XMM Survey Science Center **SSC-AIP-TN-003**
Astrophysikalisches Institut Potsdam **10/08/2007**

Georg Lamer (glamer @ aip.de)

Extended source detection in the 2XMM catalogue pipeline

The detection and characterization of extended X-ray sources is one of the major changes to the 2XMM source detection pipeline with respect to the 1XMM version. The scope of this document is to explain the algorithm employed for extended source detection and to highlight some of the implications for the resulting catalogue of extended sources.

1 Detection pipeline

Source detection is applied to a set of usually 15 XMM EPIC images (images from 3 EPIC cameras in 5 energy bands each). The algorithm in the 2XMM pipeline involves the following steps:

SAS task	description
`eexpmap`	Create exposure maps for each of the 15 input images.
`emask`	Create detection masks (one for each camera) excluding areas near chip boundaries, dead pixels etc. based on the exposure maps.
`eboxdetect`	Box detection run using local background and 5×5 image pixels.
`esplinemap`	Create 15 background maps by fitting spline surfaces to images after excluding sources found in local eboxdetect run.
`eboxdetect`	Box detection (5×5 image pixel) run using the `esplinemap` maps as background. Both `eboxdetect` searches run simultaneously on all 15 input images.
`emldetect`	Estimate source parameters by fitting the PSF convolved with an extent model to each source from the second `eboxdetect` run. This run is applied simultaneously to all 15 images. See also section 1.1.
`emldetect`	Run PSF fitting on 3 images (one per camera) in the ("XID") band 0.5-4.5 keV. The source list of the first `emldetect` run is used as input list; positions and extent parameters are kept fixed.
`srcmatch`	Merges the two `emldetect` source lists and writes final source list with one row per source.

1.1 Emldetect PSF fitting

The PSF fitting by the task `emldetect` is the most crucial step for the characterization of extended sources. In the 2XMM pipeline the CAL 'medium accuracy' PSF is convolved with a β-model for the source extent:

$$f(x,y) = \left(1 + \frac{(x-x_0)^2 + (y-y_0)^2}{r_c^2}\right)^{-3\beta+1/2}$$

The β parameter is fixed at the canonical value $\beta = \frac{2}{3}$ for the surface brightness distribution of clusters of galaxies (Jones & Forman 1984). The upper limit of the fitting range for the extent parameter is set to 80 arcsec. The convolved extent model is fitted to each position from the input list on sub-images with 60 arcsec radius. The size of the sub-images and the maximum of the extent parameter are the result of a compromise between the necessity to fit large extended sources on the one hand and on the other hand the attempt to avoid point source confusion and spurious detections.

The fitting is performed on the sub-images of each of the 15 input images simultaneously. The two positional fit parameters x, y and the extent parameter r_c are constrained to be equal for all energy bands and cameras. Count rates are fitted to each sub-image, amounting to another 15 free fit parameter for each source.

The fitting procedure of `emldetect` minimizes the C - statistic

$$C = 2\sum_{i=1}^{N}(e_i - n_i \ln e_i)$$

to find the best set of model parameters, where e_i is the expected model value and n_i the measured number of counts in pixel i (Cash 1979).

The detection likelihood for each input image n is

$L_n = C_{null} - C_{best}$

where C_{null} is the C-statistic of the Null-hypothethis model (zero source flux) and C_{best} is the C-statistic of the best fitting model. It can be used to estimate the detection likelihood after correcting for the number of free parameters with the relation

$$L = -\ln(1 - P(\frac{\nu}{2}, L')) \quad \text{with} \quad L' = \sum_{n=1}^{N} L_i$$

where P is the incomplete Gamma function, N is the number of energy bands involved, ν is the number of degrees of freedom of the fit ($\nu = 3 + N$, if the source extent is a free fit parameter and $\nu = 2 + N$ otherwise).

The detection likelihood obeys the relationship

$L = -\ln(P)$

where P is the probability that the source is spurious due to a Poissonian fluctuation.

The extent likelihood L_{ext}, determining whether the extent of a source is significant, is calculated in an analogue fashion. Here, C_{null} is determined by the best fitting point source model, and C_{best} is derived from the best extended model. Again the probability P that the extent detection is spurious due to Poissonian fluctuation is described by

$L_{ext} = -\ln(P)$

For both point sources and extended sources a detection likelihood threshold $L_{min} = 6.0$ is set to include a source in the final catalogue. The threshold for the extent likelihood is set to $L_{ext,min} = 4.0$. If L_{ext} exceeds the threshold, the parameters of the extended model is written to the source list, otherwise the parameters of the best-fitting point source model. Also if the best fit value of the extent parameter r_c is less than 1.5 image pixels (6 arcsec), the point source model is selected.

Since the confusion of two or more neighbouring sources can cause spurious extent detection, a second fitting loop tests for the presence of a second source, when the first fit prefers the extended source model. Since fitting 2 source models simultaneously is very CPU-expensive, this step is only performed for brighter sources (if the `eboxdetect` likelihood of the input source exceeds 10.0). This second loop can either result in 2 point sources or in the combination of one point source and one extended source. If it improves the likelihood with respect to the single source fit, the 2 sources with their parameters are written to the output list.

The background model for the fitting process is taken from the maps produced by `esplinemap`. After each completed source fit the background model is updated by adding the best-fit source model. Since the brightest sources have the strongest effect on the background map, the input source list is processed in descending order of `eboxdetect` likelihood.

References:

Cash, W., 1979, ApJ, 228, 939

Jones, C., & Forman, W., 1984, ApJ, 276, 38

2 Quality of the catalogue of extended sources

2.1 Causes for spurious detections

The 2XMM catalogue contains more than 20000 entries of extended detections. However, with respect to point-like sources the detection and parameterization of extended sources is much stronger affected by systematic effects, which can be inherent to the detection algorithm or caused by uncertainties of the instrument calibration. Therefore the number of spurious extended sources is much higher than expected from Poissonian fluctuations only.

As the main reasons for spurious detections of extended sources the following causes have been identified:

- **Spurious detections near bright point sources:**
  The main cause for this problem is the deviation of the CAL PSF models from the true EPIC PSF. This leads to inaccurate modelling of the internal background by `emldetect` and hence causes spurious detections near bright sources. For the pre-release catalogue (2XMMp) it is estimated that $\sim$ 25% of the 6300 extended source detections are spurious due to this problem (see Fig.1).
- **Confusion of point sources:**
  This is usually a problem for faint sources, where pairs or multiples of close point sources are hard to distinguish from an extended source. Also multiples of more then two brighter sources can result in an extended detection, since only up to two point sources are modelled simultaneously by `emldetect`.
- **Insufficient background subtraction:**
  Some spatial variations of the background cannot be accurately modelled with the 12×12 spline fit employed by `esplinemap`. Examples for small scale background variations are the scattering features in MOS images, OOT features of bright (pile-up affected) sources in PN images, edges of noisy chips (MOS), or soft background features in PN images. In areas, where the true background is underestimated by the maps, spurious detections of extended sources are possible. In many cases the extent parameter of these sources is at or near the maximum of the allowed range (80 arcsec).
- **Multiple detections of extended sources:** The surface brightness distribution of extended sources is generally more complex than the β-model fitted by `emldetect`. Therefore `emldetect` tends to fit additional extended sources to the wings of brighter extended sources. The most extreme cases are observations of complex, bright extended sources (e.g. galactic supernova remnants), accounting for a large number of extended source detections.

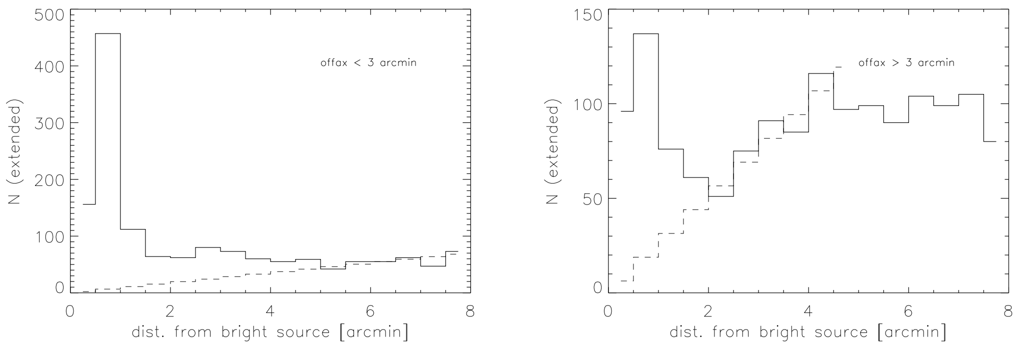

Figure 1: *left:* Extended sources close to bright ($> 10^4$ EPIC counts) on-axis point sources. The histogram shows the number of extended sources in 2XMMp as a function of distance from bright point sources. A number of 950 spurious extended sources is estimated from this histogram after subtracting the assumed "background" of genuine extended sources (dashed). *right:* Same for bright (> 5000 EPIC counts) off-axis (> 3 arcmin) point sources. The background of unaffected sources is determined by a linear fit to the histogram at larger distances. After subtracting the background of extended sources (dashed histogram), a number of 260 spurious sources is estimated.

2.2 Automatic source flagging

In order to identify sources, which are likely to be affected by the problems mentioned in section 2.1, a flagging scheme has been implemented in the 2XMM pipeline. The SAS task `dpssflag` uses information from the `emldetect` lists and writes 9 flags to the source list. These flags are the first 9 of the 12 character column `EP_FLAG` in the catalogue file.

The individual flags indicate the following conditions:

Flag	meaning	details
1	low detector coverage	`maskfrac` < 0.5
2	near other source	($d[''] < 65. \times \sqrt{(\texttt{ep\_rate})}$ `or` $d < 10''$) `and` $d < 400''$
3	within extended emisssion	$d < 3 \times$ `ep_extent` `and` $d < 200''$
4	possibly spurious due to bright source	flag #2 set and primary source has `ep_rate` $>$ 1000
5	possibly spurious in extended source	$d \leq 160''$ `and` rate $< 0.4\times$ rate of primary source
6	possibly spurious, most signal from one image only	`det_ml` of one band > 0.9 of total `det_ml`
7	possibly spurious extended source	At least one of the flags 4, 5, or 6 is set
8	On bright MOS1 corner or bright PN column	
9	Near bright MOS1 corner	

Flags 4-7 are written for extended sources only, these conditions indicate an increased probability that the source is spurious.

Figure 2 shows the distribution of the flag settings in the 2XMM catalogue for all extended source detections. For about 50% (9882/20837) of the extended detections at least one of the relevant flags 4-6 have been set.

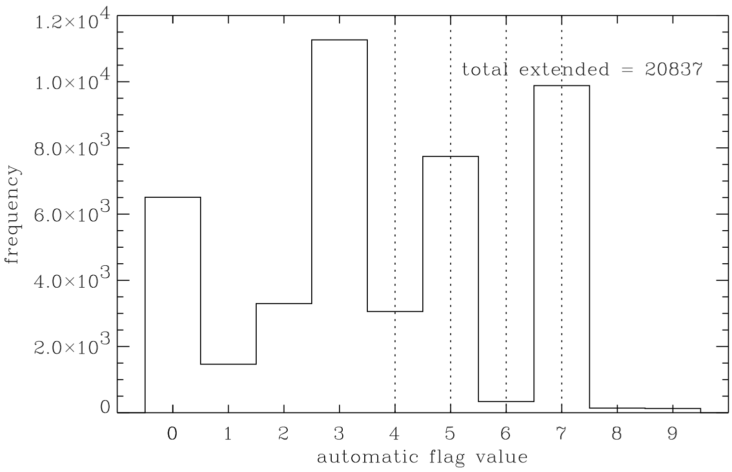

Figure 2: Distribution of automatic source flags for the extended source detections in 2XMM. The value '0' means all flags of a source are set to false, the values 1..9 correspond to the automatic flags as described in section 2.2. The dotted lines mark the flags 4-7. Flags 4-6 indicate conditions, in which the spurious detection of extended sources is likely. Flag 7, signifying that at least one of the flags 4-6 is 'true', is set in 9882 cases.

2.3 Results from 2XMMp

A fraction of the extended source detections in the 2XMMp catalogue (the pre-release version of 2XMM) has been inspected visually with the aim of verifying the existence and extent classification of each source. The inspected sample had been pre-selected according to galactic latitude ($|bII| > 20°$), off-axis angles (> 4 arcmin), and extent parameter ($r < 80''$). Fields with large extended emission (SNRs, bright clusters) and sources associated with the target of the observation have been removed. Therefore the results cannot be regarded as representative for the entire catalogue. However, the sample is rather typical for an extragalactic serendipity survey and is used here to verify the potential of the automatic flagging scheme for the elimination of spurious detections.

According to their visual appearance, the 669 sources have been grouped into the following catagories:

	visual category	total number	flag #7 set
a)	visually confirmed extended sources	494	25
b)	uncertain: not visually confirmed as extended source, but no obvious reason for spurious detection	74	2
c)	spurious due to confusion of several point sources	24	3
d)	spurious detection due to nearby (bright) point source	33	29
e)	multiple extended detection, source is part of another extended source	44	38
b-e)	likely spurious	175	72

Figure 3 shows the distribution of the automatic flag settings broken down by visual categories.

As expected, the automatic flagging scheme is very effective in indicating spurious detections due to neighbouring sources and in the wings of other extended sources. Where point sources caused extended detections, 29 of the 33 spurious detections were flagged as suspicious (flag #7), for multiple extended detections 38 of the 44 spurious sources were flagged. On the other hand, only a small fraction of the visually confirmed extended sources were flagged, here flag #7 was set in

25 of 494 cases.

According to the visual screening up to 26% of the total sample may be spurious. About 40% of the spurious detections have been caught by the automatic flagging scheme.

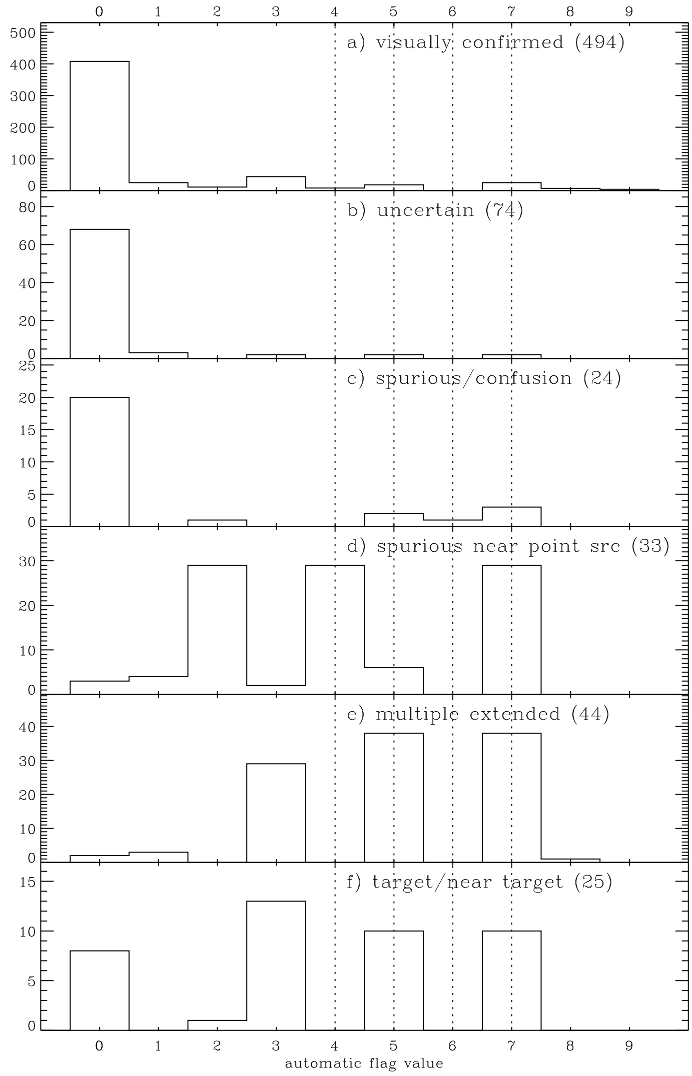

Figure 3: Distribution of 2XMMp automatic flag settings (see section 2.2) broken down by source categories after visual screening (section 2.3).

3 Conclusions and outlook

The 2XMM source catalogue contains 20837 entries of extended source detections. About 50% of these detections are flagged as possibly spurious by the task `dpssflag`. Based on the experience with the 2XMMp catalogue most of the flagged sources can indeed be considered as spurious. Since not all spurious detections are caught by the automatic flagging system, another, yet unquantified fraction of the extended source detections will be spurious. However, the results described in section 2.3 demonstrate, that the fraction of genuine extended detections can be increased substantially, when reasonable sample selections are applied.

In order to improve the future quality of the EPIC extended source detection, corrections to both the source detection algorithm and to the instrument calibration are necessary. A more sophisticated extent model for brighter sources is likely to reduce the number of multiple detections of extended sources. An improvement of the background modelling, which is expected to reduce the number of 'blank field' spurious detections, needs the calibration of detector induced background features and changes to `esplinemap` . An improved calibration of the EPIC PSF is already in preparation, a significant reduction of spurious detections near point sources has already been demonstrated.