– computation of the *probabilities* described in 3.3 and using the coefficients computed therein

– *preranking* counterpart sets on the basis of such probabilities being < 0.01 (good), $0.01 < prob < 0.03$ (fair)

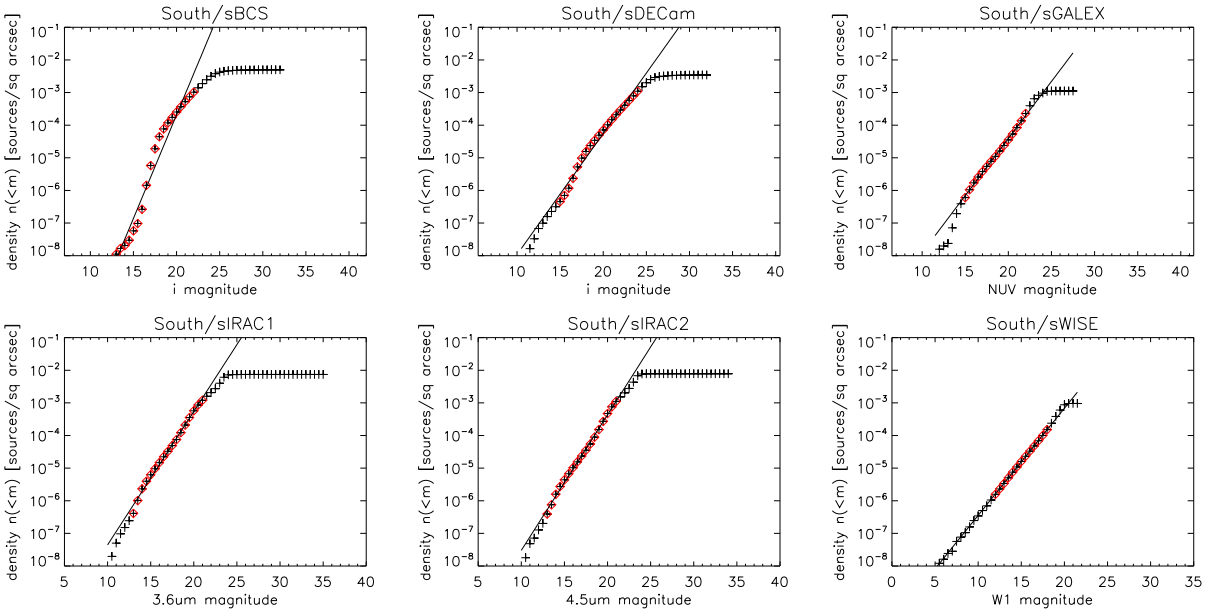

Fig. 2. Density fits in the southern area. Red diamonds indicates the points used in the fit. Coefficients resulting from the fit are reported in Table 3

Tables	CFHTLS_W1	CFHTLS_D1	CFHTLS_WA	CFHTLS_WB	CFHTLS_WC	SDSS	UDS	nWIRcam	UKIDSS	VIDEO	nWISE	IRAC1	IRAC2	GALEX	w1t7	zt7	cfhtlens	wircam	ukidssdr10	swiredr6	galexgr6	omsuss	marseillespec	ned	simbad	sdssdr10	wise	irac2v0
CFHTLS_D1	0.5																											
CFHTLS_WA	0.5	0.5																										
CFHTLS_WB	0.5	0.5	0.5																									
CFHTLS_WC	0.5	0.5	0.5	0.5																								
SDSS	0.5	0.5	0.5	0.5	0.5																							
UDS	0.5	0.5	0.5	0.5	0.5	0.5																						
nWIRcam	1.0	1.0	1.0	1.0	1.0	1.0	1.0																					
UKIDSS	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0																				
VIDEO	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0																			
nWISE	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0																		
IRAC1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0																	
IRAC2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0																
GALEX	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5															
w1t7	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1.0	1.0	1.0	1.0	1.0	1.0	1.5														
cfhtlens	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1.0	1.0	1.0	1.0	1.0	1.0	1.5	0.5	0.5												
wircam	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	1.0	1.0	1.0	1.0	1.0	1.5	1.0	1.0	1.0											
ukidssdr10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	1.0	1.0	1.0	1.0	1.5	1.0	1.0	1.0	1.0										
swiredr6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.5	1.0	1.0	1.0	1.0	1.0									
galexgr6	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	0.5	1.5	1.5	1.5	1.5	1.5	1.5								
omsuss	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5							
marseillespec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.5	1.0	1.0	1.0	1.0	1.0	1.0	1.5	1.5						
ned	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	2.0	1.5	1.5	1.5	1.5	1.5	1.5	2.0	2.0	1.5					
simbad	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	2.0	1.5	1.5	1.5	1.5	1.5	1.5	2.0	2.0	1.5	1.0				
sdssdr10	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1.0	1.0	1.0	1.0	1.0	1.0	1.5	0.5	0.5	0.5	1.0	1.0	1.0	1.5	1.5	0.5	1.5	1.5			
wise	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	1.0	1.0	1.5	1.0	1.0	1.0	1.0	1.0	1.0	1.5	1.5	1.0	1.5	1.5	1.0		
irac2v0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.5	1.5	1.0	1.0	1.0	1.0	1.0	1.0	1.5	1.5	1.0	1.5	1.5	1.0	1.0	
usno	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.5	1.0	1.0	1.0	1.0	1.0	1.0	1.5	1.5	1.0	1.0	1.0	1.0	1.0	1.0

Table 4. Correlation radii (in arcsec) used for the northern test table `testxon` (in processing order). Note that table `zt7` (sharing id's with `w1t7`) has not been used in the current exercise.

or > 0.03 (bad) (involving also rejection of the worst counterpart sets)
- *ranking* of counterpart sets according to *scores* computed according to heuristic recipes
- analysis of *ambiguities*, with the idea of seeing if an X-ray source has a single counterpart set with a valid rank or more than one
- postprocessing, making sure each X-ray source has just one (actually or nominally) preferred counterpart set while keeping track of other non-rejected candidates.

Tables	sBCS	sDECam	sWISE	sIRAC1	sIRAC2	sGALEX	bcsru	bcslmu	ssdf2v8	galexgr6	marseillespec	ned	simbad	decam	ssdf2v9	wise
sDECam	0.5															
sWISE	1.0	1.0														
sIRAC1	1.0	1.0	1.0													
sIRAC2	1.0	1.0	1.0	1.0												
sGALEX	1.5	1.5	1.5	1.5	1.5											
bcsru	0.5	0.5	1.0	1.0	1.0	1.5										
bcslmu	0.5	0.5	1.0	1.0	1.0	1.5	0.5									
ssdf2v8	1.0	1.0	1.0	0.5	0.5	1.5	1.0	1.0								
galexgr6	1.5	1.5	1.5	1.5	1.5	0.5	1.5	1.5	1.5							
marseillespec	1.0	1.0	1.0	1.0	1.0	1.5	1.0	1.0	1.0	1.5						
ned	1.5	1.5	1.5	1.5	1.5	2.0	1.5	1.5	1.5	2.0	1.5					
simbad	1.5	1.5	1.5	1.5	1.5	2.0	1.5	1.5	1.5	2.0	1.5	1.0				
decam	0.5	0.5	1.0	1.0	1.0	1.5	0.5	0.5	1.0	1.5	1.0	1.5	1.5			
ssdf2v9	1.0	1.0	1.0	0.5	0.5	1.5	1.0	1.0	0.5	1.5	1.0	1.5	1.5	1.0		
wise	1.0	1.0	0.5	1.0	1.0	1.5	1.0	1.0	1.0	1.5	1.0	1.5	1.5	1.0	1.0	
usno	1.0	1.0	1.0	1.0	1.0	1.5	1.0	1.0	1.0	1.5	1.0	1.0	1.0	1.0	1.0	1.0

Table 5. Correlation radii (in arcsec) used for the southern test table `testxos` (in processing order)

Such procedure has been applied for the first time to the XXL now. It has not been applied to the pre-existing `XXLNOPT` and `XXLSOPT` work version catalogues, but only to the new GCTs `testxon` and `testxos`, and in particular it has used only the counterparts in SD tables (the entries flagged "old only" - see section 4 below - are ignored).

In the coding of the scripts I used for the procedure, I took advantage of the opportunity to generalize them (while the original scripts used for `2XLSSOPTd` were rather ad-hoc) in such a way they can be re-used in a future production version changing items like the table list in a configuration file.

3.5.1. Actual probabilities

While the computation of probability is conceptually simple (one has to apply the formula given in 3.3 using the coefficients in Table 3), making it general is operationally complicated by the fact the list of non-X-ray tables varies from northern to southern area, or from current to future release (and the related columns, and the coefficient values vary from table to table).

The solution, similar to what described in 3.4, is to use *wrapper scripts driven by a configuration file* invoked as

```
probwrap.csh testxon glorxxln north33
probwrap.csh testxos glorxxls south33
```

to write an adhoc script of `mysql` commands and then invoke it.

Such script reads a configuration file (e.g. `testxon.prob.config`), with one (or more, see below) row per table. Each row lists:

- the member table name
- the RA and Dec column names
- the preferred magnitude column name
- the A and B coefficients of the density fit (Table 3)
- an optional repeat flag

The `mysql` script creates in the GCT a probability column named according to the member table (e.g. `sBCS` gives `probsBCS`). The probability is then computed from the magnitude and distance of the counterpart from the X-ray position. Undefined magnitudes, set to 99, may occur in two cases: if there is no counterpart in the specific member table, or if the preferred magnitude is undefined.

In the future, if a magnitude in the preferred band is undefined, one could use another band in decreasing preference order. An approach like this was hardcoded in the XMM-LSS scripts for some speficic tables like ***`swiredr6`*** *or* ***`galexgr6`****.*

The optional repeat flag is provided for this purpose. One just lists more rows for the same member table, each one with a different magnitude band, and flags all but the first. This inhibits the creation of the probability column. The existing column is filled with the new probability value only if still undefined (so the order of appearance gives the preference). Of course one would have to compute separate density fit coefficients for each magnitude band, which would require a change in the ingestion procedure.

In the current run, considering the doubt on the meaning of ***`i_old`*** *and* ***`i_new`*** *vs i' and y magnitudes for CFHTLS W1 (see end of 4.3), I have used the repeat flag for* ***`CFHTLS_W1`*** *using the same coefficients computed for what is nominally called* ***`i_old`*** *for both magnitudes.*

3.5.2. Pre-ranking

The XMM-LSS pre-ranking script used to work with *four* probabilities (a GALEX one, an UKIDSS one, a SWIRE one and an optical one obtained grouping W1 (and ABC) and D1).

To mantain the same script structure preserving a general form, while our GCTs may have a larger number of probabilities, I defined four "grouped" probabilities:

- `probXO`, the *optical* probability which takes the smallest (best) probability for all optical tables (W1, D1, WA, WB, WC, SDSS and UDS in the north, BCS and DEcam in the south)
- `probXN`, the *near IR* probability which takes the smallest for NIR tables (WIRcam, UKIDSS, VIDEO in the north, undefined in the south)
- `probXI`, the *IR* probability, smallest among WISE, IRAC1 and IRAC2 for both areas
- `probXG`, renamed from the GALEX probability for both areas

These 4 probabilities are additional columns in the GCT precomputed before invoking the preranking script, which is another *wrapper scripts driven by a configuration file* invoked as

```
prerankwrap.csh testxon glorxxln north33
prerankwrap.csh testxos glorxxls south33
```

which writes an adhoc script of `mysql` commands and then invoke it.

However the script uses three configuration files. One is the same used by the probability wrapper described in the previous section. The other two are called e.g. `testxon.prob.config.opt` and `testxon.prob.config.ir`. They just list, on separate rows, *only* the optical or respectively IR members and their preferred magnitude. *For `CFHTLS_W1` there are two rows, one for each old and new magnitudes.*

The preranking scripts performs the following steps:

- creates the necessary columns `rank` and `autorank` and initializes them as undefined
- assigns `autorank=4` to blank fields, i.e. X-ray sources with no counterpart in any of the member tables listed in the main configuration
- assigns `autorank=0` for *unique* counterparts which are good in at least one of the four probabilities and at least fair or undefined in the other
- assigns `autorank=1` for *unique* counterparts which are good in at least one probability and still undefined
- assigns `autorank=2` for *unique* counterparts which are fair in at least one probability (and still undefined)
- assigns `autorank=3` for other *unique* counterparts
- assigns `autorank` 1, 2 or 3 to the "brightest and closest" best by probability (good, fair or otherwise)
- adds 10 to `autorank` for the "best by optical magnitude"
- adds 20 to `autorank` for the "best by IR magnitude" (these two steps use the secondary configuration files)
- adds 100 to `autorank` for the "best by distance"
- at the end `autorank` may assume a set of discrete values, like e.g. 11, 12, 13 or 19, 21 etc., 31 etc., 111 etc., 121 etc., 131 etc.
- according to a set of recipes such values and the probabilities are used to assign a temporary value from 90 to 93 to the `rank`
- except that objects which an intermediate stage do not fall in any of the recipes are *rejected* i.e. flagged `rank=-1`
- Finally not rejected `rank` is reset in the range 0 to 3

For convenience at the end of this stage *entries flagged "old only", which are nominally "blank fields" since they have counterparts in the Milan DB members but not in SD members, are also flagged rejected (`rank=-1 autorank=4`).*

3.5.3. Ranking

The adaptation of the original `2XLSSOPTd` ranking script to the general XXL case has been simpler. One does not need any configuration file, but just a template SQL script `ranking.newsql` in which one has just to replace the name of the GCTs and of the main X-ray member table.

The script ignores *rejected* (`rank=-1`) and *blank field* (`autorank=4`) entries, and processes those which have been provisionally assigned a rank of 0 or 1 as well as those with a still undefined rank of 9.

It will re-rank cases of single counterpart sets (now considering singles also those with a single *surviving* set even if other counterpart sets were rejected) as ranks 0 or 1 according to *autorank* being 0,1 or 2,3.

It will re-rank as 0 or 1 the best case of multiple counterpart sets, and assign a rank of 2 to the other cases.

It will perform two more repeated passes to make sure to process all entries, the last pass sorting on probability, and doing a further *rejection* at an intermediate stage (this gets rid of some cases with undefined probabilities).

Ideally the objects with *all good* probabilities and already flagged as preferred get here a `rank=0`.

3.5.4. Ambiguity analysis

The ambiguity analysis also occurs via a couple of template scripts adapted in general form from the `2XLSSOPTd` version.

Tne template scripts are called `unambig.newsql` and `ambig.newsql`. These scripts are rather similar at least in their initial part. They ignore the rejected counterpart sets (those flagged `rank=-1`) and use only entries flagged `rank between 0 and 2`.

The first script ("unambiguous") works on the X-ray sources which have just one counterpart set with the above valid ranks. The second script ("ambiguous") on those which have more than one.

The following euristic *rules* (so called "Brera rules" dating back to the XMDS analysis) and associated scores are considered:

- the entry with `rank between 0 and 1` is flagged *selected*
- the entry having either the optical or the IR probability good (< 0.01) is said to satisfy *strongly rule 1*, which adds 1 to the score
- if the optical or the IR probability is fair (< 0.03) it is said to satisfy *weakly rule 1*, which adds 0.5 to the score
- if a counterpart set includes at least one IR counterpart (i.e. WISE or IRAC1 or IRAC2) it is said to satisfy *rule 2*, which also adds 1 to the score
- if a counterpart set has a ratio of its best probability *greater than 10 times* with respect to all other counterpart sets of the same source (including for verification those rejected), it is said to satisfy *rule 3*, which also adds 1 to the score

Of course there are possibilities of counterpart sets having an unit probability ratio. This occurs because the best probability is the same, which may either happen because all probaiblities are undefined, or because the counterpart sets are intrinsically ambiguous (e.g. the same IRAC2 counterpart associated to two WIRcam counterparts). These cases are duly flagged.

Then one verifies if the counterpart set flagged selected has a score greater or equal than all other possible counterpart sets of the same X-ray source (case flagged "PLUS"), or less (case flagged "minus") or is "solitary" (the X-ray source has no other counterpart sets).

In the second script ("ambiguous") only, one considers also the score difference between the selected counterpart set and all other "valid" (`rank=2`) counterpart sets for the same X-ray source.

The idea is that for the cases when an X-ray source is *not unambiguous*, i.e. it has more than one counterpart set with a non-rejected rank, the best one is flagged `rank=0` if it is definitely better than all others (which are only nominally possible secondaries), while it is flagged `rank=1` if it is only nominally, and perhaps marginally, better in terms of probability of (at least) one of the secondaries (`rank=2`).

The distinction between ranks 0 and 1 has no such specific distinction for unambiguous cases. The `autorank` keeps track of all the previous procedures, and is also used to flag ambiguous and unambiguous cases. For ambiguous cases the autorank covers the range from 0 to 3. For unambiguous cases instead the range from 10 to 13. Therefore:

- a counterpart set with `rank=0, autorank`≥ 10 is a single good quality counterpart for its X-ray source
- a counterpart set with `rank=1, autorank`≥ 10 is a single counterpart of somewhat lesser quality
- a counterpart set with `rank=0, autorank`< 10 is the preferred and highly likely counterpart for its X-ray source, which however can have other nominal secondaries with `rank=2`
- a counterpart set with `rank=1, autorank`< 10 is the nominally preferred counterpart for its X-ray source, which however can have other secondaries of which one comparable to the preferred one

3.5.5. Postprocessing

In the `2XLSSOPTd` case one more script was enough to conclude the work. This script has also been generalized in a template script, but it has been found it was insufficient, so a second (new) template script was added after manual inspection of the results.

The original template `resetranks.newsql` should have verified the above condition, and in particular reset ranks 0 to 1 or viceversa for ambiguous cases.

However it has been found that it did not ensure that each X-ray sources has at least one and only one counterpart set with `rank between 0 and 1`. There was a limited number of X-ray sources where the best counterpart still was ranked 2. An adhoc template script `finalranks.newsql` was written to adjust those cases to the convention described above at the end of 3.5.4.

The statistics of the associations after all previous steps are presented in section 4.4.

4. Result summary

The XMM-LSS-like incremental procedure (see 3.4) "multiplies" the entries in the GCT. E.g. one starts with 14134 `XXLN` X-ray sources, and just after the correlation with the CFHTLS W1 (first member) table one has 33370 entries (counterpart sets). 11837 X-ray sources have 31073 counterparts (31040 distinct) and 2297 have none. For `XXLS` with 11863 X-ray sources, the correlation with the first member, sBCS, gives 23015 counterpart sets, 10176 X-ray sources have 21328 counterparts (21289 distinct) and 1687 have none.

The number of counterpart sets gradually increases adding more member tables, and for the northern area one reaches 58710 counterpart sets (of which 301 unidentified) once one has processed all tables in SD (which we recall are processed first).

This figure can be compared with the number of entries in `XXLNOPT`, 55525 (of which 267 unidentified) including the correlation with "external" catalogues like the Marseille spectroscopic one and NED, SIMBAD and USNO.

The equivalent figures for the southern area are 51498 counterpart sets (of which 61 unidentified) after processing of all tables in SD.

This can be compared with `XXLSOPT` with just 28271 entries (1178 unidentified) including external catalogues. Here the order of magnitude is somewhat different, I sus-

pect this could be due to the different processing of BCS and/or the absence of DECam in the old `XXLSOPT`.

Anyhow the exercise did not end here, but continued with the processing of the older tables in Milan DB (the ones already used for `XXLNOPT` and `XXLSOPT`) and of the newer tables added to Milan DB since September. This causes (somewhat unexpectedly) the addition of more counterpart sets, for a total of 71752 in the north, and 58645 in the south. The unidentified cases remain sort of stable (7 in the north, 51 in the south).

In the remainder the counterpart sets with a `seq` below 58710 (included, north) or 51498 (included, south) will be called ***new*** *(they have counterparts in at least one of Sotiria's tables and may or may not have counterparts in the Milan DB tables). The counterpart set with higher `seq` will be called* ***old only*** *i.e. they have counterparts only in (old or new) Milan DB tables.*

Using the database one can rather easily produce a *statistics of how many countepart sets include which combination of which catalogues*. An example for the southern area (which in the production phase will include 6 member tables from SD) is reported in Table 18. This table lists 61 possible (actual) combinations. One can see that only 1543 counterpart sets have counterparts in all 6 catalogues. The larger number of counterparts is 12026 in DECam only, etc. etc.

One could compare it with the equivalent Table 19 for `XXLSOPT` with 4 member tables in the old Milan DB set.

However an equivalent table for the northern area (14 member tables from SD) is already of prohibitive size, with 627 actual combinations. I could produce it, but it will occupy several pages. Old `XXLNOPT` with 8 members would have 130 combinations (barely at the limit of sensible representation).

Tables giving the possible counterparts in both the new (Sotiria's) and old (and new Milan DB tables) will be of prohibitive size. `testxos` would have 1551 combinations for 17 members, while in the north `testxon` even 6718 combinations for 28 members !

It is perhaps more productive to compare new (Sotiria) and old (Milan DB) tables of related origin one at a time. The comparison lists the cases of counterparts in two related tables (e.g. Sotiria's `nWIRcam` and Milan DB `wircam`) which appear to be associated (in the same counterpart set) to the same X-ray source. This means they will both be within $6''$ from the X-ray position, and could either be closer between them than the correlation radius given in Tables 4 or 5, or farther, but closer to some other counterpart.

In the northern area

- For WIRcam (Table 6) one has that most of SD counterparts are confirmed in old Milan DB, but a good fraction of old Milan DB are not present in SD WIRcam table although most of them are associated to other new entries. *Reason* ? different processing ?
- The situation for SDSS (Table 7) is even worse. The objects which are in Milan DB are more than those confirmed in SD. *Reason* ? different processing or different version ?
- The situation for UKIDSS (Table 8) is sort of similar to WIRcam. *Reason* ? different processing or different version ?
- The situation for WISE (Table 9) is more complex. There are objects confirmed in both, but even more present only in Milan DB, and some present only in SD. *Reason* ? different processing or different version ?
- The situation for IRAC (Table 10) is complicated by the fact SD includes two tables (3.6 and 4.5μm selected) which Milan DB includes only the latter. Nevertheless there is a good fraction of confirmed cases but also many present only in one catalogue. *Reason* ? different processing or different version ?
- The situation for GALEX (Table 11) is also variegated, and similar to most of the cases above. Note that in Milan DB only cases not flagged as potential tiling artifact were considered. *Reason* ? different processing or different version ? or tiling handling ?
- The case of CFHTLS (Table 12) is considered last, because it is complicated by the fact SD includes separate tables for the ABC fields (which were duly reprocessed but presumably without overlap removal) while in Milan DB a single table `w1t7` includes all of W1 as well as ABC.

In the southern area

- The situation for BCS (Table 13) is complicated by the presence of two alternate analyses in Milan DB (`bcsru` and `bcslmu`. Anyhow there is the usual large number of confirmed cases and non negligible number of cases present in one table only. *Reason* ? different processing ?
- The situation for DECam (Table 14) shows the usual mix of confirmed and unconfirmed cases. *Reason* ? different processing ?
- The situation for WISE (Table 15) and GALEX (Table 17) is similar to the one for the northern area.
- For IRAC the situation is complicated by the usage of different bands and version. SD includes two tables (3.6 and 4.5μm selected) of undocumented version, while Milan DB included originally only the 4.5μm-selected from SSDF version 8, which has been recently superseded by the "private" version 9.

4.1. Kind of counterparts

Besides the general comparison presented in the tables of the section above, one can also do a different kind of comparison on the GCTs.

This involves generating what I call *unid*s (for "unique identifiers') or "partial *unid*s" for a coun-

nWIRcam	wircam	new or old	count	notes
	✓	old only	134	old only not Sotiria
		old only	12908	no WIRcam at all
✓	✓	new	13285	confirmed
	✓	new	2699	old only, other Sotiria
		new	42726	no WIRcam at all, other Sotiria

Table 6. Comparison of counterparts in Sotiria's `nWIRcam` table with original Milan DB `wircam` table. Northern area `testxon` GCT.

SDSS	sdssdr10	new or old	count	Notes
	✓	old only	4711	old only no Sotiria
		old only	8331	no SDSS at all
✓	✓	new	7502	confirmed
	✓	new	9501	old only other Sotiria
		new	41707	no SDSS at all other Sotiria

Table 7. Comparison of counterparts in Sotiria's `SDSS` table with original Milan DB `sdssdr10` table. Northern area `testxon` GCT.

UKIDSS	ukidssdr10	new or old	counts	Notes
	✓	old only	141	old only not in Sotiria
		old only	12901	no UKIDSS at all
✓	✓	new	8162	confirmed
	✓	new	2648	old only other Sotiria
		new	47900	no UKIDSS at all other Sotiria

Table 8. Comparison of counterparts in Sotiria's `UKIDSS` table with original Milan DB `ukidssdr10` table. Northern area `testxon` GCT.

terpart set. I.e. I take the `seq`s of all counterparts in a given counterpart set (or in a subset of member tables, hence "partial"), replace null values with a zero, and concatenate everything in a string like e.g. 207902:19923:14524:11702:46941:0:0:0:131655.

In particular I generated *unid*s for `north33`, `w1t7`, `cfhtlens`, `wircam`, `ukidssdr10`, `swiredr6`, `galexgr6`, `omsuss`, `marseillespec`, `ned`, `simbad`, `usno` in the northern area or `south33`, `bcsru`, `bcslmu`, `ssdf2v8`, `galexgr6`, `marseillespec`, `ned`, `simbad`, `usno` in the southern area, both from the original work version catalogues in Milan DB (`XXLNOPT` and `XXLSOPT`), and for the GCTs of the current exercise (`testxon` and `testxos`).

Otherwise said, I compare the counterparts in the *old Milan DB tables*, in both cases they are flagged "new" or "old only".

Of course (multiple) entries having *unid*s of the form north33:0:0:0:0:0:0:0:0, i.e. "only Sotiria" counterparts but none in Milan DB, have to be removed.

In the northern area one has that 49939 cases have some *unid* (41086 flagged new, 8853 flagged old only); most of them coincide (i.e. the two identification procedures are sort of reproducible), but 2986 old *unid*s from `XXLNOPT` are **not** present in `testxon`, 3539 new old-style *unid*s from `testxon` are **not** present in `XXLNOPT`, of which 2623 are flagged new (have Sotiria counterpart) and 916 as old only (no Sotiria counterpart).

In the southern area one has that 21347 cases have some *unid* (18444 flagged new, 2903 flagged old only); 4389 old *unid*s from `XXLSOPT` are **not** present in `testxos`, 1916 new old-style *unid*s from `testxos` are **not** present in `XXLSOPT`, of which 1570 are flagged new (have Sotiria counterpart) and 346 as old only (no Sotiria counterpart).

Although in general most *unid*s coincide (i.e. the two identification procedures are sort of reproducible), there is a non negligible number of discrepancies, which are too many for a detailed examination. I therefore inspected only a few cases by sample.

For instance in the northern area the first two "new" cases are 200057:355:0:2131389:0:0:0:0:0:0:0:0 and 200074:299:1742:0:0:0:1304:0:0:32:0:1565, and the first two "old only" cases are 200057:355:0:0:0:0:0:0:0:0:0:0 and 200074:299:1742:0:0:0:0:0:0:32:0:1565.

They look as "crossed" cases, referring to the same X-ray source.

In the first example, `XXLNOPT` had nothing associated with `w1t7=355`, while the current procedure has a WIRcam object at $0.67''$. In this case the difference is due to the procedure (according to Table 4 the radius to correlate W1 and WIRcam should be $1.0''$, while the old procedure used $0.5''$).

In the second example, the current procedure generates *a single counterpart set* for the X-ray source 200074, while `XXLNOPT` had two. The new entry is just the *merger* of the two old ones. `galexgr6=1304` is at $1.59''$ from `w1t7=299`, i.e. above the $1.5''$ radius of Table 4. The same above-threshold distance occurs between the corresponding objects in SD tables (`GALEX=87` and `CFHTLS_W1=310`), but `GALEX=87` is at $1.4''$ from `SDSS=47`, i.e. SDSS "pulls in" `galexgr6` too !

Corresponding examples for the southern area are instead independent.

nWISE	wise	new or old	count	Notes
	✓	old only	3850	only old not Sotiria
		old only	9192	no WISE at all
✓	✓	new	7086	confirmed
✓		new	451	only Sotiria no old wise
	✓	new	7794	only wise but other in Sotiria
		new	43379	no WISE at all but other in Sotiria

Table 9. Comparison of counterparts in Sotiria's `nWISE` table with original Milan DB `wise` table. Northern area `testxon` GCT.

IRAC1	IRAC2	irac2v0	new or old	count	Notes
		✓	old only	4392	old only not in Sotiria
			old only	8650	no IRAC at all
✓	✓	✓	new	13028	confirmed (also in ch1)
✓		✓	new	97	old unconfirmed in ch2 but in ch1 ?
✓			new	1762	only Sotiria ch1
	✓	✓	new	1427	confirmed in ch2 only
		✓	new	6913	old only but in no other IRAC Sotiria

Table 10. Comparison of counterparts in Sotiria's `IRAC2` (and `IRAC1` too) tables with original Milan DB `irac2v0` table. Northern area `testxon` GCT.

GALEX	galexgr6	new or old	count	Notes
	✓	old only	2153	only old not in Sotiria
		old only	10889	no GALEX at all
✓	✓	new	8822	confirmed
✓		new	9248	only Sotiria not in old
	✓	new	502	no GALEX in Sotiria but in old and other Sotiria
		new	40138	no GALEX at all but other Sotiria

Table 11. Comparison of counterparts in Sotiria's `GALEX` with original Milan DB `galexgr6` table. Northern area `testxon` GCT.

200047:931:164:78:0:0:80270:0:153798 is a first "new" example. Another one is 200105:1001:0:174:0:0:0:0:0. They both happen to be the merger of two `XXLSOPT` counterpart sets "pulled in" by some of SD objects.

207902:19923:14524:11702:0:0:0:0: is a first "old only" example which has a new correspondent of the form 207902:19923:14524:11702:0:0:0:0:131655, i.e. has also a USNO counterpart. In `XXLSOPT` `usno=131655` was associated to a different BCS object `bcsru=19922` but here the USNO object is catched first by a Sotiria `sGALEX` object (not in `galexgr6`).

200325:598:476:466:42262:0:0:0:0 is another "old only" case which associates `bcsru=598` with `bcslmu=476`. In the new procedure the same `bcsru` associates to a different `bcslmu=475`. The reason is that `bcsru` 599 and 598 both associate with a `sBCS`, but `bcslmu=475` associates also with a `sDECam` which does not associate with `sBCS` (note the association `bcslmu` with `decam` is univocal but with `sBCS` is not !).

All this is indicative of the caution one should take when associating a large number of tables among them.

4.2. Distance statistics

A counterpart set can contain counterparts from equivalent tables in SD and Milan DB (e.g. BCS and `bcsru`, GALEX and `galexgr6`, etc.). These counterparts can be the result of a *direct association* (they are both associated within 6″ with the same X-ray source, and associated among themselves within the smaller correlation radius, which is 0.5″ for tables of same origin), but they might also be associated indirectly (associated with same X-ray source, and associated to some other counterpart within the radii in Tables 4 and 5).

We report the distance histograms (in semilog scale) in Fig. 3 and 4. This is of course relevant to counterpart sets flagged "new" (i.e. with counterparts both in SD and Milan DB).

The histograms have usually a sharp peak at less than 0.1″. We report also the percentage of entries within the direct association radius of 0.5″. Distributions which are fully or strongly peaked means probably the same data is used in SD and Milan DB. Regular distributions probably means reprocessing was done in SD, or different versions were used, but results, as far as position is concerned, are anyhow well compatible with Milan DB. Tails and residual peaks may either indicate indirect associations (crowded areas ?) or peculiarities (for instance I doubt that the GALEX tails indicate residual non removed tiling artifacts (in SD ?).

The most compatible cases are in the north CFHTLS W1, UKIDSS, WIRcam and IRAC2, in the south IRAC2 with SSDF V9, hinting to a common version used. The WA, WB, WC cases seems to indicate a slight position offset with respect to the old analysis.

W1	WA	WB	WC	w1t7	w1t7(ABC)	new or old	count	notes
				✓		old only	183	old only flagged W1
					✓	old only	569	old only flagged ABC
						old only	12290	no CFHTLS at all
✓				✓		new	31073	confirmed W1
	✓	✓			✓	new	24	confirmed WA and WB
	✓	✓				new	2	Sotiria AB only
	✓				✓	new	582	confirmed WA
	✓					new	286	Sotiria A only
		✓			✓	new	817	confirmed WB
		✓				new	363	Sotiria B only
			✓		✓	new	140	confirmed WC
			✓			new	98	Sotiria C only
				✓		new	6586	old W1 only other Sotiria
					✓	new	1074	old ABC only other Sotiria
						new	17665	no CFHTLS at all other Sotiria

Table 12. Comparison of counterparts in Sotiria's CFHTLS W1, WA, WB and WC with original Milan DB `w1t7` table (which includes also ABC which can be differentiated using the `tile` column. Northern area `testxon` GCT.

sBCS	bscru	bcslmu	new or old	count	Notes
	✓	✓	old only	63	this lot is
	✓		old only	126	in bcsru and/or bcslmu
		✓	old only	1406	but not in Sotiria BCS
			old only	7207	not BCS nor old neither new
✓	✓	✓	new	10033	Sotiria BCS confirmed both bcsru bcslmu
✓	✓		new	1314	ditto confirmed only bcsru
✓		✓	new	1452	ditto confirmed only bcslmu
✓			new	8529	ditto unconfirmed
	✓	✓	new	1936	somehow present in bcsru/bcslmu
	✓		new	687	but not in Sotiria BCS
		✓	new	784	however in some other Sotiria's table
			new	26763	not BCS

Table 13. Comparison of counterparts in Sotiria's `sBCS` with original Milan DB `bcsru` and `bcslmu` tables. Southern area `testxos` GCT.

sDECam	decam	new or old	count	notes
	✓	old only	4330	only in old not in Sotiria
		old only	2817	not in Sotiria no DECam at all
✓	✓	new	22547	confirmed
✓		new	6353	only Sotiria no old decam
	✓	new	3256	only decam but other in Sotiria
		new	19342	no DECam at all but other in Sotiria

Table 14. Comparison of counterparts in Sotiria's `sDECam` with original Milan DB `decam` table. Southern area `testxos` GCT.

4.3. Magnitude comparison

Another possible comparison between associated counterparts in equivalent tables in the same counterpart set concerns the *reference magnitudes* (and potentially all other magnitudes).

Fig. 5 and 7 plot the magnitude in Sotiria's tables vs the equivalent one in the corresponding Milan DB table. In most cases these figures do not show enough detail, with just some more or less large scatter around the locus of apparent equal magnitude. Only in cases like UKIDSS, WIRcam, IRAC and WISE there is clear evidence of a *large systematic offset*, presumably due to the yet undocumented standard corrections applied by Sotiria.

However such an offset, even if of smaller measure, is present also in other cases, and can be appreciated in the alternate representation of Fig. 6 and 8, plotting the *difference* between Sotiria's and Milan magnitude.

In all plots black crosses correspond to associated counterparts which are distant less than 0.5″ (i.e. presumably direct associations, see previous section 4.2). They may concentrate in a narrow strip for tables which were presumably taken from the same data source, or show some more or less large scatter for those where yet undocumented reprocessing (other than standard corrections) was presumably applied in SD.

In all plots red diamonds correspond to counterparts associated with a distance larger than 0.5″, i.e. presumably indirect and possibly spurious ones. In fact they usually show a larger scatter.

One important thing shown by those figures for the CFHTLS W1 case is a major misunderstanding concerning the i magnitudes in SD. I know that CFHTLS used the i′ filter for most of its earliest exposures, and that this was later replaced by the so-called y filter. In Milan DB table `w1t7` *uses a single column to store both magnitudes, and a column* `ugriz`*, which can assume values* `ugriz` *or* `ugryz` *to tell which is which (on tile basis). However table* `CFHTLS_W1` *in SD does not preserve any*

sWISE	wise	new or old	count	notes
	✓	old only	1641	only in old not in Sotiria
		old only	5506	not in Sotiria no WISE at all
✓	✓	new	5635	confirmed
✓		new	421	only Sotiria no old wise
	✓	new	3181	only wise but other in Sotiria
		new	42261	no WISE at all but other in Sotiria

Table 15. Comparison of counterparts in Sotiria's `sWISE` with original Milan DB `wise` table. Southern area `testxos` GCT.

sIRAC1	sIRAC2	ssdf2v8	ssdf2v9	new or old	count	notes
		1	1	old only	71	in SSDFV8 or also V9
		1		old only	44	but not Sotiria
			1	old only	1531	in SSDFV9 not in Sotiria
				old only	5401	no SSDF nor old neither new
1	1	1	1	new	12353	confirmed in all catalogues
1	1		1	new	1585	confirmed all but V8
1		1	1	new	21	
1		1		new	31	
1			1	new	58	
1				new	5498	only Sotiria ch 1
	1	1	1	new	936	confirmed in all ch 2 catalogues
	1		1	new	5474	only Sotiria ch 2
		1	1	new	1178	not in Sotiria
		1		new	120	but in some
			1	new	221	of the SSDF versions
				new	24023	not SSDF

Table 16. Comparison of counterparts in Sotiria's IRAC tables (both bands) with original Milan DB SSDF table (both versions 8 and 9, but only for the 4.5 μm-selected data. Southern area `testxos` GCT.

sGALEX	galexgr6	new or old	count	notes
	✓	old only	795	only in old not in Sotiria
		old only	6351	not in Sotiria no GALEX at all
✓	✓	new	4895	confirmed
✓		new	1491	only Sotiria
	✓	new	335	only galexgr6 but other in Sotiria
		new	44777	no GALEX at all but other in Sotiria

Table 17. Comparison of counterparts in Sotiria's `sGALEX` ith original Milan DB `wise` table. Southern area `testxos` GCT.

tile information, and for each object it has two columns `i_old_MAG_AUTO` and `i_new_MAG_AUTO`, only one of which is defined at a time (the other assumes the undefined (-99) value. I assumed `i_old` referred to i' and `i_new` to y, but it appears it is the other way round !!

The preliminary draft ended here. The next sections are added in this final version.

4.4. Probability ranking

The results of the procedures described in 3.5 are presented here.

Probabilities (see 3.5.1) were computed for 14 "Sotiria's" member tables in the north, and 6 in the south. The sources flagged "old only" in section 4 (i.e. with counteparts only in Milan DB tables) obtain by construction undefined probabilities and are ignored in the remainder of this report.

So we start with 58710 counterpart sets for 14134 X-ray sources in the north, and 51498 for 11863 X-ray-sources in the south.

At the end of the entire procedure we remain in the north with 29810 non-rejected counterpart sets, of which 15676 `rank=2` secondaries. Of the 13134 preferred counterpart sets, 6159 are singles (3331 rank 0, and 2828 rank 1), 4032 are reliable (rank 0) preferred counterpart with nominal secondaries, and 3642 are intrinsically ambiguous cases (nominally preferred to possible secondaries). 301 are blank fields (no nominal catalogued counterpart).

I stress that the so-called blank fields (or unidentified sources) shall be visually inspected, because they just report the lack of catalogued counterparts, but sometimes very bright (saturated ?) sources clearly visible in an image are missed in photometric catalogues !

In the south we remain with 18991 non-rejected counterpart sets (7128 secondaries). Of 11963 preferred counterpart sets, 7182 are singles (3864 rank 0 and 3318 rank 1), 2373 are reliable rank 0 with nominal secondaries, 2247 are intrinsically ambiguous and 61 are blank fields.

Considering intermediate results in the probability step, in the north 96% of all CFHTLS W1 counterparts have a valid (not undefined) probability *considering it has been computed with the same coefficients for both the old and new i' magnitudes*, 96% of the D1, 93% of WA, 90% of WB and WC, more than 99% of SDSS, UDS, WIRcam,

nctps	BCS	DECam	WISE	IRAC1	IRAC2	GALEX	count
6	✓	✓	✓	✓	✓	✓	1543
5	✓	✓	✓	✓	✓		2181
5	✓	✓	✓	✓		✓	4
5	✓	✓	✓		✓	✓	6
5	✓	✓		✓	✓	✓	670
5	✓		✓	✓	✓	✓	199
5		✓	✓	✓	✓	✓	217
4	✓	✓	✓	✓			14
4	✓	✓	✓		✓		30
4	✓	✓	✓			✓	4
4	✓	✓		✓	✓		3980
4	✓	✓		✓		✓	33
4	✓	✓			✓	✓	41
4	✓		✓	✓	✓		318
4	✓		✓		✓	✓	1
4	✓			✓	✓	✓	83
4		✓	✓	✓	✓		370
4		✓	✓		✓	✓	3
4		✓		✓	✓	✓	50
4			✓	✓	✓	✓	140
3	✓	✓	✓				16
3	✓	✓		✓			521
3	✓	✓			✓		459
3	✓	✓				✓	172
3	✓		✓	✓			8
3	✓		✓		✓		3
3	✓		✓			✓	4
3	✓			✓	✓		718
3	✓			✓		✓	3
3	✓				✓	✓	15
3		✓	✓	✓			5
3		✓	✓		✓		6
3		✓	✓			✓	3
3		✓		✓	✓		2114
3		✓		✓		✓	3
3		✓			✓	✓	9
3			✓	✓	✓		206
3			✓	✓		✓	5
3			✓		✓	✓	35
3				✓	✓	✓	31
2	✓	✓					2940
2	✓		✓				18
2	✓			✓			196
2	✓				✓		242
2	✓					✓	172
2		✓	✓				8
2		✓		✓			653
2		✓			✓		725
2		✓				✓	94
2			✓	✓			11
2			✓		✓		54
2			✓			✓	145
2				✓	✓		1118
2				✓		✓	30
2					✓	✓	25
1	✓						6734
1		✓					12026
1			✓				499
1				✓			4122
1					✓		4756
1						✓	2646
none							61

Table 18. Statistics of the identification in the southern area in the current exercise (`testxos` GCT). The last column gives the number of **new** counterpart sets with as many non-null counterparts in Sotiria's tables as given in the first column. Non-null entries are marked by a checkmark.

96% of UKIDSS, 75% of VIDEO, more than 99% of WISE, IRAC1 and IRAC2 and 93% of GALEX.
55937 counterpart sets (95%) have at least one probability defined, 27818 have at least one "fair" (< 0.03) and 18988 have at least one "good" (< 0.01).

In the south 90% of BCS have a valid probability, but only 36% of DECam, and also more than 99% of WISE, IRAC1 and IRAC2 and 93% of GALEX.
41130 (80%) have at least one probability defined, 16820 have at least one "fair", and 12109 have at least one "good". The lack of DECam counterparts is presumably due to the fact (according to the `decam` table in Milan DB) about 25% of the sources are in *gz* tiles not in *griz* tiles.

The numerology of the pre-ranking step (see 3.5.2) has been inspected and found plausible, but is not of general

nctps	BCS Rutgers	BCS LMU	SSDF 2V8	GALEX GR6	count
4	✓	✓	✓	✓	7018
3	✓	✓	✓		5081
3	✓	✓		✓	204
3	✓		✓	✓	475
3		✓	✓	✓	164
2	✓	✓			2272
2	✓		✓		1169
2	✓			✓	59
2		✓	✓		1026
2		✓		✓	98
2			✓	✓	84
1	✓				828
1		✓			2845
1			✓		3045
1				✓	2421
none					1183

Table 19. Statistics of the identification in the southern area in the previous `XXLSOPT` catalogue. Conventions as in Table 18, but for the old tables in Milan DB.

interest, except to note that at this stage 22587 counterpart sets are rejected in the northern area, and 17829 in the south.

At this stage I attempted an analysis of *intrinsic ambiguities among counterparts*, like checking cases having e.g. optical-optical distance (in different catalogues) greater than the logged threshold for the correlation radii (reported in Tables 4 and 5), and corresponding to counterpart sets which, for a given X-ray source, have a multiplicity greater than one in one of the member catalogues. However while a reduced number of intrinsic ambiguities of the latter type was found in an inspection, they generally did not appear to exceed the thresholds. I preferred therefore not to apply any pre-filtering on such cases, and let them show up in a future visual inspection.

The statistics of the ranking step (see 3.5.3 and 3.5.4) for the northern area gives for the unambiguous cases 928 solitaries, 4192 cases flagged PLUS and just 24 flagged minus, and for the ambiguous cases 7546 PLUS and 31 minus.

In the southern area one has for the unambiguous 512 solitaries, 5591 PLUS and just 36 minus; for the ambiguous cases 4531 PLUS and 34 minus.

The post-processing (see 3.5.5) original rank resetting affects 94 rank 0 and 234 rank 1 in the north and 634 rank 0 in the south, but it does not ensure that each X-ray source has at least one (and just one) preferred counterpart set with rank 0-1.

This instigated adding the final rank setting step, which handles the residual case with only rank 2 counterparts (988 in the north and 1159 in the south). Most of them are actually singles (with all other counterpart sets `rank=-1` rejected). In a few cases (97 in the north and 55 in the south) there are multiple rank 2 entries for the same X-ray source, but it is easy to spot the best one using `autorank`. All these cases are promoted to `rank=1`.

This results in the final statistics reported at the beginning of the present subsection, and in the next one.

4.5. Summary statistics

We report here also some summary statistics mainly on the preferred counterparts (rank 0-1). I consider here the presence of counterparts grouped grossly by band as in 3.5.2, i.e. optical, IR, NIR and UV (IR applies only to the northern area).

In the northern area of 14134 X-ray sources, only 2941 have a rank 0-1 counterpart in all 4 coarse bands (of these 2733 have a not-undefined probability in all bands). 12133 have an optical countepart, 8228 a NIR one, 7480 both optical and NIR, 10367 have an IR counterpart, 9394 both optical and IR, 6957 both NIR and IR, and just 5386 a GALEX one.

In the southern area of 11863 X-ray sources, only 2374 have a preferred counterpart in all **3** coarse bands (of which 2091 with not-undefined probability in all bands). 9743 have an optical counterpart, 10807 an IR one, 8991 both optical and IR, and just 2927 a GALEX one.

Tables 20 and 21 report the breakdown of the preferred counterparts per coarse band, both in absolute numbers and in percentage. In the north more than 20% of the counterpart sets are in all four bands, and 25% in all bands except UV. In the south (no NIR band) 20% of the counterpart sets are in all three bands, and moret han 55% in all bands except UV.

The statistics for *single* counterparts (not shown) are similar, while for rank 2 secondaries (also not shown) there is a prevalence of GALEX-only (29% in the north, and 37% in the south), and, in the south of optical-IR (39%) or optical-only (23%). Note that the percentages refer to the number of distinct X-ray sources with at least one rank 2 counterpart (7674 in the north and 4620 oin the south), but there is a cfertain degree of multiplicity with more than one secondary, the total number of secondaries is 15676 in the north and 7128 in the south.

5. Future work

Possible future activities can be grouped in two areas: one are the improvements to the files in SD, necessary for a

optical	NIR	IR	UV	count	percent
				301	2.1%
			✓	414	2.9%
		✓		434	3.1%
		✓	✓	104	0.7%
	✓			296	2.1%
	✓		✓	17	0.1%
	✓	✓		385	2.7%
	✓	✓	✓	50	0.4%
✓				1590	11.2%
✓			✓	191	1.4%
✓		✓		1395	9.9%
✓		✓	✓	1477	10.4%
✓	✓			766	5.4%
✓	✓		✓	192	1.4%
✓	✓	✓		3581	25.3%
✓	✓	✓	✓	2941	20.8%

Table 20. Statistics of the preferred (`rank between 0 and 1`) counterparts in the northern area. Optical refers to at least one of CFHTLS W1, D1, WA, WB, WC, SDSS or UDS; NIR to at least one of WIRcam, UKIDSS or VIDEO; IR to one of WISE, IRAC1 or IRAC2; and UV to GALEX. Both the absolute number of entries and the percentage (100%=14134 X-ray sources) are shown.

optical	IR	UV	count	percent
			61	0.5%
		✓	243	2.0%
	✓		1590	13.4%
	✓	✓	226	1.9%
✓			668	5.6%
✓		✓	84	0.7%
✓	✓		6617	55.8%
✓	✓	✓	2374	20.0%

Table 21. Statistics of the preferred (`rank between 0 and 1`) counterparts in the southern area. Optical refers to at least one of BCS or DECam; IR to one of WISE, IRAC1 or IRAC2; and UV to GALEX. Both the absolute number of entries and the percentage (100%=11863 X-ray sources) are shown.

production reprocessing probably with any method, not just with or in the Milan DB; another one are possible tuning to the current XMM-LS-like procedure.

I stress however that such procedure has never been intended as giving definitive results, but just as an initial working tool, to be followed either by further ranking and validation based on science considerations (e.g. X-ray fits, SEDs etc.) and/or by visual inspection and validation.

5.1. Recommendations

The suggested improvements to a next release of SD are summarized as follows:

- Provide detailed documentation (see 2.1)
- Consider more datasets (OM-SUSS, SWIRE etc. see Table 1) and updated datasets (see red text in 1)
- Pre-merge IRAC1 and IRAC2 tables
- Pre-merge (handle overlaps) W1 with WA/WB/WC
- Make sure tile overlap is handled for any other case (GALEX etc.)
- Clarify reasons for differences with tables currently in Milan DB (specially on position, see 4, and id's, see 2.5.1)
- Clarify what is `i_old` and `i_new` (see 4.3 at end)
- Choose table names to avoid clashes with Milan DB and include version reference (see 2.2)
- Drop table name prefix from colum names, make column names all in same case (lower, upper, mixed), and provide units and description for columns in FITS header (see 2.3)
- Match identifiers with those used in public websites (and in Milan DB), see list in 2.5.1.
- Include additional columns (see list in 2.5.2 and possibly also original magnitudes along with standard ones ?)

5.2. Tuning options

Possible tuning might occur in the following areas.

- The correlation radii in Tables 4 and 5 might be adjusted (made smaller ? specially for GALEX ?) case by case.
- One could consider to handle all possible magnitude bands (and not just a preferred one) in probability computation. Or (also to solve the cases with undefined magnitudes) to replace undefined magnitudes with the tile limiting magnitude (as done in table `w1t7`).
- One might also consider possible more detailed fits to the density $n(brighter\ than\ m)$ (e.g. twin power law fits, different fitting ranges, etc.).
- One might also consider the use of *capped probabilities* (e.g. if the distance to the X-ray source is less than $2''$, one might fix the distance to $2''$, to avoid overweighting sources closer than a typical X-ray position uncertainty).
- Finally one could adjust all the recipes used in the various ranking steps.

Acknowledgements. I acknowledge the copious work done by S.Fotopoulou to prepare the datasets distributed in February 2014 and used as primary input for this work. I acknowledge also earlier correspondence with N.Fourmanoit, and recent correspondence with M.Pierre and S.Paltani.

References

Chiappetti, L., 2013, The XXLN and XXLS catalogues, Preliminary release for internal use, XMM-LSS Internal Report N. 12-Mi (Report XII)

Chiappetti, L., Clerc, N., Pacaud, F., et al. 2013, MNRAS, 429, 1652

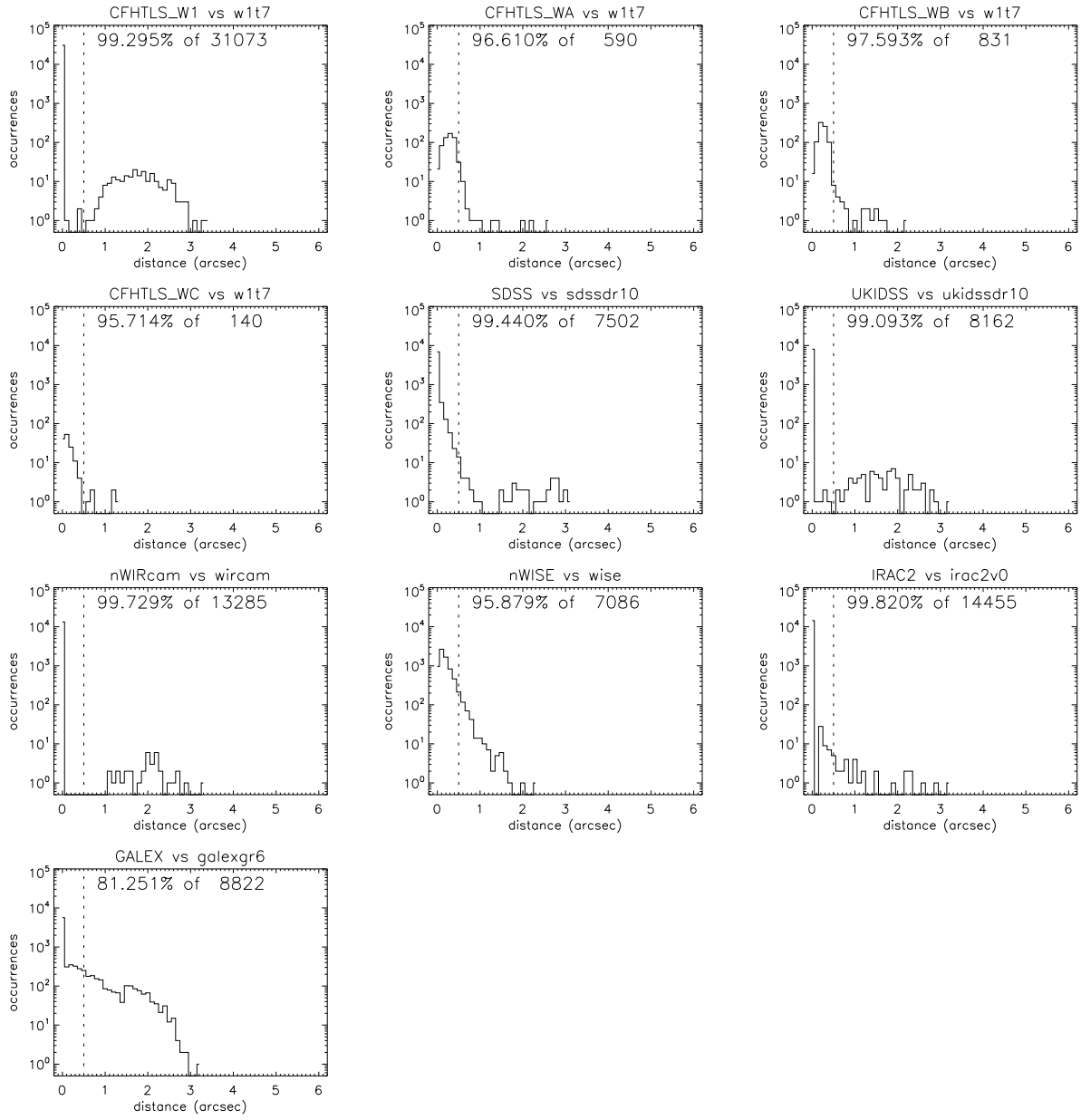

Fig. 3. Distance between counterparts in the same counterpart set taken from equivalent tables in SD and Milan DB for the northern area. The dashed vertical line corresponds to the $0.5''$ threshold used for direct association.

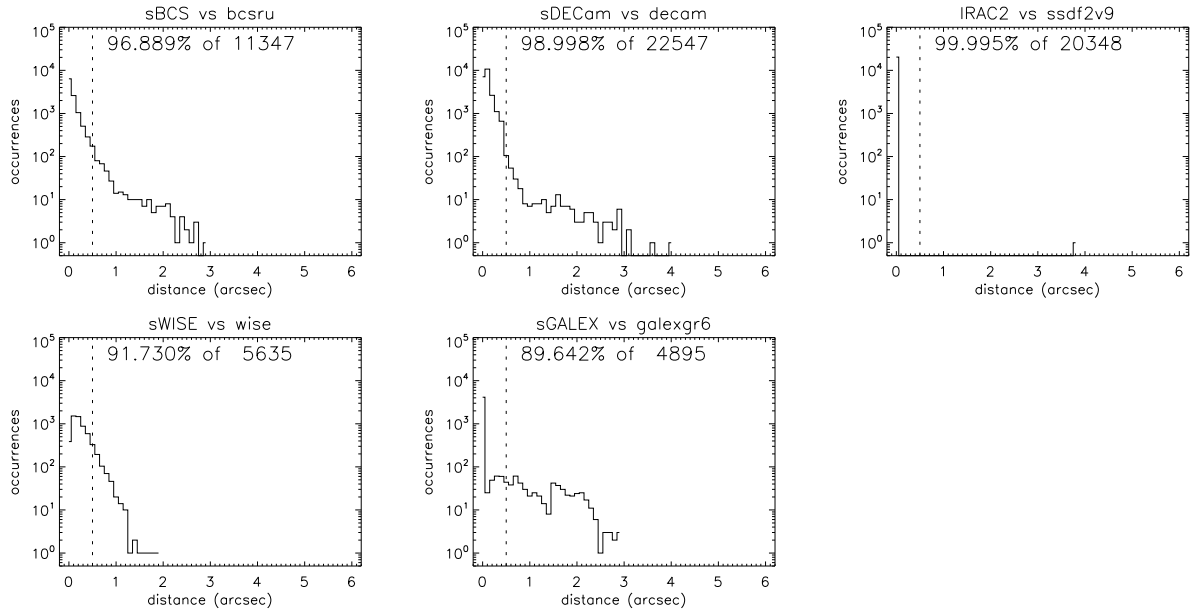

Fig. 4. Distance between counterparts in the same counterpart set taken from equivalent tables in SD and Milan DB for the southern area. Same conventions as for Fig. 3.

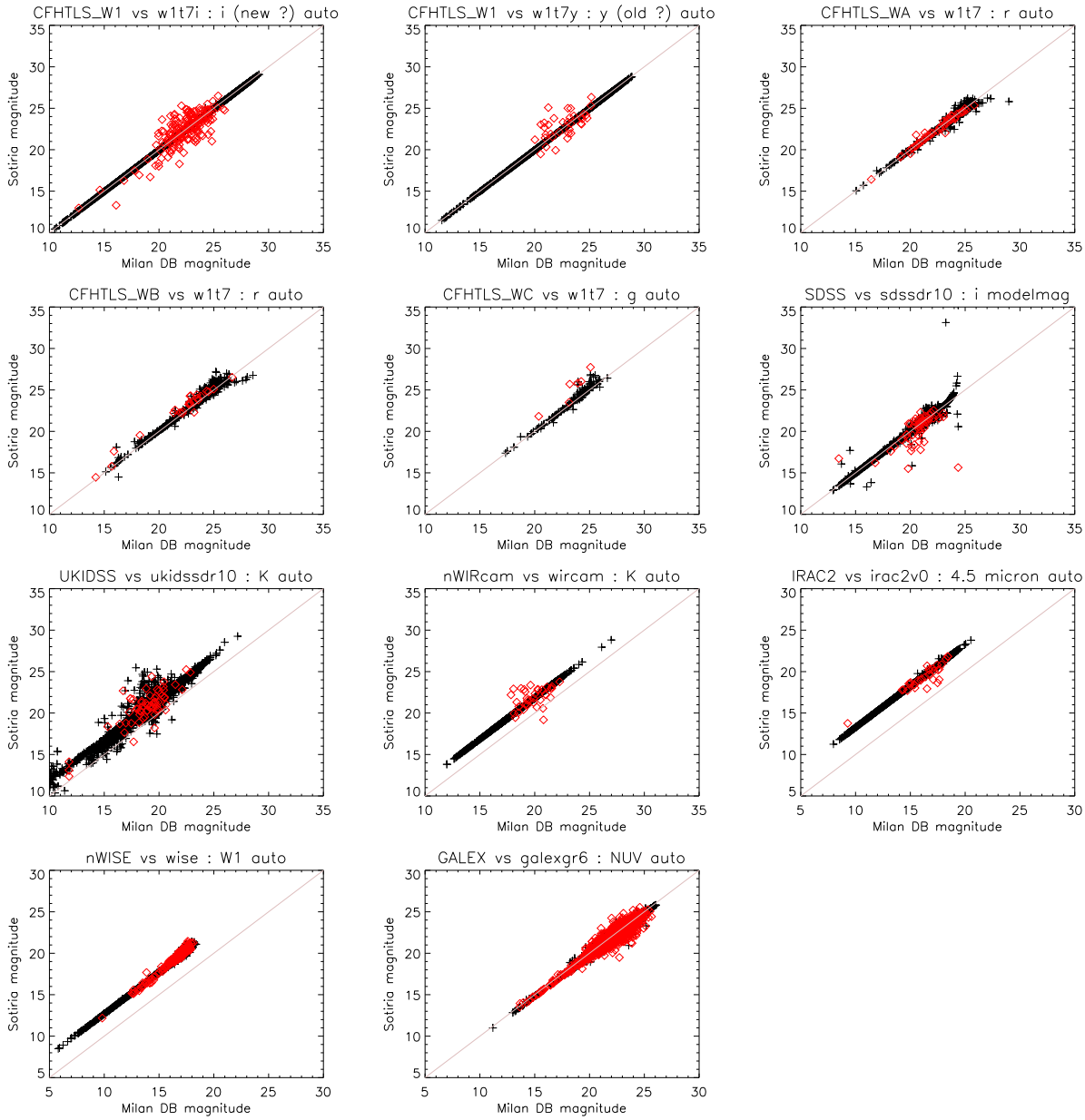

Fig. 5. Comparison between the magnitudes for counterparts in the same counterpart set taken from equivalent tables in SD and Milan DB for the northern area. The diagonal pink line is the locus of equal magnitude. Black crosses correspond to direct associations or anyhow associations within a distance of 0.5″. Red diamonds correspond to associations with a larger distance.

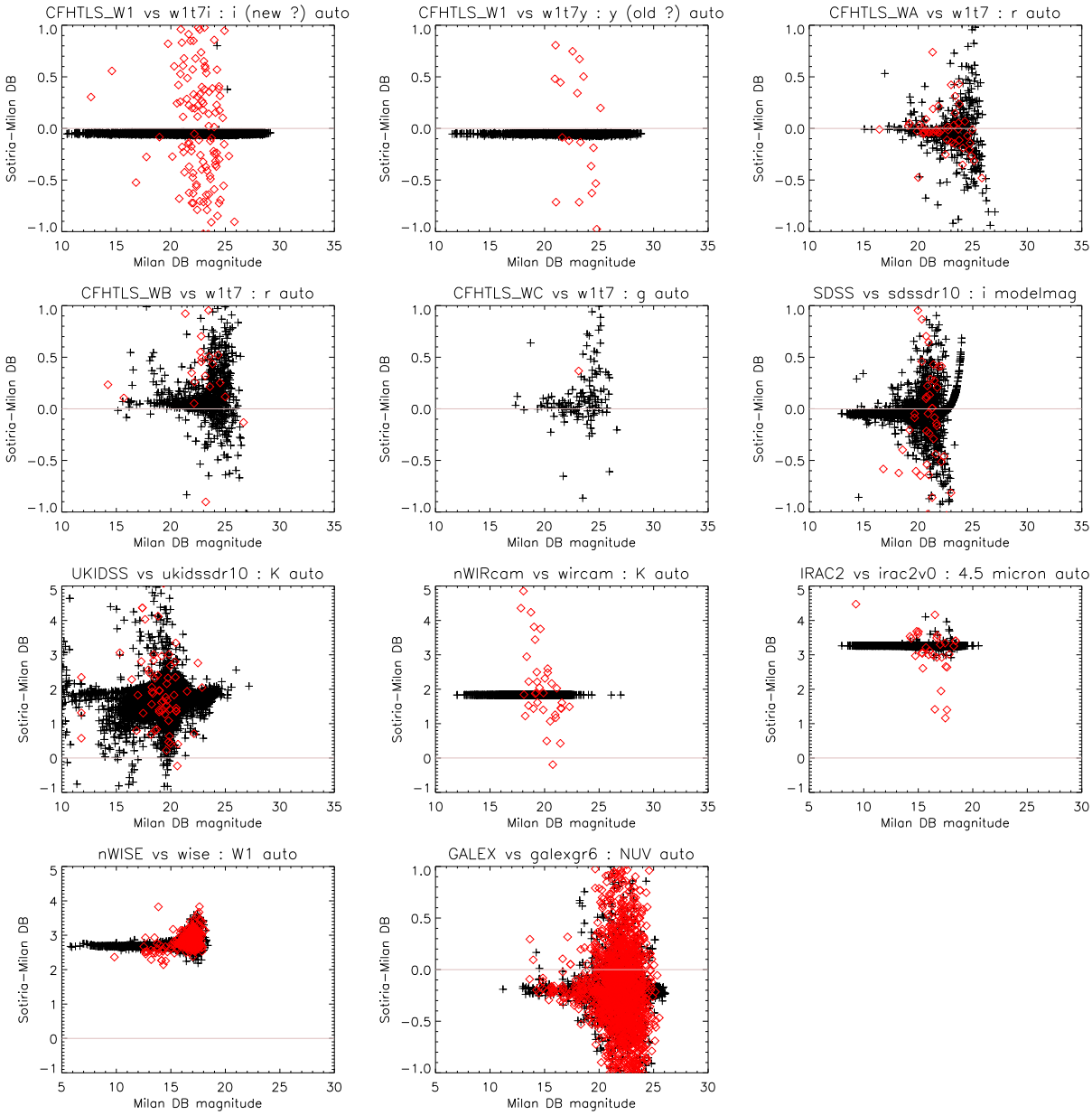

Fig. 6. Alternate comparison between the magnitudes for counterparts in the same counterpart set taken from equivalent tables in SD and Milan DB for the northern area, plotting the magnitude difference vs the magnitude in Milan DB. The horizontal pink line is the locus of equal magnitude. A vertical offset indicates some sort of standard corrections applied in SD. Symbols and colour conventions as for Fig. 5.

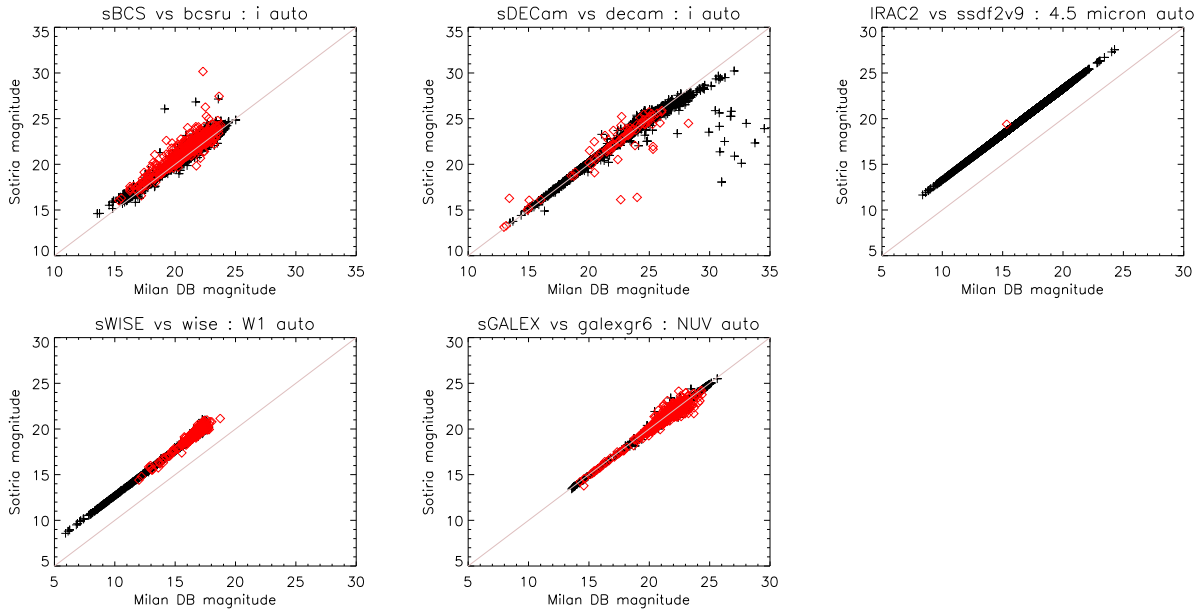

Fig. 7. Comparison between the magnitudes for counterparts in the same counterpart set taken from equivalent tables in SD and Milan DB for the southern area. Same conventions as for Fig. 5.

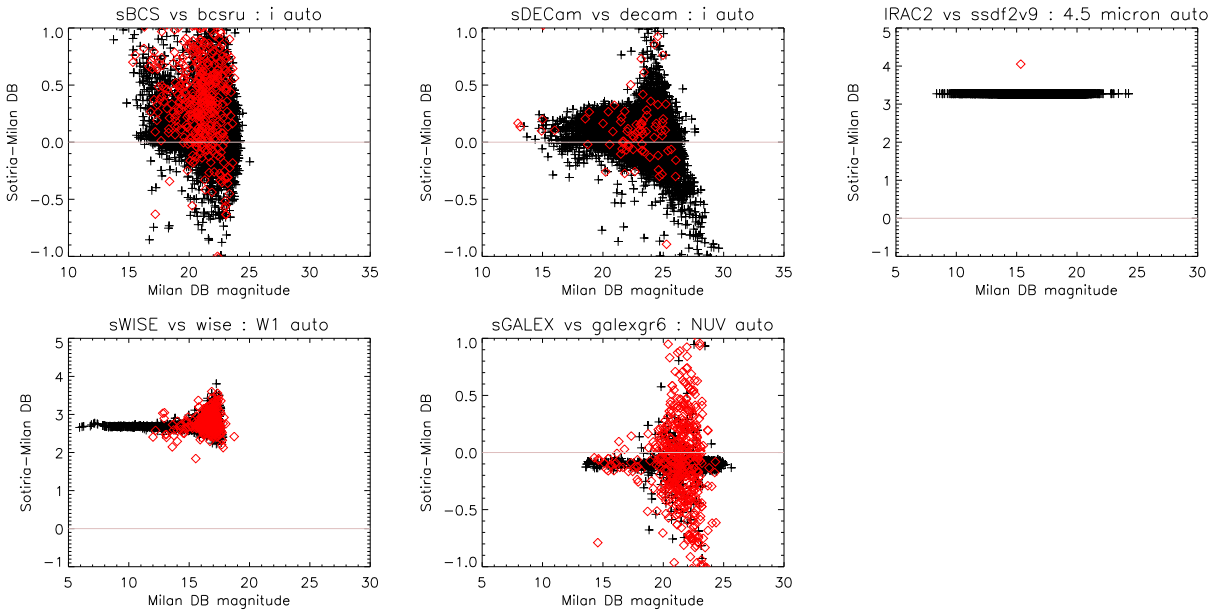

Fig. 8. Alternate comparison between the magnitudes for counterparts in the same counterpart set taken from equivalent tables in SD and Milan DB for the southern area, plotting the magnitude difference vs the magnitude in Milan DB. Same conventions as for Fig. 6.