XMM Survey Science Centre
University of Cambridge
Cambridge, CB3 OHA, UK

SSC-CAM-TN-0003
draft 1.2
2001 Feb 16

Suzanne Aigrain, Richard G. McMahon (rgm@ast.cam.ac.uk)

Likelihood ratio method for SSC XID work

Issue	Author	Date	Comment
1.1	S. Aigrain	2001 January 28	Draft Issue
1.2	S. Aigrain	2001 February 16	Added paragraph on Rutledge et al. Reliability calculation

Printed February 20, 2001 at 10:53

1 Summary

This document outlines the likelihood ratio technique for cross identification between different wave-bands as described in available literature and its application to SCC XID work.

The WWW pages containing the work are in:

http://www.ast.cam.ac.uk/~xmmssc/xidops/

Contents

2 Introduction

XMM source lists are to be matched with source lists from other instruments, including INT WFC, APMCAT and CIRSI. In this document, the matching of XMM lists with WFC data is described.

As the pipelines for processing XMM data are still being finalised, the XMM source lists used have varying formats, although this is expected to be resolved in the near future. In some cases systematic position offsets became apparent. These were corrected using an initial comparison with WFC data, after which the matching process was repeated. Some XMM source lists contained individual position errors, some not.

Both unmerged (*i.e.* band by band) and merged WFC data was used. The matching was performed using merged source lists to enable, for example, colour–colour or colour magnitude plots of matched optical sources to be produced. However the likelihood ratio analysis presented below was performed using objects detected in the i–band only (usually the deepest, and also available for all fields for which any WFC data was available).

Quality control plots for WFC data can be found at the following location:

http://www.ast.cam.ac.uk/~xmmssc/xid/data_status.php

For each field, the small gif image in the 'plots' column, if green, leads to the corresponding quality control plots. These are independent of any X–ray information and publicly accessible.

3 Initial matching on the basis of positional coincidence only

For each XMM source in a given field, the nearest WFC object was found using a rigorous great circle distance calculation and was declared a match if within a specified search radius.

Histograms of total, RA and DEC offsets out to a larger radius (30"), offset scatter plots and spatial distribution plots of XMM sources and offsets overlaid with the optical or IR image boundaries were produced. These are available in the following directory for each field and source_list:

*http://www.ast.cam.ac.uk/~xmmssc/xidops/***field/source_list***/html/match/*

(see document SSC–CAM–TN–0002 for a full description of the directory structure in use for SSC XID work).

Fig. 1 is a typical example of the results. A Gaussian was fitted to each RA and Dec offset histogram and the optimal search radius was taken to be 3σ. For XMM / WFC or XMM / CIRSI matching this optimal radius was found to be 5 arcsecs, for XMM / APM matching 6 arcsecs. σ is significantly reduced if only objects detected in the i–band are included, which prompted us to perform the likelihood ratio analysis on the unmerged WFC i–band data (see section 5).

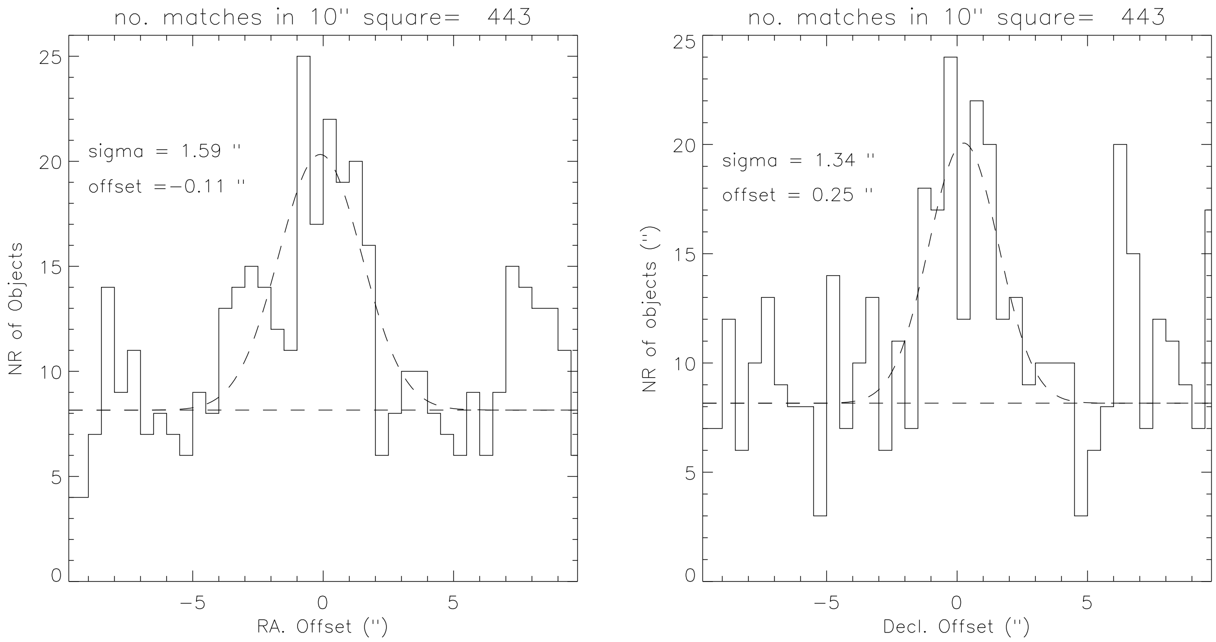

Figure 1: Histograms of a) RA offset and b) Dec offset in arcsecs for the Lockman Hole field XMM data matched with merged WFC data. The horizontal dashed line shows the background level.

4 Calculating likelihood ratio and reliability

Cross–identification of source lists in different wave-bands using maximum likelihood techniques has been widely used, and the definitions for the quantities evaluated vary. In this section we present the definitions we chose to use, and put forward the reasoning behind such choices in the light of previous work and the characteristics of our dataset.

4.1 Notation

The following notation is used throughout this section and the next:

- XMM objects are referred to as 'sources' and the index i runs from 1 to N_x where N_x is the number of XMM sources in the sample under consideration.
- Optical (or Infrared) objects within the search radius of an XMM source are referred to as 'candidates' and the index j runs from 1 to N_i where N_i is the number of candidates for the i^{th} XMM source.
- The position errors for in our sample are much larger for the XMM data (a few arcsecs) than for the optical or Infrared data, thus in the following we will only include the XMM position errors in the calculations. Strictly speaking they should be added in quadrature with the optical or infrared errors. We will also assume for simplicity that RA and Dec (or x and y) errors are the same, and we give the label σ_i to the position error of the i^{th} XMM source in either direction. When an RMS position error is defined for the whole sample it is referred to as σ.

- $r_{i,j}$ is the normalised distance between the i^{th} XMM source and the j^{th} candidate for that source in arcsecs, *i.e.* $r_{i,j} = d_{i,j}/\sigma_i$ The actual distance $d_{i,j}$ is often taken as the offsets along 2 perpendicular axes added in quadrature, although in our calculations rigorous great circle distances are computed.

- r_s is the search radius, *i.e.* the maximum value of $d_{i,j}$ for an object to be considered as a candidate for that source.

- $p(A|B)$ represents the probability of event A given event B. Hence $p(r_{i,j}|id)$ is the probability that the j^{th} candidate, found at distance $r_{i,j}$ from the i^{th} XMM source, is a true identification for this source. Similarly $p(r_{i,j}|c)$ is the probability that the same candidate is a chance candidate, *i.e. not* the true identification. On the other hand, $P(id|r_{i,j})$ is the probability that a true identification exists at distance $r_{i,j}$ from the i^{th} XMM source. Note that in the probabilities mentioned here, in the cases where magnitude as well as separation is taken into account, we should use (for example) $p(r_{i,j}, M_j|id)$ (where M_j is the magnitude of the j^{th} candidate) rather than $p(r_{i,j}|id)$.

- The background density of optical or infrared objects is sometimes defined independent of magnitude, in which case it is referred to as ρ. However in some cases it is desirable to take into account the magnitude of the optical or infrared candidate as a bright X–ray source is likely to have a bright counterpart. Thus we define $\rho(< M)$ as the background density of optical or infrared objects with magnitude less than or equal to M. When considering a candidate labeled j, $\rho(< M_j)$ is used.

- The likelihood ratio for a given candidate is referred to as $LR_{i,j}$.

- The reliability for a given candidate is referred to as $R_{i,j}$.

4.2 Definition of likelihood ratio

The likelihood ratio for candidate j for a given source i is (de Ruiter et al. 1977 [2],Prestage & Peacock 1983 [4], Wolstencroft et al. 1986 [5],Irwin 1996 [7]):

$$LR_{i,j} = \frac{dp(r_{i,j}|id)}{dp(r_{i,j}|c)} \tag{1}$$

with the general expressions

$$dp(r|id) = r\exp(r^2/2) \tag{2}$$

and

$$dp(r|c) = 2\pi\sigma^2\rho r dr = 2\lambda r dr \tag{3}$$

such that, as in [4]

$$LR_{i,j} = \frac{\exp(r_{i,j}^2/2)}{2\lambda} \tag{4}$$

Note that a number of authors (Benn 1983 [3], [5], Sutherland & Saunders 1992 [6]) have included the factor Q (*a priori* probability that an identification exists above the plate limit) in the expressions for $dp(r|id)$ and hence $LR_{i,j}$. How ever this factor Q is unknown and has to be guessed, which goes against the idea that a likelyhood ratio should be derived from the data and make no *a priori* assumptions. For this reason we chose no to include it, although it (or rather $p(id)$, which is very similar to it) appears in the definition of reliability (see 4.3

4.2.1 How to estimate Q

Chosing a value for the prior Q is necessary at some point in the process of evaluating the identifications. Even using equation refeq:lhr, the value of $p(id)$ is required to estimate the reliability $p(id|r)$. In this section we present some methods used by the authors who chose to include Q in the expression for the likelyhood ratio.

Benn [3] uses an iterative method to obtain an estimate for Q, based on the fact that

$$\sum_j p(id|r_{i,j}) = Q_i$$

such that

$$\sum_i \sum_j p(id|r_{i,j}) =< Q > \times N_x$$

Having calculated the left–hand side of the above from equations 2 and 3, Benn iterates over non–zero values of Q until both sides converge. Estimating Q through iteration is one method we may investigate in the future.

Wolstencroft et al. [5] have put forward a method for estimating Q which is quite different. They take $Q = Q_1 Q_2$, where Q_1 is the overall probability that an identification exists for the ensemble of all sources in the sample, and Q_2 is defined to account for the fact that a bright candidate is more likely to be the true identification than a fainter one:

$$Q_{2_{i,j}} = \frac{\int_{-\infty}^{M_j} n(M')dM'}{\int_{-\infty}^{M_{\mathrm{lim}}} n(M')dM'} \tag{5}$$

where M_j is the magnitude of the j^{th} candidate, $n(M)$ is the number of candidates with magnitude in the range M to $M + dM$, and M_{lim} is the plate limit of the optical sample.

We have preferred to include information on the magnitude of the candidate only in the value taken for ρ, the background density, as in Rutledge et al. 2000 [9].

4.2.2 Chance candidates and background density

The expression for the probability that a given candidate is *not* the true identification (equation 3) is not followed exactly be all authors: de Ruiter et al. [2] replace it by

$$dp(r|c) = 2\lambda r \exp(-\lambda r^2) dr$$

However, this approach is rejected by subsequent papers [5, 6] because it relies on the assumption that the closest candidate is always the true identification, which is not the can when the background density is high.

The need to consider the magnitude of the candidate in estimating the background density is widely recognised [5, 6, 9] and in equation 3 we will use $\rho(< M_j)$, the density of background objects with magnitude less than or equal to that of the candidate.

4.3 Reliability

Reliability is usually meant as the probability that the true identification is the candidate being considered [3, 6, 9], *i.e.* $R_{i,j} = p(id|r_{i,j}, M_{i,j})$, and this is the definition we will take.

The simplest estimate for $R_{i,j}$ is [3]:

$$R_{i,j} = p(id|r_{i,j}) = \frac{p(r_{i,j}|id)p(id)}{p(r_{i,j}|id)p(id) + p(r_{i,j}|c)p(c)} \tag{6}$$

from Bayes' theorem.

To calculate this quantity in the first instance we will assume $p(id) = 1$, and later work will aim to improve this first estimate in ways outlined in section 6.

However, the above expression ignores any information about the fact that there may be more that one candidate for a given source, which is the case for all XMM sources due to the depth of the optical data. To overcome this limitation, Sutherland & Saunders [6] derive an alternative expression, which we will call $R'_{i,j}$:

$$R'_{i,j} = \frac{Q \times LR_{i,j}}{\sum_j (Q \times LR_{i,j}) + (1 - Q)} \tag{7}$$

(note this expression differs from that in the Sutherland and Saunders paper because we have opted for a different definition of likelyhood ratio.)

Another way to take into account the possibility of multiple candidates is proposed by Rutledge et al. [9]. The reliability they define is purely a function of the likelihood ratio of the candidate under consideration (and hence does not reflect the eventuality that another candidate may have a similar or higher likelihood ratio). However they go on to evaluate the probability that the candidate is the emphunique identification for the source , and the probability that there is no identification at all, in a separate step, leading to the expressions below.

The probability that there is no identification at all for the i^{th} source is given by:

$$P_{no-id,i} = \frac{\prod_{j=1}^{N_i}(1 - R_{i,j})}{S} \tag{8}$$

$$P_{id,i,j} = \left[\frac{R_i}{1 - R_i} \prod_{j=1}^{N_i} (1 - R_{i,j}) \right] / S \tag{9}$$

where S is a normaisation factor such that $P_{no-id,i} + \sum_{j=1}^{N_i} P_{id,i,j} = 1$.

We have carried out the calculation for both methods and it appears that the second one gives more contrasted results, while R' is nearly always very close to one for the closest or most likely match.

5 Results for XMM / WFC matching

5.1 Practical evaluation of likelihood ratio and reliability

When matching two source lists (on the basis of positional coincidence, as described in section 3), one runs through each XMM source and compares it with every WFC object in the field, computing the distance to it along great circles. RA and Dec offsets and magnitude are stored for all WFC candidates out to 50 arcsecs for each XMM source, and likelihood ratios are computed as described in the previous section. We then evaluate reliabilities according to equation 7.

This was done using WFC i–band source lists as this is usually the deepest band and data is consistently available for all fields with XMM data.

A database of cumulative magnitude counts was created for the WFC data: bins are 0.25 magnitude wide and contain the number of objects per square arcsec with magnitude less than or equal to the *upper* limit of the bin. The area cover by one WFC pointing (4 CCD's) is taken to be around 0.2 square degrees. This was then used to estimate the background density as a function of magnitude for each candidate,

5.2 Results for two example fields

The field with the most complete set of data, and the hence the most complete analysis, is the Lockman Hole field.

Fig.ĩeffig:tblo shows differential histograms of $p(r|id)$ and $p(r|c)$. XMM position errors were taken to be 2.0 arcsecs.

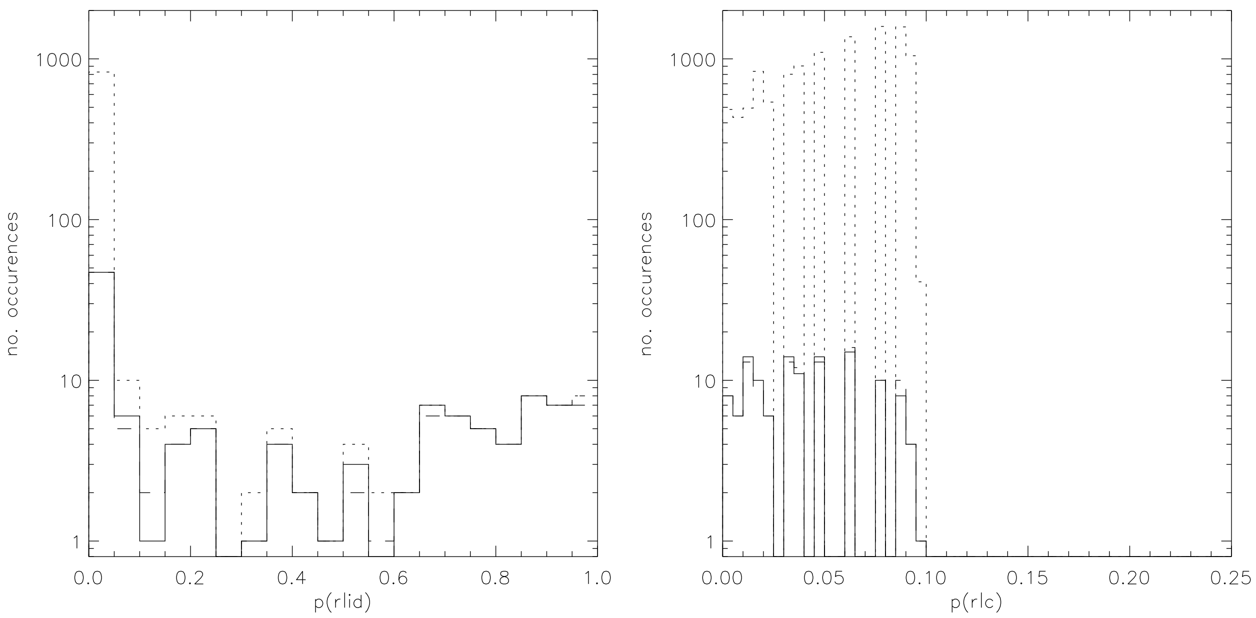

Figure 2: Histograms of a) $p(r|id)$ and b) $p(r|c)$ for the Lockman Hole field. Dotted line: all candidates out to 50 arcsecs. Solid line: most likely candidate for each source. Dashed line: closest candidate for each source.

The 'binning' effect in Fig.ĩeffig:tblo(b) is due to the finite width of the magnitude bins in the background density calculation.

Restricting the optical catalogue to objects brighter than a certain limit (say $M = 20$) removes a large fraction of the candidates with low likelihood ratio as expected.

Although in most cases (116 out of 121 sources) the closest object is the most likely, explaining why the solid and dashed curves mostly coincide in figures 2 and 3, the discrepancy is more apparent in the reliability histogram, which put an emphasis on the fact that there are more than one candidate per source. Fig. 4 shows the difference between the two methods of calculating reliability outlined in section 4.3. Which method should we choose? At the moment we are using the Sutherland & Saunders method but the other gives more contrasted results.

In this case the identification fraction with a search radius of 5 arcsecs was 0.74. Tables of all XMM sources in this field with corresponding WFC matches and the likelihood ratio and reliability for

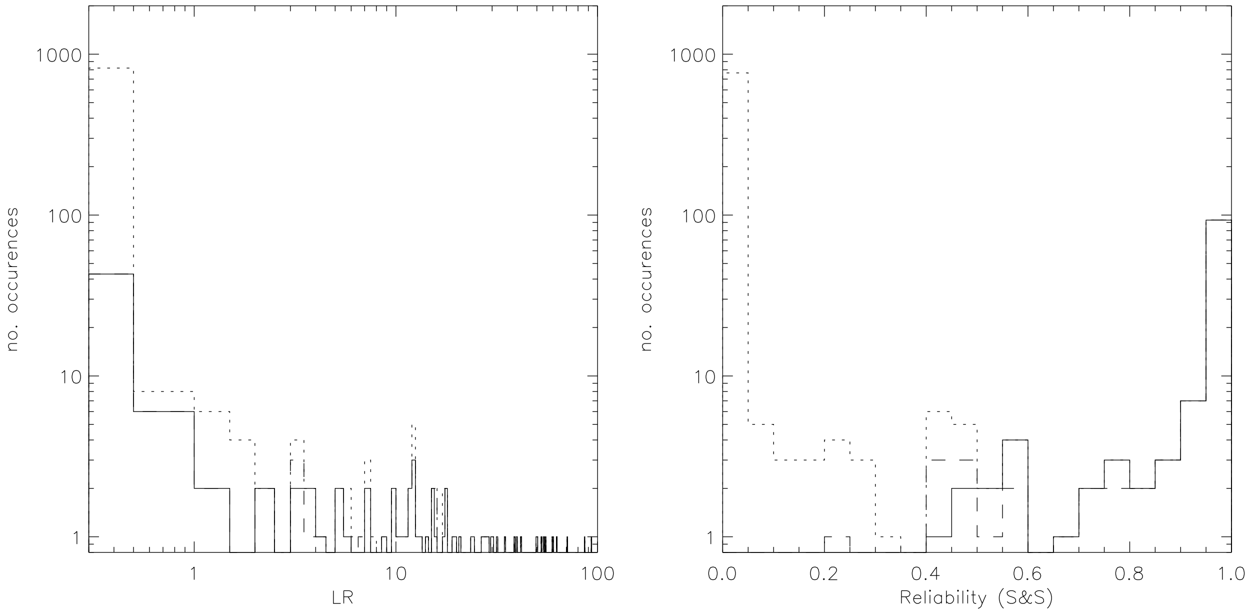

Figure 3: Histograms of a) likelihood ratio and b) reliability for the Lockman Hole field. Dotted line: all candidates out to 50 arcsecs. Solid line: most likely candidate for each source. Dashed line: closest candidate for each source.

these associations can be found here:

http://www.ast.cam.ac.uk/xmmssc/internal/lockman-hole/ lockman-hole/gh/html/

This page also contains links to colour–colour and flux–magnitude plots for the matched objects and more control plots.

5.3 Results for merged dataset

Histograms very produced for the merged dataset (*i.e.* all fields for which both XMM data and WFC i–band data were available) by calculating the likelihood ratios and reliabilities for each field and concatenating them.

The fields used were:

No.	Field	RA (arcsecs)	Dec (arcsecs)	BII (deg.)
1	Lockman Hole	10 52 40.80	+ 57 27 16.2	+ 53.15
2	SDS 1	02 17 60.00	− 05 00 00.0	− 59.75
3	PHL 1092	01 39 56.00	+ 06 19 21.0	− 54.55
4	G 133	01 04 00.00	− 06 42 00.0	− 69.35
5	MKN 205	12 21 44.00	+ 75 18 37.0	+ 41.67
6	A 399	02 58 25.00	+ 13 18 00.0	− 39.17
7	EQ PEG	23 31 50.00	+ 19 56 17.0	− 39.14
8	HCG 16	02 09 38.50	− 10 08 49.0	− 64.94
9	GRB 001025	08 36 36.50	− 13 04 30.0	+ 16.69

The role of reliability in exposing the few cases where the nearest object is not the most likely is

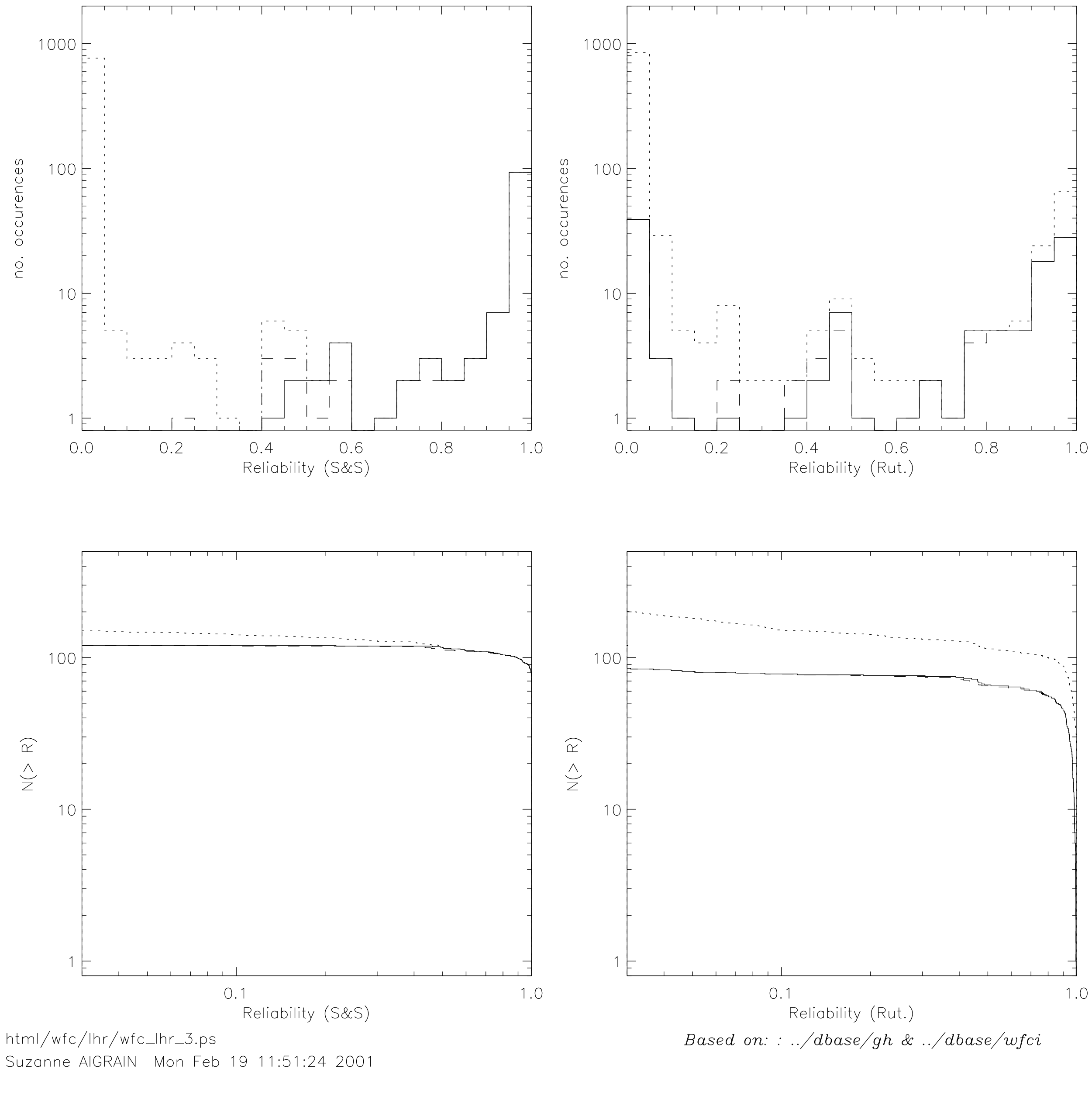

Figure 4: Histograms of reliability comparing the Sutherland & Saunders (S&S) and Rutledge et al. (Rut.) methods: (a) & (b) differential, (c) & (d) cumulative.

again visible in Fig. 5. Over the merged sample of 607 sources, this happened for 18 sources.

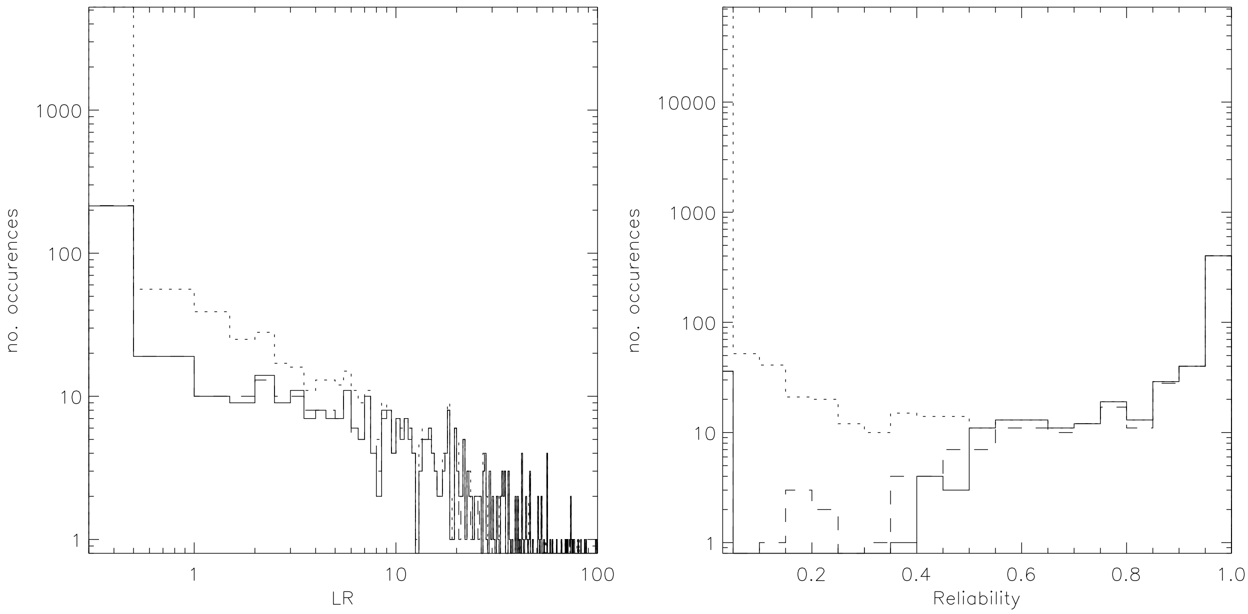

Figure 5: Histograms of a) likelihood ratio and b) reliability (S&S) for the Lockman Hole field. Dotted line: all candidates out to 50 arcsecs. Solid line: most likely candidate for each source. Dashed line: closest candidate for each source.

In Fig. 6 these cases have been highlighted. The definition of reliability provided by Sutherland & Saunders [6] provides a powerful tool for detecting instances where the positional coincidence method is insufficient.

Fig. 6(a) shows that the most likely matches have a separation of around 1 arcsec. However a significant number of associations with high likelihood ratio and reliability have a separation larger than 5 arcsecs, the optimal search radius identified using offset histograms, although they all lie within 15 arcsecs.

6 Discussion and future work

As seen above, in a small but significant number of cases the closest object is not the most likely. The quality of the matching would therefore be improved by taking likelihood ratio and Reliability into account in deciding which candidate is the true identification, rather than crudely taking the closest. It also allows us to make fuller use of the information available, as the magnitude of the candidates is also considered. The likelihood ratios and reliabilities should be calculated for all candidates within 15 arcsecs of an XMM sources (this proves sufficient to detect all best matches over the 9 fields available so far).

Before carrying this out, however, further investigation is needed into the following:

- The influence of position errors in the XMM data has not yet been investigated (position errors were set to 2.0 arcsecs to obtain the results presented above). In the case where individual errors are available for each source, these should be included. In the case where

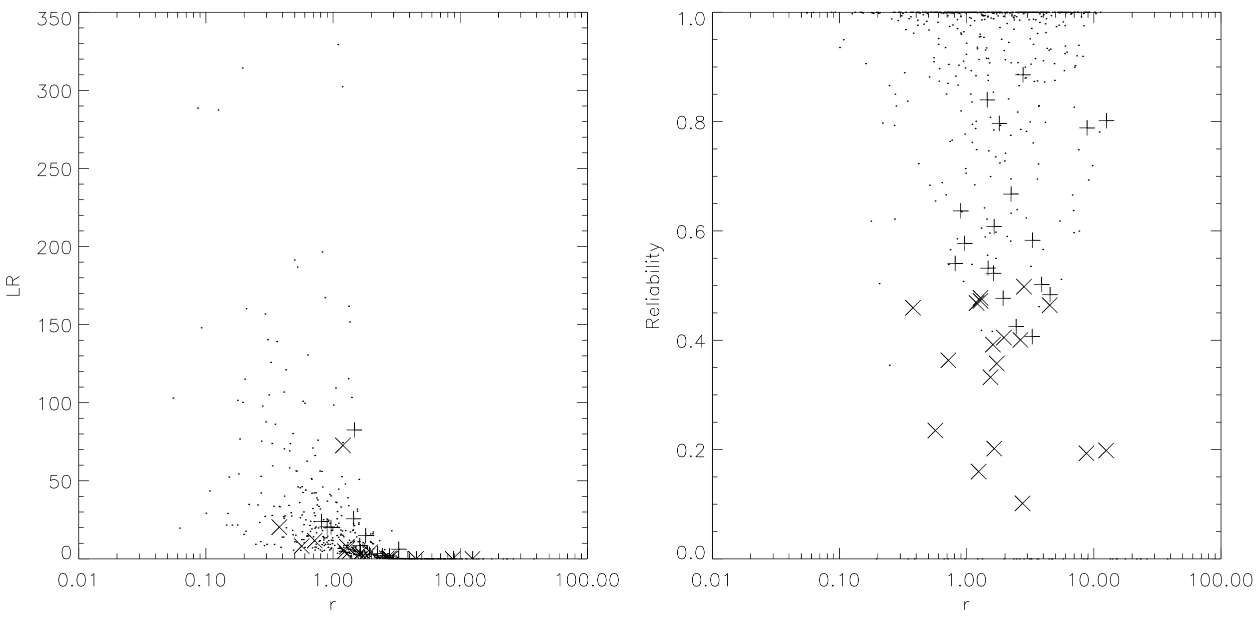

Figure 6: a) likelihood ratio and b) reliability (S&S) versus separation for the merged dataset. Dots: most likely candidate. +: most likely candidate if different from closest. $\times$: closest candidate if different from most likely.

one error value is assigned to he entire sample, one could vary this value to assess its impact on the likelihood ratios.

- A better estimate of Q or $p(id)$ is needed to compute the reliabilities, as we have simply been setting it to 1 so far, be it with the Sutherland & Saunders method or with the Rutledge et al. method. The iterative method employed by Benn [3] to estimate Q has been described in section 4.2.1, and a similar calculation could be carried out in this case.

The effect of possible offsets in the X–ray positions, or of confusion, can be important (as outlined by Hogg 2000 [8]).

A technique similar to that employed by Rutledge et al. [9] to determine the relationship between likelihood ratio and reliability could be used to calculate reliabilities from the likelyhood ratios without having to guess a value for $p(id)$. For each source field in the optical data, Rutledge et al. define a number of background fields which are circles of diameter 75 arcsecs offset from the centre of the source field by varying amounts to the east and west. The X–ray sources are then matched to optical sources in each field, and this is used to determine the effective relationship between reliability and likelihood ratio, which they then use to calculate the probability that a candidate is the *unique* identification for a given source (see section 4.3).

Similarly, we propose to generate a set of simulated random background optical objects for a region of a given size, to a certain magnitude limit, with a power law density. A set of X–ray sources can be separately generated for the same region together with the magnitude of their optical counterparts. The X–ray source positions are then shifted by a small offset which is a function of the *r.m.s.* error in the X–ray positions, to simulate confusion.

We would then have generated optical data for a 'source' field (using the true X–ray positions and the optical magnitudes of their counterparts) and for background fields, as well as shifted X–ray data for the same field. Performing the likelihood ratio calculation on these data will allow us to test the ability of this technique to pick out confusion effects and, perhaps, will identify a given method for choosing $p(id)$ as more suitable than another.

References

[1] J. J. Condon, T. J. Balonek, D. L. Jauncey. *Optical identifications of sources in the NRAO 5Ghz deep survey*, Astronomical Journal vol. 80, Nov. 1975, p. 887–894.

[2] H. R. de Ruiter, H. C. Arp, A. G. Willis. *A Westerbork 1415 MHz survey of background radio sources. II - Optical identifications with deep IIIA-J plates*, Astronomy & Astrophysics Supplement Series, vol. 28, May 1977, p. 211–293.

[3] C. R. Benn. *Optical identification of radio sources in the presence of optical confusion*, The Observatory, vol. 103, 1983, p. 150–154.

[4] R. M. Prestage, J. A. Peacock. *Optical identifications of Parkes radio sources using UK Schmidt plates*, Royal Astronomical Society Monthly Notices, vol. 204, July 1983, p. 355–364.

[5] R. D. Wolstencroft, A. Savage, R. G. Clowes, H. T. MacGillivray,S. K. Leggett, M. Kalafi. *The Identification of IRAS Point Sources - Part One - A 304-DEGREE Field Centred on the South Galactic Pole*, Royal Astronomical Society Monthly Notices, vol. 223, Nov. 1986, p. 279–302.

[6] W. Sutherland, W. Saunders *On the likelihood ratio for source identification*, Royal Astronomical Society Monthly Notices, vol. 259, Dec. 1992, p. 413–420.

[7] M. J. Irwin *Detectors and data analysis techniques for wide field optical imaging*, Lectures given at the 7th Canary Islands Winter School, 1996.

[8] D. W. Hogg *Confusion errors in astrometry and counterpart association*, submitted to Astrohpysical Journal March 14, 2000.

[9] R. E. Rutledge, R. J. Brunner, T. A. Prince, C. Lonsdale *XID: Cross-association of ROSAT/Bright source catalog X–ray sources with UNSO A-2 optical point sources*, The Astrophysical Journal Supplement Series, vol. 131, issue 1, Nov. 2000, p. 335–353