Scientific near real-time analysis software for eROSITA

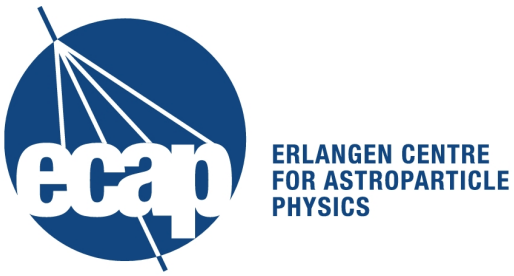
Diploma Thesis

Author:
Peter M. Friedrich

Supervisor:
Prof. Dr. Jörn Wilms

Dr. Remeis-Sternwarte Bamberg
(Astronomical Institute of the Friedrich-Alexander University of Erlangen-Nuremberg, ECAP)
Sternwartstr. 7, 96049 Bamberg, Germany

September 25, 2013



Preface

Abstract

The purpose of this work was to create a near real-time analysis software for eROSITA, a X-ray telescope, which will be aboard the Russian Spectrum-X-Gamma satellite. The software presented here is intended to process the data received by the satellite's ground station shortly after reception. The main goal is to provide some scientific evaluation before the final analysis, because there are possible events like new and/or unforeseen discoveries, which should be payed attention to immediately.
To achieve that, scientific tools were developed. The main components of the processing software are source identification and automated hypothesis testing.
The source identification chooses a good selection of sources from reference catalogs for the detected sources by approximately optimizing the KL divergence between the corresponding distributions. Subsequent the sources are weighted with cross-correlation.
The hypothesis testing makes use of Monte Carlo simulations and checks the reference catalog against the detected sources. If it is likely that the measurement contains new information, the results are rated and alerts generated when indicated.
Parts of the software for the final analysis are used for source detection.
For selected testcases measurements were simulated and processed with the software proposed by this thesis.
Furthermore, a web-based graphical user interface was created to allow users to easily access the data created during the analysis.

Zusammenfassung

Das Ziel dieser Arbeit war, eine Analysesoftware für eROSITA zu entwickeln, die in gewissem Sinne in Echtzeit auswertet. eROSITA ist ein Röntgenteleskop, das an Bord des Russischen Satelliten Spectrum-X-Gamma installiert sein wird. Die vorgestellte Software soll die von der Bodenstation empfangenen Daten kurz nach deren Empfang verarbeiten. Hauptsächlich geht es darum, eine wissenschaftliche Vorabauswertung bereitzustellen, bevor die eigentliche Endauswertung durchgeführt wird, da Ereignisse auftreten könnten, denen sofort Beachtung geschenkt werden sollte, z.B. neue und/oder unvorhergesehene Entdeckungen.
Um dies zu erreichen wurden wissenschaftliche Werkzeuge bereitstellt. Die Hauptkomponenten der Software sind Quellidentifikation und automatisierte Hypothesentests.
Die Quellidentifikation wählt für die gemessenen Quellen eine gute Auswahl aus Referenzkatalogen aus, indem die KL-Divergenz zwischen den entsprechenden Verteilungen näherungsweise optimiert wird. Anschließend werden die Quellen mittels Kreuzkorrelation gewichtet.
Die Hypothesentests werden unter Verwendung von Monte-Carlo Simulationen durchgeführt. Sie überprüfen, ob angenommen werden kann, dass die Messungen neue Informationen enthalten. Die Ergebnisse werden bewertet und ggf. Benachrichtungen generiert.
Es wurden Teile der Software für die entgültige Auswertung verwendet, um Quellen zu detektieren.
Für ausgewählte Testfälle wurden Messungen simuliert und mit der hier vorgestellten Software verarbeitet.
Desweiteren wurde eine webbasierte grafische Benutzeroberfläche erstellt, die den Benutzern erlaubt, auf einfachem Weg auf die Daten zuzugreifen, die während der Datenanalyse generiert wurden.

Contents

List of Acronyms

ARF Ancillary Response File

AGN Active Galactic Nucleus

CCD Charge-Coupled Devices

CV Cataclysmic Variable

ECAP Erlangen Center of Astroparticle Physics

eRASS eROSITA All Sky Survey

eROSITA extended ROentgen Survey with an Imaging Telescope Array

FITS Flexible Image Transport System

FOV Field Of View

FPGA Field Programmable Gate Array

GPS Global Positioning System

GTI Good Time Interval

GSL GNU Scientific Library

HEW Half Energy Width

HK Housekeeping

HMXB High-mass X-ray Binary

ICM Intra Cluster Medium

ISM Interstellar Medium

LMXB Low-mass X-ray Binary

MPE Max-Planck Institute for extraterrestrial Physics

NRTA Near Real Time Analysis

PSPC Position Sensitive Proportional Counters

PDF Probability Density Function

PIL Parameter Interface Library

PSF Point Spread function

RASS ROSAT All Sky Survey

RMF Redistribution Matrix File

SASS eROSITA Standard Analysis Software System

SRG Russian Spectrum-X-Gamma

SNR Supernova Remnant

1. Introduction

1.1. X-ray astronomy

Astronomy is maybe the oldest natural science at all. There is archaeological evidence that dates back 30000 years. Although often practiced in connection with religion, astronomy was also necessary for secular needs like a good knowledge of the seasons for agriculture or predictions of floods. Even until the invention of atomic clocks in the 20th century, the movement of celestial bodies was the most accurate basis for time measurements.
The list of past and current useful applications of astronomy is huge. As example, consider the challenge of determining the position on board of a ship far away from land. Besides solutions like inertial guidance or radio-bearing, observing the sky is a proven method for this task. And even the satellite based systems like the Global Positioning System (GPS) rely on results of researches related to astronomy (Karttunen et al., 2003).

Astronomical observations were restricted to optical light for a long time. But with new technical options, other wavelengths became a focal point of interest a few decades ago. At the moment, the covered energies range from the radio band over infrared radiation to very hard γ-rays. The results pointed out, that observing the cosmos only in the visible light only covers a tiny field of it and many new scientific insights were gathered (Carroll and Ostlie, 2007).

X-ray astronomy is one of such recent branches of astronomy. As the name suggests, it focuses on objects which emit radiation in the X-ray band, which roughly covers energies from 100 eV to 100 keV and more. However, most telescopes are only sensible for soft X-rays, which have energies up to around 20 keV.

The earth's atmosphere is opaque for high-energy photons like X-rays (see figure 1.1). This is pleasant for human life, but therefore, precise observations in the X-ray regime are only reasonable from space.
In 1942 and the following years, Friedmann at al. used rockets to reach sufficient altitudes to detect X-rays originated from the sun. Theoretical work was done prior to that, e.g. by Hulbert at al., who suggested already 1929 to use rockets for the detection of X-rays in the upper atmosphere.
For a long time, the sun was the only object of interest, because of it's relative luminosity. 1962, Giacconi at al. were the first to discover an extrasolar source, namely Sco X-1. Subsequent experiments revealed the great potential of X-ray astronomy by exploring a lot of such sources – many of them orders of magnitudes brighter than the sun. This results – and the exploration of other regimes of the electromagnetic spectrum – yielded to many new insights in the world.
The age of X-ray telescopes mounted on satellites began 1970 with the launch of UHURU. A few years later, 1978, the first imaging telescope in space, HEAO-2/Einstein, was commissioned. Further past missions are EXOSAT (1983-1986), which was designed espe-

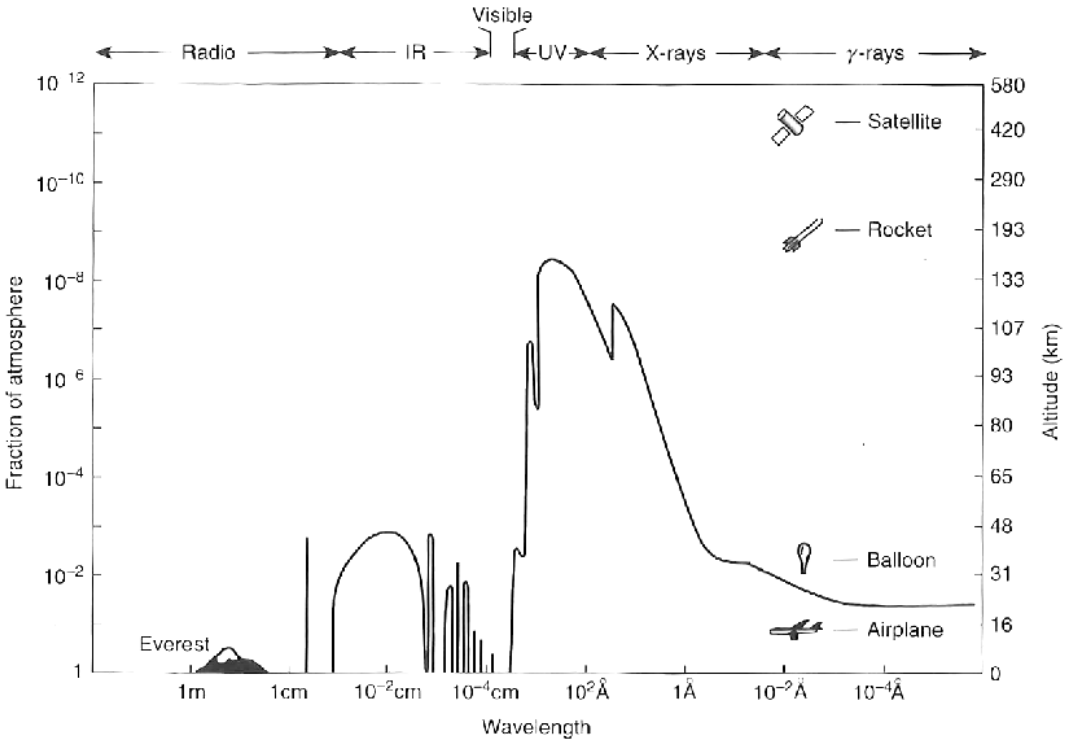


Figure 1.1.: Energy-dependent absorption of photons in the earth's atmosphere. The black line indicates in which altitude half of the incoming photons are passing through. (Seward and Charles, 2010)

cially for detecting time-dependance of sources. The ROSAT-mission (1999-2011) deserves special mention, as besides pointed observations it also performed the ROSAT All Sky Survey (RASS). eROSITA will perform such a survey too and the pre-analysis of the data it will generate is the main topic of this thesis. And obviously, ROSAT motivated the naming of eROSITA.
There are several current missions, e.g. XMM-Newton, INTEGRAL, Chandra or RXTE, and some projects which are in under construction like eROSITA or in planning state (Seward and Charles, 2010), (Giovannelli and Sabau-Graziati, 2004).
As X-ray astronomy is young field of research, it is very likely that there are still many interesting things to discover.

1.2. X-ray emmiting objects

As all electromagnetic radiation, X-rays can be thermally emitted by black bodies, which spectrum is described theoretically by the well-known Planck's law, which is also suitable for the estimation of the conditions needed to emit X-rays. A peak of the spectrum at 1 keV corresponds to a temperature of around $4 \cdot 10^6$ K, so usually emission is occurring in very hot environments (Schmid, 2012). The surface of stars is too cold to be the source of X-rays, e.g. the sun has a temperature of $6 \cdot 10^3$ K. But stars are common X-ray sources, because of their coronas, which have a temperature of millions of degrees. Besides thermal emission, ionized particles interact with other particles, resulting in excited states which

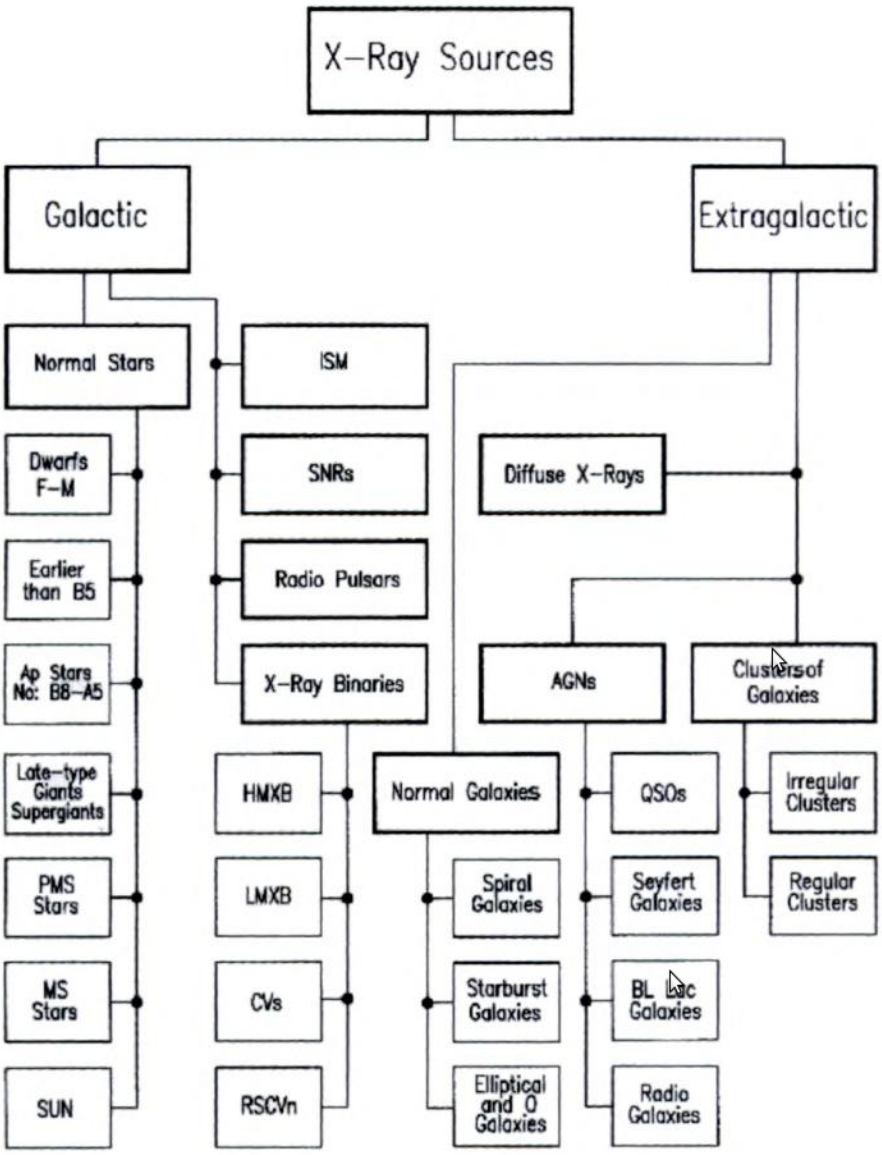


Figure 1.2.: Classification of X-ray sources (Seward and Charles, 2010)

relaxate under the emission of X-ray photons.

In addition to the sun, other objects like planets in the solar system are illuminated, mostly because they reflect X-rays from the sun. There are other processes, but as the focus of this thesis is on the eROSITA All Sky Survey (eRASS) which is intended to cover mainly extrasolar sources, here it is not gone further into it.
Another objects of interest are galaxy clusters. They diffusely emit X-rays between the galaxies as there is hot gas, called Intra Cluster Medium (ICM). Also, infalling gas can impact with resting gas. Ionized gas particles then can emit X-rays due to bremsstrahlung. Objects which aggregate matter form are a large class of X-ray emitting objects. Heavy masses attract material, mostly hydrogen gas as hydrogen is the most common element in the universe and also in stars. Infalling material generates radiation by converting gravitational energy.

As angular momentum has to be conserved, the material does not fall directly into the accreting objects and accretion discs are formed. Due to friction such discs are very hot and therefore they emit a lot of photons, also in the X-ray regime. So also black holes which do are not illuminated itself can indirectly be very bright sources of X-rays.
An example are Active Galactic Nucleus (AGN). They contain a supermassive black hole in their center and are one of the brightest objects in the universe. Because of mechanisms not fully understood yet, they often also eject hot plasma perpendicular to the accretion disc, called jets.
Another example of accreting objects are X-ray binary systems, in which a compact object like an neutron star consumes material from a companion star. They are classified in low-mass (LMXB) and high-mass (HMXB) X-ray binaries, according as the mass of the companion is smaller or much greater than that of the sun. In LMXBs the mass in the Roche-lobe, a region in which there is a balance of gravity forces, is falling through the inner Lagrangian point to the attracting object. In LMXBs, matter is mainly transferred through solar wind.
Neutron stars, which are Supernova Remnants (SNRs), can also be pulsars. They have very strong magnetic fields because of their high angular momentum. In a binary system, accreted matter is following the magnetic flux, so synchrotron radiation is emitted. Finally it impacts at the poles and emits bremsstrahlung. In this case, they are called accretion powered pulsars. There is another kind, the rotation powered pulsars, where the spin axis is different from the magnetic dipole axis. Because of the rotation, charged particles are accelerated which results in radiation originated from the poles. Depending on the angle of view it can be measured as a periodical lightcurves.
Cataclysmic Variable (CV) stars are another kind of binary transients, where the accretor is a white dwarf. The accretion disc is unstable and sometimes a part of it is falling to the surface the white dwarf causing a thermonuclear reaction. Such novae occur irregularly.
Interstellar Medium (ISM), the matter between the stars of a galaxy can also emit X-rays in regions where it contains hot gas. Such can be generated by shock heating in Supernova Remnants.
γ-ray bursts are phenomena with the highest energy flux known. The maximum intensity is only reached for a few milliseconds, where mainly γ-photons are emitted. That is followed by an afterglow with a duration up to several seconds. While the whole process, also X-ray photons are emitted. There are also X-ray bursts which are similar to γ-ray bursts, but they have a much softer spectrum. The processes causing such bursts are not completely known and currently researched.
Only an introductory overview was given here – for more comprehensive information see Trümper and Hasinger (2008) and Giovannelli and Sabau-Graziati (2004), which was also the basic source for this section.
Figure 1.2 shows an overview of X-ray emitting sources.

1.3. X-ray telescopes

There are currently two different kinds of X-ray telescopes, imaging and non-imaging ones. Non-imaging means that it is possible to detect X-ray photons, but not to determine from which direction they came. With collimators it is possible to at least drop photons which are not originated from a wanted region of the sky.

Imaging X-ray telescopes like eROSITA have some sort of optics and are able to detect the photons spacial origin.

1.3.1. Optics

In optical telescopes the imaging is done with reflection or refraction. Refraction of light on a transition from a medium with refraction index n_1 to a medium with refraction index n_2 can be described by Snell's law

$$\frac{\sin \phi_1}{\sin \phi_2} = \frac{n_2}{n_1}, \tag{1.1}$$

where ϕ_1 is the angle of incoming photons measured from the normal of the border between the two media. ϕ_2 is the angle of the refracted photons. Total reflection occurs at the critical angle ϕ_c for the incoming photons:

$$\phi_c = \frac{\arcsin n_2}{\arcsin n_1} \tag{1.2}$$

Stöcker (2004)

X-ray telescopes are today mounted in satellites, so $n_1 = 1$ because of the vacuum in space. Unfortunately in the regime of X-rays n_2 is almost 1, so ϕ_c is near to 90°. Photons have to incline very flat to be reflected. Refraction is also problematic as $\phi_1 \approx \phi_2$.
So traditional optics as used for optical astronomy are not applicable and new techniques have to be developed.
An example is a technology called coded masks. In front of the detector, a mask is mounted, which blocks the photons in some areas. As the mask is known it is possible to reconstruct the incident photons based on the measurements at the detector.
An approach which relies on total reflection are Wolter telescopes. One single mirror does not focus X-ray photons as desired, because it violates Abbe's sine condition:

$$\frac{d}{\sin \Theta} = f, \tag{1.3}$$

where f is the focal length (see Figure 1.3).
Wolter arranged multiple mirrors in a way that the condition is fulfilled. He proposed three types of telescopes, Wolter-1, Wolter-2 and Wolter-3. Figure 1.4 shows a schematic view explaining how a Wolter-1 optics work. The basic idea is to reflect the photons more than once on paraboloid and hyperboloid mirrors (Trümper and Hasinger, 2008).
Imaging X-ray optics can be characterized by several properties. First of all, there is the Field Of View (FOV) which specifies which angular area is covered by the optics. It only gives a very rough approximation, as optical properties worsen on the boundaries of the FOV and optical systems are not perfect, so it is even possible to measure photons originated outside the FOV. But it is useful for restricting the domain of coordinates covered by a measurement.
In most cases and for eROSITA too, the FOV is a square in polar coordinates and it's center is intersected by the optical axis, so it is completely specified by one angle.
As the amount of photons reflected by the optics' mirror varies with energy, a quantity

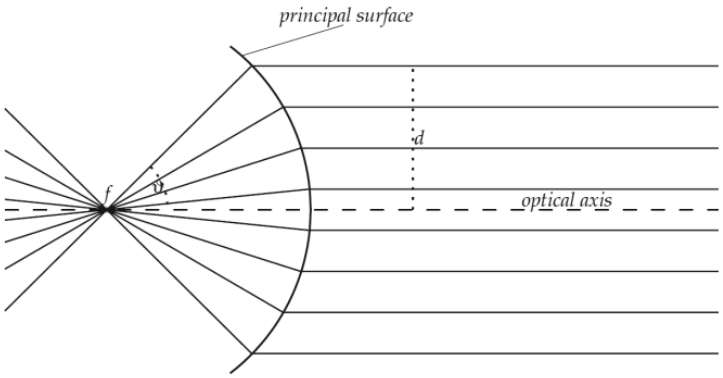


Figure 1.3.: Abbe's sine condition (Trümper and Hasinger, 2008)

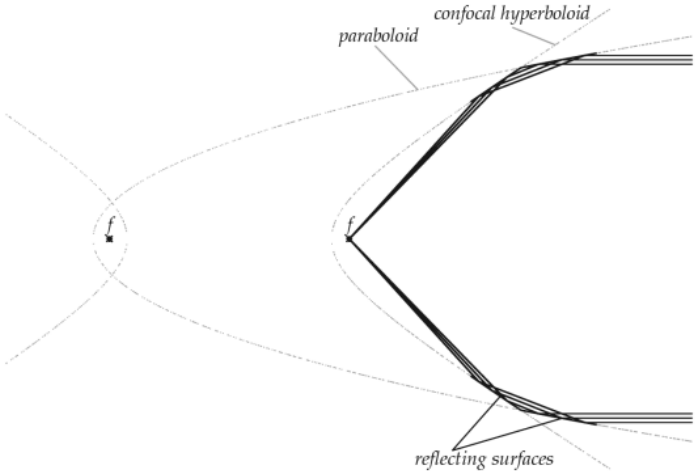


Figure 1.4.: Scheme of a Wolter-1 optics (Trümper and Hasinger, 2008)

called effective area is used. It is the product of the ratio between incoming and reflected photons with the area which is projected by the optics. The effective area can be stored in an Ancillary Response File (ARF) as specified by George et al. (2007).

The effective area changes with the photons' energy and the off-axis angle. The dependance can be described using a vignetting function v.

A Point Spread function (PSF) describes how incoming on-axis photons are imaged on the detector. For a given distribution for incident photons $I(x, y)$, the corresponding distribution on the detector $D(x, y)$ is given by the convolution of I and the PSF:

$$D(x,y) = (I * PDF)(x,y) = \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy'\, I(x',y')\, v(x',y')\, PSF(x-x', y-y') \quad (1.4)$$

As I and D are defined to be Probability Density Functions (PDFs), so they have to be normalized and because of that, also the PSF.

The underlying idea is to describe how a point source $\delta^2(x - x_0, y - y_0)$ will be imaged. In this case, $D(x, y)$ is just the PDF at $(x - x_0, y - y_0)$. At this point it is clear that the PSF itself can be considered as an PDF. Often approximations are used, e.g. by setting

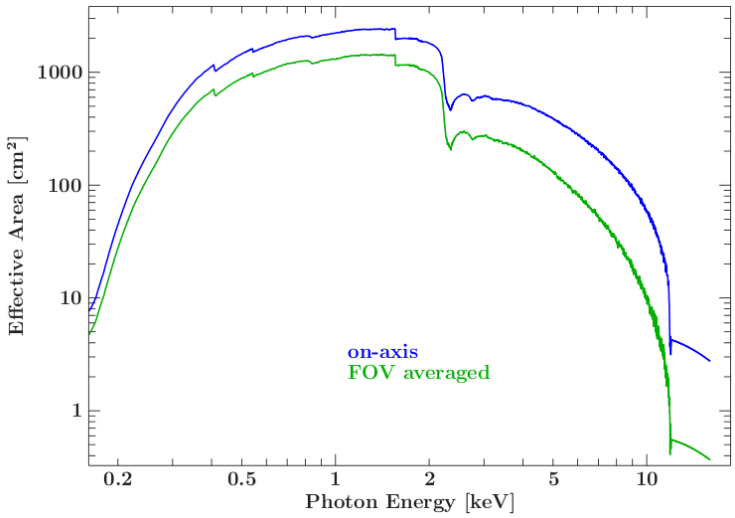


Figure 1.5.: ARF of eROSITA (Schmid (2012))

$v \equiv 1$.
In general, the PSF depends on the energy of photons and often it is necessary to include off-axis incoming photons, so additional parameters have to be introduced, but the keynotes of the PSF are conserved.
One should note that the parameters used here are Cartesian coordinates, but it can be easily transformed to polar coordinates.
The PSF can be used to calculate angular resolutions, usually the Half Energy Width (HEW) is used, which is defined by:

$$\int_0^{HEW/2} dr \int_0^{2\pi} d\varphi \; r \cdot PSF(r, \varphi) = \frac{1}{2} \tag{1.5}$$

This is the diameter of the circle containing 50% of the incoming photons at the focus. Schmid (2012) Unfortunately, in reality a PSF is a very complicated object, especially for the optics used in eROSITA. Using a set of PSFs for particular parameters and interpolation suggests itself.

1.3.2. Detectors

The simplest detector consists of a capacitor with voltage applied. The capacitor is filled with gas which can be ionized by X-rays. The desorbed electrons are moved by the electric field and result in a measurable current.
An improvement are proportional counters. They have a cylindrical cathode with a window through which photons can enter. The anode consists of one ore more thin taut wires in the inside. If photons cause free electrons in the gas, they move to the wires. Their number is proportional to the energy of the photon which caused the naming of the detector. If photons are near to the wire they can be accelerated so fast that they again produce secondary electrons through interaction with the gas, so a cascade of electrons occurs which improves the sensitivity of the detector.

It is possible to get Position Sensitive Proportional Counters (PSPC) by measuring the signal on both ends of the wires and then to estimate at which position the photon ionized the gas.
Another type of detector are Scintillation counters. Instead of gas a solid material is used which molecules can absorb X-ray photons. After an absorption, molecules are in excited states which relax after a short time with the generation of photons which can be measured with photo-detectors. The advantage is that even very hard X-rays with energies greater than 20 keV can be detected as the cross section of the used materials is much higher than that of gas.
Charge-Coupled Devicess (CCDs) are almost everywhere in today's world, e.g. modern cellphones are usually equipped with a camera with a CCD detector. Originally designed as computer memory, it became clear very fast that they are sensitive for light.
Basically they are semiconductors consisting of arrays of pixels, usually quadratic. When a photon impacts at a pixel, electron-hole pairs are generated. In contrast to photo-diodes the generated charges do not drain directly outside the device. Instead they are trapped in an electric field.
The pixels are read out by shifting the charges from pixel to pixel in one line of the array. Behind the last pixel of one line is a readout buffer which temporary stores the pixel charges which fall out of the array because of the shifting. There it can be measured using an Analog/Digital converter (A/D) (Karttunen et al., 2003)

Now possible problems of CCDs are decribed (see Wille (2011)):

- *Out of time events:* When photons arrive during the readout process, it is impossible to determine where they did impact. This disadvantage can be decreased by using a frame store area which temporarily stores the charges (see Merloni et al. (2012)).

- *Split events:* The CCD's pixels are usually not isolated in readout direction. If a photon hits the CCD near between the border between two pixels, it is possible that it affects the adjoining pixels too. And even the thin isolator between different lines are not always sufficient to prevent that.

- *Blooming:* The electric field which holds the charge in the pixel is limited. So if the exposure is too long the charge can overflow to neighborhood pixels.

- *Pile up:* Normally the energy of the measured photons is of interest. But if two or more photons are detected in one pixel, it is impossible to determine their separate energies. It is not possible to distinguish between one photon with energy E and multiple photons, whose energies sum is equal to E.

- *Particle background:* Not only X-ray photons are measured by the CCD. Many other kinds of particles can cause charges in the pixels and it is impossible to shield the detector against all of them.

- *Dark current:* Thermal energy can cause unwanted generations of electron-hole pairs.

The detector response can be modeled by using an Redistribution Matrix File (RMF) and ARF (see George et al. (2007)). With them the expected value for the counts $C(h)$ of a

specific channel h can be calculated with

$$C(h) = \int dE \; RMF(h, E) \cdot ARF(E) \cdot M(E) \tag{1.6}$$

M is the PDF of the photons arriving at the detector.

2. eROSITA

2.1. Mission

extended ROentgen Survey with an Imaging Telescope Array (eROSITA) is a German imaging X-ray telescope which will be on-board the Russian Russian Spectrum-X-Gamma (SRG) satellite. It is currently under construction, supervised by the Max-Planck Institute for extraterrestrial Physics (MPE) in Munich. The development is done by several institutes, including the Dr. Remeis Observatory located in Bamberg, which is affiliated to the Erlangen Center of Astroparticle Physics (ECAP).
The start of the spacecraft from Baikonur, Russia was delayed multiple times and now planned for end of 2014 [1]. It will perform the eROSITA All Sky Survey (eRASS), a deep survey of the entire sky. However it's angular resolution is in the scale of ROSAT, the spectral resolution is impressive and the sensitivity in the soft X-ray band (0.5-2 keV) will be 20 times as high and in the hard band (2-10 keV) it will be the first telescope measuring the whole sky. After the eRASS, selected pointed observations will be performed.

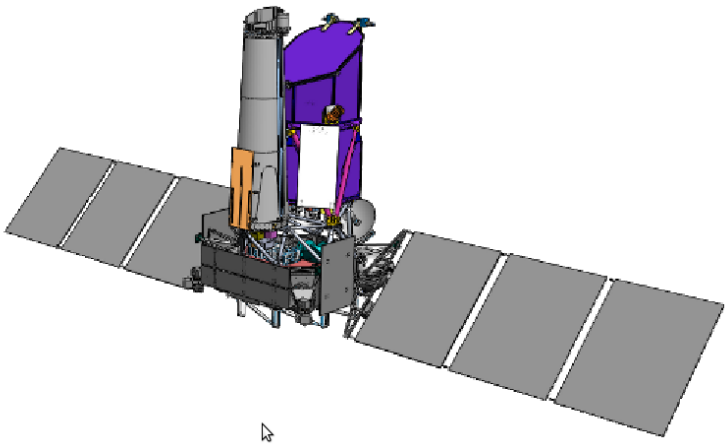


Figure 2.1.: SRG with ART-XC (front left) and eROSITA (back right) (Merloni et al., 2012)

Russian Spectrum-X-Gamma (SRG) will be placed in an halo orbit around the second Lagrangian point (L2)[2] of the sun-earth-system. All it's telescopes are looking an the same direction, which is almost perpendicular to the solar panels. The satellite rotates around an axis pointing towards the sun or several degrees away from it, so the solar panels do not have to be adjusted much. While one rotation interval, which is approximately 4

[1] See The eROSITA Bulletin, No. 3, May 2013
[2] L2 is approximately located 1.5 million kilometers behind the earth when looking from the center of the sun.

hours long and called eROSITA day (eroday), the telescopes scan one great circle of the sky. As the rotation axis is pointed to the sun, the circle's right ascension changes over time resulting that the whole sky is covered after a half year.

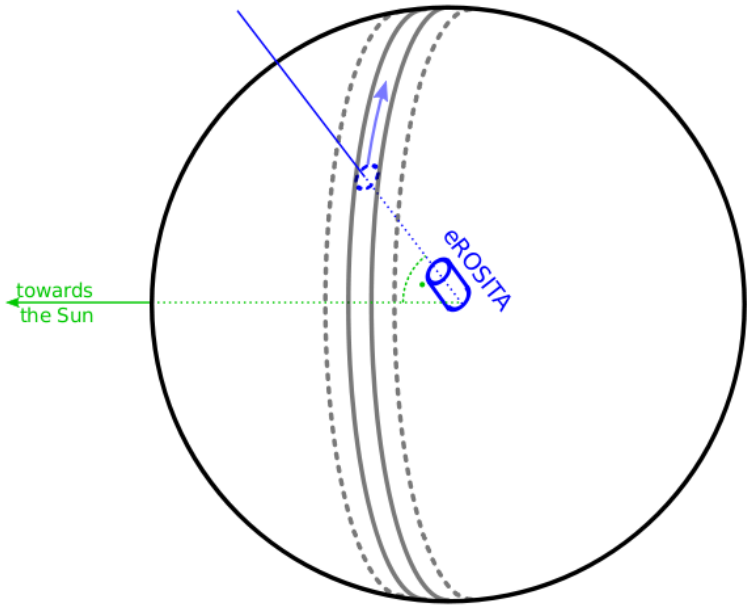


Figure 2.2.: eROSITA's attitude during eRASS (Schmid, 2012)

One of the main scientific goals is to study the universe at large scales by observing many galaxy clusters. As mentioned in 1.2, matter between galaxies can be observed in the X-ray regime. The distribution of this matter depends mainly on gravitational potential, so it is well suitable to verify cosmological models and the investigation of dark energy and dark matter.
Also it is expected to detect a huge amount of AGNs which allow the analyze how supermassive black holes evolute. Further goals are e.g. to better understand galactic compact objects, stars and SNRs (Merloni et al., 2012).
Also the detection of completely unknown phenomena is possible.

2.2. Specifications

The eROSITA telescope is in fact a collection of seven Wolter type I telescopes, all constructed in the same way and pointing in the same direction. Their main properties are collected in table 2.1
The CCDs consist of two parts each. One image area where incoming X-ray photons are measured and a frame store area shielded against X-rays. Both have 384 x 384 pixels and are on the same chip. After exposure the charges can be shifted very fast ($\approx 0.1\mu s$) into the frame store area, where it can be read out without the danger of the generation of new charges caused by photons which impact after the process began. (Merloni et al., 2012)

focal length	1.6m
on-axis resolution	15" HEW at 1.5 keV
detector	framestore pn-CCD, 384^2 pixel
FOV	61'
mirror coating	Au
pixel size	75μm (corresponding to $\approx$ 9.7 arcsec)
time resolution	50 ms
mirror shells per module	54
energy resolution	138 eV at 6 keV
energy range	$\approx$ 0.2 - 10 keV

Table 2.1.: eROSITA specifications

2.3. Data analysis

This section is intended to overview the kinds of appearing data, the dataflow and the planned analysis steps and their objectives. Also, this thesis is associated within this context.

2.3.1. Dataflow and acteurs

It is appropriate to begin with a short description of the data processing in the satellite itself and it's communication with the earth station, as that is part of the fundamental experimental setup.
The on-board soft- and hardware of eROSITA control the telescopes based on instructions from earth, but it does not just acts as an interface. It is capable of doing some early preprocessing like the subtraction of an offset-map. Also, because the communication has limited speed, is not strongly reliable and is only possible in intervals, when the satellite is visible from an earth station. So, there is a mass memory to temporary store data. The computing devices of eROSITA mainly consist of an Field Programmable Gate Array (FPGA) and two classical PowerPC processors (Merloni et al., 2012).
Furthermore it is obvious that some tasks such keeping the solar panels directed perpendicular towards the sun can and should be performed quite independently. This is done by the satellite itself which has independent computer hardware which is not further focused on here.
The data sent to earth can be classified into two main categories, Housekeeping (HK) and telemetry. Housekeeping data contains information about the states and conditions of the hardware, e.g. temperatures, power consumptions, positions of cover plates or the rotation speed. Telemetry data consists of the telescopes' measurements, i.e. the observations – in case of eROSITA pixel-data readout from CCDs with corresponding times and telescope identifiers. The satellite's attitude belongs to another category, auxiliary data. It is specified not only by the orientation of hardware components relatively to others but also to stars, which is determined by observing them with a star-tracker.
The data is buffered until it can be transmitted. It's not guaranteed that it is received completely and in proper sequence.
The earth station will forward received data to the Near Real Time Analysis (NRTA)

software operated at the Dr. Remeis Observatory in Bamberg, where the raw data is preprocessed for the final processing, which will be mainly performed by the MPE.

2.3.2. Data preprocessing

The first step done by the NRTA is the conversion of incoming data to the Flexible Image Transport System (FITS)-format according to well-defined schemes. As the incoming data is not completely specified yet, it is possible to configure the software in a very flexible way by defining the structure of incoming data in an XML file. The converted data is then persistently stored in the raw data archive. For the case of fatal errors, also the raw input data is stored so it cannot get lost. The tool which performs this tasks is named TM2FITS. Afterwards, the data is arranged for scientific analysis and stored. New output of TM2FITS is continuously fetched and merged in a preprocessing database. The data is chronologically sorted and split into erodays, which are the basic units for analysis. Another program checks the completeness of the data, i.e. it determines Good Time Intervals (GTIs) in which all observation data was received. It forwards the events of complete erodays to the archive injector if the eroday is completely overlapped by GTIs or if a specific deadline time elapsed. Missing data which arrives afterwards is stored in the archive indeed, but in a special way, because "complete" data is maybe already forwarded to the final analysis and the results must be equally reproducible at any time. The archive injector copies the results into the archive for the final analysis (Grossberger and Wille, 2010).
Figure 2.3 shows an overview of this tasks.
All tasks are performed cumulative, so in case of restarts, only relevant new data is processed, not the whole database.

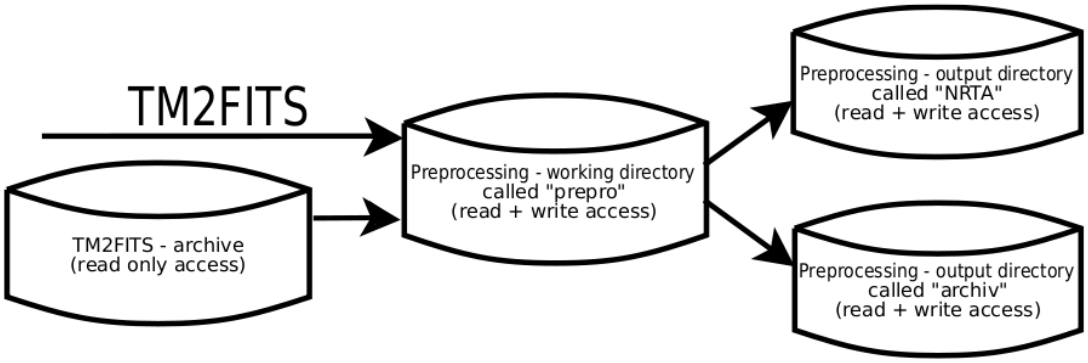


Figure 2.3.: Schematic of the different directories, where the arrows represent the direction in which files can be copied (Grossberger and Wille, 2010).

2.3.3. Housekeeping

The NRTA software is also responsible for keep track of the status of eROSITA, so one of it's tasks analyzes incoming housekeeping data immediately and creates alerts when un-

expected conditions occur. The monitored data includes state information, temperatures, voltages, rate information and diagnostic maps like noise maps.(Wilms et al.)

2.3.4. Final analysis

The final analysis will be done with the eROSITA Standard Analysis Software System (SASS). It provides tools for performing pipeline processing and for interactive analysis of the data, which also involves standard software like XSPEC. Mainly validated data products for further breakdown are created.

The SASS is based on the software created for ROSAT, Abrixas and XMM-Newton. The main data products which will be created are Brunner (2009):

- calibrated event lists
- sky images
- exposure maps
- background maps
- source lists
- cross-correlation lists
- source spectra
- source time series

Although the term SASS often incorrectly does not refer to the NRTA, actually the NRTA is part of the SASS. Therefore here it is suggested to introduce a new definition, F-SASS, to indicate that only the components designed for the final analysis are meant.
Figure 2.4 shows an overview of the SASS pipeline processing.

2.3.5. Scientific near realtime analysis

Besides the tools for data preprocessing, further called NRTA-P, the NRTA also contains software for performing a preliminary scientific analysis, further called NRTA-S. This is no hard separation as e.g. they share some program functions.
The purpose of NRTA-S is to analyze incoming data immediately, even if it's not considered complete enough for processing through F-SASS. The data format of the inputs is identical for both (Wilms et al.). See also figure 2.5.
Especially it should be able to detect unexpected measurements like new X-ray sources which were or could not have been detected by past observations of other missions or eROSITA itself.
This thesis suggests some approaches to archieve that.

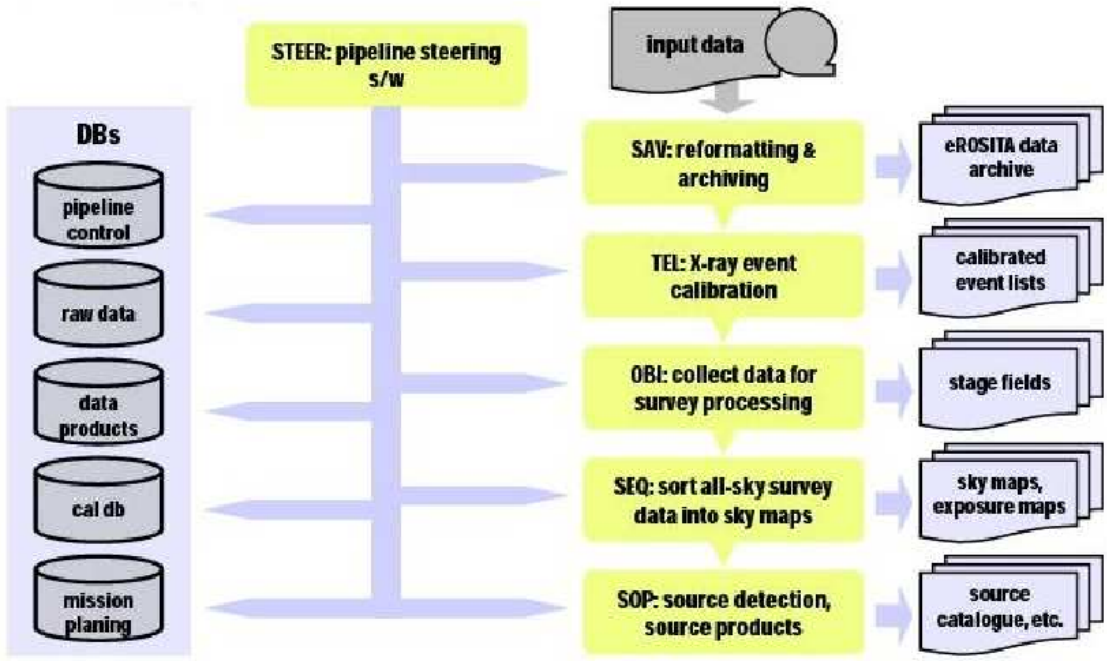


Figure 2.4.: Overview of the SASS pipeline processing (data processing program chains, control software, and database system)(Brunner, 2009)

2.3.6. Used software

The operation system on which the software will run is Linux, but in principle it should be able to compile it on other platforms, too. The F-SASS is almost entirely written in the programming language Fortran 77, the NRTA – except TM2FITS – in C.
For easy compilation of the NRTA-S tools the GNU autotools are used.
All generated data is stored in the data format Flexible Image Transport System (FITS), which was especially designed for astronomical data. It is time-tested since several decades and the de facto standard in many branches of astronomy today. The used API for working with such files is CFITSIO. It is part of HEASoft software package which also contains the Parameter Interface Library (PIL) also used in NRTA.
For numerical calculations the GNU Scientific Library (GSL) is applied.
The team responsible for eROSITA's NRTA decided to use FPIPE. FPIPE was originally developed by Schwarzburg (2005) in the context of his Diploma thesis titled "A software for realtime analysis of experimental data in Flexible Image Transport System (FITS)"[3]. The title of his work describes concisely the purpose of FPIPE. As the name suggests, it is based on pipelines, which transfer data in the FITS format between processes with well-defined interfaces. The pipelines can easily described and configured with FITS or XML files.
So FPIPE is using the dataflow programming paradigm, a concept which is used by many analysis frameworks like Labview or ON[4]. Maybe one of the greatest features is that tools

[3]own translation – the original German title is "Eine Software zur Echtzeitanalyse von experimentellen Daten im Flexible Image Transport System (FITS)"
[4]Schwarzburg mentioned them among others as examples for realtime analysis software which obviously inspired FPIPE.

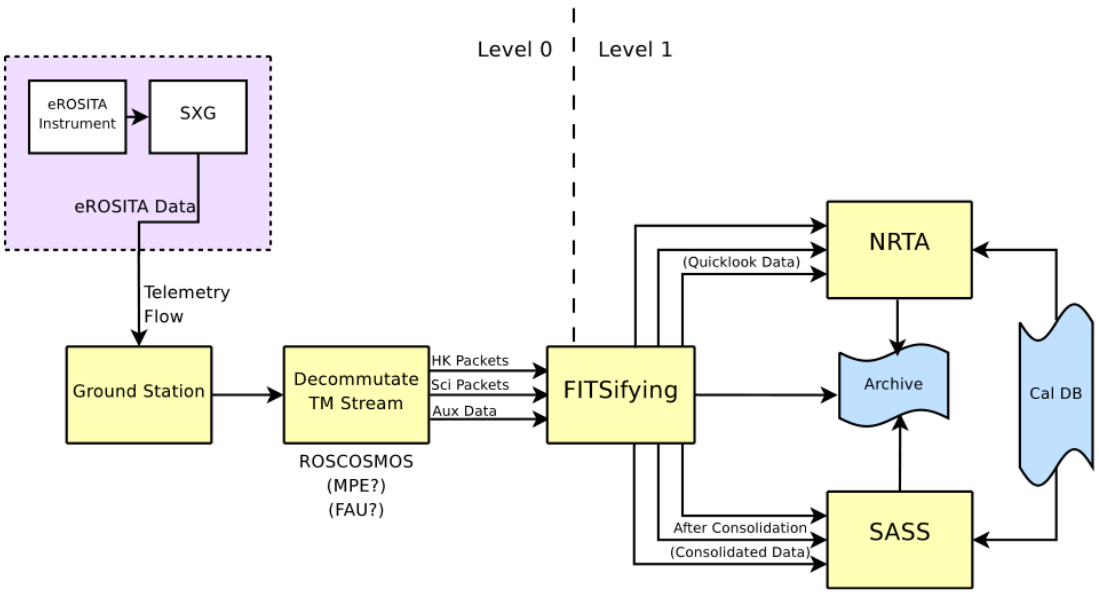


Figure 2.5.: Data flow from the measurement to the ground segment and the NRTA and SASS pipelines (Wilms et al.)

in a pipeline can be run in server-mode. That means they are not restarted every time new input data is available. Instead they get their instructions by interprocess communication (IPC). That has the great advantage that if a tool needs to initialize data structures, that has not to be done every time it has to process new data, which can lead to enormous performance benefits. To this day, FPIPE was developed further by several people.
The software details of the user interface will be described in chapter 4.

3. Scientific near realtime analysis

3.1. Overview

The NRTA-S receives data from the NRTA-P binned in erodays. Contrary to the final analysis the data has not to be complete.
Firstly the data is sent to the orbit prediction which generates a table containing the expected positions of the satellite in the future. Then, the attitude prediction generates an attitude file under usage of this information.
Afterwards the source detection generates a source list, maybe it uses the predicted attitudes if necessary. For each source, candidate sources from a reference catalog and past NRTA-S results are selected. The candidate mapping is further refined in the source identification task, which determines which and how much reference sources belong to the most likely detected sources.
Subsequently it is checked with hypothesis tests if the measurement contains new information, i.e. if a detected source does not belong to the selected candidates. The results are classified and rated. If the rating exceeds some limit, alerts are generated and indicated sources are inserted into the NRTA-S source catalog to prevent duplicate alerts in the future. This catalog is also useful for an overview of discoveries made so far.
Figure 3.1 shows the suggested pipeline configuration. The candidate selection, source identification and hypothesis test tasks do not generate separate output files, instead they append new HDUs to the source detection FITS-files. So the whole output of one NRTA-S run is stored in one single FITS-file. All information about involved reference sources are stored in the HDU containing the candidates. This is done for convenience if the results will be further analyzed by scientists. Only one FITS-file is required to comprehend thrown alerts.

3.2. Expected exposure times

During eRASS the interesting observations for NRTA-S are performed. eROSITA scans the sky with 7 independent subtelescopes, each directing in the same direction. The satellite rotates once every so-called eROSITA-day (eroday), which is around four hour long. This causes the FOV to uniformly propagate over a great circle on the celestial sphere. The FOV is a circular area with a diameter of $d = 61 arcmin$.
While the satellite moves around the sun, the great circle changes, but all of them are intersecting at two poles, so it is suitable to speak about latitude Φ and longitude ϕ, but here for simplicity, neither a origin of coordinates for longitude nor the north/south-orientation are defined.
As the satellite stays on L2, an eROSITA-year is an earth-year, so the FOV's center's

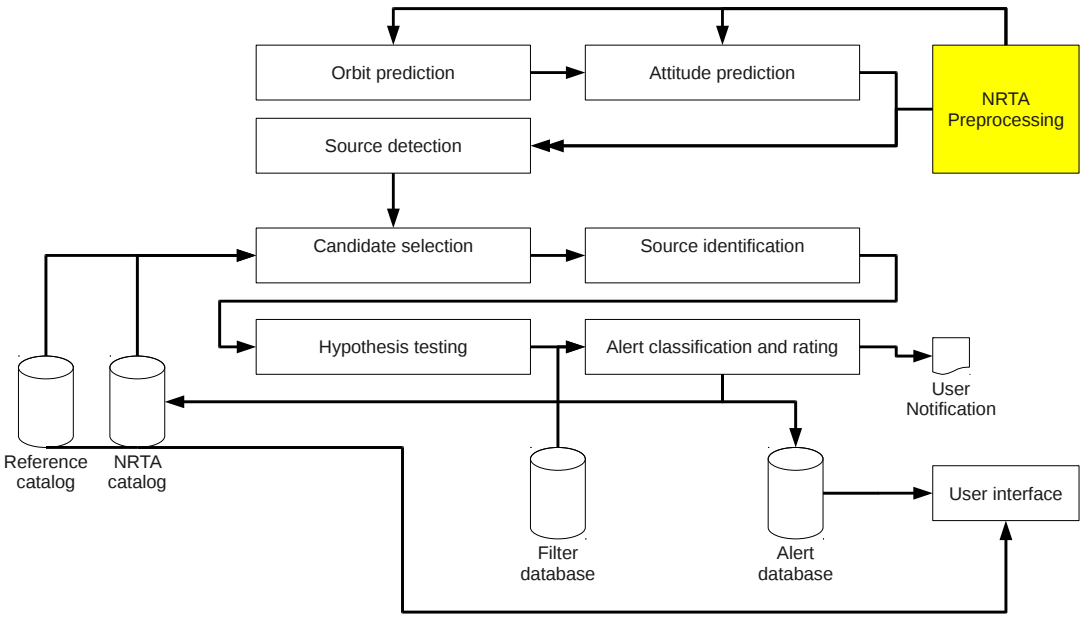


Figure 3.1.: Suggested pipeline and dataflow for NRTA-S

position moves according to

$$\dot{\Phi} \approx \frac{2\pi}{4h} \tag{3.1}$$

$$\dot{\phi} \approx \frac{2\pi}{1a} \tag{3.2}$$

During one eroday there is no overlap of exposed ares, so the exposure time $E(x)$ for a covered sky position which is an angle of x away from the FOV's center perpendicular to the scan direction is approximately (see figure 3.2):

$$E(x) = \frac{2\sqrt{(d/2)^2 - x^2}}{\dot{\Phi}} \tag{3.3}$$

For $x = 0$ the exposure time is around 40 seconds which seems enough to reasonable try to perform source detection, which will be confirmed in the simulation chapter. It should be noted that eROSITA consists of 7 identical telescopes which ideally all contribute that exposure.

3.3. Handling of incomplete data

Here it is distinguished between three types of incomplete data. Telemetry data comes in packets consisting of the pixel-readout of the CCD-detectors, but there is no guarantee they are received in the right order. They can even get lost or arrive very late.

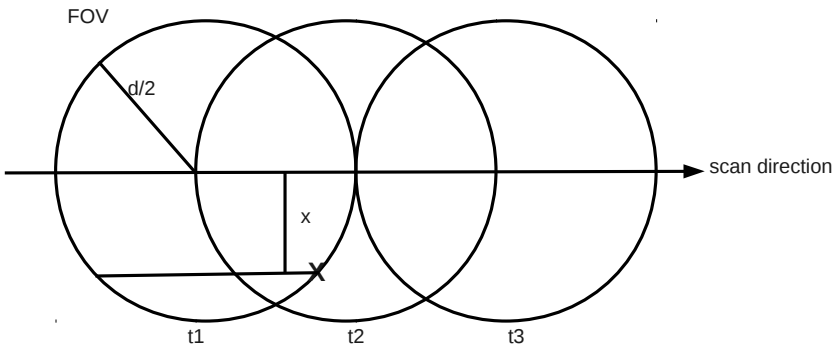


Figure 3.2.: Exposure during scan

The telemetry data can also be permanently affected by systematic errors like bad pixels or failures of the readout circuits.
Both cases maybe will handled by the NRTA-P. The third one is when no attitude information is available for a measurement. In this case it will be predicted.

Without relativistic effects the satellites trajectory x is described by:

$$\frac{d\underline{x}}{dt^2} = \sum_i m_i \frac{\underline{r} - \underline{x}}{|\underline{r}_i - \underline{x}|^3} \tag{3.4}$$

Where each m_i and $\underline{r}$ corresponds to properties respectively trajectories of relevant objects, mainly planets. Their trajectories could be described by Newton's law, too, but the orbit prediction[1] is using SPICE, a toolkit generated by the NASA, for gathering that information. So only the given equation has to be solved for given initial values, i.e. a recent known position and velocity. The solution is generated using capabilities of the GSL and stored in a FITS file containing nicely interpolatable time/position-pairs.
It should be mentioned that all orbits around Lagrangian points are not completely stable, so from time to time the satellites thrusters have to be used to correct the position. If that happens, the orbit and therefore the attitude prediction tasks fails as they can not know about it.

The tool for attitude prediction performs a linear interpolation between the two points next to a given time. If the available data does not permit that, an error is thrown. From now on, assume we have a function $\underline{x}(t)$, which provide the predicted satellite position in Cartesian solar barycenter coordinates with ecliptic coordinate system axis.
The optical axes of the telescopes are uniformly rotating perpendicularly to an imaginary line directed directly to the sun. Not a method to predict the attitude is presented. Assume at time t_0 that $\underline{a_0} \in \mathbb{R}^3$ is a normalized vector that points in the telescopes viewing direction. Let $\underline{x_0}$ and $\underline{x_1}$ be two unit vectors pointing from the sun to the satellites position at times t_0 respectively t_1. If the satellite does not rotate, but $\underline{a_0}$ is perpendicular

[1] developed by Wiebke Eikmann

to the orbit's vector at all times, then the attitude $\underline{a_1}$ at time t_1 can be determined by rotating $\underline{a_0}$ in the same way as $\underline{x_0}$ is rotated to $\underline{x_1}$.
The normalized axis n and the angle α of the joint rotation can be calculated using scalar and cross products:

$$|\underline{x_0} \times \underline{x_1}| = |x_0||x_1| \sin\alpha = \sin\alpha \tag{3.5}$$

$$\underline{n} = \frac{\underline{x_0} \times \underline{x_1}}{|\underline{x_0} \times \underline{x_1}|} = \frac{\underline{x_0} \times \underline{x_1}}{\sin\alpha} \tag{3.6}$$

Actually the satellite rotates, so there is an additional rotation. In the mathematical sense it does not matter if the satellite rotates uniformly with an angular momentum $\frac{d}{dt}\beta$ or performing the whole rotation at once, e.g. at time t_0 by $\beta = (t_1 - t_0) \cdot \frac{d}{dt}\beta$ around the axis $\underline{x_0}$.
$\underline{a_1}$ is calculated by applying Rodrigues' rotation formula two times for the two rotations. It states that if $\underline{v}$ is rotated around an normalized rotation axis $\underline{z}$ by an angle φ, the resulting vector $\underline{v'}$ is given by

$$\underline{v'} = \underline{v}\cos\varphi + (\underline{z} \times \underline{v})\sin\varphi + \underline{z}(\underline{z} \cdot \underline{v})(1 - \cos\varphi) \tag{3.7}$$

The change of the roll angle from t_0 to t_1, i.e.the rotation of the FOV, is determined by calculating how much the rotation around $\underline{n}$ contributes to a rotation around $\underline{a_0}$.
The attitude prediction requires as input the requested time interval, the output of the orbit prediction covering that interval and some initial values (see appendix B.1). Details of the output can be found in appendix D.
Analogous to orbit prediction it's output is an attitude file containing predicted attitudes. The time step size can be configured.

3.4. Models and reference catalog

It is assumed that there is a standardized PDF s which represents the photon distribution on the sky.
As usual, polar coordinates are used to specify positions on the celestial sphere, α is the declination and δ the right ascension[2]. The complete PDF s depends on this positions, the photon energy E and the time t, so $s = s(\alpha, \delta, E, t)$. The reference catalog will not contain lightcurves, so $s = s(\alpha, \delta, E)$. An approach how time-dependent sources can be partially modeled will be presented later.
In the strict sense the PDF has an additional degree of freedom to specify the position in space at which s is valid, but here it is assumed that all sources are far away. Although ignored in this work it could be possible to measure X-rays originated from planets or that extrasolar sources are masked by them[3].
It is assumed that s is a finite sum of N independent PDFs, each corresponding to a physical photon source. For a source identified by $i \in \mathbb{N}$, let s_i be it's normalized PDF

[2]It is necessary to provide an epoch in which the coordinates are valid. The software presented here expects that all coordinates are given for a fixed epoch.
[3]Maybe the scan strategy will avoid that the telescope detects planets. It clearly avoids that the optical axis directed nearby the sun, as it is almost perpendicular to the direction to the sun at all times.

and r_i the photon rate. Then s is given by

$$I := \sum_{i=1}^{N} r_i \tag{3.8}$$

$$s = \frac{1}{I} \sum_{i=1}^{N} r_i \cdot s_i \tag{3.9}$$

It is presumed that a sources' PDF $s_i(\alpha, \delta, E, t)$ can be separated in two independent PDFs, a PDF describing the spacial distribution $I_i(\alpha, \delta, t)$ and one describing the spectrum $S_i(E, t)$. So s_i can be written as $I_i \cdot S_i$. It should be noted that actually the existence of sources which do not allow such a model is not impossible. If really needed it is thinkable to model them approximately by a finite superposition of multiple s_i.
To determine the number of incoming photons a Poisson distribution can be assumed, which gives the probability to detect k photons:

$$PDF_{Poisson}(k) := \frac{\lambda^k \exp(-\lambda)}{k!} \tag{3.10}$$

λ is the rate which can be determined from s by integrating over the regions of interest, i.e. a set of photon properties, and multiplying with I and the measurement duration.
The spatial PDF of every source is given by a two-dimensional Gaussian distribution in the model without correlation between the two axes, so the covariance matrix is diagonal. Extended sources which are not rotational symmetric and described by Gaussians are not modeled. A source at position (α_i, δ_i) with extend σ is then described by

$$I_i(\alpha, \delta; \alpha_i, \delta_i, \sigma_i) = \frac{1}{2\pi\sigma_i} \exp\left(-\frac{1}{2\sigma_i^2}\left((\alpha - \alpha_i)^2 + (\delta - \delta_i)^2\right)\right) \tag{3.11}$$

For a point-source, we define

$$I_i(\alpha, \delta; \alpha_i, \delta_i, 0) := \lim_{\sigma_i \to 0} I({}_i\alpha, \delta; \alpha_i, \delta_i, \sigma_i), \tag{3.12}$$

which is just a delta distribution:

$$I_i(\alpha, \delta; \alpha_i, \delta_i, 0) = \delta(\alpha - \alpha_i)\delta(\delta - \delta_i) = \delta^2(\alpha - \alpha_i, \delta - \delta_i) \tag{3.13}$$

Actually the Source spectra $S_i(E)$ can be complicated objects. For bright sources it is maybe possible to extract spectra and lightcurves even in the date provided to the NRTA-S. For this reason it is designated to process such information.
The SOU chain of the F-SASS will contain tools for that, so they should be used if it seems appropriate to analyze such things in the NRTA-S. Currently the tools are not ready, yet[4].
At the moment the following compromise between not analyzing spectra and using the extracting tools of the final analysis is suggested: The used source detection programs can split sources into energy bands, i.e. a source which has contributions in different energy

[4]See the internal document `http://www2011.mpe.mpg.de//erosita/internal/SASS-devel/SASStasks.html` for getting status information of SASS tasks.

bands will result in multiple entries in the generated catalog (Brunner, 2012). That can be interpreted as a spectrum with very large bins.
So the basic idea of the catalog is that sources are conglomerates of parts. S_i is modeled as a sum of M Gaussians with contributions r_{ij}:

$$S_i(E) := \frac{\sum_{j=1}^{M} r_{ij} \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left(-\frac{(E-E_{ij})^2}{2\sigma_{ij}^2}\right)}{\sum_{j=1}^{M} r_{ij}} \tag{3.14}$$

Lightcurves are not modeled at all. Because of the short exposure times, the NRTA-S is not able to resolve the sources timing during one eroday. The only lightcurves it generates are binned to erodays. But sources for which is known that their flux extremely changes over time, multiple sources at the same position can be inserted into the reference catalog. This will avoid that the NRTA-S is not able to identify such sources.
Very long transients or objects which were illuminated in the past and then went off for undetermined time can be flagged in the reference catalog. If so, the NRTA-S will generate an alert if such a source went on and was detected.

The model presented so far does not allow the specification of errors. The reference catalog as well as the results of source detections as used in NRTA-S contain errors.
According to the central limit theorem, the mean of a list of real-valued samples, generated by an arbitrary distribution, follows a Gaussian distribution if the sample size goes to infinity. The estimation of source properties are based on measurements of photons. Although the number of measured photons is small for the measurements analyzed by NRTA-S, it is assumed here that the errors are Gaussian ones. This leads to some inexactness, especially at the tails of the Gaussians. Furthermore, performing source detection means performing curve fitting, so the resulting errors are actually not statistical ones but residuals.

If errors are involved, it is reasonable to define a distributions E_s and E_I which describe the probability of specific s and I respectively. So E_s and E_I are distributions of distributions. The basis for the models introduced in the next sections are the expected distributions. They are introduced here exemplarily for the case of a extended source in one dimension without spectral information at position $x_0 \pm \Delta x_0$ and extend e. Please note that $\pm$ denotes statistical errors in this work, not error limits or other quantities.
The probability that the source is located at x is given by $\mathcal{N}(x_0, (\Delta x_0)^2)(x)$, i.e. that the photon distribution is given by $\mathcal{N}(x, e^2)$. So the random variable of interest is itself a distribution.

The expected photon distribution $D(x)$ is

$$\begin{aligned} D(x) &= \mathbb{E}_{x' \sim \mathcal{N}(x_0,(\Delta x_0)^2)}[\mathcal{N}(x', e^2)(x)] && (3.15)\\ &= \int_{-\infty}^{\infty} dx' \quad \mathcal{N}(x_0, (\Delta x_0)^2)(x') \cdot \mathcal{N}(x', e^2)(x) && (3.16)\\ &= \int_{-\infty}^{\infty} dx' \quad \mathcal{N}(x_0, (\Delta x_0)^2)(x') \cdot \mathcal{N}(0, e^2)(x - x') && (3.17)\\ &= (\mathcal{N}(x_0, (\Delta x_0)^2) * \mathcal{N}(0, e^2))(x) && (3.18)\\ &= \mathcal{N}(x_0, (\Delta x_0)^2 + e^2)(x) && (3.19) \end{aligned}$$

Of course using that means loosing some information as in this context it is not distinguishable between statistical error and real extend anymore. But this has the advantage that they do not have to be specified separately in the reference catalog.

It seemed reasonable to use the format of the output of the used source detection also for the reference catalog, as that is what is compared in NRTA-S[5]. The source detection output table and so also the reference catalog specification is described in appendix A.
The unique primary key identifying a source part is the tuple (ID, ID_INSTR, ID_BAND). In the source detection output, ID_INSTR specifies the telescope and ID_BAND the energy band. One source identified by it's ID can have multiple entries for different instruments. And each instrument/energy band combination can again have multiple entries. ID_INSTR=0 or ID_BAND=0 means that the entry specified a combined detection over multiple instruments or bands respectively. The NRTA-S ignores all entries with ID_INSTR $\neq 1$ and ID_BAND $= 0$, so (ID, ID_INSTR, ID_BAND) is unique with that. So is necessary to take care that the source detection performs the detection combined for all instruments.
ID_BAND is not used further, only the flux of a source is used to calculate energy information. It can be used as liked to specify source parts.

A possible reference catalog can be generated e.g. from the RASS bright source catalog (see Voges et al. (1999)).

3.5. Source detection

This section describes source detection by sliding boxes as used in the analysis software of the XMM-Newton project. The SASS uses an adapted version of the relevant tasks, but the algorithm is in principle the same (see also Brunner (2009)). The NRTA-S uses the tools from the F-SASS for source detection. The source detection in the SASS is accomplished in several steps:

1. expmap[6]: vignetted and unvignetted exposure maps creation, creates exposure maps (per energy band/telescope/total, vignetted/non-vignetted)

[5] With an additional column FLAGGED in the reference catalog specifying that the notification flag as described above is set if $\neq 0$

[6] The original tool from the XMM-Newton SAS is named eexpmask

2. ermask[7] creates an detection mask by using exposure maps. If the exposure of a pixel is above a specified cutoff, the detection mask of that pixel is set to 1, otherwise to 0. Alternatively, the exposure gradient can be used. Additionally, a blacklist can be applied, resulting in the detection mask to be set to zero for all pixels which are in that list. This tool is not used in NRTA-S, as blacklists can be generated by using the filtering feature.

3. erbox[8], run in local mode searches in the raw pixel data for sources by using the algorithm described below. The detection mask created by ermask selects which regions are analyzed.

4. erbackmap[9] firstly removes the sources detected by erbox from the data and weightenes it with the exposure. Then, two-dimensional splines of configurable order are fitted to the image. So a smooth background map is created.

5. erbox, at this time run in map mode again searches for sources using the background map which was generated by erbackmap.

6. ermldet[10] eROSITA ML PSF fitting tool, performs maximum likelihood PSF and extent fitsxx

The following descriptions are mainly a summary of information provided by the Users Guide to the XMM-Newton Science Analysis System, Issue 10.0 and Documentation of the XMM-Newton Science Analysis System 8.0.0.
Now, the algorithm implemented in erbox is presented. A sliding box is used, i.e. a connected subset of the cells of a two-dimensional lattice. Two such boxes are used, where one, the inner or source box, is a subset of the other, the background area.
Usually the used boxes are square, so let n denote the source boxes size, and m the background boxes size. The energy fractions α and β for the source respectively the background box are:

$$\alpha = \sum_{n \times n} PSF \tag{3.20}$$

$$\beta = \sum_{m \times m} PSF - \alpha \tag{3.21}$$

Secondary, the raw source box count c_s and the weighted raw background map c_b is calculated:

$$c_s = \sum_{n \times n} image \tag{3.22}$$

$$c_b = \frac{\left(\sum_{m \times m} image \right) - c_s}{m^2 - n^2} \tag{3.23}$$

[7]The original tool from the XMM-Newton SAS is named emask
[8]The original tool from the XMM-Newton SAS is named eboxdetect
[9]The original tool from the XMM-Newton SAS is named esplinemap
[10]The original tool from the XMM-Newton SAS is named emldetect

The source counts are corrected by applying the PSF-related energy fractions α and β and the background calculated from the outer box is subtracted. Analogously, a corrected background map is computed:

$$c_s' = \frac{c_s - c_b n^2}{\alpha - \frac{\beta n^2}{m^2 - n^2}} \tag{3.24}$$

$$c_b' = \frac{c_b - c_s' \beta}{m^2 - n^2} \tag{3.25}$$

Also, count number errors are determined by assuming a Poissonian statistics, i.e. the error is the square root of the number of counts, background and source count errors are propagated in the usual way.
The used detection likelihood function is $L = -\log p$, where p is the probability that a random fluctuation caused the observed number of counts in the source box or more, so the basic idea is to determine if a source sliding box contains significant excess from the surrounding background.
p is calculated using an regularized upper incomplete Gamma function which directly corresponds to the a cumulative Poisson distribution:

$$p = P(c_s, n^2 c_b) := \frac{\int_{n^2 c_b}^{\infty} dt\, t^{c_s - 1} \exp(-t)}{\Gamma(c_s)} \tag{3.26}$$

If source detection is performed over several energy categories, then the overall likelihood is generated using the incomplete Gamma function again to merge the single likelihoods L_i:

$$L = P(n, \sum_{i=1}^{n} L_i) \tag{3.27}$$

The sliding boxes are moved over all image pixels and energy bands. If the likelihood exceeds a given cutoff value, the results are included in the output.
Extended sources are detected by successively increasing the used box sizes.
The description so far is for the local mode of erbox. In the second run in the map mode it uses the background map instead of the raw pixel values in background boxes to get more precise results.
In the last step, ermldet performs a maximum likelihood PSF fit, which is simultaneously done for all energies and telescopes. Free fit parameters are source positions, extends and the count rates per energy band.
Adjacent sources are fitted simultaneously if their number does not exceed some cutoff value and if their PSF overlap. The fit is done with the Levenberg–Marquardt algorithm by optimizing the likelihood.
Sources which do not significantly improve the goodness of fit are rejected.
Further information about the algorithm can be found in Cruddace et al. (1988). There are other approaches for source detection, like Bayesian background source separation (Guglielmetti, 2010) or wavelet source detection (Valtchanov et al., 2001), which will be implemented in the SASS too. As the corresponding tools will all generate same structured output files, they can be easily integrated into the NRTA-S if needed.

3.6. Source identification

An approach how to assign given reference sources to a newly discovered source is given in this section. It is performed by finding a good selection of sources for a detected source. The NRTA-S should be able to detect if a measurement discovered new sources or it can be meaningful explained by prior knowledge. So it is required to assign measured sources to sources provided by the reference catalog.
The used hypothesis test described in 3.7 determines afterwards if such an identification is reasonable. If the results suggest that a new source is discovered it is also useful for further analysis to know which reference sources come into consideration if it is suspected that the source is not a new source.
It should be made clear that sources are modeled as sets of parts (see 3.4). The word "source" refers to such a part in this section.

3.6.1. Candidate selection

It is useful to select candidate sources from the reference catalog for detected sources to restrict the set of source which come into question. The source identification approach relies on such a preselection, because without, it would be far too performance expensive. And it is possible to suitable adjust the preselection so that it does not affect the results' accuracy.
The candidate selection is identical to the one presented by Pineau et al. (2010). They performed cross-correlation between a X-ray and an optical source catalog, but many concepts can be adopted here.
Per definition, a source is a candidate for another source, if the probability that they are located at the same position is equal or greater than a given critical value.
Let (α_1, δ_1) and (α_2, δ_2) be the positions of two sources. The errors on $\alpha_i \cos(\delta_i)$ and δ_i are given by σ_{α_i} and σ_{δ_i} respectively. In the general case, correlations between σ_{α_i} and σ_{δ_i} require an additional parameter, but fortunately, the SASS source-detection algorithm as well as the reference catalogs only provide a combined error, i.e. $\sigma_{\alpha_i} = \sigma_{\delta_i} =: \sigma_i$. That are the radii of the well-known error circles. If the errors would be different on the two axes, the circles turn into error ellipses and their handling is mathematically laborious, especially at the poles.
The spherical problem is transformed in a Cartesian plane one in the following way. The first source is located at the center $(0, 0)$ of the new frame, the second at $(d, 0)$, where d is the angular distance between the two sources, so the x-axis corresponds to a great circle connecting them. If the projection is equidistant, i.e. Euclidean distances in the new frame directly represent angular distances, then the position PDF of the first source is $\mathcal{N}(0, 0, \sigma_1^2)$ and that of the second source is $\mathcal{N}(d, 0, \sigma_2^2)$, where $\mathcal{N}$ are bivariate normal distributions without correlation and equal standard deviations for each axis:

$$\mathcal{N}(\mu_x, \mu_y, \sigma^2) := \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x-\mu_1)^2 + (y-\mu_2)^2}{2\sigma^2}\right) \tag{3.28}$$

The desired probability density that the two sources are located both located at the same position when the second source is located at (x, y) is given by

$$\int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' \mathcal{N}(0,0,\sigma_1^2)(x',y') \cdot \mathcal{N}(x,y,\sigma_2^2)(x',y') \tag{3.29}$$

$$= \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' \mathcal{N}(0,0,\sigma_1^2)(x',y') \cdot \mathcal{N}(0,0,\sigma_2^2)(x'-x,y'-y) \tag{3.30}$$

$$= (\mathcal{N}(0,0,\sigma_1^2) * \mathcal{N}(0,0,\sigma_2^2))(x',y') \tag{3.31}$$

$$= \mathcal{N}(0,0,\sigma_1^2+\sigma_2^2)(x',y') \tag{3.32}$$

A proof for the last equality can be found in Grinstead and Snell (2003). A maybe interesting additional information is that if $X \sim f$ and $Y \sim g$ are two independent random variables distributed by f respectively g, then $X + Y \sim f * g$ in the signal processing sense Hogg et al. (2012).
By switching to polar coordinates where the radial coordinate is $d(x, y) := \sqrt{x^2 + y^2}$, the probability density that an object at an angular distance of $d \geq 0$ from $(0, 0)$ is in fact at $(0, 0)$ is given by the Rayleigh distribution $\mathcal{R}(\sqrt{\sigma_1^2 + \sigma_2^2})$:

$$\int_{\{(x,y):|(x,y)|=d\}} dxdy\, \mathcal{N}(\sigma_1^2+\sigma_2^2)(x,y) \tag{3.33}$$

$$= \int_{\{(x,y):\sqrt{x^2+y^2}=d\}} dxdy\, \frac{1}{2\pi(\sigma_1^2+\sigma_2^2)} \exp\left(-\frac{x^2+y^2}{2(\sigma_1^2+\sigma_2^2)}\right) \tag{3.34}$$

$$= \int_0^{2\pi} d\varphi\, \frac{d}{2\pi(\sigma_1^2+\sigma_2^2)} \exp\left(-\frac{d^2}{2(\sigma_1^2+\sigma_2^2)}\right) \tag{3.35}$$

$$= \frac{d}{\sigma_1^2+\sigma_2^2} \exp\left(-\frac{d^2}{2(\sigma_1^2+\sigma_2^2)}\right) \tag{3.36}$$

$$=: \mathcal{R}(\sqrt{\sigma_1^2+\sigma_2^2})(d) \tag{3.37}$$

It is obvious that it makes no difference if the two sources are interchanged.
A potential question is why the maximum of this distribution is not at $d = 0$, but at $(0, 0)$ for the Cartesian version. The simple answer is that it is not the distribution that something is at a particular position (x, y), but that it is in an arc $\{(x, y) : |(x, y)| = d\}$, so one would have to divide by the arc's radius and calculate a limit value if necessary.
The probability that if the two sources are located at the same position, they have a distance greater than d between them is

$$\int_d^{\infty} \mathcal{R}(d) = \exp(-\frac{1}{2}\frac{d^2}{\sigma_1^2+\sigma_2^2}) =: k(d) \tag{3.38}$$

If $1 - k(d)$ is greater than a given significance level, candidates are rejected, otherwise retained. It is sufficient to calculate a critical value for $\frac{d^2}{\sigma_1^2+\sigma_2^2}$ instead of calculating the

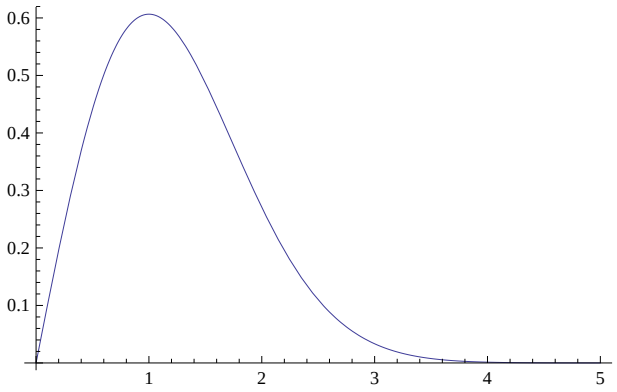


Figure 3.3.: PDF of $\mathcal{R}(1)$

whole integral every time, so by doing that the performance can be increased.
Of course it would be a waste of computing time to consider all reference sources. So, before doing the computation described so far, another preselection of sources is performed. Sources are only included if their angular distance does not exceed some configurable cutoff value.
This is accomplished by using a three-dimensional k-d tree in the way as Schmid (2012) did to select sources in a given circular region. A k-d tree is data structure which stores points located in R^k. It was introduced by Bentley (1975).
A k-d tree is a binary tree, i.e. a set of nodes where each node contains some specific data – here a source, in general at least the position of the node in R^k is required for k-d trees – and a reference to a left and right child node, both in the same tree. The child nodes are optional, e.g. leaf nodes do not have any of them. A node which is not the child of another node is called the root node.
A recursive procedure for the construction of the tree with N elements is described now. At first the list of elements is sorted by their first coordinates. Here this is done with the well-known quicksort algorithm.
Then the median element of the resulting ordered list $(e_0, \ldots, e_{N-1})$ is selected, i.e. $e_{floor(N/2)}$, whereupon floor is a function which rounds downwards it's argument if it is not integer. Afterwards the list is split into two new lists $(e_0, \ldots, e_{floor(N/2)-1})$ and $(e_{floor(N/2)+1}, \ldots, e_{N-1})$. Please note that lists can be empty when the median element is the first or last in the list. Furthermore the median element itself is not in any of the two lists.
The selected median element, also called pivot element, is the root node of the tree. Now recursion is performed by repeating the whole procedure for non-empty sublists, but now the elements are sorted by their second coordinates and in the next recursion level by their third coordinates and so on until there are no dimensions left. The again the first coordinate is used. This is done until both sublists are empty in one step.
Every time new sublists are generated the created root nodes of the subtrees created afterwards are the left respectively right child nodes of the current node.
The building of the tree runs in time $O(NlogN)$.
As recursion is difficult to describe with spoken language, here the function in pseudocode:

```
FUNCTION kdtree(list of points, depth)
```

```
BEGIN
 return null if list of points is empty
 axis = depth mod k
 sort list of points
 select median element
 create node
 node.data = median element
 node.left = kdtree(list of points left from the median element, depth+1)
 node.right = kdtree(list of points right to the median element, depth+1)
 return node
END
```

Because of the properties of k-d trees it is possible to perform a fast range search. Again that is done by a recursive function. The function has a parameter containing subtree node in which the search is currently performed and a parameter indicating the current recursion level. At the start that is the root node of course.
First it is checked if the current node is in the specified search range. If so, it is included in the result set. Analogous to the construction process the current axis is determined by calculating recursion level modulo the tree's dimension. Then it is checked which subtrees might contain nodes which are in the given search circle. For subtrees with match this criterion the function is recursively invoked. For example in when the current node is m in Figure 3.4 and the current axis is the x axis, the partition S_0 can be neglected as the search circle does not intersect the y axis and is not covered by S_0. It is obvious that at least one tree might contain requested nodes if both subtrees exist.
Lee and Wong (1977) showed that in the worst case where both subtrees have to be searched, the algorithm run in time $O(k \cdot N^{1-1/k})$.
Actually the sources' are located on an two dimensional surface of a sphere. But an 3-d tree is used by embedding that surface in the R^3:

$$x = \cos(\alpha)\cos(\delta) \tag{3.39}$$

$$y = \cos(\alpha)\sin(\delta) \tag{3.40}$$

$$z = \sin(\alpha) \tag{3.41}$$

The big advantage is that it is not necessary to take care about the boundaries of the domains of α and δ. The range search would be much more complicated otherwise, e.g. consider a source at $(0, 0)$. Then the search circle has to contain points not only close to $(0, 0)$, but also close to $(2\pi, \pi)$ as the k-d tree does not know anything like $(0, 0)$ has the same meaning as $(2\pi, \pi)$.

3.6.2. Maximum likelihood method

This section presents the basic idea behind the approach how to identify which combination of reference sources is the most likely choice to explain a measurement result. This is the first step for source identification and also necessary for the hypothesis tests which will be described later. It will be introduced why approximations are used, which are prerequisites for the understanding of the subsequent methods.
The challenge is to find a combination S of reference sources R which maximizes the likelihood of S when the result x of a measurement is given:

$$\max_{S \in \mathcal{P}_1(R)} L(S|x) = \max_{S \in \mathcal{P}_1(R)} P(x|S) \tag{3.42}$$

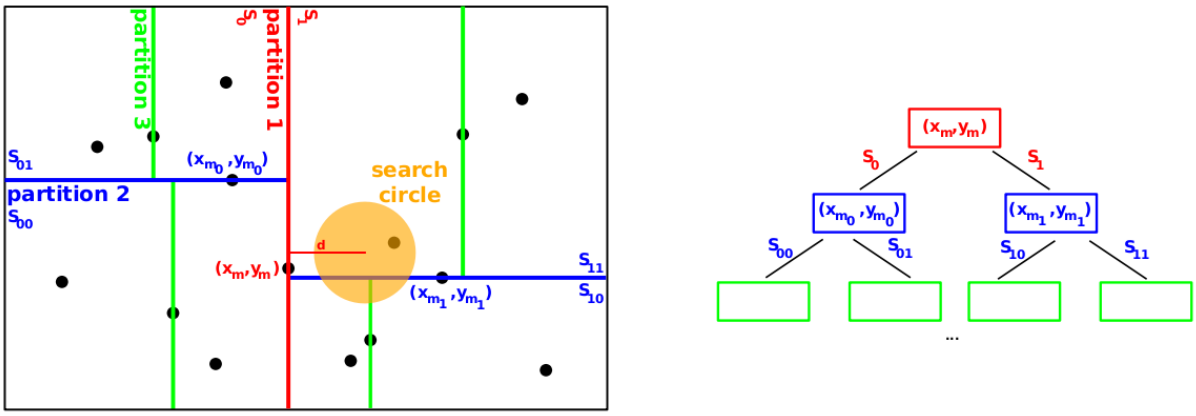


Figure 3.4.: "Structure of a 2-d tree. The total search volume containing the source positions (black dots) is iteratively divided in sub-volumes. Each partition of a volume is performed at the median of the subset of points with respect to either the x or y coordinate in an alternating sequence. The orange circle denotes the search circle [...]" Schmid (2012)

$\mathcal{P}_1(R) := \mathcal{P}(R) \setminus \emptyset$ denotes the power set of the set of reference sources R without the empty set. It is possible that this maximum is not unique, especially if the measurement does not cover the whole sky and therefore P is independent from sources which were outside the FOV during the whole exposure. So it is reasonable to ask for the smallest set S which optimizes L. For the unlikely case that even that is not unique, one should be randomly chosen.
The elements in R and S are reference sources, i.e. a set of source properties for which a photon distribution $PDF(S)$ and a total photon rate $I(S)$ can be specified as described in section 3.4.
The measurement result x should be considered as a set of readout events. One such event contains the images from the CCDs together with the exposure time and attitude information, i.e. the position and orientation of the FOVs.
$P(x|S)$ is a very complicated object which is basically a product of probabilities, each corresponding to one readout event:

$$P(x|S) = \prod_{e \in x} P(e|S) \tag{3.43}$$

The $P(e|S)$ are basically products of the single probabilities for each pixel value x_{ij} in the image provided by e:

$$P(e|S) = \prod_{ij} P(x_{ij}(e)|S, t(e), a(e)) \tag{3.44}$$

$P(x_{ij}|S, t, a)$ is the probability under the condition of a exposure time of t and an attitude of a. a is basically used for a coordinate transformation $T(a)$ of $PDF(S)$ to match the pixel coordinates. The photon distribution D on the detector if S is true is basically a convolution of $T(a)PDF(S)$ with the PSF. But other properties of the optics like the ones described by the FOV, ARF or the vignetting function are involved, too. Then to

get the probability of a pixel value, D has to be integrated over the pixel of interest. The result would be multiplied with the total photon rate $I(S)$ and the exposure time t to get the expected number of photons in the pixel. Then a Poisson distribution can be used to determine the desired probability.
Furthermore $P(x_{ij}|S,t,a)$ does not only depend on D, but also on the detector response, e.g. modeled by an RMF.
This explanations show how complicated such a quite exact approach would be. Another option would be to simulate the measurement often enough to approximate $L(S|x)$.
The NRTA-S is using the results of the source detection instead of raw data. They reflect the most likely parameters of a model of the photon distribution on the sky, but due to the data reduction, some information is lost. E.g. it does not directly supply information how likely an in fact existing source was not been detected. This is the reason why the NRTA-S does not recognize the disappearance of sources.
More pleasant is that the background is already subtracted by the source detection.
The measurement result x is lost due to the data reduction and replaced with the source detection results s. The question is how likely s is under the condition that S is true. Actually the exact calculation of $L(S|s)$ would be even more complicated than calculating $L(S|x)$ as if $s(x)$ denotes the result of the source detection for a measurement x, it is given by

$$L(S|s) = \frac{\sum_{x:s(x)=s} L(S|x)}{|\{x : s(x) = s\}|} \tag{3.45}$$

Methods how to get to some value which is related to $L(S|s)$ are described in the next sections.

3.6.3. KL divergence

Before the description of it's application for source identification, the so called KL-divergence is introduced now.
Consider N samples generated by a multinomial distribution and define n_i to be the number of samples in the i-th bin, so $N := \sum_{i=1}^{k} n_i$. For such multinomial histograms the likelihood $L(\underline{p}|\underline{n})$ that a measurement $\underline{n}$ came out of the distribution P is given by

$$L(\underline{p}|\underline{n}) = P(\underline{n}|\underline{p}) = N! \cdot \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!} \tag{3.46}$$

$L(\underline{p}|\underline{n}) \to 0$ in the limit $N \to \infty$, but the average likelihood $\overline{L} := L^{1/N}$ can be defined which does not surely converges to zero.
Now with the usage of Stirling's approximation $\log n! \approx n \log n - n$, $\log \overline{L}$ for large n, the

average log likelihood can be simplified:

$$\begin{aligned}
\lim_{N\to\infty} \log \overline{L} &= \log\left(N! \cdot \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!}\right)^{1/N} && (3.47)\\
&= \frac{1}{N}\left(\log N! + \sum_{i=1}^{k} \log \frac{p_i^{n_i}}{n_i!}\right) && (3.48)\\
&= \frac{1}{N}\left(\log N! + \sum_{i=1}^{k} n_i \log p_i - \sum_{i=1}^{k} \log n_i!\right) && (3.49)\\
&= \frac{1}{N}\left(N\log N - N + \sum_{i=1}^{k} n_i \log p_i - \sum_{i=1}^{k} (n_i \log n_i - n_i)\right) && (3.50)\\
&= \log N - 1 + \frac{1}{N}\sum_{i=1}^{k} n_i \log p_i - \frac{1}{N}\sum_{i=1}^{k} (n_i \log n_i - n_i) && (3.51)\\
&= \log N - 1 + \frac{1}{N}\sum_{i=1}^{k} n_i \log p_i - \frac{1}{N}\sum_{i=1}^{k} n_i \log n_i + \frac{1}{N}\sum_{i=1}^{k} n_i && (3.52)
\end{aligned}$$

According to the strong law of large numbers, it can be assumed that there is a probability q_i so that n_i is it's expected value $n_i = q_i \cdot N$ in the investigated limit:

$$\begin{aligned}
\lim_{N\to\infty} \log \overline{L} &= \log N - 1 + \frac{1}{N}\sum_{i=1}^{k} q_i N \log p_i - \frac{1}{N}\sum_{i=1}^{k} q_i N \log q_i N + \frac{1}{N}\sum_{i=1}^{k} q_i N && (3.53)\\
&= \log N - 1 + \sum_{i=1}^{k} q_i \log p_i - \sum_{i=1}^{k} q_i \log q_i N + \sum_{i=1}^{k} q_i && (3.54)\\
&= \log N + \sum_{i=1}^{k} q_i \log p_i - \sum_{i=1}^{k} q_i \log q_i N && (3.55)\\
&= \log N + \sum_{i=1}^{k} q_i \log p_i - \sum_{i=1}^{k} q_i \log q_i - \log N \sum_{i=1}^{k} q_i && (3.56)\\
&= \sum_{i=1}^{k} q_i \log p_i - \sum_{i=1}^{k} q_i \log q_i && (3.57)\\
&= \sum_{i=1}^{k} q_i \log \frac{q_i}{p_i} && (3.58)\\
&= D_{KL}(q||p) && (3.59)
\end{aligned}$$

$D_{KL}(q||p)$ is called the Kullback–Leibler (KL) divergence from p to q, originally investigated by Kullback and Leibler (1951).
The presented derivation is similar to that ones in Shlens (2007) and Nowak (2009), where in last also the neat comment that "[t]he central intuition is that the KL divergence effectively measures the average likelihood of observing (infinite) data with the distribution [q] if the particular model [p] actually generated the data." can be found.

Besides this interpretation with probability theory, the KL-divergence has applications in information theory, where it measures the information loss if a distribution is used to approximate another.
If q and p are continuous, the sums become integrals, but the statistical meaning is analogous.

$$D_{KL}(q||p) = \int_{-\infty}^{\infty} d\underline{x} \quad q(\underline{x}) \log \frac{q(\underline{x})}{p(\underline{x})} \tag{3.60}$$

The KL divergence is obviously not symmetric in general [11]. One interesting property is also that it is ≥ 0 for all p, q and it is $= 0$ iff. $p = q$.

The KL divergence is also closely related to entropy, so it is often called relative entropy:

$$\mathbb{E}_{X\sim Q}[\log P_\Theta(X)] = -\left(-\sum_X Q(X) \log P_\Theta(X)\right) \tag{3.61}$$

$$= -\left(\underbrace{-\sum_X Q(X) \log Q(X)}_{H(Q)} + \underbrace{\sum_X Q(X) \log\left(\frac{Q(X)}{P_\Theta(X)}\right)}_{D_{KL}(Q||P_\Theta)}\right) \tag{3.62}$$

where H_Q is the entropy of Q and $\mathbb{E}_{X\sim Q}[\log P_\Theta(X)]$ the expected value of $[\log P_\Theta(X)]$ over the distribution Q, also called cross entropy.
In NRTA-S the KL divergence compares distributions which are mixture Gaussians, i.e. a sum of bivariate normal distributions. There is no analytic closed form for it, but it can be calculated up to arbitrary accuracy by performing Monte Carlo simulations as done here:

$$\begin{aligned}
\mathbb{E}_{X\sim H_1}[\log \Lambda(X)] &= \int d\underline{x} H_1(\underline{x}) \log \frac{H_0(\underline{x})}{H_1(\underline{x})} && (3.63)\\
&= -\int d\underline{x} H_1(\underline{x}) \log \frac{H_1(\underline{x})}{H_0(\underline{x})} && (3.64)\\
&= -D_{KL}(H_1||H_0) && (3.65)
\end{aligned}$$

It can be shown that the KL-divergence and so the algorithm is convergent for mixture Gaussians Hershey and Olsen (2007).

3.6.4. Source matching by KL divergence optimization

Firstly it is assumed that both, the considered reference sources S and the detected sources s are known exactly, i.e. there are no errors.
The key idea is to approximate the maximization of $L(S||s)$ by maximizing something like $D_{KL}(PDF(s)||PDF(S))$ which is directly related to average log likelihood maximization

[11] There are some exceptions, the most prominent example is the KL divergence between two normal distributions with same variance.

as described in the last section. Maximizing the log likelihood is equivalent to maximizing the plain likelihood as all logarithms are monotonically increasing. So it is also not important which base the logarithm has. In information theory it is mostly 2, but here just the more common choice in Physics e was made.
The photon distributions $PDF(S)$ and $PDF(s)$ alone are not sufficient, as the total photon rate has to be compared, too. They are just multiplied with the PDFs of total photon rates. Without errors, these PDFs are delta distributions $P(I|s) = \delta(I(s) - I)$.
To get the errors and parameter extends involved, the expected photon distributions as described in 3.4 could be used and $P(I|s)$ replaced by a Gaussian distribution. But it is possible that the errors in the source detection are smaller or bigger than in the reference catalog which results in that the wrong models are compared. That is because the used photon distributions are not reflecting measurement probabilities directly. So another approximation is done by creating some sort of a combined error.
One reference source and one single detected source are considered. Under the assumption that the source detection really measured a point source which is known exactly in the reference catalog, it is reasonable to convolve the delta distribution with the Gaussian describing the positional error of the source detection's result to adapt the reference model. This means changing the reference photon distribution to something which is assumed to be similar to possible outcomes of the source detection if it holds. This is done with all errors.
Contrariwise if a point source was detected without any statistical errors (impossible, here only considered as an example) and the reference catalog contains errors, it should be convoluted with the reference catalogs error too.
If multiple sources are investigated, it is suggested here to use a weighted average of errors in a region around each source to determine the Gaussians for the convolutions. It is not only of interest which reference sources were detected, but also which sources belong to one specific detected source. For now, only single detected sources are considered at once. This is again an approximation, because e.g. if a reference source was resolved into two detected sources by the measurement, it belongs to both of them.
Only the selected candidate sources are used as possible reference sources for a detected source.
This considerations lead to the models Q for the measurement $(\alpha_0, \delta_0, \sigma, e, \Delta e, r_0, \Delta r, E_0, \Delta E)$

and P for the candidate reference sources $(\alpha_i, \delta_i, \sigma, e_i, \Delta e_i, r_i, \Delta r, E_i, \Delta E_i)$:

$$\sigma_{ref}^2 = \left(\frac{\sum_i c_i r_i \sigma_i}{\sum_i c_i r_i}\right)^2 \tag{3.66}$$

$$e_{ref}^2 = \left(\frac{\sum_i c_i r_i e_i}{\sum_i c_i r_i}\right)^2 \tag{3.67}$$

$$(\Delta E_{ref})^2 = \left(\frac{\sum_i c_i r_i \Delta E_i}{\sum_i c_i r_i}\right)^2 \tag{3.68}$$

$$(\Delta r_{ref})^2 = \left(\sum_i c_i \Delta r_i\right)^2 \tag{3.69}$$

$$\begin{aligned} Q(\alpha, \delta, r, E) &:= \mathcal{N}(\alpha_0, \sigma^2 + e^2 + \sigma_{ref}^2 + e_{ref}^2)(\alpha) \cdot & (3.70)\\ &\quad \mathcal{N}(\delta_0, \sigma^2 + e^2 + \sigma_{ref}^2 + e_{ref}^2)(\delta) \cdot & (3.71)\\ &\quad \mathcal{N}(r_0, (\Delta r)^2 + (\Delta r_{ref})^2)(r) \cdot & (3.72)\\ &\quad \mathcal{N}(E_0, (\Delta E)^2 + (\Delta E_{ref})^2)(E) & (3.73)\\ P_i(\alpha, \delta, E) &:= \mathcal{N}(\alpha_i, \sigma_i^2 + e_i^2 + \sigma^2 + e^2(\alpha) \cdot \mathcal{N}(\delta_i, \sigma_i^2 + e_i^2 + \sigma^2 + e^2)(\delta) \cdot & (3.74)\\ &\quad \mathcal{N}(E_i, (\Delta E_i)^2 + (\Delta E)^2)(E) & (3.75)\\ P(\alpha, \delta, r, E) &:= \frac{\sum_i c_i r_i P_i(\alpha, \delta, E)}{\sum_i c_i r_i} \cdot \mathcal{N}\left(\sum_i c_i r_i, (\Delta r_{ref})^2 + (\Delta r)^2\right) & (3.76) \end{aligned}$$

The contributions are given by the rates r_i and additional specific weights c_i. For the source identification process, where only one detected source is considered at once and where it is assumed that a reference source either contributes completely or not at all, the c_i are just set to 1, for the hypothesis tests this will be refined.
The average of rate errors $(\Delta r_{ref})^2$ needs special treatment as the rate distributions are combined to a total rate distribution. Due to the linearity of the total rate $\sum_i c_i r_i$, their error is just given by eq. 3.69.
It should be mentioned that later it will be necessary to allow reference sources to not contribute completely, i.e. there are $c_i < 1$. The distribution of the total rate is using the absolute contributions c_i to reflect that. But the P_i are weighted using relative contributions to get a normalized PDF.
The source detection provide a source flux $f \pm \Delta f$ instead of an energy E, but E can be easily calculated with $E = f/r$. ΔE could be determined by performing simple error propagation by approximating $\Delta E = E(f + \Delta f, r + \Delta r) - E(f, r)$. But then the errors on f and r can cancel out each other which maybe results in $\Delta E = 0$. They are calculated with $\Delta f/r$ instead.
The source identification is done by minimizing $D_{KL}(Q||P)$ for combinations of candidate sources. If two are equal, the first one found will be selected as it will not affect the hypothesis tests significantly. If there are many candidates, the set of choices which otherwise would have a power of $2^n - 1$ is restricted. This is done with a cutoff value which specifies the maximum number of counterparts of one source.
So P are sums of Gaussians, i.e. mixture Gaussians. If a spectrum is provided, one source consists of multiple parts which are treated like different sources here. It should be remembered that transients are modeled as a set of possible distributions. Here this

distributions are treated like independent sources to find the state in which the source currently is.
The used approach is inspired by the ideas of Goldberger et al. (2003), who used the KL divergence to measure the similarity of images.
Very important is that the source identification is not the final product, it is possible and desirable that the hypothesis tests performed afterwards will reveal that the identification is wrong.

For illustration, some simple examples are discussed. Figure 3.5 shows some simplified

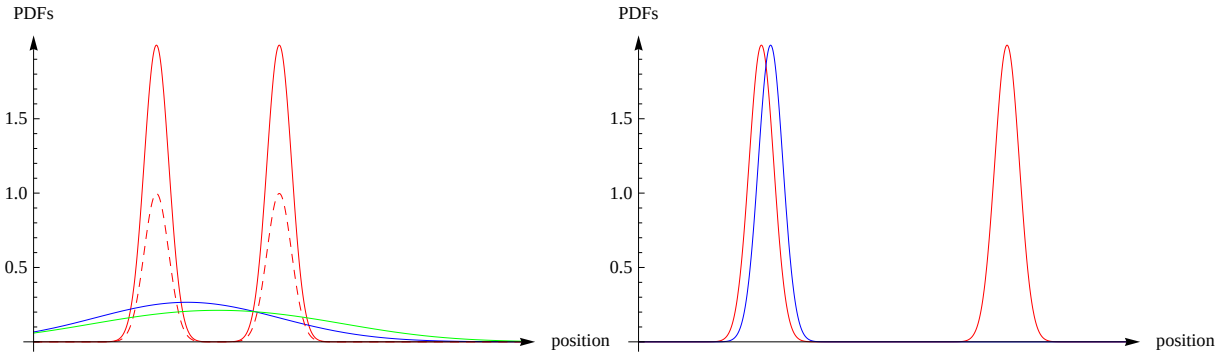


Figure 3.5.: Simplified example distributions. Red ones are candidate reference sources, dashed red are the result of the KL divergence optimization and blue are detected sources. Left two reference sources could not been separated, right only one of two reference was detected.

one-dimensional position PDFs. Other parameters are ignored here. The red ones are candidate reference sources and the blue ones are detected sources. It should be noted that the candidates were not summed yet. Consider that all other parameters are described by delta distributions which are identical for all sources. The shown dimension could be e.g. the declination of the sources.

Example 1 In the left diagram it seems likely that the two sources could not been separated by the measurement. The KL divergence optimization results in that $1/2PDF_1 + 1/2PDF_2$ is the best reference PDF, in the figure shown dashed red. That is because if the errors are convoluted with the measurement error, the resulting PDF (green line) does not allow the separation the two sources any more and it is very similar to the PDF of the detected source. No further weighting is necessary as the two sources belong to only one detected source.

Example 2 Now it is assumed that in the left plot the blue ones are the candidates and the red ones the detected sources the situation is much different. In this simple case no optimization is necessary, because there is only one candidate. But the contributions of candidate to the two detected sources are different. The cross correlation, which will be described in the next section, will result in that the candidate contributes a little bit more to the left one as it is nearer.

Example 3 In the right graphic a reference source could not have been detected, but the detected source matches quite good to the other candidate. The optimization

will just drop the not detected source.

To confirm that the software really generates the qualitatively described results, the tool was run with inputs corresponding to the examples. See chapter 5 for details.

3.6.5. Cross correlation

One might consider to maximize the cross correlation between P_Θ and Q, i.e.

$$\overline{\Theta}_{CC} = \arg\max_{\Theta}(P \star Q)(0) \tag{3.77}$$

$$=: \left[\arg\max_{\Theta} \int_{-\infty}^{+\infty} d\underline{x} p(\underline{x}) q(\underline{t} + \underline{x})\right]_{\underline{t}=0} \tag{3.78}$$

$$= \arg\max_{\Theta} \int_{-\infty}^{+\infty} d\underline{x} p(\underline{x}) q(\underline{x}) \tag{3.79}$$

Here Θ denotes a combination of candidate sources. It should be noted that the usual time-offset parameter is set to 0 here, as it's use in time series analysis is replaced by the parameter Θ.
To see why that approach fails here becomes clear if it is interpreted by probability theory. The cross correlation of two PDFs P and Q gives the distribution of $X - Y$, where $X \sim P$, $Y \sim Q$, in this case where only $t = 0$ is considered, the infinitesimal probability for $X = Y$, which equals $E_{X\sim P}[P(X)]$.
The interpretation of such a infinitesimal probability makes only sense if it is compared with another such probability, e.g. with a likelihood quotient. The KL divergence calculates the expected value of such a quotient (see 3.7.2). If the cross correlation is optimized, no likelihood in the required sense is maximized, rather an unphysical assumption that the two PDFs could be equal is applied.
For example, if two reference sources could not been resolved by the used instrument, it is obviously wrong to select only the nearest reference source which would be done.

Although the cross correlation is used to estimate how much a reference source contributes to matching detected sources. The expected value of $Q(X)$ if $X \sim P$ is given by:

$$w_i := \mathbb{E}_{X\sim P}[Q_i(X)] = \int_{-\infty}^{+\infty} d\underline{x}\, P(\underline{x}) Q(\underline{x}) = (P \star Q)(0) \tag{3.80}$$

Here Q_i is like in eq. 3.70, but for a detected source i. P is set so that only the desired reference source contributes.
As the P and Q are multivariate Gaussians without correlation in this specific case, they are just products of Gaussians. Gaussians have the following symmetry:

$$\mathcal{N}(\mu, \sigma^2)(-x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(-x-\mu)^2}{\sigma^2}\right) \tag{3.81}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x+\mu)^2}{\sigma^2}\right) \tag{3.82}$$

$$= \mathcal{N}(-\mu, \sigma^2)(x) \tag{3.83}$$

Because of that, the cross correlation between two multivariate normal distributions (μ_1, Σ_1) and (μ_2, Σ_2) without correlation can be expressed by a convolution:

$$\mathcal{N}(\mu_1, \Sigma_1^2) \star \mathcal{N}(\mu_2, \Sigma_2^2) = \mathcal{N}(\mu_1, \Sigma_1^2) * \mathcal{N}(-\mu_2, \Sigma_2^2) \tag{3.84}$$

Vinga (2004) has shown that

$$\mathcal{N}(\mu_1, \Sigma_1^2) * \mathcal{N}(\mu_2, \Sigma_2^2) = \mathcal{N}(\mu_1 + \mu_2, \Sigma_1^2 + \Sigma_2^2) \tag{3.85}$$

That formula is used to calculate the cross correlation.
The final weighting $\overline{w}_i$ is calculated by normalizing:

$$\overline{w}_i = \frac{w_i}{\sum_j w_j} \tag{3.86}$$

It has some meaning again because two infinitesimal probabilities are divided.

3.6.6. Implementation

The entry point of NRTA-S is the candidate selection tool "csel". If started, it builds a kd-tree for the reference catalog.
The kd-tree implementation is based on that of Schmid (2012), but a little bit extended so it can be used for other purposes that for simulation.
It's input is the source detection's output file and the NRTA-S catalog, for which again a kd-tree will be build for each run. For each detected source, a range search in both, the reference catalog and the NRTA-S catalog is performed. The p-value for all found reference source is compared with a specific significance and the mapping is stored in the HDU "CANDIDATES" if it is high enough. All reference sources which are assigned to any detected sources are copied in the HDU "REFERENCE_SOURCES", the detected source in "DETECTED_SOURCES".
"sid" performs the described source identification for all detected sources. It calculates the KL divergence for all combinations of candidate sources and selects that with the lowest. The maximum number of sources in such a combination can be specified.
The KL divergences are calculated according to eq. 3.63 by sampling from Gaussians with methods of the GSL. It is possible that the likelihood quotient can not be calculated for some samples due to numerical problems when the probabilities are too low. In this case additional samples are generated up to a maximum total number of samples[12].
If no sample was valid or if there are no candidates for a given source, no source identification result is written, otherwise it is inserted in the HDU "IDENTIFICATION".
In addition, "sid" calculates the cross correlations as described in 3.6.5.
Details on the output data products can be found in appendix D, details on the input parameters in appendix B.

3.7. Hypothesis testing

Hypothesis tests are performed after the source identification. The objective is to determine if something new was discovered, i.e. the detected source is actually a new source.

[12]The maximum number of samples can be configured with the parameter "MAX_MC_CALLS", the minimum number with "MIN_MC_CALLS"

A classification of the results is done afterwards. The matched reference source models (P in eq. 3.76) are tested against the detected source's models (Q in 3.70). But this time, the c_i are the results of the cross correlations as described in section 3.6.5.

3.7.1. Neyman-Pearson lemma

Neyman and Pearson (1933) investigated which tests are the most powerful ones[13]. Their research yielded in the nowadays well-known Neyman-Pearson lemma, which states that likelihood-ratio tests (LRT) between two point hypotheses are most powerful, i.e. when the null-hypothesis H_0 is rejected in favor of H_1 for a given sample x and significance α, if

$$\Lambda(x) = \frac{L(H_0|x)}{L(H_1|x)} \leq \eta \tag{3.87}$$

$$P(\Lambda(X) \leq \eta \,|\, X \sim H_0) = \alpha \tag{3.88}$$

That is a nice and mighty theorem, but unfortunately determining the critical value η is difficult in general, but an approach for arbitrary distributions is presented in the next section. So it's major application is to approximate special problems. For example the χ^2-tests are usually derived by this way.
Mostly the log likelihood quotient $\log \Lambda(x)$ is used, here too.

3.7.2. Monte-Carlo method

Dimitrov et al. (2003) proposed a "procedure for calculating critical level and power of likelihood ratio test[s], based on a Monte-Carlo simulation method", which is the basis of the algorithm presented here.
The main challenge is to get the cumulative distribution function $\alpha(\eta)$ in equation 3.88. This can be done by Monte Carlo by generating a sequence of N iid. random samples $(x_1, \ldots, x_n)$ distributed by H_0.
Then the value w_i of Λ is calculated for each sample:

$$w_i = \log \Lambda(\underline{x}) = \log \prod_{i=1}^{n} \frac{H_0(x_i)}{H_1(x_i)} = \sum_{i=1}^{n} (\log H_0(x_i) - \log H_1(x_i)) \tag{3.89}$$

The first transformation is correct because the samples are independently generated. Please note that here the log likelihood is used because numerical reasons as the value of Λ can be very small.
Afterwards the empirical distribution is generated, which will converge to the needed distribution for large N:

$$\overline{\alpha(\eta)} = \frac{\text{number of } w_i\text{s less or equal } \eta}{N} \tag{3.90}$$

The counting is done be first sorting the w_is and then iterating it.
Then the p-value of the test which is directly connected to significance can be calculated:

$$\text{p-value} = 1 - \overline{\alpha(\log \Lambda(\underline{x}))} \tag{3.91}$$

[13] A hypothesis test is called most powerful, if it's a test which has the highest probability of rejection of the null-hypothesis compared to all other possible tests.

In the context used here, there is no sample x, but an PDF H_1 corresponding to the model of a detected source. Hence a single sample size of $n = 1$ is used and the $\log \Lambda(\underline{x})$ is replaced with the expected value $\mathbb{E}_{X\sim H_1}[\log \Lambda(X)]$ of the likelihood quotient, which can be calculated with the KL divergence:

$$\begin{aligned} \mathbb{E}_{X\sim H_1}[\log \Lambda(X)] &= \int d\underline{x} H_1(\underline{x}) \log \frac{H_0(\underline{x})}{H_1(\underline{x})} && (3.92) \\ &= -\int d\underline{x} H_1(\underline{x}) \log \frac{H_1(\underline{x})}{H_0(\underline{x})} && (3.93) \\ &= -D_{KL}(H_1||H_0) && (3.94) \end{aligned}$$

3.7.3. Implementation

The tool "htest" requests the output of "sid" as input. In performs a hypothesis test for each detected source.
It is checked if there are candidates for a given detected source. If there are none, a p-value of 0 is written for the detected source immediately. Elsewise, $t := \mathbb{E}_{X\sim H_1}[\log \Lambda(X)]$ is calculated using the methods for the KL divergence which were already described in the implementation of the source identification.
Afterwards N samples are generated according to H_0[14]. A rate can be easily sampled from a Gaussian distribution with the GSL. The position and energy is sampled by first selecting a P_i according to their relative probabilities $c_i r_i / \sum_i c_i r_i$ and then sampling from Gaussians. The log likelihood quotient of each sample is inserted in a list.
When the list is complete, it is sorted and iterated to determine the desired empirical distribution which is stored as a list of pairs (number of w_i's $\leq \eta, \eta$). Then the p-value is calculated as described in the last section and written to the output table.
Details on the output data products can be found in appendix D, details on the input parameters in appendix B.

3.8. Automated classification of results

In the final step, the tool "rclass" categorizes and rates the results of the previous tools and creates user-notifications.

3.8.1. Alert generation and rating

The tool which performs the classification of the hypothesis tests results creates alerts which will afterwards be further filtered and rated. The filtering also involves plausibility checks. Another purpose of the classification tool is to insert new sources and parts in the internal NRTA-S catalog if the corresponding alerts passed the plausibility checks. This catalog has the same format as the reference catalog. The reason for this catalog is to prevent the alert creation for already handled events and an easy access to the data of newly detected sources. The source identifiers (ID + energy band) are expected to be

[14] N can be specified with the parameter PIL-parameter MC_CALLS.

unique over both, the NRTA-S and the reference catalog. Be convention, NRTA-S sources have negative IDs and reference source positive ones, so it can be easily determined which catalog contains a given source.
The hypothesis tests checks if is reasonable to reject the reference catalog for detected sources. Rejecting it means to assume that a new source was detected. The cut-off for the tests' p-value, i.e. the significance can be specified with a parameter (see appendix B).
When a reference catalog does not fit to new measurements anymore that can be caused by some reasons. The catalog will not include all known sources. Maybe the telescopes whose measurements were the basis of the catalog were not able to spatially or spectral resolve sources. Also, their sensitivity in some energy bands could have been different than that of eROSITA. At last, eventually some regions of the sky were not or too shortly exposed. It could be considered to declare all such things which involved the measurements in the reference catalog, but that would complicate things.
All that things are caused because of the measurements which were the basis for the reference catalog. Another category of reasons is that really something on the sky changed. The NRTA-S is not able to discriminate between them.
There are two possibilities, either the properties, mainly the spectrum of a source changed or a new source was newly illuminated.
As already described, sources are considered to be conglomerates of parts and it is not investigated by NRTA-S if sources or parts of them went off, so it can only be detected if a source's spectrum got additional components.
It is not possible to distinguish between if a new source or new parts of a source were detected. In both cases, an alert of type "NEWSOURCE"' will be thrown.
When a source changes it's state that can be known in the reference catalog or unknown. If it is known, it will contain an entry for each known state. It depends on flags if the transition will be reported. For sources whose lightcurves are good known it is not of interest. But sources which were on in the past but then went off without a measurement of a additional illumination that is a very interesting fact. If the set of identified sources for a detected source contains a flag an alert with flag STATECHANGE is thrown.
If multiple detected sources belong to one reference source according to hypothesis test, they are flagged with RESOLVED to indicate that the measurement was able to resolve a known source into multiple parts (position and/or spectrum). Also that occurrence leads to the insertion in the NRTA-S catalog.
Besides the rough classification into the main categories

- NEWSOURCE
- STATECHANGE
- RESOLVED,

a classification catalog is used. It has basically the same format as the reference catalog, but some additional columns for classification and rating. Again, the detected new source parts are identified with sources in it. This time the position is ignored and only one catalog source is investigated. This faciliate an easy classification definition. The classification catalogs source ID is used for the subcategory identifier and a human-readable description should be provided. If the classification fails because of numerical problems

when no classification source fits enough, the artificial ID 0 with description "unknown" is used instead.
For each classification source, a rating value has to be specified in the table. It will be multiplied with the ratings of the main categories which can be configured using PIL.
For the format of the classification catalog see appendix A.

3.8.2. Plausibility checks

It is possible that the null hypothesis is rejected because of incorrect input data and not because statistics. The most obvious eventuality for such an error is the failure of the attitude prediction due to an intended orbit or attitude correction initiated by the earth station. It is impossible to correct such an error as it would require exact information about the maneuver which are not scheduled to be transmitted to the NRTA. Also attitudes which are not predicted potentially can be wrong. There are other potential sources of errors, too.
When such an error occurs the NRTA would generate a huge amount of unjustified alerts. That happened on an other NRTA software used in the past for a different X-ray mission. To avoid that, a simple plausibility check is made.
If in a region of the sky are too much alerts in one run, the alerts in that region are rejected, but an alert of type PLAUSIBILITY is thrown. Such an alert is generated at most once per run. As it is probable that an error affects more than one alert, this procedure seems reasonable. Please note that the input product of the filter remains untouched, so inappropriately refused alerts are not lost.
If a run generated too much alerts, i.e. if their number exceeds a configurable threshold[15], the alert is rejected.
As there can be alerts which are so grave that they should never been rejected, another threshold can be defined[16]. If the threshold of the alert is larger than that, it will be ignored by the plausibility check.
At this point it should be noted that at least an attitude offset could be detected in principle, e.g. by maximizing the cross correlation between a reference catalog and the measurement. It is not gone further into it here, as it requires more detected sources per run than expected to work. Further information is available in the documentation of the tool "eposcorr" of the XMM-Newton Science Analysis System (SAS)[17]. If really needed, that tool could be adopted for the NRTA in the future – maybe it could also be useful for the final analysis.

3.8.3. Alert filter, database and notifications

The filter checks if the coordinates or other properties of a detected source belonging to an alert fall into regions denied by a blacklist.
The blacklist will not cause the rejection of all alerts, instead it specifies a rating cutoff below which the alerts falling in the criteria will be rejected.

[15] configuration parameter REJECTION_THRESHOLD
[16] column KEEPING_RATING_THRESHOLD
[17] available at `http://xmm.esa.int/sas/10.0.0/doc/eposcorr/node4.html`

Blacklist entries are defined in a FITS-table according to the specification in table C.1. Each entry defines a coordinate range $(\alpha, \delta) : \alpha_min \leq \alpha \leq \alpha_max, \delta_min \leq \delta \leq \delta_min$, an energy range $[E_min, E_max]$, an rate range $[r_min, r_max]$ and the noticed rating cutoff. All source properties have to be in the corresponding ranges for the activation of the filter.

All alert are inserted in the alert database event the rejected ones which are accordingly flagged. The database table is specified in table D.
If an alert passed all filters, a script will be executed. It gets the alert data via command line arguments (see table D.3).
Currently only a simple notification by mail is implemented exemplary.

4. User interface

All data is stored in FITS files, so the users just need access to them. This can be enabled by providing classical accounts on the server where the files are stored. But it is maybe unhandy if just a quick look on the data is required or external persons should get guest access to some data. With the proceedings of web technology it became possible to create rich graphical user interfaces (GUIs) which run in almost any modern browser without additional plugins by using the standardized HTML format and JavaScript. It requires no installation and is usable immediately from every computer which has network access to the server which runs the tools' backend. So a web-based application for viewing and writing FITS files was created.
Because such a tool can be of use for other purposes too, it was created as an application which is independent from the NRTA. It is named JavaScript FITS Viewer (JsFv), where the letters Fv indicate that it was inspired by the well-known Fv[1]. Figure 4.1 shows a screenshot of the GUI.
The frontend of the web interface is based on ExtJS, a comprehensive JavaScript-framework which includes not only GUI components but also implements the Model-View-Controller (MVC) design pattern and has support for AJAX together with an API which allows an easy access to data stores located at the server. JsFv makes use of all of that features.
It does not support all capabilities of the FITS format, but is able to handle the files generated or needed by the software described here. For further information refer to the documentation of JsFv. At the moment, it also does only provide read only access to the files.
Although not really needed for NRTA, some diagrams can be created. The API which displays them is Highstock, which was originally designed for stock exchange diagrams. A possible application for NRTA could be the plotting of housekeeping information like device temperatures.
The backend is object-oriented written in Perl and uses some modules, e.g. one which provides JSON support, which is used for the AJAX communication between client and server. Another examples are wrappers to the CFITSIO and the ImageMagick libraries for accessing FITS files and generating images respectively.

[1]Fv is an interactive FITS file editor which is part of the FTOOLS software developed by the NASA

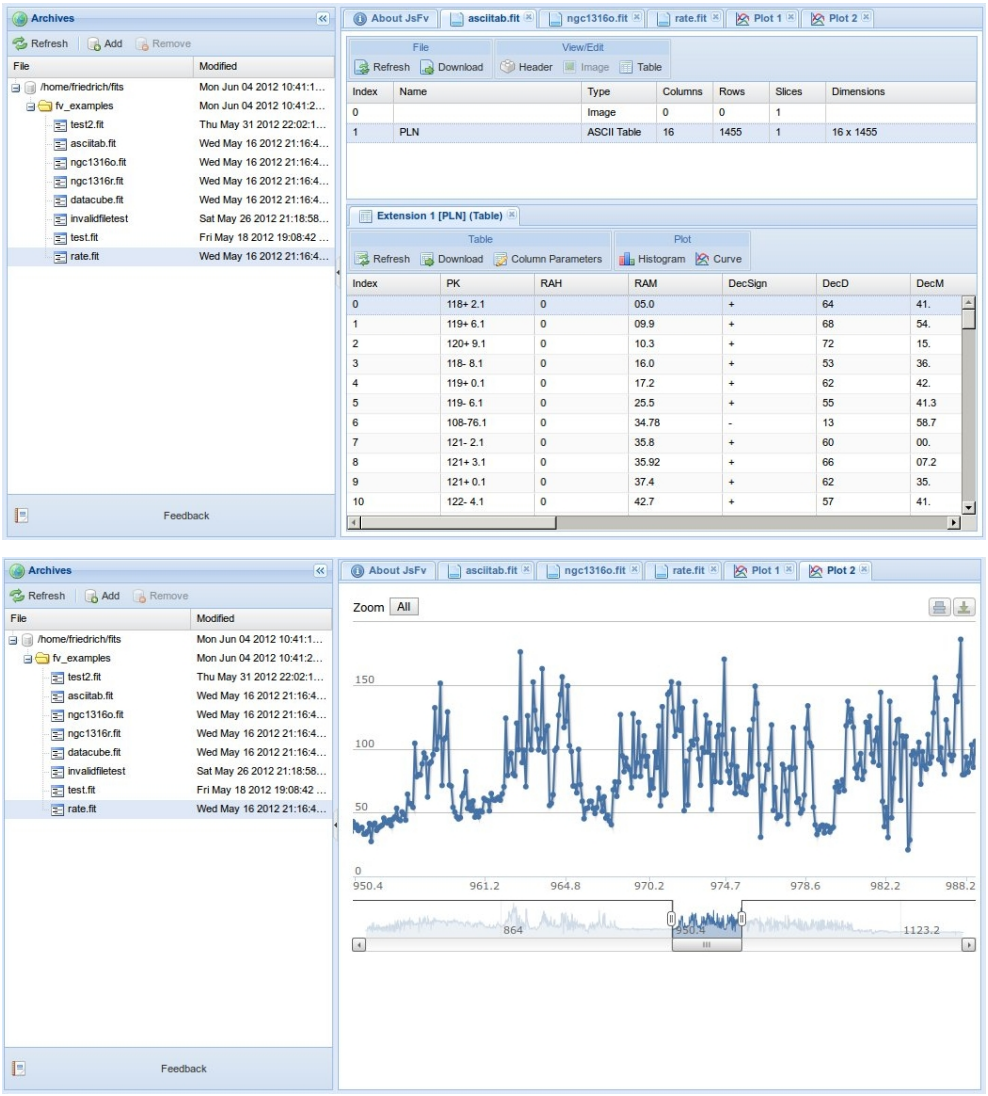


Figure 4.1.: Screenshots of the GUI[2]

[2]Because of copyright reasons it should be noted that some used icons are from famfamfam.com.

5. Simulations and testing

For the validation of the software and to determine their expected usability during eRASS, simulations and tests were performed.
Schmid (2012) created an extensive and adaptable software for simulating X-ray telescopes. It samples photons for sources specified in a SIMPUT-catalog (Schmid et al., 2011). After the photon generation, the telescope's mirror and detector models are applied. The models are highly configurable and even background and detector response models can be simulated. Furthermore it enables the user to supply attitude files, so e.g. it can be used for whole surveys like eRASS.

The NRTA-S is intended to analyze only one eroday in one run. As described in section 3.2, there is no overlap between the exposed slices during one eroday, so for the simulation it is sufficient to cover only an area of 2x2 degrees. This is a little bit bigger than needed as the FOV has a diameter of around one degree. The attitude proceeds uniformly with the speed specified in eq. 3.1 over a declination interval of 10 degrees while keeping the right ascension fixed. According to the length of one eroday the exposure time is 400s. The simulated data was cut to the addressed area size at a position where the slew of the FOV is fully covered.
Figure 5.1 shows the simulated exposure map.
As a first test, sources with fluxes between 10^{-8} and 10^{-15} $erg\,s^{-1}\,cm^{-2}$ were simulated.

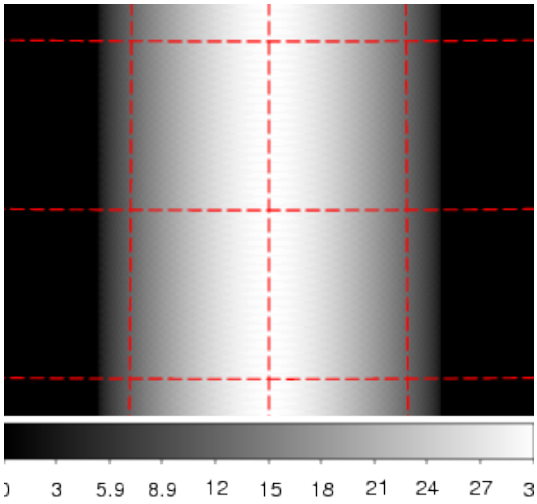


Figure 5.1.: Exposure map of the region of interest. Scan direction is upwards. The grid's width is 1/2 degrees.

All located at the center of the analyzed region, i.e. also located at the scan curve and so at the best position possible. The simulated sources are specified in table 5.1 (simulations 1-6).

The same absorbed power law as Schmid (2012) suggested for eRASS simulations is assumed for the spectrum of all simulated sources.
The results are shown in figure 5.2. Several similar simulation were done for sources with equal flux but different photon energies in the main energy range of eROSITA. All results were similar.

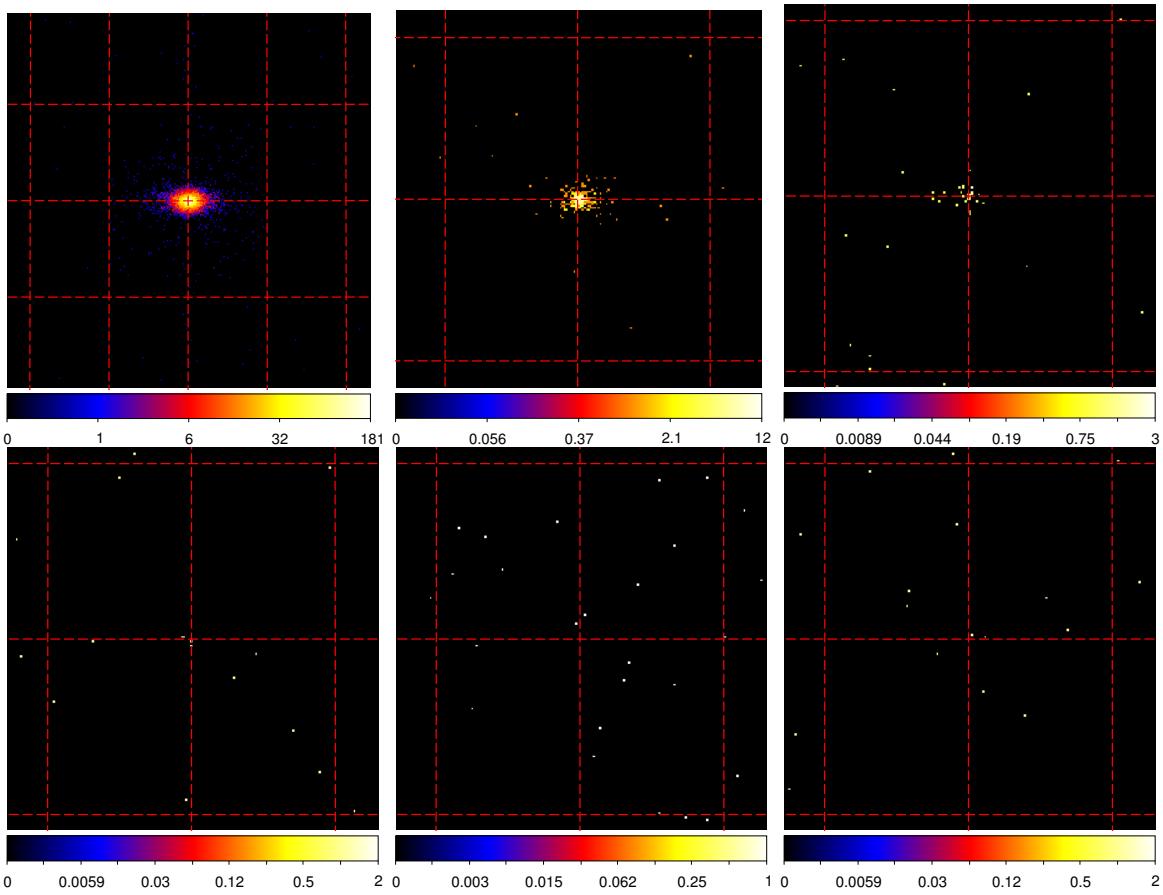


Figure 5.2.: Pixel maps of simulated sources, from top left to bottom right $10^{-9}, 10^{-11}, 10^{-12}, \ldots, 10^{-15}\, erg\, s^{-1}\, cm^{-2}$ The grid is 1/10 degree in all images. The position of the simulated source is always specifies at center and zoom was selected to cover the interesting areas.

Two further simulations were performed with sources with flux $10^{-12}\, erg\, s^{-1}\, cm^{-2}$. In the first, a source which is not at the perfect location was investigated (simulation 7 in table 5.1). In the seconds, two sources where involved (8a and 8b in table 5.1). The results are shown in figure 5.3.

Then the source detection tools were run for each image[1]. Unfortunately they were not able to handle the coordinate systems which are available as output for the simulations. It seems like ermldet does not care about the WCS FITS header at all and assumes some parallel orthographic projection for determining sky coordinates. As in the small area covered deformations because of the coordinate transformation are not significant,

[1]Only one single energy band was assumed.

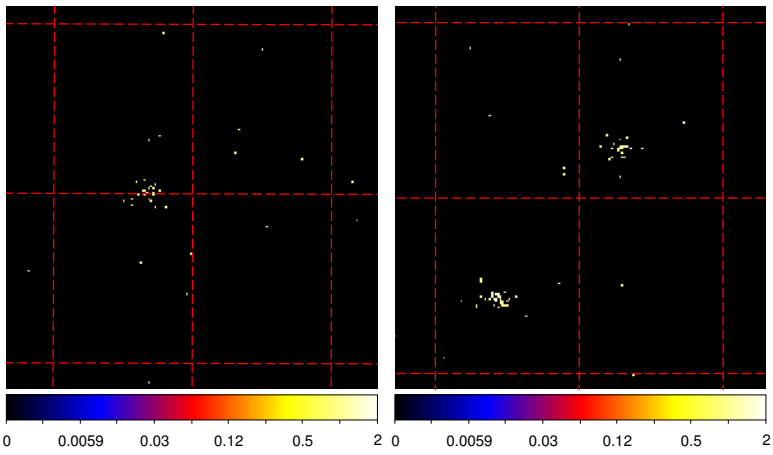


Figure 5.3.: Pixel maps of simulated sources. Left a source at 1/3 degrees outside the FOV center, right two sources located near to the FOV center. Grid spacing is 1/10 degrees, the centers of the images are at right ascension 34.7° resp. 35.0° and declination 35.0°

ermldet was just changed so that the reference point is at the image's center and the pixel directly represent angular distances. It is equal to the assumption that a sphere is locally nearly Euclidean. The used workaround leads to erroneous but suitable results for the following tests. In the future, the SASS tasks will generate the images for the source detection, and as the SASS tools are developed to work together, this problem will supposedly vanish.
Here it should be mentioned that for the eRASS the generated images should be either split for source detection or the projection should be wisely chosen. One eroday covers a whole great circle and the source detection is working on pixel level, so the pixels should not be deformed by the mapping.

Except for the simulated sources with flux lesser than $10^{-12}\,erg\,s^{-1}\,cm^{-2}$, it was possible to detect sources. That is a reasonable result according to the simulated images.
Some problems occurred:

- The brightest simulated source with a flux of $10^{-9} ergs^{-1} cm^{-2}$ was incorrectly separated which makes the following NRTA-S results meaningless for that source[2].
- The detected rates were equal to the fluxes all times and differed too much from the simulated ones. Maybe wrong telescope specifications were used in the selected source detection configuration. To although test the NRTA-S software, the source detection results were duplicated and adapted. Also, the reference catalog contains equal values for rates and fluxes.

Afterwards, the NRTA-S tools were run. Their and the source detection results are shown in table 5.3. Reference catalogs can be found in table 5.2, for the simulations, the first one was used. The source positions are the same as for the simulated sources. Source 1-7 correspond directly to the simulated sources, sources 8 and 9 are the simulated sources

[2]For one of the detected sources, an extend of 10.07227 was detected. The unit should be degrees, but it seems as the source detection provided it in pixels.

8a and 8b respectively. As the source detection returned the same value for flux and rate, they were set equal. The rate was calculated by the following way:

Schmid (2012) calculated an approximate reference photon count rate of $13500 s^{-1}$ for the often used standard candle Crab nebula when observed with eROSITA.
The rate is used here to estimate the source's rates which are needed to generate a reference catalog for testing purposes according to the source definitions in the SIMPUT files. The used catalog is presented in table 5.2, which also contains the simulation inputs.
The rate $r(f)$ corresponding to a source with flux f is then given by:

$$r(f) = 13500 \frac{f}{1\,crab} \tag{5.1}$$

The reference catalog for testing purposes was created to match the SIMPUT definitions, so it contains no errors, because a simulation of exactly known virtual sources is performed. In contrast, a real reference catalog will contain errors.

As already mentioned, the detected rates are strange. The NRTA-S results are reasonable, at least for most sources the identification already contains the expected source. Multiple identifications are caused to get the needed total rate.
Therefore, the rates were adapted to match the reference catalog. The results can be found in table 5.3, too.
As expected, the p-value is greater than any reasonable significance level and the identified sources always contain the reference source which belongs to the simulated source. That sometimes additional sources were assigned can be argued and ignored, because they have always a much lesser rate and so they contribute relatively nothing relevant, but because of the specified errors, it is possible that they are involved.
Although only one energy band was used for source detection, it was shown that the source identification will work, because the reference catalog contains multiple sources with different rates at the same spacial location. In principle this is the same as multiple source parts.

The result table also shows some additional testcases:

- *Testcase 1:* The scenario is that a unknown source was detected. The source is located a little bit away from the reference sources. The source identification correctly has correctly chosen the nearest reference source, but the hypothesis tests p-value clearly shows that this source can not explain the assumed detection.

- *Testcase 2:* The second reference catalog was used. The testcase shows what happens of two reference sources could not be resolved. Both reference sources are assigned and the hypothesis test confirms that assumption.

- *Testcase 3:* The third reference catalog was used. Here, a known source was revolved into two sources. The scenario is inverse to that of testcase 2.

Simulation	right ascension [deg.]	declination [deg.]	source flux $[erg\,s^{-1}\,cm^{-2}]$
1	35.000000	35.000000	10^{-9}
2	35.000000	35.000000	10^{-11}
3	35.000000	35.000000	10^{-12}
4	35.000000	35.000000	10^{-13}
5	35.000000	35.000000	10^{-14}
6	35.000000	35.000000	10^{-15}
7	34.666667	35.000000	10^{-12}
8a	34.943431	34.943431	10^{-12}
8b	35.028285	35.028285	10^{-12}

Table 5.1.: Simulated sources

Catalog	Source ID	flux = rate $[ergs^{-1}cm^{-2}$ resp. $1/s]$	right ascension [deg.]	declination[deg.]
1	1	$5.625 \cdot 10^{2}$	35.000000	35.000000
	2	$5.625 \cdot 10^{0}$	35.000000	35.000000
	3	$5.625 \cdot 10^{-1}$	35.000000	35.000000
	4	$5.625 \cdot 10^{-2}$	35.000000	35.000000
	5	$5.625 \cdot 10^{-3}$	35.000000	35.000000
	6	$5.625 \cdot 10^{-4}$	35.000000	35.000000
	7	$5.625 \cdot 10^{-1}$	34.666667	35.000000
	8	$5.625 \cdot 10^{-1}$	34.943431	34.943431
	9	$5.625 \cdot 10^{-1}$	35.028285	35.028285
2	10	0.5625 ± 0.0563	34.952 ± 0.01	34.942 ± 0.01
	11	0.5625 ± 0.0563	35.023 ± 0.01	35.026 ± 0.01
3	12	1.1250 ± 0.1125	34.988 ± 0.3	34.985 ± 0.3

Table 5.2.: Used reference catalogs. Sources with no errors specified have no errors. No source has an extend.

Simulation ID	RA [deg.]	DEC [deg.]	error RA/DEC [deg.]	rate = flux $[ergs^{-1}cm^{-2}$ resp. $1/s]$	rate error = flux error $[ergs^{-1}cm^{-2}$ resp. $1/s]$	identified reference source IDs	hypothesis test p-value
1	34.99969	34.99916	0.00364	340.16498	4.62950	-	0
	35.00285	34.99900	0.11246	101.12585	0.26956	-	0
	35.00185	35.17469	3.14270	0.25035	0.145101	4 ($\approx$ 100%), 5 ($\approx$ 100%)	0.10236
	34.99458	35.14449	6.28539	0.15384	0.014571	4 ($\approx$ 0%), 5 ($\approx$ 0%)	0.00012
2	34.99953	34.99939	0.02082	11.89464	0.99826	2, 9	0.00012
3	34.99977	34.99966	0.08509	1.08812	0.30204	3, 9	0.42256
7	34.72764	35.00160	0.11879	1.16602	0.36161	7, 8	0.37164
8	34.95261	34.94271	0.07225	1.52297	0.35916	3, 8	0.12184
	35.02390	35.02696	0.10709	0.86279	0.26562	4, 9	0.17816

Simulation ID	RA [deg.]	DEC [deg.]	error RA/DEC [deg.]	rate = flux $[ergs^{-1}cm^{-2}$ resp. $1/s]$	rate error = flux error $[ergs^{-1}cm^{-2}$ resp. $1/s]$	identified reference source IDs	hypothesis test p-value
2	34.99953	34.99939	0.02082	5.625	0.5625	2	0.4836
3	34.99977	34.99966	0.08509	0.5625	0.05625	3, 4	0.46904
7	34.72764	35.00160	0.11879	0.5625	0.05625	5, 7	0.32052
8	34.95261	34.94271	0.07224	0.5625	0.05625	6 (63.4%), 8 (100%)	0.45516
	35.02390	35.02696	0.10709	0.5625	0.05625	6 (36.6%), 9 (100%)	0.48264

Testcase	RA [deg.]	DEC [deg.]	error RA/DEC [deg.]	rate = flux $[ergs^{-1}cm^{-2}$ resp. $1/s]$	rate error = flux error $[ergs^{-1}cm^{-2}$ resp. $1/s]$	identified reference source IDs	hypothesis test p-value
1	34.80000	34.80000	0.08509	0.56250	0.05625	8	0.00888
2	34.98826	34.98483	0.30000	1.12500	0.11250	10, 11	0.67364
3	34.95261	34.94271	0.01000	0.56250	0.05625	12 ($\approx$ 50%)	0.42436
	35.02390	35.02696	0.01000	0.56250	0.05625	12 ($\approx$ 50%)	0.43068

Table 5.3.: Results of source detection and NRTA-S software for the simulations. The contribution of identified sources is specified in brackets if $< 100\%$. The first table shows the real detected sources, the second the same with adapted rates and the last one additional testcases.

6. Conclusions

A scientific near real-time analysis software for eROSITA measurements was proposed in this work.
In the final analysis of eROSITA's experimental data, humans will investigate it's scientific importance and the datasets will be widely complete. During the measurement process, especially the eRASS, the results will be researched too, but the time intervals in which new data will be examined are potentially too long for the recognition of events which should be payed attention to immediately. For example it may be useful to look at phenomenons with other telescopes, but if they are very short-term incidents it could be too late if there is no tool which notifies the monitoring scientists. And even if people are looking all the time on new data that would require an acceptable preparation and presentation of the information. Also it would be nice to have a software which provides support by filtering maybe important from less important data.
The software presented in this work is proposed for this purpose. It provides tools for the prediction of missing data and suggests concepts for the automatic execution of parts of the software used for the final analysis. So it is possible to have a quick-look on the measurements. For more comfort a web-based application was developed which allows users an easy access to pre-analyzed as well as raw data.
But the main challenge was to investigate algorithms for automatic detection of potentially interesting parts of measurements. Three types of interesting events were defined:

- The discovery of new sources or the detection of new spectral parts of sources
- The detection of long-term transients which were off for a long time
- An information gain by the resolution of a known source into multiple parts

To detect such events, detected sources are mapped to known sources in a reference catalog. This is done in several steps. Firstly, candidates are selected based on their angular distance from detected sources. Candidates which are unlikely at the same position as the detected ones are rejected.
Afterwards an algorithm was implemented which tries to refine the candidate selection for a given detected source. It basically minimizes the KL divergence between models for the candidates and the detected source. It also approximately determines how much reference source contribute to detected ones.
To test if the resulting source identification matches the measurement, hypothesis tests are performed with a Monte Carlo method.
At last, a tool was developed which performs a classification and rating of the results.
Simulations and testcases have shown that the used approaches should be applicable.
Besides for source detection, no raw data is used, so a simple usage of the source identification in other circumstances or for other purposes is possible.

Bibliography

J. Bentley. Multidimensional binary search trees used for associative searching. 1975.

H. Brunner. Preliminary design of the erosita sass, rev. 3 (unpublished). 2009.

H. Brunner. *Column layout of eROSITA FITS table source catalogues (internal document)*, 2012.

B. Carroll and D. Ostlie. *An Introduction to Modern Astrophysics*. 2007.

R.G. Cruddace, G.R Hasinger, and J.H Schmitt. The application of a maximum likelihood analysis to detection of sources in the rosat data base. 1988.

B. Dimitrov, D. Green, et al. On statistical hypothesis testing via simulation method. 2003.

I. George, K. Arnaud, B. Pence, et al. The calibration requirements for spectral analysis (definition of rmf and arf file formats). 2007.

F. Giovannelli and L. Sabau-Graziati. *The Impact of Space Experiments on Our Knowledge of the Physics of the Universe*. 2004.

J. Goldberger, S. Gordon, and H. Greenspan. An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. 2003.

C. Grinstead and J. Snell. *Introduction to Probability*. 2003.

C. Grossberger and M. Wille. erosita: Preprocessing of nrta software, design document (unpublished). 2010.

F. Guglielmetti. Background–source separation in astronomical images with bayesian probability theory. 2010.

J. Hershey and P. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. 2007.

R. Hogg, J. McKean, and A. Craig. *Introduction to Mathematical Statistics*. 2012.

H. Karttunen et al. *Fundamental Astronomy*. 2003.

S. Kullback and R. Leibler. On information and sufficiency. 1951.

D. Lee and C. Wong. Worst-case analysis for region and partial region searches in multidimensional binary search trees and balanced quad trees. 1977.

A. Merloni, P. Predehl, , et al. erosita science book: Mapping the structure of the energetic universe. 2012.

J. Neyman and K. Pearson. On the problem of the most efficient tests of statistical hypotheses. 1933.

R. Nowak. Lecture notes on statistical learning theory, lecture 13: Maximum likelihood estimation. 2009.

F.-X. Pineau, C. Motch, et al. Cross-correlation of the 2xmmi catalogue with data release 7 of the sloan digital sky survey. 2010.

C. Schmid. *X-ray Telescopes in the Digital Lab: Instrument Performance Simulations.* PhD thesis, 2012.

C. Schmid, R. Smith, and J. Wilms. Simput: A file format for simulation input. 2011.

S. Schwarzburg. Eine software zur echtzeitanalyse von experimentellen daten im flexible image transport system (fits). 2005.

F. Seward and P. Charles. *Exploring the X-ray Universe.* 2010.

J. Shlens. Notes on kullback-leibler divergence and likelihood theory. 2007.

H. Stöcker. *Taschenbuch der Physik.* 2004.

J. Trümper and G. Hasinger. *The Universe in X-Rays.* 2008.

I. Valtchanov, M. Pierre, and R. Gastaud. Comparison of source detection procedures for xmm-newton images. 2001.

S. Vinga. Convolution integrals of normal distribution functions. 2004.

W. Voges, B. Aschenbach, T. Boller, et al. The rosat all-sky survey bright source catalogue. 1999.

M. Wille. Detector performance of erosita. 2011.

J. Wilms, I. Kreykenbohm, et al. erosita: Near real time data analysis software design (unpublished).

A. Source catalogs

All source catalogs contain at least the columns specified in table A.1.

Column name	Datatype	Description
ID	Integer	Source ID
ID_BAND	Integer	Energy band ID
RA	Double	Right ascension
DEC	Double	Declination
RADEC_ERR	Double	Position error
EXT	Double	Source extend
EXT_ERR	Double	Source extend error
SCTS	Double	Source counts
SCTS_ERR	Double	Source counts error
RATE	Double	Count rate
RATE_ERR	Double	Count rate error
FLUX	Double	Source flux
FLUX_ERR	Double	Source flux error

Table A.1.: Table columns, mainly a subset of the output of ermldet

The table is expected to be in the first binary-table HDU of the FITS-file. Please note that SCTS and SCTS_ERR are not really used by the NRTA-S, but they are looped through the outputs. The different catalogs have to contain all that columns and some additional ones:

- *Source detection results* contain an additional integer column ID_INSTR specifying the ID of the telescope.
- The *reference and NRTA-S catalogs* have an integer column FLAGGED to specify that the detection of the source should cause an alert of type STATECHANGE. NRTA-S catalogs moreover save the FILENAME of the hypothesis tests output which caused a source to be inserted and this source's ID and ID_BAND fields in the columns DETECTION_ID and DETECTION_ID_BAND respectively to assure the traceability.
- The *classification catalog* contains a double-valued column RATING which should reflect the importance if such a source is newly discovered. Also a human-readable description should be stored in the string column DESCRIPTION.

Other column may be present, but they will be ignored.

B. Input parameters

Parameter	Datatype	Description
FILENAME_ORBIT	String	file containing the predicted orbit
T_START	Double	Modified Julian Date (MJD) specifying the begin of the time period for which the attitude will be predicted
T_END	Double	MJD specifying the end of the time period for which the attitude will be predicted
T_STEP	Double	Time interval in seconds between two predicted attitudes
INITIAL_RA	Double	known right ascension in degrees at time T_START
INITIAL_DEC	Double	known declination in degrees at time T_START
INITIAL_ROLL_ANGLE	Double	known roll angle in degrees at time T_START
ANGULAR_SPEED	Double	angular speed in degrees per second of the satellite's rotation

Table B.1.: Input parameters for the attitude prediction tool

Parameter	Datatype	Description
SOURCELIST_FILENAME	String	file containing the source detection's output
OUTPUT_FILENAME	String	output file's filename
REFERENCE_CATALOG	String	reference catalog filename
NRTA_CATALOG	String	NRTA-S catalog filename
DISTANCE_CUTOFF	Double > 0	angular distance cutoff
SIGNIFICANCE	$0 \leq$ Double ≤ 1	p-value cutoff

Table B.2.: Input parameters for the candidate selection tool

Parameter	Datatype	Description
FILENAME	String	input/output filename
MIN_MC_CALLS	Integer > 0	minimum number of Monte Carlo samples
MAX_MC_CALLS	Integer > 0	maximum number of Monte Carlo samples
MAX_COMBINATION_K	String	Maximum number of considered candidate sources

Table B.3.: Input parameters for the source identification tool

Parameter	Datatype	Description
FILENAME	String	input/output filename
MC_CALLS	Integer > 0	number of Monte Carlo samples

Table B.4.: Input parameters for the hypothesis testing tool

Parameter	Datatype	Description
INPUT_FILENAME	String	output of the hypothesis test tool
NRTA_CATALOG	String	NRTA-S catalog filename
ALERT_DATABASE	String	alert database filename
NEWSOURCE_RATING	Double	the basic rating of an alert of type NEWSOURCE
RESOLVED_RATING	Double	the basic rating of an alert of type RESOLVED
STATECHANGE_RATING	Double	the basic rating of an alert of type STATECHANGE
PLAUSIBLITY_RATING	Double	the basic rating of an alert of type PLAUSIBILITY
CLASSIFICATION_CATALOG	String	filename of the classification catalog
REJECTION_THRESHOLD	Double	maximum number of alerts per run which are still considered as plausible
KEEPING_RATING_THRESHOLD	Double	minimum rating of alerts which will be never rejected by the plausibility check
P_VALUE_CUTOFF	$0 \leq$ Double ≤ 1	Maximum p-value of the hypothesis tests which leads to the rejection of the null hypothesis
NOTIFICATION_SCRIPT	String	Path the the script which will be executed for each alert which was not filtered out

Table B.5.: Input parameters for the result classification tool

C. Filter specfication

Column name	Datatype	Description
RA_MIN	Double	minimum right ascension for the alert's detected source
RA_MAX	Double	maximum right ascension for the alert's detected source
DEC_MIN	Double	minimum declination for the alert's detected source
DEC_MAX	Double	maximum declination for the alert's detected source
E_MIN	Double	minimum energy for the alert's detected source
E_MAX	Double	maximum energy for the alert's detected source
RATE_MIN	Double	minimum rate for the alert's detected source
RATE_MAX	Double	maximum rate for the alert's detected source
RATING_CUTOFF	Double	maximum rating of the alert

Table C.1.: Filter table, the descriptions explain the criteria on the alert for rejection which all have to be matched for the activation of the filter.

D. Data products

Column name	Datatype	Description
t	Double	MJD specifying the time at which the attitude is valid
ra	Double	predicted right ascension in degrees at time t
dec	Double	predicted declination in degrees at time t
rollangle	Double	predicted roll-angle in degrees at time t

Table D.1.: Output table of the attitude prediction

Column name	Datatype	Description
MAIN_CATEGORY_ID	Double	ID of the main category of the event (0=NEW-SOURCE, 1=RESOLVED, 2=STATECHANGE, 3=PLAUSIBLITY)
MAIN_CATEGORY_DESCRIPTION	String	human-readable representation of MAIN_CATEGORY_ID
SUB_CATEGORY_ID	Integer	ID of the sub category of the event, equals the source ID in the classification catalog
SUB_CATEGORY_DESCRIPTION	String	human-readable description of the sub category as specified in the classification catalog
NRTA_CATALOG_SOURCE_ID	Integer	generated ID of the source in the NRTA-S catalog
RATING	Integer	the rating of the alert
NRTA_CATALOG_SOURCE_BAND_ID	Integer	energy band ID of the source
REJECTED	Integer	$\neq 0$ if the filter or plausibility
rollangle	Double	predicted roll-angle in degrees at time t

Table D.2.: Alert database table

Parameter name	Description
MAIN_CATEGORY_DESC	human-readable description of the alert's main category
SUB_CATEGORY_DESC	human-readable description of the alert' sub category
RATING	the rating of the alert
NRTA_CATALOG_SOURCE_ID	ID of the detected source in the NRTA-catalog
NRTA_CATALOG_SOURCE_BAND_ID	energy band ID of the detected source
SOURCE_RA	right ascension of the detected source
SOURCE_DEC	declination of the detected source
SOURCE_RADEC_ERR	positional error of the detected source
SOURCE_FLUX	flux of the detected source
SOURCE_FLUX_ERR	flux error of the detected source
SOURCE_RATE	rate of the detected source
SOURCE_RATE_ERR	rate error of the detected source

Table D.3.: Command line arguments for the notification script. They have all the form <parameter name>=<value> and are always in the sequence as presented here.

HDU	Column name	Datatype	Description
2 (DETECTED)	The same table format as specified in table A.1		
3 (REFERENCE_SOURCES)	The same table format as that one of the reference catalog		
4 (CANDIDATES)	SOURCE_ID	Integer	ID of the detected source
	SOURCE_BAND_ID	Integer	energy band ID of the detected source part
	CANDIDATE_ID	Integer	ID of the candidate source
	CANDIDATE_BAND_ID	Integer	energy band ID of the candidate source part
	PVALUE	Double	p-value of the candidate selection
	DISTANCE	Double	angular distance between the candidate and the detected source in degrees
4 (IDENTIFICATION)	SOURCE_ID	Integer	ID of the detected source
	SOURCE_BAND_ID	Integer	energy band ID of the detected source part
	CANDIDATE_ID	Integer	ID of the candidate source
	CANDIDATE_BAND_ID	Integer	energy band ID of the candidate source part
	WEIGHT	Double	assumed contribution of the identified source to the detected sources
5 (TEST)	SOURCE_ID	Integer	ID of the detected source
	SOURCE_BAND_ID	Integer	energy band ID of the detected source part
	PVALUE	Double	The p-value of the test

Table D.4.: Tables of the final data product of the hypothesis testing

Declaration

I hereby declare that this thesis is my own work and only the denoted resources were used.

Erlangen, September 25, 2013

Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Erlangen, 25.09.2013